Addressing challenges of molecular precision diagnostics for cancer patients in the genomics era



Pauline Robbe

Radcliffe department of Medical Sciences

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

St Cross College 2017

To my husband, Marc.

To my parents, family and friends.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work, except where specifically indicated in the text. This dissertation contains approximately 41,000 words and 250 pages.

Pauline Robbe

2017

Acknowledgements

Foremost, I would like to express my deepest thanks to my main supervisor Professor Anna Schuh for making me the scientist I am today. I am forever grateful for her guidance, generosity, her unparalleled knowledge and for letting me find my own path. She will remain a true inspiration and role model to me.

Thank you to Professor Francesca Buffa, my co-supervisor for her invaluable intellectual contributions to this thesis as well as her warm welcome in her group, her constant support and encouragement.

My special words of thanks go to Dr Ruth Clifford, my friend, my desk (and house) neighbour, for adopting me in the lab on day one, and teaching me all about CLL and so many other things in life, and to Dr Laura López Pascua for convincing me that this crazy idea to start a DPhil in Oxford was totally reasonable and being there every step of the way.

A special mention for my MSc mentors, Dr Jean-Claude Ourlin and Mr Jean-Luc Peiffer who taught me critical thinking, and shared with me their passion of science and music.

I would like to thank my library partners and fellow students Mimi Sheikh and Reem Alsolami, for keeping me motivated during the final months of writing.

I am sincerely thankful to all my labmates and colleagues (past and present) in the JR, the WTCHG and ORCRB, with special thanks to Hélène Dréau, Maite Cabes, Adele Timbs, Samantha Knight and Adam Burns for including me in the team six years ago, David Bruce

for our many fascinating conversations about science and more and Niamh Appleby for her invaluable comments on my thesis.

In addition, I would like to thank my bioinformatics colleagues for their patience: Dr Katie Ridout for her kindness, encouragements and great advices, Drs Pavlos Antoniou and Niko Popitsch for their continuous help from the start, showing me that bioinformatics was the answer to all my questions, and Dr Dimitris Vavoulis for always kindly putting up with me and my stats queries.

Drs Romain Guièze and Olivier Tournilhac for their constant enthusiasm and being great collaborators. Thank you to my all collaborators in Illumina Chesterford especially Drs Jennifer Becq, Miao He, Mark Ross and David Bentley for their helpful assistance on my work.

On a personal note, a massive thank you to Divija Jatavallabhula for her continued friendship, support, hugs, FaceTime chats and laughs. I have been lucky to be surrounded by many other intelligent and thoughtful people, my Oxford family and academic "big brothers and sisters" Drs Stéphanie Trouche, Emmanuel Perisse, Sebastian Bonilla, Ariel Gómez Díaz, Thomas Lessinnes and Johannes Noller. Thanks to my great friends Laura, Sophie and Jean-Roch, for reminding me there is a whole world outside of my DPhil.

Lastly, I would like to thank all my family, in particular my grandparents and my brother David for always being there for me despite the distance.

Marc my husband, for never letting me doubt myself, thank you for accompanying me on this adventure, I look forward to our next one!

And finally my parents for everything, I am forever grateful. You made me who I am.

Abstract

Genomic technologies have proven successful in research-led studies and their potential for characterising prognostic and predictive biomarkers in cancer patients has been recognised. However, such techniques and findings are yet to be applied in routine diagnostic environments. Indeed, many challenges remain before being able to establish a comprehensive genomics classification for all cancer patients, not least pre-analytical difficulties and variants interpretation. One of the most important pre-analytical challenges is the use of formalin-fixed paraffin-embedded (FFPE) tissues which are far more widely available in diagnostics compared with alternative fresh-frozen (FF) tissue samples. The key issue is that FFPE treatment of the tissue greatly impacts on the DNA quality and downstream analysis. This is important because genomic studies have discovered a significant number of mutations in multiple cancer driver genes. The presence of these mutations informs on patients' prognoses or responses to treatment and guides clinical management. With an increasing number of targeted therapies available it is critical to stratify patients using genomics technologies. However, current clinically available tests fail to identify known driver mutations in the majority of patients.

The overall objective of the work presented in this thesis was to address current challenges and limitations of FFPE testing by improving knowledge in genomic technology and by providing a means to offer genomic-guided precision medicine to a greater number of cancer patients. The specific aims were (i) to explore the feasibility of using FFPE tissue from

patients with solid tumour for clinical grade whole genome sequencing (WGS) and (ii) to use WGS and targeted sequencing data to search for novel genomic biomarkers of aggressive disease in CLL patients.

In Chapter 3, a prospective WGS study of cancer patients using 52 matching FF and FFPE samples was performed. This highlighted 71% agreement in somatic single nucleotide variants and revealed sub-optimal copy number alteration (CNA) detection from FFPE samples (44% median correlation with FF) due to non-uniform coverage. In Chapter 4, after pre-analytical investigation and experimental optimisation, CNA detection was improved significantly using lower reverse crosslinking temperature in FFPE DNA extraction (80°C or 65°C depending the methods) or by controlling formalin fixation. The results showed that WGS data derived from FFPE tissue derived DNA is capable of detecting 96% of clinically actionable variants (including 97% of CNAs) and therefore the approach is considered suitable for implementation in the routine clinical diagnostic environment.

Furthermore, the study of several clinical trial CLL cohorts presented in this thesis highlighted multiple novel genomic biomarkers of aggressive disease in CLL patients. In Chapter 5 relapse/refractory patients (R/R) were clustered according to the driver mutations identified using targeted sequencing; a multiple-hit profile defined as recurrent combinations of ≥ 2 mutations of *TP53*, *SF3B1* and *ATM* was identified. This multiple-hit profile, found in 19% of patients was associated with a median progression-free survival of 12 months compared with 22.5 months in the remaining patients (p=0.003). Other potential biomarkers were identified by exploring the genome of 320 CLL patients, such as mutational signatures.

Collectively, the work performed and the data generated provide new insights to overcome current challenges linked to the establishment of genomics technologies in the clinic in order to offer precision medicine to cancer patients.

Publications arising from this work

Chapter 3 and 4

Robbe P, Popitsch N, Knight SJL, Antoniou P, Becq J, He M, Kanapin A, Samsonova A, Vavoulis D, Ross MT, Zya Kingsbury Z, Cabes M, Ramos SDC, Page S, Dreau H, Ridout K, Jones L, Tuff-Lacey A, Henderson S, Mason J, Buffa FM, Verrill C, Maldonado-Perez D, Roxanis I, Collantes E, Browning L, Dhar S, Damato S, Davies S, Caulfield M, Bentley DR, Taylor JC, Turnbull C, Schuh A, On behalf of the 100,000 Genomes Project. Clinical whole genome sequencing from routine Formalin-fixed paraffin embedded (FFPE) specimens: pilot study for the 100,000 Genomes Project. Genetics in Medicine *(In press)*

Chapter 5

Guièze R*, **Robbe P*,** Clifford R, de Guibert S, Pereira B, Timbs A, Dilhuydy MS, Cabes M, Ysebaert L, Burns A, Nguyen-Khac F, Davi F, Véronèse L, Combes P, Le Garff-Tavernier M, Leblond V, Merle-Béral H, Alsolami R, Hamblin A, Mason J, Pettitt A, Hillmen P, Taylor J, Knight SJ, Tournilhac O, Schuh A. Presence of multiple recurrent mutations confers poor trial outcome of relapsed/refractory CLL. Blood. 2015; 126(18):2110-

Chapter 6

Clifford R*, **Robbe P***, Vavoulis D, Weller S, Timbs A, Burns A, Stamatopoulos B, Titsias M, Antoniou P, Cabes M, Blakemore S, Alsolami R, Hamblin A, Dreau H, Mason J, Henderson S, Howard DR, Chalmers A, Oates M, Pettitt A, Catovsky D, Strefford J, Yau C, Holmes C, Hillmen P, Taylor J, Knight SJL ‡, Schuh A‡. Towards Response Prediction using Integrated Genomics in Chronic Lymphocytic Leukemia: Results from frontline UK Clinical Trials. (*Submitted).*

Klintman J*, Barmpouti K*, Knight SJL, **Robbe P**, Dreau H, Clifford R, Ridout K, Burns A, Timbs A, Bruce D, Antoniou P, Sosinsky A, Becq J, Bentley D, Hillmen P, Taylor J, Caulfield M‡ and Schuh A‡. Clinical-grade Validation of Whole Genome Sequencing Reveals Robust Detection of Low-Frequency Variants and Copy Number Alterations in CLL. (*Submitted)*

* co-first authors

Contents

Abstract v

Publications arising from this work viii

Contents x

List of Figures xxi

List of Tables xxvii

Nomenclature xxix

Chapter 1 Introduction 1

1.1 Cancer is a disease of the genome 1

1.1.1 From phenotype to genotype 1

1.1.2 Cancer as an evolutionary concept 2

1.1.3 Cancer as a mutation driven disease 3

1.2 Massively parallel sequencing to investigate genomic mutations 6

1.2.1 Library preparation 9

1.2.2 Template generation approaches 11

1.2.3 Sequencing approaches 15

1.2.4 Bioinformatics analysis for resequencing strategies and somatic variant detection 20

1.2.5 Whole genome vs. targeted/exome approaches 24

1.2.6 Other techniques for genomes analysis 27

1.3 Genomic technologies revolutionised the cancer genetics field 29

1.3.1 The genomic landscape of cancers demonstrates inter-tumour heterogeneity 29

1.3.2 Intra-tumour heterogeneity expanded the cancer evolutionary concept 30

1.3.3 Characterisation of mutational mechanisms 32

1.4 Genomics era paving the way for precision medicine 34

1.5 Chronic Lymphocytic Leukaemia as a model for genomics studies 36

1.5.1 Existing genomic biomarkers 37

1.5.2 Discovery of novel biomarkers using MPS technology 40

1.6 MPS in the clinic 42

1.6.1 Current and future practices 42

1.6.2 Challenges to implement WGS in the clinic for cancer patients 43

1.6.3 Uncover molecular biomarkers to provide molecular diagnosis to all patients 45

1.7 Aims of the thesis 47

Chapter 2 Materials and methods 49

2.1 Thawing frozen peripheral blood mononuclear cells 49

2.2 Histopathology preparation from resected tissues 49

2.3 Extraction of genomic DNA using spin columns 50

2.3.1 From fresh frozen tissues 51

2.3.2 From formalin-fixed, paraffin-embedded tissues 51

2.3.3 From thawed peripheral blood mononuclear cells and buffy coat 54

2.3.4 From saliva samples 54

2.4 DNA quality control 55

2.4.1 Nanodrop 55

2.4.2 Qubit 56

2.4.3 Agarose gel electrophoresis 57

2.4.4 Real-time PCR assay 58

2.5 Sanger sequencing 62

2.6 Library preparation 63

2.6.1 Whole genome libraries 64

2.6.2 Illumina TruSeq Custom Amplicon preparation 69

2.6.3 Ampliseq Cancer panel 73

2.7 Massively parallel sequencing 75

2.7.1 Whole genome paired-end sequencing on HiSeq2500 76

2.7.2 Targeted paired-end sequencing on MiSeq 76

2.7.3 Targeted paired-end sequencing on PGM 77

2.8 Whole genome sequencing bioinformatics analysis 77

2.8.1 Reads mapping and metrics calculation 77

2.8.2 Somatic variant calling 79

2.9 Targeted massively parallel sequencing bioinformatics analysis 84

2.9.1 Illumina MiSeq analysis 84

2.9.2 Ampliseq cancer panel analysis 85

2.10 Single Nucleotide Polymorphism microarray 85

2.10.1 Illumina Genome-wide SNP platform hybridisations 86

2.10.2 Analysis pipeline with Nexus Copy Number 8.0 (BioDiscovery) 88

Chapter 3 Assessment of whole genome sequencing data from routine formalin-fixed, paraffin-embedded (FFPE) specimens 89

3.1 Introduction 89

3.2 Aims 90

3.3 Materials and methods 91

3.3.1 Patients and samples 91

3.3.2 Nucleic Acid Extraction 91

3.3.3 Sequencing 92

3.3.4 Bioinformatics analysis 92

3.3.5 Statistical analysis 93

3.3.6 Individual contributions 93

3.4 Results 94

3.4.1 A significant number of samples were lost to the study due to poor quality 94

3.4.2 FFPE DNA quality control revealed short and denatured DNA 94

3.4.3 FFPE DNA presented shorter fragments and data revealed non-uniform coverage 98

3.4.4 A different number of somatic SNVs and indels could be detected in FF and FFPE samples and the agreement was dependent on the variant calling algorithm used 102

3.4.5 Tumour heterogeneity and sampling heterogeneity explained differences between FF and FFPE samples 111

3.4.6 Somatic mutations in cancer driver genes were accurately detected in FFPE datasets 118

3.4.7 Copy number intensity signal was wavy in FFPE dataset 120

3.5 Limitations of the study 124

3.6 Discussion 126

3.6.1 Prospective FF tissue collection confirmed challenging in routine clinical testing 126

3.6.2 Confidence in somatic SNV and indel detection in coding and specific non-coding regions 127

3.6.3 FFPE preparation influenced the poor sequencing coverage signal and CNA detection 130

3.7 Conclusions 130

Chapter 4 Optimisation of FFPE sample quality for whole genome sequencing 133

4.1 Introduction 133

4.2 Aims and objectives 135

4.3 Material and methods 136

4.3.1 Patients and samples 136

4.3.2 DNA extraction 138

4.3.3 Sequencing 140

4.3.4 Bioinformatics 141

4.3.5 Variant annotation and reporting 141

4.3.6 Statistical analysis 143

4.3.7 Individual contributions 144

4.4 Results 144

4.4.1 No FFPE QC metrics predicted sequencing quality outcome 144

4.4.2 Optimisation of FFPE DNA extraction improved somatic variant detection 146

4.4.3 Investigation of tissue handling effects on sequencing metrics highlighted key preparation steps 153

4.4.4 New FFPE tissue preparation method did not benefit from optimised DNA extraction protocol 159

4.4.5 WGS data from optimised FFPE DNA samples generated comparable clinical reports to matching FF samples 165

4.5 Limitations of the study 171

4.6 Discussion 172

4.6.1 Different combinations of parameters lead to different types of sequencing artefact 172

4.6.2 Protocol to generate high quality FFPE specimens and detect CNAs 173

4.6.3 Clinical WGS reporting of FFPE samples 174

4.6.4 Conclusions 174

Chapter 5 Multiple recurrent mutations confer poor trial outcome of relapsed/refractory CLL patients 177

5.1 Introduction 177

5.2 Aims and objectives 179

5.3 Material and methods 179

5.3.1 Patients and samples 179

5.3.2 Extraction of genomic DNA from peripheral blood mononuclear cells and quality control 181

5.3.3 TruSeq Custom Amplicon targeted library preparation 181

5.3.4 MiSeq massively parallel targeted sequencing 181

5.3.5 Bioinformatics pipeline 181

5.3.6 Single nucleotide polymorphism microarray 182

5.3.7 Statistical analysis 182

5.3.8 Individual contributions 183

5.4 Results 183

5.4.1 Target sequencing coverage was acceptable for variants analysis 183

5.4.2 Reads re-alignment allowed the detection of clinically relevant variants 184

5.4.3 Mutational landscape revealed recurrently mutated genes and associations between alterations 186

5.4.4 Mutated genes distribution revealed recurrent associations, or multiple hit profile (MH-profile) 188

5.4.5 Patients harbouring the MH-profile displayed a significant poor response to treatment 189

5.4.6 MH-profile was found to be a prognostic marker of shorter survival 191

5.4.7 MH profile preserved its prognostic impact in *TP53* disrupted patients 194

5.4.8 Recurrent patterns of clonal evolution were discovered 195

5.5 Limitations of the study 201

5.6 Discussion 203

5.6.1 Mutation frequency comparisons highlighted similarities and differences with other CLL cohorts 203

5.6.2 Specific genes showed an impact on prognosis 205

5.6.3 Clonal evolution was heterogeneous in the cohort 206

5.7 Conclusion 207

Chapter 6 WGS in the clinic identifies CLL patients with high risk of relapse 209

6.1 Introduction 209

6.2 Aims and objectives 212

6.3 Material and methods 213

6.3.1 Patients and samples 213

6.3.2 Extraction of genomic DNA from peripheral blood mononuclear cells and quality control 214

6.3.3 MiSeq massively parallel targeted sequencing 215

6.3.4 PCR-free whole genome sequencing 215

6.3.5 Single nucleotide polymorphism microarray 215

6.3.6 Bioinformatics pipeline 216

6.3.7 Statistical analysis 216

6.3.8 Individual contributions 216

6.4 Results 217

6.4.1 Genomic heterogeneity in FCR non responders 217

6.4.2 Acceptable alignment metrics in tumour samples and lower alignment rate in germline samples 219

6.4.3 A subset of somatic variants was validated by targeted sequencing 222

6.4.4 Genomic landscape of pre-treatment and relapse refractory CLL 223

6.4.5 Operative mutational signatures 226

6.4.6 Structural alteration as a measurement of genomic complexity 228

6.4.7 Exploration of *IgHV* mutational status and genomic features 229

6.5 Limitation of the study 233

6.6 Discussion 234

6.6.1 Investigation of FCR responder predictors by targeted sequencing and SNP array 234

6.6.2 Investigation of more predictive markers in CLL using WGS 235

6.6.3 Surrogate marker to *IgHV* status 237

6.7 Conclusions 237

Chapter 7 Discussion 239

7.1 Introduction 239

7.2 Summary of the findings 240

7.3 The use of MPS technology in clinical practice 241

7.4 Intra- and inter-tumour heterogeneity in cancer 243

7.5 Mutational markers and genomics complexity linked with poor-prognosis in CLL 245

7.6 Technical consideration 246

7.7 Future work and directions 246

7.8 Conclusions 247

References A-1

Appendix A Overview of patients' characteristics A-56

Appendix B Coverage metrics for all samples B-58

Appendix C Description of the scopes used to investigate genome data C-63

Appendix D Pairwise comparison of variant overlap in different tissues D-64

Appendix E Validation of somatic SNVs and indels detected in WGS data using Ion AmpliSeq Cancer Hotspot Panel v2 ..... E-66

Appendix F Different conditions applied to the FFPE tissue preparation F-70

Appendix G DNA quality control results for FF, FFPE and blood samples G-75

Appendix H Log_2R plots from Nexus of whole genome for experimental FFPE samples ..H-77

Appendix I Clinical report of nine cancer cases I-80

Appendix J Univariate analyses for complete response rate (CRR), overall response rate (ORR), progression-free survival (PFS) and overall survival (OS).....J-89

Appendix K Mutational signatures in different CLL groups K-91

List of Figures

Figure 1.1 Model of clonal evolution in neoplasia (adapted from Nowell 1976).....2

Figure 1.2 Transitions and transversions base patterns. 4

Figure 1.3 Overview of the general steps of molecular biology and computational analysis required for MPS.....8

Figure 1.4 Main library preparation approaches used in cancer genomics projects, adapted from (Illumina Inc. 2015; Illumina Inc. 2017b).....12

Figure 1.5 Most common methods of template generation approaches with amplification step (adapted from (Goodwin et al. 2016)) 13

Figure 1.6 Sequencing by synthesis approaches: cyclic reversible termination (CRT) (top) and single-nucleotide addtion (SNA) (bottom) adapted from (Goodwin et al. 2016).....19

Figure 1.7 Flow diagram illustrating the MPS bioinformatics pipeline 21

Figure 1.8 Slow progression of Chronic Lymphocytic leukaemia, from (Guièze & Wu 2015). 36

Figure 1.9 Different types of DNA damages cause by FFPE tissue preparation (from (Do & Dobrovic 2014)).....44

Figure 1.10 Predictive biomarkers in CLL from (Rossi et al. 2017) 46

Figure 3.1 Patient dropout at different quality control steps 96

Figure 3.2 DNA integrity quality control 98

Figure 3.3 Distribution of sequencing metrics 99

Figure 3.4 Sequencing depth statistics 101

Figure 3.5 Correlation between pathologist assessment of proportion of invasive tumour in the sample and purity assessment by WGS data 102

Figure 3.6 SNVs and indels detection in FF and FFPE samples 103

Figure 3.7 Somatic variant overlap between FF and FFPE samples 104

Figure 3.8 Distribution of the number of high-risk variants for FF and FFPE samples using different variant callers 106

Figure 3.9 Overlap of somatic single nucleotide variants between FF and FFPE samples in different genomic regions. 107

Figure 3.10 Absolute count of SNVs per sample, per tissue types for FF and FFPE samples 108

Figure 3.11 SNV comparison and indel comparison between matching FF and FFPE samples according to the sample tissue type 109

Figure 3.12 Correlation of the proportion of C>T base substitutions between FF and FFPE samples 110

Figure 3.13 Somatic SNVs overlap between FF and FFPE samples 112

Figure 3.14 Estimation of the tumour sample purity illustrated by the distribution of the allelic fraction of FF and FFPE samples. 114

Figure 3.15 Evolution of the proportion of overlapping SNVs between FF and FFPE samples for different AF thresholds.....115

Figure 3.16 SNV overlap when underpowered variants are filtered out. 116

Figure 3.17 Correlation between proportion of FF-FFPE SNV overlap and median allelic fractions of FFPE somatic SNVs 117

Figure 3.18 Somatic SNVs and indels comparison between FF and FFPE samples in cancer related genes.....119

Figure 3.19 Log_2R and B-allele frequency across chromosome 6 for case 039 121

Figure 3.20 Correlation of Log_2R calculated with BIC-seq between FF and FFPE samples of six representative samples.....122

Figure 3.21 Distribution of Spearman's rank correlation coefficient of FF Log_2R and FFPE Log_2R 123

Figure 3.22 Number of somatic copy number alterations detected by BIC-seq 124

Figure 4.1 Variant classification decision process for clinical reporting.....143

Figure 4.2 Spearman correlation matrixes between sequencing metrics and DNA quality control. 146

Figure 4.3 Evolution of the proportion of variant overlap between matching initial and experimental FFPE samples.....147

Figure 4.4 Comparison of five alignment metrics in five patients between FF samples experimental FFPE and initial FFPE samples 148

Figure 4.5 FFPE DNA extraction optimisation effect depth of coverage.....150

Figure 4.6 CNA detection improvement using different DNA extraction conditions from the same FFPE sample 151

Figure 4.7 Distribution of spearman correlation coefficient of FF Log_2R and FFPE Log_2R 152

Figure 4.8 DNA quality control for different preparation conditions 155

Figure 4.9 Sequencing metrics for different preparation conditions 156

Figure 4.10 Effect size of FFPE preparation parameters on sequencing metrics when taking all parameters in consideration 158

Figure 4.11 FFPE DNA quality control comparison of four preparation methods 160

Figure 4.12 Alignment metrics for the FF and matching FFPE samples 162

Figure 4.13 Log_2R across all chromosomes for five samples of patient 402 164

Figure 4.14 Spearman correlation coefficient values of log_2R from FF and matching FFPE samples 165

Figure 4.15 Clinically relevant variants reported in the nine patients 167

Figure 4.16 Comparison of allelic fractions of clinically relevant variants in FF and matching FFPE samples 169

Figure 5.1 *TP53* deletion of 29 bp visualised in Integrative Genomics Viewer (IGV) 185

Figure 5.2 Type of mutations (left) and number of patients with (right) for each gene 186

Figure 5.3 Distribution of mutations across *TP53*, *SF3B1*, *ATM* and *NOTCH1* protein structure 187

Figure 5.4 Patient clustering according to gene mutations 189

Figure 5.5 Impact of the multiple hit profile on progression free survival 192

Figure 5.6 Impact of the multiple-hit profile on overall survival 193

Figure 5.7 Impact of the multiple-hit profile on response (left) and survival (right) in the *TP53* disruption subgroup.....195

Figure 5.8 Cancer cell fraction estimated for mutations reported in cluster 3 patients. 196

Figure 5.9 Clonal evolution pattern 1: *ATM* and *SF3B1* 198

Figure 5.10 Clonal evolution pattern 2: *ATM* first, *TP53* second.....199

Figure 5.11 Clonal evolution pattern 3: *ATM* first, *SF3B1* second.....199

Figure 5.12 Clonal evolution pattern 4: *SF3B1* first, *TP53* second.....200

Figure 5.13 Likely subclonal composition of pre-treatment samples in MH profile patients201

Figure 6.1 Proportion of patients carrying copy number changes, SNVs or indels in known CLL genes/cytobands.....218

Figure 6.2 Distributions of alterations in FCR non responder patients (n=56). 219

Figure 6.3 Median sequencing coverage for 320 germline samples from saliva and 320 tumour samples from peripheral blood 220

Figure 6.4 Histogram showing the distribution of the proportion of reads aligned to the reference human genome (mapping rate (%)) of GL and tumour samples.....221

Figure 6.5 Correlation between variant allelic fractions of mutations detected by targeted MPS and WGS for all variants investigated (left, n=79) and low VAF variants only (right, n=43). 222

Figure 6.6 Copy number alterations validation by FISH and by SNP array.....223

Figure 6.7 Number of patients with functional mutation.....224

Figure 6.8 Proportion of patients with functional mutation in different groups.....225

Figure 6.9 Number of patient with CNAs in different groups.....226

Figure 6.10 Proportion of the major mutational signatures in six CLL groups.....227

Figure 6.11Structural variants investigation in the different groups 229

Figure 6.12 Proportion of the major mutational signatures in *IgHV* mutated vs. unmutated 230

Figure 6.13 Correlation between the proportion of signature 9 and percentage of sequence homologous to germline (defined by Sanger sequencing).....231

Figure 6.14 Putative regions of Kataegis per patient and per group.....232

Figure 6.15 SNV density in genomic regions of the Ig loci, described regions of kataegis..233

List of Tables

Table 1.1 List of MPS technologies available 7

Table 1.2 Comparison of characteristics of WGS, WES and targeted sequencing for somatic variant investigations 25

Table 2.1 Input DNA quantity requirements for FFPE library preparation 60

Table 2.2 DNA standard concentrations of real-time PCR Kapa Library Quantification kit (KK4835, KapaBiosystems, Woburn, MA, USA) 61

Table 2.3 CLL targeted sequencing panel 69

Table 2.4 Targeted sequencing variants classification 85

Table 3.1 Comparison of DNA metrics in FF and matching FFPE tissue datasets 97

Table 3.2 Sensitivity and PPV in SNVs and indels using four different detection methods . 105

Table 3.3 Validation of SNVs and indels detected by WGS using the AmpliSeq Cancer Hotspot Panel v2 119

Table 4.1 DNA and sequencing quality control metrics for five FFPE samples 136

Table 4.2 FFPE preparation and DNA extraction conditions for experimental validation ... 137

Table 4.3 Conditions applied to the FFPE tissue specimens to optimise DNA extraction 139

Table 4.4 Patients' diagnoses 166

Table 4.5 SNV, indel and copy number alteration overlap of clinical report between FF and FFPE 170

Table 5.1 Patient characteristics 180

Table 5.2 Sanger sequencing validation of *TP53* mutations 184

Table 5.3 Impact of the multiple-hit profile on response by univariate analysis 190

Table 5.4 Impact of the multiple-hit profile on survival by univariate analysis 191

Table 5.5 Multivariate analysis for progression-free survival (treatment adjusted) 192

Table 5.6 Multivariate analyses for overall survival (treatment adjusted) 194

Table 6.1 Patients groups 214

Table 6.2 Distribution of WGS alignment metrics in tumour and germline samples 221

Table 0.1 ..... I-80

Nomenclature

°C	degree Celsius
1KGP	1000 genomes project
ABL	Abelson murine leukemia
AF	allelic fraction
AID	activation-induced cytidine deaminase
APC	Adenomatous polyposis coli
APOBEC	apolipoprotein B mRNA editing enzyme catalytic polypeptide
ARID1A	AT-Rich Interaction Domain 1A
ATAC-seq	assay for transposase-accessible chromatin using sequencing
ATM	ATM Serine/Threonine Kinase
ATR	ataxia telangiectasia and Rad3-related
BAF	B-allele frequency
BAM	binary alignment map
BAP1	BRCA1 Associated Protein 1
BCL	binary basecall
BCR	breakpoint cluster region
BGI	Beijing Genomics Institute
BIRC3	Baculoviral inhibition of apoptosis repeat containing 3
bp	base pair
BRCA1	breast cancer 1
BRCA2	breast cancer 2
BRIP1	BRCA1 interacting protein C-terminal helicase 1
CCF	cancer cell fraction
CDH1	Cadherin 1
CDKN2A	cyclin-dependent kinase Inhibitor 2A
CHEK	checkpoint kinase
ChIP-Seq	chromatin immunoprecipitation followed by sequencing

CIT	chemo-immunotherapy
CLL	chronic lymphocytic leukaemia
CMOS-IFSET	complementary metal-oxide semiconductor - ion-sensitive field-effect transistor
CNA	copy number alteration
cnLOH	copy neutral loss of heterozygocity
COSMIC	Catalogue Of Somatic Mutations In Cancer
CR	Colorectal
CRR	complete response rate
CRT	cyclic reversible termination
CSV	comma-separated values
Ct	cycle threshold
ddNTP	dideoxynucleotide
DDR	DNA damage response
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic
dNTP	deoxynucleotide triphosphate
dsDNA	double-stranded DNA
EDTA	Ethylene diamine tetra acetic acid
EGFR	Epidermal growth factor receptor
EGR2	Early Growth Response 2
ENCODE	Encyclopaedia of DNA elements
EPCAM	Epithelial Cell Adhesion Molecule
ERBB3	Erb-B2 Receptor Tyrosine Kinase 3
ERCC3	ERCC Excision Repair 3, TFIIH Core Complex Helicase Subunit
ERIC	European research initiative on CLL
FANCA	Fanconi Anemia Complementation Group A
FANCI	Fanconi Anemia Complementation Group I
FBXW7	F-Box And WD Repeat Domain Containing 7
FCR	Fludarabine, Cyclophosphamide, Rituximab
FF	fresh frozen
FFPE	formalin-fixed-paraffin-embedded
FGFR	Fibroblast Growth Factor Receptor
FISH	Fluorescence *in situ* hybridisation
g	gram
Gb	Gigabyte
GC	genomic complexity
gDNA	genomic DNA
GIAB	Genome in a bottle
GL	germline
GRCh	Genome Reference Consortium human

HCl	Hydrogen chloride
HDAC2	Histone Deacetylase 2
HER2	human epidermal growth factor receptor 2
HGF	Hepatocyte Growth Factor
IAP	index amplification plate
ICGC	The International Cancer Genome Consortium
Ig	immunoglobulin
IgHV	immunoglobulin heavy chain variable region
IGV	Integrative Genomic Viewer
indel	insertion and deletion
ISP	Ion sphere particles
JAK2	Janus Kinase 2
kb	kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIF5B	Kinesin Family Member 5B
KMT2C	Lysine Methyltransferase 2C
KRAS	KRAS Proto-Oncogene, GTPase
LASSO	Least Absolute Shrinkage and Selection Operator
Log2R	Log2 Ratio
LOH	loss of heterozygocity
Mb	megabase
MBL	monoclonal B-cell lymphocytosis
MED12	Mediator Complex Subunit 12
mg	milligram
MH	multiple hit
ml	millilitre
MLH	MutL Homolog
mM	milimolar
MMR	mismatch repair
MOA	mechanism of action
MPS	massively parallel sequencing
MRE11	MRE11 Homolog, Double Strand Break Repair Nuclease
MSH2	MutS Homolog 2
mTOR	Mechanistic Target Of Rapamycin Kinase
MYD88	Myeloid differentiation primary response 88
NaCl	Sodium Chloride
NaOH	sodium hydroxide
NBF	neutral buffered formalin
NBN	Nibrin
ng	nanogram

NGS	next-generation sequencing
nm	nanometre
NS	normal saline
OCT	optimal cutting temperature
ORR	overall response rate
OS	overall survival
PAL	pooled amplicon library
PALB2	Partner And Localizer Of BRCA2
PARP1	Poly(ADP-Ribose) Polymerase 1
PAX5	paired box 5
PBS	Phosphate buffered saline
PCR	polymerase chain amplification
PE	paired-end
PF	pass filter
Pfam	protein families database
PFS	progression free survival
PGM	personal genome machine
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
pM	picomolar
PMKB	Precision Medicine Knowledgebase
POT1	Protection Of Telomeres 1
PPV	positive predictive value
PTEN	Phosphatase And Tensin Homolog
QC	quality control
qPCR	quantitative polymerase chain reaction
R/R	relapse refractory
RNA	Ribonucleic acid
RPS15	ribosomal protein S15
RTA	real-time analysis
SAM	Sequence Alignment Map
SAMHD1	SAM And HD Domain Containing Deoxynucleoside Triphosphate Triphosphohydrolase 1
SBS	sequencing by synthesis
SD	standard deviation
SDS	Sodium dodecyl sulfate
SETD2	SET Domain Containing 2
SF3B1	splicing factor 3B subunit 1
SIFT	sorts intolerant from tolerant
SMRT	single-molecule real time
SMT	somatic mutation theory
SNA	single-nucleotide addition

SNP	Single nucleotide polymorphism
SNV	single nucleotide variant
SP	sequencing primer
SSC	Saline-sodium citrate
ssDNA	single-stranded DNA
SV	structural variant
TCGA	The cancer genome atlas
TERT	telomerase reverse transcriptase
TP53	Tumor Protein 53
TSS	transcription start site
UTR	untranslated region
UV	Ultraviolet
VAF	variant allelic fraction
VCF	variant call format
VEGF	vascular endothelial growth factor
VEP	variant effect predictor
VHL	Von Hippel-Lindau Tumor Suppressor
WES	whole exome sequencing
WGBS	Whole genome bisulfite sequencing
WGS	whole genome sequencing
WT	wild type
XPO1	Exportin 1
ΔCq	deltaCq
µg	microgram
µl	microlitre
µM	micromolar

Chapter 1 Introduction

1.1 Cancer is a disease of the genome

1.1.1 From phenotype to genotype

There are many differences between normal and tumour tissues. Cancer cells are capable of excessive proliferation leading to altered organisation of the tissue as the cells replicate uncontrollably. Furthermore, tumour cells are immortal, evading signals from growth suppressors, cell repair and cell death. As a result, they have a self-induced proliferation and invasive behaviour (Hanahan & Weinberg 2000). In order to survive and proliferate, cancer cells are required to interact with their microenvironment (De Palma et al. 2017). Solid tumours require oxygen and nutrients which are supplied through the blood. Under hypoxic conditions, tumour cell can induce angiogenesis by secreting vascular endothelial growth factor (VEGF) (Knudson 1971; Potente et al. 2011). Multiple pathways, including the VEGF pathway are commonly deregulated in cancer to mediate signals of proliferation in the absence of the correct stimulus (Lemmon & Schlessinger 2010), and the cause of this deregulation is alterations in the DNA sequence.

The somatic mutation theory (SMT) first linked the phenotype of a cell to its genotype in the context of carcinogenesis (Boveri 2008). Historical studies have firstly observed large tissue specific karyotype changes, such as trisomy 12 in chronic lymphocytic leukaemia

(CLL) (Gahrton et al. 1980) or t(14;18)(q32;q21) in follicular lymphoma (Yunis et al. 1984), or more commonly the t(9;22) translocation in chronic myeloid leukaemia (Nowell & Hungerford 1960).

Whilst emerging theories postulated that the acquisition of somatic mutation is a late event during carcinogenesis (Brücher & Jamall 2016), the SMT largely remains the basis of modern cancer research where genomic instability is the primary event (Loeb 1991).

1.1.2 Cancer as an evolutionary concept

Cancer is characterised by genome instability leading to the accumulation of mutations over time. The first theory of cancer evolution was postulated by Nowell in 1976 (Nowell 1976) (Figure 1.1). This theory uses the well described Darwinian model on the evolution of species postulated more than 100 years before (Piyasatian et al. 2008).

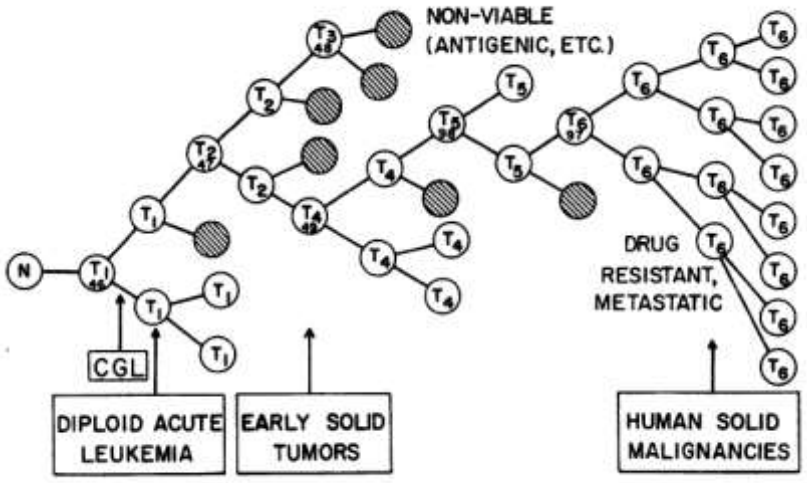


Figure 1.1 Model of clonal evolution in neoplasia (adapted from Nowell 1976)

In his model, Nowell describes that all cancers arise from a single clone initiating the cancer. By successive cell division, this initial clone gives rise to several subclones that acquired several different chromosomal alterations or mutations. Most of these subclones actually have a metabolic or immunologic disadvantage because these newly acquired mutations do not drive the tumour progression. However occasionally, one subclone obtains a mutation giving it a selective advantage within the microenvironment and cells originating from this clone proliferate. The main sources of selective pressure in the microenvironment (Sidow & Spies 2015) include nutrient availability, the immune system and geographic barriers.

This model has been demonstrated in colorectal cancer (Fearon & Vogelstein 1990; Kinzler & Vogelstein 1996). In this disease the accumulation of genomic alterations can be shown to be early events such as mutations in the gene *APC*, whilst some are required for the tumour to evolve from the early to late adenocarcinomas, such as *KRAS* mutations, and others are late events in the evolution to carcinomas like *TP53* mutations. This accumulation of genomic alterations in the cancer genome drives disease progression.

1.1.3 Cancer as a mutation driven disease

The main characteristic of cancer cells is their genomic instability (Yates & Campbell 2012; Burrell et al. 2013). Mutations are defined as somatic when they are specific to a cell population. This is in contrast to inherited mutations which are called germline (GL) (Stratton et al. 2009). Cancer cells have acquired additional mutations that are unique to them. Exogenous sources (such as ultraviolet (UV) light, chemicals and ionizing radiations), endogenous sources (such as reactive oxygen species or aldehydes) or mitotic

errors can cause genomic instability. These mutations are sometimes not repaired or incorrectly repaired and accumulate in the genome of the cells (Martincorena & Campbell 2015). The rate of mutation accumulation differs between different cancers types, with juvenile cancers and leukaemia having the lowest rates and skin cancers the highest mutation rates (Alexandrov et al. 2013; Lawrence et al. 2013; Chalmers et al. 2017). In addition, somatic mutations can be characterised by their prevalence within the cells of a tumour. Clonal mutations are present in all cells of the cancer cells whereas mutations existing in a minority of cancer cells are referred as subclonal.

There are two main classes of mutations types: base substitutions and structural variants (SVs) (Meyerson et al. 2010). Nucleotide substitutions are small genomic events where one nucleotide is substituted by another, namely single nucleotide variants (SNVs) or where one or several nucleotides are inserted or deleted (indel). SNVs are the most common type of alterations. They can be transitions when the base change replaces one purine with another (A>G or G>A), or one pyrimidine with another (C>T or T>C); or transversions when the base change replaces a purine with a pyrimidine or a pyrimidine with a purine (Figure 1.2).

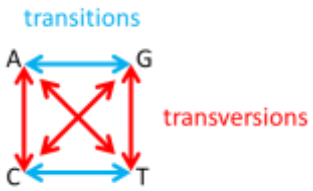


Figure 1.2 Transitions and transversions base patterns.

A: Adenine, G: Guanine, C: Cytosine, T: Thymine

SVs compose the second class of variants, which affect the genome at a chromosomal scale. These can be inversions (where the sequence is inversed on the chromosome), copy number alterations (CNAs) (such as deletions or amplifications of a segment or the whole of a chromosome), or translocations where parts of one or several chromosomes are rearranged, sometimes leading to fusion genes, formed from two distinct genes.

Occasionally, a few somatic mutations provide a selective advantage to the cancer cells and have been positively selected as they initiate tumour growth or progression. These are called driver mutations (Stratton et al. 2009). Conversely, the majority of somatic mutations do not contribute to the cancer initiation or progression. They are described as having no functional consequence and are called passenger mutations (Stratton et al. 2009). Of note, the word driver is also used to refer to a gene carrying driver mutations (Vogelstein et al. 2013).

Some genes are known to have mutations which are found predominantly in a specific domain, amino acid or nucleotide, namely hotspot mutations. Early sequencing studies have identified some of them, such as KRAS codon 12 and 13 in various cancers (Tsuchida et al. 1982; Kranenburg 2005), *BRAF* V600E mutations in melanoma and hairy cell leukaemia (Davies et al. 2002), *EGFR* L858R mutations in non–small cell lung cancer (Paez 2004; Lynch et al. 2004), *JAK2* V617F in myeloproliferative neoplasms (Levine et al. 2005), and *PIK3CA* mutations which are prevalent in the helical domain, mainly H1047R (exon 9), and the kinase domain, mainly E545K (exon 20), in various cancers (Samuels 2004; Bachman et al. 2004). Other genes do not have a specific hotspot. The most illustrative example is the gene *TP53* where mutations are distributed throughout the coding sequence, especially in the DNA binding domain of the protein (Hollstein et al.

1991; Petitjean et al. 2007). Some of these driver genes, such as *TP53* are frequently mutated in many types of cancers whereas other are specifically mutated in one type of cancer.

These driver genes are classified as either oncogenes (*BRAF*, *PIK3CA*, *EGFR*, and JAK2 (Davies et al. 2002; Samuels & Waldman 2010; Lynch et al. 2004; Kralovics et al. 2005; Baker et al. 1989)) or tumour suppressor genes (TP53 and PTEN (Li 1997; Steck et al. 1997)). Generally, tumour suppressor genes negatively regulate the cell cycle promoting proliferation and/or stimulate apoptosis responses (Knudson 2001). Some tumour suppressor genes are haploinsufficient: their mutation patterns follow the "two-hit hypothesis*"* (Knudson 1971). In this model, a bi-allelic alteration of the gene is necessary to affect the cell function. Other tumour suppressor genes are haplosufficient, and only require one hit. Oncogenes on the other hand, only require one activating mutation to deregulate its pathway. Proteins coded by oncogene induce signal of cell survival through growth, differentiation or proliferation (Croce 2008).

1.2 Massively parallel sequencing to investigate genomic mutations

The Human Genome Project (Venter et al. 2001; Lander et al. 2001; McPherson et al. 2001), a 10 years collaborative effort, provided the reference human genome sequencing, using a combination of capillary sequencing and DNA microarrays. In the years following this first draft of the human genome, sequencing techniques have further improved, lowering costs and increasing both throughput and sensitivity. These techniques,

commonly referred as "next-generation sequencing", or NGS, are massively parallel sequencing (MPS) techniques. They allowed the sequencing of billions of reads per run and made it possible to answer many biological questions. This technology has evolved throughout the years, with various platforms offering different reads length, template generation, sequencing, and secondary measurements approaches (Table 1.1).

Table 1.1 List of MPS technologies available

Technology	Year	Reads length	Template generation	Sequencing	Measure	Reference
454 (Roche)*	2005	Short reads	Emulsion PCR	Sequencing by synthesis: SNA (pyrosequencing)	bioluminescence	(Margulies et al. 2005)
Helicos*	2008	Short reads	Amplification-free solid-phase	Single molecule sequencing by synthesis: SNA	fluorescence	(Bowers et al. 2009; Harris et al. 2008)
SOLiD (Thermo Fisher)	2008	Short reads	Emulsion PCR	Sequencing by ligation	fluorescence	(Valouev et al. 2008)
Illumina	2008	Short reads	Solid-phase bridge amplification	Sequencing by synthesis: CRT	fluorescence	(Bentley et al. 2008)
Pacific Biosciences	2009	Real-time long-read sequencing	Hairpin library structure	SMRT sequencing	fluorescence	(Eid et al. 2009)
Oxford Nanopore Technologies	2009	Real-time long-read sequencing	Hairpin library structure	SMRT sequencing	ionic current (nanopore)	(Clarke et al. 2009)
Complete Genomics (BGI)	2010	Short reads	In-solution DNA nanoball generation	Sequencing by ligation	fluorescence	(Drmanac et al. 2010)
Ion Torrent (Thermo Fisher)	2011	Short reads	Emulsion PCR	Sequencing by synthesis: SNA	pH (Ion semiconductor)	(Rothberg et al. 2011)
GenoCare (Direct Genomics)	2016	Short reads	Amplification-free solid-phase	Single molecule Sequencing by synthesis	fluorescence	(Gao et al. 2016; Zhao et al. 2017)
GeneReader (Qiagen)	2017	Short reads	Emulsion PCR	Sequencing by synthesis: CRT	fluorescence	(Darwanto et al. 2017; Koitzsch et al. 2017)

* discontinued; CRT: cyclic reversible termination; SNA: single-nucleotide addition; SMRT: single-molecule real time, PCR: polymerase chain reaction, BGI: Beijing Genomics Institute

All methods have a common a two-step process: a molecular biology (wet lab) step followed by a computational analysis (dry lab) step (Figure 1.3).

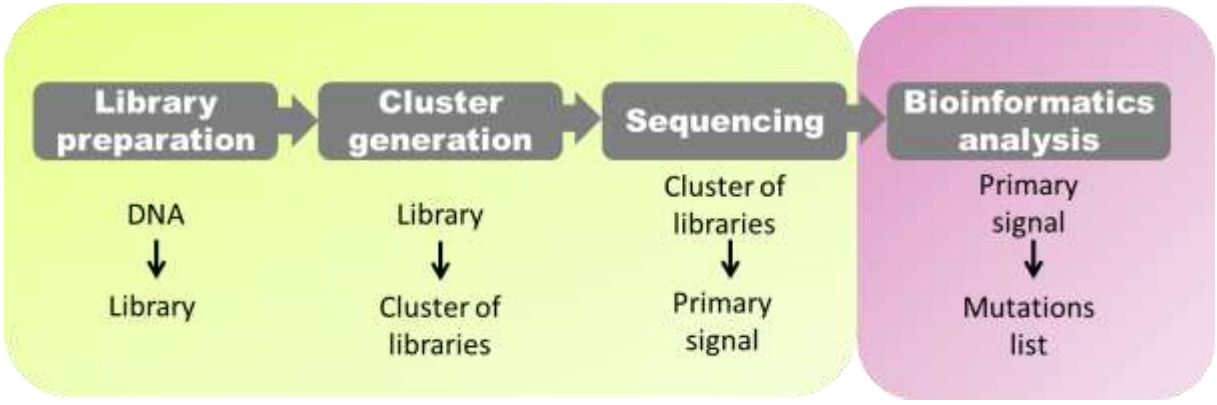


Figure 1.3 Overview of the general steps of molecular biology and computational analysis required for MPS The green section represents the wet lab part and the purple section represents the dry lab step.

The wet lab part generally consists first in a library preparation step that turns genomic DNA into fragments of DNA constructs. This process involves adding adapters and sequencing primers (SPs) to short fragments (for instance Illumina or Ion Torrent) or long fragments (i.e. Oxford Nanopore, Pacific Biosciences) of DNA that are required for sequencing. Then, during cluster generation, the constructs are hybridised onto a medium (such as a slide or beads) and might be amplified. This process can also be amplification free as firstly designed by Helicos. These amplification free methods are named in single molecule sequencing. Finally, the last stage of the wet lab part is the MPS where the libraries are read in the sequencer. Most sequencing approaches, including the Illumina methodology measure fluorescence signals, but exceptions exist such as Oxford Nanopore measuring ionic current and Ion Torrent (Thermo Fisher) quantifying pH changes.

In early years, the Illumina technology has increased the MPS capacity by developing reversible terminator chemistry (Bentley et al. 2008). This technique consists in a solid phase amplification followed by sequencing by synthesis (SBS) of short reads and enabled the first cancer genome, from an acute myeloid leukaemia patient to be sequenced (Ley et al. 2008). Illumina sequencing is now the MPS technique most widely employed in cancer genomics studies (Timmerman 2015). The reduction of cost and processing times has enabled the Illumina technology to characterise a large number of cancer genomes.

Several MPS methodologies exist and are commonly employed: whole genome sequencing (WGS), whole exome sequencing (WES) and targeted sequencing approaches. WGS refers to sequencing both coding and non-coding regions of the genome whereas WES denotes the technique elucidating the sequence of protein-coding regions only. Finally, targeted MPS allow the investigation of specific targets of interest independently of their genomic location.

1.2.1 Library preparation

The first step in the workflow consists in preparing library constructs for sequencing. Generally, this step includes DNA fragmentation followed by ligation of oligonucleotide sequences called adaptors, necessary for subsequent amplifications and sequencing. These adaptors contain oligonucleotides for cluster generation and amplification (P5 and P7) and SP. In the case of DNA libraries from several individuals being pooled and sequenced together, one or two indexes (unique 6 to 8 base pairs (bp) oligonueotides) are incorporated into the DNA product of each individual's DNA.

Several commercially available approaches can be used to generate the libraries depending on the application. For short read sequencing, depending on the sequencing technology, the DNA is fragmented to a size between 30 to 300 bp approximately. This size varies with the genomic regions targeted (WGS, WES, or targeted sequencing, reviewed in Chapter 1.2.5)

The most commonly used WGS library preparation is the Illumina TruSeq workflow (Figure 1.4). Random fragmentation of genomic DNA is followed by size selection of DNA fragments using beads. Finally, the adaptors are ligated in 5' and 3' of the DNA template.

Several exome library preparation methods are available to perform WES and the most common ones are in-solution capture methods. Most of these methods starts with libraries prepared using a commercially available kit such as one of the Illumina TruSeq workflow. Following the process described for WGS libraries (Figure 1.4) DNA templates containing exons are specifically selected using a pool of oligonucleotides bound to magnetic beads. After a pull-down step, the DNA template containing the regions of interest are amplified, producing exonic sequence enriched targets. However, other in-solution capture approaches start from genomic DNA (gDNA) as opposed to libraries. It is the case of the Agilent SureSelect method, offering an end-to-end library preparation solution, and the Illumina Nextera Rapid capture exomes. In this method, fragmented gDNA is directly pulled-down. Overall the three main in-solution capture methods are commercialised by Illumina (42 Mb coding + 20 Mb UTR), Agilent (50 Mb coding + 25 Mb UTR) and NimbleGen (64 Mb coding + 32 Mb UTR) (Chiara & Pavesi 2017). In addition to varying in size, they also vary in performances as described in Chapter 1.2.5.

Finally, smaller targeted methods are also popular approaches. These targeted sequencing panels are designed to specifically select regions of interest. Two main types of targeted sequencing are commercially available: capture hybridization-based and amplicon-based. Capture hybridization are those previously described for WES. Amplicon-based approaches are widely used as they require only a limited amount of gDNA input (Samorodnitsky et al. 2015). Several methods are available. One option is circular polymerase chain reaction (PCR) amplification such as HaloPlex (Agilent technologies 2015), where gDNA fragments digested by restriction enzymes are ligated in a circular form with adaptors necessary for amplification and sequencing. Another option is linear PCR amplification such as Ampliseq (Ion torrent) (Thermo Fisher Scientific 2015) and Truseq amplicon (Illumina) (Illumina Inc. 2015) (Figure 1.4). This methodology allows targets to be selected and amplified from gDNA by multiplex-PCR. Pairs of oligonucleotides complementary of the targets' flanking regions are used to amplify all the regions of interest in one reaction. A second amplification step is performed to add sequencing adaptors and indexes necessary for MPS.

1.2.2 Template generation approaches

Two template generations for short reads are widely used: the emulsion PCR (Ion Torrent, Thermo Fisher) and solid-phase bridge amplification (Illumina) (Figure 1.5).

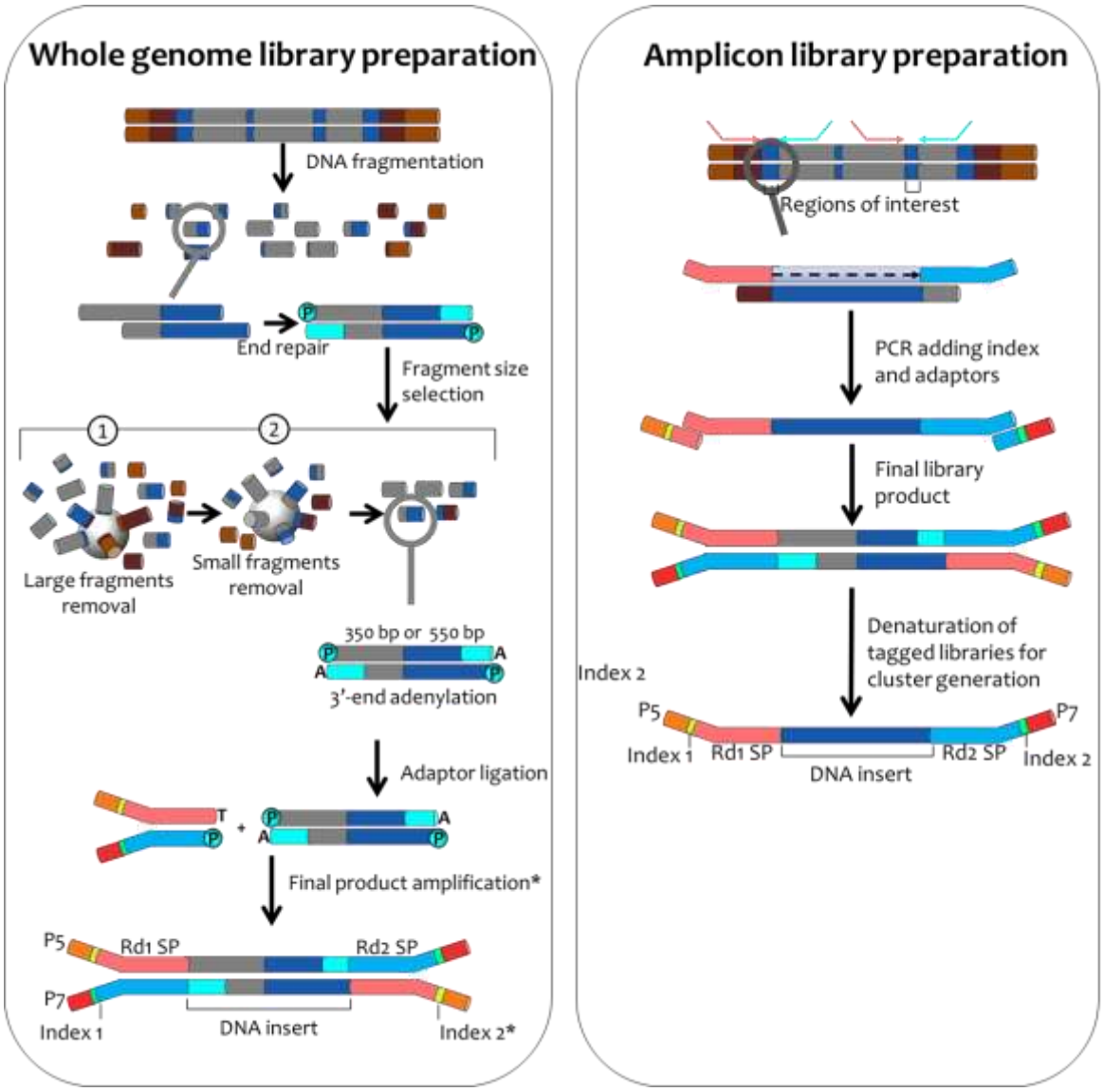


Figure 1.4 Main library preparation approaches used in cancer genomics projects, adapted from (Illumina Inc. 2015; Illumina Inc. 2017b)

For the whole genome library preparation, first the DNA is fragmented and the ends are repaired. Size selection of fragments is performed using magnetic beads. The 3' ends of the fragments are adenylated followed by adaptor ligations (in order to add sequencing primers, index and flow cell adaptors). The final product can be amplified if required.

For amplicon library preparation, regions of interest are amplified in a multiplex reaction; adaptors are then added to the DNA fragments by PCR. Final fragments are denatured before cluster generation.

P: 5'phosphate; Rd1/2 SP: read 1/2 sequencing primer. *optional

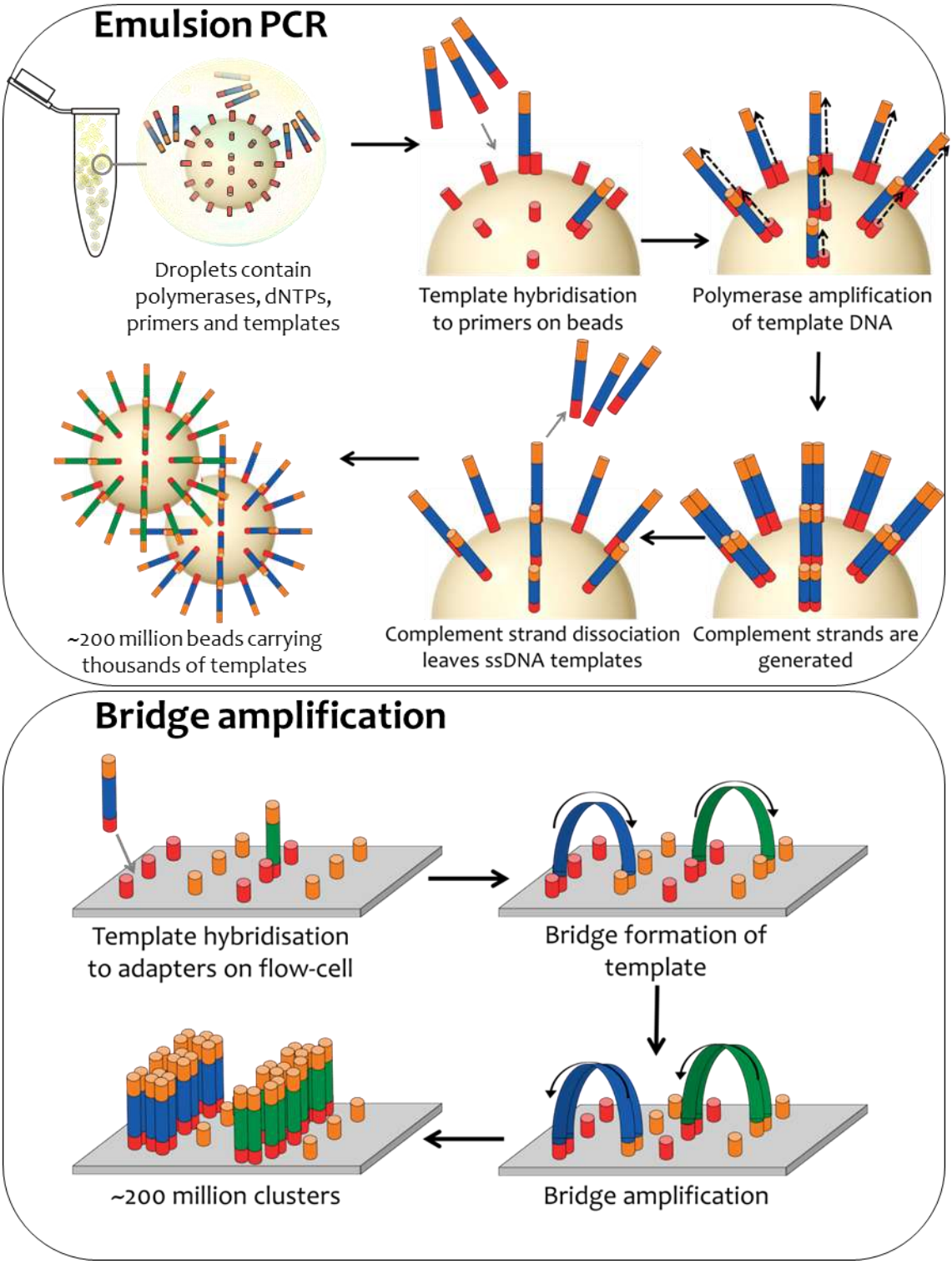


Figure 1.5 Most common methods of template generation approaches with amplification step (adapted from (Goodwin et al. 2016))

In the emulsion PCR, the reaction takes place in microdroplets containing polymerase, dNTPs, primers and templates. DNA library fragments are hybridised on the beads and amplified to generate the complement strand. The initial strand is dissociated to leave ssDNA templates on the beads.

In the bridge amplification, DNA library fragments are hybridised onto a flow cell. The library fragments are hybridised from the other end to form a bridge required for the amplifications where each DNA library will form a cluster of libraries.

Ion torrent sequencing consists of template generation by emulsion PCR. The DNA amplification takes place on beads in micro-droplets made of water and oil containing all PCR reagents required. Micro-droplets contain a maximum of one bead. The DNA fragments end with an adapter sequence complementary to oligonucleotide fragments immobilised on the bead surface. Clonal amplification takes place leading to the generation of 100-200 million beads carrying thousands of DNA templates (Shendure 2005).

Solid-phase bridge amplification is another template generation approach commonly used with the Illumina technology. Single-stranded DNA (ssDNA) constructs are bound to oligonucleotide fragments immobilised on a slide surface (flow cell). Optimal template concentration allows each DNA construct to be hybridised in non-overlapping clonal clusters. The free end of the DNA template is then bound to the flow cell by bridge formation. The second strand of DNA is created which in turn will be used as template for another step of bridge amplification. After several rounds of amplification, each initial molecule forms a cluster on the flow cell and a total of 100-200 million clusters are created on a flow cell.

Another commonly used template generation, amplification free, was developed by Helicos and still employed by GenoCare (Direct Genomics). Single-stranded DNA

templates contain a poly-A tail on its 3' end that will bind to the flow-cell which is covalently bound with poly-T oligonucleotides. Using this method, no amplification takes place and fragments are sequenced without the formation of clusters. Finally, the fourth main approach consists of ligating hairpin structures in the DNA template. This method is used by real-time long-read sequencing techniques such as Oxford Nanopore, operating with a single leader-hairpin structure at the 3' end of the template, and Pacific Biosciences, using a hairpin (SMRTbell) at each end of the DNA template (Biosciences 2014).

1.2.3 Sequencing approaches

Several sequencing methods have been developed and can be classified as sequencing by ligation (such as SOLiD platform and Complete Genomics) (Valouev et al. 2008), single molecule real-time sequencing (SMRT), and SBS (Metzker 2010).

Sequencing by ligation relies on the hybridisation and ligation on the library of labelled molecules made of an anchor complementary to the SP sequencing, one or two known nucleotides, followed by several universal nucleotides binding to any base that is linked to a fluorophore. The construct containing the one or two nucleotides complementary to the DNA library is ligated and the emission light of the fluorophore is imaged. A new cycle start after the fluorophore is cleaved. To sequence all bases of the DNA fragment single-nucleotide offsets are incorporated to cover the next set of nucleotides.

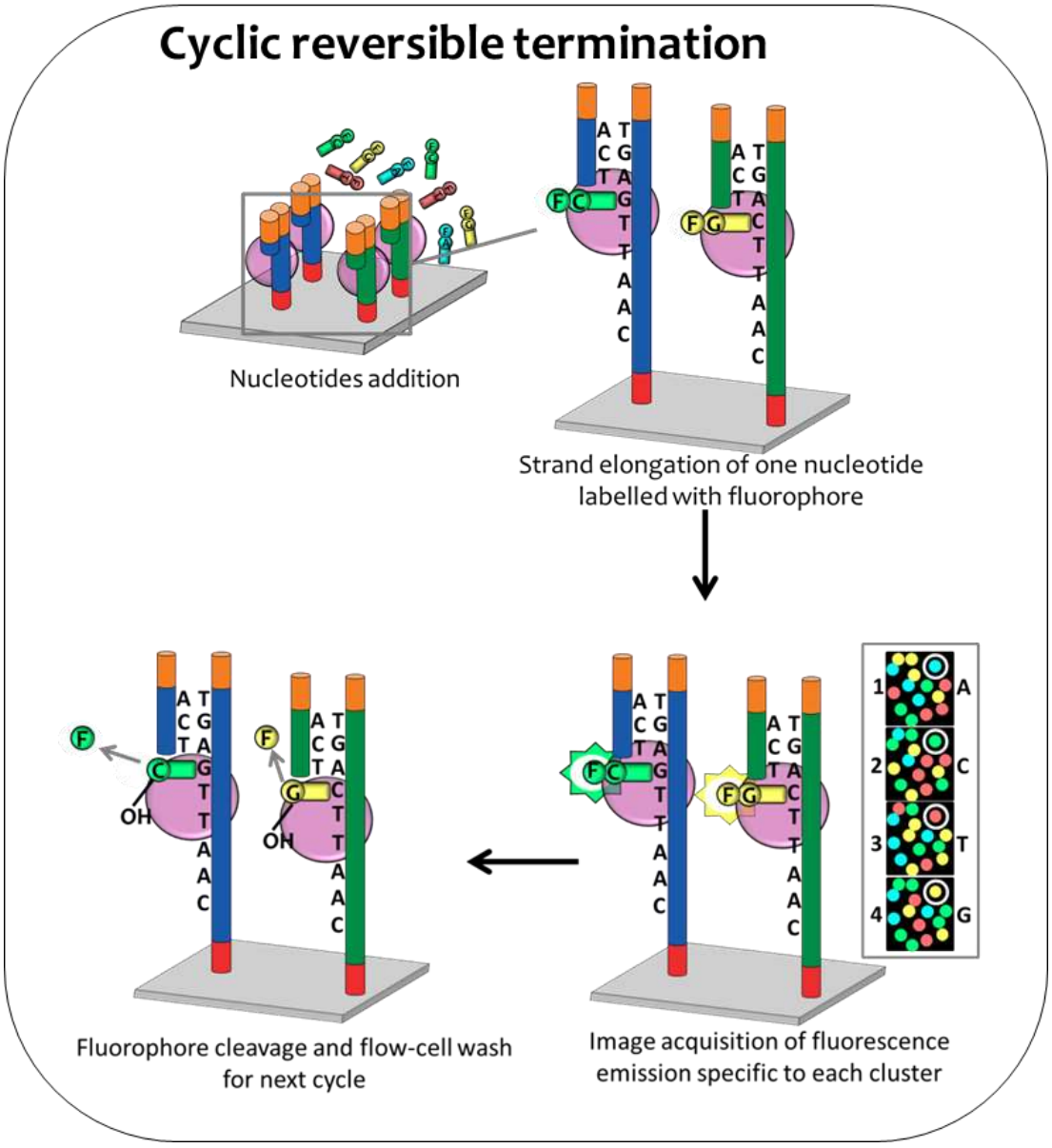
SBS are defined as polymerase-dependent sequencing methods and can be classified as two methodologies. Cyclic reversible termination (CRT) was developed by the Illumina technology and also used by the Qiagen GeneReader (Bentley et al. 2008; Darwanto et al.

2017; Koitzsch et al. 2017) (Figure 1.6). It is based on the principle that the 3'-OH group of the nucleotides incorporate during sequencing is blocked and then unblocked. In the first step, the flow-cell contains SP sequences hybridised to the ssDNA fragments arranged in clusters. All four nucleotides, labelled by a fluorophore and with 3'-OH blocked are added and the polymerase incorporates one nucleotide on each DNA fragment. After washing out unbound nucleotide, in the second step the instrument images the fluorophore emission light for each cluster. Finally, the fluorophore is washed out, and the nucleotide is unblocked revealing the 3'-OH end in order to start another cycle of nucleotide incorporation-imaging-wash.

The second approach is single-nucleotide addition (SNA), established by the Roche 454 and now broadly used by Ion torrent sequencing (Figure 1.6) (Margulies et al. 2005; Rothberg et al. 2011). In this case, unlike the CRT method, only one nucleotide is incorporated in the reaction and therefore four reactions with each individual nucleotide is necessary to complete one cycle. However, since the nucleotides are incorporated one by one, it is not necessary for them to be blocked. After each nucleotide addition, the signal is measured to detect or not the presence of an incorporated base. One effect of this type of sequencing is that homopolymers (sequences of the same repeated nucleotide) can be incorporated in one cycle. The detection method can detect different signal intensity revealing the number of bases incorporated at once. The ion torrent technology is the first method that does not rely on fluorescence detection, but actually measures a change pH as H^+ ions concentration changes. This change is in theory proportional to the number of nucleotides integrated.

Finally SMRT sequencing is based on long-read sequencing, and is now mainly used by Oxford nanopore and Pacific Biosciences (Eid et al. 2009). For these techniques, the enzyme or polymerase is fixed and the DNA fragment progress through it. In both cases, long fragment of several kilobases can be sequenced. The Pacific Biosciences approach measures, in real-time, the colour and duration of the fluorophore's emission light, and can sequence the same fragment several times as it is circular, generating a consensus sequence and reducing noise in the signal. The Oxford nanopore does not measure a secondary signal and directly detects the DNA composition of the fragment. The sequencing does not happen base by base like the Illumina technology, but as a fragment of several nucleotides, the change in current, characteristic of the sequence, is measured. Several thousands of possible signals can be measured depending on the composition and length of the fragment.

In most cases, paired-end (PE) sequencing is employed as opposed to single read DNA sequencing (Fullwood et al. 2009). This consists of sequencing one DNA template in both directions and can be particularly useful to obtain longer reads which is necessary for the resolution of repetitive regions of the genome or regions containing DNA variants. PE sequencing has been extensively used since it was developed in 1981 (Hong 1981), for early initiatives (Venter et al. 1998; Lander et al. 2001) and more recent work (Bovee et al. 2008).


Cyclic reversible termination
Nucleotides addition
Strand elongation of one nucleotide labelled with fluorophore
Image acquisition of fluorescence emission specific to each cluster
Fluorophore cleavage and flow-cell wash for next cycle
1 A
2 C
3 T
4 G
OH

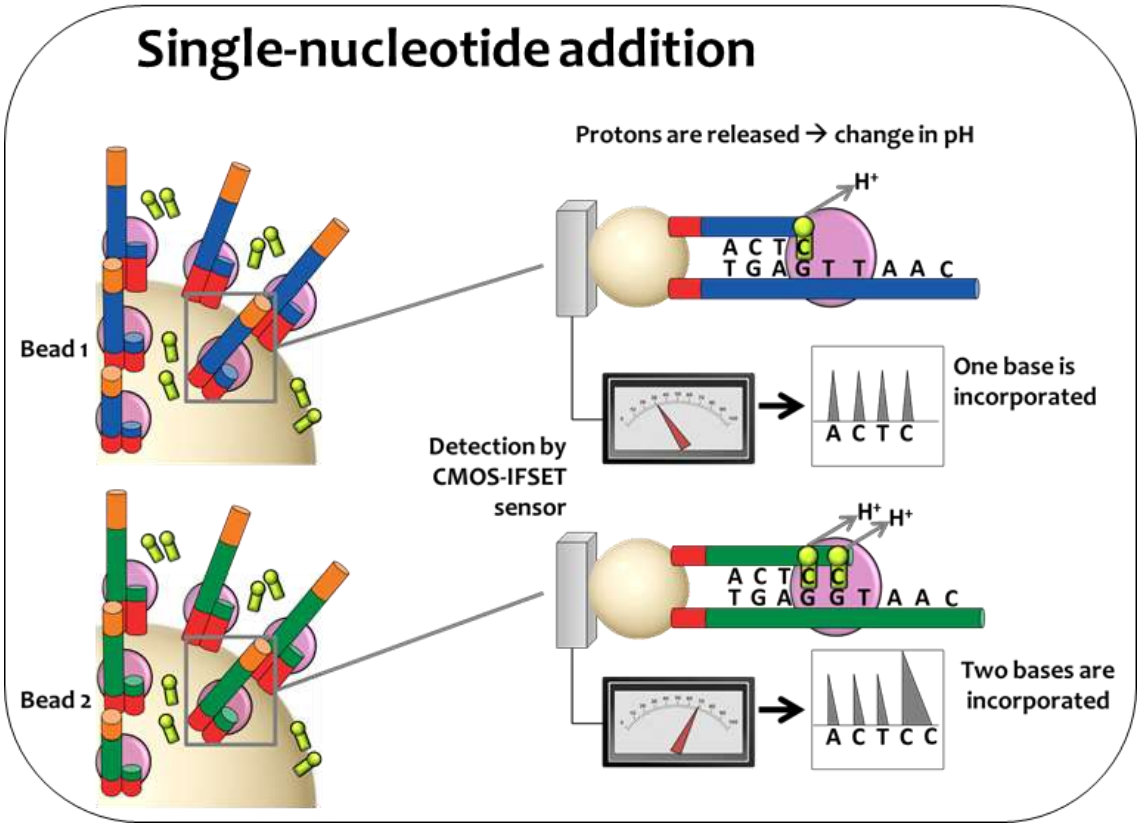


Figure 1.6 Sequencing by synthesis approaches: cyclic reversible termination (CRT) (top) and single-nucleotide addtion (SNA) (bottom) adapted from (Goodwin et al. 2016)

The cyclic reversible termination consists in the addition of the four labelled nucleotides where only the complementary one will bind the sequence. Then imaging of the one nucleotide incorporated takes place on all the different tiles of the flow cell where each dot represent one cluster. Finally the fluorophore is cleaved before a new cycle starts.

The single-nucleotide addition sequencing approach consists in adding one type of nucleotide at a time (four cycles, each with one type of nucleotide, are required to complete the sequencing of one base). When one unlabelled unprotected nucleotide is incorporated, a proton is released leading to a change in pH which is measured by the detector. If the sequence is composed to several identical nucleotides in a row, several bases are incorporated and the pH signal measured is proportional to the number of bases.

F: cleavable fluorophore; CMOS-IFSET: complementary metal-oxide semiconductor - ion-sensitive field effect transistor

1.2.4 Bioinformatics analysis for resequencing strategies and somatic variant detection

Bioinformatics analysis is essential to interpret MPS data. A multitude of tools have been developed to turn raw output data from the sequencer to sequences of DNA and to identify variants inserted in the DNA sample (Figure 1.7).

Each MPS instrument measures a signal that can be fluorescence, change of pH or other which is stored as metrics such as images. The first step of the analysis is base calling where the signal is decoded to assign which nucleotide was incorporated and it is specific to each platform. A quality score, similar to the Phred score for Sanger sequencing, is associated with each call determining the probability of error (meaning the confidence in the call) (Ewing & Green 1998). Typically, the quality score reached with MPS data is 30 (Q30), meaning a probability of 1 in 1000 of incorrect base call (Illumina Inc. 2011). Most modern MPS technology allow real-time base calling while sequencing is taking place. Therefore, additional quality controls (QCs) are applied such as the Illumina quality filtering of cluster, indicating "pass filter" reads (PF). It represents the ratio of the highest of the four (base type) intensities to the sum of highest two. It allows to filter out entire reads where at least two bases out of the 25 first base called are failing this quality score usually when two clusters are too close (Illumina Inc. 2010). Read sequences are generated in the FASTQ format, containing the actual sequence of each read and the quality score associated to each base call. For PE sequencing, one FASTQ file is generated for each sequencing direction. When several samples are pooled together and

sequenced using different indexes, a demultiplex step is necessary to divide reads belonging to different samples.

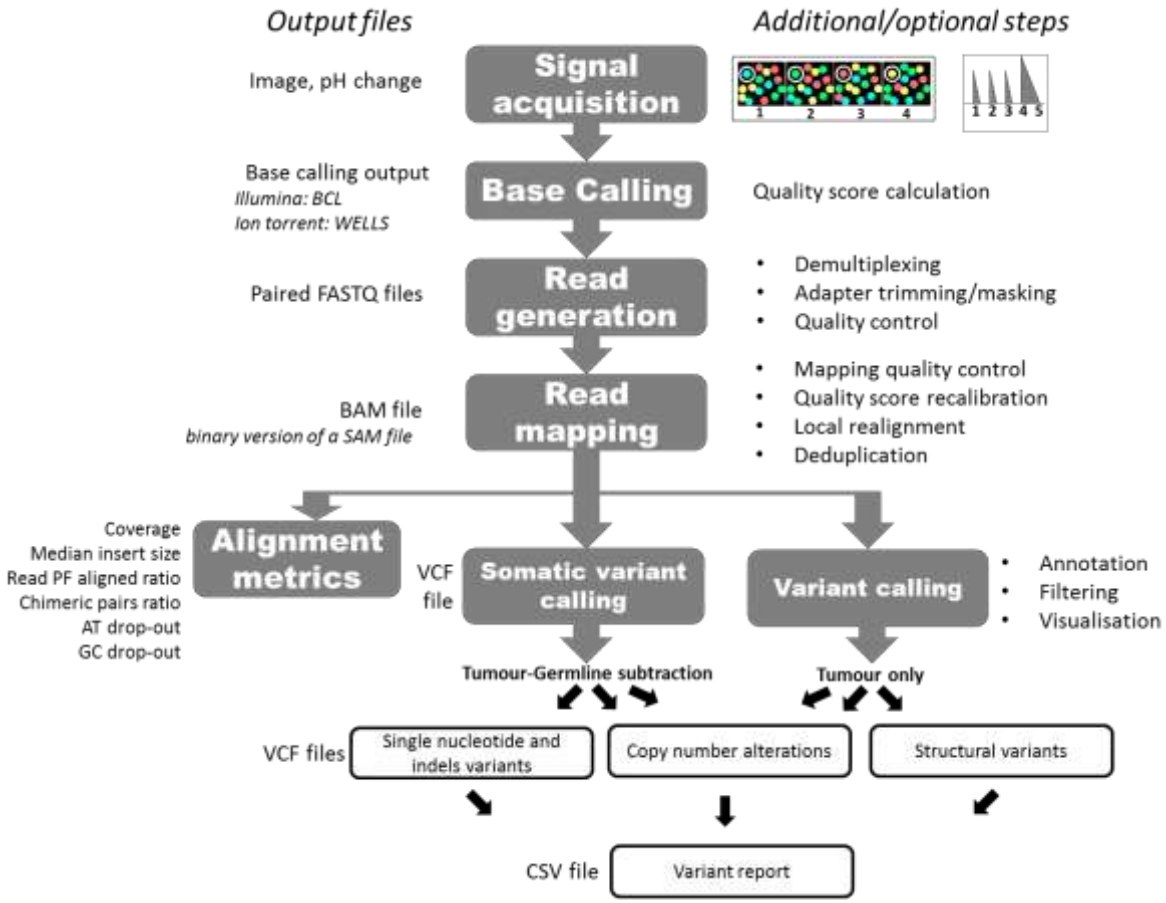


Figure 1.7 Flow diagram illustrating the MPS bioinformatics pipeline

BCL: binary basecall; SAM: Sequence Alignment Map; PF: pass filter; VCF: Variant call format, CSV: comma-separated values

The reads are then mapped to a reference sequence, or reference genome to reconstruct the original sequence of the sample. This data is stored in a binary alignment map (BAM) file. Multiple intermediate steps can be applied in order to remove PCR duplicates (deduplication), improve alignment in regions containing insertions and deletions (local realignment), correction of base quality score systematic errors (quality score

recalibration) and mapping QC. Several mapping QC can be calculated to make sure the sequencing data is reliable as read mapping can affect variant detection (Broad institute 2017). The median insert size is the median length between extremities of paired reads. The ratio of reads aligned passing the filter (Read PF aligned) represents the fractions of sequencing reads that have a position on the reference genome. The chimeric pair ratio reflects paired reads which align to different regions or chromosomes. Measures of AT dropout show how less covered regions with GC content < 50% are compared to the mean and GC dropout shows how less covered regions with GC content > 50% are compared to the mean. Percent aligned to the human genome reference and High AT/GC normalised coverage indicate coverage uniformity. Insert sizes and chimeric reads values correlate with the number of potential somatic false positive variants. In addition, sequencing depth of coverage is crucial to avoid the presence of gaps in the sequencing (under covered regions) which could lead to sub-optimal variant detection performances (Sims et al. 2014). From short read sequencing, some regions are not well aligned, especially repetitive regions and therefore it is crucial to measure the sequencing coverage uniformity across the whole genome.

The mapped reads are analysed to uncover the difference between the sample and the reference (GL variants) or the sample and normal control (somatic variants). Multiple somatic variant callers have been developed to detect SNVs and indels, some based on a Bayesian approach (Mutect (Cibulskis et al. 2013), Strelka (Saunders et al. 2012)), others based on Fischer's exact statistics (Shimmer (Hansen et al. 2013), VarScan (Koboldt et al. 2012)). A recent study, comparing nine somatic variant callers highlighted important differences in the variants detected between the different tools tested (Krøigård et al.

2016). Methods called intersect and combine approaches have been developed to overcome these differences. They use several somatic variant callers to generate a consensus list (intersection) or a combined list (union) of variants from sequencing data (Rashid et al. 2013; Kim et al. 2014; Callari et al. 2017).

Despite the numerous tools available, some variants remain challenging to detect. Some variants are only carried in a limited number of cells (subclonal variants), and therefore when bulk samples are sequenced only a small fraction of the reads will carry that variants. The allelic fraction (AF) represents the fraction of reads presenting with a variant and the cancer cell fraction (CCF) represents the proportion of cell carrying that variant. A heterozygous clonal variant will present an AF of 0.5 and a CCF of 1. In order to detect variants with low AF a sufficient depth of coverage (number of reads covering a given nucleotide) is required (Shin et al. 2017; Lee et al. 2017). Other variants are challenging to uncover as they are in repetitive regions of the genome generally filtered out of the analysis or in homopolymers. Several soft filters are used to annotate the variants in order to easily filter out these variants if necessary without increasing false negatives (Illumina Inc. 2016).

Although more challenging, larger variants (>50bp), called SVs (balanced and unbalanced) can also be detected from MPS data (Tattini et al. 2015). There are different technique to detect SV as detailed by Alkan and colleagues: read pair (where the expected distance between pairs and observed distance is different), read depth (where a region shows exceptional lower or higher depth than adjacent regions), split read (where a read overlapping with a breakpoint is split), assembly (where the contigs inform on the presence of SV) (Alkan et al. 2011)

In order to investigate the significance of each variant and generate a final report, variants require annotations. Several annotations tools are available for SNVs and indels annotations, such as ANNOVAR (Wang et al. 2010), or Variant effect predictor (VEP) (McLaren et al. 2010). Annotation tools generally provide information on the variant location and consequence, predictions on consequence of the variant (deleterious or benign using tools such as SIFT (sorting intolerant from tolerant) (Kumar et al. 2009)) and cross-references with databases such as the catalogue of somatic mutations in cancer (COSMIC) (Forbes et al. 2011), 1000 genomes project (1KGP) (1000 Genomes Project Consortium et al. 2012), and ClinVar (Landrum et al. 2016).

1.2.5 Whole genome vs. targeted/exome approaches

Multiple MPS methodologies are available to investigate genomic mutations in cancer genomes (Table 1.2).

WGS approaches allow the detection all types of variants, including SNVs, small insertions and deletions (indels), CNAs and other SVs (Pang et al. 2014) in both coding and non-coding regions of the genome (such as splice sites, promoter regions and most microRNAs). In addition, it is possible to use a PCR-free approach which improves coverage uniformity by avoiding GC-bias.

