Impact of Molecular Subtyping in Muscle Invasive Bladder Cancer by mRNA Expression Clustering on Predicting Survival and Response of Treatment

Tanan Bejrananda

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Health Sciences
Prince of Songkla University
2022

Impact of Molecular Subtyping in Muscle Invasive Bladder Cancer by mRNA Expression Clustering on Predicting Survival and Response of Treatment

Tanan Bejrananda

A Thesis Submitted in Fulfillment of the Requirements for the

Degree of Doctor of Philosophy in Health Sciences

Prince of Songkla University

2022

Thesis Title	Impact of Molecular Subtyping in Muscle Invasive Bladder Cancer by mRNA expression Clustering on Predicting Survival and Response of Treatment
Author	Mr.Tanan Bejrananda
Major Program	Health Sciences

Major Advisor

....
(Prof. Dr. Surasak Sangkhathat)

Co-advisor (If any)

....
(Assoc.Prof.Dr. Paramee Thongsuksai)

Co-advisor (If any)

....
(Asst. Prof. Dr. Kanet Kanjanapradit)

Examining Committee :

....Chairperson
(Prof. Wachira Kochakarn)

....Committee
(Prof. Dr.Surasak Sangkhathat)

....Committee
(Assoc.Prof.Dr. Paramee Thongsuksai)

....Committee
(Dr.Pasuree Sangsupawanich)

The Graduate School, Prince of Songkla University, has approved this thesis as fulfillment of the requirements for the Doctor of Philosophy Degree in Health Sciences

....
(Prof. Dr. Damrongsak Faroongsarng)
Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate’s own investigations. Due acknowledgement has been made of any assistance received.

....Signature
(Prof.Dr.Surasak Sangkhathat)
Major Advisor

....Signature
(Assoc. Prof. Paramee Thongsuksai)
Co-advisor

....Signature
(Ass.Prof.Kanet Kanjanapradit)
Co-advisor

....Signature
(Mr.Tanan Bejrananda)
Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

....Signature
(Mr.Tanan Bejrananda)
Candidate

ชื่อวิทยานิพนธ์ ผลจากการแบ่งกลุ่มย่อยระดับโมเลกุลในมะเร็งกระเพาะปัสสาวะที่รุกเข้าชั้นกล้ามเนื้อ โดยการจัดกลุ่มระดับการแสดงออกของยีนต่อการทำนายการรอดชีพและการตอบสนองต่อการรักษา

ผู้เขียน นายธนัญญ์ เพชรานนท์

สาขาวิชา วิทยาศาสตร์สุขภาพ

ปีการศึกษา 2564

บทคัดย่อ

บทนำ:

มะเร็งกระเพาะปัสสาวะชนิดที่รุกเข้ากล้ามเนื้อ (muscle invasive bladder cancer) เป็นมะเร็งกระเพาะปัสสาวะชนิดที่มีความรุนแรง ยากต่อการรักษา และอัตราการรอดชีพต่ำ ในปัจจุบัน การศึกษาในระดับโมเลกุลเพื่อหารูปแบบของมะเร็งที่สัมพันธ์กับอัตรารอดชีพและการตอบสนองต่อยาจึงมีความสำคัญต่อการรักษา ในงานวิจัยชิ้นนี้จึงสนใจศึกษาการแบ่งกลุ่มย่อยของมะเร็งกระเพาะปัสสาวะชนิดที่รุกเข้ากล้ามเนื้อเพื่อหารูปแบบความสัมพันธ์ที่อาจช่วยในการรักษาที่มากยิ่งขึ้น การศึกษาถึงการแสดงออกของ mRNA ในรูปแบบ transcriptome จะให้ผลการศึกษาที่มีความแม่นยำสูง ทางทีมวิจัยจึงได้ทำการศึกษารูปแบบการแสดงออกในรูปแบบ transcriptome ควบคู่กับการแสดงออกระดับโปรตีนของตัวบ่งชี้ทางชีวภาพ 4 ชนิดที่มีการรายงานว่ามีความสัมพันธ์กับการแบ่งชนิดของเซลล์มะเร็งกระเพาะปัสสาวะชนิดรุกเข้าชั้นกล้ามเนื้อในรูปแบบ transcriptome analysis ได้แก่ GATA3, CK20, CK5/6 และ CK14

วิธีการศึกษา:

รูปแบบการศึกษารายย้อนหลังโดยเก็บรวบรวมข้อมูลผู้ป่วยมะเร็งกระเพาะปัสสาวะที่รุกเข้าชั้นกล้ามเนื้อ เก็บรวบรวมมูลปัจจัยทางคลินิกจากเวชระเบียนและปัจจัยทางคลินิก ทำการนำชิ้นเนื้อมะเร็งมาทำการย้อมการแสดงออกระดับโปรตีน 4 ตัว GATA3, CK20, CK5/6 และ CK14 ด้วยการทำแบบ tissue microarray และมีการส่งตรวจชิ้นเนื้อมะเร็งกระเพาะปัสสาวะส่งตรวจ mRNA ด้วยการเก็บชิ้นเนื้อชนิด Fresh frozen tissue และส่งตรวจหาการแสดงออกของยีนระดับ RNA ทั้งแยกแบ่งกลุ่มย่อยระดับโมเลกุลในมะเร็งกระเพาะปัสสาวะที่รุกเข้าชั้นกล้ามเนื้อ (molecular subtypings of MIBC) โดยการจัดกลุ่มระดับการแสดงออกของยีนต่อการทำนายการรอดชีพ และการตอบสนองต่อการรักษา

ผลการศึกษา:

โดยผลการศึกษาพบว่า จากจำนวนชิ้นเนื้อที่ใช้ย้อม IHC จำนวน 132 ตัวอย่าง อายุเฉลี่ยของผู้ป่วยคือ 65.6 ปี อัตราการแสดงออกของ IHC ที่เป็นบวกของ GATA3, CK20, CK5/6 และ CK14 คือ 80.3%, 50.8%, 42.4% และ 28.0% ตามลำดับ มีเพียง GATA3 และ CK5/6 เท่านั้นที่มีความสัมพันธ์อย่างมีนัยสำคัญกับผลลัพธ์การรอดชีวิต (ค่า log-rank p-values = 0.004 และ 0.02) จากนั้น GATA3 และ CK5/6 ถูกใช้เพื่อสร้างชนิดย่อย ซึ่งได้แก่ กลุ่ม luminal (GATA+ และ CK5/6−, 38.6%) กลุ่ม basal (GATA− และ CK5/6+, 12.9%) กลุ่มผสม (GATA+ และ CK5/6+, 37.9%) และกลุ่ม double negative (GATA− และ CK5/6−, 10.6%) ผู้ป่วยที่เป็นชนิดย่อยแบบผสมมีอัตราการรอดชีวิตที่ 5 ปีที่ดีขึ้นอย่างมีนัยสำคัญที่ 42.8% ในขณะที่ผู้ป่วยที่เป็นชนิดย่อยแบบ double-negative มีการพยากรณ์โรคที่แย่ที่สุดในสี่กลุ่ม มีอัตราการรอดชีวิตที่ 5 ปี 7.14%

ผลการศึกษาเบื้องต้นของชิ้นเนื้อมะเร็งกระเพาะปัสสาวะที่มีการรุกเข้าชั้นกล้ามเนื้อจำนวน 30 ตัวอย่าง การแบ่งกลุ่มย่อย mRNA เป็น unsupervised clustering ออกเป็น 3 กลุ่มคือ cluster 1 ถึง 3 ซึ่งแต่ละกลุ่มมีการแสดงออกของยีนที่มีความแตกต่างกัน และทางผู้วิจัยได้ใช้ข้อมูลผู้ป่วยจากฐานข้อมูล TCGA เข้าไปเพิ่มเติม ซึ่งพบว่าการจัดกลุ่ม transcriptome ในรูปแบบนี้มีความสัมพันธ์กับอัตราการรอดชีวิตในผู้ป่วยมะเร็งกระเพาะปัสสาวะแบบรุกเข้ากล้ามเนื้อ

สรุปผลการศึกษา:

การแบ่งกลุ่มย่อยโดยใช้ GATA3 และ CK5/6 ใช้ได้กับ MIBCs และผู้ป่วยที่มี subtype แบบ double-negative มีความเสี่ยงสูงสุด ส่วนการแบ่งกลุ่มย่อยของ mRNA จากการวิเคราะห์โดยใช้ข้อมูลจากตัวอย่างที่ศึกษาแบ่งออกกลุ่มใหม่ได้เป็น 3 กลุ่มย่อมที่มีนัยสำคัญและประยุกต์ใช้กับข้อมูลผู้ป่วยจากฐานข้อมูลอื่น สามารถพบรูปแบบความสัมพันธ์กับอัตราการรอดชีพได้อย่างมีนัยสำคัญทางสถิติ

Thesis Title	Impact of Molecular Subtyping in Muscle Invasive Bladder Cancer by mRNA expression clustering on Predicting Survival and Response of Treatment
Author	Mr.Tanan Bejrananda
Major Program	Health Sciences
Academic year	2021

ABSTRACT

Abstract

Purpose

Recently discovered molecular classifications for urothelial bladder cancer appeared to be promising prognostic and predictive biomarkers. It is a major challenge for clinical work to study the molecular subtypes of BC. Outcome of bladder cancer (BC) treatment still need establish and explored the molecular subtypes of bladder cancer and potential clusters. The present study was conducted to evaluate the prognostic impact of molecular subtypes assessed by mRNA expression in a consecutively collected, mono-institutional muscle-invasive bladder cancer (MIBC) cohort, performed by unsupervised clustering and validate subtypes of our institutional cohort with data from The Cancer Genome Atlas (TCGA) and possible to correlate the mRNA expression with tumor molecular subtype membership. Our overall goal was to determine whether mRNA expression have shown significant difference in specific molecular subtypes and correlation with clinical outcomes. Molecular subtyping of muscle-invasive bladder cancer (MIBC) predicts disease progression and treatment response. However, present subtyping techniques are based primarily on transcriptomic analysis, which is relatively expensive. Subtype classification of protein levels by immunohistochemistry (IHC) are more affordable and feasible to perform in a general pathology laboratory. Recent data demonstrated that GATA3, CK20, CK5/6, and CK14 protein levels were correlated with MIBC molecular subtypes. We aimed to evaluate the correlation of those IHC markers with survival outcomes after radical cystectomy in Thai patients. Moreover, we aim to evaluate molecular subtypings by mRNA expression analysis.

Method

30 MIBC were pathologically re-evaluated and molecular subtypes were assessed on mRNA. Fresh-frozen primary tumor samples from a single cohort in Songklanagarind hospital who underwent radical cystectomy between 2015 and 2020. First, we screened the expression profiles of differentially genes expression and of BC by comparing DEG and principle component analysis with K-mean clustering. Moreover, external validation set from the Cancer Genome Atlas (TCGA) database was done by using significant gene expression. We used the complete TCGA dataset with our subtype gene expression and assign TCGA's bladder cancers to molecular subtypes. Taken together, we explored the molecular subtypes and their outcome treatment of BC. Institutional cohort (n= 30 MIBC) and The Cancer Genome Atlas (TCGA)-dataset (n=231 MIBC) were subtyped using unsupervised genes and analyzed for predicting of survival, cancer-specific survival (CSS), overall-survival (OS), and recurrence–free survival (RFS). Moreover, we evaluated the IHC-based subtypes in MIBC, as classified by GATA3, CK20, CK5/6, and CK14 expression in 132 MIBC patients who underwent radical cystectomy followed by adjuvant chemotherapy (2008–2016). All individual markers and clinicopathological parameters were analyzed against treatment outcomes after radical cystectomy and some selected tissues were sent for whole transcriptome sequencing and clustering from mRNA expression.

Result

Unsupervised consensus hierarchical clustering applied to gene expression data and identified 3 molecular subtypes. These subtypes were associated with distinct clinicopathological characteristics and molecular expression. The clustering was validated in the TCGA dataset. We identified different clinical characteristics and identified 3 molecular subtypes MIBC specimens from cohort dataset successfully. In multivariable analyses, N-stage, T-stage, M-stage and/or age predicted CSS/OS and/or cisplatin- based adjuvant-chemotherapy response. In the TCGA-dataset, publications report that subtypes risk-stratify patients for OS. For IHC study section, the result showed that the mean patient age was 65.6 years, and the male to female ratio was 6.8:1. Positive IHC expression rates of GATA3, CK20, CK5/6, and CK14 were 80.3%, 50.8%, 42.4%, and 28.0%, respectively. The 5-year overall survival (OS) was 27.0% (95% confidence interval (CI) 19.6%–35.0%). Only GATA3 and CK5/6 were significantly associated with survival outcome (log-rank p-values = 0.004 and 0.02). GATA3 and CK5/6 were then used to establish subtypes, which were luminal (GATA+

and CK5/6−, 38.6%), basal (GATA− and CK5/6+, 12.9%), mixed (GATA+ and CK5/6+, 37.9%), and double-negative (GATA− and CK5/6−, 10.6%). Patients with the mixed subtype had a significantly better 5-year OS at 42.8%, whereas patients with the double-negative subtype had the worst prognosis among the four groups (5-year OS 7.14%). In the multivariable analysis, lymph node status and subtype independently predicted survival probability. The double-negative subtype had a hazard ratio of 3.29 (95% CI 1.71–6.32).

Conclusion

The results further reinforce the conclusion that the molecular subtypes of bladder cancer are distinct disease entities with specific molecular subtype. In our cohorts/subtyping-classifications, clinical and novel molecular subtypes for predicting outcome of treatment. For immunohistochemistry subtyping using GATA3 and CK5/6 was applicable in MIBCs, and patients with the double-negative subtype were at the highest risk and may require more intensive therapy and mRNA subtyping by mRNA expression must showed the significant relationship with survival rate.

ACKNOWLEDGEMENT

The successful completion of this thesis would not have been possible without the help of many people in so many ways. First, I would like to express my deepest gratitude to Prof. Dr. Surasak Sangkhathat, Dr. Kanet Kanjanapradit, Assoc. Prof. Paramee Thongsuksai, and Dr. Pasuree Sangsupawanich. Without their guidance, constructive criticism this research would not have been possible.

Appreciation is also expressed to staffs and members in Division of Urology, Department of Surgery, Faculty of Medicine, Prince of Songkla University for their practical support.

I would also like to thank Dr.Jirakrit Saetung, Dr.Wison Laochareonsuk and all the laboratory staffs for helping me with laboratory work. Without their assistance, my work won’t have been complete.

Finally, I would like to express my indebtedness to my wife and my son who have given me constant support, considerable encouragements, and unlimited love during the completion of the thesis.

Tanan Bejrananda

CONTENTS

	Page
Contents	xi
List of Tables	xiv
List of Figures	xv
List of Abbreviations and Symbols	xix
Chapter	
Chapter 1: Introduction	1
1.1 Background and Rationale	1
1.2 Research Questions	4
1.3 Hypotheses	4
1.4 Objectives	5
1.5 Literature review	5
1.5.1 Gene expression profiling of bladder cancer	5
1.5.2 Molecular classification of MIBC	6
1.5.3 Transcriptional regulation of bladder cancer subtypes	9
1.5.4 Prognostic and predictive associations of MIBC subtypes	10
1.5.5 Immunohistochemistry-based profiling	12
Chapter 2: Research methodology	16
2.1 Methodology: Part I	16
2.1.1 Study design and targeted population	16
2.1.2 Sample selection	16
2.1.3 RNA isolation	17
2.1.4 Data acquisition	17
2.1.5 Gene expression analysis	18

2.1.6 Principle component analysis (PCA) 22

2.1.7 External validate TCGA cohort 23

2.1.8 Statistical analysis 23

2.2 Methodology: Part II 25

2.2.1 Cohort composition 25

2.2.2 Study design and targeted population 25

2.2.3 IHC study by tissue microarray 26

2.2.4 Statistical analysis 27

2.2.5 Sample size calculation 28

Chapter 3: Results 29

3.1 Results: Part I 29

3.1.1 Clinical characteristics 29

3.1.2 Transcriptome profiling and classification of Thai MIBC 30

3.1.3 Differential gene expression (DEG) analysis revealed 37 genes expressed with the different levels among clusters 35

3.1.4 ROC analysis of 37 differentially expressed genes found in MIBC tissues 37

3.1.5 Clinical characteristic of MIBC in our cohort associated with treatment Outcome 40

3.1.6 Molecular subtypes of MIBC are associated with response of perioperative chemotherapy 41

3.1.7 The transcriptomic classification using PCA analysis of tissue sample 43

3.1.8 Identification of differentially expressed genes of each cluster of MIBC 46

3.1.9 The certain signaling pathways were associated with each type of cluster 47

3.2 Results: Part II 52

3.2.1 Demographic and clinicopathological data 52

3.2.2 Association between molecular subtypes and survival outcomes 59

3.3 Results: Part III 61

3.3.1 Integrative result of molecular subtyping from mRNA expressions using 4 markers from immunohistochemistry in our MIBC cohort 61

3.3.2 Validate each signature gene expression in TCGA data 63

3.3.3 AUC score of selected signature genes of molecular subtypes 65

3.3.4 Optimal cutpoint, accuracy, sensitivity, specificity and AUC of 37 difference gene in each cluster 68

Chapter 4: Discussion 72

Chapter 5: Conclusion 78

Bibliography 80

Appendices 88

Appendix A: The ethics committee approval 89

Appendix B: Publication 91

Appendix C: Shell command; Alignment and transcript count 102

Appendix D: R code for analysis 105

Vitae 132

LIST OF TABLES

Table **Page**

1. Clinical data summary of studied MIBC dataset 29

2. The area under the curve (AUC) from specificity and sensitivity of the genes expressed differently for each MIBC cluster 38

3. Univariate and multivariate logistic regression analyses of clinical data of MIBC Patients 40

4. Unique significant differential expression 47

5. The demographic characteristics of the patients and immunoreactivity each marker 52

6. Immunopositivity of the 4 markers studied and their correlation with clinicopathological parameters 53

7. Patient’s characteristic classified by IHC subtypes categorized by GATA3 and CK5/6 56

8. Univariable and multivariable regression analysis of clinical outcomes 58

9. AUC score of selected signature genes classified clusters 65

10. Sensitivity, specificity and cut point of selected signature genes between cluster B-C 66

11. Optimal cutpoint, accuracy, sensitivity, specificity and AUC of 37 difference gene in each cluster 68

LIST OF FIGURES

Figure **Page**

1. Evolving schemes of molecular classification of urothelial carcinoma of the bladder 3

2. Comparison of the molecular subtyping schemes of urothelial carcinoma 8

3. Immunohistochemical staining of MIBC tissue: GATA3, CK20, CK5/6, and CK14 14

4. Sample of bladder tumor preparation for mRNA sequencing 17

5. Transcriptomic pipeline after mRNA sequencing 19

6. library preparation 20

7. Massive parallel sequencing 20

8. Alignment and transcript count 21

9. Boxplot with normalized counts 30

10. MA plot analysis demonstrated of genes were found to be up-regulated and downregulated in MIBC 31

11. Volcano plot analysis demonstrated more than a hundred of genes in MIBC 32

12. Elbow plot analysis was applied for Thai MIBC transcriptome data 33

13. The transcriptomic data of all MIBC tissue samples by K-mean clustering 33

14. Heatmap analysis of gene expression of MIBC patients 34

15. Boxplot of examples gene expression in 3 clusters 34

16. Venn diagram; The data from differential gene expression (DEG) analysis also revealed the number of genes that expressed differently between two clusters 35

