Utility of machine learning approaches for cancer diagnosis and analysis from RNA sequencing

by

Jasleen K Grewal

B.Sc., Simon Fraser University, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

August 2020

© Jasleen K Grewal 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Utility of machine learning approaches for cancer diagnosis and analysis from RNA sequencing

submitted by **Jasleen K Grewal** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Bioinformatics**.

Examining Committee:

Dr. Steven JM Jones, Bioinformatics
Supervisor

Dr. Ryan Morin, Bioinformatics
Supervisory Committee Member

Dr. Marianne Sadar, Pathology and Laboratory Medicine
University Examiner

Dr. Wan Lam, Pathology and Laboratory Medicine
University Examiner

Additional Supervisory Committee Members:

Dr. Inanc Birol, Bioinformatics
Supervisory Committee Member

Dr. Ryan Morin, Bioinformatics
Supervisory Committee Member

Dr. Sohrab Shah, Bioinformatics
Supervisory Committee Member

Dr. Stephen Yip, Pathology and Laboratory Medicine
Supervisory Committee Member

Abstract

The highest number of cancer-associated deaths are attributable to metastasis. These include rare cancer types that lack established treatment guidelines, or cancers that become resistant to established lines of therapy. Precision oncology projects aim to develop treatment options for these patients by obtaining a detailed molecular view of the cancer. Scientists use sequencing data like whole-genome sequencing and RNA-sequencing to understand the biology of the cancer. A significant challenge in this process is diagnosing the cancer type of the sample since the observed measurements are best understood with this context.

Routine histopathology relies on tissue morphology and can fail to provide a determinative diagnosis when the cancer metastasizes, presents biology attributable to multiple different cancer types, or presents as a rare cancer type. Molecular data has revealed differences in the genetic makeup of cancers that appear morphologically similar, motivating the use of molecular diagnostics. Nevertheless, no existing tools utilize the output from these sequencing modalities in its entirety (that is, without feature selection). There is also limited work evaluating the utility of pan-cancer molecular diagnostics in a precision oncology trial.

In this work we review an ongoing precision oncology trial and identify the impact of sequencing-based approaches on cancer diagnosis. We develop SCOPE, a machine-learning method that uses RNA-Seq profiles of tumours for automated cancer diagnosis. We show that this method, which uses over 17,688 gene measurements as input, has better classification accuracy than when using statistically prioritized marker genes, can deconvolve cancer-types with mixed histology, and has high performance in metastatic cancers and cancers of unknown origin. In precision oncology, manual analysis of the tumour's genomic profile is used to understand tumour biology and driver pathways. We find that by assessing the classifier's dependence on gene subsets, we can automatically calculate the importance of various biological programs in individual tumours. Pathways prioritized

through this tool - called PIE - show a high overlap with manual integrative analysis performed by expert bioinformaticians to identify clinically important genomic changes. Lastly, we demonstrate that PIE facilitates cohort-wide cancer analysis and discovery of novel sub-groups in advanced cancers.

Lay summary

Diagnosis is an important early step in the management of cancer, and is usually provided by expert doctors based on the cells' appearance. This process can be easy or difficult based on various factors. Cancers are diseases of the genome. In this thesis we show that in many advanced cancers we can use molecular measurements like DNA and RNA sequencing to provide a diagnosis. We develop a computational cancer diagnosis tool that uses expression measurements of all the genes in a cancer. We then show that this method can be used to learn which biological changes are important for an individual cancer. We compare these automatically identified changes with what expert computational biologists found manually, and find a significant overlap. This approach automates the way we use sequencing data for diagnosing and understanding cancers, and expands our ability to understand rare and understudied cancers.

Preface

All the work presented herein was conducted at Canada's Michael Smith Genome Sciences Centre, part of BC Cancer, in the laboratory of Dr. Steven J.M. Jones. Data from Personalized OncoGenomics (POG) clinical trial was obtained after institutional review board (IRB) approval. This work was approved by and conducted under the University of British Columbia – British Columbia Cancer Agency Research Ethics Board (H12-00137, H14-00681), and approved by the institutional review board (IRB). The POG program is registered under clinical trial number NCT02155621. As part of this trial, cancer patients with advanced disease who failed conventional treatment and fulfilled the inclusion criteria were consented for tumour profiling using RNA-Seq (tumour) as well as whole-genome sequencing (tumour and blood).

Patients were referred to the POG program through their treating oncologist and enrolled into the program through a POG trained oncologist or study nurse. Sample collection was performed by the overseeing surgical oncologist. Dr. Andrew Mungall was responsible for the processing and library construction of the samples. Dr. Richard Moore oversaw the sequencing of the samples. Eric Chuah, Karen Mungall, Tina Wong and Reanne Bowlby supervised the alignment and variant calling of the samples.

A version of Chapter 2 has been published in Cold Spring Harbor's Molecular Case Studies and the citation is below. A license to reuse the text and figures was not necessary since the authors retain the copyright for the publication, licensed under CC-BY. Drs. SJM Jones, Marra, Laskin and Karnezis contributed to the conception and design of the study. Dr. Anna V Tinker referred the patient to the study. Initial pathology at the Vancouver General Hospital was led by Dr. Chen Zhou. Validation pathology was led by Dr. Anthony N Karnezis and assisted by Drs. Kenrry Chiu and Basile Tessier-Cloutier. Dr. Andy Mungall contributed to the collection and assembly of data. Drs. Martin Jones and Peter Eirew contributed to data analysis and interpretation. I led the manuscript writing efforts in

collaboration with Drs. Kenrry Chiu, Basile Tessier-Cloutier, Martin Jones, Anna V Tinker, and Anthony N Karnezis. All authors approved the final manuscript.

Grewal JK, Eirew P, Jones M, Chiu K, Tessier-Cloutier B, Karnezis AN, Karsan A, Mungall A, Zhou C, Yip S et al. Detection and genomic characterization of a mammary-like adenocarcinoma. Molecular Case Studies. 2017 November 21.

The work described in Chapter 3 was written entirely by myself and developed jointly by Dr. Basile Tessier-Cloutier and myself. I was the main bioinformatics analyst and Dr. Tessier-Cloutier led the pathology analysis and evaluation. The study was jointly designed by me and Dr. Tessier-Cloutier with supervision from Dr. Steven J.M. Jones and Dr. Stephen Yip. All other authors contributed equally to study design, implementation, interpretation, and writing. A version of Chapter 3 highlighting the clinical implications will be submitted for publication as follows. '*' indicates co-first authors.

Tessier-Cloutier B*, Grewal JK*, Jones M, Pleasance E, Shen Y, Cai E, Dunham C, Hoang L, Horst B, Huntsman D, Ionescu D, Karnezis AN, Lee A, Lee CH, Lee TH, Mungall A, Mungall K, Naso JR, Ng T, Schaeffer DF, Sheffield BS, Skinnider B, Smith T, Williamson L, Zhong E, Laskin J, Marra M, Gilks CB, Jones SJM, Yip S. The impact of whole genome and transcriptome sequencing on diagnostic accuracy and treatment planning.

A version of Chapter 4 has been published in JAMA Network Open. A license to reuse the text and figures was not necessary since the article was published under the CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the author and journal are credited. The study was conceptualized and designed by Drs. SJM Jones, Martin Jones, Marco Marra, Michael Taylor and myself. I assisted in the collection and interpretation of the data in conjunction with Drs. Tessier-Cloutier, M Jones, Gakkhar, Ma, Moore, Mungall, Zhao, Mungall, Gelmon, Lim, Renouf, Laskin, and Yip. I performed the statistical analysis in collaboration with Sita Gakkhar. I created the experimental design for the classifier development, undertook the analysis and wrote the full initial draft. Dr. Jones devised the concept of the project. Early versions of the research design and assessment were developed in collaboration with Dr. SJM Jones, Sita Gakkhar, and Dr. Basile Tessier-Cloutier. All other authors contributed equally to the manuscript.

Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, Mungall AJ, Zhao Y, Taylor MD, Gelmon K et al. Application of a neural network whole transcriptome–based pan-cancer method for diagnosis of primary and metastatic cancers. JAMA network open. 2019 April 26.

The published case study referenced and re-analyzed in Chapter 5 was originally published in Cold Spring Harbor's Molecular Case Studies and the citation is provided below. I assisted in collation of the genomic findings and provided expertise on interpretation of results from the supervised cancer-type classifier used in this analysis. A license to reuse the text and figures was not necessary since the authors retain the copyright for the publication, licensed under CC-BY.

Ko JJ, Grewal JK, Ng T, Lavoie JM, Thibodeau ML, Shen Y, Mungall AJ, Taylor G, Schrader KA, Jones SJM et al. Whole-genome and transcriptome profiling of a metastatic thyroid-like follicular renal cell carcinoma. Molecular Case Studies. 2018 December 17.

I conceptualized the study in Chapter 5 jointly with Dr. Jones. I was the main researcher for this work and developed all the presented code and analysis. This work was written entirely by myself, supervised by Dr. Jones. Drs. Pleasance, Csizmok and Williamson provied guidance for interpretation of results and development of statistical analysis methods. All other authors contributed equally to the editing and review. A version of this chapter has been submitted for publication as follows:

Grewal JK, Pleasance E, Csizmok V, Williamson L, Wee K, Bleile D, Shen Y, Tessier-Cloutier B, Yip S, Renouf DJ, Laskin J, Marra M, Jones SJM. Single-sample pathway analysis using Pathway Impact Evaluation (PIE) of machine-learning based cancer classifiers.

The Introduction and Conclusion chapters are original work and have not been published or submitted for publication elsewhere.

Table of Contents

Abstract iii

Lay summary v

Preface vi

Table of Contents ix

List of Tables xiii

List of Figures xv

List of Abbreviations xxv

Acknowledgements xxviii

Dedication xxxi

1 Introduction 1
- 1.1 Pathology and cancer diagnosis 2
  - 1.1.1 The history of cancer diagnosis 3
  - 1.1.2 Cancer classification using histopathology 6
  - 1.1.3 Pathology in recent decades 9
  - 1.1.4 Diagnostic challenges in pathology 12
  - 1.1.5 Impact of diagnosis on treatment 14
- 1.2 Genomics and cancer diagnosis 15
  - 1.2.1 Computational algorithms for cancer classification . . 24
  - 1.2.2 Beyond the diagnosis - identifying biological changes in individual tumours 29
- 1.3 Objectives and chapters overview 34

2 Background 38
- 2.1 Case study 39

2.1.1 Clinical background . . . 40
2.1.2 Methods . . . 42
2.1.3 Pathology analysis and findings . . . 44
2.1.4 Genomic analyses . . . 48
2.1.5 Clinical decision and outcome . . . 54
2.2 Summary . . . 55
2.3 Conclusion . . . 56

3 Impact of genomics on diagnostic pathology in a precision oncology trial . . . 58
3.1 Methods . . . 60
3.1.1 Consent and institutional review board process . . . 60
3.1.2 Tissue biopsy and processing . . . 60
3.1.3 Library construction and sequencing . . . 60
3.1.4 Determination of tumour type . . . 61
3.1.5 Assessment of clinical input of whole-genome and transcriptome analysis in pathology . . . 62
3.2 Results . . . 63
3.2.1 Cohort demographics, clinical metrics, and sequencing data . . . 63
3.2.2 Correlation of histopathologic diagnosis and next generation sequencing results . . . 65
3.3 Discussion . . . 76
3.4 Conclusion . . . 78

4 Development and validation of SCOPE - supervised cancer origin prediction using expression . . . 80
4.1 Background . . . 81
4.2 Methods . . . 83
4.2.1 Training data . . . 83
4.2.2 Test data . . . 85
4.2.3 Model training . . . 87
4.2.4 Algorithmic model selection . . . 91
4.2.5 Ensemble selection . . . 92
4.2.6 Feature weights analysis for neural network . . . 93
4.3 Results . . . 93
4.3.1 Association of classification anomalies and biological similarities in held-out set . . . 94
4.3.2 Prioritization of known diagnostic gene features without prior knowledge . . . 97

4.3.3 External validation on primary cancers 99
4.3.4 Providing diagnosis for pre-treated metastases 100
4.3.5 Identification of putative primary tumour type for cancers of unknown primary 102
4.3.6 Impact of feature removal on classification 104
4.4 Conclusion 105

5 Enabling cancer transcriptome analysis from SCOPE using single-sample pathway impact evaluation (PIE) 107
5.1 Background 108
5.2 Methods 109
5.2.1 Test Data 109
5.2.2 Classifier used for PIE measurements 110
5.2.3 Pathway analysis for individual samples 110
5.2.4 Cohort-level pathway analysis 111
5.2.5 Statistical selection of top pathways associated with each cancer type 113
5.2.6 Statistical identification of important pathways for single-sample analysis 113
5.3 Results 114
5.3.1 Pathway impact profiles allow clustering and analysis of samples by cancer type 114
5.3.2 Pathway impact scores reveal prostate cancer subgroups 124
5.3.3 PIE independently recovers sample-level findings from integrative genomic analysis 126
5.3.4 PIE enables sample-level genomic analysis of cancers with unknown primary 129
5.4 Discussion 134

6 Conclusions 136
6.1 Contributions 137
6.1.1 Impact of genomic information on diagnosis of advanced cancers 137
6.1.2 Algorithmic advances in cancer classifier development 138
6.1.3 Interpreting cancer classification decisions and performing single-sample pathway analysis 139
6.2 Limitations of developed tools 140
6.3 Broader challenges in clinical translation 141

6.3.1 Management of diagnostic inaccuracies in clinical practice . . . 142
6.3.2 Facilitating adoption in routine practice . . . 144
6.3.3 Ensuring equitable access to developed tools . . . 146
6.3.4 Keeping classifiers up-to-date . . . 147
6.3.5 Incorporating other -omics technologies in automated diagnosis . . . 148
6.3.6 Utilizing single-cell sequencing for interrogation of cancer genomes . . . 149
6.4 Final words . . . 150

Bibliography . . . 152

Appendices

Appendix . . . 177

List of Tables

1.1 Diagnostic challenges in cancer histopathology 14

2.1 Details of sequencing experiments. 44
2.2 SNVs of interest are listed, along with details on the counts of the supporting reads spanning the tumour genome at the mutated and reference bases, in the tumour genome (transcriptome). 49
2.3 Copy number variants of interest in the tumour genome are listed, along with percentile values and fold changes calculated from the respective RPKMs against a background of TCGA Breast cancers. 50

3.1 Classification outcome from SCOPE for the cancer cohorts. . 72

4.1 Cancer types used for training, with abbreviations referenced in text. 83
4.2 Breakdown of cancer types in the external metastatic cohort. 86
4.3 Architecture, identifying names, and additional information for each neural network in the SCOPE ensemble. 93
4.4 Performance of SCOPE on the Genentech cohort of primary mesotheliomas. The training cohort was composed of epithelioid mesotheliomas, whereas the testing cohort was composed of epithelioid mesotheliomas and sarcoma-like mesotheliomas. Mesotheliomas that also show sarcoma-like histology are either predicted correctly as part sarcoma, part mesothelioma ("sarcomatoid mesothelioma"), or otherwise, usually as mesothelioma alone ("epithelioid mesothelioma"), or as sarcoma alone ("sarcoma"). 99
4.5 Performance of SCOPE on the metastatic cohort. Number of mis-predictions are listed in brackets if more than one. 101

1 Important genes based on frequency analysis of gene weights for each neural network in SCOPE. 177
2 Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. 183
3 Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. 237

List of Figures

1.1 Histologic classification of cancers based on their organ-system of origin. Various organ-systems of origin have multiple cancer-types associated with them, differing by the cell-type they originate from. 8

1.2 Thesis overview and key contributions. In this thesis we explore the utility of bulk RNA-Seq as a diagnostic and analysis aid in personalized oncogenomics initiatives. In a detailed retrospective study we review the frequency of diagnostic changes motivated by genomic data and molecular observations. We develop an automated, open-access tool (SCOPE) for cancer classification using large, representative RNA-Seq profiles. We then extend this method to provide pathway-level profiles of individual cancer samples, also made available as an open-access tool (PIE). 36

2.1 Clinical history and pathology sampling timepoints for MLAV patient. Initial treatment is indicated in orange, tumour biopsies at various time-points following metastasis indicated with purple lines, and treatments provided based on genomic analysis are shown with purple drug symbols over dark-grey timeline bars. Tumour biopsies on which immunohistochemistry was performed are shown with open circle termination of corresponding line. Abbreviations: IHC - Immunohistochemistry test, POG - Personalized OncoGenomics clinical trial (Clinical Trial number: NCT02155621). 42

2.2 Histopathology of biopsies retrieved from MLAV Patient. A) The biopsy of the vulvar mass shows a poorly differentiated tumour composed of nests and cords of pleomorphic tumour cells. B) The HER2 immunostain on the initial vulvar mass biopsy is equivocal, compatible with score 2+ based on predominantly incomplete, weak and moderate membrane staining within greater than 10% of tumour cells. C) The fine needle aspirate of the recurrence lesion from the supraclavicular lymph node shows clusters of pleomorphic tumour cells (H&E stain). D) The HER2 immunostain of the supraclavicular lymph node shows tumour cells with complete, intense membrane staining in greater than 10% of tumour cells compatible with score 3+. 47
2.3 ERBB2 gene's genomic locus is shown in the patient's tumour. A) A lollipop plot showing the coordinates of the S310F gain-of-function mutation observed in this case. B) A plot of the copy number landscape of Chromosome 17 in the tumour. The ERBB2 copy-number gain is indicated. 52
2.4 Correlation plots of the cancer's RNA-Seq profile with TCGA cancer datasets. A) Boxplot distribution of the pairwise Spearman correlation of the recurrence biopsy's gene expression profile and all TCGA samples. The x-axis represents cancer types following TCGA naming conventions. TCGA breast cancer cohort is indicated by BRCA. B) Boxplot distribution of the pairwise Spearman correlation between the recurrence biopsy and the TCGA breast cancer cohort based on the PAM50 set of genes. The pairwise correlations with adjacent normal are shown in blue. 53

3.1 Cohort selection for the assessment of the impact of DNA and RNA sequencing analysis on histopathologic diagnosis in the POG clinical trial. 64
3.2 Tumour types in the cohort are shown, along with the type of genomic data guiding major outcomes from the retrospective analysis evaluating the diagnostic utility of RNA-Seq and WGS. 66

3.3 Detection of clinically relevant molecular alterations by whole-genome and RNA sequencing in the POG cohort. (A-C) Detection of HER2 amplification in a colorectal carcinoma is shown, as indicated by immunohistochemistry (IHC) staining for HER2 (overexpression, 3+) in the tumour sections in panels A) and B), and with FISH testing for additional copies of HER2 (HER2 to chromosome 17 centromere (CEP17) ratios > 2.0) in panel C). (D-F) ALK fusion identified in a lung adenocarcinoma, missed on initial FISH analysis. H&E staining of the tumour sample is shown in D). ALK IHC testing results showing equivocal ALK staining are represented in E), with the original negative FISH results (break apart probe test, less than 15% of cells showed break apart probes) shown in F). (G and H) Detection of an IDH1 mutation in a CUP supported the putative diagnosis of cholangiocarcinoma. The H&E staining is shown in G). Panel H) shows a snapshot of the Integrative Genomics Viewer track for the mutation location with proportional read-counts supporting the reference (G, in orange) and mutation (A, in green) in the tumour genome. This supported the putative diagnosis of this CUP as a cholangiocarcinoma in the clinical context, as aided by RNA-Seq analysis. 68

3.4 The outcome from genomic analysis is shown separated by A) the site of biopsy of the tumour, and B) the organ-system of origin of the cancer. M and P indicated the number of metastatic and primary/relapse samples respectively. 69

3.5 The final diagnoses for the 15 CUP cases and 2 cases with revised diagnosis are shown, along with the type of genomic data guiding each of the outcomes. WGS = Whole-genome sequencing. 70

3.6 Impact of tumour content on the ability of RNA-Seq to provide the correct putative diagnosis in the POG cohort. The majority of samples arose from 3 biopsy sites - lymph node, lung, and liver, indicated in each of the panels. Wilcox test for significance between SCOPE outcome matching final diagnosis, versus each of the other categories: * $p =< 0.05$; ** $p =< 0.01$; *** $p =< 0.001$; ns $p > 0.05$ 74

3.7 Impact of tumour content on the ability of RNA-Seq to provide the correct putative diagnosis in the POG cohort, agnostic of biopsy site. Wilcox test for significance between SCOPE outcome matching final diagnosis, versus each of the other categories: * p =< 0.05; ** p =< 0.01; *** p =< 0.001; ns p > 0.05 75

4.1 Performance of SMOTE as compared to other class expansion methods. Cross-validation results on the TCGA training dataset are shown. Abbreviations: dup - duplication of samples in small classes, none - no class expansion applied, weight - inverse cost for misclassification of smaller classes during training. 88

4.2 Results from algorithm and feature selection experiments, and performance on held-out test set. A) Feature selection does not improve pan-cancer classification. B) Comparison of algorithms - performance of single neural network on held-out set is higher than other algorithms. C) Validation of SCOPE on TCGA held-out set demonstrates high discriminatory power amongst most cancer types. Point with bar represents average F1-score and standard deviation spread for corresponding category. Incorrect predictions for more than 10% of samples belonging to a given cancer type are shown by curved directed edges. Curve width indicates relative fraction of samples in misprediction set. Mispredictions occur amongst cancer types with the same organ-system of origin. Specific trends are discussed further in Section 4.3.1. 95

4.3 Performance of various models that make up SCOPE, on the cross-validation and held-out sets. The x-axis is ordered by increasing class size. Performance is reported as precision for the test-folds from CV *inblack* and for all samples in the held-out set *inyellow*. Number of samples in training are shown in the upper histogram panel. Cancer codes follow TCGA nomenclature and are defined in Table A.1, with _TS samples indicating tumours and _NS samples indicating adjacent normal tissues. The difference between CV-fold performance and held-out performance is typically larger for small classes. The difference become insignificant as class size approaches N > 100. When the classifier is augmented with addition of synthetic samples in the training folds (last panel), we observe an overall increase in performance for the smaller classes with a concomitant reduction in the performance gap between mean-CV-precision and heldout precision. The line of best fit (loess) is indicated for each model, with standard error bounds in grey. The spread of performance across different CV folds is shown by the black point (mean) with 1 standard deviation bars. 96

4.4 t-SNE plot of transcriptomic data in TCGA training cohorts. The relevant gynecologic and gastrointestinal cancer types are shown, and reflect the trends of cross-calling observed in SCOPE. Esophageal adenocarcinoma $ESCA\_EAC$ and stomach adenocarcinoma $STAD$ cluster together, as do uterine carcinosarcomas UCS with uterine corpus endometrial carcinomas $UCEC$. 98

4.5 Performance of SCOPE on external metastatic cohort. A) Two-sided t-tests show a significant association of tumour content on general diagnosis as organ system, for biopsies samples from site of metastasis. B) Two-sided t-tests show no effect of tumour content on misclassification to organ system, for biopsies sampled from the cancer's site of origin. C) SCOPE has improved performance compared with baseline linear comparator trained from a statistically filtered feature subset. Abbreviations: AC - adenocarcinoma, CA - carcinoma, SCC - squamous cell carcinoma, CESC AC - cervical/endocervical adenocarcinoma, UCEC - uterine corpus endometrial carcinoma. 103

4.6 SCOPE prediction and putative primary for cancers with unknown primary site. A confusion matrix of predictions is shown, where the size of the circles represents relative number of samples in each category. Case count for CUPs by putative origin is shown with a histogram on the right. Correct predictions are indicated in yellow whereas incorrect ones are shown in black. Salivary carcinoma, neuroendocrine tumours, and ewing sarcomas were not present in SCOPE training, explaining the inability of the method to identify these accurately. Abbreviations: CA - carcinoma, AC - adenocarcinoma. 104

5.1 UMAP projections of PIE profiles for 3,963 biochemical pathways, for samples in the TCGA cohort of primary tumours. For ease of readability, the projections in panel A) show TCGA tumour types coloured by their organ system of origin. The spread of sample-specific silhouette indices, grouped by cancer type, is shown in panel B). 115

5.2 Pathways commonly associated with multiple cancer types in the TCGA cancers are shown. Grey bars indicate total number of tumour samples evaluated, whereas coloured bars indicate the number of tumour samples from the respective organ-system of origin. Panel A) shows the most common cell-function pathways. Panel B) shows the most common cancer-associated pathways. 117

5.3 Statistically significant pathways in TCGA cancers. Panels show important pathways for each tumour and normal category, grouped by organ system of origin. Each group shows the top-5 pathways associated exclusively with cancers in the relevant organ-systems of origin, ordered by number of samples in which the pathway had a positive PIE score. Coloured bars indicate fraction of tumour samples from the organ-system where the respective pathway was positively scored by PIE. 118

5.4 Determination of pathway-level activities for TCGA primary cancers, using PIE. Panel A) shows the number of statistically significant positively associated with each cancer-type, from the group of 3,963 pathways evaluated using PIE. Panel B) shows the number of pathways with statistically significant PIE scores per sample. 119

5.5 Clustering of POG cohort samples by cancer-type using PIE profiles for 3,963 biochemical pathways. Cancer types with at-least 10 (N = 510/602) are shown for ease of readability. A) UMAP projections of pathway profiles are shown. Using pathway importance scores, samples cluster by their diagnosed cancer type. B) Silhouette indices of samples are shown, grouped by cancer type. A positive silhouette index indicates sample clusters with assigned cancer-type. . . 121
5.6 Clustering of MET500 cohort samples by cancer-type using PIE profiles for 3,963 biochemical pathways. For ease of readability, the projections only show cancer types with at-least 10 samples in the MET500 cohort (N = 259/375). A) UMAP projections of pathway profiles are shown. Using pathway importance scores, samples cluster by their diagnosed cancer type. Of note, we observe 3 distinct clusters of prostate adenocarcinoma (PRAD, in dark-green). B) Silhouette indices of samples are shown, grouped by cancer type. A positive silhouette index indicates sample clusters with assigned cancer-type. 122
5.7 Silhouette index spread for the MET500 cohort subtypes. Silhouette metrics are calculated from the UMAP projections initialized with the first two principal components; clusters evaluated based on cancer type annotation. A positive silhouette index indicates sample clusters with assigned cancer-type. Abbreviations: BLCA – Bladder cancer, BRCA – Breast cancer, IDC – invasive ductal carcinoma, ILC – invasive lobular carcinoma, CHOL – Cholangiocarcinoma, EHCH – extrahepatic CHOL, IHCH – intrahepatic CHOL, COADREAD – colorectal adenocarcinoma, ESCA – esophageal carcinoma, SCC – squamous cell carcinoma, EAC – adenocarcinoma, OV – ovarian cancer, PRAD – prostate adenocarcinoma, SARC – sarcoma, RHBD – rhabdoid, LMS – leiomyosarcoma, EW – Ewings Sarcoma, UPS – Undifferentiated pleomorphic carcinoma, DDL – dedifferentiated sarcoma, SKCM – subcutaneous melanoma. 123

5.8 Cohort comparison between the top 25 pathways associated with breast cancer, for The Cancer Genome Atlas (TCGA) cohort of primary cancers, the POG cohort of metastatic tumours, and the MET500 cohort of metastatic tumours. Panel A) shows the number of unique and shared pathways between each of the cohorts. The MET500 and POG cohorts are grouped as 'metastatic'. Pathways common between primary and metastatic cancers (in purple), exclusive to primary cancers (in orange), and common within the metastatic cohorts (in light blue) are shown in panel B) with the corresponding mean PIE score across samples on the y-axis. 125
5.9 UMAP projections of the MET500 cohort are shown, filtered to view only the prostate adenocarcinoma samples. UMAP projections (initialized by the first two principal components) are calculated based on A) sample pathway importance profiles calculated automatically by PIE for 3,963 pathways, and B) gene expression profiles of the samples (RPKM values). Panel B) also suggests a non-random separation of the samples. 127
5.10 Top 25 pathways driving the 3 distinct clusters observed for the prostate adenocarcinomas in the MET500 cohort. 128
5.11 Top 25 pathways from PIE-based pathway analysis of a mammary-like vulvar adenocarcinoma. 40% of the pathways shown here overlap with the integrative pathway analysis (in yellow, green), and 16% are associated with paclitaxel therapy that the patient had received previously (in green, blue). Size of pathways is indicated in brackets next to the pathway name on the y-axis. The right panel shows the number of genes shared between the integrative analysis (N = 50) and the indicated pathways. Distribution of PIE scores for the remaining 65 classes is shown in grey, for each pathway. 130

5.12 Top 25 pathways identified by automated pathway impact analysis using PIE, for a cancer of unknown primary that was later diagnosed as a rare thyroid-like follicular renal cell carcinoma. Size of pathways is indicated in brackets next to the pathway name on the y-axis. Panel on the right shows the number of genes from integrative analysis (N = 34) that overlap with the genes in each of the pathways. 48% of the pathways in the main panel overlap with manual integrative pathway analysis findings (in yellow, red), of which 12% associated with the actual rare cancer type that this cancer represented (in red). Distribution of PIE scores for the remaining 65 output classes is shown in grey, for each pathway. 132

5.13 Comparison of pathway importance scores for two different output categories – renal clear cell carcinoma (KIRC) and renal papillary carcinoma (KIRP). Scores were calculated by PIE using the SCOPE classifier output. The input was the RNA-Seq profile of a cancer of unknown primary, later diagnosed as a rare follicular renal cell carcinoma that molecularly aligned with KIRP. The pathways that were important for classification of the sample as KIRP instead of KIRC are highlighted in yellow. Pathways important for classification of the sample as KIRC instead of KIRP are shown in blue. As is evident, the magnitude of the pathway importance is higher for pathways driving the classification of KIRP over KIRC. Relevant pathways have been labelled. . 133

1 Example output from SCOPE for a sarcomatoid mesothelioma, predicted with split confidence as mesothelioma and sarcoma. 181

2 Mean prediction accuracy of SCOPE as RPKM values of various fractions of genes are set to 0 in the input RNA-Seq data. Grey bars around mean points indicate standard error bounds. Black line indicates the line of best fit (loess). At a given threshold n% genes in input were randomly set to zero. This was repeated 10 times for each n in (10, 20, 30, 40, 50, 60, 70, 80, 90, 99). 181

3 UMAP projections of PIE profiles for 3,963 biochemical pathways, for samples in the TCGA cohort of primary tumours.The projections are coloured by tumour-type. . . . 182

4 Pathway importance for various Androgen Receptor associated pathways for the MET500 Prostate Adenocarcinoma samples, separated by observed cluster groups. 250
5 Manual integrative analysis of a mammary-like vulvar adenocarcinoma. Colour of circles shows fold expression change of the respective gene in the sample, relative to a background of all healthy normal tissues from GTEx. Box adjacent to circle indicates percentile expression compared to the Cancer Genome Atlas' cohort of breast cancers. Over-expression is shown in red, and loss of expression in blue. The key oncogenic pathways impacted in this case are shown with grey boxes and red border. Manual analysis identified activation of ERBB2/ERBB3, mTOR pathway, and the MAPK pathway. Overexpression of various genes participating in transcriptional regulation and metabolism was also identified (shown in red borders). 251
6 Manual integrative analysis of a cancer with unknown primary, which was diagnosed as a thyroid-like follicular renal cell carcinoma molecularly similar to renal papillary carcinoma. Colour of circles shows fold expression change of the respective gene in the sample, relative to a background of healthy renal tissues. Box adjacent to circle indicates percentile expression compared to the Cancer Genome Atlas' cohort of renal papillary carcinomas. Over-expression is shown in red, and loss of expression in blue. The key oncogenic pathways impacted in this case are shown with grey boxes and red border. 252

List of Abbreviations

Short	Long
ACC	Adrenocortical Carcinoma
ACYC	Adenoid Cystic Carcinoma
BCCA	British Columbia Cancer Agency
BLCA	Bladder Carcinoma
BRCA	Breast Carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
CI	Confidence interval
CIHR	Canadian Institutes of Health Research
CNV	Copy Number Variant
COADREAD	Colorectal adenocarcinoma
COSMIC	Catalog of Somatic Mutations in Cancer
CUP	Cancer with Unknown Primary
CV	Cross-validation
DLBC	Diffuse Large B-cell Lymphoma
DLBC-BM	Diffuse Large B-cell Lymphoma (Bone Marrow)
EMPD	Extra-mammary Pagets disease
ESCA	Esophageal carcinoma
ESCA-EAC	Esophageal adenocarcinoma
ESCA-SCC	Esophageal Squamous Cell Carcinoma
ET	Extra-trees
FL	Follicular lymphoma
GBM	Glioblastoma multiforme
GSC	Genome Sciences Centre
GTEx	Genotype-Tissue Expression
HNSC	Head and Neck squamous cell carcinoma
HPV	Human Papillomavirus
ICGC	International Cancer Genome Consortium
IHC	Immunohistochemistry

(continued)

Short	Long
INDEL	Insertion/Deletion event
IQR	Inter-quartile range
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LOH	Loss of heterozygosity
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MB-Adult	Adult medulloblastoma
MESO	Mesothelioma
MISC	Miscellaneous
MLAV	Mammary-like adenocarcinoma of the vulva
NCI-GPH-DLBCL	Diffuse Large B-cell Lymphoma (National Cancer Institute cohort)
NN	Neural network
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PIE	Pathway Impact Evaluation
POG	Personalized OncoGenomics
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
RF	Random forest
RNA-Seq	RNA Sequencing
RPKM	Reads Per Kilobase of transcript per Million mapped reads
SARC	Sarcoma
SCOPE	Supervised Cancer Origin Prediction using Expression
SKCM	Skin Cutaneous Melanoma
SMOTE	Synthetic Minority Oversampling Technique
SNV	Single Nucleotide Variant
STAD	Stomach adenocarcinoma
SVM	Support vector machine
TCGA	The Cancer Genome Atlas

(continued)

Short	Long
TFRI-NCL-GBM	Glioblastoma multiforme (Terry Fox Research Institude cohort, non-cell line)
TGCT	Testicular Germ Cell Tumours
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma
WES	Whole exome sequencing
WGS	Whole genome sequencing

Acknowledgements

I would like to thank my supervisor, Dr. Steven Jones, for the supportive environment and mentorship he has provided me throughout my degree. The independence granted me to pursue collaborations, research directions, and other opportunities beyond academia have enabled me to explore my research and teaching interests to their limits, and for that, I am extremely grateful. Steve's wisdom and foresightedness - percolated through numerous long hours in the office and at the pub - continue to serve as life-long lessons in mentorship, team management, problem-solving, and interpersonal relationships. Special thanks to Louise Clarke and Sharon Ruschkowski for their administrative support through the years. Meetings would be impossible to schedule and paperwork a torturous exercise if not for Louise's ever-smiling presence and incredible organizational skills.

I would also like to thank my committee members, Drs. Stephen Yip, Ryan Morin, Sohrab Shah, and Inanc Birol, for their mentorship during these years. Their unique perspectives and eternal wisdom have helped me appreciate the practicalities of interdisciplinary research and provided me the motivation to keep learning beyond what I know. Dr. Morin was instrumental in giving me my first proper taste of bioinformatics research in my undergraduate days - without his support and advice during that time, I would not have been nearly as confident embarking upon my PhD. Thankyou to Dr. Birol for constantly teaching me the humble lesson that every 'cancer sample' we study is a person with dreams and aspirations. Dr. Shah's meticulousness, eye for detail, and command over mathematics principles have kept me on my toes, providing impetus on those rare days of procrastination. My thanks to Dr. Yip in particular for being akin to a second supervisor to me, guiding me into the world of histopathology and spearheading our various collaborations. The visits to the cancer genomics laboratory and dissections have provided me with better understanding of cancer genomics in the clinic and a deep appreciation for the medical profession. The opportunity to work with residents like Drs. Basile Tessier-Cloutier and Adrian Levine has been

a privilege and contributed significantly to this research.

The last few years would have been very dull without the guidance and cheerleading abilities of several current and former lab members, particularly Drs. Jake Lever, Erin Pleasance, Yaoqing Shen, Martin Jones, Laura Williamson, Zoltán Bozóky, Martin Krzywinski, and Eric Zhao. The long days spent coding, debugging, and thinking about thinking about writing were made shorter by Jake and Eric's constant concern and encouragement. Thankyou, Erin, Yaoqing, Laura, Zoltan, and both Martins, for providing me the much needed guidance in research efforts, faith in my abilities, and abstract musings to keep the gears well-oiled along this journey. To Martin K in particular, thankyou for opening up my eyes to the importance of data visualization, providing me the opportunity to contribute to the Points of Significance series, relating much-needed philosophical quips and musings during stressful times, and often, as a consequence of our invigorating long chats, making me late for other meetings.

My fellow graduate trainees in the lab, including Luka Culibrk, Michael Disyak, Emre Erhan, Jenny Yang, and Vahid Akbari, have been a constant source of inspiration. I am particularly thankful to have worked alongside Luka, Micha, Emre, and Jenny in their research endeavors.Their curiousity, intelligence, and enthusiasm have helped preserve the wide-eyed child in me throughout. Drs. Robin Coope, Andy Mungall, Gordon Robertson and Kieran O'Neill have been invaluable in helping me understand the larger world of scientific research and the gears that keep it moving. I am grateful to them for their valuable feedback and advice in situations where I could not map out a clear way forward myself. I hope we continue to keep in touch as we all advance through life.

My friends and family have been vital in making me the person I am today. My immense love to my *nanaji* (grandfather) Jaswant Sran and my *naniji* (grandmother) Gurdev Sran, who have constantly believed in me. My deepest gratitude to my parents, Jaswinder and Gurmukh, for their vision and guidance, and to my *mamiji* and *mamaji*, Sarb and Jasdev, who have been excellent aunt and uncle + proxy parents for the last 10 years. This degree is as much theirs as it is mine. I would not have nearly the same drive and enthusiasm for my research areas if not for the mentorship of my high-school teachers Ms. Ekta Bali, Ms. Hampreet Sidana, and Mr. Doug Barham. Their dedication to the profession and their eagerness to do best by their students gives me lots to aspire to. My sister Baldeep and my

cousins Shub, Sehaj, and Seerat have brightened up my years with their laughter and smiles - an important factor in rainy Vancouver. The company of life-long friends - Kern, Archit, Onishma, Bhavit, and Varun - and that of my peers in the bioinformatics training program has also served as vital motivation and support. Most crucially, a big thanks to my boyfriend and rock, John Dupuis, for his love, patience, and confidence in me throughout comps, thesis writing, and the years in-between. His beyond-par baking skills and Wednesday beer club have fuelled me throughout this degree, and his family has been a warm and welcoming refuge from the worries of research obligations.

Lastly, I must thank the many members of the personalized oncogenomics (POG) team, including the patients and their families who put their trust in this initiative. The exemplar leadership by Drs. Marco Marra, Steven Jones, and Robyn Roscoe have been an inspiration to become the best scientist version of myself. Thankyou to Jessica Nelson for putting up with and resolving my frequently obtuse inquiries about the clinical data. The vast majority of this work would not be possible without the various online contributors and developers of StackOverflow, RStudio, BioRender, and Python, and the various members of the GSC's systems team. Various funding agencies have supported this research, including a UBC Four Year Fellowship and travel fellowships from Canada's Michael Smith Genome Sciences Centre, CIHR, and the Canadian Cancer Society. Part of the presented work was supported by the BC Cancer Foundation and Genome British Columbia (project B20POG). I also acknowledge contributions towards equipment and infrastructure from Genome Canada and Genome BC (projects 202SEQ, 212SEQ, and 12002), Canada Foundation for Innovation (projects 20070, 30981, 30198, and 33408), and the BC Knowledge Development Fund.

Dedication

To my father Gurmukh, my mother Jaswinder, my sister Baldeep, and my boyfriend John. This thesis is sponsored by the many hugs, conversations, and delicious food supplied by them.

Chapter 1

Introduction

At the microscopic level, organs are distinguished from each other by the cells that make them up. Cells in different tissues and organs behave differently from each other. This behaviour can be characterized by differences in gene expression. It is an inherent biological property of each cell type. Tumours begin when healthy cells start dividing in an uncontrolled manner. This can happen due to exposure to carcinogens that cause DNA damage, or from small changes in DNA that accumulate over time. Tumours can be benign, as is the case with moles that arise from skin cells called melanocytes. In certain cases, tumours can become cancerous, starting to compete with surrounding healthy cells for space and resources. Cancerous lesions also exhibit local and distant invasion. When cancers move away from their site of origin (primary site) to other sites in the body, they are said to have metastasized. Metastasis can be facilitated by the lymphatic or circulatory systems. Cancer metastasis is the primary cause of cancer morbidity and mortality. Advances in cancer treatment have resulted in effective management or even complete cure of cancers if detected and diagnosed prior to metastasis [25]. However, metastatic cancers are challenging to treat, resulting in about 90% of all cancer-associated deaths in North America [25].

Cancer treatment is conventionally based on the site of origin of the tumour. Primary and metastatic cancers usually retain certain biological features characteristic of their primary cell type. These features can be physical (the shape of cells, how cells organize themselves), or molecular (expression patterns of genes). The identification of the site and cell type a cancer originated from is dependent on various factors. In most cases, a cancer diagnosis - characterizing the type of cancer and accompanying treatment option - is provided to patients using guidelines that consider the tumour cells' appearance and the patient's clinical history. However, morphology-based diagnosis is a challenging task, involving sequential interrogation of the tissue sample using various established histopathology methods. Previous studies have found that histopathology

based misdiagnoses can range from 10-20% simply due to differences in interpretation of results by different pathologists [70]. Other studies have found that misdiagnosis rate can go up to 50% depending on the type of cancer [39, 69, 198].

Cancer can be considered a disease of the genome. Presumably then, the genomic profile of a cancer provides more reliable diagnostic assessment than manual inspection of tissue morphology. The work presented in this thesis explores the relevance of cancer diagnosis in precision oncology and outlines the utility of gene expression data in diagnostic pathology. It then proposes machine-learning approaches to integrate RNA-sequencing experiments into diagnostic workflows and genomic analysis of rare and advanced cancers.

In this chapter the reader will be familiarized with the essential aspects of cancer diagnosis - how it is obtained, why it is needed, and what some of the recent technical and biological advances in the field are. The first section deals with the origin and evolution of diagnostic pathology over the last 3 centuries. The second section outlines the contribution of genomics to cancer diagnosis in recent decades and introduces the reader to technical concepts of relevance in subsequent chapters. We conclude the chapter by outlining the key objectives of this thesis accompanied by a brief overview of the following chapters.

1.1 Pathology and cancer diagnosis

Evidence-based medicine bases itself on the bedrock of medical guidelines. These constantly evolving guidelines form the criteria for diagnosis and management of diseases. Particularly in cancer, the appearance of cells and nuclei determines whether abnormal tissue growths are cancerous or not. The field of diagnostic pathology establishes these morphology-based guidelines in cancer care.

The widely used Pap smear test is the most common example of the primary means for cancer diagnosis. Dr. George Papanicolau developed the test in 1923 to identify cervical cancer. It involves exfoliating cells from the cervix through scraping and examining these cells microscopically for cancerous behaviour. The low cost, ease of administration, and accurate interpretation made this test accessible, significantly reducing global cervical cancer incidence [190]. The approach itself is known as histopathology and forms the basis for modern-day cancer diagnosis. Two

key ideas that established the field - cell staining and cell theory - emerged from extensive investigation of disease origin and characterization in the 17th to 19th centuries.

1.1.1 The history of cancer diagnosis

By the early 1600s, it was well-recognized that many illnesses are correlated with changes inside the patient's body. Giovanni Morgagni, an Italian pathologist and anatomist, was the first to use broad anatomical findings for routine diagnosis of several diseases - including tumours. In his 1769's seminal work, "The Seats and Causes of Diseases Investigated by Anatomy", he noted tumourous growths on the maxillary gland of a patient who had trouble swallowing. Dr. Morgagni's work also documented the spread of cancer, observing that the same patient had some tumours in the pharynx and larynx. In 1775, the occurrence of cancers of the scrotum in chimney sweeps' was reported - the first incidence of an occupational cancer [78]. The biological cause of cancer was as yet unknown.

The 1800s saw the emergence of two notable theories on the origin of cells. In 1838, a German pathologist, Dr. Johannes Müller proposed that cancer is made up of abnormal cells originating from the body itself. Through examination of several microscopic samples of tumours, he determined that cancers possess distinct microscopic features that can be used to identify them, thereby establishing the field of histopathology [79]. In 1852, Dr. Robert Remak, a prominent embryologist working with Dr. Müller, leveraged membrane staining to observe cellular changes. Tracing the origin of new cells using this membrane staining technique, Dr. Remak was able to show that all cells arise from cells - 'omnis cellula a cellula'. This was a significant advancement on Dr. Müller's original hypothesis that cancerous cells arose from bodily fluids spontaneously. The ensuing debate, accompanied by the introduction of the compound microscope in medical research in the 1850s, shifted the assessment protocol for cancers from gross anatomic features to cellular changes [119]. Combined with Dr. Rudolph Virchow's subsequent work supporting this cell theory, the findings established the domain of cellular pathology [119].

Tissue excision

The use of tissue excisions to understand diseases was not entirely novel for the time. The first documented case of physicians studying the appearance

of a disease dates back to the 900's, when the Arab physician Albucasis (936-1013) used a thin needle to retrieve a tissue sample from the throat of a goiter patient - an approach known today as the fine-needle aspiration biopsy [136]. A few centuries later, in 1848, a German dermatopathologist used microscopic studies of tissue excisions to distinguish normal and abnormal skin [136]. The term 'biopsy', used commonly today to refer to tumour tissue specimens taken for histopathology analysis, was coined in 1879 by French dermatologist Ernest Besnier [136].

Over the last few decades, various approaches for biopsying an abnormal lesion have been established [40]. Cytologic examination of scraped cells can provide a quick overview of the tissue morphology. Detailed histologic examination is done using core-needle biopsies, whereby a special hollow needle is used to take a small cylinder-shaped (core) sample of the lesion. If the lesion cannot be excised easily or if the excision could lead to functional impairment, an incisional biopsy or a fine-needle aspirate is drawn to permit evaluation. In some cases, the tumour can be excised using surgical tools, providing an excisional biopsy for pathology. In challenging cases, the process can be guided by an imaging procedure like X-ray or computerized tomography.

Tissue sectioning

Once the tissue is biopsied, it needs to be treated such that it could endure long-term storage and study. Until the 1860s, tissue specimens would be prepared with fluid smears, or by scrapping the cut-surface of tissues [79]. As the microscopic resolution of lenses increased, so did the need to improve the preservation quality of tissue specimens and facilitate fine-grained study.

Edwin Klebs introduced paraffin embedding of tissues in 1869 [79]. Since then, various embedding techniques using waxes and resins have been developed to analyze different tissue specimens, but paraffin remains most suitable for embedding the broadest range of tissues. If the tissue is calcified (for example, bone), minerals have to be removed via decalcification first.

Prior to paraffin embedding, tissues need to be fixed to impart mechanical rigidity and withstand subsequent processing. No ideal fixative has been found to date - that is, one that preserves cellular morphology perfectly without compromising the specimen's composition or reactivity of proteins in the cell [194]. Formalin, a fixative discovered in the early 1900's, is used most commonly in the fixation process. Currently, formalin-fixed paraffin-embedded (FFPE) tissue sectioning is one of the two prominent

methods used to prepare tissue for analysis in pathology laboratories.

Dr. Virchow's student, Julius Cohnheim, was able to preserve tissue excisions as frozen sections in 1864. Unlike FFPE samples, frozen sections could be prepared within minutes and used to undertake additional work-ups (for example, special fixation and tissue staining). Today, frozen sections are preferred over FFPE for genomic analysis (DNA or RNA sequencing, for example) since the FFPE process damages nucleic acids, causing artefacts that compromises their quality [40]. FFPE tissues require 12-24 hours to prepare, but have better morphologic quality and are hence preferred over frozen tissues for archival purposes [40].

Several pre-analytical factors also influence the morphologic quality of prepared tissue sections from the FFPE process [7]. The time interval between the tissue's removal from the patient to the time it is fixed in formalin - known as the cold ischemia time - impacts antigen viability for various protein binding assays. The fixation time in formalin also impacts the availability of the antigens detected by pathology assays like immunohistochemistry. After this, the time and temperature of formalin fixation impact how well the preserved tissue can be used for molecular tests like *in situ* hybridization (discussed later) [7].

Tissue staining

After tissue sectioning (FFPE or frozen), thin sections of the tissue can be cut using a microtome, and stained to facilitate histopathology analysis. These stains aid in tumour differential diagnosis and classification [40]. Based on the type of stain used, various morphologic attributes in the tissue specimen can be highlighted. In the 1800s, Dr. Cohnheim used a silver stain to outline frozen sections of nerve endings. A few years later, it was found that hematoxylin and eosin stain different parts of the cell [4]. Hematoxylin is a naturally occurring chemical discovered in 1502. It was found in the 1800s that the dye binds to nuclear proteins (specifically, histones) in its oxidized form, imparting a deep blue to black colour [197]. Eosin Y is an acidic dye that binds to positively charged components of the cytoplasm, imparting a pink colour to the non-nuclear components of the cell. It is used as a 'counterstain' along-with hematoxylin [4]. H&E is the most widely used mixture of dyes in histology. Hematoxylin binds to the nuclei, painting them a bright blue, and eosin stains the extracellular matrix and cytoplasm pink. Other cell organelles take on a combination of these hues. In the case of FFPE samples, the method used for decalcification can impact the effectiveness of the H&E stain [7].

Additional stains exist to specifically identify bio-molecules like mucins, amyloids, lipids, glycogen, and elastic tissues. Van Gieson's stain or the Masson trichrome method, for example, can be used to distinguish collagen and muscle [40]. The stained sections are reviewed by trained pathologists, who draw upon their expertise in the area to determine the tumour characteristics. The triad of histology, microscopy, and pathologist examination forms the crux of modern-day cancer diagnosis [4]. It is the responsibility of the surgical pathologist to use available stains as needed and synthesize the resultant findings provided by tissue morphology and by each stain to provide a comprehensive diagnosis for the cancer patient [40].

1.1.2 Cancer classification using histopathology

The assessment of cell morphology to study diseases organically led to the development of classifications for healthy and malignant cells. In 1858, Dr. Rudolph Virchow observed that some patients had an abnormal number of white blood cells, naming the condition leukamie. It would be a few more decades before leukemia is classified as cancer in 1938. In the interim, histopathology remained broadly the same.

The healthy human body has various tissue types. The primary ones are the epithelium (outer layer of skin, mucosal tissues), supportive tissue (bone, cartilage, connective tissues), nerve tissues, lymphatic tissue, and the bone marrow. Abnormal tissue changes (lesions) are classified based on the cell-type of origin, and in some cases, using additional attributes that indicate how different the cancer looks from its counterpart normal cells. The nomenclature for all tumours, benign or malignant, is based on the normal tissue the tumour originates from. Current tumour classification systems encompass terms on biologic behaviour, cellular function, histology, embryonic origin, and anatomic locations. These designations carry well-defined clinical implications and are also important for communicating a diagnosis [155].

Benign lesions can give rise to localized tissue masses. Fibromas, chondromas, and adenomas fall into this group. They are distinguished from each other based on micro- and macro-scopic patterns that characterize their cells of origin. While not invasive, these masses can compete for space and resources, requiring their removal. For example, benign tumours in the nervous system are common. Due to their location, they can be harmful to the patient but their removal is also extremely difficult.

Malignant lesions, or tumours, are broadly classified into carcinomas, sarcomas, lymphomas, and leukaemias. Cancers that arise from epithelial cells are called carcinomas, and further subtyped based on the appearance the tumour takes. Particularly, malignant epithelial tumours with a glandular growth pattern are called adenocarcinomas, whereas those with a stratified distribution of cells are called squamous cell carcinomas. Sarcomas arise from mesenchymal tissues and are further subtyped based on their histogenesis. For example, cancers arising from fibrous tissues are called fibrosarcomas. Cancers arising in lymph nodes and other parts of the lymphatic system are called lymphomas, whereas cancers arising from the bone marrow are commonly grouped under leukaemias. These classifications are further described in Figure 1.1.

The cells' appearance can be graded along a scale of 1-3 indicating the degree of differentiation. Well-differentiated tumours closely resemble the healthy cells they originated from in form and structure, and have a low grade. Anaplastic tumours - those that do not display differentiation - can be poorly differentiated (with minor resemblances to the primary tissue) or undifferentiated. These tumours have a high grade. The more anaplastic a tumour, the less likely it is to have specialized functional activity. Since these tumours also lack the defining morphologic features that well-differentiated tumours possess, they are challenging to diagnose using histopathology. Low grade tumours typically have a good prognosis, and as the grade increases, the tumours tend to grow faster, metastasize easier and have a poorer prognosis. Tumour grading is important when determining the most suitable drug therapy or any other post-operative medical treatment.

At the time a tumour is detected, the international TNM classification protocol is used to place it along 5 stages - 0, I, II, III, and IV [182]. T refers to the primary tumour's size, N denotes the extent to which it has spread to regional lymph nodes, and M indicates the existence of distant metastases. The lymph nodes are most frequently involved in distant metastases of carcinomas since cancer typically spreads to other sites through the bloodstream or the lymphatic system. The liver and lungs are also frequently involved in secondary metastasis as all portal area drainage flows to the liver, while all caval blood flows to the lungs [176].

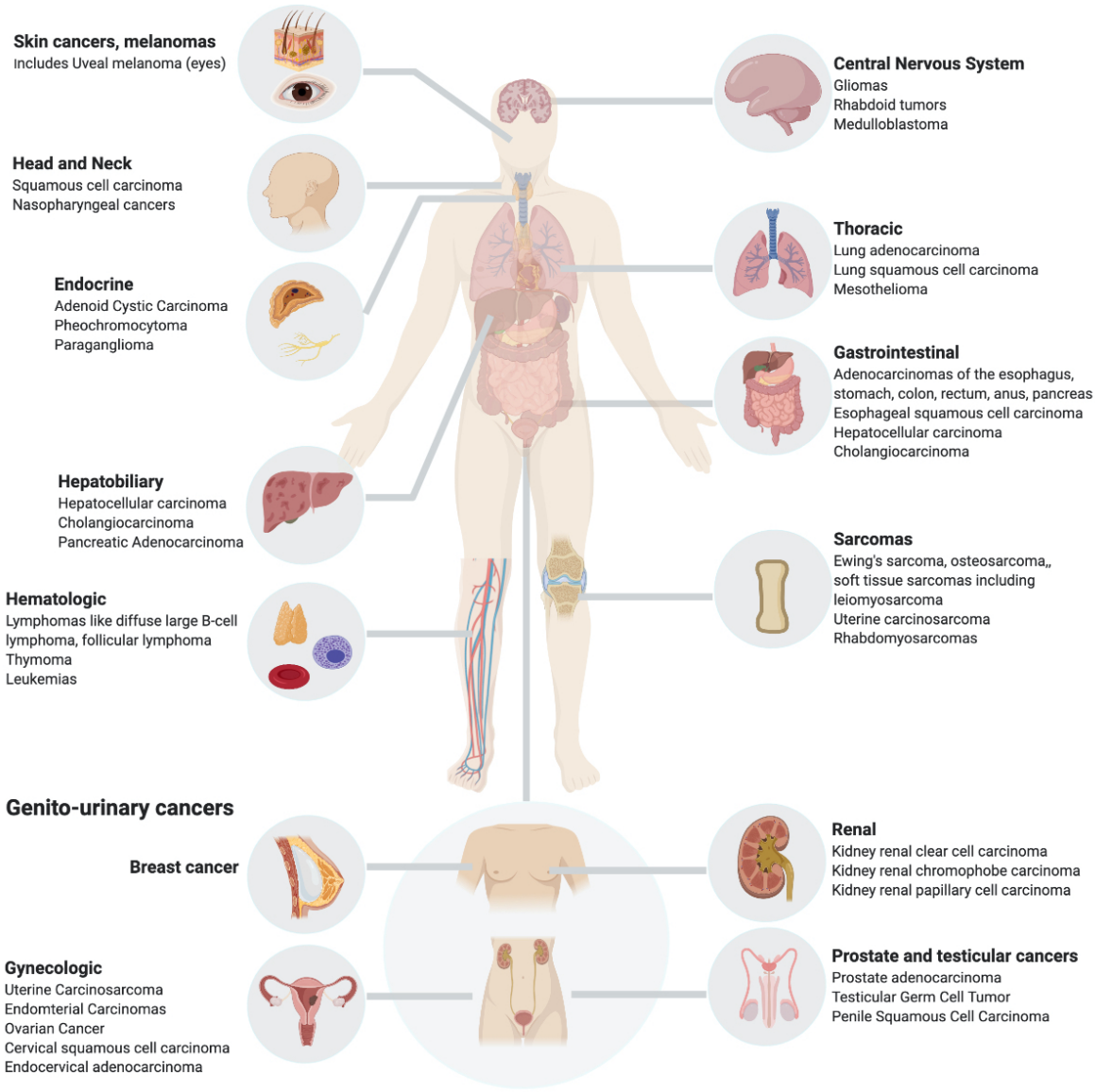

Figure 1.1: Histologic classification of cancers based on their organ-system of origin. Various organ-systems of origin have multiple cancer-types associated with them, differing by the cell-type they originate from.

1.1.3 Pathology in recent decades

Over the centuries, histomorphology has driven the establishment of clinical protocols for cancer classification. However, since the invention of microscopy in the 1600s, and the discovery of H&E stains in the late 1800s, the field itself has remained broadly unchanged [4]. Unsurprisingly, these two tools of the trade can prove insufficient for providing an accurate cancer diagnosis in today's day and age [4]. There are well-established diagnostic guidelines when the cancer presents as a well-differentiated mass of cells, with minimal invasion of blood vessels into the system (low vasculature), and high tumour content in the biopsied tissue. In the event any of these conditions are not met, cancer diagnosis using traditional histomorphology becomes a challenging task. Advances in genomics and computer technology have contributed to improvements in routine histopathology in recent decades. Three innovations in particular, namely *in-situ* hybridization, immunohistochemistry, and digital pathology, have enhanced the amount of information that can be extracted from a tissue biopsy. The following sub-sections describe these methods and summarize the benefits and drawbacks of each.

In-situ hybridization (ISH)

ISH uses a complementary DNA or RNA probe to identify whether a certain DNA or RNA sequence exists in a biological specimen. The technique, developed in the late 1960s, allows pathologists to detect changes in the DNA directly on cytology specimens or FFPE slides [53]. By the early 1980s, fluorescent tags on RNA probes could be used to identify complementary DNA sequences in prepared tissue and visualized using fluorescence microscopy. Since then, advances in microscopy, digital imaging, and genomics have led to significant improvements in the resolution, sensitivity, specificity, and accessibility of fluorescence *in-situ* hybridization (FISH) [32]. Common FISH-detected alterations include chromosomal deletions, gains, translocations, amplifications, and polysomy [32, 53].

Particularly, gene fusions arise in carcinomas and sarcomas due to genomic rearrangements. When these fusions happen between fusion partners that are otherwise distantly located on the genome, FISH can easily and reliably identify these events. However, fusions or translocations involving small chromosomal distances can be difficult to resolve at the resolution provided by a light microscope. For example, chronic myelogenous leukemia (CML) has a known marker, the Philadelphia translocation, which joins the 5'

portion of the BCR gene (chromosome 22) to the 3' portion of the ABL gene (chromosome 9). However, up to 6% of children and 20% of adults with acute lymphoblastic leukemia (ALL) also have this chimeric gene, but at a slightly different location (breakpoint). In this case, molecular techniques are the only way to reliably distinguish CML from AML (DeVita et al. [51]).

When FISH cannot resolve translocations and fusions easily, sequencing approaches like transcriptome sequencing (RNA-Seq) are favoured (Cheng et al. [32]). FISH remains the gold standard for detection of chromosomal abnormalities in routine diagnostic pathology. A key limitation of FISH is the need for sequence-specific probes, which can miss novel fusions that RNA-Seq can easily detect. As RNA-seq is also able to detect gene mutations simultaneously, provide a precise mapping of fusion break-points and translocations, and discover cryptic or novel fusions at high resolution, it is increasingly being favoured over current molecular test methods like FISH [32].

Immunohistochemistry

Depending on the tissue in which they began, cancers express certain characteristic proteins. While the detection of many clinically relevant biomarkers is best performed with nucleic acids (DNA or RNA), protein expression can be used when antibodies exist for specific proteins or mutated protein domains [51]. The protein expression is visualized through an enzyme-linked antibody that activates a fluorescent reporter, and the stain intensity used as the assessment metric [51]. Monoclonal antibodies can detect single amino-acid changes, such as the BRAF V600E mutation. By sequentially testing a series of known marker proteins, pathologists can narrow down the diagnosis to a set of possible cancer types (differential diagnosis). The differential diagnoses guide the selection of antibodies. This area of diagnostic pathology, called immunohistochemistry (IHC), is an indispensable tool of the trade [214].

Refining a broad differential with immunohistochemistry means performing a series of tiered analyses. Primary cytokeratin markers (specifically, CK7 and CK20) are used to stratify cases into subgroups of differential diagnoses, which can then be refined based on additional exclusionary markers like carcinoembryonic antigen and urothelin [61]. Candidate proteins are tested one at a time (single-plexed). It is no surprise then, that this approach can still fail in cancer types where the diagnostic protein markers are not known, or in cases where the marker stain is inaccurate or assessed with inconsistent rules across different labs. For example, dedifferentiated

melanomas can stain negative for melanoma-specific markers like S100 [51]. HER2 receptor status is an important prognostic factor in breast cancer and guides treatment selection. It has been shown that various HER2 IHC tests in breast cancer can be wrong because of differing criteria for HER2 positivity among pathologists, or because of variability in positivity across different sections of the same tumour [66].

At the analytical level, IHC protocols have to be optimized every time a new marker is included. The staining results can also vary depending on the preparation method and testing site. Immunostaining tests are generally performed at specialized independent laboratories using expensive special stains, resulting in high costs for each tested slide [4]. The amount of available tissue also places a physical limitation on immunohistochemistry experiments. The number of slides processed for diagnostic biopsies can range from a few slides to more than 20, with an average total cost of $2000 USD (varying by cancer type) [137]. IHC studies using 10-12 stains for CUP diagnosis have not been shown to increase diagnostic accuracy [81], and a meta-analysis investigating the use of IHC to diagnose metastatic samples showed that the approach led to an accurate diagnosis only 64-67% of the time [216]. Other factors impacting the reliability of IHC testing include tissue antigenicity, inter- and intra- observer variability of interpretation, and tissue heterogeneity [51]. As a result, IHC staining is typically not used as a stand-alone diagnostic tool.

Digital pathology

Advancements in commercially available digital cameras and scanners in the late 1900s led to an innovative replacement for glass histology slides. Instead of requiring extensive storage and preservation of prepared histology slides, electronic scanning could now be used to generate whole slide images (WSI). WSI systems have been tested in diagnostic, teaching, and research domains. These studies have found that scanned H&E slides can be used to render primary diagnoses, aid manual review of frozen sections, or share slides across remote locations for consultation and review [16]. In recent years, machine learning methods have been developed to use these slides for cancer classification as well. However, the prohibitive cost of scanners, additional time required for generating and reviewing clean scans, digital storage costs, and the inability to resolve tissue artifacts like folds and bubbles, all mean that this technology is still in nascent stages for clinical-scale application [16, 222].

In challenging cases, -ISH and IHC can still result in a wide differential

with many plausible cancer types. Pathologists are routinely confronted by tumours that cannot be accurately classified despite extensive work-up and expert reviews of the specimen [4]. Furthermore, all of these methods rely on histomorphology, keeping diagnostic pathology in the realm of visual inspection of fixed tissue. Depending on the specimen type and the expertise of the pathologist, a high degree of discordance exists among the diagnoses - a 2015 study revealed that breast pathologists can rack up a diagnosis discordance rate of 25% amongst themselves (N = 60 biopsies, 115 pathologists) [54]. The subjective nature of the exercise has led to an increasingly complex system of classifications based on histomorphologic differences and an increased use of sequential, single-plexed diagnostic tests that impact the patient both in terms of time and monetary costs [4]. We will now describe the biological factors impacting diagnostic pathology.