Table 1.2 Comparison of characteristics of WGS, WES and targeted sequencing for somatic variant investigations

		WGS (Illumina)	WES	Targeted sequencing
PCR amplification†		optional	yes	yes
Length/fraction of the genome covered†		In theory 3,234Mb / 100%	37-64 Mb of coding sequence + 0-32 Mb of UTR / 1-3%	1 kb-5Mb / <0.15%
Regions covered	**Coding**	yes	yes	limited
	Regulatory	yes	no	limited
	Other non-coding	yes	no	limited
DNA input†		1 ug (100 ng with PCR amplification step)	100-250 ng	5-250 ng
Recommended depth of coverage*		15-60X	100-120X	1,000-10,000X
Price for 100X sequencing*		~$1,100	~$550	
Amount of data for a 100X†		300 Gb	4-6 Gb	< 0.5 Gb
Process time from DNA to VCF file†		6 days (for 2 samples)	6 days (for 12 to 96 samples)	3 days (12 to 96 samples
Variant detection	**SNVs-indels***	yes	yes	yes
	SVs*	yes	limited	limited
	Subclonal variants (low allelic fraction)	limited	yes	yes
	Novel coding variants	yes	yes	no
	Novel non-coding variants	yes	no	no

† Manufacturer's data. WGS: whole genome sequencing, based on Illumina technology, WES: Whole exome sequencing, based on Illumina, Agilent and Nimblegen technology. Targeted sequencing based on Ion torrent, Illumina, Haloplex

* Recommended depth of coverage (Bentley et al. 2008; Fang et al. 2014; Clark et al. 2011; Meynert et al. 2013; Kim et al. 2015)
* Price for 100X sequencing (Genohub Inc. 2017)
* SNVs-indels (Fang et al. 2014)
* SVs(Tattini et al. 2015; Magi et al. 2013; Bansal et al. 2014; Boeva et al. 2014)

However other MPS methodologies are available as well. Because of the current cost of WGS technology and challenges brought by data storage and analysis, other approaches might be preferred depending on the final application, especially in large scale studies where many samples must be sequenced, or in cancer genome studies where higher sequencing depth are required to detect subclonal variants (Clark et al. 2011). To date, WES is the most popular technique in cancer genomics studies as it allows the detection of most well described and well annotated somatic variants at a reasonable cost (Gargis et al. 2012). A recent study highlighted the differences between three in-solution exome capture methods: Agilent SureSelectQXT, Illumina Nextera Rapid capture exome and NimbleGen SeqCap EZ Choice (used as gold standard). Although all methods had suitable detection of DNA variants of interest, the Agilent method showed an overall quality close to the gold standard, NimbleGen showed highest coverage and evenness and Illumina Nextera had the highest on-target levels but also had high duplicate rates (García-García et al. 2016).

It is worth noting, however, that exome sequencing also has some weaknesses. Several reports have shown the potential deleteriousness of variants present in non-coding regions (especially in regulatory regions (Puente et al. 2015; Killela et al. 2013)) and current WES capture methods do not cover these regions. In addition, it has been estimated that up to 20% of the exome is not captured by standard WES (Bamshad et al. 2011). When compared to WGS: it provides a less even coverage, strand bias, and a lower proportion of transcripts that are completely covered (Lelieveld et al. 2015). This lower performance of WES can impact on variant detection, especially copy-number variants (Medvedev et al.

2009; Szatkiewicz et al. 2013) whilst allelic bias can overestimate or underestimate the AF of a variant.

Because of the reported poor performances of WES compared to its costs, many laboratories developed targeted sequencing panel to specifically screen a limited number of genomic regions and increase coverage in these regions. This technique is cost effective as many samples can be sequenced in one run and it requires less starting material which is helpful for some applications where DNA availability is low. The library preparation and sequencing is quicker than larger scale approaches. Regions of interest are generally sequenced much deeper than WES, allowing the detection of subclonal variants in cancer samples with higher sensitivities despite allelic bias due to PCR amplification. Although several algorithms exist to detect CNAs from targeted sequencing data (Talevich et al. 2016; Demidov et al. 2016), this detection remains dependent on the spread of targets across the whole genome. Several targeted sequencing panels are commercially available, such as Ion AmpliSeq Cancer Hotspot Panel v2 (Thermo Fisher Scientific, Waltham, MA, USA) targeting hotspot regions in 50 oncogenes and tumour suppressor genes. In addition, custom amplicon panels can be designed to fit any applications using web tools such as Design Studio (Illumina Inc. 2017a), SureDesign (Agilent 2017) or Ion AmpliSeq Designer (Thermo Fisher Scientific 2017).

1.2.6 Other techniques for genomes analysis

The epigenome is another layer of information that is required to fully understand the implication of the genome in the cancer process. This includes key elements of DNA transcription such as DNA methylation, post-transcriptional histone modifications and

transcription factor occupancy, as well as chromatin accessibility, nucleosome position and three-dimensional chromosomal conformation. In recent years, following progress on DNA sequencing using MPS, great developments have been made in the technology to study such epigenetic changes.

Whole-genome bisulfite sequencing (WGBS) is considered the gold-standard to investigate the DNA methylome. High methylation levels generally indicate repressed promoters or gene expression (Gal-Yam et al. 2008; Lister et al. 2009; Ohm et al. 2007; Schlesinger et al. 2007; Wolf et al. 1984) and low levels are linked to active enhancers (Stadler et al. 2011). As the majority of methylation takes place on cytosine bases, sodium bisulfite is used to convert all non-methylated cytosine bases to uracil and leave methylated bases unaffected (Lister et al. 2009).

Post-transcriptional histone modifications and transcription factor binding sites are investigated by chromatin immunoprecipitation followed by sequencing (ChIP-Seq) (Barski et al. 2007). The chromatin and transcription factors can be fixed before being digested and purified using specific antibody coupled with magnetic beads. Double stranded DNA fragments are then isolated from the complex, and sequenced.

Open chromatin sites and transcription factor footprints can be investigated using the assay for transposase-accessible chromatin using sequencing (ATAC-seq) where a Tn5 transposase fragments accessible chromatin and adds sequencing tags to the fragments. These DNA fragments are then sequenced (Buenrostro et al. 2013).

Nucleosome positions can be determined by MNAse-Seq profiling experiments (Cui & Zhao 2012). MNAse digest all DNA except fragments of approximately 147 bp protected

by the nucleosome structure, revealing the footprint on DNA where the nucleosomes were located.

Three-dimensional chromosomal conformation, or chromatin loops, can be studied to highlight the interactions between different DNA fragments that are distant on the linear sequence, but interact with each other in the 3D DNA conformation. Circular chromosome confirmation capture (4C-Seq) is a very specific technique characterising the interactions of a specific bait (Splinter et al. 2011). HiC-seq, on the other hand provides low resolution of the 3D interactions at the whole genome scale (Dixon et al. 2012).

1.3 Genomic technologies revolutionised the cancer genetics field

1.3.1 The genomic landscape of cancers demonstrates inter-tumour heterogeneity

Large scale research-led MPS studies have dramatically changed our understanding of the cancer genome (Tomczak et al. 2015). Such initiatives allowed the analysis of numerous samples and were not confined to only a few genomic targets. Indeed, current collaborative efforts are using WES or WGS to characterise thousands of cancer genomes. The Cancer Genome Atlas (TCGA) initiative aims to draw up a comprehensive catalog of genomic abnormalities, by initially characterising the genomes of glioblastomas and subsequently collected data on 33 different tumour types in 11,000 patients (McLendon et al. 2008). The International Cancer Genome Consortium (ICGC) has a similar goal, in

order to elucidate comprehensively the genomic changes in 50 different cancer types and more than 25,000 cancer genomes (Hudson (Chairperson) et al. 2010). The final objective of both collaborative projects is to translate these findings into clinical practice.

Since their launch in 2005 and 2008, respectively, they have defined the large extent of tumour heterogeneity. Inter-tumour heterogeneity was characterised by the multitude of genomic landscape publications (Creighton et al. 2013; Getz et al. 2013; Hammerman et al. 2012; Muzny et al. 2012; Puente et al. 2011). The cancer genes census was greatly extended since the first version and now lists more than 600 genes referenced in the COSMIC database (Futreal et al. 2004; Forbes et al. 2011; Wellcome trust sanger institute 2017).

1.3.2 Intra-tumour heterogeneity expanded the cancer evolutionary concept

One of the characteristics of the cancer genome is its lack of stability leading to the accumulation of somatic mutations over time. Most cancers have an elevated mutation rate (Loeb et al. 2003), for instance triple-negative breast cancers (13.3X higher than normal cells (Y. Wang, Jill Waters, Leung, Unruh, et al. 2014)), which fuels intra-tumour heterogeneity.

The theory firstly described by Nowell that intra tumour heterogeneity drives cancer evolution has been since extensively studied using MPS by describing intra-tumour heterogeneity and the presence of subclonal mutations.

Most studies supported this concept of monoclonal origin of tumours and stepwise mutations acquisition follow by strong selection (Anderson et al. 2011; Schuh et al. 2012; Diaz Jr et al. 2012; Gerlinger et al. 2012; Gerlinger et al. 2014; Landau et al. 2015; Gerstung et al. 2017). Different complexity of this model was described as some tumours follow linear evolution and others showed branched evolution (Davis et al. 2017; Sidow & Spies 2015). Linear evolution is the simplest model where mutations are acquired one after the other where the newest driver mutations provide a selective advantage compared to the previous ones. In the branched evolution model, different subclones are derived from a common ancestor and they all co-exist in the tumour. Other studies, for example in CLL described cases of convergent evolution where subclones in different phylogenetic branches carry the same mutations (Wang et al. 2017; Ojha et al. 2015). Other concepts of cancer evolution have been postulated since and contradict the Darwinian models.

One of them is the polyclonal tumour origin where two or more clones interact with each other and initiate cancer development (Parsons 2008), another one is the punctuated equilibrium model which is challenging the gradual acquisition of mutations (Cross et al. 2016) and the last one is the neutral process without selection arguing against the strong clonal selection (Schwartz & Schäffer 2017). Sottoriva and colleagues took the previously described neutral evolution concept (Yachida et al. 2010; Jones et al. 2008) and postulated a “Big Bang” model where the totalities of clonal and subclonal mutations occur early during tumour evolution in a single clonal expansion (Sottoriva et al. 2015). Neutral process without selection was described to be of importance especially in solid tumours (Ling et al. 2015; Williams et al. 2016; Sottoriva et al. 2015), but remains controversial (Tarabichi et al. 2017).

Although all these models are different, they describe the presence of intra-tumour heterogeneity and subclonal mutations.

1.3.3 Characterisation of mutational mechanisms

The mechanism of acquisition of somatic mutation has been described as random (Welch et al. 2012; Nik-Zainal, Alexandrov, et al. 2012). A theory by Wang and colleagues highlighted that with a mutation rate of 0.9 per cell division, a tumour (commonly assumed to be composed of at least $3x10^9$ cells) could theoretically have developed a point mutation on every nucleotide of the genome (Y. Wang, Jill Waters, Leung, Unruh, et al. 2014). Even bearing this in mind, the theory of clonal evolution using a stepwise acquisition of random mutations is difficult to support in light of the presence of hotspot mutations.

However, this model can partially be explained by the mechanisms of hypermutability leading to chromosome instability, extensively detailed in a recent review (Schwartz & Schäffer 2017). Mutation development can take place within a specific pattern of nucleotides, called mutational signatures, as described by Alexandrov and colleagues (Alexandrov et al. 2013). This study highlighted the importance of passenger somatic mutations and the mutational nucleotide bias which is sometimes specific to a tissue type. To date, 30 nucleotide signatures have been described. Some of these signatures are triggered by the environment (Alexandrov & Stratton 2014), others have been associated with hypermutability such as deregulation of the apolipoprotein B mRNA editing enzyme catalytic polypeptide (APOBEC) family (signature 2 and 13), the activation-induced cytidine deaminase (AID) protein (signature 9), mismatch repair (MMR) deficiency

(signature 6, 20 and 26) or BRCA1/BRCA2 deficiency (signature 3). The origin of other signatures is unknown. Most of the mutations linked to these signatures are distant of each other, however regions where mutations are clustered, or regions of kataegis have been described in cancer (Nik-Zainal, Alexandrov, et al. 2012; Lawrence et al. 2013). They are defined as regions carrying six mutations with a distance between mutations at least two standard deviations (SDs) below the median in a genome. This mechanism has been linked to different enzymes activities depending on the tissue type: double strand breaks and APOBEC in breast cancer (Alexandrov et al. 2013), and AID-mediated deamination in CLL (Rebhandl et al. 2014). Chromosome instability and hypermutability have been also reported to be linked to chemotherapy causing double strand breaks (Johnson et al. 2014), or altering clonal evolution of cancer (Johnson et al. 2014).

The genome can also be altered at a chromosomal level with massive genomic rearrangements called chromothripsis, where chromosomes are shattered and reassembled in an apparent random fashion (Stephens et al. 2011). In addition, more recently other types of signatures, defined by SVs, have been described in breast cancer and ovarian cancers, and further refined the underlying molecular mechanisms of cancer (Nik-Zainal et al. 2016; Macintyre et al. 2017).

Following DNA damage, the DNA damage response (DDR) is activated in order to stop the cell cycle, repair defects such as base single or double strand damage to prevent cancer development. *ATM* is a key regulator of the DDR which acts by recruiting many other proteins including *TP53* as well as *MSH2* and *MLH*, which are involved in the DNA MMR (Takagi 2017). *ATM* and other DDR proteins as well as MMR effectors have been reported to be inactivated by both GL and somatic mutations. *ATM* alterations lead to

ataxia–telangiectasia syndrome and are also recurrently reported in several haematological malignancies such as CLL (Stankovic & Skowronska 2014). *TP53*, the gene with the most somatic mutations in cancer, is also found mutated in the GL of patients with Li-Fraumeni syndrome which causes a predisposition to cancer (Nichols et al. 2001). Bi-allelic alteration of the MMR genes causes a constitutional repair deficiency, and has also been associated with cancer development (Takagi 2017).

The presence of multiple mutational mechanisms as well as inter and intra-tumour heterogeneities have been presented challenges to translate genomics findings into clinical practice. Indeed, the resistance mechanisms can be different in patients with the same diagnosis by involving different mutated genes, or affected pathways which complicate potential strategy of effective therapies. However, with the development of targeted therapies to specific genes and pathways, patient stratification and precision medicine is now a reality.

1.4 Genomics era paving the way for precision medicine

In recent years, advances in genomic technologies have led to incredible progress in the discovery of molecular biomarkers, which has enable research findings to be translated into clinical diagnostics. DNA alterations, by their presence or their absence in the cancer genome, can help in defining a diagnosis, predicting prognosis or guiding treatment choice (Quintáns et al. 2014). This has been the origin of precision oncology, or precision medicine, aiming to treat a patient with the most suitable therapy available, in light of the alterations detected in the genome of the tumour cells (Collins & Varmus 2015). Several

studies have highlighted the benefit of patient management guided by genomics results (Biankin et al. 2015). In addition, the identification of new biomarkers has been able to guide the development of new targeted therapies for patients who are not responding to existing ones.

The identification of key mutations in certain cancers has already benefitted from the use of precision medicine. The most recognised example is chronic myeloid leukaemia where all of the patients carry a *BCR-ABL* translocation. Following the development of the *BCR-ABL* inhibitor, Imatinib, the patient's survival at 5 years increased to 90% (Saussele et al. 2015). Other targeted therapies added to chemotherapy regimens have been shown to be effective against human epidermal growth factor receptor 2 (HER2)-positive breast cancer (Slamon et al. 2001; Geyer et al. 2006; Verma et al. 2012; Swain et al. 2015).

Some genes are recurrently mutated specifically in one type of cancer, other are mutated across several types, such as *APC* virtually confined to colorectal cancer, or *TP53* and *PIK3CA* distributed in several cancer types (Kandoth et al. 2013). This observation has lead to pan-cancer genomics analysis and basket clinical trials (West 2017). These innovative clinical trial designs include patients with different histology classification but the same genetic mutation (André et al. 2014; Lawler et al. 2015; Meric-Bernstam et al. 2015; Le Tourneau et al. 2015; Stockley et al. 2016), and have recently demonstrated their feasibility and efficacy (Hyman et al. 2015; Horak et al. 2016).

1.5 Chronic Lymphocytic Leukaemia as a model for genomics studies

In most of the cancer types, the mutational landscape is heterogeneous and most patients do not present with the exact same mutational hotspot. This is the case with CLL which is a genetically heterogeneous cancer and to date, it remains challenging to provide a molecular diagnosis to all patients. In addition, the progression of the disease is slow from a pre-malignant monoclonal B-cell lymphocytosis (MBL) to a malignant CLL (in approximately 1% of the cases (Rawstron et al. 2008)) that can be stable for many years or refractory to the treatment, and to potentially further progression to Richter syndrome, an aggressive lymphoma (in 5% of patients) (Figure 1.8) (Parikh et al. 2014).

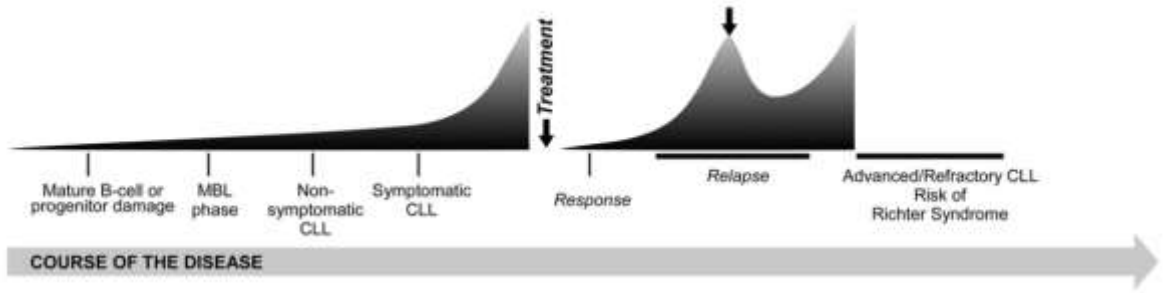


Figure 1.8 Slow progression of Chronic Lymphocytic leukaemia, from (Guièze & Wu 2015).

MBL: Monoclonal B-cell lymphocytosis, CLL: Chronic Lymphocytic leukaemia

CLL is therefore a good model to study tumour heterogeneity and evolution to uncover new biomarkers in the context of precision medicine using genomics technology.

1.5.1 Existing genomic biomarkers

Several genomic prognostic biomarkers have been discovered in the last few decades targeting specific pathways. CLL biomarkers are derived from four major pathways (Landau et al. 2015). The DDR and cell cycle control pathways are affected in CLL mainly through *TP53* and *ATM*, two key tumour suppressor genes. *TP53* mutations are generally associated with copy number loss over its locus on chr17p (reported in 80% TP53 abnormalities) (Thorsten Zenz, Eichhorst, et al. 2010; Rossi et al. 2009; Gonzalez et al. 2011). Alterations in this gene are found in 4-10% of patients before treatment and increases up to 30% at relapse (Stilgenbauer et al. 2009; Hallek et al. 2010; T Zenz et al. 2010; Rossi et al. 2009) . These alterations are associated with shorter survival and refractoriness to standard first-line chemo-immunotherapy (CIT) Fludarabine, Cyclophosphamide, Rituximab (FCR) (Stilgenbauer et al. 2009; Hallek et al. 2010), as its mechanism of action (MOA) is *TP53*-driven apoptosis (Pettitt 2003). Therefore, in the context of CIT, it is recommended to test for the presence of alterations in *TP53* prior to the commencement of treatment (Pospisilova et al. 2012; Hallek et al. 2010). However, patients with TP53 alterations can respond to targeted therapy such as Ibrutinib, a Bruton's tyrosine kinase inhibitor which targets the B-cell receptor pathway (Burger et al. 2014; Byrd et al. 2013; Farooqui et al. 2015). *ATM* alterations are highly recurrent in CLL patients with 15-20% of patients having a copy number loss over the chr11q locus and 30% of patients carrying mutations (Neilson et al. 1997; Döhner et al. 1997). Bi-allelic alterations are frequently reported, and have been associated with poor prognosis, equivalent to the one observed with TP53 alterations (Skowronska et al. 2012; Rose-

Zerilli et al. 2014). Deletions also have been historically reported as being poor prognosis (Byrd et al. 2004; Edelmann et al. 2012; Hernández et al. 2015). However, the addition of Rituximab in the first-line treatment FCR has significantly improved the prognosis of patients with del11q (Hallek et al. 2010; Tsimberidou & Keating 2009).

The RNA and ribosomal processing pathways are also altered in CLL. The most frequently mutated gene in this pathway is the splicing factor 3B subunit 1 (*SF3B1*) which was found mutated in 15% of patients, mainly in the C-terminal HEAT-repeat domain and associated with poor prognosis (Quesada et al. 2011; Wang et al. 2011; Rossi et al. 2011; Quesada et al. 2012). More recently, mutations in the ribosomal protein S15 (*RPS15*) have been reported and were associated with early relapse after CIT FCR treatment, classifying this gene as a potential new predictive biomarker in CLL (Ljungstrom et al. 2016).

The NF-kappa B signalling pathway, including inflammatory pathways is activated in CLL and several key biomarkers have been described. Baculoviral inhibition of apoptosis repeat containing 3 (*BIRC3*) is rare in early patient but significantly increased in fludarabine-refractory cases (from 3 to 24%) (Rossi, Rasi, et al. 2013) and as it is located on chr11q22 (co-localised with *ATM*). It is also recurrently affected by bi-allelic alterations (Malek 2013). Myeloid differentiation primary response 88 (*MYD88*) is rarely mutated in CLL (4%). It is concentrated on a hotspot L265P missense mutation and has been associated with good risk patients (Puente et al. 2011).

Finally, the Notch signalling pathway is also recurrently dysregulated in CLL. It is particularly affected by a 2-bp frameshift deletion (c.7544_7545delCT) the PEST domain of *NOTCH1* and has been associated with a poorer outcome (Fabbri et al. 2011; Puente et al. 2011).

In total, recent WES and WGS studies have reported SNVs and indels in more than 2000 genes in CLL, demonstrating the high genetic heterogeneity of the disease (Wang et al. 2011; Quesada et al. 2011; Fabbri et al. 2011; Landau et al. 2013; Landau et al. 2015; Puente et al. 2015; Burns et al. 2018).

Additionally, other SVs have been proven to be significant biomarkers in CLL. This is the case with chr13q deletions which are associated with good prognosis when in isolation (Van Dyke et al. 2016; Edelmann et al. 2012), as well as trisomy 12 which has been associated with different prognoses: intermediate in isolation (Döhner et al. 2000; Van Dyke et al. 2016), but associated with Richter's transformation (Chigrinova et al. 2013; Rossi, Spina, et al. 2013). The four main structural variations 17p13, 11q23.3 and 13q14 as well as trisomy 12 were linked with clinical outcome in 2000 by Döhner and colleagues and remain the most widely used prognostic model in CLL (Döhner et al. 2000).

Furthermore, a very well characterised biomarker in CLL is the immunoglobulin heavy chain variable region (*IgHV*) mutational status. *IgHV* status allows dividing the patients into two prognostic groups: patients in the good prognostic group carry a mutated *IgHV*, whereas the poor prognostic group presented with >98% homology to the GL sequence and are defined as unmutated (Damle et al. 1999; Hamblin et al. 1999).

Despite the characterisation of multiple prognosis genomic biomarkers, predictive ones are still lacking and therefore additional studies are required (Sargent et al. 2005): at present, only TP53 defects (deletions and mutations) as well as *IgHV* mutational status are recommended to be screened in routine before deciding to give CIT to a patient. However, the lack of predictors could be explained as treatments available evolve and new biomarkers are required (Hallek et al. 2010).

Many other DNA alterations have been recently detected and could potentially become biomarkers if investigated in clinical trial settings, such as subclonal variants, mutational signatures and the presence of a complex karyotype.

1.5.2 Discovery of novel biomarkers using MPS technology

With the advances in MPS, more biomarkers could be detected in CLL by exploring different aspect of the genome. Several recent studies have highlighted the importance in investigating non-coding regions of the genome, as well as global measurement of the genomic instability, namely the presence of genomic complexity (GC) or complex karyotypes and mutational signatures.

WGS offers the advantage of being able to explore the non-coding regions of the genome including regulatory regions such as promoter, enhancer and repressors. Several studies have identified recurrent non-coding variants in the promoter of the telomerase reverse transcriptase (*TERT*) gene firstly in melanoma, and subsequently in brain, endocrine, gynaecological, urological, digestive system, head and neck, soft tissue and pleural, and other tumours, as reported in more than 100 publications since 2013 (Killela et al. 2013). The *TERT* promoter mutation has been proposed to be both a prognostic and predictive biomarker. It was reported as diagnostic biomarker in benign vs. malignant thyroid tumours (Liu & Xing 2014) and as a monitoring biomarker in urine DNA in bladder and renal cancers (K. Wang, Liu, Ge, et al. 2014; K. Wang, Liu, Liu, et al. 2014; Wang et al. 2015). These mutations have been found to be independent prognostic markers associated multiple times with aggressive, high grade, progressing tumours, poor overall survival

(OS), or with metastases in several types of tumours, as well as lower response rate to some therapies (as reviewed by Liu and colleagues) (Liu et al. 2016).

In CLL, other non-coding variants have been identified as recurrently mutated. Mutations in the paired box 5 (*PAX5)* enhancer region have been identified in CLL patients and associated with a decrease in protein expression (Puente et al. 2015; Burns et al. 2018). In addition, mutations in the 3'UTR of *NOTCH1* have also been reported in CLL (Puente et al. 2015; Burns et al. 2018), and that the presence of this mutation gave the same poor outcome as the coding 2bp deletion (Larrayoz et al. 2017) .

Other types of genomic markers are global measurements of the whole genome, such as mutational signatures defined by Alexandrov and colleagues (Alexandrov et al. 2013). In CLL, the AID signature has been found enriched in all patients, but particularly associated with *IgHV* mutated patients (Burns et al. 2018; Puente et al. 2015; Kasar et al. 2015).

Another global measure on the genome is called GC and has been extensively studied in CLL during the era of single nucleotide polymorphism array (SNP array). Using this technique, GC has been defined as 3 or more CNAs and has been associated with poorer outcome (Kujawski et al. 2015; Ouillette, Collins, Shakhan, Li, Li, et al. 2011; Brown et al. 2012; Knight et al. 2012). As expected, GC is associated with defects in DDR genes *ATM* and *TP53* (Gunnarsson et al. 2010). However, in two third of patients with GC, these two genes were unaffected, suggesting a different molecular mechanism, still unidentified, leading to these chromosomal events (Malek 2013). Chromothripsis has also been described in CLL and such events were associated with poorer progression free survival in the context of first-line FCR therapy (Edelmann et al. 2012).

Finally, the presence of subclonal variants in genes well defined as prognostic markers in CLL have been investigated in recent years. Similarly to clonal variants, small subclonal mutations in *TP53* have been associated with poor survival. The clinical impact of subclonal variants in *SF3B1*, *NOTCH1*, *ATM* and *BIRC3* remains controversial (Davide Rossi et al. 2014; Nadeu et al. 2016; Rasi et al. 2016). Recent genomic studies following the tumour progression (CLL) have confirmed that the clonal composition of CLL tumour evolves over time and it has been demonstrated that molecular relapse can be detected in sequencing data before clinical relapse (Knight et al. 2012; Schuh et al. 2012; Landau et al. 2013). It has also been shown that more aggressive subclones are selected by treatment (Landau et al. 2015). One study found evidence of the presence of Ibrutinib resistance mutations in very low subclones before treatment (Burger et al. 2016).

1.6 MPS in the clinic

1.6.1 Current and future practices

Although Sanger sequencing and cytogenetics are still largely used in diagnostics laboratories, massive parallel sequencing (MPS) and in particular targeted sequencing is now becoming the gold standard method to investigate genomic mutations for cancer patients in the clinic (Koboldt et al. 2013). The laboratory can explore the presence of mutations in disease specific genes by designing custom targeted sequencing panels (Hamblin et al. 2017; Roller et al. 2013).

However, with the reduction of both cost and bioinformatics challenges, it is widely accepted that WGS may present advantages over targeted approaches (Muir et al. 2016;

Taylor et al. 2015). It can provide a comprehensive evaluation of the mutational landscape across the entire genome, in both coding and non-coding regions.

It is a single test which allows the detection of the full range of genetic alterations used to guide clinicians, including SNV, indels, and SVs such as CNAs from a single aliquot of DNA (Pang et al. 2014). This is particularly valuable as the conventional multi-modality testing currently employed in routine diagnostics can rapidly exhaust the low amount of material available from tumour specimens (Travis et al. 2011). Although several targeted sequencing methods also allow the detection of all classes of mutations, WGS presents an additional advantage of unbiased sequencing (Cottrell et al. 2014; Pritchard et al. 2014; Lionel et al. 2017). In addition, it enables mutation detection from any DNA sequence that could potentially be proven of clinical significance without the necessity of redesign and validation of future tests to identify new "actionable" variants or clinically relevant genes or non-coding regions. Moreover, WGS provides homogeneity of the data for future data mining.

1.6.2 Challenges to implement WGS in the clinic for cancer patients

WGS efforts in cancer have mainly been undertaken on high quality nucleic acids extracted from fresh tissue (FF) specimens collected through a research infrastructure (Shabihkhani et al. 2014). One study has demonstrated the feasibility of WGS for cancer patients in the clinic using fresh tumours and matching blood samples (Laskin et al. 2015). However, in routine cancer care, these specimens are generally more challenging to collect, and as a consequence this is not typically carried out.

On the other hand, formalin-fixed, paraffin-embedded (FFPE) material is used to establish a histopathological diagnosis (Ludyga et al. 2012). However, such treatment impacts on DNA quality by generating different type of alterations. Such degradation was classified into four types (Figure 1.9) (Do & Dobrovic 2014).

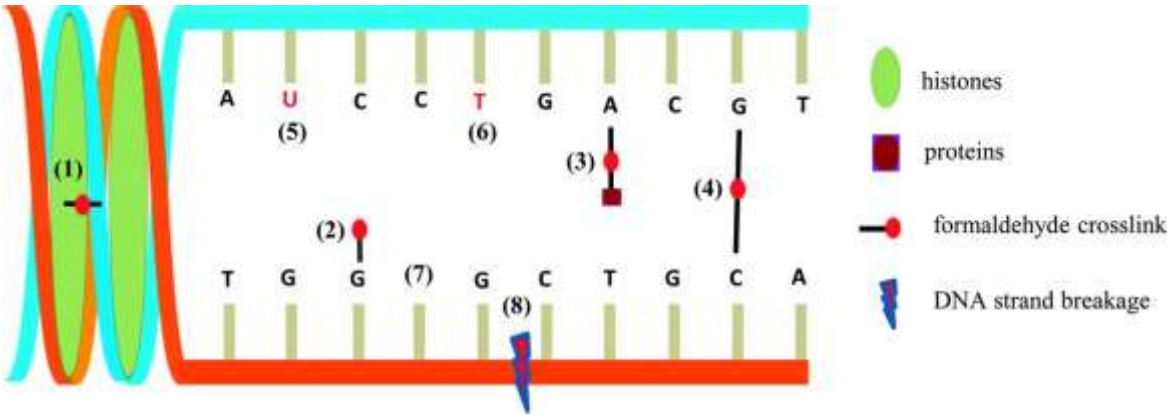


Figure 1.9 Different types of DNA damages cause by FFPE tissue preparation (from (Do & Dobrovic 2014)).

There are four types of crosslinks: (1) Histone–DNA crosslinks, (2) formaldehyde–DNA adducts, (3) DNA–protein crosslinks, (4) DNA–DNA crosslinks. Results of deamination with base changes rom cytosine to (5) uracil and (6) thymine. Loss of DNA resulting in (7) abasic sites and (8) strand breaks.

The first type of degradation is nucleic acid fragmentation caused by formalin fixation. This is adversely influenced by long formalin fixation time or use of unbuffered formalin and exacerbated by long term storage in paraffin. The second type results from when crosslinks induced by formaldehyde occur between DNA and other components (protein, other DNA strands and formaldehyde itself) and reduce the stability of the nucleic acids' structure. At low pH in formalin, the third type of alteration can occur: purine bases can be lost by hydrolysis resulting in abasic sites which in turn can lead to localised DNA denaturation and strand breaks. The fourth type of DNA degradation is the conversion of

cytosine bases to uracil by deamination, which can lead to miscalling of C to T transition mutations.

Some of these alterations can be dealt with by following adapted protocols in the laboratory such as DNA repair steps (Do & Dobrovic 2012). However, as nucleic acids from FFPE cancer specimens remain of variable and sometimes poor quality, it has impeded the implementation of WGS in cancer diagnosis.

1.6.3 Uncover molecular biomarkers to provide molecular diagnosis to all patients

With the generation of large scale data, millions of somatic variants have been identified in cancer genomes. One of the major challenges has been to interpret the clinical significance of the variants observed in a given patient (Quintáns et al. 2014). Many recurrently mutated genes and variants have been identified in different cancer types. However, as many research initiatives used targeted sequencing panels and exome sequencing, the focus has been on coding regions excluding non-coding elements of the genome (Tomczak et al. 2015). In addition, many clinical studies have presented findings focusing on single region/genes as opposed to looking at the genome in a global measurement (such as mutational signatures and genomic complexity).

In CLL, a limited number of altered genes and chromosomal regions are prognostic/predictive biomarkers. In the era of CIT as well as targeted therapies such as Ibrutinib and Venetoclax, gene specific biomarkers are indispensable. However, so far only pathogenic SNVs or deletions of *TP53* are currently used to direct an untreated patient to receive targeted therapies such as Ibrutinib and yet many patients who do not

respond to standard first-line treatment do not carry these alterations. In a recent review from Rossi and colleague proposed *IgHV* mutational status as a second predictive marker in addition to *TP53* (Figure 1.10A) but this remains controversial.

In addition, the focus of the large scale CLL studies has been on large cohorts of early stage CLL patients (Puente et al. 2015; Landau et al. 2015), and therefore patients with more advanced disease, or relapse refractory (R/R) disease have been studied less to date (Schnaiter et al. 2013; Dreger et al. 2013). This results in a lack of molecular biomarkers for relapsed patients (Figure 1.10B).

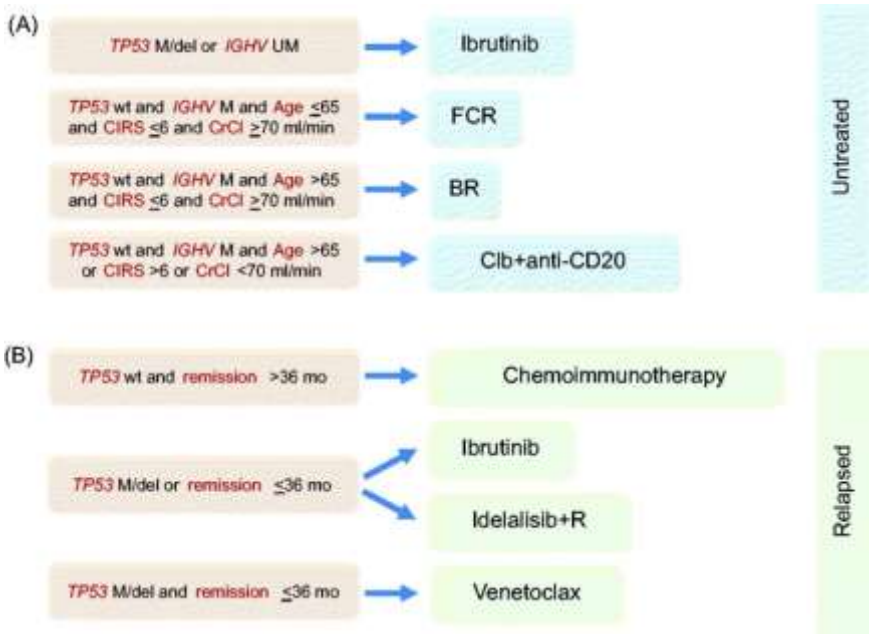


Figure 1.10 Predictive biomarkers in CLL from (Rossi et al. 2017)

(A) in the context of untreated patients, (B) in the context of relapsed patients

M: mutation; del: deletion; CIRS: cumulative illness rating scale; CrCl: creatinine clearance; *IgHV*: immunoglobulin heavy variable gene; mo: months; FCR: fludarabine, cyclophosphamide, rituximab; BR: bendamustine, rituximab; Clb: chlorambucil.

1.7 Aims of the thesis

The first aim of this thesis was to address the feasibility of implementing WGS into routine cancer diagnostics so that it may be possible to establish comprehensive genomics characterisation of cancer patients undergoing surgery with curative intent.

To do so, the first objective was to:

(i) assess the feasibility of using FFPE specimens collected via routine pathology pathways as a source of DNA for WGS

(ii) compare WGS data quality metrics between matched fresh frozen (FF) and FFPE specimens

(iii) evaluate variant detection in FFPE specimens: SNVs, indels and CNAs

The second objective was to optimise the sequencing quality of DNA from FFPE specimens by:

(i) predicting WGS quality of FFPE specimens using DNA quality metrics

(ii) improving FFPE DNA quality by optimizing the DNA extraction method

(iii) identifying key steps in the FFPE tissue preparation influencing WGS quality

(iv) generating clinical reports from WGS data using FFPE DNA samples

The second aim of the thesis was to identify genomic markers of aggressive R/R disease that could be used for risk and therapy stratification using CLL as a model.

The first objective was to identify molecular markers that would be enriched in cohorts of R/R patients by:

(i) investigating in R/R patients the incidence of mutations and their co-occurrence and comparing this to the mutation pattern of treatment-naïve/unselected patients

(ii) identifying predictive markers to treatment

(iii) studying patterns of clonal evolution by tracking driver mutations using technologies currently available in the clinic

Finally, the second objective was to investigate the significance of predictive molecular marker of relapse after first-line CIT:

(i) using the technologies currently available in the clinic

(ii) validating and using WGS as a diagnostic tool in the clinic for CLL patients

Chapter 2 Materials and methods

2.1 Thawing frozen peripheral blood mononuclear cells

Frozen cells need to be thawed with caution to avoid cell death.

Vials containing the frozen cells were incubated at 37 degrees in a water bath until thawed. Cells are transferred drop by drop in 10 ml of Dulbecco's Modified Eagle Medium (DMEM) (11995-065, Thermo Fisher Scientific, Waltham, MA, USA), centrifuged at 180g for 5 minutes. The supernatant was removed and 200 μl of phosphate buffered saline (PBS) (10010031, Thermo Fisher Scientific, Waltham, MA, USA) was added.

2.2 Histopathology preparation from resected tissues

Tumour samples were collected from resection of cancer patients in theatre. The maximum cold ischaemia time was 2 hours. The tissue samples were prepared as both FF and FFPE samples as per the usual protocol in National Health Service diagnostic laboratories depending on the tissue type. For FF specimen preparation at least one punch of 5 mm was taken and was flash frozen in liquid nitrogen. The frozen punch(es) were weighed and embedded in OCT (Embedding medium for optimal cutting temperature). H&E sections were taken and assessed by a pathologist for percentage of viable tumour

cells (out of total number of cells present) and percentage necrosis. Samples with >40% viable invasive tumour cells were suitable for the study. One 5 mm core punch was collected from the tissue to proceed to DNA extraction. The remainder of the specimen was prepared as FFPE. All samples underwent fixation with formal saline for routine sectioning: this was usually carried out the following day, but the optimal duration of fixation depends on specimen type and/or time of sample receipt. In most cases, the FFPE block was taken in the area surrounding the FF punch. Pathologists assessed the FFPE slides corresponding to the block and chose a high cellularity area which is well preserved and avoided areas of necrosis if possible. Two 1 mm core punched were collected from the specimen to proceed to DNA extraction. All FFPE samples were stored at room temperature.

2.3 Extraction of genomic DNA using spin columns

Cells are disrupted to release their content in solution and treated with enzymes. In presence of high concentration of salt, DNA binds to silica-gel membranes contained in spin columns. After being washed, elution from the membrane of pure nucleic acids is performed at a low concentration of salt. Commercially available kits were used to perform DNA extractions from different sources. Unless stated otherwise, all reagents used were provided in the kit.

2.3.1 From fresh frozen tissues

The AllPrep DNA/RNA Mini kit (80204, Qiagen, Hilden, Germany) was used to extract DNA from approximately 30 mg of fresh frozen tissue. The frozen tissue was transferred into a 15 ml tube containing 600 µl Buffer RLT Plus, containing 2-mercaptoethanol (63689-25MLF Signa-Aldrich, St Louis, MO, USA). Tissue sample was disrupted and homogenised in solution using a rotor–stator homogenizer for 30 seconds. Lysate was centrifuged 3 minutes at full speed and supernatant was transferred to an AllPrep DNA spin column placed in a 2 ml collection tube. The column was centrifuged at 8000g for 30 seconds. At this stage the flow through lysate containing RNA was stored at -80°C. The DNA spin column was placed in a new 2 ml collection tube, washed with 500 µl of buffer AW1 followed by a 15 seconds centrifugation at 8000g and washed with 500 µl of buffer AW2 followed by a 2 minutes centrifugation at full speed. The collection tube was carefully emptied after each wash. The DNA spin column was placed in a 1.5 ml microcentrifuge tube and incubated at room temperature, lid closed for 2 minutes with 50 µl of buffer EB (Q19086, Qiagen, Hilden, Germany). Finally, DNA was eluted by centrifuging the column at 8000g for 1 minute.

2.3.2 From formalin-fixed, paraffin-embedded tissues

This process requires dissociation of paraffin by treatment (mechanical or chemical), followed by a heat treatment to reverse crosslinks formed with DNA in presence of formalin and followed by DNA purification using spin columns. DNA was extracted from 2 punched cores of 1 mm diameter.

2.3.2.1 QIAamp DNA FFPE tissue kit

The QIAamp DNA FFPE tissue kit (56404, Qiagen, Hilden, Germany) was used to extract DNA from FFPE tissues. The tissue sample was placed in a 1.5 ml microcentrifuge tube. 320 µl of deparaffinization Solution (19093, Qiagen, Hilden, Germany) was added to the sample, vortexed for 10 seconds and pulse-centrifuged. The sample was incubated at 56°C for 3 minutes and incubated until reaching room temperature again. 180 µl of buffer ATL was added and the tube pulse-vortexed. The sample was centrifuged at 11000 x g for 1 minute. 20 µl of proteinase K was added into the lower, clear phase of the sample and gently mixed by pipetting. The sample was incubated overnight at 56°C to allow complete digestion followed by 1 hour at 90°C for crosslinks reversal. The sample was briefly centrifuged. The lower, clear phase was transferred into a new 2 ml microcentrifuge tube. 200 µl of buffer AL was added to the sample, mixed by vortexing and 200 µl of >96% ethanol was added and mixed again by vortexing the tube. The sample was briefly centrifuged and the entire lysate was transferred to a QIAamp MinElute column in a 2 ml collection tube. The column was centrifuged at 6000g for 1 minute. The spin column was placed in a new 2 ml collection tube, washed with 500 µl of buffer AW1 followed by a 1 minute centrifugation at 6,000g and washed with 500 µl of buffer AW2 followed by a 1 minute centrifugation at 6,000g. The collection tube was changed after each wash. The column was centrifuged at 20,000g for 3 minutes to dry the membrane completely. The column was placed to a 1.5 ml microcentrifuge tube and incubated at room temperature for 1 minute with 50 µl of buffer ATE. Finally the column was centrifuged at 20,000g for 1 minute.

2.3.2.2 truXTRAC FFPE DNA microTUBE kit

The truXTRAC FFPE DNA microTUBE kit (PN 520136, Covaris, Woburn, MA, USA) was used to extract FFPE DNA. This method is based on adaptive focused acoustics treatment to efficiently remove paraffin and rehydrate FFPE tissues at the same time using an ultrasonicator instrument (Covaris, Woburn, MA, USA)

The FFPE specimens were transferred in a microTUBE Screw-Cap containing 100 µl of Tissue Sodium dodecyl sulphate (SDS) buffer. The tube was loaded in the ultrasonicator instrument M220 and treated with 20% factor duty, 200 cycles per burst for 300 seconds at 20°C with a peak incident power of 75 Watts. 20 µl of proteinase K was added to the sample, mixed thoroughly and treated second time with the same parameters for 10 seconds. The sample was incubated overnight at 56°C to allow complete digestion followed by 1 hour at 80°C for crosslinks reversal. For DNA purification, the sample was transferred to a 1.5 ml microcentrifuge tube. 140 µl of buffer B1 was added to the tube followed by pulse-vortexing and 160 µl of >96% ethanol was added to the same tube and pulse-vortexed. The sample was centrifuged for 2 minutes at 10,000g. The liquid layer was transferred to a purification column in a 2 ml collection tube, making sure to avoid the paraffin layer visible on top of the liquid. The column was centrifuged for 1 minute at 11,000g and the flow-through discarded. 500 µl of buffer BW was added to the samples and the column centrifuged for 1 minute at 11,000g. After discarding the flow-through from the collection tube, 600 µl of buffer B5 was added and the column centrifuged for 1 minute at 11,000g. The column was placed in a new collection tube and dried by a 1 minute centrifugation at 11,000g. The column was transferred to a 1.5 ml microcentrifuge

tube. 50 µl of pre-warmed buffer BE (5mM Tris Hydrogen chloride (HCl) pH 8.5, 70°C) was added to the column for elution and incubated at room temperature for 3 minutes. The column was then centrifuged for 1 minute at 11,000g.

2.3.3 From thawed peripheral blood mononuclear cells and buffy coat

The QIAamp DNA mini kit (51306, Qiagen, Hilden, Germany) was used to proceed with extraction of DNA. 20 µl of Qiagen protease and 200 µl of buffer AL was added to a 1.5 ml microcentrifuge tube containing 200 µl of lysed sample. After pulse-vortexing, the tube was incubated at 56°C for 10 minutes. 200 µl of ethanol was added, the tube pulse-vortexed. The mix was transferred to a spin column and centrifuged at 6000g for 1 minute. After changing the 2 ml collection tube to a new one, 500 µl of buffer AW1 was added and the column was centrifuged again at 6000g for 1 minute. The column was transferred to a new 2 ml collection tube and 500 µl of buffer AW2 was added. The column was centrifuged at 20,000g for 3 minutes and transferred to a 1.5 ml microcentrifuge tube. Buffer AE was added; the column was incubated for 1 minute at room temperature and centrifuged at 6000g for 1 minute.

2.3.4 From saliva samples

Saliva samples were extracted using the Oragene DNA kit (DNA Genoted Inc, Ottawa, ON, Canada). Briefly, the saliva collected was mixed by inverting the sample 10 times and incubated at 50°C for 1 hour. The sample was then transferred to a 15 ml Falcon tube

containing a 1/25th volume of PT-L2P solution (DNA Genotek Inc, Ottawa, ON, Canada), vortexed for 10 seconds and incubated at 4°C for 10 minutes until impurities were precipitated and centrifuged at 3500xg for 10 minutes. The supernatant was transferred to a new tube and 1.2x volume of room temperature 100% ethanol was added to the sample to precipitate the DNA. The tube was mixed by 10 inversions and incubated at room temperature for 10 minutes before centrifugation at 3500xg for 10 minutes. The supernatant was removed and discarded and the pellet washed by adding 1 ml of 70% ethanol. The sample was incubated at room temperature for 1 minute. The supernatant was removed. The DNA (contained in the pellet) was rehydrated by adding 0.5 ml of TE buffer (10 mM Tris-HCl, 1 mM disodium Ethylene diamine tetra acetic acid (EDTA), pH 8.0) (Sigma-Aldrich, St Louis, MO, USA) and vortexed for 30 seconds. The sample was incubated at room temperature overnight, vortexed for 30 seconds and transferred to a 1.5 ml microcentrifuge tube.

2.4 DNA quality control

2.4.1 Nanodrop

Nucleic acids concentration and purity was assessed by spectrophotometry using a Nanodrop 2000c UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). All nucleic acids absorb UV light at a wavelength of 260 nanometres (nm). A photodetector measures the light going through the sample and gives an optical density result. The more light is absorbed, the higher the concentration in nucleic acids is. The optical density result is converted to a concentration in nucleic acids using the Beer-

Lambert law. Therefore, spectrophotometry quantifies all nucleic acids in the sample and not only DNA. Proteins on the other hand, absorb UV light at a wavelength of 280 nm. A ratio of absorbance at 260 and 280 (A260/A280) is used to determine samples contamination with proteins. An acceptable ratio is between 1.8 and 2.

A blank measurement with 1 µl elution buffer used at the DNA extraction step was done prior sample measurement. The instrument pedestal was cleaned with distilled water between measurements. 1.5 µl of sample was pipetted on the pedestal and absorbance measured. The software Nanodrop 2000 software package (v1.6.198, Thermo Fisher Scientific, Waltham, MA, USA) was used to calculate concentrations and A260/A280 ratios.

2.4.2 Qubit

Double-stranded DNA (dsDNA) was quantified by fluorimeter using a Qubit 2.0 Fluorimeter (Thermo Fisher Scientific, Waltham, MA, USA). Fluorimetric dye is added to the sample and will specifically bind dsDNA by a base intercalation process. The quantity of dye bound is directly proportional to fluorescence intensity and to dsDNA quantity in the sample. However, this measurement does not provide information on sample purity.

Concentration of undiluted DNA samples was assessed with the dsDNA broad range assay kit (Q32850, Thermo Fisher Scientific, Waltham, MA, USA) and diluted samples with dsDNA high sensitivity assay kit (Q32854, Qiagen, Hilden, Germany). A working solution was prepared by diluting Qubit reagent in Qubit buffer (dilution 1:200). Standards and test samples were diluted in 500 µl microcentrifuge tubes. The two standards were prepared by mixing 10 µl of standard solution with 190 µl of working solution. The test

samples were prepared by mixing 2 µl of DNA with 198 µl of working solution. After vortexing the tubes, the samples were incubated at room temperature for 2 minutes prior to measurement in the instrument. A standard curve was calculated by measuring the two standards first. Each test samples were measured and concentration was calculated by the software integrated in the instrument.

Both Nanodrop and Qubit are used to assess the quantity of DNA. However, the total nucleic acid yield can be assessed using the Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA), as it measures all nucleic acids. The Qubit (Thermo Fisher Scientific, Waltham, MA, USA), on the other hand specifically measures dsDNA. A ratio of Qubit / Nanodrop can be calculated to assess the proportion of dsDNA over the total amount of nucleic acids in a sample.

2.4.3 Agarose gel electrophoresis

DNA samples and PCR products size were assessed using agarose gel electrophoresis. It uses an electrical field to move the negatively charged DNA towards a positive electrode through an agarose gel matrix. The smaller the fragment to further it can migrate through the gel matrix. A 3% gel was made by dissolving 6g of agarose (Sigma-Aldrich, St Louis, MO, USA) in 200 ml of 1 x Tris-Borate-EDTA with 0.5 µg/ml ethidium bromide (Sigma-Aldrich, St Louis, MO, USA). The sample was mixed with 1/5 volume of 5x loading dye (2.5% Ficoll, 0.01% bromophenol blue, 0.01% xylene cyanole). It was run in 1x TBE running buffer with DNA ladder, selected according to the expected size of fragments of the samples applying a voltage of 120 V for 45 minutes. The DNA on the gel were

visualised in a UV transilluminator (Bio-Rad, Hercules, CA, USA) under UV light (302 nm).

2.4.4 Real-time PCR assay

In a real time PCR assay a positive reaction is detected by accumulation of a fluorescent signal, such as SYBR green. SYBR green binds to dsDNA. At every PCR cycle, the amount of dsDNA increases and the fluorescence increase with it. The measurement called cycle threshold (Ct) correspond to the number of PCR cycles required for the fluorescent signal to exceed background noise (to be higher than a set threshold). Ct levels are inversely proportional to the amount of target nucleic acid in the sample.

2.4.4.1 DNA quality control: ΔCq assay

The "availability" of the DNA for WGS was assessed by a real time PCR assay ΔCq specifically for DNA extracted from FFPE tissues (Illumina 2016) (Illumina Inc., San Diego, CA, USA). It reflects the amount of DNA molecules that are of sufficient length. Commercially available primers (unpublished sequences) amplify 200 bp fragments of a specific Alu family of which around 400 copies are found throughout the genome. It offers a genome-wide view of the DNA availability. The ΔCq value is calculated by subtracting Ct value of the test sample with the Ct value of a reference sample. A high ΔCq value reflects poor quality of the DNA sample. Depending on the result of the assay the DNA input for library preparation is adjusted.

Control DNA sample ACD1 (FC-121-9999, Illumina Inc., San Diego, CA, USA) and test DNA samples' concentrations were measured using Qubit (Chapter 2.4.2) using the dsDNA broad range assay kit (Q32850, Thermo Fisher Scientific, Waltham, MA, USA) and according to the reading were diluted to 2 ng/µl with a final volume of 20 µl using buffer EB (Q19086, Qiagen, Hilden, Germany) Tris-HCl 10 mM, pH 8.5 as diluent. The dilutions' concentration was measured using the dsDNA high sensitivity assay kit (Q32854, Qiagen, Hilden, Germany) and a dilution of 10 µl at 0.53 ng/ µl was prepared in triplicate for each sample using buffer EB (Q19086, Qiagen, Hilden, Germany) Tris-HCl 10 mM, pH 8.5 as diluent. Each PCR reaction contained 10 µl of KAPA SYBR FAST qPCR (quantitative polymerase chain reaction) Master Mix (2X) (KK4601, Kapa Biosystems, Inc., Wilmington, MA, USA), 2 µl of QC primer 204 (FC-121-9999, Illumina Inc., San Diego, CA, USA), 0.4 ROX Low (50X) (KK4601, Kapa Biosystems, Inc., Wilmington, MA, USA) and 7.6 µl of diluted DNA sample (or non-template control). The plate was sealed, briefly centrifuged and place in the real-time thermocycler 7500 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). The amplification reaction consisted of an initial denaturation step of 50°C for 2 minutes and 95°C for 10 minutes followed by 40 cycles of denaturation at 95°C for 30 seconds, primer annealing at 60°C for 30 seconds and extension at 72°C for 30 seconds. The fluorescence signal was measured during this last step during each cycle. Ct values were calculated for each reaction using the 7500 System SDS Software (Applied Biosystems, Foster City, CA, USA) and the mean Ct was calculated for each sample. ΔCq values were calculated for each test sample as follow:

$$\Delta Cq_{test\ sample} = \overline{Ct}_{test\ sample} - \overline{Ct}_{ACD1}$$

Where $\overline{Ct}$ is the mean of Ct values of the triplicate

ΔCq values were interpreted as detailed in table Table 2.1.

Table 2.1 Input DNA quantity requirements for FFPE library preparation

ΔCq value	Quantity of input DNA used for FFPE library preparation (ng)
≤ 1	250
> 1	400

2.4.4.2 Library quantification

Whole genome libraries were quantified using the real-time PCR Kapa Library Quantification kit (KK4835, KapaBiosystems, Woburn, MA, USA). Ct values of library samples are compared to a standard curve of 6 standard samples to calculate the library molar quantity.

Each purified library was diluted in 1:1000 in a microcentrifuge tube using library Dilution Buffer which contains EB buffer (Q19086, Qiagen, Hilden, Germany) and 0.05% Tween 20 (p7949-500ML, Sigma-Aldrich, St Louis, PA, USA) as follow (1 µl of library + 999 µl of Dilution Buffer). A two-fold serial dilution was prepared in a second micro centrifuge tube using 100 ul of diluted library and 100 ul of Dilution buffer. The 20 µl reaction mix was prepared as followed: 12 µl 2X KAPA SYBR FAST qPCR Master Mix containing Illumina Primer Premix, 0.4 µl low ROX, 3.6 µl of PCR-grade water and 4 µl of template (standard, test sample or non-template control). Sufficient reaction mix

was prepared to run in triplicate reactions 6 standards samples, the two dilutions of each library sample, and a non-template control in a qPCR plate. After the plate was sealed and briefly centrifuged it was placed in the real-time thermocycler 7500 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). The amplification conditions consisted of an initial denaturation step of 95°C for 5 minutes followed by 35 cycles of denaturation at 95°C for 30 seconds and primer annealing and extension at 60°C for 45 seconds. Data acquisition was performed during this last step of the cycle. Analysis of the results was performed as followed: standard curve was calculated using the DNA standard concentration values (Table 2.2).

Table 2.2 DNA standard concentrations of real-time PCR Kapa Library Quantification kit (KK4835, KapaBiosystems, Woburn, MA, USA)

Sample name	dsDNA concentration (pM)
Standard 1	20
Standard 2	2
Standard 3	0.2
Standard 4	0.02
Standard 5	0.002
Standard 6	0.0002

The concentration of each library sample is calculated as follow:

$$\text{concentration of library} = \text{A} \times \frac{452 \times \text{d}}{\text{L}}$$

where *A* is the average concentration (pM) from triplicate results, *L* is the average fragment length, and *d* the dilution factor (1000 or 2000).

The reaction efficiency was confirmed by making sure the 2-fold library DNA dilution fell within the range of 90% - 110%. The calculated concentration of library was used to prepare the appropriate dilution of the library to load on the flow cell for sequencing.

2.5 Sanger sequencing

Validation of the TP53 gene mutations was performed by Sanger sequencing. PCR amplification of exons 2 to 11 was done using the following primers:

5’-TCAGACACTGGCATGGTGTT-3’ (exon 2 and 3 forward)

5’-TCATCTGGACCTGGGTCTTC (exon 2 and 3 reverse)

5’-TCCCAAGCAATGGATGATTT-3’ (exon 4 forward)

5’-CAGGCATTGAAGTCTCATGG-3’ (exon 4 reverse)

5’-TATCTGTTCACTTGTGCCCTG-3’ (exon 5 and 6 forward)

5’-TTGCACATCTCATGGGGTTA-3’ (exon 5 and 6 reverse)

5’-CCTGCTTGCCACAGGTCT-3’ (exon 7 forward)

5’-TGATGAGAGGTGGATGGGTAG-3’ (exon 7 reverse)

5’-CAAGGGTGGTTGGGAGTAGA-3’ (exon 8 and 9 forward)

5’-CCCCAATTGCAGGTAAAACA-3’ (exon 8 and 9 reverse)

5’-GTCAGCTGTATAGGTACTTGAAGTGC-3’ (exon 10 forward)

5’-GAAGGCAGGATGAGAATGGA-3’ (exon 10 reverse)

5’-CCTTAGGCCCTTCAAAGCAT-3’ (exon 11 forward)

5’-AGACCCAAAACCCAAAATGG-3’ (exon 11 reverse)

One PCR reaction contained 1.5 µl of Buffer 10x-MgCl2 (15mM), 0.25 µl of dNTPs (2.5 µM), 0.1 µl of Taq DNA Polymerase (Roche 5U/ µl, Sigma-Aldrich, St Louis, MO, USA), 2.5 µl of primers (2 µM), 6.4 µl of molecular grade water and 2 µl of DNA template diluted to 10 ng/µl. The PCR program was as followed: 95°C for 15 minutes, then 33 cycles of 94°C for 30 seconds, 60°C for 20 seconds and 72°C for 1 minute; the final step was 72°C for 10 minutes. A post-PCR purification was completed by incubating 7.5 µl of PCR product with 1.6 µl of a Exonuclease I + calf intestinal alkaline phosphatase mix (1:2) for 15 minutes at 37°C and 15 minutes at 85°C (to deactivate the PCR enzymes). Dilutions of PCR products were performed at 1:10 for exons 2, 3, 7 and 11 and 1:20 for exons 5, 6, 8, 9 and 10. Sanger sequencing amplification was carried out using Taq DNA polymerase, dNTPs and fluorescent dideoxynucleotides (ddNTPs), primer (same as previously described and purified PCR product. Samples underwent capillary electrophoresis (3130XL Applied Biosystems, Foster City, CA, USA) and analysed with SeqScape v2.6 (Applied Biosystems, Foster City, CA, USA). The genome reference used was IARC NC_000017 version 9.

2.6 Library preparation

DNA molecules are prepared in libraries by being fragmented and tagged with specific sequences necessary for the sequencing reaction. Whole genome library preparation techniques generate DNA libraries covering the totality of the human genome whereas targeted library preparation methods select specifically genomic regions of interest. The TruSeq DNA PCR-Free Library Prep (FC-121-9006DOC, Illumina Inc., San Diego, CA,

USA) was used for all samples except DNA from FFPE specimen. In this case, the TruSeq Nano DNA Library Prep (FC-121-9010DOC, Illumina Inc., San Diego, CA, USA) was used with additional DNA treatments.

2.6.1 Whole genome libraries

Whole genome library preparation consisting in fragmenting the totality of the gDNA and ligating at its end specific DNA sequences (tags) necessary for sequencing.

2.6.1.1 DNA fragmentation and clean up

Genomic DNA was normalised with RSB buffer in a microplate to a final volume of 55 µl at 20 ng/µl for DNA from blood or fresh frozen tissues, 5 ng/µl for DNA from FFPE tissues with a ΔCq value < 1, and 8 ng/µl for DNA from FFPE tissues with a ΔCq value > 1. 52.5 µl of gDNA was transferred to a new Covaris microTUBE (520077, Covaris, Woburn, MA, USA) centrifuged at 280 g for 5 seconds and fragmented to generate 350 bp inserts using the following Covaris M220 settings: Duty factor: 20%, Peak/Displayed power: 50 W, Duration: 65 seconds. After centrifugation of the microTUBE at 280 x g for 5 seconds, 50 µl of the supernatant was transferred to the clean-up microplate. 80 µl of SPB beads were thoroughly mixed and added to the fragment DNA sample, vortexed at 1800 rpm for 2 minutes, incubated at room temperature for 2 minutes and centrifuged at 280 x g for 1 minute. The microplate was then left on a magnetic stand from 8 minutes and supernatant was discarded. Two washes were performed by adding 200 µl of freshly prepared 80% ethanol. All ethanol was then removed and the pellet was air dried on the

magnetic stand for 5 minutes. Following the TruSeq DNA PCR-Free Library Prep 52.5 µl of RSB buffer was added to the microplate and following the TruSeq Nano DNA Library Prep 82.5 µl of RSB buffer was added to the microplate. The microplate was vortexed at 1800 rpm for 2 minutes, centrifuged at 280 x g for 1 minute and placed on the magnetic stand for 5 minutes. 50 µl (PCR-Free protocol) of sample was transferred into a new microplate to be processed for ends repairs and size selection. 80 µl ("nano" protocol) were transferred into a new microplate for FFPE treatments.

2.6.1.2 FFPE DNA library treatment

10 µl of treatment buffer 1 and 10 µl of treatment 1 mix were added to the sample and gently mixed by pipetting. The microplate was incubated at 37°C for 30 minutes and cooled down to 4°C. 5 µl of STM buffer and 168 µl of sample purification beads were added to the sample. The microplate was incubated at room temperature for 5 minutes and placed on a magnetic stand for 5 minutes. 260 µl of supernatant was discarded and 2 washes were performed using 200 µl of 80% ethanol. The sample was left to air dry on the magnetic stand for 5 minutes. The sample was resuspended in 42.5 µl of RSB buffer, incubated 2 minutes at room temperature. The plate was placed on the magnetic stand and after 5 minutes incubation 40 µl of the supernatant was transferred into a new microplate for treatment 2. 5 µl of treatment buffer 2 and 5 µl of treatment 2 mix were added to the sample and mixed by pipetting. As for treatment 1, the plate was incubated at 37°C for 30 minutes, cooled down at 4°C. STM buffer and sample purification beads were added. The

sample was washed twice with 80% ethanol, resuspended in resuspension buffer and 60 µl was transferred to a new microplate for ends repair.

2.6.1.3 Ends repair and size selection

2.6.1.3.1 TruSeq DNA PCR-Free Library Prep

10 µl of CTE and 40 µl of ERP2 were added to the sample. The microplate was shaken at 1800 rpm for 2 minutes, centrifuged at 280 x g for 1 minute, incubated at 30°C for 30 minutes and cooled down on ice. 160 µl of diluted sample purification beads (dilution factor = 2:3) was added to the sample. The microplate was shaken at 1800 rpm for 2 minutes, incubated at room temperature for 5 minutes, centrifuged at 280 x g for 1 minute and placed on a magnetic stand for 5 minutes. 250 µl of the supernatant was transferred to a new microplate. 30 µl of undiluted sample purification beads were added to the microplate. The microplate was shaken at 1800 rpm for 2 minutes, incubated at room temperature for 5 minutes, centrifuged at 280 x g for 1 minute and placed on a magnetic stand for 5 minutes. After the supernatant was discarded the sample was washed twice with 200 µl of 80% ethanol and air dried for 5 minutes on the magnetic stand. 17.5 µl of resuspension buffer was added. The microplate was shaken at 1800 rpm for 2 minutes, centrifuged at 280 x g for 1 minute and placed on the magnetic stand for 5 minutes. 15 µl of the supernatant was transferred to a new microplate.

2.6.1.3.2 TruSeq Nano DNA Library Prep

40 µl of end repair mix 2 was added to the sample, mixed by pipetting, incubated at 30°C for 30 minutes and cooled down on ice. 77 µl of sample purification beads were added to the microplate, mixed by pipetting, incubated at room temperature for 5 minutes and placed on a magnetic stand for 5 minutes. 172 µl of the supernatant was discarded and the sample was washed twice with 200 µl of 80% ethanol. The plate was left to air dry for 5 minutes and 20 µl of resuspension buffer was added to the microplate. The sample was incubated for 2 minutes, placed on the magnetic stand for 5 minutes and 17.5 µl of the supernatant was transferred to a new microplate.

2.6.1.4 3’ end adenylation

Only with the TruSeq DNA PCR-Free Library Prep, 2.5 µl of CTA was added to the sample. Then, 12.5 µl of ATL was added to the microplate, shaken at 1800 rpm for 2 minutes and incubated as follow: 37°C for 30 minutes, 70°C for 5 minutes and 4°C for 5 minutes.

2.6.1.5 Adaptor ligation

The following reagents were added to the sample: 2.5 µl of CTL, 2.5 µl of LIG2 and 2.5 µl of DNA adaptors. The plate was shaken at 1800 rpm for 2 minutes, centrifuged at 280 x g for 1 minute and incubated as follow: 30°C for 10 minutes and cooled down on ice. 5 µl of STL was added to the sample, shaken at 1800 rpm for 2 minutes and centrifuged at 180

x g for 1 minute. 42.5 µl of sample purification beads were added; the plate was shaken at 1800 rpm for 2 minutes, centrifuged at 180 x g for 1 minute and placed on a magnetic stand for 5 minutes. Supernatant was discarded and the pellet was washed twice with 200 µl of 80% ethanol. After air drying for 5 minutes, 52.5 µl of resuspension buffer was added. The plate was shaken and centrifuge as previously described and placed on a magnetic stand for 5 minutes. 50 µl of the supernatant was transferred to a new microplate and the bead purification was repeated with 50 µl of sample purification beads and 22.5 µl of resuspension buffer for the TruSeq DNA PCR-Free Library Prep and 27.5 µl of resuspension buffer for the TruSeq Nano DNA Library Prep. 20 µl (DNA PCR-Free) or 25 µl (Nano DNA) were transferred to a new microplate.

2.6.1.6 DNA fragment enrichment

The TruSeq Nano DNA Library Prep was followed by a PCR amplification using the following reagents: 5 µl of PPC and 20 µl of EPM. The PCR program was as follow: 95°C for 3 minutes, then 8 cycles of 98°C for 20 seconds, 60°C for 15 seconds and 72°C for 30 seconds, then 72°C for 5 minutes. After centrifugation at 280 x g for 1 minute, the sample was cleaned up by adding 50 µl of sample purification beads. The plate was incubated for 5 minutes and placed on a magnetic stand for 5 minutes. The supernatant was removed and the pellet washed twice with 200 µl of 80% ethanol. After air drying the plate, the sample was resuspended with 32.5 µl of resuspension buffer. The plate was incubated at room temperature for 2 minutes and on a magnetic stand for 5 minutes. 30 µl of the supernatant was transferred to a new microplate.

2.6.1.7 Library normalisation

The library sample was quantified by real-time PCR as described in Chapter 2.4.4.2 and normalised to 12 pM.

2.6.2 Illumina TruSeq Custom Amplicon preparation

Amplicon library preparation is a targeted method to selectively amplify and specifically sequence genomic regions of interest. The Illumina TruSeq Custom Amplicon Library Prep kit (FC-130-1001, Illumina Inc., San Diego, CA, USA) was used to prepare 250 bp indexed PE library constructs of custom amplicons from gDNA.

2.6.2.1 Custom amplicon oligo pool design

The custom amplicon oligo pool necessary to select genomic regions of interest was designed and ordered using the Illumina web app Design Studio (Illumina Inc., San Diego, CA, USA), a sequencing assay designer. The designed pool contained oligo to select genomic targets in the genes of interest presented in Table 2.3.

Table 2.3 CLL targeted sequencing panel

Gene name	Description	targets	Cytoband
ATM	ATM Serine/Threonine Kinase	All 63 exons	11q22.3
BIRC3	Baculoviral IAP Repeat Containing 3	All 9 exons	11q22.2
MED12	Mediator Complex Subunit 12	All 45 exons	Xq13.1
MYD88	Myeloid Differentiation Primary Response	Exon 5	3p22.2

	88		
NOTCH1	Notch1	Exon 34	9q34.3
SAMHD1	SAM And HD Domain Containing Deoxynucleoside Triphosphate Triphosphohydrolase 1	All 16 exons	20q11.23
SF3B1	Splicing Factor 3b Subunit 1	Exons 14, 15, 16, 18	2q33.1
TP53	Tumor Protein P53	All 11 exons	17p13.1
XPO1	Exportin 1	Exon 15	2p15

Descriptions from Human gene database GeneCards (Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet 1997).

2.6.2.2 Library generation, clean up and normalisation

Custom oligos are hybridised upstream and downstream of gDNA targets by incubation at a high temperature. Unbound oligos are then washed out using a size selection filter. A DNA polymerase extend region of interest from the 5' oligo and ligate the fragment to the 3' oligo resulting in the production of fragment with the target of interest flanked by sequences necessary for amplification. Indexed primers are then used to amplify all regions of interest in the same reaction (multiplexed PCR). Amplified libraries are cleaned-up using AMPure XP beads (A63881, Beckman Coulter Ltd, High Wycombe, UK) which are magnetic beads specifically binding DNA. Each library is then normalised using beads and pooled together. It is possible as the multiplex PCR reaction introduced

different indices (of 8 bp) for each sample at both end of the DNA fragments. Unless otherwise stated, the reagents used for this protocol were provided in this kit.

First the DNA samples were quantified by Qubit (Chapter 2.4.2) and normalised to 25 ng/μl with a final volume of 10 μl. Then, the oligos were hybridised to the template DNA by adding 5 μl of Custom Amplicon Oligo mix to 10 μl of DNA sample and 35 μl of OHS2 buffer in a microplate and incubated for 80 minutes on heat block pre-heated at 95°C and gradually cooling down to 40°C. Once this temperature was reached, the plate was centrifuged for 1 minute a 1000 g and each sample transferred on a filter plate mounted on a midi collection plate. The plate unit was centrifuged at 2400 g for 2 minutes and wash twice with 45 μl of buffer SW1 and once with buffer UB1 followed by 2 minutes centrifugation at 2400 g. The products were then extended and ligated by adding 45ul of ELM4 to each sample and incubating at 37°C for 45mins. After incubation the plate was centrifuged for 2 minutes at 2400 g, 25 μl of freshly prepared 50 mM sodium hydroxide (NaOH) (72068, Sigma-Aldrich, St Louis, MO, USA) was added to each reaction and mixed well on the filter plate. During 5 minutes incubation at room temperature, the index amplification plate (IAP) was prepared with 4 μl i5 tag, 4 μl i7 tag, 21.6 μl PMM2 and 0.4 μl TDP1 in each well. After the 5 minutes incubation with NaOH (72068, Sigma-Aldrich, St Louis, MO, USA), 20 μl of each sample was transferred to the IAP plate and centrifuged at 1000 g for 1 minute. The amplification conditions consisted of an initial denaturation step of 95°C for 3 minutes followed by 25 cycles of denaturation at 95°C for 30 seconds, primer annealing at 66°C for 30 seconds and primer extension at 72°C for 60 seconds, followed by a final 5 minutes extension at 72°C. 5 μl of the PCR product was run on an electrophoresis gel as described in Chapter 2.4.3. The acceptable

PCR product size was 350 bp. The remaining of DNA samples were transferred to a new midi plate containing 45 μl of AMPure XP beads (A63881, Beckman Coulter Ltd, High Wycombe, UK). After shaking the plate at 1800 rpm for 2 minutes it was incubated at room temperature for 10 minutes. Then, the plate was placed on a magnetic stand and incubated until the liquid was clear. The supernatant was removed and discarded. Two washes were performed by adding 200 μl of freshly prepared 80% ethanol (20821.296, VWR Chemicals, Radnor, PA, USA), incubating on a magnetic stand for 30 seconds and discarding the supernatant. Remaining ethanol was removed and the plate was air dried for 10 minutes. The beads were suspended in 30 μl of EBT, shaken at 1800 rpm for 2 minutes and incubated at room temperature for 2 minutes. The plate was placed on a magnetic stand until the liquid was clear; 20 μl of the supernatant was transferred to a new microplate and centrifuged at 1000 g for 1 minute. The libraries were normalised to have equal representation each library in the pool by incubating a pre mixed solution of 38 μl LNA1 + 7 μl LNB1 per sample. The plate was shaken for 30 minutes at 1800 rpm and immediately placed on a magnetic stand until the liquid was cleared. The supernatant was discarded and samples were washed twice with 45 μl of LNW1 buffer, shaken at 1800 rpm for 5 minutes, placed on a magnetic stand and supernatant discarded. 30 μl of freshly prepared 0.1 NaOH (72068, Sigma-Aldrich, St Louis, MO, USA) was added to each sample and the plate shaken at 1800 rpm for 5 minutes. After placing the plate to a magnetic stand 30 μl of the supernatant is transferred to a new plate containing 30 μl of LNS2. Libraries were combined by adding 5 μl of normalised samples in a single tube called pooled amplicon library (PAL).

2.6.3 Ampliseq Cancer panel

2.6.3.1 Libraries generation

Libraries were generated from DNA samples normalised at 10 ng/µl in the following reaction mix: 4 µl of 5X Ion Ampliseq HIFI Master mix (Thermo Fisher Scientific, Waltham, MA, USA), 4 µl of 5X Ion Ampliseq Primer pool, 1 µl of DNA sample, 11 µl of nuclease free water. The sample was prepared in a 0.2 ml microcentrifuge tube and thoroughly mixed by pipetting. The amplification reaction was carried out as follow: 99°C for 2 minutes and 21 cycles of 99°C for 15 seconds and 60°C for 4 minutes. Library purification was completed by adding 2 µl of FuPa reagent to the sample and incubated following the digestion program: 50°C for 10 minutes, 55°C for 10 minutes and 60°C for 20 minutes. To each tube containing the digested sample, 4 µl of Switch solution, 1 µl of Ion P1 adapter, 1 µl of Ion Xpress Barcode, 2 µl of nuclease free water and finally 2 µl of DNA ligase. The ligation program was as follow: 22°C for 30 minutes and 72°C for 10 minutes. The sample was transferred to a new 1.5 ml microcentrifuge tube containing 45 µl of Axyprep beads (Mag-PCR-CL-250, Axygen Biosciences, Union City, CA, USA) and mixed by pipetting. After incubation for 5 minutes at room temperature, the sample was placed on a magnetic stand for 2 minutes and the supernatant was discarded. The sample was washed twice by adding 150 µl of freshly prepared 70% ethanol (20821.296, VWR Chemicals, Radnor, PA, USA) and removing the supernatant. The sample's pellet is dried on the magnetic stand at room temperature for 5 minutes. Elution is performed off the magnetic stand by adding 50 µl of Low TE to the sample and mixing the sample by

pipetting. After the microcentrifuge tube is placed on the magnetic stand and incubated for 2 minutes, the supernatant is transferred to a new 1.5 ml microcentrifuge tube. Libraries prepared were quantified by real time PCR. Three standards were prepared from E. coli DH10B Ion Control Library (68pM) at 6.8 pM, 0.68 pM and 0.068 pM. The sample was diluted in nuclease free water (1:100). The PCR reaction mixture was prepared as follow: 10 µl of 2X TaqMan MasterMix and 1 µl of 20X Ion TaqMan Assay and 9 µl of diluted library/standard/non template control. Each sample was prepared in duplicate. The amplification was carried out using an ABI 7500 Fast System real-time instrument as follow: 50°C for 2 minutes, 95°C for 20 seconds and 40 cycles of 95°C for 3 seconds and 60°C for 31 seconds. The library concentration is calculated using standards and multiplying the value by a factor 100.

2.6.3.2 Preparation of enriched template Ion sphere particles (ISPs)

First, the amplification plate was installed in the OneTouch instrument (Thermo Fisher Scientific, Waltham, MA, USA). After the disposable tubing and injector were installed and the OneTouch Reagent Tube filled with OneTouch oil the amplification solution was prepared as follow: 2 µl of each library sample, diluted to 100 pM was transferred to a 0.2 ml microcentrifuge tube. In a new 0.5 ml microcentrifuge tube 4 µl of pooled library just prepared and 46 µl of nuclease free water were combined. In a 1.5 ml microcentrifuge tube the follow reagents were mixed together: 25 µl of nuclease free water, 500 µl of Ion PGM (personal genome machine) Template 0T2 200 Reagent Mix, 300 µl Ion PGM Template 0T2 200 PCR Reagent B, 50 µl of Ion PGM Template 0T2 200 Enzyme mix and

25 µl of diluted library. 100 µl of thoroughly resuspended Ion PGM Template OT2 200 ISP (beads) was added to the tube. The entire 1000 µl was slowly transferred to the filter assembly as well as 1500 µl of Reaction Oil. The assembly was then placed on the Ion OneTouch instrument and the run started. The supernatant was removed from the two Ion OneTouch Recovery tubes leaving approximately 50 µl of solution in the tube to resuspend the beads by pipetting the beads up and down. Then, enrichment of the template-positive ISPs was carried out in the OneTouch ES instrument, by loading 130 µl of DynabeadsR MyOne Streptavidin C1 Beads resuspended in MyOne Beads wash solution, 3 x 300 µl of Ion OneTouch Wash Solution, 300 µl of freshly prepared Melt-Off solution and a 0.2 ml microcentrifuge tube containing 10 µl of Neutralisation solution. After the run, the 0.2 ml microcentrifuge tube containing the enriched material was taken out of the instrument for sequencing.

2.7 Massively parallel sequencing

MPS on Illumina instruments (Illumina Inc., San Diego, CA, USA) consists in hybridising library constructs on a flow cell to generate clusters using the Illumina adaptors ligated at each end of the construct followed by sequencing by synthesis. This method is based on elongating the DNA construct and detecting single dNTP by emitting fluorescence as they are incorporated.

2.7.1 Whole genome paired-end sequencing on HiSeq2500

The first step was cluster generation. It was carried out on a cBot system (Illumina Inc., San Diego, CA, USA). After the instrument pre-run wash, the consumables (reagent plate, flow cell and manifold) are loaded in the instrument as well as the samples. At the end of the run the hybridised flow cell was stored a 4°C in storage buffer for up to 10 days.

The second step was to perform the sequencing on the flow cell using the technique of sequencing by synthesis. The sequencing was performed on a Hiseq2500 (Illumina Inc., San Diego, CA, USA) following a High Output v4 mode, 2x125 cycles (paired end, non-indexed). The SBS reagents and PE reagents were loaded into the HiSeq instrument as well as the prime flow cell. After the flow check was conducted, the sequencing flow cell was inserted instead of the prime flow cell.

2.7.2 Targeted paired-end sequencing on MiSeq

Samples underwent next generation sequencing using the MiSeq Reagent Kit v2 300-cycle kit (MS-102-2002, Illumina Inc., San Diego, CA, USA). A diluted amplicon library sample was prepared combining 6 µl of PAL and 594 µl of buffer HT1 and loaded into the reagents cartridge. Cluster generation and sequencing by synthesis was performed subsequently in the MiSeq instrument (SY-410-1003, Illumina Inc., San Diego, CA, USA). Paired end sequencing was completed as described in the following: 150 cycles for DNA insert, 8 cycles for index i7, 8 cycles for index i5, and 150 cycles for DNA insert.

2.7.3 Targeted paired-end sequencing on PGM

After initialisation of the PGM System sequencer with wash buffers, each dNTPs reagents were loaded on the PGM instrument (Thermo Fisher Scientific, Waltham, MA, USA). The sample was then prepared as follow: 5 µl of control ISPs were directly added to the template-positive ISPs in a 0.2 ml microcentrifuge tube. After 2 minutes centrifugation at 12,800 rpm the supernatant was removed leaving 15 µl in the tube and 12 µl of SP was added. The pellet was disrupted by pipetting and the sample was incubated at 95°C for 2 minutes and 37°C for 2 minutes. Then, 3 µl of PGM 200 sequencing polymerase was added to the sample and incubated at room temperature for 5 minutes. After chip preparation, 30 µl of the sample was loaded to the chip through the loading port. The chip was centrifuged for 30 seconds and then the sample was mixed in the chip by pipetting. The process was repeated 3 times and then the liquid was removed from the chip by pipetting and centrifugation. The chip was then set up on the PGM and the sequencing run was started.

2.8 Whole genome sequencing bioinformatics analysis

2.8.1 Reads mapping and metrics calculation

Read mapping is the first step in order to detect genetic differences between different organisms of the same species. In the context of resequencing, mapping consists in comparing each read sequence with a reference genome and finding the unique (or highest match) location on the genome for that read. Unaligned raw short reads, contained in

FASTQ files, are aligned to the reference human genome using mapping algorithms producing a BAM file. Statistical metrics are calculated from the BAM file using a set of command line tools developed to manipulate such files.

2.8.1.1 Solid tumour dataset

Read alignment to Genome Reference Consortium human (GRCh) reference 37.1was achieved using a combination of Isaac aligner version 01.14.02.06 and version SAAC00776.15.01.27 (Raczy et al. 2013). QC was done with Isaac Variant Caller (Illumina 2014). Alignment metrics such as median insert size, read PF aligned ratio, chimeric pairs ratio, AT dropout and GC dropout were calculated using Picard tools:

• Median inserts size: median length between extremities of paired reads. The higher the better, there is a size selection during the library preparation for 350bp fragments.

• Read PF aligned ratio: fractions of sequencing reads that have a position on the reference genome. The higher the better, below 80% is indicative of poor quality sample.

• Chimeric pair ratio: Paired reads which align to different regions or chromosomes.

• AT dropout: a measure of how less covered regions with GC content < 50% are compared to the mean.

• GC dropout: a measure of how less covered regions with GC content > 50% are compared to the mean.

Percent aligned to the human genome reference and High AT/GC normalised coverage indicate coverage uniformity. Insert sizes and chimeric reads values correlate with the number of potential somatic false positive variants.

It is important to highlight that these measurements are slightly dependent on the library prep method. In order to increase library quantity, smaller fragments are included in the FFPE preparation. In addition, the PCR step specific to FFPE library preparation can lead to some biases. Sequencing coverage metrics were calculated using an in-house coverage analysis tool and CODOC (Popitsch 2014). We also measured the number of bases covered at different thresholds (from less than 3 x to more than 100 x) and the uniformity of sequencing coverage by calculating the SD of the sequencing depth of 100kb genomic windows. A high value indicated a highly variable sequencing coverage directly impacting on CNA detection. Poor sequencing coverage uniformity is an issue as it leads to waste of sequencing data by over sequencing some regions of the genome and impacts the quality of variant detection with lower coverage in other part of the genome.

2.8.1.2 CLL dataset

Reads were aligned using Isaac aligner v 03.16.02.19 on the reference genome GRCh38. Sequencing QC metrics were calculated using Bwtool for depth of coverage and SAMtools for other metrics such as GC and AT dropout chimeric pairs, mapping rate and size fragment mean (Pohl & Beato 2014; Li et al. 2009).

2.8.2 Somatic variant calling

Variant calling is the second step in order to detect genetic differences between different organisms of the same species. From aligned reads, variant callers can identify differences between the read sequence and the reference sequence in order to call the presence of a variant. In the context of somatic variant detection, reads from multiple tissues can be

compared in order to detect differences (in this instance: tumour vs. normal tissue). Many tools exist to identify variants, using different methods and all presenting with advantages and drawbacks. To overcome potential weaknesses of a tool, it is common practice to use several of them in parallel and to compare and unify their results to increase variant detection.

SVs are defined as variant greater than 1kb in size (Freeman et al. 2006). Tools to discover SVs generally using read depth and/or split read alignments.

2.8.2.1 Solid tumour dataset

Variant callers are employed to uncover true variants among the mismatched calls from aligned sequencing data. Somatic variants are detected by comparing alignment data of a test sample (tumour) and a matching normal sample (GL). As most bioinformatics algorithms are developed for sequencing data originating from fresh tissue DNA, diverse variant caller give different results.