17. The expression levels of each gene were demonstrated as Volcano plots for cluster A 36

LIST OF FIGURES

Figure **Page**

18. The expression levels of each gene were demonstrated as Volcano plots for cluster B 36

19. The expression levels of each gene were demonstrated as Volcano plots for cluster C 37

20. ROC curve analysis was performed for all 37 genes and all cluster 39

21. The Kaplan-Meier analysis demonstrated that cluster B displayed the highest probability of survival within 48 months of follow-up while cluster C showed the lowest value 41

22. RNA-based molecular subtypes are associated with pathological response to neoadjuvant chemotherapy 42

23. RNA-based molecular subtypes are associated with pathological response to adjuvant chemotherapy 42

24. The centroid derived from MIBC tissue samples to separate PCA coordinates of TCGA cohort into three clusters according to MIBC tissues 43

25. MA plot analysis demonstrated more than a hundred of genes were found to be up-regulated and downregulated in MIBC comparing between non-TCGA and TCGA group 44

26. The expression levels of each gene were demonstrated as Volcano plots comparing between non-TCGA and TCGA group 44

27. The Kaplan-Meier analysis demonstrated that cluster B displayed the highest probability of survival within 1,800 days of follow-up while cluster C showed the lowest value 45

28. Cumulative MIBC cases by K-mean clustering that classified by mRNA with TCGA classification 45

LIST OF FIGURES

Figure **Page**

29. Unique and common enriched genes of each cluster 46

30. Top 10 significantly enriched terms in the KEGG enrichment analysis for cluster A 48

31. Top 10 significantly enriched terms in the KEGG enrichment analysis for cluster B 48

32. Top 10 significantly enriched terms in the KEGG enrichment analysis for cluster C 49

33. GO Enrichment analyses of the differentially expressed genes in cluster A 50

34. GO Enrichment analyses of the differentially expressed genes in cluster B 50

35. GO Enrichment analyses of the differentially expressed genes in cluster C 51

36. Number of differential gene expressions among different cluster after achieved by pathway enrichment 51

37. Kaplan-Meier curves demonstrating survival probability in 132 MIBC patients according to IHC markers expression; GATA3, CK5/6, CK14 and CK 20 55

38. Correlation heatmap of immunohistochemistry markers expression among GATA3, CK20 (basal-like markers), CK5/6 and CK14 (luminal-like markers) 56

39. Kaplan-Meier survival curves of molecular subtypes of muscle invasive bladder cancer according to their immunohistological subtypes 60

40. Violin plot of GATA3 expression in 3 clusters 61

41. Violin plot of CK14 expression in 3 clusters 62

42. Violin plot of CK5/6 expression in 3 clusters 62

43. Violin plot of CK5/6 expression in 3 clusters 63

44. Violin plot of GATA3 expression in 3 cluster from TCGA data 63

LIST OF FIGURES

Figure **Page**

45. Violin plot of CK 14 expression in 3 cluster from TCGA data 64

46. Violin plot of CK 5/6 expression in 3 cluster from TCGA data 64

47. Violin plot of CK 20 expression in 3 cluster from TCGA data 65

48. Sensitivity, specificity and cut point of selected signature genes 66

49. Cut point and ROC of GATA3 genes between cluster B-C 67

50. Cut point and ROC of KRT14(CK14) genes between cluster B-C 67

51**.** Cut point and ROC of KRT5 (CK5/6) genes between cluster B-C 68

LIST OF ABBREVIATIONS AND SYMBOLS

BCa	=	Bladder cancer
NMIBC	=	non-muscle invasive bladder cancer
MIBC	=	muscle invasive bladder
LG	=	low grade
HG	=	high grade
UC	=	Urothelial carcinoma
LVI	=	lymphovascular invasion
OS	=	overall survival
TURBT	=	transurethral resection of bladder tumor
LUND	=	Lund University
UROMOL	=	UROMOL study
TCGA	=	The Cancer Genome Atlas
UNC	=	The University of North Carolina
MDACC	=	The MD Anderson Cancer Center
NAC	=	neoadjuvant chemotherapy
ICI	=	immune checkpoint inhibitors
IHC	=	immunohistochemistry
GATA3	=	GATA binding protein 3
CK20	=	cytokeratin 20
CK5/6	=	cytokeratin 5/6
CK14	=	cytokeratin 14
DNA	=	deoxyribonucleic acid
RNA	=	ribonucleic acid

LIST OF ABBREVIATIONS AND SYMBOLS (continued)

mRNA	=	messenger ribonucleic acid
ERBB2	=	Erb-B2 Receptor Tyrosine Kinase 2
ESR1	=	Estrogen Receptor 1
URO	=	Urothelial-like
GU	=	Genomically Unstable
Ing.	=	Infiltrated
Mes	=	Mesenchymal-like
Basal/SCCL	=	Basal/Squamous Cell Carcinoma-like
NE-like	=	Small-cell/Neuroendocrine-like
LumP	=	luminal papillary
LumU	=	luminal unstable
LumNS	=	luminal non-specified
FGFR3	=	Fibroblast Growth Factor Receptor 3
CCND1	=	Cyclin D1
E2F3	=	the E2F Transcription Factor 3
RB1	=	Retinoblastoma
CDKN2A	=	cyclin- dependent kinase inhibitor
KRT5	=	keratin 5
KRT6	=	keratin 6
KRT14	=	keratin 14
CD44	=	CD44 Molecule
CDH3	=	Cadherin 3
EGFR	=	Epidermal Growth Factor Receptor

LIST OF ABBREVIATIONS AND SYMBOLS (continued)

PPARG	=	Peroxisome Proliferator Activated Receptor Gamma
FOXA1	=	Forkhead Box A1
RA	=	retinoic acid
PPAR	=	proliferator-activated receptor
RXR	=	retinoid X receptor
FABP4	=	Fatty Acid Biding Protein 4
FABP5	=	Fatty Acid Biding Protein 5
RAR	=	retinoic acid receptors
STAT3	=	Signal Trasnducer And Activator Of Transription 3
ΔNp63	=	dominant-negative inhibitor of TAp63
SHH	=	Sonic Hedgehog Signaling Molecule
IHH	=	Indian Hedgehog Signaling Molecule
DHH	=	Desert Hedgehog Signaling Molecule
FGFs	=	fibroblast growth factors
BMPs	=	bone morphogenetic proteins
WNT	=	Wnt Familt Member
GLIs	=	GLI Family Zinc Finger
HOX	=	homeobox
anti-PD-L1	=	programmed death-ligand 1
QC	=	quality control
ng	=	nanogram
µl	=	microlitre

LIST OF ABBREVIATIONS AND SYMBOLS (continued)

RNA-Seq	=	rRNA sequencing
DEG	=	differential gene expression
log2FC	=	log2 fold change
T stage	=	Tumor stage
N stage	=	Nodal stage
M stage	=	Metastasis stage
PI3K	=	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase
Akt	=	AKT Serine/Threonine Kinase
FN1	=	Fibronectin 1
COL6A2	=	Collagen Type VI Alpha 2 Chain
COL1A2	=	Collagen Type I Alpha 2 Chain
MAPK	=	Mirogen-Activated Protein Kinase
TGFB1	=	Transforming Growth Factor Beta 1
MECOM	=	MDS1 And EVI1 Complex Locus
MT1A	=	Metallothionein 1A
MT2A	=	Metallothionein 2A
TUBB6	=	Tubulin Beta 6 Class V
TUBB3	=	Tubulin Beta 3 Class III
LGALS1	=	Galectin 1
IFITM3	=	Interferon Induced Transmembrane Protein 3
CDA	=	Cytidine Deaminase
MAOA	=	Monoamine Oxidase A
MGST1	=	Microsomal Glutathione S-Transferase 1
KEGG	=	Kyoto Encyclopedia of Genes and Genomes

LIST OF ABBREVIATIONS AND SYMBOLS (continued)

JAK-STAT	=	Janus Kinase- Signal Transducer And Activator Of Transcription
BDKRB1	=	Bradykinin Receptor B1
EDNRA	=	Endothelin Receptor Type A
AVPR1A	=	Arginine Vasopressin Receptor 1A
PTGER3	=	Prostaglandin E Receptor 3
PTGFR	=	Prostaglandin F Receptor
NTRK3	=	Neurotrophic Receptor Tyrosine Kinase 3
P2RX1	=	Purinergic Receptor P2X 1
ITGA8	=	Integrin Subunit Alpha 8
CREB5	=	CAMP Responsive Element Binding Protein 5
COL6A3	=	Collagen Type VI Alpha 3 Chain
FGF7	=	Fibroblast Growth Factor 7
NGF	=	Nerve Growth Factor
HGF	=	Hepatocyte Growth Factor
ANGPT1	=	Angiopoietin 1
KCNMB1	=	Potassium Calcium-Activated Channel Subfamily M Regulatory Beta Subunit 1
KCNMA1	=	Potassium Calcium-Activated Channel Subfamily M Alpha 1
ADRA2A	=	Adrenoceptor Alpha 2A
ATP1B2	=	ATPase Na+/K+ Transporting Subunit Beta 2
ADORA1	=	Adenosine A1 Receptor
PRKG1	=	Protein Kinase CGMP-Dependent 1
BDKRB1	=	Bradykinin Receptor B1

EDNRA	=	Endothelin Receptor Type A
AVPR1A	=	Arginine Vasopressin Receptor 1A
PDGFRB	=	Platelet Derived Growth Factor Receptor Beta
TNC	=	Tenascin C
PDGFRB	=	Platelet Derived Growth Factor Receptor Beta
PRKG1	=	Protein Kinase CGMP-Dependent 1
AUC	=	area under the curve
HR	=	hazard ratio
SD	=	standard deviation
95% CI	=	95% confidence interval
θ	=	hazard ratio
p	=	proportion allocated to a group
MGST1	=	Microsomal Glutathione S-Transferase 1
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
JAK-STAT	=	Janus Kinase- Signal Transducer And Activator Of Transcription
BDKRB1	=	Bradykinin Receptor B1
EDNRA	=	Endothelin Receptor Type A
AVPR1A	=	Arginine Vasopressin Receptor 1A
PTGER3	=	Prostaglandin E Receptor 3
PTGFR	=	Prostaglandin F Receptor
NTRK3	=	Neurotrophic Receptor Tyrosine Kinase 3
P2RX1	=	Purinergic Receptor P2X 1
ITGA8	=	Integrin Subunit Alpha 8
CREB5	=	CAMP Responsive Element Binding Protein 5
COL6A3	=	Collagen Type VI Alpha 3 Chain

LIST OF ABBREVIATIONS AND SYMBOLS (continued)

FGF7	=	Fibroblast Growth Factor 7
NGF	=	Nerve Growth Factor
HGF	=	Hepatocyte Growth Factor
ANGPT1	=	Angiopoietin 1
KCNMB1	=	Potassium Calcium-Activated Channel Subfamily M Regulatory Beta Subunit 1
KCNMA1	=	Potassium Calcium-Activated Channel Subfamily M Alpha 1
ADRA2A	=	Adrenoceptor Alpha 2A
ATP1B2	=	ATPase Na+/K+ Transporting Subunit Beta 2
ADORA1	=	Adenosine A1 Receptor
PRKG1	=	Protein Kinase CGMP-Dependent 1
BDKRB1	=	Bradykinin Receptor B1
EDNRA	=	Endothelin Receptor Type A
AVPR1A	=	Arginine Vasopressin Receptor 1A
PDGFRB	=	Platelet Derived Growth Factor Receptor Beta
TNC	=	Tenascin C
PDGFRB	=	Platelet Derived Growth Factor Receptor Beta
PRKG1	=	Protein Kinase CGMP-Dependent 1
AUC	=	area under the curve
HR	=	hazard ratio
SD	=	standard deviation
95% CI	=	95% confidence interval
θ	=	hazard ratio
p	=	proportion allocated to a group

CHAPTER 1

Introduction

1.1 Background and Rationale

The new cases of bladder cancer (BCa) increase by more than 500,000 per year and the deaths caused by BCa increase by approximately 200,000 per year (1,2). Traditionally, based on the degree of invasion in the bladder muscle wall, BC can be classified into either non-muscle invasive (NMIBC) or muscle invasive (MIBC) and also be divided pathologically into high-grade (HG) or low-grade (LG) tumors. HG tumors are poorly differentiated and LG tumors are usually well-differentiated, Urothelial carcinoma (UC) is a cancer that arises from the epithelial lining of the bladder wall. Basal stem cells at the stromal interface self-renew and make intermediate and superficial/umbrella cells to maintain and regenerate functioning urothelium, according to well-established differentiation research. Although the majority of urothelial cancers do not reach the submucosal stroma (lamina propria) or bladder wall (muscularis propria), those that do can have a wide range of histologic features, clinical outcomes, and molecular profiles (8).

The treatment outcomes are diverse for different BCa patients, especially muscle-invasive BCa (MIBC) (3). Result of treatment in our published radical cystectomy cohort included 111 MIBC patients reported the 5-year cancer-specific survival rate was only 36%. Of several factors examined, univariate analysis identified tumor stage, nodal status, metastasis, margin positive and lymphovascular invasion (LVI) as significant predictors of OS, of which tumor stage and nodal status appeared to be independently related to overall survival on multivariate analysis (4). However, personalized treatment of each patient need development for improve survival of patient. Molecular subtyping has been purposed for classified and specified treatment option in each subtype.

Predicting response to available peri-operative treatment and developing novel methods of targeting invasive bladder cancers are two areas of intense research where molecular profiling is thought to be useful. The availability of diagnostic trans-urethral resection of bladder tumor (TURBT) specimens before radical surgery allows molecular profiling to potentially assist patient decisions about surgery and neoadjuvant treatment (8).

There were previously five major subtyping classification systems: LUND, UROMOL, The Cancer Genome Atlas (TCGA), The University of North Carolina (UNC), and the MD Anderson Cancer Center (MDACC). These five subtyping classification systems not only evolved independently, but each taxonomy's nomenclature differs from the others (5-10). Until now, there is no consistent risk stratification for BCa. Before having radical cystectomy, patients with MIBC are usually treated with neoadjuvant chemotherapy (NAC) (11). Immune checkpoint inhibitors (ICI) or other novel drugs guided by biomarkers, such as targeted treatments, are commonly used to treat locally progressed and metastatic illness (12-21).

The genetic categorization of urothelial carcinoma tumors may provide important information for stratifying prognostically significant groups or determining the best treatment for a specific patient (15,22). The development of a molecular taxonomy for bladder cancer is possibly the most exciting and major therapeutic discovery in decades. MIBC was also separated into two primary molecular subgroups, luminal and basal, utilizing advanced approaches comparable to those used in breast cancer research (5,7,9,23-27). According to reports, there are considerable differences in prognosis and responsiveness to current therapy between these two broad categories, with the basal subtype being more aggressive than the luminal subtype (5,9,26,28,29). Many molecularly defined groups are now accessible due to differences in approach and interpretation of previous findings (7,8,10,15,22,30-35). Figure 1 shows how bladder cancer molecular classifications have changed over time.

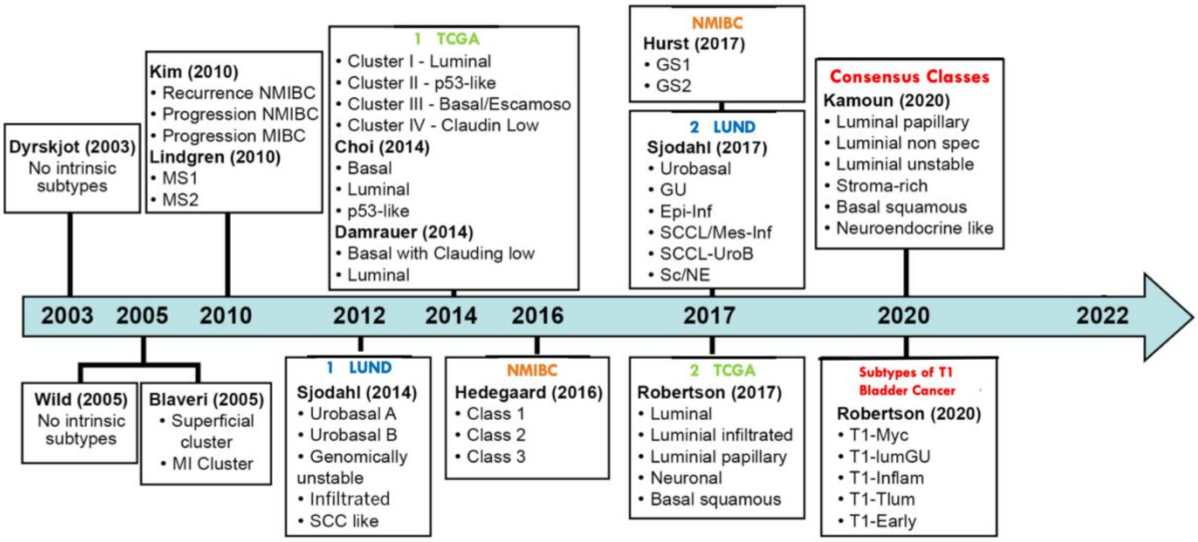


Figure 1. Evolving schemes of molecular classification of urothelial carcinoma of the bladder.

Despite recent agreement on several molecular subtypes, molecular categorization remains a complicated, expensive, and infrequently available technology (10,36–39). However, the recent introduction of gene expression analysis offers an alternative method for molecular subtyping, with the potential benefit of decreased analysis costs and the production of accurate gene classifiers with clinical relevance (33,40). Thus, a study examining novel subtypes in the context of urothelial carcinoma's molecular taxonomy using mRNA and immunohistochemistry (IHC) would be relevant and beneficial in terms of developing less expensive and repeatable tools for investigating the molecular classification of urothelial bladder carcinoma. As a result, molecular subtypes provide a context that connects tumor biology to the ability to affect and stratify patients in order to improve oncological outcomes. The goal of this study was to see how far molecular profiling had progressed in muscle-invasive bladder cancer. We hope to present a description of molecular subtypes, an enumeration of promising targeted therapeutics, and a vision of how molecular subtypes could be incorporated in routine pathology for the healthcare professional who diagnoses or treats bladder cancer. We also want to learn more about fundamental molecular pathology studies and how to interpret molecular subtypes in tissue samples with the cellular diversity of invasive bladder tumors.