1.1.4 Diagnostic challenges in pathology

Cancers can lose the expression of predictive IHC markers, prohibiting accurate diagnosis. Other ancillary histopathology tests can be confounded by several biological issues. Table 1.1 summarizes biological issues that can impact routine diagnosis. Specific challenges of cell-type mimicking, lack of established pathology criteria, and complex phenotypes are outlined briefly as follows:

- Soft tissue sarcomas arise from connective tissue like fat, muscle, nerves, and blood vessels. They often mimic other cancers like melanomas or carcinomas, particularly when they have epithelioid origins [147, 150].

- Cholangiocarcinomas arise from the bile duct epithelia. Distinguishing this pancreato-biliary cancer from pancreatic cancers and liver malignancies remains a challenge due to the absence of pathology criteria. Late detection and diverse phenotypes within the disease have led to a paucity of effective markers to distinguish cholangiocarcinomas from secondary adenocarcinomas and hepatic cancers [12].

- Several carcinomas and lymphomas can present as a mixture of different tissue morphologies, or as an intermediate phenotype between two distinct subtypes of the same cancer [59, 226].

The eventual diagnosis in these cases is usually based on the exclusion of

other possible cancer types. 3-5% of metastatic tumours identified annually still cannot be definitively diagnosed with traditional histopathology [120, 176]. These cancers with unknown primary (CUPs) are usually adenocarcinomas (90%), with one-third of these adenocarcinomas being poorly differentiated. The remaining 10% of CUPs present as poorly differentiated neoplasms - squamous cell carcinomas and undifferentiated carcinomas, and rarely, as neuroendocrine cancers or mixed tumours. Cholangiocarcinomas can also go undiagnosed if IHC findings are nonspecific.

Currently, the management of CUPs is based on serial exclusionary diagnosis with IHC protocols. These diagnostics are sequential, require a vast array of immunohistochemical stains, and the stain reactivity status is based on subjective analysis. Since tested tissue is exhaustive, this places an additional constraint on the diagnosis. This is a perplexing situation, given that CUPs are the 4th common cause of cancer-related deaths globally, with a 5-year survival rate of 11% [58, 120]. Standard IHC workups are unsuccessful in ~75% of the cases, meaning a putative diagnosis is only possible for 20-30% of these cases [171]. In the absence of a definitive primary (80% of cases), the standard treatment is broad-spectrum chemotherapy [77]. This is typically a platinum-based regimen like paclitaxel-carboplatin-gemcitabine. In about 20% of the total cases, patients can be matched with a 'treatable subset', whereby the clinical features are used to suggest a specific diagnosis despite the inability to identify a primary site. The former approach of empiric chemotherapy results in significantly poorer outcome than the latter, where the cases are treated based on a putative primary [162]. Nevertheless, the treatment outcomes for most patients in this group remains poor, with a median survival of 9 months [77]. Recent work in the area has attempted to address the development of improved diagnostics instead of evaluating new chemotherapy regimens.

Particularly for adenocarcinomas of unknown origin, there is considerable variability in the selection of IHC markers that are informative for diagnosis. Except for prostate specific antigen (PSA) - used to identify prostate cancer - individual IHC stains prove inadequate in differentiating various adenocarcinoma CUPs. Treatment decisions in such cases are largely based on clinical features that may be indicative of a specific diagnosis [77]. A uniform approach to IHC driven diagnostics for CUPs is lacking, with the expectation being that more refined molecular approaches can help identify biomarkers with diagnostic relevance.

Table 1.1: Diagnostic challenges in cancer histopathology

Cancer Type	Challenges with histopathology presentation
Cancers with known primaries	Limited/improper staining, poor differentiation
Cancers with unknown primaries	5 broad histologic categories (non origin-specific)
Gastrointestinal tumours	Shared histological features and immunophenotypes
Sarcomas	Diagnosis of exclusion

1.1.5 Impact of diagnosis on treatment

Cancer diagnosis plays an important role in selecting treatment options for a patient. Systemic cancer therapy covers a variety of treatment approaches - chemotherapy, hormonal treatment, and surgical removal of the tumour. Surgical resections are the most common course of treatment for primary cancers, if the loss of the tissue mass does not negatively impact the patient's wellbeing [165]. Chemotherapy and hormonal treatments are based on the anatomic site and histopathology of the tumour.

The NCI Molecular Analysis for Therapy Choice (NCI-MATCH) trial, initiated in April 2015, aimed to find genomic evidence that could match effective targeted therapies with patient molecular profiles. While this and several related studies have shown promise for targeting cancers based on individual mutations irrespective of their site of origin, a diagnosis remains essential for treating the vast majority of cancers. Especially in the case of rare cancers (defined as fewer than 6 diagnosed cases per 100,000 people [67]), CUPs, and other malignancies where actionable mutations may not be found, a diagnosis can provide biological context and motivate the realignment of treatment options based on the putative primary. Cell-type of origin can also influence the response of a tumour to therapy - this is demonstrated by the variable response of vemurafenib to melanomas and colorectal cancers carrying the oncogenic BRAF V600E mutation.

Clinical actionability based on genomic profiles

In 2004, it was shown that gefitinib, a drug that inhibits the activity of a gene called EGFR, worked particularly well at treating lung cancer patients if their cancer harboured certain small mutations in this gene [121]. These mutations led to increased expression of EGFR in the cancer cells. Gefitinib binds to the protein product of EGFR, preventing it from performing its role within the cell, consequently killing the cell. The high levels of EGFR caused by these mutations made cancer cells particularly susceptible to

gefitinib. The findings demonstrated the potential for clinical actionability of a cancer drug based on the cancer's molecular profile. Gefitinib, and other tyrosine kinase inhibitor drugs like it, have since been in clinical trials for other cancers. A counterpart, lapatinib, yielded positive results in EGFR-mutated breast cancer patients, and another similar drug, erlotinib, is approved for metastatic pancreatic cancers in combination with another drug, gemcitabine [177]. Contrary findings, although rare, have also emerged. In a recent clinical trial for epithelial ovarian cancer patients, it was found to have limited clinical activity when used as a stand-alone therapy [159].

Context-specific behaviour of vemurafenib

In 2010, another drug, vemurafenib, was discovered through a large-scale drug screen. Vemurafenib selectively killed skin cancer (melanoma) cells when these cells harbored a specific mutation in the kinase inhibitor gene, BRAF. This V600E mutation, which caused the 600th amino acid product of the gene to change from valine to glutamic acid, was quite common in melanomas and well known since the early 2000s [104]. Clinical trials showed the drug to be effective in 80-90% of melanomas harboring BRAF ^{V600} mutations [104]. Due to the efficacy of the drug, vemurafenib advanced rapidly through clinical trials, receiving FDA approval for BRAF mutated melanomas in 2011.

Several other cancers, including hairy cell leukemias, colorectal cancers, gastric cancers, and papillary thyroid carcinomas were found to contain similar activating BRAF gene mutations [94]. Unlike gefitinib though, vemurafenib failed to be as effective in other BRAF ^{V600} mutated cancers [94]. For example, in recently finished clinical trials, the drug yielded a response rate of <5% in metastatic BRAF ^{V600} mutated colorectal cancers [107] and 25-50% in iodine-resistant thyroid cancers [18]. It appears that at least in the case of BRAF ^{V600} mutated cancers, the histologic context determine response to this drug [94] for unknown reasons.

1.2 Genomics and cancer diagnosis

The Human Genome Project finished in 2003, and shortly thereafter in 2005, The Cancer Genome Atlas (TCGA) project was launched by the National Institute of Health in America to catalogue changes that occur at the genomic level in cancers [215]. The consortium sequenced the genomes of over 10,000 cancer patients, generating high-resolution molecular profiles of

their cancer and the healthy counterparts in their body. At the same time, short-read sequencing technologies started becoming increasingly routine. Today, it is estimated that more than 1.5 million individual human genomes have been sequenced by 2018 - a significant increase from just the one genome in 2003 [180].

Large-scale projects like TCGA and ICGC [43] have profiled thousands of primary, untreated cancers and identified changes that occur at the DNA and RNA level in over 40 different cancer types. Extensive research arising from these projects has shed light on molecular changes that characterize various cancers, which in turn has advanced our understanding of what causes these diseases. As shown by various clinical trials evaluating targets identified through these analyses, this approach can also uncover diagnostic markers and cancer subtypes with different molecular profiles.

The integration of genetic testing into clinical diagnostic pathology has benefitted cancer management. Companion diagnostic tests in the United States are usually co-developed with a particular drug, and approved by the Food and Drug Administration for prescribing the safe use of the drug [31]. For example, the MSK-IMPACT clinical assay and the Foundation One CDx diagnostic both profile clinically actionable genomic variants in a pre-defined set of biologically relevant genes to enable informative selection of FDA-approved targeted therapies [30, 158]. Prognostic multi-gene expression based panels also exist for specific common cancer types, such as MammaPrint (now used routinely to predict metastasis risk for young breast-cancer patients [181]) and Oncotype-Dx (used to evaluate risk for aggressive breast cancer and for treatment stratification in breast cancer patients [24]).

In general, these tests rely on biomarkers - known molecular events that are robustly associated with a specific cancer or predictive of favourable response to certain drugs. These molecular events can be at the gene expression level (for example, HER2 overexpression testing is required before patients can be treated with trastuzumab) or at the genomic level (BCR-ABL fusion testing for CML patients on Tasigna). Some of these changes can also be hereditary, i.e. present in the patient's healthy tissues at the time of conception. Mutations in the CDH1 gene, for example, are associated with hereditary diffuse gastric cancer. These familial syndromes are rare, and germline testing is generally not required unless a patient's family history is suggestive of their existence [53]. More commonly, the germline effects can be diffused, or absent entirely, in which case the

genetic changes within the tumour cells are more informative instead. Since the aim of these tests is to provide treatment planning support in the clinic, for most companion diagnostics prior information about the cancer diagnosis is either provided by the healthcare provider (for example in MammaPrint, Slodkowska and Ross [181]) or is not necessary for medical oncology decision-making and treatment planning (for example, using pembrolizumab for solid tumours with high microsatellite instability, Marcus et al. [123]).

Efforts in improving the efficiency of cancer diagnosis *itself* also aim to leverage molecular profiles generated from sequencing data to classify cancers. Different types of genomics data can be used to develop computational algorithms for cancer-type classification. However, 3 main data-types have been used consistently since the first classifier papers were published in 2001 [163, 187] - DNA methylation, gene expression, and genomic changes including somatic mutations and copy number changes. DNA methylation measures the epigenetic landscape - changes in methylation (addition of a methyl group) in the genome, which in turn can regulate the activity of a DNA segment without modifying the actual sequence. Alternation of DNA methylation patterns is a known hallmark of cancer, with the ability to distinguish cancer cells from normal tissue. Gene expression is a broad term encompassing two main platforms for measuring expressed genes, namely microarrays and RNA-sequencing. Both of these platforms can measure messenger RNA (mRNA) that is translated into proteins, or smaller RNA species such as microRNAs (miRNA). Genomic changes can be single-base mutations in the tumour, or large-scale gains and deletions in segments of the genome, known as copy number changes. More complex rearrangements of genomic fragments, called structural variants, are known to drive certain types of cancers. Specific structural variants typify a subset of known cancers, and can be tested easily through -ISH approaches. The next sub-section discusses the development of 'omics based classifiers, and the differences in the classification performance as noted in literature.

In practise, molecular diagnostic tests can be classified as either commercial panel-based diagnostics or laboratory developed high-throughput sequencing diagnostics that aid research efforts. Typically, the scope of commercial diagnostics is restricted to a pre-selected set of genomic changes with rigorously evaluated clinical actionability, whereas laboratory-developed tests are developed and used by specific institutions as part of pharmacogenomics-focused clinical trials and research efforts, not linked

to a particular drug [31]. Regardless of the approach though, each test can be characterized by sensitivity, specificity, and limit of detection. In the next two sub-sections we discuss the research and development behind current-day 'omics based classifiers and contextualize them against commercial assays available for cancer diagnosis.

Genomic data used for cancer classification

Existing cancer diagnostics test changes in multiple genes simultaneously [53]. The main types of molecular data commonly used for this purpose are exome sequencing, whole-genome sequencing, RNA sequencing, and bisulfite sequencing (BS-Seq). The first two interrogate the DNA in a sample, RNA sequencing can be used to capture transcribed RNA (mRNA) or small RNA species like microRNA (miRNA, long non-coding RNA), and BS-Seq is used to identify methylated regions of the DNA.

Early efforts for algorithm-based cancer classifiers focused on using gene expression measurements from tissue microarrays to find a representative feature set, and then classify a small set of tumours. These methods tried to distinguish 11-14 different types of tumours using supervised machine learning algorithms, achieving 78% (N = 218) [163] to 90% (N = 175) classification accuracy [187]. These early methods used support vector machines (SVMs) with optional recursive feature elimination. The classification groups themselves were quite high-level. For example, stomach and esophageal cancers would usually be combined into a single gastroesophageal group, and lung adenocarcinomas and squamous cell carcinomas combined as lung carcinomas.

Efforts shortly thereafter used slightly different algorithms to classify round blue cell tumours (Burkitt lymphoma, Ewing sarcoma, neuroblastoma, and rhabdomyosarcoma) and leukaemias [103, 196]. The two main algorithms used at the time were neural networks [103] and nearest-centroids [196]. The emphasis was on finding the smallest number of predictive genes that can facilitate the development of new diagnostic tools easily [196]. In 2004, it was demonstrated that neural networks can be applied to two completely different microarray platforms and build tumour classifiers with 83-85% accuracy across the board (N = 120) [13]. A key insight was that feature scaling helps build a robust classifier.

The flurry of papers around cancer-type classification in the early 2000's was followed by concerns about the reproducibility of gene expression data obtained from microarrays. Several studies at the time showed that different

commercial microarray platforms yielded different intensity measurements [50]. A 2007 study using a refined microarray approach observed that the probe sequence can impact the relationship between observed gene intensity and actual gene expression [229]. Background levels of hybridization of a probe can also limit the accuracy of expression measurements from microarrays, particularly in case of low-abundance transcripts [229].

Most of these tests, while touting high overall accuracy, also failed to improve classification accuracy where it mattered most - in cancers that were refractory to histopathology-based diagnosis. For example, in a 2013 study, which later evolved into the commercially available Cancer Genetics Incorporated Tissue of Origin Test, researchers used a gene expression panel to distinguish 15 different types of cancers [81, 157]. In their set of 160 test cases, they found that the method failed in metastases of gastric cancers (<30% accuracy) and had equivocal performance as compared to routine IHC for non-adenocarcinomas. Others yet would aggregate these challenging categories into their high-level grouping, limiting clinical benefit from the application of these methods.

RNA sequencing, particularly sequencing of transcribed RNAs through RNA-Seq, has been steadily replacing microarray based methods for interrogation of gene expression profiles. The methodology extends our ability to quantify splice variants, non-coding RNAs, isoform specific gene expression, and gain insight into variation that is not captured at the genomic level. Systematic analyses of the utility of RNA-Seq based diagnostic assays and classifiers has shown that RNA-Seq based classifiers outperform arrays in characterizing cancer transcriptomes [23]. Furthermore, retrospective meta-analyses on these projects have revealed that the large amount of biological diversity in CUPs is not fully captured within panel-based assays, emphasizing the need to build up from a global gene expression set in order to identify a comprehensive set of site of origin markers that are universally expressed [118, 162].

Besides gene expression, other data modalities like DNA methylation, somatic variations, and microRNA expression can be used to classify cancers. A microRNA-based classifier that used feature selection embedded in the Least Absolute Shrinkage and Selection Operator (LASSO) classification algorithm obtained 88% accuracy on a cohort of 48 metastatic samples, but failed to classify metastases to the liver, and to distinguish between stomach and esophageal cancers [184]. MiRview mets, an RT-PCR based miRNA assay classifying 25 different tumour types using decision trees and

K-nearest neighbours obtained 86% accuracy on the held-out set (N = 83) [127]. An independent test cohort of 80 samples only showed performance gains for distinguishing biopsies from the liver or otherwise, and biopsies of gastrointestinal tumours or otherwise [127], falling appallingly short of generalization. This test was later expanded into a commercially available kit (Cancer Origin Test by Rosetta Genomics), subsequently withdrawn from market because of bankruptcy. Another recent study leveraged gene mutations and copy number alternations to distinguish 28 different cancer types, yielding an overall accuracy of 78% [183]. Another group has shown 91% accuracy to classify 17 different tumours using miRNA data, and 95% using DNA methylation data [191].

Methods have been developed that use microRNA expression to achieve 70-90% classification accuracy for identifying a putative primary in CUPs [127, 170]. Several small-scale studies have also shown the utility of microarray panels to identify the site of origin of certain types of CUPs [77, 161, 201]. Demonstrating the potential for pan-cancer based diagnostic approaches for CUPs, Tothill et al undertook a comprehensive interrogation of 13 cases of cancers of unknown primary against a background of 229 primary and metastatic cancers spanning 14 tumour types [201]. Leveraging a support vector machine (SVM) based training model, the authors were able to demonstrate that an expression-based test could strongly characterize 11 of the 13 CUPs, validated by pathologic evaluation and clinical outcome information. The authors also found that if they left out different primary tumour types during training, and used them for testing only, the left-out class was predicted most often as a trained cancer type that was the most similar to it biologically. However, all these analyses leveraged narrow representations of tissue types, and have been unable to demonstrate generalizability and clinical utility of the diagnostic method to external validation cohorts of CUPs.

Commercial assays for cancer diagnosis

Various -omics data have been leveraged to identify the tissue of origin of metastatic cancers (including copy number variants, nonsynonymous point mutations, and single nucleotide substitution frequencies), but the performance of methods that rely on gene expression information has been shown to outperform methods based on other data types [124]. The two commercially available pan-cancer diagnostic assays rely on expression profiling using microarrays or reverse transcriptase - polymerase chain reaction (RT-PCR). These two assays provide a diagnosis for cancer

types and subtypes at varying degrees of resolution - CancerTYPE ID (bioTheranostics, Inc.) and the Tissue of Origin Test (Cancer Genetics Incorporated).

The Tissue of Origin Test leverages the gene expression profile (microarray) of 2,000 genes, covering 15 tumour types [157]. While the details of the algorithms used for gene selection and classification are unclear, the commercial assay achieved a performance of 89% on the test set (N = 462, 15 cancer types). The tumour types, in this case, indicate the anatomical site of origin instead of nuanced histologic categories, for example kidney, gastric, pancreas, sarcoma, and non-small cell lung cancer. Notably, the test performed better on metastatic samples than primary tumours, but had a notable drop in performance for lung metastases (N = 3/5).

The CancerTYPE ID assay uses RT-PCR to assay the expression of 92 genes from FFPE samples, profiling 30 main tumour types indicating the anatomical site of origin (for example, brain, breast, cervix, and skin) and 54 subtypes indicating the histologic classification (for example, pancreatic carcinoma, melanoma, renal clear cell carcinoma) [55]. In leave-one-out cross-validation, the method demonstrated 87% accuracy for the main tumour types and 85% for the histological subtypes (N = 2,206). The generalizable performance of this approach as 83% (N = 187, 28/30 main cancer types). The genes themselves were selected for optimal multi-tumour classification from a set of 578 tumour samples using a genetic algorithm, and include 5 genes for normalization in the RT-PCR experiment. While the method was able to obtain robust performance, both training and test samples were optimized to reach 80% tumour content prior to RT-PCR. The RT-PCR method requires a high RNA quantity and quality as well, which can preclude several cancer samples from analysis (95-98% success rate for meeting criteria, as reported in the paper). In routine practice, an entire FFPE block or biopsy core is required. Based on retrospective abstractions of pathology reports from the test cohort, the diagnostic utility of the assay was found to extend beyond aiding diagnosis of CUPs, to those samples where multiple differential diagnoses existed but a single definitive diagnosis was absent [55]. This highlights another important implication of pan-cancer classifiers besides providing a diagnosis for CUPs - resolving differential diagnoses that may come up routinely. A recent retrospective study using this method showed successful application to render a diagnosis in 56 patients with neuroendocrine tumours of unknown primary [27].

Two challenges still exist with direct adoption of these tests in precision

oncology. Firstly, precision oncology projects emphasize the study of molecular data, particularly sequencing of the DNA and RNA, to characterize and treat the cancer. Nucleic acid sequencing preferentially requires fresh frozen tissue, whereas the existing commercial assays necessitate FFPE blocks. Collection of additional tumour biopsies places an extra burden on the patient, can be excessively invasive (as is the case for brain cancers), and may not be feasible in many cases (for example, in pancreatic adenocarcinomas).

Secondly, only one of these tests reached a diagnostic resolution beyond the anatomical site of origin, limiting the diagnostic utility [55]. For example, in clinical practice it is not sufficient to indicate that a tumour arises from the kidney - several clinically relevant and biologically distinct cancer types exist within kidney cancers, with different treatment regimens. The test requires a large amount of input RNA for assessment, reverting to the original challenge of limited tissue and the value of building a molecular profile of a cancer. There is a current need to develop classifiers that are easy to scale, integrate within precision oncology projects that emphasize granular study of advanced cancers and can leverage the deluge of information provided by genome sequencing.

Limitations of present-day 'omics based classifiers

Despite years of work in this area, several problems remain with immediate adoption of the resulting methods to all cancer types. The selection of representative features to select for specific cancer types can lead to overfitting or restrict the application of the method to those cancer types only. As these computational methods typically require a lot of training data, as discussed in Section 1.2.1, rarer cancer types are excluded from the classification task, or amalgamated into a broader anatomic category with limited clinical actionability. These issues can also stem from the type of genomic data used for the classification, since representative datasets of whole-genome sequence, RNA-Seq, miRNA, and methylation, are not available for all cancer types.

1. *Selection of feature subsets can prohibit generalization to rare and complex aetiologies*
   Clinically predictive gene expression signatures are not always reproducible across multiple studies of the same tumour type either. This can arise from differences in technical platforms, standardization of data, and at a biological level, a complex phenotype of survival [155]. A 2010 review of gene expression based diagnostic panels for diagnosis

of CUPs noted that many available panels have decreased accuracy for poorly differentiated tumours, or for specific tumour types like lung and pancreatic adenocarcinomas [132]. This is particularly important for various heterogeneous cancer types reflected in CUPs, which may not always have the same stable markers that are enriched for during feature selection optimized to differentiate known primary tumours [201].

2. *Existing classification approaches lack granularity and exclude rare cancer types*
   Cancers arising from the hepatobiliary and pancreatic systems are particularly challenging to diagnose using both histopathology and genomics- based approaches. Some work in this domain sidesteps these diagnostic challenges when developing classifiers, excluding these cancer types from the classification task [201]. In other work, these clinically distinct cancer types are merged together. This removes the challenge of granularity in their prediction task, but negatively impacts the clinical utility of such methods [81, 221]. At a broader level in these studies, while the reported accuracy ranges from 87-93%, diagnostically challenging cohorts are simply combined, and/or samples refractory to the method are excluded from the reported calculations [221].

3. *Limited data availability across all sequencing platforms prevents efficient training of rare cancer types*
   A limitation in scaling the use of miRNA and BS-seq based classification to other cancer types has been the unavailability of these type of molecular data for rarer cancers. The challenge remains, with genome sequencing and RNA-Seq being the most commonly available molecular profiles of the maximum different types of cancers. Additionally, previous results using various data modalities show that cancers refractory to diagnosis by gene-expression based algorithms show similar trends using other data-types as well.

4. *Interpretation of diagnostic decisions from classification platforms is limited*
   Cancer classifiers based on gene-panels and molecular assays can provide a classification decision, but understanding the biological changes contributing to the decision is difficult. Methods for feature prioritization in the underlying algorithms have been proposed, but it is unclear how to interpret these at a single-sample level. Since the

algorithms themselves are based on marker selection, the breadth of biological changes that they can interrogate and prioritize remains limited.

In the next section we describe the computational approaches underpinning existing classification tools, and how these algorithms can be leveraged to address some of the challenges highlighted above.

1.2.1 Computational algorithms for cancer classification

Five main algorithms dominate the field of 'omics-based cancer type classifiers - support vector machines (SVMs) [163], random forests (RFs), k-nearest neighbours (kNNs), naive Bayes (NB), and neural networks (NNs) [103]. They broadly differ in the way they compare features to establish classification thresholds [84].

Linear classifiers distinguish two or more classes using a linear combination of features. An example algorithm is Naive Bayes. In a two-dimensional setting (when using two features for classification), a linear classifier will be a line. In higher dimensions, this line becomes a decision hyperplane. Linear classifiers attempt to learn the parameters describing slope and threshold for this hyperplane, leveraging examples from the training set [84]. For example, the θ weights for discriminating between $class_1$ and $class_2$ will be learnt through multiple training examples in the following equation,

$$b = \theta_0 + x_1\theta_1 + x_2\theta_1 + \cdots$$

such that, for a new sample $\mathbf{x}$,

$$b < \theta_0 + x_1\theta_1 + \cdots \implies \mathbf{x} \in class_1, \text{ else } \mathbf{x} \in class_2$$

Linear classifiers work quite well when a linear combination of the distinguishing features can separate the classes of interest (i.e. the classes are linearly separable). If linear separation holds, then we can find an infinite number of linear separators. Selecting the most suitable decision hyperplane, that is, one that generalizes the best to new data, is an inherent challenge when training linear classifiers. Additionally, when the training data is noisy, the decision hyperplane can easily overfit to the training samples and generalize poorly to new data.

If a classification problem is nonlinear, this means the class boundaries cannot be approximated well using linear hyperplanes [84]. In this case,

nonlinear classifiers are a suitable alternative. Some non-linear models like kNNs can also subsume linear models in special cases. kNNs have a decision boundary that can take more complex shapes than a hyperplane [200]. During classification, the test sample is assigned the majority class of k nearest neighbours. Neural networks are a supervised machine learning approach that have been shown to be quite powerful at non-linear classification tasks. These algorithms require labelled data and are trained iteratively using a training dataset to distinguish the classes of interest.

Nevertheless, there is no known universally optimal classification strategy. The optimal learning algorithm can be selected *a priori* if it is known that classifications are linearly separable or not. Further selection can be guided by parametrization, feature subsetting, and by assessing the trained models' performance on a set of held-out samples. However, the generalizability of a classification algorithm can be truly assessed only using external datasets and new samples drawn from the real world.

Evaluation of performance

The performance of a classification algorithm can be evaluated on the training data itself, on part of the training data that the algorithm has never 'seen', or on an external dataset reflecting the type of data the algorithm will encounter in practice. These three types of data can reveal different aspects of the algorithm. Performance on the training data can help the user determine optimal model parameters, especially if this is wrapped within a resampling approach like cross-validation. k-fold cross-validation can be used to train a model on k-1 folds of the data and test on the remaining fold, repeating this procedure k times and leaving a different fold out for validation each time. The held-out dataset is generated by keeping a part of the training data away from the entire training process (including resampling strategies). This dataset reflects the type of data the algorithm was trained on, and its performance indicates whether the trained model has a high variance to maximize prediction of the training samples (overfitting), or if it has learnt representative features that can generalize to new data from the same underlying data distribution. The optimal test set should ideally be an external dataset generated independent of the training data and is the true measure of the generalizability of the trained model on new data.

When comparing the classification results across multiple classes, various metrics can be used to evaluate the performance of a classifier on held-out or independent test sets [110]. We can measure the number of

correctly classified samples (accuracy). However, aggregate accuracy can be confounded by an over-representation of certain classes compared to other. Accuracy calculated after adjusting for smaller cases is called balanced accuracy. However, balanced accuracy measurements give equal weight to the presence of false negatives (FN, or Type 2 error - the number of samples from category 'A' incorrectly classified as another category) and false positives (FP, or Type 1 error - number of samples not from category 'A' incorrectly classified as category 'A'). This can be misleading when the class imbalance is severe and evaluation needs to take into account the higher number of 'negative samples' for each class.

In these cases, the F1-score is considered a suitable alternative since it takes both FP and FN into account [110]. For a given class, precision ($TP/(TP+FP)$) and recall ($TP/(TP+FN)$) measure the positive predictive value and sensitivity of a classifier respectively. Here, TP indicates true positives - the number of samples from a particular category 'A' correctly classified as such, with FP and FN as previously defined. These metrics can be combined as $2*TP/(2TP+FP+FN)$ to get the F1-Score for a class, with a high F1-Score being desirable. The mean F1-score across all classes gives a macro-F1 score metric for a multi-class classifier with class imbalance.

A common issue with cancer classification methods is the lack of demonstration of classification accuracy on samples that are not simply held-out from the training data or generated by the laboratory protocols identical to the training data. Systematic biases from data extraction and processing can easily lead to over-fitting of the trained model on the training data, wherein the classifier ends up learning the technical noise and artefacts inherent to the training dataset instead of identifying features that remain relevant in independent datasets. Cross-validation within the training dataset can demonstrate the ability of the classifier to learn about the training dataset itself, but the generalizability of these classification methods can only be assessed by their performance on independent datasets generated from different protocols and laboratory environments. Many of the reported studies for cancer classification measure the performance of their trained algorithms on the held-out samples, incorrectly reporting the resultant performance as a metric of generalizability [163, 183, 187, 191, 196]. Even more misleadingly, what many of these methods label as 'independent test sets' are in fact held-out samples from the training dataset, following the same tissue preparation and sequencing protocols [13, 187].

The use of held-out sets to represent cancer classification performance is

particularly questionable since an imperative first step in all these methods is typically to select a small subset of representative genes prior to model fitting. Feature selection can bias the performance of supervised machine learning methods. The selection is usually statistically driven, based on Pearson correlation to uncover associations between features and labels [191], Wilcoxon rank score [187], or recursive feature elimination [163]. If the features are selected to increase the separation of training data, as they typically are, then performance on held-out samples will be expected to be better by default, as compared to performance on a test set where the representative features may be different depending on the data preparation and processing protocols. We now discuss the various feature selection methods and their resultant impact on classification.

1.2.1.1 Feature reduction

Feature selection is used in classification algorithms to reduce dimensionality, discard noisy features, reduce computational costs, or to incorporate prior information about the system to increase the likelihood of generalization. Analytical methods for feature selection can be filtering based, wrapper based, or embedded within the classification algorithm. Ensemble feature selection with bootstrapping samples can be used on top of any of these strategies. The feature selection method of choice is run on several random subsamples of the training data, and different lists of variables are selected. Eventually, these lists are merged into a subset that is most representative of the various classes of interest. Alternatively, feature selection can be performed based on prior knowledge about biological programs characterizing the relevant signals. We discuss the prominent types of algorithms for both approaches here.

Analytical approaches for feature reduction

Filtering based approaches include analysis of variance (ANOVA), or statistical thresholding from pairwise t-tests between categories. These methods can lead to false positives if not careful and do not provide an intuitive cross-validation approach to optimize for a set of discriminatory features. However, they model feature dependencies independent of the type of classifier used, reducing the risk of overfitting when identifying an optimal subset of features.

Some classifiers can incorporate feature selection as part of the classifier

construction. These embedded methods include lasso regression and elastic nets. Lasso regression controls the sparsity of a solution by encouraging the selection of individual discriminatory features for each category. The method tries to avoid redundant genes in a given signature and is not expected to be stable in data where typically many genes encode for functionally related proteins. Ridge regression, on the other hand, distributes importance over multiple input features when optimizing classification. Elastic net tries to combine both these strategies to select groups of correlated genes that are useful for classification. However, an investigative study assessing the impact of feature selection methods on stability and accuracy of molecular signatures found that the performance was equivalent from any of these strategies, with very unstable outputs in all cases [85].

Wrapper methods jointly select sets of variables with good predictive power for a classifier. They perform a greedy search in the space of sets of features. Extensive cross-validation is required to estimate the accuracy of the selected feature subsets. An example of this is recursive feature elimination, a strategy used in SVMs and random forests. This method is not necessarily stable, but can be combined with ensemble feature selection to improve the stability of features. These wrapper methods are less prone to local optima, but are computationally intense with a high risk of overfitting.

Most feature selection methods make an assumption about the kind of feature coefficients that will be encountered. For example, in ridge or lasso regression, it is expected that most feature coefficients will be zero or near zero, enabling the selection of important discriminatory features. In practice though, it has been found that features selected through the Student's t-test seem to provide the most robust and stable selection of features [85]. When extended to a multi-class setting, the ANOVA implementation is used.

Biological approaches for feature reduction

Oftentimes, well known clinical and genomic features may also be utilized in the development of a cancer diagnostic. These selections are guided by observed molecular associations within cancer types of interest, which may arise from known protein markers or previous studies that aimed to characterize the genomic underpinnings of various cancers [89, 160, 192]. These features may include genomic or transcriptomic events like HER2 expression evaluation for breast cancer stratification [66], APC gene mutation status for colorectal cancer diagnosis [62], 1p19q locus mutations and IDH1 mutations for brain cancer classifications [15]. The Catalogue of Somatic Mutations in Cancer (COSMIC) database lists the various genes

mutated frequently in different types of cancers [63], and can be utilized for cancer classification and prognosis [195]. Genome-wide mutations can also be consolidated into distinct mutation signatures, many of which are associated with cancers or exposure to carcinogens [92]. The biological approaches for feature reduction for not necessarily separate from the automated approaches. Features can be selected through a combination of the two to ensure a balanced representation of prior information and automatically discovered rules in the training data [2].

Feature selection, if done correctly, can reduce computational costs and avoid overfitting to noise in the training data. However, it requires a large, representative dataset that encompasses the various heterogeneous 'genotypes' (in the form of mutations, expression, or other input data type) for each of the cancer types. Typically, sequencing data of untreated primary cancers from TCGA and ICGC is used for this purpose. These datasets have an over-representation of common cancers that have been studied extensively over the decades and have well-established molecular subtypes. Cancer types that are refractory to routine diagnosis fall outside of these well-defined criteria. This includes poorly differentiated cancers, rare cancer types, and cases presenting with mixed cancer phenotypes (like sarcomatoid mesotheliomas). In such scenarios it is possible that samples do not display the same representative markers as the well-differentiated primary counterparts.

There has also been a dearth of literature on using an approach devoid of feature selection for cancer classification. A 2011 study has shown that it is possible to retain all genes for a pan-cancer microarray expression driven classification of CUPs, and still obtain high performance - in this case, obtaining 89% accuracy on a validation set of adenocarcinoma CUPs [148]. Additional work in this area is required to evaluate if feature selection is indeed a necessity for cancer classification, and whether bypassing this step can provide any additional biological insights for cancers refractory to routine histopathology diagnosis. One useful outcome will be the ability to not just classify various cancers, but to further interrogate the trained model and obtain insights into the decision-making process.

1.2.2 Beyond the diagnosis - identifying biological changes in individual tumours

The impact of individual changes in genes plays out through various cellular pathways. Placing genomic alterations in the context of the oncogenic pathways they impact can help us understand the biology of the tumour, identify potential causal mechanisms, and prioritize therapeutically relevant targets [80]. It is particularly relevant for personalized analysis of cancer genomes and transcriptomes, as biological interactions and patient-specific sources of variability (germline influences, prior treatment etc.) can easily confound our ability to identify relevant features for diagnosis and therapy. Molecular analysis based on sets of statistically or biologically selected genes can help detect known patterns of positive selection across various tumours. On the other hand, this approach can overlook functional changes that are important for a rare subset of tumours, simply due to the limited power of a typical cancer cohort [44].

Analysis of genomic changes through the lens of cellular mechanisms can help distill the multitude of changes that happen in a single tumour into oncogenic pathways. Aggregating the molecular changes into a view of the most dysregulated pathways in an individual tumour typically requires manual prioritization of known tumour suppressors and oncogenes, extensive literature review to align observations against known biological pathways, and integration of genomic changes with expression dysregulation. One prevalent approach for automated pathway analysis is network-based, whereby genes are overlaid with observed genomic changes (evaluated against a set of controls for gene expression changes, or a healthy tissue reference for mutations) to identify 'hubs' of biological activity from known pathways [68, 208] or to recover novel interactions and topologies [167]. Another popular approach is one of statistical enrichment, whereby case-control comparisons are made to identify statistically significant pathways and gene clusters that are differentially expressed in the tumours [228]. Statistical enrichment may be combined with network analysis to prioritize pathways in a set of samples [44]. However, there are very few reference-free approaches available for automated prioritization of important pathways in individual cancers [44, 116, 208].

1.2.2.1 Findings from clinical trials utilizing genomic analysis for cancer management

Clinical trials exploring targeted therapy in advanced cancers

Various clinical trials have been conducted over recent years to evaluate the benefit of treating patients with therapies that are targeted to specific molecular alterations in their tumour. While these programs have found marginal to significant benefit in different scenarios when providing targeted therapy, the fraction of tested populations where actionable mutations were found was typically quite low. The MOSCATO 01 clinical trial in 2017 aimed to evaluate clinical benefit of targeted therapy in a cohort of 843 adult patients using RNA-Seq and whole exome sequencing techniques [125]. In the subset of 199 patients that eventually received targeted therapy, progression free survival (PFS) was 1.3-fold higher as compared to patients on prior therapy. Notably, PFS was lowest in a subset of 36 'ill defined primary tumours' (pathognomonic cancers of unknown primary) in this cohort, regardless of being on matched therapy or otherwise. The 2014 multi-center SHIVA trial screened 741 patients of any tumour type, finding a slight difference in median PFS between the matched treatment and the prior therapy arms (2.3 versus 2.0 months respectively), but a significantly higher average PFS in the matched treatment group of patients (maximum PFS 3.8 months versus 2.1 months) [109].

One key limitation of actionable mutation based clinical trials is the limited subset of patients that can eventually benefit from this approach. In the MOSCATO 01 trial, for example, scientists found actionable mutations in 411 patients but only 199 patients were eventually treated with a targeted therapy (based on a matching genomic alteration). A recent multi-center study across 2,579 patients also found that only 6% of patients were able to receive matched therapy, with a very low overall response rate (0.9%) [203]. Similar findings emerged from the Precision in Pediatric Sequencing (PIPseq) program for children with hematologic or solid cancers at Columbia University Medical Centre [146], where only a small fraction (16%) of successfully screened patients obtaiend matched therapy. These findings suggest that molecular screening is not a viable approach for routine clinical practice at present, but predictive biomarkers should continue being evaluated for efficacy [109]. Interestingly, in the PIPseq program researchers found that genomic data - particularly RNA-Seq data - was useful for prognosis, diagnosis, or pharmacogenomics in an additional

38% of cases, suggesting that detailed molecular profiling beyond mutation panels may still have wider benefits beyond treatment selection.

Clinical trials on CUPs

Ongoing clinical trials on genomics-guided cancer treatment also suggest that further research may be required to transpose existing targeted therapy approaches to cancers of unknown primary (CUPs). In 2017, a large-scale clinical trial at Memorial Sloan Kettering Cancer Center evaluated the utility of molecular/genomic profiling alongwith pathology and clinical information in improving treatment options and outcome for 333 CUP patients in particular [207]. 150 patients (45%) had 34-410 cancer-associated genes sequenced in a panel in an effort to identify clinically actionable mutations. Of the 45 patients (34%) where an actionable mutation was found, only 15 received targeted therapy. Factors limiting the use of targeted therapy included limited availability of suitable drugs, poor performance status, and/or rapid clinical decline. In the small subset that benefited though, overall survival was 13 months, as compared to historical observations of 3 to 8 months in previous studies. These findings, accompanied by other recent basket trials that aim to group and treat patients based on similar molecular profiles, have motivated the need for finding actionable mutations in CUPs.

In a systematic review of clinical trial outcomes from 2002-2009, Pentheroudakis et al [152] found assessed outcome and survival in patients with CUPs, where a putative primary could be identified using molecular platforms like gene expression arrays. They found that patients with CUPs of putative lung or pancreatic origin had equivalent tumour shrinkage and median survival as those with known metastatic lung and pancreatic cancers. However, within the metastatic breast and bowel cancer groups, putatively diagnosed CUPs had a significantly inferior response rate compared to patients with known primaries of breast and bowel, suggesting that while CUPs can be accurately classified by molecular approaches, in some cancer types they may be molecularly distinct from their known primary counterparts.

A recent clinical trial compared site-specific therapy (based on gene expression based diagnostic) with empirical chemotherapy for 113 CUP patients [86]. They found that there was no significant improvement in 1-year survival between the two groups, but median overall survival (16.6 versus 10.6 months) and progression-free survival (5.5 versus 3.9 months) were improved in patients treated based on the putative primary site.

Pancreatic cancers presenting as CUPs have a four-fold incidence of bone metastasis, and 30% higher incidence of lung metastases compared to known primary pancreatic cancers [6]. The higher incidence of metastatic occurrence (compared to putative primary cancer) has be reasoned to be driven by immunosuppression and aggressive metastatic potential of early progenitor cells of CUPs [6]. Various projects focused on molecular characterization of CUPs have found that while they do not typically display activating point mutations in oncogenes or tumour suppressors [60, 120, 207], they are typically characterized by angiogenesis activation (in 50-89% of cases), oncogene overexpression (10-30% of cases), hypoxia-related proteins (25% of cases), epithelial-mesenchymal transition markers (16% of cases) and activation of intracellular signals like AKT or MAPK (20-35% of cases) [120, Rassy et al. [166]]. Emerging evidence on putative molecular hallmarks of CUPs has also indicated the need to further elucidate the role of mechanisms like growth factor independence, immune evasion, chromosomal instability, and telomerase activity in these elusive cancers [166].

1.2.2.2 Role of cancer diagnosis in genomic analysis

When undertaking an individualized analysis approach for a cancer patient, scientists need to contextualize the genomic and transcriptomic findings from the cancer against a background of healthy tissue and tumour samples with similar histology. This is particularly true for gene expression data, where a background set of samples is used to determine if certain genes are over- or under- expressed. A precise cancer diagnosis is, therefore, an important step before a suitable background or comparators can be selected. Classification models incorporate various molecular measurements of tumours to provide a quantifiable prediction. Presumably then, if we can understand the rationale for cancer classification, we can obtain biological insights into the pathways and mechanisms driving a cancer.

Feature selection prior to training an algorithm is one approach to aid interpretation of classifier results. Manual or statistical prioritization of relevant features can indicate which genes or pathways are frequently associated with a particular group of cancers. Alternatively, correlated or orthogonal feature sets guiding the machine-learning based diagnostic can be inferred using post-hoc inference of feature importance from trained models. Given a small set of features, we can measure the impact of each

feature on the output through approaches like recursive feature elimination (for random forests, SVMs), and integrated gradients (for neural networks). However, feature selection may not always be the most robust approach, especially when studying cancer types where the markers may be redundant, confounded by another biological signal, or absent. Statistical approaches may also lead to the selection of technical confounders or covariate genes.

A comprehensive sample-level approach to cancer diagnosis and pathway analysis can obviate the need for choosing an appropriate background (a challenge when studying rare aetiologies), and avoid any bias towards known gene candidates and pathways in the integrative analysis. Such a method can also help compare results from related samples and datasets, identify new subtypes based on common patterns of network alterations [90], and propose cancer mechanisms.

1.2.2.3 Automated tools for single-sample analysis from RNA-Seq

Single-subject analysis of transcriptomes is an underappreciated approach for individual-level analysis of diseases. Current approaches leverage cohort-based population analysis either require large representative sets of each disease for comparison or rely on a case-control approach to find differentially expressed pathways [116]. In these tools, the input (gene expression values) is either used as is, or transformed either via ranking, z-score calculation or through statistical thresholding like log-likelihood. After data transformation, the pathway activity is measured by aggregating the values of all genes in a given pathway arithmetically or through the enrichment of gene-level perturbations. Results can be confounded depending on the type of statistical metric used. For example, in an extensive review done on these methods, the authors found that the enrichment based methods were highly sensitive to the way the pathway was defined [116].

The current approach for single-subject transcriptomic analysis has clear limitations. The requirement of controls and background samples makes it difficult to analyze cancers that present with mixed histology, lack suitable comparator datasets (rare cancers, post-treatment cancers), or have important individual signals that characterize the tumour. The analysis can be further impacted by platform biases, and in the event of small case/control studies, be severely underpowered [210].

1.3 Objectives and chapters overview

Studying the molecular profile of cancers is becoming standard practice for patients with advanced disease, propelling an era of precision medicine. Molecular profiling has expanded from disease-specific tests to broader panels that interrogate multiple genomic changes simultaneously and link to clinical data [42, 53]. These changes can be at the DNA, RNA, or protein level in a patient's tumour. Through numerous research consortia, thousands of high-resolution profiles of cancers have also been generated using whole-genome sequencing, exome sequencing, and RNA sequencing [180]. These modalities capture the genomic and transcriptomic landscape of each individual tumour. When associated with clinical data, this high-resolution sequencing information can provide us with a portrait of various cancer types [43, 215] and identify functionally important genes in common cancers [144, 168]. They can also automate clinical tasks like cancer classification [73, 114] and guide treatment protocols [30]. Analyzing the molecular landscape of rare tumours can indicate a rationale for transposing lines of therapy that align with widely studied and curated cancer types.

The overall objectives of this thesis are to investigate the utility of RNA-Sequencing as a standalone diagnostic modality for common and rare cancers and to develop machine learning methods that utilize all available gene expression information to resolve differential diagnoses, provide a putative diagnosis for CUPs, and guide genomic analysis of rare cancers, while also providing a molecular rationale for the resultant decision. An overview of the main contributions is shown in Figure 1.2. This work will move us closer to a world where routine cancer diagnosis is based on detailed molecular profiles of cancers, and where the diagnosis decision-making can be easily broken down in terms of biological pathways and networks driving the tumour.

The next chapter will begin with a background of an ongoing precision oncology trial at BC Cancer, the personalized oncogenomics project (POG), from which the vast majority of our research data is drawn (Clinicaltrails.gov ID: NCT02155621). This chapter will motivate the need for a cancer diagnosis in genomics-based cancer profiling. Challenges with diagnosis and their relation with cancer treatment will be highlighted through a published case-study where transcriptomics was used as a diagnostic aid to contextualize analysis and revise diagnosis [72].

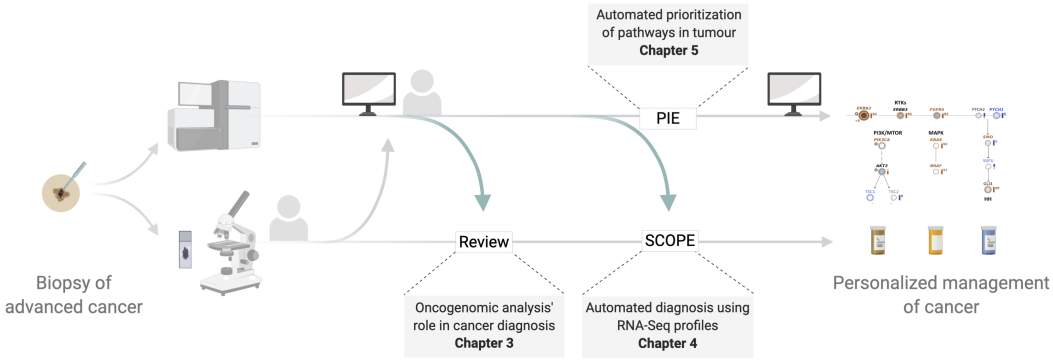

Figure 1.2: Thesis overview and key contributions. In this thesis we explore the utility of bulk RNA-Seq as a diagnostic and analysis aid in personalized oncogenomics initiatives. In a detailed retrospective study we review the frequency of diagnostic changes motivated by genomic data and molecular observations. We develop an automated, open-access tool (SCOPE) for cancer classification using large, representative RNA-Seq profiles. We then extend this method to provide pathway-level profiles of individual cancer samples, also made available as an open-access tool (PIE).

Chapter 3 scales up this investigation to the POG cohort, asking when and how whole-genome sequencing and RNA sequencing can impact cancer diagnosis. Through a retrospective analysis of >300 POG patients, we identify cancer types that are most frequently refractory to pathology-based diagnosis. We further review how sequencing information could be used to detect, review, and resolve incidences of misdiagnosis, differential diagnosis, and indeterminate diagnosis associated with advanced and rare cancers. This includes the use of SCOPE, a published neural network based cancer classification method, the development and validation of which is described in detail in Chapter 4 [73].

A pan-cancer classifier like SCOPE, that uses large transcriptomic profiles for decision-making, provides a method for quantitative, robust orthogonal cancer diagnosis. In the process, does this model learn any biological properties of each cancer? How does this automated learning compare to the manual genomic analyses commonplace in precision oncology? In Chapter 5 we address these questions using PIE, a tool for extracting pathway-level impact scores from SCOPE. We find that PIE can recover known cancer biology for primary cancers from >10,000 samples in TCGA. The resultant pathway profiles can be used to cluster cancers by their

diagnosed cancer type. We show that it can perform single-sample analysis - it identifies therapeutically-relevant pathways for the case described in Chapter 2, and in another case-study, characterizes the biology of cancers of unknown origin.

Finally, Chapter 6 concludes the thesis by discussing the strengths and limitations of the research presented in Chapters 3-5. It outlines outstanding challenges and interesting directions for future research in this area, including the utility of the methods developed herein.

Chapter 2

Background

Various genomic indications for cancer drug selection and treatment stratification have been translated to the clinic in the past decade. These include the evaluation of hormonal markers in breast cancer for selecting suitable drugs [66], identification of treatment options for colorectal cancers based on microsatellite instability status [145], and treating cancer patients with high tumour mutation burden using pembrolizumab [123]. Many of these indications rely on an accurate cancer diagnosis prior to administration. What role does a cancer diagnosis play in cancer management, and how can molecular data contribute to the same? Here we explore this question within the realms of a precision oncology trial for managing treatment-resistant cancers.

In bioinformatics analysis of cancers, an accurate diagnosis must occur after the raw sequencing data has been processed into sample-specific gene expression values, and before aligning the changes to a reference tissue type. In the case study that follows, the patient presented with an adenocarcinoma of the vulva that was refractory to established lines of therapy. The initial diagnosis of vulvar adenocarcinoma, provided by a pathologist, was re-evaluated after the analysis of gene expression data. The resultant putative diagnosis was validated against clinical records and follow-up validation through immunohistochemistry. The molecular diagnosis was compared with the diagnosis from an experienced pathologist to determine the correct cancer type, subsequently guiding the selection of treatment options for the patient.

The Personalized OncoGenomics (POG) project at BC Cancer was established in 2011 with the aim to sequence and treat patients with advanced cancers [108]. Patients were enrolled after their cancer no longer responded to standard lines of therapy. The project analyzes the genomic and gene expression profiles of each patient's cancer in order to identify drugs that can target the individual cancer. The project has

had considerable demonstrated success, guiding targeted therapy and elucidating novel resistance mechanisms in highly aggressive cancers [96, 108].

Patients in the POG project are recruited and biopsied at BC Cancer and affiliated hospitals. Clinical laboratories at BC Cancer assess the site of origin from these biopsies according to established protocols. Subsequently, standard Illumina protocols are followed for whole-genome sequencing (WGS) of the tumour and peripheral blood (as control), and transcriptome sequencing (RNA-Seq) for the tumour. Sequencing is performed using Illumina HiSeq 2000 sequencers. The raw genomic and transcriptomic sequences are processed through a series of software tools to quantify the expression of genes, identify structural variants, mutations, and copy number changes. These findings need to be contextualized against a reference tumour type before we can draw inferences about clinically relevant changes in the patient's cancer. The following case study will help the reader gain an appreciation for a routine precision oncology workflow, and to appreciate the implications of expression based cancer diagnosis.

Patients in the POG project have typically received multiple lines of chemotherapy prior to enrolment in the program. As a result, their cancers usually acquire complex molecular profiles, and oftentimes have moved away from their site of origin (metastasized). Due to this, identifying the cancer's site of origin is a major challenge.

2.1 Case study

Mammary-like glands in the vulva were first reported in 1872, and thought to be supernumerary breast tissue remnants located along the milk lines [204]. Current understanding suggests that these are modified vulvar eccrine glands, that can give rise to vulvar adenocarcinomas [204]. Vulvar cancers represent 5% of gynecologic cancers and <1% of all cancers in women [14]. Approximately 90% of vulvar cancers are squamous cell carcinomas, and associated primarily with high-risk human papilloma virus (HPV) [49]. Most of the remaining vulvar cancers are not associated with HPV, and are typically vulvar adenocarcinomas. This category includes primary tumours arising from the vulva, and metastases to the vulva [142]. A determinative diagnosis of vulvar adenocarcinomas is complicated, and can encompass primary adenocarcinomas (mammary-like, mucinous, adenoid

cystic, Bartholin gland, and extramammary Paget disease) and metastatic disease. Metastatic adenocarcinomas to the vulva forms 5-8% of all vulvar cancers [142].

Breast cancers are the most common malignancy affecting women in North America [5]. In contrast, mammary-like adenocarcinomas of the vulva (MLAV) are rare, locally aggressive tumours that arise from the vulva but strongly resemble primary breast carcinomas [5]. They were initially reported in 1875, and at the time were thought to be breast tissue remnants located along the milk line [83]. Current understanding suggests that they are modified vulvar eccrine glands that can give rise to several different tumours, including vulvar adenocarcinomas [204]. They can metastasize to lymph nodes in approximately 60% of cases, and recur frequently after treatment [5]. Treatment guidelines are traditionally the same across vulvar carcinomas, but more evidence now suggests the transposition of breast cancer treatment regimens to MLAV; these include sentinel node biopsy, molecular subtyping, and adjuvant therapy [1, 21, 153].

Herein we describe the case study of a patient who presented with a poorly differentiated vulvar adenocarcinoma. The tumour was subsequently reclassified as a HER2+ MLAV upon transcriptomic analysis. The molecular profile of the case also aligned more strongly with breast cancer over gynecologic cancers. Put together, these findings suggest molecular likeness between breast cancers and MLAV, adding further support to the transposition of breast cancer regimens to this rare cancer type [1, 153]. The case also highlights the utility of genomics in resolving complex diagnoses. To our knowledge, this is the first case of a mammary-like adenocarcinoma of the vulva being described with detailed whole-genome and transcriptome sequencing analysis [126].

2.1.1 Clinical background

A 60-year-old woman presented with a poorly differentiated, bleeding mass in the vulva. There was no family history of cancer malignancies. PET/CT scans determined it to be a stage IV malignancy with a regional spread from the bilateral inguinal area to the retrocaval lymph nodes, external iliac, and common iliac lymph nodes. In total, three masses were noted on physical examination - a 2.5 cm firm right labium maius mass, a 2.5 cm bleeding vaginal introital mass, and bilateral inguinal lymphadenopathy up to 3.0 cm. The entire vulva was severely flaky, but the no specific skin

lesion was observed. The Bartholin gland, from where Bartholin gland carcinoma can arise in 1% of all vulvar neoplasms, was not proximal to the right labium maius. The largest lymph node was a left external iliac lymph node measuring 3.6 cm. No breast masses were identified in the physical examination. Furthermore, the patient reported that a history of remote mammograms had shown no malignancy. Bilateral mammograms were also negative for breast malignancy. Chest X-ray showed no metastatic pulmonary disease. A hypermetabolic mass in the medial aspect of the right labium maius (3.2 x 2.1 cm, maximal standardized uptake value (SUV) of fluorodeoxyglucose (FDG) 24.0) and an adjacent FDG-positive circumferential mass in the vaginal introitus (3.0 x 2.0 cm, maximal SUV 23.6) were detected using positron emission tomography/computed tomography (PET/CT) scans. The uterus was enlarged but showed on physiologic uptake of FDG. High SUV of FDG is indicative of cells having a high metabolic rate, and can reveal differences in glucose consumption of various cancerous lesions. No other putative primary tumour sites were identified. The clinical and radiologic findings were consistent with a stage IV vulvar cancer, with metastasis to the bilateral inguinal, retrocaval, external iliac, and common iliac lymph nodes. An initial vulvar biopsy was taken at this point and pathology findings report "poorly differentiated infiltrating carcinoma, favor [sic] poorly differentiated adenocarcinoma" (see Section 2.1.3).

The patient was treated with four rounds of carboplatin and paclitaxel, with a positive response observed in all areas apart from the inguinal lymph nodes. The response was assessed using repeat imaging by PET/CT. Sequential radiotherapy was then given to the entire spread of the disease at baseline. This was accompanied by additional radiotherapy boosts to the still FDG-avid inguinal lymph nodes. A rapid recurrence was observed in the patient's left supraclavicular lymph nodes six weeks after the completion of radiotherapy. A fine needle aspirate confirmed this to be a poorly differentiated adenocarcinoma consistent with metastatic vulvar adenocarcinoma.

In the absence of subsequent standard treatment options, the patient was enrolled in the POG project at BC Cancer to identify actionable targets and to validate the clinical diagnosis of vulvar adenocarcinoma. At the time of enrollment in POG, the patient had no family history of cancer. No additional genetic testing was done, and no other treatment was received between the initial vulvar biopsy and presentation of metastasis in teh left supraclavicular lymph node. The sample from the left supraclavicular lymph

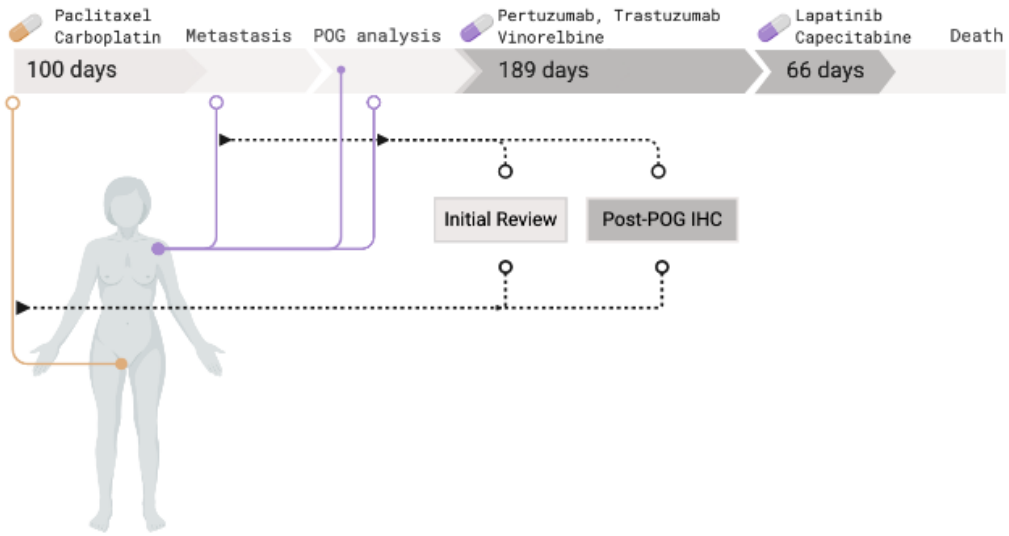

Figure 2.1: Clinical history and pathology sampling timepoints for MLAV patient. Initial treatment is indicated in orange, tumour biopsies at various time-points following metastasis indicated with purple lines, and treatments provided based on genomic analysis are shown with purple drug symbols over dark-grey timeline bars. Tumour biopsies on which immunohistochemistry was performed are shown with open circle termination of corresponding line. Abbreviations: IHC - Immunohistochemistry test, POG - Personalized OncoGenomics clinical trial (Clinical Trial number: NCT02155621).

node (subsequently referred to as the recurrence biopsy) was submitted to POG for sequencing and analysis. A detailed clinical time-line is shown in Figure 2.1.

2.1.2 Methods

Ultrasound guided core-needle biopsies were obtained for the POG study. FISH assays and IHC were performed by the clinical laboratories at BC Cancer according to established protocols. The rabbit monoclonal antibody for HER2 (clone 4B5; Ventana Medical Systems) was used for HER2 protein staining. Immunostaining was performed on the Ventana Benchmark Ultra automated system (Ventana Medical Systems) with 36 minutes of ULTRA CC1 before being incubated with the prediluted HER2

antibody for eight minutes at 36°. The ultraView DAB detection kit was used with an ultraWash step.

RNA and DNA were extracted and sequence libraries constructed using standard protocols (summarized in Table 2.1). Sequencing was performed on an Illumina HiSeq2500 platform at the Canada's Michael Smith Genome Sciences Centre (GSC). One microgram each of DNA from normal blood and tumour biopsy were separately used as input to the GSC Polymerase Chain Reaction (PCR)-free WGS protocol, and sequenced to 43x and 90x coverage respectively. 1.725 microgram of total RNA from the tumour was treated with the strand-specific messenger RNA sequencing protocol with poly-adenylated reads capture, and sequenced to a total of 291 million reads. The reads were aligned to the GRCh37 reference human genome using BWA v0.5.7 [113]. Duplicate reads were marked using Picard (v1.38, https://github.com/broadinstitute/picard/). Microbial and viral integration detection analysis was done using an in-house pipeline and BioBloom Tools [34]. WGS variants identified using Samtools v0.1.7 mpileup [113].

The tumour and normal samples were compared to identify somatic events. Somatic single nucleotide variants (SNVs) were called using Strelka v0.4.62 [174] and MutationSeq v1.0.2 [52]. Strelka v0.4.62 was also used to call small insertions and deletions. The somatic variant annotation was done with the Ensemble database (v69), and the effect calculation was assisted by annotations from snpEff 3.2 [36], COSMIC v64, and dbSNP v137. LOH events and tumour content were determined with APOLLOH v0.1.1 [75]. Copy number variants were identified using CNAseq v0.0.6 (https://www.bcgsc.ca/platform/bioinfo/software/cnaseq).

RNA-Seq data was analyzed using JAGuaR v2.0.3 [22]. The RNA-Seq data was subsequently processed by an in-house pipeline for Whole-Transcriptome Shotgun Sequencing coverage analysis, to yield exon- and transcript- level read counts and normalized expression values (Reads Per Kilobase of transcript per Million mapped reads, RPKM). Gene-level RPKM values were then calculated based on a collapsed gene model. Fold change for each gene was calculated by dividing each gene's RPKM value against an average of the RPKM values for the gene in a compendium of adjacent normal tissue samples from the Illumina Human BodyMap 2.0 project. A percentile ranking of the RPKM of each gene against the compendium of breast cancer transcriptomes from TCGA was used to identify genes with aberrant expression and to prioritize genes of interest.

Expression correlation analysis for tumour typing was undertaken relative to the entire set of normal and tumour transcriptomes in TCGA. Two-way Analysis of Variance (ANOVA) was used to identify genes that distinguished each pair of TCGA tumour types. This resulted in a set of 3,000 genes that were the most informative in explaining patterns of variance amongst all TCGA tumour types. A spearman correlation was calculated for this set of genes from the tumour sample against each TCGA sample. These pairwise correlations were clustered by the disease status (tumour or normal) and cancer type of the TCGA samples. The cancer set with the highest median correlation was determined to be representative of the closest cancer type for the sample.

Table 2.1: Details of sequencing experiments.

Sample	Type	Input micrograms	Library protocol	Coverage	Reads total
Biopsy tumour	DNA	1.000	PCR-free WGS	90x	NA
Biopsy tumour	RNA	1.725	ssRNA-Seq	NA	291 million
Normal blood	DNA	1.000	PCR-free WGS	43x	NA

2.1.3 Pathology analysis and findings

Pathology analysis was conducted on two biopsies - the initial vulvar mass, and an aspirate of the recurrence in the supraclavicular lymph node. The initial vulvar mass was evaluated twice - once at the time of initial biopsy, and subsequently following the indication of a mammary-like adenocarcinoma from the POG project's genomic analysis.

