

ABSTRACT

ASLAM, JAI K. Categorifying the Chromatic Polynomial of a Hypergraph and TDA in Cancer Genomics. (Under the direction of Radmila Sazdanović).

In Chapter 1 we generalize the categorifications of the chromatic polynomials of graphs in [60] and [48] to the chromatic polynomials of hypergraphs. We then show that the simplicial chromatic polynomial of a simplicial complex S , introduced in [35], is the chromatic polynomial of any hypergraph \mathcal{G} with independence complex S . Moreover, the simplicial chromatic state sum homology of S [37] is isomorphic to the hypergraph chromatic homology of such a \mathcal{G} . Due to this relationship between the chromatic polynomials of hypergraphs and simplicial chromatic polynomials, we show that there exists a monomial ideal K with Hilbert polynomial equal to the simplicial chromatic polynomial of S , leveraging results in [20, 65, 108].

In Chapter 2 we consider copy number changes, which are known to play an important role in the development of cancer. They are commonly associated with changes in gene expression. Persistence curves, such as Betti curves, have been used to detect copy number changes; however, it is known these curves are unstable with respect to small perturbations in the data. We consider the stability of lifespan and Betti curves by providing bounds on the distance between persistence curves of Vietoris–Rips filtrations built on data and slightly perturbed data in terms of the bottleneck distance. Next, we perform simulations to compare the predictive ability of Betti curves, lifespan curves (conditionally stable) and persistent landscapes (stable) to detect copy number aberrations. We use these methods to identify significant chromosome regions associated with the four major molecular subtypes of breast cancer: Luminal A, Luminal B, Basal and HER2 positive. Identified segments are then used as predictor variables to build machine learning models which classify patients as one of the four subtypes. We find that no single persistence curve outperforms the others and instead suggest a complementary approach using a suite of persistence curves. In this study, we identified new cytobands associated with three of the subtypes: 1q21.1-q25.2, 2p23.2-p16.3, 23q26.2-q28 with the Basal subtype, 8p22-p11.1 with Luminal B and 2q12.1-q21.1 and 5p14.3-p12 with Luminal A. These segments are validated by the TCGA BRCA cohort dataset except for those found for Luminal A. Lastly, we present preliminary results that narrow significant genomic regions to specific genes.

© Copyright 2023 by Jai K. Aslam

All Rights Reserved

Categorifying the Chromatic Polynomial of a Hypergraph and TDA in Cancer Genomics

by
Jai K. Aslam

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Mathematics

Raleigh, North Carolina
2023

APPROVED BY:

Seth Sullivan

Nathan Reading

Tye Lidman

F. Javier Arsuaga
External Member

Radmila Sazdanović
Chair of Advisory Committee

DEDICATION

To Bud and Jo.

BIOGRAPHY

The author was born in Maine to loving parents Dru and Padiath. He developed a love for math early in life and was supported by many top-notch educators. During his freshman year of high school he was introduced to programming and this has been an integral part of his problem-solving process ever since. For undergraduate studies he attended Northeastern University where he studied math and computer science. After a research project with Ivan Martino and an REU with Florian Frick, he decided he wanted to conduct more math research and attended North Carolina State to do so. In his free time the author likes to play tennis, cook and play ultimate.

ACKNOWLEDGEMENTS

I am grateful to be surrounded by wonderful people. If I failed to mention you here it was not on purpose, I am thankful for you too.

From preschool to graduate school, I've been lucky to have excellent teachers every step of the way. This thesis is as much yours as mine – I wouldn't be here without you. I would like to specifically thank Mary-Lou, Mr. Barricelli, Ms. Balbo, Ms. Frechette, Mrs. Baker, Ms. Luce, Mrs. King and Mr. Brown who all went above and beyond for me. From undergrad I am particularly grateful to Dr. Martino, Professor Frick, Professor Iarrobino and Professor Suci.

More recently, I'd like to thank Tye Lidman and Ricky Liu for providing me with a strong foundation in abstract algebra and combinatorics respectively. Their qualifying exam sequences were challenging and I grew as a mathematician because of them. I'd also like to thank Nathan Reading for teaching excellent special topics classes about Coxeter Groups and Cluster Algebras from which I learned a lot.

Thank you to my committee members: Tye Lidman, Javier Arsuaga, Nathan Reading, Seth Sullivant, Radmila Sazdanović and my graduate student representative Don Sheehy for agreeing to be on my committee, asking great questions and providing comments on my thesis.

To my advisor Radmila Sazdanović, thank you for introducing me to the beautiful field of categorification and for countless Zoom and in person meetings. I'd also like to thank Javier Arsuaga, Sergio Ardanza-Trevijano and Jingwei Xiong for many discussions about TDA and cancer genomics, as well as for being excellent collaborators on the work that appears in the second chapter.

I am so grateful to my many friends, old and new, who have helped keep me happy during graduate school.

Innumerable thanks go out to my late grandparents Bud and Jo who were the best role models I could ever hope to have. I am forever grateful to my parents, Dru and Padiath, who have provided me with perpetual support and so many opportunities. Thank you to my siblings Sharif, Sunny, Ali, Rebecca and Jess for being there for me and for more games of spades than they probably cared to play.

Last but not least, thank you to my partner Aneri for enduring me and providing me with an immense amount of support over the past five years.

TABLE OF CONTENTS

List of Tables	vii
List of Figures	x
Chapter 1 Categorifying the Chromatic Polynomial of a Hypergraph	1
1.1 Background	3
1.1.1 Preliminaries and Definitions	3
1.1.2 Chromatic Graph Homology	7
1.1.3 Graph Configuration Spaces	11
1.1.4 Simplicial Chromatic Polynomial	12
1.2 Chromatic Hypergraph Homology	15
1.2.1 State Sum Formulation	15
1.2.2 Properties of Chromatic Hypergraph Homology	19
1.3 Hypergraph Configuration Spaces	25
1.4 Comparison with the Simplicial Chromatic Constructions	28
1.5 The Simplicial Chromatic Coloring Complex	34
1.6 Future Directions	39
Chapter 2 Topological Data Analysis Applied to Cancer Genomics	41
2.1 Introduction	41
2.2 Topological Data Analysis	44
2.3 Methods	56
2.3.1 The TAaCGH Method	56
2.3.2 Cancer Subtype Predictive Models	57
2.4 Data	59
2.4.1 Horlings Dataset	59
2.4.2 TCGA BRCA Cohort Data	59
2.4.3 UCSF 500 Gene Panel	60
2.4.4 Simulation Data	60
2.5 Bounds on the Distance between Persistence Curves	61
2.6 Results	65
2.6.1 Comparison of Performance of Different Persistence Curves on Simulated Data	65
2.6.2 Comparison of Topological Summaries within the TAaCGH Framework on Horlings Data	71
2.6.3 Genomic Regions Associated to Cancer Subtypes	73
2.6.4 Cancer Subtype Classification Models	79
2.6.5 1-Dimensional TAaCGH	82
2.6.6 Narrowing Genomic Regions to Genes	86

2.7 Discussion	90
References	94
APPENDIX	104
Appendix A Chapter 2 Supplementary Figures and Tables	105

LIST OF TABLES

Table 1.1	The chromatic graph homology of θ_2 with \mathcal{A}_2	24
Table 2.1	Average sensitivity and specificity of lifespan curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.	71
Table 2.2	Average sensitivity and specificity of second landscape curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.	71
Table 2.3	HER2 Phenotype: Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the HER2 subtype on the Horlings dataset [66] using the <i>R</i> TDA package.	72
Table 2.4	HER2 Logistic Regression: The confusion matrices and accuracy of logistic regression models built to predict the HER2 phenotype from Betti, lifespan and landscape curves.	72
Table 2.5	Luminal A Phenotype: Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the Luminal A subtype on the Horlings dataset [66] using the <i>R</i> TDA package.	72
Table 2.6	Luminal A Logistic Regression: The confusion matrices and accuracy of the logistic regression models built to predict the Luminal A phenotype from lifespan curves and third landscape curves.	73
Table 2.7	Basal phenotype: Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the Basal subtype on the Horlings dataset [66] using the <i>R</i> TDA package.	74
Table 2.8	Basal Logistic Regression: The confusion matrices and accuracy of logistic regression models built to predict the Basal phenotype from Betti, lifespan and landscape curves.	82
Table 2.9	Genomic regions detected by betti curves in the Horlings data set using the TAaCGH method with 1-dimensional persistent homology. Too many regions are detected and there is also overlap between subtypes.	84
Table 2.10	Genomic regions detected in the Horlings data set by lifespan curves using the TAaCGH method with 1-dimensional persistent homology.	85
Table 2.11	Genomic regions detected in the Horlings data set by first landscape curves using the TAaCGH method with 1-dimensional persistent homology.	86
Table 2.12	Genes that were detected as significant for the basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set and are also UCSF 500 genes.	88

Table 2.13	The general signatures of oncogenes and tumor suppressor genes that are driven by copy number. Oncogenes have high copy number and high gene expression and tumor suppressor genes have low copy number and low gene expression.	88
Table 2.14	Candidate oncogenes that were detected as significant for the HER2 subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.	89
Table 2.15	Candidate oncogenes that were detected as significant for the Luminal B subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.	90
Table 2.16	The cytobands corresponding to the new regions detected by Betti curves, lifespan curves, and landscape curves compared to the regions detected in [9]. Red indicates new cytobands associated with the Basal subtype, blue with Luminal A and gray Luminal B.	91
Table A.1	Comparison of the chromosome segments detected in the original Betti-0 study [9] using JavaPlex and the new Betti-0 study using the R TDA package. The rows contain the segments that the old study detected but the new study did not for each subtype of breast cancer.	106
Table A.2	Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the Luminal B subtype on the Horlings dataset [66] using the R TDA package. Basals were not included in the control set.	107
Table A.3	Average sensitivity and specificity of Betti curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.	107
Table A.4	Average sensitivity and specificity of lifespan curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.	113
Table A.5	Average sensitivity and specificity of lifespan curves for patient classification. The length of the aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.	113
Table A.6	Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.	114
Table A.7	Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.	115

Table A.8	Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.	116
Table A.9	Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.	117
Table A.10	Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.	118

LIST OF FIGURES

Figure 1.1	A proper coloring of K_3 with 3 colors. No two adjacent vertices are assigned the same color.	4
Figure 1.2	The chromatic graph chain groups of the triangle graph K_3	9
Figure 1.3	The chromatic hypergraph chain groups of a hypergraph $\mathcal{G} = (V = [4], \{\{1, 2, 3\}, \{1, 2, 4\}, \{2, 3, 4\}\})$. The chromatic polynomial of \mathcal{G} is $\chi(\mathcal{G}, t) = t^4 - 3t^2 + 2t$	16
Figure 1.4	θ_n is an n uniform hypergraph with n vertices and 1 edge.	23
Figure 1.5	Hypergraph configuration spaces for \mathcal{G} , $\mathcal{G} \setminus e$ and \mathcal{G}/e where \mathcal{G} is the hypergraph on three vertices with hyperedge 123 and $M = [0, 1]$	31
Figure 1.6	Chromatically equivalent hypergraphs with chromatic polynomial $t^6 - t^4 - t^3 + t$	32
Figure 1.7	The 1-skeleta of the simplicial complexes with minimal nonfaces $\{\{1, 2, 4, 5\}, \{2, 3, 6\}\}$ (A) and minimal nonfaces $\{\{1, 2, 4, 5\}, \{2, 3, 5\}, \{1, 2, 3, 4, 6\}\}$ (B).	33
Figure 1.8	\mathcal{G}_1 and \mathcal{G}_2 are cochromatic hypergraphs with chromatic polynomials equal to $t^4 - 3t^3 + 2t^2$. These are the hypergraphs of minimal nonfaces corresponding to the simplicial complexes in Figure 1 of [35].	34
Figure 1.9	A graph L_3 and its associated simplicial chromatic coloring complex.	35
Figure 1.10	Bijection between S -compatible colorings with 2 colors and degree 1 monomials in K_{L_3}	37
Figure 1.11	Graphs G_1 and G_2 such that $\chi_c(G_1, t) = \chi_c(G_2, t)$	37
Figure 2.1	The TAaCGH pipeline. This workflow determines if a segment of the genome is statistically significant for a cancer subtype. Once a particular segment of study is chosen, the TAaCGH pipeline begins. In Step 1 the copy number variation data is separated into control and test patients. Copy number variation data from a single patient is pictured. Next, in Step 2, the data is converted into a sliding window point cloud for each patient. The sliding window point cloud from the sample patient's data is pictured. In Step 3 a Vietoris–Rips filtration is built on each patient's point cloud, the persistent homology of each patient's Vietoris–Rips filtration is computed, recorded in a persistence diagram and summarized into a persistence curve. As an illustration we are using a lifespan curve from the sample patient. Examples of all persistence curves are shown in Figure 2.2c–e. Lastly, in Step 4, the persistence curves of all patients in the test group and in the control group are averaged and a permutation test is run on the L^2 norm of these averaged persistence curves.	45

Figure 2.2	Sample patient data and output. The copy number data (a) , sliding window point cloud (b) , persistent landscape (c) , Betti curve (d) and lifespan curve (e) of a patient from the Horlings dataset [66] on chromosome 17q segment 3 using 0-dimensional persistence. 17qs corresponds to the cytoband range 17q21.2-q21.33.	46
Figure 2.3	The sliding window point cloud of the function $f(x) = \sin(x)$ at increments of $\frac{\pi}{4}$ from 0 to 2π	48
Figure 2.4	The Vietoris Rips complex on points sampled from a circle.	49
Figure 2.5	Vietoris-Rips Complex and Čech complex on the same point cloud.	49
Figure 2.6	A point cloud formed by the corners of a 6×4 rectangle.	52
Figure 2.7	The Vietoris-Rips filtration on the rectangle point cloud from Figure 2.6.	53
Figure 2.8	The 0-dimensional persistence diagram for the Vietoris-Rips complex in Figure 2.7.	53
Figure 2.9	The 1-dimensional persistence diagram of the Vietoris-Rips complex in Figure 2.7	54
Figure 2.10	Simulation design. Simulations are designed to test the ability of TAaCGH with various persistence curves to distinguish a set of patients with a single contiguous aberration from a set of patients without them. Each patient has 20 probes. Test patients with aberrations of length λ have copy number values sampled from a normal distribution with mean μ_t and standard deviation σ . The remaining copy number values for test patients are sampled from a normal distribution with mean $\mu_c = 0$ and standard deviation σ . Control patients have all copy number values sampled from a normal distribution with mean $\mu_c = 0$ and standard deviation σ . The <i>MIX</i> parameter controls the percentage of patients in the test set that have aberrations, the remaining patients in the test set have data drawn from a normal distribution with mean $\mu_c = 0$ and standard deviation σ . For each set of parameters, we ran 50 simulations. Each simulation consisted of a total of 120 patients, 60 in the test set and 60 in the control set. The parameters varied over the following values $\mu_t \in \{-1, 0.6, 1\}$, $\sigma \in \{0.2, 0.22, \dots, 0.5\}$, $\lambda \in \{1, 3, 5, 10, 15\}$ and $MIX \in \{20\%, 40\%, 60\%, 80\%, 100\%\}$	66
Figure 2.11	Simulation data and associated persistence curves. Simulated CNA data (a) , corresponding sliding window point cloud (b) , Betti curve (c) , lifespan curve (d) and landscape curves (e) for a hypothetical cancer patient on a segment of 20 probes with a single length $\lambda = 5$ contiguous aberration with aberration mean $\mu = 1$ and standard deviation $\sigma = 0.2$. The nonaberrant probes are sampled from a distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.2$	67

- Figure 2.12 Comparison of sensitivity values of different approaches in TAaCGH with respect to varying MIX parameter on persistence of simulated data in dimension zero. Figure shows comparisons between lifespan ℓ_0 curve and Betti β_0 (**a**), second landscape λ_2 (**b**), third landscape λ_3 (**c**), fourth landscape λ_4 curves (**d**). For a fixed value of the MIX parameter, we compute the sensitivity of our method with one of the persistence curves for each set of simulations as the standard deviation σ , the mean μ and the length λ vary over all possible values detailed in Section 2.4.4. The height of each bar represents the percentage of those simulations where the sensitivity of the lifespan curve was bigger (purple), equal to (pink) and less than the sensitivity (blue) of other persistence curves. 68
- Figure 2.13 Comparison of sensitivity values of different approaches in TAaCGH with respect to varying the standard deviation σ on persistence of simulated data in dimension zero. The figure shows comparisons between lifespan ℓ_0 curve and Betti β_0 (**a**), second landscape λ_2 (**b**), fourth landscape λ_4 curves (**c**). For a fixed value of the standard deviation σ , we compute the sensitivity of our method with one of the persistence curves for each set of simulations as the mean μ , length λ and MIX parameters vary over all possible values detailed in Section 2.4.4. The height of each bar represents the percentage of those simulations where the sensitivity of the lifespan curve was bigger (purple), equal to (pink) and less than the sensitivity (blue) of the other persistence curve. Note that the third landscape λ_3 behaves similarly to the second landscape. 69
- Figure 2.14 Comparison of sensitivities of different approaches in TAaCGH with respect to varying the length λ on persistence of simulated data in dimension zero. The figure illustrates differences between lifespan ℓ_0 curve and Betti β_0 (**a**), second landscape λ_2 (**b**), third landscape λ_3 (**c**), fourth landscape λ_4 curves (**d**). For a fixed value of the MIX parameter, we compute the sensitivity of our method with one of the persistence curves for each set of simulations as the standard deviation σ , the mean μ and the mix parameter varies over all possible values detailed in Section 2.4.4. The height of each bar represents the percentage of those simulations where the sensitivity of the lifespan curve was bigger (purple), equal to (pink) and less than the sensitivity (blue) of other persistence curves. 70

Figure 2.15	HER2 phenotype most aberrant cytobands in TCGA BRCA cohort Data Cytobands with gains and losses in the TCGA BRCA cohort dataset [21] for the HER2 phenotype. The chromosome arms 1, 8, 17 and 20 are included since they had above 10% of patients with aberrations in genes in those cytobands on average. The colors indicate which persistence curves detected those cytobands as significant in [66]. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of HER2 patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with a score -2	74
Figure 2.16	Luminal A Patient Profiles. Four Luminal A patient profiles from the Horlings dataset on cytobands 2q12-2q21.1. All share a significant copy number gain between 125 and 135 Mbp.	76
Figure 2.17	Luminal B most aberrant cytobands in TCGA BRCA cohort Data. The chromosome arms 1, 8, 11, 17 and 20 are included above 10% of patients had aberrations in the genes within this cytoband on average. The colors indicate which persistence curves detected those cytobands as significant in [66] for the Luminal B phenotype with no Basals in the control group. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2	77
Figure 2.18	Luminal B phenotype cytobands 8p22-8p11.21. Top graph: Shows the percentage of patients in the Luminal B phenotype with either a 1 or 2 score from the TCGA data (orange) as well as the percentage of all other phenotypes with these scores (gray). Bottom graph: Shows the same as the top graph but for scores of -1 , -2	77

Figure 2.19	Basal phenotype cytobands in TCGA BRCA cohort dataset. The colors indicate how many persistence curve methods detected that particular cytoband in the Horlings dataset [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Basal patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with a score -2	78
Figure 2.20	Basal phenotype most aberrant cytobands in TCGA BRCA cohort dataset. The chromosome arms 1, 3, 5, 8, 10, and 12 are included since above 10% of patients had aberrations in the genes in this cytoband. The colors indicate which persistence curves detected those cytobands as significant in [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Basal patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with a score -2	79
Figure 2.21	Basal phenotype cytobands 2p16.3-2p23.2. Shows the percentage of patients in the Basal phenotype with either a 1 or 2 score from the TCGA BRCA cohort dataset (orange) as well as the percentage of all other phenotypes with these scores (gray).	80
Figure 2.22	β_0 and β_1 curves for Basal on chromosome section 1p36.32 – p36.11.	85
Figure A.1	HER2 vs. other phenotypes combined on TCGA BRCA cohort Data. Cytobands with gains and losses in the TCGA BRCA cohort dataset [21] for the HER2 phenotype (red) as well as all phenotypes combined (black). Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort data set indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of HER2 patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2	106

Figure A.2	Luminal A Phenotypes: The plots illustrate in which cytobands there are gains and losses in the TCGA BRCA cohort dataset [21] for the Luminal A phenotype as well as all phenotypes combined. Luminal A is in red and all groups combined are in black. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .	108
Figure A.3	Luminal B Phenotype: The plots illustrate in which cytobands there are gains and losses in the TCGA BRCA cohort dataset [21] for the Luminal B phenotype as well as all phenotypes combined. Luminal B is in green and all groups combined are in black. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .	108
Figure A.4	Basal Phenotype: The plots illustrate in which cytobands there are gains and losses in the TCGA BRCA cohort dataset [21] for the Basal phenotype as well as all phenotypes combined. Basal is in orange and all groups combined are in black. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Basal patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .	109
Figure A.5	Luminal A phenotype cytobands in TCGA BRCA cohort dataset. The colors indicate how many persistence curve methods detected that particular cytoband in the Horlings dataset [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal A patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .	109

Figure A.6	Luminal A phenotype most aberrant cytobands in TCGA BRCA cohort dataset. The chromosome arms 1, 8, 11, 16, 17 and 20 are included above 10% of patients had aberrations in the genes in these cytobands. The colors indicate which persistence curves detected those cytobands as significant in [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal A patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2	110
Figure A.7	Luminal B phenotype cytobands in TCGA BRCA cohort dataset. The colors indicate how many persistence curve methods detected that particular cytoband in the Horlings dataset [66] with no Basals in the control group. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2	110
Figure A.8	Luminal B phenotype most significant cytobands in TCGA BRCA cohort dataset. The chromosome arms 1, 8, 11, 17 and 20 are included since above 10% of patients had aberrations in the genes in these cytobands. The colors indicate which persistence curves detected those cytobands as significant in [66] for Luminal B with no Basals in the control group. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2	111
Figure A.9	Basal phenotype cytobands $1q21.1 - 1q25.2$. Shows the percentage of patients in the Basal phenotype with either a 1 or 2 score from the TCGA BRCA cohort dataset (orange) as well as the percentage of all other phenotypes with these scores (gray).	111
Figure A.10	Basal phenotype cytobands $23q26.2 - 23q28$. Shows the percentage of patients in the Basal phenotype with either a 1 or 2 score from the TCGA BRCA cohort dataset (orange) as well as the percentage of all other phenotypes with these scores (gray).	112

Figure A.11 Luminal B patient profiles. Four Luminal B patient profiles from the Horlings dataset on cytobands 8p22-8p11.1. All share a significant copy number gain between 35 and 45 Mbp and a loss from 15–35 Mbp.¹¹²

CHAPTER

1

CATEGORIFYING THE CHROMATIC POLYNOMIAL OF A HYPERGRAPH

Categorification is a way to promote an algebraic object to one with more structure [74]. A basic example is the categorification of the natural numbers by finite-dimensional vector spaces. Natural numbers lift to dimensions of vector spaces and addition and multiplication are lifted by direct sum and tensor product since $\dim(V \oplus W) = \dim(V) + \dim(W)$ and $\dim(V \otimes W) = \dim(V) \cdot \dim(W)$. A more interesting example arrives in a first algebraic topology class. Given a finite CW complex X its Euler characteristic $\chi(X)$ is defined to be $\chi(X) = \sum_n (-1)^n c_n$ where c_n are the number of n -cells in X . It is later shown that χ can be computed purely in terms of the homology groups of X .

Theorem 1 ([58]). *Let X be a finite CW complex, then $\chi(X) = \sum_n (-1)^n \text{rank}(H_n(X))$.*

In other words, given the homology of a finite CW complex X , one can recover its Euler characteristic. But homology distinguishes more topological spaces than Euler characteristics do, i.e. it is a stronger invariant. Not only is homology a stronger invariant, it also

provides more structure. The Euler characteristic is just an integer, but homology is a functor. Homology takes topological spaces to homology groups and continuous maps between topological spaces to group homomorphisms between their homology groups. This extra structure can be leveraged to gain even more information about the initial topological spaces.

Khovanov took this basic concept much further by building a cohomology theory for links with graded Euler characteristic equal to the Jones polynomial of the link. This construction upgrades the Kauffman state sum for the Jones polynomial to a homology theory. Khovanov's seminal work inspired similar categorifications where polynomial invariants were lifted to cohomology theories. This is especially true for graphs, where many graph invariants were categorified using inclusion exclusion formulae such as the chromatic polynomial [60], the Tutte polynomial [69], the dichromatic polynomial [109] and the chromatic symmetric function [103] among others. More recently, a general framework was described for categorification over thin posets [26] which is further generalized in [50].

A geometric categorification of the chromatic polynomial of a graph was introduced in [48]. From graphs they build manifolds which generalize configuration spaces. They then show that these graph configuration spaces have Euler characteristics equal to the evaluations of chromatic polynomials. In [13], it was shown that the state sum and graph configuration space categorifications of the chromatic polynomial are related. In particular, there is a spectral sequence from the chromatic homology in [60] to the geometric graph homology in [48].

A third type of object from which one can extract the chromatic polynomial of a graph is the coloring ideal of a graph [108]. Brenti asked if the chromatic polynomial of a graph is Hilbert [18]. Steingrimsón showed this was true. Specifically, given a graph G , Steingrimsón constructed a monomial ideal with Hilbert polynomial equal to the chromatic polynomial of G . The coloring ideal was later shown to be a special case of monomial ideals introduced previously in [65].

A natural question to ask is if similar constructions exist for generalizations of graphs since there are many polynomial invariants of more general structures. These include chromatic polynomials of hypergraphs, characteristic polynomials of hyperplane and subspace arrangements, as well as the Bott polynomial for cell complexes [17] among others. This was answered in the affirmative for the first two constructions in [35, 37]. In these works, the authors generalize configuration spaces and state sum homology to simplicial complexes. This results in a novel polynomial invariant of simplicial complexes,

the simplicial chromatic polynomial. Since the coloring ideal is a special case of another construction, there are many generalizations. These include a generalization from graphs to hypergraphs and to subspace arrangements [20, 67]. Coloring ideals and associated coloring complexes were furthered studied in [64, 72, 79, 99].

In this work we build a new configuration space construction related to the Cooper-de Silva-Sazdanovic construction [35] for simplicial complexes and directly generalize the Eastwood-Huggett chromatic theory for graphs. This generalization arises naturally from the deletion-contraction recurrence for the chromatic polynomials of hypergraphs. We show that through the bijection between clutters and simplicial complexes, that simplicial chromatic state sum homology is isomorphic to hypergraph chromatic homology. As a corollary, we get that all simplicial chromatic polynomials are chromatic polynomials of associated hypergraphs. Furthermore, this leads to the proof of the existence of a coloring complex for the simplicial chromatic polynomial, which is the hypergraph coloring complex from [20] for its associated hypergraph. Therefore, the simplicial chromatic polynomial is the Hilbert polynomial of a polynomial ideal.

1.1 Background

1.1.1 Preliminaries and Definitions

This section provides basic definitions and explanatory examples. Further details can be found in [58], [83], [107] and [86].

Definition 1. A **graph** $G = (V, E)$ is a set of vertices V and a set of edges $E \subseteq V \times V$. In cases of ambiguity, V is referred to as $V(G)$ and E as $E(G)$.

Graphs are frequently studied through proper colorings and the related chromatic number and chromatic polynomials.

Definition 2. [16] Let $G = (V, E)$ be a graph. A **proper coloring** of G is a function $c : V \rightarrow [k]$ such that $c(v_i) \neq c(v_j)$ for every $(v_i, v_j) \in E$.

An example of a proper coloring of the complete graph on 3 vertices, K_3 , is pictured in Figure 1.1. The chromatic polynomial of a graph counts the number of proper colorings of a graph.

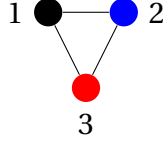


Figure 1.1: A proper coloring of K_3 with 3 colors. No two adjacent vertices are assigned the same color.

Definition 3 ([16]). Let $G = (V, E)$ be a graph, then there is a polynomial denoted $\chi_G(t)$ which counts the number of proper colorings of G with t colors. This polynomial is called the **chromatic polynomial** of G .

Example 1. For example, consider the complete graph on n vertices, K_n . Given t colors, we can assign a color to the vertices one at a time. Initially, no vertices are colored so we can color the first vertex with any of the t colors. The next vertex we color can only be colored with $t - 1$ colors since all vertices are adjacent and the previous vertex already has a color. Similarly, as we proceed to color the i th vertex there will be $t - (i - 1)$ colors left to color it with since all vertices are adjacent. Hence, in total there are $t(t - 1) \cdots (t - (n - 1))$ and so $\chi_{K_n}(t) = t(t - 1) \cdots (t - (n - 1))$.

Graphs are naturally generalized by simplicial complexes.

Definition 4. A **simplicial complex** on a vertex set V is a collection of subsets of V that is closed under taking subsets.

Simplicial complexes are determined by their maximal faces or facets. They are also determined by their minimal nonfaces.

Definition 5 ([86]). Let S be a simplicial complex on vertex set V . A subset $F \subseteq V$ is called a **minimal nonface** of S if F is not a face of S and every subset of F is a face of S . We denote the set of minimal nonfaces of S by $\text{MNF}(S)$.

Definition 6. Let S be a simplicial complex, then its **Alexander dual**, denoted by S^* is the simplicial complex with faces equal to the complements of nonfaces of S .

One way to combine two simplicial complexes to produce a new one is the join.

Definition 7 ([83]). Let S_1 and S_2 be simplicial complexes with vertex sets V_1 and V_2 . The **join** of S_1 and S_2 , denoted $S_1 * S_2$, is the simplicial complex on vertex set $V_1 \sqcup V_2$ and the set of simplices $\{F_1 \sqcup F_2 : F_1 \in S_1 \text{ and } F_2 \in S_2\}$.

The Stanley-Reisner ideal and the Stanley-Reisner ring of a simplicial complex S are defined in terms of the nonfaces of S .

Definition 8 ([86]). *Let \mathbf{k} be a field and S be a simplicial complex on $[n]$, then the **Stanley-Reisner ideal** of S is $I_S = (x_{i_1} \cdots x_{i_j} : \{i_1, \dots, i_j\} \notin S)$, the ideal of $\mathbf{k}[x_1, \dots, x_n]$ generated by square-free monomials corresponding to the nonfaces of S .*

The Stanley-Reisner ring of a simplicial complex is the polynomial ring $\mathbf{k}[x_1, \dots, x_n]$ modded out by the Stanley-Reisner ideal.

Definition 9 ([86]). *Let \mathbf{k} be a field and S be a simplicial complex on $[n]$, then the **Stanley-Reisner ring** of S is $\mathbf{k}[x_1, \dots, x_n]/I_S$.*

Simplicial complexes can be further generalized to hypergraphs.

Definition 10 ([19]). *A **hypergraph** $\mathcal{G} = (V, E)$ is a set of vertices V along with a set of subsets of the vertex set, which we denote by E . Each set in E is called a **hyperedge**. We will sometimes refer to the edges of \mathcal{G} as $E(\mathcal{G})$.*

One important subclass of hypergraphs, called clutters, are in bijection with simplicial complexes.

Definition 11 ([118]). *A hypergraph $\mathcal{G} = (V, E)$ is a **clutter** if for all hyperedges $e \in E$ there does not exist $e' \in E$ such that $e' \subset e$.*

Clutters are in bijection with their independence complexes [118].

Definition 12 ([118]). *Let $\mathcal{G} = (V, E)$ be a hypergraph, then the **independence complex** of \mathcal{G} , denoted by $\Delta(\mathcal{G})$, equals $\Delta(\mathcal{G}) = \{e \subset [n] : e' \not\subseteq e \text{ for all } e' \in E\}$.*

The faces of $\Delta(\mathcal{G})$ are the independent sets of \mathcal{H} and the minimal nonfaces of $\Delta(\mathcal{H})$ are the hyperedges of \mathcal{H} . Since the bijection between clutters and their independence complexes will be important, we define notation for the unique clutter with independence complex S .

Definition 13. *Given a simplicial complex S , denote by \mathcal{G}_S the hypergraph with the vertex set of S and hyperedges equal to the minimal nonfaces of S called the **hypergraph of minimal nonfaces** for S .*

Throughout, we will refer to the subhypergraph of \mathcal{H} with the same vertices as \mathcal{H} and hyperedges in $E' \subseteq E$ which we denote by $[\mathcal{H} : E']$.

In graph theory the chromatic polynomial of a graph $\chi_G(t)$ is a well-studied polynomial which gives information about the number of proper colorings of G with t colors. There is an analogue in hypergraph theory which restricts to the ordinary chromatic polynomial in the case that a hypergraph \mathcal{G} is a graph.

Definition 14. Let $\mathcal{G} = (V, E)$ be a hypergraph. A **proper coloring** of \mathcal{G} is a function $c : V \rightarrow [k]$ such that for every $F \in E$ there exist $v_1, v_2 \in F$ such that $c(v_1) \neq c(v_2)$. That is, no hyperedge is monochromatic.

Hereafter, we refer to proper colorings as colorings.

Definition 15 ([59]). Let \mathcal{G} be a hypergraph, then there is a polynomial called the **chromatic polynomial** of \mathcal{G} which is denoted $\chi_{\mathcal{G}}(t)$, such that $\chi_{\mathcal{G}}(t)$ counts the number of colorings of \mathcal{G} with t colors.

The chromatic polynomials of hypergraphs satisfy many generalizations of well-known properties of the chromatic polynomials of graphs. For example, the chromatic polynomial of a hypergraph has a state sum formula. To prove this we make use of the following lemma from [112].

Lemma 1 (Lemma 1.2 [112]). Let $\mathcal{G} = (V, E)$ be a hypergraph with n vertices. Then $\chi_{\mathcal{G}}(t) = t^n + a_{n-1}t^{n-1} + \dots + a_1t$, where

$$a_j = \sum_{i \geq 0} (-1)^i N(i, j) \text{ for } 1 \leq j \leq n-1$$

and $N(i, j)$ denotes the number of subhypergraphs $[\mathcal{G} : s]$ where $|s| = i$ and the number of components of $[\mathcal{G} : s]$ is j .

Proposition 1. Let $\mathcal{G} = (V, E)$ be a hypergraph. For any $s \subseteq E$ denote the number of components of $[\mathcal{G} : s]$ by $k(s)$. Then

$$\chi_{\mathcal{G}}(t) = \sum_{i \geq 0} \sum_{\substack{s \subseteq E, \\ |s|=i}} (-1)^i t^{k(s)}.$$

Proof. From Lemma 1 it follows that $\chi_{\mathcal{G}}(t) = \sum_{j \geq 0} \sum_{i \geq 0} (-1)^i N(i, j) t^j$, but this can be rewritten by summing over all subsets of E by size $\chi_{\mathcal{G}}(t) = \sum_{i \geq 0} \sum_{\substack{s \subseteq E, \\ |s|=i}} (-1)^i t^{k(s)}$. \square

Another defining property of the chromatic polynomial of a hypergraph is a deletion-contraction recurrence.

Theorem 2 ([71]). *Let \mathcal{G} be a hypergraph and fix a hyperedge f of \mathcal{G} , then $\chi_{\mathcal{G}}(t)$ satisfies the deletion-contraction recurrence for hypergraphs*

$$\chi_{\mathcal{G}}(t) = \chi_{\mathcal{G}-f}(t) - \chi_{\mathcal{G}/f}(t). \quad (1.1)$$

Lastly, the definition of weak compositions will be necessary for the succinct statement of a later result.

Definition 16 ([107]). *Given $n \in \mathbb{N}$ a solution to the equation $x_1 + x_2 + \dots + x_k = n$ using nonnegative integers is called a **weak composition** of n into k parts.*

1.1.2 Chromatic Graph Homology

Chromatic graph homology categorifies the chromatic polynomial of a graph. It is a bigraded homology theory introduced in [60]. Chromatic graph homology can be viewed in two equivalent ways. The first uses a cube complex construction similar to the approach taken in [12] for Khovanov homology. The second uses a state sum formula as in [115] for Khovanov homology. The two constructions are then shown to be equivalent. In [62], chromatic graph homology was generalized from $\mathbb{Z}[x]/(x^2)$ to a more general class of algebras. A more general definition is provided here.

The initial definition of chromatic graph homology requires an ordering on the edges of a graph. Let $G = (V, E)$ be a graph with such an ordering.

Definition 17 ([62]). *A **graded R algebra** \mathcal{A} is an R -algebra with a direct sum decomposition $\mathcal{A} = \bigoplus_{i=0}^{\infty} A_i$ such that $a_i a_j \in A_{i+j}$ for all $a_i \in A_i$ and $a_j \in A_j$. Elements in A_i are called **homogeneous of degree i** .*

For simplicity we assume the following conditions on algebras \mathcal{A} :

Assumption 1 ([62]). *$\mathcal{A} = \bigoplus_{i=0}^{\infty} A_i$ is a commutative graded algebra over \mathbb{Z} such that each A_j is a free \mathbb{Z} -module of finite rank.*

These assumptions can be relaxed, for more details see [62, 93, 13].

Definition 18 ([62]). *Given a commutative graded \mathbb{Z} -algebra \mathcal{A} its **q -dimension** is $q\dim(\mathcal{A}) = \sum_j q^j \text{rank}(A_j)$.*

Consider the following example to make this definition clear.

Example 2. Let $\mathcal{A}_k = \mathbb{Z}[x]/(x^k)$, then $\text{qdim}(\mathcal{A}_k) = 1 + q + q^2 + \dots + q^k$.

We also define graded chain complexes.

Definition 19 ([62]). Let M and N be graded \mathbb{Z} -modules. A \mathbb{Z} -module map $\phi : M \rightarrow N$ is **graded with degree d** if $\phi(M_j) \subseteq N_{j+d}$. A **graded chain complex** is a chain complex where the chain groups are graded \mathbb{Z} -modules and the differentials are degree preserving.

We now define the graded Euler characteristic.

Definition 20 ([62]). Let C be a graded cochain complex, then its **graded Euler characteristic**, denoted $\chi_q(C)$, is $\chi_q(C) = \sum_{i=0}^{\infty} (-1)^i \text{qdim}(H^i(C))$.

It is shown in [12] that under certain conditions the graded Euler characteristic of a graded chain complex can be computed directly from its graded chain groups.

Proposition 2 ([12]). Let C be a graded cochain complex with degree preserving differential and such that all its cochain groups have finite free rank, then

$$\chi_q(C) = \sum_{i=0}^{\infty} (-1)^i \text{qdim}(H^i(C)) = \sum_{i=0}^{\infty} (-1)^i \text{qdim}(C^i).$$

Definition 21 ([62]). Let \mathcal{A} be as in Assumption 1.1.2. An **enhanced state** of G is $S = (s, \ell)$ where $s \subseteq E$ and ℓ is an assignment of an element of \mathcal{A} to each connected component of $[G : s]$.

Definition 22 ([62]). Let \mathcal{A} be as in Assumption 1.1.2 and let $G = (V, E)$ be a graph. The **chromatic graph chain groups**, $C^i(G)$, are $C^i(G) = \bigoplus_{|s|=i, s \subseteq E(G)} C_s^i(G)$ where $C_s^i(G) = \mathcal{A}^{\otimes k(s)}$ and $k(s)$ is the number of components of $[G : s]$.

Remark 1. The chromatic graph homology chain groups $C^i(G)$ depend on the algebra \mathcal{A} , and therefore should be written as $C_{\mathcal{A}}^i(G)$. We omit the subscript when the algebra being used is clear.

The chromatic graph chain groups for K_3 , the complete graph on three vertices is pictured in Figure 1.2.

