

# INAUGURAL-DISSERTATION

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und  
Naturwissenschaften

der

Ruprecht-Karls-Universität  
Heidelberg

vorgelegt von

M. Sc. Saskia Haupt

aus Stuttgart

Tag der mündlichen Prüfung:



# **MATHEMATICAL MODELING OF LYNCH SYNDROME CARCINOGENESIS**

22. November 2022

**Betreuer:** Prof. Dr. Vincent Heuveline  
**Zweitbetreuer:** PD Dr. Matthias Kloor

## **Colophon**

This document was typeset with the help of [KOMA-Script](#) and [L<sup>A</sup>T<sub>E</sub>X](#) using the [kaobook](#) class.

The back cover was created with [WOMBO Dream](#).

What math and engineering and biology have in common is enormous hidden complexity.

John Doyle, California Institute of Technology



## ABSTRACT

Cancer is one of the leading causes of disease-related death worldwide. In recent years, large amounts of data on cancer genetics and molecular characteristics have become available and accumulated with increasing speed. However, the current understanding of cancer as a disease is still limited by the lack of suitable models that allow interpreting these data in proper ways. Thus, the highly interdisciplinary research field of mathematical oncology has evolved to use mathematics, modeling, and simulations to study cancer with the overall goal to improve clinical patient care.

This dissertation aims at developing mathematical models and tools for different spatial scales of cancer development at the example of colorectal cancer in Lynch syndrome, the most common inherited colorectal cancer predisposition syndrome. We derive model-driven approaches for carcinogenesis at the DNA, cell, and crypt level, as well as data-driven methods for cancer-immune interactions at the DNA level and for the evaluation of diagnostic procedures at the Lynch syndrome population level. The developed models present an important step toward an improved understanding of hereditary cancer as a disease aiming at rapid implementation into clinical management guidelines and into the development of novel, innovative approaches for prevention and treatment.

## ZUSAMMENFASSUNG

Krebs ist weltweit eine der häufigsten krankheitsbedingten Todesursachen. In den letzten Jahren sind große Datenmengen zur Krebsgenetik und zu molekularen Eigenschaften verfügbar geworden und haben sich mit zunehmender Geschwindigkeit angesammelt. Das derzeitige Verständnis von Krebs als Krankheit ist jedoch durch das Fehlen geeigneter Modelle immer noch begrenzt, die eine angemessene Interpretation dieser Daten ermöglichen. Daher hat sich das stark interdisziplinäre Forschungsgebiet der mathematischen Onkologie entwickelt, um Mathematik, Modellierung und Simulationen zur Untersuchung von Krebs mit dem übergeordneten Ziel zu verwenden, die klinische Patientenversorgung zu verbessern.

Ziel dieser Dissertation ist die Entwicklung mathematischer Modelle und Werkzeuge für unterschiedliche räumliche Skalen der Krebsentstehung am Beispiel von Darmkrebs beim Lynch-Syndrom, dem häufigsten erblichen kolorektalen Krebsprädispositionssyndrom. Wir erarbeiten modellgetriebene Ansätze zur Karzinogenese auf DNA-, Zell- und Kryptenebene sowie datengetriebene Methoden für Krebs-Immun-Interaktionen auf DNA-Ebene und zur Evaluation diagnostischer Verfahren auf Populationsebene des Lynch-Syndroms. Die entwickelten Modelle stellen einen wichtigen Schritt in Richtung eines besseren Verständnisses von erblichem Krebs als Krankheit dar, mit dem Ziel einer raschen Umsetzung in klinische Behandlungsleitlinien und in die Entwicklung neuartiger, innovativer Ansätze für Prävention und Behandlung.





# ACKNOWLEDGMENTS

*„Science is a collaborative effort. The combined results of several people working together is often much more effective than could be that of an individual scientist working alone.“* — John Bardeen, from his second Nobel Prize Banquet speech (10 Dec 1972).

With this, I am taking the opportunity to say thank you to the numerous people who accompanied and supported me during the last few years.

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Dr. Vincent Heuveline for guiding me and my research journey, starting with my Master’s thesis and continuing over my Ph.D. at the Faculty of Mathematics and Computer Science at Heidelberg University. With his trust, scientific guidance, and visionary thinking, I have had all possibilities to develop this interdisciplinary Mathematical Oncology research in the framework of my Ph.D. thesis. Vincent, un grand merci à toi!

This interdisciplinary Ph.D. would not have been possible without the highly valuable support from my second supervisor PD Dr. Matthias Kloor from the Department of Applied Tumor Biology (ATB) at Heidelberg University Hospital. Thank you very much for the very warm welcome in the collaboration and the uncountable inspiring scientific discussions we had to bring together mathematics and cancer research. Vielen herzlichen Dank dafür, Matthias.

Special thanks go to all current and former colleagues from the Engineering Mathematics and Computing Lab (EMCL) and the Data Mining and Uncertainty Quantification (DMQ) group for the friendly working environment and the many discussions we had over lunch or coffee breaks. In particular, I want to thank Lydia Mehra for her always pragmatic and efficient solutions for any organizational issues, for often making the seemingly impossible become possible, and for the lectorate of this dissertation. Further, thank you to my office mate Elaine Zaunseder for the nice chats we have and the opportunity to work together on a different medical application in newborn screening. I am also grateful to Nils Gleim for being my first advised student and an amazing collaborator for so many years.

I would like to extend my sincere thanks to all the medical colleagues at ATB, in particular, Johannes Witt and Michael Jendrusch, for their open-mindedness to make this interdisciplinary collaboration a great experience. Many thanks to Dr. Aysel Ahadova for her valuable input for the scientific collaboration and her continuous support in preparing presentations, manuscripts, press releases, blog posts, and much more. Çox sağol. I am also thankful to Prof. Dr. Magnus von Knebel Doeberitz for his future-oriented outlook of bringing together the different fields and initiating this mathematical oncology collaboration with Vincent.

Thanks should also go to all my mathematical and medical collaborations and communities I could join and worked with during the last few years. Especially, I want to thank all the amazing people from the European Hereditary Tumor Group (EHTG), the Mathematical Oncology subgroup of the Society for Mathematical Biology (SMB), the Prospective Lynch Syndrome Database (PLSD), and the Official Mathematical Oncology Website team for their community efforts bringing together research fields and people from all over the world. Thank you, Vielen Dank, Tusen takk, Kiitos, Gracias, Dank je!

Besides that, I would like to thank the Faculty of Mathematics and Computer Science, the Interdisciplinary Center for Scientific Computing (IWR) at Heidelberg University, the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp), as well as the Heidelberg Institute for Theoretical Studies (HITS) for their infrastructural and professional support. I very much appreciate the support of the Klaus Tschira Foundation (KTS) within the projects Informatics4Life and Mathematics in Oncology.

Last but definitely not least, I could not have undertaken this journey without the endless support and love from my family and friends, in particular, my parents, siblings, and my partner in research and life, Alexander Zeilmann. Ein großes Dankeschön für Eure ständige Unterstützung und die vielen kleinen Gesten.

# PREFACE

This Ph.D. work has led to the following articles that were published throughout the Ph.D. period in scientific journals and on preprint servers.

## Peer-reviewed papers

- ▶ H. Bläker, **S. Haupt**, M. Morak, E. Holinski-Feder, A. Arnold, D. Horst, J. Sieber-Frank, F. Seidler, M. von Winterfeld, E. Alwers, J. Chang-Claude, H. Brenner, W. Roth, C. Engel, M. Löffler, G. Möslein, H. Schackert, J. Weitz, C. Perne, S. Aretz, R. Hüneburg, W. Schmiegel, D. Vangala, N. Rahner, V. Steinke-Lange, V. Heuveline, M. von Knebel Doeberitz, A. Ahadova, M. Hoffmeister, M. Kloor, German Consortium for Familial Intestinal Cancer: *Age-dependent performance of BRAF mutation testing in Lynch syndrome diagnostics*. International Journal of Cancer, September 2020, [doi.org/10.1002/ijc.33273](https://doi.org/10.1002/ijc.33273).
- ▶ A. Ballhausen, M. Przybilla, M. Jendrusch, **S. Haupt**, E. Pfaffendorf, F. Seidler, J. Witt, A. Sanchez, K. Urban, M. Draxlbauer, S. Krausert, A. Ahadova, M. Kalteis, P. Pfuderer, D. Heid, D. Stichel, J. Gebert, M. Bonsack, S. Schott, H. Bläker, T. Seppälä, J. Mecklin, S. Broeke, M. Nielsen, V. Heuveline, J. Krzykalla, A. Benner, A. Riemer, M. von Knebel Doeberitz, M. Kloor: *The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution*. Nature Communications, September 2020, [doi.org/10.1038/s41467-020-18514-5](https://doi.org/10.1038/s41467-020-18514-5).
- ▶ **S. Haupt**, A. Zeilmann, A. Ahadova, H. Bläker, M. von Knebel Doeberitz, M. Kloor, V. Heuveline: *Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure*. PLOS Computational Biology, May 2021, [doi.org/10.1371/journal.pcbi.1008970](https://doi.org/10.1371/journal.pcbi.1008970).
- ▶ **S. Haupt**, N. Gleim, A. Ahadova, H. Bläker, M. von Knebel Doeberitz, M. Kloor, V. Heuveline: *A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis*. Computational and Systems Oncology, July 2021, [doi.org/10.1002/cso2.1020](https://doi.org/10.1002/cso2.1020).
- ▶ P. Møller, T. T. Seppälä, J. G Dowty, **S. Haupt**, M. Dominguez-Valentin [and 161 others, including M. von Knebel Doeberitz, A. Ahadova, M. Kloor, V. Heuveline], The European Hereditary Tumour Group (EHTG) and the International Mismatch Repair Consortium (IMRC): *Colorectal cancer incidences in Lynch syndrome: a comparison of results from the Prospective Lynch Syndrome Database and the International Mismatch Repair Consortium*. Hereditary Cancer in Clinical Practice, October 2022, [doi.org/10.1186/s13053-022-00241-1](https://doi.org/10.1186/s13053-022-00241-1).

- ▶ A. Ahadova, J. Witt, **S. Haupt**, R. Gallon, R. Hüneburg, J. Nattermann, S. ten Broeke, L. Bohaumilitzky, A. Hernandez-Sanchez, M. Santibanez-Koref, M. Jackson, M. Ahtiainen, K. Pylvänäinen, K. Andini, V. Grolmusz, G. Möslein, M. Dominguez-Valentin, P. Møller, D. Fürst, R. Sijmons, G. Borthwick, J. Burn, J. Mecklin, V. Heuveline, M. von Knebel Doeberitz, T. Seppälä, M. Kloor: *Is HLA type a possible cancer risk modifier in Lynch syndrome?*. International Journal of Cancer, October 2022, [doi.org/10.1002/ijc.34312](https://doi.org/10.1002/ijc.34312)
- ▶ J. Witt, **S. Haupt**, A. Ahadova, L. Bohaumilitzky, V. Fuchs, A. Ballhausen, M. Przybilla, M. Jendrusch, T. Seppälä, D. Fürst, T. Walle, E. Busch, G. Haag, R. Hüneburg, J. Nattermann, M. von Knebel Doeberitz, V. Heuveline, M. Kloor: *A simple approach for detecting HLA-A \*02 alleles in archival formalin-fixed paraffin-embedded tissue samples and an application example for studying cancer immunoediting*. HLA, October 2022, [doi.org/10.1111/tan.14846](https://doi.org/10.1111/tan.14846).

## Preprints

- ▶ A. Ballhausen, M. Przybilla, M. Jendrusch, **S. Haupt**, E. Pfaffendorf, M. Draxlbauer, F. Seidler, S. Krausert, A. Ahadova, M. Kalteis, D. Heid, J. Gebert, M. Bonsack, S. Schott, H. Bläker, T. Seppälä, J. Mecklin, S. Broeke, M. Nielsen, V. Heuveline, J. Krzykalla, A. Benner, A. Riemer, M. von Knebel Doeberitz, M. Kloor: *The shared neoantigen landscape of MSI cancers reflects immunoediting during tumor evolution*. bioRxiv, July 2019, [doi.org/10.1101/691469](https://doi.org/10.1101/691469).
- ▶ H. Bläker, **S. Haupt**, M. Morak, E. Holinski-Feder, A. Arnold, D. Horst, J. Sieber-Frank, F. Seidler, M. von Winterfeld, E. Alwers, J. Chang-Claude, H. Brenner, W. Roth, C. Engel, M. Löffler, G. Möslein, H. Schackert, J. Weitz, C. Perne, S. Aretz, R. Hüneburg, W. Schmiegel, D. Vangala, N. Rahner, V. Steinke-Lange, V. Heuveline, M. von Knebel Doeberitz, A. Ahadova, M. Hoffmeister, M. Kloor, German Consortium for Familial Intestinal Cancer: *BRAF mutation testing of MSI CRCs in Lynch syndrome diagnostics: performance and efficiency according to patient's age*. medRxiv, October 2019, [doi.org/10.1101/19009274](https://doi.org/10.1101/19009274).
- ▶ **S. Haupt**, A. Zeilmann, A. Ahadova, M. von Knebel Doeberitz, M. Kloor, V. Heuveline: *Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure*. bioRxiv, August 2020, [doi.org/10.1101/2020.08.14.250175](https://doi.org/10.1101/2020.08.14.250175).
- ▶ **S. Haupt**, N. Gleim, A. Ahadova, H. Bläker, M. von Knebel Doeberitz, M. Kloor, V. Heuveline: *Computational model investigates the evolution of colonic crypts during Lynch syndrome carcinogenesis*. bioRxiv, December 2020, [doi.org/10.1101/2020.12.29.424555](https://doi.org/10.1101/2020.12.29.424555).

This dissertation reuses some of these articles verbatim, indicating such use at the beginning of the corresponding sections.

During this Ph.D., also papers and preprints in another context than modeling Lynch syndrome carcinogenesis have been published.

- ▶ S. Gawlok, P. Gerstner, **S. Haupt**, V. Heuveline, J. Kratzke, P. Lösel, K. Mang, M. Schmidtbreick, N. Schoch, N. Schween, J. Schwegler, C. Song, M. Wlotzka: *HiFlow3 – Technical Report on Release 2.0*. Preprint Series of the Engineering Mathematics and Computing Lab, November 2017, [doi.org/10.11588/EMCLPP.2017.06.42879](https://doi.org/10.11588/EMCLPP.2017.06.42879).
- ▶ E. Zaunseder, **S. Haupt**, U. Mütze, S. Garbade, S. Kölker, V. Heuveline: *Opportunities and challenges in machine learning-based newborn screening – A systematic literature review*. JIMD Reports, March 2022, [doi.org/10.1002/jmd2.12285](https://doi.org/10.1002/jmd2.12285).



# CONTENTS

<b>Abstract</b>	<b>v</b>
<b>Zusammenfassung</b>	<b>v</b>
<b>Acknowledgments</b>	<b>vii</b>
<b>Preface</b>	<b>ix</b>
<b>Contents</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 What is cancer and why is mathematics necessary for further developments in oncology? . . . . .	2
1.2 Context and goals of this work in hereditary cancer development . . . . .	3
1.2.1 Outline . . . . .	5
1.3 Contribution of this work with focus on Lynch syndrome . . . . .	7
1.3.1 Main contributions at the DNA level . . . . .	8
1.3.2 Main contributions at the cell and crypt levels . . . . .	9
1.3.3 Main contributions at the population level . . . . .	13
<b>CURRENT STATUS OF LYNCH SYNDROME RESEARCH</b>	<b>15</b>
<b>2 State-of-the-art medical understanding of Lynch syndrome</b>	<b>17</b>
2.1 Basic cell and crypt biology for cancer development . . . . .	17
2.1.1 From the genome over proteins to cellular behavior . . . . .	18
2.1.2 The cell cycle . . . . .	19
2.1.3 Basic biology of colonic crypts . . . . .	21
2.2 Current understanding of cancer development . . . . .	24
2.2.1 Cancer development is a multi-step process . . . . .	24
2.2.2 Cancer develops through different pathways . . . . .	25
2.2.3 Cancer develops sporadically or in a hereditary context . . . . .	26
2.2.4 Different types of colorectal cancer . . . . .	26
2.3 State-of-the-art knowledge on Lynch syndrome colorectal cancer . . . . .	34
2.3.1 Lynch syndrome is the most common inherited colorectal cancer syndrome . . . . .	34
2.3.2 The three pathway hypothesis of Lynch syndrome colorectal carcinogenesis . . . . .	35
2.4 State-of-the-art insights in Lynch syndrome cancer immunology . . . . .	40
2.5 High clinical needs in Lynch syndrome . . . . .	43

<b>3</b>	<b>Challenges and opportunities for modeling in Lynch syndrome</b>	<b>47</b>
3.1	How to model carcinogenesis . . . . .	48
3.1.1	How to analyze the mutational history of tumors . . . . .	49
3.2	Lynch syndrome as a valuable example for modeling . . . . .	50
3.3	State-of-the-art computational modeling at the cell level . . . . .	51
3.4	State-of-the-art mathematical modeling at the crypt level . . . . .	55
 <b>MODELING LYNCH SYNDROME AT THE DNA LEVEL</b>		<b>59</b>
<b>4</b>	<b>Parametrizing mutation rates in a gene-dependent way</b>	<b>61</b>
4.1	Relevant point mutation rates depending on gene hotspot lengths . . . . .	62
4.2	Relevant LOH event rates depending on whole gene lengths . . . . .	64
<b>5</b>	<b>Quantifying HLA type-dependent immuno-editing in cancer</b>	<b>67</b>
5.1	Quantifying the landscape of frameshift mutations using ReFrame . . . . .	69
5.2	Quantifying the landscape of frameshift peptides using immunological scores	73
5.3	HLA type-dependent tumor-immune interactions . . . . .	76
 <b>MODELING LYNCH SYNDROME AT THE CELL AND CRYPT LEVELS</b>		<b>81</b>
<b>6</b>	<b>Computational cell-based model of intra-crypt dynamics</b>	<b>83</b>
6.1	Modeling cell dynamics within a crypt using a Voronoi tessellation . . . . .	85
6.1.1	Modeling the cell cycle . . . . .	86
6.1.2	Modeling cell differentiation . . . . .	88
6.1.3	Modeling cell division and mutations . . . . .	89
6.1.4	Modeling cell migration . . . . .	92
6.1.5	Modeling cell death . . . . .	95
6.1.6	Modeling stem cell dynamics . . . . .	98
6.2	Software and hardware background . . . . .	98
6.3	<i>In silico</i> numerical simulation results . . . . .	100
6.3.1	Epithelial renewal in non-mutated crypts . . . . .	101
6.3.2	The spread of stem cell and TA cell mutations . . . . .	102
6.3.3	The influence of cell location on mutation spread . . . . .	106
6.3.4	The effect of stem cell exchange on monoclonality . . . . .	109
6.4	Outcomes and discussion . . . . .	112
<b>7</b>	<b>Mathematically modeling Lynch syndrome colorectal carcinogenesis using the Kronecker structure</b>	<b>115</b>
7.1	Current medical hypotheses about multiple pathways of Lynch syndrome colorectal carcinogenesis . . . . .	117

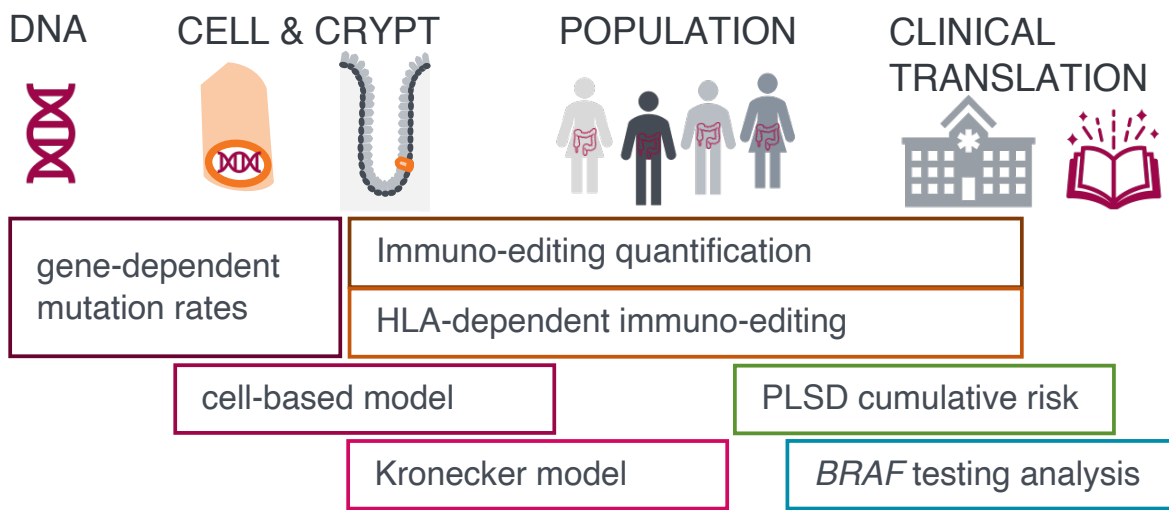


7.2	Modeling Lynch syndrome colorectal carcinogenesis using the Kronecker structure . . . . .	119
7.2.1	Defining gene mutation graphs . . . . .	120
7.2.2	Estimates for point mutation and LOH event rates per crypt per year	122
7.2.3	Fitness advantages and clonal expansion modeled by self-loops in the graph . . . . .	124
7.2.4	Combination of gene mutation graphs using the Kronecker structure	125
7.2.5	Linear dynamical system with Kronecker structure . . . . .	128
7.2.6	The Kronecker structure allows for computational feasibility . . . .	134
7.3	Modifying parameters and initial conditions to model other types of colorectal carcinogenesis . . . . .	138
7.4	Calibration and validation results of the Kronecker model . . . . .	140
7.5	Outcomes and discussion . . . . .	146

**MODELING LYNCH SYNDROME AT THE POPULATION LEVEL 149**

<b>8</b>	<b>PLSD: Modeling cumulative cancer risk in the Lynch syndrome population</b>	<b>151</b>
8.1	Computation of prospective cumulative cancer risk with confidence intervals	153
8.1.1	Definition of cumulative incidence function in survival analysis . .	153
8.1.2	Nelson-Aalen cumulative incidence estimates based on a Poisson distribution . . . . .	154
8.2	Novel computation results with comparison to previous approach . . . .	157
8.3	Outcomes and discussion . . . . .	159
<b>9</b>	<b>Age-dependent performance of <i>BRAF</i> mutation testing: Cost-benefit analysis</b>	<b>163</b>
9.1	Medical evidence for age-dependent Lynch syndrome diagnostics . . . . .	165
9.2	Data collection . . . . .	167
9.2.1	Frequency of <i>BRAF</i> mutations in LS CRCs . . . . .	167
9.2.2	Frequency of LS CRCs among all CRCs . . . . .	168
9.2.3	Frequency of <i>BRAF</i> mutations among MSI CRCs . . . . .	168
9.3	Cost-benefit analysis . . . . .	170
9.3.1	Calculated age-specific frequency of <i>BRAF</i> -mutated LS CRCs among all CRCs . . . . .	170
9.3.2	Calculation of erroneously excluded cases . . . . .	171
9.3.3	Calculated proportion of MSI CRC excluded from MMR gene germline analysis due to <i>BRAF</i> mutation . . . . .	172
9.3.4	Cost calculations for both diagnostic algorithms . . . . .	173
9.4	Outcomes and discussion . . . . .	177

<b>CONCLUSION</b>	<b>179</b>
<b>10 Conclusion</b>	<b>181</b>
10.1 Summary . . . . .	181
10.2 Outlook . . . . .	184
<b>Bibliography</b>	<b>187</b>



# 1 INTRODUCTION

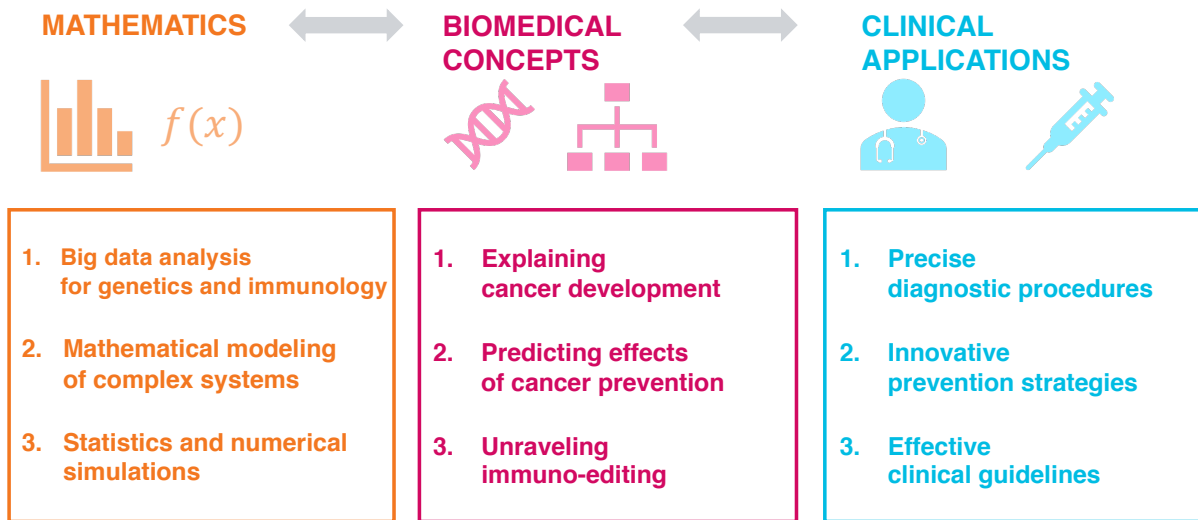
This dissertation is based in the field of mathematical oncology, a highly innovative and interdisciplinary research field that, by the help of mathematics, aims at a detailed understanding of cancer development for improving cancer diagnostics, prevention, and treatment. In this chapter, after an introduction to cancer as a disease, we provide reasoning why mathematics is needed for further developments in cancer research. We emphasize the enormous hidden complexity of cancer as a disease which is one of the main challenges but also opportunities of applying mathematics to oncology. The focus of this dissertation is on large bowel (colorectal) cancer in frame of an inherited cancer predisposition syndrome, namely Lynch syndrome (LS) which serves as a valuable example for modeling cancer development from a mathematical and medical point of view. The goal of this work is to derive first mathematical approaches at different scales ranging from the DNA over the cell and crypt levels to the population level to obtain a comprehensive understanding of Lynch syndrome colorectal cancer (CRC) development with potential to clinical translation (see chapter image above).

We describe the three main components of this goal comprised of modeling cancer development, quantifying cancer immunology, and analyzing population and clinical consequences. After an outline of this work, we conclude the chapter with emphasizing the main contributions of this dissertation for the mathematical oncology and Lynch syndrome research communities.

1.1	WHAT IS CANCER AND WHY IS MATHEMATICS NECESSARY FOR FURTHER DEVELOPMENTS IN ONCOLOGY? . . . .	2
1.2	CONTEXT AND GOALS OF THIS WORK IN HEREDITARY CANCER DEVELOPMENT . . . .	3
1.3	CONTRIBUTION OF THIS WORK WITH FOCUS ON LYNCH SYNDROME . . . . .	7

## **1.1 WHAT IS CANCER AND WHY IS MATHEMATICS NECESSARY FOR FURTHER DEVELOPMENTS IN ONCOLOGY?**

Cancer is one of the leading causes of disease-related death worldwide. In recent years, rapid increase in the molecular understanding of cancer has unraveled significant additional complexity of the disease. Although large amounts of data on cancer genetics and molecular characteristics are available and accumulating with increasing speed, adequate interpretation of these data still represents a major bottleneck. This is exactly where mathematics can be applied to oncology: Through mathematical modeling of complex biological processes, we are able to gain novel, unprecedented medical insights. The fields of application of mathematical models include the analysis of biological concepts and medical hypotheses about cancer evolution, and the prediction of clinical outcomes using existing clinical and molecular information. On the other hand, the medical applications give rise to mathematical challenges, which can lead to new methods and algorithms in various fields of mathematics, like data analysis, statistics, mathematical modeling, and numerical simulations. In subsequent approaches, the established models and methods can be applied to different scenarios. Therefore, applying mathematics in the field of oncology will facilitate data interpretation and improve our understanding of carcinogenic processes. This dissertation is part of the MATHEMATICS IN ONCOLOGY initiative in Heidelberg which intends to provide a sustainable platform at Heidelberg University to support and propel the translation of medical research into innovative cancer therapeutic and preventive approaches (Figure 1.1). A first 3-year project is officially funded by the Klaus Tschira Foundation since November 2021.



**Figure 1.1: Integration of mathematics in biomedicine and clinical management.** *From left to right:* With the help of mathematics, it is possible to analyze and explain complex biological concepts which play a key role in cancer research. This incorporates topics like cancer evolution, estimation of life-time risks, and immuno-editing, which reflects the impact of the immune system during tumor evolution. Close interaction between the disciplines of mathematics and oncology helps to improve and transform existing clinical procedures such as diagnostics, treatment and prevention. *From right to left:* In the other direction, new applications demand new mathematical solutions. Close interaction with concrete medical problems stimulate the development of adequate models, methods, and solution algorithms.

## 1.2 CONTEXT AND GOALS OF THIS WORK IN HEREDITARY CANCER DEVELOPMENT

One of the major challenges on the way to new mathematical modeling approaches is the enormous complexity of cancer as a disease. In particular, early steps of cancer development need to be understood better to guide prevention approaches. An ideal scenario reflecting general principles of cancer initiation and evolution, which is amenable to mathematical modeling, is hereditary forms of cancer. In hereditary cancers, which are responsible for 5 to 10% of the world-wide tumor burden, the causative mechanism increasing the life-time cancer risk of affected individuals is known. Based on the existence of a defined mechanism, it is possible to study the medical phenomena in the mathematical framework.

In this dissertation, mathematical modeling is implemented on the example of Lynch syndrome. It is the most common inherited cancer syndrome and predisposes affected individuals to developing cancer in the large bowel (colorectal cancer) and other organs. Lynch syndrome is ideally suited for studying different facets of cancer development because it

reflects general principles of carcinogenesis beyond the hereditary context in an exemplary manner. Due to its unique and well-defined molecular transformation mechanism, which is associated with the generation of predictable, shared and immunogenic neoantigens, Lynch syndrome also allows the study of tumor immunology in an unprecedented degree of detail. Most importantly, there is a very high medical need for better prevention and management of Lynch syndrome, and direct clinical validation of the developed models is feasible.

Thus, the goal of this dissertation and the associated research work is divided into three parts:

- ▶ **Mathematically modeling the development of Lynch syndrome colorectal cancer to unravel the black boxes of cancer development at small scales.** The view of Lynch syndrome as one homogeneous disease is a clinically detrimental oversimplification. Alterations in four different DNA mismatch repair genes can cause Lynch syndrome, and gene-dependent differences in disease manifestation relating to penetrance, disease severity and prognosis, have been detected. Recently, the existence of three distinct subgroups of Lynch syndrome-associated cancers was discovered, each with a different sequence of key events during their development, which are reflected by a clinically distinct presentation. Although this fundamental heterogeneity of Lynch syndrome cancers clearly calls for tailored treatment and prevention approaches, this is not yet reflected in clinical guidelines.

With the help of mathematical models, the understanding of Lynch syndrome cancer development at the small scales can be significantly improved. Allowing a more precise modeling of cancer development will provide direct implications for prevention strategies.

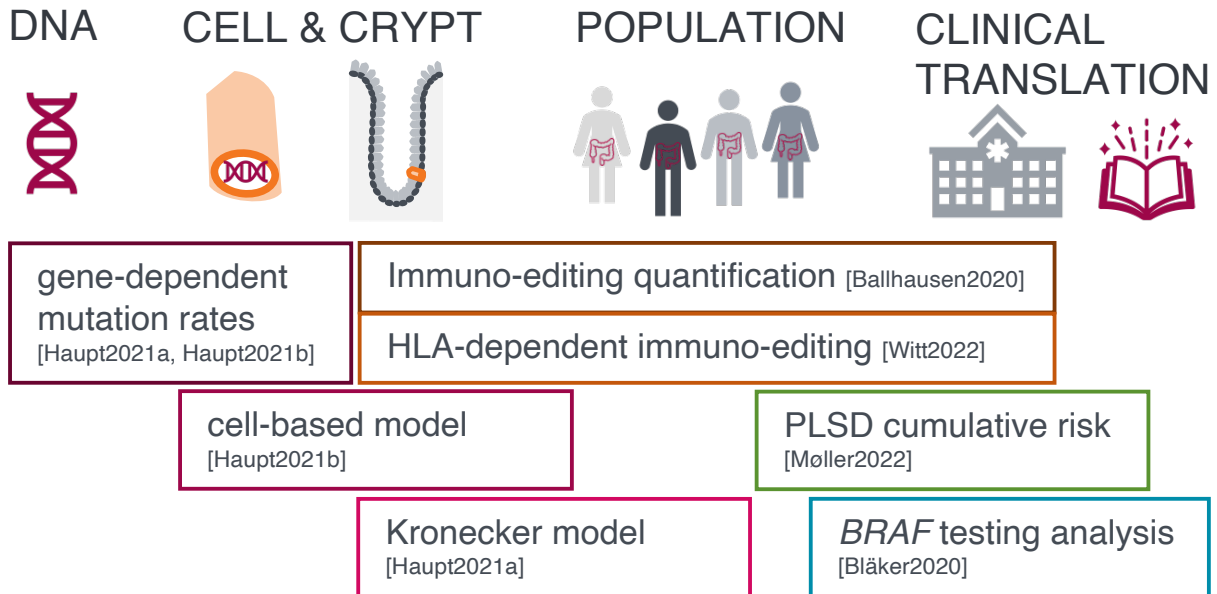
- ▶ **Quantifying cancer immunology and different influencing factors.** In the setting of Lynch syndrome-associated cancers, highly immunogenic cancer neoantigens arise as a direct result of the carcinogenic process, allowing us to precisely study the general principles of cancer immuno-editing. Lynch syndrome-associated cancers display a recurrent set of neoantigens, which occur in identical form in most cancers. This results from the well-defined mechanism of DNA mismatch

repair deficiency that causes exceedingly high numbers of random mutations, affecting specific sequences of the genome. This scenario is ideal for the generation of cancer preventive vaccine approaches, for which the selection of the right vaccination targets is essential. Here, mathematical and bioinformatics approaches will enable the prediction of the immune-relevant neoantigens also considering the individual's HLA type which is responsible for the regulation of the immune system. Thereby, mathematics guides the design of cancer therapeutic and preventive vaccines.

- **Estimating cumulative cancer risk and predicting the efficacy of clinical prevention and *BRAF*-mutation testing in Lynch syndrome diagnostics.** Statistical survival analysis with appropriate assumptions on the underlying data distribution will be used for prospective estimates of cancer risk stratified by age, mismatch repair (MMR) gene variant, and gender of Lynch syndrome individuals. These will be used to analyze the effects of cancer screening approaches, like colonoscopy, for preventing Lynch syndrome-associated cancer. Further, we will focus on the evaluation and comparison of the clinical benefit associated with *BRAF*-mutation testing in Lynch syndrome diagnostic procedures in an age-dependent way by combining and analyzing data from different sources using probabilistic approaches.

### 1.2.1 OUTLINE

This dissertation is structured as follows: We start with a part on the current status of Lynch syndrome research firstly describing the state-of-the-art medical understanding of Lynch syndrome in Chapter 2 followed by explaining the challenges and opportunities these concepts rise for mathematical modeling in Lynch syndrome in Chapter 3. In particular, we summarize state-of-the-art computational and mathematical modeling approaches at the cell and crypt levels of general colorectal cancer which serve as related work to this dissertation. Further, we highlight why Lynch syndrome is a highly valuable example for developing mathematical models with the potential to describe common phenomena in general cancer development despite hereditary colorectal cancer.



**Figure 1.2:** Overview of modeling approaches from the DNA level over the cell and crypt levels to the population level with clinical translations. This figure illustrates the outline of this dissertation in which we present different modeling approaches for different levels of Lynch syndrome cancer research. Each box represents a chapter within this dissertation. Corresponding publications are given in brackets.

The remaining chapters present modeling approaches for Lynch syndrome colorectal cancer developed in the framework of this dissertation ranging from the DNA level over the cell and crypt levels to the population level, illustrated in Figure 1.2.

Starting with the DNA level, we present in Chapter 4 parameterization approaches for modeling DNA alterations, namely point mutations and loss of heterozygosity (LOH) events in a gene-dependent way. These formulas will serve as the basis for the models at the cell and crypt levels. In Chapter 5, we present the mathematical foundations for general bioinformatics and statistics tools to quantify immuno-editing during cancer development. We further refine the presented immunological scores to account for the influence of the human leukocyte antigen (HLA) system, which encodes proteins central for the regulation of the immune system, on the process of immuno-editing and the quantification thereof.

We continue at the cell and crypt levels with developing a computational cell-based model of intra-crypt dynamics in early Lynch syndrome colorectal carcinogenesis in Chapter 6. By numerical simulations, we analyze the spread and monoclonal conversion of different mutations under varying conditions such as the type of mutation, the cell type and location, as well as the influence of stem cell dynamics.



In Chapter 7, we develop a mathematical model for Lynch syndrome colorectal carcinogenesis with whole crypts as the smallest entities. The linear ordinary differential equation model makes use of the Kronecker structure to present the mutational events in multiple pathways of carcinogenesis in a medically interpretable, mathematically analyzable and computationally fast way.

At the population level, we firstly develop and apply a statistical approach for estimating prospective cumulative cancer risk with confidence intervals in the Lynch syndrome population in Chapter 8. The approach is based on Nelson-Aalen cumulative incidence risk estimates with an underlying Poisson distribution which describe the underlying PLSD data, the largest prospective Lynch syndrome database worldwide, in a feasible way improving the currently used approach. Secondly, in Chapter 9, we develop a probabilistic approach to combine data from different existing databases to compare cost and efficacy of two currently used diagnostic procedures to detect Lynch syndrome cancers in an age-dependent way, leading to suggestions for adapting current clinical guidelines.

### **1.3 CONTRIBUTION OF THIS WORK WITH FOCUS ON LYNCH SYNDROME**

Mathematical oncology is a highly innovative research field which develops constantly due to the ongoing data and knowledge generation at different scales. Combining heterogeneous types of data and information and formalize medical knowledge in a mathematically rigorous way allows for a more comprehensive understanding of cancer development and thus supports clinical procedures in cancer prevention, diagnosis, and treatment. As pointed out above, Lynch syndrome colorectal cancer serves as a valuable example for studying various aspects of general cancer development including mutational processes, tumor-immune interactions, and the connection of different scales.

In the framework of this Ph.D. dissertation and the associated research work, we provide, to the best of our knowledge, unique modeling approaches towards a comprehensive investigation of Lynch syndrome colorectal cancer development including various important medical aspects and data at different scales ranging from the DNA over the cell and crypt to the population level. We describe in detail the main contributions at each level in the following sections.

### 1.3.1 MAIN CONTRIBUTIONS AT THE DNA LEVEL

For the DNA level, we formalize a mathematical framework for modeling gene-dependent mutation rates which serves as a basis for the models developed at the cell and crypt levels. Further, we are mathematically involved in the development and analysis of bioinformatics and data analysis approaches for quantifying immuno-editing during cancer development and the influence of the HLA type thereon.

For the **parametrization of gene-dependent mutation rates**, we derive formulas for different types of relevant gene alterations, namely point mutations and LOH events affecting whole parts of chromosomes. For each gene, the latter are assumed to be dependent on the overall gene lengths, whereas the former depends on the length of hot spots, regions that give rise to phenotypical changes. The parameters in these equations have a biomedical meaning and can be set using state-of-the-art databases. For the subsequent models at the cell and crypt levels, general parameters of the equations like gene lengths are taken from the current reference sequence database at NCBI. In addition, for specific information on hot spot lengths, mutation data from other public resources or recent medical research publications are used to estimate the remaining parameters.

For the **quantification of immuno-editing during cancer development**, the development of a bioinformatics-based algorithm, called ReFrame, was initiated at ATB Heidelberg to quantitatively detect microsatellite indel mutations with high sensitivity. We are involved in the application of this algorithm to microsatellite unstable colorectal cancers with subsequent data analysis and interpretation under different medical conditions. Therefore, immunological scores are defined which predict the frameshift peptides' possibility to

provoke an immune response, called immunogenicity. Most importantly, we discover a negative correlation between the frameshift mutations in MMR-deficient colorectal cancers and the predicted immunogenicity of the resulting frameshift peptides.

In the future, detailed analyses about the influence of the HLA system on immuno-editing processes should be addressed. However, the laboratory procedure to determine the HLA genotype of patients is currently challenging. Thus, a second laboratory study developing an easy-to-use laboratory method for these purposes is performed at ATB Heidelberg, with our mathematical support for a rigorous data analysis. In this context, the previously defined immunological scores are generalized to deal with multiple HLA types describing the landscape of HLA genotypes within individuals and the population.

The developed algorithms and frameworks serve as the basis for the international INDICATE initiative performing future comprehensive analyses on the influence of the HLA type on immuno-editing and related tumor-immune interactions.

### 1.3.2 MAIN CONTRIBUTIONS AT THE CELL AND CRYPT LEVELS

We develop a computational and a mathematical model for the cell and crypt levels to obtain first *in silico* realizations of the time evolution of colorectal cancer development in Lynch syndrome. Both models use the modeling approaches for gene-dependent mutation and LOH events described in the previous chapter to have a common foundation at the DNA level. Further, both models make use of recent experimental data and biomedical knowledge incorporating representative genes of known drivers in Lynch syndrome colorectal cancer. This includes recent experimental data [12] demonstrating that somatic *CTNNB1* mutations, if affecting both alleles of the gene, are common drivers of Lynch syndrome-associated colorectal cancers. The following narrative is adapted from [81, 82].

The **first model at the cell level** is a computational model with numerical simulations of colonic crypt evolution during Lynch syndrome carcinogenesis [81]. It is developed to translate knowledge about the effects of defined mutations from

[12]: Arnold et al. (2020), “The majority of  $\beta$ -catenin mutations in colorectal cancer is homozygous”.

[81]: Haupt et al. (2021), “A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis”.

[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

the cellular to the crypt level. Although experimental data on mutation rates in dividing cells *in vitro* are existing, it is hard to translate these numbers onto the level of crypts, the organ, or the individual. In these lines, information about (1) the likelihood of a defined mutation leading to monoclonal conversion of the surrounding crypt, and (2) the time until conversion takes place are paramount. The developed model is designed as a first step to fill this gap of knowledge.

Although organoid cultures represent a huge leap forward in the analysis of cellular alterations in a tissue context, computational models are more flexible with regard to altering certain parameters, including the implementation of environmental changes and their effects on crypt homeostasis or mutational manifestation. Moreover, computational models have the significant advantage that upscaling the number of simulations is only limited by the availability of computational resources.

The present model is an extension and adaptation of existing approaches [66, 117, 132] for modeling Lynch syndrome carcinogenesis allowing to obtain *in silico* experiments for mutational processes and intra-crypt dynamics during Lynch syndrome carcinogenesis. We take into account crucial biological features by defining stem cell dynamics with only one active stem cell at a time according to [121, 173]. In addition, the modeling incorporates feedback and resulting death mechanisms which turn out to be essential to avoid an overpopulation of the crypt.

Beside somatic *APC* mutations as known drivers of colorectal cancer in general, we incorporate MMR deficiency and related genetic dependencies such as increased mutation and death rates into the model. Further, as stated above, we include somatic biallelic *CTNNB1* mutations as common drivers of Lynch syndrome-associated colorectal cancers.

Our model is able to simulate the effect of different mutations, including non-transforming and transforming mutations, on the intra-crypt dynamics. Mutations without any survival advantage, such as the MMR deficiency-inducing second hit are of particular importance in Lynch syndrome carcinogenesis. Modeling the effects of both types of mutations is thus essential for the biological understanding. We investigate how these driver mutations influence the intra-crypt dynamics and analyze the influence of the cell location and the

[66]: Fletcher et al. (2012), “Mathematical modeling of monoclonal conversion in the colonic crypt”.

[117]: Leeuwen et al. (2009), “An integrative computational model for intestinal tissue renewal”.

[132]: Meineke et al. (2001), “Cell migration and organization in the intestinal crypt using a lattice-free model”.

[121]: Li and Clevers (2010), “Coexistence of Quiescent and Active Adult Stem Cells in Mammals”.

[173]: Sato et al. (2009), “Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche”.

effect of stem cell dynamics on the spread and monoclonal conversion of mutations within a crypt. In general, the model allows simulating effects of other, possibly yet unidentified mutations of both non-transforming and transforming type by straightforwardly extending the model.

The model parameters are based on the existing biomedical and clinical estimates allowing a comparison with available human data and new medical hypotheses for human colonic crypt evolution. Besides that, it is possible to use the modeling approach for murine colonic crypts by adapting size- and species-dependent parameters, which can support studies analyzing carcinogenic processes using animal models. Further, upon adaptation of certain parameters, this model can be rolled-out to reflect the development of sporadic colorectal cancers and colorectal cancers in the Lynch-like [43] and familial adenomatous polyposis (FAP) scenario.

[43]: Carethers (2014), “Differentiating Lynch-Like From Lynch Syndrome”.

With our numerical simulations, we obtain *in silico* estimates for crypt renewal in healthy tissue, as well as time span predictions for monoclonal conversion of non-transforming and transforming mutations under varying cellular conditions which could be experimentally validated in humans in the future.

One important finding is that we observe loss and recovery of monoclonality due to stem cell exchange mechanisms in some simulations which might explain why not all MMR-deficient crypts, which are highly likely to occur in Lynch syndrome individuals, might evolve into a carcinoma.

In summary, we provide a first computational model for intra-crypt dynamics of early Lynch syndrome colorectal cancer development including key driver mutations, stem cell mechanisms, as well as feedback mechanisms with possibilities for cell death to infer crypt homeostasis. The simulations allow for initial insights into usually unobservable processes and time span quantification of key components of early Lynch syndrome colorectal cancer development.

As a **second model at the crypt level**, we provide a general mathematical framework that describes arbitrarily complex and arbitrary numbers of pathways and mutations because the chosen Kronecker structure enables a modular construction and an analytic, computationally efficient solution. We use Lynch syndrome carcinogenesis to illustrate the flexibility of the model. Naturally, specific assumptions may vary

[111]: Komarova et al. (2002), “Dynamics of Genetic Instability in Sporadic and Familial Colorectal Cancer”.

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

[12]: Arnold et al. (2020), “The majority of  $\beta$ -catenin mutations in colorectal cancer is homozygous”.

[89]: Huels et al. (2015), “E-cadherin can limit the transforming properties of activating  $\beta$ -catenin mutations”.

for other types of cancer. We illustrate model modifications for FAP, Lynch-like and the classical microsatellite-stable colorectal carcinogenesis.

Instead of focusing on modeling *APC* inactivation and MMR deficiency as in [111], we choose a more general approach for combining mutations in different genes. Compared to [160], we take into account different modes of cancer development beside the classical adenoma-carcinoma sequence of microsatellite-stable colorectal carcinogenesis, including hereditary forms like Lynch syndrome and FAP. Further, as described above, recent data show that in Lynch syndrome-associated colorectal cancers, biallelic mutations of *CTNNB1* seem to be required to mediate an oncogenic driver effect [12, 89], which we include in the definition of the gene mutation graphs.

While the approach in [160] is a hybrid approach of linear ordinary differential equations (ODEs) and a stochastic branching process, we use a system of ODEs to model the evolution of all genotypic states which eases the computational solution process tremendously. This goes in hand with the fact that all formulas in our model are exact from a mathematical point of view without using any approximations which in turn allows for an analytical solution of the ODEs by using the matrix exponential.

Further, the model consists of different components for modeling independent and dependent mutational processes taking into account currently available clinical observations and biomedical data.

Finally, our approach makes it possible to easily include new medical insights, while preserving the other properties of the model, like the integration of the involved differential equations. This incorporates the possibility for multiple cancerous genotypic states reflecting the real world heterogeneity of cancer, the consideration of multiple driver genes, as well as the use of different initial values and parameter combinations for modeling other carcinogenesis processes.

### 1.3.3 MAIN CONTRIBUTIONS AT THE POPULATION LEVEL

At the population level, we provide the mathematical foundation for two large database analyses: one being PLSD, the epidemiological prospective description of cancer risk in Lynch syndrome individuals and the other being a systematic review and comparison regarding costs and efficacy of two current diagnostic procedures including and excluding *BRAF* mutation testing for detecting Lynch syndrome individuals.

For the **prospective Lynch syndrome database (PLSD) analyses**, we develop a mathematical analysis approach for estimating the cumulative cancer risk up to a certain age with 95% confidence intervals in Lynch syndrome individuals based on a Nelson-Aalen estimate with an underlying Poisson distribution. This novel approach replaces the one used so far which was based on the simplifying assumption of a normal distribution. However, as in PLSD, cancer occurrence is considered a dichotomous variable, and the number of cancer cases is counted in a specific age interval, the Poisson distribution should be the natural choice for all calculations. In a first study, the novel calculation method is used to compare estimates of colorectal cancer incidences obtained with PLSD and with another database from the International Mismatch Repair Consortium (IMRC). Whereas patients included in PLSD undergo regular colonoscopy, this is not the case for most of the individuals in the IMRC database allowing a comparison of cancer risks in Lynch syndrome individuals with and without regular screening to quantify the effect of the latter on colorectal cancer risk within the Lynch syndrome population. In the future, the novel method will be the standard for the next versions of the PLSD results which are regularly updated including recently collected data.

For the **cost-benefit analysis of *BRAF* mutation testing**, we combine data from a systematic literature review to compute costs of two diagnostic approaches and conditional probabilities for erroneously excluding Lynch syndrome individuals from germline variant analysis due to the presence of a *BRAF* mutation which carries a risk of missing this hereditary predisposition. We evaluate the performance of *BRAF* mutation testing in Lynch syndrome diagnostics in an age-specific way.

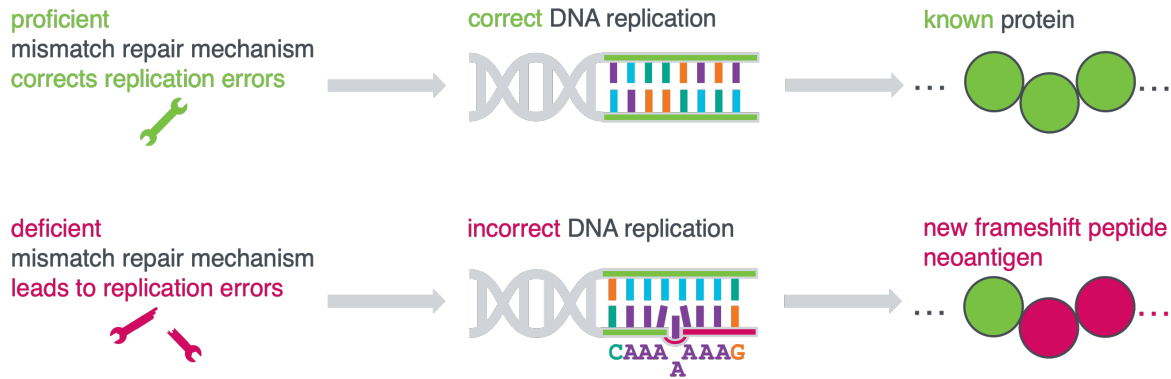
The performed calculations yield the result that the use of *BRAF* mutation testing in patients younger than 50 years of age carries a high risk of missing Lynch syndrome and is cost-inefficient. This leads to the recommendation of adapting current diagnostic guidelines by also considering the age of the patient for the decision on the diagnostic procedure to be chosen.

Thus, current risk assessment of different cancer types in Lynch syndrome as well as cost-efficacy investigations largely benefit from our mathematical analyses of population data in Lynch syndrome enabling a mathematically rigorous foundation of the performed computations. Further, by these findings, we support the further development of future clinical guidelines and provide the basis for future population-based cancer risk assessments extending the current Lynch syndrome databases.



**CURRENT STATUS OF LYNCH SYNDROME  
RESEARCH**





## 2 STATE-OF-THE-ART MEDICAL UNDERSTANDING OF LYNCH SYNDROME

In this chapter, in Section 2.1, we give an introduction to basic cell and crypt biology, based on our description in [82]. This is necessary for the understanding of cancer development in general and at the example of colorectal cancer which will be the subject of Section 2.2. The remaining sections focus on different aspects of Lynch syndrome cancer, starting with the state-of-the-art understanding of cancer development in Lynch syndrome in Section 2.3. Next, we give current insights in Lynch syndrome cancer immunology and the important concept of immuno-editing, see Section 2.4 and conclude with emphasizing the high clinical need for adequate Lynch syndrome diagnostics, prevention and treatment in Section 2.5.

### 2.1 BASIC CELL AND CRYPT BIOLOGY FOR CANCER DEVELOPMENT

Cancer is a disease caused by alterations of the genome, the carrier of the genetic information [62, 215]. This understanding dates back to the early years of the 20th century [19], even before the structure of the specific molecule which carries the genomic information, deoxyribonucleic acid (DNA) [204], was identified.

In this section, we explain how these alterations on the genome level affect proteins and by this the cellular behavior. We further give medical insights into healthy and aberrant cell

2.1	CELL AND CRYPT BIOLOGY . . . . .	17
2.2	CANCER DEVELOPMENT . . . . .	24
2.3	LYNCH SYNDROME COLORECTAL CANCER	34
2.4	LYNCH SYNDROME IMMUNOLOGY . . . . .	40
2.5	CLINICAL NEEDS . . . . .	43

[82]: Haupt et al. (2021), "Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure".

[62]: Edler and Kopp-Schneider (2005), "Origins of the mutational origin of cancer".

[215]: Wunderlich (2006), "Early references to the mutational origin of cancer".

[19]: Bauer (1957), "Mutationstheorie der Geschwulstentstehung. Berlin 1928".

[204]: Watson and Crick (1953), "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid".

and crypt behavior which are essential for the development of mathematical models later in this dissertation.

### 2.1.1 FROM THE GENOME OVER PROTEINS TO CELLULAR BEHAVIOR

The genetic material contains all information and programs relevant for life and development of a cell or an organism. It is present in every cell of a multi-cellular organism and has to be duplicated during each cell division. This process, called DNA replication (see Section 2.1.2), takes place frequently as it is required to maintain homeostasis (equilibrium) of the organism. For example, cells of the skin or the intestines proliferate at a high rate to replace cells undergoing programmed cell death (apoptosis). In the human body, DNA replication happens nearly two trillion ( $2 \cdot 10^{12}$ ) times each day [24].

[24]: Bianconi et al. (2013), "An estimation of the number of cells in the human body".

On the one hand, the constant renewal of cells allows a long lifespan and protects from the accumulation of environmental damage and other detrimental influences over time. The double helix structure of the underlying DNA molecule is designed ideally for duplication, since all information is available on both strands that can serve as a template for building the new cells' genomes. On the other hand, the replication of the DNA structure is chemically highly complex and thus error-prone.

Usually, the information contained in the DNA is, after transcription to an intermediate messenger molecule (messenger RNA or mRNA), translated into proteins using an organism-specific genetic code. To put it in numbers, the genome consists of 3.2 billion ( $3.2 \cdot 10^9$ ) base pairs, whereby only 1% represents genes that are in fact translated into proteins [188]. The number of genes in the human genome is estimated to be 23,000. The resulting proteins are made up from 20 different building blocks (amino acids) and are therefore highly diverse (Figure 2.1, top). Proteins are responsible for a vast majority of functions within organisms, including proliferation and survival of cells. For normal cell behavior, it is thus crucial that possible errors occurring during DNA replication are repaired because they can have severe consequences on the produced proteins, their structure and their function (Figure 2.1). Therefore, all organisms from bacteria

[188]: The ENCODE Project Consortium (2012), "An Integrated Encyclopedia of DNA Elements in the Human Genome".

to complex multi-cellular organisms have developed a wide range of error detection, repair and control systems. Central to this essential network, which enables healthy life, are DNA repair enzymes. Even when repair enzymes work properly, some alterations (or mutations, see Section 2.1.2) will always be detectable in newly generated cells. If DNA repair itself is impaired, which is the case in Lynch syndrome individuals, the number of these mutations will raise dramatically. This can thus increase the risk of developing cancer and we will explain the underlying mechanisms in more detail in Section 2.2.4.

### 2.1.2 THE CELL CYCLE

We have seen that in order to sustain the integrity of tissues, it is crucial that cells are able to divide and grow, which we collectively term cell proliferation. We will now have a closer look at the individual steps of this process, including the possibility of different types of mutations.

Between two divisions, each cell undergoes a series of events known as the cell cycle, which includes the preparation for DNA replication (G1 phase), DNA replication (S phase), the preparation for cell division (G2 phase), and cell division (M phase) [51]. The success of the cell cycle depends on complex signaling cascades consisting of proteins, enzymes, and hormones, each fulfilling different tasks. The inhibition of single elements of these cascades can lead to the arrest of the cell cycle, inhibiting proliferation. In this case, the cell is said to be quiescent, or in the G0 phase.

Cell differentiation describes the transition from one cell type to another, involving changes in size, shape, and responsiveness to biochemical signals [6]. In most cases, this leads to the cell becoming more specialized, that is, it fulfills more specific tasks within its respective tissue. A division of a cell which results in two new cells of the same type is called symmetric, while a division resulting in the creation of one cell of the same type and one cell of a more differentiated cell type is called asymmetric. Biochemical signaling cascades heavily influence the process of cell differentiation. One of the primary regulators might be the so-called Wnt pathway [128] which plays an important role in colorectal cancer development (see Section 2.2.4).

[51]: Cooper (2018), *The Cell: A Molecular Approach*. 8th edition.

[6]: Alberts et al. (2007), *Molecular Biology of the Cell*.

[128]: Matteis et al. (2012), “A review of spatial computational models for multi-cellular systems, with regard to intestinal crypts and colorectal cancer development”.

[118]: Leeuwen (2007), "Towards a multiscale model of colorectal cancer".

[172]: Ruddon (2007), *Cancer biology*.

[83]: Heath (1996), "Epithelial cell migration in the intestine".

The Wnt pathway describes a complex signaling cascade involved in several distinct processes across all animal species, such as embryonic development, tissue regeneration, and carcinogenesis. The ongoing activity of the Wnt pathway can prolong differentiation [118]. This is due to the role of the two proteins APC and  $\beta$ -catenin: Broadly speaking, APC is part of a complex of proteins which degrades  $\beta$ -catenin and thereby prevents it from traveling to the cell nucleus, where it can induce cell division. The activation of the Wnt pathway involves blocking APC from degrading  $\beta$ -catenin and leads to ongoing cell division. Due to this connection, mutations in the *APC* gene and in *CTNNB1*, the gene encoding for  $\beta$ -catenin, have been linked to various types of cancer [172].

Beside biochemical signaling cascades, the function of a cell is heavily influenced by the interaction with other cells and with the extracellular matrix, which are all macromolecules in the intercellular space. These two types of interactions are collectively termed cell adhesion. The attachment of cells to the extracellular matrix is necessary for the directed movement of cells, termed cell migration. In the colonic crypt, the adherence of cells to the extracellular matrix causes an upward migration. Further, the division of adjacent cells creates a so-called mitotic pressure, which also contributes to cell migration [83]. The latter is essential for the maintenance of structures, and the formation and regeneration of tissues within organisms.

Upon DNA replication, errors can occur and, if not corrected, manifest as *mutations*. These mutations occur over the whole genome, whereby we differentiate between two broad classes: So-called point mutations only affect a single nucleotide, while loss of heterozygosity (LOH) events refer to the loss of some region in one copy of the diploid genome, which can result in the deletion of whole genes.

If mutations strike in regions with a protein-encoding function, two main scenarios that can favor uncontrolled cell growth are seen: Somatic mutations can either directly activate oncogenes (typically referred to in the literature as gain-of-function mutations), which physiologically promote appropriate cell growth and proliferation, through conformational changes or impairing self-inactivation, or mutations can damage or destroy tumor suppressor genes (typically referred to in the literature as loss-of-function mutations), which physiologically limit cell growth and proliferation. In

colorectal carcinogenesis, classical examples are *KRAS* as oncogene, and *APC* and *TP53* as tumor suppressor genes (see Section 2.2.4).

Further, a *germline variant* is any detectable mutation within germ cells that can be passed down from parent to child and that is present in all cells from birth on. In contrast, a *somatic mutation* happens in a somatic cell (all cells other than germ cells) and is usually not transmitted to descendants. However, it is present in all descendants of this cell within the same organism.

### 2.1.3 BASIC BIOLOGY OF COLONIC CRYPTS

As we are focusing on colorectal cancer, we give a short introduction to the biology of the colon necessary to understand colorectal cancer development. Colonic crypts are tubular invaginations within the colonic epithelium which are believed to be the origin site of colorectal cancer [118]. According to current estimations, human colonic crypts are about 75 to 110 cells long and have an average circumference of 23 cells [15]. Naive multiplication therefore suggests the total cell number per crypt to range between 1.7 and 2.5 thousand. The cells of the colonic crypt can be divided into three groups: stem cells, transit-amplifying (TA) cells, and fully differentiated (FD) cells. They are characterized by their functions and abilities with respect to the cell proliferation, division, differentiation, and migration (for a general description, see Section 2.1.2).

Stem cells reside at the crypt bottom [15]. These cells are undifferentiated and have unlimited proliferative potential, such that they can renew themselves and give rise to more differentiated progeny. According to [121, 173], we assume that there is one active stem cell at a time populating the crypt, whereby the others are quiescent. In general, the stem cell cycle is much longer than the one of TA cells. Usually, stem cell division is *asymmetric*, leading to one stem cell and one transit-amplifying cell. During this process, mutations can happen which can also lead to mutation-induced death of the stem cell. In this case, one of the adjacent stem cells divides symmetrically leading to two new stem cells and a fixed number of stem cells over time. The relative frequency of either mode of division is a current research topic [211].

[118]: Leeuwen (2007), “Towards a multiscale model of colorectal cancer”.

[15]: Baker et al. (2014), “Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon”.

[121]: Li and Clevers (2010), “Coexistence of Quiescent and Active Adult Stem Cells in Mammals”.

[173]: Sato et al. (2009), “Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche”.

[211]: Wodarz and Komarova (2014), *Dynamics of Cancer*.

[95]: Johnston (2008), “Mathematical modelling of cell population dynamics in the colonic crypt with application to colorectal cancer”.

[165]: Potten and Loeffler (1990), “Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt”.

[42]: Buske et al. (2011), “A Comprehensive Model of the Spatio-Temporal Stem Cell and Tissue Organisation in the Intestinal Crypt”.

[51]: Cooper (2018), *The Cell: A Molecular Approach*. 8th edition.

[129]: McDonald et al. (2006), “Clonal Expansion in the Human Gut: Mitochondrial DNA Mutations Show Us the Way”.

[18]: Barker et al. (2007), “Identification of stem cells in small intestine and colon by marker gene *Lgr5*”.

[180]: Shih et al. (2001), “Top-down morphogenesis of colorectal tumors”.

[110]: Komarova and Wodarz (2004), “The optimal rate of chromosome loss for the inactivation of tumor suppressor genes in cancer”.