Therefore, to ensure that the differences originated from the samples and not from poor performance of the algorithms on FFPE data somatic SNV detection was performed with Mutect v1.1.4 (Cibulskis et al. 2013), Shimmer v0.1.1 (Hansen et al. 2013) and Strelka 2.0.14 (Saunders et al. 2012). Indel (<50bp) detection was performed using Shimmer and Strelka.

An in-house method, "VCF Intersect" (using VCF files: variant call format), was used to intersect SNV and indel calls from the three tools. In brief, VCF Intersect compares the output (VCF files) of variant calling pipelines by performing several steps: data cleaning (splitting complex variants, normalising chromosome names and order),

including/excluding genomic regions, performing scoped comparison and generating quality metrics (summary statistics, Venn diagrams and VCF files for manual inspection or follow up computations). Only calls identified by all three methods (Mutect, Shimmer and Strelka), according to VCF Intersect, were considered high confidence. These high confidence mutations cancel the potential specific bias of each algorithm as they are detected by all three variant callers. Variant calling was performed in an unbiased way (i.e. in FF and FFPE, independently) samples to efficiently assess the potential of variant calling in FFPE specimens.

Variants were annotated with Ensembl VEP GRCh37 release 85 (McLaren et al. 2010). Variant statistics, such as number of SNPs, number of transitions, number of transversions, and number of variants in each substitution class were calculated with BCFtools v1.1 (Li et al. 2009).

The unfiltered lists of somatic mutations from FF and FFPE sample datasets were compared directly using VCF Intersect to identify variants detected either only in FF or only in FFPE sample data (labelled as “FF unique” and “FFPE unique”) and variants detected in both samples (named “FF-FFPE overlap”). For this analysis, we calculated the sensitivity and positive predictive value (PPV) using FF sample data as the reference dataset. Somatic SNVs detected only in one sample of the pair were investigated further using the dataset produced with Mutect as it offered the best balance between sensitivity, PPV and number of SNVs detected in both FF and FFPE DNA samples and investigated the presence of potential deamination artefacts and the AF directly computed from the BAM alignment files using vcfBAFAnnotator. The variant detection limits for clinically

relevant somatic SNVs and indels using FFPE data were tested using the Strelka variant datasets as these showed the best sensitivity in both SNV and indel detection.

Regarding copy number changes detection, Log_2 ratio (Log_2R) were derived from read depth and calculated from tumour-normal pair data using BIC-seq v 1.2.1 (Xi et al. 2011). In addition, somatic CNAs were called with Nexus Discovery Edition 7.5 (BioDiscovery, Inc., El Segundo, CA, USA) using Log_2R and BAF data extracted from the read alignments. The Log_2R signal was extracted using the ngCGH script (Davis 2016) with a configured window size of 500 reads. BAF information was obtained from VCF files generated by Platypus 0.5.2 (Rimmer et al. 2014). The paired Log_2R and BAF signals were then visualised in Nexus Copy Number, segmented and manually curated.

Purity estimation was computed from Canvas (Roller et al. 2016). The comparison of purity assessment from WGS data and visual assessment by pathologist is challenging because of the discrepancy in defining tumour purity by these two methods. The computational assessments of tumour purity using sequencing data will identify the proportion of reads with somatic variants (SNVs, CNAs, or both, depending on the algorithms) and infer the proportion of cells in the sample carrying these somatic mutations (i.e. tumour cells). The computational purity assessment is therefore dependent on sampling and the alignment performances. On the other hand, pathologists visually quantify the proportion of cells and depending on the tissue type might consider some stromal cell infiltration as part of the malignant process. Another explanation for estimation variability for this method could be that the tumour cells are bigger than stromal cells and inflammatory cells and therefore are more visible. This “overestimation” has been previously reported in another study by Viray and colleagues (Viray et al. 2013).

2.8.2.2 CLL dataset

Somatic variants were detected using Strelka 2.4.7 (Illumina Inc., San Diego, CA, USA) for small variants detection (SNVs and indels). They were filtered as follow: depth required was greater than 10 and AF greater than 0.05, the quality filter annotation should be "PASS" and quality score greater than 30. After annotations using Ensembl VEP GRCh38 release 89.4 (McLaren et al. 2016), variants were further filtered using allele frequencies from normal populations (databases 1KGP phase 3 1405.34_GRCh38.p8 and EXAC v0.3) to keep only variants with a frequency lower than 0.05.

Regions of kataegis were identified as previously described by Lawrence and colleagues (Lawrence et al. 2013; Nik-Zainal, Alexandrov, et al. 2012) (see Chapter 1.3.3). All the filtered variants were plotted across the whole genome. The genome was binned into 100kb windows and examined for six mutations with less than two standard deviations below the median of intermutational distance of SNVs, or at least 6 variants in 1 kb.

Mutational signatures were calculated using the relative contributions of each base substitution type (A>T, A>C, A>G, C>A, C>T, C>G) on the genome of each patient. Nonnegative matrix factorisation (NMF) was performed as previously described (Nik-Zainal, Alexandrov, et al. 2012) to find the mutational signatures

CNAs were called using Canvas 1.3.1 and SVs were called using Manta 0.28.0. Variants were filtered to only keep those with filter "PASS".

2.9 Targeted massively parallel sequencing bioinformatics analysis

2.9.1 Illumina MiSeq analysis

The program MiSeq Reporter v.2.3.32.0 (Illumina Inc., San Diego, CA, USA) generated demultiplexed and paired FASTQ files. The files were aligned to the reference sequences using a banded Smith-Waterman algorithm. Variants were identified using Illumina's somatic variant caller, an algorithm specifically designed to detect subclonal variants. The data was also aligned with an alternative method: Stampy v.1.0.22 (Lunter & Goodson 2011). From this alignment variants were called using Platypus v.0.5.1 (Rimmer et al. 2014) using default parameters. After VCF files from Illumina's somatic variant caller and Platypus were merged, variants were annotated using Annovar (Wang et al. 2010). The following databases were used: hg19 refGene, dbSNP v.132 and v.137, 1KGP v. April 2012, COSMIC v.67 and prediction scores SIFT and Polyphen (from ljb2). Variants filtered out were synonymous, outside coding and splicing regions or found at position of common polymorphisms (described in 1KGP). At least 50 reads at the position of the variant were required and a variant allelic of 0.08 or more. For less covered variants (depth between 20 and 50x), the minimal variant AF required was 0.3. The remaining variants were classified as described in

Table 2.4.

Table 2.4 Targeted sequencing variants classification

group	description
1	Variants previously described in the CLL literature
2	Variants previously described in other cancers (in at least 2 cases in the COSMIC database)
3	Variant not previously described, characterised as “novel” but with a deleterious SIFT score.

2.9.2 Ampliseq cancer panel analysis

The Torrent Suite Software was used to analyse the data on a local Torrent server. Alignment was performed with default parameters against the hg19 reference genome and variant calling was carried out using the design target regions and hotspot regions supplied with the Ampliseq cancer panel. The software Ion Reporter was used to report the variants.

Variants were visually inspected using the Integrative Genomics Viewer (IGV) (Grever et al. 2007).

2.10 Single Nucleotide Polymorphism microarray

Genome wide genotyping array are reliable methods to detect of copy number variations across the whole genome. Such as amplifications, deletions and cnLOH (copy neutral loss of heterozygosity). Samples were analysed using HumanOmni1-Quad SNP chips (WG-

311-1110, Illumina Inc., San Diego, CA, USA). Unless otherwise stated, the reagents used for this protocol were provided in these kits.

2.10.1 Illumina Genome-wide SNP platform hybridisations

The first step consists in a whole genome amplification of the gDNA. Amplified fragments are then fragmented and cleaned up by alcohol precipitation. Samples are annealed to locus specific sequences linked to BeadChip. Sequence is extended at the locus of the interrogated polymorphism and stained. The fluorescence emitted from the chip is measured by an iScan instrument (Illumina Inc., San Diego, CA USA) and then images are converted into genotype calls.

DNA samples were normalised at 50 ng/μl and 4 μl were transferred into a microplate containing 20 μl of MA1. Samples were denatured by incubation in 4 μl 0.1N NaOH (72068, Sigma-Aldrich, St Louis, MO, USA) for 10 minutes. 68 μl of MA2 and 75 μl of MSM were added to the sample and the plate was vortexed for 1 minute at 1600 rpm, pulse centrifuged and incubated for 20 hours at 37°C. Next, amplified DNA samples were fragmented by adding 50 μl of FMS, vortexing the plate at 1600 rpm for 1 minute followed by pulse centrifugation and incubating 1 hour at 37°C. Then, the samples were precipitated by adding 100 μl of PMS vortexing for 1 minute at 1600 rpm followed by a pulse centrifugation, incubating 5 minutes at 37°C, adding 300 μl of 100% 2-propanol (I9516, Sigma Aldrich), mixing well, incubating at 4°C for 30mins and then centrifuging at 3,000 xg for 20 minutes at 4°C, before removing the supernatant. The pellet containing the DNA was left to air-dry for 1 hour at room temperature and resuspended by incubation

at 48°C for 1 hour in 46 µl of RA1, vortexed at 1800 rpm for 1 minute and pulse centrifuged. Samples were denatured at 95°C for 20 minutes and cooled down for 30 minutes at room temperature. After pulse centrifugation 35 µl of DNA was loaded on the BeadChip and incubated at 48°C for 16 hours. After incubation, the BeadChip was washed in two PB1 baths for 1 minute successively and processed to single base extension. The BeadChip, incubated at 44°C, was washed 5 times with 150 µl of RA1, incubated with 225 µl of LX1 for 10 minutes twice, then incubated with 225 µl of LX2 for 10 minutes twice, incubated with 300 µl of EML for 15 minutes and 250 µl of 95% formamide/1mM EDTA three times. After 5 minutes incubation, the incubation temperature was dropped to 32°C and incubated three times with 250 µl of XC3. The staining part went as follow: 10 minutes incubation in presence of 250 µl of SML, three washes with 250 µl of XC3 followed by a 5 minutes incubation, 10 minutes incubation with 250 µl of ATM and three washes with 250 µl of XC3. Another SML incubation, followed by ATM and then SML again was performed successively interspersed by XC3 washes as previously described. The BeadChip was then washed in a PB1 bath for 5 minutes and coated in XC4 by incubating it in a XC4 bath for 5 minutes. The BeadChip was dried in a vacuum desiccator for 55 minutes at 675 mm Hg (0.9 bar). Once the coating is dry, the BeadChip underside was wiped cleaned with Ethanol and placed in the iScan instrument (Illumina Inc., San Diego, CA USA) for image acquisition.

2.10.2 Analysis pipeline with Nexus Copy Number 8.0 (BioDiscovery)

Once the scan was completed, two intensity file .idat specifying the intensity values for each bead were produced for each sample. These files were loaded into the software GenomeStudio (Illumina Inc., San Diego, CA USA) to generate a final report (.txt file) containing the genotyping calls for each bead. The text file .txt generated by GenomeStudio was analysed using Nexus 8 Discovery Edition (BioDiscovery, El Segundo, CA, USA) with the following settings (SNPRank Segmentation): Significance Threshold 1x 10- 5; Max Contiguous Probe Spacing (kb) 1000.0; Min number of probes per segment 5; High Gain 0.6; Gain 0.2; Loss -0.2; Big Loss -1.0; 3:1 sex chromosome gain 1.2; Homozygous Frequency threshold 0.95; Homozygous Value Threshold 0.8; Heterozygous Imbalance Threshold 0.4; Minimum loss of heterozygosity (LOH) Length (Kb) 20; percentage outliers to remove 3%. All Nexus processed plots were also inspected visually to scan for changes not identified using these analysis settings. The minimum size of events reported were 50kb for copy number gains and losses and 1 Mb for cnLOH.

Chapter 3 Assessment of whole genome sequencing data from routine formalin-fixed, paraffin-embedded (FFPE) specimens

3.1 Introduction

FFPE specimens are broadly available for all cancer patients in routine diagnostics. However, DNA samples extracted from FFPE have alterations that can impede downstream analysis such as MPS analysis. It is well established that FFPE specimens can be used for targeted sequencing panel analysis (Wagle et al. 2012; Hamblin et al. 2017). In the past decade, several studies have compared sequencing data obtained from FF and FFPE specimens and have shown that FFPE samples can be interrogated using targeted sequencing, including whole exome approaches (Menon et al. 2012; Holley et al. 2012; Hedegaard et al. 2014; Van Allen et al. 2014; Munchel et al. 2015; Astolfi et al. 2015; Oh et al. 2015; De Paoli-Iseppi et al. 2016). However, there is a paucity of publications evaluating WGS data (Munchel et al. 2015; Wood et al. 2010; Schweiger et al. 2009). Although overall a relatively high concordance between FF and FFPE was reported with respect to SNV and indel detection, this was variable between regions and dependent on

the specific coverage. In addition, the data quality was variable across platforms and specimen types (Van Allen et al. 2014; Kerick et al. 2011). Furthermore, most of these studies did not use a matched GL sample and therefore differences in the somatic variants called between FF and FFPE have not been investigated rigorously. Only two studies performing WES considered somatic SNVs (using GL sample data) in exonic regions in four cancer samples (Astolfi et al. 2015; Oh et al. 2015). Previous studies have disagreed as to whether CNA calls between FF and FFPE were comparable (Van Allen et al. 2014; Schweiger et al. 2009) or poorer in FFPE samples (Menon et al. 2012). In the first two studies, low depth WGS or WES without matching GL samples were used and in the third study WES of one patient (FF, FFPE and matched GL samples) was reported. Therefore, there is a lack of knowledge regarding somatic mutation detection performance using WGS (in particular SVs and variants in non-coding regions) from FFPE specimens.

In the work described in this chapter, I addressed key technical questions relating to WGS analysis of DNA from FFPE derived cancer samples. The differences observed between sequencing data obtained from FF and FFPE were comprehensively detailed in 156 genomes from 52 matched FF tumour, FFPE tumour and peripheral blood GL samples collected as part of the clinical routine examinations.

3.2 Aims

The aims of this work are to (i) assess the feasibility of using FFPE specimens collected via routine pathology pathways as a source of DNA for WGS, (ii) compare WGS data

quality metrics between matched FF and FFPE specimens and (iii) evaluate variant detection in FFPE specimens: SNVs, indels and CNAs.

3.3 Materials and methods

3.3.1 Patients and samples

One hundred and eighty-four consecutive patients with cancer, undergoing surgical resection with curative intent, were recruited into the study at the Oxford University Hospitals Foundation Trust's Genomics Medicine Centre. Their surgery was carried out between March 2014 and July 2015. At the time of resection, the maximum cold ischaemia time was 2 hours. The tissue samples were prepared as both fresh frozen and FFPE samples, as described in Chapter 2.2. A peripheral blood sample (2 ml) was also collected. For each patient a trio of DNA samples was required to carry out the analysis: a GL sample from peripheral blood and two tumour specimens prepared as FF and FFPE samples.

3.3.2 Nucleic Acid Extraction

DNA extraction was carried out as described in Chapter 2.3.1 for tissues prepared as FF, Chapter 2.3.2.2 for FFPE specimens, and in Chapter 2.3.3 for buffy coats from blood. DNA quantity and quality were assessed by Qubit (Thermo Fisher Scientific, Waltham, MA, USA), Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA) and 3% agarose

gel electrophoresis as detailed in Chapter 2.4.1, 2.4.2 and 2.4.3. DNA derived from FFPE tissues was also measured by a real-time PCR assay: ΔCq (Illumina Inc., San Diego, CA, USA) as per Chapter 2.4.4.1. To pass DNA QC, the minimum yield required was 1 µg for all FF and GL samples and 250 ng for FFPE samples with a ΔCq value above 1 and 400 ng for FFPE samples with a ΔCq value above 1.

3.3.3 Sequencing

Whole genome libraries were prepared from extracted DNA samples, as described in Chapter 2.6.1 and sequencing was performed on a HiSeq2500 High Output v4 mode, 2x125 cycles (paired end, non-indexed) (Illumina Inc., San Diego, CA, USA) (Chapter 2.7.1).

Targeted sequencing libraries of the AmpliSeq Cancer Hotspot Panel (Thermo Fisher Scientific, Waltham, MA, USA) were prepared and sequenced on a PGM instrument (Thermo Fisher Scientific, Waltham, MA, USA) as described in Chapters 2.6.3 and 2.7.3.

3.3.4 Bioinformatics analysis

WGS reads mapping and quality metrics were calculated as described in Chapter 2.8.1.1. WGS variant calling was described in Chapters 2.8.2.1 and 2.9.2. Purity assessment from WGS data was detailed in Chapter 2.8.2.1.

The variant detected by the AmpliSeq Cancer Hotspot Panel were reported as described in Chapter 2.9.2.

3.3.5 Statistical analysis

Continuous data comparisons were performed using t-tests for normally distributed variables and Wilcoxon signed rank tests for asymmetrical distributions. Multiple testing adjustments of p-values were performed using Bonferroni corrections. Data transformation was performed using a Blom transformation. Linear relationships were calculated with Spearman's rank correlation coefficients r. For all statistical analysis, the null hypothesis was rejected when the p-value was less than 0.05. SNV and indel detection sensitivity was calculated as follows: number of variants detected in FF and FFPE divided by the number of all variants detected in FF. The PPV was calculated as follows: number of variants detected in FF and FFPE divided by the number of all variants detected in FFPE.

3.3.6 Individual contributions

In this chapter, sample collection and histopathology processing was performed in the Department of Cellular Pathology of the Oxford University Hospital Foundation Trust. Whole genome library preparation, sequencing and initial data alignment was performed by Illumina Cambridge Ltd.. The AmpliSeq Cancer Hotspot Panel library preparation and sequencing was performed by Ms Suzanne Page. The in-house bioinformatics tools "VCF Intersect" and "vcfBAFAnnotator" were developed by Dr Niko Popitsch.

3.4 Results

3.4.1 A significant number of samples were lost to the study due to poor quality

Patient dropout was measured for sample collection, DNA QC and library QC (Figure 3.1A-B). A significant number of patients were excluded from the analysis cohort as they had at least one sample of the trio either missing or failing QC. Importantly, as many as 87 patients were excluded due to the lack of a suitable FF sample. Of these, 60 patients did not have any FF collected at all for either biology-related reasons (sample too small) or practical reasons (sample collected out of hours), for the other 27 the purity of the FF sample collected was not sufficient (<40% tumour content). Thirty of the remaining patients were excluded because the quality of FFPE DNA samples (22 patients) or libraries (eight patients) failed QC. As another 15 patients were excluded at various other QC steps, full trio sets were available for 52 cancer patients out of the original 184. This cohort comprised 10 breasts, 12 colorectal, seven endometrial, four prostatic, 14 renal and five thoracic tissues. The diagnosis of each patient and all clinical characteristics are detailed in Appendix A.

3.4.2 FFPE DNA quality control revealed short and denatured DNA

The median time between specimen collection and DNA extraction was slightly higher for FF samples (66 days) than FFPE samples (56 days), but the variation was large so this was not significantly different (p=0.526, paired Wilcoxon signed rank tests and Bonferroni

corrected p-value). Thus, the quantity and quality of the 52 matching FF and FFPE DNAs were compared using established multiple metrics using paired Wilcoxon signed rank tests (Table 3.1).

The total nucleic acid yield assessed using a Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA), measuring all nucleic acids, was lower for FFPE than FF DNA samples, but the variability was large so this did not reach statistical significance ($p=0.0781$). The same metric measured by Qubit dsDNA Broad Range Assay kit (Thermo Fisher Scientific, Waltham, MA, USA), measuring specifically dsDNA showed a significantly lower DNA quantity for FFPE samples ($p=1.41e^{-06}$). These results supported the assertion that FFPE samples presented with similar amounts of total nucleic acids in the sample, but lower dsDNA quantity due to denaturation.

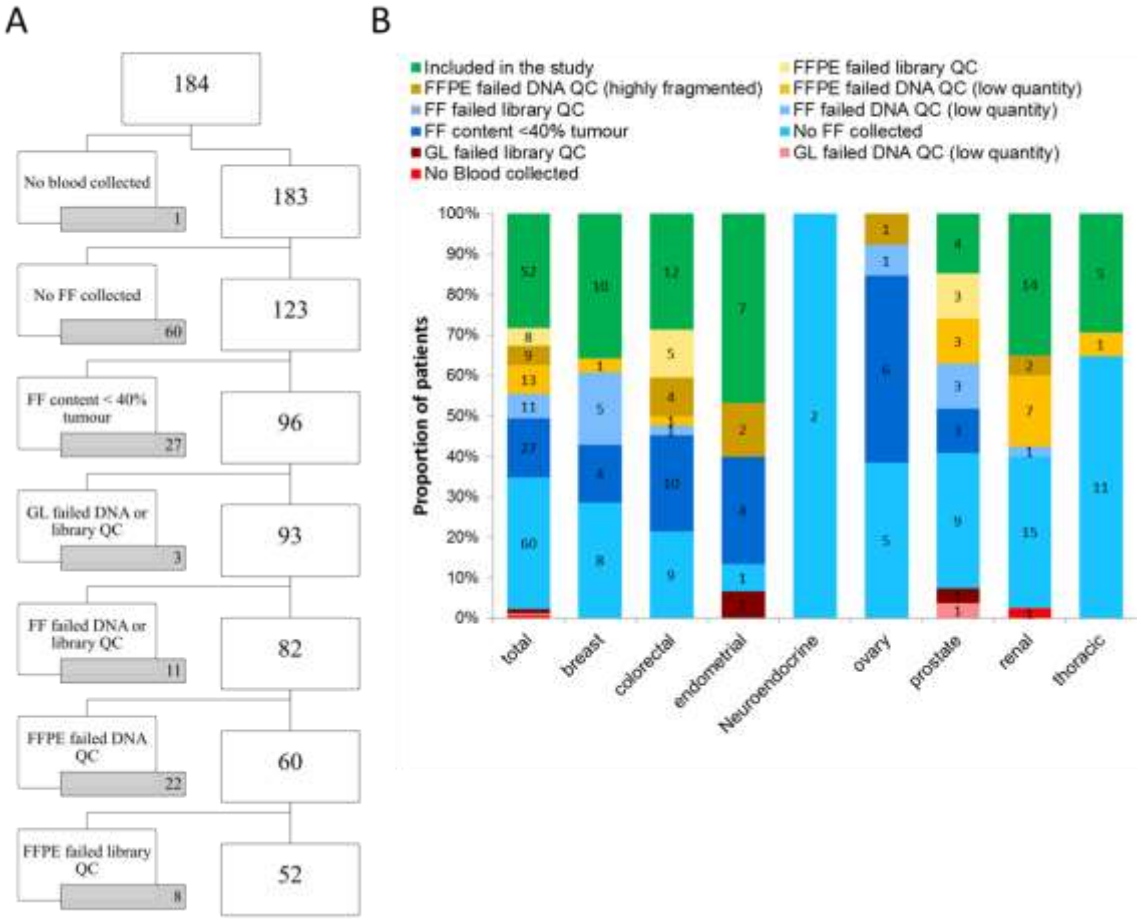


Figure 3.1 Patient dropout at different quality control steps

(A) Summary overview of the patient dropout and (B) detailed quality control dropout for each tissue type. Dropout was measured for sample collection, DNA quality control and library quality control.

FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded sample, QC: Quality Control, GL: germline (from Blood)

Table 3.1 Comparison of DNA metrics in FF and matching FFPE tissue datasets

	FF	FFPE	Corrected p-value
	Median (range)		
Time between tissue collection and DNA extraction (day)	66 (7 - 210)	56 (12 - 184)	0.526
DNA yield assessed by Qubit (μg)	7.7 (1.1 - 55.0)	3.0 (0.5 - 30.1)	0.00000141
DNA yield assessed by Nanodrop (μg)	7.9 (1.1 - 68.2)	7.1 (1.6 - 63.2)	0.0781
A260/A280 ratio	1.88 (1.67 - 1.99)	1.89 (1.62 - 1.95)	0.994

Wilcoxon signed rank test. P-value corrected with Bonferroni method. FF: Fresh frozen, FFPE: formalin-fixed, paraffin-embedded

In addition, the A260/A280 ratio, a marker of nucleic acid purity, was very similar between FFPE and FF samples with mean values of 1.89 and 1.88 respectively. However, when subjected to standard gel electrophoresis the FFPE DNA samples revealed nucleic acid smears in the range <0.5 – 40kb, indicating DNA fragmentation, whereas the matching FF samples presented a distinctive band at >40kb, the limit of electrophoresis resolution (Figure 3.2).

These results showed DNA samples extracted from FFPE samples presented the same amount of total nucleic acids and same purity as samples extracted from FF samples. However, the proportion of long and dsDNA molecules was significantly lower in FFPE samples.

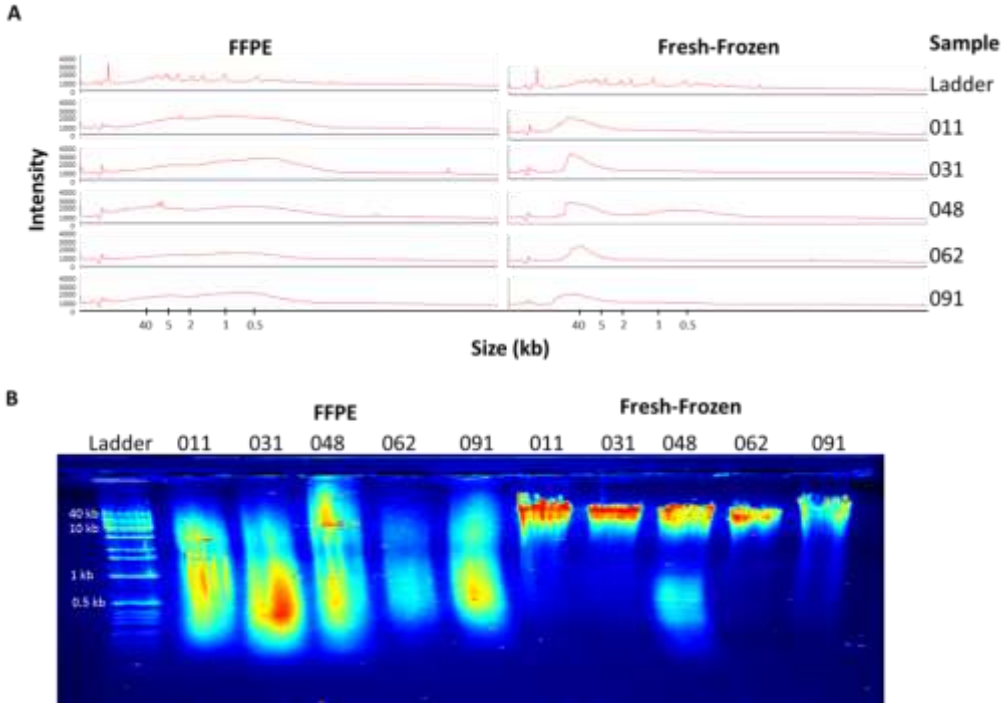


Figure 3.2 DNA integrity quality control

(A) DNA electrophoresis traces and (B) digital printout of relative intensity of DNA bands on a 3% agarose gel electrophoresis (five representative pairs of FF and FFPE samples from the patient cohort). The ladder used was the 1kb DNA Extension Ladder (Thermo Fisher Scientific, Waltham, MA, USA) (the FFPE sample 048 showed an ethidium bromide artefact at 10kb).

3.4.3 FFPE DNA presented shorter fragments and data revealed non-uniform coverage

Alignment metrics reflecting WGS performances were calculated from WGS data of the 52 patients as described in Chapter 2.8.1.1. The median insert size of the reads was 1.52 x higher in FF, the median aligned ratio of PF reads was 96% for FF and 94% for FFPE samples, and the median ratio of chimeric pairs was 0.4% for FF and 1.4% for FFPE samples (

Figure 3.3A-C). These results were significantly poorer in FFPE than FF specimens ($p=1.798e^{-09}$, $3.618e^{-09}$ and $1.330e^{-08}$, respectively, Wilcoxon signed rank test and Bonferroni corrected p-values for the total number of metrics). In addition, FFPE samples were characterised by an important increase of AT dropout (9.7x more than in FF samples) which was significantly different from FF ($p=1.804e^{-09}$), and a significantly lower CG dropout than FF ($p=1.804e^{-09}$) (Wilcoxon signed rank test and Bonferroni corrected p-values for the total number of metrics) (

Figure 3.3D and E).

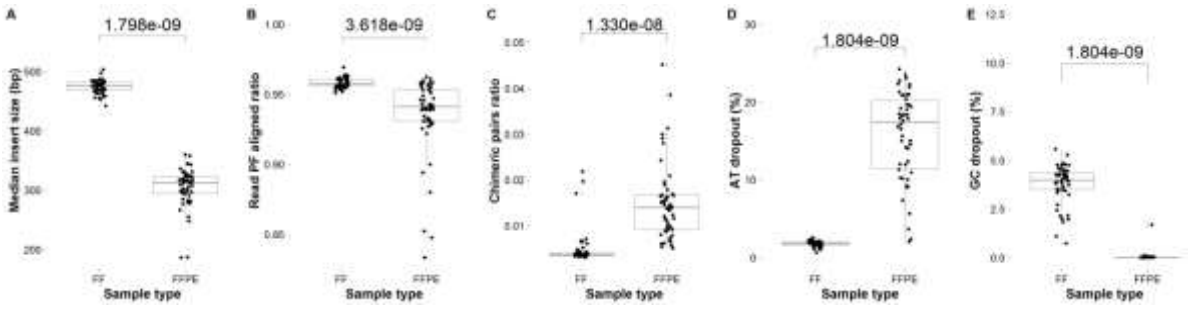


Figure 3.3 Distribution of sequencing metrics

The metrics are presented as follow: (A) median insert size in bp, (B) Read PF aligned ratio, (C) Chimeric pairs ratio, (D) proportion of AT and (E) GC dropout in percent in FF and matching FFPE samples (N=52); Wilcoxon signed rank test and Bonferroni corrected p-values.

The sequencing depth was also studied as described in Chapter 2.8.1 using CODOC. FF samples had 2.48 billion reads and a mean of 93% paired reads and FFPE presented 2.42 billion reads with 87% being mapped. FF samples had 6% PCR duplicates and FFPE samples presented 13%. The mean sequencing depth was 77x (50-100x) for FFPE samples

and 93X (78-122x) for FF samples and was variable across all chromosomes. A summary of the sequencing coverage metrics can be found in Appendix B.

The distribution of depth was calculated by measuring the proportion of bases covered at different minimum thresholds (Figure 3.4A). The depth of FFPE samples presented a lower proportion of regions reaching the targeted depth of 70X: 0.351 versus 0.782 for the FF samples. The uniformity of coverage, crucial for optimal variant detection was calculated taking the SD of read coverage in genomic windows of 100kb to measure extreme high and low signals (Figure 3.4B). A high SD describes a non-uniform coverage across the genome. The median SD was between 1.1 times and 5.8 times higher in FFPE than matching FF for the 52 patients (Wilcoxon signed rank test, $p<2.2e^{-16}$) meaning sequencing coverage was less uniform in FFPE data.

Purity determined by visual assessment of pathologists and determined from the sequencing showed a statistically significant positive correlation (Figure 3.5). The estimated purity was greater than 40% by visual assessment and was less than 40% tumour content by sequencing for 24% of cases. When studied separately, FF samples presented a significant correlation (p=0.0132, r=0.42, Spearman's rank correlation) whereas FFPE did not (p=0.371, Spearman's rank correlation). Finally, the purity of FF and matched FFPE samples was not significantly correlated (p= 0.0735, Spearman's rank correlation).

For the majority of samples, the pathologist’s visual assessment gave a greater proportion of invasive tumour compared to the purity assessed by WGS data. All samples to the left of the vertical dotted line were samples with visual estimation of tumour content greater

than 40% and computational assessment from WGS data less than 0.4 (32% of FF samples and 15% of FFPE samples).

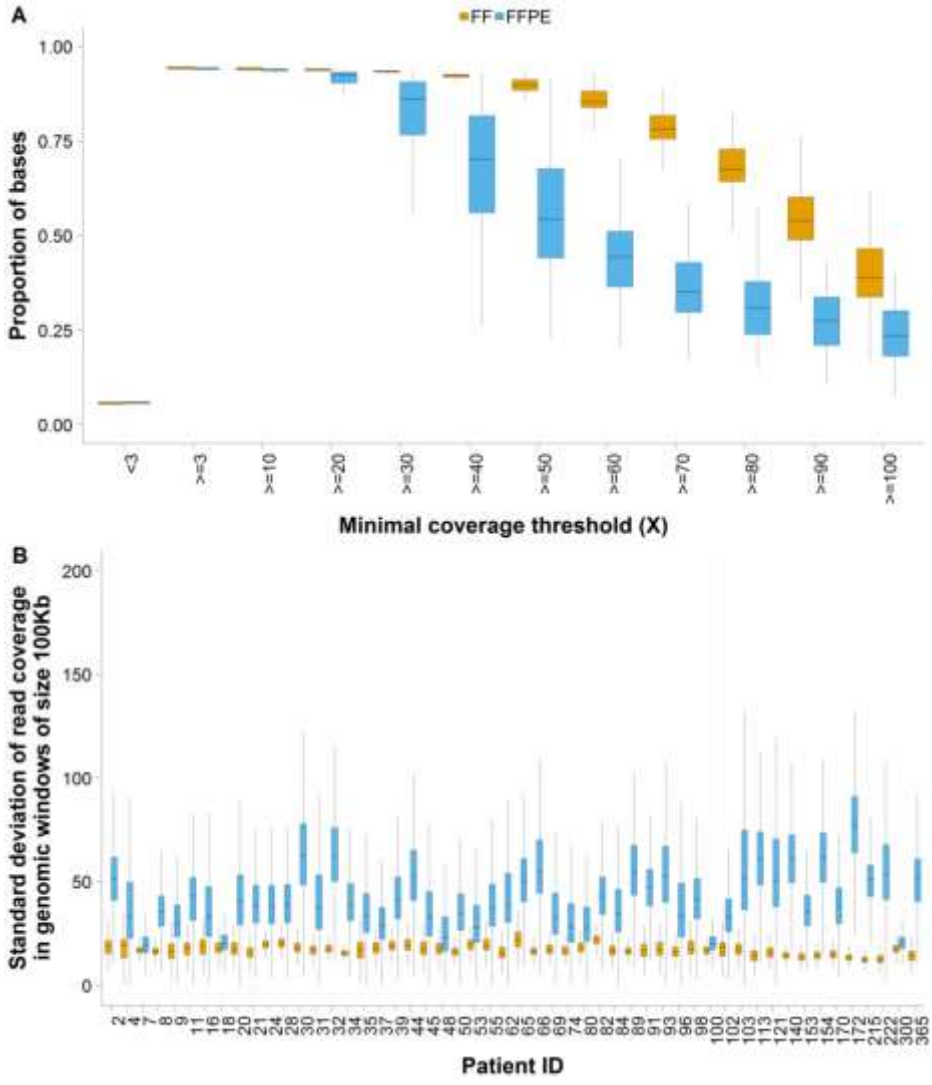


Figure 3.4 Sequencing depth statistics

(A) Distribution of sequencing depth for different minimum coverage thresholds for 52 FF and 52 matching FFPE samples and (B) distribution of standard deviations of sequencing coverage in 100kb genomic windows for FF and matching FFPE samples. FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded.

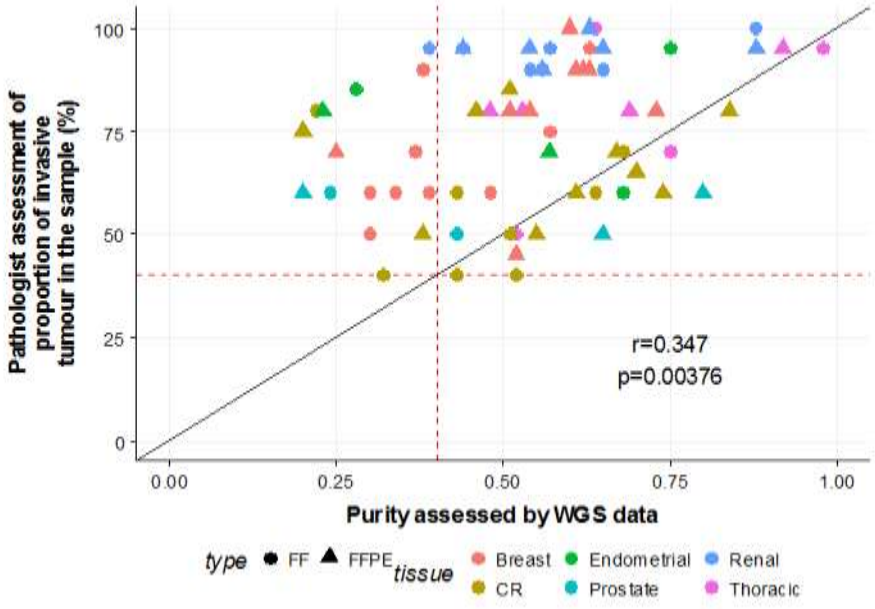


Figure 3.5 Correlation between pathologist assessment of proportion of invasive tumour in the sample and purity assessment by WGS data

Canvas was used to compute the purity from WGS data. Correlation coefficient and p values were calculated using a Spearman's rank correlation. Pathologist visual assessment (y axis) excluded all samples with a purity value lower than than 40% (red dotted horizontal line).

FF: fresh frozen, FFPE: formalin-fixed, paraffin-embedded, CR: colorectal, WGS: whole genome sequencing

3.4.4 A different number of somatic SNVs and indels could be detected in FF and FFPE samples and the agreement was dependent on the variant calling algorithm used

SNV and indel calling from FFPE WGS data was compared to the matching FF data using four major established methods, namely Strelka (Saunders et al. 2012), Mutect (Cibulskis et al. 2013), Shimmer (Hansen et al. 2013) and the combined output of these three tools

(Chapter 2.8.2.1). First, all variants reported were counted without filters applied (Figure 3.6A). Three out of four of the variant calling methods detected a higher number of somatic SNVs in FFPE compared with FF data, with 1.1 x, 1.4 x and 3.8 x additional mutations respectively. In the case of indel detection, the algorithms Shimmer and Strelka were used as Mutect does not call indels (Figure 3.6B). Both gave conflicting results: Shimmer detected 1.2 x more indels in FF than in FFPE data, whereas Strelka detected 3.3 x more indel variants in FFPE than in FF. Because the two variant callers gave such different results, the combination method did not report many variants in total.

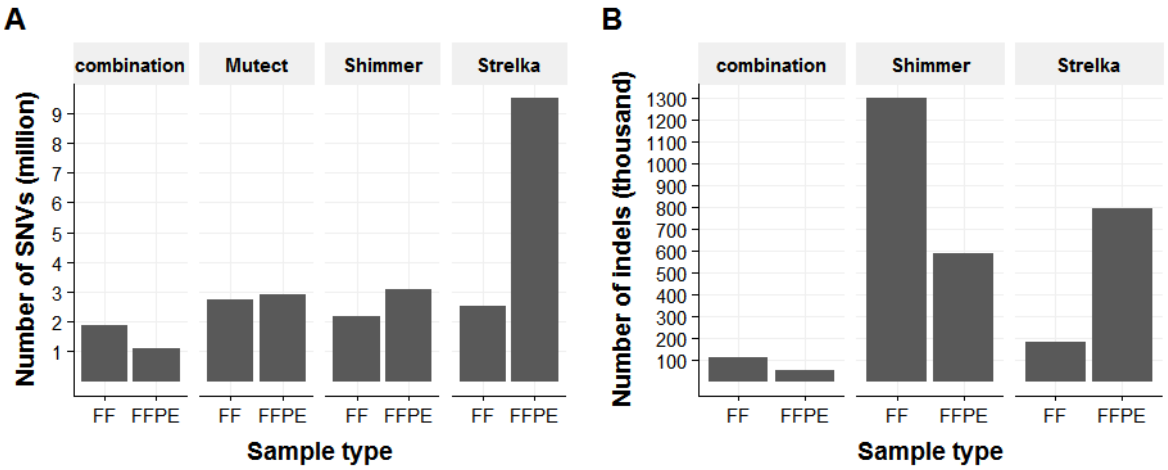


Figure 3.6 SNVs and indels detection in FF and FFPE samples

(A) Number of SNVs and (B) number of indels reported using four variant detection methods (Mutect, Shimmer, Strelka and the combination method) in FF and FFPE samples (across all samples). FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded sample.

FF and FFPE datasets were compared to identify variants detected either only in FF or only in FFPE sample data (“FF unique” and “FFPE unique”) and variants detected in both samples (“FF-FFPE overlap”) (Figure 3.7A). Mutect and Strelka discovered a similar

number of SNVs in common in FF and FFPE samples (2.08 million and 1.95 million, respectively). However, the proportion of variants detected in the FF-FFPE overlap was lower for Strelka (19%) than Mutect (58%) due to the large number of variants detected only in the FFPE samples. Shimmer and the combination method found fewer variants in the FF-FFPE overlap: 1.09 million (50%) and 0.97 million (48%) respectively.

Regarding indels, Shimmer found more FF unique indels and Strelka detected more FFPE unique indels (Figure 3.7B).

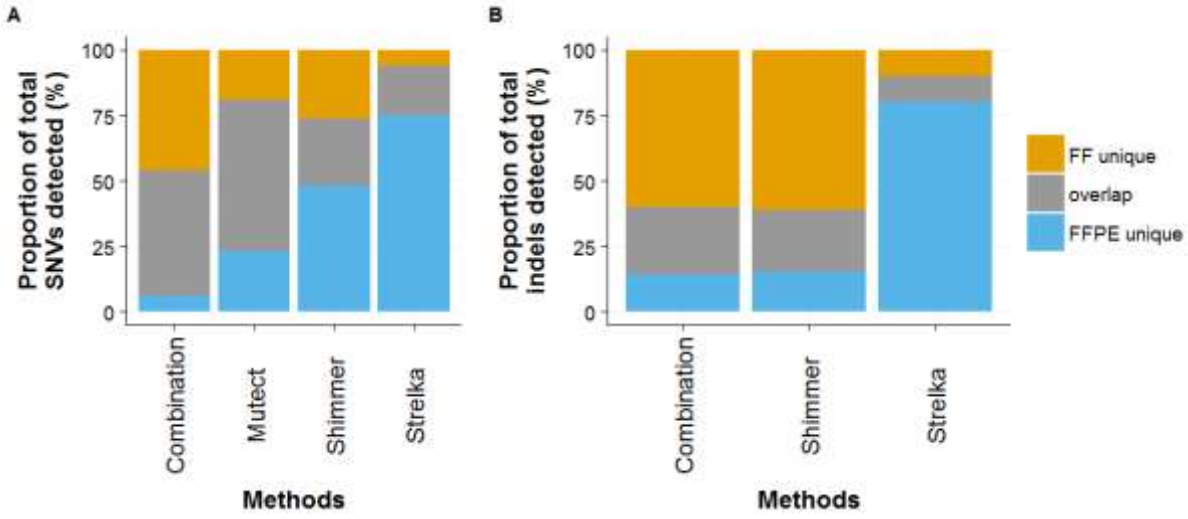


Figure 3.7 Somatic variant overlap between FF and FFPE samples

(A) SNVs and (B) indels overlap. FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded sample. SNVs: single nucleotide variants, indels: small insertion and deletions

Sensitivity and PPV were calculated using the FF sample as the gold standard following the approach described in Chapter 3.3.5 (Table 3.2). The high number of FFPE unique SNVs detected by Strelka considerably reduced the PPV (0.21) while presenting the highest sensitivity at 0.77. The best PPV of 0.89 was attained using the combined caller approach, although the sensitivity was low (0.51). In addition, similarly to the

observations in Figure 3.7, conflicting sensitivity and PPV patterns were observed for indel detection using Shimmer and Strelka.

Table 3.2 Sensitivity and PPV in SNVs and indels using four different detection methods

variant type	variant caller	sensitivity	PPV
SNV	Mutect	0.76	0.72
	Strelka	0.77	0.21
	Shimmer	0.49	0.35
	Combination	0.51	0.89
indel	Strelka	0.5	0.11
	Shimmer	0.28	0.62
	Combination	0.3	0.65

SNV: single nucleotide variant, indel: small insertion and deletion, PPV: positive predicting value

However, the number of high-risk variants (indels, start and stop codon losses and gains) was not increased in FFPE samples compared to FF (Figure 3.8).

Variants detected in different regions of the genome were also studied to identify potential regions with low SNV agreement between FF and FFPE datasets (Appendix C). The agreement in different genomic regions was explored using the Mutect dataset, as this algorithm presented the best compromise between sensitivity and PPV (Figure 3.9).

A high FF-FFPE agreement was detected in reliable regions compared to all regions (defined in Genome in a bottle, GIAB (Zook et al. 2014)) (regions of reliable calls, p-value = $6.484e^{-07}$, paired t-test) whereas the agreement was lower in regions of low complexity (which are outside of GIAB regions).

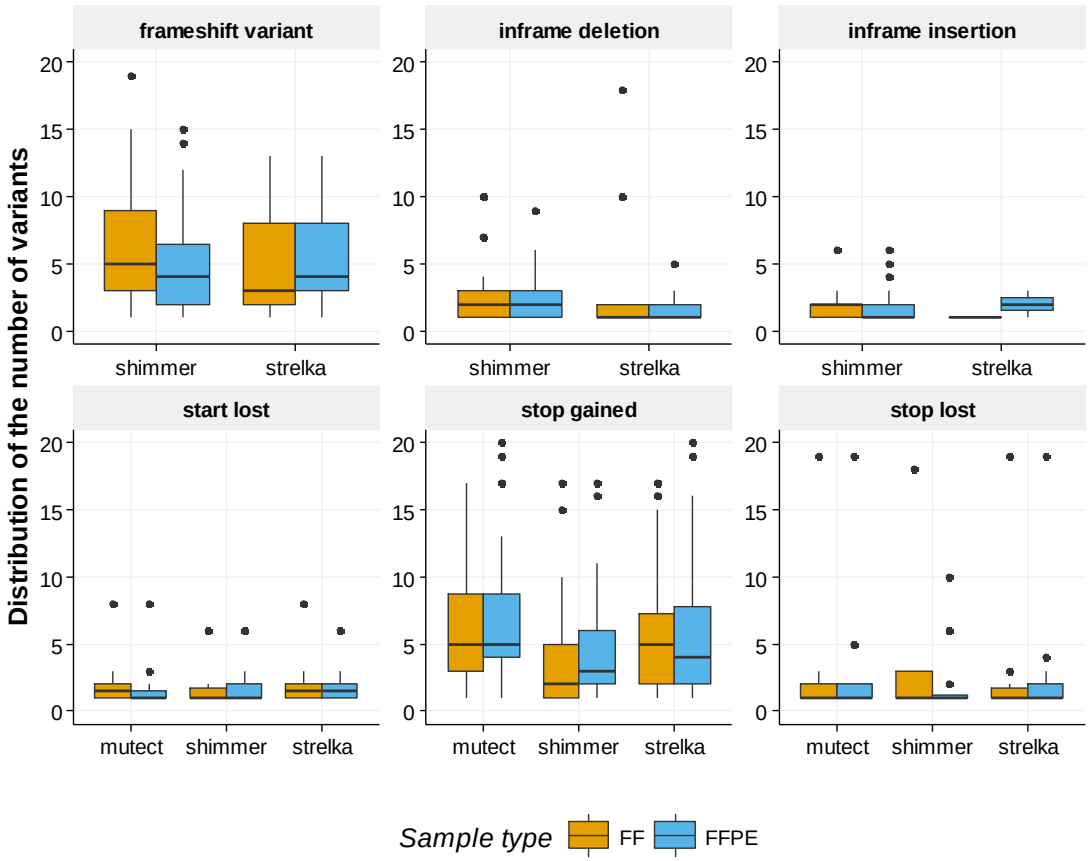


Figure 3.8 Distribution of the number of high-risk variants for FF and FFPE samples using different variant callers

High-risk variants are defined by frameshift indels, in-frame deletions, in-frame insertions, start lost, stop gained and stop lost variants. FF: fresh frozen; FFPE: formalin-fixed, paraffin-embedded

These regions outside the GIAB regions mainly represent sequences of low complexity and mappability (approximately 30% of the genome). However, there was no difference between the whole genome (all), genes (Genecode Genes V19), (p-value=0.49, paired t-test).

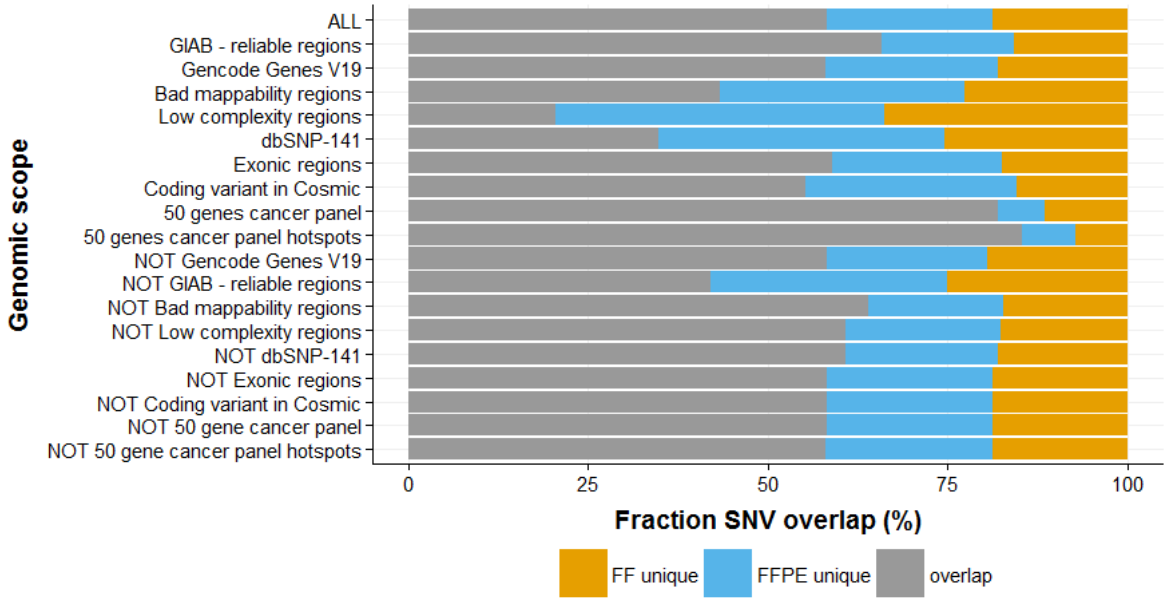


Figure 3.9 Overlap of somatic single nucleotide variants between FF and FFPE samples in different genomic regions.

The different regions are detailed in Appendix C. FF: Fresh Frozen sample, FFPE: F formalin-fixed, paraffin-embedded sample; SNV: single nucleotide variant

I compared the six tissue types included in the study. Thoracic samples presented more SNVs than other tissues (Figure 3.10) and gave the highest proportion of SNVs and indels overlap between FF and FFPE with all variant callers (Figure 3.11, Pairwise comparison in Appendix D).

For other tissue types, the method employed showed little difference (breast, pairwise comparisons using t-tests failed to show significance) or important differences (for prostate, Shimmer vs combination method: p=0.0071, and Strelka vs. combination method: p=0.0074, pairwise comparisons using t-tests with Bonferroni correction). Renal samples gave the lowest proportion of SNV and indel overlaps between FF and FFPE.

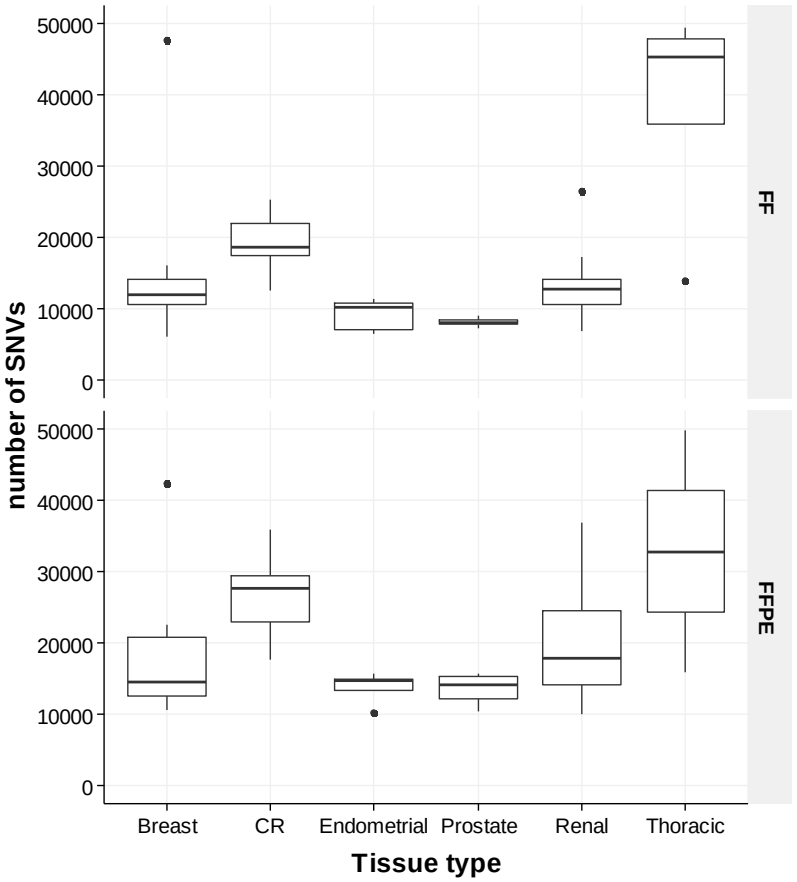


Figure 3.10 Absolute count of SNVs per sample, per tissue types for FF and FFPE samples

Case 82 was excluded from this plot as it is a hypermutated endometrial case.

CR: colorectal; FF: fresh frozen; FFPE: formalin-fixed, paraffin-embedded; SNVs: single nucleotide variants

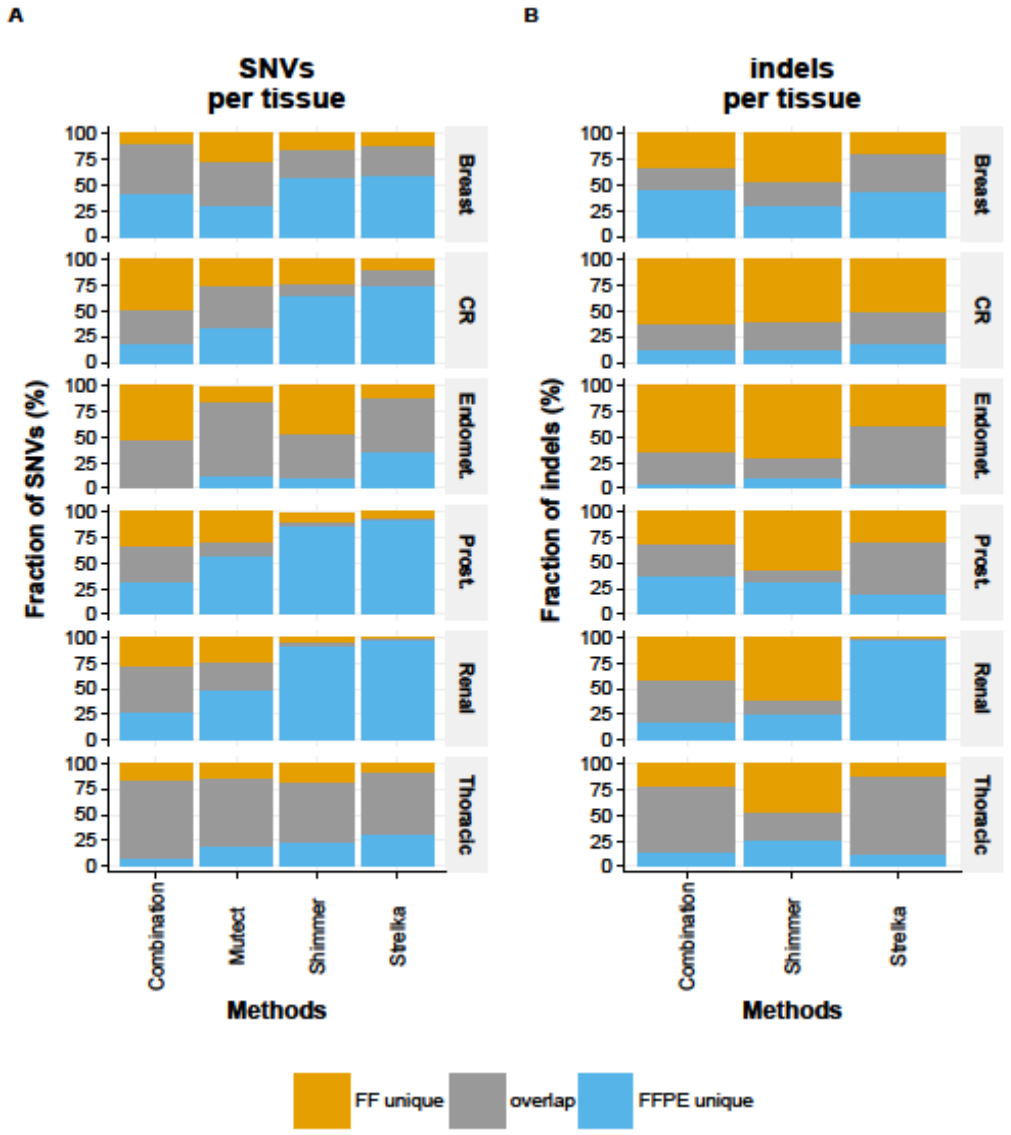


Figure 3.11 SNV comparison and indel comparison between matching FF and FFPE samples according to the sample tissue type

Measurement of overlap for somatic (A) SNVs and (B) indels represented in percent. CR: colorectal; Endomet.: endometrial; Prost.: prostate, FF: Fresh Frozen sample, FFPE: F formalin-fixed, paraffin-embedded, SNVs: single nucleotide variants

In order to explain the differences observed between FF and FFPE variants, the proportion of C>T base substitutions was calculated for all tumour samples. The proportion of C>T

base substitutions is generally known to be enriched in FFPE specimen sequencing data (Chen et al. 2014) and could elucidate the differences observed in variant detection between FF and FFPE datasets. Base substitutions were studied using the four datasets generated (Figure 3.12).

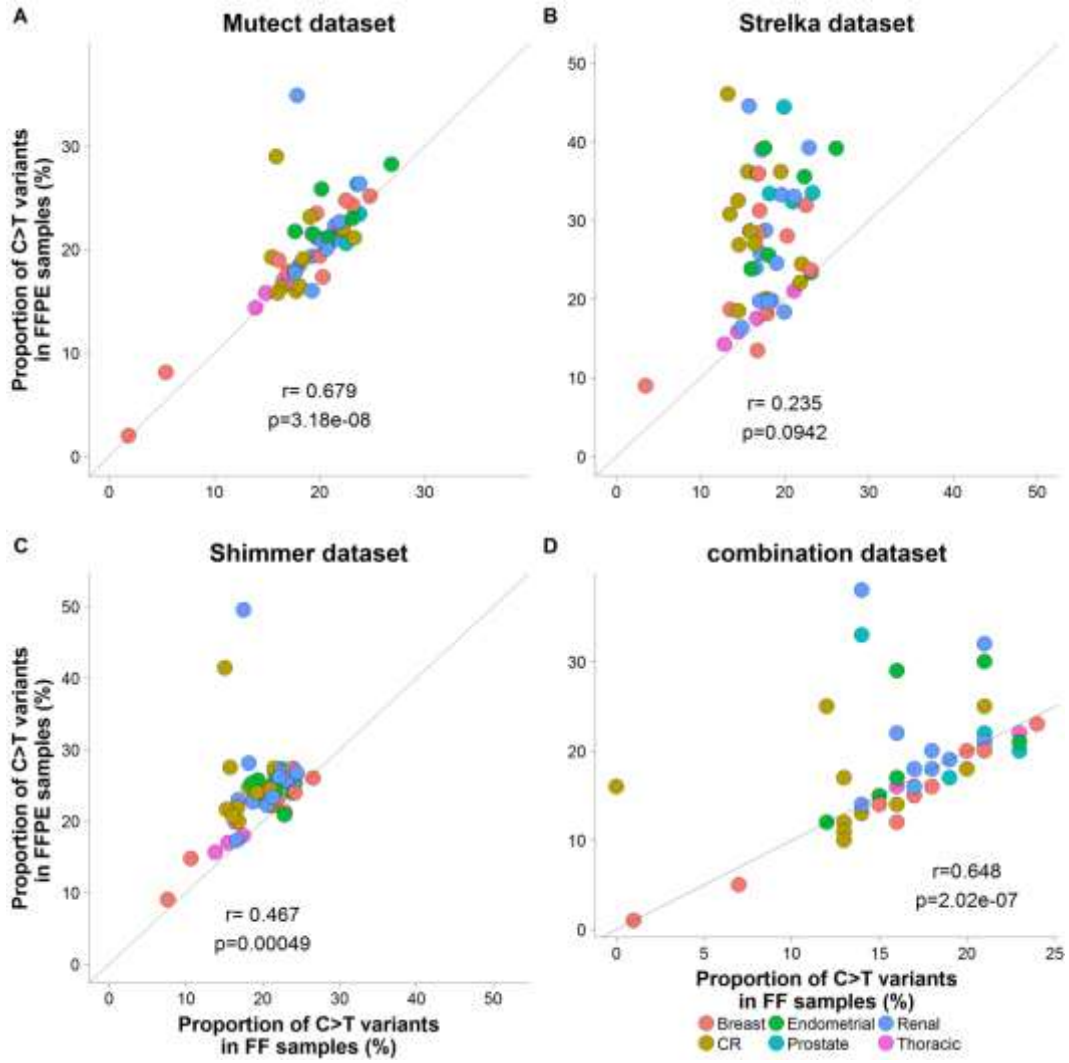


Figure 3.12 Correlation of the proportion of C>T base substitutions between FF and FFPE samples

Correlation of the proportion of C>T base substitutions (with Spearman's rank correlation) for the (A) Mutect dataset, (B) Strelka dataset, (C) Shimmer dataset and (D) combination dataset (high confidence variants). FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded sample, CR: colorectal, Endomet.: endometrial

Changes from C>T detected in FF and FFPE samples were generally correlated. However, a greater proportion of C>T variants were detected in the FFPE with respect to the FF samples (p-value=0.04 for combination method and p-value =0.02 for Mutect, paired t-test). When Strelka and Shimmer were used this was particularly evident (p-value= $1.873e^{-11}$ and $6.943e^{-07}$ respectively, paired t-test). This increase in C>T mutations observed in the Strelka dataset was correlated with the presence of more intergenic variants. Therefore, the increase of C>T variants could explain some of the differences of variants found in FF and FFPE, using Strelka for example, but not other methods such as Mutect, as the difference was so small. Further analyses are required to uncover the source of this difference in the dataset specifically obtained with Mutect.

3.4.5 Tumour heterogeneity and sampling heterogeneity explained differences between FF and FFPE samples

Somatic SNVs detected only in one sample of the FF/FFPE pair were investigated using the Mutect dataset. Fifty two percent of variants were identified in the FF-FFPE overlap, 21% were in the FF unique dataset, and 27% were in the FFPE unique dataset (Figure 3.13A).

Sixty percent of SNVs classified as "FFPE unique" had an AF ≥0.01 (median = 0.02) in the FF sample. Therefore, these might have been misclassified in the analysis due to variant calling limitations. Reciprocally, 53% of "FF unique" SNVs had an AF ≥0.01 (median = 0.02) in the FFPE sample.

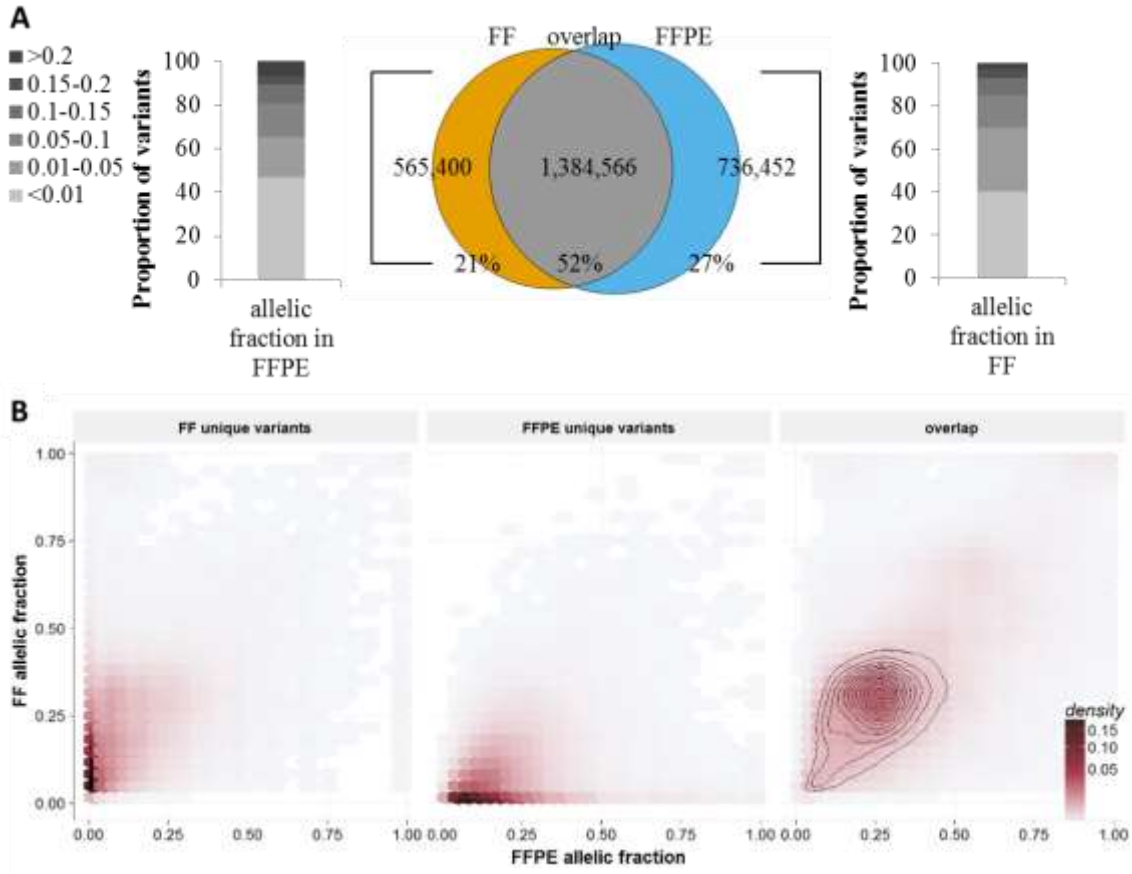


Figure 3.13 Somatic SNVs overlap between FF and FFPE samples

(A) Overlap between all FF and all FFPE variants (no filters applied). (B) AFs directly computed from the alignment data (BAM file) for all genomic positions. Allelic fraction of FF and FFPE samples at position of somatic SNV (N=2,686,418) separated in three different groups: SNVs detected in FF only (N=565,400), SNVs detected in FFPE only (N=736,452) and SNVs detected in both FF and FFPE (N=1,384,566). The shading in the left and middle panels represents variants not detected by the variant caller (VCF file), but with an allelic fraction greater than 0 in the alignment (BAM file).

The number of variants present in the sample counterpart remained substantial at higher AFs. At a threshold of 0.1 in the FF alignments, 15% of variants were classified as "FFPE unique" (but were actually also present in the FF). And at a threshold of 0.1 in the FFPE alignments, 19% of variants were misclassified "FF unique".

In addition, the FFPE unique dataset presented a median proportion of C>T substitutions of 24.2% (15.8-42.3), which was lower than in the overlap dataset (36% [17-53.9]). The

FFPE unique dataset also showed a higher C>A/G>T substitution rate. These two observations confirmed FFPE unique variants were predominantly not deamination artefacts.

The AFs of FF unique and FFPE unique variants highlighted the presence of low-level variants (Figure 3.13B). In addition, variation in the distribution of allelic fractions was used as an indication of sample purity. Some cases showed high purity of both FF and FFPE samples (Figure 3.14A) visualised by a long tail of variants in the intersection group. Other cases presented with high purity of one of the two tumour samples: FF sample for patient 113 and FFPE sample for patient 096 (Figure 3.14B-C). Finally some patients, such as patient 140 presented with low purity in both FF and FFPE samples (Figure 3.14D).

This data clearly suggested that intra-tumour heterogeneity (variants with low allelic fractions that are truly different between FF and FFPE tissue specimens) or sampling heterogeneity (rare variants present in the tissue specimens but missed by sequencing because the individual molecules with that variant were not sampled) were important explanations for the differences observed in SNV calling.

To observe the effect of tumour heterogeneity and sampling heterogeneity, I filtered out SNVs detected at low AFs using several thresholds (Figure 3.15).

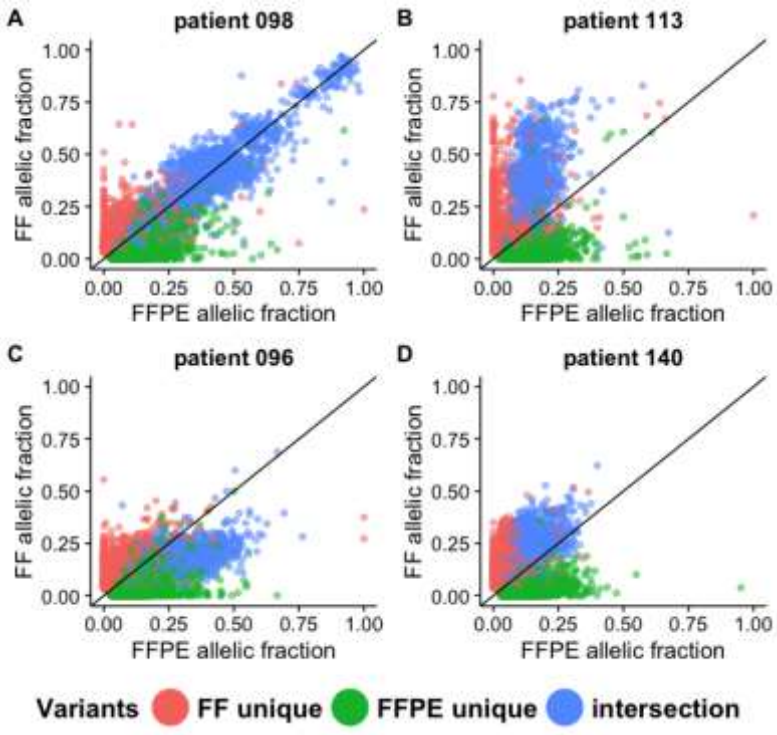


Figure 3.14 Estimation of the tumour sample purity illustrated by the distribution of the allelic fraction of FF and FFPE samples.

Representative examples of (A) high purity in both FF and FFPE samples, (B) high purity only in FF and (C) only in FFPE and (D) low purity in both FF and FFPE. FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded embedded sample

The proportion of SNV overlap between FF and FFPE increased dramatically when mutations with low AFs were excluded. The amplitude of this effect varied in the samples and different tissue types. Thoracic sample data achieved a median SNV FF-FFPE overlap of 75% when filtering out variants with AFs lower than 0.02, making this tissue type the best performing one. Other tissue types reached a median FF-FFPE overlap of 75% only if variants with AFs lower than 0.12 or 0.22 were removed, depending on tissues.

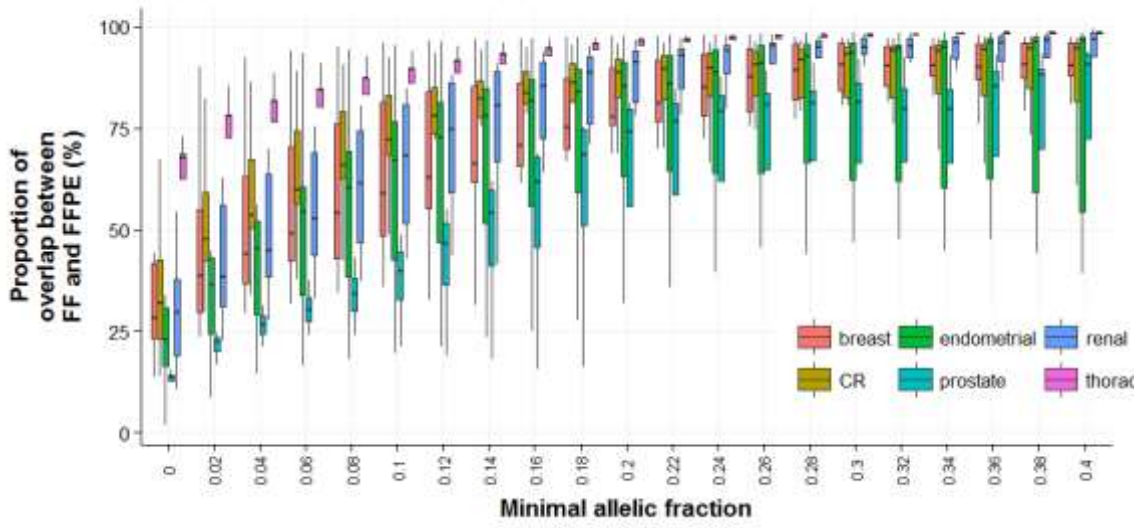


Figure 3.15 Evolution of the proportion of overlapping SNVs between FF and FFPE samples for different AF thresholds

Tissue type: renal (N=14), CR (N=12), prostate (N=4), breast (N=10), endometrial (N=7) and thoracic (N=5). FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded sample, SNVs: single nucleotide variants, CR: colorectal, thorac.: thoracic, AF: allelic fraction

In order to limit the effect of tumour heterogeneity and sampling heterogeneity, variants underpowered in matching FF or FFPE samples were filtered out of the dataset using a binomial distribution model (Figure 3.16A): the depth needed to be at least 70X at the position of the variant for both FF and FFPE samples, and the AF needed to be at least 0.067 in one of the two samples. With this approach, the SNV agreement between FF and FFPE samples increased from 52% to 63% when screening the whole genome (Figure 3.16B) and 71% when selecting variants in reliable regions (GIAB regions, representing 69% of the genome) (Figure 3.16C).

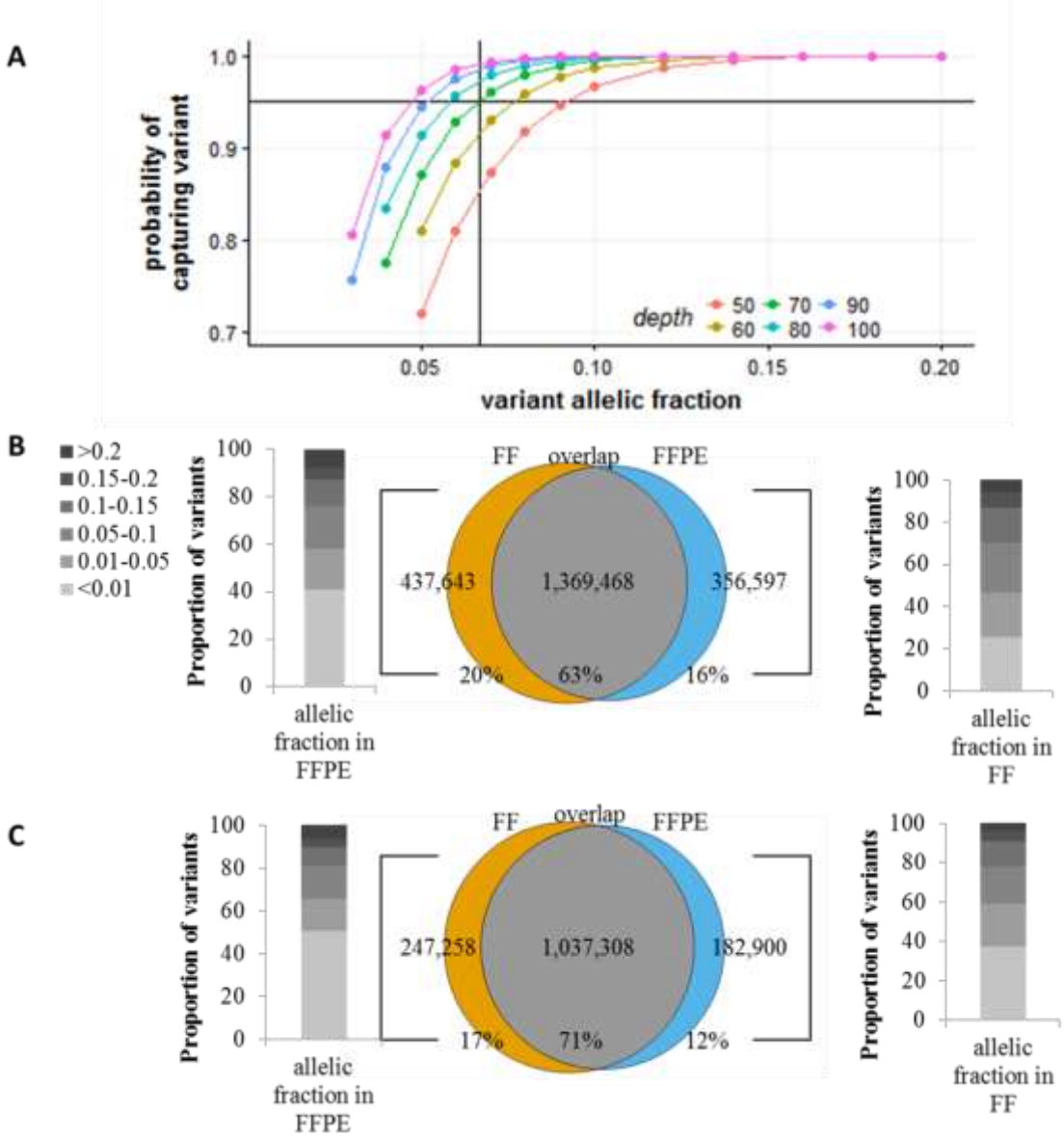


Figure 3.16 SNV overlap when underpowered variants are filtered out.

(A) Probability of observing at least 1 read with variant using binomial distribution. For a depth of 70X and the minimal allelic fraction detectable is 0.067 with a probability of 0.95. The vertical bar showed an allelic fraction of 0.067 and the horizontal bar showed a probability of 0.95. (B) The following variants were considered for this plot: variants where the depth needed to be at least 70X at the position of the variant for both FF and FFPE samples and the allelic fraction needed to be at least 0.067 in one of the two samples. (C) The following variants were considered for this plot: variants in reliable regions (GIAB regions) where the depth needed to be at least 70X at the position of the variant for both FF and FFPE samples and the allelic fraction needed to be at least 0.067 in one of the two samples.

The SNV overlap between FF and FFPE samples was found to correlate with several metrics, but particularly with the distribution of allelic fractions of somatic SNVs (r=0.731, p=1.09e^{-09}, Spearman correlation) (Figure 3.17): the higher the median allelic fractions the higher the SNV overlap between FF and FFPE samples.

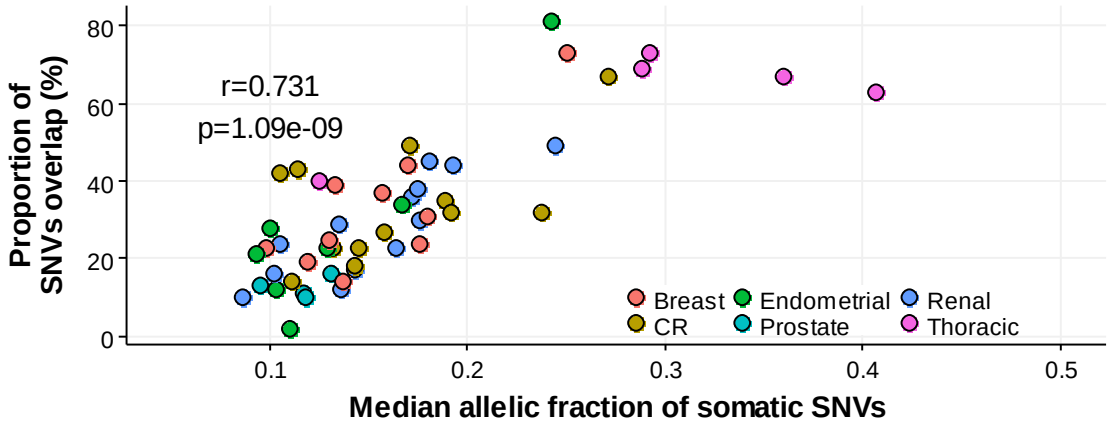


Figure 3.17 Correlation between proportion of FF-FFPE SNV overlap and median allelic fractions of FFPE somatic SNVs

The correlation coefficient and the p value were calculated using a Spearman's rank correlation.

FFPE: formalin-fixed, paraffin-embedded, SNVs, single nucleotide variants, CR: Colorectal

In this cohort, most thoracic samples presented a high distribution of allelic fractions and therefore a high SNV overlap between FF and FFPE samples (4 out of 5), although this was not confined to this tissue type (3 other cases involved) and did not correlated with tumour purity assessed by the pathologists.

3.4.6 Somatic mutations in cancer driver genes were accurately detected in FFPE datasets

So far, I have considered all variants reported by the four variant callers. To study in more detail the impact of FFPE preparation on genomic regions that are currently of clinical relevance, I selected SNVs and indels in 600 cancer-associated genes, as reported in the COSMIC cancer gene census (Futreal et al. 2004). Taking all mutation types and locations into account (coding and non-coding), 90% of genes were found mutated in more patients in the FFPE dataset than FF dataset (Figure 3.18A). Of note, genes such as *BIRC3*, *ELK4* and *KIF5B* were found to be mutated in more than 30% of FFPE datasets and were found 8 x, 6 x and 5 x more mutated than the matching FF samples, respectively. Next, we considered the most clinically relevant types of mutation in splicing regions and exons i.e. missense, stop, frameshift and in-frame indel mutations (Figure 3.18B). In this analysis, all genes were equally mutated in the FFPE and FF samples. One exception was *KMT2C* which was found to be 2.1 x more mutated in the FFPE datasets.

In addition, some known hotspot mutations discovered in FF and FFPE samples were validated using another MPS technique. SNVs and indels were validated using the Ion AmpliSeq Cancer Hotspot Panel v2 (Thermo Fisher Scientific, Waltham, MA, USA), an alternative method to the Illumina technology covering 207 targets in 46 clinically relevant genes.

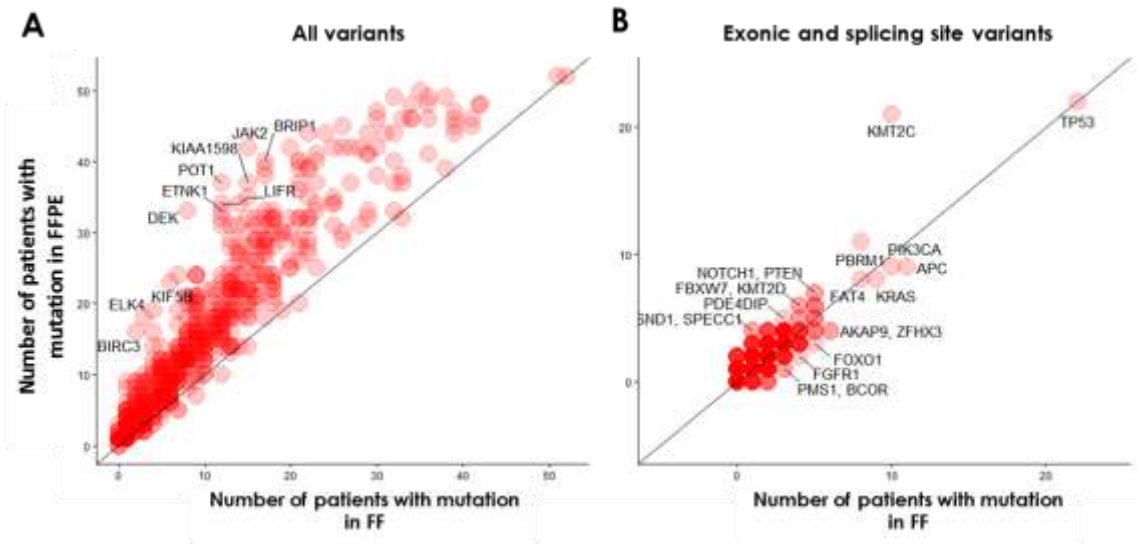


Figure 3.18 Somatic SNVs and indels comparison between FF and FFPE samples in cancer related genes.