Initial vulvar biopsy assessment

An initial biopsy of the vulvar mass showed nests and cords of large pleomorphic epithelioid cells with eosinophilic cytoplasm and hyperchromatic nuclei. Gland formation, papillary structures, or intra-luminal vacuoles were not evident, except for some occasional cells showing possible signet-ring features and intra-luminal vacuoles. Mitotic figures were easily identified. No mammary-like glands or overlying epidermis were present. Based pm immunohistochemical (IHC) workup at the receiving hospital, the tumour was assessed positive for CK7

and Ber-EP4, and negative for CEA, CK5, CK20, MART-1 and S100. PAS-diastase was negative for definitive intra-luminal murin.

These initial findings were consistent with a poorly differentiated adenocarcinoma. Moreover, the poorly differentiated morphology and nonspecific IHC profile in the initial biopsy resulted in a broad differential diagnosis that included poorly differentiated vulvar squamous cell carcinoma, poorly differentiated vulvar adenocarcinoma (Bartholin gland adenocarcinoma, MLAV, or adenocarcinoma arising from extramammary Paget disease), and metastatic adenocarcinoma (from the gastrointestinal tract or gynecologic organs), and melanoma.

Post-POG pathology analysis

Subsequent to the genomic analysis favoring a diagnosis of mammary-type carcinoma (described in Section 2.1.4), follow-up validation stains were performed at BC Cancer. This validation work was carried out on the initial vulvar biopsy and on a repeat aspirate from the left supraclavicular lymph node.

Initial biopsy

Additional validation stains on the vulvar biopsy were performed for validation at BC Cancer, to evaluate the POG-indicated diagnosis of mammary-type carcinoma. IHC for HER2, the protein product of *ERBB2* gene, was equivocal (score 2 as evaluated according to ASCO/CAP guidelines [218]), as indicated in Figure 2.2. This prompted reflex testing for HER2 amplification in the genome, using Fluorescent *In Situ* Hybridization (FISH) testing. This showed the HER2/CEP17 ratio was 2.0 with 20 cells counted, and an average HER2 copy number per cell of 6.35. Although the HER2/CEP17 ratio was equivocal, HER2 was interpreted as amplified based on HER2 copy number equal or greater than 6.0 signals per cell as per the 2013 ASCO/CAP guidelines [218]. These guidelines require "complete, intense staining" of the circumferential membranes of >10% of tumour cells.

Additional IHC testing revealed that the tumour was negative for ER, PAX8, GCDFP-15 and mammaglobin, and focally positive for vimentin. The negative CEA, GCDFP-15 and PAS-diastase excluded the diagnosis of adenocarcinoma arising from extramammary Pagets disease (EMPD). The negative CK5 ruled out squamous cell carcinoma. The negative ER, GCDFP, and mammaglobin discounted the possibility of luminal A/B-types of primary MLAV. The negative PAX8 and CK20 ruled out metastatic

gynecologic and lower gastrointestinal tract carcinomas, respectively. The negative MART-1 and S100 (tested previously) were contrary to expected indications for melanoma. The overall pathologic findings from the post-POG workup of the initial vulvar biopsy were in keeping with high-grade ER-negative, HER2-positive mammary-type carcinoma of the vulva.

Recurrence biopsy

A fine-needle aspiration (FNA) of the recurrent disease in the left supraclavicular lymph node was collected and showed poorly cohesive irregular glandular clusters of pleomorphic malignant cells consistent with metastatic carcinoma. A repeat FNA of the site, collected for IHC testing, demonstrated that the sample was strongly positive for GATA3, a relatively recent marker for breast cancer [130]. HER2 IHC was positive (score 3+) based on 30% of tumour cells showing strong circumferential membranous staining. ER was negative, identical to the initial vulvar biopsy. Histological estimates placed the tumour content of the recurrence biopsy at 69%, with 85% cellularity.

Alternative diagnoses and exclusion of differential diagnoses

HER2 overexpression can also be observed in adenocarcinomas arising from EMPD. A 2005 study of patients with mammary and extramammary Paget disease observed co-expression of *ERBB2* and *AR* in 88% (51/58) of cases with mammary Paget [115]; the genomic analysis for this case also showed high expression of *ERBB2* and *AR*. With this observation in mind, invasive carcinoma arising from primary EMPD, rare anogenital tumours with proposed precursors that include Toker cells, pluripotent germinative cells, eccrine or apocrine glands, and mammary-like glands [102, 217] were considered for differential diagnosis during the validation pathology workup of the recurrence. The positive staining for CK7, GATA3, and HER2, and the negative staining for ER overlapped with previous reports of EMPD [48, 133, 154]. However, CEA, a nonspecific immuno-stain that is positive in most cases of primary EMPD [19], was negative in this patient's tumour. Furthermore, the overlying epidermis was not seen in the vulvar biopsy. Thus, the presence of pagetoid cells, which are necessary for the diagnosis of Paget disease and carcinomas arising therein, could not be assessed. The confluence of the infiltrating tumour also favor mammary-type carcinoma over Pagets disease, which is characteristically more superficial [149]. In summary the IHC findings supported the diagnosis of MLAV.

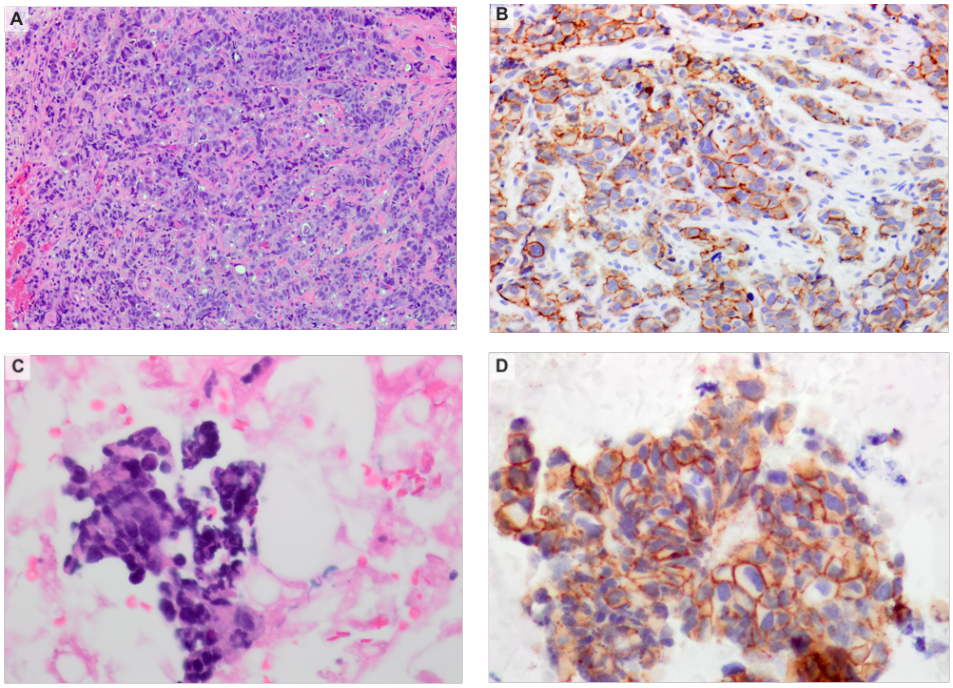

Figure 2.2: Histopathology of biopsies retrieved from MLAV Patient. A) The biopsy of the vulvar mass shows a poorly differentiated tumour composed of nests and cords of pleomorphic tumour cells. B) The HER2 immunostain on the initial vulvar mass biopsy is equivocal, compatible with score 2+ based on predominantly incomplete, weak and moderate membrane staining within greater than 10% of tumour cells. C) The fine needle aspirate of the recurrence lesion from the supraclavicular lymph node shows clusters of pleomorphic tumour cells (H&E stain). D) The HER2 immunostain of the supraclavicular lymph node shows tumour cells with complete, intense membrane staining in greater than 10% of tumour cells compatible with score 3+.

2.1.4 Genomic analyses

The recurrence sample was submitted to POG for whole-genome and transcriptome sequencing and analyses. In the absence of subsequent options for standard treatment for the patient, the aim of this exercise was to a) identify potentially actionable genomic targets, and to b) clarify the diagnosis and evaluate the validity of the initial diagnosis as vulvar adenocarcinoma. A constitutional blood sample and tumour from the lymph node biopsy were sequenced to a redundant sequence coverage depth of 43-fold and 90-fold, respectively. A transcriptome of 291 million sequence reads was also generated from the same tumour sample. The genomic and transcriptomic findings indicated the cancer was a mammary-like adenocarcinoma. The association of specific genomic events and transcriptomic changes with breast cancer was confirmed upon detailed literature review and integrative analysis.

Somatic Variants

Somatic variants were identified by comparison of paired-end WGS results from the tumour sample and the blood (germline reference). These variants were subsequently filtered to discard known artifacts and low-confidence variants, resulting in 375 single nucleotide variants (SNVs) and 15 insertion/deletion events (INDELs). A subset of 16 non-synonymous protein-coding SNVs were present in the COSMIC database and were considered to be the most relevant variants. No structural variants of clinical significance were identified. Copy number variant (CNV) analysis indicated a triploid karyotype with an estimated 68% tumour content, consistent with the pathology estimate of 69%. Focal copy number amplifications were detected on chromosomes 2, 8, 9, 17, and X. Screening for microbial and viral sequences was negative for any microbial contaminants. There was no evidence of HPV genomic integration either.

A gain of function mutation (p.S310F) was observed in *ERBB2* with prevalence level of 86%. This event was accompanied by a copy number gain (five copies), as shown in Figure 2.3. The S310F mutation has been identified in several cancers including breast, lung, and ovarian [88].

Loss of function mutations were observed in *TP53* and *RB1* tumour suppressor genes. These genes overlapped with regions of loss of heterozygosity (LOH) in the copy number landscape. Variants of unknown significance were also noted in *PIK3CA, AKT3*, and *GNAS*. All SNVs of

interest are summarized in Table 2.2.

A triploid model with an estimated 68% tumour content was inferred based on the prediction of allelic imbalance and loss of heterozygosity in the sample, as already described in Section 2.1.2. Copy-number variants were estimated with respect to this ploidy model. Focal copy-number amplifications were detected in Chromosomes 2, 8, 9, 17, and X. Of particular interest, copy number gains were observed for *ERBB2*, *AKT3*, *PIK3CA*, *CDK1*, *CCNB1*, and *AR*. Loss of heterozygosity events (LOH), arising from the loss of a single copy of the respective gene, were detected for the tumour suppressor genes *BRCA2*, *RB1*, and *TP53*. Additionally, *RB1* and *TP53* had loss of function mutations in the two remaining (homozygous) copies. These findings are summarized in Table 2.3.

Table 2.2: SNVs of interest are listed, along with details on the counts of the supporting reads spanning the tumour genome at the mutated and reference bases, in the tumour genome (transcriptome).

Gene	Chr	DNA Change	Variant	Alt/Ref (Alt_RNA/Ref_RNA)
AKT3	1	244006441 C>A	VUS	18/113 (0/11)
ERBB2	17	37868208 C>T	GoF	165/26 (4181/625)
GNAS	20	57430298 C>G	VUS	7/39 (0/4)
PIK3CA	3	178938934 G>A	VUS	69/40 (102/24)
RB1	13	49033844 C>T	LoF	35/14 (294/59)
TP53	17	7577082 C>T	LoF	30/13 (230/24)
MAP3K12	12	53877268 C>T		7/62 (4/44)
OR14A16	1	247978827 G>T		
PAF1	19	39876915 G>A		10/37 (76/208)
PCDHA6	5	140208403 G>A		25/38 (0/0)
PPM1B	2	44428594 A>G		20/72 (71/232)
SLCO3A1	15	92669422 G>A		18/37 (9/21)
SNTG2	2	1271197 G>C		13/57
THBS2	6	169629714 C>T		8/53 (0/102)
UPF3A	13	115047496 G>C		9/15 (0/25)
ZNF830	17	332893990 C>T		15/39 (51/74)
ZXDB	X	57618845 G>A		11/22 (2/16)

Table 2.2: SNVs of interest are listed, along with details on the counts of the supporting reads spanning the tumour genome at the mutated and reference bases, in the tumour genome (transcriptome). *(continued)*

Gene	Chr	DNA Change	Variant	Alt/Ref (Alt_RNA/Ref_RNA)
ZXDB	X	57618849 A>C		11/21 (3/17)

Abbreviations: AA, amino acid; Alt, coverage of alternative allele; Alt_RNA, RNA reads mapping alternative allele; Chr, chromosome; GoF, Gain of function; LoF, Loss of function; Ref, coverage of reference allele; Ref_RNA, RNA reads mapping reference allele; SNV, single-nucleotide variant; VUS, variant of unknown significance.

Table 2.3: Copy number variants of interest in the tumour genome are listed, along with percentile values and fold changes calculated from the respective RPKMs against a background of TCGA Breast cancers.

Gene	Chr	Copy type	TCGA expression percentile	Fold expression change	Copy change in ploidy corrected model (versus 3n, triploid tumour)
AKT3	1	Gain	21	-5.18	+1 (HET)
ERBB2	17	Amplification	98	32.73	+8 (ALOH)
GNAS	20	Gain	2	-1.55	+1 (NLOH)
PIK3CA	3	Gain	49	-1.40	+1 (HET)
RB1	13	Loss	88	1.76	-1 (DLOH)
TP53	17	Loss	21	-1.10	-1 (DLOH)
AR	X	Amplification	100	26.39	+8 (NLOH)
BIRC5	17	Amplification	100	40.87	+5 (BCNA)
BRCA2	13	Loss	94	2.21	-1 (DLOH)
CDK12	17	Amplification	99	5.00	+8 (ALOH)

Table 2.3: Copy number variants of interest in the tumour genome are listed, along with percentile values and fold changes calculated from the respective RPKMs against a background of TCGA Breast cancers. *(continued)*

Gene	Chr	Copy type	TCGA expression percentile	Fold expression change	Copy change in ploidy corrected model (versus 3n, triploid tumour)
CCNE2	8	Amplification	100	25.71	+17 (ALOH)

Abbreviations:
ALOH, amplification with loss of heterozygosity; BCNA, balanced amplification; Chr, chromosome; DLOH, deletion with loss of heterozygosity; exprn, expression; GoF, gain of function; HET, heterozygous; LoF, loss of function; NLOH, neutral with loss of heterozygosity.

De novo assembly of the genome and transcriptome was performed to identify structural rearrangements of potential biological and clinical significance. However, none were detected.

Transcriptomic analysis

A pairwise expression correlation analysis was undertaken to compare the gene RPKM values from the sample with The Cancer Genome Atlas tumour samples (TCGA; see Section 2.1.2). This analysis, done across 27 different tumour types available from TCGA, indicated that the tumour sample correlated the most with the breast cancer (BRCA) cohort. Based on this observation, we replicated the PAM50 test by selecting the PAM50 set of genes to correlate the sample's transcriptome against TCGA breast cancer samples with known BRCA molecular subtype status. Consistent with the amplification and gain-of-function mutation in the *ERBB2* gene, the tumour sample correlated the highest with the HER2 enriched and Luminal B subtypes. These results are shown in Figure 2.4.

Based on the findings from the correlation analysis, the genomic events and RNA-level changes were considered against a background of breast cancers. A fold-change value for each gene was calculated against a normal breast tissue transcriptome (Illumina Human BodyMap 2.0) and a percentile rank of expression calculated in comparison to the breast cancer cohort from TCGA (detailed in Section 2.1.2). A fold change of -1.1 of *TP53* gene expression corroborated with the loss of function mutation identified in

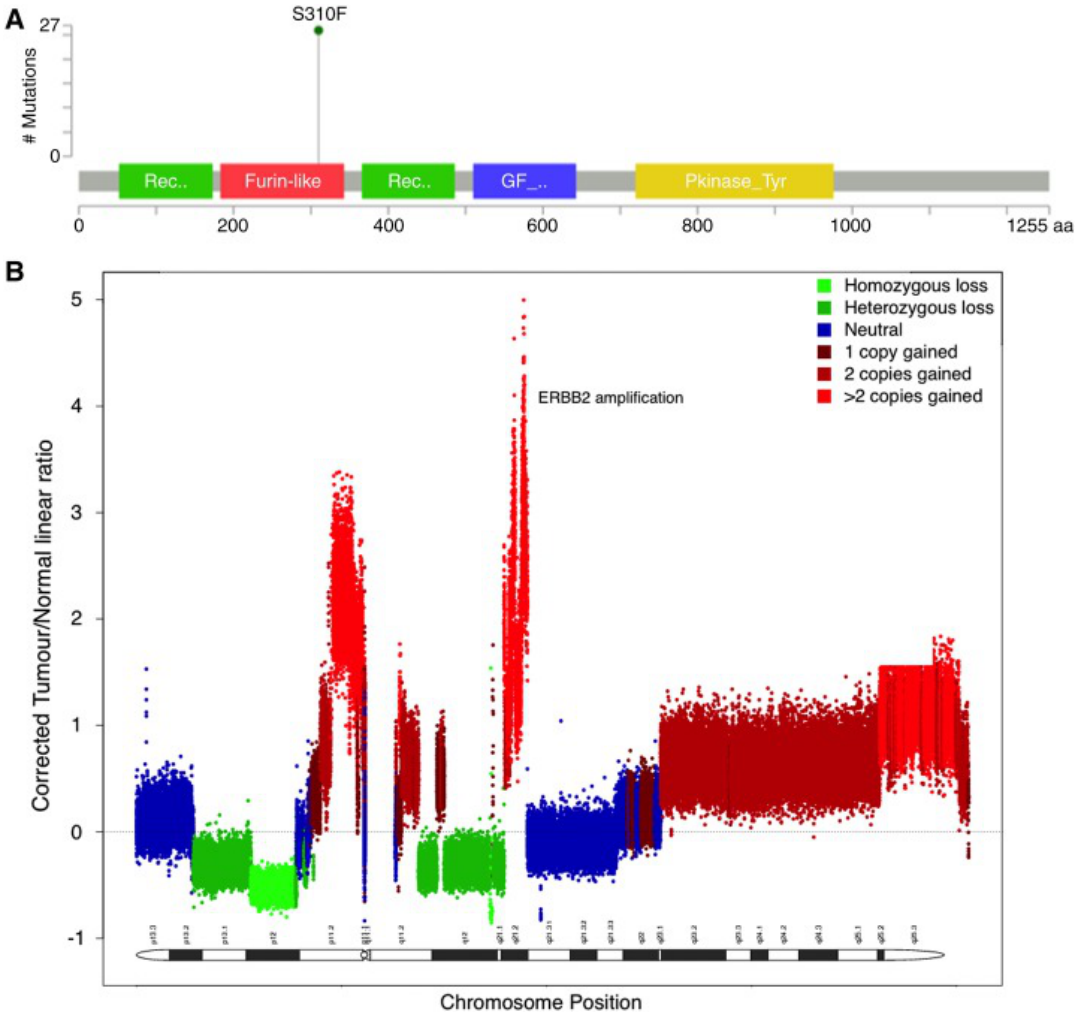

Figure 2.3: ERBB2 gene's genomic locus is shown in the patient's tumour. A) A lollipop plot showing the coordinates of the S310F gain-of-function mutation observed in this case. B) A plot of the copy number landscape of Chromosome 17 in the tumour. The ERBB2 copy-number gain is indicated.

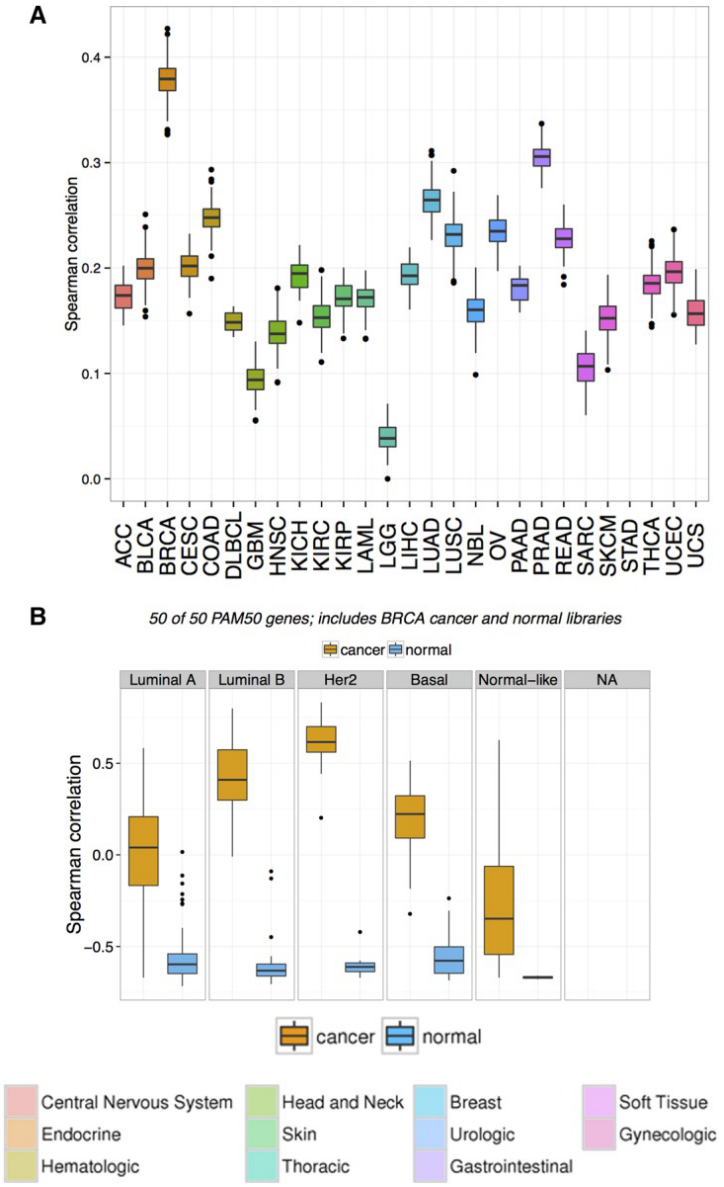

Figure 2.4: Correlation plots of the cancer's RNA-Seq profile with TCGA cancer datasets. A) Boxplot distribution of the pairwise Spearman correlation of the recurrence biopsy's gene expression profile and all TCGA samples. The x-axis represents cancer types following TCGA naming conventions. TCGA breast cancer cohort is indicated by BRCA. B) Boxplot distribution of the pairwise Spearman correlation between the recurrence biopsy and the TCGA breast cancer cohort based on the PAM50 set of genes. The pairwise correlations with adjacent normal are shown in blue.

TP53. The *ERBB2* gene, which had a gain of function mutation and copy number gains, also had an RNA-level fold change of 32.7 relative to the compendium average. Expression outliers were identified and evaluated in conjunction with mutational status, copy-number state, and known biological function. Nine genes of interest were identified having gains of more than three copies each, and also ranked in the 98th–100th percentile versus TCGA breast cancers (Table 2.3). Of particular interest among these genes were *ERBB2* (98th percentile), *CDK12* (99th percentile), *AR* (100th percentile), and *CCNE2* (100th percentile). The extreme outlier expression of *ERBB2* (33-fold overexpression, 98th percentile of BRCA) combined with the observed of the gain of function mutation (p.S310F) and estimated five copy gain (Figure 2.3) further supported a diagnosis of a HER2+ mammary-like cancer and identified HER2 as a likely driver of the disease.

Mutational signatures

To further evaluate the differential diagnosis of the tumour as a breast cancer, we considered the WGS mutational data against previously catalogued mutational signatures [97]. A strong APOBEC signature was observed (Signatures 2 and 13), and no HPV was detected. The APOBEC family of cytidine deaminases generates mutations of a specific pattern (the APOBEC signature mutation pattern), which has been reported in several cancers, including HER2+ breast cancers [134]. APOBEC activation has also been associated with HPV; however, screening for microbial and viral sequences was negative for any microbial contaminants and no evidence for genomic integration of HPV was detected.

2.1.5 Clinical decision and outcome

The diagnosis of the tumour as a mammary-like adenocarcinoma, and the presence of a clearly druggable target (HER2) led to the patient being treated with a 1st line standard of care for HER2+ breast cancer. A strong APOBEC signature is associated with HER2+ breast cancers, and because of its association with PDL1, has been positively correlated with response to immunotherapy in other cancers [134]. Unfortunately, at the time of this analysis, immunotherapy was not available as an accessible line of treatment and was not pursued further [138].

Transformed Ba/F3 cells harboring the S310F mutation in HER2 (as in this

tumour) have been shown to be sensitive to neratinib, afatinib, lapatinib and trastuzumab. In consideration of these findings, the patient was treated with trastuzumab, pertuzumab, and vinorelbine, followed by capecitabine and lapatinib. The patient had a poor clinical response to all targeted therapies tried, at best achieving short-term disease stabilization but never achieving regression of disease (see 2.3 Conclusion for potential explanation of this behaviour). Future disease progression included the development of cutaneous lesions in the left shoulder and back. An FNA sample of the region confirmed this to be a metastatic carcinoma. The patient passed away two years and five months after her initial diagnosis.

2.2 Summary

The diagnosis and classification of vulvar adenocarcinomas is a complicated and understudied area, as this is a rare histologic subtype of vulvar cancers. The differential diagnoses include MLAV, adenocarcinoma arising from EMPD, mucinal carcinoma, Bartholin gland adenocarcinomas, and metastatic adenocarcinomas from various sites. In our case, the clinical, radiologic, and histologic features indicated a poorly differentiated (high-grade) primary vulvar adenocarcinoma. The patient was enrolled in the BC Cancer POG project upon the development of new metastases, with the two aims of characterizing the underlying genomics of this poorly differentiated cancer and identifying actionable therapeutic targets.

The poorly differentiated morphology and non-specific immunoprofile of the initial biopsy had resulted in a broad pathologic differential diagnosis including mammary-type carcinoma, vulvar adenocarcinoma, and upper gastrointestinal adenocarcinoma. The subsequent bioinformatics analysis findings indicated the patient's tumour was most consistent with a HER2+ breast cancer profile. Post-hoc histopathologic investigations on the initial biopsy and a repeat aspirate from the supraclavicular lymph node site corroborated the determinative finding of HER2 overexpression from the bioinformatics analysis, and supported the diagnosis of MLAV. IHC was negative for the mammary markers mammoglobin and GCDFP-15, as is the case for most ER-negative breast carcinomas [38]. GATA3, a more recently established marker that is positive in breast carcinomas, was positive in the recurrence biopsy.

Additional genomic events, specifically the GoF S310F mutation in ERBB2,

the LoF mutations combined with LOH in TP53 and RB1, and the co-expression of ERBB2 and AR at high levels, pointed to a mammary-like cancer. It has been shown that RB1 LOH occurs at a higher frequency in basal-like and luminal B breast cancers [87], suggesting a potential role of RB1 LOH as a predictive marker in these subtypes. No definitive molecular support was found for ER over-expression.

2.3 Conclusion

As demonstrated in this case, a visual inspection of gene expression correlation patterns against independent primary cancer datasets can give us clues about the molecular behaviour of a cancer. Insights driven by RNA-Seq comparisons with established cancer types can lead to a change in management of a patient with terminal disease, providing a brief but eventually insufficient respite from a rare malignancy. In the larger scale of things, this study adds to the body of evidence for treating these rare cancers, MLAVs, as an ectopic breast tissue malignancy [126].

At the biological level, the findings emerging from this analysis lend support to existing observations in literature about mammary-like carcinomas. AR overexpression is found in 60% of breast cancers and is generally observed more frequently in ER+ breast cancers than ER- ones. However, when present, AR expression is significantly correlated with HER2 expression in ER- breast cancers, and a proliferative role for AR has been suggested recently in ER-, HER2+ patients [129, 156]. An overexpression of CDK2 and a high-percentile expression of CCNE2 were observed in the recurrence sample's transcriptomic analysis. These have been suggested as potential resistance markers for trastuzumab [199], and we can speculate on their potential role in rendering the treatment ineffective.

Additionally, post-genomic analysis findings from the pathology workup and validation tests provide evidence for the improved ability of emerging histopathology markers like AR and GATA3 in diagnosing breast cancers, as compared to more established ones like mammoglobin and GCDFP-15. A recent IHC study showed that MLAV can be classified into four breast intrinsic subtypes, including a HER2+/ER− group [193, 202]. During the initial pathology workup for the vulvar biopsy from this patient, the diagnosis of MLAV was not favored on the basis of ER, mammaglobin, and GCDFP-15 negativity. However, primary breast carcinomas with high

nuclear grade are mostly ER− [139], and a recent study has proposed GATA3 as a more sensitive marker for HER2+/ER− breast carcinomas than mammaglobin and GCDFP-15 [179]. The analyses presented herein are consistent with this body of evidence and confirm that ER negativity is consistent with the diagnosis of HER2+ MLAV.

How often does realignment of diagnosis happen as a result of detailed genomic analysis? Are there certain cancer types that are particularly refractory to routine histopathology based diagnosis? If so, can we develop diagnostic approaches that incorporate high-dimensional sequencing data to provide robust and confident assessments of a cancer type? The next chapter delves deeper into these questions through a cohort-wide analysis of cases that have been subject to similar detailed genomic analysis using whole-genome sequencing and RNA sequencing.

Chapter 3

Impact of genomics on diagnostic pathology in a precision oncology trial

```
## Warning: package 'stringr' was built under R version 3.5.2
```

Health care institutions now offer a varied selection of genetic testing to guide treatment options for cancer patients in the clinic. As discussed in Chapter 1, this includes prognostic panels like MammaPrint and Oncotype Dx [24, 181], and single-marker tests for prognostic genomic changes like mutations (for example, in KRAS, EGFR, or IDH1), and fusions like ABL-BCR (DeVita et al. [51]) or EWSR1 (for sarcomas). However, in the absence of a corollary cancer diagnosis to guide the companion test selection, or when none of the informative and actionable targets assessed by these panel-based approaches are present in a cancer, whole-genome and RNA sequencing provide an unbiased and detailed molecular view of the disease. Over the recent years, these sequencing modalities have enabled a dramatic expansion in the scale and content of cancer characterization and management within the scope of research investigations.

Precision oncology can be described as the use of DNA and RNA sequencing to facilitate discovery and analysis of molecular changes that impact patient management and treatment. The sequencing modalities vary, ranging from targeted deep sequencing of a subset of genes [30], to sequencing the mRNA-coding regions (exome), the entire genome (whole-genome), or the transcribed RNA (RNA-Seq). Precision oncology trials typically use one of exome (WES) or whole-genome (WGS) sequencing, combined with RNA-Seq [131, 146, 169]. As the price of sequencing decreases and more healthcare research facilities develop sequencing capabilities at scale, routine comprehensive DNA and RNA sequencing can be expected to make its way into cancer management [35, 169].

While advanced molecular techniques are being adopted steadily through precision oncology clinical trials, the interpretation of these assays is an ongoing challenge. Analysis and identification of therapeutically relevant genomic and transcriptomic changes is a tiered process. The raw sequencing data in the form of reads is filtered for microbial contamination, assembled and aligned to the human reference genome, and based on the sequencing modality, genomic variants are identified or a quantification of reads mapping to loci of interest (exons, transcripts) is made. Genomic variants specific to the tumour are identified through comparison with the assembled healthy normal tissue genome obtained from the patient, also known as the germline genome. The quantified reads obtained from RNA-Seq are normalized, usually to the library size and the gene length, yielding reads per kilobase per million mapped reads (RPKM) values for each gene/transcript. However, these expression measurements cannot be interpreted in isolation and have to be compared against a healthy tissue that matches the tissue of origin of the cancer type. The healthy tissue RNA-Seq reflects what a normal expression profile from that primary site would look like, helping us pinpoint genes and pathways that are aberrantly expressed. The degree of aberration is placed in context with other primary cancers reflecting the same cancer biology. Changes at the gene level (genomic or transcriptomic) are then assessed manually by a computational biologist, who summarizes and contextualizes them within biological pathway diagrams. These diagrams pin-point important biological associations that have been impacted in the tumour, and serve as putative therapeutic options.

This process provides a vast amount of information about a single cancer sample after drawing upon pre-existing datasets of healthy tissues and primary cancers. Recently published clinical trials that utilized sequencing data to do this have found that this approach can provide a hitherto unmatched level of insight into the mechanisms driving metastatic cancers [169, 225]. No studies to date have assessed the effectiveness of combining large-scale sequencing protocols with histomorphology in guiding the analysis and management of advanced cancers. Successful precision oncology based analyses require the identification of a suitable healthy tissue comparator, which in turn requires knowledge of the cancer type. We reviewed a series of cases in which whole-genome and transcriptome sequencing was performed as part of the POG trial in Canada, considering this data in the context of tumour histomorphology. Our goal was to evaluate the impact of integrating WGS and RNA-Seq on pathological

diagnosis, and to understand how this data impacted subsequent biomarker testing and selection of targeted therapy.

3.1 Methods

3.1.1 Consent and institutional review board process

This research project was approved by the BC Cancer Agency Research Ethics Board (protocols H14-00681 and H12-00137). Cancer patients with advanced disease who failed conventional treatment and fulfilled the inclusion criteria were consented for tumour profiling using RNA-Seq (tumour) as well as whole-genome sequencing (tumour and blood) (Clinicaltrials.gov ID: NCT02155621).

3.1.2 Tissue biopsy and processing

A fresh tissue biopsy was mandatory for all patients participating in this study. Samples were taken from metastatic or recurrent tumours through needle-core biopsies or surgical resection, with guidance from imaging. Matching normal DNA was extracted from peripheral blood leukocytes. The samples were snap-frozen and anchored in a small amount of optimal-cutting-temperature (OCT) compound for cryo-sectioning. These sections were used for DNA and RNA extraction, after being assessed for histologic correlation. The snap frozen tissue specimens were cryo-sectioned at 50 μm for nucleic acid and protein extraction, and at 5 μm for hematoxylin-eosin (H&E) staining every 200 μm. Cases were excluded if the tumour content of the sections was less than 40% by pathology review. The intervening sections were placed into RNAse-free Eppendorf tubes. Only a small amount of OCT compound was used to bind the tissue to the chuck of the cryostat since OCT is known to inhibit downstream nucleic acid extraction and PCR steps [95].

3.1.3 Library construction and sequencing

Paired-end DNA and RNA sequencing libraries were generated at Canada's Michael Smith Genome Sciences Centre, and sequencing was performed using the HiSeq platform (version 3: Illumina, San Diego, CA, U.S.A).

Average coverage for WGS was 80-100x on frozen tumour tissue and on germline DNA from blood. cDNA libraries for RNA sequencing were prepared from biopsy samples using strand specific RNA-Seq Sample Preparation kit (stranded, polyA+) from Illumina. Sequencing was performed on the Illumina HiSeq 2500 platform. A minimum of 200x coverage per sample was required for the targeted amplicon reads. RNA-Seq data was analyzed with JAGuar, and subsequently processed with previously published in-house pipelines to yield exon- and transcript-level read counts and RPKM values. Gene level RPKM values were calculated using a collapsed gene model. The bioinformatics analysis pipelines used for this study are as described already in Chapter 2, Section 2.1.2.

3.1.4 Determination of tumour type

Each case was reviewed during weekly tumour board meetings by a multidisciplinary group of pathologists, medical oncologists, bioinformaticians, and computational biologists. Data for determining the pathologic diagnosis of the case was gathered from all four main components of each case's comprehensive analysis: 1) clinical background, 2) histomorphology of the tumour specimen, 3) gene expression, and 4) mutation profiling and structural rearrangements.

1. Relevant clinical history was presented by the treating medical oncologist for each case. This information was provided to the genome analysts/bioinformaticians and pathologists atleast four weeks prior to the discussion of the case so as to facilitate any relevant clinical interpretation.

2. The histomorphology assessment was gathered from previous pathology reports including diagnostic biopsies and/or resection specimens collected as part of routine management of the patient. All cases were reviewed internally centrally by an expert specialty-practice pathologist prior to being analyzed by the computational biologists and being presented to the multidisciplinary board.

3. Expression correlation analysis for tumour typing was undertaken relative to the entire set of normal and tumour transcriptomes in TCGA, as described in Section 2.1.2. Supervised Cancer Origin Prediction using Expression (SCOPE), described in detail in Chapter

4, was also evaluated as an ancillary automated tool to predict the cancer type from RNA-Seq expression profiles. Based on these two approaches, a reference cancer type was established. Fold-level changes in gene expression for the sample were calculated against a background of GTEx samples whose biology most closely matched the site of origin of the reference cancer type. Individual gene changes in the sample were placed in the context of TCGA samples from the reference cancer type, and aberrant genes and pathways highlighted through manual analysis by a computational biologist.

4. Mutation profiling was used to identify mutations and other genomic events that are known to be associated with certain cancer types. These associations arose from an internal knowledgebase curated through extensive literature review performed as part of the analytic pipeline for each case. When present, COSMIC variants were evaluated to assess support or confirmation of a particular diagnosis. The number of mutations present in the sample were compared against the TCGA reference cancer dataset to determine if the mutation burden was high or low.

Overall, a combination of mutation profile, mutation burden, gene expression based cancer type classification, integrative pathway analysis of gene expression changes and mutations, histomorphology report, immunohistochemistry, imaging, and other clinical metrics was used to suggest a tumour type/site of origin to the oncology clinical team. The cases were labelled as cancers of unknown primary (CUPs) when a specific diagnosis, including site and tissue type, could not be rendered after the pathology work-up of tumour tissue from the biopsied specimen.

3.1.5 Assessment of clinical input of whole-genome and transcriptome analysis in pathology

Genomic contribution to pathological diagnosis was assessed by reviewing the POG reports presented to the tumour board. These reports highlighted the pathway-level changes and key targetable genes based on the integrative analysis from bioinformatics. They also included information on the suggested diagnoses based on genomic and transcriptomic data, pathway analysis, and recommended lines of therapy. Two pathologists asked the following three questions.

Firstly, was the tumour diagnosis confirmed or re-aligned after the POG analysis, performed as described above? This was determined by comparing the final diagnosis arising from the POG analysis (as determined from the POG reports) to the initial pathology diagnosis (see Section 3.1.4).

Secondly, was the molecular subtype significantly changed after the POG analysis? This pertained to genomic events known to be clinically associated with specific cancer subtypes, for example, a lung adenocarcinoma that tested negative for *ALK1* fusion negative through FISH, was found to harbor an actionable *ALK1* gene fusion on WGS.

Thirdly, did the POG analysis augment the pathologic diagnosis workflow? It was said to 'augment' the pathologic diagnosis if a molecular alteration led to additional pathologic assessment (not already part of routine testing guidelines) of potential prognostic or predictive value without affecting the diagnosis. Of note, cases where genetic event(s) affected the oncologic management but could not be validated by histopathology methods were not said to augment the pathologic diagnosis workflow. Only those molecular abnormalities that could be confirmed by orthogonal clinically validated biomarker assays, in an accredited clinical laboratory and using tests already in clinical use, were considered. It should be noted that potentially 'actionable' molecular abnormalities were identified in most cases, but these would arise within the context of off-label therapies, the analysis of which was outside the scope of this study.

3.2 Results

3.2.1 Cohort demographics, clinical metrics, and sequencing data

An initial 492 cases were selected for this study from the POG trial based on material availability and completion of POG analysis. Of these cases, initial pathology diagnosis could be matched to a corresponding TCGA cancer type in 389 unique cases, which constituted the study cohort. Patients were 36% males and 64% female, and the 5-year survival was 45% with a median survival of three years. The cohort selection process is summarized in Figure 3.1. Breakdown of the outcomes, separated by tumour types, is shown in Figure 3.2.

A detailed breakdown of the 389 cases included in this study, separated

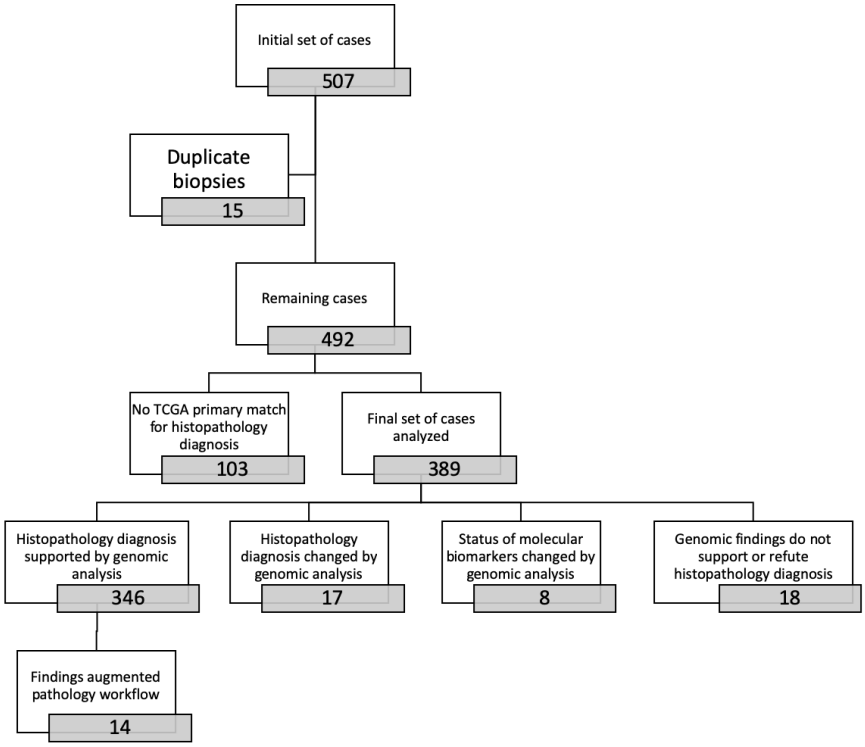

Figure 3.1: Cohort selection for the assessment of the impact of DNA and RNA sequencing analysis on histopathologic diagnosis in the POG clinical trial.

by the site of biopsy and by the diagnosed cancer type, is provided in Figure 3.4. As can be seen in this figure, the majority of tumours sampled were metastases to the liver, lung, or the lymph node. Cancer metastases accounted for 352/389 (90%) of the cases. In the remaining set, two (~1% of total) samples were of undetermined origin, 29 (7% of total) were primary tumours, and six (2% of total) were recurrences. Cancer metastases from the liver, lymph node, pelvis, soft tissue (like muscle), and the abdominal cavity, were the most likely to have the initial diagnosis revised by genomic analysis. Among the cancer types that were prevalent in the cohort, breast cancer, colorectal adenocarcinoma, sarcomas, and lung adenocarcinomas formed the majority of cases.

3.2.2 Correlation of histopathologic diagnosis and next generation sequencing results

3.2.2.1 Impact of genomic analysis on histopathologic diagnosis and prognostic marker identification

Most pathologic diagnoses were supported by the genomic analysis and clinical findings

Of the 389 total cases reviewed, 15 cases presented as CUPs. The integrated genomics analysis agreed with the original pathologic diagnosis in 346 of the remaining 374 cases (92.51%). In most cases, RNA-Seq provided evidence towards cancer diagnosis.

In 18 cases (4.81% of cases with a determinative initial pathologic diagnosis), the detailed genomic and transcriptomic analysis did not find any evidence that supported or challenged the initial histopathology diagnosis. The majority of these cases were pancreato-biliary cancers (two cholangiocarcinomas, four pancreatic adenocarcinomas), followed by five gynecologic cancers with non-specific carcino-sarcomatoid histomorphology, and three metastatic breast cancers. Tumour content of these cases, as estimated through bioinformatics analysis, ranged from 25-90%, similar to the rest of the cohort. Biopsy site and patient characteristics were also not significantly different from the rest of the cohort (Figure 3.2).

Genomic analysis serves as a robust refractory tool for identifying prognostic molecular markers

In a further eight (2.14%) cases, the molecular features of the cancer were

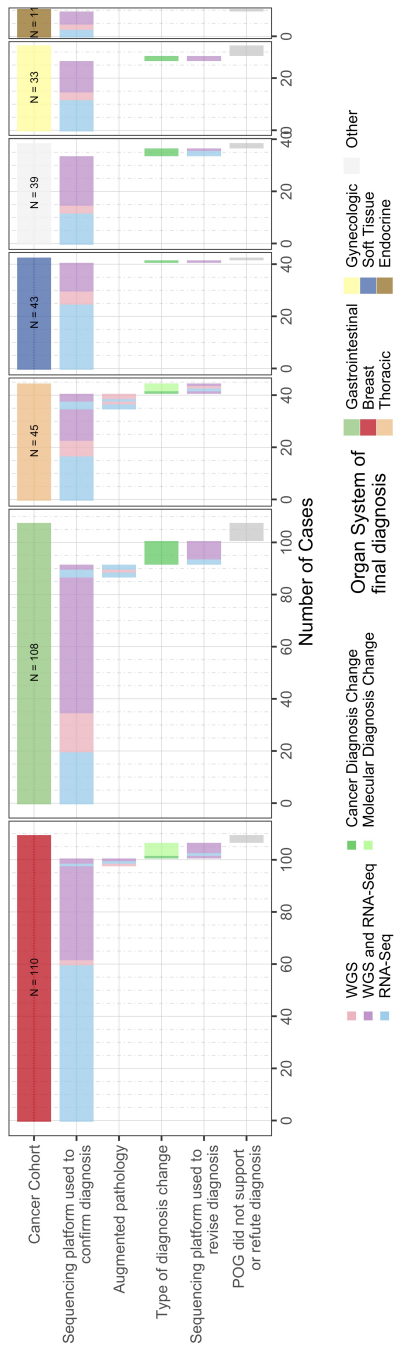

Figure 3.2: Tumour types in the cohort are shown, along with the type of genomic data guiding major outcomes from the retrospective analysis evaluating the diagnostic utility of RNA-Seq and WGS.

significantly redefined without a histopathology diagnosis change. These changes were significant to the extent of supporting different therapeutic regimens. Of those, five cases were breast carcinomas where the *HER2* status was changed after the genomic analysis, from *HER2* negative (determined using IHC, and FISH assays for equivocal cases) to *HER2* amplified based on the amplification/overexpression identified based on copy number analysis and gene expression profile. Two cases were lung adenocarcinomas where no driver mutations were identified on the routine IHC and gene sequencing panels, but where an activating mutation in *EGFR* (L858R and exon 20 insertion) was identified through the genomic analysis. The last molecularly redefined case was a lung adenocarcinoma where the anaplastic lymphoma kinase (ALK) fusion oncogene, an important marker of a molecular subtype of lung adenocarcinomas, was negative on initial pathology testing through FISH, but genomic analysis revealed an *ALK1* rearrangement (Figure 3.3).

3.2.2.2 Impact of genomic analysis on diagnosis of CUPs

Comprehensive whole-genome and transcriptome analysis identified misdiagnoses and putative primaries for CUPs

In two cases (0.53% of cases with a determinative initial pathologic diagnosis), the original pathology report diagnosis was found to be incorrect after molecular analysis and review. One case was initially diagnosed as a vulvar adenocarcinoma but gene expression analysis closely aligned with breast ductal adenocarcinomas. This triggered a detailed pathology review and validation through IHC, the diagnosis was adjusted to a *HER2* amplified mammary-like adenocarcinoma of the vulva, and the patient was treated with an *ERBB2* inhibitor (described in Chapter 2) [72]. The second case was initially diagnosed as adenocarcinoma of likely ovarian origin, but comprehensive analysis of the gene expression and mutation profiles supported the diagnosis of ovarian clear cell carcinoma. Evidence included high expression of *HNF1β*, *NAPSA* (Napsin A), *GPC3* (Glypican-3), an inactivating mutation in *ARID1A*, and copy gains in *HNF1β* and *ERBB2* genes.

In 15 cases (3.9% of total cases), the initial pathology workup and clinical assessments could not confidently assign a tumour site of origin or a histomorphologic category. Nine cases were initially diagnosed as adenocarcinomas of unknown origin, two as squamous cell carcinomas, three as carcinomas, and one as an unclassifiable malignancy. The

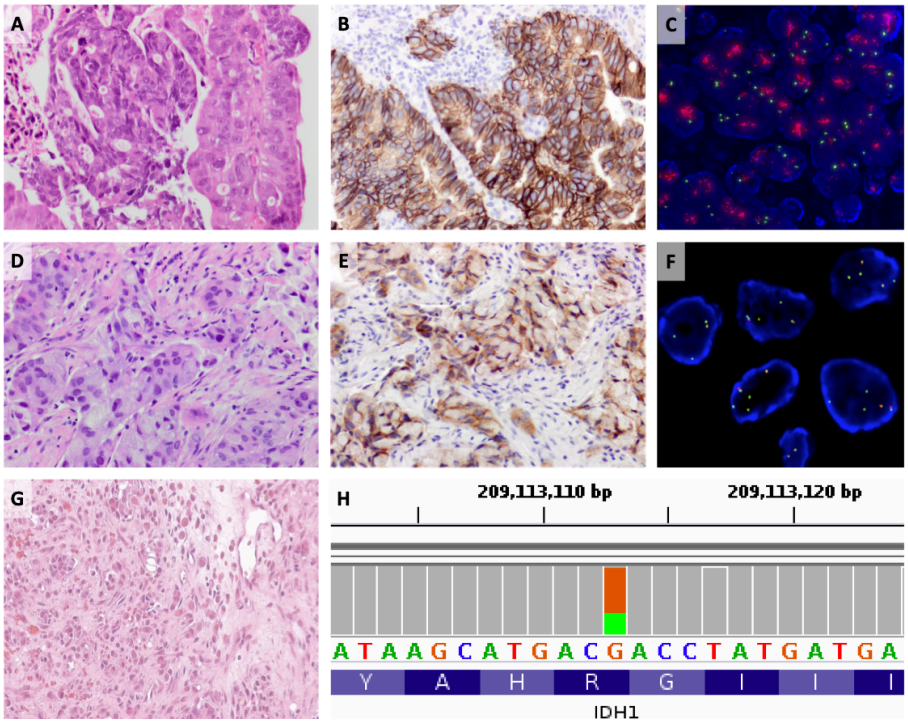

Figure 3.3: Detection of clinically relevant molecular alterations by whole-genome and RNA sequencing in the POG cohort. (A-C) Detection of HER2 amplification in a colorectal carcinoma is shown, as indicated by immunohistochemistry (IHC) staining for HER2 (overexpression, 3+) in the tumour sections in panels A) and B), and with FISH testing for additional copies of HER2 (HER2 to chromosome 17 centromere (CEP17) ratios > 2.0) in panel C). (D-F) ALK fusion identified in a lung adenocarcinoma, missed on initial FISH analysis. H&E staining of the tumour sample is shown in D). ALK IHC testing results showing equivocal ALK staining are represented in E), with the original negative FISH results (break apart probe test, less than 15% of cells showed break apart probes) shown in F). (G and H) Detection of an IDH1 mutation in a CUP supported the putative diagnosis of cholangiocarcinoma. The H&E staining is shown in G). Panel H) shows a snapshot of the Integrative Genomics Viewer track for the mutation location with proportional read-counts supporting the reference (G, in orange) and mutation (A, in green) in the tumour genome. This supported the putative diagnosis of this CUP as a cholangiocarcinoma in the clinical context, as aided by RNA-Seq analysis.

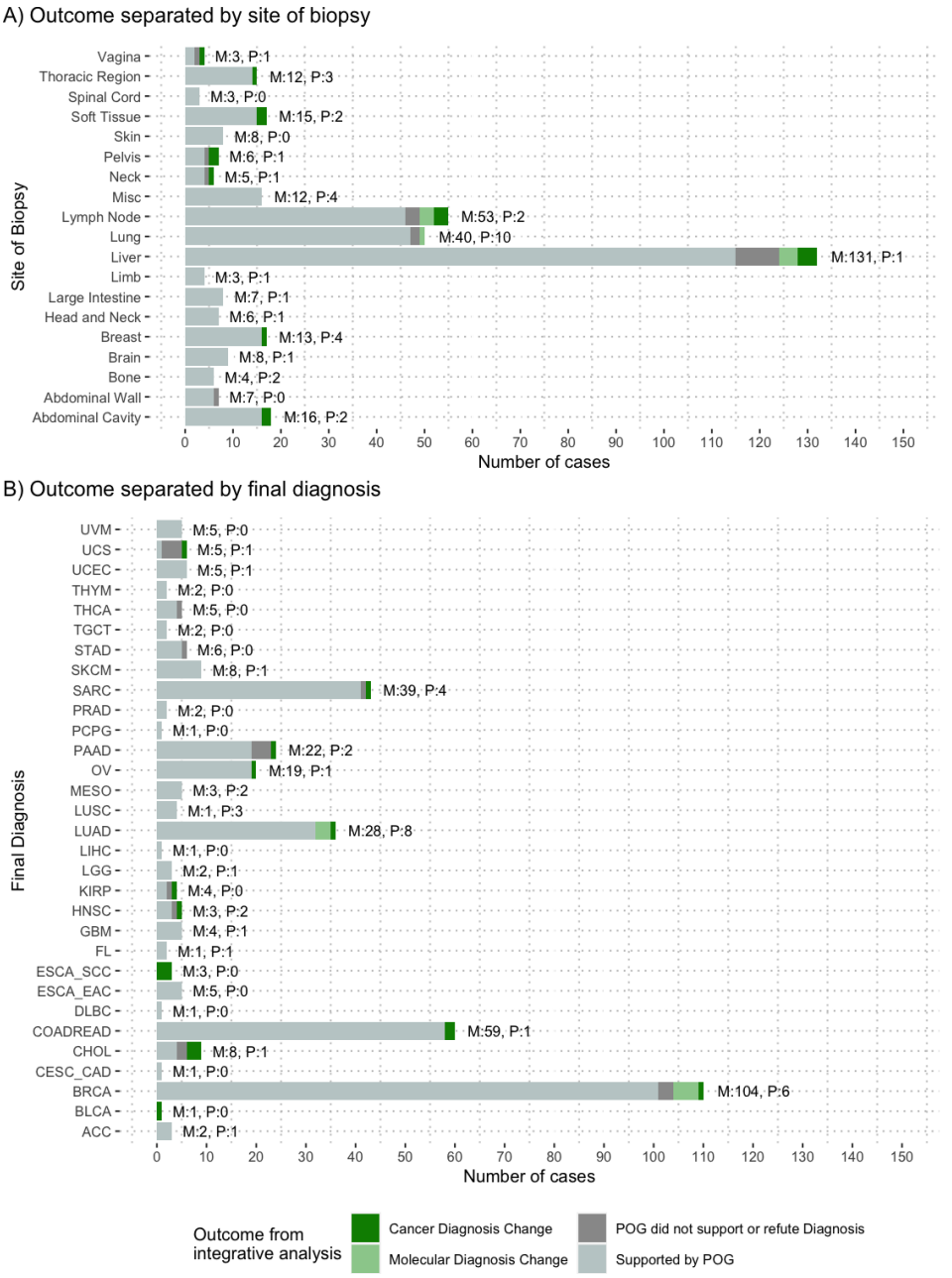

Figure 3.4: The outcome from genomic analysis is shown separated by A) the site of biopsy of the tumour, and B) the organ-system of origin of the cancer. M and P indicated the number of metastatic and primary/relapse samples respectively.

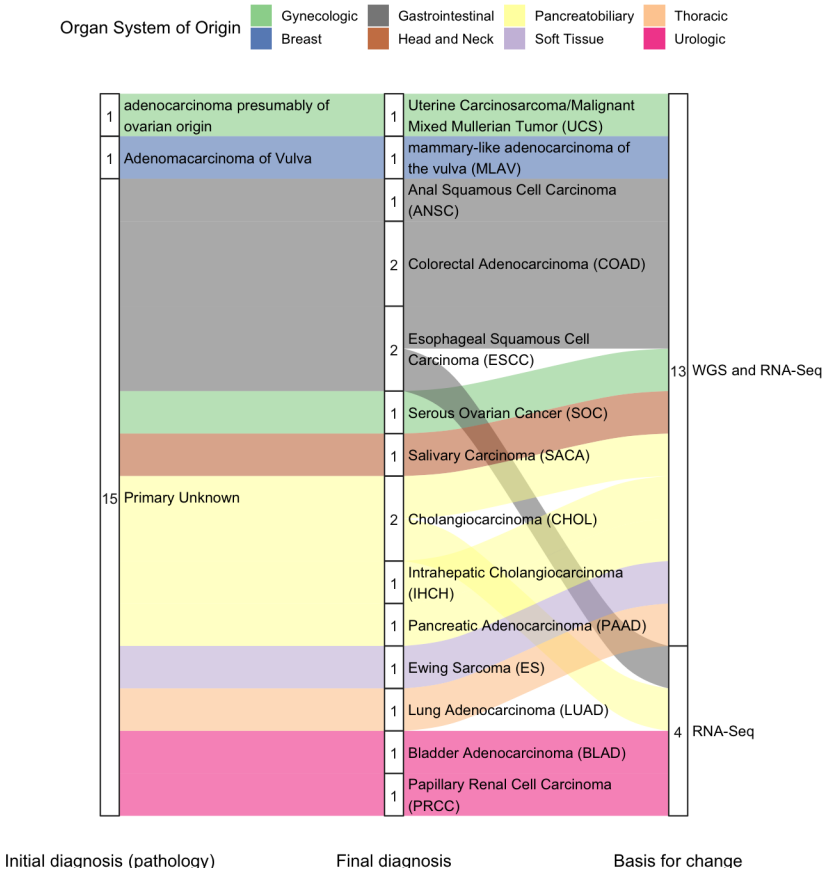

Figure 3.5: The final diagnoses for the 15 CUP cases and 2 cases with revised diagnosis are shown, along with the type of genomic data guiding each of the outcomes. WGS = Whole-genome sequencing.

comprehensive genomic analysis was able to pinpoint the site of origin of these cancers confidently.

Overall, within the highly frequent cancer types, breast cancer (N = 110) was mostly always correctly diagnosed. Cholangiocarcinomas, esophageal squamous cell carcinomas, and colorectal adenocarcinomas were among the cancer types most refractory to initial histopathology-based diagnosis. The type of genomic information used to determine a diagnosis for all 17 cases (including the 15 CUPs), and the resultant cancer diagnoses, are indicated in Figure 3.5.

Utility of SCOPE, an automated RNA-Seq based cancer diagnostic, for confirming diagnoses

The SCOPE algorithm was used retrospectively on these cases to assess the potential of automated tools in aligning diagnoses from RNA-Seq data in precision oncology workflows. SCOPE matched the final diagnosis in 273 of the 374 cases where an initial histopathology diagnosis was available (73%). When looking at tumour types with 10 or more cases, the SCOPE algorithm alone had the highest rate of success with breast carcinoma (BRCA, 87% accuracy, N = 96/110), ovarian carcinoma (OV, 84% accuracy, N = 16/19), lung adenocarcinoma (LUAD, 80% accuracy, N = 28/35), as opposed to pancreatic adenocarcinomas (PAAD, 22% accuracy, N = 5/23) which were missed most often by the method.

Eleven tumour types had an accuracy of <50% when using SCOPE (N = 46 samples), including pancreatic adenocarcinoma (PAAD, 22% accuracy), cholangiocarcinoma (CHOL, 33% accuracy), uterine carcinosarcoma (UCS, 33% accuracy), head and neck squamous cell carcinoma (HNSC, 25% accuracy), lower grade gliomas (LGG, 33% accuracy), uveal melanoma (UVM, 20% accuracy), esophageal adenocarcinoma (ESCA_EAC, 20% accuracy), and thyroid carcinoma (THCA, 40% accuracy), follicular lymphoma (FL, 0% accuracy). Among these cancers, we found that the incorrect predictions were typically mis-identification of the cancer as a histologically similar cancer (46% of mispredictions, N = 21/46, Table 3.1). For example, the two follicular lymphoma (FL) cases were predicted as diffuse large B-Cell lymphomas (DLBCL) from the National Cancer Institute's cohort (NCI_DLBCL) which contains some DLBCLs with FL-like features. Two head and neck malignancies were predicted as other types of squamous cell carcinomas instead.

SCOPE is impacted by low tumour content in liver biopsies

Eight of the 46 mis-predictions matched the site of biopsy instead. Particularly, in case of some liver biopsies it was observed that the highest prediction would be hepatocellular carcinoma (LIHC), and the second-highest prediction would be the correct cancer-type. Hypothesizing that these observations may be due to dilution of signal from the tumour itself, we decided to investigate the impact of biopsy site and tumour content on SCOPE's outcome.

We observed a significant association between biopsy site and SCOPE outcome (p-value 0.008, chi-square test, only biopsy sites with minimum

of 10 samples considered, N = 282/374). In biopsy sites with at least 10 samples, tumour content was found to be significantly associated with SCOPE prediction in liver biopsies only (N = 124/282, p.adjusted = 0.017, unpaired t-test). In Figure 3.6 we show the differences in outcome from SCOPE for the three most frequent sites of metastasis - the lymph node, lung, and liver.

SCOPE is a suitable automated method for providing a diagnosis for CUPs

In practice the effect of possible tissue contamination from the biopsy site could be accounted for during the case discussion meetings where molecular data was reviewed and SCOPE classifications became a useful tool in combination with orthogonal analyses, especially to establish cell lineage in CUPs (RNA-Seq contributed to resolution in all 15, where SCOPE's predictions were accurate in 8/15 cases, and in 3/15 additional cases matched the revised diagnosis confidently when accounting for biopsy site bias). When considering the performance of cancer types that had low accuracy (nine cancer types), SCOPE's predictions were not found to be significantly influenced by tumour content (Figure 3.7), reflecting a systemic inability in classifying these cancer types accurately.

Table 3.1: Classification outcome from SCOPE for the cancer cohorts.

Cancer type	Class size	SCOPE outcome matched			
		Final diagnosis	Cancer type from same organ system	Biopsy site	Other cancer type
BRCA	110	96 (87%)	0 (0%)	10 (9%)	4 (4%)
COADREAD	58	41 (71%)	7 (12%)	8 (14%)	2 (3%)
SARC	42	29 (69%)	2 (5%)	2 (5%)	9 (21%)
LUAD	35	28 (80%)	0 (0%)	0 (0%)	7 (20%)
PAAD	23	5 (22%)	6 (26%)	5 (22%)	7 (30%)
OV	19	16 (84%)	0 (0%)	0 (0%)	3 (16%)
SKCM	9	9 (100%)	0 (0%)	0 (0%)	0 (0%)
CHOL	6	2 (33%)	3 (50%)	0 (0%)	1 (17%)
STAD	6	5 (83%)	1 (17%)	0 (0%)	0 (0%)
UCEC	6	4 (67%)	0 (0%)	0 (0%)	2 (33%)
UCS	6	2 (33%)	1 (17%)	0 (0%)	3 (50%)
ESCA_EAC	5	1 (20%)	4 (80%)	0 (0%)	0 (0%)
GBM	5	4 (80%)	1 (20%)	0 (0%)	0 (0%)
MESO	5	4 (80%)	0 (0%)	0 (0%)	1 (20%)

Table 3.1: Classification outcome from SCOPE for the cancer cohorts. *(continued)*

Cancer type	Class size	SCOPE outcome matched			
		Final diagnosis	Cancer type from same organ system	Biopsy site	Other cancer type
THCA	5	2 (40%)	0 (0%)	1 (20%)	2 (40%)
UVM	5	1 (20%)	3 (60%)	1 (20%)	0 (0%)
HNSC	4	1 (25%)	2 (50%)	0 (0%)	1 (25%)
LUSC	4	4 (100%)	0 (0%)	0 (0%)	0 (0%)
ACC	3	3 (100%)	0 (0%)	0 (0%)	0 (0%)
KIRP	3	2 (67%)	0 (0%)	0 (0%)	1 (33%)
LGG	3	1 (33%)	0 (0%)	0 (0%)	2 (67%)
FL	2	0 (0%)	2 (100%)	0 (0%)	0 (0%)
PRAD	2	2 (100%)	0 (0%)	0 (0%)	0 (0%)
TGCT	2	1 (50%)	0 (0%)	0 (0%)	1 (50%)
THYM	2	2 (100%)	0 (0%)	0 (0%)	0 (0%)
CESC_CAD	1	0 (0%)	0 (0%)	0 (0%)	1 (100%)
DLBC	1	1 (100%)	0 (0%)	0 (0%)	0 (0%)
LIHC	1	1 (100%)	0 (0%)	0 (0%)	0 (0%)
PCPG	1	0 (0%)	0 (0%)	1 (100%)	0 (0%)
Total	***374***	***267***	***32***	***28***	***47***

Percent of cases in cancer type shown in brackets with corresponding outcome.