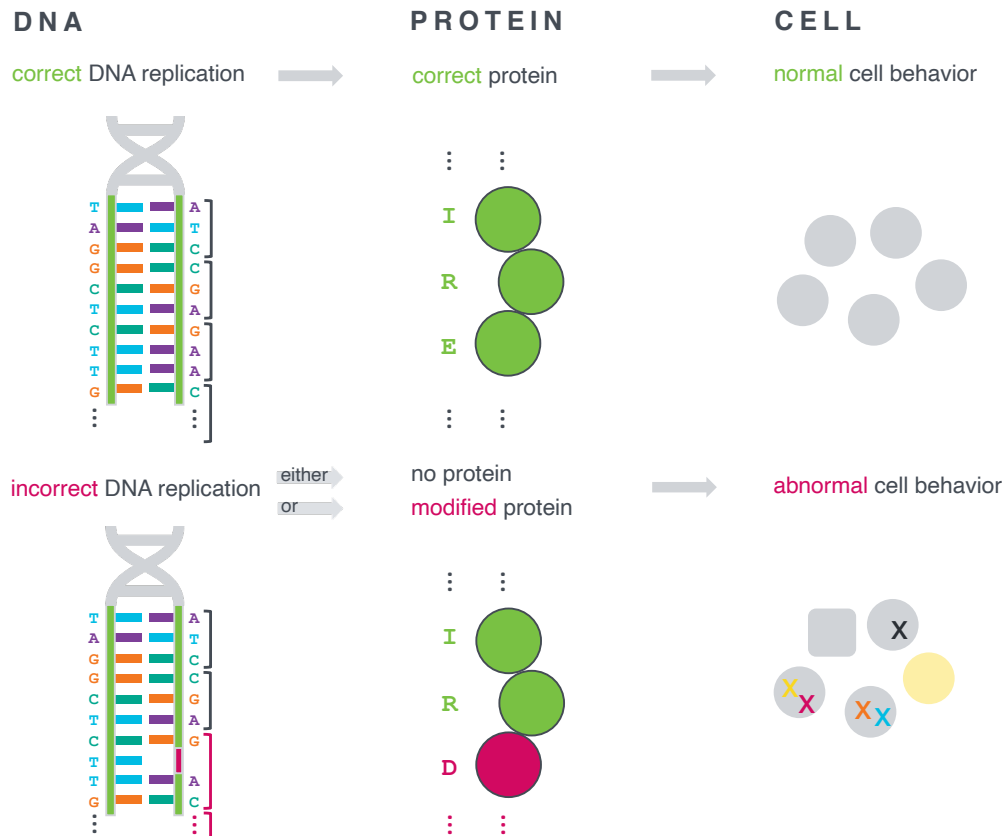
TA cells are located above the stem cells in the lower and lower middle part of the colonic crypt. These cells are thought to possess certain properties of both stem cells and fully differentiated cells, depending on how far along they are on the path to differentiation [95]. TA cells lack the ability to endlessly regenerate as they only divide a certain number of times before becoming fully differentiated [165]. We ignore the possibility of de-differentiation of TA cells in the current modeling approach. However, TA cells divide more frequently than stem cells. The mode of division is, among others, determined by the activity of the Wnt pathway within these cells. According to [42], constitutive activation of Wnt in cells leads to an expansion of the populations of undifferentiated cells, and reduced Wnt signaling results in a complete loss of undifferentiated cells. Further, mutations can happen during each division leading to potential mutation-induced death. Besides that, we assume that TA cells can die due to high mitotic pressure.

FD cells are located above the TA cells and never divide [51], thus no mutations are generated. Due to cell adhesion and mitotic pressure, they migrate to the top of the crypt, where they undergo apoptosis and are shed into the colonic lumen.

Whenever a mutation occurs in a cell of a crypt, this cell has the potential to pass on the mutation to its progeny, possibly resulting in almost all cells in the crypt showing the mutation after a certain amount of time. This process is called *monoclonal conversion* and has been observed experimentally for mutations in mitochondrial DNA [129] and for the progeny of stem cells [18]. The latter gives rise to the hypothesis that colorectal cancer originates in stem cells, as in these cells the mutations are least susceptible to be washed out of the crypt. The corresponding type of monoclonal conversion starting at the bottom of the crypt is called *bottom-up morphogenesis*.

In contrast, the existence of colonic crypts showing morphologically normal bottom regions paired with APC inactivation at the top and dysplastic epithelium lining the luminal surface [180] suggests that mutants at the top of the crypt expand downwards. Additionally, a mathematical model [110] predicts that at least the second *APC* hit occurs in the migrating population, as opposed to both hits occurring in stem cells. This type is summarized as *top-down morphogenesis*.





**Figure 2.1:** A change in the genomic information caused by incorrect DNA replication can change the protein production and thus the cell behavior. *Top:* In cell division, the DNA has to be replicated, meaning that each DNA building block (called base or nucleotide) is matched with its corresponding one (illustrated by **A** to **T** and **C** to **G**). Certain parts of the DNA, the genes, are then translated into proteins, each carrying out the genomic information contained in these parts. Thereby, the genetic code specifies how the sequences of nucleotide triplets are translated into amino acids (illustrated by a chain of green circles), which in turn form the building blocks for proteins. If no error occurs during DNA replication, the correct protein is built and thus the cells (gray circles) will behave in a normal way. *Bottom:* Different errors can occur during DNA replication: Either the wrong base is incorporated into the newly synthesized DNA strand, an additional base is inserted, or an existing base is deleted (missing **A** marked with red) during DNA replication. As a result, the protein production is changed leading either to a modified protein, the wrong amount of proteins or no protein at all. As illustrated in this example, an insertion or deletion leads to a complete shift of the subsequent reading of the bases and thus the subsequent amino acids are wrong, too (illustrated by a red circle and **D** instead of **E**). Any of these changes can affect the cell behavior and potentially induce disease.

## 2.2 CURRENT UNDERSTANDING OF CANCER DEVELOPMENT

[154]: Nowell and Hungerford (1960), "Chromosome studies on normal and leukemic human leukocytes".

In the early stages of cancer research, it was unknown whether or not the development of cancer was a purely chaotic process of random mutations. However, in 1960, Nowell and Hungerford [154] made the observation of a specific recurrent alteration across different cancers of the same type. This suggested the existence of at least a certain degree of order in the assumed chaos.

In this section, we give a short historic overview on main achievements in understanding cancer development and explain key concepts such as multi-step and multi-pathway cancer development that shape medical cancer research nowadays. We explain different types of cancer in general and at the example of colorectal cancer, the main focus of our work.

### 2.2.1 CANCER DEVELOPMENT IS A MULTI-STEP PROCESS

[201]: Vogelstein and Kinzler (1993), "The multistep nature of cancer".

[202]: Vogelstein et al. (1988), "Genetic Alterations during Colorectal-Tumor Development".

In the decades following the observation of a specific recurrent alteration in one cancer type, evidence emerged that one single mutation is normally insufficient to drive a cell into malignancy because cells possess multiple control mechanisms which protect the organism from uncontrolled growth of single cells. Thus, Vogelstein, Fearon and Kinzler [201, 202] established a step-wise hypothesis of cancer formation in the colon postulating that several mutations are required for the development of cancer cells. This Adenoma-Carcinoma Hypothesis describes the formation of certain precancerous lesions and their progression into a manifest cancer. The model implies that adenomas are the precursor lesions of most colorectal cancers and it describes typical molecular events associated with progression to cancer. The existence of adenomas as precursor lesions and their visibility and removability in colonoscopy examinations allow for effective prevention of colorectal cancer. In Germany, a colonoscopy every 10 years after the age of 55 is recommended for the general population, which in fact leads to a significant prevention effect [32]. So, the adenoma-carcinoma sequence of

[32]: Brenner et al. (2010), "Low Risk of Colorectal Cancer and Advanced Adenomas More Than 10 Years After Negative Colonoscopy".

colorectal cancer development fits very well into the scientific observations regarding colorectal cancer prevention by colonoscopy [32] (see also Section 2.5).

The step-wise hypothesis of cancer development has been validated subsequently in many independent studies for many different cancer types. Currently, it is expected for most cancer types that a minimum number of three mutation events is required to transform a normal cell into a cancer cell. This hypothesis is called the *three strike hypothesis* [191].

### 2.2.2 CANCER DEVELOPS THROUGH DIFFERENT PATHWAYS

We have seen in Section 2.1 that not all mutations may lead to a change of the cell behavior. Thus, from all the possible mutations that can occur, the coding mutations have to be identified, as they might have a functional impact on the cell. As described previously, this includes the identification of oncogenes and tumor suppressor genes, but there are many more mutations to be identified. Moreover, only a certain combination of these mutations will lead to cancer in the end. This might be due to the fact that some mutations have a growth-repressing effect and lead to cell death. Further, there is the possibility of controlling cancer by non-cell autonomous mechanisms, like immune surveillance, which is especially important for the example of Lynch syndrome [101] (see Section 2.4). Apart from that, current data raise the possibility that the immune system may not only remove precursor lesions but also may infiltrate cancers, as described for Lynch syndrome-associated cancers [177].

Different combinations of key mutations result in several distinct pathways of carcinogenesis to be distinguished by the involved genes and the ordering thereof. It is especially important for a comprehensive understanding of cancer development and thus for a successful clinical management to investigate which of these pathways can arise in human cancer development, a process called carcinogenesis. In other words, which of the latter are defined pathways of carcinogenesis.

Different pathways of carcinogenesis might lead to molecularly and clinically different types of cancer (see Figure 2.2, middle). Identifying the clinical consequences of individual

[191]: Tomasetti et al. (2014), “Only three driver gene mutations are required for the development of lung and colorectal cancers”.

[101]: Kloor and von Knebel Doeberitz (2016), “The Immune Biology of Microsatellite-Unstable Cancer”.

[177]: Seppälä et al. (2019), “Lack of association between screening interval and cancer stage in Lynch syndrome may be accounted for by over-diagnosis: a prospective Lynch syndrome database report”.

pathways of carcinogenesis for patients is crucial not only for the design of new cancer treatment strategies, but also for efficient cancer prevention approaches (more details in Section 2.5).

### 2.2.3 CANCER DEVELOPS SPORADICALLY OR IN A HEREDITARY CONTEXT

Most of the cancers in the general population occur by chance. These cancers are called sporadic. However, in some families certain types of cancer appear more frequently. This is either a familial or a hereditary form of cancer. The former are due to a combination of genetic and environmental factors but in contrast to hereditary cancers there is not a specific pattern of altered genes which is passed down in the family from parent to child.

From a modeling point of view, the advantage of focusing on hereditary tumors is that there are clearly defined molecular events determining the onset of the disease and thereby representing a known mechanism underlying carcinogenesis (see Figure 2.2). We will further explain these concepts in Section 3.2.

### 2.2.4 DIFFERENT TYPES OF COLORECTAL CANCER

Colorectal cancer is the third most common type of cancer worldwide [184] accounting for about 10% of all cancer cases in 2020. Further, with respect to mortality, it is the second most deadly cancer type with over 576,000 deaths in 2020 [184].

Classical sporadic colorectal cancer is the most common type of colorectal cancer and develops following the adenoma-carcinoma-sequence, introduced in Section 2.2.1. In colon cancer, the three-strike-hypothesis [191] corresponds to the alteration of three molecular signaling pathways: the activation of Wnt signaling with the tumor suppressor gene *APC* being involved, the activation of the EGFR pathway responsible for uncontrolled cell growth with e.g., the *KRAS* gene involved as classical oncogene, and the inactivation of programmed cell death (apoptosis) control mechanisms via

[184]: Sung et al. (2021), “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries”.

[191]: Tomasetti et al. (2014), “Only three driver gene mutations are required for the development of lung and colorectal cancers”.

e.g., mutations in the tumor suppressor gene *TP53*. These alterations on a genome level are linked to phenotypic changes on a cell, crypt and tissue level: Mutations in the *APC* gene in normal epithelium lead to the formation of polyps and early adenomas. An additional mutation in the *KRAS* gene leads to further progression to a late adenoma, and finally mutations in *TP53* may result in a manifest colorectal cancer [150].

Besides that, there are many other types of colorectal cancer occurring under different conditions [113]. An Euler diagram of the different subtypes of colorectal cancer is given in Figure 2.3. A possible molecular distinction of colorectal cancer cases can be done based on the functioning or deficiency of one of the DNA repair mechanisms and the resulting molecular fingerprint of microsatellite stability (MSS) or microsatellite instability (MSI).

A deficient repair mechanism in affected patients is not able to repair errors which occur during DNA replication [108] (see Figure 2.4c). The individuals have a pathogenic germline variant in one of the four so-called mismatch repair genes *MLH1*, *MSH2*, *MSH6* and *PMS2* [53]. This germline variant, in combination with a second somatic hit inactivating the functional allele of the respective affected gene, leads to mismatch repair (MMR) deficiency in the affected cell. Disruption of the normal function of the MMR system leads to the accumulation of uncorrected replication errors. Particularly vulnerable to errors during DNA replication are areas of the genome termed *microsatellites*. These microsatellites are characterized by the same short base pair sequences repeated several times. Therefore, in MMR-deficient cells, insertion of additional bases or deletion of existing bases at these vulnerable microsatellites are the earliest and most prominent changes upon MMR inactivation. The resulting phenotype, called *microsatellite instability* [94] (see Figure 2.4c) is therefore also used to clinically diagnose MMR deficiency (see Section 2.5). As for all other mutations, most of the insertion/deletion mutations occur in non-coding regions of the genome, where they are believed to have no functional consequences. However, an insertion/deletion mutation occurring in a microsatellite which lies in coding regions of the genome, can have severe consequences: It leads to a shift in the reading and interpretation of the subsequent bases, called a frameshift. Such frameshift mutations can trigger the synthesis of completely different, functionally inactive proteins

[150]: Nguyen and Duong (2018), "The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy (Review)".

[113]: Ladabaum (2020), "What Is Lynch-like Syndrome and How Should We Manage It?"

[108]: Kolodner (1996), "Biochemistry and genetics of eukaryotic mismatch repair."

[53]: de la Chapelle (2003), "Microsatellite Instability".

[94]: Jiricny (2013), "Postreplicative Mismatch Repair".

[7]: Alhopuro et al. (2011), “Candidate driver genes in microsatellite-unstable colorectal cancer”.

[60]: Duval et al. (2001), “Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers”.

[213]: Woerner et al. (2001), “Systematic identification of genes with coding microsatellites mutated in DNA mismatch repair-deficient cancer cells”.

[122]: Linnebacher et al. (2001), “Frameshift peptide-derived T-cell epitopes: A source of novel tumor-specific antigens”.

[101]: Kloor and von Knebel Doeberitz (2016), “The Immune Biology of Microsatellite-Unstable Cancer”.

[208]: Wimmer et al. (2014), “Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium Care for CMMRD (C4CMMRD)”.

(see Figure 2.4c). Thereby, such insertion/deletion mutations can inactivate tumor-suppressor genes [7, 60, 213]. The identification of driving insertion/deletion mutations has been done previously using SelTarBase (see also Section 2.4).

Although being wrong and non-functional, proteins generated as a result of a frameshift mutation have a very special feature: They are completely novel to the immune system and, at the same time, highly specific to mismatch repair-deficient cells. This results in a strong response from the patient’s immune system against such frameshift peptides (FSP) and opens up possibilities for novel prevention approaches against these cancers [122], which will be discussed in further detail in Section 2.4.

About 85% of colorectal cancer cases are MMR-proficient and display microsatellite stability meaning that there is no evidence of a deficient mismatch repair system and thus no increase in mutations at microsatellites. This is also true for the above mentioned example of classical sporadic colorectal cancer. In contrast, approximately 15% of all colorectal cancer cases are microsatellite unstable [101] (see Figure 2.4b).

Further, as explained in Section 2.2.3, cancer can occur sporadically or in a hereditary context, which is also true for colorectal cancer. It is estimated that up to 5% to 10% of colorectal cancer is hereditary. Lynch syndrome (LS) is by far the most common inherited colorectal cancer syndrome accounting for around 3% of all colorectal cancer cases, followed by familial adenomatous polyposis (FAP) which is estimated to be up to 30 times less likely than Lynch syndrome. Lynch syndrome individuals have one germline variant in one of the MMR genes such that only a second hit on the other allele of the affected MMR gene is necessary for MMR deficiency. Thus, most of the tumors in those individuals are MSI. Constitutional MMR deficiency (CMMRD) syndrome is defined by biallelic germline variants in one of the MMR genes with MMR deficiency present in all cells from birth on, heavily increasing the risk of developing one or more types of cancer in childhood and early adolescence [208]. FAP individuals have a germline variant in the *APC* gene leading to the development of hundreds to thousands of colorectal polyps and thus to an accelerated cancer development displaying microsatellite stability.

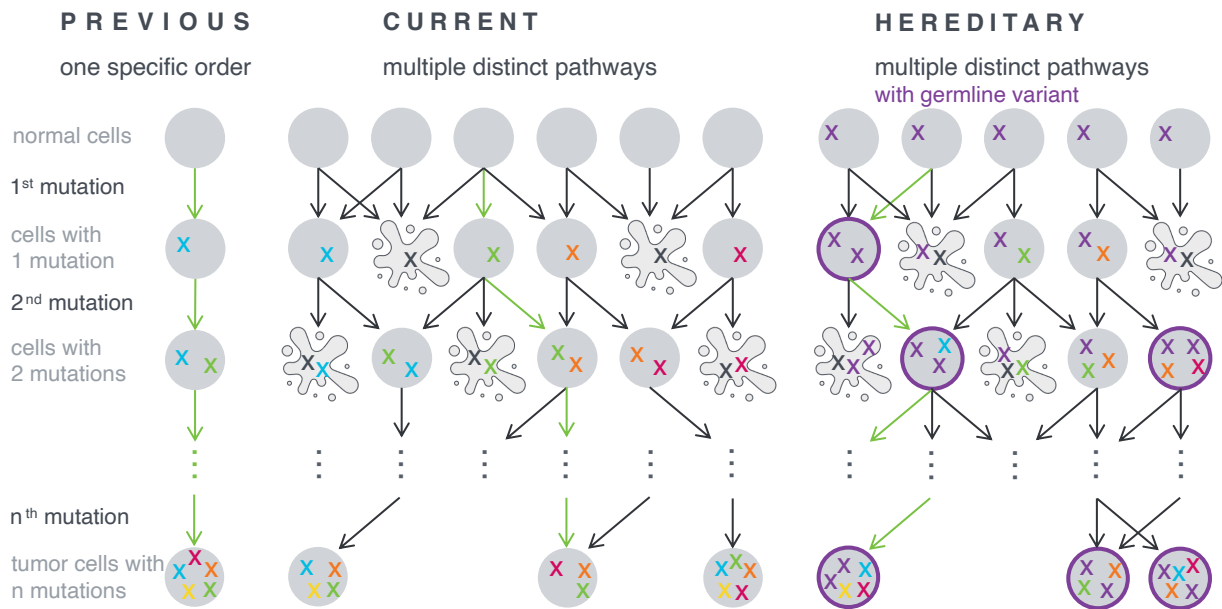
In general, approximately 80% of all MSI colorectal cancers are sporadic, whereas 20% are hereditary, almost always due to Lynch syndrome (Figure 2.4b). Many of the sporadic MSI colorectal tumors possess acquired somatic hypermethylation of the promoter region of the DNA MMR gene *MLH1* [43]. Due to the biallelic methylation of the *MLH1* promoter, the corresponding protein is not transcribed resulting in MMR deficiency. For distinguishing those sporadic MSI and Lynch syndrome-associated colorectal cancers, the following criteria can be used: a) the presence of *BRAF* mutations in sporadic cancers, b) the older age at diagnosis in sporadic cancer patients, c) the lack of significant family history suggesting of Lynch syndrome in sporadic patients, and d) the presence of methylation on the *MLH1* promoter in sporadic patients [43]. Based on these criteria, a separation is usually possible in practice [43], although for formal proof of Lynch syndrome, the detection of a germline variant in one of the MMR genes is required. We evaluate the performance of *BRAF* mutation testing for detecting Lynch syndrome in an age-dependent way later in this dissertation (see Chapter 9).

Another type of cancer, called Lynch-like cancer, is quite difficult to distinguish from Lynch syndrome according to the before mentioned guidelines, as they are quite similar in those aspects. However, Lynch-like cancers show MSI, have no MMR gene germline variant, and show no hypermethylation of *MLH1*. Possible causes for cancer in Lynch-like syndrome are currently discussed [43, 133] with one explanation being two somatic mutations in one of the MMR genes. This makes Lynch-like colorectal cancer a true sporadic counterpart of Lynch syndrome-associated cancer and broadens the possible application of new therapeutic approaches to cancers beyond Lynch syndrome.

[43]: Carethers (2014), "Differentiating Lynch-Like From Lynch Syndrome".

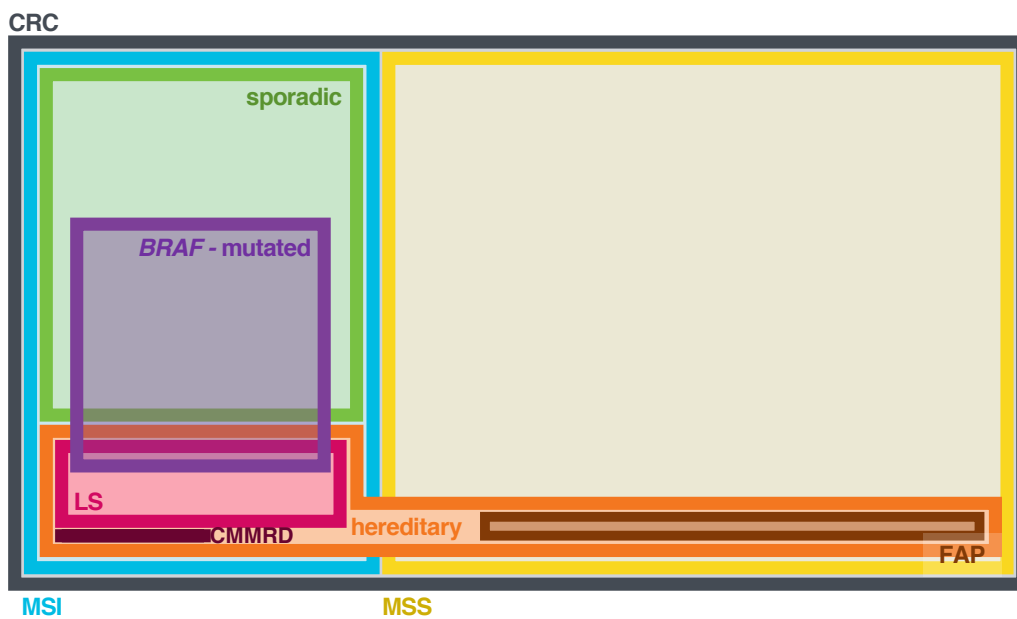
[133]: Mensenkamp et al. (2014), "Somatic Mutations in *MLH1* and *MSH2* Are a Frequent Cause of Mismatch-Repair Deficiency in Lynch Syndrome-Like Tumors".

## CONCEPTS OF CARCINOGENESIS

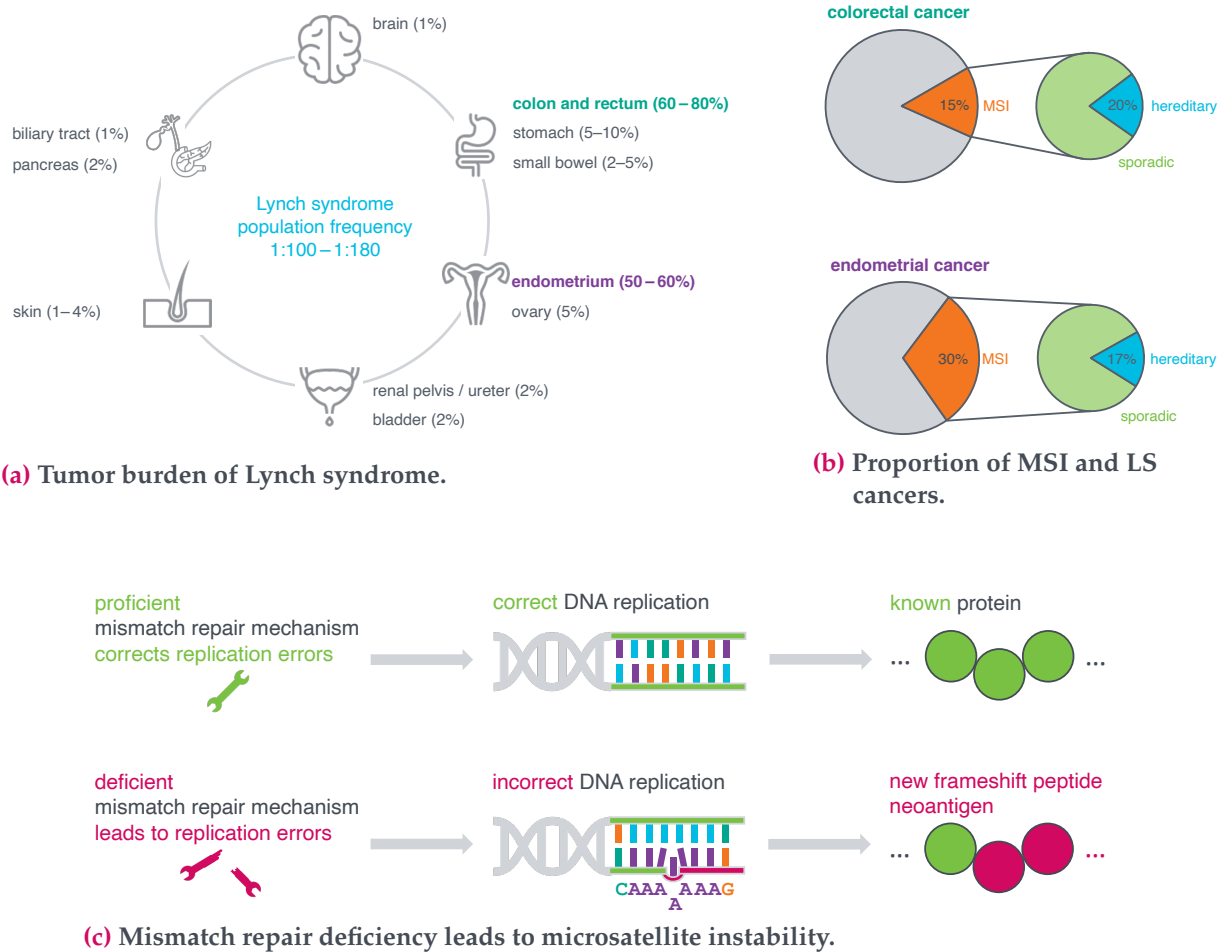


**Figure 2.2: Previous and current concepts of carcinogenesis with hereditary tumors as an example cancer type for mathematical modeling.** The hypothesis of how a tumor develops has changed over the years. *Left:* In the early stages of cancer research, the understanding was that a normal cell (gray circle) obtains several predefined key mutations (marked as X, X, X, X, X) in one specific order (marked by green arrows) as time progresses. This in the end leads to a first malignant cell (marked here with 5 crosses as key mutations), which can outgrow to a manifest tumor. *Middle:* Nowadays, it is hypothesized that there are multiple distinct pathways of carcinogenesis which can lead to a first malignant cell. This hypothesis is inclusive of the fact that the order of the key events in cancer development given by the previous understanding is only explainable in the context of a network of mutational events and pathways. The different pathways in this network exist due to the fact that the order of the key mutations can vary. Further, during the process of tumor development, several mutations (marked as X) can occur which damage the cell in a way which leads directly to cell death (gray deformed cells with no further progression). This means that in total only very few of all these mutated cells will lead to cancer. *Right:* Hereditary tumors are ideal examples for modeling carcinogenesis, as they share a defined causative mechanism, whereas for sporadic tumors there are multiple, uncertain possibilities of cancer initiation, which complicates mathematical modeling approaches. In Lynch syndrome, the most common inherited tumor syndrome, this origin is the inactivation of the mismatch repair system. A first variant affecting one copy of a mismatch repair gene is already present in the germline (marked by X in all cells on the top level). As soon as the second mutation inactivates the second copy of the same mismatch repair gene, the mismatch repair system becomes functionally deficient (cell marked with purple circle and X X), which leads to many more mutations in subsequent cell divisions. As these mutations occur randomly, they mostly destroy the affected cells. However, in cells that develop into tumors, we find certain patterns of mutations enriched with mutations favoring cell survival and tumor formation, termed driver mutations. These are important factors for the ultimate development of the first malignant tumor cell.

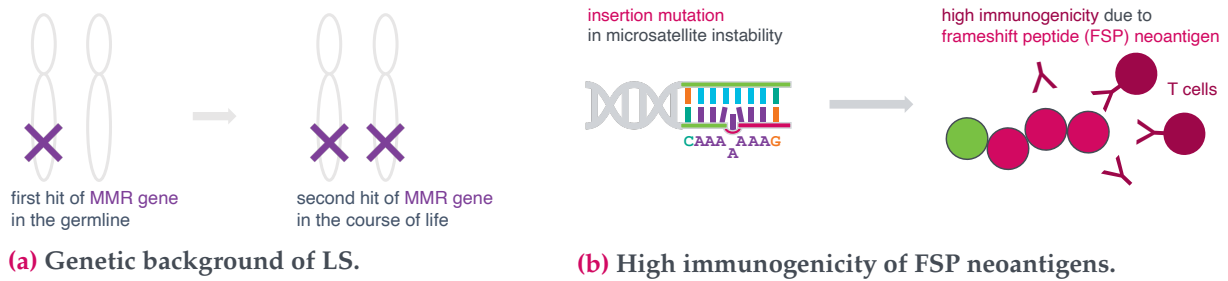




**Figure 2.3:** Euler diagram for the different subtypes of colorectal cancer. First, colorectal cancer (CRC) is separated in microsatellite stable (MSS CRC) and microsatellite unstable (MSI CRC), where the latter is again split in sporadic MSI CRC and hereditary MSI CRC. Most of the CRCs in LS are hereditary, whereas a small number may be sporadic. Approximately 50% of the sporadic MSI CRCs are BRAF-mutated, where the rest of CRCs is BRAF-wild type. Other hereditary forms of CRC include CMMRD (almost always MSI) and FAP (mostly MSS).



**Figure 2.4:** Lynch syndrome as the most common inherited cancer syndrome highly increases the life-time cancer risk of affected individuals. **(a)** Lynch syndrome is estimated to occur with a prevalence of 1 out of 180 to 1 out of 100 in the general population [67] and it includes an increased life-time risk for cancers of the following organs: endometrium, ovary (only for women), renal pelvis and ureter, brain, skin, biliary tract, pancreas, stomach, colon and rectum as well as small bowel. The estimated life-time risks are given in brackets for each cancer type. Figure adapted from [102]. **(b) Top:** 15% of all colorectal cancers show microsatellite instability (MSI), whereby 1 out of 5 of these MSI cancers are hereditary, mainly caused by Lynch syndrome. **Bottom:** 30% of all endometrial cancers show MSI, whereby 1 out of 6 of these MSI cancers are caused by Lynch syndrome. Figure adapted from [101]. **(c) Top:** A proficient mismatch repair system is able to correct errors which occur during DNA replication. Thus, the DNA is usually replicated in a correct way, which in turn is translated into a known and correct protein. **Bottom:** If the mismatch repair mechanism is deficient due to a mutation of one of the underlying mismatch repair genes, called *MLH1*, *MSH2*, *MSH6* and *PMS2*, then DNA replication errors which occur at repetitive sequences (the same base pairs occur successively) cannot be corrected by the mismatch repair system. These mutations are either insertion or deletion mutations (in the figure an additional base **A** is inserted) leading to a shift of the subsequent reading frame. This results in a completely new protein being induced by the frameshift, namely a frameshift peptide neoantigen.



**Figure 2.5: Structured mechanisms of Lynch syndrome carcinogenesis facilitate the development of mathematical models.** (a) The genetic background of Lynch syndrome carcinogenesis is precisely known, since Lynch syndrome patients already have an inherited mutation of one of the mismatch repair genes in the germline (× at the gray chromosome above). For an inactivation of the mismatch repair mechanism, a second hit of the remaining allele of the mismatch repair gene is necessary (second × at the same location on the chromosome at the bottom). In this way, Lynch syndrome carcinogenesis follows Knudson’s two hit hypothesis. (b) Lynch syndrome cancers show a high level of microsatellite instability due to insertion and deletion mutations at microsatellites. As an example, a microsatellite consisting of 6 A bases with an insertion mutation is shown. This leads to a frameshift of all subsequent bases, which in turn leads to completely new frameshift peptide (FSP) neoantigens (protein on the right with changed structure in red). The latter can be detected by the immune system and will lead to a high immune response (T cells on the right).

## 2.3 STATE-OF-THE-ART KNOWLEDGE ON LYNCH SYNDROME COLORECTAL CANCER

As illustrated in the last section, there are different types of colorectal cancer, one of which are Lynch syndrome-associated colorectal cancers. We will explain in this section why we focus on this cancer type within this dissertation and the associated research work.

### 2.3.1 LYNCH SYNDROME IS THE MOST COMMON INHERITED COLORECTAL CANCER SYNDROME

[53]: de la Chapelle (2003), “Microsatellite Instability”.

[100]: Klimstra et al. (2019), “Classification of neuroendocrine neoplasms of the digestive system”.

[102]: Kloor et al. (2005), “Molecular testing for microsatellite instability and its value in tumor characterization”.

[93]: Jaspersion et al. (2010), “Hereditary and Familial Colon Cancer”.

[170]: Robert Koch-Institut (2016), “Cancer in Germany 2011/2012”.

Lynch syndrome is the most common inherited colorectal cancer syndrome [53]. The currently estimated population frequency of Lynch syndrome is one person out of 180 [100]. Individuals with Lynch syndrome are predisposed to developing certain malignancies with a substantially higher life-time risk compared to the general population. The most common Lynch syndrome manifestations are colorectal cancer (60–80% [102] compared to 6% in the normal population) and endometrial cancer (50–60% compared to 2.6% in women without Lynch syndrome) [93, 170]. Further, individuals have an increased life-time risk for many other types of cancer such as in the stomach, small bowel, brain, skin, pancreas, biliary tract, ovary (only for women) and upper urinary tract (see Figure 2.4a). In this dissertation, we derive novel methods to obtain prospective cancer risk estimates for the Lynch syndrome population based on the worldwide largest prospective Lynch syndrome database (PLSD, see Chapter 8). Because of the high tumor burden, it is crucial to develop adequate treatment and prevention approaches for Lynch syndrome individuals (see also Section 2.5).

As explained in Section 2.2.4, tumors of Lynch syndrome carriers serve as an example of tumors showing a deficient repair mechanism, not being able to repair errors which occur during DNA replication [108].

[108]: Kolodner (1996), “Biochemistry and genetics of eukaryotic mismatch repair.”

### 2.3.2 THE THREE PATHWAY HYPOTHESIS OF LYNCH SYNDROME COLORECTAL CARCINOGENESIS

As depicted in Section 2.2.1, a number of key events are necessary in order to develop cancer, which can be combined in different ways leading to different pathways of carcinogenesis. For colorectal cancer, there is this one dominant adenoma-carcinoma sequence hypothesis for carcinogenesis [201, 202] which implies that adenomas are the precursor lesions of most colorectal cancers and it describes typical molecular events associated with tumor progression, namely alterations in *APC*, *KRAS*, and *TP53*. Further, the adenoma-carcinoma model of colorectal cancer development fits very well into the scientific observations regarding colorectal cancer prevention by colonoscopy [32] in Germany.

Due to the high risk of cancer in Lynch syndrome, the recommended colonoscopy intervals are shorter (every 1–3 years) — see Section 2.5 for further explanation of the calculations underlying such clinical guidelines. However, in Lynch syndrome patients, incident cancers occur which apparently cannot be prevented even by further shortening the screening intervals [63]. This means that colorectal carcinogenesis in Lynch syndrome patients differs from that of the general population. Therefore, the model of carcinogenesis in Lynch syndrome needs to be adapted to the empirical observations.

The first step towards an adaptation of the model is the acceptance of different pathways of Lynch syndrome colorectal carcinogenesis, not necessarily always starting with or at all involving an adenoma phase [2]. This theory was first based on the discovery of MMR-deficient crypts, that can be detected in the normal colonic mucosa of Lynch syndrome mutation carriers but not in sporadic colorectal cancer patients [104, 105]. MMR-deficient crypts are histologically normal crypts from normal colonic mucosa that lack the expression of the MMR gene affected in the germline. Due to the loss of MMR function, these crypts present with MSI and thereby possess the potential to transform into malignant lesions [158, 182] (see Figure 2.6).

[201]: Vogelstein and Kinzler (1993), “The multistep nature of cancer”.

[202]: Vogelstein et al. (1988), “Genetic Alterations during Colorectal-Tumor Development”.

[32]: Brenner et al. (2010), “Low Risk of Colorectal Cancer and Advanced Adenomas More Than 10 Years After Negative Colonoscopy”.

[63]: Engel et al. (2018), “No Difference in Colorectal Cancer Incidence or Stage at Detection by Colonoscopy Among 3 Countries With Different Lynch Syndrome Surveillance Policies”.

[2]: Ahadova et al. (2018), “Three molecular pathways model colorectal carcinogenesis in Lynch syndrome”.

[104]: Kloor et al. (2011), “Analysis of EPCAM Protein Expression in Diagnostics of Lynch Syndrome”.

[105]: Kloor et al. (2012), “Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study”.

[158]: Pai et al. (2018), “DNA mismatch repair protein deficient non-neoplastic colonic crypts: a novel indicator of Lynch syndrome”.

[182]: Staffa et al. (2015), “Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters”.

[65]: Fearon (2011), “Molecular Genetics of Colorectal Cancer”.

[2]: Ahadova et al. (2018), “Three molecular pathways model colorectal carcinogenesis in Lynch syndrome”.

[175]: Sekine et al. (2017), “Mismatch repair deficiency commonly precedes adenoma formation in Lynch Syndrome-Associated colorectal tumorigenesis”.

[1]: Ahadova et al. (2016), “CTNNB1-mutant colorectal carcinomas with immediate invasive growth: a model of interval cancers in Lynch syndrome”.

As described previously, in general, key events in carcinogenesis are the mutations in tumor suppressor genes and oncogenes. In Lynch syndrome, an additional key event is MMR inactivation, which occurs due to a second somatic mutation in one of the four MMR genes (Figure 2.5a). The essential question is: In which order do the defined mutations required for cancer development occur? At first, MMR inactivation was not regarded as an initiating event of Lynch syndrome carcinogenesis, but merely as an accelerator initiated by other somatic mutations of tumor suppressor genes, like *APC*, or oncogenes such as *KRAS* [65]. However, recent studies on Lynch syndrome-associated colorectal cancer have revealed that even those early somatic mutations carry very specific mutational signatures characteristic of MMR deficiency [2, 175]. This observation, together with the existence of MMR-deficient crypt foci, suggests initiation of colorectal cancer development in Lynch syndrome by MMR inactivation as one major pathway of carcinogenesis.

As MMR-deficient crypt foci cannot even be microscopically detected based on morphology alone, these lesions are likely hard to detect or not detectable at all by colonoscopy, thereby potentially explaining the observation of incident cancers in Lynch syndrome. Of course, this pathway of carcinogenesis does not exclude the possibility of colorectal cancer development in Lynch syndrome with MMR inactivation as a secondary event, or the development of an adenoma later from an MMR-deficient crypt. Altogether, this leads to the current *three pathway hypothesis* of Lynch syndrome cancer development [2] illustrated in Figure 2.7. The relative proportion of one or the other pathway of carcinogenesis and the contribution of certain molecular events is thereby an open question to be addressed by mathematical cancer modeling [1, 2].

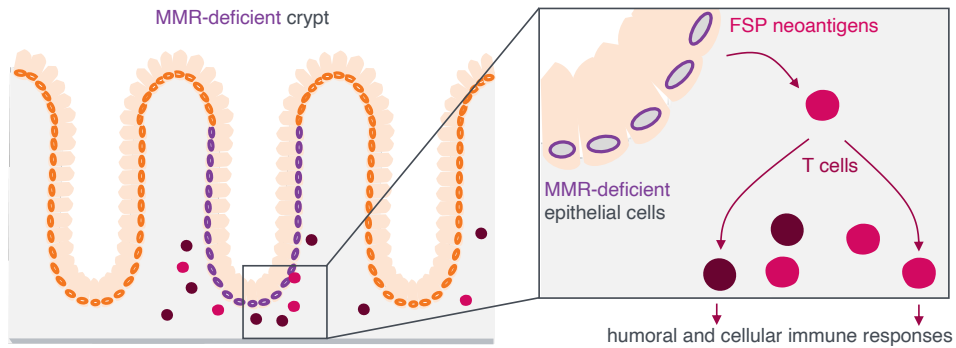
Solving the problem of distinct pathways in Lynch syndrome has immediate clinical implications because polypous precursors initiated by *APC* mutations can be detected by colonoscopy, whereas cancers initiated by MMR deficiency, not progressing through a polypous phase, may not be detectable in this way (see Figure 2.7). There are, however, alternative prevention strategies which have been evaluated in clinical trials. These include chemoprevention approaches, and specific stimulation of the immune system by vaccination in order to recognize and eliminate MMR-deficient cells.

Safety and clinical feasibility of such a vaccination approach have been demonstrated in a clinical study [40, 169]. Protective immunity with reduced tumor burden and improved overall survival was shown in a Lynch syndrome mouse model [70].

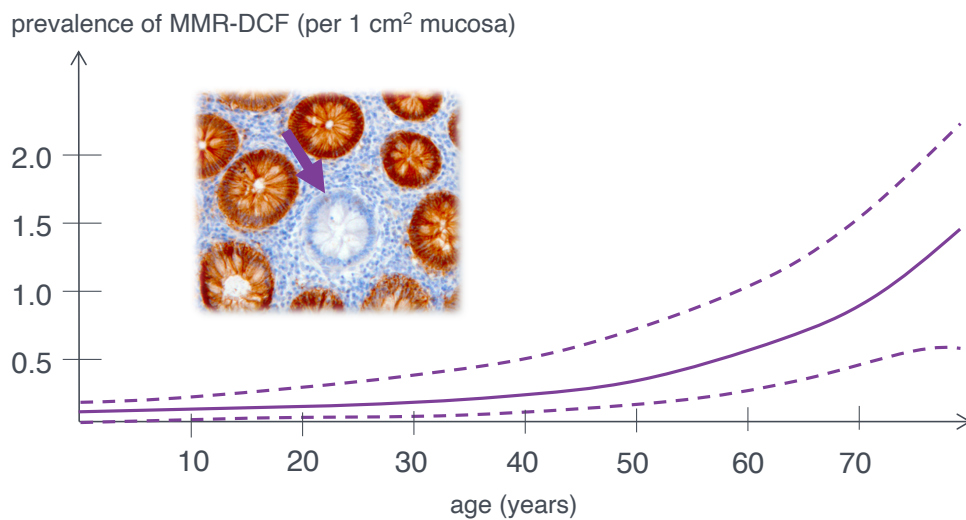
[40]: Burn et al. (2011), "Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial".

[169]: Reuschenbach et al. (2014), "A multiplex method for the detection of serum antibodies against in silico-predicted tumor antigens".

[70]: Gebert et al. (2021), "Recurrent Frameshift Neoantigen Vaccine Elicits Protective Immunity With Reduced Tumor Burden and Improved Overall Survival in a Lynch Syndrome Mouse Model".



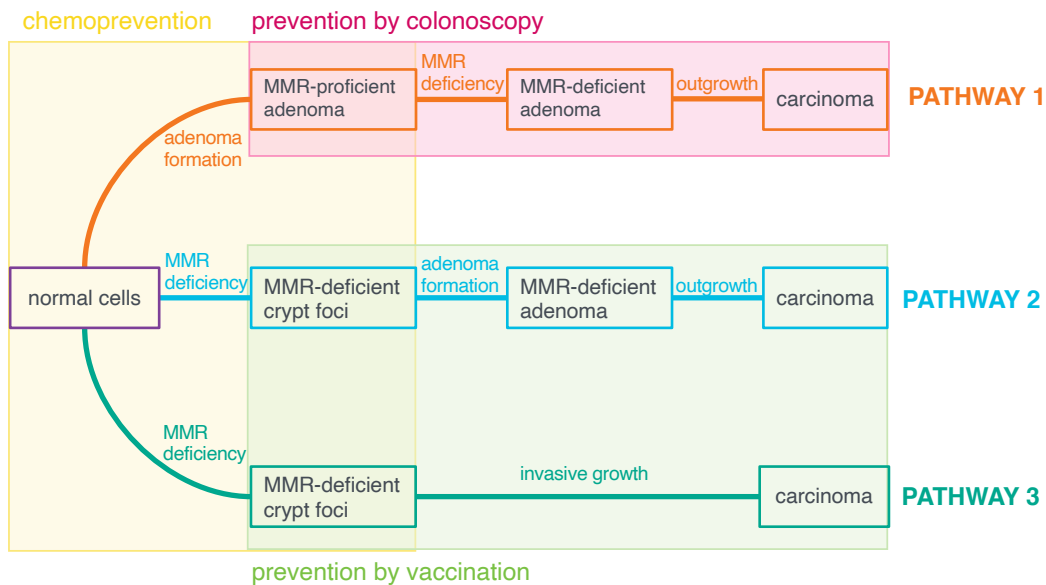
(a) A schematic MMR-deficient crypt in the colonic mucosa of a Lynch syndrome individual.



(b) Age-dependent prevalence of MMR-deficient crypt foci.

**Figure 2.6:** MMR-deficient crypt foci occur at a high frequency in phenotypically normal tissue in Lynch syndrome patients. (a) MMR-deficient crypts (affected cells colored in purple) may be recognized by the host's immune system (T cells in pink and dark red). They harbor coding microsatellite mutations that can give rise to the generation of FSP neoantigens, even before a clinically manifest tumor develops. Humoral and cellular immune responses against FSP neoantigens have been detected in healthy, tumor-free Lynch syndrome carriers. (b) The prevalence of MMR-deficient crypt foci, which are clusters of MMR-deficient crypts, increases with age (dashed line: 95% confidence interval, derived from [182]). This may be responsible for the increased incidence of colorectal cancer with higher age in Lynch syndrome. The inlay shows immunohistochemical staining of an MMR-deficient crypt in a Lynch syndrome carrier with loss of the mismatch repair mechanism, which is marked by a purple arrow. The precise consequences of MMR-deficient crypts on the induction of immune responses over time in Lynch syndrome are not yet known and require further research.





**Figure 2.7: The Three Pathway Hypothesis [2] is the current understanding of Lynch syndrome colorectal cancer development.** In principle, colorectal cancer in Lynch syndrome can only develop in three different ways: **PATHWAY 1:** the classical adenoma-carcinoma pathway of colorectal carcinogenesis, which also occurs sporadically in the general population, with MMR deficiency as a late event. This pathway might be prevented by regular colonoscopy (pink box). **PATHWAY 2:** Here, MMR deficiency is the initiating event followed by adenoma formation and outgrowth to a carcinoma. It might be prevented by a vaccine (light green box). **PATHWAY 3:** Starting with MMR deficiency, MMR deficient crypt foci build up and grow directly into a malignant tumor without developing into an adenoma. This pathway corresponds to the incident cancer cases and might be prevented by a vaccine as well (light green box). Chemoprevention might be a useful approach at the beginning of all of the three pathways of carcinogenesis (yellow box). The model has further implications beyond the prevention aspect as it indicates a biological heterogeneity of Lynch syndrome colorectal cancer, which is rooted in their evolutionary history and which could be reflected by their clinical behavior, prognosis and therapy response. Figure adapted from [2].

## 2.4 STATE-OF-THE-ART INSIGHTS IN LYNCH SYNDROME CANCER IMMUNOLOGY

[101]: Kloor and von Knebel Doeberitz (2016), “The Immune Biology of Microsatellite-Unstable Cancer”.

[212]: Woerner et al. (2003), “Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes”.

[214]: Woerner et al. (2009), “SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology”.

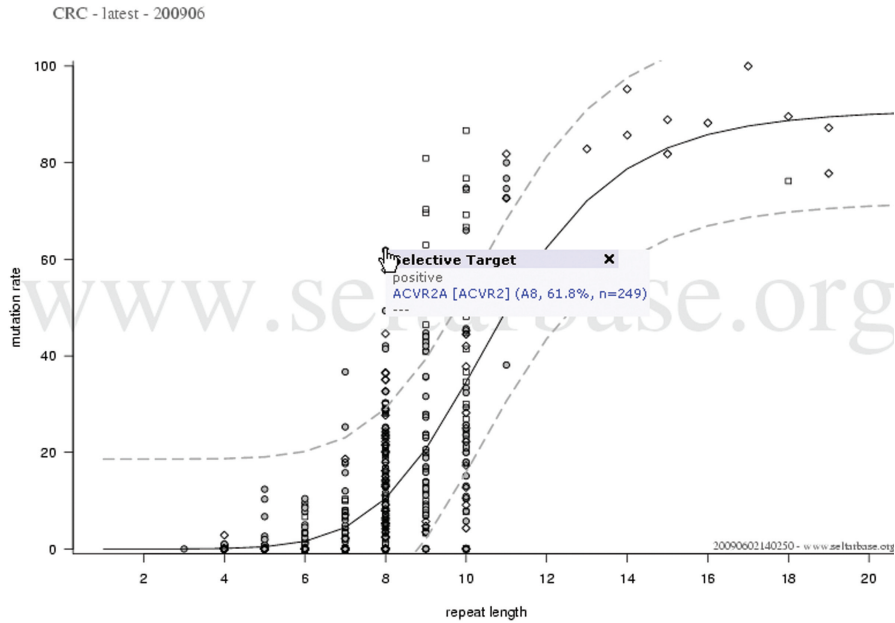
[17]: Ballhausen et al. (2020), “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoeediting during tumor evolution”.

As illustrated in Figure 2.5b, mismatch repair deficiency results in the accumulation of mutations at microsatellites, leading to the generation of highly immunogenic frameshift peptide neoantigens [101]. This means there is a strong link and interaction between cancer development and immunology which we will describe in more detail in this section.

The likelihood of microsatellite mutations should theoretically increase with the length of the respective microsatellite. This was demonstrated by a model by Woerner et al. [212] showing the mutation rate of microsatellites in relation to their length: longer microsatellites are more frequently mutated than shorter ones (Figure 2.8). Interestingly, certain microsatellites show a higher mutation rate than predicted by the regression model, pointing to their selective advantage during microsatellite unstable carcinogenesis (adapted from [214]).

Our research partner in the Mathematics in Oncology project, ATB, initiated a study to analyze the neoantigen landscape of microsatellite unstable cancers (see Chapter 5). The author of this Ph.D. thesis contributed the mathematical analysis and interpretation of the results. It showed that during the microsatellite unstable carcinogenesis, mutations resulting in the generation of highly immunogenic frameshift peptides are counterselected. Therefore, in microsatellite unstable colorectal cancer, they occur less frequently than those that result in weakly immunogenic frameshift peptide neoantigens [17]. This study was made possible by a newly developed bioinformatics-based tool, called ReFrame which is publicly available on GitHub ([github.com/atb-data/neoantigen-landscape-msi](https://github.com/atb-data/neoantigen-landscape-msi)). Importantly, a set of mutations can occur frequently despite giving rise to highly immunogenic frameshift peptides, confirming the tumor-promoting effect of those mutations [17]. Thus, the evolution of Lynch syndrome-associated cancers is mainly guided by two factors: the effect of the acquired mutation on the cell growth, and the immunogenicity of the frameshift peptide generated as a result of this mutation.

Elimination of cell clones with highly immunogenic frameshift peptides occurs naturally due to the immune reaction



**Figure 2.8:** Regression analysis results of colorectal cancer microsatellite mutation rates. A sigmoid regression analysis was performed, where the fitted regression line is drawn as solid black line, the upper and lower prediction lines as bold dashed gray lines (adapted from [214]; screenshot of the SelTarBase website [www.selTarBase.org](http://www.selTarBase.org), 18/11/2022).

of the body against the frameshift peptides [38, 57, 122, 138, 163]. However, 70% of microsatellite unstable tumors acquire immune evasion mechanisms allowing these tumors to thrive despite dense immune infiltration [157]. Most frequently, these mechanisms involve the impairment of the antigen presentation machinery, such as mutations of the *Beta-2-Microglobulin (B2M)* gene [103] (see Figure 2.9). The outgrowth of cell clones with impaired antigen presentation is particularly common in an active immune microenvironment [61, 91, 162].

At the same time, high immunogenicity of frameshift peptide neoantigens gives an opportunity for the design of primary preventive measures, such as vaccines, that would be able to stop the progression at the point when MMR deficiency-induced frameshift peptide neoantigens first occur in colonic crypts, most likely long before malignant transformation. In order to design an effective vaccine, it is important to account for immuno-editing during cancer development: frameshift peptides detected in manifest tumors may not be of highest immunogenicity, as the cell clones with highly immunogenic frameshift peptides could have been eliminated at the initial stages of carcinogenesis. Here, mathematical modeling could facilitate the understanding of immuno-editing and deliver answers that would have otherwise been obtained by

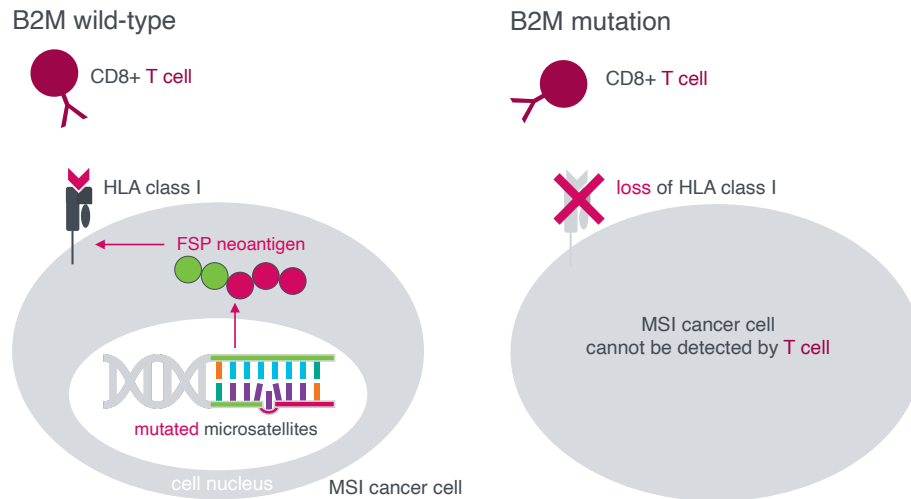
[157]: Ozcan et al. (2018), “Complex pattern of immune evasion in MSI colorectal cancer”.

[103]: Kloor et al. (2007), “Beta2-microglobulin mutations in microsatellite unstable colorectal tumors”.

[61]: Echterdiek et al. (2015), “Low density of FOXP3-positive T cells in normal colonic mucosa is related to the presence of beta2-microglobulin mutations in Lynch syndrome-associated colorectal cancer”.

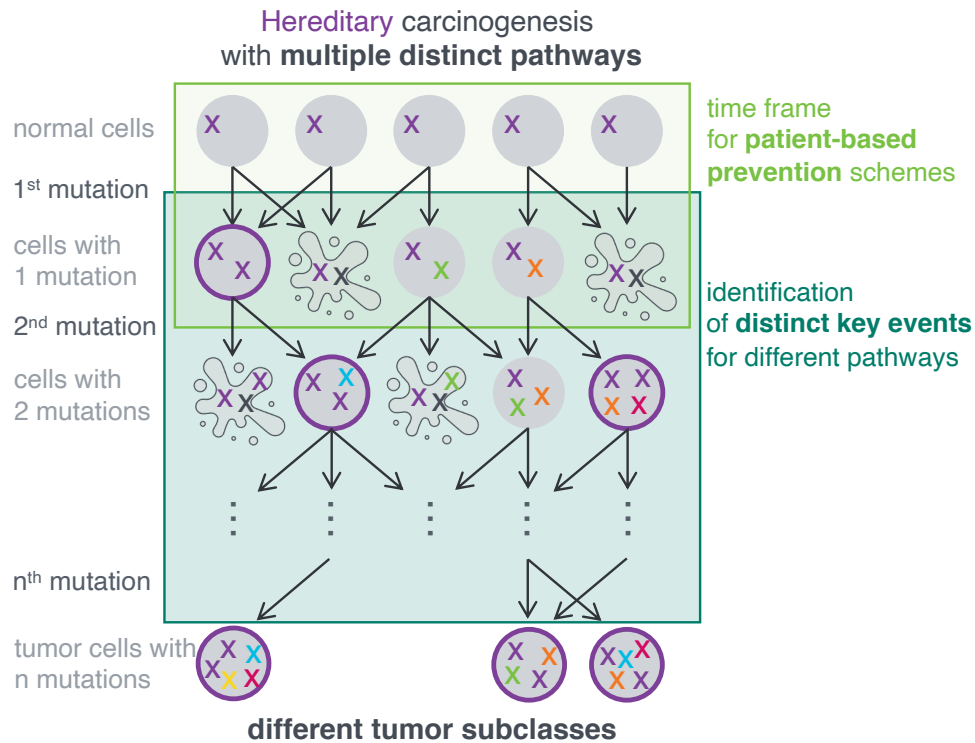
[91]: Janikovits et al. (2017), “High numbers of PDCD1 (PD-1)-positive T cells and *B2M* mutations in microsatellite-unstable colorectal cancer”.

[162]: Pfuderer et al. (2019), “High endothelial venules are associated with microsatellite instability, hereditary background and immune evasion in colorectal cancer”.



**Figure 2.9: Mechanisms of immune evasion of microsatellite unstable tumors by B2M mutations.** *Left:* Normally, microsatellite unstable cancer cells present frameshift peptide neoantigens, which are mediated by the Human Leukocyte Antigen (HLA) class I molecules. The latter play an essential role in the human immune system, as they are responsible for its regulation. A specific type of T cells, CD8-positive T cells, is attracted by the frameshift peptide neoantigens presented by HLA class I and can thus destroy the cancer cells. *Right:* Beta-2-Microglobulin (B2M) mutations are the most common alterations leading to immune evasion through a complete breakdown of HLA class I-mediated presentation of frameshift peptide neoantigens in microsatellite unstable cancer. This means B2M mutations induce a complete lack of assembled HLA class I antigens on the tumor cell surface. As a consequence, CD8-positive T cells cannot attack B2M-mutant microsatellite unstable cancer cells.

numerous animal studies and clinical trials. Thus, mathematical modeling of evolutionary processes in Lynch syndrome carcinogenesis will deliver answers not only with regard to carcinogenic pathways and risk delineation in mutation carriers, but also support the development of effective cancer-preventive approaches (see Figure 2.10), a research direction that is currently of high interest in the mathematical oncology community (see Chapter 3).



**Figure 2.10:** The mathematical identification of key events in carcinogenesis is helpful to sharpen the time frame for patient-based cancer prevention schemes. There are multiple distinct pathways leading to different final tumor subclasses. The pathways incorporate distinct key events, which have to be identified (dark green box) using a mathematical modeling approach. When the key events with their time of occurrence are identified, it is possible to sharpen the time frame in which a prevention scheme is applicable (light green box). As the process of carcinogenesis is patient-specific, this is also true for the corresponding prevention scheme, which can then be translated into the clinical context by use of appropriate modeling techniques.

## 2.5 HIGH CLINICAL NEEDS IN LYNCH SYNDROME

As outlined in Section 2.3.2, the hereditary cancer scenario gives a good way to study the multiple-strike model of carcinogenesis, which generally provides a valuable representation of human cancer development in all common cancer types including breast, colon, lung, kidney cancer etc. Lynch syndrome represents a reasonable model disease due to the well-characterized steps of the carcinogenic mechanism further explained in Section 3.2. In addition to serving as a model disease, there is also an increased medical need to be met for individuals affected by Lynch syndrome. This most common cancer predisposition syndrome with half a million people affected in Germany alone is associated with a dramatically elevated risk of developing colon cancer (see Figure 2.4b), and surveillance and prevention strategies currently available are only partially effective. This situation is

[3]: Ahadova et al. (2020), “The unnatural history of colorectal cancer in Lynch syndrome: Lessons from colonoscopy surveillance”.

[4]: Ahadova et al. (2021), “Distinct Mutational Profile of Lynch Syndrome Colorectal Cancers Diagnosed under Regular Colonoscopy Surveillance”.

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

[41]: Busch et al. (2021), “Beta-2-microglobulin Mutations Are Linked to a Distinct Metastatic Pattern and a Favorable Outcome in Microsatellite-Unstable Stage IV Gastrointestinal Cancers”.

[64]: Engel et al. (2020), “Associations of Pathogenic Variants in *MLH1*, *MSH2*, and *MSH6* With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome”.

[112]: Kuntz et al. (2011), “A Systematic Comparison of Microsimulation Models of Colorectal Cancer”.

[31]: Brenner et al. (2011), “Sojourn Time of Preclinical Colorectal Cancer by Sex and Age: Estimates From the German National Screening Colonoscopy Database”.

[32]: Brenner et al. (2010), “Low Risk of Colorectal Cancer and Advanced Adenomas More Than 10 Years After Negative Colonoscopy”.

[92]: Järvinen et al. (1995), “Screening reduces colorectal cancer rate in families with hereditary nonpolyposis colorectal cancer”.

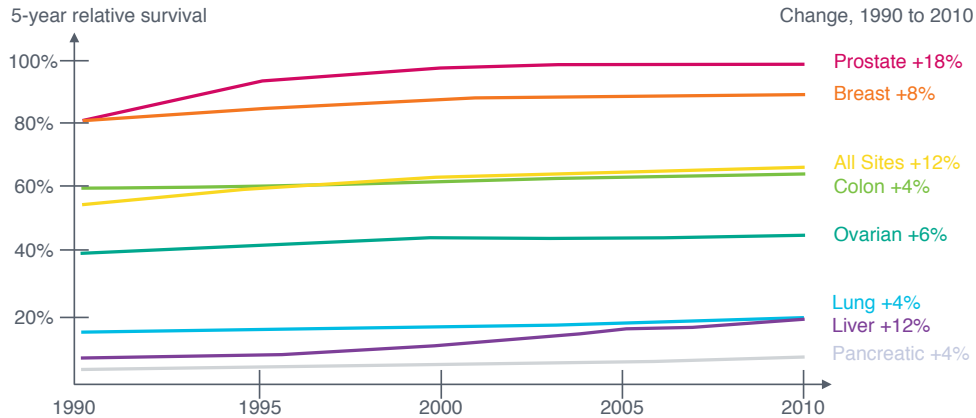
[196]: Vasen et al. (2007), “Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer)”.

mainly caused by the heterogeneity of Lynch syndrome [3, 4, 27, 41, 64], as a disease from a genotypic and phenotypic point of view which is currently not reflected in clinical guidelines. Thus, there is a substantial need for improvement to alleviate the severe psychological burden that Lynch syndrome confers to affected individuals and their relatives.

One of the characteristics that renders colorectal cancer a highly suitable model of solid cancer development is the existence of certain detectable and removable precursor lesions, i.e., adenomas, which can be subject to molecular analysis. The three-strike-hypothesis, which is the core of the adenoma-carcinoma model in sporadic colorectal cancer described in Section 2.2.1, commonly encompasses the *APC-KRAS-TP53* sequence. As described in Section 2.2.4, there is one additional feature involved in Lynch syndrome: MMR deficiency. Although the three major functional alterations resemble those of sporadic colorectal cancer, Lynch syndrome cancers lose the capacity of DNA mismatch repair at a certain time point during their development. In order to adapt colorectal cancer prevention protocols from the general population to Lynch syndrome, understanding MMR deficiency and its onset — as the factor differentiating Lynch syndrome from the general population — is essential.