1.2 Research Questions

1.2.1 Can we perform molecular subtyping of muscle invasive bladder cancer by mRNA expression and immunohistochemistry?

1.2.2 What is the impact of molecular subtyping in muscle invasive bladder cancer on predicting survival and response of treatment?

1.3 Hypotheses

1.3.1 Molecular subtypes of the muscle-invasive bladder have different expressions in mRNA and protein level.

1.3.2 Molecular subtypes of bladder cancer can predict the outcome and survival of treatment.

Our study aimed to study NGS-genomic/transcriptomic profiling used to generate molecular data in bladder cancer and provide clinically meaningful datasets for the molecular classification of bladder cancer. For IHC, our study generated a four-gene classifier, incorporating GATA3 and CK20 (typically related to luminal molecular subtype) and CK5/6 and CK14 (typically related to basal molecular subtype). This methodology allowed us to explore differences in clinicopathological parameters and potential sensitivities to treatment in urothelial carcinomas of bladder patients. The purpose of this study was to use IHC to identify molecular subgroups in four distinct MIBC cohorts and to investigate their link to prognosis and treatment results. We expected that IHC will uncover significant groups of tumors that mimic known molecular subtypes and have relevant clinical relationships, with patients with Basal or double negative subtypes having worse outcomes than patients with the other subtype. Moreover, we performed a transcriptomic study of MIBC to perform the unsupervised clustering of the novel molecular subtypes and validate with TCGA, different molecular subtypes impact survival.

1.4 Objectives

1.4.1 Primary objective

- ❖ To cluster the novel molecular subtypes by transcriptomic profiling and evaluate clinically significant of data of MIBC
- ❖ To validate the subtypes with TCGA dataset

1.4.2 Secondary objective

- ❖ To use the 4 markers (GATA3, CK5/6, CK14 and CK20) classified into molecular subtypes and evaluate the clinically significant of MIBC patients

1.5 Literature review

1.5.1 Gene expression profiling of bladder cancer

A significant difficulty in molecular oncology is interpreting the cumulative biological effect of the many genetic abnormalities and dysregulated cellular processes observed in each given tumor. The RNA transcriptome serves as a link between the molecular foundations and the cellular phenotype, and as such, global gene expression profiling is one of the most powerful methods available for biological characterization. Early research in bladder cancer demonstrated that low-grade NMIBC and MIBC could be separated by their gene expression patterns. Numerous studies provided mRNA expression profiles with purported clinical prognostic value, such as predicting overall survival, disease-free survival, or progression; nevertheless, these signatures are difficult to test in independent datasets, frequently performing no better than chance. Perou and Srlie pioneered the use of hierarchical clustering to deduce unique molecular subgroups exhibiting distinct gene expression patterns associated with a range of biological activities and pathways, exhibited a correlation with pathogenic factors such as ERBB2 and ESR1 expression and shown survival differences (41).

1.5.2 Molecular classification of MIBC

The molecular characterisation of MIBC has been a major focus of recent bladder cancer research. Internal subtypes with intriguing prognostic and predictive capabilities have been discovered by large-scale mRNA profiling studies (5,6,9). These subgroups are analogous to those observed in breast cancer, for which a basal-luminal molecular categorization scheme identifying five subtypes has been devised (41–43). These subgroups have been defined as key components of breast cancer treatment stratification due to their prognostic and predictive relationships. University of North Carolina (UNC) researchers discovered basal-luminal differentiation of tumors as a critical axis that contributes to the formation of two separate main subtypes of MIBC (44). Following this, genomic and proteomic tumor profiling have resulted in the development of a succession of revised and overlapping subtyping taxonomies for bladder cancer subtyping, with UNC, MD Anderson University (MDA), The Cancer Genome Atlas (TCGA), and Lund University pioneering these efforts. All of these subtyping approaches have in common the identification of luminal-like and non-luminal-like (basal-like) subtypes at the highest hierarchical level, corresponding to differential paths of urothelial differentiation (45). Thus, while the various taxonomies for subtyping were devised independently, they all agree on the identification of basal and luminal subtypes that can be divided from three to seven groupings. Notably, these subtypes have exhibited disparities in clinical outcomes, with basal subtype cancers being more aggressive and having a poor prognosis, whereas patients with luminal subtype tumors had an improved overall survival (5,6,45). Each taxonomy's subcategories also have specific prognostic connections. While each of these classifications gives valuable insight into the genetic variety and clinical behavior of these malignancies, discrepancies in these taxonomies due to methodological variances in subtyping have limited the impact of this work. The varied nomenclatures, definitions, and distinctions in subtyping taxonomies, as well as their clinical importance, have hampered the interpretation of this data. Thus, in 2019, key leaders from each of these groups collaborated to develop a consensus methodology and taxonomy for subtyping, pooling transcriptomic data from 1750 patients to delineate six consensus subtypes: luminal-papillary, luminal-unstable, luminal-unspecified, basal-squamous, and stroma-rich. (10). (Fig. 2)

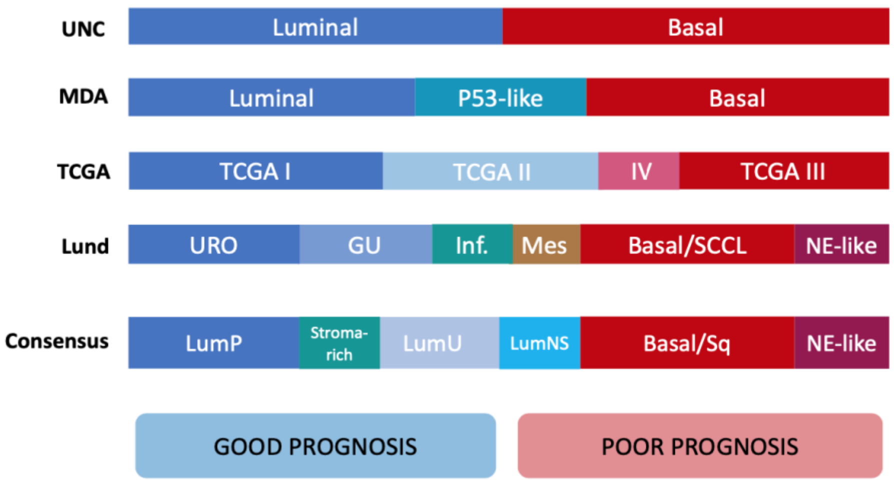


Figure 2. Comparative analysis of urothelial cancer molecular subtyping schemes. UNC = Univesity of North Carolina group; MDA = MD Anderson group, TCGA = The Cancer Genome Atlas; Lund, The Lund Bladder Cancer Research group. URO = Urothilial-like; GU = Genomically Unstable; Ing. = Infiltrated; Mes = Mesenchymal-like; Basal/SCCL = Basal/Squamous Cell Carcinoma-like; NE-like = Small-cell/Neuroendocrine-like; LumP = luminal papillary; LumU = luminal unstable; LumNS = luminal non-specified (10,26).

The substantial effort necessary to classify MIBCs into distinct categories is unsurprising, given their high biological heterogeneity, which may be a result of their high mutation rates - one of the highest of all human malignancies (44). In general, UC develops via one of two mutually exclusive genetic pathways: the Fibroblast Growth Factor Receptor 3 (FGFR3)/Cyclin D1 (CCND1) system or the E2F Transcription Factor 3 (E2F3)/Retinoblastoma (RB1) pathway (47,48). Lower stage and grade tumors have been linked to mutations in the FGFR3/CCND1 pathway. These tumors are defined by hyperactivation of FGFR3, overexpression of CCND1, and deletions of genes on 9p and 9q, including the cyclin-dependent kinase inhibitor (CDKN2A) gene, which produces the p16 protein (47). Tumors with a higher stage or grade have been documented to have mutations in the E2F2/RB1 pathway, as well as increased CDKN2A expression. Whichever of these two routes is disrupted contributes to a tumor's molecular landscape, with changes in either pathway determining molecular subtypes. Across all categorization schemes, widely classified luminal MIBC tumors are enriched for mutations in FGFR3, uroplakins, KRT20, ERBB2, and CCND1, as well as differentiation markers forkhead box A1 (FOXA1) and GATA-binding protein 3 (GATA3) (5,6,9). Whereas basal cancers exhibit basal differentiation-associated cytokeratins such as KRT5, KRT6, and KRT14, as well as CD44 and CDH3. Enrichment of Epidermal Growth Factor Receptor (EGFR) mutations further characterizes basal cancers (10,49).

1.5.3 Transcriptional regulation of bladder cancer subtypes

Divergent differentiation is a well-known characteristic of urothelial malignancies, as seen by the range of subtypes reported. Corruption of the normal urothelial stratification and differentiation regulatory pathways appears to be at the root of various molecular subtypes. PPARG, FOXA1, and GATA3, as well as other key transcription factors involved in the development and differentiation of normal urothelium, have been repeatedly demonstrated to be defining factors in tumor subtypes that retain some degree of normal urothelial differentiation or expression of urothelial markers, whereas their loss is strongly associated with non-urothelial-like subtypes. Similarly, retinoic acid (RA) signaling is a critical component of urothelial development, and dysregulations of this signaling have been observed in bladder cancer proliferator-activated receptor (PPAR) proteins from heterodimers with retinoid X receptor (RXR) proteins, as well as the expression pattern of several genes involved in ligand shuttling to these nuclear hormone receptor dimers (e.g. FABP4, FABP5, and CRABP2) EGFR, in future research, in BaSq-like cancers. Both retinoic acid receptors (RAR) and peroxisome, STAT3, and Np63 appear to be significant drivers of the observed expression patterns that exhibit subtype-specific expression. The interaction of hedgehog proteins (SHH, IHH, and DHH), fibroblast growth factors (FGFs), bone morphogenetic proteins (BMPs), WNTs, GLIs, HOX, and TBF-signaling is well researched in embryonic biology, where gradient expression and feedback loops are crucial for organ formation. When studying components of these pathways, one cannot rely merely on gene expression from a tumor biopsy, since the spatial arrangement of signaling gradients and interaction between stratified urothelium and stroma is critical and must be taken into account. Each of the regulatory components listed in this section has been thoroughly investigated in both normal and cancer environments; nevertheless, a comprehensive understanding of how they each contribute to the molecular biology of bladder cancer is still lacking. It will be critical to appropriately evaluate the dysregulation seen in cancer. Numerous components of the hedgehog indicate that knowledge of the bladder's natural embryonic biology has been included (41).

1.5.4 Prognostic and predictive associations of MIBC subtypes

The connections between molecular subtypes and prognosis and response to therapy suggest that subtyping has a wide range of clinical uses (5,6,13,22,26,50). Several therapeutically meaningful correlations have been postulated by various organizations based on these various taxonomies; however, these need to be validated further. As a result, determining a uniform, high-throughput subtyping process would speed up the identification and confirmation of these applications. In the end, this simplified process would make molecular subtyping for patient stratification more realistic for doctors.

The first of these potential uses is the use of systemic medicines such as NACT, which should ideally be prioritized for patients who are at high risk. Molecular subtypes of the Lund taxonomy, as well as other subtyping methods, have been shown to have strong predictive value. In terms of prognosis, luminal cancers have a better overall survival rate than their more aggressive basal counterparts (5,6,44). One of the most appealing uses of bladder cancer subtyping is the development of a predictive biomarker for treatment response (6,22,25,26,50). A number of studies have found that patients with the Basal/SCCL subtype benefit more from treatment (6,22,25,26,50). Notably, Seiler et al. demonstrated that NACT treatment improves the prognosis of basal malignancies from the worst to the best, but patients with luminal tumours experienced no change in survival (26). In a clinical trial conducted in 2016, McConkey and colleagues found that 90 percent of patients with bladder cancer of the basal subtype received a 5-year survival benefit from NACT, compared to 70 percent of patients with bladder cancer of the luminal subtype (22). Numerous studies have now corroborated these findings in relation to NACT, demonstrating benefits for the basal subtype in terms of survival or pathologic response (10,50). However, a new multi-omics study by Taber and colleagues has found that the basal subtype is linked to poor NAC response as measured by pathologic response, directly contradicting previous research (51). Taber et al. show that immunological infiltration and genomic instability caused by a large number of chromosomal abnormalities and/or BRCA2 mutations are linked to treatment response in this study (51). Although basal/luminal subtyping has been linked to NACT response and survival outcomes, recent contradictory data suggest that more research is needed to fully understand these interactions.

Another potentially intriguing application of these molecular groupings is the occurrence of targetable mutations in specific subtypes, which suggests the possibility of stratifying patients for targeted therapy based on their subtype. Uro malignancies, which are enriched for FGFR3 activating mutations, and Basal/SCCL cancers, which are enriched for EGFR mutations, could be appealing targets for targeted FGFR3 and EGFR inhibitors, respectively (10,25). Furthermore, recent research has revealed that certain MIBC subtypes are linked to immune checkpoint blockade response or response biomarkers (13,32,52). Mariathasan et al. discovered an enrichment in the GU subtype for patients responding to the anti-PD-L1 (programmed death-ligand 1) drug atezolizumab (52). Following anti-PD1 treatment with pembrolizumab, Necchi et al. discovered that basal tumors with high immunological scores have the highest 2-year progression-free survival (13).

Overall, MIBC subgroups have shown a multitude of connections with prognosis, targetable changes, and medication responsiveness, providing a viable route for improving patient treatment stratification. Despite their importance in stratifying patients for NACT, targeted treatment, and ICB, these findings need to be validated in retrospective and clinical trial research. Importantly, a consistent, clinically practical, and robust methodology for identifying these MIBC subtypes is required to both determine these relationships and ease their clinical adoption.

1.5.5 Immunohistochemistry-based profiling

Despite advancements in bladder cancer categorization, large-scale transcriptome investigations and insights from genetic subtyping have proven minimal benefit to patients. Biomarkers of prognosis or chemotherapeutic response, as well as basal/luminal profiling of bladder cancer, have yet to be used in clinical practice and play no role in treatment decision-making. The intricacy of RNA-based profiling approaches, which are expensive and time-consuming, has hampered implementation. This has made it difficult to construct a single, consistent, and straightforward methodology for determining the clinical consequences of various subtypes. Furthermore, infiltrative signals from benign stromal and immunological cells have caused confusion and discordance among subtyping systems that use this methodology. As a result, a number of studies have demonstrated the efficacy of IHC in identifying tumor intrinsic molecular subgroups (5,24,25,30,53). IHC has the limitation of often only looking at one gene product (protein) per sample, which pales in comparison to transcriptome profiling, which may look at up to 40,000 transcripts per sample. IHC, on the other hand, has the advantage of being a simple, clinically available instrument that pathologists utilize on a regular basis, and it is now widely employed in the clinical diagnosis and prognosis of a range of malignancies. IHC also avoids the drawbacks of transcriptome profiling, which does not distinguish between cancer cells and non-cancer cells. IHC allows pathologists to distinguish and analyse signals from only the tumor cells of interest when evaluating protein expression.

As a result, recent research has focused on confirming IHC-based bladder cancer subtypes by identifying putative proteome characteristics that distinguish subtype. Several of these studies use basal-luminal transcriptome profiling of MIBCs and IHC to confirm subtyping. Many of the luminal markers FOXA1, GATA3, and KRT20, as well as basal markers KRT5 and KRT14 for defining basal-luminal subtypes, are recapitulated in the Lund group's tumor cell phenotypic classifications (30,32,34,54). GATA3 and KRT5 immunohistochemistry have been discovered as the two best indicators for distinguishing between basal and luminal cancers with over 90% accuracy (30,34), and p16 expression as a marker for identifying GU cancers by distinguishing between luminal subtypes (55).

However, none of these studies employing IHC to validate subtype identification define luminal tumor subtypes, such as the GU subtype reported by the Lund taxonomy. Furthermore, many of these address the links between transcriptome and IHC phenotypes, highlighting significant aspects, but do not provide a step-by-step approach for using these stains to identify subtypes (25).

In our cohort translated these findings and methods into a 4-antibody, which uses antibodies against GATA3, CK5/6, CK 14 and CK 20 to identify intrinsic molecular subtypes of MIBC (Figure 3). This method focuses on key MIBC subtype characteristics and uses antibodies commonly found in clinical pathology labs to speed up research and clinical application of molecular subtyping.

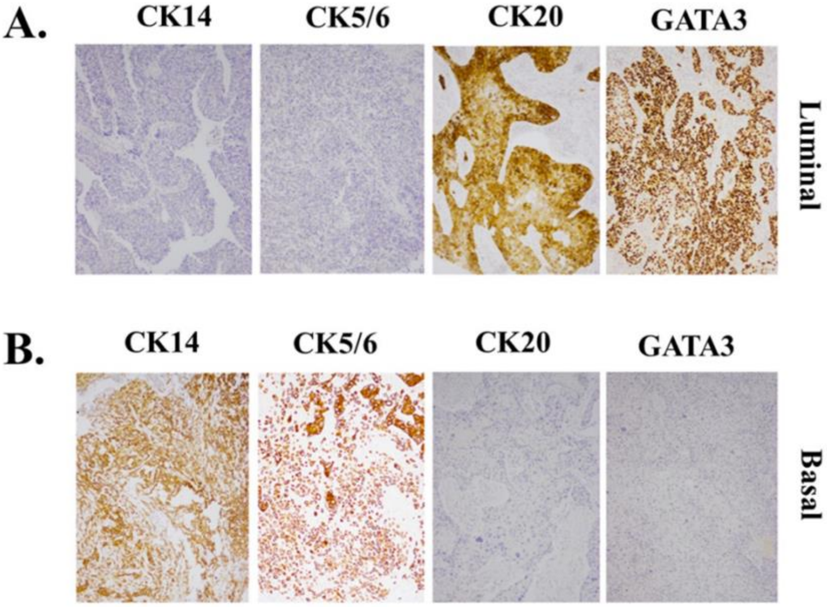


Figure 3. Immunohistochemical staining of MIBC tissues for GATA3, CK20, CK5/6, and CK14. (**A**) Luminal type, (**B**) Basal type.