3.2.2.3 Impact of genomic analysis on clinical workflow

In 14 cases (3.6%), the genomic and transcriptomic analysis prompted downstream testing by pathologists for markers that had the potential to alter patient management. This group was composed of six lung adenocarcinomas, three breast carcinomas, three colorectal adenocarcinomas, one esophageal adenocarcinoma, and one anal squamous cell carcinoma. Lung adenocarcinomas had the highest fraction (N = 6/36, 17%) of cases where additional testing was suggested through this analysis. All other major tumour groups (class size > 10) had rates of less than 5%. Among all cases where the biomarker status was called under review as a consequence of the integrative analysis, 12 (86%) had HER2 overexpression (Figure 3.3). The remaining two cases included the identification of a ROS1 fusion in a lung adenocarcinoma, at the time when screening for ROS1 was not

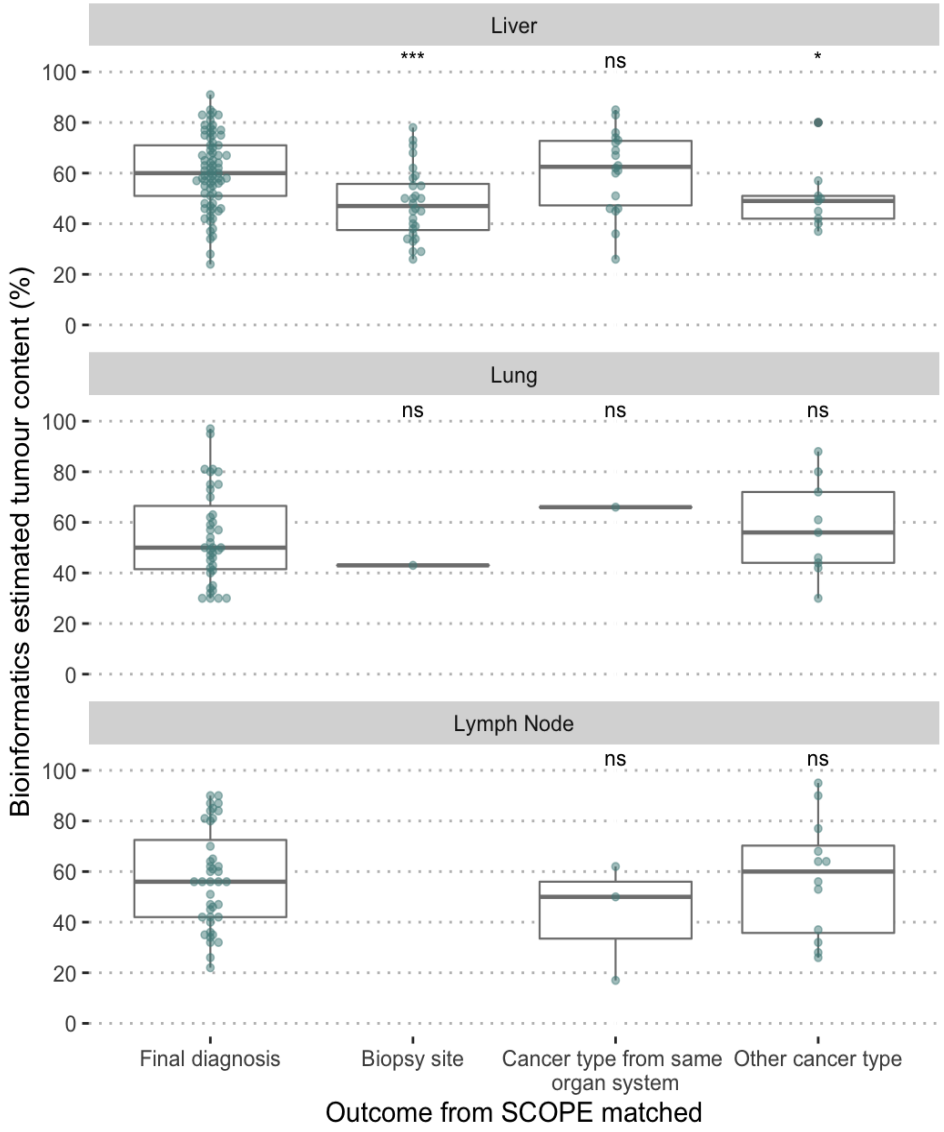

Figure 3.6: Impact of tumour content on the ability of RNA-Seq to provide the correct putative diagnosis in the POG cohort. The majority of samples arose from 3 biopsy sites - lymph node, lung, and liver, indicated in each of the panels. Wilcox test for significance between SCOPE outcome matching final diagnosis, versus each of the other categories: * p =< 0.05; ** p =< 0.01; *** p =< 0.001; ns p > 0.05

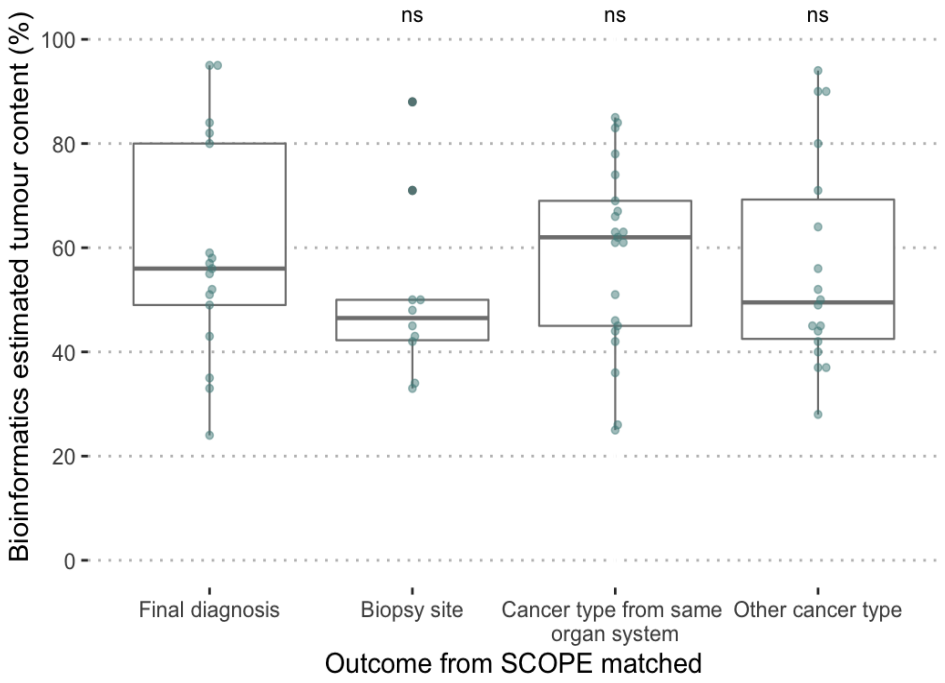

Figure 3.7: Impact of tumour content on the ability of RNA-Seq to provide the correct putative diagnosis in the POG cohort, agnostic of biopsy site. Wilcox test for significance between SCOPE outcome matching final diagnosis, versus each of the other categories: * p =< 0.05; ** p =< 0.01; *** p =< 0.001; ns p > 0.05

standard of care yet, and the detection of HPV integration in the DNA for an anal squamous cell carcinoma.

3.3 Discussion

In this study we assessed how whole-genome and RNA sequencing impacts tumour diagnosis and biomarker assessment in the current clinical laboratory environment. We studied this within an ongoing clinical trial that leveraged sequencing data to profile advanced, treatment-resistant cancers and suggest alternative, targeted lines of therapy. We showed that this approach has particular benefit for diagnosing and managing cancers of unknown primary. At the clinical level, integrative genomic analysis did not provide additional routine clinical guidance with the exception of rare genomic events (ROS1 fusion, HPV integration) and HER2 marker status.

Integrative genomic analysis is important for diagnosing and assessing complex presentations like CUPs

Our capacity to generate high-resolution multi-omic data from tumour tissues has enhanced to the point where genomic and transcriptomic analysis can be considered for integration in routine clinical oncology. This precision oncology approach has shown itself to be extremely valuable for understanding disease onset and advancement, determining clinically valuable molecular subtypes, and identifying targeted treatment hypotheses that have potential for clinical translation [30, 35, 146, 169]. It also has immense potential for resolving CUPs - indeed, in our series it was able to identify a tissue of origin in all our CUP cases. These cases formed a significant proportion of the analyzed cases (3.9%), and based on our analysis, RNA-Seq was important in determining a putative primary in all 15 of these cases. It is important to note that due to the often poorly differentiated and advanced nature of these cases, there were no gold standard diagnoses. Although post-POG testing using IHC markers could be done, some uncertainty about the tissue of origin could still remain. Nevertheless, knowing the site of origin of a cancer facilitates treatment decisions, helps the patient understand their disease, and provides some closure to those impacted.

Gene expression profiling for CUP diagnosis is not a novel approach. Commercial assays have been developed numerous times but struggled to reach the clinic. The ESMO clinical guideline for CUP diagnosis and

treatment does not recommend gene expression profiling as an ancillary diagnostic test, but earlier studies in this domain agree that gene expression profiles are a useful prognostic marker [61]. The microarray platform used in the previous studies, however, are eclipsed by the dynamic range and quantifiable measurements obtained from RNA-Seq. These studies have also discounted the integration of gene expression and mutation analysis together for resolving CUPs. Our data shows that mutational analysis was an added value in determining tissue of origin for most CUPs. The clinical value in CUPs from whole transcriptome RNA-Seq and mutation analysis is unclear, given the absence of approved therapies that match potentially actionable changes that this process uncovers, but should be evaluated further.

In addition to considering the gene expression measurements and mutation profiles by themselves, we explored the utility of SCOPE, an automated method that could provide a diagnosis from the RNA-Seq data without requiring an expert bioinformatician as an intermediary. While the method was quite successful for a variety of tumours, it failed to perform well on certain challenging gastric malignancies like pancreatic adenocarcinomas and esophageal adenocarcinomas. We were able to determine that in these cases the method was often confounded by low tumour content, and upon accounting for the effect of the biopsy site, many of the predictions from the method reflected the ground truth.

Integrative genomic analysis provides improved testing options in patients already assayed through conventional histopathology tests

In this series of cases, integrative genomic analysis had the potential to improve management options over the conventional workup in 6% of cases, most of which consisted of identifying false-positive and false-negative findings from conventional assays. The contribution of genomics to HER2 status determination is particularly valuable considering its importance as a prognostic factor, and given the high prevalence of false-positives and false-negatives in IHC-based screening of HER2 status in the clinic [66]. In general, by identifying potentially actionable targets, the analysis uncovered non-traditional routes for cancer management (in the form of clinical trials for unapproved therapies, off-label drugs). Outcome analysis and assessment of off-label treatment options was outside of the scope of our study.

As expected, when considering clinically validated treatment options only,

our results suggest that integrative genomic analysis did not offer up any additional insights to supplement the current standard of evidence in cancer management. While these findings may point towards the efficiency of current testing guidelines for well-defined tumour histologies, based on our experience and observations within this cohort, we believe they underscore the gap between the small number of therapeutic options for which there is level 1 evidence to support usage, and the huge number of potential treatment targets determined through integrative genomic analysis for which there is no established efficacy data yet.

It is quite likely that as there are advancements in our understanding of oncologic mechanisms at the level of an individual, we will reach the point where targeted panels and single-target gene assays will prove insufficient to evaluate the sheer number of evidence-based actionable targets. In that scenario, whole-genome and transcriptome sequencing may become the most effective test, accompanied by automated analysis methods like SCOPE that complement this type of data. Our findings show promise in this direction, with 2% of our genomically analyzed cases being redefined histopathologically and 3.6% of cases impacting clinical workflow directly through additional testing and revised patient management. Large-scale genomic and transcriptomic clinical trials, like the one described here and others conducted across the world [30, 169], will be an essential part of the progression of whole-genome and transcriptome sequencing into a clinical standard.

3.4 Conclusion

Our experience with whole-genome and transcriptome sequencing as part of a clinical trial has defined its strengths and limitations in the area of cancer diagnosis and resultant impact on clinical management. As previously demonstrated by other projects like this one, integrative genomic analysis is an excelled hypothesis generating tool. It also has extreme value in identifying a putative primary for CUPs, guiding the management of these patients. What we find in addition to these previous studies is that in a noteworthy fraction of cases whole-genome and transcriptome sequencing also revised the findings of ancillary histopathology tests (FISH, IHC), revised patient management, motivated additional downstream testing through histopathology, and realigned the molecular diagnosis for well-established entities like breast cancer subtypes. Translation of this

approach in the clinic requires evolution in the predictive testing landscape to the point where integrative analyses like these add significantly more benefit to the current histopathology based approach.

There were several rare cancers that were excluded from this case-series as the representative genomic datasets most suitable for their analysis are currently unavailable. The study of rare and complex cancers based on a reference dataset of cancers is a challenge in itself, with no evident solutions. The use of automated tools like SCOPE can help find the closest cancer-type with available data, as has been shown in a published case study discussed later in Chapter 5. Aligning those findings to a clinical outcome still requires the generation and integrative analysis of these rare cancer types. The development and validation of SCOPE is now discussed in Chapter 4.

Chapter 4

Development and validation of SCOPE - supervised cancer origin prediction using expression

Identification of the site of origin of a tumour in a patient is currently used to guide cancer treatment. It also informs any subsequent analysis through alignment with relevant tumour literature and expected molecular background. Currently, established pathology approaches are used for cancer diagnosis and are considered the gold standard. In most cases, this includes morphology- and histochemistry- guided approaches which also determine eligibility to drug regimens and clinical trials. Modern pathology is a process of sequential exclusion and prioritization across candidate diagnoses, but an exhaustive search is rendered unpossible by limited tissue and diagnostic stains.

The efficiency of cancer diagnostics can be vastly improved if an automated method can be developed to approach this task with some knowledge of cancer biology, similar to a pathologist. A machine-learning method trained across diverse tumours and normal tissues will learn what characterizes each cancer, rather than its tissue site. Training on high-resolution molecular data will allow it to discover such tissue- and tumour- specific biological patterns from the entire transcriptome.

The use of gene expression data has outperformed traditional pathology workflows for cancer diagnosis in several landmark studies [3, 127, 132, 152]. Recent studies have also shown that transcriptome-wide profiling offers greater information about tumours than microarrays [74, 186], with utility in precision oncology [30, 96]. We can therefore use high-resolution transcriptomic data as an orthogonal approach to improve diagnostic

accuracy in many cancers [58, 146]. While analyzing such high-dimensional data within a diagnostic workflow is not manually feasible, machine-learning methods can be trained to do so instead.

Here we describe the methods underlying Supervised Cancer Origin Prediction using Expression (SCOPE), a set of neural networks that use the transcriptome to identify the closest match for a tumour from amongst 40 cancer types and 26 normal tissues, and which was used as an ancillary tool for cancer diagnosis in the previous Chapter. We account for the influence of differentiation and biopsy site by including normal tissues (classes) from TCGA in our training dataset [164]. We determine genes weighted heavily for decision-making and show that SCOPE is able to prioritize genes relevant to each class without any prior information.

SCOPE is trained devoid of any feature selection, and is able to achieve high precision and recall within The Cancer Genome Atlas (TCGA) primary cancer and adjacent normal cohorts. Our results suggest that using the entire transcriptome in a pan-cancer classification approach performs better than using feature selection. We also validate the classifier on an independent cohort of primary mesotheliomas, where we achieve classification accuracy of up to 100%. Lastly, we show high performance when using this method in external use cases, to identify the site of origin of (a) treatment-resistant metastatic cancers, biopsied from their site of origin, (b) cancers that are refractory to standard histopathology techniques for diagnosis, and (c) treatment-resistant metastatic cancers, biopsied from their site of metastasis. Another valuable application of this method lies in providing an objective, orthogonal source of differential diagnoses in cancers that are refractory to standard diagnostic practice.

4.1 Background

Pathology protocols for cancer diagnosis work best when the tissue specimens display high quality and recognizable histological features in a substantial number of cells. Generic histological features alone are often not sufficient to subtype a tumour, hence the confirmation of cell-of-origin - typically via IHC - remains the bedrock of modern pathology practice [214]. Therefore, diagnosis can become a challenging task of tiered, single-plex IHC analyses for lineage-specific proteins, iteratively evaluating the next likely diagnostic candidates. Limited tissue availability and a limited list

of unambiguous IHC antibodies restrict the extent of validation work-ups. Inter-observer variability in pathology based diagnoses, sample related challenges, and limited tissue samples for immunohistochemical analyses can further restrict the ability to identify the underlying pathology of a biopsy sample [209]. This is especially true for metastatic and poorly differentiated (high-grade) cancers.

Misdiagnosis rates for metastases in clinical practice can range between 45-94% in the event of challenging presentation (suboptimal sample quality, histologic similarity between tissues, poor differentiation) [128]. This is concerning since metastases can form up to 60% of distant recurrences and cause upwards of 90% of cancer associated deaths for cancers detected in the gastrointestinal tract and across certain gynecological cancers [33, 144, 209]. Biomarker conversion in metastases can confound diagnosis from IHC and from biomarker based assays [186]. The site of biopsy is yet another confounder, particularly in case of the liver [169]. Previous work utilizing expression microarrays has indicated that the microenvironment can contribute to the enrichment of hepatic genes' expression in liver metastases, confounding an accurate diagnosis [37]. These issues are magnified in CUPs, where developing specific diagnostic protocols remains a challenge for pathology [10, 132, 206].

Inclusion of rare cancer types and providing a refined diagnosis remain challenges for current computational diagnostics. In order to optimize training, rare cancer types are often excluded, and geographically proximal cancers are merged. This inevitably leads to loss of granularity and limited scope in the application of the models trained [55, 132]. Performance is evaluated on the test set, which can either be held-out from the initial cohort, or preferably (but rarely) a cohort of samples generated and processed at different centers.

RNA-Seq has largely replaced microarrays for transcriptome-wide profiling. However, the current repertoire of diagnostics does not draw upon the high dynamic range and comprehensive coverage provided by RNA-Seq [30, 224]. Large-scale sequencing projects (The Cancer Genome Atlas, TCGA [215], International Cancer Genome Consortium, ICGC [43]) have amassed RNA-Seq data from upwards of 10,000 patients with untreated primary cancers. This provides unprecedented opportunity to apply machine-learning approaches to improve the classification of all cancer types. With the availability of high-performance computing systems, now it is also possible to train models using information about the transcriptional

status of all genes.

4.2 Methods

4.2.1 Training data

Multi-platform RNA-Seq data was obtained from TCGA (multi-platform - Illumina Hi-Seq 2000 and Genome Analyzer II, processed with TCGA RNA-Seq v2 RSEM processing pipeline), the National Cancer Institute (NCI) non-Hodgkin lymphoma dataset [135] (sequenced with Illumina Genome Analyzer II, median normalized), and non- cell-line primary tumour data from the Terry Fox Research Institute's Glioblastoma Multiforme (GBM) project. 2 in-house cancer cohorts, adult medulloblastoma (MB-Adult) and follicular lymphoma (FL), further supplemented this dataset (sequenced with Illumina HiSeq 2500). Colon and rectum adenocarcinomas from TCGA were combined into a single cohort (COADREAD) due to their geographical proximity in primary lesions, supported by findings from our initial quality control that showed insufficient decomposition of these two cancer types based on their transcriptomic data. The TCGA RNA-Seq libraries were prepared by various different sequencing centers, but to facilitate harmonization across samples, the TCGA RNASeq v2 RSEM processing pipeline aligned all RNA-Seq reads in an unstranded manner.

Table 4.1: Cancer types used for training, with abbreviations referenced in text.

Code Name	Full Name	Normal	Tumour
ACC	Adrenocortical Carcinoma		79
BLCA	Urothelial Bladder Carcinoma	19	408
BRCA	Breast Ductal Carcinoma	113	1095
CESC_CAD	Cervical and Endocervical Adenocarcinoma	3	47
CESC_SCC	Cervical Squamous Cell Carcinoma	6	257
CHOL	Cholangiocarcinoma	27	36
COADREAD	Colorectal Adenocarcinoma	51	372
DLBC	Diffuse Large B-Cell Lymphoma		48
DLBC_BM	DLBCL Blood/Bone Marrow		11

Table 4.1: Cancer types used for training, with abbreviations referenced in text. *(continued)*

Code Name	Full Name	Normal	Tumour
ESCA	Esophageal carcinoma	3	15
ESCA_EAC	Esophageal Adenocarcinoma	24	79
ESCA_SCC	Esophageal Squamous Cell Carcinoma	6	90
FL	Follicular Lymphoma		50
GBM	Glioblastoma Multiforme	15	161
HNSC	Head and Neck Squamous Cell Carcinoma	44	520
KICH	Kidney Chromophobe Carcinoma	25	66
KIRC	Clear Cell Kidney Carcinoma	72	533
KIRP	Papillary Kidney Carcinoma	32	290
LAML	Acute Myeloid Leukemia		173
LGG	Lower Grade Glioma		516
LIHC	Liver Hepatocellular Carcinoma	50	371
LUAD	Lung Adenocarcinoma	59	515
LUSC	Lung Squamous Cell Carcinoma	50	501
MB-Adult	Adult Medulloblastoma		143
MESO	Mesothelioma		87
NCI_GPH_DLBCL	Diffuse Large B-Cell Lymphoma (NCI cohort)		111
OV	Ovarian Serous Cystadenocarcinoma		305
PAAD	Pancreatic Ductal Adenocarcinoma	12	178
PCPG	Paraganglioma & Pheochromocytoma	9	179
PRAD	Prostate Adenocarcinoma	52	497
SARC	Sarcoma	6	259
SKCM	Cutaneous Melanoma	3	469
STAD	Stomach Adenocarcinoma	35	415
TFRI_GBM_NCL	Glioblastoma Multiforme (TFRI cohort)		52
TGCT	Testicular Germ Cell Cancer		150
THCA	Thyroid Carcinoma	59	505
THYM	Thymoma	6	120
UCEC	Uterine Corpus Endometrial Carcinoma	24	177
UCS	Uterine Carcinosarcoma		57

Table 4.1: Cancer types used for training, with abbreviations referenced in text. *(continued)*

Code Name	Full Name	Normal	Tumour
UVM	Uveal Melanoma		80

Abbreviations:
NCI - National Cancer Institute; TFRI - Terry Fox Research Institute

This resulted in a dataset of 10,822 transcriptomes spanning 40 different untreated primary tumour types and 26 adjacent normal tissue types (66 'classes'), with individual class sizes ranging from three to 1095 samples. No feature selection was done on the consolidated set of transcriptomes besides filtering for (a) genes with a recorded RPKM value in every sample (N = 21,220), and (b) genes that overlapped with available annotations for our independent test sets (N = 17,688). This resulted in a set of 10,822 samples, spanning 66 different tumour and adjacent normal classes, and with each sample represented by 17,688 distinct median normalized gene RPKM values. Table 4.1 shows the annotations used, following TCGA nomenclature. The table also shows the number of training samples available for each cancer type.

4.2.2 Test data

The trained ensemble of neural networks was validated retrospectively on a set of primary mesotheliomas (MESO) published as part of an independent study by Genentech. Test sets for adult metastatic disease and 15 cancers of unknown primary were obtained retrospectively from the Personalized OncoGenomics clinical trial at BC Cancer [108]. The attributes of these datasets are as follows.

Genentech primary mesothelioma dataset

Mesothelioma is a rare and aggressive cancer arising in the linings of lung, abdomen, or the heart. An independent set of 211 adult primary untreated mesotheliomas cancers was obtained from the Genentech Mesothelioma cohort [20]. 126 of these samples are classic epithelioid mesotheliomas, while 85 are sarcomatoid variants. As the training set of mesotheliomas was histologically classic epithelioid mesotheliomas, testing was as follows: For the epithelioid mesotheliomas, we tested whether the classification was

exclusively for mesothelioma. For the biphasic and sarcomatoid variants, we tested whether the classification was split between sarcomas and mesotheliomas, as would be expected based on mixed histology of the samples. The Genentech Mesothelioma dataset has 211 transcriptomes from untreated, primary lung biopsies of mesothelioma, spanning 4 distinct molecular subtypes of mesothelioma – sarcomatoid (N = 29), epithelioid (N = 54), biphasic-epithelioid (N = 72), and biphasic-sarcomatoid (N = 56). The RNA-seq libraries were prepared using TruSeq RNA Sample Preparation kit (unstranded, polyA+) from Illumina, and sequenced on the HiSeq 2500 (~66 million paired-end reads per sample) [20].

Personalized OncoGenomics (POG) trial metastatic disease dataset

Cases of adult metastatic disease were selected based on the following criteria – (a) a primary of origin was identified, based on a joint consideration of clinical/pathology/genomic data, and (b) cDNA libraries prepared from the biopsy sample passed in-house quality control. Based on these criteria, we identified 201 samples spanning 26 different cancer types, summarized in Table 4.2 shows the annotations used, following TCGA nomenclature. The table also shows the number of training samples available for each cancer type. 168 of the 201 metastases were biopsied from their site of metastasis (24 cancer types), and the remaining 33 from their site of origin (12 cancer types).

Table 4.2: Breakdown of cancer types in the external metastatic cohort.

Code Name	Organ-System	Full Name	Count
ACC	Endocrine	Adrenocortical Carcinoma	2
BRCA	Breast	Breast Ductal Carcinoma	70
CESC_CAD	Gynecologic	Cervical and Endocervical Adenocarcinoma	1
CHOL	Gastrointestinal	Cholangiocarcinoma	5
COADREAD	Gastrointestinal	Colorectal Adenocarcinoma	22
SKCM	Skin	Cutaneous Melanoma	3
DLBC	Hematologic	Diffuse Large B-Cell Lymphoma	1
ESCA_EAC	Gastrointestinal	Esophageal Adenocarcinoma	2
ESCA_SCC	Gastrointestinal	Esophageal Squamous Cell Carcinoma	4
FL	Hematologic	Follicular Lymphoma	1
GBM	CNS	Glioblastoma Multiforme	4
LIHC	Gastrointestinal	Liver Hepatocellular Carcinoma	1

Table 4.2: Breakdown of cancer types in the external metastatic cohort. *(continued)*

Code Name	Organ-System	Full Name	Count
LGG	CNS	Lower Grade Glioma	2
LUAD	Thoracic	Lung Adenocarcinoma	18
LUSC	Thoracic	Lung Squamous Cell Carcinoma	1
MESO	Thoracic	Mesothelioma	5
OV	Gynecologic	Ovarian Serous Cystadenocarcinoma	7
PAAD	Gastrointestinal	Pancreatic Ductal Adenocarcinoma	11
KIRP	Urologic	Papillary Kidney Carcinoma	2
PRAD	Urologic	Prostate Adenocarcinoma	1
SARC	Soft Tissue	Sarcoma	23
STAD	Gastrointestinal	Stomach Adenocarcinoma	3
TGCT	Urologic	Testicular Germ Cell Cancer	1
THYM	Hematologic	Thymoma	1
UCS	Gynecologic	Uterine Carcinosarcoma	5
UCEC	Gynecologic	Uterine Corpus Endometrial Carcinoma	6

Abbreviations: CNS - Central Nervous System

Cancers of unknown primary

Additionally, the POG cohort contained 15 cases where the primary site of origin could not be determined by initial pathology analysis. Genomic and transcriptomic analysis as part of the POG project determined the corresponding cancer type for 15 of these cases, which was used as gold standard for assessing the prediction from the classifier. The classification was performed retroactively after the closest suitable cancer type had been determined based on detailed pathway-level and genomic analysis of the cancer.

4.2.3 Model training

For the initial selection of the optimal classification algorithm, gene RPKMs were used as input. Support vector machines, random forests, extra trees, and a fully connected neural network were compared. Five-cross validation with grid search was used to identify the best parameters for each of these

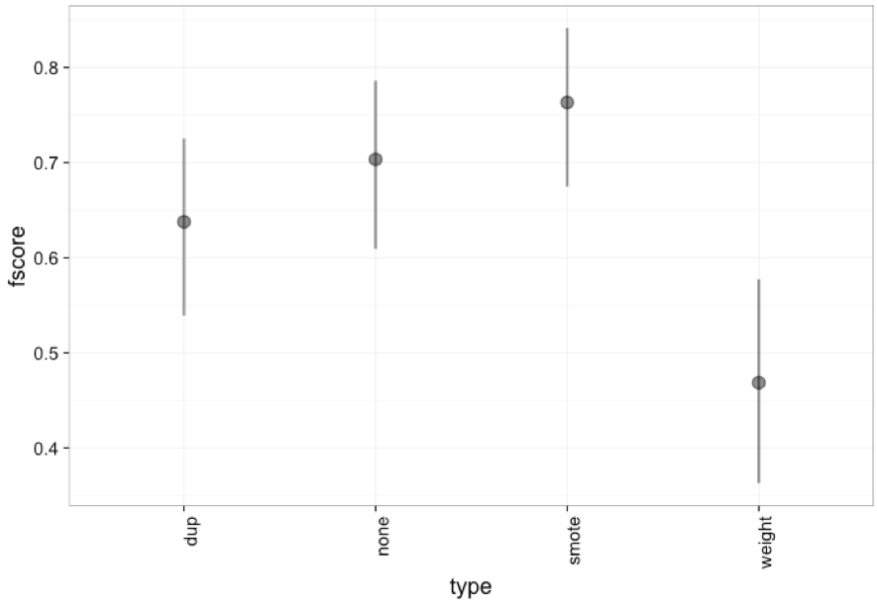

Figure 4.1: Performance of SMOTE as compared to other class expansion methods. Cross-validation results on the TCGA training dataset are shown. Abbreviations: dup - duplication of samples in small classes, none - no class expansion applied, weight - inverse cost for misclassification of smaller classes during training.

algorithms. The trained models were subsequently tested on the one-fifth held-out set.

Because the other ensemble models (random forest, extra trees) had near-equivalent 5-cross validation results with the neural network during training, we evaluated the utility of extending the neural network model. An ensemble was developed by training multiple neural networks with different linear transformations of the data. The resultant classifier (SCOPE) contained five neural networks. For one of these neural networks, we synthetically generated additional samples to expand the rarer classes during training (SMOTE - Synthetic Minority Oversampling Technique [28]). We compared SMOTE with other commonly used class expansion strategies in machine learning and found it to outperform the others (Figure 4.1). The differences in the five networks are described in detail in Table 4.3.

Training data was randomly split up into 4/5th 'model training' data, and 1/5th 'held-out test' data. The training and test splits maintained relative class frequencies. In classes with less than five samples (six classes, all adjacent normal), one sample was randomly assigned to the held-out test set, and the remaining samples were kept in the model training set. All models discussed in this paper were trained on the 4/5th model training data, with the held-out test set used as the first external validation of performance of the fully trained models. Stratified 5-fold Cross Validation (5-CV) was used for hyperparameter selection for each algorithm. The StratifiedKFold function in the scikit-learn package in Python was used to generate class-balanced CV folds [151].

4.2.3.1 Data normalization and feature selection

Data transformation and feature selection

Technical artefacts in the training data can cause over-fitting while training a classifier. This results in a classifier that performs quite well on the training data, but does not generalize to samples that it has not seen during training. Two main approaches to overcome over-fitting prior to training are (a) data pre-processing, and (b) feature selection. Data pre-processing is generally done by re-scaling the input data to fall within a certain range of values, or by forcing it to follow a certain distribution (ex. normal distribution for expression data). Feature selection can be done by several methods, but usually, a subset of features that are critical to distinguishing the training cohorts are selected using feature reduction methods like Principal Component Analysis (PCA) or pair-wise analysis of variance (ANOVA). This subset is then used to train the classifier.

We assessed the utility of data transformation and feature selection in improving the best performing model in the previous step, the shallow neural network. To this end, various scaling and data transformation methods, namely minmax scaling, L2 norm scaling, and rank normalization (average), were assessed separately. The performance of each approach was assessed by stratified 5-CV.

Subsequent to selection of the optimal algorithm as described, we tested the utility of feature selection in improving classification performance. Guided by previous work [85], we used pair-wise ANOVA of log-transformed training data to identify a subset of 3,000 genes that are statistically

significant at discriminating the training classes. We also trained a classifier using COSMIC's list of 552 genes harboring somatic mutations [63]. Neural network architectures optimal for each input space were identified using grid search across parameters, and trained with 5-CV for comparison.

Class expansion using Synthetic Minority Oversampling

A supervised machine learning based classifier works by seeing multiple different samples representing each cancer/tissue type and steadily learning which genes (features) are most valuable in identifying each type of interest. A common problem with this approach is that a classifier can sometimes fail to appreciate the features that characterize the smaller cancer/tissue types. This class imbalance can be overcome by pre-processing the training set in specific ways – by duplicating some of the samples in the smaller class(es), by 'punishing' the classifier more for making a mistake with the smaller classes, or by supplementing the smaller classes with synthetic samples. One such method for adding synthetic samples to smaller classes is Synthetic Minority Oversampling (SMOTE).

We trained and assessed the performance of the RPKM-based neural network classifier method using three different class expansion approaches, (a) duplicating samples randomly in the smaller cohorts to inflate their total sample size to the largest class, (b) adding an inverse weight factor for mis-classification of smaller classes (i.e. making it more expensive for the classifier to mislabel a sample from a smaller class during training), (c) adding synthetic samples using SMOTE, and compared these three approaches to (d) doing no class expansion. Duplicated/synthetic samples were only added to the training folds, so that the cross-validation test fold always only contained non-synthetic samples that were absent in the training folds. The synthetic sampling algorithm of SMOTE was adopted from Chawla et al [28].

4.2.3.2 Metrics for evaluation

Since the training cohorts have a wide range of representative samples (N = 3-1025), using accuracy as a metric of performance of the classifiers would not necessarily reflect the ability of the classifier to discriminate between all 66 output classes. Precision, recall, and F1 score were used to evaluate models and demonstrate their performance. Aggregate precision and F1 scores, where reported in text, are accompanied by 95% CIs. Precision is

defined as (true-positives)/(true-positives + false-positives), and intuitively represents the classifier's ability to distinguish between positive and negative cases. Recall is defined as (true-positives)/(true-positives + false-negatives), and intuitively represents the classifier's ability to correctly identify all positive cases. The F1 score is the harmonic mean of the precision and recall. These metrics are calculated for each individual class, and the mean reported as the cohort metric. Accuracy is reported as (true-positives + true-negatives)/(total cases), and is calculated for the entire cohort.

A paired χ^2 test for association between prediction accuracy and tumour content was performed on the metastatic test cohort, with the null hypothesis being, "the classification accuracy of SCOPE is independent of tumour content." Tumour content was determined by pathology analysis. Students paired t-test was used to test the association between prediction accuracy and confidence score (null hypothesis: no correlation exists between prediction accuracy and confidence score). The level of significance was 2-sided $P = 0.05$ for all tests of association. Pearson correlation was used to evaluate association between class-specific accuracy and training class size. Statistical tests were conducted using the base statistics package available in R (R version 3.5.0; RStudio API version 1.1.442; R Project for Statistical Computing).

For a given input, the ensemble generates a pooled confidence score for each of the 66 output classes. Predicted classes are jointly ordered by the confidence score and number of machines in agreement. This max vote-pooling method was used to obtain a quantitative confidence score for each category. This confidence score was taken as a proxy for differential diagnosis when assessing metastatic samples. Thus, in the event that the prediction from the ensemble classifier was split between different cancer types, the correctness of the prediction was assessed by comparing the diagnosed cancer type against the pool of confident predictions.

4.2.4 Algorithmic model selection

For the initial selection of the optimal classification algorithm, RPKMs were used as input. Support Vector Machines (SVM), Random Forests (RF), Extra Trees (ET), and a fully-connected neural network (NN) were compared. 5-cross validation (5-CV) with grid-search was used to identify the best parameters for each of these algorithms. The trained models were subsequently tested on the held-out set of 1/5th of the total samples.

A shallow neural network (with a hidden layer of 17,000 genes, tanh activation, learning rate = 0.001, L2 regularization cost = 0.0001), was found to be the top performing model on the held-out test set. As the other ensemble models (RF, ET) had near-equivalent 5-CV results with the NN during training, we evaluated the utility of extending the NN model. An ensemble was developed by training multiple neural networks with different linear transformations of the data. The resultant classifier (SCOPE) contained five neural networks. The F1-Score was used as the main metric of assessment, to account for class imbalances in the training and test sets.

4.2.5 Ensemble selection

Based on our observations from Section 4.2.3, we built an ensemble of neural networks that used both RPKM- and rank normalized- training data as input across varying architectures and regularizations. This extended our selected classification model to include five additional neural network architectures. The additional neural networks were selected using the 5-CV approach discussed already. Furthermore, these neural networks were evaluated on the held-out test set (1/5th of the training data) which was set aside prior to cross-validation, and networks (machines) that had a performance at par/greater than the RPKM-only, transcriptome-wide neural network on the held-out test set were used to build an ensemble classifier. The resultant ensemble classifier contained five neural networks, with each 'neural network machine' in the ensemble assigning a confidence score (as represented by class probability) for each output class.

For the assessment of metastatic samples biopsied from the site of metastasis, the confidence of prediction was taken into account as an evaluation of differential diagnosis. In the event that the prediction from the classifier was split between different cancer types, the 'correctness' of the prediction was assessed by comparing the diagnosed cancer type against the split pool of predictions. This cohort represents advanced, treatment-resistant metastatic disease that has undergone multiple rounds of selective pressures from its local environment and chemotherapy regimens. We included a previously published baseline linear comparator for our classification method on this dataset, in order to identify the lower bound for transcriptomics-based characterization of these cases using primary cancer data [72].

Table 4.3: Architecture, identifying names, and additional information for each neural network in the SCOPE ensemble.

Model.name	Architecture	Data.pre.processing	Additional.rules
none17k	17688 x 17000 x 66	None (RPKM)	None
none17kdropout	17688 x 17000 x 17000 x 66	None (RPKM)	Dropout (10%) input in training
smotenone17k	17688 x 17000 x 66	None (RPKM) with SMOTE samples in training	None
rm500	17688 x 500 x 66	Rank norm + minmax(0,1) scaling	None
rm500dropout	17688 x 500 x 500 x 66	Rank norm + minmax(0,1)	Dropout (10%) input in training

4.2.6 Feature weights analysis for neural network

Following training, the weights and biases for each layer were extracted using the lasagne.layers.*get_all_param_values*(network) function. Subsequently, following the rules of weight propagation in fully connected neural networks, a forward multiplication loop was evaluated, resulting in a matrix of dimensions [Number of genes, Number of output categories]. For each output category, the resultant network weights were sorted, and the top-100 genes with the highest weights for the class were saved. This was done over 5-cross validation models for each neural network, resulting in 25 lists of top-100 genes. For a given neural network, genes found to be top-ranked in at least three out of five CV folds were identified. Subsequently, for each category, the NN-specific top genes were filtered for occurrence in at least 3/5 neural networks, resulting in a set of important genes for each cancer type and normal tissue in the classification categories (Appendix Table 1).

4.3 Results

A total of 10,688 adult patient samples representing 40 untreated primary tumour types and 26 adjacent-normal tissues were used for training. Among the training data set, 5,157 of 10,244 (50.3%) were male and the mean (SD) age was 58.9 (14.5) years. Testing was performed on 211 patients with untreated primary mesothelioma (173 [82.0%] male; mean [SD] age 64.5 [11.3] years); 201 patients with treatment-resistant cancers (141 [70.1%]

female; mean [SD] age, 55.6 [12.9] years); and 15 patients with cancers of unknown primary of origin; among the treatment-resistant cancers, 168 were metastatic, and 33 were the primary presentation. In our study, SCOPE achieved 97% accuracy and a macro F1-score of 0.92 on the 2,780 cases in the TCGA held-out set.

The transcriptome had improved performance over the COSMIC cancer gene set and ANOVA-selected genes (Figure 4.2 A). The single neural network outperformed other machine-learning algorithms (Figure 4.2 B). For 46 out of the 66 classes, 80-100% of the samples in each class were correctly classified (Figure 4.2 C). We found that seven classes were refractory to appropriate classification, among which three were cancer types (esophageal carcinomas and adenocarcinomas and cervical cancers), and all seven had fewer than 50 training examples (class size range, 3-50). On closer investigation of the five neural networks in the ensemble, we found that the neural network trained with SMOTE-supplemented training examples showed improved performance on smaller classes compared with the other four (Figure 4.3).

The performance of the model on the held-out set was better than that quantified through cross-validation on the training dataset (Figure 4.2 A). This is because of the difference in the training size of cross-validation and for held-out. The class-specific metrics show a positive difference between the cross-validation runs and the held-out set, but only for the classes where the total number of samples are extremely low ($N < 50$, Figure 4.3). As cross-validation only happens on 80% of the data, which in turn is split into 80% training and 20% cross-validation test fold, smaller classes have fewer samples to train on in a cross-validation run. As a result, the performance of the classifier is poorer on the cross-validation test folds for such classes. However, when testing on the 20% heldout set, we are training the model on the entire remaining 80% of the data. While this is of little consequence to classes that are well represented, the smaller classes are more thoroughly learnt during training.

4.3.1 Association of classification anomalies and biological similarities in held-out set

Among the poor-performing classes in the TCGA held-out set, certain patterns were evident. The three kidney adjacent-normal classes (KICH, KIRP, KIRP) had significant cross-calling, which was as expected because

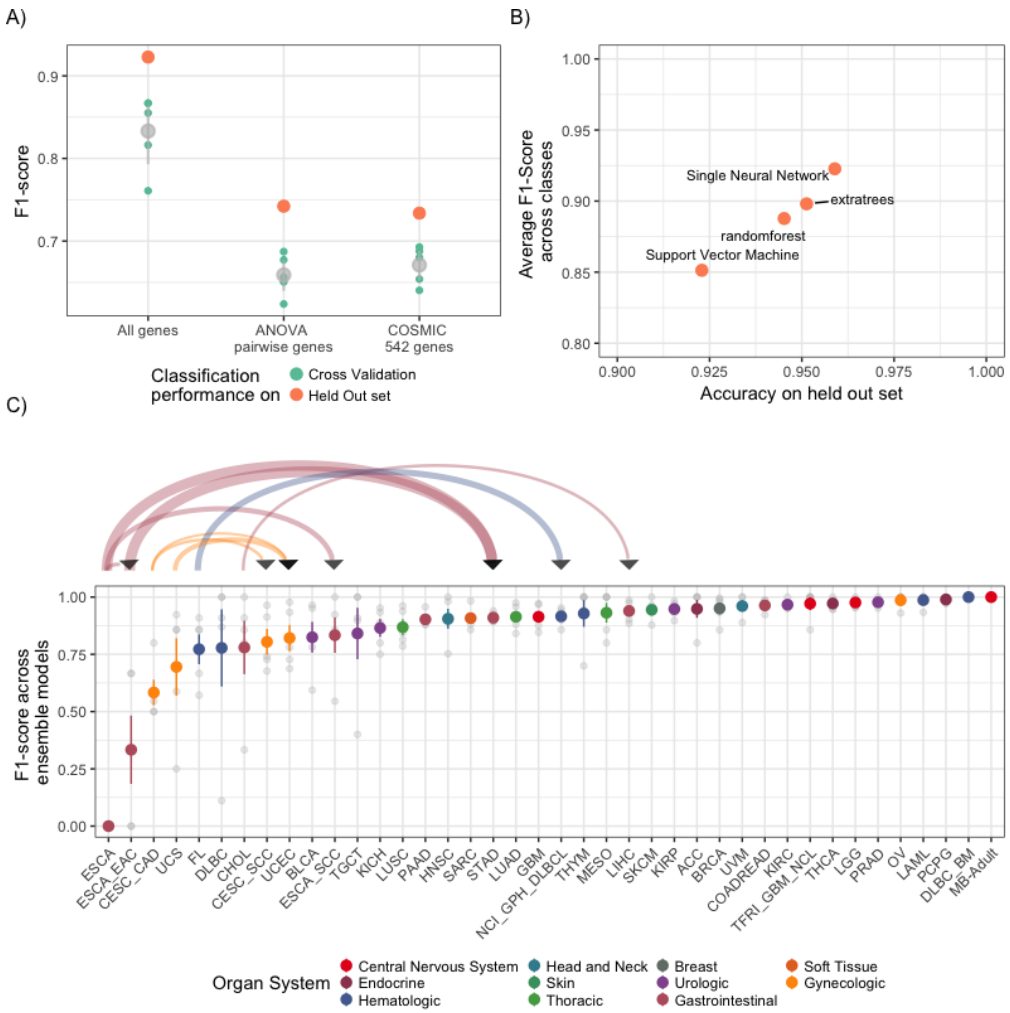

Figure 4.2: Results from algorithm and feature selection experiments, and performance on held-out test set. A) Feature selection does not improve pan-cancer classification. B) Comparison of algorithms - performance of single neural network on held-out set is higher than other algorithms. C) Validation of SCOPE on TCGA held-out set demonstrates high discriminatory power amongst most cancer types. Point with bar represents average F1-score and standard deviation spread for corresponding category. Incorrect predictions for more than 10% of samples belonging to a given cancer type are shown by curved directed edges. Curve width indicates relative fraction of samples in misprediction set. Mispredictions occur amongst cancer types with the same organ-system of origin. Specific trends are discussed further in Section 4.3.1.

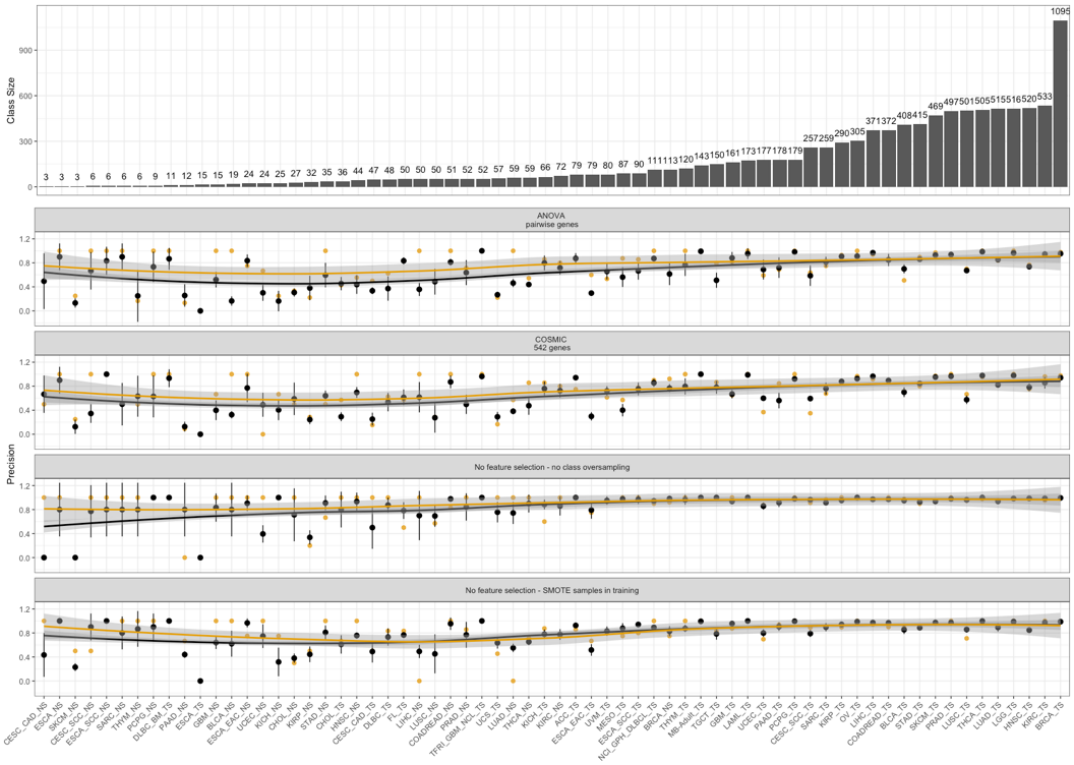

Figure 4.3: Performance of various models that make up SCOPE, on the cross-validation and held-out sets. The x-axis is ordered by increasing class size. Performance is reported as precision for the test-folds from CV *inblack* and for all samples in the held-out set *inyellow*. Number of samples in training are shown in the upper histogram panel. Cancer codes follow TCGA nomenclature and are defined in Table A.1, with _TS samples indicating tumours and _NS samples indicating adjacent normal tissues. The difference between CV-fold performance and held-out performance is typically larger for small classes. The difference become insignificant as class size approaches N $>$ 100. When the classifier is augmented with addition of synthetic samples in the training folds (last panel), we observe an overall increase in performance for the smaller classes with a concomitant reduction in the performance gap between mean-CV-precision and heldout precision. The line of best fit (loess) is indicated for each model, with standard error bounds in grey. The spread of performance across different CV folds is shown by the black point (mean) with 1 standard deviation bars.

all three represent healthy kidney tissue. Esophageal carcinomas and adenocarcinomas were often misclassified as stomach adenocarcinomas. For cervical cancers, which can be squamous, adenosquamous, and adenocarcinomas, subtypes were also challenging to distinguish by SCOPE. We found these trends were replicated in unsupervised clustering of the RNA sequencing data, suggesting biological rationale for the same (Figure 4.4).

As further evidence, we observed other molecular patterns previously noted in literature in our results. The endometrium is a common site of occurrence for uterine carcinosarcomas, and an endometrioid carcinoma-like profile is a well-documented molecular subtype of uterine carcinosarcomas. We found that uterine carcinosarcoma was frequently misclassified as uterine corpus endometrial carcinoma. The Cancer Genome Atlas analysis has found that a majority of uterine carcinosarcoma samples had serous-like endometrial carcinoma precursors [33]. This cross-calling was also observed by another group using this data set for classification [114].

4.3.2 Prioritization of known diagnostic gene features without prior knowledge

Manual review of the high-importance genes summarized in Appendix Table 1 showed that the genes prioritized for each class were biologically relevant to the corresponding cancer or normal tissue type. For example, two kidney-specific genes, *UMOD* and *AQP2*, were exclusively associated with the adjacent normal tissues from all three renal cancer types in training. Known diagnostic markers for renal clear cell carcinoma, namely *CA9* and *CA12*, were associated with renal clear cell carcinoma. Important genes for testicular germline cancers, *POU5F1*, *GDF3*, and *NANOG*, are known and proposed biomarkers. High *POU5F1* (OCT4) and *NANOG* expression is associated with spermatogenesis dysregulation [185]. Unexpectedly, in the absence of a healthy tissue class corresponding to a primary tumour type, some important genes for the cancer reflect biological characteristics of the progenitor healthy tissue, such as *DPPA3/5* for testicular germline cancers, and *TYR* and *MLANA* for uveal melanomas. These observations underscore the value of including adjacent normal tissues for a high-dimensional pan-cancer classifier.

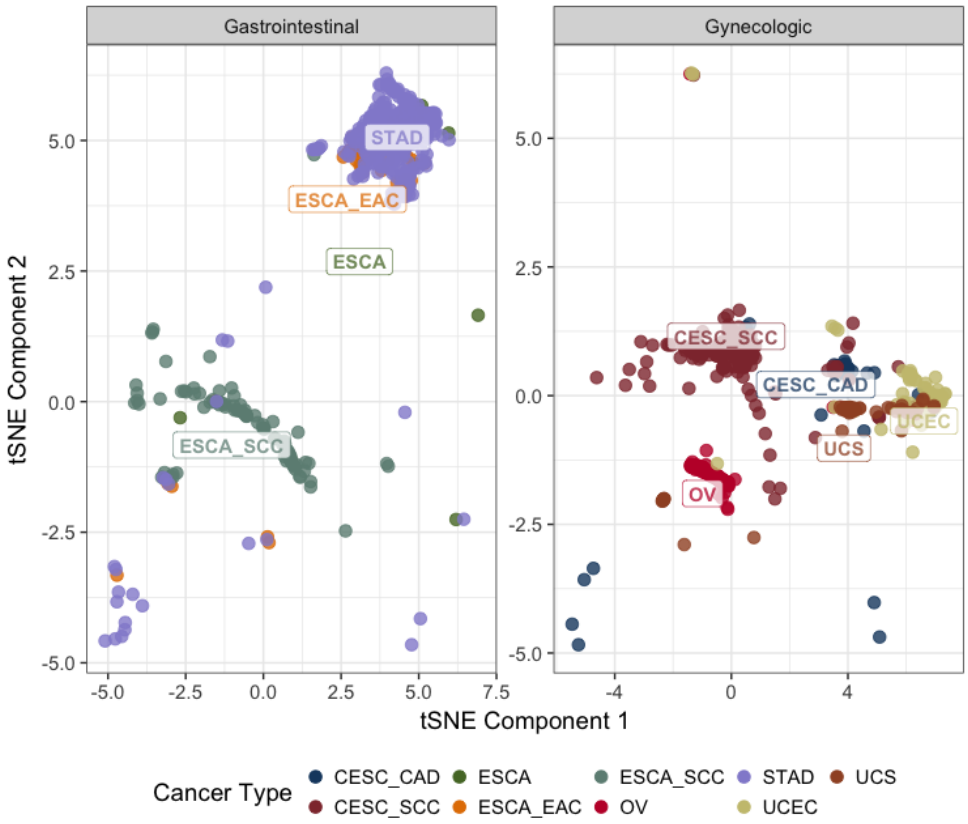

Figure 4.4: t-SNE plot of transcriptomic data in TCGA training cohorts. The relevant gynecologic and gastrointestinal cancer types are shown, and reflect the trends of cross-calling observed in SCOPE. Esophageal adenocarcinoma $ESCA\_EAC$ and stomach adenocarcinoma $STAD$ cluster together, as do uterine carcinosarcomas UCS with uterine corpus endometrial carcinomas $UCEC$.

4.3.3 External validation on primary cancers

Mesothelioma is a cancer that arises in the pleura, which lines the lungs. Three main histologic categories have been defined within mesothelioma: epithelioid, sarcomatoid, and a biphasic type that presents a combination of features from the former [93]. Subtype diagnosis in mesothelioma influences patient prognosis and disease management, but without specialized histopathologist training, there is low agreement between diagnoses [17]. We applied SCOPE on a previously published cohort of primary, untreated mesothelioma subtypes.

Characterizing cancers with mixed histology

We obtained 99.2% accuracy (125 of 126) in identifying epithelioid mesotheliomas and biphasic-epithelioid cancers in this cohort. This is as expected, because SCOPE was trained to identify epithelioid mesotheliomas (this subtype was exclusively represented in the mesothelioma training set). Twenty-three of 29 sarcomatoid mesotheliomas (79.3%) and 55 of 56 biphasic-sarcomatoid mesotheliomas (98.2%) were predicted with split confidence between mesothelioma and sarcoma (Table 4.4). In addition, four of the remaining six sarcomatoid subtype samples were predicted confidently as sarcomas. Appendix Figure 1 shows an example of what these split predictions look like as an output from SCOPE.

Table 4.4: Performance of SCOPE on the Genentech cohort of primary mesotheliomas. The training cohort was composed of epithelioid mesotheliomas, whereas the testing cohort was composed of epithelioid mesotheliomas and sarcoma-like mesotheliomas. Mesotheliomas that also show sarcoma-like histology are either predicted correctly as part sarcoma, part mesothelioma ("sarcomatoid mesothelioma"), or otherwise, usually as mesothelioma alone ("epithelioid mesothelioma"), or as sarcoma alone ("sarcoma").

Mesothelioma Subtype	Case Count	Precision	Recall	F1-Score	Predicted category	Count
Biphasic epitheliod-like	72	1	1.00	1.00	Epithelioid mesothelioma	72
Epithelioid	54	1	0.98	0.99	Epithelioid Mesothelioma	53
					Sarcomatoid Mesothelioma	18

Table 4.4: Performance of SCOPE on the Genentech cohort of primary mesotheliomas. The training cohort was composed of epithelioid mesotheliomas, whereas the testing cohort was composed of epithelioid mesotheliomas and sarcoma-like mesotheliomas. Mesotheliomas that also show sarcoma-like histology are either predicted correctly as part sarcoma, part mesothelioma ("sarcomatoid mesothelioma"), or otherwise, usually as mesothelioma alone ("epithelioid mesothelioma"), or as sarcoma alone ("sarcoma"). *(continued)*

Mesothelioma Subtype	Case Count	Precision	Recall	F1-Score	Predicted category	Count
Sarcomatoid	29				Epithelioid Mesothelioma	5
					Sarcoma	4
					Other	2
Biphasic sarcoma-like	56				Epithelioid mesothelioma	38
					Sarcomatoid mesothelioma	17
					Other	1

Abbreviations: NA - not applicable; SCOPE - Supervised Cancer Origin Prediction using Expression

4.3.4 Providing diagnosis for pre-treated metastases

In an independent set of 201 post-treatment metastatic cancers, SCOPE performed well above the baseline linear classifier, achieving an overall accuracy (SD) of 86% (11%), and a mean (SD) F1 score of 0.79 (0.12) (Figure 4.5 A; Table 4.5). Among the 41 mispredictions, seven (17.1%) matched the site of biopsy (for example, predicting hepatocellular carcinoma for a breast cancer biopsy specimen from the liver), and 13 of the 41 (31.7%) matched a cancer type with same organ system of origin instead (for example, predicting uterine carcinosarcoma as ovarian cancer, predicting stomach adenocarcinoma as esophageal adenocarcinoma). For the remaining 21 cases, no obvious explanation was found for misclassification. Because our method provided a confidence score for each prediction, we found that in the set of confident diagnoses from the ensemble (118 of 201, confidence score of 80%, spanning 20 cancer types) accuracy went up to 92%.

Table 4.5: Performance of SCOPE on the metastatic cohort. Number of mis-predictions are listed in brackets if more than one.

		Cohort metrics			Cases predicted as			
Cancer type	Cases	Precision	Recall	F1-Score	Diagnosis	Biopsy Site	Organ System	Other
Metastatic biopsies								
Adenocortical CA	1	1.00	1.00	1.00	1	-	-	-
Follicular Lymphoma	1	1.00	1.00	1.00	1	-	-	-
Mesothelioma	1	1.00	1.00	1.00	1	-	-	-
Prostate AC	1	1.00	1.00	1.00	1	-	-	-
Testicular Germ Cell Tumour	1	1.00	1.00	1.00	1	-	-	-
Thymoma	1	1.00	1.00	1.00	1	-	-	-
Colorectal AC	21	1.00	0.81	0.89	17	LIHC	STAD (2)	CHOL_n
Papillary Kidney AC	2	1.00	0.50	0.67	1	-	-	LUAD
UCEC	5	1.00	0.40	0.57	2	-	BRCA	”BLCA,STAD”
Uterine Carcinosarcoma	4	1.00	0.25	0.40	1	-	”OV, SARC”	HNSC
Breast CA	65	0.97	0.97	0.97	63	LIHC_n	-	BLCA
Lung AC	14	0.93	1.00	0.97	14	-	-	-
Sarcoma	17	0.90	0.53	0.67	9	LIHC	-	”BRCA, DLBC (2), GBM, SKCM (2), KIRC”
Ovarian CA	7	0.86	0.86	0.86	6	-	-	PAAD
Prostate AC	9	0.75	0.33	0.46	3	LIHC	”CHOL (3), LUSC”	BLCA
Cholangio-CA	5	0.67	0.80	0.73	4	-	STAD	-
Cutanous Melanoma	2	0.50	1.00	0.67	2	-	-	-
Diffuse Large B-Cell Lymphoma	1	0.33	1.00	0.50	1	-	-	-
Stomach AC	3	0.25	0.67	0.36	2	LIHC	-	-
CESC-AC	1	0.00	0.00	0.00	-	-	-	STAD
Esophageal AC	2	0.00	0.00	0.00	-	LIHC	STAD	-
Esophageal SCC	4	0.00	0.00	0.00	-	LUSC (1)	-	”CESC_SCC (2), LUSC (1)”
Primary site biopsies								
Adrenocortical CA	1	1.00	1.00	1.00	1	-	-	-
Breast CA	4	1.00	1.00	1.00	4	-	-	-
Colorectal AC	1	1.00	1.00	1.00	1	-	-	-

Table 4.5: Performance of SCOPE on the metastatic cohort. Number of mis-predictions are listed in brackets if more than one. *(continued)*

Cancer type	Cases	Cohort metrics			Cases predicted as			
		Precision	Recall	F1-Score	Diagnosis	Biopsy Site	Organ System	Other
Glioblastoma Multiforme	4	1.00	1.00	1.00	4	-	-	-
Brain Glioma	2	1.00	1.00	1.00	2	-	-	-
Liver Hepatocarcinoma	1	1.00	1.00	1.00	1	-	-	-
Pancreatic AC	2	1.00	1.00	1.00	2	-	-	-
Cutaneous Melanoma	1	1.00	1.00	1.00	1	-	-	-
Uterine Carcinosarcoma	1	1.00	1.00	1.00	1	-	-	-
Sarcoma	6	1.00	0.83	0.91	5	-	-	HNSC
Lung AC	4	1.00	0.75	0.86	3	-	LUSC	-
Mesothelioma	4	1.00	0.75	0.86	3	-	-	KIRC
Lung SCC	1	0.50	1.00	0.67	1	-	-	-
UCEC	1	0.00	0.00	0.00	-	-	CESC_SCC	
Total								
	201	0.80	0.76	0.75	160	7	13	21

Abbreviations: Prediction categories: Cases where predicted cancer type matched pathology diagnosis (Diagnosis), was same as tissue type of biopsy site (Biopsy Site), matched a cancer type with same organ-system of origin (Organ-system), or did not match any of the above (Other). Abbreviations: AC - adenocarcinoma, CA - carcinoma, SCC - squamous cell carcinoma, CESC AC - cervical/endocervical adenocarcinoma, UCEC - uterine corpus endometrial carcinoma

In our assessment of this cohort, we found no association between classification accuracy and tumour content ($P = 0.59$), and a weak correlation with the size of training class (Pearson correlation coefficient, 0.39). There was an association between classification accuracy and confidence score (N = 201; $P < 0.001$). In metastatic site biopsies (N = 168), an association was found between low tumour content and the diagnosis of another cancer type with the same organ system of origin (Figure 4.5C). This association was absent in primary site biopsies (Figure 4.5B, Table 4.5).

4.3.5 Identification of putative primary tumour type for cancers of unknown primary

We retrospectively predicted the cancer type for 15 cancers where the primary site of origin was unknown after initial pathology assessment.