In the general population, the time from normal colonic mucosa to colon cancer is estimated to range from 10.6 to 25.8 years [112], whereas time from a precancerous lesion to colon cancer was estimated to vary between 4.5 to 5.8 years [31]. Therefore, an interval of 10 years is accepted as safe and effective for preventing colorectal cancer [32]. Due to the accelerating factor of MMR deficiency, it is recommended that colonoscopy intervals for Lynch syndrome patients should be shorter compared to the general population (depending on the country, between every year and every 3 years [92, 196]).

However, after more than 10 years of observation using prospective data, evidence suggests that despite colonoscopy prevention, colorectal cancer remains the most frequently observed cancer in Lynch syndrome mutation carriers [142]. Lynch syndrome patients have a risk of more than 15% of developing an incident cancer over 10 years [63, 131, 197], even under colonoscopy surveillance, indicating that current programs are not suited to substantially reduce the disease burden (see Figure 2.11).



**Figure 2.11:** The US 5-year relative survival for all ages, races and gender has only slightly increased from 1990 to 2010 for all types of cancer, especially for colorectal cancer. Adapted from SEER Survival Statistics, Period Analysis for 2010, 18 Registry Data Set.

This is mainly attributable to the fact that appropriate models reflecting the heterogeneity of genetic predisposition and the complexity of cancer evolution have been lacking. For example, the inherited MMR gene variant may have an impact on the pathogenic pathway and clinical phenotype of the manifest cancer. A recent prospective study analyzing the adenoma and carcinoma risk in Lynch syndrome carriers under colonoscopy surveillance shows significant differences between the carriers of *MLH1*, *MSH2*, and *MSH6* pathogenic variants [64]. Importantly, these differences in the clinical manifestation of Lynch syndrome were also reflected in the mutational characteristics of the manifest cancers between different MMR gene carriers [64]: Whereas *MLH1* carriers showed low adenoma risk and high incident cancer risk, and presented with *CTNNB1* mutations, *MSH2* carriers presented with high adenoma and high incident cancer risk, and often displayed *APC* mutations in cancers. In contrast to those, *MSH6* carriers showed high adenoma and low incident cancer risk, and exclusively *APC* mutations. This means, Lynch syndrome cancers are heterogeneous, and both carcinogenesis and clinical manifestation may differ depending on several factors, which have to be accounted for in a mathematical model.

[64]: Engel et al. (2020), "Associations of Pathogenic Variants in *MLH1*, *MSH2*, and *MSH6* With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome".





## Novel medical hypothesis testing uses mathematical modeling



## 3 CHALLENGES AND OPPORTUNITIES FOR MODELING IN LYNCH SYNDROME

Mathematical oncology is a quite young research field that is constantly increasing. One of the main advantages of including mathematics in cancer research is the fast and resource-saving implementation. In contrast to performing various *in vitro* and *in vivo* clinical studies in a sequential way, mathematical models can evaluate different hypotheses about tumor evolution simultaneously in a time- and cost-efficient way. This in turn can be used to analyze and optimize different approaches for tumor prevention including chemoprevention and vaccination. Thus, mathematical oncology leads to a fast and resource-saving clinical implementation of the modeling results (see chapter image).

In this chapter, we give a short overview of the field, including example publications for modeling carcinogenesis and analyzing the mutational history of different types of cancer which is still one of the main research activities. We describe state-of-the-art computational and mathematical modeling approaches at the cell and crypt levels respectively to describe different aspects of colorectal carcinogenesis which will serve as related work to our developed modeling approaches. We highlight some biomedical concepts and hypotheses which are in our view key for a detailed understanding of colorectal cancer development in Lynch syndrome. Thus, we describe important ingredients and challenges for model development in the context of Lynch syndrome colorectal carcinogenesis at different scales. Further, we emphasize general aspects for model selection and adaptation to make a mathematically rigorous analysis possible and to allow for a generalization to

3.1	HOW TO MODEL CARCINOGENESIS . . .	48
3.2	LS IS VALUABLE FOR MODELING . . . . .	50
3.3	STATE-OF-THE-ART COMPUTATIONAL CELL MODELING . . . . .	51
3.4	STATE-OF-THE-ART MATHEMATICAL CRYPT MODELING . . . . .	55

[10]: Armitage and Doll (1954), “The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis”.

[11]: Armitage and Doll (1957), “A Two-stage Theory of Carcinogenesis in Relation to the Age Distribution of Human Cancer”.

[98]: Kendall (1960), “Birth-and-Death Processes, and the Theory of Carcinogenesis”.

[178]: Serio (1984), “Two-stage stochastic model for carcinogenesis with time-dependent parameters”.

[185]: Tan and Brown (1988), “A nonhomogeneous two-stage model of carcinogenesis”.

[186]: Tan and Hanin (2008), *Handbook of Cancer Models with Applications*.

[15]: Baker et al. (2014), “Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon”.

[16]: Baker et al. (2019), “Crypt fusion as a homeostatic mechanism in the human colon”.

[26]: Binder et al. (2017), “Genomic and transcriptomic heterogeneity of colorectal tumours arising in Lynch syndrome”.

[55]: DESPER et al. (1999), “Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data”.

[71]: Gerstung et al. (2009), “Quantifying cancer progression with conjunctive Bayesian networks”.

[73]: Gerstung et al. (2020), “The Evolutionary History of 2,658 Cancers”.

[213]: Woerner et al. (2001), “Systematic identification of genes with coding microsatellites mutated in DNA mismatch repair-deficient cancer cells”.

other carcinogenic scenarios. We conclude by pointing why Lynch syndrome serves as a valuable example for studying various aspects of carcinogenesis.

### 3.1 HOW TO MODEL CARCINOGENESIS

First attempts to build mathematical models in cancer research were made in the middle of the 20th century. Armitage and Doll [10, 11] proposed and analyzed one of the first multistage models of carcinogenesis, which are based on the hypothesis that there are multiple subsequent steps before a cancer is formed. The model was extended in the following years [98, 178]. Among the first to consider a model of multiple pathways of carcinogenesis were Tan et al. [185, 186]. These are based on the hypothesis that there are several possible ways in which cancer can develop. With the increasing medical knowledge about cancer development, it became more and more evident that a single model describing the whole process of carcinogenesis from the genomic, over the cell, up to the tissue, organ and organism-level is too complex to build. Nowadays, there exist different types of models describing individual aspects of carcinogenesis (in an unordered list of example publications):

- ▷ Modeling **healthy tissue formation**, such as the evolution of colonic crypts [15, 16, 26],
- ▷ detecting **driver genes** [55, 71, 73, 213],
- ▷ estimating the most likely **temporal order of key mutations** [136, 191],
- ▷ modeling the **cancer-immune system interactions**, including neoantigen presentation [17, 39, 115],
- ▷ predicting **effects of intervention strategies** on tumor growth and patient survival, such as the effect of screening on adenoma risk [189].

From a mathematical point of view, modeling makes use of different approaches, such as ordinary differential equations [13, 111], partial differential equations [123], stochastic processes [90, 153], graph theory [22, 148, 193], and statistics [37, 47], to only name a few publications.

### 3.1.1 HOW TO ANALYZE THE MUTATIONAL HISTORY OF TUMORS

Studying key events of carcinogenesis is not only important in Lynch syndrome, but also central to cancer research in general. Which mutations are necessary for cancer to develop, and in which chronological orders can these mutations occur? To answer this question, genome sequencing data is required. Such data can then be used to build mathematical models estimating the mutational history of different cancers.

There are several models for sequence data-based reconstruction of tumor development, which have been published in the literature. Desper et al. [55] proposed a model based on the mathematical concept of directed acyclic graphs, which they called oncogenetic trees. This was generalized in different ways: First, the assumptions on the graph structure were weakened [20, 21]. Then, several extensions were established to deal with time dependency of cell duplication processes as well as noisy and partially observed data [72]. The application of these approaches to real data is limited, as they are either too complex or need too much storage and computing time to be feasible for implementation in a clinical procedure.

For hereditary colorectal cancers, Komarova et al. [109, 111] proposed a model for the occurrence and ordering of key events during carcinogenesis based on ordinary differential equations, which was adapted to sporadic carcinogenesis. In particular, it addresses the question of the extent of genetic instability as an early event in carcinogenesis. The modeling approach was based on synthetic data because the parameters used are hard to measure in vivo. We will have a closer look at this model for the crypt level in Section 3.4.

Williams et al. [207] proposed a mathematical approach which is able to distinguish mutations that happened early in carcinogenesis from those that occurred at a later time, using real-world sequencing data with a computationally feasible model. It was first designed for the case of neutral tumor evolution, meaning that the occurring mutations do not change the growth behavior of the cells. Subsequently, it has been adapted to the non-neutral case with positive [183] and negative selection pressure [115]. Positive selection leads to an increased cell proliferation rate, whereas negative selection leads to a decreased cell proliferation rate.

[55]: DESPER et al. (1999), “Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data”.

[20]: Beerenwinkel and Sullivant (2009), “Markov models for accumulating mutations”.

[21]: Beerenwinkel et al. (2006), “Evolution on distributive lattices”.

[72]: Gerstung et al. (2011), “The Temporal Order of Genetic and Pathway Alterations in Tumorigenesis”.

[109]: Komarova et al. (2003), “Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability”.

[111]: Komarova et al. (2002), “Dynamics of Genetic Instability in Sporadic and Familial Colorectal Cancer”.

[207]: Williams et al. (2016), “Identification of neutral tumor evolution across cancer types”.

[183]: Sun et al. (2017), “Between-region genetic divergence reflects the mode and tempo of tumor evolution”.

[115]: Lakatos et al. (2020), “Evolutionary dynamics of neoantigens in growing tumors”.

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

[201]: Vogelstein and Kinzler (1993), “The multistep nature of cancer”.

A recent paper by Paterson et al. [160] presents a model for quantifying the evolutionary dynamics of sporadic MSS colorectal cancer initiation and progression based on describing the occurrence of key driver mutations, following the three-strike-hypothesis [201]. We will further explore this model in Section 3.4.

All these models are capable of identifying the coarse structure of the tumor evolution process for the studied tumors by focusing on one particular aspect of carcinogenesis. However, a unifying model incorporating different biological mechanisms, like the probability of cell death or the interplay of different genetic mutations is lacking. Especially the latter is essential in order to take into account the molecular level of carcinogenesis including the role of different genes in signaling pathways.

### 3.2 LYNCH SYNDROME AS A VALUABLE EXAMPLE FOR MODELING

The Ph.D. thesis and the related research work focuses on modeling the carcinogenesis of Lynch syndrome-associated cancers. They serve as a valuable example cancer type due to the following main reasons:

- ▷ **Well-defined carcinogenesis:** Humans have two copies (alleles) of each gene in each cell. Thus, most genes require two mutations to impair protein function (Knudson’s two hit hypothesis [107]). This is especially true for tumor suppressor genes, as a single functional tumor suppressor gene is often sufficient.

In Lynch syndrome, the first mutation, or hit, is already passed on in the family from parent to child (*germline variant*). Therefore, all body cells including cancer cells and precursors harbor one defined identical alteration of the genome. This reduces the number of unknowns and allows for tailored, effective and testable mathematical modeling. The enhanced number of pre-cancerous events in Lynch syndrome carriers due to the presence of the germline variant opens the possibility of studying very early steps of cancer development, which usually represent a black box in cancer research (see Figure 2.2).

[107]: Knudson (1971), “Mutation and Cancer: Statistical Study of Retinoblastoma”.

Tumor manifestation in Lynch syndrome requires a second hit inactivating the remaining functional allele (Figure 2.5a). The time point when this second hit occurs is random; therefore, different pathways of carcinogenesis are observed in Lynch syndrome, enabling studies on the competition between genetic alterations during progression of normal cells to cancer. One of the most striking advantages related to focusing on Lynch syndrome as a model for carcinogenesis stems from the fact that the mechanism driving carcinogenesis is known and linked to a deficiency of the MMR system.

- ▷ **The same mutations at coding microsatellites leading to carcinogenic transformation are the ones that give rise to highly immunogenic cancer antigens, called frameshift peptide neoantigens.** The insertion/deletion mutations lead to a shift of the translational reading frame. This means that all subsequent base triplets, which are then translated into one component of the corresponding protein, are completely novel, and therefore the protein is highly immunogenic [122]. The likelihood of frameshift peptide neoantigen recognition by the immune system (Figure 2.5b) can be modeled mathematically and then used to design adequate prevention approaches for detecting and preventing the accumulation of cells containing those mutations and neoantigens. Modeling immunoediting is only possible because of the presence of predictable, recurrent neoantigens characteristic of Lynch syndrome-associated tumors.

[122]: Linnebacher et al. (2001), “Frameshift peptide-derived T-cell epitopes: A source of novel tumor-specific antigens”.

### 3.3 STATE-OF-THE-ART COMPUTATIONAL MODELING AT THE CELL LEVEL

We start with an overview of state-of-the-art computational modeling at the cell level which will serve as the foundation for the developed computational model for Lynch syndrome colorectal carcinogenesis at the cell level (see Chapter 6). The descriptions here heavily rely on those in [81]. In general, the mathematical approaches used to model cell populations can be broadly divided into three categories: 1) *Spatial models*, which take into account the specific location of individual cells or the location of a population of cells, 2) *compartmental*

[81]: Haupt et al. (2021), “A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis”.

[95]: Johnston (2008), “Mathematical modelling of cell population dynamics in the colonic crypt with application to colorectal cancer”.

[116]: Leeuwen et al. (2006), “Crypt dynamics and colorectal cancer: advances in mathematical modelling”.

[126]: Lowengrub et al. (2009), “Nonlinear modelling of cancer: bridging the gap between cells and tumours”.

[128]: Matteis et al. (2012), “A review of spatial computational models for multi-cellular systems, with regard to intestinal crypts and colorectal cancer development”.

[134]: Metzcar et al. (2019), “A Review of Cell-Based Computational Modeling in Cancer Biology”.

[156]: Osborne et al. (2017), “Comparing individual-based approaches to modelling the self-organization of multicellular tissues”.

*models*, which describe the transition between cell types, irrespective of their position within the population, and 3) non-spatial *stochastic models*, a more general class of models, all involving stochasticity as a main feature. For detailed reviews, we refer to [95, 116, 126, 128, 134, 156]. It is important to note that the models are distinguished based on their basic setup, not on the mathematical tools they use. For instance, one can couple a setup with ordinary (ODEs) or partial differential equations (PDEs) as well as stochastic processes. This gives rise to a *system*, whereby the methods are used to describe the *state* of the system.

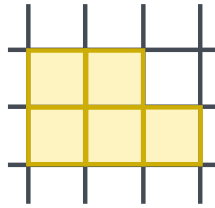
As our focus is on spatial models for the description of intracrypt dynamics at early stages of cancer development in Lynch syndrome, we will explain the main ideas of such an approach following the reviews mentioned above. We would like to point out that we are referring to cell-based spatial models in the following using a discrete spatial representation of cells rather than a continuous representation by, e.g., PDE-based approaches.

**Cell-based spatial models** treat cells as discrete entities bearing specific characteristics as internal states which change in discrete time steps. At each time step, these internal states are updated according to certain rules (usually governed by equations), such that the whole state of the system is recomputed. Cell-based spatial models can be further divided into two subcategories: *in-lattice models* and *off-lattice models*, see Figure 3.1. The main structural difference lies in whether or not the cells are assumed to be positioned on a rigid grid.

In off-lattice models, cells are loosely located in space, yielding cell shapes which are more biologically realistic. Models within this class are further distinguished depending on how exactly cells are labeled within the underlying space, such that we can track their position over time. *Vertex models* represent cells as polygons and the vertices of the cells are tracked in space and time, while *overlapping spheres models* and *Voronoi tessellation models* describe cells through the position of their nuclei and the nuclei are tracked in space and time. OS models regard cells as spheres with a certain time-dependent radius. Voronoi tessellations give rise to a cell body consisting of all points in space whose distance to the nucleus is less than or equal to their distance to any other nucleus in the cell population. Each model is equipped with a unique representation of cell division. Voronoi tessellation models have

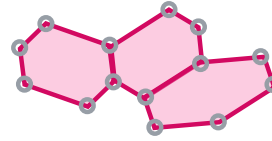
## CELL-BASED SPATIAL MODELS

## In-lattice models

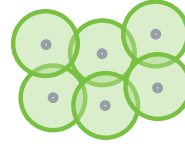


## Off-lattice models

vertex models



overlapping spheres



Voronoi tessellation



**Figure 3.1: Overview of some spatial cell-based models.** Spatial cell-based models can be divided into in-lattice (right) and off-lattice models (left). Using in-lattice models, cells are assumed to be positioned on a rigid grid, where for off-lattice models, cells are loosely located in space. Common examples for off-lattice models are vertex models (top right), overlapping spheres (middle right), and Voronoi tessellation models (bottom right). In vertex models, cells are defined as polygons and labeled by their vertices. For overlapping spheres models, cells are represented as spheres with changing radii, which interact whenever there is an overlapping. Further, using Voronoi tessellation models, cells are defined as polygons as in vertex models. However, they are labeled by their centers which also define the shape. Proximate cell centers are connected via the Delaunay triangulation. We will use the latter modeling approach for our model of colorectal cancer development at the cell level.

been used extensively to model cellular migration [132], or in multi-scale models [145], which used Voronoi tessellation to model the tissue architecture of stratified epithelium.

Not all modeling approaches which concern intestinal crypts can clearly be assigned to exactly one of the three major classes. Many models concentrate on single aspects rather than cell populations and can further be used in multi-scale models.

We describe the aspects which are in our understanding fundamental for modeling intra-crypt dynamics at early stages of cancer development in Lynch syndrome, and explain recent modeling attempts.

- **Measuring monoclonal conversion of different mutations.** In order to obtain temporal estimates for the duration of cancer development, it is essential to quantitatively analyze the process of monoclonal conversion (also called mutation fixation) in colonic crypts. To the best of our knowledge, there are two approaches, which investigate this aspect [8, 9, 66]. Both focus on the process of monoclonal conversion, using multi-scale

[132]: Meineke et al. (2001), “Cell migration and organization in the intestinal crypt using a lattice-free model”.

[145]: Morel et al. (2001), “A Proliferation Control Network Model: The Simulation of Two-Dimensional Epithelial Homeostasis”.

[8]: Araujo et al. (2018), “Testing three hypotheses of the contribution of geometry and migration dynamics to intestine crypt evolution”.

[9]: Araujo et al. (2019), “Investigating the Origins of Cancer in the Intestinal Crypt with a Gene Network Agent Based Hybrid Model”.

[66]: Fletcher et al. (2012), “Mathematical modeling of monoclonal conversion in the colonic crypt”.

[66]: Fletcher et al. (2012), “Mathematical modeling of monoclonal conversion in the colonic crypt”.

[8]: Araujo et al. (2018), “Testing three hypotheses of the contribution of geometry and migration dynamics to intestine crypt evolution”.

[9]: Araujo et al. (2019), “Investigating the Origins of Cancer in the Intestinal Crypt with a Gene Network Agent Based Hybrid Model”.

[132]: Meineke et al. (2001), “Cell migration and organization in the intestinal crypt using a lattice-free model”.

[42]: Buske et al. (2011), “A Comprehensive Model of the Spatio-Temporal Stem Cell and Tissue Organisation in the Intestinal Crypt”.

models based on spatial approaches. In order to obtain a comprehensive understanding of these processes, various analyses are necessary which also account for different types of mutations.

- ▶ **Choosing an appropriate geometric modeling framework.** While Fletcher et al. [66] uses a Voronoi tessellation model, the work by Araujo et al. [8, 9] is based on a cellular automaton model, an example of an in-lattice model, rather than an off-lattice model, where the latter entails limitations regarding the cells’ geometric representation. Further, Fletcher et al. [66] represents the crypt in a simplified way as a two-dimensional surface of revolution. The assumptions on crypt size and cell cycle lengths are chosen in such a way to be suitable for application to mouse data. However, human crypts are larger in size leading to more cells and thus operations which are necessary to compute in every time step. Thus, choosing a geometric modeling framework which is computationally feasible such that it can deal with many numbers of cells is desired.
- ▶ **Incorporating cell migration.** Cell migration is a fundamental process in colonic crypts ensuring the integrity and constant renewal of the tissue. Meineke et al. [132] developed a corresponding model which is currently used in many approaches [66] and should be used in the future.
- ▶ **Incorporating gene dependencies and interactions.** As the individual driver genes and their interactions are essential for the overall selection process taking place during cancer development, these should be considered in each model. Araujo et al. incorporate various gene-gene interactions in great detail via the use of a gene regulatory network. However, the implementation of sporadic mutations is lacking. In order to allow for computational feasibility, it is also here important to focus on the most relevant genetic events in our case tailored for Lynch syndrome carcinogenesis.
- ▶ **Implementing the role of the Wnt pathway.** As described later in more detail 2.1.3, the Wnt pathway is assumed to be one of the key determinants for cell differentiation and thus should be addressed in these kind of modeling attempts. This assumption is represented in the model by Araujo et al. as well as in modeling approaches by Buske et al. [42] and De Mat-



teis et al. [128]. Further, [42] take into account that the colonic crypt has a cap, and by this, assume that Wnt activity is determined by the local curvature of the basal membrane. This increases the modeling complexity and computational costs of the solving process. Finding a sweet spot between accurate description of the underlying processes and computational feasibility is key for these cell-based computational models.

### 3.4 STATE-OF-THE-ART MATHEMATICAL MODELING AT THE CRYPT LEVEL

Next, we want to highlight two recent publications describing colorectal carcinogenesis at the crypt level. This means that individual crypts form the basic unit of interest, thus modeling a larger scale than before. Both models are based on the concept of multi-step carcinogenesis and will serve as related work for the developed mathematical model at the crypt level in Chapter 7. The development follows partly descriptions in [82].

The first paper by Paterson et al. [160] presents a model for quantifying the evolutionary dynamics of sporadic microsatellite-stable colorectal cancer initiation and progression based on describing the occurrence of key driver mutations. Those represent the classical adenoma-carcinoma sequence of mutations in *APC*, *KRAS*, and *TP53*. By defining mutational graphs for each of the genes, considering *APC* and *TP53* as classical tumor suppressor genes and *KRAS* as a classical oncogene. By allowing mutations to occur in any order, the authors obtain a network of possible evolutionary pathways leading to cancer. The mutational graphs are built using a general approach of gene-specific numbers of driver positions and by assuming *APC* and *KRAS* provide fitness advantage but not *TP53*. The latter assumption is based on several independent studies [15, 118, 151]. Based on these assumptions, a stochastic model is developed and parametrized using recent experimental data [160]. The model predictions are compared to the reported lifetime colorectal cancer risk.

For hereditary colorectal cancers, in particular, Komarova et al [109, 111] proposed a model for the occurrence and

[128]: Matteis et al. (2012), “A review of spatial computational models for multi-cellular systems, with regard to intestinal crypts and colorectal cancer development”.

[42]: Buske et al. (2011), “A Comprehensive Model of the Spatio-Temporal Stem Cell and Tissue Organisation in the Intestinal Crypt”.

[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

[15]: Baker et al. (2014), “Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon”.

[118]: Leeuwen (2007), “Towards a multiscale model of colorectal cancer”.

[151]: Nicholson et al. (2018), “Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium”.

[109]: Komarova et al. (2003), “Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability”.

[111]: Komarova et al. (2002), “Dynamics of Genetic Instability in Sporadic and Familial Colorectal Cancer”.

ordering of key events during carcinogenesis based on ordinary differential equations. In particular, it addresses the question of the extent of genetic instability, namely chromosomal instability and microsatellite instability, as an early event in carcinogenesis. Therefore, the transition between normal cells and chromosomal or microsatellite unstable cells is described by mutation-selection networks assuming certain baseline mutation rates for the different cell status. By changing some of the cell status and parameters, different scenarios like tumor initiation in the sporadic case or in FAP and Lynch syndrome are modeled.

While these approaches serve as valuable starting points for modeling colorectal carcinogenesis in some settings, in our understanding, some key aspects have to be addressed in future models:

► **Incorporating gene dependencies and interactions.**

In Komarova et al. [111], mutation rates are considered depending on the general cell status (normal, chromosomal or microsatellite unstable). In Paterson et al. [160], the mutation rates are determined in a gene-dependent way which is an important aspect for analyzing the influence of the respective gene on cancer development. However, in contrast to the current approaches, there are dependencies reported in the literature between some of the mutational events, as an alteration in one gene may lead to an increase or decrease of alterations in another gene. For example, a non-functioning mismatch repair gene (MMR deficiency) leads to a generally increased point mutation rate which may lead to loss-of-function mutations in tumor suppressor genes like *APC* or *TP53*. Future mathematical models should thus incorporate gene dependencies and gene interactions.

► **Fitting simulations to age-dependent data.** Both approaches [111, 160] compare the simulation results to *in vivo* single data points, mostly to available patient measurements at 70 years of age, e.g., comparing the simulated cancer risk in [160] to reported life-time colorectal cancer risks. However, an age-dependent fitting of the parameters or comparison to human data for the whole life of a patient is currently lacking. To obtain an age-dependent picture of cancer development, both more clinical and biomedical data as well

[111]: Komarova et al. (2002), “Dynamics of Genetic Instability in Sporadic and Familial Colorectal Cancer”.

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

as appropriate models are needed such that mathematical parameter learning and sensitivity analysis are possible.

- ▶ **Modularity of the modeling approach.** The approach by Paterson et al. [160] is designed to model the classical case of sporadic microsatellite-stable colorectal cancer development meaning that the presented equations only fit to this specific scenario and a possible adaptation to other scenarios is not mentioned. In Komarova et al. [111], different scenarios of colorectal cancer development are addressed. However, as the knowledge on cancer heterogeneity and on different pathways of carcinogenesis is constantly increasing, a mathematical description of modular components that allow to easily adapt the model to other scenarios is desirable and should be addressed in future modeling approaches.
- ▶ **Mathematically rigorous analysis.** In general, every modeler should aim for a mathematically rigorous description and analysis of the model. This makes it possible to explore different scenarios with varying parameters and initial conditions to obtain an overview of what might and might not be compatible with life. It further opens the possibility for future analyses of different treatment and prevention effects on cancer development.
- ▶ **Computational feasibility.** Another general aim is computational feasibility of the model in order to allow for long-term predictions and the ability to analyze different scenarios. Therefore, the model should be chosen as simple as possible with exploiting the full range of numerical schemes for an efficient solving process.

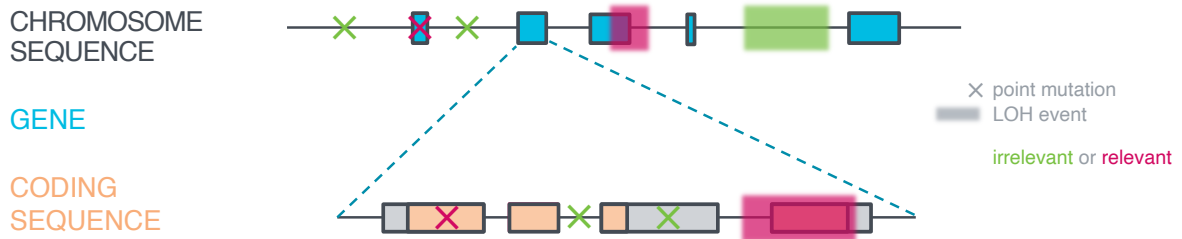
[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

[111]: Komarova et al. (2002), “Dynamics of Genetic Instability in Sporadic and Familial Colorectal Cancer”.



**MODELING LYNCH SYNDROME AT THE DNA  
LEVEL**





## 4 PARAMETRIZING MUTATION RATES IN A GENE-DEPENDENT WAY

In this chapter, we present an approach for defining and parametrizing mutation rates for two common types of mutations namely point mutations and chromosomal changes leading to loss of heterozygosity events in a gene-dependent way. These gene-dependent mutation rates will serve as a basis for both models of colorectal cancer development on the cell and crypt levels later explained in Chapter 6 and Chapter 7. The following chapter heavily relies on [81, 82].

As depicted in Section 2.1.2, mutations can occur over the whole genome during DNA replication. A first educated guess is that those mutations are more or less randomly distributed over the whole genome. It follows that the probability of a specific gene becoming mutated is depending on its length, whereby this assumption is made in many modeling approaches [15, 118, 151, 160]. The length of a region considered as relevant depends on the type of mutational event: As explained previously, we assume that point mutations are only relevant for cancer development if they occur in regions that give rise to a phenotypical change. Often, this refers to (parts of) protein-encoding regions, called *coding regions* or *exons*, and we will consider these corresponding region lengths for point mutations. LOH events refer to the loss of some region in one copy of the diploid genome independent of the function of the gene region. Thus, we consider the coding and non-coding regions, i.e. exons and introns, of genes as possible targets of LOH events. Their length is given by the actual gene length. A schematic illustration is given in Figure 4.1.

4.1 POINT MUTATIONS IN GENE HOTSPOTS . . . . .	62
4.2 LOH EVENTS IN WHOLE GENES . . . . .	64

[81]: Haupt et al. (2021), “A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis”.

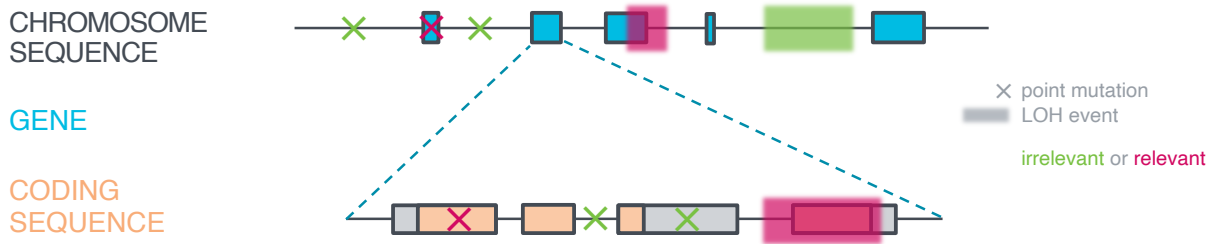
[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

[15]: Baker et al. (2014), “Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon”.

[118]: Leeuwen (2007), “Towards a multiscale model of colorectal cancer”.

[151]: Nicholson et al. (2018), “Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium”.

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.



**Figure 4.1:** Schematic illustration of genes located on chromosomes affected by point mutations and LOH events. *Bottom:* A gene usually consists of coding sequences and non-coding sequences. *Top:* Several genes form a chromosome sequence (top). All parts can be affected by LOH events (semi-transparent colored boxes) or point mutations (×) whereby we differentiate between alterations that are irrelevant or relevant for cancer development. Relevant point mutations (×) appear in coding sequences and LOH events are relevant (semi-transparent boxes) if they affect any part of a considered gene as we assume that all detectable LOH events are large enough such that the affected gene becomes inactivated.

## 4.1 RELEVANT POINT MUTATION RATES DEPENDING ON GENE HOTSPOT LENGTHS

We want to model the rate  $\pi_{\text{pt}}(\text{gene})$  of relevant point mutations in a specific gene for each cell and each cell division. During each cell division, we assume to accumulate  $n_{\text{pt}}$  point mutations in one cell. The mutations are assumed to be uniformly distributed over the base pairs on the entire genome, where there are  $n_{\text{bp, genome}}$  base pairs on the genome. As explained above, we only focus on the point mutations relevant for cancer development as we are interested in which genes and mutations thereof drive the cancer development process. Those are the point mutations which occur in regions that give rise to a phenotypical change. We call those regions hotspots. The length of the hotspots  $n_{\text{hs}}(\text{gene})$  is gene dependent. Further, there might be more than one relevant point mutation on a gene. However, we assume that there is no significant additional effect on the phenotype in case of multiple relevant point mutations. Besides that, we assume that the two copies of the genome are independent of each other, meaning that a mutation in one allele does not influence the mutation probability of the second allele. Thus, the point mutation rate  $\pi_{\text{pt}}(\text{gene})$  is twice as large if there is no mutated allele ( $n_{\text{mut}}(\text{gene}) = 0$ ) compared to the state where one allele is already mutated ( $n_{\text{mut}}(\text{gene}) = 1$ ).



**Definition 4.1** Rate of relevant point mutations

Under the above assumptions, the rate of a relevant point mutation per cell division for a specific gene of a cell is given by

$$\pi_{\text{pt}}(\text{gene}) = n_{\text{pt}} \frac{n_{\text{hs}}(\text{gene})}{n_{\text{bp,genome}}} \cdot \left(1 - \frac{1}{2} n_{\text{mut}}(\text{gene})\right) \quad (4.1)$$

We note that the number of base pairs on the genome  $n_{\text{bp,genome}}$  is by definition independent of the considered gene and only depends on the species we are looking at. Further, the total relevant number of point mutations per cell division  $n_{\text{pt}}$  is gene-independent. However, it may depend on the species and organ of interest, as well as the cell type of cancer origin. The number of mutated alleles  $n_{\text{mut}}(\text{gene}) \in \{0, 1, 2\}$  is a categorical variable and depends on the current mutational state of the considered gene.

The important parameter which has to be determined for each gene in an organ- and species-specific context is the length of the hotspots  $n_{\text{hs}}(\text{gene})$  measured in base pairs. For some genes, only a very few defined base pairs have been found to become usually mutated during cancer development in one specific organ and species. In this case, we propose to use this specific number of base pairs for an estimate of  $n_{\text{hs}}(\text{gene})$ . In Lynch syndrome-associated colorectal carcinogenesis, this is true for *CTNNB1* with 5 relevant mutations on 12 base pairs [1], and for the oncogene *KRAS* with 7 relevant mutations according to [2].

However, for other genes, such specific mutation hotspots have not been identified thus far, since the mutations appear to be rather uniformly distributed over the whole genome. In this case, the full coding sequence length of the considered gene could be used as an estimate for  $n_{\text{hs}}(\text{gene})$ . An appropriate choice in this context is to use the reference sequence database at NCBI for coding sequence lengths [155]. We did so for the two considered MMR genes *MLH1* ( $n_{\text{hs}}(\text{MLH1}) = 2,270$ ) and *MSH2* ( $n_{\text{hs}}(\text{MLH1}) = 2,800$ ), as well as for the tumor suppressor gene *TP53* ( $n_{\text{hs}}(\text{TP53}) = 1,180$ ).

If specific information is lacking, mutation data from publicly available databases, like the DFCI database using the cBioPortal website [44, 69], could be used to obtain estimates for the hotspot length. In our setting, we make use of this data source of about 4,000 colorectal cancer samples for the

[1]: Ahadova et al. (2016), “CTNNB1-mutant colorectal carcinomas with immediate invasive growth: a model of interval cancers in Lynch syndrome”.

[2]: Ahadova et al. (2018), “Three molecular pathways model colorectal carcinogenesis in Lynch syndrome”.

[155]: O’Leary et al. (2015), “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”.

[44]: Cerami et al. (2012), “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1.”

[69]: Gao et al. (2013), “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal”.

tumor suppressor gene *APC* to identify approximately 2,400 hotspot base pairs.

## 4.2 RELEVANT LOH EVENT RATES DEPENDING ON WHOLE GENE LENGTHS

Similar to the rate of relevant point mutations, we want to model the rate of relevant LOH events per cell and per cell division. Here, we assume that all detectable LOH events are large enough such that the affected gene becomes inactivated. This means that if a gene is hit by an LOH event, then a coding sequence (exon) is lost, the gene is hit in a region relevant for cancer development and the gene therefore becomes inactivated. Again, we assume that multiple LOH events on the same gene allele have the same functional consequences as a single LOH event. Thus, also in this context, the probability of an LOH event  $p_{\text{LOH}}(\text{gene})$  for a given gene is proportional to its length, whereby we denote the full gene length including introns and exons by  $n_{\text{bp}}(\text{gene})$ .

### Definition 4.2 Rate of relevant LOH events

*Under the above stated assumption, the rate per cell division of a relevant LOH event for a specific gene of a cell per cell division is given by*

$$p_{\text{LOH}}(\text{gene}) = \left(1 - \frac{1}{2}n_{\text{mut}}(\text{gene})\right) \alpha n_{\text{bp}}(\text{gene}) \quad (4.2)$$

*with a parameter  $\alpha \in \mathbb{R}_{>0}$  independent of the considered gene which has to be estimated.*

We note that in contrast to the length of the hotspots  $n_{\text{hs}}(\text{gene})$ , the full gene length  $n_{\text{bp}}(\text{gene})$  is independent of the considered organ. However, it depends on the considered species. Further, in the context of colorectal carcinogenesis in Lynch syndrome, we assume that LOH events cannot affect oncogenes like *KRAS* as oncogenes typically need an activating point mutation for a phenotypic change. Thus, only the genes where usually two hits are required for an inactivation, can be affected by LOH events. A reasonable source for full gene lengths is again given by the reference sequence database at NCBI [155].

[155]: O’Leary et al. (2015), “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”.

gene	$n_{\text{bp}}(\text{gene})$
<i>MLH1</i>	57,500
<i>MSH2</i>	80,000
<i>CTNNB1</i>	41,000
<i>APC</i>	139,000
<i>TP53</i>	19,200

**Table 4.1:** Estimates for the full gene lengths  $n_{\text{bp}}(\text{gene})$  of the genes *MLH1*, *MSH2*, *CTNNB1*, *APC*, and *TP53*, as used in our models for Lynch syndrome colorectal carcinogenesis. Those estimates are necessary for the computation of the rates of relevant LOH events for the individual genes. They are based on the reference sequence database at NCBI [155]. Table reprinted from [82].

As an example, for the models developed within this Ph.D. project, we used the estimates for the full gene lengths  $n_{\text{bp}}(\text{gene})$  of the genes *MLH1*, *MSH2*, *CTNNB1*, *APC*, and *TP53* as available from the database at NCBI [155] (see Table 4.1).

To determine the parameter  $\alpha$ , medical knowledge on the relative proportion of point mutations and LOH events during cancer development could be useful. As direct time-dependent *in vivo* measures are hardly feasible in humans, we use a snapshot in time of tumor sequencing data for specific genes to estimate this proportion constant  $\alpha$ .

For the example of Lynch syndrome colorectal carcinogenesis, we use available data for *MLH1* suggesting that inactivation via LOH events is twice as likely to occur than via point mutations [164]. In formulas, we thus assume

$$p_{\text{LOH}}(\text{MLH1}) = 2 \cdot \pi_{\text{pt}}(\text{MLH1}). \quad (4.3)$$

Using the derived equations for the rates of relevant point mutations and LOH events, we obtain

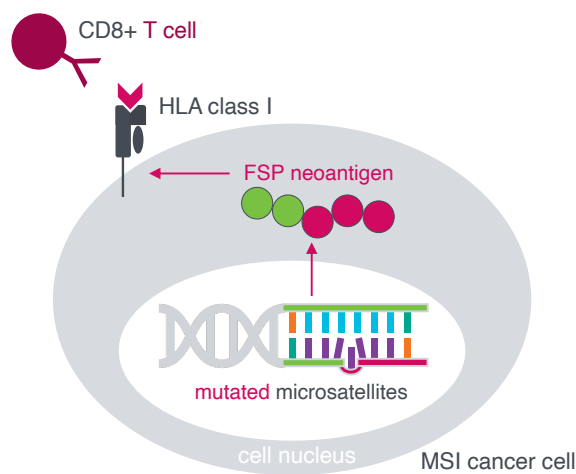
$$\alpha = 2 \frac{n_{\text{hs}}(\text{MLH1})}{n_{\text{bp}}(\text{MLH1})} \frac{n_{\text{pt}}}{n_{\text{bp, genome}}} \quad (4.4)$$

which is used as an estimate in our models of Lynch syndrome colorectal carcinogenesis (see Chapter 6 and Chapter 7) and could be further refined as soon as more detailed data become available.

[155]: O’Leary et al. (2015), “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”.

[164]: Porkka et al. (2017), “Sequencing of Lynch syndrome tumors reveals the importance of epigenetic alterations”.





## 5 QUANTIFYING HLA TYPE-DEPENDENT IMMUNO-EDITING IN CANCER

In this chapter, we will focus on quantifying immuno-editing during the development of microsatellite unstable colorectal cancer in general including the Lynch syndrome case. Hereby, we first analyze the tumor-immune interactions in a general way followed by a detailed exploration of the influence of the HLA system on these processes.

As shown in Figure 2.5b and in Section 2.4, MMR-deficient cancers accumulate an exceptionally high load of insertion/deletion (indel) mutations at coding microsatellites. Those indel mutations affecting coding microsatellites in genomic regions encoding tumor-suppressor genes are considered major drivers of microsatellite unstable carcinogenesis. At the same time, these indel mutations lead to a shift of the translational reading frame, generating unique frameshift peptides, a major source of neoantigens [101, 174]. This makes the cancer cells with a high load of mutation-induced neoantigens recognizable and attackable for the immune system. For the recognition of neoantigens by the immune system, processing through the cellular antigen machinery and presentation by human leukocyte antigen (HLA) class I molecules on the tumor cell surface are essential prerequisites [17]. These HLA class I molecules consist of a heavy chain and a non-covalently bound light chain where the latter is encoded by the *Beta-2-microglobulin (B2M)* gene [17] (see Section 2.4). Further, the likelihood of HLA binding for a defined peptide depends on the HLA genotype, as every individual harbors six alleles (HLA-A, HLA-B, HLA-C, two alleles each) that encode for HLA class I heavy chains [17, 96].

<b>5.1 FRAMESHIFT MUTATION</b>	
LANDSCAPE . . . . .	69
<b>5.2 FRAMESHIFT PEPTIDE</b>	
LANDSCAPE . . . . .	73
<b>5.3 HLA TYPE INFLUENCE</b>	76

[101]: Kloor and von Knebel Doeberitz (2016), “The Immune Biology of Microsatellite-Unstable Cancer”.

[174]: Schwitalle et al. (2008), “Immune Response Against Frameshift-Induced Neopeptides in HNPCC Patients and Healthy HNPCC Mutation Carriers”.

[17]: Ballhausen et al. (2020), “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution”.

[17]: Ballhausen et al. (2020), "The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution".

During this Ph.D. time, a new algorithm, called ReFrame [17], was developed at ATB Heidelberg to quantitatively detect microsatellite indel mutations with high sensitivity in a collection of tumors. Using ReFrame, mutations that are shared by most microsatellite unstable colorectal and endometrial cancers respectively are identified. This was the foundation to study the concept of immunoediting during cancer evolution in microsatellite unstable colorectal cancers. For the algorithm development and the subsequent analyses, a collaborative study was initiated and coordinated by our collaboration partners at ATB, Heidelberg University Hospital, with our contributions representing EMCL and HITS, with groups at the DKFZ Heidelberg, as well as international collaboration partners. The results are published in Nature Communications [17], where the following presentation is based on.

All study results rely on a dataset of frameshift mutation frequencies in 41 coding microsatellites residing in 40 target genes among 139 microsatellite unstable colorectal cancers and 28 microsatellite unstable endometrial cancers which found the basis of the subsequent analyses. We discovered a negative correlation between the frameshift mutations in MMR-deficient colorectal cancers and the predicted immunogenicity of the resulting frameshift peptides, that is their possibility to provoke an immune response. Further, this correlation is absent in *B2M*-mutated tumors, and dependent on the HLA type. The overall study strongly supports the concept of continuous immunoediting during cancer development and provides new evidence for the hypothesis that immunogenic cancers and precancerous cell clones can be attacked and potentially eradicated by the immune system. In a translational context, these findings underline the potential of neoantigen-based cancer-preventive vaccines that may in the future help to reduce tumor risk in Lynch syndrome individuals.

To make future detailed analyses possible about the influence of the HLA system as a central component of the antigen presentation machinery on the previously described negative correlation between mutation frequencies and the predicted immunogenicity, a second laboratory study was performed at ATB Heidelberg, with our mathematical support for a rigorous data analysis. The aim of the study was to develop a laboratory procedure to determine the HLA type from

formalin-fixed paraffin-embedded (FFPE) tissue as a great source of archival and historic tissue samples made accessible for molecular biological studies. In those samples, the HLA type is usually difficult to analyze due to fragmentation of DNA, hindering the application of commonly used assays that require long DNA stretches. The refined approach developed at ATB Heidelberg with our mathematical support, including validation data and application to the above mentioned Ballhausen et al. [17] microsatellite unstable colorectal cancer data is published in [210].

## 5.1 QUANTIFYING THE LANDSCAPE OF FRAMESHIFT MUTATIONS USING ReFRAME

In this section, we focus on the data analysis performed in Ballhausen et al. [17] for the identification and quantification of frameshift mutations in coding microsatellites. This serves as the basis for the subsequent analyses in this chapter. As a first step, we have to understand the data generation process. This understanding helps to build a mathematical tool, the ReFrame algorithm, to quantify the frameshift mutations.

In general, short-read next-generation sequencing approaches are not ideally suited for frameshift mutation analyses of MSI cancers [17, 147, 192, 195] because, in particular, long coding microsatellites with a high number of repeats which are often affected by mutations during MSI carcinogenesis are missed by these approaches. Thus, the current gold standard for the detection of these mutations and hence, for the detection of MSI is fragment length analysis which is also used at ATB Heidelberg. Fragment length analysis is a genetic analysis method which consists of four general steps: DNA extraction, polymerase chain reaction (PCR) amplification, capillary electrophoresis, and data analysis. In general, PCR amplification of microsatellite loci with a specific length generates fragments of different lengths around the true length because either indel mutations in MMR-deficient cells or polymerase slippage events can occur. The resulting patterns of the number of fragments of different lengths than the original microsatellite length are called stutter band artifacts.

[17]: Ballhausen et al. (2020), “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution”.

[210]: Witt et al. (2022), “A simple approach for detecting *HLA-A\*02* alleles in archival formalin-fixed paraffin-embedded tissue samples and an application example for studying cancer immunoediting”.

[147]: Nakano et al. (2017), “Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area”.

[192]: Treangen and Salzberg (2011), “Repetitive DNA and next-generation sequencing: computational challenges and solutions”.

[195]: Vanderwalde et al. (2018), “Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11, 348 patients”.

They cause overlays of peak patterns and hamper data interpretation making it hard to determine the true length distribution of the underlying sample. Thus, the aim is to build an algorithm that computes the true peak distribution prior to PCR amplification based on the observed peak distribution measured after PCR amplification. ReFrame is developed at the ATB Heidelberg to exactly reach this goal to allow quantitative analysis of microsatellite mutations by removing the stutter band artifacts.

It can be assumed that the process of PCR amplification, if performed correctly according to current standard procedures, is a linear operator in the mathematical sense because the PCR amplification of two microsatellites combined in one tube equals the PCR amplification of the individual microsatellites in separate tubes. Further, starting with a certain proportion of the usual amount of DNA strands within a tube, leads to exactly this same proportion of the PCR product after PCR amplification. In mathematical terms, it holds for the process of PCR amplification for the amounts of microsatellites  $A$  and  $B$

$$\begin{aligned}\text{PCR}(A + B) &= \text{PCR}(A) + \text{PCR}(B), \\ \text{PCR}(\lambda A) &= \lambda \text{PCR}(A), \quad \lambda \in \mathbb{R}_+.\end{aligned}$$

We will represent this process by a matrix  $C^L$  mapping the true length distribution  $p_{\text{true}}$  in the sample to the observed length distribution  $p_{\text{observed}}$  including stutter band artifacts. The whole process depends on the microsatellite length  $L$ . We want to solve the following optimization problem with constraints

$$\begin{aligned}\min_{p_{\text{true}}} & \|C^L p_{\text{true}} - p_{\text{observed}}\|_2^2 \\ \text{s.t.} & \langle \mathbb{1}, p_{\text{true}} \rangle = 1, \\ & 0 \leq p_{\text{true}} \leq 1.\end{aligned}$$

The optimization problem is solved in  $\mathbb{R}$  using a quadratic solver with an active-set method for solving quadratic programming problems.

In the following, we describe how  $C^L$  is constructed. First of all, only a certain amount of possible microsatellite lengths  $\Delta \in \{-4, -3, \dots, 4\}$  around the main peak, denoted by  $\Delta = 0$ ,



are considered. Thus,  $C^L \in [0, 1]^{9 \times 9}$ . This assumption also leads to  $C^L$  being a band matrix with bandwidth 4.

It is known from the analyses in SelTarBase that one determining factor for the mutation rate in microsatellites is their length [212]. In Ballhausen et al. [17], main peak fractions of different microsatellites are obtained with the corresponding microsatellite length  $L$  and a logistic function, called  $p(L)$ , is fitted to these data. It describes the microsatellite length contribution on the peak distribution. Further, reference relative peak heights  $p_{\text{ref}}$  are used for each microsatellite of interest which are computed by taking the median of relative peak heights from each microsatellite locus in MMR-proficient control samples. This is a measure for the baseline stutter band distribution occurring during PCR amplification from normal tissue samples.

Those two measurements are used to calculate the diagonal entries of the matrix  $C^L$ , that is the proportions of length-specific relative peak heights that do not change during PCR amplification which is called the effective length  $p_{\text{effective}}$ . To calculate these values for each possible length  $L$  of the considered coding microsatellite, the reference value of the main peak  $p_{\text{ref}}(0)$  is inserted into the inverse logistic fit function, shifted by the considered length shifts  $\Delta \in \{-4, -3, \dots, 4\}$ , and inserted into the logistic function again. In formulas, this reads

$$C_{\Delta\Delta}^L = p_{\text{effective}}(\Delta) = p(p^{-1}(p_{\text{ref}}(0)) + \Delta).$$

For the off-diagonal elements, it is ensured that  $C^L$  is a band matrix by using the indicator function  $\mathbb{1}_{\leq x}$  defined by

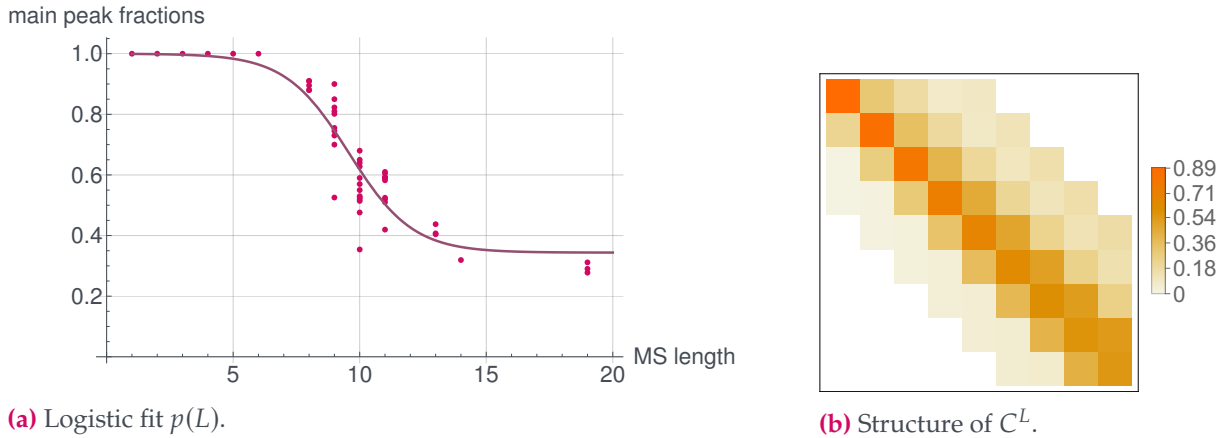
$$\mathbb{1}_{\leq x}(y) = \begin{cases} 1, & \text{if } |y| \leq x, \\ 0, & \text{else,} \end{cases}$$

with setting  $x = 4$ . The off-diagonal matrix elements  $C_{\Delta'\Delta}^L$ ,  $\Delta' \neq \Delta$  represent the proportions of length-specific relative peak heights that change during PCR amplification. They can be computed as follows for  $\Delta' \neq \Delta$

$$C_{\Delta'\Delta}^L = \mathbb{1}_{\leq 4}(\Delta' - \Delta) \cdot \frac{p_{\text{ref}}(\Delta' - \Delta)}{1 - p_{\text{ref}}(0)} \cdot (1 - p_{\text{effective}}(\Delta)).$$

[212]: Woerner et al. (2003), "Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes".

[17]: Ballhausen et al. (2020), "The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoeediting during tumor evolution".



**Figure 5.1: Components of the ReFrame algorithm.** (a) The data points are main peak fractions for varying microsatellite lengths  $L$  of different microsatellites to which a logistic function  $p(L)$  is fitted. (b) The matrix  $C^L$  is a band matrix with bandwidth 4. Further, the diagonal entries become smaller with increasing length resembling the length-dependent proportion of the mutation probabilities. Data from GitHub [github.com/atb-data/neoantigen-landscape-msi](https://github.com/atb-data/neoantigen-landscape-msi).

The expression for the off-diagonal matrix elements can be explained in the following way:

- ▶ As stated above,  $\mathbb{1}_{\leq 4}$  ensures the band structure of the matrix with bandwidth 4.
- ▶ The factor  $1 - p_{\text{effective}}(\Delta)$  is the proportion that remains for the off-diagonal elements as the diagonal elements proportion is  $p_{\text{effective}}(\Delta)$ .
- ▶ This proportion now has to be distributed among the shifts according to what is expected from the normal samples, that is the reference relative peak heights  $p_{\text{ref}}(\Delta' - \Delta)$  which are normalized to the corresponding proportion  $1 - p_{\text{ref}}(0)$  that remains when already considering the reference relative peak height  $p_{\text{ref}}(0)$  of the main peak.

[17]: Ballhausen et al. (2020), “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution”.

[214]: Woerner et al. (2009), “SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology”.

In Ballhausen et al. [17], ReFrame is used in a series of MSI colorectal cancers ( $n = 139$ ) to quantify the mutation frequency for mutations in 41 coding microsatellites residing in 40 target genes derived from SelTarBase, version 201307 [214]. By doing so, a large set of coding microsatellite mutations is found that is shared by the majority of MSI colorectal cancer samples.

Further, ReFrame is used to distinguish indel mutation types, which is crucial for the prediction of the frame of the resulting frameshift peptides. Here, we distinguish between M1 frameshifts (deletions of one nucleotide, m1, or insertions of two nucleotides, p2) and M2 frameshifts (deletions of

two nucleotides, m2, or insertions of one nucleotide, p1). The M1/M2 distribution is analyzed showing significantly different patterns across different coding microsatellites.

## 5.2 QUANTIFYING THE LANDSCAPE OF FRAMESHIFT PEPTIDES USING IMMUNOLOGICAL SCORES

Using NetMHCpan 4.0, a state-of-the-art MHC ligand prediction tool based on artificial neural networks, neopeptides are predicted that are possibly presented as epitopes by HLA class I antigens encoded by the most important HLA super-types [17, 96, 168, 171]. The predicted epitopes are subdivided into three classes based on commonly accepted thresholds: high-affinity binders ( $IC_{50} < 50$  nM), low-affinity binders ( $50$  nM  $< IC_{50} < 500$  nM), and very low-affinity binders ( $500$  nM  $< IC_{50} < 5000$  nM).  $IC_{50}$  value is called the half maximal inhibitory concentration. It is a general quantitative measure indicating how much of an inhibitory substance is necessary to *in vitro* inhibit a given biological process or biological component by 50%. It is typically expressed as molar concentration.

To identify frameshift peptides with potentially the highest relevance for immune recognition, a general epitope likelihood score (GELS) is defined. It accounts for MHC ligand prediction and the prevalence of the respective HLA allele in a defined population, as the latter influences the probability of a frameshift peptide to encompass an MHC ligand recognized by the immune system in a patient of this population [17, 76, 96]. In order to compute this quantity, we have to define several probabilities for the given candidate frameshift peptides to produce immune reactions: the epitope likelihood score (ELS) per HLA type, the general epitope likelihood score (GELS) comprising all HLA types under consideration, and the immune relevance score (IRS).

**Definition 5.1** Epitope likelihood score, ELS [17]

*The epitope likelihood score (ELS) is defined as the probability of a given frameshift neoantigen to be effective across a population, for*

[17]: Ballhausen et al. (2020), “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution”.

[96]: Jurtz et al. (2017), “NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data”.

[168]: Reche and Reinherz (2005), “PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands”.

[171]: Robinson et al. (2016), “The IPD-IMGT/HLA Database – New developments in reporting HLA variation”.

a single HLA type

$$\text{ELS}_H(n) = \left(1 - (1 - f_H)^2\right) \cdot \left(1 - (1 - p_{\text{binding}})^{|E_H(n)|}\right),$$

where  $H \in S$  is a given HLA among the set  $S$  of HLA types,  $n \in \text{cMS}$  is a given frameshift peptide,  $f_H$  the allele frequency of a given HLA allele obtained from [allelefrequencies.net](http://allelefrequencies.net),  $p_{\text{binding}}$  the probability that a given predicted epitope is actually bound, and  $E_H(n)$  the set of all epitopes predicted for a given HLA  $H$  and frameshift peptide  $n$ . Taken together,  $\text{ELS}_H$  is the probability of a given candidate frameshift peptide  $n$  having at least one MHC ligand for an HLA  $H$  and a random person from a given population having at least one allele of  $H$ .

**Definition 5.2** General epitope likelihood score, GELS [17]

The general epitope likelihood score (GELS) is defined as the probability of a candidate frameshift peptide  $n$  having at least one MHC ligand among all HLAs, for which the given HLA is also present in a randomly selected individual

$$\text{GELS}_H(n) = 1 - \prod_{H \in S_X} (1 - \text{ELS}_H(n)), \quad X \in \{A, B\},$$

$$\text{GELS}(n) = \text{GELS}_A(n) + \text{GELS}_B(n) - \text{GELS}_A(n) \cdot \text{GELS}_B(n),$$

where  $S_X$  is the set of HLA types considered for locus  $X \in \{A, B\}$ .

**Definition 5.3** Immune relevance score, IRS [17]

The immune relevance score (IRS) is the joint probability of a given frameshift peptide and its underlying coding microsatellite mutation being present in an individual and at least one predicted binder existing for an HLA present in that individual, assuming independence between the presence of HLA alleles and present frameshift peptides

$$\text{IRS}(n) = p_{\text{mut}}(n) \cdot \text{GELS}(n).$$

[17]: Ballhausen et al. (2020), “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution”.

[76]: González-Galarza et al. (2014), “Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations”.

In [17], the GELS is computed for all frameshift peptides using HLA allele frequencies for US and European Caucasians, where the latter are obtained from the Allele Frequency Net Database [76]. By doing so, we notice that the most commonly mutated coding microsatellite shows a very low GELS, whereas very high GELS candidates seem to be associated with a low mutation frequency. This observation is underlined by a statistically significant inverse correlation between the GELS and the mutation frequency. Further, hierarchical

clustering of all candidate frameshift peptides on all tumor samples reveals the existence of three distinct populations displaying a trend in their mean GELS, with a computed clustering dissimilarity threshold of 5.7.

These observations suggest that emerging tumor cell clones with highly immunogenic frameshift peptides are counterselected during MSI carcinogenesis. Interestingly, when considering tumor subgroups of *B2M*-wild type and *B2M*-mutant tumors, the inverse correlation is not significant among *B2M*-mutant tumors. In these tumors, immune selection on the basis of HLA class I antigen presentation should not apply. The observable trend possibly reflects immune surveillance prior to *B2M* mutation.

Besides that, we account for the possible confounder of the microsatellite length, we do not find a significant relation between the GELS and the microsatellite length. Further, we repeat the analyses for length-adjusted relative mutation frequencies retaining the negative correlation with the GELS in *B2M*-wild type tumors.

Despite the negative correlation of the GELS and the frameshift mutation frequency, also some outliers are observed which show a high GELS and a high frameshift mutation frequency. As hypothesized in [17], this could reflect distinct effects of cell survival like growth advantages of coding microsatellite mutations in tumor suppressor genes. Candidate coding microsatellites with a high GELS and a high mutation frequency are potentially of great importance for the interaction of the immune system and MMR-deficient tumor cells. For a quantitative analysis of these candidates, the immune relevance score (IRS) is computed which combines these two factors. Using this analysis, Ballhausen et al. [17] uncovered various frameshift peptide candidates with predicted importance for the immune biology of MMR-deficient cancers. Interestingly, candidate genes possibly acting as tumor-suppressors are common among the high-IRS genes. This observation may suggest that highly immunogenic frameshift peptides are tolerated preferentially if the cells gain a compensatory survival advantage from the mutation by switching off a tumor-suppressive pathway and thus supporting MSI carcinogenesis [17].

[17]: Ballhausen et al. (2020), "The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoeediting during tumor evolution".

### 5.3 HLA TYPE-DEPENDENT TUMOR-IMMUNE INTERACTIONS

[210]: Witt et al. (2022), “A simple approach for detecting *HLA-A\*02* alleles in archival formalin-fixed paraffin-embedded tissue samples and an application example for studying cancer immunoediting”.

In this section, we consider the data analysis performed in [210] for the quantification of the influence of the HLA type on negative selection in MSI colorectal cancers. Therefore, we split for each HLA type, the group of all patients  $G_{\text{all}}$  into two groups  $G_{\text{pr}}$  consisting of patients with at least one allele of the considered HLA type present and  $G_{\text{non}}$  representing patients without the considered HLA type. We refined the immunological scores introduced above to estimate the immunogenicity of the considered 41 frameshift peptides to the particular HLA-A constellation in the  $G_{\text{pr}}$  and  $G_{\text{non}}$  groups.

[181]: Sidney et al. (2008), “HLA class I supertypes: a revised and updated classification”.

As described above, allele frequency data were obtained from [allelefrequencies.net](http://allelefrequencies.net) for the German population, including all HLA alleles  $A$  with an allele frequency  $f_A > 0.0001$ . These frequencies are summarized according to the supertype classification of [181] to obtain allele frequencies  $f_S$  for the HLA-A supertypes  $S$

$$f_S = \sum_{A \in S} f_A,$$

where  $f_A$  is the frequency of the HLA-A allele  $A$  which belongs to the supertype  $S$ . This can be used to determine for both  $G_{\text{pr}}$  and  $G_{\text{non}}$  groups, the probability  $p(S)$  that at least one allele of the supertype  $S$  is present in a patient of this group.

#### Definition 5.4 Ligand likelihood, LL [210]

Using  $p(S)$ , we compute for each supertype  $S$ , in both groups  $G \in \{G_{\text{pr}}, G_{\text{non}}\}$  the ligand likelihood  $LL_{S,G}$  that is defined to be the likelihood that at least one peptide, derived from the FSP  $m$ , is presented by an allele of the supertype  $S$ . In formulas, this reads

$$LL_{S,G}(m) = p(S) \cdot \left( 1 - \prod_{n \in M} (1 - EL_S(n)) \right),$$

where  $n$  is a peptide derived from an FSP  $m$ ,  $M$  is the set of all possible peptides derived from  $m$ ,  $EL_S(n)$  is the likelihood of  $n$  to be a ligand for the representative of the considered supertype  $S$ , as determined by NetMHCpan-4.1.

**Definition 5.5** Overall ligand likelihood, OLL [210]

Using the above notation, we summarize the ligand likelihood of all supertypes of the HLA-A locus in both groups separately by defining the overall ligand likelihood  $OLL_G$  for group  $G \in \{G_{pr}, G_{non}\}$  by

$$OLL_G(m) = 1 - \prod_{S \in S_G} (1 - LL_{S,G}(m)).$$

It estimates the probability that at least one FSP-derived peptide of  $m$  is presented by an HLA-A allele of the supertype  $S \in S_G$  from the set of supertypes in a patient belonging to the group  $G$ .