The study discussed here aims to decrease the complexity of tumor intrinsic subtyping to a manageable number of antibodies, making IHC a viable alternative to transcriptome profiling. Our goal is to show that a simplified IHC subtyping assay preserves critical prognostic correlations established with more complicated profiling approaches. Future views for bladder cancer molecular classification Although different organizations now utilize different classification approaches, they all capture essential characteristics of bladder cancer biology. Clinical trials and other research investigations are increasingly using RNA and DNA sequencing and molecular categorization systems to get new insights. The reanalysis of this plethora of created data will definitely yield a much improved understanding of bladder cancer and how to treat it. Remember that existing classification systems are still evolving. More work is needed on categorization algorithm, accounting for tumor microenvironment, multi-level data integration, and clinical usefulness.

CHAPTER 2

Research methodology

2.1 Methodology: Part I

2.1.1 Study design and targeted population

A prospective study was performed. The inclusion criteria were patients, who were diagnosed with muscle invasive bladder cancer, were admitted to the university hospital in Thailand between January 2015 and December 2020 were included. Additionally, the patients had a histologically-confirmed diagnosis by a pathologist. The patients were excluded as follows: 1) unavailable and inaccessible medical record, or 2) unavailable tissue specimens for RNA sequencing.

2.1.2 Sample selection

We studied 30 MIBC cases, collected fresh frozen tissues from consecutive patients who underwent radical cystectomy in Songklanagarind hospitals, Thailand from 2015 until 2020. All tumor specimens reviewed by an experienced pathologist in bladder cancer diagnosis. Fresh-frozen tissues were collected at the time of surgical resection, and samples with size 0.5 cm were snap frozen with RNA later and kept in tube kept at -80 C for long term storage or liquid nitrogen until RNA extraction. These samples were used as quality controls, since they are a source of high-quality RNA. Seven samples of non-tumorous urinary bladder obtained during a cystectomy were used as controls. Informed consent was obtained from all patients, and the study was approved by the Ethical Committee of Songklanagarind hospital, Prince of Songkla University, in accordance with the Helsinki Declaration. (REC 61-222-10-1).

2.1.3 RNA isolation

All fresh frozen tissues were used for isolation. RNA was isolated with DNA/RNA AllPrep kit (QIAGEN). RNA was measured with Qubit® fluorometer or NanoDrop^{TM} spectrophotometer. Digital quality control (QC) analysis for mRNA was performed using the NanoString^{TM} PanCancer Progression Panel. The samples were loaded (10–35 ng RNA in a total of 30 µl loading mixture) on a cartridge and proceeded according to the manufacturers' instructions. RNA extracted from fresh-frozen tissues was assessed for quantity using Nanodrop 1000 (Nanodrop), and for quality using the 2100 Bioanalyzer (Agilent Technologies, Canada). (Figure 4)

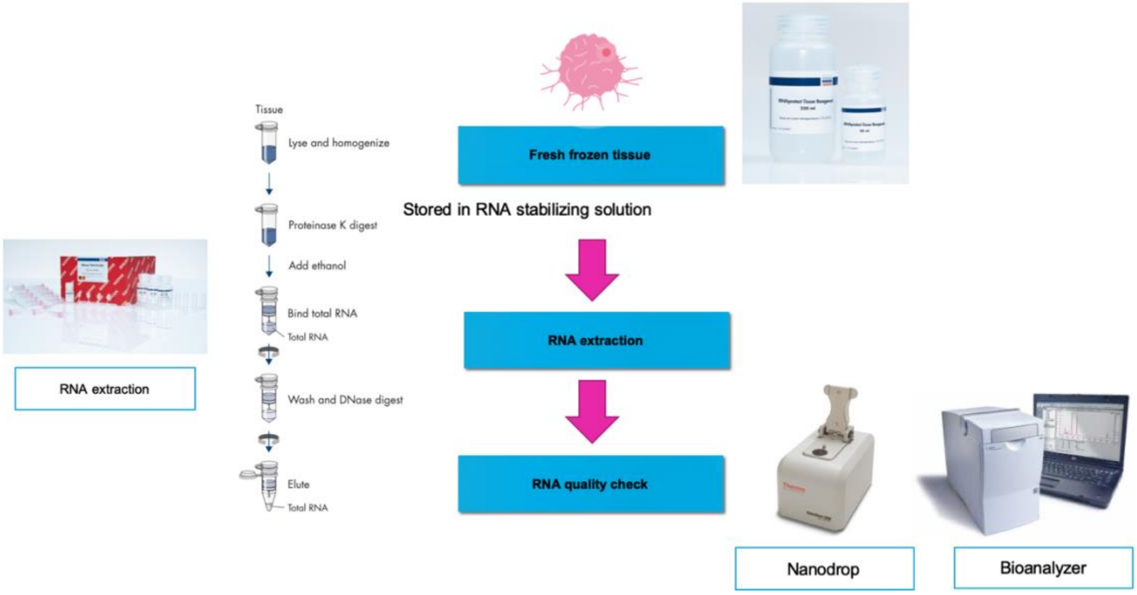


Figure 4. Sample of bladder tumor preparation for mRNA sequencing

2.1.4 Data acquisition

Our cohort: RNA-Seq data on MIBC specimens was accessed through our institutional cohort included 30 MIBC specimens. Additionally, we further confirmed the results by analyzing the 231 MIBC specimens from the TCGA was accessed. The mRNA-seq data (counts format), clinical data of 30 MIBC patients and 231 data were downloaded from the TCGA database (https://cancergenome.nih. gov/).

2.1.5 Gene expression analysis

From each RNA sample, 3 ug of total RNA was used for strand-specific library preparation. Illumina Stranded mRNA preparation kit (Illumina) was used to generate the sequencing libraries, according to the manufacturer's protocol. cDNA was prepared with random hexamer primer. The Illumina NovaSeq 6000 platform was used for transcriptome sequencing following the manufacturer's instructions.

Paired-end raw data in FASTQ format from the sequencing machine was checked for read quality, size and GC content using FASTQC program. The pipeline began with alignment step that can be done and reads were aligned to the reference genome using STAR version 2.7.8. Total mapping rate and mapped read number were analyzed using HTSeq version 0.13.5. After we re-build full mRNA sequence, the number of mRNA of each gene will be counted and the number of mRNA of each gene will represent gene expression level. Finally, the gene expression level from two group of samples will be compared using differential gene expression analysis by DESeq2 software (R package) then will get the differently expressed gene that can be candidate gene marker for bladder cancer.

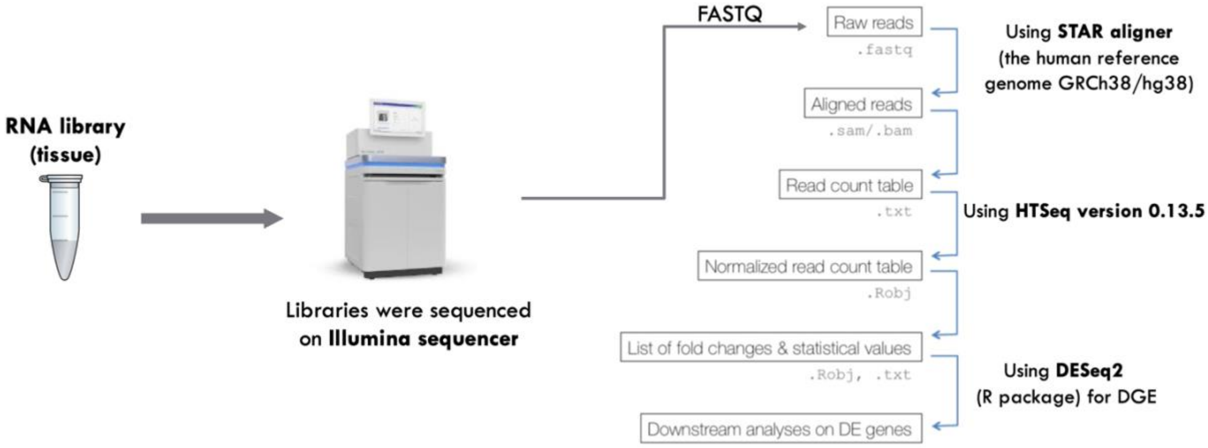


Figure 5. Transcriptomic pipeline after mRNA sequencing

Total number of mapped reads and fragement per Kilobase of exon model per million mapped reads (FPKM) were calculated for each annotated gene. The differentially expressed genes (DEGs) for 30 BC samples, 7 non-tumorous bladder tissue and 231 samples from TCGA were analyzed with the DESeq2 package, and $|\log_2 FC| > 2$ and $p < 0.05$ were set as the cutoff for DEGs. Venn algorithm was performed on the obtained DEGs and obtained differentially expressed genes in BC. The top 30 up-regulated genes in each subtype were selected and subjected to heatmap analysis and three-dimensional principal component analysis (PCA) to distinguish different molecular subtypes. The false discovery rate (FDR) measures the proportion of false discoveries among a set of hypothesis tests called significant. This quantity is typically estimated based on p-values or test statistics. In some scenarios, there is additional information available that may be used to more accurately estimate the FDR.

For further investigation, the gene expression value from mRNA-seq was log2-transformed. (Figure 6-8)

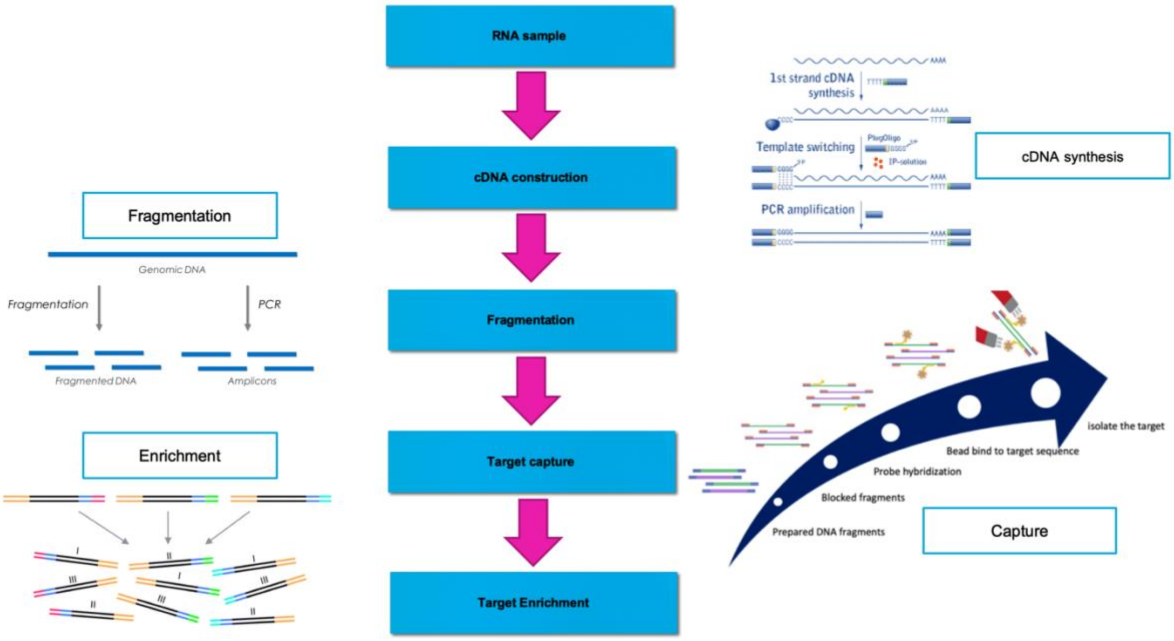


Figure 6. library preparation

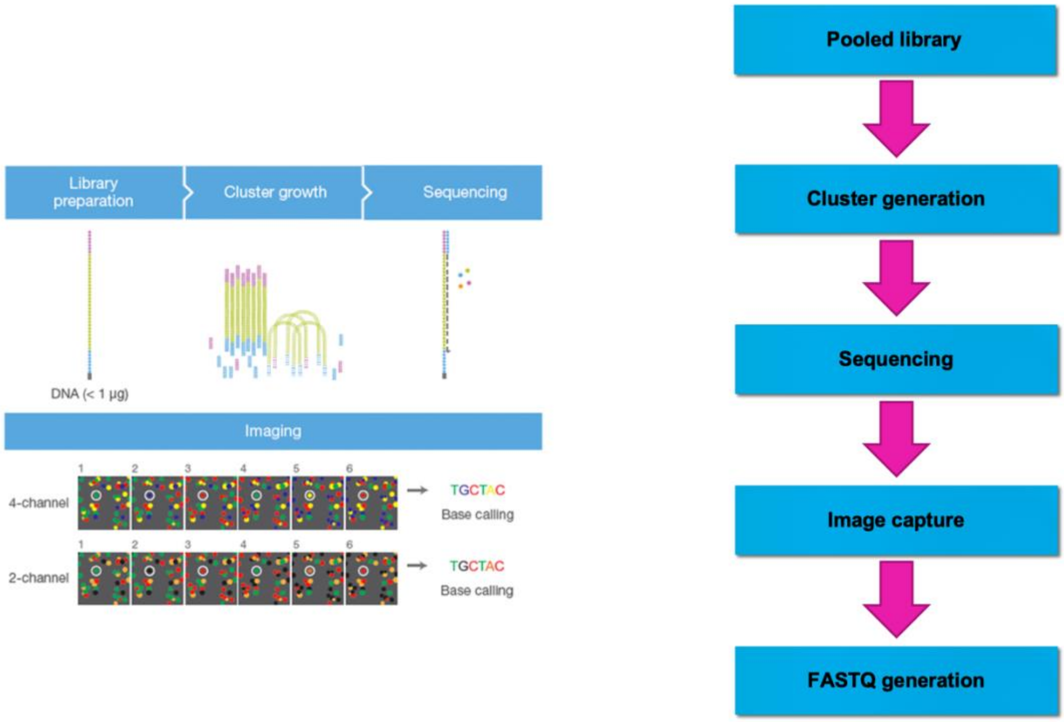


Figure 7. Massive parallel sequencing

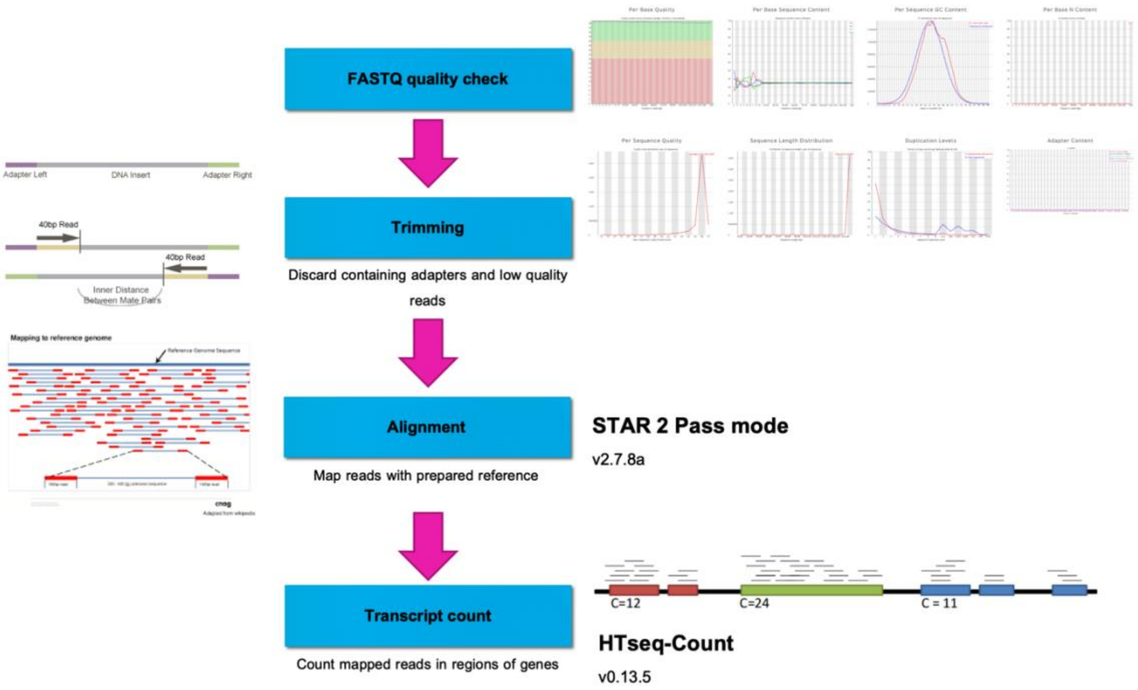

FASTQ quality check
Adapter Left
DNA Insert
Adapter Right
40bp Read
40bp Read
Inner Distance Between Mate Pairs
Trimming
Discard containing adapters and low quality reads
Mapping to reference genome
Reference Genome Sequence
cnag
Adapted from wikipedia
Alignment
Map reads with prepared reference
STAR 2 Pass mode
v2.7.8a
C=12
C=24
C = 11
Transcript count
Count mapped reads in regions of genes
HTseq-Count
v0.13.5

Figure 8. Alignment and transcript count

Log2 fold change (logFC) expression and normalized mean counts are shown in the MA-plot for each gene in comparison to the control group. Depending on the logFC threshold the user specifies and the expression directionality, different colors are utilized to denote distinct characteristics (UP or DOWN). Volcano plots, which stress both rate of expression (logFC) and statistical significance, are widely used and maybe the most instructive (p-value). Gene-specific tests (y-axis) versus logFC have negative log10-transformed p-values (y-axis) (on the x-axis). There is a distinct clustering of data points with low p-values at the top of the graph. Equidistant points' direction shift (up and down) is calculated using logFC values. Features that are more prominent than others are highlighted in red according to the selected cut-off values (83).

2.1.6 Principle component analysis (PCA)

Principal Component Analysis (PCA) is one of the techniques in unsupervised learning that is used to reduce the dimensions of the data with minimum loss of information. PCA is usually used for data that has many features.

Generally, the Principal Component Analysis (PCA) steps are: 1) Scaling our data. This is important because PCA is an algorithm that is strongly influenced by the size of each column, 2) Calculate Covariance Matrix, 3) Calculate Eigenvalues and Eigenvector, 4) Sort Eigenvalues and Eigenvector, 5) Pick top-2 or top-3 (or any amount of Principal Components that you want) eigenvalues and 6) Transform the original data.

PCA plots are a great way to display the combined effect of experimental variables and batch effects. PCA depicts groups of samples that, in an ideal world, would correspond to each of the RNA-Seq conditions. First, the most important group is clustered, followed by groupings that are less important. It is advisable to remove a repetition from the analysis if there are enough other samples from different situations to do so (at least two). It could also demonstrate if there is a batch effect problem, where samples in the same condition are spread out over a large area of a plot. To determine which samples are from a different batch in this circumstance, the user can simply rerun the analysis. Double-checking with data from the wet lab sample preparation is still recommended (83).

The *K-Means Clustering* is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data. K-means clustering is an unsupervised learning technique to group data by considering the centroid of each data group. In other words, the data will be grouped by the nearest centroid.

