University of Glasgow

A comprehensive study of extrachromosomal circular DNA in pancreatic ductal adenocarcinoma

Daniel Schreyer, MSc

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

SCHOOL OF CANCER SCIENCES

COLLEGE OF MEDICAL, VETERINARY & LIFE SCIENCES

UNIVERSITY OF GLASGOW

December, 2023

Abstract

Introduction

Extrachromosomal circular DNAs (eccDNA) vary in sizes and drive many aspects of tumour biology, including drug resistance and oncogene amplification. However, little is known about their landscape, role, and association with specific features of cancer in pancreatic ductal adenocarcinoma (PDAC).

Approach

To better understand their properties in PDAC, I combine multiple sources of genomic data with transcriptomic data from a large library of PDAC patient-derived organoids (PDOs), including some with matched primary tumours or cell lines, and a panel of PDAC patient-derived cell lines.

Results

Here, I report that large and amplified eccDNAs (ecDNAs) are prevalent in PDAC, revealing a diverse landscape containing PDAC-specific oncogenes. The frequency of tumours harbouring ecDNAs was increased when genomic instability and TP53 alterations were present, suggesting the association of ecDNA occurrence with unstable genomes. Furthermore, integrating experimental work, performed by collaborators at the University of Verona, and computational analysis, ecDNAs were found to drive PDO adaptation to harsh environments.Investigation of the complete eccDNA landscape revealed that eccDNAs predominantly originate from regions with a high gene- and enhancer-density, active transcription, and accessible chromatin. I identified 61 recurrent eccDNA hotspots associated with increased transcription suggesting a direct link between transcription and eccDNA formation. Finally, as eccDNA analysis involves the use of multiple tools, I present nf-core/circdna, an open-source workflow that integrates the tools of eccDNA research and data processing. This workflow offers ease of use, portability, comprehensive documentation and full reproducibility.

Conclusions

EccDNAs are an important feature of PDAC genomic. This in-depth characterisation of multiple features of eccDNAs provides a comprehensive overview of their properties in PDAC and a valuable resource for future research, hopefully uncovering potential avenues for eccDNA-based precision medicine.

“The fuel on which science runs is ignorance.”

MATT RIDLEY

Contents

Abstract **I**

List of Tables **VIII**

List of Figures **X**

List of accompanying Material **XIII**

List of Publications **XIV**

Acknowledgements **XV**

Author's declaration **XVII**

List of Abbreviations **XVIII**

1 Introduction **1**

1.1 Cancer . . . 1

1.1.1 Cancer genomics . . . 2

1.2 Pancreatic ductal adenocarcinoma . . . 6

1.2.1 Overview and clinical presentation . . . 6

1.2.2 Genomics . . . 7

1.2.3 Transcriptomics reveals PDAC subtypes . . . 8

1.3 Extrachromosomal circular DNA . . . 9

1.3.1 EccDNA vs. EcDNAs - History, nomenclature, and definition . . . 9

1.3.2 Structural features of ecDNA . . . 10

1.3.3 Biological implications of ecDNAs in cancer . . . 12

1.3.4 EcDNA maintenance, targeting, and elimination . . . 15

1.3.5 The biogenesis of ecDNAs . . . 16

1.3.6 EccDNAs - Molecular characteristics and biogenesis . . . 19

1.3.7 Function of small non-amplified eccDNAs . . . 21

1.3.8 EccDNA research in pancreatic cancer . . . 22

1.3.9 Techniques for the identification and analysis of eccDNAs . . . 23

1.4 Aims & objectives . . . 27

2 Methods **29**

2.1 nf-core/circdna . . . 29
2.2 Cell culture methods . . . 30
2.2.1 Cell culture . . . 30
2.2.2 Cell harvest . . . 30
2.3 Human specimens and clinical data . . . 31
2.4 PDO establishment and culture . . . 31
2.5 Organoids metaphase spreads . . . 32
2.6 DNA fluorescence *in situ* hybridisation . . . 32
2.7 PDO and primary tissue DNA Isolation . . . 32
2.8 DNA Panel Sequencing . . . 32
2.9 Whole genome sequencing (WGS) . . . 33
2.10 Circle-seq . . . 33
2.10.1 HMW DNA extraction . . . 34
2.10.2 Volume 1 - Linear DNA removal . . . 34
2.10.3 Volume 2 - Linear DNA removal . . . 35
2.10.4 Rolling-circle amplification . . . 35
2.10.5 Sequencing . . . 35
2.11 Data processing and quality control . . . 36
2.12 Sequence alignment and duplicate removal . . . 36
2.13 EccDNA calling . . . 36
2.14 EccDNA filtering . . . 36
2.15 Visual inspection of eccDNA calls . . . 36
2.16 Read pileup . . . 36
2.17 Generating random eccDNA data sets . . . 37
2.18 Permutation test . . . 37
2.19 EccDNA annotation . . . 37
2.20 EccDNA hotspot analysis . . . 38
2.21 Overrepresentation analysis . . . 38
2.22 EccDNA *de novo* assembly . . . 38
2.23 PDCL copy number and expression data . . . 39
2.24 ATAC-seq data analysis . . . 39
2.25 ICGC PDAC data . . . 39
2.26 WGS data pre-processing and alignment . . . 39
2.27 Amplicon characterisation . . . 40
2.28 Sample classification . . . 40
2.29 EcDNA analysis . . . 40
2.30 RNA-seq of HCMI PDOs . . . 40
2.31 Gene set enrichment analysis . . . 41
2.32 Public datasets . . . 41
2.33 Restricted Datasets . . . 41
2.34 Statistical analysis . . . 42

2.35 Survival analysis . . . 42
2.36 Molecular biology methods . . . 42
2.36.1 Primer design for eccDNA candidate validation . . . 42
2.36.2 Inverse PCR targeting candidate eccDNA junctions . . . 42
2.36.3 Sequencing of candidate eccDNA junctions . . . 43
2.37 Graphic design and illustration . . . 43
2.38 Extended Data . . . 43
2.39 List of software and algorithms . . . 43
2.40 List of reagents . . . 47
2.41 Media formulation . . . 49
2.42 Primer sequences . . . 51

3 nf-core/circdna: *A Nextflow pipeline for the detection of ecDNA and eccDNA in genomic data sets* 53
3.1 Pipeline structure . . . 54
3.2 Portability . . . 55
3.3 Continuous integration testing of nf-core/circdna . . . 55
3.4 Adapting the nf-core/circdna pipeline for user-specific needs . . . 57
3.4.1 Input data . . . 57
3.4.2 User-friendly, yet highly customisable: The parameters of nf-core/circdna . . . 57
3.4.3 Data format and samplesheet styles . . . 58
3.5 Installation and usage . . . 59
3.6 Output description and utility . . . 59
3.6.1 Pipeline reports and quality control . . . 59
3.6.2 Intermediate Files . . . 60
3.6.3 EcDNA/EccDNA Information . . . 60
3.7 Results . . . 61
3.7.1 AmpliconArchitect branch identifies ecDNAs . . . 62
3.7.2 Single-fragment eccDNA identification . . . 64
3.7.3 Multi-fragment eccDNA identification . . . 65
3.8 Conclusion & discussion . . . 66

4 Investigating ecDNAs in PDAC 67
4.1 Study samples . . . 68
4.2 EcDNAs are common in PDAC . . . 70
4.3 EcDNAs are retained in PDAC PDOs . . . 72
4.4 The amplicon landscape of PDAC PDOs . . . 74
4.5 Distinct transcriptomic profiles in ecDNA+ tumours and PDOs . . . 76
4.5.1 EcDNAs are associated with a Basal-like signature in PDAC tumours 77
4.5.2 Differential transcriptomic signatures in ecDNA+ PDAC tumours and PDOs . . . 78

4.6 Presence of ecDNA in PDAC is linked to genomic instability 80
4.6.1 Mutational landscape defining ecDNA+ and ecDNA– PDAC 80
4.6.2 EcDNA+ tumours are associated with an unstable genome 81
4.6.3 EcDNA-positivity is associated with whole-genome duplications . . 82
4.6.4 Transcriptomic chromosomal instability signature is enriched in ecDNA+ PDAC 83
4.7 The role of ecDNAs in oncogene amplification and expression 84
4.8 Copy number alterations in ecDNA+ PDAC samples 85
4.9 EcDNA selection and evolution during PDO adaptation 86
4.10 Circle-seq validates *MYC*-ecDNA 91
4.11 EcDNAs are maintained in metastatic PDOs 92
4.12 Prognostic implications of ecDNA presence 93
4.13 Investigating ecDNAs in PDAC cell lines 95
4.13.1 Commonly used PDAC cell lines lack ecDNAs 95
4.13.2 Retention of ecDNAs in PDO-derived cell lines 96
4.14 Discussion 98

5 Investigating the eccDNA landscape in PDAC 101
5.1 Establishing Circle-seq 102
5.2 Circle-seq on 8 PDAC PDCLs 105
5.3 Retention of high quality eccDNAs 107
5.4 EccDNAs in PDAC PDCLs: Size distribution and origin 107
5.5 EccDNA origins: Gene expression and chromatin accessibility 108
5.6 Specific genomic features are enriched on eccDNAs 112
5.7 Genes are rarely fully incorporated on eccDNAs 114
5.8 Identification of multi-fragment eccDNAs 115
5.8.1 Identification of high-quality *de novo* assembled eccDNAs 116
5.8.2 Multi-fragment eccDNAs in PDAC PDCLs 116
5.9 Computational validation of *de novo* assembly approach 119
5.10 Validation of Circle-seq results 120
5.11 The retention of eccDNAs in PDAC PDCLs 123
5.11.1 Candidate eccDNA junctions are not present across different passages 124
5.11.2 Circle-seq of two consecutive passages of 7 PDAC PDCLs 125
5.11.3 EccDNA Landscape of two consecutive passages 126
5.11.4 Shared eccDNAs are identified in other PDCLs 128
5.11.5 EcDNAs are replicated and retained during passaging and Gemcitabine treatment 128
5.12 EccDNA-formation hotspots in PDAC 130
5.12.1 Identification of common eccDNA hotspots 130
5.12.2 Common eccDNA hotspots are gene-dense and located in specific genomic regions 132

5.12.3 EccDNA hotspots are associated with increased gene expression and chromatin accessibility 133
5.12.4 Consensus eccDNA-formation hotspots 135
5.12.5 *MYC* is located inside a consensus eccDNA hotspot 136
5.13 Discussion 138

6 Conclusion **144**

A Investigating ecDNAs in PDAC **147**

B Investigating the eccDNA landscape in PDAC **148**

Bibliography **153**

List of Tables

1.1 Difference between ecDNAs and eccDNAs 10

2.1 Software and their respective versions used in the nf-core/circdna pipeline . 29
2.2 PDCL and PDO samples used for Circle-seq and their unique identifier (ID) 33
2.3 List of software and algorithms 43
2.4 List of resources 47
2.5 Media formulation 49
2.6 Primer sequences 51

3.1 Minimal input requirements of nf-core/circdna 58
3.2 Samplesheet file column description with `--input_format` 'FASTQ' and 'BAM' 58
3.3 Classified amplicons identified by the 'ampliconarchitect' branch of five commonly used cell lines 62

4.1 Statistical examination of *MYC*-ecDNA similarities identified in the VR01 and VR06 primary tumour and PDO 74
4.2 Amplicon similarity examination of the two circular amplicons identified in both VR11 PDOs 93
4.3 PDAC CCLE cell lines and their respective number of amplicons per amplicon class identified 96

5.1 Mean fragment sizes after fragmentation of circular DNA enriched samples 104
5.2 Fold change decrease of linear DNA compared to circular DNA after DNase digestion of HMW DNA of 8 PDAC PDCLs 106
5.3 EccDNA candidates used for validation 123
5.4 Fold change decrease of linear DNA compared to circular DNA after DNase digestion of two consecutive passages of seven PDAC PDCLs 126
5.5 PDCL passage shared eccDNAs that are also identified in other PDCLs . . 128
5.6 Genes identified within consensus eccDNA hotspots 136

A.1 CCLE PDAC cell lines with available WGS data 147

B.1 EccDNA candidates, their chromosomal origin, and their content used for validation 149

B.2 Wilcoxon-rank sum test results of number of genomic elements identified in the common hotspots compared to universal eccDNA hotspots 152

List of Figures

1.1 Pancreatic cancer statistics compared to other cancer types 6
1.2 Segregation schema of chromosomal and extrachromosomal circular DNA . 12
1.3 EcDNA formation mechanisms 20
1.4 Circle-seq procedure 24
1.5 EcDNA & EccDNA Identification with short-read sequencing data 26

2.1 Schematic representation of the permutation test 37

3.1 nf-core/circdna Branch Overview 56
3.2 Number of eccDNAs identified in the three nf-core/circdna branches identifying putative eccDNA junctions 64
3.3 Number of unique single-or multi-fragment eccDNAs identified using the 'Unicycler' branch of nf-core/circdna 65

4.1 ICGC and HCMI sample overview 69
4.2 Amplicon and sample classification of ICGC primary and HCMI PDO samples 71
4.3 No association between PDAC stage and ecDNA presence 72
4.4 Retention and possible evolution of *MYC*-ecDNAs in PDAC PDOs 73
4.5 Amplicon landscape of PDAC PDOs 75
4.6 EcDNAs are associated with a Basal/Squamous signature in PDAC tumours 77
4.7 Hallmark pathway activation in ecDNA+ and ecDNA– PDAC tumours and PDOs 79
4.8 TP53 inactivation in ecDNA+ and ecDNA– ICGC tumours with transcriptomic data. 80
4.9 Overview of key PDAC gene alterations in ecDNA+ and ecDNA– tumours and PDOs 81
4.10 EcDNA presence is associated with chromothripsis events and an abundance of SVs 82
4.11 Whole-genome duplication status in ecDNA+ and ecDNA– samples 83
4.12 Chromosomal instability signature is enriched in ecDNA+ tumours and PDOs 84
4.13 Copy number and transcription of extrachromosomal and chromosomal amplified oncogenes 84
4.14 Genomic overview of copy number gains and losses in ecDNA+ and ecDNA– tumours and PDOs 86
4.15 Overview of gene mutations and copy number alterations in the Verona PDOs 87

4.16 Adaptation and propagation of six PDOs grown in -WR media 88
4.17 *MYC*-ecDNA amplification and evolution during adaptation to environmental changes 88
4.18 *MYC* expression in and genomic analysis of parental and adapted (-WR) PDOs 89
4.19 Genomic and transcriptomic analysis of *MYC* expression in VR01 and VR06 89
4.20 Evolution of *MYC*-ecDNA in VR01-WRb 90
4.21 Circle-seq coverage confirms VR01-O *MYC*-ecDNA 92
4.22 Large similarity between ecDNA amplicons of VR11-O and VR11-LNO . . 94
4.23 Kaplan-Meyer survival analysis of ecDNA+ and ecDNA– PDAC tumours . 95
4.24 EcDNA landscape of PDOs with matching PDO-derived cell lines 97
4.25 MYC ecDNA is retained in PDO-derived cell line 97

5.1 Circular DNA-safe DNase decreases greatly reduces chromosomal DNA content 103
5.2 Circular DNA is amplified by rolling-circle amplification 104
5.3 Bioanalyzer verifies DNA fragmentation size 104
5.4 Validation of Circle-seq library preparation 105
5.5 Representative genomic overview of mapped Circle-seq reads 106
5.6 Sequential filtering steps to discern high-quality eccDNAs 107
5.7 Characteristics of high-quality eccDNAs in eight PDAC PDCLs 108
5.8 Genomic overview of selected eccDNAs with matching ATAC-seq data . . 109
5.9 EccDNAs originate from genes with elevated expression and regions with high increased accessibility 110
5.10 EccDNAs originate from PDAC and cancer related pathway genes 111
5.11 EccDNAs contain specific genomic features 113
5.12 EccDNAs in PDAC rarely contain full genes 115
5.13 Unicycler eccDNA filtering steps. 117
5.14 Large eccDNAs are comprised of multiple fragments 117
5.15 Read alignment views of eccDNAs identified by Unicycler 118
5.16 Most *de novo* assembled single-fragment eccDNAs are validated by Circle-Map 119
5.17 Inverse PCR validates Circle-seq results 121
5.18 Sanger sequencing of candidate eccDNA junctions confirms Circle-seq results 122
5.19 Graphical description of Circle-seq set-up from two consecutive passages of PDAC PDCLs 124
5.20 Validation failure of eccDNA junctions in different passages 125
5.21 EccDNA abundance in two consecutive passages of seven PDAC PDCLs . 127
5.22 *MYC*-ecDNA is retained VR01-O during Gemcitabine treatment and passaging 129
5.23 EccDNA Hotspot identification in the three Circle-seq datasets 131
5.24 Genomic view of common eccDNA hotspots, coldspots, and normal regions 132
5.25 Common eccDNAs hotspots contain specific genomic elements 133
5.26 Increased gene expression and chromatin accessibility in common eccDNA hotspots 135
5.27 Consensus eccDNA hotspots highlight the *MYC* locus 137

5.28 EccDNA abundance on Chromosome 8 137
5.29 Identified genomic characteristics associated with eccDNAs and eccDNA formation hotspots 141

B.1 Size selection is associated with huge DNA loss 148
B.2 Validation of library preparation success of 10 Circle-seq samples 149
B.3 EccDNA abundance on chromosome 8 151
B.4 EccDNA abundance and sizes in each sample of the three Circle-seq data sets 151

List of accompanying Material

1. Extended data, scripts, raw data, and processed data, as well as results, are accessible through the Open Science Framework (https://osf.io/xhbev). Access to the repository will be granted to supervisors, thesis examiners, and upon reasonable request. Please reach out to the thesis author via email (ds.danielschreyer@gmail.com) to request access.

List of Publications

1. **Schreyer, D.**, Neoptolemos, J. P., Barry, S. T., & Bailey, P. (2022). Deconstructing Pancreatic Cancer Using Next Generation-Omic Technologies–From Discovery to Knowledge-Guided Platforms for Better Patient Management. Frontiers in Cell and Developmental Biology, 9. https://www.frontiersin.org/articles/10.3389/fcell.2021.795735

2. Malinova, A., **Schreyer, D.**, Fiorini, E., Pasini, D., Bevere, M., D'Agosto, S., Andreani, S., Lupo, F., Veghini, L., Grimaldi, S., Pedron, S., Nourse, C., Salvia, R., Malleo, G., Ruzzenente, A., Guglielmi, A., Milella, M., Lawlor, R. T., Luchini, C., … Corbo, V. (2023). ecDNA amplification of MYC drives intratumor copy-number heterogeneity and adaptation to stress in PDAC. bioRxiv. https://doi.org/10.1101/2023.09.27.559717

3. Bailey, P., Ridgway, R. A., Cammareri, P., Treanor-Taylor, M., Bailey, U.-M., Schoenherr, C., Bone, M., **Schreyer, D.**, Purdie, K., Thomson, J., Rickaby, W., Jackstadt, R., Campbell, A. D., Dimonitsas, E., Stratigos, A. J., Arron, S. T., Wang, J., Blyth, K., Proby, C. M., … Inman, G. J. (2023). Driver gene combinations dictate cutaneous squamous cell carcinoma disease continuum progression. Nature Communications, 14(1), Article 1. https://doi.org/10.1038/s41467-023-40822-9

4. Geyer, M., **Schreyer, D.**, Gaul, L.-M., Pfeffer, S., Pilarsky, C., & Queiroz, K. (2023). A microfluidic-based PDAC organoid system reveals the impact of hypoxia in response to treatment. Cell Death Discovery, 9(1), Article 1. https://doi.org/10.1038/s41420-023-01334-z

5. Zeng, S., Lan, B., Ren, X., Zhang, S., **Schreyer, D.**, Eckstein, M., Yang, H., Britzen-Laurent, N., Dahl, A., Mukhopadhyay, D., Chang, D., Kutschick, I., Pfeffer, S., Bailey, P., Biankin, A., Grützmann, R., & Pilarsky, C. (2022). CDK7 inhibition augments response to multidrug chemotherapy in pancreatic cancer. Journal of Experimental & Clinical Cancer Research, 41(1), 241. https://doi.org/10.1186/s13046-022-02443-w

Acknowledgements

First and foremost, I'd like to thank my main supervisor, Peter Bailey, for taking me on as his PhD student, guiding me through my PhD, helping me with my endless questions and shaping my biological and scientific understanding. His expertise in computational biology and cancer genomics has been invaluable to my time as a PhD student. I also appreciate the feedback I have received over the past few months.

I would also like to take this opportunity to thank the European Union and its Horizon 2020 Research and Innovation Programme Marie Skłodowska-Curie grant designated for PRECODE (No 861196). Without this grant, my work would not have been possible and you wouldn't be reading this thesis.

I'd also like to thank my second supervisor, Gareth Inman, for bringing me into his lab environment and letting me be part of the R01. It has been a great pleasure to participate in lab meetings or (more importantly) lab nights with the whole R01. Speaking of the R01 and its collaborators: Of course, I also have to thank everyone in the lab (my colleagues*) from my time at the Beatson. In particular, Carlotta, Max and Mairi for the typical PhD thingies, including pub nights, talking about nonsense and Italian food, and sometimes a bit of science. Please remember that "we need stats" in the thesis. Also, thanks to Michela, Christina, John and Craig for their endless help in the lab. Without you, my bioinformatics brain would have gone absolutely haywire in the lab. It was also wonderful to have the occasional game night with all of you. Finally, thanks to Irati and Jasbani for helping me settle into the Beatson and guiding me around the lab at the beginning of my PhD. You all have a place in my Glaswegian heart.

During my PhD and my time in PRECODE, I was fortunate to work with and visit an amazing lab at the University of Verona. Here, thanks to Vincenzo for hosting me and guiding my computational analyses and a huge thanks to the full lab, but most importantly Toni, Elena, and Davide, who answered all my questions and performed some wonderful experimental work. Without you, my PhD thesis would miss 1/3 of its content.

Additionally, I got to see the industry side of research at AstraZeneca, where I was hosted by Simon Barry. Thus, also a huge thanks to Simon and AstraZeneca for inviting me and shape my future plans.

I would also like to thank my extended supervisory committee of Professor Joanne Edwards

and Professor Daniel Murphy. Their input each year has guided my work and improved my overall understanding of it.

I would like to extend my sincere gratitude to the examiners of this thesis for their time, effort, and expertise. I am looking forward to discuss my thesis with you and debate its contributions to the scientific community.

Moving to a new country, a new city and a new place is not easy. Especially when the initial period is intertwined with several Covid-19 lockdowns. Without Maria and Adam, who took me into their extended household, this time would not have been as much fun as it was. Being a person who likes to be alone at home was a treat during the lockdowns, but the occasional gaming/drinking nights were definitely needed to keep me in a good spirit. When the lockdown was lifted, I finally started to see the real Glasgow and meet new people. In particular, the French and the Scottish/Irish have taken me into their friend groups and made Glasgow a wonderful place to be. I would also like to thank the pubs for providing me with a fresh pint of Tennent's whenever I needed it.

Speaking of Scots, I have to mention my DnD people, with whom I spend endless hours rolling dice, defeating imaginary monsters and discussing whether or not to be reasonable. Let's roll a D20 (+4 modifier) to see how many weeks it will take us to finally kill Strahd, or die trying.

Also, thanks to Flo for his keen eye for detail in identifying grammatical and structural errors.

I must thank my parents for their incredible support during my time away. By constantly checking up on me and supporting me in tricky situations, they made it possible for me to concentrate fully on my PhD in Glasgow and make my life as easy as possible.

Importantly, before I forget, I must thank my girlfriend Franzi for supporting me in my decision to move to Glasgow and complete my PhD. I know the time apart has not been easy for us, but I think we made it in the end and can now focus on our time together. You have been a tremendous support, giving me strength from afar at the most difficult times. It was also wonderful to see that you enjoyed Glasgow as much as I did.

Finally, of course, there are many other people to thank, and I think they will know when they read this. So thank you. Also to you, dear reader, with whom I will now take you on a journey through three years of doctoral work.

Author’s declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Daniel Schreyer

December, 2023

List of Abbreviations

List of Abbreviation	Meaning
Cancer Types and Models	
PAEN	Pancreatic Cancer Endocrine Neoplasm
PC	Pancreatic Cancer
PDAC	Pancreatic Ductal Adenocarcinoma
PDCL	Patient-Derived Cell Line
PDO	Patient-Derived Organoid
Genomic and Molecular Biology Terms	
BFB	Breakage-Fusion Bridge Cycle
bp	Base Pair
CA19-9	Carbohydrate Antigen 19-9
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CT	Cycle Threshold
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
DM	Double Minute
dsDNA	Double-Stranded DNA
eccDNA	Extrachromosomal Circular DNA
ecDNA	Extrachromosomal Circular DNA
FOLFIRINOX	5-fluorouracil, folinic acid, irinotecan, and oxaliplatin
FoSTeS	Replication fork stalling and template switching
gneg	Giemsa-negative
gpos	Giemsa-positive
HDR	Homology-Directed Repair
HMW	High Molecular Weight
HSR	Homogeneously Staining Regions
kbp	Kilo Base Pair
lncRNA	Long non-coding RNA

Continued on next page

Abbreviation	Definition
LINE	Long-Interspersed Nuclear Element
LTR	Long-Terminal Repeat
Mbp	Mega Base Pair
miRNA	MicroRNA
MMEJ	Microhomology-Mediated End-Joining
nab	Nanoparticle-Albumin Bound
NHEJ	Non-Homologous End-Joining
PARP	Poly (ADP-Ribose) polymerase
RCA	Rolling-Circle Amplification
RNA	Ribonucleic Acid
SCNA	Somatic Copy Number Alterations
SINE	Short-Interspersed Nuclear Element
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variant
TKI	Tyrosine Kinase Inhibitor
Sequencing and Bioinformatics	
BAM	Binary Alignment Map
BWA	Burrows-Wheeler Algorithm
CPU	Central Processing Unit
FASTA	Fast All
GATK	Genome Analysis Toolkit
GRCh19	Genome Reference Consortium Human Build 19
GRCh38	Genome Reference Consortium Human Build 38
GSEA	Gene Set Enrichment Analysis
hg19	Human Genome, Version 19
hg38	Human Genome, Version 38
IGV	Integrative Genomics Viewer
NGS	Next-Generation Sequencing
RPKM	Reads Per Kilobase Per Million Reads
SAM	Sequence Alignment Map
seq	Sequencing
WGS	Whole-Genome Sequencing
Cancer Research and Analysis Tools	
CCLE	Cancer Cell Line Encyclopedia
CIN	Chromosomal Instability
ecDNA+	ecDNA-Positive
ecDNA–	ecDNA-Negative
EMT	Epithelial-To-Mesenchymal Transition

Continued on next page

Abbreviation	Definition
HCMI	Human Cancer Model Initiative
ICGC	International Cancer Genome Consortium
No-fSCNA	No Focal Somatic Copy Number Alteration
PRECODE	Pancreatic Cancer Organoids Research
TCGA	The Cancer Genome Atlas
General Laboratory and Measurement Terms	
°C	Degree Celsius
h	Hour
L	Liter
M	Molar
mL	Milliliter
min	Minute
nm	Nano Molar
ng	Nanogram
PBS	Phosphate-Buffered Saline
RFC	Relative Centrifugal Force
s	Second
µg	Microgram
µL	Microliter
V	Volt
Data Repositories and Bioinformatics Resources	
NCBI	National Center for Biotechnology Information
SRA	Sequence Read Archive
KEGG	Kyoto Encyclopedia of Genes and Genomes

Chapter 1

Introduction

> Biology is the study of complicated things that have the appearance of having been designed with a purpose.
>
> *Richard Dawkins*

Pancreatic ductal adenocarcinoma (PDAC) is a highly lethal disease and the most prevalent form of pancreatic cancer. The primary curative treatment is surgical resection, but unfortunately, it is often not feasible for many patients (Mizrahi et al., 2020). While targeted therapies have improved patient survival, a significant number of PDAC patients have tumours that lack actionable alterations, rendering personalised therapies ineffective (Pishvaian et al., 2020). This highlights the need for a deeper understanding of PDAC at the molecular level to guide future therapeutic strategies.

A promising area of research that has recently gained significant attention involves extrachromosomal circular DNAs (ecDNAs). These genetic elements are present in almost all cancer types and are known to contribute to oncogene amplification, cancer progression, and drug resistance (Kim et al., 2020; Nathanson et al., 2014). However, their comprehensive landscape, functional roles, characteristic features, and biogenesis in PDAC have not yet been fully delineated. This thesis aims to address these critical knowledge gaps by conducting an in-depth characterisation of ecDNAs in PDAC. The investigation seeks to establish a foundation for understanding ecDNAs' broader implications in PDAC, potentially guiding future research towards novel diagnostic and therapeutic approaches.

1.1 Cancer

Cancer is defined by an uncontrolled division of abnormal cells in specific parts of the human body, which can have the characteristic to invade proximal and distant tissue forming metastatic niches. In normal tissues, cells are under constant control mechanism to maintain

homeostasis between dividing stem cells and differentiated cells. This is governed by molecular mechanisms inhibiting cell proliferation, activating cell differentiation or cell death. Through their lifespan, cancer cells accumulate genetic and epigenetic alterations that allow it to become abnormal and escape the normal cell regulations. This can include resistance to cell death, increased genomic instability, defects in DNA repair, immune escape, increased cell proliferation, or acquiring the ability to metastasise. Tumour growth and metastatic formation in different types of cancer can lead to various health complications, significantly impacting the health of the patient. These complications include organ failure, internal bleeding, increased susceptibility to infections, and respiratory failure, all of which can have life-threatening consequences (Hanahan, 2022; Bishop, 1987; Weinberg, 1996, 2013).

1.1.1 Cancer genomics

Cancer genomes are characterised by a complex landscape of genetic alterations, comprising of both driver alterations that contribute to tumour initiation and progression, and passenger alterations that have minimal impact on cancer growth. Depending on the type of genomic alteration, gene activity and transcription of the affected genes can be positively or negatively modulated. Besides single-nucleotide mutations, which are the most common form of genomic alterations, tumours can also be affected by larger alterations such as DNA deletions, insertions, amplifications, and translocations. While small deletions and insertions in genes can disrupt the normal function of the encoded proteins, deletions or amplifications spanning entire genes or contiguous regions of the genome can result in abnormal transcription and activity of the protein products (Bishop, 1987; Vogelstein et al., 2013).

Genomic alterations arise from a combination of intrinsic and extrinsic factors. Intrinsic factors include errors during DNA replication and repair, or spontaneuos DNA damage. Extrinsic factors involve the exposure to mutagenic agents that induce DNA mutations (Modrich, 1994; Wu et al., 2016).

In some cases, individuals can inherit tumour-promoting mutations in their germline, which are passed on from their parents. These germline mutations can predispose the indviduals to an increased risk of developing certain types of tumours. However, it is important to note that most alterations in cancer are somatically acquired over time (Bishop, 1987; Vogelstein et al., 2013; Weinberg, 2013).

The consequences of genomic alterations on the cancer biology depend on the affected genes. Protein inactivating alterations are usually found in tumour suppressor genes, such as tumour protein 53 (*TP53*), cyclin dependent kinase inhibitor 2A (*CDKN2A*), breast-cancer gene 1 and 2 (*BRCA1*/*BRCA2*), or retinoblastoma 1 (*RB1*). Inactivation is often facilitated by protein-altering mutations, such as single-nucleotide variations, insertions, or deletions, or the loss of gene copies. In contrast, activating alterations, such as gene amplifications or protein-activating mutations, drive the over-activation of proto-oncogenes, transforming them into cancer driving oncogenes. Proto-oncogenes encode proteins with roles in cell

proliferation, cell survival, or both. Some of the most frequently deregulated oncogenes in cancer are genes from the Myc family, *MYC* (c-Myc), *MYCL* (L-Myc), and *MYCN* (N-Myc), the Kirsten rat sarcoma viral oncogene Homolog (*KRAS*), the epidermal growth factor receptor (*EGFR*), and the B-cell lymphoma 2 (*BCL2*) (Weinberg, 1996, 2013; Vogelstein et al., 2013; Croce, 2008).

Understanding the molecular mechanisms that drive cancer development and progression are essential for the development of effective therapeutic approaches for individual patients. Through world-wide efforts, thousands of cancer genomes have been analysed using next-generation sequencing technologies. Coupled with advances in computational algorithms, these studies have uncovered cancer-specific recurrent gene alterations, which have revealed complex heterogeneity in the genetic makeup of tumours (Tomczak, Czerwińska & Wiznerowicz, 2015; Aaltonen et al., 2020; Vogelstein et al., 2013; Andor et al., 2016). Interestingly, such heterogeneity is not only observed between different cancer types, but also between patients affected by the same disease. This heterogeneity possess significant challenges for developing therapeutic approaches, as not all cancer can be treated equally (Bedard et al., 2013). Adding another layer to this complexity, single-cell analyses have unraveled the multifaceted architecture of tumours, identifying varied cell populations with unique genetic and phenotypic profiles. This evolving understanding describes cancer not as a static disease, but as a dynamic system that can continuously evolve (Keller & Pantel, 2019; Lawson et al., 2018).

Tumour heterogeneity and evolution

Most tumour populations are extensively heterogeneous in their phenotypic and genotypic architecture (Campbell et al., 2010; Andor et al., 2016). This does not only include the genetic alterations which shape the tumour populations, but also the interplay between non-cancerous cells. One of the main drivers of intra-tumour heterogeneity is genomic instability. Genomic instability leads to the accumulation of mutations with either neutral, beneficial or detrimental effect for the cancer cells. The effect of the mutation is also highly dependent on the environmental circumstances. While specific alterations can have negative effects on cell growth and survival at the tissue of origin, beneficiary effects can be observed in other environments such as pre-metastatic niches (Dagogo-Jack & Shaw, 2018).

The cancer initiation process is driven by an acquisition of multiple mutations that activate oncogenic programs and inactivate tumour suppressor genes, which decouple the cells from their normal cell programs (Hanahan, 2022; Vogelstein et al., 2013). Affected by genomic instability, cancer cells are not only mutated in specific genes, but contain somatic mutations in hundreds of other regions of the genome, which don't seem to have a large impact on the cancer biology and fitness. During cancer progression, cancer cells become intensively unstable and accumulate an abundance of somatic mutations and alterations, which do not only affect local regions of the genome, but can affect whole chromosomes. Due to the

continuous accumulation of mutations during cancer progression, single cancer cells acquire novel driver mutations that leads to the formation of sub-populations inside the tumour. These sub-populations have vastly diverse characteristics in terms of cell growth rates, response to therapies, or genetic and transcriptional landscape (Caiado, Silva-Santos & Norell, 2016; Dagogo-Jack & Shaw, 2018; Williams et al., 2023). Therefore, sub-populations inside tumour cells might behave differently during various stress or environmental conditions making the tumour highly adaptable.

Depending on different selective pressures, sub-populations in heterogeneous tumours can emerge or can be extinguished. This is especially critical in therapeutic circumstances where small sub-populations show high tolerance to a given drug or can evolve to exhibit increased drug tolerance (Dagogo-Jack & Shaw, 2018). Studies in various cancer types identified that sub-populations pre-exist before treatment and drive drug resistances (Jamal-Hanjani et al., 2017; Bhang et al., 2015). Therefore, finding combinatorial approaches that target the majority of cancer cells, but also sub-populations need to be identified to completely eradicate cancer cells from the body.

Genomic instability

Genomic instability is one of the hallmarks of cancer and is characterised by a high frequency of large structural variations (SVs), increased mutational burden, and complex genome rearrangements such as chromothripsis or breakage-fusion bridge (BFB) cycles (Negrini, Gorgoulis & Halazonetis, 2010). Genomic instability is observed in nearly all cancer types and contributes to tumour progression and acquired resistance to therapy (Negrini, Gorgoulis & Halazonetis, 2010; Nowell, 1976; Dharanipragada et al., 2023; Shoshani et al., 2021).

Genomic instability can arise from several sources, including the inactivation of genome caretaker genes or the activation of oncogenes. Genome caretaker genes such as *BRCA1/2*, *XRCC7*, and *MSH2*, are involved in DNA damage repair. When these caretaker genes are inactivated, DNA damage accumulates. However, cancer genome landscape studies have shown that caretaker genes are infrequently mutated in sporadic cancers. Therefore, it has been suggested that many sporadic cancers are not affected by caretaker-inactivation-induced genomic instability (Negrini, Gorgoulis & Halazonetis, 2010). On the other hand, oncogene-induced genomic instability is more commonly observed in sporadic cancer due to the high alteration frequency of oncogenes. This type of genomic instability is often caused by oncogene-induced replication stress. Several genes, including *CCND1*, *CCNE1*, *KRAS*, and *MYC*, have been identified to induce replication stress and DNA damage. Activation of these genes can lead to premature cell cycle entry or increased replication origin firing, resulting in DNA damage and the activation of the DNA damage response pathways (O'Connor, 2015; Hills & Diffley, 2014). Tumour suppressor genes such as *TP53* and *CDKN2A*, which are frequently mutated in cancer, play a role in regulating these mechanisms and maintaining genomic stability by stalling cell cycle progression or activating cell apoptosis (Negrini,

Gorgoulis & Halazonetis, 2010; Aaltonen et al., 2020; Mijit et al., 2020).

Somatic copy number alterations

Cancer initiation and progression is driven by the acquisition of somatic driver alterations, including somatic copy number alterations (SCNAs). Technological advances and cost reduction enabled the large-scale identification of copy number changes in cancer samples. Common techniques include the use of single-nucleotide polymorphism arrays or whole-genome sequencing (WGS), which identify the full SCNA landscape. This allowed the identification of driver events and cancer driver genes, such as tumour suppressor genes and oncogenes, which play a role in the disease biology by altering gene dosage (Zack et al., 2013; Beroukhim et al., 2010; Steele et al., 2022; Almal & Padh, 2012).

SCNAs are extremely common in cancer and especially in specific regions of the genome. These regions contain tumour suppressor genes or oncogenes, which are inactivated or activated due to DNA deletion or amplification, respectively. Recurrence of gene amplifications and deletions is dependent on the cancer type, but common SCNA are found broadly among many cancer types (Beroukhim et al., 2010).

Amplifications of genomic regions increases gene copy number levels and can consequently over-express proto-oncogenes to drive tumour malignancy. Genomic amplifications may manifest as simple tandem duplications to complex chromosomal rearrangements. Catastrophic events leading to complex chromosomal rearrangements are chromothripsis or BFB cycles. These rearrangements can give rise to focal amplifications, including homogeneously staining regions (HSRs), and extrachromosomal circular DNAs (ecDNAs) (McClintock, 1941; Shoshani et al., 2021; Rosswog et al., 2021).

HSRs do no exhibit the usual chromosomal pattern in G banding images and are homogeneously stained, thus their name. This specific type of amplification is associated with an increased DNA content and gene amplifications (Cowell, 1982; Balaban-Malenbaum, Grove & Gilbert, 1979; Storlazzi et al., 2010). EcDNAs are also associated with massive copy number amplifications, which can exceed copy number levels observable on chromosomes (Kim et al., 2020). In a study by Kim et al. (2020), ecDNAs were identified to be among one of four amplicon classes (circular amplicons (ecDNAs), BFB amplicons, heavily-rearranged amplicons, and linear amplicons) commonly identified in cancer and were found to exist in almost all cancer types. Other amplicon classes are defined based on their genomic characteristics (Kim et al., 2020; Deshpande et al., 2019). Linear amplicons are simple focal somatic amplifications of proximal regions. Heavily-rearranged amplicons describe distant heavily-rearranged regions that jointly form an amplicon. And lastly, BFB amplicons harbour genomic signatures of a BFB (Kim et al., 2020; Deshpande et al., 2019). The analysis by Kim et al. (2020) showed the importance of ecDNAs in cancer and was one of the cornerstone for my investigation of the role of ecDNAs in PDAC.

1.2 Pancreatic ductal adenocarcinoma

1.2.1 Overview and clinical presentation

Pancreatic cancer (PC) is one of the deadliest solid tumour malignancies with a 5-year survival rate of less than around 15% (Siegel et al., 2023; "Pancreatic cancer statistics", 2015). In the UK, PC is the 11th most common cancer type, but the 5th most common cause of cancer-related death ("Cancer Statistics for the UK", 2015) (Figure 1.1). The most common form of PC is PDAC, which contributes to around 90% of all PC cases. Due to the rise of obesity and type 2 diabetes in the western population, the PDAC incidence rate is predicted to increase over the next decade (Mizrahi et al., 2020; Sarantis et al., 2020; Siegel et al., 2023).

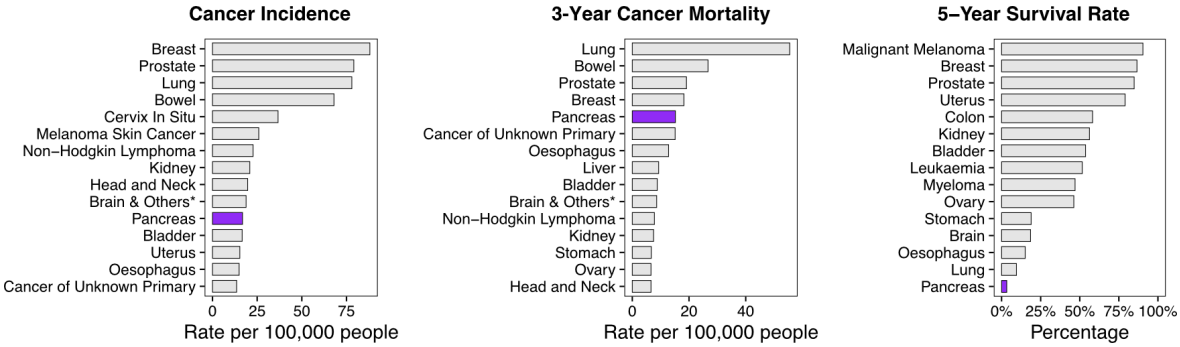


Fig. 1.1 | Pancreatic cancer statistics compared to other cancer types. Cancer Incidence: UK cancer incidence rates of the top 15 most common cancer types (2017). 3-Year Cancer Mortality: UK 3-year cancer mortality of the top 15 most common cancer types (2017-2019). 5-Year Survival Rate: Cancer survival rate in percent of selected cancer types. Data from England and Wales (2010-2011). Brain & Others*: Brain, Other CNS and Intracranial Tumours. Source: "Cancer Statistics for the UK" (2015), accessed on 1 Jun. 2023.

PDAC is usually identified at an advanced stage due to missing standardised methods for early-detection and late onset of symptoms. The only universally applied biomarker used for PDAC diagnosis is carbohydrate antigen 19-9 (CA19-9), which has several limitations and lacks sensitivity and specificity for early detection (Swords et al., 2016). PDAC patients with late-stage disease are mostly faced with a poor prognosis and treated with chemotherapeutic agents such as Gemcitabine or FOLFIRINOX (5-fluorouracil, folinic acid, irinotecan, and oxaliplatin) based on their fitness. In the last few years, more treatment options were tested in clinical trials and Gemcitabine is nowadays used in combination with nanoparticle-albumin bound paclitaxel (nab-paclitaxel). However, the overall survival and survival time only marginally improved over the last years, despite the introduction of new treatment strategies (Mizrahi et al., 2020).

The success of standard-of-care chemotherapy is strongly dependent on the underlying genomics and transcriptomics of PDAC. Novel trials are now investigating the use of precision medicine on actionable genetic alterations. Despite their success in increasing patient survival time, only limited number of patients harbour actionable mutations, which makes them eligible for current personalised treatment options (Springfeld et al., 2023; Aung et al., 2018; Pishvaian et al., 2020). Therefore, novel actionable genetic alterations need to be identified to make personalised medicine available to a broader PDAC patient population.

1.2.2 Genomics

Large genomic sequencing studies have provided valuable insights into the genomic landscape of PDAC, revealing recurrent mutations, copy number alterations, and SVs. These genomic alterations contribute to the tumour initiation and progression, shape the tumour biology, and underpin intra-tumour heterogeneity.

PDAC is defined by a core set of driver mutations, along with a diverse range of less prevalent mutated genes. Among the recurrent mutations, activating *KRAS* mutations are the most prevalent, occurring in around 90% of all PDAC cases and arising early during PDAC development. In addition to *KRAS*, three other key driver genes are frequently altered in PDAC: *CDKN2A* (30 - 50% of cases), *TP53* (60 - 70% of cases) and *SMAD4* (20 - 50% of cases). Inactivating mutations of these drivers play crucial roles in the initiation and progression of PDAC by deregulating the cell cycle, DNA damage response, apoptosis, or cell differentiation (Hruban et al., 2000; Witkiewicz et al., 2015; Bardeesy & DePinho, 2002; Jones et al., 2008; Hu et al., 2021; Biankin et al., 2012).

Beyond these four core driver genes, PDAC is defined by a long tail of low-prevalence mutations, which affect genes such as *KDM6A*, *MLL3*, *ARID1A*, *TGFBR2*, *BRCA1/2*, or *PALB2*. These genes also have an important role in cancer pathways like DNA damage response, TGF-beta signalling, chromatin remodelling (Biankin et al., 2012; Jones et al., 2008; Collisson et al., 2019).

Furthermore, the PDAC genomic landscape extends beyond simple single-nucleotide mutations and is characterised by extensive copy number and structural alterations. These large genomic alterations play a significant role in deregulating key pathways and contributing to PDAC progression (Waddell et al., 2015; Notta et al., 2016).

Copy number alterations, including gene amplifications and deletions, are prevalent in PDAC and have large impacts on gene transcription and ultimately protein levels. Among the most highly affected genes of copy number increases are the oncogenes *MYC* and *ERBB2* resulting in the over-expression of their respective proteins. *MYC* encodes for a transcription factor, which plays a pivotal role in activating the transcription of genes involved in cell growth, proliferation, apoptosis, or cell differentiation (Dang, 1999, 2012). In PDAC, MYC activity is enriched in a PDAC subtype showing the worst patient outcome and is amplified and highly expressed in PDAC metastases, revealing its involvement in disease progression (Sodir et al., 2020; Maddipati et al., 2022; Bailey et al., 2016). Similarly, the amplification and over-expression of *ERBB2*, a regulator of cell proliferation and apoptosis, is associated with a more aggressive disease and poor patient outcome (Ménard et al., 2003; Ortega et al., 2022). On the other side of the copy number spectrum, copy number loss is common in tumour suppressor genes, such as *TP53*, *CDKN2A*, and *SMAD4*, which are also frequently mutated.

Genomic instability represents a common feature of PC defined by an abundance of copy number and structural alterations or the occurrence of catastrophic genomic rearrangements such as chromothripsis. Chromothripsis, a phenomenon where chromosomes undergo shattering and rearrangement, occurs frequently in PDAC with a frequency exceeding 50% (Notta et al., 2016; Cortés-Ciriano et al., 2020; Waddell et al., 2015). These genomic events can result in amplifications, deletions, or knockouts of cancer driver genes leading to fast progression along the disease continuum (Notta et al., 2016). Especially in advanced stage disease, genomic instability in combination with polyploidisiation is common which can drive metastases and progression (Chan-Seng-Yue et al., 2020; Notta et al., 2016). These events are linked to be associated with disrupted *TP53* activity, highlighting the importance of *TP53* in the stability of PDAC tumours (Rausch et al., 2012; Notta et al., 2016).

Overall, understanding the genomics of PDAC is crucial to understand the biology regarding treatment effectiveness. However, with regards to the actionable mutations identified, precision medicine cannot be applied to each patient. Further exploration of the complex genomic landscape and its functional implications is needed to uncover potential therapeutic targets. This can inform the development of more effective treatment strategies for patients with PDAC.

1.2.3 Transcriptomics reveals PDAC subtypes

Over the past few decades, omic technologies have identified a range of signatures that define the PDAC biology. The most notable studies have defined PDAC subtypes from transcriptomic data. While initial research into the PDAC transcriptome employed array-based methodologies, current research overwhelmingly utilises the advanced capabilities of RNA sequencing (RNA-seq) (Collisson et al., 2011; Moffitt et al., 2015; Bailey et al., 2016; Chan-Seng-Yue et al., 2020; Mantione et al., 2014). Transcriptomic subtyping studies have used gene expression profiles to define between two and five distinct subtypes. While the number and the names of the subtypes differ between studies, the actual gene programmes that define each subtype are highly concordant. In all studies one subtype (Basal-like, Squamous, Quasimesenchymal, Pure Basal-like) expressed Basal-like or Squamous-like features and is associated with significantly worse patient outcome. Additionally, another overlapping subtype termed Classical, Pancreatic-Progenitor, or Pure-Classical, expresses endoderm specification transcription factors such as *GATA6*, *HNF1A*, *FOXA2/3*, *HNF4A*, or *PDX1*. The majority of PDAC tumours show a Classical/Pancreatic-Progenitor expression signature which is associated with a better outcome and decreased aggressiveness compared compared to the Basal-like subtype (Bailey et al., 2016; Collisson et al., 2011; Moffitt et al., 2015; Puleo et al., 2018; Collisson et al., 2019).

Overall, defining PDAC subtypes is a prominent technique to reveal tumour groups that may respond to specific chemotherapies (Aung et al., 2018). However, by including other omics technologies, a more comprehensive overview about the PDAC subtypes can be

generated. While each subtype can reveal information about drug susceptibilities, profiling of PDAC tumours may provide individual treatment strategies.

1.3 Extrachromosomal circular DNA

1.3.1 EccDNA vs. EcDNAs - History, nomenclature, and definition

Extrachromosomal circulars DNAs were identified and described under different names starting from the 1960s. The earliest research is dated back to 1964 when Hotta and Bassel (1965) identified DNA circles in wheat nuclei and boar sperm. In the following year, (Cox, Yuncken & Spriggs, 1965) identified paired extrachromosomal chromatin bodies, called double minutes (DMs), in pediatric tumours. By preparing and visualising chromosomes during their metaphase, DNA bodies were identified outside the chromosomes. Due to their miniature size, their chromatin content, and their occurrence in pairs, Cox, Yuncken and Spriggs (1965) termed these elements DMs. It was debated if those extrachromosomal DNAs were of bacterial origin, but they could be found in multiple cells of the same patient, under different conditions, and were absent from interphases or the cell background. Therefore, it was concluded that these are from chromosomal origin. Later research identified that DMs are circular structures and have a large size with more than 100 kbp, which makes them observable in cell metaphases with a light microscope (Cox, Yuncken & Spriggs, 1965; Hahn, 1993). Additionally, DMs do not contain centromeres but do normally replicate during the cell cycle, leading to uneven segregation into the daughter cells (Levan & Levan, 1978; Hamkalo et al., 1985). Overall, DMs were associated with tumour heterogeneity, gene amplification, and drug resistance (Cowell, 1982; Barker, 1982). Rather than DMs, the large, usually amplified, extrachromosomal circular DNAs are now named extrachromosomal circular DNA with the abbreviation ecDNA. Common identification techniques of ecDNAs are whole-genome sequencing (WGS), flourescence *in situ* hybridisation (FISH) of cell metaphases, DAPI staining of cell metaphases, or Circle-seq (sequencing of extrachromosomal circular DNA) (Koche et al., 2020; Cox, Yuncken & Spriggs, 1965; Henssen et al., 2019a; Kim et al., 2020; Turner et al., 2017; Wu et al., 2019; deCarvalho et al., 2018).

On the other hand, many smaller extrachromosomal circular DNA elements were identified which were named based on their structure and sequence. Small polydispersed circular DNA (spcDNA) were identified by electron microscopy and describes small DNA circles with heterogeneous size between 150 and a few thousand bp (Smith & Vinograd, 1972). SpcDNA contain mostly repetitive sequences and occur more frequently in unstable genomes (Gaubatz & Flores, 1990; Cohen, Regev & Lavi, 1997). Another term for small circular DNAs is microDNA, which was established in 2012 to define small circularised DNAs that were isolated from mouse and human cell lines (Shibata et al., 2012). Those have a similar size distribution than spcDNA with an average size of around 200-400 bp (Paulsen et al., 2018; Shibata et al., 2012; Dillon et al., 2015). Despite the size similarities, in comparison to

spcDNA, microDNAs do mostly arise from non-repetitive regions and originate from all parts of the genome (Shibata et al., 2012). However, it is unclear whether the contents and origin is different of spcDNA and microDNAs or if these findings are dependent on the experimental methodology used (Noer et al., 2022). Another type of smaller circles are telomeric circles, which occur in cells with an active alternative lengthening of telomeres (ALT) pathway (Cesare & Griffith, 2004). These circles contain repetitive telomeric nucleotide sequences and can serve as a template for the telomere lengthening, an important step of the immortalisation of tumour cells (Liao et al., 2020; Tomaska, McEachern & Nosek, 2004).

While different nomenclatures were defined for extrachromosomal circular DNA depending on their size, physical features, or appearance, they all can be grouped in two broad classes (Table 1.1). One class contains the large amplified extrachromosomal circular DNAs with specific genes and the diverse roles in cancer, and the other contains all smaller circular DNA classes for which the role is only sparsely defined. In this thesis, I will use the term ecDNA for large amplified circular DNAs that can be identified by either WGS, FISH, or DAPI staining and eccDNA for the whole landscape of eccDNA which can be abundantly identified by Circle-seq. These two groups are also mostly studied individually as their roles are vastly different and they are also separately investigated in this thesis.

Tab. 1.1 | Difference between ecDNAs and eccDNAs. Functions that are still debated are marked with an asterisk. Fluorescence *in situ* hybridisation (FISH), 4′,6-diamidino-2-phenylindole (DAPI).

Function	Size	Copy number	Detection methods	Incidence
EcDNA				
Oncogene Amplification, Elevated Gene Expression, Super-Enhancers, Tumour Heterogeneity, Drug Resistance	Large, usually between 100 kbp and a few Mbp	high	WGS, Long-read sequencing, FISH or DAPI staining	Cancer Cells
EccDNA				
Activate Immune Response, Telomere maintenance*, transcription factor sponging*, miRNA and small gene expression	small, typically ranging from less than 100 bp to 10 kbp	variable, typically low	Circle-seq, ATAC-seq, Long-read sequencing, Inverse PCR	Both Normal and Cancer Cells

1.3.2 Structural features of ecDNA

EcDNA is double-stranded DNA with a circular structure that exist outside the chromosomes, but inside the nucleus (Cox, Yuncken & Spriggs, 1965; Cowell, 1982; Wu et al., 2019). EcDNAs derive from single or multiple chromosomal fragments, which can be rearranged, before circularisation (Kim et al., 2020; Helmsauer et al., 2020; Deshpande et al., 2019). In

most cases, the formation of ecDNA is preceded by chromosomal deletions which provide the necessary material for their formation. Their size ranges from around 100 kbp to a few Mbp, making them observable by light or electron microscopy (Turner et al., 2017; Cowell, 1982; Wu et al., 2019). The genomic content is similar to the chromosomal fragments they are comprised of and their large size allows ecDNAs to bare full genes, especially oncogenes (Vogt et al., 2004; Luebeck et al., 2023). Additionally, ecDNAs are packaged into chromatin consisting of nucleosomes showing their large similarities to chromosomal DNA (Wu et al., 2019). However, despite these similarities, ecDNAs are found to have no centromere, leading to random segregation during the cell cycle (Figure 1.2) (Levan & Levan, 1978; Yi et al., 2022).

EcDNAs are clearly visible in cell metaphases, but are also present in all stages of the cell cycle. During the metaphase, ecDNAs are observed to be located closely around the chromosomes, especially in proximity to their telomeres (Barker & Hsu, 1978). During interphase, however, ecDNAs may form hubs inside the nucleus. These hubs comprise up to 100 ecDNAs and create spatial proximity of enhancers and oncogenes resulting in additional interactions compared to diffused ecDNAs (Hung et al., 2021; Zhu et al., 2021). Zhu et al. (2021) also identified that the transcription of ecDNA-genes is markedly increased when hubs are formed. However, a recent study demonstrated that the ecDNA-gene expression is mostly affected by the copy number amplification whereas hubs do not have an impact on transcription, suggesting that more research is necessary to decipher the roles of ecDNA hubs (Purshouse et al., 2022). Next to ecDNA-ecDNA interactions, ecDNAs are also interacting with chromosomes. The ecDNA-chromosome contacts are associated with transcriptional activity induced by enhancer elements on ecDNAs. EcDNA-based enhancers are mobile inside the nucleus and have the potential to activate genome-wide gene transcription (Zhu et al., 2021). Furthermore, ecDNAs can rewire the chromatin topology leading to transcription activation of oncogenes due to co-occurrence with enhancer sequences on ecDNAs. On chromosomes the normal chromatin topology did not allow enhancer-oncogene contacts, however, by circularisation, the topology is changed mediating enhancer-oncogene contacts and activated oncogene transcription (Morton et al., 2019).

EcDNAs are amplified circular chromatin elements that show highly elevated transcriptional activity compared to other amplification classes (Kim et al., 2020). The high transcriptional activity of ecDNA-residing genes is not only copy number dependent but is also enhanced by an increased chromatin accessibility of ecDNAs compared to their chromosomal counterparts (Wu et al., 2019). In addition, the observed amplification levels of ecDNAs also exceed amplification levels of chromosomal origin (Kim et al., 2020).

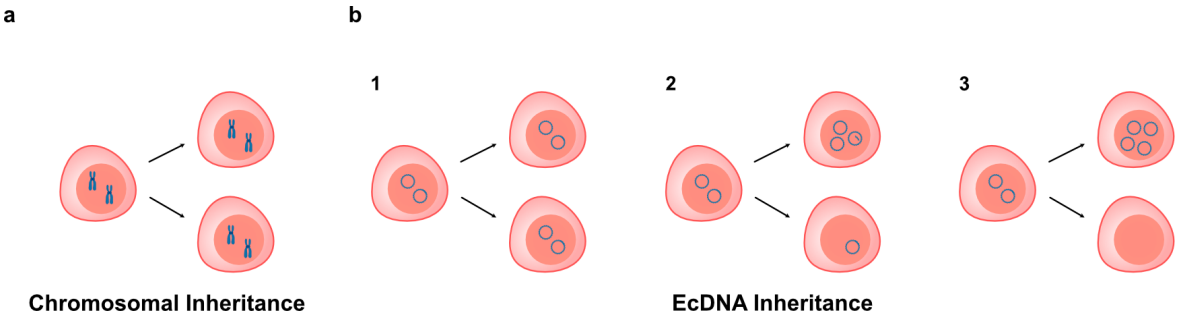


Fig. 1.2 | Segregation schema of chromosomal and extrachromosomal circular DNA. a, Chromosomes are segregated evenly into the daughter cells. Each daughter cell inherits the same amount of chromosomes. **b**, Three potential inheritance options are possible with two ecDNAs in the mother cell. The uneven ecDNA inheritance can lead to ecDNA accumulation or reduction in the daughter cells.

1.3.3 Biological implications of ecDNAs in cancer

Occurrence in cancer

Pan-cancer studies identified that ecDNAs are common in almost all cancer types, albeit with varying frequency levels. Glioblastoma, sarcoma, esophageal, cancer types showing a highly aggressive phenotype, exhibit the highest ecDNA frequency, whereas in normal tissue, ecDNAs are mostly absent (Kim et al., 2020; Turner et al., 2017; Cowell, 1982). In glioblastoma more than 50% of tumours harbour ecDNAs, while cancers of the immune system show minimal ecDNA occurrence. However, these cancer types also have general low amplicon levels suggesting a dependency of ecDNA occurrence with large copy number alterations. Similarly, the highest ecDNA-frequency was found in cancer types where most tumours had at least one amplicon. Here, an amplicon is defined as an highly amplified region with a copy number level above four. PC, comprising different pancreas cancer types, exhibited a 12% ecDNA frequency (Kim et al., 2020).

Importantly, ecDNAs have been detected not only in primary tumours but also in various tumour models, including cell lines, genetically-engineered mouse models, or patient-derived xenografts (Turner et al., 2017; Koche et al., 2020; Cowell, 1982). A study by deCarvalho et al. (2018) showed that ecDNAs are retained in xenograft mice models and neurospheres, which originated from ecDNA-harbouring primary tumours (deCarvalho et al., 2018). These models provide the platform for experimental ecDNA studies investigating the consequences of ecDNA amplification and testing potential therapeutic approaches in ecDNA-harbouring tumours.

Recent technological advances in large-scale ecDNA detection have enabled the comprehensive characterisation of the ecDNA landscape across various cancer types (Kim et al., 2020; Turner et al., 2017; Koche et al., 2020; Luebeck et al., 2023). However, for low-prevalence cancers or cancers without a high ecDNA frequency, such as PC, the characterisation remains limited.

Elevating oncogene transcription

Their chromosomal origin and their large size allows ecDNAs to harbour full protein-coding genes and enhancer elements (Koche et al., 2020; Wu et al., 2019; Kim et al., 2020). From all genes, especially oncogenes are significantly enriched on ecDNAs compared to non-oncogenes (Luebeck et al., 2023). Notably, the oncogenes identified on ecDNAs is dependent on the cancer type investigated (Kim et al., 2020; Koche et al., 2020; Turner et al., 2017). *MYCN*-ecDNAs are common in neuroblastoma (Kohl et al., 1983; Koche et al., 2020), *EGFR*, *CDK4*, and *MYC* on ecDNA are identified in glioblastoma (Zhu et al., 2021), *HER2*-ecDNAs are prevalent in *HER2*-positive breast cancer (Vicario et al., 2015), and *JUP*- and *ERBB2*-ecDNAs are commonly found in oesophageal cancer (Luebeck et al., 2023).

A comprehensive pan-cancer study demonstrated that ecDNA amplifications result in a gene expression increase in comparison to different types of chromosomal amplifications (Kim et al., 2020). The open chromatin structure of ecDNAs is considered to be one possible reason for this increased transcription, as the chromatin shows greater accessibility than the chromosomal origin (Wu et al., 2019). While the gene expression increase is also observed by normalising for copy number levels, significantly higher copy number levels can be observed on ecDNA compared to chromosomal amplicons (Kim et al., 2020; Wu et al., 2019; Luebeck et al., 2023). It is not unusual to observe copy numbers of more than 20 on ecDNAs, which is rare in simple amplicons of chromosomal origin (Luebeck et al., 2023; Wu et al., 2022a; Kim et al., 2020). More complex amplicons, generated by catastrophic events such as BFB, can also lead to these massive copy number levels. However, ecDNAs can also surpass BFB copy number levels, making them a hallmark for massive copy number amplifications (Kim et al., 2020).

EcDNAs mediate tumour evolution and drug resistance

Intra-tumour heterogeneity is affecting patient outcome and drug resistance (Fisher, Pusztai & Swanton, 2013). Tumour sub-populations are defined by genomic alterations that can make them more susceptible or resistant to given therapies (Bedard et al., 2013; Fisher, Pusztai & Swanton, 2013). EcDNAs levels are highly variable based on uneven distribution during the cell cycle, which can increase tumour heterogeneity and determine cell fitness for ongoing survival (Yi et al., 2022; Turner et al., 2017). The fitness of a cell is defined by its underlying genomics and the selection pressure arising from the constantly changing microenvironment, environmental changes resulting from drug treatments, or environment present in metastatic niches (Vanharanta & Massagué, 2013; Vishwakarma & Piddini, 2020; Salgia & Kulkarni, 2018).

EcDNA-dependent tumour heterogeneity originates from the creation of a pool of cancer cells with varying degrees of ecDNA copy number levels (Turner et al., 2017). This pool of cells is challenged by environmental factors removing cells without the required fitness. A study by Nathanson et al. (2014) investigated the dynamic regulation ecDNAs containing

mutant *EGFR*. The study found that glioblastoma cells can dynamically regulate the mutant *EGFR* transcription by having mutant *EGFR* on an ecDNA. During treatment with the *EGFR* tyrosine kinase inhibitor (TKI) erlotinib, the *EGFR* mutant ecDNAs copy number levels are decreased to evade drug-induced cell death. Conversely, after withdrawal of TKI, *EGFR* mutant ecDNAs re-emerge and amplify to drive tumour progression (Nathanson et al., 2014).

The earliest studies describing drug resistance mediated by ecDNA amplification date back to the late 1970s and early 1980s. By treating various cell lines with increasing dosages of methotrexate, amplification of the dihydrofolate reductase (*DHFR*) gene on ecDNA was observed (Kaufman, Brown & Schimke, 1979; Da & Rt, 1981; Kaufman, Brown & Schimke, 1981; Beverley et al., 1984). Methotrexate is and old anti-cancer drug which inhibits *DHFR*, a key enzyme for cell proliferation and growth (Huennekens, 1994). During methotrexate treatment, *DHFR*-containing ecDNAs amplifications arise to substitute the *DHFR* inhibition exhibited by methotrexate (Kaufman, Brown & Schimke, 1979; Da & Rt, 1981; Kaufman, Brown & Schimke, 1981; Beverley et al., 1984). A further study revealed that most methotrexate resistant cell lines harbouring *DHFR*-ecDNAs have partially or completely lost their chromosomal *DHFR* gene, suggesting that the loss of the *DHFR* region preceded the formation of the *DHFR*-ecDNA. This was highlighted by cells displaying intermediate states of *DHFR* containing HSRs, followed by the chromosomal breakage and the generation of a loose *DHFR*-containing fragment, which ultimately was used for ecDNA-generation. (Singer et al., 2000). Interestingly, removal of methotrexate or prolonged methotrexate treatment leads to ecDNA decrease and incorporation of the ecDNA into the genome (Haber & Schimke, 1981). The reintegration of ecDNAs was also observed by Nathanson et al. (2014) and displays one of the mechanisms to decrease gene transcription. However, it seems that most ecDNAs are not reintegrated but lost due to uneven segregation of ecDNAs and the accompanied cell fitness changes (Beverley et al., 1984; Nathanson et al., 2014; Kaufman, Brown & Schimke, 1981; Da & Rt, 1981).

In a study of *HER2*-positive breast cancer models containing *HER2*-ecDNAs, it was found that anti-*HER2* treatment did not cause an ecDNA-dependent *HER2* copy number decrease. While *HER2* protein levels did decrease, the *HER2* ecDNA amplifications remained stable. This suggests that decreasing ecDNA copy number levels is not a universal approach to decrease protein levels and facilitate drug resistances (Vicario et al., 2015).

In summary, multiple studies have shown that ecDNA-dependent gene amplifications can be a major driver of drug resistance and tumour heterogeneity. The potential for dynamically adapting to environmental changes can impact cancer cell fitness and drug resistance highlighting the need for further studies investigating the underlying mechanism and the potential for targeted therapy.

Association with poor patient outcome

EcDNA plays a massive role in cancer and is driving tumour progression. In the pan-cancer study by Kim et al. (2020) it was also found that patients with tumours harbouring ecDNAs are associated with worse patient outcome in comparison to patients with tumours harbouring other chromosomal amplifications (Kim et al., 2020). Furthermore, in a study investigating ecDNAs in Barrett's oesophagus, ecDNAs were significantly enriched in late-stage disease. These findings highlight the importance of further research on ecDNAs as potential therapeutic targets for cancer treatment, which may be applicable for patients with bad prognosis and late-stage cancer.

1.3.4 EcDNA maintenance, targeting, and elimination

EcDNAs are replicated during the cell cycle, like normal chromosomes, which maintains the information of the ecDNA-containing genes (Yi et al., 2022). However, the uneven segregation of ecDNAs during cell division can facilitate accumulation or depletion of genetic information in specific daughter cells, which can lead to cell fitness changes and high intra-tumour heterogeneity (Yi et al., 2022). EcDNAs carry an abundance of cancer driver genes important for tumour progression or drug resistance (Kim et al., 2020; Luebeck et al., 2023). Therefore, cells are required to maintain or enrich ecDNA levels to progress in the disease continuum or resist ongoing drug treatment.

As previously described, ecDNAs are rapidly amplified when under positive selection pressure originating from environmental changes or drug treatment. In contrast, when removing selection pressure ecDNAs and the ecDNA-containing genes levels are decreased (Kaufman, Brown & Schimke, 1981; Schimke et al., 1981; Carroll et al., 1987; Carroll et al., 1988). In a study by Von Hoff et al. (1991) several cell lines containing ecDNAs were analysed. It has been noted that using hydroxyurea, a common anti-cancer drug, on all cell lines, decreased the number of ecDNAs in a concentration-dependent fashion. Higher concentrations of hydroxyurea lead to a greater copy number reduction of ecDNAs and their incorporated genes. On the contrary, chromosomal gene copy numbers were not affected by hydroxyurea (Von Hoff et al., 1991). In a follow-up study by Von Hoff et al. (1992) it was observed that a decrease of ecDNA levels by hydroxyurea can also affect tumourigenesis by reducing *MYC*-ecDNA copy number levels. This effect is, again, not observed when using hydroxyurea on cell lines with chromosomal *MYC* amplifications. It was also reported that the ecDNA copy number loss is associated with the entrapment of ecDNAs within micronuclei (Von Hoff et al., 1992).

Micronuclei are formed by encapsulation of whole chromosomes, chromosomal fragments, or ecDNAs. Micronuclei mostly form through chromosome breakage or mis-segregation of chromosomes, but can also form by encapsulating damaged ecDNAs (Norppa & Falck, 2003; Shimizu, Misaka & Utani, 2007; Schoenlein et al., 2003; Oobatake & Shimizu, 2020). The entrapment of ecDNAs in micronuclei can ultimately lead to ecDNA loss, ecDNA rearrangement, or reintegration into the chromosome (Oobatake & Shimizu, 2020; Shoshani

et al., 2021).

The effect of ecDNA loss is also observed by other cancer therapies. Radiation therapy can significantly decrease the ecDNA and ecDNA-gene copy number levels, which is also associated with micronuclei formation. In a study by Yu et al. (2013) gemcitabine, a cell-death inducing chemotherapeutic drug commonly used in many cancer types, including PC, was used to assess its capabilities to reduce ecDNA copy number levels in comparison to hydroxyurea. While it was found that both can decrease the ecDNA levels, gemcitabine was much more potent and a lower dose was required compared to hydroxyurea (Yu et al., 2013).

Evidently, some anti-cancer therapies have the potential to decrease ecDNA copy number levels to decrease tumour progression and increase drug susceptibility. However, the full mechanism of micronuclei encapsulation and ecDNA loss is still not fully understood. EcDNAs also have the potential to reintegrate into the genome, leading to genomic rearrangements and HSRs. Such reintegration events can be observed with high frequency when DNA damage is induced by Poly (ADP-ribose) polymerase (*PARP*) inhibition (Shoshani et al., 2021).

With regards to their ecDNA dependency, decreasing ecDNA levels and simultaneous reversal of oncogene or drug resistance gene copy number amplifications can inhibit tumour progression and could potentially be interesting for patient therapy. However, the complete mechanism of the dynamic regulation of ecDNA copies has yet to be fully uncovered.

1.3.5 The biogenesis of ecDNAs

The origin of ecDNA is heavily discussed and it appears that multiple forms of DNA damage with subsequent DNA repair contribute to the biogenesis. In the following section, the main types will be discussed, including their roles in cancer (Figure 1.3).

Chromothripsis

Chromothripsis is a complex shattering of chromosome parts or complete chromosomes, which leads to massive chromosomal rearrangements by religation (Leibowitz, Zhang & Pellman, 2015). During the process, shattered chromosomal fragments can also be religated to form an ecDNA (Zhang et al., 2015). This creation is usually accompanied by the loss of the original region on the chromosomes and can form complex ecDNAs comprised of fragments from multiple chromosomes (Stephens et al., 2011; Rausch et al., 2012; Francis et al., 2014). This distinguishes the ecDNAs generated by chromothripsis to ecDNAs originated from more simple DNA damaging events in which the original genomic regions are in close proximity (Storlazzi et al., 2010). Chromothripsis-dependent ecDNA generation can occur through the enclosure of missegregated chromosomes in micronuclei. In a study by Zhang et al. (2015) out of 9 daughter cells derived from cells with ruptured micronuclei, one daughter cell showed evidence for multiple ecDNAs after segregation. Chromothripsis is a common event in cancer and is strikingly associated with 36% of all ecDNAs, suggesting that chromothripsis can

be an ecDNA initiating event for at least 1/3 of all ecDNAs. It is also significantly more abundant in ecDNAs (circular amplicons) than chromosomal amplicons highlighting the effect of ecDNA-based copy number amplification arising from chromothripsis (Cortés-Ciriano et al., 2020; Kim et al., 2020). Chromothripsis can also succeed a BFB cycle event. Shoshani et al. (2021) demonstrated that multiple BFB cycles generated a HSR, which shattered by a chromothriptic event and was rearranged to an ecDNA (Shoshani et al., 2021).

Breakage-fusion bridge cycle

BFB cycle is, similar to chromothripsis, a catastrophic genomic event leading to severe genomic rearrangements and copy number alterations. A BFB cycle is initiated by the telomere loss of a chromosome. Subsequent to the telomere loss, the open chromatid ends can fuse during the cell cycle to form a dicentric chromosome. During anaphase, the dicentric chromosome is pulled by the spindle to the opposite poles of the cell, creating chromatin bridges. Subsequent breakage of the chromatin bridges can lead to either cell death or create gene amplifications (Lo et al., 2002; Murnane & Sabatier, 2004; Guérin & Marcand, 2022). This process can be repeated multiple times creating severe amplifications of genes close to the telomere breakage site, which are exhibited by either HSRs or ecDNAs (Cowell & Miller, 1983). As previously described, BFB cycles followed by chromothripsis can give rise to complex ecDNAs (Shoshani et al., 2021). However, it still needs to be clarified if BFB cycles can give rise to ecDNAs solely from chromosomal deletions during the chromatin bridge breakage without a succeeding chromothriptis event (Hahn, 1993; Noer et al., 2022).