We use those definitions, to derive the group-specific quantities in particular by specifying  $p(S)$  in both groups. Exemplarily, this is done for determining the HLA-A\*02 status. Most of the HLA-A\*02 alleles are assigned to the supertype A02 using the supertype classification by [181]. The remaining alleles have a very low allele frequency  $f_A < 0.001$  such that we assume that patients of the  $G_{pr}$  group have at least one allele of the supertype A02 and thus,  $p(A02) = 1$  in this group. The second HLA-A allele may belong to other supertypes ( $S_{G_{pr}} = \{A01, A02, A03, A24\}$ ) and we consider the supertypes A01, A03, A24 with the corresponding supertype frequencies  $f_S$ . This means, for these supertypes,  $f_S$  can be used to approximate the likelihood  $p(S)$  that an allele of these three supertypes is actually present in patients of the  $G_{pr}$  group.

For the  $G_{non}$  group, the probability that a specific HLA-A supertype  $S \in S_{G_{non}} = \{A01, A03, A24\}$  which is not A02 by the group's definition, is present in this group can be estimated by

$$p(S) = 1 - \left(1 - \frac{f_S}{1 - f_{A02}}\right)^2,$$

where the supertype frequencies  $f$  in the general population are used, as derived from [allelefrequencies.net](http://allelefrequencies.net).

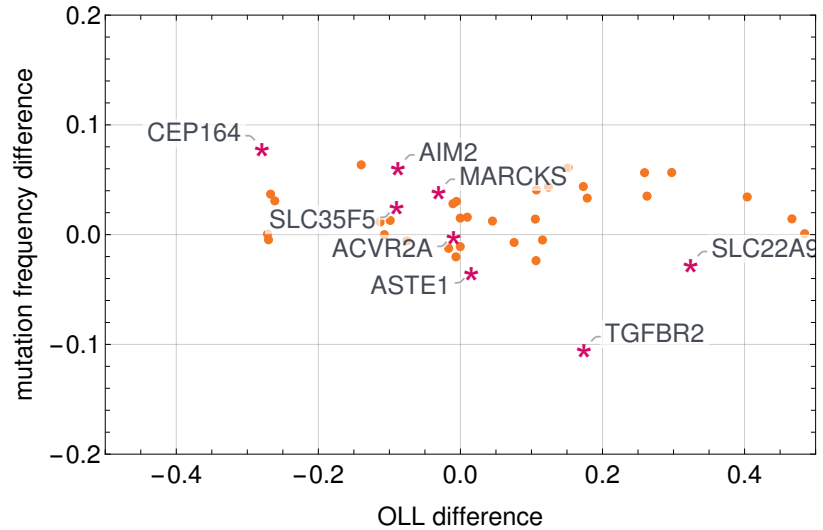
As a first step towards a comprehensive understanding of HLA type-dependent negative selection in MSI colorectal cancers, we focus on HLA-A\*02 and refine the immunogenicity analyses performed in [17] to the particular HLA-A constellation in the  $G_{pr}$  and  $G_{non}$  groups.

We calculate the differences in the overall ligand likelihood  $OLL_{G_{pr}} - OLL_{G_{non}}$  and the average mutation frequencies

[181]: Sidney et al. (2008), "HLA class I supertypes: a revised and updated classification".

[17]: Ballhausen et al. (2020), "The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoeediting during tumor evolution".

**Figure 5.2:** Correlation between the overall ligand likelihood  $OLL_{G_{pr}} - OLL_{G_{non}}$  and the average mutation frequencies  $\bar{f}_{m1,G_{pr}} - \bar{f}_{m1,G_{non}}$  of both groups. The candidates with an average frequency of m1-mutations  $\bar{f}_{m1,G_{all}} > 0.25$  are analyzed separately and labeled by  $\star$  and their gene names. Adapted from [210].



$\bar{f}_{m1,G_{pr}} - \bar{f}_{m1,G_{non}}$  between both groups. By doing so, other factors influencing the coding microsatellite mutation frequencies and immune selection like the length of the underlying microsatellite are eliminated as they are approximately the same in both groups.

[210]: Witt et al. (2022), “A simple approach for detecting *HLA-A\*02* alleles in archival formalin-fixed paraffin-embedded tissue samples and an application example for studying cancer immunoediting”.

For being able to detect a possible *HLA-A\*02*-dependent negative selection of coding microsatellite mutations, the average mutation frequency, irrespective of the *HLA-A\*02* status, should not be too low. Thus, in [210], the coding microsatellites with an average frequency of m1-mutations  $\bar{f}_{m1,G_{all}} > 0.25$  are in addition separately analyzed to exclude candidates with a possibly random distribution of mutations between the two groups.

Indeed, no correlation is detectable when considering all coding microsatellites. However, when only examining the candidates with  $\bar{f}_{m1,G_{all}} > 0.25$ , an inverse correlation is observed ( $p \approx 0.01$ , Pearson’s  $r \approx -0.77$ ) [210] possibly reflecting an influence of the *HLA-A\*02* status on immune selection in MSI colorectal cancers (see Figure 5.2).

In [210], we focused on one *HLA-A* supertype, namely *A02*, allowing the classification of MSI colorectal cancer patients into two groups of patients with and without at least one allele belonging to this supertype. However, the presented approach can be applied to other *HLA-A* superotypes and thus enables an additional subdivision of patients in future studies, leading to higher discriminative power.



HLA-dependent presentation of frameshift peptides and the deviating allele frequencies between populations might explain the varying cancer penetrance in Lynch syndrome individuals, reported in different studies [79, 93, 101, 166]. In patients with an effective presentation of particular antigens by HLA molecules, cell clones with respective cMS mutations in major driver genes may be eliminated by the host's immune system. In this case, the progression of these precancerous cells to a manifest cancer would be impaired by the immune response. For investigating this hypothesis, ATB Heidelberg recently founded the INDICATE initiative (see Figure 5.3) together with other groups from Finland (Helsinki), UK (Newcastle) and the Netherlands (Groningen) [5].

Identifying the HLA type as a possible modulator of cancer risk would allow the development of personalized screening strategies in Lynch syndrome, considering the patient-specific molecular characteristics of emerging tumor cells.

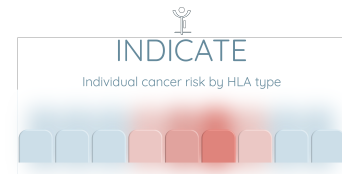
[79]: HAMPEL et al. (2005), "Cancer Risk in Hereditary Non-polyposis Colorectal Cancer Syndrome: Later Age of Onset".

[93]: Jasperson et al. (2010), "Hereditary and Familial Colon Cancer".

[101]: Kloor and von Knebel Doeberitz (2016), "The Immune Biology of Microsatellite-Unstable Cancer".

[166]: Quehenberger (2005), "Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment".

[5]: Ahadova et al. (2022), "Is HLA type a possible cancer risk modifier in Lynch syndrome?"

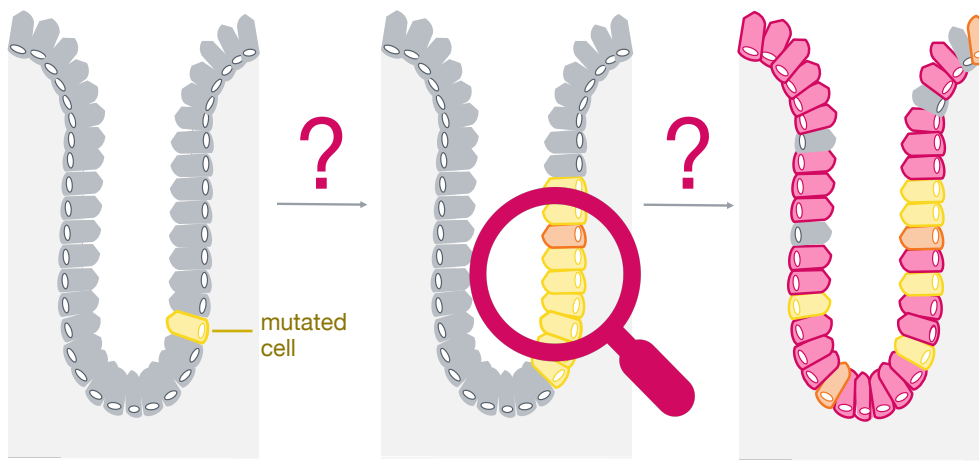


**Figure 5.3:** The INDICATE initiative, [indicate-lynch.org](http://indicate-lynch.org)



**MODELING LYNCH SYNDROME AT THE CELL  
AND CRYPT LEVELS**





## 6 COMPUTATIONAL CELL-BASED MODEL OF INTRA-CRYPT DYNAMICS

This chapter presents our developed computational model at the cell level describing intra-crypt dynamics of early colorectal carcinogenesis in Lynch syndrome.

In general, cancer develops due to the accumulation of different driver mutations which occur during cell division in single cells. For the colon, colonic crypts are believed to be the origin site of colorectal cancer [118]. Crypts are small collections of about 2,000 cells each in humans, with three main types of cells: stem cells, transit-amplifying and fully-differentiated cells (for a more comprehensive biomedical introduction, see Section 2.1.3). Proliferation is ongoing in stem cells and occurs multiple times in transit-amplifying cells near the crypt base giving rise to mutations in both types of cells [16, 95, 165]. The mutations can then spread throughout the crypt and possibly take over the whole crypt meaning that the mutation is present in all cells, a process called monoclonal conversion. However, a detailed understanding of the intra-crypt dynamics underlying both healthy and aberrant crypts and monoclonal conversion in particular is still lacking. In other words, there is a gap in understanding how a mutation in a single cell takes over a colonic crypt and how this contributes to Lynch syndrome carcinogenesis and thus results in tumor risk predictions on a population level. A reason for this is that *in vivo* data with a temporal evolution are very hard to obtain in practice.

We thus developed a computational model with numerical *in silico* simulations of the intra-crypt dynamics during Lynch syndrome colorectal carcinogenesis. The main goal of the

6.1	MODELING CELL DYNAMICS WITHIN A CRYPT USING A VORONOI TESSELLATION . . . . .	85
6.2	SOFTWARE AND HARDWARE BACKGROUND . . . . .	98
6.3	<i>IN SILICO</i> NUMERICAL SIMULATION RESULTS . . . . .	100
6.4	OUTCOMES AND DISCUSSION . . . . .	112

[118]: Leeuwen (2007), “Towards a multiscale model of colorectal cancer”.

[16]: Baker et al. (2019), “Crypt fusion as a homeostatic mechanism in the human colon”.

[95]: Johnston (2008), “Mathematical modelling of cell population dynamics in the colonic crypt with application to colorectal cancer”.

[165]: Potten and Loeffler (1990), “Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt”.

modeling approach is to translate knowledge about the effects of defined mutations from the cellular to the crypt level. Although experimental data on mutation rates in dividing cells *in vitro* are existing, it is hard to translate these numbers onto the level of crypts, the organ, or the individual. In these lines, information about (1) the likelihood of a defined mutation leading to monoclonal conversion of the surrounding crypt, and (2) the time until conversion takes place are paramount. The present model has been designed as a first step to fill this knowledge gap. We aim at answering the following questions:

- ▶ How do mutations spread throughout the crypt?
- ▶ How does a crypt become mutated, i.e. how can a mutation take over the entire crypt?
- ▶ How long does it take until monoclonal conversion occurs?
- ▶ Are these processes dependent on the type of mutation?
- ▶ Is there an influence of cell location or stem cell dynamics on these results?

[66]: Fletcher et al. (2012), “Mathematical modeling of monoclonal conversion in the colonic crypt”.

[117]: Leeuwen et al. (2009), “An integrative computational model for intestinal tissue renewal”.

[132]: Meineke et al. (2001), “Cell migration and organization in the intestinal crypt using a lattice-free model”.

[81]: Haupt et al. (2021), “A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis”.

We present a computational cell-based model as an extension of existing approaches [66, 117, 132] adapted for modeling Lynch syndrome carcinogenesis allowing to obtain *in silico* experiments for mutational processes and intra-crypt dynamics during Lynch syndrome carcinogenesis and thus answering the before mentioned questions. We model the cell cycle including cell division with possible mutations in the early Lynch syndrome colorectal driver genes *APC*, *CTNNB1* and one of the MMR genes, as well as different cell death and feedback mechanisms. Here, the parametrization of mutational events depending on the length of specific gene regions developed in Chapter 4 is used.

The present modeling approach with simulations using the Chaste software is published in Computational and Systems Oncology [81]. In this chapter, we closely follow the lines of thoughts therein. The corresponding code is made publicly available on GitHub ([github.com/Mathematics-in-Oncology/ComputationalColonicCrypts](https://github.com/Mathematics-in-Oncology/ComputationalColonicCrypts)).

## 6.1 MODELING CELL DYNAMICS WITHIN A CRYPT USING A VORONOI TESSELLATION

We assume that a crypt can be geometrically represented by an open cylinder. This is a geometric simplification where other approaches [42, 66] provide corresponding extensions. Instead of considering this two-dimensional surface in 3D, we implement a two-dimensional rectangular surface  $\Omega = [0, 2\pi r_{\text{crypt}}] \times [0, h_{\text{crypt}}]$  with surface area  $L_{\Omega} = 2\pi r_{\text{crypt}} h_{\text{crypt}}$  and periodic boundary conditions.

For the description of the intra-crypt dynamics, we use an off-lattice model, where the cells are described by the positions of their nuclei which are tracked in space and time. Therefore, we assume a Voronoi tessellation model, which has been shown to be a well-suited approach to model epithelia [86]. It is suited for both short-range and long-range dynamics due to its smooth transitions and the ability to easily verify cell neighborhoods using its dual graph, the Delaunay triangulation. By this means, the body of the cell is given by the Voronoi tessellation using Euclidean distances.

### Definition 6.1 Voronoi tessellation

Consider a population of  $k \in \mathbb{N}$  cells denoted as the set  $\{1, \dots, k\}$ , together with the positions of their nuclei  $(r_i)_{i=1, \dots, k} \subseteq \Omega$ . The Voronoi cell (cell body) is then defined as

$$C_i := \{x \in \Omega \mid \|x - r_i\| \leq \|x - r_j\| \forall j \in [k]\}.$$

The collection of cells  $(C_i)_{i=1, \dots, k}$  is called Voronoi tessellation.

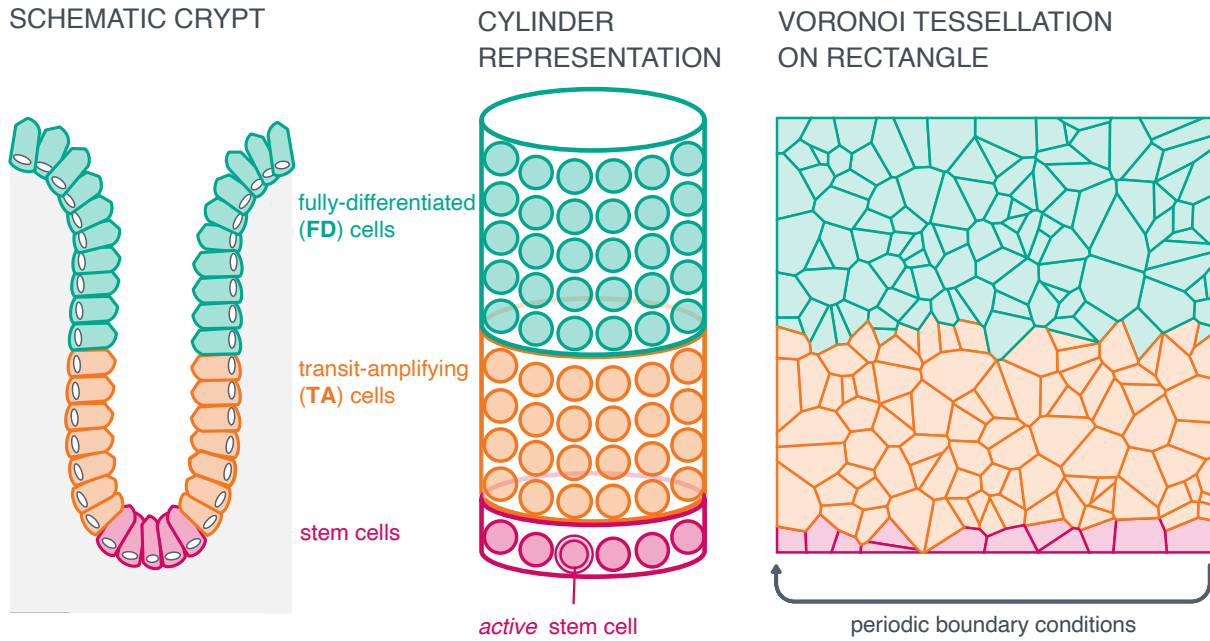
Note that the term *cell nucleus* in our model does not refer to a physical subcellular compartment or structure but is defined as the geometric center of the Voronoi tessellation. It also has a biological meaning regarding the modeling of division by placing the defined nuclei of cells after division at a fixed distance to each other which will be explained in more detail in Section 6.1.3. The geometric representation of the crypt with its Voronoi tessellation ansatz is illustrated in Figure 6.1.

For the human simulations, we initialize the computational model with an equidistant grid of  $80 \times 20$  (length  $\times$  height) nodes and compute the Voronoi tessellation. This leads to a

[42]: Buske et al. (2011), “A Comprehensive Model of the Spatio-Temporal Stem Cell and Tissue Organisation in the Intestinal Crypt”.

[66]: Fletcher et al. (2012), “Mathematical modeling of monoclonal conversion in the colonic crypt”.

[86]: Honda et al. (1996), “Spontaneous Architectural Organization of Mammalian Epidermis from Random Cell Packing”.



**Figure 6.1:** Illustration of the geometric representation and Voronoi tessellation of the simulated crypt. The colonic crypt is represented by a cylinder consisting of **fully-differentiated (FD)** cells at the top of the crypt, **transit-amplifying cells (TA)** in the middle and **stem cells** at the bottom. An active stem cell populates the crypt at any point in time [121, 173]. As we model Lynch syndrome, all cells are initialized with a germline variant in exactly one of the MMR genes. The cylinder is transformed into a rectangle with periodic boundary conditions, where the cells are represented by a Voronoi tessellation. Parts reprinted from [81].

mesh resembling a honeycomb, with symmetrical hexagonal cell shapes.






### 6.1.1 MODELING THE CELL CYCLE

For each cell constructed by the Voronoi tessellation, we consider a cell cycle model which describes cell proliferation, cell differentiation and possibly mutations where the exact circumstances depend on the cell type. For an introduction to the underlying cell biology, we refer to Sections 2.1.2 and 2.1.3. An overview of the biological components included in the computational model is given in Figure 6.2.

**Duration of the cell cycle.** The cell cycle consists of different phases, namely the G1, S, G2, and M phase (for further information see Section 2.1.2). The duration of each cell cycle is assumed to vary stochastically around a mean length for different cells. This is implemented by assuming the G1 phase to vary from cell to cell while the other phase times are assumed to be fixed. Experimentally, the exact distribution is hard to determine [46]. Following [46], we assume that the

[46]: Chao et al. (2019), “Evidence that the human cell cycle is a series of uncoupled, memoryless phases”.



	STEM CELLS 	TA CELLS 	FD CELLS 
<b>CELL CYCLE</b>	couple of weeks	one day	quiescent cell
<b>DIVISION</b>	asymmetric or symmetric after cell death 	Wnt level determines mode asymmetric or symmetric 	no cell division, only upwards migration
<b>MUTATIONS</b>	in <i>APC</i> , <i>CTNNB1</i> and MMR possibly lethal mutations	in <i>APC</i> , <i>CTNNB1</i> and MMR possibly lethal mutations	no additional mutations
<b>DEATH</b>	only due to lethal mutations	due to lethal mutations or mitotic pressure	apoptosis at top of a crypt

**Figure 6.2: Overview of the biological components included in the computational model.** For each cell type, we model the cell cycle including cell proliferation and division, and possible mutations in one of the MMR genes, in *APC* and *CTNNB1*, as well as multiple death mechanisms. Reprinted from [81].

G1 phase follows a normal distribution, i.e., for its length, we assume  $t_{G1} \sim \mathcal{N}(m_{G1}, \sigma_{G1})$ . Here, the values for  $m_{G1}$  and  $\sigma_{G1}$  may vary for different cell types.

For the human simulations, we set the parameters as follows: We assume the TA cell cycle to last for approximately  $m_{t_{cc}} = 24$  hours [51], approximating the duration as follows: The G1 phase lasts about 11 hours, the S phase about 8 hours, the G2 phase about 4 hours, and the M phase about 1 hour. To be precise, we assume for the duration of the G1 phase to be normally distributed with  $\mathcal{N}(11 \text{ hrs}, 0.5 \text{ hrs})$ . Stem cells are assumed to only divide every couple of weeks, while FD cells are quiescent and therefore are permanently in the G0 phase, hence non-dividing.

[51]: Cooper (2018), *The Cell: A Molecular Approach*. 8th edition.

**The Wnt pathway determines the cell type.** One of the main factors in the cell cycle model is the activity of the Wnt pathway which we assume to distinguish the different cell types, namely the dividing cells, i.e., stem cell and TA cells from the non-dividing FD cells. In other words, the activity of the Wnt pathway determines the cell type. Thus, we assume that each cell is assigned a Wnt level  $l_{wnt}$ , where the lower the cell is located in the crypt, the higher is its Wnt level.

**Definition 6.2** Wnt level

The Wnt level  $l_{\text{wnt}}$  is defined by

$$l_{\text{wnt}}: [0, h_{\text{crypt}}] \rightarrow [0, 1],$$

$$h_i(t) \mapsto 1 - \frac{h_i(t)}{h_{\text{crypt}}}, \quad (6.1)$$

where  $h_{\text{crypt}} \in \mathbb{R}_{>0}$  denotes the height of the crypt and the current cell height  $h_i(t)$  is given by the  $y$ -coordinate of the position of the cell nucleus.

We assume that there is only one active stem cell at a time where the remaining stem cells are quiescent, following current biomedical hypotheses [121, 173]. Only the active stem cell is assumed to be responsive to Wnt signaling while the non-dividing quiescent stem cells are non-responsive. The active stem cell resides at the bottom of the crypt surrounded by the quiescent stem cells and thus is assumed to have the highest Wnt level. More details are provided below.

For further distinguishing the FD cells from TA cells, we introduce a Wnt threshold  $\tau_{\text{wnt}}$ , where

$$\text{cell } i \text{ is a } \begin{cases} \text{FD cell,} & \text{if } l_{\text{wnt}}(h_i(t)) < \tau_{\text{wnt}}, \\ \text{TA cell,} & \text{if } l_{\text{wnt}}(h_i(t)) \geq \tau_{\text{wnt}}. \end{cases} \quad (6.2)$$

For the simulations, we set the general Wnt threshold  $\tau_{\text{wnt}} = 0.75$  in order to be consistent with existing schematics of the human colonic crypt, see e.g., [205].

### 6.1.2 MODELING CELL DIFFERENTIATION

The modes of cell differentiation differ for the different cell types, where we refer to Figure 6.2 for a summary. For stem cells, cell differentiation is assumed to be asymmetric, always leading to one stem cell and one TA cell. Only in the case that one stem cell dies, one neighboring until then quiescent stem cell divides symmetrically such that the total number of stem cells remains fixed in time. Either this or another neighboring stem cell then is selected to continue populating the crypt. For TA cells, the Wnt level introduced above determines the cell differentiation mode leading either to two TA cells or to one TA and one FD cell. FD cells are assumed to never divide, and thus cannot differentiate any further.

[121]: Li and Clevers (2010), “Co-existence of Quiescent and Active Adult Stem Cells in Mammals”.

[173]: Sato et al. (2009), “Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche”.

[205]: Weinberg (2013), *The biology of cancer*.

### 6.1.3 MODELING CELL DIVISION AND MUTATIONS

**Cell division.** Cell division is modeled by the creation of a new daughter cell from a mother cell. In Voronoi tessellations, this is done by placing the daughter cell nucleus  $i_d$  at a close, fixed distance  $\epsilon$  from the mother cell nucleus  $i_m$  in a random direction, i.e.,  $\|i_d - i_m\| = \epsilon$ . Subsequently, the tessellation is recomputed. We note that the random placement of daughter cell nuclei in all directions in our model is a more general approach of cell division compared to the fixed lateral placement to the right or to the left in [30]. As there seems to be no biological evidence for a restriction of daughter nucleus placement, we have chosen the more general approach in this respect.

The cell cycle length  $t_{cc}(i_m)$  of the cell with nucleus  $i_m$  is fixed, whereas a new cell cycle length is assigned to  $i_d$  with the G1 phase length sampled from  $\mathcal{N}(m_{G1}, \sigma_{G1})$ . The ages are set to  $\text{age}_{i_m} = \text{age}_{i_d} = 0$  for both cells with cell nuclei  $i_m$  and  $i_d$ .

**Mutational events.** During each cell division, the new daughter cell can acquire mutations in one of the following considered driver genes: one of the MMR genes according to the underlying germline variant of the Lynch syndrome individual, *APC*, and *CTNNB1*. We consider point mutations and LOH events for each of those involved driver genes. The point mutation rate  $p_{\text{pt}}(\text{gene})$  and LOH event rate  $p_{\text{LOH}}(\text{gene})$  are assumed to depend on the corresponding hot spot and full gene length, respectively, as introduced in Chapter 4. For the parameters in Definition 4.1, we assume to accumulate  $n_{\text{pt}} = 10$  point mutations per cell division [206], where there are  $n_{\text{bp, genome}} = 3.2 \cdot 10^9$  base pairs on the genome. The corresponding hot spot lengths and full gene lengths are given in Chapter 4.

Further, the mutational events of each cell can be independent of and dependent on other alterations, as explained in more detail below. The same modeling ansatz of gene length-dependent alteration rates and a connected network of mutational events will be applied to the crypt level model using the Kronecker structure (see Chapter 7).

[30]: Bravo and Axelrod (2013), “A calibrated agent-based computer model of stochastic cell dynamics in normal human colon crypts useful for in silico experiments”.

[206]: Werner et al. (2019), “Measuring single cell divisions in human cancers from multi-region sequencing data”.

**Genotypic status of a cell.** The genotypic status  $g(i, t)$  of a cell  $i$  at time  $t$  is determined by the combination of the mutation status of the three genes (one of the MMR genes, *APC* and *CTNNB1*) at this time given by the triple

$$g(i, t) = (g_{\text{MMR}}(i, t), g_{\text{APC}}(i, t), g_{\text{CTNNB1}}(i, t)),$$

where for all cells  $i$ , for all time points  $t$

$$\begin{aligned} g_{\text{MMR}}(i, t) &\in \{m, l, mm, ml, ll\}, \\ g_{\text{APC}}(i, t) &\in \{\emptyset, m, l, mm, ml, ll\}, \\ g_{\text{CTNNB1}}(i, t) &\in \{\emptyset, m, l, mm, ml, ll\}, \end{aligned}$$

where we introduce the following mutation status for single genes:

- ▶ **State  $\emptyset$ :** None of the alleles is affected by any point mutation nor LOH event.
- ▶ **States  $m$  and  $mm$ :** Point mutations affecting one (respectively two) allele(s).
- ▶ **States  $l$  and  $ll$ :** LOH events affecting one (respectively two) allele(s).
- ▶ **State  $ml$ :** One of the alleles is affected by a point mutation and the other by an LOH event, not differentiating which allele has which alteration and in which order they happened.

As we focus on Lynch syndrome, all cells have a first hit in the respective MMR gene, i.e.,  $g_{\text{MMR}}(i, 0) = m \forall i$  or  $g_{\text{MMR}}(i, 0) = l \forall i$ . We neglect the possibility that two somatic mutations occur in one of the other MMR genes. Further, we assume that two LOH events in *APC* or *CTNNB1* damage the cell in such a way that it directly leads to cell death [160].

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

[64]: Engel et al. (2020), “Associations of Pathogenic Variants in *MLH1*, *MSH2*, and *MSH6* With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome”.

**Mutational dependencies.** According to Engel et al. [64], somatic *CTNNB1* mutations are significantly more frequent in *MLH1*-associated Lynch syndrome colorectal cancer compared to colorectal cancers associated with the other MMR genes. It might be explained by a simultaneous inactivation of *MLH1* and *CTNNB1* triggered by a common LOH event affecting both genes. We incorporate this dependency in our model with an occurrence rate  $r_{\text{effLOH}}$  per LOH event of either gene, where we set  $r_{\text{effLOH}} = 0.8$  for the numerical simulations.

We assume that the second hit at time  $t_2$  in an MMR gene leads to an increased point mutation rate  $p_{\text{pt}}$  by a factor  $\lambda_{\text{pt}} > 1$  compared to MMR-proficient cells for all other genes in this cell. This is modeled by

$$\pi_{\text{pt}}(i, t, \text{gene}) = \lambda_{\text{pt}}(i, t) \cdot p_{\text{pt}}(\text{gene}) \quad \forall t > t_2, \quad (6.3a)$$

where

$$\lambda_{\text{pt}}(i, t) \begin{cases} > 1, & \text{if } g_{\text{MMR}}(i, t) \in \{\text{mm}, \text{m1}, \text{11}\}, \\ = 1, & \text{else.} \end{cases} \quad (6.3b)$$

For the simulations, we set  $\lambda_{\text{pt}} = 100$  as the mismatch repair system, when functioning properly, only fails to detect 1 out of 100 base pair mismatches [205, Section 7.11].

[205]: Weinberg (2013), *The biology of cancer*.

Further, mutations in both *APC* and *CTNNB1* are assumed to increase the activity of the Wnt pathway and thus leading to a prolonged cell proliferation and an increased resistance to cell death. This is implemented in the model by decreasing the Wnt threshold  $\tau_{\text{wnt}}$  by a predefined factor  $\lambda_{\text{wnt}}(i, t) \in [0, 1]$ , which is equivalent to increasing the Wnt level for those cells. The factor varies for different genotypic states. It is assumed to be significantly decreased for biallelically *APC*- or *CTNNB1*-mutated cells, and slightly decreased for monoallelically *APC*-mutated and monoallelically *CTNNB1*-mutated cells. In formulas, this reads

$$t_{\text{wnt}}(i, t) = \lambda_{\text{wnt}}(i, t) \cdot \tau_{\text{wnt}} \quad \forall t, \quad (6.4a)$$

where

$$\lambda_{\text{wnt}}(i, t) \begin{cases} = 0, & \text{if } g_{\text{APC}}(i, t) \text{ or } g_{\text{CTNNB1}}(i, t) \in \{\text{mm}, \text{m1}\}, \\ \in (0, 1), & \text{if } g_{\text{APC}}(i, t) \in \{\text{m}, \text{1}\} \text{ or } g_{\text{CTNNB1}}(i, t) = \text{m}, \\ = 1, & \text{else.} \end{cases} \quad (6.4b)$$

The effects of the different mutational status on the intra-crypt dynamics are summarized in Table 6.1.

**Table 6.1: Effects of mutations on intra-crypt dynamics.** We summarize the effects on mutation rates, differentiation and cell death for MMR deficiency, monoallelic and biallelic *APC* or *CTNNB1* mutations, which are implemented in the model. Reprinted from [81].

Mutation	Effect on mutation rate	Effect on differentiation	Effect on death
MMR deficiency	increase	none	increase of mutation-induced death
monoallelic <i>APC</i> or <i>CTNNB1</i>	none	delay	none
biallelic <i>APC</i> or <i>CTNNB1</i>	none	inhibition	partial apoptotic resistance, cell death if complete gene loss

For the exact parameter values used in the numerical simulations, we assume

$$\lambda_{\text{wnt}}(i, t) = \begin{cases} 0, & \text{if } g_{APC}(i, t) \text{ or } g_{CTNNB1}(i, t) \in \{\text{mm}, \text{m}\}, \\ 0.8, & \text{if } g_{APC}(i, t) \in \{\text{m}, \text{l}\}, \\ 0.9, & \text{if } g_{CTNNB1}(i, t) = \text{m}, \\ 0.8 \cdot 0.9, & \text{if } g_{APC}(i, t) \in \{\text{m}, \text{l}\} \text{ and } g_{CTNNB1}(i, t) = \text{m}, \\ 1, & \text{else.} \end{cases}$$

#### 6.1.4 MODELING CELL MIGRATION

We assume that the cells of the colonic crypt exert pressure on each other. As cell division happens only in the lower part of the crypt, the pressure in this region is higher compared to the upper part of the crypt. The pressure which occurs due to cell division is called mitotic pressure. This pressure results in an upward movement of the cells along the vertical crypt axis. It is suggested that mitotic pressure is the main contributor to upward migration [161]. We will not explicitly model this pressure at a tissue level but rather implicitly incorporate mitotic pressure into a force-based cell mechanics model which will be explained in this section. We use a *linear spring force*, meaning that virtually all cells are connected via mechanical springs, as defined in [132]. Here, we make two assumptions:

- ▶ The forces between the cells are modeled by a network of springs, where the latter are the vectors connecting one cell to another.
- ▶ Each force linearly depends on this vector.

[161]: Paulus et al. (1992), “A model of the control of cellular regeneration in the intestinal crypt after perturbation based solely on local stem cell regulation”.

[132]: Meineke et al. (2001), “Cell migration and organization in the intestinal crypt using a lattice-free model”.

**Definition 6.3** Spring

Let  $r_i(t), r_j(t) \in \Omega$  denote the positions of the nuclei of two cells  $i$  and  $j$ . Then, the vector  $r_{ij}(t) = r_j(t) - r_i(t)$  is called spring from cell  $i$  to cell  $j$  at time  $t$ .

**Definition 6.4** Resting spring length

We consider cells  $i$  and  $j$  at time  $t$ . Then, the resting spring length, denoted by  $s_{ij}(t)$ , is the ‘natural length’ of the spring between those two cells. It corresponds to the minimal length at which cell  $i$  does not exert pressure on cell  $j$ , and is defined in the following way

$$s_{ij}(t) = \begin{cases} \epsilon + (s - \epsilon) \cdot \frac{\text{age}_i(t)}{t_g}, & \text{if both cells are newly divided,} \\ & \text{i.e., } \text{age}_i(t) < t_g, \text{ and } \text{age}_j(t) < t_g, \\ s_i(t) + s_j(t), & \text{else,} \end{cases} \quad (6.5a)$$

where

$$s_i(t) = \begin{cases} \frac{s}{2} \frac{t_d(i,t)}{t_a}, & \text{if cell } i \text{ undergoes apoptosis, i.e., } t_d(i,t) < t_a, \\ \frac{s}{2}, & \text{else.} \end{cases} \quad (6.5b)$$

Here,  $\epsilon \in \mathbb{R}_{>0}$  is the distance between two cell nuclei after division,  $\text{age}_i(t) \in \mathbb{R}_{\geq 0}$  is the age of cell  $i$  at time  $t$ ,  $t_g \in \mathbb{R}_{>0}$  is the time needed for a newly divided cell to grow to its original size,  $t_d(i,t) \in \mathbb{R}_{\geq 0}$  is the time until the death of cell  $i$ , and  $t_a \in \mathbb{R}_{>0}$  the duration of apoptosis. Further, the parameters  $s \in \mathbb{R}_{>0}$ ,  $\epsilon$ ,  $t_g$  and  $t_a$  are assumed to be time-independent and fixed for each cell.

Directly after cell division, the daughter cell is rather small and the distance between the two cells is  $\epsilon$ . After cell growth, this distance is increased to  $s \geq \epsilon$ . As the age of the cells  $\text{age}_i(t)$  is linearly increasing with time  $t$ , this is also true for the resting spring length  $s_{ij}(t)$  until  $\text{age}_i(t) = t_g$ .

Further, if cell  $i$  undergoes apoptosis, that is  $t_d(i,t) < t_a$ , it holds

$$\begin{aligned} \lim_{t \rightarrow t_a} s_{ij}(t) &= \lim_{t \rightarrow t_a} \frac{s}{2} \overbrace{\frac{t_d(i,t)}{t_a}}^{\rightarrow 0} + \frac{s}{2} \\ &= \frac{s}{2}. \end{aligned}$$

**Definition 6.5** Linear spring force (i) The force exerted by cell  $i$  on cell  $j$  at time  $t$  is defined by the vector

$$f_{ij}(t) := \mu \cdot \frac{r_{ij}(t)}{\|r_{ij}(t)\|} (s_{ij}(t) - \|r_{ij}(t)\|), \quad (6.6)$$

where  $\mu \in \mathbb{R}_{\geq 0}$  is the so-called elasticity constant.

(ii) We further define the sum of all forces exerted on cell  $i$  at time  $t$  as

$$F_i(t) := \sum_{j \in \mathcal{N}(i)} f_{ji}(t), \quad (6.7)$$

where  $\mathcal{N}(i)$  denotes the Delaunay neighborhood of cell  $i$ , corresponding to the dual graph of the Voronoi tessellation of cell  $i$ .

If the cells have not reached their natural distance, i.e.,  $s_{ij}(t) > \|r_{ij}(t)\|$  in Equation (6.6), the force is called *repulsive*, driving the cells further away from each other. If  $s_{ij}(t) < \|r_{ij}(t)\|$ , the force is called *attractive*.

Further, the parameter  $\mu$  describes the elasticity of the cell. Large values of  $\mu$  lead to larger forces, in other words, the cells are easier to push away or pull nearer, respectively.

Almost always, the force is antisymmetric, that is  $f_{ij} = -f_{ji}$ . Only if both cells are of different ages less than  $t_g$ , i.e., both are newly divided cells from distinct mother cells, this property is invalid.

The force can now be used to define the motion of cells over time, where we assume Brownian dynamics, as in the model by [132]. However, we extend the originally proposed equation of motion in [132] by a mechanism causing upward migration. This can be achieved by incorporating a basement membrane flow, additionally to mitotic pressure [83]. The former is represented by an additive term increasing the second component of the cell position vector  $r_i$ , as defined in the following.

**Definition 6.6** Equation of motion

The change of the position  $r_i$  of cell  $i$  over time  $t$  is described by the ordinary differential equation

$$\frac{d}{dt} r_i(t) = \frac{1}{v} F_i(t) + \begin{pmatrix} 0 \\ \gamma \end{pmatrix}, \quad (6.8)$$

[132]: Meineke et al. (2001), "Cell migration and organization in the intestinal crypt using a lattice-free model".

[83]: Heath (1996), "Epithelial cell migration in the intestine".



where  $\nu \in \mathbb{R}_{>0}$  is the so-called damping constant of a cell and  $F_i(t)$  are the forces exerted on cell  $i$  at time  $t$  defined in Definition 6.5. Further,  $\gamma \in \mathbb{R}_{\geq 0}$  is the additional increase in height caused by the basement membrane flow.

The tissue-dependent parameter  $\nu$  describes cell-matrix adhesion. The greater the value of  $\nu$ , the stronger the cell adheres to the extracellular matrix, which is in our case the basement membrane. This leads to decreased cell mobility. The latter is described by  $\delta = \frac{\mu}{\nu}$  with  $\mu$  defined in Definition 6.5.

In our case, the Equation of Motion (6.8) is discretized in time using the forward Euler method

$$r_i(t + \Delta t) = r_i(t) + \Delta t \cdot \left( \frac{F_i(t)}{\nu} + \begin{pmatrix} 0 \\ \gamma \end{pmatrix} \right), \quad (6.9)$$

where  $\Delta t$  denotes the default time step. This cell migration update is called at every time step, before checking for cell division.

In our simulations, we set  $\Delta t = \frac{1}{45}$  hours. Further, to the best of our knowledge, no experimental estimates exist for the parameters of the spring force model. We have chosen a parameter combination leading to simulation results which are most consistent with the biological reality. In summary, we set  $s = 1$  length unit,  $\epsilon = 0.5$  length units,  $t_g = 3$  hours,  $t_a = 0.5$  hours,  $\mu = 14$ ,  $\gamma = 0.0675$  length units per hour, and  $\nu = 1$ .

### 6.1.5 MODELING CELL DEATH

Modeling cell death is quite important for tissue homeostasis. We will incorporate three different mechanisms how a cell of a crypt can die, where in each case, the respective node is removed from the Voronoi tessellation and after this process, all associated springs are deleted. The incorporated mechanisms are: 1) The cell reaches the top of the crypt and is sloughed into the colonic lumen, 2) the cell falls victim to homeostatic mechanisms, or 3) the cell dies due to the acquisition of a disadvantageous mutation.

**Modeling cell sloughing.** The cells regularly undergo apoptosis after they have finished the migration through the crypt. This is incorporated by letting all cells die when they reach the crypt height, i.e., if for cell  $i$

$$h_i(t) = h_{\text{crypt}}. \quad (6.10)$$

**Modeling feedback mechanisms via homeostasis.** Feedback mechanisms within a crypt are modeled by homeostatic cell death in order to avoid an overpopulation of the crypt with TA cells. For this, we assume exemplarily that cells with a surface area below a certain threshold  $\tau_{\text{size}}$  are non-viable and will die. The exact value of this threshold regulates the number of cells present upon the crypt's homeostatic equilibrium. In other words, the minimum cell size determines the number of cells present in crypt homeostasis.

Cells which have acquired either two *CTNNB1* mutations or two *APC* mutations are assigned a lower threshold, since these cells are partially able to ignore biochemical signals which would normally induce apoptosis. This means these cells are viable with a smaller size than those without these mutations which drastically increases the chance of survival and eases proliferation for these cells. In formulas, this reads

$$t_{\text{size}}(i, t) = \lambda_{\text{size}}(i, t) \cdot \tau_{\text{size}} \quad (6.11a)$$

where

$$\lambda_{\text{size}}(i, t) \begin{cases} \in (0, 1), & \text{if } g_{\text{APC}}(i, t) \text{ or } g_{\text{CTNNB1}}(i, t) \in \{\text{mm}, \text{m1}\}, \\ = 1, & \text{else.} \end{cases} \quad (6.11b)$$

We set the parameter in our simulations

$$\tau_{\text{size}}(i, t) = \begin{cases} 0.3, & \text{if } g_{\text{APC}}(i, t) \text{ or } g_{\text{CTNNB1}}(i, t) \in \{\text{mm}, \text{m1}\}, \\ 0.43, & \text{else.} \end{cases}$$

We assume further homeostatic mechanisms within smaller subpopulations, like Wnt-induced senescence, and assign an additional probability  $p_{cc}$  of homeostatic death to all cells in the following way.

**Proposition 6.7**

Let  $p_{cc} \in [0, 1]$  denote the probability of homeostatic cell death per cell cycle. Further, let  $\Delta t$  denote the default time step and  $m_{t_{cc}}$  the average duration of the cell cycle, both in hours. Then, the probability of cell death at each time step is given by

$$p_t = 1 - (1 - p_{cc})^{\frac{\Delta t}{m_{t_{cc}}}}. \quad (6.12)$$

*Proof.* The number of time steps per hour is given by  $\frac{1}{\Delta t}$ . Therefore, there are  $\frac{m_{t_{cc}}}{\Delta t}$  time steps per cell cycle. The probability of not dying during one cell cycle duration is

$$1 - p_{cc} = (1 - p_t)^{\frac{m_{t_{cc}}}{\Delta t}}.$$

Rearranging for  $p_t$  yields the statement.  $\square$

Here, only the parameters for modeling cell mobility and TA cell cycle length determine the size of the effect, which demonstrates the importance of feedback mechanisms for tissue organization and homeostasis. During simulations, we set the rate of cell death induced by all other homeostatic processes per cell cycle  $p_{cc} = 0.0005$ .

**Modeling mutation-induced death.** Mutations damaging a cell in such a way that it is not viable anymore are called *lethal* mutations. We incorporate this mutation-induced death as apoptosis of the daughter cell as mutations are introduced in the daughter cells during cell division. As introduced earlier, LOH events on both alleles of *APC* or *CTNNB1* are assumed to be lethal in every case. Further, the probability of a lethal point mutation,  $p_{\text{mutdeath}}$ , is higher in MMR-deficient cells, as those have a higher mutation rate. In formulas, we obtain

$$p_{\text{mutdeath}}(i, t) = \begin{cases} 1.0, & \text{if } g_{APC}(i, t) = \mathbb{1}\mathbb{1} \\ & \text{or } g_{CTNNB1}(i, t) = \mathbb{1}\mathbb{1}, \\ \lambda_{\text{pt}}(i, t) \cdot \pi_{\text{mutdeath}}, & \text{else} \end{cases} \quad (6.13)$$

with  $\lambda_{\text{pt}}$  defined in Equation (6.3b) and the rate of mutation-induced cell death per cell division  $\pi_{\text{mutdeath}} = 0.0001$  in the simulations. Further, the effects of the different mutations on cell death are summarized in Table 6.1.

[121]: Li and Clevers (2010), “Co-existence of Quiescent and Active Adult Stem Cells in Mammals”.

[173]: Sato et al. (2009), “Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche”.

[15]: Baker et al. (2014), “Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon”.

## 6.1.6 MODELING STEM CELL DYNAMICS

We assume that each colonic crypt is populated by a single stem cell at a time [121, 173], which populates the crypt for the time  $t_{\text{stem}}$  of stem cell cycle length. This active stem cell divides asymmetrically, renewing itself and giving rise to a new TA cell. If a mutation occurs upon cell division, the TA cell passes on the mutation to its descendants. In our current Chaste implementation, stem cells are modeled in a simplified way and not visualized explicitly. To be more precise, if the active stem cell becomes mutated, all cells in the current bottom row of the Voronoi tessellation become mutated, assuming this row of cells to be the progeny of the mutated TA cell.

Adjacent to the active stem cell, there are  $S - 1 \in \mathbb{N}$  quiescent stem cells, which do not divide. Over time, the  $S$  stem cells alternate at populating the crypt with a probability  $p_{\text{change}}$  of stem cell exchange per stem cell division. The new active stem cell is chosen with uniform probability among the  $S - 1$  quiescent stem cells. Stem cell death is only possible due to a lethal mutation with a probability given in Equation (6.13). In this case, symmetric division of an adjacent stem cell compensates the dead stem cell, in order to maintain a constant number  $S$  of stem cells over time.

For the numerical simulations, we assume  $S = 6$  stem cells per crypt [15] with a stem cell cycle duration of ten weeks. The probabilities for stem cell mutations and mutation-induced death are the same as for TA cells. The probability of stem cell exchange per stem cell division is set to  $p_{\text{change}} = 0.5$ . Thus, on average, a stem cell will populate the crypt for about five months.

## 6.2 SOFTWARE AND HARDWARE

### BACKGROUND

We state a summary of the computational model in form of a pseudocode in Algorithm 1. It includes all parameters which have to be defined prior to the simulations, the necessary initialization steps, as well as the main computations during each time step and stem cell cycle.

---

**Algorithm 1: Summary of the presented computational model given as pseudocode.** We state the necessary input parameters and initialization steps as well as the main computational procedure performed in each time step and stem cell cycle. The corresponding equations are given in brackets. Reprinted from [81].

---

```

1 PARAMETERS:
2   crypt size  $h_{\text{crypt}}, 2\pi r_{\text{crypt}}$ ;
3   stem cell number  $S$ ;
4   stem cell exchange  $p_{\text{change}}$ ;
5   cell cycle lengths  $m_{t_{\text{cc}}}, t_{\text{stem}}$ ;
6   death rates  $p_{\text{cc}}, \pi_{\text{mutdeath}}$ ;
7   mutation rates  $\lambda_{\text{mut}}, \pi_{\text{mut}}$ ;
8   Wnt threshold  $\tau_{\text{wnt}}$ ;
9   size threshold  $\lambda_{\text{size}}, \tau_{\text{size}}$ ;
10  time step  $\Delta t$ ;
11  migration parameters  $\epsilon, s, t_g, t_a, \mu, \nu, \gamma$ ;
12 INITIALIZATION:
13  create honeycomb mesh;
14  create location indices;
15  initialize cells by Voronoi (Def.6.1);
16  cell cycle model and Wnt level (6.1);
17 for stem cell cycle  $i = 1$  to  $n$  do
18   for every time step do
19     kill cells after sloughing (6.10);
20     kill cells with completed apoptosis;
21     cell migration (6.5)–(6.9);
22     recompute Voronoi and Delaunay;
23     check for differentiation (6.2), (6.4);
24     update cell cycle phase;
25     check for cell division;
26     if cell divides then
27       check for mutations (6.3a);
28       check for mutation-induced death;
29       mark cells for apoptosis (6.13);
30       create daughter cell with node and properties;
31     mark cells for apoptosis (6.11a), (6.12);
32     visualize;
33   check for stem cell loss, exchange, mutations;

```

---

All model simulations and *in silico* experiments were conducted within the Chaste framework (Version 2019.1)[135] ([www.cs.ox.ac.uk/chaste](http://www.cs.ox.ac.uk/chaste)).

[135]: Mirams et al. (2013), “Chaste: An Open Source C++ Library for Computational Physiology and Biology”.

The simulations were run on a modern workstation. As parallelization is not yet available in Chaste, each crypt was computed in a sequential way. For simulating a single crypt with 1600 cells for one year of human life-time based on our parameter setting, approximately 630 million operations had to be computed. This took approximately 44 hours of computation time.

### 6.3 *IN SILICO* NUMERICAL SIMULATION RESULTS

For our computational model of Lynch syndrome crypts, we adapted and extended the existing Chaste implementation of a general crypt model with main changes in the cell cycle model and additional features for the Lynch syndrome-related mutations, for stem cell dynamics, and for the feedback mechanisms affecting homeostatic cell death, where the latter were implemented from the ground up. In particular, we added the three mutations of interest, and included the effects of each on the cell cycle and cell differentiation models. Further, we included our own stem cell model and extended the already existing cell migration model, as described in Section 6.1.4. Finally, the feedback mechanism was implemented in Chaste. The implementation was mainly done by Nils Gleim and is accessible on GitHub [github.com/Mathematics-in-Oncology/ComputationalColonicCrypts/releases/tag/v1.0](https://github.com/Mathematics-in-Oncology/ComputationalColonicCrypts/releases/tag/v1.0), release v1.0. A pseudocode of the computational model is given in Section 6.2. Videos of some simulation runs are provided online<sup>1</sup> and referenced accordingly in this section.

1: [SaskiaHaupt.de/phd-thesis](https://saskiahaupt.de/phd-thesis)

In order to answer the main questions raised at the beginning of this chapter, we analyzed epithelial renewal times in mice and humans, monoclonal conversion of different types of mutations, as well as the influence of cell location and stem cell dynamics on the spread of mutations within a crypt. By this, we are able to gain a better understanding of the biomedical mechanisms leading to the spread of advantageous mutations within a crypt, which in turn are known as driver events in colorectal carcinogenesis.

Further, we obtained *in silico* estimates for the duration of epithelial renewal and of monoclonal conversion of different mutations in these driver genes. The results were previously published in [81] which we closely follow in this section.

### 6.3.1 EPITHELIAL RENEWAL IN NON-MUTATED CRYPTS

The renewal time of a crypt is the duration of the complete exchange of all cells within this crypt, which is rather short in many tissues. While in the murine small intestine, reliable estimates of less than one week have been established [48], estimations for the human colon are less precise.

In our model, the renewal time is most significantly influenced by the crypt size and the parameters describing cell mobility, in particular by the interplay of the overall cell mobility parameter  $\delta$  and the basement membrane flow parameter  $\gamma$  introduced in Definition 6.6.

We have calibrated the model in such a way that using the parameters described in Section 6.1, the *in silico* renewal times for murine crypts obtained in our numerical simulations are in concordance with the available estimates in mice [49]. To be precise, a crypt which initially consists of about 200 cells, resulting in an equilibrium state of about 220 cells, which is representative of a murine crypt, is renewed every 6 days, in concordance with the estimates in mice [49].

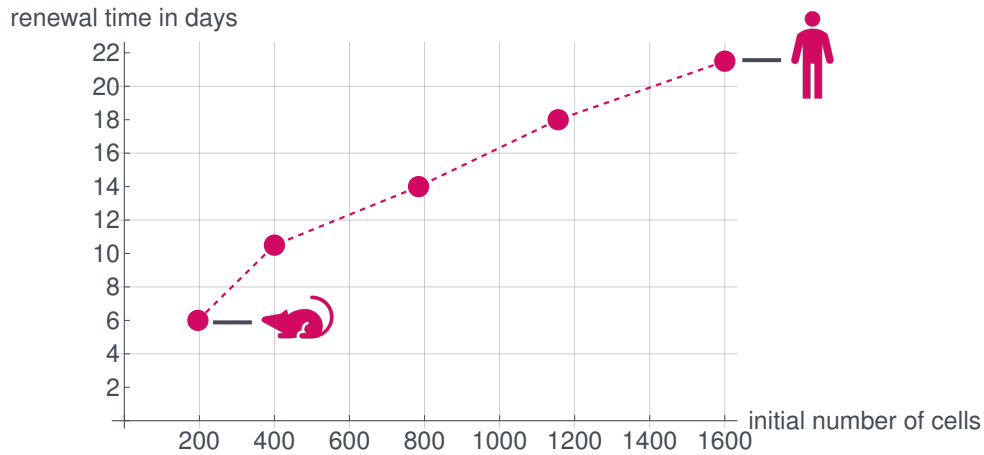
From this point onwards, we used the calibrated parameters for simulations of human colonic crypts, where measurements are typically scarce. For human crypts initially consisting of 1600 cells, the *in silico* estimates for the renewal time increase to three weeks. This suggests that if the parameters are comparable between mice and humans, the process of epithelial renewal takes a number of weeks in humans. The estimates are illustrated in Figure 6.3.

We want to highlight that the renewal times for different initial cell numbers are not strictly linear. This is in concordance with our expectation since 1) the crypt height, which is one determining factor, does not increase linearly with the total cell number and 2) the increase in the number of FD cells inhibits crypt renewal more than the same (relative) increase in the number of TA cells accelerates it.

[81]: Haupt et al. (2021), “A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis”.

[48]: Clevers (2013), “The intestinal crypt, a prototype stem cell compartment”.

[49]: Cole and McKalen (1961), “Observations of cell renewal in human rectal mucosa in vivo with thymidine-H3”.



**Figure 6.3: Renewal times depending on the initial cell number.** Single simulations for an initial number of 196, 400, 784, 1156 and 1600 cells, respectively, result in equilibrium cell numbers of about 220, 460, 910, 1350 and 1850 cells. The initial setup in all cases followed a 4:1 ratio of crypt height to crypt circumference, measured by the number of cells. All other parameters were set as in Section 6.1. All performed simulations showed renewal times only differing by a few hours. Adapted from [81].

### 6.3.2 THE SPREAD OF STEM CELL AND TA CELL MUTATIONS

When studying different mutational events, we observe differences between different types of mutational processes, namely non-transforming which do not directly confer a significant fitness advantage, and transforming mutations. Exemplarily, we consider the following non-transforming mutations

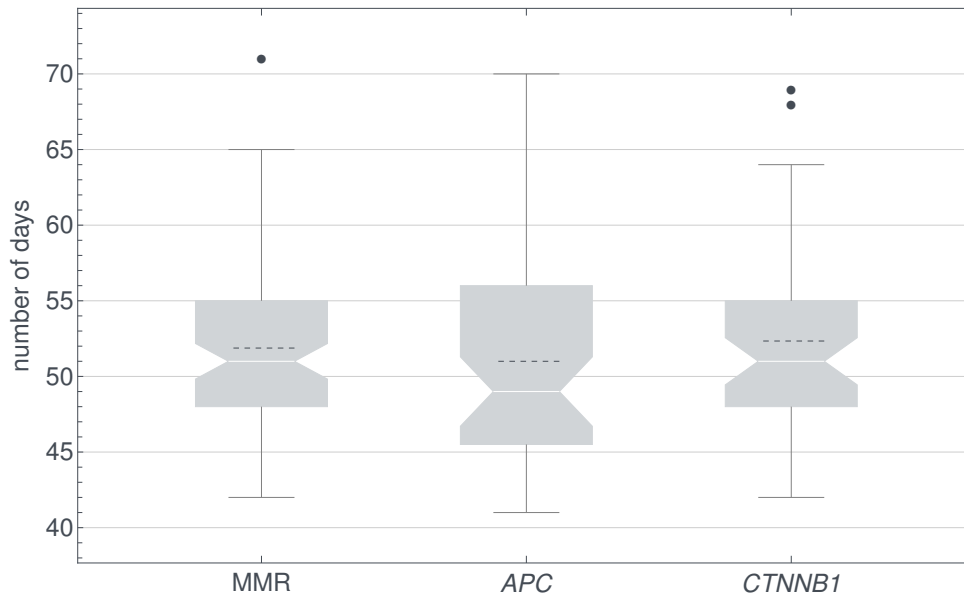
- ▶ two inactivating MMR mutations:  
 $g_{\text{MMR}}(i, t) \in \{\text{mm}, \text{m}\bar{1}, \bar{1}\bar{1}\},$
- ▶ one inactivating *APC* mutation:  
 $g_{\text{APC}}(i, t) \in \{\text{m}, \bar{1}\},$
- ▶ one activating *CTNNB1* mutation:  
 $g_{\text{CTNNB1}}(i, t) = \text{m}.$

As examples of transforming mutations, we consider the double-hit *APC* and the double-hit *CTNNB1* mutation, i.e.,

- ▶ two inactivating *APC* mutations:  
 $g_{\text{APC}}(i, t) \in \{\text{mm}, \text{m}\bar{1}\},$
- ▶ double-hit *CTNNB1* mutation:  
 $g_{\text{CTNNB1}}(i, t) \in \{\text{mm}, \text{m}\bar{1}\}.$

We analyzed the spread and monoclonal conversion of these two types of mutations in the following.





**Figure 6.4: Comparison of monoclonal conversion of MMR deficiency, monoallelic *APC* and *CTNNB1* mutations.** The results are illustrated by notched Box-Whisker plots, where the notches show the 95% confidence intervals and the dashed lines indicate the means. The respective medians (white line) amounted to 51 days (95% CI: [49.8; 52.2]) for MMR deficiency, 51 days (95% CI: [49.4; 52.6]) for monoallelic *CTNNB1* mutations, and 49 days (95% CI: [46.7; 51.3]) for monoallelic *APC* mutations. Reprinted from [81].

**Spread of non-transforming mutations.** For analyzing the spread and monoclonal conversion of non-transforming mutations, we initialized the computational model with the corresponding mutation in the active stem cell and ran several numerical simulations.

**An MMR-deficient stem cell almost always leads to monoclonal conversion between 6 and 9 weeks.** For analyzing the spread of MMR deficiency throughout the crypt, we used an MMR-deficient stem cell as initial condition and run 98 simulations. 88 of them predicted monoclonal conversion to be completed after between 42 and 71 days, with an average value of 51.9 days within this subset of the simulations, illustrated in Figure 6.4. In four of the remaining simulations, a biallelic *APC* mutation occurred before monoclonal conversion was completed. The predictions of the other six remaining simulations are discussed in Section 6.3.3.

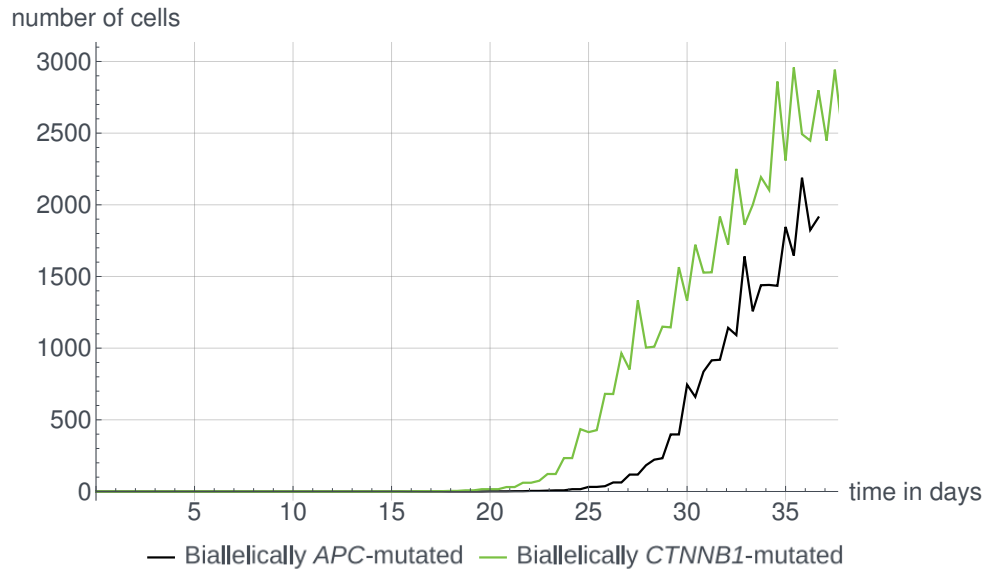
**The kinetics of the spread of monoallelic *APC* and *CTNNB1* mutations resemble the one of MMR deficiency.** After a stem cell mutation in either *APC* or *CTNNB1*, the mutation spreads throughout the crypt in the great majority of cases (*APC*: 52/53, *CTNNB1*: 50/54). For both types of mutations, our model predicts monoclonal conversion to take about

52 days on average (*APC*: 51 days, *CTNNB1*: 52.3 days), see Figure 6.4.

Based on a Kruskal-Wallis test, the null hypothesis that the means for MMR deficiency, monoallelic *APC* and monoallelic *CTNNB1* mutations are the same is not rejected at the 5% significance level ( $p = 0.266$ ) indicating a highly similar duration time of monoclonal conversion for non-transforming mutations. The similarity resides in the fact that the delay of differentiation of cells harboring the monoallelic *APC* or *CTNNB1* mutation does not provide any advantage regarding the *speed* of the spread of the mutation, as it does not change the cell's proliferative behavior, survival, or mobility. This also implies a high similarity of the duration time of monoclonal conversion for non-transforming mutations and wild-type crypts. The effect on differentiation only increases the probability that a mutation spreads at all. This probability is indeed very high for stem cell mutations, but rather low for TA cell mutations. Here, most of the expansions of a mutated clone can be prevented by the feedback mechanism described by homeostatic death mechanisms, and by the short renewal time of a crypt. The latter process results in such clones being frequently washed out of a crypt, which becomes inevitable as soon as all cells of the clone complete differentiation. This underlines the importance of stem cells regarding the origin of colorectal cancer.

**Monoallelic *APC* mutations in TA cells** are frequent in our simulations due to the high number of hot spot regions. As we assume no dominant-negative effects of *APC* mutations for modeling, monoallelic *APC* mutations are rather non-transforming since the second allele continues to produce a sufficient amount of the *APC* protein. In addition, differentiation is prolonged due to the lower Wnt threshold. The expansion of a monoallelically mutated *APC* clone can be prevented in most cases by the feedback mechanisms and the fast renewal of the crypt resulting in a wash-out.

In the case of MMR deficiency, we assume point mutations to be much more likely than in MMR-proficient cells. In the case of an MMR-deficient stem cell, this results in the formation of many clones with monoallelic *APC* mutations, most of which are however washed out. For instance, we observe up to 45 monoallelic *APC* mutations in a single crypt over the course of 30 weeks, none of which gives rise to a second inactivation event.