Proportion of patients mutated in 600 genes from the COSMIC census in 52 FF samples compared with the matching 52 FFPE samples. (A) All SNV and indel types and (B) functional mutations (specific splicing regions and exons: missense, stop, frameshift and in-frame indels) (N=600).

When analysing the targeted data from FF samples, the top 5 genes mutated in this cohort were *TP53* (18 mutations), *KRAS* (9 mutations), *PIK3CA* (9 mutations), *PTEN* (7 mutations) and *SMAD4* (7 mutations) (Appendix E). Using the results from the targeted panel as a gold standard, sensitivity and PPV were calculated (Table 3.3).

Table 3.3 Validation of SNVs and indels detected by WGS using the AmpliSeq Cancer Hotspot Panel v2

	Number of variants Covered by the AmpliSeq Cancer Hotspot Panel v2	WGS Sensitivity	WGS PPV
FF	69	0.86	1
FFPE	61	0.82	1

WGS: whole genome sequencing, PPV: positive predictive value, FF: Fresh Frozen sample, FFPE: formalin-fixed, paraffin-embedded sample

The sensitivities of WGS using FF and FFPE were found to be 0.86 and 0.82, respectively and the PPVs were both 1. Out of the 10 variants missed in FF and 11 in FFPE by WGS, 8 had an AF <0.1 in the panel data. These variants were more challenging to detect by WGS as the depth is approximately 100X. In comparison the sequencing depth for the targeted panel was approximately 2000x. After inspection of WGS data in IGV (Robinson et al. 2011), a deletion in *VHL* was confirmed in FF and FFPE data as well as one *TP53* deletion in FFPE data.

3.4.7 Copy number intensity signal was wavy in FFPE dataset

Although SNVs and indels have been studied extensively, larger acquired SVs such as CNAs are also clinically relevant. Therefore, the ability to identify these from the FFPE WGS datasets was evaluated. The Log2R values showed increased intensity for genome amplifications and decreased intensity for deletions. The corresponding B-allele frequency (BAF) plots indicated LOHs. Tumour-normal pairs were compared for FF and matching FFPE samples. Data was visually inspected using the software Nexus 8 Discovery Edition (BioDiscovery, El Segundo, CA, USA) (see Chapter 2.8.2.1) (Figure 3.19). In this

representative example, patient 039 presented with a cnLOH on the p arm of chromosome 6 and a copy number loss on the q arm.

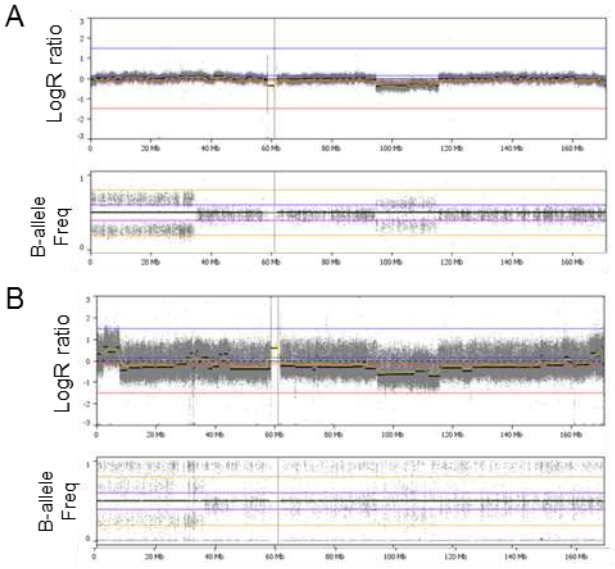


Figure 3.19 Log_2R and B-allele frequency across chromosome 6 for case 039

Plot from Nexus 8 Discovery Edition (BioDiscovery, El Segundo, CA, USA) showing CNAs for (A) the FF sample and (B) the FFPE sample

This example illustrated that the data points are highly wavy in FFPE samples, characterised by a larger spread of the Log_2R values. The signal seemed to follow a wavy pattern. The impact of this signal in FFPE samples was quantified using genome-wide Log_2R values calculated by BIC-seq on tumour-normal pairs. I examined the Spearman's rank correlation coefficients of the data in FF and matching FFPE specimens (Figure 3.20). This correlation coefficient was low in all samples.

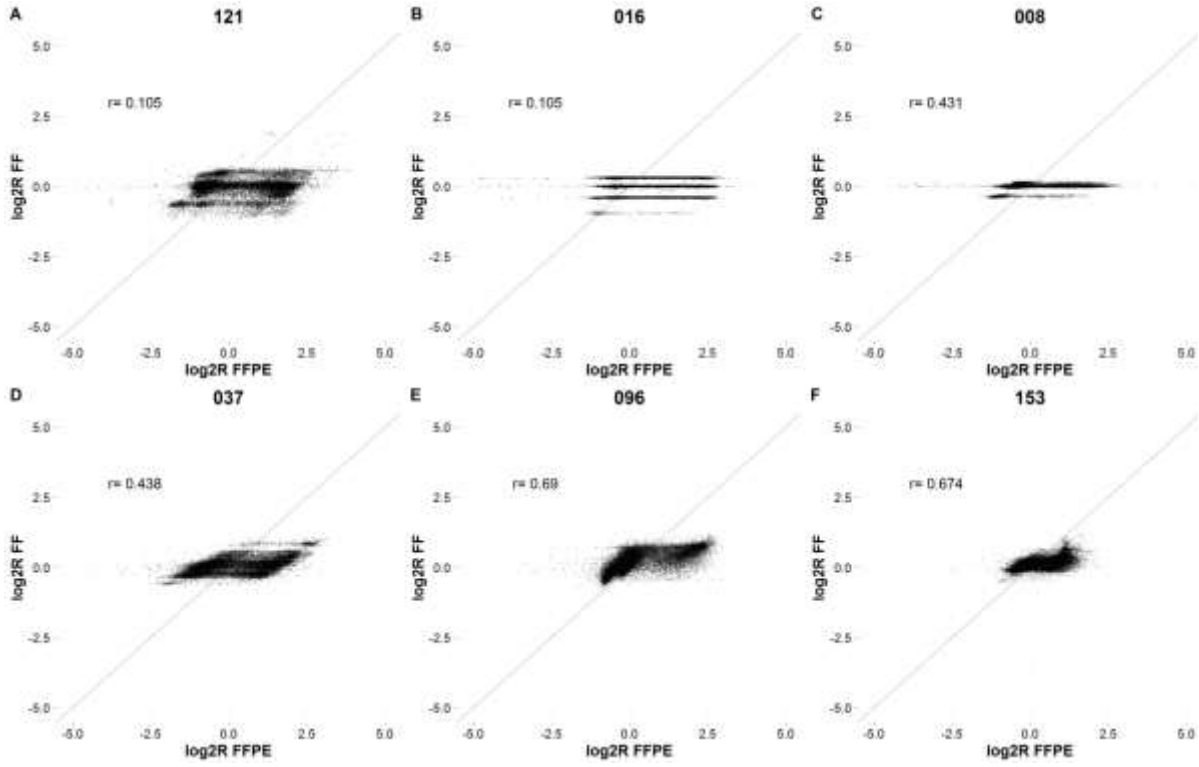


Figure 3.20 Correlation of Log_2R calculated with BIC-seq between FF and FFPE samples of six representative samples

Log_2R: Log2 Ratio, FF: fresh frozen, FFPE: formalin-fixed, paraffin-embedded, r: Spearman's rank correlation coefficients.

Indeed, the median correlation coefficient across the cohort was 0.435, and the highest correlation was achieved by case 96 (r=0.69) (Figure 3.21).

As described in Chapter 3.4.3, the depth of coverage was less uniform across the genome in FFPE than FF samples, especially in AT rich regions. This lack of uniformity explained the disrupted signal in the FFPE data.

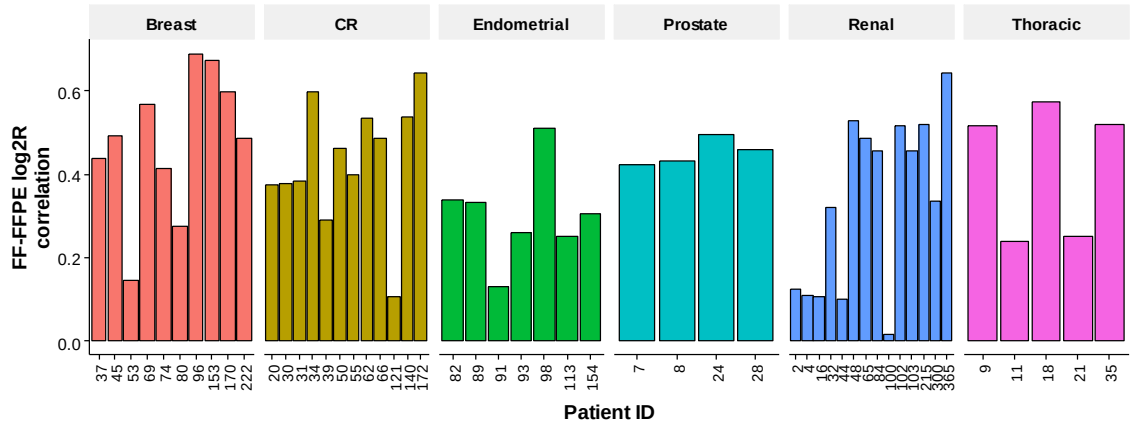


Figure 3.21 Distribution of Spearman's rank correlation coefficient of FF Log_2R and FFPE Log_2R

CR: colorectal, FF: fresh frozen, FFPE: formalin-fixed, paraffin-embedded

Therefore, when BIC-seq was employed to call CNAs, the number of variants detected was approximately 100 x higher in FFPE than FF (Figure 3.22).

Overall, this highly disrupted signal made it impossible for the algorithm to detect copy number losses and gains reliably from FFPE/GL pairs. By contrast, CNAs and regions of cnLOH were identified from FF/GL paired data using a combination of BIC-seq calls and visual inspection in Nexus.

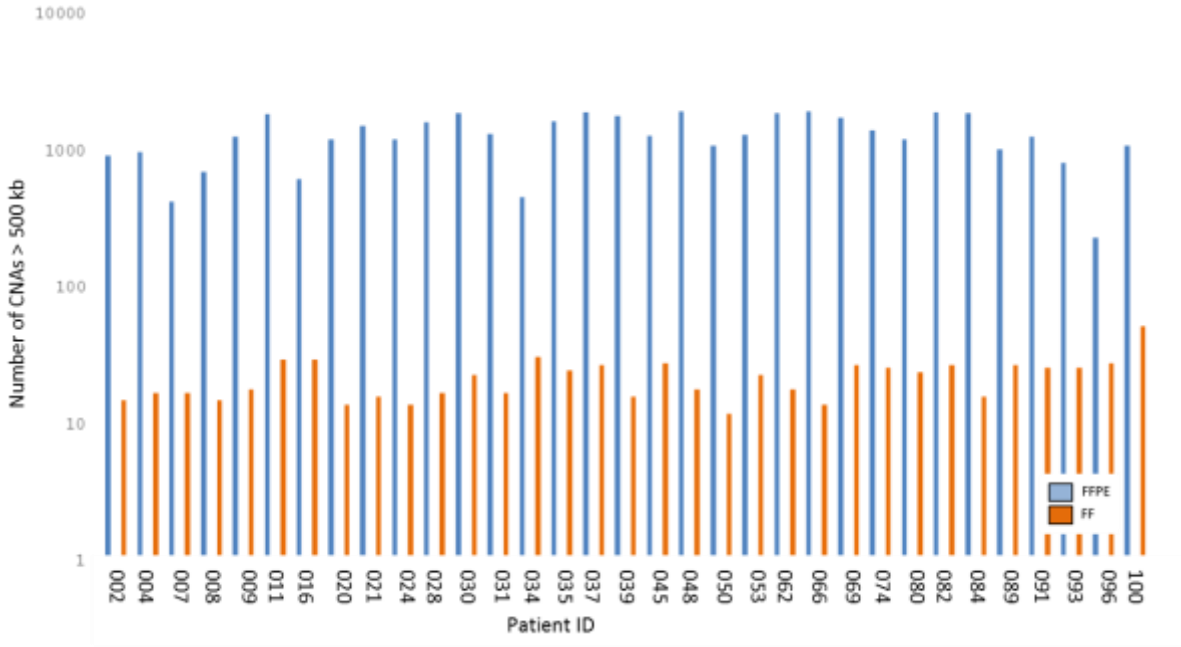


Figure 3.22 Number of somatic copy number alterations detected by BIC-seq

CNAs: copy number alterations

3.5 Limitations of the study

This study has provided many insights into the challenges of testing cancer patient samples using FFPE specimens collected in a routine clinical environment. Whilst advances have been made, the study had limitations, such as the type of control samples collected (FF and GL), limit of sequencing coverage and bioinformatics tools available to date.

The first limitation was to use data derived from FF samples as the 'gold standard'. The FFPE specimens might have included true subclonal somatic variants not present in the matching FF samples, variants that would be misclassified as false positive mutations in the cohort. However, the results with respect to FFPE and FF sensitivity and PPV were

consistent across a large number of samples. In addition, I tried to minimise this effect by collecting FF and FFPE samples from adjacent sections of tissue blocks.

Another limitation of the study was that using matched normal specimen derived from blood rather than from matched normal FF or FFPE tissues might have resulted in the production of artefact signal in the coverage subtracted data. Indeed, I observed that combining coverage data from tumour with normal data derived from blood DNA samples resulted in a fluctuation of Log_2R values.

In addition, SVs other than CNAs were not investigated due to the difficulty on calling such complex variants such as translocations or inversion on short insert size (Ahdesmäki et al. 2017). More investigations are required to investigate the potential differences between FF and FFPE samples in SV discovery.

Furthermore, given the current cost of WGS, sequencing coverage was limited and this technique cannot be as sensitive as targeted approaches that are able to detect reliably low subclonal variants (AF of less than 0.1). Indeed, in our 70-100 x tumour data sets, eight clinically relevant variants with an AF under 0.1 were not detected in either FF or FFPE tissue samples. In future years, the reduction in costs will allow for higher depth, further increasing WGS sensitivity. Improved software performance is also expected as most current bioinformatics tools were developed for data generated from FF tissues or blood samples. New statistical and bioinformatics tools that take specific features of FFPE derived cancer sequencing data into account (such as shorter fragment lengths with non-uniformed coverage) and that detect better other forms of SVs are required.

3.6 Discussion

This study answered the crucial question of whether WGS can be reliably introduced in the clinic. WGS of cancer specimens obtained from FFPE material and matching FF samples were generated and compared.

3.6.1 Prospective FF tissue collection confirmed challenging in routine clinical testing

Of the 184 cases recruited to the study, the full trio of datasets (GL, FF and FFPE) was obtained only for 52. This study demonstrated the challenges of collecting FF tissues prospectively in routine clinical testing. Out of the 48% of patients excluded because of a lack of FF sample available: 33% were due to practical reasons (weekend surgery or out of hours), and 15% were due to biological reasons (the sample did not reach the required tumour content of 40%).

Sample purity assessment by computational method generally calculated lower tumour purity than pathologist visual assessment of the specimen. Therefore, initial assessment of purity of the tissue and rigorous purity inclusion criteria are crucial to insure a sufficient detection power of subclonal variants.

For FFPE tissues, the exclusion of samples was not due to sample collection, which is well established in routine clinical practice, but instead attributable mainly to DNA extraction and library QC fail rates (37% for FFPE vs 12% for FF). However, the success rate of library preparation from FFPE DNA was higher than reported in a previous exome sequencing study (Hedegaard et al. 2014) (80% vs. 29.5%), due to fast processing from

surgery to DNA extraction (6 months or less), and adapting the amount of input DNA based on DNA QC data (ΔCq assay) to assess DNA fragment size (Table 2.1).

Whilst FF samples are the optimal source of tumour DNA for WGS, as confirmed by this work their availability is limited in the clinic. This lack of availability is one of the main barriers to the widespread application of WGS in clinical practice. Therefore, the ability to use FFPE tissues as a viable alternative source of DNA for WGS was confirmed to be of significant value clinically.

3.6.2 Confidence in somatic SNV and indel detection in coding and specific non-coding regions

Many previous studies have compared SNV concordance and indel concordance between matching FF and FFPE data without subtracting GL sample data (Holley et al. 2012; Van Allen et al. 2014; Munchel et al. 2015; Oh et al. 2015; De Paoli-Iseppi et al. 2016; Wood et al. 2010; Schweiger et al. 2009; Kerick et al. 2011; Spencer et al. 2013; Bourgon et al. 2014; Betge et al. 2015). As such, efforts to date have focused on GL variants making up the majority of mutations found in the tumour sample. By contrast, we have compared the detection of somatic SNVs and indel between FF and FFPE data. Our results have demonstrated a high discrepancy between these datasets depending on the analytical toolset employed, validating previous observations (Alioto et al. 2015; Betge et al. 2015). Different methods showed different benefits and drawbacks, indicating that tool choice may need to be guided by the specific application. However, at best, the sensitivities and PPVs were 0.77 and 0.5 for SNVs and 0.89 and 0.65 for indels, using FF samples as a

gold standard. Strelka limited the number of false negative calls. Although, the combination method limited false positive calls, it also lost any potential true positive variants only detected by one of two out of three variant callers. Mutect (SNVs only) offered the best balance of both, as it selectively avoids calling mutations in highly repetitive regions.

It has been shown previously that extended storage of specimens decreased variant overlap between FF and FFPE samples due to mutational artefacts (Hedegaard et al. 2014) like for example an increase of predicted C>T transition mutations (Oh et al. 2015; Spencer et al. 2013). With the exception of very few patients, the short storage of our samples before DNA extraction coupled with the use of DNA repair steps (described in Chapter 2.6.1.2) reduced the fraction of C>T mutations called by our pipeline. Results were comparable to matching FF samples, as observed previously by Munchel and colleagues (Munchel et al. 2015).

Consistent with previous reports, the high discrepancy between the datasets was due to the presence of low AF variants (Munchel et al. 2015) caused by intra-tumour heterogeneity and sampling heterogeneity leading to wrongly calling false negative and false positive variants in FFPE data (even though efforts were made to minimise this effect by collecting FF and FFPE samples from adjacent region of tissue blocks). However, it could not be ruled out that the false positive variants in FFPE samples were only detected due to the presence low quality reads. Kerick and colleagues suggested the fraction of discordant SNVs and indels between FF and FFPE could be reduced by thoroughly filtering out variants in low coverage regions (Kerick et al. 2011).

In addition, different tissue types presented different overlaps in term of SNVs and indels, an observation that I speculate may be associated with tissue specific variation of preparation within histopathology departments. This could also be due to the low number of subclonal variants. For example, thoracic tissues, which are generally smaller and fixed for shorter times presented better performance than other tissues. At best, after power calculation and filtering the somatic SNV agreement reached 71% with 12% of variants detected only in FFPE and 17% of variants missed in FFPE. Although this result would be insufficient in a clinical setting most of these variants are found in non-coding regions, which are not clinically relevant to date.

Our study also highlighted poor concordance between FF and FFPE somatic variants in regions of sequence complexity and regions with reduced read mappability. In the GIAB regions (representing 69% of the genome) higher concordance were observed. In addition, exonic and splicing regions in cancer related genes as well as specific clinically relevant regions were found mutated similarly between FF and FFPE tissue datasets. These observations demonstrated that WGS data generated from FFPE extracted DNA presented advantages compared to WES, restricted to coding regions.

WGS coverage is limited due to cost and this technique cannot be as sensitive as targeted approaches to detect low subclonal variants (AF of less than 0.1). Indeed, in our tumour data sets, nine clinically relevant variants with AF< 0.1 called by the Ampliseq Cancer panel were not detected in either FF or FFPE tissue samples. In future years, the reduction in costs will allow for higher depth and a further increase in WGS sensitivity.

3.6.3 FFPE preparation influenced the poor sequencing coverage signal and CNA detection

In addition to identifying SNVs and indels, CNAs are also an important component of cancer analyses. The CNA analysis of FFPE data presented variable Log_2R values attributable to non-uniform coverage depth, especially AT dropout. It was also characterised by sub-optimal alignment metrics such as percentage of reads aligned, chimeric pairs and median insert size, as described previously in multiple studies (Spencer et al. 2013; Hedegaard et al. 2014; Schweiger et al. 2009; Oh et al. 2015). However, these metrics were somewhat biased by lower stringency of fragment size selection during the alignment process for FFPE derived sequences and by the PCR step included in the FFPE library preparation method (described in Chapter 2.6.1.6). These steps were however necessary to insure sufficient material was available for sequencing. Overall, the CNA detection results presented in this study confirmed previous work (Menon et al. 2012) also performing normal subtraction on exome data but contradicted a more recent study using WES data (Van Allen et al. 2014).

3.7 Conclusions

In conclusion, this work was the first prospective WGS study of cancer patients. Datasets generated from FF and paired FFPE specimens collected in a routine clinical environment were compared comprehensively.

This work has highlighted the difficulties of collecting FF samples in routine diagnostic for cancer patients, reinforcing the need for adapted methods to use FFPE specimens for genomics investigations. When data was analysed, different somatic variant callers detected different sets of variants in FF and FFPE samples. Strelka was more appropriate to limit false negative variants, and Mutect appeared suitable to get a good balance between sensitivity and PPV. The difference between matched FF and FFPE SNVs was mainly not caused by deamination generally produced in FFPE samples, but was most likely the result of tumour heterogeneity of some samples, which are completely independent of the FFPE preparation.

Sequencing summary statistics were poorer in FFPE than FF datasets, in particular depth of coverage that was found non-uniform with lower coverage over AT-rich regions. This led to the impossibility to clearly detect CNAs in the FFPE dataset. To be able to call such SVs, it is crucial to improve the coverage of the FFPE libraries. To do so, a careful examination of the FFPE preparation is required and an optimisation of such preparation would allow CNAs investigation by WGS.

Chapter 4 Optimisation of FFPE sample quality for whole genome sequencing

4.1 Introduction

Formaldehyde prepared as formalin is the most widely used fixative in pathology because it conserves the morphology of the tissue. However, degradation of nucleic acids due to FFPE specimen processing (Bass et al. 2014) has been extensively described in cancer genetics studies (Do et al. 2013; Agell et al. 2008; Williams et al. 1999). Such degradations has been classified into four types: nucleic acid fragmentation, crosslinks, abasic sites and deamination (Do & Dobrovic 2014). These modifications of the nucleic acid molecules are likely to impede downstream molecular analyses.

Before proceeding to WGS, the quality of nucleic acids is routinely assessed in order to exclude the samples that are likely to generate poor quality data. In the context of assessing DNA from FFPE specimens, the ΔCq assay has been used as a QC method (Betge et al. 2015) and provides insight on the amount of DNA molecules that are of sufficient length (Illumina 2016) (Chapter 2.4.4.1). Subsequently, WGS data performance is assessed using sequencing alignment metrics (Wysoker et al. 2015). Such post-sequencing measurements are required to evaluate the quality of the data and confirm that

variants can be called reliably (Chapter 2.8.1). Samples exhibiting poor sequencing quality metrics require a repeat specimen to be sequenced, leading to double the cost. Therefore, pre-analytical measures of DNA samples are required to identify specimens that yield poor WGS data.

A more ideal scenario is to generate high quality FFPE samples in the first place so that pre-analytical measures are not required. The FFPE sample preparation pathway in routine diagnostics is highly complex and includes multiple variables such as time from biopsy or resection to fixation, time in fixative, tissue size and penetrance of formalin, pH of formalin, paraffin embedding conditions and DNA extraction methods (Bass et al. 2014). Previous studies have examined the effect of FFPE tissue preparation in the context of Fluorescence in situ hybridisation (FISH) or PCR based assays and suggested that the formalin fixation duration should not exceed 72 hours (Greer et al. 1991; Petersen et al. 2004; Rogers et al. 1990). However, FFPE preparation has not yet been formally investigated in the context of generating high quality WGS data in a routine diagnostic laboratory. As all samples from 0 were prepared in a routine pathology laboratory, it can be very challenging to optimise established standard operation procedure. Therefore, in the first instance, DNA extraction was optimised as it is a simpler and more accessible step of the process.

In Chapter 3, I have demonstrated that AT-rich regions were under-represented in FFPE data compared to GC-rich regions (measured as high AT dropout). As chromosomal sequences are composed of alternate AT-rich and GC-rich regions, this effect led to an unbalanced overall depth of coverage, which in turn led to the wavy patterns of Log_2R values and difficulties in calling copy number variants. Heterochromatin is more AT-rich

whereas euchromatin containing coding genes is GC-rich (Gilbert et al. 2004). In addition, heterochromatin is generally a compact structure with low flexibility. My hypothesis was that during the reverse crosslinking step of DNA extraction, heterochromatin structures that were more embedded in formaldehyde crosslinks were more fragmented than less compact DNA structures. Therefore, the optimisation of the de-crosslinking step might lead to more preserved nucleic acids. That way, the samples might be less fragmented and enriched in dsDNA.

In the work described in this chapter, the origin of FFPE poor quality sequencing data has been investigated using the 52 FFPE samples detailed in 0 and a method to optimise the quality of FFPE derived WGS data has been devised. It is anticipated that this will allow the acquisition of genome-wide data for all patients with cancer, including those for whom only FFPE material is available.

4.2 Aims and objectives

The aims of this work are to (i) predict WGS quality of FFPE specimens using DNA quality metrics, (ii) improve FFPE DNA quality by optimising the DNA extraction method, (iii) identify key steps in the FFPE tissue preparation influencing WGS quality and (iv) generate clinical reports from WGS data using FFPE DNA samples.

4.3 Material and methods

4.3.1 Patients and samples

In the first part of this study, the 52 patients from 0 were selected (Appendix A). The tissue samples were prepared as both fresh frozen and FFPE samples, as described in Chapter 2.2. To optimise the DNA extraction process for FFPE specimens, five patients were selected, representative of different FFPE DNA and sequencing qualities (Table 4.1).

Table 4.1 DNA and sequencing quality control metrics for five FFPE samples

Patient ID	4	32	65	300	365
A260/A280	1.98	1.78	1.90	1.94	1.91
ΔCq	2.75	1.60	1.49	2.47	1.19
median insert size (bp)	283	187	298	327	336
Read PF aligned ratio	0.93	0.85	0.95	0.96	0.96
Chimeric pairs ratio	0.016	0.028	0.009	0.005	0.007
AT dropout (%)	18.4	23.0	18.3	2.4	19.6
GC dropout (%)	0.0	0.0	0.0	1.7	0.0
median reads covered > 70X	$6.90e^{+08}$	$5.60e^{+08}$	$1.40e^{+09}$	$2.30e^{+09}$	$1.40e^{+09}$
median SD of depth in 100kb windows	33	62	51	21	52

PF: pass filter, SD: standard deviation, SNV: single nucleotide variant, PPV: positive predictive value

In another experiment, FFPE tissue preparation was investigated in a cohort of 203 FFPE samples collected in five UK centres during the 100,000 genomes study work between 2014 and 2015 (Appendix F).

Finally, to further explore the results from the previous analysis, experiments were conducted on new additional trio samples (FF, FFPE, GL) from 10 patients. For these samples, GL and FF samples were prepared as described previously in Chapter 2.2. The FFPE tissue preparation of the 40 samples was recorded (Table 4.2).

Table 4.2 FFPE preparation and DNA extraction conditions for experimental validation

patient ID	Tissue of origin	De-crosslinking condition	fixation condition	Fixation time (hours)	DNA QC performed	WGS performed
432	Colorectal	80°C for 1 hour	NBF	24	Yes	Yes
432	Colorectal	80°C for 1 hour	NS	24	Yes	No
432	Colorectal	65°C for 1 hour with SSC	NBF	24	Yes	No
432	Colorectal	65°C for 1 hour with SSC	NS	24	Yes	No
320	Colorectal	80°C for 1 hour	NBF	24-48	Yes	No
320	Colorectal	80°C for 1 hour	NS	24-48	Yes	No
320	Colorectal	65°C for 1 hour with SSC	NBF	24-48	Yes	No
320	Colorectal	65°C for 1 hour with SSC	NS	24-48	Yes	No
373	Breast	80°C for 1 hour	NBF	23	Yes	No
373	Breast	80°C for 1 hour	NS	23	Yes	No
373	Breast	65°C for 1 hour with SSC	NBF	23	Yes	No
373	Breast	65°C for 1 hour with SSC	NS	23	Yes	No
376	Renal	80°C for 1 hour	NBF	<12	Yes	No
376	Renal	80°C for 1 hour	NS	<12	Yes	No
376	Renal	65°C for 1 hour with SSC	NBF	<12	Yes	No
375	Renal	65°C for 1 hour with SSC	NS	<12	Yes	No
382	Endometrial	80°C for 1 hour	NBF	>72	Yes	Yes
382	Endometrial	80°C for 1 hour	NS	>72	Yes	No
382	Endometrial	65°C for 1 hour with SSC	NBF	>72	Yes	No
382	Endometrial	65°C for 1 hour with SSC	NS	>72	Yes	No
366	Endometrial	80°C for 1 hour	NBF	>72	Yes	No
366	Endometrial	80°C for 1 hour	NS	>72	Yes	No
366	Endometrial	65°C for 1 hour with SSC	NBF	>72	Yes	No
366	Endometrial	65°C for 1 hour with SSC	NS	>72	Yes	No
402	Colorectal	80°C for 1 hour	NBF	25	Yes	Yes
402	Colorectal	80°C for 1 hour	NS	N/A	Yes	Yes
402	Colorectal	65°C for 1 hour with SSC	NBF	25	Yes	Yes
402	Colorectal	65°C for 1 hour	NS	N/A	Yes	yes

		with SSC				
478	Renal	80°C for 1 hour	NBF	24	Yes	No
478	Renal	80°C for 1 hour	NS	48	Yes	No
478	Renal	65°C for 1 hour with SSC	NBF	24	Yes	No
478	Renal	65°C for 1 hour with SSC	NS	48	Yes	No
480	Renal	80°C for 1 hour	NBF	25	Yes	No
480	Renal	80°C for 1 hour	NS	49	Yes	No
480	Renal	65°C for 1 hour with SSC	NBF	25	Yes	No
480	Renal	65°C for 1 hour with SSC	NS	49	Yes	No
393	Breast	80°C for 1 hour	NBF	29	Yes	Yes
393	Breast	80°C for 1 hour	NS	29	Yes	No
393	Breast	65°C for 1 hour with SSC	NBF	29	Yes	No
393	Breast	65°C for 1 hour with SSC	NS	29	Yes	no

FFPE: formalin-fixed, paraffin-embedded, SSC: Saline-sodium citrate, NBF: neutral buffered formalin, NS: normal saline

4.3.2 DNA extraction

DNA extraction from FF tissues and blood samples were carried out as described in Chapters 2.3.1 and 2.3.3. DNA extraction from FFPE tissues was optimised by varying different parameters: reverse cross-linking temperature (90°C, 80°C, 65°C and 56°C), incubation time (1 hour and 3 hours) and buffer addition (Saline-Sodium Citrate, SSC) using two commonly used commercially available DNA extraction kits, as in Chapter 2.3.2 (Table 4.3). The buffer 20x SSC was prepared as follow: 175.3g of sodium chloride (NaCl) (VWR Chemicals, Radnor, PA, USA) and 88.2g trisodium dihydrate citrate (Sigma-Aldrich, St. Louis, MO, USA) in 1L of distilled water, pH adjusted to 7.0. Working solution was used as 2x adding 1:10 of final volume in the digested sample

(approximately 11 µl in 100 µl, but require the volume to be adjusted in case of evaporation).

Table 4.3 Conditions applied to the FFPE tissue specimens to optimise DNA extraction

Patient ID	sample type	De-crosslinking temperature (°C)	Incubation time (min)	Additional buffer	DNA extraction kit
4	FFPE	90	60	none	Qiagen
4	FFPE	80	60	none	Qiagen
4	FFPE	80	60	SSC	Qiagen
4	FFPE	65	60	none	Qiagen
4	FFPE	65	60	SSC	Qiagen
32	FFPE	80	60	none	Covaris
32	FFPE	65	60	none	Covaris
32	FFPE	65	180	none	Covaris
32	FFPE	80	60	SSC	Covaris
32	FFPE	65	60	SSC	Covaris
32	FFPE	65	180	SSC	Covaris
65	FFPE	80	60	none	Covaris
65	FFPE	65	60	none	Covaris
65	FFPE	65	180	none	Covaris
65	FFPE	80	60	SSC	Covaris
65	FFPE	65	60	SSC	Covaris
300	FFPE	80	60	SSC	Covaris
300	FFPE	65	60	none	Covaris
300	FFPE	65	60	SSC	Covaris
300	FFPE	90	60	none	Qiagen
300	FFPE	80	60	none	Qiagen
300	FFPE	80	60	SSC	Qiagen
300	FFPE	65	60	none	Qiagen
300	FFPE	65	60	SSC	Qiagen
365	FFPE	80	60	none	Covaris
365	FFPE	80	60	SSC	Covaris
365	FFPE	65	60	SSC	Covaris
365	FFPE	90	60	none	Qiagen
365	FFPE	80	60	none	Qiagen
365	FFPE	65	60	none	Qiagen
365	FFPE	65	180	none	Qiagen

365	FFPE	56	60	none	Qiagen
365	FFPE	56	180	none	Qiagen

SSC: Saline-Sodium Citrate, Covaris: truXTRAC FFPE DNA, Qiagen: QIAamp DNA FFPE Tissue kit

The QIAamp DNA FFPE Tissue kit (Qiagen, Hilden, Germany) was used with deparaffinisation solution for the paraffin removal step and incubated overnight with proteinase K, as per Chapter 2.3.2.1. TruXTRAC FFPE DNA kit (Covaris, Inc., Woburn, MA, USA) extraction was performed as previously described (Chapter 2.3.2.2). To ensure patient's tumour homogeneity of the different sample conditions extracted, all proteinase K digested aliquots were pooled together, homogenised and split across several tubes before the reverse crosslinking step.

DNA samples from the 203 FFPE tissues were extracted using different methods detailed in Chapter 4.4.3.

The experimental exploration of tissue fixation for 10 patients included FFPE samples extracted using the Covaris truXTRAC FFPE DNA using de-crosslinking conditions as detailed in Table 4.2. Additional FF and GL samples were extracted as described in Chapters 2.3.1 and 2.3.3.

4.3.3 Sequencing

Whole genome libraries were prepared from extracted DNA samples as described in Chapter 2.6.1 and sequencing was performed using a HiSeq2500 instrument, High Output v4 mode, 2x125 cycles (paired end, non-indexed) (see Chapter 2.7.1).

4.3.4 Bioinformatics

Read mapping and quality metrics were calculated as described in Chapter 2.8.1.1. Somatic SNV and indel calling was performed using Strelka 2.0.14 (Saunders et al. 2012) as detailed in Chapter 2.8.2.1. Somatic CNAs calling was performed using Nexus Discovery Edition 7.5 (BioDiscovery Inc., El Segundo, CA, USA) as described in Chapter 2.8.2.1.

4.3.5 Variant annotation and reporting

Clinical reports were generated for nine patients using data obtained from the optimised FFPE protocol and independently, using matching FF data. SNVs and indels from Strelka 2.0.14 (Saunders et al. 2012) were annotated using VEP GRCh37 release 85 (McLaren et al. 2016) and only variants in the COSMIC cancer gene census referencing 600 genes (Futreal et al. 2004) and genes of the specific cancer type in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways were kept for further analysis (Kanehisa et al. 2016). Log_2R and BAF plots were visually inspected and CNA calls manually curated using Nexus Discovery Edition 7.5 (BioDiscovery Inc., El Segundo, CA, USA) and annotated with COSMIC cancer gene census and the cancer specific KEGG pathways genes. Alterations were divided into tiers to determine if they would be clinically actionable (drugable/predictive/prognostic and/or with diagnostic/classification implications, as described in a previous study (Sukhai et al. 2016)), (Figure 4.1):

- Tier 1 included variants in actionable genes (meaning such variants detected can influence the patient's treatment, survival, or ability to gain access to a clinical

trial, as described in https://clinicaltrials.gov/ or https://www.mycancergenome.org/). These variants had to follow the same MOA as described in the literature (gain of function or loss of function of the protein affected from cancer.sanger.ac.uk/cosmic, www.cbioportal.org/ and www.cancer-genetics.org/) and in the case of a cnLOH, they needed to be associated with a mutation. The studies showing evidence of gene "actionability" were derived from the Precision Medicine Knowledgebase (PMKB) (Huang et al. 2017).

- Tier 2 included cnLOH not associated with a mutation in an actionable gene known to be affected by loss of function. This tier also included other types of mutations in actionable genes with an unknown MOA as described in the literature, that were likely to alter the protein (copy number changes, stop codon gains and losses and indels) or recurrently found mutated in TCGA (Cerami et al. 2012; Gao et al. 2013). Finally, tier 2 also included non-actionable variants if the variants presented the same MOA as described in the literature. In cases where the MOA was not known in the literature, the variant had to be likely to alter the protein (copy number changes, stop codon gains and losses and indels) or recurrently found mutated in TCGA and affecting a protein domain described in the protein families database (Pfam) (Finn et al. 2014).
- Tier 3 gathered actionable genes with variants not following the same MOA as that described in the literature. In cases where the MOA was unknown in the literature, variants were reported in tier 3 if they were missense variants or in UTRs, not recurrently found mutated in TCGA or not affecting a protein domain. Finally,

non-actionable variants, with unknown MOA in the literature, recurrently found mutated in TCGA but not affecting a protein domain, were also classified as tier 3.

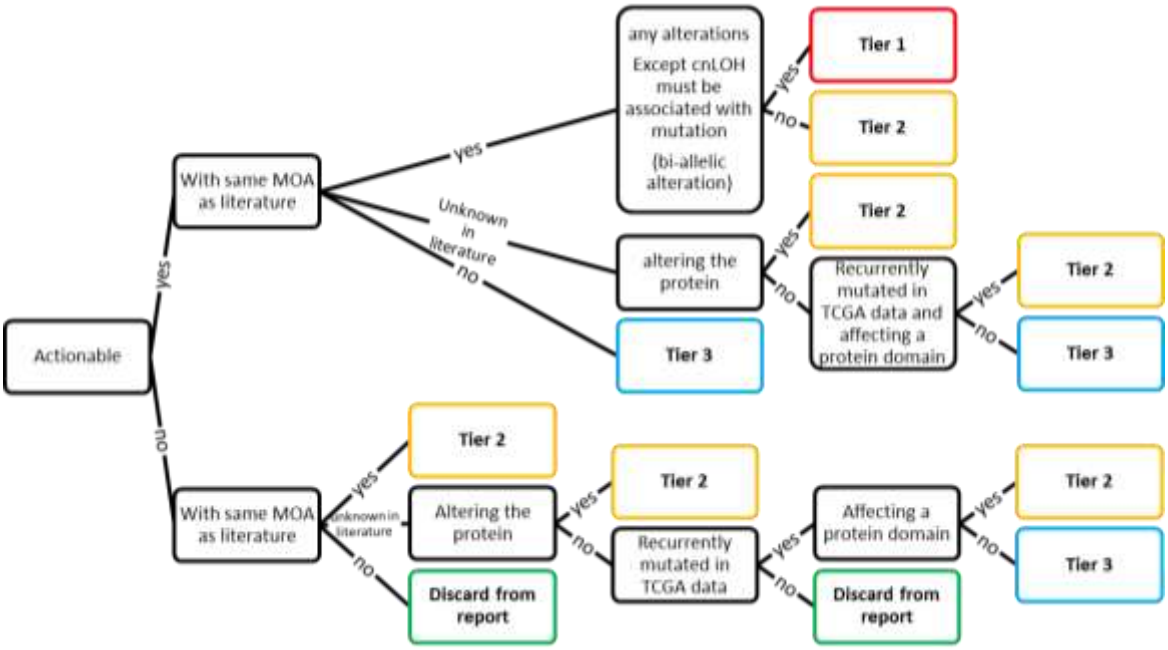


Figure 4.1 Variant classification decision process for clinical reporting

MOA: mechanism of action (gain of function or loss of function of the protein affected from cancer.sanger.ac.uk/cosmic, cbioportal.org and cancer-genetics.org/); Actionable: can influence the patient's treatment, survival or to gain access to a clinical trial, as described in clinicaltrial.gov or mycancergenome.org; cnLOH: copy neutral loss of heterozygosity, altering the protein: copy number changes, stop codon gains and losses and indels. TCGA: The cancer genome atlas

4.3.6 Statistical analysis

Continuous data comparisons were performed using Wilcoxon signed rank tests for asymmetrical distributions of paired data and Mann-Whitney U tests for two independent groups. Correlation matrices and linear relationships were calculated with Spearman correlation coefficients. Multiple testing adjustments of p-values were performed using a

Bonferroni correction. The correlation between FFPE tissue preparation and both DNA and WGS quality metrics was examined using a linear regression on Blom transformed data. In this analysis each preparation step was evaluated against each quality metric individually. In a second analysis, all variables were analysed together in order to identify the most predictive ones using another regression model, the LASSO (Least Absolute Shrinkage and Selection Operator) technique (Tibshirani 2011). Coefficients representing the effect size of the preparation step on the sequencing metric were calculated using the *glmnet* package in R (Friedman et al. 2010). The coefficients close to 0 (between -0.01 and 0.01) were excluded for further analysis.

For all statistical analysis, the null hypothesis was rejected when p-value was less than 0.05.

4.3.7 Individual contributions

In this chapter, in addition to the contributions from chapter 3, the statistical analyses were performed in collaboration and under the supervision of Dimitris Vavoulis.

4.4 Results

4.4.1 No FFPE QC metrics predicted sequencing quality outcome

Some FFPE samples presented a higher proportion of C>T variants, showed a poor overlap of SNVs and indels with matching FF samples, or revealed a non-uniform depth of coverage resulting in suboptimal CNA detection. To determine whether the FFPE

specimens were poorly performing in one of these categories only or whether all categories were affected concurrently, the FFPE data were investigated for correlations (Figure 4.2A). No correlation was identified between alignment metrics and SNV and indel detection sensitivity or PPV. However, one of the strongest positive correlations was between the coverage uniformity (median SD of the coverage) and AT dropout ($r=0.86$, $p=5.86e^{-16}$, Spearman correlation, Bonferroni correction). One of the strongest negative correlations was between the percentage of C>T variants and sensitivity of SNV calling from the Strelka dataset ($r= -0.86$, $p=6.03e^{-15}$ (Spearman correlation, Bonferroni correction).

An accurate pre-sequencing quality prediction to identify poorly performing FFPE samples prior WGS would significantly decrease costs by placing WGS on hold until a DNA sample of suitable quality might be obtained. DNA QC metrics were investigated to identify potential predictors of poor SNV detection, indel detection or sequencing metrics by performing a correlation analysis (Figure 4.2B). With the exception of ΔCq and median insert size ($r=-0.44$, $p=9.85e^{-05}$, Spearman correlation, Bonferroni correction), none of the DNA metrics were predictive of poor sequencing outcomes.

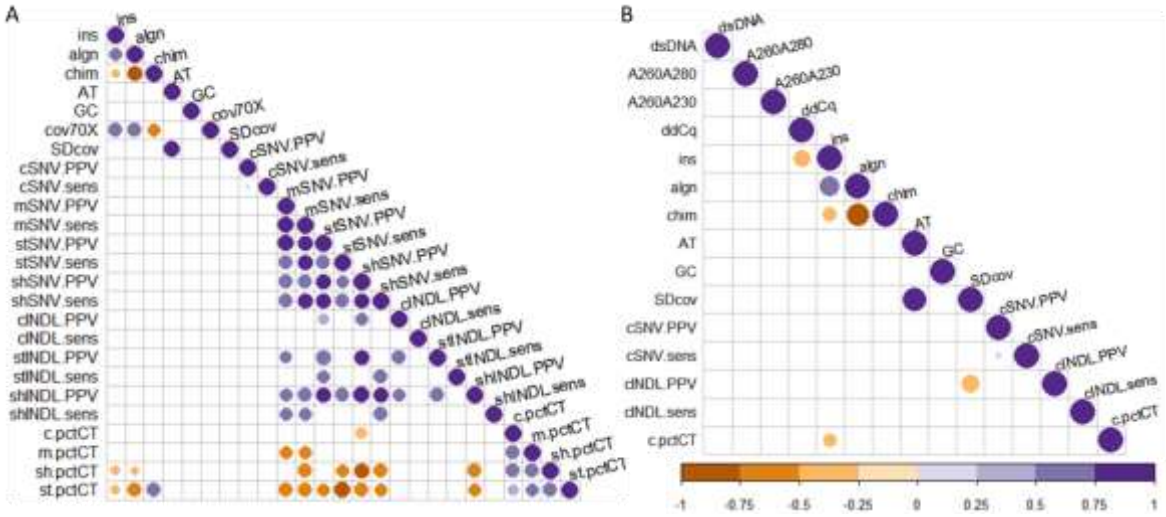


Figure 4.2 Spearman correlation matrixes between sequencing metrics and DNA quality control.

(A) The parameters analysed were the following: sequencing metrics (alignment metrics, PPV and sensitivity of SNV and indel calling measured by the four methods and percentage C>T variants measured by the four methods) and (B) DNA quality control and a selected list of sequencing metrics.

Empty cells represent non-significant correlations. The diameter of the circle represents the significance levels from corrected p-values (Bonferroni correction).

dsDNA: double stranded DNA estimation, A260A280: ratio A260/A280, A260A230: ratio A260/A230, ddCq: ΔCq, ins: median insert size, algn: Read PF aligned ratio, chim: Chimeric pairs ratio, AT: AT dropout, GC: GC dropout, cov70x: number of bases covered by at least 70X, SDcov: median standard deviation of the coverage, c: combination method, m: Mutect, st: Strelka, sh: Shimmer, PPV: positive predictive value, sens: sensitivity, pctCT: percentage C>T variants.

4.4.2 Optimisation of FFPE DNA extraction improved somatic variant detection

As it was not possible to predict poor sequencing quality using DNA metrics I wanted to optimise DNA integrity of FFPE samples to improve coverage uniformity and therefore CNA detection. The hypothesis was that modifying the reverse crosslinking step would yield less degraded nucleic acids. To test this hypothesis, five representative FFPE samples were selected from those presented in 0 (Table 4.1).

First, the detection of SNVs in optimised FFPE samples was evaluated. The proportion of somatic variant call overlap between (i) FF and initial FFPE datasets and (ii) FF and experimental FFPE datasets were compared (Figure 4.3). This comparison showed that the variant overlap was higher for the FF vs experimental FFPE datasets. One exception was the condition [Qiagen 80°C/1 hour/no buffer addition] for patient 004, where the FFPE indel overlap with the FF dataset was 74% for the FF vs experimental FFPE datasets compared with 76% for the FF vs initial FFPE datasets.

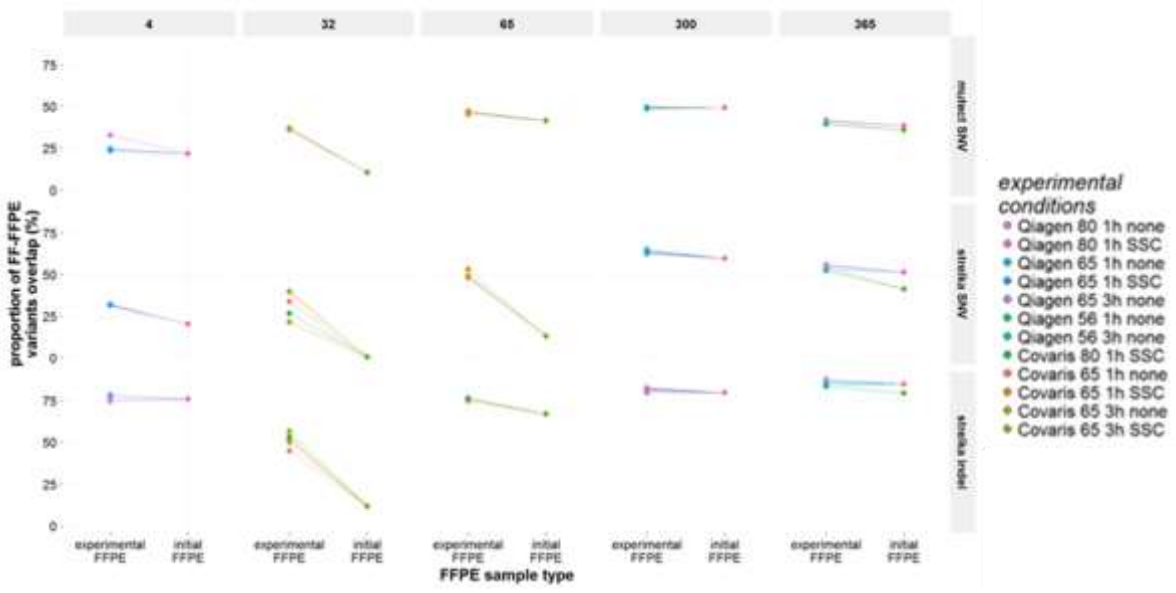


Figure 4.3 Evolution of the proportion of variant overlap between matching initial and experimental FFPE samples

Each patient is represented in a column and different DNA extraction conditions are showed by different coloured dots. The experimental conditions are described in the following order: DNA extraction kit (Qiagen=QIAamp DNA FFPE Tissue, Covaris=truXTRAC FFPE DNA Kit), reverse crosslinks step temperature and incubation time, addition of buffer (SSC = Saline-Sodium Citrate), Initial = FFPE DNA extracted according to the manufacturer's instructions, experimental = FFPE DNA extracted following a different reverse crosslinking step.

Then, comparisons of sequencing alignment metrics were made to estimate the potential improvement of alignment performance and coverage uniformity. Overall, alignment metrics such as Read PF aligned ratio, chimeric pairs ratio, AT and GC dropout were equivalent or improved in the experimental FFPE datasets compared to the initial FFPE datasets (Figure 4.4).

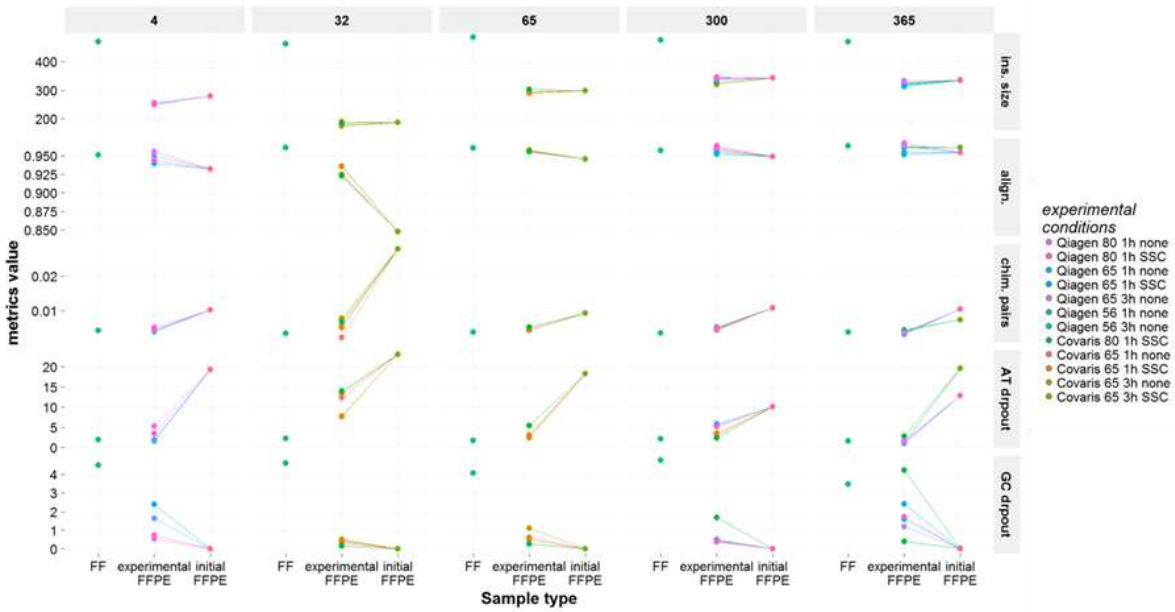


Figure 4.4 Comparison of five alignment metrics in five patients between FF samples experimental FFPE and initial FFPE samples

Each patient is represented in a column and different DNA extraction conditions are showed by different coloured dots. The experimental conditions are described in the following order: DNA extraction kit (Qiagen=QIAamp DNA FFPE Tissue, Covaris=truXTRAC FFPE DNA Kit), reverse crosslinks step temperature and incubation time, addition of buffer (SSC = Saline-Sodium Citrate), Initial = FFPE DNA extracted according to the manufacturer's instructions, experimental = FFPE DNA extracted following a different reverse crosslinking step. Ins. Size = median insert size (bp), align. = Read PF aligned ratio, chim. pairs = Chimeric pairs ratio. AT and GC dropout are expressed in percent.

Global sequencing coverage was improved in all experimental FFPE compared to initial FFPE datasets (Figure 4.5A). Four out of five patients had at least two experimental FFPE

samples with at least 50% of 100kb genomic windows showed a median coverage ≥ 70 x. For two patients, this value was increased to 75% of the windows for at least one condition. Furthermore, experimental FFPE datasets from another patient gave better results for this coverage metric than the FF samples. Coverage SD were also clearly improved by the altered reverse-crosslinking conditions (Figure 4.5B).

The experimental FFPE specimens presented a distribution of SD of sequencing coverage lower than the matching initial FFPE samples. For four out of five patients the SD measurements were equal to or lower than the matching FF specimens.

Finally, CNA detection was assessed, firstly by visualization using Nexus as described in Chapter 2.8.2.1. An improvement of the uniformity of the Log_2R values was observed for all five patients. For example, in patient 300 small deletions of 600kb and 800kb can be detected for chromosome 1 using the optimised FFPE samples (Figure 4.6).

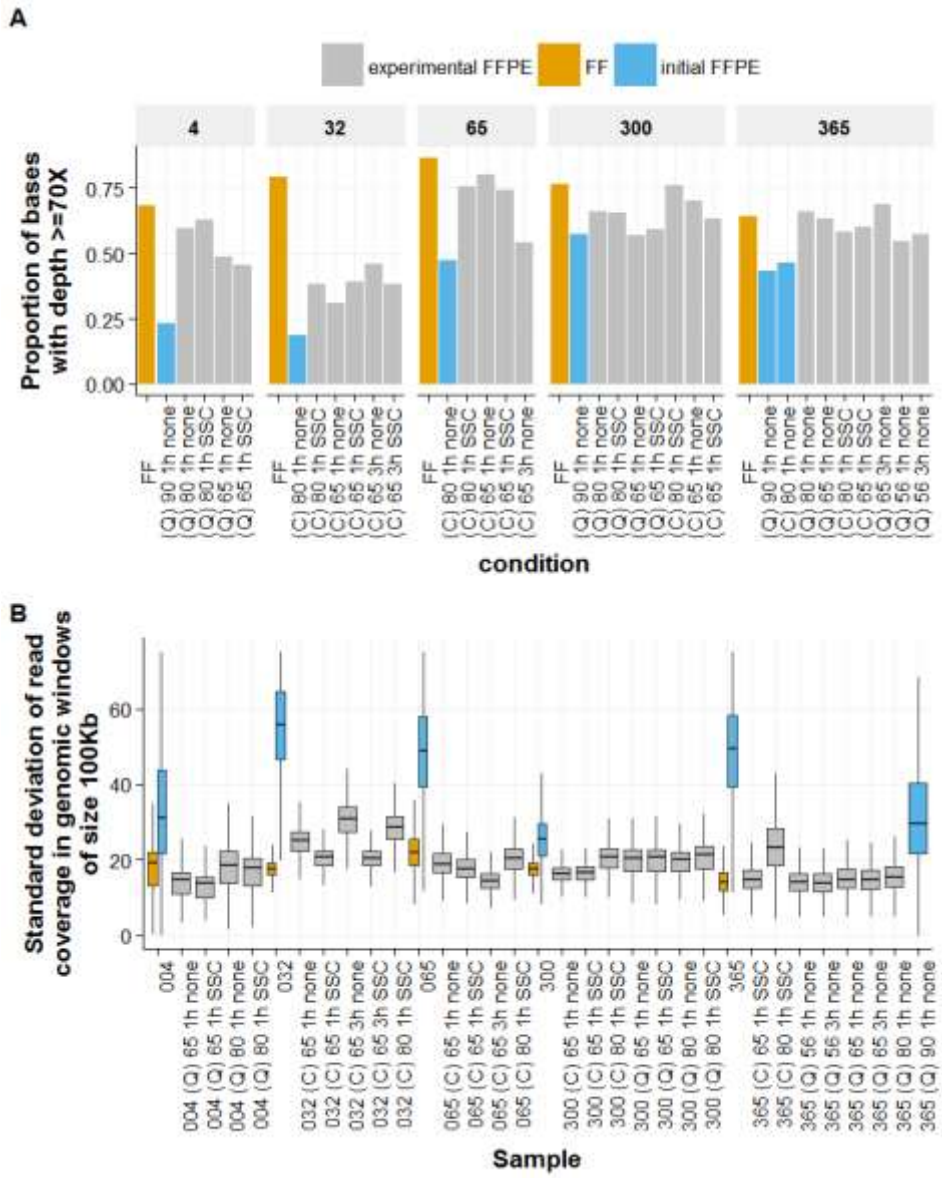


Figure 4.5 FFPE DNA extraction optimisation effect depth of coverage

(A) Proportion of 100kb windows across the whole genome with a median coverage ≥ 70 x for FF and (B) distribution of standard deviations of sequencing coverage in 100kb windows for initial FFPE and experimental FFPE samples for five patients. The experimental conditions are described in the following order: DNA extraction kit (Qiagen=QIAamp DNA FFPE Tissue, Covaris=truXTRAC FFPE DNA Kit), reverse crosslinks step temperature and incubation time, addition of buffer (SSC= Saline-Sodium Citrate), Initial= FFPE DNA extracted according to the manufacturer's instructions, experimental= FFPE DNA extracted following a different reverse crosslinking step. Covaris kit = (C) and Qiagen kit (Q).

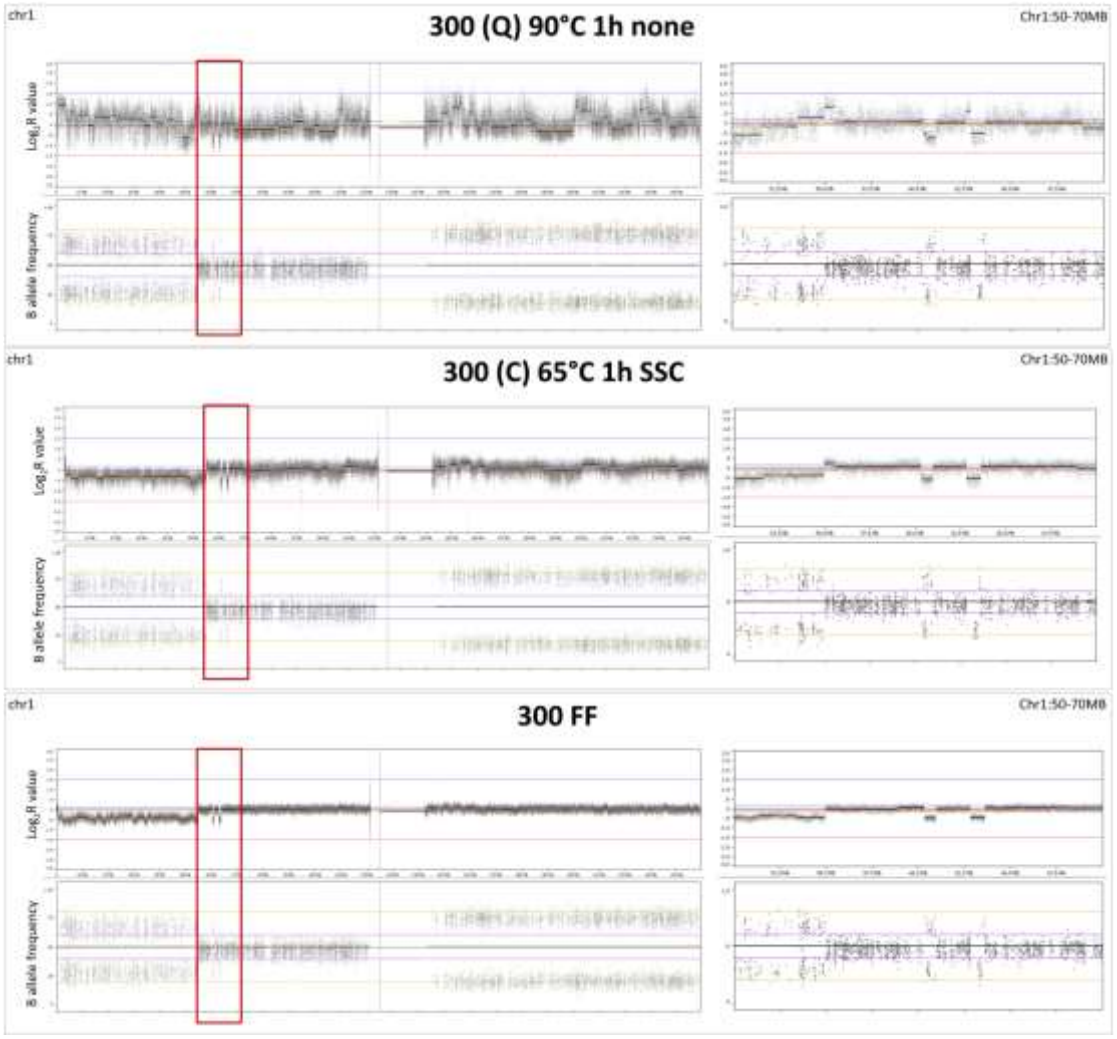


Figure 4.6 CNA detection improvement using different DNA extraction conditions from the same FFPE sample

Only a representative example of patient 300 is presented. The left panels show the Log_2R values and B allele frequencies plotted for chromosome 1, the right panels show a zoomed in view of the region in the red box (chr1:50000000-70000000) presenting two small deletions on chromosome 1. 300 (QC) 90°C 1h none = FFPE sample from patient 300 extracted with Qiagen kit (de-crosslinking conditions 90C for 1h without any additional buffer). 300 (C) 65°C 1h SSC = FFPE sample from patient 300 extracted with Covaris kit (de-crosslinking conditions 65°C for 1h with additional SSC buffer). 300 FF = FF sample from patient 300. SSC= Saline-Sodium Citrate

CNA detection was also assessed by calculating the correlation coefficient of log_2R between FF and experimental FFPE samples. Correlation coefficients were greatly

improved for all experimental FFPE samples and were above 0.85 for three out of five patients (patients 004, 065 and 365) (Figure 4.7, Appendix G). These results confirmed that WGS data from poor quality FFPE samples can be improved to enable more accurate detection of small and large variants.

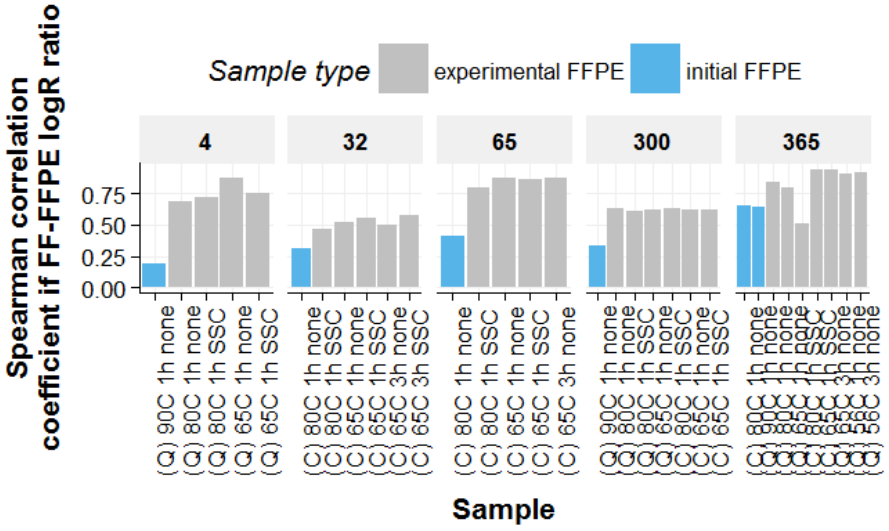


Figure 4.7 Distribution of spearman correlation coefficient of FF Log_2R and FFPE Log_2R

The experimental conditions are described in the following order: DNA extraction kit (Qiagen=QIAamp DNA FFPE Tissue, Covaris=truXTRAC FFPE DNA Kit), reverse crosslinks step temperature and incubation time, addition of buffer (SSC= Saline-Sodium Citrate), Initial= FFPE DNA extracted according to the manufacturer's instructions, experimental= FFPE DNA extracted following a different reverse crosslinking step. Covaris kit = (C) and Qiagen kit (Q)

4.4.3 Investigation of tissue handling effects on sequencing metrics highlighted key preparation steps

Going forward, the standard operation procedures of the FFPE sample pathway will require a transformation to be adapted to genomics technologies. The multiple steps of tissue preparation were studied to identify those impacting on the WGS data quality in a cohort of 203 FFPE samples collected in five centres in the UK. First, when comparing the tissue preparation between the five UK centres, a high variability of the protocol was observed (Appendix F). For instance, the size of the tissue block, the processing protocol, the fixation time, the fixative type, the processing time, the time in hot wax, the tumour sample cellularity and the sample type varied between different laboratories. This observation demonstrated the FFPE preparation procedure is not standardised across routine diagnostics laboratories. The DNA extraction procedure was less variable between different laboratories. Indeed, most centres used commercially available DNA extraction kits such as truXTRAC FFPE DNA and QIAamp DNA FFPE Tissue kits.

Then, several observations were made from the linear regression analysis comparing FFPE preparation protocols and DNA QC (Figure 4.8). A significant relationship was found between low A260/A280 values and the reverse crosslinking temperature of 65°C (linear regression model, $p= 2.24e^{-06}$) as well as high A260/A280 values and the reverse crosslinking temperature of 90°C ($p= 5.56e^{-03}$). In addition, when studying the ΔCq metrics, high values were found to be associated with the fixation time ">48 hours" ($p=2.46e^{-02}$), the sample type "scroll" ($p= 3.80e^{-02}$), the time in hot wax "490 or 540

minutes" ($p= 1.76e^{-02}$) and tissue block size "mega-block" ($p=1.44e^{-02}$). Conversely, low values of ΔCq were associated with the sample type "section mounted" ($p= 3.11e^{-02}$).

Next, the sequencing metrics (percent of reads aligned, median insert size and AT dropout) and FFPE preparation steps were analysed by linear regression to identify potential correlations between the two variables (Figure 4.9). Most of the preparation steps were associated with different sequencing metrics (the significances of the linear regressions were presented on Figure 4.9). Overall, the parameters leading to improved sequencing metrics were; small tissue block size, short fixation time (24-48 hours), in neutral buffered formalin with short time in hot wax (140 to 360 minutes), using the QIamp FFPE DNA extraction kit, with paraffin removal on slide using Histoclear, with an RNA removal step and a low de-crosslinking temperature (65°C).

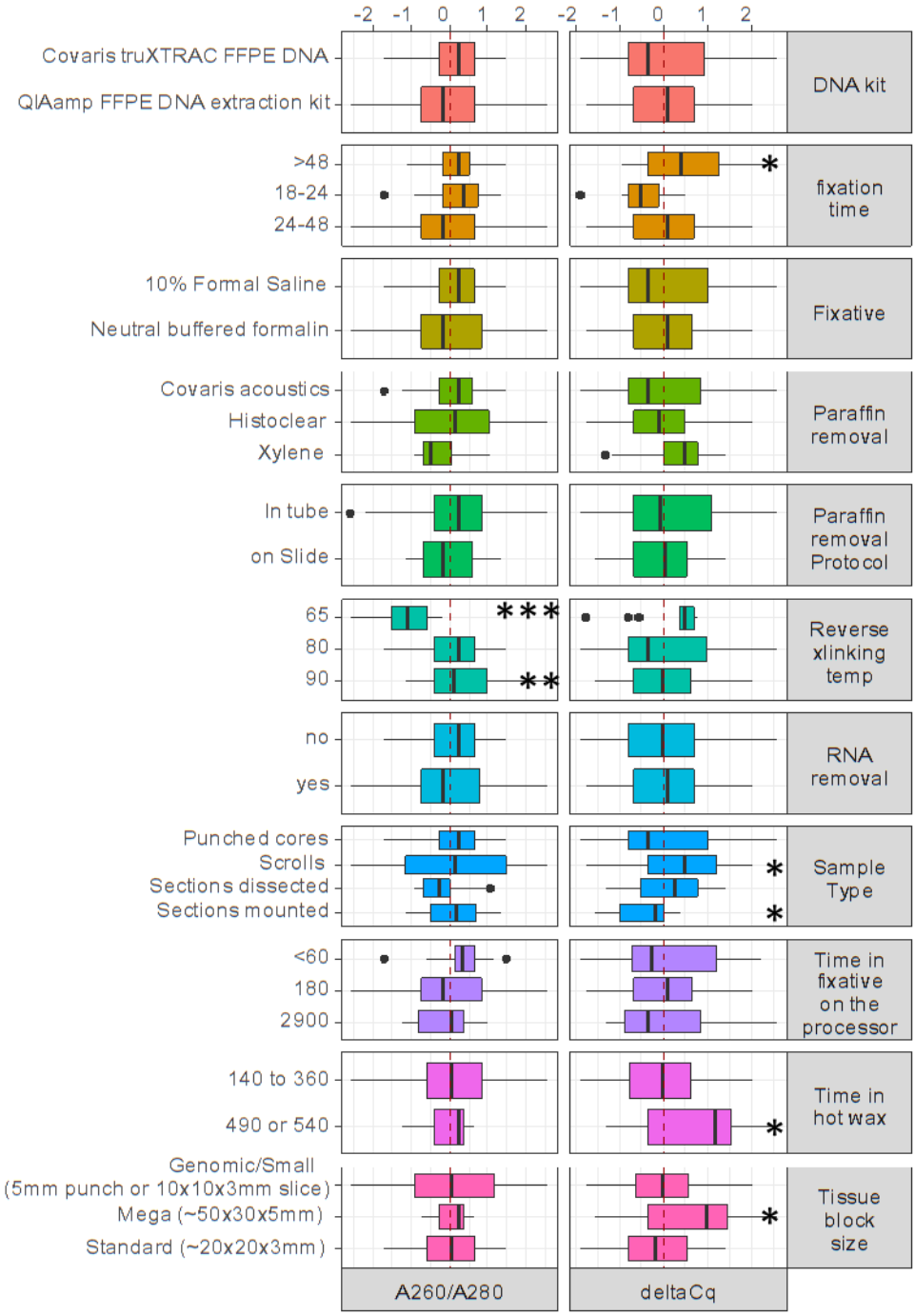


Figure 4.8 DNA quality control for different preparation conditions

Fixation time in hours, Reverse crosslinking temperature in degrees celsius, Time in fixative on the processor in minutes, Time in hot wax in minutes. Only variables with more than 10 datapoints were kept for the analysis. Linear regression p-value codes: '***' <0.001; '**' <0.01; '*' <0.05

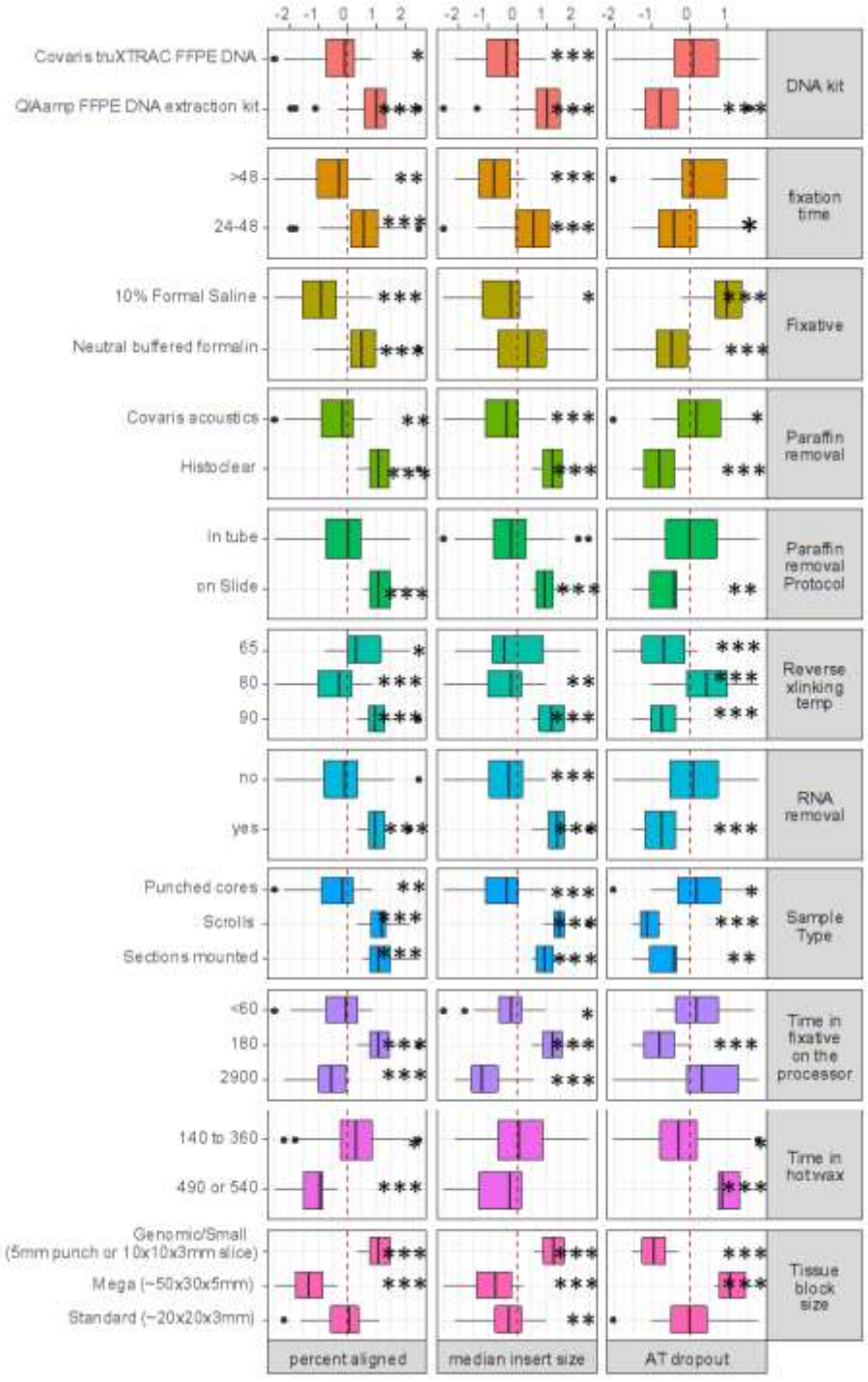


Figure 4.9 Sequencing metrics for different preparation conditions

Fixation time in hours, Reverse crosslinking temperature in degrees celsius, Time in fixative on the processor in minutes, Time in hot wax in minutes. Only variables with more than 10 datapoints were kept for the analysis. Linear regression p-value codes: '***' <0.001; '**' <0.01; '*' <0.05

In order to reduce the potential confounding bias between different parameters, I used a LASSO technique, which allowed the reduction of the number of covariates to the most important ones (Figure 4.10).

Major signals were observed from this analysis. The 10% Formal Saline fixative was associated with increased AT dropout and decreased percent of reads aligned. On the other hand, the small and genomic tissue blocks were associated with a decreased AT dropout and an increased median insert size. The time in fixative on the processor also influenced the quality of sequencing data: prolonged times (2900 minutes) were associated with lower median insert size and shorter times (180 minutes) with increased percent of reads aligned. The conclusion of this analysis is controlled fixation for shorter time is preferred using smaller tissue blocks to allow fast and optimal fixation throughout the tissue. In the case some tissues are large, a small genomic block should be prepared as FFPE sample separately from the main specimen to ensure fast and efficient fixation without compromising the quality of DNA.

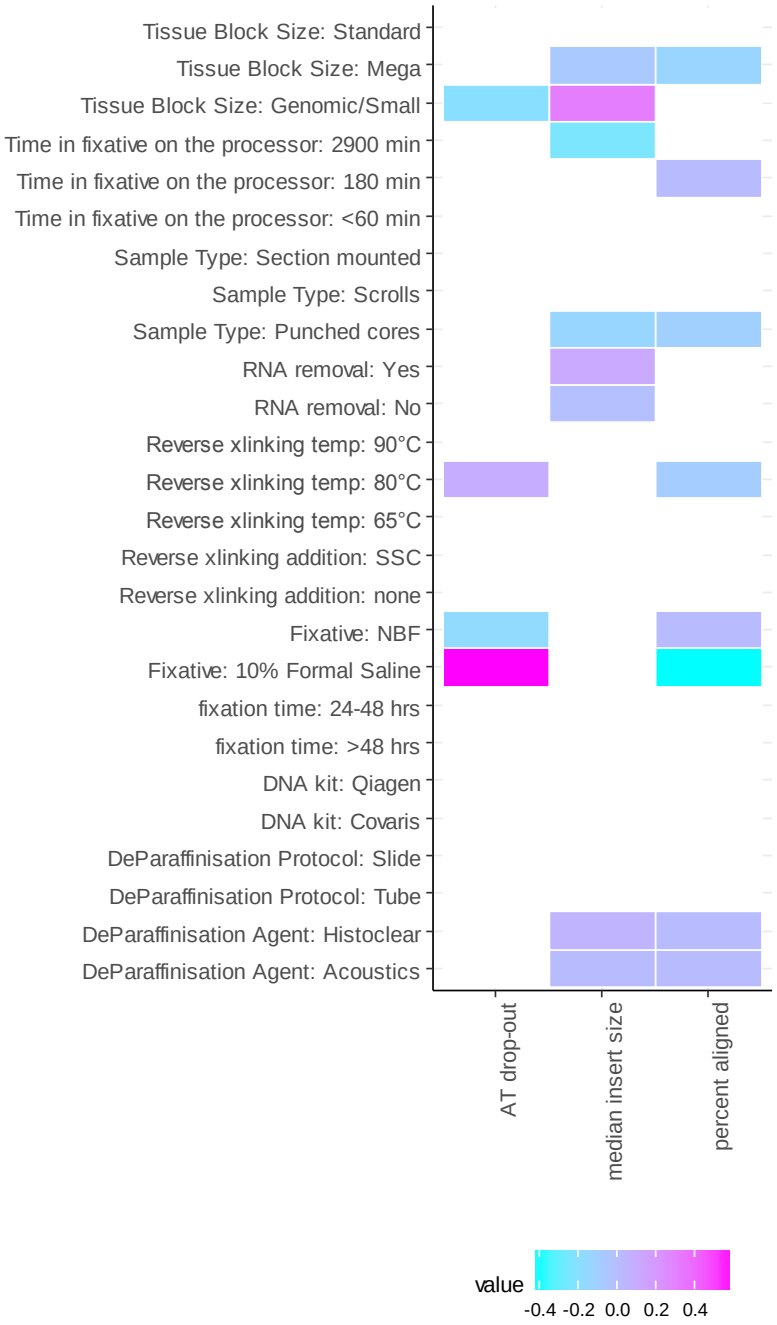


Figure 4.10 Effect size of FFPE preparation parameters on sequencing metrics when taking all parameters in consideration

The value represents the effect size on the sequencing metrics, in other words how much influence each FFPE preparation step had on sequencing metrics. A positive value represents a high sequencing metric value and a negative value a low sequencing metric value. White cells represent non-significant results where the absolute value of the effect size is less than 0.01.

Xlinking: crosslinking; NBF: neutral buffered formalin

4.4.4 New FFPE tissue preparation method did not benefit from optimised DNA extraction protocol

As highlighted by the previous multivariate analysis, the formalin fixation conditions (type of fixative and time of fixation) are crucial to improve sequencing metrics. Indeed, control of these conditions ensured an unbiased representation of AT rich regions as well as longer insert size and percentage of read aligned. These metrics suggested that CNAs detection could be improved by optimising the fixation. However, these observations were made using FFPE samples collected in non-controlled experiments in five different centres and therefore not in a controlled experiment.

In addition, it has not been confirmed whether the combination of optimisation of both fixation and DNA extraction presented in Chapter 4.4.2 could lead to even better data quality.

To investigate both tissue preparation and DNA extraction conditions, I performed WGS on 10 patients where the FFPE samples were prepared by varying both the fixation condition and the reverse crosslinking condition. The four conditions compared were the following: normal saline formalin fixation and de-crosslinking at 80°C (NS 80°C), normal saline formalin fixation and de-crosslinking at 65°C in the presence of SSC (NS 65°C+SSC), neutral buffered formalin fixation and de-crosslinking at 80°C (NBF 80°C), and neutral buffered formalin fixation and de-crosslinking at 65°C in the presence of SSC (NBF 65°C+SSC). Selected samples for five patients were subsequently sequenced (Table 4.2).

4.4.4.1 Quality control

Nucleic acids were extracted from FFPE tissues of 10 patients (four samples per patient) and quality of the DNA was assessed using a Nanodrop, Qubit and real-time PCR to calculate the ΔCq value (Figure 4.11).

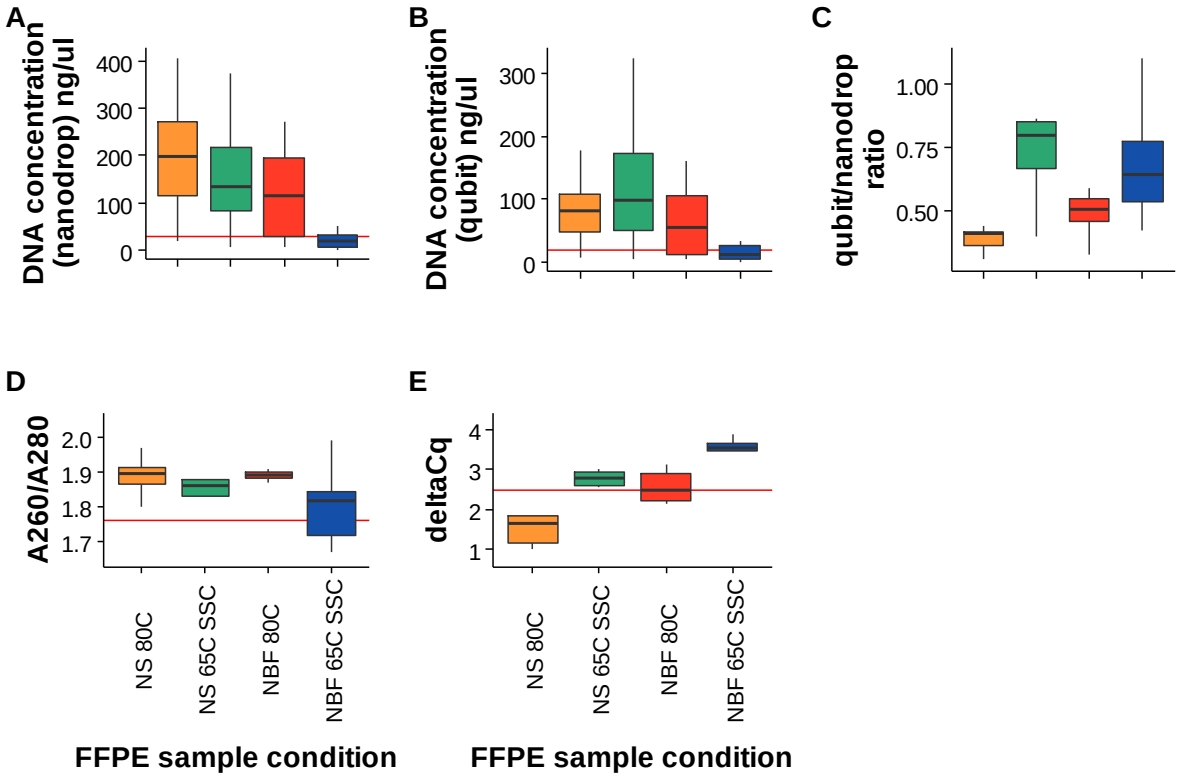


Figure 4.11 FFPE DNA quality control comparison of four preparation methods

Each box represents the distribution of the DNA metrics for 10 samples. DNA concentration in ng/µl measured by (A) Nanodrop and (B) Qubit. (C) Ratio of Qubit (dsDNA) and Nanodrop (all nucleic acids) concentration measurements. (D) A260/A280 ratio measured by Nanodrop. (E) ΔCq value measured by real-time PCR. The presence of a red horizontal line denotes the threshold of acceptable quality (above the line for concentrations and A260/A280, and below the line for ΔCq). NS 80C: normal saline formalin fixation and de-crosslinking at 80°C, NS 65C SSC: normal saline formalin fixation and de-crosslinking at 65°C in presence of saline-sodium citrate, NBF 80C: neutral buffered formalin fixation and de-crosslinking at 80°C, NBF 65C SSC: neutral buffered formalin fixation and de-crosslinking at 65°C in presence of SSC.