Episome model

The episome model describes the continuous enlargement of smaller circular DNAs, called episomes, to form large ecDNAs. Limited evidence is available for the initial episome formation. One method involves bidirectional replication which leads to the looping out of double-stranded DNA and the formation of an episome (Schimke et al., 1986; Carroll et al., 1988). (Carroll et al., 1987) suggested that episomes are replicating ecDNA precursors containing a replication origin and chromosomal genes. The episome formation is accompanied by the loss of the chromosomal region. Subsequently, the episome can gradually enlarge and become an ecDNA (Carroll et al., 1987; Carroll et al., 1988). While the episome model was originally described in hamster cells, ecDNAs, specifically *MYC*- and *MYCN*-ecDNAs, generated by episome enlargement were also detected in human acute myeloid leukaemia, small cell lung carcinoma, and neuroblastoma cancer cells (Storlazzi et al., 2006; Storlazzi et al., 2010).

Simple DNA damage

The episome model described the successive enlargement of small circular DNAs to ecDNAs. This process is initiated by the deletion, release, and circularisation of a chromosomal fragment (Carroll et al., 1988). A deletion of a chromosomal fragment can be facilitated by simple DNA damage, such as two double-strand breaks flanking the region. In theory, when two breakpoints

are present, the genetic material between them can be head-to-tail repaired, resulting in the formation of an ecDNA. Notably, in an experimental setting, induction of chromosomal deletions via CRISPR-Cas technology has shown to give rise to ecDNAs exceeding 200 kbp in size (Møller et al., 2018b). However, not all ecDNA chromosomal origins are deleted prior ecDNA formation. In a study analysing glioma tumours, *EGFR*-ecDNAs were formed from simple DNA damaging events without the loss of the chromosomal loci (Vogt et al., 2004). This suggests that following post-replicative chromosomal deletions, homology-directed repair (HDR) can repair the deletions using the intact sister chromatid. Additionally, most of the ecDNAs were also formed from a single chromosomal fragment, which was also observed in many ecDNAs in pan-cancer WGS studies (Vogt et al., 2004; Turner et al., 2017; Kim et al., 2020). Nevertheless, further clarification is still needed to identify if the episome model is required for further enlargement to achieve large ecDNA sizes or if simple DNA damage can instantaneously give rise to ecDNAs.

Replication fork stalling and template switching

Replication fork stalling and template switching (FoSTeS) was termed by (Lee, Carvalho & Lupski, 2007) and is a DNA replication mechanism that can create complex SVs, including deletions and rearrangements, in the human genome. During DNA replication the replication machinery can stall caused by DNA lesions and switch to an adjacent DNA template by microhomology, leading to the formation of a single-stranded DNA fragment. This mechanism is thought to be involved in the formation of some ecDNAs in glioblastoma tumours (Yang et al., 2013). A further study by Watanabe et al. (2017) has shown that FoSTeS can lead to the rearrangements of ecDNAs (Watanabe et al., 2017).

Genome instability

Cancer is frequently characterised by genomic instability, a hallmark common in many cancer types (Negrini, Gorgoulis & Halazonetis, 2010). The emergence of ecDNA in cancer is becoming increasingly popular and understanding the underlying mechanism of ecDNA formation is becoming crucial to identify cancer with higher likelihood of ecDNA biogenesis. Recent studies describe that ecDNAs are common in many cancer types, especially cancers with an increasingly unstable genome (Kim et al., 2020; Turner et al., 2017). Chromothripsis, one of the main mechanisms involved in ecDNA formation and a feature of genomic instability, is highly common in samples containing ecDNAs. In a recent study of Barrett's oesophagus, *TP53* mutations were found to be significantly more frequent in tumours containing ecDNAs compared to tumours without (Luebeck et al., 2023). *TP53* inactivation has also been linked to chromothripsis and the formation of ecDNAs (Negrini, Gorgoulis & Halazonetis, 2010; Rausch et al., 2012; Shoshani et al., 2021). Moreover whole-genome doubling, another feature of genomic instability, was commonly observed in ecDNA-positive tumours (Luebeck et al., 2023; Dewhurst et al., 2014).

Replication stress is debated to be also considered one of cancer hallmarks due to its

common occurrence and its implications in genomic instability and tumour progression (Macheret & Halazonetis, 2015). Replication stress can be induced by over-expression of oncogenes driving DNA replication and cell cycle progression (Macheret & Halazonetis, 2015). In a study by Tang et al. (2005), ecDNA formation was identified following the over-expression of *SERTAD1*, a driver of cell cycle progression and DNA repair inhibition. The observed ecDNA formation is likely the result of the genomic instability induced by *SERTAD1* (Sugimoto et al., 1999; Tang et al., 2005; You et al., 2017).

A normal activity of *SIRT1* is required to prevent DNA damage induced by replication stress. By using the replication stress inducer aphodicolin, Utani et al. (2017) observed a significant increase of genomic instability and occurrence of ecDNAs in cells with inactive *SIRT1*, revealing the importance of replication stress and compromised DNA repair in ecDNA formation (Utani et al., 2017).

Overall, the emergence of ecDNAs is associated with genomic instability and replication stress, and understanding the underlying mechanism can potentially help in identify cancers with increased likelihood of ecDNA amplifications to assist in treatment decisions.

1.3.6 EccDNAs - Molecular characteristics and biogenesis

Size

EccDNAs exist in various sizes ranging from a few hundred base pairs to up to a few Mb. While large eccDNAs, called ecDNAs, are considered to have major roles in cancer, they only represent a tiny fraction of the full eccDNA landscape (Møller et al., 2015; Wang et al., 2021; Koche et al., 2020). Most eccDNAs are small with a length of less than a 1,000 bp. Recent Circle-seq studies, characterised the eccDNA landscape in various organisms and a median eccDNA size of around 200 - 1500 bp was identified depending on the organism and cells analysed (Wang et al., 2021; Møller et al., 2018a; Shibata et al., 2012). These small eccDNAs can originate from all parts of the genome and are abundantly found in cancer and normal cells (Møller et al., 2015; Møller et al., 2018a; Shibata et al., 2012).

Origin and formation patterns

EcDNAs are described to contain and amplify oncogenes to drive cancer progression (Kim et al., 2020; Turner et al., 2017; Luebeck et al., 2023). However, the origin of eccDNAs, in general, is more diverse and eccDNAs can originate from all parts of the genome (Møller et al., 2015; Wang et al., 2021; Møller et al., 2020). Studies investigating the broad eccDNA landscape use the power of Circle-seq, the sequencing of eccDNA-enriched DNA isolates.

In these studies it has been identified that eccDNAs form across the whole genome, but it has also been noted that eccDNAs are preferentially formed from specific genomic regions (hotspots), genomic elements, or gene elements (Møller et al., 2015; Møller et al., 2018a; Shibata et al., 2012). EccDNAs commonly arise from genic regions and are enriched in

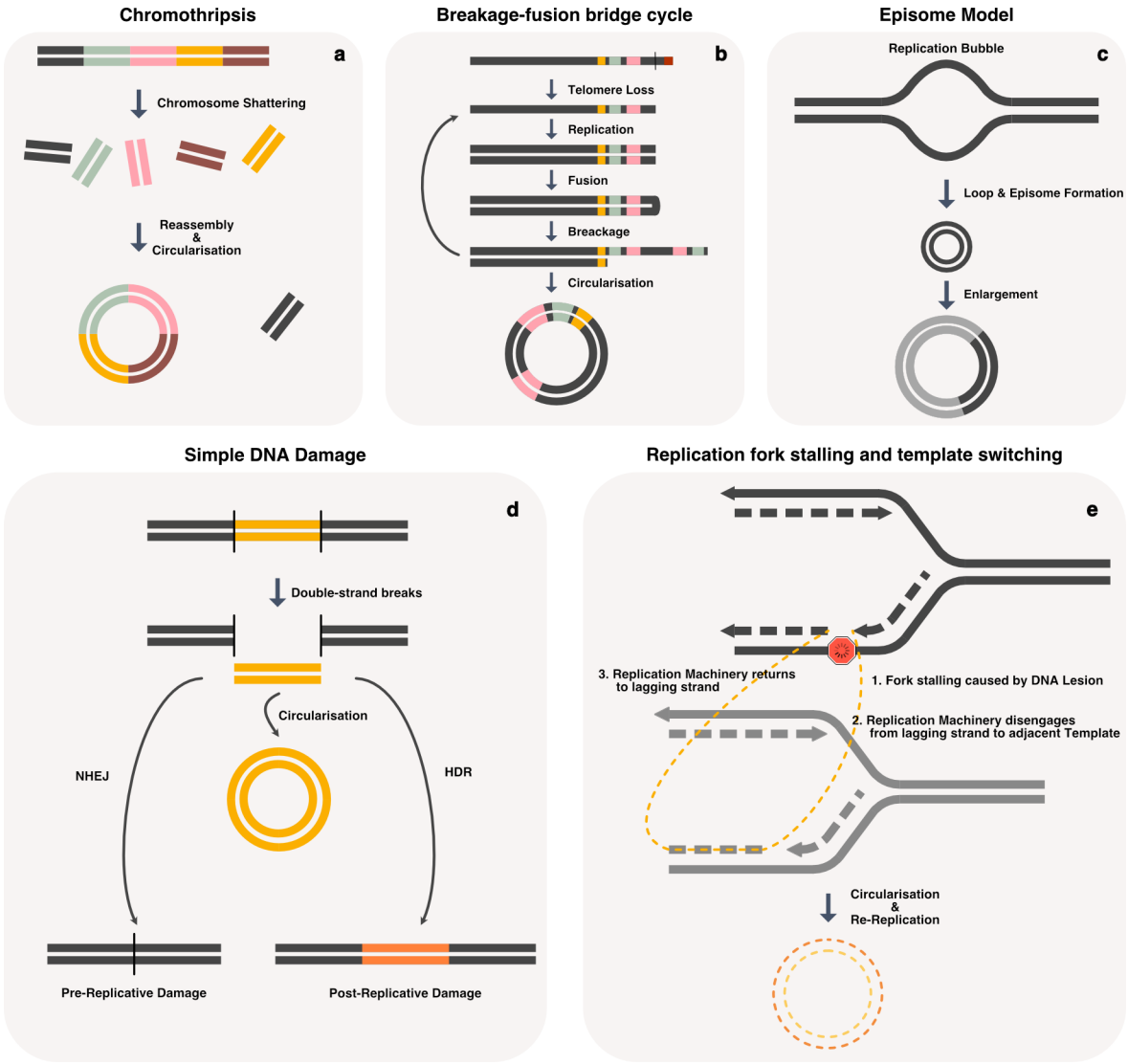


Fig. 1.3 | EcDNA formation mechanisms. **a**, chromothripsis shattering of whole or parts of chromosomes generates chromosomal fragments resulting in massive rearrangements and can form ecDNAs. **b**, Breakage-fusion bridge (BFB) cycle is initiated by telomere loss leading to rearrangements of telomere adjacent regions. The cycle of replication, fusion, and breakage can be repeated multiple times, creating opportunities for the formation of an ecDNA by circularisation of chromosomal fragments. **c**, The episome model describes the episome formation by looping out of a replication bubble followed by the successive enlargement of the episome to form an ecDNA. **d**, Simple DNA damage with two double-strand breaks can create a free chromosomal fragment that can be head-to-tail religated to generate an ecDNA. Due to the timing of the DNA damage, different DNA repair mechanisms are employed to repair the open chromosome ends. Pre-replicative damage can be repaired by non-homologous end-joining (NHEJ), which may result in the loss of the chromosomal region. On the other hand, post-replicative chromosomal damage will be completely repaired by homology directed repair (HDR) due to the existence of a sister chromatid. **e**, Replication fork stalling and template switching is discussed (Wu et al., 2022a) to also lead to ecDNA formation. During the process, a single-stranded DNA fragment is generated which can circularise by religation and become double-stranded DNA through replication.

specific genic elements like the 5' untranslated region (5'UTR) or exons (Shibata et al., 2012). A majority of eccDNAs also contain repeat elements, including long- and short-interspersed nuclear elements (LINEs and SINEs) (Møller et al., 2018a; Møller et al., 2015). Furthermore, eccDNAs contain an increased GC content compared to non-eccDNA flanking regions and are associated with an enrichment of CpG islands (Shibata et al., 2012).

Møller et al. (2018a) also identified the association between gene richness and eccDNA

formation and the association with transcription. Gene-rich chromosomes form more eccDNAs than gene-poor chromosomes and eccDNAs are commonly identified from highly transcribed genes (Møller et al., 2018a).

These Circle-seq studies have shown that eccDNAs are not random, but originate from specific hotspots. However, a study by Wang et al. (2021) did not find such patterns. Wang et al. (2021) used long-read sequencing technology in conjunction with the Circle-seq protocol and identified 1.6 million single- and multi-fragment eccDNAs. These eccDNAs were found to map across the whole genome without any noticeable pattern or region of eccDNA hotspots. Furthermore, it was noted that inducing apoptosis enhances eccDNA formation suggesting that eccDNAs are apoptotic products formed after apoptosis-induced DNA breakdown (Wang et al., 2021).

Taking together, Circle-seq studies defined the eccDNA landscape and potentially identified hotspots of preferential eccDNA formation. However, the origin and the mechanism is not yet completely understood. It is also unclear whether the cell type or the tissue of origin affect the eccDNA hotspots. Interestingly, Koche et al. (2020) reported that neuroblastoma samples have an abundance of eccDNA around the *MYCN* locus, which is one of the oncogenes recurrently located on ecDNAs and driving neuroblastoma progression. This suggests that hotspots for ecDNA and eccDNA origin might overlap and could be cancer type specific (Koche et al., 2020; Huang & Weiss, 2013).

1.3.7 Function of small non-amplified eccDNAs

While the roles of ecDNAs in cancer are broadly established, the role of smaller non-amplified eccDNAs is not yet fully understood and needs to be further investigated. EccDNA research in cancer cells is also underrepresented and most studies use yeast, plant, or non-cancerous human or mouse cells (Wang et al., 2021; Gaubatz & Flores, 1990; Noer et al., 2022). However, these findings might be also applicable for their roles in cancer.

Most eccDNAs do not contain whole protein-coding genes, in contrast to ecDNAs, due to their small sizes. However, they can contain gene elements, such as exons, or small micro RNAs (miRNAs) (Koche et al., 2020). In a study by Paulsen et al. (2019), artificially synthesised eccDNAs containing gene parts were used to study eccDNA-based transcription of partial genes without promoter sequences. It has been reported that transcription occurs on eccDNA in vitro and in vivo, in contrast to linear DNA, even if the gene is not fully contained. The transcription of partial genes can produce miRNAs or small interfering RNAs (siRNAs) that regulate gene expression (Paulsen et al., 2019). Furthermore, eccDNAs can contain enhancer sequences to activate gene transcription. The mobility of eccDNAs in the nucleus allows eccDNA-bound enhancers to activate gene transcription genome-wide without any spatial restrictions. This was identified for ecDNAs, but could also be applicable to smaller eccDNAs (Zhu et al., 2021).

Telomere maintenance is important for the long-term cell division of cells. In cancer, stable telomere maintenance be facilitated by expression of telomerase, which is usually active in stem cells to prevent senescence (Shay & Wright, 2011). If telomerase activity is low, cancer cells can also use the alternative lengthening of telomeres (ALT) pathway to maintain the telomere. This mechanism is dependent on the presence of extrachromosomal DNA containing repetitive telomere sequences that are used as template to lengthen the chromosomal telomeres. This extrachromosomal DNA can be linear or circular (eccDNAs) (Cesare & Reddel, 2010; Huang et al., 2017).

EccDNAs may also play a role in activating the innate immune system. Wang et al. (2021) identified that eccDNAs are apoptotic products which stimulate the innate immune system in a more potent fashion in contrast to comparable linear DNA. EccDNAs can be found in the medium of apoptotic cells and activates proteins of the innate immune pathway (Wang et al., 2021). These results opened a potential new way of improving patient treatment by activating cancer cell apoptosis, resulting in eccDNA formation, and ultimately activating the innate immune system to activate cancer cell killing.

1.3.8 EccDNA research in pancreatic cancer

PC is a complex disease commonly characterised by genomic instability and chromosome abnormalities (Notta et al., 2016; Campbell et al., 2010; Griffin et al., 1995). By cytogenetic investigation of cell metaphase images, Griffin et al. (1995) described the existence of ecDNAs in 8 of 62 primary pancreatic adenocarcinoma tumours (Griffin et al., 1995). This revealed that ecDNAs are also a common feature (12.9% of cases) in PC. Kim et al. (2020) identified similar occurrence rate by investigating amplicon classes in cancers of the pancreas (Kim et al., 2020).

A further study by (Notta et al., 2016) identified ecDNAs in the primary tumour and the matching metastasis of a PC patient. The ecDNA identified harboured *MYC*, which was massively amplified in the primary tumour and the metastasis suggesting its role in disease continuum (Notta et al., 2016; Maddipati et al., 2022).

Next to these two studies, only limited evidence of the role of ecDNAs and the abundance is available for PC. Despite its common occurrence ($>$ 10% of PC cases), ecDNAs are mostly investigated in other cancer types, which show higher ecDNA occurrence rates and have ecDNA-positive model system (Koche et al., 2020; Turner et al., 2017; Kim et al., 2020).

PC is a highly lethal disease which is usually treated with standard-of-care chemotherapy, which prolongs the life of patient marginally. Targeted therapy, which increases survival time compared to standard-of-care therapy, is available only to a quarter of patients harbouring actionable mutations (Kamisawa et al., 2016; Pishvaian et al., 2018; Pishvaian et al., 2020). Identifying susceptibilities of ecDNA-positive pancreatic tumours might make targeted therapy available to more patients. However, as mentioned earlier, the ecDNA research is limited in

PC and a complete characterisation of ecDNAs in PC is necessary to identify the potential use in clinical settings.

1.3.9 Techniques for the identification and analysis of eccDNAs

Cytogenetic methods

EcDNAs are large extrachromosomal elements already visible by light microscopy (Cox, Yuncken & Spriggs, 1965). However, during most parts of the cell cycle, ecDNAs are intermingled with the chromatids and are not easily distinguishable from chromosomal regions (deCarvalho et al., 2018). Therefore, cells need to be imaged in the metaphase to determine the extrachromosomal nature of ecDNAs and their specific genomic content. To enrich for metaphase cells, cells are usually prepared with a cell cycle inhibitor, such as colcemid, arresting the cells in the metaphase (Cox, Yuncken & Spriggs, 1965). These metaphase spreads are usually stained using DAPI (4′,6-diamidino-2-phenylindole), a blue fluorescent DNA-binding molecule, to distinguish between DNA and contaminations or artefacts. However, DAPI staining is not sufficient to identify the ecDNA content (Turner et al., 2017; deCarvalho et al., 2018). To determine the genomic content of ecDNAs, the cell metaphase spreads can be combined with fluorescence in-situ hybridisation (FISH) probes, which bind to specific genes or chromosomal regions. FISH is one of the main techniques to identify a gene on an ecDNA and quantify the ecDNA and gene copy number levels (Rayeroux & Campbell, 2009; Wu et al., 2019).

In combination with computational methods, ecDNA detection can be automated to facilitate high-throughput ecDNA identification and quantification. EcSeg, a deep learning algorithm trained on DAPI and FISH metaphase spreads, is able to identify and quantify ecDNAs and gene amplifications on ecDNAs (Rajkumar et al., 2019). Similarly, ECdetect can quantify ecDNAs from DAPI stained metaphases (Turner et al., 2017). Despite the promises, the output quality is highly dependent on the input dataset for both datasets, which could lead to error-prone quantification and false-positives or false-negatives (Rajkumar et al., 2019; Turner et al., 2017).

Sequencing approaches

Different sequencing approaches have been used to study ecDNAs. While sequencing technologies have revolutionised many areas of cancer research, the ecDNA field required the development of new computational methodologies for accurate identification. In 2017, Turner et al. (2017) presented 'AmpliconArchitect', which uses WGS data and identifies ecDNAs by detecting amplicon segments that form a head-to-tail fusion (Figure 1.5a) (Turner et al., 2017; Deshpande et al., 2019). This head-to-tail structure is characteristic for ecDNAs based on their circular structure, but can also be generated by tandem duplications. However, by investigating only amplified regions, ecDNAs can be accurately determined with high frequency. Additionally, the investigation of amplicon regions allows the use of low-coverage WGS data,

of around 1-10x, which significantly reduces the cost-per-genome (Kim et al., 2020; Turner et al., 2017; Deshpande et al., 2019). With the development of AmpliconArchitect, ecDNAs were detected in many cancer types and widened the field for new ecDNA research (Kim et al., 2020; Luebeck et al., 2023; Wu et al., 2019; Zhu et al., 2021; Hung et al., 2021).

By combining the power of AmpliconArchitect's ecDNA detection using WGS data with optical mapping data of long DNA fragments (< 150 kb), AmpliconReconstructor is able to resolve the full ecDNA structure (Luebeck et al., 2020). By using WGS data alone, the complete ecDNA structure can only be partially resolved and AmpliconArchitect outputs many potential structures, which could co-exist or could be fused to generate a larger ecDNA (Deshpande et al., 2019; Kim et al., 2020; Luebeck et al., 2020).

All computational tools analysing sequencing data to identify ecDNAs are based on SV identification generated by head-to-tail fusions. These head-to-tail fusions, which are generated through ecDNA formation, are found across the whole genome making whole-exome sequencing (WES) impracticable for ecDNA identification (Kim et al., 2020). A high ecDNA detection accuracy can then be achieved by using WGS data and investigating only amplified regions (Turner et al., 2017; Deshpande et al., 2019). However, multiple software were developed that call ecDNAs based on simple identification of head-to-tail fusions without the need of copy number information. This can lead to falsely classifying ecDNAs due to its structural similarities to tandem duplications (Figure 1.5b) (Kumar et al., 2020; Møller, 2020). Therefore, it is advised to use sequencing data that is enriched with reads originating from eccDNAs. By using an plasmid-safe exonuclease, linear (chromosomal) DNA will be removed and circular DNA (eccDNA, ecDNA, mitochondrial DNA) will be retained. Subsequent circular DNA amplification, using rolling-circle-amplification, achieves high enrichment of eccDNAs which can then be subjected to sequencing (Figure 1.4). This established method is named Circle-seq (Møller, 2020; Koche et al., 2020; Møller et al., 2015). By integrating Circle-seq data with WGS data, Koche et al. (2020) identified that Circle-seq achieves high accuracy in determining the eccDNA and ecDNA landscape (Koche et al., 2020). However, eccDNA enrichment can also be incorporated into other sequencing methods such as ATAC-seq (Kumar et al., 2020).

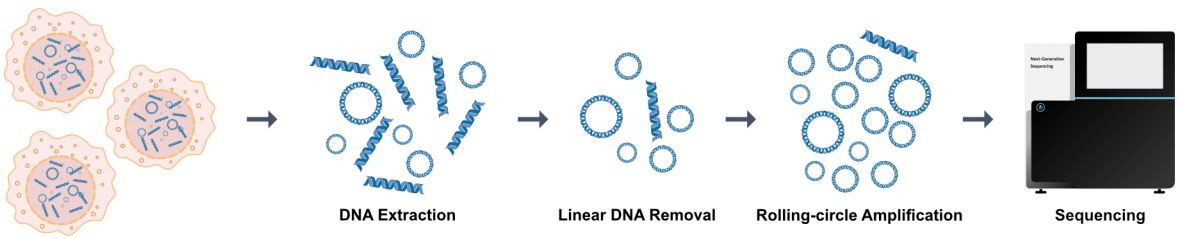


Fig. 1.4 | Circle-seq procedure. Circle-seq describes an eccDNA-specific sequencing method. During the procedure, extracted DNA will be treated with plasmid-safe exonuclease to remove linear DNA. Subsequent rolling-circular amplification amplified the remaining eccDNAs which are used for sequencing.

The analysis of Circle-seq data can be performed using various software which identify the head-to-tail fusions (eccDNA junctions) (Figure 1.5b). Three of the most prominent

tools are 'Circle_finder' (Kumar et al., 2020), 'Circle-Map' (Prada-Luengo et al., 2019), or 'CIRCexplorer2' (Zhang et al., 2016). By only identifying a putative head-to-tail fusion which can be generated by eccDNA formation, they are ideally used to identify single-fragment eccDNAs. Single-fragment eccDNAs are eccDNAs that were generated from a sole loose fragment that is circularised by fusing the start and the end sequence. If an eccDNA originated from multiple fragments, the eccDNA junction does reveal the accurate eccDNA content and therefore other methods need to be employed.

Multi-fragment eccDNAs are eccDNAs that are generated from more than one chromosomal fragment, which are joined together through multiple DNA repair events. To my knowledge, no method or software has been established that can identify these easily. In theory, a similar algorithm to the AmpliconArchitect algorithm could be employed that uses regions of Circle-seq read coverage and joins them by identifying region-spanning SVs. However, an adaptation still needs to be developed. One way of identifying multi-fragment eccDNAs could be the use of software that *de novo* assembles sequencing reads to identify the full sequence of an eccDNA (Figure 1.5c). This method is usually performed to identify novel genomes for which no reference genome is available (Paszkiewicz & Studholme, 2010). Due to the circular structure of eccDNAs, only *de novo* assembled circular sequences should be considered whereas linear sequences need to be removed. Hence, *de novo* assembly software developed for bacterial genomes, which share the circular characteristic, can be used for identifying eccDNAs, specifically eccDNAs originating from multiple fragments. For short-read sequencing data, this can be performed by 'Unicycler' (Wick et al., 2017). However, validation of using this method for eccDNA detection still needs to be performed.

De novo assembly using short-read sequencing data is sub-optimal due to the high coverage necessary to achieve the full assembly of an eccDNA. If a conjoining read is missing, the full sequence can not be assembled resulting in eccDNA detection failure. Therefore, recent studies use long-read sequencing data from Nanopore or PacBio technologies to identify the correct eccDNA structure (Koche et al., 2020; Wang et al., 2021; Chitwood et al., 2023). Long-read sequencing is still markedly more expensive than short-read sequencing, but a shift towards long-read sequencing is expected to become the gold-standard for the eccDNA landscape identification as it mitigates some of the challenges of short-read sequencing (De Maio et al., 2019; Amarasinghe et al., 2020).

With the introduction of analysis tools the identification of ecDNAs and eccDNAs from sequencing data became accessible for a broader range of researchers. Currently, several state-of-the art tools exist that can accurately detect ecDNAs and eccDNAs from various sequencing datasets (Møller, 2020; Kim et al., 2020; Prada-Luengo et al., 2019; Zhang et al., 2016; Kumar et al., 2020). However, these tools often require distinct input data formats and rely on specific software dependencies to generate their respective outputs. Additionally, with the rise of big data, it is becoming essential to parallelise software processes to increase efficiency and reduce computational time. Unfortunately, most existing software in the circular

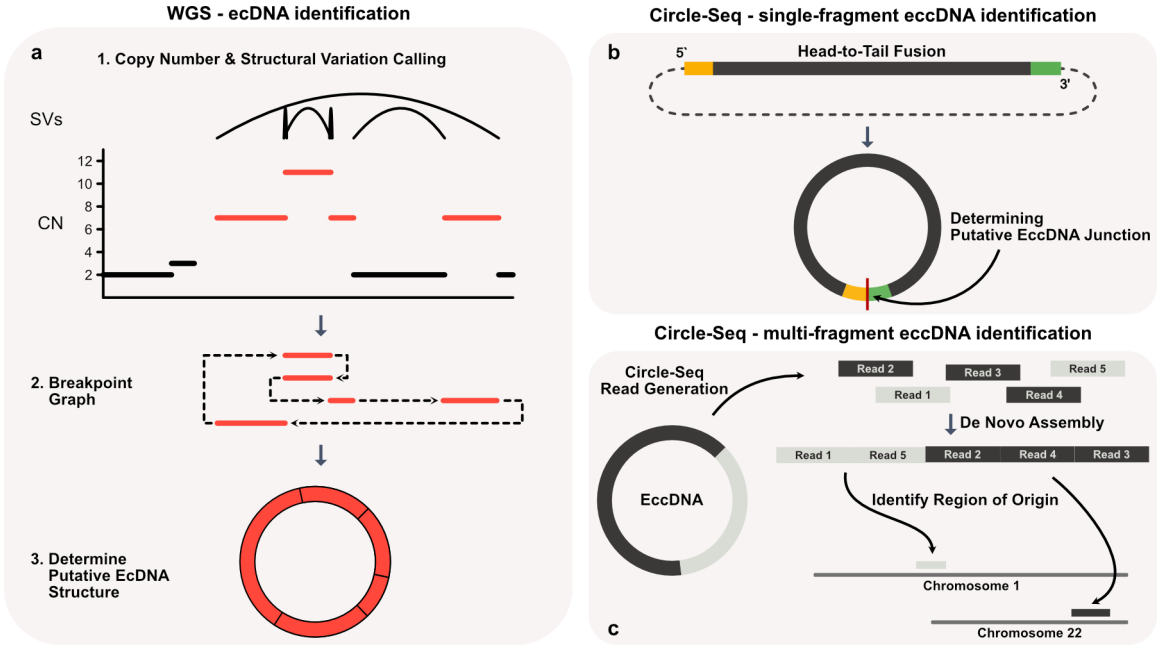


Fig. 1.5 | EcDNA & EccDNA Identification with short-read sequencing data. **a**, The ecDNA structure is delineated by identifying amplified regions, from copy number (CN) calls, and structural variations (SVs) connecting the regions to each other. This generates a breakpoint graph of the region arrangement. A breakpoint graph creating a circular amplicon could be originating from an underlying ecDNA. This method was developed by Deshpande et al. (2019). **b**, Single-fragment eccDNAs can be identified from Circle-seq data by identifying a putative eccDNA junction which originated from a head-to-tail fusion. **c**, The potential identification multi-fragment eccDNA can be facilitated by *de novo* assembly of eccDNAs with subsequent reference genome mapping to identify the eccDNA origin.

DNA field is not optimised for large datasets and typically runs samples individually.

In conclusion, to tackle those challenges, new workflows need to be developed that efficiently manage software dependencies and enable process parallelisation. By organising and streamlining the analysis pipeline, these workflows will enable the efficient analysis of large-scale datasets and increase user accessibility for researchers of all backgrounds.

1.4 Aims & objectives

Within this introduction several knowledge gaps about eccDNAs have been identified, specifically regarding PDAC. PDAC is under-represented in eccDNA studies and it has become evident that eccDNAs can have cancer-specific contents or roles. Therefore, the first aim of this thesis is to identify the ecDNA/eccDNA landscape in PDAC and characterise their origin, their association with characteristic features of PDAC, and their potential roles in the disease. This can help in identifying potential therapeutic opportunities and enhance our understanding of the complex genomics of PDAC. To achieve this aim several objectives were defined that will be pursued and described throughout this thesis.

- **Determine the prevalence of ecDNAs in PDAC:** This objective aims to identify the prevalence rate of ecDNAs in PDAC to determine how common they are. Additionally, integrative analysis aims to reveal patient subgroups with a higher likelihood of containing ecDNA-positive PDAC tumours. This will provide insight into the clinical relevance of ecDNAs in PDAC.
- **Characterise the landscape of ecDNAs in PDAC:** This objective is to determine the location, size, and occurrence of genomic feature on ecDNAs, including the presence of PDAC-specific cancer drivers.
- **Investigate the potential roles of ecDNAs in PDAC:** This objective is to analyse the ecDNA landscape and integrate it with relevant biological and clinical data to identify the potential roles in PDAC. This will include the correlation of ecDNA information with gene expression data and copy number data to verify established roles of ecDNAs in other cancer types. Furthermore, integration with clinical data will provide insights into the association between ecDNA presence and clinical outcomes or disease stage. Lastly, using sequencing data generated from organoids under stress conditions will explore the roles of ecDNAs as an adaptation mechanism.
- **Determine the usability of PDAC model systems in ecDNA research:** This objective is to utilise sequencing data from PDAC patient-derived organoids (PDOs), PDO-derived cell lines, and matching primary tissue to characterise ecDNAs in different PDAC samples. This includes the comprehensive comparison of the ecDNA landscape in matching samples to assess if model systems are usable for ecDNA research in PDAC.
- **Characterise the landscape of eccDNAs in PDAC:** This objective is to assess the general landscape of eccDNAs in PDAC by characterising multiple PDAC models on their eccDNA occurrence and identifying putative roles of eccDNAs.
- **Determine the characteristics of eccDNAs in PDAC:** This objective aims to evaluate specific characteristics of eccDNAs and compare it to previous literature about other cancer types. This includes examining the association with different PDAC features and eccDNA biogenesis, identifying hotspots of eccDNA abundance, and determining if eccDNAs are retained to play a role in the tumour biology.

My second aim is to make eccDNA research accessible to a wider proportion of the scientific community and increase reproducibility, transparency, and efficiency. To achieve this, a comprehensive workflow will be developed that includes a range of ecDNA/eccDNA software tools to generate ecDNA/eccDNA information from short-read sequencing datasets. To accomplish these aims, several objectives have been determined and will be chased throughout this thesis:

- **Create a dynamic and scalable workflow of several software packages used in the ecDNA/eccDNA field:** This objective is to utilise the workflow manager Nextflow (Di Tommaso et al., 2017) to write a workflow concatenating relevant software tools to identify ecDNA/eccDNA.
- **Enable customisation and flexibility in the pipeline:** This objective aims to include options to adjust for the user's needs. This will involve the inclusion of user-definable parameters inside the pipeline which can be separately specified.
- **Include control measurements to ensure the correct configuration and deployment of the workflow:** This objective is to verify that the all parameters, input files, and output options are correctly defined. Additionally, the workflow aims to have several quality control tools in place to check data quality.
- **Verify the workflows performance on real sequencing datasets:** This objective is to ensure that the workflow was correctly developed and verifies the usability of all ecDNA/eccDNA tools.
- **Enable rigorous testing and peer reviewing of the workflow:** This objective is to allow continuous development and assess the quality of the workflow. This will include making the workflow available for open review and requiring stringent testing and reviewing by independent individuals.
- **Provide documentation and a user-friendly interface:** This objective aims to increase usability by a wider audience. This will include providing sufficient documentation for all stages of the pipeline, such as input data, temporary, and output data, but also on how to effectively run the data on individual computational resources.
- **Validate the accuracy of tools not defined for eccDNA analysis:** This objective aims to verify software tools which are currently not published to be used for ecDNA/eccDNA data analysis. This will include validating the computational results by laboratory techniques.

Chapter 2

Methods

> Experiments are the only means of knowledge at our disposal. The rest is poetry, imagination.
>
> *Max Planck*

2.1 nf-core/circdna

nf-core/circdna is a Nextflow pipeline written with the nf-core (https://nf-co.re/) template. The pipeline is described in Results Chapter 3. The pipeline requires Nextflow version >= ’22.10.1’. Table 2.1 describes the version of each software used in the pipeline.

Tab. 2.1 | Software and their respective versions used in the nf-core/circdna pipeline.

Software	Version	Reference
Programming Language		
NEXTFLOW	22.10.1	Di Tommaso et al. (2017)
Quality Control		
FASTQC	0.11.9	Andrews et al. (2010)
MULTIQC	1.12	Ewels et al. (2016)
Read & Adapter Trimming		
TRIM GALORE	0.6.7	Krueger (2015)
CUTADAPT	4.1	Martin (2011)
Mapping & Processing		
BWA	0.7.17-r1188	Li (2013)
SAMTOOLS	1.15.1	Danecek et al. (2021)

Continued on next page

Software	Version	Reference
PICARD	2.26.10	http://broadinstitute.github.io/picard/
Extrachromosomal circular DNA calling		
CIRCLE-MAP	1.1.4	Prada-Luengo et al. (2019)
CIRCEXPLORER2	2.3.8	Zhang et al. (2016)
CIRCLE_FINDER	commit: 3eb333d	Kumar et al. (2017)
UNICYCLER	0.5.0	Wick et al. (2017)
SEQTK	1.3	Li (2012)
MINIMAP2	2.21	Li (2018)
CNVKIT	0.9.9	Talevich et al. (2016)
AMPLICONARCHITECT	1.3_r1	Deshpande et al. (2019)
AMPLICONCLASSIFIER	0.4.5	Luebeck et al. (2023)

2.2 Cell culture methods

All cell tissue handling procedure that require aseptic techniques were conducted under sterile conditions within class II biological safety cabinets. Prior to use, all work surfaces and equipment were sterilised using 70% ethanol.

2.2.1 Cell culture

The Kinghorn Cancer Centre (TKCC) patient-derived cell lines (PDCLs) were cultured in their respective media and conditions described in Hardie et al. (2017). PaCaDD137 was cultured in the media formulated by Rückert et al. (2012). Mayo-4636 was cultured in the Mayo Media detailed in Table 2.5. The media components for all PDCLs are detailed in Methods Table 2.5. The cells were grown at 37°C and 5% CO_2. Cell lines containing low oxygen (LO) specification (e.g. TKCC-15-LO) were grown under low oxygen conditions at 37°C with 5% O_2 an% CO_2 (Table 2.2). Cells were split every 3-5 days with a 1:3/1:4 ratio. The PDCLs were a kind gift from Holly Brunton, Irati Ricón Santoyo, and Carlotta Cattolico, Cancer Research UK Beatson Institute. All PDCL cultures were originally seeded with around 500,000 cells.

2.2.2 Cell harvest

The cells were harvested at approximately 80% confluency through trypsinisation. In detail, the cell media was aspirated and cells were washed with warm (37°C) PBS (Sigma-Aldrich) piror trypsinisation. Trypsinisation was performed using 3-10 mL warm (37°C) 1x Trypsin (Gibco™) for 3-10 minutes at 37°C. Complete cell detachement was confirmed using a light microscope, and the cell count was determined. The detached cells were diluted 1:3 with their respective media and centrifuged at 400 relative centrifugal force (RCF) for 5 minutes. The supernatant was removed, and the cell pellets were resuspended into PBS and transferred into one to three 2 mL Eppendorf tubes, aiming for a total cell count ranging from 1 to 10 millions

cells per tube. Subsequently, the cell suspensions were centrifuged at 400 RCF for 5 minutes, and the supernatant was aspirated. The final cell pellets were stored at -80°C until further use.

2.3 Human specimens and clinical data

PDAC tissue samples were acquired from the General and Pancreatic Surgery Unit of the University of Verona. Prior to the tissue acquisition, patients provided written informed consent. The collection of fresh tissues for patient-derived organoids (PDOs) were performed by the Vincenzo Corbo Lab, Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy, under a study that received approval from the Integrated University Hospital Trust (AOUI) Ethics Committee (Comitato Etico Azienda Ospedaliera Universitaria Integrata): approval number 1911 (Prot. n 61413, Prog 1911 on 19/09/2018). Formalin-fixed and paraffin-embedded tissues were also collected from the ARC-NET Biobank in accordance with protocol number 1885 approved by the AOUI Ethics Committee.

2.4 PDO establishment and culture

The following procedure was performed by the Vincenzo Corbo Lab, Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy. PDAC PDOs were established using procedures previously published by Boj et al. (2015) (Boj et al., 2015). Prior to establishment, pathologists examined the specimens used to generate PDOs to verify the existence of neoplastic cells. To propagate organoids, confluent organoids were removed from Matrigel®, dissociated into small clusters by pipetting and resuspended in an appropriate volume of fresh Matrigel®. All organoid models were acquired as part of the Human Cancer Model Initiative (HCMI) (https://ocg.cancer.gov/programs/HCMI) and are available for access from the American Type Culture Collection (ATCC). The corresponding IDs and clinical data are listed in the (Extended Data Table 1). Organoid cultures were passaged once a week with a splitting ratio of 1:3 in +WR (Wnt3A and R-spondin 1 containing human complete media) or human depleted Media (-WR, human complete media without Wnt3A and R-spondin 1 conditioned media) (Boj et al., 2015). To establish WR (Wnt3A and R-spondin 1) independent PDOs, organoids established and propagated in human complete media (+WR) were placed and maintained in -WR for several passages. Due to cell death induced by -WR, the media was refreshed every 3 days and Matrigel® was refreshed every 14 days without propagating the cultures until WR-independent PDOs emerged. To obtain 'late-passage' PDOs, organoids were passaged at least 40 times in +WR medium after establishment. Organoids were routinely tested for the presence of mycoplasma contamination using the Mycoalert Mycoplasma Detection Kit (Lonza). For Gemcitabine treatment, VR01-O was in culture for 56 days before pellet collection with adding 2.5 nM Gemcitabine every third day. PDO spliiting was performed when necessary.

2.5 Organoids metaphase spreads

The following procedure was performed by the Vincenzo Corbo Lab, Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy. Organoids were incubated with Colcemid (1 µg/mL, Gibco) in culture medium at 37°C, 5% CO_2 overnight. After incubation, organoids were dissociated as previously described. Briefly, single cells were incubated in hypotonic solution (potassium chloride 0.56% and sodium citrate 0.8%) for 20 minutes at room temperature. The metaphases were then fixed in ice-cold methanol-acetic acid (3:1), washed with methanol-acetic acid (2:1) and dropped onto adhesive slides.

2.6 DNA fluorescence *in situ* hybridisation

The following procedure was performed by the Vincenzo Corbo Lab, Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy. DNA fluorescence *in situ* hybridisation (FISH) on methanol-acetic acid-fixed nuclei was performed using the ZytoLight SPEC MYC/CEN8 Dual Color FISH probe (ZytoVision). The probes were applied to the slides, sealed with rubber cement and incubated in a humidified atmosphere (Thermobrite system) at 80°C for 10 minutes to allow denaturation of the probes and DNA target. The slides were then incubated overnight at 37°C to allow hybridisation. The rubber cement and coverslip were then removed and the slides were washed in 2X SSC/0.3% NP40 for 15 minutes at RT and then at 72°C for two minutes. After post-hybridisation washes, slides were counterstained with DAPI 1 µg/mL (Kreatech, Leica).

2.7 PDO and primary tissue DNA Isolation

The following procedure was performed by the Vincenzo Corbo Lab, Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy. Cells were incubated in Cell Recovery Medium (Corning) for 30 minutes at 4°C to remove Matrigel® and pelleted by 10,000g centrifugation at 4°C for 5 minutes. For tissue, sections of PDAC snap frozen tissue were scored by a pathologist for the percentage of neoplastic cellularity and only tissue with >20% neoplastic cellularity was used. For WGS and panel DNA sequencing, DNA was isolated using the DNeasy Blood & Tissue Kit (Qiagen).

2.8 DNA Panel Sequencing

The following procedure was performed by the Vincenzo Corbo Lab, Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy. The SureSelectXT HS Target Enrichment System (Agilent) was used for library preparation. Pair-end 2x150 sequencing of the panel was performed using NextSeq 550 (Illumina). Genes present in the panel are listed in Extended Data Table 2.

2.9 Whole genome sequencing (WGS)

The following procedure was performed by the Vincenzo Corbo Lab, Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy. DNA quality was evaluated using the DNF-467 50 kb DNA Kit on a Bioanalyzer 2100 (Agilent). The library was prepared and sequenced using the NovaSeq 6000 S4 Reagent kit v1.5 (300 cycles) at 15x coverage, generating around 160 million reads per sample.

2.10 Circle-seq

In this thesis, two Circle-seq runs were performed with modifications in the original Circle-seq protocol. The protocol was originally developed by Henssen et al. (2019a) and modified to fit our samples and laboratory equipment. After the first Circle-seq run (Volume 1) the protocol was updated based on the Volume 1 results and recommendations described in Møller (2020). Modifications to the original protocol by Henssen et al. (2019a) are noted in the following methods description for both Circle-seq runs.

The samples, including their passage number and their cell count which were used for Circle-Seq are detailed in Table 2.2.

Tab. 2.2 | PDCL and PDO samples used for Circle-seq and their unique identifier (ID).

Sample ID	Unique ID	Passage	Cell count [M]
Volume 1: PDAC PDCLs			
TKCC-2.1	TKCC-2.1	45	1.30
TKCC-09	TKCC-09	40	6.30
TKCC-10	TKCC-10	25	3.37
TKCC-15-LO	TKCC-15	36	7.50
TKCC-17-LO	TKCC-17	30	0.79
TKCC-18	TKCC-18	28	3.28
TKCC-22	TKCC-22	34	8.85
TKCC-26	TKCC-26	47	3.74
TKCC-27-LO	TKCC-27	37	0.62
PaCaDD137	PaCaDD137	31	3.02
Volume 2: 2 consecutive passages of 7 PDAC PDCLs			
Mayo-4636	Mayo-4636_P1	27	3.50
Mayo-4636	Mayo-4636_P2	28	2.00
PaCaDD137	PaCaDD137_P1	32	4.80
PaCaDD137	PaCaDD137_P2	33	9.00
TKCC-2.1-LO	TKCC-2.1_P1	36	3.20
TKCC-2.1-LO	TKCC-2.1_P2	37	3.00

Continued on next page

Sample ID	Unique ID	Passage	Cell count
TKCC-09	TKCC-09_P1	46	2.40
TKCC-09	TKCC-09_P2	47	1.68
TKCC-10-LO	TKCC-10_P1	28	3.80
TKCC-10-LO	TKCC-10_P2	29	3.00
TKCC-15-LO	TKCC-15_P1	42	6.00
TKCC-15-LO	TKCC-15_P2	43	6.00
TKCC-22-LO	TKCC-22_P1	35	4.80
TKCC-22-LO	TKCC-22_P2	36	5.20
Volume 2: PDAC PDOs			
HCM-CSHL-0080-C25-O	VR01-O	29	-
HCM-CSHL-0182-C25-O	VR30-O	11	-
HCM-CSHL-0600-C25-O	VR19-O	16	-
HCM-CSHL-0077-C25-O	VR02-O	24	-
HCM-CSHL-0084-C25-O	VR06-Oa	41	-
HCM-CSHL-0084-C25-O	VR06-Ob	45	-
HCM-CSHL-0089-C25-O	VR23-O	41	-
Volume 2: PDAC PDOs - Gemcitabine Treatment			
HCM-CSHL-0080-C25-O-GEM	VR01-O-GEM	26	-

2.10.1 HMW DNA extraction

The high molecular weight (HMW) DNA of each Circle-seq sample cell pellet, stored at -80°C, was isolated using the MagAttract HMW DNA (Qiagen). The HMW DNA extraction was performed following the Henssen et al. (2019a) protocol. The HMW DNA extraction from patient-derived organoids (PDOs), and PDO-derived cell lines was performed accordingly and provided by the Corbo Lab, Department of Diagnostics and Public Health, University of Verona, Italy.

2.10.2 Volume 1 - Linear DNA removal

To enrich the HMW DNA samples for eccDNAs, the removal of linear DNA was achieved by digesting it with an ATP-dependent plasmid-safe DNase (Lucigen) following the Henssen et al. (2019a) protocol. The DNase digestion was conducted daily for five consecutive days, with 20 Units of DNase, 4 µL of a 25 mM ATP solution, and reaction buffer added each day. After five days, the DNase was heat-inactivated at 70°C for 30 minutes. The effectiveness of linear DNA removal and the retention of circular DNA were assessed using quantitative PCR (qPCR) with primers targeting a chromosomal (linear) gene (*HBB* or *COX5B*, linear DNA control) and a mitochondrial gene (*MT-CO1*, circular DNA control). A fold change decrease of linear DNA content by at least 200-fold compared to circular DNA was considered as sufficient linear DNA removal. If the fold change decrease was below 200, the DNA was

subjected to an additional three days of treatment following the same protocol. The linear DNA fold change decrease was calculated as followed:

$$\Delta CT_{HBB} = CT_{HBB}(\text{post-DNase DNA}) - CT_{HBB}(\text{pre-DNase DNA})$$
$$\Delta CT_{MT\text{-}CO1} = CT_{MT-CO1}(\text{post-DNase DNA}) - CT_{MT\text{-}CO1}(\text{pre-DNase DNA})$$
$$\Delta\Delta CT = \Delta CT_{HBB} - \Delta CT_{MT\text{-}CO1}$$
$$FC = 2^{\Delta\Delta CT}$$

2.10.3 Volume 2 - Linear DNA removal

Similarly to the procedure described in Methods Section 2.10.2, HMW DNA was subjected to linear DNA removal by using the plasmid-safe DNase (Lucigen). The samples were treated for a minimum of seven days with 20 and 30 Units of DNase added daily. ATP solution and reaction buffer were added based on the recommended enzyme units provided by the manufacturer. After seven days, the DNase was heat-inactivated, and the effectiveness of linear DNA removal was assessed as described in Methods Section 2.10.2. If the fold change decrease of linear DNA content was below 200, an additional two-day treatment was performed by adding DNase, ATP, and reaction buffer. During this run, the final DNA enriched for circular DNAs was further concentrated by reducing the initial volume three to four fold using the using Savant™ DNA SpeedVac® DNA120 (Thermo Scientific).

2.10.4 Rolling-circle amplification

Rolling-circle amplification of the remaining DNA, enriched for circular DNA, was carried out using a Phi29 polymerase and the Repli-G Mini Kit (Qiagen), following the protocol described by Henssen et al. (2019a). The DNA concentration was measured using the Qubit® 2.0 Fluorometer (Invitrogen™) and the Qubit™ dsDNA BR Assay Kit (Invitrogen™). Upon successful DNA amplification, the amplified DNA was purified using Agencourt AMPure XP Beads (Beckman Coulter).

2.10.5 Sequencing

Approximately 500 to 550 ng of circular DNA-enriched DNA were sheared to a mean length of around 450 bp using a M220 Focused-ultrasonicator (Covaris). Library preparation was performed using the NEBNext® Ultra II DNA Library Prep Kit for illumina®, which involved sequencing adapter addition and amplification. DNA Clean-up was conducted using the Agencourt AMPure XP Beads (Beckamn Coulter). All prepared libraries were sequenced on the illumina® NextSeq500 platform using either the NextSeq 500/550 Mid Output Kit v2.5 (300 Cycles) (Volume 1) or the NextSeq 500/550 High Output Kit v2.5 (300 Cycles) (Volume 2), generating approximately 10-15 million paired-end 150 bp reads per sample.

2.11 Data processing and quality control

Raw Circle-seq data was processed using Illumina® 'bcl2fastq' to generate paired FASTQ files for each sample. FASTQ quality control was performed using FastQC (Andrews et al., 2010). Adapter and low-quality bases were trimmed using 'Trim Galore' (Krueger, 2015), with default values, which utilises the functionality of 'cutadapt' (Martin, 2011). After trimming, complete adapter and low quality base removal was verified using 'FastQC'.

2.12 Sequence alignment and duplicate removal

Trimmed sequencing data was mapped to the GRCh38 reference genome using 'BWA' (Li, 2013). Duplicate reads were marked and removed using 'Picard MarkDuplicates' (http://broadinstitute.github.io/picard/).

2.13 EccDNA calling

'Circle-Map Readextractor' (Prada-Luengo et al., 2019) was used to prepare mapped reads for the identification of putative eccDNA junctions. Putative eccDNA junctions were called using 'Circle-Map Realign' with default values.

2.14 EccDNA filtering

Putative eccDNA junctions were filtered based on several criteria to retain only high-quality eccDNAs. A high-quality eccDNA had a circle-score (defined by 'Circle-Map') above 200, at least 5 split reads covering the eccDNA junction, and an overall coverage of the full eccDNA region of at least 80%. EccDNAs overlapping with a blacklist region defined by 'ENCODE' were also removed (Amemiya, Kundaje & Boyle, 2019).

2.15 Visual inspection of eccDNA calls

High-quality eccDNAs and sequencing background were visually inspected using IGV (Integrative Genomics Viewer) (Thorvaldsdóttir, Robinson & Mesirov, 2013) to ensure accurate eccDNA calling and sequencing quality.

2.16 Read pileup

BigWig files were generated and normalised using 'deeptools bamCoverage' (Ramírez et al., 2014). Normalisation to reads per kb per million mapped reads (RPKM) was performed. Region-specific pileups and base information was generated using 'Bigly' (Pedersen, 2022a).

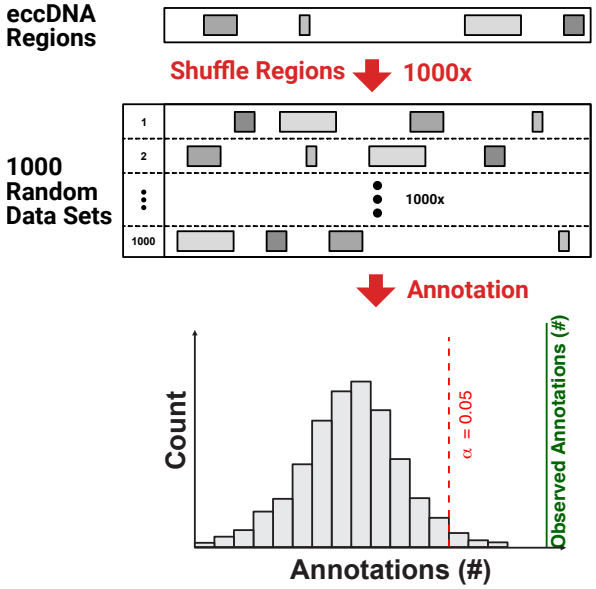


Fig. 2.1 | Schematic representation of the permutation test. EccDNA regions are shuffled 1,000 times and annotated with individual genomic elements. The distribution of the number of annotations of the random data set is then compared to the observed number of annotations in the eccDNA region. A P value below $\alpha < 0.05$ is considered significant.

2.17 Generating random eccDNA data sets

To perform a permutation analysis and identify eccDNA hotspots and coldspots, 1,000 random data sets were generated for each Circle-seq cohort. In detail, the high-quality eccDNA regions identified by 'Circle-Map Realign' were randomly permuted alongside their original GRCh38 chromosome, excluding regions overlapping with blacklist regions defined by 'ENCODE' (Amemiya, Kundaje & Boyle, 2019). This was performed using 'bedtools shuffle' (Quinlan & Hall, 2010) and the GRCh38 genome ranges, excluding regions overlapping the 'ENCODE' GRCh38 blacklist.

2.18 Permutation test

A permutation test was performed to identify annotations that are significantly enriched or lacking on eccDNA. The random regions in the 1,000 random data sets and the true eccDNA regions were annotated with various genomic elements (Methods Section 2.19). P values were determined based on the number of random data sets with more extreme number of annotations compared to the observed number of annotations.

2.19 EccDNA annotation

EccDNAs were annotated using all GRCh38 annotations provided by the R-package 'annotatr' (Cavalcante & Sartor, 2017). This includes annotations of enhancers defined by 'FANTOM' (https://fantom.gsc.riken.jp/), CpG islands, gene elements, different gene types, and repeat elements. Furthermore, eccDNAs were also annotated with GRCh38 genes and the repeat elements defined by 'RepeatMasker' (https://www.repeatmasker.org/).

2.20 EccDNA hotspot analysis

Recurrent eccDNA hotspots, coldspots, or regions of 'normal' eccDNA numbers are identified by shuffling the eccDNAs, identified in the three Circle-seq datasets, 1,000 times randomly along their original chromosome disregarding ENCODE blacklist regions (Amemiya, Kundaje & Boyle, 2019). Permutation of the eccDNAs is described in Methods Section 2.17. Subsequently, the genome is divided into around 1 Mbp bins, with the true and random eccDNA numbers counted for each Circle-seq dataset. If the number of eccDNAs in a bin exceeds the number in all 1,000 randomly generated datasets, the bin is classified as an eccDNA 'hotspot'. On the other hand, if fewer eccDNAs are found than in all randomly generated datasets, the bin would be identified as an eccDNA 'coldspot'. If neither of these two conditions is met, the bin is considered a region with a normal level of eccDNA occurrence. Recurrent eccDNA hotspots, coldspots, and normal regions are identified as those that appear in at least two of the three Circle-seq datasets outlined in Table 2.2. On the other hand, universal hotspots are hotspots that are identified in each Circle-seq data set. Regions falling in chromosome X and Y were removed prior downstream analysis.

To determine associations with specific genomic or transcriptomic features, all bins were annotated with genomic elements (Methods Section 2.19), RNA-seq, ATAC-seq, and methylation data. The number of each genomic element per bin was counted and compared between bin classes. To simplify visualisation, the average counts per bin were Z-score normalised. The median shift, for each genomic element, is calculated by subtracting the median normalised counts of the hotspot or coldspot regions with the median normalised counts of the normal region.

2.21 Overrepresentation analysis

Overrepresentation analysis of 'KEGG' pathways (Kanehisa & Goto, 2000), provided by 'MSigDB' (Liberzon et al., 2011), was performed using the R-package 'clusterProfiler' (Yu, 2022). Gene sets with a Benjamini-Hochberg P adjusted value less than 0.1 were considered significant.

2.22 EccDNA *de novo* assembly

Multi-fragment eccDNAs were *de novo* assembled (contigs) using 'Unicycler' (Wick et al., 2017) and mapped to the GRCh38 reference genome using 'Minimap2' (Li, 2018). To retain high-quality eccDNAs, sole eccDNA fragments that mapped to multiple regions were removed and the minimum mapping quality was set to 60. To account for small deletions, insertions, or mismatches within the eccDNA fragments (size < 50 bp), mapped contig fragments that were less than 50 bp apart were merged. This step helps reconstruct and consolidate fragmented eccDNAs, improving their representation and accuracy. The length of the mapped contigs

was compared to the total contig length. Contigs with a deviation of more than 10% of the total contig length were removed. This step ensures that the majority of the eccDNA sequence aligns to the reference genome, reducing the inclusion of partially mapped or misaligned fragments.

2.23 PDCL copy number and expression data

The PDCL copy number and expression data was obtained from Brunton et al. (2020).

2.24 ATAC-seq data analysis

The raw PDCL ATAC-seq data used in the thesis was published and provided by Brunton et al. (2020). To process the ATAC-seq data, the pipeline nf-core/atacseq (https://nf-co.re/atacseq) was utilised. The pipeline offers a standardised and comprehensive analysis workflow for ATAC-seq data.

After processing, broad peaks were called using 'MACS2' (Gaspar, 2018) and normalised using 'DESeq2' (Love, Huber & Anders, 2014).

2.25 ICGC PDAC data

Amplicon information from the 'PACA-CA' and 'PACA-AU' projects of the International Cancer Genome Consortium (ICGC) was obtained from Kim et al. (2020). This included a total of 142 samples, of which 81 were sequenced by PACA-CA and 61 by PACA-AU. Additional matching clinical, copy number, mutational, and transcriptomic data were retrieved from the ICGC database (release 28, https://dcc.icgc.org/). To focus specifically on PDAC, only PDAC tumours with histological types '8500/3', '8560/3', '8140/3', 'Adenosquamous carcinoma' and 'Pancreatic Ductal Adenocarcinoma' were used in the downstream analysis. Furthermore, chromothripsis data were also obtained from Cortés-Ciriano et al. (2020). For additional comparison with another form of pancreatic tumours, amplicon information from the pancreatic endocrine neoplasm projects PAEN-AU (n = 38) and PAEN-IT (n = 33) was also extracted from Kim et al. (2020).

2.26 WGS data pre-processing and alignment

The WGS data were pre-processed and aligned to the GRCh38 reference genome using the 'nf-core/sarek' pipeline (Garcia et al., 2020). This pipeline incorporates various tools and steps for processing and analysing WGS data. 'Fastp' (Chen et al., 2018) was used to remove low-quality bases and adapters from the raw reads. 'BWA Mem' (Li, 2013) was employed to map the trimmed trimmed reads to the GRCh38 reference genome, provided by the Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/grc). Subsequently,

'Picard MarkDuplicates' was utilised to mark duplicate reads, and 'GATK BaseRecalibartor' and 'GATK ApplyBQSR' were employed to recalibrate the base quality scores of the reads (McKenna et al., 2010).

2.27 Amplicon characterisation

The nf-core/circdna (version 1.0.1, https://nf-co.re/circdna) pipeline branch 'AmpliconArchitect' was used to define amplicon classes in each WGS sample. nf-core/circdna calls copy number using 'cnvkit' (Talevich et al., 2016) and identified amplified seeds with a copy number greater than 4.5 for 'AmpliconArchitect' by utilising functionality of the 'AmpliconSuite-Pipeline' (https://github.com/jluebeck/AmpliconSuite-pipeline). 'AmpliconArchitect' was run on the aligned reads and the amplified seeds to delineate the amplicon structures. The identified amplicons were further classified into 'circular' (ecDNA), 'linear', 'complex', or 'BFB' (amplicon with a breakage-fusion-bridge signature) using 'AmpliconClassifier' (Luebeck et al., 2023). Both software tools, 'AmpliconArchitect' and 'AmpliconClassifier', utilise reference genome data which needs to match the reference genome version used in the alignment step. Therefore, the newly generated WGS data was analysed with the GRCh38, and the Cancer Cell Line Encylopedia (CCLE) WGS data (Methods Section 2.32) with the GRCh37 reference genome data.

2.28 Sample classification

Samples containing at least one circular amplicon (ecDNA) were classified as 'ecDNA-positive'/ecDNA+', whereas samples without ecDNA amplicons were classified as 'ecDNA-negative'/'ecDNA–'. Based on the types of amplicons they contained, samples were further classified into 'Circular', 'Linear', 'Complex', 'BFB', or 'no-fSCNA' (no-focal somatic copy number amplification detected) (Kim et al. (2020) for more information). Samples with multiple amplicons were classified based on the amplicon with the highest priority, following the order of Circular > BFB > Complex > Linear.

2.29 EcDNA analysis

Putative ecDNA cycle plots were generated using the cycle information obtained from 'AmpliconArchitect'. The cycles were annotated using the GRCh38 gene annotation and plotted using the 'circlize' R-package.

2.30 RNA-seq of HCMI PDOs

RNA-seq data was generated for 14 pancreatic cancer (PC) PDOs that had matching WGS data. The sequencing and the count matrix generation was performed by the lab of Corbo Lab,

Department of Diagnostics and Public Health, University of Verona, Italy. Normalisation was performed using the 'DESeq2' function 'rlog' (Love, Anders & Huber, 2021).

2.31 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed using the 'fgsea' R-package. The analysis utilised the Hallmark pathways database (Liberzon et al., 2015) provided by the 'msigdbr' R-package (Dolgalev, 2022). Additionally, PDAC subtype gene signatures defined in previous studies by Moffitt et al. (2015), Bailey et al. (2016) and Raghavan et al. (2021), and Chan-Seng-Yue et al. (2020) were included in the analysis.

2.32 Public datasets

In addition to the ICGC PACA-CA and PACA-AU data sets, several public data sets were acquired for the study:

Methylation profiles from 24 PDAC tissue grown as patient-derived tumour xenografts (PDTXs) was obtained from EMBL-EBI Biostudies with the accession number E-MTAB-5571 (Lomberk et al., 2018). The profiles were originally mapped to the GRCh37 reference genome. To integrate them with GRCh38 aligned data, the genomic locations were converted to GRCh38 using the 'liftOver' tool in the 'rtracklayer' R-package (Lawrence, Carey & Gentleman, 2021). The hg19 (GRCh37) to hg38 (GRCh38) conversion chain was downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/).

RNA-seq count data of 44 PDAC PDOs was obtained from the National Cancer Institute's (NCI's) Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/) (Tiriac et al., 2018). The dataset is available under the project identifier 'ORGANOID-PANCREATIC'. The raw counts were log-normalised using the 'DESeq2' 'rlog' function.

The normalised RNA-sec counts of 150 PDAC samples from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) PDAC project 'TCGA-PAAD' were obtained from the GDC Data Portal. The dataset is accessible under the project name 'TCGA-PAAD'. Prior to downstream analysis, the normalised counts were log-transformed.

Ten WGS data sets from PDAC cell lines of the Cancer Cell Line Encyclopedia (CCLE) project were downloaded from the NCBI Sequence Read Archive (SRA) under the project ID SRP186687.

2.33 Restricted Datasets

Mutational calls for all PDOs are provided by the HCMI consortium. This dataset is embargoed until official publication. Furthermore, identified gene mutations for certain PDOs can also

be found on the HCMI Searchable Catalog (https://hcmi-searchable-catalog.nci.nih.gov/).

2.34 Statistical analysis

All statistical analyses were performed using R (v4.1.2). Various statistical tests were employed depending on the specific analysis requirements. A Fisher's exact test or a chi-squared test was used to evaluate independence between two variables. The Wilcoxon rank-sum test, also known as the Mann Whitney U test, was utilised in two-group comparisons. The relationship between quantitative variables was measured using the Pearson correlation coefficient. Other statistical tests were conducted as described in the figures or figure captions, depending on the specific analysis and hypothesis being tested.

2.35 Survival analysis

The Kaplan-Meier survival analysis was performed using the R-package 'survival' and the results were visualised using the R-package 'survminer'(Therneau & Lumley, 2015; Kassambara et al., 2017).

2.36 Molecular biology methods

2.36.1 Primer design for eccDNA candidate validation

Outward-directed primers for validating candidate eccDNA junctions were designed using Benchling's primer wizard (Benchling, 2023). The primer wizard utilises 'Primer3' (Untergasser et al., 2012) to identify the optimal primer pair. Primers were selected based on the lowest penalty score generated by 'Primer3', aiming for a PCR fragment size of 200-800 bp, a GC content of 40-60%, a primer length of 20-25 bp, and a primer melting temperature of 58-62°C. The primer sequences are provided in Methods Table 2.6.

2.36.2 Inverse PCR targeting candidate eccDNA junctions

Inverse PCR was performed using outward-directed primers and Invitrogen™ Platinum™ Pfx DNA Polymerase (Thermo Fisher Scientific) following the manufacturer's instructions. The reaction conditions included an initial denaturation at 94°C for 10 minutes, followed by 35 cycles of 15 s at 95°C, 30 s at 58°C, and 60 s at 68°C, followed by 15 s at 94°C and 15 s at 68°C. The length length of the PCR products was verified by conventional agarose gel electrophoresis using 1-2% agarose in TAE buffer. Visualisation was achieved by adding the recommended amount of SYBR™ Safe DNA Gel Stain (Invitrogen), and a 100 bp DNA Ladder (Invitrogen) was included as a size marker. Gel electrophoresis was conducted at 80-150 V until the dye line reached approximately 80% of the total gel length. Gel images were acquired using a ChemiDoc™ Imaging System (Bio-Rad), and 'Fiji' with 'ImageJ2'

(version 2.9.0/1.53t) (Abramoff, Magalhães & Ram, 2004) was used to invert image colors, adjust brightness, and contrast.

2.36.3 Sequencing of candidate eccDNA junctions

Sequencing was performed to verify the sequences of the amplified PCR products. Prior sequencing, 10 µL of the PCR products were purified using a Zymoclean Gel DNA recovery Kit (Zymo Research Europe GmbH). The purified PCR products were sequenced by the Beatson Institute for Cancer Research Molecular Technology Service on an Applied Biosystems 3130xl (16 capillary) sequencer (Fisher Scientific). The resulting sequences were aligned to the reference sequence using the 'MUSCLE' aligner (Edgar, 2004) implemented in the 'Unipro UGENE' software (Okonechnikov et al., 2012).

2.37 Graphic design and illustration

Affinity Designer (version 1.10.6) was used to consolidate individual figures, enhance their visual presentation, and generate schematic illustrations. Unless stated otherwise, all figures and illustrations were created by myself.

2.38 Extended Data

Extended data including tables, figures, scripts, and processed data are available on an Open Science Framework (OSF) repository. The access to the specific repository for this thesis needs to be requested and granted by the author of this thesis. A mail detailing the request needs to be sent to ds.danielschreyer@gmail.com.

2.39 List of software and algorithms

Tab. 2.3 | List of software and algorithms.

Software	Version	Source
R-Packages		
annotate	1.72.0	Gentleman (2021)
annotatr	1.20.0	Cavalcante (2021)
bedtoolsr	2.30.0-1	Patwardhan et al. (2021)
Biobase	2.54.0	Gentleman et al. (2021)
biomaRt	2.50.3	Durinck and Huber (2022)
Biostrings	2.62.0	Pagès et al. (2021)
ChIPseeker	1.30.3	Yu (2021)
chromoMap	4.1.1	Anand (2022)

Continued on next page

Software	Version	Source
circlize	0.4.15	Gu (2022)
clusterProfiler	4.2.2	Yu (2022)
colorblindr	0.1.0	McWhite and Wilke (2021)
colorspace	2.0-3	Ihaka et al. (2022)
ComplexHeatmap	2.10.0	Gu (2021)
DESeq2	1.34.0	Love, Anders and Huber (2021)
DiffBind	3.4.11	Stark and Brown (2022)
dplyr	1.0.10	Wickham et al. (2022a)
fgsea	1.20.0	Korotkevich, Sukhov and Sergushichev (2021)
forcats	0.5.2	Wickham (2022a)
GenomicDistributions	1.2.0	Kupkova et al. (2021)
GenomicFeatures	1.46.5	Carlson et al. (2022)
GenomicRanges	1.46.1	Aboyoun, Pagès and Lawrence (2021)
ggalluvial	0.12.3	Brunson and Read (2020)
ggbeeswarm	0.6.0	Clarke and Sherrill-Mix (2017)
ggbio	1.42.0	Yin, Lawrence and Cook (2021)
ggforce	0.4.1	Pedersen (2022b)
gghighlight	0.3.3	Yutani (2022)
ggplot2	3.3.6	Wickham et al. (2022b)
ggpubr	0.4.0	Kassambara (2020)
ggrepel	0.9.1	Slowikowski (2021)
ggsci	2.9	Xiao (2018)
ggupset	0.3.0	Ahlmann-Eltze (2020)
GSEA	1.2	Subramanian, Tamayo and Castanza (2019)
GSVA	1.42.0	Guinney and Castelo (2021)
karyoploteR	1.20.3	Gel (2022)
liftOver	1.18.0	Bioconductor Package Maintainer (2021)
maftools	2.10.05	Mayakonda (2022)
patchwork	1.1.2	Pedersen (2022c)
plyranges	1.14.0	Lee, Lawrence and Cook (2021)
RColorBrewer	1.1-3	Neuwirth (2022)
Rsamtools	2.10.0	Morgan et al. (2021)

Continued on next page

Software	Version	Source
stringr	1.4.1	Wickham (2022b)
StructuralVariantAnnotation	1.10.1	Cameron and Dong (2021)
survival	3.4-0	Therneau (2022)
survminer	0.4.9	Kassambara, Kosinski and Biecek (2021)
tidyr	1.2.1	Wickham and Girlich (2022)
tidyverse	1.3.2	Wickham (2022c)
VariantAnnotation	1.40.0	Maintainer et al. (2021)
vcfR	1.13.0	Knaus and Grunwald (2022)
WGCNA	1.71	Langfelder et al. (2022)
Command Line Tools		
FastQC	0.11.9	Andrews et al. (2010)
MultiQC	1.12	Ewels et al. (2016)
Trim Galore	0.6.7	Krueger (2015)
Cutadapt	4.1	Martin (2011)
BWA	0.7.17-r1188	Li (2013)
Samtools	1.15.1	Li et al. (2009)
Picard	2.26.10	http://broadinstitute.github.io/picard/
Circle-Map	1.1.4	Prada-Luengo et al. (2019)
CIRCexplorer2	2.3.8	Zhang et al. (2016)
Circle	git commit: 3eb333d	Kumar et al. (2017)
Unicycler	0.5.0	Wick et al. (2017)
Seqtk	1.3	Li (2012)
Minimap2	2.21	Li (2018)
CNVkit	0.9.9	Talevich et al. (2016)
PrepareAA	0.1032.2	Kim et al. (2020)
AmpliconArchitect	1.3_r1	Deshpande et al. (2019)
AmpliconClassifier	0.4.11	Kim et al. (2020)
deepTools	3.5.1	Ramírez et al. (2014)
bcl2fastq	2.19.0.316	Illumina
bedtools	2.30.0	Quinlan and Hall (2010)
R	4.1.2	https://www.r-project.org/
nf-core/atacseq	1.2.0	Patel et al. (2020)
nf-core/sarek	3.0.2	Garcia et al. (2020)
Fastp	0.23.2	Chen et al. (2018)
GATK	4.2.6.1	McKenna et al. (2010)
MACS2	2.2.7.1	Zhang et al. (2008)

Continued on next page

Software	Version	Source
Stand-Alone Software		
QuantStudio Design & Analysis Software	2.6.0	Thermo Fisher Scientific
2100 Expert Software	B.02.09.SI725	Agilent Technologies
Affinity Designer	1.10.6	Serif
RStudio	2022.07.0	RStudio
Fiji with ImageJ2	2.9.0/1.53t	Abramoff, Magalhães and Ram (2004)
Unipro UGENE	4.4.0	Okonechnikov et al. (2012)

2.40 List of reagents

Tab. 2.4 | List of resources.