An enhanced state $S = (s, \ell)$ of a graph G is called homogeneous if $\ell(E_i)$ is a homogeneous element of \mathcal{A} for all components E_1, \dots, E_k of $[G : s]$. Given a homogeneous enhanced state

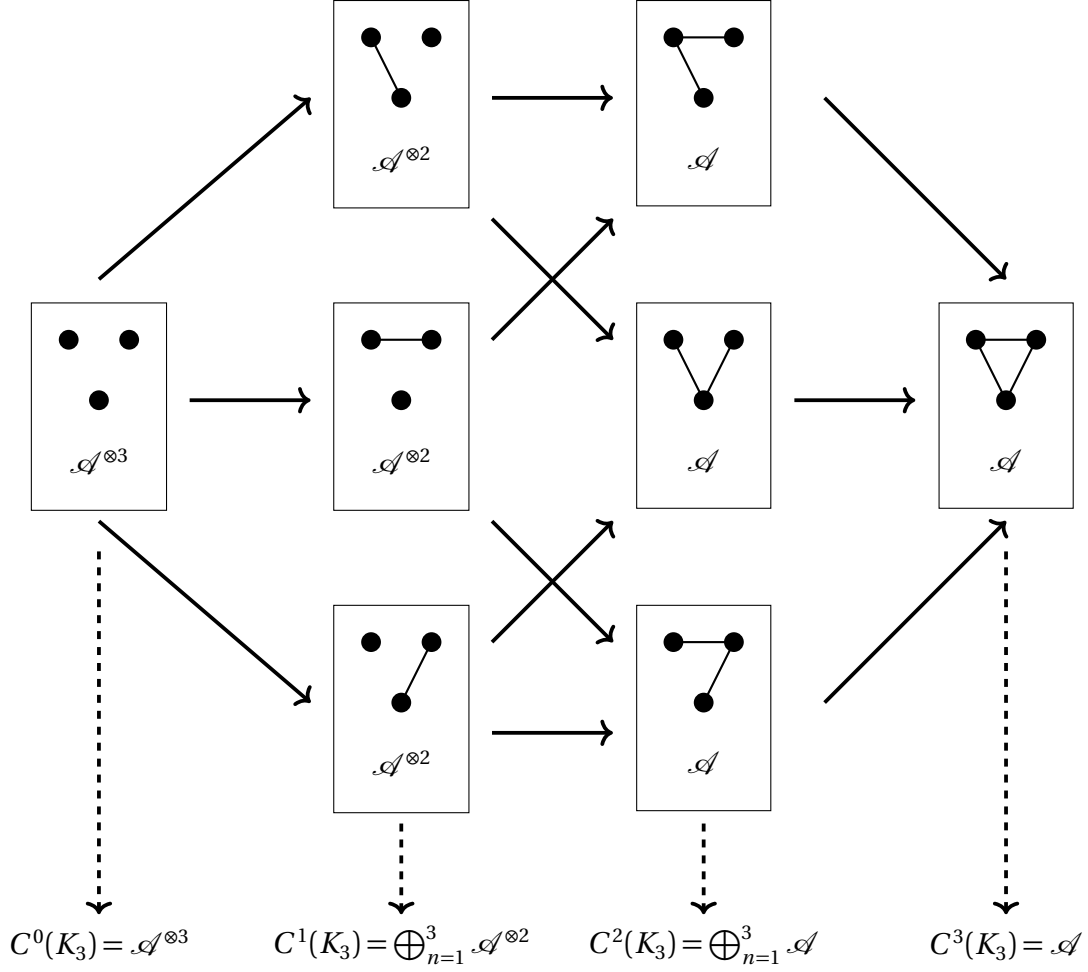


Figure 1.2: The chromatic graph chain groups of the triangle graph K_3 .

From S we can define $j(S) = \sum_i \deg(\ell(E_i))$. The grading of \mathcal{A} induces a second grading j on $C^i(G)$ such that $C^{i,j}(G)$ is generated by homogeneous enhanced states where $|s| = i$ and $j(S) = j$.

To define the differential, a few pieces of notation need to be introduced. Given an edge $e \in E$ and $s \subseteq E$, let $n_s(e)$ denote the number of edges in s preceding e in the given ordering. From $S = (s, \ell)$ define a new enhanced state $S_e = (s_e, \ell_e)$ where $s_e = s \cup \{e\}$ and ℓ_e is defined in the following way. Let K_1, \dots, K_n be the connected components of $[G : s]$. The edge e in $[G : s]$ either connects a component to itself or to another component. If it connects a component to itself, without loss of generality K_1 , then the components of $[G : s \cup \{e\}]$ are $K_1 \cup \{e\}, K_2, \dots, K_n$. We then define $\ell_e(K_1 \cup \{e\}) = \ell(K_1), \ell_e(K_2) = \ell(K_2), \dots, \ell_e(K_n) = \ell(K_n)$. If adding e connects, without loss of generality, K_1 to K_2 , then we define $\ell_e(K_1 \cup K_2) =$

$\ell(K_1) \cdot \ell(K_2), \dots, \ell_e(K_n) = \ell(K_n)$. If $\ell_e(K_1 \cup K_2) = 0$, then we define $S_e = 0$.

Definition 23. [60] Let $G = (V, E)$ be a graph. The differential $d : C^{i,j}(G) \rightarrow C^{i+1,j}(G)$ is defined on an enhanced state $S \in C^{i,j}(G)$ by

$$d(S) = \sum_{e \in E \setminus s} (-1)^{n_s(e)} S_e.$$

A somewhat tedious calculation or more general argumentation from [26] shows that $d^2 = 0$, i.e. d is a differential. It is also shown in [60] that chromatic graph homology is independent of the ordering of the edges of the graph. This is done by constructing an isomorphism between chain complexes of a graph with two different orderings of its edges. The argument is omitted here, but the analogous argument for hypergraphs is made later.

Definition 24. [60] Given a graph G the **chromatic graph homology** of G are the abelian groups that form the homology of the chain complex $(C^{**}(G), d)$ denoted $H^{**}(G)$.

In [60] it is shown that if an edge e from a graph $G = (V, E)$ is fixed, then $C^{i-1,j}(G \setminus e)$, $C^{i,j}(G)$ and $C^{i,j}(G/e)$ fit into a short exact sequence

$$0 \longrightarrow C^{i-1,j}(G/e) \xrightarrow{\alpha_{ij}} C^{i,j}(G) \xrightarrow{\beta_{ij}} C^{i,j}(G \setminus e) \longrightarrow 0$$

where α_{ij} and β_{ij} (which we refer to as α and β) are defined as follows. Let e be an edge of G with endpoints v_e and w_e . Let $S = (s, \ell)$ be an enhanced state of G/e for some edge e of G . Set $\tilde{s} = s \cup \{e\}$. Since the components of $[G/e : s]$ and $[G : \tilde{s}]$ are the same except the one where $v_e \in G/e$ is expanded to e in G , ℓ yields a labeling of the components of $[G : \tilde{s}]$. We denote this labeling by $\tilde{\ell}$. Then set $\alpha(S) = (\tilde{s}, \tilde{\ell})$ which is an enhanced state in $C^{i,j}(G)$. Extending α linearly yields a homomorphism $\alpha : C^{i-1,j}(G/e) \rightarrow C^{i,j}(G)$.

To define β let $S = (s, \ell)$ be an enhanced state of G . If e is not in s then S is an enhanced state of $G \setminus e$ and we set $\beta(S) = S$. If $e \in s$ then we set $\beta(S) = 0$. Extending β linearly yields a homomorphism $\beta : C^{i,j}(G) \rightarrow C^{i,j}(G \setminus e)$.

Applying the snake lemma to this short exact sequence yields a long exact sequence which categorifies the chromatic polynomial for graphs upon computing its graded Euler characteristic.

Theorem 3 ([60]). Let G be a graph and \mathcal{A} be as in 1.1.2. Chromatic graph homology with \mathcal{A} categorifies the chromatic polynomial, that is, $\chi_q(H^i(G)) = \chi_G(\text{qdim}(\mathcal{A}))$.

1.1.3 Graph Configuration Spaces

Chromatic graph homology is not the only categorification of the chromatic polynomial. Given a graph G and an input manifold¹ M , Eastwood and Huggett built a new manifold M_G with Euler characteristic equal to the chromatic polynomial of G evaluated at the Euler characteristic of M [48].

These new manifolds, M_G , called graph configuration spaces were introduced by Eastwood and Huggett and are defined below. Let M be an m -dimensional manifold and let $G = (V, E)$ be a graph with n vertices. Let $e = (v_i, v_j) \in E$ then we define the diagonal corresponding to e by

$$\Delta_e = \{(x_1, \dots, x_n) \in M^{\times n} : x_i = x_j\}.$$

Definition 25 ([48]). *Let $G = (V, E)$ be a graph and M be a manifold. Then the **graph configuration space** of G is*

$$M_G = M^{\times n} \setminus \bigcup_{e \in E} \Delta_e. \quad (1.2)$$

M_G is referred to as a graph configuration space because it generalizes configuration spaces as introduced in [51]. To see this, recall the definition of a configuration space.

Definition 26. *Let X be a topological space, then its **nth configuration space** is $\text{Conf}_n(X) = X^n \setminus \{(x_1, \dots, x_n) \in X^n : x_i = x_j \text{ for } i \neq j\}$.*

Let X be a topological space and let $G = K_n$ be the complete graph on n vertices. Then the graph configuration space of G using the topological space X is equal to $\text{Conf}_n(X)$. Essentially, while the regular configuration space requires $x_i \neq x_j$ for all $i \neq j$, the graph configuration space relaxes this condition to $x_i \neq x_j$ only when v_i and v_j are connected by an edge.

Eastwood and Huggett use the Leray sequence [76] to show that their construction obeys a deletion-contraction formula.

Proposition 3. [76] *Let X be an orientable manifold and Y be a closed submanifold of X of codimension m with orientable normal bundle. The homology of X , Y and $X \setminus Y$ are related by the Leray long exact sequence*

$$\dots \rightarrow H_i(X \setminus Y) \xrightarrow{f} H_i(X) \xrightarrow{g} H_{i-m}(Y) \rightarrow H_{i-1}(X \setminus Y) \xrightarrow{f} \dots \quad (1.3)$$

¹Manifolds are used for convenience, but more general spaces can be used e.g. simplicial spaces.

The map $f : H_i(X \setminus Y) \rightarrow H_i(X)$ is induced by the inclusion of $X \setminus Y$ into X and the map $g : H_i(X) \rightarrow H_{i-m}(Y)$ comes from taking a cycle in X and intersecting it with Y to obtain a cycle in Y .

Theorem 4 ([48]). *Let M be an orientable m -dimensional manifold with m even² and let $G = (V, E)$ be a graph with n vertices. Fix an edge $f \in E$. Then*

$$\chi(M_G) - \chi(M_{G \setminus f}) + \chi(M_{G/f}) = 0.$$

Proof. Fix an edge $f \in E$ and let $X - Y = M_G$, $X = M^{\times n} \setminus \bigcup_{e \in E, e \neq f} \Delta_e$, and $Y = X \cap \Delta_f$ which implies $M_{G \setminus f} = X$ and $M_{G/f} = Y$. Then apply the Leray sequence (1.3) to X and Y to obtain

$$0 \rightarrow H_{mn}(M_G) \rightarrow H_{mn}(M_{G \setminus f}) \rightarrow H_{mn-m}(M_{G/f}) \rightarrow H_{mn-1}(M_G) \rightarrow \dots \quad (1.4)$$

Applying the Euler characteristic to this long exact sequence yields the result. □

From the multiplicativity of the Euler characteristic it follows that $\chi(M_G)$ is equal to $\chi(G, t)$ evaluated at $t = \chi(M)$.

Theorem 5 ([48]). *Let M be an orientable even dimensional manifold and G be a graph. Then $\chi(M_G) = \chi(G, t)$ when $t = \chi(M)$.*

If $M = \mathbb{C}\mathbb{P}^{t-1}$, then the chromatic polynomial of G is recovered.

Corollary 1 ([48]). *Let $G = (V, E)$ be a graph and let $M = \mathbb{C}\mathbb{P}^{t-1}$, then $\chi_G(t) = \chi(M_G)$.*

1.1.4 Simplicial Chromatic Polynomial

In [35], Cooper-de Silva-Sazdanović generalized the Eastwood and Huggett by building simplicial configuration spaces. This generalizes [48] in the sense that given a graph G the graph configuration space of G is equal to the simplicial configuration space of the independence complex of G . The construction leads to a novel invariant of simplicial complexes called the simplicial chromatic polynomial. In [37], Cooper-Sazdanović construct a state sum formulation for the simplicial chromatic polynomial.

²Even dimension is required to get the correct sign relating the Euler characteristics that categorify the deletion-contraction recurrence.

Let M be an m -dimensional manifold and let S be a simplicial complex on a vertex set V of size n . Let $\sigma = \{v_{i_1}, \dots, v_{i_k}\}$ be a subset of V , then we define the diagonal corresponding to σ to be $D_\sigma = \{(x_1, \dots, x_n) \in M^{\times n} : x_{i_1} = x_{i_2} = \dots = x_{i_k}\}$.

Definition 27 ([35]). *Let S be a simplicial complex on vertex set V and let M be a manifold. Then the **simplicial configuration space** is defined to be*

$$M_S = M^n \setminus \bigcup_{\sigma \in \text{NF}(S)} D_\sigma = M^n \setminus \bigcup_{\sigma \in \text{MNF}(S)} D_\sigma$$

where $\text{NF}(S)$ is the set of nonfaces of S and $\text{MNF}(S)$ is the set of minimal nonfaces of S .

The starting point of Eastwood and Hugget's categorification of the chromatic polynomial of a graph was the deletion-contraction recurrence [48]. In the simplicial complex case this notion had to be introduced. Let S be a simplicial complex and let σ be a face of S .

Definition 28. *The **deletion** of σ , denoted $S \setminus \sigma$, is the simplicial complex S with σ and any face that contains σ deleted.*

Definition 29. *The **contraction** of σ , denoted S/σ , is the simplicial complex on the vertex set $V/\sigma = (V \setminus \sigma) \cup v$. A subset $T \subseteq V/\sigma$ either contains v or not. If it contains v , then T is a face of S/σ if and only if $(T \setminus v) \cup \sigma$ is a face of S . If it does not contain v , then T is a face of S/σ if and only if T is a face of S .*

Using the Leray sequence and a similar argument to [48], a deletion-contraction long exact sequence is described for configuration spaces of simplicial complexes.

Theorem 6 ([35]). *Let S be a simplicial complex with n vertices and σ a k -simplex of S . Let M be an orientable even dimensional manifold of dimension m , then*

$$0 \rightarrow H_{mn}(M_{S \setminus \sigma}) \rightarrow H_{mn}(M_S) \rightarrow H_{mn-mk}(M_{S/\sigma}) \rightarrow H_{mn-1}(M_{S \setminus \sigma}) \rightarrow \dots \quad (1.5)$$

is a long exact sequence.

Fixing $M = \mathbb{C}\mathbb{P}^{t-1}$ as in [48] and taking the Euler characteristic of M_S yields a new polynomial invariant of simplicial complexes called the simplicial chromatic polynomial.

Definition 30 ([35]). *Let S be a simplicial complex and let $M = \mathbb{C}\mathbb{P}^{t-1}$, then the **simplicial chromatic polynomial**, denoted $\chi_c(S, t)$, is $\chi_c(S, t) = \chi(M_S)$.*

Through the independence complex, the simplicial chromatic polynomial generalizes the chromatic polynomial of a graph [35].

Proposition 4 ([35]). *Let G be a graph, then $\chi_G(t) = \chi_c(\Delta(G), t)$.*

Further properties of this polynomial are explored in [35, 36, 37]. In particular, a state sum construction for the simplicial chromatic polynomial analogous to [60] is introduced in [37].

This construction can be viewed in more general terms.

Definition 31. *Let V be a finite vertex set with $|V| = n$ and f be a function with domain V . Then the **equality complex** of f , $\text{Eq}(f)$ is the simplicial complex on V where $T \subseteq V$ forms a face of $\text{Eq}(f)$ if and only if f restricts to a constant function on T .*

Points $x = (x_1, \dots, x_n)$ in M^n are in bijection with maps $f_x : V \rightarrow M$ where $f_x(v_i) = x_i$ for all i . The simplicial configuration space can then be redefined in terms of the equality complex.

Definition 32. *Let S be a simplicial complex on vertex set V with $|V| = n$. The **simplicial configuration space** of S is $M_S = \{x \in M^n : \text{Eq}(f_x) \text{ is a subcomplex of } S\}$.*

Given vertex sets V and W a map $g : V \rightarrow W$ induces a map $M^g : M^{|W|} \rightarrow M^{|V|}$ defined by $M^g(x)(v) = f_x(g(v))$ for $v \in V$.

Lemma 2. *Let $f : V \rightarrow W$ be a map between vertex sets. Then the pullback of $M_S \rightarrow M^{|V|} \leftarrow M^{|W|}$ is $M(T)$ where T is the simplicial complex on W whose simplices are subsets of W whose preimage under f is a simplex of S . Equivalently, T is the largest simplicial complex on W whose pullback to V (defined to be the set of subsets of V whose image under f is a simplex of T) is contained in S .*

Proof. For x in $M^{|W|}$, the pullback of $\text{Eq}(x)$ to V is $\text{Eq}((M^f)(x))$. We seek the points x such that $\text{Eq}((M^f)(x))$ is a subcomplex of S . Thus we seek the points x such that the pullback of $\text{Eq}(x)$ to V is a subcomplex of S . In other words, we seek the points x such that $\text{Eq}(x)$ is contained in T as described above, this is M_T . \square

Let S be a simplicial complex on V that contains some simplex σ . Consider the quotient map $f : V \rightarrow V/\sigma$ that collapse the vertices of σ to a single vertex. Let $W := V/\sigma$. Then $M^f : M^{|W|} \rightarrow M^{|V|}$ is a submanifold embedding whose image is the diagonal $\Delta_\sigma = \{x : \text{Eq}(x) \text{ contains } \sigma\}$. Lemma 1.1.4 identifies $M_S \cap \Delta_\sigma$ with M_T and $M_S \setminus \Delta_\sigma = M_{S \setminus \sigma}$.

1.2 Chromatic Hypergraph Homology

In this chapter chromatic hypergraph homology is defined. This generalizes chromatic graph homology as originally introduced in [60] and briefly described in Section 1.1.2. Chromatic hypergraph homology categorifies the chromatic polynomial of a hypergraph.

1.2.1 State Sum Formulation

Graphs are hypergraphs where hyperedge sizes are restricted to be of size 2. The chromatic graph homology construction does not depend on this difference in any meaningful way.

Definition 33. Let \mathcal{A} be as in Assumption 1.1.2. An **enhanced state** of a hypergraph $\mathcal{G} = (V, E)$ is $S = (s, \ell)$ where $s \subseteq E$ and ℓ is an assignment of an element of \mathcal{A} to each connected component of $[\mathcal{G} : s]$.

Definition 34. Let \mathcal{A} be as in Assumption 1.1.2. Let $\mathcal{G} = (V, E)$ be a hypergraph, then $C^i(\mathcal{G}) = \bigoplus_{|s|=i, s \subseteq E(\mathcal{G})} C_s^i(\mathcal{G})$ where $C_s^i(\mathcal{G}) = \mathcal{A}^{k(s)}$ and $k(s)$ is the number of components of $[\mathcal{G} : s]$. The $C^i(\mathcal{G})$ are called the **chromatic hypergraph chain groups**.

An example of the chromatic hypergraph chain groups for a hypergraph on 4 vertices is pictured in Figure 1.3.

An enhanced state $S = (s, \ell)$ of a hypergraph \mathcal{G} is called homogeneous if $\ell(E_i)$ is a homogeneous element of \mathcal{A} for all components E_1, \dots, E_k of $[\mathcal{G} : s]$. Given a homogeneous enhanced state S we can define $j(S) = \sum_i \deg(\ell(E_i))$. The grading of \mathcal{A} induces a second grading j on $C^i(\mathcal{G})$ such that $C^{i,j}(\mathcal{G})$ is generated by homogeneous enhanced states where $|s| = i$ and $j(S) = j$.

We now order the hyperedges of \mathcal{G} as follows, F_1, \dots, F_n . This allows us to define the differential

$$d : C^{i,j}(\mathcal{G}) \rightarrow C^{i+1,j}(\mathcal{G})$$

which takes each labeled state $S = (s, \ell) \in C^{i,j}(\mathcal{G})$ to:

$$d(S) = \sum_{F \in E \setminus s} (-1)^{n(F)} S_F \tag{1.6}$$

where $n(F)$ and S_F are defined as follows. We define $n(F)$ to be the number of hyperedges in s which come before F in the ordering of the hyperedges. S_F is either a labeled state

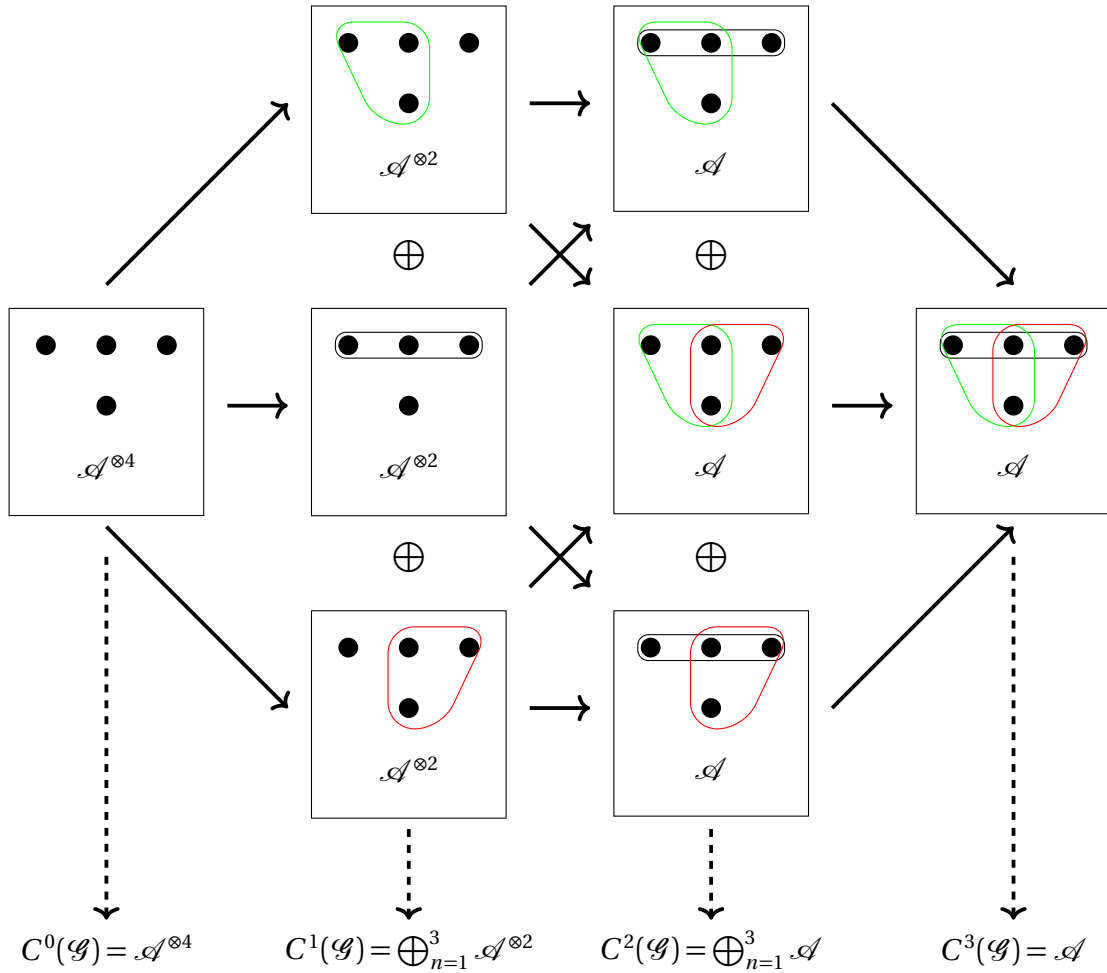


Figure 1.3: The chromatic hypergraph chain groups of a hypergraph $\mathcal{G} = (V = [4], \{\{1, 2, 3\}, \{1, 2, 4\}, \{2, 3, 4\}\})$. The chromatic polynomial of \mathcal{G} is $\chi(\mathcal{G}, t) = t^4 - 3t^2 + 2t$

or 0 as follows. Let $s_F = s \cup \{F\}$ and denote the components of $[\mathcal{G} : s]$ by E_1, \dots, E_k . If F connects some E_i to itself, without loss of generality, E_1 , then the components of $[\mathcal{G} : s \cup \{F\}]$ are $E_1 \cup F, E_2, \dots, E_k$. In this case, $\ell_F(E_1 \cup F) = \ell(E_1), \ell_F(E_2) = \ell(E_2), \dots, \ell_F(E_k) = \ell(E_k)$. That is, when F connects one component back to itself then ℓ_F maintains the labelings of all the components. The other possibility is that F connects some number of different components together, without loss of generality say, E_1, \dots, E_r . Then the components are $\bigcup_{i=1}^r E_i, E_{r+1}, \dots, E_k$ and we define $\ell_F(\bigcup_{i=1}^r E_i) = \prod_{i=1}^r \ell(E_i), \ell_F(E_{r+1}) = \ell(E_{r+1}), \dots, \ell_F(E_k) = \ell(E_k)$. That is to say, when F connects components together then ℓ_F preserves the labelings of ℓ on all the components not involved in the connection and takes the product of the labels of all the components connected together by F to be the label of this new component. Note that $\ell_F(\bigcup_{i=1}^r E_i)$ may be equal to 0, and in this case we set $S_F = 0$ and otherwise $S_F = (s_F, \ell_F)$.

Proposition 5. *The map d , defined in (1.6), is a differential.*

Proof. Let $S = (s, \ell)$ be a labeled state. Then by definition

$$d(d(S)) = \sum_{F' \in E \setminus s \cup \{F\}} \sum_{F \in E \setminus s} (-1)^{n(F') + n(F)} S_F.$$

The state $s \cup \{F\} \cup \{F'\}$ appears twice in $d(d(S))$. Once from adding F to S in the internal sum and then adding F' and once from adding the edges in the reverse order. We wish to show that the two instances of the state appear with opposite signs in the sum. Without loss of generality let F precede F' in the given ordering on the edges. One must consider the 4 cases when there are an even/odd number of edges preceding F and F' in s . Assume an even number of edges precede F in s and an odd number precede F' in s . The number of edges preceding F' in $s \cup \{F\}$ must therefore be even. So the sign on $S_{s \cup \{F\} \cup \{F'\}}$ which results from adding F and then F' is even. Since an even number of edges precede F in s and F' comes after F there are an even number of edges preceding F in $s \cup \{F'\}$. Therefore the sign on $S_{s \cup \{F'\} \cup \{F\}}$ coming from adding F' and then F must be odd. The other three cases are similar. \square

Remark 2. *Proposition 5 also follows from [26] since this is a categorification defined over a boolean poset, which therefore admits a balanced coloring.*

Definition 35. *Given a hypergraph \mathcal{G} and an ordering of its hyperedges, the **chromatic hypergraph chain complex** of \mathcal{G} is $(C^{**}(\mathcal{G}), d)$.*

The chain complex as currently defined depends on the ordering of the hyperedges. To show it is independent of this ordering, we prove that for any ordering of hyperedges the resulting chain complexes are isomorphic. This shows that chromatic hypergraph homology is a hypergraph invariant.

Theorem 7. *The homology of $(C^{**}(\mathcal{G}), d)$ is independent of the ordering of hyperedges.*

Proof. Let \mathcal{G} be a hypergraph with some ordering of hyperedges. Let σ be a permutation on the set of hyperedges of \mathcal{G} . Let \mathcal{G}_σ denote the hypergraph \mathcal{G} with the new ordering of hyperedges induced by σ . Since adjacent transpositions generate the symmetric group, we show that $C(\mathcal{G}) \cong C(\mathcal{G}_\sigma)$ for σ an adjacent transposition.

Let e_i and e_{i+1} be the hyperedges whose orders are swapped by σ . Let $f : C(\mathcal{G}) \rightarrow C(\mathcal{G}_\sigma)$ be the map defined by sending a state S to $-S$ if it contains both e_i and e_{i+1} , the identity otherwise and extending linearly. This map is clearly an isomorphism, so it just remains to show that f commutes with the differentials d .

There are 4 cases to check: when S contains both e_i and e_{i+1} , when S contains neither e_i nor e_{i+1} , when it contains e_i but not e_{i+1} and the reverse. Let S contain e_i but not e_{i+1} then

$$d(S) = \sum_{F \in E \setminus S} (-1)^{n(F)} S_F = \sum_{F \in E \setminus (S \cup \{e_{i+1}\})} (-1)^{n(F)} S_F + (-1)^{n(e_{i+1})} S_{e_{i+1}}.$$

By definition f is the identity on S_F for any $F \in E \setminus (S \cup \{e_{i+1}\})$ but $f((-1)^{n(e_{i+1})} S_{e_{i+1}}) = (-1)^{n(e_{i+1})+1} S_{e_{i+1}}$. Since S does not contain both e_i and e_{i+1} , applying f to S is just the identity. In \mathcal{G}_σ , $n(e_{i+1})$ is one less than in \mathcal{G} since e_i comes after e_{i+1} in the ordering of edges in \mathcal{G}_σ . Since $f(d(S))$ has an additional minus sign on $S_{e_{i+1}}$ compared to $d(S)$, $f(d(S)) = d(f(S))$ as required. The other three cases are similar. \square

Since the homology of $(C^{**}(\mathcal{G}), d)$ is independent of the ordering of the hyperedges of \mathcal{G} , the chromatic hypergraph homology of \mathcal{G} can be defined without a fixed ordering of hyperedges.

Definition 36. *Given a hypergraph \mathcal{G} the **chromatic hypergraph homology** of \mathcal{G} is the homology of the chain complex $(C^{**}(\mathcal{G}), d)$, denoted $H^{**}(\mathcal{G})$.*

The chromatic hypergraph homology of \mathcal{G} categorifies the chromatic polynomial of a hypergraph.

Theorem 8. *Let \mathcal{G} be a hypergraph and \mathcal{A} be an algebra satisfying Assumption 1.1.2, then $\chi_q(C^*(G)) = \chi_G(\text{qdim}(\mathcal{A}))$.*

The hypergraph construction for graphs is the same as the graph construction [60] for graphs which yields the following theorem.

Theorem 9. *The chromatic graph homology of a graph G is equal to the chromatic hypergraph homology of G .*

1.2.2 Properties of Chromatic Hypergraph Homology

Chromatic hypergraph homology satisfies many of the same properties as chromatic graph homology but also has a more complicated structure. For example, the chromatic graph homology of a connected graph is known to lie along two diagonals, this is not true for chromatic hypergraph homology as is shown in Example 3 and Theorem 11. In this section we explore the properties of chromatic hypergraph homology.

The deletion-contraction recurrence is a defining recurrence for the chromatic polynomial of a hypergraph just like for the chromatic polynomial of a graph. This recurrence can be viewed as a shadow of the short exact sequence of chain groups associated to \mathcal{G} , $\mathcal{G} \setminus f$ and \mathcal{G}/f . The proof of Proposition 6 is essentially the same as that of Lemma 3.1 in [60], since the proof in no way relies on the number of vertices in (hyper)edges.

Proposition 6. *There exists a short exact sequence of chain groups*

$$0 \longrightarrow C^{i-1,j}(\mathcal{G}/F) \xrightarrow{\alpha} C^{i,j}(\mathcal{G}) \xrightarrow{\beta} C^{i,j}(\mathcal{G}-F) \longrightarrow 0$$

where α and β are defined as follows.

- *Let \mathcal{G} be a hypergraph and F a hyperedge of \mathcal{G} . Denote by v_F the vertex in \mathcal{G}/F that F contracted to. Order the edges of \mathcal{G} so that F comes last, then there are induced orderings on \mathcal{G}/F and $\mathcal{G}-F$ given by removing F from the ordering on \mathcal{G} . We begin by defining α . Let $S = (s, \ell)$ be a labeled state of \mathcal{G}/F . Let $\tilde{s} = s \cup \{F\}$. The components of $[\mathcal{G}/F : s]$ are exactly the components of $[\mathcal{G} : \tilde{s}]$ except the component containing v_F in \mathcal{G}/F which is replaced by F in \mathcal{G} . We can therefore define $\tilde{\ell}$ by setting it equal to ℓ on the components of $[\mathcal{G} : \tilde{s}]$ that agree and equal to ℓ of the component containing v_F for the component containing F . Then $(\tilde{s}, \tilde{\ell})$ is a labeled state on \mathcal{G} and we set $\alpha(S) = (\tilde{s}, \tilde{\ell})$. Extending α linearly yields a homomorphism.*

- Let $S = (s, \ell)$ be a labeled state of \mathcal{G} . If $F \notin s$ then S is a labeled state of $\mathcal{G} - F$ and we set $\beta(S) = S$. Otherwise, if $F \in s$, then we set $\beta(S) = 0$. Extending linearly, as before, yields a homomorphism.

Proof. We first show that α is a chain map, i.e. that the following square

$$\begin{array}{ccc} C^{i-1,j}(\mathcal{G}/F) & \xrightarrow{\alpha} & C^{i,j}(\mathcal{G}) \\ \downarrow d_{\mathcal{G}/F} & & \downarrow d_{\mathcal{G}} \\ C^{i,j}(\mathcal{G}/F) & \xrightarrow{\alpha} & C^{i+1,j}(\mathcal{G}) \end{array}$$

commutes. Let (s, ℓ) be a labeled state of \mathcal{G}/F . Then

$$\begin{aligned} d_{\mathcal{G}}(\alpha((s, \ell))) &= d_{\mathcal{G}}((s \cup \{F\}, \tilde{\ell})) \\ &= \sum_{F_i \in E(\mathcal{G}) \setminus (s \cup \{F\})} (-1)^{n_{\mathcal{G}}(F_i)} (s \cup \{F, F_i\}, (\tilde{\ell})_{F_i}) \end{aligned}$$

where $n_{\mathcal{G}}(e_i)$ denotes the number of edges in $s \cup \{F\}$ that precede F_i in the ordering of edges of \mathcal{G} and $(\tilde{\ell})_{F_i}$ is the coloring built from $\tilde{\ell}$ as in the definition of the differential. We also compute $\alpha \circ d_{\mathcal{G}/F}$:

$$\begin{aligned} \alpha(d_{\mathcal{G}/F}(s, \ell)) &= \alpha\left(\sum_{F_i \in E(\mathcal{G}/F) \setminus s} (-1)^{n_{\mathcal{G}/F}(F_i)} (s \cup \{F_i\}, \ell_{F_i})\right) \\ &= \sum_{F_i \in E(\mathcal{G}/F) \setminus s} (-1)^{n_{\mathcal{G}/F}(F_i)} (s \cup \{F_i, F\}, \tilde{\ell}_{F_i}) \end{aligned}$$

where $n_{\mathcal{G}/F}(F_i)$ is the number of edges in s that precede F_i in the ordering of edges in \mathcal{G}/F induced by the ordering on $E(\mathcal{G})$. Since $E(\mathcal{G}) \setminus (s \cup \{F\}) = E(\mathcal{G}/F) \setminus s$, the summations in $d_{\mathcal{G}} \circ \alpha$ and $\alpha \circ d_{\mathcal{G}/F}$ are taken over the same set of edges. The labelings $(\tilde{\ell})_{F_i}$ and $\tilde{\ell}_{F_i}$ are easily seen to be the same. Lastly, $n_{\mathcal{G}}(F_i) = n_{\mathcal{G}/F}(F_i)$ since F came last in the ordering of \mathcal{G} . Since $d_{\mathcal{G}} \circ \alpha = \alpha \circ d_{\mathcal{G}/F}$, α is a chain map as desired. Next, we show that β is a chain map, i.e. that the following square

$$\begin{array}{ccc} C^{i,j}(\mathcal{G}) & \xrightarrow{\beta} & C^{i,j}(\mathcal{G} \setminus F) \\ \downarrow d_{\mathcal{G}} & & \downarrow d_{\mathcal{G} \setminus F} \\ C^{i+1,j}(\mathcal{G}) & \xrightarrow{\beta} & C^{i+1,j}(\mathcal{G} \setminus F) \end{array}$$

commutes. Let (s, ℓ) be an enhanced state of \mathcal{G} . There are two cases, the first when $F \in s$

and the second when $F \notin s$. If $F \in s$, then $\beta((s, \ell)) = 0$ by definition which implies that $d_{\mathcal{G} \setminus F}(\beta((s, \ell))) = 0$. Similarly,

$$\beta(d_{\mathcal{G}}((s, \ell))) = \beta\left(\sum_{F_i \in E(\mathcal{G}) \setminus s} (-1)^{n_{\mathcal{G}}(F_i)}(s \cup \{F_i\}, \ell_{F_i})\right) = 0$$

since each state in the sum contains s and hence F . So in this case $d_{\mathcal{G} \setminus F} \circ \beta = \beta \circ d_{\mathcal{G}}$. If $F \notin s$, then

$$d_{\mathcal{G} \setminus F}(\beta((s, \ell))) = d_{\mathcal{G} \setminus F}((s, \ell)) = \sum_{F_i \in E(\mathcal{G} \setminus F) \setminus s} (-1)^{n_{\mathcal{G} \setminus F}(F_i)}(s \cup \{F_i\}, \ell_{F_i}).$$

We also compute $\beta \circ d_{\mathcal{G}}$ and see

$$\beta(d_{\mathcal{G}}((s, \ell))) = \beta\left(\sum_{F_i \in E(\mathcal{G}) \setminus s} (-1)^{n_{\mathcal{G}}(F_i)}(s \cup \{F_i\}, \ell_{F_i})\right)$$

by definition. But β kills off the summand where $F_i = F$ and is the identity on the rest so we can rewrite the sum as

$$\beta(d_{\mathcal{G}}((s, \ell))) = \sum_{F_i \in E(\mathcal{G}) \setminus (s \cup F)} (-1)^{n_{\mathcal{G}}(F_i)}(s \cup \{F_i\}, \ell_{F_i}).$$

Since $n_{\mathcal{G} \setminus F}(F_i) = n_{\mathcal{G}}(F_i)$, $\beta \circ d_{\mathcal{G}} = d_{\mathcal{G} \setminus F} \circ \beta$ in this case also, β is a chain map.

Next we show that the sequence is short exact. α is injective due to the correspondence between components of $[\mathcal{G}/F : s]$ and $[\mathcal{G} : \tilde{s}]$. Since $\tilde{s} = s \cup \{F\}$, $\text{im}(\alpha) \subseteq \ker(\beta)$. Let $(s, \ell) \in \ker(\beta)$, then $F \in s$. Then α sends a labeled state from \mathcal{G}/F with hyperedges $s \setminus F$ to a labeled state with hyperedges s , so it just remains to find a labeling for the components of $[\mathcal{G}/F : s \setminus F]$ that will map to ℓ . We just note again that the components in $[\mathcal{G}/F : s \setminus F]$ will be precisely the components of $[\mathcal{G}/F : s]$ except for the component of \mathcal{G}/F containing the vertex v_F which is replaced by F in \mathcal{G} . Hence we can label the corresponding components of $[\mathcal{G}/F : s \setminus F]$ in the same way that ℓ labeled them in \mathcal{G} . Thus, $\ker(\beta) \subseteq \text{im}(\alpha)$ and so $\text{im}(\alpha) = \ker(\beta)$. Since β is a projection map, it is surjective. The sequence is therefore short exact as desired. \square

The snake lemma then yields a long exact sequence in cohomology which is recorded in Theorem 10, analogous to Theorem 3.2 in [60].

Theorem 10. *Let \mathcal{G} be a hypergraph and let F be a hyperedge of \mathcal{G} . For each j there is a long exact sequence*

$$0 \rightarrow H^{0,j}(\mathcal{G}) \xrightarrow{\beta^*} H^{0,j}(\mathcal{G} \setminus F) \xrightarrow{\gamma^*} H^{0,j}(\mathcal{G}/F) \xrightarrow{\alpha^*} H^{1,j}(\mathcal{G}) \xrightarrow{\beta^*} H^{1,j}(\mathcal{G} \setminus F) \rightarrow \dots \quad (1.7)$$

and taking a direct sum over j yields a long exact sequence

$$0 \rightarrow H^0(\mathcal{G}) \xrightarrow{\beta^*} H^0(\mathcal{G} \setminus F) \xrightarrow{\gamma^*} H^0(\mathcal{G}/F) \xrightarrow{\alpha^*} H^1(\mathcal{G}) \xrightarrow{\beta^*} H^1(\mathcal{G} \setminus F) \rightarrow \dots \quad (1.8)$$

Taking the graded Euler characteristic of this long exact sequence recovers the deletion-contraction recurrence for the chromatic polynomial of hypergraphs.