The stages of K-means are 1) Determine the number of clusters (k), 2) The algorithm will choose ‘k’ objects randomly from the data as the center of the cluster, 3) The rest of the data will be entered into the cluster. It will belong to the cluster whose center is closest, 4) After cluster ‘k’ is created, the algorithm will select a new center by calculating the average value of all data in the cluster, 5) Then, the rest of the data will be iteratively updated again (because the centers are now different — there may be data points closer to the center of the new cluster than the center of the original cluster). This step is carried out until no more members have moved clusters and determine the value of k using the elbow method.

2.1.7 External validate TCGA cohort

The Cancer Genome Atlas provided data for the muscle invasive bladder cancer (MIBC) cohort used in this study, which was used in previous research. The clinicopathological and mRNA data from the cohort were downloaded using the open access site cBioPortal (https://www.cbioportal.org), leaving a total of 231 patients in the study for further analysis.

2.1.8 Statistical analysis

Statistical analyses were performed using R software version 4.1.10. Subtypes’ association with clinical outcome was analyzed by univariate (single parameter logistic regression) analysis. ROC curves were used to compute sensitivity and specificity. Mean, median, and 95% CI of sensitivity and specificity were calculated. Hosmer-Lemeshow test was performed to confirm the model's goodness of fit. The Cox proportional hazard model with stepwise selection was used to assess subtypes' correlation OS. One of the most popular regression techniques for survival outcomes is Cox proportional hazards regression analysis. There are several important assumptions for appropriate use of the Cox proportional hazards regression model, including independence of survival times between distinct individuals in the sample, a multiplicative relationship between the predictors and the hazard (as opposed to a linear one as was the case with multiple linear regression analysis, discussed in more detail below), and a constant hazard ratio over time. Tests of hypothesis are used to assess whether there are statistically significant associations between predictors and time to event.

The Cox proportional hazards model is called a semi-parametric model, because there are no assumptions about the shape of the baseline hazard function. There are other regression models used in survival analysis that assume specific distributions for the survival times such as the exponential, Weibull, Gompertz and log-normal distributions. The exponential regression survival model, for example, assumes that the hazard function is constant. Other distributions assume that the hazard is increasing over time, decreasing over time, or increasing initially and then decreasing.

Kaplan-Meier plots with log-rank statistics categorized MIBC patients into outcome risk categories. Molecular subtypes and age were compared. The Bonferroni adjustment was employed to correct for multiple testing. The significance of univariable Kaplan-Meier regressions was assessed using the log-rank and Wilcoxon tests. Multivariable analyses using Cox proportional hazard regression. For results from the univariable analysis a p value cut-off of <0.2 was chosen to include relevant clinical or pathologic parameters that would have been missed with a more restrictive p value of <0.05. Contingency analyses of nominal variables were performed with Pearson’s chi-squared test. Variables for the multivariable analysis included significant (p<0.2) clinicopathological characteristics on univariable analysis (pT-Stage, pN-Stage, age, gender,) and genes (Statistical analyses of numeric continuous variables were performed with non-parametric tests (Wilcoxon rank-sum test, Kruskal-Wallis test).

2.2 Methodology: Part II

2.2.1 Cohort composition

Criteria for patient inclusion was diagnosis of MIBC (Stage T2+) at radical cystectomy. Patients included in this study underwent radical cystectomy following a diagnosis of muscle invasive bladder cancer (stage pT2+) or high risk NMIBC (high grade, recurrent tumors with aggressive features). Specimens were collected from formalin-fixed paraffin-embedded (FFPE) tissue obtained from TURBT and cystectomy procedures. MIBC tumour samples were obtained from archived tissue samples from Songklanagarind hospital (cystectomies). All samples of our cohorts were obtained from cystectomy procedures performed at Songklanagarind hospital. Clinical and pathologic information for all patients and samples was compiled into a research database. Hematoxylin and Eosin (H&E) samples underwent pathologic review and were annotated by pathologists to select regions of interest for tissue coring and tissue microarray construction. High grade muscle-invasive tumor regions were selected, with samples excluded on the basis of lack of muscle-invasive disease, insufficient amount of tissue or sample unavailability.

2.2.2 Study design and targeted population

This study included 132 patients with urinary bladder cancer who underwent radical cystectomy and who received standard adjuvant chemotherapy at Songklanagarind Hospital, Thailand from 2008 to 2016. Inclusion criteria were patients with bladder cancer aged older than 15 years who underwent surgery primarily at our institute and who completed adjuvant treatment according to the standard of the Thai Urological Association. All eligible cases were reviewed for clinical stage, and their histopathology was confirmed by a pathologist. Staging was performed according to the TNM classification, whereas stage grouping was performed according to the eighth version of the American Joint Committee on Cancer Staging Manual.

Cases without muscularis propria invasion and those with subtypes other than non-urothelial carcinoma were excluded. Clinical data were extracted from the electronic medical records of the hospital (HIS system). Data on survival status combined with the clinical follow-up records and death registry data from the Thai citizen registration system were analyzed and archived by the Cancer Unit, Songklanagarind Hospital. Cases with operative mortality were excluded from the survival analysis. The study protocol was approved by the Human Research Ethic Committee of the Faculty of Medicine, Prince of Songkla University (REC61-222-10-1). All methods were carried out in accordance with the World Medical Association Declaration of Helsinki. Informed consent was obtained from all patients or legally authorized representatives included in the study.

2.2.3 IHC study by tissue microarray

Sampling of the tumor part for this pilot study was performed by a collaborative work between the attending surgeon who know the orientation of the specimen and the pathologist who examined the histopathology. Bladder carcinoma in situ and flat lesions were excluded in this study. Several areas of tumor in the same patients for the pathological morphology and selected the representative areas that have both richness in tumor cells and the morphology was like other areas in the same cases were selected for examination. Archived pathological specimens from all included cases were retrieved as formalin-fixed paraffin- embedded tissue blocks, which were then selected and prepared as 5-μm sections for a tissue microarray (TMA) using a tissue arrayer (Quick-Ray, UT06; UNITMA, Seoul, Korea). Immunostaining procedures were conducted with 3 (triplicate) TMA cores per section by a pathology technician who specializes in this technique. In cases of multiple foci, all foci were selected for examinations. Subtype-specific primary antibodies used here are as follows: GATA3 (UMAB218, 1:100 dilution; OriGene, MD, USA), CK5/6 (D5/16, 1:50 dilution; Dako, Glostrup, Denmark), CK14 (OIT4A7, 1:100 dilution; OriGene), and CK20 (OTI4A, 21:50

dilution; OriGene). These antibodies were used to identify potential markers to establish molecular subtypes in the tissue sections contained in the TMA. A pathologist (Kanet Kanjanapradit) blinded to the clinical outcomes examined the results using a light microscope and scored all TMA sections. For mixed and/or borderline cases, the positive immunostains were interpreted 2 times with the consensus of a pathologist. The positivity and intensity of tumor cell nuclei stained for GATA3 and membranous or cytoplasmic staining for CK20, CK5/6, and CK14 were recorded. Staining intensity was assessed as 0 (negative; 0–10%) or 1 (positive; 10–100%).

2.2.4 Statistical analysis

Categorical and continuous parameters were compared using the Chi-square test and were analyzed using the Spearman rank correlation test. The median differences between groups for non- normally distributed variables were evaluated by independent sample Kruskal–Wallis test. Differences in the percentages of IHC staining between or among comparable groups were analyzed using the Student's *t* test and one-way analysis of variance. The hazard ratios (HRs) and 95% confidence intervals (CIs) were also calculated. In all patients who underwent radical cystectomy with perioperative chemotherapy, the OS after radical cystectomy was calculated using the Kaplan–Meier method. Survival probabilities were estimated using the Kaplan–Meier method, whereas the log-rank test was adopted to compare survival probabilities between each variable. All variables with $p \leq 0.1$ in the univariable analyses were entered into the multivariable regression analysis. Multivariable analyses were also performed using Cox regression as described in section 2.1.8. Two-sided p values < 0.05 were considered statistically significant. The R program (version 4.0.1) was used for statistical analyses.

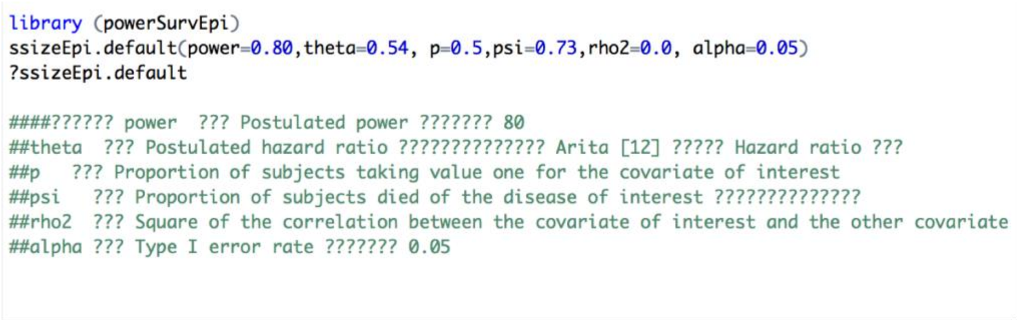
2.2.5 Sample size calculation

The DEPTh model is approaches categorized clinical research questions into the four types. In detail, D stands for Diagnosis, E for Etiology, P for Prognosis, and or Therapy (or intervention). For objective studied the survival of each subtype as therapy research in the DEPTh model. Sample size calculation was performed using Schoenfeld's formula

Schoenfeld's formula as follows:

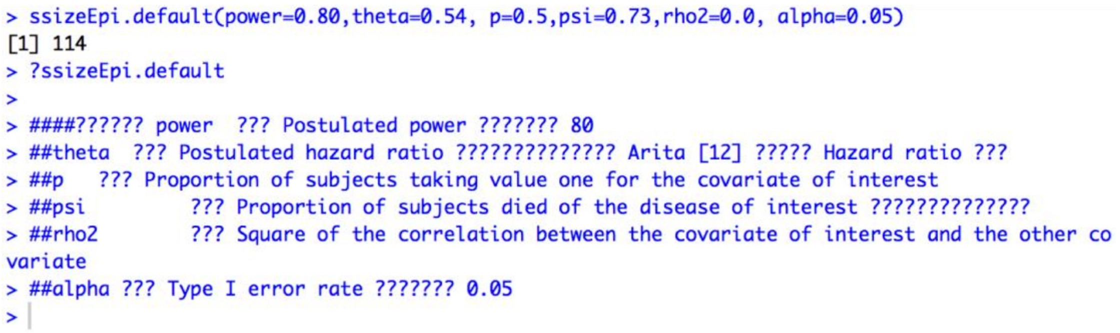
$$n = \frac{4(Z_{1-\alpha} + Z_{1-\beta})^2}{(\log\theta)^2}$$

Where *Z 1-α* equals 1.96 (confidence level 95%), *Z 1-β* equals 0.84 (power 80%), and *θ* denotes the hazard ratio



```
library (powerSurvEpi)
ssizeEpi.default(power=0.80,theta=0.54, p=0.5,psi=0.73,rho2=0.0, alpha=0.05)
?ssizeEpi.default

####?????? power  ??? Postulated power ??????? 80
##theta  ??? Postulated hazard ratio ?????????????? Arita [12] ????? Hazard ratio ???
##p   ??? Proportion of subjects taking value one for the covariate of interest
##psi   ??? Proportion of subjects died of the disease of interest ??????????????
##rho2  ??? Square of the correlation between the covariate of interest and the other covariate
##alpha ??? Type I error rate ??????? 0.05
```



```
> ssizeEpi.default(power=0.80,theta=0.54, p=0.5,psi=0.73,rho2=0.0, alpha=0.05)
[1] 114
> ?ssizeEpi.default
>
> ####?????? power  ??? Postulated power ??????? 80
> ##theta  ??? Postulated hazard ratio ?????????????? Arita [12] ????? Hazard ratio ???
> ##p   ??? Proportion of subjects taking value one for the covariate of interest
> ##psi         ??? Proportion of subjects died of the disease of interest ??????????????
> ##rho2        ??? Square of the correlation between the covariate of interest and the other co
variate
> ##alpha ??? Type I error rate ??????? 0.05
> |
```

CHAPTER 3

Results

3.1 Results: Part I

3.1.1 Clinical characteristics

All tissue samples were from patients recruited at the Songklanagarind Hospital, Songkhla, Thailand (Table 1). These included tumor tissue from 26 males and 4 females with ages between 52-92 years old. The data of 231 MIBC cohorts retrieved from The Cancer Genome Atlas (TCGA) was also shown in Table 1 that included the clinical information from 304 males and 108 females between the age of 32-90 years old. It should be noted that there is a quite difference in the proportion of T stages and N stages between data from tissue samples and TCGA cohort. Moreover, no metastasis was found in all Thai MIBC patients.

Table 1. Clinical data summary of studied MIBC datasets

	Thai patient dataset	**Percentage**	**TCGA dataset**	**Percentage**
Samples	**30**		**231**	
Average of age (range)	**67.5 (52- 92)**		**69 (46-90)**	
Gender				
Male	**26**	**86.2**	**169**	**73.16**
Female	**4**	**13.8**	**62**	**26.64**
ECOG				
0	**21**	**70**	**158**	**68.5**
1	**9**	**30**	**73**	**31.5**
T stage				
T 2	**24**	**80**	**75**	**32.47**
T 3	**6**	**20**	**123**	**53.25**

T 4	**0**	**0**	**33**	**14.28**
N stage				
N 0	**22**	**73.3**	**143**	**58.01**
N 1	**7**	**23.3**	**28**	**10.68**
N 2	**1**	**3.3**	**42**	**18.2**
N 3	**0**	**0**	**4**	**1.94**
N x	**0**	**0**	**14**	**6.06**
M stage				
M 0	**30**	**100**	**116**	**50.22**
M 1	**0**	**0**	**5**	**2.16**
Not available	**0**	**0**	**110**	**47.62**

3.1.2 Transcriptome profiling and classification of Thai MIBC

The transcriptome sequencing of all tissue samples was performed based on the strand-specific library preparation to identify the expression levels of all genes. Figure 9 presents the boxplots which provide an easy way to visualize the count distribution in each sample.

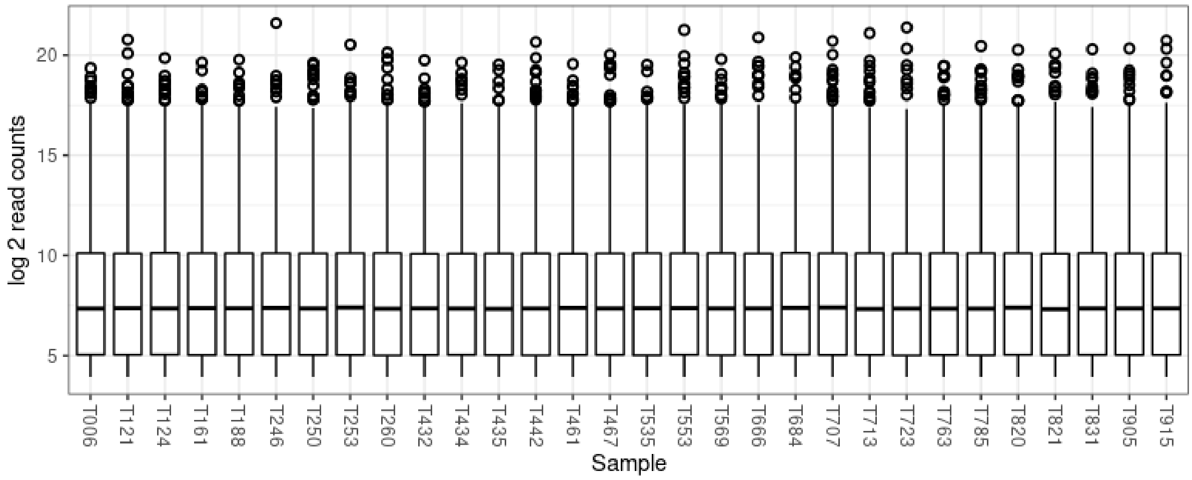


Figure 9. Boxplot with normalized counts. The frequency distribution and some statistics like mean, median and outliers are represented in these plots of log2 normalized counts

MA and Volcano plot analysis demonstrated more than a hundred of genes were found to be up-regulated and downregulated in MIBC compared to normal bladder tissue (Fig. 10 and 11).