Name	Catalog Number	Source
Molecular Biology Reagents & Consumables		
microTUBE-15 AFA Beads Screw-Cap	520145	Covaris
Microamp Fast Optical 96 Well Reaction Plate, 0.1mL	ST0140	Thermo Scientific
MicroAmp™ Optical Adhesive Film	4360954	Applied Biosystems™
Invitrogen™ Platinum™ Pfx DNA Polymerase	10532693	Thermo Fisher Scientific
SYBR™ Safe DNA Gel Stain	S33102	Invitrogen™
Thermo Scientific DyNAmo HS SYBR Green qPCR Kit	F410L	Thermo Scientific
Karyomax™ Colcemid™ Solution in PBS	15212012	Thermo Scientific
100 bp DNA Ladder	15628-019	Invitrogen™
Microamp Fast Optical 96 Well Reaction Plate, 0.1mL	43-469-07	Fisher Scientific
Gel Loading Dye, Purple (6X)	B7024S	New England Biolabs
Agarose	15510-027	Invitrogen™
Agencourt AMPure XP beads	A63880	Beckman Coulter
Cell Culture Media, Supplements, and Reagents		
Phosphate Buffered Saline (PBS)	Sigma-Aldrich	P4417
M199	31150022	Thermo Fisher Scientific
Ham's F-12 Nutrient Mix	21765029	Thermo Fisher Scientific
HEPES, 1M Buffer Solution	15630049	Thermo Fisher Scientific
L-Glutamine (200 mM)	25030024	Thermo Fisher Scientific
apo-Transferrin human	T1147-500MG	Sigma-Aldrich
Human EGF Recombinant Protein	PHG0311L	Thermo Fisher Scientific
Hydrocortisone 21-hemisuccinate sodium salt	H4881	Sigma-Aldrich
Insulin, human recombinant, zinc solution	12585014	Gibco™
D-(+)-Glucose solution	G8644-100ML	Sigma-Aldrich
Fetal Bovine Serum, qualified, Brazil	10270106	Gibco™
3,3′,5-Triiodo-L-thyronine sodium salt	T6397-100MG	Sigma-Aldrich
MEM Vitamin Solution (100X)	11120037	Gibco™
O-phosphorylethanolamine	P0503	Sigma-Aldrich
Penicillin-Streptomycin (10,000 U/mL)	15140122	Gibco™
Gentamicin solution	G1272	Sigma-Aldrich
IMDM	21980032	Gibco™

Continued on next page

Name	Catalog Number	Source
DMEM/F-12	11320-074	Gibco™
Trypsin (2.5%), no phenol red	15090046	Gibco™
DMEM, high glucose	11965092	Gibco™
Keratinocyte SFM (1X)	17005042	Gibco™
Keratinocyte-SFM Medium (Kit) with L-glutamine, EGF, and BPE	17005075	Gibco™
Commercial Assays & Kits		
Agilent DNA 1000 Kit	5067-1504	Agilent Technologies
Qubit™ dsDNA BR Assay Kit, 500 assays	Q32853	Invitrogen™
NextSeq 500/550 Mid Output Kit v2.5 (300 Cycles)	20024905	Illumina
NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®	E7645S	New England Biolabs
NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1)	E7335S	New England Biolabs
PlasmidSafe™ ATP-Dependent DNase	E3110K	Cambio
REPLI-g Mini Kit (25)	150023	Qiagen
MagAttract HMW DNA Kit (48)	67563	Qiagen
Zymoclean Gel DNA Recovery Kit (un-capped)	D4001	Zymo Research Europe GmbH
Critical Instruments		
Savant™ DNA SpeedVac® DNA120	DNA120-115	Thermo Scientific
QuantStudio™ 3 Real-Time PCR System	A28567	Applied Biosystems™
ChemiDoc™ Imaging System	17001401	Bio-Rad
Applied Biosystems™ 3130xl/3100 Genetic Analyzer 16-Capillary Array	15771816	Fisher Scientific
NextSeq500	SY-415-1001	Illumina
2100 Bioanalyzer Instrument	G2939BA	Agilent
M220 Focused-ultrasonicator	500295	Covaris
Incubator BD 53	9010-0081	Binder
NanoDrop 2000c Spectrophotometer	ND-2000C	Thermo Scientific
Qubit® 2.0 Fluorometer	Q32866	Invitrogen

2.41 Media formulation

Tab. 2.5 | Media formulation.

Name	Volume	Catalog Number	Source
M199/F12 Media			
M199	215.5 mL	31150022	Thermo Fisher Scientific
Ham's F-12 Nutrient Mix	215.5 mL	21765029	Thermo Fisher Scientific
HEPES, 1M Buffer Solution	7.5 mL	15630049	Thermo Fisher Scientific
L-Glutamine (200 mM)	5 mL	25030024	Thermo Fisher Scientific
apo-Transferrin human	5 mL	T1147-500MG	Sigma-Aldrich
Human EGF Recombinant Protein	10 µL	PHG0311L	Thermo Fisher Scientific
Hydrocortisone 21-hemisuccinate sodium salt	5 µL	H4881	Sigma-Aldrich
Insulin, human recombinant, zinc solution	1 mL	12585014	Gibco™
D-(+)-Glucose solution	3 mL	G8644-100ML	Sigma-Aldrich
Fetal Bovine Serum, qualified, Brazil	37.5 mL	10270106	Gibco™
3,3′,5-Triiodo-L-thyronine sodium salt	2.5 µL	T6397-100MG	Sigma-Aldrich
MEM Vitamin Solution (100X)	5 mL	11120037	Gibco™
O-phosphorylethanolamine	50 µL	P0503	Sigma-Aldrich
Penicillin-Streptomycin (10,000 U/mL)	5 mL	15140122	Gibco™
Gentamicin solution	250 µL	G1272	Sigma-Aldrich
IMDMrich Media			
IMDM	389 mL	21980032	Gibco™
Fetal Bovine Serum, qualified, Brazil	100 mL	10270106	Gibco™
Penicillin-Streptomycin (10,000 U/mL)	5 mL	15140122	Gibco™
MEM Vitamin Solution (100X)	2.5 mL	11120037	Gibco™
Human EGF Recombinant Protein	10 µL	PHG0311L	Thermo Fisher Scientific
apo-Transferrin human	500 µL	T1147-500MG	Sigma-Aldrich
Insulin, human recombinant, zinc solution	1 mL	12585014	Gibco™
Gentamicin solution	250 µL	G1272	Sigma-Aldrich
Mayo Media			
DMEM/F-12	450 mL	11320-074	Gibco™
Fetal Bovine Serum, qualified, Brazil	50 mL	10270106	Gibco™
L-Glutamine (200 mM)	5 mL	25030024	Thermo Fisher Scientific

Continued on next page

Name	Volume	Catalog Number	Source
Penicillin-Streptomycin (10,000 U/mL)	5 mL	15140122	Gibco™
Gentamicin solution	250 µL	G1272	Sigma-Aldrich
Pacadd Media			
DMEM, high glucose	266 mL	11965092	Gibco™
Keratinocyte-SFM Medium (Kit) with L-glutamine, EGF, and BPE	-	17005075	Gibco™
Keratinocyte SFM (1X)	145 mL	17005042	Gibco™
Fetal Bovine Serum, qualified, Brazil	100 mL	10270106	Gibco™
Penicillin-Streptomycin (10,000 U/mL)	5 mL	15140122	Gibco™
Gentamicin solution	250 µL	G1272	Sigma-Aldrich

2.42 Primer sequences

Tab. 2.6 | Primer sequences.

Primer Name	Sequence 5' to 3'	Reference
Chromosomal DNA Removal Control		
MT-CO1_F	GCCCACTTCCACTATGTCCT	Møller (2020)
MT-CO1_R	GATTTTGGCGTAGGTTTGGTCT	Møller (2020)
COX5B_F	GGGGCACCATTTTCCTTGATCAT	Møller (2020)
COX5B_R	AGTCGCCTGCTCTTCATCAG	Møller (2020)
HBB_F	TATTGGTCTCCTTAAACCTGTCTTG	Henssen et al. (2019a)
HBB_R	CTGACACAACTGTGTTCACTAGC	Henssen et al. (2019a)
Circle-seq Validation: Single-Fragment EccDNA		
PaCaDD137_16_88-62_F	TGACCTTGCTGAGGCCCATCCA	
PaCaDD137_16_88-62_R	GCTGTTATTCTCCGCTGGCGCT	
TKCC-2.1_7_38-52_F	TCCCTGGGGCTCCCGAAAGAAA	
TKCC-2.1_7_38-52_R	TTGGGACGCCCTCTGTTGTTGC	
TKCC-10_9_11-09_F	AGCAGGGGCCATCTGATCCCAA	
TKCC-10_9_11-09_R	AAGCACTGACCCGCTGCTGTTC	
TKCC-15_7_97-98_F	ATTTCAGCCCTGCTCAGAGCCC	
TKCC-15_7_97-98_R	CGTGGTGTCAGCATGGTCTGGT	
TKCC-15_19_02-25_F	GGTCGGTTGGAAATCCCTGGCA	
TKCC-15_19_02-25_R	CGCTGGGTGCCCTTTCTTTCCA	
TKCC-18_5_21-80_F	ACATGTGGCATGCTGGTGTGCT	
TKCC-18_5_21-80_R	AAGAAATGGCACTGGGGGAGCG	
TKCC-22_6_99-19_F	TTAGAGGCCTTGGCCAGCACCT	
TKCC-22_6_99-19_R	ACCCAGGCTGGAGTGCAGTGAT	
TKCC-22_6_40-41_F	TGGCAGCCATTCCCCATTGTCC	
TKCC-22_6_40-41_R	GCAACAGCGCAACAAACTATGGCA	
TKCC-26_6_42-82_F	AACCTCCTCGGCCTCCCAAAGT	
TKCC-26_6_42-82_R	TTCTGCCATCCTGGGGGTTCCA	
Circle-seq Validation: Multi-Fragment EccDNA		
PaCaDD137_47_87_10_F	AGATTGCACGGCCAACCCACAA	
PaCaDD137_47_87_10_R	ACAGAGGGGAAGAGGGTGGCTT	
TKCC-2_1_12_13_20_F	CGTACATCCGGAAATGGAATCTGC	
TKCC-2_1_12_13_20_R	ACACTTGGGCTGCAGCACTGAG	
TKCC-2_1_55_54_15_F	TGGGGAAAGTGTGAGTGGTGCT	
TKCC-2_1_55_54_15_R	ACAGAGGCCAATGTGTCAGCCC	
TKCC-09_13_41_18_F	AAGGGGCCAGGCTCCCTCTTTT	

Continued on next page

Primer Name	Sequence 5’ to 3’	Reference
TKCC-09_13_41_18_R	AGTTGGCTGGTCAGGGGTTGTG	
TKCC-10_73_3_58_F	TGTCTGCTCCATTCGGCTGTGA	
TKCC-10_73_3_58_R	CCCACTGCCATTTCCCCATTCCT	
TKCC-18_18_90_7_F	TCTTTGGGGCCTCAGGATGGCT	
TKCC-18_18_90_7_R	ACAGCTCTGTGTGCCTAGGCCA	
TKCC-18_40_2_99_12_F	GCTCCCACTGTAGCCTCTGGAACA	
TKCC-18_40_2_99_12_R	ACCTGTCAACCTCCCAGAGCCA	
TKCC-22_68_12_15_F	AGGGTCCGCGGCTTCATTCTTG	
TKCC-22_68_12_15_R	ACCTCTTGCCCACAAAGAGGCT	

Chapter 3

nf-core/circdna

A Nextflow pipeline for the detection of ecDNA and eccDNA in genomic data sets

Where is the 'any' key?

Homer Simpson

The introduction of novel sequencing technologies has revolutionised the field towards big data, allowing the processing of hundreds of samples simultaneously. While this transition offers unprecedented opportunities, it also revealed new challenges that need to be addressed. With the increase in throughput, computational power must be increased accordingly, to facilitate a fast processing time. While most data processing software already allow CPU parallelisation for single samples, running multiple samples in parallel is mostly not configured. Additionally, depending on the compute, different independent processes can be run in parallel, to accelerate processing time (Leipzig, 2017; Wratten, Wilm & Göke, 2021).

To tackle these and other challenges, different pipeline frameworks such as Nextflow (Di Tommaso et al., 2017), BigDataScript (Cingolani, Sladek & Blanchette, 2015), and Snakemake (Köster & Rahmann, 2012) were developed. Those function as a workflow management system and handle software containers, software dependencies, computational resources, and parallelisation. This creates the portability of pipelines to different computational systems and achieves high reproducibility (Di Tommaso et al., 2017; Köster & Rahmann, 2012; Ewels et al., 2020; Wratten, Wilm & Göke, 2021).

Open source projects created a vast library of processing and analysis pipelines. Also in the biological framework, open source pipelines allow users from different experience levels to process and analyse their own biological datasets (Wratten, Wilm & Göke, 2021). One

prominent open-source initiative that built and is building pipelines for various biological datasets is named nf-core (Ewels et al., 2020).

The nf-core pipelines are built with the Nextflow language (Di Tommaso et al., 2017) and an integral pipeline framework that defines code structure and documentation, to enhance readability, portability, reproducibility, and usability. Each nf-core pipeline must be open source, thoroughly documented, peer-reviewed, built with the nf-core pipeline template, and usable with the software management systems Docker, Singularity, and Conda (Ewels et al., 2020). Currently, more than 50 pipelines are officially released and part of nf-core (https://nf-co.re/). These include pipelines for the processing of diverse biological datasets such as ATAC-seq (nf-core/atacseq), RNA-seq (nf-core/rnaseq), genome/exome sequencing (nf-core/sarek), CHiP-seq (nf-core/chipseq). Most pipelines are under continuous development to remove bugs or change the pipeline composition depending on the gold-standard in the field.

While many biological datasets already have a respective nf-core pipeline, a pipeline to identify eccDNAs from sequencing data was missing. Furthermore, the eccDNA field was revolutionised in recent years with the release of novel algorithms and software. Therefore, I aimed to develop a nf-core pipeline that contains the most used software for the analysis of eccDNA from WGS, ATAC-seq, or Circle-seq data. The pipeline is aimed to bring all the potential of Nextflow and nf-core workflows to the eccDNA field making eccDNA research available for a broader community.

3.1 Pipeline structure

In recent years, several tools studying circular DNAs have been developed for the identification and analysis of eccDNA junctions (Kumar et al., 2017; Prada-Luengo et al., 2019; Zhang et al., 2016), amplified eccDNAs (ecDNAs) (Deshpande et al., 2019), or circular genomes (Wick et al., 2017). However, each of these tools require unique input files and have diverse software requirements. To address this issue, a pipeline was developed within the nf-core framework, which enables easy investigation of various eccDNA research avenues.

To further enhance accessibility of eccDNA research, nf-core/circdna was developed, a Nextflow pipeline written with the nf-core framework including tools for eccDNA research. This pipeline, officially released as version 1.0.0 on June 1 2022, currently operates on version 1.0.4 (released on June 5 2023) and runs on Nextflow version >= 22.10.1. It comprises five major branches, each incorporating distinct software for the identification of eccDNAs (Figure 3.1). Included are Circle-Map (Prada-Luengo et al., 2019), CIRCexplorer2 (Zhang et al., 2016), and Circle_finder (Kumar et al., 2017) for the detection of putative eccDNA junctions from single-fragment eccDNAs (Figure 1.5 top right), Unicycler (Wick et al., 2017) and Minimap2 (Li, 2018) for *de novo* assembly and mapping of multi-fragment eccDNAs (Figure 1.5 bottom right), and AmpliconArchitect (Deshpande et al., 2019) for the identification

of amplified eccDNAs (ecDNAs; Figure 1.5 left). The selection of these tools is rooted in the widespread acceptance and frequent use within the eccDNA or circular DNA research community.

Several other tools are included in the pipeline alongside the eccDNA detection tools, each serving a specific purpose. The adapter sequence and low quality bases are removed from the raw data by the trimming software 'Trim Galore' (Krueger, 2015). 'FastQC' carries out quality control of the FASTQ data, and 'BWA' indexes the reference genome and maps the sequence data (Andrews et al., 2010; Li, 2013). And lastly, 'Picard MarkDuplicates' marks duplicate sequences, which can be removed before eccDNA detection (http://broadinstitute.github.io/picard/). All tools are essential to ensure accurate input data that is usable by the eccDNA detection tools and furthermore check sequencing and processing quality.

With this pipeline structure, the only data the user needs to supply is a sequencing dataset and a reference genome.

3.2 Portability

Each software component within nf-core/circdna has unique dependencies, which can potentially interfere with other software. To manage this, Nextflow enables the use of conda (https://docs.conda.io/en/latest/) as well as other container environments like docker (https://www.docker.com/) and singularity (https://sylabs.io/singularity/). These platforms handle software dependencies and mitigate the issue of dependency conflicts. For each process inside the pipeline, a dedicated environment is created in which the respective software and its dependencies are installed. The creation of individual environments for each process does not only avoid dependency conflicts, but also ensures consistent performance across different computing systems, resulting in a highly portable pipeline.

3.3 Continuous integration testing of nf-core/circdna

Validation and robustness are two key features of developing reliable software pipelines. For nf-core/circdna, a key method of ensuring these features is the implementation of continuous integration (CI) tests, powered by GitHub Actions (https://github.com/features/actions). These tests are designed to automatically verify that any code changes made inside the pipeline do not negatively affect its performance or output.

In detail, nf-core/circdna employs two main CI tests. The first CI test verifies the functionality of the 'ampliconarchitect' branch, while the second examines functionality and the performance with a test data of rest of the branches: Circle-Map ('circle_map_realign' and 'circle_map_repeats'), 'unicycler', 'circle_finder', and 'circexplorer2'.

Generally, each CI test requires a modest test dataset containing sequencing data and its

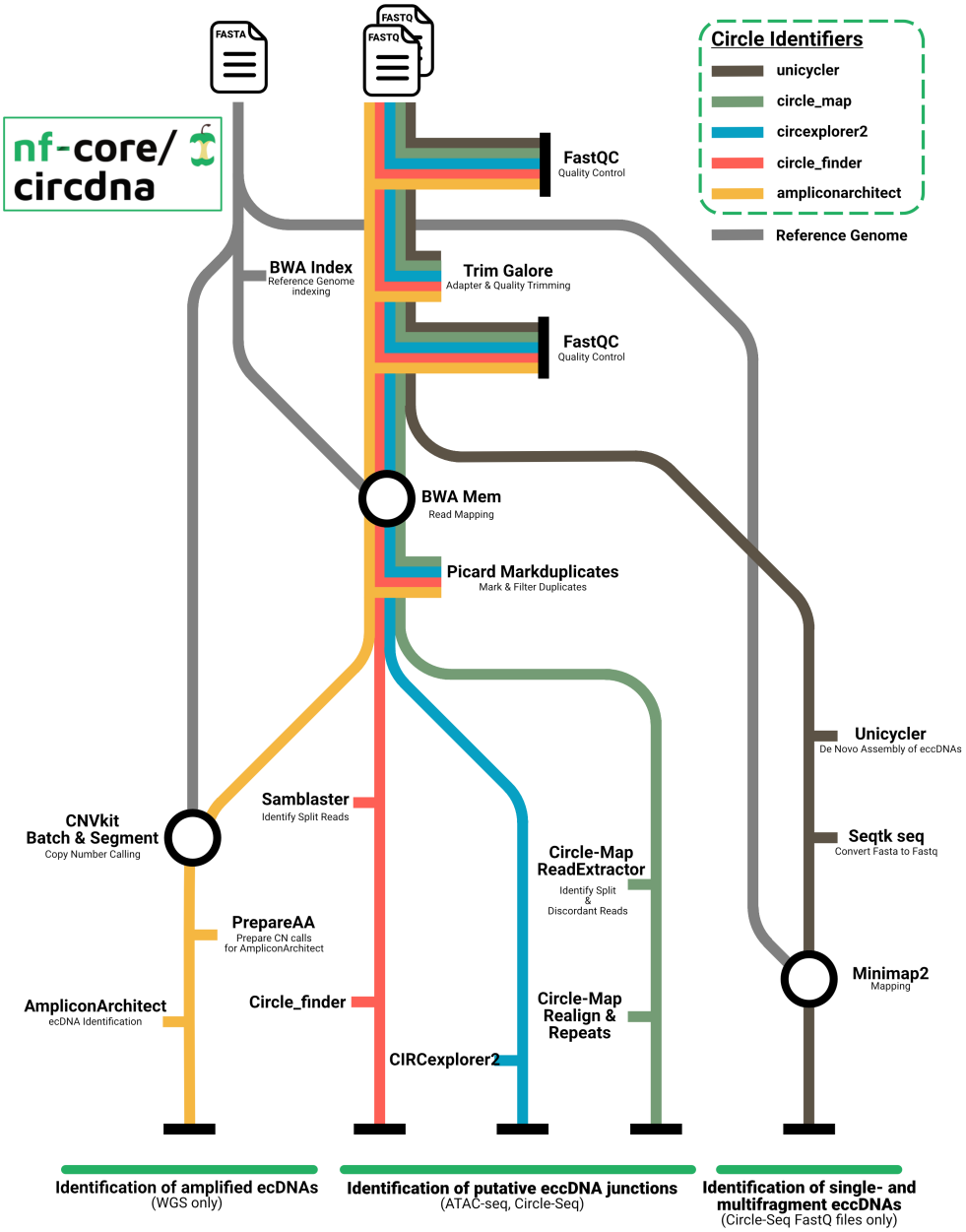


Fig. 3.1 | nf-core/circdna Branch Overview. The pipeline contains five branches, which are named after the eccDNA identifying tool. The pipeline uses paired-end sequencing data and needs a FASTA file from an appropriate reference genome. Different datasets are required for different branches: The 'ampliconarchitect' branch requires WGS data. 'circle_finder', 'circexplorer2', and 'circle_map' can be used with either ATAC-seq or Circle-seq data. And the 'unicycler' branch is recommended to be used with Circle-seq data only. Depending on the branch used, different outputs are generated. 'ampliconarchitect' identifies amplified ecDNAs, 'unicycler' de-novo aligns eccDNAs and can identify single-and multi-fragment eccDNAs, and 'circle_map', 'circexplorer2', and 'circle_finder' identify putative eccDNA junctions.

corresponding reference genome. However, the 'ampliconarchitect' branch requires additional resources, specifically prepared reference genome data files (github.com/virajbdeshpande/AmpliconArchitect) and a Mosek license (mosek.com). As these resources cannot be freely shared, they need to be individually acquired or downloaded. Therefore, the CI test for 'ampliconarchitect' uses the 'stub' method from Nextflow to check the correct installation and general functionality of all software within the branch. This method generates empty files and does not conduct a test dataset nor produce amplicon data, limiting its ability to identify bugs within the software.

On the other hand, the second key CI test uses simulated Circle-seq data to examine the other branches. This dataset was simulated with 'Circle-Map Simulate' (Prada-Luengo et al., 2019) using the parameters '-c 50 -N 400000 -r 150' and the Saccharomyces cerevisiae (R64-1-1) reference genome. This generated 200,000 simulated 150 bp paired-end FASTQ Circle-seq data with a mean coverage of 50 reads within the eccDNA coordinates. A high coverage is important for the 'unicycler' branch to fully detect an eccDNA sequence. The full test dataset is uploaded onto github (https://github.com/nf-core/test-datasets/tree/circdna) and will be downloaded when the CI tests are executed. With the use of this dataset and the execution of the CI test, all branches, but 'ampliconarchitect' can be fully validated.

In summary, CI tests form a vital aspect of open source projects, essentially contributing to keeping the integrity of the pipeline intact during continuous development by various developers. In total, integration of these tests verify functionality, reliability, and robustness of the nf-core/circdna, which ensures its successful application in eccDNA research.

3.4 Adapting the nf-core/circdna pipeline for user-specific needs

3.4.1 Input data

Incorporating various branches and software into the nf-core/circdna pipeline allows users to customise each workflow run based on the specifics of the input dataset and project requirements. For example, the 'ampliconarchitect' branch is designed for use with WGS data to detect amplified ecDNAs (Deshpande et al., 2019). 'circle_map', 'circexplorer2', and 'circle_finder' are optimally used with Circle-seq data, but may also work with ATAC-seq data as indicated by Kumar et al. (2020) (Prada-Luengo et al., 2019; Zhang et al., 2016; Kumar et al., 2020). However, the 'unicycler' branch should strictly be used with Circle-seq data to prevent false positives or incorrect assemblies from read originating from non-circular regions.

3.4.2 User-friendly, yet highly customisable: The parameters of nf-core/circdna

In an effort to allow users with varying programming skills to use nf-core/circdna, each run requires only three parameters, '`input`', '`input_format`', and '`output`' (Table 3.1). This will execute nf-core/circdna with default parameters on the 'circexplorer2' branch.

While nf-core/circdna can be launched with minimal parameters, it offers over 50 customisable parameters to fit users' individual requirements. A critical parameter to adjust prior to a workflow run is '`--circle_identifier`'. This parameter guides the pipeline on which branch and tool to execute, such as 'circle_map_realign' (uses the 'Circle-Map Realign' tool inside the 'circle_map' branch) or 'ampliconarchitect'. Multiple branches can be run

Tab. 3.1 | Minimal input requirements of nf-core/circdna

Parameter	Description
`--input`	comma-separated text file containing a row for each sample, listing the sample name, and the paired-end FASTQ or BAM file paths (samplesheet).
`--input_format`	Defines the file format of the input files. The pipeline accepts either paired-end FASTQ or BAM files.
`--output`	Path to the output directory in which all generated results will be stored.

simultaneously by specifying their identifiers in a comma-separated list. To illustrate the '`--circle_identifier`' can be set to 'circle_map_realign,circexplorer2' to analyse data through both the Circle-Map Realign and CIRCexplorer2 branches concurrently.

All pipeline parameters are detailed and explained on the nf-core website under https://nf-co.re/circdna/1.0.4/parameters. This resource is automatically updated with every new release, ensuring users always have access to the latest parameter documentation.

For users requiring in-depth customisation, nf-core/circdna allows for modification of each process using a user-defined configuration file. This file details the process and any additional parameters for the software within the process. This option further enables the user to modify each process, even without a without a predefined customisable parameter.

3.4.3 Data format and samplesheet styles

Depending on the available data, nf-core/circdna users can choose between 'FASTQ' or 'BAM' input formats. It's important to note that only paired-end sequencing data is supported by the pipeline. If users are working with FASTQ files, users need to prepare a 3-column comma-separated text file (samplesheet) that specifies the sample names along with the location of the paired-end FASTQ files. The necessary column names are 'sample', 'fastq_1', and 'fastq_2'. On the other hand, when BAM files are used as input, a 2-column samplesheet is required, comprising the columns 'sample', and 'bam'. Detailed explanations of these columns and their content for FASTQ and BAM input formats are detailed in Table 3.2.

Tab. 3.2 | Samplesheet file column description with `--input_format` 'FASTQ' and 'BAM'

Column	Description
`--input_format` 'FASTQ'	
sample	User-defined sample name.
fastq_1	Full path to Illumina short read FASTQ file 1. The specified file needs to be gzipped with the extension '.fastq.gz' or '.fq.gz'.
fastq_2	Full path to Illumina short read FASTQ file 2. The specified file needs to be gzipped with the extension '.fastq.gz' or '.fq.gz'.
`--input_format` 'BAM'	
sample	User-defined sample name.
bam	Full path to BAM file generated from Illumina paired-end short-read sequencing files.

3.5 Installation and usage

Ease of installation and usage are key to ensuring accessibility of software pipelines, especially for non-experienced users. Consequently, multiple methods for installing and executing the nf-core/circdna pipeline have been implemented.

One way to access nf-core/circdna is via GitHub, which hosts several branches each representing a specific state of the pipeline. Cloning the repository will download all necessary scripts and configuration files for the pipeline execution using Nextflow (https://seqera.io/). The pipeline is then executed using the following command 'nextflow run main.nf ...'.

Alternatively, the Nextflow or nf-core command line tool Ewels et al. (2020) can be used. These tools automatically download and run nf-core/circdna when specified using the following command line arguments: 'nextflow run nf-core/circdna -r 1.0.3 ...' or 'nf-core launch nf-core/circdna -r 1.0.3 ...'. By specifying the '-r' parameter, a specific version can be downloaded; otherwise the latest stable version will be downloaded by default.

Lastly, there's the Nextflow Tower (https://cloud.tower.nf/), a graphical user interface to launch and monitor Nextflow pipelines.

A complete guide to installation and usage commands is available on the nf-core/circdna website https://nf-co.re/circdna.

3.6 Output description and utility

This section discusses the output files generated by nf-core/circdna. Each pipeline run generates various output files depending on the branch, input format, and parameters specified. This can include the generation of reports which detail the quality of the input data and the analysis or the pipeline run. Furthermore, each branch generates individual intermediate and final outputs that can be used for further downstream analysis. All outputs are also reported on the nf-core/circdna website under https://nf-co.re/circdna/1.0.4/docs/output.

3.6.1 Pipeline reports and quality control

Next-generation sequencing data achieves a high accuracy in determining the read sequence. However, factors such as sample quality, library preparation technique, and human or technical errors can lead to a higher error rate and, consequently, to poor base accuracy (Koboldt et al., 2010). As a result, quality control checks and reports are essential for verifying data integrity and quality.

nf-core/circdna is designed to produce user-friendly outputs, including multiple quality control metrics. This ensures that pre-processing steps were performed correctly and the sequencing quality is sufficient for subsequent downstream analysis.

FastQC, a frequently used tool for high-throughput sequencing data, is employed for various quality parameters for FASTQ reads, including base sequencing quality, adapter content, GC content, and duplication levels (Andrews et al., 2010). FastQC is executed both before and after read trimming to verify the complete removal of low quality bases and adaptors.

The mapping process and the resultant alignment file undergo quality checking using Samtools stats (Danecek et al., 2021). This identifies and quantifies unmapped reads, mapping quality, duplicated reads, reads properly mapped, among other parameters.

Upon completion of each quality control process, MultiQC consolidates all raw quality control files and reports to compile a user-friendly, comprehensive report (Ewels et al., 2016).

Moreover, each pipeline run generates a report detailing the pipeline's execution including all its parameters, and the software versions used. This is particularly critical for ensuring reproducibility and transparency.

3.6.2 Intermediate Files

Software pipelines are designed to concatenate multiple tools in a linear fashion. In this structure, the output of one tool is utilised as the input for the succeeding tool. As the pipeline progresses, intermediate files are generated, which serve as inputs for the creation of the final output that contains relevant information for biological interpretation or subsequent downstream analysis. However, these intermediate files might also be valuable for further analyses not directly related to the pipeline's objective. Therefore, it might be beneficial to preserve these files.

nf-core/circdna is designed with user adaptability in mind and retains intermediate files as needed. This can include, trimmed FASTQ reads, BAM files, BAM files post-duplicate marking, or BAM files post-duplicate removal. Additionally, useful software's generated files are saved in their respective output folders. This facilitates the re-running of certain processes, quality control, and additional downstream analyses.

3.6.3 EcDNA/EccDNA Information

The integral part of every nf-core pipeline is to generate output for further investigations or analyses. For instance, the primary goal of nf-core/rnaseq (https://nf-co.re/rnaseq) is to generate a count matrix containing gene or transcript counts per sample. Conversely, nf-core/sarek is specialised in calling and annotating copy number alterations, SNVs, and SVs from genomic data (Garcia et al., 2020). And the newly developed nf-core/circdna aims to generate information about the eccDNA and ecDNA content within sequenced samples.

The five main branches of nf-core/circdna are categorised into three different functional groups:

1. AmpliconArchitect: This branch identifies amplified eccDNAs (ecDNAs)

2. Unicycler: This branch identifies single- and multi-fragment eccDNAs by eccDNA *de novo* assembly

3. Circle_finder, Circle-Map, and CIRCexplorer2: These branches identify eccDNA junctions

Each functional group should be reviewed and carefully selected based on the biological and technical data at hand. For instance, Circle-seq data is optimally utilised with eccDNA junction detection branches or the Unicycler branch, whereas WGS should only be employed with AmpliconArchitect.

Each eccDNA junction detection branch produces a final text output file, containing the chromosomal location of the identified putative eccDNA junctions. These files can be readily used with programming languages like R or Python for subsequent downstream analysis.

Unicycler, in contrast, generates a FASTA file containing the sequence of *de novo* assembled eccDNAs. To identify the eccDNA origin, the FASTA file is transformed into a FASTQ file and analysed with Minimap2, which maps the eccDNA sequences to a reference genome and generates a pairwise mapping format (PAF) file containing mapping information for each identified eccDNA. Again, the output of the Unicycler branch can be analysed using diverse programming languages.

Lastly, AmpliconArchitect produces cycles and graph files for each amplicon, which are later used to delineate the amplicon class by AmpliconClassifier. An amplicon that is classified as 'cyclic', exhibits an ecDNA signature, and has no detected BFB signature is finally classified as a circular amplicon, an ecDNA. More details about the method and the output can be found under https://nf-co.re/circdna/1.0.4/docs/output or in the papers Deshpande et al. (2019) and Luebeck et al. (2023) which describe the use and utility of AmpliconArchitect and AmpliconClassifier, respectively.

In essence, nf-core/circdna aims to produce readable and user-friendly outputs for various datasets investigated for their eccDNA information. These outputs are readily usable for downstream analysis or biological examination and can give insights into the eccDNA landscape of each sample.

3.7 Results

To test the performance of all nf-core/circdna branches, the Circle-seq dataset (Volume 1) of eight PDAC patient-derived cell lines (PDCLs), and two publicly available datasets were analysed. The branch AmpliconArchitect was tested by using WGS datasets of a total of five cell lines. The four other main branches, Circle_finder, Circle-Map, Unicycler, and

CIRCexplorer2 were tested with our own Circle-seq data generated from eight PDAC PDCLs.

3.7.1 AmpliconArchitect branch identifies ecDNAs

To verify the functionality of the AmpliconArchitect branch, WGS FASTQ files from five commonly used cell lines was analysed. These datasets were chosen based on their known presence or absence of ecDNAs. The datasets included low-pass WGS data of four cell lines, COLO205, GBM39, OVCAR8, and PC3, which were previously published and analysed by Turner et al. (2017) under the NCBI Sequence Read Archive (SRA) accession number 'PRJNA338012'. Additionally, the WGS data of COLO320DM, a commonly used cell line in ecDNA research, was acquired from SRA under the accession number 'PRJNA506071' analysed and published by Wu et al. (2019).

In this test, nf-core/circdna version 1.0.4 was used with the branch 'ampliconarchitect' and the five WGS datasets. The final output was generated by AmpliconClassifier summarising the classified amplicons, which were identified in each sample:

Tab. 3.3 | Classified amplicons identified by the 'ampliconarchitect' branch of five commonly used cell lines. The amplicons are classified into 'ecDNA', 'Linear', 'Complex', and 'BFB' based on their structure. All amplicons identified in a cell line are enumerated (N). Oncogenes identified on an amplicon are depicted.

N	Class	Location	Oncogenes
COLO205			
1	Linear	chr6:37851801-42907499	*CCND3*, *GLO1*, *TFEB*
2	Linear	chr6:51269024-56333928	
3	Linear	chr6:63307927-63957117, chr6:63958453-65312798	*PTP4A1*
4	Linear	chr6:65362739-67555799	
5	Linear	chr6:133087500-133947841, chr6:133948586-138114599	*AHI1*, *MYB*, *SGK1*, *TNFAIP3*
6	Linear	chr9:117204900-118360499	
7	Linear	chr12:124425303-124735303	
COLO320DM			
1	BFB	chr2:126394061-126917181, chr2:126919696-128339386	*ERCC3*
2	Complex	chr2:130184981-130197365, chr2:130199814-130459369, chr2:131372603-131372825, chr2:131488950-131489308	
3	ecDNA	chr8:126425747-127997818, chr8:129265936-129274477, chr6:371964-374266	*MYC*, *PVT1*
4	Linear	chr13:72303005-73503002	*DIS3*, *KLF5*
5	Linear	chr16:32287679-32359704	
GBM39			
1	ecDNA	chr7:54763279-55127269, chr7:55155020-56049370	*EGFR*

Continued on next page

N	Class	Location	Oncogenes
OVCAR8			
1	Linear	chr5:14125441-15870492, chr5:97786598-97807181	*TRIO*
2	Linear	chr8:109388988-109478987	
PC3			
1	Linear	chr1:171434357-173851884, chr1:199654128-200439123, chr1:200734119-201584114	
2	Linear	chr5:148663641-149388637	
3	ecDNA	chr8:111334757-111348272, chr8:118555462-118566017, chr8:119051678-119054057, chr8:121853389-121954292, chr8:122774690-122776268, chr8:130455524-130855582, chr8:132851686-132892345, chr8:133661572-133786290, chr8:137238950-137273896, chr8:138244840-138547408, chr8:139324742-139366500, chr8:139587901-139636511, chr8:141195337-141237974, chr8:141238340-141277958	
4	Linear	chr10:33014920-33599919, chr10:37849322-38049667	
5	Linear	chr10:36569911-37769910	
6	Linear	chr10:48855370-51315371	
7	Linear	chr10:61105383-63685385	
8	Linear	chr10:65305397-71065499, chr10:72748800-79521899, chr10:79799400-80227499	*PRF1*, *SIRT1*
9	Linear	chr12:27859522-28184520	
10	Linear	chr12:28554510-29584496	
11	Linear	chr19:58242603-58607612	*TRIM28*

As expected, COLO205 and OVCAR8, which have no known ecDNAs, contained only chromosomal amplicons classified as ’linear’ (Turner et al., 2017). In contrast, in COLO320 and GBM39, both of which have been reported to contain ecDNAs, the previously reported *MYC*-ecDNA and *EGFR*-ecDNA has been identified, respectively, alongside other chromosomal amplicons (Turner et al., 2017; Wu et al., 2019).

Interestingly, while an ecDNA has been detected for PC3, which has been reported to contain a *MYC*-ecDNA, the chromosomal fragment carrying the *MYC* gene could not been identified on this ecDNA. An ecDNA carrying regions from chromosome eight around the *MYC* locus has been discovered, but the *MYC* region was not included. This unexpected result could be due to the limitations of low-pass WGS in detecting ecDNAs or fully describing their structure. Indeed, the maintainers of AmpliconArchitect recommend using WGS data with a coverage of at least 5x, while Turner et al. (2017) used a median coverage of 1.19x. This may have not been sufficient for accurate ecDNA detection.

In conclusion, this analysis demonstrates that the ’ampliconarchitect’ branch, used within the nf-core/circdna pipeline version 1.0.4 with default parameters, is capable of identifying

and classifying amplicons from WGS data. The data processing is incorporated by utilising the tools AmpliconArchitect and AmpliconClassifier, which use copy number calls from cnvkit and delineate the amplicon structure (Deshpande et al., 2019; Luebeck et al., 2023; Talevich et al., 2016). Notably, this branch identified two ecDNAs that have been characterised in multiple previous studies (Turner et al., 2017; Wu et al., 2019; Hung et al., 2021). Additionally, the 'ampliconarchitect' branch can generate all necessary output files for downstream analysis, including detailed amplicon structures and the amplicon class (Table 3.3). Moreover, by default, the branch saves intermediate files generated during the process by default, providing a resource for further investigation A detailed information about the output can be found in the output documentation on the nf-core/circdna website (https://nf-co.re/circdna/1.0.4/docs/output/).

3.7.2 Single-fragment eccDNA identification

The three branches 'circle_finder', 'circexplorer2', and 'circle_map_realign' were tested using Circle-seq data of eight PDAC PDCLs, which are used for eccDNA junction identification. The nf-core/circdna pipeline was executed with default values, generating three distinct files, one for each branch, containing eccDNA information. After pipeline execution, the eccDNA information was further processed by removing eccDNAs, whose eccDNA junction was supported by fewer than five reads. The remaining eccDNAs that passed the filtering are counted and visualised for each branch.

Each branch successfully identified eccDNAs, albeit in varying quantities (Figure 3.2). Circle_finder detected approximately 46,500 eccDNAs, while 'CIRCexplorer2' and 'Circle-Map Realign' both identified roughly 75,000. This underlines that each branch's capacity to independently detect eccDNAs is dependent on the branch of choice, the biological query, and the software selected.

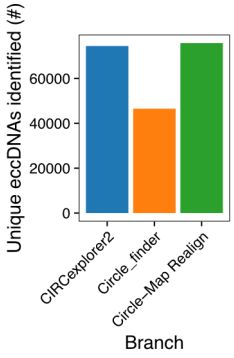


Fig. 3.2 | Number of eccDNAs identified in the three nf-core/circdna branches identifying putative eccDNA junctions.

The generated output can be subsequently analysed with programming languages such as 'R' or 'Python'. The role of nf-core/circdna is confined to process sequencing data to generate eccDNA information, but does not perform any secondary analyses.

In summary, this basic run demonstrates the capability of these three nf-core/circdna

branches in processing sequencing data and detecting eccDNAs.

3.7.3 Multi-fragment eccDNA identification

Single-fragment eccDNAs are simple eccDNAs generated from one specific fragment and locus of the chromosomes. However, the eccDNA landscape is much more complex and eccDNAs can comprise multiple different chromosomal fragments from one or multiple chromosomes (Kim et al., 2020; Deshpande et al., 2019). Identifying multi-fragment eccDNAs is key to identify the full eccDNA landscape. However, current methods are not adapted to identify eccDNAs from multiple chromosomal regions, but are optimised for identifying putative eccDNA junctions (Prada-Luengo et al., 2019; Kumar et al., 2020).

To discover eccDNAs from multiple chromosomes, the eccDNA-specific sequencing technique, Circle-seq, needs to be utilised as other sequencing methods are also generating reads from chromosomal regions, impeding the eccDNA identification process. With sequencing data enriched for eccDNA reads, *de novo* assembly algorithms can reconstruct the complete eccDNA sequence, which can then be aligned to a reference genome to determine the eccDNA origin.

Within nf-core/circdna this process is implemented using Unicycler (Wick et al., 2017) and Minimap2 (Li, 2018) included in the 'unicycler' branch. To test this experimental eccDNA identification procedure, the branch was used with the Circle-seq Volume 1 dataset. Then, the mapped *de novo* assembled eccDNA sequences were filtered based on several criteria to identify high-quality eccDNAs (Methods Section 2.22).

Running the Circle-seq dataset revealed more than 25,000 (26,582) *de novo* assembled eccDNAs. The majority of those were defined as originating from one single chromosomal fragment (22,371 of 26,582; 84.2%). However, also multi-fragment eccDNAs were identified with 12.6% having two fragments and 3.21% containing three or more fragments from different parts of the chromosome.

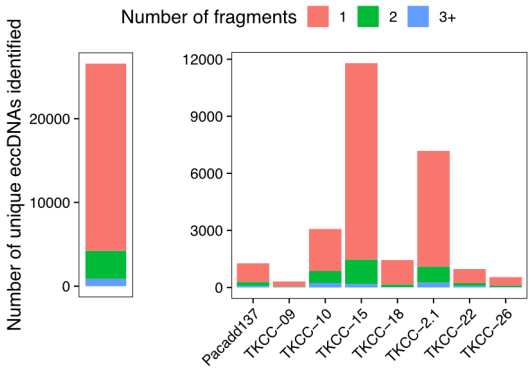


Fig. 3.3 | Number of unique single-or multi-fragment eccDNAs identified using the 'Unicycler' branch of nf-core/circdna. The total number (left) and the number of eccDNAs per sample (right) are divided into the number of fragments of each eccDNA. Three or more than three fragments were grouped into 3+.

In summary, the nf-core/circdna branch 'unicycler' can identify single- and multi-fragment

eccDNAs using the software Unicycler in combination with Minimap2. Further validation of this technique and branch is described in the Results Chapter 5.

3.8 Conclusion & discussion

The nf-core/circdna pipeline, utilising Nextflow and the nf-core framework, facilitates sequencing data analysis for eccDNA content. The pipeline is accessible from the nf-core website and GitHub, and supports multiple branches. Documentation for the pipeline is comprehensive, easing its application and modification to meet specific needs.

Leveraging Nextflow's domain-specific language has led to optimisation for ease of use, accuracy, speed, and adaptability (Ewels et al., 2020; Di Tommaso et al., 2017). While nf-core/circdna offers a comprehensive framework for detecting eccDNA, user expertise and input are still critical for optimal use. Additional processing and parameter modifications may be necessary to ensure accuracy, taking into consideration biological hypotheses and data. The pipeline provides a convenient method for researchers with diverse expertise to conduct eccDNA identification. At present, nf-core/circdna is optimised for analysing short read sequencing data. Nevertheless, with the growing usage of long read sequencing data for eccDNA research, relevant adaptations will need to be implemented in the pipeline framework in the future.

The nf-core guidelines impose stringent guidelines on the functionality of nf-core/circdna (Ewels et al., 2020). nf-core/circdna has been evaluated using WGS and Circle-seq data, and all software tools meet the necessary standards for effortless deployment across various computational systems.

The current version of nf-core/circdna (1.0.4) undergoes consistent refinement and maintenance. nf-core pipelines uphold elevated standards for reproducibility, improvement, and compliance with optimum procedures (Ewels et al., 2020). With nf-core/circdna, it is anticipated that research on eccDNA will be more widely accessible, facilitating advancements towards a thorough comprehension of patients' genomics. Furthermore, nf-core/circdna contributes to the extensive collection of analytical pipelines being developed for various types of biological data.

Chapter 4

Investigating ecDNAs in PDAC

> In biology, nothing is clear, everything is too complicated, everything is a mess, and just when you think you understand something, you peel off a layer and find deeper complications beneath. Nature is anything but simple.
>
> *Richard Preston*

The pancreatic ductal adenocarcinoma (PDAC) genomic landscape is intricate, marked by a limited number of highly prevalent driver gene mutations and an array of infrequent ones. In addition, copy number alterations, which lead to the deletion or amplification of cancer driver genes, play a significant role in PDAC. Copy number increases primarily affect oncogenes, leading to elevated activation of oncogene-specific pathways. In PDAC, many of the recurrently amplified oncogenes, including *KRAS*, *MYC*, *CDK6*, *MET*, or *GATA6*, control cell-cycle, subtype state, apoptosis, genomic instability, all of which can drive PDAC progression (Waddell et al., 2015; Chan-Seng-Yue et al., 2020; Maddipati et al., 2022; Lomberk et al., 2018).

Extrachromosomal circular DNAs (ecDNAs) are important players in the mechanism of oncogene amplification. Unlike normal chromosomes, ecDNAs lack a centromere, resulting in random inheritance during the cell cycle. This process can drive amplifications and intratumour heterogeneity, leading to tumour progression or drug resistances (Kim et al., 2020; Hung et al., 2021; Nathanson et al., 2014; Yi et al., 2022). While ecDNAs have been identified in almost all types of cancer, including pancreatic cancer (PC), further research on ecDNAs in PDAC is limited (Kim et al., 2020; Notta et al., 2016).

To address this gap of knowledge, the primary aim of this study is to characterise and describe the occurrence, associations, and potential implications of ecDNAs in PDAC. Through

expanding our comprehension of the complex genomics of this disease, these findings may prove beneficial for PDAC patients.

The study integrated PDAC amplicon information retrieved from Kim et al. (2020) with clinical, mutational, and transcriptomic data from the ICGC projects PACA-CA and PACA-AU. Additionally, WGS and RNA-seq data from PDAC patient-derived organoids (PDOs) were analysed. These PDOs were generated by the Vincenzo Corbo lab, University of Verona, in collaboration with the Human Cancer Model Initiative (HCMI). Post-characterisation, selected PDOs were subjected to media condition changes to investigate the involvement of ecDNAs in adaptation mechanisms, offering insights into the role of ecDNAs in adaptation mechanisms. And lastly, I also analysed WGS data from PDAC cell lines published by the Cancer Cell Line Encyclopedia (CCLE) to identify their use in ecDNA research.

4.1 Study samples

In order to investigate the specific characteristics of ecDNAs in PDAC, I obtained and analysed the amplicon data from the ICGC PC samples, as reported by Kim et al. (2020). The study methodically identified and categorised the amplicons for each WGS sample from the TCGA and ICGC. Samples containing circular amplicons (ecDNAs) were categorised as 'Circular' or 'ecDNA+'. However, it is worth noting that the study did not differentiate between the different types of PC (Kim et al., 2020).

ICGC PC projects classify samples based on the specific pancreas cancer of interest. For instance, 'PACA' projects primarily include PDAC tumours, while 'PAEN' projects contain pancreatic cancer endocrine neoplasms (PAEN) tumours. By dividing all PC tumour samples into their respective sub-projects, PACA and PAEN, and investigating the amplicon content, a slightly higher, but non-significant (Fisher's exact test, P value = 0.51), ecDNA+ frequency was identified in the in the tumours of the PACA project (13.4%) compared to the PAEN tumours (9.9%, Figure 4.1a). This finding underscores the prevalence in PDAC tumours and sets the basis for a more detailed examination of ecDNAs in PDAC.

The two ICGC PACA projects, PACA-CA and PACA-AU, comprise samples from various PC types with a focus on PDAC. To ascertain characteristics specific to PDAC, histologically identifiable non-PDAC samples were excluded prior further analysis (Figure 4.1b and Methods Section 2.25). Out of the initial 142 PC tumours in the PACA projects, 127 were classified as PDAC and employed for further study. These 127 PDAC samples primarily originated from early-stage primary tumour tissues with varying degrees of tumour purity, ranging from 0.18 to 0.99 (Figure 4.1c,f,g).

The HCMI aims to generate up to 1,000 PDO models from different types of cancer. In collaboration with the HCMI, the lab of Vincenzo Corbo at the University of Verona, produced PC PDOs. In short, PC PDOs were created by mechanically fragmenting the collected

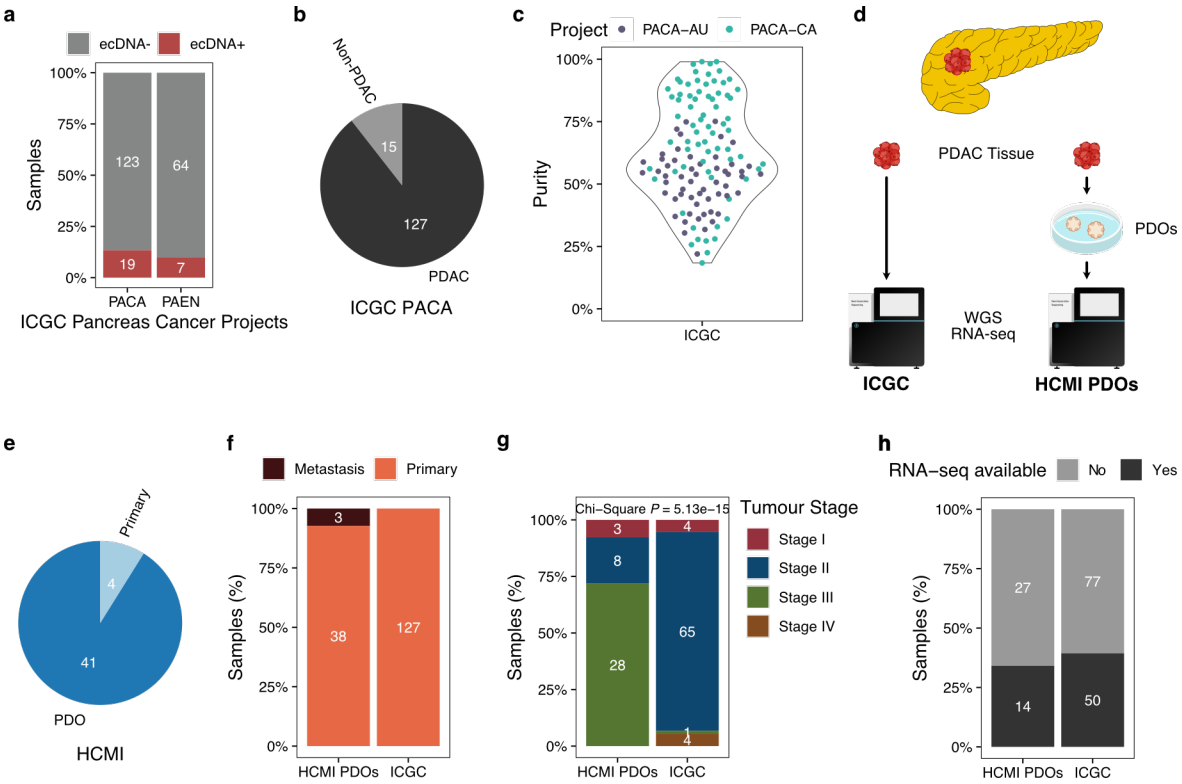


Fig. 4.1 | ICGC and HCMI sample overview. **a**, The ICGC pancreas cancer projects Pancreatic Cancer (PACA) and Pancreatic Cancer Endocrine neoplasms (PAEN) analysed by Kim et al. (2020) and their ecDNA+ and ecDNA– sample frequency. **b**, The ICGC PACA samples grouped into PDAC and non-PDAC depending on tumour histology type (Methods Section 2.25). **c**, Purity levels of PDAC tumours analysed in ICGC PACA-AU and PACA-CA. **d**, Schematic overview of project strategy to generate tumour sequencing data. **e**, WGS samples analysed from the HCMI cohort. **f**, Specimen origin by project type. **g**, Comparison of tumour stage in the two project groups. Stage known or assignable for 39/41 HCMI PDOs and 74/127 ICGC primary tumours. **e**, Matching RNA-seq data available for the HCMI PDOs and the ICGC PDAC samples.

tumour specimen and then embedding it in Matrigel. After organoid establishment and selective outgrowth of cancer organoids, the PC PDOs comprise a heterogeneous population of neoplastic cells (Boj et al., 2015; Seino et al., 2018). For a comprehensive genomic analysis, encompassing the identification of amplicons, including ecDNAs, a collection of 41 PC PDOs and four PDO-matching primary tumours underwent WGS (Figure 4.1e). To note, these PC PDOs were primarily established from PDAC and will be referred to as PDAC PDOs hereafter (Extended Data Table 1).

The sequencing coverage in the HCMI WGS data averaged approximately 15x for each sample, which is sufficient for copy number and amplicon analysis. No matching normal samples were simultaneously sequenced. The PDOs were generated from tissue biopsies of PDAC primary tumours ($n = 38$) and lymph node metastases ($n = 3$), and they were sequenced after successful establishment (Figure 4.1f). Compared to the PDAC tumours in ICGC, a notable proportion of HCMI tumours, which were the source of PDOs, were found in advanced stages of the disease (Figure 4.1g).

To identify ecDNAs and other amplicons, the nf-core/circdna pipeline was utilised, incorporating the functions of AmpliconArchitect and AmpliconClassifier (Results Chapter 3). Full amplicon information was generated for the sequenced HCMI samples, enabling comprehens-

ive downstream analysis. However, the ICGC PDAC data only provided accurate information regarding amplicon types and numbers in each sample. The identification of ecDNA content was not feasible using the published data by Kim et al. (2020). Hence, differing analyses were conducted on the respective datasets depending on the available information.

Numerous studies analysing the ICGC datasets have been published, and many findings are publicly accessible on the ICGC data portal (dcc.icgc.org/). This included RNA-seq, clinical, and mutational data. Here, matching RNA-seq data was available of 50 ICGC PACA-CA samples, while the PACA-AU dataset lacked matching WGS and RNA-seq samples. In addition, RNA-seq was performed on 14 HCMI PDOs (Figure 4.1h).

As previously mentioned, the HCMI WGS data was generated without normal samples, which are essential for distinguishing germline single nucleotide polymorphisms (SNPs) or mutations from somatic mutations. For this reason, the newly generated WGS data was not utilised for mutational calling. Instead, mutational calls for all PDOs are provided by the HCMI consortium, who performed genomic sequencing on the corresponding PDOs along with their matching normals. This dataset was available for analysis but under embargo until official publication. To bypass publication restrictions, the PDOs underwent re-sequencing. Only the mutational calls from the original HCMI data were utilised. All other analyses were carried out on the WGS data generated by the Corbo Lab.

4.2 EcDNAs are common in PDAC

Recent revolutionary studies have identified ecDNAs in almost all cancer types, indicating that ecDNAs are a common feature in cancer genomes (Turner et al., 2017; Kim et al., 2020). The collective analysis of PC revealed that more than 10% of PC tumours contain ecDNAs (Kim et al., 2020). To identify the PDAC-specific occurrence, the samples in the ICGC and HCMI datasets were classified based on their amplicon types as 'Circular', 'BFB', 'Complex' or 'Heavily-rearranged', and 'Linear'. Circular amplicons are ecDNA-based amplicons that form one or more circular structures. BFB amplicons refer to amplicons with a BFB signature. Complex amplicons contain distal or interchromosomal segments, while linear amplicons relate to focal amplifications. For samples with multiple amplicons, the classification was based on the amplicon with the highest priority. Priority was defined as follows: Circular > BFB > Complex/Heavily-rearranged > Linear. Samples containing no amplicons were classified as having no focal somatic copy number amplification (No-fSCNA), as originally defined by Kim et al. (2020).

Analysis of the ICGC data revealed that 14.2% of PDAC primary tumours contain ecDNA (Figure 4.2a). This is a slightly higher frequency compared to the overall PACA cohort, which included non-PDAC samples (14.2% vs. 13.4%). The HCMI PDOs exhibited an even higher frequency of ecDNA, with almost 30% of the PDAC PDOs (12 out of 41, 29.27%) containing at least one ecDNA. In contrast, the frequency of samples with no identified

amplicons (No-fSCNA) is comparable between the ICGC and HCMI PDAC samples (26.83% vs. 33.9%).

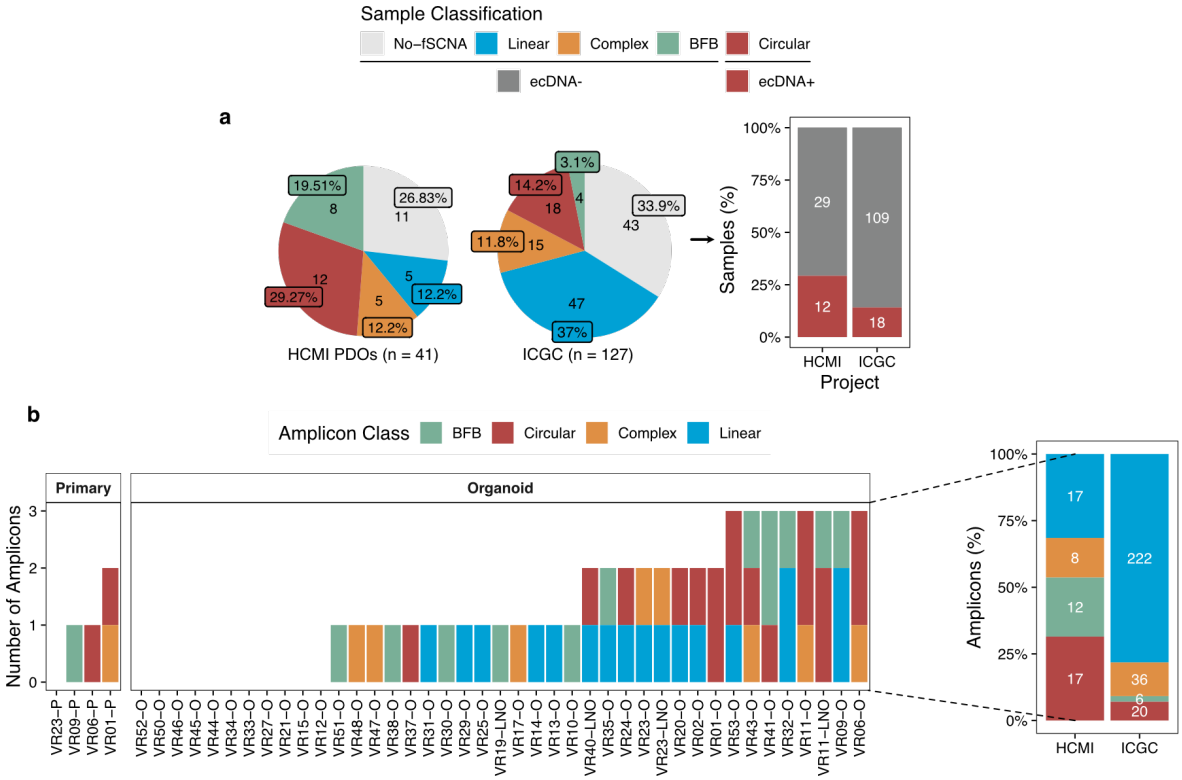


Fig. 4.2 | Amplicon and sample classification of ICGC primary and HCMI PDO samples. a, HCMI PDO and ICGC primary tumour sample classifications based on their amplicon types (Methods Section 2.28). Samples without any amplicon are termed No-fSCNA (no focal somatic copy number amplification detected). Sample and amplicon classification for ICGC primary tumours was obtained from Kim et al. (2020). Samples with a detected circular amplicon were classified as ecDNA+, all other samples were termed ecDNA–. **b**, Number of amplicons per sample in the HCMI cohort (left). Number of amplicons per amplicon type in the HCMI PDO and ICGC samples (right).

A closer look at the amplicon landscape within the ICGC cohort revealed a significant prevalence of linear amplicons (Figure 4.2b). Similarly, the HCMI PDOs showed a high frequency of linear amplicons (n = 17, 31.4% of all amplicons). However, in contrast to the ICGC primary tumours, the HCMI PDOs also present an abundance of circular amplicons (n = 17).

This comparison of the different amplicon classes in the ICGC and HCMI cohorts revealed a substantial discrepancy between the amplicon types identified. It is noteworthy that the methodology for amplicon classification by AmpliconClassifier has changed between the legacy version (Kim et al., 2020) and the version used in the HCMI analysis (Luebeck et al., 2023). However, it remains uncertain whether these changes significantly affect amplicon classes called, and no considerable differences are expected between ecDNA calling in the old and new versions.

Tumour purity plays an important role in somatic mutation and copy number calling (Xu et al., 2014; Zare et al., 2017; Alioto et al., 2015). PDOs are purely neoplastic cells and achieve high levels of tumour purity, making them an ideal tool for identifying cancer-specific features (Seino et al., 2018). In contrast, primary tumours are a complex mixture of normal and

cancer cells. Consequently, primary tumours exhibit varying degrees of purity, as observed in the ICGC PDAC primary tumour dataset, potentially affecting copy number calling methods, structural variation (SV) detection, and amplicon determination (Figure 4.1c) (Carter et al., 2012). Therefore, the observed amplicon differences between the ICGC dataset and the HCMI PDOs could also be explained by a higher tumour cell purity in the PDOs that allow a more accurate amplicon determination.

Furthermore, as described in Results Section 4.1, the tumour stage of the HCMI PDAC cohort differs significantly from that of the ICGC cohort. In the study by Luebeck et al. (2023), an association between ecDNAs and late stage disease was identified in oesophageal adenocarcinoma. Accordingly, I investigated whether a similar association would be observed in PDAC by integrating tumour stage and ecDNA occurrence. However, no such association was found (Figure 4.3) (Luebeck et al., 2023).

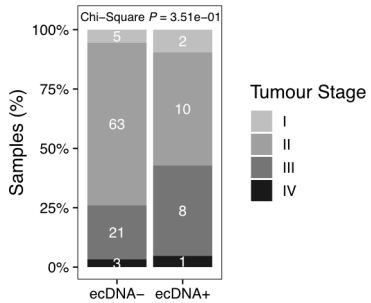


Fig. 4.3 | No association between PDAC stage and ecDNA presence. EcDNA presence is not significantly associated with tumour stage (P value = 3.51e-01, chi-squared test, two sided) in the HCMI PDO and ICGC samples. The tumour stage, as defined by the AJCC staging system, was obtained from the clinical information of the PDAC tumours. To simplify, the tumour stage was broadly categorised into a four stage classification without dividing into the tumour stage subcategories.

These findings from the analysis of the HCMI PDOs and the ICGC primary tumours suggest that ecDNAs may be more prevalent than originally believed. Moreover, circular amplicons make up a considerable portion of all amplicons in the HCMI PDOs, hinting at the possibility of ecDNAs serving as a huge driver for oncogene amplifications in PDAC.

4.3 EcDNAs are retained in PDAC PDOs

In cancer research, model systems are critical for representing specific disease aspects. While primary tumour analysis is the most accurate method in reflecting the cancer genotype and phenotype, PDAC PDOs have been shown to closely mimic the genomic and transcriptomic landscape (Tuveson & Clevers, 2019; Tiriac et al., 2018; Nam et al., 2022). This characteristic makes PDOs ideal for functional genomic studies. While other cancer model systems, such as patient-derived neurospheres or cell lines, have been used previously to investigate the roles of ecDNAs in multiple cancer types, ecDNAs, to my knowledge, have not yet been detected or analysed in PDOs (Nathanson et al., 2014; Wu et al., 2019; Turner et al., 2017).

The foregoing analysis revealed that ecDNAs are prevalent in PDAC PDOs. To establish whether these ecDNAs stem from the primary tumours, a comparative analysis of the WGS

data was carried out on the four patients with corresponding primary tumours and PDOs.

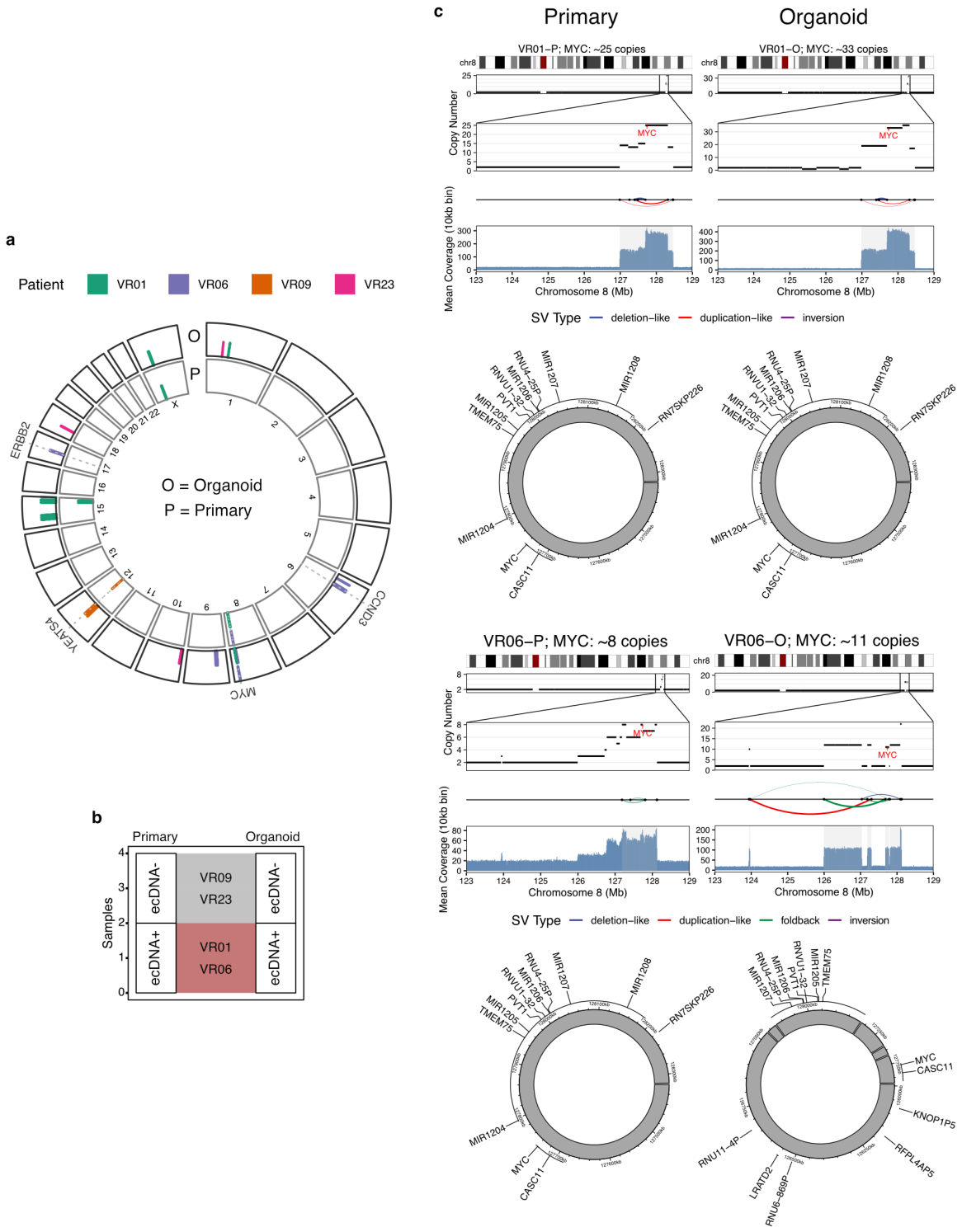


Fig. 4.4 | Retention and possible evolution of *MYC*-ecDNAs in PDAC PDOs. **a**, Circos plot displays the amplicon regions that were identified in the primary tumours and their corresponding PDOs in four patients. Although PDOs exhibit a greater number of amplicon regions, they are still able to retain the same amplicon regions as the matching primary tumour. **b**, EcDNA status in both the primary tumour and matching PDOs. **c**, Retention and potential evolution of *MYC*-ecDNAs in primary tumour and PDO of the two ecDNA+ patients (VR01, VR06).

The analysis uncovered that the amplicons identified in the primary tumours were also identifiable in the PDOs (Figure 4.4a). Specifically, three of four primary tumours contained similar amplicon regions as those found in the PDOs. Remarkably, the PDOs exhibited a

larger number of amplicons when compared to their corresponding primary tumours. This suggests that the pure neoplastic nature of the PDOs might provide a more comprehensive view of the genomic landscape compared to primary tumours, which generally contain a mixture of neoplastic and non-neoplastic cells.

Additionally, the ecDNA-positivity in the primary tumours was retained in the corresponding PDOs. Two of the four primary tumours (VR01-P and VR06-P) were identified to be ecDNA+, while the other two, VR09-P and VR23-P, either contained a BFB amplicon or no amplicon (Figure 4.4b and Figure 4.3b). This indicates that the ecDNA status is likely conserved during PDO establishment from the primary tumour.

Of significance, both patients with ecDNAs carried a *MYC*-bearing ecDNA (Figure 4.4c). The amplicon structure and landscape are entirely overlapping in VR01-P and VR01-O, emphasising the correlation between these ecDNAs and indicating that the VR01-P ecDNA was retained in VR01-O. In contrast, VR06-P and VR06 exhibited distinct amplicon structures. Despite their similar gene content, including *MYC*, the SV and ecDNA composition are vastly different. When comparing the amplicon similarities using AmpliconClassifier (Luebeck et al., 2023), no significant relationship was observed (P value = 0.12, Similarity Score = 0.265, Table 4.1). Therefore, it is unclear whether both *MYC*-ecDNAs are related. In a study by Shoshani et al. (2021), ecDNA evolution was noted under selection pressure, which might explain the structural evolution of the *MYC*-ecDNA during the PDO establishment. However, the available evidence is limited and further analysis and research is needed to identify ecDNA evolution and the actual relationship.

Tab. 4.1 | Statistical examination of *MYC*-ecDNA similarities identified in the VR01 and VR06 primary tumour and PDO.

Primary Tumour	PDO	Similarity Score	P value
VR01-P	VR01-O	0.527	0.013
VR06-P	VR06-O	0.265	0.123

Taking together, despite the absence of a significant similarity between the *MYC*-ecDNA in VR06-P and VR06-O, it is suggested that the PDOs reflect the amplicon landscape of PDAC primary tumours, and also retain their ecDNA characteristics. These findings underline the potential of PDOs as a representative model system to studying the role of ecDNA in PDAC.

4.4 The amplicon landscape of PDAC PDOs

Comprehending the genomic landscape of PDAC is pivotal for tailoring personalised therapeutic approaches (Pishvaian et al., 2020). To improve our understanding of the genomic complexity and identify potential targetable alterations, we analysed the amplicon heterogeneity and landscape, specifically the ecDNA landscape, in the PDAC PDOs (Figure 4.2).