All 10 FFPE samples prepared as "NS 80°C" presented the highest DNA concentration measured by Nanodrop (Figure 4.11A). However, these samples presented the lowest Qubit to Nanodrop ratio with a mean of 0.39 [range 0.31-0.44] vs 0.63 [range 0.33-1] (Mann-Whitney U Test, $p=5.56e^{-05}$) (Figure 4.11C). All samples with a de-crosslinking step at 65°C in presence of SSC presented a higher Qubit to Nanodrop ratio compared to samples with a de-crosslinking step at 80°C (mean of 0.67 vs. 0.44, Wilcoxon Signed Rank Test, $p=9.65e^{-04}$) (Figure 4.11C).

In contrast, all 10 FFPE samples prepared with "NBF 65°C+SSC" showed the lowest DNA concentration values (measured by both Nanodrop and Qubit, Figure 4.11A-B) and most of these samples presented DNA concentrations under the acceptable threshold (mean = 16.4 ng/ul [0-57 ng/ul]). In addition, two out of 10 samples prepared with this protocol presented a suboptimal A260/A280 ratio below the QC threshold (Figure 4.11D).

The ΔCq values were found below the threshold of 2.5 in the majority of "NS 80°C" FFPE samples but failed to show statistical significance (mean = 2.7 [range 1.0-7.8] vs. 3.3 [range 1.7-10.1], Mann-Whitney U Test, p=0.082, Figure 4.11E).

Altogether, the FFPE samples prepared as "NS 80°C" and "NBF 65°C+SSC" presented extreme DNA QC results, and "NBF 80°C" and "NS 65°C+SSC" presented similar DNA QC data (except the Qubit/Nanodrop ratio). As the FFPE preparation "NBF 65°C+SSC" failed DNA QC most of the time, this condition was not investigated further. Given these DNA QC results, the FFPE condition "NBF 80°C" was investigated further over "NBF 65°C+SSC" which almost systematically failed DNA QC.

WGS was subsequently performed on four patients (four GL samples, four FF samples and seven FFPE samples) to assess the quality of FFPE samples preparation (Table 4.2).

Relevant alignment metrics were calculated for FFPE samples and compared to matched FF samples (Figure 4.12).

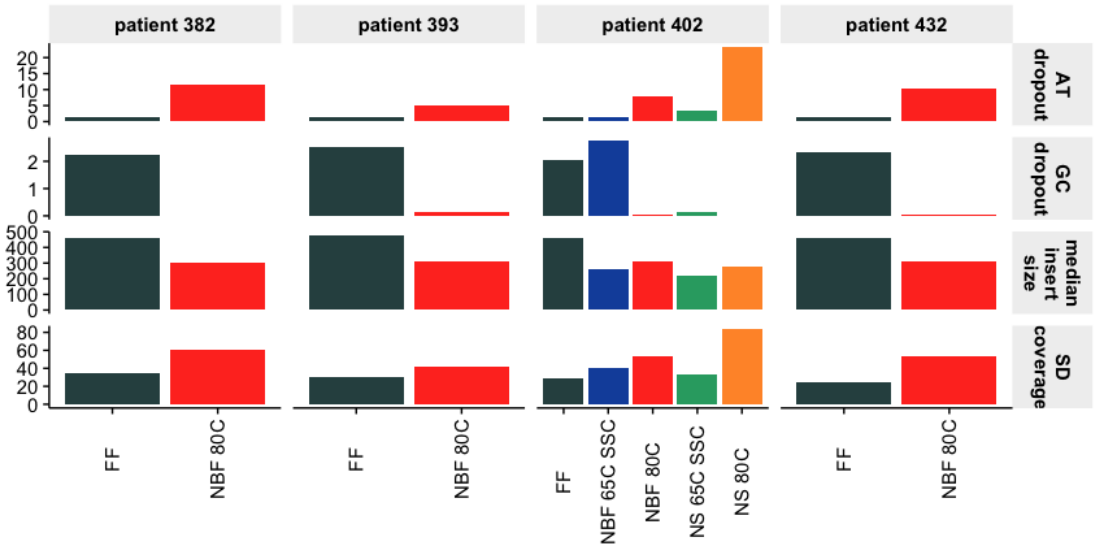


Figure 4.12 Alignment metrics for the FF and matching FFPE samples

AT and GC dropout are in percent, median insert size in bp and SD coverage in depth of coverage units (X). SD coverage: standard deviation of the depth of coverage

The AT dropout was still higher for the FFPE samples compared to the FF, however it was much higher for the condition NS 80C than the other FFPE conditions. This result reflected the GC dropout values. Interestingly, the FFPE sample NBF 65C SSC presented a GC dropout higher than the matching FF sample. The median insert size was lower for all FFPE samples compared to their respective matching FF samples and no difference was noted between the different FFPE conditions of patient 402. Finally, the SD of the depth of coverage, reflecting the wavy pattern of the sequencing depth of coverage was

higher in the FFPE samples. When comparing the different FFPE conditions in patient 402, NS 65C SSC had the lowest SD whereas NS 80C has the highest.

4.4.4.2 Copy number alterations

The alignment metrics indicated that FFPE samples prepared as "NBF 80°C" were still sub-optimal compared to matching FF samples, but improved in patient 402 compared to non-optimised FFPE sample (NS 80C), especially in term of AT dropout and SD of depth of coverage, which tends to correlate with more uniform Log_2R, as previously described in Chapter 4.4.2. Tumour-normal pairs were compared for FF and matching FFPE samples and data was visually inspected using the software Nexus (see Chapter 2.8.3.1) (Figure 4.13 and Appendix H). The data for patient 402 illustrated the wavy pattern for the FFPE sample "NS 80°C" as previously observed (Chapter 3.4.7). All three other FFPE samples showed an improvement in this pattern.

The correlation of Log_2R signals between FF and FFPE samples were calculated using the BIC-seq data on tumour-normal pairs (Figure 4.14). The Spearman's rank correlation coefficients of the data in FF and matching FFPE specimens showed variable results between the different patients, especially patient 393 with a correlation coefficient of 0.195. This low correlation was mainly due to chromothripsis events involving chromosome 1 (Appendix H).

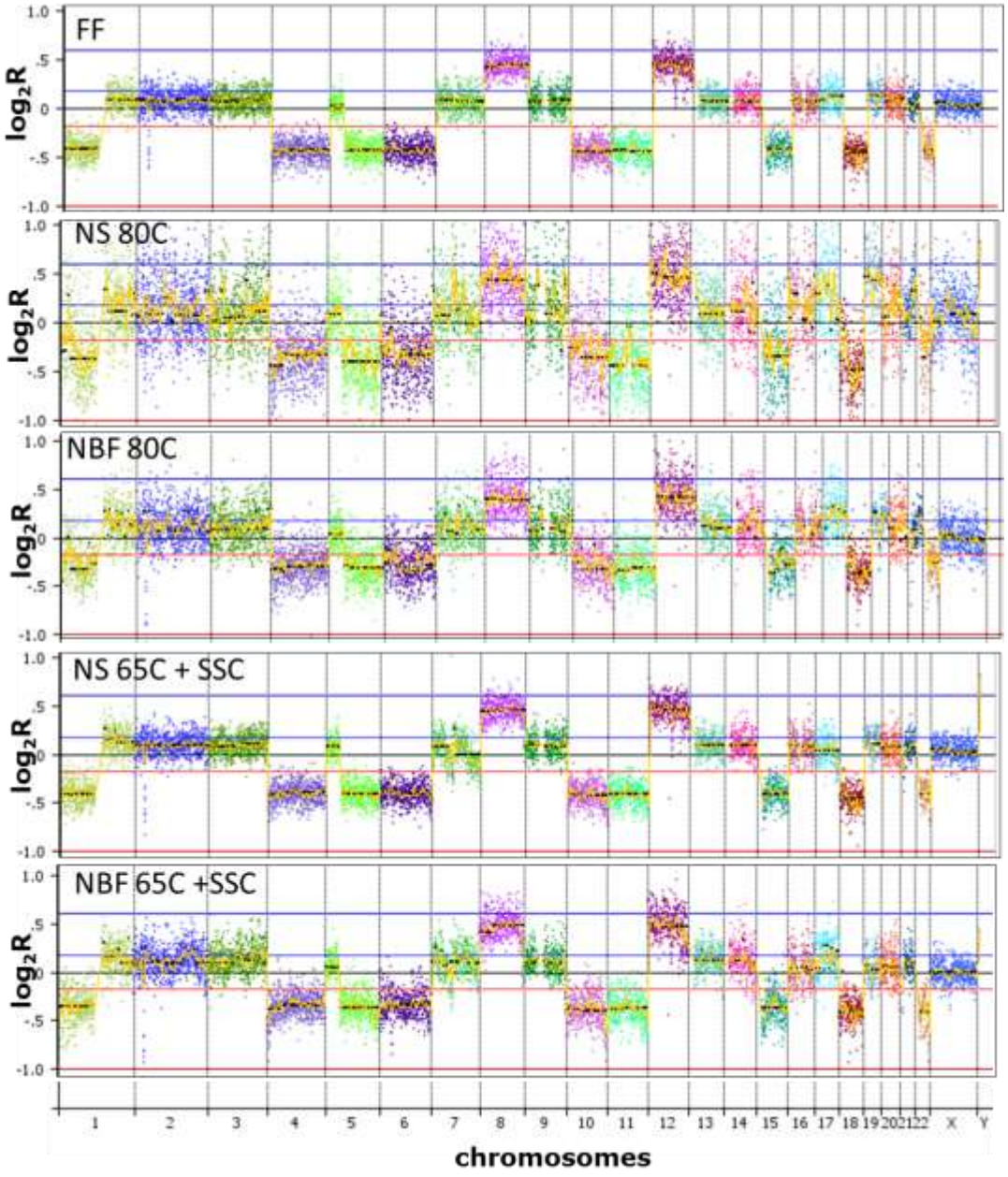


Figure 4.13 Log_2R across all chromosomes for five samples of patient 402

NS 80C: Normal saline formalin fixation and de-crosslinking at 80°C, NS 65C SSC: Normal saline formalin fixation and de-crosslinking at 65°C in presence of saline-sodium citrate, NBF 80C: neutral buffered formalin fixation and de-crosslinking at 80°C, NBF 65C SSC: neutral buffered formalin fixation and de-crosslinking at 65°C in presence of saline-sodium citrate.

For patient 402, the FFPE samples NS 65°C+SSC revealed the highest correlation of Log_2R with matching FF sample. The Log_2R correlation of NBF 65°C+SSC with FF was lower than NBF 80°C, suggesting that the optimisation of reverse crosslinking to 65°C+SSC was beneficial for FFPE samples prepared in NS but not in NBF.

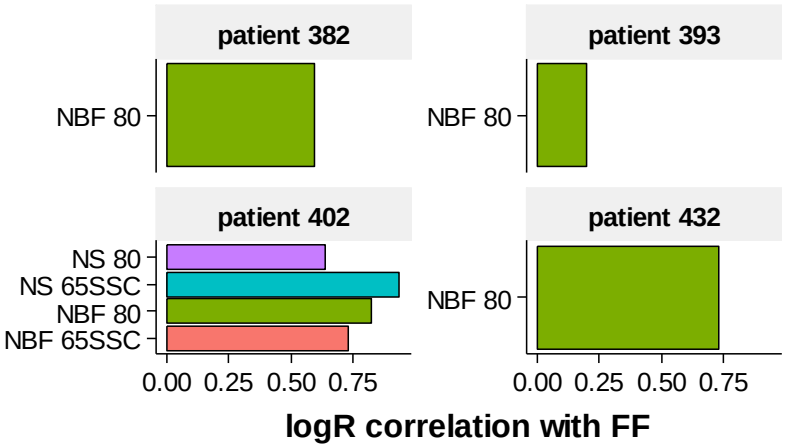


Figure 4.14 Spearman correlation coefficient values of log_2R from FF and matching FFPE samples

NS 80C: Normal saline formalin fixation and de-crosslinking at 80°C, NS 65C SSC: Normal saline formalin fixation and de-crosslinking at 65°C in presence of saline-sodium citrate, NBF 80C: neutral buffered formalin fixation and de-crosslinking at 80°C, NBF 65C SSC: neutral buffered formalin fixation and de-crosslinking at 65°C in presence of saline-sodium citrate.

4.4.5 WGS data from optimised FFPE DNA samples generated comparable clinical reports to matching FF samples

Clinical whole genome reports were generated for the five patients with optimised DNA extraction and the four patients with optimised tissue preparation protocols (Table 4.4).

Table 4.4 Patients' diagnoses

patient	tissue	diagnosis	Gender	age at surgery (years)
004	Renal	chromophobe renal cell carcinoma	M	65
032	Renal	clear cell renal cell carcinoma	F	76
065	Renal	Type II papillary renal cell carcinoma	M	83
300	Renal	Clear cell renal cell carcinoma	M	80
365	Renal	Clear cell renal cell carcinoma	M	74
382	Endometrial	Papillary serous adenocarcinoma	F	64
393	Breast	Lobular	F	60
402	Colorectal	Adenocarcinoma	F	66
432	Colorectal	Adenocarcinoma	M	48

M: male, F: female

These reports were compared to an independent analysis of the FF data. We reported somatic SNVs, indels and CNAs as well as cnLOH. However for patient 382, copy number changes were not reported and simply annotated as "complex karyotype" (Appendix I). Between 13 and 34 variants were reported in the patients (Figure 4.15). In five patients (including patient 382) the majority of variants detected were CNAs: 8 out of 16 for patient 004; 18 out of 32 for patient 365; 13 out of 21 for patient 393 and 19 out of 25 for patient 402. Two other patients presented a majority of frameshift variants (indels): 11 out of 28 for patient 032 and 11 out of 30 for patient 300.

The variants detected in the five renal patients were further detailed. Four patients (032, 065, 300 and 365) had a 3p alteration, leading to loss of function of the genes *VHL*, *MLH1*, *SETD2*, *BAP1*, which could be associated with sensitivity to PARP inhibitors (Peña-Llopis et al. 2012). Four patients (004, 065, 300 and 365) presented with an *ARID1A* alteration that could also be associated with sensitivity to PARP inhibitors (Shen et al. 2015).

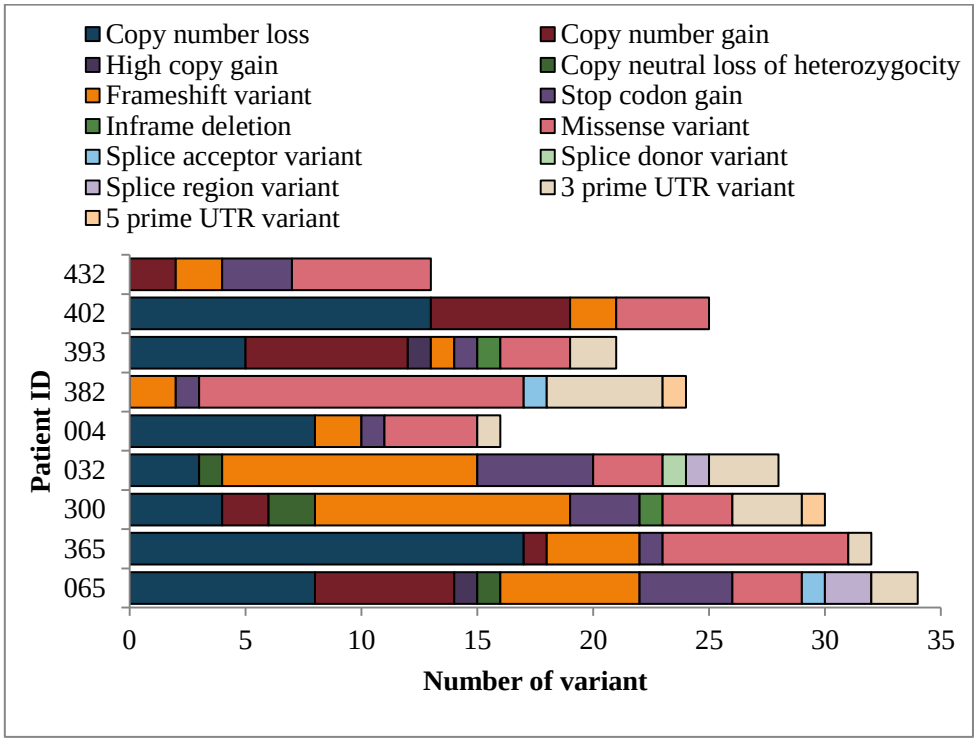


Figure 4.15 Clinically relevant variants reported in the nine patients

Copy number loss, gain, high copy gain and copy neutral loss of heterozygosity are structural variants. UTR: untranslated region.

In addition, three patients (004, 065 and 365) had a *PTEN* alteration, three patients (004, 065 and 365) showed an *EPCAM* alteration, two (300 and 365) had a *PIK3CA* variant and two (004 and 065) exhibited a *TP53* alteration. Four out of five patients showed bi-allelic alterations in driver genes leading to possible loss of function of the protein. Patient 004 presented a bi-allelic alteration of *TP53*: a heterozygous copy loss of the p arm of chromosome 17 associated with a G334W missense variant. Patient 032 and 365 showed a bi-allelic alteration of *VHL* with a heterozygous copy loss of chromosome 3 associated with a splice donor variant for patient 032 and a missense variant V130F for patient 365.

Patient 065 had a bi-allelic alteration of the gene *SETD2*: a heterozygous copy loss of chromosome 3 and a frameshift variant on amino acid R554PPX (Appendix I).

Two patients presented with colorectal adenocarcinomas. Both carried a G12V/D *KRAS* mutation. Patient 402 also had a copy number gain associated with that mutation. This type of alteration has been associated with sensitivity to treatments such as mTOR inhibitors or MEK inhibitors in association with other therapies and other options (Corcoran et al. 2013; Migliardi et al. 2012; Misale et al. 2014; Faber et al. 2014). In addition, patient 402 revealed other alterations that were potentially actionable, such as an *APC* frameshift mutation, a copy number loss over the *ATM* and *PTEN* loci, copy number gains of *ERBB3* and *MYC* and a *PIK3CA* missense variant.

Patient 393 presented with a lobular breast cancer carrying amplifications in genes potentially associated with response or resistance to treatment of *FGFR2* (FGFR inhibitors (Soria et al. 2014)), *HGF* and *MET* (resistance to Trastuzumab (Minuti et al. 2012)). This patient also presented a copy number loss of *ARID1A* (a possible other predictive biomarker for Trastuzumab resistance (Berns et al. 2016)), a frameshift variant of *CDH1* (potentially associated with sensitivity to Bicalutamide (Huang & Vakoc 2016)) and a subclonal *PIK3CA* missense variant only detected in the FFPE sample (possibly associated with response to Everolimus + Trastuzumab + chemotherapy (André et al. 2016)).

Patient 382 presented with papillary serous adenocarcinoma patient. Several CNAs were detected involving all chromosomes (Appendix H) and a high number of somatic SNVs and indels (859,203 in FF and 1,834,590 in FFPE pre filtering and 35,516 in FF and 177,731 in FFPE post filtering). However, this patient did not carry any SNV or indel classified in tier 1.

One hundred and ten SNVs and indels were reported in the nine FF samples and 130 in the nine corresponding FFPE samples. The median allelic fractions reported were 0.28 for FF and 0.29 for the FFPE datasets. Forty-eight percent of FF variants and 46% of FFPE variants had an AF between 0.3 and 0.6, which could be broadly assimilated to heterozygous clonal variants. Overall there was no significant difference between allelic fractions of variants reported in FF and FFPE datasets (Paired t-test, p= 0.34) and the correlation between the two groups was significantly positive (Spearman correlation, r=0.643, p=9.44e^{-17}) (Figure 4.16).

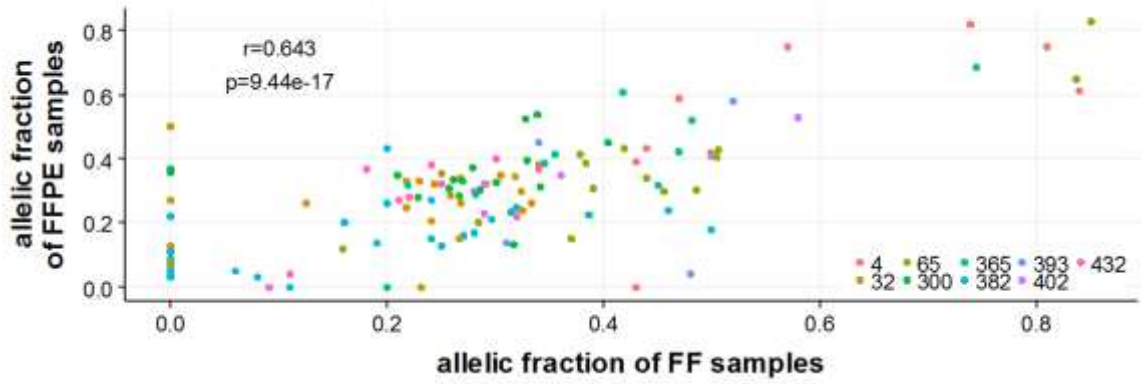


Figure 4.16 Comparison of allelic fractions of clinically relevant variants in FF and matching FFPE samples

Correlation coefficient r and p values were calculated using a Spearman's rank correlation. FF: fresh frozen, FFPE: formalin-fixed, paraffin-embedded

All variants reported were then classified into three tiers as described in Chapter 4.3.5. Fifty-four alterations were classified as tier 1 and clinically actionable, 106 as tier 2 and 62 as tier 3 across the nine patients (Table 4.5).

Table 4.5 SNV, indel and copy number alteration overlap of clinical report between FF and FFPE

		total	unique FF	unique FFPE	intersect	% intersect all variants
Sample	**tier**			*total (SNV + indel / CNA)*		
004	1	6 (2/4)	1 (0/1)	0 (0/0)	5 (2/3)	83 (100/75)
004	2	8 (4/4)	1 (1/0)	0 (0/0)	7 (3/4)	88 (75/100)
004	3	2 (2/0)	0 (0/0)	1 (1/0)	1 (1/0)	50 (50/NA)
032	1	4 (2/2)	0 (0/0)	0 (0/0)	4 (2/2)	100 (100/100)
032	2	17 (17/0)	1 (1/0)	3 (3/0)	13 (13/0)	76 (76/NA)
032	3	7 (5/2)	1 (0/1)	3 (3/0)	3 (2/1)	43 (40/50)
065	1	11 (1/10)	0 (0/0)	0 (0/0)	11 (1/10)	100 (100/100)
065	2	12 (10/2)	0 (0/0)	1 (1/0)	11 (9/2)	92 (90/100)
065	3	11 (7/4)	1 (0/1)	1 (1/0)	9 (6/3)	82 (86/75)
300	1	8 (2/6)	0 (0/0)	0 (0/0)	8 (2/6)	100 (100/100)
300	2	16 (14/2)	0 (0/0)	2 (2/0)	14 (12/2)	88 (86/100)
300	3	6 (6/0)	0 (0/0)	1 (1/0)	5 (5/0)	83 (83/NA)
365	1	11 (2/9)	0 (0/0)	0 (0/0)	11 (2/9)	100 (100/100)
365	2	11 (8/3)	1 (1/0)	0 (0/0)	10 (7/3)	91 (88/100)
365	3	10 (4/6)	0 (0/0)	0 (0/0)	10 (4/6)	100 (100/100)
382	1	0 (0/NA)	0 (0/NA)	0 (0/NA)	0 (0/NA)	NA
382	2	6 (6/NA)	0 (0/NA)	3 (3/NA)	3 (3/NA)	50 (50/NA)
382	3	17 (17/NA)	1 (1/NA)	6 (6/NA)	10 (10/NA)	59 (59/NA)
393	1	5 (2/3)	0 (0/0)	1 (1/0)	4 (1/3)	80 (50/100)
393	2	13 (3/10)	0 (0/0)	2 (2/0)	11 (1/10)	85 (34/100)
393	3	3 (3/0)	0 (0/0)	1 (1/0)	2 (2/0)	67 (67/NA)
402	1	8 (3/5)	0 (0/0)	0 (0/0)	8 (3/5)	100 (100/100)
402	2	14 (2/12)	0 (0/0)	1 (0/1)	13 (2/11)	93 (100/92)
402	3	3 (1/2)	0 (0/0)	0 (0/0)	3 (1/2)	100 (100/100)
432	1	1 (1/0)	0 (0/0)	0 (0/0)	1 (1/0)	100 (100/NA)
432	2	9 (7/2)	1 (1/0)	1 (0/1)	7 (6/1)	78 (86/50)
432	3	3 (3/0)	0 (0/0)	0 (0/0)	3 (3/0)	100 (100/NA)
Total	1	54 (15/39)	1 (0/1)	1 (1/0)	52 (14/38)	96 (93/97)
Total	2	106 (71/35)	4 (4/0)	13 (11/2)	89 (56/33)	84 (79/94)
Total	3	62 (48/14)	3 (1/2)	13 (13/0)	46 (34/12)	74 (71/86)

FF = fresh frozen, FFPE = formalin-fixed, paraffin-embedded, SNV = single nucleotide variants, CNA = copy number alteration, NA: not applicable

Of these alterations, 96% of those in tier 1, 84% in tier 2 and 74% in tier 3 were reported in both FF and FFPE. Twenty-seven variants were only detected in the FFPE data; 24 were SNVs and indels (with a majority in the UTRs) and 19 out of 25 were subclonal, with AF lower than 0.2. Eight variants, of which three were CNAs, were detected only in FF tissue data. In the end, 83 out of 88 CNAs were detected from both FFPE and FF, and only one not detected in the FFPE sample was classified as tier 1. These results confirmed the improved ability to detect CNAs from optimised FFPE samples.

4.5 Limitations of the study

In this study, all samples underwent "real-life" routine processing to preserve tumour architecture and ability to perform immunohistochemistry staining required for diagnostics. However, there is currently no standardization of FFPE processing and conditions used vary considerably depending on different cancer tissue types, resection or biopsy sizes and also according to different laboratories and technical equipment used. These variations in the preparation can lead to different types of DNA alterations likely to require different types of preparation optimisation.

In the statistical analysis of the FFPE tissue handling process some parameters were always combined together for specific tissue types, and therefore it was not always possible to identify single parameters influencing WGS quality metrics.

Finally, although the final experiment did not present a representative number of samples from different tissue types it provided some illustrative examples of clinical reports from FFPE samples. Visual inspection of Log_2R and BAF was necessary to confirm the

presence of a variant. Indeed, in genomes with many copy number changes and only few diploid regions, algorithms are not always accurate in calibrating the signal to CN=2. This resulted in the necessity to visually recalibrate the Log_2R baseline.

4.6 Discussion

The initial comparison of FF and FFPE WGS data highlighted the variability in sequencing quality of FFPE and the poor ability to call CNAs from FFPE samples preventing the application of WGS for cancer patients in the clinic. In this chapter, I investigated potential QC methods to predict poor quality of sequencing data in FFPE samples. In addition, I examined the FFPE tissue preparation and DNA extraction methods in order to optimise the sample quality. Clinical reports for the patients with optimised FFPE samples were compared to matching FF samples.

4.6.1 Different combinations of parameters lead to different types of sequencing artefact

This study showed that FFPE samples presented with different types and severities of degradation, such as high C>T transition rate, low SNV concordance with FF, sub-optimal alignment metrics or non-uniform coverage. However, it is crucial to predict which FFPE samples will yield poor quality WGS data independently of the type of degradation and where possible, to reduce any biases due to sample preparation (Schweiger et al. 2009; Oh et al. 2015). To this end, the only correlations found for the different effects of degradation in our datasets were the median insert size and the ΔCq values. This suggests

that combinations of several types of degradation lead to different types of sequencing artefact. As it was not possible to predict poor sequence quality using DNA metrics, it became essential to optimise DNA integrity of FFPE samples to improve coverage uniformity and therefore CNA detection.

4.6.2 Protocol to generate high quality FFPE specimens and detect CNAs

To increase CNA detection accuracy in FFPE samples, I optimised the DNA extraction methods for five patients' samples. By adapting the reverse crosslinking step during DNA extraction, the coverage uniformity was improved, particularly in AT rich genomic regions. Next, the statistical analysis of 203 FFPE samples comparing the DNA and sequencing metrics for different protocols highlighted that fixation conditions must be controlled. In particular, shorter fixation on smaller tissue size in neutral buffered formalin increased the alignment sequencing quality. Conflicting reports were previously published regarding the use of NBF and superior data quality (Miething et al. 2006; Barcelos et al. 2008). In this study, I found that NS fixation followed by DNA extraction with a reverse crosslinking temperature of 80°C generated WGS data of higher quality than with previous fixation protocols. However, the DNA extraction optimisation developed for FFPE samples fixed in 10% formal saline led to very low DNA yields in FFPE samples fixed with NBF, and worse CNA signal in the only sample tested.

Several crucial recommendations can be suggested in order to generate high quality FFPE samples for WGS. For archival samples, where FFPE tissues were already fixed using an uncontrolled protocol and where unbuffered formalin was most likely used for prolonged

times, the optimised DNA extraction protocol can be employed to rescue long strands of dsDNA molecules. For new patients included in a study, optimised tissue preparation should be followed. When possible, a small amount of genomic block should be collected and separated from the main specimen and fixed in buffered formalin to insure quick and even fixation. This quick step allows the reduction of DNA damage done by formaldehyde. DNA extraction should be performed as described in the manufacturer's protocol to ensure higher DNA yields. Further experiments are required to optimise DNA extraction from this type of FFPE samples.

4.6.3 Clinical WGS reporting of FFPE samples

As the coverage was more uniform, copy number calling was possible for nine patients. As a result, 83 out of 86 copy number changes identified in FF samples were independently called in FFPE samples. Importantly, 96% of clinically actionable (tier 1) variants were identified in the WGS data generated via optimised FFPE samples. In conclusion, several actionable variants were detected for all nine reported patients using FFPE samples.

4.6.4 Conclusions

The primary aim of this work was to screen FFPE samples using pre-analytical metrics to identify those that will likely produce poor quality data. This information could be used to save the cost of unnecessary WGS and allow for another FFPE sample to be prepared if available. However, none of the pre-established DNA quality measurements could successfully predict the WGS data quality.

The whole genome sequences from DNA prepared using the FFPE protocol optimised as part of this work can be interrogated for SNVs and indels affecting specific cancer related genes, as well as CNAs that are clinically relevant. The statistical work and validation analysis pointed out that fixation conditions are crucial to obtain FFPE data of acceptable quality for clinical grade WGS and also allowed for clinical WGS reports to include CNAs.

These results support the use of optimised FFPE cancer tissues as an alternative source of DNA for WGS cancer diagnostic applications when FF specimens are not collected or not available as a part of patients' routine clinical care.

Chapter 5 Multiple recurrent mutations confer poor trial outcome of relapsed/refractory CLL patients

5.1 Introduction

One of the significant challenges of molecular precision diagnostics for cancer patients is the lack of molecular predictors. Indeed, known driver alterations influencing prognosis or intervention is currently not detected in all cancer patients. Although some cancer types present a homogeneous genetic profile (Nowell & Hungerford 1960), most are characterised by the presence of multiple recurrently altered genomic features. CLL is one of these genetically heterogeneous diseases (Puente et al. 2011; Landau et al. 2015). As the cancerous cells of the disease are mainly found in the peripheral blood, sample collection and processing are straightforward which lead to comprehensive genomics characterization. In the next two chapters, CLL will be used as a cancer model to explore the discovery of molecular predictors. In this chapter, the genomic landscape of R/R patients, generally understudied, is characterised using MPS.

Since the emergence of MPS technique, CLL genomes have been comprehensively characterised in multiple studies (Wang et al. 2011; Puente et al. 2011; Schuh et al. 2012; Quesada et al. 2011; Landau et al. 2013; Landau et al. 2015; Puente et al. 2015; Burns et al. 2018; Ramsay et al. 2013; Fabbri et al. 2011; Fabbri et al. 2013). These studies led to the identification of multiple recurrently mutated genes, of which *SF3B1*, *ATM*, *TP53* and *NOTCH1* are the most frequent (Landau et al. 2015). The clinical impact of mutations in these genes was investigated in several cohorts of patients, especially untreated and first-line treated patients. *TP53*, *NOTCH1* and *SF3B1* mutations were associated with shorter OS in retrospective studies of untreated patients (Baliakas et al. 2014; Jeromin et al. 2014; Rossi, Rasi, et al. 2013) and shorter progression free survival (PFS) in prospective studies of first-line treated patients (Oscier et al. 2013; Stilgenbauer et al. 2014; Landau et al. 2015). In R/R patients, however, the impact of these mutations is unclear. In the CLL2H (Schnaiter et al. 2013) study, *NOTCH1* mutated patients treated with Alemtuzumab experienced a longer PFS than wild type (WT) patients, whereas in the CLL3X (Dreger et al. 2013) trial the presence of the mutation did not influence either PFS or OS of patients receiving allogeneic hematopoietic stem cell transplantation, although both trials were screening for coding mutations in the protein PEST domain. In both studies *SF3B1* mutations had no impact on survival. *TP53* mutations were shown to have no impact either in the CLL3X study and did not correlate with disease progression in another R/R study (Dufour et al. 2013). Generally, R/R patients have been overlooked because the cohorts are smaller, making it more difficult to carry out statistical testing. Therefore, the clinical impact of known CLL driver mutations needs to be investigated further.

In this work, the genomic landscape of R/R CLL patients was explored to identify potential predictive markers of patient outcome and to describe the evolution of the CLL genome at relapse. A cohort of 114 R/R patients recruited in three different clinical trials was studied by performing targeted MPS of known CLL driver genes.

5.2 Aims and objectives

The aims of this work were (i) to investigate in R/R patients the incidence of mutations and their co-occurrence and comparing this to the mutation pattern of treatment-naïve/unselected patients (ii) to identify predictive markers to treatment and (iii) to study patterns of clonal evolution by tracking driver mutations.

5.3 Material and methods

5.3.1 Patients and samples

114 R/R CLL patients treated by conventional CIT were included in this study: 55 were enrolled in the ICLL01 trial by the French CLL Intergroup prior to a planned interim analysis (phase two salvage treatment with Bendamustine, Ofatumumab, and methylprednisolone in relapsed CLL (de Guibert et al. 2014)); 37 were included in the CLL201 trial (a randomised phase 2 trial of Fludarabine, Cyclophosphamide, and Mitoxantrone with or without rituximab in previously treated CLL (Hillmen et al. 2011)); and 26 in the CLL202 trial (a phase 2 study of subcutaneous Alemtuzumab plus

Fludarabine in patients with Fludarabine refractory CLL (Varghese et al. 2010)) (Table 5.1).

Table 5.1 Patient characteristics

	median [range] **n (%)**
no. of patients	114
Age	66 [32-79]
Gender	
Male	92 (81)
Female	22 (19)
Clinical staging	
Binet A, Rai 0	15 (13)
Binet B, Rai I-II	47 (41)
Binet C, Rai III-IV	52 (46)
No. of previous therapeutic lines	
1	52 (46)
>1	62 (54)
Fludarabine-refractory	
yes	35 (31)
no	79 (69)
Cytogenetics (FISH)	
17p deletion	26 (23)
11q deletion	29 (26)
missing	4 (3.5)
***IGHV* mutational status**	
mutated	18 (15.8)
unmutated	78 (68.4)
missing	18 (15.8)

Peripheral blood mononuclear cells were obtained from the French CLL Intergroup for the patients in the ICLL01 trial and from the UK National Cancer Research Institute CLL Trials biobanks for the patients enrolled in the CLL201 and CLL202 trials.

5.3.2 Extraction of genomic DNA from peripheral blood mononuclear cells and quality control

Genomic DNA samples were extracted from peripheral blood mononuclear cells using the QIAamp DNA mini kit (Qiagen, Hilden, Germany), see Chapter 2.3.3. DNA QC was then performed using a Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA) for quantification of all nucleic acids and Qubit dsDNA Broad Range Assay kit (Thermo Fisher Scientific, Waltham, MA, USA) for quantification of double stranded nucleic acids as described in methods Chapter 2.4.1 and 2.4.2.

5.3.3 TruSeq Custom Amplicon targeted library preparation

Specific genomic targets of interest were amplified from DNA samples using the TruSeq Custom Amplicon preparation (Illumina Inc., San Diego, CA, USA) according to previously described methods (Chapters 2.6.2 and Table 2.3).

5.3.4 MiSeq massively parallel targeted sequencing

Libraries prepared were sequenced on a MiSeq instrument (Illumina Inc., San Diego, CA, USA) by 300 bp PE sequencing as described previously (Chapter 2.7.2).

5.3.5 Bioinformatics pipeline

Sequencing reads were aligned and variants called as described in Chapter 2.9.1.

5.3.6 Single nucleotide polymorphism microarray

SNP array was performed as described in Chapter 2.10.

5.3.7 Statistical analysis

5.3.7.1 Associations, response and survival analysis

Associations between alterations were tested using Pearson's Chi-squared or Fisher's exact test as appropriate. Univariate analysis on overall response rate (ORR) and complete response rate (CRR) with genetic alterations was performed using logistic regression. The Kaplan-Meier method was used to study patients' PFS and OS and comparisons were performed using log-rank testing. Univariate analysis of survival (PFS and OS) with genetic alterations was performed using proportional hazards models (Cox proportional-hazards model).

All variables with $P<0.1$ on univariate analysis were included in a multivariate analysis. Multivariate Cox models were defined after backward selection. The models were validated by bootstrapping: means estimates of each covariate coefficient and standard error were calculated.

5.3.7.2 Subclonal analysis

The likely subclonal distribution was calculated using variant allelic fractions (VAF) from the targeted sequencing results. Data was corrected for copy number using 11q (*ATM*) and 17p (*TP53*) status by SNP array. Where two mutations in the same patient showed a VAF

difference of <5%, the average was taken to calculate the percentage of cell carrying the mutations assuming they occurred in the same clone; in cases where the VAF difference was >5%, the two mutations were presumed to be independent (the mutation with the highest VAF occurred prior the second lower VAF mutation).

5.3.8 Individual contributions

In this chapter, Sanger sequencing experiments and analysis of the *TP53* gene was performed by Dr Véronique Leblond. Statistical analysis was performed by Dr Bruno Pereira. The analysis was done in collaboration with Dr Romain Guieze.

5.4 Results

5.4.1 Target sequencing coverage was acceptable for variants analysis

Nine genes recurrently mutated in CLL (*TP53, SF3B1, ATM, NOTCH1, XPO1, SAMHD1, MED12, BIRC3* and *MYD88)* were investigated by targeted sequencing (Table 2.3). Prior to the mutation analysis, sequencing performances were assessed using the application real-time analysis (RTA) v.1.18.42 (Illumina, San Diego, CA, USA). A mean of 14.1M reads were obtained per run, of which 96.8% were identified PF reflecting an acceptable signal to noise ratio. The yield was 4.1Gb and 95.9% of these reads had a Phred score above 30. The median depth coverage was 2266X. The mean depth of coverage across all genes was greater than 100X for 97% and 500X for 92% of patients. Only one gene,

NOTCH1 presented a sub-optimal coverage. However, the known driver hotspot mutation in this gene (7541_7542delCT) was sufficiently covered, with 96% of patients with at least 100X coverage and 88% with at least 500X coverage. The sequencing metrics were therefore acceptable to allow analysis of variants.

5.4.2 Reads re-alignment allowed the detection of clinically relevant variants

A subset of *TP53* mutations identified by targeted sequencing was validated using conventional Sanger sequencing (Table 5.2). All 13 variants were confirmed to be present by the alternative method.

Table 5.2 Sanger sequencing validation of *TP53* mutations

Chromosome	Position (hg19 reference)	Reference allele	Variant allele	Allelic fraction	Validated by Sanger sequencing
17	7577108	C	A	0.39	yes
17	7577100	T	C	0.51	yes
17	7577580	-	A	0.94	yes
17	7578212	G	A	0.22	yes
17	7577559	G	A	0.38	yes
17	7574003	G	-	0.68	yes
17	7577538	C	T	0.26	yes
17	7578219	-	T	0.16	yes
17	7573934	G	A	0.38	yes
17	7577025	T	A	0.16	yes
17	7577120	C	T	0.22	yes
17	7577538	C	T	0.28	yes
17	7577121	G	A	0.78	yes

In addition, the Sanger sequencing analysis revealed a 29 bp deletion in *TP53*, not detected by the MiSeq Reporter bioinformatics solution (v.2.3.32.0) in the targeted

sequencing data. Therefore, the targeted reads of all samples were re-aligned using the alignment tool Stampy (v.1.0.22) and the 29 bp deletion was then found alongside three additional indels and two stop gain variants in other samples' genes (Figure 5.1). These variants were added to the dataset for all further analyses.

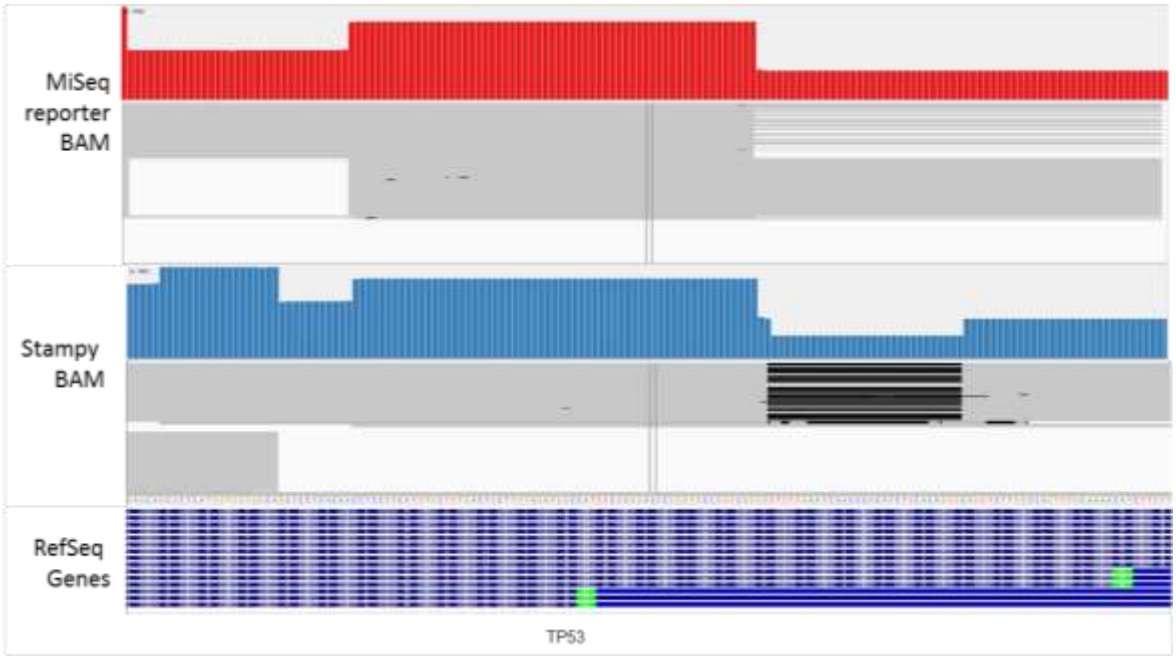


Figure 5.1 *TP53* deletion of 29 bp visualised in Integrative Genomics Viewer (IGV)

The upper panel represents the alignment data from MiSeq reporter and the middle panel shows the reads aligned using Stampy. For these two panels, the top tracks in blue and red show the depth of coverage and the bottom tracks show read alignment viewed as pairs. The lower panel shows the RefSeq coding sequence of the gene *TP53*. Black horizontal bars represent a deletion.

5.4.3 Mutational landscape revealed recurrently mutated genes and associations between alterations

191 mutations were identified in the 114 patients (mean of 1.7 mutations per patient): 135 missense mutations, 40 substitutions/indels, 12 nonsense (codon stop losses or gains) and four splicing mutations. *SF3B1*, *ATM* and *TP53* were the most mutated genes, found in 32 (28%), 30 (26%) and 16 (23%) patients, respectively, and were largely affected by missense mutations. *NOTCH1* was mutated in 17 (15%) patients and was mostly affected by indels. The incidence of other genes' mutations was distributed as follows: *XPO1* mutations in 17 (15%), *SAMHD1* mutations in 11 (10%), *MED12* mutations in 10 (9%), *BIRC3* mutations in six (5%), and *MYD88* mutations in three (3%) patients (Figure 5.2).

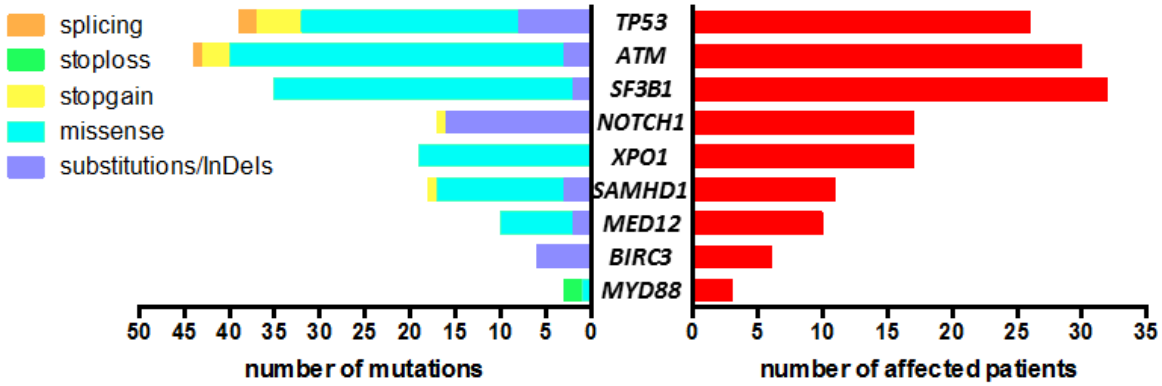


Figure 5.2 Type of mutations (left) and number of patients with (right) for each gene

Mutations in *TP53* mainly occurred in the DNA binding domain, those in *SF3B1* were predominantly p.K700E/c.A2098G (46%). 88% of *NOTCH1* variants were the 2-bp deletion (c.7541_7542delCT) in exon 34. By contrast, *ATM* mutations were scattered

across the gene, with four in the telomere-length maintenance and DDR domain (TAN), five in the FAT domain and three in the phosphatidylinositol 3- and 4-kinase domain (PI3 PI4 kinase) (Figure 5.3).

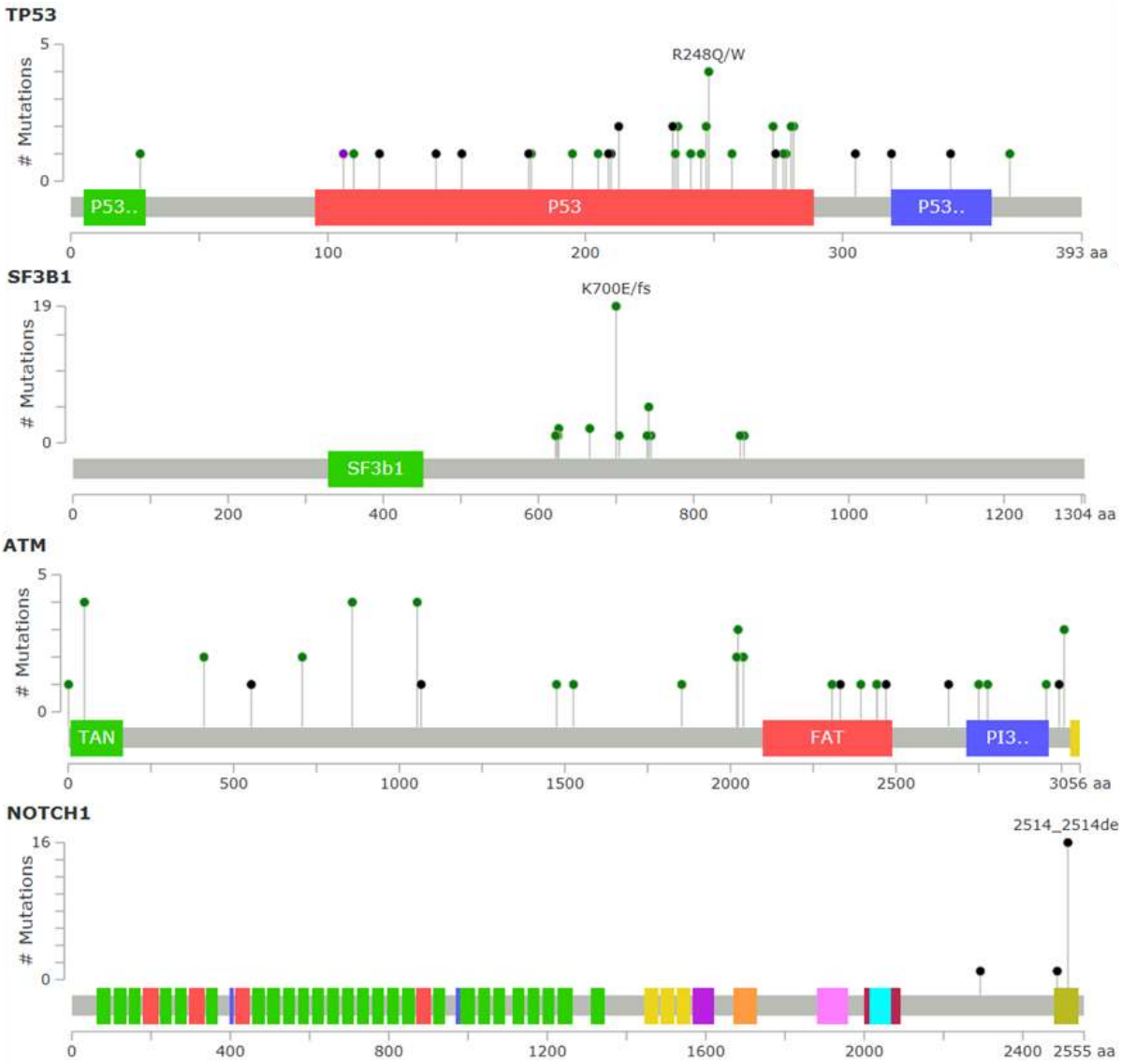


Figure 5.3 Distribution of mutations across *TP53*, *SF3B1*, *ATM* and *NOTCH1* protein structure

Protein domains represented by coloured boxes were obtained from the Pfam database

Associations between mutations and copy number changes were investigated. This revealed the well-known associations between *TP53* mutations and 17p deletions (65.4% vs 9.4%, P < 0.001). 55% of 11q deleted patients had either mutated *SF3B1* (n=9) and/or mutated *ATM* (n=7), without reaching statistical significance. Associations between mutation and *IgHV* mutational status were also studied. The well characterised association between TP53 mutations and unmutated *IgHV* was confirmed in this cohort (26.9% vs 0%, p=0.01, Fisher's exact test). All patients with mutated *MYD88* had mutated *IgHV* (n=3) and all *BIRC3* mutations occurred in *IgHV* unmutated patients (n=6).

5.4.4 Mutated genes distribution revealed recurrent associations, or multiple hit profile (MH-profile)

Associations between different mutated genes were investigated. Patients were clustered together according to (i) the number of mutated genes detected (zero in cluster 1, one in cluster 2, ≥two in clusters 3 and 4) and (ii) the recurrence of combinations (≥5% of cases in cluster 3 or <5% of cases in cluster 4) (Figure 5.4). 19% of patients were found in cluster 1, 43% in cluster 2, 19% in cluster 3 and 18% in cluster 4. Cluster 3 presented mutation(s) in at least two of the following genes: *TP53*, *SF3B1* and *ATM*, and was defined as the multiple hit profile (MH-profile). These three genes and *XPO1* were more frequently mutated when co-occurring with each other than when occurring as a sole alteration. In cluster 4, five patients presented mutations in both *XPO1* and *NOTCH1* and four patients had a *NOTCH1* and *SF3B1* mutations. However, these combinations of events were too rare to perform statistical analysis separately.

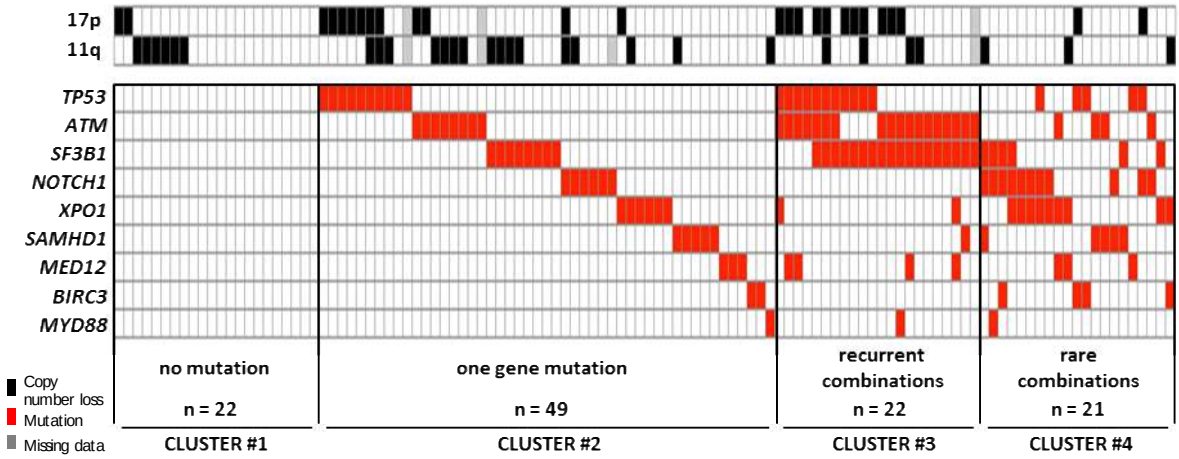


Figure 5.4 Patient clustering according to gene mutations

The upper panel shows copy number losses over 17p and 11q, the lower panel presents gene mutations. Each column represents one patient. Recurrent combinations of mutations are found in > 5% of the cohort; rare combinations in < 5%.

Association of these clusters with other genetic and clinical features were studied. Clusters were found independent of *IgHV* mutational status and clinical trials. MH-profile was significantly associated with 17p deletions (p=0.002, Pearson's Chi-square test).

5.4.5 Patients harbouring the MH-profile displayed a significant poor response to treatment

Clinical and mutation data were explored by statistical analysis to assess their significance on response to treatment. ORR was first considered. When genes were analysed individually, *TP53* the only disrupted (mutated or deleted) gene to be predictive of ORR (57% vs. 81%, p=0.007, logistic regression). When the four clusters were studied, the MH profile was associated with poorer ORR (42.9% for MH profile vs. cluster 1: 86.4%,

cluster 2: 83.3% and cluster 4: 71.4%, p=0.002, logistic regression) (Table 5.3). Fludarabine refractoriness status was the only other marker of poor ORR (55.9% vs. 81.8%, p=0.004). The full results of univariate analysis in Appendix J. Of note, TP53 mutated patient presented a significant shorter PFS compared to unmutated (12 vs. 20.7 months, p=0.004, logistic regression) and 17p deleted patients presented lower ORR and CRR compared to patient with this deletion (ORR: 57.1 vs. 81.3$, p=0.007; CRR: 8.6 vs. 26.7%, p=0.03, logistic regression). A multivariate analysis using logistic regression highlighted only two independent markers of poor ORR: the MH profile (p=0.006) and Fludarabine refractoriness status (p=0.002). *TP53* disruption did not retain its significance in this multivariate analysis.

Table 5.3 Impact of the multiple-hit profile on response by univariate analysis

	ORR (%) n = 111	p-value	CRR (%) n = 111	p-value
Clusters				
#1	86.4		27.3	
#2	83.3	0.002	27.7	0.058
#3 (MH profile)	42.9		0	
#4	71.4		19	

Test: logistic regression, ORR = overall response rate, CRR = complete response rate, MH: multiple hit

CRR was also studied by univariate analysis. *TP53* disrupted and *ATM* mutated patients presented with a significantly lower CRR (p=0.03 and p=0.045) (Appendix J). Interestingly, no patient with an MH profile reached complete response, compared to 27.3% for cluster 1, 27.7% for cluster 2 and 19% for cluster 4 (p=0.058) (Table 5.3).

5.4.6 MH-profile was found to be a prognostic marker of shorter survival

In addition to response to treatment, survival was also studied by statistical analysis.

By univariate analysis (Cox proportional-hazards model), PFS was significantly lower for *TP53* mutated patients (12 months vs. 21.7 months, p=0.004), Fludarabine refractoriness status (12.3 months for refractory patients vs 22.5 months, p=0.001) and the MH profile (12 months for MH profile vs 19.1 for cluster 1, 23 for cluster 2 and 17.5 for cluster 4, p=0.02) (Table 5.4, Figure 5.5 and Appendix J).

Table 5.4 Impact of the multiple-hit profile on survival by univariate analysis

	Median PFS (months) n = 114	p-value	Median OS (months) n = 114	p-value
Clusters				
#1	19.1		NR	
#2	23	0.018	46.6	0.019
#3 (MH profile)	12		28.2	
#4	17.5		27.1	

Cox proportional-hazards model, PFS = progression free survival, OS = overall survival, NR: not reached

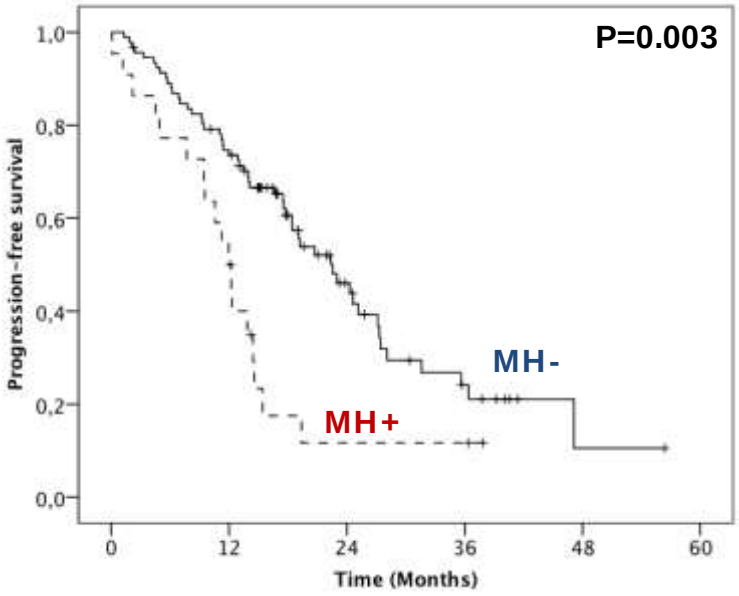


Figure 5.5 Impact of the multiple hit profile on progression free survival

Log-rank test. MH = multiple-hit profile.

Multivariate analysis results of PFS were in line with response results. Indeed, Fludarabine refractoriness status and MH profile were found to be the only two independent prognostic markers (p=0.004 and p=0.015, respectively, Cox proportional-hazards model), where *TP53* mutation was not found significant after bootstrapping process (Table 5.5).

Table 5.5 Multivariate analysis for progression-free survival (treatment adjusted)

Prognostic factors	HR	95% CI	p-value	p-value*
Fludarabine-refractoriness	3.36	1.65-6.86	0.001	0.004
TP53 mutation	1.81	1.02-3.21	0.043	0.077
Multiple-hit profile	2.24	1.24-4.03	0.007	0.015

*After bootstrapping process (1000 samples). Cox proportional-hazards model, HR = Hazard ratio; CI = confidence interval, P: p-value

The last marker to be assessed was OS. By univariate analysis, Binet stage and Fludarabine refractoriness status were found to be of adverse prognosis (46.6 months for

stage A, 63.6 for stage B and 27.1 for stage C, p=0.018; 21.9 months for Fludarabine refractoriness vs. 40.4 months, P < 0.001 Cox proportional-hazards model). Patients in cluster 1 (no mutation) presented with significantly good prognosis (median OS not reached vs. 46.6 months for cluster 2, 28.2 for the MH profile and 27.1 for cluster 4, p=0.019, Cox proportional-hazards model) (Figure 5.6, Appendix J).

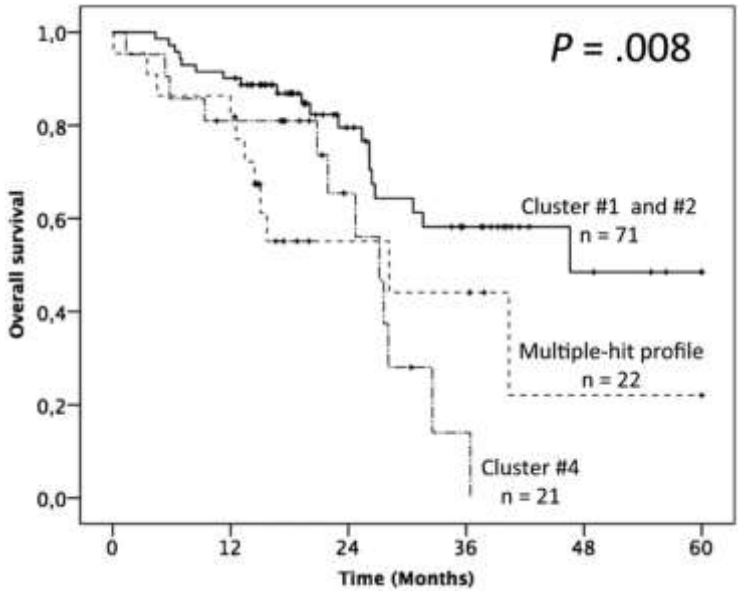


Figure 5.6 Impact of the multiple-hit profile on overall survival

Two models were calculated by multivariate analysis (Table 5.6). In model A, >1 mutations (MH profile and cluster 4) presented an independent poor impact (p=0.001, Cox proportional-hazards model), whereas in model B, the MH profile alone retained significance (p=0.025, Cox proportional-hazards model). In addition, in both models Fludarabine refractoriness status was found to be a predictor of poor OS (p=0.001 in model A and p=0.003 in model B, Cox proportional-hazards model). *BIRC3* mutations

were also found to be a predictor of poor OS in both models, although this marker lost its significance after the bootstrapping process.

Table 5.6 Multivariate analyses for overall survival (treatment adjusted)

Multivariate models	Prognostic factors	HR	95% CI	P*
Model A	Fludarabine-refractoriness	4.1	1.75-9.64	0.005
	BIRC3 mutation	3.41	1.15-10.12	0.835
	MH profile & #4	2.91	1.52-5.60	0.003
Model B	Fludarabine-refractoriness	3.53	1.51-8.22	0.014
	BIRC3 mutation	4.79	1.57-14.61	0.786
	MH profile	2.85	1.37-5.94	0.025

Cox proportional-hazards model. Model A includes clusters 3 and 4 vs 1 and 2, and model B includes cluster 3 (multiple-hit profile) vs 1, 2, and 4. *After bootstrapping process (1000 samples). HR: Hazard ratio; CI: confidence interval, MH: multiple hit, P: p-value

5.4.7 MH profile preserved its prognostic impact in *TP53* disrupted patients

The MH profile was found to be a marker of response and survival, independent of *TP53* disruption. This independent nature was investigated by a second analysis. To do so, a statistical analysis was performed on the sub cohort of *TP53* disrupted patients (n = 35) divided into two groups: MH profile (dis-*TP53* MH+, n = 14) and non-MH profile (dis-*TP53* MH-, n = 21). ORR in dis-*TP53* MH+ was found to be poorer than in dis-*TP53* MH-

or than *TP53* WT patients (p=0.01, logistic regression). In addition, PFS was shorter for dis-*TP53* MH+ than dis-*TP53* MH- (p=0.03, Cox regression) (Figure 5.7).

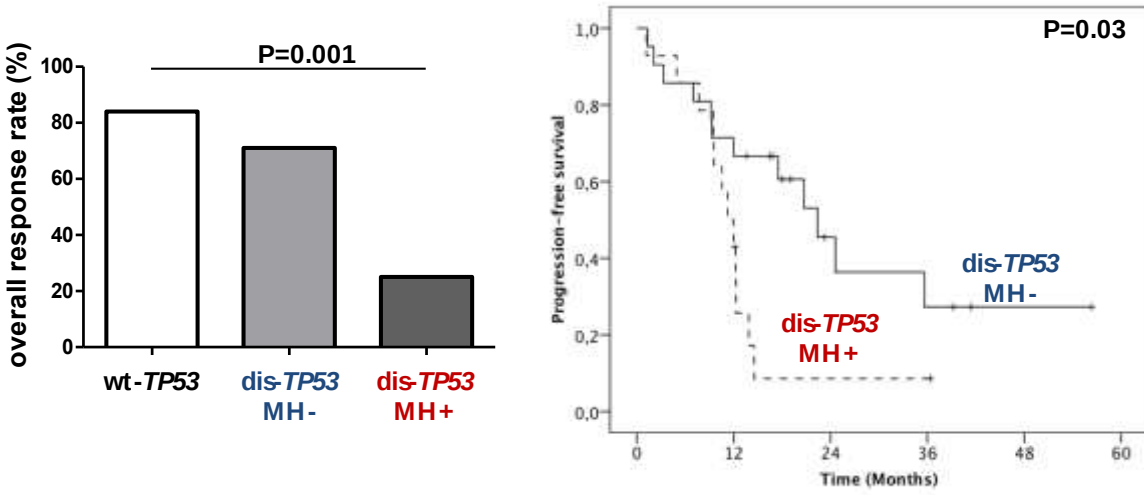


Figure 5.7 Impact of the multiple-hit profile on response (left) and survival (right) in the *TP53* disruption subgroup

wt = wild type, dis- = disrupted, MH = multiple hit profile. Log-rank test.

5.4.8 Recurrent patterns of clonal evolution were discovered

The MH profile included patients with mutations in at least two of *SF3B1*, *ATM* or *TP53*. The likely subclonal distribution of MH profile patients (n = 22) was investigated by combining VAF information from targeted sequencing data and copy number deletions of chromosome 17p and 11q by SNP array. CCFs were estimated by correcting the VAFs with allelic copy number status according to the equation:

$$Cancer\ cell\ fraction\ variant\ allele\ frequency\ x\ copy\ number\ at\ the\ same\ locus$$

CCFs of different mutations in the same patient were studied (Figure 5.8).

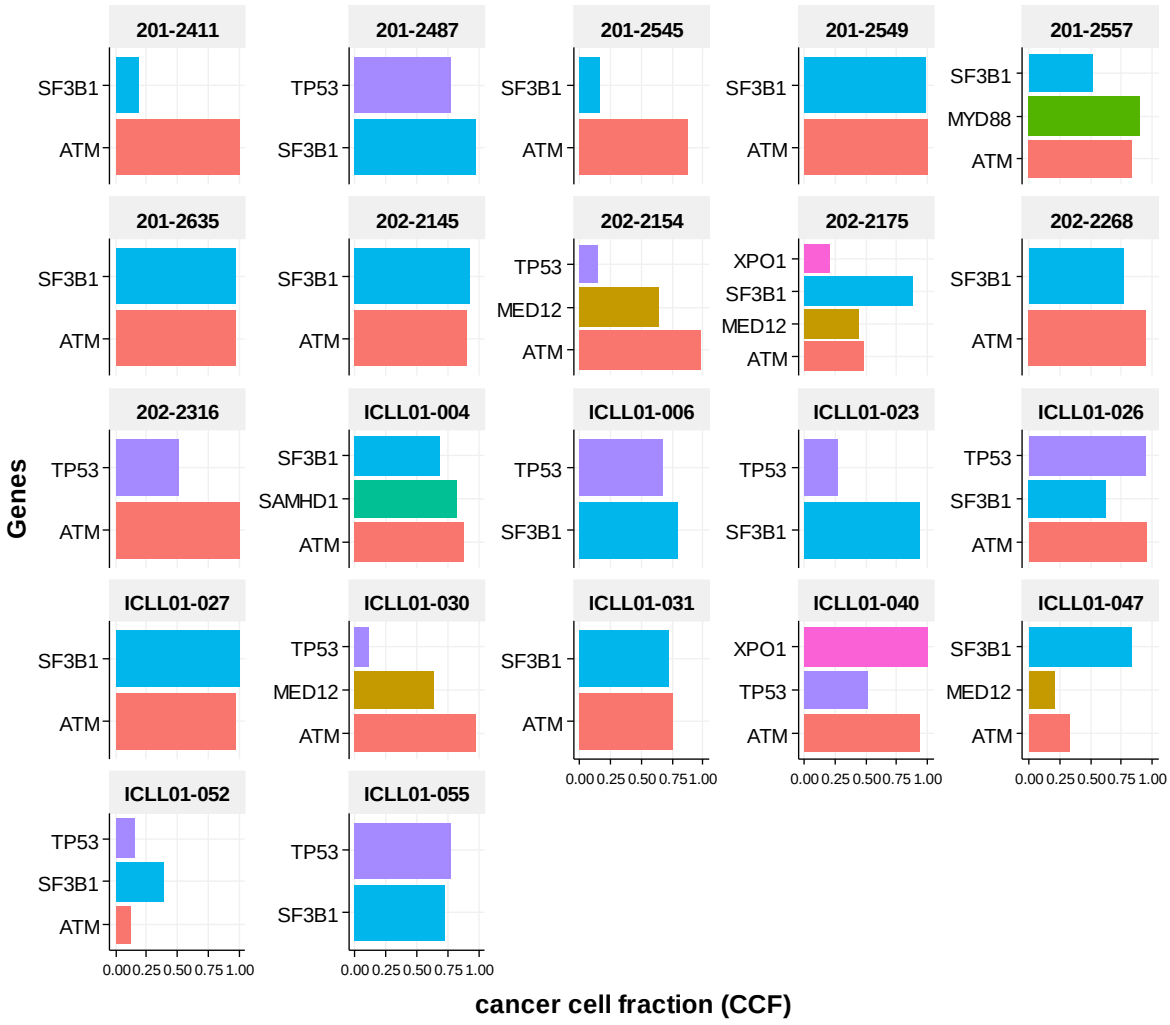


Figure 5.8 Cancer cell fraction estimated for mutations reported in cluster 3 patients.

Variants with CCFs > 0.95 were defined as clonal and all others were defined as subclonal. Ten patients presented with at least one clonal mutation in the genes screened for and four presented with two genes mutated with CCF > 95%. These patients all carried both *SF3B1* and *ATM* mutations.

Applying the "pigeonhole principle" (Nik-Zainal, Van Loo, et al. 2012), first and secondary mutation events were inferred using CCFs. *ATM* mutations were found as first events in 15 and *SF3B1* in 11 out of 22 patients. *TP53* occurred more frequently as a secondary event. Indeed, it was the case for eight of the 10 patients with *TP53* mutations. Four recurrent patterns were discovered in the cohort fitting 18 out of 22 patients. In pattern 1, found in five patients, both *SF3B1* and *ATM* mutations arose from the same clone (Figure 5.9). Most of these patients presented several alterations in the same gene. In this group, patient ICLL01-031 presented low CCFs. Therefore two different models of clonal evolution were hypothesised. In model 1, the mutation in *SF3B1* and *ATM* with the high CCF were the main driver mutations. The cells carrying the other three mutations (two *SF3B1* and one *ATM*) also carried the two main driver mutations. In model 2, some cells carried one *ATM* and *SF3B1* mutations and other carried the other three mutations.

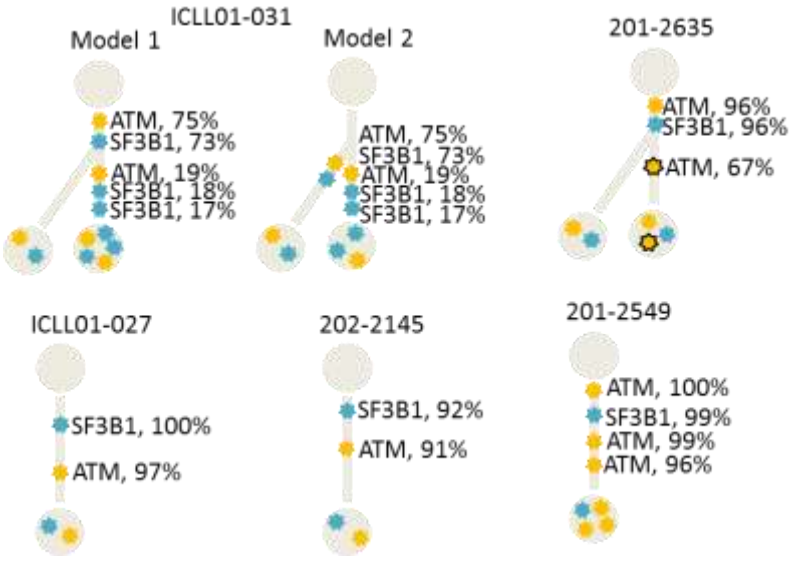


Figure 5.9 Clonal evolution pattern 1: *ATM* and *SF3B1*

The percent represent the cancer cell fraction. Stars represent mutations, a dark outline represents a homozygous mutation and diagonal patterns show a copy number deletion associated to the mutation.

In pattern 2, *ATM* mutations arose first with *TP53* mutations following (4 patients; Figure 5.10). In two of these patients, a homozygous *MED12* mutation was associated with several *TP53* mutations observed as secondary events. These two patients also presented with two possible models of subclonal distribution due to low CCFs.

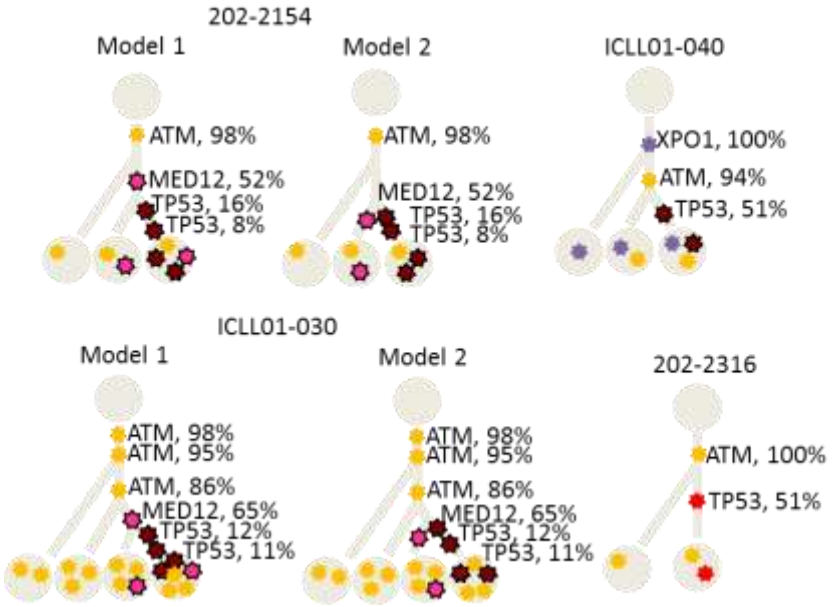


Figure 5.10 Clonal evolution pattern 2: *ATM* first, *TP53* second

In pattern 3, *ATM* mutations were first events and *SF3B1* mutations secondary events (5 patients; Figure 5.11). Like in pattern 1, several *ATM* mutations are acquired in these patients.

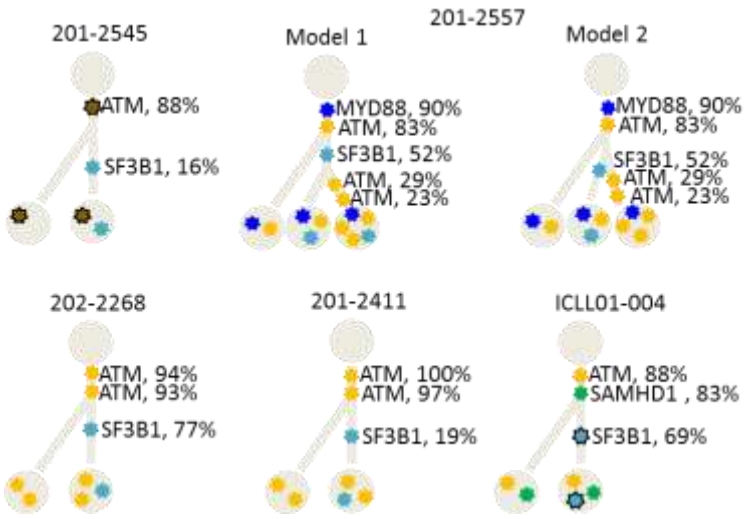


Figure 5.11 Clonal evolution pattern 3: *ATM* first, *SF3B1* second

In pattern 4, *SF3B1* mutations originated first, followed by a *TP53* mutation (4 patients; Figure 5.12). In this group, two patients presented two models because they carried variants with low CCFs.

In this analysis, the top circle represented the most recent common ancestor cell that was unmutated. The bottom circles symbolised the likely most prevalent subclones in the sample collected pre-treatment. The proportion of each of these subclones, or subclonal composition was estimated for the 22 patients of the MH profile (Figure 5.13).

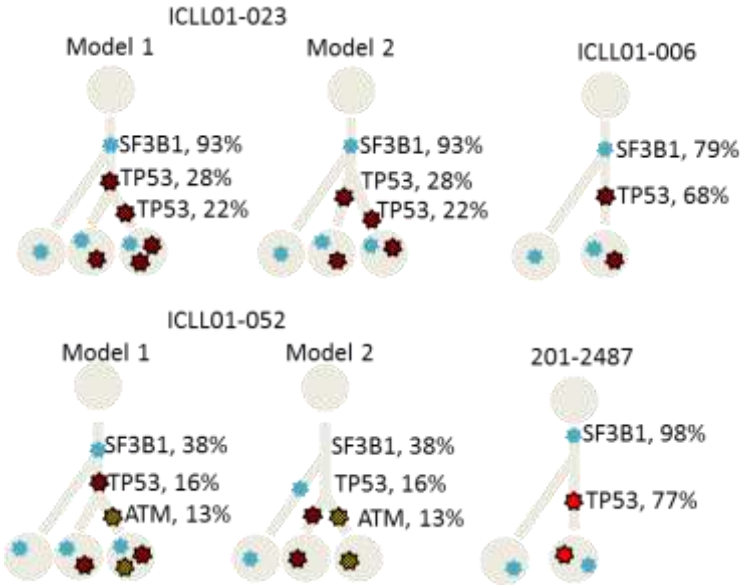


Figure 5.12 Clonal evolution pattern 4: *SF3B1* first, *TP53* second

Subclone 1 represented the initial subclone (the most recent common ancestor cell) that does not carry any driver mutations. Four patients showed two major subclones, 11 patients had three subclones; six patients had four subclones and finally one patient had five major subclones.

To summarise, the MH profile presented with 22 patients with mutations in the genes *ATM*, *SF3B1* and/or *TP53*. However, these patients did not all show the same clonal evolution that led to that particular subclonal composition at the time they entered the clinical trial. Indeed, some patients presented with more complex types of clonal evolution with multiple mutations in the same gene (especially in *ATM* and *SF3B1*) and multiple major subclones (up to 5 in one patient).

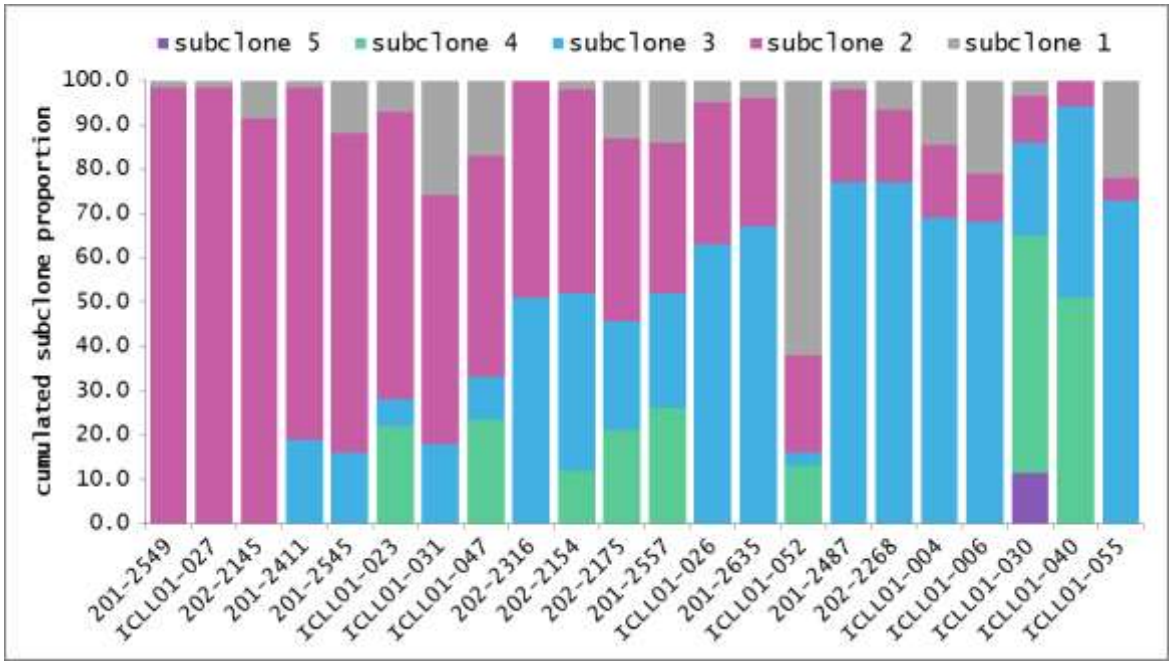


Figure 5.13 Likely subclonal composition of pre-treatment samples in MH profile patients

5.5 Limitations of the study

This study described the genomic landscape of R/R CLL patients, with some interesting results, although, there is scope to expand the study.

The custom amplicon oligo pool (Table 2.3) was designed before major studies in the field were published (Landau et al. 2015; Puente et al. 2015) which identified some important target genes generally less frequently observed in CLL including *POT1*, *CDKN2A*, *EGR2*, RPS15 and *FBXW7*.

Optimisation and validation of the analysis pipeline was performed in order to detect true positive mutations and filter out false positive ones. Although two aligners and two variant callers were used to identify variants, no method of somatic detection will accurately

detect all variants. Therefore, some undetected true variants might remain. Also, in the absence of GL samples matching the CLL sample, variant filtering is challenging and relies mainly on annotations from publically available databases and prediction scores. This method is therefore limited by the quality of the databases. Of note, dbSNP is known to contain pathogenic variants (Mitchell et al. 2004) and COSMIC includes mutational artefacts (from TCGA for instance (Buckley et al. 2017)). The variant annotations available at the time of data analysis will influence the filtering process and the final list of variants.

The mutation distribution revealed recurrent associations of mutated genes (MH profile with *TP53*, *ATM* and *SF3B1*), but also rare associations (in less than 5% of the cohort) grouped in cluster 4. Of note, some of these patients presented combinations of *XPO1* and *NOTCH1* mutations and *SF3B1* and *NOTCH1* mutations. Cluster 4 presented an even poorer OS than the MH profile. However, because of the cohort size, it was not possible to test statistically the impact of each rare combination separately. Larger cohorts of patients are required to address this question.