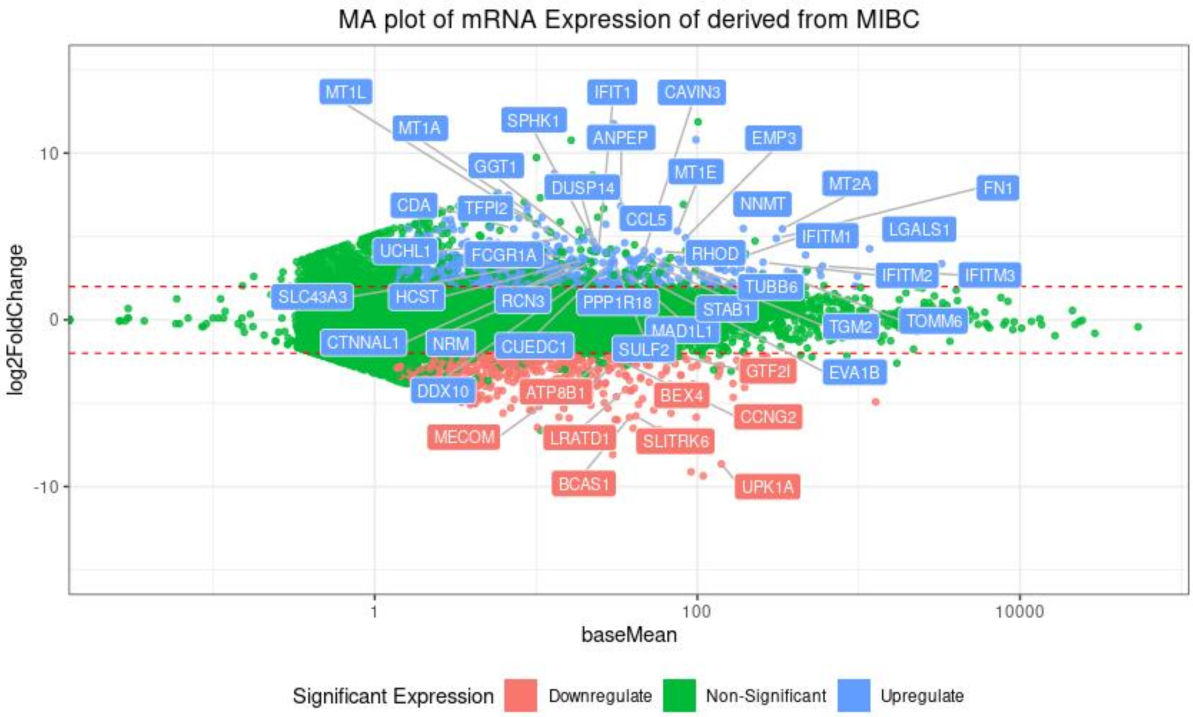


Figure. 10 MA plot analysis demonstrated more than a hundred of genes were found to be up-regulated and downregulated in MIBC

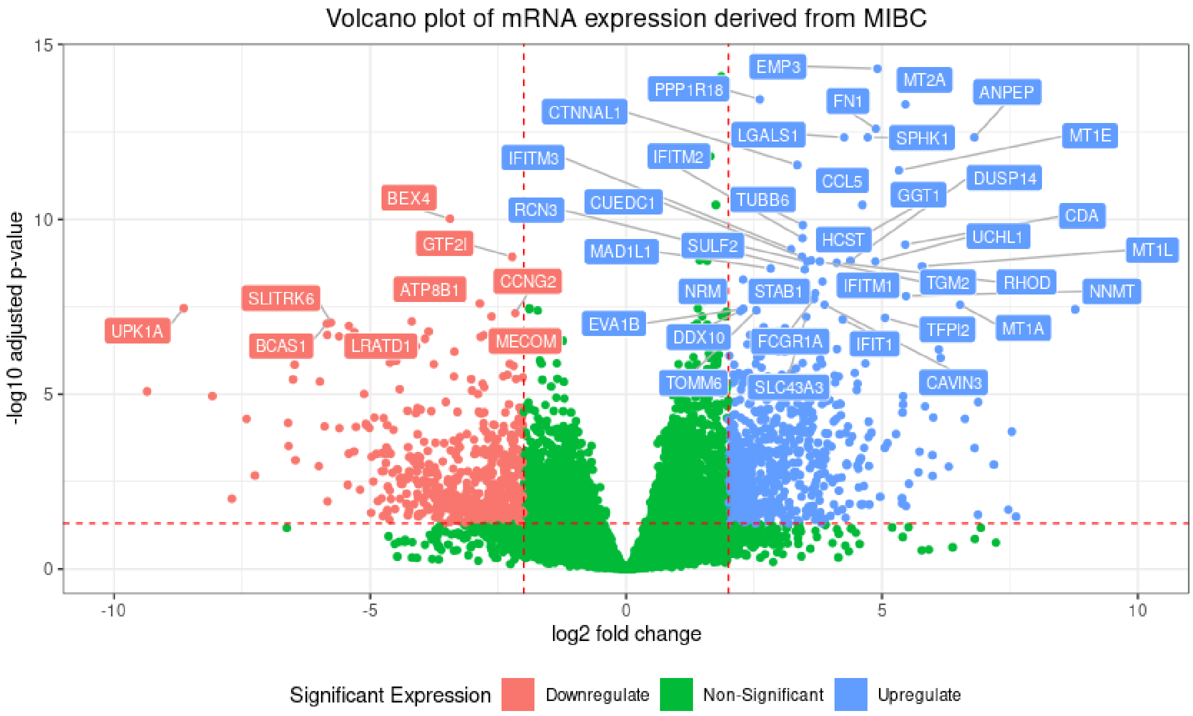


Figure 11. Volcano plot analysis demonstrated more than a hundred of genes were found to be up-regulated and downregulated in MIBC

The thirty most significantly changed genes included PI3K-Akt signaling molecules (*FN1*, *COL6A2*, and *COL1A2*), MAPK pathway related molecules (*TGFB1* and *MECOM*), mitochondrial biogenesis regulators (*MT1A* and *MT2A*), exosomal proteins (*TUBB6*, *TUBB3*, *LGALS1* and *IFITM3*), biomolecules metabolism (*CDA*, *SPHK1*, *MAOA* and *MGST1*), and others. To identify the optimal number of clusters based on transcriptomic classification, Elbow plot analysis was applied for Thai MIBC transcriptome data. The result showed that the three clusters were found to be optimal as demonstrated in Fig. 12.

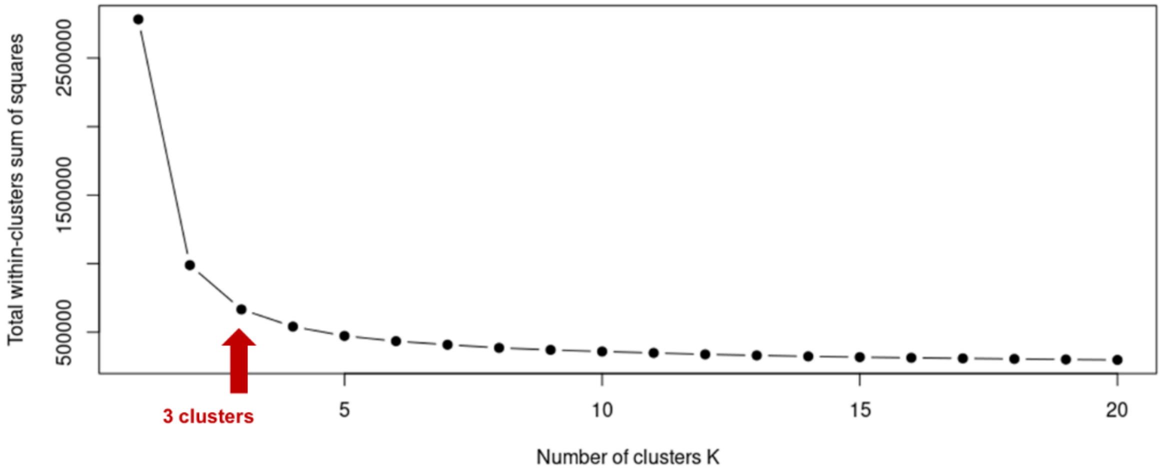


Figure 12. Elbow plot analysis was applied for Thai MIBC transcriptome data. The result showed that the three clusters were found to be optimal

The transcriptomic data of all MIBC tissue samples were then subjected to the classification of the MIBC groups by using principal component analysis by K-mean clustering (Fig. 13).

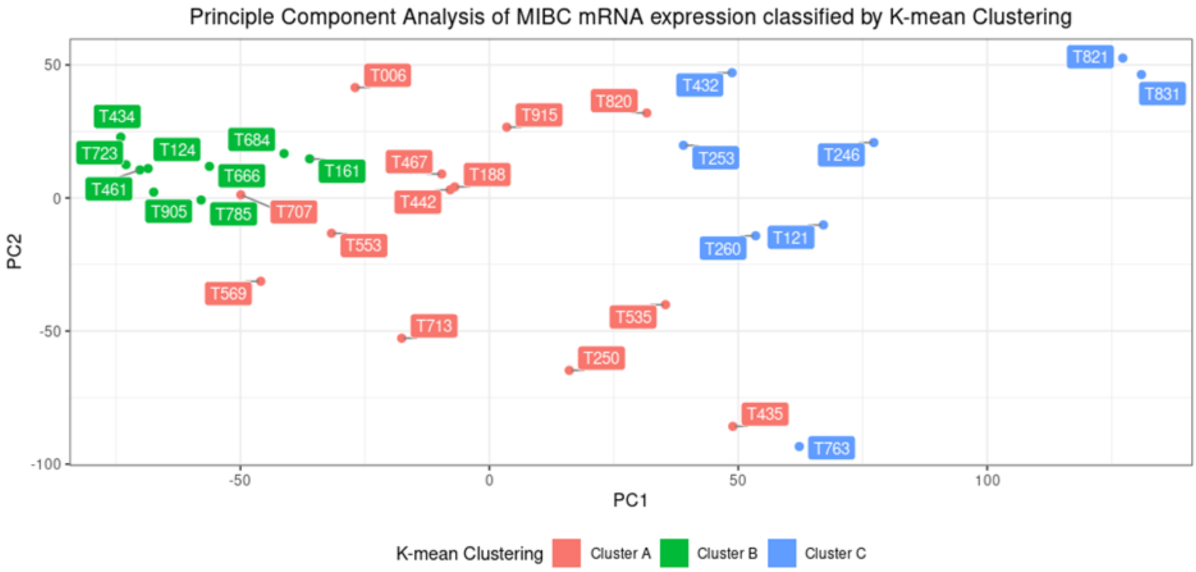


Figure 13. The transcriptomic data of all MIBC tissue samples were then subjected to the classification of the MIBC groups by using principal component analysis by K-mean clustering

Heatmap analysis of gene expression derived from the thirty most significantly changed genes revealed the obvious unique pattern of each MIBC cluster confirming the specific character of each group of MIBC patients (Fig. 14).

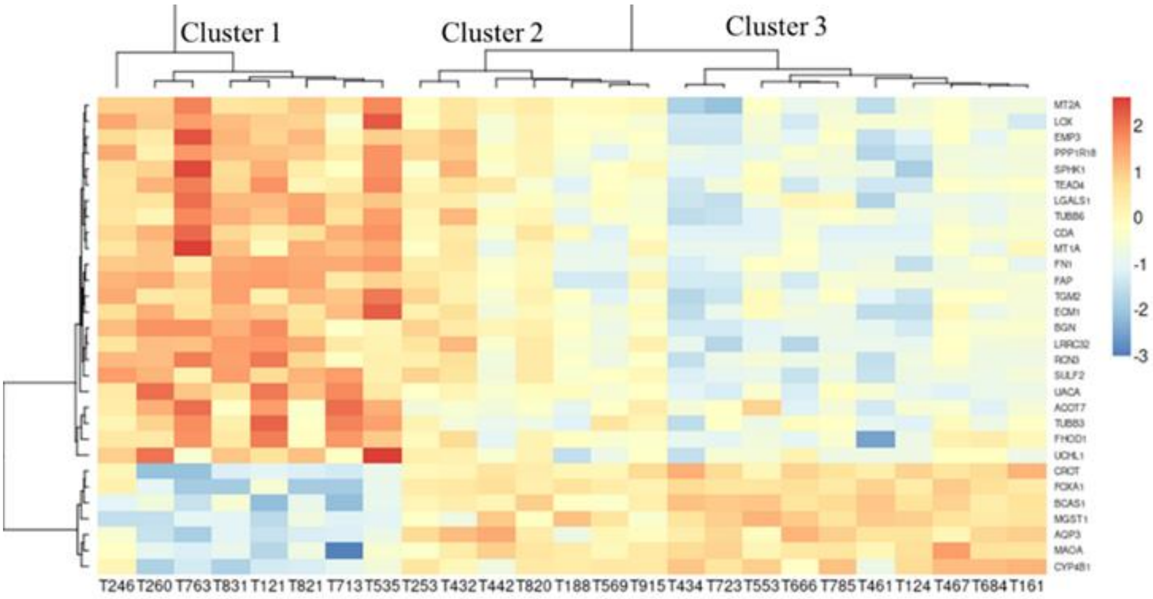


Figure 14. Heatmap analysis of 30 gene expression derived from the thirty most significantly changed genes revealed the obvious unique pattern of each MIBC cluster confirming the specific character of each group of MIBC patients. The genes were ranked following by adjust p-value.

Example of each gene expression in difference 3 clusters were shown in Fig. 15

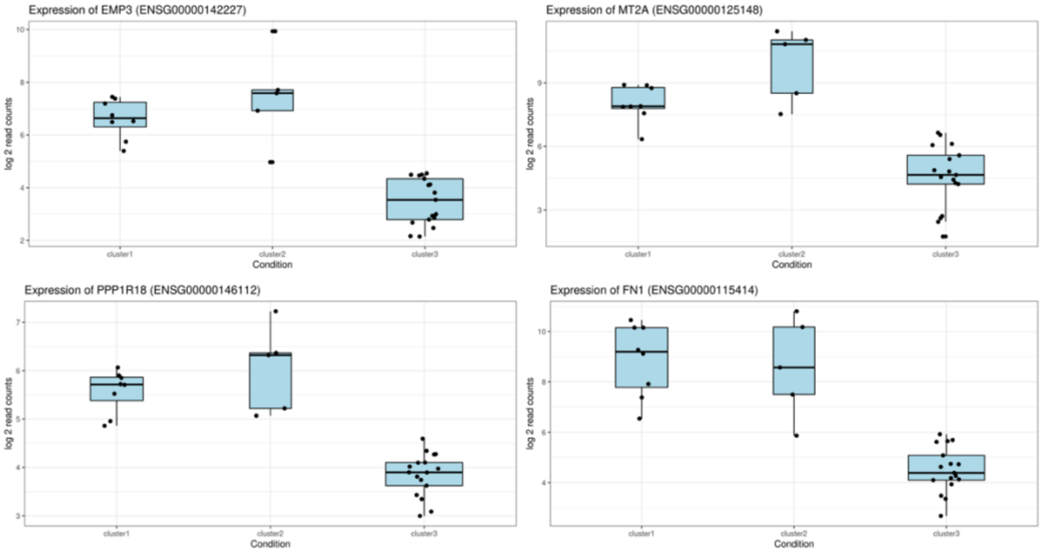


Figure 15. Boxplot of examples gene expression in 3 clusters

3.1.3 Differential gene expression (DEG) analysis revealed 37 genes expressed with the different levels among clusters

In addition to the enrichment study, the data from differential gene expression (DEG) analysis also revealed the number of genes that expressed differently between two clusters, as displayed in the Venn diagram (Fig. 16).

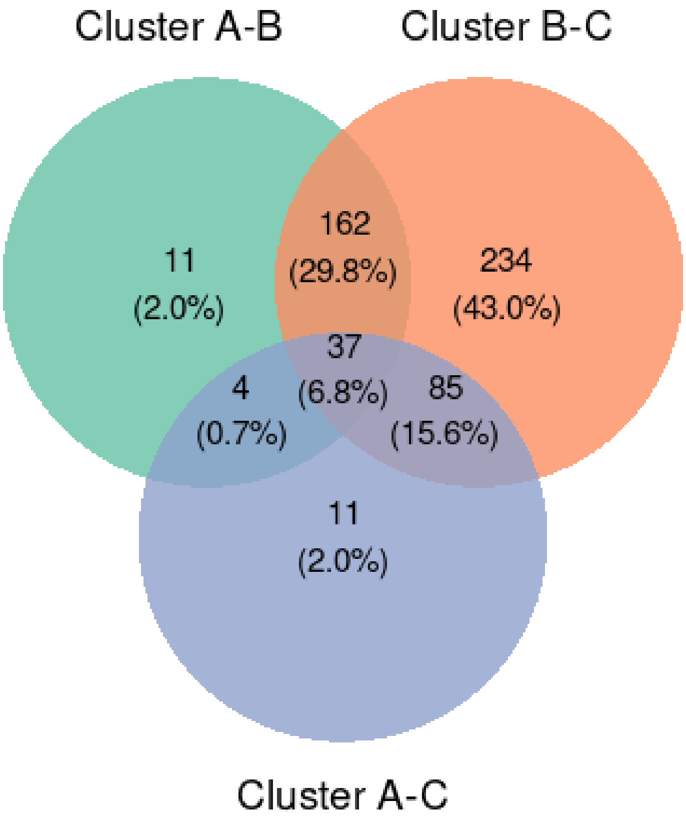


Figure 16. The data from differential gene expression (DEG) analysis also revealed the number of genes that expressed differently between two clusters, as displayed in the Venn diagram

Importantly, 37 genes were observed to be expressed differently in all clusters. These included the genes related to calcium signaling pathway (*BDKRB1*, *EDNRA*, *AVPR1A*, *PTGER3*, *PTGFR*, *NTRK3*, *P2RX1*, etc.*)*, PI3K-Akt signaling pathway (*COL6A2*, *COL1A2*, *ITGA8*, *CREB5*, *COL6A3)*, MAPK signaling pathway (*FGF7*, *NGF*, *HGF*, *ANGPT1*), or cGMP-PKG signaling pathway (*KCNMB1*, *KCNMA1*, *ADRA2A*, *ATP1B2*, *ADORA1*, *PRKG1).* The expression levels of each gene were demonstrated as Volcano plots for each cluster (Fig. 17, 18, and 19). Interestingly, all 37 genes were significantly up-regulated in clusters A and C, but down-regulated in cluster B. The most significantly expressed genes in cluster A included

BDKRB1, *EDNRA*, *AVPR1A*, *PDGFRB*, and *TNC*, while *COL6A3*, *COL1A1*, *COL6A2*, *PDGFRB*, and *PRKG1* were found to be the top 5 genes highly expressed in cluster C. For cluster B, the collagen-related genes (*COL6A3*, *COL1A2*, *COL6A2)*, tenascin C (*TNC*), and fibroblast growth factor 2 (*FGF2*) were the transcripts statistically suppressed.