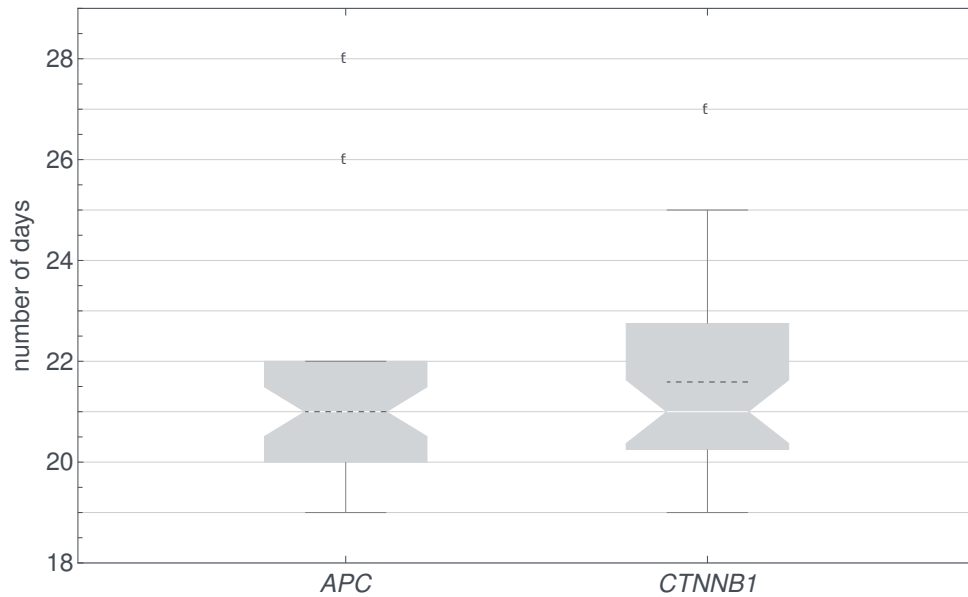


**Figure 6.5:** Exponential expansion of a biallelically mutated clone. Evolution of the number of biallelically APC-mutated cells and biallelically CTNNB1-mutated cells. While the simulation for CTNNB1 ran for approximately 60 days and it reached the plateau after 35 days, we only show the initial 35 days to ensure a one-to-one comparison. Videos of the complete simulations are provided online [SaskiaHaupt.de/phd-thesis#video2](https://SaskiaHaupt.de/phd-thesis#video2) and [SaskiaHaupt.de/phd-thesis#video3](https://SaskiaHaupt.de/phd-thesis#video3). Reprinted from [81].

**Spread of transforming mutations.** If a cell with one mutation in either *APC* or *CTNNB1* acquires a second hit, it becomes partially resistant to apoptosis (excluding 11 which is assumed to be not compatible with survival). Further, the minimum size threshold for homeostatic death is lowered, which both result in a heavily increased chance of survival. The latter is reflected by the spread of such mutations.

**Biallelic APC and CTNNB1 mutations always lead to monoclonal conversion.** The evasion of the feedback mechanism allows the clone with biallelic *APC* or *CTNNB1* mutation to initially grow exponentially and always take over the crypt in our simulations. Exemplary figures and simulation videos illustrating the evolution of cells with biallelic *APC* or *CTNNB1* mutations are shown in Figure 6.5.

As an important contrast to non-transforming mutations, the location of the first mutated cell does not play a role regarding the mutation's ability to spread. Further, monoclonal conversion is completed significantly faster compared to the non-transforming mutations: The means are 21 days and 21.6 days for biallelic *APC* and *CTNNB1* mutations, respectively, see Figure 6.6. However, based on a Kruskal-Wallis test, there is no significant difference in the means between the biallelic *APC* and *CTNNB1* mutations ( $p = 0.09$ ).



**Figure 6.6: Comparison of monoclonal conversion of biallelic *APC* and *CTNNB1* mutations.** The results are illustrated by notched Box-Whisker plots, as in Figure 6.4. For both biallelic *APC* and *CTNNB1* mutations, the median over our simulations (41 for *APC*, 39 for *CTNNB1*) amounted to 21 days with 95% confidence intervals of [20.5; 21.5] and [20.4; 21.6], respectively. Reprinted from [81].

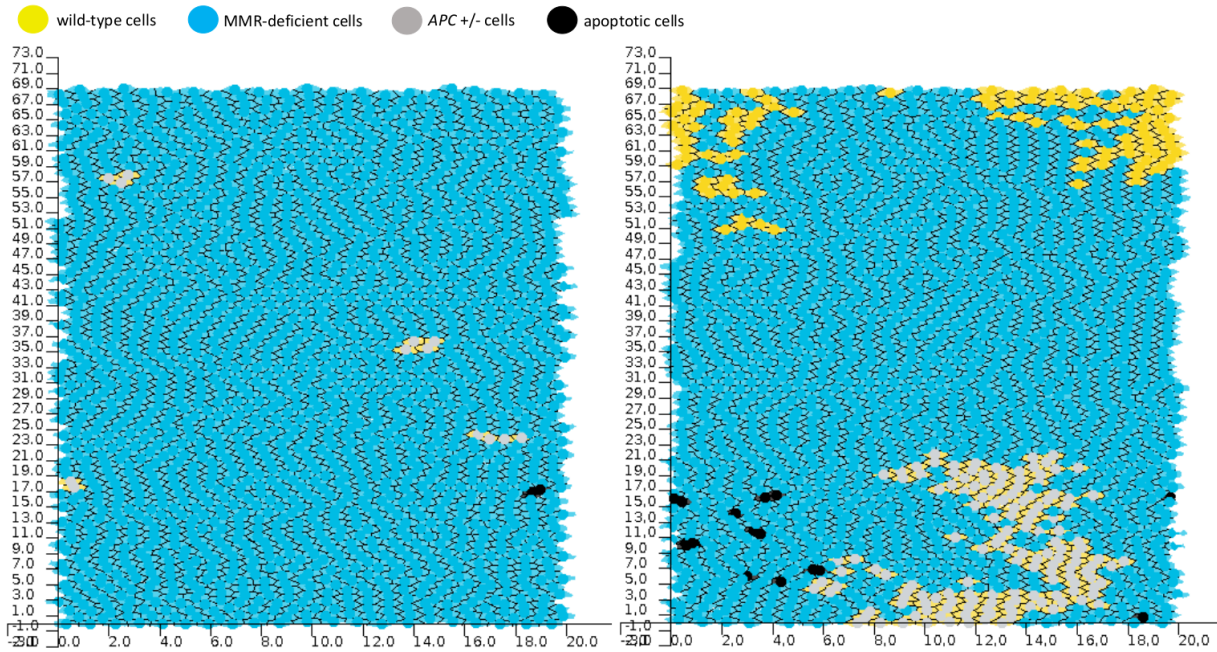
It might be feasible to assume that each monoclonal crypt will give rise to an aberrant lesion such as an adenomatous polyp. Importantly, this would directly imply that any biallelic *APC* or *CTNNB1* mutation occurring in a cell of the crypt results in the development of such a lesion, a hypothesis that has to be studied further in experiments.

### 6.3.3 THE INFLUENCE OF CELL LOCATION ON MUTATION SPREAD

[179]: Shahriyari et al. (2016), “The role of cell location and spatial gradients in the evolutionary dynamics of colon and intestinal crypts”.

The spread of a mutation is predicted to depend on the location of the cell in which the mutation first occurs [179]. Within our short-term simulations, a clear trend in favor of the crypt base could be observed, which will be analyzed in the following paragraphs.

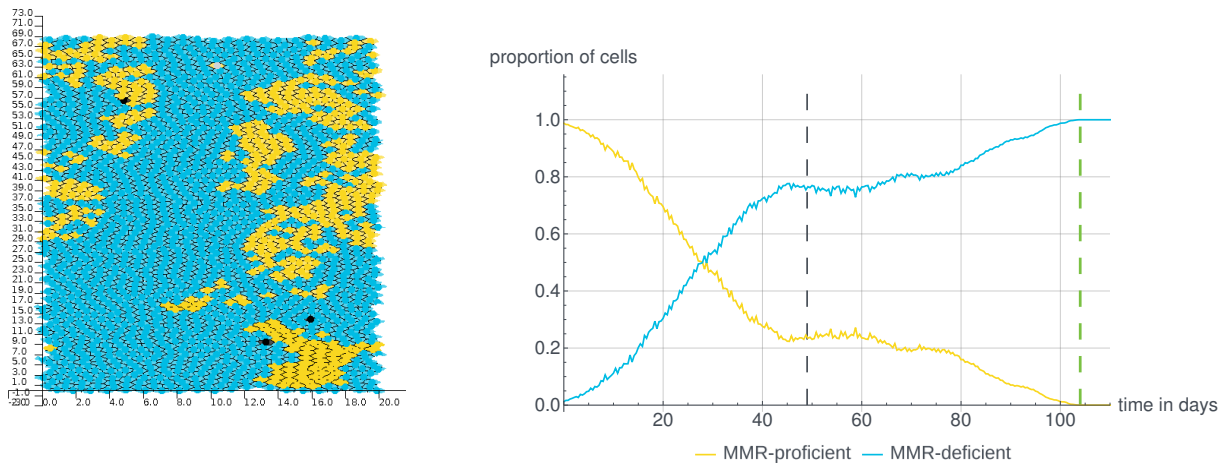
**The crypt base is the most stable environment within the crypt for mutated and wild-type cells.** Mutated clones which originate in the lowest region of the crypt usually have more time to expand without facing complete differentiation. Consequently, the ability that mutations, especially non-transforming ones, manage to spread throughout the crypt



**Figure 6.7:** Spread of APC mutations dependent on cell position. Video frames of two separate simulations are shown, both crypts are populated by an MMR-deficient stem cell by initialization. In all simulations, the following colors are used: MMR-deficient cells, MMR-proficient cells, monoallelically APC-mutated cells, biallelic APC-mutated cells with a black nucleus, biallelically *CTNNB1*-mutated cells and apoptotic cells. **Left:** Limited spread of APC mutations in an MMR-deficient crypt. Four small monoallelically APC-mutated clones are visible 64 days after the start of the simulation. All mutations occurred in the lower middle part of the crypt. However, the higher the APC mutations occur, the smaller are the resulting clones. **Right:** Successful spread of APC mutations in an almost MMR-deficient crypt. A monoallelic APC mutation occurred after 20 days in the lowest cell row of TA cells. This led to the formation and temporal persistence of a large monoallelic APC-mutated clone, here shown after 40 days. Videos of the complete simulations are provided online [SaskiaHaupt.de/phd-thesis#video4](http://SaskiaHaupt.de/phd-thesis#video4) an [SaskiaHaupt.de/phd-thesis#video5](http://SaskiaHaupt.de/phd-thesis#video5). Adapted from [81].

at all increases when decreasing the position of the initial mutated cell. In our *in silico* experiments, we observe this for stem cell mutations, which are located at the bottom of the crypt and in most cases become monoclonal. As an additional example, a cell which acquires a monoallelic APC mutation near the bottom of the crypt consistently gives rise to larger clones, compared to cells located higher in the crypt. Examples are illustrated in Figure 6.7.

However, the property of the crypt base being a stable environment is not limited to mutated cells. It can equally serve as a beneficial region for wild-type cells. In our simulations, a mutated stem cell only rarely does not give rise to monoclonal conversion. In this rare case, a clone of wild-type cells is able to persist for a sufficiently long period of time at the crypt base. This slows down the spread of the mutation, such that monoclonal conversion is either completed much later



(a) 49 days after an MMR-deficient stem cell starts populating the crypt.

(b) Proportion of MMR-proficient and -deficient cells over time.

**Figure 6.8: Prolonged monoclonal conversion of an MMR stem cell mutation.** Color legend and initial condition are as before. A video of the complete simulation is provided online [SaskiaHaupt.de/phd-thesis#video6](https://SaskiaHaupt.de/phd-thesis#video6). (a) After 49 days (gray dashed line), close to the average duration of monoclonal conversion of MMR deficiency (see Figure 6.4), a large clone of MMR-proficient TA cells still resides near the crypt base. Note that many other differentiated MMR-proficient cells also have not been washed out at this point in time. As can be seen in the online simulation video [SaskiaHaupt.de/phd-thesis#video6](https://SaskiaHaupt.de/phd-thesis#video6), three weeks later, the progeny of the same clone has lost the position near the crypt base and is subsequently washed out of the crypt. Monoclonal conversion was completed after 104 days (dashed green line). (b) The relative evolution of MMR-proficient and MMR-deficient cells over time. Note that the duration of monoclonal conversion is substantially longer compared to the average in Figure 6.4. Adapted from [81].

than on average, or never completed at all, explaining the outliers in Figures 6.4 and 6.6. An example of the former case is illustrated in Figure 6.8.

We observed those delays of monoclonal conversion at least once for monoallelic stem cell mutations in both *APC* and *CTNNB1*, as well as in an MMR-deficient stem cell. If monoclonal conversion of these mutations is prolonged until the end of the cell cycle of the stem cell which populates the crypt after a stem cell mutation, the process might not be completed due to a stem cell exchange.

**Top-down vs bottom-up morphogenesis is determined by the Wnt threshold.** During our simulations, we observed two ways of morphogenesis of biallelically *APC*-mutated cells: 1) The second hit occurs in parts of the crypt where wild-type cells already have completed differentiation. In this case, as a result of the direction of cell migration, the mutation spreads toward the top of the crypt before taking over the bottom half, see Figure 6.9, top. This is consistent with *top-down morphogenesis*, as discussed in Section 2.1.3, and can be

seen in the simulation video available online [SaskiaHaupt.de/phd-thesis#video7](https://www.saskiahaupt.de/phd-thesis#video7). 2) Due to the proliferative zone of the crypt being located in its lowest quarter, the majority of biallelic *APC* mutations are predicted to originate in the bottom region. This is known as *bottom-up morphogenesis*, illustrated in Figure 6.9, bottom, and in the simulation video provided online [SaskiaHaupt.de/phd-thesis#video8](https://www.saskiahaupt.de/phd-thesis#video8). In our simulations, the spread of a biallelically *APC*-mutated clone follows top-down morphogenesis in 22% of cases (9/41) and bottom-up morphogenesis in 66% of cases (27/41), while the remaining five cases could not be clearly identified as one or the other. Altogether, our *in silico* experiments show both modes of morphogenesis to be possible, whereby bottom-up morphogenesis was more frequent. These findings are in concordance with [30].

Based on our analyses, the proportion of biallelically *APC*-mutated crypts, which develop in a top-down manner, depends most notably on the Wnt threshold  $\tau_{\text{wnt}}$  for monoallelically *APC*-mutated cells. The lower this value, the later in the process of migration these cells differentiate and thus the higher the frequency of top-down monoclonal conversions. However, lowering the Wnt threshold also gives rise to more biallelically *APC*-mutated crypts overall.

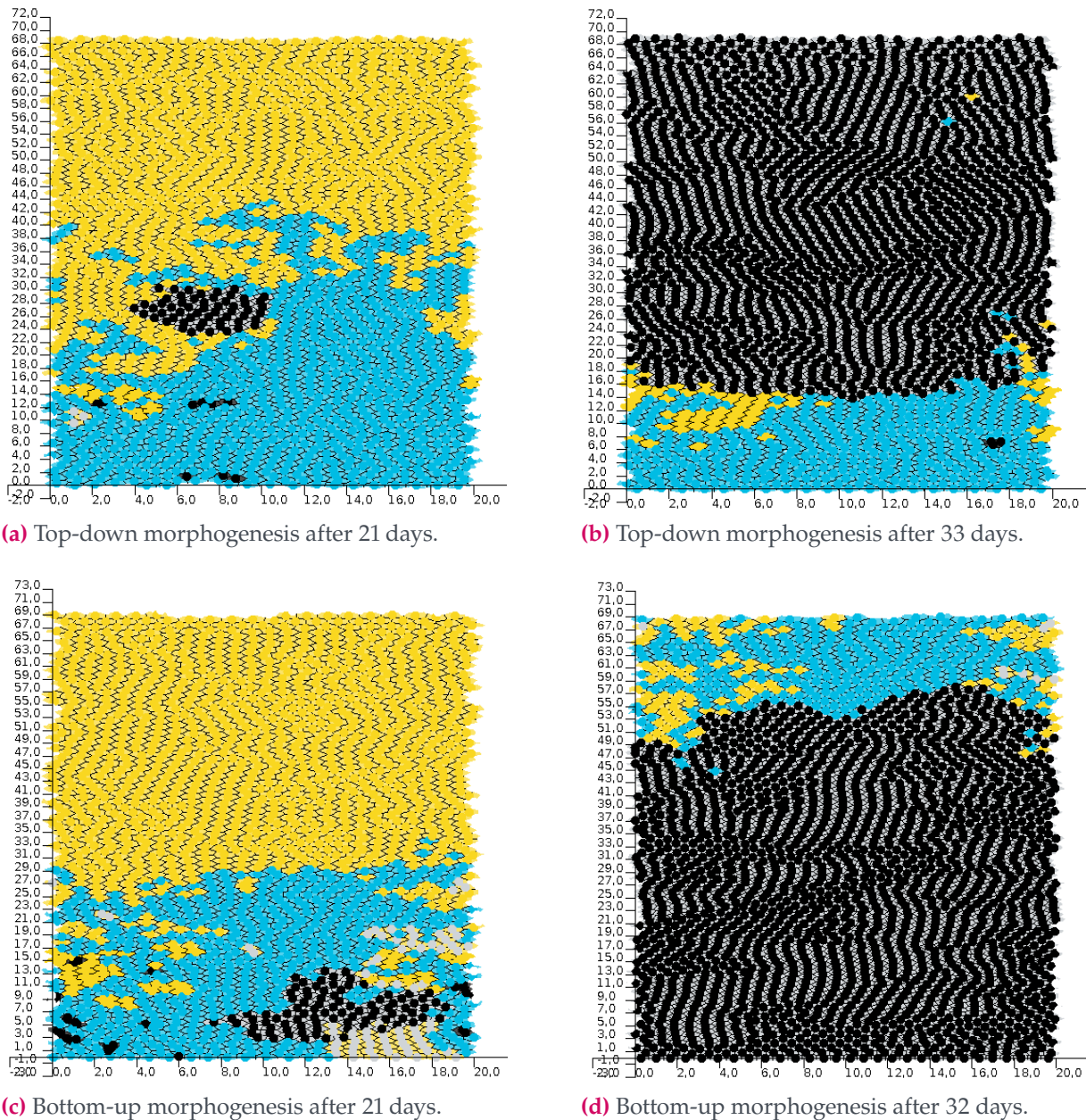
[30]: Bravo and Axelrod (2013), "A calibrated agent-based computer model of stochastic cell dynamics in normal human colon crypts useful for *in silico* experiments".

### 6.3.4 THE EFFECT OF STEM CELL EXCHANGE ON MONOCLONALITY

The effect of a stem cell exchange event on the monoclonality of a crypt depends on the respective mutation.

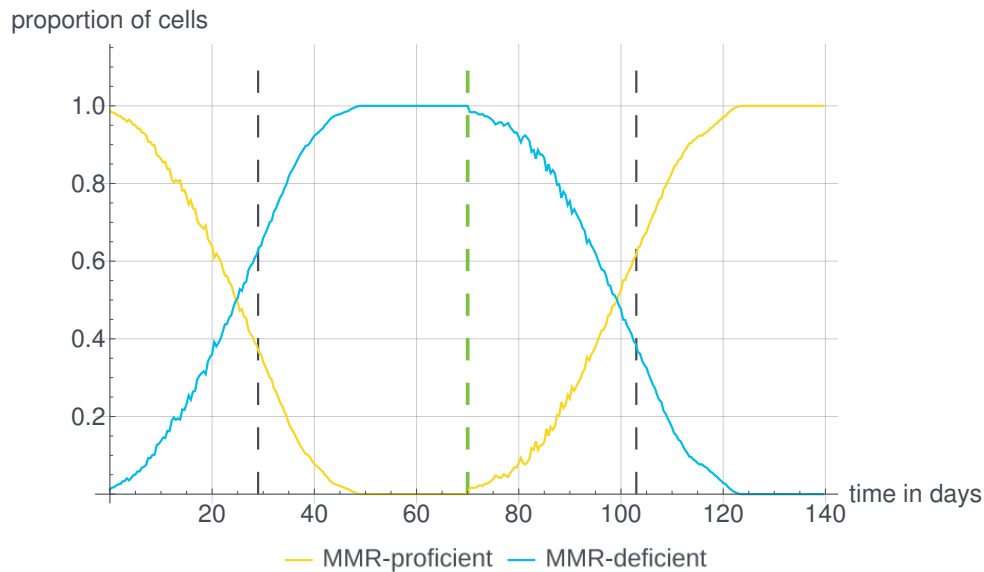
#### **Stem cell exchanges can lead to transient polyclonality.**

Our model simulations suggest that monoclonal biallelic MMR mutations are regularly washed out of the crypt after a stem cell exchange event. This process is illustrated in Figure 6.10. The observations are in concordance with our biological understanding, since MMR deficiency per se is not expected to provide a proliferative advantage to the cells and the transition to an alternative monoclonal status with transient polyclonality is likely to appear.

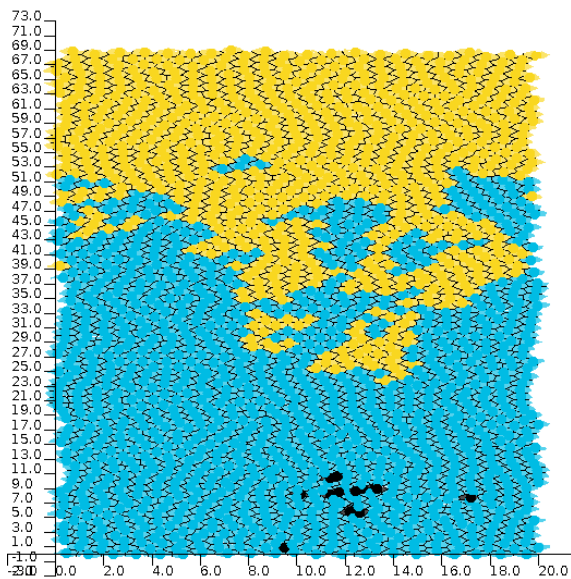


**Figure 6.9:** Different modes of morphogenesis of a biallelically *APC*-mutated crypt. Color legend and initial condition are as before. **Top:** Top-down morphogenesis. The second *APC* mutation occurred in the region of wild-type FD cells after 15 days. The biallelically *APC*-mutated clone consists of 56 cells after 21 days (a.) and of about 1800 cells 12 days later (b.). The upper parts of the crypt are populated first, before the bottom quarter. **Bottom:** Bottom-up morphogenesis. The second *APC* mutation occurred after 14.5 days in a cell within a monoallelically *APC*-mutated clone consisting of 24 cells, which can be seen below the biallelically *APC*-mutated clone. The latter consists of 90 cells after 21 days (c.) and of about 1900 cells 11 days later (d.). Videos of the complete simulations are provided online [SaskiaHaupt.de/phd-thesis#video7](http://SaskiaHaupt.de/phd-thesis#video7) and [SaskiaHaupt.de/phd-thesis#video8](http://SaskiaHaupt.de/phd-thesis#video8). Adapted from [81].

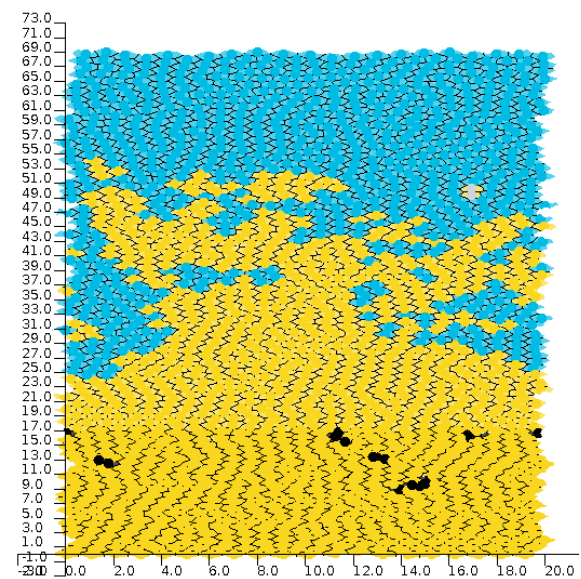




(a) Cyclic behavior of MMR-deficient and MMR-proficient cell numbers.



(b) Cellular composition after 29 days.



(c) Cellular composition after 103 days.

**Figure 6.10: Development and loss of MMR deficiency due to stem cell exchange.** (a) Initialization with an MMR-deficient stem cell leads to monoclonal conversion of MMR-deficient cells after 49 days, while the MMR-proficient cells are washed out of the crypt. After a first stem cell cycle of 70 days (green line), an MMR-proficient stem cell populates the crypt and the process of monoclonal conversion is reversed. After an additional 54 days, no MMR-deficient cells are left in the crypt. (b) and (c) show the cellular composition of the crypt after 29 days and 103 days, respectively. The time points are illustrated in a. by gray lines with colors as before. A video of the complete numerical simulation is provided online [SaskiaHaupt.de/phd-thesis#videol](https://SaskiaHaupt.de/phd-thesis#videol). Adapted from [81].

For monoclonal monoallelically *APC*-mutated or *CTNNB1*-mutated crypts, we observe similar scenarios. The underlying reason is again that the delay of differentiation caused by these mutations does not yield a significant advantage compared to the expansion of MMR-deficient or wild-type cells.

**Monoclonality can persist upon stem cell exchanges.** This is true for the monoclonality of biallelically *APC*-mutated or *CTNNB1*-mutated crypts. In our *in silico* experiments, the division of a wild-type stem cell populating the crypt after a stem cell exchange gives rise to a small population of wild-type TA cells. However, this population becomes extinct after 2–3 days. Since the biallelic mutations provide a significant proliferative advantage, wild-type cells are inferior to such cells and do not have the ability to recapture the crypt. Together with our observations from the previous sections, this implies that biallelic *APC* or *CTNNB1* mutations in any cell within the crypt result in the development of a monoclonal crypt, and that this crypt persists over time.

## 6.4 OUTCOMES AND DISCUSSION

We presented a computational model describing the evolution of colonic crypts in Lynch syndrome scenario to answer important questions for the very first steps of carcinogenesis. Besides studying the evolution of non-mutated crypts, we obtain insights into the evolution of crypts carrying possible driver mutations and initiating events of Lynch syndrome carcinogenesis.

By a suitable choice of parameters and further programming, the model could also be used to simulate the development of other types of colorectal cancer, analogously to what was done in Chapter 7. By initializing the crypts with no pathogenic germline variants or adapting the set of driver genes, Lynch-like colorectal carcinogenesis [43] and sporadic MSS colorectal carcinogenesis could be modeled. By initializing with a pathogenic germline variant in *APC*, the model could be used to model colorectal cancer initiation in FAP.

[43]: Carethers (2014), “Differentiating Lynch-Like From Lynch Syndrome”.

Concerning the dynamics of non-mutated tissue, the simulation results predicted that feedback mechanisms are necessary to ensure crypt homeostasis. However, the specific mechanism is currently not known from a biological point of view and our implementation (i.e., the death of small cells) should be regarded as exemplary, although the loss of cell integrity due to overwhelming pressure appears to be reasonable. Once further biological hypotheses are available, they can be included in the model.

Furthermore, our estimates obtained for both human and murine crypts suggest that the process of crypt renewal might take several weeks in humans, where we identified the magnitude of cell mobility and the crypt size as the two determining parameters. While the estimates are in concordance with experiments in mice, the *in silico* duration in humans is yet to be experimentally validated.

The reported predictions of time span required for a monoclonal conversion of a crypt carrying certain mutations provide the basis for future studies, in particular addressing the time required for an aberrant crypt to become an endoscopically visible lesion.

Our estimates for non-transforming mutations differ from the mean value of 18.6 days reported in [66]. However, in the latter study, the spread of a non-transforming mutation throughout a crypt consisting of 250 cells was examined with a significantly shorter average stem cell cycle duration of 24 hours compared to a few weeks in our approach. These factors may explain the difference between the estimates.

We further discussed the influence of stem cell dynamics on monoclonality. With our simulations, stem cell exchange can restore the integrity of crypts and contribute to the elimination of mutations without a directly transforming effect, including MMR gene mutations. This suggests that it serves as a mechanism which can inhibit carcinogenesis. Furthermore, the results of our simulations of MMR-deficient crypts may help explain why many MMR-deficient crypts do not progress to larger lesions: Such crypts might be detectable via staining, but frequently lose their status due to stem cell exchange later on. However, it is important to note that the monoclonality might be regained, in case the mutated stem cell populates the crypt again. This pattern of loss and recovery of monoclonality might occur repetitively,

[66]: Fletcher et al. (2012), "Mathematical modeling of monoclonal conversion in the colonic crypt".

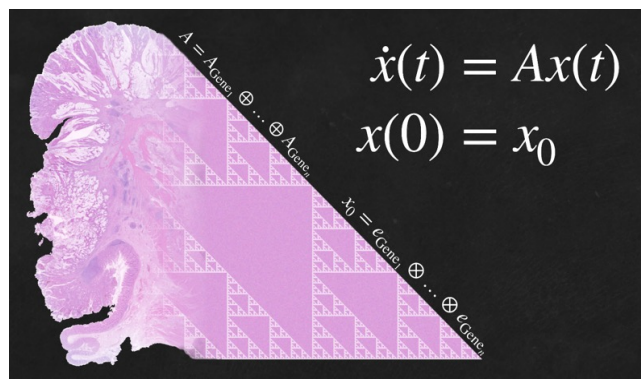
as long as no further advantageous mutations occur, and the mutated stem cell does not die. As in Lynch syndrome, a cell only needs a second hit to be MMR-deficient, this finding is of particular importance. MMR-deficient crypts are much more likely to occur but not all of them might evolve into a carcinoma later due to the stem cell exchange mechanisms.

Although our approach takes into account many different processes and mechanisms occurring within colonic crypts, several aspects are neglected or not modeled in great detail. While we concentrated on the Wnt pathway with high evidence regarding its role in colorectal cancer, the implementation of other signal gradients like Eph/ephrin and Notch signaling is possible with the presented model framework and is due to future work. Further, as we focused on the initiating events of Lynch syndrome carcinogenesis, mutations considered to occur at more advanced stages and/or less frequently in Lynch syndrome carcinogenesis, such as *KRAS* or *TP53*, were not implemented in our model, although in principle our model would allow the implementation of additional mutational events.

We carefully evaluated the level of complexity with the computational costs of our model regarding computation time and storage. Future work shall add further modeling complexity where needed while trying to not increase the computational expense to allow for long-term simulations. For example, this could include a dynamic cell cycle model, a detailed model for feedback mechanisms and further, possibly yet unidentified mutations frequently occurring in colorectal carcinogenesis.

In summary, our results provide first mathematical clues for effective surveillance protocols for Lynch syndrome carriers. First, the duration of monoclonal conversion for different driver events is a first hint for the duration of single steps in Lynch syndrome carcinogenesis. By long-term simulations, we will address the question when these events happen and by this, our estimates could contribute to the question of surveillance intervals. Second, the possibility of disappearing MMR-deficient crypts is currently a medical hypothesis linked to the question of overdiagnosis in Lynch syndrome [177]. During our simulations, we observed this scenario supporting the medical hypothesis. Future work will include more simulations for analyzing this effect in more detail.

[177]: Seppälä et al. (2019), "Lack of association between screening interval and cancer stage in Lynch syndrome may be accounted for by over-diagnosis: a prospective Lynch syndrome database report".



## 7 MATHEMATICALLY MODELING LYNCH SYNDROME COLORECTAL CARCINOGENESIS USING THE KRONECKER STRUCTURE

In this chapter, we present our developed mathematical model for Lynch syndrome colorectal carcinogenesis at the crypt level with focus on the concept of multiple pathways of carcinogenesis. As described in Section 2.3.2, in Lynch syndrome, one of the central medical hypotheses is the three pathway hypothesis of colorectal Lynch syndrome carcinogenesis. As different pathways of carcinogenesis are linked to different molecular features and thus need tailored prevention, diagnosis, and treatment approaches, understanding these pathways of carcinogenesis in more detail, determining unique characteristics, and quantifying the relative distribution of these pathways among Lynch syndrome individuals are essential. We thus aim at developing a mathematical model that

- ▶ offers a **simultaneous description** of the multiple pathways of colorectal carcinogenesis in Lynch syndrome,
- ▶ is **medically interpretable** in the sense that the parameters and model components have a biomedical meaning,
- ▶ can be **analyzed systematically** meaning that the influence of different components can be examined in a mathematically rigorous way,
- ▶ is **computationally feasible**, i.e. also large model systems can be computed efficiently,
- ▶ is **modular** such that we can add and remove driver genes, mutational dependencies, and whole pathways of carcinogenesis easily.

7.1	CURRENT MEDICAL HYPOTHESES . . . . .	117
7.2	MODELING WITH THE KRONECKER STRUCTURE . . . . .	119
7.3	MODIFICATIONS FOR OTHER CANCER TYPES . . . . .	138
7.4	CALIBRATION AND VALIDATION RESULTS	140
7.5	OUTCOMES AND DISCUSSION . . . . .	146

## MODELING WORKFLOW

### 1 MEDICAL KNOWLEDGE

- ▶ define pathways of carcinogenesis
- ▶ identify driver genes
- ▶ explore mutational dependencies

### 2 PARAMETER VALUES

- ▶ define gene-dependent point mutation and LOH event rates
- ▶ determine possible fitness changes and fixation affinities

### 3 GRAPH REPRESENTATION

- ▶ build gene mutation graphs for each driver gene
- ▶ build graphs for mutational dependencies

### 4 ADJACENCY MATRICES

- ▶ derive adjacency matrices corresponding to graphs using the Kronecker structure

### 5 LINEAR ODE SOLUTION

- ▶ set initial condition
- ▶ solve linear ODE explicitly using the matrix exponential
- ▶ extract mutational status of interest from the solution vector

**Figure 7.1:** Overview of the general modeling workflow for modeling multiple pathways of carcinogenesis using the Kronecker structure. In this chapter, each step of the workflow will precisely be defined and illustrated at the example of Lynch syndrome colorectal carcinogenesis.

We develop a general mathematical framework that fulfills these requirements and thus, can describe arbitrarily complex pathway networks of carcinogenesis and arbitrary numbers of mutations and pathways of carcinogenesis. The proposed framework is based on a linear dynamical system using the Kronecker structure for the system matrix, illustrated in Figure 7.1.

We use Lynch syndrome colorectal carcinogenesis to illustrate the applicability of the model. From the model solution vector, we extract specific mutational status describing defined precursor lesions like MMR-deficient crypts, early and late adenomatous, as well as cancerous states to obtain an age-dependent evolution of the number of crypts of specific mutational status within a Lynch syndrome individual which is hardly possible with current medical data. Further, we can analyze the influence of different MMR gene variants to further support gene-dependent clinical decision-making. The influence of other medical parameters is studied to better understand cancer development and how it is currently re-

flected by the mathematical model. In addition, first results to quantify the distribution among the three pathways of Lynch syndrome-associated colorectal carcinogenesis depending on the patient's age are derived. In the future, this could help to test the effectiveness of different prevention and treatment schemes on a population level.

The general modeling framework with the Kronecker structure used in the system matrix can be used to describe other types of cancer. Naturally, specific assumptions on the mutations and pathways of carcinogenesis may vary for other types of cancer. We illustrate model modifications for FAP, Lynch-like, and the classical sporadic MSS colorectal carcinogenesis, as well as the possibility to apply the model structure to carcinogenesis in other organs.

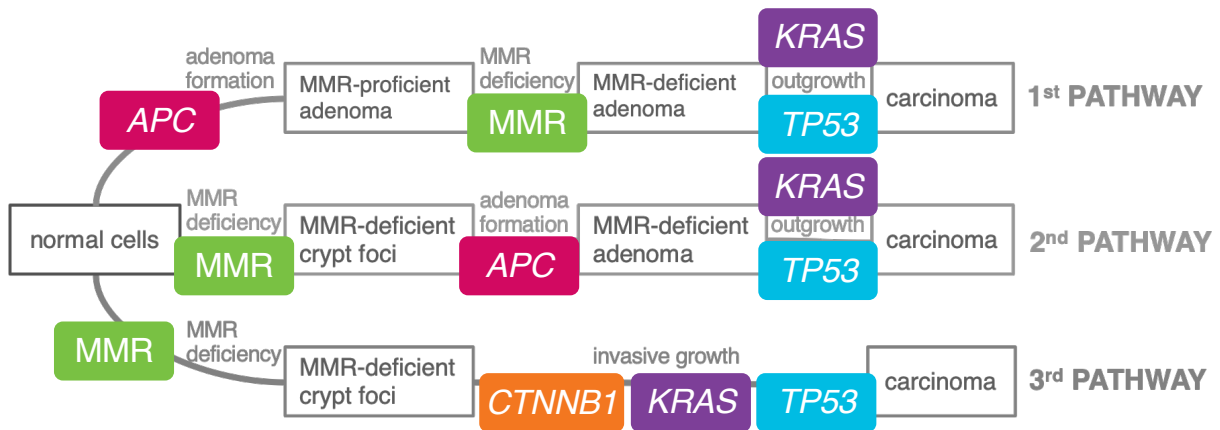
This chapter is based on our work previously published in PLOS Computational Biology [82]. A blog post for the general Mathematical Oncology community was published on the official MathOnco blog [mathematical-oncology.org/blog/modeling-carcinogenesis-using-kronecker-structure.html](https://mathematical-oncology.org/blog/modeling-carcinogenesis-using-kronecker-structure.html). Besides that, the presented work was the inspiration for the Cover Art of the official Mathematical Oncology newsletter, Week 166, which also serves as the cover image for this chapter.

[82]: Haupt et al. (2021), "Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure".

## 7.1 CURRENT MEDICAL HYPOTHESES ABOUT MUTATIONAL EVENTS IN MULTIPLE PATHWAYS OF LYNCH SYNDROME COLORECTAL CARCINOGENESIS

As explained in Section 2.3.2, colorectal cancer in Lynch syndrome is currently hypothesized to develop via three different pathways of carcinogenesis which we wanted to describe simultaneously with a mathematical model. Ahadova et al. [2] showed that these pathways of carcinogenesis are linked to different mutational processes, e.g., *CTNNB1*-mutated colorectal carcinomas are associated with immediate invasive growth, following the third presented pathway. Further, *APC* is linked to a polypous growth and the formation of adenomas, following pathway 1 or 2. The mutation-pathway associations are illustrated in Figure 7.2.

[2]: Ahadova et al. (2018), "Three molecular pathways model colorectal carcinogenesis in Lynch syndrome".



**Figure 7.2:** Pathways of carcinogenesis in Lynch syndrome-associated colorectal cancer are linked to different mutational events. While we differentiate pathways 1,2 from 3 by the involved driver genes, i.e., mutational events in *APC* or *CTNNB1*, for a distinction between pathway 1 and 2 the order of mutations seems to play an important role, having either *APC* inactivation (pathway 1) or *MMR* deficiency (pathway 2) as initiating event of carcinogenesis.

[3]: Ahadova et al. (2020), “The unnatural history of colorectal cancer in Lynch syndrome: Lessons from colonoscopy surveillance”.

[63]: Engel et al. (2018), “No Difference in Colorectal Cancer Incidence or Stage at Detection by Colonoscopy Among 3 Countries With Different Lynch Syndrome Surveillance Policies”.

[64]: Engel et al. (2020), “Associations of Pathogenic Variants in *MLH1*, *MSH2*, and *MSH6* With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome”.

However, the relative distribution of patients among these pathways of carcinogenesis is an important open question in current Lynch syndrome research with significant clinical implications (see also Section 2.5): Recent independent studies (analyzed in [3]) demonstrated that a substantial proportion of Lynch syndrome individuals develops colorectal cancer despite regular colonoscopy. Further, there is no difference in colorectal cancer incidence or stage at detection by colonoscopy with respect to different Lynch syndrome surveillance intervals [63]. Besides that, *MMR* gene-dependent differences are observed regarding the risk of colorectal adenomas and carcinomas, and regarding somatic mutations in patients with Lynch syndrome [64]. The mathematical model should be able to reflect these molecular characteristics. Further, it should be easy to change these characteristics for different simulations to test their consequences for adapted clinical prevention and treatment approaches, where the latter requires that a rigorous analysis is possible.

To address the point of medical interpretability on a genome level, the probability of mutational events, like point mutations or LOH events, was modeled in a gene-dependent way, as depicted in Chapter 4. Further, we considered all mutational events to be in a network of mutations that reflects the relations between distinct driver mutations. As a baseline, we assumed all mutations to be independent of each other, which is either due to independence indicated by medical data or due to missing medical insight suggesting otherwise.



However, the considered mutations relevant for cancer development by definition might change the functional behavior of a cell. In particular, some specific mutations affect the probability of certain other mutations. In other words, there are mutations which are mutually exclusive or mutations which increase the probability of mutations in other genes [120]. For the presented approach, we assumed and modeled the following mutational dependencies [82]:

- ▶ **Increased point mutation rate of APC after MMR deficiency.** MMR deficiency leads to an increased mutation rate, especially in microsatellites [53]. Among others, this is true for the point mutation rate of *APC*.
- ▶ **LOH event affecting CTNNB1 and MLH1 simultaneously.** According to [64], somatic *CTNNB1* mutations are significantly higher in *MLH1*-cancers than in the other MMR gene-associated colorectal cancers.
- ▶ **Increased LOH event rate after APC inactivation.** An increased LOH event rate of *APC*-inactivated crypts is assumed to be the case in many cancers [153].
- ▶ **Enhancement of the last two effects.** *APC* inactivation increases the LOH event rate of other genes, including *MLH1*. Further, there is a positive association of *MLH1* and *CTNNB1* alterations due to LOH events affecting those genes simultaneously. Thus, the first effect enhances the second effect.
- ▶ **Increased mutation rate of KRAS after MMR deficiency.** Further, *KRAS* is an oncogene with one point mutation sufficient for activation, where mainly codon 12 or 13 are hit. Codon 13 mutations are known to be associated with and enriched in MMR-deficient cancers, as these mutations are more likely to occur under the influence of MMR deficiency [2].

[120]: Leiserson et al. (2015), “CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer”.

[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

[53]: de la Chapelle (2003), “Microsatellite Instability”.

[64]: Engel et al. (2020), “Associations of Pathogenic Variants in *MLH1*, *MSH2*, and *MSH6* With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome”.

[153]: Nowak et al. (2002), “The role of chromosomal instability in tumor initiation”.

[2]: Ahadova et al. (2018), “Three molecular pathways model colorectal carcinogenesis in Lynch syndrome”.

## 7.2 MODELING LYNCH SYNDROME COLORECTAL CARCINOGENESIS USING THE KRONECKER STRUCTURE

We introduce our model for colorectal carcinogenesis in Lynch syndrome. The model consists of a dynamical system given in the form of a linear ordinary differential equation. The choice of the system matrix  $M$  is crucial to the approach.

This matrix is built additively using adjacency matrices of gene mutation graphs describing the joint process of mutations in several genes, including mutations independent of and depending on other mutations. The additive structure of the system matrix  $M$  underlines the model's medical interpretability. All mutations are assumed to be present in the whole crypt, clearly distinguishing this approach from the cell-based crypt model described in Chapter 6. Mutations which occur in one cell but are washed out as they reach the top of the crypt and undergo apoptosis are not considered in the model. Therefore, the model concentrates on crypts as the smallest unit of interest, rather than the cells.

We introduce a matrix  $A$  for the independent processes and matrices  $B, C, D, E$  and  $F$  for the dependent processes, i.e.,  $M = A + B + C + D + E + F$ . All the latter are based on three main assumptions leading to the Kronecker sum and product as underlying structures of the matrices in a natural way. The following chapter is mainly based on [82].

[82]: Haupt et al. (2021), "Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure".

### 7.2.1 DEFINING GENE MUTATION GRAPHS

We want to find a graph representation of the system matrix describing all mutational status and gene alterations that are possible in the process of carcinogenesis from wild type to cancerous crypts. Therefore, mutational status and the alteration rates between them have to be represented. This can be done with the help of graphs  $G = (\mathcal{V}, \mathcal{E})$ , i.e., a set  $\mathcal{V}$  of mutation status (vertices) which are combined with edges  $\mathcal{E}$  if an alteration (point mutation or LOH event) between the status is possible. To be precise, we connect the mutation status that differ by only one alteration, namely one point mutation or LOH event. This means we assume that only one alteration happens at any specific time point. We call the resulting graphs gene mutation graphs and define them for each involved driver gene or involved combined mutational process separately. For the mutation status of the genes, we distinguish between tumor suppressor genes where usually two hits are required for a phenotypic change, and oncogenes where usually one hit is sufficient for activation (see Section 2.1.2). In addition, the rates for point mutations and LOH events are considered to be gene-dependent as previously introduced in Chapter 4.

For the resulting dynamical system, we represent each graph as an adjacency matrix  $A_G \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ , that is, a matrix which has as many rows and columns as there are vertices in the graph. The entry  $a_{i,j} \in \mathbb{R}$  at position  $(i, j)$  of the matrix  $A_G$  indicates whether there is no possible alteration ( $a_{i,j} = 0$ ), a low rate of alterations ( $|a_{i,j}|$  is small) or a high rate of alterations ( $|a_{i,j}|$  is large) between the vertices  $i$  and  $j$ . The value  $a_{i,j}$  is often referred to as the weight of the edge  $ij$ . In general, vertices  $i$  and  $j$  which are connected by an edge, i.e., for which  $a_{i,j} \neq 0$  are called adjacent and are denoted by  $i \sim j$ . Vertices may also be connected to themselves, meaning that  $a_{i,i} \neq 0$ . In this case, the edge is called a self-loop and may model fitness advantages and disadvantages of mutational status, as explained in more detail below. In our setting, the edges are directed because the alteration happens in a defined direction. This is done by letting  $a_{i,j}$  (weight of the edge from  $i$  to  $j$ ) differ from  $a_{j,i}$  (weight of the edge from  $j$  to  $i$ ). A directed graph with no directed cycles is called a directed acyclic graph (DAG). This means we make the assumption that once a mutation happened it cannot be reversed by another mutation. In this case, the vertices of a DAG can be ordered such that the adjacency matrix is an upper triangular matrix which reduces the computational costs of the solving process.

In detail, for Lynch syndrome colorectal carcinogenesis, we consider the following driver genes: the MMR gene with the inherited germline variant, *CTNNB1*, *APC*, *KRAS*, and *TP53*, as those are typical representatives of the oncogenes and tumor suppressor genes affected in the corresponding pathways of Lynch syndrome-associated colorectal carcinogenesis.

We use the mutation status notation for single genes introduced in Section 6.1.3. We assume that 11 in *CTNNB1*, *APC*, and *TP53* is incompatible with cell survival [160], as already done for *CTNNB1* and *APC* in Chapter 6. As we model the evolution of genotypic states of crypts (not cells), we do not consider the 11 status for *CTNNB1*, *APC*, and *TP53*.

Further, we assume for the driver genes in Lynch syndrome-associated colorectal carcinogenesis:

- **MMR:** Two hits are necessary for inactivation. In Lynch syndrome, one germline variant is present in all cells and thus all crypts.

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

[12]: Arnold et al. (2020), “The majority of  $\beta$ -catenin mutations in colorectal cancer is homozygous”.

[89]: Huels et al. (2015), “E-cadherin can limit the transforming properties of activating  $\beta$ -catenin mutations”.

[107]: Knudson (1971), “Mutation and Cancer: Statistical Study of Retinoblastoma”.

[56]: Dihlmann et al. (1999), “Dominant negative effect of the APC1309 mutation: a possible explanation for genotype-phenotype correlations in familial adenomatous polyposis”.

- ▶ **CTNNB1**: Two hits are assumed to be necessary to mediate an oncogenic driver effect as recent data showed [12, 89]. These two hits in *CTNNB1* are one of the transforming mutations we analyzed in Chapter 6.
- ▶ **APC, TP53**: Two hits are necessary for inactivation, as both are assumed tumor suppressor genes which dates back to Knudson et al. in 1971 [107]. In particular, we ignore a possibly dominant-negative effect of *APC* and *TP53* mutations resulting in a single hit necessary for inactivation [56].
- ▶ **KRAS**: One hit is necessary for activation as it is assumed to be a classical oncogene.

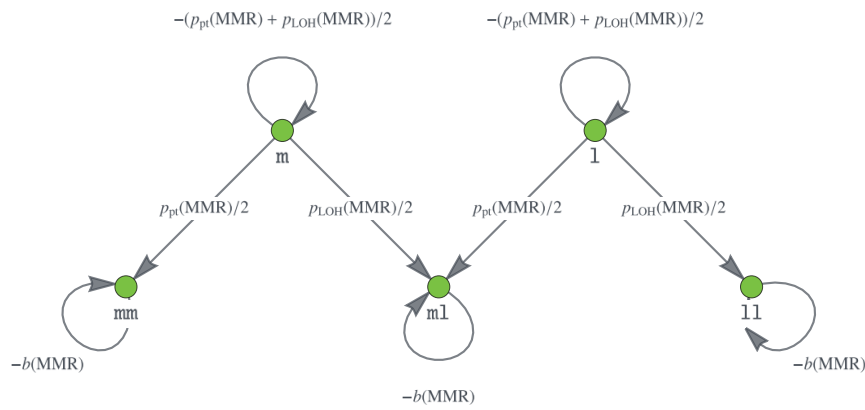
Altogether this leads to the vertex sets

$$\begin{aligned}\mathcal{V}_{\text{MMR}} &= \{m, l, mm, ml, ll\}, \\ \mathcal{V}_{\text{CTNNB1}} &= \{\emptyset, m, l, mm, ml\}, \\ \mathcal{V}_{\text{APC}} &= \{\emptyset, m, l, mm, ml\}, \\ \mathcal{V}_{\text{KRAS}} &= \{\emptyset, m\}, \\ \mathcal{V}_{\text{TP53}} &= \{\emptyset, m, l, mm, ml\}.\end{aligned}$$

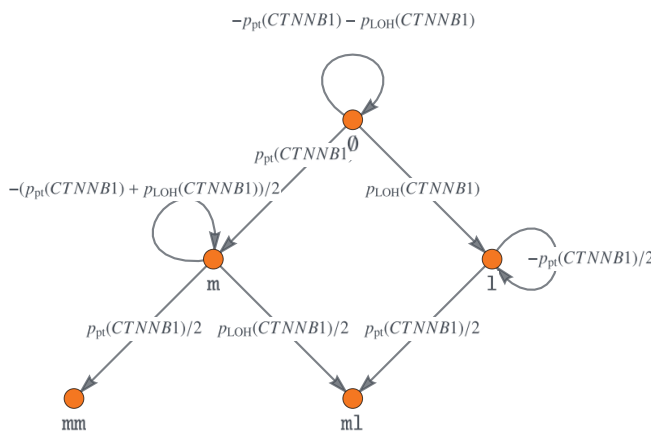
The resulting graphs for the individual genes are illustrated in Figure 7.3, whereby the definitions for the MMR genes, *CTNNB1* and *APC* are in concordance with those in Chapter 6. In Figure 7.3, we also display the edge weights of each gene mutation graph, i.e., the likelihood that we transfer from one mutation status to another. The exact parameter values for the gene-dependent alteration rates will be explained in the following section.

## 7.2.2 ESTIMATES FOR POINT MUTATION AND LOH EVENT RATES PER CRYPT PER YEAR

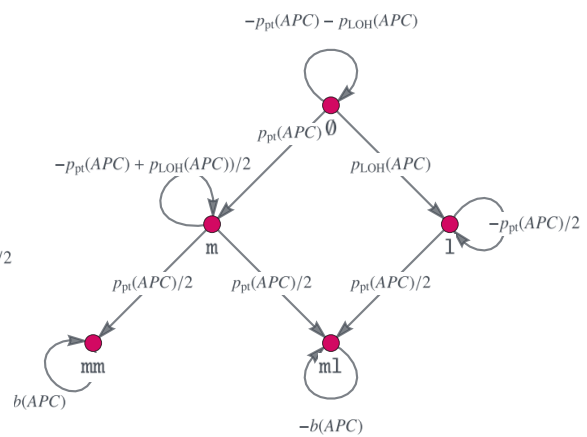
As illustrated in Chapter 4, point mutation and LOH event rates per cell division for individual cells are modeled depending on the length of the corresponding gene portion. To model the likelihoods  $\tilde{p}_{\text{pt}}(\text{gene})$  and  $\tilde{p}_{\text{LOH}}(\text{gene})$  for crypts being affected by point mutations and LOH events, respectively, in a specific gene, we have to convert the rates accordingly.



(a) MMR gene mutations.



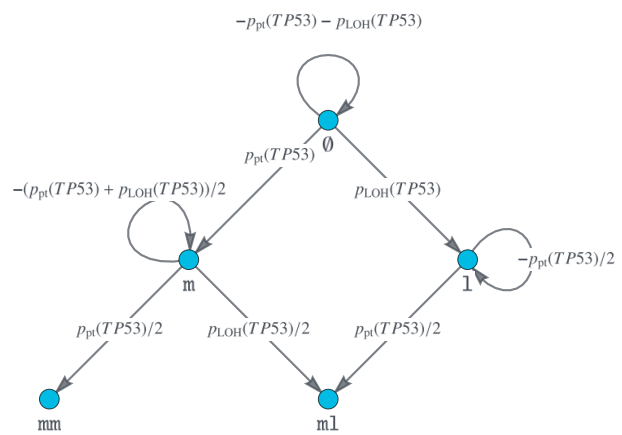
(b) CTNNB1 mutations.



(c) APC mutations.



(d) KRAS mutations.



(e) TP53 mutations.

**Figure 7.3: Gene mutation graphs.** These graphs represent the possible mutation states, i.e. which mutations the alleles of the gene can have accumulated, as vertices  $\emptyset$ ,  $m$ ,  $1$ ,  $mm$ ,  $11$  and  $m1$ . The edges connecting different vertices represent mutations, whereas self-loops, i.e. edges that connect a vertex with itself, describe no mutation occurring at the current point in time. The edges are labeled by the amount of change which happens at each point in time.

Further, as we want to model the evolution of crypts over years but many measurements and estimates are given in days, we use the factor 365 to convert the measurements per day to measurements per year.

For the parameters in Definition 4.1, we assume to accumulate  $n_{\text{pt}} = 1.2$  point mutations per cell division [206], where a cell division is assumed to take one day [51]. There are  $n_{\text{bp, genome}} = 3.2 \cdot 10^9$  base pairs on the genome.

Each crypt is estimated [15] to consist of approximately  $1.7 \cdot 10^3$  to  $2.5 \cdot 10^3$  cells, whereas only approximately 75% of them can divide. Thus, we use  $n_{\text{cells}} = 1500$  as an approximation to the number of dividing cells per crypt.

Not all point mutations which appear in a crypt take over the entire crypt [151]. We model this with a gene-dependent fixation affinity  $f(\text{gene})$ , i.e., the tendency of a cell with a mutation in a gene to take over the whole crypt.

#### Definition 7.1 Alteration rates in crypts

The above assumptions together with Definition 4.1 lead to the following formula for the likelihood  $\tilde{p}_{\text{pt}}(\text{gene})$  of point mutations per crypt per year

$$\tilde{p}_{\text{pt}}(\text{gene}) = 365 \cdot n_{\text{cells}} \cdot f(\text{gene}) \cdot p_{\text{pt}}(\text{gene}).$$

The likelihood  $\tilde{p}_{\text{LOH}}(\text{gene})$  of LOH events per crypt per year is defined similarly by

$$\tilde{p}_{\text{LOH}}(\text{gene}) = 365 \cdot n_{\text{cells}} \cdot f(\text{gene}) \cdot p_{\text{LOH}}(\text{gene}).$$

In general, medical data is hardly measurable to set precise values of these parameters, in particular for the fixation affinities. Thus, we will calibrate the model with these parameters such that the model results are quantitatively comparable to current clinical data (see Section 7.4).

### 7.2.3 FITNESS ADVANTAGES AND CLONAL EXPANSION MODELED BY SELF-LOOPS IN THE GRAPH

There is the possibility of introducing fitness changes described by the numerical value  $b(\text{gene})$  for individual mutation status of a gene. As we model the evolution of mutations at the crypt level, this corresponds to the clonal

[206]: Werner et al. (2019), "Measuring single cell divisions in human cancers from multi-region sequencing data".

[51]: Cooper (2018), *The Cell: A Molecular Approach*. 8th edition.

[15]: Baker et al. (2014), "Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon".

[151]: Nicholson et al. (2018), "Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium".

expansion of the crypts with one of the considered mutations. A fitness advantage is ensured by  $b(\text{gene}) > 0$  and a disadvantage with  $b(\text{gene}) < 0$ . By using the notion of graphs, this corresponds to a self-loop of the respective genotypic state node with a weight equal to the fitness change. We assume that MMR deficiency leads to a fitness disadvantage [68], i.e.,  $b(\text{MMR}) < 0$ , and APC inactivation and KRAS activation lead to a fitness advantage, i.e.,  $b(\text{APC}) > 0$  and  $b(\text{KRAS}) > 0$ , in concordance with recent measurements [14, 151] and other modeling approaches [160]. We set  $b(\text{MMR}) = -0.01$ ,  $b(\text{APC}) = 0.10$ ,  $b(\text{KRAS}) = 0.01$ .

In other words, the proliferation and disappearance of certain genotypic states is jointly modeled by the self-loops in the graph. This largely reduces the number of probability parameters necessary to be determined, accounting for the fact that there were not enough prospective data available to estimate or learn all the parameters. However, once there are enough data available, this assumption can be relaxed and additional states for dead or disappearing lesions can be introduced (see [82, Supplementary material]).

#### 7.2.4 COMBINATION OF GENE MUTATION GRAPHS USING THE KRONECKER STRUCTURE

We defined the gene mutation graphs for each involved driver gene independently. The next step is to combine these graphs to model the occurrence of mutations in different genes of the same crypt which we want to represent as one single process. For this combination, we make the following key assumptions:

- ▶ **Existence of states:** The states in the combined graph should exactly be all possible combinations of states from the underlying graphs. This means that all combinations of mutations in the different genes are possible, no mutations are prevented by other mutations and there are no additional states. This also implies that the order in which mutations are accumulated is ignored.
- ▶ **Edge connectivity:** We require that in the combined graph only one mutational event can happen at any point in time. In other words, no two alterations can occur at the exactly same point in time.

[68]: Galeota-Sprung et al. (2019), “The fitness cost of mismatch repair mutators in *Saccharomyces cerevisiae*: partitioning the mutational load”.

[14]: Baker and Graham (2016), “Quantifying human intestinal stem cell and crypt dynamics: the implications for cancer screening and prevention”.

[151]: Nicholson et al. (2018), “Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium”.

[160]: Paterson et al. (2020), “Mathematical model of colorectal cancer initiation”.

[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

- **Independence of the mutational events** We require that the mutational events are independent of each other. This entails that one mutational event does not change the probability for other mutational events.

**Theorem 7.2** Cartesian graph product

Consider two mutational processes that are represented by the gene mutation graphs  $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$  and  $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ . The combined process, which satisfies our key assumptions, is then represented by a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , that is given by the Cartesian product [78], denoted by a small square  $\square$ , of the graphs  $\mathcal{G}_1$  and  $\mathcal{G}_2$

[78]: Hammack et al. (2011), *Handbook of Product Graphs*.

$$\mathcal{G} = \mathcal{G}_1 \square \mathcal{G}_2.$$

*Proof.* The set of vertices  $\mathcal{V}$  of the Cartesian graph product is given by the Cartesian product  $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$  of the vertex sets  $\mathcal{V}_1$  and  $\mathcal{V}_2$ . This means the vertex set  $\mathcal{V}$  consists of all possible combinations of vertices from the first graph with vertices from the second graph

$$\mathcal{V} = \{(v_1, v_2) | v_1 \in \mathcal{V}_1 \text{ and } v_2 \in \mathcal{V}_2\}.$$

In other words, the first requirement is satisfied.

The edge set  $\mathcal{E}$  is made up of edges of the forms

$$(v_1, v_2) \sim (w_1, v_2), \quad (v_1, v_2) \sim (v_1, w_2),$$

where the vertices  $v_1, w_1 \in \mathcal{V}_1$  are adjacent in the graph  $\mathcal{G}_1$  and similarly for  $v_2, w_2 \in \mathcal{V}_2$ . This means that we connect the states in our combined process such that each edge corresponds to a single edge in exactly one of the underlying processes. Thus, our second assumption is also fulfilled.

As each edge in  $\mathcal{E}$  corresponds to exactly one of the edges in  $\mathcal{E}_1 \cup \mathcal{E}_2$ , we can transfer the edge weights from the graphs  $\mathcal{G}_1$  and  $\mathcal{G}_2$  to  $\mathcal{G}$ . From this, we conclude the satisfaction of the independence assumption.  $\square$

This combination of two gene mutation graphs can be extended to more than two mutation graphs by iteratively applying the theorem to two of the processes.

The last two key assumptions are clearly true for the independent mutational events. However, they can also be applied to the dependent mutational events in the following sense: Although there are dependencies between two mutational



events, the combination of these processes is again independent of the other mutational events. For example, the increased point mutation rate in *APC* after MMR deficiency is independent of the mutation status of the remaining genes *CTNNB1*, *KRAS*, and *TP53*. Thus, we build mutation graphs for the dependent mutational events and combine them with the other genes using the Cartesian graph structure, which will be explained in more detail for each dependency below.

The next step is to relate the Cartesian product of graphs to their adjacency matrices, where the Kronecker product and Kronecker sum of matrices play a crucial role.

**Definition 7.3** Kronecker product and Kronecker sum

The Kronecker product  $A \otimes B \in \mathbb{R}^{mp \times nq}$  of two matrices  $A \in \mathbb{R}^{m \times n}$  and  $B \in \mathbb{R}^{p \times q}$  is defined [87, 124] by the block matrix

$$A \otimes B = \begin{pmatrix} a_{1,1}B & \dots & a_{1,n}B \\ \vdots & \ddots & \vdots \\ a_{m,1}B & \dots & a_{m,n}B \end{pmatrix}.$$

For square matrices  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{m \times m}$ , we further define the Kronecker sum

$$A \oplus B = A \otimes I_m + I_n \otimes B \in \mathbb{R}^{mn \times mn},$$

where  $I_m$  (resp.  $I_n$ ) denotes the identity matrix of size  $m \times m$  (respectively  $n \times n$ ).

**Theorem 7.4** Cartesian graph product and Kronecker sum of matrices

Let  $A_1$  and  $A_2$  be the adjacency matrices of the graphs  $\mathcal{G}_1$  and  $\mathcal{G}_2$ . The adjacency matrix of the Cartesian graph product  $\mathcal{G}_1 \square \mathcal{G}_2$  is given by the Kronecker sum  $A_1 \oplus A_2$  [97].