If a graph contains a loop, then its cohomology groups are trivial as shown in Proposition 3.4 of [60]. The proof uses the long-exact sequence categorifying the deletion-contraction recurrence for graphs and generalizes directly to hypergraphs.

Proposition 7. *Let \mathcal{G} be a hypergraph that contains a loop i.e. a hyperedge of size 1, then its cohomology groups are trivial.*

It then follows from the deletion-contraction long exact sequence and Proposition 1.2.2 that repeated hyperedges do not affect the chromatic hypergraph homology. This directly generalizes Proposition 3.5 [60] for graphs. The proof is the same as in the graph case.

Proposition 8. *Let \mathcal{G} be a hypergraph with repeated hyperedges, the chromatic hypergraph homology is preserved if all repeated hyperedges are replaced by a single edge.*

Chromatic graph homology was originally introduced with $\mathcal{A}_2 = \mathbb{Z}[x]/(x^2)$, but was generalized in [62] and shown to work with a more general class of algebras. It was shown in [93] that there are cochromatic graphs (graphs with the same chromatic polynomial) such that the chromatic graph homology of these graphs with $\mathcal{A}_3 = \mathbb{Z}[x]/(x^3)$ is not equal. More examples of cochromatic graphs that are distinguished by their chromatic graph homology with \mathcal{A}_3 were found in [102]. It was not known whether such graphs existed over \mathcal{A}_2 until Lowrance and Sazdanović showed that chromatic graph homology with \mathcal{A}_2 is entirely determined by the chromatic polynomial of a graph [80]. That is, if $\chi_{G_1}(t) = \chi_{G_2}(t)$ then $H^i(G_1) \cong H^i(G_2)$ for all i where homology is taken over \mathbb{Z} with \mathcal{A}_2 . In [62], Helme-Guizon and Rong also introduced a “twisted” version of their cohomology theory. Let $\mathcal{A} = \bigoplus_{i=0}^{\infty} A_i$ be a commutative graded \mathbb{Z} -algebra with 1 such that each A_i is free of finite rank. Let $f : \mathcal{A} \rightarrow \mathcal{A}$ be a degree preserving algebra homomorphism. We slightly redefine the differential for



Figure 1.4: θ_n is an n uniform hypergraph with n vertices and 1 edge.

the chain complex from the perspective of the cube construction. Let ξ be an edge from the cube construction. Let α_1 be its tail and α_2 be its head with e the edge of G such that $s_2 = s_1 \cup \{e\}$. Then if e joins a component of G to itself, the components of $[G : s_1]$ and $[G : s_2]$ are in bijection in the obvious way. We define $d_\xi = f^\otimes : A^{\otimes k_1} \rightarrow A^{\otimes k_2}$ where k_1 is the number of components of $[G : s_1]$ and k_2 is the number of components of $[G : s_2]$, clearly $k_1 = k_2$. If f is the identity, then this gives the original cohomology. If f is not the identity, then the homology groups can differ. A twisted version of hypergraph cohomology also exists and is defined in the obvious way.

Proposition 9. *Chromatic hypergraph homology with A_m ($m \neq 2$) is stronger than the chromatic polynomial at distinguishing hypergraphs that are not graphs.*

Proof. It is known that chromatic graph homology over A_m can distinguish cochromatic graphs G_1 and G_2 [93, 102] Construct hypergraphs \mathcal{G}_1 and \mathcal{G}_2 by adding in a hyperedge to the corresponding graphs that contains all the vertices in each graph. Since chromatic hypergraph homology is determined by the minimal hyperedges by inclusion, the homology of \mathcal{G}_1 is equal to the homology of G_1 and similarly for \mathcal{G}_2 and G_2 . Therefore, the hypergraph homology distinguishes \mathcal{G}_1 from \mathcal{G}_2 , but $\chi_{G_1}(t) = \chi_{\mathcal{G}_1}(t) = \chi_{G_2}(t) = \chi_{\mathcal{G}_2}(t)$. \square

Given a graph G with n vertices, chromatic graph homology with A_2 is known to lie along the two diagonals $i + j = n$ and $i + j = n - 1$ [61]. The following example shows that, unlike for chromatic graph homology, chromatic hypergraph homology is not thin even over A_2 .

Example 3. *Let θ_n denote the n -uniform hypergraph with one hyperedge for $n > 1$ pictured in Figure 1.4, then with \mathcal{A}_2*

$$H^i(\theta_n) = \begin{cases} \mathbb{Z}^{n-1}\{1\} \oplus \bigoplus_{j=2}^n \mathbb{Z}^{(j)}\{j\} & i = 0 \\ 0 & i > 0. \end{cases}$$

To see this, we construct a matrix representation of the differential $d^0 : \mathcal{A}_2^{\otimes n} \rightarrow \mathcal{A}_2$ when

j			
2	\mathbb{Z}		
1	\mathbb{Z}		
0			
	0	1	i

Table 1.1: The chromatic graph homology of θ_2 with \mathcal{A}_2 .

$n = 3$:

$$d^0 = \begin{matrix} & 111 & 11x & 1x1 & 1xx & x11 & x1x & xx1 & xxx \\ \begin{matrix} 1 \\ x \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

and notice that d^0 applied to $(a_1, \dots, a_8)^T$ is

$$d^0((a_1, \dots, a_8)^T) = (a_1, a_2 + a_3 + a_5)^T.$$

This shows that elements of $\ker(d^0)$ are vectors $(0, a_2, \dots, a_8)^T$ where $a_2 + a_3 + a_5 = 0$. For arbitrary n , elements of $\ker(d^0)$ look like $(0, a_2, \dots, a_{2n})^T$ where the sum of all n coordinates corresponding to degree 1 elements is 0. This space has basis elements e_k for k an index corresponding to an element of degree greater than 1 and $(0, 1, 0, \dots, -1, 0, \dots, 0)$ with -1 at the $n-1$ coordinates corresponding to degree 1 elements. In other words, the $\ker(d^0)$ is generated by $n-1$ degree 1 elements and $\binom{n}{j}$ degree j elements for $2 \leq j \leq n$.

Note that when $n = 2$ this yields the homology in Table 1.1 and θ_2 is just a tree with one edge. Let T_n be a tree with n edges. In Example 33 of Helme-Guizon's Dissertation, it is shown that with \mathcal{A}_2 , $H^0(T_n) \cong \mathbb{Z}\{n\} \oplus \mathbb{Z}\{n+1\}$ and $H^i(T_n) = 0$ for $i \neq 0$. This formula applied to T_1 agrees with the formula in Example 1.2.2 for θ_2 .

The previous example shows that unlike the connected graph case for a graph with n vertices, where $H^{i,j}(G)$ is nontrivial only for $n-1 \leq i+j \leq n$, connected hypergraphs can have nontrivial homology outside of these two diagonals.

Theorem 11. *Chromatic hypergraph homology is not homologically thin over A_2 .*

Proof. The hypergraph θ_3 is a connected hypergraph with 3 vertices. Example 3 shows that $H^0(\mathcal{G})$ is nontrivial along the three diagonals $i+j=1$, $i+j=2$ and $i+j=3$ and is therefore not thin. □

In particular, θ_n shows that the lower bound on $i + j$ for nontrivial homology does not hold. Since the homology of θ_n is nontrivial when $i = 0$ and $j = 1, \dots, n-2$. This computation is generalized to A_k in the following example.

Example 4. Let $n > 1$ and consider the hypergraph θ_n and the algebra A_k . Let $\kappa(a, b, c)$ be the number of weak compositions of a into b parts with each part less than c . By a similar argument to Example 3 about the structure of the matrix representation of d^0

$$H^i(\theta_n) = \begin{cases} \bigoplus_{j=1}^{k-1} \mathbb{Z}^{\kappa(j,n,k)}\{j\} \oplus \bigoplus_{m \geq k}^{n(k-1)} \mathbb{Z}^{\kappa(m,n,k)}\{m\} & i = 0 \\ 0 & i > 0. \end{cases}$$

Example 4 shows that the homology of θ_n over A_k lies on $n(k-1)$ diagonals.

Proposition 10. The homology of θ_n for $n \geq 3$ over A_k for $k \geq 2$ is not thin and lies on $n(k-1)$ diagonals.

1.3 Hypergraph Configuration Spaces

Eastwood and Huggett define graph configuration spaces and consider these spaces for graphs G , $G \setminus e$, G/e and a manifold M [48]. They then show that these spaces fit into an exact sequence so that their Euler characteristics are equal to the evaluation of the chromatic polynomial of G at $\chi(M)$, thus categorifying the deletion-contraction recurrence for the chromatic polynomial of graphs. Cooper-de Silva-Sazdanović generalize this construction to a simplicial complex S in [35], recovering the deletion-contraction formula for a graph in the case that $S = \Delta(G)$, the independence complex of the graph. Here, we generalize the construction to hypergraphs \mathcal{G} , recovering the Eastwood and Huggett construction when \mathcal{G} is a graph and the Cooper-de Silva-Sazdanović construction when S is the independence complex of the hypergraph \mathcal{G} .

Let M be an even dimensional orientable manifold of dimension m and let $\mathcal{G} = (V, E)$ be a hypergraph with n vertices. For each hyperedge $e = v_{i_1} v_{i_2} \cdots v_{i_k}$, we define the diagonal corresponding to e by

$$\Delta_e = \{(x_1, \dots, x_n) : x_{i_1} = x_{i_2} = \dots = x_{i_k}\}.$$

The hypergraph configuration space for \mathcal{G} is then defined as

$$M_{\mathcal{G}} = M^{\times n} \setminus \bigcup_{e \in E} \Delta_e.$$

Remark 3. Let \mathcal{G} be a hypergraph and let \mathcal{G}' be the subhypergraph of \mathcal{G} with only the hyperedges of \mathcal{G} that are minimal under inclusion. If \mathcal{G} contains hyperedges $f, f' \in E(\mathcal{G})$ where $f \subset f'$ then $\Delta_f \supset \Delta_{f'}$. This implies that $M_{\mathcal{G}} = M_{\mathcal{G}'}$.

The deletion-contraction recurrence can then be categorified through the Leray long exact sequence as in Section 1.1.3 for graphs and Section 1.1.4 for simplicial complexes. Fix a hyperedge f in E and define X and Y as follows

$$X = M^{\times n} \setminus \bigcup_{\substack{e \in E \\ e \neq f}} \Delta_e \text{ and } Y = X \cap \Delta_f.$$

This follows the setup in [48] except now f is a hyperedge. Next they assert that the Leray sequence relates $M_{\mathcal{G}}$, $M_{\mathcal{G}/f}$ and $M_{\mathcal{G} \setminus f}$ which implies that $X - Y = M_{\mathcal{G}}$ and $Y \cong M_{\mathcal{G}/f}$. The assertion that $X - Y = M_{\mathcal{G}}$ is clear, but the second assertion requires proof.

Lemma 3. Let $\mathcal{G} = (V, E)$ be a hypergraph. If $Y = \left(M^{\times n} \setminus \bigcup_{\substack{e \in E \\ e \neq f}} \Delta_e \right) \cap \Delta_f$, then $Y \cong M_{\mathcal{G}/f}$.

Proof. If $f \supset f'$ for $f' \in E$, then $Y = X \cap \Delta_f = \emptyset$. But \mathcal{G}/f contains a loop, so $M_{\mathcal{G}/f} = \emptyset$. Otherwise, f does not contain f' for any $f' \in E$. Let f be a hyperedge of \mathcal{G} containing k vertices. Label the vertices of this hyperedge so that they correspond to the first k copies of M in $M^{\times n}$. Then define $h : Y \rightarrow M_{\mathcal{G}/f}$ by $h(x, x, \dots, x, x_1, \dots, x_{n-k}) = (x, x_1, \dots, x_{n-k})$ where $M_{\mathcal{G}/f}$ has vertices labeled so that the first component corresponds to the vertex v that all of the vertices of f were identified to and the rest of the vertices are in the same order. We first need to check that this map actually maps into $M_{\mathcal{G}/f}$ as claimed. If it didn't map into $M_{\mathcal{G}/f}$, then there would be some $x_{i_1} = x_{i_2} = \dots = x_{i_j}$ such that the vertices v_{i_1}, \dots, v_{i_j} are in a hyperedge e of \mathcal{G}/f . If e does not contain v , then it is a hyperedge in $\mathcal{G} \setminus f$ and hence the corresponding coordinates could not have been equal. If one of the vertices is equal to v , then this edge corresponds to a hyperedge in $M_{\mathcal{G}/f}$ where v is replaced by v_1, \dots, v_k . These must all be equal as $M_{\mathcal{G}/f}$ has been intersected with Δ_f and therefore all coordinates corresponding to this edge must be equal, which is not possible. So h maps into $M_{\mathcal{G}/f}$. It is clear that h is injective. The argument for surjectivity is a reversal of the argument that h maps into $M_{\mathcal{G}/f}$ and h is continuous because it is a projection map. \square

Theorem 12. *Let $\mathcal{G} = (V, E)$ be a hypergraph. Let M be an even dimensional orientable manifold of dimension m and let $e = v_{i_1} \cdots v_{i_k}$ be a hyperedge of \mathcal{G} . Then there is a long exact sequence*

$$0 \rightarrow H_{mn}(M_{\mathcal{G}}) \rightarrow H_{mn}(M_{\mathcal{G} \setminus f}) \rightarrow H_{mn-m(k-1)}(M_{\mathcal{G}/f}) \rightarrow H_{mn-1}(M_{\mathcal{G}}) \rightarrow \dots \quad (1.9)$$

Proof. Choose X and Y as above and apply Lemma 3 and Equation (1.3). \square

Applying the Euler characteristic to this sequence yields a deletion-contraction relation.

Corollary 2. *Let $\mathcal{G} = (V, E)$ be a hypergraph. Let M be an even dimensional orientable manifold of dimension m and let $e = v_{i_1} \cdots v_{i_k}$ be a hyperedge of \mathcal{G} , then*

$$\chi(M_{\mathcal{G}}) - \chi(M_{\mathcal{G} \setminus f}) + \chi(M_{\mathcal{G}/f}) = 0.$$

From the multiplicativity of the Euler characteristic it follows that $\chi(M_{\mathcal{G}})$ is equal to $\chi_{\mathcal{G}}(t)$ evaluated at $t = \chi(M)$.

Theorem 13. *Let M be an orientable even dimensional manifold and \mathcal{G} be a hypergraph. Then $\chi(M_{\mathcal{G}}) = \chi_{\mathcal{G}}(t)$ when $t = \chi(M)$.*

This recovers graph configuration spaces and their categorification of the chromatic polynomial of a graph as a special case.

Theorem 14. *The categorification of the chromatic polynomial of a hypergraph via hypergraph configuration spaces generalizes the categorification of the chromatic polynomial of a graph via graph configuration spaces.*

If $M = \mathbb{C}\mathbb{P}^{t-1}$ and a hypergraph \mathcal{G} consists of a single vertex with no hyperedges then $\chi(M_{\mathcal{G}})$ is the number of ways to color the vertices of \mathcal{G} with t colors.

Theorem 15. *Let \mathcal{G} be a hypergraph and let $M = \mathbb{C}\mathbb{P}^{t-1}$. Let $M_{\mathcal{G}}$ be the hypergraph configuration space as defined previously, then $\chi_{\mathcal{G}}(t) = \chi(M_{\mathcal{G}})$.*

Proof. Let \mathcal{G}' be the hypergraph on n vertices with no edges. Since the Euler characteristic is multiplicative, $\chi_{M_{\mathcal{G}'}} = t^n$. Therefore, since $\chi(M_{\mathcal{G}})$ satisfies the same initial condition and recurrence relation as the chromatic polynomial of a hypergraph, we have $\chi_{\mathcal{G}}(t) = \chi(M_{\mathcal{G}})$. \square

Just as in the simplicial configuration space case, hypergraph configuration spaces can be defined in terms of equality complexes.

Definition 37. *Let \mathcal{G} be a hypergraph on vertex set V with n vertices. Then the **hypergraph configuration space** of \mathcal{G} is $M_{\mathcal{G}} = \{x \in M^n : \text{Eq}(f_x) \text{ is a subcomplex of } \Delta(\mathcal{G})\}$.*

1.4 Comparison with the Simplicial Chromatic Constructions

A geometric construction which generalizes [48] from graphs to simplicial complexes was introduced in [35]. Decategorification of this homology theory led to a new invariant of simplicial complexes called the simplicial chromatic polynomial. In [37], Cooper-Sazdanović introduced a state-sum homology theory analogous to [60] which categorified the simplicial chromatic polynomial. Here we explore the relationship between hypergraph chromatic homology and the hypergraph Eastwood-Huggett construction with their simplicial counterparts.

Recall the definition of a clutter from Definition 1.1.1. There is a bijection φ between simplicial complexes and clutters defined in the following way. Given a simplicial complex S , $\varphi(S)$ is \mathcal{G}_S , its hypergraph of minimal nonfaces. φ^{-1} sends a clutter \mathcal{G} to $\Delta(\mathcal{G})$. The relationship between hypergraph chromatic homology and simplicial chromatic homology can be viewed as a lifting of this bijection.

Comparing the state-sum construction for the simplicial chromatic polynomial in [37] and the state-sum construction for the hypergraph chromatic homology shows that the chain complexes are the same for S and a hypergraph on the same vertex set as S with hyperedges equal to the nonfaces of S . The constructions are identical except that in the simplicial chromatic case, the states are sets of nonfaces and in the hypergraph construction the states are sets of hyperedges. We record this relationship in the following theorem.

Theorem 16. *Let S be a simplicial complex, then $(C^{i,j}(S), d) \cong (C^{i,j}(\mathcal{G}), d)$ where \mathcal{G} has the same vertex set as S and hyperedges equal to the nonfaces of S .*

Corollary 3. *Let S be a simplicial complex, then $H^*(S) \cong H^*(\mathcal{G})$ where \mathcal{G} has the same vertex set as S and hyperedges equal to the nonfaces of S .*

Denote by $H_{\mathcal{A}}^{mnf}(S)$ the homology of the subcomplex of $C^i(S)$ built only from states with minimal nonfaces. Then Theorem 5.5 of [37] shows that given a simplicial complex S the simplicial state-sum homology $H_{\mathcal{A}}(S) \cong H_{\mathcal{A}}^{mnf}(S)$.

Theorem 17. *Let S be a simplicial complex, then $H_{\mathcal{A}}(\mathcal{G}_S) \cong H_{\mathcal{A}}^{mnf}(S) \cong H_{\mathcal{A}}(S)$.*

Proof. Something slightly stronger is true, their associated chain complexes are equal. States for simplicial complexes consist of minimal nonfaces and states for hypergraphs consist of hyperedges. Since $\mathcal{G}_{\mathcal{A}}$ has hyperedges equal to the minimal nonfaces of S , the result follows. \square

Simplicial complexes are a special case of hypergraphs. Specifically, they are hypergraphs $\mathcal{G} = (V, E)$ such that if a hyperedge $e \in E$ then $e' \in E$ for all $e' \subseteq e$. This could suggest that hypergraph chromatic homology contains more information than simplicial chromatic homology. Theorem 17 shows this is not the case since it implies that many hypergraphs have the same chromatic hypergraph homology. Consider a hypergraph and identify its smallest hyperedges with respect to inclusion. These uniquely determine a simplicial complex because the only restriction on minimal nonfaces in a simplicial complex is that they do not contain other minimal nonfaces. All of the remaining hyperedges of the hypergraph correspond to nonfaces in the simplicial complex that are not minimal and therefore do not contribute to homology.

Theorem 18. *Let \mathcal{G} and \mathcal{G}' be hypergraphs such that their minimal hyperedges with respect to inclusion are the same, then $H_{\mathcal{A}}(\mathcal{G}) \cong H_{\mathcal{A}}(\mathcal{G}')$.*

Remark 4. *Theorem 18 mirrors the situation for chromatic polynomials of hypergraphs. The chromatic polynomial of a hypergraph \mathcal{G} is determined by its minimal hyperedges with respect to inclusion. Consider a hypergraph \mathcal{G} that contains two hyperedges e and e' such that $e \subseteq e'$. A proper coloring of \mathcal{G} must not make e monochromatic. Since e' contains e , this proper coloring automatically makes e' not monochromatic. This means e' is not adding any additional restrictions on a proper coloring and can be ignored for the sake of computing the chromatic polynomial of \mathcal{G} . Hyperedges like e' are sometimes referred to as **chromatically inactive** [117].*

Next we consider the Eastwood-Huggett style constructions for simplicial complexes [35] and for hypergraphs.

Definition 38. Let S be a simplicial complex, then the **simplicial configuration space** M_S is defined to be

$$M_S = M^n \setminus \bigcup_{\sigma \in \text{NF}(S)} D_\sigma = M^n \setminus \bigcup_{\sigma \in \text{MNF}(S)} D_\sigma.$$

In other words, the diagonals corresponding to minimal nonfaces of S are deleted. The hypergraph configuration space was defined so that the diagonals corresponding to the hyperedges are deleted. This yields the following proposition which follows directly from the definitions.

Proposition 11. Let S be a simplicial complex and let \mathcal{G} be its hypergraph of nonfaces, then $M_S \cong M_{\mathcal{G}}$ where M_S is the simplicial configuration space and $M_{\mathcal{G}}$ is the hypergraph configuration space.

Proposition 12. Let S be a simplicial complex and let \mathcal{G}_S be its hypergraph of nonfaces, then $\chi(M_{\mathcal{G}}) = \chi_{\mathcal{G}}(t)$.

Proof. By Proposition 11, $M_S \cong M_{\mathcal{G}}$ and therefore $\chi(M_{\mathcal{G}}) = \chi(M_S) = \chi(S, t)$. We know that $\chi(S, t) = \chi_{\mathcal{G}_S}(t)$ where \mathcal{G}_S is the hypergraph of minimal nonfaces of S . Since the hyperedges of \mathcal{G} are the nonfaces of S , the minimal hyperedges of \mathcal{G} by inclusion must be the minimal nonfaces of S . Therefore, by Theorem 18 $H_{\mathcal{A}}(\mathcal{G}) \cong H_{\mathcal{A}}(\mathcal{G}_S)$ and hence $\chi_{\mathcal{G}_S}(t) = \chi_{\mathcal{G}}(t)$. \square

Theorem 19. Let \mathcal{G} be a simplicial complex, let \mathcal{G}_S be its hypergraph of minimal nonfaces and let $M = \mathbb{C}\mathbb{P}^{t-1}$, then $\chi(M_S) = \chi(M_{\mathcal{G}_S}) = \chi_{\mathcal{G}_S}(t)$.

Proof. This follows from the fact that $\chi_s(S, t) = \chi_{\mathcal{G}_S}(t)$. \square

We now consider an example of a hypergraph configuration space and its associated simplicial configuration space.

Example 5. Consider the hypergraph configuration space from the simplest hypergraph which is not a graph. Let $\mathcal{G} = ([3], E)$ where E consists of the hyperedge 123 and $M = [0, 1]$. Then $M_{\mathcal{G}}, M_{\mathcal{G} \setminus e}$ and $M_{\mathcal{G}/e}$ are depicted in Figure 1.5. The red line in Figure 1.5 represents what has been deleted from the cube. In other words, $M_{\mathcal{G}}$ is the complement of this line within the unit cube. In this case, $\Delta(\mathcal{G})$ is $S = \{12, 13, 23\}$, the boundary of the 2-simplex and its simplicial configuration space is also pictured.

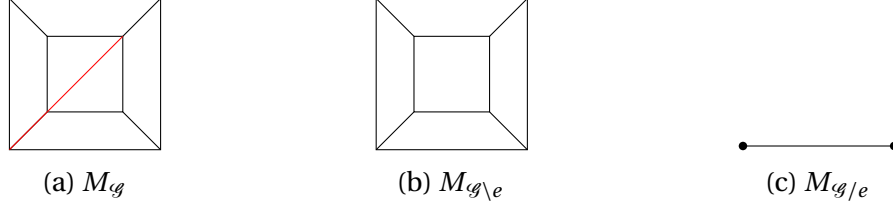


Figure 1.5: Hypergraph configuration spaces for \mathcal{G} , $\mathcal{G} \setminus e$ and \mathcal{G}/e where \mathcal{G} is the hypergraph on three vertices with hyperedge 123 and $M = [0, 1]$.

The new interpretation for the simplicial chromatic polynomial in terms of hypergraphs allows for simplified and purely combinatorial proofs of previously established results in [35].

Proposition 13 (Corollary 6.6 in [35]). *Let S_1 and S_2 be simplicial complexes, then $\chi_c(S_1 * S_2, t) = \chi_c(S_1, t) \cdot \chi_c(S_2, t)$.*

Proof. Let $\text{MNF}(S_1)$ and $\text{MNF}(S_2)$ be the sets of minimal nonfaces of S_1 and S_2 respectively. It is easy to see that $\text{MNF}(S_1 * S_2) = \text{MNF}(S_1) \cup \text{MNF}(S_2)$. Therefore, $\mathcal{G}_{S_1 * S_2} = \mathcal{G}_{S_1} \sqcup \mathcal{G}_{S_2}$ so $\chi_{\mathcal{G}_{S_1 * S_2}}(t) = \chi_{\mathcal{G}_{S_1}}(t) \cdot \chi_{\mathcal{G}_{S_2}}(t)$ and the result follows. \square

Proposition 14 (Corollary 6.7 in [35]). *Let S be a simplicial complex, then $\chi_c(S \sqcup \{pt\}, t) = t \cdot \chi_c(S, t - 1)$.*

Proof. The minimal nonfaces of $S \sqcup \{pt\}$ are the minimal nonfaces of S as well as $\{\{pt\}, x\}$ for every vertex x of S . This means the hyperedges in $\mathcal{G}_{S \sqcup \{pt\}}$ are the hyperedges in \mathcal{G}_S as well as a hyperedge between every vertex of \mathcal{H}_S and the new vertex $\{pt\}$. All colorings of $\mathcal{G}_{S \sqcup \{pt\}}$ with t colors can be constructed by coloring $\{pt\}$ of which there are t ways to do so and then coloring the rest of the vertices with only $t - 1$ colors since $\{pt\}$ is connected to every other vertex by a hyperedge of which there are $\chi_{\mathcal{G}_S}(t - 1)$ ways to do so. The result follows. \square

Since a simplicial complex is a hypergraph satisfying the condition that if E is a hyperedge, then so is every subset of E , we can compute the chromatic polynomial of a hypergraph for a simplicial complex. The chromatic polynomial of a hypergraph can distinguish some simplicial complexes that the simplicial chromatic polynomial cannot as shown in Example 6.



Figure 1.6: Chromatically equivalent hypergraphs with chromatic polynomial $t^6 - t^4 - t^3 + t$

Example 6. Figure 1.6 shows two hypergraphs with the chromatic polynomial $t^6 - t^4 - t^3 + t$. These hypergraphs are defined as $\mathcal{G}_1 = ([6], E_{\mathcal{G}_1})$ and $\mathcal{G}_2 = ([6], E_{\mathcal{G}_2})$ where $E_{\mathcal{G}_1} = \{\{1, 2, 4, 5\}, \{2, 3, 6\}\}$ and $E_{\mathcal{G}_2} = \{\{1, 2, 4, 5\}, \{2, 3, 5\}, \{1, 2, 3, 4, 6\}\}$. Since the simplicial chromatic polynomial counts the number of ways to color a simplicial complex with t colors such that no minimal nonface is monochromatic, the simplicial complexes S_1 and S_2 with minimal nonfaces

$$MNF(S_1) = \{\{1, 2, 4, 5\}, \{2, 3, 6\}\}$$

$$MNF(S_2) = \{\{1, 2, 4, 5\}, \{2, 3, 5\}, \{1, 2, 3, 4, 6\}\}$$

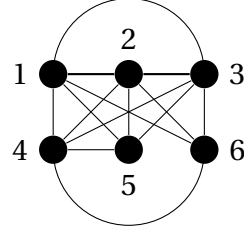
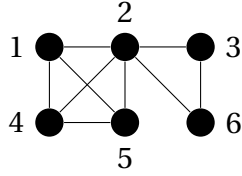
have the same simplicial chromatic polynomial: $\chi_c(S_1) = \chi_c(S_2) = t^6 - t^4 - t^3 + t$.

To compute the chromatic polynomial of S_1 and S_2 as hypergraphs we compute their facets:

$$FACETS(S_1) = \{\{2, 3\}, \{2, 6\}, \{3, 6\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 4, 5\}, \{2, 4, 5\}\}$$

$$FACETS(S_2) = \{\{1, 2, 3, 4\}, \{1, 2, 3, 6\}, \{1, 2, 4, 6\}, \{1, 3, 4, 6\}, \{2, 3, 4, 6\}, \{3, 5\}, \{1, 2, 5\}, \{2, 4, 5\}, \{1, 4, 5\}\}.$$

Note that the chromatic polynomial of a hypergraph in the special case of a simplicial complex is determined by the 1-skeleton of the simplicial complex. These graphs are pictured in Figure 1.7. The chromatic polynomial of the 1-skeleton of S_1 and hence of S_1 itself is $\chi_{S_1}(t) = t(t-1)^2(t-2)^2(t-3)$. Since K_5 is contained inside the 1-skeleton of S_2 , there are no colorings of S_2 with 4 colors. Hence, $t-4$ must be a factor of $\chi_{S_2}(t)$. Since $t-4$ is not a factor of $\chi_{S_1}(t)$, it must be that the chromatic polynomials of S_1 and S_2 are distinct.



(a) The 1-skeleton of S_1 from Example 6. (b) The 1-skeleton of S_2 from Example 6.

Figure 1.7: The 1-skeleta of the simplicial complexes with minimal nonfaces $\{\{1, 2, 4, 5\}, \{2, 3, 6\}\}$ (A) and minimal nonfaces $\{\{1, 2, 4, 5\}, \{2, 3, 5\}, \{1, 2, 3, 4, 6\}\}$ (B).

Since hypergraphs generalize simplicial complexes, the simplicial chromatic homology and the hypergraph chromatic homology can be computed for a simplicial complex. As Example 1.7 shows, in general they will not agree. But if the simplicial complex S has a specific structure, then they will.

Proposition 15. *Let S be a simplicial complex, then \mathcal{G}_S is isomorphic to the 1-skeleton of S if and only if the 1-skeleton of S is a self-complementary graph and if S contains the boundary of an $(n \geq 2)$ -simplex, then it must contain the entire simplex.*

Proof. \implies) Assume \mathcal{G}_S is isomorphic to the 1-skeleton of S . Since the minimal nonfaces of S form a graph, there cannot be any boundaries of $n \geq 2$ simplices in S . Since the minimal nonfaces of S are 1-dimensional, \mathcal{G}_S must be the complement of the 1-skeleton of S . Hence, the 1-skeleton of S is self-complementary.

\impliedby) Assume the 1-skeleton of S is a self-complementary graph and that S does not contain the boundary of an n -simplex for $n \geq 2$. Since there are no boundaries of n -simplices for $n \geq 2$, the minimal nonfaces of S are 1-dimensional. Therefore, the minimal nonfaces of S are equal to the graph complement of the 1-skeleton of S . Since the 1-skeleton of S is a self-complementary graph, the graph of minimal nonfaces of S is isomorphic to the 1-skeleton of S . \square

Given two co-simplicial chromatic simplicial complexes, the homology of their simplicial configuration spaces can distinguish them for some choices of M . This was shown in Theorem 5.2 of [35] for two specific simplicial complexes and $M = \mathbb{C}P^2$. Computing the hypergraphs of minimal nonfaces for these simplicial complexes shows that two cochromatic hypergraphs can be distinguished by their hypergraph configuration spaces for some choices of M .



(a) A hypergraph \mathcal{G}_1 that is cochromatic with \mathcal{G}_2 . (b) A hypergraph \mathcal{G}_2 that is cochromatic with \mathcal{G}_1 .

Figure 1.8: \mathcal{G}_1 and \mathcal{G}_2 are cochromatic hypergraphs with chromatic polynomials equal to $t^4 - 3t^3 + 2t^2$. These are the hypergraphs of minimal nonfaces corresponding to the simplicial complexes in Figure 1 of [35].

Proposition 16. *The hypergraphs \mathcal{G}_1 and \mathcal{G}_2 in Figure 1.8 have the same chromatic polynomial, but are distinguished by the homology of their hypergraph configuration spaces for $M = \mathbb{C}\mathbb{P}^2$.*

Proof. Follows from Theorem 5.2 in [35]. □

1.5 The Simplicial Chromatic Coloring Complex

Coloring ideals were introduced by Steingrímsson for graphs [108]. Given a graph G , a polynomial ideal is constructed so that the Hilbert polynomial of this ideal is equal to the chromatic polynomial of G . This construction was later generalized to the characteristic polynomials of hyperplane arrangements in [67] and chromatic polynomials of hypergraphs in [20]. Through the relationship between $\chi_c(S, t)$ and the chromatic polynomial of \mathcal{G}_S a coloring complex for the simplicial chromatic polynomial is constructed.

Independently, Park constructed a different coloring complex for the simplicial chromatic polynomial of a class of simplicial complexes [95]. Park finds a specific family of simplicial complexes with simplicial chromatic polynomial equal to the Hilbert series of the Stanley-Reisner rings of associated simplicial complexes. This shows that simplicial chromatic polynomials of this class of simplicial complexes can be determined by the h -vectors of their associated simplicial complexes.

There are a number of different ways to define coloring complexes, we use the approach and terminology from [72] to define a coloring complex for the simplicial chromatic polynomial.

Definition 39. *Let S be a simplicial complex on vertex set $[n]$. A subset of $[n]$ is S -stable if no nonface of S is contained in it. Define Δ_S , the **simplicial coloring complex** of S , by*

$\{X_1, X_2, \dots, X_k\}$ is in Δ_S if and only if

$$\emptyset = X_0 \subsetneq X_1 \subsetneq X_2 \subsetneq \dots \subsetneq X_k \subsetneq X_{k+1} = [n]$$

and at least one of the sets $Y_i = X_i \setminus X_{i-1}$ for $1 \leq i \leq k+1$ is not S -stable.

Following [72], we refer to $\{X_1, X_2, \dots, X_k\}$ as a **chain** and Y_1, \dots, Y_{k+1} as the **components** of the chain. We now define the accompanying ideal which will be necessary to show that Δ_S has the correct Hilbert polynomial.

Definition 40. Let S be a simplicial complex and k be a field. Define $A = k[x_T : T \subseteq [n]]$, $I = \{x_{T_1} x_{T_2} : T_1 \not\subseteq T_2, T_2 \not\subseteq T_1\}$ and $R = A/I$. The **simplicial coloring ideal** of S , K_S , is the ideal of R generated by monomials $x_{X_1}^{e_1} x_{X_2}^{e_2} \cdots x_{X_k}^{e_k}$ ($e_i > 0$) such that $\{X_1, \dots, X_k\}$ is a chain whose components are S -stable.

Example 7. Let $S = L_3$ be the path graph on 3 vertices as in Figure 1.9. The only nonface of this simplicial complex is 13. There cannot be a chain of length 2, because each component would then have cardinality 1. Therefore, Δ_S is at most 0-dimensional. Since $13 \setminus \emptyset$ contains the nonface 13 and $123 \setminus 2$ contains the nonface 13, Δ_S is the simplicial complex with two vertices labeled 2 and 13 as in 1.9b.

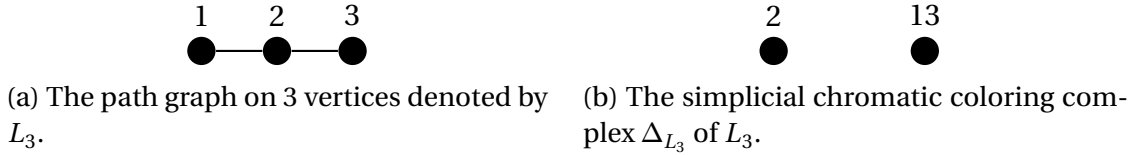


Figure 1.9: A graph L_3 and its associated simplicial chromatic coloring complex.

To compute K_S , we need to identify the chains of S whose components are S -stable. We have already identified the chains whose components are not S -stable, so the remaining chains must have S -stable components. These chains are $\{1\}$, $\{3\}$, $\{12\}$, $\{23\}$, $\{1, 12\}$ and $\{3, 23\}$ so $K_S = \langle x_1, x_3, x_{12}, x_{23}, x_1 x_{12}, x_3 x_{23} \rangle = \langle x_1, x_3, x_{12}, x_{23} \rangle$.

This definition allows an analogous argument to Steingrímsson's proof of Theorem 9 in [108] to show that K_S counts the colorings of the vertices of S such that no minimal nonface is monochromatic.

Theorem 20. *Let S be a simplicial complex on the vertex set $[n]$. The monomials of degree d in K_S are in bijection with the S -compatible colorings of S with $(d + 1)$ colors.*

Proof. Every monomial in K_S has the form $x_{T_1}^{e_1} x_{T_2}^{e_2} \cdots x_{T_k}^{e_k}$ for a chain $T_1 \subsetneq T_2 \subsetneq \cdots \subsetneq T_k \subseteq [n]$ where $T_i \setminus T_{i-1}$ is an S -stable set. Given a monomial $M \in K_S$ of degree d , we will construct an S -compatible coloring of S . Since each component $T_i \setminus T_{i-1}$ is an S -stable set, associating a color to each of these will yield an S -compatible coloring of S . This is done in the following way: the vertices in $T_j \setminus T_{j-1}$ are assigned the color $(\sum_{i=1}^{j-1} e_i) + 1$ for $j \geq 2$. If $T_1 \neq \emptyset$, then the vertices in T_1 are assigned the color 1. If $[n] \setminus T_k \neq \emptyset$ then these vertices are assigned the color $(\sum_i e_i) + 1$. To show that this map is a bijection, we construct the inverse map. Consider an S -compatible coloring of the vertices of S with $d + 1$ colors (not all of which need be used). Group the vertices into sets $C_{i_1}, C_{i_2}, \dots, C_{i_k}$ where the vertices in C_{i_j} have color i_j and $i_1 < i_2 < \dots < i_k$. That is, the vertices are grouped into sets according to color and these sets are then ordered by color. Define T_j to be the union of the first j of these. If $i_1 > 1$, then send the coloring to the monomial:

$$x_{\emptyset}^{i_1-1} x_{T_1}^{i_2-i_1} x_{T_2}^{i_3-i_2} \cdots x_{T_{k-1}}^{i_k-i_{k-1}} x_{T_k}^{d-i_k}$$

and otherwise, if $i_1 = 1$, send the coloring to the monomial:

$$x_{T_1} x_{T_2}^{i_2-i_1} x_{T_3}^{i_3-i_2} \cdots x_{T_{k-1}}^{i_{k-1}-i_{k-2}} x_{T_k}^{d-i_{k-1}}.$$

This is the inverse map by construction and the result follows. \square

Corollary 4. *Let S be a simplicial complex on the vertex set $[n]$, then the Hilbert polynomial $H(K_S, t) = \chi_c(S, t + 1)$.*

Proof. The Hilbert polynomial of a monomial ideal is, by definition, the unique polynomial which is equal to the Hilbert function of the ideal for sufficiently large t . Since the number of monomials of degree t in K_S is in bijection with the S -compatible colorings of S with $t + 1$ colors, $\chi_c(S, t + 1)$ must be the Hilbert polynomial. \square

To demonstrate the bijection in Theorem 20, we return to the coloring complex from Example 7.