The analysis included 41 PDOs, with a diverse set of amplicons (Figure 4.2a,b). Of the 41

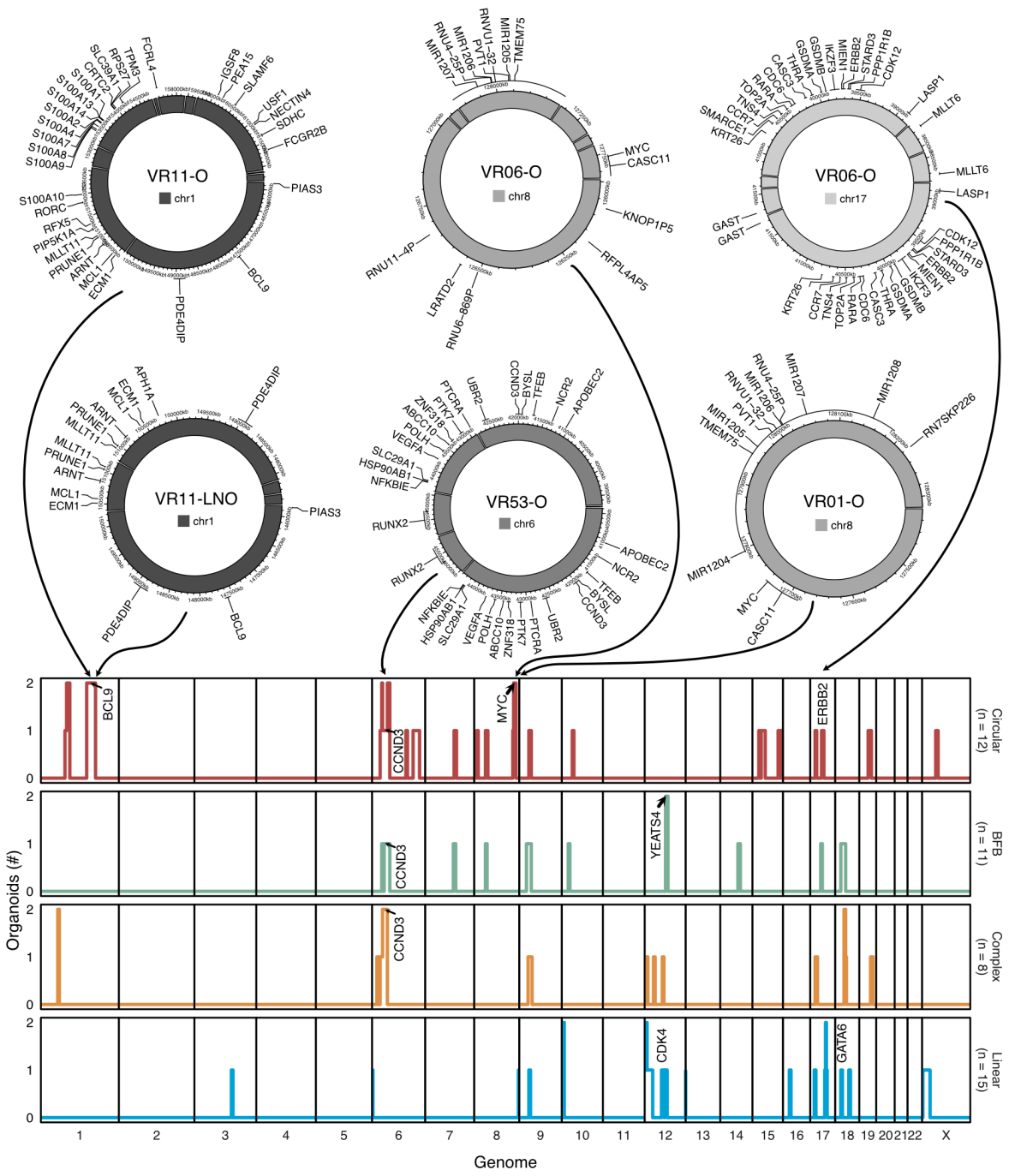


Fig. 4.5 | Amplicon landscape of PDAC PDOs. Genome-wide distribution of amplicons by amplicon class in the HCMI PDOs (n = 41). The amplicons were counted for each 5 Mbp bin and are shown as the total number of PDOs containing a given amplicon class. Recurrent or PDAC driver genes are highlighted. Selected putative ecDNAs (circular amplicons), their structure and gene content are displayed above the genome plot. To improve readability, only cancer driver genes defined in the allOnco gene database are plotted for the VR06-O chr17-ecDNA, the VR53-O chr6-ecDNA, and the VR11-O chr1-ecDNA. The number of PDOs with an amplicon of the respective class is denoted by n below each amplicon class title.

PDOs, 30 had amplicons, with 16 of these containing multiple amplicons (Figure 4.2b). Overall, 54 amplicons were identified, exhibiting an extensive genomic distribution (Figure 4.5).

The most common cancer driver gene found on amplicons was *CCND3* (n = 4) a well-known activator of the cell cycle in PDAC (Radulovich et al., 2010). Furthermore, one of the four amplicons containing *CCND3* was also classified as circular, revealing a *CCND3*-bearing ecDNA. Recurrent PDAC driver genes were less frequent, observed in a maximum of two

PDOs, indicating a remarkably diverse amplicon landscape in PDAC.

Twelve ecDNA+ samples were identified, containing a total of 17 circular amplicons. *MYC* was found on ecDNAs in two separate PDOs. Two PDOs from a single patient (VR11) were found to contain the *BCL9* gene, one originating from a primary tumour (VR11-O) and the other from a lymph node metastasis (VR11-LNO). While additionally, *CCND3* and *ERBB2* were each present on a single ecDNA.

Upon examining the putative structure of the *MYC*-ecDNAs in PDOs VR01-O and VR06-O, it was observed that while ecDNAs may contain similar genes, they consist of diverse genomic fragments. The chromosomal breakpoints that give rise to ecDNA formation appear to be randomly distributed and do not display any specific patterns. However, recent research discovered that the contents of ecDNA are non-random and overwhelmingly consist of cancer-specific driver genes – a conclusion that this analysis of PDAC PDOs further supports (Kim et al., 2020; Luebeck et al., 2023). Moreover, the putative structure of *BCL9*-ecDNAs of VR11-O and VR11-LNO appears to have slight differences. However, a subsequent analysis will be conducted to scrutinise both amplicons for any similarities or differences to determine the likelihood of a shared origin.

In brief, these findings underline the potential significance of ecDNAs in PDAC by amplifying PDAC-specific cancer driver genes, such as *MYC*, *ERBB2*, and *CCND3*. The ecDNA landscape heterogeneity in PDAC hints that ecDNAs may have diverse roles that are possibly impacted by the particular cancer genes they comprise. This emphasises the potential necessity for detecting and determining ecDNA content for personalised therapeutic strategies for PDAC.

4.5 Distinct transcriptomic profiles in ecDNA+ tumours and PDOs

Studies have indicated that ecDNA is linked to aggressiveness, enhanced proliferation, and an unfavourable prognosis for patients (Kim et al., 2020; Koche et al., 2020). Given these connections, this sections aims to elucidate the biological mechanisms associated with ecDNA+ PDAC tumours and PDOs, respectively. For this, an integrative transcriptomic analysis was conducted using matching RNA-seq data. The ICGC dataset included 50 primary tumours with amplicon information and corresponding transcriptomic data, of which seven were identified as ecDNA+ and 43 as ecDNA–. Additionally, RNA-seq was performed on 14 of the HCMI PDOs, seven of which were classified as ecDNA+ and seven as ecDNA–.

4.5.1 EcDNAs are associated with a Basal-like signature in PDAC tumours

Presently, PDAC is divided into two to five subtypes by various bulk tumour subtyping schemes (Bailey et al., 2016; Moffitt et al., 2015; Chan-Seng-Yue et al., 2020; Collisson et al., 2011; Raghavan et al., 2021; Puleo et al., 2018; Collisson et al., 2019). However, despite the existence of diverse subtypes, these schemes generally classify PDAC tumours into two broad subtypes. One subtype, called 'Classical' or 'Pancreatic Progenitor' is associated with the the activation of pancreatic lineage genes and a favourable patient outcome. The other subtype, termed 'Basal' or 'Squamous', is linked to poor patient outcomes, and the activation of epithelial-to-mesenchymal transition (EMT), *MYC*, and proliferation pathways (Bailey et al., 2016; Collisson et al., 2019).

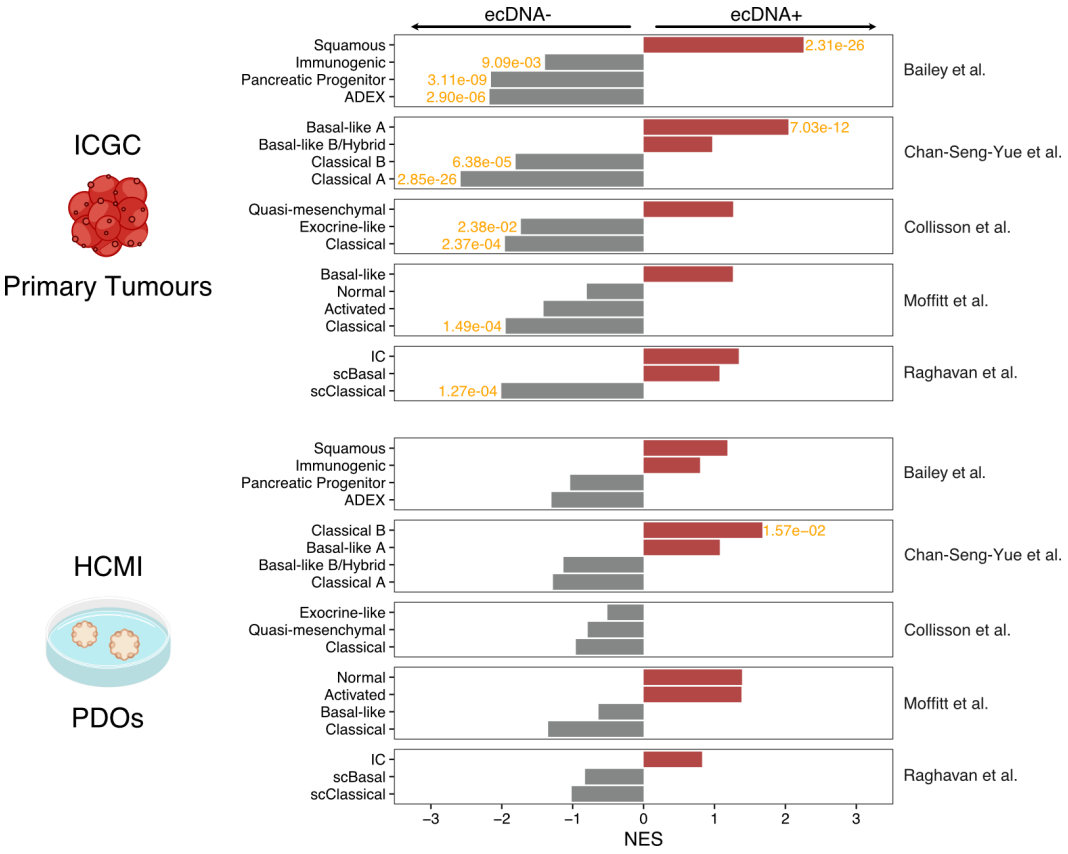


Fig. 4.6 | EcDNAs are associated with a Basal/Squamous signature in PDAC tumours. Barplot displaying gene set enrichment analysis results of subtype gene signatures in ecDNA+ (ICGC: $n = 7$, HCMI: $n = 7$) and ecDNA– (ICGC: $n = 43$, HCMI: $n = 7$). Significant association (P value < 0.05) is displayed in orange. Subtype gene signatures are defined by Bailey et al. (2016), Chan-Seng-Yue et al. (2020), Collisson et al. (2011) and Raghavan et al. (2021), and Moffitt et al. (2015). The Chan-Seng-Yue et al. (2020) signatures are termed based on their expression in specific subtypes: Classical A (Signature 1), Classical B (Signature 6), Basal-like A (Signature 2), Basal-like B/Hybrid (Signature 10) (Chan-Seng-Yue et al., 2020).

To determine whether a PDAC subtype correlates with ecDNA+ samples, a transcriptomic analysis was conducted on the ICGC primary tumours and the HCMI PDOs. This comprised of a differential expression analysis between ecDNA+ and ecDNA– primary tumours and HCMI PDOs, revealing deregulated genes between the different states (Extended Data Table 4). Following this, a gene set enrichment analysis was conducted using various subtyping gene set signatures with the objective of unveiling subtype associations.

In the ICGC dataset, tumours positive for ecDNAs exhibited notable correlations with the Squamous subtype of the Bailey classification and the Basal-like A subtype of the Chan-Seng-Yue signatures (Figure 4.6). While other Basal gene signatures, such as Quasi-mesenchymal or Basal-like, showed enrichment, their activation was not established as statistically significant (P value < 0.05). In contrast, tumours lacking ecDNA revealed significant enrichment for gene signatures associated with the Classical subtypes (Pancreatic Progenitor, Classical, scClassical). This suggests that ecDNA+ tumours are associated with a Basal phenotype, whereas ecDNA– tumours exhibit characteristics of Classical gene signatures.

However, an equivalent analysis of the HCMI PDOs failed to establish a comparable correlation (Figure 4.6). Only the Classical B subtyping was found to be enriched in ecDNA+ PDOs, which contradicts the results from primary tumours. Additionally, an enrichment of the Squamous subtype was observed in ecDNA+ PDOs, but no significant association was found.

PDAC PDOs often shift from their Basal subtype to a more Classical phenotype due to the presence of growth factors and chemokines in the organoid media (Raghavan et al., 2021). These utilisation of such additives, which are necessary for prolonged growth, seems to impact the transcriptomic composition of PDOs (Boj et al., 2015; Raghavan et al., 2021). Furthermore, while early-passage analysis might recapitulate the primary disease subtype, extended propagation leads to the expression of Classical gene programmes (Tiriac et al., 2018; Raghavan et al., 2021). Sequencing of the HCMI PDOs was carried out at various passage levels, meaning the subtyping accuracy may have been compromised (Extended Data Table 6).

In summary, these findings demonstrate a link between ecDNA+ PDAC primary tumours and Basal/Squamous gene signatures, which are recognised to align with increased tumour aggressiveness and poor patient outcomes (Collisson et al., 2019; Bailey et al., 2016). While the PDO analysis yielded no comparable connections, the inadequacy of the PDO RNA-seq data may have impeded subtype analysis. Therefore, it is not possible to definitively exclude an association between the Basal subtype and ecDNA+ tumours. Nevertheless, more research is required to validate this association, using additional datasets generated, ideally, from PDAC primary tumours.

4.5.2 Differential transcriptomic signatures in ecDNA+ PDAC tumours and PDOs

Transcriptomic profiling has uncovered diverse biological processes that impact the growth and progression of PDAC tumours (Bailey et al., 2016; Peng et al., 2019). However, the connection between ecDNAs and distinct biological programs in the context of PDAC remains unexplored. Therefore, a gene set enrichment analysis was conducted on the transcriptomic data of the ICGC PDAC primary tumours and the HCMI PDOs, with the aim of identifying

active gene programmes described in the Hallmark gene sets (Liberzon et al., 2015).

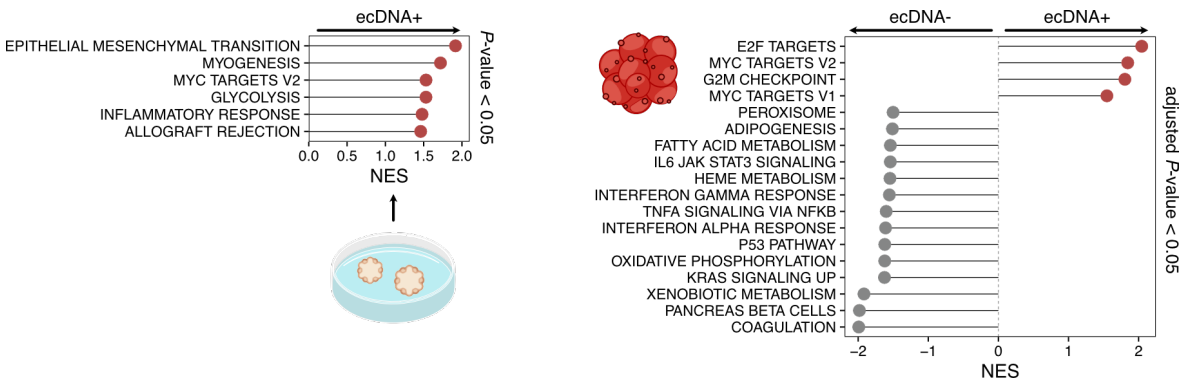


Fig. 4.7 | Hallmark pathway activation in ecDNA+ and ecDNA– PDAC tumours and PDOs. Hallmark pathway gene set enrichment analysis compares ecDNA+ and ecDNA– samples within ICGC primary tumours (n = 50, right) and HCMI PDOs (n = 14, left). Significantly enriched pathways (HCMI: P value < 0.05, ICGC: adjusted P value < 0.05) for both groups are displayed. For the ICGC data, the P values were adjusted using the Benjamini-Hochberg method to highlight highly significant pathways.

The analysis of the primary tumours in the ICGC data revealed upregulation of cell cycle pathways, E2F targets, G2M checkpoints, and MYC targets in ecDNA+ tumours (Figure 4.7, Extended Data Table 4). This trend of increased MYC activity was also significantly evident (P value < 0.05) in ecDNA+ PDOs.

Interestingly, primary tumours that are ecDNA+ display a downregulation of innate immune pathways, including TNFA signalling via NF-KB (genes regulated by NF-KB in response to TNF) or IL6 JAK STAT3 signalling (genes upregulated by IL6 via STAT3) (Kumari et al., 2016; Dolcet et al., 2005). This finding is consistent with a pan-cancer study that linked ecDNA presence with reduced immune activity (Wu et al., 2022b). Thus, the findings indicate that immune activity may also be inhibited in ecDNA+ PDAC tumours.

The P53 pathway, which is integral to the DNA repair system, was upregulated in ecDNA– compared to ecDNA+ tumours. However, it was also found that a higher *TP53* mutation frequency was observed in ecDNA+ tumours (85.7%, 6 out of 7) compared to ecDNA– tumours (55.8%, 24 out of 43). However, TP53 downregulation in ecDNA+ tumours cannot be excluded, as no significant association between TP53 inactivation and ecDNA positivity was found in this subset of samples (P value = 0.219, Figure 4.8). This, together with the activation of pathways involved in proliferation (namely E2F targets, MYC targets, and G2M checkpoint) in ecDNA+ tumours, could potentially exacerbate replication stress and result in extensive DNA damage (Macheret & Halazonetis, 2015).

This analysis concludes that several gene programmes are linked to ecDNA+ and ecDNA– PDAC samples. Most notably, MYC targets were enriched in ecDNA+ tumours and PDOs suggesting a link between ecDNA-positivity and MYC activation. Alongside with the down-regulation of innate immune pathways and the activation of cell proliferation pathways, ecDNA+ PDAC tumours might exhibit higher proliferation rates and enhanced metastatic potential in PDAC (Maddipati et al., 2022; Malumbres & Barbacid, 2009; Hagerling, Casbon & Werb, 2015). Of note, the P values resulting from the PDO analysis were not adjusted for

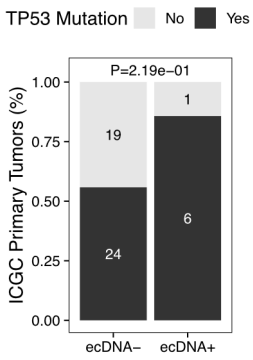


Fig. 4.8 | TP53 inactivation in ecDNA+ and ecDNA– ICGC tumours with transcriptomic data. P values were calculated using two-sided Fisher's exact test.

multiple testing since no pathways were found to be significantly enriched (P adjusted value < 0.05) after adjustment. Consequently, to ensure their validity, the PDO analysis should only be evaluated in conjunction with the primary tumour analysis.

4.6 Presence of ecDNA in PDAC is linked to genomic instability

Cancer genomes with genomic instability exhibit high mutation rates, complex genomic rearrangements, and unique transcriptomic profiles (Negrini, Gorgoulis & Halazonetis, 2010; Shoshani et al., 2021; Carter et al., 2006). The occurrence of ecDNA has been linked to several genomic instability characteristics, such as chromothripsis or BFB cycles (Shoshani et al., 2021; Kim et al., 2020). Moreover, a recent study conducted by Luebeck et al. (2023) found that whole-genome duplications and *TP53* alterations are associated with ecDNA+ oesophageal cancer, although a definite link in PDAC remains to be established.

The transcriptomic analysis of ecDNA+ and ecDNA– PDAC tumours revealed a downregulation of the *TP53* pathway in ecDNA+ PDAC tumours, suggesting that genomic instability may play a role in ecDNA formation in PDAC. To identify its role, several analyses were conducted using WGS and RNA-seq data from the ICGC PDAC primary tumours and the HCMI PDAC PDOs.

4.6.1 Mutational landscape defining ecDNA+ and ecDNA– PDAC

The genetic mutations defining cancer tumours provide insights into altered biological processes. PDAC typically exhibits the recurrence of four key driver genes, *KRAS*, *TP53*, *SMAD4*, and *CDKN2A*, and a long tail of infrequent gene alterations (Waddell et al., 2015).

Upon comparing the mutational landscape of ecDNA+ and ecDNA– tumours and PDOs (Figure 4.9a), it was found that all recurrent PDAC driver genes, except for *TP53* (P value = 0.00574, Fisher's exact test, Figure 4.9b), had a non-significant different alteration frequency (P value > 0.05, Fisher's exact test). In ecDNA– PDAC, *TP53* was found to be mutated in

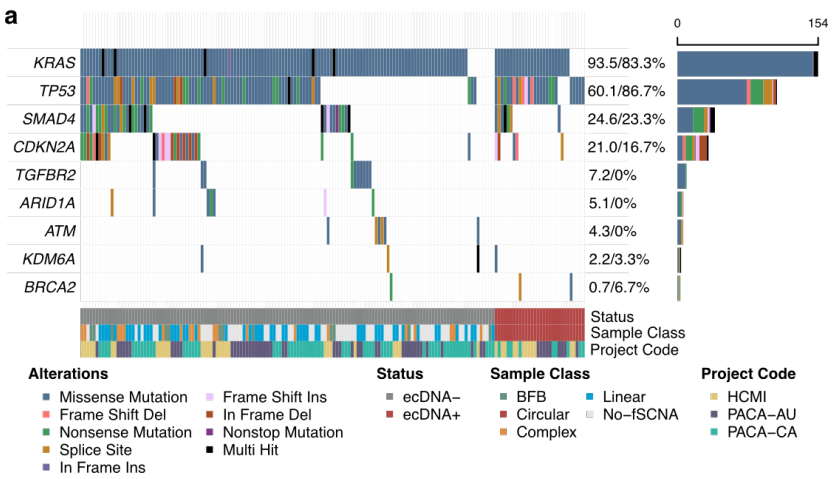


Fig. 4.9 | Overview of key PDAC gene alterations in ecDNA+ and ecDNA– tumours and PDOs. a, Oncoplot displaying recurrent alterations identified in each sample of the HCMI PDOs and the ICGC PACA-CA and PACA-AU primary tumours. The samples are divided by ecDNA status. The proportion of altered genes in ecDNA– and ecDNA+ samples are shown on the right. **b**, *TP53* was identified to be significantly altered in ecDNA+ tumours and PDOs compared to ecDNA– tumours and PDOs. The *P* value was calculated using a Fisher's exact test.

60.1% (83 of 138), while in ecDNA+ PDAC, it had an alteration frequency of 86.7% (26 of 30). Notably, all PDOs with ecDNA+ exhibited *TP53* mutations. This provides evidence for a link between *TP53* mutation and ecDNA occurrence in PDAC and supports the evidence identified in oesophageal cancer (Luebeck et al., 2023).

4.6.2 EcDNA+ tumours are associated with an unstable genome

SV detection offers a reliable method for measuring genomic instability and identifying unstable genomes (Waddell et al., 2015). Chromothripsis, a complex shattering of chromosomes, is a second metric for genomic instability that is frequent in PDAC and has also been linked with ecDNA formation in cancer (Cortés-Ciriano et al., 2020; Shoshani et al., 2021). The notable pan-cancer study, conducted by Kim et al. (2020), has discovered that around 36% of ecDNAs display a chromothripsis signature, surpassing the prevalence of all the other amplicon types.

To establish whether there exists a correlation between ecDNA presence and either chromothripsis or SV abundance, a comprehensive analysis was performed using ICGC amplicon data, chromothripsis data from (Cortés-Ciriano et al., 2020), and SV data from the ICGC database (release 28, https://dcc.icgc.org/).

Interestingly, PDAC tumours harbouring ecDNA were significantly impacted by high confidence chromothripsis events (Figure 4.10a). Specifically, 68% of ecDNA+ samples were affected by at least one high confidence chromothripsis event, the highest proportion as compared to other sample classes. Compared to samples lacking amplicons (No-fSCNA) and samples containing solely linear amplicons (Linear), a significantly greater proportion of ecDNA+ PDAC tumours (circular) were affected by chromothripsis events (No-fSCNA: $P = 4.85 \times 10^{-5}$, Linear: $P = 0.0292$). Despite slightly missing significance ($P = 0.0502$) when

compared to samples classified as Complex, there is also a similar trend present.

This analysis explored the correlation between chromothripsis-affected samples and the presence of ecDNA. However, it is unclear if an observed chromothripsis event contributed to the formation of the respective ecDNAs in the sample. Due to the limitations of the available datasets, precise integration of the chromosomal location of the chromothripsis event and the ecDNA was not feasible.

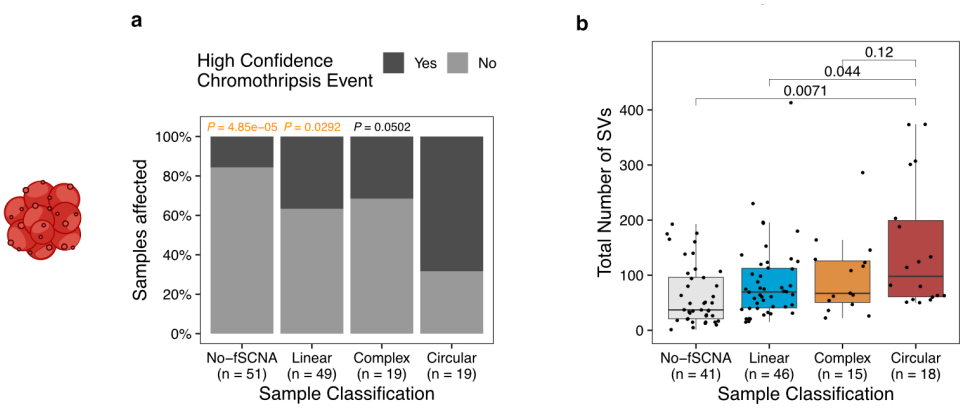


Fig. 4.10 | EcDNA presence is associated with chromothripsis events and an abundance of SVs. a, Barplot displaying the proportion of samples with high-confidence chromothripsis events, organised by sample class. A pairwise Fisher's exact test was performed to compare the statistical significance among each sample class. Circular samples (ecDNA+ tumours) serving as the reference. P values are displayed above each sample class compared against the reference, with significant results (P value < 0.05) highlighted in orange. **b**, Comparison of the total number of SVs identified in each sample grouped by sample class. Statistical significance was evaluated using a Student's t-tests. **a & b**, BFB samples were excluded due to low sample size ($n = 4$). Only samples with available chromothripsis or SV information are included. No-fSCNA, No focal somatic copy number amplification detected.

Patterns and SV numbers can define distinct genomic subtypes. Notably, a high number of SVs are linked with unstable PDAC genomes (Waddell et al., 2015). Analysing the integration of SV and amplicon data for ICGC primary tumours confirmed that PDAC tumours, which contain ecDNAs (Circular), have the highest total number of SVs compared to the other three sample classes, No-fSCNA, Linear, and Complex. Significance was found in Circular vs Linear ($P = 0.0071$) and Circular vs No-fSCNA ($P = 0.044$). Similarly to the results of the chromothripsis analysis, a similar trend was observed between Circular and Complex samples, but the statistical significance was slightly missed ($P = 0.12$).

Overall, although statistical significance was lacking in comparisons between Circular and Complex samples, the analyses demonstrate an association between ecDNA presence and significant genomic instability in PDAC. In general, chromothripsis and high numbers of SVs were more frequent in ecDNA+ PDAC. These results emphasise the relationship between genomic instability and the tendency for ecDNA creation in PDAC tumours.

4.6.3 EcDNA-positivity is associated with whole-genome duplications

Polyploidy is a common event in cancer, linked to genomic instability, *TP53* mutation, or cell cycle disruption (Bielski et al., 2018). Previous analysis has demonstrated the association between ecDNA occurrence and *TP53* mutation as well as the link to genomic instability in

ecDNA+ samples. Therefore, a subsequent analysis investigating the tumour ploidy, particularly the whole-genome duplication status, was performed. For the primary tumours, the whole-genome duplication status was obtained from the ICGC data portal (dcc.icgc.org/) and the HCMI PDO whole-genome duplication status was predicted through the utilisation of AMBER, COBALT, and PURPLE in tumour-only mode (github.com/hartwigmedical/hmftools).

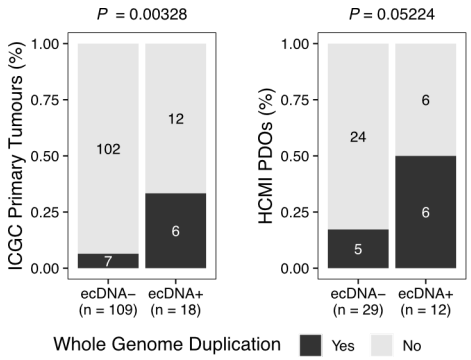


Fig. 4.11 | Whole-genome duplication status in ecDNA+ and ecDNA– samples. P values were calculated using two-sided Fisher's exact test. n equals the number of samples in each group.

Whole-genome duplications were prominently enriched in ecDNA+ samples (Figure 4.11). Enrichment was observed with statistical significance in the ICGC dataset ($P = 0.00328$), with the HCMI PDOs demonstrating a similar trend, albeit without achieving statistical significance ($P = 0.05224$), likely due to the smaller sample size. Notably, half of ecDNA+ PDAC PDOs had undergone whole-genome duplication. Combining both PDAC datasets, a significant association ($P = 8.2 \times 10^{-5}$, Fisher's exact test) is observed between ecDNA-positivity and whole-genome duplication, indicating that whole-genome duplication is commonly associated with ecDNA+ PDAC.

4.6.4 Transcriptomic chromosomal instability signature is enriched in ecDNA+ PDAC

Previous analyses have focused on genomic features that highlight the presence of genomic instability. However, transcriptomic data can also be used to determine genomic instability, particularly chromosomal instability. Carter et al. (2006) identified a transcriptomic signature consisting of 70 genes (CIN70 signature), which is highly expressed in tumours with high chromosomal instability scores. This signature has the capability to predict both chromosomal instability and a sub-optimal outcome (Carter et al., 2006). Here, I used the matching transcriptomic data from the HCMI PDOs and the ICGC primary tumours and conducted gene set enrichment analysis with the CIN70 signature, to identify a link between chromosomal instability and ecDNA-positivity.

The gene set enrichment analysis has shown, in both datasets, a significant enrichment of the CIN70 signature in ecDNA+ tumours ($P = 5.52 \times 10^{-5}$) and PDOs ($P = 0.025$, Figure 4.12). With this transcriptomic analysis, the previous results are further supported highlighting the link between diverse genomic instability characteristics and the presence of ecDNA. Overall,

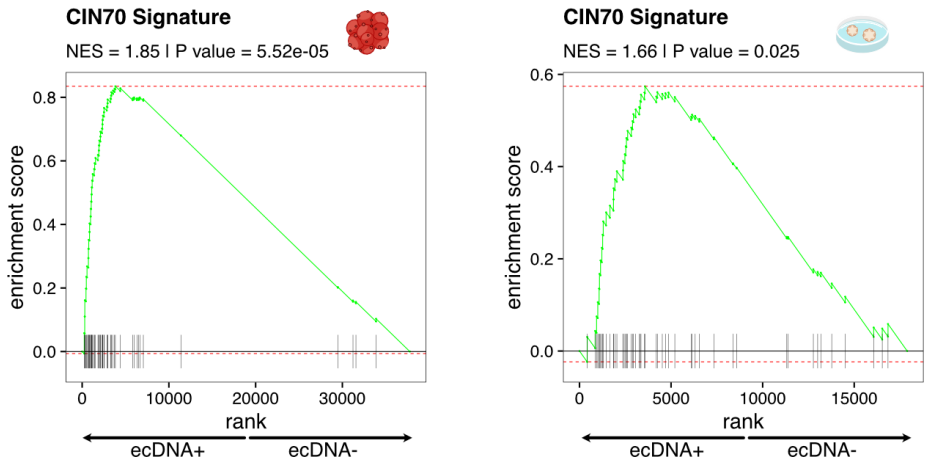


Fig. 4.12 | Chromosomal instability signature is enriched in ecDNA+ tumours and PDOs. The transcriptomic chromosomal instability signature CIN70 (Carter et al., 2006) was used for a gene set enrichment analysis of the ecDNA+ and ecDNA– ICGC primary tumours (left, $n = 50$) and HCMI PDOs (right, $n = 14$).

it has been established that PDAC tumours with an unstable genome are more prone to carry ecDNAs than tumours with a stable genome.

4.7 The role of ecDNAs in oncogene amplification and expression

It is evident that ecDNAs possess a distinct topological structure and can accumulate through uneven segregation during the cell cycle. As a result, genes found on ecDNAs are expressed to a massive extent (Wu et al., 2019; Kim et al., 2020; Yi et al., 2022). Additionally, it has been identified that oncogenes are highly enriched on ecDNAs in cancer cells (Luebeck et al., 2023). To investigate the possible association between ecDNAs and amplified oncogene expression and copy number levels in PDAC, an integrative analysis with copy number and RNA-seq data performed. For the copy number analysis, the entire PDO cohort was utilised. Conversely, the expression analysis was limited to the 14 PDOs that underwent RNA-seq.

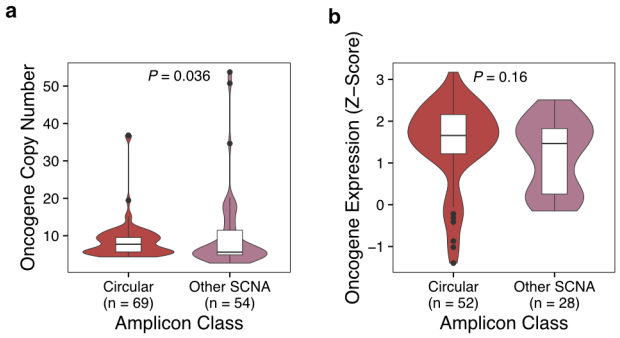


Fig. 4.13 | Copy number and transcription of extrachromosomal and chromosomal amplified oncogenes. Comparative analysis of copy number levels (**a**) and expression levels (**b**) of oncogenes identified within Circular amplicons (ecDNAs) and other types of somatic copy number amplifications (Other SCNA: BFB, Complex, and Linear amplicons). Amplicons were identified using AmpliconArchitect (Deshpande et al., 2019). Statistical significance was evaluated using a Wilcoxon Rank Sum Test. **a**, 41 samples; **b**, 14 samples

A total of 69 oncogenes, from the ONGene database (Liu, Sun & Zhao, 2017), were found to be located on ecDNAs (Figure 4.13). In line with previous studies, these 69 ecDNA-

based oncogenes had significantly higher copy number levels compared to the 54 oncogenes identified on chromosomal amplicons (Kim et al., 2020; Luebeck et al., 2023). Increased copy number levels generally suggest an elevation in transcription. Moreover, ecDNAs have been linked with increasingly accessible chromatin compared to chromosomal DNA (Wu et al., 2019). However, a significant association was not observed between ecDNAs and an increased transcription ($P = 0.16$). While the heightened copy number levels likely influence transcription in PDAC, this could not be conclusively confirmed with the existing dataset. To fully comprehend the mechanisms in PDAC, RNA-seq analysis of all PDOs is required to increase the sample size, or a further dataset must be utilised. However, this was not within the scope of this thesis.

4.8 Copy number alterations in ecDNA+ PDAC samples

Copy number alterations, which involve increasing or decreasing gene copy levels, impact gene expression and consequently the cancer biology. In PDAC, gene amplifications or deletions in important drivers can promote tumour progression and are correlated with early onset and late-stage disease (Hu et al., 2021; Chan-Seng-Yue et al., 2020). While the PDAC copy number landscape is already deciphered, we know little about copy number alterations in ecDNA+ tumours (Jones et al., 2008; Waddell et al., 2015). Thus, to investigate copy number alterations in PDAC, copy number information and ecDNA status has been integrated and analysed.

A Fisher's exact test was conducted to determine which genes are significantly amplified or deleted in ecDNA+ or ecDNA– tumours and PDOs. Whilst the frequency of amplifications (copy number ≥ 3) and deletions (copy number ≤ 1) in specific regions appears relatively similar between ecDNA+ and ecDNA–PDOs, ecDNA+ PDOs exhibit heightened gain in regions on the p arm of chromosome 6 and the q arm of chromosome 7 (Figure 4.14 top). These regions contain the oncogenes and cell cycle promoters *CCND3* ($P = 0.0053$, 5/12 ecDNA+ vs. 1/29 ecDNA–) and *CDK6* ($P = 0.05$, 4/12 ecDNA+ vs. 2/29 ecDNA–). In contrast, a region positioned on the p arm of chromosome 9 shows a considerably high frequency of copy number loss in ecDNA+ PDOs. This region includes the tumour suppressor and cell cycle inhibitor *CDKN2A* ($P = 0.0026$, 10/12 ecDNA+ vs. 14/29 ecDNA–).

After conducting the same analysis on the ICGC PDAC tumours, there were notable discrepancies in the copy number gain and loss landscape when compared to the HCMI PDOs (Figure 4.14 bottom). In general, the copy number loss frequencies in ecDNA+ and ecDNA– tumours was vastly similar. In comparison, the ecDNA+ tumours exhibit strong enrichment of copy number gains compared to ecDNA– tumours. Specifically, this analysis showed high copy number gain frequency of *MYC* ($P = 0.00156$) in ecDNA+ (14 of 17) compared to ecDNA– primary tumours (44 of 108). Additionally, *CDK6* has also been observed to be significantly amplified in ecDNA+ tumours ($P = 0.049$, 9/17 ecDNA+ vs. 30/108 ecDNA–).

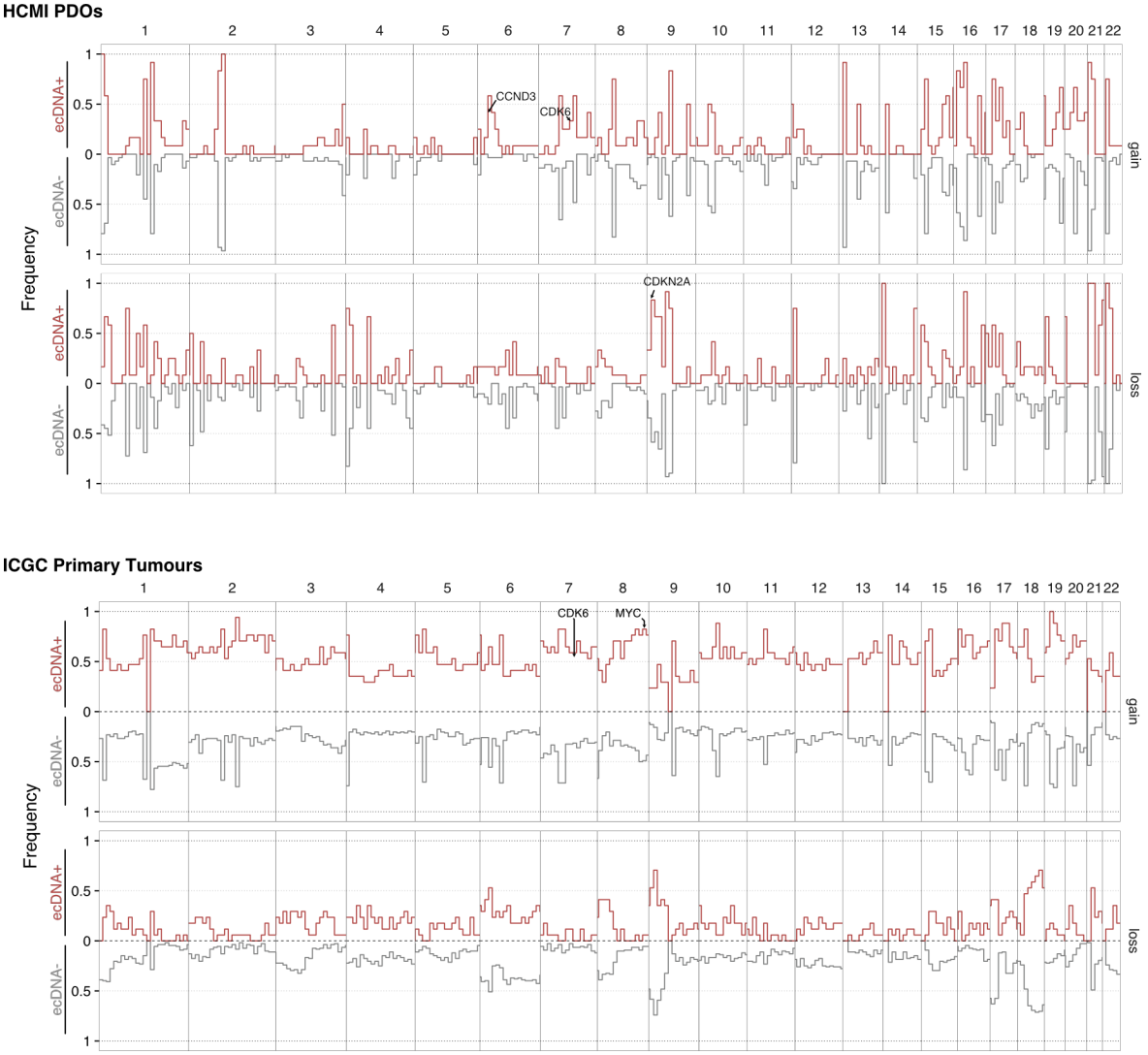


Fig. 4.14 | Genomic overview of copy number gains and losses in ecDNA+ and ecDNA– tumours and PDOs. Copy number gain and loss frequency of ecDNA+ (PDOs: $n = 12$; ICGC: $n = 17$) and ecDNA– (PDOs: $n = 29$; ICGC: $n = 108$) samples. The genome is divided and visualised in 10 Mbp bins. The frequency is calculated based on samples having either loss or gain segments inside a bin divided by the total number of samples. Altered PDAC drivers ($P < 0.05$, Fisher's exact test) between ecDNA+ and ecDNA– samples are labelled. PDO copy number loss: copy number ≤ 1; PDO copy number gain: copy number ≥ 3. ICGC copy number calls were downloaded from the ICGC Data Portal (Zhang et al., 2019b).

In conclusion, the presence of ecDNAs in PDAC samples correlates with distinct patterns of copy number alterations. Notably, genes such as *MYC*, *CCND3*, and *CDK6* are particularly amplified in ecDNA+ samples, suggesting a potential role in tumour aggressiveness and cell cycle activation. Conversely, the loss of *CDKN2A* in ecDNA+ PDOs may suggest a mechanism for evading cell cycle control. Therefore, it appears that cell cycle regulation is altered in ecDNA+, which may provide insights into the formation of ecDNAs by cell cycle and DNA damage repair deregulation.

4.9 EcDNA selection and evolution during PDO adaptation

Gene amplifications have been implicated in cancer cell drug resistance, impacting targeted therapy. EcDNA elements, specifically, are dynamically regulated in cancer cells (Nathanson

et al., 2014). Nathanson et al. (2014) demonstrated that multiple glioblastomas, identified by an EGFR-vIII mutation situated on an ecDNA, can regulate their EGFR-vIII copies by increasing or decreasing ecDNA levels respectively during treatment with tyrosine kinase inhibitor drugs, and after discontinuing treatment. This dynamic regulation can be driven by the uneven segregation during cell division and the resulting acquired cell fitness (Yi et al., 2022; Nathanson et al., 2014).

To explore the dynamic regulation of ecDNAs, we sought to examine how *MYC*-ecDNA bearing PDOs respond to microenvironmental stressors. By removing Wnt3a and R-spondin 1 (WR) from the PDO growth media, we exploited the known dependency of PDAC PDOs on WNT signalling to create an artificial selection pressure (Boj et al., 2015; Seino et al., 2018). *MYC*, a known target gene of the Wnt pathway, is presumed to be a crucial regulator for PDO survival under such stress conditions (Hao et al., 2019; Rennoll & Yochum, 2015). The previous analysis identified two *MYC*-ecDNA+ PDOs (VR01-O, VR06-O). These two PDOs, in addition with four *MYC*-ecDNA– PDOs, including three PDOs with intra-chromosomal *MYC* amplification (VR02, VR20, VR23, Figure 4.15) and one with normal *MYC* copy levels (VR29), were artificially stressed by WR removal from the normal media.

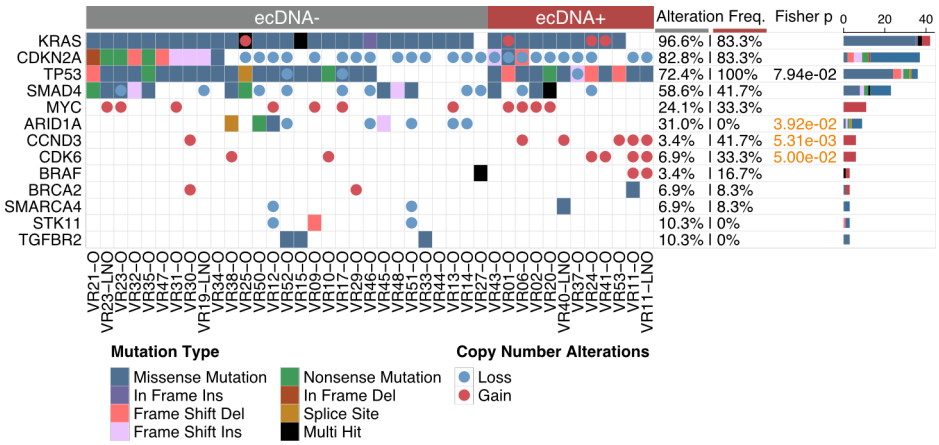


Fig. 4.15 | Overview of gene mutations and copy number alterations in the Verona PDOs. Oncoplot of mutations and copy number alterations in the HCMI organoids. Copy number alterations of gain (copy number ≥ 3) and loss (copy number ≤ 1) are displayed as dots. A Fisher's exact test was performed to compare the frequency of alteration in ecDNA– vs. ecDNA+ PDOs. *P* values below 0.1 are displayed and significant *P* values (< 0.05) are highlighted in orange.

After the withdrawal of WR, three PDOs, specifically VR02-O, VR20-O, and, VR29-O, died quickly, whereas the remaining three, VR01-O, VR06-O, and VR23-O, adapted to the environmental changes over an extended period (Figure 4.16).

By investigating the genomic landscape of the adapted PDOs, specifically their amplicon landscape, it was observed that *MYC*-ecDNA levels had significantly increased after adaptation to the -WR media. This was evident in both *MYC*-ecDNA PDOs, VR01-O and VR06-O, which increased their *MYC* copy number levels by at least four-fold in two biological replicates (Figure 4.17). Importantly, this rise in copy number coincided with an increase in *MYC* transcription (Figure 4.19a).

In contrast, the bulk *MYC* copy number levels of VR23-O did not change during and after

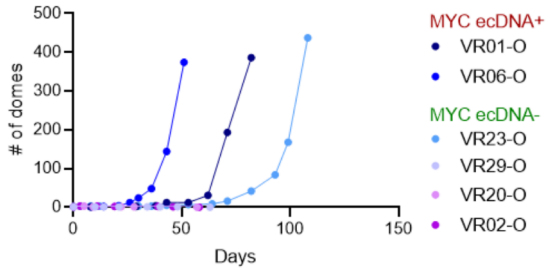


Fig. 4.16 | Adaptation and propagation of six PDOs grown in -WR media. Growth curve of *MYC*-ecDNA+ ($n = 2$) and *MYC*-ecDNA– ($n = 4$) PDOs in -WR media. Culture growth is represented as number of domes (50 µl Matrigel/dome). Antonia Malinova and Elena Fiorini, Vincenzo Corbo Lab, University of Verona, conducted the experiment and generated the graph.

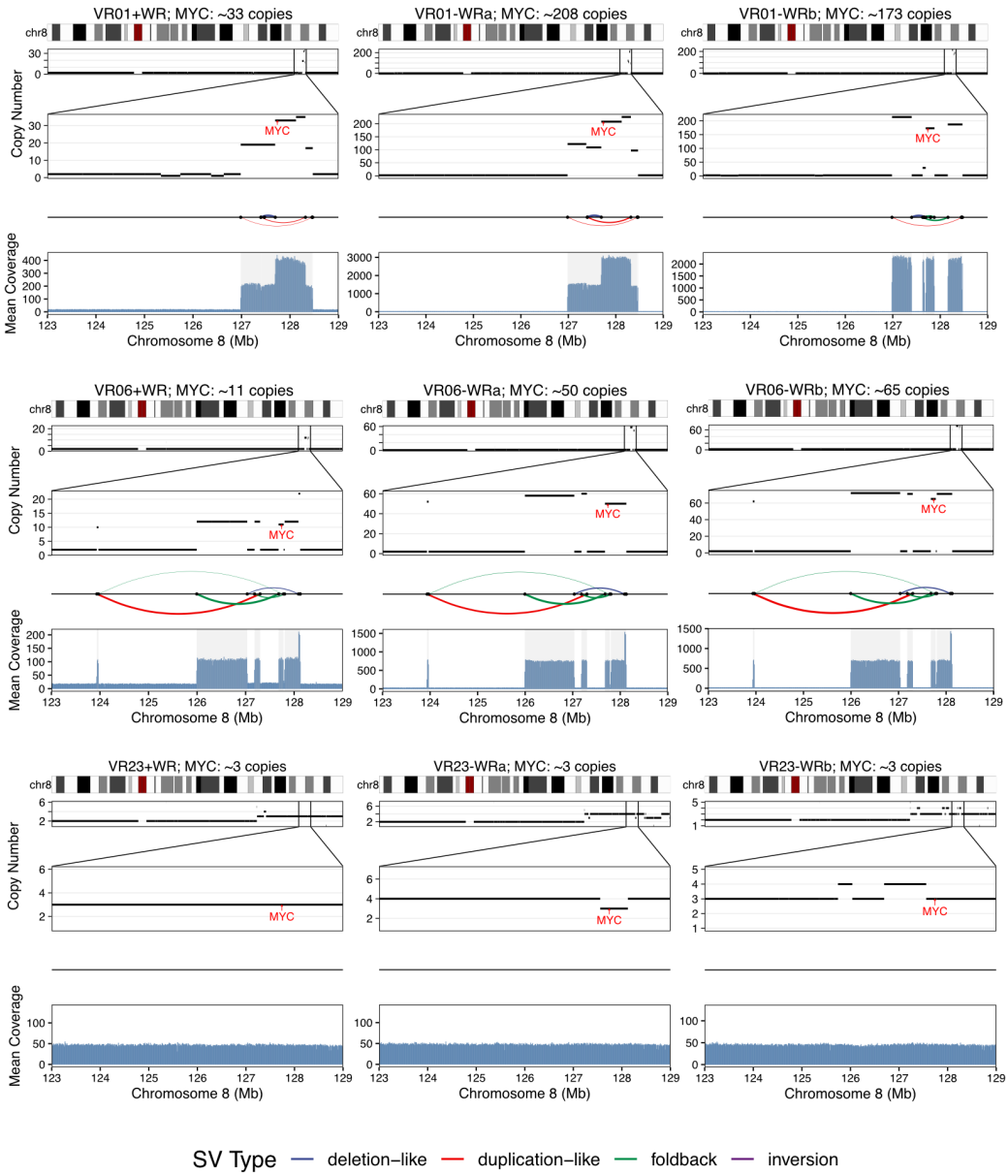


Fig. 4.17 | *MYC*-ecDNA amplification and evolution during adaptation to environmental changes. Genomic overview of copy number calls, ecDNA structural variations, and read coverage shows high similarities between the *MYC*-ecDNA prior adaptation (+WR) and after adaptation process to the human complete media without Wnt3a and R-spondin 1 (-WR) of two *MYC*-ecDNA containing PDOs VR01-O and VR06-O. *MYC* copy levels are massively increased after adaptation in both -WR replicates of both PDOs (-WRa and -WRb) in comparison to the parental line (+WR). In contrary, *MYC* copy number levels are not affected of VR23-O in both -WR replicates.

adaptation. The transcriptomic analysis of VR23-WRa also did not uncover an increase in *MYC* expression, indicating that the adaptation of VR23-O is not influenced by *MYC* activity (Figure 4.18a). Moreover, no additional genomic alterations have been detected in driver genes or WNT pathway genes (Figure 4.18b). This finding is consistent in both VR01-O and VR06-O analyses, indicating that the amplification of the *MYC*-ecDNA and the consequent increase in *MYC* transcription drives the adaptation process in both PDOs.

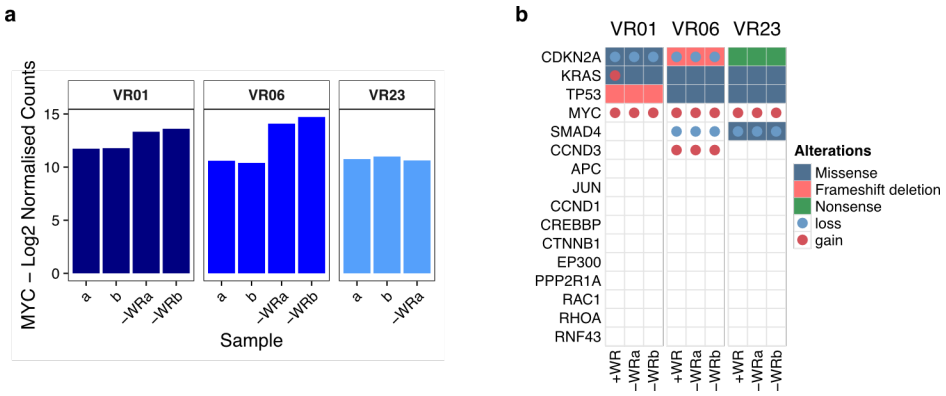


Fig. 4.18 | *MYC* expression in and genomic analysis of parental and adapted (-WR) PDOs. a, *MYC* expression analysis of two biological replicates (a, b) of the parental lines and adapted lines (-WRa and -WRb) of VR01-O, VR06-O, and VR23-O. The gene expression values are normalised and log2-transformed. **b**, Genomic alterations in cancer driver and WNT pathway genes in parental (+WR) and adapted (-WR) PDOs. No additional alteration in adapted lines has been identified.

A thorough examination of the ecDNA structure of VR01-O and VR06-O revealed large structural differences. The *MYC* gene was present entirely on both ecDNAs, but the constituent genomic segments and joining breakpoints differed substantially (Figure 4.17). Interestingly, despite having a lower initial bulk copy number of *MYC*, VR06-O-WR PDOs demonstrated a greater *MYC* transcriptional activity than the VR01-O-WR PDOs, suggesting a diverging effect of *MYC* copy number on the two PDOs (Figure 4.19a).

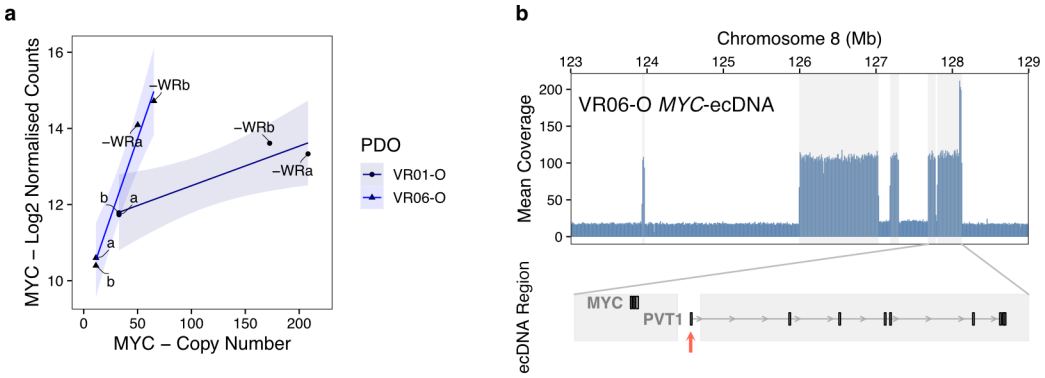


Fig. 4.19 | Genomic and transcriptomic analysis of *MYC* expression in VR01 and VR06. a, Correlation analysis of *MYC* expression and copy number levels in VR01-O and VR06-O parental PDOs (a, b) and -WR adapted PDOs (-WRa, -WRb). **b**, Genomic view of the VR06-O *MYC*-ecDNA segments and the location of *MYC* and *PVT1* regions. The absence of the *PVT1* starting region on the *MYC*-ecDNA is shown with an orange arrow.

PVT1 co-amplification and transcription are crucial factors in *MYC* transcription and tumourigenesis (Tseng et al., 2014). Moreover, it has been shown that enhancer elements in the *PVT1* site can enhance *MYC* transcription when the *PVT1* promoter is absent, implying that the *PVT1* promoter has a tumour suppressive effect (Cho et al., 2018). Therefore, a more

detailed view of the loci was generated. In the VR06-O *MYC*-ecDNA of both parental and the adapted lines, the starting region of *PVT1*, comprising the promoter, is truncated and not located on the ecDNA, suggesting that the tumour suppressive role of the *PVT1* promoter is diminished and the enhancer elements inside the *PVT1* gene are directly promoting *MYC* transcription (Figure 4.19b) (Cho et al., 2018). In contrast, the full gene is present on the *MYC*-ecDNA in VR01-O (Figure 4.5). Therefore, the phenomenon observed in VR06-O may be the reason for the increased *MYC* transcription in VR06-O, despite its lower *MYC* copy number levels compared to VR01-O.

Comparing the parental and adapted *MYC*-ecDNAs in a PDO line, revealed significant similarities in their genomic structure and ecDNA breakpoints. However, VR01-WRb had a divergent structure from the baseline, indicating ecDNA evolution during the adaptation process (Figure 4.17). Despite the similarities in breakpoints between VR01-WRb and its parental line, the modified ecDNA structure of VR01-WRb lacks a distinct genomic locus comprising *TMEM75* and parts of *PVT1* (Figure 4.20a). The absence of *TMEM75* directly impacted the transcription in VR01-WRb. While the genes present on ecDNA showed similar levels of expression in both -WR replicates, the expression of *TMEM75* was entirely absent in VR01-WRb (Figure 4.20b). Therefore, this transcriptomic analysis confirms the genomic structure of the altered *MYC*-ecDNA. It is worth noting that despite a significant fraction of the *PVT1* gene being absent, the expression of *PVT1* was largely unaffected. Furthermore, a slight increase in *MYC* expression is observed in VR01-WRb compared to VR01-WRa, despite the lower *MYC* copy number levels (Figure 4.19a). Therefore, the evolved ecDNA structure might further enhance *MYC* transcription, which is independent of the *MYC* copy number levels.

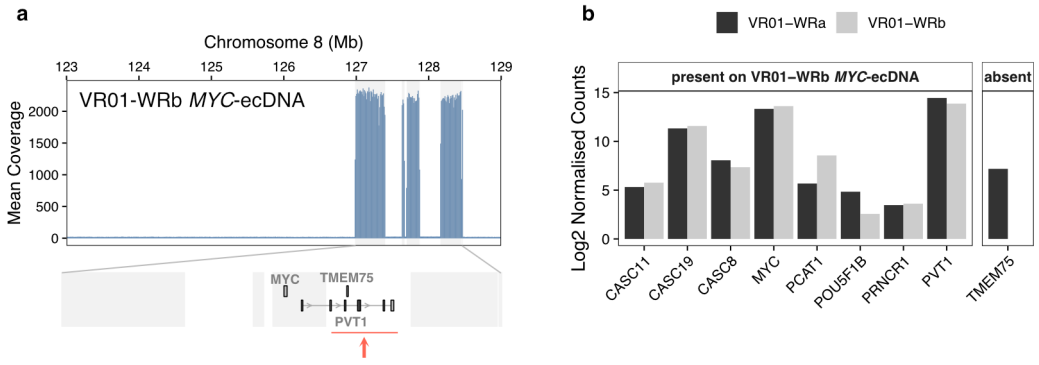


Fig. 4.20 | Evolution of *MYC*-ecDNA in VR01-WRb. a, Genomic structure of the *MYC*-ecDNA identified in VR01-WRb. The ecDNA segments are coloured in grey and the location of the genes *MYC*, *PVT1*, and *TMEM75* are displayed. *TMEM75* and a large proportion of the *PVT1* gene are absent on the ecDNA. **b**, Expression of genes present or absent on VR01-WRa and VR01-WRb *MYC*-ecDNA. *TMEM75* was identified to be present on VR01-WRa ecDNA, but absent on VR01-WRb ecDNA. The gene expression values are normalised and log2-transformed.

In summary, our research sheds light on the dynamic nature of ecDNAs and their potential contribution to cell adaptation and resistance. Specifically, PDOs carrying *MYC*-ecDNA showed significant adaptability to environmental stressors, as evidenced by substantial amplification of *MYC*-ecDNAs and a corresponding surge in *MYC* transcription following WR removal. Conversely, the lack of a *MYC*-ecDNA, a *MYC* copy number increase, or additional

driver alterations in VR23-O suggests the presence of an alternative adaptation mechanism, which could not directly be uncovered by WGS or RNA-seq data analysis. Interestingly, the structural variations within *MYC*-ecDNAs, including the lack of the *PVT1* promoter in VR06-O and an evolved ecDNA structure in VR01-WRb, seem to influence the transcriptional activity of *MYC*. This implies a complicated interplay between the ecDNA-based gene amplification, the intrinsic ecDNA structure, and the ecDNA-based gene-enhancer interactions to facilitate adaptation in challenging environments.

4.10 Circle-seq validates *MYC*-ecDNA

Validating computational analyses is essential for the biological interpretation of the results and for future research. EcDNAs can be detected through the high-throughput WGS method and are typically confirmed by performing fluorescence *in situ* hybridisation (FISH) on metaphase cells, or Circle-seq. Therefore, to confirm the identified *MYC*-ecDNAs, Circle-seq and FISH were carried out on VR01-O and VR06-O. To note, FISH has been performed by the Vincenzo Corbo Lab, University of Verona.

Increased Circle-seq coverage has been found near the *MYC* locus within the ecDNA region detected via WGS analysis in VR01-O (Figure 4.21a). Additionally, only minimal Circle-seq coverage is observed outside the ecDNA region, verifying the circularity and presence of the *MYC*-ecDNA in VR01-O. It also demonstrates that this technique is applicable in PDAC PDOs to validate and detect large ecDNAs. Additionally, FISH on VR01-O also demonstrated the extrachromosomal nature of *MYC* (Figure 4.21b). Interestingly, in VR01-O, multiple *MYC* FISH probes were identified that clustered together, suggesting either ecDNA hubs or multiple *MYC* genes on the same ecDNA. However, based on the WGS data analysis, a clear delineation of the ecDNA structure cannot be made and a potential VR01-O *MYC*-ecDNA, identified by AmpliconArchitect, has only one *MYC* copy (Figure 4.5). Therefore, this could suggest that *MYC*-ecDNA hubs are formed. However, ecDNA hubs are currently highly debated and more research on VR01-O is needed (Zhu et al., 2021; Purshouse et al., 2022).

In contrast, two Circle-seq samples of VR06-O did not exhibit an increase of Circle-seq reads in and around the *MYC* locus, suggesting its absence (Figure 4.21c). It appears that VR06-O did not a *MYC*-ecDNA during DNA extraction, which was carried out at passages 41 and 45. Upon examining *MYC* FISH images revealed that the of VR06-O cells collected during the later passages (passage number > 40) do not contain *MYC*-ecDNAs in contrast to those collected during earlier passages of VR06-O (Figure 4.21b,d). This suggests that the *MYC*-ecDNA might have only be present in the primary tumour and early passages of VR06-O and was lost after continuous PDO passaging. This absence of the *MYC*-ecDNA is also validated by the Circle-seq analysis (Figure 4.21a). Unfortunately, Circle-seq was not carried out on early-passage VR06-O.

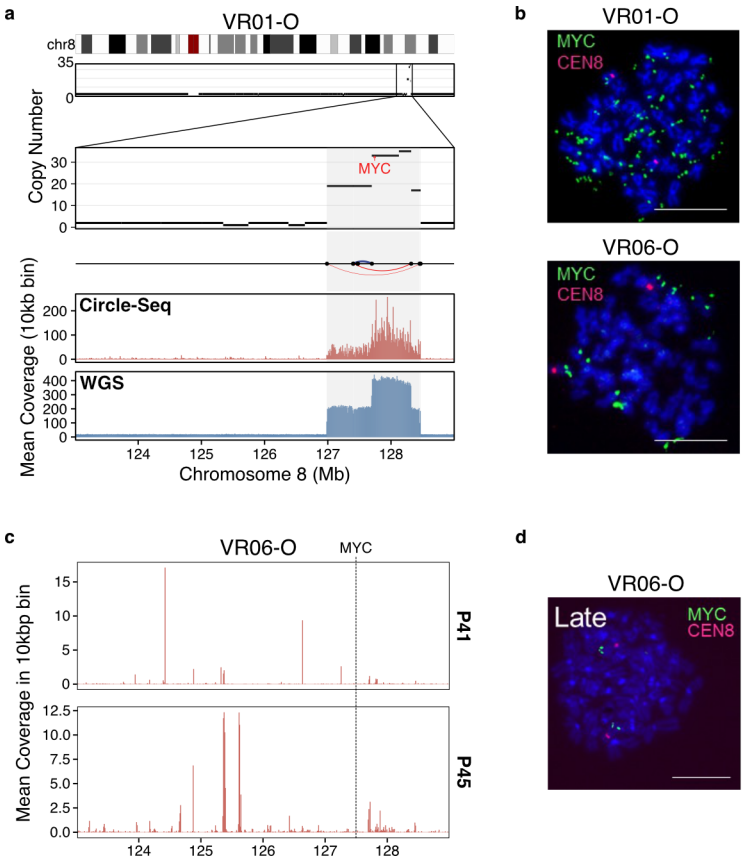


Fig. 4.21 | Circle-seq coverage confirms VR01-O *MYC*-ecDNA. a, Visualisation of copy number levels, ecDNA structural variations and coverage of WGS and Circle-seq reads around the *MYC* locus of the *MYC*-ecDNA containing VR01-O. **b**, Representative FISH images validating the presence of *MYC*-ecDNA in the VR01-O and VR06-O PDO. **c**, Circle-seq coverage of late-passage (P41 & P45) VR06-O. No coverage increase observed in or around *MYC* locus. **d**, Representative FISH image of a late-passage VR06-O (passage number ≥ 40) validating the absence of a *MYC*-ecDNA. **b** & **d** FISH has been performed by the Vincenzo Corbo Lab, University of Verona, Italy.

In conclusion, our validation efforts emphasise on the robustness of high-throughput WGS in detecting ecDNAs, as both Circle-seq and FISH successfully verified the presence of *MYC*-ecDNAs in VR01-O. Furthermore, Circle-seq emerged as a valuable tool for confirming large ecDNAs in PDAC PDOs. Intriguingly, the absence of *MYC*-ecDNAs in later passages of VR06-O suggests a dynamic nature of ecDNAs during PDO passaging. This highlights the importance of temporal monitoring and multimodal validation for ecDNA studies, as their presence or absence may have significant implications in the context of tumour evolution and treatment response.

4.11 EcDNAs are maintained in metastatic PDOs

Our cohort consisted of 41 distinct PDOs, some of which were derived from the same patients but from different tissue types. Biopsies from the primary tumour and lymph node metastasis were taken from two patients, namely VR11 and VR23, and organoids were successfully established. The amplicon analysis revealed the presence of various amplicons across all four PDOs. VR23-O and VR23-LNO harboured one complex and one linear amplicon, respectively. VR11-O and VR11-LNO each contained two circular amplicons and an additional

chromosomal amplicon. The chromosomal amplicon of VR11-LNO was classified as BFB, while the VR11-O amplicon was classified as complex. With the two circular amplicons depicted by VR11-O and VR11-LNO an opportunity arose to perform a comparative analysis on common origin and structure of these ecDNA containing amplicons. The presence of two circular amplicons in VR11-O and VR11-LNO provided the chance to carry out a comparative analysis of the common origin and structure of ecDNA-containing amplicons.

Tab. 4.2 | Amplicon similarity examination of the two circular amplicons identified in both VR11 PDOs. Similarity score and its P values are calculated based on the amplicon regions and breakpoint overlap (Method described in Luebeck et al. (2023)).

VR11-O	VR11-LNO	Similarity Score	P value
Amplicon 1	Amplicon 1	0.735	0.001
Amplicon 3	Amplicon 3	0.999	< 0.001

A significant level of similarity was observed between both ecDNA amplicons of VR11-O and VR11-LNO by performing an amplicon similarity analysis (Table 4.2). The genomic structure of the amplicons also indicates similar copy number levels and genomic breakpoints (Figure 4.22). AmpliconArchitect is unable to fully distinguish the ecDNA structure and produces many possible ecDNA structures that may coexist or combine to create larger ecDNAs. For each amplicon one putative structure is presented alongside its genomic view. Here, large similarities can be observed in the gene and ecDNA composition between the matching VR11-O and VR11-LNO amplicons (Figure 4.22). It appears that the lymph node metastasis preserved the ecDNA amplicon from the initial tumour and the cells carrying these ecDNAs separated from the primary tumour and formed the respective metastasis in the lymph node.

4.12 Prognostic implications of ecDNA presence

A comprehensive pan-cancer study has previously established that patients diagnosed with tumours that contain ecDNAs have, on average, significantly shorter survival periods compared to patients with tumours containing chromosomal amplicons or no amplicons at all Kim et al. (2020). This analysis included close to 30 cancer types, including tumours of the pancreas. Although a general association between ecDNA-positivity and poor outcome was established, survival times were not compared for individual cancer types. Therefore, I examined the ecDNA status and survival time of 56 and 71 patients, respectively, from the ICGC PACA-AU and PACA-AU cohort (Kim et al., 2020).

Firstly, the two PDAC cohorts were analysed separately, showing contradictory results. The ecDNA+ patients who were part of the ICGC PACA-AU cohort displayed a significantly shorter survival time, in comparison to their ecDNA– counterparts ($P = 0.022$). Conversely, no significance was found amongst patients from the PACA-CA cohort ($P = 0.68$). To increase statistical power, both cohorts were also collectively analysed, which resulted in no significant association ($P = 0.067$) between the presence of ecDNAs and a worse outcome.

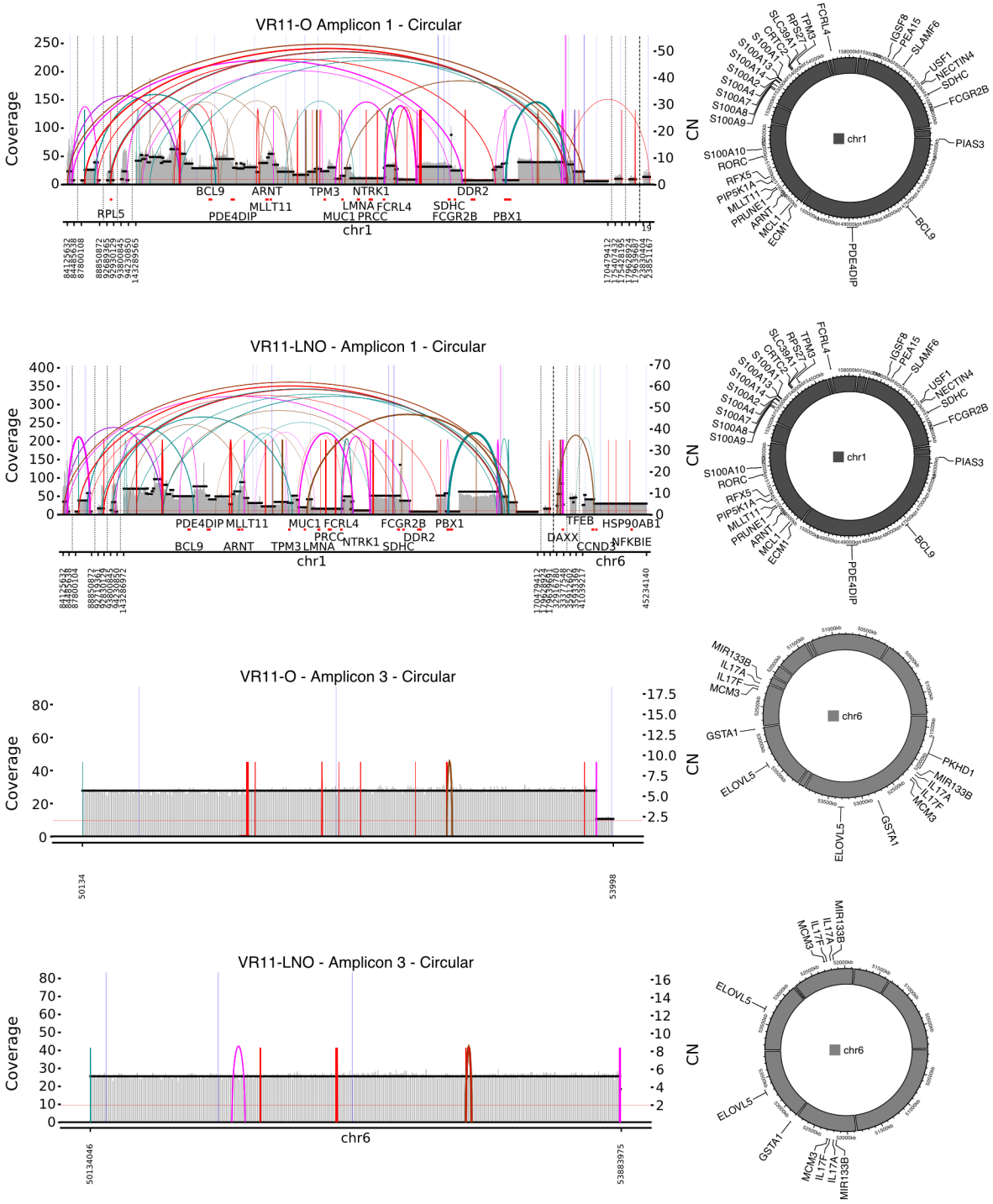


Fig. 4.22 | Large similarity between ecDNA amplicons of VR11-O and VR11-LNO. Amplicon genomic overview shows copy number, coverage, structural variations and oncogene content. The genomic overview is generated by AmpliconArchitect (Deshpande et al., 2019). One putative ecDNA structures of each amplicon is displayed next to the amplicon genomic overview. Cancer driver genes located on the putative ecDNAs are labelled.