Clonal evolution was also analysed in this study using targeted sequencing of one timepoint in 114 patients. This analysis provided some insights on subclonal composition of the CLL samples; however, it also presented some limitations. Most algorithms available require WES data to guaranty a sufficient number of datapoints, such as ABSOLUTE (Carter et al. 2012). Some signal correction was not possible given the type of data available. Indeed, AF normalization could not be performed because of the lack of datapoints (number of mutations reported) for each patient. In addition, the clonal composition described is only a partial representation as only nine genes were screened.

Finally, the different patterns of clonal composition described were not tested in regard to clinical outcome as the number of patients in each group was too low.

5.6 Discussion

In this work, the mutational landscape of R/R CLL patients was studied, including examination of the incidence and co-occurrence of mutations in key CLL drivers and their utility as potential response predictors to therapy. Finally, patterns of clonal evolution were studied using the same sequencing data.

5.6.1 Mutation frequency comparisons highlighted similarities and differences with other CLL cohorts

Nine recurrently mutated genes were investigated in 114 R/R patients. The mutation frequency reported in this cohort was compared with previously published data. However, most previous studies have focused their analyses on heterogeneous and predominantly treatment naïve cohorts of patients (Baliakas et al. 2014; Jeromin et al. 2014; Rossi, Rasi, et al. 2013; Oscier et al. 2013; Stilgenbauer et al. 2014; Schnaiter et al. 2013; Dreger et al. 2013; Dufour et al. 2013). Furthermore, R/R patients have mainly been examined in retrospective studies using Sanger sequencing. Therefore, these studies focused on only a limited number of genes such as *TP53*, *SF3B1* and *NOTCH* (Baliakas et al. 2014; Schnaiter et al. 2013; Dreger et al. 2013). The frequency of *NOTCH1* mutations was comparable between studies (12-15%). Different incidences of *TP53* mutations were

found in different publications: 23% of cases in this cohort (n=114), 13% in the European research initiative on CLL (ERIC) study (n=3490 patients, mainly before treatment) (Baliakas et al. 2015), 30% in the CLL3X trial (n=90) (Dreger et al. 2013) and 39% in the CLL2H (n=97) (Schnaiter et al. 2013). For *SF3B1,* differences were also reported: 16% mutated cases in ERIC study, 17.5% in CLL2H, 26% in CLL3X and 28.1% in this cohort. These differences could be explained by the difference in methods employed to screen for mutations. The targeted sequencing panel used in this work was more sensitive and therefore better able to detect low frequency variants compared to Sanger sequencing. In addition, this panel covered larger regions of the genes, in particular covering all exons of *TP53*, rather than just the DNA binding domain from exons 4 to 9, and four full exons of *SF3B1* around the hotspot amino acid residue K700. Differences in mutation frequencies could also be explained by different inclusion criteria due to a different definition of R/R patients for each trial.

This work reported new insights on the distribution of the other genes targeted in this panel. Indeed, these genes have not been previously investigated in R/R patients. Therefore, they were compared with other CLL cohorts in first-line treatment. Generally, the frequency of the other genes was higher in the R/R cohort. Of note, *ATM* was found mutated in 26% in this cohort, and 14%, 15% and 15% in other study (Guarini et al. 2012; Skowronska et al. 2012; Landau et al. 2015). The incidence of *SAMHD1* was higher in this cohort than in an untreated cohort, as well as *XPO1* (15% vs. 3%, 5%) (Jeromin et al. 2014; Landau et al. 2015). Also, the prevalence of *BIRC3* mutation in our cohort was twice that reported by Baliakas and colleagues (Baliakas et al. 2014) and Rossi et al. (Rossi, Rasi, et al. 2013). Finally, *MED12* mutations were also enriched in this cohort

(9%) compared to other cohort (5%, 2%) (Kämpjärvi et al. 2014; Landau et al. 2015). Conversely, *MYD88* was less frequently mutated in the cohort, consistent with its association with mutated *IgHV* status.

5.6.2 Specific genes showed an impact on prognosis

Mutations detected in genes were aligned with outcome data such as response to treatment and survival times in order to identify predictive markers to treatment. As all the patients were enrolled in prospective trial studies, the clinical outcome data was well annotated avoiding possible bias in the way these metrics are collected.

In this cohort of R/R patients, *SF3B1* and *NOTCH1* did not influence outcome, confirming a previous similar study (Schnaiter et al. 2013). Indeed, *NOTCH1* mutations were even found to be associated with better PFS than WT *NOTCH1* patients. In this study, additional genes were tested that are not routinely tested, such as *ATM*, *XPO1* or *SAMHD1*. However, with the exception of *TP53* no gene alone was correlated with poor outcome. Only co-occurrence of mutated genes, and in particular at least two out of *ATM*, *SF3B1* and *TP53* were associated with poorer survival. This work changes the conventional routine type of analysis of single gene testing towards an integrated approach. This study showed that several genes should be tested in addition to different type of alterations: SNVs, indels and CNAs. In addition, this cohort presented many patients with more than one driver gene mutated, contradicting a previous study (Rossi et al. 2011) but agreeing with other studies (J. Wang, Hossein Khiabanian, Rossi, Fabbri, et

al. 2014; Landau et al. 2015). This work illustrated that alterations in known drivers are not mutually exclusive at relapse.

5.6.3 Clonal evolution was heterogeneous in the cohort

Cancer as an evolutionary concept has been extensively described for the past few decades, and it has also been detailed in CLL with the emergence of MPS techniques (Landau et al. 2013; Schuh et al. 2012; Knight et al. 2012; Ahn et al. 2017; Rose-Zerilli et al. 2016; Ouillette et al. 2013; Stilgenbauer et al. 2007; Kadri et al. 2016; Zhao et al. 2016). In this analysis of 114 patients, one bulk sample taken at the time of treatment was used to infer subclonal composition. Despite the obvious limitations of such analysis using a sequencing targeted approach on a single timepoint, several patterns of clonal evolution were unravelled. *TP53* was confirmed to be a late event in most patients (J. Wang, Hossein Khiabanian, Rossi, Fabbri, et al. 2014) as opposed to *ATM* and *SF3B1* which were early events. This data also suggested that the accumulation of alterations in several driver genes leading to genomic complexity or complex karyotypes and subclonal heterogeneity drives CLL evolution. The link between genomic complexity and poor outcome of CLL patients was demonstrated in many studies (Ouillette, Collins, Shakhan, Li, Peres, et al. 2011; Yu et al. 2017; Anderson et al. 2017). In addition, the presence of subclonal driver mutations was also was also shown to adversely influence survival of patients (Landau et al. 2013; Schuh et al. 2012; Landau et al. 2015).

This work highlights what could be currently feasible in routine diagnostics for CLL patients as it does not require more than one aliquot of DNA. Larger cohorts of patients

are required to confirm the patterns discovered and the clinical implication of such subclonal composition.

5.7 Conclusion

This work presented the genomic landscape of 114 R/R CLL patients enrolled in prospective clinical trials. DNA samples were screened for nine known CLL drivers using a targeted sequencing panel. After establishment of the bioinformatics pipeline and validation of a subset of variants by Sanger sequencing, we revealed recurrently mutated genes (the main ones being *SF3B1*, *ATM* and *TP53*) and previously described co-occurrence of lesions such as 17p deletion and *TP53* mutations, or mutated *IgHV* and *MYD88* mutations.

We have identified that the combination of mutations in at least two of the three genes *ATM*, *SF3B1* and *TP53*, defined as the MH profile, conferred a poorer response to conventional chemotherapy and shortened survival. Clonal evolution of relapse/refractory CLL patients was found to be extremely heterogeneous but with some defined patterns emerging, generally with *SF3B1* and *ATM* mutations as early events and *TP53* mutations as late events. Some patients presented a highly complex CLL genome with up to five major subclones detected.

In this work, the lack of molecular predictors was addressed in a cohort of R/R CLL patients, a patient group generally understudied due to the low number of patients recruited in clinical trials. Further work is required to study larger cohorts of relapsing

patients, different types of alterations, and a more comprehensive list of genomic regions (such as non-coding regions).

Chapter 6 WGS in the clinic identifies CLL patients with high risk of relapse

6.1 Introduction

In chapter 5, molecular predictors were detected in R/R patients, a generally less studied group of CLL patients. However, the majority of patients requiring first-line therapy lack molecular predictors that can be tested in the clinic.

CIT with FCR is the standard of care in first-line treatment of CLL for fit patients and for patients with good performance status (Fischer et al. 2016). With this approach great differences in response and survival have been measured (Hallek et al. 2010; Tam et al. 2014; Fischer et al. 2016; Thompson et al. 2016). On one hand, more than 50% of *IgHV* mutated patients never relapse, even after median follow-up of 12.8 years after CIT, which could make this group functionally cured and ideal candidates to receive FCR in first-line therapy (Thompson et al. 2016). On the other hand, 25% of FCR-treated patients relapse within 24 months and therefore do not benefit from this treatment (Hallek et al. 2010). Of these patients, the known adverse prognostic marker *TP53* is mutated or deleted in only a fraction of R/R patients (Stilgenbauer et al. 2014). So far, mutations and/or deletions of *TP53* remain the only predictive marker screened for in routine clinical practice (Thorsten

Zenz, Mertens, et al. 2010; Hallek et al. 2008), accounting for only one third of patients relapsing early after FCR (Stilgenbauer et al. 2014; Landau et al. 2015). A recent study has revealed a novel candidate predictor of early relapse: somatic mutations in the gene *RPS15* (Landau et al. 2015; Ljungstrom et al. 2016). Taken together with *TP53* disruption, these only occur in a subset of high-risk patients. Therefore, molecular predictors for patients receiving FCR treatment are still required.

As described in Chapter 5, most CLL cohorts studied so far focus on patients in the early phase of the disease (including MBL and CLL at diagnosis). This allows recruiting a high number of patients and performing robust statistical analysis. However, these are non-uniform cohorts of patients receiving heterogeneous therapies, with variable response assessment strategies in routine clinical practice (Wang et al. 2011; Landau et al. 2013; Puente et al. 2015; Burns et al. 2018). The exception was the study from Landau and colleagues, investigating the CLL8 cohort (first-line treated patients) using exome sequencing (Landau et al. 2015). Therefore, it is necessary to perform additional MPS studies on different groups of CLL patients.

Current practice to provide a molecular classification to a CLL patient requires several distinct laboratory-based tests. This multi-testing approach involves techniques such as conventional metaphase karyotyping (Han et al. 1984), fluorescence *in situ* hybridisation (FISH) (Döhner et al. 2000), and a sequencing technique such as Sanger sequencing or targeted MPS. These techniques present of number of limiting factors. Conventional metaphase karyotyping informs on genomic complexity of the whole genome. However, CLL cells are rarely in mitotic state when cultured *in vitro* and despite the development of novel techniques with B-cell mitogens, its use remains limited in routine practice (Gahrton

et al. 1980; Juliusson et al. 1990; Muthusamy et al. 2011). FISH can reliably detect chromosomal alterations on non-dividing CLL cells (Döhner et al. 2000), but only the regions of interest are targeted with probes and global genomic complexity is not detected. In addition, reporting the presence or absence of an alteration is dependent on the analyst (Smoley et al. 2010; Ellis 2000). The limitations of the sequencing technique used in routine for CLL have been detailed in Chapters 1.2.5 and 1.6.1. In brief, Sanger sequencing and targeted MPS are affordable and robust techniques, but only allow the screening of a few limited targets (Jethwa et al. 2013; Sutton et al. 2015; Vollbrecht et al. 2015; Nadeu et al. 2016; Rigolin et al. 2016; Rasi et al. 2016; D. Rossi et al. 2014; Baliakas et al. 2015). In the case of Sanger sequencing, the sensitivity of the method does not allow the detection of clinically important subclonal variants (D. Rossi et al. 2014; Landau et al. 2015; Nadeu et al. 2016). Therefore, current practice of genomic testing requires transformation in order to provide one high-throughput test exploring all the different types of genetic alterations.

One of the first prognostic markers defined in CLL was GC, found in patients with a complex karyotype (defined as more than 3 chromosomal rearrangements) (Han et al. 1984; Juliusson et al. 1985; Juliusson et al. 1990; Mayr et al. 2006; Van Den Neste et al. 2007; Haferlach et al. 2007; Kipps 2008; Ouillette, Collins, Shakhan, Li, Peres, et al. 2011; Herling et al. 2016). However, with the exception of recent work from Yu and colleagues using WES, MPS based studies fail to define a global measure of GC and therefore this biomarker requires revisiting in the context of WGS studies (Yu et al. 2016).

In recent years, genomic studies have defined new alterations with potential clinical value such as non-coding variants in *NOTCH1* 3'UTR, mutations in the enhancer region of

PAX5, as well as mutational signatures, such as signature 9 which is associated with *IgHV* mutations (Alexandrov et al. 2013; Puente et al. 2015; Kasar et al. 2015; Burns et al. 2018). Other types of signatures have been described and remain to be further explored to explain the mutational mechanisms behind the disease. Despite these few reports, the clinical implications of non-coding variants and mutational signatures are still largely unexplained.

Here, I present a comprehensive analysis of FCR non-responder patients using conventional targeted MPS in combination with SNP array, and then using WGS. CLL patients in clinical trials were investigated, comparing FCR non-responders, to FCR responders as well as other first-line treated patients and R/R patients allowing exploration of more aspects of the CLL genome.

6.2 Aims and objectives

I investigated predictive molecular markers of relapse after first-line chemo-immunotherapy using (i) the same technologies currently available in the clinic and (ii) using WGS. In addition, I (iii) validated WGS as a diagnostic tool for routine reporting of CLL patients and using this technique, I (iv) detailed the genome of 320 patients enrolled in different clinical trials.

6.3 Material and methods

6.3.1 Patients and samples

I initially investigated 224 patients using targeted sequencing and SNP array enrolled in first-line FCR-based clinical trials (ARCTIC and AdMIRe) to identify common alterations associated with early relapse. Patients were classified into two groups: 168 were FCR responders, as they did not relapse within 24 months; 56 were FCR non-responders as they relapsed or died following FCR treatment.

Then, using WGS I investigated 321 CLL patients from 6 national UK clinical trials: CLEAR (n=12), CLL210 (n=30), ARCTIC (n=61) and AdMIRe (n=64), RIAltO (n=88) and FLAIR (n=66) (Table 6.1). The 125 patients recruited in the trials ARCTIC and AdMIRe have been previously investigated with targeted sequencing and SNP array as part of the initial cohort of 224 patients.

The CLEAR trial studied the effect of using antibiotics for early CLL patients. These patients are annotated with CLL at diagnosis. In the AdMIRe and ARCTIC trials patients received frontline FCR-based treatment. These cohorts are described as FCR responders vs. FCR non-responders. The CLL210 trial included high risk CLL patients requiring treatment after relapsing and who received monoclonal antibody based treatment (alemtuzumab or ofatumumab). These patients are annotated with R/R. These four trials were phase 2 studies.

The RIAltO and FLAIR studies are current phase 3 trials. In RIAltO, CLL patients requiring treatment for the first time and who are not fit for standard first line therapy

(FCR) were included. They received ofatumumab with bendamustine or chlorambucil chemotherapy. These patients are annotated 1st-line Ofa Benda/Clb. In the FLAIR study first-line treated patients are receiving ibrutinib alone or in combination (with rituximab or venetoclax) or standard first-line FCR treatment. These patients are called 1st-line FLAIR. *IgHV* mutational status was available for a subset of 97 patients in the ARCTIC and AdMIRe patients. Minimal residual disease (MRD) was available for 104 patients and treatment response assessment for 108 patients.

Table 6.1 Patients groups

CLL group	Trial	Number of patients	Additional data
FCR responder	ARCTIC AdMIRe	82	IgHV (43 mutated vs. 54 unmutated) MRD (54 negative vs. 50 positive) treatment response assessment (77 complete remission vs. 31 partial remission or stable/progressive disease) age: 61 (33-80)
FCR non responder	ARCTIC AdMIRe	32	
R/R	CLL 210	30	none
1st-line Ofa Benda/Clb	RIAltO	88	none
1st-line FLAIR	FLAIR	66	none
CLL at diagnosis	CLEAR	12	none

FCR: Fludarabine, Cyclophosphamide, Rituximab, Ofa: Ofatumamab, Benda: bendamustine, Clb: chlorambucil

6.3.2 Extraction of genomic DNA from peripheral blood mononuclear cells and quality control

Genomic DNA samples were extracted from unsorted peripheral blood mononuclear cells using the QIAamp DNA mini kit (Qiagen), see Chapter 2.3.3. DNA QC was then

performed using a Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA) for quantification of all nucleic acids and Qubit dsDNA Broad Range Assay kit (Thermo Fisher Scientific, Waltham, MA, USA) for quantification of double stranded nucleic acids as described in methods Chapters 2.4.1 and 2.4.2.

6.3.3 MiSeq massively parallel targeted sequencing

Targeted sequencing libraries were prepared and sequenced using the CLL targeted sequencing panel as described in Chapters 2.6.2, Table 2.3 and 2.7.2 on 224 patients enrolled in first-line FCR-based clinical trials. Targeted libraries were sequenced on a MiSeq instrument (Illumina Inc., San Diego, CA, USA) by 300 bp PE sequencing as described previously (Chapter 2.7.2). The mean depth of sequencing was 2030X.

6.3.4 PCR-free whole genome sequencing

Whole genome libraries were prepared from extracted DNA samples of cohort of 320 patients as described in Chapter 2.6.1 and sequencing was performed on a HiSeq2500 High Output v4 mode, 2x125 cycles (paired end, non-indexed) (Chapter 2.7.1).

6.3.5 Single nucleotide polymorphism microarray

SNP array was performed as described in Chapter 2.10.

6.3.6 Bioinformatics pipeline

Targeted sequencing reads were aligned and variants called, filtered and annotated as described in Chapter 2.9.1.

WGS reads were aligned and variants called, filtered and annotated as described in Chapters 2.8.1.2 and 2.8.2.2.

6.3.7 Statistical analysis

The proportions of mutated and non-mutated patients, for specific alterations, were compared in different patient groups using Fisher's exact tests. Multiple testing adjustments of p-values were performed using false discovery rate (FDR) corrections. Group comparison was performed using ANOVA to compare all the groups to the total cohort, or t test to compare two independent groups. Correlations were calculated with a Spearman’s rank correlation test.

6.3.8 Individual contributions

In this Chapter, whole genome library preparation, sequencing and initial data alignment was performed by Illumina Cambridge Ltd.. Alignment metrics were calculated by Dr Alona Sosinsky. Kataegis and mutational signatures data were generated by Dr Kate Ridout.

6.4 Results

6.4.1 Genomic heterogeneity in FCR non responders

First, 224 patients enrolled in first-line FCR-based clinical trials were investigated using a clinical targeted sequencing panel of nine genes and SNP array to determine CNAs across the genome. After variants filtering and reporting, 60% of patients carried at least one mutation in the nine genes screened and 20% had more than one mutation. The patients were grouped in “FCR responders” and “FCR non-responders”. In the FCR non-responder cohort, 13 out of 56 patients had a *TP53* alteration, 13 patients had an *ATM* alteration (mutations and/or copy number deletion), 13 patients had a *SF3B1* mutation and 34 patients have an unmutated *IgHV*. The proportion of patients with specific known alterations in CLL was compared in the responder and non-responder groups. (Figure 6.1). Some alterations were significantly more detected in one of the two groups using the Fisher's exact test (FDR correction), such as 17p deletions (p=0.0022), *SAMHD1* mutations (p=0.042) and *TP53* mutations (p=0.00018). These alterations were enriched in the FCR non-responders. In addition, a trend towards enrichment of *IgHV* mutated patients in the FCR responder (p=0.091) and an enrichment of 20q deletions in FCR non-responders (p=0.0615) were observed.

Some of these alterations tended to occur in patients with short time to progression (TTP) (Figure 6.2). It was the case for alterations such as 17p deletions (23% of alterations in patients with TTP < 24 months, vs. 8% in patient with TTP> 24 months), but also *TP53*

mutations (33% vs. 12%), *SAMHD1* mutations (17% vs. 0%) and presence of other CNV alterations (83% vs. 50%). On the other hand, *SF3B1* mutations were detected in 38% of patients with TTP > 24 months vs. 10% of patients with TTP < 24 months, as well as trisomy, enriched in patients with later TTP (27% vs. 10%). However, due to low number, these observations could not be confirmed by statistical tests and require additional samples to be validated.

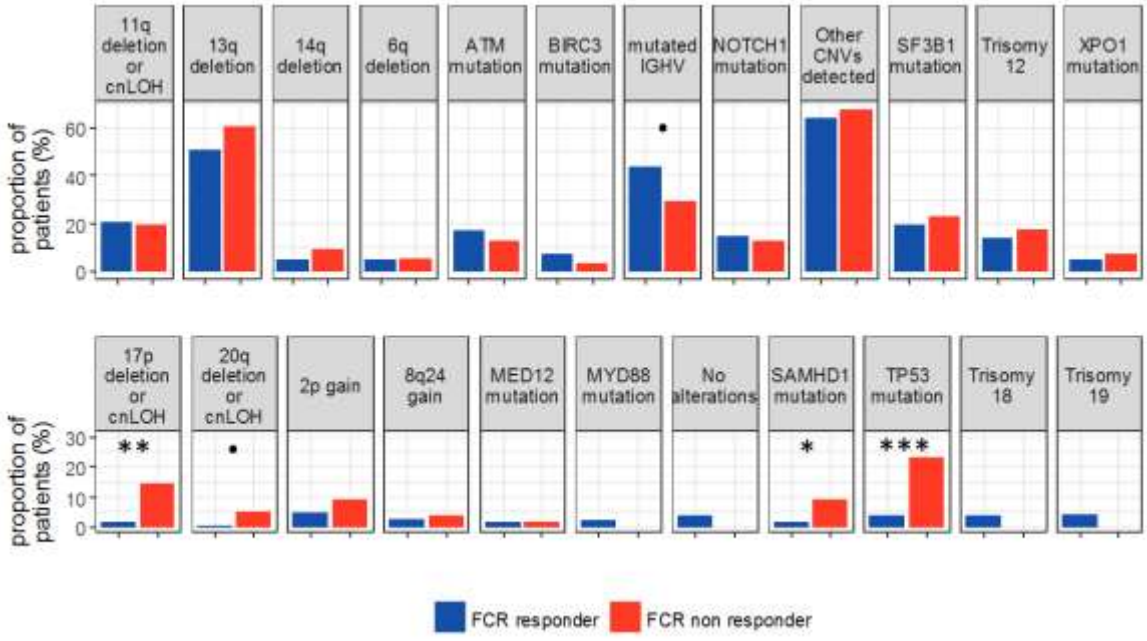


Figure 6.1 Proportion of patients carrying copy number changes, SNVs or indels in known CLL genes/cytobands.

The cohort is divided in FCR responders (patients who did not relapse within the first 2 years of treatment) and FCR non responders (relapsing patients). The statistical test used was a Fisher's exact test p-value codes were as follow: '***' <0.001; '**' <0.01; '*' <0.05; '.' < 0.1.

cnLOH: copy neutral loss of heterozygosity. Other CNVs detected: other than all the other ones detailed on the plot.

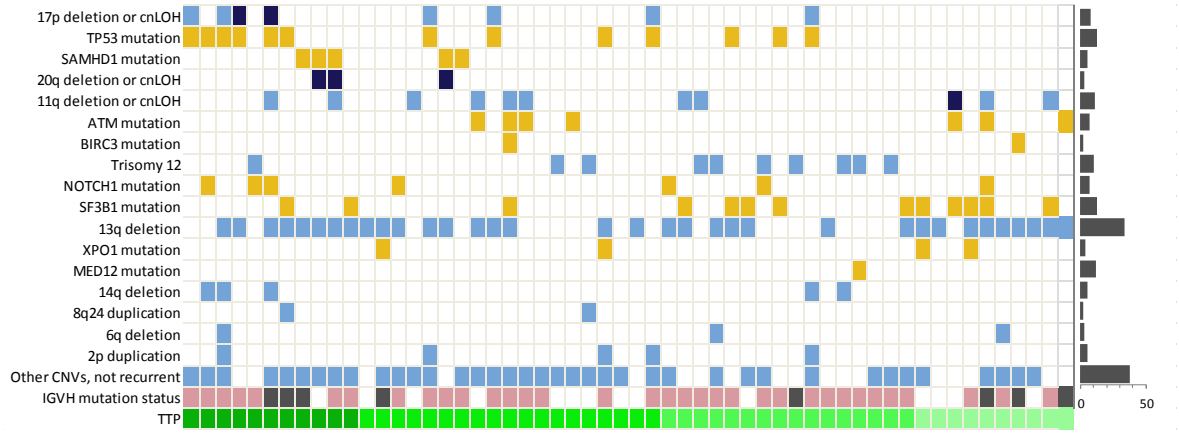


Figure 6.2 Distributions of alterations in FCR non responder patients (n=56).

Each column represents the distribution of mutation in a patient. Yellow cells characterise gene mutations, light blue cells denote copy number changes, dark blue ones represent cnLOH, pink ones show unmutated *IgHV* and dark grey ones show missing data. The four shades of green represent < 12 months, 12-24 months, 24-36 months, >36 months.

cnLOH: copy neutral loss of heterozygosity, TTP: time to progression.

In this cohort, patients carrying a *TP53* or *SAMHD1* mutation or a 17p or 20q deletion were enriched in FCR non responders and seemed to be associated with early relapse. Therefore, these patients are unlikely to obtain durable disease control from FCR treatment. However, only 18 out of 56 patients in this cohort carried one of these alterations and further studies are required to reveal other alterations significantly enriched in the FCR non-responder cohort.

6.4.2 Acceptable alignment metrics in tumour samples and lower alignment rate in germline samples

Subsequently, WGS was performed on a subset of the FCR treated patients (n=125), as well as other patients that can be classified as low risk and high risk patients (n=196), bringing the total to 321 CLL genomes.

After the generation of WGS data for tumour samples from PBMCs and GL samples from saliva, the sequencing depth was confirmed to be within range of the aimed depth: 90X for tumour and 30X for GL samples. The coverage summary data was calculated using bwtool. The median depths of coverage were 94.2X (67.9-144.4) for tumour and 28.5X (20.5-55.8) for GL samples (Figure 6.3).

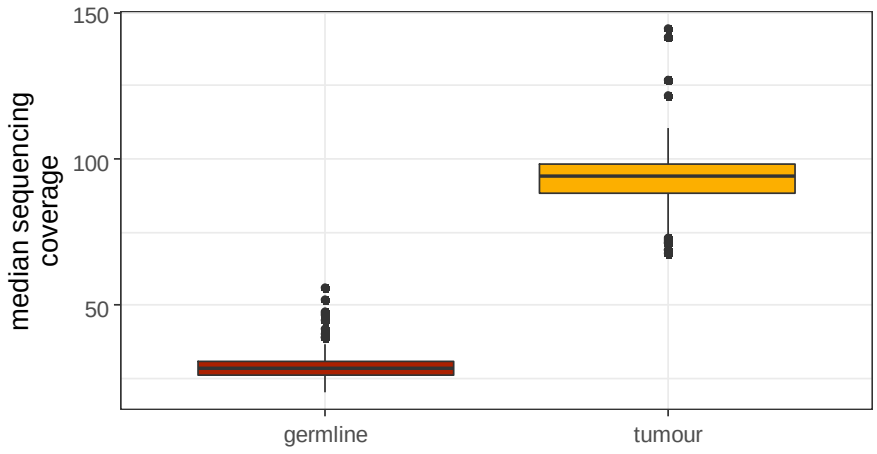


Figure 6.3 Median sequencing coverage for 320 germline samples from saliva and 320 tumour samples from peripheral blood

Other sequencing metrics were all found to be within range of acceptable values: dropout < 5%, percentage of chimeric pairs < 1% and fragment size > 450 bp (Table 6.2). However, the proportion of reads aligned to the reference genome was found suboptimal in the GL sample: it was significantly lower than in tumour samples (median of 88.0% vs. 95.2%, Wilcoxon rank sum test $p<2.2e^{-16}$).

Table 6.2 Distribution of WGS alignment metrics in tumour and germline samples

	tumour samples	germline samples
	median (min-max)	
GC dropout (%)	3.9 (-0.65-8.0)	2.3 (-5.7-7.5)
AT dropout (%)	3.3 (0.6-6.7)	4.6 (1.3-9.5)
Chimeric pair (%)	0.3 (0.2-1.5)	0.3 (0.1-1.1)
size fragment mean (bp)	484 (418-595)	466 (397-579)
Mapping rate (%)	95.3 (93.5-96.2)	86.3 (43.1-94.6)

The distribution of the proportion of reads aligned to the reference human genome showed that 75 out of 321 GL samples with sub-optimal data (mapping rate < 80%) and that nine presented an extreme low rate (<60%) (Figure 6.4).

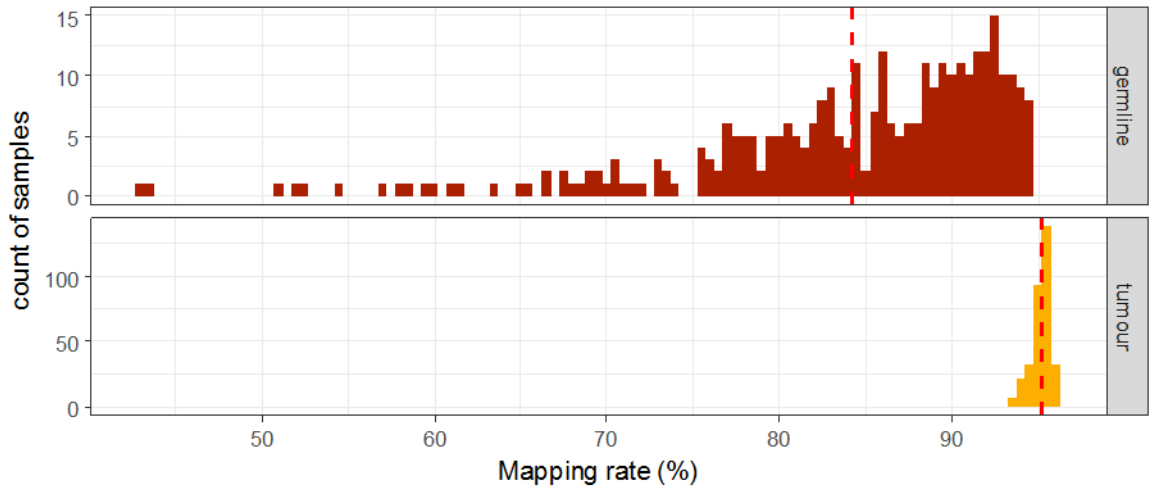


Figure 6.4 Histogram showing the distribution of the proportion of reads aligned to the reference human genome (mapping rate (%)) of GL and tumour samples.

The red vertical dotted line represents the median value.

The number of variants pre and post filtering was counted. A median of 364,043 SNVs, 1045 indels, 114 CNAs were detected per patient in unfiltered variants lists. After filtering, a median of 9172 SNVs, 2348 indels, 8 CNAs were detected per patient.

6.4.3 A subset of somatic variants was validated by targeted sequencing

A subset of somatic variants was validated using a targeted sequencing panel (described in Chapter 2.6.2.1) in 64 samples. In addition, another targeted panel including the more recently discovered mutated gene *RPS15* was used to validate samples with a mutation in this gene. Overall, 79 variants were validated in 9 genes and a high positive correlation of the variants VAFs were observed between the two techniques (r=0.963, p=1.29e^{-45}, Spearman's rank correlation, Figure 6.5). This high positive correlation was also observed for somatic variants with low VAFs (less than 0.3).

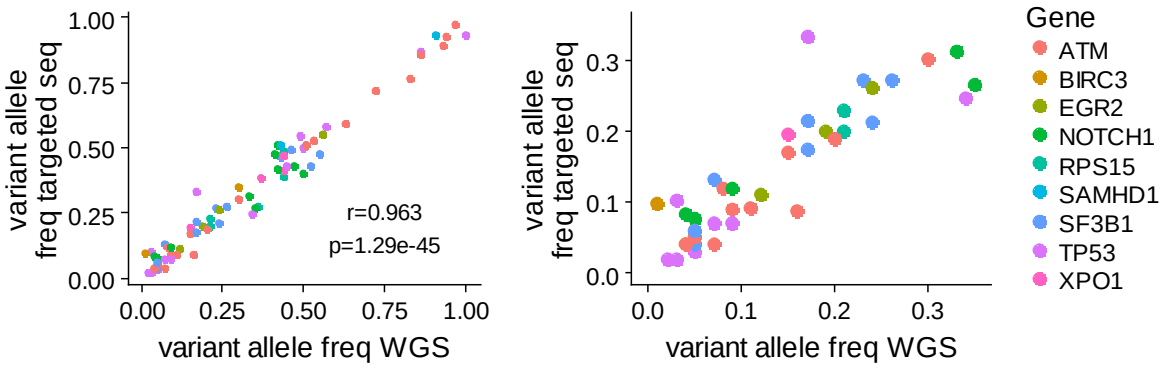


Figure 6.5 Correlation between variant allelic fractions of mutations detected by targeted MPS and WGS for all variants investigated (left, n=79) and low VAF variants only (right, n=43).

The plot on the left shows the correlation coefficient and p value of the Spearman's rank correlation. WGS: whole genome sequencing.

A subset of well-defined CNAs was validated in the same cohort of patients (Figure 6.6) using two techniques: FISH (provided by the Haematological Malignancy Diagnostic Service in Leeds) and SNP array. Using FISH, a concordance of 83% with WGS was found by examining deletions of 11q, 13q, 17p and trisomy 12. SNP array and WGS data were concordant at 93%.

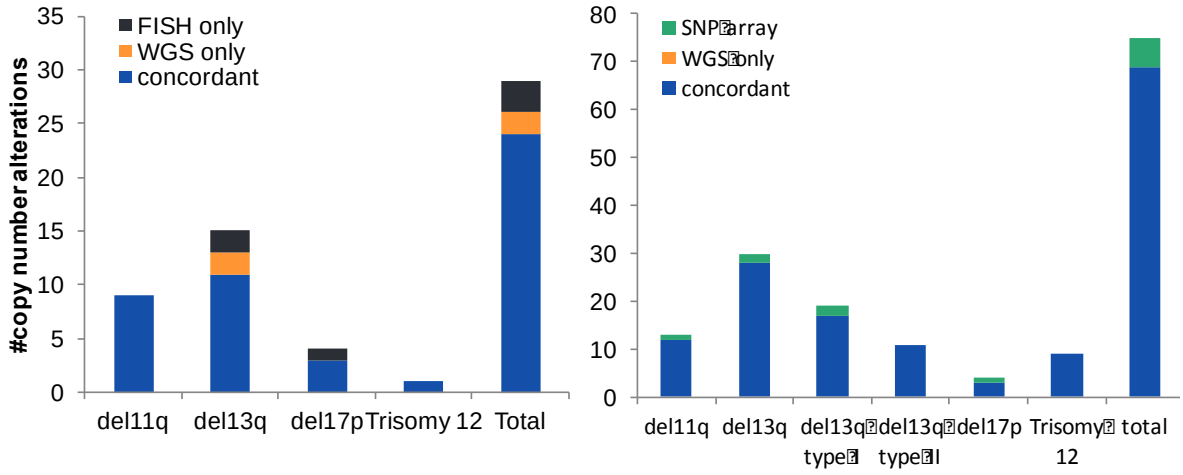


Figure 6.6 Copy number alterations validation by FISH and by SNP array

CNA concordance between WGS and FISH (left panel) and WGS and SNP array (right panel). FISH: fluorescence *in situ* hybridisation, WGS: whole genome sequencing, del: deletion

Overall WGS was validated against current practice of the laboratory and more novel genomic features were explored next.

6.4.4 Genomic landscape of pre-treatment and relapse refractory CLL

Multiple genes have been previously reported recurrently mutated in cohorts of CLL patients (Figure 6.7). In this cohort of 320 patients, I have studied the presence of functional mutations in recurrently mutated genes. The top mutated genes were *SF3B1* (17% of patients), *TP53* (13%), *NOTCH1* (12%), *IGLL5* (12%), and *ATM* (11%).

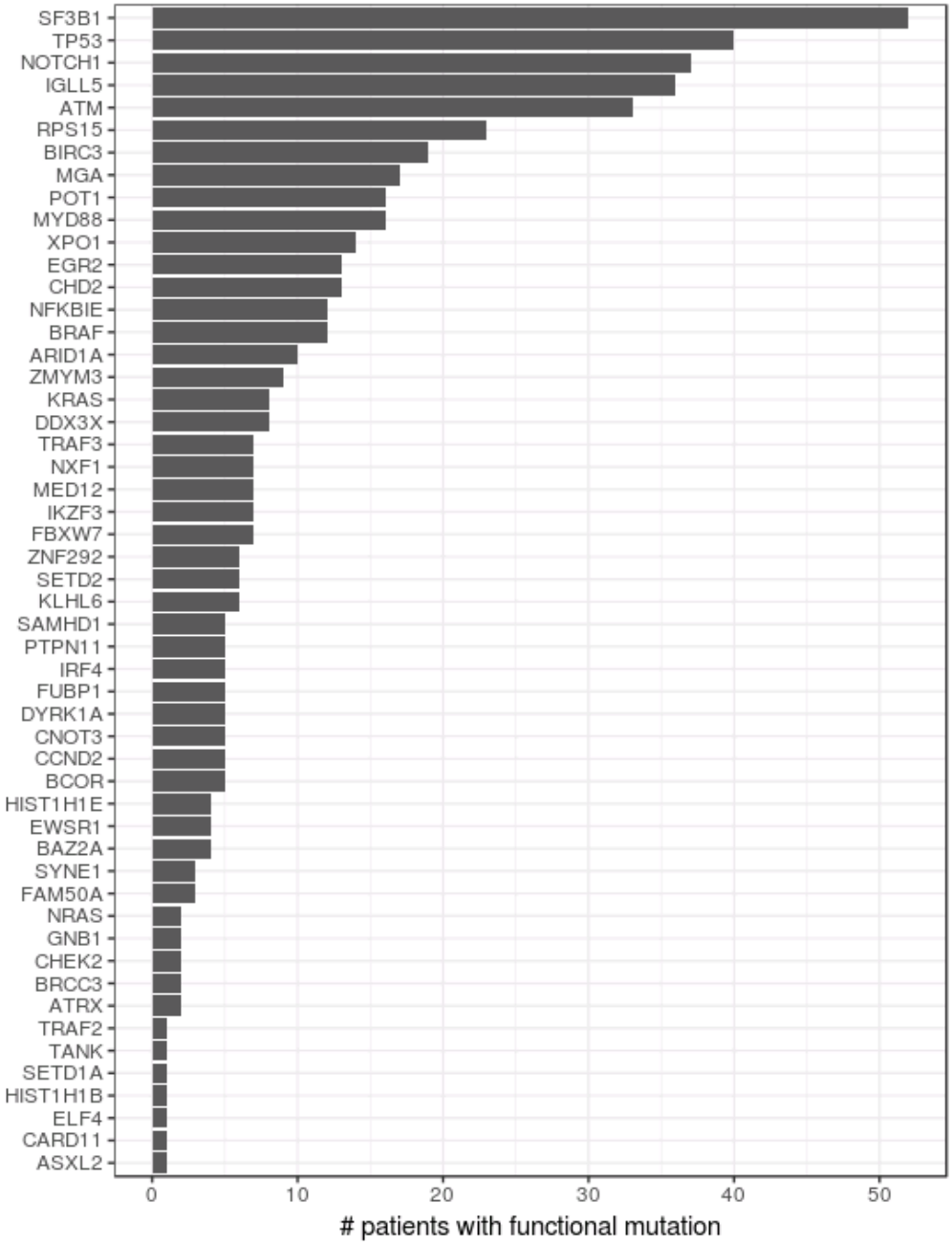


Figure 6.7 Number of patients with functional mutation

Functional mutations: mutation in exons with one of the following consequence: missense, stop, frameshift and in-frame indels.

The distributions of the 10 most recurrently mutated genes (in more than 5% of the cohort) were compared in different patient groups. However, except for *TP53*, enriched in R/R

patients ($p=3.7.0e^{-10}$, Fisher exact test), none of the other genes were enriched in a particular group Figure 6.8)

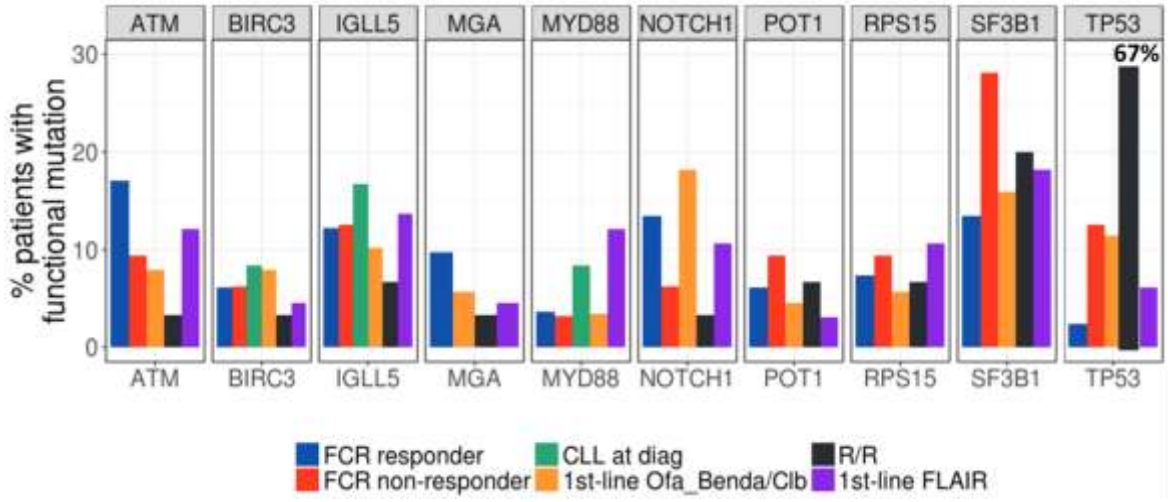


Figure 6.8 Proportion of patients with functional mutation in different groups

The cohort is divided in FCR responders (patients who did not relapse within the first 2 years of treatment) and FCR non-responders (relapsing patients). FCR: Fludarabine, Cyclophosphamide, Rituximab, Ofa: Ofatumamab, Benda: bendamustine, Clb: chlorambucil

When comparing patients with complete response and patients with partial response or stable/progressive disease only *TP53* was enriched in the second group (1.3% vs. 13.0%, p=0.023, Fisher's exact test).

The number of CNAs, namely gains and losses were compared in the different groups (Figure 6.9). Amplifications of the p arm of chromosome 2 were enriched in FCR non-responders and R/R patients, without reaching statistical significance. In addition, as expected the number of de17p in the R/R cohort was enriched in the R/R group ($p=1.7.0e^{-08}$, Fisher exact test).

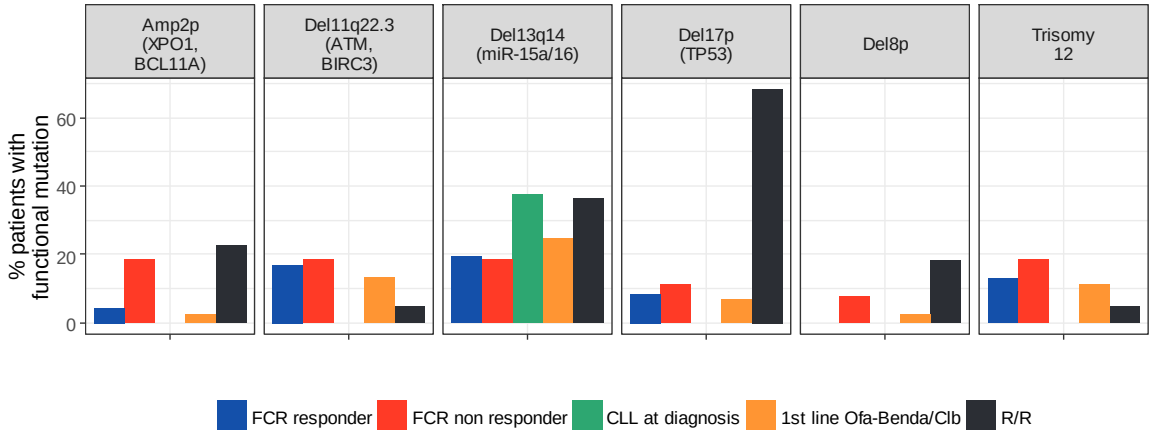


Figure 6.9 Number of patient with CNAs in different groups

The cohort is divided in FCR responders (patients who did not relapse within the first 2 years of treatment) and FCR non-responders (relapsing patients). FCR: Fludarabine, Cyclophosphamide, Rituximab, Ofa: Ofatumamab, Benda: bendamustine, Clb: chlorambucil

6.4.5 Operative mutational signatures

Mutational signatures were studied in the cohort of patients using signatures previously defined (Alexandrov et al. 2013) (see Chapter 6.3.6). First, the signatures were computed in each group of patients separately and six mutational signatures were identified: Signatures 1, 3, 5, 9, 12 and 20 (Figure 6.10).

Signatures 5, 9, 12 and 20 were different in some groups, especially in the CLL at diagnosis (CLEAR patients). There was no significant difference between the groups for signature 1 and signature 3. Signature 1 has been correlated with age of cancer diagnosis (Alexandrov et al. 2013), however this signature was not found enriched in the patients of the 1st-line Ofa Benda/Clb group (group of patients generally older (Lymphoma

Association 2017)). In addition, signature 1 was not correlated with age in this cohort (r=-0.09, p=0.45, Spearman correlation).

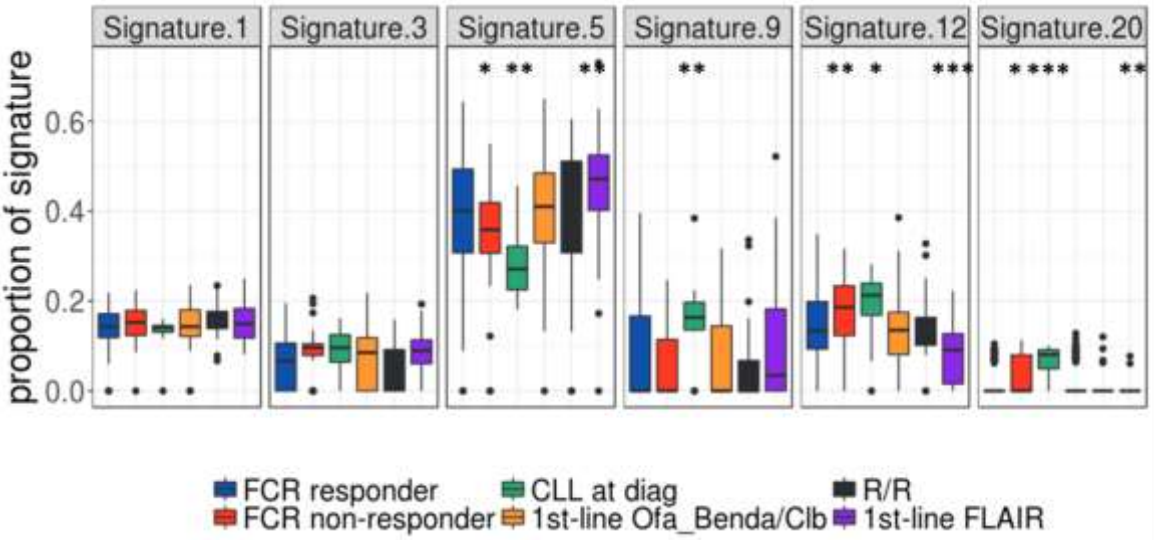


Figure 6.10 Proportion of the major mutational signatures in six CLL groups

The cohort is divided in FCR responders (patients who did not relapse within the first 2 years of treatment) and FCR non-responders (relapsing patients). The statistical test used was ANOVA test comparing each group to the total cohort, p-value codes were as follow: '***' <0.001; '**' <0.01; '*' <0.05. FCR: Fludarabine, Cyclophosphamide, Rituximab, Ofa: Ofatumamab, Benda: bendamustine, Clb: chlorambucil

However when comparing only FCR responders vs. non responders signature 3 was enriched in non-responders (6.7% [0-19.6] vs 9.6% [0-20.6], p=0.008, t-test). Signature 3 was reported associated with germline and somatic BRCA1 and BRCA2 mutations in other cancers (Nik-Zainal, Alexandrov, et al. 2012; Alexandrov et al. 2013).

No differences in signatures were observed in the MRD positive vs MRD negative groups and between the complete response vs the partial response or stable or progressive disease (Appendix K).

6.4.6 Structural alteration as a measurement of genomic complexity

CNAs and SVs were explored to identify patients with GC. In the cohort, a median of 14.9 Mb of losses and 2.4 Mb of gains per patients were reported. In addition, other SVs were observed: a median of five segmental duplications per patient, 15 inversions and 108 translocations were counted. The different SVs were compared in the different groups (Figure 6.11). In the R/R group, more gains and losses were reported. In addition, more gains were reported in the FCR non-responders compared to the responders, without reaching statistical significance.

Given these results, it seems more likely that GC should be defined by using CNAs as opposed to other SVs. By defining GC using the cumulative length of CNAs, we found a median length of alteration of 26.8 Mb of alteration in the whole cohort.

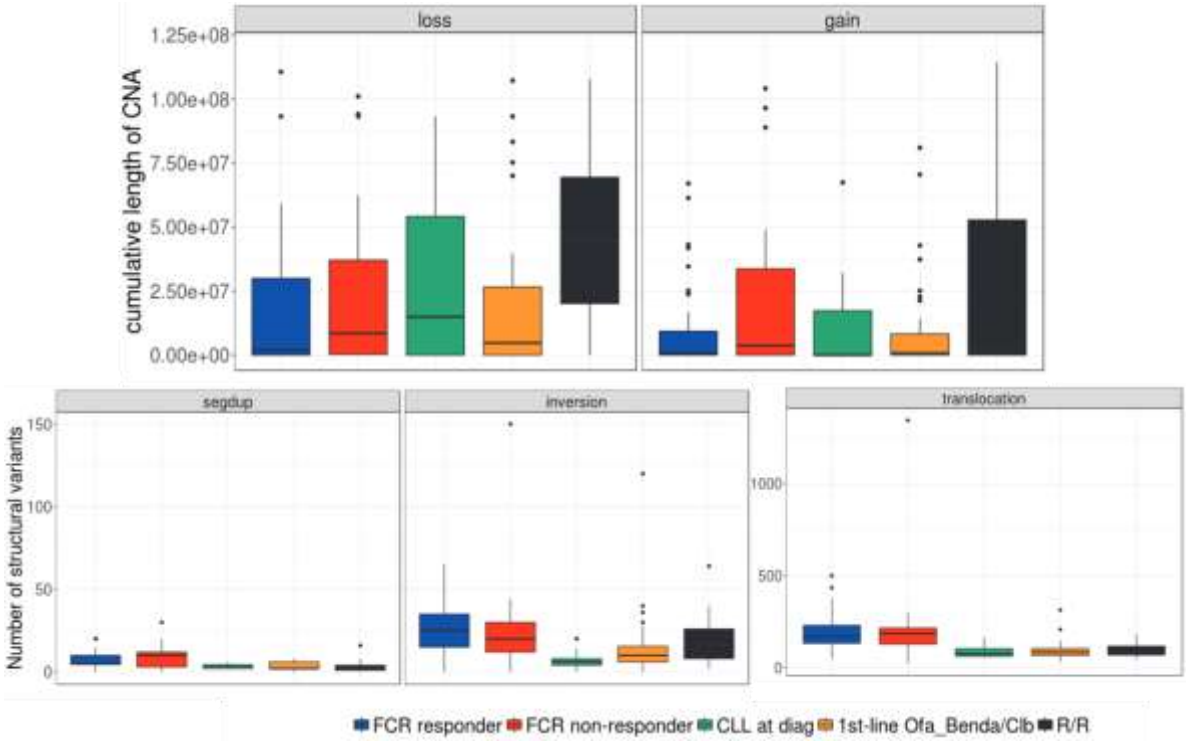


Figure 6.11Structural variants investigation in the different groups

Cumulative copy number losses and gains in each patient (top) and number of structural variants in each patient (bottom)

Loss: copy number loss, gain: copy number gain, segdup: segmental duplication.

6.4.7 Exploration of *IgHV* mutational status and genomic features

The *IgHV* sequence of approximately 400 bp is highly variable, containing multiple base substitutions, insertions and deletions and therefore it is generally not aligned to the reference genome by mappers, especially with WGS data generated from short Illumina reads (van Dongen et al. 2003). Using genomic measures, the *IgHV* mutational status was explored in the cohort.

The somatic mutations detected in *IgHV* mutated (n=43) and unmutated (n=54) highlighted that the gene *RPS15* was enriched in patients with *IgHV* unmutated status (in eight patients vs none, p=0.0082, Fisher's exact test). No other genes were enriched in these groups.

The mutational signatures of the 97 patients with *IgHV* mutational status information were studied (Figure 6.12). In these two groups, signature 3 and signature 20 were not different. Signatures 1 and 9 were particularly different with an enrichment of signature 1 in unmutated patients and an enrichment of signature 9 in unmutated patients. Signature 9 is associated with AID-mediated IGHV mutation and linked to non-canonical AID activity (Alexandrov et al. 2013; Kasar et al. 2015).

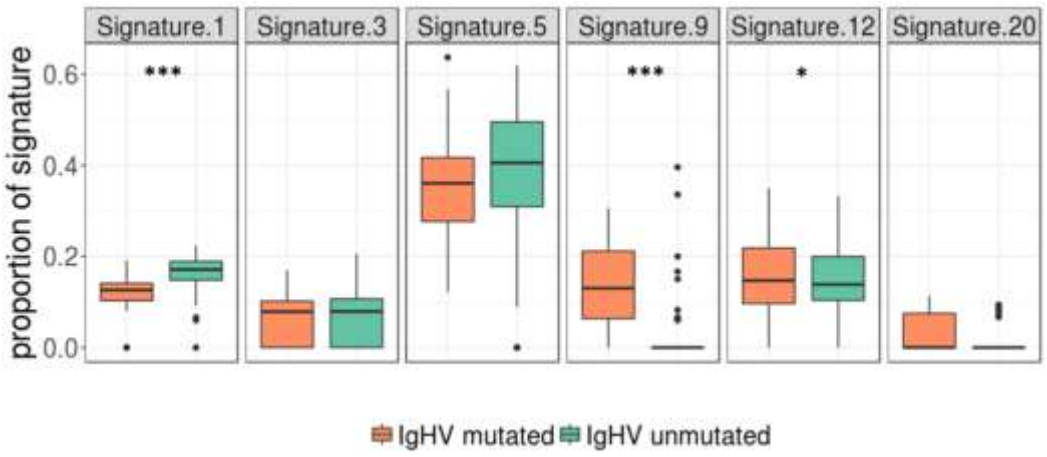


Figure 6.12 Proportion of the major mutational signatures in *IgHV* mutated vs. unmutated

The ARCTIC and AdMIRe cohort is divided in *IgHV* mutated vs. unmutated. The statistical test used was a t test comparing the two groups for each signature, p-value codes were as follow: '***' <0.001; '**' <0.01; '*' <0.05.

In addition, the *IgHV* mutated and unmutated groups were investigated further for signature 9 (Figure 6.13).

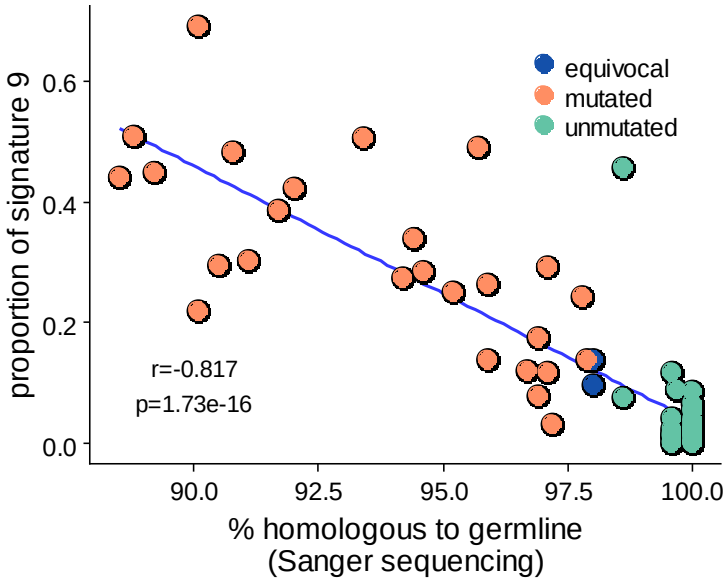


Figure 6.13 Correlation between the proportion of signature 9 and percentage of sequence homologous to germline (defined by Sanger sequencing)

Correlation statistics were calculated using a Spearman correlation. Colour codes were defined by the Sanger sequencing results: < 0.98 is mutated, >0.98 is unmutated and 0.98 is equivocal.

A negative correlation was observed between the proportion of signature 9 and the percentage homologous to germline (r=-0.817, $p=1.73e^{-16}$). To the exception of one sample, the rest of the unmutated samples presented a proportion of signature 9 below 12%.

In addition, the regions of localised hypermutations or "kataegis" were investigated in the cohort (Figure 6.14). The most highly mutated regions were *IGLL5* as well as the immunoglobulin (Ig) loci on chromosomes 2 (*IgK*), 14 (*IgH*) and 22 (*IgL*).

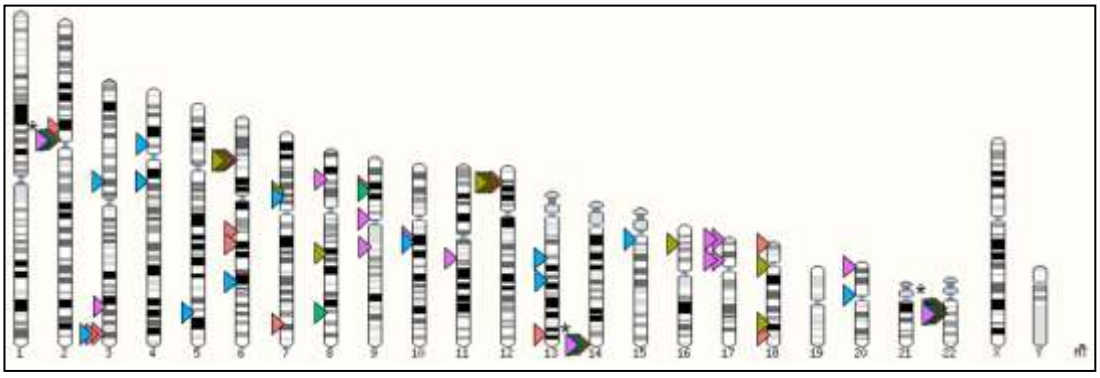


Figure 6.14 Putative regions of Kataegis per patient and per group.

* indicates Immunoglobulin loci on chromosomes 2, 14 and 22.

In these regions, the density of SNVs in these regions was found different in mutated and unmutated patients (Figure 6.15). Many peaks observed in these genomic regions are generally found higher in the *IgHV* mutated patients than in the unmutated patients.

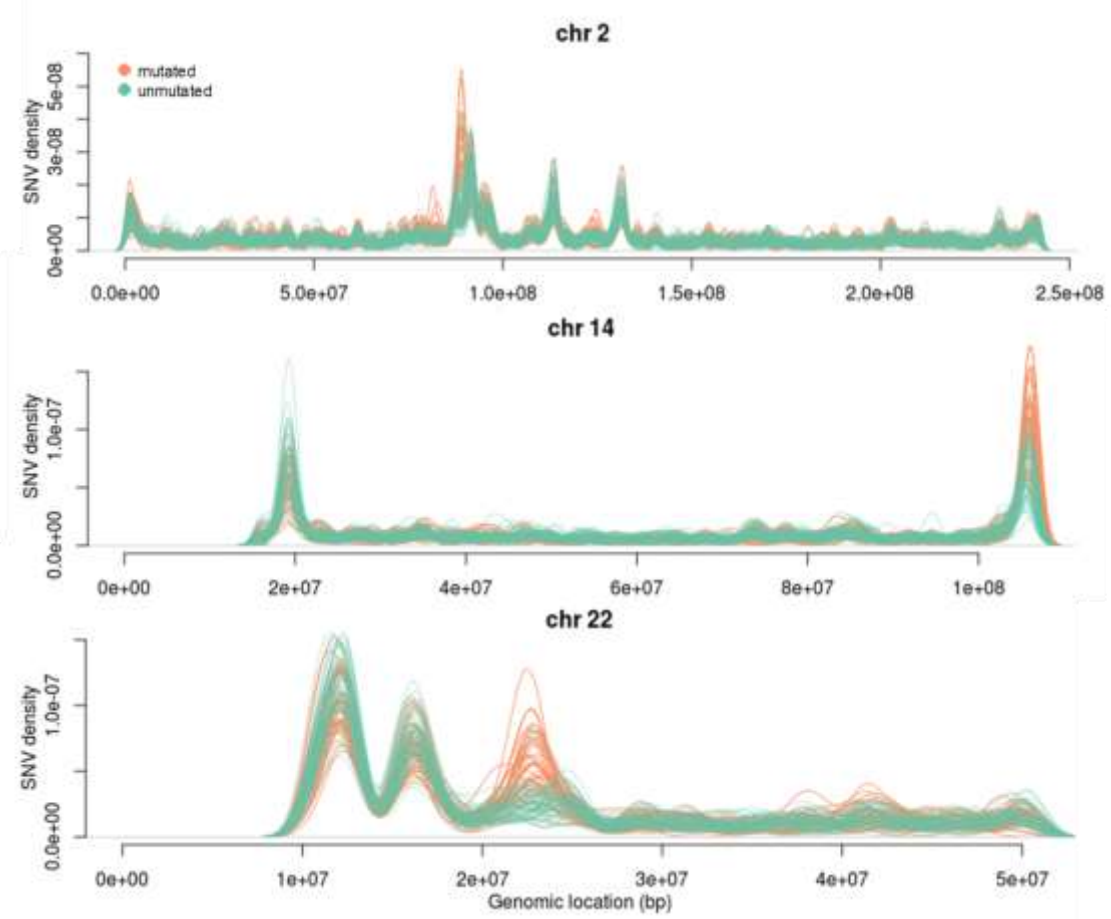


Figure 6.15 SNV density in genomic regions of the Ig loci, described regions of kataegis

6.5 Limitation of the study

In this study, several germline samples from saliva presented a low mapping rate to the human reference genome. This was certainly due to the presence of non-human DNA contamination in the sample, as often observed in DNA isolated from saliva samples (bacteria, fungi or food) (Nunes et al. 2012; Sun & Reichenberger 2014; Durdiaková et al.

2012). The proportion of contamination has been previously reported up to 63% in one study (Nishita et al. 2009). Although previous studies, mainly performing genotyping did not report that the saliva DNA quality led to poorer data (Sun & Reichenberger 2014; Durdiaková et al. 2012), this has not been tested in WGS data.

One of the main limitations of this study is the low number of patients in each group studied (for instance 32 patients in the FCR non-responder group and 30 patients in the R/R group). These numbers makes any statistical comparison of infrequently occurring mutations very challenging, especially in exploring non-coding regions of the genome. More global measures of the genome would provide continuous variables to each patient instead of binary variables (mutated vs unmutated) and improve the power of the analysis.

6.6 Discussion

6.6.1 Investigation of FCR responder predictors by targeted sequencing and SNP array

In the cohort of 224 patients enrolled in first-line FCR-based clinical trials, two genes were found enriched in non-responders: *TP53* and *SAMHD1*. *TP53* has been known for a long time to be associated with early relapse following FCR therapy (Stilgenbauer et al. 2009; Hallek et al. 2010). Alterations of *SAMHD1* lead to abnormal dNTPs pools and abnormal cell cycles, which can promote cancer (Li et al. 2017). *SAMHD1* mutations have been linked to chemo-refractoriness in CLL and response to DNA damage (Clifford et al. 2014) and have been reported in other studies (Landau et al. 2013; Landau et al. 2015).

The results in this study link, for the first time, early relapse after FCR treatment and *SAMHD1* alterations. *ATM* alterations were not found enriched in non-responders, confirming previous results that patients with *ATM* alterations (including del11q) benefit from the addition of Rituximab in first-line therapy (Stilgenbauer et al. 2014).

Both *TP53* and *SAMHD1* were found mutated by targeted sequencing and deleted by SNP array. This result highlighted the necessity of using an integrated genomic approach in order to identify all possible patients with a response predictor.

6.6.2 Investigation of more predictive markers in CLL using WGS

WGS is another integrated approach allowing the investigation of both mutations (SNVs and indels) and CNAs. In this work, WGS data was compared to the targeted sequencing data as well as SNP array and FISH data. When evaluating SNVs and indels, most of these were detected by WGS, even at low VAF (including low subclones of TP53 mutations down to 2%). This result showed an improvement compared to previous studies (Rehm 2013; Frampton et al. 2013). With the addition of the CNA results, this analysis showed that WGS provides accurate detection of at least the clinically relevant alterations in CLL. Previous studies have reported in total 254 CLL genomes: 46 patients in the most recent study by Burns and colleagues (Burns et al. 2018), 30 patients reported by Kasar and colleagues (Kasar et al. 2015) and 150 patients in Puente's work (Puente et al. 2015). Therefore this study presented the largest cohort of CLL genomes reported so far.

As the targeted sequencing approach combined with SNP array only allowed the detection of 8% of FCR non-responders, the data generated from WGS was explored in order to discover additional genomic biomarkers.

In this cohort the proportions of patients mutated in *SF3B1*, *NOTCH1* and *ATM* were generally higher than a cohort of early stage CLL patients in a study by Puente and colleague (Puente et al. 2015) and similar to the first-line FCR treated CLL8 cohort (Landau et al. 2015). This is consistent with previous observations that more mutations are acquired with disease progression.

When studying mutational signatures previously described, signature 3, associated with germline and somatic BRCA1 and BRCA2 mutations (Nik-Zainal, Alexandrov, et al. 2012; Alexandrov et al. 2013), was enriched in the FCR non-responder cohort. However, this observation was not confirmed when the patients were grouped according to MRD or response assessment to therapy. In solid tumours, multiple bi-allelic DDR pathway defects have been reported and are targeted by novel therapies (Pritchard et al. 2016; Stover et al. 2016). This could benefit CLL patients affected by such signature. Therefore, novel therapies developed for other cancers could be considered in pre-clinical development for CLL.

Until recently, SNP arrays were considered the gold standard for genome-wide CNA analysis of cancer genomes, but recent studies have used WGS data for copy number profiling with fine resolution, along with chromosomal translocations analysis (Burns et al. 2018; Puente et al. 2015). In this study, we have found an enrichment of CNAs in the R/R cohort, confirming the presence of higher GC at relapse. A threshold needs to be defined to further describe GC by WGS. Recent work reported that patients with complex karyotypes were more likely to relapse after therapy with the BCL2 inhibitor Venetoclax and the Bruton's tyrosine kinase inhibitor Ibrutinib, highlighting the need to accurately

identify patients having GC regardless of the treatment received (CIT or more recent targeted therapies) (Thompson et al. 2015; Anderson et al. 2017)

6.6.3 Surrogate marker to *IgHV* status

In order to overcome the challenges of MPS data assembly at the Ig loci, surrogate genomic markers were explored using WGS data. Mutational signature 9 was found enriched in *IgHV* mutated patients. This signature is associated with non-canonical AID activity (Alexandrov et al. 2013). In addition, kataegis analysis can help to differentiate patients with mutated or unmutated status by measuring the SNV density at the Ig loci on chromosomes 2, 14 and 22. These results should be combined in a statistical model in order to find a refined surrogate marker to *IgHV* status.

More recently, novel mutation signatures linked to canonical AID activity has been described and could refine further the model (Kasar et al. 2015). New mutational signatures could be explored in this dataset in order to understand the mutational mechanisms in CLL.

6.7 Conclusions

The exploration of the initial cohort of 224 patients identified *TP53* and *SAMHD1* alterations as predictive markers to early relapse after FCR treatment. After validation of WGS in the same cohort, I explored new biomarkers (such as mutational signatures) and reconsidered already defined ones such as GC using this technique. WGS allowed a comprehensive study of the CLL genome using only one reliable test. I compared the

genome of CLL patients at different stages of the disease in a systematic way. The comparison of FCR responders and non-responders identified mutational signature 3 as enriched in non-responders. The comparison of *IgHV* mutated and unmutated patients highlighted several parameters that, once combined in a sophisticated model, could provide a surrogate marker to the mutational status tested by Sanger sequencing. In conclusion, this cohort represents the largest cohort of WGS CLL compared to all previous studies together, and as all patients were enrolled in a clinical trial the potential of biomarker's exploration remains open.

Chapter 7 Discussion

7.1 Introduction

Multiple cancer initiation and evolution theories have been described, but the SMT, as step-wise acquisition of mutations due to genomic instability, remains the basis or most genomic research (Boveri 2008; Loeb 1991). This genomic instability and the presence of recurrently mutated genes have led to way to precision medicine. Over the past decade MPS has really broaden our understanding of the mechanism of the origin of cancer as well as its progression (Creighton et al. 2013; Getz et al. 2013; Hammerman et al. 2012; Muzny et al. 2012; Puente et al. 2011). However, most large scale studies remain research-led and their findings are not straightforwardly applied in the clinic. Despite the reduction in cost of genomics technology, multiple challenges persist to apply molecular precision diagnostics for cancer patients. It has been anticipated that a reduction of the pre-analytical challenges as well as the characterization of an increasing number of cancer genome may allow the use of genomics technologies in routine cancer diagnostics.

The key findings described in the preceding chapters lay the foundation and pave the way for us to start considering using MPS, and in particular WGS in the clinic.

7.2 Summary of the findings

In this thesis we have particularly studied two main challenges preventing MPS technologies to be applied in the clinic: the use of FFPE specimen for clinical WGS reporting and the lack of prognostic and predictive biomarkers in CLL patients.

In **0**, the feasibility of reporting clinical somatic variants from WGS using FFPE samples was assessed by comparing the data using matching FF samples. This work confirmed the difficulty of collecting FF samples in routine diagnostics and the necessity of using FFPE samples. In addition, after thorough filtering of the data, somatic SNVs were proven to be confidently called from FFPE including in selected non-coding regions of the genome (defined as GIAB regions). Difference between FF and FFPE somatic SNVs were partially due to intra-tumour heterogeneity or sampling heterogeneity. CNAs detection was found sub-optimal in FFPE specimen and poorly correlated with FF data due to non-uniform coverage. This observation motivated the questions explored in Chapter 4.

In **Chapter 4**, I explored the different steps of FFPE preparation and DNA extraction to identify the cause of poor CNAs detection in WGS data. Using real-life cancer samples from patient included in a prospective study, DNA extraction was optimised leading to confident identification of somatic variants including CNAs. In addition, after exploration of the different steps to prepare FFPE samples, a new complete protocol was proposed to use FFPE specimen for clinical WGS reports.

In **Chapter 5**, a MPS targeted panel was designed to explore the genomic landscape of 114 R/R CLL patients and identify biomarkers of poor outcome. Associations between recurrently mutated genes defined the MH profile in 22 patients. This profile was

associated with poor response to therapy and shorter survival. Co-occurrence in mutated genes allowed to explore the subclonal composition and to infer the evolution of these clones, highlighting several defined patterns. The presence of mutations in *TP53* was confirmed to be a late event, whereas *ATM* and *SF3B1* were early events.

In **Chapter 6**, 224 first-line FCR-treated patients were explored using the same targeted panel as used in the previous chapter as well as SNP array. The identification of biomarkers by these techniques remained limited to the identification of *TP53* and *SAMHD1* as predictive markers in 18 out of 56 patients. Therefore, this cohort and additional patients were studied by WGS. The 320 patients were recruited in 6 different clinical trials and other potential biomarkers were identified, such as mutational signature 3. In addition, surrogate markers to IgHV mutational status were explored such as mutational signature 9 and regions of hypermutations located in the Ig loci.

7.3 The use of MPS technology in clinical practice

With the development of targeted therapies, the demand for MPS technology in clinical practice has increased dramatically (Deans et al. 2017). However, the necessity to use FFPE specimens has delayed the application of large scale technologies such as WGS for solid tumours. Since the start of the work presented in 0 and 4, very few cancer diagnostic laboratories have developed methods to screen for cancer genomes using FFPE samples. One of the most significant has been the Exome Cancer Test v1.0, the first Health-approved WES-based test for precision cancer care in the state of New-York and has been using FFPE samples (Rennert et al. 2016). However, in their work, the authors showed

that approximately half of the FFPE samples presented low quality DNA and that 1µg was necessary to investigate SNVs and CNAs by WES. In other recent studies in clinical laboratories, MPS cancer panels remain the preferred choice to investigate clinically relevant genes as PCR-based assay generate results of superior quality when the starting material is FFPE (Uzilov et al. 2016; Hamblin et al. 2017; Radovich et al. 2016; Sholl et al. 2016). Any other clinical initiative providing a more integrated approach was performed on FF samples (Uzilov et al. 2016).

In CLL, although technical challenges are less important, the implementation of MPS technique in the clinic remains limited to targeted panels as well. Most reports so far described sequencing panels of 9-20 genes (Sutton et al. 2015; Vollbrecht et al. 2015; Nadeu et al. 2016; Rigolin et al. 2016). As *TP53* is the only biomarker screened for in the clinic, the sequencing of this gene is offered as a service in routine diagnostics using MPS (Oxford Molecular Diagnostics centre 2017). WGS offered as a service in routine is still lacking and remains to date a research tool for variant discovery. However, I have shown in this thesis that in addition to variant discovery, WGS can replace several molecular tests and describe multiple aspect of the genome particularly described in Chapter 6 and also in Chapter 5 such as SVs, mutational signatures, GC, and subclonal evolution that can be of clinical significance and that solid tumours can also be screened using FFPE samples with adequate optimisations.

7.4 Intra- and inter-tumour heterogeneity in cancer

Generally, cancers are heterogeneous and carry multiple subclonal mutations acquired during the course of disease progression. In Chapter 3, we have discussed the presence of intra-tumour heterogeneity through the concept of variants with low allelic fractions that are truly different between FF and FFPE tissue specimens and which was one of the cause of differences observed in paired FF and FFPE samples. Other studies have reported such differences in term of subclonal variant composition by sequencing different part of a same tumour or metastasis of a tumour (Gundem et al. 2015; Nik-Zainal, Van Loo, et al. 2012; Yates et al. 2015; Hong et al. 2015; Suzuki et al. 2017). These observations lead to reasonable doubts that one biopsy is sufficient to provide a clinical report of variants. As the sample studied only represents the subclonal composition of that part of the tumour, other sublcones carrying mutations of potential importance might be missed and lead to the development of resistance mechanism (Landau et al. 2013). However, in Chapter 3 I have shown that there is very little difference between two sample of the same tumour (FF and FFPE samples) in term of functional variants in coding regions of cancer genes. This indicates that many of the driver genes are well established in the cancer cells and not small subclones.

CLL is another cancer exhibiting intra-tumour heterogeneity. However, to study this heterogeneity, instead of spatial, most studies described the longitudinal subclonal evolution (Knight et al. 2012; Schuh et al. 2012; Landau et al. 2015; Zhao et al. 2016; Rose-Zerilli et al. 2016). In Chapter 5, the possible order of mutation occurrence was inferred using the AF of driver mutations, although this simplistic method requires

validation. More sophisticated approached has studied this phenomenon at a single cell level and revealed novel CLL drivers (Vickovic et al. 2016; Wang et al. 2017). However, single cell analysis costs and technical aspects make difficult to consider this technique in routine diagnostic at the moment but like WGS, could expand in future years.

In addition, inter-tumour heterogeneity is another key concept in cancer and has led the research for precision medicine. Indeed, as patients carry different type of mutations, it is possible to treat their tumour with different types of targeted therapy specifically directed against a mutated cell type. This is the case in CLL as it is a highly genetically heterogeneous cancer type. Indeed, patients carry mutations in different driver genes (Wang et al. 2011; Fabbri et al. 2011; Puente et al. 2011; Mansouri et al. 2013; Landau et al. 2015; Burns et al. 2018). In Chapter 5, I have discussed the necessity of implementing an MPS integrated approach to screen for the different types of alterations (SNVs, indels, CNAs) and the multiple driver genes, in routine diagnostics, corroborating other studies' conclusions (Sutton et al. 2015; Vollbrecht et al. 2015; Baliakas et al. 2014; Jeromin et al. 2014; Doménech et al. 2012). During the course of the thesis this idea has been confirmed multiple times and it is now well accepted in the community (Rigolin et al. 2016; Roos-Weil et al. 2016). Since then, multiple genomic studies have highlighted the presence of more than 2000 recurrently mutated genes, increasing our understanding of this complexity (Wang et al. 2011; Landau et al. 2013; Landau et al. 2015; Puente et al. 2015; Quesada et al. 2011; Burns et al. 2018). In Chapter 6, I confirmed these observations.

Inter-tumour heterogeneity of cancer guide patient stratification and management. In Chapter 4, all nine patients with a clinical report (from the FFPE data) had different clinically actionable mutations (although the treatment might not be currently available in

the UK). It was also the case in Chapter 5 where 19 patients out of 114 had a MH profile associated with poor outcome, and in Chapter 6 where 18 FCR-treated patients had a *TP53* of *SAMHD1* alteration and 20 out of 32 FCR non-responders had an increased mutational signature 3.

7.5 Mutational markers and genomics complexity linked with poor-prognosis in CLL

CLL is also clinically heterogeneous which is explained by the genetic heterogeneity previously described in Chapter 7.4. In this thesis, I have highlighted important difference between different high-risk CLL groups. In Chapter 5, I have discovered the MH profile in R/R patients associated with adverse response to treatment and survival. This MH profile has been since confirmed by another group (Ljungstrom et al. 2016). In Chapter 6, the FCR non-responder had more *TP53* and *SAMHD1* alterations. *SAMHD1* downregulation is associated with several cancer, and has been previously described in CLL (Clifford et al. 2014).

GC has been described for the first time in CLL more than 30 years ago (Han et al. 1984). Since then, despite the technical improvement of the karyotyping techniques, genomic technologies have not yet redefined the concept of GC. In Chapter 5 and Chapter 6, the R/R patients presented higher degree of GC, characterised with higher number of driver mutations (Chapter 5) and a higher number of CNAs (Chapter 6). These observations have been previously described (Knight et al. 2012). The MH profile of the recurrently mutated

driver genes *ATM, SF3B1* and *TP53,* described in Chapter 5 also highlighted this increase in GC in patients in an advanced stage of the disease.

7.6 Technical consideration

With the increase of mutational landscape cancer studies looking at an increasing number of patients, the altered gene discovery might reach saturation (Chalmers et al. 2017). However, by screening different group of patients, such as the CLL R/R patients in Chapter 5, novel biomarkers can be discovered. Such studies are generally challenging as it is difficult to collect large number of samples for very specific cancer subgroups. In order to gather more than 100 samples from this type of high-risk patients, international collaborations are required (such as the French-UK collaboration described in Chapter 5). Genomic discovery can also be further enhanced by integrated approaches using for instance epigenetic techniques in association with WGS (Kretzmer et al. 2015; Woo et al. 2017). Such integrated methodologies allow in particular the exploration of non-coding variants until now under-studied in cancer (Zhao et al. 2015).

7.7 Future work and directions

The experimental work presented in Chapter 4 including formalin fixation and DNA extraction of FFPE samples requires additional validations in larger cohort of samples. Controlled timed formalin fixation depending on the tissue type and size might reveal an

optimal fixation condition of sample for downstream genomic applications. Furthermore, optimisation of DNA extraction of NBF fixed tissues should be explored. The data generated from FFPE could also benefit from new computational methods to correct for noisy signal in order to more robustly detected CNAs, without having to perform visual inspection. By performing these optimisations, FFPE samples could be routinely used in the clinic for WGS testing, opening up this methodology to more patients for them to benefit from precision medicine.

The subclonal evolution of the CLL genomes has guided biomarker discovery (Landau et al. 2015; Rose-Zerilli et al. 2016). Such studies could benefit from the additional information brought by WGS (Schuh et al. 2012). Additional samples could be collected at relapse of FCR-treated patients to determine the subclone driving the resistance mechanism to CIT. As opposed to previous studies using WES, WGS could allow to measure the genomic changes, especially in terms of mutational signatures and other non-coding features unexplored so far.

WGS of relapse samples will highlight which clone grow out after CIT and drives the resistance mechanism.

7.8 Conclusions

The overall objective of the work presented in this thesis was to address current challenges and limitations of FFPE testing by improving knowledge in genomic technology and by providing a means to offer genomic-guided precision medicine to a greater number of cancer patients. The first aim was to explore the feasibility of using FFPE tissue from

patients with solid tumour for clinical grade whole genome sequencing (WGS). The work highlighted the multiple challenges faced by routine laboratories in order to establish such technology. It also provided the first WGS clinical reports from FFPE samples including clinically actionable genes for most patients. The second aim was to use WGS and targeted sequencing data to search for novel genomic biomarkers of aggressive disease in CLL patients. A MH profile was discovered in high risk R/R patients and global genomic measured, especially in non-coding regions of the genome could explain some of the resistance mechanism to treatment still unknown to date.

References

1000 Genomes Project Consortium et al., 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), pp.56–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23128226.

Agell, L. et al., 2008. KLF6 and TP53 mutations are a rare event in prostate cancer: distinguishing between Taq polymerase artifacts and true mutations. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 21(12), pp.1470–1478.

Agilent, 2017. Agilent SureDesign. Available at: https://earray.chem.agilent.com/suredesign/.

Agilent technologies, 2015. HaloPlex Target Enrichment System. Available at: https://www.agilent.com/cs/library/usermanuals/public/G9900-90001_HaloPlex_Target_Enrichment _Manual.pdf.

Ahdesmäki, M.J. et al., 2017. Prioritisation of structural variant calls in cancer genomes. *PeerJ*, 5, p.e3166. Available at: https://peerj.com/articles/3166.

Ahn, I.E. et al., 2017. Clonal evolution leading to ibrutinib resistance in chronic lymphocytic leukemia. *Blood*, 129(11), pp.1469–1479. Available at: http://www.bloodjournal.org/lookup/doi/10.1182/blood-2016-06-719294.

Alexandrov, L.B. et al., 2013. Signatures of mutational processes in human cancer. *Nature*, 500(7463), pp.415–421. Available at: http://www.nature.com/doifinder/10.1038/nature12477.

Alexandrov, L.B. & Stratton, M.R., 2014. Mutational signatures: the patterns of somatic

mutations hidden in cancer genomes. *Current opinion in genetics & development*, 24, pp.52–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24657537.

Alioto, T.S. et al., 2015. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, 6, p.10001. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26647970.

Alkan, C., Coe, B.P. & Eichler, E.E., 2011. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5), pp.363–376. Available at: http://www.nature.com/doifinder/10.1038/nrg2958.

Van Allen, E.M. et al., 2014. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature Medicine*, 20(6), pp.682–688. Available at: http://www.nature.com/articles/nm.3559.

Anderson, K. et al., 2011. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, 469(7330), pp.356–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21160474.

Anderson, M.A. et al., 2017. Clinico-pathological features and outcomes of progression of CLL on the BCL2 inhibitor venetoclax. *Blood*, 129(25), pp.3362–3371. Available at: http://www.nature.com/doifinder/10.1038/leu.2016.378.

André, F. et al., 2014. Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). *The Lancet Oncology*, 15(3), pp.267–274. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1470204513706119.

André, F. et al., 2016. Molecular alterations and everolimus efficacy in human epidermal growth factor receptor 2-overexpressing metastatic breast cancers: Combined

exploratory biomarker analysis from BOLERO-1 and BOLERO-3. *Journal of Clinical Oncology*, 34(18), pp.2115–2124.

Astolfi, A. et al., 2015. Whole exome sequencing (WES) on formalin- fixed , paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST). *BMC Genomics*, pp.1–11. Available at: http://dx.doi.org/10.1186/s12864-015-1982-6.