Example 8. *In Example 7, we showed that for $S = L_3$, $K_S = \langle x_1, x_3, x_{12}, x_{23} \rangle$. The monomials of degree 1 are the 4 generators of K_S . By Theorem 20, we would expect there to be 4 S -compatible*

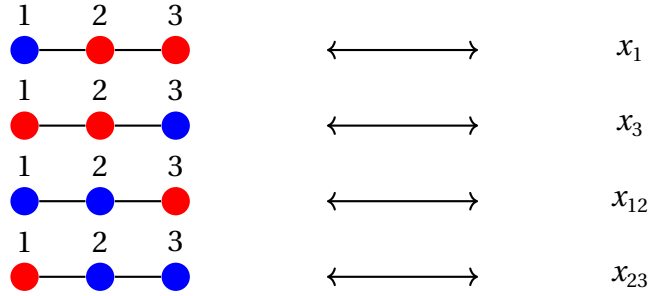


Figure 1.10: Bijection between S -compatible colorings with 2 colors and degree 1 monomials in K_{L_3}

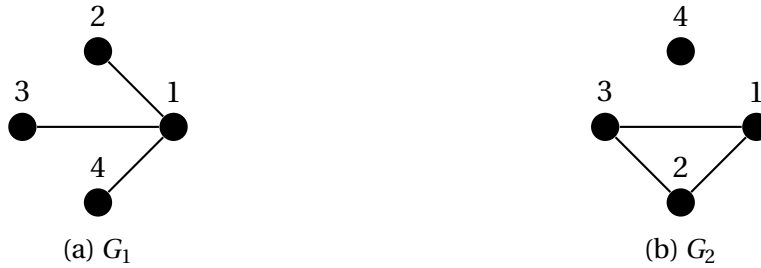


Figure 1.11: Graphs G_1 and G_2 such that $\chi_c(G_1, t) = \chi_c(G_2, t)$

colorings using 2 colors. The only nonface of S is 13, so the only restriction on a coloring of the vertices of S is that 1 and 3 are colored differently. There are two choices to color vertex 1 which fixes the color of vertex 3. There are then 2 choices left for vertex 2 for a total of 4 colorings as expected. Let color 1 be blue and color 2 be red, then the bijection as described in Theorem 20 is pictured in Figure 1.10.

Since the simplicial coloring complex can be viewed as a categorification of the simplicial chromatic polynomial, it can distinguish at least as many simplicial complexes as the simplicial chromatic polynomial. This is shown in the following example.

Example 9. In Theorem 5.2 of [35], Cooper-de Silva-Sazdanović show that the graphs G_1 and G_2 as in Figure 1.11 are distinguished by their simplicial chromatic cohomology theory with $M = \mathbb{C}\mathbb{P}^2$, even though $\chi_c(G_1, t) = \chi_c(G_2, t) = t^4 - 3t^2 + 2t^2$. The minimal nonfaces of G_1 are 23, 24 and 34. Since they are all of the same dimension, Δ_{G_1} is pure. The minimal nonfaces of G_2 are 14, 24, 34 and 123. Since these minimal nonfaces are not of the same dimension, Δ_{G_2} is not pure, so $\Delta_{G_1} \not\cong \Delta_{G_2}$.

Proposition 17. The simplicial coloring complex of a simplicial complex is a strictly stronger

invariant than its simplicial chromatic polynomial.

Since the simplicial chromatic coloring complex is a special case of the hypergraph coloring complex, many of its properties follow from facts about the hypergraph coloring complex. In this section, we examine some of these properties particularly those introduced in [20]. The reader is referred to [20] for most of the proofs, as we are interested in the new interpretations of this coloring complex in terms of simplicial complexes and the simplicial chromatic polynomial.

Proposition 18. *Let S be a simplicial complex with n vertices and smallest minimal nonface dimension m , then Δ_S is an $(n - m - 1)$ -dimensional simplicial complex.*

The proof follows from the following observations. The dimension of the faces of the coloring complex are determined by the number of sets in a chain. This means that facets correspond to the longest possible chains. The way to maximize chains is to make sure one component is a minimal nonface and that all other components are single vertices.

The previous discussion of the facets of Δ_S provides a simple description of them. To each minimal nonface F of S there are $(n - |F| + 1)!$ associated facets, since the minimal nonfaces could be in any component of a chain. In [108], the subcomplex of facets associated to an edge are referred to as edge-spheres. Following the notation introduced in [20], we define the analogue for the simplicial chromatic coloring complex.

Definition 41. *Let S be a simplicial complex and let F be a minimal nonface of S . Then the subcomplex of faces of Δ_S such that F is contained in one of the components of Δ_S , denoted Q_F , is the **minimal-nonface-sphere** of F .*

Whether or not a simplicial complex is pure is another natural piece of information to desire. In [20], Breuer-Dall-Kubitzke show that the coloring complex of a hypergraph is not pure in general and the same is true for the simplicial coloring complex. Indeed, clutters need not be uniform hypergraphs, the condition for the hypergraph coloring complex to be pure. In terms of a simplicial complex S , the equivalent of having a uniform hypergraph is that all the minimal nonfaces of S be of the same dimension. This can be phrased in terms of the Alexander dual as well.

Proposition 19. *Let S be a simplicial complex, then Δ_S is pure if and only if all its minimal nonfaces are of the same dimension. Equivalently, Δ_S is pure if and only if its Alexander dual, S^* , is pure.*

Proof. The first part of the proposition follows from the definition of the coloring complex. Since the facets of S^* are the complements of the nonfaces of S , the Alexander dual will be pure if and only if all the minimal nonfaces of S are of the same dimension. \square

As previously mentioned, there are a number of ways to build the hypergraph coloring complex. These include the chain and component approach [72], a related partition of vertices approach, an arrangement approach and a monomial ideal approach from [65]. The last formulation allows for a natural description of Δ_S . Let J be a monomial ideal in $\mathbb{R}[x_1, \dots, x_n]$, then we define the complex Δ_J to have $(l-1)$ faces $\{A_1, \dots, A_l\}$ where

$$\emptyset \neq A_1 \subsetneq A_2 \subsetneq \dots \subsetneq A_l \neq [n]$$

and $\prod_{r \in A_i \setminus A_{i-1}} x_r \in J$ for at least one $1 \leq i \leq l+1$ where $A_0 = \emptyset$ and $A_{l+1} = [n]$. This general framework from [65] encapsulates the graph and hypergraph coloring complexes when J is the edge or hyperedge ideal. The simplicial chromatic coloring complex is similar.

Proposition 20. *Let S be a simplicial complex with vertex set $[n]$. Then $\Delta_S \cong \Delta_{I_S}$ where I_S is the Stanley-Reisner ideal of S .*

1.6 Future Directions

Question 1. *Baranovsky and Sazdanović showed that Helme-Guizon and Rong's chromatic graph homology is related to Eastwood and Huggett's construction via a spectral sequence [13]. It follows that two cochromatic graphs that are not distinguished by their chromatic graph homology are also not distinguished by the homology of their associated graph configuration spaces. Do there exist cochromatic graphs that are not distinguished by their Eastwood and Huggett homology but are distinguished by their chromatic graph homology?*

Question 2. *Lowrance and Sazdanović showed that the chromatic graph homology of a graph G over \mathbb{Z} with \mathcal{A}_2 is determined by its chromatic polynomial [80]. This is not true with other algebras, for example \mathcal{A}_3 . Since chromatic hypergraph homology generalizes chromatic graph homology, it follows that chromatic hypergraph homology is also not determined with all algebras \mathcal{A} . Is the chromatic hypergraph homology of \mathcal{G} also determined with \mathcal{A}_2 when \mathcal{G} is not a graph?*

Question 3. *Given a connected graph G with n vertices it has been shown that the non-trivial cohomology groups (with rational coefficients) come in isomorphic pairs in [30]. These pairs are said to be related by a knight move. Is there a generalization of the knight move to chromatic hypergraph homology?*

Question 4. *Whitney's broken circuit theorem for graphs shows that the chromatic polynomial of a graph $G = (V, E)$ can be written as an alternating sum over specific $s \subset E$ rather than all s . Chandler and Sazdanović categorify this theorem in [27]. Whitney's broken circuit theorem has a generalization to hypergraphs [46, 113]. Can this generalized broken circuit theorem also be categorified?*

Question 5. *In [80] it is shown that using \mathcal{A}_2 the only possible torsion in chromatic graph homology is of order 2. Does this extend to hypergraph chromatic homology?*

Question 6. *Do there exist hypergraphs \mathcal{G}_1 and \mathcal{G}_2 such that $H_i(M_{\mathcal{G}_1}) \cong H_i(M_{\mathcal{G}_2})$ but $M_{\mathcal{G}_1} \not\cong M_{\mathcal{G}_2}$?*

CHAPTER

2

TOPOLOGICAL DATA ANALYSIS APPLIED TO CANCER GENOMICS

2.1 Introduction

Cancer is a set of polygenic diseases with many driving factors and these factors can be measured through DNA methylation, copy-number data, gene expression data and micro RNA expression data among others [1, 78]. In this work we focus primarily on copy-number aberrations, one driving factor in cancers [55, 82, 85], as well as gene expression.

Copy number changes are found in most cancers and since they are known to be key regulators of gene expression, are frequently used as markers to identify cancer-driving genes [42]. Microarray and sequencing technologies have been used to detect copy number changes in cancer [41, 116, 38, 81]. These experimental methods have identified key cancer driver genes and revealed that not all copy number changes drive gene expression. Instead, many copy number changes that are apparently unrelated to gene regulation also accumulate as cancer evolves. The accumulation of these “passenger” copy number aberrations

make the identification of cancer-driving copy number changes a challenging task.

The standard approach to differentiate cancer-driving copy number aberrations from passenger copy number changes is through association studies. Topological methods replace standard measurements with topological measurements. TDA in cancer genomics is not a new idea. In fact, there are even now survey papers on TDA applied to oncology [23]. Measurements from topological data analysis have been used effectively as inputs into statistical methods to detect copy number changes [44]. Topological data analysis has more generally been successful in cancer genomics. For example, the Mapper algorithm was used to detect a new subgroup of breast cancer with a high survival rate [90]. More recently, it has been used to detect potential tumor-producing genes, some of which have been confirmed in mouse models [97].

In a typical study, array Comparative Genomic Hybridization data (aCGH) [33] is first segmented (most commonly using circular binary segmentation or similar approaches [91]) and each of the segments is then tested for association to a previously selected phenotype. In [44], a topological data analysis method was introduced to identify copy number changes associated with a given cancer phenotype. In this approach, called Topological Analysis of array CGH (TAaCGH), copy number data measured using array CGH is mapped into a point cloud from which topological signatures are extracted. Then an association study between these topological signatures and the desired phenotype is conducted. Figure 2.1 shows the TAaCGH workflow. One difference between topological approaches and traditional approaches is that their multiscale character allows for multiresolution analysis of each copy number change, making pre-selected cutoffs unnecessary. Additionally, because of the global character of topological signatures, they can capture combined effects from different copy number changes, as is expected in a polygenic disease.

TAaCGH was used to identify copy number changes associated with breast cancer molecular subtypes [9]. The study used Betti curves in dimension 0, β_0 curves, to detect copy number changes and it showed an overall agreement with other current methods to identify chromosome aberrations. This study was later combined with statistical learning methods to build logistic regression models as classifiers for cancer subtypes [57]. Betti curves in dimension 1 were also studied in [7] and used to identify co-occurring copy number changes, that is copy number changes that tend to appear in combination with other copy number changes but do not appear independently as observed in [8, 75].

Betti curves are known to be unstable with respect to small perturbations of the data [119]. We provide bounds on the distance between the Betti and lifespan curves that come from

the Vietoris–Rips complex and the one arising from the time series with a slight perturbation for both Betti and lifespan curves (Section 2.5) leveraging results from [31, 25].

Given copy number data, which we treat as time series data with position along the genome taking the place of time, we build a Vietoris–Rips complex on the associated sliding window point cloud. We expand the TAaCGH method to lifespan and persistent landscape curves and perform an exhaustive comparison of the performance of these curves on simulated and patient data [66]. Lifespan and persistent landscape curves have the potential to capture different properties of data than Betti curves. This was shown in [11], where the authors compared the persistent entropy function, Betti curves and other related summaries. They showed that these curves provide complementary information for the task of image classification. We therefore hypothesize that other persistence curves may discover relevant genes complementary to the ones found by Betti curves in [9].

Simulation results (Section 2.6.1) indicate that lifespan curves outperform Betti and landscape curves for the task of distinguishing a group of patients with single contiguous aberrations from a group of patients with no aberrations. In particular, lifespan curves are less sensitive to noise in the data of individual patients than the other persistence curves. Additionally, lifespan curves perform better on focal aberrations than Betti or landscape curves.

The performance of TAaCGH with lifespan and landscape curves on the dataset published by Horlings and colleagues [66] is presented in Section 2.6.2. The performance of all curves is comparable in detecting significant regions across molecular subtypes. In the HER2 subtype, all three curves detected segments in 17q, the long arm of chromosome 17, and importantly they all detected cytobands 17q12-q21.31 which contain the ERBB2 gene. For the Luminal A subtype, Betti and lifespan curves detected a subset of the regions detected by persistence landscapes. In the Luminal B subtype only one region of the short arm of chromosome 8, 8p22-p11.1, was detected and it was detected by Betti curves. In the Basal subtype the three curves detected similar regions, with landscapes detecting some different regions from the other two curves. Newly detected regions include 1q21.1-q25.2, 2p23.2-p16.3, 23q26.2-q28 for the Basal subtype, 8p22-p11.1 for Luminal B and 2q12.1-q21.1 and 5p14.3-p12 for Luminal A. The TCGA BRCA cohort data supports the new regions associated with the Basal and Luminal B phenotypes.

As in [57], we build predictive models using logistic regression and find that all approaches perform similarly, despite finding different predictor variables. In the Luminal A subtype only the region 5p14.3-p12 was found to have predictive power. For HER2 either

one of overlapping segments 17q11.1-q12 and 17q12-q21.31 was found to have predictive power depending on the persistence curve used. The predictor variables for Basal predictive models differed greatly between persistence curves. The only repeated predictor variable was 10p12.31-p11.1 which was a predictor variable for the lifespan and fourth landscape curve predictive models.

2.2 Topological Data Analysis

Forty years ago it would have seemed unlikely to many that abstract concepts from algebraic topology – such as homology – would be used in a real world application. Now, however, topological data analysis is a very exciting and active area of research. A quick search of Google Scholar for the phrase topological data analysis yields over 4.3 million hits with survey articles about applications of TDA to time series analysis [104], oncology [23], biomedicine [105], data science [29] and many more. Here we provide a basic introduction to persistent homology. To fill in more details consider Carlsson’s paper on topology and data [24] or Harer and Edelsbrunner’s very readable textbook [49]. For a refresher on algebraic topology take a look at [58].

The first step in a persistence-based topological data analysis approach is determining the point cloud on which to construct a filtration and then apply persistent homology. The point cloud may already exist in the form of multi-dimensional vector representations of data points or it may need to be constructed as in the case of initial time-series data. Next, simplicial complexes are built from this point cloud and the persistent homology of these simplicial complexes then yields topological information about the shape of the original data. We present a schematic of the process below:

$$\text{data} \xrightarrow{\gamma} \text{point cloud} \xrightarrow{\delta} \text{simplicial complex} \xrightarrow{\phi} \text{Persistent Homology} \xrightarrow{\psi} \text{TDA Summary.}$$

One of the advantages of this framework is its flexibility. In the case of one-dimensional input data, there are many potential conversions into a point cloud, one such example is the sliding window point cloud. The simplicial complex can be chosen such as the Vietoris-Rips, Čech, alpha or witness complexes. The types of topological summary can also be chosen such as persistence curves [31], persistent landscapes [22] or persistence images [2]. By varying these choices one can hone in on specific properties of the data.

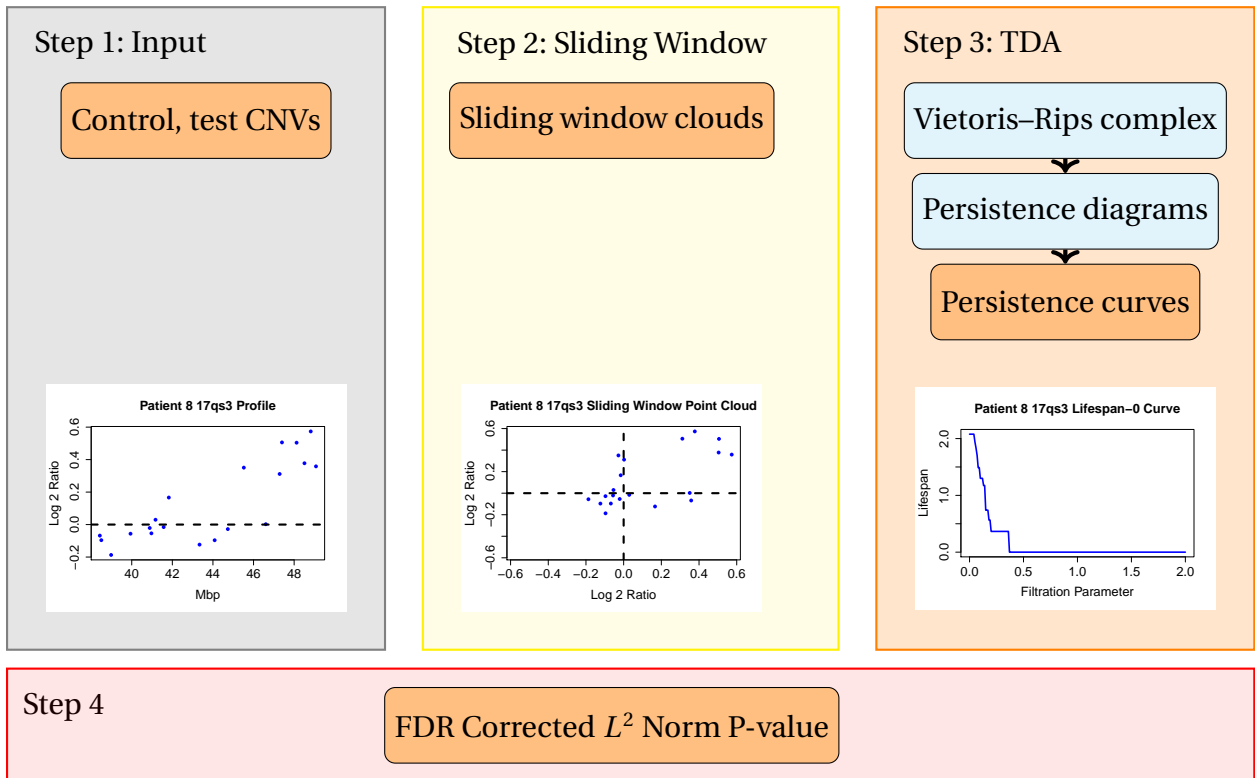


Figure 2.1: **The TAaCGH pipeline.** This workflow determines if a segment of the genome is statistically significant for a cancer subtype. Once a particular segment of study is chosen, the TAaCGH pipeline begins. In Step 1 the copy number variation data is separated into control and test patients. Copy number variation data from a single patient is pictured. Next, in Step 2, the data is converted into a sliding window point cloud for each patient. The sliding window point cloud from the sample patient’s data is pictured. In Step 3 a Vietoris–Rips filtration is built on each patient’s point cloud, the persistent homology of each patient’s Vietoris–Rips filtration is computed, recorded in a persistence diagram and summarized into a persistence curve. As an illustration we are using a lifespan curve from the sample patient. Examples of all persistence curves are shown in Figure 2.2c–e. Lastly, in Step 4, the persistence curves of all patients in the test group and in the control group are averaged and a permutation test is run on the L^2 norm of these averaged persistence curves.

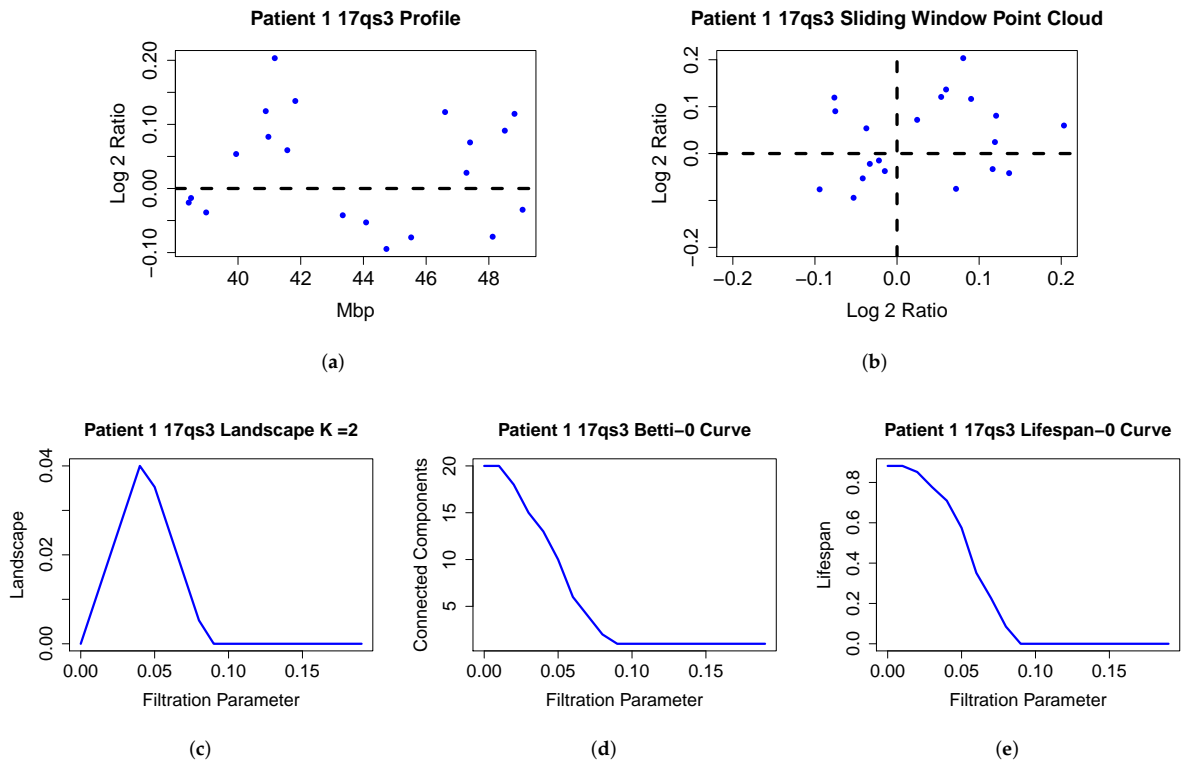


Figure 2.2: Sample patient data and output. The copy number data (a), sliding window point cloud (b), persistent landscape (c), Betti curve (d) and lifespan curve (e) of a patient from the Horlings dataset [66] on chromosome 17q segment 3 using 0-dimensional persistence. 17qs corresponds to the cytoband range 17q21.2-q21.33.

γ Choice of point cloud

Sometimes data comes prepackaged in a point cloud ready for topological data analysis techniques to be applied. Other times the data needs to be converted into a point cloud so that meaningful topological features can be extracted. For example, Harer-Perea introduce a method for detecting periodicity of time series data using persistent homology in [96]. In their algorithm they first convert the time series data to what is known as the sliding window point cloud, construct a Vietoris-Rips complex from the point cloud and finally apply persistent homology.

Since then the sliding window point cloud has been used as a step in TAaCGH to detect CNAs in cancer patients ([44], [9], [57], [7]).

Definition 42. *Given a time series x_1, x_2, \dots, x_n , the sliding window point cloud with window size s of this time series is the set of points $\{(x_1, x_2, \dots, x_s), (x_2, x_3, \dots, x_{s+1}), \dots, (x_n, x_1, \dots, x_{s-1})\}$.*

Example 10. *For example, consider the time series generated by the function $f(x) = \sin(x)$ generated by increments of $\frac{\pi}{4}$ from 0 to 2π . The time series is*

$$0, \frac{\sqrt{2}}{2}, 1, \frac{\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2}, -1, -\frac{\sqrt{2}}{2}, 0.$$

The sliding window point cloud with window size 2 constructed from this time series consists of the points $(0, \frac{\sqrt{2}}{2}), (\frac{\sqrt{2}}{2}, 1), (1, \frac{\sqrt{2}}{2}), (\frac{\sqrt{2}}{2}, 0), (0, -\frac{\sqrt{2}}{2}), (-\frac{\sqrt{2}}{2}, -1), (-1, -\frac{\sqrt{2}}{2}), (-\frac{\sqrt{2}}{2}, 0), (0, 0)$. The time series as well as its sliding window point cloud are pictured in Figure 2.3.

Since nearby genes tend to have similar copy numbers, combining the copy number ratios from adjacent points into one point makes sense heuristically. Many copy number aberrations can occur throughout the genome of a patient. In [7], it is noted that 1-dimensional cycles of the Vietoris-Rips complexes built from the sliding window point cloud of copy number data can capture co-occurring copy number aberrations. This makes both 0 and 1-dimensional persistence of the sliding window point cloud useful and we study both here.

δ Choice of simplicial complex

The Vietoris-Rips (VR) filtration is frequently used due to its interpretability and computability.

Sliding Window Point Cloud of $f(x) = \sin(x)$

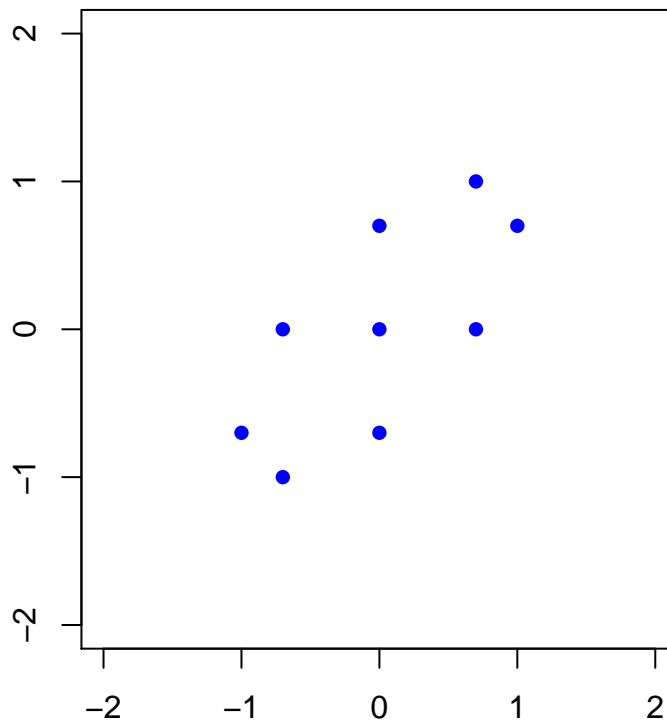


Figure 2.3: The sliding window point cloud of the function $f(x) = \sin(x)$ at increments of $\frac{\pi}{4}$ from 0 to 2π .

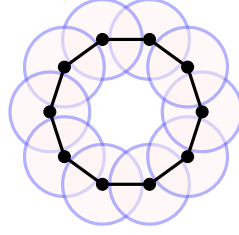


Figure 2.4: The Vietoris Rips complex on points sampled from a circle.

Vietoris-Rips Complex Čech Complex

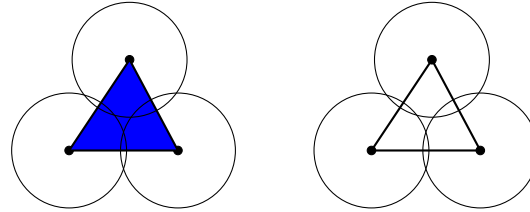


Figure 2.5: Vietoris-Rips Complex and Čech complex on the same point cloud.

Definition 43. Let X be a point cloud and $\epsilon > 0$, then the **Vietoris-Rips complex** on X , $VR(X, \epsilon)$ is the simplicial complex defined on the points of X as follows. A set of points $\{x_1, \dots, x_k\} \in X$ form a face of $VR(X, \epsilon)$ if and only if $B(x_i, \epsilon) \cap B(x_j, \epsilon) \neq \emptyset$ for all $i, j \in [k]$.

For example, consider the points in Figure 2.4 sampled from a circle and distance 1 from each other. Then the Vietoris-Rips complex for $\epsilon = .45$ is pictured. The similarly defined Čech complex more accurately captures the topology of the space, but is computationally more expensive.

Definition 44. The **Čech complex**, $\check{C}(P, \epsilon)$, has $\{p_1, \dots, p_k\}$ as a simplex if and only if $\bigcap_{i=1}^k B(p_i, \epsilon)$ is nonempty.

To see the difference between a Čech complex and a Vietoris-Rips complex consider Figure 2.5. Once a Vietoris-Rips complex has the edges of a complete graph that complete graph is filled in by a face. The Čech complex, however, requires a n -way intersection of discs around 3 points to include the full n simplex from those points. The Čech complex is an example of a nerve complex.

Definition 45. Let $C = \{U_i\}_{i \in I}$ be a finite collection of sets. Then the **nerve** of C , denoted $N(C)$, is

$$N(C) = \{J \subseteq I : \bigcap_{j \in J} U_j \neq \emptyset\}$$

Specifically, the Čech complex is the nerve of the collection of disks of radius epsilon around each point in a given point cloud the complex is built on. Another filtration that is sometimes used is the alpha complex.

Definition 46. Let $P = \{p_1, \dots, p_m\} \subseteq \mathbb{R}^n$ be a point cloud. To each point $p_i \in P$ we associate the **Voronoi cell** of p_i , denoted V_{p_i} , which is

$$V_{p_i} = \{x \in \mathbb{R}^n : d(x, p_i) \leq d(x, p_j) \text{ for all } p_j \in P\},$$

all of the points in \mathbb{R}^n that are closer to p than the other points in the point cloud.

Around each $p_i \in P$ we define $\tilde{B}(p_i, \epsilon) = B(p_i, \epsilon) \cap V_{p_i}$.

Definition 47. The **alpha complex** on P , denoted $\alpha(P, \epsilon)$ is then the simplicial complex on the vertices of P such that $\{p_1, \dots, p_k\}$ forms a simplex if and only if $\bigcap_{i=1}^k \tilde{B}(p_i, \epsilon)$ is nonempty.

The alpha complex is homotopy equivalent to the Čech filtration, but in low dimensions can be easier to compute.

It is simple to check that these complexes are actually filtered simplicial complexes, that is, for $\epsilon < \delta$ $\text{VR}(P, \epsilon)$ is a subcomplex of $\text{VR}(P, \delta)$ and $\check{C}(P, \epsilon)$ is a subcomplex of $\check{C}(P, \delta)$.

For this work it is also important to mention a separate interpretation for the VR filtration. Since a VR filtration is completely determined by its 1-skeleton it can be encoded by a sequence of graphs G_i on the vertices of a point cloud. The simplicial complex associated to each graph G_i is the clique complex of the graph denoted $X(G_i)$.

Definition 48. The **clique complex** of a graph is the simplicial complex on the vertices of the graph such that a set of vertices $\{p_1, \dots, p_k\}$ forms a $k - 1$ face if and only if it is a k -clique of the graph.

ϕ Persistent homology

After a filtered simplicial complex is constructed from the point cloud, the next step is to extract topological information e.g. the number of connected components or 1-cycles for each value of ϵ . To get this information the homology groups H_n of the complex at each ϵ are computed over \mathbb{Z}_2 . The homology of a simplicial complex X is derived from a chain complex. This chain complex is defined in the following way. First an order is established on the vertices of X .

Definition 49. Let X be a simplicial complex, a **p -chain**, $c = \sum_i a_i \sigma_i$ is a formal sum of p -simplices σ_i in X with $a_i \in \mathbb{Z}_2$. Given $c = \sum_i a_i \sigma_i$ and $c' = \sum_i b_i \sigma_i$ define $c + c' = \sum_i (a_i + b_i) \sigma_i$. Then the **group of p -chains**, $C_p(X)$ is the set of all p -chains of X under the addition operation.

Next the boundary homomorphism is defined.

Definition 50. Let $\sigma = [v_0, \dots, v_p] \in X$. Then define the **boundary homomorphism**, $\partial_p : C_p(X) \rightarrow C_{p-1}(X)$, by $\partial_p(\sigma) = \sum_{j=0}^p [v_0, \dots, \hat{v}_j, \dots, v_p]$ and extending linearly where \hat{v}_j means that vertex has been omitted.

Given the chain complex defined by the group of p -chains and the associated boundary homomorphism homology can be defined.

Definition 51. The **p th homology group** of X , $H_p(X)$, is $H_p(X) = \ker(\partial_p) / \text{im}(\partial_{p+1})$.

Let X be a simplicial complex and let $f : X \rightarrow \mathbb{R}$ be nondecreasing on increasing chains of faces. Then $f^{-1}((-\infty, a])$ is a subcomplex of X for each $a \in \mathbb{R}$. Applying f to the simplices of X yields $a_1 < a_2 < \dots < a_n$ and fix $a_0 = -\infty$ then $X_i = f^{-1}(a_i)$ for each i . This sequence of complexes is the filtration associated to f . This filtration essentially adds simplices until the full simplicial complex X is reached. The Vietoris-Rips, Čech and alpha complexes can all be formed in this way, see [49] for more details.

Given a simplicial complex X , and a monotonic $f : X \rightarrow \mathbb{R}$, functoriality of homology can be used to define persistent homology. Let $X = (X_0 \subseteq X_1 \dots \subseteq X_n)$ be a filtered simplicial complex.

Definition 52. For each $i \leq j$ denote the map induced by the inclusion $f_p^{i,j} : H_p(X_i) \rightarrow H_p(X_j)$. The **p th persistent homology groups** are $H_p^{i,j}(X) = \text{im}(f_p^{i,j})$. $\gamma \in H_p(X_i)$ is **born** at i if $\gamma \notin H_p^{i-1,i}(X)$ and that it **dies** entering j if it merges with a class born before it in X_j , i.e. $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}(X)$ but $f_p^{i,j}(\gamma) \in H_p^{i-1,j}(X)$.

The intuition behind this definition is that n -dimensional "holes" that exist at lower filtration values will be filled in by simplices at higher filtration values. An important measure of the p -th persistent homology groups are the p th persistent Betti numbers.

Definition 53. The **p th persistent Betti numbers** are $\beta_p^{i,j}(X) = \text{rank}(H_p^{i,j}(X))$.

Given a class $\gamma \in H_p(X_i)$ its persistence can be defined.

6



6

Figure 2.6: A point cloud formed by the corners of a 6×4 rectangle.

Definition 54. Let $\gamma \in H_p(X_i)$ be born at X_i and die at X_j , then its **persistence** or **lifespan** is $a_j - a_i$. If γ never dies, then its persistence is set to ∞ .

The information in persistent homology groups can be encoded in persistence diagrams.

Definition 55. Let X be a simplicial complex and $f : X \rightarrow \mathbb{R}$ be nondecreasing on increasing chains of simplices in X . Then the **p th persistence diagram** of the filtration on X induced by f is $\text{Dgm}_p(f)$ is the multiset of points (a_i, a_j) corresponding to independent p -dimensional classes that are born at X_i and die entering X_j along with the diagonal in $(\mathbb{R}_{\geq 0} \cup \{\pm\infty\})^2$. The points in the p th persistence diagram (b, d) are called **birth-death pairs**.

As an example, consider the point cloud in Figure 2.6. It has the Vietoris-Rips filtration pictured in Figure 2.7. At filtration parameter $\epsilon = 0$ there are 4 connected components. We say these components are born at $\epsilon = 0$. At $\epsilon = 2$ the balls around the left and right points in the point cloud intersect and therefore these vertices are connected in the Vietoris-Rips complex. This means two of the 4 original connected components have now combined with two other connected components. We say these connected components have died and record the birth-death pair $(0, 2)$ for each of these connected components. At $\epsilon = 3$ the top and bottom points in the point cloud become connected, joining all the points into one connected component. This means a connected component dies at $\epsilon = 3$ so we record the birth-death pair $(0, 3)$ in the persistence diagram. Finally, at $\epsilon = \frac{\sqrt{52}}{2}$ all points are connected to each other in the Vietoris-Rips complex forming a tetrahedron. There is still only one connected component so there are no more birth-death pairs. The 0-dimensional persistence diagram for this filtration is pictured in Figure 2.8.

Persistence diagrams are similar for 1-dimensional persistent homology, except now instead of connected components they record the birth-death pairs of 1-cycles. In Figure 2.7

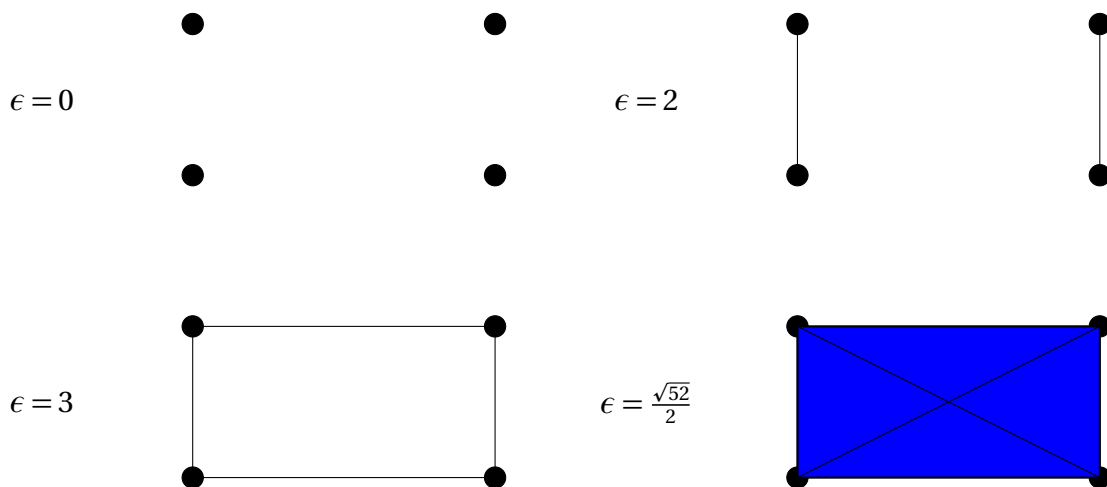


Figure 2.7: The Vietoris-Rips filtration on the rectangle point cloud from Figure 2.6.

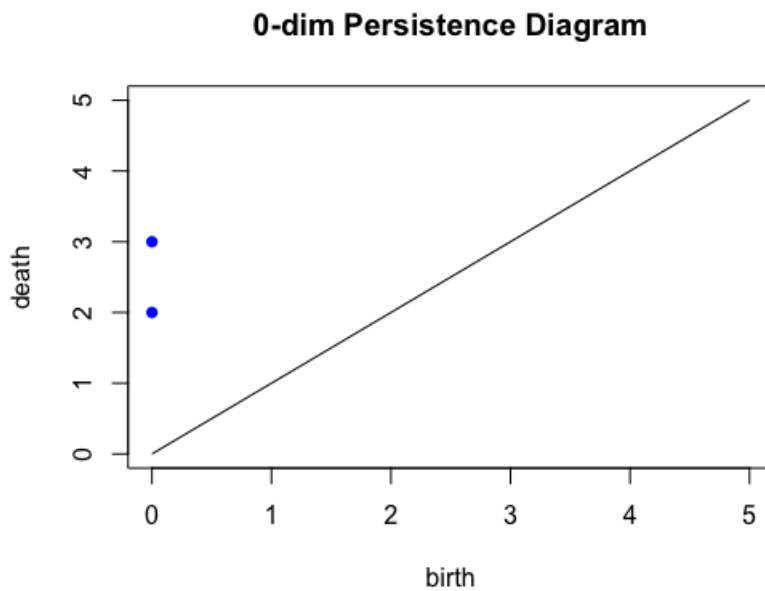


Figure 2.8: The 0-dimensional persistence diagram for the Vietoris-Rips complex in Figure 2.7.

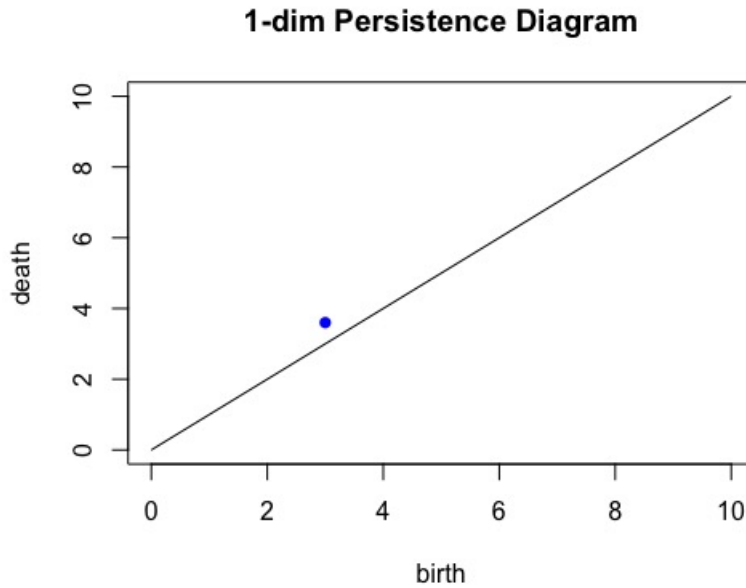


Figure 2.9: The 1-dimensional persistence diagram of the Vietoris-Rips complex in Figure 2.7

a single 1-cycle was born at $\epsilon = 3$ and died at $\epsilon = \frac{\sqrt{52}}{2}$. This is recorded in the 1-dimensional persistence diagram in Figure 2.9.