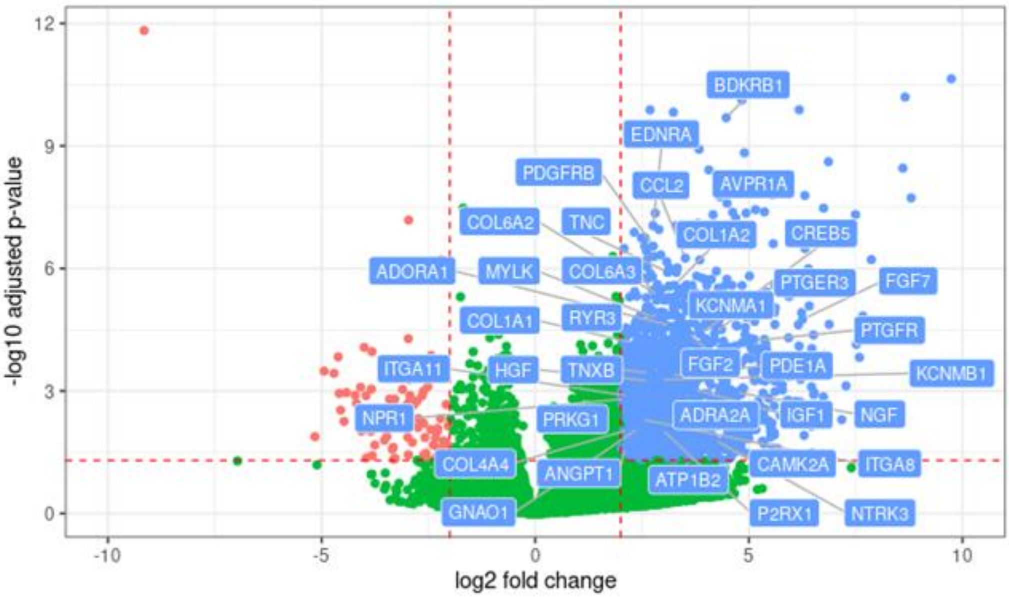


Figure 17. The expression levels of each gene were demonstrated as Volcano plots for cluster A

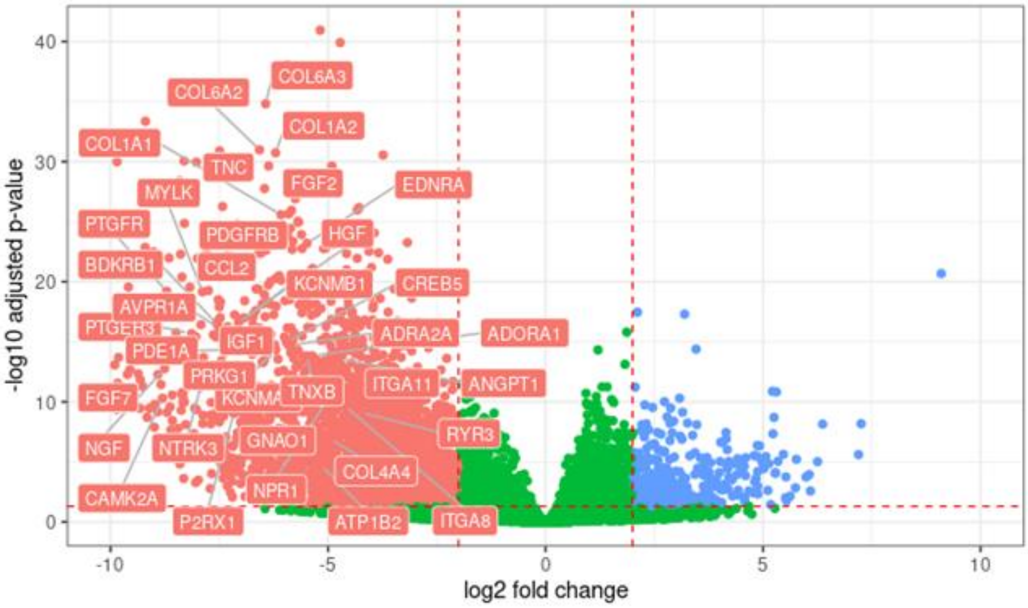


Figure 18. The expression levels of each gene were demonstrated as Volcano plots for cluster B

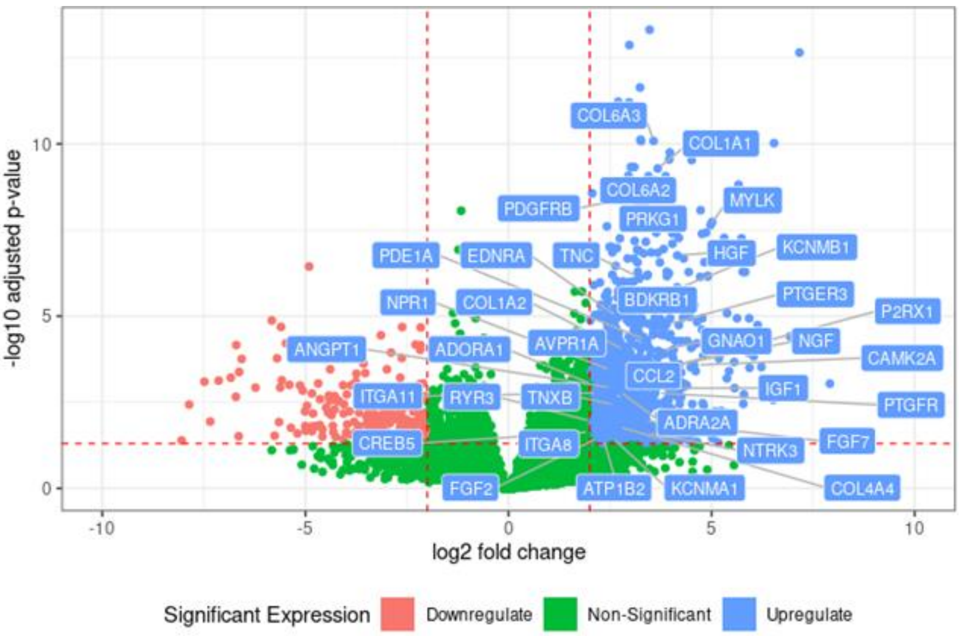


Figure 19. The expression levels of each gene were demonstrated as Volcano plots for cluster C

3.1.4 ROC analysis of 37 differentially expressed genes found in MIBC tissues

To evaluate the specificity and sensitivity of the genes expressed differently for each MIBC cluster, ROC curve analysis was performed for all 37 genes and all cluster. Interestingly, the corresponding areas under the ROC curve (AUCs) with the value more than 0.8, 0.9, and 0.95 were found in 33, 25, and 14 genes from 37 genes for cluster B (Table 2 and Fig. 20). While the AUC values above 0.9 were observed only in 5 genes for cluster C, no one could be found for cluster A. The highest AUC value of cluster B was 0.988 for *PDGFRB* and *COL6A2* genes meanwhile the lowest AUC was 0.72 from *ITGA8* gene (Table 2 and Fig. 20). *KCNMB1* is the gene presented the highest AUC with the value of 0.936 in cluster C while *ITGA11* showed the lowest AUC with 0.67. Cluster A displayed the lowest value of mean AUC, the range of AUC values of differential expressed genes in this cluster were between 0.52-0.869 which *RYR3* gene showed the lowest and *COL1A1* are the highest AUC.

Table 2. The area under the curve (AUC) from specificity and sensitivity of the genes expressed differently for each MIBC cluster

Genes	**cluster A**	**cluster B**	**cluster C**
CCL2	0.7680	0.941	0.859
FGF2	0.6940	0.947	0.886
RYR3	**0.5200**	0.721	0.735
MYLK	0.8040	0.955	0.914
EDNRA	0.8110	0.951	0.833
FGF7	0.7890	0.976	0.927
BDKRB1	0.7030	0.872	0.727
PDGFRB	0.8380	**0.988**	0.838
HGF	0.7310	0.945	0.867
AVPR1A	0.6200	0.774	0.708
NGF	0.6650	0.893	0.783
PTGFR	0.7230	0.947	0.912
PDE1A	0.6170	0.916	0.86
PTGER3	0.7790	0.951	0.821
CAMK2A	0.8200	0.971	0.862
NTRK3	0.5380	0.828	0.831
P2RX1	0.7710	0.96	0.907
TNC	0.8530	0.947	0.746
COL4A4	0.5890	0.852	0.785
COL6A2	0.8640	**0.988**	0.886
ITGA8	0.6030	**0.72**	0.799
COL6A3	0.8430	0.979	0.849
CREB5	0.8200	0.946	0.79
TNXB	0.6460	0.875	0.815
ANGPT1	0.6820	0.826	0.686
IGF1	0.689	0.919	0.879
COL1A1	**0.869**	0.95	0.736
COL1A2	0.857	0.967	0.787
ITGA11	0.824	0.915	**0.67**
NPR1	0.628	0.885	0.808
KCNMB1	0.749	0.957	**0.936**
ADORA1	0.761	0.89	0.762
PRKG1	0.777	0.964	0.889
ATP1B2	0.542	0.743	0.751
ADRA2A	0.582	0.915	0.845
KCNMA1	0.799	0.954	0.88
GNAO1	0.739	0.939	0.85

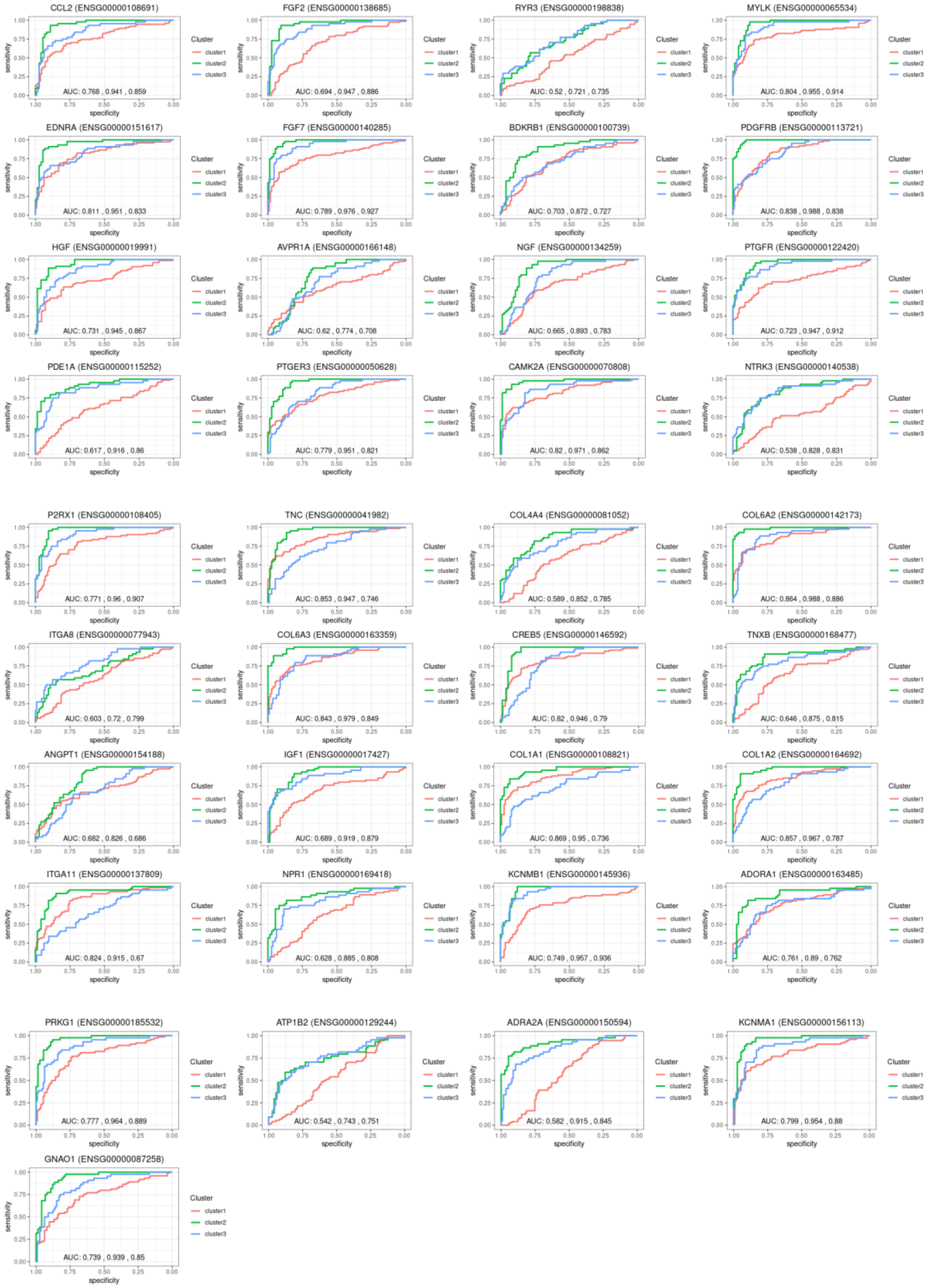


Figure 20. ROC curve analysis was performed for all 37 genes and all cluster. Interestingly, the corresponding areas under the ROC curve (AUCs) with the value more than 0.8, 0.9, and 0.95 were found in 33, 25, and 14 genes from 37 genes for cluster B

3.1.5 Clinical characteristic of MIBC in our cohort associated with treatment outcome

To identify factors associated with MIBC, logistic regression analysis was performed, including prognostic scores, patient characteristics and tumour characteristics (Table 1). Univariate analysis identified tumour stage and nodal status as significant predictors of overall suvival. The multivariate logistic regression model identified tumour stage (HR, 25.64; 95% CI, 2.31-284.04; p=0.006) as independent predictor of overall survival. For cluster C exhibited showed higher hazard ratio without statistically significant. (HR, 2.63; 95% CI, 0.44-15.79; p=0.291) (Table 3)

Table 3 Univariate and multivariate logistic regression analyses of clinical data of MIBC patients

Variables	**Univariate analysis**			**Multivariate analysis**		
	HR	95% Cl	P value	HR	95% Cl	P value
T stage						
2	Ref					
3-4	6.62	1.09–40.1	0.041	25.64	2.31-284.04	0.006
N stage						
0	Ref		0.036			
1	4.05	0.45–5.46				
2	6.46	0.35–6.86				
Age (yrs)						
<=65	Ref					
>65	2.42	0.27–21.67	0.392			
Lymph node metastasis						
negative	Ref					
positive	2.39	1.29–4.41	0.01			
LVI						
negative	Ref					
positive	1.08	0.18-6.5	0.929			
Ureteric margin						
negative	Ref					
positive	5.32	0.59–48.17	0.212			
Cluster						
Cluster A	Ref					
Cluster B	0	1.6–4.94	0.108			
Cluster C	2.63	0.44-15.79	0.291			

Univariate analysis showed no significant differences in survival between the molecular subtypes (*p* =0.108). Pairwise comparisons using log-rank tests also showed that survival not differed significantly between each molecular subtype, with patients with the cluster B subtype experiencing the longest survival, followed by those with the cluster A. The poorest survival was observed among patients with cluster C (Figure 21).

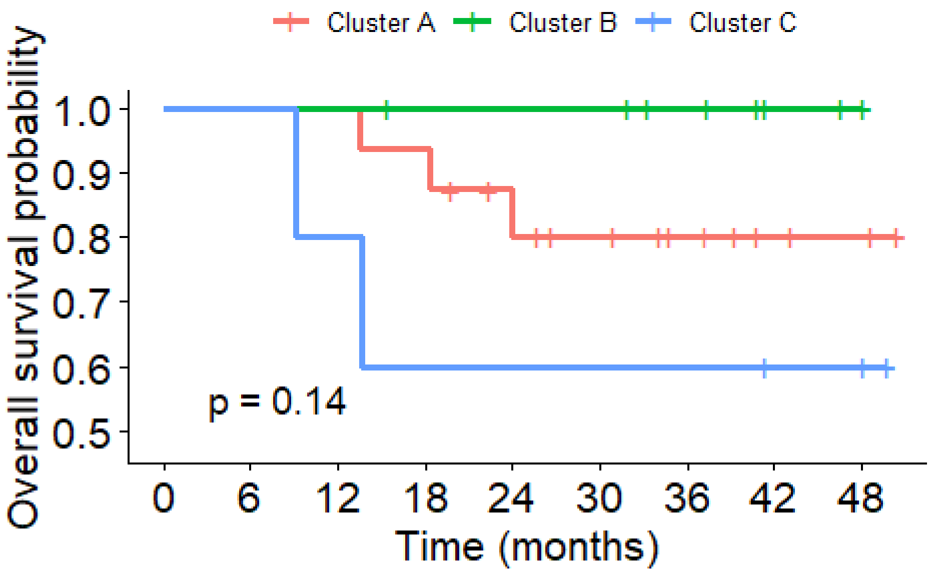


Figure 21. The Kaplan-Meier analysis demonstrated that cluster B displayed the highest probability of survival within 48 months of follow-up while cluster C showed the lowest value

3.1.6 Molecular subtypes of MIBC are associated with response of perioperative chemotherapy

We applied molecular subtype classification to tumors from 30 patients treated with preoperative chemotherapy. We henceforth analyzed response in the neoadjuvant chemotherapy. Among these patients, cluster B had better pathologic response to neoadjuvant chemotherapy (40%). (Fig. 22)

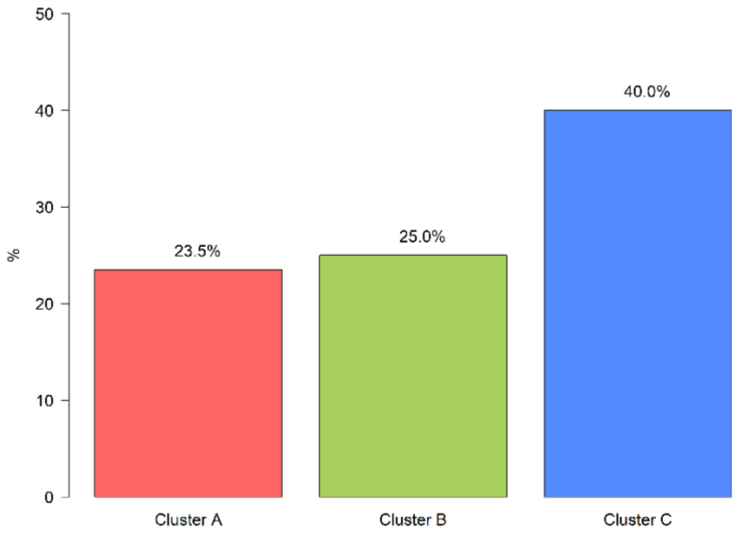


Figure 22. RNA-based molecular subtypes are associated with pathological response to neoadjuvant chemotherapy

Moreover, MIBC in cluster B also exhibited better outcome after adjuvant chemotherapy for progression or metastatic free survival. (Fig. 23)

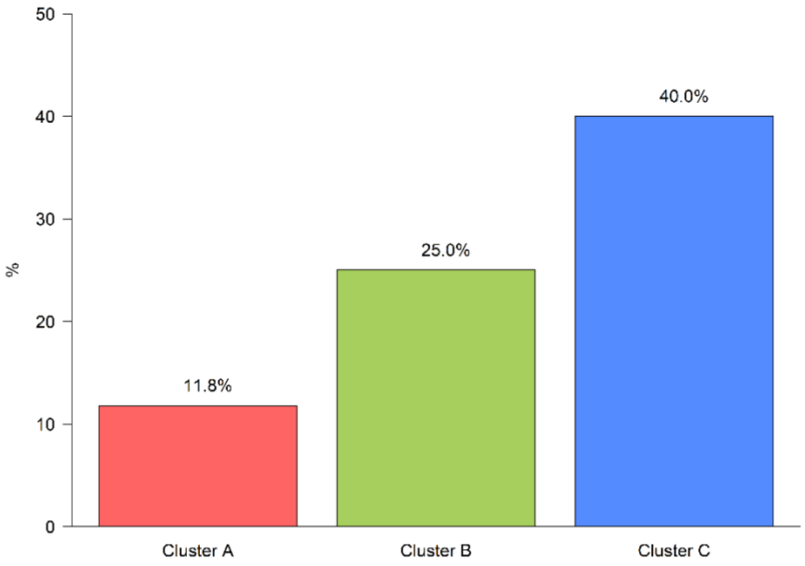


Figure 23. RNA-based molecular subtypes are associated with pathological response to adjuvant chemotherapy

3.1.7 The transcriptomic classification using PCA analysis of tissue sample with the TCGA data provided the significant prognostic value of MIBC overall survival

To increase the number of samples used in this study, we included 231 transcriptomic data from TCGA database for classification by using PCA analysis and unsupervised K-mean clustering. We decided to apply the centroid derived from MIBC tissue samples to separate PCA coordinates of TCGA cohort into three clusters according to MIBC tissues (Fig. 24).