Bachman, K.E. et al., 2004. The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biology & Therapy*, 3(8), pp.772–775. Available at: http://www.tandfonline.com/doi/abs/10.4161/cbt.3.8.994.

Baker, S. et al., 1989. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, 244(4901), pp.217–221. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.2649981.

Baliakas, P. et al., 2015. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia*, 29(2), pp.329–336. Available at: http://www.nature.com/doifinder/10.1038/leu.2014.196.

Baliakas, P. et al., 2014. Recurrent mutations refine prognosis in chronic lymphocytic leukemia. *Leukemia*, 29(April), pp.1–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24943832.

Bamshad, M.J. et al., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics*, 12(11), pp.745–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21946919.

Bansal, V. et al., 2014. Outlier-Based Identification of Copy Number Variations Using Targeted Resequencing in a Small Cohort of Patients with Tetralogy of Fallot C. Liu, ed. *PLoS ONE*, 9(1), p.e85375. Available at: http://dx.plos.org/10.1371/journal.pone.0085375.

Barcelos, D., Franco, M.F. & Leão, S.C., 2008. Effects of tissue handling and processing steps on PCR for detection of Mycobacterium tuberculosis in formalin-fixed paraffin-embedded samples. *Revista do Instituto de Medicina Tropical de Sao Paulo*, 50(6), pp.321–326. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0036-46652008000600002&lng=en&nrm=iso&tlng=en.

Barski, A. et al., 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4), pp.823–837. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867407006009.

Bass, B.P. et al., 2014. A Review of Preanalytical Factors Affecting Molecular, Protein, and Morphological Analysis of Formalin-Fixed, Paraffin-Embedded (FFPE) Tissue: How Well Do You Know Your FFPE Specimen? *Archives of Pathology & Laboratory Medicine*, 138(11), pp.1520–1530. Available at: http://www.archivesofpathology.org/doi/abs/10.5858/arpa.2013-0691-RA.

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), pp.53–59. Available at: http://www.nature.com/doifinder/10.1038/nature07517.

Berns, K. et al., 2016. Loss of ARID1A Activates ANXA1, which Serves as a Predictive Biomarker for Trastuzumab Resistance. *Clinical Cancer Research*, 22(21), pp.5238–5248. Available at: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-15-2996.

Betge, J. et al., 2015. Amplicon Sequencing of Colorectal Cancer: Variant Calling in Frozen and Formalin-Fixed Samples J.-S. Seo, ed. *PLOS ONE*, 10(5), p.e0127146. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4444292&tool=pmcentrez&

rendertype=abstract.

Biankin, A. V., Piantadosi, S. & Hollingsworth, S.J., 2015. Patient-centric trials for therapeutic development in precision oncology. *Nature*, 526(7573), pp.361–370. Available at: http://www.nature.com/doifinder/10.1038/nature15819.

Biosciences, P., 2014. Introduction to SMRTbell™ Template Preparation. Available at: http://files.pacb.com/Training/IntroductiontoSMRTbellTemplatePreparation/story_content/external_files/Introduction to SMRTbell™ Template Preparation.pdf.

Boeva, V. et al., 2014. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics (Oxford, England)*, 30(24), pp.3443–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25016581.

Bourgon, R. et al., 2014. High-Throughput Detection of Clinically Relevant Mutations in Archived Tumor Samples by Multiplexed PCR and Next-Generation Sequencing. *Clinical Cancer Research*, 20(8), pp.2080–2091. Available at: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-13-3114.

Bovee, D. et al., 2008. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nature genetics*, 40(1), pp.96–101. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18157130.

Boveri, T., 2008. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science*, 121(Supplement 1), pp.1–84. Available at: http://jcs.biologists.org/cgi/doi/10.1242/jcs.025742.

Bowers, J. et al., 2009. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature methods*, 6(8), pp.593–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19620973.

Broad institute, 2017. Picard command line tools. *Github*. Available at:

https://broadinstitute.github.io/picard/picard-metric-definitions.html#AlignmentSummaryMetrics.

Brown, J.R. et al., 2012. Integrative genomic analysis implicates gain of PIK3CA at 3q26 and MYC at 8q24 in chronic lymphocytic leukemia. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 18(14), pp.3791–802. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22623730.

Brücher, B.L.D.M. & Jamall, I.S., 2016. Somatic Mutation Theory - Why it's Wrong for Most Cancers. *Cellular Physiology and Biochemistry*, 38(5), pp.1663–1680. Available at: https://www.karger.com/Article/FullText/443106.

Buckley, A.R. et al., 2017. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC Genomics*, 18(1), p.458. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28606096%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5467262.

Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12), pp.1213–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24097267.

Burger, J.A. et al., 2016. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nature Communications*, 7, p.11589. Available at: http://www.nature.com/doifinder/10.1038/ncomms11589.

Burger, J.A. et al., 2014. Safety and activity of ibrutinib plus rituximab for patients with high-risk chronic lymphocytic leukaemia: a single-arm, phase 2 study. *The Lancet Oncology*, 15(10), pp.1090–1099. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1470204514703353.

Burns, A. et al., 2018. Whole-genome sequencing of chronic lymphocytic leukaemia reveals distinct differences in the mutational landscape between IgHVmut and IgHVunmut subgroups. *Leukemia*, 32(2), pp.332–342. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28584254.

Burrell, R. a et al., 2013. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467), pp.338–45. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24048066.

Byrd, J.C. et al., 2013. Targeting BTK with Ibrutinib in Relapsed Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 369(1), pp.32–42. Available at: http://www.nejm.org/doi/10.1056/NEJMoa1215637.

Byrd, J.C., Stilgenbauer, S. & Flinn, I.W., 2004. Chronic lymphocytic leukemia. *Hematology. American Society of Hematology. Education Program*, pp.163–83. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15561682.

Callari, M. et al., 2017. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Medicine*, 9(1), p.35. Available at: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0425-1.

Carter, S.L. et al., 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5), pp.413–421. Available at: http://www.nature.com/doifinder/10.1038/nbt.2203.

Cerami, E. et al., 2012. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5), pp.401–404.

Chalmers, Z.R. et al., 2017. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, 9(1), p.34. Available at:

http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0424-2.

Chen, G. et al., 2014. Cytosine Deamination Is a Major Cause of Baseline Noise in Next-Generation Sequencing. *Molecular Diagnosis & Therapy*, 18(5), pp.587–593. Available at: http://link.springer.com/10.1007/s40291-014-0115-2.

Chiara, M. & Pavesi, G., 2017. Evaluation of Quality Assessment Protocols for High Throughput Genome Resequencing Data. *Frontiers in Genetics*, 8. Available at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00094/full.

Chigrinova, E. et al., 2013. Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood*, 122(15), pp.2673–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24004666.

Cibulskis, K. et al., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3), pp.213–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3833702&tool=pmcentrez&rendertype=abstract.

Clark, M.J. et al., 2011. Performance comparison of exome DNA sequencing technologies. *Nature biotechnology*, 29(10), pp.908–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21947028.

Clarke, J. et al., 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4), pp.265–270. Available at: http://www.nature.com/doifinder/10.1038/nnano.2009.12.

Clifford, R. et al., 2014. SAMHD1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to DNA damage. *Blood*, 123(7), pp.1021–1031.

Collins, F.S. & Varmus, H., 2015. A new initiative on precision medicine. *The New England journal of medicine*, 372(9), pp.793–5. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/25635347.

Corcoran, R.B. et al., 2013. Synthetic Lethal Interaction of Combined BCL-XL and MEK Inhibition Promotes Tumor Regressions in KRAS Mutant Cancer Models. *Cancer Cell*, 23(1), pp.121–128. Available at: http://dx.doi.org/10.1016/j.ccr.2012.11.007.

Cottrell, C.E. et al., 2014. Validation of a next-generation sequencing assay for clinical molecular oncology. *Journal of Molecular Diagnostics*, 16(1), pp.89–105. Available at: http://dx.doi.org/10.1016/j.jmoldx.2013.10.002.

Creighton, C.J. et al., 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456), pp.43–49. Available at: http://www.nature.com/doifinder/10.1038/nature12222.

Croce, C.M., 2008. Oncogenes and Cancer. *New England Journal of Medicine*, 358(5), pp.502–511. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMra072367.

Cross, W.C., Graham, T.A. & Wright, N.A., 2016. New paradigms in clonal evolution: punctuated equilibrium in cancer. *The Journal of Pathology*, 240(2), pp.126–136. Available at: http://doi.wiley.com/10.1002/path.4757.

Cui, K. & Zhao, K., 2012. Genome-Wide Approaches to Determining Nucleosome Occupancy in Metazoans Using MNase-Seq. In pp. 413–419. Available at: http://link.springer.com/10.1007/978-1-61779-477-3_24.

Damle, R.N. et al., 1999. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood*, 94(6), pp.1840–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10477712.

Darwanto, A. et al., 2017. Use of the QIAGEN GeneReader NGS system for detection of KRAS mutations, validated by the QIAGEN Therascreen PCR kit and alternative NGS platform. *BMC Cancer*, 17(1), p.358. Available at:

http://bmccancer.biomedcentral.com/articles/10.1186/s12885-017-3328-z.

Davies, H. et al., 2002. Mutations of the BRAF gene in human cancer. *Nature*, 417(6892), pp.949–954. Available at: http://www.nature.com/articles/nature00766.

Davis, A., Gao, R. & Navin, N., 2017. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1867(2), pp.151–161. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0304419X17300197.

Davis, S., 2016. ngCGH. Available at: http://github.com/seandavi/ngCGH.

Deans, Z.C. et al., 2017. Integration of next-generation sequencing in clinical diagnostic molecular pathology laboratories for analysis of solid tumours; an expert opinion on behalf of IQN Path ASBL. *Virchows Archiv*, 470(1), pp.5–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27678269.

Demidov, G. et al., 2016. A statistical approach to detection of copy number variations in PCR-enriched targeted sequencing data. *BMC Bioinformatics*, 17(1), p.429. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1272-6.

Diaz Jr, L.A. et al., 2012. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*. Available at: http://www.nature.com/doifinder/10.1038/nature11219.

Dixon, J.R. et al., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), pp.376–380. Available at: http://www.nature.com/doifinder/10.1038/nature11082.

Do, H. et al., 2013. Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. *Clinical Chemistry*, 59(9), pp.1376–1383.

Do, H. & Dobrovic, A., 2012. Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil-DNA glycosylase. *Oncotarget*, 3(5), pp.546–558. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3388184&tool=pmcentrez&rendertype=abstract.

Do, H. & Dobrovic, A., 2014. Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization. *Clinical Chemistry*, 61(1), pp.64–71. Available at: http://www.clinchem.org/cgi/doi/10.1373/clinchem.2014.223040.

Döhner, H. et al., 1997. 11q deletions identify a new subset of B-cell chronic lymphocytic leukemia characterized by extensive nodal involvement and inferior prognosis. *Blood*, 89(7), pp.2516–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9116297.

Döhner, H. et al., 2000. Genomic aberrations and survival in chronic lymphocytic leukemia. *The New England journal of medicine*, 343(26), pp.1910–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11136261.

Doménech, E. et al., 2012. New Mutations in Chronic Lymphocytic Leukemia Identified by Target Enrichment and Deep Sequencing R. A. Panepucci, ed. *PLoS ONE*, 7(6), p.e38158. Available at: http://dx.plos.org/10.1371/journal.pone.0038158.

van Dongen, J.J.M. et al., 2003. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*, 17(12), pp.2257–317. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14671650.

Dreger, P. et al., 2013. TP53, SF3B1, and NOTCH1 mutations and outcome of allotransplantation for chronic lymphocytic leukemia: six-year follow-up of the

GCLLSG CLL3X trial. *Blood*, 121(16), pp.3284–3289.

Drmanac, R. et al., 2010. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*, 327(5961), pp.78–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19892942.

Dufour, A. et al., 2013. Inactivation of TP53 correlates with disease progression and low miR-34a expression in previously treated chronic lymphocytic leukemia patients. *Blood*, 121(18), pp.3650–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23525797.

Durdiaková, J. et al., 2012. Comparison of different collection procedures and two methods for DNA isolation from saliva. *Clinical chemistry and laboratory medicine*, 50(4), pp.643–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22149741.

Van Dyke, D.L. et al., 2016. The Dohner fluorescence in situ hybridization prognostic classification of chronic lymphocytic leukaemia (CLL): the CLL Research Consortium experience. *British journal of haematology*, 173(1), pp.105–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26848054.

Edelmann, J. et al., 2012. High-resolution genomic profiling of chronic lymphocytic leukemia reveals new recurrent genomic alterations. *Blood*, 120(24), pp.4783–4794.

Eid, J. et al., 2009. Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, 323(5910), pp.133–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19023044.

Ellis, I.O., 2000. Recommendations for HER2 testing in the UK. *Journal of Clinical Pathology*, 53(12), pp.890–892. Available at: http://jcp.bmj.com/cgi/doi/10.1136/jcp.53.12.890.

Ewing, B. & Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3), pp.186–94. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/9521922.

Fabbri, G. et al., 2011. Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation. *The Journal of experimental medicine*, 208(7), pp.1389–401. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21670202%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3135373.

Fabbri, G. et al., 2013. Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. *The Journal of Experimental Medicine*, 210(11), pp.2273–88. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24127483.

Faber, A.C. et al., 2014. mTOR inhibition specifically sensitizes colorectal cancers with KRAS or BRAF mutations to BCL-2/BCL-XL inhibition by suppressing MCL-1. *Cancer Discovery*, 4(1), pp.42–52.

Fang, H. et al., 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine*, 6(10), p.89. Available at: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-014-0089-z.

Farooqui, M.Z.H. et al., 2015. Ibrutinib for previously untreated and relapsed or refractory chronic lymphocytic leukaemia with TP53 aberrations: a phase 2, single-arm trial. *The Lancet Oncology*, 16(2), pp.169–176. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1470204514711829.

Fearon, E.R. & Vogelstein, B., 1990. A genetic model for colorectal tumorigenesis. *Cell*, 61(5), pp.759–767. Available at: http://linkinghub.elsevier.com/retrieve/pii/009286749090186I.

Finn, R.D. et al., 2014. Pfam: The protein families database. *Nucleic Acids Research*, 42(D1), pp.211–222.

Fischer, K. et al., 2016. Long-term remissions after FCR chemoimmunotherapy in previously untreated patients with CLL: updated results of the CLL8 trial. *Blood*, 127(2), pp.208–215. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26486789.

Forbes, S. a. et al., 2011. COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 39(SUPPL. 1), pp.945–950.

Frampton, G.M. et al., 2013. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology*, 31(11), pp.1023–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24142049.

Freeman, J., Perry, G. & Feuk, L., 2006. Copy number variation: new insights in genome diversity. *Genome research*, (617), pp.949–961. Available at: http://genome.cshlp.org/content/16/8/949.short.

Friedman, J., Hastie, T. & Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1), pp.1–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20808728.

Fullwood, M.J. et al., 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research*, 19(4), pp.521–532. Available at: http://genome.cshlp.org/cgi/doi/10.1101/gr.074906.107.

Futreal, P. et al., 2004. A census of human cancer genes. *Nat Rev Cancer*, 4(3), pp.177–183.

Gahrton, G. et al., 1980. Nonrandom chromosomal aberrations in chronic lymphocytic leukemia revealed by polyclonal B-cell-mitogen stimulation. *Blood*, 56(4), pp.640–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6968231.

Gal-Yam, E.N. et al., 2008. Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35), pp.12979–84. Available

at: http://www.ncbi.nlm.nih.gov/pubmed/18753622.

Gao, J. et al., 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269), pp.1–20. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4160307/.

Gao, Y. et al., 2016. Single molecule targeted sequencing for cancer gene mutation detection. *Scientific Reports*, 6(1), p.26110. Available at: http://www.nature.com/articles/srep26110.

García-García, G. et al., 2016. Assessment of the latest NGS enrichment capture methods in clinical context. *Scientific Reports*, 6(1), p.20948. Available at: http://www.nature.com/articles/srep20948.

Gargis, A.S. et al., 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature biotechnology*, 30(11), pp.1033–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23138292.

Genohub Inc., 2017. https://genohub.com. Available at: https://genohub.com.

Gerlinger, M. et al., 2014. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46(3), pp.225–233. Available at: http://www.nature.com/doifinder/10.1038/ng.2891.

Gerlinger, M. et al., 2012. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10), pp.883–892. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa1113205.

Gerstung, M. et al., 2017. The evolutionary history of 2,658 cancers. *bioRxiv*. Available at: http://biorxiv.org/content/early/2017/07/11/161562.abstract.

Getz, G. et al., 2013. Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447), pp.67–73. Available at:

http://www.nature.com/doifinder/10.1038/nature12113.

Geyer, C.E. et al., 2006. Lapatinib plus Capecitabine for HER2-Positive Advanced Breast Cancer. *New England Journal of Medicine*, 355(26), pp.2733–2743. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa064320.

Gilbert, N. et al., 2004. Chromatin Architecture of the Human Genome. *Cell*, 118(5), pp.555–566. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15339661.

Gonzalez, D. et al., 2011. Mutational status of the TP53 gene as a predictor of response and survival in patients with chronic lymphocytic leukemia: Results from the LRF CLL4 trial. *Journal of Clinical Oncology*, 29(16), pp.2223–2229.

Goodwin, S., McPherson, J.D. & McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), pp.333–351. Available at: http://www.nature.com/doifinder/10.1038/nrg.2016.49.

Greer, C.E., Lund, J.K. & Manos, M.M., 1991. PCR amplification from paraffin-embedded tissues: recommendations on fixatives for long-term storage and prospective studies. *Genome Research*, 1(1), pp.46–50. Available at: http://www.genome.org/cgi/doi/10.1101/gr.1.1.46.

Grever, M.R. et al., 2007. Comprehensive assessment of genetic and molecular features predicting outcome in patients with chronic lymphocytic leukemia: Results from the US intergroup phase III trial E2997. *Journal of Clinical Oncology*, 25(7), pp.799–804.

Guarini, A. et al., 2012. ATM gene alterations in chronic lymphocytic leukemia patients induce a distinct gene expression profile and predict disease progression. *Haematologica*, 97(1), pp.47–55. Available at: http://www.haematologica.org/cgi/doi/10.3324/haematol.2011.049270.

de Guibert, S. et al., 2014. Bendamustine, Ofatumumab and High-Dose Methylprednisolone

(BOMP) in Relapsed/Refractory CLL: Results of a Planned Interim Analysis of the French CLL Intergroup ICLL01 Phase II Trial. *Blood*, 124(21), p.3341 LP-3341. Available at: http://www.bloodjournal.org/content/124/21/3341.abstract.

Guièze, R. & Wu, C.J., 2015. Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia. *Blood*, 126(4), pp.445–453.

Gundem, G. et al., 2015. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547), pp.353–357. Available at: http://www.nature.com/doifinder/10.1038/nature14347.

Gunnarsson, R. et al., 2010. Large but not small copy-number alterations correlate to high-risk genomic aberrations and survival in chronic lymphocytic leukemia: a high-resolution genomic screening of newly diagnosed patients. *Leukemia*, 24(1), pp.211–215. Available at: http://www.nature.com/doifinder/10.1038/leu.2009.187.

Haferlach, C. et al., 2007. Comprehensive genetic characterization of CLL: a study on 506 cases analysed with chromosome banding analysis, interphase FISH, IgV(H) status and immunophenotyping. *Leukemia*, 21(12), pp.2442–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17805327.

Hallek, M. et al., 2010. Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *Lancet (London, England)*, 376(9747), pp.1164–74. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0140673610613815.

Hallek, M. et al., 2008. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood*, 111(12), pp.5446–5456. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2007-

06-093906.

Hamblin, A. et al., 2017. Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: Retrospective validation and prospective audit in the UK National Health Service M. Ladanyi, ed. *PLOS Medicine*, 14(2), p.e1002230. Available at: http://dx.plos.org/10.1371/journal.pmed.1002230.

Hamblin, T.J. et al., 1999. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood*, 94(6), pp.1848–54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10477713.

Hammerman, P.S. et al., 2012. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417), pp.519–525. Available at: http://www.nature.com/doifinder/10.1038/nature11404.

Han, T. et al., 1984. Prognostic Importance of Cytogenetic Abnormalities in Patients with Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 310(5), pp.288–292. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6690952.

Hanahan, D. & Weinberg, R.A., 2000. The hallmarks of cancer. *Cell*, 100(1), pp.57–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10647931.

Hansen, N.F. et al., 2013. Shimmer: Detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics*, 29(12), pp.1498–1503. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23620360.

Harris, T.D. et al., 2008. Single-molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, 320(5872), pp.106–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18388294.

Hedegaard, J. et al., 2014. Next-Generation Sequencing of RNA and DNA Isolated from Paired Fresh-Frozen and Formalin-Fixed Paraffin-Embedded Samples of Human Cancer

and Normal Tissue Z. Zuo, ed. *PLoS ONE*, 9(5), p.e98187. Available at: http://dx.plos.org/10.1371/journal.pone.0098187.

Herling, C.D. et al., 2016. Complex karyotypes and KRAS and POT1 mutations impact outcome in CLL after chlorambucil-based chemotherapy or chemoimmunotherapy. *Blood*, 128(3), pp.395–404. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27226433.

Hernández, J.Á. et al., 2015. A Low Frequency of Losses in 11q Chromosome Is Associated with Better Outcome and Lower Rate of Genomic Mutations in Patients with Chronic Lymphocytic Leukemia. *PloS one*, 10(11), p.e0143073. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26630574.

Hillmen, P. et al., 2011. A randomized phase II trial of fludarabine, cyclophosphamide and mitoxantrone (FCM) with or without rituximab in previously treated chronic lymphocytic leukaemia. *British Journal of Haematology*, 152(5), pp.570–578.

Holley, T. et al., 2012. Deep Clonal Profiling of Formalin Fixed Paraffin Embedded Clinical Samples M. R. Emmert-Buck, ed. *PLoS ONE*, 7(11), p.e50586. Available at: http://dx.plos.org/10.1371/journal.pone.0050586.

Hollstein, M. et al., 1991. p53 mutations in human cancers. *Science*, 253(5015), pp.49–53. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1905840.

Hong, G.F., 1981. A method for sequencing single-stranded cloned DNA in both directions. *Bioscience reports*, 1(3), pp.243–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6271278.

Hong, M.K.H. et al., 2015. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nature Communications*, 6(1), p.6605. Available at: http://www.nature.com/doifinder/10.1038/ncomms7605.

Horak, P., Fröhling, S. & Glimm, H., 2016. Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls. *ESMO open*, 1(5), p.e000094. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27933214.

Huang, L. et al., 2017. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *Journal of the American Medical Informatics Association : JAMIA*, 24(3), pp.513–519. Available at: https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocw148.

Huang, Y.-H. & Vakoc, C.R., 2016. A Biomarker Harvest from One Thousand Cancer Cell Lines. *Cell*, 166(3), pp.536–537. Available at: http://dx.doi.org/10.1016/j.cell.2016.07.010.

Hudson (Chairperson), T.J. et al., 2010. International network of cancer genome projects. *Nature*, 464(7291), pp.993–998. Available at: http://www.nature.com/doifinder/10.1038/nature08987.

Hyman, D.M. et al., 2015. Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *New England Journal of Medicine*, 373(8), pp.726–736. Available at: http://www.nejm.org/doi/10.1056/NEJMoa1502309.

Illumina, 2016. Illumina FFPE QC Kit for TruSeq Amplicon - Cancer Panel and TruSeq Custom Amplicon v1.5. *Document #15033035 v01*. Available at: https://support.illumina.com/content/dam/illumina-support/documents/myillumina/255af045-4832-44e6-af20-ee4f28956b5e/truseq_amplicon_cancer_panel_ffpe_qc_insert_15033035_a.pdf.

Illumina, 2014. Isaac Genome Alignment and Isaac Variant Caller. Available at: https://www.illumina.com/documents/products/whitepapers/whitepaper_isaac_workflow.pdf.

Illumina Inc., 2010. De Novo Assembly Using Illumina Reads. *Technical Note: Sequencing*. Available at: https://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf.

Illumina Inc., 2017a. Design Studio. Available at: https://designstudio.illumina.com/.

Illumina Inc., 2011. Quality Scores for Next-Generation Sequencing. *Technical Note: Sequencing*. Available at: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf.

Illumina Inc., 2015. TruSeq® Custom Amplicon v1.5. *Data Sheet: Targeted Resequencing*. Available at: https://support.illumina.com.cn/content/dam/illumina-marketing/documents/products/datasheets/datasheet_truseq_custom_amplicon.pdf.

Illumina Inc., 2017b. TruSeq™ DNA PCR-Free. *Prepare Library*. Available at: https://www.illumina.com/documents/products/datasheets/datasheet_truseq_dna_pcr_free_sample_prep.pdf.

Illumina Inc., 2016. Whole Genome Sequencing v5.0 VCF File Annotations. *Whole Genome Sequencing v5.0 App Online Help*. Available at: https://support.illumina.com/help/BaseSpace_App_WGS_v5_OLH_15050955_02/Content/Source/Informatics/Apps/VCFFileAnnotations_appENR_appISCWGS.htm.

Jeromin, S. et al., 2014. SF3B1 mutations correlated to cytogenetics and mutations in NOTCH1, FBXW7, MYD88, XPO1 and TP53 in 1160 untreated CLL patients. *Leukemia*, 28(1), pp.108–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24113472.

Jethwa, A. et al., 2013. Targeted resequencing for analysis of clonal composition of recurrent gene mutations in chronic lymphocytic leukaemia. *British Journal of Haematology*,

163(4), pp.496–500. Available at: http://doi.wiley.com/10.1111/bjh.12539.

Johnson, B.E. et al., 2014. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science (New York, N.Y.)*, 343(6167), pp.189–193. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24336570.

Jones, S. et al., 2008. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11), pp.4283–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18337506.

Juliusson, G. et al., 1985. Prognostic information from cytogenetic analysis in chronic B-lymphocytic leukemia and leukemic immunocytoma. *Blood*, 65(1), pp.134–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3871161.

Juliusson, G. et al., 1990. Prognostic subgroups in B-cell chronic lymphocytic leukemia defined by specific chromosomal abnormalities. *The New England journal of medicine*, 323(11), pp.720–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2201915.

Kadri, S. et al., 2016. Clonal Evolution Pattern of Leukemia Progression and Richter Transformation in Ibrutinib-Relapsed CLL Patients. *Blood*, 128(22), pp.3961–3961. Available at: http://www.bloodjournal.org.gsk.idm.oclc.org/content/128/22/3961.

Kämpjärvi, K. et al., 2014. Somatic MED12 mutations are associated with poor prognosis markers in chronic lymphocytic leukemia. *Oncotarget*, 6(3), pp.1884–1888.

Kandoth, C. et al., 2013. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471), pp.333–339. Available at: http://www.nature.com/doifinder/10.1038/nature12634.

Kanehisa, M. et al., 2016. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(November 2016), pp.1–15. Available at: http://fdslive.oup.com/www.oup.com/pdf/production_in_progress.pdf%5Cnhttp://www.

ncbi.nlm.nih.gov/pubmed/27899565.

Kasar, S. et al., 2015. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications*, 6, p.8866. Available at: http://www.nature.com/doifinder/10.1038/ncomms9866.

Kerick, M. et al., 2011. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC medical genomics*, 4(1), p.68. Available at: http://www.biomedcentral.com/1755-8794/4/68.

Killela, P.J. et al., 2013. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proceedings of the National Academy of Sciences*, 110(15), pp.6021–6026. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1303607110.

Kim, K. et al., 2015. Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. *Genomics & Informatics*, 13(2), p.31. Available at: http://genominfo.org/journal/view.php?doi=10.5808/GI.2015.13.2.31.

Kim, S., Jacob, L. & Speed, T.P., 2014. Combining calls from multiple somatic mutation-callers. *BMC Bioinformatics*, 15(1), p.154. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-154.

Kinzler, K.W. & Vogelstein, B., 1996. Lessons from Hereditary Colorectal Cancer. *Cell*, 87(2), pp.159–170. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867400813331.

Kipps, T.J., 2008. Genomic complexity in chronic lymphocytic leukemia. *Blood*, 112(5), pp.1550–1550. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2008-

06-156729.

Knight, S.J.L. et al., 2012. Quantification of subclonal distributions of recurrent genomic aberrations in paired pre-treatment and relapse samples from patients with b-cell chronic lymphocytic leukemia. *Leukemia*, 26(7), pp.1564–1575. Available at: http://www.nature.com/doifinder/10.1038/leu.2012.13.

Knudson, A.G., 1971. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4), pp.820–3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5279523.

Knudson, A.G., 2001. Two genetic hits (more or less) to cancer. *Nature reviews. Cancer*, 1(2), pp.157–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11905807.

Koboldt, D.C. et al., 2013. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1), pp.27–38. Available at: http://dx.doi.org/10.1016/j.cell.2013.09.006.

Koboldt, D.C. et al., 2012. VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3), pp.568–576.

Koitzsch, U. et al., 2017. Use of the GeneReader NGS System in a clinical pathology laboratory: a comparative study. *Journal of clinical pathology*, 70(8), pp.725–728. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28400467.

Kralovics, R. et al., 2005. A Gain-of-Function Mutation of JAK2 in Myeloproliferative Disorders. *New England Journal of Medicine*, 352(17), pp.1779–1790. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa051113.

Kranenburg, O., 2005. The KRAS oncogene: Past, present, and future. *Biochimica et*

Biophysica Acta (BBA) - Reviews on Cancer, 1756(2), pp.81–82. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0304419X05000624.

Kretzmer, H. et al., 2015. DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nature Genetics*, 47(11), pp.1316–1325. Available at: http://www.nature.com/doifinder/10.1038/ng.3413.

Krøigård, A.B. et al., 2016. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data I. K. Jordan, ed. *PLOS ONE*, 11(3), p.e0151664. Available at: http://dx.plos.org/10.1371/journal.pone.0151664.

Kujawski, L. et al., 2015. Genomic complexity identifies patients with aggressive chronic lymphocytic leukemia. *Neoplasia*, 112(5), pp.1993–2004.

Kumar, P., Henikoff, S. & Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(8), pp.1073–1081. Available at: http://www.nature.com/doifinder/10.1038/nprot.2009.86.

Landau, D.A. et al., 2013. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell*, 152(4), pp.714–726. Available at: http://dx.doi.org/10.1016/j.cell.2013.01.019.

Landau, D.A. et al., 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature*, 526(7574), pp.525–530. Available at: http://www.nature.com/doifinder/10.1038/nature15395.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921. Available at: http://www.nature.com/nature/journal/v409/n6822/abs/409860a0.html%5Cnhttp://www.

ncbi.nlm.nih.gov/pubmed/11237011.

Landrum, M.J. et al., 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1), pp.D862–D868. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26582918.

Larrayoz, M. et al., 2017. Non-coding NOTCH1 mutations in chronic lymphocytic leukemia; their clinical impact in the UK CLL4 trial. *Leukemia*, 31(2), pp.510–514. Available at: http://www.nature.com/doifinder/10.1038/leu.2016.298.

Laskin, J. et al., 2015. Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. *Molecular Case Studies*, 1(1), p.a000570. Available at: http://molecularcasestudies.cshlp.org/lookup/doi/10.1101/mcs.a000570.

Lawler, M. et al., 2015. Changing the Paradigm--Multistage Multiarm Randomized Trials and Stratified Cancer Medicine. *The Oncologist*, 20(8), pp.849–851. Available at: http://theoncologist.alphamedpress.org/cgi/doi/10.1634/theoncologist.2015-0014.

Lawrence, M.S. et al., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457), pp.214–8. Available at: http://dx.doi.org/10.1038/nature12213.

Lee, C. et al., 2017. A Method to Evaluate the Quality of Clinical Gene-Panel Sequencing Data for Single-Nucleotide Variant Detection. *The Journal of Molecular Diagnostics*, 19(5), pp.651–658. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1525157817300740.

Lelieveld, S.H. et al., 2015. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation*, 36(8), pp.815–822. Available at: http://doi.wiley.com/10.1002/humu.22813.

Lemmon, M.A. & Schlessinger, J., 2010. Cell signaling by receptor tyrosine kinases. *Cell*,

141(7), pp.1117–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20602996.

Levine, R.L. et al., 2005. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell*, 7(4), pp.387–397. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1535610805000942.

Ley, T.J.T.J. et al., 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218), pp.66–72. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2603574.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19505943.

Li, J., 1997. PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer. *Science*, 275(5308), pp.1943–1947. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.275.5308.1943.

Li, M. et al., 2017. Roles of SAMHD1 in antiviral defense, autoimmunity and cancer. *Reviews in Medical Virology*, 27(4), p.e1931. Available at: http://doi.wiley.com/10.1002/rmv.1931.

Ling, S. et al., 2015. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proceedings of the National Academy of Sciences*, 112(47), pp.E6496–E6505. Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1519556112.

Lionel, A.C. et al., 2017. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *GENETICS in MEDICINE*, 0(June), pp.1–9. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/28771251%0Ahttp://www.nature.com/doifinder/10.1038/gim.2017.119.

Lister, R. et al., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), pp.315–322. Available at: http://www.nature.com/doifinder/10.1038/nature08514.

Liu, R. & Xing, M., 2014. Diagnostic and prognostic TERT promoter mutations in thyroid fine-needle aspiration biopsy. *Endocrine-related cancer*, 21(5), pp.825–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25121551.

Liu, T., Yuan, X. & Xu, D., 2016. Cancer-Specific Telomerase Reverse Transcriptase (TERT) Promoter Mutations: Biological and Clinical Implications. *Genes*, 7(7), p.38. Available at: http://www.mdpi.com/2073-4425/7/7/38.

Ljungstrom, V. et al., 2016. Whole-exome sequencing in relapsing chronic lymphocytic leukemia: clinical impact of recurrent RPS15 mutations. *Blood*, 127(8), pp.1007–1016. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2015-10-674572.

Loeb, L.A., 1991. Mutator phenotype may be required for multistage carcinogenesis. *Cancer research*, 51(12), pp.3075–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2039987.

Loeb, L.A., Loeb, K.R. & Anderson, J.P., 2003. Multiple mutations and cancer. *Proceedings of the National Academy of Sciences*, 100(3), pp.776–781. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.0334858100.

Ludyga, N. et al., 2012. Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses. *Virchows Archiv*, 460(2), pp.131–140.

Lunter, G. & Goodson, M., 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6), pp.936–939.

Lymphoma Association, 2017. RIAltO: a phase 3 trial to see how well ofatumumab works with chemotherapy for people with chronic lymphocytic leukaemia who can't have more intensive treatment. Available at: https://www.lymphomas.org.uk/lymphoma-trialslink/find-trial/lymphoma-rialto-trial.

Lynch, T.J. et al., 2004. Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *New England Journal of Medicine*, 350(21), pp.2129–2139. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa040938.

Macintyre, G. et al., 2017. Copy-number signatures and mutational processes in ovarian carcinoma. *bioRxiv*. Available at: http://biorxiv.org/content/early/2017/09/04/174201.abstract.

Magi, A. et al., 2013. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome biology*, 14(10), p.R120. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24172663.

Malek, S.N., 2013. The biology and clinical significance of acquired genomic copy number aberrations and recurrent gene mutations in chronic lymphocytic leukemia. *Oncogene*, 32(23), pp.2805–2817. Available at: http://dx.doi.org/10.1038/onc.2012.411.

Mansouri, L. et al., 2013. Short telomere length is associated with NOTCH1/SF3B1/TP53 aberrations and poor outcome in newly diagnosed chronic lymphocytic leukemia patients. *American Journal of Hematology*, 88(8), pp.647–651.

Margulies, M. et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp.376–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16056220.

Martincorena, I. & Campbell, P.J., 2015. Somatic mutation in cancer and normal cells.

Science (New York, N.Y.), 349(6255), pp.1483–1489.

Mayr, C. et al., 2006. Chromosomal translocations are associated with poor prognosis in chronic lymphocytic leukemia. *Blood*, 107(2), pp.742–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16179374.

McLaren, W. et al., 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)*, 26(16), pp.2069–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20562413.

McLaren, W. et al., 2016. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), p.122. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4.

McLendon, R. et al., 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216), pp.1061–1068. Available at: http://www.nature.com/doifinder/10.1038/nature07385.

McPherson, J.D. et al., 2001. A physical map of the human genome. *Nature*, 409(6822), pp.934–941. Available at: http://www.nature.com/doifinder/10.1038/35057157.

Medvedev, P., Stanciu, M. & Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11s), pp.S13–S20. Available at: http://www.nature.com/doifinder/10.1038/nmeth.1374.

Menon, R. et al., 2012. Exome Enrichment and SOLiD Sequencing of Formalin Fixed Paraffin Embedded (FFPE) Prostate Cancer Tissue. *International Journal of Molecular Sciences*, 13(12), pp.8933–8942. Available at: http://www.mdpi.com/1422-0067/13/7/8933.

Meric-Bernstam, F. et al., 2015. Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. *Journal of Clinical Oncology*,

33(25), pp.2753–2762. Available at: http://ascopubs.org/doi/10.1200/JCO.2014.60.4165.

Metzker, M.L., 2010. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1), pp.31–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19997069.

Meyerson, M., Gabriel, S. & Getz, G., 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10), pp.685–696. Available at: http://dx.doi.org/10.1038/nrg2841.

Meynert, A.M. et al., 2013. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics*, 14(1), p.195. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-195.

Miething, F. et al., 2006. Effect of fixation to the degradation of nuclear and mitochondrial DNA in different tissues. *Journal of Histochemistry & Cytochemistry*, 54(3), pp.371–374. Available at: http://journals.sagepub.com/doi/10.1369/jhc.5B6726.2005.

Migliardi, G. et al., 2012. Inhibition of MEK and PI3K/mTOR suppresses tumor growth but does not cause tumor regression in patient-derived xenografts of RAS-mutant colorectal carcinomas. *Clinical Cancer Research*, 18(9), pp.2515–2525.

Minuti, G. et al., 2012. Increased MET and HGF gene copy numbers are associated with trastuzumab failure in HER2-positive metastatic breast cancer. *British Journal of Cancer*, 107(5), pp.793–799. Available at: http://www.nature.com/doifinder/10.1038/bjc.2012.335.

Misale, S. et al., 2014. Blockade of EGFR and MEK Intercepts Heterogeneous Mechanisms of Acquired Resistance to Anti-EGFR Therapies in Colorectal Cancer. *Science Translational Medicine*, 6(224), p.224ra26-224ra26. Available at: http://stm.sciencemag.org/cgi/doi/10.1126/scitranslmed.3007947.

Mitchell, A.A. et al., 2004. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics*, 20(7), pp.1022–1032.

Muir, P. et al., 2016. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17(1), p.53. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0917-0.

Munchel, S. et al., 2015. Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. *Oncotarget*, 6(28), pp.25943–61. Available at: http://www.oncotarget.com/fulltext/4671.

Muthusamy, N. et al., 2011. Enhanced detection of chromosomal abnormalities in chronic lymphocytic leukemia by conventional cytogenetics using CpG oligonucleotide in combination with pokeweed mitogen and phorbol myristate acetate. *Cancer Genetics*, 204(2), pp.77–83. Available at: http://linkinghub.elsevier.com/retrieve/pii/S2210776210006046.

Muzny, D.M. et al., 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407), pp.330–337. Available at: http://www.nature.com/doifinder/10.1038/nature11252.

Nadeu, F. et al., 2016. Clinical impact of clonal and subclonal TP53, SF3B1, BIRC3, NOTCH1, and ATM mutations in chronic lymphocytic leukemia. *Blood*, 127(17), pp.2122–2130. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2015-07-659144.

Neilson, J.R. et al., 1997. Deletions at 11q identify a subset of patients with typical CLL who show consistent disease progression and reduced survival. *Leukemia*, 11(11), pp.1929–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9369428.

Van Den Neste, E. et al., 2007. Chromosomal translocations independently predict treatment failure, treatment-free survival and overall survival in B-cell chronic lymphocytic leukemia patients treated with cladribine. *Leukemia*, 21(8), pp.1715–1722. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17541398.

Nichols, K.E. et al., 2001. Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 10(2), pp.83–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11219776.

Nik-Zainal, S. et al., 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605), pp.47–54. Available at: http://www.nature.com/doifinder/10.1038/nature17676.

Nik-Zainal, S., Alexandrov, L.B., et al., 2012. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5), pp.979–993. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867412005284.

Nik-Zainal, S., Van Loo, P., et al., 2012. The life history of 21 breast cancers. *Cell*, 149(5), pp.994–1007.

Nishita, D.M. et al., 2009. Clinical trial participant characteristics and saliva and DNA metrics. *BMC medical research methodology*, 9, p.71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19874586.

Nowell, P. & Hungerford, D., 1960. National Academy of Sciences. *Science*, 132(3438), pp.1488–1501. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.132.3438.1488.

Nowell, P.C., 1976. The clonal evolution of tumor cell populations. *Science*, 194(4260),

pp.23–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/959840.

Nunes, A.P. et al., 2012. Quality of DNA extracted from saliva samples collected with the Oragene^{TM} DNA self-collection kit. *BMC Medical Research Methodology*, 12(1), p.65. Available at: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-65.

Oh, E. et al., 2015. Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples J. S. Castresana, ed. *PLOS ONE*, 10(12), p.e0144162. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4671711&tool=pmcentrez&rendertype=abstract.

Ohm, J.E. et al., 2007. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature genetics*, 39(2), pp.237–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17211412.

Ojha, J. et al., 2015. Deep sequencing identifies genetic heterogeneity and recurrent convergent evolution in chronic lymphocytic leukemia. *Blood*, 125(3), pp.492–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25377784.

Oscier, D.G. et al., 2013. The clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF CLL4 trial. *Blood*, 121(3), pp.468–75. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23086750.

Ouillette, P., Collins, R., Shakhan, S., Li, J., Peres, E., et al., 2011. Acquired genomic copy number aberrations and survival in chronic lymphocytic leukemia. *Blood*, 118(11), pp.3051–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21795749.

Ouillette, P. et al., 2013. Clonal evolution, genomic drivers, and effects of therapy in chronic lymphocytic leukemia. *Clin Cancer Res*, 19(11), pp.2893–2904. Available at:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=23620403.

Ouillette, P., Collins, R., Shakhan, S., Li, J., Li, C., et al., 2011. The prognostic significance of various 13q14 deletions in chronic lymphocytic leukemia. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 17(21), pp.6778–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21890456.

Oxford Molecular Diagnostics centre, 2017. Specialist Integrated Haematological Malignancy Diagnostic Service. *Molecular Haemato-Oncology*. Available at: http://www.oxford-translational-molecular-diagnostics.org.uk/content/molecular-haemato-oncology.

Paez, J.G., 2004. EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science*, 304(5676), pp.1497–1500. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1099314.

De Palma, M., Biziato, D. & Petrova, T. V., 2017. Microenvironmental regulation of tumour angiogenesis. *Nature Reviews Cancer*, 17(8), pp.457–474. Available at: http://www.nature.com/doifinder/10.1038/nrc.2017.51.

Pang, A.W.C. et al., 2014. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. *G3*, 4(1), pp.63–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3887540&tool=pmcentrez&rendertype=abstract.

De Paoli-Iseppi, R. et al., 2016. Comparison of whole-exome sequencing of matched fresh and formalin fixed paraffin embedded melanoma tumours: implications for clinical decision making. *Pathology*, 48(3), pp.261–266. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0031302515300076.

Parikh, S.A., Kay, N.E. & Shanafelt, T.D., 2014. How we treat Richter syndrome. *Blood*, 123(11), pp.1647–1657. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2013-11-516229.

Parsons, B., 2008. Many different tumor types have polyclonal tumor origin: Evidence and implications. *Mutation Research/Reviews in Mutation Research*, 659(3), pp.232–247. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1383574208000768.

Peña-Llopis, S. et al., 2012. BAP1 loss defines a new class of renal cell carcinoma. *Nature Genetics*, 44(7), pp.751–759. Available at: http://www.nature.com/doifinder/10.1038/ng.2323.

Petersen, B.L. et al., 2004. Fluorescence in situ hybridization on formalin-fixed and paraffin-embedded tissue: optimizing the method. *Applied immunohistochemistry & molecular morphology : AIMM*, 12(3), pp.259–65. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=15551741&retmode=ref&cmd=prlinks%5Cnpapers3://publication/uuid/7b462510-431d-44de-9b3d-279344666d73.

Petitjean, A. et al., 2007. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database. *Human Mutation*, 28(6), pp.622–629.

Pettitt, A.R., 2003. Mechanism of action of purine analogues in chronic lymphocytic leukaemia. *British Journal of Haematology*, 121(5), pp.692–702. Available at: http://doi.wiley.com/10.1046/j.1365-2141.2003.04336.x.

Piyasatian, N., Fernando, R.L. & Dekkers, J.C.M., 2008. Introgressing multiple QTL in breeding programmes of limited size. *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie*, 125(1), pp.50–6. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/18254826.

Pohl, A. & Beato, M., 2014. bwtool: a tool for bigWig files. *Bioinformatics*, 30(11), pp.1618–1619. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu056.

Popitsch, N., 2014. CODOC: Efficient access, analysis and compression of depth of coverage signals. *Bioinformatics*, 30(18), pp.2676–2677.

Pospisilova, S. et al., 2012. ERIC recommendations on TP53 mutation analysis in chronic lymphocytic leukemia. *Leukemia*, 26(7), pp.1458–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22297721.

Potente, M., Gerhardt, H. & Carmeliet, P., 2011. Basic and Therapeutic Aspects of Angiogenesis. *Cell*, 146(6), pp.873–887. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867411010099.

Pritchard, C.C. et al., 2016. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. *The New England journal of medicine*, 375(5), pp.443–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27433846.

Pritchard, C.C. et al., 2014. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *The Journal of molecular diagnostics : JMD*, 16(1), pp.56–67. Available at: http://dx.doi.org/10.1016/j.jmoldx.2013.08.004.

Puente, X.S. et al., 2015. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 526(7574), pp.519–524. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26200345%5Cnhttp://www.nature.com/doifinder/10.1038/nature14666.

Puente, X.S. et al., 2011. Whole-genome sequencing identifies recurrent mutations in chronic

lymphocytic leukaemia. *Nature*, 475(7354), pp.101–105. Available at: http://www.nature.com/doifinder/10.1038/nature10113.

Quesada, V. et al., 2011. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature genetics*, 44(1), pp.47–52. Available at: http://www.nature.com/articles/ng.1032.

Quesada, V., Ramsay, A.J. & Lopez-Otin, C., 2012. Chronic Lymphocytic Leukemia with SF3B1 Mutation. *New England Journal of Medicine*, 366(26), pp.2530–2530. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMc1204033.

Quintáns, B. et al., 2014. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Applied & translational genomics*, 3(3), pp.60–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27284505.

Raczy, C. et al., 2013. Genome analysis Isaac : ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*, 29(16), pp.2041–2043.

Radovich, M. et al., 2016. Clinical benefit of a precision medicine based approach for guiding treatment of refractory cancers. *Oncotarget*, 7(35), pp.56491–56500. Available at: http://www.oncotarget.com/fulltext/10606.

Ramsay, A.J. et al., 2013. POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet*, 45(5), pp.526–530. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23502782%5Cnhttp://www.nature.com/ng/journal/v45/n5/pdf/ng.2584.pdf.

Rashid, M. et al., 2013. Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics*, 29(17), pp.2208–2210. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt371.

Rasi, S. et al., 2016. Clinical impact of small subclones harboring NOTCH1 , SF3B1 or BIRC3 mutations in chronic lymphocytic leukemia. *Haematologica*, 101(4), pp.e135–e138. Available at: http://www.haematologica.org/lookup/doi/10.3324/haematol.2015.136051.

Rawstron, A.C. et al., 2008. Monoclonal B-Cell Lymphocytosis and Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 359(6), pp.575–583. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMoa075290.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D., 1997. GeneCards : kmegrat g information about genes , prote and diseases. *Trends in Genetics*, 13(163), p.4.

Rebhandl, S. et al., 2014. APOBEC3 signature mutations in chronic lymphocytic leukemia. *Leukemia*, 28(9), pp.1929–1932. Available at: http://www.nature.com/doifinder/10.1038/leu.2014.160.

Rehm, H.L., 2013. Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews Genetics*, 14(4), pp.295–300. Available at: http://www.nature.com/doifinder/10.1038/nrg3463.

Rennert, H. et al., 2016. Development and validation of a whole-exome sequencing test for simultaneous detection of point mutations, indels and copy-number alterations for precision cancer care. *npj Genomic Medicine*, 1(1), p.16019. Available at: http://www.nature.com/articles/npjgenmed201619.

Rigolin, G.M. et al., 2016. Extensive next-generation sequencing analysis in chronic lymphocytic leukemia at diagnosis: clinical and biological correlations. *Journal of Hematology & Oncology*, 9(1), p.88. Available at: http://jhoonline.biomedcentral.com/articles/10.1186/s13045-016-0320-z.

Rimmer, A. et al., 2014. Integrating mapping-, assembly- and haplotype-based approaches

for calling variants in clinical sequencing applications. *Nature genetics*, 46(November 2013), pp.1–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25017105.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24–26. Available at: http://www.nature.com/articles/nbt.1754.

Rogers, B.B. et al., 1990. Analysis of DNA in fresh and fixed tissue by the polymerase chain reaction. *The American journal of pathology*, 136(3), pp.541–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1877479&tool=pmcentrez&rendertype=abstract.

Roller, A. et al., 2013. A 13-Gene Panel Targeted To Investigate CLL By Next-Generation Amplicon Deep-Sequencing Can Be Successfully Implemented In Routine Diagnostics A. Kohlmann, ed. *Blood*, 122(21), p.867 LP-867. Available at: http://www.bloodjournal.org/content/122/21/867.abstract.

Roller, E. et al., 2016. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*, 32(15), pp.2375–2377. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw163.

Roos-Weil, D. et al., 2016. Targeted Sequencing for Clinical Investigation in Chronic Lymphocytic Leukemia. *Blood*, 128(22), p.3202 LP-3202. Available at: http://www.bloodjournal.org/content/128/22/3202.abstract.

Rose-Zerilli, M.J.J. et al., 2014. ATM mutation rather than BIRC3 deletion and/or mutation predicts reduced survival in 11q-deleted chronic lymphocytic leukemia: data from the UK LRF CLL4 trial. *Haematologica*, 99(4), pp.736–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24584352.

Rose-Zerilli, M.J.J. et al., 2016. Longitudinal copy number, whole exome and targeted deep

sequencing of “good risk” IGHV-mutated CLL patients with progressive disease. *Leukemia*, 30(6), pp.1301–1310. Available at: http://www.nature.com/doifinder/10.1038/leu.2016.10.

Rossi, D., Spina, V., et al., 2013. Association between molecular lesions and specific B-cell receptor subsets in chronic lymphocytic leukemia. *Blood*, 121(24), pp.4902–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23637131.

Rossi, D. et al., 2014. Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. *Blood*, 123(14), pp.2139–2147. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2013-11-539726.

Rossi, D. et al., 2014. Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. *Blood*, 123(14), pp.2139–2147. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2013-11-539726.

Rossi, D., Rasi, S., et al., 2013. Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. *Blood*, 121(8), pp.1403–1412.

Rossi, D. et al., 2011. Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: Association with progression and fludarabine-refractoriness. *Blood*, 118(26), pp.6904–6908.

Rossi, D. et al., 2009. The prognostic value of TP53 mutations in chronic lymphocytic leukemia is independent of Del17p13: implications for overall survival and chemorefractoriness. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15(3), pp.995–1004. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19188171.

Rossi, D., Gerber, B. & Stüssi, G., 2017. Predictive and prognostic biomarkers in the era of new targeted therapies for chronic lymphocytic leukemia. *Leukemia & Lymphoma*,

58(7), pp.1548–1560. Available at: https://www.tandfonline.com/doi/full/10.1080/10428194.2016.1250264.

Rothberg, J.M. et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), pp.348–352. Available at: http://www.nature.com/doifinder/10.1038/nature10242.

Samorodnitsky, E. et al., 2015. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Human Mutation*, 36(9), pp.903–914. Available at: http://doi.wiley.com/10.1002/humu.22825.

Samuels, Y., 2004. High Frequency of Mutations of the PIK3CA Gene in Human Cancers. *Science*, 304(5670), pp.554–554. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1096502.

Samuels, Y. & Waldman, T., 2010. Oncogenic Mutations of PIK3CA in Human Cancers. In pp. 21–41. Available at: http://link.springer.com/10.1007/82_2010_68.

Sargent, D.J. et al., 2005. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23(9), pp.2020–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15774793.

Saunders, C.T. et al., 2012. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14), pp.1811–1817. Available at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts271.

Saussele, S. et al., 2015. Impact of comorbidities on overall survival in patients with chronic myeloid leukemia: results of the randomized CML study IV. *Blood*, 126(1), pp.42–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25918346.

Schlesinger, Y. et al., 2007. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature Genetics*, 39(2), pp.232–236. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17200670.

Schnaiter, A. et al., 2013. NOTCH1, SF3B1, and TP53 mutations in fludarabine-refractory CLL patients treated with alemtuzumab: results from the CLL2H trial of the GCLLSG. *Blood*, 122(7), pp.1266–1271.

Schuh, A. et al., 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20), pp.4191–4196. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22915640.

Schwartz, R. & Schäffer, A.A., 2017. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4), pp.213–229. Available at: http://www.nature.com/doifinder/10.1038/nrg.2016.170.

Schweiger, M.R. et al., 2009. Genome-Wide Massively Parallel Sequencing of Formaldehyde Fixed-Paraffin Embedded (FFPE) Tumor Tissues for Copy-Number- and Mutation-Analysis I. K. Jordan, ed. *PLoS ONE*, 4(5), p.e5548. Available at: http://dx.plos.org/10.1371/journal.pone.0005548.

Shabihkhani, M. et al., 2014. The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. *Clinical Biochemistry*, 47(4–5), pp.258–266. Available at: http://dx.doi.org/10.1016/j.clinbiochem.2014.01.002.

Shen, J. et al., 2015. ARID1A Deficiency Impairs the DNA Damage Checkpoint and Sensitizes Cells to PARP Inhibitors. *Cancer discovery*, 5(7), pp.752–767.

Shendure, J., 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741), pp.1728–1732. Available at:

http://www.sciencemag.org/cgi/doi/10.1126/science.1117389.

Shin, H.-T. et al., 2017. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nature Communications*, 8(1), p.1377. Available at: http://www.nature.com/articles/s41467-017-01470-y.

Sholl, L.M. et al., 2016. Institutional implementation of clinical tumor profiling on an unselected cancer population. *JCI Insight*, 1(19). Available at: https://insight.jci.org/articles/view/87062.

Sidow, A. & Spies, N., 2015. Concepts in solid tumor evolution. *Trends in genetics : TIG*, 31(4), pp.208–14. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0168952515000189.

Sims, D. et al., 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2), pp.121–132. Available at: http://www.nature.com/doifinder/10.1038/nrg3642.

Skowronska, A. et al., 2012. Biallelic ATM Inactivation Significantly Reduces Survival in Patients Treated on the United Kingdom Leukemia Research Fund Chronic Lymphocytic Leukemia 4 Trial. *Journal of Clinical Oncology*, 30(36), pp.4524–4532. Available at: http://ascopubs.org/doi/10.1200/JCO.2011.41.0852.

Slamon, D.J. et al., 2001. Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *New England Journal of Medicine*, 344(11), pp.783–792. Available at: http://www.nejm.org/doi/abs/10.1056/NEJM200103153441101.

Smoley, S.A. et al., 2010. Standardization of fluorescence in situ hybridization studies on chronic lymphocytic leukemia (CLL) blood and marrow cells by the CLL Research Consortium. *Cancer Genetics and Cytogenetics*, 203(2), pp.141–148. Available at:

http://linkinghub.elsevier.com/retrieve/pii/S0165460810004516.

Soria, J.-C. et al., 2014. Phase I/IIa study evaluating the safety, efficacy, pharmacokinetics, and pharmacodynamics of lucitanib in advanced solid tumors. *Annals of Oncology*, 25(11), pp.2244–2251. Available at: https://academic.oup.com/annonc/article-lookup/doi/10.1093/annonc/mdu390.

Sottoriva, A. et al., 2015. A Big Bang model of human colorectal tumor growth. *Nat Genet*, 47(3), pp.209–216. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=25665006.

Spencer, D.H. et al., 2013. Comparison of Clinical Targeted Next-Generation Sequence Data from Formalin-Fixed and Fresh-Frozen Tissue Specimens. *The Journal of Molecular Diagnostics*, 15(5), pp.623–633. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1525157813000913.

Splinter, E. et al., 2011. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & Development*, 25(13), pp.1371–1383. Available at: http://genesdev.cshlp.org/cgi/doi/10.1101/gad.633311.

Stadler, M.B. et al., 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*. Available at: http://www.nature.com/doifinder/10.1038/nature10716.

Stankovic, T. & Skowronska, A., 2014. The role of ATM mutations and 11q deletions in disease progression in chronic lymphocytic leukemia. *Leukemia & lymphoma*, 55(6), pp.1227–39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23906020.

Steck, P.A. et al., 1997. Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers. *Nature Genetics*,

15(4), pp.356–362. Available at: http://www.nature.com/articles/ng0497-356.

Stephens, P.J. et al., 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1), pp.27–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21215367.

Stilgenbauer, S. et al., 2007. Clonal evolution in chronic lymphocytic leukemia: acquisition of high-risk genomic aberrations associated with unmutated VH, resistance to therapy, and short survival. *Haematologica*, 92(9), pp.1242–1245. Available at: http://www.haematologica.org/cgi/doi/10.3324/haematol.10720.

Stilgenbauer, S. et al., 2014. Gene mutations and treatment outcome in chronic lymphocytic leukemia: results from the CLL8 trial. *Blood*, 123(21), pp.3247–3255.

Stilgenbauer, S. et al., 2009. Subcutaneous alemtuzumab in fludarabine-refractory chronic lymphocytic leukemia: clinical results and prognostic marker analyses from the CLL2H study of the German Chronic Lymphocytic Leukemia Study Group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27(24), pp.3994–4001. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19597025.

Stockley, T.L. et al., 2016. Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial. *Genome medicine*, 8(1), p.109. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27782854.

Stover, E.H. et al., 2016. Biomarkers of Response and Resistance to DNA Repair Targeted Therapies. *Clinical Cancer Research*, 22(23), pp.5651–5660. Available at: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-16-0247.

Stratton, M.R. et al., 2009. The cancer genome. *Nature*, 458(7239), pp.719–724. Available at: http://www.nature.com/doifinder/10.1038/nature07943.

Sukhai, M.A. et al., 2016. A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. *Genetics in Medicine*, 18(2), pp.128–136. Available at: http://www.nature.com/doifinder/10.1038/gim.2015.47.

Sun, F. & Reichenberger, E.J., 2014. Saliva as a source of genomic DNA for genetic studies: review of current methods and applications. *Oral health and dental management*, 13(2), pp.217–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24984625.

Sutton, L.-A. et al., 2015. Targeted next-generation sequencing in chronic lymphocytic leukemia: a high-throughput yet tailored approach will facilitate implementation in a clinical setting. *Haematologica*, 100(3), pp.370–376. Available at: http://www.haematologica.org/cgi/doi/10.3324/haematol.2014.109777.

Suzuki, Y. et al., 2017. Multiregion ultra-deep sequencing reveals early intermixing and variable levels of intratumoral heterogeneity in colorectal cancer. *Molecular Oncology*, 11(2), pp.124–139. Available at: http://doi.wiley.com/10.1002/1878-0261.12012.

Swain, S.M. et al., 2015. Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic Breast Cancer. *New England Journal of Medicine*, 372(8), pp.724–734. Available at: http://www.nejm.org/doi/10.1056/NEJMoa1413513.

Szatkiewicz, J.P. et al., 2013. Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. *Molecular psychiatry*, 18(11), pp.1178–84. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23938935.

Takagi, M., 2017. DNA damage response and hematological malignancy. *International Journal of Hematology*, 106(3), pp.345–356. Available at: https://doi.org/10.1007/s12185-017-2226-0.

Talevich, E. et al., 2016. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology*, 12(4), p.e1004873.

Available at: http://dx.plos.org/10.1371/journal.pcbi.1004873.

Tam, C.S. et al., 2014. Long-term results of first salvage treatment in CLL patients treated initially with FCR (fludarabine, cyclophosphamide, rituximab). *Blood*, 124(20), pp.3059–3064. Available at: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2014-06-583765.

Tarabichi, M. et al., 2017. Neutral tumor evolution? *bioRxiv*. Available at: http://biorxiv.org/content/early/2017/06/30/158006.abstract.

Tattini, L., D'Aurizio, R. & Magi, A., 2015. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3. Available at: http://journal.frontiersin.org/Article/10.3389/fbioe.2015.00092/abstract.

Taylor, C. et al., 2015. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature genetics*, 47(7), p.717.

Thermo Fisher Scientific, 2015. Ion Ampliseq Cancer Hotspot Panel v2. Available at: https://assets.thermofisher.com/TFS-Assets/LSG/brochures/Ion-AmpliSeq-Cancer-Hotspot-Panel-Flyer.pdf.

Thermo Fisher Scientific, 2017. Ion AmpliSeq Designer. Available at: https://www.ampliseq.com.

Thompson, P.A. et al., 2015. Complex karyotype is a stronger predictor than del(17p) for an inferior outcome in relapsed or refractory chronic lymphocytic leukemia patients treated with ibrutinib-based regimens. *Cancer*, 121(20), pp.3612–3621. Available at: http://doi.wiley.com/10.1002/cncr.29566.

Thompson, P.A. et al., 2016. Fludarabine, cyclophosphamide, and rituximab treatment achieves long-term disease-free survival in IGHV-mutated chronic lymphocytic leukemia. *Blood*, 127(3), pp.303–309. Available at:

http://www.bloodjournal.org/cgi/doi/10.1182/blood-2015-09-667675.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), pp.273–282. Available at: http://doi.wiley.com/10.1111/j.1467-9868.2011.00771.x.

Timmerman, L., 2015. DNA Sequencing Market Will Exceed $20 Billion, Says Illumina CEO Jay Flatley. *Forbes*. Available at: https://www.forbes.com/sites/luketimmerman/2015/04/29/qa-with-jay-flatley-ceo-of-illumina-the-genomics-company-pursuing-a-20b-market/#1863e77642e7.

Tomczak, K., Czerwińska, P. & Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland)*, 19(1A), pp.A68-77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25691825.

Le Tourneau, C. et al., 2015. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The Lancet Oncology*, 16(13), pp.1324–1334. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1470204515001886.

Travis, W.D. et al., 2011. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 6(2), pp.244–85. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21252716.

Tsimberidou, A.-M. & Keating, M.J., 2009. Treatment of fludarabine-refractory chronic

lymphocytic leukemia. *Cancer*, 115(13), pp.2824–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19402170.

Tsuchida, N., Ohtsubo, E. & Ryder, T., 1982. Nucleotide sequence of the oncogene encoding the p21 transforming protein of Kirsten murine sarcoma virus. *Science*, 217(4563), pp.937–939. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.6287573.

Uzilov, A. V. et al., 2016. Development and clinical application of an integrative genomic approach to personalized cancer therapy. *Genome Medicine*, 8(1), p.62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27245685.

Valouev, A. et al., 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18(7), pp.1051–1063. Available at: http://www.genome.org/cgi/doi/10.1101/gr.076463.108.

Varghese, A.M. et al., 2010. Long Term Survival Report of the UKCLL02 Trial: A Phase II Study of Subcutaneous Alemtuzumab In Patients with Fludarabine Refractory CLL (on Behalf of the NCRI CLL Trials Sub-Group) [abstract]. *Blood*, 116(21), p.Abstract 922.

Venter, J.C. et al., 1998. Shotgun sequencing of the human genome. *Science (New York, N.Y.)*, 280(5369), pp.1540–2. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9644018.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11181995.

Verma, S. et al., 2012. Trastuzumab Emtansine for HER2-Positive Advanced Breast Cancer. *New England Journal of Medicine*, 367(19), pp.1783–1791. Available at: http://www.nejm.org/doi/10.1056/NEJMoa1209124.

Vickovic, S. et al., 2016. Massive and parallel expression profiling using microarrayed single-cell sequencing. *Nature Communications*, 7, p.13182. Available at:

http://www.nature.com/doifinder/10.1038/ncomms13182.

Viray, H. et al., 2013. A prospective, multi-Institutional diagnostic trial to determine pathologist accuracy in estimation of percentage of malignant cells. *Archives of Pathology and Laboratory Medicine*, 137(11), pp.1545–1549.

Vogelstein, B. et al., 2013. Cancer Genome Landscapes. *Science*, 339(6127), pp.1546–1558. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1235122.

Vollbrecht, C. et al., 2015. Comprehensive Analysis of Disease-Related Genes in Chronic Lymphocytic Leukemia by Multiplex PCR-Based Next Generation Sequencing C. Hatzis, ed. *PLOS ONE*, 10(6), p.e0129544. Available at: http://dx.plos.org/10.1371/journal.pone.0129544.

Wagle, N. et al., 2012. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer discovery*, 2(1), pp.82–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22585170.

Wang, J., Khiabanian, H., Rossi, D., Fabbri, G., et al., 2014. Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia. *eLife*, 3, pp.1–23. Available at: http://elifesciences.org/lookup/doi/10.7554/eLife.02869.

Wang, K. et al., 2015. TERT Promoter Mutations and TERT mRNA but Not FGFR3 Mutations Are Urinary Biomarkers in Han Chinese Patients With Urothelial Bladder Cancer. *The Oncologist*, 20(3), pp.263–269. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25657201.

Wang, K., Liu, T., Ge, N., et al., 2014. TERT promoter mutations are associated with distant metastases in upper tract urothelial carcinomas and serve as urinary biomarkers detected by a sensitive castPCR. *Oncotarget*, 5(23), pp.12428–39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25474136.

Wang, K., Liu, T., Liu, L., et al., 2014. TERT promoter mutations in renal cell carcinomas and upper tract urothelial carcinomas. *Oncotarget*, 5(7), pp.1829–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24742867.

Wang, K., Li, M. & Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16), pp.e164–e164. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq603.

Wang, L. et al., 2017. Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. *Genome Research*, 27(8), pp.1300–1311. Available at: http://genome.cshlp.org/lookup/doi/10.1101/gr.217331.116.

Wang, L. et al., 2011. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *The New England journal of medicine*, 365(26), pp.2497–506. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3685413&tool=pmcentrez&rendertype=abstract.

Wang, Y., Waters, J., Leung, M.L., Unruh, A., et al., 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513), pp.155–60. Available at: http://www.nature.com/doifinder/10.1038/nature13600.

Welch, J.S. et al., 2012. The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell*, 150(2), pp.264–278. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867412007775.

Wellcome trust sanger institute, 2017. The Cancer Gene Census (CGC). Available at: http://cancer.sanger.ac.uk/census/.

West, H. (Jack), 2017. Novel Precision Medicine Trial Designs. *JAMA Oncology*, 3(3), p.423. Available at:

http://oncology.jamanetwork.com/article.aspx?doi=10.1001/jamaoncol.2016.5299.

Williams, C. et al., 1999. A high frequency of sequence alterations is due to formalin fixation of archival specimens. *The American journal of pathology*, 155(5), pp.1467–1471.

Williams, M.J. et al., 2016. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3), pp.238–244. Available at: http://www.nature.com/doifinder/10.1038/ng.3489.

Wolf, S.F. et al., 1984. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, 81(9), pp.2806–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6585829.

Woo, H.G. et al., 2017. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nature Communications*, 8(1), p.839. Available at: http://www.nature.com/articles/s41467-017-00991-w.

Wood, H.M. et al., 2010. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Research*, 38(14), pp.e151–e151. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq510.

Wysoker, A., Tibbetts, K. & Fennell, T., 2015. Picard tool. Available at: http://picard.sourceforge.net.

Xi, R. et al., 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences*, 108(46), pp.E1128–E1136. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1110574108.

Yachida, S. et al., 2010. Distant metastasis occurs late during the genetic evolution of

pancreatic cancer. *Nature*, 467(7319), pp.1114–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20981102.

Yates, L.R. et al., 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine*, 21(7), pp.751–759. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26099045.

Yates, L.R. & Campbell, P.J., 2012. Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11), pp.795–806. Available at: http://www.nature.com/doifinder/10.1038/nrg3317.

Yu, L. et al., 2017. Survival of Del17p CLL Depends on Genomic Complexity and Somatic Mutation. *Clinical Cancer Research*, 23(3), pp.735–745. Available at: http://clincancerres.aacrjournals.org/lookup/doi/10.1158/1078-0432.CCR-16-0594.

Yu, X.-M. et al., 2016. Notch1 Signaling Regulates the Aggressiveness of Differentiated Thyroid Cancer and Inhibits SERPINE1 Expression. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 22(14), pp.3582–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26847059.

Yunis, J.J. et al., 1984. Recurrent chromosomal defects are found in most patients with non-Hodgkin's-lymphoma. *Cancer genetics and cytogenetics*, 13(1), pp.17–28. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6467179.

Zenz, T., Mertens, D., et al., 2010. From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nature Reviews Cancer*, 10(1), pp.37–50. Available at: http://dx.doi.org/10.1038/nrc2764.

Zenz, T., Eichhorst, B., et al., 2010. TP53 mutation and survival in chronic lymphocytic leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(29), pp.4473–9. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/20697090.

Zenz, T. et al., 2010. TP53 mutation profile in chronic lymphocytic leukemia: evidence for a disease specific profile from a comprehensive analysis of 268 mutations. *Leukemia*, 24(12), pp.2072–2079. Available at: http://www.nature.com/doifinder/10.1038/leu.2010.208.

Zhao, L. et al., 2017. Resequencing the Escherichia coli genome by GenoCare single molecule sequencing platform. *bioRxiv*. Available at: http://biorxiv.org/content/early/2017/07/13/163089.abstract.

Zhao, T. et al., 2015. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. *Molecular BioSystems*, 11(1), pp.126–136. Available at: http://xlink.rsc.org/?DOI=C4MB00478G.

Zhao, Z. et al., 2016. Evolution of multiple cell clones over a 29-year period of a CLL patient. , (May).

Zook, J.M. et al., 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3), pp.246–251. Available at: http://www.nature.com/nbt/journal/v32/n3/full/nbt.2835.html%5Cnhttp://www.nature.com/nbt/journal/v32/n3/pdf/nbt.2835.pdf.

Appendix A Overview of patients' characteristics

						Fresh frozen		FFPE	
Patient ID	**Age**	**Gender**	**Tissue of origin**	**Diagnosis**	**Date Surgery**	**Percentage invasive tumour in sample (%)**	**Percentage necrosis (%)**	**Percentage invasive tumour in sample (%)**	**Percentage necrosis (%)**
2	63	M	Renal	Clear cell renal cell carcinoma	07/03/2014	95	0	95	0
4	65	M	Renal	chromophobe renal cell carcinoma	18/03/2014	100	0	95	0
7	NA	M	Prostate	Acinar adenocarcinoma	28/04/2014	60	N/A	60	0
8	57	M	Prostate	Acinar adenocarcinoma	29/04/2014	50	0	50	0
9	71	M	Thoracic	squamous cell carcinoma	02/06/2014	100	30	80	5
11	71	F	Thoracic	small cell carcinoma	25/04/2014	95	0	95	0
16	65	F	Renal	clear cell RCC	20/05/2014	90	0	100	0
18	80	F	Thoracic	squamous cell carcinoma	27/08/2014	>40	N/A	40	5
20	71	M	CR	adenocarcinoma	09/06/2014	20	0	60	10
21	72	M	Thoracic	squamous cell carcinoma	04/07/2014	50	N/A	80	20
24	44	M	Prostate	prostatic acinar Adenocarcinoma	28/07/2014	>40	N/A	60	0
28	NA	M	Prostate	prostatic acinar adenocarcinoma	02/07/2014	>40	N/A	70	0
30	56	F	CR	adenocarcinoma	01/07/2014	60	0	65	10
31	81	M	CR	adenocarcinoma	22/07/2014	40	5	60	10
32	76	F	Renal	clear cell RCC	18/07/2014	95	0	95	10
34	77	F	CR	adenocarcinoma	14/07/2014	60	0	75	0
35	75	F	Thoracic	squamous cell carcinoma	25/07/2014	70	N/A	80	5
37	53	F	Breast	invasive ductal carcinoma	21/07/2014	60	N/A	100	0
39	73	M	CR	adenocarcinoma	23/07/2014	40	0	50	0
44	68	M	Renal	clear cell RCC	07/11/2014	70	N/A	70	0
45	53	F	Breast	invasive ductal carcinoma	04/08/2014	60	5	90	5
48	68	M	Renal	Renal cell carcinoma	13/08/2014	>40	N/A	>40	N/A
50	87	M	CR	adenocarcinoma	01/09/2014	40	N/A	50	5
53	40	F	Breast	pleomorphic invasive lobular carcinoma	18/08/2014	95	0	90	0

55	75	M	CR	adenocarcinoma	15/08/2014	80	0	60	5
62	73	M	CR	adenocarcinoma	05/09/2014	50	0	80	5
65	83	M	Renal	Type II papillary renal cell carcinoma	04/09/2014	95	0	95	0
66	59	M	CR	adenocarcinoma	11/09/2014	70	0	70	0
69	53	F	Breast	invasive carcinoma with ductal NST and papillary features	10/09/2014	50	10	80	0
74	68	F	Breast	invasive ductal carcinoma	17/09/2014	70	5	80	5
80	75	F	Breast	invasive ductal carcinoma	23/09/2014	60	5	70	15
82	56	F	Endometrial	endometrioid endometrial adenocarcinoma	24/09/2014	60	20	70	5
84	78	M	Renal	clear cell renal cell carcinoma	30/09/2014	95	0	90	10
89	NA	F	Endometrial	Endometrioid adenocarcinoma	15/04/2014	80	0	75	0
91	NA	F	Endometrial	N/A	N/A	50	20	75	0
93	NA	F	Endometrial	N/A	N/A	80	10	85	5
96	58	F	Breast	invasive metaplastic carcinoma with squamous and ductal NST	13/10/2014	90	5	90	0
98	89	F	Endometrial	carcinosarcoma and endometrioid adenocarcinoma	08/10/2014	80	N/A	75	0
100	50	M	Renal	clear cell renal cell carcinoma	18/10/2014	90	0	95	0
102	NA	M	Renal	N/A	N/A	>40	N/A	>40	N/A
103	65	M	Renal	clear cell renal cell carcinoma	21/10/2014	>40	N/A	>40	N/A
113	76	F	Endometrial	High grade endometrial serous adenocarcinoma	21/10/2014	95	5	>40	N/A
121	88	M	CR	adenocarcinoma	05/11/2014	85	0	80	0
140	62	F	CR	adenocarcinoma	28/11/2014	80	0	85	0
153	41	F	Breast	invasive ductal carcinoma	02/12/2014	60	0	45	0
154	72	F	Endometrial	Endometrioid adenocarcinoma	02/12/2014	85	0	80	0
170	64	F	Breast	invasive ductal carcinoma	15/12/2014	70	0	90	0
172	80	F	CR	adenocarcinoma	24/12/2014	60	N/A	80	5
215	85	F	Renal	Clear cell renal cell carcinoma	17/02/2015	90	0	95	0
222	58	F	Breast	invasive ductal carcinoma	17/02/2015	75	0	80	5
300	80	M	Renal	Clear cell renal cell carcinoma	20/05/2015	>40	N/A	>40	N/A
365	74	M	Renal	Clear cell renal cell carcinoma	01/07/2015	95	0	95	0

N/A = not available, CR= colorectal, RCC= renal cell carcinoma

Appendix B Coverage metrics for all samples

patient id	sample type	total number of reads	paired reads	PCR duplicates (%)	singletons (%)	average depth
2	FF	2,527,372,188	93.0	4.0	1.2	110.4
2	GL	964,717,368	92.8	2.0	1.2	42.1
2	FFPE	2,346,890,942	70.8	17.8	3.3	102.5
4	GL	940,130,258	92.4	2.3	1.2	41.1
4	FF	2,511,805,356	92.1	4.9	1.3	109.7
4	FFPE	2,299,778,458	85.6	18.2	2.0	100.5
7	GL	917,110,822	92.8	2.1	1.1	40.1
7	FF	2,295,702,778	92.6	4.0	1.2	100.3
7	FFPE	2,188,366,634	84.9	15.9	1.9	95.6
8	GL	892,384,664	92.7	2.0	1.2	39.0
8	FF	2,270,130,586	92.4	3.1	1.4	99.2
8	FFPE	2,426,632,978	73.1	21.4	3.1	106.0
9	FF	2,305,799,410	93.0	4.6	1.2	100.7
9	FFPE	2,401,226,914	91.8	8.8	1.3	104.9
9	GL	969,256,738	92.7	2.0	1.2	42.3
11	FF	2,435,417,314	93.7	4.8	1.0	106.4
11	FFPE	2,313,254,082	86.4	15.8	1.6	101.1
11	GL	1,009,905,044	93.7	2.2	1.0	44.1
16	FF	2,603,941,288	93.4	4.6	1.1	113.8
16	GL	958,468,048	93.3	2.1	1.2	41.9
16	FFPE	2,243,934,526	86.9	18.0	1.6	98.0
18	GL	1,000,709,758	92.7	2.1	1.4	43.7
18	FF	2,498,996,030	93.2	3.9	1.2	109.2
18	FFPE	2,465,714,112	92.4	6.4	1.3	107.7
20	FF	2,475,787,556	93.2	4.4	1.2	108.2
20	GL	921,580,490	93.1	2.0	1.1	40.3
20	FFPE	2,209,235,866	86.3	15.4	1.6	96.5
21	GL	1,179,535,012	92.8	2.6	1.2	51.5
21	FF	2,194,347,238	92.9	3.4	1.2	95.9

21	FFPE	2,344,349,308	91.4	10.4	1.3	102.4
24	FF	2,519,041,716	91.9	4.3	1.5	110.0
24	GL	1,048,754,202	92.6	2.4	1.3	45.8
24	FFPE	2,285,289,342	86.3	10.4	2.1	99.8
28	FF	2,661,264,148	92.6	4.1	1.2	116.3
28	GL	1,013,669,230	92.4	2.3	1.2	44.3
28	FFPE	2,165,360,148	89.0	8.8	1.5	94.6
30	FF	2,507,343,880	92.7	4.3	1.3	109.5
30	FFPE	2,368,262,360	91.8	16.8	1.3	103.5
30	GL	979,611,550	92.7	2.5	1.2	42.8
31	GL	895,790,522	93.0	2.1	1.2	39.1
31	FF	2,394,434,286	92.8	3.9	1.3	104.6
31	FFPE	2,291,471,168	89.1	13.2	1.6	100.1
32	GL	1,468,319,436	93.2	3.1	1.1	64.1
32	FF	2,409,053,552	93.1	5.6	1.2	105.2
32	FFPE	2,663,058,964	63.4	28.8	2.8	116.3
34	GL	1,006,224,936	93.0	2.7	1.1	44.0
34	FF	2,223,252,206	92.5	4.4	1.4	97.1
34	FFPE	2,148,679,288	88.6	17.9	1.6	93.9
35	FF	2,485,356,280	93.2	4.1	1.3	108.6
35	FFPE	2,279,399,946	92.2	6.6	1.3	99.6
35	GL	982,545,718	93.2	2.0	1.2	42.9
37	FF	2,416,634,088	93.4	3.8	1.1	105.6
37	FFPE	2,014,748,666	86.1	19.0	1.5	88.0
37	GL	905,368,044	93.1	2.0	1.1	39.6
39	FF	2,535,589,364	92.7	3.8	1.2	110.8
39	FFPE	2,206,438,452	86.1	16.8	1.8	96.4
39	GL	991,993,326	92.3	2.4	1.3	43.3
44	GL	1,010,507,394	92.9	2.1	1.3	44.1
44	FF	2,590,826,826	92.3	4.2	1.4	113.2
44	FFPE	2,528,139,526	88.1	9.1	1.8	110.4
45	FF	2,452,875,962	93.1	4.3	1.2	107.2
45	FFPE	2,294,319,594	90.9	10.8	1.4	100.2
45	GL	991,958,776	92.9	2.4	1.2	43.3
48	GL	915,142,616	93.0	2.5	1.1	40.0
48	FF	2,465,852,216	93.1	5.0	1.2	107.7
48	FFPE	2,245,537,880	88.0	16.9	1.6	98.1
50	GL	871,380,376	93.0	2.1	1.2	38.1
50	FF	2,299,786,140	92.7	3.6	1.1	100.5
50	FFPE	2,151,770,516	90.7	8.2	1.4	94.0
53	FF	2,501,551,528	92.5	3.8	1.4	109.3
53	GL	1,024,399,446	93.0	2.1	1.3	44.8

53	FFPE	2,122,558,702	87.0	10.0	1.7	92.7
55	FF	2,598,948,852	92.9	4.1	1.2	113.5
55	FFPE	2,024,932,304	86.6	11.8	1.9	88.5
55	GL	1,082,447,752	92.2	2.6	1.5	47.3
62	GL	916,697,228	92.0	2.5	1.3	40.0
62	FF	2,296,354,888	92.5	4.1	1.3	100.3
62	FFPE	2,209,355,276	85.6	10.3	2.0	96.5
65	GL	1,612,534,344	92.8	3.0	1.2	70.4
65	FF	3,261,558,974	93.2	4.6	1.2	142.5
65	FFPE	2,947,708,398	89.2	11.4	1.5	128.8
66	GL	1,270,006,436	92.6	2.4	1.2	55.5
66	FF	2,264,385,054	92.6	3.1	1.3	98.9
66	FFPE	3,028,563,456	59.0	33.3	3.2	132.3
69	GL	839,936,328	93.5	2.1	1.1	36.7
69	FF	2,453,283,424	93.6	3.7	1.1	107.2
69	FFPE	2,344,804,628	92.0	9.1	1.3	102.4
74	GL	848,776,468	93.1	2.1	1.3	37.1
74	FF	2,404,439,520	93.3	3.6	1.2	105.0
74	FFPE	2,452,929,996	88.1	13.7	1.8	107.2
80	FF	2,439,373,004	92.8	2.8	1.3	106.6
80	GL	1,028,160,050	93.1	2.3	1.2	44.9
80	FFPE	2,125,999,774	88.5	9.7	1.6	92.9
82	FF	2,732,010,696	92.3	4.1	1.4	119.4
82	GL	1,052,927,618	92.8	2.6	1.2	46.0
82	FFPE	2,124,331,122	85.2	9.4	2.1	92.8
84	FF	2,224,352,914	92.4	2.9	1.2	97.2
84	GL	1,049,542,246	92.1	2.1	1.4	45.9
84	FFPE	2,229,727,358	88.7	8.2	1.7	97.4
89	GL	877,886,598	93.1	2.2	1.2	38.4
89	FF	2,309,299,928	93.5	3.1	1.1	100.9
89	FFPE	2,887,988,232	77.8	14.5	2.6	126.2
91	GL	895,220,210	93.3	2.2	1.2	39.1
91	FF	2,323,779,720	93.7	3.3	1.1	101.5
91	FFPE	2,169,275,586	78.3	14.8	2.5	94.8
93	FF	2,385,325,112	94.3	3.0	0.9	104.2
93	GL	967,064,324	93.2	2.5	1.1	42.2
93	FFPE	2,286,151,202	91.8	7.9	1.3	99.9
96	FF	2,411,582,754	93.1	3.4	1.3	105.4
96	GL	944,517,030	92.6	2.2	1.3	41.3
96	FFPE	2,461,023,186	89.7	9.8	1.4	107.5
98	GL	1,013,796,592	92.7	2.3	1.4	44.3
98	FF	2,539,190,070	93.2	4.0	1.2	110.9

98	FFPE	2,529,628,574	90.4	8.9	1.3	110.5
100	FF	2,437,303,134	92.6	3.4	1.4	106.5
100	GL	938,102,308	93.0	2.1	1.1	41.0
100	FFPE	2,534,865,510	92.5	6.6	1.2	110.7
102	FF	2,521,597,694	93.8	3.7	1.0	110.2
102	GL	1,022,408,928	93.0	2.0	1.1	44.7
102	FFPE	2,571,390,362	85.8	8.8	1.7	112.3
103	GL	995,004,226	92.3	2.1	1.4	43.5
103	FF	2,473,477,420	92.8	3.8	1.2	108.1
103	FFPE	2,526,069,002	90.9	11.3	1.4	110.4
113	GL	969,346,181	92.5	26.3	1.6	42.3
113	FF	2,727,022,697	89.8	11.7	2.4	119.1
113	FFPE	2,649,104,563	89.0	16.1	1.9	115.7
121	GL	1,027,311,594	92.2	2.3	1.4	44.9
121	FF	2,052,494,252	92.8	3.2	1.2	89.7
121	FFPE	2,863,393,056	89.6	10.4	1.7	125.1
140	GL	976,078,256	92.3	30.4	1.8	42.6
140	FF	2,754,463,239	90.5	13.0	2.2	120.3
140	FFPE	2,754,397,538	90.2	19.4	1.8	120.3
153	GL	890,080,472	92.9	17.0	1.6	38.9
153	FF	2,868,312,914	93.2	22.8	1.4	125.3
153	FFPE	2,694,453,915	89.7	16.3	2.2	117.7
154	GL	910,686,718	92.8	17.6	1.7	39.8
154	FF	2,812,331,126	90.9	14.9	2.1	122.9
154	FFPE	2,742,962,319	89.8	20.8	2.0	119.8
170	GL	1,032,957,864	92.8	2.2	1.4	45.1
170	FF	2,082,544,826	93.4	3.7	1.1	91.0
170	FFPE	2,775,974,502	92.9	5.7	1.2	121.3
172	GL	848,289,173	92.8	10.7	1.6	37.1
172	FF	2,837,284,011	92.8	22.1	1.6	123.9
172	FFPE	2,741,519,484	91.7	17.5	1.5	119.8
215	FFPE	2,517,761,770	91.1	6.6	1.4	110.0
215	GL	906,738,416	92.4	14.4	1.8	39.6
215	FF	2,700,478,636	92.3	24.3	1.8	118.0
222	GL	861,699,749	92.5	22.3	1.8	37.6
222	FF	2,679,070,161	93.0	23.8	1.6	117.0
222	FFPE	2,777,253,554	90.1	22.0	1.7	121.3
300	FFPE	2,515,102,878	91.9	5.3	1.4	109.9
300	GL	939,545,826	93.1	2.4	1.1	41.0
300	FF	2,392,423,780	93.0	4.8	1.1	104.5
365	FFPE	2,434,929,354	92.8	5.9	1.1	106.4
365	GL	843,862,450	93.4	2.2	1.0	36.9

365	FF	2,192,429,282	93.8	6.0	0.9	95.8

FF: Fresh frozen, FFPE: formalin-fixed, paraffin-embedded; GL: Germline

Appendix C Description of the scopes used to investigate genome data

Scope	Description	Genome coverage
50 gene cancer panel & hotspots	Commercial cancer panel Ion AmpliSeq Cancer Hotspot Panel v2 (Thermo Fisher Scientific, Waltham, MA, USA)	<0.001%
Coding variants in COSMIC	Positions of somatic coding mutations reported to the Catalogue Of Somatic Mutations In Cancer (COSMIC)	<0.001%
Gencode Genes V19	GENCODE gene annotations, release 19	46%
Exonic regions	Targeted regions of the SeqCap EZ Exome Kit v3.0	2%
dbSNP-141	Genomic positions of variants reported in NCBI dbSNP Build 141	<0.001%
Low complexity regions	Regions of low sequence complexity as published in *(37)*	9%
Bad mappability regions	Regions of low read mappability as calculated by ARGOS *(38)*	23%
GIAB – reliable regions	Regions that produced reliable calls in the Genome in a bottle (GIAB) project	69%
Any scope labelled “NOT”	Report all genomic positions except the ones included in the specified scope	-

Appendix D Pairwise comparison of variant overlap in different tissues

Corrected p-values (Bonferroni) of pairwise comparisons using t-tests of percentage of SNV overlap between FF and FFPE of different tissue types (Blom transformation applied on data).