These uncertain findings fail to validate any association between a worse outcome and ecDNA in PDAC. However, it should be noted that the total number of patients with ecDNAs may be insufficient to provide a comprehensive evaluation. Therefore, additional investigations involving larger sample sizes are necessary for further analysis.

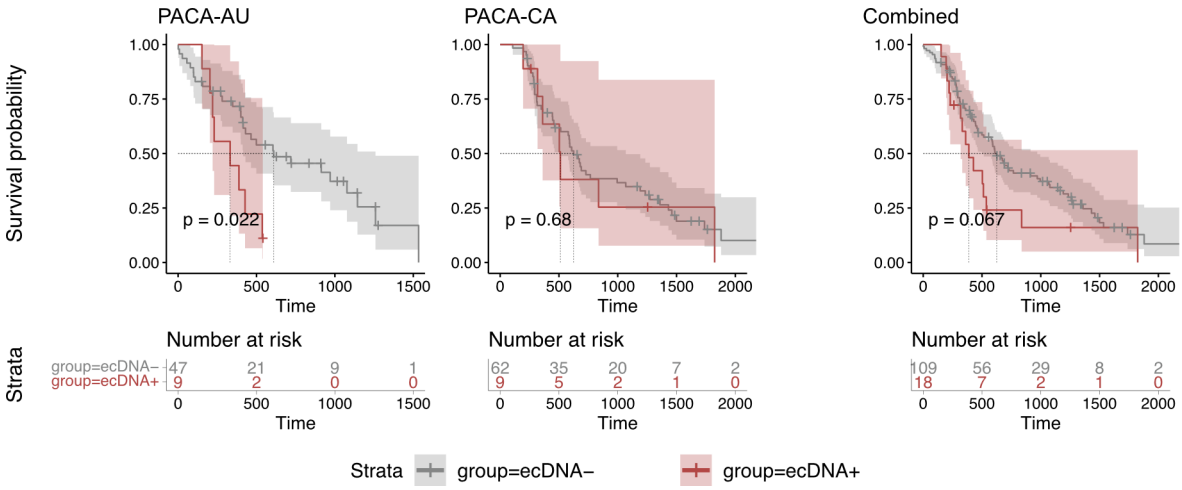


Fig. 4.23 | Kaplan-Meyer survival analysis of ecDNA+ and ecDNA– PDAC tumours. The ICGC primary tumours from PACA-AU and PACA-CA cohorts were individually and collectively analysed for their survival time. The survival curves were compared using the log-rank test.

4.13 Investigating ecDNAs in PDAC cell lines

Although PDOs are recognised as superior in recapitulating phenotypic and genetic features compared to tumour-derived cell lines, they present logical challenges such as higher maintenance costs and an increased cultivation difficulty. In contrast, cell lines offer ease of maintenance and scalability, which might be valuable for ecDNA research in PDAC (Drost & Clevers, 2018).

Our WGS analysis demonstrated that ecDNAs are widespread in PDOs and correspond with ecDNAs present in matching primary tumours. Many studies on ecDNA focus on known ecDNA+ cell lines (Turner et al., 2017; Hung et al., 2021; Wu et al., 2019). However, to my knowledge, there have been no studies on ecDNA+ PDAC cell lines, which are both readily available and frequently researched. Therefore, I aimed to characterise frequently utilised PDAC cell lines using accessible WGS data from the CCLE. Additionally, three cell lines from three PDOs, of the Verona HCMI cohort, were established. Among the three PDOs, two were ecDNA+. These newly established cell lines were also subjected to WGS and amplicon analysis. Thus, this allows for a direct comparison of the ecDNA landscape within the PDOs and the corresponding PDO-derived cell lines.

4.13.1 Commonly used PDAC cell lines lack ecDNAs

The CCLE encompasses ten PDAC cell lines that underwent WGS as part of the project number PRJNA523380 (Table A.1). This dataset was analysed using the nf-core/circdna pipeline, with the tools AmpliconArchitect (Deshpande et al., 2019) and AmpliconClassifier (Luebeck et al., 2023), to identify amplicons, including ecDNAs, and their corresponding amplicon class. All amplicons and their respective class identified in the ten PDAC cell lines are detailed in the Table 4.3.

Unfortunately, the analysis showed that none of the 10 PDAC cell lines contain ecDNAs

Tab. 4.3 | PDAC CCLE cell lines and their respective number of amplicons per amplicon class identified.

Cell Line ID	Amplicon Class	Count
Capan-1	Linear	3
DAN-G	Linear	14
HPAC	–	–
MIA PaCa-2	Linear	1
PA-TU-8988T	Linear	3
Panc 03.27	Linear	1
Panc 10.05	Linear	1
PANC-1	Linear	1
SUIT-2	Linear	3
SW 1990	Linear	3

that can be detected by WGS data analysis. In particular, only linear amplicons were detected, while other chromosomal amplifications such as Complex or BFB amplicons were not detected.

These findings illustrate that the widely used PDAC cell lines do not possess ecDNAs, suggesting their unsuitability for ecDNA research. It is worth noting that while finalising the thesis a repository was published describing the CCLE amplicon landscape (https://ampliconrepository.org/). This repository described the identification of ecDNAs in two PDAC cell lines, SUIT-2 and DAN-G, when using the same ecDNA detection tools. Consequently, the difference between the two methodologies requires further assessment. Furthermore, this might also reveal the presence of ecDNA+ PDAC cell lines commonly used in research. However, it first needs to be established why the same tools classify the amplicons differently.

4.13.2 Retention of ecDNAs in PDO-derived cell lines

To examine the amplicon landscape of the PDO-derived cell lines (VR02-2D, VR06-2D, and VR23-2D), WGS was performed and the amplicons and their classes was determined using the nf-core/circdna pipeline in a similar manner as described for the WGS PDO analysis.

To explore the persistence of ecDNAs in the PDO-derived cell lines, I conducted a comparison of the ecDNA regions identified through WGS analysis of both the PDOs and their corresponding cell lines (Figure 4.24). The results demonstrated the presence of all ecDNA regions identified in the PDO as well as additional ecDNA regions exclusively detected in 2D, indicating a minor shift in the ecDNA profiles between the PDO and the PDO-derived cell line. Importantly, concordant regions were identified in VR02 and VR06, both with existing ecDNAs in their PDOs. It remains unclear whether these potential ecDNAs still exist in the PDO but are undetected because of the copy number thresholds implemented for amplicon detection. Importantly, the two ecDNAs carrying the PDAC drivers, namely *ERBB2* and *MYC*, that were identified in VR06-O are also found in VR06-2D.

Considering the potential genomic rearrangements and evolution of ecDNAs during environmental changes, an in-depth analysis of the *MYC*-ecDNA of VR06-O and VR06-2D was performed (Shoshani et al., 2021). The investigation revealed that VR06-2D conserved

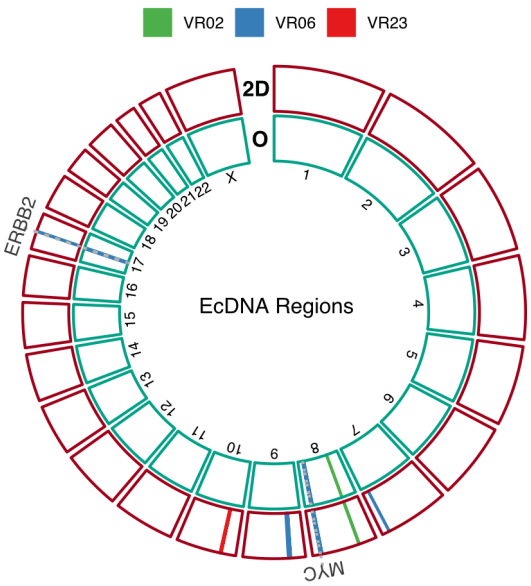


Fig. 4.24 | EcDNA landscape of PDOs with matching PDO-derived cell lines. Cell lines were derived from three PDOs, namely VR02-O, VR06-O, and VR23-O. Circos plot showing the identified ecDNA regions in either the the PDO (O) or the PDO-derived cell line (2D). No ecDNA was detected for VR23-O.

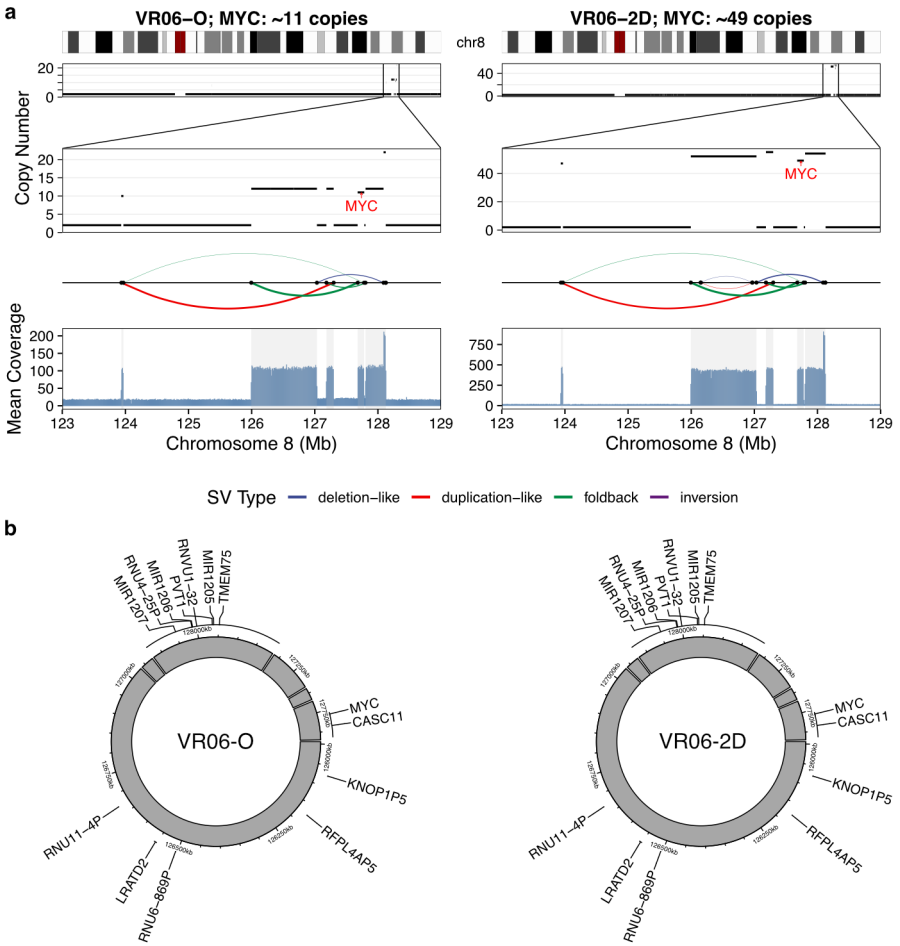


Fig. 4.25 | MYC ecDNA is retained in PDO-derived cell line. a, Copy number and coverage view of the *MYC* locus. EcDNA forming SVs are displayed in the middle section. The mean coverage is calculated for 10 kb bins. **b**, Putative structure of the MYC-ecDNA from VR06-O and the VR06-O-derived cell line VR06-2D.

MYC-ecDNA during and after establishment, maintaining a similar structure and breakpoints as its parent PDO (Figure 4.25a,b). Notably, The *MYC* copy number levels experienced a significant surge from 11 to 49 copies, indicating an extensive amplification of the *MYC*-bearing

ecDNA in VR06-2D. This suggests a potent selection pressure favouring high *MYC*-ecDNA amplification in VR06-2D. Copy number calling from bulk WGS presents the mean copy number of all sequenced cells. Given that PDOs may exhibit a broader cell heterogeneity compared to cell lines, copy number levels can be influenced and result in decreased amplification levels (Drost & Clevers, 2018). However, PDO heterogeneity was not evaluated and it is unclear whether this has an effect or the *MYC* copy number increase is driven by the altered cultivation conditions.

In summary, the PDO-derived cell lines VR02-2D and VR06-2D have retained ecDNAs from their respective PDOs. As PDOs display considerable heterogeneity and require substantial time and resources, generating cell lines from PDOs provides an attractive avenue for studying ecDNAs while also reducing costs and workload (Drost & Clevers, 2018; D'Agosto et al., 2019).

4.14 Discussion

The amplification of oncogenes situated on ecDNAs is a crucial factor promoting tumour evolution, allowing the tumour to adapt to challenging events, including drug treatment. The investigation and characterisation of ecDNAs can uncover possible resistance mechanisms, providing opportunities for a better personalised therapy (Nathanson et al., 2014; Wu et al., 2022a). Furthermore, ecDNA has been associated with unfavourable patient outcomes and tumour progression, highlighting the need to enhance our understanding of these entities (Kim et al., 2020; Luebeck et al., 2023). As research on the relationship between ecDNAs and PDAC is currently lacking, I conducted a comprehensive analysis of the genomic and transcriptomic profile of ecDNA+ tumours and PDOs. The analysis showed that ecDNAs are widespread in PDAC, and the ecDNA presence is linked to genomic instability, *TP53* inactivation, and a Squamous signature. Moreover, we discovered indications that ecDNA crucially contributes to adaptation mechanisms in response to environmental pressures. As PDAC maintains one of the poorest survival rates amongst cancer patients, obtaining a comprehensive understanding of its underlying genomics is crucial (Neoptolemos et al., 2023; Siegel et al., 2023).

An initial study identified that PDAC exhibited low levels of ecDNA, with a proportion of approximately 14%, indicating its affiliation with cancer types with a low ecDNA incidence rate (Kim et al., 2020). Our PDO analysis found that ecDNAs can occur in PDAC at an increased rate of nearly 30%. Consequently, PDAC could be found among the cancer types with elevated ecDNA frequency (Kim et al., 2020). The ecDNA landscape in PDAC comprises common PDAC oncogenes, including *MYC*, *CCND3*, or *ERBB2*. This investigation proposes that often amplified drivers in PDAC may potentially reside on ecDNAs instead of the chromosomes. The implications of this discovery could significantly impact the understanding of the PDAC tumour biology and patient outcomes (Waddell et al., 2015; Kim et al., 2020). However, several ecDNAs identified in the PDOs did not exhibit any overlap with PDAC-specific oncogenes. This suggests that some ecDNAs may have a separate role from oncogene

amplification (Wu et al., 2022a).

Accurately replicating the primary disease is essential for experimental work (Tiriac et al., 2018; Drost & Clevers, 2018). Through our use of PDOs and matched primary tissue, we have shown that PDOs retain ecDNAs from primary tumours and can replicate the ecDNA landscape. This suggests that utilising model systems is a feasible method to replicate the ecDNA spectrum of primary PDAC. Previous research by deCarvalho et al. (2018) has also revealed comparable results in neurospheres originated from glioblastomas. While there were some variances between the ecDNA profiles of the PDOs and their related primary tumours, it is unclear whether these outcomes were influenced by factors such as sample collection and data analysis methods, given the ecDNA detection AmpliconArchitect's stringent copy number cutoff for amplicon determination. It is plausible that ecDNA detection is more attainable in purely neoplastic organoids as opposed to primary tumours, wherein a significant fraction of stromal cells are present (Deshpande et al., 2019; Kim et al., 2020; Tiriac et al., 2018; Chu et al., 2007).

Our findings provide compelling evidence of a correlation between ecDNA+ PDAC, *TP53* alterations, and genomic instability. This study provides support for research on ecDNAs in Barrett's oesophagus, as it linked ecDNA+ with altered *TP53* and polyploidy. The research proposed that inactivated *TP53*, combined with genomic instability, facilitates ecDNA formation (Luebeck et al., 2023). Therefore, it is hypothesised that unstable PDAC genomes are more prone to harbouring ecDNAs, by increasing the chance of ecDNA formation, which then potentially can indicate adaptability to drug treatment or environmental changes (Shoshani et al., 2021).

While a study spanning multiple types of cancer has established the association between the presence of ecDNA and poor outcomes, such correlation has not yet been confirmed in our analysis in the case of PDAC specifically (Kim et al., 2020). While patient outcomes play a significant role in determining the relevance of genomic traits, they cannot be the sole determinant. Various subtypes have been identified in PDAC that define cancer phenotypes and drug tolerance (Bailey et al., 2016; Collisson et al., 2019; Raghavan et al., 2021). Analysis of primary PDAC tumours has identified a correlation between the presence of ecDNA and gene signatures of the Squamous subtype. This subtype is linked with *TP53* alterations and activation of the MYC pathway, associations that we also identified in ecDNA+ tumours (Bailey et al., 2016). Squamous tumours are characterised by high proliferation rates and generally have a poor prognosis for patients. These features were also previously identified in ecDNA+ cancers (Bailey et al., 2016; Kim et al., 2020). It is currently unclear if ecDNAs and their related genes play a role in the aggressiveness of PDAC, but it is likely that the biology of Squamous tumours favours ecDNA formation.

Targeted therapy resistance presents a challenge for individuals with cancer (Vogelstein et al., 2013). Recent studies have recognised ecDNAs as a crucial mediator of treatment resistance

in cancer (Nathanson et al., 2014; Lange et al., 2022). In our analysis, when challenging *MYC*-ecDNA PDOs by removing WR, a significant amplification of *MYC* and consequent activation of *MYC* gene transcription was observed. While the massive amplification influenced *MYC* transcription in -WR PDOs, *MYC* levels also appear to rely on other factors, such as the general ecDNA structure. In one PDO, the *MYC*-ecDNA contained a truncated *PVT1*, which could have potentially boosted *MYC* transcription by enabling contact between enhancers within *PVT1* and the *MYC* promoter (Cho et al., 2018). Moreover, a structural evolution of an ecDNA has been observed in VR01-WRb that could have been driven by the applied selection pressure in the -WR conditions. The underlying mechanism responsible for the ecDNA evolution is still uncertain. Shoshani et al. (2021) suggested that chromothripsis initiates the evolution of ecDNA, but this hypothesis was not evaluated, and only slight modifications in the evolved ecDNA were detected. Our research highlights the significance of ecDNAs for PDAC cells subjected to selection pressure. Directly targeting ecDNAs may be necessary for treating cells that harbour such ecDNAs to prevent their adaptation.

The potential to improve patient outcomes through targeted therapy is widely recognised, and genomic and transcriptomic biomarkers are employed to define the necessary therapeutic approaches for better patient response (Neoptolemos et al., 2023). EcDNA may act as a biomarker for therapeutic tolerance of tumours as the abundance of ecDNA can rapidly fluctuate due to its random segregation during the cell cycle and the selection pressure applied to the cells (Nathanson et al., 2014; Lange et al., 2022). Our observations in PDAC PDOs showed that ecDNAs are not synthesised *de novo* under selection pressure and the absence of an adaptation mechanism led to rapid cell death. Although *de novo* formation of ecDNAs has been observed in some cases, it was noted that cells can adapt more easily when the relevant genes are already present on the ecDNAs (Singer et al., 2000). Thus, the use of ecDNAs as a biomarker must be viewed in the context of the genes that they contain. Acquisition of resistances is plausible when ecDNAs already bear resistance genes.

In conclusion, I conducted a primary comprehensive examination of ecDNAs in PDAC. EcDNAs, a common source of high-level amplifications in PDAC, comprise numerous PDAC driver genes that can potentially alter tumour biology. The complete analysis revealed that genomic instability, *TP53* mutations, and Squamous transcriptomic profiles are associated with ecDNA+ tumours. We were also able to demonstrate that applying specific selection pressures to *MYC*-ecDNAs highlights their role in driving adaptation and suggests the potential of ecDNAs to facilitate the tumour's adaptation to challenging environmental pressures.

Chapter 5

Investigating the eccDNA landscape in PDAC

> I cannot think of a single field in biology or medicine in which we can claim genuine understanding, and it seems to me the more we learn about living creatures, especially ourselves, the stranger life becomes.
>
> *Lewis Thomas*

EcDNAs can be observed using microscopy techniques as well as diverse sequencing methods (Turner et al., 2017; Yi et al., 2022; Kumar et al., 2020). However, detecting smaller and non-amplified eccDNAs is more difficult. Recently, Circle-seq has been used in eccDNA studies as a technique for enriching eccDNAs and concurrently removing linear DNAs from extracted DNA before sequencing. These studies have characterised the eccDNA landscape in human somatic and cancer cells, defining general eccDNA characteristics (Møller, 2020; Møller et al., 2018a; Koche et al., 2020; Wang et al., 2021). While the broad eccDNA characterisations have had a crucial impact on our understanding of these features, the roles of eccDNAs and characteristics in individual cancer types remain unclear. EccDNAs have yet to be studied in PDAC and a comprehensive overview of the eccDNAs identified in this disease is lacking.

Therefore, this study aims to investigate the prevalence of eccDNAs in PDAC and their genomic characteristics utilising sequencing data generated through the Circle-seq method. The analysis utilises several newly generated datasets from PDAC model systems, including patient-derived cell lines (PDCLs) and patient-derived organoids (PDOs). Computational analysis is undertaken to identify high-quality eccDNAs from sequencing data in order to determine the eccDNA landscape and its genomic characteristics in PDAC. The resulting data allows for a comprehensive comparison to other studies and cancer types, which may reveal

PDAC-related and common eccDNA characteristics.

5.1 Establishing Circle-seq

Next-generation sequencing (NGS) is shaping cancer genomics and transcriptomics. Due to advances in technology, throughput is massively increased and cost decreased in recent years. Multiple protocols have been established to sequence and characterise the eccDNA landscape in eukaryotic cells. These include DNA extraction, digestion of linear/chromosomal DNA with an ATP-dependent circular DNA-safe exonuclease, and amplification of the remaining circular DNA by rolling-circle amplification (RCA) using the bacteriophage phi29 DNA polymerase. Additionally, some protocols propose the removal of mitochondrial DNA by using additional restriction enzymes, linearising the mitochondrial DNA. Linear mitochondrial DNA will then also be digested by a circular DNA-safe exonuclease. To ensure high removal efficiency of linear DNA, exonuclease digestion is usually performed over multiple days and additional exonuclease, ATP, and buffer is added throughout the period Møller et al. (2015), Møller (2020) and Koche et al. (2020).

Møller et al. (2015) adapted previous protocols for eukaryote plasmid DNA isolation to isolate and sequence eccDNA from yeast. To specifically isolate eccDNA, a column-based plasmid isolation kit was used. Henssen et al. (2019a), on the other hand, used the Qiagen high-molecular weight DNA extraction kit, to ensure extraction of large eccDNAs. The HMW DNA extraction kit is optimised to isolate also DNA with up-to a few hundred kbp lengths. Therefore, In our study, the DNA extraction was performed by the MagAttract HMW DNA Kit (Qiagen) to retain large circular DNAs in our samples.

To create additional DNA ends for exonuclease digestion, Møller et al. (2015) treated the purified DNA with the rare-cutting endonuclease NotI. Henssen et al. (2019b) suggest the usage of of the rare-cutting endonuclease MssI/PmeI to linearise mitochondrial DNA. The use of endonucleases improves the linear DNA removal by exonuclease digestion, however, also a rarely cutting endonuclease can fragment circular DNA (Koche et al., 2020). Large eccDNAs, which have the potential to contain and amplify cancer-specific oncogenes, have a higher chance of containing a restriction enzyme cut site. This would lead to endonuclease cleavage of the eccDNA, generating a linearised DNA, which is further digested in the subsequent exonuclease step. Therefore, some eccDNAs could not be identified by Circle-seq, when including an endonuclease step. Based on this, I decided to not not digest the DNA using restriction enzymes and use the unprocessed extracted DNA for exonuclease digestion.

The Circle-seq procedure in detail: Linear DNA removal was performed as described by Henssen et al. (2019a) for 5-7 consecutive days, adding 20 Units of DNA exonuclease enzyme, 4 µL ATP (25 mM), and reaction buffer. Linear DNA degradation was then confirmed using quantitative PCR and primers amplifying the mitochondrial gene *MT-CO1* and the chromosomal gene *HBB*(Figure 5.1 b). To validate the primer bindings and the linear DNA

removal, agarose gel electrophoresis confirmed the absence of chromosomal DNA after the exonuclease digestion and the presence of circular DNA (Figure 5.1 a). Despite the initial study by Koche et al. (2020) which used the Henssen et al. (2019a) protocol and reported a 10^{10}-fold removal of linear DNA, low amounts of linear/chromosomal DNA still remained in most samples. To achieve a high enrichment of circular DNA in our samples, samples were further processed if a 200 fold-change (FC) reduction of linear DNA compared to circular DNA was achieved. The FC decrease of linear DNA in a sample was calculated as described in Methods Section 2.10.2.

If the FC threshold was not achieved during the 5-day DNase digestion, the process was repeated or 2 additional days of DNase digestion were added following the same procedure as the first five days.

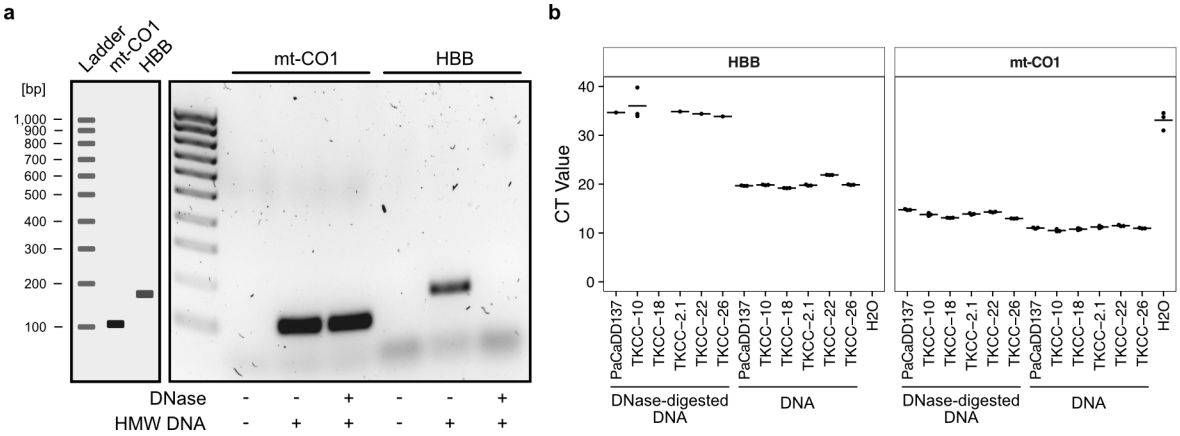


Fig. 5.1 | Circular DNA-safe DNase greatly reduces chromosomal DNA content. a, Agarose gel with PCR fragments amplified using the *MT-CO1* and HBB primer pairs before and after DNase digestion of TKCC-22 HMW DNA. The schema on the left-hand side depicts the ladder sizes and the expected band heights for the *MT-CO1* (114 bp) and *HBB* (173 bp) fragment. **b**, Quantitative PCR (qPCR) results of DNA and DNase-digested DNA of six PDCLs with linear DNA control *HBB*, and circular DNA control *MT-CO1*. **a & b**, The HMW DNA was treated for five days with a circular DNA-safe DNase. Each sample was run in three technical replicates. The bar represents the mean of the three replicates. H2O was used in a no template control.

After removing linear DNA successfully, RCA was carried out on the remaining DNA. Moreover, a few samples were utilised to evaluate eccDNA enrichment after RCA (Figure 5.2). RCA resulted in lower CT values for all samples, confirming the amplification of circular DNA. Furthermore, two out of three samples showed a high *HBB* CT value or *HBB* was not identified, indicating that the linear DNA was not amplified significantly during the RCA process. One sample, PaCaDD137, exhibited an elevated CT value for *HBB* subsequent to RCA, and therefore, was discontinued. Overall, a significant degree of variability was observed in the removal of linear DNA and amplification of circular DNA (data not shown). Samples that demonstrated unusual outcomes, for example those retaining excessive amounts of linear DNA, were discontinued and the Circle-seq procedure was repeated.

Around 500 ng of the purified, quality-controlled, and circular DNA-enriched samples were fragmented by the M220 Focused-ultrasonicator (Covaris). The final libraries were designed for paired-end sequencing, totalling 300 cycles per read and producing reads with a 150 bp length. To optimise read information, a mean fragment size of 350 to 450 bp

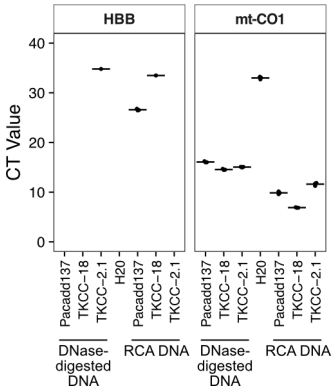


Fig. 5.2 | Circular DNA is amplified after rolling-circle amplification. Efficacy of rolling-circle amplification was tested on DNase-digested DNA of three PDCLs. Quantitative PCR (qPCR) was performed on 5-day DNase-treated DNA prior (DNase-digested DNA) and post rolling-circle amplification (RCA DNA) with linear DNA control HBB, and circular DNA control *MT-CO1*. Each sample was run in three technical replicates. The bar represents the mean of the three replicates. H2O was used in a no template control

was determined. The shearing time was adjusted for the desired fragment time using the microTUBE-15 AFA Beads Screw-Cap and the M220 sonicator. A uniform fragmentation to between 350 and 450 bp was attained through a fragmentation period of approximately 46 s in our initial Circle-seq run (Figure 5.3).

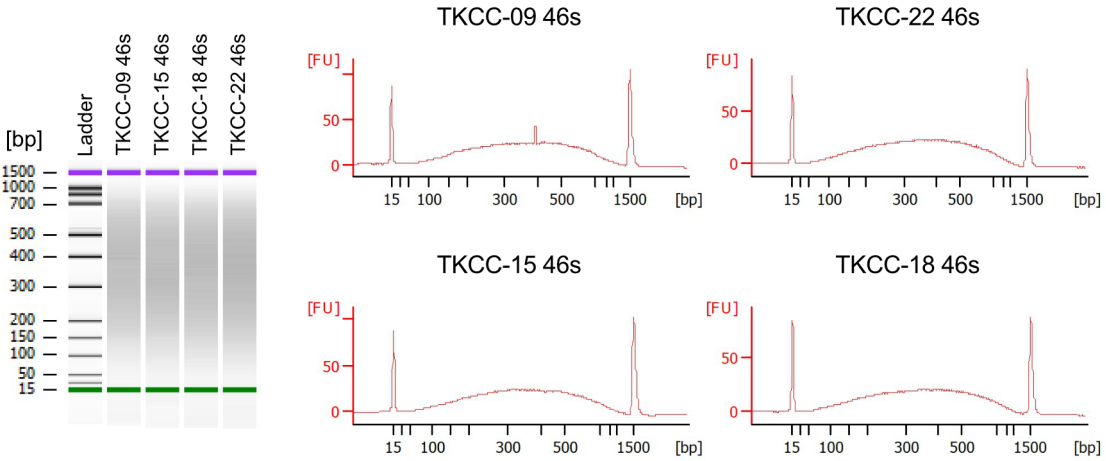


Fig. 5.3 | Bioanalyzer verifies DNA fragmentation size. Agilent DNA 1000 Bioanalyzer (Agilent Technologies) run shows distribution of DNA fragments after a shearing time of 46 s with the M220 sonicator (Covaris) and the microTUBE-15 AFA Beads Screw-Cap (Covaris). Fragmentation test was performed with four PDCL DNA samples enriched for circular DNA. Fragmentation time was adjusted to have increased DNA fragment quantity between 300 and 500 bp and a mean peak size between 350 and 450 bp. The Bioanalyzer output was generated using the 2100 Expert software (Agilent Technologies, version B.02.09.SI725).

Tab. 5.1 | Mean fragment sizes after fragmentation of circular DNA enriched samples. Fragmentation was performed for 46 s with the M220 sonicator (Covaris) and the microTUBE-15 AFA Beads Screw-Cap (Covaris).

PDCL	Mean Fragment Size in bp (50 - 950 bp Region)
TKCC-09	380
TKCC-15	371
TKCC-18	375
TKCC-22	375

After successful fragmentation, the fragmented DNA was prepared for sequencing using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (New England Biolabs) and

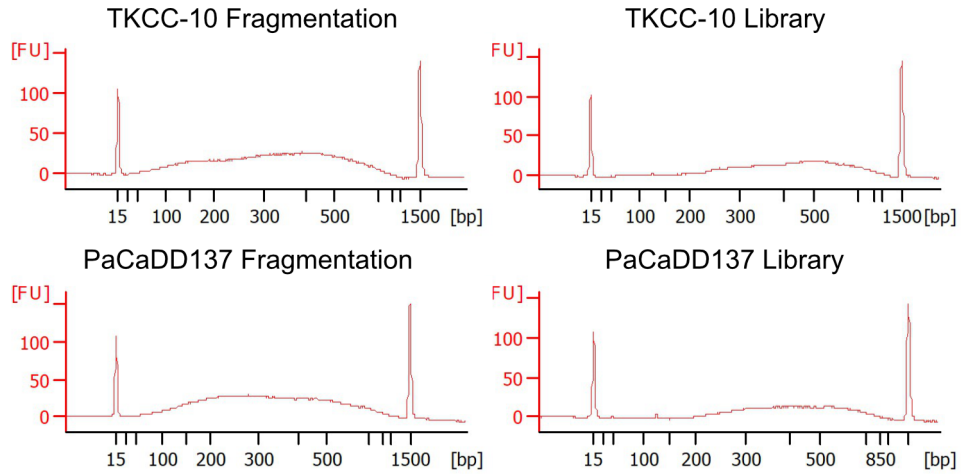


Fig. 5.4 | Validation of Circle-seq library preparation. Library preparation success was validated using the 2100 Bioanalyzer and the Agilent DNA 1000 Kit (Agilent Technologies). Fragmentation of circular DNA-enriched DNA was performed using the M220 sonicator (Covaris) with the microTUBE-15 AFA Beads Screw-Cap (Covaris). The fragmentation time was 46 s for each sample. 1 µL of sample was loaded onto the Bioanalyzer. Library preparation success was evaluated based on DNA concentration and a shift of around 100-150 bp to the original fragment size due to adapter and index addition.

the appropriate library preparation protocol. Size selection was not performed on the final libraries due to the large amount of DNA lost in the process (Figure B.1). The final libraries were PCR amplified for a total of 4 cycles and quality checked on the Bioanalyzer. A shift of approximately 100-150 bp from the original fragmentation size to the final library size was expected due to the addition of sequencing adapters and indexes. Two samples were tested before and after library preparation and validated the success of the library preparation (Figure 5.4). All other libraries were quality checked after completion of library preparation and showed similar fragment sizes and expected distribution (Figure B.2). In summary, the original circle-seq protocols from Henssen et al. (2019a), Møller et al. (2015) or Møller (2020) were carefully reviewed to adapt the Henssen et al. (2019a) to our samples and resources. Overall, the circle-seq protocol achieved a high enrichment of circular DNA in our samples and generated good quality libraries suitable for NGS.

5.2 Circle-seq on 8 PDAC PDCLs

The described protocol adaptation resulted in high quality libraries suitable for NGS. The adapted protocol was subsequently utilised to generate libraries from a total of eight PDAC PDCLs (PaCaDD137, TKCC-2.1, TKCC-09, TKCC-10, TKCC-15, TKCC-18, TKCC-22, TKCC-26). Briefly, HMW DNA was extracted and digested with plasmid-safe DNase for five days. The remaining DNA was then amplified using RCA. Approximately 500 ng of amplified DNA was sheared to a fragment size of approximately 450 bp, which was then used for library preparation and NGS.

Quality controls were performed after DNase digestion and library preparation to ensure the quality of the DNA. Linear DNA removal was verified by qPCR (Table 5.2).

The final libraries containing DNA fragments enriched for eccDNA were sequenced. Qual-

Tab. 5.2 | Fold change decrease of linear DNA compared to circular DNA after DNase digestion of HMW DNA of 8 PDAC PDCLs.

PDCL	Passage	Linear DNA Decrease (FC)
PaCaDD137	31	2418.7
TKCC-2.1	45	1082.4
TKCC-09	40	Linear DNA not detected
TKCC-10	25	7696.6
TKCC-15	36	Linear DNA not detected
TKCC-18	28	Linear DNA not detected
TKCC-22	34	2435.5
TKCC-26	47	4067.7

ity control was performed using FastQC for each sample after sequencing. The nf-core/circdna pipeline was then used to align and trim the sequencing data. Within nf-core/circdna Circle-Map Realign or Unicycler and Minimap2 were utilised used to call eccDNAs. The ultimate outcome includes tables with information on eccDNA and alignment data of the reads (Extended Data Table 5). To validate the sequencing data's quality and the enrichment of eccDNAs, I utilised the Integrative Genomics Viewer (IGV), a tool that visualises genomics data (Thorvaldsdóttir, Robinson & Mesirov, 2013).

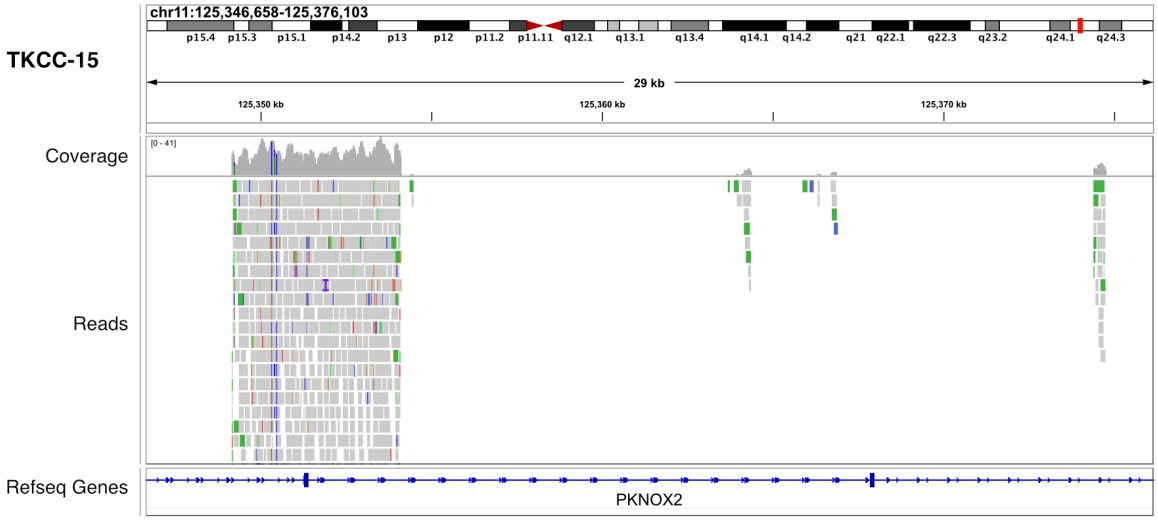


Fig. 5.5 | Representative genomic overview of mapped Circle-seq reads. The figure represents a representative overview of TKCC-15 Circle-seq reads mapped to the GRCh38 reference genome. The view displays two eccDNA at the start and the end of the view with discordant reads (green) representing the eccDNA fusion. The middle part is mostly empty showcasing most chromosomal DNA was removed prior sequencing.

A genomic representation of TKCC-15 Circle-seq reads has been presented in Figure 5.5, showing the significant removal of chromosomal/linear DNA and the abundance of eccDNA reads. Nevertheless, filtering thresholds are required to recognise high-quality eccDNAs and differentiate them from tandem duplications, as some background reads are still being identified.

5.3 Retention of high quality eccDNAs

The identification of putative eccDNA junctions from short-read sequencing data is usually necessary for detecting eccDNAs accurately (Møller et al., 2015; Prada-Luengo et al., 2019; Kumar et al., 2020). Here, I employed Circle-Map Realign to identify these eccDNA junctions from our Circle-seq datasets. EccDNA junctions are typically detected using data from reads that map discordantly and suggest the presence of circular DNA. However, such a signal can also be produced by tandem duplications, which are not distinguishable by eccDNA detection tools. To address this issue, a stringent filtering procedure was implemented, which aimed to exclude eccDNA junctions that lack sufficient quality and certainty (Figure 5.6 and Methods Section 2.14). This included, quality, coverage, and ENCODE blacklist region filtering.

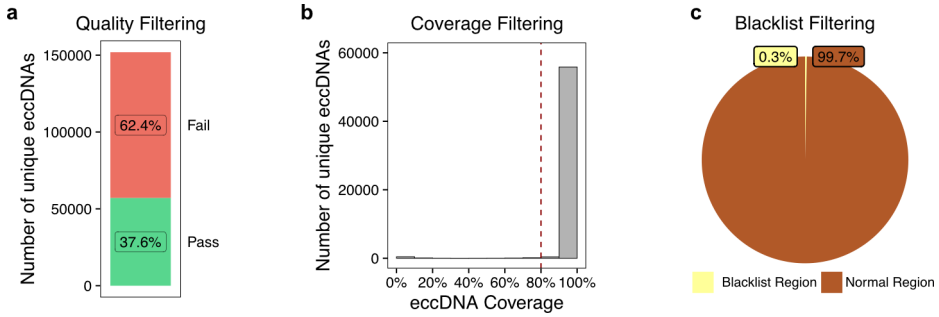


Fig. 5.6 | Sequential filtering steps to discern high-quality eccDNAs. a, The first filtering step focused on eccDNA quality metrics, retaining eccDNAs with a circle-score above 200 and at least five split reads supporting the eccDNA junction (n = 57,049). **b**, The subsequent filtering step eliminated eccDNAs that were inadequately covered (less than 80%) by Circle-seq reads. **c**, The final filtering step considered blacklist regions as defined by ENCODE, discarding any eccDNAs that intersected with these blacklisted areas (Amemiya, Kundaje & Boyle, 2019).

Through these three filtering stages, the initial dataset of around 150,000 putative eccDNA junctions was reduced to 56,092 high-quality eccDNAs used for downstream analysis (Figure 5.6).

5.4 EccDNAs in PDAC PDCLs: Size distribution and origin from gene-rich chromosomes

In this Circle-seq study, a total of 56,092 high-quality eccDNAs were identified which covered 4.1% (125,646,252 bp) of the human genome. These eccDNAs originated from all chromosomes (Figure 5.7b). Their size distribution was notably variable, ranging from a few bp to up to 90 kbp, suggesting that large ecDNAs with a length of more than 100 kbp are not prevalent in these eight PDAC PDCLs. Notably, there were several prominent peaks, specifically around sizes of 350, 700, 1000, and 1050 bp (Figure 5.7c). The average size of 2,323 bp of all eccDNAs is in accordance to the mean eccDNA size identified in neuroblastoma tumours (2,403 bp, Koche et al. (2020)) suggesting similar median sizes between different tumour types, whereas shorter sizes can be expected in normal tissue (Møller et al., 2018a).

Furthermore, investigating the number of eccDNAs per chromosome revealed that long

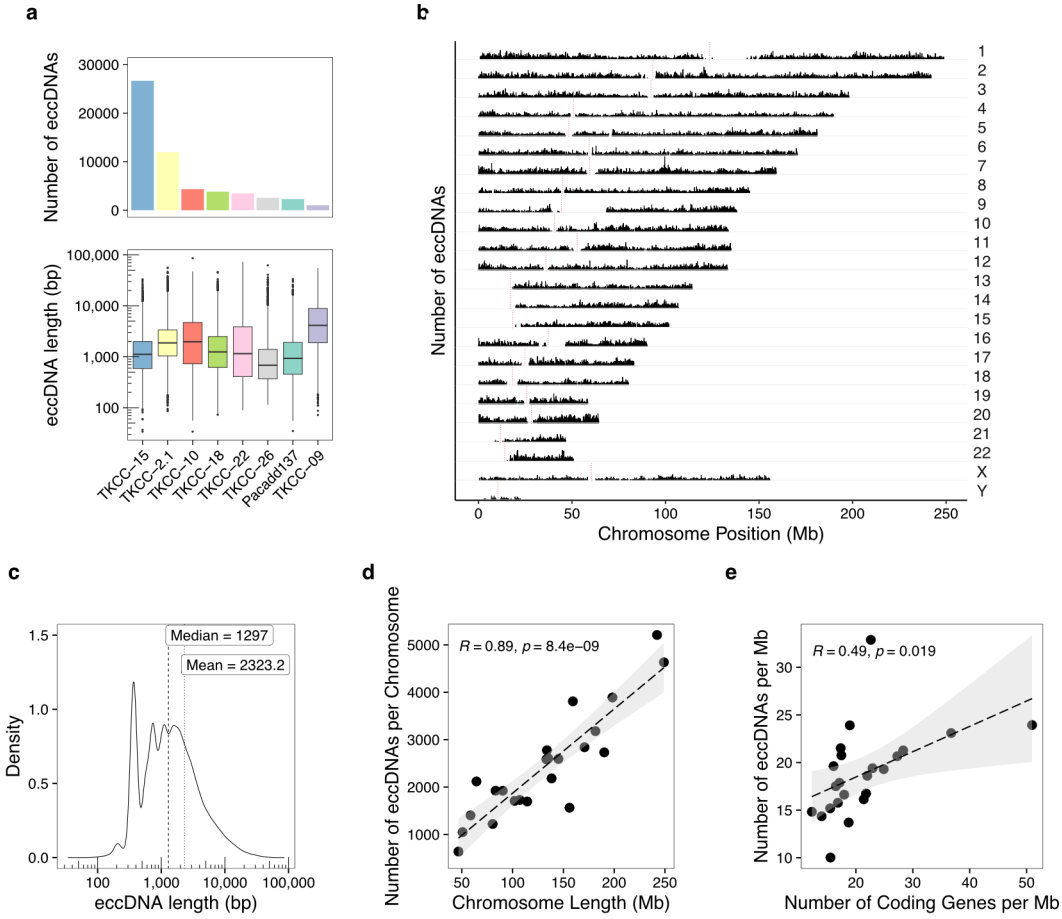


Fig. 5.7 | Characteristics of high-quality eccDNAs in eight PDAC PDCLs. **a**, Representation of the number and size distribution of high-quality eccDNAs discerned in the 9 PDAC PDCLs. **b**, Genome overview displaying the distribution of the high-quality eccDNAs, wherein the genome is divided into 100 kbp segments. The most concentrated segment contained 21 eccDNAs. **c**, Size distribution of all high-quality eccDNAs. **d**, Correlation plot displaying the number of eccDNAs per chromosome against the length of the chromosome. **d**, Correlation plot delineating the relationship between the number of eccDNAs and the density of coding genes per Mb. **c & d**, Correlation was assessed using the Pearson method. Chromosome Y was removed prior analysis.

and gene-rich chromosomes generate the most eccDNAs (Figure 5.7d,e). Notably, this phenomenon was previously observed in healthy human somatic tissue (Møller et al., 2018a). The propensities of longer chromosomes to generate more eccDNAs could be attributed to the higher susceptibility of longer chromosomes to DNA damage, due to the increase in size, compared to their shorter counterparts. Conversely, gene-dense chromosomes might contribute to more eccDNAs due to their increased transcriptional activity, which is widely postulated as one of the main drivers of eccDNA formation (Møller et al., 2018a; Hull et al., 2019; Dillon et al., 2015).

5.5 EccDNA origins: Ties with gene expression and chromatin accessibility

EccDNAs are omnipresent across the genome, but are enriched for specific genomic features (Møller et al., 2018a). While their miniature size mostly prevents them from encompassing whole genes, they often contain partial genes, which also might be transcriptionally active

(Koche et al., 2020; Paulsen et al., 2019). Additionally, transcription is thought to be a main mechanisms of eccDNA formation. Interestingly, eccDNAs tend to contain a higher GC content compared to their neighbouring regions and are marked with active histone marks, a further sign of transcriptional activity in proximity to the eccDNA origin (Dillon et al., 2015; Shibata et al., 2012; Kudla et al., 2006).

As most of these PDCLs have previously been characterised for their chromatin accessibility (Brunton et al., 2020) and gene expression levels (Dreyer et al., 2021), a novel approach of integrating Circle-seq with RNA-seq and ATAC-seq data can strengthen the relation between active transcription and open chromatin to eccDNA formation. ATAC-seq is a genomic technique used to assess chromatin accessibility. It uses a hyperactive Tn5 transposase to cut open nucleosome-free regions of the genome and simultaneously insert sequencing adapters. This key step not only marks the accessible DNA fragments, but also prepares them for subsequent amplification and sequencing. The analysis of the sequencing data provides crucial insights into the open chromatin landscape of the sample (Buenrostro et al., 2013).

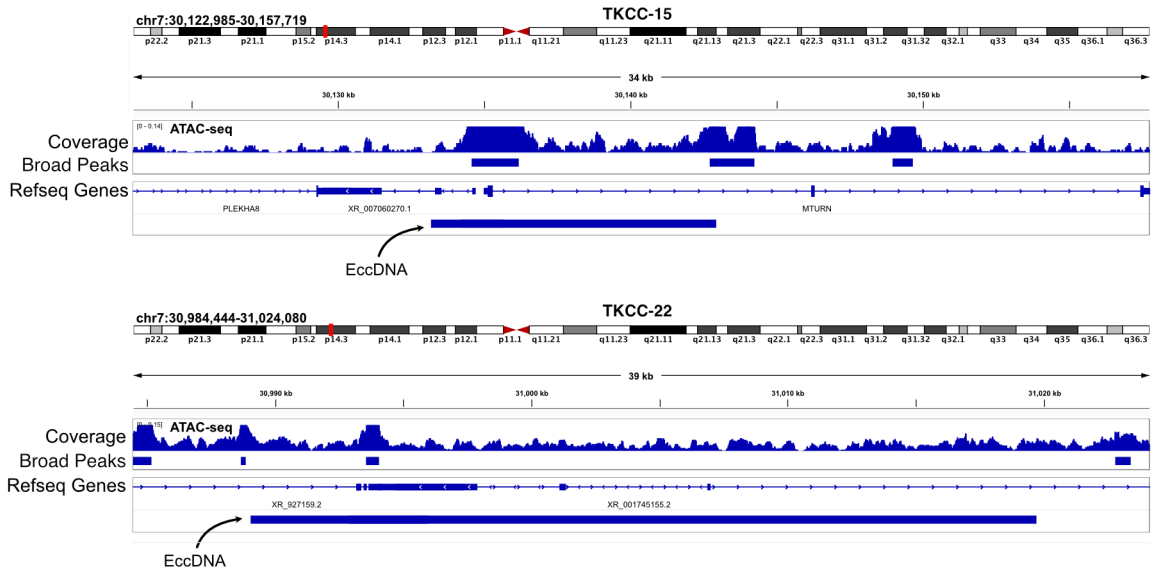


Fig. 5.8 | Genomic overview of selected eccDNAs with matching ATAC-seq data. One eccDNA from the PDCL TKCC-15 (top) and one from TKCC-22 (top) were selected to display the overlap to the matching ATAC-seq data. Broad ATAC-seq peaks were identified within an eccDNA region and in proximity to the eccDNA junction. The ATAC-seq data was previously generated by Brunton et al. (2020).

To get an overview of the eccDNAs and the corresponding ATAC-seq data, a genomic view was generated of two eccDNAs with matching ATAC-seq data and their identified peaks. The two cases examined showed peaks overlapping with the eccDNA region, but also additional peaks adjacent to the eccDNA junctions (Figure 5.8). This showed that some eccDNAs can overlap with ATAC-seq peaks and have peaks in proximity to eccDNA junctions. These single genomic views can provide a deeper insight into how eccDNAs are formed and where chromatin is accessible within or around eccDNA origins. However, a more detailed investigation of the location of ATAC-seq peaks in the vicinity of eccDNA junctions was not performed. Instead, a genome-wide analysis was performed to investigate whether, in general, increased open chromatin is associated with eccDNA regions compared to non-eccDNA, chromosomal (chr) regions.

Performing a genome-wide matching of the the ATAC-seq with the eccDNA data revealed an association between elevated chromatin accessibility and the origins of eccDNA (Figure 5.9a). A majority of the samples (6 of 7) indicated that ATAC-seq peaks overlapping with eccDNAs had a higher weight compared to non-eccDNA overlapping peaks (chrDNA), suggesting that regions described by a more accessible chromatin are more prone to generate eccDNAs.

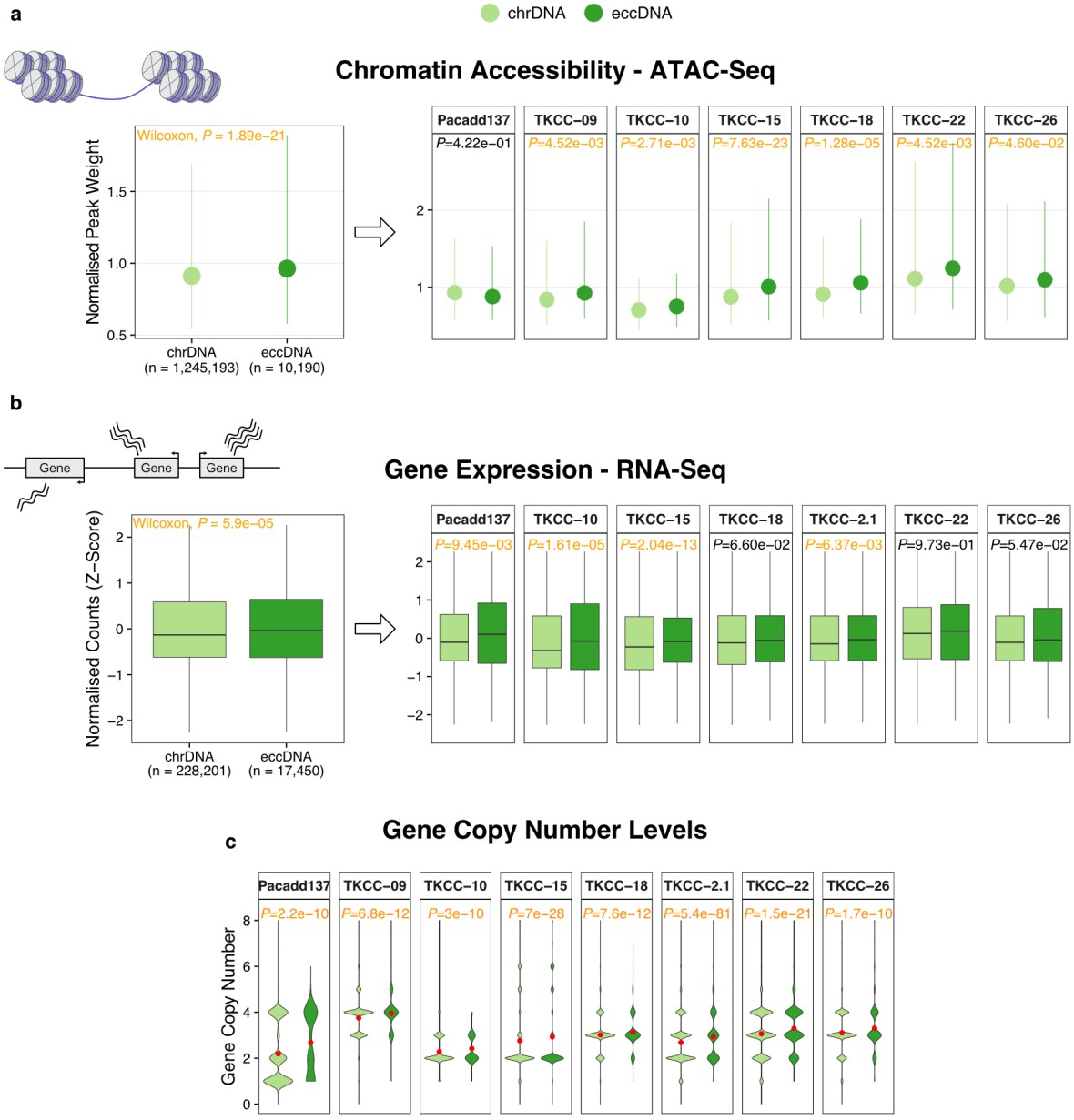


Fig. 5.9 | EccDNAs originate from genes with elevated expression and regions with high increased accessibility. a, Simplified boxplot displaying 25% and 75% normalised peak weight range (line) and the median (point). EccDNA regions were integrated with ATAC-seq data of seven PDAC PDCLs published by Brunton et al. (2020). Peaks inside an eccDNA were annotated as 'eccDNA' and peaks not included inside an eccDNA were annotated as chromosomal DNA (chrDNA). **b**, Boxplot showing the comparison of normalised gene expression values of genes overlapping with an eccDNAs (eccDNA) and genes without eccDNA overlap (chrDNA). Gene expression data of seven PDAC PDCLs was available from Dreyer et al. (2021). **c**, Violin plot displaying copy number levels of eccDNA overlapping (eccDNA) and non-overlapping genes (chrDNA). The mean copy number level is displayed as a red dot. **a, b, & c**, Statistical difference was assessed using a Wilcoxon-rank sum test. P values for all samples were additionally adjusted using the Benjamini-Hochberg method. Significant association (P value or P adjusted < 0.05) is displayed in orange.

Nevertheless, while the chromatin of ecDNA is considerably more accessible compared to its chromosomal counterpart, the increased chromatin accessibility of small eccDNAs remains ambiguous. These findings do not suggest a more open chromatin in eccDNAs as the ATAC-seq data does not directly match to the Circle-seq data and eccDNAs have been established to be highly variable in individual cells (Wu et al., 2019; Møller et al., 2018a). However, the analysis can suggest that eccDNA-generating regions have a more open chromatin structure compared to regions in the genome that are sparsely generating eccDNAs.

In order to determine the consequence of open chromatin on gene transcription, ATAC-seq peaks are usually annotated by the nearest regulatory elements or genes (Yan et al., 2020). Consequently, the analysis suggests that the open chromatin structure from chromosomal regions generating eccDNAs is transcriptionally active. This suggestion is further evaluated by analysing the available RNA-seq data of seven PDAC PDCLs. Interestingly, genes that overlapped with an eccDNA had an increased expression relative to those that did not intersect with an eccDNA (Figure 5.9b). To note, almost all of the eccDNAs did not incorporate a full gene region, but contain only parts of genes (Figure 5.12a). Therefore, it is not expected that these eccDNAs drive the expression of the incorporated genes. In contrast, this comparison suggests that parts of genes with heightened transcriptional activity are more likely to be located on eccDNAs.

Furthermore, an increase in the copy number levels of genes tied to eccDNA origin was uncovered (Figure 5.9c). This reinforces the association between transcriptional activity and eccDNA formation.

To further refine the specific genes associated with eccDNA origin, a pathway overrepresentation analysis leveraging the KEGG pathways was performed (Figure 5.10). This pathway analysis identified that genes from PDAC- and cancer-related pathways are significantly enriched to be overlapping with eccDNAs (Biankin et al., 2012; Kanehisa & Goto, 2000; Yu & Hung, 2000).

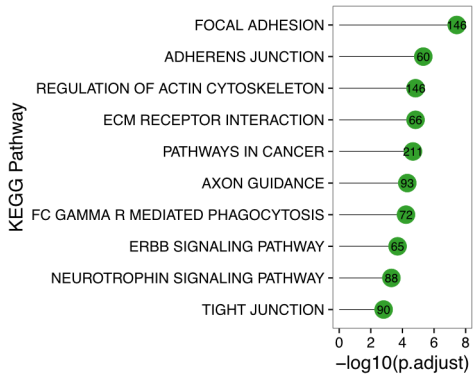


Fig. 5.10 | EccDNAs originate from PDAC and cancer related pathway genes. Overrepresentation analysis results of the genes overlapping an eccDNA in the KEGG pathways. Pathways are sorted based on the P adjust value with the gene count for each pathway displayed in the respective dot. Statistical analysis was performed using a hypergeometric test with P values adjusted using the Benjamini-Hochberg method. The 10 pathways with the lowest P adjusted value are displayed.

In sum, these findings underscore the intricate ties between eccDNAs, gene expression,

and chromatin accessibility. This underscores earlier findings of eccDNA hotspots at regions with active transcription and open chromatin and underpin the genomic specificities of the eccDNA origin.

5.6 Specific genomic features are enriched on eccDNAs

The previous analysis established that the eccDNAs present in PDAC PDCLs arise from all parts of the genome, with a particular enrichment in gene-dense chromosomes, and in region with actively transcribed genes and open chromatin marks. In order to delve deeper into the composition of these eccDNAs, various genomic elements have been used for annotation.

The annotation of eccDNA with genomic features indicates that the proportion of eccDNAs containing specific genomic features is relatively stable across all samples, as shown in Figure 5.11a. However, the sample TKCC-09 deviated from this trend by exhibiting a higher proportion of the majority of genomic features. Specifically, gene elements and individual repeat elements display a significant overrepresentation in TKCC-09 compared to the other seven samples. It is noteworthy that TKCC-09 displayed the largest average eccDNA size, potentially elevating the probability of containing specific genomic elements (Figure 5.7a). Excluding TKCC-09, the seven remaining PDAC PDCLs exhibited similar proportions, indicating a uniform distribution of genomic elements on eccDNAs.

Diving into the annotated eccDNA genomic features, a strikingly high proportion of eccDNAs are annotated with at least one repetitive element (around 90% of eccDNAs, Figure 5.11a). The most abundant repetitive elements are short-interspersed nuclear elements (SINEs) which are prevalent on more than 60% of all eccDNAs. Furthermore, over half of the identified eccDNAs encompass gene regions, predominantly from protein-coding genes. However, considering the prevalence of all these genomic elements, in the human genome, is vital for a comprehensive understanding of common eccDNA origins.

To fully assess whether specific genomic elements are enriched or absent on eccDNAs, a permutation test was utilised with randomised eccDNA datasets. In this approach, 1,000 random datasets were generated by shuffling the locations of all high-quality eccDNA regions on their respective chromosome, creating a total of 56,092,000 random eccDNAs spread across 1,000 datasets (Methods Section 2.18 or Figure 5.11b). Subsequent annotations of both random datasets and the actual eccDNAs with genomic elements facilitated the basis for the statistical permutation test.

The permutation analyses underscored that eccDNAs are not random features, but encompass distinct genomic elements (Figure 5.11c). Repeat elements are significantly represented on eccDNAs. Given that up to 69% of the human genome consists of repetitive elements, an abundance of repetitive elements on eccDNAs is anticipated (Koning et al., 2011). Yet, eccDNAs exhibited a significant increased number of repeat element annotations compared to their

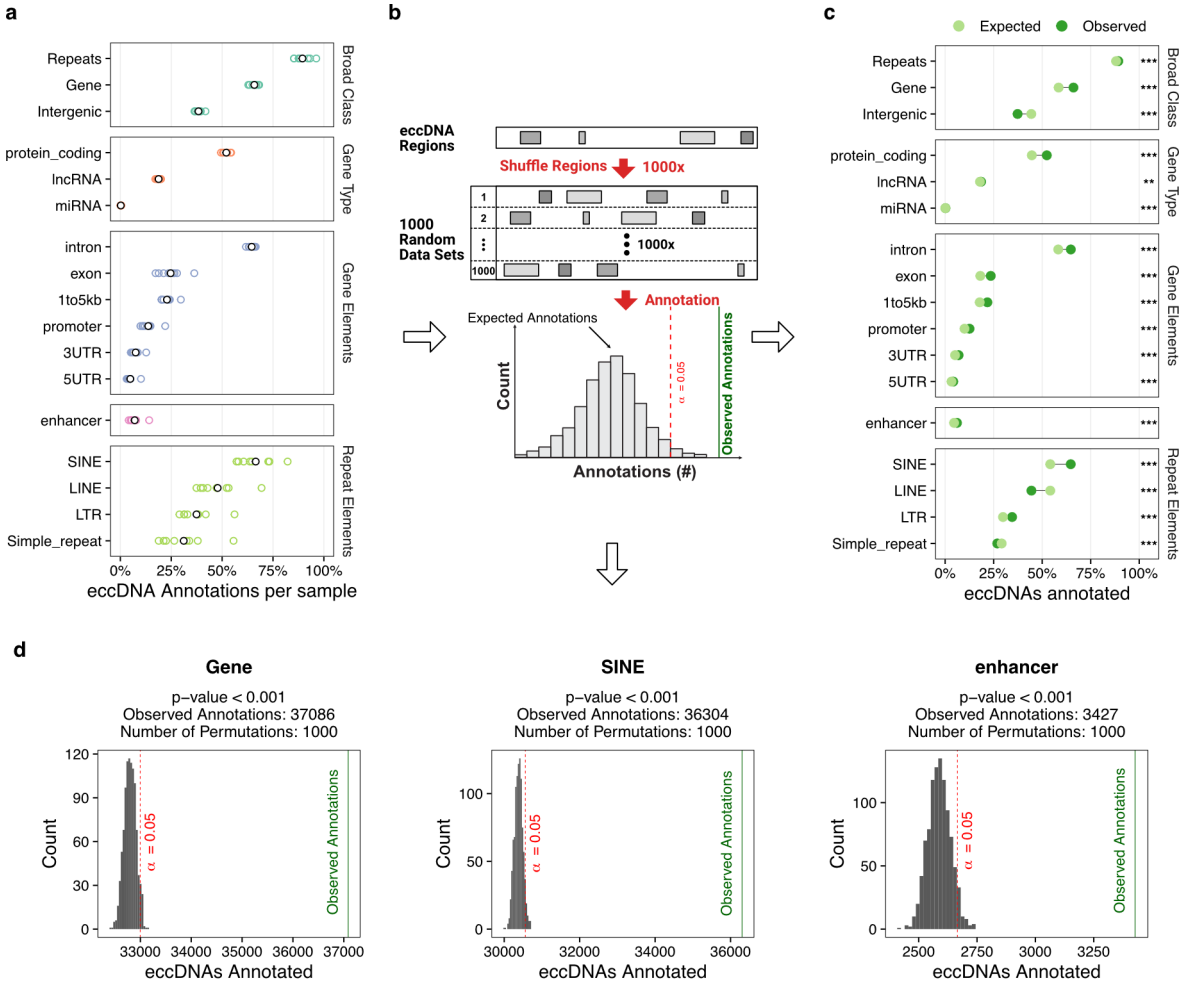


Fig. 5.11 | EccDNAs contain specific genomic features. a, Genomic feature abundance is similar between the eight PDAC PDCLs. **b**, Schematic illustration of permutation test performed to compare the number of expected and observed annotations of each genomic feature. To identify the expected number of annotations, a 1000 random datasets were generated and annotated. **c**, Dumbbell plot displaying difference between expected and observed number of annotations for diverse genomic features. **d**, Example permutation test of the three genomic features: genes, SINEs, and enhancers. **c & d**, P values are calculated through the permutation test performed for each genomic feature. ***: P value < 0.001, **: P value < 0.01.

randomised counterparts. Among the repeat elements, SINEs and long-interspersed nuclear elements (LINEs) were the most abundant elements identified on eccDNAs. Interestingly, SINEs were more abundant on eccDNAs, whereas LINEs were comparatively rarer. This hinted the potential roles of specific repeats in the eccDNA genesis or their location inside and close to eccDNA formation hotspots.

Moreover, genes and their respective elements exhibit significant overlap with eccDNAs. In contrast, intergenic regions are markedly underrepresented. Dividing genes into the three gene types, protein-coding, long non-coding RNA (lncRNA), and microRNA (miRNA), an enrichment is observed for all. These observations underline the specificity of eccDNA biogenesis within genes.

A recent study showed that nuclear ecDNAs are mobile and can enhance gene expression if they harbour super-enhancers (Zhu et al., 2021). These super-enhancers, when on ecDNAs, can interact with regions distant from their original chromosomal region to form

transcriptional hubs and activate gene expression. Our analysis revealed that enhancers are also enriched on smaller eccDNAs in the PDAC PDCLs, a feature observed similarly in ecDNAs (Figure 5.11c,d) (Zhu et al., 2021). This suggests that even smaller eccDNAs might be shaping the tumour transcription by encompassing enhancers. Nonetheless, additional experiments are paramount to confirm this hypothesis.

In summary, eccDNAs are not random, but arise from specific genomic elements. Most genomic elements depicted are overrepresented on eccDNAs than randomly expected, which highlights the predominant origins of eccDNAs. This includes gene regions, enhancer elements, and SINE elements, advancing our understanding of the origin and the overlap of eccDNAs in cancer cells (Dillon et al., 2015).

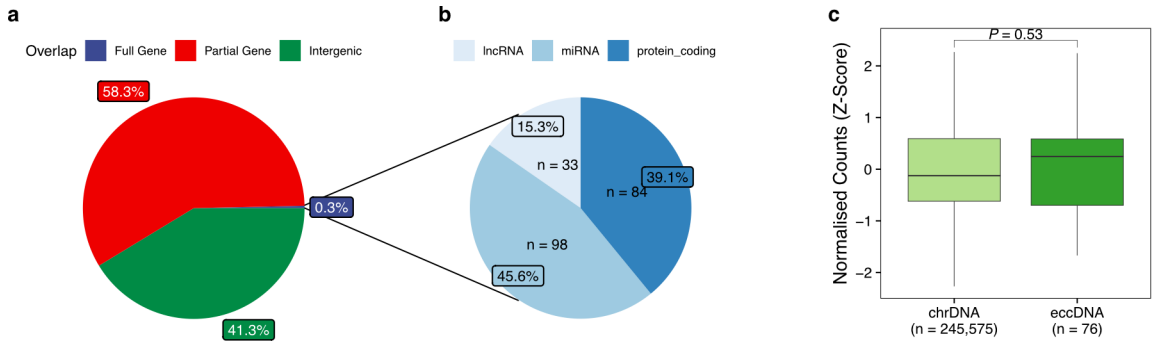
5.7 Genes are rarely fully incorporated on eccDNAs

EccDNAs were identified in neuroblastoma as short, non-amplified fragments that mostly contain partial genes. Additionally, if a gene is fully incorporated, it does not exhibit increased expression (Koche et al., 2020). Given these characteristics, I investigated whether we observe similar eccDNA features in our PDAC PDCLs.

Unsurprisingly, eccDNAs mostly contain partial genes, and only a minor fraction of eccDNAs did harbour the full gene locus. Out of all these genes that are fully residing on eccDNAs (n = 215), 84 encoded for a protein and 98 encoded for a miRNA. Within these genes, 12 were identified to be cancer driver genes based on the allOnco cancer driver list (Figure 5.12b,e,f, www.bushmanlab.org). Although none of these genes were found to be major drivers of PC, they have been shown to play roles in other types of cancer (Waddell et al., 2015). For instance, in lung cancer cells, the miRNA *MIR23A* is regulating TGF-β-induced epithelial-mesenchymal transition (Cao et al., 2012). *MIR27A* regulates immune response and chemoresistance, while *PF4* can promote lung cancer growth (Zhang et al., 2019a; Xie et al., 2014; Pucci et al., 2016). It is unclear whether these cancer driver genes located on eccDNAs also impact PDAC, and further research is needed. However, with their presence on eccDNAs the dynamic regulation of their copy number levels is provided which can impact the activity of these oncogenes (Yi et al., 2022).

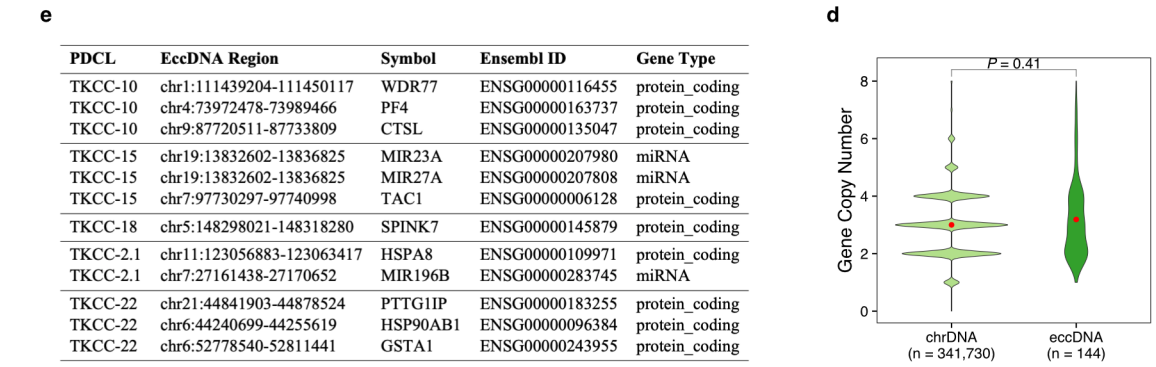
In order to evaluate any increases in copy number levels and expression on genes residing on eccDNAs, we compared the expression and copy number levels of these genes to the non-eccDNA genes. The analysis found no evidence of increased expression or copy number levels for genes found on eccDNAs (Figure 5.12c,d). This highlights the contrast between eccDNAs and ecDNAs, with ecDNAs playing a significant role in amplifying and expressing oncogenes, while eccDNAs have minimal impact on gene transcription or copy numbers.