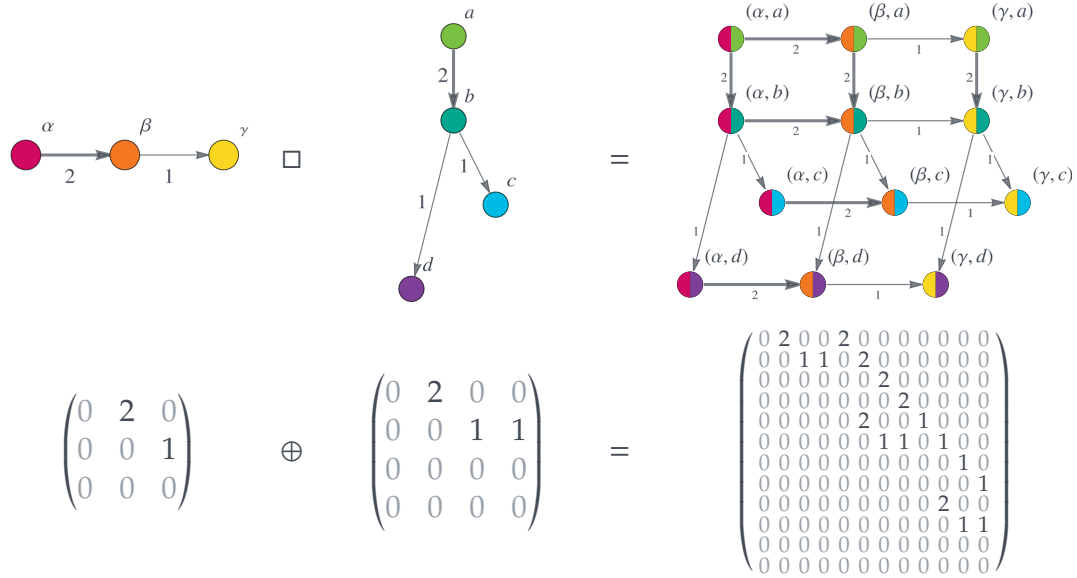
This connection of the Cartesian graph product and the Kronecker sum of the adjacency matrices is visualized in Figure 7.4.

Note that the Kronecker product is not commutative, i.e., in general we have  $A_1 \otimes A_2 \neq A_2 \otimes A_1$ . Accordingly, the graph products  $\mathcal{G}_1 \square \mathcal{G}_2$  and  $\mathcal{G}_2 \square \mathcal{G}_1$  are not equal to each other. However, as they are isomorphic to each other, we are free to choose an ordering of the matrices as long as we use the same ordering in all computations.

[87]: Horn and Johnson (1991), *Topics in Matrix Analysis*.

[124]: Loan (2000), “The ubiquitous Kronecker product”.

[97]: Kaveh and Rahami (2005), “A unified method for eigendecomposition of graph products”.



**Figure 7.4:** The Cartesian product of graphs corresponds to the Kronecker sum of their adjacency matrices. The upper row shows two graphs and their Cartesian graph product. Notice how each vertex  $(\alpha, \beta, \gamma)$  of the first graph is combined with each vertex  $(a, b, c, d)$  from the second graph, yielding a total of 12  $(= 3 \cdot 4)$  vertices in the Cartesian graph product. The edge weights (indicated by numbers next to the edges and the edge thickness) of the graphs on the left and middle transfer to the corresponding edges in the Cartesian graph product. The bottom row displays the adjacency matrices corresponding to the graphs in the upper row as an equation involving the Kronecker sum of the matrices. Reprinted from [82, Supplementary material].

### 7.2.5 LINEAR DYNAMICAL SYSTEM WITH KRONECKER STRUCTURE

As we have defined the overall structure of the system matrix components, we will now state our mathematical model of multiple pathways in Lynch syndrome-associated colorectal carcinogenesis.

It is given by a system of linear ordinary differential equations

$$\dot{x}(t) = \left( \underbrace{A}_{\text{basic mutation rate}} + \underbrace{B}_{\text{increased APC point mutation rate after MMR deficiency}} + \underbrace{C}_{\text{increased LOH event rates after APC inactivation}} + \underbrace{D}_{\text{effect combination of C and D}} + \underbrace{E}_{\text{simultaneous hit of MLH1 and CTNNB1}} + \underbrace{F}_{\text{increased KRAS mutation rate after MMR deficiency}} \right)^T x(t), \quad x(0) = x_0,$$

MMR gene germline variant

where  $A$  describes the basic independent mutational processes,  $B$  the increased  $APC$  point mutation rate after MMR deficiency,  $C$  the increased LOH event rates for the other genes after  $APC$  inactivation,  $D$  the simultaneous hit of  $MLH1$  and  $CTNNB1$  by an LOH event,  $E$  the effect combination of  $C$  and  $D$ ,  $F$  the increased  $KRAS$  mutation rate after MMR deficiency and the initial value is chosen in such a way that all cells (and thus all crypts) show a germline variant in one of the MMR genes.

We now derive how the individual matrices are defined, and how the initial condition is set.

**Basic independent mutational processes.** For matrix  $A$ , we use the gene mutation graphs shown in Figure 7.3 and construct the corresponding adjacency matrices  $A_{MMR}$ ,  $A_{CTNNB1}$ ,  $A_{APC}$ ,  $A_{KRAS}$ ,  $A_{TP53}$  for the different driver genes. The parameters of  $A_{MMR}$  depend on the hot spot and gene length of the considered MMR gene, where we focus on  $MLH1$  and  $MSH2$ . As stated in Theorem 7.4, the adjacency matrices of the individual genes are combined using the Kronecker sum to obtain the matrix  $A$  for independent mutational processes:

$$A = A_{MMR} \oplus A_{CTNNB1} \oplus A_{APC} \oplus A_{KRAS} \oplus A_{TP53}.$$

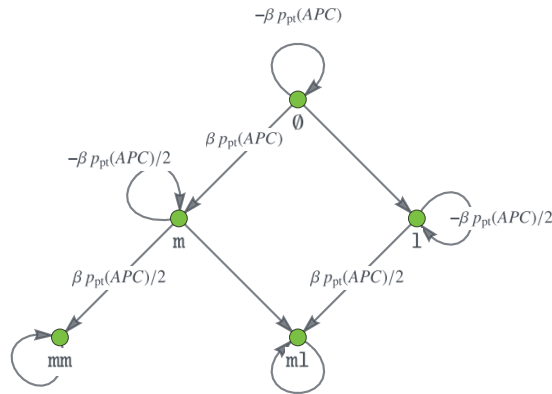
**Increased point mutation rate of  $APC$  after MMR deficiency.**

For the matrix  $B$ , we assume that the point mutation rate of  $APC$  is increased by a factor  $\beta + 1 \in \mathbb{R}$  if the crypt mutation status is MMR-deficient. This is assumed to be independent of the state of the other genes.

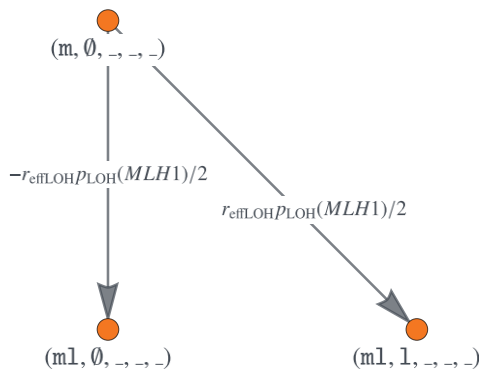
As we want to leave the matrix  $A$  unchanged, we add a matrix  $B$  to  $A$  with the entries corresponding to this mutational process multiplied by  $\beta$  instead of multiplying the corresponding entries in  $A$  by  $\beta + 1$ . The matrix  $B$  is defined by

$$B = B_{MMR} \otimes B_{CTNNB1} \otimes B_{APC} \otimes B_{KRAS} \otimes B_{TP53},$$

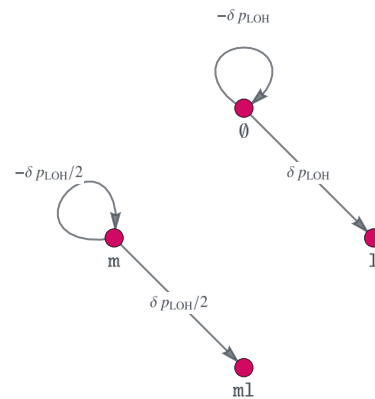
where  $B_{APC}$  is the adjacency matrix of the gene mutation graph in Figure 7.5 with the factor  $\beta$  for increased point



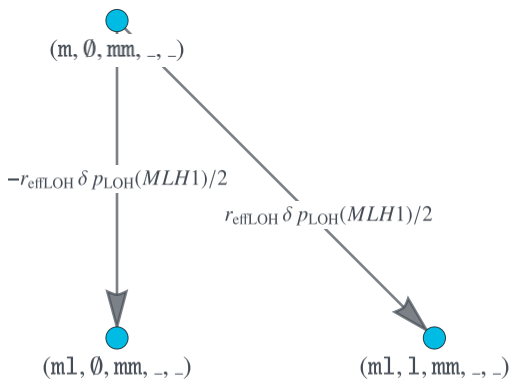
**Matrix B:** Component for increasing the point mutation rate of APC after MMR deficiency.



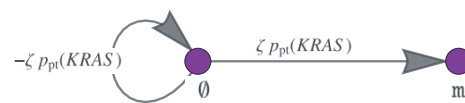
**Matrix C:** Component for the LOH event simultaneously affecting *MLH1* and *CTNNB1*.



**Matrix D:** Component for increasing the LOH event rate after APC inactivation.



**Matrix E:** Component for the mutual enhancement of the effects C and D.



**Matrix F:** Component for increasing the mutation rate of *KRAS* after MMR deficiency.

**Figure 7.5:** Graphs for building the model matrices B, C, D, E, and F. In the following,  $-$  denotes an arbitrary state of the corresponding gene. **Matrix C:** This graph is a part of the combined gene mutation graph for *CTNNB1* and *MLH1* of the matrix C. The graphs for the other possible mutation status  $MLH1 \in \{1, 11\}$ ,  $CTNNB1 \in \{m, m1\}$  are defined analogously. **Matrix D:** The graphs for both, *CTNNB1* and *TP53* are shown. The graph for MMR is defined analogously. **Matrix E:** This graph is a part of the combined gene mutation graph for *CTNNB1* and *MLH1* after APC inactivation of the matrix E. The graphs for the other possible mutation status  $MLH1 \in \{1, 11\}$ ,  $CTNNB1 \in \{m, m1\}$ ,  $APC = m1$  are defined analogously. **Matrix F:** This is the gene mutation graph of *KRAS* for the matrix F with the *KRAS* mutation rate increased by a factor  $\zeta$  after MMR deficiency.

mutation rates after MMR deficiency and

$$B_{\text{MMR}} = \text{diag}(0, 0, 1, 1, 1), \quad B_{\text{CTNNB1}} = I_5 = B_{\text{TP53}}, \\ B_{\text{KRAS}} = I_2.$$

Here,  $\text{diag}(d_1, d_2, \dots, d_n) \in \mathbb{R}^{n \times n}$  denotes a diagonal matrix with entries  $d_i, i \in \{1, 2, \dots, n\}$  on its diagonal. We set  $\beta = 1,000$  which is a factor 10 larger than in the cell-based crypt model due to calibration.

The definition of the matrix  $B$  yields the desired result of increasing the point mutation rate of  $APC$  after MMR deficiency. This can be explained intuitively: We only want to increase the point mutation rate after MMR deficiency, meaning that the MMR state is  $mm, m1$  or  $11$ , leading to the matrix  $B_{\text{MMR}}$ . Further, this influence of MMR on  $APC$  is independent of the other genes, meaning that it should hold for all states of the other genes. Thus, we choose the respective identity matrices for  $CTNNB1, KRAS$  and  $TP53$  and connect all matrices via the Kronecker product.

### Simultaneous LOH event affecting $MLH1$ and $CTNNB1$ .

Further, when considering the MMR gene  $MLH1$ , we assume an occurrence rate of  $r_{\text{effLOH}} = 0.9$  for a simultaneous hit of  $MLH1$  and  $CTNNB1$  which is a slightly different value than in the cell-based crypt model. However, currently, no biomedical measurements are available and both parameter values are rough estimates.

We build a combined gene mutation graph for  $MLH1$  and  $CTNNB1$  and connect it with the remaining genes via the Kronecker product. This connection is possible because we introduced the correct order of the genes for the Kronecker product and use this order for all matrices. The matrix  $C$  modeling this effect is given by

$$C = C_{\text{MLH1}, \text{CTNNB1}} \otimes C_{\text{APC}} \otimes C_{\text{KRAS}} \otimes C_{\text{TP53}},$$

where  $C_{\text{APC}} = C_{\text{TP53}} = I_5$  and  $C_{\text{KRAS}} = I_2$ . The matrix  $C_{\text{MLH1}, \text{CTNNB1}}$  is the adjacency matrix corresponding to the combined gene mutation graph for  $MLH1$  and  $CTNNB1$ . We explain in the following how this combined gene mutation graph is built and illustrate it in Figure 7.5.

Let  $_$  denote an arbitrary state of the corresponding gene. Instead of multiplying the edge weight  $\tilde{p}_{\text{LOH}}(\text{MLH1})/2$  of

the edge  $(m, \emptyset, -, -, -) \rightarrow (m1, \emptyset, -, -, -)$  by  $(1 - r_{\text{effLOH}})$  in the original matrix  $A$ , we add a matrix  $C$  with a corresponding edge weight  $-r_{\text{effLOH}} \tilde{p}_{\text{LOH}}(MLH1)/2$ . The following edges are added to the matrix  $C$  with the same weight:

$$\begin{aligned} (1, \emptyset, -, -, -) &\rightarrow (11, \emptyset, -, -, -), \\ (m, m, -, -, -) &\rightarrow (m1, m, -, -, -), \\ (1, m, -, -, -) &\rightarrow (11, m, -, -, -). \end{aligned}$$

Furthermore, we need to insert the following new edges with edge weight  $r_{\text{effLOH}} \tilde{p}_{\text{LOH}}(MLH1)/2$

$$\begin{aligned} (m, \emptyset, -, -, -) &\rightarrow (m1, 1, -, -, -), \\ (1, \emptyset, -, -, -) &\rightarrow (11, 1, -, -, -), \\ (m, m, -, -, -) &\rightarrow (m1, m1, -, -, -), \\ (1, m, -, -, -) &\rightarrow (11, m1, -, -, -). \end{aligned}$$

All other entries of  $C$  are zero, leading to a sparse matrix with only 400 non-zero entries.

**Increased LOH event rate after APC inactivation.** The matrix  $D$  describes the increased LOH event rate of those crypts which show *APC* inactivation due to two point mutations (having now the mutational status  $mm$ ) or due to one point mutation and one LOH event (having now the mutational status  $m1$ ). In this status, further LOH events can occur for *MMR*, *CTNNB1*, and *TP53* which will be modeled by individual matrices for each effect leading to  $D = D_1 + D_2 + D_3$ , where

$$\begin{aligned} D_1 &= D_{\text{MMR}} \otimes I_5 \otimes \text{diag}(0, 0, 0, 1, 1) \otimes I_2 \otimes I_5, \\ D_2 &= I_5 \otimes D_{\text{CTNNB1}} \otimes \text{diag}(0, 0, 0, 1, 1) \otimes I_2 \otimes I_5, \\ D_3 &= I_5 \otimes I_5 \otimes \text{diag}(0, 0, 0, 1, 1) \otimes I_2 \otimes D_{\text{TP53}}. \end{aligned}$$

Analogous to the model component  $B$ , we define a gene mutation graph of *MMR*, *CTNNB1* and *TP53* with parameter  $\delta$  such that the LOH event rate is increased by a factor  $\delta + 1$ , where we set  $\delta = 100$ . This is illustrated in Figure 7.5 for *CTNNB1* and *TP53*, where the gene mutation graph for *MMR* is defined analogously.

**Enhancement of effect  $C$  by  $D$ .** Since the additive structure allows for a biological interpretation of the single components, also the enhancement of the effect  $C$  by  $D$  is modeled by an additional matrix  $E$ . As for the matrix  $C$ , we build the combined adjacency matrix for  $MLH1$  and  $CTNNB1$  and combine it with the other genes via the Kronecker product, i.e.,

$$E = E_{MLH1,CTNNB1} \otimes \text{diag}(0, 0, 0, 1, 1) \otimes I_2 \otimes I_5,$$

where again, the order of the matrices is essential to enable an efficient implementation.

This enhancement only affects the  $APC$ -inactivated crypts, thus we use  $\text{diag}(0, 0, 0, 1, 1)$  for the  $APC$  matrix. Analogous to Figure 7.5, we illustrate parts of the gene mutation graph for the combination of  $MLH1$  and  $CTNNB1$  after  $APC$  inactivation in Figure 7.5.

**Increased mutation rate of  $KRAS$  after MMR deficiency.**

We will consider the association of an increased mutation rate of  $KRAS$  after MMR deficiency by increasing the  $KRAS$  mutation rate after MMR deficiency by a factor  $\zeta + 1$ . For this, the matrix  $F$  is defined analogously to the matrix  $B$  with the corresponding matrix entries multiplied by  $\zeta$ , which is set to 100. The gene mutation graph of  $KRAS$  is given in Figure 7.5.

**Initialization of the model.** We are left with setting the initial condition. We assume that the Lynch syndrome patients have no mutations at birth except for an MMR gene germline variant due to a point mutation (90–95 % of patients) or due to an LOH event (5–10 % of patients) [105]. We differentiate these two groups of patients by using different initial values for the differential equation. The initial value  $x_0$  for the first group of patients is

$$x_0 = n_{\text{crypts}} e_m \otimes \underbrace{e_0 \otimes e_0 \otimes e_0 \otimes e_0}_{\substack{\text{no mutations in } CTNNB1, \\ APC, KRAS \text{ and } TP53}}, \quad (7.1)$$

where  $n_{\text{crypts}} = 9.95 \cdot 10^6$  is the estimated [88] number of crypts in the colon and  $e_m$  (respectively  $e_0$ ) denotes the unit vector, which is zero everywhere, except for a 1 at the entry

[105]: Kloor et al. (2012), “Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study”.

[88]: Hounnou and Destrieux (2002), “Anatomical study of the length of the human intestine”.

corresponding to the mutation status  $m$  (respectively  $\emptyset$ ). This initial value can also be described as a vector which has the entry  $n_{\text{crypts}}$  at the position corresponding to the genotype  $(m, \emptyset, \emptyset, \emptyset, \emptyset)$  and is zero everywhere else.

Accordingly, the initial value for the second group of patients is given by

$$x_0 = n_{\text{crypts}} e_1 \otimes \underbrace{e_\emptyset \otimes e_\emptyset \otimes e_\emptyset \otimes e_\emptyset}_{\text{no mutations in CTNNB1, APC, KRAS and TP53}}. \quad (7.2)$$

In summary, by this choice of the mathematical model, we can simultaneously describe multiple pathways of carcinogenesis by defining gene mutation graphs of involved driver genes with gene-dependent alteration rates and dependencies between mutational events as far as medical data suggest so. All involved parameters and components have a biomedical meaning because that is how they have been developed and defined. Here, the gene length-dependent definition of alteration rates as well as the additive structure of the system matrix play an important role. The latter also supports the possibility to analyze the model systematically because the influence of the different components can be studied independently. With this, the first three requirements for the mathematical model stated at the beginning of this chapter are fulfilled. The remaining chapter shows the computational feasibility and modularity of the chosen Kronecker model.

## 7.2.6 THE KRONECKER STRUCTURE ALLOWS FOR COMPUTATIONAL FEASIBILITY

**Explicit solution using the matrix exponential.** In general, linear differential equations have a unique solution [187, p. 60], which also is true for our Kronecker model. The solution is given by

$$x(t) = \expm(t(A + B + C + D + E + F)^\top) x_0 \quad \forall t \in \mathbb{R}, \quad (7.3)$$

where  $\expm(A + B + C + D + E + F)$  describes the matrix exponential of the system matrix  $A + B + C + D + E + F$ ,

[187]: Teschl (2012), *Ordinary Differential Equations and Dynamical Systems*.



which is defined by

$$\begin{aligned} \text{expm}: \mathbb{R}^{n \times n} &\longrightarrow \mathbb{R}^{n \times n} \\ M &\longmapsto \sum_{k=0}^{\infty} \frac{1}{k!} M^k. \end{aligned}$$

Computing the matrix exponential [140] can be done in a variety of different ways. However, as the matrix exponential is multiplied with the vector  $x_0$ , we do not need to compute the matrix exponential as a full matrix, but only the action of the matrix exponential on the vector  $x_0$ . Algorithms for this task are studied in the context of exponential integrators [139, 152].

The definition of the matrix exponential directly yields that the matrix exponential of a triangular matrix is again a triangular matrix, which is the case for our system matrix because we assume that alterations are irreversible, leading to a DAG structure, as explained above.

**Sparsity of the system matrix.** Further, the system matrix has  $1200 = 5 \times 5 \times 5 \times 2 \times 5$  rows and columns each, corresponding to all possible genotypes. However, it is very sparse, as illustrated in Figure 7.6a. Also the matrix  $\text{expm}(A + B + C + D + E + F)$  is sparse, where the structure is reminiscent of a Sierpiński fractal (Figure 7.6b).

**Connection of the matrix exponential and the Kronecker structure.** When only considering independent mutational processes, our model reduces to

$$\dot{x}(t) = A^\top x(t), \quad x(0) = x_0. \quad (7.4)$$

In particular, the matrix  $A$  is defined as the Kronecker sum of several smaller matrices. The matrix exponential of such a matrix simplifies according to [84, Theorem 10.9, p. 237] to

$$\text{expm} \left( \bigoplus_{i \in [n]} A_i \right) = \bigotimes_{i \in [n]} \text{expm} (A_i), \quad (7.5)$$

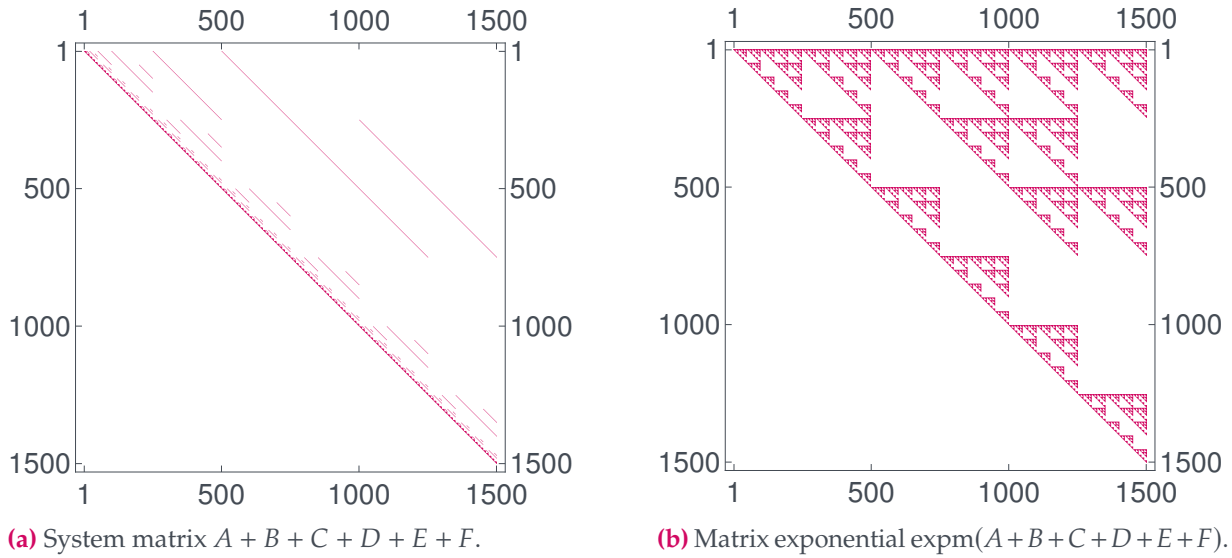
where  $[n]$  denotes the set of integers from 1 to  $n$ .

[140]: Moler and Van Loan (2003), "Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later".

[139]: Al-Mohy and Higham (2011), "Computing the Action of the Matrix Exponential, with an Application to Exponential Integrators".

[152]: Niesen and Wright (2012), "Algorithm 919: A Krylov Subspace Algorithm for Evaluating the  $\varphi$ -Functions Appearing in Exponential Integrators".

[84]: Higham (2008), *Functions of Matrices*.



**Figure 7.6: Sparse matrix structure.** (a) The system matrix  $A + B + C + D + E + F$  of the linear model is a very sparse matrix, i.e. only a few entries are nonzero. These **nonzero entries** are colored red in the plot, which also illustrates the fact that  $A + B + C + D + E + F$  is an upper triangular matrix. (b) The sparsity structure of the matrix  $\expm(A + B + C + D + E + F)$ , which is reminiscent of a Sierpiński fractal, is due to the individual matrices being the Kronecker product and sum of matrices. The two plots also illustrate nicely how modeling sparse local interactions in the matrix  $A + B + C + D + E + F$  can have a more global effect in  $\expm(A + B + C + D + E + F)$ .

Thus, instead of computing the matrix exponential of one large matrix, we can compute the matrix exponentials of several small matrices and connect them with the Kronecker product, which gives an additional performance boost.

**The initial value as Kronecker product.** In our case, the initial value  $x_0$  itself can be written as a Kronecker product (see equations (7.1) and (7.2)), i.e.

$$x_0 = \bigotimes_{i \in [n]} x_i, \tag{7.6}$$

where the  $x_i$  are vectors with sizes corresponding to the sizes of the matrices  $A_i$ . With this, the solution of the dynamical system (7.4) can be written as

$$x(t) = \left( \bigotimes_{i \in [n]} \expm(tA_i^T) \right) \left( \bigotimes_{i \in [n]} x_i \right) \tag{7.7a}$$

$$= \bigotimes_{i \in [n]} \expm(tA_i^T) x_i, \tag{7.7b}$$

where the last equality is due to the mixed product property of Kronecker products [124].

[124]: Loan (2000), “The ubiquitous Kronecker product”.

Thus, in our case, when only considering the matrix  $A$ , the solution for the Lynch syndrome individuals with an MMR gene germline variant caused by a point mutation reads

$$\begin{aligned} x(t) = & \expm(tA_{\text{MMR}}^{\top})e_m \otimes \expm(tA_{\text{CTNNB1}}^{\top})e_0 \otimes \\ & \otimes \expm(tA_{\text{APC}}^{\top})e_0 \otimes \expm(tA_{\text{KRAS}}^{\top})e_0 \otimes \\ & \otimes \expm(tA_{\text{TP53}}^{\top})e_0 n_{\text{crypts}}. \end{aligned}$$

### Extracting several mutation status from the solution vector.

In most cases, we are not only interested in a single entry of the solution vector  $x(t)$ , but in the sum of several entries. To achieve this, we consider the scalar product  $v^{\top}x(t)$  of the vector  $x(t)$  with a vector  $v$  of the same size which has a 1 in all states we want to consider and a 0 everywhere else.

Often this vector can, similarly to the initial value  $x_0$  in equation (7.6), be written as the Kronecker product of vectors  $v_i$  with sizes corresponding to the matrices  $A_i$

$$v = \bigotimes_{i \in [n]} v_i.$$

In this case, the accumulation simplifies to

$$v^{\top}x(t) = \bigotimes_{i \in [n]} v_i^{\top} \expm(tA_i^{\top}) x_i \quad (7.8a)$$

$$= \prod_{i \in [n]} v_i^{\top} \expm(tA_i^{\top}) x_i, \quad (7.8b)$$

where the first equality follows as above from the mixed product property of Kronecker products and the second one is due to the fact that the Kronecker product of real numbers (here:  $v_i^{\top} \expm(tA_i) x_i$ ) is the standard product of real numbers.

**Solution of the full system.** Now, we want to focus on the whole system matrix with its additive components. The equation (7.5) can be seen as a generalization of  $e^{a+b} = e^a e^b$  for real numbers  $a$  and  $b$ . However, it is important to note that this statement does not hold for the standard matrix addition and product. Thus, in general, we have [84, Theorem 10.2, p. 235]

$$\expm(A + B) \neq \expm(A) \expm(B),$$

[84]: Higham (2008), *Functions of Matrices*.

[130]: McLachlan and Quispel (2002), "Splitting Methods".

[23]: Biagi and Bonfiglioli (2018), *An Introduction to the Geometrical Analysis of Vector Fields*.

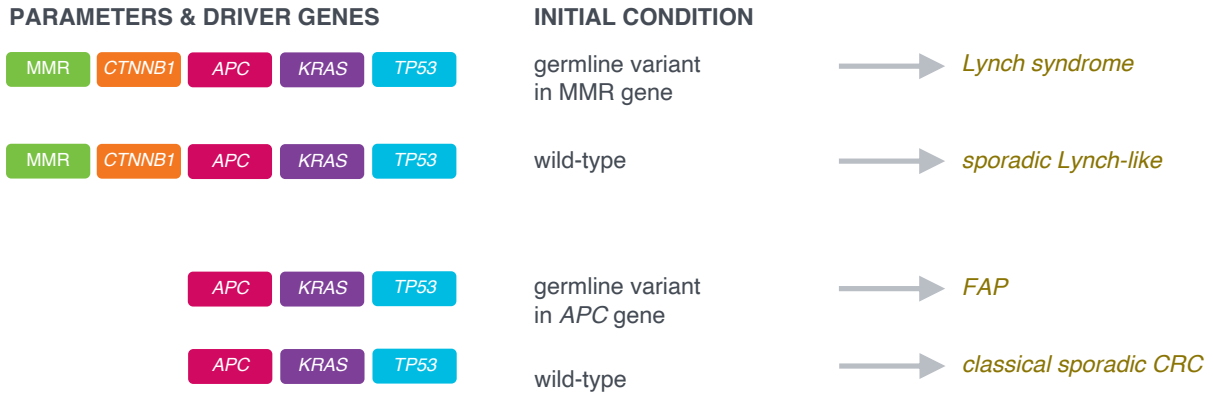
meaning that we can not apply (7.5) to the additive components of the system matrix. This means the matrices cannot be simplified as much as for only considering the matrix  $A$  built by the Kronecker sum. However, we can still use the exponential integration mentioned above. Further mathematical analysis is possible in the context of operator splitting [130] or the Baker-Campbell-Hausdorff formula [23]. This underlines the potential of fast and cheap computations using the Kronecker structure in the system matrix for modeling multiple pathways of carcinogenesis, thus, fulfilling requirement 4 (computational feasibility) at the beginning of the chapter.

### 7.3 MODIFYING PARAMETERS AND INITIAL CONDITIONS TO MODEL OTHER TYPES OF COLORECTAL CARCINOGENESIS

We introduced how the driver genes, mutation parameters and initial condition are set in order to model Lynch syndrome-associated colorectal carcinogenesis. By changing some of these, we are able to model other types of colorectal carcinogenesis which we explain in this section to emphasize the modularity of the derived Kronecker model. We will present modifications to handle Lynch-like, FAP-associated, and classical sporadic microsatellite-stable colorectal carcinogenesis. An overview on the possible modifications is given in Figure 7.7.

In general, the Kronecker modeling framework is also applicable to carcinogenesis in other organs which is based on the accumulation of point mutations and LOH events affecting specific driver genes. This will check the last requirement from the beginning of this chapter on the modularity of the modeling framework.

**Lynch-like colorectal carcinogenesis.** Lynch-like carcinogenesis is currently considered the sporadic counterpart of Lynch syndrome. Thus, the main difference between Lynch-like and Lynch syndrome carcinogenesis is the absence or presence of a monoallelic MMR gene germline variant. Thus, only the initial condition has to be changed to model Lynch-like carcinogenesis. We do so by introducing the additional



**Figure 7.7: Modifications to the parameters, included driver genes and initial conditions made to model other types of colorectal carcinogenesis.** Initial conditions can be varied to model either sporadic or hereditary mode of colorectal cancer. By changing driver genes and corresponding parameters, different pathways and thus different types of colorectal carcinogenesis, like MSI or MSS, can be modeled.

vertex  $\emptyset$  in  $\mathcal{V}_{\text{MMR}} = \{\emptyset, m, 1, mm, m1, 11\}$  with point mutation and LOH event rates described in Section 7.2. The initial value changes by  $x_0 = 0$  except for the entry corresponding to  $(m, \emptyset, \emptyset, \emptyset, \emptyset)$  or  $(1, \emptyset, \emptyset, \emptyset, \emptyset)$  in the hereditary case and  $(\emptyset, \emptyset, \emptyset, \emptyset, \emptyset)$  in the sporadic case for which the value is set to  $n_{\text{crypts}}$ .

**Sporadic MSS colorectal carcinogenesis.** By not including MMR genes and *CTNNB1* in the vertex set, we model the classical adenoma-carcinoma sequence including *APC*, *KRAS*, and *TP53*, and thus, describe the classical sporadic MSS colorectal carcinogenesis.

**FAP colorectal carcinogenesis.** We could also describe colorectal carcinogenesis in another hereditary syndrome, namely FAP. Those patients have a single germline variant in *APC*, which is known to be a point mutation in almost all cases [146, 167]. Thus, the dynamical system starts with all crypts in the state  $(\emptyset, \emptyset, m, \emptyset, \emptyset)$ .

As reported in [77], we assume that the germline variants are not equally distributed among the base pairs of the *APC* gene. Instead, they are concentrated at specific codons leading to the fact that we change the number of hotspot base pairs in the FAP case. Due to [99], the classical FAP case is associated with germline variants in codons 1250 – 1464, leading to the assumption of approximately setting  $n_{\text{hs}} = 600$  in our model for FAP simulations.

[146]: Nagase and Nakamura (1993), “Mutations of the APC (adenomatous polyposis coli) gene”.

[167]: Rashid et al. (2016), “Adenoma development in familial adenomatous polyposis and MUTYH-associated polyposis: somatic landscape and driver genes”.

[77]: Gryfe (2009), “Inherited colorectal cancer syndromes”.

[99]: Kinzler and Vogelstein (1996), “Lessons from hereditary colorectal cancer”.

[99]: Kinzler and Vogelstein (1996), “Lessons from hereditary colorectal cancer”.

[75]: Goldstein et al. (2003), “Hyperplastic-like Colon Polyps That Preceded Microsatellite-Unstable Adenocarcinomas”.

[182]: Staffa et al. (2015), “Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters”.

The common regions of germline variants described above are also correlated with the most occurring polyps (more than 5,000) [99] in FAP patients. With an estimated diameter of 4.8 mm per polyp [75] and 0.09 mm per crypt [182], this would result in  $10^7$  crypts in a polypous state. Thus, our model simulations should also reflect that the number of polyps, assumed to consist of *APC*-inactivated crypts, should be much higher than in the sporadic case.

**Carcinogenesis in other organs.** The presented modeling framework of using gene-dependent point mutation and LOH event rates, building gene mutation graphs and corresponding adjacency matrices using the Kronecker structure, and solving the resulting linear dynamical system is very general. It could be applied to carcinogenesis in other organs by possibly changing the incorporated driver genes. The tissue structure will be different not describing crypts but other structures with different properties, e.g., regarding mutation fixation. Therefore, the point mutation and LOH event rates have to be adapted accordingly to account for different cell properties.

## 7.4 CALIBRATION AND VALIDATION

### RESULTS OF THE KRONECKER MODEL

We present the simulations results for modeling Lynch syndrome-associated colorectal carcinogenesis using the Kronecker structure which were used for model calibration and validation. To be precise, the age-resolved evolution of the number of crypts in the various combined mutational status of the driver genes of a Lynch syndrome individual are simulated. We focus on the results for *MLH1* and *MSH2*, which are the MMR genes that are related to the highest CRC incidence in Lynch syndrome [64].

By adapting the initial conditions, the exact same model can be used for results on the distribution of Lynch syndrome individuals among the pathways. The results have also been presented in [82], where the following section is strongly based on.

[64]: Engel et al. (2020), “Associations of Pathogenic Variants in *MLH1*, *MSH2*, and *MSH6* With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome”.

[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

### Age-dependent evolution of crypts with mutation status of clinical interest.

As previously described, we extracted and combined several mutational status which define clinical states such as early precursor lesions as well as more advanced states in the process of carcinogenesis. Those are of clinical importance because they are either the first measurable signs of carcinogenesis, like MMR-deficient crypts, or they are associated to a specific pathway and thus give evidence for which treatment option might be promising. The clinical states are defined by:

- ▶ **MMR-deficient crypts:** MMR-deficient; *CTNNB1*, *APC*, *KRAS*, *TP53* intact, i.e.  $(mm, \emptyset, \emptyset, \emptyset, \emptyset) + (m1, \emptyset, \emptyset, \emptyset, \emptyset) + (11, \emptyset, \emptyset, \emptyset, \emptyset)$
- ▶ **State 1:** MMR-proficient or MMR-deficient, *CTNNB1* activated; *APC* inactivated; *KRAS* and *TP53* intact (called early adenomatous)
- ▶ **State 2:** MMR-proficient or MMR-deficient, *CTNNB1* activated; *APC* inactivated; *KRAS* activated; *TP53* intact (called late adenomatous)
- ▶ **State 3:** MMR-proficient or MMR-deficient, *CTNNB1* activated; *APC* and *TP53* inactivated; *KRAS* activated (called cancerous)

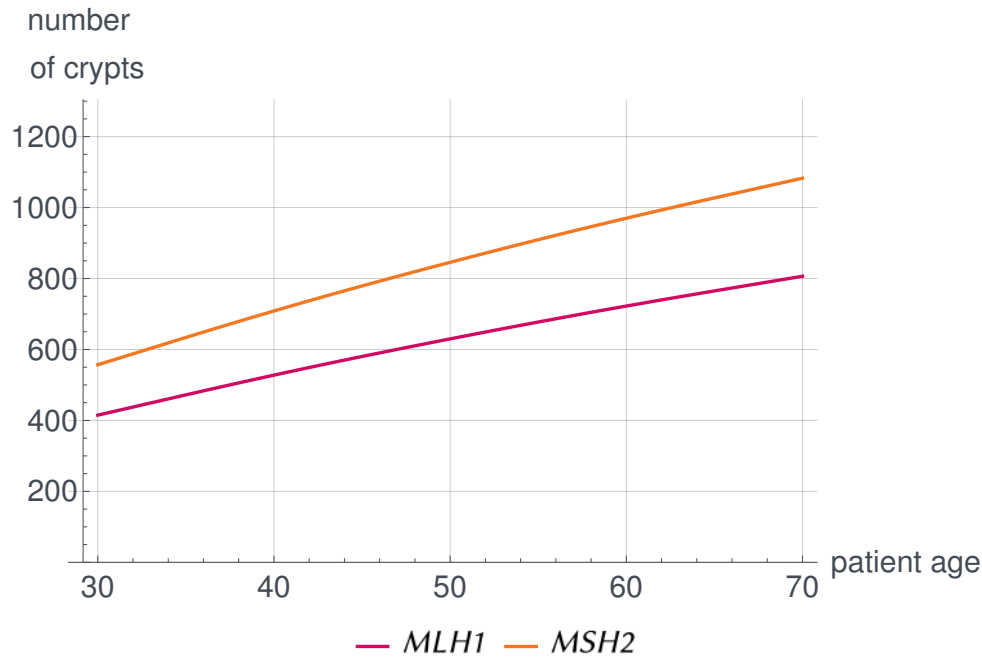
We calibrated the model in such a way that the number of MMR-deficient crypts is quantitatively comparable to the clinical data available [182] with results shown in Figure 7.8.

As suitable medical data are missing to determine the fixation affinity parameters, parameter learning was not performed in a mathematically rigorous way. Thus, some of the numbers of crypts presented here may not match the real numbers if measurable. This seems reasonable for the more advanced clinical states, shown in Figure 7.9. As soon as further data are available either for the parameters or for the evolution of crypt numbers or both, parameter learning will be possible.

### Influences of different MMR gene variants on carcinogenesis.

Clinical data and population-based studies currently show MMR gene variant-dependent differences in adenoma and carcinoma risk of Lynch syndrome individuals. Further, the distribution among the pathways of carcinogenesis may depend on the MMR gene.

[182]: Staffa et al. (2015), “Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters”.



**Figure 7.8:** Number of MMR-deficient crypts over the life of a typical Lynch syndrome individual for *MLH1* and *MSH2*. The model parameters for fixation affinities are calibrated in such a way that the simulation results are in concordance with published data [182]. In our model, differences among genes are due to differences in coding region and gene lengths as well as due to the different mutational dependencies. Figure reprinted from [82].

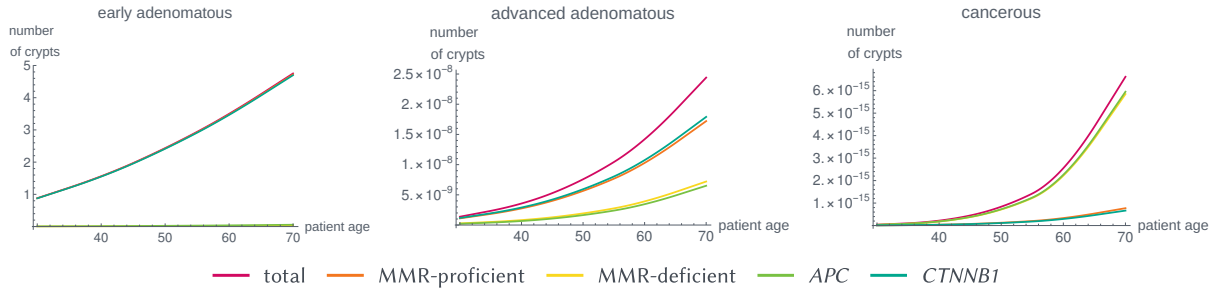
[64]: Engel et al. (2020), “Associations of Pathogenic Variants in *MLH1*, *MSH2*, and *MSH6* With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome”.

[182]: Staffa et al. (2015), “Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters”.

It is essential to examine these gene-specific associations with the pathways of carcinogenesis, as the different pathways of carcinogenesis show different needs for treatment and surveillance [64].

As an example, we simulated the age-dependent evolution of MMR-deficient crypts, the earliest detectable precursor lesions of pathways 2 and 3 in Lynch syndrome-associated colorectal carcinogenesis. Current data show that a typical Lynch syndrome individual has about 500-1,000 of MMR-deficient crypts in the whole colon [182]. However, age-dependent data are, to the best of our knowledge, currently not reliably available. Our simulation results, shown in Figure 7.8, are in good concordance with these data, showing an *in silico* age-dependent evolution. In the model, the differences are due to different properties of the MMR genes, such as coding region and gene lengths, and because dependent mutational processes influence the evolution of the crypts differently. This is in particular true for the chromosomal changes simultaneously affecting *MLH1* and *CTNNB1* but not *MSH2*. However, as more data become available, additional MMR gene-dependent differences can be included in the model.





**Figure 7.9:** Age-dependent evolution of the number of crypts of specific advanced clinical states in a typical *MLH1* carrier, like early adenomatous, advanced adenomatous and cancerous states. Due to the mutational dependencies included in the model, the distribution of MMR-deficient and MMR-proficient, as well as the contribution of *CTNNB1* and *APC* change for the different states. Due to the lack of suitable medical data, parameter learning was not performed in a rigorous way. As soon as data are available, this can be done using different mathematical techniques. Figure reprinted from [82].

### Initial analyses of the distribution among the pathways of carcinogenesis.

In general, with the model, we are able to obtain an age-dependent evolution of the number of crypts of a Lynch syndrome individual among the pathways of carcinogenesis simultaneously. The distribution among the pathways of carcinogenesis on a population level can be approximated by multiplying the initial vector by the number of Lynch syndrome patients we are interested in, say 100. Then, we compare the percentages of people in different states which are unique to specific pathways of carcinogenesis to obtain estimates for the age-dependent distribution among the pathways of carcinogenesis on a population level.

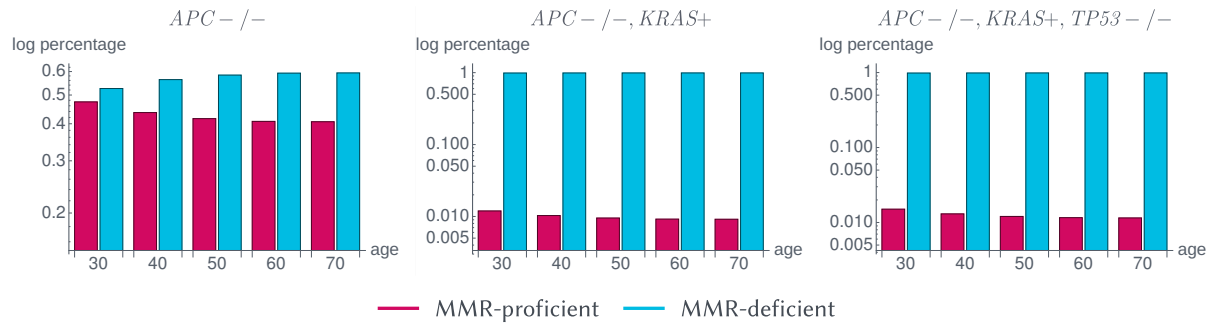
As an example, we analyzed the proportion of MMR-proficiency and MMR-deficiency in *APC*-inactivated crypts to determine the proportion in which MMR deficiency occurred as an initial event in carcinogenesis of Lynch syndrome carriers. The results are shown in Figure 7.10 and are similar to the currently available data [2] ( $\approx 75\%$ ) with a slight underestimation of MMR-deficient *APC*-inactivated crypts ( $\approx 60\%$ ) compared to MMR-proficient ones. In our simulations, more of the *APC*-inactivated crypts are MMR-deficient, supporting the hypothesis that MMR deficiency is often an initial event in Lynch syndrome colorectal carcinogenesis [2, 175, 182].

These relative distributions change for further advanced clinical states as we included mutational dependencies in the model. In general, for independent mutational processes only including matrix  $A$ , the distributions in Figure 7.10 are the same for all states.

[2]: Ahadova et al. (2018), “Three molecular pathways model colorectal carcinogenesis in Lynch syndrome”.

[175]: Sekine et al. (2017), “Mismatch repair deficiency commonly precedes adenoma formation in Lynch Syndrome-Associated colorectal tumorigenesis”.

[182]: Staffa et al. (2015), “Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters”.



**Figure 7.10:** Age-dependent proportion of MMR-proficient and MMR-deficient crypts in a typical *MLH1* carrier in different clinical states corresponding to the states in the classical adenoma-carcinoma sequence by Vogelstein [201]. Among the *APC*-inactivated (*APC*<sup>-/-</sup>) crypts (left), the number of MMR-deficient crypts is up to 20% higher than the number of MMR-proficient ones. This difference largely increases with the subsequent *KRAS* activation (*KRAS*<sup>+</sup>) (middle) and *TP53* inactivation (*TP53*<sup>-/-</sup>) (right) leading to the fact that almost all crypts in the last state, corresponding to a cancerous state, are MMR-deficient. These simulation results are in concordance with available data with a slight underestimation of MMR-deficient *APC*-inactivated crypts [2]. Figure reprinted from [82].

In our setting, the relative distribution heavily changes from *APC*-inactivated to additionally *KRAS*-activated crypts. In particular, almost all *APC*-inactivated, *KRAS*-activated crypts are MMR-deficient.

**Influence of medical parameters on carcinogenesis.** Analyzing influences of parameters on the model solutions are important, in particular, in the case of uncertain or missing data.

Firstly, the number of point mutations  $n_{pt}$ , the number of cells  $n_{cells}$ , and the number of crypts  $n_{crypts}$  determine the absolute values of the analyzed numbers.

Further, the relation of the hotspot length and the gene length determines the relative frequency of point mutations and LOH events for the individual genes, which can be changed by the mutational dependencies for specific mutational states. Here, the magnitude of the parameters  $r_{effLOH}$ ,  $\beta$ ,  $\delta$ , and  $\zeta$  determines the effect size of the individual mutational dependencies.

The fitness parameter  $b(\text{gene})$  affects the slope of the crypt evolution curve. In our case,  $b(\text{MMR}) < 0$  leads to the fact that further MMR-deficient crypts are disadvantageous for the crypt survival leading to fewer additional MMR-deficient crypts with increasing age (Figure 7.8).

In contrast, *APC* inactivation is modeled as an advantage for the crypts such that  $b(\text{APC}) > 0$  leads to more additional *APC*-inactivated crypts with increasing age.

Furthermore, the relation of the fixation affinities  $f(\text{gene})$  for different genes seems to influence the mutation order. A larger value of  $f(\text{gene})$  leads to a faster fixation in this gene and thus to an earlier event in carcinogenesis (Figure 7.10). According to the calibration of the model, those parameters are currently set to  $f(\text{MMR}) = 2.3 \cdot 10^{-6}$ ,  $f(\text{CTNNB1}) = 1.2 \cdot 10^{-3}$ ,  $f(\text{APC}) = 8.3 \cdot 10^{-7}$ ,  $f(\text{KRAS}) = 2.5 \cdot 10^{-8}$ , and  $f(\text{TP53}) = 1.2 \cdot 10^{-5}$ .

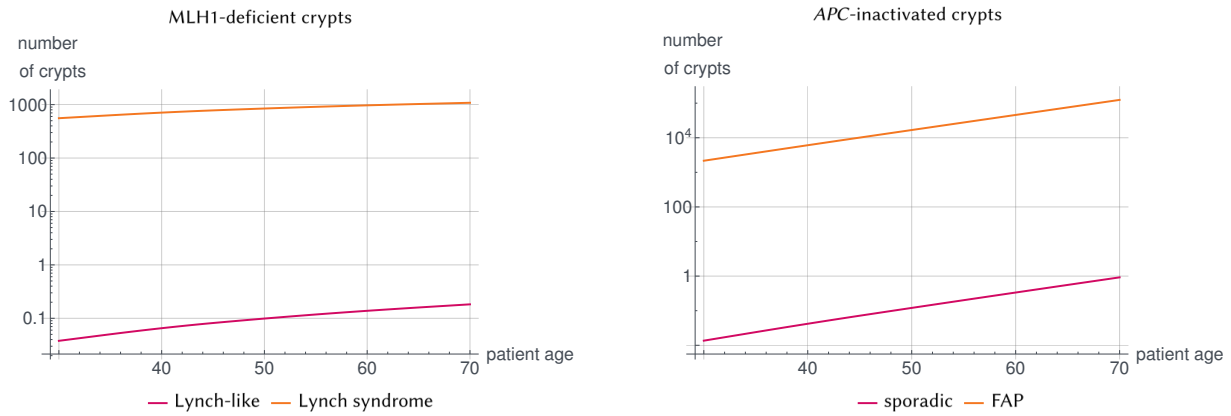
As soon as there are more molecular data available, parameter learning could be applied to the model in order to get a deeper understanding of the underlying mechanisms in Lynch syndrome carcinogenesis. In particular, there is still uncertainty in the data about the fitness advantages and disadvantages of individual genetic changes as well as on the fixation affinities of mutations. General information on mutational dependencies and how they affect the phenotype of the cells is crucial to extend the model with further biomolecular mechanisms.

**Age-dependent simulation results for Lynch-like and FAP colorectal carcinogenesis.** As explained in Section 7.3, the modular structure of the Kronecker model allows to change parameters and initial conditions to describe other types of colorectal carcinogenesis. We present the results comparing Lynch-like and Lynch syndrome colorectal carcinogenesis as well as sporadic MSS and FAP colorectal carcinogenesis, closely following [82].

For Lynch-like and Lynch syndrome colorectal carcinogenesis, we compared the age-dependent evolution of the number of MMR-deficient crypts as early precursor lesions. There are much more MMR-deficient crypts in Lynch syndrome individuals than in Lynch-like individuals, which corresponds to the medical findings in [182]. The simulation results are illustrated in Figure 7.11. For the absolute numbers in Lynch-like individuals and the evolution over age, further clinical data have to be collected to possibly adapt the model parameters accordingly.

[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

[182]: Staffa et al. (2015), “Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters”.



**Figure 7.11: Comparison of the age-dependent evolution of crypts in sporadic and hereditary modes of colorectal carcinogenesis.** Left: MMR-deficient crypts in Lynch-like and Lynch syndrome individuals. The number of MMR-deficient crypts (log-scale on y-axis) is significantly higher in Lynch syndrome individuals compared to Lynch-like individuals, which matches the findings in [182]. Right: *APC*-inactivated crypts in the sporadic MSS case and in FAP individuals, where we changed the initial value of the dynamical system as well as  $n_{hs}(APC) = 600$  for FAP. Our simulation results yield numbers below estimates found in the literature [75, 99, 182]. With improved measurements, future work will adapt the parameters accordingly. Figure adapted and reprinted from [82].

Further, we compared the *APC*-inactivated crypt evolution of a typical FAP individual with a sporadic MSS case without a germline variant in *APC* for all crypts. With the previously described parameter set, our model simulations yield between  $10^4 - 10^5$  *APC*-inactivated crypts, which is below the estimates of  $10^7$  calculated from available literature (see Section 7.3). The age evolution of the number of crypts is shown in Figure 7.11. Also in this case, age-dependent data would be necessary to adapt the model parameters accordingly.

## 7.5 OUTCOMES AND DISCUSSION

We presented a general modeling framework using the Kronecker structure that 1) offers a simultaneous description of multiple pathways of carcinogenesis, 2) is medically interpretable with respect to the parameters and the model components, 3) can be analyzed systematically, 4) is computationally feasible, and 5) is modular. We model carcinogenesis on the basis of the number of crypts being present with specific mutational status defined by the involved driver genes. The latter can be aligned to clinically defined stages such as early adenoma in colorectal carcinogenesis, although we are fully aware of the fact that the congruence between clinical and molecular definitions will be limited due to the

dynamics of cancer development and the limited availability of comprehensive data. In particular, age-dependent data are missing for a systematic parameter learning and calibration of the model to Lynch syndrome colorectal carcinogenesis which is subject of future work. This would allow to check the hypothesis of gene-length dependent mutation rates, further align the clinical and molecular definitions of important carcinogenesis states and pathways of carcinogenesis.

Limitations of data also concern the topic of overdiagnosis and disappearing lesions. From a mathematical point of view, it is straightforward to include spontaneous disappearance of lesions in the modeling approach, as shown in [82, Supplementary material]. However, there are currently not enough prospective data available to estimate or learn the necessary parameters, e.g., the probability of spontaneous crypt loss for each mutation status. This is the reason why we have chosen a simpler model jointly modeling the proliferation and disappearance by the self-loops in the graph, largely reducing the number of parameters that need to be determined. If more molecular data with the analysis of all possibly relevant genes are available, a comparison of the model with these data will allow for parameter learning of the yet unmeasurable parameters. In addition, a systematic sensitivity analysis of the involved parameters would be possible which might allow even for procedures in the context of optimal experimental design.

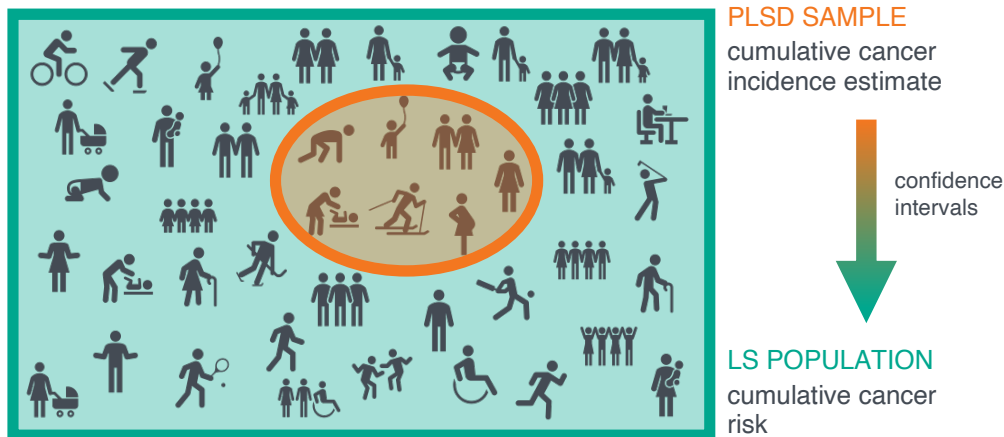
[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.



**MODELING LYNCH SYNDROME AT THE  
POPULATION LEVEL**







## 8 PLSD: MODELING CUMULATIVE CANCER RISK IN THE LYNCH SYNDROME POPULATION

In this chapter, we develop a statistical method for estimating the cumulative risk of Lynch syndrome individuals to develop cancer up to a certain age, i.e. the cumulative cancer risk in the Lynch syndrome population, based on data from the Prospective Lynch Syndrome Database (PLSD). These estimates are used to evaluate the effect of colonoscopy on cancer risk in affected individuals.

PLSD is a shared database of international prospective follow-up studies of Lynch syndrome families. The aim of PLSD was to describe cancer incidences in all organs in Lynch syndrome individuals, characterized as carriers of pathogenic MMR (*path\_MMR*) variants, who were undergoing follow-up according to the internationally advocated clinical guidelines. A stratification by age, gene, and sex should complement these analyses. Once sufficient numbers of carriers and follow-up years were collated, the intention was to use the information obtained to assess whether the results were compatible with current assumptions about carcinogenesis and the expected effects of interventions in Lynch syndrome.

From a mathematical analysis point of view, PLSD compiles observed cancers in *path\_MMR* carriers from the first prospectively planned and performed colonoscopy. It considers all cancers that occur before or at the same age as the first colonoscopy as prior or prevalent cancers, and from that point onwards it counts new primary cancers as events. Data collection was made from age 25 years at earliest, and cumulative incidence of CRC at age 25 years was set to zero. When CRC was counted as the event, all carriers who already

8.1	COMPUTATION OF CANCER RISK . . . . .	153
8.2	COMPUTATION RESULTS . . . . .	157
8.3	OUTCOMES AND DISCUSSION . . . . .	159

[141]: Møller (2020), “The Prospective Lynch Syndrome Database reports enable evidence-based personal precision health care”.

[143]: Møller et al. (2022), “Colorectal cancer incidences in Lynch syndrome: a comparison of results from the prospective lynch syndrome database and the international mismatch repair consortium”.

had CRC prior to or at inclusion in the study were excluded, and observation time was right-censored at the first event, last observation or death, whichever came first. Using the number of observed cancers and the number of observation years in predefined 5-year age intervals starting from 25 years to 75 years, annual incidence rates and cumulative incidence risks are calculated. The used methods have been discussed in detail in [141]. However, as depicted in more detail below, the currently used methods are based on the simplifying assumption of a Normal distribution which is often done in statistics. We develop a method for calculating the cumulative incidence risks based on a Nelson-Aalen estimate with an underlying Poisson distribution. The latter represents the data more appropriately as cancer occurrence is a dichotomous variable, and we count the number of cancer cases in a specific age interval which is the definition of Poisson distributed data. Further, we present an approach to calculate the corresponding confidence intervals to obtain an impression about how well the cumulative cancer incidence in the PLSD data sample represents the cumulative cancer risk in the overall Lynch syndrome population. The methods were previously published in [143], which we closely follow in this chapter for the derivation of the methods. We will compare the previous and the newly defined method for calculating the cumulative incidence risks and 95% confidence intervals for the different types of cancer stratified by gene and sex. Further, we shortly discuss other possibilities to compute the confidence intervals in this setting.

The novel computation method is used for comparing estimates of colorectal cancer incidences of PLSD and of the International Mismatch Repair Consortium (IMRC). IMRC was established around the same time as PLSD was developed with the aim at compiling data on as many Lynch syndrome families as possible for a retrospective segregation analysis to also obtain cumulative cancer incidences in *path\_MMR* carriers but by a retrospective analysis including family members in former generations. Further, the novel method will be the standard for the next versions of the PLSD results which are regularly updated including recently collected data.

## 8.1 COMPUTATION OF PROSPECTIVE CUMULATIVE CANCER RISK WITH CONFIDENCE INTERVALS

The PLSD data consist of a number of subjects, each followed for a known time period, either until an event occurs or until they leave the study for other reasons. In other words, cancer is assumed to be a dichotomous variable (cancer occurs yes or no). The number of events, i.e., cancer cases, and time at risk are then aggregated per 5-year age interval starting from 25 to 75 years of age. This is mathematically described by a Poisson distribution.

### 8.1.1 DEFINITION OF CUMULATIVE INCIDENCE FUNCTION IN SURVIVAL ANALYSIS

We introduce standard notation in survival analysis, a branch of statistics for analyzing the expected duration of time until one event occurs, e.g., cancer in Lynch syndrome individuals.

By  $F(t)$  we denote the probability that the event occurs between time 0 and  $t$ . This corresponds to the cumulative incidence (denoted by  $Q(\text{age})$  in [59]) which is the quantity we are interested in. The survival function at time  $t$  is defined by

$$S(t) = 1 - F(t).$$

The hazard function  $h(t)$  can be written by

$$h(t) = \frac{-d \log S(t)}{dt}.$$

The cumulative hazard function  $H(t)$  (denoted by  $CH(\text{age})$  in [59]) is the integration of the hazard function from time 0 to  $t$ , in formulas

$$\begin{aligned} H(t) &= -\log(S(t)) \\ &= -\log(1 - F(t)). \end{aligned}$$

[59]: Dominguez-Valentin et al. (2019), "Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database".

It is connected to the survival function by

$$S(t) = \exp(-H(t))$$

and thus, it holds

$$F(t) = 1 - \exp(-H(t)).$$

### 8.1.2 NELSON-AALEN CUMULATIVE INCIDENCE ESTIMATES BASED ON A POISSON DISTRIBUTION

We derive estimates for the cumulative incidence risks based on a Poisson distribution, where the following is based on [50, Chapter 2] and we closely follow the derivation in [143, Supplementary material]. We use the number of observed cancer cases  $d$  and the number of observation years  $y_{\text{obs}}$  within the 5-year age intervals to compute the *incidence rate*, denoted by AIR in [59]. Typically, in survival analysis, we use data on the number of observed events and the number of patients under risk  $n$  to compute the *incidence risk* IR. Those quantities can be computed via

$$\begin{aligned} \text{IR} &= \frac{d}{n}, \\ \text{AIR} &= \frac{d}{y_{\text{obs}}}. \end{aligned}$$

The following connection holds between those quantities which we will make use of in the subsequent analysis

$$\text{IR within age interval} = \frac{d}{y_{\text{obs}}} \cdot \text{length of age interval}.$$

Thus, the incidence risk IR is approximated by  $\text{IR} = \text{AIR} \cdot 5\text{yrs}$ .

#### Proposition 8.1 Cumulative cancer incidence estimate

Assuming a Poisson distribution, our quantity of interest in the PLSD setting, the cumulative cancer incidence Nelson-Aalen estimate is given by

$$\hat{F}(t_k) = \prod_{j=1}^k \exp\left(-\frac{d_j}{(y_{\text{obs}})_j} \cdot 5 \text{ years}\right).$$

[50]: Collett (2015), *Modelling Survival Data in Medical Research*.

[143]: Møller et al. (2022), "Colorectal cancer incidences in Lynch syndrome: a comparison of results from the prospective lynch syndrome database and the international mismatch repair consortium".

[59]: Dominguez-Valentin et al. (2019), "Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database".

*Proof.* To make use of the Poisson distribution, we use the Nelson-Aalen estimator of the cumulative hazard function in age interval  $k$ , which is given by

$$\hat{H}(t_k) = \sum_{j=1}^k \left( \frac{d_j}{(y_{\text{obs}})_j} \cdot 5 \text{ yrs} \right).$$

Thus, by definition,

$$\begin{aligned} \hat{S}(t_k) &= \exp(-\hat{H}(t_k)) \\ &= \exp\left(-\sum_{j=1}^k \left( \frac{d_j}{(y_{\text{obs}})_j} \cdot 5 \text{ yrs} \right)\right) \\ &= \prod_{j=1}^k \exp\left(-\frac{d_j}{(y_{\text{obs}})_j} \cdot 5 \text{ years}\right). \end{aligned}$$

With  $F(t) = 1 - S(t)$ , we obtain the stated formula for  $F(t)$ .  $\square$

**Remark 8.2** In statistical terms,  $\hat{F}(t_k)$  is the Nelson-Aalen estimate for the distribution function of the random variable  $T$  associated with the survival time. In general, the Nelson-Aalen estimate is based on a Poisson distribution of the number of cancer cases within the given age interval. The Kaplan-Meier estimate which is often used in survival analysis approximates the Nelson-Aalen estimate [50]. Further, the Nelson-Aalen estimate has been shown to perform better than the Kaplan-Meier estimate in small samples.