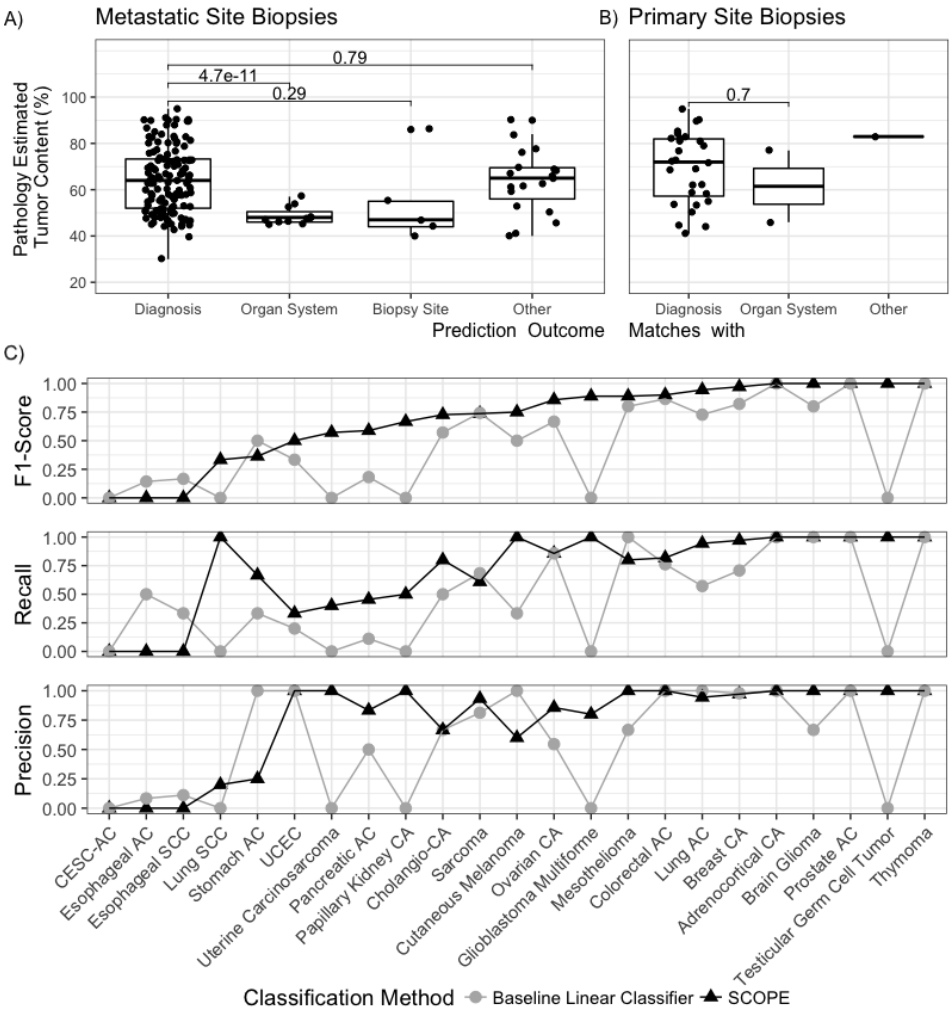

Figure 4.5: Performance of SCOPE on external metastatic cohort. A) Two-sided t-tests show a significant association of tumour content on general diagnosis as organ system, for biopsies samples from site of metastasis. B) Two-sided t-tests show no effect of tumour content on misclassification to organ system, for biopsies sampled from the cancer's site of origin. C) SCOPE has improved performance compared with baseline linear comparator trained from a statistically filtered feature subset. Abbreviations: AC - adenocarcinoma, CA - carcinoma, SCC - squamous cell carcinoma, CESC AC - cervical/endocervical adenocarcinoma, UCEC - uterine corpus endometrial carcinoma.

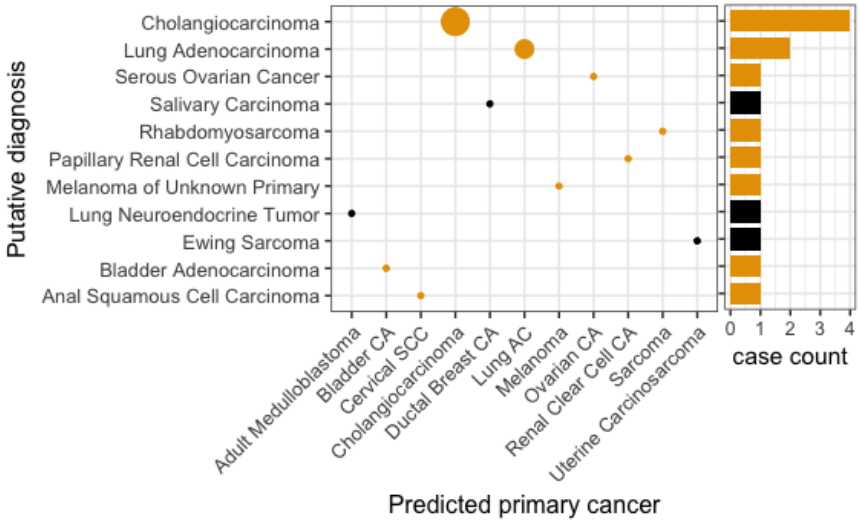

Figure 4.6: SCOPE prediction and putative primary for cancers with unknown primary site. A confusion matrix of predictions is shown, where the size of the circles represents relative number of samples in each category. Case count for CUPs by putative origin is shown with a histogram on the right. Correct predictions are indicated in yellow whereas incorrect ones are shown in black. Salivary carcinoma, neuroendocrine tumours, and ewing sarcomas were not present in SCOPE training, explaining the inability of the method to identify these accurately. Abbreviations: CA - carcinoma, AC - adenocarcinoma.

These tumours were therefore refractory to standard pathology protocols. Subsequent diagnosis was determined by analysis of whole-genome sequencing and RNA-Seq data, and validated by pathology review and immunohistochemistry. The prediction by SCOPE was compared against this putative diagnosis. As shown in Figure 4.6, the classifier's prediction matched all putative diagnoses except one Ewing sarcoma, one neuroendocrine tumour, and one salivary carcinoma; these three cancer types were not present in training.

4.3.6 Impact of feature removal on classification

In order to evaluate whether there is a thresholding effect on accuracy based on the number of genes provided as input to SCOPE, we performed a

gene 'blanking' experiment for the 81 samples in the metastatic cohort that SCOPE predicted with high confidence. The RPKM values of a percentage of genes were randomly set to 0 before the RNA-Seq data was passed as input to SCOPE, and this process repeated 10 times for each percentage threshold. 10 percentage thresholds were tested, in the range 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 99%. The resultant accuracy was measured for all 81 samples. We found that as more genes are set to zero in the input, fewer samples were predicted correctly (pearson correlation -0.80) (Appendix Figure 2). The decline in performance happens sharply when 40% of the genes are set to 0. Surprisingly, accuracy did not entirely go to 0 when 99% of the genes were set to 0 (only the expression of 177 random genes passed as input), staying at around 12.5%.

4.4 Conclusion

In this chapter, we describe the development and validation of a cancer-type classifier that leverages the entire gene-expression profile of a tumour sample to correctly identify its site of origin. Out method achieves 97% overall accuracy and a mean (SD) F1-score of 0.92 (0.06) on our held-out set. This performance level is maintained on external cohorts, with an overall accuracy of 99% on primary mesotheliomas and mean (SD) accuracy of 86% (11%) for a dataset of various metastatic cancers. We use the confidence score values (equivalent to probabilities) for predictions to characterize cancers with mixed histology.

Metastatic cancers form 12-15% of cancer diagnoses worldwide, but account for 90% of cancer-associated deaths. While in some cases this is because of a paucity of research in identifying appropriate targets for rare cancers, often this can occur as a result of delayed diagnosis or misdiagnosis (in 15-28% of the cases) [39, 69]. SCOPE can be easily deployed for automated diagnosis from RNA-Seq data, facilitate analysis of rare cancers [26] and support re-alignment of diagnosis [105, retrospectively evaluated for case-study described in Chapter 2; 72]. As shown by its performance on CUPs, it is particularly useful in expediting precision oncology workflows and in clinical laboratories where access to a plethora of immunostains for sequential diagnosis may be limited. The method is available online as a python package, *cancerscope*.

Since the method spans 40 established primary cancer types and 26 normal

tissue types, it is able to consider multiple differential diagnoses and provide a quantifiable prediction of the most likely primary of origin, guiding our efforts in using personalized cancer genomics to investigate the underlying biology of morphologically challenging cancers and cancers with multiple differential diagnoses from histopathology. Since it leverages whole transcriptome profiles, the impact of known biological programs can be characterized at a sample-specific level as well. This application of SCOPE is described in greater detail in the next chapter, along with published case-studies where SCOPE was used for biological analysis and interpretation.

Chapter 5

Enabling cancer transcriptome analysis from SCOPE using single-sample pathway impact evaluation (PIE)

Precision oncology necessitates detailed characterization of aberrant pathways and driver biology of individual tumours. This process requires an accurate cancer diagnosis to align the observed changes against the appropriate background and select comparators. Cancer classifiers leveraging transcriptome and genomic data can optimize the task of distinguishing various cancers, but understanding the biological underpinning of the diagnosed cancer still remains a complex and laborious manual task. We have developed a method that uses classifiers trained with whole-transcriptome data to provide single-sample biological pathway importance scores. This unsupervised exclusion analysis approach for pathway impact evaluation (PIE) recapitulates cancer-specific biology and clustering of the classifier training data from The Cancer Genome Atlas, performs single-sample analysis of treatment-resistant cancers to help explain diagnosis and subtyping, identifies biological pathways associated with drug response, and reflects known biology for metastatic cancers. PIE provides a score of each biological pathway across forty cancer types, for a given sample. It is available as a python package ('cancerscope') that includes an RNA-Seq based pan-cancer classifier.

5.1 Background

Cancers acquire oncogenic potential through somatic mutations, epigenetic modifications, genomic rearrangements, copy number alterations, and gene expression changes [80]. Many genes have been identified that either promote or restrict the growth of cancer cells. Their downstream effectors and upstream regulators have been curated and studied, and the key oncogenic genes and their interaction partners have been placed in the context of regulatory networks [211]. Researchers are able to identify genes that may be driving an individual cancer (and hence serve as therapeutic targets) by aligning observed changes against this complex set of oncogenic pathways [68, 213]. Evaluation of the scope and impact of these changes on specific cellular pathways and protein networks at the single-sample level is a key challenge.

In most existing approaches for single-sample analysis, comparator cohorts and samples are required every time an analysis is performed, and results can vary depending on the type of statistical metric used [116, 228]. The requirement of controls and background samples for every single-sample analysis makes it difficult to analyze cancers that lack suitable comparator datasets (rare cancers, post-treatment cancers), present with mixed histology, or have important individual signals that characterize the tumour. The analysis can be further impacted by platform biases, and in the event of small case/control studies, be severely underpowered [210]. PARADIGM, the only equivalent approach for measuring patient-specific pathway activities pathway network of interest, requires users to manually define the nature of interactions between each of the member genes for each pathway [208]. The tool itself has since been commercialized, and the utility of the publicly available implementation is limited in the absence of accompanying pathway networks.

We have developed an approach for single-sample pathway impact evaluation (PIE) by quantification of the impact of various gene sets (representing pathways) on classification confidence from pan-cancer classifiers trained with large feature representations. By encapsulating all required comparator information into a classifier, we forego the need for comparator datasets, and provide quantification of pathway impact across a large number of cancer types simply using pathway gene lists. In a previous work we had trained SCOPE, an ensemble of neural networks, to distinguish forty primary cancer types (represented by the site of origin and

well-established cancer subtypes) [73]. As described in Chapter 4, SCOPE uses large gene expression profiles from bulk RNA-Seq as input (over 17,688 genes). Using this tool as the core classification model, we tested the impact on classification for 3,963 biological pathways and gene groups that represent various regulatory and biochemical functions in eukaryotic cells. For a given sample, pathway impact scores were calculated by setting the expression values of the pathway-specific genes to zero, and then calculating the difference in classification performance against the original sample.

We demonstrate the utility of PIE for pathway-level analysis and interpretation of classification results from SCOPE through four analyses – a) validating that the method recovers relevant biological pathways from the training data of the underlying classifier, b) conducting cohort-wide pathway analysis and recovering clustering by site of origin based on pathway importance scores for two independent cohorts of metastatic cancers that were not included in classifier training, c) independently identifying oncogenic pathways important for cancer maintenance and a pathway associated with paclitaxel treatment by analyzing a previously published case of vulvar adenocarcinoma [72], and d) refining the diagnosis and recovering known biological programs driving the cancer for previously published case-study of a cancer with unknown primary (CUP) that was originally diagnosed using SCOPE [105]. Lastly, after using PIE to generate pathway-level representations of samples, we discover new subtypes in a previously analyzed cohort of metastatic prostate adenocarcinoma [169].

5.2 Methods

5.2.1 Test Data

Cohort-level validation of PIE was performed using three publicly available cohorts of advanced cancers. In all these cases, the gene expression data (bulk RNA-Seq) was filtered to select the 17,688 genes that overlapped with the required input for SCOPE, unless otherwise indicated below. No other normalization or pre-processing was done on the RPKM values. The first cohort of 10,156 primary tumours and healthy tissues was drawn from The Cancer Genome Atlas [215]. Additional details about this data are included already in Chapter 4, Section 4.2.1.

A cohort of 651 advanced cancers was obtained from the personalized

oncogenomics (POG) clinical trial at BC Cancer. These cases were selected based on the following criteria – (a) a primary of origin was identified, based on a joint consideration of clinical/pathology/genomic data, and (b) cDNA libraries prepared from the biopsy sample passed in-house quality control.

The second cohort was the MET500 cohort of advanced, metastatic patients [169]. Only 375 of the 500 patients in this cohort had available RNA-Seq data with a confirmed diagnosis that mapped to a TCGA category. 17,347 of the 58,450 annotated genes overlapped with the 17,688 genes required as input to SCOPE. The missing genes were set to 0 for all subsequent analyses.

For sample-level analysis, PIE was used retrospectively to profile two individual cases where detailed integrative pathway analysis based on whole-genome and transcriptomic sequencing was available. These two cases - a vulvar adenocarcinoma, and a rare thyroid-like renal carcinoma - have been previously published [72, 105].

5.2.2 Classifier used for PIE measurements

Since PIE's scores are calculated by blanking representative pathway genes and measuring changes in quantified classification scores from a multi-class classifier, we wanted to use a previously published, open-access classifier that leverages large transcriptomic profiles (so as to facilitate blanking of all relevant gene sets) and provides classification probabilities across a vast number of cancer types (so as to provide insights about as many cancer types as possible). SCOPE is a previously validated cancer-type classifier [73]. It is trained on primary cancers and adjacent normal samples from The Cancer Genome Atlas [215]. SCOPE is an ensemble of five different neural network classifiers, each of which provide a probability value between 0.0 and 1.0 for output class (40 cancer types, 26 healthy tissues). The average probability value across the five ensemble members was used as the confidence score for a given class. The baseline confidence score for each class was calculated using the default sample input (RPKM values for 17,688 genes).

5.2.3 Pathway analysis for individual samples

Calculation of pathway importance

Pathways were curated manually at the Michael Smith Genome Sciences Centre from KEGG [99], Reactome [45], PathCards [9], TarBase [178], Consensus Pathway Database [98], and the Pathways Interaction Database [175]. This resulted in a set of 3,952 pathways. An additional 11 canonical oncogenic pathways representing common signaling cascades that are disrupted in cancer were curated in-house, resulting in a set of 3,963 pathways. Pathways were represented as the set of their member genes. The impact of each pathway on classification was calculated by setting the RPKM values for the pathway genes to 0.0 in the input and calculating the resultant confidence scores across the 66 output classes. Pathways that were important for classification of the sample as class 'm' would have a reduced confidence score for class 'm' when the relevant genes were removed from the input. Inversely, pathways that were preventing the sample from being predicted as class 'm' would have a higher confidence score for class 'm' upon being blanked in the input.

For a given pathway-sample pair, the pathway confidence score was subtracted from the sample confidence score to obtain the *pathway importance* score for each output class. A positive score for a given output class 'm' meant the pathway was important for classification of the sample as class 'm'. A negative score meant the pathway was confounding the sample classification as class 'm'.

Calculation of number of important sample-level pathways in TCGA cohorts

For each sample, positively scoring pathways (PIE score > 0.0) were selected. The inter-quartile range (IQR) was calculated as the difference between the 25th and 75th quantile of these scores. Pathways with a PIE score > 1.5 * IQR were selected as being important in the sample.

5.2.4 Cohort-level pathway analysis

Pathway profiles of each sample were generated for 3,963 pathways using the pathway importance scores. A sample * (pathway, output class) matrix was generated from 651 POG samples, resulting in a matrix of size [651, 261558]. Similarly, a matrix of size [375, 261558] was generated from the 375 MET500 samples.

Visualization of samples in the TCGA, POG, and MET500 metastatic cancer cohorts

The pathway, output class pairs where the output class matched the cancer-type of diagnosis were selected for each sample. This reduced the number of 'features' per sample to 3,963, matching the number of unique pathways. Uniform Manifold Approximation and Projection (UMAP) was used for dimensionality reduction and visualization of this high-dimensional data. UMAP decomposition of each cohort was generated using the umap package in Python using Manhattan distance, n_neighbours set to 15, n_components set to 2, and initialized with the first two Principal Components of the PCA decomposition of the matrix, as recommended elsewheres [106]. All other function arguments were set to the default. No normalization of the input was done prior to generating the decomposition as the input values were already scaled measurements in a range between [-1,1].

Measurement of cluster metrics by cancer-type in the TCGA, POG, and MET500 cancer cohorts

Silhouette indices [101] were used to quantify the quality of clusters obtained from UMAP projections of cancer cohorts. Given cluster assignments, silhouette index measures how similar a given sample is to its own cluster, compared to other clusters. A high positive value indicates the sample is well-placed in its present cluster, whereas a high negative value indicates it is poorly placed in the current cluster (i.e. has more similarity to a different cluster). For the silhouette scores presented in this analysis, the diagnosed cancer-type was used as the default cluster label, and the Principal-Component initialized UMAP projections used as the sample measurements.

The 'silhouette_samples' function in sklearn's metrics [151] was used to calculate the sample-level cluster correspondence score from the PCA-initialized UMAP projections, as measured against the cancer-types. Euclidean distance was used along with all other defaults in the function.

Identification of important pathways distinguishing the various prostate adenocarcinoma (PRAD) clusters in the MET500 cohort

For each cluster of PRAD samples, the cancer-type specific pathway importance scores were selected. The mean importance of each pathway across all the samples in the cluster was calculated. Pathways that had a positive mean importance compared to the other two PRAD clusters were filtered and sorted by decreasing mean importance. The mean pathway importance of top pathways in each of the clusters were then plotted.

$P^{\star\hat{}}(\mu_j^\star) = (1/B)\sum_{j=1}^{B} I(\mu\hat{}_X > \hat{\mu})$

5.2.5 Statistical selection of top pathways associated with each cancer type

For a given pan-cancer cohort **X** (X being TCGA, POG, or MET500), pathways positively associated with each cancer type were identified using one-sided test of significance across bootstrapped samples. For a given cancer type C, the subset of samples in **X** belonging to C were identified, and bootstrapped datasets were generated from the resultant subset of sample * pathway matrix, X_c. Sampling with replacement was done to form bootstrapped datasets with 20% the samples in X_c. 1000 iterations were performed, indexed with *j*.

Next, for each pathway, the P-value for rejecting the null hypothesis (null hypothesis being that PIE score of pathway is not significantly higher in cancer type C, compared to the entire cohort X) was calculated as follows:

$\hat{P}^\star(\mu_j^\star) = (1/B)\sum_{j=1}^{B} I(\hat{\mu}_X > \hat{\mu}_j)$

$\hat{\mu}_X$ is the mean PIE score across all samples in **X**
$\hat{\mu}_j$ is the mean PIE score of samples in the j^{th} bootstrap from X_c
B is the number of bootstrap iterations sampled from X_c
I is an indicator function that equals 1 when the arguments compute to True, and 0 otherwise.

The resultant pathway p-values were ordered by increasing p-value and then by decreasing mean value in X_c and the top n pathways (of highest statistical significance and high quantitative impact) selected. Q-values were calculated using the 'qvalue' package [46] at a false discovery rate (FDR) threshold of 0.05. Statistically significant, positively associated pathways for a given cancer type were selected with the criteria Q-value <= 0.001 and mean PIE score in cancer type > 0.01.

5.2.6 Statistical identification of important pathways for single-sample analysis

For a given sample, for each pathway, a two-sided Grubbs test was applied to identify single positive or negative outliers from across the 66 classification categories. Positive (right-tail) outlier pathway, cancer-type

pairs were filtered to identify outlier pathways in the classification category of interest. These pathways were then ordered by their PIE scores and the top-25 pathways used to generate the associated visualizations.

5.3 Results

SCOPE uses large gene expression profiles from bulk RNA-Seq as input (over 17,688 genes). Using this tool as the core classification model, we tested the impact on classification for 3,963 biological pathways and gene groups that represent high-level cellular functions. For a given sample, pathway impact scores were calculated by setting the expression values of the pathway-specific genes to zero, and then calculating the difference in classification performance against the original sample.

5.3.1 Pathway impact profiles allow clustering and analysis of samples by cancer type

Firstly, we tested whether pathway profiles generated by PIE reflected the biology of samples in SCOPE's training data from The Cancer Genome Atlas (TCGA; Weinstein et al. [215]). A uniform manifold approximation and projection (UMAP) from the 3,963 pathways for each sample retrieved the expected clustering for 71% (N = 25/35) of the cancer types (mean silhouette score > 0.0, positive value indicates support for clustering with other samples in the same cancer-type), demonstrating that the pathway profiles were sufficient in recovering most cancer types (Figure 5.1, Appendix Figure 3). Box plots illustrate the median (centre black line) with the lower and upper hinges indicating the 25th and 75th percentiles respectively. The upper whisker shows the largest value at-most 1.5 times the interquartile range (IQR) from the hinge, and the lower whisker shows the smallest value at most 1.5 times the IQR of the hinge. The IQR is calculated as the distance between the first and third quartiles. Data points outside these ranges are plotted as individual points. Poorly clustered cancer-types primarily included gastrointestinal and gynecologic malignancies.

Pathways common to multiple cancer types across various organ-systems of origin reflected cancer biology and pathways relevant to cellular function (Figure 5.2). The latter included pathways such as translation initiation

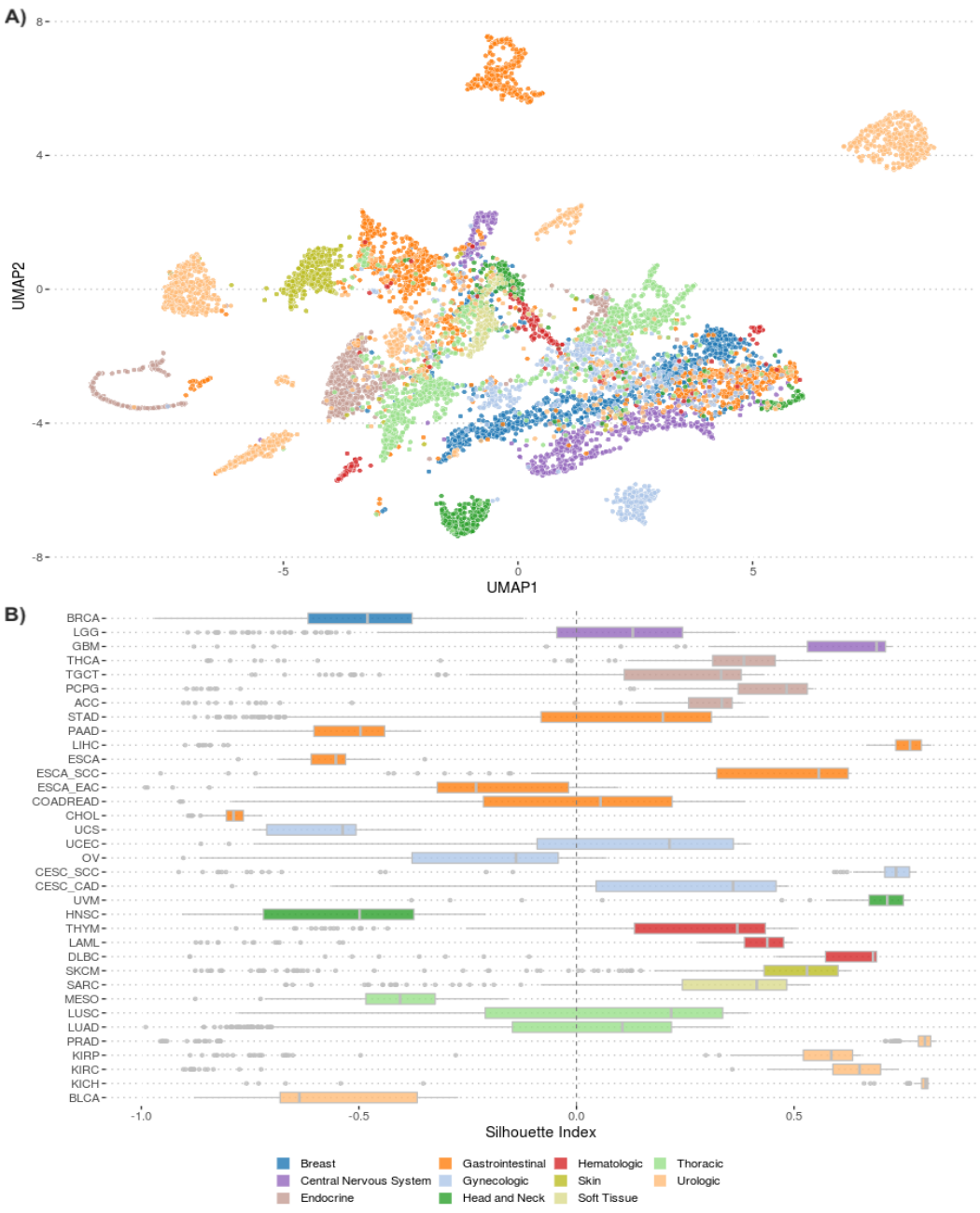

Figure 5.1: UMAP projections of PIE profiles for 3,963 biochemical pathways, for samples in the TCGA cohort of primary tumours. For ease of readability, the projections in panel A) show TCGA tumour types coloured by their organ system of origin. The spread of sample-specific silhouette indices, grouped by cancer type, is shown in panel B).

and termination, metabolism, gene expression, adaptive immune system, and signal transduction. Various oncogenic pathways were also globally represented, notably cell cycle, NOTCH, MAPK, VEGF and TNFalpha pathways.

For individual cancer types, statistically significant pathways from PIE (FDR adjusted Q-value < 0.001) overlapped with known biology (Figure 5.3, Appendix Table 2). Examples included the C-MYB and estrogen receptor networks in breast cancer, steroid hormone pathways in endocrine tumours, tubulin folding and neural crest differentiation in gliomas, androgen receptor and prostate cancer pathways for prostate adenocarcinoma, and T-Cell Receptor signaling in hematologic cancers.

Next we investigated trends for pathway complexity in different cancer-types. We observed that complex and heterogeneous cancers like glioblastoma multiforme (GBM), sarcoma (SARC) and various hematologic malignancies had over 200 unique statistically significant pathways associated with them, whereas other tumours like pheochromocytoma (PCPG) and thyroid cancer (THCA) needed fewer than 20 pathways (Figure 5.4a). We found a similar trend when looking at the number of important pathways at the sample-level (Figure 5.4b). Comparing this distribution with mutation burden data for these cohorts, we found that sample-level pathway counts separated the cancer types by known mutation frequencies [92]. On average, several cancer types with low mutation burden also had fewer important pathways per sample (PCPG, Uveal Melanoma (UVM), Lower Grade Glioma (LGG), and acute myeloid leukemia (LAML)), and ones with high mutation burden had a higher number of pathways per-sample (esophageal squamous (ESCA_SCC), lung squamous (LUSC), subcutaneous melanoma (SKCM)). Notable exceptions were testicular germ cell tumours (TGCT) and thyroid cancer (THCA), that harbor low mutation burden, but had a high count of sample-level pathways, and on the inverse trend (having very few important pathways per sample, but typically having a high mutation burden), stomach adenocarcinoma (STAD) and lung adenocarcinoma (LUAD).

5.3.1.1 PIE profiles enable cohort analysis of metastatic cancers

We also interrogated PIE profiles for a cohort of 651 metastatic cancers from the Personalized OncoGenomics (POG) project and 375 metastatic cancers from the MET500 cohort [169]. UMAP profiles showed consistent ability to

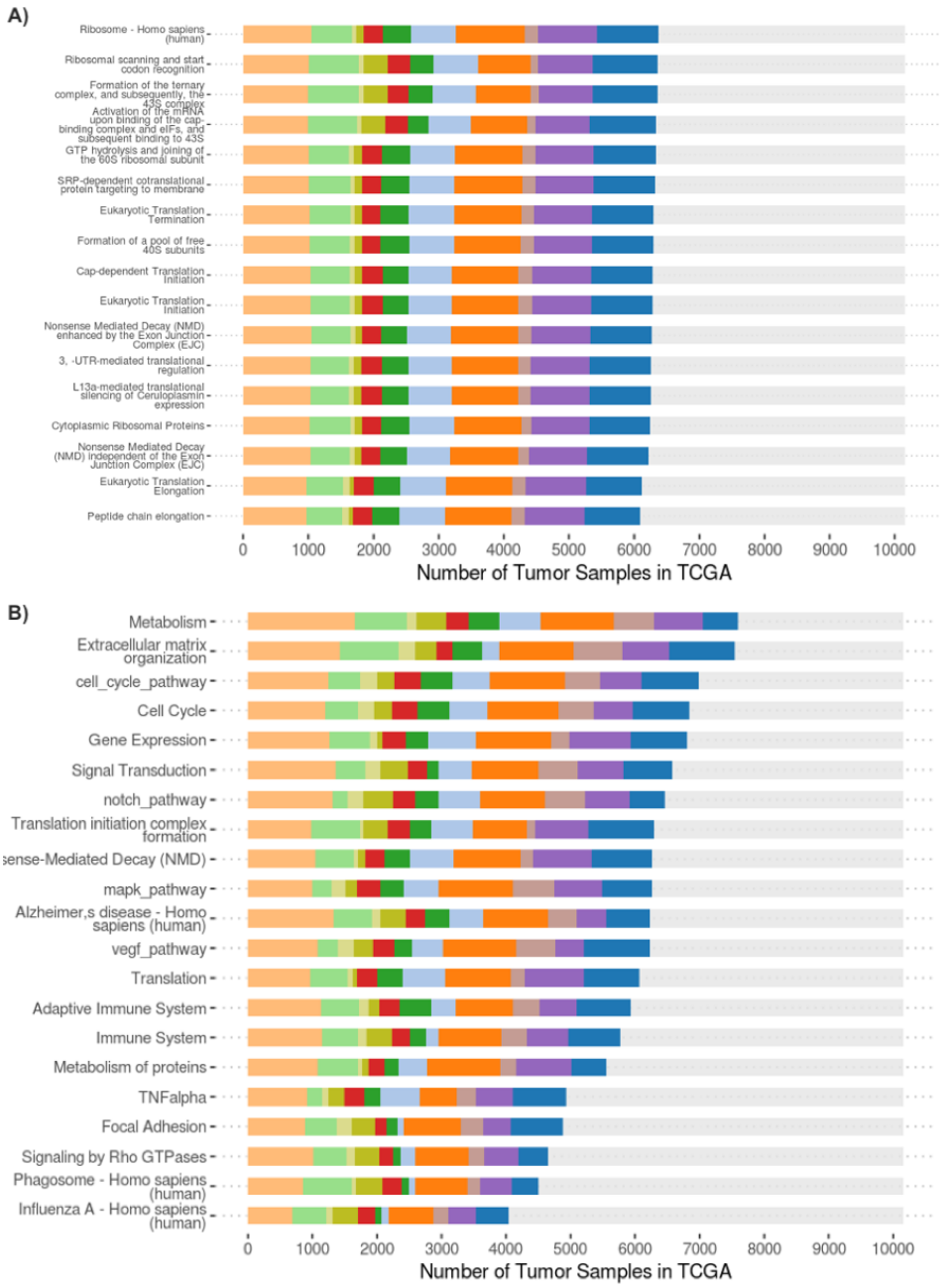

Figure 5.2: Pathways commonly associated with multiple cancer types in the TCGA cancers are shown. Grey bars indicate total number of tumour samples evaluated, whereas coloured bars indicate the number of tumour samples from the respective organ-system of origin. Panel A) shows the most common cell-function pathways. Panel B) shows the most common cancer-associated pathways.

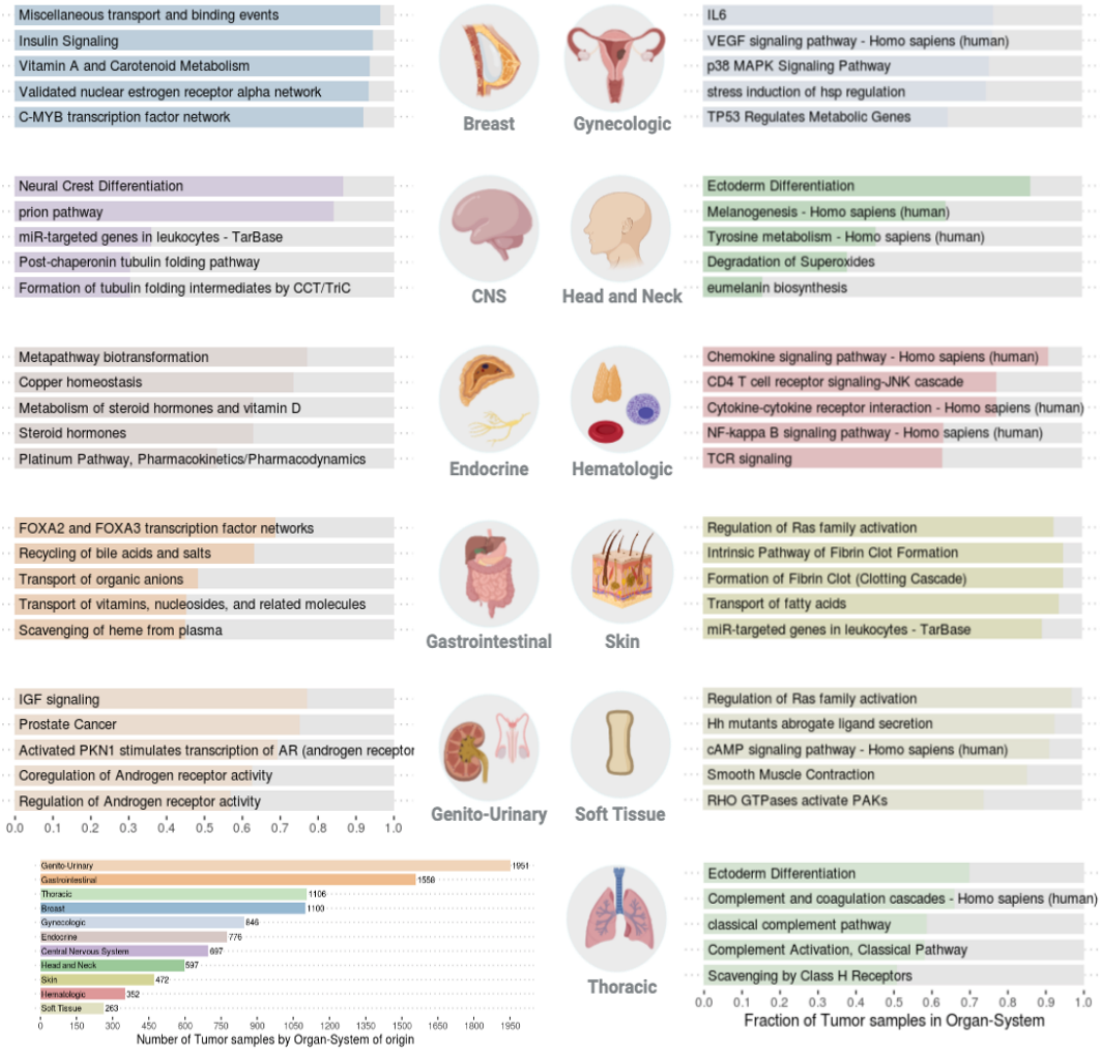

Figure 5.3: Statistically significant pathways in TCGA cancers. Panels show important pathways for each tumour and normal category, grouped by organ system of origin. Each group shows the top-5 pathways associated exclusively with cancers in the relevant organ-systems of origin, ordered by number of samples in which the pathway had a positive PIE score. Coloured bars indicate fraction of tumour samples from the organ-system where the respective pathway was positively scored by PIE.

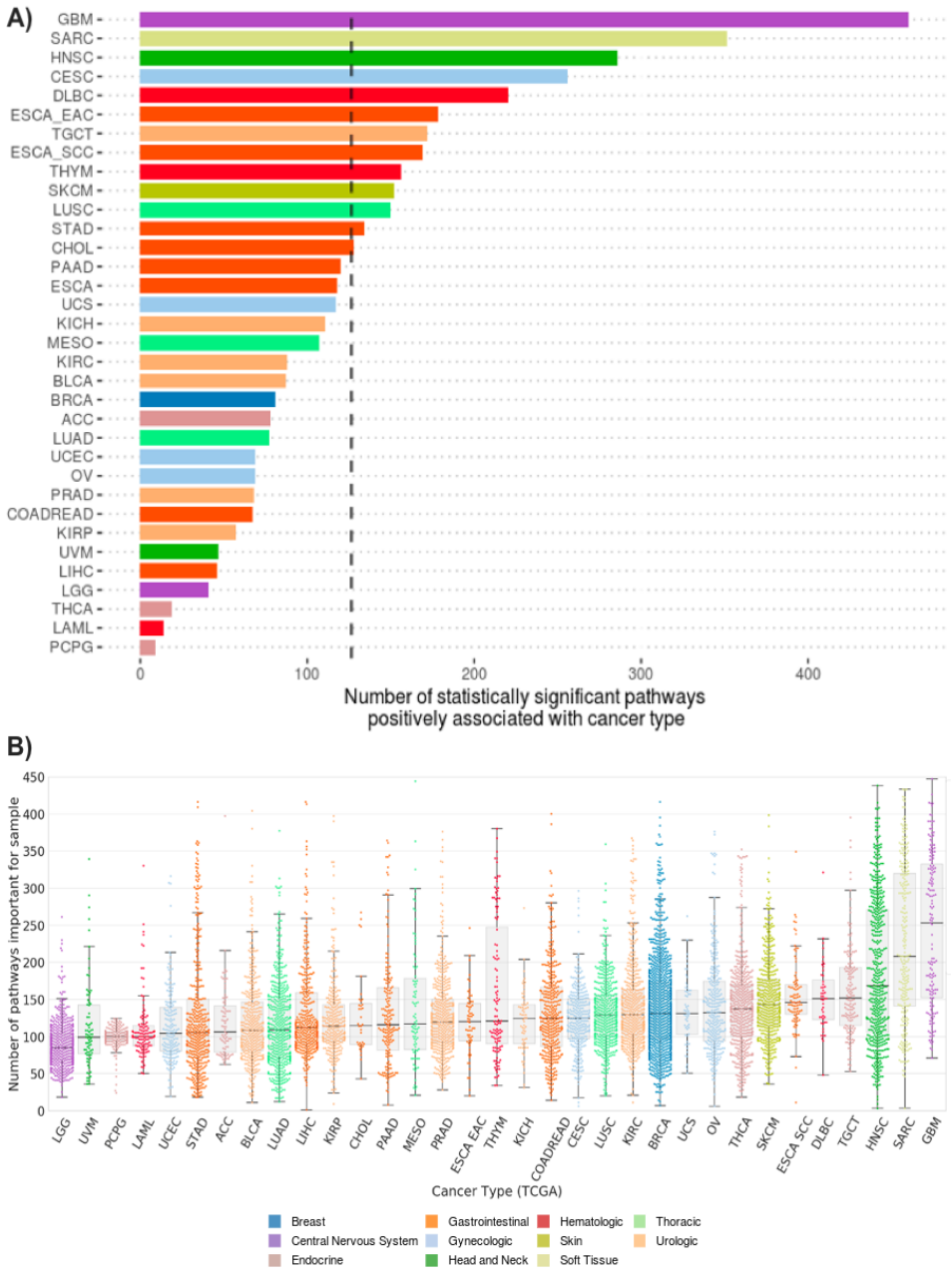

Figure 5.4: Determination of pathway-level activities for TCGA primary cancers, using PIE. Panel A) shows the number of statistically significant positively associated with each cancer-type, from the group of 3,963 pathways evaluated using PIE. Panel B) shows the number of pathways with statistically significant PIE scores per sample.

cluster samples by cancer types in both these cohorts, with an average cohort silhouette distance of 0.15 (+/- 0.35), 0.13 (+/- 0.41) and mean silhouette score > 0.0 for 7/9, 6/9 common cancer-types in the POG and MET500 cohorts respectively (Figure 5.5, Figure 5.6). Not surprisingly, malignancies absent in the SCOPE classifier had poorer clustering with the corresponding cancer-type (silhouette index < 0.0), primarily fusion-associated sarcomas (Figure 5.7). Top pathways for all cancer types in both cohorts are listed in Appendix Table 3.

5.3.1.2 Important pathways for breast cancer within and across primary and metastatic cohorts

Examining the statistically significant pathways for breast cancer samples in the TCGA, POG and MET500 cohorts (N = 1100 for TCGA, N = 160 in POG, N = 60 in MET500) revealed features of breast cancer biology (Figure 5.8). The prioritization of estrogen signaling in all three cohorts is consistent with the role of the estrogen receptor and the estrogen signaling pathway in breast cancer progression [65]. The PI3K-Akt pathway has an important role in cell growth and tumour proliferation, and is dysregulated in most common cancers [223]. PIE identified the extracellular matrix (ECM), which plays a vital role in breast cancer progression and metastasis [140], as being relevant to both primary (TCGA) and metastatic (POG, MET500) breast cancers. It also recapitulates recent findings about differing roles of insulin signaling [172] and retinoic acid signaling [41] in breast cancer biology. Interestingly, the Reactome pathway 'Miscellaneous transport and binding events', which contains the AZGP1 gene associated with breast cancer and lipoprotein regulation, was also prioritized in all three cohorts.

Several inflammation-associated pathways were associated only with metastatic breast cancer samples, including IL27 mediated signaling and inflammasome pathways, particularly the NLRP3 inflammasome. These pathways reflect emerging knowledge about the role of the NLRP3 inflammasome and recruitment of myeloid cells through interleukin signaling in metastatic breast cancers [56]. This indicates that the underlying classifier, SCOPE, had learnt the importance of tumour-associated inflammation in this tumour type despite being trained on primary tumours.

While no established support could be found for endochondral ossification and warfarin metabolism (other important pathways in the metastatic group), previous research utilizing the TCGA breast cancers highlighted

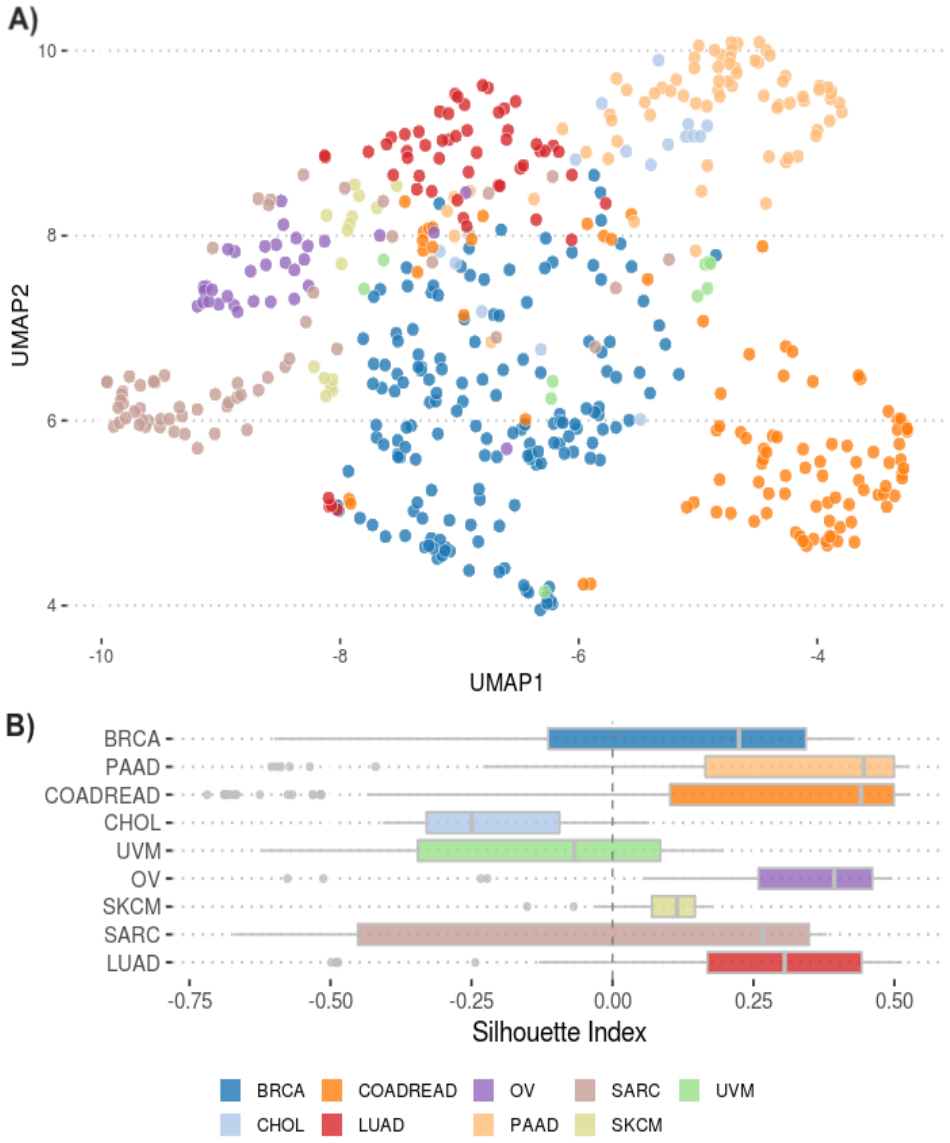

Figure 5.5: Clustering of POG cohort samples by cancer-type using PIE profiles for 3,963 biochemical pathways. Cancer types with at-least 10 (N = 510/602) are shown for ease of readability. A) UMAP projections of pathway profiles are shown. Using pathway importance scores, samples cluster by their diagnosed cancer type. B) Silhouette indices of samples are shown, grouped by cancer type. A positive silhouette index indicates sample clusters with assigned cancer-type.

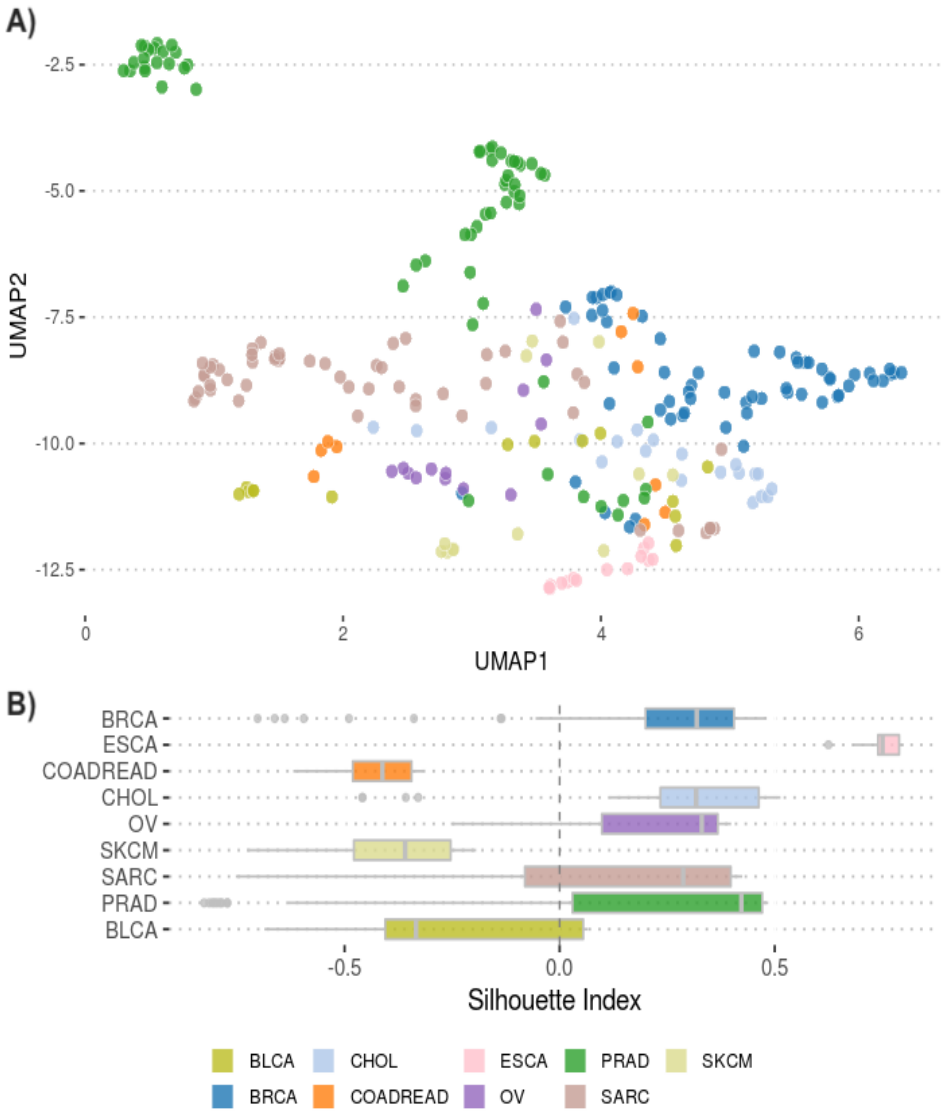

Figure 5.6: Clustering of MET500 cohort samples by cancer-type using PIE profiles for 3,963 biochemical pathways. For ease of readability, the projections only show cancer types with at-least 10 samples in the MET500 cohort (N = 259/375). A) UMAP projections of pathway profiles are shown. Using pathway importance scores, samples cluster by their diagnosed cancer type. Of note, we observe 3 distinct clusters of prostate adenocarcinoma (PRAD, in dark-green). B) Silhouette indices of samples are shown, grouped by cancer type. A positive silhouette index indicates sample clusters with assigned cancer-type.

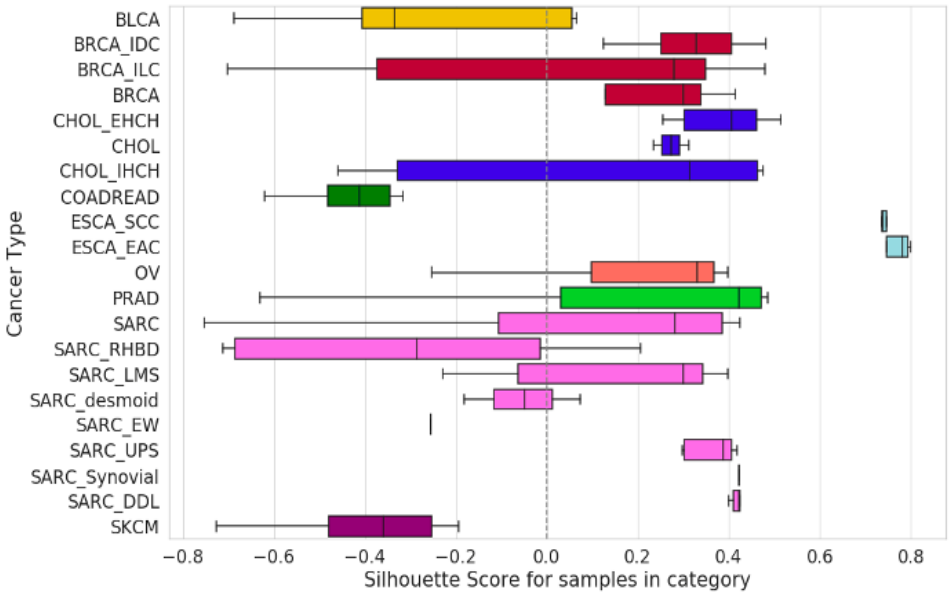

Figure 5.7: Silhouette index spread for the MET500 cohort subtypes. Silhouette metrics are calculated from the UMAP projections initialized with the first two principal components; clusters evaluated based on cancer type annotation. A positive silhouette index indicates sample clusters with assigned cancer-type. Abbreviations: BLCA – Bladder cancer, BRCA – Breast cancer, IDC – invasive ductal carcinoma, ILC – invasive lobular carcinoma, CHOL – Cholangiocarcinoma, EHCH – extrahepatic CHOL, IHCH – intrahepatic CHOL, COADREAD – colorectal adenocarcinoma, ESCA – esophageal carcinoma, SCC – squamous cell carcinoma, EAC – adenocarcinoma, OV – ovarian cancer, PRAD – prostate adenocarcinoma, SARC – sarcoma, RHBD – rhabdoid, LMS – leiomyosarcoma, EW – Ewings Sarcoma, UPS – Undifferentiated pleomorphic carcinoma, DDL – dedifferentiated sarcoma, SKCM – subcutaneous melanoma.

the prognostic value of COL10A1, a key member of the endochondral ossification pathway [227], and dysregulation of Vitamin K pathway genes, which act antagonistically to warfarin [8]. These pathways could indicate potential new directions to better understand breast cancer biology.

5.3.2 Pathway impact scores reveal prostate cancer subgroups

We observed three distinct clusters of prostate adenocarcinoma (PRAD) samples in the UMAP projection of PIE profiles from the MET500 cohort (Figure 5.9a, Figure 5.6a). We compared these groups against a previously published analysis [219] that included 15/62 PRADs analyzed here, and found that the Group-1 samples all harbored fusions associated with prostate cancer. The overlapping Group-2 samples had a high frequency of CDK12 copy loss and TP53 mutations (5/6, where 6/32 had published mutation data in Wu et al. [219]). We further validated these clusters through unsupervised Principal Component Analysis and UMAP decomposition visualizations of the samples' gene expression profiles. We observed separation of the Group-1 and Group-2 in both the principal component analysis and the UMAP of the original gene expression data, suggesting that the observed differences have a biological basis (Figure 5.9b).

For the Group-1 samples with available data (N = 7/20), 100% of the samples harbored ETS fusions (primarily TMRPSS2-ERG) and 4/7 harbored PTEN copy loss and mutations. This was in comparison to only 3/8 samples with ETS fusions and 3/8 samples with PTEN copy loss and mutation in the other two groups combined (N = 8/42). A comparison of the top 25 pathways distinguishing each of the clusters from the other two (Figure 5.10) revealed that the Group-1 samples had a high impact of immune signaling (N = 6/25 pathways) and cell-surface signaling pathways (N = 4/25 pathways), including the T-cell receptor complexes and ECM-receptor interaction respectively [220]. Recent work has suggested a strong association between high fusion burden and immunogenicity in prostate cancer [212]. Additional key oncogenic pathways characterizing this group were PI3K-Akt signaling axis, Notch pathway, and Jak pathway.

The Group-2 samples were strongly driven by high PIE scores for the Celecoxib pathway and calnexin/calreticulin cycle. Celecoxib is a COX-2 inhibitor drug commonly used in treatment of relapsed patients with prostate adenocarcinomas [57]. It inhibits androgen receptor (AR) and

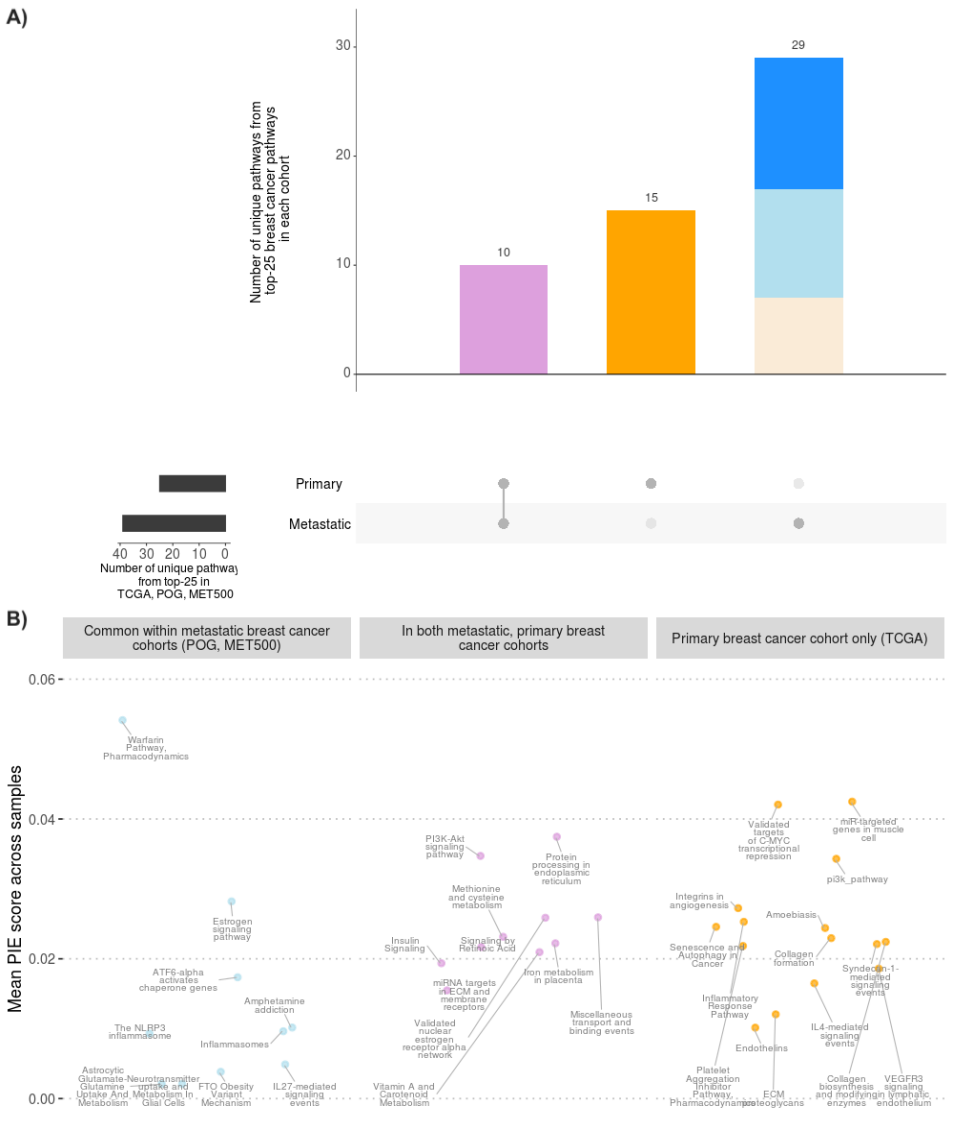

Figure 5.8: Cohort comparison between the top 25 pathways associated with breast cancer, for The Cancer Genome Atlas (TCGA) cohort of primary cancers, the POG cohort of metastatic tumours, and the MET500 cohort of metastatic tumours. Panel A) shows the number of unique and shared pathways between each of the cohorts. The MET500 and POG cohorts are grouped as 'metastatic'. Pathways common between primary and metastatic cancers (in purple), exclusive to primary cancers (in orange), and common within the metastatic cohorts (in light blue) are shown in panel B) with the corresponding mean PIE score across samples on the y-axis.

ErbB signaling; notably, we also observe a relatively higher average pathway importance for various pathways associated with androgen receptor signaling in this group (Appendix Figure 4). Suppression of calnexin by Celecoxib has been observed in cell lines [91], providing rationale for why this pathway was also considered important for this group. Other oncogenic pathways associated with Group-2 included p53 effectors, Wingless-related integration site (WNT) signaling, and MAP kinase pathway. Group-3 was characterized by immune response to viral infections, allografts, and DAP12 signaling.

5.3.3 PIE independently recovers sample-level findings from integrative genomic analysis

We used PIE with the SCOPE classifier to identify prominent pathways for a previously studied rare mammary-like vulvar adenocarcinoma (described in Chapter 2, [72]). This rare tumour initially presented as a poorly differentiated malignancy of the vulva. Subsequent genomic analysis and expression comparison with TCGA tumours determined it to be most similar to a HER2+ breast cancer.

5.3.3.1 Pathways prioritized by PIE overlap with integrative analysis

Integrative manual pathway analysis was performed at the time of case presentation, aggregating all observed changes into important pathways that could explain the oncogenesis and provide potential therapeutic options (Appendix Figure 5). We used PIE to retrospectively identify the most important pathways driving the classification of this tumour as a breast cancer by SCOPE. The top 25 pathways, statistically prioritized through the Grubbs test for outliers, recovered many of the pathways identified through genomic analysis (Figure 5.11). Quantitatively, 48% of the top-25 pathways included genes identified through the manual analysis. This included signal transduction, *ErbB*, *MAPK*, c-*MYC*, and various pathways involved with metabolism. The FOXA1 transcription factor network was also prioritized, supporting observed overexpression of *AR* and *CDKN1B* in the sample (these two genes are important members of this pathway (Belinky et al. [9] - pathway 3538)). The automated approach of PIE in determining the involvement of cancer-driving pathways appeared

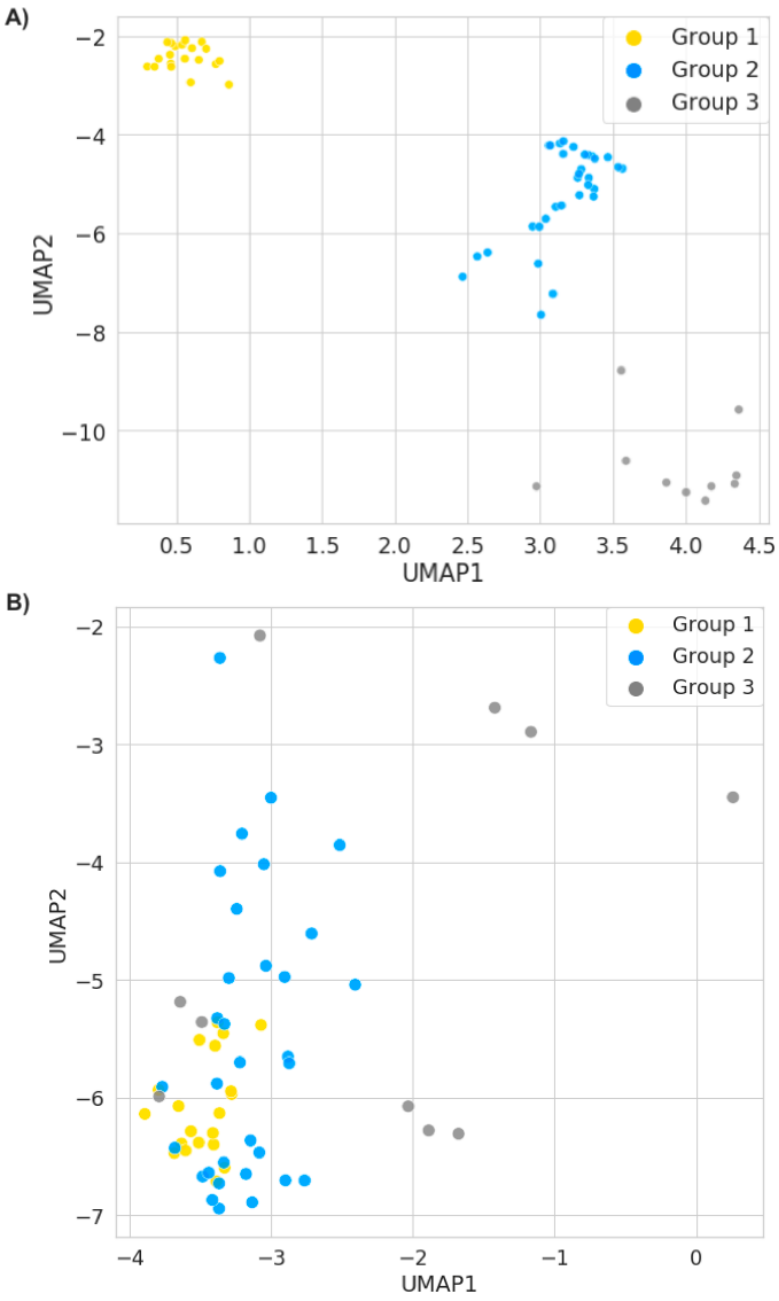

Figure 5.9: UMAP projections of the MET500 cohort are shown, filtered to view only the prostate adenocarcinoma samples. UMAP projections (initialized by the first two principal components) are calculated based on A) sample pathway importance profiles calculated automatically by PIE for 3,963 pathways, and B) gene expression profiles of the samples (RPKM values). Panel B) also suggests a non-random separation of the samples.