These findings highlight the distinct characteristics of eccDNAs in PDAC PDCLs, especially when compared to ecDNAs, which often drive oncogene amplification and expression.



e

PDCL	EccDNA Region	Symbol	Ensembl ID	Gene Type
TKCC-10	chr1:111439204-111450117	WDR77	ENSG00000116455	protein_coding
TKCC-10	chr4:73972478-73989466	PF4	ENSG00000163737	protein_coding
TKCC-10	chr9:87720511-87733809	CTSL	ENSG00000135047	protein_coding
TKCC-15	chr19:13832602-13836825	MIR23A	ENSG00000207980	miRNA
TKCC-15	chr19:13832602-13836825	MIR27A	ENSG00000207808	miRNA
TKCC-15	chr7:97730297-97740998	TAC1	ENSG00000006128	protein_coding
TKCC-18	chr5:148298021-148318280	SPINK7	ENSG00000145879	protein_coding
TKCC-2.1	chr11:123056883-123063417	HSPA8	ENSG00000109971	protein_coding
TKCC-2.1	chr7:27161438-27170652	MIR196B	ENSG00000283745	miRNA
TKCC-22	chr21:44841903-44878524	PTTG1IP	ENSG00000183255	protein_coding
TKCC-22	chr6:44240699-44255619	HSP90AB1	ENSG00000096384	protein_coding
TKCC-22	chr6:52778540-52811441	GSTA1	ENSG00000243955	protein_coding



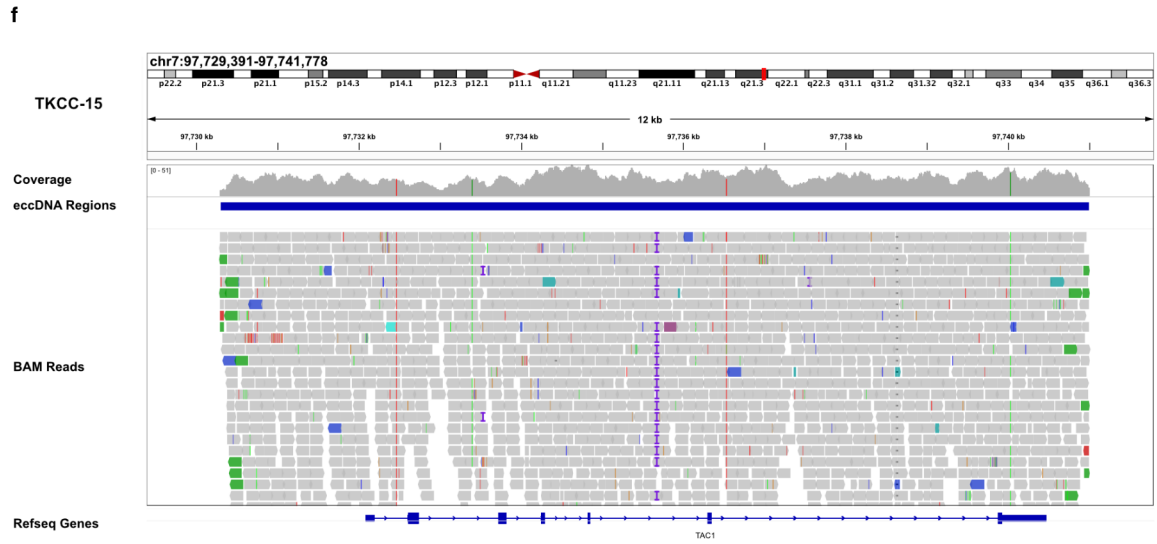


Fig. 5.12 | EccDNAs in PDAC rarely contain full genes. **a**, Fraction of eccDNAs containing full genes, partial genes, or intergenic regions. **b**, Fraction of full genes residing on eccDNAs grouped into their gene type. **c**, Relative normalised gene expression (Z-Score) of genes fully on eccDNAs (eccDNA) and others (chrDNA). **d**, Copy number levels of genes fully located on eccDNAs compared to genes not completely on eccDNAs (chrDNA). **c & d**, Statistical comparison was performed using a Wilcoxon-rank sum test. **e**, Cancer driver genes from the allOnco list (www.bushmanlab.org) completely residing on eccDNAs. **f**, Representative figure of an eccDNA fully incorporating a complete gene locus identified by Circle-seq. Here, *TAC1* is incorporated in a TKCC-15 eccDNA.

5.8 Identification of multi-fragment eccDNAs

Standard tools for eccDNA identification, optimised for short-read sequencing data, are effective in identifying single-fragment eccDNAs (Prada-Luengo et al., 2019; Kumar et al., 2020). However, these tools are limited in their ability to uncover multi-fragment eccDNAs, which constitute approximately 11% of all eccDNAs (Wang et al., 2021). While short-read sequencing data is sub-optimal for identifying multi-fragment eccDNAs, a work-around using *de novo* assembly and long-read alignment is thought to be able to generate this information.

Here, I utilised the functionality of Unicycler (Wick et al., 2017), which assembles short-read sequences *de novo* to identify circular structures. Unicycler is commonly used for bacterial genomes, but, due to the circular structure of eccDNAs, it is potentially useful for eccDNA detection. Following the *de novo* assembly, identified eccDNA sequences are mapped to the GRCh38 reference genome using Minimap2 (Li, 2018) to determine their genomic origin. Both, Unicycler and Minimap2, are incorporated into the nf-core/circdna pipeline within the 'unicycler' branch.

This section evaluates this approach for a more comprehensive identification of eccDNAs and characterises the multi-fragment eccDNAs to provide a complete overview within the PDAC PDCLs.

5.8.1 Identification of high-quality *de novo* assembled eccDNAs

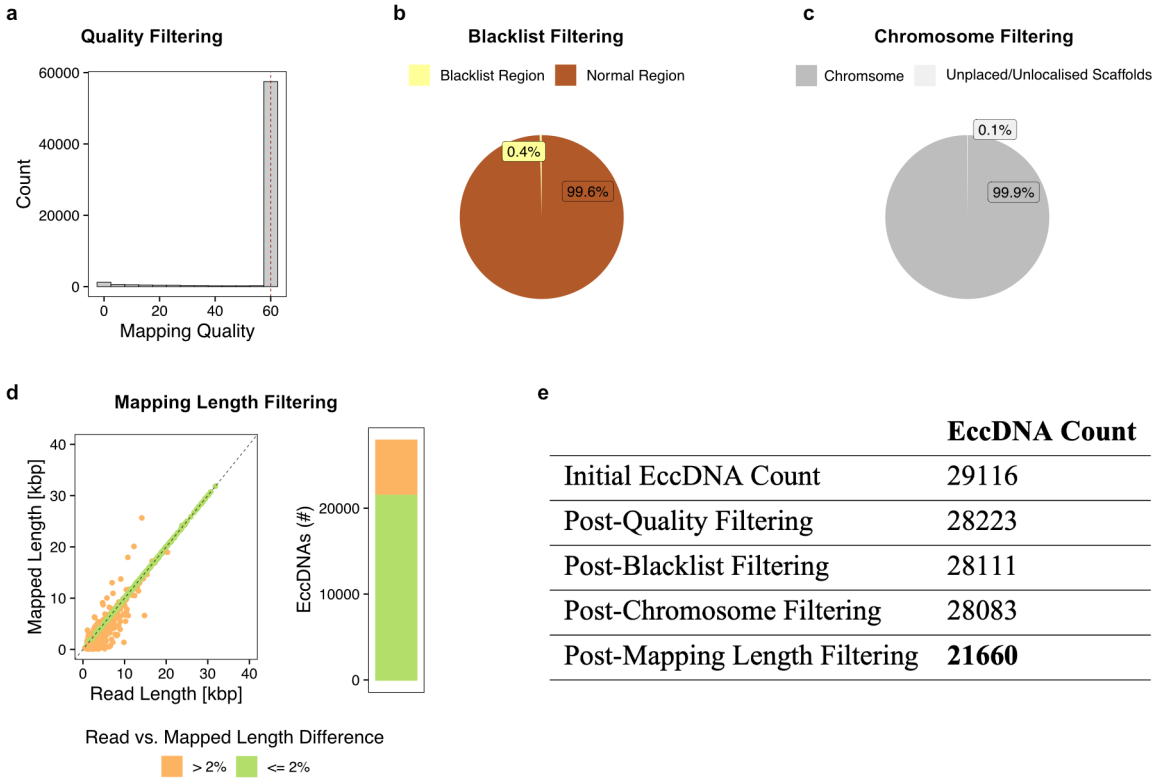
Given the novelty of using a *de novo* assembly approach, with short-read sequencing data, for identifying multi-fragment eccDNAs, a stringent quality filtering process was implemented. This process encompassed multiple filtering steps: read mapping quality filtering, blacklist region filtering, chromosome filtering, and mapping length filtering. In detail, quality filtering removes all reads that did not map with the highest mapping quality (mapping quality = 60) to the reference genome, blacklist filtering removed eccDNAs overlapping with blacklist regions defined by ENCODE (Amemiya, Kundaje & Boyle, 2019), chromosome filtering discarded eccDNAs mapped to unplaced or unlocalised scaffolds, and mapping length filtering was evaluated by comparing the length of the mapped reads to their actual read length, ensuring that the majority of each read are mapped accurately to the genome (Figure 5.13).

After applying these filters, a reduction of 25.6% in the number of identified eccDNAs has been observed. The initial count of 29,116 unique eccDNAs, identified via Unicycler and mapped with Minimap2, was reduced to 21,660 that passed all filtering criteria (Figure 5.13e). The most eccDNAs were removed based on the discrepancy between the read and the mapped length (Figure 5.13d). Although other filters led to fewer exclusions, their implementation was still important to retain high-quality eccDNAs.

In summary, the implementation of stringent quality control measures effectively filtered out low-quality eccDNAs, reducing the initial set from 29,116 to 21,660. These high-quality eccDNAs, which survived multiple levels of scrutiny, serve as a robust dataset for subsequent analyses aimed at deciphering their potential roles and origins.

5.8.2 Multi-fragment eccDNAs in PDAC PDCLs

Out of the 21,660 high-quality *de novo* assembled eccDNAs from eight PDAC PDCLs, 89.73% were found to originate from a single genomic fragment, revealing that around 10% of the eccDNAs harbour two or more fragments (multi-fragment eccDNAs, Figure 5.14a). Similar proportions were determined by Wang et al. (2021), using long-read sequencing, suggesting



e

	EccDNA Count
Initial EccDNA Count	29116
Post-Quality Filtering	28223
Post-Blacklist Filtering	28111
Post-Chromosome Filtering	28083
Post-Mapping Length Filtering	**21660**

Fig. 5.13 | Unicycler eccDNA filtering steps. a, Exclusion of reads with a mapping quality below 60. **b**, Pie plot showing proportion of eccDNAs overlapping with an ENCODE blacklist region (Amemiya, Kundaje & Boyle, 2019). **c**, EccDNAs mapping to chromosomes or unlocalised/unplaced scaffolds. **d**, Comparison between Unicycler-generated read length and actual length mapped to chromosomes, with a cutoff discrepancy of more than 2%. **e**, Number of eccDNAs retained at each filtering step. The final count is highlighted in bold.

that using the combination of Unicycler and Minimap2 can identify multi-fragment eccDNAs. Notably, longer eccDNAs were more likely to consist of multiple fragments (Figure 5.14b,c). This suggests that larger eccDNAs may be composite structures, assembled from multiple genomic fragments (Figure 4.5).

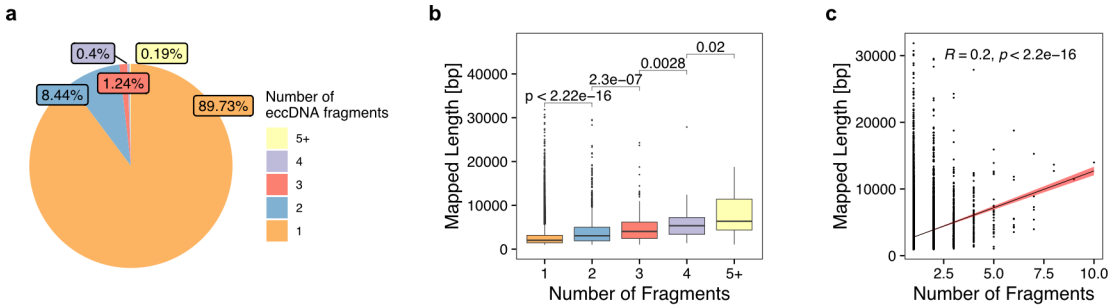


Fig. 5.14 | Large eccDNAs are comprised of multiple fragments. a, Pie chart showing the proportion of the number of fragments of the high-quality eccDNAs identified with Unicycler. **b**, Length difference of eccDNAs with single- or multiple fragments. Pairwise statistical comparison was performed using a Wilcoxon-rank sum test. **c**, Correlation analysis of the number of eccDNA fragments compared to their length. Correlation was assessed using the Pearson method.

Investigation of eccDNAs identified by Unicycler and subsequently mapped by Minimap2 showed extensive coverage by Circle-seq reads (Figure 5.15). Notably, certain single-fragment eccDNAs, which were revealed by *de novo* assembly using Unicycler, could also be independently verified by Circle-Map Realign. This lends credibility to the methodology employed. However, when it comes to multi-fragment eccDNAs, these were not detected by Circle-Map Realign or failed to meet the necessary filtering criteria (Figure 5.6). This highlights the

potential advantage of a combinatorial approach using both, *de novo* assembly and eccDNA junction identification, to capture a more diverse set of eccDNAs.

Fig. 5.15 | Read alignment views of eccDNAs identified by Unicycler. IGV figures are generated of representative single-fragment (top) and two-fragment eccDNAs (bottom) identified by Unicycler (citrus) in the TKCC-15 PDCL. As a reference single-fragment eccDNAs identified by Circle-Map Realign are depicted (lagoon).

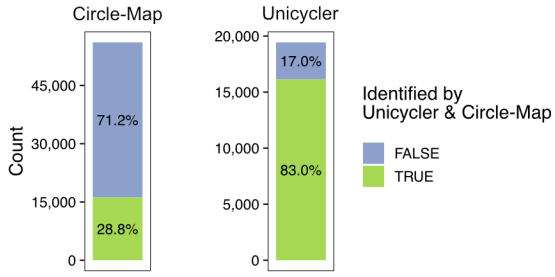


Fig. 5.16 | Most *de novo* assembled single-fragment eccDNAs are validated by Circle-Map. Overlap of high-quality eccDNAs identified by Circle-Map (n = 56,092) and single-fragment *de novo* assembled eccDNAs identified by Unicycler (n = 19,435). The bars depict if an eccDNA was identified by both eccDNA callers (overlap > 95%). Only around a quarter of eccDNAs identified by Circle-Map are also identified by Unicycler. However, most eccDNAs identified by Unicycler are also found in the Circle-Map dataset.

While Unicycler-guided *de novo* assembly successfully identified multi-fragment eccDNAs some limitations were evident (Figure 5.15 bottom). First, smaller eccDNAs (below 1,000 bp) are not reported by Unicycler. Therefore, only larger eccDNAs are identified. Second, the computational intensity of *de novo* assembly restricts the application to low-coverage Circle-seq data and smaller eccDNAs (data not shown). Third, eccDNAs with regions not covered by Circle-seq reads may be incompletely assembled and subsequently missed. In this study, short-read sequencing data was used to identify eccDNAs from various origins. However, third generation long-read sequencing technology could provide a more comprehensive understanding of both single- and multi-fragment eccDNAs (Wang et al., 2021).

Nevertheless, while this approach might be limited by certain aspects, around 10% of eccDNAs were identified to be comprised of multiple fragments in these PDAC PDCLs. Additionally, multi-fragment eccDNAs also show increased sizes compared to eccDNAs made out of one genomic fragment.

5.9 Computational validation of *de novo* assembly approach for eccDNA identification

The identification of eccDNAs via *de novo* assembly generated findings that were consistent with those of Circle-Map Realign, a tool commonly used for detecting eccDNA junctions in Circle-seq data (Figure 5.15) (Prada-Luengo et al., 2019). To further evaluate the efficacy of *de novo* assembly in identifying eccDNAs, a comparative analysis was conducted between the two methods, focusing on single-fragment eccDNAs due to limitations in eccDNA junction detection tools in identifying multi-fragment eccDNAs. As *de novo* assembly and subsequent mapping does not always identify the correct eccDNA junction, a method was chosen that compares the complete regions of the *de novo* assembled eccDNAs to the eccDNA regions identified via eccDNA junction detection. An overlap of at least 95% was declared to be sufficient for both eccDNA calling methods to depict the same eccDNA.

The 56,092 high-quality eccDNAs, identified by Circle-Map Realign, were compared

to the 19,435 high-quality single-fragment *de novo* assembled eccDNAs (Unicycler). The results indicate a considerable difference between the two methods. Nearly 75% of eccDNAs identified by Circle-Map Realign were not captured by *de novo* assembly (Figure 5.16 left). One plausible explanation could be the minimum sequence length of 1,000 bp imposed by Unicycler. However, it is noteworthy that around 83% of single-fragment *de novo* assembled eccDNAs are also identified by Circle-Map Realign.

In summary, although *de novo* assembly covers a portion of the single-fragment eccDNA landscape, it provides the complete eccDNA structure. This cannot be accomplished solely through eccDNA junction detection tools, which solely detect the eccDNA junction. Notably, the use of this method allows for complete validation of single-fragment eccDNAs or the identification of their entire structures. Moreover, the accuracy of identifying single-fragment eccDNAs assembled *de novo* is confirmed in over 80% of cases. Therefore, combining *de novo* assembly with eccDNA junction detection methods enhances both the accuracy and comprehensiveness of eccDNA identification, particularly with the ability to identify multi-fragment eccDNAs.

5.10 Validation of Circle-seq results

The Circle-seq analysis identified thousands of eccDNAs across eight PDAC PDCLs varying in size, genomic origin, and chromosomal content. While Circle-seq is a prominent and commonly used method to study the eccDNA landscape, an experimental validation of the results is recommended (Koche et al., 2020; Kumar et al., 2017; Møller et al., 2015; Møller et al., 2018a). This validation can be performed by multiple methods such as inverse PCR, Sanger sequencing, DAPI imaging, or FISH imaging (Kumar et al., 2020; Møller, 2020; Koche et al., 2020; Turner et al., 2017). Since standard imaging methods do not have the required resolution for smaller eccDNAs (< 100 kbp) and their content, I selected inverse PCR and subsequent Sanger sequencing to validate some of the eccDNAs that were identified in the PDCLs (Figure 5.17a).

In total, I selected 17 eccDNA candidates: nine single-fragment eccDNAs identified by Circle-Map Realign (C) and eight multi-fragment eccDNAs identified by Unicycler (U). These candidates were chosen for their size, content, complexity, and PDCL of origin. For each PDCL, at least one eccDNA was aimed to be validated. The complete list of chosen eccDNAs is detailed in Table B.1. Among these eccDNA candidates, some encompassed full protein-coding genes or miRNAs, which could be actively transcribed in the cells.

For each candidate, outward-pointing primers were designed to amplify an individual eccDNA junction (Table 2.6). Outward-pointing primers used in an inverse PCR are usually employed to identify circular nucleic acids (Su et al., 2021; Kumar et al., 2020; Zhao et al., 2019). In linear/chromosomal DNA, these primers face away from each other; however, in a circular DNA, they can amplify a specific junction. While tandem duplications can create

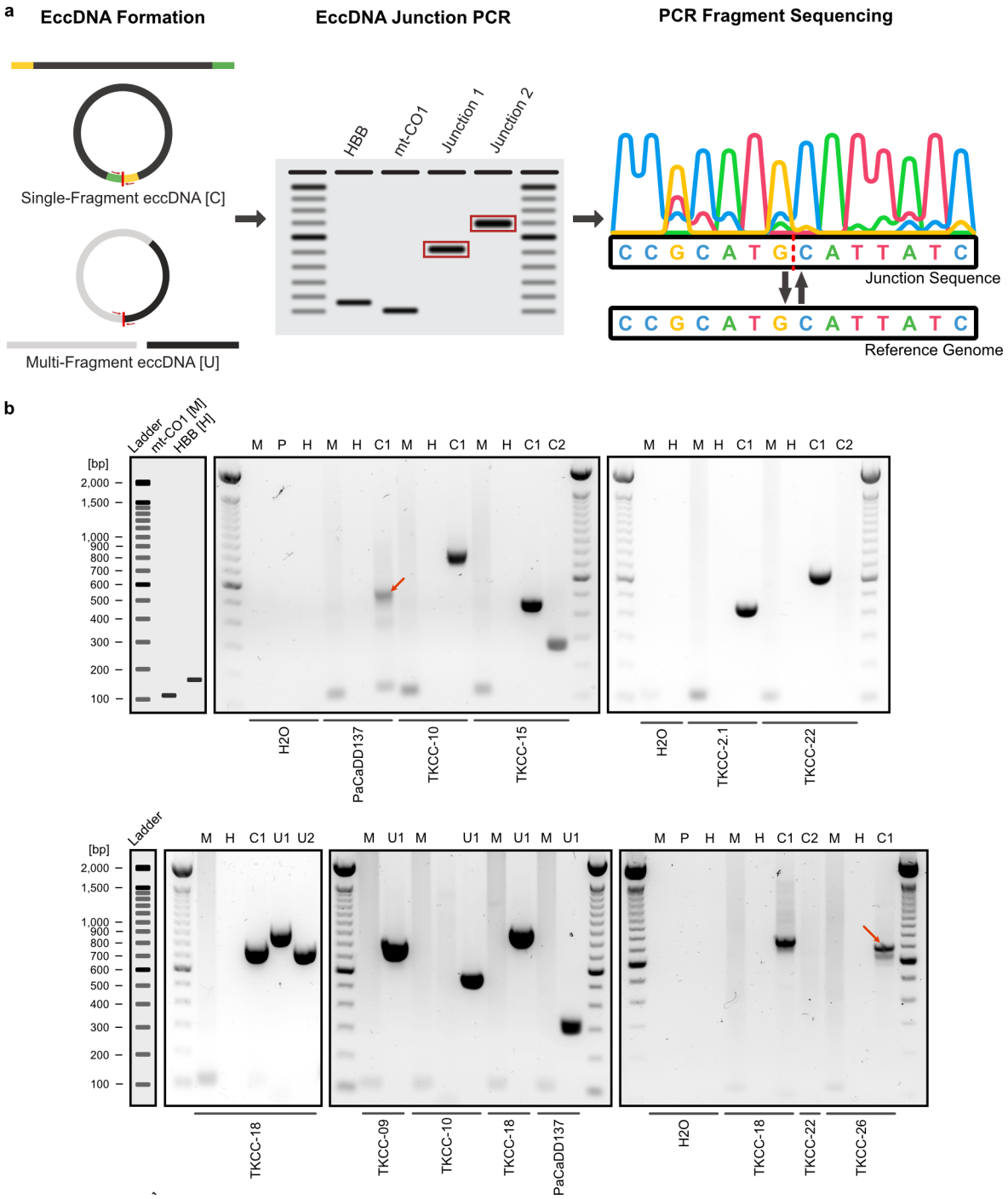


Fig. 5.17 | Inverse PCR validates Circle-seq results. a, Experimental set-up of Circle-seq validation. Candidate single-fragment eccDNAs identified using Circle-Map (C) and multi-fragment eccDNAs identified using Unicycler (U) were validated using inverse PCR and subsequent Sanger sequencing of the PCR fragment. Inverse PCR uses inverse-directed primers to amplify the eccDNA junction. **b**, Agarose gel electrophoresis of PCR fragments of single- and multi-fragment eccDNA candidates. The correct band size of candidates showing multiple bands is highlighted with a red arrow. *MT-CO1* and *HBB* were used as controls for the circular DNA and linear DNA level in the prepared DNA, respectively. Optimal bands of the 100 bp ladder and the two controls are displayed on the left side.

similar structures, which will be amplified by inverse PCR, our DNA preparation process minimised chromosomal DNA, and thereby reducing the chances of false positives. The expected fragment size for each candidate is detailed in Table 5.3.

A previous analysis confirmed the selective removal of linear DNA and retention of circular DNA prior to sequencing (Figure 5.1 and Table 5.2). Additionally, in most inverse

PCR experiments, controls for circular DNA (*MT-CO1*) and linear DNA (*HBB*) were added to validate the DNA preparation process. In all PCR experiments, for each PDCL, the same Circle-seq prepared DNA was used which was also subjected to sequencing.

Overall, inverse PCR validated 16 of 17 candidate eccDNAs, as evidenced by clear bands at expected sizes (Figure 5.17b). One candidate (TKCC-22 C2) yielded no PCR amplification with the designed primers. The positive control, amplifying *MT-CO1*, was also present in all DNA samples. In contrast, *HBB* was absent verifying that chromosomal DNA was largely reduced before the PCR procedure.

To verify the junction sequences, all PCR products underwent purification and Sanger sequencing. 13 of these 16 validated eccDNAs showed high sequence similarity to their respective origin in the human reference genome (Figure 5.18). The remaining three failed quality control or did not match the reference sequence. To investigate the sequence similarities, the first 25 bp upstream and downstream of each eccDNA junction sequence (top) and compared to the matching reference genome sequence (bottom). Minor alterations were observed near the junctions, potentially due to insertions, deletions, or single nucleotide variations (SNVs). The source of these alterations could not be determined as germline sequence data were unavailable. However, despite the observed alterations, the validated junctions closely resembled the reference sequence. These results corroborate the computational identification of the eccDNA candidates by Circle-Map Realign and Unicycler.

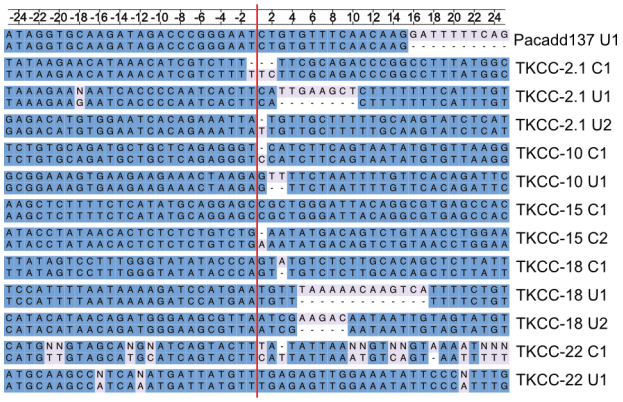


Fig. 5.18 | Sanger sequencing of eccDNA junctions confirms Circle-seq results. High concordance between sequence of the eccDNA candidate fragments amplified by inverse PCR (top) and the respective reference genome sequence (bottom). Concordant bases are highlighted in blue and base mismatches in grey. 25 bp upstream and downstream to the eccDNA junction are displayed for each candidate. Only sequences of candidates are displayed that were successfully sequenced and matched the reference genome with a high percentage.

Validation is key to ensure the accuracy of these methods, but can only be performed in a low throughput setting. Here, I chose some of the largest eccDNAs present, including some containing cancer-specific oncogenes. The full validation process was successful in approximately 76.5% (13/17) of the chosen candidates, but the inverse PCR, which already confirmed the existence of the eccDNA junction, was successful in all but one candidate (94.1%; 16/17). Further investigation is necessary, why the inverse PCR for TKCC-22 C2 failed. Given that PCR efficiency largely hinges on primer design (Dieffenbach, Lowe, Dveksler et al., 1993), a new primer set may help validate this particular eccDNA.

Tab. 5.3 | EccDNA candidates used for validation.

PDCL	ID	Expected PCR Fragment Size	PCR Validation	Junction Sequencing
Single-Fragment EccDNA identified by Circle-Map Realign (C)				
PaCaDD137	C1	508	✓	✗
TKCC-2.1	C1	442	✓	✓
TKCC-10	C1	778	✓	✓
TKCC-15	C1	449	✓	✓
TKCC-15	C2	285	✓	✓
TKCC-18	C1	757	✓	✓
TKCC-22	C1	657	✓	✓
TKCC-22	C2	643	✗	✗
TKCC-26	C1	693	✓	✗
Multi-Fragment EccDNA identified by Unicycler (U)				
PaCaDD137	U1	303	✓	✓
TKCC-2.1	U1	579	✓	✓
TKCC-2.1	U2	657	✓	✓
TKCC-09	U1	704	✓	✗
TKCC-10	U1	542	✓	✓
TKCC-18	U1	874	✓	✓
TKCC-18	U2	719	✓	✓
TKCC-22	U1	354	✓	✓

On the other hand, Sanger sequencing was successful in 13 of the 16 validated candidates. This shortfall could be attributed to the PCR DNA purification method employed. Rather than purifying the DNA directly from the agarose gel, the full PCR products were purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research Europe GmbH). While this approach removed primers, unwanted PCR fragments were not eliminated. This issue became evident as Sanger sequencing failed for two of candidates with multiple gel bands (PaCaDD137 C1, TKCC-26 C1; Figure 5.17b), suggesting that multiple sequences could have led to the sequencing failure. Only one candidate, TKCC-09 U1, showed a clear gel band yet failed sequencing, necessitating additional investigation.

In summary, both inverse PCR and Sanger sequencing verified the existence and the computational analysis of most candidate eccDNA junctions (Table 5.3). This supports the validity of eccDNA identification methods applied to the generated Circle-seq data and underscores the importance of these methods in accurate eccDNA identification.

5.11 The retention of eccDNAs in PDAC PDCLs

The replication and distribution of large, amplified ecDNAs during the cell cycle are well-characterised (Yi et al., 2022). However, the fate of smaller eccDNAs remains unclear. Previous work on leukocytes from the same patient found minimal eccDNA overlap even within these genetically identical cells (Møller et al., 2018a). Contrarily, eccDNAs in yeast (*Saccharomyces cerevisiae*) were found to often contain autonomously replicating sequences

suggesting eccDNA replication is possible during the cell cycle (Møller et al., 2015). Additionally, some evidence points to the sharing of eccDNAs across individual esophageal squamous cell carcinoma specimen of different individuals (Sun et al., 2021). Therefore, replication and retention of smaller eccDNAs, with most having only a few hundred or thousand bps, remains to be ambiguous. To dissect the retention and replication of small eccDNAs in PDAC PDCLs, two distinct methodologies were employed.

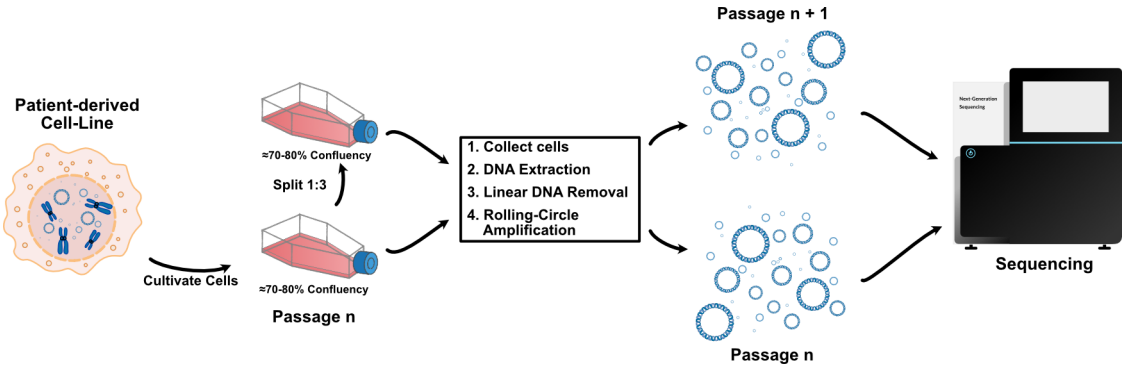


Fig. 5.19 | Graphical description of Circle-seq set-up from two consecutive passages of PDAC PDCLs. PDAC PDCLs are cultivated until a confluency of around 70-80% is reached. 2/3 of the cells are collected for DNA extraction and the remaining 1/3 is transferred into a new flask and supplied with new media. After 70-80% confluency is reached, again, cells are collected from the second passage and DNA is extracted. The extracted DNA of both passages is then prepared following the Circle-seq protocol, to enrich the circular DNA content. In short, the linear DNA is removed by circular DNA-safe DNase and the circular DNA is amplified by rolling-circle amplification using the Phi29 polymerase. Afterwards, the remaining DNA is sequenced using paired-end NGS.

First, a subset of validated candidate eccDNA junctions, identified by Circle-seq, were aimed to be validated in DNA extracted from a different passage of the same PDCL. Here, the DNA from a different passage was prepared similarly to the Circle-seq DNA, and both DNA samples were subjected to inverse PCR. For the second method, Circle-seq data were generated from seven PDAC PDCLs across two consecutive passages each. A schematic outline of the experimental set-up is provided in Figure 5.19. In short, the PDCLs are cultivated until 70-80% confluency and subsequently split with a 1:3 ratio into a fresh flask. The rest of the cells that are not used for splitting and are subjected to DNA extraction and Circle-seq. When the second passage reaches 70-80% confluency, the cells are harvested and their DNA is extracted for Circle-seq. This methodology generates Circle-seq datasets from cells that are highly related to each other. In both methods, a passage is defined as a transfer of a fraction of the cells into a new cell culture flask.

5.11.1 Candidate eccDNA junctions are not present across different passages

To validate the retention of candidate eccDNA junctions, seven candidates from three PDCLs, PaCaDD137, TKCC-2.1, and TKCC-22, were examined using inverse PCR (Møller et al., 2015). Although eccDNA junctions were validated in the Circle-seq DNA, none were detected in DNA from earlier or later passages of the same PDCLs (Figure 5.20). This absence was observed even in cases where DNA was extracted from consecutive passages (PaCaDD137,

TKCC-22), negating the influence of prolonged cultivation time. Notably, the DNA preparation technique has worked as expected detailed in the retention of circular DNA (*MT-CO1* presence) and linear DNA removal (*HBB* absence).

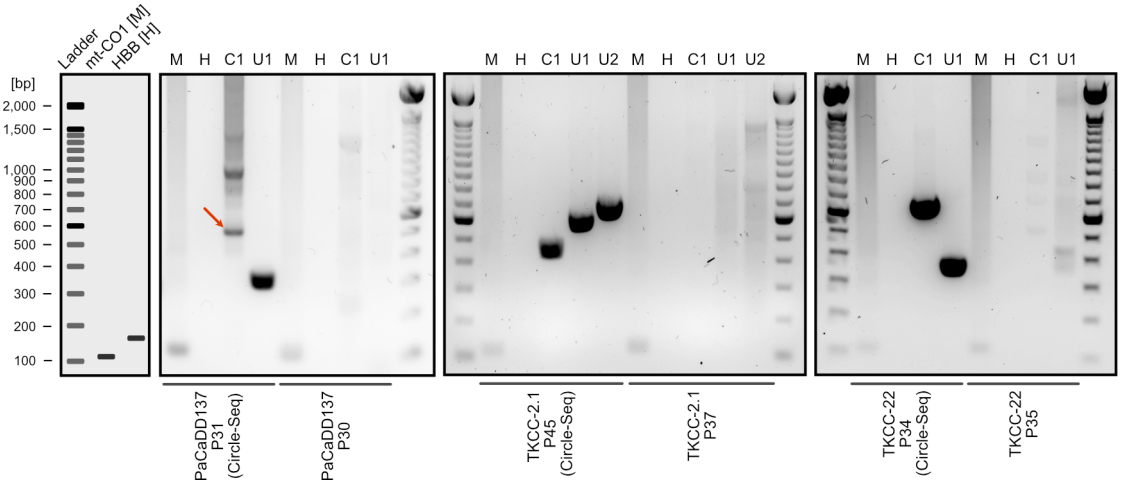


Fig. 5.20 | Validation failure of eccDNA junctions in different passages. Inverse PCR and subsequent agarose gel electrophoresis was performed on a total of seven eccDNA candidate junctions from three individual PDCLs. EccDNA junction was identified in the DNA used for sequencing (left in each gel image; marked with 'Circle-seq'), but was absent in the DNA extracted and enriched for circular DNA from a different passage (right-hand side of each gel image). The corresponding passage number is given below the PDCL name. *MT-CO1* and *HBB* were used as controls for the circular DNA and linear DNA level in the prepared DNA, respectively. Optimal bands of the 100 bp ladder and the two controls are displayed on the left side.

Generally, these preliminary data indicate a lack of eccDNA retention. However, the low-throughput nature of this study limits the generalisability of the findings. Therefore, a high-throughput approach was undertaken to determine if eccDNAs are retained in the PDAC PDCLs.

5.11.2 Circle-seq of two consecutive passages of 7 PDAC PDCLs

As the initial Circle-seq run generated high quality sequences and an initial overview of the eccDNA landscape in PDAC PDCLs. To further expand on open questions, validate some initial findings, and identify eccDNA retention rate, a second Circle-seq dataset was generated. This dataset contained a total of 14 Circle-seq samples from seven PDCLs (2 samples for each PDCL). These samples, extracted from consecutive cell culture passages, were chosen to mitigate potential genetic drift and other long-term culturing effects, in accordance with established guidelines (Hughes et al., 2007; Maitra et al., 2005; Wenger et al., 2004). This approach was also chosen by the variable behaviour of eccDNAs and the uncertainty of eccDNA replication and retention mechanisms (Møller et al., 2018a).

For this Circle-seq study, each PDCL was cultured until reaching 70-80% confluency before being split in a 1/3 ratio, resulting in 1/3 of the cells transferred into a new flask and 2/3 of the cells available for DNA extraction (Figure 5.19). The extracted DNA of all 14 samples was then subjected to the adjusted Henssen et al. (2019a) Circle-seq protocol described in Methods Section 2.10. This protocol was further adjusted for optimal linear DNA removal, by increasing the DNase digestion time to seven days with additional adding of 3 µL DNase

(30 units), 6 µL ATP (25 mM), and 0.9 µL 10x reaction buffer every 24 hours. Following the extended DNase digestion, a fold change removal of at least 200 fold was achieved for each sample, with a median fold change reduction of 1,785.50 (Table 5.4). After further DNA processing and library preparation detailed in Methods Section 2.10 the DNA was then sequenced via NextSeq500, generating around 15 million paired-end reads per sample.

Data processing was carried out using nf-core/circdna (v 1.0.1) and its 'circle_map_realign' branch, which facilitated eccDNA junction identification.

Tab. 5.4 | Fold change decrease of linear DNA compared to circular DNA after DNase digestion of two consecutive passages of seven PDAC PDCLs.

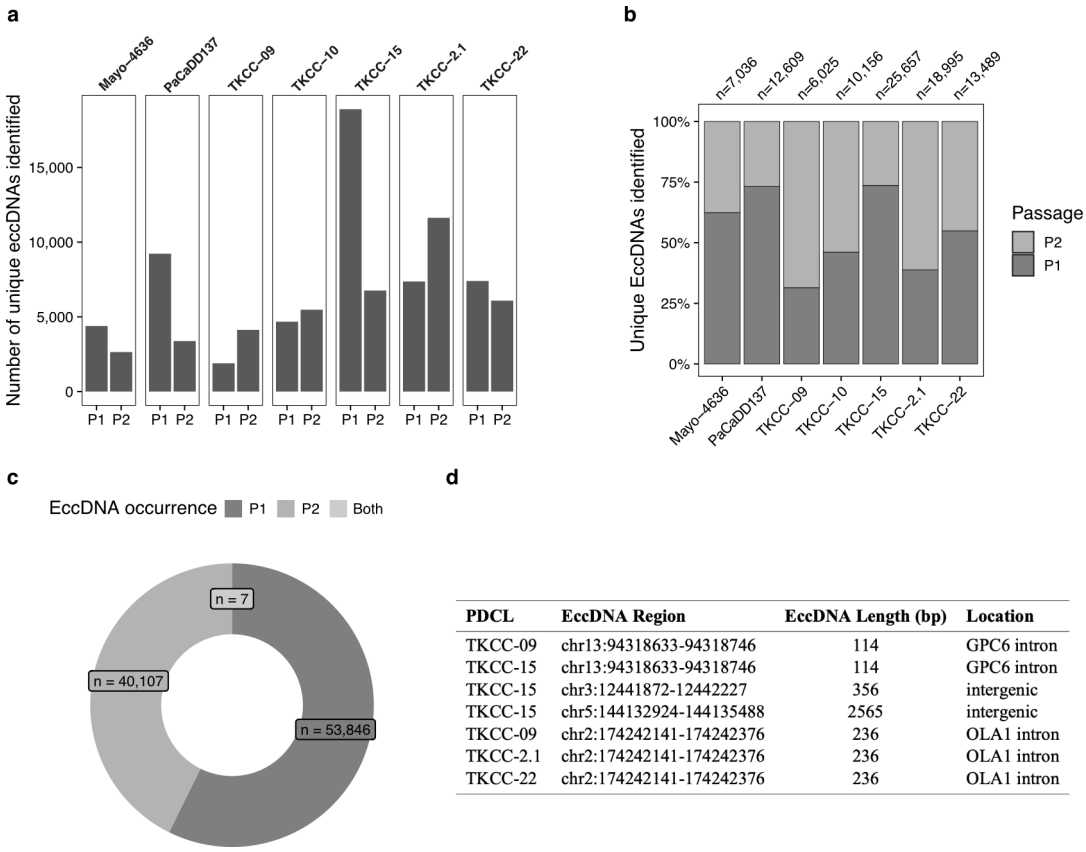
PDCL	Passage #	ID	Linear DNA Decrease (FC)
Mayo-4636	27	P1	1995.0
Mayo-4636	28	P2	224.0
PaCaDD137	32	P1	939.0
PaCaDD137	33	P2	644.0
TKCC-2.1	36	P1	13158.0
TKCC-2.1	37	P2	1923.0
TKCC-09	46	P1	1498.0
TKCC-09	47	P2	1262.0
TKCC-10	28	P1	2762.0
TKCC-10	29	P2	287.0
TKCC-15	42	P1	2112.0
TKCC-15	43	P2	6426.0
TKCC-22	35	P1	3791.0
TKCC-22	36	P2	1648.0

5.11.3 EccDNA Landscape of two consecutive passages

Stringent quality controls (Methods Section 2.14) yielded 93,967 unique eccDNAs across the 14 samples. This number is comprised out of 53,853 and 40,114 unique eccDNAs identified in passage 1 (P1) and passage 2 (P2) PDCLs, respectively (Figure 5.21a). A unique eccDNA is defined as an eccDNA with a unique eccDNA junction in an individual sample. The most eccDNAs were identified in TKCC-15, which had a total of 25,657 different eccDNAs, 18,895 in P1 and 6,762 in P2. The least were identified in TKCC-09, which had a total number of 6,025 unique eccDNAs. Interestingly, the initial Circle-seq analysis also identified the most eccDNAs in TKCC-15 and the least in TKCC-09, suggesting that the number of eccDNAs is dependent on the cell line and the underlying genomics.

By comparing P1 and P2 for each PDCL, the distribution of unique eccDNAs varied considerably across individual PDCLs, highlighting their inherent variability of eccDNA presence in these cell lines (Figure 5.21b). Specifically, TKCC-10 or TKCC-22 exhibited a relatively balanced distribution of eccDNAs across both passages. In contrast, approximately

75% of the total eccDNAs were found in P1 for both PaCaDD137 (73.2%) and TKCC-15 (73.6%), whereas 68.5% were observed in P2 of TKCC-09. To identify eccDNAs that



PDCL	EccDNA Region	EccDNA Length (bp)	Location
TKCC-09	chr13:94318633-94318746	114	GPC6 intron
TKCC-15	chr13:94318633-94318746	114	GPC6 intron
TKCC-15	chr3:12441872-12442227	356	intergenic
TKCC-15	chr5:144132924-144135488	2565	intergenic
TKCC-09	chr2:174242141-174242376	236	OLA1 intron
TKCC-2.1	chr2:174242141-174242376	236	OLA1 intron
TKCC-22	chr2:174242141-174242376	236	OLA1 intron

Fig. 5.21 | EccDNA abundance in two consecutive passages of seven PDAC PDCLs. a, Number of unique eccDNAs in both passages (P1 & P2) of each PDCL. **b**, Proportion of eccDNAs identified in each passage per PDCL. **c**, Proportion of eccDNAs identified in passage 1 (P1), passage 2 (P2), or in both passages of a PDCL. **d**, Table displaying and characterising the seven eccDNAs that were shared in both passages of a PDCL.

exist in both passages of a PDCL, all eccDNAs were compared based on their individual junctions. Only eccDNAs sharing the exact same junction start and end counted. Interestingly, a minuscule fraction (less than 0.01%) of eccDNAs were shared between the two passages of each PDCL. This is deflected by the more than 90,000 individual eccDNAs identified and the seven eccDNAs which are shared in P1 and P2 of an individual PDCL (Figure 5.21c). This suggests that eccDNAs are not typically propagated through cell generations but are likely generated *de novo* and lost during cell cycle progression.

Although Circle-seq offers a snapshot of a large cell population, it is plausible that an increase in sequencing depth might yield additional common eccDNAs between passages. However, the likelihood of a significant increase in shared eccDNAs remains minimal as a high number of eccDNAs is likely already identified at the current sequencing depth (Møller et al., 2015).

Of the seven shared eccDNAs, two arose in TKCC-09, three in TKCC-15, one in TKCC-2.1 and one in TKCC-22 (Figure 5.21d). In contrast, none were shared in both passages of Mayo-4636, PaCaDD137, and TKCC-10. Most shared eccDNAs were short (6 of 7; < 500 bp) and localised in an intron or intergenic regions (7 of 7), which leave their functional role

ambiguous. Therefore it is unclear whether they serve a purpose in the PDCL biology or are just a byproduct of DNA damage or apoptosis (Wang et al., 2021).

Taken together, the comparative analysis of both passages of the seven PDCLs revealed that almost all eccDNAs are unique in an individual passage suggesting that they are not replicated and passed to the following generations, but are lost, expelled, or eliminated rapidly during cell progression.

5.11.4 Shared eccDNAs are identified in other PDCLs

Based on the previous analysis, seven eccDNAs are shared between P1 and P2 of individual PDCLs. Interestingly, two eccDNAs, 'chr13:94318633-94318746' and 'chr2:174242141-174242376', were identified to be shared in more than one PDCL (Figure 5.21d). Similarly, by further investigating the shared eccDNAs, multiple of these were also found in other PDCLs (Table 5.5). This suggests a potential commonality in the genomic breakpoints leading to the formation of eccDNAs. Such commonalities contrasts with existing literature for ecDNAs, which largely describes eccDNA breakpoints as random occurrences (Kim et al., 2020). However, similar eccDNAs might be shared between individuals (Sun et al., 2021).

Tab. 5.5 | PDCL passage shared eccDNAs that are also identified in other PDCLs.

PDCL	Passage	EccDNA Region
Mayo-4636	P1	chr13:94318633-94318746
TKCC-10	P1	chr13:94318633-94318746
TKCC-2.1	P2	chr13:94318633-94318746
TKCC-22	P2	chr13:94318633-94318746
Mayo-4636	P2	chr2:174242141-174242376
PaCaDD137	P2	chr2:174242141-174242376
TKCC-15	P1	chr2:174242141-174242376
TKCC-10	P2	chr3:12441872-12442227

Although random eccDNA formation does have a tendency to occur in transcriptionally active regions or regions with a high GC content (Shibata et al., 2012; Møller et al., 2018a). This introduces the possibility of hotspots within the genome where eccDNA formation is more likely and consequently leading to the shared eccDNAs observed. However, additional validation is necessary in these cases, using inverse PCR or other methodologies. These occurrences may also result from issues with sequence mapping or eccDNA detection.

5.11.5 EcDNAs are replicated and retained during passaging and Gemcitabine treatment

The previous analysis revealed that the majority of eccDNA are not retained during cell line passaging. However, ecDNAs are driving tumour development and tumour adaptation by replication and random segregation (Yi et al., 2022; Kim et al., 2020). Therefore, unlike smaller eccDNAs, large ecDNAs might be differently regulated. To investigate if ecDNAs

alter their structure or disappear during cell passaging or treatment, a PDO, with an existing *MYC*-ecDNA, was examined under normal conditions (VR01-O) and under Gemcitabine treatment (VR01-O-GEM). Here, both samples were subjected to Circle-seq to identify the full eccDNA landscape. Additionally, the passage numbers of these PDOs differed; Gemcitabine treatment began at passage 26, while the original PDO was sequenced at passage 29. Therefore, multiple passages are between those sample states, which enables an assessment of the impact of propagation and treatment on ecDNA structure and presence.

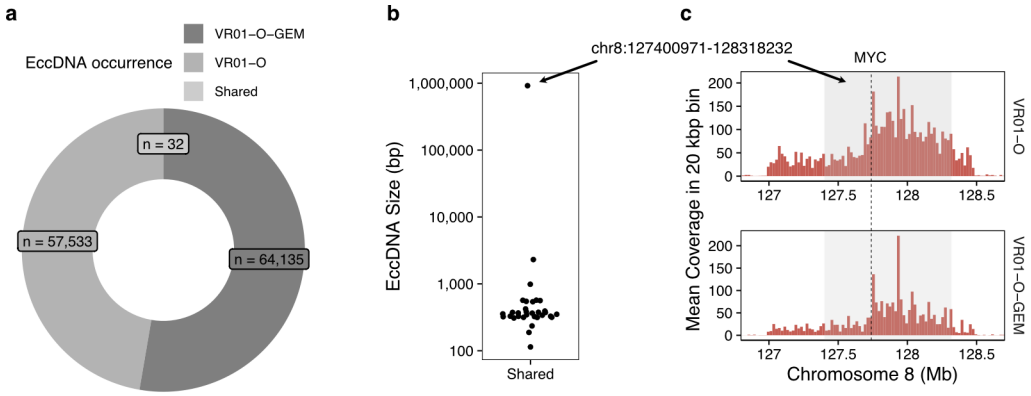


Fig. 5.22 | *MYC*-ecDNA is retained in VR01-O during Gemcitabine treatment and passaging. a, Proportion of eccDNAs categorised as unique to either VR01-O or VR01-O-GEM, or shared between both. **b**, Distribution of sizes for eccDNAs found in both VR01-O and VR01-O-GEM, represented in bp. **c**, Circle-seq coverage view focusing on the *MYC* locus on chromosome 8. A specific region, chr8:127400971-128318232, shared between VR01-O and VR01-O-GEM is highlighted in red. The position of the *MYC* gene is marked by a dashed line. The Y-axis represents the mean sequence coverage calculated in 20 kbp bins.

EccDNA junction analysis in both PDO samples revealed a substantial number of eccDNAs (Figure 5.22a). Compared to PDCLs, VR01-O and VR01-O-GEM had at least three times as many eccDNA junctions to any PDCL (Figure 5.21a). In total, 121,732 eccDNAs were identified in both PDOs - 64,167 in VR01-O-GEM and 57,565 in VR01-O. Out of the large number of eccDNAs identified in both samples, only 32 were common in both. This highlights that, like in the PDAC PDCLs, most eccDNAs are not retained during PDO propagation.

Again, almost all shared eccDNAs have a size of less than 5,000 bp (Figure 5.22b). Interestingly, the only shared eccDNA with a length greater than 5,000 was identified to be an ecDNA containing *MYC*. This ecDNA had a length of 917,262 bp and showed similar Circle-seq coverage distributions in both VR01-O and VR01-O-GEM, revealing that the ecDNA breakpoints did not differ and the same ecDNA is identified in both samples (Figure 5.22).

Earlier WGS data analysis revealed that the VR01-O *MYC*-ecDNA is most likely comprised of multiple chromosomal fragments and encompasses a wider region (Figure 4.21a). However, due to the limitations of the used eccDNA detection software, the complete structure can not be identified using Circle-seq. This shows that secondary validation methods need to be used to fully validate a complete structure of an eccDNA.

In summary, the data shows that larger eccDNAs, which are also detectable by other techniques like FISH or WGS, are retained and replicated during extended passaging and are also not eliminated by Gemcitabine treatment. In contrast, shorter eccDNAs appear to be

transient.

5.12 EccDNA-formation hotspots in PDAC

The genesis of eccDNA is heavily debated. Some argue for the significance of eccDNA hotspots, while others describe their random occurrence (Wang et al., 2021; Koche et al., 2020; Møller et al., 2015; Møller et al., 2018a). Our previous analysis has shown that the identified eccDNAs in our PDAC PDCLs did not follow a random distribution but carried or overlapped specific genome elements. Furthermore, it seems that they are also associated with increased chromatin accessibility or transcriptional activity. While eccDNAs are mostly too small to carry full genes, their existence might still have an effect on cancer biology. Therefore, identifying specific hotspots of increased eccDNA formation is of vast importance for the eccDNA field.

While, mostly two Circle-seq datasets were examined in the previous Circle-seq analyses, a third dataset originated from seven PDAC patient-derived organoid (PDO) samples, collected from six individual PDAC PDOs, was generated in the volume 2 Circle-seq study (Table 2.2). These three distinct datasets provide a unique opportunity to investigate the occurrence of eccDNA-formation hotspots across different PDAC models.

5.12.1 Identification of common eccDNA hotspots

To explore eccDNA hotspots in PDAC, an integrative analysis was performed of the three Circle-seq dataset comprised of eight PDCLs, seven PDCLs across two consecutive passages (7 PDCLs x 2 Passages), and seven PDOs. EccDNA 'hotspots', 'coldspots', and regions with 'normal' eccDNA abundance were identified using a permutation-based method. Specifically, the eccDNA regions for each dataset were randomly permuted alongside their original chromosomes. This permutation was performed 1,000 times generating a 1,000 datasets containing random eccDNA origins for each Circle-seq dataset. Subsequently, the genome was divided into approximately 1 Mbp bins and the number of eccDNAs and random eccDNAs in each bin was counted. Bins with more eccDNA counts compared the random datasets (P value < 0.001) were classified as 'hotspots'. Conversely, bins with fewer eccDNAs were classified as 'coldspots' (P value < 0.001). Bins with comparable counts (P value > 0.001) to the random datasets were designated as 'normal' regions. For more details on the methodology refer to Methods Section 2.20.

With this analysis, over 300 hotspots and coldspots were identified across the three Circle-seq datasets (Figure 5.23a,b). Despite variability between the datasets, many regions appeared as either hotspots or coldspots in multiple datasets. However, only six hotspots were identified in all three datasets (consensus eccDNA hotspots), showing that eccDNA hotspots may vary between the analysed cells, the underlying model system, or are data specific. In contrast, coldspots showed considerable overlap, particularly in the acrocentric chromosomes 13 and

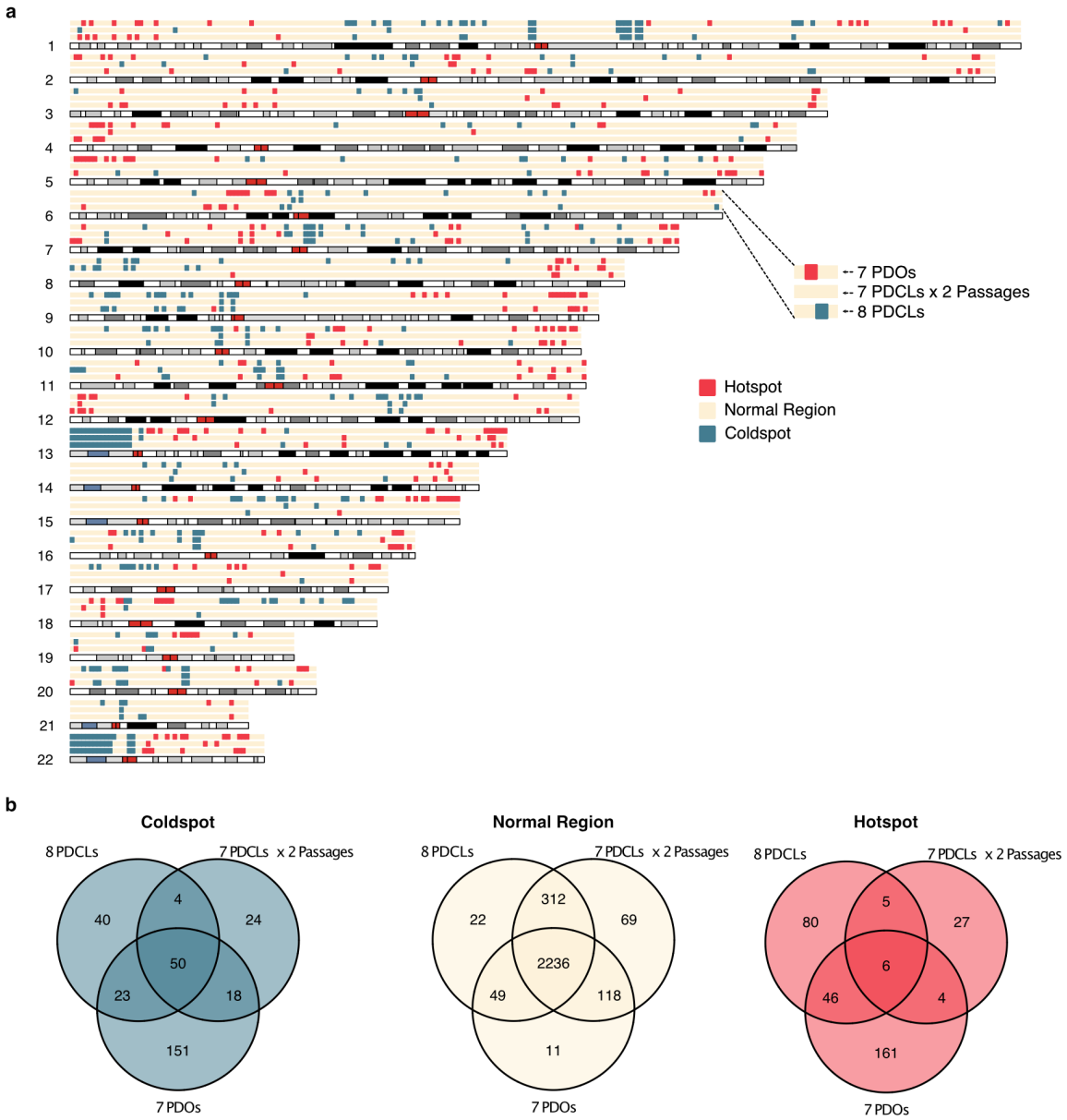


Fig. 5.23 | EccDNA Hotspot identification in the three Circle-seq datasets. a, EccDNA hotspots, coldspots, and normal regions were identified based on the number of eccDNAs in a 1 Mbp region compared to random generated datasets (Methods Section 2.20). The 1 Mbp bins are coloured based on their region class and located based on their genomic location. The three Circle-seq datasets are individually visualised (see figure legend). **b**, Venn diagram showing the overlap of coldspot (blue), normal (beige), and hotspot (red) regions in all three datasets.

22 and centromeric regions. This shows that eccDNA hotspots might be variable, but eccDNA coldspots can be universally identified, suggesting common regions where eccDNA formation is absent.

While only a few hotspots are identified in all three Circle-seq datasets, many are identified in at least two of those. To further expand on different eccDNA regions and their characteristics, recurrent hotspots, coldspots, and normal regions are further analysed which occur in at least two Circle-seq datasets. This generated a list of common regions containing 95 coldspots, 2,715 normal regions, and 61 hotspots (Figure 5.23b and Figure 5.24).

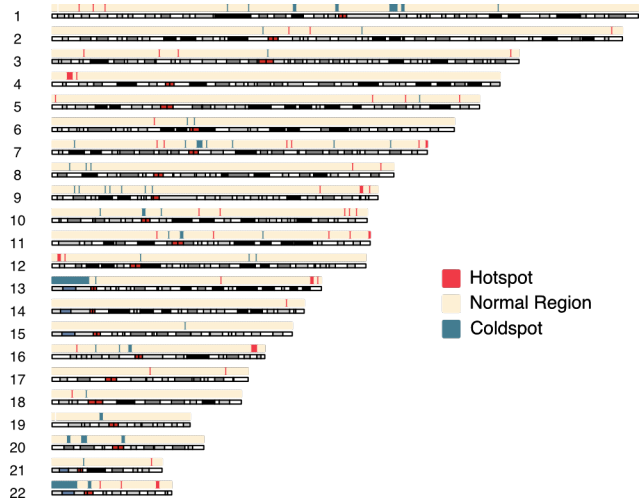


Fig. 5.24 | Genomic view of common eccDNA hotspots, coldspots, and normal regions. A common region (hotspot, coldspot, or normal) is defined as being identified twice in the three Circle-seq datasets, 8 PDCLs, 7 PDCLs x 2 Passages, and 7 PDOs.

5.12.2 Common eccDNA hotspots are gene-dense and located in specific genomic regions

The analysis of the common eccDNA hotspots and coldspots revealed patterns in their distributions, indicating that these regions are not randomly spread across the genome (Figure 5.24). Instead, a further computational analysis shows that eccDNA hotspots are located preferentially in Giemsa-negative (gneg) stained regions of the genome and are found in proximity to chromosome ends (telomeres) (Figure 5.25a,b). Giemsa-negative areas, contrary to Giemsa-positive bands (gpos), are gene-rich and associated with increased transcriptional activity (Grewal & Jia, 2007; Sumner, 1982; Fungtammasan et al., 2012; Laird et al., 1987; Furey & Haussler, 2003). With regards to the eccDNA coldspots, these regions are dispersed throughout the chromosomes, but show an increase in or around centromeres (Figure 5.25b). In addition, a large proportion of coldspots overlap with Giemsa-positive regions, especially regions with the darkest Giemsa staining bands, gpos100 or gpos75. These findings underscore the non-random nature of eccDNA hotspots and association with specific genomic characteristics such as the location in sites of active transcription.

Further analysis of these regions revealed significant correlations between eccDNA hotspots and specific genomic elements. A comparison to the normal regions and their abundance of genomic features revealed that eccDNA hotspots and coldspots displayed opposing associations with genomic feature. While hotspots were enriched with almost all genomic elements analysed, coldspots are sparsely populated by those (Figure 5.25c). In particular, coldspots are gene-sparse, and have low abundance of repeat elements and enhancer sequences. This scarcity is consistent with their proximity to centromeres or their association with heterochromatin as those regions are defined as gene-sparse and low of other genomic features (Sumner, 1982; Grady et al., 1992; Miga, 2020).

In contrast, eccDNA hotspots demonstrate a significant enrichment of various genomic

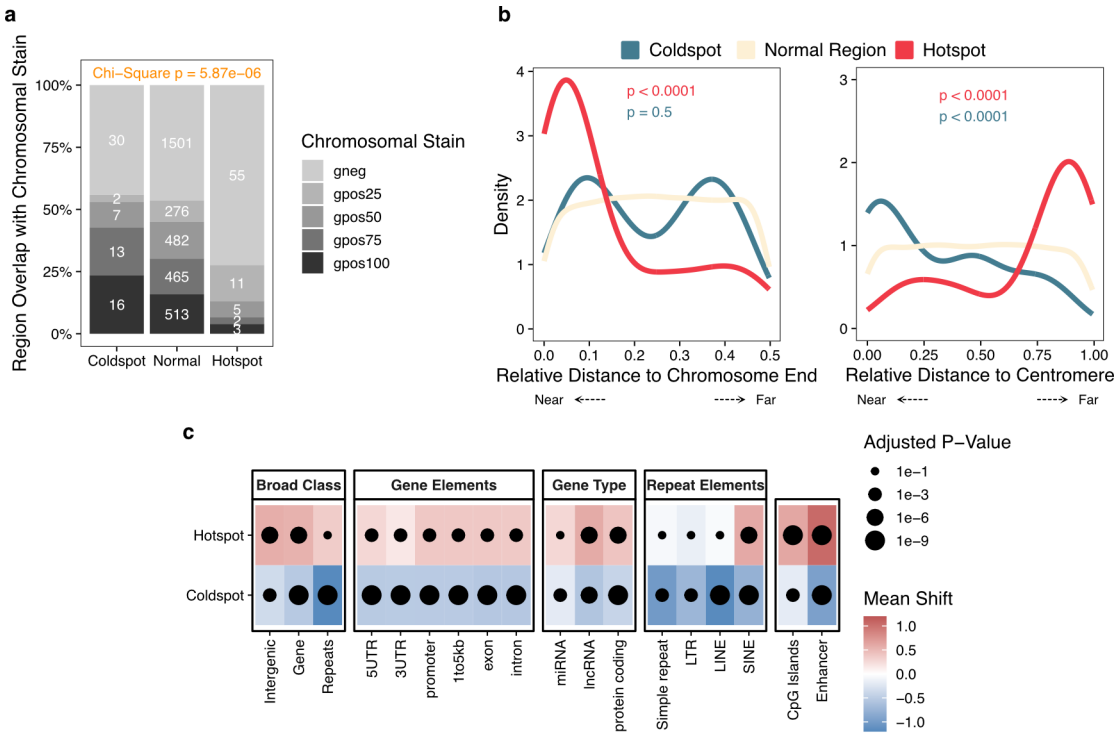


Fig. 5.25 | Common eccDNAs hotspots contain specific genomic elements. **a**, Cytoband staining patterns at common regions. gneg, Giemsa-negative; gpos, Giemsa-positive. **b**, Density plot of relative distances to chromosome ends or centromeres of common regions. A Kolmogorov-Smirnov test was used to calculate a significant difference between the distributions identified in the hotspots (red) and coldspots (blue) compared to the normal regions (beige). **c**, Median shift of the normalised counts of specific genomic features within hotspots and coldspots compared to normal regions. Statistical significance is assessed via Wilcoxon rank-sum tests, with *P* values adjusted by the Benjamini-Hochberg method.

elements. Notably, hotspots exhibit a high gene density for all three gene types, protein-coding, miRNAs, and lncRNAs. Similarly, all individual gene elements are enriched in eccDNA hotspots. Furthermore there is a modest enrichment of genomic repeat elements in eccDNA hotspots compared to normal regions. Upon further examination, the sub-classification of repeat elements reveals a strong enrichment of SINEs. Interestingly, the other three repeat subclasses, namely LINEs, long-terminal repeats (LTRs), and simple repeats, appear to be slightly less prevalent on eccDNA hotspots compared to normal regions (Figure 5.25d). Lastly, eccDNA hotspots show a massive enrichment for CpG islands and enhancers.

In conclusion, these findings provide evidence that eccDNA hotspots exhibit a high density of diverse genomic elements and are predominantly located in regions of open chromatin. These observations align with previous studies, which identified Giemsa-negative regions as having the highest gene, CpG and SINE density (Furey & Haussler, 2003; Gilbert et al., 2004).

5.12.3 EccDNA hotspots are associated with increased gene expression and chromatin accessibility

Basic annotation of eccDNA hotspots with genomic properties revealed the association with open chromatin regions and specific genomic elements. Giemsa-negative regions are considered to be gene-dense with an increased transcriptional activity (Furey & Haussler,

2003; Morrison & Thakur, 2021). An eccDNA study by Dillon et al. (2015) also noted that small eccDNAs (microDNAs) abundance might be linked to transcription and transcriptional activity (Dillon et al., 2015). To identify if the eccDNAs hotspots are also associated with increased gene expression, an integration with gene expression data of three PC datasets was performed. These three datasets were chosen based on the model system (Brunton et al. (2020), $n = 48$) and the large cohorts (TCGA PAAD, $n = 150$; ICGC PACA-AU, $n = 269$), which display the extensive transcriptomic landscape of PC.

Across all three analysed datasets, a consistent pattern emerged regarding the gene expression levels in relation to eccDNA hotspots, coldspots and normal regions. Specifically, the coldspot regions consistently exhibited significantly lower average gene expression compared to normal and hotspots regions, supporting that eccDNAs are mainly coming from regions with active transcription (Figure 5.26). In contrast, eccDNA hotspots exhibited significantly increased gene expression in the TCGA and ICGC dataset. However, in the dataset by Brunton et al. (2020), comprised of 48 PDAC PDCLs, including some also used for Circle-seq, a significant difference in gene expression was not observed. Despite the discrepancy, the overall trend of a near step-wise increase of gene expression from coldspots, over normal regions, and ultimately to hotspots was consistently observed in all three datasets. These findings underscore the strong association between the abundance of eccDNAs and transcriptionally active genomic regions.

Open chromatin is associated with increased chromatin accessibility for transcription factors, facilitating the activation and initiation of transcription factors (Klemm, Shipony & Greenleaf, 2019). To investigate the epigenetic characteristics and chromatin accessibility of the identified common regions, additional analysis was performed using ATAC-seq data of 9 PDAC PDCLs and methylation profiles from 24 PDAC patient-derived tumour xenografts.

Consistent with the gene expression analysis, a similar step-wise increase is observed in the chromatin accessibility and methylation score across the coldspots, normal regions, and hotspots (Figure 5.26). Coldspots exhibited the lowest values, while hotspots displayed the highest. The ATAC-seq analysis provided the further support for the association between eccDNA hotspots and increased chromatin accessibility, reinforcing our previous findings. This suggests that regions with higher eccDNA abundance are defined by a more accessible chromatin, potentially facilitating the transcriptional activity of genes.

In contrast, the analysis of the methylation profiles revealed that hotspots also display increased methylation levels at CpG islands, genes, and enhancer sites. Notably, hypermethylation of CpG islands is often associated with gene silencing (Bird, 2002). Therefore, an inverse association between methylation and gene expression would be expected. Unfortunately, the RNA-seq data for the 24 patient-derived xenografts analysed by Lomberk et al. (2018) was not openly available. Therefore, it is unclear whether the increased methylation levels did indeed affect gene expression or if the data would support earlier findings showing increased

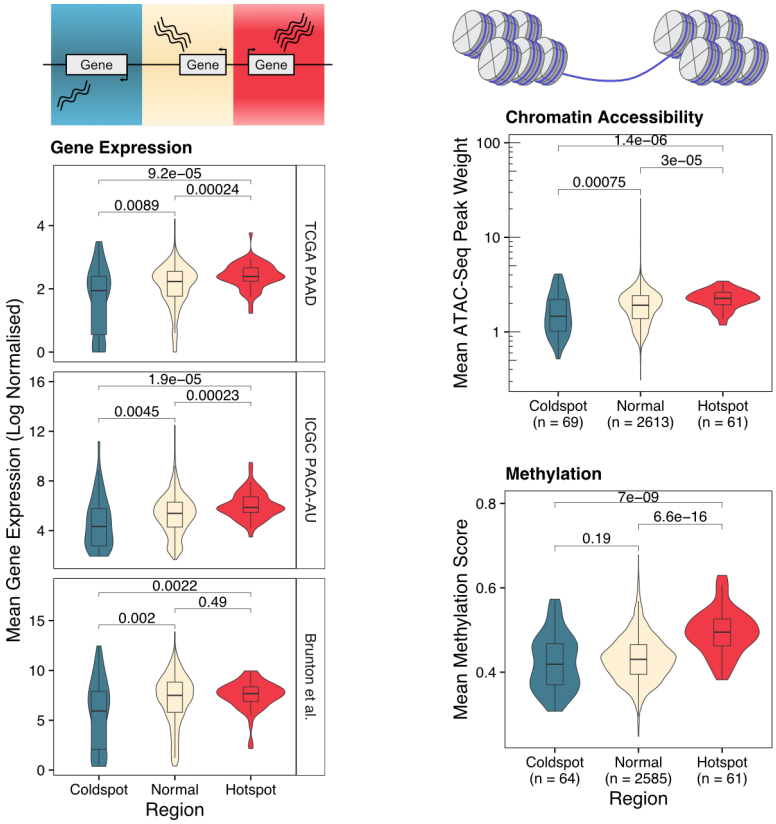


Fig. 5.26 | Increased gene expression and chromatin accessibility in common eccDNA hotspots. a, The common eccDNA coldspots, normal regions, and hotspots were integrated with gene expression data from the TCGA PAAD (n = 150), ICGC PACA-AU (n = 269), and the Brunton et al. (2020) PDAC PDCL (n = 48) data. **b**, Epigenomic characteristics at the different regions was assessed using the ATAC-seq dataset of PDAC PDCLs (n = 9, two biological replicates each, Brunton et al. (2020)) and PDAC methylation profiling data (n = 27, Lomberk et al. (2018)). **a & b**, Statistical analysis was performed using a Wilcoxon-rank sum test.

transcription at eccDNA hotspots. These findings raise intriguing questions about the functional implications of increased methylation at eccDNA hotspots. Further investigation is needed to support or refuse these observations.

Taking together, the integrative analysis revealed distinct characteristics associated with eccDNA coldspots and hotspots. Coldspots were found to be associated with low transcriptional activity and a condensed chromatin state. In contrast, hotspots exhibited high gene expression levels and an open chromatin structure. These findings align with previous research by Møller et al. (2018a), which identified a correlation between increased transcription of specific genes and the abundant occurrence of eccDNAs. Collectively, these findings enhance our understanding about the occurrence of eccDNA hotspots, highlighting the frequent occurrence of eccDNAs in regions of active transcription and open chromatin.

5.12.4 Consensus eccDNA-formation hotspots

The comprehensive eccDNA hotspot and eccDNA origin analysis revealed that eccDNAs predominantly arise in regions characterised by active transcription and open chromatin architecture. In the hotspot analysis, recurrent eccDNA hotspots, coldspots, and normal regions were identified. However, only six of the more than 60 recurrent eccDNA hotspots are identified hotspots in all three distinct Circle-seq datasets (consensus eccDNA hotspots,

Figure 5.27). Those six are located on chromosome 3, 7, 8, 12, 13, and 18 (Table 5.6). Many genes were identified to be located within these hotspots, with some of them having cancer driver properties. However, it is unclear why those hotspots were universally identified as no significant difference has been found in the number of genomic features compared to recurrent hotspots (Table B.2).

Tab. 5.6 | Genes identified within consensus eccDNA hotspots. Cancer driver genes from the allOnco list (www.bushmanlab.org) are highlighted in bold.