A key question that was posed soon after persistent homology was introduced is its stability with respect to small perturbations in the input data. If slightly different data produced vastly different persistence diagrams then persistent homology would not be useful for data analysis. In order to measure stability, different metrics were introduced on the space of persistence diagrams. The two main distances we will be interested in for this work are the bottleneck distance and 1-Wasserstein distance.

Definition 56. Let C and D be persistence diagrams. Then the **bottleneck distance** between C and D is

$$W_\infty(C, D) = \inf_{f \in \Gamma} \sup_{p \in C} \|p - f(p)\|_\infty$$

where Γ is the set of all bijections from C to D .

In other words, the bottleneck distance is the smallest distance between points in a matching between persistence diagrams such that all matched points are at most that

distance measured with the infinity norm on \mathbb{R}^2 . The 1-Wasserstein distance is defined similarly.

Definition 57. *Let C and D be persistence diagrams along with the diagonal in $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$. Then the **1-Wasserstein** between C and D is*

$$W_1(C, D) = \inf_{f \in \Gamma} \sum_{p \in C} \|p - f(p)\|_{\infty}^1 \quad (2.1)$$

where Γ is the set of all bijections from C to D .

The bottleneck distance was shown to be stable to small perturbations in the initial point cloud with respect to the Gromov-Hausdorff distance between point clouds [28]. The 1-Wasserstein distance was shown to be stable with respect to the Wasserstein distance on point clouds [106].

Many libraries now exist to compute persistent homology efficiently including Dionysus [88], Gudhi [111], Perseus [89], Eirene [63], and Ripser [15]. We use the R package TDA [53] which makes use of the Gudhi and Dionysus libraries.

ψ Persistence Curves

In this section, we discuss persistence curves, an important tool for summarizing topological information, which we can then combine with statistical methods. Since the number of points in persistence diagrams varies based on the values of the data, there is no well-defined mean of persistence diagrams and they cannot be treated as a fixed-length vector. This makes methods from statistics and machine learning difficult to apply. In order to overcome this, the topological information from persistent homology is frequently summarized using tools such as persistence curves [31], kernel SVM for persistence [98], persistence landscapes [22], and persistence images [2] among others. We focus on persistence curves including Betti and lifespan curves, and persistence landscapes. It is worth noting that, in the 0-dimensional case, one generator could have an infinite lifespan. We therefore choose to consider reduced persistent homology for lifespan curves which amounts to removing the infinite generator in dimension 0 or assert that the infinite generator dies at a predetermined filtration value. We begin by defining the persistence curves of interest.

Definition 58. *Given an n -dimensional persistence diagram D the n th **Betti curve**, denoted $\beta_n(D, t)$, is equal to the number of birth-death pairs $(b, d) \in D$ such that $t \in (b, d)$.*

The n th lifespan curve is defined similarly.

Definition 59. The n th **lifespan curve** denoted $\ell_n(D, t)$ is equal to the sum of the lifespans of all birth-death pairs $(b, d) \in D$ (where $d \neq \infty$) such that $t \in (b, d]$.

The lifespan curve can be thought of as a Betti curve where the birth-death pairs are weighted by their lifespan. Lastly, we define persistence landscapes as introduced in [22].

Definition 60. The k th **persistence landscape** of D , denoted $\lambda(k, t)$, is

$$\lambda(k, t) = k\max_p([\min(t - b, d - t)]_+)$$

where $p = (b, d) \in D$, $[c]_+ = \max(c, 0)$ and $k\max$ is the k th highest value.

Persistence curves were originally introduced as a general framework under which previously studied summaries of persistence diagrams lie [31]. A similar framework was also built and studied in [25]. The work in [31] allows for easy generation of new summaries including lifespan curves, which we consider. Lastly, the persistence curve framework provides a way to make general arguments about the stability of the curves with respect to the bottleneck distance between persistence diagrams. Persistence landscapes were shown to be stable in [22].

2.3 Methods

2.3.1 The TAaCGH Method

The Topological Analysis of Array CGH (TAaCGH) method is a form of genetic association study which uses copy number data to associate segments of the genome with particular subtypes of breast cancer. It differs from standard association studies by using topological information as a test statistic. An overview of the pipeline is pictured in Figure 2.1. The method begins by splitting the copy number data from each chromosome arm into consecutive segments of 20 probes each. Each segment overlaps with the next segment in 10 probes. Next, the data set is split into a test and control set. The process tests for the significance of each individual segment. The test set consists of all patients from one cancer subtype and the control set consists of all other patients. The data for a fixed segment is converted into a point cloud in \mathbb{R}^2 using the sliding window mapping [44, 96] for each

patient. A Vietoris–Rips complex is built on each point cloud and the persistent homology of each Vietoris–Rips complex is taken. The persistent homology is then summarized into persistence curves for each patient. The data, sliding window point cloud, Betti curve, lifespan curve, and 2nd persistent landscape curve on 0-dimensional persistence are pictured in Figure 2.2 for a particular patient. Next, the persistence curves from all patients in the test set are averaged together and similarly for the control set. The L^2 norm of the difference between the average test and the control persistence curves is used as the test statistic. Permutation testing with FDR correction for multiple testing is used to test the significance of the statistic and thus of each segment for each cancer subtype.

It is worth noting that in [9], persistence was calculated using JavaPlex [3] which in some cases outputs a shorter lifespan than the R TDA package used in this study. To illustrate the difference, consider 1-dimensional persistence of a Vietoris–Rips filtration on the four vertices of a unit square. If the JavaPlex discretization increment is set to 0.01, a 1-dimensional cycle connecting all the vertices is born at filtration parameter value $t = 1$ which agrees with the R TDA output. However, R TDA computes its persistence as precisely $[1, \sqrt{2})$ while JavaPlex gives $[1, 1.41)$, due to the choice of discretization.

2.3.2 Cancer Subtype Predictive Models

In [57], the authors build a logistic regression predictive model to quantify the predictive power of the detected copy number aberrations. The predictor variables were built using TAaCGH and consisted of two types of predictors. The full chromosome arm copy number changes that were detected by the displacement of the center of mass of the chromosome arm and segment copy number changes whose significance was determined by Betti curves. The center of mass technique was developed to complement the detection of local aberrations by Betti curves in [9]. This approach associated arms 1p, 16p and 16q to the Luminal A subtype, arm 9p to Luminal B, and arms 1p, 2p, 3q, 4p, 5q, 6p, 6q, 8q, 10p, 10q, 12p and 14q to Basal.

We expand this approach to use predictor variables associated with the significant segments detected by lifespan and landscape curves. Next, we compare the predictive power of the models built from each type of curve. In order to make the comparison, we implemented the corresponding logistic models and computed their accuracy. We also computed a confusion matrix for each model.

First, we find the set of chromosome arms A with a significant displacement in their

centers of mass and the set of maximally non-intersecting significant chromosome sections K as described in [9]. Patients are then classified as either positive 1 or negative 0 for indicator variables $I_{a,i}^{CM}$ and $I_{k,i}^S$ for significant arms $a \in A$ and sections $k \in K$. The subscript i denotes the patient number. To specifically define these indicator variables, we introduce the notation $P_{i,k}^{Test}, P_{i,k}^{Ctrl}$ denoting the average persistence curve from the test and control sets with patient i removed from the set of all patients. $P_{i,k}$ denotes the persistence curve of patient i on segment k . We also introduce the similarity between persistence curves

$$SS_{k,i}^G = \sum_{\epsilon} (P_{i,k} - P_{i,k}^G)^2$$

for $G \in \{Test, Ctrl\}$ and ϵ the filtration parameter. Then the indicator variable $I_{k,i}^S$ for patient i and section k which determines if patient i has the aberration for that section is

$$I_{k,i}^S = \begin{cases} 1 & \text{if } SS_{k,i}^{Test} < SS_{k,i}^{Ctrl} \\ 0 & \text{if } SS_{k,i}^{Test} \geq SS_{k,i}^{Ctrl}. \end{cases}$$

Similarly, if the center of mass of the point cloud of the patient is outside the confidence interval for the control group center of mass, then the indicator variable $I_{i,a}^{CM} = 1$ otherwise it is 0. Specifically, we have if the center of mass for the arm is a gain for the arm $a \in A$ then

$$I_{a,i}^{CM} = \begin{cases} 1 & \text{if } \bar{x}_i^a > \mu + \frac{t_{\alpha}\sigma}{\sqrt{n}} \text{ with } n-1 \text{ d.f.} \\ 0 & \text{otherwise} \end{cases}$$

and if the center of mass for the arm is a loss then

$$I_{a,i}^{CM} = \begin{cases} 1 & \text{if } \bar{x}_i^a < \mu - \frac{t_{\alpha}\sigma}{\sqrt{n}} \text{ with } n-1 \text{ d.f.} \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{x}_i^a = \sum_{probes} \frac{x_i^a}{n_a}$ where n_a is the number of probes in the arm a and n is the number of patients. Next, we fit a logistic regression model for each phenotype and persistent curve type over indicator variables $I_{k,i}^S$ and $I_{a,i}^{CM}$ using the patient data from the Horlings dataset and a classification threshold of ≥ 0.5 for positive classification of the phenotype. Note that a set of $I_{k,i}^S$ exists for each type of persistence curve which could lead to some ambiguity,

the context of these variables will make it clear which persistence curve they are defined for. The logistic regression model for a specific persistence curve is then

$$\text{Logit} = \text{Intercept} + \sum_{k \in K} w_k I_{s,i}^S + \sum_{a \in A} w_a I_{a,i}^{CM}.$$

In order to prevent overfitting, we use forward addition and the Akaike Information Criterion (AIC) for model selection. This criterion introduces a penalty with respect to the number of covariates in the final model

$$AIC := 2k - 2\ln(\hat{L})$$

where k is the number of variables in the model and \hat{L} is the maximum of the likelihood function for the model. Forward addition begins with a null model and adds covariates until the best model is found. We evaluate these models by using leave-one-out cross-validation.

2.4 Data

2.4.1 Horlings Dataset

The dataset used in this study is from [66]. This is the same dataset used in [44, 9, 57]. It consists of BAC Microarrays from the genome with an average spacing of 1 Mb. Each BAC clone was spotted in triplicate on each slide (Code Link Activated Slides, Amersham Biosciences). The dataset contains 68 patient samples from the 4 most common molecular subtypes of breast cancer: Luminal A, Luminal B, basal-like and HER2+. There are 21 Luminal A samples, 12 Luminal B samples, 21 basal-like samples and 14 HER2+ samples. We consider each molecular subtype separately as the test group and the remaining patients as the control group.

2.4.2 TCGA BRCA Cohort Data

The TCGA BRCA cohort data was collected from the Firehose dataset with the disease name: breast invasive carcinoma [21]. The dataset consists of 1098 tumors hybridized to the Affymetrix SNP 6.0 array platform, using the GRCh38 assembly of the human genome as a reference. Circular binary segmentation (CBS) was applied and the copy number values

were estimated for each segment. Results were then $\log_2(\text{copy number}/2)$ transformed and used to assign focal scores to protein-coding genes. A cutoff was further considered to discretize the focal score into values of $-2, -1, 0, 1,$ and 2 , where -2 means complete deletion, -1 means loss, 0 means normal, 1 means gain, 2 means amplification. This dataset contains 185 Basal samples, 549 Luminal A samples, 206 Luminal B samples and 81 HER2+ samples.

2.4.3 UCSF 500 Gene Panel

The TAaCGH method identifies regions of the genome but does not pinpoint specific genes. To further validate the method and pinpoint specific genes of interest we considered the UCSF 500 gene panel [114]. This is a panel of around 500 genes that is used in the clinic. This panel helps to identify the type of cancer the patient has and to direct personalized therapies and treatment. Specifically, it has been used to help identify targeted therapies for patients with pediatric brain cancer [73]. We intersected the TCGA BRCA data set with the genes from the UCSF 500 and considered the genes that fall within regions identified within the Horlings dataset.

2.4.4 Simulation Data

In the simplest hypothetical case, a cancer patient has a single contiguous copy number aberration of some length. As in real cancer patients, this aberration will have copy number values either above or below 0 around the same positive or negative value representing either a gain or a loss. In order to perform well on real data, the method must be capable of distinguishing the copy number aberrations from random noise. To test this, we generate data mimicking patients with single contiguous aberrations and control patients without them (see [9, 57]). We used the patients with the aberration as the test set and the rest of the patients as the control set. Specifically, we generated aberrant profiles in the test set with copy number values within the aberration drawn from a normal distribution with mean $\mu \in \{-1, 0.6, 1\}$ and standard deviation $\sigma \in \{0.2, 0.22, 0.24, \dots, 0.5\}$. The rest of the copy number values in the aberrant test profiles were drawn from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma \in \{0.2, 0.22, 0.24, \dots, 0.5\}$ matching the standard deviation of the aberrant values. The total length of each profile was 20 probes, chosen to match the length of segments in [9]. The length of the contiguous section of aberrant values was $\lambda \in \{1, 2, 3, 5, 10, 15\}$. Control patient profiles had values drawn from a normal distribution

with mean $\mu = 0$ and standard deviation $\sigma \in \{0.2, 0.22, 0.24, \dots, 0.5\}$, matching the standard deviation of their test counterparts. Since there is no guarantee that all patients will contain an aberration, simulations were also run with the MIX parameter which determined the penetrance of the aberration in the population, that is the percentage of patients within the test set that had aberrations. The rest of the patients in the test set consisted of control profiles. The MIX parameter was in $\{20\%, 40\%, \dots, 100\%\}$. 50 simulations were run for each set of variables and each simulation consisted of 120 patient profiles. 60 of the profiles were in the test set and the other 60 were in the control set. For each combination of parameters, we computed the sensitivity of the TAaCGH method for correctly determining the significance of the test set from the control set. Sensitivity is defined to be $\frac{TP}{TP+FN}$ where TP is the number of true positives and FN is the number of false negatives. The second set of simulations was performed as well, using the same kind of data as the first set of simulations across all parameters. This time, for each of the 120 patients in a simulation we calculated its persistence curve. Then we calculated the distance from the current patient's persistence curve to the average curve from the test and control sets. Each was classified as a test profile if its distance from the test persistence curve was smaller than its distance from the control persistence curve and as a control curve otherwise. Given these classifications, sensitivity and specificity was calculated for each simulation. Specificity is defined as $\frac{TN}{TN+FP}$ where TN is true negative and FP is false positive. An average sensitivity and specificity were then computed over the 50 iterations of each simulation with given parameters.

2.5 Bounds on the Distance between Persistence Curves

In this section, we address the statistical properties of persistence curves. Theorem 1 from [31] provides general upper bounds on the difference between two persistence curves under the L^1 norm in terms of the bottleneck (W_∞) and 1-Wasserstein distances (W_1). These upper bounds can provide stability results for some persistence curves. It is possible, however, that these upper bounds are not tight and that curves which appear unstable under these general upper bounds are actually stable. Recently, Dlotko and Gurnari improved on the bounds in [31] to show that distance between Betti curves under the L^1 norm is stable with respect to the 1-Wasserstein distance [45].

Theorem 21 ([45]). *Let C and D be two k -dimensional persistence diagrams. Then their*

Betti curves are stable with respect to the 1-Wasserstein distance,

$$|\beta_k(C, t) - \beta_k(D, t)| \leq 2W_1(C, D).$$

The difference in Betti curves under a different norm, the ∞ -norm, has been shown to be unstable with respect to the 1-Wasserstein (W_1) and bottleneck distance (W_∞) [70].

Here we apply theorem 1 from [31] to Betti and lifespan curves, to measure how far away two persistence curves could theoretically get under certain constraints.

Theorem 22 ([31]). *Let C and D be persistence diagrams, W_∞ denote the bottleneck distance, n^C be the number of birth-death pairs in C and L^C denote the sum of the lifespans of all birth-death pairs in C . Then*

$$\|\beta(C, t) - \beta(D, t)\|_1 \leq 2 \max(n^C, n^D) W_\infty(C, D) + \min(L^C, L^D) \quad (2.2)$$

$$\|\ell(C, t) - \ell(D, t)\|_1 \leq 2(L^C + L^D) W_\infty(C, D). \quad (2.3)$$

The main challenge in the application of persistence curves, including Betti and lifespan curves, comes from the fact that small perturbations in the initial point cloud can lead to large changes in the curves [119]. The main result of this section is a bound on the L^1 norm between two Betti or two lifespan curves built from finite and bounded point clouds with respect to the bottleneck distance. Consider the bounds on the L^1 norm between two Betti curves or two lifespan curves from Theorem 22. The bottleneck distance is already stable with respect to small perturbations in the initial point clouds [34], so our bounds will be in terms of it. We need to find bounds on the maximal number of birth-death pairs in a persistence diagram, as well as the maximal lifespan of any birth-death pair.

First, we establish the existence of bounds on these quantities under the given constraints for i -dimensional persistent homology. Then we explicitly compute bounds in the case of 1-dimensional persistent homology of the Vietoris–Rips complex. We also compute bounds in the case of 0-dimensional persistent homology for both the Vietoris–Rips and Čech complex. The existence results given here are for Vietoris–Rips and Čech complexes, but essentially the same arguments work for the various forms of witness complexes described in [24]. They also hold for the alpha complex since it is filtered homotopy equivalent to the Čech complex.

Proposition 21. *Let $P \subseteq \mathbf{R}^n$ be a finite point cloud, then there exists a bound on the maximal number of birth-death pairs in the persistence diagrams of $\text{VR}(P, \epsilon)$ and $\check{C}(V, \epsilon)$.*

Proof. Since P is finite, there are a finite number of simplicial complexes that can be built on P . Both the VR and Čech filtrations change a finite number of times, hence the persistence diagrams from these complexes have a maximal number of generators. \square

Recall we consider reduced 0-dimensional persistent homology to avoid the issue of an infinite lifespan generator.

Proposition 22. *Let $P \subseteq \mathbf{R}^n$ be a finite point cloud with diameter d , then the maximal lifespan of a birth-death pair from the i -dimensional persistence diagram of $\text{VR}(P, \epsilon)$ or $\check{C}(P, \epsilon)$ is at most d .*

Proof. Let $P = \{p_1, \dots, p_k\}$ be a finite point cloud in \mathbf{R}^n . Consider the epsilon balls $B(p, \epsilon)$ for $p \in P$ and $\epsilon > d$. Since the diameter of P is d , each of these balls must contain all other points in P . Therefore, $B(p_i, \epsilon) \cap B(p_j, \epsilon) \neq \emptyset$ and $\bigcap_{i=1}^k B(p_i, \epsilon) \neq \emptyset$ so both $\text{VR}(P, \epsilon)$ and $\check{C}(P, \epsilon)$ are $(k-1)$ -simplices. Since simplices are contractible, there is no i -dimensional persistent homology for $i \geq 1$ and therefore the maximal lifespan is bounded above by m . In the 0-dimensional case we consider reduced homology. \square

The following theorem can be proved by combining Theorem 22 with Propositions 21 and 22.

Theorem 23. *Let $P \subseteq \mathbf{R}^n$ be a finite point cloud with diameter d , then the i -dimensional Betti and lifespan curves of $\text{VR}(P, \epsilon)$ and $\check{C}(P, \epsilon)$ are bounded with respect to small perturbations of P .*

The next results provide explicit bounds on the maximal number of independent classes 1-dimensional persistent homology in a Vietoris–Rips filtration. To do this, we require that the given point clouds have pairwise distinct distances between points. The following two results were made available to the authors through private correspondence with David Moon [87]. The proofs provided here are different from those provided to the authors.

Proposition 23. *Let $G = (V, E)$ be a simple graph on n nodes, then the maximal 1st Betti number of the clique complex of G , $X(G)$, is $\lfloor \frac{n}{2} \rfloor \lfloor \frac{n}{2} \rfloor - (n-1)$.*

Proof. The first Betti number of G is $\beta_1(G) = |E| - n + 1$. Subtracting the number of triangles T in G from this quantity yields $\beta_1(X(G)) = |E| - n + 1 - T$. Let G_1 be a graph containing at least one triangle. Remove an edge from a triangle in G_1 to obtain G_2 . G_2 has at least one less triangle than G_1 , but only one less edge so $\beta_1(X(G_2)) \geq \beta_1(X(G_1))$. The graph G

for which $\beta_1(X(G))$ is maximized must therefore be triangle-free. By Mantel's theorem [6], the triangle-free graph with the maximal number of edges is the complete bipartite graph $K_{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$. This graph has $\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil$ edges which completes the proof. \square

Proposition 24. *Let $P \subseteq \mathbb{R}^d$ be a finite point cloud with n vertices such that the pairwise distances between points are distinct. Then the maximal number of birth-death pairs in the 1-dimensional persistence diagram of $\text{VR}(P, \epsilon)$ is $\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil - (n - 1)$.*

Proof. The information from a VR filtration can be encoded by a sequence of graphs such that $G_i \subseteq G_{i+1}$. Since the distances between points in P are pairwise distinct, G_i differs from G_{i+1} by a single edge. If adding an edge $e = 12$ to G_i to form G_{i+1} completes a triangle 123 , then e does not birth a new cycle in $H_1(X(G_{i+1}))$. To see this note that if e completes a cycle say $a_1, \dots, a_k, 1, 2$ then this cycle is homologous to $a_1, \dots, a_k, 1, 3, 2$ in $X(G_{i+1})$ since the two cycles differ by the triangle 123 . Since the edges 13 and 32 were in G_i the cycle represented by $a_1, \dots, a_k, 1, 3, 2$ was already in $H_1(X(G_i))$ and hence e did not birth a new cycle. Since triangles do not birth new cycles, any VR filtration which has the maximum number of birth-death pairs in a persistence diagram can have all generators alive at once. Therefore the maximum number of birth-death pairs over the entire filtration is the same as the maximum number of cycles that can be alive at a fixed filtration parameter. Proposition 23 completes the proof. \square

Proposition 24 improves a special case of Theorem 3.1 from [56] for $n < 24$, which says that the maximal 1st Betti number of a Vietoris–Rips complex at a fixed filtration value is $5n$.

The point clouds considered in this work have a maximum diameter. In particular, if the minimal and maximal copy number ratios are c_{\min} and c_{\max} , then all sliding window point clouds with window size s are contained in $[c_{\min}, c_{\max}]^s$. Therefore, the maximum diameter of these point clouds is $d = \sqrt{s}(c_{\max} - c_{\min})$.

Theorem 24. *Let P and P' be sliding window point clouds built from copy number ratio data with window sizes s . Let P and P' each consist of n points. Let C and D be the 1-dimensional persistence diagrams coming from the Vietoris–Rips filtration built on P and P' . Then*

$$\begin{aligned} \|\beta_1(C, t) - \beta_1(D, t)\|_1 &\leq \left(\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil - (n - 1) \right) (2W_\infty(C, D) + \sqrt{s}(c_{\max} - c_{\min})) \\ \|\ell_1(C, t) - \ell_1(D, t)\|_1 &\leq 4 \left(\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil - (n - 1) \right) \sqrt{s}(c_{\max} - c_{\min}) W_\infty(C, D). \end{aligned}$$

In the 0-dimensional case for both the Vietoris–Rips and Čech filtrations, it is clear that the maximal number of connected components in a point cloud with n vertices is n . This yields the following explicit bounds for β_0 and ℓ_0 curves in the case of copy number variation sliding window point clouds.

Theorem 25. *Let P and P' be sliding window point clouds built from copy number ratio data with window sizes s . Let P and P' each consist of n points. Let C and D be the 0-dimensional persistence diagrams coming from either Vietoris–Rips or Čech filtrations built on P and P' . Then*

$$\begin{aligned}\|\beta_0(C, t) - \beta_0(D, t)\|_1 &\leq n(2W_\infty(C, D) + \sqrt{s}(c_{max} - c_{min})) \\ \|\ell_0(C, t) - \ell_0(D, t)\|_1 &\leq 4n\sqrt{s}(c_{max} - c_{min})W_\infty(C, D).\end{aligned}$$

The bounds from Theorem 25 apply to persistence diagrams that come from a single test patient and a single control patient. In the TAaCGH pipeline the curves from all of the test patients are averaged together, then the curves from all of the control patients are averaged together and finally, these average curves are compared. Since a mean cannot be defined on persistence diagrams, this bound cannot be extended to a bound on average persistence curves.

2.6 Results

2.6.1 Comparison of Performance of Different Persistence Curves on Simulated Data

We used simulations to understand the properties of TAaCGH with various persistence curves as we vary aberrations. As outlined in more detail in Section 2.4.4 and Figure 2.10, we created simulated data for patients with single contiguous aberrations defined by the parameters: μ the mean of the aberration, σ the standard deviation of the aberration and λ the length of the aberration. We then studied the sensitivity of the TAaCGH method to distinguish groups of these patients with aberrations from patients with no aberrations. Lastly, we introduced the MIX parameter which determines the percentage of patients in the test set that have the aberration, the rest of the patients have nonaberrant profiles. In particular we consider Betti curves, lifespan curves, and landscape curves within the

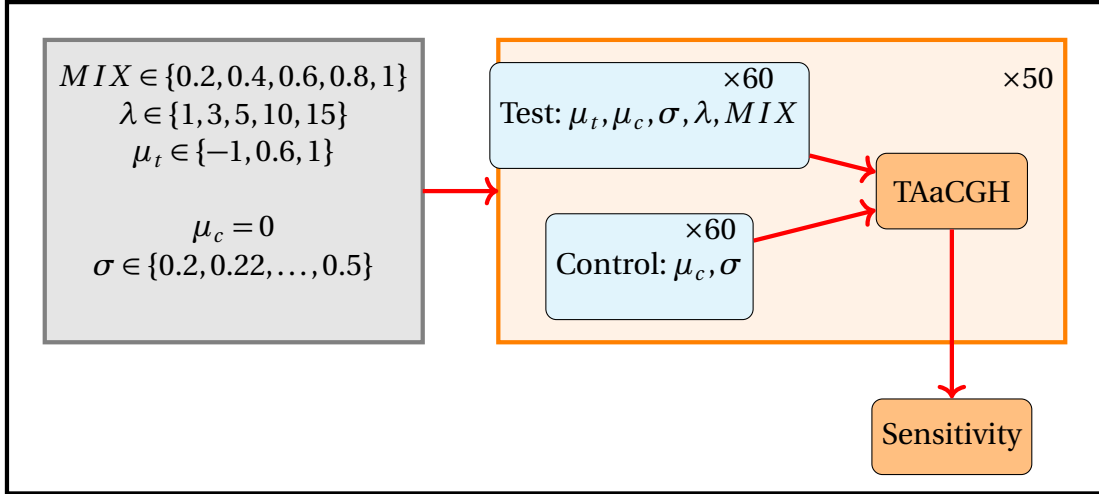


Figure 2.10: Simulation design. Simulations are designed to test the ability of TAaCGH with various persistence curves to distinguish a set of patients with a single contiguous aberration from a set of patients without them. Each patient has 20 probes. Test patients with aberrations of length λ have copy number values sampled from a normal distribution with mean μ_t and standard deviation σ . The remaining copy number values for test patients are sampled from a normal distribution with mean $\mu_c = 0$ and standard deviation σ . Control patients have all copy number values sampled from a normal distribution with mean $\mu_c = 0$ and standard deviation σ . The MIX parameter controls the percentage of patients in the test set that have aberrations, the remaining patients in the test set have data drawn from a normal distribution with mean $\mu_c = 0$ and standard deviation σ . For each set of parameters, we ran 50 simulations. Each simulation consisted of a total of 120 patients, 60 in the test set and 60 in the control set. The parameters varied over the following values $\mu_t \in \{-1, 0.6, 1\}$, $\sigma \in \{0.2, 0.22, \dots, 0.5\}$, $\lambda \in \{1, 3, 5, 10, 15\}$ and $MIX \in \{20\%, 40\%, 60\%, 80\%, 100\%\}$.

TAaCGH pipeline on this simulated data, see Figure 2.11.

We begin by investigating the effect of the MIX parameter on the various topological summaries. In Figure 2.12, we compare the sensitivity of the lifespan curve to Betti and landscape curves. We do this by fixing the MIX parameter, and considering all simulations with that MIX parameter as the other parameters vary across their ranges as defined in Section 2.4.4. Then we compute the sensitivity of each persistence curve for each set of simulation parameters. Lastly, we calculate the percentage of all simulations with this fixed MIX parameter for which the sensitivity of the lifespan curve is larger, smaller or equal to the sensitivity of each type of persistence curve.

As the MIX parameter decreases, the lifespan curve outperforms the Betti curve Figure 2.12a.

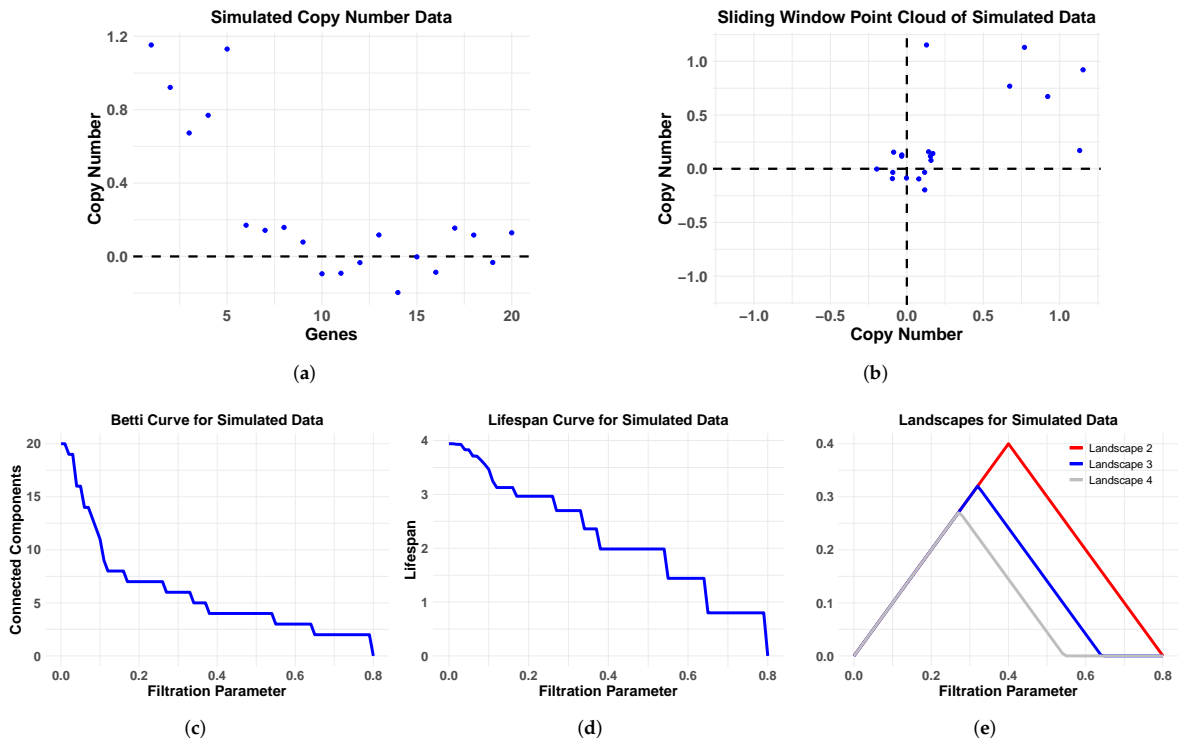


Figure 2.11: Simulation data and associated persistence curves. Simulated CNA data **(a)**, corresponding sliding window point cloud **(b)**, Betti curve **(c)**, lifespan curve **(d)** and landscape curves **(e)** for a hypothetical cancer patient on a segment of 20 probes with a single length $\lambda = 5$ contiguous aberration with aberration mean $\mu = 1$ and standard deviation $\sigma = 0.2$. The nonaberrant probes are sampled from a distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.2$.

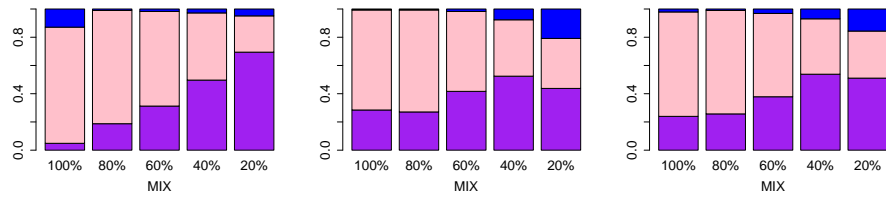


Figure 2.12: Comparison of sensitivity values of different approaches in TAAcGH with respect to varying MIX parameter on persistence of simulated data in dimension zero. Figure shows comparisons between lifespan ℓ_0 curve and Betti β_0 (a), second landscape λ_2 (b), third landscape λ_3 (c), fourth landscape λ_4 curves (d). For a fixed value of the MIX parameter, we compute the sensitivity of our method with one of the persistence curves for each set of simulations as the standard deviation σ , the mean μ and the length λ vary over all possible values detailed in Section 2.4.4. The height of each bar represents the percentage of those simulations where the sensitivity of the lifespan curve was bigger (purple), equal to (pink) and less than the sensitivity (blue) of other persistence curves.

For the lifespan curve compared to the landscape curves the results are similar Figure 2.12b–d. As the MIX parameter decreases, for the most part, the percentage of simulations for which the lifespan curve has a higher sensitivity than the landscapes increases or stays the same. In all cases, the lifespan curve has a higher percentage of sensitivity values for which the lifespan sensitivity is higher than the landscape sensitivities. In summary, the lifespan curve outperforms Betti curves and landscape curves as the mix parameter decreases.

Next, we compare the lifespan curve to the Betti curve for a fixed standard deviation σ and vary the other parameters (see Figure 2.13). For every value of the standard deviation, the lifespan curve has a higher percentage of simulations in which it has a higher sensitivity than Betti and landscape curves. When the lifespan curve is compared to each of the landscape curves in general as the standard deviation increases, the percentage of simulations where the lifespan curve outperforms the landscape curve increases, see Figure 2.13b,c. The same general trend is visible in Figure 2.13a, where lifespan curves are compared to Betti curves. The peak percentage is lower than for the comparison to the three landscape curves. Overall, the lifespan curve outperforms the other curve types at higher standard deviations.

The sensitivity values of the lifespan curves compared with the sensitivity values of the Betti and landscape curves with the length λ fixed as the other parameters vary are

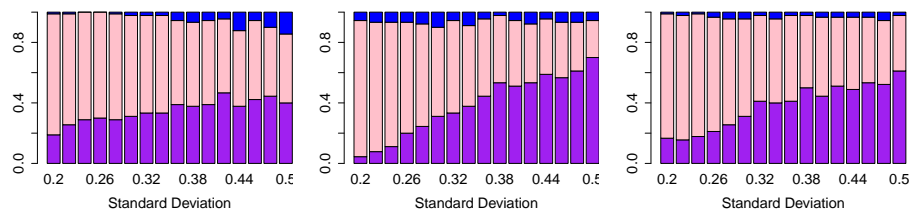
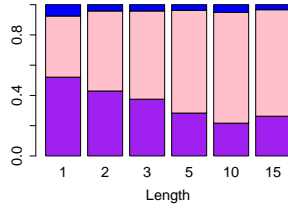


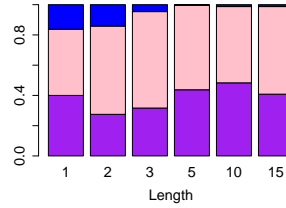
Figure 2.13: Comparison of sensitivity values of different approaches in TAaCGH with respect to varying the standard deviation σ on persistence of simulated data in dimension zero. The figure shows comparisons between lifespan ℓ_0 curve and Betti β_0 (a), second landscape λ_2 (b), fourth landscape λ_4 curves (c). For a fixed value of the standard deviation σ , we compute the sensitivity of our method with one of the persistence curves for each set of simulations as the mean μ , length λ and MIX parameters vary over all possible values detailed in Section 2.4.4. The height of each bar represents the percentage of those simulations where the sensitivity of the lifespan curve was bigger (purple), equal to (pink) and less than the sensitivity (blue) of the other persistence curve. Note that the third landscape λ_3 behaves similarly to the second landscape.

shown in Figure 2.14. As expected, the shorter the aberration, the worse each type of curve performs. For example, Figure 2.14(a) shows that lifespan curves outperform Betti curves for all aberration lengths. Similarly, Figures 2.14 (b,c,d) show that lifespan curves outperform second, third and fourth landscape curves respectively except for the third landscape at length 1. It is interesting that the number of cases in which Betti outperforms lifespan is independent from the length of the aberration.

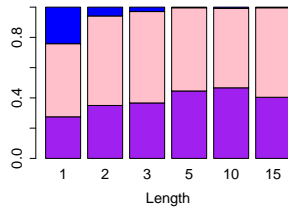
In summary, as we fix the MIX, standard deviation σ or length λ parameters and allow the other parameters to vary, the lifespan curve outperforms the Betti and landscape curves on 0-dimensional persistence. This indicates that lifespan curves are less sensitive to noise in the data of individual patients (as determined by σ and by MIX) than the other persistence curves. Additionally, lifespan curves perform better when aberrations have a small length. The only exception was for Betti curves when the MIX parameter was equal to 1 (Figure 2.12), where all test patients have the aberration. Therefore, if there is reason to believe that most patients in the test set contain the aberration, then Betti curves could be the better choice. In the second set of simulations we tested patient classification with Betti curves, lifespan curves and various landscape curves. This is a key step in building subtype classifier models. In Tables 2.1 and 2.2 the average sensitivity and specificity are pictured for simulation data



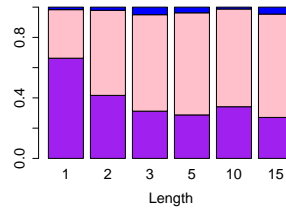
(a) Lifespan compared to Betti



(b) Lifespan compared to λ_2



(c) Lifespan compared to λ_3



(d) Lifespan compared to λ_4

Figure 2.14: Comparison of sensitivities of different approaches in TAaCGH with respect to varying the length λ on persistence of simulated data in dimension zero. The figure illustrates differences between lifespan ℓ_0 curve and Betti β_0 (a), second landscape λ_2 (b), third landscape λ_3 (c), fourth landscape λ_4 curves (d). For a fixed value of the MIX parameter, we compute the sensitivity of our method with one of the persistence curves for each set of simulations as the standard deviation σ , the mean μ and the mix parameter varies over all possible values detailed in Section 2.4.4. The height of each bar represents the percentage of those simulations where the sensitivity of the lifespan curve was bigger (purple), equal to (pink) and less than the sensitivity (blue) of other persistence curves.

Table 2.1: Average sensitivity and specificity of lifespan curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.

Lifespan					
$\mu = 1$	20% mix	40% mix	60% mix	80% mix	100% mix
$\sigma = 0.2$	TPR: 32.00%	42.00%	60.00%	80.00%	99.00%
	SPC: 85.00%	96.00%	99.00%	100.00%	100.00%
$\sigma = 0.5$	TPR: 49.00%	56.00%	63.00%	71.00%	78.00%
	SPC: 62.00%	70.00%	76.00%	79.00%	84.00%
Total	TPR: 40.50%	49.00%	61.50%	75.50%	88.50%
	SPC: 73.50%	83.00%	87.50%	89.50%	92.00%

Table 2.2: Average sensitivity and specificity of second landscape curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.

Landscape 2					
$\mu = 1$	20% mix	40% mix	60% mix	80% mix	100% mix
$\sigma = 0.2$	TPR: 27.00%	41.00%	59.00%	78.00%	93.00%
	SPC: 92.00%	98.00%	99.00%	100.00%	100.00%
$\sigma = 0.5$	TPR: 49.00%	48.00%	53.00%	58.00%	63.00%
	SPC: 60.00%	68.00%	70.00%	73.00%	76.00%
Total	TPR: 38.00%	44.50%	56.00%	68.00%	78.00%
	SPC: 76.00%	83.00%	84.50%	86.50%	88.00%

with mean $\mu = 1$ and length $\lambda = 10$ as standard deviation and MIX vary. For both types of curves as the mix parameter increases both the sensitivity and specificity do as well. When comparing the two kinds of curves, lifespan curves have a higher sensitivity than second landscape curves across all parameters. When the mix parameter is at 20% or 40% second landscape curves have better specificity than or equal specificity to lifespan curves. Once the MIX parameter hits 60% lifespan curves have higher specificity.