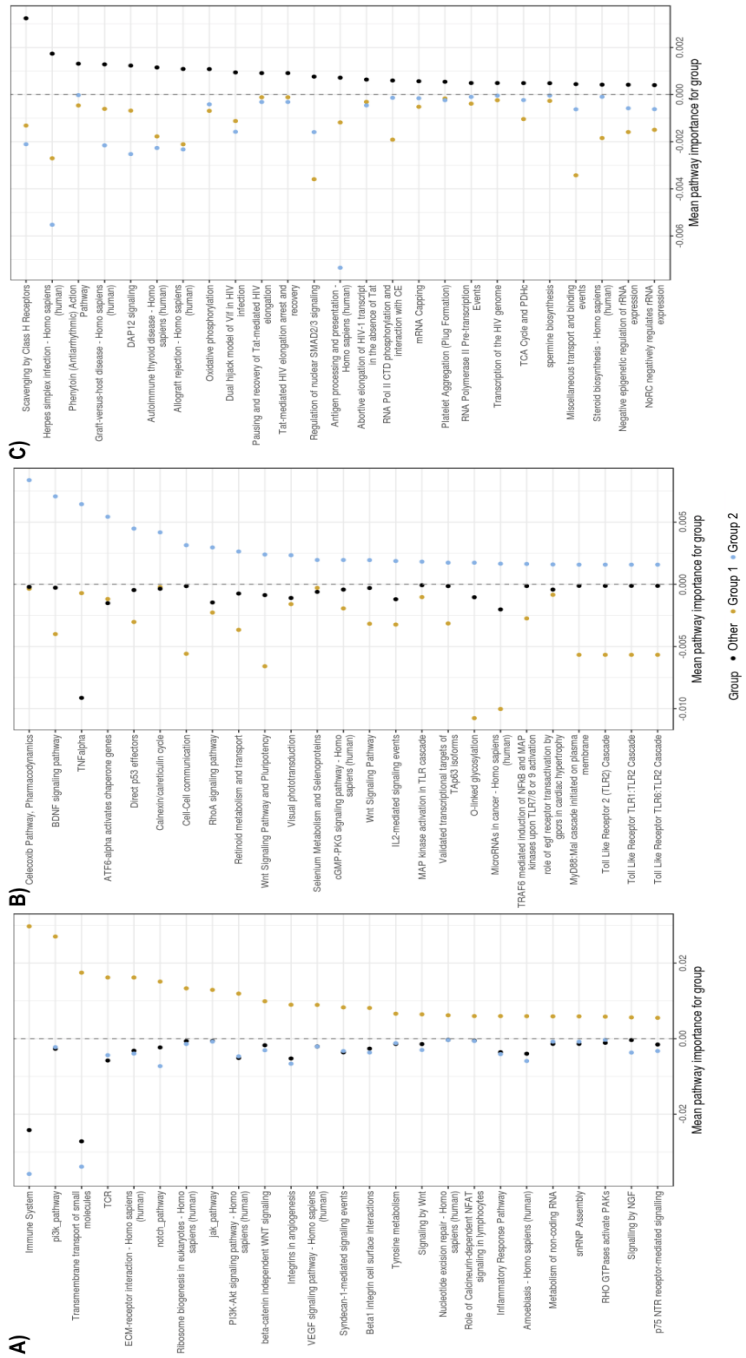

Figure 5.10: Top 25 pathways driving the 3 distinct clusters observed for the prostate adenocarcinomas in the MET500 cohort.

to be consistent with manual expert analyses of the molecular profiles. This offers up the potential of this approach being a relevant starting point for such analyses.

5.3.3.2 PIE discovers pathways associated with paclitaxel treatment

We also observed a high PIE score for endoplasmic reticulum (ER) stress signaling mediated by ATF6, 'Protein processing in endoplasmic reticulum – Homo sapiens (human)', and HTLV-I infection pathways, which were not described in the manual genomic analysis. Review of clinical history revealed that the patient had been treated with paclitaxel, a microtubule stabilizing/ER stress inducer agent used widely for breast cancer treatment, and subsequently developed resistance. A 2013 study has shown association between ER stress response and resistance to paclitaxel through activation of TRAP1 [122]. ATF6 also controls expression of ER chaperone GRP78, which prevents activation of AKT by upstream kinases. AKT3 was under-expressed in the tumour. A recent study found that certain TCGA breast cancers (included in training the underlying classifier) show gene-expression profiles consistent with those of breast cancer cell-lines harboring paclitaxel resistance [29], potentially explaining why SCOPE was able to recognize the ER stress response pathways as important for breast cancer classification.

5.3.4 PIE enables sample-level genomic analysis of cancers with unknown primary

In a previously published study, we described the clinical and genomic presentation of a rare thyroid-like follicular renal cell carcinoma in a 27 year old [105]. The tumour initially presented as a bone metastasis of unknown primary, and no chemotherapy was given prior to genomic sequencing. Using a pairwise expression correlation approach, the tumour sample correlated strongly with both the renal clear cell carcinomas (KIRC) and renal papillary carcinomas (KIRP) from TCGA, while SCOPE strongly indicated this to be similar to KIRP. We retrospectively evaluated the important pathways driving the classification of this case using PIE.

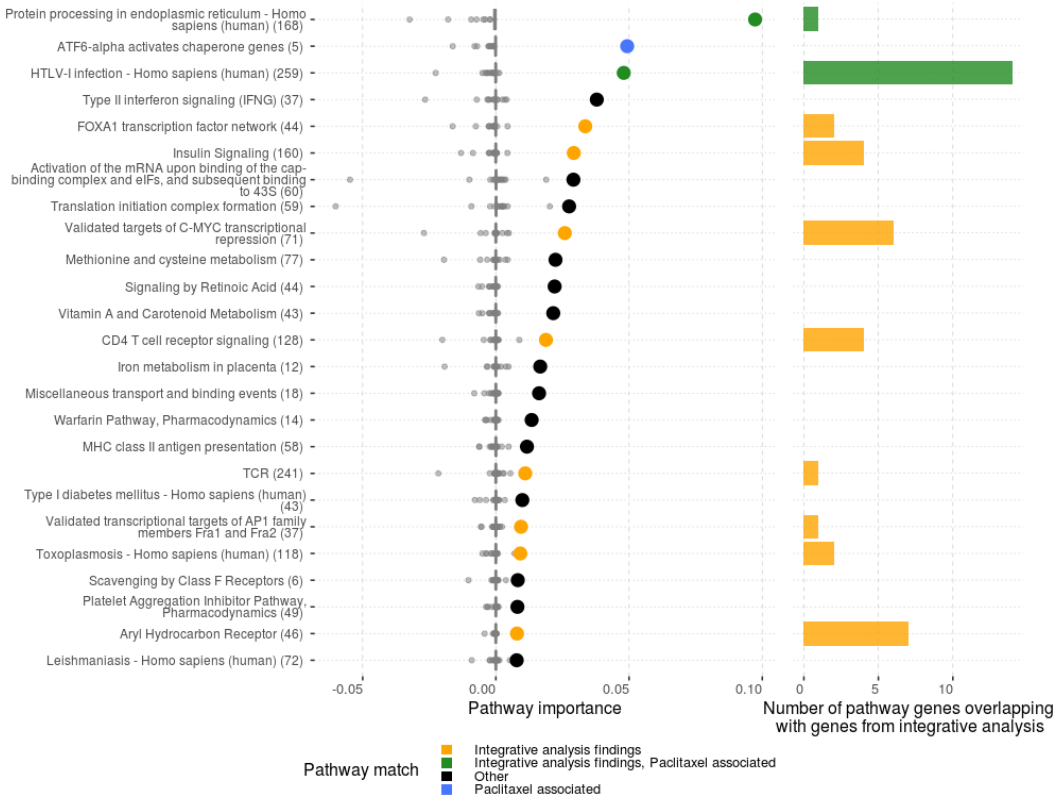

Figure 5.11: Top 25 pathways from PIE-based pathway analysis of a mammary-like vulvar adenocarcinoma. 40% of the pathways shown here overlap with the integrative pathway analysis (in yellow, green), and 16% are associated with paclitaxel therapy that the patient had received previously (in green, blue). Size of pathways is indicated in brackets next to the pathway name on the y-axis. The right panel shows the number of genes shared between the integrative analysis (N = 50) and the indicated pathways. Distribution of PIE scores for the remaining 65 classes is shown in grey, for each pathway.

5.3.4.1 PIE correctly recovers putative driver pathways for rare cancer

Integrative genomic analysis of this tumour at the time of case presentation showed a single copy loss in the tumour suppressor gene *TP53* and aberrant expression or copy changes in *CDK6*, *MYC*, *AR*, *PDGFRA*, *PDGFRB*, and *MAP2K2*. The MAPK pathway, WNT pathway, and cell cycle pathway were identified as putative drivers (Appendix Figure 6).

We used PIE to retrospectively identify the most important pathways driving SCOPE's classification of this tumour as a KIRP. As shown in Figure 5.12, 60% of the top 25 pathways overlapped with the genes and pathways prioritized through manual genomic analysis - specifically the p53, PDGF signaling, and TGF pathway (activator of MAPK pathway among others). Upon filtering for signaling pathways important for oncogenesis, we found a high impact of p53, TGF, and PI3K pathways, consistent with genomic findings and previously known attributes of renal papillary carcinomas [143].

Toll-like receptor pathways were also prioritized by PIE. These pathways are known to be upregulated in follicular thyroid cancers [76], induced by MAPK signaling. The MAPK pathway was also observed to be highly dysregulated in the genomic analysis of this case. Subsequent histologic analysis guided by the genomic alignment of the CUP as KIRP had also classified the cancer as a thyroid-like follicular renal cell carcinoma, lending further biological rationale for the high rank of this pathway from PIE.

5.3.4.2 PIE provides biological insights into subtyping of cancer of unknown origin

Comparison of the resultant pathway scores between the two subtypes using PIE revealed that PI3K/Akt pathway and NRF2 pathway were both important for the classification of the sample as KIRP over KIRC (Figure 5.13). TCGA reported NRF2/ARE pathway mutations in both KIRC and KIRP (3.2%, 9.0% respectively), and frequent mutations in the PI3K/AKT pathway within these subtypes (16.2% of KIRC and 9.8% of KIRP) [168]. Comparison of PIE scores also suggested that compared to the NRF2 pathway, the PI3K signaling pathway has a stronger negative impact on classification of this sample as KIRC.

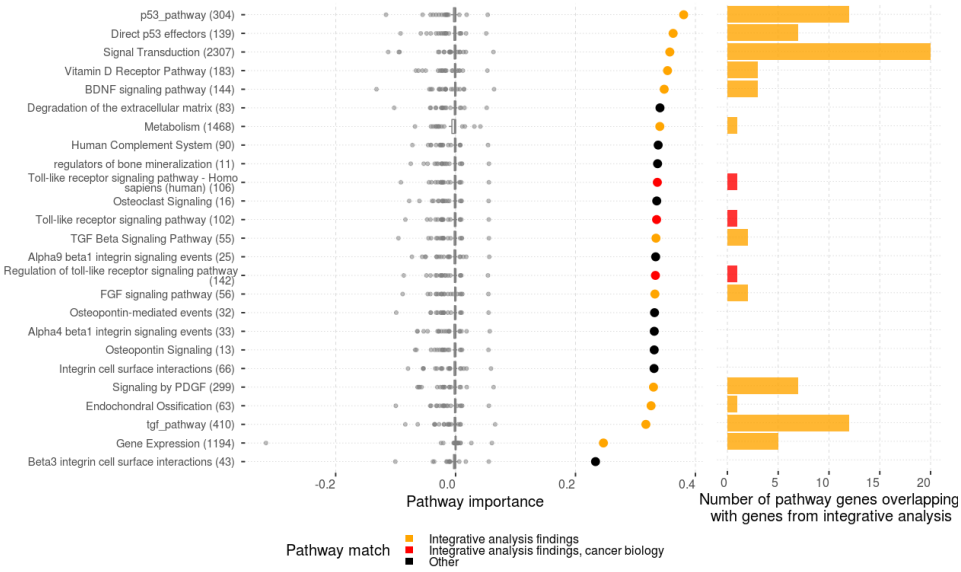

Figure 5.12: Top 25 pathways identified by automated pathway impact analysis using PIE, for a cancer of unknown primary that was later diagnosed as a rare thyroid-like follicular renal cell carcinoma. Size of pathways is indicated in brackets next to the pathway name on the y-axis. Panel on the right shows the number of genes from integrative analysis (N = 34) that overlap with the genes in each of the pathways. 48% of the pathways in the main panel overlap with manual integrative pathway analysis findings (in yellow, red), of which 12% associated with the actual rare cancer type that this cancer represented (in red). Distribution of PIE scores for the remaining 65 output classes is shown in grey, for each pathway.

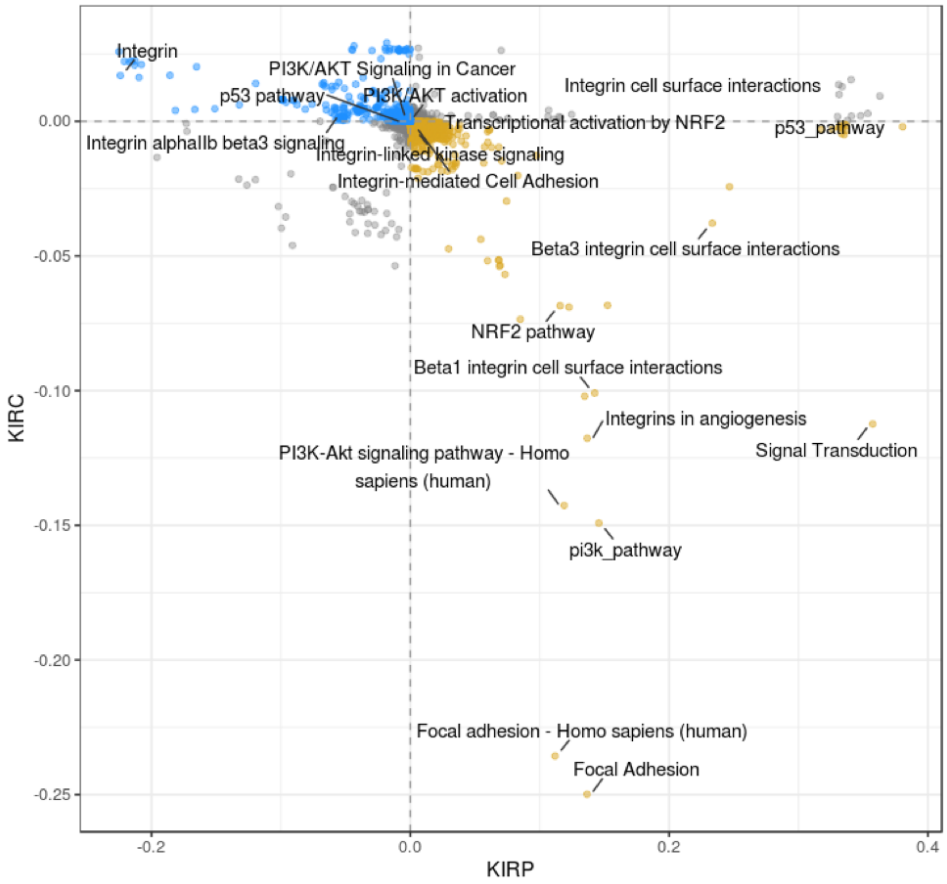

Figure 5.13: Comparison of pathway importance scores for two different output categories – renal clear cell carcinoma (KIRC) and renal papillary carcinoma (KIRP). Scores were calculated by PIE using the SCOPE classifier output. The input was the RNA-Seq profile of a cancer of unknown primary, later diagnosed as a rare follicular renal cell carcinoma that molecularly aligned with KIRP. The pathways that were important for classification of the sample as KIRP instead of KIRC are highlighted in yellow. Pathways important for classification of the sample as KIRC instead of KIRP are shown in blue. As is evident, the magnitude of the pathway importance is higher for pathways driving the classification of KIRP over KIRC. Relevant pathways have been labelled.

5.4 Discussion

PIE performs single-sample pathway analysis using classifiers trained with representative cancer profiles. It recovers biological pathways of relevance to a large set of cancer types. We show that these pathways are relevant to the biology of primary and metastatic cancers, enabling clustering by cancer-type and identifying known biological programs. It also has potential in identifying therapeutic targets in individual cancer samples – we shown examples where it recovers known important pathways in two published case-studies of rare cancers.

We show that PIE recovers biologically relevant pathways of primary cancers from TCGA, and of metastatic cancers from two large cohorts of advanced, post-treatment tumours. Not only do the sample-specific pathway scores allow clustering of the samples by their cancer type, but pathways with high importance for individual cancer types accurately reflect known biological mechanisms of action in the disease. We also identified new sub-groups in a previously studied cohort of prostate adenocarcinomas using sample-clustering patterns observed from PIE scores and validated in the gene expression space. The observed groupings seem to separate based on high immune signaling and previous exposure to Celecoxib. Put together, these findings suggest that PIE also has utility in automated pathway profiling and for discovery of new subtypes.

As we found with the breast cancer case study, PIE was able to identify pathways that were prioritized through independent expert analysis of the genomic and transcriptomic profiles of the patient's tumour. It also identified pathways associated with prior paclitaxel treatment. This is an interesting finding since the classifier that PIE calculated the pathway impact scores from was trained using primary, untreated cancers.

Thyroid-like follicular carcinoma of the kidney (TLFCK) are a rare renal cancer subtype that bear a high histologic resemblance to follicular thyroid carcinomas [112]. In this case, PIE was able to recover findings from manual genomic analysis and flag pathways that could explain this rare cancer's follicular-like histology. While little is known about the characteristic molecular biology of TLFCKs, recent work has indicated that Toll-like receptor signaling is overexpressed in follicular thyroid carcinomas and in KIRPs. PIE independently identified toll-like receptor pathways as a vital component of the tumour's molecular identity.

Here we demonstrate a novel approach for using classifiers to obtain biological insights and identify biological programs that characterize individual cancers. As long as a classifier trained with a sufficiently representative feature set is available, the analysis can be performed at a single-sample level without needing a matching normal or a dataset of cancers against which the pathway changes need to be aligned. Our findings also show that the underlying classifier does not necessarily have to be trained with post-treatment cancers in order to prioritize pathways associated with prior therapy and cancer metastasis. This is a surprising finding, since it suggests that classifiers trained with these large, informative feature representations can simultaneously learn characteristics of the cancers that are not part of the optimization function. The tool for pathway impact evaluation from such classifiers, PIE, is a straightforward algorithm and can be easily repurposed to use different classifiers as the backend, such as random forests, SVMs, or linear classifiers. It is also available as an extension in the original python package for the SCOPE classifier, www.github.com/jasgrewal/cancerscope.

Chapter 6

Conclusions

Genomic analysis of cancers has prompted a new design and implementation paradigm for clinical trials in recent years [11]. The inclusion of genomic profiling in clinical cancer care has shown demonstratable benefits and the generation of molecular data is aiding our understanding of cancer. We are moving towards an era where the molecular experiments driving these efforts become accessible and affordable enough to enable routine clinical usage for advanced cancer patients. For this to happen, the development of technologies that enable quick and insightful interpretation of individual cancer genomes and transcriptomes is essential. In routine cancer management, cancer diagnosis is one common task that has remained manual with limited incorporation of molecular changes. Precision oncology has helped us realise the need to go beyond gross cytologic observations to molecular changes, but the complex task of identifying and assessing important molecular changes has also remained largely manual. We can prioritize diagnostic workflows based on decision charts and pre-select relevant genes based on existing cohort-wide analyses, but both these tasks become severely complicated when presented with rare cancers or cancers that might present as a mix of two or more established histo-types (mixed histology cancers).

The tools presented in this thesis enable better diagnoses based on molecular profiles and expedite personalized cancer analysis efforts by automatically identifying important molecular events driving the cancer classification. We curated the frequency with which personalized oncology initiatives encounter rare cancers or require a review of the initial pathology-based diagnosis (Chapter 3). We found that, as also observed in the literature, RNA-Seq is a common modality driving the diagnosis reviews and changes. In Chapter 4 we developed SCOPE, a pan-cancer classifier that is trained with large feature representations of the cancer samples, and validate it on advanced cancers. The method uses an RNA-Seq profile without additional processing or feature selection, and provides a quantitative

classification decision from across 40 cancer-types in under two minutes. Lastly, in Chapter 5 we showed that by using a pathway impact evaluation tool we developed (PIE), we can identify pathway-level importance scores from these classifier results, automatically generating pathway profiles of advanced cancers on a per-sample basis. Our results provide an impetus for building classifiers with more comprehensive feature representations, instead of pre-selecting features that optimize the desired class separation. They also suggest that a cancer classifier that uses high-dimensional feature representations can learn the biological underpinnings of the classified cancers.

This final chapter will summarize and discuss the implications of the work presented in this thesis. We will review the current limitations of the developed tools and suggest interesting future directions of research to help redress some of these limitations. We will also comment on broader challenges in practice and highlight lessons learn in the process of integrating this work in precision oncology and diagnostic pathology.

6.1 Contributions

6.1.1 Impact of genomic information on diagnosis of advanced cancers

Our assessment of the role of sequencing data in guiding diagnosis changes (Chapter 3) highlights an important component of precision oncology that often goes unreviewed. As per our findings, rigorous evaluation of pathologist-provided diagnoses using molecular data led to changes from histopathologic findings for 2-4% of advanced cancer cases. While forming a small fraction overall, at the individual level these pieces of molecular evidence can severely impact downstream choices for biological comparators and resultant interpretation of genomic findings. Using -omics information to revise diagnosis was particularly relevant in 15 cancers of unknown origin - a rare set of cancers that are typically refractory to routine histopathology analysis.

Our findings from SCOPE (Chapter 4) support observations in literature that physiologically proximal and morphologically similar cancer types - such as stomach adenocarcinomas and esophageal adenocarcinomas arising near the gastric junction - are also quite similar at the transcriptome-wide level

despite having distinct clinical designations [33, 144]. It also reflects the existing challenge with using glass-based pathology (even with the aid of immunohistochemistry) to discern these tumours.

CUPs form 3-5% of all cancer diagnoses, and the identification of a putative primary in these cancers has important implications for the management and treatment of these cancers [120]. Immunohistochemistry based assessment of these cancers is often inconclusive, or yields a wide differential diagnosis. Our method provides an orthogonal approach to narrow down the diagnostic candidates and identify the most likely primary site of such cancers. Since we include adjacent normals as separate classes in our trained ensemble classifier, the system learns to distinguish different tumours, and also identifies expression patterns that are indicative of a normal tissue profile.

6.1.2 Algorithmic advances in cancer classifier development

Computational advances that reduce training time costs for high dimensional datasets, and deep learning algorithms that automatically prioritize relevant features during the training process, are making it even easier to scale computational solutions to broad classification tasks without requiring feature selection. We demonstrate that it is possible to train large neural networks with >17,000 genes as input and maintain a high level of classification accuracy.

Rare cancers form a difficult subset of cancers to try to classify using supervised machine learning methods, simply because of the limited availability of training datasets reflecting these cancer types. To address this issue, we demonstrate the success of using SMOTE as a viable approach to generate additional training samples in these under-represented classes, resulting in improved classification ability compared to using the best-performing models trained on the class-imbalanced data alone. We also observe performance gains for well-represented classes when using rank representations of the large transcriptome inputs.

Feature selection is not always a stable process – the selected features can vary based on the selection dataset, criteria, and algorithms. Biological changes like biomarker conversion can occur in the transition from primary to metastatic disease, for example in the case of HER2/ERBB2 [186]. This can confound results from gene expression panels that were built from a

subset of features discriminating primary tumour types and is particularly detrimental when attempting to diagnose rare cancers where characteristic biological features are still being researched or are ill-defined. We find that SCOPE performs better than neural networks trained with feature sets selected either through pairwise t-tests or curated based on prevalent knowledge about cancer biology.

6.1.3 Interpreting cancer classification decisions and performing single-sample pathway analysis

In routine clinical practice, a cancer diagnosis carries a vast amount of information about cancer biology with it. For example, knowing a breast cancer is HER2+ suggests that pathways associated with the *ErbB* family of transmembrane receptor tyrosine kinases are activated. Using PIE, we extend the inference obtained from SCOPE to automatically identify and provide activity scores for biological pathways driving the classification decision. These pathway representations recapitulate known biology and allow clustering of samples by tumour-type. In the event sub-clusters are identified within a group of samples, we show that we can easily revert to the individual pathway scores and identify the biological rationale behind each sub-group.

PIE analysis also empowers precision oncology efforts by allowing prioritization of pathways reflecting the biology and therapeutically-relevant pathways of a single tumour sample. We demonstrate its utility in a precision oncology workflow with the analysis of two different rare cancer cases. The first was a rare vulvar adenocarcinoma that was found to be more similar to breast cancers than other gynecologic cancers. We had described the actual manual analysis process for this tumour in Chapter 2. PIE was able to recover the oncogenic pathways known to be impacted in this tumour. It also independently uncovered pathways associated strongly with the patient's prior treatment with paclitaxel, a finding that we then validated with a larger cohort of breast cancers. The second case was a cancer of unknown origin, where genomic analysis and application of SCOPE determined a putative diagnosis of thyroid-like follicular renal carcinoma. PIE was again able to pinpoint the known oncogenic pathways associated with this sample in particular, and with this rare cancer-type at large. Since an early point of contention in diagnosing this case had been similarity with another renal cancer subtype, we also compared the pathway scores

between these two renal cancer subtypes. We found evidence of known pathways that made this sample look like the diagnosed subtype over the other, and prioritized other pathways that distinguish the profile of this cancer between the two.

Single-sample analysis is an essential and largely manual component of personalized medicine. The prioritization of driver changes and therapeutic targets remains a complex and vital exercise performed by expert computational biologists. Using PIE we can extrapolate classifier results into interpretable sets of pathways, identify biological programs that distinguish different samples and cohorts, and generate automated biological profiles of advanced cancers on a per-sample basis. The proposed approach is a proof-of-concept for automated interpretation of gene expression data with minimal requirements for expert manual review, comparator datasets, and controls. As additional training datasets become available across other -omics data-types, we expect to scale the scope and depth of interpretation provided by PIE.

6.2 Limitations of developed tools

A limitation of SCOPE is the lack of external validation sets for all classes. A challenge for general application of this method is transcriptomic data that has been generated from RNA extracted from formalin-fixed paraffin-embedded (FFPE) tissue, rather than from snap-frozen tissue. Formalin-fixed paraffin-embedded specimens are persistent morphologic records of tissue biopsies, and highly prevalent in pathology laboratories worldwide. However, controllable and uncontrollable variables, including tissue characteristics, fixation technique, and storage conditions, can affect the yield and quality of total RNA in FFPE blocks. We obtained 100% accuracy on five in-house primary FFPE samples (three primary pancreatic ductal adenocarcinomas, one sarcoma, and one primary colon/esophageal adenocarcinoma), and were able to correctly identify the primary cancer type for all five cases. Nonetheless, FFPE application of this method will require additional validation.

Unlike prior work in the development of cancer diagnostics using molecular profiles, we did not pre-select our testing samples to have uniformly high tumour content. As a consequence, we obtained insights into the role of tumour content in empowering automated cancer diagnostics. Specifically,

we observed that SCOPE has difficulty discriminating metastatic cancers that share the same organ system of origin, if the tumour content of the sequenced sample is low. It is possible that although there is diluted signal for the correct cancer type, low tumour content limits an accurate prediction. Interestingly, based on our observations it appears this bias is only significant in metastatic cancers to the liver (Figures 3.6, 3.7). Other areas where SCOPE performed poorly in pretreated metastases may be driven by known biological differences in the metastatic space (for example, pancreatic cancers). Training classification models with metastatic tumours and examples of advanced cancers is a possible workaround for these problems in the future.

The pathway analysis tool, PIE, does not take into account directional regulatory effects of various genes in a pathway. Nor is it able to propose rationale for prioritized pathways, requiring manual interpretation to determine which prioritized pathways associated with cancer biology versus prior therapy or site of biopsy. The method is also limited in its ability to flag novel or rare cancer-types that may not be present in the core classifier. While this can be overcome with a combination of highly-discriminant classifier and manual expertise, automating the process remains an active area of research.

The output from PIE is a function of the quality of the classifier that is used as the foundation. It can flag biological artefacts output by a given classifier, but the effectiveness of the tool can potentially be impacted if such artefacts dominate the classification results. The impact of sequencing protocols on how different pathways are marked important or otherwise remains another unexplored area in this body of work. Since the core classifier used to demonstrate PIE in these analyses (SCOPE) was trained with RNA-Seq data from poly-adenylated transcripts, we anticipate that the impact of pathways where members of the ribosomal RNA families are overrepresented may not be as easily quantified in RNA-Seq data from ribo-depletion protocols.

6.3 Broader challenges in clinical translation

During the work summarized in this thesis, we came across some common hurdles in data curation, methods validation and application in precision oncology. Some of these relate to clinical practice and diagnostic pathology,

whereas others stem from the need for understanding the rationale behind classification decisions from algorithms. The ability to share tools through simplistic programming libraries and distributed healthcare networks like CanDIG is vital for widespread access, evaluation, and adoption of these methods. The following section summarizes some of the key insights related to these based on our experience. Technologies like single-cell sequencing and methylome profiling offer exciting prospects for sample-level characterisation of tumour biology. Understanding the definition of a 'cancer type' and incorporating cell-of-origin information, when available, will also help improve our ability to identify and characterize new cancer etiologies. These potential areas of future work are also highlighted below.

6.3.1 Management of diagnostic inaccuracies in clinical practice

Dr. Eric Topol, Director of the Scripps Research Translational Institute, identifies certain attributes of doctors that contribute to diagnostic inaccuracy - cognitive bias through representative heuristic (taking shortcuts to decisions based on past experiences), availability bias (diagnosing based on the options "available" to them mentally), overconfidence, and confirmation bias (embracing information supporting one's beliefs and rejecting contradictory information) [198]. Making it even more difficult to recognize these pitfalls is the feedback mechanism - or lack thereof - in modern medicine.

A retrospective study on diagnostic uncertainty was conducted by the Mayo Clinic, evaluating diagnostic agreement between the referral and second opinion diagnoses for 286 patients referred from primary care practitioners [205]. They found that the second opinion diagnosis agreed with the referral physician diagnosis in only 12% of patients. This is incredibly low, but even more concerning is the observation that second opinions are rare enough as is - either due to cost, access, or availability of expert physicians. In cancer management for patients with aggressive, treatment-resistant disease, second opinions can also not be an option due to the extra time needed to identify, communicate, and reach a decision.

Incorporating diagnostic changes from genomic analysis

Obtaining feedback on the accuracy of one's decision is limited in medicine, and it takes years of experience for doctors to gain this perspective [198].

Dr. Michael Lewis, a medical doctor from the University of Toronto, summarises this concisely - "The entire profession had arranged itself as if to confirm the wisdom of its decisions" [111]. Relying on broad-based histopathology assessments as a way to reject granular diagnostic insights is dangerous. It enables practitioners to discount information obtained from molecular analysis and diminishes the demonstratable impact of precision oncology efforts that genuinely advance the standard of care in this area.

Acknowledging diagnostic changes arising from genomic analysis is another challenge that we often observed within the framework of a personalized oncology clinical trial (POG). The creation of a list of suspect diseases, also called a differential diagnosis, is a commonplace clinical practice. By definition, the differential diagnosis facilitates the inclusion of common and rare suspect diagnoses into a single list. This hypothesis-generating method is intended to guide pathologists in their attempt to classify a tumour sample following established histopathologic rules for cancer classification. The presence of a wide range of cancer categories in a differential diagnoses does not indicate that any of them is the final diagnosis. The exclusion of a cancer from the differential diagnosis does not discount its possibility.

We observed an alternative use of the differential diagnoses within POG. In some cases, the genomics-guided diagnosis, while disagreeing with the final diagnosis from the pathologist, overlapped with the differential diagnoses list instead. In such cases, the differential diagnosis list was typically assimilated into the gold-standard for comparison between pathologist diagnosis and genomics-guided diagnosis. While doing so is a reasonable approach for pooling various sources of knowledge together, anecdotally we found that it opened up room for practitioners to repudiate or ignore informative contributions towards diagnosis from genomics. The work presented in Chapter 3 attempted to avoid this bias by including indications for when and how genomic data indicated the final diagnosis. Formalizing this process in other personalized oncology trials would provide valuable evidence-based orthogonal diagnostics and help quantify the potential contributions of genomic analysis to routine diagnostic pathology. It can also support and advance the pathology practice by providing an automated feedback mechanism and evidence-based arguments for the adoption of emerging molecular markers or high-throughput sequencing modalities for complex presentations.

Curating diagnostic changes from genomic analysis

Anecdotally, we observed that the curation of diagnosis changes in precision

oncology efforts faces organizational hurdles and aversion to accepting observations from molecular data. In our grand review of diagnostics in the POG program, we undertook detailed manual review of pathologist notes, tumour board findings, and clinical management system records in order to confidently determine cases that underwent a change in diagnosis or molecular status through genomic analysis. The dispersion of records across multiple sites and organizations is an evident barrier to recording such changes.

6.3.2 Facilitating adoption in routine practice

In practice, physician uptake of multiplexed sequencing approaches for cancer diagnosis remains a challenge. A 2014 review of 160 physicians at the Dana Farber Cancer Institute, an NCI-designated comprehensive cancer center revealed that physicians have low confidence in findings from high-throughput genomic sequencing [71]. The researchers found that high genomic confidence was associated with being a medical oncologist, a researcher, and using available genomic tests more frequently. These observations are consistent with previous studies on the subject. In our experience as well, the gains in diagnostic assessment provided by SCOPE were best evaluated by scientists, pathologists, and oncologists working together. An important driver was the ability to support observed cases of refractory diagnoses with established genomic events associated with the revised histopathology and recognized in clinical oncology.

Curating clinical annotations for improved cancer analysis and management

Findings from genomic analysis can be maximally utilized for patient care if the important molecular changes are linked with patient clinical data such as history of previous treatments and malignancies, duration of previous treatments, and other clinicopathological information [11]. Particularly in case of treatment-resistant cancers, the availability of detailed clinical annotations is severly lacking, consequently impacting the validation and clinical translation of actionable genomic changes. Clinical data recorded over a sustained duration of follow-up with treatment details recorded in harmonized formats and standardized languages can facilitate research efforts across cancer centres, aid text-mining and natural language processing efforts for cancer characterization and significantly improve our ability to utilize genomic data for clinical care in general [11]. A good

model for this is the Genomics Evidence Neoplasia Information Exchange (GENIE), a pan-cancer registry initiative launched in 2016 to link existing and future clinical sequencing efforts (typically panel-based sequencing) with longitudinal patient outcomes to empower clinical decision-making [42]. Replicating this model in research consortia that utilize more comprehensive sequencing data like WGS and RNA-Seq is essential for advancing our understanding of treatment resistance and improving the management of advanced cancers.

Making machine-learning models interpretable

Predictive ML algorithms have certain benefits compared to a trained physician - they can be queried inexhaustibly with no impact on performance, have high consistency between each re-run on the same input, and provide a mathematical representation of the resultant output. A physician can have varying degrees of expertise in couching their clinical decision-making with known facts and reasoning, but they possess the ability to provide a narrative that explains their actions and behaviour [64]. The inability to provide a flow of logic from the input to the output is a fundamental aspect limiting the uptake of present-day machine learning-based diagnostics in clinical practice.

Providing a biological rationale for classification decisions from algorithms can be extremely valuable for analysis of individual cancers and our understanding of cancer in general. In Chapter 4 we extract important genes associated with each classification category of SCOPE using integrated gradients [188]. We show that these genes reflect tumour and heatlhy tissue biology, and overlap with known diagnostic markers of different cancer types. We extend the interpretation beyond individual genes in Chapter 5, using PIE to automatically calculate the impact of groups of genes on classification by SCOPE. Distilling the causative features into pathways enables immediate characterization of individual tumours without the need to manually determine biological associations behind genes important for classification.

Adding interpretability for classification decisions can also reveal the biological pathways influencing anomalous classification decisions. For example, a key pitfall of SCOPE is its inability to confidently resolve diagnoses in liver biopsies with low tumour content (Sections 3.2.2.2 and Section 4.3.4). We hypothesized that contamination from healthy tissue might be leading to this effect. Using PIE we found that for cases where the biopsy site (liver) had confounded prediction, liver-associated pathways

had significantly high PIE scores for hepatocellular carcinoma (LIHC) classification. On the other hand, pathways reflecting the expected tumour biology drove the sample's classification as the correct cancer-type. These observations, while preliminary, enable a unique insight into the rationale for anomalous classification results, and can help characterize the biological implications of low tumour content in bulk RNA-Seq experiments.

6.3.3 Ensuring equitable access to developed tools

Research is not limited to discovery. Technology development from findings is an essential component of innovative research, allowing better access and benefits to tax-payers from discoveries arising through funded projects. The machine-learning methods developed in this project had an immediate benefit for a large fraction of cancers with unknown primary, and for several presentations of advanced cancers refractory to histopathology. At the very minimum, patients were able to obtain a diagnosis - valuable for personal closure, guiding treatment, or a better understanding of a rare etiology - because of their enrollment in the POG trial. How do we ensure access to and support for using such tools for patients in remote communities? How do we enable clinical adoption of these observations into clinical practice?

Precision oncology is a fast-growing area of research, and several bioinformatics tools emerge from these cross-disciplinary investigations with potentially widespread benefit. In our work we have made the developed tools available as plug-and-play python packages, available through GitHub or the traditional 'pip' install option. In the long term, ensuring equal access to the underlying sequencing platforms and resultant analysis tools requires the development of strong collaboration networks across the country with a heavy emphasis on preserving patient privacy. Projects like CanDIG, which aims to develop a national platform for distributed analysis of locally-controlled private genome data, provide the essential infrastructure for enabling equitable access to these tools in urban and rural communities alike. However, currently in Canada there are no clear policies in place to support commercialization or public translation of research software and technologies of value in personalized healthcare.

At present, researchers have limited access to time and resources for developing new technologies, validating them, and ensuring that their use continues beyond the term of the lead graduate student or post-doctoral scholar. Funding mechanisms need to be developed to support the

deployment, maintenance, and updating these technologies in the long run. Institute-level initiatives like the Canada Foundation for Innovation funds are now emerging to support the evolving need of Canadian researchers. Another area for future work is the development of policy frameworks that support researchers beyond the early phases of research and discovery, assisting them in fulfilling the continual requirements of developing and maintaining bioinformatics tools for universal access.

Cancer classifiers are one example of bioinformatics tools that have limited longevity in the research domain. A vast number of published work in the field lack open-access releases of the developed models and become out-of-date quickly as classifications and subtypes with nuanced clinical implications are discovered. Few of the published tools have become commercialized, and the limited set that are open-access are rendered unusable due to deprecated dependencies and deployment platforms. Some of these issues can be addressed by funding policies that encourage and support the maintenance of bioinformatics tools with a demonstratable impact in personalized healthcare. Ensuring continued relevance of the information provided by the tools is another challenge altogether.

6.3.4 Keeping classifiers up-to-date

Training cancer classification algorithms continuously to accommodate edge-cases and out-of-distribution test samples also pose computational and algorithmic challenges. In recent years, promising advances in machine learning have been made that can overcome these issues. Using methods like active learning and reinforcement learning, machine learning models can be trained with new data and effectively learn from their mistakes. Recent work has also suggested improvements in neural networks that enable them to reliably accommodate new examples and classes without 'forgetting the past'. Adopting these best practices for model development can help ensure that classification models remain dynamic and can robustly learn from past mistakes and new data without requiring extensive retraining.

As we found in our analysis, studying the outcomes from machine learning algorithms in the context of known biological artifacts and confounding sources of noise is important before the output from these tools can be taken at face value. Re-training models with examples that reflect these edge-cases is one approach for fixing such issues as low tumour-content and rare cancer-types not included in training the original classifiers. The ability

of these systems to make reliable decisions can only improve as more quality data becomes available and the edge-cases get absorbed into training.

6.3.5 Incorporating other -omics technologies in automated diagnosis

A key limitation of the work presented in this thesis is the limited ability to compare efficacy of mRNA sequencing against other modalities like methylation, miRNA, and lnc-RNA. The precision oncology project used to evaluate the contributions of genomics in cancer diagnosis (Chapter 3) was limited to whole-genome sequencing and poly-adenylated RNA sequencing. The training data used in Chapter 4 and which formed the basis for Chapter 5's analysis was limited to RNA-Seq, partly because this was the sequencing modality for which the most number of cancers had representative data. A multi-omic approach for cancer diagnosis can capture novel subtypes and prognostic groups based on miRNA signatures or methylome profiles, potentially improving the diagnositc ability of associated methods [230]. Emerging work has also shown that methylation profiles captured from circulating tumour DNA can effectively classify brain, gastrointestinal, and gynecologic cancers [82, 117]. Another benefit of alternative sequencing modalities is that they can extend beyond tissue biopsies to any bodily fluid, better capturing the inherent heterogeneity in a tumour, and enabling less invasive early detection and diagnosis for the patient [117].

Till date, these discoveries have been limited in their translational value because of limited analysis focused on primary cancers and due to the absence of large, representative validation datasets for accurate evaluation of generalizability of underlying machine learning models. Existing pan-cancer research in these alternative modalities and multi-omic diagnostics has also been extremely small-scale and limited to common cancer-types like breast cancer, gliomas, and pancreatic cancers. Nevertheless, the increasing exploration and generation of cancer profiles using alternative -omics technologies only serves to increase future availability of such datasets. Going forwards, the emerging body of work in this area will enable us to rigorously train and evaluate multi-omics models for classification of rare cancer-types and metastatic, treatment-resistant cancers as well.

6.3.6 Utilizing single-cell sequencing for interrogation of cancer genomes

In the pathology review (Chapter 3) we observed that compared to whole-genome sequencing, RNA-Seq information was utilized more often in determining the vast majority of revised diagnoses, in alignment with its acceptance as a suitable molecular experiment for cancer diagnostics. Comprehensive, upfront RNA sequencing can eliminate the need for individual assays, providing transcriptome-wide measurements of gene expression [35]. This is particularly attractive for cases where transcriptomic signatures might not have been established (such as rare cancers, cancers of unknown origin) or in retrospective clinical trials [35]. RNA-Seq is also able to connect genotypes (DNA profile) with the resultant phenotypes (cancer subtypes, drug response). It can be used to establish transcriptomic signatures related to observed cell types. It can also profile genetic changes like structural rearrangements, mutations, fusions, and viral integration [35]. As a stand-alone experimental modality, RNA-Seq goes beyond what can be learnt from genetic testing alone, be it through comparative genomic hybridization or DNA sequencing.

So far, bulk short-read sequencing-based genomic analyses have shaped most of our understanding of oncogenesis and treatment resistance in cancer patients. These methods simultaneously profile the tumour and the diverse microenvironment including normal, immune, and stromal components. Deconvolution of these components has been an important but challenging area of research. Over the last decade, single-cell profiling has revealed new and interesting directions beyond bulk tissue sequencing for studying cancers [189]. Cancers are composed of a heterogeneous set of cell types and cell states. The complex interaction of these cells (which can be malignant or otherwise) impacts inter-tumour and intra-tumour heterogeneity, which in turn has implications for treatment selection and prognosis [100, 141]. Single-cell sequencing provides a powerful new approach to understand the fine-grained differences between various cancers.

Combining single-cell sequencing with multiplexed imaging techniques can also provide interesting insights into cancer subtypes with implications on therapeutic strategies. This approach can preserve spatial techniques and help us better understand the role of the microenvironment in tumour progression [173]. Understanding cancer heterogeneity at the single-cell level can also provide a granular view of the differentiation hierarchy

and reveal differences between cells of origin for histologically similar cancer-types [47, 100]. Once pan-cancer datasets become available and the technology itself becomes more robust to noise and drop-out, two immediate areas of utility would be accurate diagnosis of tumours from low tumour content biopsies, and determination of the cell-of-origin and its functional differences in different cancer types.

6.4 Final words

The work presented in this thesis provides proof-of-principle for the utility of machine learning methods in profiling advanced cancers using bulk RNA sequencing data, and embiggens the potential of RNA-Seq as a stand-alone diagnostic and single-sample cancer analysis tool. We find that in practice, when treating advanced cancers, RNA-Seq profiles are used much more frequently than genomic information to determine the site of origin of the cancer. Researchers have long shied away from using high-dimensional profiles for cancer classification, rightly arguing for the negative impact this approach can have on overfitting to noise, poor generalization, and high computational costs. We show that these potential fallouts can be addressed when training with a sufficiently large training set, utilizing synthetic samples to supplement under-represented classes, using appropriate measures like regularization and early stopping to prevent overfitting during training, and leveraging graphical processing units for training models. Recognizing that not all research and healthcare centres will have access to identical bioinformatics pipelines and compute infrastructure, we train the models without extensive feature selection or necessitating any other prior processing.

When a pathologist provides a diagnosis, they are able to associate certain expectations of tumour biology with the assessment, based either on the morphology or based on the findings of ancillary tests like FISH and IHC. This immediate biological interpretation is lacking in present-day computational classifiers used for cancer diagnosis. The validation and interpretation of the diagnosis from molecular data is another challenging task for precision oncology. Understanding the molecular changes characterizing each individual tumour has been a manually-intensive downstream task, that uses the diagnosis to make comparisons and identify important molecular events. We address these

interpretability problems by developing PIE. We show that this approach can help delineate the biology, prior exposures, and therapeutic targets for individual patients. When taken at scale in large cancer cohorts, we uncover novel subtypes and provide biological rationale for such groupings. We make the resultant tools available for common use through GitHub and Zenodo. It is reasonable to suspect that the usage of such machine learning-based interpretable diagnostics in clinical care will have to become part of best practices if these tools continue being demonstrably better than human-level assessment and become accessible at low cost [64].

Bibliography

[1] Abbott, J. J. and Ahmed, I. (2006). Adenocarcinoma of mammary-like glands of the vulva: report of a case and review of the literature. *The American journal of dermatopathology*, 28(2):127–133.

[2] Adetiba, E. and Olugbara, O. O. (2015). Improved classification of lung cancer using radial basis function neural network with affine transforms of voss representation. *PloS one*, 10(12):e0143542.

[3] Agwa, E. and Ma, P. C. (2013). Overview of various techniques/platforms with critical evaluation of each. *Current treatment options in oncology*, 14(4):623–633.

[4] Ahmed, A. A. and Abedalthagafi, M. (2016). Cancer diagnostics: the journey from histomorphology to molecular profiling. *Oncotarget*, 7(36):58696.

[5] Alligood-Percoco, N. R., Kessler, M. S., and Willis, G. (2015). Breast cancer metastasis to the vulva 20 years remote from initial diagnosis: a case report and literature review. *Gynecologic oncology reports*, 13:33.

[6] Alshareeda, A. T., Al-Sowayan, B. S., Alkharji, R. R., Aldosari, S. M., et al. (2020). Cancer of unknown primary site: Real entity or misdiagnosed disease? *Journal of Cancer*, 11(13):3919.

[7] Bass, B. P., Engel, K. B., Greytak, S. R., and Moore, H. M. (2014). A review of preanalytical factors affecting molecular, protein, and morphological analysis of formalin-fixed, paraffin-embedded (ffpe) tissue: how well do you know your ffpe specimen? *Archives of pathology and laboratory medicine*, 138(11):1520–1530.

[8] Beaudin, S., Kokabee, L., and Welsh, J. (2019). Divergent effects of vitamins k1 and k2 on triple negative breast cancer cells. *Oncotarget*, 10(23):2292.

[9] Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran,

M., and Lancet, D. (2015). Pathcards: multi-source consolidation of human biological pathways. *Database*, 2015.

[10] Bender, R. A. and Erlander, M. G. (2009). Molecular classification of unknown primary cancer. In *Seminars in oncology*, volume 36, pages 38–43. Elsevier.

[11] Berger, M. F. and Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology*, 15(6):353–365.

[12] Blechacz, B., Komuta, M., Roskams, T., and Gores, G. J. (2011). Clinical diagnosis and staging of cholangiocarcinoma. *Nature reviews. Gastroenterology & hepatology*, 8(9):512–22.

[13] Bloom, G., Yang, I. V., Boulware, D., Kwong, K. Y., Coppola, D., Eschrich, S., Quackenbush, J., and Yeatman, T. J. (2004). Multi-platform, multi-site, microarray-based human tumor classification. *The American journal of pathology*, 164(1):9–16.

[14] Board, P. A. T. E. (2017). Vulvar cancer treatment (pdq®): Health professional version. *PDQ*.

[15] Boots-Sprenger, S. H., Sijben, A., Rijntjes, J., Tops, B. B., Idema, A. J., Rivera, A. L., Bleeker, F. E., Gijtenbeek, A. M., Diefes, K., Heathcock, L., et al. (2013). Significance of complete 1p/19q co-deletion, idh1 mutation and mgmt promoter methylation in gliomas: use with caution. *Modern Pathology*, 26(7):922–929.

[16] Boyce, B. (2015). Whole slide imaging: uses and limitations for surgical pathology and teaching. *Biotechnic & Histochemistry*, 90(5):321–330.

[17] Brcic, L., Vlacic, G., Quehenberger, F., and Kern, I. (2018). Reproducibility of malignant pleural mesothelioma histopathologic subtyping. *Archives of pathology & laboratory medicine*, 142(6):747–752.

[18] Brose, M. S., Cabanillas, M. E., Cohen, E. E., Wirth, L. J., Riehl, T., Yue, H., Sherman, S. I., and Sherman, E. J. (2016). Vemurafenib in patients with brafv600e-positive metastatic or unresectable papillary thyroid cancer refractory to radioactive iodine: a non-randomised, multicentre, open-label, phase 2 trial. *The Lancet Oncology*, 17(9):1272–1282.

[19] Brown, H. M. and Wilkinson, E. J. (2002). Uroplakin-iii to distinguish

primary vulvar paget disease from paget disease secondary to urothelial carcinoma. *Human pathology*, 33(5):545–548.

[20] Bueno, R., Stawiski, E. W., Goldstein, L. D., Durinck, S., De Rienzo, A., Modrusan, Z., Gnad, F., Nguyen, T. T., Jaiswal, B. S., Chirieac, L. R., et al. (2016). Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nature genetics*, 48(4):407.

[21] Butler, B., Leath III, C. A., and Barnett, J. C. (2014). Primary invasive breast carcinoma arising in mammary-like glands of the vulva managed with excision and sentinel lymph node biopsy. *Gynecologic oncology case reports*, 7:7.

[22] Butterfield, Y. S., Kreitzman, M., Thiessen, N., Corbett, R. D., Li, Y., Pang, J., Ma, Y. P., Jones, S. J., and Birol, I. (2014). Jaguar: junction alignments to genome for rna-seq reads. *PloS one*, 9(7):e102398.

[23] Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., and Craig, D. W. (2016). Translating rna sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 17(5):257.

[24] Carlson, J. J. and Roth, J. A. (2013). The impact of the oncotype dx breast cancer assay in clinical practice: a systematic review and meta-analysis. *Breast cancer research and treatment*, 141(1):13–22.

[25] Chaffer, C. L. and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *science*, 331(6024):1559–1564.

[26] Chahal, M., Pleasance, E., Grewal, J., Zhao, E., Ng, T., Chapman, E., Jones, M. R., Shen, Y., Mungall, K. L., Bonakdar, M., et al. (2018). Personalized oncogenomic analysis of metastatic adenoid cystic carcinoma: using whole-genome sequencing to inform clinical decision-making. *Molecular Case Studies*, 4(2):a002626.

[27] Chauhan, A., Farooqui, Z., Silva, S. R., Murray, L. A., Hodges, K. B., Yu, Q., Myint, Z. W., Raajesekar, A. K., Weiss, H., Arnold, S., et al. (2019). Integrating a 92-gene expression analysis for the management of neuroendocrine tumors of unknown primary. *Asian Pacific journal of cancer prevention: APJCP*, 20(1):113.

[28] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.

(2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[29] Chen, J., Ge, X., Zhang, W., Ding, P., Du, Y., Wang, Q., Li, L., Fang, L., Sun, Y., Zhang, P., et al. (2020). Pi3k/akt inhibition reverses r-chop resistance by destabilizing sox2 in diffuse large b cell lymphoma. *Theranostics*, 10(7):3151.

[30] Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z. Y., Won, H. H., Scott, S. N., et al. (2015). Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (msk-impact): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *The Journal of molecular diagnostics*, 17(3):251–264.

[31] Cheng, L., Lopez-Beltran, A., Massari, F., MacLennan, G. T., and Montironi, R. (2018). Molecular testing for braf mutations to inform melanoma treatment decisions: a move toward precision medicine. *Modern Pathology*, 31(1):24.

[32] Cheng, L., Zhang, S., Wang, L., MacLennan, G. T., and Davidson, D. D. (2017). Fluorescence in situ hybridization in surgical pathology: principles and applications. *The Journal of Pathology: Clinical Research*, 3(2):73–99.

[33] Cherniack, A. D., Shen, H., Walter, V., Stewart, C., Murray, B. A., Bowlby, R., Hu, X., Ling, S., Soslow, R. A., Broaddus, R. R., et al. (2017). Integrated molecular characterization of uterine carcinosarcoma. *Cancer cell*, 31(3):411–423.

[34] Chu, J., Sadeghi, S., Raymond, A., Jackman, S. D., Nip, K. M., Mar, R., Mohamadi, H., Butterfield, Y. S., Robertson, A. G., and Birol, I. (2014). Biobloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. *Bioinformatics*, 30(23):3402–3404.

[35] Cieślik, M. and Chinnaiyan, A. M. (2018). Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2):93.

[36] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

[37] Clark, A. M., Ma, B., Taylor, D. L., Griffith, L., and Wells, A. (2016). Liver metastases: Microenvironments and ex-vivo models. *Experimental Biology and Medicine*, 241(15):1639–1652.

[38] Clark, B. Z., Beriwal, S., Dabbs, D. J., and Bhargava, R. (2014). Semiquantitative gata-3 immunoreactivity in breast, bladder, gynecologic tract, and other cytokeratin 7–positive carcinomas. *American Journal of Clinical Pathology*, 142(1):64–71.

[39] Clegg, L. X., Feuer, E. J., Midthune, D. N., Fay, M. P., and Hankey, B. F. (2002). Impact of reporting delay and reporting error on cancer incidence rates and trends. *Journal of the National Cancer Institute*, 94(20):1537–1545.

[40] Connolly, J. L., Schnitt, S. J., Wang, H. H., Longtine, J. A., Dvorak, A., and Dvorak, H. F. (2003). Role of the surgical pathologist in the diagnosis and management of the cancer patient. In *Holland-Frei Cancer Medicine. 6th edition*. BC Decker.

[41] Connolly, R. M., Nguyen, N. K., and Sukumar, S. (2013). Molecular pathways: current role and future directions of the retinoic acid pathway in cancer prevention and treatment. *Clinical Cancer Research*, 19(7):1651–1659.

[42] Consortium, A. P. G. et al. (2017). Aacr project genie: powering precision medicine through an international consortium. *Cancer discovery*, 7(8):818–831.

[43] Consortium, I. C. G. et al. (2010). International network of cancer genome projects. *Nature*, 464(7291):993.

[44] Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., et al. (2015). Pathway and network analysis of cancer genomes. *Nature methods*, 12(7):615.

[45] Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., et al. (2014). The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477.

[46] Dabney, A., Storey, J. D., and Warnes, G. (2010). qvalue: Q-value estimation for false discovery rate control. *R package version*, 1(0).

[47] Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., et al. (2017). Single-cell rna-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell reports*, 21(5):1399–1410.

[48] De Leon, E. D., Carcangiu, M. L., Prieto, V. G., McCue, P. A., Burchette, J. L., To, G., Norris, B. A., Kovatich, A. J., Sanchez, R. L., Krigman, H. R., et al. (2000). Extramammary paget disease is characterized by the consistent lack of estrogen and progesterone receptors but frequently expresses androgen receptor. *American journal of clinical pathology*, 113(4):572–575.

[49] Del Pino, M., Rodriguez-Carunchio, L., and Ordi, J. (2013). Pathways of vulvar intraepithelial neoplasia and squamous cell carcinoma. *Histopathology*, 62(1):161–175.

[50] Dennis, J. L., Vass, J. K., Wit, E. C., Keith, W. N., and Oien, K. A. (2002). Identification from public data of molecular markers of adenocarcinoma characteristic of the site of origin. *Cancer research*, 62(21):5999–6005.

[51] DeVita, V. T., Rosenberg, S. A., and Lawrence, T. S. (2018). *DeVita, Hellman, and Rosenberg's cancer*. Lippincott Williams & Wilkins.

[52] Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M. A., Condon, A., et al. (2011). Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics*, 28(2):167–175.

[53] El-Deiry, W. S., Goldberg, R. M., Lenz, H.-J., Shields, A. F., Gibney, G. T., Tan, A. R., Brown, J., Eisenberg, B., Heath, E. I., Phuphanich, S., et al. (2019). The current state of molecular testing in the treatment of patients with solid tumors, 2019. *CA: a cancer journal for clinicians*.

[54] Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A. N., Nelson, H. D., Pepe, M. S., Allison, K. H., Schnitt, S. J., et al. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132.

[55] Erlander, M. G., Ma, X.-J., Kesty, N. C., Bao, L., Salunga, R., and Schnabel, C. A. (2011). Performance and clinical evaluation of the 92-gene real-time pcr assay for tumor classification. *The Journal of Molecular Diagnostics*, 13(5):493–503.

[56] Ershaid, N., Sharon, Y., Doron, H., Raz, Y., Shani, O., Cohen, N., Monteran, L., Leider-Trejo, L., Ben-Shmuel, A., Yassin, M., et al. (2019). Nlrp3 inflammasome in fibroblasts links tissue damage with inflammation in breast cancer progression and metastasis. *Nature communications*, 10(1):1–15.

[57] Etheridge, T., Liou, J., Downs, T. M., Abel, E. J., Richards, K. A., and Jarrard, D. F. (2018). The impact of celecoxib on outcomes in advanced prostate cancer patients undergoing androgen deprivation therapy. *American journal of clinical and experimental urology*, 6(3):123.

[58] Ettinger, D. S., Agulnik, M., Cates, J. M. M., Cristea, M., Denlinger, C. S., Eaton, K. D., Fidias, P. M., Gierada, D., Gockerman, J. P., Handorf, C. R., Iyer, R., Lenzi, R., Phay, J., Rashid, A., Saltz, L., Shulman, L. N., Smerage, J. B., Varadhachary, G. R., Zager, J. S., Zhen, W. K., and National Comprehensive Cancer Network (2011). NCCN Clinical Practice Guidelines Occult primary. *Journal of the National Comprehensive Cancer Network : JNCCN*, 9(12):1358–95.

[59] Evens, A. M., Kanakry, J. A., Sehn, L. H., Kritharis, A., Feldman, T., Kroll, A., Gascoyne, R. D., Abramson, J. S., Petrich, A. M., Hernandez-Ilizaliturri, F. J., Al-Mansour, Z., Adeimy, C., Hemminger, J., Bartlett, N. L., Mato, A., Caimi, P. F., Advani, R. H., Klein, A. K., Nabhan, C., Smith, S. M., Fabregas, J. C., Lossos, I. S., Press, O. W., Fenske, T. S., Friedberg, J. W., Vose, J. M., and Blum, K. A. (2015). Gray zone lymphoma with features intermediate between classical Hodgkin lymphoma and diffuse large B-cell lymphoma: Characteristics, outcomes, and prognostication among a large multicenter cohort. *American Journal of Hematology*, 90(9):778–783.

[60] Fizazi, K., Greco, F., Pavlidis, N., Daugaard, G., Oien, K., and Pentheroudakis, G. (2015a). Cancers of unknown primary site: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 26(suppl_5):v133–v138.

[61] Fizazi, K., Greco, F. A., Pavlidis, N., Daugaard, G., Oien, K., and Pentheroudakis, G. (2015b). Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO*, 26:v133–v138.

[62] Fodde, R. (2002). The apc gene in colorectal cancer. *European journal of cancer*, 38(7):867–871.

[63] Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., et al. (2017). Cosmic: somatic cancer genetics at high-resolution. *Nucleic acids research*, 45(D1):D777–D783.

[64] Froomkin, A. M., Kerr, I., and Pineau, J. (2019). When ais outperform doctors: confronting the challenges of a tort-induced over-reliance on machine learning. *Ariz. L. Rev.*, 61:33.

[65] Frost, A. R., Hurst, D. R., Shevde, L. A., and Samant, R. S. (2012). The influence of the cancer microenvironment on the process of metastasis. *International journal of breast cancer*, 2012.

[66] Garrison Jr, L. P., Babigumira, J. B., Masaquel, A., Wang, B. C., Lalla, D., and Brammer, M. (2015). The lifetime economic burden of inaccurate her2 testing: estimating the costs of false-positive and false-negative her2 test results in us patients with early-stage breast cancer. *Value in Health*, 18(4):541–546.

[67] Gatta, G., Van Der Zwan, J. M., Casali, P. G., Siesling, S., Dei Tos, A. P., Kunkler, I., Otter, R., Licitra, L., Mallone, S., Tavilla, A., et al. (2011). Rare cancers are not so rare: the rare cancer burden in europe. *European journal of cancer*, 47(17):2493–2511.

[68] Good, B. M., Ainscough, B. J., McMichael, J. F., Su, A. I., and Griffith, O. L. (2014). Organizing knowledge to enable personalization of medicine in cancer. *Genome biology*, 15(8):438.

[69] Graber, M. L. (2013). The incidence of diagnostic error in medicine. *BMJ Qual Saf*, 22(Suppl 2):ii21–ii27.

[70] Graber, M. L., Wachter, R. M., and Cassel, C. K. (2012). Bringing diagnosis into the quality and safety equations. *Jama*, 308(12):1211–1212.

[71] Gray, S. W., Hicks-Courant, K., Cronin, A., Rollins, B. J., and Weeks, J. C. (2014). Physicians' attitudes about multiplex tumor genomic testing. *Journal of Clinical Oncology*, 32(13):1317.

[72] Grewal, J. K., Eirew, P., Jones, M., Chiu, K., Tessier-Cloutier, B., Karnezis, A. N., Karsan, A., Mungall, A., Zhou, C., Yip, S., et al. (2017). Detection and genomic characterization of a mammary-like adenocarcinoma. *Molecular Case Studies*, 3(6):a002170.

[73] Grewal, J. K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., Mungall, A. J., Zhao, Y., Taylor, M. D., Gelmon, K., et al.

(2019). Application of a neural network whole transcriptome–based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA network open*, 2(4):e192597–e192597.

[74] Gröschel, S., Bommer, M., Hutter, B., Budczies, J., Bonekamp, D., Heining, C., Horak, P., Fröhlich, M., Uhrig, S., Hübschmann, D., et al. (2016). Integration of genomics and histology revises diagnosis and enables effective therapy of refractory cancer of unknown primary with pdl1 amplification. *Molecular Case Studies*, 2(6):a001180.

[75] Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., et al. (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome research*, 22(10):1995–2007.

[76] Hagström, J., Heikkilä, A., Siironen, P., Louhimo, J., Heiskanen, I., Mäenpää, H., Arola, J., and Haglund, C. (2012). Tlr-4 expression and decrease in chronic inflammation: indicators of aggressive follicular thyroid carcinoma. *Journal of clinical pathology*, 65(4):333–338.

[77] Hainsworth, J. D. and Greco, F. A. (2014). Gene expression profiling in patients with carcinoma of unknown primary site: from translational research to standard of care. *Virchows Archiv*, 464(4):393–402.

[78] Hajdu, S. I. (2006). Thoughts about the cause of cancer. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 106(8):1643–1649.

[79] Hajdu, S. I. (2011). Microscopic contributions of pioneer pathologists. *Annals of Clinical & Laboratory Science*, 41(2):201–206.