[50]: Collett (2015), *Modelling Survival Data in Medical Research*.

**Proposition 8.3** Variance of the cumulative hazard estimate  
It holds for the variance of the Nelson-Aalen estimator of the cumulative hazard function

$$\text{Var}\left(\hat{H}(t_k)\right) = \sum_{j=1}^k \frac{25d_j}{(y_{\text{obs}})_j^2}.$$

*Proof.* With the rules for calculating variances and the Delta method, it holds for 5 year age intervals:

$$\begin{aligned}\text{Var}\left(\hat{H}(t_k)\right) &= \text{Var}\left(\sum_{j=1}^k \frac{5d_j}{(y_{\text{obs}})_j}\right) \\ &= \sum_{j=1}^k \text{Var}\left(\frac{5d_j}{(y_{\text{obs}})_j}\right) \\ &= \sum_{j=1}^k \frac{25d_j}{p_j^2}.\end{aligned}$$

□

[25]: Bie et al. (1987), “Confidence Intervals and Confidence Bands for the Cumulative Hazard Rate Function and Their Small Sample Properties”.

**Point-wise confidence intervals.** A point-wise confidence interval can be obtained by assuming that the Nelson-Aalen estimate at a given point in time is a sample from a normal distribution. We use the logarithmic transformation which was shown empirically to perform well for this kind of data, in particular for small sample sizes [25]. For each point estimate in age interval  $k$ , we first compute the two-sided  $1 - \alpha$  confidence interval for the cumulative hazard function

$$\begin{aligned}& [\hat{H}_{\text{lower}}(t_k), \hat{H}_{\text{upper}}(t_k)] \\ &= \left[ \hat{H}(t_k) \exp\left(-z_{1-\alpha/2} \frac{\text{Var}(\hat{H}(t_k))^{\frac{1}{2}}}{\hat{H}(t_k)}\right), \right. \\ & \quad \left. \hat{H}(t_k) \exp\left(z_{1-\alpha/2} \frac{\text{Var}(\hat{H}(t_k))^{\frac{1}{2}}}{\hat{H}(t_k)}\right) \right],\end{aligned}$$

which is feasible for  $\hat{H}(t_k) \neq 0$  and thus,

$$\begin{aligned}& [\hat{F}_{\text{lower}}(t_k), \hat{F}_{\text{upper}}(t_k)] \\ &= \left[ 1 - \exp\left(-\hat{H}(t_k)_{\text{lower}}\right), 1 - \exp\left(-\hat{H}(t_k)_{\text{upper}}\right) \right].\end{aligned}$$

In PLSD, we are interested in 95% confidence intervals and thus,  $\alpha = 0.05$  and  $z_{0.975} = 1.96$  is the 97.5% percentile of the standard normal distribution.

**Remark 8.4** (Small sample properties) Instead of a logarithmic transformation, also an arcsine-transformation is possible [25]. Following [25], error rates for standard confidence interval have been shown to be too high in simulations,

especially for  $n = 25$ . Considerable improvement was obtained by applying one of the transformed intervals. The logarithmic confidence interval seems to give slightly too low error rates, the arcsine-transformation slightly too high, but both acceptable.

In general, confidence intervals should be symmetric around the mean. Here, transformed intervals perform clearly better. Arcsine-transformed intervals seem to be somewhat better but similar to logarithmic transformation. However, in [25, Table 4], for no censoring and small sample sizes, logarithmic-transformed interval performs better such that we have chosen this type of confidence interval.

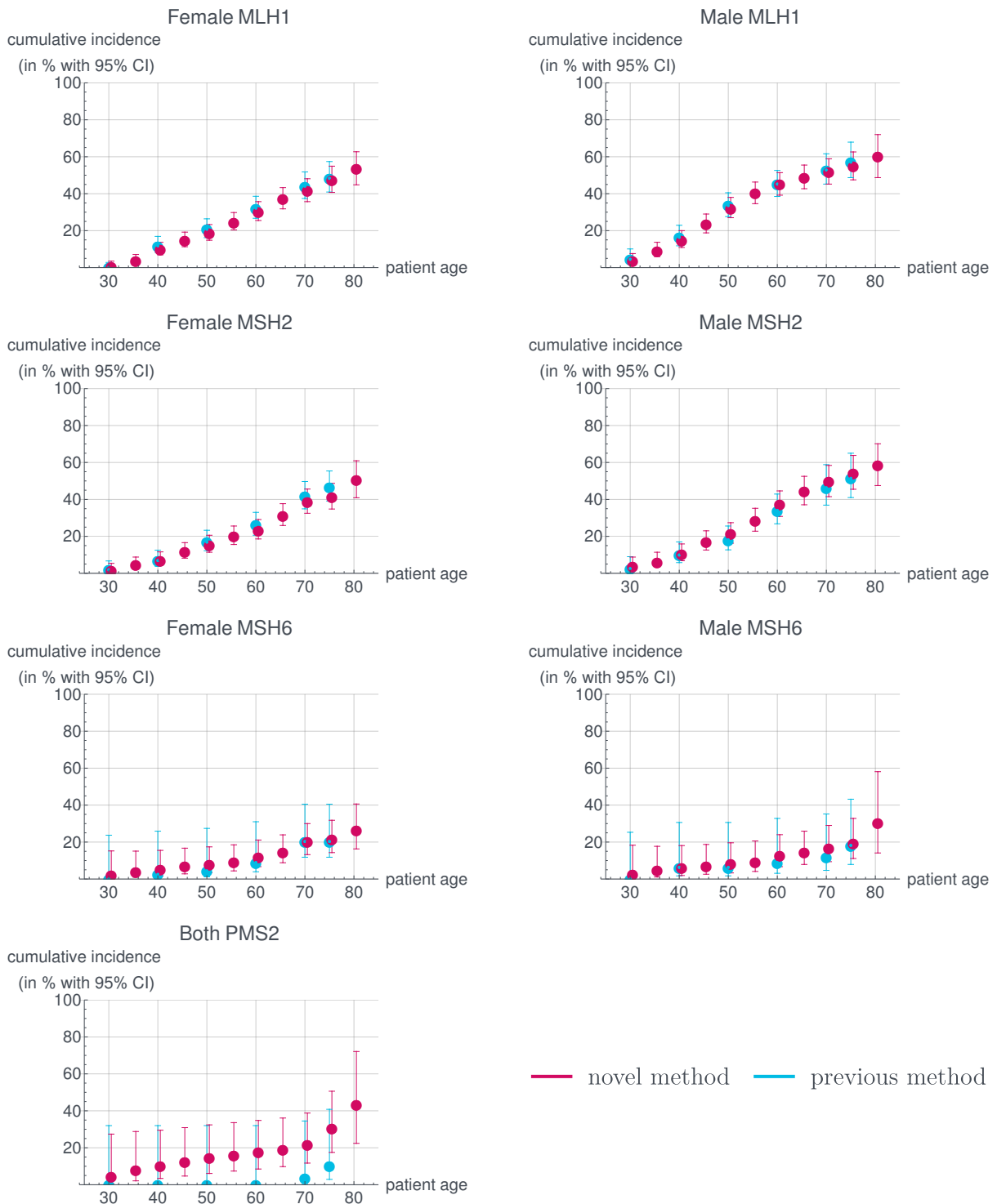
[25]: Bie et al. (1987), "Confidence Intervals and Confidence Bands for the Cumulative Hazard Rate Function and Their Small Sample Properties".

## 8.2 NOVEL COMPUTATION RESULTS WITH COMPARISON TO PREVIOUS APPROACH

We present the results for cumulative incidence risk of colorectal cancer in the Lynch syndrome population based on the previous and the novel method. The results stratified by gene and sex are presented in Figure 8.1.

Further results for other affected organs are obtained in the same way using the PLSD data, which is currently in preparation for publication and thus, not shown here.

As expected, the cumulative incidences calculated by the previous and novel methods were close to identical, while the Poisson distribution gave slightly different confidence intervals. This is likely due to the fact that the previously assumed Normal distribution does not explain the data well. The Poisson distribution is the natural choice for the considered data and thus, the presented novel approach will be used for further publications.



**Figure 8.1:** PLSD cumulative incidence risk estimates for colorectal cancer in the Lynch syndrome population. We show the results for the **previous method** and the **novel method** indicating the means ●, ●, and the 95% confidence intervals (error bars) per patient age. The results are slightly jittered in age for illustration purposes. The results are stratified by gene (*MLH1*, *MSH2*, *MSH6*, and *PMS2*) and sex (male and female). The differences between the two methods are not of practical significance.



## 8.3 OUTCOMES AND DISCUSSION

We presented an approach for calculating cumulative cancer risk and corresponding 95% confidence intervals for the PLSD data which is based on Nelson-Aalen estimates with underlying Poisson distribution. This is the natural mathematical choice for the present data consisting of the number of cancer cases occurring during a certain number of observation years, which is exactly the Poisson distribution definition.

From now on, this approach will be used for the upcoming versions of the PLSD database. A first application with implications for clinical guideline discussions is demonstrated in [143], which we closely follow. We tested the assumption that colonoscopy reduces colorectal cancer incidence in Lynch syndrome. Therefore, we compared the colorectal cancer cumulative incidences from the PLSD database, which were computed by the presented approach, with retrospective cohort data from IMRC. In the prospective PLSD cohort all were subjected to regular colonoscopy whereas the IMRC cohort included carriers who did not all receive regular colonoscopy.

We observed the cumulative incidences of colorectal cancer in *path\_MLH1* and *path\_MSH2* carriers of both genders were significantly higher in the prospective PLSD cohort than in the IMRC cohort. No significant differences were observed for *MSH6*, for which fewer patients and events were available in both cohorts. The point estimates for the mean for *path\_PMS2* carriers below 50 years of age indicated a lower colorectal cancer incidence in the PLSD cohort when compared to the IMRC cohort, but this was not statistically significant.

Thus, the hypothesis tested that surveillance colonoscopy would reduce colorectal cancer incidence [197], which is the paradigm underlying current health care for Lynch syndrome, was rejected.

Consideration of the methodologies used and the associated statistical concepts and confounders is indicated to explore the possibility that the results we obtained might reflect methodological biases, particularly as they were the opposite of what was expected.

[143]: Møller et al. (2022), “Colorectal cancer incidences in Lynch syndrome: a comparison of results from the prospective lynch syndrome database and the international mismatch repair consortium”.

[197]: Vasen et al. (1995), “Interval cancers in hereditary non-polyposis colorectal cancer (Lynch syndrome)”.

[141]: Møller (2020), “The Prospective Lynch Syndrome Database reports enable evidence-based personal precision health care”.

[209]: Win et al. (2021), “Variation in the risk of colorectal cancer in families with Lynch syndrome: a retrospective cohort study”.

[176]: Seppälä et al. (2021), “European guidelines from the EHTG and ESCP for Lynch syndrome: an updated third edition of the Mallorca guidelines based on gene and gender”.

[3]: Ahadova et al. (2020), “The unnatural history of colorectal cancer in Lynch syndrome: Lessons from colonoscopy surveillance”.

[197]: Vasen et al. (1995), “Interval cancers in hereditary non-polyposis colorectal cancer (Lynch syndrome)”.

[177]: Seppälä et al. (2019), “Lack of association between screening interval and cancer stage in Lynch syndrome may be accounted for by over-diagnosis: a prospective Lynch syndrome database report”.

[45]: Chalabi et al. (2020), “Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers”.

[200]: Versluis et al. (2020), “Learning from clinical trials of neoadjuvant checkpoint blockade”.

[106]: Kloor et al. (2020), “A Frameshift Peptide Neoantigen-Based Vaccine for Mismatch Repair-Deficient Cancers: A Phase I/IIa Clinical Trial”.

The PLSD methods have been described previously and discussed in detail [141] and the IMRC results were produced using commonly accepted methods, as previously described [209].

The *path\_PMS2* carriers in the PLSD cohort had lower incidence of colorectal cancer before 50 years of age than those reported by the IMRC, but the difference was not significant. That is, the assumption that colonoscopy does reduce colorectal cancer incidence may be true for young *path\_PMS2* carriers. If this finding is confirmed, the recently revised clinical guidelines for *path\_PMS2* carriers [176] that advocate postponing surveillance compared to other groups with Lynch syndrome should be reconsidered. Observations in larger numbers of *path\_PMS2* carriers are needed to clarify this.

A recent overview of current knowledge on carcinogenetic mechanisms in Lynch syndrome colorectal cancers [3] validated the former assumption that these may follow the adenoma-carcinoma pathway in some cases [197], but there are now additional carcinogenetic mechanisms to be considered as well. Five hypotheses on carcinogenetic mechanisms were described. These included 1) adenomas that are overlooked during colonoscopy, 2) fast progression of adenomas to carcinomas [197], 3) colorectal cancers developing without a macroscopically visible adenoma phase, 4) over-diagnosis or disappearing cancers [177], and 5) colonoscopy inducing cancer in *path\_MMR* carriers via damage of the colonic epithelium. Hypothesis 1 and 2 cannot explain the results described in this paper as we found higher rates of colorectal cancer incidence in those receiving colonoscopy. Although hypothesis 5 is consistent with our results, we have no method to evaluate this. We are left with hypotheses 3 and 4 that colorectal cancer may develop directly from the MMR deficient crypts without a macroscopically visible precursor and that microsatellite unstable crypts, or more advanced cancers, may be invaded by immunocompetent cells leading to eradication of the lesion. The latter underlies the principle of neoadjuvant checkpoint inhibitor therapy that has shown marked success in recent trials in MMR deficient colorectal cancers [45, 200] and current studies exploring the feasibility of vaccines to prevent or cure Lynch syndrome cancers [106].

An adult *path\_MLH1* or *path\_MSH2* carrier is thought to have over 1000 microsatellite unstable crypts in his/her colon [29, 105, 158, 182]. It is known that apparently healthy *path\_MMR* carriers have measurable immune responses against frameshift-induced neo-peptides suggesting their immune systems can detect and potentially attack microsatellite unstable crypts [174]. The probabilities for such crypts persisting, disappearing, or developing into infiltrating cancers are not known. The biology of colorectal cancer in *path\_PMS2* carriers may be different from carriers of the pathogenic variants of the three other genes [35, 36].

Although the focus of the results in [143] is on colorectal cancer incidence, we consider prevention of death due to colorectal cancer to be the main goal of surveillance colonoscopy in *path\_MMR* carriers, and the good prognosis of colorectal cancer detected in *path\_MMR* carriers who are subjected to colonoscopy every three year or more frequently has been described in previous PLSD reports [59]. This is a strong argument to continue surveillance of *path\_MMR* carriers by colonoscopy. The work presented in [143] does not call this into question, but its findings do support a change in the message to be communicated to *path\_MMR* carriers, namely that the purpose of surveillance colonoscopy is not to prevent colorectal cancer from occurring but to detect it early.

[29]: Brand et al. (2020), "Detection of DNA mismatch repair deficient crypts in random colonoscopic biopsies identifies Lynch syndrome patients".

[105]: Kloor et al. (2012), "Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study".

[158]: Pai et al. (2018), "DNA mismatch repair protein deficient non-neoplastic colonic crypts: a novel indicator of Lynch syndrome".

[182]: Staffa et al. (2015), "Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters".

[174]: Schwitalle et al. (2008), "Immune Response Against Frameshift-Induced Neopeptides in HNPCC Patients and Healthy HNPCC Mutation Carriers".

[35]: Broeke et al. (2018), "Molecular Background of Colorectal Tumors From Patients With Lynch Syndrome Associated With Germline Variants in PMS2".

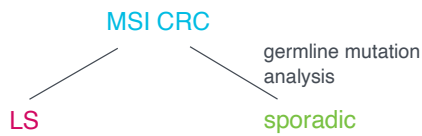
[36]: Broeke et al. (2021), "The coding microsatellite mutation profile of PMS2-deficient colorectal cancer".

[143]: Møller et al. (2022), "Colorectal cancer incidences in Lynch syndrome: a comparison of results from the prospective lynch syndrome database and the international mismatch repair consortium".

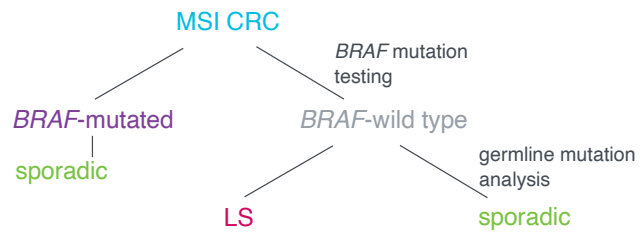
[59]: Dominguez-Valentin et al. (2019), "Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database".



### DIAGNOSTIC ALGORITHM 1



### DIAGNOSTIC ALGORITHM 2



## 9 AGE-DEPENDENT PERFORMANCE OF *BRAF* MUTATION TESTING: COST-BENEFIT ANALYSIS

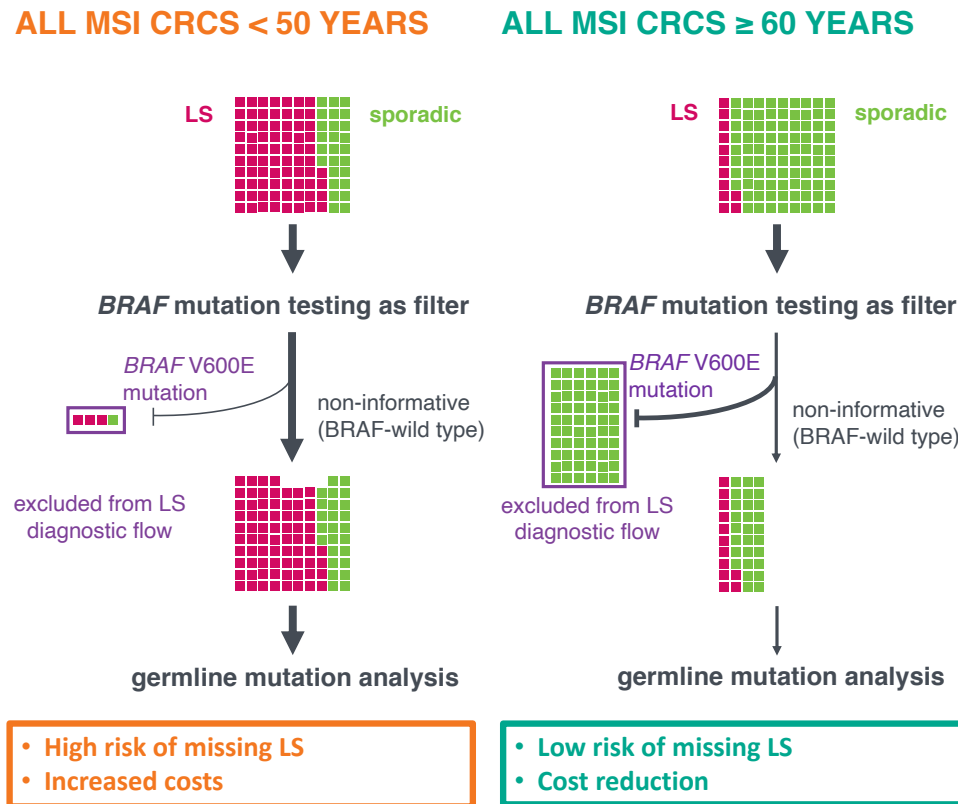
In this chapter, we present a probabilistic approach to evaluate cost and efficacy of two diagnostic procedures to detect Lynch syndrome among microsatellite unstable cancer samples depending on the age of the patient to be tested.

The accurate diagnosis of the type of colorectal cancer is of great importance for each patient. This is due to the different measures regarding prevention and treatment required for different cancer types. In particular, it is necessary to differentiate between sporadic and hereditary types of colorectal cancer as they are linked to different molecular conditions.

For distinction, efficient and sensitive diagnostic algorithms are needed and several established diagnostic algorithms exist. Here, we compare two of them which differentiate between sporadic MSI and Lynch syndrome-associated hereditary colorectal cancer with respect to cost efficiency and sensitivity. The two approaches, in the following called diagnostic algorithm 1 and 2, are schematically illustrated in the chapter image above. Diagnostic algorithm 1 is to test all MSI CRCs for Lynch syndrome via germline mutation analysis. Another in diagnostic guidelines often recommended approach, diagnostic algorithm 2, includes *BRAF* mutation testing in MSI CRC patients before germline mutation analysis to potentially reduce patients' mental stress and healthcare costs.

We evaluated the age-specific performance of *BRAF* mutation testing in Lynch syndrome diagnostics. We systematically compared the prevalence of *BRAF* mutations in Lynch

9.1	MEDICAL EVIDENCE FOR AGE-DEPENDENT LYNCH SYNDROME DIAGNOSTICS . . . . .	165
9.2	DATA COLLECTION . . . . .	167
9.3	COST-BENEFIT ANALYSIS . . . . .	170
9.4	OUTCOMES AND DISCUSSION . . . . .	177



**Figure 9.1:** Graphical summary. *BRAF* mutation testing correctly identifies sporadic MSI CRC patients and saves germline analysis costs in patients aged  $\geq 60$ , whereas in patients aged  $< 50$  years, *BRAF* mutation testing leads to misclassification of true Lynch syndrome carriers as sporadic MSI CRC patients and to increased costs of analyses. Adapted from [27].

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

syndrome-associated colorectal cancers and unselected MSI colorectal cancers in different age groups as available from published studies, databases and population-based patient cohorts [27]. Calculations for the risk of actual Lynch syndrome mutation carriers to be erroneously excluded from further germline mutation analysis as well as cost calculations were performed.

As illustrated in Figure 9.1, in patients of age  $< 50$  years, the use of *BRAF* mutation tests as a filter led to a high risk of erroneously excluding Lynch syndrome patients and increased healthcare costs for tumor type analyses. We thus concluded that *BRAF* mutation testing of diagnostic algorithm 2 in patients of age  $< 50$  years carries a high risk of missing a hereditary cancer predisposition and is cost-inefficient. In summary, MSI colorectal cancer patients younger than 50 years should directly be referred to genetic counseling without *BRAF* testing (diagnostic algorithm 1). The remaining chapter follows closely the work published in [27].

## 9.1 MEDICAL EVIDENCE FOR AGE-DEPENDENT LYNCH SYNDROME DIAGNOSTICS

Usually, Lynch syndrome-associated cancers show MMR deficiency and microsatellite instability, as introduced in Section 2.2.4. Thus, testing for those conditions is commonly the first step in diagnosing Lynch syndrome [194]. However, as most MSI tumors occur sporadically, MMR deficiency or microsatellite instability alone do not prove Lynch syndrome. Sporadic MSI tumors commonly occur in older patients with predominance for female gender, lack MLH1 protein expression due to *MLH1* promoter methylation and are strongly associated with the CpG island methylator phenotype and the serrated route of carcinogenesis related to the activating hotspot oncogenic mutation in the *BRAF* gene (c.1799T>A p.Val600Glu, also called V600E) [119] (see Section 2.2.4). The presence of MLH1-deficient MSI CRCs is therefore not a highly sensitive criterion to diagnose Lynch syndrome, particularly those occurring in the elderly. According to the NICE guidelines ([www.nice.org.uk/guidance/dg27/chapter/1-Recommendations](http://www.nice.org.uk/guidance/dg27/chapter/1-Recommendations)), MSI tumor testing for potential Lynch syndrome should not only be done for patients fulfilling the Bethesda criteria [28, 196] but in all colorectal cancers diagnosed before the age of 70 [198]. Thus, additional markers are required to differentiate Lynch syndrome and sporadic MSI CRC. Such markers should reduce the number of MSI CRC patients referred to germline mutation analysis, thus also reducing patients' mental stress and healthcare costs [58, 74, 196]. Deng et al [54] suggested the *BRAF* V600E mutation as a possible marker occurring in sporadic but not in Lynch syndrome-associated MSI CRC. This finding was later supported by studies reporting 100% specificity [125],[58, 137]. However, others, rarely detected *BRAF* mutations in Lynch syndrome-associated CRC [114, 149, 203], with a frequency of 1.4% determined in a meta-analysis [159]. Thompson et al estimated a frequency of 2.9% for the presence of *BRAF* mutations in LS CRC [190] in a clinic-based cohort.

The little number of *BRAF*-mutated MSI CRCs in Lynch syndrome may be not related to a tumor developing because of Lynch syndrome, but rather developing in a sporadic way

[194]: Umar et al. (2004), "Revised Bethesda Guidelines for Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndrome) and Microsatellite Instability".

[119]: Leggett and Whitehall (2010), "Role of the Serrated Pathway in Colorectal Cancer Pathogenesis".

[28]: Boland et al. (1998), "A National Cancer Institute Workshop on Microsatellite Instability for Cancer Detection and Familial Predisposition: Development of International Criteria for the Determination of Microsatellite Instability in Colorectal Cancer".

[196]: Vasen et al. (2007), "Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer)".

[198]: Vasen et al. (2013), "Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts".

[58]: Domingo (2004), "BRAF screening as a low-cost effective strategy for simplifying HNPCC genetic testing".

[74]: Giardiello et al. (2014), "Guidelines on Genetic Evaluation and Management of Lynch Syndrome: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer".

[54]: Deng (2004), "BRAF Mutation Is Frequently Present in Sporadic Colorectal Cancer with Methylated hMLH1, But Not in Hereditary Nonpolyposis Colorectal Cancer".

[125]: Loughrey et al. (2007), "Incorporation of somatic BRAF mutation testing into an algorithm for the investigation of hereditary non-polyposis colorectal cancer".

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

[33]: Brenner et al. (2014), “Reduced Risk of Colorectal Cancer Up to 10 Years After Screening, Surveillance, or Diagnostic Colonoscopy”.

[34]: Brenner et al. (2016), “Survival of patients with symptom- and screening-detected colorectal cancer”.

[85]: Hoffmeister et al. (2015), “Statin Use and Survival After Colorectal Cancer: The Importance of Comprehensive Confounder Adjustment”.

1: German Consortium for Hereditary Non-Polyposis Colorectal Cancer, [www.health-atlas.de/projects/13](http://www.health-atlas.de/projects/13)

2: (Darmkrebs: Chancen der Verhütung durch Screening/-Colorectal cancer: chances for prevention through screening), <http://dach.s.dkfz.org/dachs/>

3: Dana Farber Cancer Institute, <http://www.cbioportal.org/>

[44]: Cerami et al. (2012), “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1.”

[69]: Gao et al. (2013), “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal”.

irrespective of the hereditary predisposition of the patient. The probability of such a sporadic MSI CRC in Lynch syndrome patients is small, as the probability of getting an MSI CRC because of Lynch syndrome in younger ages is much higher.

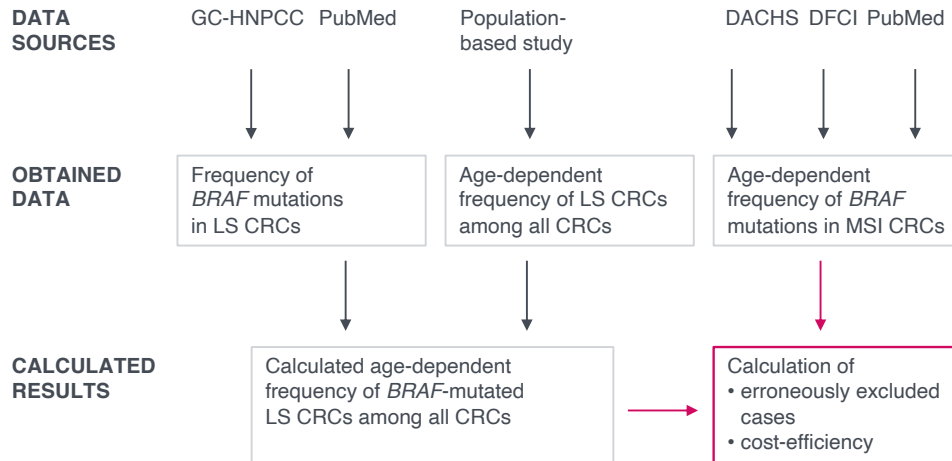
As illustrated in the chapter image at the beginning of this chapter, when performing diagnostic algorithm 2, all MSI CRCs are first tested for *BRAF* mutations. Then, the MSI CRCs without *BRAF* mutations, called *BRAF*-wild type (*BRAF*-wt) MSI CRCs, are tested for Lynch syndrome via germline mutation analysis. Using this procedure, the *BRAF*-mutated LS CRCs among all MSI CRCs may be erroneously excluded from germline mutation analysis. Therefore, these cases may be missed in the sense that these patients are not diagnosed as LS carriers. As sporadic MSI CRCs occur more often with increasing age, we hypothesized that the predictive value of *BRAF* V600E mutation for the exclusion of Lynch syndrome may depend on the age at diagnosis [27].

We analyzed Lynch syndrome- and population-based databases to determine the prevalence of *BRAF* mutations and the presence of Lynch syndrome germline variants in patients of different age groups. The data sets incorporate data from the GC-HNPCC<sup>1</sup>, which has been founded in 1999 by the German Cancer Aid and wants to improve the clinical care of patients with Lynch syndrome in Germany. Further, data collected by the DACHS<sup>2</sup> cohort study [33, 34, 85], which is an epidemiological case-control study of the German Cancer Research Center (DKFZ), was used. In addition, we included data from the DFCI<sup>3</sup> cancer genomics database [44, 69].

By using the three mentioned data sets, we computed the frequency of *BRAF*-mutated LS CRCs among all MSI CRCs. This in turn was used together with the prevalence of *BRAF*-mutations in MSI CRCs to compute the erroneously excluded cases using diagnostic algorithm 2. Further, we computed the costs of diagnostic algorithm 2 in comparison to diagnostic algorithm 1 for different age groups.

As all these probabilities are age-dependent, the goal was to make recommendations in which age groups *BRAF* mutational testing should be used as diagnostic algorithm with respect to cost efficiency and sensitivity. The overall procedure is illustrated in Figure 9.2.





**Figure 9.2:** Flow diagram for the calculation of the sensitivity and cost efficiency of *BRAF* mutation testing of MSI CRCs in LS diagnostics. This includes the data sources with corresponding data sets as well as the intermediate stages in the calculations with the desired output. Adapted from [27].

## 9.2 DATA COLLECTION

An overview of the used data sources is given in Figure 9.2. Further information about the literature review performed on NCBI PubMed with a schematic illustration of this process is given in [27]. In the following subsections, we will have a closer look at the single data sources.

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

### 9.2.1 FREQUENCY OF *BRAF* MUTATIONS IN LS CRCs

As a first input for our model, we collected data for the frequency of *BRAF* mutations in LS CRCs. This corresponds to the small part of *BRAF*-mutated MSI CRC overlapping with the LS CRCs in Figure 2.3. In formulas, this is the conditional probability of *BRAF*-mutations given Lynch syndrome colorectal cancers, i.e.  $\mathbb{P}(\text{BRAF-mut} \mid \text{LS, CRC})$ . For this, we used the database of the German HNPCC Consortium as well as a literature review on NCBI PubMed following PRISMA guidelines. These data sets are summarized in Table 9.1 with information about the MMR gene affected by a germline variant. A detailed description of the individual literature studies can be found in the supplementary material of [27].

As these studies did not give a detailed description of *BRAF* mutation status in LS CRCs with respect to age, we assumed  $\mathbb{P}(\text{BRAF-mut} \mid \text{LS, CRC}) = 1.6\%$  to be constant for all age groups.

**Table 9.1:** Prevalence of *BRAF* mutations in LS CRCs from the German HNPCC Consortium and the literature review including 30 studies. Information on the MMR gene affected by germline variant was available in 832 of the 969 cases. LS: Lynch syndrome. Not reported: MMR gene affected by germline variant was not specified. Reprinted from [27].

	All LS	<i>MLH1</i>	<i>MSH2</i>	<i>MSH6</i>	<i>PMS2</i>	not reported
German HNPCC database	98	74	14	4	6	0
Literature (28 studies)	871	408	255	23	48	137
Total	969	482	269	27	54	137
<i>BRAF</i> -mutations	15 (1.6%)	8 (1.7%)	2 (0.7%)	0	5 (9.3%)	0

**Table 9.2:** Frequency of LS among all CRCs. Reprinted from [144, Supplementary table 2].

	< 50 years	50 – 59 years	60 – 69 years	≥ 70 years
$\mathbb{P}(\text{LS} \mid \text{CRC})$	8.4%	2.9%	1.4%	0.8%

### 9.2.2 FREQUENCY OF LS CRCs AMONG ALL CRCs

Next, we needed data on the frequency of LS CRCs among all CRCs, which we obtained from the largest population-based study so far on this topic (Supplementary table 2 in [144]). In formulas, this reads  $\mathbb{P}(\text{LS} \mid \text{CRC})$  and corresponds to the **red box** in Figure 2.3 compared to the gray box of all CRCs. We calculated all probabilities for the age groups < 50, 50 – 59, 60 – 69 and ≥ 70 years as most of the studies summarize the data in these age groups. Thus, we also summarized the data in [144] as given in Table 9.2. This procedure included adding the data given in 5-year intervals to the used four age groups and computing the corresponding frequencies.

[144]: Moreira et al. (2012), “Identification of Lynch Syndrome Among Patients With Colorectal Cancer”.

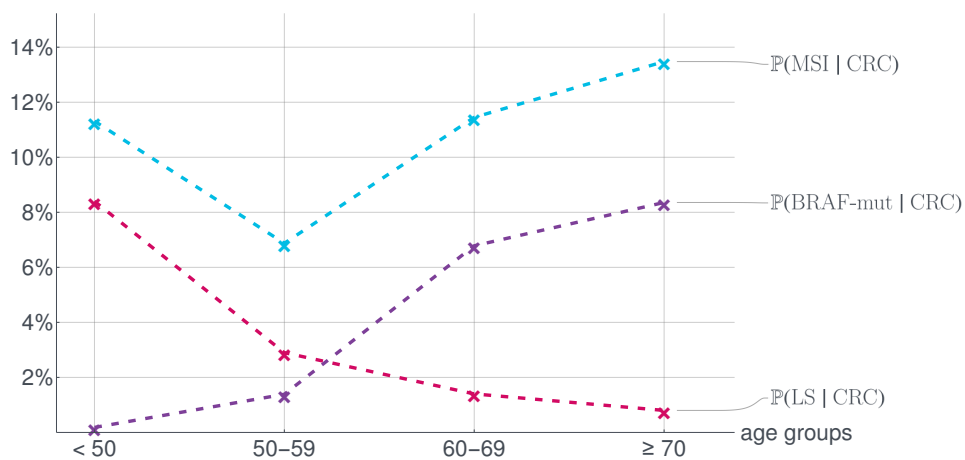
### 9.2.3 FREQUENCY OF *BRAF* MUTATIONS AMONG MSI CRCs

As last input, we used data on the age-specific prevalence of *BRAF* mutations in MSI CRCs from the DACHS and DFCI studies as well as from a literature review including seven studies (for further information, see [27]). In formulas, we collected age-dependent information on  $\mathbb{P}(\text{BRAF-mut} \mid \text{MSI, CRC})$ . From these data, we also obtained the frequency of *BRAF*-wild type MSI CRCs among all MSI CRCs, i.e.  $\mathbb{P}(\text{BRAF-wt} \mid \text{MSI, CRC})$ , and for all age groups, the frequencies  $\mathbb{P}(\text{BRAF-mut, MSI} \mid \text{CRC})$  and  $\mathbb{P}(\text{BRAF-wt, MSI} \mid \text{CRC})$ .

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

**Table 9.3:** Overview of data obtained by DACHS, DFCI as well as from seven published studies reporting on age-specific prevalence of molecular CRC subtypes stratified by *BRAF* and MSI analysis. This includes the frequency of *BRAF* mutations and *BRAF*-wild type among MSI CRCs, as well as the frequency of *BRAF*-wild type and *BRAF*-mutated MSI CRCs among all CRCs. Adapted from [27, Supplementary Table 4].

	< 50	50 – 59	60 – 69	≥ 70
$\mathbb{P}(\text{BRAF-mut} \mid \text{MSI, CRC})$	1.6%	20.2%	59.2%	62.0%
$\mathbb{P}(\text{BRAF-wt} \mid \text{MSI, CRC})$	98.4%	79.8%	40.8%	38.0%
$\mathbb{P}(\text{BRAF-mut, MSI} \mid \text{CRC})$	0.2% (2/1096)	1.4% (24/1731)	6.8% (109/1607)	8.4% (204/2441)
$\mathbb{P}(\text{BRAF-wt, MSI} \mid \text{CRC})$	11.1%	5.5%	4.7%	5.1%



**Figure 9.3:** Age-dependent frequency of MSI CRC, LS CRC and *BRAF*-mutated MSI CRC among all CRC. Reprinted from [27].

The corresponding probabilities are summarized in Table 9.3 and given in more detail in the supplementary material of [27]. For a summary of the collected data, we refer to Figure 9.3.

As a minor remark, the percentage of *BRAF*-mutated MSI CRCs among all MSI CRCs corresponds to the percentage of tests which are excluded from MMR gene germline mutation analysis by performing diagnostic algorithm 2. This means that here a certain percentage of MMR gene germline mutation analysis tests can be saved in comparison to diagnostic algorithm 1. We will have a closer look at this in Section 9.3.3.

We also state the corresponding number of patients of the included studies in Table 9.1. This is used in the next section to obtain insights into the potentially erroneously excluded cases by performing diagnostic algorithm 2.

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

## 9.3 COST-BENEFIT ANALYSIS

The following section describes the fundamental computational part of [27], namely how the data analysis was performed to obtain estimates about the performance and costs of diagnostic algorithm 2 in comparison to diagnostic algorithm 1.

### 9.3.1 CALCULATED AGE-SPECIFIC FREQUENCY OF *BRAF*-MUTATED LS CRCs AMONG ALL CRCs

Following the scheme illustrated in Figure 9.2, we used the data obtained in the last section to calculate the frequency of *BRAF*-mutated LS CRCs among all CRCs for all considered age groups. This was done by using the definition of conditional probabilities and we computed for each age group:

$$\mathbb{P}(BRAF\text{-mut, LS} \mid CRC) = \mathbb{P}(BRAF\text{-mut} \mid LS, CRC) \cdot \mathbb{P}(LS \mid CRC),$$

where  $\mathbb{P}(BRAF\text{-mut} \mid LS, CRC) = 1.6\%$  was assumed for all age groups (see Table 9.1).

We wanted to find out for each age group if there is a considerable risk of misdiagnosis. In other words, we quantified the risk of erroneously excluding an actual Lynch syndrome patient by *BRAF* testing as the patient shows *BRAF* mutations and the diagnostic algorithm 2 only tests *BRAF*-wild type MSI CRCs for Lynch syndrome. But there is a small proportion of *BRAF*-mutated LS CRCs among all LS CRCs, which may be due to the fact that they are actually sporadic MSI CRCs arising in Lynch syndrome patients independent of their Lynch syndrome carrier.

By comparing the calculated number of *BRAF*-mutated LS CRCs with the observed number of *BRAF*-mutated MSI CRCs in each age group, we obtained insights if there is a possible risk of erroneously excluding Lynch syndrome patients from germline mutation analysis. For this, we applied the calculated frequency of *BRAF*-mutated LS CRCs to the number of patients observed in the considered study cohorts and compared it to the number of *BRAF*-mutated MSI CRC patients observed.

**Table 9.4:** Comparison of the calculated number of *BRAF*-mutated LS CRCs and the number of observed *BRAF*-mutated MSI CRCs according to age groups. Adapted from [27].

	< 50	50 – 59	60 – 69	≥ 70
Total number of observed CRC cases $n_{\text{CRC}}$	1096	1731	1607	2441
Calculated <i>BRAF</i> -mut. LS CRC $n_{\text{CRC}} \cdot \mathbb{P}(\textit{BRAF}\text{-mut, LS} \mid \text{CRC})$	1.5	0.8	0.4	0.3
Observed <i>BRAF</i> -mut. MSI CRC $n_{\text{CRC}} \cdot \mathbb{P}(\textit{BRAF}\text{-mut, MSI} \mid \text{CRC})$	2	24	109	204
$\frac{\textit{BRAF}\text{-mut. LS CRC}}{\textit{BRAF}\text{-mut. MSI CRC}}$	75.0%	3.3%	0.4%	0.2%

If the two numbers are almost the same, the risk of misdiagnosis is high as the Lynch syndrome patients account for the majority of *BRAF*-mutated MSI CRCs which are excluded from further germline mutation analysis by diagnostic algorithm 2. On the other hand, if the number of *BRAF*-mutated LS CRCs is much smaller than the number of *BRAF*-mutated MSI CRCs, this risk is less significant. We state the results of this comparison in Table 9.4 and explain the consequences in Section 9.4.

### 9.3.2 CALCULATION OF ERRONEOUSLY EXCLUDED CASES

In order to predict the risk of erroneously excluded Lynch syndrome cases, we calculated

$$\begin{aligned}
 & \mathbb{P}(\textit{BRAF}\text{-mut, LS} \mid \text{MSI, CRC}) \\
 &= \frac{\mathbb{P}(\textit{BRAF}\text{-mut, LS, MSI, CRC})}{\mathbb{P}(\text{MSI, CRC})} \\
 &= \frac{\mathbb{P}(\textit{BRAF}\text{-mut, LS, CRC})}{\mathbb{P}(\text{MSI, CRC})} \\
 &= \frac{\mathbb{P}(\textit{BRAF}\text{-mut, LS} \mid \text{CRC}) \cdot \mathbb{P}(\text{CRC})}{\mathbb{P}(\text{MSI} \mid \text{CRC}) \cdot \mathbb{P}(\text{CRC})} \\
 &= \frac{\mathbb{P}(\textit{BRAF}\text{-mut, LS} \mid \text{CRC})}{\mathbb{P}(\text{MSI} \mid \text{CRC})},
 \end{aligned}$$

where we first used the definition of conditional probabilities. The second equality holds because the probability that a Lynch syndrome patient can get MSS CRC is neglectable.

**Table 9.5:** Calculated proportion of erroneously excluded Lynch syndrome mutation carriers among all MSI CRCs for different age groups.

	< 50	50 – 59	60 – 69	≥ 70
Calculated proportion of erroneously excluded LS cases	1.2%	0.7%	0.2%	0.1%

Note that the denominator can also be computed on the basis of the data obtained in the last section:

$$\begin{aligned}
& \mathbb{P}(\text{MSI} \mid \text{CRC}) \\
&= \frac{\mathbb{P}(\text{MSI}, \text{CRC})}{\mathbb{P}(\text{CRC})} \\
&= \frac{\mathbb{P}(\text{BRAF-mut}, \text{MSI}, \text{CRC}) + \mathbb{P}(\text{BRAF-wt}, \text{MSI}, \text{CRC})}{\mathbb{P}(\text{CRC})} \\
&= \frac{\mathbb{P}(\text{BRAF-mut}, \text{MSI} \mid \text{CRC}) \cdot \mathbb{P}(\text{CRC})}{\mathbb{P}(\text{CRC})} + \\
&\quad + \frac{\mathbb{P}(\text{BRAF-wt}, \text{MSI} \mid \text{CRC}) \cdot \mathbb{P}(\text{CRC})}{\mathbb{P}(\text{CRC})} \\
&= \mathbb{P}(\text{BRAF-mut}, \text{MSI} \mid \text{CRC}) + \mathbb{P}(\text{BRAF-wt}, \text{MSI} \mid \text{CRC}).
\end{aligned}$$

As it turned out, the proportion of erroneously excluded Lynch syndrome mutation carriers is age-dependent. It is low in older age groups, but substantially higher in younger age groups. The results are given in Table 9.5 and in Figure 9.4a.

### 9.3.3 CALCULATED PROPORTION OF MSI CRC EXCLUDED FROM MMR GENE GERMLINE ANALYSIS DUE TO BRAF MUTATION

When performing diagnostic algorithm 1, a germline mutation analysis is needed for all present MSI CRCs. Using diagnostic algorithm 2, not all MSI CRCs have to be considered for a germline mutation analysis, but only the ones which are BRAF-wild type. This leads to a certain percentage of MSI CRCs for each age group where a germline mutation analysis is **not** needed.

**Table 9.6:** Calculated percentage of saved MMR gene germline mutation tests among all MSI CRCs due to a positive *BRAF* mutation status for all age groups.

	< 50	50 – 59	60 – 69	≥ 70
Calculated percentage of saved germline mutation tests	1.6%	20.2%	59.2%	62.0%

This is the percentage of *BRAF*-mutated MSI CRCs among all MSI CRCs and we calculated it by

$$\begin{aligned}
 & \mathbb{P}(\text{BRAF-mut} \mid \text{MSI, CRC}) \\
 &= \frac{\mathbb{P}(\text{BRAF-mut, MSI, CRC})}{\mathbb{P}(\text{MSI, CRC})} \\
 &= \frac{\mathbb{P}(\text{BRAF-mut, MSI} \mid \text{CRC}) \cdot \mathbb{P}(\text{CRC})}{\mathbb{P}(\text{MSI} \mid \text{CRC}) \cdot \mathbb{P}(\text{CRC})} \\
 &= \frac{\mathbb{P}(\text{BRAF-mut, MSI} \mid \text{CRC})}{\mathbb{P}(\text{MSI} \mid \text{CRC})}.
 \end{aligned}$$

The results for this calculation are given in Table 9.6 and Figure 9.4b.

### 9.3.4 COST CALCULATIONS FOR BOTH DIAGNOSTIC ALGORITHMS

We compared the costs of the two diagnostic algorithms assuming a constant number of MSI CRC cases  $n_{\text{MSICases}}$  to be tested in each age group.

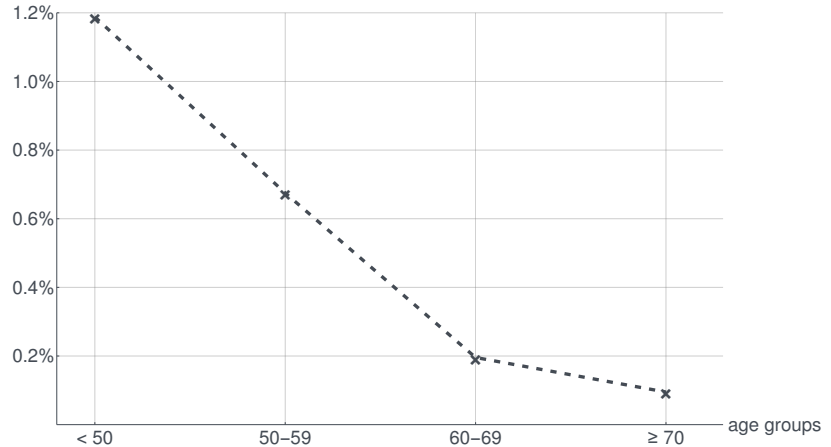
For the performance of diagnostic algorithm 1, all MSI CRC cases have to be analyzed for germline mutation analysis with the overall costs

$$C_{\text{Diag1}} = n_{\text{MSICases}} \cdot C_{\text{germlineTest}},$$

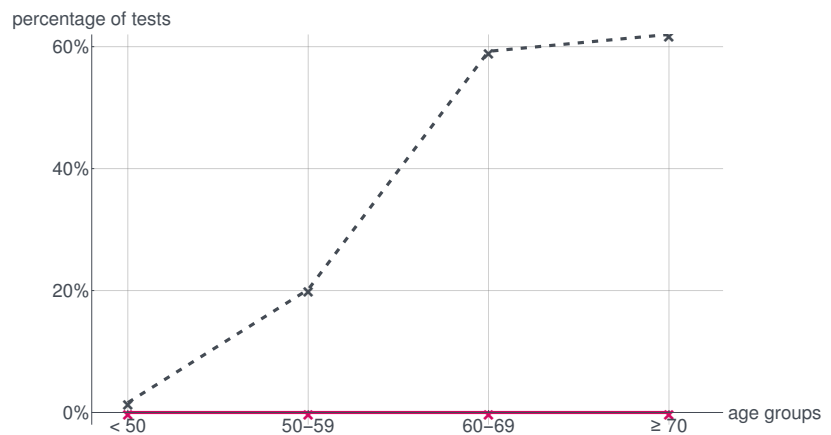
where  $C_{\text{germlineTest}}$  are the costs for the germline mutation analysis of one MSI CRC patient. We note that these costs are constant for all age groups, as we assumed a constant number of MSI CRC patients to be tested in a clinical setting in each age group.

Performing diagnostic algorithm 2 requires firstly *BRAF* mutation testing for all MSI CRC patients in each age group, where a single *BRAF* mutation testing has  $C_{\text{BRAFTest}}$  costs.

**Figure 9.4: Performance of the diagnostic algorithm 2 for the exclusion of LS according to age groups. (a)** The percentage of erroneously excluded Lynch syndrome mutation carriers is low in higher age groups. In patients younger than 60 years at diagnosis, the risk of missing Lynch syndrome by using diagnostic algorithm 2 with *BRAF* mutation testing increases. **(b)** *BRAF* mutation testing of MSI CRC only leads to a marginal reduction of MMR gene germline mutation analysis in younger age groups as only a very small proportion of MSI CRC can be excluded. A substantial reduction of required analyses is achieved in older age groups. Using diagnostic algorithm 1, all MSI CRCs undergo MMR gene germline mutation analyses which is used as a reference (0% of tests excluded from MMR gene germline mutation analysis, red line.) Adapted from [27].



**(a)** Calculated percentage of erroneously excluded cases given a constant number of MSI CRCs in each age group.



**(b)** Percentage of MSI CRC excluded from MMR gene germline mutation analysis due to *BRAF* mutation.

Then, the percentage of MSI CRCs having no *BRAF* mutation had to be analyzed for germline variants with the same costs per case as in the diagnostic algorithm 1. This led to the following overall costs for diagnostic algorithm 2:

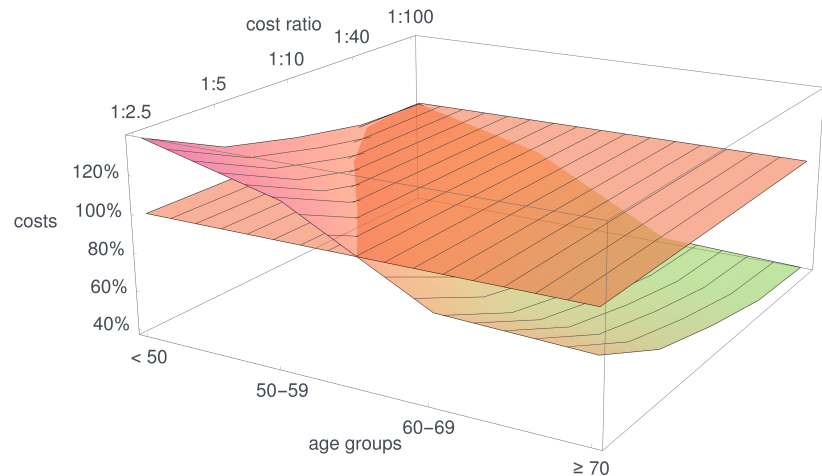
$$C_{\text{Diag2}} = n_{\text{MSI Cases}} \cdot (C_{\text{BRAF Test}} + \mathbb{P}(\text{BRAF-wt} \mid \text{MSI, CRC}) \cdot C_{\text{germline Test}}).$$

As seen in the last subsection, not all MSI CRCs have to be analyzed for germline mutations. Therefore, in clinical diagnostic procedures, a certain amount of money can probably be saved when performing diagnostic algorithm 2. This is only the case if the percentage of saved germline mutation analyses is such high that the additional costs for *BRAF* mutation testing are compensated. This depends on the cost ratio  $q$  of *BRAF* mutation testing and germline mutation analysis. In practice, this ratio differs across countries and even



across healthcare centers from 1 : 2.5 up to 1 : 100. Therefore, we calculated the costs for both diagnostic algorithms for different cost ratios  $q \in [2.5, 100]$ .

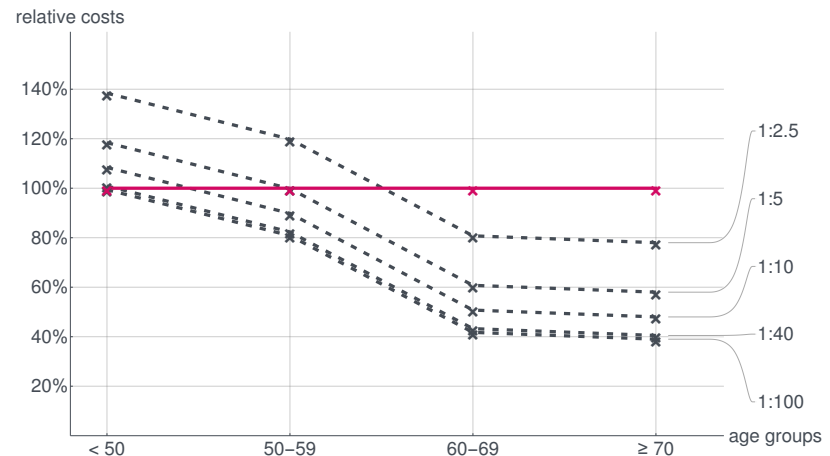
By assuming a constant number of MSI CRC patients in each age group, we used the costs for diagnostic algorithm 1 as a reference, which are constant for all age groups. The costs for diagnostic algorithm 2 decrease by age, as the proportion of MSI CRC cases with *BRAF*-wild type does so as well (see Figure 9.5b). Further, if *BRAF* mutation analysis is cheap, the overall costs for diagnostic algorithm 2 are lowest comparing the different cost ratios  $q$ , which is illustrated in Figure 9.5b. In patients of age < 50 years, the diagnostic algorithm 2 is more expensive than the diagnostic algorithm 1. This behavior changes for patients older than 60 years. For patients aged 50–59, this depends on the cost ratio  $q$ . The cost performance for distinct age groups is given in Figure 9.5c.



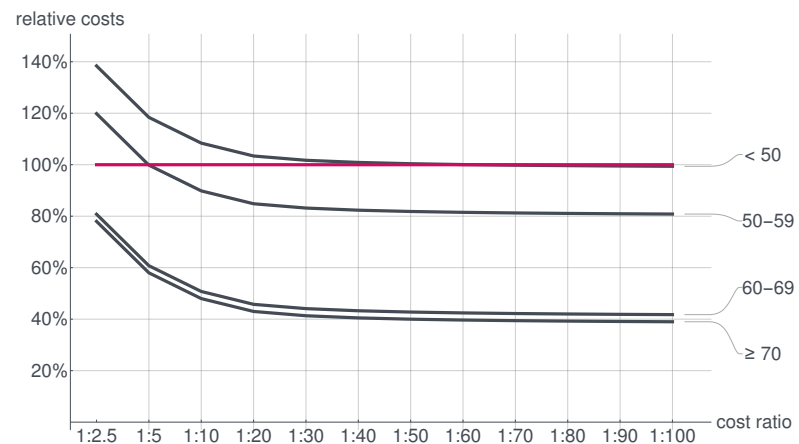
(a) 3D contour plot for all age groups and cost ratios.

**Figure 9.5: Cost calculations for both diagnostic algorithms.**

The costs for diagnostic algorithm 1 for performing MMR gene germline mutation analysis for all MSI CRC patients serves as a reference (reference plane in (a), and red line in (b), (c)). The costs for performing diagnostic algorithm 2, i.e. *BRAF* mutation testing of all MSI CRCs followed by MMR gene germline mutation analysis for *BRAF*-wild type MSI CRCs only are shown by the mountain surface in (a) and by the gray lines in (b), (c). *BRAF* mutations lead to a significant reduction of germline mutation analysis in older age groups, whereas the implementation of *BRAF* mutation testing for the exclusion of Lynch syndrome in patients younger than 50 years at diagnosis leads to a cost increase for all ratios of *BRAF* mutation costs relative to costs of MMR gene germline variant analysis. In addition to failure of *BRAF* mutation testing with regard to cost reduction, implementation of *BRAF* mutation testing in patients younger than 50 years also has the risk of missing hereditary cancer patients. Adapted from [27].



(b) Selective 2D illustration for different cost ratios.



(c) Selective 2D illustration for different age groups.

## 9.4 OUTCOMES AND DISCUSSION

Several systematic reviews and cost-effectiveness studies have analyzed different diagnostic algorithms for Lynch syndrome, also accounting for the age of the patient. However, only upper age limits of testing tumors for MSI have been considered and we are not aware of any studies analyzing the impact of patient's age on the performance of *BRAF* mutation testing.

Our results demonstrate that *BRAF* mutation testing in MSI CRC from patients below the age of 50 is not justified in Lynch syndrome diagnostics. As *BRAF*-mutated MSI CRCs can also occur in Lynch syndrome individuals, a present *BRAF* V600E mutation should not lead to exclusion of patients younger than 50 years from MMR gene germline analysis. By this means, *BRAF* mutation testing in patients < 50 has a substantial risk of erroneously excluding actual Lynch syndrome mutation carriers. Even more, *BRAF* mutation testing in patients < 50 is not cost-effective but may substantially delay the Lynch syndrome diagnostics flow (see Figure 9.1, left). For those patients, besides the increased healthcare costs, the physical burden and mental stress might be even higher due to a delayed and probably inappropriate treatment.

In patients aged between 50 and 60, a potential cost reduction depends on the cost ratio between *BRAF* mutation testing and MMR gene germline analysis. We found that a cost reduction is possible whenever *BRAF* mutation testing costs less than 20% of the price for MMR gene germline mutation analysis.

For patients > 60 years, *BRAF* mutation testing is confirmed to be a useful tool in Lynch syndrome diagnostics as the risk of erroneously excluding Lynch syndrome individuals is small and the cost reduction across all cost ratios is large (see Figure 9.1, right).

In general, the difficulty with these diagnostic approaches is the general purpose of determining the MMR gene germline status based on indirect measures such as *BRAF* mutation status or *MLH1* promoter methylation status [52, 127, 149, 199]. Thus, several studies have suggested direct MMR gene sequencing from tumor tissue [80] which is expected to gain more importance in the future due to decreased costs for NGS-based tumor sequencing.

[52]: Cunningham et al. (1998), "Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability".

[127]: Lynch et al. (2009), "Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications".

[149]: Newton et al. (2014), "Tumour MLH1 promoter region methylation testing is an effective prescreen for Lynch Syndrome (HNPCC)".

[199]: Veigl et al. (1998), "Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers".

[80]: Hampel et al. (2018), "Assessment of Tumor Sequencing as a Replacement for Lynch Syndrome Screening and Current Molecular Tests for Patients With Colorectal Cancer".

A strength of the study is the large number of data from CRCs collected in population-based cohorts. One limitation due to the available data is the assumption that the *BRAF* mutation frequency in LS CRCs is constant over time. Few cases with available age information showed a high proportion of *BRAF*-mutated LS CRCs in patients aged < 50. Besides that, cost calculations assumed a constant number of MSI CRCs to be tested in each age group which might not reflect the daily clinical practice. Future studies for clinical practice could base these calculations on age prevalence data of MSI CRCs in the general population. Further, some overlaps of tumors reported in different studies can not fully be excluded.

In conclusion, using the available data resources and introducing corresponding cost and efficiency analyzes, we recommend to directly refer MSI CRC patients younger than 50 years to genetic counseling without prior *BRAF* mutation testing.

## CONCLUSION



## 10.1 SUMMARY

10.1 SUMMARY . . . . . 181

10.2 OUTLOOK . . . . . 184

In this dissertation, several mathematical, computational, and statistical models have been developed to model various aspects of colorectal cancer development at different scales in the context of Lynch syndrome, the most common inherited colorectal cancer predisposition syndrome.

Throughout this dissertation, we have seen that the application of mathematics to oncology allows for **(1) unraveling the black boxes of cancer development at small scales** as in particular many processes at the smaller scales are hard to observe *in vivo*. This is in particular true for the DNA, cell and crypt level in colorectal cancer development for which we developed parametrizations for different types of genetic alterations in Chapter 4 all parameters of which have a medical interpretation and where current biomedical databases could be used for calibration (see also [81, 82]).

These parametrizations were subsequently used for the computational model at the cell level in Chapter 6 describing very early processes of cancer development namely the spread of different types of mutations within individual crypts [81]. The latter gives insight into the duration of monoclonal conversion for different types of mutations being drivers in Lynch syndrome colorectal carcinogenesis as first *in silico* estimates for key components of cancer initiation.

Besides that, the DNA alteration parametrizations built the basis for the mathematical model at the crypt level describing Lynch syndrome colorectal carcinogenesis with crypts as the smallest entities, developed in Chapter 7. By using the Kronecker structure for the model matrix of the ordinary differential equation system, each component has a medical interpretation which is crucial for the interaction with tumor biologists and clinicians. It further enables a rigorous mathematical analysis, allows for resource saving computations and a straight forward extension and modification for other driver mutations and pathways of carcinogenesis [82].

[81]: Haupt et al. (2021), “A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis”.

[82]: Haupt et al. (2021), “Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure”.

To sum up, by developing the mathematical modeling approaches at the DNA, cellular and crypt level, we were able to unravel some of the black boxes of cancer development at the small scales which was an important goal of this dissertation.

In addition to an improved understanding of cancer development with the help of mathematical modeling, using mathematical and bioinformatics tools, we have been able to make sense of a large amount of different biomedical and clinical data and to extract important information to support the development of future cancer prevention, diagnosis, and treatment procedures. As a first approach in this direction, we developed mathematical and bioinformatics tools for **(2) quantifying cancer immunology and different influencing factors**. Especially at the DNA level, we focused on immunoediting during carcinogenesis in Chapter 5. As a first essential step, our collaboration partners at ATB Heidelberg initiated a project to develop the ReFrame algorithm for quantifying the landscape of frameshift mutations (see also [17]). Further, in collaboration, we analyzed the extent of immuno-editing for different frameshift peptides using immunological scores which have been adapted to quantify the influence of the HLA type on tumor-immune interactions (see also [210]).

[17]: Ballhausen et al. (2020), “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution”.

[210]: Witt et al. (2022), “A simple approach for detecting *HLA-A\*02* alleles in archival formalin-fixed paraffin-embedded tissue samples and an application example for studying cancer immunoediting”.

Further, as the third part of this dissertation, we used and developed statistical methods for **(3) estimating the cumulative cancer risk of Lynch syndrome individuals and predicting the efficacy of clinical prevention and *BRAF*-mutation testing in Lynch syndrome diagnostics**. For the former, we used large databases on the population level to derive statistical cumulative cancer risk estimates in the Lynch syndrome population from the Prospective Lynch Syndrome Database (PLSD) in Chapter 8. These estimates have been subsequently used to compare the colorectal cancer risk with and without colonoscopy assuming theoretical values of colonoscopy sensitivity to quantify the effect of colonoscopy screening as a preventive measure for the Lynch syndrome population (see also [143]).

[143]: Møller et al. (2022), “Colorectal cancer incidences in Lynch syndrome: a comparison of results from the prospective lynch syndrome database and the international mismatch repair consortium”.

Furthermore, for the latter, we performed a cost-benefit analysis in Chapter 9 comparing two currently used diagnostic procedures for the detection of Lynch syndrome taking into account the patient’s age. We combined data from different databases using a probabilistic approach. By this, we could suggest a refinement of current clinical guidelines for Lynch



syndrome to consider the patient age when deciding on the diagnostic procedure to be used (see also [27]).

In summary, we could address in this dissertation some important aspects of current cancer research care, generating *in silico* estimates, simulations and knowledge of colorectal cancer development in Lynch syndrome in concordance with currently available biomedical and clinical data that have to be further validated by future *in vitro* and *in vivo* studies.