Hotspot	Genes
chr3:194289588-195291080	*ACAP2-IT1, ATP13A3, CPN2, FAM43A, GP5, LINC00884, LINC00887, LINC01968, LINC01972, LRRC15, LSG1, MIR3137, RN7SL36P, RNU6-1101P, RNU6-25P, RPL23AP93, TMEM44, TMEM44-AS1,* ***XXYLT1****, XXYLT1-AS1, XXYLT1-AS2*
chr7:47102269-48104444	*C7orf57, C7orf65, C7orf69,* ***HUS1****, LINC00525, LINC01447, PKD1L1, SUN3, TNS3*
chr8:127121426-128122381	*CASC11, CASC8, CCAT2, MIR1204, MIR1205, MIR1206, MIR1207,* ***MYC****, POU5F1B, RNU4-25P, RNVU1-32, TMEM75*
chr12:2004141-3006210	*CACNA1C-AS1, CACNA1C-AS2, CACNA1C-AS3, CACNA1C-AS4, CACNA1C-IT1, CACNA1C-IT2, CACNA1C-IT3, CBX3P4, FKBP4,* ***FOXM1****, IQSEC3P1, ITFG2, ITFG2-AS1, ITFG2-AS1, LINC02371, NRIP2, RHNO1, RNU6-1315P, RPL23AP14, TEX52, TULP3*
chr13:110351545-111354739	*ANKRD10, ANKRD10-IT1, ARHGEF7, ARHGEF7-AS1, ARHGEF7-AS2, ARHGEF7-IT1, CARS2, COL4A2-AS1, COL4A2-AS2,* ***ING1****, LINC00368, LINC00431, LINC00567, NAXD, PARP1P1, PRECSIT, RAB20, RPL21P107, TEX29*
chr18:8037329-9041994	*AKR1B1P6, COP1P1, GACAT2, MTCL1, RAB12, RN7SL50P, RPS4XP19, THEMIS3P, TOMM20P3*

5.12.5 *MYC* is located inside a consensus eccDNA hotspot

An in-depth investigation of the hotspots and their gene contents revealed that *MYC* is located within one of these consensus eccDNA hotspots (Figure 5.27). None of the remaining five hotspots contained any other known PDAC driver genes.

MYC is an important driver of PDAC progression, regulating many aspects of PDAC biology (Sodir et al., 2020; Maddipati et al., 2022). Our analysis added that *MYC* is also recurrently amplified on ecDNAs in PDAC PDOs and in their respective primary tumours. A focused examination of chromosome 8 revealed that the *MYC* locus and its adjacent region collectively harboured 566 eccDNAs, vastly exceeding eccDNA counts in other regions of the chromosome (Figure 5.28a). Interestingly, the upstream region to the *MYC* eccDNA hotspot contained the second most eccDNAs (n = 267), indicating a localised increase in eccDNA formation around the *MYC* locus (Figure 5.28b). In contrast to these two regions, other regions on chromosome 8 mostly display only modest numbers of eccDNAs.

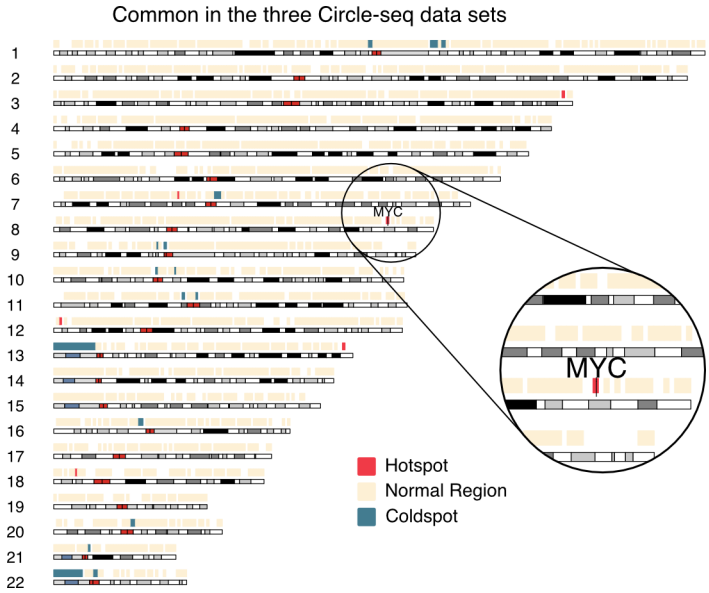


Fig. 5.27 | Consensus eccDNA hotspots highlight the *MYC* locus. Genomic view of EccDNA hotspots, coldspots, and normal regions universally identified across all three Circle-seq. Regions that showed dissimilar region class in the three Circle-seq datasets are absent. The *MYC* locus is focused revealing its location inside an eccDNA hotspots.

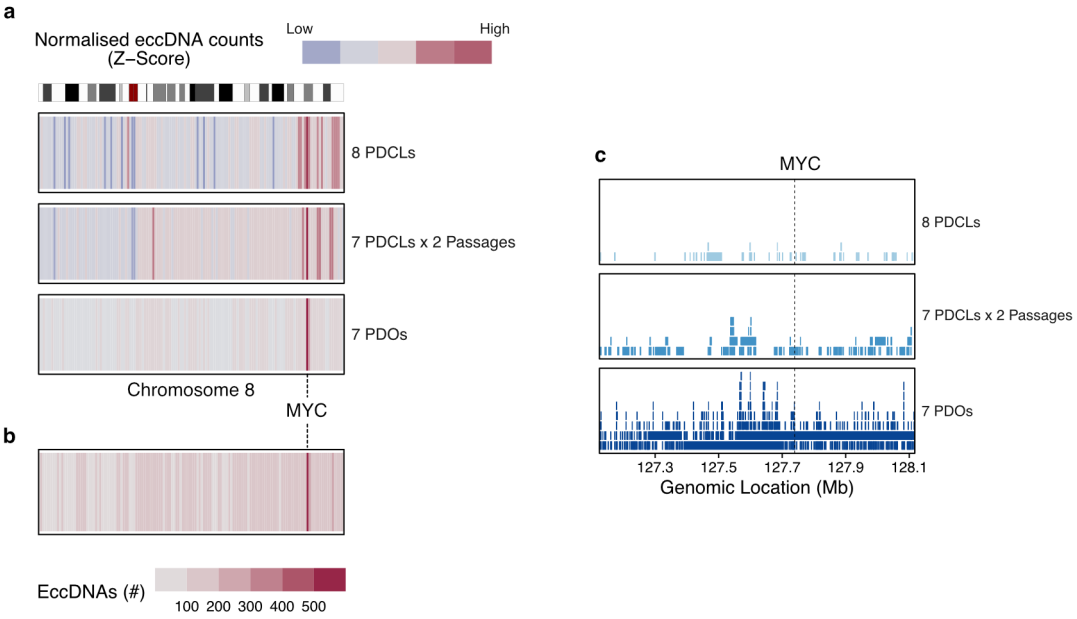


Fig. 5.28 | EccDNA abundance on Chromosome 8. **a**, Z-score normalised eccDNA counts of all three Circle-seq datasets on chromosome 8. Low z-score values define low eccDNA levels and high values define high eccDNA levels in a distinct dataset. Z-scores were calculated for each dataset individually and simplified for display. The original graphs containing the accurate legends are visible in Figure B.3. **b**, Total number of eccDNAs identified in all three Circle-seq datasets. **a & b**, The number of eccDNAs are counted in a 1 Mbp bin similar to the eccDNA hotspot analysis. **c**, EccDNA locations in and around *MYC* locus. Regional view of eccDNA hotspots containing *MYC* and all the eccDNAs identified in the three distinct Circle-seq datasets. *MYC* location is highlighted with a dashed line.

Further scrutiny of the *MYC* eccDNA hotspot revealed that large proportion of eccDNAs were identified in the seven PDOs (Figure 5.28c). Notably, the most eccDNAs are also identified in the seven PDOs dataset, which could influence the number of eccDNAs in this locus (Figure B.4). While *MYC* is located inside the named eccDNA hotspots, only a few eccDNAs directly overlap with the complete *MYC* gene. One prominent *MYC*-comprising eccDNA is identified in VR01-O, which is described earlier. Other eccDNAs are located

around the *MYC* gene or comprise only *MYC* gene elements.

To summarise, *MYC* is not only recurrently identified on ecDNAs in various cancer types and in our PDAC PDOs, it also is located inside an eccDNA hotspot which comprises plenty of eccDNAs with various sizes (Kim et al., 2020; Hung et al., 2021). This suggests that the formation of eccDNAs and the uprising of ecDNAs might be associated. While the characteristics of eccDNAs are largely different, it is debated whether small eccDNAs can give rise to large ecDNAs by steadily incorporating chromosomal fragments or other eccDNAs (Schimke et al., 1986; Carroll et al., 1988). While ecDNAs are rarely identified in our PDAC samples, the eccDNA abundance around the *MYC* locus could be an indicator to the role of *MYC* and its surrounding genes in eccDNA formation in PDAC by impediment of replication forks (Watanabe et al., 2017). In a study by Koche et al. (2020), neuroblastoma tumours were investigated and *MYCN* ecDNAs were recurrently identified. Additional Circle-seq studies verified the *MYCN* ecDNAs, but also uncovered the eccDNA landscape of the neuroblastoma tumours. While the eccDNAs spread throughout the genome, a major number of those are located around *MYCN*, which is recurrently amplified in neuroblastoma (Koche et al., 2020; Huang & Weiss, 2013). This suggests that eccDNA formation hotspots might be cancer-type specific and might be located around common sites of ecDNA origin.

5.13 Discussion

For decades, large amplified eccDNAs (ecDNAs) have been studied and are well-understood in their role of influencing the cancer biology. Most roles are based on the genomic content of the eccDNAs, but smaller eccDNAs may not not contain any gene elements, and it is unclear how they can influence the cancer biology (Carroll et al., 1988; Cowell, 1982; Wu et al., 2019; Kim et al., 2020; Ling et al., 2021; Wu et al., 2022a; Møller et al., 2018a). Also significant differences are reported between eccDNA and ecDNA landscapes. The former can be present in healthy and cancerous cells, while the latter is mostly found in cancer cells (Kim et al., 2020; Møller et al., 2018a; Koche et al., 2020; Paulsen et al., 2018). In our investigation utilising several Circle-seq datasets, it emerged that eccDNAs are abundantly present in PDAC model systems and originate non-randomly, contradicting the results of random biogenesis via cell apoptosis (Wang et al., 2021). The initial Circle-seq analysis, comprising eight PDAC PDCLs, demonstrated the detectability of eccDNAs in every PDAC PDCL in differing degrees. TKCC-15 exhibited the most, while TKCC-09 exhibited the least eccDNAs within the first dataset. This finding was further supported by the second Circle-seq investigation of seven PDAC PDCLs, which included two consecutive passages each. The results indicate that the amount of eccDNAs in a sample may be influenced by the genomic background, as eccDNAs are produced by various forms of DNA damage. It is likely that unstable genomes produce more eccDNAs, highlighting the potential to utilise eccDNA abundance as a biomarker (Cohen, Regev & Lavi, 1997; Paulsen et al., 2021). However, the study has limitations as genomic data integration was not performed, which leaves room for

further research. Additionally, in some instances, there was a significant variation in eccDNA levels between passages in the PDAC PDCLs. Therefore, it may be essential to increase the sample size and conduct future eccDNA characterisation with multiple biological replicates to consider the variability.

In accordance with a study in neuroblastoma, Circle-seq analysis identified thousands of eccDNAs, which are mostly small, and do not typically contain complete genes (Koche et al., 2020). Additionally, their size distribution strongly overlaps with a study on eccDNAs as an apoptotic byproduct (Wang et al., 2021). The study by Wang et al. (2021) revealed that eccDNAs originate randomly from all parts of the genome. However, our results show that the eccDNAs are not completely random as specific genomic elements are prominently enriched on eccDNAs. Although apoptosis could be a contributing factor to the formation of certain eccDNAs, complete random biogenesis is not observed in the PDAC PDCLs.

Notably, some of the genes identified on eccDNAs possess cancer-driving properties. Therefore, elevated transcription of those, resulting from eccDNA-based amplification, could promote tumour progression. Functional small regulatory RNAs that can modulate gene expression were found to be expressed by incomplete genes on eccDNA (Paulsen et al., 2019). As such, eccDNAs lacking complete genes also hold potential to impact PDAC biology. However, a definitive conclusion about their influence cannot be drawn from the present data.

Carrying complete genes, particularly cancer-specific oncogenes, is a defining characteristic of ecDNAs (Luebeck et al., 2023; Kim et al., 2020). These ecDNAs are typically over 100 kbp in length, significantly amplified, and present in a large proportion of cancerous cells (Turner et al., 2017; Yi et al., 2022). However, no ecDNAs with ecDNA-like properties were identified in the PDCLs of PDAC. Analysis of an ecDNA-carrying PDO (VR01-O) demonstrated that Circle-seq has the ability to identify ecDNAs. This suggests that ecDNAs either do not exist in the PDAC PDCLs, or they are too complex to be detected using the current methods. EcDNAs have the ability to create complex structures that integrate fragments from distal regions. However, the techniques applied can only detect single-fragment eccDNAs or fully covered multi-fragment eccDNAs (Koche et al., 2020; Shoshani et al., 2021). WGS been demonstrated to be an optimal data type for the analysis of ecDNA. The PDAC PDCLs utilised in this investigation were previously subjected to WGS and are accessible for analysis if desired (Dreyer et al., 2021). Nonetheless, in this dissertation, the data was not obtained nor examined, thereby presenting an opportunity for the further identification of ecDNA-carrying models.

As multi-fragment ecDNAs can have complex structures that do not always consist of a single DNA fragment, a novel approach was employed to detect them from short-read sequencing data (Wang et al., 2021; Møller et al., 2018a; Koche et al., 2020). According to a recent study using long-read sequencing, the *de novo* assembly of eccDNA has shown that about 10% of eccDNA come from multiple fragments, which is also supported by our

findings (Wang et al., 2021). Furthermore, our results show that multi-fragment eccDNA have larger sizes than single-fragment eccDNA, emphasising the need to identify complex eccDNA. This methodology can be employed to achieve a more comprehensive understanding of the eccDNA landscape, thereby potentially revealing additional gene-containing eccDNA. Nevertheless, *de novo* assembly is restricted by the computational resources required and its reliance on the uninterrupted coverage of an eccDNA. During the Circle-seq procedure, larger eccDNAs exhibited lower levels of amplification in comparison to their smaller counterparts. This phenomenon can be attributed to the bias of the phi29 amplification towards small and more abundant eccDNAs (Norman et al., 2014). As a result, it remains unclear whether high coverage sequencing can guarantee accurate identification of all eccDNAs. Alternative tools, such as AmpliconArchitect (Deshpande et al., 2019) and the recently developed Circlehunter (Yang et al., 2023), may be a superior choice, in the long term, as they can identify regions with enriched sequencing reads and connect eccDNA regions by utilising split and discordant reads. Nevertheless, these tools are not optimised for smaller eccDNAs and Circle-seq data. To address these issues, the use of long-read sequencing data is preferred for eccDNA research. This approach allows for the detection of eccDNA origins, complex structures, complete sequences, and eccDNA breakpoints (Li et al., 2023; Wanchai et al., 2022; Koche et al., 2020; Wang et al., 2021).

This investigation of different PDAC model systems demonstrated that practically every part of the genome can add to the formation of eccDNAs. These eccDNAs do not arise randomly but mainly stem from eccDNA hotspots, which are typically found surrounding genes and close to chromosome ends. The hotspots contain a high abundance of genes, SINEs, and CpG islands, and are enriched with enhancers (Figure 5.29). Furthermore, they demonstrate an open chromatin and are active transcription sites. Much of this has already been identified in other cells, both normal and cancerous (Møller et al., 2015; Møller et al., 2018a; Koche et al., 2020; Dillon et al., 2015). The emerging pattern of eccDNA biogenesis suggests an association with increased transcription, based on multiple analyses. This finding supports the hypothesis that active transcription may lead to increased DNA damage caused by elevated recombination rates or R-loop formation. R-loops are hybrids made of DNA and RNA and expose single-stranded DNA. However, R-loop formation is particularly prevalent at 5’ and 3’ untranslated regions (UTR), which are similarly enriched on extrachromosomal circular DNA (eccDNA) as other genic elements (Thomas & Rothstein, 1989; Skourti-Stathaki & Proudfoot, 2014; Dillon et al., 2015). Transcription-based R-loop formation may lead to the biogenesis of eccDNA, although it is unlikely to be the sole contributor. Shibata et al. (2012) observed a distinct enrichment of the 5’UTR of microDNAs, which could be generated by DNA-RNA hybrids (Shibata et al., 2012). Therefore, microDNAs smaller than 500 bp may differ in their genomic origins from larger eccDNAs and should be investigated further. In yeast, it has been discovered that an increase in gene transcription leads to a rise in double-strand break formation and a subsequent increase in eccDNA accumulation around the gene’s location (Hull et al., 2019). The analysis of eccDNA contents has revealed that

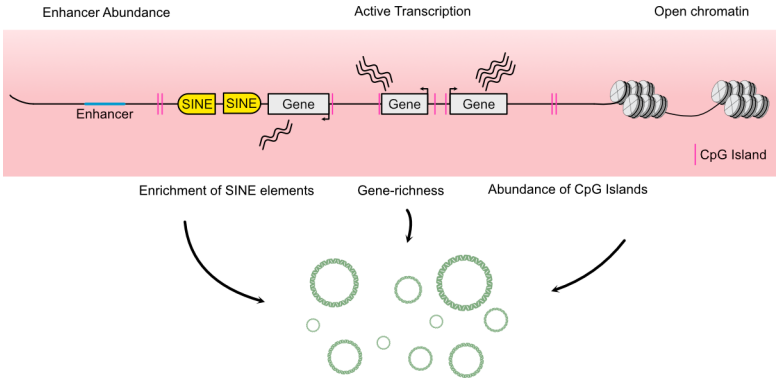


Fig. 5.29 | Identified genomic characteristics associated with eccDNAs and eccDNA formation hotspots. Schematic representation of genomic elements and characteristics identified within eccDNA hotspots or overlapping with eccDNAs.

transcriptionally active genes, including cancer-driver genes, and open chromatin regions are enriched on eccDNAs. This implies that transcription is a major factor driving eccDNA formation in the PDAC PDCLs. However, the mechanism underlying formation of eccDNAs remains unclear, thus necessitating further investigation.

The hotspot analysis showed that the *MYC* locus is identified in an eccDNA hotspot across all three Circle-seq datasets. The *MYC* gene has recurrently been discovered on ecDNAs, occurring in two of twelve ecDNA-positive PDOs, and is implicated in shaping tumour biology. *MYC* is a crucial driver of PDAC associated with the Squamous subtype and the formation of metastases. Its significance in the progression of PDAC cannot be overstated (Sodir et al., 2020; Bailey et al., 2016; Maddipati et al., 2022). Building on previous research in neuroblastoma, it appears that the *MYCN* locus featured significant eccDNA coverage in neuroblastoma, leading to the hypothesis that there may be cancer-type-specific eccDNA formation hotspots associated with the cancer gene expression (Koche et al., 2020). This may reveal the reason for the formation of ecDNAs containing drug resistance genes inside cells when exposed to drugs: increased transcription of drug resistance genes induces ecDNA and eccDNA formation (Kaufman, Brown & Schimke, 1979; Singer et al., 2000).

EcDNAs replicate and are inherited by the daughter cells during the cell cycle (Yi et al., 2022). It remains unclear whether the replication and separation characteristics are shared among all eccDNAs, particularly smaller ones. Assumptions about the role of eccDNAs, including their ability to transcribe miRNAs or express gene fragments, are based on the notion that eccDNAs are stable, transcribable over an extended period, and present in numerous tumour cells (Paulsen et al., 2019). In our second Circle-seq study examining two successive passages of seven PDAC PDCLs, it was determined that nearly all eccDNAs were not replicated and transmitted to their daughter cells. Although it was anticipated that many cells in both passages were genetically related due to their seeding at low confluency, resulting in strong genetic ties, the ecDNA profiles differed significantly between the first and second passage of each cell line. A study conducted by Møller et al. (2018a) found low eccDNA overlap between leukocyte populations from the same individual. This suggested that the constant turnover of leukocytes leads to the elimination of existing eccDNAs. Our analysis now adds

that the eccDNA landscape is influenced by constant turnover due to the low general stability of eccDNAs or the absence of eccDNA replication in combination with the formation of novel eccDNAs. After analysing the VR01-O PDOs in-depth, it is apparent that ecDNAs are retained and their retention can be confirmed through Circle-seq. Hence, the chosen method does not seem to significantly affect the eccDNA landscape. These findings provide fresh insights into the retention rate of eccDNAs, but do not explore the mechanisms involved in their removal. Several mechanisms may be involved, including the reintegration of eccDNAs into the genome, excision and degradation within the cytoplasm, complete excision from the cell, or degradation in micronuclei (Nathanson et al., 2014; Shimizu, Shimura & Tanaka, 2000; Kumar et al., 2017; Paulsen et al., 2018). It is likely that multiple processes are at play, but further research is necessary.

The clinical implications of eccDNAs remain largely unknown when disregarding ecDNAs. The immunostimulatory effect of eccDNAs is observed upon their recognition by the cGAS-STING pathway, which activates innate immune pathways (Wang et al., 2021). However, our findings do not reveal the location of the eccDNAs required for suggesting the activation of cGAS. Nonetheless, some PDCLs showcase an abundance of eccDNAs, which may be released into the cytoplasm to stimulate the activation of the cGAS-STING pathway, leading to an increase in immune activity. This finding presents a potential avenue for immunotherapy (Wang et al., 2021). Pancreatic cancer (PC) is known to be an immune cold tumour and hence has shown poor response to immune checkpoint inhibitors. A high tumour mutational burden is indicative of an improved response to immunotherapy (Cattolico, Bailey & Barry, 2022; Samstein et al., 2019). EccDNA formation due to DNA damage may render tumours that are rich in eccDNA more sensitive to immunotherapy (Paulsen et al., 2021). However, my discoveries do not uncover if eccDNAs are actually released or are degraded in the nucleus. Therefore, it is uncertain whether eccDNA abundance contributes to innate immune activation or whether other factors are involved. Nevertheless, additional characterisation of the eccDNA landscape, combined with clinical data integration, could yield new insights into its usability for personalised medicine.

To summarise, eccDNAs are typically small entities in PDAC, containing mainly specific gene elements rather than complete genes. The characteristics of eccDNAs in PDAC significantly overlap with other cancer types, including their formation hotspots and their content. Nonetheless, this research proposes a straightforward association between active transcription and open chromatin with eccDNA formation. As the low retention rate of eccDNAs becomes evident, it is suggested that eccDNAs are primarily byproducts of various biological processes and do not directly affect cancer biology. However, many avenues of eccDNA research are yet to be investigated.

In conclusion, it is evident that this study on eccDNAs in PDAC merely addresses the surface of the potential analyses to identify their characteristic features and roles. Nonetheless, my analysis revealed a number of distinctive features of eccDNAs that raise new questions,

which can be tackled by integrating existing datasets or generating novel data. Additionally, laboratory research can provide further understanding of the function of eccDNA abundance in activating the immune system or transcribing genetic elements. Within the intricate field of PDAC, eccDNAs may prove to be a crucial piece of the puzzle towards better personalised therapies.

Chapter 6

Conclusion

> If you thought that science was certain - well, that is just an error on your part.
>
> *Richard P. Feynman*

Three central aspects of eccDNA research were investigated in this thesis. The first results chapter covered the processing of high-throughput sequencing data into eccDNA information. The second chapter investigated large and amplified eccDNA in PDAC, its characteristics, associations, and its implications for cell adaptation. Finally, the third chapter delved into the complex world of smaller and highly abundant eccDNAs in PDAC model systems.

nf-core/circdna

nf-core/circdna has laid the foundation for the ecDNA and eccDNA chapters by providing a reproducible, scalable and adaptable workflow. Based on the structure of the pipeline, each dataset in an analysis was treated equally. nf-core/circdna is adaptable when analysing different datasets, performing ecDNA detection using WGS data or eccDNA detection when dealing with Circle-seq data. It also handles quality control for the initial quality check, providing the user with all the necessary information to make an assessment of the eccDNA calling and raw sequencing quality. Using a pipeline to process sequencing data increases the integrity and reproducibility of research, as all parameters used and modifications made can be easily checked. In comparison, running all the included software individually is more prone to human error as each dataset has to be handled in the same way. Therefore, nf-core/circdna can provide peace of mind to the user.

A number of verification steps have been undertaken in the thesis and incorporated into the pipeline to identify the correct installation, use and output of the pipeline. Each branch can be activated by adjusting the workflow parameters, and the inclusion of other parameters makes the pipeline highly adaptable. The pipeline is extensively documented on the nf-core website (https://nf-co.re/circdna) to enable users to understand and run the pipeline with their

own datasets.

Overall, nf-core/circdna follows best practice recommendations in workflow development and provides a state-of-the-art pipeline for processing sequencing data to identify eccDNAs. This pipeline will help to broaden the availability of eccDNA research to the scientific community and aims to increase the knowledge of eccDNAs in cancer. However, as current software is updated or new software and algorithms are developed, and with the rise of long-read sequencing technologies in eccDNA research, this pipeline will require constant updates. Therefore, the journey of nf-core/circdna continues.

EcDNAs in PDAC

The central questions of the ecDNA study in PDAC were to determine prevalence, potential associations with genomic or transcriptomic characteristics, and the potential role of ecDNAs. As little was known about ecDNAs in PDAC, a baseline characterisation was undertaken to investigate multiple avenues. My analysis using publicly available datasets, datasets generated by collaborators at the University of Verona, and self-generated datasets revealed that ecDNAs are abundant in PDAC and may play an important role in PDAC biology, progression, or adaptation.

The results, which highlight genomic instability and the Basal-like subtype in association with ecDNA occurrence, may provide potential avenues for therapeutic intervention. The identification of ecDNAs as biomarkers may indicate potential drug tolerance mechanisms or a potential worse patient outcome. The WR study highlighted the dynamic nature of ecDNA abundance, which can be influenced by external factors by increasing or decreasing selection pressure on cells. While these results were obtained in model systems in a specific environment, analogue environments can exist in patients. Therefore, our results suggest the potential use of ecDNAs as a cell adaptation mechanism to changing environments.

Overall, the comprehensive ecDNA study improves our understanding of these features in PDAC. As a tumour type, PDAC is under-represented in ecDNA research and it has become clear that ecDNAs may be a common instance in PDAC. However, further research with larger patient cohorts is needed to evaluate the potential role of ecDNAs as biomarkers, as actionable targets or to assess therapeutic resistance.

EccDNAs in PDAC

Small abundant eccDNAs, unlike large amplified ecDNAs, are still a mystery with many potential roles. My eccDNA study using PDAC model systems has improved our understanding of two specific eccDNA characteristics - their origin and maintenance - which were important aims of my PhD. The eccDNAs identified do not appear to arise randomly, but from regions of high transcription and open chromatin. They are also unlikely to be maintained and are thought to be byproducts of transcription or DNA damage.

However, much remains unknown about eccDNAs, such as the actual process of their biogenesis, which DNA repair proteins are involved, or how they are lost or deleted. In my analysis, each of the PDAC models was characterised for the presence of eccDNAs and their potential abundance. In future studies, these results can be used to further refine our understanding of eccDNA abundance in relation to biological processes such as immune activation. If eccDNA abundance can be linked to specific processes, a potential use of eccDNAs as biomarkers is conceivable. However, it is clear that our knowledge of eccDNA is still sparse and the study of larger cohorts and the use of novel technologies such as long-read sequencing are advisable for further studies.

Summary

In conclusion, this work lays the foundation for further ecDNA and eccDNA research in PDAC by characterising model systems and patient tumours. This scientific groundwork is necessary for a comprehensive understanding of circular DNA landscapes in order to identify novel potential hypotheses for roles and their use in personalised medicine. In conjunction with nf-core/circdna, the work of this thesis significantly advances the field of eccDNA by advancing the understanding of these features and enabling a wider proportion of the scientific community to study eccDNAs with their own datasets.

Appendix A

Investigating ecDNAs in PDAC

Tab. A.1 | CCLE PDAC cell lines with available WGS data. The cell lines are divided into their site of origin.

SRR Number	Cell Line ID	Cellosaurus ID
Primary tumour		
SRR8639156	HPAC	CVCL_3517
SRR8652122	MIA PaCa-2	CVCL_0428
SRR8670730	Panc 10.05	CVCL_1639
SRR8670731	Panc 03.27	CVCL_1635
SRR8670712	SW 1990	CVCL_1723
SRR8670733	PANC-1	CVCL_0480
SRR8788980	DAN-G	CVCL_0243
Metastasis		
SRR8639189	Capan-1	CVCL_0237
SRR8670709	SUIT-2	CVCL_3172
SRR8670732	PA-TU-8988T	CVCL_1847

Appendix B

Investigating the eccDNA landscape in PDAC

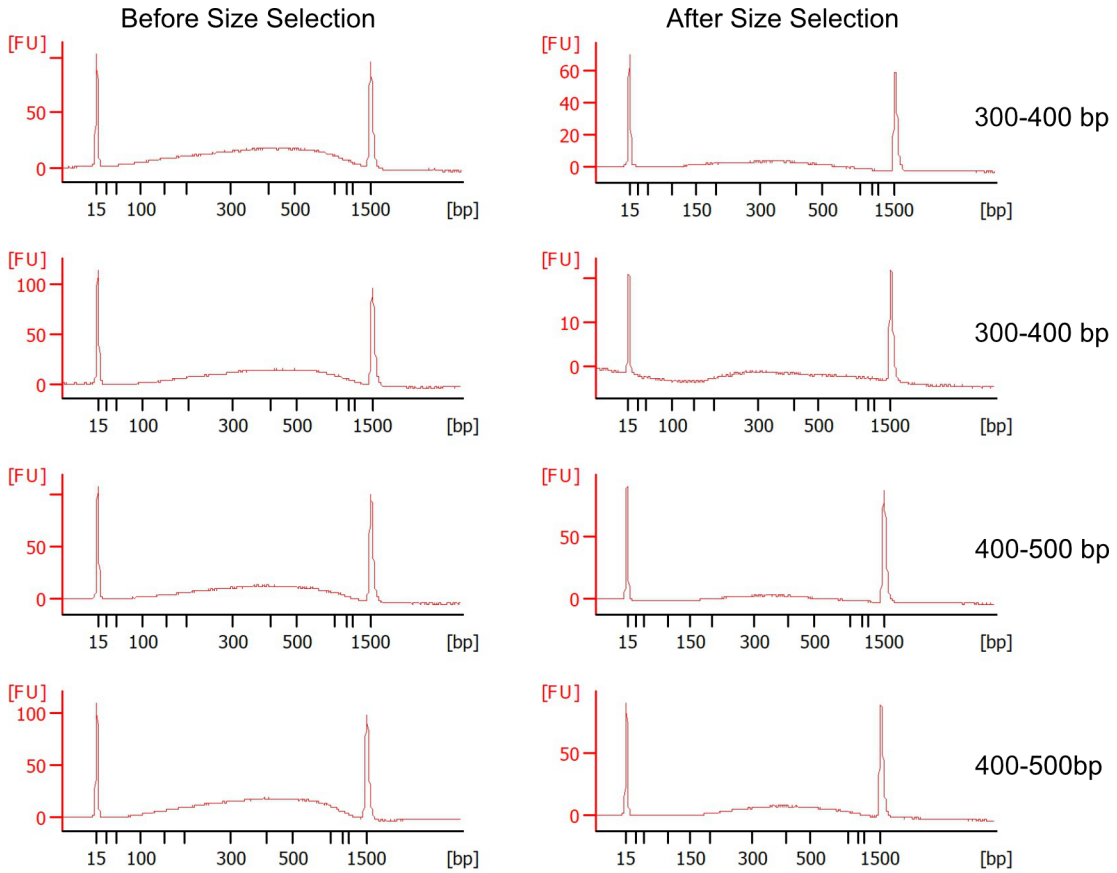


Fig. B.1 | Size selection is associated with huge DNA loss. Size selection test was performed using circular DNA enriched TKCC-2.1 DNA after it was sheared with the M220 sonicator (Covaris) and the microTUBE-15 AFA Beads Screw-Cap (Covaris) for 42 s to achieve an average fragment size of around 400 bp (Before Size Selection). Two size selection procedures were tested to achieve an approximate insert size of around 350-400 bp or 400-500 bp using AMPure XP beads and following the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® recommendations. Size selected DNA was further concentrated for 15 minutes at medium temperature using the Savant™ DNA SpeedVac® DNA120. After all liquid evaporated, DNA was re-eluted in 2 µL DNase-free H2O. 1 µL of sample was loaded onto the DNA 1000 chip (Agilent Technologies) and analysed using the 2100 Bioanalyzer.

Validation of Circle-Seq Results

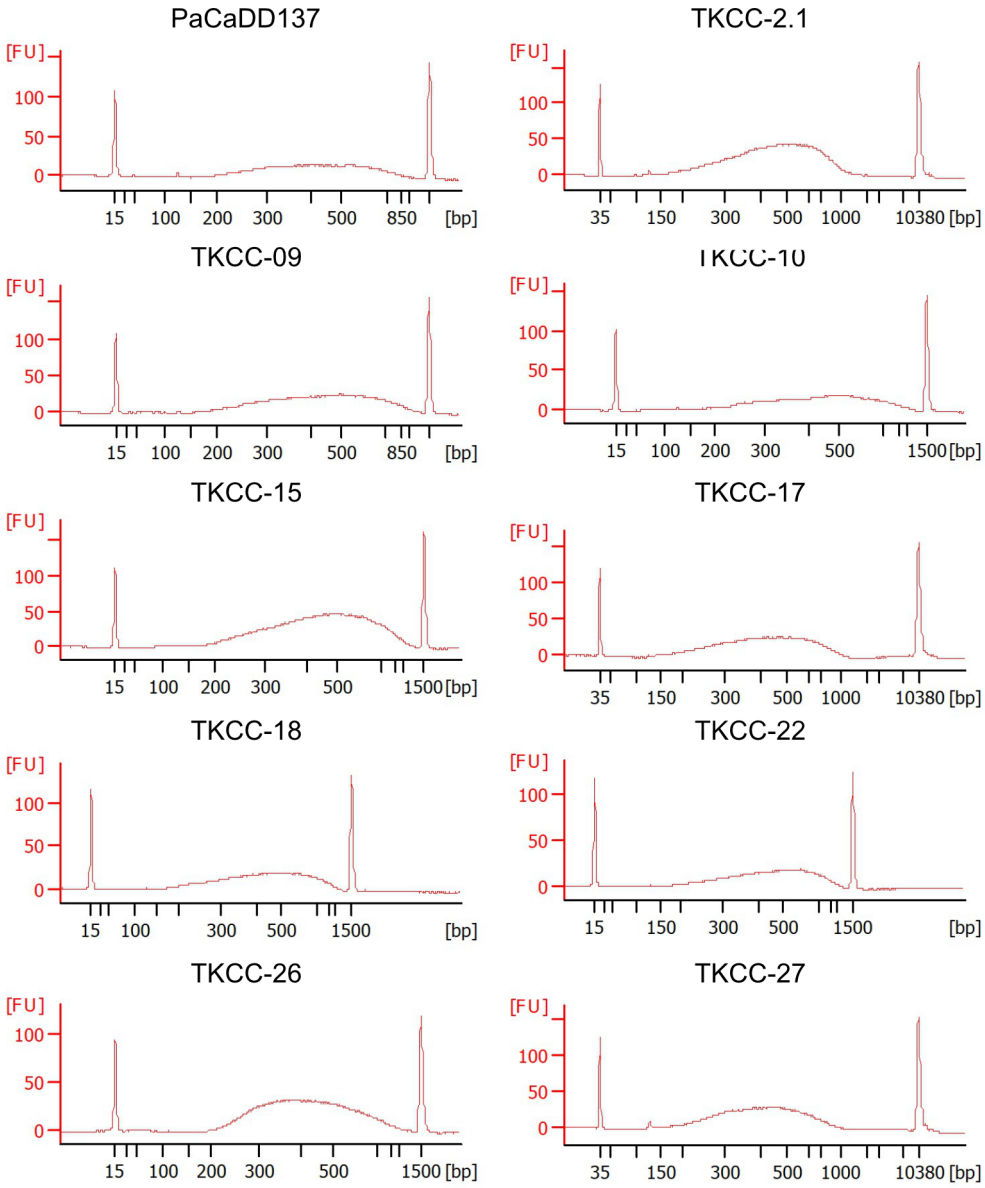


Fig. B.2 | Validation of library preparation success of 10 Circle-seq samples. Libraries were prepared from circular DNA-enriched DNA of 10 PDCLs and validated with the Bioanalyzer (Agilent Technologies).

Tab. B.1 | EccDNA candidates, their chromosomal origin, and their content used for validation.

Cell Line	ID	Region	Genes
Single-Fragment EccDNA identified by Circle-Map Realign (C)			
PaCaDD137	C1	chr16:2090088-2090462	MIR1225
TKCC-2.1	C1	chr7:27161438-27170652	MIR196B
TKCC-10	C1	chr9:87720511-87733809	CTSL
TKCC-15	C1	chr19:13832602-13836825	MIR23A, MIR24-2, MIR27A
TKCC-15	C2	chr7:97730297-97740998	TAC1
TKCC-18	C1	chr5:148298021-148318280	SPINK7
TKCC-22	C1	chr6:52778540-52811441	GSTA1
TKCC-22	C2	chr6:44240699-44255619	MIR4647, HSP90AB1
TKCC-26	C1	chr6:90308442-90317382	MIR4464

Continued on next page

Cell Line	ID	Region	Genes
Multi-Fragment EccDNA identified by Unicycler (U)			
PaCaDD137	U1	chr5:169556621-169557065, chr7:139719221-139721147, chr8:119130626-119142276	
TKCC-2.1	U1	chr12:16282570-16292115, chr14:30369567-30374559, chr17:22126759-22134110, chr5:153417975-153423954	
TKCC-2.1	U2	chr10:94869957-94870934, chr14:59803683-59803815, chr16:60316909-60317024, chr2:48544104-48544607	
TKCC-09	U1	chr12:19708289-19709090, chr12:19709154-19727404, chr12:44398192-44403419	RNU1-146P
TKCC-10	U1	chr14:51560143-51560448, chr14:60296852-60298009, chr3:107668746-107686036	
TKCC-18	U1	chr3:34301581-34301976, chr3:42360873-42361024, chr3:67239278-67262414	
TKCC-18	U2	chr13:81493392-81503331, chr13:81503405-81507562, chr13:85523839-85526761	
TKCC-22	U1	chr10:106313147-106313257, chr17:22078210-22078662, chr3:162988384-162988652, chr5:29160358-29178465	

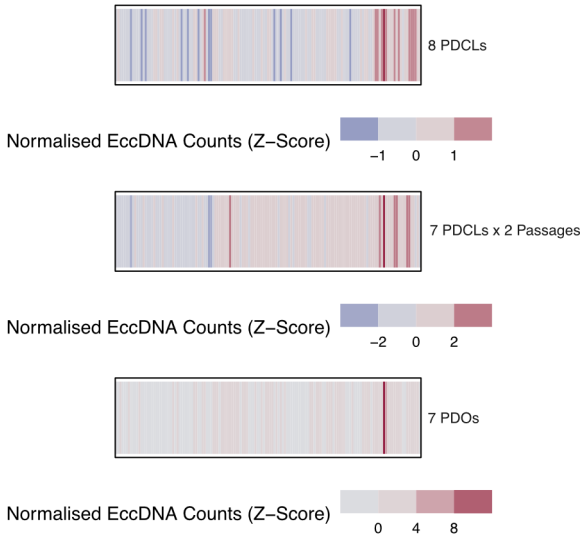


Fig. B.3 | EccDNA abundance on chromosome 8. Z-score normalised eccDNA abundance on chromosome 8 of all three Circle-seq data sets. Each data set contains its own legend below its respective plot.

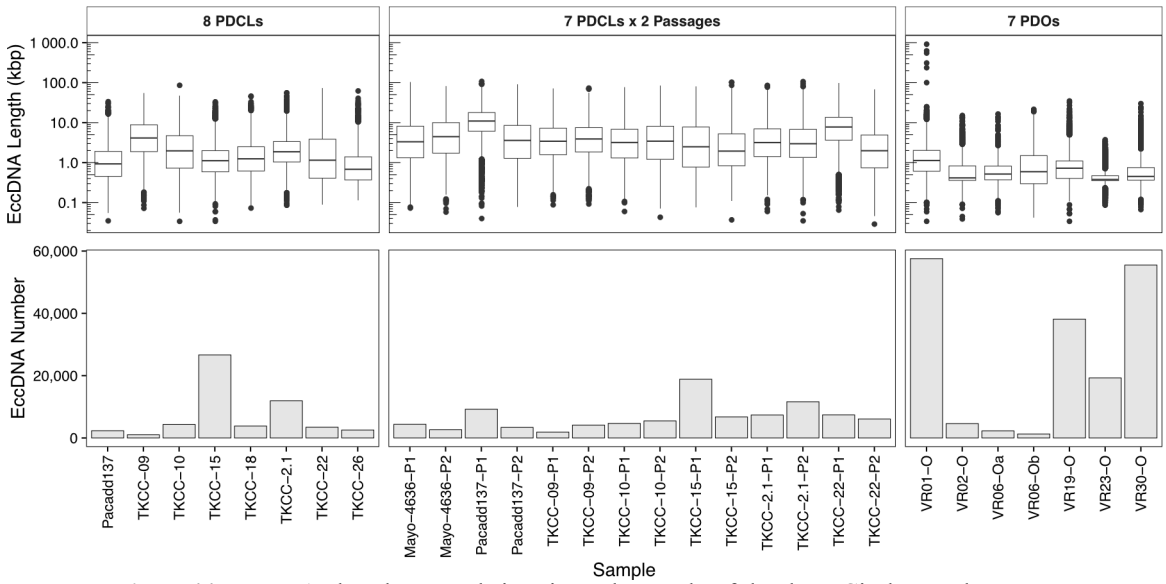


Fig. B.4 | EccDNA abundance and sizes in each sample of the three Circle-seq data sets.

Tab. B.2 | Wilcoxon-rank sum test results of number of genomic elements identified in the common hotspots compared to universal eccDNA hotspots. A common eccDNA hotspot is identified in two Circle-seq data sets, whereas a universal hotspot is identified in all three. *P* values were adjusted using the Benjamini-Hochberg method.

Genomic Feature	Statistic	*P* value	*P* adjusted
Intergenic	250.5	0.039	0.748
hAT	243.5	0.056	0.748
TcMar-Mariner	94.5	0.0874	0.748
RTE-X	97.5	0.102	0.748
enhancers fantom	97	0.102	0.748
protein coding	229.5	0.12	0.754285714
cpg islands	224	0.156	0.754769231
cpg shores	223	0.164	0.754769231
lncrna gencode	110.5	0.19	0.754769231
cpg shelves	219	0.195	0.754769231
cpg inter	214	0.239	0.754769231
hAT-Blackjack	118	0.258	0.754769231
exons	121	0.292	0.754769231
intronexonboundaries	121	0.292	0.754769231
TcMar-Tigger	122	0.303	0.754769231
introns	122	0.303	0.754769231
exonintronboundaries	124	0.327	0.754769231
1to5kb	130	0.403	0.754769231
3UTRs	200	0.403	0.754769231
firstexons	130	0.403	0.754769231
promoters	130	0.403	0.754769231
Simple repeat	199	0.417	0.754769231
Gypsy	131.5	0.42	0.754769231
repeats	197.5	0.438	0.754769231
SINE	197	0.446	0.754769231
MIR	195.5	0.467	0.761037037
LTR	137.5	0.513	0.80262069
L2	191.5	0.529	0.80262069
5UTRs	188	0.586	0.8591
hAT-Tip100	186.5	0.61	0.8591
ERVL	185.5	0.628	0.8591
ERV1	146	0.654	0.8591
ERVL-MaLR	148	0.689	0.8591
CR1	151.5	0.753	0.8591
Low complexity	178.5	0.753	0.8591
DNA	152	0.762	0.8591
L1	152.5	0.771	0.8591
Genes	177	0.78	0.8591
hAT-Charlie	153	0.781	0.8591
cds	173	0.856	0.918634146
Unknown	169.5	0.92	0.963809524
LINE	168.5	0.942	0.963906977
Alu	166	0.99	0.99

Bibliography

Aaltonen, Lauri A. et al. (Feb. 2020). “Pan-cancer analysis of whole genomes”. en. In: *Nature* 578.7793. Number: 7793 Publisher: Nature Publishing Group, pp. 82–93. issn: 1476-4687. doi: 10.1038/s41586-020-1969-6. url: https://www.nature.com/articles/s41586-020-1969-6 (visited on 06/06/2023).

Aboyoun, P., H. Pagès and M. Lawrence (2021). *GenomicRanges: Representation and manipulation of genomic intervals*. url: https://bioconductor.org/packages/GenomicRanges.

Abramoff, M. D., Paulo J. Magalhães and Sunanda J. Ram (2004). “Image processing with ImageJ”. en. In: *Biophotonics international* 11.7. Accepted: 2011-05-12T10:39:50Z Publisher: Laurin Publishing, pp. 36–42. issn: 1081-8693. url: https://dspace.library.uu.nl/handle/1874/204900 (visited on 07/06/2023).

Ahlmann-Eltze, Constantin (2020). *ggupset: Combination Matrix Axis for ggplot2 to Create UpSet Plots*. url: https://github.com/const-ae/ggupset.

Alioto, Tyler S. et al. (Dec. 2015). “A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing”. en. In: *Nature Communications* 6.1. Number: 1 Publisher: Nature Publishing Group, p. 10001. issn: 2041-1723. doi: 10.1038/ncomms10001. url: https://www.nature.com/articles/ncomms10001). (visited on 17/04/2023).

Almal, Suhani H. and Harish Padh (Jan. 2012). “Implications of gene copy-number variation in health and diseases”. en. In: *Journal of Human Genetics* 57.1. Number: 1 Publisher: Nature Publishing Group, pp. 6–13. issn: 1435-232X. doi: 10.1038/jhg.2011.108. url: https://www.nature.com/articles/jhg2011108 (visited on 24/04/2023).

Amarasinghe, Shanika L. et al. (Feb. 2020). “Opportunities and challenges in long-read sequencing data analysis”. In: *Genome Biology* 21.1, p. 30. issn: 1474-760X. doi: 10.1186/s13059-020-1935-5. url: https://doi.org/10.1186/s13059-020-1935-5 (visited on 09/05/2023).

Amemiya, Haley M., Anshul Kundaje and Alan P. Boyle (June 2019). “The ENCODE Blacklist: Identification of Problematic Regions of the Genome”. en. In: *Scientific Reports* 9.1. Number: 1 Publisher: Nature Publishing Group, p. 9354. issn: 2045-2322. doi: 10.1038/s41598-019-45839-z. url: https://www.nature.com/articles/s41598-019-45839-z (visited on 18/06/2021).

Anand, Lakshay (2022). *chromoMap: Interactive Genomic Visualization of Biological Data*. url: https://CRAN.R-project.org/package=chromoMap.

Andor, Noemi et al. (Jan. 2016). "Pan-cancer analysis of the extent and consequences of intratumor heterogeneity". en. In: *Nature Medicine* 22.1. Number: 1 Publisher: Nature Publishing Group, pp. 105–113. issn: 1546-170X. doi: 10.1038/nm.3984. url: https://www.nature.com/articles/nm.3984 (visited on 25/05/2023).

Andrews, Simon et al. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Aung, Kyaw L. et al. (Mar. 2018). "Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial". en. In: *Clinical Cancer Research* 24.6, pp. 1344–1354. issn: 1078-0432, 1557-3265. doi: 10.1158/1078-0432.CCR-17-2994. url: https://aacrjournals.org/clincancerres/article/24/6/1344/474/Genomics-Driven-Precision-Medicine-for-Advanced (visited on 27/06/2022).

Bailey, Peter et al. (Mar. 2016). "Genomic analyses identify molecular subtypes of pancreatic cancer". en. In: *Nature* 531.7592. Number: 7592 Publisher: Nature Publishing Group, pp. 47–52. issn: 1476-4687. doi: 10.1038/nature16965. url: https://www.nature.com/articles/nature16965 (visited on 16/06/2021).

Balaban-Malenbaum, G., G. Grove and F. Gilbert (Mar. 1979). "Increased DNA content of HSR-marker chromosomes of human neuroblastoma cells". en. In: *Experimental Cell Research* 119.2, pp. 419–423. issn: 0014-4827. doi: 10.1016/0014-4827(79)90376-8. url: https://www.sciencedirect.com/science/article/pii/0014482779903768 (visited on 24/04/2023).

Bardeesy, Nabeel and Ronald A. DePinho (Dec. 2002). "Pancreatic cancer biology and genetics". en. In: *Nature Reviews Cancer* 2.12. Number: 12 Publisher: Nature Publishing Group, pp. 897–909. issn: 1474-1768. doi: 10.1038/nrc949. url: https://www.nature.com/articles/nrc949 (visited on 31/05/2023).

Barker, P. E. and T. C. Hsu (May 1978). "Are double minutes chromosomes?" en. In: *Experimental Cell Research* 113.2, pp. 457–458. issn: 0014-4827. doi: 10.1016/0014-4827(78)90391-9. url: https://www.sciencedirect.com/science/article/pii/0014482778903919 (visited on 25/04/2023).

Barker, P.E. (Feb. 1982). "Double minutes in human tumor cells". en. In: *Cancer Genetics and Cytogenetics* 5.1, pp. 81–94. issn: 01654608. doi: 10.1016/0165-4608(82)90043-7. url: https://linkinghub.elsevier.com/retrieve/pii/0165460882900437 (visited on 30/03/2022).

Bedard, Philippe L. et al. (Sept. 2013). "Tumour heterogeneity in the clinic". en. In: *Nature* 501.7467. Number: 7467 Publisher: Nature Publishing Group, pp. 355–364. issn: 1476-4687. doi: 10.1038/nature12627. url: https://www.nature.com/articles/nature12627 (visited on 25/05/2023).

Benchling (2023). *Benchling*. url: https://benchling.com/.

Beroukhim, Rameen et al. (Feb. 2010). "The landscape of somatic copy-number alteration across human cancers". en. In: *Nature* 463.7283. Number: 7283 Publisher: Nature Pub-

lishing Group, pp. 899–905. issn: 1476-4687. doi: 10.1038/nature08822. url: https://www.nature.com/articles/nature08822 (visited on 24/04/2023).

Beverley, S. M. et al. (Sept. 1984). "Unstable DNA amplifications in methotrexate-resistant Leishmania consist of extrachromosomal circles which relocalize during stabilization". eng. In: *Cell* 38.2, pp. 431–439. issn: 0092-8674. doi: 10.1016/0092-8674(84)90498-7.

Bhang, Hyo-eun C. et al. (May 2015). "Studying clonal dynamics in response to cancer therapy using high-complexity barcoding". en. In: *Nature Medicine* 21.5. Number: 5 Publisher: Nature Publishing Group, pp. 440–448. issn: 1546-170X. doi: 10.1038/nm.3841. url: https://www.nature.com/articles/nm.3841 (visited on 03/04/2023).

Biankin, Andrew V. et al. (Nov. 2012). "Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes". en. In: *Nature* 491.7424. Number: 7424 Publisher: Nature Publishing Group, pp. 399–405. issn: 1476-4687. doi: 10.1038/nature11547. url: https://www.nature.com/articles/nature11547 (visited on 29/06/2022).

Bielski, Craig M. et al. (Aug. 2018). "Genome doubling shapes the evolution and prognosis of advanced cancers". eng. In: *Nature Genetics* 50.8, pp. 1189–1195. issn: 1546-1718. doi: 10.1038/s41588-018-0165-1.

Bioconductor Package Maintainer (2021). *liftOver: Changing genomic coordinate systems with rtracklayer::liftOver*. url: https://www.bioconductor.org/help/workflows/liftOver/.

Bird, Adrian (Jan. 2002). "DNA methylation patterns and epigenetic memory". en. In: *Genes & Development* 16.1. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 6–21. issn: 0890-9369, 1549-5477. doi: 10.1101/gad.947102. url: http://genesdev.cshlp.org/content/16/1/6 (visited on 24/05/2023).

Bishop, J. Michael (Jan. 1987). "The Molecular Genetics of Cancer". In: *Science* 235.4786. Publisher: American Association for the Advancement of Science, pp. 305–311. doi: 10.1126/science.3541204. url: https://www.science.org/doi/abs/10.1126/science.3541204 (visited on 25/05/2023).

Boj, Sylvia F. et al. (Jan. 2015). "Organoid Models of Human and Mouse Ductal Pancreatic Cancer". en. In: *Cell* 160.1, pp. 324–338. issn: 0092-8674. doi: 10.1016/j.cell.2014.12.021. url: https://www.sciencedirect.com/science/article/pii/S009286741401592X (visited on 25/06/2022).

Brunson, Jason Cory and Quentin D. Read (2020). *ggalluvial: Alluvial Plots in ggplot2*. url: http://corybrunson.github.io/ggalluvial/.

Brunton, Holly et al. (May 2020). "HNF4A and GATA6 Loss Reveals Therapeutically Actionable Subtypes in Pancreatic Cancer". en. In: *Cell Reports* 31.6, p. 107625. issn: 2211-1247. doi: 10.1016/j.celrep.2020.107625. url: http://www.sciencedirect.com/science/article/pii/S2211124720305787 (visited on 30/05/2020).

Buenrostro, Jason D. et al. (Dec. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". en. In: *Nature Methods* 10.12. Publisher: Nature Publishing Group, pp. 1213–

1218. issn: 1548-7105. doi: 10.1038/nmeth.2688. url: https://www.nature.com/articles/nmeth.2688 (visited on 15/04/2024).

Caiado, Francisco, Bruno Silva-Santos and Håkan Norell (2016). "Intra-tumour heterogeneity – going beyond genetics". en. In: *The FEBS Journal* 283.12. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/febs.13705, pp. 2245–2258. issn: 1742-4658. doi: 10.1111/febs.13705. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/febs.13705 (visited on 31/03/2023).

Cameron, Daniel and Ruining Dong (2021). *StructuralVariantAnnotation: Variant annotations for structural variants*.

Campbell, Peter J. et al. (Oct. 2010). "The patterns and dynamics of genomic instability in metastatic pancreatic cancer". en. In: *Nature* 467.7319. Number: 7319 Publisher: Nature Publishing Group, pp. 1109–1113. issn: 1476-4687. doi: 10.1038/nature09460. url: https://www.nature.com/articles/nature09460 (visited on 12/05/2023).

Cancer Statistics for the UK (May 2015). en. url: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk (visited on 01/06/2023).

Cao, Mengru et al. (Sept. 2012). "MiR-23a regulates TGF-β-induced epithelial-mesenchymal transition by targeting E-cadherin in lung cancer cells". In: *International Journal of Oncology* 41.3. Publisher: Spandidos Publications, pp. 869–875. issn: 1019-6439. doi: 10.3892/ijo.2012.1535. url: https://www.spandidos-publications.com/10.3892/ijo.2012.1535 (visited on 19/10/2023).

Carlson, M. et al. (2022). *GenomicFeatures: Conveniently import and query gene models*. url: https://bioconductor.org/packages/GenomicFeatures.

Carroll, S M et al. (May 1987). "Characterization of an episome produced in hamster cells that amplify a transfected CAD gene at high frequency: functional evidence for a mammalian replication origin." In: *Molecular and Cellular Biology* 7.5, pp. 1740–1750. issn: 0270-7306. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC365275/ (visited on 26/04/2023).

Carroll, S M et al. (Apr. 1988). "Double minute chromosomes can be produced from precursors derived from a chromosomal deletion". In: *Molecular and Cellular Biology* 8.4. Publisher: American Society for Microbiology, pp. 1525–1533. doi: 10.1128/mcb.8.4.1525-1533.1988. url: https://journals.asm.org/doi/abs/10.1128/mcb.8.4.1525-1533.1988 (visited on 26/04/2023).

Carter, Scott L. et al. (Sept. 2006). "A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers". en. In: *Nature Genetics* 38.9. Number: 9 Publisher: Nature Publishing Group, pp. 1043–1048. issn: 1546-1718. doi: 10.1038/ng1861. url: https://www.nature.com/articles/ng1861 (visited on 05/07/2023).

Carter, Scott L. et al. (May 2012). "Absolute quantification of somatic DNA alterations in human cancer". en. In: *Nature Biotechnology* 30.5. Number: 5 Publisher: Nature Publishing Group, pp. 413–421. issn: 1546-1696. doi: 10.1038/nbt.2203. url: https://www.nature.com/articles/nbt.2203 (visited on 27/06/2023).

Cattolico, Carlotta, Peter Bailey and Simon T. Barry (2022). “Modulation of Type I Interferon Responses to Influence Tumor-Immune Cross Talk in PDAC”. eng. In: *Frontiers in Cell and Developmental Biology* 10, p. 816517. issn: 2296-634X. doi: 10.3389/fcell.2022.816517.

Cavalcante, Raymond G and Maureen A Sartor (Aug. 2017). “annotatr: genomic regions in context”. In: *Bioinformatics* 33.15, pp. 2381–2383. issn: 1367-4803. doi: 10.1093/bioinformatics/btx183. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860117/ (visited on 29/06/2022).

Cavalcante, Raymond G. (2021). *annotatr: Annotation of Genomic Regions to Genomic Annotations*.

Cesare, Anthony J. and Jack D. Griffith (Nov. 2004). “Telomeric DNA in ALT Cells Is Characterized by Free Telomeric Circles and Heterogeneous t-Loops”. In: *Molecular and Cellular Biology* 24.22. Publisher: American Society for Microbiology, pp. 9948–9957. doi: 10.1128/MCB.24.22.9948-9957.2004. url: https://journals.asm.org/doi/full/10.1128/MCB.24.22.9948-9957.2004 (visited on 04/04/2023).

Cesare, Anthony J. and Roger R. Reddel (May 2010). “Alternative lengthening of telomeres: models, mechanisms and implications”. en. In: *Nature Reviews Genetics* 11.5. Number: 5 Publisher: Nature Publishing Group, pp. 319–330. issn: 1471-0064. doi: 10.1038/nrg2763. url: https://www.nature.com/articles/nrg2763 (visited on 11/05/2023).

Chan-Seng-Yue, Michelle et al. (Feb. 2020). “Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution”. en. In: *Nature Genetics* 52.2. Number: 2 Publisher: Nature Publishing Group, pp. 231–240. issn: 1546-1718. doi: 10.1038/s41588-019-0566-9. url: https://www.nature.com/articles/s41588-019-0566-9 (visited on 07/05/2021).

Chen, Shifu et al. (Sept. 2018). “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17, pp. i884–i890. issn: 1367-4803. doi: 10.1093/bioinformatics/bty560. url: https://doi.org/10.1093/bioinformatics/bty560 (visited on 09/02/2023).

Chitwood, Dylan G. et al. (Jan. 2023). “Microevolutionary dynamics of eccDNA in Chinese hamster ovary cells grown in fed-batch cultures under control and lactate-stressed conditions”. en. In: *Scientific Reports* 13.1. Number: 1 Publisher: Nature Publishing Group, p. 1200. issn: 2045-2322. doi: 10.1038/s41598-023-27962-0. url: https://www.nature.com/articles/s41598-023-27962-0 (visited on 09/05/2023).

Cho, Seung Woo et al. (May 2018). “Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element”. eng. In: *Cell* 173.6, 1398–1412.e22. issn: 1097-4172. doi: 10.1016/j.cell.2018.03.068.

Chu, Gerald C. et al. (2007). “Stromal biology of pancreatic cancer”. en. In: *Journal of Cellular Biochemistry* 101.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcb.21209, pp. 887–907. issn: 1097-4644. doi: 10.1002/jcb.21209. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcb.21209 (visited on 10/10/2023).

Cingolani, Pablo, Rob Sladek and Mathieu Blanchette (Jan. 2015). “BigDataScript: a scripting language for data pipelines”. In: *Bioinformatics* 31.1, pp. 10–16. issn: 1367-4803. doi:

10.1093/bioinformatics/btu595. url: https://doi.org/10.1093/bioinformatics/btu595 (visited on 10/06/2023).

Clarke, Erik and Scott Sherrill-Mix (2017). *ggbeeswarm: Categorical Scatter (Violin Point) Plots*. url: https://github.com/eclarke/ggbeeswarm.

Cohen, S., A. Regev and S. Lavi (Feb. 1997). "Small polydispersed circular DNA (spcDNA) in human cells: association with genomic instability". en. In: *Oncogene* 14.8. Number: 8 Publisher: Nature Publishing Group, pp. 977–985. issn: 1476-5594. doi: 10.1038/sj.onc.1200917. url: https://www.nature.com/articles/1200917 (visited on 15/06/2021).

Collisson, Eric A. et al. (Apr. 2011). "Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy". en. In: *Nature Medicine* 17.4. Number: 4 Publisher: Nature Publishing Group, pp. 500–503. issn: 1546-170X. doi: 10.1038/nm.2344. url: https://www.nature.com/articles/nm.2344 (visited on 03/06/2020).

Collisson, Eric A. et al. (Apr. 2019). "Molecular subtypes of pancreatic cancer". en. In: *Nature Reviews Gastroenterology & Hepatology* 16.4, pp. 207–220. issn: 1759-5045, 1759-5053. doi: 10.1038/s41575-019-0109-y. url: http://www.nature.com/articles/s41575-019-0109-y (visited on 27/06/2022).

Cortés-Ciriano, Isidro et al. (Mar. 2020). "Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing". en. In: *Nature Genetics* 52.3. Number: 3 Publisher: Nature Publishing Group, pp. 331–341. issn: 1546-1718. doi: 10.1038/s41588-019-0576-7. url: https://www.nature.com/articles/s41588-019-0576-7 (visited on 13/01/2023).

Cowell, John K. (1982). "DOUBLE MINUTES AND HOMOGENEOUSLY STAINING REGIONS: Gene Amplification in Mammalian Cells". In: *Annual Review of Genetics* 16.1. _eprint: https://doi.org/10.1146/annurev.ge.16.120182.000321, pp. 21–59. doi: 10.1146/annurev.ge.16.120182.000321. url: https://doi.org/10.1146/annurev.ge.16.120182.000321 (visited on 30/03/2022).

Cowell, John K. and Orlando J. Miller (Sept. 1983). "Occurrence and evolution of homogeneously staining regions may be due to breakage-fusion-bridge cycles following telomere loss". en. In: *Chromosoma* 88.3, pp. 216–221. issn: 1432-0886. doi: 10.1007/BF00285623. url: https://doi.org/10.1007/BF00285623 (visited on 26/04/2023).

Cox, David, Catherine Yuncken and ArthurI. Spriggs (July 1965). "MINUTE CHROMATIN BODIES IN MALIGNANT TUMOURS OF CHILDHOOD". en. In: *The Lancet*. Originally published as Volume 2, Issue 7402 286.7402, pp. 55–58. issn: 0140-6736. doi: 10.1016/S0140-6736(65)90131-5. url: https://www.sciencedirect.com/science/article/pii/S0140673665901315 (visited on 30/03/2022).

Croce, Carlo M. (Jan. 2008). "Oncogenes and Cancer". In: *New England Journal of Medicine* 358.5. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMra072367, pp. 502–511. issn: 0028-4793. doi: 10.1056/NEJMra072367. url: https://doi.org/10.1056/NEJMra072367 (visited on 25/05/2023).

D'Agosto, Sabrina et al. (Jan. 2019). "Preclinical Modelling of PDA: Is Organoid the New Black?" en. In: *International Journal of Molecular Sciences* 20.11. Number: 11 Publisher:

Multidisciplinary Digital Publishing Institute, p. 2766. issn: 1422-0067. doi: 10.3390/ijms20112766. url: https://www.mdpi.com/1422-0067/20/11/2766 (visited on 22/04/2023).

Da, Haber and Schimke Rt (Nov. 1981). "Unstable amplification of an altered dihydrofolate reductase gene associated with double-minute chromosomes". en. In: *Cell* 26.3 Pt 1. Publisher: Cell. issn: 0092-8674. doi: 10.1016/0092-8674(81)90204-x. url: https://pubmed.ncbi.nlm.nih.gov/7326744/ (visited on 29/04/2023).

Dagogo-Jack, Ibiayi and Alice T. Shaw (Feb. 2018). "Tumour heterogeneity and resistance to cancer therapies". en. In: *Nature Reviews Clinical Oncology* 15.2. Number: 2 Publisher: Nature Publishing Group, pp. 81–94. issn: 1759-4782. doi: 10.1038/nrclinonc.2017.166. url: https://www.nature.com/articles/nrclinonc.2017.166 (visited on 31/03/2023).

Danecek, Petr et al. (Feb. 2021). "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.2, giab008. issn: 2047-217X. doi: 10.1093/gigascience/giab008. url: https://doi.org/10.1093/gigascience/giab008 (visited on 21/06/2023).

Dang, Chi V. (Jan. 1999). "c-Myc Target Genes Involved in Cell Growth, Apoptosis, and Metabolism". In: *Molecular and Cellular Biology* 19.1. Publisher: Taylor & Francis _eprint: https://doi.org/10.1128/MCB.19.1.1, pp. 1–11. issn: null. doi: 10.1128/MCB.19.1.1. url: https://doi.org/10.1128/MCB.19.1.1 (visited on 31/05/2023).

— (Mar. 2012). "MYC on the Path to Cancer". en. In: *Cell* 149.1, pp. 22–35. issn: 0092-8674. doi: 10.1016/j.cell.2012.03.003. url: https://www.sciencedirect.com/science/article/pii/S0092867412002966 (visited on 31/05/2023).

De Maio, Nicola et al. (Sept. 2019). "Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes". In: *Microbial Genomics* 5.9, e000294. issn: 2057-5858. doi: 10.1099/mgen.0.000294. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6807382/ (visited on 09/05/2023).

deCarvalho, Ana C. et al. (May 2018). "Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma". en. In: *Nature Genetics* 50.5. Number: 5 Publisher: Nature Publishing Group, pp. 708–717. issn: 1546-1718. doi: 10.1038/s41588-018-0105-0. url: https://www.nature.com/articles/s41588-018-0105-0 (visited on 25/04/2023).

Deshpande, Viraj et al. (Jan. 2019). "Exploring the landscape of focal amplifications in cancer using AmpliconArchitect". en. In: *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group, p. 392. issn: 2041-1723. doi: 10.1038/s41467-018-08200-y. url: https://www.nature.com/articles/s41467-018-08200-y (visited on 27/06/2022).

Dewhurst, Sally M. et al. (Feb. 2014). "Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution". eng. In: *Cancer Discovery* 4.2, pp. 175–185. issn: 2159-8290. doi: 10.1158/2159-8290.CD-13-0285.

Dharanipragada, Prashanthi et al. (Apr. 2023). "Blocking Genomic Instability Prevents Acquired Resistance to MAPK Inhibitor Therapy in Melanoma". In: *Cancer Discovery* 13.4, pp. 880–909. issn: 2159-8274. doi: 10.1158/2159-8290.CD-22-0787. url: https://doi.org/10.1158/2159-8290.CD-22-0787 (visited on 06/06/2023).

Di Tommaso, Paolo et al. (Apr. 2017). “Nextflow enables reproducible computational workflows”. en. In: *Nature Biotechnology* 35.4. Number: 4 Publisher: Nature Publishing Group, pp. 316–319. issn: 1546-1696. doi: 10.1038/nbt.3820. url: https://www.nature.com/articles/nbt.3820%5C%7B (visited on 09/06/2023).

Dieffenbach, CW, TM Lowe, GS Dveksler et al. (1993). “General concepts for PCR primer design”. In: *PCR methods appl* 3.3, S30–S37.

Dillon, Laura W. et al. (June 2015). “Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity”. en. In: *Cell Reports* 11.11, pp. 1749–1759. issn: 22111247. doi: 10.1016/j.celrep.2015.05.020. url: https://linkinghub.elsevier.com/retrieve/pii/S221112471500546X (visited on 04/04/2023).

Dolcet, Xavier et al. (May 2005). “NF-kB in development and progression of human cancer”. en. In: *Virchows Archiv* 446.5, pp. 475–482. issn: 1432-2307. doi: 10.1007/s00428-005-1264-9. url: https://doi.org/10.1007/s00428-005-1264-9 (visited on 04/07/2023).

Dolgalev, Igor (2022). *msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format*. url: https://igordot.github.io/msigdbr/.

Dreyer, Stephan B. et al. (Jan. 2021). “Targeting DNA Damage Response and Replication Stress in Pancreatic Cancer”. en. In: *Gastroenterology* 160.1, 362–377.e13. issn: 0016-5085. doi: 10.1053/j.gastro.2020.09.043. url: https://www.sciencedirect.com/science/article/pii/S001650852035229X (visited on 26/03/2021).

Drost, Jarno and Hans Clevers (July 2018). “Organoids in cancer research”. en. In: *Nature Reviews Cancer* 18.7. Number: 7 Publisher: Nature Publishing Group, pp. 407–418. issn: 1474-1768. doi: 10.1038/s41568-018-0007-6. url: https://www.nature.com/articles/s41568-018-0007-6 (visited on 14/09/2023).

Durinck, Steffen and Wolfgang Huber (2022). *biomaRt: Interface to BioMart databases (i.e. Ensembl)*.

Edgar, Robert C. (Mar. 2004). “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32.5, pp. 1792–1797. issn: 0305-1048. doi: 10.1093/nar/gkh340. url: https://doi.org/10.1093/nar/gkh340 (visited on 07/06/2023).

Ewels, Philip et al. (2016). “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19, pp. 3047–3048.

Ewels, Philip A. et al. (Mar. 2020). “The nf-core framework for community-curated bioinformatics pipelines”. en. In: *Nature Biotechnology* 38.3. Number: 3 Publisher: Nature Publishing Group, pp. 276–278. issn: 1546-1696. doi: 10.1038/s41587-020-0439-x. url: https://www.nature.com/articles/s41587-020-0439-x (visited on 27/06/2022).

Fisher, R., L. Pusztai and C. Swanton (Feb. 2013). “Cancer heterogeneity: implications for targeted therapeutics”. en. In: *British Journal of Cancer* 108.3. Number: 3 Publisher: Nature Publishing Group, pp. 479–485. issn: 1532-1827. doi: 10.1038/bjc.2012.581. url: https://www.nature.com/articles/bjc2012581 (visited on 28/04/2023).

Francis, Joshua M. et al. (Aug. 2014). “EGFR Variant Heterogeneity in Glioblastoma Resolved through Single-Nucleus Sequencing”. In: *Cancer Discovery* 4.8, pp. 956–971. issn: 2159-

8274. doi: 10.1158/2159-8290.CD-13-0879. url: https://doi.org/10.1158/2159-8290.CD-13-0879 (visited on 26/04/2023).

Fungtammasan, Arkarachai et al. (Jan. 2012). "A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome?" en. In: *Genome Research* 22.6. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 993–1005. issn: 1088-9051, 1549-5469. doi: 10.1101/gr.134395.111. url: https://genome.cshlp.org/content/22/6/993 (visited on 23/05/2023).

Furey, Terrence S. and David Haussler (May 2003). "Integration of the cytogenetic map with the draft human genome sequence". eng. In: *Human Molecular Genetics* 12.9, pp. 1037–1044. issn: 0964-6906. doi: 10.1093/hmg/ddg113.

Garcia, Maxime et al. (Sept. 2020). "Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants". In: *F1000Research* 9, p. 63. issn: 2046-1402. doi: 10.12688/f1000research.16665.2. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7111497/ (visited on 09/02/2023).

Gaspar, John M. (Dec. 2018). *Improved peak-calling with MACS2*. en. Pages: 496521 Section: Contradictory Results. doi: 10.1101/496521. url: https://www.biorxiv.org/content/10.1101/496521v1 (visited on 23/06/2022).

Gaubatz, James W. and Sonia C. Flores (Jan. 1990). "Tissue-specific and age-related variations in repetitive sequences of mouse extrachromosomal circular DNAs". en. In: *Mutation Research/DNAging* 237.1, pp. 29–36. issn: 0921-8734. doi: 10.1016/0921-8734(90)90029-Q. url: https://www.sciencedirect.com/science/article/pii/092187349090029Q (visited on 03/06/2021).

Gel, Bernat (2022). *karyoploteR: Plot customizable linear genomes displaying arbitrary data*. url: https://github.com/bernatgel/karyoploteR.

Gentleman, R. (2021). *annotate: Annotation for microarrays*.

Gentleman, R. et al. (2021). *Biobase: Base functions for Bioconductor*. url: https://bioconductor.org/packages/Biobase.

Gilbert, Nick et al. (Sept. 2004). "Chromatin Architecture of the Human Genome: Gene-Rich Domains Are Enriched in Open Chromatin Fibers". en. In: *Cell* 118.5, pp. 555–566. issn: 0092-8674. doi: 10.1016/j.cell.2004.08.011. url: https://www.sciencedirect.com/science/article/pii/S0092867404007883 (visited on 23/05/2023).