2.6.2 Comparison of Topological Summaries within the TAaCGH Framework on Horlings Data

As detailed in Section 2.3.1, JavaPlex, used for the persistent homology computation in [9], differs from the TDA package in *R*. Therefore, we have repeated the study with 0-dimensional Betti curves β_0 . Results are shown in Tables 2.3–2.7 and indicate that both studies agree for

Table 2.3: HER2 Phenotype: Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the HER2 subtype on the Horlings dataset [66] using the R TDA package.

HER2+	
Betti-0	17q11.1-q21.31, 17q21.31-q22
Lifespan-0	17q11.1-q22
Landscape	λ_2 : 17q11.1-q21.31, λ_3 : 17q11.1-q21.33, λ_4 : 17q11.1-q21.33

Table 2.4: HER2 Logistic Regression: The confusion matrices and accuracy of logistic regression models built to predict the HER2 phenotype from Betti, lifespan and landscape curves.

Betti HER2		Lifespan HER2		λ_2 HER2		λ_3 HER2		λ_4 HER2	
9	5	9	5	8	6	7	7	6	8
2	48	0	50	0	50	2	48	0	50
Accuracy: 89%		92%		90%		86%		88%	

Luminal A, while the new study detected 8p22-p11.1 for Luminal B with Basal patients in the control set, missed 17q21.2-q21.33 from the previously detected segments 17q11.1-q22 for HER2+, while for Basal the new study missed some segments and detected an additional one: 2p23.2-p16.3. The missing segments from the new Betti curve study compared to the old Betti curve study are shown in Table A.1. We applied our 3 curves of study, Betti, lifespan and persistence landscape in dimension 0.

No significant regions for the Luminal B subtype were detected in [9] when the Basal

Table 2.5: Luminal A Phenotype: Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the Luminal A subtype on the Horlings dataset [66] using the R TDA package.

Luminal A	
Betti-0	11q 22.1-q23.2
Lifespan-0	2q12.1-q21.1, 5p14.3-p12
Landscape	λ_3 : 2q12.1-q21.1, 5p14.3-p12, 11q22.1-q23.2

Table 2.6: Luminal A Logistic Regression: The confusion matrices and accuracy of the logistic regression models built to predict the Luminal A phenotype from lifespan curves and third landscape curves.

Luminal A Lifespan	Luminal A Landscape 3
11 7	9 9
6 40	2 44
Accuracy: 80%	Accuracy: 83%

subtype was included in the control group. This was hypothesized to be because Luminal B is known to share similar aberrations with other breast cancer subtypes [5]. Therefore, TAaCGH was run with Luminal B as the test set but only HER2+ and Luminal A as the control set in [9]. We repeated this with our three curves of study.

2.6.3 Genomic Regions Associated to Cancer Subtypes

In this section, we report regions significant for each breast cancer subtype based on Betti, lifespan, and persistence landscape curves computed with the R TDA software package, GUDHI for computing persistence and Dionysus for detecting generators. Results are compared against TCGA BRCA cohort data set.

HER2

For the HER2+ subtype, the results agree with previous methods. The results are summarized in Table 2.3. Every method besides lifespan curves missed at least one of the segments detected in the original study. The original study detected 17q11.1-q22, new Betti curves missed 17q21.2-21.33, second landscape functions missed 17q11.1-q22 and third and fourth landscape functions missed 17q21.31-q22. We used the TCGA BRCA cohort data [21] to validate some of the significant regions that the persistence curves detected. For the HER2 phenotype, all significant sections were contained in chromosome arm 17q. The four chromosome arms with the most aberrant cytobands for the HER2 phenotype are pictured in Figure 2.15. These are arms 1, 8, 17 and 20. The most aberrant cytobands in the TCGA BRCA cohort dataset are in arm 17q and are detected by TAaCGH. No cytobands are detected in arms 1, 8 or 20 for the HER2 phenotype. One reason for this may be because

Table 2.7: Basal phenotype: Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the Basal subtype on the Horlings dataset [66] using the R TDA package.

Basal	
Betti-0	1p36.32-p33, 1p32.3-p31.1, 1p22.2-p12, 2p23.2-p16.3, 2p15-p11.2, 3p26.3-p24.3, 3p21.2-p13, 4p15.1-p11, 4q21.21-q34.1 5p15.33-p15.1, 5q11.1-q13.1, 6p25.3-p22.1, 6p21.33-p11.2, 6q24.1-q27, 7p21.3-p14.2, 9p24.3-p22.3, 10p15.3-p11.1, 10q21.1-q22.1, 10q22.2-q26.11, 12p13.31-p11.21, 13q12.2-q31.2, 13q31.2-q34, 14q24.3-q32.33, 15q11.2-q22.31, 15q23-q26.3, 18q12.1-q21.2, 23p22.33-p11.21
Lifespan-0	1p36.32-p36.11, 1p32.3-p31.1, 2p23.2-p16.3, 2p15-p11.2, 3p26.3-p25.1, 4q24-q27, 4q28.3-q31.3, 4q31.3-q34.1, 6p21.33-p11.2, 10p15.3-p11.1, 10q23.1-q24.2, 13q21.1-q31.2, 15q14-q22.31, 23p13.2-p12
Landscape	λ_4 : 1p32.1-p31.1, 1q21.1-q25.2, 2p15-p11.2, 3p26.3-p25.1, 4p15.1-p11, 4q24-q28.3, 4q31.21-q34.1, 5p15.33-p15.1, 10p15.3-p12.31, 10p12.31-p11.1, 10q23.1-q25.1, 12p13.31-p11.21, 13q31.2-q34, 14q31.3-q32.33, 23p22.33-p21.3, 23q26.2-q28

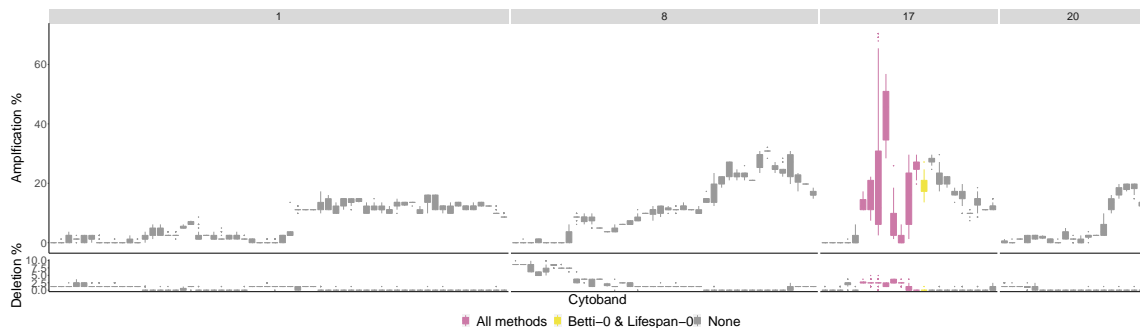


Figure 2.15: HER2 phenotype most aberrant cytobands in TCGA BRCA cohort Data Cytobands with gains and losses in the TCGA BRCA cohort dataset [21] for the HER2 phenotype. The chromosome arms 1, 8, 17 and 20 are included since they had above 10% of patients with aberrations in genes in those cytobands on average. The colors indicate which persistence curves detected those cytobands as significant in [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of HER2 patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with a score -2 .

multiple phenotypes contain significant aberrations in the same chromosome arms in the TCGA BRCA cohort dataset. This can be seen in Figure A.1 where the only arm in which the HER2 percentages are significantly higher than the percentages from all patients is in arm 17q.

Luminal A

The significant regions detected by the various methods for the Luminal A subtype are contained in Table 2.5. There were only two newly detected regions 2q12.1-q21.1 and 5p14.3-p12 which were both detected by lifespan and the third landscape curves. The only other significant region that was detected was 11q22.1-q23.2 which was detected by Betti curves and the third landscape but not lifespan curves or other landscape curves.

The difference between Luminal A compared to all other phenotypes is in Figure A.2 in the Appendix. The Luminal A subtype is associated with 11q22.1-q23.2, 2q12.1-q21.1 and 5p14.3-p12 by the TAaCGH method. These sections were not validated by the TCGA BRCA cohort data at the cytoband level. The arms with the most aberrant cytobands in the TCGA BRCA cohort data along with the methods which detect significant sections within them are pictured in Figure A.6. However, there is a clear signal in the Horlings dataset. Figure 2.16 shows four examples of typical Luminal A patients from 2q12-2q21.1.

There is a large copy number gain between base pairs 125 million to 130 million.

Luminal B

For the case of Luminal B, only one significant section was detected by any of the methods, including the original study. The new Betti 0 study detected 8p22-p11.1. As noted in Section 2.6.2, it was hypothesized that no significant sections were detected for this subtype because Luminal B is known to contain similar aberrations to the other cancer types. Therefore, we repeated the study while removing the basal patients from the control set. In this case, many new sections were detected, particularly by the lifespan and landscape curves. Cytobands 1q32.1-q41 and 12q21.31-q23.2 were detected by all three methods. The first region, 1q32.1-q41, was a copy number gain in the Horlings dataset. The second region, 12q21.31-q23.2, was driven by one patient in the Luminal B phenotype with significant copy number aberrations, while all other profiles were centered at 0. This matches the TCGA BRCA cohort dataset, where this particular aberration is rare. All newly detected segments are in Table A.2 where red indicates newly detected segments.

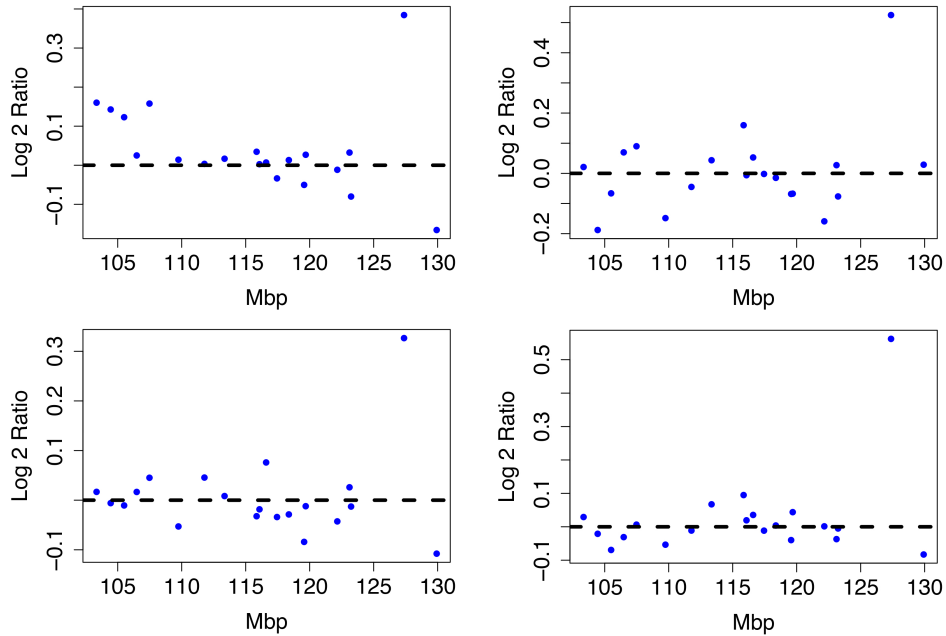


Figure 2.16: Luminal A Patient Profiles. Four Luminal A patient profiles from the Horlings dataset on cytobands 2q12-2q21.1. All share a significant copy number gain between 125 and 135 Mbp.

The difference between Luminal B compared to all other phenotypes is in Figure A.3 in the appendix. The most aberrant chromosome arms for the Luminal B phenotype in the TCGA BRCA cohort dataset are 1, 8, 11, 17 and 20 and are pictured in Figure A.8. The significant regions detected by the TAaCGH method with no basals in the control set are colored in Figure A.8. These same arms are pictured in Figure 2.17 where significant cytobands are colored from the Horlings dataset with Basals in the control set. The new region detected for the Luminal B phenotype is 8p22-8p11.1. This section is validated by the TCGA BRCA cohort data in Figure 2.18.

The TCGA BRCA cohort data show that a higher percentage of Luminal B patients have a copy number gain of 8p11.21-8p11.23. It also shows that for genes in 8p12-8p22, Luminal B patients have a higher percentage of copy number losses than the other phenotypes. This matches Luminal B patients from the Horlings dataset. Some sample patients from the Horlings dataset are shown in Figure A.11 matching the TCGA results.

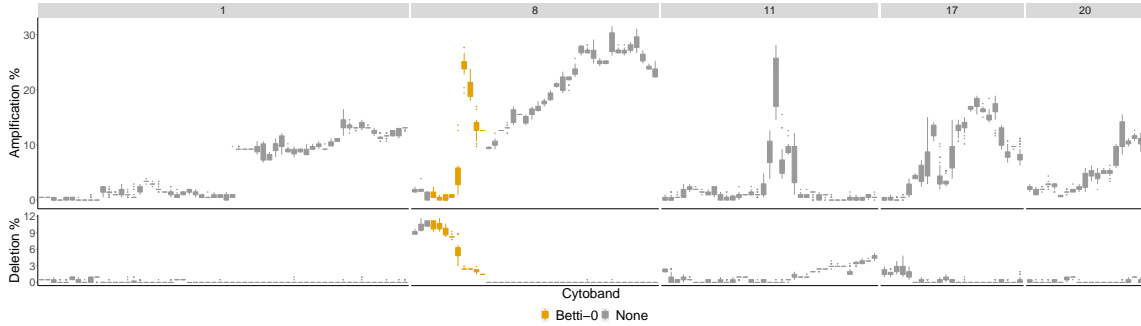


Figure 2.17: Luminal B most aberrant cytobands in TCGA BRCA cohort Data. The chromosome arms 1, 8, 11, 17 and 20 are included above 10% of patients had aberrations in the genes within this cytoband on average. The colors indicate which persistence curves detected those cytobands as significant in [66] for the Luminal B phenotype with no Basals in the control group. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

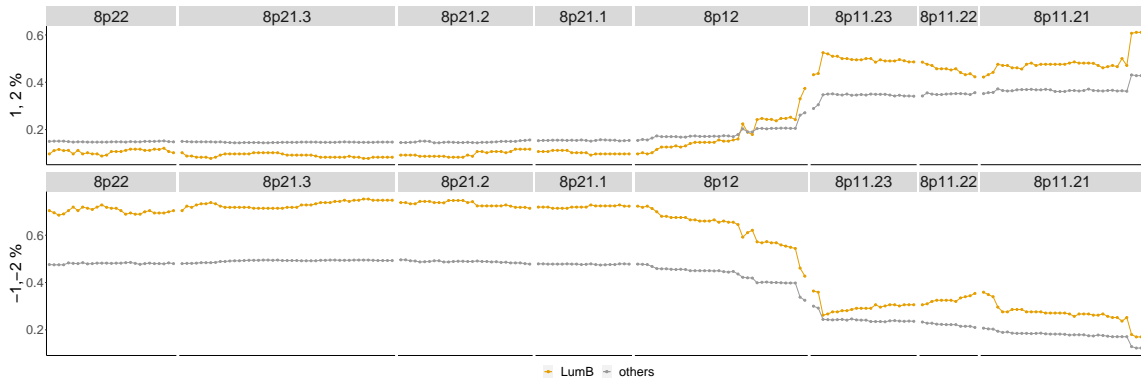


Figure 2.18: Luminal B phenotype cytobands 8p22-8p11.21. Top graph: Shows the percentage of patients in the Luminal B phenotype with either a 1 or 2 score from the TCGA data (orange) as well as the percentage of all other phenotypes with these scores (gray). Bottom graph: Shows the same as the top graph but for scores of -1 , -2 .

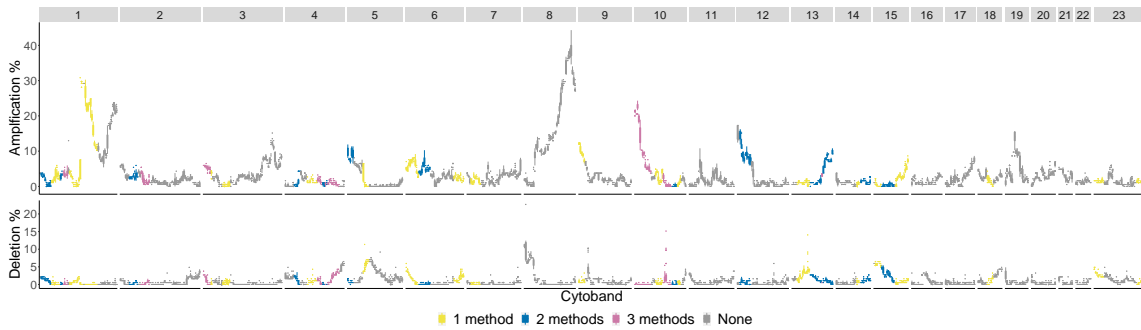


Figure 2.19: Basal phenotype cytobands in TCGA BRCA cohort dataset. The colors indicate how many persistence curve methods detected that particular cytoband in the Horlings dataset [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Basal patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with a score -2 .

Basal

The significant regions detected by the various methods for the Basal subtype are contained in Table 2.7 with newly detected regions in red. Three new segments are detected as significant compared to the original study: 1q21.1-q25.2, 2p23.2-p16.3 and 23q26.2-q28. The new Betti curve and the lifespan curve both detect 2p23.2-p16.3, but 23q26.2-q28 is only detected by the 4th landscape curve. Both 1q21.1-q25.2 and 23q26.2-q28 are copy number gains in the Horlings dataset, whereas 2p23.2-p16.3 is driven by an undetermined combination of gains and losses within the region. Notably, the second and third landscape functions do not detect any significant segments, suggesting that significance for the Basal subtype in dimension 0 is driven by less persistent connected components.

The significant cytobands for the Basal subtype are in Figure 2.19 colored by the number of persistence curves that detect them.

The chromosome arms with the most aberrant cytobands for the Basal subtype in the TCGA BRCA cohort dataset are 1, 3, 5, 8, 10, 12 and are pictured in Figure 2.20.

Newly detected cytobands 1q21.1-1q24.2, 2p16.3-2p23.2 and 23q26.2-23q28 were detected as copy number gains in the Horlings dataset which is validated by the TCGA BRCA cohort dataset for individual genes within these cytobands. This can be seen in Figures 2.21, A.9 and A.10. The difference between basal compared to all other phenotypes are in Figure A.4 in the appendix.

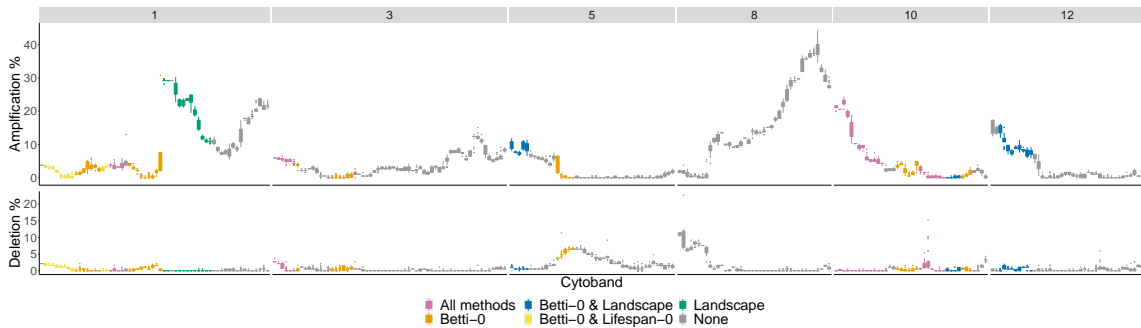


Figure 2.20: Basal phenotype most aberrant cytobands in TCGA BRCA cohort dataset. The chromosome arms 1, 3, 5, 8, 10, and 12 are included since above 10% of patients had aberrations in the genes in this cytoband. The colors indicate which persistence curves detected those cytobands as significant in [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Basal patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with a score -2 .

In Figure A.9, we see that over 80% of patients in the TCGA BRCA cohort dataset from the Basal phenotype have copy number gains for most of the genes in cytobands 1q21.1-1q25.2. The other phenotypes only have around 70% of their patients with a copy number gain in these cytobands.

In Figure 2.21, around 60% of the patients with the Basal phenotype have copy number gains compared to around 10% of patients from the other phenotypes. This supports the detection of 2p16.3-2p23.2 as significant in the Horlings dataset for the Basal phenotype.

In Figure A.10, around 30% of the patients with the Basal phenotype have copy number gains compared to around 10% of patients from the other phenotypes. This supports the detection of 23q26.2-23q28 as significant in the Horlings dataset for the Basal phenotype.

2.6.4 Cancer Subtype Classification Models

HER2

In our predictive model for the HER2+ subtype, both Betti curves and life-span curves detected 17q12-q21.31, abbreviated 17qs2, with 89% and 92% accuracy, respectively. Landscape 2 and landscape 3, however, detected 17q11.1-q12 abbreviated 17qs1, as the predictor

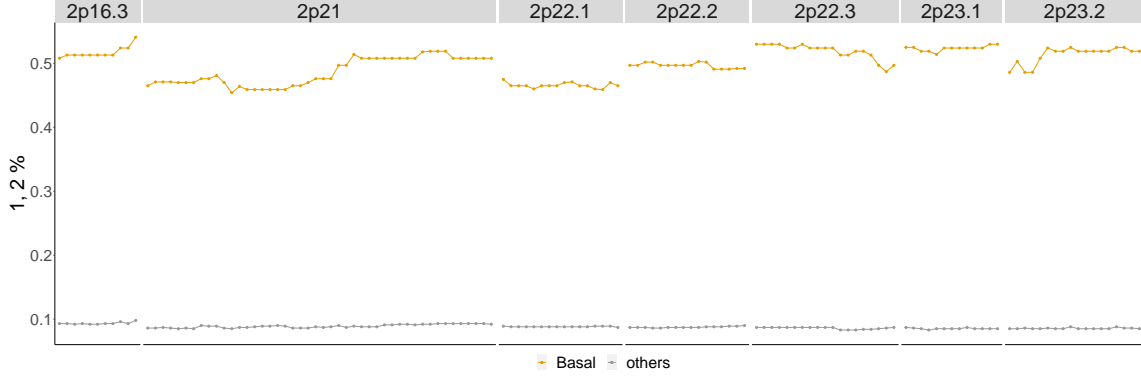


Figure 2.21: Basal phenotype cytobands 2p16.3-2p23.2. Shows the percentage of patients in the Basal phenotype with either a 1 or 2 score from the TCGA BRCA cohort dataset (orange) as well as the percentage of all other phenotypes with these scores (gray).

with 90% and 86% accuracy, respectively. These results are consistent since these two regions of the genome overlap in cytoband 17q12, which contains the gene ERBB2. The models for the HER2 subtype are as follows:

$$\begin{aligned}
 L_{Betti} &= -2.262 + 3.766I_{17qs2,i}^S \\
 L_{lifespan} &= -2.303 + 20.869I_{17qs2,i}^S \\
 L_{land2} &= -2.12 + 20.69I_{17qs2,i}^S \\
 L_{land3} &= -1.925 + 3.178I_{17qs1,i}^S \\
 L_{land4} &= -1.833 + 20.399I_{17qs1,i}^S.
 \end{aligned}$$

The confusion matrices of the HER2 models are contained in Table 2.4.

To further evaluate the HER2 models, we used leave-one-out cross-validation. The average Mean square error (MSE) was 0.104 for the Betti curve model, 0.074 for the lifespan curve model, 0.102 for the second landscape model, 0.130 for the third landscape model and 0.125 for the fourth landscape model. For the HER2 subtype the lifespan curve model has the lowest MSE which agrees with the simulation results.

Luminal A

Next we build a predictive model using the significant cytobands detected for Luminal A.

In the Luminal A subtype, only Betti, lifespan and landscape 2 curves detected significant

copy number changes in segments 2q12.1-q21.1 and 5p14.3-p12 (although 2qs2 was not validated in the TCGA data set). However, only 5p14.3-p12, which we abbreviate to 5ps3, was found to have predictive power. Our results show 80% accuracy. The model from lifespan curve and third landscape curves predicting the Luminal A phenotype are

$$L_{lifespan} = -1.743 + 2.349I_{5ps3,i}^S$$

$$L_{land3} = -1.587 + 3.091I_{5ps3,i}^S.$$

The confusion matrix of this model is contained in Table 2.6.

The Betti curve logistic regression model did not include any binary predictor variables, so we do not include it.

To evaluate the Luminal A models, we used leave-one-out cross-validation. The average MSE was 0.193 for the lifespan curve model and 0.192 for the third landscape model.

Luminal B

For the Luminal B subtype, only Betti curves detected a significant segment. This meant there was no comparison to be made between predictive models from different curves, so a predictive model was not built.

Basal

In the Basal-like subtype, Betti, lifespan and landscape 4 detected many significant chromosome regions. The variables which had the most predictive power are 1p36.32-p36.11, 1p22.2-p13.3, 3p26.3-p25.1, 4q21.21-q24, 4q31.21-q32.3, 6p25.3-p22.3, 10p15.3-p12.31, 10p12.31-p11.1, 10q23.31-q25.1, 13q21.33-q31.2, 13q31.2-q34 and 15q14-q21.3 abbreviated 1ps1, 1ps10, 3ps1, 4qs4, 4qs10, 6ps1, 10ps1, 10ps3, 10qs6, 13qs6, 13qs8 and 15qs2, respectively, with models:

$$L_{Betti} = -8.425 + 2.376I_{1p10,i}^S + 3.605I_{4qs4,i}^S + 2.510I_{6ps1,i}^S + 3.507I_{10ps1,i}^S + 2.539I_{13qs6,i}^S$$

$$L_{life} = -4.791 + 2.289I_{1ps1,i}^S + 4.183I_{10ps3,i}^S + 3.711I_{13qs8,i}^S + 2.689I_{15qs2,i}^S$$

$$L_{land4} = -4.566 + 2.681I_{3ps1,i}^S + 2.331I_{4qs10,i}^S + 2.803I_{10ps3,i}^S + 2.528I_{10qs6,i}^S.$$

The confusion matrices for the Basal subtype models are contained in Table 2.8.

Table 2.8: Basal Logistic Regression: The confusion matrices and accuracy of logistic regression models built to predict the Basal phenotype from Betti, lifespan and landscape curves.

	Betti Basal		Lifespan Basal		λ_4 Basal
17	2	14	5	16	3
3	42	4	1	6	39
	Accuracy: 92%		Accuracy: 86%		Accuracy: 86%

All three models had an accuracy above 85% (Betti had 92%, lifespan had 86% and landscape 4 had 86%). Consistent with the heterogeneity of the basal subtype the models detected different regions as predictors. All three methods identified chromosome arm 10p as a predictor. Betti detected 10p1 while lifespan and landscape detected 10p3. These are consecutive regions in chromosome 10. Both Betti and Lifespan curves detected regions in chromosome 1 and in chromosome 13. Regions 1qs1 and 1qs10 are far from each other and can be considered as independent predictors. Regions 13qs6 and q8 are consecutive in the genome. Two methods also identified chromosome 4q as a predictor. Betti curves detected region 4qs4 while fourth landscapes detected 4qs10.

To further evaluate the Basal model, we used leave-one-out cross-validation. The average MSE was 0.256 for the Betti curve model, 0.210 for the lifespan curve model, and 0.185 for the fourth landscape model. Since Basal patients share many copy number aberrations with other subtypes of cancer, they tend to be more difficult to distinguish from the other subtypes. This could be the cause for higher MSE values and is a direction for future study.

2.6.5 1-Dimensional TAaCGH

A future direction of research is to expand the TAaCGH method to 1-dimensional persistent homology. Some progress was made in [7] to use 1-dimensional persistent homology to detect co-occurring copy-number aberrations. In this work they consider the generators of first persistent homology for different filtration parameters. They then map these generators back to the genome and find the particular locations of the genome that contribute to first homology. They then consider the distribution of generators for the test and control. They see that for HER2 positive patients, a higher percentage of patients have generators coming from the section of the genome containing the gene ERBB2. One potential issue

with this method is that the representative cycles of homology are not necessarily minimal. This means the persistent homology software used to calculate generators could produce different homologous generators if the order of points in the point cloud matrix is changed. This could potentially affect the distribution of generators. Recent methods to find minimized cycle representatives with respect to a loss function could potentially be used to mitigate this issue [77].

In theory, 1-dimensional persistent homology can be directly substituted into the TAaCGH pipeline. But Betti curves end up detecting far more significant regions than expected for most of the subtypes.

1-dimensional Betti curves are more complicated than 0-dimensional Betti curves. For example, 0-dimensional Betti curves are monotonic because a filtration must have the same number of connected components or less as the filtration parameter increases. That property does not hold for 1-dimensional Betti curves. Similarly, the condition for two points to be in a connected component of a Vietoris-Rips filtration at filtration parameter ϵ is much simpler than for being in a 1-cycle. Two points x and y in a point cloud X are contained inside the same connected component of $VR(X, \epsilon)$ if there exists a sequence of points $x = z_1, \dots, y = z_n$ where $d(z_i, z_{i+1}) \leq 2\epsilon$ for all $i \in [n]$. The condition (for simplicity) for just 4 points x_1, x_2, x_3, x_4 to form a 1-cycle in $VR(X, \epsilon)$ is that $d(x_i, x_{i+1}) \leq 2\epsilon$ for all i where $i + 1$ is taken modulo 4 and that $d(x_1, x_3), d(x_2, x_4) > \epsilon$. This can be seen in Figure 2.7 between $\epsilon = 3$ and $\epsilon = \frac{\sqrt{52}}{2}$ labeling the points of the rectangle clockwise from the top left as x_1, \dots, x_4 .

The additional complexity of β_1 curves makes it much more difficult to distinguish significant regions which are driven by the test as opposed to the control. For example, consider the β_0 and β_1 curves pictured in Figure 2.22. In the β_0 curve, the test curve remains mostly above the control curve, indicating that the significant difference between the curves is driven by the Basal subtype. For the β_1 curves there are many intersections between the test and the controls and it is unclear which curve might drive a significant difference in this case. This is because the genesis of a 1-cycle is more complicated and its interpretation within a sliding window point cloud is not clear.

These theoretical difficulties become practical difficulties when TAaCGH is implemented on the Horlings data set using 1-dimensional persistence. The significant sections detected by β_1 curves are shown in Table 2.6.5. Too many regions are detected for each subtype and there is a lot of overlap between subtypes. For example, aberrations in the arm 17q are detected for both the Luminal A and HER2 subtypes. The 1-dimensional lifespan

Table 2.9: Genomic regions detected by betti curves in the Horlings data set using the TAaCGH method with 1-dimensional persistent homology. Too many regions are detected and there is also overlap between subtypes.

	LumA	Basal	HER2
β_1	<p>1p36.32-p36.11, 1p36.22-p35.1, 1p36.11-p34.2, 1p35.1-p33, 1p34.2-p32.1, 1p32.3-p31.1, 1p32.1-p31.1, 1p31.1-p22.2, 1p22.2-p13.3, 2p15-p11.2, 3p26.1-p24.3, 3q24-q26.2, 4p15.1-p11, 4q24-q27, 4q25-q28.3, 4q27-q31.21, 6p21.33-p11.2, 9p24.3-p22.3, 9p23-p21.3, 9p22.2-p21.1, 9p21.3-p11.2, 9q21.13-q22.1, 9q21.33-q22.32, 9q31.1-q33.1, 9q32-q33.3, 10q21.2-q23.1, 10q24.2-q26.11, 10q25.2-q26.3, 11p15.5-p15.1, 11q22.1-q23.2, 12p13.33-p12.3, 13q13.3-q21.1, 14q24.1-q31.2, 15q14-q21.3, 17q21.2-q21.33, 20q11.21-q13.12, 20q11.23-q13.2, 23p22.33-p21.3, 23p22.2-p11.4</p>	<p>1p36.32-p36.11, 1p36.22-p35.1, 1p36.11-p34.2, 1p35.1-p33, 1p32.3-p31.1, 1p32.1-p31.1, 1p22.2-p13.3, 1p21.2-p12, 1q23.1-q31.1, 2p23.2-p16.3, 2p15-p11.2, 3p26.3-p25.1, 3p26.1-p24.3, 3p25.1-p23, 3p22.1-p14.3, 3p21.2-p13, 3p14.2-p11.2, 4p15.1-p11, 4q13.3-q22.1, 4q21.21-q24, 4q22.1-q25, 4q24-q27, 4q25-q28.3, 4q27-q31.21, 4q28.3-q31.3, 4q31.21-q32.3, 4q31.3-q34.1, 4q32.3-q35.2, 5p15.33-p15.1, 5p15.2-p13.3, 5p14.3-p12, 5q11.1-q13.1, 6p25.3-p22.3, 6p24.2-p22.1, 6p21.33-p11.2, 6q24.1-q25.3, 6q25.1-q27, 7p21.3-p14.2, 9p24.3-p22.3, 9p23-p21.3, 9q21.13-q22.1, 9q21.33-q22.32, 9q32-q33.3, 9q33.1-q34.3, 10p15.3-p12.31, 10p14-p12.1, 10p12.31-p11.1, 10q11.21-q21.2, 10q21.1-q22.1, 10q22.2-q23.31, 10q23.1-q24.2, 10q23.31-q25.1, 10q24.2-q26.11, 10q25.2-q26.3, 11q22.1-q23.2, 12p13.33-p12.3, 12p13.31-p11.21, 13q12.2-q14.13, 13q13.3-q21.1, 13q14.2-q21.32, 13q21.1-q22.3, 13q21.33-q31.2, 13q31.2-q34, 14q12-q21.3, 14q21.1-q23.1, 14q24.1-q31.2, 14q24.3-q32.13, 14q31.3-q32.33, 15q11.2-q15.3, 15q14-q21.3, 15q21.1-q22.31, 15q23-q26.3, 18q11.1-q12.3, 18q12.1-q21.2, 18q12.3-q21.33, 18q21.2-q23, 23p22.33-p21.3, 23p22.2-p11.4, 23p21.3-p11.21</p>	<p>17q11.1-q12, 17q12-q21.31, 17q21.31-q22</p>

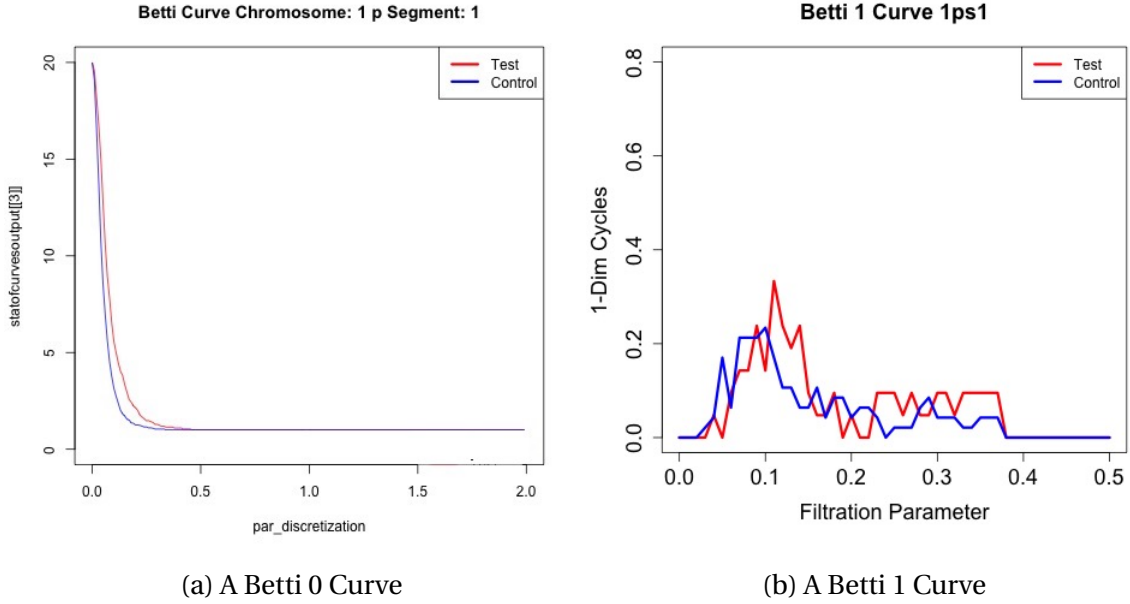


Figure 2.22: β_0 and β_1 curves for Basal on chromosome section 1p36.32 – p36.11.

Table 2.10: Genomic regions detected in the Horlings data set by lifespan curves using the TAAcGH method with 1-dimensional persistent homology.

	LumA	Basal	HER2
ℓ_1		1p36.22-p35.1, 1p21.2-p12, 4q21.21-q24, 4q25-q28.3, 5p14.3-p12, 6p21.33-p11.2, 10p15.3-p12.31, 10q23.31-q25.1, 13q13.3-q21.1	17q12-q21.31

curve detects significantly fewer regions, but the difficulty of determining whether significance is driven by the test or the control remains. The regions detected by ℓ_1 curves are in Figure 2.6.5. 1-dimensional persistent landscape curves also detect fewer regions than 1-dimensional Betti curves. Since there are not many 1-dimensional cycles that appear in small point clouds, λ_k for $k > 1$ did not detect significant regions. The regions that were detected by λ_1 are in Table 2.6.5.

Table 2.11: Genomic regions detected in the Horlings data set by first landscape curves using the TAaCGH method with 1-dimensional persistent homology.

	LumA	Basal	HER2
λ_1		1p36.22-p35.1, 1p21.2-p12, 4q13.3-q22.1, 4q25-q28.3, 5p14.3-p12, 6p21.33-p11.2, 10p15.3-p12.31, 10q11.21-q21.2, 13q13.3-q21.1	17q12-q21.31

2.6.6 Narrowing Genomic Regions to Genes

The TAaCGH method has been used to associate regions of the genome with the four subtypes of breast cancer. But it does not associate specific genes to those subtypes. To narrow these regions to specific genes, we consider the TCGA BRCA cohort which has copy number data and gene expression data at the gene level of resolution. This subsection contains preliminary analysis and further analysis is the subject of future work.

To detect significant genes, we first fix a test subtype of breast cancer filter the TCGA BRCA data by genes that are contained in genomic regions that were detected as significant for that subtype in the Horlings data set. Next, for each gene we compute the mean copy number measurements for the test subtype μ_t and for the other breast cancer subtypes μ_c . We then perform permutation tests on the statistic $|\mu_t - \mu_c|$ to detect significant genes using FDR correction for multiple testing and a significance threshold of .05. We repeat this process for all subtypes and then repeat with gene expression data as well.

The particular statistical tests used here could be improved in future work. Here copy number data and gene expression data are treated as independent variables, but they are actually related. One potential improvement could be to consider the correlation between copy number data and gene expression data for each gene across all patients in the test set and the control set. Then do a permutation test on the difference in correlations.

UCSF 500 Genes in Significant Horlings Regions

First, UCSF 500 genes from genomic regions that were detected in the Horlings data set were considered. This serves three main purposes: validating the genomic regions that were detected, narrowing those regions to specific genes and paving the road for further analysis of genes that are not in the UCSF 500 panel.

For the Luminal A subtype there were 8 candidate genes that are from a significant region detected in the Horlings data set and in the UCSF 500 and contained in the TCGA data set. From these 8 genes 6 genes had significant copy number and gene expression compared to the other subtypes. These genes were PAX8, GLI2, FGF10, RICTOR, NIPBL and DROSHA.

The Luminal B subtype contained 7 candidate UCSF 500 genes. Five of these genes had significant copy number and gene expression for the Luminal B subtype in comparison to the other subtypes combined. These genes were FGFR1, KAT6A, DUSP4, ASH2L and ZNF703.

The HER2 subtype contained 17 candidate UCSF 500 genes and 9 of these genes had significant copy number and gene expression data for the HER2 subtype compared to the other combined subtypes. These genes were ERBB2, CDK12, RAD51D, IKZF3, EZH1, TOP2A, STAT3, HOXB13 and RNF43.

Lastly, the Basal subtype contained 144 candidate genes. Of these candidate genes 96 had significant copy number and gene expression in comparison to the other combined subtypes. These genes are listed in Table 2.6.6.

Genes in Significant Horlings Regions

We also considered genes that lie in significant regions from the Horlings data set that are not in the UCSF 500. Thousands of genes lie in these regions for some of the subtypes of breast cancer. We assume that because these significant regions were detected using copy number data, that any significant genes we detect are driven by copy number.