[80] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

[81] Handorf, C. R., Kulkarni, A., Grenert, J. P., Weiss, L. M., Rogers, W. M., Kim, O. S., Monzon, F. A., Halks-Miller, M., Anderson, G. G., Walker, M. G., et al. (2013). A multicenter study directly comparing the diagnostic accuracy of gene expression profiling and immunohistochemistry for primary site identification in metastatic tumors. *The American journal of surgical pathology*, 37(7):1067.

[82] Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., Flagg, K., Hou, J., Zhang, H., Yi, S., et al. (2017). Dna methylation markers for

diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences*, 114(28):7414–7419.

[83] Hartung, E. (1875). *Über einen Fall von Mamma accessoria*. E. Th. Jacob.

[84] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[85] Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, 6(12):e28210.

[86] Hayashi, H., Kurata, T., Takiguchi, Y., Arai, M., Takeda, K., Akiyoshi, K., Matsumoto, K., Onoe, T., Mukai, H., Matsubara, N., et al. (2019). Randomized phase ii trial comparing site-specific treatment based on gene expression profiling with carboplatin and paclitaxel for patients with cancer of unknown primary site. *Journal of Clinical Oncology*, 37(7):570–579.

[87] Herschkowitz, J. I., He, X., Fan, C., and Perou, C. M. (2008). The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal b breast carcinomas. *Breast Cancer Research*, 10(5):R75.

[88] Herter-Sprie, G. S., Greulich, H., and Wong, K.-K. (2013). Activating mutations in erbb2 and their impact on diagnostics and treatment. *Frontiers in oncology*, 3:86.

[89] Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.

[90] Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944.

[91] Huang, K.-H., Kuo, K.-L., Chen, S.-C., Weng, T.-I., Chuang, Y.-T., Tsai, Y.-C., Pu, Y.-S., Chiang, C.-K., and Liu, S.-H. (2012). Down-regulation of glucose-regulated protein (grp) 78 potentiates

cytotoxic effect of celecoxib in human urothelial carcinoma cells. *PLoS One*, 7(3).

[92] Huang, P.-J., Chiu, L.-Y., Lee, C.-C., Yeh, Y.-M., Huang, K.-Y., Chiu, C.-H., and Tang, P. (2018). msignaturedb: a database for deciphering mutational signatures in human cancers. *Nucleic acids research*, 46(D1):D964–D970.

[93] Hylebos, M., Van Camp, G., van Meerbeeck, J. P., and de Beeck, K. O. (2016). The genetic landscape of malignant pleural mesothelioma: results from massively parallel sequencing. *Journal of Thoracic Oncology*, 11(10):1615–1626.

[94] Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J.-Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A., et al. (2015). Vemurafenib in multiple nonmelanoma cancers with braf v600 mutations. *New England Journal of Medicine*, 373(8):726–736.

[95] Jamshidi, F., Pleasance, E., Li, Y., Shen, Y., Kasaian, K., Corbett, R., Eirew, P., Lum, A., Pandoh, P., Zhao, Y., et al. (2014). Diagnostic value of next-generation sequencing in an unusual sphenoid tumor. *The oncologist*, 19(6):623.

[96] Jones, S. J., Laskin, J., Li, Y. Y., Griffith, O. L., An, J., Bilenky, M., Butterfield, Y. S., Cezard, T., Chuah, E., Corbett, R., Fejes, A. P., Griffith, M., Yee, J., Martin, M., Mayo, M., Melnyk, N., Morin, R. D., Pugh, T. J., Severson, T., Shah, S. P., Sutcliffe, M., Tam, A., Terry, J., Thiessen, N., Thomson, T., Varhol, R., Zeng, T., Zhao, Y., Moore, R. A., Huntsman, D. G., Birol, I., Hirst, M., Holt, R. A., and Marra, M. A. (2010). Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome biology*, 11(8):R82.

[97] Ju, Y. S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L. B., Rahbari, R., Wedge, D. C., Davies, H. R., Ramakrishna, M., Fullam, A., et al. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714.

[98] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). Consensuspathdb: toward a more complete picture of cell biology. *Nucleic acids research*, 39(suppl_1):D712–D717.

[99] Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

[100] Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., Specht, M. C., Bernstein, B. E., Michor, F., and Ellisen, L. W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in tnbc through single-cell rna-seq. *Nature communications*, 9(1):1–10.

[101] Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125.

[102] Kazakov, D. V., Spagnolo, D. V., Kacerovska, D., and Michal, M. (2011). Lesions of anogenital mammary-like glands: an update. *Advances in anatomic pathology*, 18(1):1–28.

[103] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673.

[104] Kim, A. and Cohen, M. S. (2016). The discovery of vemurafenib for the treatment of braf-mutated metastatic melanoma. *Expert opinion on drug discovery*, 11(9):907–916.

[105] Ko, J. J., Grewal, J. K., Ng, T., Lavoie, J.-M., Thibodeau, M. L., Shen, Y., Mungall, A. J., Taylor, G., Schrader, K. A., Jones, S. J., et al. (2018). Whole-genome and transcriptome profiling of a metastatic thyroid-like follicular renal cell carcinoma. *Molecular Case Studies*, 4(6):a003137.

[106] Kobak, D. and Linderman, G. C. (2019). Umap does not preserve global structure any better than t-sne when using the same initialization. *bioRxiv*.

[107] Kopetz, S., Desai, J., Chan, E., Hecht, J. R., O'Dwyer, P. J., Maru, D., Morris, V., Janku, F., Dasari, A., Chung, W., et al. (2015). Phase ii pilot study of vemurafenib in patients with metastatic braf-mutated colorectal cancer. *Journal of clinical oncology*, 33(34):4032.

[108] Laskin, J., Jones, S., Aparicio, S., Chia, S., Ch'ng, C., Deyell, R., Eirew, P., Fok, A., Gelmon, K., Ho, C., Huntsman, D., Jones, M., Kasaian, K., Karsan, A., Leelakumari, S., Li, Y., Lim, H., Ma, Y., Mar, C., Martin, M., Moore, R., Mungall, A., Mungall, K., Pleasance, E., Rassekh, S. R., Renouf, D., Shen, Y., Schein, J., Schrader, K., Sun, S., Tinker, A., Zhao, E., Yip, S., and Marra, M. A. (2015). Lessons learned from the application

of whole-genome analysis to the treatment of patients with advanced cancers. *Cold Spring Harbor molecular case studies*, 1(1):a000570.

[109] Le Tourneau, C., Delord, J.-P., Gonçalves, A., Gavoille, C., Dubot, C., Isambert, N., Campone, M., Trédan, O., Massiani, M.-A., Mauborgne, C., et al. (2015). Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (shiva): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The lancet oncology*, 16(13):1324–1334.

[110] Lever, J., Krzywinski, M., and Altman, N. (2016). Classification evaluation.

[111] Lewis, M. (2019). Bias in the er: Doctors suffer from the same cognitive distortions as the rest of us. *Nautilus*.

[112] Li, C., Dong, H., Fu, W., Qi, M., and Han, B. (2015). Thyroid-like follicular carcinoma of the kidney and papillary renal cell carcinoma with thyroid-like feature: comparison of two cases and literature review. *Annals of Clinical & Laboratory Science*, 45(6):707–712.

[113] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595.

[114] Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., and Li, L. (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC genomics*, 18(1):508.

[115] Liegl, B., Horn, L.-C., and Moinfar, F. (2005). Androgen receptors are frequently expressed in mammary and extramammary paget's disease. *Modern pathology*, 18(10):1283.

[116] Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2020). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in bioinformatics*, 21(1):36–46.

[117] Locke, W. J., Guanzon, D., Ma, C., Liew, Y. J., Duesing, K. R., Fung, K. Y., and Ross, J. P. (2019). Dna methylation cancer biomarkers: translation to the clinic. *Frontiers in Genetics*, 10.

[118] Löffler, H., Pfarr, N., Kriegsmann, M., Endris, V., Hielscher, T., Lohneis, P., Folprecht, G., Stenzinger, A., Dietel, M., Weichert, W., et al. (2016). Molecular driver alterations and their clinical relevance in cancer of unknown primary site. *Oncotarget*, 7(28):44322.

[119] Loison, L. (2016). The microscope against cell theory: Cancer research in nineteenth-century parisian anatomical pathology. *Journal of the history of medicine and allied sciences*, 71(3):271–292.

[120] Losa, F., Soler, G., Casado, A., Estival, A., Fernández, I., Giménez, S., Longo, F., Pazo-Cid, R., Salgado, J., and Seguí, M. (2018). Seom clinical guideline on unknown primary cancer (2017). *Clinical and Translational Oncology*, 20(1):89–96.

[121] Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Harris, P. L., Haserlat, S. M., Supko, J. G., Haluska, F. G., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139.

[122] Maddalena, F., Sisinni, L., Lettini, G., Condelli, V., Matassa, D. S., Piscazzi, A., Amoroso, M. R., La Torre, G., Esposito, F., and Landriscina, M. (2013). Resistance to paclitxel in breast carcinoma cells requires a quality control of mitochondrial antiapoptotic proteins by trap1. *Molecular oncology*, 7(5):895–906.

[123] Marcus, L., Lemery, S. J., Keegan, P., and Pazdur, R. (2019). Fda approval summary: pembrolizumab for the treatment of microsatellite instability-high solid tumors. *Clinical Cancer Research*, 25(13):3753–3758.

[124] Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., Ferté, C., Jamal-Hanjani, M., Wilson, G. A., Shafi, S., Swanton, C., André, F., Szallasi, Z., and Eklund, A. C. (2015). TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Medical Genomics*, 8(1):58.

[125] Massard, C., Michiels, S., Ferté, C., Le Deley, M.-C., Lacroix, L., Hollebecque, A., Verlingue, L., Ileana, E., Rosellini, S., Ammari, S., et al. (2017). High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: results of the moscato 01 trial. *Cancer discovery*, 7(6):586–595.

[126] McMaster, J., Dua, A., and Dowdy, S. C. (2013). Primary breast adenocarcinoma in ectopic breast tissue in the vulva. *Case reports in obstetrics and gynecology*, 2013.

[127] Meiri, E., Mueller, W. C., Rosenwald, S., Zepeniuk, M., Klinke, E., Edmonston, T. B., Werner, M., Lass, U., Barshack, I., Feinmesser, M.,

et al. (2012). A second-generation microrna-based assay for diagnosing tumor tissue origin. *The oncologist*, 17(6):801–812.

[128] Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., and Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA internal medicine*, 173(21):1952–1958.

[129] Micello, D., Marando, A., Sahnane, N., Riva, C., Capella, C., and Sessa, F. (2010). Androgen receptor is frequently expressed in her2-positive, er/pr-negative breast cancers. *Virchows Archiv*, 457(4):467–476.

[130] Miettinen, M., Cue, P. A. M., Sarlomo-Rikala, M., Rys, J., Czapiewski, P., Wazny, K., Langfort, R., Waloszczyk, P., Biernat, W., Lasota, J., et al. (2014). Gata 3–a multispecific but potentially useful marker in surgical pathology–a systematic analysis of 2500 epithelial and non-epithelial tumors. *The American journal of surgical pathology*, 38(1):13.

[131] Mody, R. J., Wu, Y.-M., Lonigro, R. J., Cao, X., Roychowdhury, S., Vats, P., Frank, K. M., Prensner, J. R., Asangani, I., Palanisamy, N., et al. (2015). Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. *Jama*, 314(9):913–925.

[132] Monzon, F. A. and Koen, T. J. (2010). Diagnosis of metastatic neoplasms: molecular approaches for identification of tissue of origin. *Archives of pathology & laboratory medicine*, 134(2):216–224.

[133] Morbeck, D., Tregnago, A. C., Netto, G. B., Sacomani, C., Peresi, P. M., Osório, C. T., Schutz, L., Bezerra, S. M., de Brot, L., and Cunha, I. W. (2017). Gata 3 expression in primary vulvar paget disease: a potential pitfall leading to misdiagnosis of pagetoid urothelial intraepithelial neoplasia. *Histopathology*, 70(3):435–441.

[134] Morganella, S., Alexandrov, L. B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A. M., Brinkman, A. B., Martin, S., Ramakrishna, M., et al. (2016). The topography of mutational processes in breast cancer genomes. *Nature communications*, 7:11383.

[135] Morin, R. D., Johnson, N. A., Severson, T. M., Mungall, A. J., An, J., Goya, R., Paul, J. E., Boyle, M., Woolcock, B. W., Kuchenbauer, F., et al. (2010). Somatic mutations altering ezh2 (tyr641) in follicular and diffuse large b-cell lymphomas of germinal-center origin. *Nature genetics*, 42(2):181.

[136] Mufti, A. and Jackson, R. (2016). Biopsy—what's in the name? *JAMA dermatology*, 152(2):190–190.

[137] Muirhead, D., Aoun, P., Powell, M., Juncker, F., and Mollerup, J. (2010). Pathology economic model tool: a novel approach to workflow and budget cost analysis in an anatomic pathology laboratory. *Archives of pathology & laboratory medicine*, 134(8):1164–1169.

[138] Mullane, S. A., Werner, L., Rosenberg, J., Signoretti, S., Callea, M., Choueiri, T. K., Freeman, G. J., and Bellmunt, J. (2016). Correlation of apobec mrna expression with overall survival and pd-l1 expression in urothelial carcinoma. *Scientific reports*, 6:27702.

[139] Nadji, M., Gomez-Fernandez, C., Ganjei-Azar, P., and Morales, A. R. (2005). Immunohistochemistry of estrogen and progesterone receptors reconsidered: experience with 5,993 breast cancers. *American journal of clinical pathology*, 123(1):21–27.

[140] Nallanthighal, S., Heiserman, J. P., and Cheon, D.-J. (2019). The role of the extracellular matrix in cancer stemness. *Frontiers in Cell and Developmental Biology*, 7.

[141] Navin, N. E. (2015). The first five years of single-cell cancer genomics and beyond. *Genome research*, 25(10):1499–1507.

[142] Neto, A. G., Deavers, M. T., Silva, E. G., and Malpica, A. (2003). Metastatic tumors of the vulva: a clinicopathologic study of 66 cases. *The American journal of surgical pathology*, 27(6):799–804.

[143] Network, C. G. A. R. (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. *New England Journal of Medicine*, 374(2):135–145.

[144] Network, C. G. A. R. et al. (2017). Integrated genomic characterization of oesophageal carcinoma. *Nature*, 541(7636):169–175.

[145] Nojadeh, J. N., Sharif, S. B., and Sakhinia, E. (2018). Microsatellite instability in colorectal cancer. *EXCLI journal*, 17:159.

[146] Oberg, J. A., Bender, J. L. G., Sulis, M. L., Pendrick, D., Sireci, A. N., Hsiao, S. J., Turk, A. T., Cruz, F. S. D., Hibshoosh, H., Remotti, H., et al. (2016). Implementation of next generation sequencing into pediatric hematology-oncology practice: moving beyond actionable alterations. *Genome medicine*, 8(1):133.

[147] Obiorah, I. E. and Ozdemirli, M. (2019). Clear cell sarcoma in unusual sites mimicking metastatic melanoma. *World journal of clinical oncology*, 10(5):213.

[148] Ojala, K. A., Kilpinen, S. K., and Kallioniemi, O. P. (2011). Classification of unknown primary tumors with a data-driven method based on a large microarray reference database. *Genome medicine*, 3(9):63.

[149] Onaiwu, C. O., Salcedo, M. P., Pessini, S. A., Munsell, M. F., Euscher, E. E., Reed, K. E., and Schmeler, K. M. (2017). Paget's disease of the vulva: A review of 89 cases. *Gynecologic oncology reports*, 19:46–49.

[150] Peccerillo, F., Mandel, V. D., Di Tullio, F., Ciardo, S., Chester, J., Kaleci, S., De Carvalho, N., Del Duca, E., Giannetti, L., Mazzoni, L., et al. (2019). Lesions mimicking melanoma at dermoscopy confirmed basal cell carcinoma: Evaluation with reflectance confocal microscopy. *Dermatology*, 235(1):35–44.

[151] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[152] Pentheroudakis, G., Greco, F., and Pavlidis, N. (2009). Molecular assignment of tissue of origin in cancer of unknown primary may not predict response to therapy or outcome: a systematic literature review. *Cancer treatment reviews*, 35(3):221–227.

[153] Perrone, G., Altomare, V., Zagami, M., Vulcano, E., Muzii, L., Battista, C., Rabitti, C., and Muda, A. O. (2009). Breast-like vulvar lesion with concurrent breast cancer: a case report and critical literature review. *in vivo*, 23(4):629–634.

[154] Perrotto, J., Abbott, J. J., Ceilley, R. I., and Ahmed, I. (2010). The role of immunohistochemistry in discriminating primary from secondary extramammary paget disease. *The American Journal of Dermatopathology*, 32(2):137–143.

[155] Pfeifer, J. D. and Wick, M. R. (1995). The pathologic evaluation of neoplastic diseases. *Clinical Oncology, 2nd ed. Edited by Murphy GP, Lawrence W, Lenhard RE. Atlanta: American Cancer Society*, 75:95.

[156] Pietri, E., Conteduca, V., Andreis, D., Massa, I., Melegari, E., Sarti,

S., Cecconetto, L., Schirone, A., Bravaccini, S., Serra, P., et al. (2016). Androgen receptor signaling pathways as a target for breast cancer treatment. *Endocrine-related cancer*, 23(10):R485–R498.

[157] Pillai, R., Deeter, R., Rigl, C. T., Nystrom, J. S., Miller, M. H., Buturovic, L., and Henner, W. D. (2011). Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens. *The Journal of molecular diagnostics*, 13(1):48–56.

[158] Piros, E., Petak, I., Erdos, A., Hautman, J., and Lisziewicz, J. (2016). Market opportunity for molecular diagnostics in personalized cancer therapy. *Handbook of clinical nanomedicine. Law, business, regulation, safety, and risk. Stanford: Taylor & Francis*, pages 273–301.

[159] Posadas, E. M., Liel, M. S., Kwitkowski, V., Minasian, L., Godwin, A. K., Hussain, M. M., Espina, V., Wood, B. J., Steinberg, S. M., and Kohn, E. C. (2007). A phase ii and pharmacodynamic study of gefitinib in patients with refractory or recurrent epithelial ovarian cancer. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 109(7):1323–1330.

[160] Pstrąg, N., Ziemnicka, K., Bluyssen, H., and Wesoły, J. (2018). Thyroid cancers of follicular origin in a genomic light: in-depth overview of common and unique molecular marker candidates. *Molecular cancer*, 17(1):116.

[161] Quon, G. and Morris, Q. (2009). Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, 25(21):2882–2889.

[162] Raghav, K., Mhadgut, H., McQuade, J. L., Lei, X., Ross, A., Matamoros, A., Wang, H., Overman, M. J., and Varadhachary, G. R. (2016). Cancer of unknown primary in adolescents and young adults: Clinicopathological features, prognostic factors and survival outcomes. *PLoS One*, 11(5).

[163] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154.

[164] Rapin, N., Bagger, F. O., Jendholm, J., Mora-Jensen, H., Krogh, A., Kohlmann, A., Thiede, C., Borregaard, N., Bullinger, L., Winther,

O., et al. (2014). Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in aml patients. *Blood, The Journal of the American Society of Hematology*, 123(6):894–904.

[165] Rapiti, E., Verkooijen, H. M., Vlastos, G., Fioretta, G., Neyroud-Caspar, I., Sappino, A. P., Chappuis, P. O., and Bouchardy, C. (2006). Complete excision of primary breast tumor improves survival of patients with metastatic breast cancer at diagnosis. *J clin oncol*, 24(18):2743–2749.

[166] Rassy, E., Assi, T., and Pavlidis, N. (2020). Exploring the biological hallmarks of cancer of unknown primary: where do we stand today? *British Journal of Cancer*, pages 1–9.

[167] Reyna, M. A., Leiserson, M. D., and Raphael, B. J. (2018). Hierarchical hotnet: identifying hierarchies of altered subnetworks. *Bioinformatics*, 34(17):i972–i980.

[168] Ricketts, C. J., De Cubas, A. A., Fan, H., Smith, C. C., Lang, M., Reznik, E., Bowlby, R., Gibb, E. A., Akbani, R., Beroukhim, R., et al. (2018). The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell reports*, 23(1):313–326.

[169] Robinson, D. R., Wu, Y.-M., Lonigro, R. J., Vats, P., Cobain, E., Everett, J., Cao, X., Rabban, E., Kumar-Sinha, C., Raymond, V., et al. (2017). Integrative clinical genomics of metastatic cancer. *Nature*, 548(7667):297–303.

[170] Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., Benjamin, H., Shabes, N., Tabak, S., Levy, A., et al. (2008). Micrornas accurately identify cancer tissue origin. *Nature biotechnology*, 26(4):462.

[171] Ross, J. S., Wang, K., Gay, L., Otto, G. A., White, E., Iwanik, K., Palmer, G., Yelensky, R., Lipson, D. M., Chmielecki, J., Erlich, R. L., Rankin, A. N., Ali, S. M., Elvin, J. A., Morosini, D., Miller, V. A., and Stephens, P. J. (2015). Comprehensive Genomic Profiling of Carcinoma of Unknown Primary Site: New Routes to Targeted Therapies. *JAMA oncology*, 1(1):40–9.

[172] Rostoker, R., Abelson, S., Bitton-Worms, K., Genkin, I., Ben-Shmuel, S., Dakwar, M., Orr, Z. S., Caspi, A., Tzukerman, M., and LeRoith,

D. (2015). Highly specific role of the insulin receptor in breast cancer progression. *Endocrine-related cancer*, 22(2):145–157.

[173] Saadatpour, A., Lai, S., Guo, G., and Yuan, G.-C. (2015). Single-cell analysis in cancer genomics. *Trends in Genetics*, 31(10):576–586.

[174] Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817.

[175] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl_1):D674–D679.

[176] Schroten-Loef, C., Verhoeven, R., de Hingh, I., van de Wouw, A., van Laarhoven, H., and Lemmens, V. (2018). Unknown primary carcinoma in the netherlands: decrease in incidence and survival times remain poor between 2000 and 2012. *European Journal of Cancer*, 101:77–86.

[177] Seshacharyulu, P., Ponnusamy, M. P., Haridas, D., Jain, M., Ganti, A. K., and Batra, S. K. (2012). Targeting the egfr signaling pathway in cancer therapy. *Expert opinion on therapeutic targets*, 16(1):15–31.

[178] Sethupathy, P., Corda, B., and Hatzigeorgiou, A. G. (2006). Tarbase: A comprehensive database of experimentally supported animal microrna targets. *Rna*, 12(2):192–197.

[179] Shaoxian, T., Baohua, Y., Xiaoli, X., Yufan, C., Xiaoyu, T., Hongfen, L., Rui, B., Xiangjie, S., Ruohong, S., and Wentao, Y. (2017). Characterisation of gata3 expression in invasive breast cancer: differences in histological subtypes and immunohistochemically defined molecular subtypes. *Journal of clinical pathology*, 70(11):926–934.

[180] Shendure, J., Findlay, G. M., and Snyder, M. W. (2019). Genomic medicine–progress, pitfalls, and promise. *Cell*, 177(1):45–57.

[181] Slodkowska, E. A. and Ross, J. S. (2009). Mammaprint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert review of molecular diagnostics*, 9(5):417–422.

[182] Sobin, L. H., Gospodarowicz, M. K., and Wittekind, C. (2011). *TNM classification of malignant tumours*. John Wiley & Sons.

[183] Soh, K. P., Szczurek, E., Sakoparnig, T., and Beerenwinkel, N. (2017).

Predicting cancer type from tumour dna signatures. *Genome medicine*, 9(1):104.

[184] Søkilde, R., Vincent, M., Møller, A. K., Hansen, A., Høiby, P. E., Blondal, T., Nielsen, B. S., Daugaard, G., Møller, S., and Litman, T. (2014). Efficient identification of mirnas for classification of tumor origin. *The Journal of Molecular Diagnostics*, 16(1):106–115.

[185] Song, H.-W. and Wilkinson, M. F. (2014). Transcriptional control of spermatogonial maintenance and differentiation. In *Seminars in cell & developmental biology*, volume 30, pages 14–26. Elsevier.

[186] Stefanovic, S., Wirtz, R., Deutsch, T. M., Hartkopf, A., Sinn, P., Varga, Z., Sobottka, B., Sotiris, L., Taran, F.-A., Domschke, C., et al. (2017). Tumor biomarker conversion between primary and metastatic breast cancer: mrna assessment and its concordance with immunohistochemistry. *Oncotarget*, 8(31):51416.

[187] Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., et al. (2001). Molecular classification of human carcinomas by use of gene expression signatures. *Cancer research*, 61(20):7388–7393.

[188] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org.

[189] Suvà, M. L. and Tirosh, I. (2019). Single-cell rna sequencing in cancer: lessons learned and emerging challenges. *Molecular cell*, 75(1):7–12.

[190] Tan, S. Y. and Tatsumura, Y. (2015). George papanicolaou (1883–1962): discoverer of the pap smear. *Singapore medical journal*, 56(10):586.

[191] Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2017). Tumor origin detection with tissue-specific mirna and dna methylation markers. *Bioinformatics*, 34(3):398–406.

[192] Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific mirna and dna methylation markers. *Bioinformatics*, 34(3):398–406.

[193] Tessier-Cloutier, B., Asleh-Aburaya, K., Shah, V., McCluggage, W. G., Tinker, A., and Gilks, C. B. (2017). Molecular subtyping of mammary-like

adenocarcinoma of the vulva shows molecular similarity to breast carcinomas. *Histopathology*, 71(3):446–452.

[194] Thavarajah, R., Mudimbaimannar, V. K., Elizabeth, J., Rao, U. K., and Ranganathan, K. (2012). Chemical and physical basics of routine formaldehyde fixation. *Journal of oral and maxillofacial pathology: JOMFP*, 16(3):400.

[195] Tian, R., Basu, M. K., and Capriotti, E. (2014). Contrastrank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics*, 30(17):i572–i578.

[196] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

[197] Titford, M. (2005). The long history of hematoxylin. *Biotechnic & histochemistry*, 80(2):73–78.

[198] Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

[199] Tormo, E., Adam-Artigues, A., Ballester, S., Pineda, B., Zazo, S., González-Alonso, P., Albanell, J., Rovira, A., Rojo, F., Lluch, A., et al. (2017). The role of mir-26a and mir-30b in her2+ breast cancer trastuzumab resistance and regulation of the ccne2 gene. *Scientific reports*, 7:41309.

[200] Toth, C. D., O'Rourke, J., and Goodman, J. E. (2017). *Handbook of discrete and computational geometry*. Chapman and Hall/CRC.

[201] Tothill, R. W., Shi, F., Paiman, L., Bedo, J., Kowalczyk, A., Mileshkin, L., Buela, E., Klupacs, R., Bowtell, D., and Byron, K. (2015). Development and validation of a gene expression tumour classifier for cancer of unknown primary. *Pathology*, 47(1):7–12.

[202] Tran, T. A., Deavers, M. T., Carlson, J. A., and Malpica, A. (2015). Collision of ductal carcinoma in situ of anogenital mammary-like glands and vulvar sarcomatoid squamous cell carcinoma. *International Journal of Gynecological Pathology*, 34(5):487–494.

[203] Trédan, O., Wang, Q., Pissaloux, D., Cassier, P., de la Fouchardière, A., Fayette, J., Desseigne, F., Ray-Coquard, I., de la Fouchardière, C., Frappaz, D., et al. (2019). Molecular screening program to select

molecular-based recommended therapies for metastatic cancer patients: analysis from the profiler trial. *Annals of Oncology*, 30(5):757–765.

[204] Van der Putte, S. (1994). Mammary-like glands of the vulva and their disorders. *International Journal of Gynecological Pathology*, 13(2):150–160.

[205] Van Such, M., Lohr, R., Beckman, T., and Naessens, J. M. (2017). Extent of diagnostic agreement among medical referrals. *Journal of evaluation in clinical practice*, 23(4):870–874.

[206] Varadhachary, G. R., Raber, M. N., Matamoros, A., and Abbruzzese, J. L. (2008). Carcinoma of unknown primary with a colon-cancer profile—changing paradigm and emerging definitions. *The lancet oncology*, 9(6):596–599.

[207] Varghese, A., Arora, A., Capanu, M., Camacho, N., Won, H., Zehir, A., Gao, J., Chakravarty, D., Schultz, N., Klimstra, D., et al. (2017). Clinical and molecular characterization of patients with cancer of unknown primary in the modern era. *Annals of Oncology*, 28(12):3015–3021.

[208] Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245.

[209] Vennalaganti, P., Kanakadandi, V., Goldblum, J. R., Mathur, S. C., Patil, D. T., Offerhaus, G. J., Meijer, S. L., Vieth, M., Odze, R. D., Shreyas, S., et al. (2017). Discordance among pathologists in the united states and europe in diagnosis of low-grade dysplasia for patients with barrett's esophagus. *Gastroenterology*, 152(3):564–570.

[210] Vitali, F., Li, Q., Schissler, A. G., Berghout, J., Kenost, C., and Lussier, Y. A. (2019). Developing a 'personalome'for precision medicine: emerging methods that compute interpretable effect sizes from single-subject transcriptomes. *Briefings in bioinformatics*, 20(3):789–805.

[211] Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–799.

[212] Wagle, M.-C., Castillo, J., Srinivasan, S., Holcomb, T., Yuen, K. C., Kadel, E. E., Mariathasan, S., Halligan, D. L., Carr, A. R., Bylesjo, M., et al. (2020). Tumor fusion burden as a hallmark of immune infiltration in prostate cancer. *Cancer Immunology Research*.

[213] Wagner, A. H., Walsh, B., Mayfield, G., Tamborero, D., Sonkin, D., Krysiak, K., Pons, J. D., Duren, R., Gao, J., McMurry, J., et al. (2018). A harmonized meta-knowledgebase of clinical interpretations of cancer genomic variants. *BioRxiv*, page 366856.

[214] Wang, H. L., Kim, C. J., Koo, J., Zhou, W., Choi, E. K., Arcega, R., Chen, Z. E., Wang, H., Zhang, L., and Lin, F. (2017). Practical immunohistochemistry in neoplastic pathology of the gastrointestinal tract, liver, biliary tract, and pancreas. *Archives of Pathology and Laboratory Medicine*, 141(9):1155–1180.

[215] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113.

[216] Weiss, L. M., Chu, P., Schroeder, B. E., Singh, V., Zhang, Y., Erlander, M. G., and Schnabel, C. A. (2013). Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. *The Journal of Molecular Diagnostics*, 15(2):263–269.

[217] Willman, J. H., Golitz, L. E., and Fitzpatrick, J. E. (2005). Vulvar clear cells of toker: precursors of extramammary paget's disease. *The American journal of dermatopathology*, 27(3):185–188.

[218] Wolff, A. C., Hammond, M. E. H., Hicks, D. G., Dowsett, M., McShane, L. M., Allison, K. H., Allred, D. C., Bartlett, J. M., Bilous, M., Fitzgibbons, P., et al. (2013). Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline update. *Archives of Pathology and Laboratory Medicine*, 138(2):241–256.

[219] Wu, Y.-M., Cieślik, M., Lonigro, R. J., Vats, P., Reimers, M. A., Cao, X., Ning, Y., Wang, L., Kunju, L. P., de Sarkar, N., et al. (2018). Inactivation of cdk12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell*, 173(7):1770–1782.

[220] Xia, A., Zhang, X.-Y., Wang, J., Yin, T., and Lu, X.-J. (2019). T cell dysfunction in cancer immunity and immunotherapy. *Frontiers in immunology*, 10:1719.

[221] Xu, Q., Chen, J., Ni, S., Tan, C., Xu, M., Dong, L., Yuan, L., Wang,

Q., and Du, X. (2016). Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Modern Pathology*, 29(6):546.

[222] Yagi, Y. and Gilbertson, J. R. (2008). A relationship between slide quality and image quality in whole slide imaging (wsi). In *Diagnostic pathology*, volume 3, page S12. BioMed Central.

[223] Yang, J., Nie, J., Ma, X., Wei, Y., Peng, Y., and Wei, X. (2019). Targeting pi3k in cancer: mechanisms and advances in clinical trials. *Molecular cancer*, 18(1):26.

[224] Zararsız, G., Goksuluk, D., Korkmaz, S., Eldem, V., Zararsiz, G. E., Duru, I. P., and Ozturk, A. (2017). A comprehensive simulation study on classification of rna-seq data. *PloS one*, 12(8).

[225] Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S. M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*, 23(6):703.

[226] Zhang, F., Chen, X. P., Zhang, W., Dong, H. H., Xiang, S., Zhang, W. G., and Zhang, B. X. (2008). Combined hepatocellular cholangiocarcinoma originating from hepatic progenitor cells: Immunohistochemical and double-fluorescence immunostaining evidence. *Histopathology*, 52(2):224–232.

[227] Zhang, M., Chen, H., Wang, M., Bai, F., and Wu, K. (2020). Bioinformatics analysis of prognostic significance of col10a1 in breast cancer. *Bioscience Reports*, 40(2).

[228] Zhang, W., Chien, J., Yong, J., and Kuang, R. (2017). Network-based machine learning and graph theory algorithms for precision oncology. *NPJ precision oncology*, 1(1):1–15.

[229] Zilliox, M. J. and Irizarry, R. A. (2007). A gene expression bar code for microarray data. *Nature methods*, 4(11):911.

[230] Zubor, P., Kubatka, P., Dankova, Z., Gondova, A., Kajo, K., Hatok, J., Samec, M., Jagelkova, M., Krivus, S., Holubekova, V., et al. (2018). mirna in a multiomic context for diagnosis, treatment monitoring and personalized management of metastatic breast cancer. *Future Oncology*, 14(18):1847–1867.

Appendix

Additional materials for Chapter 4

Table 1: Important genes based on frequency analysis of gene weights for each neural network in SCOPE.

Cancer Code	Tissue	Organ-System	Genes
ACC	Tumour	Endocrine	CYP11A1, CYP17A1, CYP21A2, DLK1, GSTA1, IGF2, NPTX2, STAR
BLCA	Adjacent Normal	Urologic	ACTG2, CNN1, DES, DHRS2, GPX2, KRT13, KRT5, LY6D, OLFM4, PLA2G2A, S100P, SPRR3, UPK2
BLCA	Tumour	Urologic	AKR1C2, DES, DHRS2, GATA3, GPX2, KRT13, KRT17, KRT5, PSCA, S100P, SPINK1, UPK1B, UPK2
BRCA	Adjacent Normal	Breast	ADH1B, ADIPOQ, APOD, AZGP1, GATA3, KRT14, LPL, MUCL1, PIP, PLIN1, S100A1, SAA1, SCGB1D2, SCGB2A2, TFF1
BRCA	Tumour	Breast	AGR3, AZGP1, CALML5, CRABP2, EFHD1, FABP4, GATA3, KRT14, KRT6B, LTF, MMP11, MUCL1, NPY1R, PIP, SCGB2A2, SERPINA3, SPDEF, TFF1
CESC_CAD	Adjacent Normal	Gynecologic	DES
CESC_CAD	Tumour	Gynecologic	CEACAM5, CLDN3, KRT7, MMP11, PIGR, SCGB2A1
CESC_SCC	Adjacent Normal	Gynecologic	CNN1
CESC_SCC	Tumour	Gynecologic	CALML3, KRT13, KRT14, KRT19, KRT5, KRT6A, MMP11
CHOL	Tumour	Gastrointestinal	AGT, ALB, AMBP, CEACAM6, CRP, FGA, FGB, FGG, ORM1, REG1A, TM4SF4, TTR
COADREAD	Adjacent Normal	Gastrointestinal	AQP8, CA1, CEACAM7, CLCA1, DES, FABP1, FAM3D, GPX2, GUCA2A, KRT20, SLC26A3, SPINK4, ZG16

Table 1: Important genes based on frequency analysis of gene weights for each neural network in SCOPE. *(continued)*

Cancer Code	Tissue	Organ-System	Genes
COADREAD	Tumour	Gastrointestinal	CDH17, CDX2, CEACAM5, CEACAM6, DPEP1, FABP1, FAM3D, GPX2, LGALS4, MUC13, MUC2, PLA2G2A, PPP1R1B, REG4, S100P, SPINK4, TSPAN8, VIL1
ESCA_EAC	Adjacent Normal	Gastrointestinal	ACTG2, DES, LIPF, PGA3, PGA4
ESCA_EAC	Tumour	Gastrointestinal	CEACAM5, CST1, KRT13, LGALS4, MALAT1, MUC13, PIGR, PLA2G2A, S100A7, SPRR3, TSPAN8, UBD
ESCA	Adjacent Normal	Gastrointestinal	KRT13
ESCA_SCC	Adjacent Normal	Gastrointestinal	SPRR1B
ESCA_SCC	Tumour	Gastrointestinal	CALML3, CST1, DES, KRT14, KRT5, LY6D, MALAT1, S100A7, SPRR1B, SPRR3, TRIM29
ESCA	Tumour	Gastrointestinal	CLDN18, CST1, MALAT1, REG1A, REG3A, SPINK1
FL	Tumour	Hematologic	CCL21
GBM	Tumour	CNS	AQP4, CHI3L1, GFAP
HNSC	Adjacent Normal	Head and Neck	ACTA1, CALML5, CKM, KRT13, KRT4, MB, MUC7, MYH2, MYL1, MYL2, MYLPF, PIP, PRB3, SAA1, SCGB3A1, SMR3B, STATH, TCAP, TGM3, TNNC2
HNSC	Tumour	Head and Neck	ACTA1, CALML3, CALML5, KRT13, KRT14, LGALS7, MMP1, SPRR2A, SPRR3
KICH	Adjacent Normal	Urologic	ALDOB, AQP2, FXYD2, UMOD
KICH	Tumour	Urologic	ATP6V0A4, ATP6V0D2, CDH16, DEFB1, RHCG, SPINK1, SPP1, TMEM213
KIRC	Adjacent Normal	Urologic	AQP2, CDH16, SLC34A2, UMOD
KIRC	Tumour	Urologic	ANGPTL4, CA12, CA9, DEFB1, EGLN3, ESM1, FXYD2, GSTA1, NAT8
KIRP	Adjacent Normal	Urologic	AQP2, PIGR, UMOD
KIRP	Tumour	Urologic	C19orf33, MAL, MMP7, PIGR, SST, WFDC2
LAML	Tumour	Hematologic	AZU1, CSF3R, FOSB, MPO, PRTN3, RNASE2, S100A8
LGG	Tumour	CNS	EEF1A1P9, GFAP, PTPRZ1
LIHC	Adjacent Normal	Gastrointestinal	IGFBP1

Table 1: Important genes based on frequency analysis of gene weights for each neural network in SCOPE. *(continued)*

Cancer Code	Tissue	Organ-System	Genes
LIHC	Tumour	Gastrointestinal	ALB, APCS, APOA2, APOC3, CRP, FGA, FGB, GC, HULC, ITIH2, RBP4, TF, TM4SF4, UBD, VTN
LUAD	Adjacent Normal	Thoracic	HBA1, NAPSA, SCGB1A1, SCGB3A1, SCGB3A2, SFTPA1, SFTPB, SFTPC, SFTPD, SLPI
LUAD	Tumour	Thoracic	C8orf4, CEACAM5, CRABP2, FGG, NAPSA, PGC, S100P, SCGB1A1, SCGB3A1, SCGB3A2, SFTA2, SFTPA1, SFTPA2, SFTPB, SLC34A2, SPINK1
LUSC	Adjacent Normal	Thoracic	CCL21, NAPSA, RPS4Y1, SCGB1A1, SCGB3A1, SCGB3A2, SFTA2, SFTPA1, SFTPA2, SFTPB, SFTPC, SFTPD, SLC34A2
LUSC	Tumour	Thoracic	AKR1C2, CALML3, CES1, KRT15, KRT16, KRT19, KRT5, KRT6A, KRT6B, NAPSA, NTS, SCGB1A1, SCGB3A2, SFTPA1, SFTPA2, SFTPB, SFTPC, SPRR2A
MB-Adult	Tumour	CNS	GFAP, STMN2
MESO	Tumour	Thoracic	C19orf33, CALB2, EFEMP1, ITLN1, KRT19, KRT7, MSLN, UPK3B
OV	Tumour	Gynecologic	CHI3L1, CLDN3, FOLR1, KLK6, KLK7, MALAT1, MSLN, PAX8, SCGB2A1, SOX17, SST
PAAD	Adjacent Normal	Gastrointestinal	CELA3A, CPA1, CPB1, CRP, CTRB1, CTRB2, CTRC, CTSE, GCG, INS, PNLIP, PPY, PRSS1, REG1A, REG3A, TTR
PAAD	Tumour	Gastrointestinal	AGR2, CEACAM5, CHGB, CTRB1, CTRB2, GCG, INS, PNLIP, PPY, REG1A, REG4, S100P, SFRP2, SPINK1, SST, TFF1, TFF2, TTR
PCPG	Adjacent Normal	Endocrine	CYP11B1, CYP17A1, DLK1, GSTA1, STAR
PCPG	Tumour	Endocrine	CHGA, CHGB, DBH, DLK1, NPY, PENK
PRAD	Adjacent Normal	Urologic	ACPP, KLK2, KLK3, KLK4, NPY, OLFM4, PIP, SEMG1
PRAD	Tumour	Urologic	ACTG2, AZGP1, DES, FOLH1, FOXA1, KLK2, KLK3, KLK4, NKX3-1, NPY, PLA2G2A, SLC45A3
SARC	Tumour	Soft Tissue	DLK1
SKCM	Adjacent Normal	Skin	DCT, MLANA, PRAME, TYR
SKCM	Tumour	Skin	APOD, DCT, EDNRB, KRT6B, MLANA, PLP1, PRAME, S100A1, SERPINE2, SOX10, TYR, TYRP1, VGF

Table 1: Important genes based on frequency analysis of gene weights for each neural network in SCOPE. *(continued)*

Cancer Code	Tissue	Organ-System	Genes
STAD	Adjacent Normal	Gastrointestinal	ACTG2, APOA1, APOA4, CLDN18, DES, GKN1, HSPB6, PGA4, PGC, PI3, REG3A
STAD	Tumour	Gastrointestinal	ACTG2, CEACAM6, CST1, MALAT1, PGC, REG4, SPINK1, TFF1, TFF3
TFRI_GBM_N(	Tumour	CNS	MALAT1, PCDHGA1, PCDHGA8, PCDHGC4, PMP2
TGCT	Tumour	Urologic	DPPA3, DPPA5, GDF3, NANOG, POU5F1
THCA	Adjacent Normal	Endocrine	CCL21, HBA2, MT1G, PAX8, TG, TPO
THCA	Tumour	Endocrine	C16orf89, CLIC3, NKX2-1, S100A1, SFTA3, SFTPB, TG, TPO, ZCCHC12
THYM	Adjacent Normal	Hematologic	CALML3, CCL25, KRT5
THYM	Tumour	Hematologic	CALML3, CCL25, DNTT, KRT14, KRT15, KRT17, KRT19, KRT5, PAX1
UCEC	Adjacent Normal	Gynecologic	CNN1, DES
UCEC	Tumour	Gynecologic	MMP11, MSX1, PAX8, SCGB1D2, SCGB2A1, SFN, VTCN1
UCS	Tumour	Gynecologic	CRABP1, DLK1, PCOLCE, PRAME
UVM	Tumour	Head and Neck	CITED1, MLANA, SOX10, TYR, TYRP1

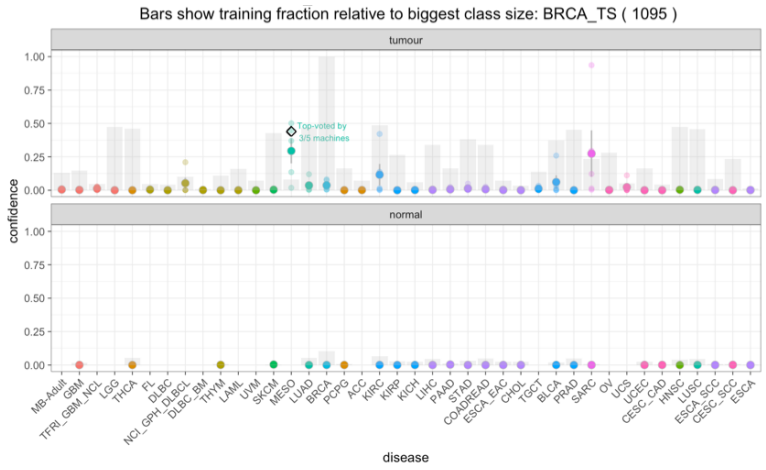

Figure 1: Example output from SCOPE for a sarcomatoid mesothelioma, predicted with split confidence as mesothelioma and sarcoma.

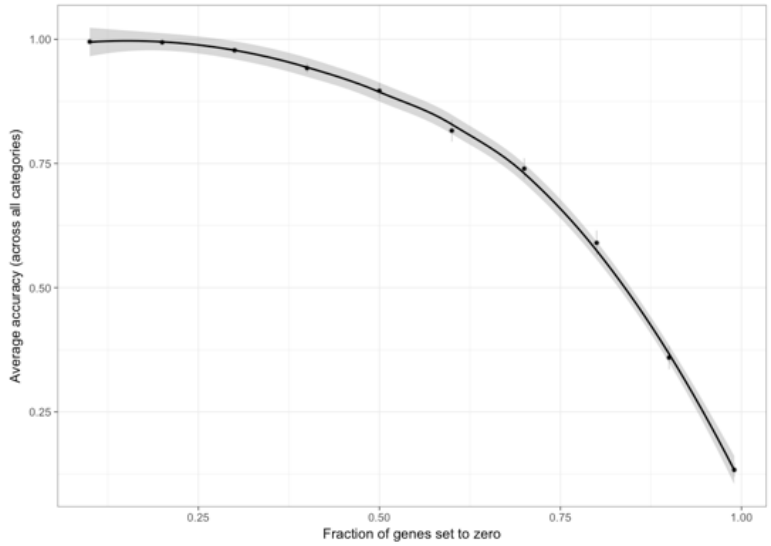

Figure 2: Mean prediction accuracy of SCOPE as RPKM values of various fractions of genes are set to 0 in the input RNA-Seq data. Grey bars around mean points indicate standard error bounds. Black line indicates the line of best fit (loess). At a given threshold n% genes in input were randomly set to zero. This was repeated 10 times for each n in (10, 20, 30, 40, 50, 60, 70, 80, 90, 99).

Additional materials for Chapter 5

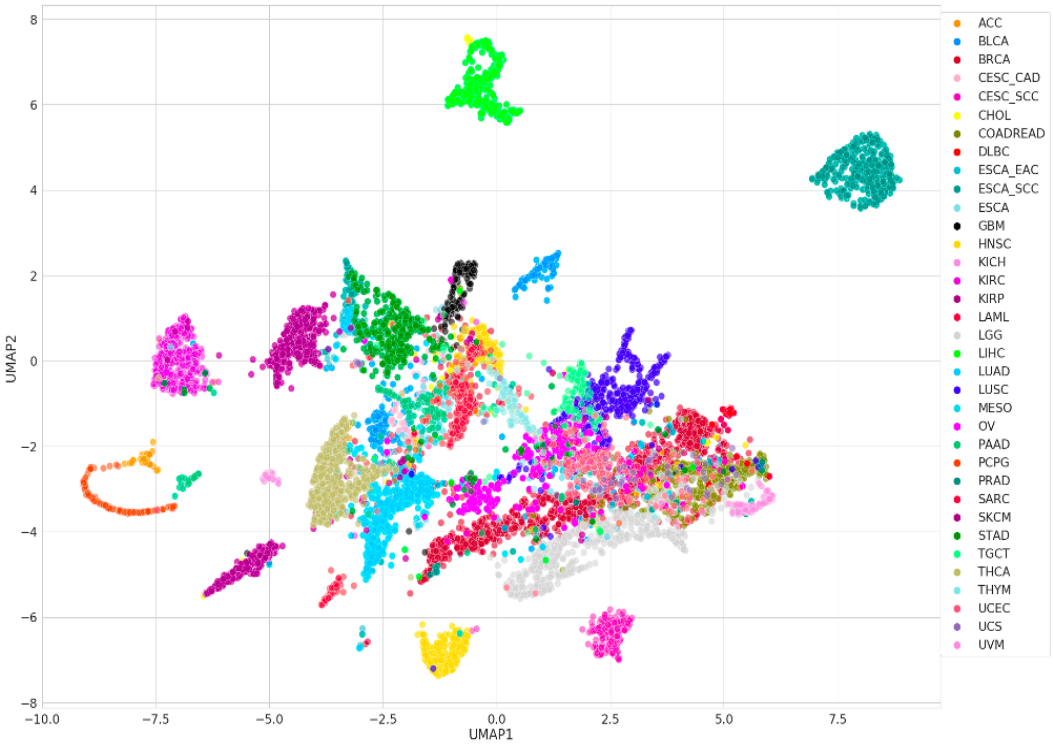

Figure 3: UMAP projections of PIE profiles for 3,963 biochemical pathways, for samples in the TCGA cohort of primary tumours.The projections are coloured by tumour-type.

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores.

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Breast	BRCA	Normal	A tetrasaccharide linker sequence is required for GAG synthesis, AP-1 transcription factor network, bone remodeling, cadmium induces dna synthesis and proliferation in macrophages, Chondroitin sulfate biosynthesis, Chondroitin sulfate/dermatan sulfate metabolism, Dermatan sulfate biosynthesis, EGFR Inhibitor Pathway, Pharmacodynamics, Estrogen signaling pathway, Heparan sulfate/heparin (HS-GAG) metabolism, Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis, il 3 signaling pathway, Lipid digestion, mobilization, and transport, Miscellaneous transport and binding events, nerve growth factor pathway (ngf), PPAR signaling pathway - Homo sapiens (human), Quercetin and Nf-kB- AP-1 Induced Cell Apoptosis, Regulation of lipolysis in adipocytes - Homo sapiens (human), RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways, TRAF6 mediated NF-kB activation, Transcriptional regulation of white adipocyte differentiation, Translation Factors, Transport of fatty acids, tsp-1 induced apoptosis in microvascular endothelial cell, west nile virus

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Breast	BRCA	Tumour	Amoebiasis - Homo sapiens (human), Collagen biosynthesis and modifying enzymes, Collagen formation, ECM proteoglycans, Endothelins, IL4-mediated signaling events, Inflammatory Response Pathway, Insulin Signaling, Integrins in angiogenesis, Iron metabolism in placenta, Methionine and cysteine metabolism, miR-targeted genes in muscle cell - TarBase, miRNA targets in ECM and membrane receptors, Miscellaneous transport and binding events, pi3k_pathway, PI3K-Akt signaling pathway - Homo sapiens (human), Platelet Aggregation Inhibitor Pathway, Pharmacodynamics, Protein processing in endoplasmic reticulum - Homo sapiens (human), Senescence and Autophagy in Cancer, Signaling by Retinoic Acid, Syndecan-1-mediated signaling events, Validated nuclear estrogen receptor alpha network, Validated targets of C-MYC transcriptional repression, VEGFR3 signaling in lymphatic endothelium, Vitamin A and Carotenoid Metabolism

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Central Nervous System	GBM	Normal	Acetylcholine Neurotransmitter Release Cycle, Asymmetric localization of PCP proteins, Autodegradation of the E3 ubiquitin ligase COP1, beta-catenin independent WNT signaling, Budding and maturation of HIV virion, CREB phosphorylation through the activation of Ras, Dopamine Neurotransmitter Release Cycle, Effects of Botulinum toxin, GABA synthesis, release, reuptake and degradation, Gastrin-CREB signalling pathway via PKC and MAPK, Generic Transcription Pathway, Ion channel transport, Neuronal System, Neurotransmitter Release Cycle, NGF signalling via TRKA from the plasma membrane, Norepinephrine Neurotransmitter Release Cycle, Serotonin Neurotransmitter Release Cycle, signal dependent regulation of myogenesis by corepressor mitr, Signaling by EGFR, Signaling by ERBB2, Synaptic vesicle cycle - Homo sapiens (human), Synaptic Vesicle Pathway, Toxicity of botulinum toxin type C (BoNT/C), Transmission across Chemical Synapses, Uptake and actions of bacterial toxins

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Central Nervous System	GBM	Tumour	Anchoring of the basal body to the plasma membrane, Assembly of the primary cilium, Cell Cycle, Cell Cycle, Mitotic, cell_cycle_pathway, EPO signaling, G2/M Transition, Gap junction - Homo sapiens (human), IL-7 signaling, JAK STAT pathway and regulation, Loss of Nlp from mitotic centrosomes, Loss of proteins required for interphase microtubule organizationfrom the centrosome, Mammary gland development pathway - Involution (Stage 4 of 4), miR-targeted genes in epithelium - TarBase, miR-targeted genes in leukocytes - TarBase, miR-targeted genes in squamous cell - TarBase, Mitotic G2-G2/M phases, Organelle biogenesis and maintenance, p73 transcription factor network, Parkin-Ubiquitin Proteasomal System pathway, Pyrimidine metabolism, Regulation of PLK1 Activity at G2/M Transition, RMTs methylate histone arginines, stathmin and breast cancer resistance to antimicrotubule agents, tgf_pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Central Nervous System	LGG	Tumour	3, -UTR-mediated translational regulation, Cap-dependent Translation Initiation, Chylomicron-mediated lipid transport, Cytoplasmic Ribosomal Proteins, Eukaryotic Translation Initiation, Eukaryotic Translation Termination, GTP hydrolysis and joining of the 60S ribosomal subunit, HDL-mediated lipid transport, L13a-mediated translational silencing of Ceruloplasmin expression, Lipid digestion, mobilization, and transport, Lipoprotein metabolism, Neural Crest Differentiation, Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC), Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC), Nonsense-Mediated Decay (NMD), Platelet degranulation , prion pathway, Response to elevated platelet cytosolic Ca2+, Retinoid metabolism and transport, Spinal Cord Injury, SRP-dependent cotranslational protein targeting to membrane, Statin Pathway, Statin Pathway, Pharmacodynamics, Visual phototransduction, Vitamin B12 Metabolism

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Endocrine	ACC	Tumour	11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), 21-hydroxylase deficiency (CYP21), 3-Beta-Hydroxysteroid Dehydrogenase Deficiency, Activated NOTCH1 Transmits Signal to the Nucleus, Adipogenesis, Adrenal Hyperplasia Type 3 or Congenital Adrenal Hyperplasia due to 21-hydroxylase Deficiency, Adrenal Hyperplasia Type 5 or Congenital Adrenal Hyperplasia due to 17 Alpha-hydroxylase Deficiency, Apparent mineralocorticoid excess syndrome, Cardiac Progenitor Differentiation, Congenital Lipoid Adrenal Hyperplasia (CLAH) or Lipoid CAH, Corticosterone methyl oxidase I deficiency (CMO I), Corticosterone methyl oxidase II deficiency - CMO II, Cytochrome P450 - arranged by substrate type, Endogenous sterols, FOXA2 and FOXA3 transcription factor networks, IGF-Core, Metabolism of steroid hormones and vitamin D, Notch signaling pathway, Ovarian steroidogenesis - Homo sapiens (human), Posttranslational regulation of adherens junction stability and dissassembly, Pregnenolone biosynthesis, SHC-related events triggered by IGF1R, Steroid hormones, superpathway of steroid hormone biosynthesis

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Endocrine	PCPG	Normal	11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), 17-Beta Hydroxysteroid Dehydrogenase III Deficiency, 21-hydroxylase deficiency (CYP21), 3-Beta-Hydroxysteroid Dehydrogenase Deficiency, 3, -UTR-mediated translational regulation, Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S, Adaptive Immune System, Adherens junction - Homo sapiens (human), Adrenal Hyperplasia Type 3 or Congenital Adrenal Hyperplasia due to 21-hydroxylase Deficiency, Adrenal Hyperplasia Type 5 or Congenital Adrenal Hyperplasia due to 17 Alpha-hydroxylase Deficiency, Androgen and estrogen biosynthesis and metabolism, Androgen and Estrogen Metabolism, androgen biosynthesis, Androgen biosynthesis, antigen processing and presentation, Antigen processing and presentation - Homo sapiens (human), Apparent mineralocorticoid excess syndrome, Aromatase deficiency, Arrhythmogenic Right Ventricular Cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy (ARVC) - Homo sapiens (human), Axon guidance, Bacterial invasion of epithelial cells - Homo sapiens (human), BCR, Biological oxidations

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Endocrine	PCPG	Tumour	FGFR2, FGFR3, FGFR4, -arrestins in gpcr desensitization, 11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), 17-Beta Hydroxysteroid Dehydrogenase III Deficiency, 2-Methyl-3-Hydroxybutryl CoA Dehydrogenase Deficiency, 21-hydroxylase deficiency (CYP21), 3-Beta-Hydroxysteroid Dehydrogenase Deficiency, 3-Hydroxy-3-Methylglutaryl-CoA Lyase Deficiency, 3-hydroxyisobutyric acid dehydrogenase deficiency, 3-hydroxyisobutyric aciduria, 3-Methylcrotonyl Coa Carboxylase Deficiency Type I, 3-Methylglutaconic Aciduria Type I, 3-Methylglutaconic Aciduria Type III, Activated NOTCH1 Transmits Signal to the Nucleus, Adipogenesis, FOXA2 and FOXA3 transcription factor networks, Metabolism, Notch signaling pathway, notch_pathway, Signal Transduction, Signaling by NOTCH, Signaling by NOTCH1

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Endocrine	THCA	Normal	Autoimmune thyroid disease - Homo sapiens (human), BCR, cadmium induces dna synthesis and proliferation in macrophages, Calcineurin-regulated NFAT-dependent transcription in lymphocytes, calcium signaling by hbx of hepatitis b virus, Calcium signaling in the CD4+ TCR pathway, Choline metabolism in cancer - Homo sapiens (human), Colorectal cancer - Homo sapiens (human), HIF-1-alpha transcription factor network, IL2-mediated signaling events, IL4-mediated signaling events, LPA receptor mediated events, mets affect on macrophage differentiation, Oncostatin M Signaling Pathway, oxidative stress induced gene expression via nrf2, Quercetin and Nf-kB- AP-1 Induced Cell Apoptosis, RANKL-RANK Signaling Pathway, repression of pain sensation by the transcriptional regulator dream, role of egf receptor transactivation by gpcrs in cardiac hypertrophy, T cell receptor signaling pathway - Homo sapiens (human), Thyroid hormone synthesis, Thyroid hormone synthesis - Homo sapiens (human), Thyroxine (Thyroid Hormone) Production, trefoil factors initiate mucosal healing, tsp-1 induced apoptosis in microvascular endothelial cell

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Endocrine	THCA	Tumour	FGFR2, FGFR3, FGFR4, -arrestins in gpcr desensitization, 11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), Assembly of collagen fibrils and other multimeric structures, Collagen degradation, Collagen formation, Complement and Coagulation Cascades, cpdb_cancer_related_pathway, Degradation of the extracellular matrix, Extracellular matrix organization, Gastric cancer network 2, Hemostasis, Lysosome - Homo sapiens (human), Platelet activation, signaling and aggregation, Platelet degranulation , Prostaglandin Synthesis and Regulation, Renin secretion - Homo sapiens (human), Response to elevated platelet cytosolic Ca2+, Thyroid hormone synthesis - Homo sapiens (human), Trafficking and processing of endosomal TLR, Validated targets of C-MYC transcriptional repression, Vitamin D Receptor Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	CHOL	Normal	ABC transporters in lipid homeostasis, ABC-family proteins mediated transport, Bile acid and bile salt metabolism, Binding and Uptake of Ligands by Scavenger Receptors, EGFR1, Folate Metabolism, FOXA2 and FOXA3 transcription factor networks, HDL-mediated lipid transport, Hemostasis, Human Complement System, Lipid digestion, mobilization, and transport, Lipoprotein metabolism, Metabolism of lipids and lipoproteins, Platelet activation, signaling and aggregation, Platelet degranulation , Recycling of bile acids and salts, Response to elevated platelet cytosolic Ca2+, Scavenging of heme from plasma, Selenium Micronutrient Network, SLC-mediated transmembrane transport, Transmembrane transport of small molecules, Transport of organic anions, Transport of vitamins, nucleosides, and related molecules, Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling, Vitamin B12 Metabolism

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	CHOL	Tumour	Activation of C3 and C5, Alpha4 beta1 integrin signaling events, Alpha9 beta1 integrin signaling events, Alternative complement activation, alternative complement pathway, BDNF signaling pathway, Complement cascade, EGFR1, FGF signaling pathway, G alpha (i) signalling events, GPCR downstream signaling, Human Complement System, IL-6 signaling pathway, Immune System, Initial triggering of complement, Innate Immune System, lectin induced complement pathway, Osteopontin Signaling, Osteopontin-mediated events, Overview of nanoparticle effects, Regulation of toll-like receptor signaling pathway, regulators of bone mineralization, Toll-like receptor signaling pathway, Validated nuclear estrogen receptor beta network, Vitamin D Receptor Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	COADREAD	Normal	cAMP signaling pathway - Homo sapiens (human), cGMP-PKG signaling pathway - Homo sapiens (human), DAP12 interactions, DAP12 signaling, EPH-Ephrin signaling, EPHA-mediated growth cone collapse, Host Interactions of HIV factors, IL12-mediated signaling events, IL4-mediated signaling events, Integrin-linked kinase signaling, Intestinal immune network for IgA production - Homo sapiens (human), Muscle contraction, Nef-mediates down modulation of cell surface receptors by recruiting them to clathrin adapters, Pancreatic secretion - Homo sapiens (human), RHO GTPases activate CIT, RHO GTPases activate PAKs, RHO GTPases activate PKNs, RHO GTPases Activate ROCKs, Sema4D in semaphorin signaling, Sema4D induced cell migration and growth-cone collapse, Semaphorin interactions, Smooth Muscle Contraction, The role of Nef in HIV-1 replication and disease pathogenesis, Vascular smooth muscle contraction - Homo sapiens (human), Vitamin D Receptor Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	COADREAD	Tumour	3, -UTR-mediated translational regulation, Cap-dependent Translation Initiation, Cytoplasmic Ribosomal Proteins, Eukaryotic Translation Elongation, Eukaryotic Translation Initiation, Eukaryotic Translation Termination, Formation of a pool of free 40S subunits, Formation of the ternary complex, and subsequently, the 43S complex, GTP hydrolysis and joining of the 60S ribosomal subunit, IL11, IL2 signaling events mediated by PI3K, Interleukin-11 Signaling Pathway, L13a-mediated translational silencing of Ceruloplasmin expression, mtor signaling pathway, mTORC1-mediated signalling, Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC), Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC), Nonsense-Mediated Decay (NMD), Peptide chain elongation, PI3K Cascade, PKB-mediated events, Ribosomal scanning and start codon recognition, Ribosome - Homo sapiens (human), SRP-dependent cotranslational protein targeting to membrane, Translation

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	ESCA	Tumour	Adaptive Immune System, Antigen processing and presentation - Homo sapiens (human), B cell receptor signaling pathway - Homo sapiens (human), Class I MHC mediated antigen processing & presentation, DAP12 signaling, Disease, Glycine, serine, alanine and threonine metabolism, GPCR ligand binding, HIV Infection, HIV Life Cycle, Host Interactions of HIV factors, Immune System, Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell, Infectious disease, Integrin-mediated Cell Adhesion, Metabolism of non-coding RNA, Notch Signaling Pathway, notch_pathway, Pyrimidine metabolism, Regulation of Telomerase, Signaling by GPCR, snRNP Assembly, TGF-beta signaling pathway - Homo sapiens (human), Vascular smooth muscle contraction - Homo sapiens (human), Vitamin E metabolism