SNVs - Mutect

	Breast	CRC	Endometrial	Prostate	Renal
CRC	1	-	-	-	-
Endometrial	1	1	-	-	-
Prostate	0.10682	0.07945	0.99718	-	-
Renal	1	1	1	0.24082	-
Thoracic	0.26985	0.25276	0.04563	0.00072	0.05562

SNVs - Strelka

	Breast	CRC	Endometrial	Prostate	Renal
CRC	1	-	-	-	-
Endometrial	1	1	-	-	-
Prostate	1	1	1	-	-
Renal	1	1	1	1	-
Thoracic	0.304	0.077	0.017	0.02	0.045

SNVs - Shimmer

	Breast	CRC	Endometrial	Prostate	Renal
CRC	1	-	-	-	-
Endometrial	1	1	-	-	-
Prostate	1	1	1	-	-
Renal	1	1	1	1	-

Thoracic	0.027	0.0057	0.0326	0.0058	0.0294

SNVs - Combination

	Breast	CRC	Endometrial	Prostate	Renal
CRC	1	-	-	-	-
Endometrial	1	1	-	-	-
Prostate	1	1	1	-	-
Renal	1	0.362	1	1	-
Thoracic	0.323	0.023	0.411	0.279	1

Indels - Strelka

	Breast	CRC	Endometrial	Prostate	Renal
CRC	1	-	-	-	-
Endometrial	1	1	-	-	-
Prostate	1	1	1	-	-
Renal	1	0.227	1	1	-
Thoracic	0.176	0.023	0.122	0.335	1

Indels - Shimmer

	Breast	CRC	Endometrial	Prostate	Renal
CRC	1	-	-	-	-
Endometrial	1	1	-	-	-
Prostate	1	1	1	-	-
Renal	1	1	1	1	-
Thoracic	1	0.22	0.55	0.21	0.36

Appendix E Validation of somatic SNVs and indels detected in WGS data using Ion AmpliSeq Cancer Hotspot Panel v2

						FF panel		FF WGS					FFPE panel		FFPE WGS				
Gene	chr	position	ref	alt	Sample ID	detection	AF	evaluated	detection	AF	number of ref reads	number of alt reads	detection	AF	evaluated	detection	AF	number of ref reads	number of alt reads
FGFR3	4	1807871	A	G	4	yes	5	yes	*no*	*0*	*82*	*0*	N/A	N/A	yes	*no*	*0*	*336*	*0*
ERBB2	17	37881453	G	C	172	yes	5	yes	*no*	*0*	*93*	*0*	N/A	N/A	yes	*no*	*0*	*290*	*1*
ERBB2	17	37881453	G	C	170	yes	6	yes	*no*	*0*	*68*	*0*	N/A	N/A	yes	*no*	*0*	*187*	*0*
RET	10	43615572	A	T	4	yes	6	yes	*no*	*0*	*39*	*0*	N/A	N/A	yes	*no*	*0*	*158*	*0*
SMAD4	18	48584594	A	T	4	yes	6	yes	*no*	*0*	*71*	*0*	N/A	N/A	yes	*no*	*0*	*46*	*0*
SMAD4	18	48591897	G	C	4	yes	7	yes	*no*	*0*	*105*	*0*	N/A	N/A	yes	*no*	*0*	*50*	*0*
GNAQ	9	80409443	C	A	4	yes	7	yes	*no*	*0*	*141*	*0*	N/A	N/A	yes	*no*	*0*	*79*	*0*
EGFR	7	55259559	G	A	4	yes	8	yes	*no*	*0*	*93*	*0*	N/A	N/A	yes	*no*	*0*	*167*	*0*
APC	5	112175650	-	A	28	yes	14	yes	*no*	*0*	*59*	*0*	N/A	N/A	yes	*no*	*0*	*22*	*0*
PIK3CA	3	178952085	A	G	89	yes	48	yes	yes	52	46	49	no	0	no	*no*	*0*	*62*	*0*
VHL	3	10183772	CG	-	16	yes	65	yes	*no*	*0*	*39*	*19*	N/A	N/A	yes	*no*	*0*	*120*	*50*
TP53	17	7574027	C	A	4	yes	83	yes	yes	81	9	39	N/A	N/A	yes	yes	73	21	57
TP53	17	7577022	G	A	93	yes	68	yes	yes	59	20	29	N/A	N/A	yes	yes	52	78	86
PTEN	10	89717672	C	T	11	yes	12	yes	yes	7	54	4	no	0	no	*no*	*0*	*19*	*0*
TP53	17	7577046	C	A	9	yes	64	yes	yes	58	23	32	N/A	N/A	yes	yes	59	46	67
TP53	17	7577139	G	C	35	yes	63	yes	yes	63	26	25	N/A	N/A	yes	yes	60	32	48
TP53	17	7577528	G	A	91	yes	65	yes	yes	72	22	57	N/A	N/A	yes	yes	20	119	29
TP53	17	7577538	C	T	50	yes	37	yes	yes	27	44	16	N/A	N/A	yes	yes	72	28	71
TP53	17	7578190	T	C	69	yes	20	yes	yes	13	56	8	N/A	N/A	yes	yes	70	28	65
TP53	17	7578212	C	T	45	yes	44	yes	yes	40	29	19	N/A	N/A	yes	yes	64	44	78
TP53	17	7578212	G	C	11	yes	97	yes	yes	98	1	45	N/A	N/A	yes	yes	87	8	54
TP53	17	7578383	CCCCACCATGAGCGCTGC	-	98	yes	84	yes	yes	66	25	48	N/A	N/A	yes	yes	84	25	128
TP53	17	7578394	A	G	113	yes	66	yes	yes	50	43	43	N/A	N/A	yes	yes	18	224	50
TP53	17	7578406	C	T	39	yes	31	yes	yes	52	11	12	N/A	N/A	yes	yes	29	110	44
TP53	17	7578406	G	A	121	yes	59	yes	yes	70	11	26	N/A	N/A	yes	yes	74	48	136
TP53	17	7578413	C	T	31	yes	43	yes	yes	39	35	22	N/A	N/A	yes	yes	50	89	88
TP53	17	7578431	G	A	20	yes	55	yes	yes	59	18	26	N/A	N/A	yes	yes	73	38	105
TP53	17	7579361	A	G	30	yes	27	yes	yes	26	46	16	N/A	N/A	yes	yes	34	90	46
KRAS	12	25398281	C	T	39	yes	21	yes	yes	18	84	18	N/A	N/A	yes	yes	29	20	8
KRAS	12	25398281	C	T	50	yes	27	yes	yes	18	85	19	N/A	N/A	yes	yes	48	22	20
KRAS	12	25398281	G	A	55	yes	40	yes	yes	25	76	26	N/A	N/A	yes	yes	13	35	5
KRAS	12	25398284	C	A	30	yes	19	yes	yes	15	99	18	N/A	N/A	yes	yes	21	11	3
KRAS	12	25398284	G	T	121	yes	37	yes	yes	36	47	26	N/A	N/A	yes	yes	35	31	17
KRAS	12	25398284	C	A	93	yes	57	yes	yes	51	68	71	N/A	N/A	yes	yes	48	26	24
APC	5	112174631	C	T	172	yes	21	yes	yes	16	103	20	N/A	N/A	no	*no**	*0*	*20*	*0*
KRAS	12	25398284	G	T	172	yes	59	yes	yes	57	49	66	N/A	N/A	yes	yes	46	7	6
KRAS	12	25398284	C	A	66	yes	68	yes	yes	62	38	62	N/A	N/A	yes	yes	73	3	8
APC	5	112175347	TGAATTTTCTTCAGGAGC	T	30	yes	21	yes	yes	14	59	10	N/A	N/A	no	*no**	*0*	*60*	*5*
CTNNB1	3	41266103	G	C	154	yes	27	yes	yes	27	68	25	N/A	N/A	yes	yes	29	25	10
CTNNB1	3	41266137	C	T	98	yes	28	yes	yes	28	68	27	N/A	N/A	yes	yes	35	32	17

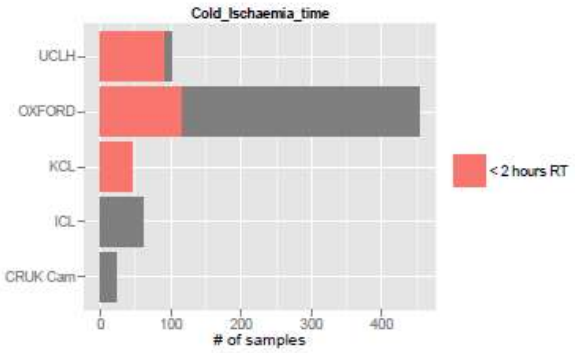
PIK3CA	3	178927974	G	A	172	yes	21	yes	yes	34	83	42	N/A	N/A	no	*no* *	*0*	*16*	*0*
SMAD4	18	48581209	G	A	89	no	0	no	*no*	*0*	*109*	*0*	N/A	N/A	no	yes *	12	68	8
SMAD4	18	48603033	G	A	140	yes	27	yes	yes	25	83	27	N/A	N/A	yes	yes	11	172	21
SMAD4	18	48603033	G	A	82	yes	33	yes	yes	34	86	45	N/A	N/A	yes	yes	17	97	20
SMAD4	18	48604665	G	A	31	yes	24	yes	yes	23	48	14	N/A	N/A	yes	yes	51	22	23
EGFR	7	55241717	C	T	82	yes	34	yes	yes	34	61	32	N/A	N/A	yes	yes	33	88	43
KDR	4	55961059	G	A	82	yes	30	yes	yes	36	58	32	N/A	N/A	yes	yes	21	22	13
RB1	13	48953841	C	T	8	no	0	no	*no*	*0*	*62*	*0*	N/A	N/A	no	yes *	11	122	15
PTEN	10	89624263	A	G	96	yes	45	yes	yes	13	73	11	N/A	N/A	yes	yes	30	205	86
PTEN	10	89685319	G	A	89	no	1	no	*no*	*0*	*110*	*1*	N/A	N/A	no	yes *	15	68	12
PTEN	10	89692905	G	A	82	yes	34	yes	yes	32	84	39	N/A	N/A	yes	yes	29	17	7
PTEN	10	89711899	C	T	82	yes	36	yes	yes	37	85	49	N/A	N/A	yes	yes	19	21	5
PTEN	10	89692905	G	A	89	no	1	no	*no*	*0*	*111*	*2*	N/A	N/A	no	yes *	12	66	9
PTEN	10	89720726	G	T	172	yes	21	yes	yes	62	33	53	N/A	N/A	yes	yes	41	10	7
APC	5	112173917	C	T	55	yes	33	yes	yes	35	73	39	N/A	N/A	yes	yes	13	52	8
TP53	17	7577127	C	G	74	yes	18	yes	yes	14	56	9	no	0	no	*no*	*0*	*56*	*0*
SMAD4	18	48581229	C	G	39	yes	27	yes	yes	33	62	30	no	0	no	*no*	*0*	*23*	*0*
APC	5	112175639	C	T	39	yes	40	yes	yes	31	61	27	N/A	N/A	yes	yes	40	31	21
APC	5	112175757	T	A	50	yes	27	yes	yes	31	50	22	N/A	N/A	yes	yes	65	11	20
PTPN11	12	112888198	G	A	140	yes	28	yes	yes	21	81	22	N/A	N/A	yes	yes	10	28	3
BRAF	7	140453136	A	T	34	yes	24	yes	yes	21	59	16	N/A	N/A	yes	yes	15	29	5
BRAF	7	140453136	A	T	31	yes	42	yes	yes	34	87	45	N/A	N/A	yes	yes	44	36	28
FBXW7	4	153249384	G	A	140	yes	28	yes	yes	66	34	65	N/A	N/A	yes	yes	84	6	32
NPM1	5	170837543	C	A	82	yes	45	yes	yes	40	64	42	N/A	N/A	yes	yes	37	27	6
PIK3CA	3	178916854	G	A	82	yes	34	yes	yes	30	92	39	N/A	N/A	yes	yes	21	23	6
PIK3CA	3	178921549	T	C	154	yes	26	yes	yes	33	83	40	N/A	N/A	yes	yes	24	26	8
PIK3CA	3	178921549	T	G	93	yes	78	yes	yes	75	31	92	N/A	N/A	yes	yes	51	25	26
KRAS	12	25398287	-	GAG	89	yes	30	yes	yes	35	63	34	no	0	no	*no*	*0*	*53*	*0*
PIK3CA	3	178936093	G	C	31	yes	27	yes	yes	32	65	30	N/A	N/A	yes	yes	29	42	17
PIK3CA	3	178952084	C	AT	170	yes	36	yes	yes	31	62	31	N/A	N/A	yes	yes	32	51	24
PIK3CA	3	178952085	A	G	74	yes	33	yes	yes	29	55	22	N/A	N/A	yes	yes	35	56	30
PIK3CA	3	178952085	A	G	80	yes	31	yes	yes	38	74	46	N/A	N/A	yes	yes	10	63	7
TP53	17	7579854	T	-	55	yes	39	yes	yes	39	23	15	yes	7	yes	*no*	*0*	*128*	*24*

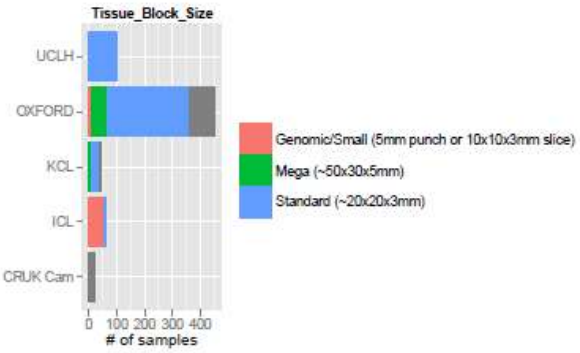
*in absence of data for cancer panel FFPE sample, we cannot rule out tumour heterogeneity between FF and FFPE. Variants highlighted in light grey were not detected in FF and FFPE by WGS, but were detected by targeted sequencing. One variant in dark grey was not detected in FFPE by WGS but was detected in FF by WGS and in both FF and FFPE by targeted sequencing.

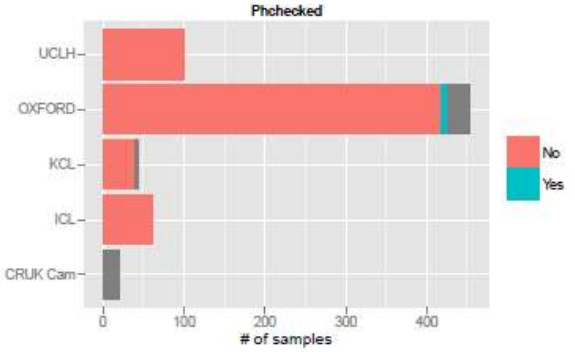
In case a variant was not detected in the sample, number of reference reads (ref reads) and alternate reads (alt reads) were derived from inspection in IGV. In case it was detected in the sample, these values were taken from the variant caller output.

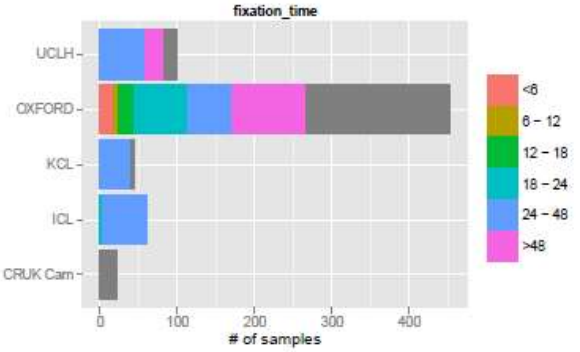
FF: Fresh frozen, FFPE: formalin-fixed, paraffin-embedded, Chr: chrosomome, position: hg19 coordinate, ref: reference allele, alt: alternate allele, AF: allelic fraction. N/A: not available

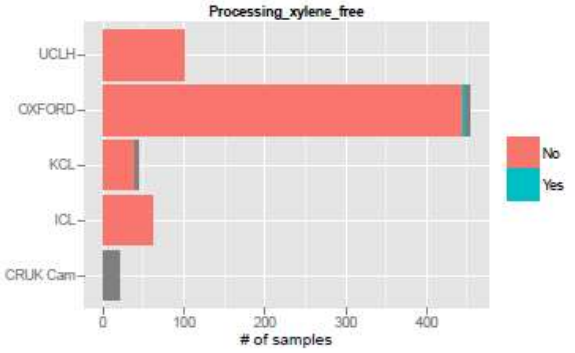
Appendix F Different conditions applied to the FFPE tissue preparation

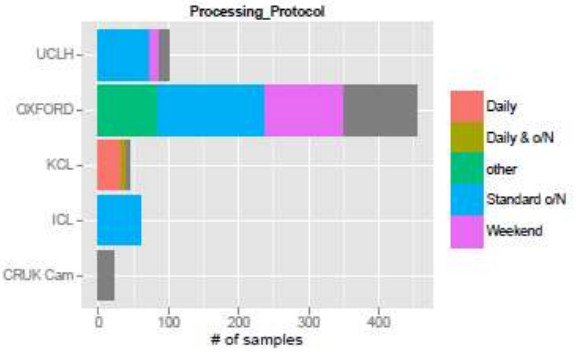

Cold_Ischaemia_time
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
< 2 hours RT

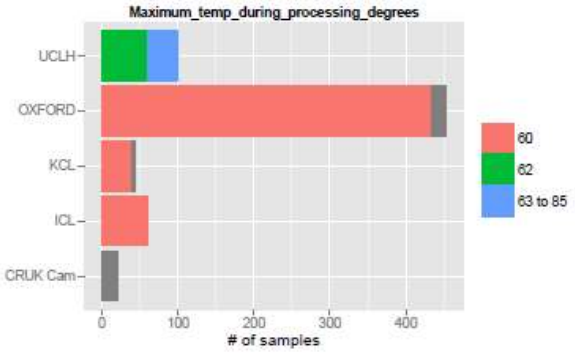

Tissue_Block_Size
UCLH
OXFORD
KCL
ICL
CRUK Cam
0 100 200 300 400
# of samples
Genomic/Small (5mm punch or 10x10x3mm slice)
Mega (~50x30x5mm)
Standard (~20x20x3mm)

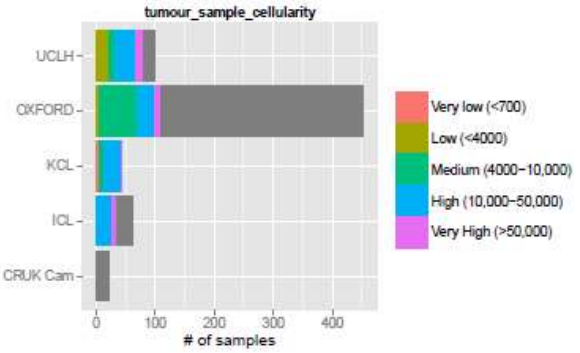

Phchecked
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
No
Yes

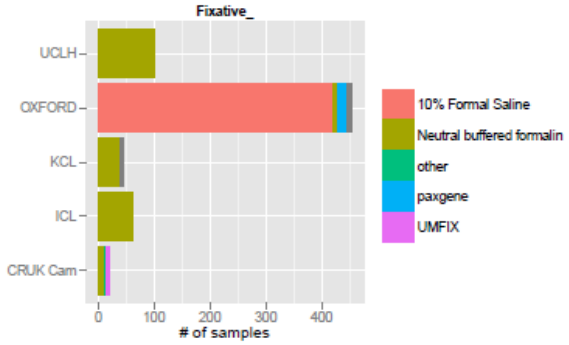

fixation_time
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
<6
6 – 12
12 – 18
18 – 24
24 – 48
>48

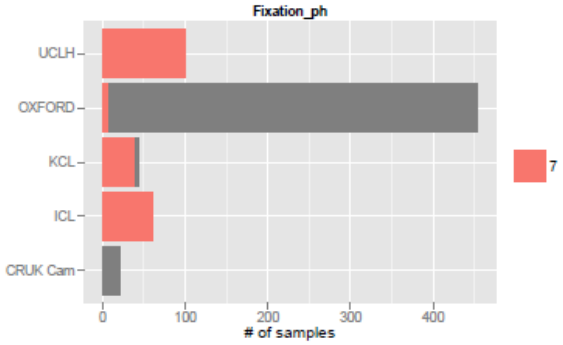

Processing_xylene_free
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
No
Yes

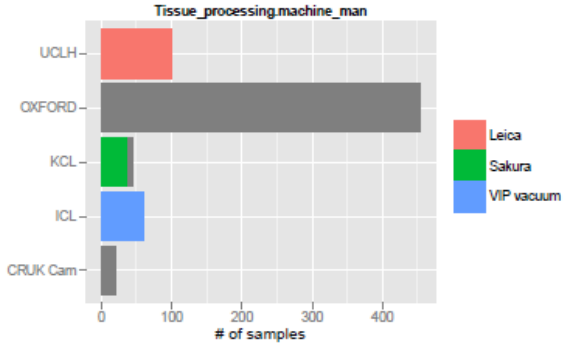

Processing_Protocol
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
Daily
Daily & o/N
other
Standard o/N
Weekend

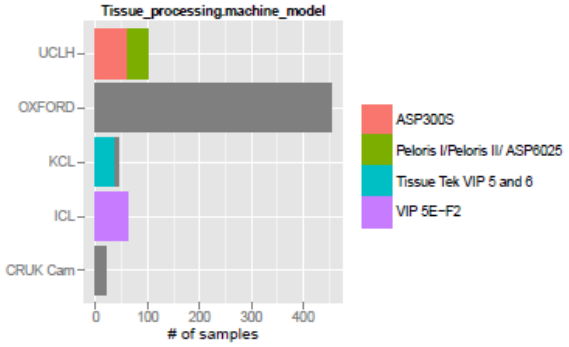

Maximum_temp_during_processing_degrees
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
60
62
63 to 85

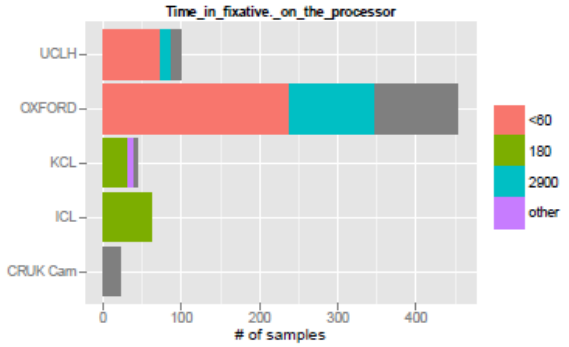

tumour_sample_cellularity
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
Very low (<700)
Low (<4000)
Medium (4000−10,000)
High (10,000−50,000)
Very High (>50,000)

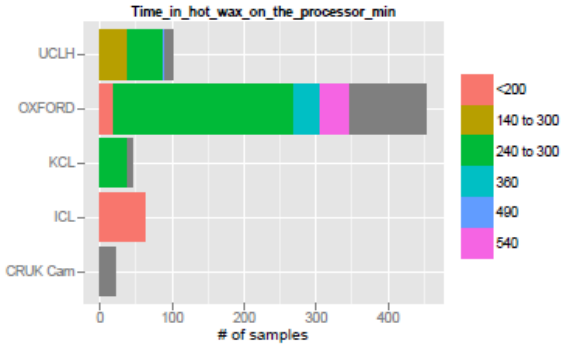

Fixative_
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
10% Formal Saline
Neutral buffered formalin
other
paxgene
UMFIX

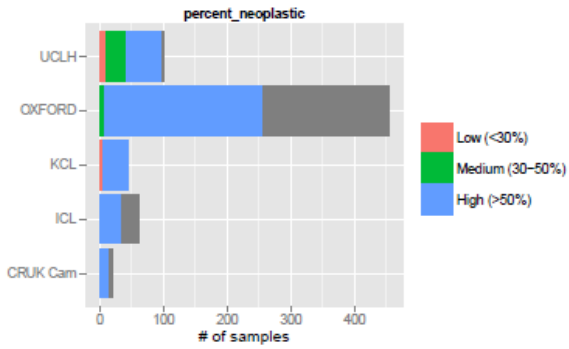

Fixation_ph
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
7

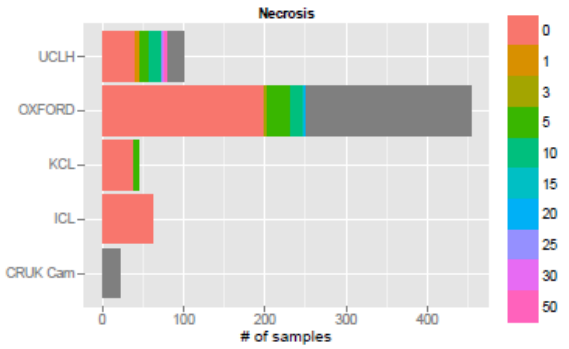

Tissue_processing.machine_man
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
Leica
Sakura
VIP vacuum

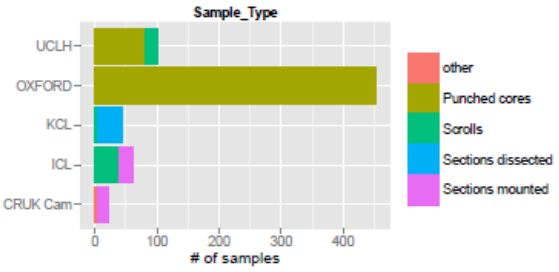

Tissue_processing.machine_model
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
ASP300S
Peloris I/Peloris II/ ASP6025
Tissue Tek VIP 5 and 6
VIP 5E-F2

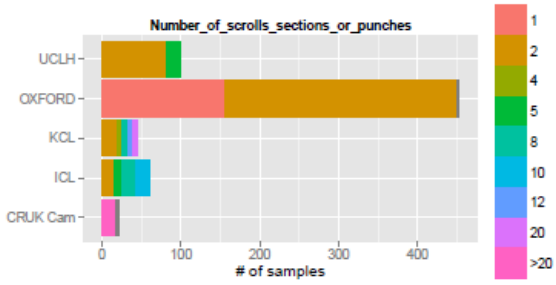

Time_in_fixative._on_the_processor
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
<60
180
2900
other

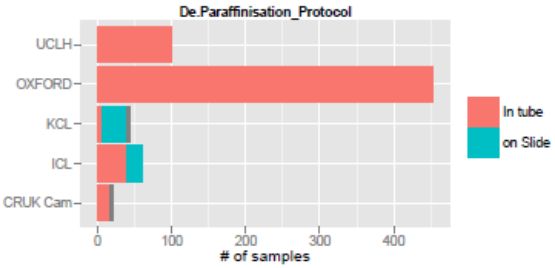

Time_in_hot_wax_on_the_processor_min
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
<200
140 to 300
240 to 300
360
490
540

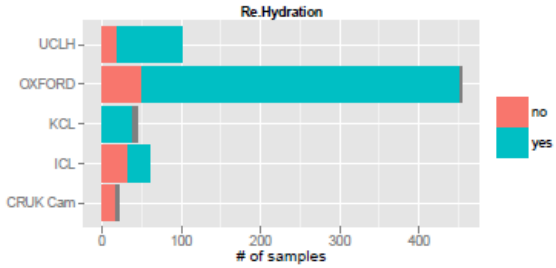

percent_neoplastic
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
Low (<30%)
Medium (30-50%)
High (>50%)

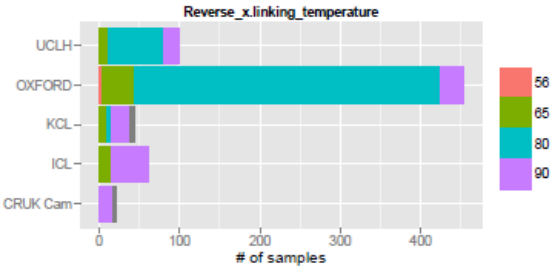

Necrosis
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
0
1
3
5
10
15
20
25
30
50

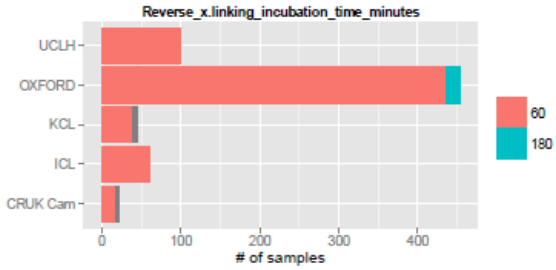

Sample_Type
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
other
Punched cores
Scrolls
Sections dissected
Sections mounted

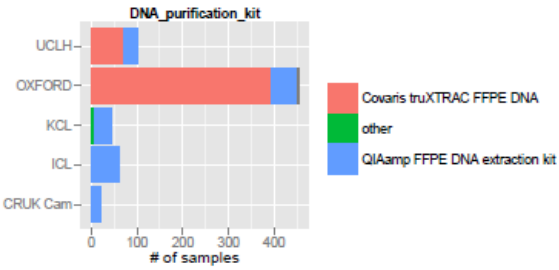

Number_of_scrolls_sections_or_punches
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
1
2
4
5
8
10
12
20
>20

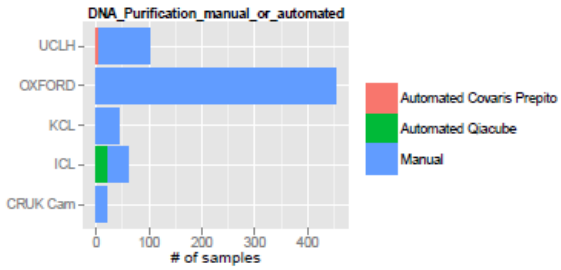

De.Paraffinisation_Protocol
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
In tube
on Slide

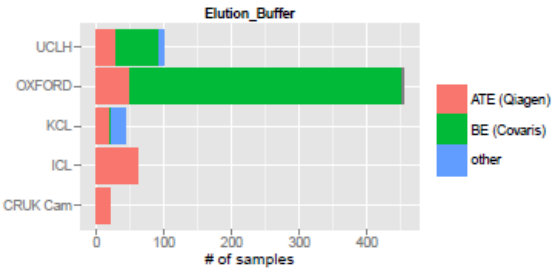

Re.Hydration
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
no
yes

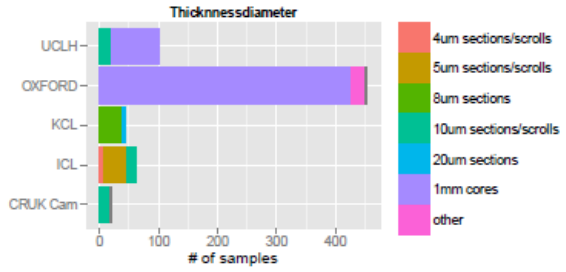

Reverse_x.linking_temperature
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
56
65
80
90

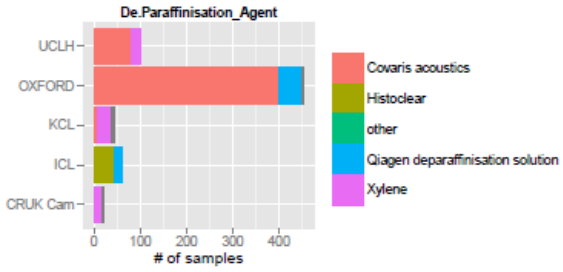

Reverse_x.linking_incubation_time_minutes
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
60
180

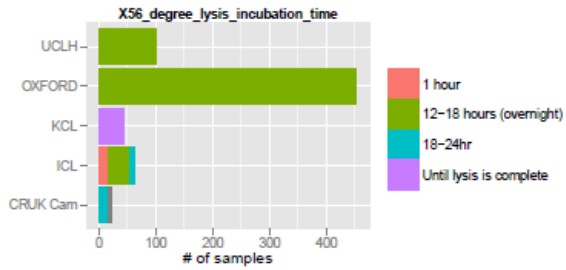

DNA_purification_kit
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
Covaris truXTRAC FFPE DNA
other
QIAamp FFPE DNA extraction kit

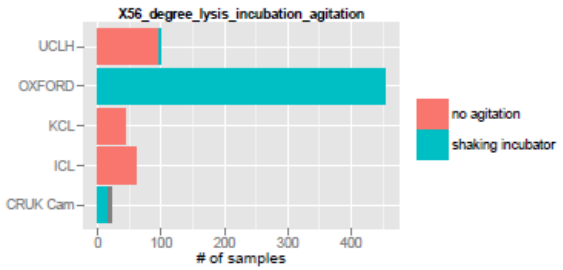

DNA_Purification_manual_or_automated
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
Automated Covaris Prepito
Automated Qiacube
Manual

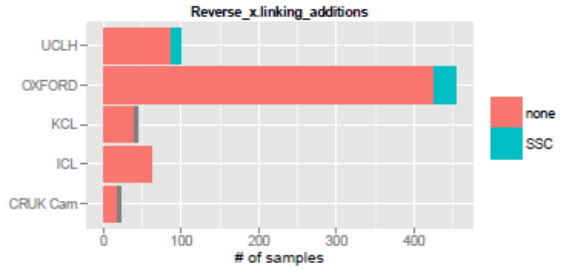

Elution_Buffer
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
ATE (Qiagen)
BE (Covaris)
other

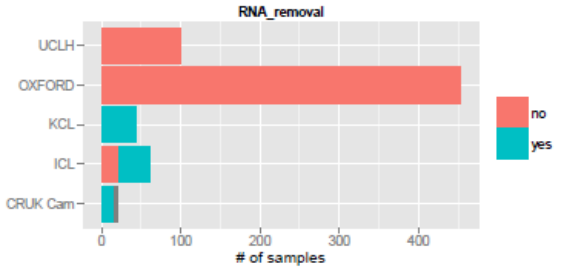

Thicknnessdiameter
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
4um sections/scrolls
5um sections/scrolls
8um sections
10um sections/scrolls
20um sections
1mm cores
other

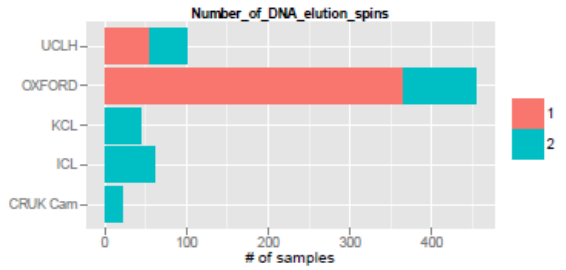

De.Paraffinisation_Agent
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
Covaris acoustics
Histoclear
other
Qiagen deparaffinisation solution
Xylene

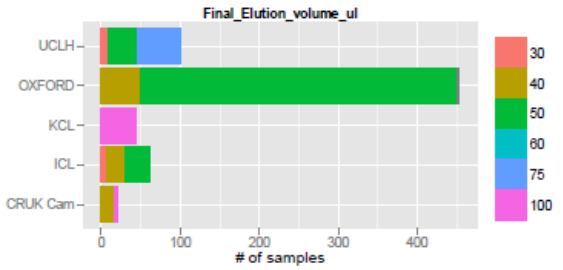

X56_degree_lysis_incubation_time
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
1 hour
12−18 hours (overnight)
18−24hr
Until lysis is complete


X56_degree_lysis_incubation_agitation
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
no agitation
shaking incubator


Reverse_x.linking_additions
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
none
SSC


RNA_removal
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
no
yes


Number_of_DNA_elution_spins
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
1
2


Final_Elution_volume_ul
UCLH
OXFORD
KCL
ICL
CRUK Cam
0
100
200
300
400
# of samples
30
40
50
60
75
100

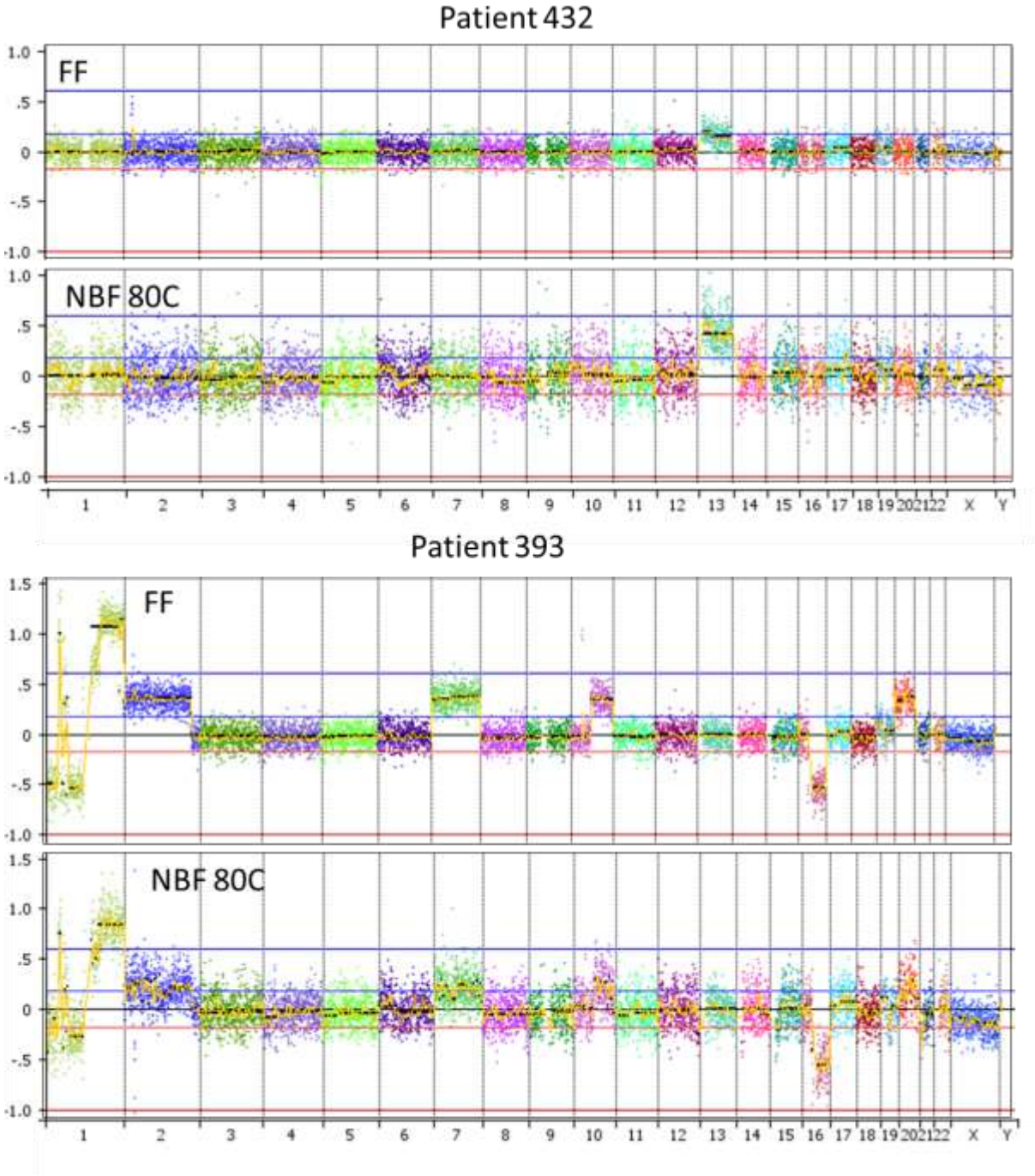
Appendix G DNA quality control results for FF, FFPE and blood samples

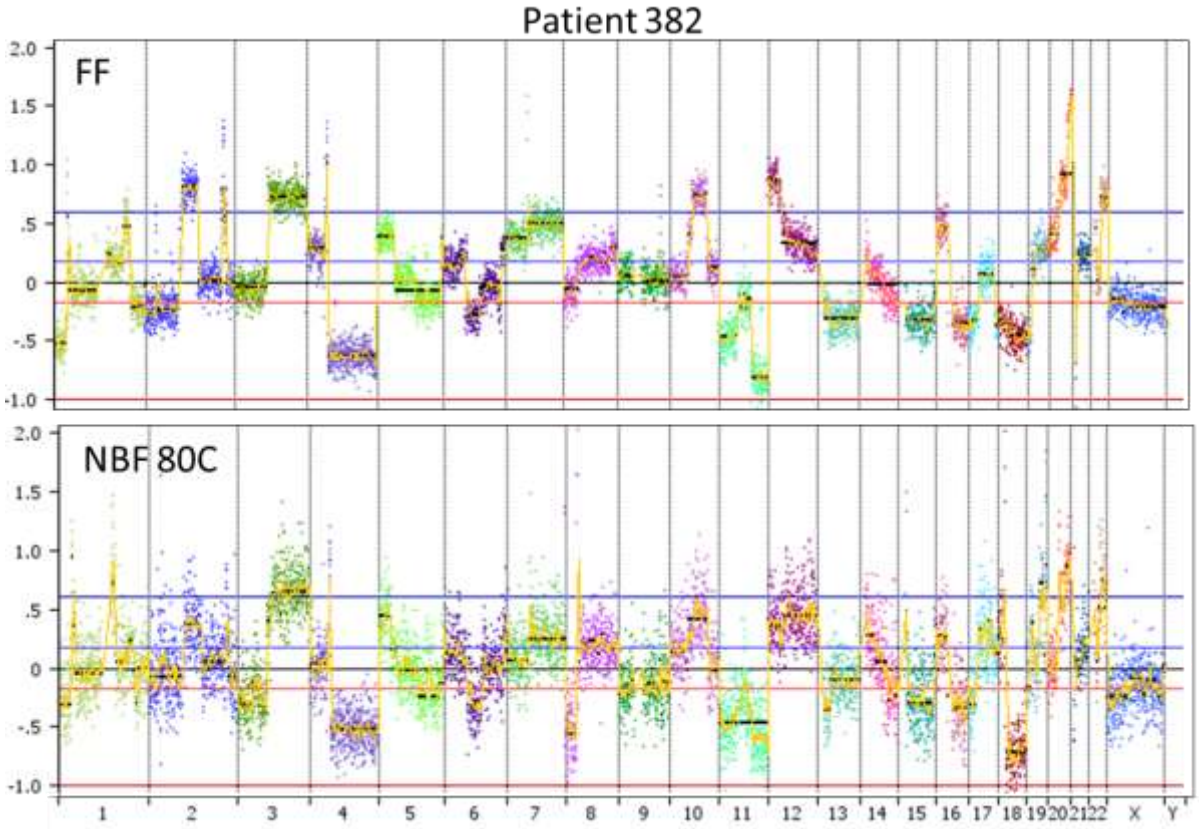
	FF			FFPE				Germline		
Patient ID	Yield assessed by qubit (μg)	Yield assessed by nanodrop (μg)	A260/A280	Yield assessed by qubit (μg)	Yield assessed by nanodrop (μg)	A260/A280	ΔCq	Yield assessed by qubit (μg)	Yield assessed by nanodrop (μg)	A260/A280
2	32.3	38.2	1.91	1.96	5.4	1.87	0.87	10.7	18.6	1.85
4	8.8	5.6	1.67	0.59	1.8	1.90	1.29	9.6	15.6	1.85
7	6.5	21.2	1.88	1.26	3.2	1.89	1.87	6.3	8.9	1.88
8	12.9	10.1	1.90	1.06	3.4	1.93	2.67	18.8	18.1	1.87
9	20.2	18.5	1.91	5.04	8.1	1.88	1.87	11.1	17.4	1.87
11	19.4	50.6	1.88	15.68	24.1	1.87	2.47	10.8	11.8	1.90
16	9.2	6.7	1.90	1.56	4.6	1.88	0.82	19.3	16.1	1.87
18	12.8	15.7	1.86	1.50	4.0	1.93	0.79	7.9	9.8	1.89
20	16.7	22.1	1.90	7.68	15.2	1.88	2.45	15.5	12.5	1.89
21	8.0	7.2	1.88	8.64	17.2	1.89	2.55	11.1	14.5	1.89
24	1.9	1.1	1.88	3.49	6.5	1.88	1.70	13.5	14.3	1.89
28	3.8	2.6	1.81	3.38	11.0	1.62	1.17	12.1	15.5	1.87
30	7.6	6.7	1.91	30.08	63.2	1.91	1.92	5.8	4.2	1.88
31	12.8	11.9	1.89	3.16	8.0	1.83	5.47	14.4	21.9	1.98
32	7.0	9.1	1.91	1.40	2.8	1.78	1.60	12.2	10.5	1.99
34	18.3	17.3	1.87	5.04	10.7	1.87	2.61	7.4	6.8	1.92
35	6.5	5.5	1.86	3.75	7.5	1.89	1.66	6.9	4.6	1.91
37	6.8	7.5	1.86	5.16	9.5	1.89	2.50	4.9	4.1	1.92
39	8.2	7.3	1.86	3.89	9.2	1.87	2.44	14.2	11.4	1.93
44	5.2	6.5	1.88	3.55	9.2	1.89	1.44	8.3	10.3	1.89
45	22.7	21.7	1.87	3.74	9.0	1.88	1.86	7.7	10.3	1.88
48	12.2	21.0	1.99	1.97	4.2	1.80	1.57	5.6	4.4	1.95
50	5.7	4.6	1.94	3.42	7.9	1.90	1.18	5.6	7.6	1.86
53	2.5	2.4	1.93	1.18	5.4	1.95	2.82	10.6	13.8	1.85

55	14.3	17.5	1.88	3.00	7.0	1.87	1.86	9.2	11.8	1.86
62	9.5	7.8	1.91	6.04	10.5	1.89	1.95	17.4	22.7	1.89
65	7.3	7.9	1.90	2.40	4.1	1.90	1.49	4.3	6.1	1.87
66	6.8	8.8	1.88	6.84	13.3	1.88	4.28	17.9	14.6	1.85
69	1.4	1.8	1.95	2.43	5.7	1.92	1.11	7.7	8.2	1.85
74	4.7	5.9	1.87	1.16	3.3	1.90	1.91	3.9	4.6	1.79
80	4.2	5.8	1.90	4.20	7.2	1.91	2.08	5.8	5.1	1.89
82	2.2	1.4	1.82	2.22	4.2	1.90	-1.52	7.8	8.1	1.86
84	11.3	11.9	1.87	1.49	5.1	1.89	0.03	3.6	4.5	1.90
89	1.1	1.4	1.84	2.26	8.7	1.91	2.60	7.6	9.8	1.84
91	18.4	16.6	1.88	9.44	20.5	1.88	2.50	3.9	10.3	1.62
93	17.1	18.1	1.88	6.00	18.5	1.92	1.48	1.7	2.7	1.94
96	1.9	5.5	1.86	0.51	1.6	1.88	0.06	8.1	9.4	1.92
98	38.6	42.0	1.83	9.68	18.7	1.88	1.55	15.9	17.7	1.88
100	7.4	7.8	1.84	0.76	2.0	1.91	1.10	9.5	11.5	1.86
102	6.0	4.9	1.84	0.70	3.1	1.97	0.93	10.7	15.0	1.85
103	3.4	5.8	1.86	1.32	3.0	1.87	1.58	9.2	13.0	1.85
113	55.0	68.2	1.85	4.00	11.0	1.91	1.42	12.1	16.9	1.85
121	9.3	10.3	1.88	4.12	11.6	1.88	1.55	21.5	26.4	1.85
140	14.9	13.7	1.86	5.68	14.3	1.89	1.65	8.1	8.0	1.90
153	2.1	2.5	1.93	2.73	7.1	1.90	1.73	24.8	24.9	1.86
154	17.1	25.2	1.91	3.90	12.2	1.88	0.73	16.3	16.5	1.86
170	1.1	1.3	1.91	2.50	5.8	1.87	0.13	14.7	14.2	1.87
172	7.5	8.0	1.86	4.44	11.4	1.89	0.98	14.1	15.0	1.85
215	3.2	4.4	1.85	1.20	3.0	1.95	1.85	17.9	18.3	1.85
222	8.2	8.8	1.85	2.48	6.6	1.89	1.79	16.5	21.5	1.83
300	7.7	10.3	1.86	0.86	1.7	1.94	2.47	10.4	13.6	1.87
365	18.7	23.0	1.87	1.15	2.6	1.91	1.19	26.2	24.7	1.85

FF: Fresh frozen, FFPE: formalin-fixed, paraffin-embedded, GL: germline

Appendix H Log_2R plots from Nexus of whole genome for experimental FFPE samples


Patient 432
FF
NBF 80C
1.0
.5
0
-.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Patient 393
FF
NBF 80C
1.5
1.0
.5
0
-.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y



Log_2R of matched FF and FFPE samples plotted across the whole genome.

Appendix I Clinical report of nine cancer cases

Table 0.1

Patient ID	Tier	Alteration	Cytoband / Genome position (hg19)	Protein consequence	Gene(s)	reported in FF	AF FF	DP FF	reported in FFPE	AF FFPE	DP FFPE
GeL004	1	Copy number loss	10p15.3 - q26.3	-	1411 genes (PTEN)	yes			yes		
GeL004	1	Missense variant	17:7574027	G334W	TP53	yes	0.81	48	yes	0.75	64
GeL004	1	Copy number loss	17p13.3 - q25.3	-	1876 genes (TP53)	yes			yes		
GeL004	1	Copy number loss	18p11.23 - q23	-	449 genes (PHLPP1)	yes			no		
GeL004	1	Copy number loss	1p36.33 - q44	-	3235 genes (ARID1A)	yes			yes		
GeL004	1	Missense variant	9:133333909	A99G	ASS1	yes	0.43	63	yes	0.39	57
GeL004	2	Frameshift variant	1:39751002-39751009	QNG437-439X	MACF1	yes	0.74	62	yes	0.82	49
GeL004	2	Copy number loss	13q11 - q34	-	747 genes (RB1)	yes			yes		
GeL004	2	Missense variant	15:91341555	D1116N	BLM	yes	0.44	117	yes	0.43	81
GeL004	2	Frameshift variant	16:30749939-30749939	R2860RX	SRCAP	yes	0.43	81	no	0.00	
GeL004	2	Stop codon gain	17:72233639	Y207*	TTYH2	yes	0.84	44	yes	0.61	33
GeL004	2	Copy number loss	18p11.32 - p11.22	-	88 genes (LAMA1)	yes			yes		
GeL004	2	Copy number loss	2p25.3 - q37.3	-	2122 genes (EPCAM)	yes			yes		
GeL004	2	Copy number loss	6p25.3 - q27	-	1862 genes (CDKN1A)	yes			yes		
GeL004	3	3 prime UTR variant	3:178957415	-	PIK3CA	no	0.00		yes	0.08	115
GeL004	3	Missense variant	8:93029512	M67I	RUNX1T1	yes	0.47	122	yes	0.59	99
GeL032	1	Stop codon gain	11:108175475	S1857*	ATM	yes	0.27	101	yes	0.34	67
GeL032	1	Splice donor variant	3:10183872	-	VHL	yes	0.33	24	yes	0.26	35
GeL032	1	Copy number loss	3p26.3-q11.1	-	797 genes (VHL, MLH1, SETD2, BAP1)	yes			yes		
GeL032	1	Copy number loss	9q21.11-q34.3	-	982 genes (TSC1)	yes			yes		
GeL032	2	Missense variant	1:205592872	I47V	ELK4	yes	0.27	97	yes	0.26	62
GeL032	2	Frameshift variant	11:61048362-61048362	R378PX	VWCE	yes	0.23	78	no	0.00	

GeL032	2	Missense variant	12:6145641	R820I	VWF	yes	0.27	94	yes	0.15	48
GeL032	2	Frameshift variant	14:99723932-99723932	S101SX	BCL11B	no	0.00		yes	0.50	68
GeL032	2	Frameshift variant	15:34355440-34355441	K175X	CHRM5	yes	0.25	84	yes	0.35	71
GeL032	2	Frameshift variant	19:36832341-36832343	I129X	ZFP14	yes	0.24	111	yes	0.32	50
GeL032	2	Frameshift variant	2:54155257-54155270	DFSKC496-500X	PSME4	yes	0.24	79	yes	0.21	53
GeL032	2	Stop codon gain	20:47266573	G997*	PREX1	no	0.00		yes	0.05	96
GeL032	2	Stop codon gain	20:60893902	R2347*	LAMA5	yes	0.23	61	yes	0.33	58
GeL032	2	Frameshift variant	3:108272574-108272575	S776X	KIAA1524	yes	0.26	62	yes	0.29	42
GeL032	2	Frameshift variant	3:49699734-49699735	L3486X	BSN	yes	0.32	44	yes	0.34	35
GeL032	2	Stop codon gain	4:103832611	R305*	SLC9B1	no	0.00	0	yes	0.37	102
GeL032	2	Stop codon gain	4:77660137	E271*	SHROOM3	yes	0.30	69	yes	0.35	48
GeL032	2	Frameshift variant	6:101253640-101253642	L306X	ASCC3	yes	0.29	110	yes	0.32	59
GeL032	2	Frameshift variant	6:2783660-2783661	I503X	WRNIP1	yes	0.33	43	yes	0.24	63
GeL032	2	Frameshift variant	8:99039817-99039817	E706EX	MATN2	yes	0.22	92	yes	0.33	63
GeL032	2	Frameshift variant	9:71006504-71006505	N251X	PGM5	yes	0.32	68	yes	0.30	37
GeL032	3	Splice region variant	12:49442550	-	KMT2D	yes	0.22	101	yes	0.25	71
GeL032	3	Missense variant	14:92472723	K533Q	TRIP11	yes	0.13	96	yes	0.26	66
GeL032	3	Copy neutral loss of heterozygocity	22q11.22 - q11.23	-	67 genes (BCR)	yes			no		
GeL032	3	3 prime UTR variant	3:178957371	-	PIK3CA	no	0.00		yes	0.12	69
GeL032	3	Copy number loss	4p16.3-q27	-	890 genes (FGFR3, PDGFRA, KIT, KDR, CXXC4, ABCG2)	yes			yes		
GeL032	3	3 prime UTR variant	9:80333166	-	GNAQ	no	0.00		yes	0.27	71
GeL032	3	3 prime UTR variant	9:80333235	-	GNAQ	no	0.00		yes	0.13	80
GeL065	1	Copy number loss	10p15.3 - q26.3	-	1270 genes (PTEN)	yes			yes		
GeL065	1	Copy number gain	12p13.33 - p11.1	-	443 genes (CCND2)	yes			yes		
GeL065	1	Copy number gain	12q11 - q24.33	-	1074 genes (CDK4, MDM2, PTPN11)	yes			yes		
GeL065	1	Copy number gain	17q12 - q25.3	-	1052 genes (BRCA1, ERBB2)	yes			yes		
GeL065	1	Copy number loss	1p36.33 - p11.2	-	1565 genes (ARID1A)	yes			yes		
GeL065	1	Copy number loss	22q11.21 - q13.33	-	736 genes (NF2, EP300)	yes			yes		
GeL065	1	Frameshift variant	3:47164465-47164465	R554PPX	SETD2	yes	0.84	135	yes	0.65	114
GeL065	1	Copy neutral loss of heterozygocity	3p26.3 - q29	-	1601 genes (VHL, MLH1, SETD2, BAP1)	yes			yes		
GeL065	1	Copy number gain	8p12 - q13.2	-	205 genes (FGFR1)	yes			yes		
GeL065	1	High copy gain	8q13.2 - q24.3	-	205 genes (MYC)	yes			yes		

GeL065	1	Copy number loss	9p24.3 - q34.3	-	1343 genes (TSC1, CDKN2A)	yes			yes		
GeL065	2	missense variant	10:101950709	M709V	CHUK	yes	0.51	97	yes	0.40	94
GeL065	2	Frameshift variant	10:115896972-115896973	N433X	C10orf118	yes	0.28	102	yes	0.20	104
GeL065	2	Frameshift variant	12:131590409-131590410	T629X	GPR133	yes	0.44	141	yes	0.34	142
GeL065	2	Copy number gain	16q11.2 - q24.3	-	555 genes (CDH1)	yes			yes		
GeL065	2	Stop codon gain	19:38951109	R819*	RYR1	no	0.00		yes	0.08	85
GeL065	2	Copy number loss	2p25.3 - q37.3	-	1994 genes (EPCAM)	yes			yes		
GeL065	2	Stop codon gain	4:85623606	W2932*	WDFY3	yes	0.50	76	yes	0.42	86
GeL065	2	Stop codon gain	5:110782393	E157*	CAMK4	yes	0.49	136	yes	0.30	89
GeL065	2	Frameshift variant	5:140603615-140603615	I554IX	PCDHB14	yes	0.34	144	yes	0.38	94
GeL065	2	Stop codon gain	5:140603618	K181*	PCDHB14	yes	0.38	141	yes	0.39	96
GeL065	2	Frameshift variant	6:100988045-100988045	V1923VX	ASCC3	yes	0.42	136	yes	0.43	99
GeL065	2	Frameshift variant	9:95013098-95013099	F775X	IARS	yes	0.51	87	yes	0.43	89
GeL065	3	Splice region variant	10:98369470	-	PIK3AP1	yes	0.37	73	yes	0.15	91
GeL065	3	Missense variant	13:41134882	S249Y	FOXO1	yes	0.46	136	yes	0.30	104
GeL065	3	Copy number gain	16p13.3 - p11.1	-	752 genes (TSC2, CREBBP)	yes			yes		
GeL065	3	3 prime UTR variant	17:37688355	-	CDK12	no	0.00	0	yes	0.07	180
GeL065	3	Splice region variant,intron variant	17:62499302	-	DDX5	yes	0.16	214	yes	0.12	144
GeL065	3	Copy number loss	18q11.1 - q23	-	319 genes (BCL2)	yes			yes		
GeL065	3	Copy number loss	19p13.3-p13.11	-	725 genes (MAP2K2)	yes			no		
GeL065	3	Splice acceptor variant	3:52682460	-	PBRM1	yes	0.85	128	yes	0.83	75
GeL065	3	Copy number loss	4p16.3 - q32.3	-	1029 genes (FGFR3, PDGFRA, KIT, KDR, IL2, FGF2, GAB1, IL21, CXXC4, ABCG2)	yes			yes		
GeL065	3	3 prime UTR variant	5:176525128	-	FGFR4	yes	0.39	110	yes	0.31	71
GeL065	3	missense variant	6:169648967	V52L	THBS2	yes	0.38	111	yes	0.41	133
GeL300	1	Frameshift variant	1:27089685-27089710	SGMCPPPGG881-889X	ARID1A	yes	0.34	56	yes	0.54	50
GeL300	1	Copy number loss	1p36.33 - p32.3	-	977 genes (ARID1A)	yes			yes		
GeL300	1	Copy number loss	3p26.3 - p11.1	-	736 genes (VHL, SETD2, BAP1, MLH1)	yes			yes		
GeL300	1	Copy number loss	3q21.3	-	35 genes (GATA2)	yes			yes		
GeL300	1	Copy number gain	5q23.2 - q31.1	-	89 genes (IL4)	yes			yes		
GeL300	1	Copy number gain	5q31.1 - q35.3	-	581 genes (PDGFRB, FLT4)	yes			yes		
GeL300	1	Missense variant,splice region variant	9:135787669	G305R	TSC1	yes	0.00		yes	0.00	

GeL300	1	Copy neutral loss of heterozygocity	9q21.11 - q34.3	-	901 genes (TSC1)	yes			yes		
GeL300	2	Stop codon gain	12:41337427	E470*	CNTN1	yes	0.32	123	yes	0.13	52
GeL300	2	Frameshift variant	13:99038021-99038032	TKIN238-241X	FARP1	no	0.00		yes	0.37	87
GeL300	2	Frameshift variant	15:76430085-76430085	I26IX	C15orf27	yes	0.26	74	yes	0.31	100
GeL300	2	Frameshift variant	16:1691169-1691171	N270X	CRAMP1L	yes	0.34	79	yes	0.31	83
GeL300	2	Inframe deletion	17:57759018-57759027	EHIG1087-1090E	CLTC	yes	0.00		yes	0.00	
GeL300	2	Frameshift variant	17:61623021-61623021	S915SX	KCNH6	no	0.00		yes	0.36	70
GeL300	2	Frameshift variant	18:72344519-72344536	SLTVKPA515-521SX	ZNF407	yes	0.29	77	yes	0.30	96
GeL300	2	Stop codon gain	19:15350238	K1181*	BRD4	yes	0.23	57	yes	0.28	50
GeL300	2	Copy neutral loss of heterozygocity	1q21.1 - q44	-	1427 genes (FH)	yes			yes		
GeL300	2	Frameshift variant	3:193029729-193029730	E774X	ATP13A5	yes	0.27	104	yes	0.33	85
GeL300	2	Frameshift variant	3:4725087-4725088	L1045X	ITPR1	yes	0.33	64	yes	0.52	65
GeL300	2	Frameshift variant	3:52389002-52389003	F1209X	DNAH1	yes	0.40	42	yes	0.45	42
GeL300	2	Copy number loss	3q11.1 - q21.3	-	263 genes (DIRC2)	yes			yes		
GeL300	2	Frameshift variant	6:157519976-157519977	Q1349X	ARID1B	yes	0.21	86	yes	0.35	97
GeL300	2	Frameshift variant	8:21853069-21853069	-768-769X	XPO7	yes	0.33	82	yes	0.40	86
GeL300	2	Stop codon gain	8:42563936	W43*	CHRNB3	yes	0.27	75	yes	0.29	77
GeL300	3	Missense variant	1:231503330	V401I	EGLN1	no	0.00		yes	0.09	91
GeL300	3	Missense variant	19:15350239	Q1180H	BRD4	yes	0.23	57	yes	0.28	50
GeL300	3	3 prime UTR variant	22:22111400	-	MAPK1	yes	0.30	60	yes	0.33	83
GeL300	3	3 prime UTR variant	3:178954226	-	PIK3CA	yes	0.28	111	yes	0.37	86
GeL300	3	3 prime UTR variant	3:178954228-178954231	-	PIK3CA	yes	0.26	111	yes	0.34	83
GeL300	3	5 prime UTR variant	5:1295133	-	TERT	yes	0.00		yes	0.00	
GeL365	1	Copy number loss	10p15.3 - q26.3	-	1270 genes (PTEN)	yes			yes		
GeL365	1	Copy number loss	11p15.5 - q25	-	1929 genes (ATM)	yes			yes		
GeL365	1	Copy number loss	17p13.3 - q25.3	-	1786 genes (TP53)	yes			yes		
GeL365	1	Copy number loss	18p11.32 - q23	-	509 genes (PHLPP1)	yes			yes		
GeL365	1	Copy number loss	1p36.33 - p22.2	-	1255 genes (ARID1A)	yes			yes		
GeL365	1	Copy number loss	1q22 - q44	-	1025 genes (AKT3, EGLN1, PRCC, TGFB2)	yes			yes		
GeL365	1	Missense variant	2:25470583	W297C	DNMT3A	yes	0.42	55	yes	0.61	64
GeL365	1	Missense variant	3:10188245	V130F	VHL	yes	0.75	51	yes	0.69	54
GeL365	1	Copy number loss	3p26.3 - q26.1	-	1297 genes (VHL, MLH1, SETD2, BAP1)	yes			yes		
GeL365	1	Copy number gain	5p15.33 - q35.3	-	1421 genes (PDGFRB, PIK3R1)	yes			yes		

GeL365	1	Copy number loss	9p24.3 - q34.3	-	1343 genes (TSC1, CDKN2A)	yes			yes		
GeL365	2	Missense variant	10:88659604	L129F	BMPR1A	yes	0.47	81	yes	0.42	90
GeL365	2	Frameshift variant	12:49177122-49177123	G32X	ADCY6	yes	0.36	90	yes	0.41	70
GeL365	2	Frameshift variant	14:56078925-56078925	-159-160X	KTN1	yes	0.39	75	yes	0.23	75
GeL365	2	Frameshift variant	18:51886964-51886965	H8X	C18orf54	yes	0.48	83	yes	0.52	104
GeL365	2	Frameshift variant	2:44201374-44201375	I357X	LRPPRC	yes	0.30	98	yes	0.21	100
GeL365	2	Copy number loss	2p25.3 - q37.3	-	1994 genes (EPCAM)	yes			yes		
GeL365	2	Missense variant	3:10116284	D929V	FANCD2	yes	0.20	55	no	0.00	
GeL365	2	Copy number loss	6p25.3 - q27	-	1707 genes (CDKN1A)	yes			yes		
GeL365	2	Copy number loss	8p23.3 - q24.3	-	1105 genes (TNFRSF10B, CLU)	yes			yes		
GeL365	2	Missense variant	9:139391076	R2372Q	NOTCH1	yes	0.28	32	yes	0.29	38
GeL365	2	Stop codon gain	MT:7321	W473*	MT-CO1	yes	0.24	9958	yes	0.15	10127
GeL365	3	3 prime UTR variant	1:204524399	-	MDM4	yes	0.45	60	yes	0.32	72
GeL365	3	Copy number loss	14p13 - q32.33	-	1147 genes (AKT1, HIF1A)	yes			yes		
GeL365	3	Missense variant	17:37686994	A1300T	CDK12	yes	0.31	54	yes	0.23	86
GeL365	3	Copy number loss	1p22.2 - q22	-	723 genes (MTOR, NRAS)	yes			yes		
GeL365	3	Copy number loss	20p13 - q13.33	-	847 genes (MMP9)	yes			yes		
GeL365	3	Copy number loss	3q26.1 - q29	-	305 genes (PIK3CA, PAK2)	yes			yes		
GeL365	3	Copy number loss	4p16.3 - q35.2	-	1167 genes (FGFR3, PDGFRA, KIT, KDR, IL2, FGF2, GAB1, IL21, CXXC4, ABCG2)	yes			yes		
GeL365	3	Missense variant	5:176562876	F258L	NSD1	yes	0.22	146	yes	0.32	123
GeL365	3	Copy number loss	7p22.3 - q36.3	-	1647 genes (BRAF, MET, EGFR, RAC1, HGF, ABCB1, PIK3CG, HILPDA, EZH2)	yes			yes		
GeL365	3	Missense variant	9:15474025	D280E	PSIP1	yes	0.34	58	yes	0.38	65
GeL382	2	Missense variant	12:56087864	H228R	ERBB3	yes	0.20	70	yes	0.43	260
GeL382	2	Frameshift variant	14:92004946	K1010KX	TRIP11	no	0.00		yes	0.05	85
GeL382	2	Frameshift variant	16:30123172	E13GX	MAPK3	no	0.00		yes	0.08	147
GeL382	2	Missense variant	19:18156035	Q52H	PIK3R2	yes	0.32	66	yes	0.25	135

GeL382	2	Missense variant	22:19188057	R1453Q	CLTCL1	no	0.00		yes	0.03	211
GeL382	2	Missense variant	5:75584773	G358V	POLK	yes	0.28	61	yes	0.17	122
GeL382	2	Stop codon gain	X:47147417	W44*	RBM10	yes	0.25	57	yes	0.13	170
GeL382	3	3 prime UTR variant	1:243504636	-	AKT3	yes	0.08	60	yes	0.03	65
GeL382	3	Missense variant	11:118506200	Q3436H	KMT2A	yes	0.20	55	yes	0.26	152
GeL382	3	5 prime UTR variant	12:56080106	-	ERBB3	no	0.00		yes	0.03	284
GeL382	3	Missense variant	12:57749208	E256K	CDK4	yes	0.19	74	yes	0.14	291
GeL382	3	3 prime UTR variant	12:68840695	-	MDM2	no	0.00		yes	0.03	343
GeL382	3	Missense variant	14:99174542	S765L	BCL11B	yes	0.11	47	no	0.00	
GeL382	3	3 prime UTR variant	15:87863974	-	NTRK3	no	0.00		yes	0.11	62
GeL382	3	Missense variant	16:3773928	M762I	CREBBP	yes	0.16	83	yes	0.20	228
GeL382	3	Missense variant	16:3773928	M762I	CREBBP	yes	0.16	83	yes	0.20	228
GeL382	3	Splice acceptor variant	17:40412972	-	TOP2A	yes	0.24	58	yes	0.27	90
GeL382	3	Missense variant	17:50189472	E912Q	COL1A1	yes	0.50	52	yes	0.18	177
GeL382	3	3 prime UTR variant	19:45350802	-	ERCC2	no	0.00		yes	0.05	222
GeL382	3	Missense variant	19:6216447	G422V	MLLT1	no	0.00		yes	0.22	109
GeL382	3	Missense variant	22:19233306	E461K	CLTCL1	no	0.00		yes	0.13	198
GeL382	3	Missense variant	22:41164060	E1246Q	EP300	yes	0.46	94	yes	0.24	156
GeL382	3	3 prime UTR variant	5:80875983	-	MSH3	yes	0.27	59	yes	0.16	83
GeL382	3	Missense variant	6:128067649	P676L	PTPRK	yes	0.06	50	yes	0.05	165
GeL393	1	Copy number gain	10q21.3 - q26.3	-	FGFR2	yes			yes		
GeL393	1	Frameshift variant	16:68815649	-485-486X	CDH1	yes	0.34	38	yes	0.45	108
GeL393	1	Copy number loss	1p36.33 - p34.1	-	845 genes (ARID1A)	yes			yes		
GeL393	1	Missense variant	3:179219634	C604R	PIK3CA	no			yes	0.07	102
GeL393	1	Copy number gain	7p11.1 - q36.3	-	HGF MET	yes			yes		
GeL393	2	Missense variant	12:56097861	S846I	ERBB3	yes	0.31	75	yes	0.14	145
GeL393	2	Copy number loss	16q11.2 - q24.3	-	CYLD, HERPUD1, CDH11, CBFB, CDH1, MAF, CBFA2T3, FANCA	yes			yes		
GeL393	2	Copy number loss	1p31.1 - p12	-	FUBP1, BCL10, RBM15, TRIM33, NRAS, FAM46C	yes			yes		
GeL393	2	Copy number loss	1p31.3	-	JAK1	yes			yes		
GeL393	2	Copy number gain	1p32.2 - p31.3	-	JUN	yes			yes		
GeL393	2	Copy number loss	1p33 - p32.2	-	TAL1	yes			yes		
GeL393	2	High Copy Gain	1p34.3 - p33	-	MUTYH	yes			yes		
GeL393	2	Copy number gain	1q12 - q44	-	PDE4DIP, BCL9, ARNT, TPM3, MUC1, PRCC, NTRK1, SDHC,	yes			yes		

					FCGR2B, PBX1, ABL2, TPR, MDM4, ELK4, SLC45A3, H3F3A, FH						
GeL393	2	Inframe deletion	20:47651092	Q1255-	NCOA3	no	0.00		yes	0.03	193
GeL393	2	Copy number gain	20q11.1 - q13.33	-	ASXL1, MAFB, TOP1, SDC4, GNAS, SS18L1	yes			yes		
GeL393	2	Copy number gain	2p25.3 - q35	-	MYCN, C2orf44, NCOA1, DNMT3A, ALK, EML4, MSH2, MSH6, FBXO11, BCL11A, REL, XPO1, IGL@, TTL, PAX8, ERCC3, CHN1, HOXD13, HOXD11, NFE2L2, PMS1, SF3B1, CREB1, IDH1, ATIC	yes			yes		
GeL393	2	Stop codon gain	5:132899615	S387N*X	AFF4	no	0.00		yes	0.07	108
GeL393	2	Copy number gain	7p22.3 - p11.2	-	CARD11, PMS2, ETV1, HNRNPA2B1, HOXA9, HOXA11, HOXA13, JAZF1, IKZF1, EGFR	yes			yes		
GeL393	3	3 prime UTR variant	15:87863992	-	NTRK3	no	0.00		yes	0.08	102
GeL393	3	3 prime UTR variant	2:211377471	-	ERBB4	yes	0.48	29	yes	0.04	101
GeL393	3	Missense variant	21:34859494	D198G	RUNX1	yes	0.52	52	yes	0.58	195
GeL402	1	Copy number loss	10q11.21 - q26.3	-	PTEN	yes			yes		
GeL402	1	Copy number loss	11q11 - q25	-	ATM	yes			yes		
GeL402	1	Missense variant	12:25245350	G12V	KRAS	yes	0.50	96	yes	0.41	96
GeL402	1	Copy number gain	12p13.33 - p11.1	-	KRAS	yes			yes		
GeL402	1	Copy number gain	12q11 - q21.1	-	ERBB3	yes			yes		
GeL402	1	Missense variant	3:179218303	E545K	PIK3CA	yes	0.29	68	yes	0.32	82
GeL402	1	Frameshift variant	5:112792509	S237X	APC	yes	0.36	39	yes	0.35	55
GeL402	1	Copy number gain	8q11.1 - q24.3	-	MYC	yes			yes		
GeL402	2	Frameshift variant	10:113165557	R465RX	TCF7L2	yes	0.28	43	yes	0.30	54
GeL402	2	Copy number loss	10p15.3 - p11.1	-	GATA3, MLLT10, KIF5B	yes			yes		
GeL402	2	Copy number gain	12q21.2 - q24.33	-	BTG1, ALDH2,	yes			yes		

GeL402	2	Copy number loss	15q11.1 - q26.3	-	PTPN11, BCL7A C15orf55, BUB1B, FLJ27352, TCF12, PML, NTRK3, IDH2, CRTC3, BLM	yes	yes
GeL402	2	Copy number loss	18q11.1 - q23	-	ZNF521, SS18, MALT1, BCL2	yes	yes
GeL402	2	Copy number loss	1p36.33 - p11.2	-	TNFRSF14, PRDM16, RPL22, CAMTA1, SDHB, PAX7, MDS2, ARID1A, LCK, SFPQ, THRAP3, MYCL1, MPL, MUTYH, TAL1, CDKN2C, EPS15, JUN, JAK1, FUBP1, BCL10, RBM15, TRIM33, NRAS, FAM46C, NOTCH2	yes	yes
GeL402	2	Copy number loss	22q11.1 - q13.33	-	CLTCL1, BCR, SMARCB1, MN1, CHEK2, EWSR1, NF2, MYH9, PDGFB, MKL1, EP300	yes	yes
GeL402	2	Copy number loss	4p16.3 - p11	-	FGFR3, WHSC1, SLC34A2, PHOX2B	yes	yes
GeL402	2	Copy number loss	5q11.1 - q35.3	-	IL6ST, PIK3R1, APC, PDGFRB, CD74, ITK, EBF1, RANBP17, TLX3, NPM1, NSD1	yes	yes
GeL402	2	Copy number loss	6p25.3 - p11.1	-	IRF4, DEK, HIST1H4I, TRIM27, POU5F1, DAXX, HMGA1, FANCE, PIM1, TFEB, CCND3	yes	yes
GeL402	2	Copy number loss	6q11.1 - q27	-	PRDM1, ROS1, GOPC, STL,	yes	yes

					MYB, TNFAIP3, ECT2L, EZR, FGFR1OP, MLLT4						
GeL402	2	Copy number gain	7q11.22 - q21.11	-	ELN, HIP1	no			yes		
GeL402	2	Copy number gain	8p23.3 - p11.1	-	PCM1, WRN, WHSC1L1, FGFR1, HOOK3	yes			yes		
GeL402	2	Missense variant	X:67546345	A400V	AR	yes	0.58	76	yes	0.53	43
GeL402	3	Copy number loss	11p15.5 - p11.12	-	IGF2	yes		56	yes		
GeL402	3	Missense variant	19:17843036	K186R	JAK3	yes	0.29	56	yes	0.23	77
GeL402	3	Copy number loss	q412 - q35.2	-	EREG AREG	yes			yes		
GeL432	1	Missense variant	12:25245350	G12D	KRAS	yes	0.30	81	yes	0.40	80
GeL432	2	Frameshift variant	12:22625379	-72-73X	ETNK1	yes	0.09	67	no	0.00	
GeL432	2	Copy number gain	13p11.1 - q34	-	CDX2, FLT3, BRCA2, LHFP, LCP1, RB1, ERCC5	yes			yes		
GeL432	2	Copy number gain	17q22 - q24.2	-	CLTC, BRIP1, CD79B, DDX5	no			yes		
GeL432	2	Missense variant	22:20993973	R468H	LZTR1	yes	0.24	98	yes	0.38	110
GeL432	2	Stop codon gain	3:187729597	R270*	BCL6	yes	0.22	55	yes	0.28	151
GeL432	2	Stop codon gain	3:52587430	E991*	PBRM1	yes	0.21	77	yes	0.27	93
GeL432	2	Frameshift variant	5:256368	TL648-649TX	SDHA	yes	0.11	79	yes	0.04	169
GeL432	2	Missense variant	7:140739856	E695K	BRAF	yes	0.32	82	yes	0.22	69
GeL432	2	Stop codon gain	X:40055424	S1562*	BCOR	yes	0.57	28	yes	0.75	68
GeL432	3	Missense variant	2:29251160	E717K	ALK	yes	0.34	65	yes	0.37	116
GeL432	3	Missense variant	4:152328356	V424L	FBXW7	yes	0.25	76	yes	0.32	75
GeL432	3	Missense variant	4:186708872	A319V	FAT1	yes	0.18	62	yes	0.37	189

Appendix J Univariate analyses for complete response rate (CRR), overall response rate (ORR), progression-free survival (PFS) and overall survival (OS)

	ORR (%) n = 111	P	CRR (%) n = 111	P	median PFS (months) n = 114	P	median OS (months) n = 114	P
All (n=114)	73.9		20.7		19.1		31.6	
Age, years								
<65	78.2	0.306	21.8	0.777	18.4	0.558	40.4	0.297
≥65	69.6		19.6		19.2		28.2	
Clinical staging								
Binet A, Rai 0	78.6		21.4		20.7		46.6	
Binet B, Rai I-II	78.3	0.510	28.3	0.211	19.4	0.479	63.6	**0.018**
Binet C, Rai III-IV	68.6		13.7		14		27.1	
Characteristics of relapse								
no. of previous lines								
1	82.4	0.061	27.5	0.107	25.2	0.101	NR	0.130
>1	66.7		15		17.5		28.1	
fludarabine refractory								
Yes	55.9	**0.004**	20.6	0.982	12.3	**0.001**	21.9	**<0.001**
No	81.8		20.8		22.5		40.4	
Clinical trial								
ICLL01	78.2		20		18.4		NR	
CLL201	79.4	0.07	14.7	0.299	22.5	0.065	40.4	0.062
CLL202	54.5		31.8		12.3		21.9	
Cytogenetics (FISH)								
17p deletion	50		7.7		12.2		32.6	
11q deletion	91.7	**0.003**	25	0.183	27.3	0.353	NR	0.748
other	75.9		24.1		18.4		31.6	
***IGHV* status**								
Mutated	72.2	0.768	16.7	0.755	22.5	0.890	NR	0.266
Unmutated	75.6		23.1		18.4		32.6	
Gene mutation								
***TP53* mutation**								
Yes	53.8	**0.008**	7.4	0.061	12	**0.004**	32.6	0.455
No	80		24.7		20.7		31.6	
***ATM* mutation**								
Yes	65.5	0.233	3.4	**0.008**	14.4	0.084	26.2	0.138
No	76.8		26.8		22.4		36.4	
SF3B1 mutation								
Yes	63.3	0.124	13.3	0.243	14.5	0.631	28.1	0.159
No	77.8		23.5		19.2		36.4	
***NOTCH1* mutation**								
Yes	64.7	0.375	41.2	**0.045**	18.4	0.531	28.1	0.475
No	75.5		17		19.2		40.4	
***XPO1* mutation**								
Yes	76.5	0.791	11.8	0.517	18.4	0.953	26.1	0.136
No	73.4		22.3		19.1		36.4	
***SAMHD1* mutation**								
Yes	72.7	1	27.3	0.695	19.4	0.705	25.4	0.361
No	74		20		18.4		32.6	
***MED12* mutation**								
Yes	77.8	1	22.2	1	25.2	0.522	NR	0.941
No	73.5		20.6		18.4		31.6	
***BIRC3* mutation**								
Yes	66.7	0.651	16.7	1	5	0.947	9.4	0.096
No	74.3		21		18.4		32.6	
***MYD88* mutation**								
Yes	100	0.566	66.7	0.108	27.3	0.205	NR	0.174
No	73.1		19.4		18.4		30.4	
TP53 disruption								
Yes	57.1	**0.007**	8.6	**0.03**	13.9	0.280	32.6	0.915
No	81.3		26.7		19.2		31.6	
Clustering of gene mutation								
#1	86.4		27.3		19.1		NR	
#2	83.3	**0.002**	27.7	0.058	23	**0.018**	46.6	**0.019**
#3 multiple hit	42.9		0		12		28.2	
#4	71.4		19		17.5		27.1	

Appendix K Mutational signatures in different CLL groups