Cancer-related genes can be split into three categories: oncogenes, tumor suppressor genes and genes related to either oncogenes or tumor suppressor genes. Oncogenes are genes that stimulate cell growth, tumor suppressor genes are genes which inhibit cell growth and other cancer genes work in concert with either oncogenes or tumor suppressor genes. Oncogenes driven by copy number changes have high copy number and high gene expression. Tumor suppressor genes have low copy number and low gene expression. This is summarized in Table 2.6.6. Since there are many more genes not contained in the UCSF 500 within the Horlings regions, we need to filter genes to make preliminary analysis more manageable. We decide to focus on detecting candidate oncogenes. To do this, we consider only genes for which the TCGA data set has expression and copy number data. We also only consider genes that have a higher mean gene expression value for the test subtype than

Table 2.12: Genes that were detected as significant for the basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set and are also UCSF 500 genes.

Basal
GATA3, LARP4B, ARID5B, PTEN, SUFU, NT5C2, SMC3, TCF7L2, CDKN1B, ETV6, CHD4, FLT3, FLT1, BRCA2, FOXO1, CYSLTR2, RB1, TSHR, FOS, MLH3, DICER1, TRAF3, AKT1, GREM1, NUTM1, RAD51, KNSTRN, TCF12, MAP2K1, SETBP1, NRAS, GFI1, GLMN, JAK1, CDKN2C, MUTYH, PTCH2, MYCL, MPL, CSF3R, RRAGC, ARID1A, ID3, CDC42, SDHB, ERFFI1, CAMTA1, TNFRSF14, MCL1, ETV3, SDHC, DDR2, SMC1A, GATA1, ARAF, RBM10, GPC3, PHF6, ATP6AP1, XPO1, MSH6, EPCAM, MSH2, SOS1, ALK, FOXP1, MITE, PBRM1, BAP1, PHOX2B, TET2, LEF1, MAML3, INPP4B, FBXW7, TERT, SDHA, IL6ST, MAP3K1, PIK3R1, VEGFA, NFKBIE, HSP90AB1, TFEB, CCND3, CDKN1A, FANCE, DAXX, DDR1, IRF4, ESR1, ARID1B, IGF2R, PDCD1LG2, CD274, SMARCA2

Table 2.13: The general signatures of oncogenes and tumor suppressor genes that are driven by copy number. Oncogenes have high copy number and high gene expression and tumor suppressor genes have low copy number and low gene expression.

	Low Expression	High Expression
Low Copy Number	Tumor Suppressor, Cancer-Related Gene	Not Driven By Copy-Number
High Copy Number	Not Driven By Copy-Number	Oncogene Cancer-Related Gene

Table 2.14: Candidate oncogenes that were detected as significant for the HER2 subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.

HER2
RPL19, PPP1R1B, ORMDL3, PSMD3, MIEN1, GRB7, PGAP3, PSMB3, RPL23, STARD3, TOB1, PRR15L, PNMT, LASP1, LRRC59, MED24, MED1, MRPL45, UBE2Z, CASC3, MSL1, POLDIP2, SCPEP1, CBX1, RPL23A, EPN3, CISD3, PHB, HOXB2, GSDMB, PIP4K2B, MRPL27, CDC6, TRAF4, SLC35B1, UTP18, WIPF2, ABCC3, PPP1R9B, CALCOCO2, PCGF2, GIT1, SKA2, TNFAIP1, NUFIP2, ERAL1, NME1, HOXB7, ZNF652, ACSF2, MIR4728, CWC25, ITGA3, ACACA, TLCD1, TMEM97, B4GALNT2, SPAG5, MLLT6, AATE, PCTP, SNF8, THRA, SRCIN1, TRIM37, TAOK1, KIAA0100, ANKRD40, FBXL20, SPATA20, MYO18A, RSAD1, KAT7, XYLT2, SPAG9, TP53I13, SOCS7, NSRP1, SUPT6H, SDF2, HOXB9, SMG8, SNX11, YPEL2, ABHD15, PHF12, COIL, SSH2, IFT20, FAM222B, CA10, EME1, ARHGAP23, DLX3, PIPOX, NEUROD2, HOXB6, DHRS13, TMEM92, LINC00672, GSDMA, ZBPB2, SNORD124, DLX4, LRRC3C, GPR179, NME2, RNF126P1, EFCAB5, TIAF1, C17orf98, FBXO47, WFIKKN2, LRRC37A11P, CRYBA1, ARL5C, TBC1D3P5

the other subtypes and similarly for copy-number. Lastly, we only choose genes that have significant copy number and gene expression compared to the other subtypes.

Within the HER2 regions detected in the Horlings data set there are 576 candidate genes. After filtering as described above there are 118 candidate oncogenes. When ordered by the magnitude of the difference in average gene expression between the test subtype and the other subtypes the top 10 genes are RPL19, PPP1R1B, ORMDL3, PSMD3, MIEN1, GRB7, PGAP3, PSMB3, RPL23 and STARD3. The last gene, STARD3, was shown to have prognostic significance in HER2+ breast cancer [52]. It has also been proposed as a potential target for cancer therapy [10]. The full list of candidate oncogenes is contained in Table 2.6.6.

Within the Luminal A regions detected in the Horlings data set there were 2010 candidate genes. After filtering as described above there were no candidate oncogenes detected. This agrees with attempts at validating the Luminal A regions using the TCGA BRCA cohort. Though there is a clear signal in Luminal A patients in the Horlings data set, the particular

Table 2.15: Candidate oncogenes that were detected as significant for the Luminal B subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.

Luminal B
EIF4EBP1, VDAC3, RAB11FIP1, ERLIN2, LSM1, BAG4, DDHD2, TM2D2, BRF2, AP3M2, ADAM9, SMIM19, RPS20P22, ZMAT4, ANK1, RNF170, LETM2, C8orf86, KCNU1, RN7SL709P, RNA5SP264, RN7SL149P

detected aberration seems to be rare in the TCGA BRCA cohort.

Within the Luminal B regions detected in the Horlings data set there were 208 candidate genes. After filtering there were 22 candidate oncogenes. These genes are listed in Table 2.6.6. The first gene in the table EIF4BP1 or Eukaryotic Initiation Factor 4E-Binding Protein, has been previously identified as an oncogene, specifically in luminal type breast cancers [100].

For the Basal subtype there were 7266 candidate genes. After filtering to find candidate oncogenes we found 1323 candidate oncogenes. Since there were so many candidate genes they were split into multiple tables and placed in the appendix. The genes are in Table A, Table A, Table A, Table A and Table A. The second gene, TMSB10 has been identified as a regulator of tumorigenesis and has been specifically associated with the Basal and HER2 subtypes [120].

This analysis is preliminary and these lists of genes need further exploration. Only candidate oncogenes have been considered, but tumor suppressor genes should also be considered. Furthermore, the signatures of UCSF 500 genes can be used to help categorize non-UCSF 500 genes into oncogenes, tumor suppressor genes and other cancer-related genes.

2.7 Discussion

In this chapter, we consider the stability of Betti curves and expand the TAaCGH method by incorporating lifespan and persistent landscape curves. We then compare the performance of these at identifying copy number changes through simulations and apply them to the Horlings dataset. On simulated data, lifespan curves outperform Betti and persistent landscape curves. On the Horlings dataset, all persistence curves are similarly success-

Table 2.16: The cytobands corresponding to the new regions detected by Betti curves, lifespan curves, and landscape curves compared to the regions detected in [9]. Red indicates new cytobands associated with the Basal subtype, blue with Luminal A and gray Luminal B.

Cytoband Ranges of Newly Detected Segments	
Betti-0	2p23.2-p16.3, 8p22-p11.1
Lifespan-0	2p23.2-p16.3, 2q12.1-q21.1, 5p14.3-p12
Landscape	λ_3 : 2q12.1-q21.1, 5p14.3-p12, λ_4 : 1q21.1-q25.2, 23q26.2-q28

ful at associating chromosome arm segments to phenotypes of breast cancer. Across the four phenotypes, different curves detect different segments, suggesting a complementary approach using the three different methods may provide the most information. From a theoretical perspective, the fact that Betti curves perform comparably to more stable curves in both simulations and on real data is unexpected. It does, however, match results in [11] where Betti curves were shown to be more resistant to certain kinds of noise for the task of image classification than other persistence curves. The cytoband ranges corresponding to the newly detected segments are pictured in Table 2.16 organized by type of persistence curve. The ranges are colored by subtype: red for Basal, blue for Luminal A and gray for Luminal B. All of the newly detected segments were the result of copy number gains except for 2p23.2-p16.3 for the Basal subtype which was inconclusive for being driven by a gain or a loss.

Within the HER2 phenotype, Betti curves detect 17q11.1-q21.31 and 17q21.31-q22, lifespan curves detect 17q11.1-q22 and persistence landscapes all miss one of the four segments in 17q, either 17q11.1-q22 or 17q21.31-q22. Importantly, in each case they detect 17q12 which contains ERBB2, a well-known driver gene for HER2+ breast cancer.

Three chromosome segments were associated with the Luminal A phenotype: 2q12.1-q21.1, 5p14.3-p12 and 11q22.1-q23.2. For this phenotype, persistent landscapes detected all three segments and Betti and lifespan curves detect a subset of these segments. Gains in chromosome arm 11q have been associated with the Luminal A subtype [32], though in different cytobands from the ones detected here. None of the three regions were validated by the TCGA BRCA cohort dataset. This could be due to the fact that only a small percentage of patients (less than 15% with the Luminal A phenotype) have aberrations in any chromosome arm. The highest percentage of aberrations in chromosome arms in the TCGA BRCA cohort

dataset occur in arms 1, 8, 11, 17 and 20 which are also the arms with the largest spikes in other phenotypes. In the Horlings dataset, patients in the control set have some type of breast cancer. This means that if multiple phenotypes have a similar signature, TAaCGH may miss these regions.

In the Luminal B phenotype, only one significant segment was detected: 8p22-p11.1. It was detected by Betti curves, but not the other persistence curves. Part of this region was also identified and associated with Luminal B in [39]. In the Horlings dataset most Luminal B patients had a copy number gain of 8p11.21-8p11.23 and a loss of 8p12-8p22. A large number of patients have this same signature in the TCGA BRCA cohort dataset. In general, detecting segments for the Luminal B phenotype is difficult using TAaCGH, because many of its aberrations are shared with the Basal phenotype. To deal with this problem, we removed the Basals from the control group and ran the same experiment. Many significant segments were then detected by the three curves.

For the Basal phenotype, Betti curves detect the most significant segments, followed by lifespan curves and then landscape curves. Only fourth landscape curves detect significant segments. The TCGA BRCA cohort dataset validates many of the detected cytobands. Figure 2.19 shows the many significant sections identified and also indicates the difficulty of detecting the Basal phenotype since it has many different aberrations which are shared with other phenotypes. In spite of these difficulties, the TAaCGH method identified three new chromosome segments 1q21.1-q25.2, 2p23.2-p16.3 and 23q26.2-q28 associated with the Basal phenotype that are confirmed by the the TCGA BRCA cohort dataset.

The logistic regression predictive models perform fairly similarly across all persistence curves, differing by a few percentage points in accuracy within each phenotype. Even though full chromosome arms were used as potential predictor variables together with chromosome segments, only chromosome segments were used in the final logistic regression models.

Within the HER2 phenotype, the logistic regression equations from Betti, lifespan and second landscape curves all chose the 17q12-q21.31 predictor variable which contains the ERBB2 gene. The third and fourth landscape curves, however, chose 17q11.1-q12. These models are both slightly less accurate than the models that use 17q12-q21.31.

For the Luminal A phenotype lifespan curves and third landscape curves performed with 80% and 83% accuracy on the Horlings dataset. The logistic regression models for each curve used the binary predictor variable associated with 5p14.3-p12.

For the Basal subtype the predictor variables in the models differed significantly from

each other. The only repeated predictor variables were 10p15.3-p12.31 and 10p12.31-p11.1. These are adjacent in the genome. In all cases the accuracy is above 85%.

We evaluated the logistic regression models using leave-one-out cross-validation. For the HER2 models and Luminal A models the MSE values were low. In the HER2 case the lifespan curve model had the lowest MSE matching the simulation results. The Basal models had the highest MSE values which could be due to the heterogeneity of the copy number changes in the Basal subtype.

Next we investigated whether the detected cytobands harbor cancer-related genes. We performed a literature search and also consulted the Sanger Cancer Data Base (COSMIC) [110]. In the Basal subtype, cytoband 1q21.1-q25.2 contains cancer genes *HORMAD1*, *LOC92312*, *SNG5*, *TMEM79*, *CCT3*, *IQGAP3*, *HDGF*, *PRCC* in [4], cytoband 2p23.2-p16.3 contains cancer genes *PLB1* and *WDR43* [40, 84] and cytoband 23q26.2-q28 contains the cancer genes *ISR4* and *FLNA* [68, 92]. In the case of the Luminal A subtype, cytoband 2q12.1-q21.1 contains gene *ECRG4* [101]. Cytoband 5p 14.3-12 contains *TERT* [47, 43, 54] and gains in this cytoband have been associated with this breast cancer [94] and also with recurrence [14]. In the Luminal B subtype, we found cytoband 8p22-p11. These cytobands are commonly associated with the Luminal subtype and contain genes *ZNF703*, *PROSC*, *BRF2*, *RAB11FIP1*, *ASH2L*, *DDHD2*, *LETM2* in [4].

REFERENCES

- [1] Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vestein 31 Zhang Wei 33 Shmulevich Ilya 31, et al. Comprehensive molecular portraits of human breast tumours. Nature, 490(7418):61–70, 2012.
- [2] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. The Journal of Machine Learning Research, 18(1):218–252, 2017.
- [3] Henry Adams, Andrew Tausz, and Mikael Vejdemo-Johansson. Javaplex: A research software package for persistent (co) homology. In International congress on mathematical software, pages 129–136. Springer, 2014.
- [4] José Adélaïde, Pascal Finetti, Ismahane Bekhouche, Laetitia Repellini, Jeannine Geneix, Fabrice Sircoulomb, Emmanuelle Charafe-Jauffret, Nathalie Cervera, Jérôme Desplans, Daniel Parzy, et al. Integrated profiling of basal and luminal breast cancers. Cancer research, 67(24):11565–11575, 2007.
- [5] Felipe Ades, Dimitrios Zardavas, Ivana Bozovic-Spasojevic, Lina Pugliano, Debora Fumagalli, Evandro De Azambuja, Giuseppe Viale, Christos Sotiriou, and Martine Piccart. Luminal b breast cancer: molecular characterization, clinical management, and future perspectives. Journal of Clinical Oncology, 32(25):2794–2803, 2014.
- [6] Martin Aigner, Günter M Ziegler, Karl H Hofmann, and Paul Erdos. Proofs from the Book, volume 274. Springer, 2010.
- [7] Sergio Ardanza-Trevijano, Georgina Gonzalez, Tyler Borrman, Juan Luis Garcia, and Javier Arsuaga. Topological analysis of amplicon structure in comparative genomic hybridization (cgh) data: an application to erbb2/her2/neu amplified tumors. In International Workshop on Computational Topology in Image Context, pages 113–129. Springer, 2016.
- [8] Edurne Arriola, Caterina Marchio, David SP Tan, Suzanne C Drury, Maryou B Lambros, Rachael Natrajan, Socorro Maria Rodriguez-Pinilla, Alan Mackay, Narinder Tamber, Kerry Fenwick, et al. Genomic analysis of the her2/top2a amplicon in breast cancer and breast cancer cell lines. Laboratory investigation, 88(5):491–503, 2008.

- [9] Javier Arsuaga, Tyler Borrmann, Raymond Cavalcante, Georgina Gonzalez, and Catherine Park. Identification of copy number aberrations in breast cancer subtypes using persistence topology. Microarrays, 4(3):339–369, 2015.
- [10] Kanwal Asif, Lorenzo Memeo, Stefano Palazzolo, Yahima Frión-Herrera, Salvatore Parisi, Isabella Caligiuri, Vincenzo Canzonieri, Carlotta Granchi, Tiziano Tuccinardi, and Flavio Rizzolio. Stard3: A prospective target for cancer therapy. Cancers, 13(18):4693, 2021.
- [11] Nieves Atienza, Rocío Gonzalez-Diaz, and M Soriano-Trigueros. A new entropy based summary function for topological data analysis. Electronic Notes in Discrete Mathematics, 68:113–118, 2018.
- [12] Dror Bar-Natan. On khovanov’s categorification of the jones polynomial. Algebraic & Geometric Topology, 2(1):337–370, 2002.
- [13] Vladimir Baranovsky and Radmila Sazdanovic. Graph homology and graph configuration spaces. Journal of Homotopy and Related Structures, 7(2):223–235, 2012.
- [14] Stephan Bartels, Jana Lisa van Luttikhuisen, Matthias Christgen, Lavinia Mägel, Angelina Luft, Sonja Hänzelmann, Ulrich Lehmann, Brigitte Schlegelberger, Fabian Leo, Doris Steinemann, et al. Cdkn2a loss and pik3ca mutation in myoepithelial-like metaplastic breast cancer. The Journal of Pathology, 245(3):373–383, 2018.
- [15] Ulrich Bauer. Ripser: efficient computation of vietoris–rips persistence barcodes. Journal of Applied and Computational Topology, 5(3):391–423, 2021.
- [16] George D Birkhoff. A determinant formula for the number of ways of coloring a map. The Annals of Mathematics, 14(1/4):42–46, 1912.
- [17] Raoul Bott. Two new combinatorial invariants for polyhedra. Portugaliae mathematica, 11(1):35–40, 1952.
- [18] Francesco Brenti. Hilbert polynomials in combinatorics. Journal of Algebraic Combinatorics, 7(2):127–156, 1998.
- [19] Alain Bretto. Hypergraph theory. An introduction. Mathematical Engineering. Cham: Springer, 2013.
- [20] Felix Breuer, Aaron Dall, and Martina Kubitzke. Hypergraph coloring complexes. Discrete mathematics, 312(16):2407–2420, 2012.
- [21] broadinstitute. Broad gdac firehose.
- [22] Peter Bubenik. Statistical topological data analysis using persistence landscapes. The Journal of Machine Learning Research, 16(1):77–102, 2015.

- [23] Anuraag Bukkuri, Noemi Andor, and Isabel K Darcy. Applications of topological data analysis in oncology. Frontiers in artificial intelligence, 4:659037, 2021.
- [24] Gunnar Carlsson. Topology and data. Bulletin of the American Mathematical Society, 46(2):255–308, 2009.
- [25] Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In International Conference on Artificial Intelligence and Statistics, pages 2786–2796. PMLR, 2020.
- [26] Alex Chandler. Thin posets, cw posets, and categorification. arXiv preprint arXiv:1911.05600, 2019.
- [27] Alex Chandler and Radmila Sazdanovic. A broken circuit model for chromatic homology theories. European Journal of Combinatorics, 104:103538, 2022.
- [28] Frédéric Chazal, Vin De Silva, and Steve Oudot. Persistence stability for geometric complexes. Geometriae Dedicata, 173(1):193–214, 2014.
- [29] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Frontiers in artificial intelligence, 4:667963, 2021.
- [30] Michael Chmutov, Sergei Chmutov, and Yongwu Rong. Knight move in chromatic cohomology. European Journal of Combinatorics, 29(1):311–321, 2008.
- [31] Yu-Min Chung and Austin Lawson. Persistence curves: A canonical framework for summarizing persistence diagrams. Advances in Computational Mathematics, 48(1):1–42, 2022.
- [32] Giovanni Ciriello, Rileen Sinha, Katherine A Hoadley, Anders S Jacobsen, Boris Reva, Charles M Perou, Chris Sander, and Nikolaus Schultz. The molecular diversity of luminal a breast tumors. Breast cancer research and treatment, 141(3):409–420, 2013.
- [33] J Climent, JL Garcia, JH Mao, J Arsuaga, and J Perez-Losada. Characterization of breast cancer by array comparative genomic hybridization. Biochemistry and Cell Biology, 85(4):497–508, 2007.
- [34] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. Discrete & computational geometry, 37(1):103–120, 2007.
- [35] Andrew A Cooper, Vin de Silva, and Radmila Sazdanović. On configuration spaces and simplicial complexes. New York J. Math, 25:723–744, 2019.

- [36] Andrew A Cooper, Joseph Mueller, and Radmila Sazdanović. The chromatic polynomial of a simplicial complex. [preprint](#), 2020.
- [37] Andrew A Cooper and Radmila Sazdanović. State sum homology for simplicial complexes. [preprint](#), 2020.
- [38] Colin S Cooper. Applications of microarray technology in breast cancer research. *Breast Cancer Research*, 3(3):1–18, 2001.
- [39] Stéphanie Cornen, Arnaud Guille, José Adélaïde, Lynda Addou-Klouche, Pascal Finetti, Marie-Rose Saade, Marwa Manai, Nadine Carbuccion, Ismahane Bekhouche, Anne Letessier, et al. Candidate luminal b breast cancer genes identified by genome, gene expression and dna methylation profiling. *PloS one*, 9(1):e81843, 2014.
- [40] Fergus J Couch, Karoline B Kuchenbaecker, Kyriaki Michailidou, Gustavo A Mendoza-Fandino, Silje Nord, Janna Lilyquist, Curtis Olswold, Emily Hallberg, Simona Agata, Habibul Ahsan, et al. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. *Nature communications*, 7(1):1–13, 2016.
- [41] Curtis R Coughlin, Gunter H Scharer, and Tamim H Shaikh. Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome medicine*, 4(10):1–12, 2012.
- [42] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.
- [43] Edaise M da Silva, Pier Selenica, Mahsa Vahdatinia, Fresia Pareja, Arnaud Da Cruz Paula, Lorenzo Ferrando, Andrea M Gazzo, Higinio Dopeso, Dara S Ross, Ariya Bakhteri, et al. Tert promoter hotspot mutations and gene amplification in metaplastic breast cancer. *NPJ breast cancer*, 7(1):1–8, 2021.
- [44] Daniel DeWoskin, Joan Climent, I Cruz-White, Mariel Vazquez, Catherine Park, and Javier Arsuaga. Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology and its Applications*, 157(1):157–164, 2010.
- [45] Paweł Dłotko and Davide Gurnari. Euler characteristic curves and profiles: a stable shape invariant for big data problems. [arXiv preprint arXiv:2212.01666](#), 2022.
- [46] Klaus Dohmen and Martin Trinks. An abstraction of whitney’s broken circuit theorem. [arXiv preprint arXiv:1404.5480](#), 2014.

- [47] Marta Dratwa, Barbara Wysoczanska, Wioletta Brankiewicz, Martyna Stachowicz-Suhs, Joanna Wietrzyk, Rafał Matkowski, Marcin Ekiert, Jolanta Szelachowska, Adam Maciejczyk, Mariusz Szajewski, et al. Relationship between telomere length, tert genetic variability and tert, tp53, sp1, myc gene co-expression in the clinicopathological profile of breast cancer. International journal of molecular sciences, 23(9):5164, 2022.
- [48] Michael Eastwood and Stephen Huggett. Euler characteristics and chromatic polynomials. European Journal of Combinatorics, 28(6):1553–1560, 2007.
- [49] Herbert Edelsbrunner and John L Harer. Computational topology: an introduction. American Mathematical Society, 2022.
- [50] Brent Everitt and Paul Turner. Cellular cohomology of posets with local coefficients. Journal of Algebra, 439:134–158, 2015.
- [51] Edward Fadell and Lee Neuwirth. Configuration spaces. Mathematica Scandinavica, 10:111–118, 1962.
- [52] Abdul Fattah Salah Fararjeh, Ali Al Khader, Ezidin Kaddumi, Maher Obeidat, et al. Differential expression and prognostic significance of stard3 gene in breast carcinoma. International Journal of Molecular and Cellular Medicine, 10(1):34, 2021.
- [53] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. Introduction to the r package tda. arXiv preprint arXiv:1411.1830, 2014.
- [54] Mathilde Gay-Bellile, Lauren Veronese, Patricia Combes, Eléonore Eymard-Pierre, Fabrice Kwiatkowski, Marie-Mélanie Dauplat, Anne Cayre, Maud Privat, Catherine Abrial, Yves-Jean Bignon, et al. Tert promoter status and gene copy number gains: effect on tert expression and association with prognosis in breast cancer. Oncotarget, 8(44):77540, 2017.
- [55] Maya Ghoussaini and Paul DP Pharoah. Polygenic susceptibility to breast cancer: current state-of-the-art. Future oncology, 5(5):689–701, 2009.
- [56] Michael Goff. Extremal betti numbers of vietoris–rips complexes. Discrete & Computational Geometry, 46(1):132–155, 2011.
- [57] Georgina Gonzalez, Arina Ushakova, Radmila Sazdanovic, and Javier Arsuaga. Prediction in cancer genomics using topological signatures and machine learning. In Topological Data Analysis, pages 247–276. Springer, 2020.
- [58] Allen Hatcher. Algebraic topology. Cambridge University Press,, 2005.
- [59] Thorkell Helgason. Aspects of the theory of hypermatroids. In Hypergraph seminar, pages 191–213. Springer, 1974.

- [60] L. Helme-Guizon and Y. Rong. A categorification for the chromatic polynomial. Algebraic & Geometric Topology, 5(4):1365–1388, 2005.
- [61] Laure Helme-Guizon, Józef H Przytycki, and Yongwu Rong. Torsion in graph homology. arXiv preprint math/0507245, 2005.
- [62] Laure Helme-Guizon and Yongwu Rong. Graph cohomologies from arbitrary algebras. arXiv preprint math/0506023, 2005.
- [63] Gregory Henselman and Robert Ghrist. Matroid filtrations and computational persistent homology. arXiv preprint arXiv:1606.00199, 2016.
- [64] Patricia Hersh and Ed Swartz. Coloring complexes and arrangements. Journal of Algebraic Combinatorics, 27(2):205, 2008.
- [65] Jürgen Herzog, Vic Reiner, and Volkmar Welker. The koszul property in affine semi-group rings. Pacific Journal of Mathematics, 186(1):39–65, 1998.
- [66] Hugo M Horlings, Carmen Lai, Dimitry SA Nuyten, Hans Halfwerk, Petra Kristel, Erik van Beers, Simon A Joosse, Christiaan Klijn, Petra M Nederlof, Marcel JT Reinders, et al. Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients. Clinical Cancer Research, 16(2):651–663, 2010.
- [67] Axel Hultman. Link complexes of subspace arrangements. European Journal of Combinatorics, 28(3):781–790, 2007.
- [68] Gerjon J Ikin, Mandy Boer, Elvira RM Bakker, and John Hilkens. Irs4 induces mammary tumorigenesis and confers resistance to her2-targeted therapy through constitutive pi3k/akt-pathway hyperactivation. Nature communications, 7(1):1–15, 2016.
- [69] Edna F Jasso-Hernandez and Yongwu Rong. A categorification for the tutte polynomial. Algebraic & Geometric Topology, 6(5):2031–2049, 2006.
- [70] Megan Johnson and Jae-Hun Jung. Instability of the betti sequence for persistent homology and a stabilized version of the betti sequence. arXiv preprint arXiv:2109.09218, 2021.
- [71] RP Jones. Some results of chromatic hypergraph theory proved by reduction to graphs. PROBLEMES COMBINATOIRES ET THEORIE DES GRAPHERS., 1978.
- [72] Jakob Jonsson. The topology of the coloring complex. Journal of Algebraic Combinatorics, 21(3):311–329, 2005.
- [73] Cassie N Kline, Nancy M Joseph, James P Grenert, Jessica van Ziffle, Eric Talevich, Courtney Onodera, Mariam Aboian, Soonmee Cha, David R Raleigh, Steve Braunstein, et al. Targeted next-generation sequencing of pediatric neuro-oncology patients

- improves diagnosis, identifies pathogenic germline mutations, and directs targeted therapy. Neuro-oncology, 19(5):699–709, 2017.
- [74] Aaron D Lauda and Joshua Sussan. An invitation to categorification. Notices of the American Mathematical Society, 69(01), 2022.
- [75] Mark DM Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics, 47(2):106–114, 2015.
- [76] Jean Leray. Le calcul différentiel et intégral sur une variété analytique complexe.(problème de cauchy. iii.). Bulletin de la Société mathématique de France, 87:81–180, 1959.
- [77] Lu Li, Connor Thompson, Gregory Henselman-Petrusek, Chad Giusti, and Lori Ziegelmeier. Minimal cycle representatives in persistent homology using linear programming: An empirical study with user’s guide. Frontiers in artificial intelligence, 4:681117, 2021.
- [78] Warwick J Locke, Dominic Guanzon, Chenkai Ma, Yi Jin Liew, Konsta R Duesing, Kim YC Fung, and Jason P Ross. Dna methylation cancer biomarkers: translation to the clinic. Frontiers in genetics, 10:1150, 2019.
- [79] Jane Holsapple Long and Sarah Crown Rundell. The hodge structure of the coloring complex of a hypergraph. Discrete mathematics, 311(20):2164–2173, 2011.
- [80] Adam M Lowrance and Radmila Sazdanović. Chromatic homology, khovanov homology, and torsion. Topology and its Applications, 222:77–99, 2017.
- [81] Elaine R Mardis and Richard K Wilson. Cancer genome sequencing: a review. Human molecular genetics, 18(R2):R163–R168, 2009.
- [82] Nina Mars, Elisabeth Widén, Sini Kerminen, Tuomo Meretoja, Matti Pirinen, Pietro della Briotta Parolo, Priit Palta, Aarno Palotie, Jaakko Kaprio, Heikki Joensuu, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. Nature communications, 11(1):1–9, 2020.
- [83] Jiří Matoušek, Anders Björner, and Günter M Ziegler. Using the Borsuk-Ulam theorem: lectures on topological methods in combinatorics and geometry, volume 2003. Springer, 2003.
- [84] Gustavo Mendoza-Fandiño, Paulo Cilas M Lyra, Thales C Nepomuceno, Carly M Harro, Nicholas T Woods, Xueli Li, Leticia B Rangel, Marcelo A Carvalho, Fergus J Couch,

- and Alvaro NA Monteiro. Two distinct mechanisms underlie estrogen-receptor-negative breast cancer susceptibility at the 2p23. 2 locus. European Journal of Human Genetics, 30(4):465–473, 2022.
- [85] Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Sidhartha Kar, Audrey Lemaçon, Penny Soucy, Dylan Glubb, Asha Rostamianfar, et al. Association analysis identifies 65 new breast cancer risk loci. Nature, 551(7678):92–94, 2017.
- [86] Ezra Miller and Bernd Sturmfels. Combinatorial commutative algebra, volume 227. Springer Science & Business Media, 2004.
- [87] David Moon, John Harer, and Rann Bar-On. Maximum number of nonzero persistence cycles in a vietoris-rips filtration. Private Communication.
- [88] Dmitriy Morozov. Dionysus. Software available at <http://www.mrzv.org/software/dionysus>, 2012.
- [89] Vidit Nanda. Perseus: the persistent homology software. Software available at <http://www.sas.upenn.edu/~vnanda/perseus>, 2012.
- [90] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences, 108(17):7265–7270, 2011.
- [91] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics, 5(4):557–572, 2004.
- [92] Abderrahman Ouban. Filamin-a expression in triple-negative breast cancer and its clinical significance. Biotechnology & Biotechnological Equipment, 35(1):1409–1419, 2021.
- [93] Milena D Pabiniak, Józef H Przytycki, and Radmila Sazdanović. On the first group of the chromatic cohomology of graphs. Geometriae Dedicata, 140(1):19, 2009.
- [94] Anna D Panani. Isochromosome 5p, a novel recurrent abnormality in breast cancer: is it a common abnormality in cancer? in vivo, 24(5):715–717, 2010.
- [95] Soohyun Park. Simplicial chromatic polynomials as hilbert series of stanley–reisner rings. arXiv preprint arXiv:2203.11927, 2022.
- [96] Jose A Perea and John Harer. Sliding windows and persistence: An application of topological methods to signal analysis. Foundations of Computational Mathematics, 15(3):799–838, 2015.

- [97] Raúl Rabadán, Yamina Mohamedi, Udi Rubin, Tim Chu, Adam N Alghalith, Oliver Elliott, Luis Arnés, Santiago Cal, Álvaro J Obaya, Arnold J Levine, et al. Identification of relevant genetic alterations in cancer using topological data analysis. Nature communications, 11(1):1–10, 2020.
- [98] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4741–4748, 2015.
- [99] Sarah Crown Rundell. The coloring complex and cyclic coloring complex of a complete k -uniform hypergraph. Journal of Combinatorial Theory, Series A, 119(5):1095–1109, 2012.
- [100] Alexandria C Rutkovsky, Elizabeth S Yeh, Stephen T Guest, Victoria J Findlay, Robin C Muise-Helmericks, Kent Armeson, and Stephen P Ethier. Eukaryotic initiation factor 4e-binding protein as an oncogene in breast cancer. Bmc Cancer, 19(1):1–15, 2019.
- [101] Renaud Sabatier, Pascal Finetti, Jose Adelaide, Arnaud Guille, Jean-Paul Borg, Max Chaffanet, Lydie Lane, Daniel Birnbaum, and François Bertucci. Down-regulation of *ecrg4*, a candidate tumor suppressor gene, in human breast cancer. PloS one, 6(11):e27656, 2011.
- [102] Radmila Sazdanovic and Daniel Scofield. Patterns in khovanov link and chromatic graph homology. Journal of Knot Theory and Its Ramifications, 27(03):1840007, 2018.
- [103] Radmila Sazdanovic and Martha Yip. A categorification of the chromatic symmetric function. Journal of Combinatorial Theory, Series A, 154:218–246, 2018.
- [104] Lee M Seversky, Shelby Davis, and Matthew Berger. On time-series topological data analysis: New data and opportunities. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 59–67, 2016.
- [105] Yara Skaf and Reinhard Laubenbacher. Topological data analysis in biomedicine: A review. Journal of Biomedical Informatics, page 104082, 2022.
- [106] Primož Skraba and Katharine Turner. Wasserstein stability for persistence diagrams. arXiv preprint arXiv:2006.16824, 2020.
- [107] Richard P Stanley. Enumerative combinatorics volume 1 second edition. Cambridge studies in advanced mathematics, 2011.
- [108] Einar Steingrímsson. The coloring ideal and coloring complex of a graph. Journal of Algebraic Combinatorics, 14(1):73–84, 2001.
- [109] Marko Stošić. Categorification of the dichromatic polynomial for graphs. Journal of Knot Theory and Its Ramifications, 17(01):31–45, 2008.

- [110] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. Nucleic acids research, 47(D1):D941–D947, 2019.
- [111] The GUDHI Project. GUDHI User and Reference Manual. GUDHI Editorial Board, 3.4.1 edition, 2021.
- [112] Ioan Tomescu. Chromatic coefficients of linear uniform hypergraphs. Journal of Combinatorial Theory, Series B, 72(2):229 – 235, 1998.
- [113] Martin Trinks. A note on a broken-cycle theorem for hypergraphs. Discussiones Mathematicae Graph Theory, 34(3):641–646, 2014.
- [114] ucsf. Ucsf500 cancer gene panel.
- [115] Oleg Viro. Remarks on definition of khovanov homology. arXiv preprint math/0202199, 2002.
- [116] Carl Virtanen and James Woodgett. Clinical uses of microarrays in cancer research. Clinical Bioinformatics, pages 87–113, 2008.
- [117] Manfred Walter. Some results on chromatic polynomials of hypergraphs. the electronic journal of combinatorics, pages R94–R94, 2009.
- [118] Russ Woodroffe. Chordal and sequentially cohen-macaulay clutters. arXiv preprint arXiv:0911.4697, 2009.
- [119] Lu Xian, Henry Adams, Chad M Topaz, and Lori Ziegelmeier. Capturing dynamics of time-varying data via topology. arXiv preprint arXiv:2010.05780, 2020.
- [120] Xin Zhang, Dong Ren, Ling Guo, Lan Wang, Shu Wu, Chuyong Lin, Liping Ye, Jinrong Zhu, Jun Li, Libing Song, et al. Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer. Breast Cancer Research, 19(1):1–15, 2017.

APPENDIX

APPENDIX

A

CHAPTER 2 SUPPLEMENTARY FIGURES
AND TABLES

Table A.1: Comparison of the chromosome segments detected in the original Betti-0 study [9] using JavaPlex and the new Betti-0 study using the R TDA package. The rows contain the segments that the old study detected but the new study did not for each subtype of breast cancer.

Previously Detected Segments not Detected with R TDA	
Basal	1p34.2-p32.1, 1q23.1-q31.1, 3p25.1-p23, 3p22.1-p14.3, 3p14.2-p11.2, 4q13.3-q22.1, 4q32.3-q35.2, 5p15.2-p12, 9p23-p21.3, 9q32-q34.3, 10q11.21-q21.2, 10q25.2-q26.3, 12p13.33-p12.3, 14q12-q21.3, 18q11.1-q12.3, 18q12.3-q23
Luminal B (No Basal in Control)	1p35.1-p33, 1q41-q44, 4q24-q27, 8p23.3-p22, 9p22.2-p21.1, 9q13-q22.1, 13q12.2-q21.1, 13q31.1-q32.2
HER2	17q21.2-q21.33

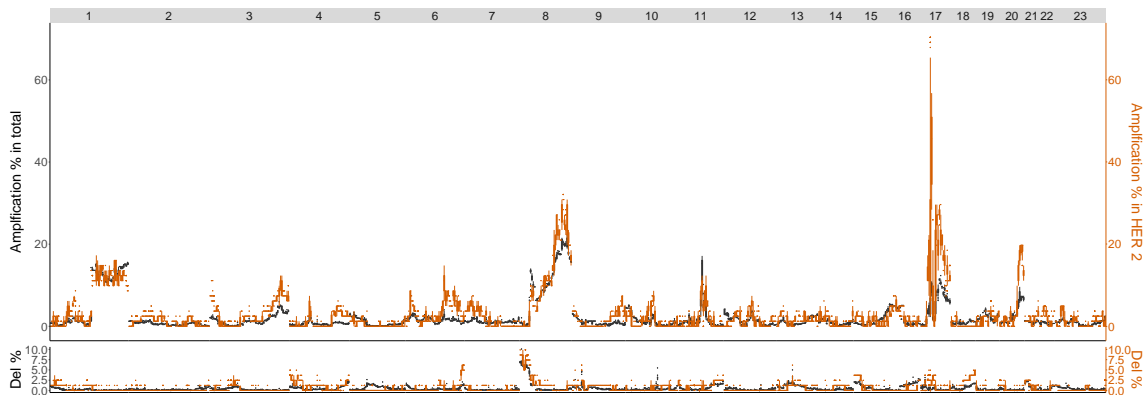


Figure A.1: HER2 vs. other phenotypes combined on TCGA BRCA cohort Data. Cytobands with gains and losses in the TCGA BRCA cohort dataset [21] for the HER2 phenotype (red) as well as all phenotypes combined (black). Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort data set indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of HER2 patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

Table A.2: Cytobands detected by 0-dimensional Betti, lifespan and persistence landscape curves for the Luminal B subtype on the Horlings dataset [66] using the *R* TDA package. Basals were not included in the control set.

Luminal B (No Basal in Control)	
Betti-0	1p36.32-p34.2, 1q32.1-q41, 8p22-p11.1, 8q24.11-q24.3, 9p24.3-p21.3, 9q21.33-q22.32, 9q31.1-q33.1, 12q21.31-q23.2, 21q11.2-q22.3,
Lifespan-0	1p36.11-p34.2, 1p31.1-p22.2, 1q31.1-q41, 1q32.1-q41, 3p22.3-p13, 3q24-q26.2, 4p16.3-p15.2, 4q31.3-q34.1, 6q24.1-q25.3, 8p23.3-p22, 8p22-p11.1, 9q21.33-q22.32, 10q23.1-q24.2, 12p13.33-p12.3, 12q21.1-q24.33, 15q23-q26.3, 21q11.2-q22.3, 23q11.1-q21.33
Landscape	λ_2 : 1p32.1-p31.1, 2q31.1-q32.2, 3p22.1-p13, 4q31.21-q34.1, 5q23.1-q31.2, 6q22.31-q24.1, 12q21.31-q24.11, 21q11.2-q22.3, 23q11.1-q21.33
	λ_3 : 1p32.1-p31.1, 1q31.1-q41, 2p25.3-p23.2, 2q31.1-q33.1, 3p22.3-p13, 4p16.3-p15.2, 4q22.1-q25, 4q25-q28.3, 4q31.21-q34.1, 5q23.1-q31.2, 6q22.31-q24.1, 8p22-p11.1, 10q21.2-q23.1, 12q21.31-q24.11, 12q24.11-q24.33, 15q21.3-q25.2, 16q11.2-q21, 21q11.2-q22.3, 23p21.3-p11.21, 23q11.1-q12.3, 23q24-q27.2
	λ_4 : 1p36.22-p34.2, 1q31.1-q41, 3p22.1-p14.3, 4p16.3-p15.2, 6q24.1-q25.3, 8p23.1-p12, 8p22-p11.1, 8q22.1-q23.3, 9q21.33-q22.32, 10q21.2-q23.1, 12p13.33-p12.3, 12q21.1-q24.23, 14q11.2-q21.1, 16q12.2-q22.1

Table A.3: Average sensitivity and specificity of Betti curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.