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	ESCA_EAC	Tumour	a6b1 and a6b4 Integrin signaling, Adherens junction - Homo sapiens (human), Aminosugars metabolism, Arrhythmogenic Right Ventricular Cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy (ARVC) - Homo sapiens (human), Bacterial invasion of epithelial cells - Homo sapiens (human), Collagen degradation, Dilated cardiomyopathy - Homo sapiens (human), eukaryotic protein translation, FAS pathway and Stress induction of HSP regulation, Fibroblast growth factor-1, Focal Adhesion, Focal adhesion - Homo sapiens (human), Hippo signaling pathway - Homo sapiens (human), Hypertrophic cardiomyopathy (HCM) - Homo sapiens (human), Integrin, Mitotic Prometaphase, Myometrial Relaxation and Contraction Pathways, Pathways in cancer - Homo sapiens (human), Protein processing in endoplasmic reticulum - Homo sapiens (human), Proteoglycans in cancer - Homo sapiens (human), Rap1 signaling pathway - Homo sapiens (human), Resolution of Sister Chromatid Cohesion, RHO GTPases Activate Formins, Shigellosis - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	ESCA_SCC	Tumour	Adherens junction - Homo sapiens (human), AndrogenReceptor, Arrhythmogenic Right Ventricular Cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy (ARVC) - Homo sapiens (human), AUF1 (hnRNP D0) destabilizes mRNA, Axon guidance, Bacterial invasion of epithelial cells - Homo sapiens (human), Developmental Biology, Dilated cardiomyopathy - Homo sapiens (human), EPH-Ephrin signaling, EPHB-mediated forward signaling, FAS pathway and Stress induction of HSP regulation, Fcgamma receptor (FCGR) dependent phagocytosis, Focal Adhesion, Focal adhesion - Homo sapiens (human), Hippo signaling pathway - Homo sapiens (human), Hypertrophic cardiomyopathy (HCM) - Homo sapiens (human), IL6, Influenza A - Homo sapiens (human), Leukocyte transendothelial migration - Homo sapiens (human), Myometrial Relaxation and Contraction Pathways, Oxytocin signaling pathway - Homo sapiens (human), p38 mapk signaling pathway, p38 MAPK Signaling Pathway, Signal Transduction

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	LIHC	Normal	Adenine phosphoribosyltransferase deficiency (APRT), Adenosine Deaminase Deficiency, Adenylosuccinate Lyase Deficiency, AICA-Ribosiduria, Androgen and estrogen biosynthesis and metabolism, Chemical carcinogenesis - Homo sapiens (human), Chylomicron-mediated lipid transport, Complement and Coagulation Cascades, Complement and coagulation cascades - Homo sapiens (human), Drug metabolism - cytochrome P450 - Homo sapiens (human), Galactose metabolism, Gout or Kelley-Seegmiller Syndrome, Heart Development, Lesch-Nyhan Syndrome (LNS), mechanism of gene regulation by peroxisome proliferators via ppara, Metabolism of nucleotides, Metabolism of xenobiotics by cytochrome P450 - Homo sapiens (human), Mitochondrial DNA depletion syndrome, Molybdenum Cofactor Deficiency, Myoadenylate deaminase deficiency, PPAR Alpha Pathway, Purine Metabolism, Purine Nucleoside Phosphorylase Deficiency, Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha), Retinoid metabolism and transport

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	LIHC	Tumour	Adaptive Immune System, Alzheimer,s disease - Homo sapiens (human), Alzheimers Disease, Apoptosis-related network due to altered Notch3 in ovarian cancer, Cell Cycle, cell_cycle_pathway, Chylomicron-mediated lipid transport, Clathrin derived vesicle budding, Complement and Coagulation Cascades, cpdb_cancer_related_pathway, Golgi Associated Vesicle Biogenesis, Immune System, Innate Immune System, Integrated Pancreatic Cancer Pathway, Iron uptake and transport, Membrane Trafficking, Metabolism of proteins, Mineral absorption - Homo sapiens (human), notch_pathway, NRF2 pathway, Nuclear Receptors Meta-Pathway, Peptide chain elongation, Scavenging by Class A Receptors, Statin Pathway, trans-Golgi Network Vesicle Budding

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	PAAD	Normal	FGFR2, FGFR3, FGFR4, 3, -UTR-mediated translational regulation, activation of csk by camp-dependent protein kinase inhibits signaling through the t cell receptor, Activation of gene expression by SREBF (SREBP), Adaptive Immune System, Adherens junction - Homo sapiens (human), African trypanosomiasis - Homo sapiens (human), Alendronate Action Pathway, Alpha4 beta1 integrin signaling events, Alpha9 beta1 integrin signaling events, Alzheimer,s disease - Homo sapiens (human), Amyotrophic lateral sclerosis (ALS), Amyotrophic lateral sclerosis (ALS) - Homo sapiens (human), Androgen receptor signaling pathway, AndrogenReceptor, Angiogenesis overview, antigen processing and presentation, Antigen processing and presentation - Homo sapiens (human), Arginine and proline metabolism - Homo sapiens (human), Arrhythmogenic Right Ventricular Cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy (ARVC) - Homo sapiens (human), Aryl Hydrocarbon Receptor Pathway, Atorvastatin Action Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	PAAD	Tumour	AGE-RAGE pathway, Anti-diabetic Drug Potassium Channel Inhibitors Pathway, Pharmacodynamics, Arf6 trafficking events, Differentiation Pathway, FoxO signaling pathway - Homo sapiens (human), Gastric cancer network 2, Glibenclamide Action Pathway, Gliclazide Action Pathway, GPCR signaling-cholera toxin, GPCR signaling-G alpha i, GPCR signaling-G alpha q, GPCR signaling-G alpha s Epac and ERK, GPCR signaling-G alpha s PKA and ERK, GPCR signaling-pertussis toxin, growth hormone signaling pathway, Insulin Pathway, Insulin processing, Insulin receptor recycling, Insulin secretion - Homo sapiens (human), insulin signaling pathway, Insulin Signalling, IRS activation, Leucine Stimulation on Insulin Signaling, Maturity onset diabetes of the young - Homo sapiens (human), Senescence and Autophagy in Cancer

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	STAD	Normal	Alcoholism - Homo sapiens (human), Alpha-defensins, Arachidonic acid metabolism, Aryl Hydrocarbon Receptor Pathway, Cytokine-cytokine receptor interaction - Homo sapiens (human), EPHA-mediated growth cone collapse, Fatty acid, triacylglycerol, and ketone body metabolism, Gastric pepsin release, GPCR signaling-cholera toxin, GPCRs, Class A Rhodopsin-like, Leukotriene metabolism, Metabolism, Metabolism of lipids and lipoproteins, Mitochondrial translation, Mitochondrial translation elongation, Mitochondrial translation initiation, Mitochondrial translation termination, Neuroactive ligand-receptor interaction - Homo sapiens (human), RHO GTPases activate CIT, RHO GTPases activate PAKs, RHO GTPases activate PKNs, RHO GTPases Activate ROCKs, Smooth Muscle Contraction, Transcription, Vascular smooth muscle contraction - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gastro-intestinal	STAD	Tumour	Adipogenesis, Alzheimer,s disease - Homo sapiens (human), Alzheimers Disease, Fanconi-bickel syndrome, Fructose-1,6-diphosphatase deficiency, gluconeogenesis, Glycerol Phosphate Shuttle, Glycogen Storage Disease Type 1A (GSD1A) or Von Gierke Disease, Glycogenosis, Type IA. Von gierke disease, Glycogenosis, Type IB, Glycogenosis, Type IC, Glycogenosis, Type VII. Tarui disease, glycolysis, Glycolysis, Glycolysis Gluconeogenesis, Huntington,s disease - Homo sapiens (human), Mitochondrial Electron Transport Chain, NADH repair, Oxidative phosphorylation - Homo sapiens (human), PCP/CE pathway, Phosphoenolpyruvate carboxykinase deficiency 1 (PEPCK1), repair_pathway, SIDS Susceptibility Pathways, Signaling by Wnt, Triosephosphate isomerase
Gynecologic	CESC_CAD	Tumour	Cell Cycle, Cell Cycle, Mitotic, cell_cycle_pathway, EGF-EGFR Signaling Pathway, Fatty acid, triacylglycerol, and ketone body metabolism, Gastric cancer network 2, IL6, insulin Mam, Insulin Signaling, Interferon type I signaling pathways, JAK STAT pathway and regulation, jak_pathway, Kit receptor signaling pathway, Leptin signaling pathway, M Phase, MAPK Signaling Pathway, MAPK signaling pathway - Homo sapiens (human), mapk_pathway, Metabolism of amino acids and derivatives, miR-targeted genes in lymphocytes - TarBase, miR-targeted genes in squamous cell - TarBase, Prostaglandin Synthesis and Regulation, Signaling mediated by p38-alpha and p38-beta, Signalling by NGF, Vitamin D Receptor Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gynecologic	CESC_SCC	Tumour	a6b1 and a6b4 Integrin signaling, Activation of BAD and translocation to mitochondria , Activation of BH3-only proteins, Cell Cycle, Cell cycle - Homo sapiens (human), cell cycle: g2/m checkpoint, cell_cycle_pathway, Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex, Class I PI3K signaling events mediated by Akt, DNA Damage Response, DNA strand elongation, E2F transcription factor network, EGF, Epigenetic regulation of gene expression, estrogen responsive protein efp controls cell cycle and breast tumors growth, Fas, FGF, FoxO family signaling, G1 to S cell cycle control, Intrinsic Pathway for Apoptosis, Meiosis, miRNA Regulation of DNA Damage Response, Mitotic Prometaphase, Nucleotide Excision Repair, Nucleotide excision repair - Homo sapiens (human)
Gynecologic	OV	Tumour	Alzheimer,s disease - Homo sapiens (human), Alzheimers Disease, Cori Cycle, downregulated of mta-3 in er-negative breast tumors, Fanconi-bickel syndrome, Fructose-1,6-diphosphatase deficiency, Gluconeogenesis, Glucose metabolism, Glycerol Phosphate Shuttle, Glycogen Storage Disease Type 1A (GSD1A) or Von Gierke Disease, Glycogenosis, Type IA. Von gierke disease, Glycogenosis, Type IB, Glycogenosis, Type IC, Glycogenosis, Type VII. Tarui disease, glycolysis, Glycolysis / Gluconeogenesis - Homo sapiens (human), Glycolysis Gluconeogenesis, HIF-1 signaling pathway - Homo sapiens (human), Iron metabolism in placenta, Methionine and cysteine metabolism, Mitochondrial Electron Transport Chain, NADH repair, Phosphoenolpyruvate carboxykinase deficiency 1 (PEPCK1), Triosephosphate isomerase, Validated targets of C-MYC transcriptional repression

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gynecologic	UCEC	Normal	Alpha6Beta4Integrin, Apoptosis-related network due to altered Notch3 in ovarian cancer, Apoptotic cleavage of cellular proteins, Apoptotic execution phase, Aurora B signaling, Caspase Cascade in Apoptosis, Caspase-mediated cleavage of cytoskeletal proteins, Cellular response to heat stress, EGFR1, HSF1 activation, Legionellosis - Homo sapiens (human), MicroRNAs in cancer - Homo sapiens (human), Muscle contraction, notch_pathway, PDGFR-beta signaling pathway, PLK1 signaling events, Primary Focal Segmental Glomerulosclerosis FSGS, Programmed Cell Death, Spinal Cord Injury, Striated Muscle Contraction, TCR Signaling Pathway, west nile virus, Wnt Canonical, Wnt Mammals, Wnt signaling pathway - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gynecologic	UCEC	Tumour	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S, Fanconi-bickel syndrome, Fructose-1,6-diphosphatase deficiency, Glucose metabolism, Glycerol Phosphate Shuttle, Glycogen Storage Disease Type 1A (GSD1A) or Von Gierke Disease, Glycogenosis, Type IA. Von gierke disease, Glycogenosis, Type IB, Glycogenosis, Type IC, Glycogenosis, Type VII. Tarui disease, glycolysis, Glycolysis, Glycolysis and Gluconeogenesis, Glycolysis Gluconeogenesis, Human Complement System, Iron metabolism in placenta, Methionine and cysteine metabolism, NADH repair, Phosphoenolpyruvate carboxykinase deficiency 1 (PEPCK1), Protein processing in endoplasmic reticulum - Homo sapiens (human), superpathway of conversion of glucose to acetyl CoA and entry into the TCA cycle, Triosephosphate isomerase, Validated targets of C-MYC transcriptional activation, Validated targets of C-MYC transcriptional repression, Warburg Effect

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Gynecologic	UCS	Tumour	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S, Adipogenesis, Amoebiasis - Homo sapiens (human), Apoptosis, Apoptosis-related network due to altered Notch3 in ovarian cancer, Cardiac Progenitor Differentiation, Collagen biosynthesis and modifying enzymes, Collagen formation, Endochondral Ossification, Extracellular matrix organization, Formation of the ternary complex, and subsequently, the 43S complex, GPCR signaling-cholera toxin, GPCR signaling-G alpha i, GPCR signaling-G alpha q, GPCR signaling-G alpha s Epac and ERK, GPCR signaling-G alpha s PKA and ERK, GPCR signaling-pertussis toxin, IGF-Core, miRNA targets in ECM and membrane receptors, Posttranslational regulation of adherens junction stability and dissassembly, Protein digestion and absorption - Homo sapiens (human), Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs), SHC-related events triggered by IGF1R, Syndecan-1-mediated signaling events, Translation initiation complex formation

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Head and Neck	HNSC	Normal	Class A/1 (Rhodopsin-like receptors), Corticotropin-releasing hormone, Endogenous TLR signaling, Fatty acid, triacylglycerol, and ketone body metabolism, G alpha (q) signalling events, Huntington,s disease - Homo sapiens (human), IL1 and megakaryotyces in obesity, Metabolism, Metabolism of amino acids and derivatives, Metabolism of lipids and lipoproteins, Metabolism of proteins, Mitochondrial translation, Mitochondrial translation initiation, Non-alcoholic fatty liver disease (NAFLD) - Homo sapiens (human), O-linked glycosylation, O-linked glycosylation of mucins, Organelle biogenesis and maintenance, Phase 1 - Functionalization of compounds, Phospholipid metabolism, Respiratory electron transport, Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins., RhoA signaling pathway, TCR, The citric acid (TCA) cycle and respiratory electron transport, Vitamin D Receptor Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Head and Neck	HNSC	Tumour	Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic proteins, APC/C:Cdc20 mediated degradation of mitotic proteins, Assembly Of The HIV Virion, Assembly of the pre-replicative complex, Association of licensing factors with the pre-replicative complex, CDK-mediated phosphorylation and removal of Cdc6, CDT1 association with the CDC6:ORC:origin complex, Cell Cycle Checkpoints, Corticotropin-releasing hormone, degradation of AXIN, degradation of DVL, DNA Replication Pre-Initiation, Ectoderm Differentiation, Fanconi Anemia pathway, G1/S DNA Damage Checkpoints, G2/M Checkpoints, G2/M DNA damage checkpoint, Glucocorticoid receptor regulatory network, Hh mutants abrogate ligand secretion, Hh mutants that don,t undergo autocatalytic processing are degraded by ERAD, IKK complex recruitment mediated by RIP1, M/G1 Transition, Membrane binding and targetting of GAG proteins, NOTCH2 Activation and Transmission of Signal to the Nucleus, Validated transcriptional targets of deltaNp63 isoforms

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Head and Neck	UVM	Tumour	3, -UTR-mediated translational regulation, Apoptotic cleavage of cellular proteins, Apoptotic execution phase, Cap-dependent Translation Initiation, Caspase-mediated cleavage of cytoskeletal proteins, Cytoplasmic Ribosomal Proteins, Degradation of Superoxides, Eukaryotic Translation Elongation, Eukaryotic Translation Initiation, Eukaryotic Translation Termination, eumelanin biosynthesis, Formation of a pool of free 40S subunits, Gene Expression, GTP hydrolysis and joining of the 60S ribosomal subunit, L13a-mediated translational silencing of Ceruloplasmin expression, Melanogenesis - Homo sapiens (human), Metabolism of proteins, MicroRNAs in cancer - Homo sapiens (human), Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC), Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC), Nonsense-Mediated Decay (NMD), Peptide chain elongation, Ribosome - Homo sapiens (human), SRP-dependent cotranslational protein targeting to membrane, Translation

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Hematologic	DLBC	Tumour	Antigen activates B Cell Receptor (BCR) leading to generation of second messengers, B cell receptor signaling, B Cell Receptor Signaling Pathway, BCR signaling pathway, Cleavage of Growing Transcript in the Termination Region , ctcf: first multivalent nuclear factor, DNA Damage Response, erk and pi-3 kinase are necessary for collagen binding in corneal epithelia, Gene Expression, Herpes simplex infection - Homo sapiens (human), IL4, MHC class II antigen presentation, miRNA Regulation of DNA Damage Response, mRNA Processing, mRNA Splicing, mRNA Splicing - Major Pathway, Post-Elongation Processing of the Transcript, Processing of Capped Intron-Containing Pre-mRNA, rho cell motility signaling pathway, RNA Polymerase II Transcription, RNA Polymerase II Transcription Termination, Signaling by the B Cell Receptor (BCR), Transcription, Transport of Mature Transcript to Cytoplasm, Tuberculosis - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Hematologic	LAML	Tumour	FGFR2, FGFR3, FGFR4, -arrestins in gpcr desensitization, 11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), 17-Beta Hydroxysteroid Dehydrogenase III Deficiency, 2-Methyl-3-Hydroxybutryl CoA Dehydrogenase Deficiency, 21-hydroxylase deficiency (CYP21), 3-Beta-Hydroxysteroid Dehydrogenase Deficiency, 3-Hydroxy-3-Methylglutaryl-CoA Lyase Deficiency, African trypanosomiasis - Homo sapiens (human), C-MYB transcription factor network, Erythrocytes take up carbon dioxide and release oxygen, Erythrocytes take up oxygen and release carbon dioxide, Folate Metabolism, hemoglobins chaperone, Malaria - Homo sapiens (human), Metabolism, O2/CO2 exchange in erythrocytes, RNA transport - Homo sapiens (human), Salivary secretion - Homo sapiens (human), Scavenging of heme from plasma, Selenium Micronutrient Network, Vitamin B12 Metabolism

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Hematologic	THYM	Normal	FGFR2, FGFR3, FGFR4, -arrestins in gpcr desensitization, 11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), 17-Beta Hydroxysteroid Dehydrogenase III Deficiency, 2-Methyl-3-Hydroxybutryl CoA Dehydrogenase Deficiency, 21-hydroxylase deficiency (CYP21), 3-Beta-Hydroxysteroid Dehydrogenase Deficiency, 3-Hydroxy-3-Methylglutaryl-CoA Lyase Deficiency, 3-hydroxyisobutyric acid dehydrogenase deficiency, 3-hydroxyisobutyric aciduria, 3-Methylcrotonyl Coa Carboxylase Deficiency Type I, 3-Methylglutaconic Aciduria Type I, 3-Methylglutaconic Aciduria Type III, 3-Methylglutaconic Aciduria Type IV, 3-Methylthiofentanyl Action Pathway, 3-Phosphoglycerate dehydrogenase deficiency, 3, -UTR-mediated translational regulation, A tetrasaccharide linker sequence is required for GAG synthesis, a6b1 and a6b4 Integrin signaling, Abacavir Pathway, Pharmacokinetics/Pharmacodynamics, ABC transporters - Homo sapiens (human), ABC transporters in lipid homeostasis

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Hematologic	THYM	Tumour	Adrenergic signaling in cardiomyocytes - Homo sapiens (human), Alcoholism - Homo sapiens (human), Amphetamine addiction - Homo sapiens (human), Calcium signaling pathway - Homo sapiens (human), Chemokine receptors bind chemokines, Chemokine signaling pathway - Homo sapiens (human), Circadian entrainment - Homo sapiens (human), Cytokine-cytokine receptor interaction - Homo sapiens (human), Dopaminergic synapse - Homo sapiens (human), G alpha (i) signalling events, Glioma - Homo sapiens (human), Glucagon signaling pathway - Homo sapiens (human), GnRH signaling pathway - Homo sapiens (human), GPCR ligand binding, Inflammatory mediator regulation of TRP channels - Homo sapiens (human), Intestinal immune network for IgA production - Homo sapiens (human), Long-term potentiation - Homo sapiens (human), Melanogenesis - Homo sapiens (human), Neurotrophin signaling pathway - Homo sapiens (human), Olfactory transduction - Homo sapiens (human), Peptide ligand-binding receptors, Phosphatidylinositol signaling system - Homo sapiens (human), Phototransduction - Homo sapiens (human), PPAR Alpha Pathway, Renin secretion - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Skin	SKCM	Normal	FGFR2, FGFR3, FGFR4, -arrestins in gpcr desensitization, 11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), 17-Beta Hydroxysteroid Dehydrogenase III Deficiency, 2-Methyl-3-Hydroxybutryl CoA Dehydrogenase Deficiency, 21-hydroxylase deficiency (CYP21), 3-Beta-Hydroxysteroid Dehydrogenase Deficiency, 3-Hydroxy-3-Methylglutaryl-CoA Lyase Deficiency, 3-hydroxyisobutyric acid dehydrogenase deficiency, 3-hydroxyisobutyric aciduria, 3-Methylcrotonyl Coa Carboxylase Deficiency Type I, 3-Methylglutaconic Aciduria Type I, 3-Methylglutaconic Aciduria Type III, 3-Methylglutaconic Aciduria Type IV, 3-Methylthiofentanyl Action Pathway, 3-Phosphoglycerate dehydrogenase deficiency, 3, -UTR-mediated translational regulation, A tetrasaccharide linker sequence is required for GAG synthesis, a6b1 and a6b4 Integrin signaling, Abacavir Pathway, Pharmacokinetics/Pharmacodynamics, ABC transporters - Homo sapiens (human), ABC transporters in lipid homeostasis

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Skin	SKCM	Tumour	Allograft Rejection, Alpha6Beta4Integrin, Apoptosis, Apoptosis-related network due to altered Notch3 in ovarian cancer, Apoptotic cleavage of cellular proteins, Apoptotic execution phase, Aurora B signaling, Caspase Cascade in Apoptosis, Caspase-mediated cleavage of cytoskeletal proteins, Common Pathway of Fibrin Clot Formation, Dissolution of Fibrin Clot, Epstein-Barr virus infection - Homo sapiens (human), Inflammasomes, Intrinsic Pathway of Fibrin Clot Formation, MicroRNAs in cancer - Homo sapiens (human), notch_pathway, Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways, Primary Focal Segmental Glomerulosclerosis FSGS, Programmed Cell Death, Regulation of Ras family activation, Sema3A PAK dependent Axon repulsion, Spinal Cord Injury, TCR Signaling Pathway, The NLRP3 inflammasome, Transport of fatty acids

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Soft Tissue	SARC	Normal	FGFR2, FGFR3, FGFR4, -arrestins in gpcr desensitization, 11-beta-hydroxylase deficiency (CYP11B1), 17-alpha-hydroxylase deficiency (CYP17), 17-Beta Hydroxysteroid Dehydrogenase III Deficiency, 2-Methyl-3-Hydroxybutryl CoA Dehydrogenase Deficiency, 21-hydroxylase deficiency (CYP21), 3-Beta-Hydroxysteroid Dehydrogenase Deficiency, 3-Hydroxy-3-Methylglutaryl-CoA Lyase Deficiency, 3-hydroxyisobutyric acid dehydrogenase deficiency, 3-hydroxyisobutyric aciduria, 3-Methylcrotonyl Coa Carboxylase Deficiency Type I, 3-Methylglutaconic Aciduria Type I, 3-Methylglutaconic Aciduria Type III, 3-Methylglutaconic Aciduria Type IV, 3-Methylthiofentanyl Action Pathway, 3-Phosphoglycerate dehydrogenase deficiency, 3, -UTR-mediated translational regulation, A tetrasaccharide linker sequence is required for GAG synthesis, a6b1 and a6b4 Integrin signaling, Abacavir Pathway, Pharmacokinetics/Pharmacodynamics, ABC transporters - Homo sapiens (human), ABC transporters in lipid homeostasis

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Soft Tissue	SARC	Tumour	Activation of Matrix Metalloproteinases, Alpha6Beta4Integrin, Amoebiasis - Homo sapiens (human), Apoptotic cleavage of cellular proteins, Apoptotic execution phase, Beta1 integrin cell surface interactions, Caspase Cascade in Apoptosis, Caspase-mediated cleavage of cytoskeletal proteins, Classical antibody-mediated complement activation, Collagen biosynthesis and modifying enzymes, Collagen formation, ECM-receptor interaction - Homo sapiens (human), Extracellular matrix organization, Inflammatory Response Pathway, Integrins in angiogenesis, MicroRNAs in cancer - Homo sapiens (human), miRNA targets in ECM and membrane receptors, NCAM1 interactions, pi3k_pathway, PI3K-Akt signaling pathway - Homo sapiens (human), Platelet Aggregation Inhibitor Pathway, Pharmacodynamics, Protein digestion and absorption - Homo sapiens (human), Regulation of Ras family activation, Syndecan-1-mediated signaling events, TCR Signaling Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Thoracic	LUAD	Normal	a6b1 and a6b4 Integrin signaling, Angiopoietin receptor Tie2-mediated signaling, Antigen Presentation: Folding, assembly and peptide loading of class I MHC, Apoptosis - Homo sapiens (human), Aryl Hydrocarbon Receptor, Bladder Cancer, C-type lectin receptors (CLRs), Chronic myeloid leukemia - Homo sapiens (human), Class I MHC mediated antigen processing & presentation, DNA Damage Response (only ATM dependent), ErbB Signaling Pathway, ErbB signaling pathway - Homo sapiens (human), Factors involved in megakaryocyte development and platelet production, Hedgehog, miR-targeted genes in epithelium - TarBase, miR-targeted genes in lymphocytes - TarBase, Nef mediated downregulation of MHC class I complex cell surface expression, p75(NTR)-mediated signaling, ras-independent pathway in nk cell-mediated cytotoxicity, Signaling Pathways in Glioblastoma, Sphingolipid signaling pathway - Homo sapiens (human), Wnt Canonical, Wnt Mammals, Wnt Signaling Pathway, Wnt signaling pathway - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Thoracic	LUAD	Tumour	Allograft rejection - Homo sapiens (human), Antigen processing and presentation - Homo sapiens (human), Autoimmune thyroid disease - Homo sapiens (human), Binding and Uptake of Ligands by Scavenger Receptors, Cell adhesion molecules (CAMs) - Homo sapiens (human), Clathrin derived vesicle budding, Gastric cancer network 2, Golgi Associated Vesicle Biogenesis, Graft-versus-host disease - Homo sapiens (human), Herpes simplex infection - Homo sapiens (human), Integrated Pancreatic Cancer Pathway, Iron uptake and transport, Lysosome - Homo sapiens (human), Membrane Trafficking, MHC class II antigen presentation, Mineral absorption - Homo sapiens (human), NRF2 pathway, Nuclear Receptors Meta-Pathway, Prostaglandin Synthesis and Regulation, Scavenging by Class A Receptors, trans-Golgi Network Vesicle Budding, Transmembrane transport of small molecules, Tuberculosis - Homo sapiens (human), Type I diabetes mellitus - Homo sapiens (human), Viral myocarditis - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Thoracic	LUSC	Normal	Activation of gene expression by SREBF (SREBP), Alpha-synuclein signaling, Bile secretion - Homo sapiens (human), Chemokine receptors bind chemokines, Class A/1 (Rhodopsin-like receptors), Formation of Fibrin Clot (Clotting Cascade), FOXA1 transcription factor network, Fructose Mannose metabolism, G alpha (i) signalling events, G alpha (q) signalling events, Glycine, serine and threonine metabolism - Homo sapiens (human), GPCR downstream signaling, GPCR ligand binding, Human Complement System, Intrinsic Pathway of Fibrin Clot Formation, Peptide ligand-binding receptors, Pertussis - Homo sapiens (human), Phagosome - Homo sapiens (human), Primary immunodeficiency - Homo sapiens (human), Proton Pump Inhibitor Pathway, Pharmacodynamics, Regulation of CDC42 activity, Regulation of cholesterol biosynthesis by SREBP (SREBF), Regulation of toll-like receptor signaling pathway, SREBP signalling, Steroid biosynthesis - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Thoracic	LUSC	Tumour	Alzheimer,s disease - Homo sapiens (human), Alzheimers Disease, Chromosome Maintenance, Cori Cycle, Ectoderm Differentiation, Fanconi-bickel syndrome, Fructose-1,6-diphosphatase deficiency, Glucagon signaling pathway - Homo sapiens (human), gluconeogenesis, Gluconeogenesis, Glucose metabolism, Glucose transport, Glycerol Phosphate Shuttle, Glycogen Storage Disease Type 1A (GSD1A) or Von Gierke Disease, Glycogenosis, Type IA. Von gierke disease, Glycogenosis, Type IB, Glycogenosis, Type IC, Glycogenosis, Type VII. Tarui disease, glycolysis, Glycolysis, Glycolysis / Gluconeogenesis - Homo sapiens (human), Hexose transport, hypoxia-inducible factor in the cardivascular system, Inositol phosphate metabolism - Homo sapiens (human), Olfactory transduction - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Thoracic	MESO	Tumour	A tetrasaccharide linker sequence is required for GAG synthesis, Acenocoumarol Action Pathway, Activation of C3 and C5, Alteplase Action Pathway, Alternative complement activation, Classical antibody-mediated complement activation, classical complement pathway, Collagen biosynthesis and modifying enzymes, Collagen formation, Complement Activation, Classical Pathway, Complement and Coagulation Cascades, Complement and coagulation cascades - Homo sapiens (human), Complement cascade, Copper homeostasis, Creation of C4 and C2 activators, Dermatan sulfate biosynthesis, ECM proteoglycans, Inflammatory Response Pathway, Initial triggering of complement, Osteoblast Signaling, Regulation of Complement cascade, Scavenging by Class H Receptors, Senescence and Autophagy in Cancer, Signaling mediated by p38-alpha and p38-beta, Syndecan-1-mediated signaling events

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	BLCA	Normal	ATF-2 transcription factor network, Cholinergic synapse - Homo sapiens (human), EPO signaling, erbb_pathway, Hepatitis B - Homo sapiens (human), IL-6 signaling pathway, IL-7 signaling, IL2-mediated signaling events, IL6-mediated signaling events, JAK STAT pathway and regulation, jak_pathway, Jak-STAT signaling pathway - Homo sapiens (human), MAPK signaling pathway - Homo sapiens (human), Oncostatin_M, PI3K/AKT activation, Prolactin, Prolactin signaling pathway - Homo sapiens (human), Signaling by EGFR, Signaling by ERBB4, Signaling by the B Cell Receptor (BCR), Signalling by NGF, TGF_beta_Receptor, TGF-beta signaling pathway - Homo sapiens (human), TNF signaling pathway - Homo sapiens (human), VEGF

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	BLCA	Tumour	AUF1 (hnRNP D0) destabilizes mRNA, cell_cycle_pathway, Deadenylation-dependent mRNA decay, Direct p53 effectors, FoxO signaling pathway - Homo sapiens (human), Gastric Cancer Network 1, Gastric cancer network 2, Gastrin, IL1 and megakaryotyces in obesity, Insulin signaling pathway - Homo sapiens (human), mRNA Processing, mRNA Splicing, mRNA Splicing - Major Pathway, mtor_pathway, p73 transcription factor network, Processing of Capped Intron-Containing Pre-mRNA, Prostaglandin Synthesis and Regulation, Regulation of mRNA stability by proteins that bind AU-rich elements, Signaling mediated by p38-alpha and p38-beta, skeletal muscle hypertrophy is regulated via akt-mtor pathway, TP53 Regulates Metabolic Genes, Transcriptional Regulation by TP53, Validated transcriptional targets of TAp63 isoforms, Viral carcinogenesis - Homo sapiens (human), Vitamin D Receptor Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	KICH	Normal	Basal transcription factors - Homo sapiens (human), Cell cycle - Homo sapiens (human), cell_cycle_pathway, Cellular Senescence, DNA Damage Response, Eukaryotic Transcription Initiation, G1 to S cell cycle control, HIV Life Cycle, HIV Transcription Initiation, Late Phase of HIV Life Cycle, MAPK family signaling cascades, mapk_pathway, miRNA Regulation of DNA Damage Response, Pyrimidine metabolism - Homo sapiens (human), Pyrimidine nucleotides nucleosides metabolism, RNA Polymerase II HIV Promoter Escape, RNA Polymerase II Pre-transcription Events, RNA Polymerase II Promoter Escape, RNA Polymerase II Transcription, RNA Polymerase II Transcription Initiation, RNA Polymerase II Transcription Initiation And Promoter Clearance, RNA Polymerase II Transcription Pre-Initiation And Promoter Opening, Senescence and Autophagy in Cancer, Senescence-Associated Secretory Phenotype (SASP), Transcription

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	KICH	Tumour	adenosine ribonucleotides de novo biosynthesis, Beta defensins, Ceramide signaling pathway, Collecting duct acid secretion - Homo sapiens (human), Defensins, Electron Transport Chain, Epithelial cell signaling in Helicobacter pylori infection - Homo sapiens (human), Formation of ATP by chemiosmotic coupling, Huntington,s disease - Homo sapiens (human), Latent infection of Homo sapiens with Mycobacterium tuberculosis, LKB1 signaling events, Metabolism of Angiotensinogen to Angiotensins, Oxidative phosphorylation - Homo sapiens (human), Peptide hormone metabolism, Phagosomal maturation (early endosomal stage), Prolactin, Prolactin Signaling Pathway, purine nucleotides de novo biosynthesis, Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins., Sphingolipid signaling pathway - Homo sapiens (human), Synaptic vesicle cycle - Homo sapiens (human), The citric acid (TCA) cycle and respiratory electron transport, thyroid hormone biosynthesis, Transferrin endocytosis and recycling, Validated nuclear estrogen receptor alpha network

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	KIRC	Normal	3-Methylthiofentanyl Action Pathway, Alfentanil Action Pathway, Alvimopan Action Pathway, Amiloride Action Pathway, Anileridine Action Pathway, Basigin interactions, Bendroflumethiazide Action Pathway, Benzocaine Action Pathway, Bile secretion - Homo sapiens (human), Blue diaper syndrome, Bumetanide Action Pathway, Bupivacaine Action Pathway, Buprenorphine Action Pathway, Carbohydrate digestion and absorption - Homo sapiens (human), Carfentanil Action Pathway, Chloroprocaine Action Pathway, Chlorothiazide Action Pathway, Chlorthalidone Action Pathway, Cocaine Action Pathway, Cyclothiazide Action Pathway, Cystinuria, Desipramine Action Pathway, Dezocine Action Pathway, Dibucaine Action Pathway, Dihydromorphine Action Pathway

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	KIRC	Tumour	Allograft Rejection, Apoptotic cleavage of cellular proteins, Apoptotic execution phase, Aurora B signaling, Caspase-mediated cleavage of cytoskeletal proteins, Celecoxib Pathway, Pharmacodynamics, Cori Cycle, Fanconi-bickel syndrome, Fructose-1,6-diphosphatase deficiency, Glutathione metabolism, glutathione redox reactions I, Glycogen Storage Disease Type 1A (GSD1A) or Von Gierke Disease, Glycogenosis, Type IA. Von gierke disease, Glycogenosis, Type IB, Glycogenosis, Type IC, Glycogenosis, Type VII. Tarui disease, glycolysis, Glycolysis / Gluconeogenesis - Homo sapiens (human), Glycolysis and Gluconeogenesis, Glycolysis Gluconeogenesis, HIF-1 signaling pathway - Homo sapiens (human), MicroRNAs in cancer - Homo sapiens (human), PLK1 signaling events, reactive oxygen species degradation, Validated targets of C-MYC transcriptional activation

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	KIRP	Normal	ABC transporters - Homo sapiens (human), Aflatoxin activation and detoxification, Clathrin derived vesicle budding, Detoxification of Reactive Oxygen Species, Folate metabolism, Glycine Serine metabolism, Glycine, serine and threonine metabolism - Homo sapiens (human), Glyoxylate and dicarboxylate metabolism - Homo sapiens (human), Golgi Associated Vesicle Biogenesis, Integrated Pancreatic Cancer Pathway, Iron uptake and transport, LKB1 signaling events, Lysosome - Homo sapiens (human), Membrane Trafficking, Metabolism of amino acids and derivatives, Metabolism of Angiotensinogen to Angiotensins, N-Glycan biosynthesis, NRF2 pathway, Nuclear Receptors in Lipid Metabolism and Toxicity, Nuclear Receptors Meta-Pathway, Scavenging by Class A Receptors, SLC-mediated transmembrane transport, thyroid hormone biosynthesis, trans-Golgi Network Vesicle Budding, Transmembrane transport of small molecules

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	KIRP	Tumour	Alpha4 beta1 integrin signaling events, Alpha9 beta1 integrin signaling events, BDNF signaling pathway, Beta1 integrin cell surface interactions, Beta3 integrin cell surface interactions, Degradation of the extracellular matrix, Direct p53 effectors, ECM-receptor interaction - Homo sapiens (human), Endochondral Ossification, FGF signaling pathway, Human Complement System, Integrin cell surface interactions, Integrins in angiogenesis, Osteoclast Signaling, Osteopontin Signaling, Osteopontin-mediated events, p53_pathway, PI3K-Akt signaling pathway - Homo sapiens (human), Protein processing in endoplasmic reticulum - Homo sapiens (human), Regulation of toll-like receptor signaling pathway, regulators of bone mineralization, Signaling by PDGF, TGF Beta Signaling Pathway, Toll-like receptor signaling pathway, Toll-like receptor signaling pathway - Homo sapiens (human)

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	PRAD	Normal	Alendronate Action Pathway, Atorvastatin Action Pathway, Cerivastatin Action Pathway, CHILD Syndrome, Cholesterol biosynthesis, Cholesterol Biosynthesis, Cholesteryl ester storage disease, Chondrodysplasia Punctata II, X Linked Dominant (CDPX2), Circadian entrainment - Homo sapiens (human), Desmosterolosis, EPHA-mediated growth cone collapse, Fluvastatin Action Pathway, FOXM1 transcription factor network, Glutathione conjugation, glutathione-mediated detoxification, Miscellaneous transport and binding events, Muscle contraction, RHO GTPases activate CIT, RHO GTPases activate PAKs, RHO GTPases Activate ROCKs, Sema4D in semaphorin signaling, Sema4D induced cell migration and growth-cone collapse, Smooth Muscle Contraction, Viral RNP Complexes in the Host Cell Nucleus, Vitamin B2 (riboflavin) metabolism

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	PRAD	Tumour	Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3, Acyl chain remodelling of PC, Acyl chain remodelling of PE, Acyl chain remodelling of PG, Acyl chain remodelling of PI, Acyl chain remodelling of PS, alpha-Linolenic acid metabolism - Homo sapiens (human), Androgen receptor signaling pathway, antigen processing and presentation, Cardiac Hypertrophic Response, Coregulation of Androgen receptor activity, DroToll-like, Ether lipid metabolism - Homo sapiens (human), FOXA1 transcription factor network, Glycerophospholipid metabolism - Homo sapiens (human), IGF signaling, Linoleic acid metabolism - Homo sapiens (human), Metabolism of proteins, Pathways in cancer - Homo sapiens (human), phospholipases, Prostate Cancer, Prostate cancer - Homo sapiens (human), Regulation of Androgen receptor activity, Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs), RHO GTPases activate PKNs

Table 2: Top 25 statistically identified pathways for each common cancer and normal tissue category in TCGA, based on PIE scores. *(continued)*

Organ-System of origin	Cancer Code	Type	Top 25 pathways
Urologic	TGCT	Tumour	Alzheimer,s disease - Homo sapiens (human), Alzheimers Disease, Cori Cycle, downregulated of mta-3 in er-negative breast tumors, Fanconi-bickel syndrome, Fructose-1,6-diphosphatase deficiency, gluconeogenesis, Gluconeogenesis, Glucose metabolism, Glycerol Phosphate Shuttle, Glycogen Storage Disease Type 1A (GSD1A) or Von Gierke Disease, Glycogenosis, Type IA. Von gierke disease, Glycogenosis, Type IB, glycolysis, Glycolysis, Glycolysis / Gluconeogenesis - Homo sapiens (human), Glycolysis and Gluconeogenesis, Glycolysis Gluconeogenesis, Metabolism of carbohydrates, POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation, POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation, superpathway of conversion of glucose to acetyl CoA and entry into the TCA cycle, Transcriptional regulation of pluripotent stem cells, Validated targets of C-MYC transcriptional activation, Warburg Effect

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores.

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	BRCA	Breast	60	Amphetamine addiction - Homo sapiens (human), Aromatase Inhibitor Pathway (Breast Cell), Pharmacodynamics, Astrocytic Glutamate-Glutamine Uptake And Metabolism, ATF6-alpha activates chaperone genes, Dopaminergic synapse - Homo sapiens (human), Estrogen signaling pathway - Homo sapiens (human), FTO Obesity Variant Mechanism, IL27-mediated signaling events, Inflammasomes, Inflammatory bowel disease (IBD) - Homo sapiens (human), Inflammatory mediator regulation of TRP channels - Homo sapiens (human), Insulin Signaling, miRNA targets in ECM and membrane receptors, Miscellaneous transport and binding events, Neurotransmitter uptake and Metabolism In Glial Cells, Oocyte meiosis - Homo sapiens (human), PI3K-Akt signaling pathway - Homo sapiens (human), Progesterone-mediated oocyte maturation - Homo sapiens (human), Protein processing in endoplasmic reticulum - Homo sapiens (human), Signaling by Retinoic Acid, Signaling events mediated by VEGFR1 and VEGFR2, The NLRP3 inflammasome, Validated nuclear estrogen receptor alpha network, Vitamin A and Carotenoid Metabolism, Warfarin Pathway, Pharmacodynamics

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
POG	BRCA	Breast	160	Amphetamine addiction - Homo sapiens (human), Aromatase Inhibitor Pathway (Breast Cell), Pharmacodynamics, Astrocytic Glutamate-Glutamine Uptake And Metabolism, ATF6-alpha activates chaperone genes, Dopaminergic synapse - Homo sapiens (human), Estrogen signaling pathway - Homo sapiens (human), FTO Obesity Variant Mechanism, IL27-mediated signaling events, Inflammasomes, Inflammatory bowel disease (IBD) - Homo sapiens (human), Inflammatory mediator regulation of TRP channels - Homo sapiens (human), Insulin Signaling, miRNA targets in ECM and membrane receptors, Miscellaneous transport and binding events, Neurotransmitter uptake and Metabolism In Glial Cells, Oocyte meiosis - Homo sapiens (human), PI3K-Akt signaling pathway - Homo sapiens (human), Progesterone-mediated oocyte maturation - Homo sapiens (human), Protein processing in endoplasmic reticulum - Homo sapiens (human), Signaling by Retinoic Acid, Signaling events mediated by VEGFR1 and VEGFR2, The NLRP3 inflammasome, Validated nuclear estrogen receptor alpha network, Vitamin A and Carotenoid Metabolism, Warfarin Pathway, Pharmacodynamics

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	CHOL	Gastrointes	21	ABC transporter disorders, alternative complement pathway, Anchoring fibril formation, Basal transcription factors - Homo sapiens (human), Beta oxidation of myristoyl-CoA to lauroyl-CoA, BMP Signalling and Regulation, BMP2 signaling TAK1, choline degradation, control of skeletal myogenesis by hdac and calcium/calmodulin-dependent kinase (camk), Defective CFTR causes cystic fibrosis, Disorders of transmembrane transporters, Felodipine Metabolism Pathway, folate polyglutamylation, geranylgeranyldiphosphate biosynthesis, glycine cleavage, Glycine Metabolism, glycine/serine biosynthesis, Heroin metabolism, Leukotriene modifiers pathway, Pharmacodynamics, LRR FLII-interacting protein 1 (LRRFIP1) activates type I IFN production, lysine degradation I (saccharopine pathway), lysine degradation II (pipecolate pathway), Resolution of AP sites via the single-nucleotide replacement pathway, Trafficking of myristoylated proteins to the cilium, WNT ligand secretion is abrogated by the PORCN inhibitor LGK974

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
POG	CHOL	Gastrointestinal	17	ABC transporter disorders, alternative complement pathway, Anchoring fibril formation, Basal transcription factors - Homo sapiens (human), Beta oxidation of myristoyl-CoA to lauroyl-CoA, BMP Signalling and Regulation, BMP2 signaling TAK1, choline degradation, control of skeletal myogenesis by hdac and calcium/calmodulin-dependent kinase (camk), Defective CFTR causes cystic fibrosis, Disorders of transmembrane transporters, Felodipine Metabolism Pathway, folate polyglutamylation, geranylgeranyldiphosphate biosynthesis, glycine cleavage, Glycine Metabolism, glycine/serine biosynthesis, Heroin metabolism, Leukotriene modifiers pathway, Pharmacodynamics, LRR FLII-interacting protein 1 (LRRFIP1) activates type I IFN production, lysine degradation I (saccharopine pathway), lysine degradation II (pipecolate pathway), Resolution of AP sites via the single-nucleotide replacement pathway, Trafficking of myristoylated proteins to the cilium, WNT ligand secretion is abrogated by the PORCN inhibitor LGK974

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	COADR	Gastrointes	10	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S, Advanced glycosylation endproduct receptor signaling, bupropion degradation, Carnitine Synthesis, Conjugation of phenylacetate with glutamine, creatine-phosphate biosynthesis, DEx/H-box helicases activate type I IFN and inflammatory cytokines production , Dorso-ventral axis formation - Homo sapiens (human), EGFR Transactivation by Gastrin, Eicosanoid Synthesis, ErbB receptor signaling network, Formation of the ternary complex, and subsequently, the 43S complex, G-protein mediated events, Glycosphingolipid biosynthesis - lactoseries, HIF-1-alpha transcription factor network, icosapentaenoate biosynthesis II (metazoa), insulin Mam, mucin core 1 and core 2 O-glycosylation, Mucin type O-Glycan biosynthesis - Homo sapiens (human), PCSK9-mediated LDLR degradation, pentose phosphate pathway, Pentose phosphate pathway (hexose monophosphate shunt), Phenylacetate Metabolism, PLC beta mediated events, yaci and bcma stimulation of b cell immune responses

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
POG	COADREAD	Gastrointestinal	98	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S, Advanced glycosylation endproduct receptor signaling, bupropion degradation, Carnitine Synthesis, Conjugation of phenylacetate with glutamine, creatine-phosphate biosynthesis, DEx/H-box helicases activate type I IFN and inflammatory cytokines production , Dorso-ventral axis formation - Homo sapiens (human), EGFR Transactivation by Gastrin, Eicosanoid Synthesis, ErbB receptor signaling network, Formation of the ternary complex, and subsequently, the 43S complex, G-protein mediated events, Glycosphingolipid biosynthesis - lactoseries, HIF-1-alpha transcription factor network, icosapentaenoate biosynthesis II (metazoa), insulin Mam, mucin core 1 and core 2 O-glycosylation, Mucin type O-Glycan biosynthesis - Homo sapiens (human), PCSK9-mediated LDLR degradation, pentose phosphate pathway, Pentose phosphate pathway (hexose monophosphate shunt), Phenylacetate Metabolism, PLC beta mediated events, yaci and bcma stimulation of b cell immune responses

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	ESCA	Gastrointes	14	Adherens junction - Homo sapiens (human), Arrhythmogenic Right Ventricular Cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy (ARVC) - Homo sapiens (human), Chaperonin-mediated protein folding, chromatin remodeling by hswi/snf atp-dependent complexes, Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding, EGFR Transactivation by Gastrin, Folding of actin by CCT/TriC, Gastric acid secretion - Homo sapiens (human), Gastric Histamine Release, Hippo signaling pathway - Homo sapiens (human), hop pathway in cardiac development, Influenza A - Homo sapiens (human), Interleukin-6 signaling, IRAK1 recruits IKK complex, IRAK1 recruits IKK complex upon TLR7/8 or 9 stimulation, MAP kinase cascade, MAPK1 (ERK2) activation, miRs in Muscle Cell Differentiation, NOTCH2 intracellular domain regulates transcription, pkc-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase, Rap1 signaling pathway - Homo sapiens (human), Regulation of Actin Cytoskeleton, the information processing pathway at the ifn beta enhancer, WNT mediated activation of DVL

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	OV	Gynecologic	13	Abciximab Action Pathway, Adipogenesis, Biosynthesis of A2E, implicated in retinal degradation, BMP signaling Dro, Cell-Cell communication, Diseases associated with visual transduction, Dopamine receptors, Eptifibatide Action Pathway, ERKs are inactivated, Gastric pepsin release, Generic Transcription Pathway, Glycosylphosphatidylinositol(GPI)-anchor biosynthesis - Homo sapiens (human), Inhibition of PKR, Lipoic acid metabolism - Homo sapiens (human), Monoamine GPCRs, NHR, NOTCH2 intracellular domain regulates transcription, Retinoic acid receptors-mediated signaling, Retinoid cycle disease events, reversal of insulin resistance by leptin, S1P4 pathway, Signal attenuation, signal dependent regulation of myogenesis by corepressor mitr, TCA Cycle Nutrient Utilization and Invasiveness of Ovarian Cancer, thymine degradation
POG	OV	Gynecologi	33	Abciximab Action Pathway, Adipogenesis, Biosynthesis of A2E, implicated in retinal degradation, BMP signaling Dro, Cell-Cell communication, Diseases associated with visual transduction, Dopamine receptors, Eptifibatide Action Pathway, ERKs are inactivated, Gastric pepsin release, Generic Transcription Pathway, Glycosylphosphatidylinositol(GPI)-anchor biosynthesis - Homo sapiens (human), Inhibition of PKR, Lipoic acid metabolism - Homo sapiens (human), Monoamine GPCRs, NHR, NOTCH2 intracellular domain regulates transcription, Retinoic acid receptors-mediated signaling, Retinoid cycle disease events, reversal of insulin resistance by leptin, S1P4 pathway, Signal attenuation, signal dependent regulation of myogenesis by corepressor mitr, TCA Cycle Nutrient Utilization and Invasiveness of Ovarian Cancer, thymine degradation

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	SKCM	Skin	12	APEX1-Independent Resolution of AP Sites via the Single Nucleotide Replacement Pathway, Apoptosis-related network due to altered Notch3 in ovarian cancer, BMP receptor signaling, Codeine and Morphine Metabolism, Codeine and Morphine Pathway, Pharmacokinetics, Common Pathway of Fibrin Clot Formation, Conversion from APC/C:Cdc20 to APC/C:Cdh1 in late anaphase, Drug Induction of Bile Acid Pathway, Fanconi anemia pathway, FGFR4 mutant receptor activation, Formation of Fibrin Clot (Clotting Cascade), Inactivation of APC/C via direct inhibition of the APC/C complex, Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components, Intrinsic Pathway of Fibrin Clot Formation, lanosterol biosynthesis, Maturity onset diabetes of the young - Homo sapiens (human), Melanin biosynthesis, NICD traffics to nucleus, Nicotine Pathway, Pharmacokinetics, Notch-HLH transcription pathway, NOTCH2 intracellular domain regulates transcription, Post-transcriptional silencing by small RNAs, spermine biosynthesis, Transport of fatty acids, WNT ligand secretion is abrogated by the PORCN inhibitor LGK974

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
POG	SKCM	Skin	15	APEX1-Independent Resolution of AP Sites via the Single Nucleotide Replacement Pathway, Apoptosis-related network due to altered Notch3 in ovarian cancer, BMP receptor signaling, Codeine and Morphine Metabolism, Codeine and Morphine Pathway, Pharmacokinetics, Common Pathway of Fibrin Clot Formation, Conversion from APC/C:Cdc20 to APC/C:Cdh1 in late anaphase, Drug Induction of Bile Acid Pathway, Fanconi anemia pathway, FGFR4 mutant receptor activation, Formation of Fibrin Clot (Clotting Cascade), Inactivation of APC/C via direct inhibition of the APC/C complex, Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components, Intrinsic Pathway of Fibrin Clot Formation, lanosterol biosynthesis, Maturity onset diabetes of the young - Homo sapiens (human), Melanin biosynthesis, NICD traffics to nucleus, Nicotine Pathway, Pharmacokinetics, Notch-HLH transcription pathway, NOTCH2 intracellular domain regulates transcription, Post-transcriptional silencing by small RNAs, spermine biosynthesis, Transport of fatty acids, WNT ligand secretion is abrogated by the PORCN inhibitor LGK974

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	SARC	Soft Tissue	53	Activation of BMF and translocation to mitochondria, Amoebiasis - Homo sapiens (human), AMPK Signaling, Assembly of collagen fibrils and other multimeric structures, Asymmetric localization of PCP proteins, Beta1 integrin cell surface interactions, BMP Signalling Pathway, Collagen biosynthesis and modifying enzymes, Collagen formation, ECM-receptor interaction - Homo sapiens (human), Extracellular matrix organization, miRNA targets in ECM and membrane receptors, NCAM signaling for neurite out-growth, NCAM1 interactions, Post-translational protein modification, role of erk5 in neuronal survival pathway, Signaling by FGFR2 in disease, Signaling by FGFR2 mutants, Signaling by FGFR3 in disease, Signaling by FGFR3 mutants, Signaling by FGFR4 mutants, Smooth Muscle Contraction, Stimuli-sensing channels, Syndecan-1-mediated signaling events, Uptake and actions of bacterial toxins
POG	SARC	Soft Tissue	60	Activation of BMF and translocation to mitochondria, Amoebiasis - Homo sapiens (human), AMPK Signaling, Assembly of collagen fibrils and other multimeric structures, Asymmetric localization of PCP proteins, Beta1 integrin cell surface interactions, BMP Signalling Pathway, Collagen biosynthesis and modifying enzymes, Collagen formation, ECM-receptor interaction - Homo sapiens (human), Extracellular matrix organization, miRNA targets in ECM and membrane receptors, NCAM signaling for neurite out-growth, NCAM1 interactions, Post-translational protein modification, role of erk5 in neuronal survival pathway, Signaling by FGFR2 in disease, Signaling by FGFR2 mutants, Signaling by FGFR3 in disease, Signaling by FGFR3 mutants, Signaling by FGFR4 mutants, Smooth Muscle Contraction, Stimuli-sensing channels, Syndecan-1-mediated signaling events, Uptake and actions of bacterial toxins

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
POG	LUAD	Thoracic	53	-oxidation (unsaturated, odd number), Aflatoxin B1 metabolism, Arachidonate production from DAG, Cilostazol Action Pathway, Clathrin derived vesicle budding, Dipyridamole (Antiplatelet) Action Pathway, dolichol and dolichyl phosphate biosynthesis, Estrogen metabolism, extrinsic prothrombin activation pathway, Golgi Associated Vesicle Biogenesis, Inhibition of PKR, Insulin processing, Iron uptake and transport, Lipoic acid metabolism - Homo sapiens (human), Membrane Trafficking, mucin core 1 and core 2 O-glycosylation, Mucin type O-Glycan biosynthesis - Homo sapiens (human), Nicotine Metabolism, Nicotine Metabolism Pathway, NRF2 pathway, Opsins, pkc-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase, Termination of O-glycan biosynthesis, trans-Golgi Network Vesicle Budding, Transmembrane transport of small molecules
MET500	BLCA	Urologic	14	(S)-reticuline biosynthesis, [2Fe-2S] iron-sulfur cluster biosynthesis, adenine and adenosine salvage II, biotin-carboxyl carrier protein assembly, Cap-dependent Translation Initiation, CDC6 association with the ORC:origin complex, estradiol biosynthesis I, estradiol biosynthesis II, Estrogen biosynthesis, Eukaryotic Translation Initiation, Joubert syndrome, L-dopachrome biosynthesis, Myometrial Relaxation and Contraction Pathways, NAD phosphorylation and dephosphorylation, Phosphatidylinositol Phosphate Metabolism, phospholipase c delta in phospholipid associated cell signaling, Porphyrin metabolism, Reuptake of GABA, Ribosome - Homo sapiens (human), RNA degradation - Homo sapiens (human), Shigellosis - Homo sapiens (human), Thiamine metabolism - Homo sapiens (human), Thyroxine (Thyroid Hormone) Production, Vitamin A (retinol) metabolism, Xenobiotics metabolism

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
MET500	PRAD	Urologic	62	-oxidation (unsaturated, odd number), Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3, Amino Acid conjugation, Androgen receptor signaling pathway, antigen processing and presentation, Conjugation of benzoate with glycine, Conjugation of carboxylic acids, Conjugation of phenylacetate with glutamine, Coregulation of Androgen receptor activity, DroToll-like, FOXA1 transcription factor network, IGF signaling, Metabolism of proteins, Pathways in cancer - Homo sapiens (human), Phenylacetate Metabolism, Prostate Cancer, Prostate cancer - Homo sapiens (human), Regulation of Androgen receptor activity, Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs), Renin-angiotensin system - Homo sapiens (human), RHO GTPase Effectors, RHO GTPases activate PKNs, Secretion of Hydrochloric Acid in Parietal Cells, Signaling by Rho GTPases, spermine and spermidine degradation I

Table 3: Top 25 statistically identified pathways for each common cancer category in the POG and MET500 cohorts, based on PIE scores. *(continued)*

Cohort	Cancer Type	Organ-System of origin	Number of samples	Top 25 pathways
POG	PRAD	Urologic	74	-oxidation (unsaturated, odd number), Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3, Amino Acid conjugation, Androgen receptor signaling pathway, antigen processing and presentation, Conjugation of benzoate with glycine, Conjugation of carboxylic acids, Conjugation of phenylacetate with glutamine, Coregulation of Androgen receptor activity, DroToll-like, FOXA1 transcription factor network, IGF signaling, Metabolism of proteins, Pathways in cancer - Homo sapiens (human), Phenylacetate Metabolism, Prostate Cancer, Prostate cancer - Homo sapiens (human), Regulation of Androgen receptor activity, Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs), Renin-angiotensin system - Homo sapiens (human), RHO GTPase Effectors, RHO GTPases activate PKNs, Secretion of Hydrochloric Acid in Parietal Cells, Signaling by Rho GTPases, spermine and spermidine degradation I

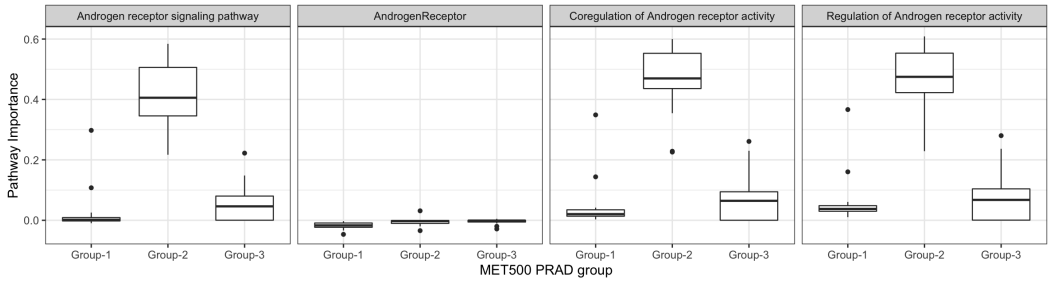

Figure 4: Pathway importance for various Androgen Receptor associated pathways for the MET500 Prostate Adenocarcinoma samples, separated by observed cluster groups.

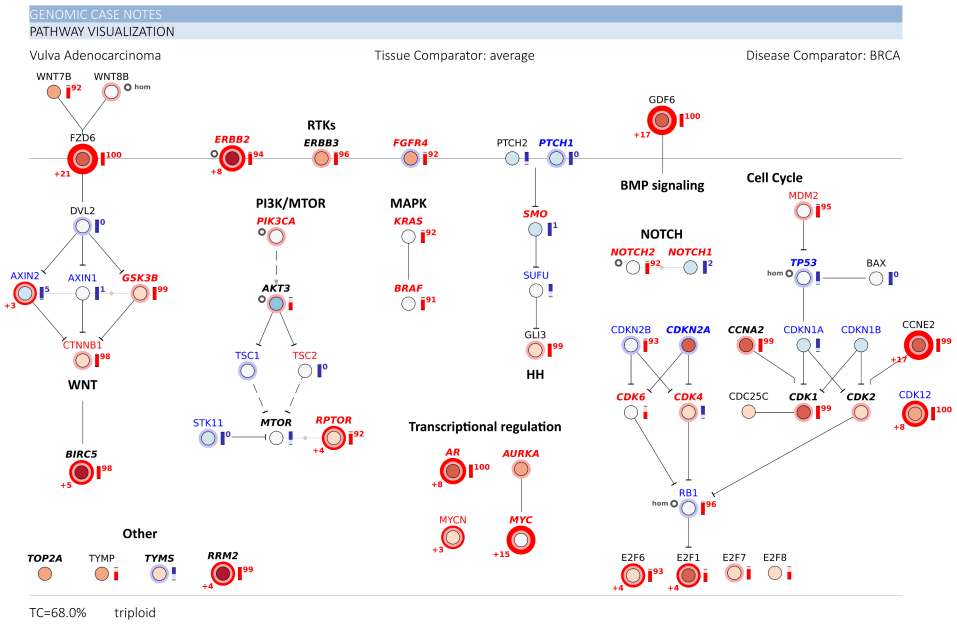

Figure 5: Manual integrative analysis of a mammary-like vulvar adenocarcinoma. Colour of circles shows fold expression change of the respective gene in the sample, relative to a background of all healthy normal tissues from GTEx. Box adjacent to circle indicates percentile expression compared to the Cancer Genome Atlas' cohort of breast cancers. Over-expression is shown in red, and loss of expression in blue. The key oncogenic pathways impacted in this case are shown with grey boxes and red border. Manual analysis identified activation of ERBB2/ERBB3, mTOR pathway, and the MAPK pathway. Overexpression of various genes participating in transcriptional regulation and metabolism was also identified (shown in red borders).

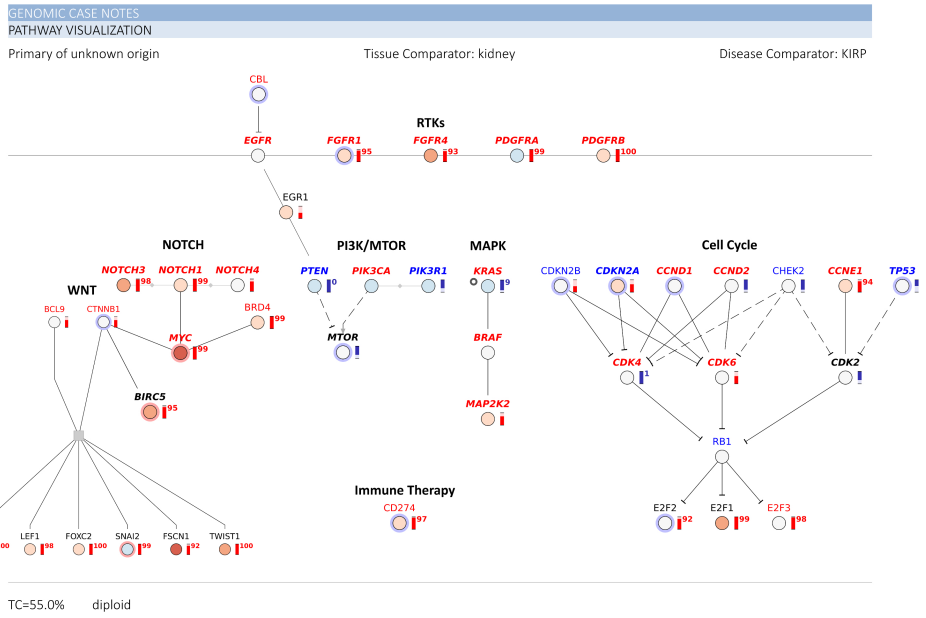

Figure 6: Manual integrative analysis of a cancer with unknown primary, which was diagnosed as a thyroid-like follicular renal cell carcinoma molecularly similar to renal papillary carcinoma. Colour of circles shows fold expression change of the respective gene in the sample, relative to a background of healthy renal tissues. Box adjacent to circle indicates percentile expression compared to the Cancer Genome Atlas' cohort of renal papillary carcinomas. Over-expression is shown in red, and loss of expression in blue. The key oncogenic pathways impacted in this case are shown with grey boxes and red border.