Grady, D L et al. (Mar. 1992). "Highly conserved repetitive DNA sequences are present at human centromeres." In: *Proceedings of the National Academy of Sciences* 89.5. Publisher: Proceedings of the National Academy of Sciences, pp. 1695–1699. doi: 10.1073/pnas.89.5.1695. url: https://www.pnas.org/doi/abs/10.1073/pnas.89.5.1695 (visited on 28/03/2023).

Grewal, Shiv I. S. and Songtao Jia (Jan. 2007). "Heterochromatin revisited". en. In: *Nature Reviews Genetics* 8.1. Number: 1 Publisher: Nature Publishing Group, pp. 35–46. issn: 1471-0064. doi: 10.1038/nrg2008. url: https://www.nature.com/articles/nrg2008 (visited on 12/09/2023).

Griffin, Constance A. et al. (June 1995). "Consistent Chromosome Abnormalities in Adenocarcinoma of the Pancreas1". In: *Cancer Research* 55.11, pp. 2394–2399. issn: 0008-5472.

Gu, Zuguang (2021). *ComplexHeatmap: Make Complex Heatmaps*.

— (2022). *circlize: Circular Visualization*. url: https://CRAN.R-project.org/package=circlize.

Guérin, Thomas M. and Stéphane Marcand (July 2022). "Breakage in breakage–fusion–bridge cycle: an 80-year-old mystery". en. In: *Trends in Genetics* 38.7, pp. 641–645. issn: 0168-9525. doi: 10.1016/j.tig.2022.03.008. url: https://www.sciencedirect.com/science/article/pii/S0168952522000695 (visited on 26/04/2023).

Guinney, Justin and Robert Castelo (2021). *GSVA: Gene Set Variation Analysis for microarray and RNA-seq data*. url: https://github.com/rcastelo/GSVA.

Haber, Daniel A. and Robert T. Schimke (Nov. 1981). "Unstable amplification of an altered dihydrofolate reductase gene associated with double-minute chromosomes". English. In: *Cell* 26.3. Publisher: Elsevier, pp. 355–362. issn: 0092-8674, 1097-4172. doi: 10.1016/0092-8674(81)90204-X. url: https://www.cell.com/cell/abstract/0092-8674(81)90204-X (visited on 06/09/2023).

Hagerling, Catharina, Amy-Jo Casbon and Zena Werb (Apr. 2015). "Balancing the innate immune system in tumor development". English. In: *Trends in Cell Biology* 25.4. Publisher: Elsevier, pp. 214–220. issn: 0962-8924, 1879-3088. doi: 10.1016/j.tcb.2014.11.001. url: https://www.cell.com/trends/cell-biology/abstract/S0962-8924(14)00194-9 (visited on 04/07/2023).

Hahn, Peter J. (1993). "Molecular biology of double-minute chromosomes". en. In: *BioEssays* 15.7. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.950150707, pp. 477–484. issn: 1521-1878. doi: 10.1002/bies.950150707. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.950150707 (visited on 15/11/2023).

Hamkalo, B. A. et al. (Feb. 1985). "Ultrastructural features of minute chromosomes in a methotrexate-resistant mouse 3T3 cell line". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 82.4, pp. 1126–1130. issn: 0027-8424. doi: 10.1073/pnas.82.4.1126.

Hanahan, Douglas (Jan. 2022). "Hallmarks of Cancer: New Dimensions". In: *Cancer Discovery* 12.1, pp. 31–46. issn: 2159-8274. doi: 10.1158/2159-8290.CD-21-1059. url: https://doi.org/10.1158/2159-8290.CD-21-1059 (visited on 30/03/2023).

Hao, Yi-Heng et al. (Oct. 2019). "Induction of LEF1 by MYC activates the WNT pathway and maintains cell proliferation". In: *Cell Communication and Signaling* 17.1, p. 129. issn: 1478-811X. doi: 10.1186/s12964-019-0444-1. url: https://doi.org/10.1186/s12964-019-0444-1 (visited on 01/02/2023).

Hardie, Rae-Anne et al. (Jan. 2017). "Mitochondrial mutations and metabolic adaptation in pancreatic cancer". In: *Cancer & Metabolism* 5.1, p. 2. issn: 2049-3002. doi: 10.1186/s40170-017-0164-1. url: https://doi.org/10.1186/s40170-017-0164-1 (visited on 09/06/2021).

Helmsauer, Konstantin et al. (Nov. 2020). "Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma". en. In: *Nature Communications*

11.1. Number: 1 Publisher: Nature Publishing Group, p. 5823. issn: 2041-1723. doi: 10.1038/s41467-020-19452-y. url: https://www.nature.com/articles/s41467-020-19452-y (visited on 17/06/2021).

Henssen, Anton et al. (Jan. 2019a). "Purification and Sequencing of Large Circular DNA from Human Cells". In: *Protocol Exchange*. doi: 10.1038/protex.2019.006.

Henssen, Anton et al. (Jan. 2019b). "Purification and Sequencing of Large Circular DNA from Human Cells". en. In: *Protocol Exchange*. issn: 2043-0116. doi: 10.1038/protex.2019.006. url: http://www.nature.com/protocolexchange/protocols/7203 (visited on 16/03/2023).

Hills, Stephanie A. and John F. X. Diffley (May 2014). "DNA Replication and Oncogene-Induced Replicative Stress". en. In: *Current Biology* 24.10, R435–R444. issn: 0960-9822. doi: 10.1016/j.cub.2014.04.012. url: https://www.sciencedirect.com/science/article/pii/S0960982214004126 (visited on 06/06/2023).

Hotta, Yasuo and Alix Bassel (Feb. 1965). "MOLECULAR SIZE AND CIRCULARITY OF DNA IN CELLS OF MAMMALS AND HIGHER PLANTS". en. In: *Proceedings of the National Academy of Sciences* 53.2, pp. 356–362. issn: 0027-8424, 1091-6490. doi: 10.1073/pnas.53.2.356. url: https://pnas.org/doi/full/10.1073/pnas.53.2.356 (visited on 03/04/2023).

Hruban, Ralph H. et al. (Aug. 2000). "Progression Model for Pancreatic Cancer1". In: *Clinical Cancer Research* 6.8, pp. 2969–2972. issn: 1078-0432.

Hu, Hai-feng et al. (Nov. 2021). "Mutations in key driver genes of pancreatic cancer: molecularly targeted therapies and other clinical implications". en. In: *Acta Pharmacologica Sinica* 42.11. Number: 11 Publisher: Nature Publishing Group, pp. 1725–1741. issn: 1745-7254. doi: 10.1038/s41401-020-00584-2. url: https://www.nature.com/articles/s41401-020-00584-2 (visited on 31/05/2023).

Huang, Chenhui et al. (June 2017). "The human CTC1/STN1/TEN1 complex regulates telomere maintenance in ALT cancer cells". en. In: *Experimental Cell Research* 355.2, pp. 95–104. issn: 0014-4827. doi: 10.1016/j.yexcr.2017.03.058. url: https://www.sciencedirect.com/science/article/pii/S0014482717301866 (visited on 11/05/2023).

Huang, Miller and William A. Weiss (Oct. 2013). "Neuroblastoma and MYCN". In: *Cold Spring Harbor Perspectives in Medicine* 3.10, a014415. issn: 2157-1422. doi: 10.1101/cshperspect.a014415. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3784814/ (visited on 13/09/2023).

Huennekens, F. M. (Jan. 1994). "The methotrexate story: A paradigm for development of cancer chemotherapeutic agents". en. In: *Advances in Enzyme Regulation* 34, pp. 397–419. issn: 0065-2571. doi: 10.1016/0065-2571(94)90025-6. url: https://www.sciencedirect.com/science/article/pii/0065257194900256 (visited on 29/04/2023).

Hughes, Peyton et al. (Nov. 2007). "The costs of using unauthenticated, over-passaged cell lines: how much more data do we need?" In: *BioTechniques* 43.5. Publisher: Future Science, pp. 575–586. issn: 0736-6205. doi: 10.2144/000112598. url: https://www.future-science.com/doi/full/10.2144/000112598 (visited on 23/03/2023).

Hull, Ryan M. et al. (Dec. 2019). "Transcription-induced formation of extrachromosomal DNA during yeast ageing". en. In: *PLOS Biology* 17.12. Publisher: Public Library of Science, e3000471. issn: 1545-7885. doi: 10.1371/journal.pbio.3000471. url: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000471 (visited on 28/03/2023).

Hung, King L. et al. (Dec. 2021). "ecDNA hubs drive cooperative intermolecular oncogene expression". en. In: *Nature* 600.7890, pp. 731–736. issn: 0028-0836, 1476-4687. doi: 10.1038/s41586-021-04116-8. url: https://www.nature.com/articles/s41586-021-04116-8 (visited on 25/04/2023).

Ihaka, Ross et al. (2022). *colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes*. url: https://CRAN.R-project.org/package=colorspace.

Jamal-Hanjani, Mariam et al. (June 2017). "Tracking the Evolution of Non-Small-Cell Lung Cancer". eng. In: *The New England Journal of Medicine* 376.22, pp. 2109–2121. issn: 1533-4406. doi: 10.1056/NEJMoa1616288.

Jones, Siân et al. (Sept. 2008). "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses". eng. In: *Science (New York, N.Y.)* 321.5897, pp. 1801–1806. issn: 1095-9203. doi: 10.1126/science.1164368.

Kamisawa, Terumi et al. (July 2016). "Pancreatic cancer". en. In: *The Lancet* 388.10039, pp. 73–85. issn: 0140-6736. doi: 10.1016/S0140-6736(16)00141-0. url: https://www.sciencedirect.com/science/article/pii/S0140673616001410 (visited on 12/05/2023).

Kanehisa, Minoru and Susumu Goto (Jan. 2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1, pp. 27–30. issn: 0305-1048. doi: 10.1093/nar/28.1.27. url: https://doi.org/10.1093/nar/28.1.27 (visited on 28/06/2022).

Kassambara, Alboukadel (2020). *ggpubr: ggplot2 Based Publication Ready Plots*. url: https://rpkgs.datanovia.com/ggpubr/.

Kassambara, Alboukadel, Marcin Kosinski and Przemyslaw Biecek (2021). *survminer: Drawing Survival Curves using ggplot2*. url: https://rpkgs.datanovia.com/survminer/index.html.

Kassambara, Alboukadel et al. (2017). "Package 'survminer'". In: *Drawing Survival Curves using 'ggplot2'.(R package version 0.3. 1.)*

Kaufman, R J, P C Brown and R T Schimke (Nov. 1979). "Amplified dihydrofolate reductase genes in unstably methotrexate-resistant cells are associated with double minute chromosomes." In: *Proceedings of the National Academy of Sciences* 76.11. Publisher: Proceedings of the National Academy of Sciences, pp. 5669–5673. doi: 10.1073/pnas.76.11.5669. url: https://www.pnas.org/doi/abs/10.1073/pnas.76.11.5669 (visited on 29/04/2023).

— (Dec. 1981). "Loss and stabilization of amplified dihydrofolate reductase genes in mouse sarcoma S-180 cell lines". eng. In: *Molecular and Cellular Biology* 1.12, pp. 1084–1093. issn: 0270-7306. doi: 10.1128/mcb.1.12.1084-1093.1981.

Keller, Laura and Klaus Pantel (Oct. 2019). "Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells". en. In: *Nature Reviews Cancer* 19.10. Number: 10 Publisher: Nature Publishing Group, pp. 553–567. issn: 1474-1768. doi: 10.1038/s41568-

019-0180-2. url: https://www.nature.com/articles/s41568-019-0180-2 (visited on 25/05/2023).

Kim, Hoon et al. (Sept. 2020). "Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers". en. In: *Nature Genetics* 52.9. Number: 9 Publisher: Nature Publishing Group, pp. 891–897. issn: 1546-1718. doi: 10.1038/s41588-020-0678-2. url: https://www.nature.com/articles/s41588-020-0678-2 (visited on 09/06/2021).

Klemm, Sandy L., Zohar Shipony and William J. Greenleaf (Apr. 2019). "Chromatin accessibility and the regulatory epigenome". en. In: *Nature Reviews Genetics* 20.4. Number: 4 Publisher: Nature Publishing Group, pp. 207–220. issn: 1471-0064. doi: 10.1038/s41576-018-0089-8. url: https://www.nature.com/articles/s41576-018-0089-8 (visited on 24/05/2023).

Knaus, Brian J. and Niklaus J. Grunwald (2022). *vcfR: Manipulate and Visualize VCF Data*. url: https://CRAN.R-project.org/package=vcfR.

Koboldt, Daniel C. et al. (Sept. 2010). "Challenges of sequencing human genomes". In: *Briefings in Bioinformatics* 11.5, pp. 484–498. issn: 1467-5463. doi: 10.1093/bib/bbq016. url: https://doi.org/10.1093/bib/bbq016 (visited on 17/11/2023).

Koche, Richard P. et al. (Jan. 2020). "Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma". en. In: *Nature Genetics* 52.1. Number: 1 Publisher: Nature Publishing Group, pp. 29–34. issn: 1546-1718. doi: 10.1038/s41588-019-0547-z. url: https://www.nature.com/articles/s41588-019-0547-z (visited on 16/10/2020).

Kohl, N. E. et al. (Dec. 1983). "Transposition and amplification of oncogene-related sequences in human neuroblastomas". eng. In: *Cell* 35.2 Pt 1, pp. 359–367. issn: 0092-8674. doi: 10.1016/0092-8674(83)90169-1.

Koning, A. P. Jason de et al. (Dec. 2011). "Repetitive Elements May Comprise Over Two-Thirds of the Human Genome". en. In: *PLOS Genetics* 7.12. Publisher: Public Library of Science, e1002384. issn: 1553-7404. doi: 10.1371/journal.pgen.1002384. url: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002384 (visited on 27/06/2022).

Korotkevich, Gennady, Vladimir Sukhov and Alexey Sergushichev (2021). *fgsea: Fast Gene Set Enrichment Analysis*. url: https://github.com/ctlab/fgsea/.

Köster, Johannes and Sven Rahmann (Oct. 2012). "Snakemake—a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19, pp. 2520–2522. issn: 1367-4803. doi: 10.1093/bioinformatics/bts480. url: https://doi.org/10.1093/bioinformatics/bts480 (visited on 09/06/2023).

Krueger, Felix (2015). "Trim galore". In: *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* 516.517.

Kudla, Grzegorz et al. (May 2006). "High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells". en. In: *PLOS Biology* 4.6. Publisher: Public Library of Science, e180. issn: 1545-7885. doi: 10.1371/journal.pbio.0040180. url: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040180 (visited on 02/08/2023).

Kumar, Pankaj et al. (Sept. 2017). “Normal and Cancerous Tissues Release Extrachromosomal Circular DNA (eccDNA) into the Circulation”. en. In: *Molecular Cancer Research* 15.9, pp. 1197–1205. issn: 1541-7786, 1557-3125. doi: 10.1158/1541-7786.MCR-17-0095. url: https://aacrjournals.org/mcr/article/15/9/1197/268041/Normal-and-Cancerous-Tissues-Release (visited on 29/06/2022).

Kumar, Pankaj et al. (May 2020). “ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines”. en. In: *Science Advances* 6.20. Publisher: American Association for the Advancement of Science Section: Research Article, eaba2489. issn: 2375-2548. doi: 10.1126/sciadv.aba2489. url: https://advances.sciencemag.org/content/6/20/eaba2489 (visited on 19/08/2020).

Kumari, Neeraj et al. (Sept. 2016). “Role of interleukin-6 in cancer progression and therapeutic resistance”. eng. In: *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine* 37.9, pp. 11553–11572. issn: 1423-0380. doi: 10.1007/s13277-016-5098-7.

Kupkova, Kristyna et al. (2021). *GenomicDistributions: fast analysis of genomic intervals with Bioconductor*. url: http://code.databio.org/GenomicDistributions.

Laird, Charles et al. (Jan. 1987). “Fragile sites in human chromosomes as regions of late-replicating DNA”. en. In: *Trends in Genetics* 3, pp. 274–281. issn: 0168-9525. doi: 10.1016/0168-9525(87)90268-X. url: https://www.sciencedirect.com/science/article/pii/016895258790268X (visited on 23/05/2023).

Lange, Joshua T. et al. (Oct. 2022). “The evolutionary dynamics of extrachromosomal DNA in human cancers”. en. In: *Nature Genetics* 54.10. Number: 10 Publisher: Nature Publishing Group, pp. 1527–1533. issn: 1546-1718. doi: 10.1038/s41588-022-01177-x. url: https://www.nature.com/articles/s41588-022-01177-x (visited on 11/10/2023).

Langfelder, Peter et al. (2022). *WGCNA: Weighted Correlation Network Analysis*. url: http://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/.

Lawrence, Michael, Vince Carey and Robert Gentleman (2021). *rtracklayer: R interface to genome annotation files and the UCSC genome browser*.

Lawson, Devon A. et al. (Dec. 2018). “Tumour heterogeneity and metastasis at single-cell resolution”. en. In: *Nature Cell Biology* 20.12. Number: 12 Publisher: Nature Publishing Group, pp. 1349–1360. issn: 1476-4679. doi: 10.1038/s41556-018-0236-7. url: https://www.nature.com/articles/s41556-018-0236-7 (visited on 25/05/2023).

Lee, Jennifer A., Claudia M. B. Carvalho and James R. Lupski (Dec. 2007). “A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders”. eng. In: *Cell* 131.7, pp. 1235–1247. issn: 0092-8674. doi: 10.1016/j.cell.2007.11.037.

Lee, Stuart, Michael Lawrence and Dianne Cook (2021). *plyranges: A fluent interface for manipulating GenomicRanges*.

Leibowitz, Mitchell L., Cheng-Zhong Zhang and David Pellman (2015). “Chromothripsis: A New Mechanism for Rapid Karyotype Evolution”. eng. In: *Annual Review of Genetics* 49, pp. 183–211. issn: 1545-2948. doi: 10.1146/annurev-genet-120213-092228.

Leipzig, Jeremy (May 2017). “A review of bioinformatic pipeline frameworks”. In: *Briefings in Bioinformatics* 18.3, pp. 530–536. issn: 1467-5463. doi: 10.1093/bib/bbw020. url: https://doi.org/10.1093/bib/bbw020 (visited on 08/06/2023).

Levan, Albert and GöRAN Levan (1978). “Have double minutes functioning centromeres?” en. In: *Hereditas* 88.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1601-5223.1978.tb01606.x, pp. 81–92. issn: 1601-5223. doi: 10.1111/j.1601-5223.1978.tb01606.x. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1601-5223.1978.tb01606.x (visited on 25/04/2023).

Li, Fuyu et al. (June 2023). *Detecting Full-Length EccDNA with FLED and long-reads sequencing*. en. Pages: 2023.06.21.545840 Section: New Results. doi: 10.1101/2023.06.21.545840. url: https://www.biorxiv.org/content/10.1101/2023.06.21.545840v1 (visited on 04/10/2023).

Li, Heng (2012). “seqtk Toolkit for processing sequences in FASTA/Q formats”. In: *GitHub* 767, p. 69.

— (2013). “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: *arXiv preprint arXiv:1303.3997*.

— (Sept. 2018). “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18, pp. 3094–3100. issn: 1367-4803. doi: 10.1093/bioinformatics/bty191. url: https://doi.org/10.1093/bioinformatics/bty191 (visited on 28/06/2022).

Li, Heng et al. (Aug. 2009). “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16, pp. 2078–2079. issn: 1367-4803. doi: 10.1093/bioinformatics/btp352. url: https://doi.org/10.1093/bioinformatics/btp352 (visited on 28/06/2022).

Liao, Zhenyu et al. (Aug. 2020). “Classification of extrachromosomal circular DNA with a focus on the role of extrachromosomal DNA (ecDNA) in tumor heterogeneity and progression”. en. In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1874.1, p. 188392. issn: 0304-419X. doi: 10.1016/j.bbcan.2020.188392. url: https://www.sciencedirect.com/science/article/pii/S0304419X20301116 (visited on 04/04/2023).

Liberzon, Arthur et al. (June 2011). “Molecular signatures database (MSigDB) 3.0”. In: *Bioinformatics* 27.12, pp. 1739–1740. issn: 1367-4803. doi: 10.1093/bioinformatics/btr260. url: https://doi.org/10.1093/bioinformatics/btr260 (visited on 28/06/2022).

Liberzon, Arthur et al. (Dec. 2015). “The Molecular Signatures Database Hallmark Gene Set Collection”. English. In: *Cell Systems* 1.6. Publisher: Elsevier, pp. 417–425. issn: 2405-4712. doi: 10.1016/j.cels.2015.12.004. url: https://www.cell.com/cell-systems/abstract/S2405-4712(15)00218-5 (visited on 04/07/2023).

Ling, Xiaoxuan et al. (Sept. 2021). “Small extrachromosomal circular DNA (eccDNA): major functions in evolution and cancer”. In: *Molecular Cancer* 20.1, p. 113. issn: 1476-4598. doi: 10.1186/s12943-021-01413-8. url: https://doi.org/10.1186/s12943-021-01413-8 (visited on 07/09/2023).

Liu, Yining, Jingchun Sun and Min Zhao (Feb. 2017). “ONGene: A literature-based database for human oncogenes”. en. In: *Journal of Genetics and Genomics* 44.2, pp. 119–121. issn:

1673-8527. doi: 10.1016/j.jgg.2016.12.004. url: https://www.sciencedirect.com/science/article/pii/S1673852716302053 (visited on 10/07/2023).

Lo, Anthony W. l. et al. (Jan. 2002). "DNA Amplification by Breakage/Fusion/Bridge Cycles Initiated by Spontaneous Telomere Loss in a Human Cancer Cell Line". en. In: *Neoplasia* 4.6, pp. 531–538. issn: 1476-5586. doi: 10.1038/sj.neo.7900267. url: https://www.sciencedirect.com/science/article/pii/S1476558602800586 (visited on 26/04/2023).

Lomberk, Gwen et al. (May 2018). "Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes". en. In: *Nature Communications* 9.1. Number: 1 Publisher: Nature Publishing Group, p. 1978. issn: 2041-1723. doi: 10.1038/s41467-018-04383-6. url: https://www.nature.com/articles/s41467-018-04383-6 (visited on 12/01/2023).

Love, Michael, Simon Anders and Wolfgang Huber (2021). *DESeq2: Differential gene expression analysis based on the negative binomial distribution*. url: https://github.com/mikelove/DESeq2.

Love, Michael I., Wolfgang Huber and Simon Anders (Dec. 2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550. issn: 1474-760X. doi: 10.1186/s13059-014-0550-8. url: https://doi.org/10.1186/s13059-014-0550-8 (visited on 09/06/2021).

Luebeck, Jens et al. (Jan. 2020). "AmpliconReconstructor: Integrated analysis of NGS and optical mapping resolves the complex structures of focal amplifications in cancer". en. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.01.22.916031. doi: 10.1101/2020.01.22.916031. url: https://www.biorxiv.org/content/10.1101/2020.01.22.916031v2 (visited on 12/10/2020).

Luebeck, Jens et al. (Apr. 2023). "Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus". en. In: *Nature*. Publisher: Nature Publishing Group, pp. 1–8. issn: 1476-4687. doi: 10.1038/s41586-023-05937-5. url: https://www.nature.com/articles/s41586-023-05937-5 (visited on 13/04/2023).

Macheret, Morgane and Thanos D. Halazonetis (2015). "DNA Replication Stress as a Hallmark of Cancer". In: *Annual Review of Pathology: Mechanisms of Disease* 10.1. _eprint: https://doi.org/10.1146/annurev-pathol-012414-040424, pp. 425–448. doi: 10.1146/annurev-pathol-012414-040424. url: https://doi.org/10.1146/annurev-pathol-012414-040424 (visited on 04/05/2023).

Maddipati, Ravikanth et al. (Feb. 2022). "MYC Levels Regulate Metastatic Heterogeneity in Pancreatic Adenocarcinoma". In: *Cancer Discovery* 12.2, pp. 542–561. issn: 2159-8274. doi: 10.1158/2159-8290.CD-20-1826. url: https://doi.org/10.1158/2159-8290.CD-20-1826 (visited on 30/05/2023).

Maintainer, Bioconductor Package et al. (2021). *VariantAnnotation: Annotation of Genetic Variants*.

Maitra, Anirban et al. (Oct. 2005). "Genomic alterations in cultured human embryonic stem cells". en. In: *Nature Genetics* 37.10. Number: 10 Publisher: Nature Publishing Group, pp. 1099–1103. issn: 1546-1718. doi: 10.1038/ng1631. url: https://www.nature.com/articles/ng1631 (visited on 23/03/2023).

Malumbres, Marcos and Mariano Barbacid (Mar. 2009). “Cell cycle, CDKs and cancer: a changing paradigm”. en. In: *Nature Reviews Cancer* 9.3. Number: 3 Publisher: Nature Publishing Group, pp. 153–166. issn: 1474-1768. doi: 10.1038/nrc2602. url: https://www.nature.com/articles/nrc2602 (visited on 04/07/2023).

Mantione, Kirk J et al. (Aug. 2014). “Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq”. In: *Medical Science Monitor Basic Research* 20, pp. 138–141. issn: 2325-4394. doi: 10.12659/MSMBR.892101. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152252/ (visited on 31/05/2023).

Martin, Marcel (May 2011). “Cutadapt removes adapter sequences from high-throughput sequencing reads”. en. In: *EMBnet.journal* 17.1. Number: 1, pp. 10–12. issn: 2226-6089. doi: 10.14806/ej.17.1.200. url: https://journal.embnet.org/index.php/embnetjournal/article/view/200 (visited on 28/06/2022).

Mayakonda, Anand (2022). *maftools: Summarize, Analyze and Visualize MAF Files*. url: https://github.com/PoisonAlien/maftools.

McClintock, B. (Mar. 1941). “The Stability of Broken Ends of Chromosomes in Zea Mays”. eng. In: *Genetics* 26.2, pp. 234–282. issn: 0016-6731. doi: 10.1093/genetics/26.2.234.

McKenna, Aaron et al. (Jan. 2010). “The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data”. en. In: *Genome Research* 20.9. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1297–1303. issn: 1088-9051, 1549-5469. doi: 10.1101/gr.107524.110. url: https://genome.cshlp.org/content/20/9/1297 (visited on 09/02/2023).

McWhite, Claire D. and Claus O. Wilke (2021). *colorblindr: Simulate colorblindness in R figures*. url: https://github.com/clauswilke/colorblindr.

Ménard, Sylvie et al. (Sept. 2003). “Biologic and therapeutic role of HER2 in cancer”. en. In: *Oncogene* 22.42. Number: 42 Publisher: Nature Publishing Group, pp. 6570–6578. issn: 1476-5594. doi: 10.1038/sj.onc.1206779. url: https://www.nature.com/articles/1206779 (visited on 31/05/2023).

Miga, Karen H. (Sept. 2020). “Centromere studies in the era of ‘telomere-to-telomere’ genomics”. en. In: *Experimental Cell Research* 394.2, p. 112127. issn: 0014-4827. doi: 10.1016/j.yexcr.2020.112127. url: https://www.sciencedirect.com/science/article/pii/S0014482720303748 (visited on 28/03/2023).

Mijit, Mahmut et al. (Mar. 2020). “Role of p53 in the Regulation of Cellular Senescence”. In: *Biomolecules* 10.3, p. 420. issn: 2218-273X. doi: 10.3390/biom10030420. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7175209/ (visited on 06/06/2023).

Mizrahi, Jonathan D et al. (June 2020). “Pancreatic cancer”. en. In: *The Lancet* 395.10242, pp. 2008–2020. issn: 0140-6736. doi: 10.1016/S0140-6736(20)30974-0. url: https://www.sciencedirect.com/science/article/pii/S0140673620309740 (visited on 31/05/2023).

Modrich, Paul (Dec. 1994). “Mismatch Repair, Genetic Stability, and Cancer”. en. In: *Science* 266.5193, pp. 1959–1960. issn: 0036-8075, 1095-9203. doi: 10.1126/science.7801122. url: https://www.science.org/doi/10.1126/science.7801122 (visited on 25/05/2023).

Moffitt, Richard A et al. (Oct. 2015). “Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma”. en. In: *Nature Genetics* 47.10, pp. 1168–1178. issn: 1061-4036, 1546-1718. doi: 10.1038/ng.3398. url: http://www.nature.com/articles/ng.3398 (visited on 27/06/2022).

Møller, Henrik D. et al. (June 2015). “Extrachromosomal circular DNA is common in yeast”. en. In: *Proceedings of the National Academy of Sciences* 112.24. Publisher: National Academy of Sciences Section: PNAS Plus, E3114–E3122. issn: 0027-8424, 1091-6490. doi: 10.1073/pnas.1508825112. url: https://www.pnas.org/content/112/24/E3114 (visited on 03/06/2021).

Møller, Henrik Devitt (2020). “Circle-Seq: Isolation and Sequencing of Chromosome-Derived Circular DNA Elements in Cells”. en. In: *DNA Electrophoresis: Methods and Protocols*. Ed. by Katsuhiro Hanada. Methods in Molecular Biology. New York, NY: Springer US, pp. 165–181. isbn: 978-1-07-160323-9. doi: 10.1007/978-1-0716-0323-9_15. url: https://doi.org/10.1007/978-1-0716-0323-9_15 (visited on 24/11/2022).

Møller, Henrik Devitt et al. (2018a). “Circular DNA elements of chromosomal origin are common in healthy human somatic tissue”. In: *Nature communications* 9.1, p. 1069.

Møller, Henrik Devitt et al. (Dec. 2018b). “CRISPR-C: circularization of genes and chromosome by CRISPR in human cells”. In: *Nucleic Acids Research* 46.22, e131. issn: 0305-1048. doi: 10.1093/nar/gky767. url: https://doi.org/10.1093/nar/gky767 (visited on 26/04/2023).

Møller, Henrik Devitt et al. (Feb. 2020). “Near-Random Distribution of Chromosome-Derived Circular DNA in the Condensed Genome of Pigeons and the Larger, More Repeat-Rich Human Genome”. In: *Genome Biology and Evolution* 12.2, pp. 3762–3777. issn: 1759-6653. doi: 10.1093/gbe/evz281. url: https://doi.org/10.1093/gbe/evz281 (visited on 23/03/2023).

Morgan, Martin et al. (2021). *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import*. url: https://bioconductor.org/packages/Rsamtools.

Morrison, Olivia and Jitendra Thakur (Jan. 2021). “Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin”. en. In: *International Journal of Molecular Sciences* 22.13. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute, p. 6922. issn: 1422-0067. doi: 10.3390/ijms22136922. url: https://www.mdpi.com/1422-0067/22/13/6922 (visited on 23/05/2023).

Morton, Andrew R. et al. (Nov. 2019). “Functional Enhancers Shape Extrachromosomal Oncogene Amplifications”. English. In: *Cell* 179.6. Publisher: Elsevier, 1330–1341.e13. issn: 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.10.039. url: https://www.cell.com/cell/abstract/S0092-8674(19)31216-4 (visited on 29/04/2023).

Murnane, John P. and Laure Sabatier (2004). “Chromosome rearrangements resulting from telomere dysfunction and their role in cancer”. en. In: *BioEssays* 26.11. _eprint: https://on-

linelibrary.wiley.com/doi/pdf/10.1002/bies.20125, pp. 1164–1174. issn: 1521-1878. doi: 10.1002/bies.20125. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.20125 (visited on 26/04/2023).

Nam, Chehyun et al. (Jan. 2022). "Genomic and Epigenomic Characterization of Tumor Organoid Models". en. In: *Cancers* 14.17. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute, p. 4090. issn: 2072-6694. doi: 10.3390/cancers14174090. url: https://www.mdpi.com/2072-6694/14/17/4090 (visited on 17/04/2023).

Nathanson, David A. et al. (Jan. 2014). "Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA". en. In: *Science* 343.6166. Publisher: American Association for the Advancement of Science Section: Report, pp. 72–76. issn: 0036-8075, 1095-9203. doi: 10.1126/science.1241328. url: https://science.sciencemag.org/content/343/6166/72 (visited on 16/02/2021).

Negrini, Simona, Vassilis G. Gorgoulis and Thanos D. Halazonetis (Mar. 2010). "Genomic instability — an evolving hallmark of cancer". en. In: *Nature Reviews Molecular Cell Biology* 11.3. Number: 3 Publisher: Nature Publishing Group, pp. 220–228. issn: 1471-0080. doi: 10.1038/nrm2858. url: https://www.nature.com/articles/nrm2858 (visited on 21/06/2021).

Neoptolemos, John P. et al. (Oct. 2023). "Personalized treatment in localized pancreatic cancer". en. In: *European Surgery*. Company: Springer Distributor: Springer Institution: Springer Label: Springer Publisher: Springer Vienna, pp. 1–17. issn: 1682-4016. doi: 10.1007/s10353-023-00814-x. url: https://link.springer.com/article/10.1007/s10353-023-00814-x (visited on 10/10/2023).

Neuwirth, Erich (2022). *RColorBrewer: ColorBrewer Palettes*. url: https://CRAN.R-project.org/package=RColorBrewer.

Noer, Julie B. et al. (July 2022). "Extrachromosomal circular DNA in cancer: history, current knowledge, and methods". en. In: *Trends in Genetics* 38.7, pp. 766–781. issn: 0168-9525. doi: 10.1016/j.tig.2022.02.007. url: https://www.sciencedirect.com/science/article/pii/S0168952522000348 (visited on 03/04/2023).

Norman, Anders et al. (Aug. 2014). "An Improved Method for Including Upper Size Range Plasmids in Metamobilomes". en. In: *PLOS ONE* 9.8. Publisher: Public Library of Science, e104405. issn: 1932-6203. doi: 10.1371/journal.pone.0104405. url: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104405 (visited on 06/10/2023).

Norppa, Hannu and Ghita C.-M. Falck (May 2003). "What do human micronuclei contain?" In: *Mutagenesis* 18.3, pp. 221–233. issn: 0267-8357. doi: 10.1093/mutage/18.3.221. url: https://doi.org/10.1093/mutage/18.3.221 (visited on 02/05/2023).

Notta, Faiyaz et al. (Oct. 2016). "A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns". en. In: *Nature* 538.7625. Number: 7625 Publisher: Nature Publishing Group, pp. 378–382. issn: 1476-4687. doi: 10.1038/nature19823. url: https://www.nature.com/articles/nature19823 (visited on 16/02/2021).

Nowell, Peter C. (Oct. 1976). "The Clonal Evolution of Tumor Cell Populations". In: *Science* 194.4260. Publisher: American Association for the Advancement of Science, pp. 23–28.

doi: 10.1126/science.959840. url: https://www.science.org/doi/abs/10.1126/science.959840 (visited on 06/06/2023).

O'Connor, Mark J. (Nov. 2015). "Targeting the DNA Damage Response in Cancer". en. In: *Molecular Cell* 60.4, pp. 547–560. issn: 1097-2765. doi: 10.1016/j.molcel.2015.10.040. url: https://www.sciencedirect.com/science/article/pii/S109727651500831X (visited on 06/06/2023).

Okonechnikov, Konstantin et al. (Apr. 2012). "Unipro UGENE: a unified bioinformatics toolkit". In: *Bioinformatics* 28.8, pp. 1166–1167. issn: 1367-4803. doi: 10.1093/bioinformatics/bts091. url: https://doi.org/10.1093/bioinformatics/bts091 (visited on 07/06/2023).

Oobatake, Yoshihiro and Noriaki Shimizu (Mar. 2020). "Double-strand breakage in the extrachromosomal double minutes triggers their aggregation in the nucleus, micronucleation, and morphological transformation". eng. In: *Genes, Chromosomes & Cancer* 59.3, pp. 133–143. issn: 1098-2264. doi: 10.1002/gcc.22810.

Ortega, Miguel A. et al. (Apr. 2022). "Implication of ERBB2 as a Predictive Tool for Survival in Patients with Pancreatic Cancer in Histological Studies". en. In: *Current Oncology* 29.4. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, pp. 2442–2453. issn: 1718-7729. doi: 10.3390/curroncol29040198. url: https://www.mdpi.com/1718-7729/29/4/198 (visited on 31/05/2023).

Pagès, H. et al. (2021). *Biostrings: Efficient manipulation of biological strings*. url: https://bioconductor.org/packages/Biostrings.

Pancreatic cancer statistics (May 2015). en. url: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer (visited on 31/05/2023).

Paszkiewicz, Konrad and David J. Studholme (Sept. 2010). "De novo assembly of short sequence reads". In: *Briefings in Bioinformatics* 11.5, pp. 457–472. issn: 1467-5463. doi: 10.1093/bib/bbq020. url: https://doi.org/10.1093/bib/bbq020 (visited on 09/05/2023).

Patel, Harshil et al. (July 2020). *nf-core/atacseq: nf-core/atacseq v1.2.0 - Iron Swan*. doi: 10.5281/zenodo.3928140. url: https://doi.org/10.5281/zenodo.3928140.

Patwardhan, Mayura et al. (2021). *bedtoolsr: Bedtools Wrapper*.

Paulsen, Teressa et al. (Apr. 2018). "Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells". en. In: *Trends in Genetics* 34.4, pp. 270–278. issn: 0168-9525. doi: 10.1016/j.tig.2017.12.010. url: https://www.sciencedirect.com/science/article/pii/S0168952517302305 (visited on 18/06/2021).

Paulsen, Teressa et al. (May 2019). "Small extrachromosomal circular DNAs, microDNA, produce short regulatory RNAs that suppress gene expression independent of canonical promoters". eng. In: *Nucleic Acids Research* 47.9, pp. 4586–4596. issn: 1362-4962. doi: 10.1093/nar/gkz155.

Paulsen, Teressa et al. (Nov. 2021). "MicroDNA levels are dependent on MMEJ, repressed by c-NHEJ pathway, and stimulated by DNA damage". In: *Nucleic Acids Research* 49.20, pp. 11787–11799. issn: 0305-1048. doi: 10.1093/nar/gkab984. url: https://doi.org/10.1093/nar/gkab984 (visited on 28/09/2023).

Pedersen, Brent (Dec. 2022a). *bigly: a pileup library that embraces the huge*. original-date: 2016-11-09T19:30:20Z. url: https://github.com/brentp/bigly (visited on 16/01/2023).

Pedersen, Thomas Lin (2022b). *ggforce: Accelerating ggplot2*. url: https://CRAN.R-project.org/package=ggforce.

— (2022c). *patchwork: The Composer of Plots*. url: https://CRAN.R-project.org/package=patchwork.

Peng, Junya et al. (Sept. 2019). "Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma". en. In: *Cell Research* 29.9. Number: 9 Publisher: Nature Publishing Group, pp. 725–738. issn: 1748-7838. doi: 10.1038/s41422-019-0195-y. url: https://www.nature.com/articles/s41422-019-0195-y (visited on 04/07/2023).

Pishvaian, Michael J. et al. (Oct. 2018). "Molecular Profiling of Patients with Pancreatic Cancer: Initial Results from the Know Your Tumor Initiative". In: *Clinical Cancer Research* 24.20, pp. 5018–5027.

Pishvaian, Michael J. et al. (Apr. 2020). "Overall survival in patients with pancreatic cancer receiving matched therapies following molecular profiling: a retrospective analysis of the Know Your Tumor registry trial". English. In: *The Lancet Oncology* 21.4. Publisher: Elsevier, pp. 508–518. issn: 1470-2045, 1474-5488. doi: 10.1016/S1470-2045(20)30074-7. url: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(20)30074-7/fulltext (visited on 21/09/2021).

Prada-Luengo, Iñigo et al. (Dec. 2019). "Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads". In: *BMC Bioinformatics* 20.1, p. 663. issn: 1471-2105. doi: 10.1186/s12859-019-3160-3. url: https://doi.org/10.1186/s12859-019-3160-3 (visited on 27/06/2022).

Pucci, Ferdinando et al. (Nov. 2016). "PF4 Promotes Platelet Production and Lung Cancer Growth". en. In: *Cell Reports* 17.7, pp. 1764–1772. issn: 22111247. doi: 10.1016/j.celrep.2016.10.031. url: https://linkinghub.elsevier.com/retrieve/pii/S221112471631436X (visited on 19/10/2023).

Puleo, Francesco et al. (Dec. 2018). "Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features". en. In: *Gastroenterology* 155.6, 1999–2013.e3. issn: 0016-5085. doi: 10.1053/j.gastro.2018.08.033. url: https://www.sciencedirect.com/science/article/pii/S0016508518349199 (visited on 24/11/2022).

Purshouse, Karin et al. (Dec. 2022). "Oncogene expression from extrachromosomal DNA is driven by copy number amplification and does not require spatial clustering in glioblastoma stem cells". In: *eLife* 11. Ed. by Jessica K Tyler. Publisher: eLife Sciences Publications, Ltd, e80207. issn: 2050-084X. doi: 10.7554/eLife.80207. url: https://doi.org/10.7554/eLife.80207 (visited on 28/04/2023).

Quinlan, Aaron R. and Ira M. Hall (Mar. 2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842. issn: 1367-4803. doi: 10.1093/bioinformatics/btq033. url: https://doi.org/10.1093/bioinformatics/btq033 (visited on 28/06/2022).

Radulovich, Nikolina et al. (Feb. 2010). “Differential roles of cyclin D1 and D3 in pancreatic ductal adenocarcinoma”. In: *Molecular Cancer* 9.1, p. 24. issn: 1476-4598. doi: 10.1186/1476-4598-9-24. url: https://doi.org/10.1186/1476-4598-9-24 (visited on 30/06/2023).

Raghavan, Srivatsan et al. (Dec. 2021). “Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer”. en. In: *Cell* 184.25, 6119–6137.e26. issn: 0092-8674. doi: 10.1016/j.cell.2021.11.017. url: https://www.sciencedirect.com/science/article/pii/S0092867421013325 (visited on 30/06/2023).

Rajkumar, Utkrisht et al. (Nov. 2019). “EcSeg: Semantic Segmentation of Metaphase Images Containing Extrachromosomal DNA”. en. In: *iScience* 21, pp. 428–435. issn: 2589-0042. doi: 10.1016/j.isci.2019.10.035. url: https://www.sciencedirect.com/science/article/pii/S2589004219304158 (visited on 05/05/2023).

Ramírez, Fidel et al. (July 2014). “deepTools: a flexible platform for exploring deep-sequencing data”. In: *Nucleic Acids Research* 42.W1, W187–W191. issn: 0305-1048. doi: 10.1093/nar/gku365. url: https://doi.org/10.1093/nar/gku365 (visited on 25/11/2022).

Rausch, Tobias et al. (Jan. 2012). “Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations”. en. In: *Cell* 148.1, pp. 59–71. issn: 0092-8674. doi: 10.1016/j.cell.2011.12.013. url: https://www.sciencedirect.com/science/article/pii/S0092867411015169 (visited on 26/04/2023).

Rayeroux, Kathleen C. and Lynda J. Campbell (Aug. 2009). “Gene amplification in myeloid leukemias elucidated by fluorescence in situ hybridization”. en. In: *Cancer Genetics and Cytogenetics* 193.1, pp. 44–53. issn: 0165-4608. doi: 10.1016/j.cancergencyto.2009.04.006. url: https://www.sciencedirect.com/science/article/pii/S0165460809002088 (visited on 04/05/2023).

Rennoll, Sherri and Gregory Yochum (Nov. 2015). “Regulation of MYC gene expression by aberrant Wnt/β-catenin signaling in colorectal cancer”. In: *World Journal of Biological Chemistry* 6.4, pp. 290–300. issn: 1949-8454. doi: 10.4331/wjbc.v6.i4.290. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4657124/ (visited on 01/02/2023).

Rosswog, Carolina et al. (Dec. 2021). “Chromothripsis followed by circular recombination drives oncogene amplification in human cancer”. en. In: *Nature Genetics* 53.12. Number: 12 Publisher: Nature Publishing Group, pp. 1673–1685. issn: 1546-1718. doi: 10.1038/s41588-021-00951-7. url: https://www.nature.com/articles/s41588-021-00951-7 (visited on 24/04/2023).

Rückert, Felix et al. (Jan. 2012). “Five Primary Human Pancreatic Adenocarcinoma Cell Lines Established by the Outgrowth Method”. en. In: *Journal of Surgical Research* 172.1, pp. 29–39. issn: 0022-4804. doi: 10.1016/j.jss.2011.04.021. url: https://www.sciencedirect.com/science/article/pii/S0022480411003714 (visited on 29/11/2022).

Salgia, Ravi and Prakash Kulkarni (Feb. 2018). “The Genetic/Non-genetic Duality of Drug ‘Resistance’ in Cancer”. en. In: *Trends in Cancer* 4.2, pp. 110–118. issn: 2405-8033. doi: 10.1016/j.trecan.2018.01.001. url: https://www.sciencedirect.com/science/article/pii/S2405803318300013 (visited on 28/04/2023).

Samstein, Robert M. et al. (Feb. 2019). “Tumor mutational load predicts survival after immunotherapy across multiple cancer types”. en. In: *Nature Genetics* 51.2. Number: 2 Publisher: Nature Publishing Group, pp. 202–206. issn: 1546-1718. doi: 10.1038/s41588-018-0312-8. url: https://www.nature.com/articles/s41588-018-0312-8 (visited on 09/10/2023).

Sarantis, Panagiotis et al. (Feb. 2020). “Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy”. In: *World Journal of Gastrointestinal Oncology* 12.2, pp. 173–181. issn: 1948-5204. doi: 10.4251/wjgo.v12.i2.173. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7031151/ (visited on 31/05/2023).

Schimke, R. T. et al. (Jan. 1981). “Chromosomal and Extrachromosomal Localization of Amplified Dihydrofolate Reductase Genes in Cultured Mammalian Cells”. en. In: *Cold Spring Harbor Symposia on Quantitative Biology* 45. Publisher: Cold Spring Harbor Laboratory Press, pp. 785–797. issn: 0091-7451, 1943-4456. doi: 10.1101/SQB.1981.045.01.097. url: http://symposium.cshlp.org/content/45/785 (visited on 03/05/2023).

Schimke, R. T. et al. (Apr. 1986). “Overreplication and recombination of DNA in higher eukaryotes: potential consequences and biological implications”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 83.7, pp. 2157–2161. issn: 0027-8424. doi: 10.1073/pnas.83.7.2157.

Schoenlein, Patricia V. et al. (Mar. 2003). “Radiation therapy depletes extrachromosomally amplified drug resistance genes and oncogenes from tumor cells via micronuclear capture of episomes and double minute chromosomes”. eng. In: *International Journal of Radiation Oncology, Biology, Physics* 55.4, pp. 1051–1065. issn: 0360-3016. doi: 10.1016/s0360-3016(02)04473-5.

Seino, Takashi et al. (Mar. 2018). “Human Pancreatic Tumor Organoids Reveal Loss of Stem Cell Niche Factor Dependence during Disease Progression”. English. In: *Cell Stem Cell* 22.3. Publisher: Elsevier, 454–467.e6. issn: 1934-5909, 1875-9777. doi: 10.1016/j.stem.2017.12.009. url: https://www.cell.com/cell-stem-cell/abstract/S1934-5909(17)30510-6 (visited on 30/03/2020).

Shay, Jerry W. and Woodring E. Wright (Dec. 2011). “Role of telomeres and telomerase in cancer”. en. In: *Seminars in Cancer Biology*. Cellular Senescence - A Barrier Against Tumor Development? 21.6, pp. 349–353. issn: 1044-579X. doi: 10.1016/j.semcancer.2011.10.001. url: https://www.sciencedirect.com/science/article/pii/S1044579X11000642 (visited on 11/05/2023).

Shibata, Yoshiyuki et al. (Apr. 2012). “Extrachromosomal MicroDNAs and Chromosomal Microdeletions in Normal Tissues”. In: *Science* 336.6077. Publisher: American Association for the Advancement of Science, pp. 82–86. doi: 10.1126/science.1213307. url: https://www.science.org/doi/10.1126/science.1213307 (visited on 31/08/2022).

Shimizu, Noriaki, Naomi Misaka and Koh-ichi Utani (2007). “Nonselective DNA damage induced by a replication inhibitor results in the selective elimination of extrachromosomal double minutes from human cancer cells”. en. In: *Genes, Chromosomes and Cancer* 46.10. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gcc.20473, pp. 865–874. issn:

1098-2264. doi: 10.1002/gcc.20473. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.20473 (visited on 02/05/2023).

Shimizu, Noriaki, Tsutomu Shimura and Tsubasa Tanaka (Mar. 2000). "Selective elimination of acentric double minutes from cancer cells through the extrusion of micronuclei". en. In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 448.1, pp. 81–90. issn: 0027-5107. doi: 10.1016/S0027-5107(00)00003-8. url: https://www.sciencedirect.com/science/article/pii/S0027510700000038 (visited on 22/03/2022).

Shoshani, Ofer et al. (Mar. 2021). "Chromothripsis drives the evolution of gene amplification in cancer". en. In: *Nature* 591.7848. Number: 7848 Publisher: Nature Publishing Group, pp. 137–141. issn: 1476-4687. doi: 10.1038/s41586-020-03064-z. url: https://www.nature.com/articles/s41586-020-03064-z (visited on 13/01/2023).

Siegel, Rebecca L. et al. (2023). "Cancer statistics, 2023". en. In: *CA: A Cancer Journal for Clinicians* 73.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21763, pp. 17–48. issn: 1542-4863. doi: 10.3322/caac.21763. url: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21763 (visited on 31/05/2023).

Singer, M. J. et al. (July 2000). "Amplification of the human dihydrofolate reductase gene via double minutes is initiated by chromosome breaks". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 97.14, pp. 7921–7926. issn: 0027-8424. doi: 10.1073/pnas.130194897.

Skourti-Stathaki, Konstantina and Nicholas J. Proudfoot (July 2014). "A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression". eng. In: *Genes & Development* 28.13, pp. 1384–1396. issn: 1549-5477. doi: 10.1101/gad.242990.114.

Slowikowski, Kamil (2021). *ggrepel: Automatically Position Non-Overlapping Text Labels with ggplot2*. url: https://github.com/slowkow/ggrepel.

Smith, Charles Allen and Jerome Vinograd (Aug. 1972). "Small polydisperse circular DNA of HeLa cells". en. In: *Journal of Molecular Biology* 69.2, pp. 163–178. issn: 0022-2836. doi: 10.1016/0022-2836(72)90222-7. url: https://www.sciencedirect.com/science/article/pii/0022283672902227 (visited on 03/04/2023).

Sodir, Nicole M. et al. (Apr. 2020). "MYC Instructs and Maintains Pancreatic Adenocarcinoma Phenotype". In: *Cancer Discovery* 10.4, pp. 588–607. issn: 2159-8274. doi: 10.1158/2159-8290.CD-19-0435. url: https://doi.org/10.1158/2159-8290.CD-19-0435 (visited on 31/05/2023).

Springfeld, Christoph et al. (May 2023). "Neoadjuvant therapy for pancreatic cancer". eng. In: *Nature Reviews. Clinical Oncology* 20.5, pp. 318–337. issn: 1759-4782. doi: 10.1038/s41571-023-00746-1.

Stark, Rory and Gord Brown (2022). *DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data*. url: https://www.cruk.cam.ac.uk/core-facilities/bioinformatics-core/software/DiffBind.

Steele, Christopher D. et al. (June 2022). "Signatures of copy number alterations in human cancer". en. In: *Nature* 606.7916. Number: 7916 Publisher: Nature Publishing Group,

pp. 984–991. issn: 1476-4687. doi: 10.1038/s41586-022-04738-6. url: https://www.nature.com/articles/s41586-022-04738-6 (visited on 24/04/2023).

Stephens, Philip J. et al. (Jan. 2011). "Massive genomic rearrangement acquired in a single catastrophic event during cancer development". eng. In: *Cell* 144.1, pp. 27–40. issn: 1097-4172. doi: 10.1016/j.cell.2010.11.055.

Storlazzi, Clelia Tiziana et al. (Mar. 2006). "MYC-containing double minutes in hematologic malignancies: evidence in favor of the episome model and exclusion of MYC as the target gene". In: *Human Molecular Genetics* 15.6, pp. 933–942. issn: 0964-6906. doi: 10.1093/hmg/ddl010. url: https://doi.org/10.1093/hmg/ddl010 (visited on 26/04/2023).

Storlazzi, Clelia Tiziana et al. (Jan. 2010). "Gene amplification as double minutes or homogeneously staining regions in solid tumors: Origin and structure". en. In: *Genome Research* 20.9. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1198–1206. issn: 1088-9051, 1549-5469. doi: 10.1101/gr.106252.110. url: https://genome.cshlp.org/content/20/9/1198 (visited on 24/04/2023).

Su, Zhangli et al. (2021). "ATAC-Seq-based Identification of Extrachromosomal Circular DNA in Mammalian Cells and Its Validation Using Inverse PCR and FISH". en. In: *BIO-PROTOCOL* 11.9. issn: 2331-8325. doi: 10.21769/BioProtoc.4003. url: https://bio-protocol.org/e4003 (visited on 05/09/2023).

Subramanian, Aravind, Pablo Tamayo and Anthony Castanza (2019). *GSEA: Gene Set Enrichment Analysis (GSEA)*. url: http://gsea-msigdb.org.

Sugimoto, Masataka et al. (Nov. 1999). "Regulation of CDK4 activity by a novel CDK4-binding protein, p34SEI-1". en. In: *Genes & Development* 13.22. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 3027–3033. issn: 0890-9369, 1549-5477. url: http://genesdev.cshlp.org/content/13/22/3027 (visited on 04/05/2023).

Sumner, A. T. (May 1982). "The nature and mechanisms of chromosome banding". en. In: *Cancer Genetics and Cytogenetics* 6.1, pp. 59–87. issn: 0165-4608. doi: 10.1016/0165-4608(82)90022-X. url: https://www.sciencedirect.com/science/article/pii/016546088290022X (visited on 23/05/2023).

Sun, Zhenguo et al. (Sept. 2021). "Extrachromosomal circular DNAs are common and functional in esophageal squamous cell carcinoma". In: *Annals of Translational Medicine* 9.18, p. 1464. issn: 2305-5839. doi: 10.21037/atm-21-4372. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8506789/ (visited on 08/11/2021).

Swords, Douglas S et al. (Dec. 2016). "Biomarkers in pancreatic adenocarcinoma: current perspectives". In: *OncoTargets and Therapy* 9. Publisher: Dove Medical Press _eprint: https://www.tandfonline.com/doi/pdf/10.2147/OTT.S100510, pp. 7459–7467. issn: null. doi: 10.2147/OTT.S100510. url: https://www.tandfonline.com/doi/abs/10.2147/OTT.S100510 (visited on 31/05/2023).

Talevich, Eric et al. (Apr. 2016). “CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing”. en. In: *PLOS Computational Biology* 12.4. Publisher: Public Library of Science, e1004873. issn: 1553-7358. doi: 10.1371/journal.pcbi.1004873. url: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004873 (visited on 28/06/2022).

Tang, Dong-Jiang et al. (Aug. 2005). “Oncogenic Transformation by SEI-1 Is Associated with Chromosomal Instability”. In: *Cancer Research* 65.15, pp. 6504–6508. issn: 0008-5472. doi: 10.1158/0008-5472.CAN-05-0351. url: https://doi.org/10.1158/0008-5472.CAN-05-0351 (visited on 04/05/2023).

Therneau, Terry M and Thomas Lumley (2015). “Package ‘survival’”. In: *R Top Doc* 128.10, pp. 28–33.

Therneau, Terry M. (2022). *survival: Survival Analysis*. url: https://github.com/therneau/survival.

Thomas, B. J. and R. Rothstein (Feb. 1989). “Elevated recombination rates in transcriptionally active DNA”. eng. In: *Cell* 56.4, pp. 619–630. issn: 0092-8674. doi: 10.1016/0092-8674(89)90584-9.

Thorvaldsdóttir, Helga, James T. Robinson and Jill P. Mesirov (Mar. 2013). “Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration”. In: *Briefings in Bioinformatics* 14.2, pp. 178–192. issn: 1467-5463. doi: 10.1093/bib/bbs017. url: https://doi.org/10.1093/bib/bbs017 (visited on 30/08/2023).

Tiriac, Hervé et al. (Sept. 2018). “Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer”. In: *Cancer Discovery* 8.9, pp. 1112–1129. issn: 2159-8274. doi: 10.1158/2159-8290.CD-18-0349. url: https://doi.org/10.1158/2159-8290.CD-18-0349 (visited on 17/04/2023).

Tomaska, Lubomir, Michael J. McEachern and Jozef Nosek (2004). “Alternatives to telomerase: keeping linear chromosomes via telomeric circles”. en. In: *FEBS Letters* 567.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.febslet.2004.04.058, pp. 142–146. issn: 1873-3468. doi: 10.1016/j.febslet.2004.04.058. url: https://onlinelibrary.wiley.com/doi/abs/10.1016/j.febslet.2004.04.058 (visited on 04/04/2023).

Tomczak, Katarzyna, Patrycja Czerwińska and Maciej Wiznerowicz (2015). “Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”. In: *Contemporary Oncology/Współczesna Onkologia* 2015.1, pp. 68–77.

Tseng, Yuen-Yi et al. (Aug. 2014). “PVT1 dependence in cancer with MYC copy-number increase”. en. In: *Nature* 512.7512. Number: 7512 Publisher: Nature Publishing Group, pp. 82–86. issn: 1476-4687. doi: 10.1038/nature13311. url: https://www.nature.com/articles/nature13311 (visited on 24/07/2023).

Turner, Kristen M. et al. (Mar. 2017). “Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity”. en. In: *Nature* 543.7643. Number: 7643 Publisher: Nature Publishing Group, pp. 122–125. issn: 1476-4687. doi: 10.1038/nature21356. url: https://www.nature.com/articles/nature21356 (visited on 07/10/2020).

Tuveson, David and Hans Clevers (June 2019). “Cancer modeling meets human organoid technology”. In: *Science* 364.6444. Publisher: American Association for the Advancement of Science, pp. 952–955. doi: 10.1126/science.aaw6985. url: https://www.science.org/doi/full/10.1126/science.aaw6985 (visited on 27/06/2023).

Untergasser, Andreas et al. (Aug. 2012). “Primer3—new capabilities and interfaces”. In: *Nucleic Acids Research* 40.15, e115. issn: 0305-1048. doi: 10.1093/nar/gks596. url: https://doi.org/10.1093/nar/gks596 (visited on 07/06/2023).

Utani, Koichi et al. (July 2017). “Phosphorylated SIRT1 associates with replication origins to prevent excess replication initiation and preserve genomic stability”. eng. In: *Nucleic Acids Research* 45.13, pp. 7807–7824. issn: 1362-4962. doi: 10.1093/nar/gkx468.

Vanharanta, Sakari and Joan Massagué (Oct. 2013). “Origins of Metastatic Traits”. en. In: *Cancer Cell* 24.4, pp. 410–421. issn: 1535-6108. doi: 10.1016/j.ccr.2013.09.007. url: https://www.sciencedirect.com/science/article/pii/S1535610813004170 (visited on 28/04/2023).

Vicario, Rocio et al. (2015). “Patterns of HER2 Gene Amplification and Response to Anti-HER2 Therapies”. eng. In: *PloS One* 10.6, e0129876. issn: 1932-6203. doi: 10.1371/journal.pone.0129876.

Vishwakarma, Medhavi and Eugenia Piddini (Mar. 2020). “Outcompeting cancer”. en. In: *Nature Reviews Cancer* 20.3. Number: 3 Publisher: Nature Publishing Group, pp. 187–198. issn: 1474-1768. doi: 10.1038/s41568-019-0231-8. url: https://www.nature.com/articles/s41568-019-0231-8 (visited on 28/04/2023).

Vogelstein, Bert et al. (Mar. 2013). “Cancer Genome Landscapes”. In: *Science* 339.6127. Publisher: American Association for the Advancement of Science, pp. 1546–1558. doi: 10.1126/science.1235122. url: https://www.science.org/doi/10.1126/science.1235122 (visited on 03/04/2023).

Vogt, Nicolas et al. (Aug. 2004). “Molecular structure of double-minute chromosomes bearing amplified copies of the epidermal growth factor receptor gene in gliomas”. In: *Proceedings of the National Academy of Sciences* 101.31. Publisher: Proceedings of the National Academy of Sciences, pp. 11368–11373. doi: 10.1073/pnas.0402979101. url: https://www.pnas.org/doi/abs/10.1073/pnas.0402979101 (visited on 25/04/2023).

Von Hoff, D. D. et al. (Sept. 1992). “Elimination of extrachromosomally amplified MYC genes from human tumor cells reduces their tumorigenicity”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 89.17, pp. 8165–8169. issn: 0027-8424. doi: 10.1073/pnas.89.17.8165.

Von Hoff, Daniel D. et al. (Dec. 1991). “Hydroxyurea Accelerates Loss of Extrachromosomally Amplified Genes from Tumor Cells”. In: *Cancer Research* 51.23 Part 1, pp. 6273–6279.

Waddell, Nicola et al. (Feb. 2015). “Whole genomes redefine the mutational landscape of pancreatic cancer”. en. In: *Nature* 518.7540. Number: 7540 Publisher: Nature Publishing Group, pp. 495–501. issn: 1476-4687. doi: 10.1038/nature14169. url: https://www.nature.com/articles/nature14169 (visited on 15/06/2021).

Wanchai, Visanu et al. (Nov. 2022). “CReSIL: accurate identification of extrachromosomal circular DNA from long-read sequences”. eng. In: *Briefings in Bioinformatics* 23.6, bbac422. issn: 1477-4054. doi: 10.1093/bib/bbac422.

Wang, Yuangao et al. (Nov. 2021). “eccDNAs are apoptotic products with high innate immunostimulatory activity”. en. In: *Nature* 599.7884. Number: 7884 Publisher: Nature Publishing Group, pp. 308–314. issn: 1476-4687. doi: 10.1038/s41586-021-04009-w. url: https://www.nature.com/articles/s41586-021-04009-w (visited on 27/06/2022).

Watanabe, Takaaki et al. (Nov. 2017). “Impediment of Replication Forks by Long Non-coding RNA Provokes Chromosomal Rearrangements by Error-Prone Restart”. English. In: *Cell Reports* 21.8. Publisher: Elsevier, pp. 2223–2235. issn: 2211-1247. doi: 10.1016/j.celrep.2017.10.103. url: https://www.cell.com/cell-reports/abstract/S2211-1247(17)31583-8 (visited on 27/01/2022).

Weinberg, Robert A. (1996). “How Cancer Arises”. In: *Scientific American* 275.3. Publisher: Scientific American, a division of Nature America, Inc., pp. 62–70. issn: 0036-8733. url: https://www.jstor.org/stable/24993349 (visited on 25/05/2023).

— (May 2013). *The Biology of Cancer*. en. Google-Books-ID: MzMmAgAAQBAJ. Garland Science. isbn: 978-1-317-96346-2.

Wenger, Sharon L. et al. (Dec. 2004). “Comparison of Established Cell Lines at Different Passages by Karyotype and Comparative Genomic Hybridization”. en. In: *Bioscience Reports* 24.6, pp. 631–639. issn: 1573-4935. doi: 10.1007/s10540-005-2797-5. url: https://doi.org/10.1007/s10540-005-2797-5 (visited on 23/03/2023).

Wick, Ryan R. et al. (June 2017). “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads”. en. In: *PLOS Computational Biology* 13.6. Publisher: Public Library of Science, e1005595. issn: 1553-7358. doi: 10.1371/journal.pcbi.1005595. url: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005595 (visited on 27/06/2022).

Wickham, Hadley (2022a). *forcats: Tools for Working with Categorical Variables (Factors)*. url: https://CRAN.R-project.org/package=forcats.

— (2022b). *stringr: Simple, Consistent Wrappers for Common String Operations*. url: https://CRAN.R-project.org/package=stringr.

— (2022c). *tidyverse: Easily Install and Load the Tidyverse*. url: https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley and Maximilian Girlich (2022). *tidyr: Tidy Messy Data*. url: https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley et al. (2022a). *dplyr: A Grammar of Data Manipulation*. url: https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley et al. (2022b). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. url: https://CRAN.R-project.org/package=ggplot2.

Williams, Hannah L. et al. (Feb. 2023). “Spatially Resolved Single-Cell Assessment of Pancreatic Cancer Expression Subtypes Reveals Co-expressor Phenotypes and Extensive Intratumoral Heterogeneity”. In: *Cancer Research* 83.3, pp. 441–455. issn: 0008-5472. doi:

10.1158/0008-5472.CAN-22-3050. url: https://doi.org/10.1158/0008-5472.CAN-22-3050 (visited on 31/03/2023).

Witkiewicz, Agnieszka K. et al. (Apr. 2015). "Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets". eng. In: *Nature Communications* 6, p. 6744. issn: 2041-1723. doi: 10.1038/ncomms7744.

Wratten, Laura, Andreas Wilm and Jonathan Göke (Oct. 2021). "Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers". en. In: *Nature Methods* 18.10. Number: 10 Publisher: Nature Publishing Group, pp. 1161–1168. issn: 1548-7105. doi: 10.1038/s41592-021-01254-9. url: https://www.nature.com/articles/s41592-021-01254-9 (visited on 09/06/2023).

Wu, Sihan et al. (Nov. 2019). "Circular ecDNA promotes accessible chromatin and high oncogene expression". en. In: *Nature* 575.7784. Number: 7784 Publisher: Nature Publishing Group, pp. 699–703. issn: 1476-4687. doi: 10.1038/s41586-019-1763-5. url: https://www.nature.com/articles/s41586-019-1763-5 (visited on 05/08/2020).

Wu, Sihan et al. (2022a). "Extrachromosomal DNA: An Emerging Hallmark in Human Cancer". In: *Annual Review of Pathology: Mechanisms of Disease* 17.1. _eprint: https://doi.org/10.1146/annurev-pathmechdis-051821-114223, pp. 367–386. doi: 10.1146/annurev-pathmechdis-051821-114223. url: https://doi.org/10.1146/annurev-pathmechdis-051821-114223 (visited on 26/04/2023).

Wu, Song et al. (Jan. 2016). "Substantial contribution of extrinsic risk factors to cancer development". en. In: *Nature* 529.7584. Number: 7584 Publisher: Nature Publishing Group, pp. 43–47. issn: 1476-4687. doi: 10.1038/nature16166. url: https://www.nature.com/articles/nature16166 (visited on 25/05/2023).

Wu, Tao et al. (Mar. 2022b). "Extrachromosomal DNA formation enables tumor immune escape potentially through regulating antigen presentation gene expression". en. In: *Scientific Reports* 12.1. Number: 1 Publisher: Nature Publishing Group, p. 3590. issn: 2045-2322. doi: 10.1038/s41598-022-07530-8. url: https://www.nature.com/articles/s41598-022-07530-8 (visited on 12/04/2023).

Xiao, Nan (2018). *ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for ggplot2*. url: https://CRAN.R-project.org/package=ggsci.

Xie, Na et al. (July 2014). "miR-27a Regulates Inflammatory Response of Macrophages by Targeting IL-10". In: *The Journal of Immunology* 193.1, pp. 327–334. issn: 0022-1767. doi: 10.4049/jimmunol.1400203. url: https://doi.org/10.4049/jimmunol.1400203 (visited on 19/10/2023).

Xu, Huilei et al. (Mar. 2014). "Comparison of somatic mutation calling methods in amplicon and whole exome sequence data". In: *BMC Genomics* 15.1, p. 244. issn: 1471-2164. doi: 10.1186/1471-2164-15-244. url: https://doi.org/10.1186/1471-2164-15-244 (visited on 17/04/2023).

Yan, Feng et al. (Feb. 2020). "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis". In: *Genome Biology* 21.1, p. 22. issn: 1474-760X. doi: 10.1186/s13059-020-1929-3. url: https://doi.org/10.1186/s13059-020-1929-3 (visited on 08/08/2023).

Yang, Lixing et al. (May 2013). “Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes”. English. In: *Cell* 153.4. Publisher: Elsevier, pp. 919–929. issn: 0092-8674, 1097-4172. doi: 10.1016/j.cell.2013.04.010. url: https://www.cell.com/cell/abstract/S0092-8674(13)00451-0 (visited on 26/04/2023).

Yang, Manqiu et al. (May 2023). “Circlehunter: a tool to identify extrachromosomal circular DNA from ATAC-Seq data”. en. In: *Oncogenesis* 12.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–12. issn: 2157-9024. doi: 10.1038/s41389-023-00476-0. url: https://www.nature.com/articles/s41389-023-00476-0 (visited on 04/10/2023).

Yi, Eunhee et al. (Feb. 2022). “Live-Cell Imaging Shows Uneven Segregation of Extrachromosomal DNA Elements and Transcriptionally Active Extrachromosomal DNA Hubs in Cancer”. In: *Cancer Discovery* 12.2, pp. 468–483. issn: 2159-8274. doi: 10.1158/2159-8290.CD-21-1376. url: https://doi.org/10.1158/2159-8290.CD-21-1376 (visited on 27/06/2022).

Yin, Tengfei, Michael Lawrence and Dianne Cook (2021). *ggbio: Visualization tools for genomic data*. url: https://lawremi.github.io/ggbio/.

You, Jia et al. (Jan. 2017). “SEI1 induces genomic instability by inhibiting DNA damage response in ovarian cancer”. en. In: *Cancer Letters* 385, pp. 271–279. issn: 0304-3835. doi: 10.1016/j.canlet.2016.09.032. url: https://www.sciencedirect.com/science/article/pii/S0304383516305742 (visited on 04/05/2023).

Yu, Dihua and Mien-Chie Hung (Dec. 2000). “Overexpression of ErbB2 in cancer and ErbB2-targeting strategies”. en. In: *Oncogene* 19.53. Number: 53 Publisher: Nature Publishing Group, pp. 6115–6121. issn: 1476-5594. doi: 10.1038/sj.onc.1203972. url: https://www.nature.com/articles/1203972 (visited on 08/08/2023).

Yu, Guangchuang (2021). *ChIPseeker: ChIPseeker for ChIP peak Annotation, Comparison, and Visualization*. url: https://guangchuangyu.github.io/software/ChIPseeker.

— (2022). *clusterProfiler: A universal enrichment tool for interpreting omics data*.

Yu, Lisa et al. (Aug. 2013). “Gemcitabine Eliminates Double Minute Chromosomes from Human Ovarian Cancer Cells”. In: *PLoS ONE* 8.8. issn: 1932-6203. doi: 10.1371/journal.pone.0071988. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3750019/ (visited on 15/12/2020).

Yutani, Hiroaki (2022). *gghighlight: Highlight Lines and Points in ggplot2*. url: https://github.com/yutannihilation/gghighlight/.

Zack, Travis I. et al. (Oct. 2013). “Pan-cancer patterns of somatic copy number alteration”. en. In: *Nature Genetics* 45.10. Number: 10 Publisher: Nature Publishing Group, pp. 1134–1140. issn: 1546-1718. doi: 10.1038/ng.2760. url: https://www.nature.com/articles/ng.2760 (visited on 24/04/2023).

Zare, Fatima et al. (May 2017). “An evaluation of copy number variation detection tools for cancer using whole exome sequencing data”. In: *BMC Bioinformatics* 18.1, p. 286. issn: 1471-2105. doi: 10.1186/s12859-017-1705-x. url: https://doi.org/10.1186/s12859-017-1705-x (visited on 17/04/2023).

Zhang, Cheng-Zhong et al. (June 2015). “Chromothripsis from DNA damage in micronuclei”. en. In: *Nature* 522.7555. Number: 7555 Publisher: Nature Publishing Group, pp. 179–184. issn: 1476-4687. doi: 10.1038/nature14493. url: https://www.nature.com/articles/nature14493 (visited on 14/06/2021).

Zhang, Jingcheng et al. (2019a). “MicroRNA-27a (miR-27a) in Solid Tumors: A Review Based on Mechanisms and Clinical Observations”. In: *Frontiers in Oncology* 9. issn: 2234-943X. url: https://www.frontiersin.org/articles/10.3389/fonc.2019.00893 (visited on 19/10/2023).

Zhang, Junjun et al. (Apr. 2019b). “The International Cancer Genome Consortium Data Portal”. en. In: *Nature Biotechnology* 37.4. Number: 4 Publisher: Nature Publishing Group, pp. 367–369. issn: 1546-1696. doi: 10.1038/s41587-019-0055-9. url: https://www.nature.com/articles/s41587-019-0055-9 (visited on 12/04/2023).

Zhang, Xiao-Ou et al. (Sept. 2016). “Diverse alternative back-splicing and alternative splicing landscape of circular RNAs”. eng. In: *Genome Research* 26.9, pp. 1277–1287. issn: 1549-5469. doi: 10.1101/gr.202895.115.

Zhang, Yong et al. (2008). “Model-based analysis of ChIP-Seq (MACS)”. eng. In: *Genome Biology* 9.9, R137. issn: 1474-760X. doi: 10.1186/gb-2008-9-9-r137.

Zhao, Jiawei et al. (May 2019). “Transforming activity of an oncoprotein-encoding circular RNA from human papillomavirus”. en. In: *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group, p. 2300. issn: 2041-1723. doi: 10.1038/s41467-019-10246-5. url: https://www.nature.com/articles/s41467-019-10246-5 (visited on 05/09/2023).

Zhu, Yanfen et al. (May 2021). “Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription”. en. In: *Cancer Cell* 39.5, 694–707.e7. issn: 1535-6108. doi: 10.1016/j.ccell.2021.03.006. url: https://www.sciencedirect.com/science/article/pii/S1535610821001641 (visited on 13/05/2021).