As a final side note, we believe that science communication is an important part of research. Thus, we have been quite active in this field, publishing press releases on several papers and the overall collaboration project. Further, we want to emphasize the blog posts we have written for the official *Mathematical Oncology Blog* explaining our research to a wider scientific audience interested in mathematical oncology. First, in the blog post for the Kronecker model [mathematical-oncology.org/blog/modeling-carcinogenesis-using-kronecker-structure.html](https://mathematical-oncology.org/blog/modeling-carcinogenesis-using-kronecker-structure.html) written by Saskia Haupt, we explain why the Kronecker structure is useful for modeling carcinogenic processes and most importantly how it can be applied to other types of cancer to also advertise and broaden the application field beyond Lynch syndrome. A second blog post written by Aysel Ahadova and Saskia Haupt [mathematical-oncology.org/blog/mathematics-for-deciphering-molecular-pathways.html](https://mathematical-oncology.org/blog/mathematics-for-deciphering-molecular-pathways.html) provides insights into our collaborative work on Lynch syndrome carcinogenesis in general, what is Lynch syndrome and why is it such useful to apply mathematics to this research field. In this context, Aysel elaborates from a medical and clinical point of view which questions in current Lynch syndrome cancer research could be answered by mathematics. In our opinion, fostering close collaborations between mathematicians and oncologists is essential to bring cancer research forward to improve cancer treatment and prevention in particular. As a final word, we quote The New York Times from 05/12/2011 „Computer scientists [and mathematicians] may have what it takes to help cure cancer“.

[27]: Bläker et al. (2020), “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”.

## 10.2 OUTLOOK

As mathematical oncology and in particular the focus on Lynch syndrome is quite a young research field mainly developed in the last twenty years, there are plenty of possibilities for future research directions. We want to highlight some of them with focus on the work we have done within this Ph.D.

- ▶ **Calibration and validation data for the modeling approaches.** The modeling approaches developed in this dissertation tried to use as much currently available information as possible to calibrate and validate the modeling results. However, in particular with increasing complexity of the mathematical models, more data are necessary to allow for realistic simulations of the underlying carcinogenic processes. Here, being able to include different types of data ranging from omics data over clinical parameters to population-based databases could be highly interesting and innovative for both the medical and mathematical research communities. In this direction, also optimal experimental design (OED) strategies could be useful for efficient data generation reducing laboratory costs and work.
- ▶ **Uncertainty Quantification and sensitivity analyses.** Uncertainty Quantification is a very important research field to on the one hand quantify the impact of uncertain input parameters like uncertainties in omics data, inexact laboratory measurements, incomplete clinical information, on the model solution which in our case might be the risk for developing colorectal cancer in Lynch syndrome. On the other hand, determining factors that might influence the different pathways of carcinogenesis and identifying parameters that have to be measured in an accurate way to obtain reliable simulation results are important tasks for future mathematical oncology research.
- ▶ **Development of data-driven multi-scale modeling approaches combining artificial intelligence and mathematical modeling.** The current approaches mainly model processes at single scales. However, for a comprehensive picture of cancer development and related processes, multi-scale models should be developed which goes in hand with largely increasing complex-

ity and computational resources. For the calibration and validation of these models, feature extraction and learning techniques from artificial intelligence could be highly beneficial. Thus, in our opinion, mathematical modeling and artificial intelligence should not be two contrary research fields but quite the opposite, the connection and interaction of both fields should be the long-term research direction not only for cancer modeling but for any kind of modeling applications in medicine and beyond.



## BIBLIOGRAPHY

- [1] A. Ahadova et al. "CTNNB1-mutant colorectal carcinomas with immediate invasive growth: a model of interval cancers in Lynch syndrome". In: *Familial Cancer* 15.4 (Mar. 2016), pp. 579–586. doi: [10.1007/s10689-016-9899-z](https://doi.org/10.1007/s10689-016-9899-z).
- [2] A. Ahadova et al. "Three molecular pathways model colorectal carcinogenesis in Lynch syndrome". In: *International Journal of Cancer* 143.1 (Feb. 2018), pp. 139–150. doi: [10.1002/ijc.31300](https://doi.org/10.1002/ijc.31300).
- [3] A. Ahadova et al. "The unnatural history of colorectal cancer in Lynch syndrome: Lessons from colonoscopy surveillance". In: *International Journal of Cancer* 148.4 (Aug. 2020), pp. 800–811. doi: [10.1002/ijc.33224](https://doi.org/10.1002/ijc.33224).
- [4] A. Ahadova et al. "Distinct Mutational Profile of Lynch Syndrome Colorectal Cancers Diagnosed under Regular Colonoscopy Surveillance". In: *Journal of Clinical Medicine* 10.11 (June 2021), p. 2458. doi: [10.3390/jcm10112458](https://doi.org/10.3390/jcm10112458).
- [5] A. Ahadova et al. "Is HLA type a possible cancer risk modifier in Lynch syndrome?" In: *International Journal of Cancer* (Oct. 2022). doi: [10.1002/ijc.34312](https://doi.org/10.1002/ijc.34312).
- [6] B. Alberts et al. *Molecular Biology of the Cell*. W.W. Norton & Company, Dec. 2007.
- [7] P. Alhopuro et al. "Candidate driver genes in microsatellite-unstable colorectal cancer". In: *International Journal of Cancer* 130.7 (Aug. 2011), pp. 1558–1566. doi: [10.1002/ijc.26167](https://doi.org/10.1002/ijc.26167).
- [8] A. Araujo et al. "Testing three hypotheses of the contribution of geometry and migration dynamics to intestine crypt evolution". In: *Artificial Life Conference Proceedings*. MIT Press. 2018, pp. 420–427.
- [9] A. Araujo et al. "Investigating the Origins of Cancer in the Intestinal Crypt with a Gene Network Agent Based Hybrid Model". In: *Artificial Life Conference Proceedings*. MIT Press. 2019, pp. 195–202.
- [10] P. Armitage and R. Doll. "The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis". In: *British Journal of Cancer* 8.1 (Mar. 1954), pp. 1–12. doi: [10.1038/bjc.1954.1](https://doi.org/10.1038/bjc.1954.1).
- [11] P. Armitage and R. Doll. "A Two-stage Theory of Carcinogenesis in Relation to the Age Distribution of Human Cancer". In: *British Journal of Cancer* 11.2 (June 1957), pp. 161–169. doi: [10.1038/bjc.1957.22](https://doi.org/10.1038/bjc.1957.22).
- [12] A. Arnold et al. "The majority of  $\beta$ -catenin mutations in colorectal cancer is homozygous". In: *BMC Cancer* 20.1 (Oct. 2020). doi: [10.1186/s12885-020-07537-2](https://doi.org/10.1186/s12885-020-07537-2).
- [13] R. Ashkenazi, S. N. Gentry, and T. L. Jackson. "Pathways to Tumorigenesis: Modeling Mutation Acquisition in Stem Cells and Their Progeny". en. In: *Neoplasia* 10.11 (Nov. 2008), 1170–IN6. doi: [10.1593/neo.08572](https://doi.org/10.1593/neo.08572).

- [14] A.-M. Baker and T. A. Graham. “Quantifying human intestinal stem cell and crypt dynamics: the implications for cancer screening and prevention”. In: *Expert Review of Gastroenterology & Hepatology* 10.3 (2016), pp. 277–279. DOI: [10.1586/17474124.2016.1134314](https://doi.org/10.1586/17474124.2016.1134314).
- [15] A.-M. Baker et al. “Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon”. In: *Cell Reports* 8.4 (Aug. 2014), pp. 940–947. DOI: [10.1016/j.celrep.2014.07.019](https://doi.org/10.1016/j.celrep.2014.07.019).
- [16] A.-M. Baker et al. “Crypt fusion as a homeostatic mechanism in the human colon”. In: *Gut* (2019), gutjnl–2018–317540. DOI: [10.1136/gutjnl-2018-317540](https://doi.org/10.1136/gutjnl-2018-317540).
- [17] A. Ballhausen et al. “The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution”. In: *Nature Communications* 11.1 (Sept. 21, 2020), pp. 1–13. DOI: [10.1038/s41467-020-18514-5](https://doi.org/10.1038/s41467-020-18514-5).
- [18] N. Barker et al. “Identification of stem cells in small intestine and colon by marker gene *Lgr5*”. In: *Nature* 449.7165 (Oct. 2007), pp. 1003–1007. DOI: [10.1038/nature06196](https://doi.org/10.1038/nature06196).
- [19] K.-H. Bauer. “Mutationstheorie der Geschwulstentstehung. Berlin 1928”. In: *Aktuelle Krebsfragen. Langenbecks Arch. klin. Chir* 287 (1957), p. 19.
- [20] N. Beerenwinkel and S. Sullivant. “Markov models for accumulating mutations”. In: *Biometrika* 96.3 (June 2009), pp. 645–661. DOI: [10.1093/biomet/asp023](https://doi.org/10.1093/biomet/asp023).
- [21] N. Beerenwinkel, N. Eriksson, and B. Sturmfels. “Evolution on distributive lattices”. In: *Journal of Theoretical Biology* 242.2 (Sept. 2006), pp. 409–420. DOI: [10.1016/j.jtbi.2006.03.013](https://doi.org/10.1016/j.jtbi.2006.03.013).
- [22] N. Beerenwinkel et al. “Learning Multiple Evolutionary Pathways from Cross-Sectional Data”. In: *Journal of Computational Biology* 12.6 (July 2005), pp. 584–598. DOI: [10.1089/cmb.2005.12.584](https://doi.org/10.1089/cmb.2005.12.584).
- [23] S. Biagi and A. Bonfiglioli. *An Introduction to the Geometrical Analysis of Vector Fields: With Applications to Maximum Principles and Lie Groups*. WORLD SCIENTIFIC, Dec. 2018.
- [24] E. Bianconi et al. “An estimation of the number of cells in the human body”. In: *Annals of Human Biology* 40.6 (July 2013), pp. 463–471. DOI: [10.3109/03014460.2013.807878](https://doi.org/10.3109/03014460.2013.807878).
- [25] O. Bie, Ø. Borgan, and K. Liestøl. “Confidence Intervals and Confidence Bands for the Cumulative Hazard Rate Function and Their Small Sample Properties”. In: *Scandinavian Journal of Statistics* 14.3 (1987), pp. 221–233.
- [26] H. Binder et al. “Genomic and transcriptomic heterogeneity of colorectal tumours arising in Lynch syndrome”. In: *The Journal of Pathology* 243.2 (Sept. 2017), pp. 242–254. DOI: [10.1002/path.4948](https://doi.org/10.1002/path.4948).
- [27] H. Bläker et al. “Age-dependent performance of *BRAF* mutation testing in Lynch syndrome diagnostics”. In: *International Journal of Cancer* 147.10 (Sept. 2020), pp. 2801–2810. DOI: [10.1002/ijc.33273](https://doi.org/10.1002/ijc.33273).

- [28] C. R. Boland et al. "A National Cancer Institute Workshop on Microsatellite Instability for Cancer Detection and Familial Predisposition: Development of International Criteria for the Determination of Microsatellite Instability in Colorectal Cancer". In: *Cancer Research* 58.22 (1998), pp. 5248–5257.
- [29] R. E. Brand et al. "Detection of DNA mismatch repair deficient crypts in random colonoscopic biopsies identifies Lynch syndrome patients". In: *Familial Cancer* 19.2 (Jan. 2020), pp. 169–175. DOI: [10.1007/s10689-020-00161-w](https://doi.org/10.1007/s10689-020-00161-w).
- [30] R. Bravo and D. E. Axelrod. "A calibrated agent-based computer model of stochastic cell dynamics in normal human colon crypts useful for in silico experiments". In: *Theoretical Biology and Medical Modelling* 10.1 (Nov. 2013). DOI: [10.1186/1742-4682-10-66](https://doi.org/10.1186/1742-4682-10-66).
- [31] H. Brenner et al. "Sojourn Time of Preclinical Colorectal Cancer by Sex and Age: Estimates From the German National Screening Colonoscopy Database". In: *American Journal of Epidemiology* 174.10 (Oct. 2011), pp. 1140–1146. DOI: [10.1093/aje/kwr188](https://doi.org/10.1093/aje/kwr188).
- [32] H. Brenner et al. "Low Risk of Colorectal Cancer and Advanced Adenomas More Than 10 Years After Negative Colonoscopy". In: *Gastroenterology* 138.3 (Mar. 2010), pp. 870–876. DOI: [10.1053/j.gastro.2009.10.054](https://doi.org/10.1053/j.gastro.2009.10.054).
- [33] H. Brenner et al. "Reduced Risk of Colorectal Cancer Up to 10 Years After Screening, Surveillance, or Diagnostic Colonoscopy". In: *Gastroenterology* 146.3 (Mar. 2014), pp. 709–717. DOI: [10.1053/j.gastro.2013.09.001](https://doi.org/10.1053/j.gastro.2013.09.001).
- [34] H. Brenner et al. "Survival of patients with symptom- and screening-detected colorectal cancer". In: *Oncotarget* 7.28 (May 2016), pp. 44695–44704. DOI: [10.18632/oncotarget.9412](https://doi.org/10.18632/oncotarget.9412).
- [35] S. W. ten Broeke et al. "Molecular Background of Colorectal Tumors From Patients With Lynch Syndrome Associated With Germline Variants in PMS2". In: *Gastroenterology* 155.3 (Sept. 2018), pp. 844–851. DOI: [10.1053/j.gastro.2018.05.020](https://doi.org/10.1053/j.gastro.2018.05.020).
- [36] S. W. B. .-. ten Broeke et al. "The coding microsatellite mutation profile of PMS2-deficient colorectal cancer". In: *Experimental and Molecular Pathology* 122 (Oct. 2021), p. 104668. DOI: [10.1016/j.yexmp.2021.104668](https://doi.org/10.1016/j.yexmp.2021.104668).
- [37] J. J. Buckley. "Fuzzy markov chains". In: *Fuzzy Probabilities*. Springer, 2003, pp. 71–83.
- [38] A. Buckowitz. "Mikrosatelliteninstabilität, lokale lymphozytäre Infiltration und ihre Bedeutung für Stadium und Prognose kolorektaler Karzinome". In: (2005).
- [39] D. Burini, E. Angelis, and M. Lachowicz. "A Continuous-Time Markov Chain Modeling Cancer-Immune System Interactions". In: *Communications in Applied and Industrial Mathematics* 9 (Dec. 2018), pp. 106–118. DOI: [10.2478/caim-2018-0018](https://doi.org/10.2478/caim-2018-0018).
- [40] J. Burn et al. "Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial". In: *The Lancet* 378.9809 (Dec. 2011), pp. 2081–2087. DOI: [10.1016/s0140-6736\(11\)61049-0](https://doi.org/10.1016/s0140-6736(11)61049-0).

- [41] E. Busch et al. "Beta-2-microglobulin Mutations Are Linked to a Distinct Metastatic Pattern and a Favorable Outcome in Microsatellite-Unstable Stage IV Gastrointestinal Cancers". In: *Frontiers in Oncology* 11 (June 2021). doi: [10.3389/fonc.2021.669774](https://doi.org/10.3389/fonc.2021.669774).
- [42] P. Buske et al. "A Comprehensive Model of the Spatio-Temporal Stem Cell and Tissue Organisation in the Intestinal Crypt". In: *PLoS Computational Biology* 7.1 (Jan. 2011). Ed. by D. Lauffenburger, e1001045. doi: [10.1371/journal.pcbi.1001045](https://doi.org/10.1371/journal.pcbi.1001045).
- [43] J. M. Carethers. "Differentiating Lynch-Like From Lynch Syndrome". In: *Gastroenterology* 146.3 (Mar. 2014), pp. 602–604. doi: [10.1053/j.gastro.2014.01.041](https://doi.org/10.1053/j.gastro.2014.01.041).
- [44] E. Cerami et al. "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1." In: *Cancer Discovery* 2.5 (May 2012), pp. 401–404. doi: [10.1158/2159-8290.cd-12-0095](https://doi.org/10.1158/2159-8290.cd-12-0095).
- [45] M. Chalabi et al. "Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers". In: *Nature Medicine* 26.4 (Apr. 2020), pp. 566–576. doi: [10.1038/s41591-020-0805-8](https://doi.org/10.1038/s41591-020-0805-8).
- [46] H. X. Chao et al. "Evidence that the human cell cycle is a series of uncoupled, memoryless phases". In: *Molecular Systems Biology* 15.3 (Mar. 2019). doi: [10.15252/msb.20188604](https://doi.org/10.15252/msb.20188604).
- [47] H. Chen and F. Zhang. "The expected hitting times for finite Markov chains". In: *Linear Algebra and its Applications* 428.11-12 (2008), pp. 2730–2749.
- [48] H. Clevers. "The intestinal crypt, a prototype stem cell compartment". In: *Cell* 154.2 (2013), pp. 274–284. doi: [10.1016/j.cell.2013.07.004](https://doi.org/10.1016/j.cell.2013.07.004).
- [49] J. W. Cole and A. McKalen. "Observations of cell renewal in human rectal mucosa in vivo with thymidine-H<sup>3</sup>". In: *Gastroenterology* 41.2 (1961), pp. 122–125. doi: [10.1016/S0016-5085\(19\)35158-3](https://doi.org/10.1016/S0016-5085(19)35158-3).
- [50] D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, May 2015.
- [51] G. M. Cooper. *The Cell: A Molecular Approach. 8th edition*. Sinauer Associates, Oxford University Press, 2018.
- [52] J. M. Cunningham et al. "Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability". In: *Cancer research* 58.15 (1998), pp. 3455–3460.
- [53] A. de la Chapelle. "Microsatellite Instability". In: *New England Journal of Medicine* 349.3 (July 2003), pp. 209–210. doi: [10.1056/nejmp038099](https://doi.org/10.1056/nejmp038099).
- [54] G. Deng. "BRAF Mutation Is Frequently Present in Sporadic Colorectal Cancer with Methylated hMLH1, But Not in Hereditary Nonpolyposis Colorectal Cancer". In: *Clinical Cancer Research* 10.1 (2004), pp. 191–195. doi: [10.1158/1078-0432.ccr-1118-3](https://doi.org/10.1158/1078-0432.ccr-1118-3).
- [55] R. DESPER et al. "Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data". In: *Journal of Computational Biology* 6.1 (Jan. 1999), pp. 37–51. doi: [10.1089/cmb.1999.6.37](https://doi.org/10.1089/cmb.1999.6.37).



- [56] S. Dihlmann et al. "Dominant negative effect of the APC1309 mutation: a possible explanation for genotype-phenotype correlations in familial adenomatous polyposis". In: *Cancer research* 59.8 (1999), pp. 1857–1860.
- [57] R. Dolcetti et al. "High Prevalence of Activated Intraepithelial Cytotoxic T Lymphocytes and Increased Neoplastic Cell Apoptosis in Colorectal Carcinomas with Microsatellite Instability". In: *The American Journal of Pathology* 154.6 (June 1999), pp. 1805–1813. DOI: [10.1016/s0002-9440\(10\)65436-3](https://doi.org/10.1016/s0002-9440(10)65436-3).
- [58] E. Domingo. "BRAF screening as a low-cost effective strategy for simplifying HNPCC genetic testing". In: *Journal of Medical Genetics* 41.9 (2004), pp. 664–668. DOI: [10.1136/jmg.2004.020651](https://doi.org/10.1136/jmg.2004.020651).
- [59] M. Dominguez-Valentin et al. "Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database". In: *Genetics in Medicine* 22.1 (July 2019), pp. 15–25. DOI: [10.1038/s41436-019-0596-9](https://doi.org/10.1038/s41436-019-0596-9).
- [60] A. Duval et al. "Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers". In: *Human Molecular Genetics* 10.5 (Mar. 2001), pp. 513–518. DOI: [10.1093/hmg/10.5.513](https://doi.org/10.1093/hmg/10.5.513).
- [61] F. Echterdiek et al. "Low density of FOXP3-positive T cells in normal colonic mucosa is related to the presence of beta2-microglobulin mutations in Lynch syndrome-associated colorectal cancer". In: *OncoImmunology* 5.2 (Nov. 2015), e1075692. DOI: [10.1080/2162402x.2015.1075692](https://doi.org/10.1080/2162402x.2015.1075692).
- [62] L. Edler and A. Kopp-Schneider. "Origins of the mutational origin of cancer". In: *International Journal of Epidemiology* 34.5 (July 2005), pp. 1168–1170. DOI: [10.1093/ije/dyi134](https://doi.org/10.1093/ije/dyi134).
- [63] C. Engel et al. "No Difference in Colorectal Cancer Incidence or Stage at Detection by Colonoscopy Among 3 Countries With Different Lynch Syndrome Surveillance Policies". In: *Gastroenterology* 155.5 (Nov. 2018), 1400–1409.e2. DOI: [10.1053/j.gastro.2018.07.030](https://doi.org/10.1053/j.gastro.2018.07.030).
- [64] C. Engel et al. "Associations of Pathogenic Variants in MLH1, MSH2, and MSH6 With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome". In: *Gastroenterology* 158.5 (Apr. 2020), pp. 1326–1333. DOI: [10.1053/j.gastro.2019.12.032](https://doi.org/10.1053/j.gastro.2019.12.032).
- [65] E. R. Fearon. "Molecular Genetics of Colorectal Cancer". In: *Annual Review of Pathology: Mechanisms of Disease* 6.1 (Feb. 2011), pp. 479–507. DOI: [10.1146/annurev-pathol-011110-130235](https://doi.org/10.1146/annurev-pathol-011110-130235).
- [66] A. G. Fletcher, C. J. Breward, and S. J. Chapman. "Mathematical modeling of monoclonal conversion in the colonic crypt". In: *Journal of Theoretical Biology* 300 (May 2012), pp. 118–133. DOI: [10.1016/j.jtbi.2012.01.021](https://doi.org/10.1016/j.jtbi.2012.01.021).
- [67] W. Frankel et al. "Lynch Syndrome: Genetic Tumour Syndromes of the Digestive System". In: *World Health Organization Classification of Tumours of the Digestive System*. 5th ed. IARC Press, 2019.

- [68] B. Galeota-Sprung, B. Guindon, and P. Sniegowski. "The fitness cost of mismatch repair mutators in *Saccharomyces cerevisiae*: partitioning the mutational load". In: *Heredity* 124.1 (Sept. 2019), pp. 50–61. doi: [10.1038/s41437-019-0267-2](https://doi.org/10.1038/s41437-019-0267-2).
- [69] J. Gao et al. "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal". In: *Science Signaling* 6.269 (Apr. 2013). doi: [10.1126/scisignal.2004088](https://doi.org/10.1126/scisignal.2004088).
- [70] J. Gebert et al. "Recurrent Frameshift Neoantigen Vaccine Elicits Protective Immunity With Reduced Tumor Burden and Improved Overall Survival in a Lynch Syndrome Mouse Model". In: *Gastroenterology* 161.4 (Oct. 2021), 1288–1302.e13. doi: [10.1053/j.gastro.2021.06.073](https://doi.org/10.1053/j.gastro.2021.06.073).
- [71] M. Gerstung et al. "Quantifying cancer progression with conjunctive Bayesian networks". In: *Bioinformatics* 25.21 (Aug. 2009), pp. 2809–2815. doi: [10.1093/bioinformatics/btp505](https://doi.org/10.1093/bioinformatics/btp505).
- [72] M. Gerstung et al. "The Temporal Order of Genetic and Pathway Alterations in Tumorigenesis". In: *PLoS ONE* 6.11 (Nov. 2011). Ed. by A. E. Toland, e27136. doi: [10.1371/journal.pone.0027136](https://doi.org/10.1371/journal.pone.0027136).
- [73] M. Gerstung et al. "The Evolutionary History of 2,658 Cancers". en. In: *Nature* 578.7793 (Feb. 2020), pp. 122–128. doi: [10.1038/s41586-019-1907-7](https://doi.org/10.1038/s41586-019-1907-7).
- [74] F. M. Giardiello et al. "Guidelines on Genetic Evaluation and Management of Lynch Syndrome: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer". In: *American Journal of Gastroenterology* 109.8 (Aug. 2014), pp. 1159–1179. doi: [10.1038/ajg.2014.186](https://doi.org/10.1038/ajg.2014.186).
- [75] N. S. Goldstein et al. "Hyperplastic-like Colon Polyps That Preceded Microsatellite-Unstable Adenocarcinomas". en. In: *American Journal of Clinical Pathology* 119.6 (2003), pp. 778–796. doi: [10.1309/DRFQ0WFUF1G13CTK](https://doi.org/10.1309/DRFQ0WFUF1G13CTK).
- [76] F. F. González-Galarza et al. "Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations". In: *Nucleic Acids Research* 43.D1 (Nov. 2014), pp. D784–D788. doi: [10.1093/nar/gku1166](https://doi.org/10.1093/nar/gku1166).
- [77] R. Gryfe. "Inherited colorectal cancer syndromes". In: *Clinics in colon and rectal surgery* 22.4 (2009), p. 198. doi: [10.1055/s-0029-1242459](https://doi.org/10.1055/s-0029-1242459).
- [78] R. Hammack, W. Imrich, and S. Klavžar. *Handbook of Product Graphs*. en. 0th ed. CRC Press, June 2011.
- [79] H. HAMPEL et al. "Cancer Risk in Hereditary Nonpolyposis Colorectal Cancer Syndrome: Later Age of Onset". In: *Gastroenterology* 129.2 (Aug. 2005), pp. 415–421. doi: [10.1016/j.gastro.2005.05.011](https://doi.org/10.1016/j.gastro.2005.05.011).
- [80] H. Hampel et al. "Assessment of Tumor Sequencing as a Replacement for Lynch Syndrome Screening and Current Molecular Tests for Patients With Colorectal Cancer". In: *JAMA Oncology* 4.6 (June 2018), p. 806. doi: [10.1001/jamaoncol.2018.0104](https://doi.org/10.1001/jamaoncol.2018.0104).

- [81] S. Haupt et al. "A computational model for investigating the evolution of colonic crypts during Lynch syndrome carcinogenesis". In: *Computational and Systems Oncology* 1.2 (July 4, 2021), e1020. doi: [10.1002/cso2.1020](https://doi.org/10.1002/cso2.1020).
- [82] S. Haupt et al. "Mathematical Modeling of Multiple Pathways in Colorectal Carcinogenesis Using Dynamical Systems with Kronecker Structure". In: *PLOS Computational Biology* 17.5, e1008970 (May 18, 2021). Ed. by J. Chen, e1008970. doi: [10.1371/journal.pcbi.1008970](https://doi.org/10.1371/journal.pcbi.1008970).
- [83] J. P. Heath. "Epithelial cell migration in the intestine". In: *Cell biology international* 20.2 (1996), pp. 139–146. doi: [10.1006/cbir.1996.0018](https://doi.org/10.1006/cbir.1996.0018).
- [84] N. J. Higham. *Functions of Matrices*. Society for Industrial and Applied Mathematics, Jan. 2008.
- [85] M. Hoffmeister et al. "Statin Use and Survival After Colorectal Cancer: The Importance of Comprehensive Confounder Adjustment". In: *JNCI: Journal of the National Cancer Institute* 107.6 (Mar. 2015). doi: [10.1093/jnci/djv045](https://doi.org/10.1093/jnci/djv045).
- [86] H. Honda, M. Tanemura, and S. Imayama. "Spontaneous Architectural Organization of Mammalian Epidermis from Random Cell Packing". In: *Journal of Investigative Dermatology* 106.2 (Feb. 1996), pp. 312–315. doi: [10.1111/1523-1747.ep12342964](https://doi.org/10.1111/1523-1747.ep12342964).
- [87] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Apr. 1991.
- [88] G. Hounnou and C. Destrieux. "Anatomical study of the length of the human intestine". In: *Surgical and radiologic anatomy* 24.5 (2002), pp. 290–294. doi: [10.1007/s00276-002-0057-y](https://doi.org/10.1007/s00276-002-0057-y).
- [89] D. J. Huels et al. "E-cadherin can limit the transforming properties of activating  $\beta$ -catenin mutations". In: *The EMBO Journal* 34.18 (Aug. 2015), pp. 2321–2333. doi: [10.15252/embj.201591739](https://doi.org/10.15252/embj.201591739).
- [90] Y. Iwasa, F. Michor, and M. A. Nowak. "Stochastic Tunnels in Evolutionary Dynamics". In: *Genetics* 166.3 (Mar. 2004), pp. 1571–1579. doi: [10.1534/genetics.166.3.1571](https://doi.org/10.1534/genetics.166.3.1571).
- [91] J. Janikovits et al. "High numbers of PDCD1 (PD-1)-positive T cells and B2M mutations in microsatellite-unstable colorectal cancer". In: *OncoImmunology* 7.2 (Nov. 2017), e1390640. doi: [10.1080/2162402x.2017.1390640](https://doi.org/10.1080/2162402x.2017.1390640).
- [92] H. J. Järvinen, J.-P. Mecklin, and P. Sistonen. "Screening reduces colorectal cancer rate in families with hereditary nonpolyposis colorectal cancer". In: *Gastroenterology* 108.5 (May 1995), pp. 1405–1411. doi: [10.1016/0016-5085\(95\)90688-6](https://doi.org/10.1016/0016-5085(95)90688-6).
- [93] K. W. Jasperson et al. "Hereditary and Familial Colon Cancer". In: *Gastroenterology* 138.6 (May 2010), pp. 2044–2058. doi: [10.1053/j.gastro.2010.01.054](https://doi.org/10.1053/j.gastro.2010.01.054).
- [94] J. Jiricny. "Postreplicative Mismatch Repair". In: *Cold Spring Harbor Perspectives in Biology* 5.4 (Apr. 2013), a012633–a012633. doi: [10.1101/cshperspect.a012633](https://doi.org/10.1101/cshperspect.a012633).
- [95] M. D. Johnston. "Mathematical modelling of cell population dynamics in the colonic crypt with application to colorectal cancer". PhD thesis. Oxford University, UK, 2008.

- [96] V. Jurtz et al. "NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data". In: *The Journal of Immunology* 199.9 (Oct. 2017), pp. 3360–3368. doi: [10.4049/jimmunol.1700893](https://doi.org/10.4049/jimmunol.1700893).
- [97] A. Kaveh and H. Rahami. "A unified method for eigendecomposition of graph products". In: *Communications in Numerical Methods in Engineering* 21.7 (Mar. 2005), pp. 377–388. doi: [10.1002/cnm.753](https://doi.org/10.1002/cnm.753).
- [98] D. G. Kendall. "Birth-and-Death Processes, and the Theory of Carcinogenesis". In: *Biometrika* 47.1/2 (June 1960), p. 13. doi: [10.2307/2332953](https://doi.org/10.2307/2332953).
- [99] K. W. Kinzler and B. Vogelstein. "Lessons from hereditary colorectal cancer". In: *Cell* 87.2 (1996), pp. 159–170. doi: [10.1016/s0092-8674\(00\)81333-1](https://doi.org/10.1016/s0092-8674(00)81333-1).
- [100] D. Klimstra et al. "Classification of neuroendocrine neoplasms of the digestive system". In: *WHO Classification of tumours, 5th Edition. Digestive system tumours* (2019), pp. 16–19.
- [101] M. Kloor and M. von Knebel Doeberitz. "The Immune Biology of Microsatellite-Unstable Cancer". In: *Trends in Cancer* 2.3 (2016), pp. 121–133. doi: [10.1016/j.trecan.2016.02.004](https://doi.org/10.1016/j.trecan.2016.02.004).
- [102] M. Kloor, M. von Knebel Doeberitz, and J. F. Gebert. "Molecular testing for microsatellite instability and its value in tumor characterization". In: *Expert Review of Molecular Diagnostics* 5.4 (2005), pp. 599–611. doi: [10.1586/14737159.5.4.599](https://doi.org/10.1586/14737159.5.4.599).
- [103] M. Kloor et al. "Beta2-microglobulin mutations in microsatellite unstable colorectal tumors". In: *International Journal of Cancer* 121.2 (2007), pp. 454–458. doi: [10.1002/ijc.22691](https://doi.org/10.1002/ijc.22691).
- [104] M. Kloor et al. "Analysis of EPCAM Protein Expression in Diagnostics of Lynch Syndrome". In: *Journal of Clinical Oncology* 29.2 (Jan. 2011), pp. 223–227. doi: [10.1200/jco.2010.32.0820](https://doi.org/10.1200/jco.2010.32.0820).
- [105] M. Kloor et al. "Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study". In: *The Lancet Oncology* 13.6 (June 2012), pp. 598–606. doi: [10.1016/s1470-2045\(12\)70109-2](https://doi.org/10.1016/s1470-2045(12)70109-2).
- [106] M. Kloor et al. "A Frameshift Peptide Neoantigen-Based Vaccine for Mismatch Repair-Deficient Cancers: A Phase I/IIa Clinical Trial". In: *Clinical Cancer Research* 26.17 (June 2020), pp. 4503–4510. doi: [10.1158/1078-0432.ccr-19-3517](https://doi.org/10.1158/1078-0432.ccr-19-3517).
- [107] A. G. Knudson. "Mutation and Cancer: Statistical Study of Retinoblastoma". In: *Proceedings of the National Academy of Sciences* 68.4 (Apr. 1971), pp. 820–823. doi: [10.1073/pnas.68.4.820](https://doi.org/10.1073/pnas.68.4.820).
- [108] R. Kolodner. "Biochemistry and genetics of eukaryotic mismatch repair." In: *Genes & Development* 10.12 (June 1996), pp. 1433–1442. doi: [10.1101/gad.10.12.1433](https://doi.org/10.1101/gad.10.12.1433).
- [109] N. L. Komarova, A. Sengupta, and M. A. Nowak. "Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability". In: *Journal of Theoretical Biology* 223.4 (Aug. 2003), pp. 433–450. doi: [10.1016/s0022-5193\(03\)00120-6](https://doi.org/10.1016/s0022-5193(03)00120-6).

- [110] N. L. Komarova and D. Wodarz. “The optimal rate of chromosome loss for the inactivation of tumor suppressor genes in cancer”. In: *Proceedings of the National Academy of Sciences* 101.18 (Apr. 2004), pp. 7017–7021. doi: [10.1073/pnas.0401943101](https://doi.org/10.1073/pnas.0401943101).
- [111] N. L. Komarova et al. “Dynamics of Genetic Instability in Sporadic and Familial Colorectal Cancer”. In: *Cancer Biology & Therapy* 1.6 (Nov. 2002), pp. 685–692. doi: [10.4161/cbt.321](https://doi.org/10.4161/cbt.321).
- [112] K. M. Kuntz et al. “A Systematic Comparison of Microsimulation Models of Colorectal Cancer”. In: *Medical Decision Making* 31.4 (June 2011), pp. 530–539. doi: [10.1177/0272989x11408730](https://doi.org/10.1177/0272989x11408730).
- [113] U. Ladabaum. “What Is Lynch-like Syndrome and How Should We Manage It?” In: *Clinical Gastroenterology and Hepatology* 18.2 (Feb. 2020), pp. 294–296. doi: [10.1016/j.cgh.2019.08.009](https://doi.org/10.1016/j.cgh.2019.08.009).
- [114] K. Lagerstedt Robinson et al. “Lynch Syndrome (Hereditary Nonpolyposis Colorectal Cancer) Diagnostics”. In: *JNCI Journal of the National Cancer Institute* 99.4 (2007), pp. 291–299. doi: [10.1093/jnci/djk051](https://doi.org/10.1093/jnci/djk051).
- [115] E. Lakatos et al. “Evolutionary dynamics of neoantigens in growing tumors”. In: *Nature Genetics* 52.10 (Sept. 2020), pp. 1057–1066. doi: [10.1038/s41588-020-0687-1](https://doi.org/10.1038/s41588-020-0687-1).
- [116] I. M. M. van Leeuwen et al. “Crypt dynamics and colorectal cancer: advances in mathematical modelling”. In: *Cell Proliferation* 39.3 (June 2006), pp. 157–181. doi: [10.1111/j.1365-2184.2006.00378.x](https://doi.org/10.1111/j.1365-2184.2006.00378.x).
- [117] I. M. M. van Leeuwen et al. “An integrative computational model for intestinal tissue renewal”. In: *Cell Proliferation* 42.5 (Oct. 2009), pp. 617–636. doi: [10.1111/j.1365-2184.2009.00627.x](https://doi.org/10.1111/j.1365-2184.2009.00627.x).
- [118] I. M. van Leeuwen. “Towards a multiscale model of colorectal cancer”. In: *World Journal of Gastroenterology* 13.9 (2007), p. 1399. doi: [10.3748/wjg.v13.i9.1399](https://doi.org/10.3748/wjg.v13.i9.1399).
- [119] B. Leggett and V. Whitehall. “Role of the Serrated Pathway in Colorectal Cancer Pathogenesis”. In: *Gastroenterology* 138.6 (May 2010), pp. 2088–2100. doi: [10.1053/j.gastro.2009.12.066](https://doi.org/10.1053/j.gastro.2009.12.066).
- [120] M. D. Leiserson et al. “CoMET: a statistical approach to identify combinations of mutually exclusive alterations in cancer”. In: *Genome biology* 16.1 (2015), pp. 1–20. doi: [10.1186/s13059-015-0700-7](https://doi.org/10.1186/s13059-015-0700-7).
- [121] L. Li and H. Clevers. “Coexistence of Quiescent and Active Adult Stem Cells in Mammals”. In: *Science* 327.5965 (Jan. 2010), pp. 542–545. doi: [10.1126/science.1180794](https://doi.org/10.1126/science.1180794).
- [122] M. Linnebacher et al. “Frameshift peptide-derived T-cell epitopes: A source of novel tumor-specific antigens”. In: *International Journal of Cancer* 93.1 (2001), pp. 6–11. doi: [10.1002/ijc.1298](https://doi.org/10.1002/ijc.1298).
- [123] Z. Liu et al. “Modeling and Analysis of a Nonlinear Age-Structured Model for Tumor Cell Populations with Quiescence”. In: *Journal of Nonlinear Science* 28.5 (May 2018), pp. 1763–1791. doi: [10.1007/s00332-018-9463-0](https://doi.org/10.1007/s00332-018-9463-0).

- [124] C. F. Loan. “The ubiquitous Kronecker product”. In: *Journal of Computational and Applied Mathematics* 123.1-2 (Nov. 2000), pp. 85–100. doi: [10.1016/s0377-0427\(00\)00393-9](https://doi.org/10.1016/s0377-0427(00)00393-9).
- [125] M. B. Loughrey et al. “Incorporation of somatic BRAF mutation testing into an algorithm for the investigation of hereditary non-polyposis colorectal cancer”. In: *Familial Cancer* 6.3 (2007), pp. 301–310. doi: [10.1007/s10689-007-9124-1](https://doi.org/10.1007/s10689-007-9124-1).
- [126] J. S. Lowengrub et al. “Nonlinear modelling of cancer: bridging the gap between cells and tumours”. In: *Nonlinearity* 23.1 (Dec. 2009), R1–R91. doi: [10.1088/0951-7715/23/1/r01](https://doi.org/10.1088/0951-7715/23/1/r01).
- [127] H. Lynch et al. “Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications”. In: *Clinical Genetics* 76.1 (July 2009), pp. 1–18. doi: [10.1111/j.1399-0004.2009.01230.x](https://doi.org/10.1111/j.1399-0004.2009.01230.x).
- [128] G. D. Matteis, A. Graudenzi, and M. Antoniotti. “A review of spatial computational models for multi-cellular systems, with regard to intestinal crypts and colorectal cancer development”. In: *Journal of Mathematical Biology* 66.7 (May 2012), pp. 1409–1462. doi: [10.1007/s00285-012-0539-4](https://doi.org/10.1007/s00285-012-0539-4).
- [129] S. A. McDonald et al. “Clonal Expansion in the Human Gut: Mitochondrial DNA Mutations Show Us the Way”. In: *Cell Cycle* 5.8 (Apr. 2006), pp. 808–811. doi: [10.4161/cc.5.8.2641](https://doi.org/10.4161/cc.5.8.2641).
- [130] R. I. McLachlan and G. R. W. Quispel. “Splitting Methods”. In: *Acta Numerica* 11 (Jan. 2002), pp. 341–434. doi: [10.1017/S0962492902000053](https://doi.org/10.1017/S0962492902000053).
- [131] J. Mecklin et al. “Development of Colorectal Tumors in Colonoscopic Surveillance in Lynch Syndrome”. In: *Gastroenterology* 133.4 (Oct. 2007), pp. 1093–1098. doi: [10.1053/j.gastro.2007.08.019](https://doi.org/10.1053/j.gastro.2007.08.019).
- [132] F. A. Meineke, C. S. Potten, and M. Loeffler. “Cell migration and organization in the intestinal crypt using a lattice-free model”. In: *Cell Proliferation* 34.4 (Aug. 2001), pp. 253–266. doi: [10.1046/j.0960-7722.2001.00216.x](https://doi.org/10.1046/j.0960-7722.2001.00216.x).
- [133] A. R. Mensenkamp et al. “Somatic Mutations in MLH1 and MSH2 Are a Frequent Cause of Mismatch-Repair Deficiency in Lynch Syndrome-Like Tumors”. In: *Gastroenterology* 146.3 (Mar. 2014), 643–646.e8. doi: [10.1053/j.gastro.2013.12.002](https://doi.org/10.1053/j.gastro.2013.12.002).
- [134] J. Metzcar et al. “A Review of Cell-Based Computational Modeling in Cancer Biology”. In: *JCO Clinical Cancer Informatics* 3 (2019), pp. 1–13. doi: [10.1200/cci.18.00069](https://doi.org/10.1200/cci.18.00069).
- [135] G. R. Mirams et al. “Chaste: An Open Source C++ Library for Computational Physiology and Biology”. In: *PLoS Computational Biology* 9.3 (Mar. 2013). Ed. by A. Prlic, e1002970. doi: [10.1371/journal.pcbi.1002970](https://doi.org/10.1371/journal.pcbi.1002970).
- [136] T. J. Mitchell et al. “Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal”. In: *Cell* 173.3 (Apr. 2018), 611–623.e17. doi: [10.1016/j.cell.2018.02.020](https://doi.org/10.1016/j.cell.2018.02.020).

- [137] M. Miyaki et al. "Both BRAF and KRAS mutations are rare in colorectal carcinomas from patients with hereditary nonpolyposis colorectal cancer". In: *Cancer Letters* 211.1 (2004), pp. 105–109. doi: [10.1016/j.canlet.2004.01.027](https://doi.org/10.1016/j.canlet.2004.01.027).
- [138] B. Mlecnik et al. "Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability". In: *Immunity* 44.3 (Mar. 2016), pp. 698–711. doi: [10.1016/j.immuni.2016.02.025](https://doi.org/10.1016/j.immuni.2016.02.025).
- [139] A. H. Al-Mohy and N. J. Higham. "Computing the Action of the Matrix Exponential, with an Application to Exponential Integrators". In: *SIAM Journal on Scientific Computing* 33.2 (Jan. 2011), pp. 488–511. doi: [10.1137/100788860](https://doi.org/10.1137/100788860).
- [140] C. Moler and C. Van Loan. "Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later". en. In: *SIAM Review* 45.1 (Jan. 2003), pp. 3–49. doi: [10.1137/S00361445024180](https://doi.org/10.1137/S00361445024180).
- [141] P. Møller. "The Prospective Lynch Syndrome Database reports enable evidence-based personal precision health care". In: *Hereditary Cancer in Clinical Practice* 18.1 (Mar. 2020). doi: [10.1186/s13053-020-0138-0](https://doi.org/10.1186/s13053-020-0138-0).
- [142] P. Møller et al. "Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database". In: *Gut* 66.3 (Dec. 2017), pp. 464–472. doi: [10.1136/gutjnl-2015-309675](https://doi.org/10.1136/gutjnl-2015-309675).
- [143] P. Møller et al. "Colorectal cancer incidences in Lynch syndrome: a comparison of results from the prospective lynch syndrome database and the international mismatch repair consortium". In: *Hereditary Cancer in Clinical Practice* 20.1 (Oct. 2022). doi: [10.1186/s13053-022-00241-1](https://doi.org/10.1186/s13053-022-00241-1).
- [144] L. Moreira et al. "Identification of Lynch Syndrome Among Patients With Colorectal Cancer". In: *JAMA* 308.15 (Oct. 2012), p. 1555. doi: [10.1001/jama.2012.13088](https://doi.org/10.1001/jama.2012.13088).
- [145] D. Morel, R. Marcelpoil, and G. Brugal. "A Proliferation Control Network Model: The Simulation of Two-Dimensional Epithelial Homeostasis". In: *Acta Biotheoretica* 49.4 (2001), pp. 219–234. doi: [10.1023/a:1014201805222](https://doi.org/10.1023/a:1014201805222).
- [146] H. Nagase and Y. Nakamura. "Mutations of the APC (adenomatous polyposis coli) gene". In: *Human mutation* 2.6 (1993), pp. 425–434. doi: [10.1002/humu.1380020602](https://doi.org/10.1002/humu.1380020602).
- [147] K. Nakano et al. "Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area". In: *Human Cell* 30.3 (Mar. 2017), pp. 149–161. doi: [10.1007/s13577-017-0168-8](https://doi.org/10.1007/s13577-017-0168-8).
- [148] K. Naxerova et al. "Origins of lymphatic and distant metastases in human colorectal cancer". In: *Science* 357.6346 (July 2017), pp. 55–60. doi: [10.1126/science.aai8515](https://doi.org/10.1126/science.aai8515).
- [149] K. Newton et al. "Tumour MLH1 promoter region methylation testing is an effective prescreen for Lynch Syndrome (HNPCC)". In: *Journal of Medical Genetics* 51.12 (2014), pp. 789–796. doi: [10.1136/jmedgenet-2014-102552](https://doi.org/10.1136/jmedgenet-2014-102552).

- [150] H. Nguyen and H.-Q. Duong. “The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy (Review)”. In: *Oncology Letters* (May 2018). DOI: [10.3892/ol.2018.8679](https://doi.org/10.3892/ol.2018.8679).
- [151] A. M. Nicholson et al. “Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium”. In: *Cell Stem Cell* 22.6 (June 2018), 909–918.e8. DOI: [10.1016/j.stem.2018.04.020](https://doi.org/10.1016/j.stem.2018.04.020).
- [152] J. Niesen and W. M. Wright. “Algorithm 919: A Krylov Subspace Algorithm for Evaluating the  $\varphi$ -Functions Appearing in Exponential Integrators”. In: *ACM Transactions on Mathematical Software* 38.3 (Apr. 2012), p. 19. DOI: [10.1145/2168773.2168781](https://doi.org/10.1145/2168773.2168781).
- [153] M. A. Nowak et al. “The role of chromosomal instability in tumor initiation”. In: *Proceedings of the National Academy of Sciences* 99.25 (2002), pp. 16226–16231. DOI: [10.1073/pnas.202617399](https://doi.org/10.1073/pnas.202617399).
- [154] P. Nowell and D. Hungerford. “Chromosome studies on normal and leukemic human leukocytes”. In: *Journal of the National Cancer Institute* 25 (July 1960), pp. 85–109.
- [155] N. A. O’Leary et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic Acids Research* 44.D1 (Nov. 2015), pp. D733–D745. DOI: [10.1093/nar/gkv1189](https://doi.org/10.1093/nar/gkv1189).
- [156] J. M. Osborne et al. “Comparing individual-based approaches to modelling the self-organization of multicellular tissues”. In: *PLOS Computational Biology* 13.2 (Feb. 2017). Ed. by Q. Nie, e1005387. DOI: [10.1371/journal.pcbi.1005387](https://doi.org/10.1371/journal.pcbi.1005387).
- [157] M. Ozcan et al. “Complex pattern of immune evasion in MSI colorectal cancer”. In: *OncImmunity* 7.7 (2018), e1445453. DOI: [10.1080/2162402x.2018.1445453](https://doi.org/10.1080/2162402x.2018.1445453).
- [158] R. K. Pai et al. “DNA mismatch repair protein deficient non-neoplastic colonic crypts: a novel indicator of Lynch syndrome”. In: *Modern Pathology* 31.10 (June 2018), pp. 1608–1618. DOI: [10.1038/s41379-018-0079-6](https://doi.org/10.1038/s41379-018-0079-6).
- [159] M. T. Parsons et al. “Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification”. In: *Journal of Medical Genetics* 49.3 (Feb. 2012), pp. 151–157. DOI: [10.1136/jmedgenet-2011-100714](https://doi.org/10.1136/jmedgenet-2011-100714).
- [160] C. Paterson, H. Clevers, and I. Bozic. “Mathematical model of colorectal cancer initiation”. In: *Proceedings of the National Academy of Sciences* 117.34 (Aug. 2020), pp. 20681–20688. DOI: [10.1073/pnas.2003771117](https://doi.org/10.1073/pnas.2003771117).
- [161] U. Paulus, C. S. Potten, and M. Loeffler. “A model of the control of cellular regeneration in the intestinal crypt after perturbation based solely on local stem cell regulation”. In: *Cell Proliferation* 25.6 (Nov. 1992), pp. 559–578. DOI: [10.1111/j.1365-2184.1992.tb01460.x](https://doi.org/10.1111/j.1365-2184.1992.tb01460.x).



- [162] P. L. Pfuderer et al. "High endothelial venules are associated with microsatellite instability, hereditary background and immune evasion in colorectal cancer". In: *British Journal of Cancer* 121.5 (July 2019), pp. 395–404. doi: [10.1038/s41416-019-0514-6](https://doi.org/10.1038/s41416-019-0514-6).
- [163] S. Popat, R. Hubner, and R. Houlston. "Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis". In: *Journal of Clinical Oncology* 23.3 (Jan. 2005), pp. 609–618. doi: [10.1200/jco.2005.01.086](https://doi.org/10.1200/jco.2005.01.086).
- [164] N. Porkka et al. "Sequencing of Lynch syndrome tumors reveals the importance of epigenetic alterations". In: *Oncotarget* 8.64 (Nov. 2017), pp. 108020–108030. doi: [10.18632/oncotarget.22445](https://doi.org/10.18632/oncotarget.22445).
- [165] C. Potten and M. Loeffler. "Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt". In: *Development* 110.4 (Dec. 1990), pp. 1001–1020. doi: [10.1242/dev.110.4.1001](https://doi.org/10.1242/dev.110.4.1001).
- [166] F. Quehenberger. "Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment". In: *Journal of Medical Genetics* 42.6 (June 2005), pp. 491–496. doi: [10.1136/jmg.2004.024299](https://doi.org/10.1136/jmg.2004.024299).
- [167] M. Rashid et al. "Adenoma development in familial adenomatous polyposis and MUTYH-associated polyposis: somatic landscape and driver genes". In: *The Journal of pathology* 238.1 (2016), pp. 98–108. doi: [10.1002/path.4643](https://doi.org/10.1002/path.4643).
- [168] P. A. Reche and E. L. Reinherz. "PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands". In: *Nucleic Acids Research* 33.Web Server (July 2005), W138–W142. doi: [10.1093/nar/gki357](https://doi.org/10.1093/nar/gki357).
- [169] M. Reuschenbach et al. "A multiplex method for the detection of serum antibodies against in silico-predicted tumor antigens". In: *Cancer Immunology, Immunotherapy* 63.12 (Aug. 2014), pp. 1251–1259. doi: [10.1007/s00262-014-1595-y](https://doi.org/10.1007/s00262-014-1595-y).
- [170] Robert Koch-Institut. "Cancer in Germany 2011/2012". en. In: (2016). doi: [10.17886/RKIPUBL-2016-015](https://doi.org/10.17886/RKIPUBL-2016-015).
- [171] J. Robinson et al. "The IPD-IMGT/HLA Database – New developments in reporting HLA variation". In: *Human Immunology* 77.3 (Mar. 2016), pp. 233–237. doi: [10.1016/j.humimm.2016.01.020](https://doi.org/10.1016/j.humimm.2016.01.020).
- [172] R. W. Ruddon. *Cancer biology*. Oxford University Press, 2007.
- [173] T. Sato et al. "Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche". In: *Nature* 459.7244 (Mar. 2009), pp. 262–265. doi: [10.1038/nature07935](https://doi.org/10.1038/nature07935).
- [174] Y. Schwitalle et al. "Immune Response Against Frameshift-Induced Neopeptides in HNPCC Patients and Healthy HNPCC Mutation Carriers". In: *Gastroenterology* 134.4 (Apr. 2008), pp. 988–997. doi: [10.1053/j.gastro.2008.01.015](https://doi.org/10.1053/j.gastro.2008.01.015).
- [175] S. Sekine et al. "Mismatch repair deficiency commonly precedes adenoma formation in Lynch Syndrome-Associated colorectal tumorigenesis". In: *Modern Pathology* 30.8 (May 2017), pp. 1144–1151. doi: [10.1038/modpathol.2017.39](https://doi.org/10.1038/modpathol.2017.39).

- [176] T. T. Seppälä et al. “European guidelines from the EHTG and ESCP for Lynch syndrome: an updated third edition of the Mallorca guidelines based on gene and gender”. In: *British Journal of Surgery* 108.5 (May 2021), pp. 484–498. doi: [10.1002/bjs.11902](https://doi.org/10.1002/bjs.11902).
- [177] T. T. Seppälä et al. “Lack of association between screening interval and cancer stage in Lynch syndrome may be accounted for by over-diagnosis: a prospective Lynch syndrome database report”. In: *Hereditary Cancer in Clinical Practice* 17.1 (Feb. 2019). doi: [10.1186/s13053-019-0106-8](https://doi.org/10.1186/s13053-019-0106-8).
- [178] G. Serio. “Two-stage stochastic model for carcinogenesis with time-dependent parameters”. In: *Statistics & Probability Letters* 2.2 (Mar. 1984), pp. 95–103. doi: [10.1016/0167-7152\(84\)90057-9](https://doi.org/10.1016/0167-7152(84)90057-9).
- [179] L. Shahriyari, N. L. Komarova, and A. Jilkine. “The role of cell location and spatial gradients in the evolutionary dynamics of colon and intestinal crypts”. In: *Biology Direct* 11.1 (Aug. 2016). doi: [10.1186/s13062-016-0141-6](https://doi.org/10.1186/s13062-016-0141-6).
- [180] I.-M. Shih et al. “Top-down morphogenesis of colorectal tumors”. In: *Proceedings of the National Academy of Sciences* 98.5 (Feb. 2001), pp. 2640–2645. doi: [10.1073/pnas.051629398](https://doi.org/10.1073/pnas.051629398).
- [181] J. Sidney et al. “HLA class I supertypes: a revised and updated classification”. In: *BMC Immunology* 9.1 (Jan. 2008). doi: [10.1186/1471-2172-9-1](https://doi.org/10.1186/1471-2172-9-1).
- [182] L. Staffa et al. “Mismatch Repair-Deficient Crypt Foci in Lynch Syndrome – Molecular Alterations and Association with Clinical Parameters”. In: *PLOS ONE* 10.3 (Mar. 2015). Ed. by J. S. Castresana, e0121980. doi: [10.1371/journal.pone.0121980](https://doi.org/10.1371/journal.pone.0121980).
- [183] R. Sun et al. “Between-region genetic divergence reflects the mode and tempo of tumor evolution”. In: *Nature Genetics* 49.7 (June 2017), pp. 1015–1024. doi: [10.1038/ng.3891](https://doi.org/10.1038/ng.3891).
- [184] H. Sung et al. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries”. In: *CA: A Cancer Journal for Clinicians* 71.3 (Feb. 2021), pp. 209–249. doi: [10.3322/caac.21660](https://doi.org/10.3322/caac.21660).
- [185] W.-Y. Tan and C. Brown. “A nonhomogeneous two-stage model of carcinogenesis”. In: *Mathematical and Computer Modelling* 11 (1988), pp. 445–448. doi: [10.1016/0895-7177\(88\)90531-6](https://doi.org/10.1016/0895-7177(88)90531-6).
- [186] W.-Y. Tan and L. Hanin. *Handbook of Cancer Models with Applications*. WORLD SCIENTIFIC, June 2008.
- [187] G. Teschl. *Ordinary Differential Equations and Dynamical Systems*. Providence, R.I: American Mathematical Society, Sept. 2012.
- [188] The ENCODE Project Consortium. “An Integrated Encyclopedia of DNA Elements in the Human Genome”. en. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. doi: [10.1038/nature11247](https://doi.org/10.1038/nature11247).

- [189] E. Thiis-Evensen et al. "The effect of attending a flexible sigmoidoscopic screening program on the prevalence of colorectal adenomas at 13-year follow-up". In: *The American Journal of Gastroenterology* 96.6 (June 2001), pp. 1901–1907. doi: [10.1111/j.1572-0241.2001.03891.x](https://doi.org/10.1111/j.1572-0241.2001.03891.x).
- [190] B. A. Thompson et al. "A Multifactorial Likelihood Model for MMR Gene Variant Classification Incorporating Probabilities Based on Sequence Bioinformatics and Tumor Characteristics: A Report from the Colon Cancer Family Registry". In: *Human Mutation* 34.1 (2012), pp. 200–209. doi: [10.1002/humu.22213](https://doi.org/10.1002/humu.22213).
- [191] C. Tomasetti et al. "Only three driver gene mutations are required for the development of lung and colorectal cancers". In: *Proceedings of the National Academy of Sciences* 112.1 (Dec. 2014), pp. 118–123. doi: [10.1073/pnas.1421839112](https://doi.org/10.1073/pnas.1421839112).
- [192] T. J. Treangen and S. L. Salzberg. "Repetitive DNA and next-generation sequencing: computational challenges and solutions". In: *Nature Reviews Genetics* 13.1 (Nov. 2011), pp. 36–46. doi: [10.1038/nrg3117](https://doi.org/10.1038/nrg3117).
- [193] S. Turajlic, N. McGranahan, and C. Swanton. "Inferring mutational timing and reconstructing tumour evolutionary histories". In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1855.2 (Apr. 2015), pp. 264–275. doi: [10.1016/j.bbcan.2015.03.005](https://doi.org/10.1016/j.bbcan.2015.03.005).
- [194] A. Umar et al. "Revised Bethesda Guidelines for Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndrome) and Microsatellite Instability". In: *JNCI Journal of the National Cancer Institute* 96.4 (Feb. 2004), pp. 261–268. doi: [10.1093/jnci/djh034](https://doi.org/10.1093/jnci/djh034).
- [195] A. Vanderwalde et al. "Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11, 348 patients". In: *Cancer Medicine* 7.3 (Feb. 2018), pp. 746–756. doi: [10.1002/cam4.1372](https://doi.org/10.1002/cam4.1372).
- [196] H. F. A. Vasen et al. "Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer)". In: *Journal of Medical Genetics* 44.6 (Feb. 2007), pp. 353–362. doi: [10.1136/jmg.2007.048991](https://doi.org/10.1136/jmg.2007.048991).
- [197] H. Vasen, F. Nagengast, and P. M. Khan. "Interval cancers in hereditary non-polyposis colorectal cancer (Lynch syndrome)". In: *The Lancet* 345.8958 (May 1995), pp. 1183–1184. doi: [10.1016/s0140-6736\(95\)91016-6](https://doi.org/10.1016/s0140-6736(95)91016-6).
- [198] H. F. A. Vasen et al. "Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts". In: *Gut* 62.6 (Feb. 2013), pp. 812–823. doi: [10.1136/gut.jn1-2012-304356](https://doi.org/10.1136/gut.jn1-2012-304356).
- [199] M. L. Veigl et al. "Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers". In: *Proceedings of the National Academy of Sciences* 95.15 (July 1998), pp. 8698–8702. doi: [10.1073/pnas.95.15.8698](https://doi.org/10.1073/pnas.95.15.8698).
- [200] J. M. Versluis, G. V. Long, and C. U. Blank. "Learning from clinical trials of neoadjuvant checkpoint blockade". In: *Nature Medicine* 26.4 (Apr. 2020), pp. 475–484. doi: [10.1038/s41591-020-0829-0](https://doi.org/10.1038/s41591-020-0829-0).

- [201] B. Vogelstein and K. W. Kinzler. "The multistep nature of cancer". In: *Trends in Genetics* 9.4 (Apr. 1993), pp. 138–141. doi: [10.1016/0168-9525\(93\)90209-z](https://doi.org/10.1016/0168-9525(93)90209-z).
- [202] B. Vogelstein et al. "Genetic Alterations during Colorectal-Tumor Development". In: *New England Journal of Medicine* 319.9 (Sept. 1988), pp. 525–532. doi: [10.1056/nejm198809013190901](https://doi.org/10.1056/nejm198809013190901).
- [203] M. D. Walsh et al. "Immunohistochemical testing of conventional adenomas for loss of expression of mismatch repair proteins in Lynch syndrome mutation carriers: a case series from the Australasian site of the colon cancer family registry". In: *Modern Pathology* 25.5 (2012), pp. 722–730. doi: [10.1038/modpathol.2011.209](https://doi.org/10.1038/modpathol.2011.209).
- [204] J. D. Watson and F. H. C. Crick. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356 (Apr. 1953), pp. 737–738. doi: [10.1038/171737a0](https://doi.org/10.1038/171737a0).
- [205] R. A. Weinberg. *The biology of cancer*. Garland science, 2013.
- [206] B. Werner et al. "Measuring single cell divisions in human cancers from multi-region sequencing data". In: (2019). doi: [10.1101/560243](https://doi.org/10.1101/560243).
- [207] M. J. Williams et al. "Identification of neutral tumor evolution across cancer types". In: *Nature Genetics* 48.3 (Jan. 2016), pp. 238–244. doi: [10.1038/ng.3489](https://doi.org/10.1038/ng.3489).
- [208] K. Wimmer et al. "Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium Care for CMMRD (C4CMMRD)". In: *Journal of Medical Genetics* 51.6 (Apr. 2014), pp. 355–365. doi: [10.1136/jmedgenet-2014-102284](https://doi.org/10.1136/jmedgenet-2014-102284).
- [209] A. K. Win et al. "Variation in the risk of colorectal cancer in families with Lynch syndrome: a retrospective cohort study". In: *The Lancet Oncology* 22.7 (July 2021), pp. 1014–1022. doi: [10.1016/s1470-2045\(21\)00189-3](https://doi.org/10.1016/s1470-2045(21)00189-3).
- [210] J. Witt et al. "A simple approach for detecting *HLA-A\*02* alleles in archival formalin-fixed paraffin-embedded tissue samples and an application example for studying cancer immunoediting". In: *HLA* (Oct. 2022). doi: [10.1111/tan.14846](https://doi.org/10.1111/tan.14846).
- [211] D. Wodarz and N. L. Komarova. *Dynamics of Cancer*. WORLD SCIENTIFIC, June 2014.
- [212] S. M. Woerner et al. "Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes". In: *Oncogene* 22.15 (Apr. 2003), pp. 2226–2235. doi: [10.1038/sj.onc.1206421](https://doi.org/10.1038/sj.onc.1206421).
- [213] S. M. Woerner et al. "Systematic identification of genes with coding microsatellites mutated in DNA mismatch repair-deficient cancer cells". In: *International Journal of Cancer* 93.1 (2001), pp. 12–19. doi: [10.1002/ijc.1299](https://doi.org/10.1002/ijc.1299).
- [214] S. M. Woerner et al. "SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology". In: *Nucleic Acids Research* 38.suppl\_1 (Oct. 2009), pp. D682–D689. doi: [10.1093/nar/gkp839](https://doi.org/10.1093/nar/gkp839).

- [215] V. Wunderlich. "Early references to the mutational origin of cancer". In: *International Journal of Epidemiology* 36.1 (Dec. 2006), pp. 246–247. doi: [10.1093/ije/dyl272](https://doi.org/10.1093/ije/dyl272).