Betti Curves					
$\mu = 1$	20% mix	40% mix	60% mix	80% mix	100% mix
$\sigma = 0.2$	TPR: 43.00%	50.00%	63.00%	81.00%	99.00%
	SPC: 71.00%	83.00%	93.00%	97.00%	99.00%
$\sigma = 0.5$	TPR: 52.00%	58.00%	66.00%	73.00%	82.00%
	SPC: 60.00%	67.00%	73.00%	77.00%	83.00%
Total	TPR: 47.50%	54.00%	64.50%	77.00%	90.50%
	SPC: 65.50%	75.00%	83.00%	87.00%	91.00%

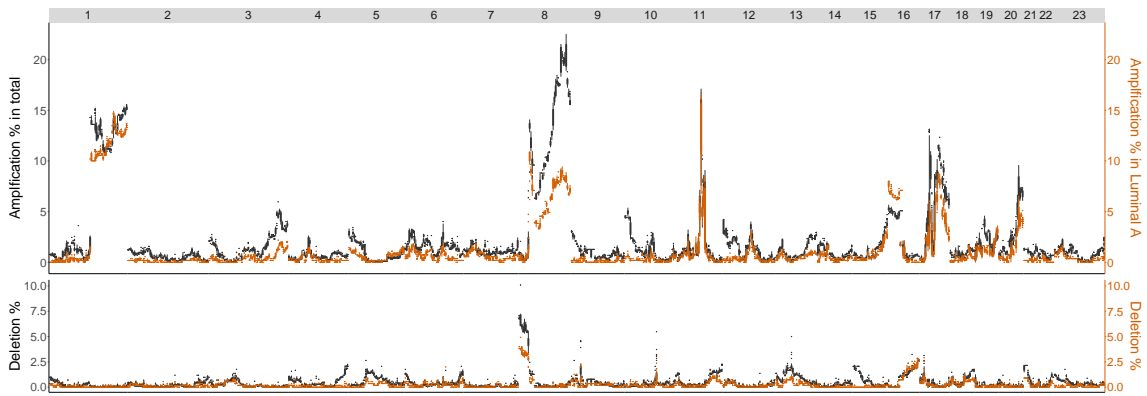


Figure A.2: Luminal A Phenotypes: The plots illustrate in which cytobands there are gains and losses in the TCGA BRCA cohort dataset [21] for the Luminal A phenotype as well as all phenotypes combined. Luminal A is in red and all groups combined are in black. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal A patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

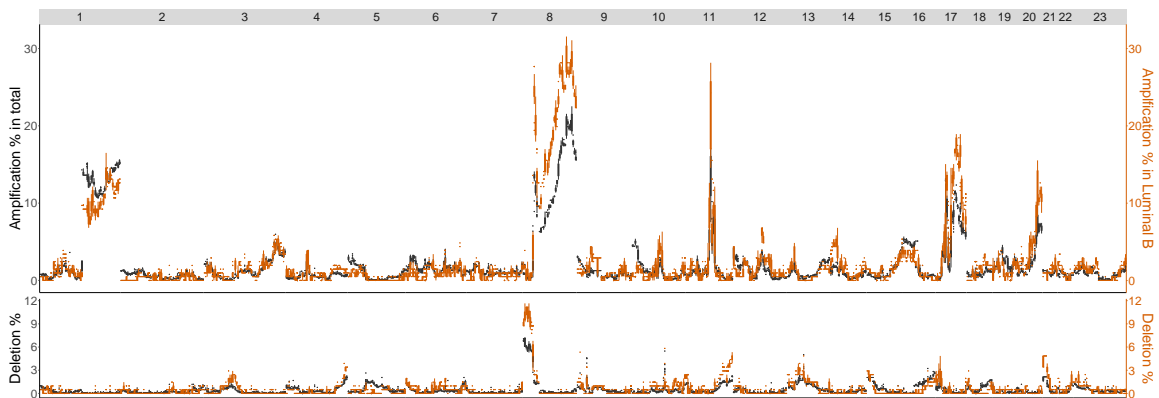


Figure A.3: Luminal B Phenotype: The plots illustrate in which cytobands there are gains and losses in the TCGA BRCA cohort dataset [21] for the Luminal B phenotype as well as all phenotypes combined. Luminal B is in green and all groups combined are in black. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

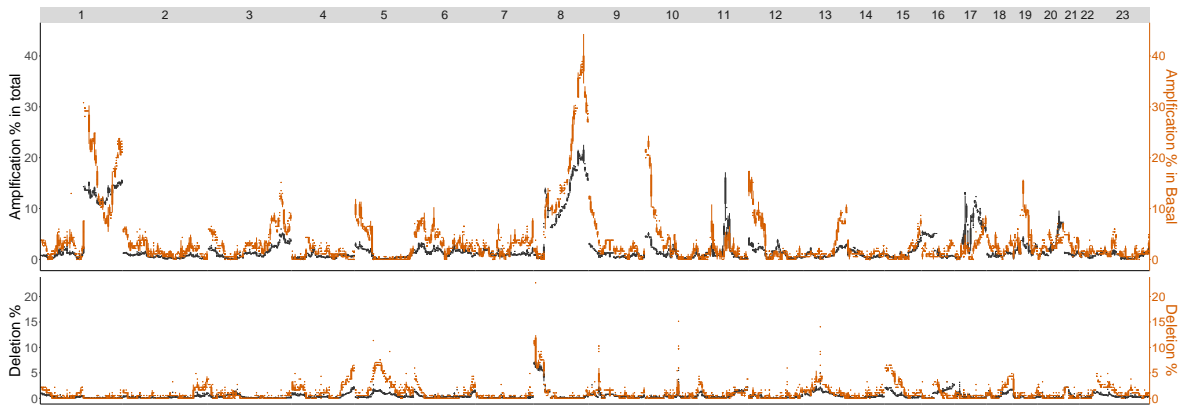


Figure A.4: Basal Phenotype: The plots illustrate in which cytobands there are gains and losses in the TCGA BRCA cohort dataset [21] for the Basal phenotype as well as all phenotypes combined. Basal is in orange and all groups combined are in black. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Basal patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

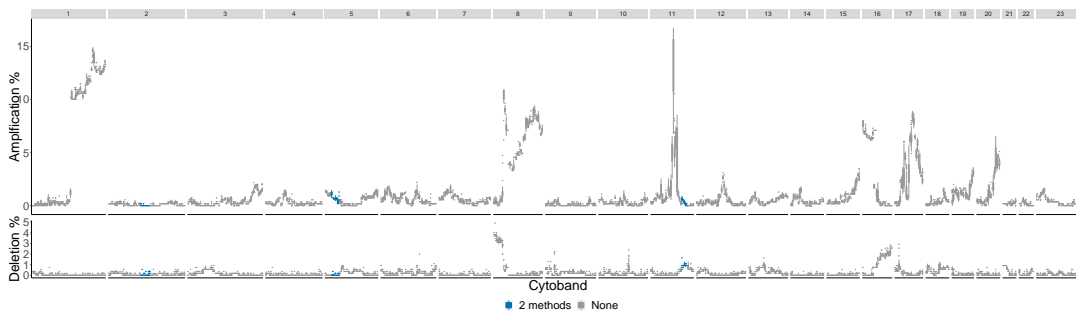


Figure A.5: Luminal A phenotype cytobands in TCGA BRCA cohort dataset. The colors indicate how many persistence curve methods detected that particular cytoband in the Horlings dataset [66]. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal A patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

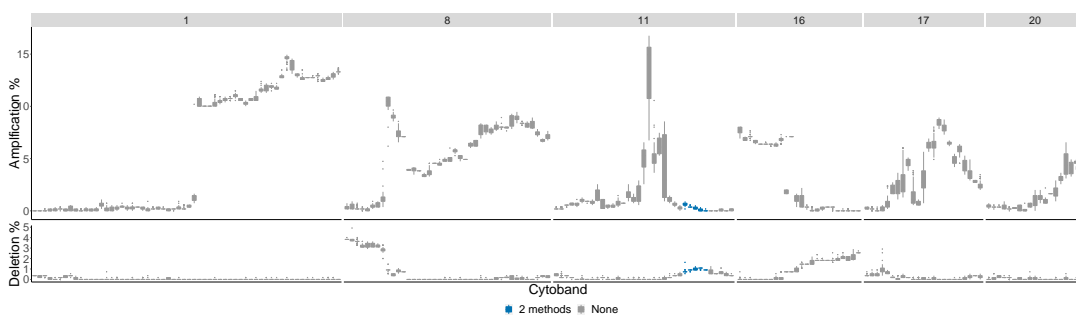


Figure A.6: Luminal A phenotype most aberrant cytobands in TCGA BRCA cohort dataset. The chromosome arms 1, 8, 11, 16, 17 and 20 are included above 10% of patients had aberrations in the genes in these cytobands. The colors indicate which persistence curves detected those cytobands as significant in [66]. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal A patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

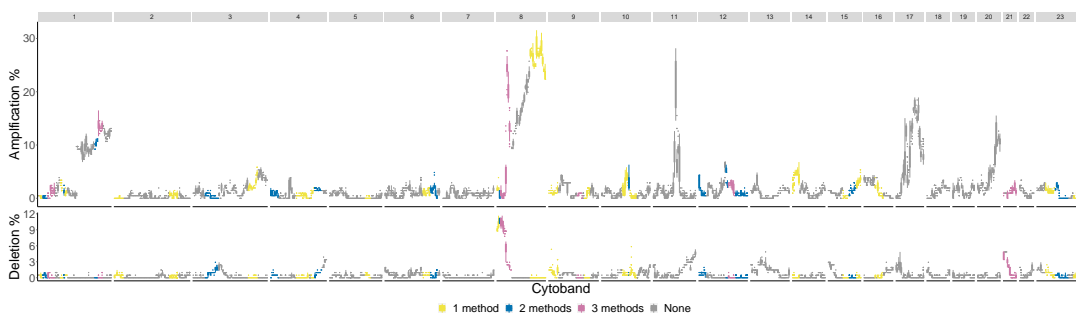


Figure A.7: Luminal B phenotype cytobands in TCGA BRCA cohort dataset. The colors indicate how many persistence curve methods detected that particular cytoband in the Horlings dataset [66] with no Basals in the control group. Each gene within a cytoband has a score of $-2, -1, 0, 1, 2$ in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

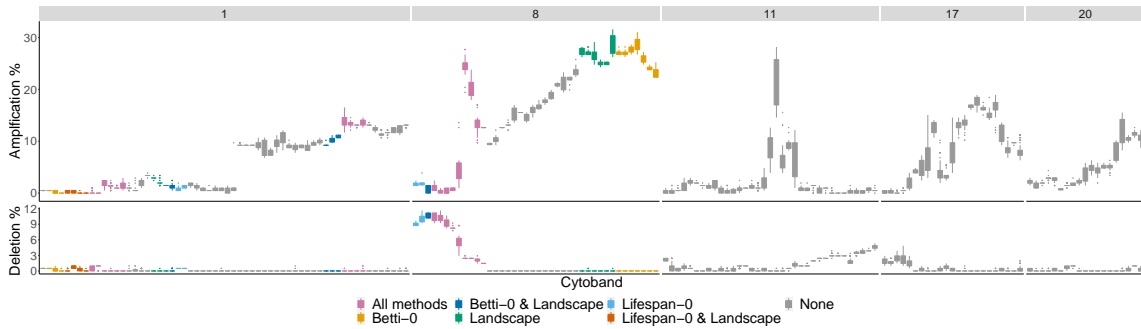


Figure A.8: Luminal B phenotype most significant cytobands in TCGA BRCA cohort dataset. The chromosome arms 1, 8, 11, 17 and 20 are included since above 10% of patients had aberrations in the genes in these cytobands. The colors indicate which persistence curves detected those cytobands as significant in [66] for Luminal B with no Basals in the control group. Each gene within a cytoband has a score of -2 , -1 , 0 , 1 , 2 in the TCGA BRCA cohort dataset [21] indicating major deletions, mild deletions, normal, mild copy number gain and major copy number gain. In the top plot for each cytoband we plot a boxplot of the percentages of Luminal B patients with a score of 2 for each gene in the cytoband. The bottom plot is the same except we calculate the percentage of genes with score -2 .

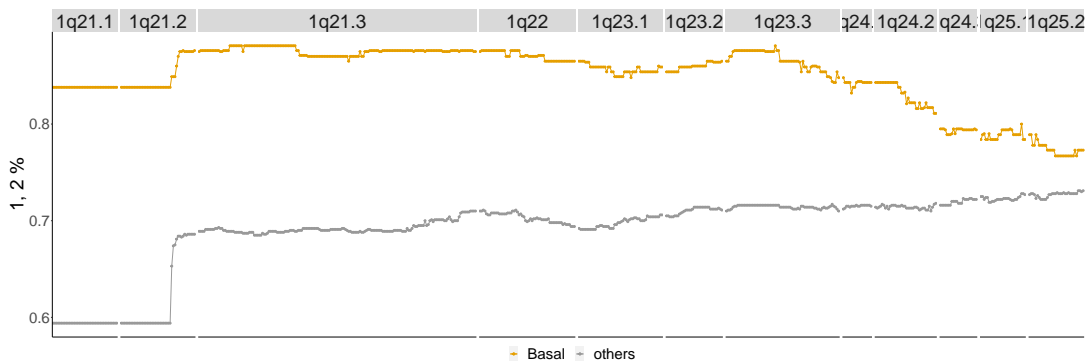


Figure A.9: Basal phenotype cytobands $1q21.1 - 1q25.2$. Shows the percentage of patients in the Basal phenotype with either a 1 or 2 score from the TCGA BRCA cohort dataset (orange) as well as the percentage of all other phenotypes with these scores (gray).

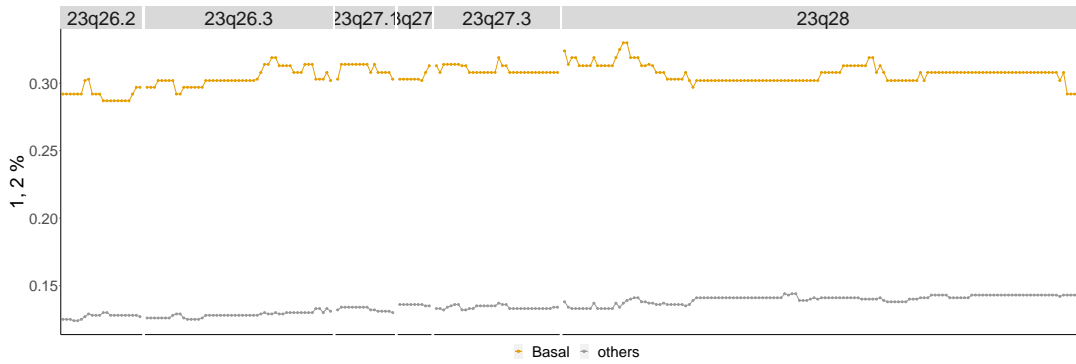


Figure A.10: Basal phenotype cytobands 23q26.2–23q28. Shows the percentage of patients in the Basal phenotype with either a 1 or 2 score from the TCGA BRCA cohort dataset (orange) as well as the percentage of all other phenotypes with these scores (gray).

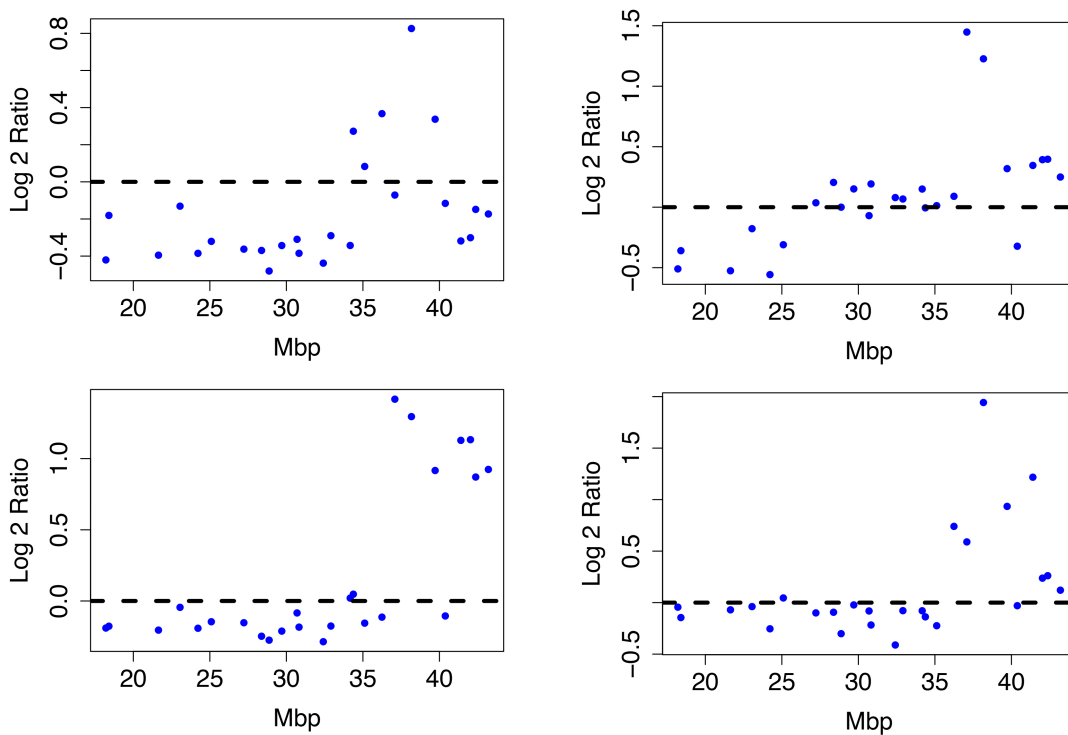


Figure A.11: Luminal B patient profiles. Four Luminal B patient profiles from the Horlings dataset on cytobands 8p22-8p11.1. All share a significant copy number gain between 35 and 45 Mbp and a loss from 15–35 Mbp.

Table A.4: Average sensitivity and specificity of lifespan curves for patient classification. The length of aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.

Landscape 3					
$\mu = 1$	20% mix	40% mix	60% mix	80% mix	100% mix
$\sigma = 0.2$	TPR: 26.00% SPC: 93.00%	TPR: 41.00% SPC: 97.00%	TPR: 59.00% SPC: 99.00%	TPR: 76.00% SPC: 99.00%	TPR: 90.00% SPC: 100.00%
$\sigma = 0.5$	TPR: 45.00% SPC: 60.00%	TPR: 50.00% SPC: 69.00%	TPR: 55.00% SPC: 71.00%	TPR: 61.00% SPC: 74.00%	TPR: 66.00% SPC: 77.00%
Total	TPR: 35.50% SPC: 76.50%	TPR: 45.50% SPC: 83.00%	TPR: 57.00% SPC: 85.00%	TPR: 68.50% SPC: 86.50%	TPR: 78.00% SPC: 88.50%

Table A.5: Average sensitivity and specificity of lifespan curves for patient classification. The length of the aberration in aberrant profiles, λ , is fixed at 10 of 20 total probes for this table.

Landscape 4					
$\mu = 1$	20% mix	40% mix	60% mix	80% mix	100% mix
$\sigma = 0.2$	TPR: 24.00% SPC: 94.00%	TPR: 40.00% SPC: 99.00%	TPR: 58.00% SPC: 99.00%	TPR: 73.00% SPC: 100.00%	TPR: 86.00% SPC: 100.00%
$\sigma = 0.5$	TPR: 45.00% SPC: 64.00%	TPR: 51.00% SPC: 71.00%	TPR: 58.00% SPC: 75.00%	TPR: 66.00% SPC: 78.00%	TPR: 72.00% SPC: 81.00%
Total	TPR: 34.50% SPC: 79.00%	TPR: 45.50% SPC: 85.00%	TPR: 58.00% SPC: 87.00%	TPR: 69.50% SPC: 89.00%	TPR: 79.00% SPC: 90.50%

Table A.6: Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.

Basal 1
IGKC, TMSB10, S100A11, S100A9, S100A6, GAPDH, S100A8, RPS18, TAGLN2, S100A10, TUBB, CALML5, S100A7, NDUFS5, YBX1, TPI1, ILF2, HMGA1, HDGF, S100A4, S100A2, PSMB4, PFDN2, A2M, MARCKSL1, RPS8, UQCRH, RPL5, ACTG2, RPS27, FLNA, TACSTD2, VIM, KLHDC3, GDI2, CDC20, CCT3, ANP32E, CHI3L2, IRAK1, VGLL1, ABRACL, C1S, FOXC1, MDFI, MTCH1, MCM3, PRDX1, PDZK1IP1, SNRPC, GBP1, SF3B5, PHGDH, MLF2, TAP1, STRAP, LMNA, PPP1CB, LSM2, MRPL51, DEK, MRPL14, DSG3, MRPS21, NES, TFDP1, CAPG, RSU1, SERBP1, CYP1B1, CLIC1, S100A16, CCT4, TAPBP, NCAPD2, SLC2A1, TNFRSF21, C6orf15, DSP, PHB2, VPS72, DSC2, HMGN4, SOX4, YBX3, IFI44, DSG2, PTPRF, USP1, CDC123, PPP1R14C, MRPL37, HSD17B10, EFNA1, CCT5, SF3B4, MPZL1, CSDE1, KCNK5, STK38, PLP2, C1R, SMG5, MAGOH, ID4, APH1A, FAM171A1, BYSL, TCF7L1, PSMB2, CDCA8, OPTN, KIFC1, ELOVL1, PTK7, CUTA, SFT2D2, PIM1, CCDC167, FOXP4, CNN3, PSMD4, NDUFB11, CCT7, ATL2, DKC1, SEPHS1, MEX3A, SLC39A7, DSC3, CMAS, NDUFS6, IGKJ1, SLC16A1, MAGED1, CTSS, NFIB, CMPK1, PGM1, PAK1IP1, DNPH1, SNRPG, GTPBP4, SPRR1B, DPM3, IFI16, GSTA1, KIF2C, SF3A3, F11R, PFKF, VANGL2, CREG1, LAMP1, CDK16, HPRT1, PXDC1, EFNA4, IRX1, NASP, LTBR, TMEM14C, COL4A2, STK38L, SSR2, FAM136A, UGP2, A2ML1, ACTR2, COPA, VCAM1, MDH1, SLC35B2, FCER1G, ATN1, TCF19, GPX7, RBM17, CITED4, MPZ, ABCF1, SLC29A1, ADAM15, PPP1R18, HMGB3, PSMB9, ERI3, RIOK1, CALM2, MRPL9, BAG6, RPIA, MOGS, XAGE2, PCSK1N, QPCT, PELI1, TUBB2B, PRRC2A, TRIP13, MTHFD1L, PPIL1, NELFE, HORMAD1, CDK1, BCAS2, RRP36, E2F3, FLAD1, ATP1A1, COL11A2, PSMB8, LRRC8D, MRPS15, UAP1, SRE, EMP1, B4GALT2, MYO10, CCNY, IPO5, GOLT1B, PIM2, HTATSF1, CNPY3, NFYA, CD83, NUDT5, RAP2A, UBQLN4, XPO5, PRTFDC1, GJB3, WDR43, TIPRL, LTB, RNF19B, IQGAP3, NET1, LRRC42, RPL7L1, DHTKD1, PSIP1, FAM189B, WDR46, TMEM54, ARHGEF2, TAF11, SSR4, MRPS10, TEX261, NUF2, WRNIP1, GMDS, TBCC, GNL2, UBE2Q1, FHL3, KANK4, DAP3, HENMT1, SCNM1, GBP5

Table A.7: Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.

Basal 2
AKIRIN1, ABI1, RAB32, USP5, TAP2, AUP1, DSG1, HAX1, EBNA1BP2, SRSF3, CLEC7A, SLC6A8, RTKN, PABPC4, SRPK1, BAK1, ARHGAP21, DPH2, HSPA14, TNFSF13B, GBP4, TINAGL1, SLC10A3, PRCC, DUSP23, DGUOK, LTBP1, SNRPA1, MAGEA4, CRTC2, CD53, DAP, PPAR, TMEM79, AK2, EML4, SPRR2A, CLPTM1L, NDC1, EDN1, USP39, TMEM14A, PRPF3, UBAP2L, DERA, GTPBP2, NUP153, SCUBE3, LDHB, NRN1, LRP6, ZWINT, CDKAL1, GMNN, IL12RB2, PQBP1, ARNTL2, MRPL36, MAGEA3, GJB5, UCK2, LTV1, KCMF1, NRM, GSTA4, CDCA3, LRPPRC, BOLA1, PHC2, TTC7A, TM7SF3, CD2, ANKH, ITPR3, GCNT2, UBL4A, TRAPPC3, SIX3, MCM10, PAGE2, MED8, PLBD1, IFI44L, EMD, NSUN2, CTSP1, TAF13, PI4KB, COL9A2, NOL7, TGFA, YRDC, PSRC1, FAM98A, TAPBPL, CD163, FAF1, SPRR1A, PDE4B, MEA1, MAPRE2, IGKJ5, UPF2, POGK, PLEKHO1, VBP1, ING4, ETV7, B4GALT3, CDC7, LGALS1, CDC5L, PLA2G7, MAGEA6, SRD5A1, EBP, CLSTN3, SSR1, TMEM39B, ECHDC3, SSBP3, IGKJ4, RNF138, FDPS, MPHOSPH10, COL4A1, DEDD, EDN2, PLEKHG6, PRKD3, SUV39H2, ARL5B, MAP7D1, ERP27, HPDL, CDC42SE1, PNO1, TPGS2, SPRR2E, ABT1, ORC1, MTHFD2, CCHCR1, PHYH, ELK1, STK40, DKK1, PARD3, PLEK, WDR77, RAB1A, HUWE1, MGST1, PSORS1C2, ISG20L2, PCP4L1, RFX5, GBP1P1, THBS3, LSM10, ABCD1, TEAD3, SLC6A9, SNX7, SPRR2D, DPY30, PDSS1, PRPF38A, TBC1D22B, MED20, MGST3, MCUR1, MOCOS, PPP3R1, MLLT11, CD4, PPP2R5D, HECA, PLEKHG1, PFDN6, SLAMF7, L1CAM, FTSJ1, KATNA1, GPR161, HDAC1, SFPQ, DEPDC1, STIL, SLC15A1, RBM8A, MGMT1, IDH3G, MASTL, UXT, PEX5, LAG3, HSPA6, CSF1, TTC22, DNM1L, C18orf21, FKBPL, MAGOHB, IGKJ3, KANK1, ATP11A, TPM3, TNF, MRPL33, IL15RA, FAM50A, RNF182, WDR3, IPO13, TBX19, LRRC8B, SLC12A7, RAD54L, CDC42EP3, RAET1L, SERPINB9, MAMLD1, CTTNBP2NL, ZNF644, TEX30, RECQL, WAC, DHX16, TRIM27, C1RL, BMP6, CLPSL1, BCL2L14, HCG11, SUV39H1, TUBB2A, HHLA3, BTN2A1, IL2RA, CISD1, GPATCH4, DNTTIP2, PIP5K1A, TIA1, ULBP3, ZFP57, HCP5, LCK, PAQR8, GNAI3, DUSP22, ULBP2, BTN3A1, PCID2, SUCLG1, C6orf136, IVL

Table A.8: Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.

Basal 3
<p>DDX11, C1orf122, PRIM2, C2CD4D, CCDC3, EHBP1, PRSS16, DYSE, LIX1L, CAMK1D, RTCA, GALNT14, CYP26B1, ALDH1A3, BCAT1, TRIT1, ARHGAP30, TMEM14B, LYRM4, GNLY, REPS1, S100A1, MTPAP, CARS2, ARHGEF11, ANKRD35, LRRC40, FAM3A, EIF2AK2, LRRC8C, SMYD5, CUL7, TSPAN2, DZIP1, NUDT11, CD207, HNRNPLL, MAP7D3, GPR180, TXLNA, CLDN10, S100A3, TPRKB, GABRE, NAA10, RNF220, FAM83B, MOB1A, TBC1D7, MAEL, SCMH1, POLH, AEBP2, CEBPZ, AMPD2, UHRF2, GNL1, CD1A, STAM, MED21, GABARAPL1, LAMTOR5, DENND4B, OXGR1, UBE2D1, ADAMTS4, ZRANB2, IRX4, VLDLR, WAS, MRS2, RAB23, DOCK7, GCLM, PKDCC, AARS2, HCFC1, LEMD2, TJAP1, CD8A, APEX2, ADD2, PPIH, CLSPN, CHTOP, REG1A, MAGEA12, PKP2, PCGF1, GPSM2, RLF, SLC2A3, TTC27, DUSP11, DEF6, XAGE3, RIMS3, GATAD2B, USP24, HTRA2, GFOD1, LRRC14B, FAM50B, ABCC10, DCLRE1C, TDRKH, CLEC4E, EIF3I, CELF2, ATAT1, ST6GALNAC5, MLLT10, ULBP1, AK4, CSNK2B, IL6R, HSPB11, CC2D1B, ADAMTSL4, ROR1, ST8SIA1, C1orf112, LINC00680, CCDC28B, OSBPL9, GPR19, ERGIC2, PNMA3, SLC9A6, PARS2, VMA21, TTF2, LMO3, MMACHC, IGSF9, C2, BCAN, PTGFR, ZNF318, JTB, BTN3A3, CTAG2, POLR1A, FCGR2A, ATP6V1E2, TOMM40L, CLK2, ITGA10, PRPF18, AGO1, MNDA, FRMD4A, VAX2, FMR1, CDYL, MAPK14, FAM167B, PRKCQ, TRIM38, PRR3, PGLYRP4, C1orf109, SLAMF6, STYK1, LINC00518, SLC30A6, LRP8, SEMA6C, DTNBP1, AGPAT1, CRIPT, PPIE, BTN2A2, CLCN5, MAGEC2, TGDS, MTF2, GNL3L, KIN, SOCS5, PBX2, SLAMF8, CENPQ, SIRT5, ZNF184, RBM15, MSTO1, SLC6A17, ADTRP, PTC3, MRPL19, NHLRC1, VIT, FAM78B, JARID2, ZNF391, ANKRD16, FOXP3, SLC38A5, BICD1, LINC00707, BEND7, GBAP1, MAGEA10, ALG6, LRRK1, TNFAIP8L2, NQO2, PLXNB3, STON1, KTI12, PNMA6A, FOXI3, POU5F1, CEP72, PSORS1C3, DNAJC6, FAM90A1, PMF1, AGO4, FGFR1OP2, RGN, USP49, CCDC18, NDUFAF7, DHX57, STRN, ZNF384, CSAG3, GSPT2, SEMA4A, ARHGEF7, GLDC, LYSD1, ANXA4, SPAST, C2orf68, FOXJ2, SASS6, RASGRP3, CD247, CD27, PAQR6, RTKN2, SLC25A27, ZNF165, ZNF438, PGLYRP3, ATXN7L2, BRD9, DNAJB4, ETV3L, SFMBT2, TFAP2E, KIAA1586, KDM4A, CD163L1, TACR1, PPP1R21, ZFP69B, F8A1, MOSPD1, GJA5, EXO5, ZBTB12, PRKAB2, GLP1R, ZYG11A, UST, CLEC4F, BOLA3, DMRT3</p>

Table A.9: Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.

Basal 4
<p>TUBGCP3, FBXO5, NMT2, NFKBIL1, MED10, AKR1C3, PLAGL1, DUSP9, USP6NL, CLIP4, THNSL1, AKR1E2, SETDB1, ADAT2, EPHX4, OLAH, CREM, MOG, PTPN22, FOXF2, BIVM, STAMBP, GCFC2, AKR1C1, S100A12, LCLAT1, SMIM12, PLB1, DSTNP2, LST1, CAMKMT, CD8B, THUMPD2, C6orf52, NOLA, MEIG1, GPR158, SPSB2, EEF1E1, AP4B1, LINC00622, ANKRD33B, TOE1, TREM1, CD48, FCGR1A, RBMXL1, ARTN, ABCA4, RNY3P8, CACHD1, GSTA2, PTBP2, XCL2, NGF, MRPL2, CPNE5, MBNL3, ARHGAP25, ANKRD36BP2, OXER1, PCSK9, ATG4C, LTA, FGD4, PIK3C2G, SMIM10, ADCY2, WNK3, FAM72D, YIPF4, PYROXD1, PDE3A, TREX2, LINC00987, LINC00460, GEMIN6, BMP8B, SPRR2G, TMEM17, TAF8, HIVEP2, PSMG4, CD1B, SLAMF1, GABRA3, APLF, SEC61A2, ZBTB8OS, FOXD3, TRDMT1, KDM4C, FCRL5, GPR173, DMRT1, MYLK4, WDR37, CLPS, FRS3, LOH12CR2, GBP6, VWA7, LCE3D, ITGB3BP, SUPT3H, MYO3A, TCP11, C1orf94, INO80C, ZIC5, DMRTA2, KCNC4, SPRR2B, S100A5, TMEM61, B4GALT6, XCL1, DENND2C, FAM72C, KRBOX4, AGER, TRIM46, PYHIN1, MAGEA11, STX11, HCG15, ADCY10P1, GPR18, SLCO1A2, ZNF451, LINC00404, BEND6, TXNDC12, FGD2, MIR4645, SLC39A12, GPR89B, F5, CCNB3, FCGR2B, SLC2A14, CSMD2, MIR17HG, MIR4258, RN7SL130P, CD40LG, TRIM39, NALCN, USP27X, DLGAP3, ANKRD53, LRRC37A6P, MAGEA9B, TRIM15, KCNJ10, CD244, RAET1G, TMEM234, INPP5B, GJB4, TMEM217, ALX3, FAM161A, AFF2, TRABD2A, RPS10, PRRG3, TMLHE, PHACTR1, OFCC1, GPRC5D, SYCP2L, FCRL3, LCE3A, INSL6, RAET1K, KCNA3, CD101, HCG18, UBL4B, IPCEF1, ACOT11, PHYHIPL, LHFPL5, ZC3H12D, NR1I3, TDRD6, WDR92, CLEC6A, CALML3, ITLN2, RNA5SP39, UBE2U, KLRD1, LHX8, ACSM4, DMRT2, SPRR2F, CLEC12A, AKR1C7P, SMCO2, HAUS7, GNB3, CUBN, AGO3, PRB2, KCND1, CAPN14, RPL13P5, ARHGGEF33, BTN2A3P, CARM1P1, MIR765, SH2D1B, LINC00565, UCMA, PPP1R3G, PTCHD3, ZBTB8B, SPRR4, LCE3E, PRB1, KLRC1, TTC24, FOXE3, NHLH1, CLYBL, SLC39A1, BARHL2, TREML2, TULP1, PPP1R2P1, SLCO1B3, PRB4, RPE65, MEMO1, SLFNL1, LINC00240, MIR4641, PAGE4, LINC00705, GSTA5, MLIP, PRR26, SOX21, PNPLA1, GPLD1, BEST4, DMBX1, FIGLA, TEX29, ATP4B, ZIC3, TEX37, ACTBP12 EFNA3, DUSP12, USF1, EMG1, TNFRSF1A, SH2D2A, PITRM1, NFYC, RAVER2, RENBP, RNF8, SRSF7, BTN3A2, SAYSD1, SMIM13, CSAG1, LDOC1, PLEKHG4B, BTF3L4, CUL2, CYP39A1, TRIM26, POLR1C, SIX2, YARS2, FGD1, PSMA5, LRRC41, AIM2, SIKE1, LBH, MRPS18A, KCNQ4, ZNF362, AIF1, RHOQ, EIF2B3, FOXQ1, STK24, ANKS1A, RCSD1, RXRB, BAG2, SLC35A2, PRADC1, ATP11C, GPR183, EVA1A, OTX1, UBAC2, RHBDL2, NSDHL, KCTD20, DCLRE1B, ACBD7, TET3, FOXN2, RCL1, CD58, EHMT2</p>

Table A.10: Candidate oncogenes that were detected as significant for the Basal subtype in the TCGA BRCA cohort that come from genomic regions that were detected as significant in the Horlings data set that are not among the UCSF 500 genes.

Basal 5
RPP38, YME1L1, SVIL, SERTAD2, POLR3C, THEM5, VPS54, IMMT, MDC1, RPP40, SORT1, KDM1B, GMPR, TMEM151B, ZNF684, SFTA1P, FBXO48, NCR2, INSL4, OXCT2, LY6G5B, RNA5SP221, MYBPHL, HFM1, RNA5SP515, RPP21, TAS2R14, ERAS, UBD, SAPCD1, ATP2B3, ARMC4P1, CD5L, ZNF157, A3GALT2, KCNA2, CSAG2, FAM151A, NOP2, NANOGNB, MIR4675, TRIM10, SLITRK2, HPCAL4, RXFP4, MCF2, FNDC7, RNA5SP305, CAPZA3, MSH5, RN7SKP272, WNT2B, LYZL1, MAGEA9, KCNJ9, PRL, POM121L2, C2orf91, TXNDC5, CDKL4, NUDT3, SNORD53, PXT1, GRK1, RPTN, PNMA6B, RN7SKP240, LRP11, LINC00626, SPDYA, LRRTM4, FAM122C, C2orf74, IR5187, PASD1, MIR4436A, PRR9, LINC00629, HULC, MIR938, OR2H2, RPS4XP5, PDE6H, RN7SL372P, GRIN2B, RN7SKP293, NUP210L, GUCA2A, METTL21C, LINC00379, RNU6ATAC39P, MIR1915, KLRK1, CHD1L, SNORD92, LINC00354, TTC4, SNORD38A, FCRL4, DMRTB1, TFAP2D, MLN, LINC00567, MAGED4B, LINC00583, SOX1, SLC17A1, KCNV2, PRH1, LINC00937, MIR137HG, GJD4, GAGE10, RNA5SP207, LRRC7, C12orf77, RN7SL748P, SUMO4, BTNL2, ST3GAL3, HCG24, GUCY2C, RN7SL273P, CER1, CTAG1B, PRH2, IDI2, RNA5SP65, TREML4, GPX6, RN7SKP241, LINC00336, AMY1B, HDGFL1, MAGED4, IL17A, MAGEA2, REG1B, LINC00551, LCE3C, LINC00703, PPIAL4A, AKR1C4, RN7SL123P, CT45A3, IL23R, LINC00612, CCDC168, IL17E, RN7SL538P, RNA5SP179, PPIAL4G, CTAG1A, ANGPTL3, GTF2H4, FAM156A, RNA5SP89, LINC00452, MIR557, TINAG, MUC22, F8A3, SPANXN3, MIR4255, RN7SKP39, MAGEA2B, TPD52L3, MIR5004, DPPA3, C1orf185, DBIL5P2, FUNDC2P2, TAS2R13, NANOG, OR2J3, RN7SL440P, PTF1A, LCA10, MIR877, RN7SKP73, VSIG8, LINC00443, RN7SL580P, RN7SL98P, RGPD1, KCNA10, LINC00477, DDX11L9, SLC10A2, RNA5SP49, TMEM247, DEFB133, RN7SL380P, BMP10, RN7SL51P, DEFB110, RN7SKP186, FAM138E, RN7SKP224, OR2B3, MIR548AS, SPATS1, HMGB4, GTF2A1L, RN7SL516P, OPN5, OR2H1, OR13H1, GPR101, MIR934, LINC00309, AMY1C, CDCP2, MIR513A1, RNA5SP178