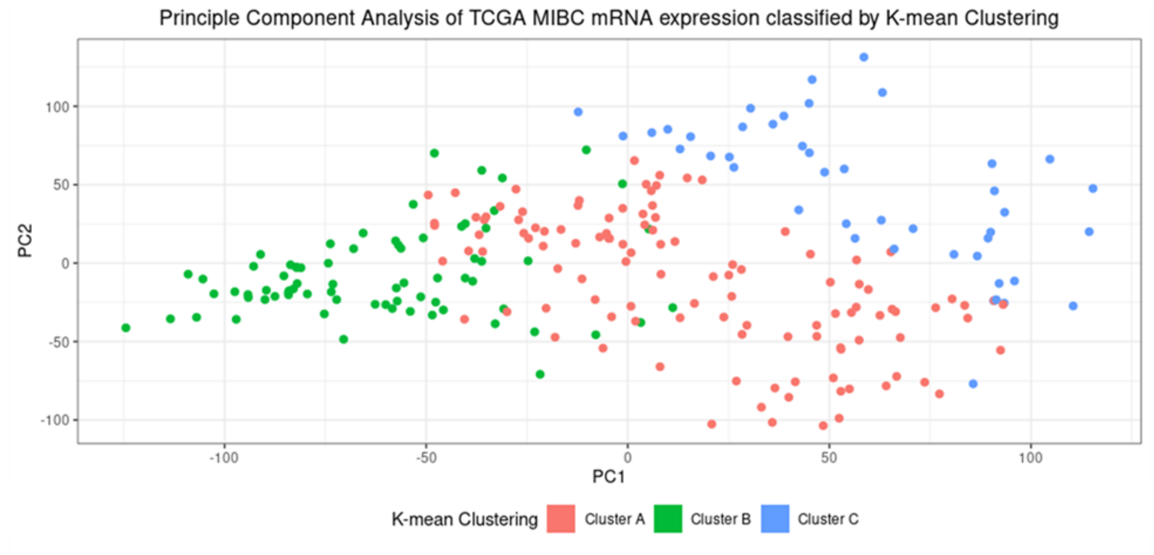


Figure 24. The centroid derived from MIBC tissue samples to separate PCA coordinates of TCGA cohort into three clusters according to MIBC tissues

MA and Volcano plot analysis demonstrated up-regulated and downregulated in MIBC comparing between non-TCGA and TCGA group (Fig. 25 and 26).

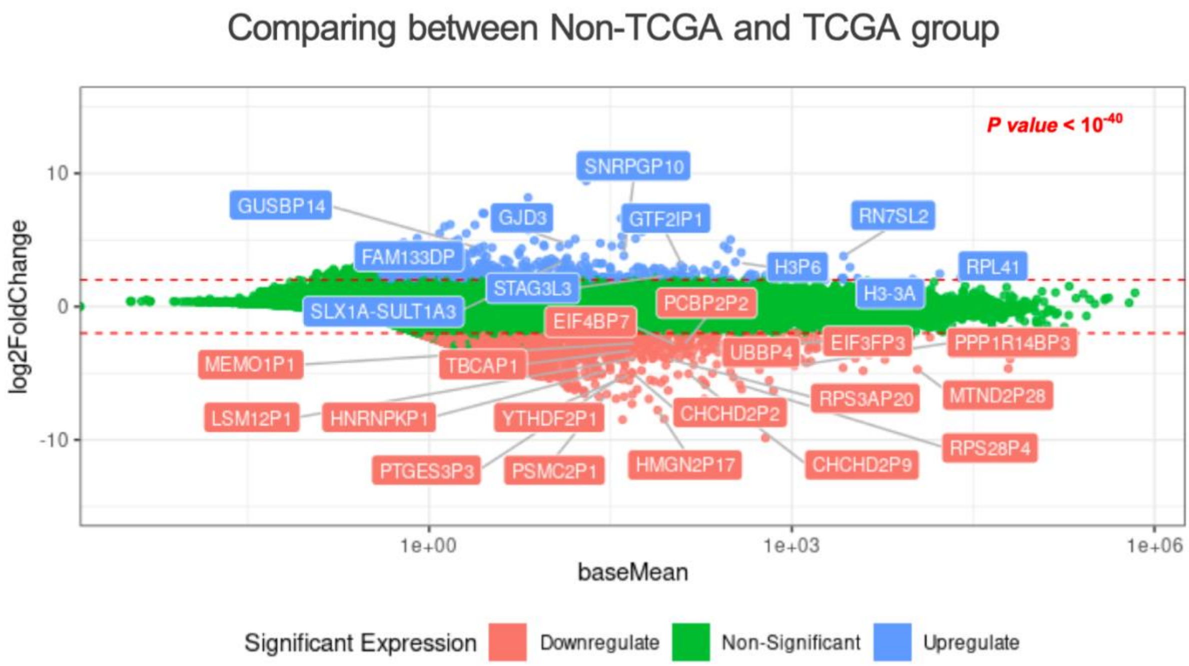


Figure 25. MA plot analysis demonstrated more than a hundred of genes were found to be up-regulated and downregulated in MIBC comparing between non-TCGA and TCGA group

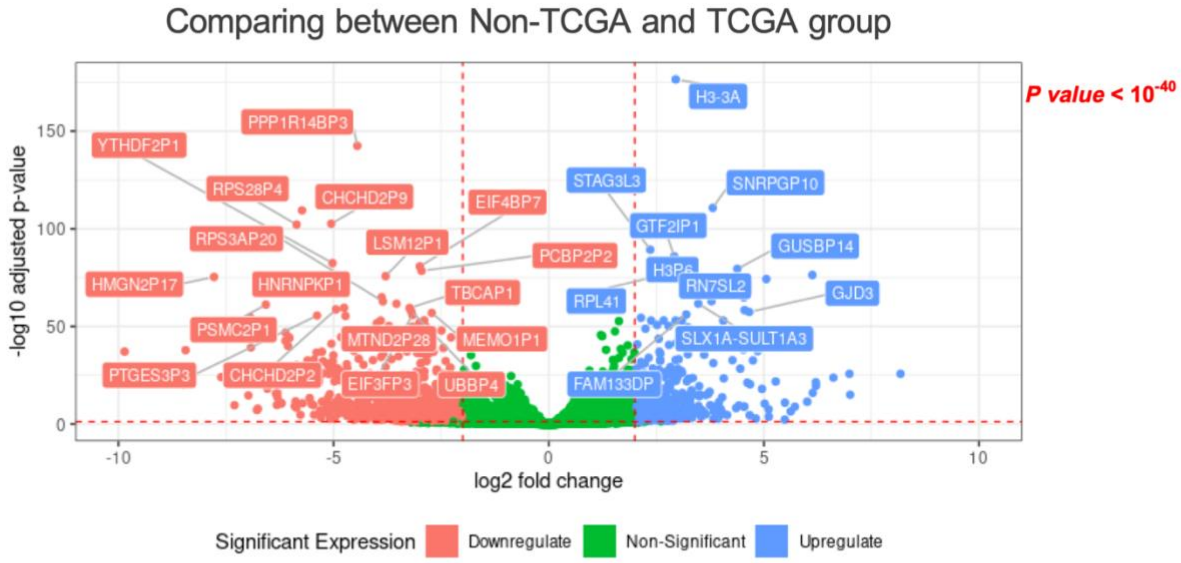


Figure 26. The expression levels of each gene were demonstrated as Volcano plots comparing between non-TCGA and TCGA group

We also determined the relationship between each cluster and survival data. The Kaplan-Meier analysis demonstrated that cluster B displayed the highest probability of survival within 1,800 days of follow-up while cluster C showed the lowest value (*p* = 0.028; Fig. 27).

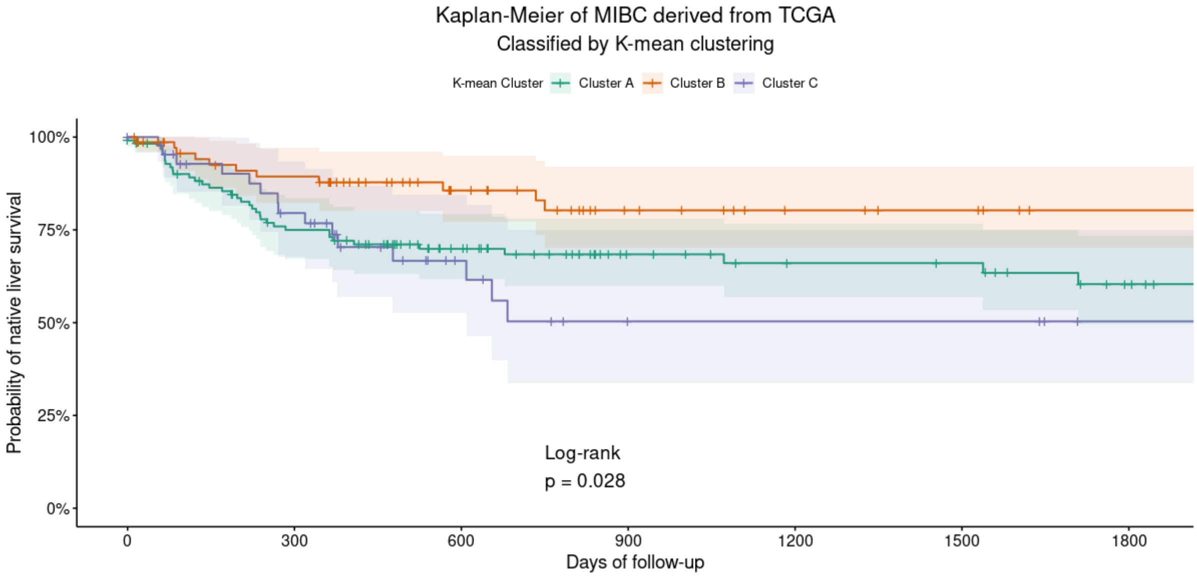


Figure 27. The Kaplan-Meier analysis demonstrated that cluster B displayed the highest probability of survival within 1,800 days of follow-up while cluster C showed the lowest value

However, we found some overlaps of gene expressions which classified by TCGA that clustering in 3 clusters. (Fig. 28)

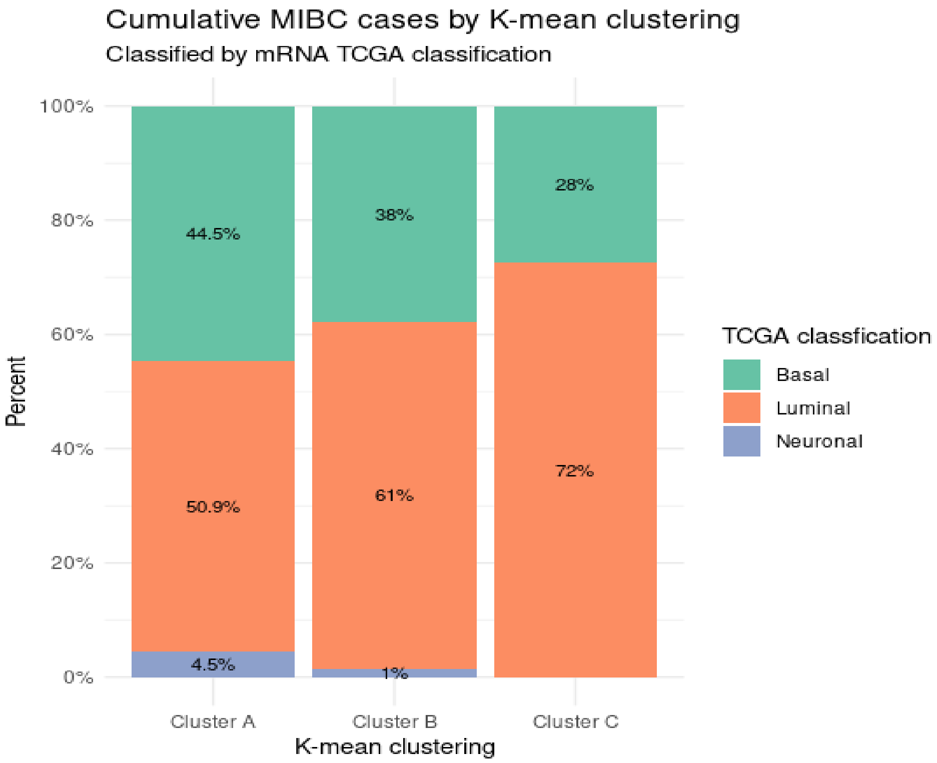


Figure 28. Cumulative MIBC cases by K-mean clustering that classified by mRNA with TCGA classification

3.1.8 Identification of differentially expressed genes of each cluster of MIBC

We compared the differentially expressed genes in the 3 subtypes and found that 15 differentially expressed genes were unique to the cluster A, 78 were unique to cluster B, and 106 were unique to cluster B muscle invasive bladder cancer samples using Filtered by log2normalized count > 15. (Fig. 29 and Table 4)

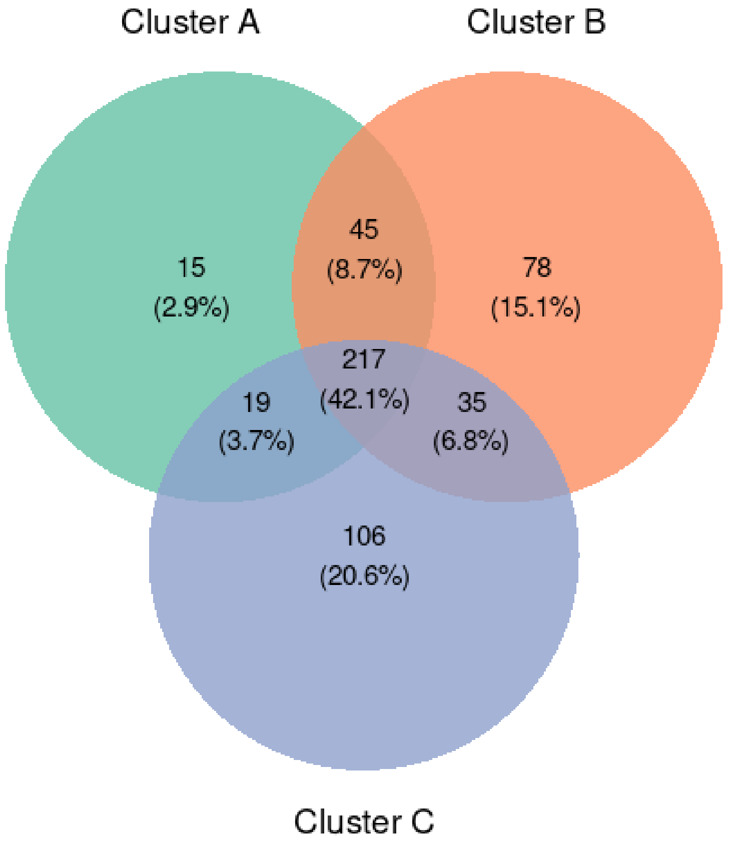


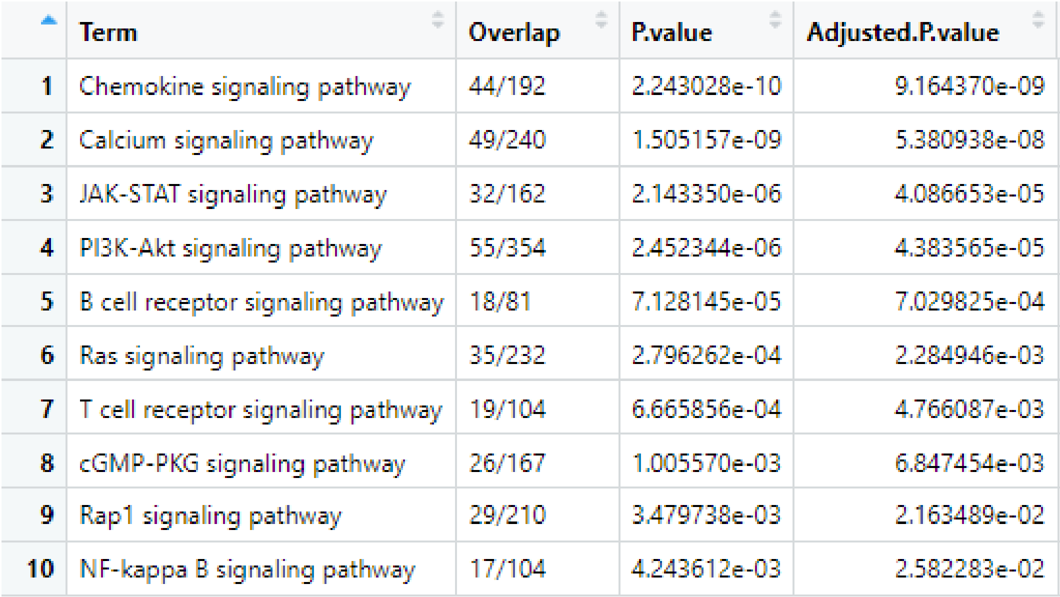
Figure 29. Unique and common enriched genes of each cluster

Table 4. Unique significant differential expression

Cluster	Genes
A	CPSF7,ALYREF,PABPC1L,ITGA2,EFNA1,KRT23,KRT16,F3,CCL5,SIK1B,G6PD,IRF1,S100A8,CD36,ACSL1
B	PPP2R2A,HK2,IGF1R,EP300,VHL,CAMK2G,EGLN3,SLC9A1,TSC2,NCOR1,TP53,ATF6B,KRT20,CREB3L2,PRKCD,HSPA1A,IRS1,CRKL,MAP3K5,RBL2,COL4A6,COL4A5,MET,IFNAR1,PRKAA1,GNG5,COL9A2,GNA15,MMP1,CAB39L,ADIPOR1,ELAVL1,ULK1,PATJ,TEAD3,FZD6,FRMD6,CAT,PGM2,PGLS,XIAP,NLRP1,ERBIN,EHMT2,SOD1,EIF4EBP2,TELO2,GNA11,GTF2I,EML4,VAV2,ELK4,MECOM,DUSP4,DUSP2,TAOK2,NF1,SGPL1,SPTLC2,ACER2,CCNG1,DBI,ACSL3,FABP4,SCP2,FABP5,ACOX1,UBC,PLIN2,HMGCS2,ACAA1,PLTP,SKIL,ZFHX3,PCGF3,TBX3,ID3,TCF3
C	ITGB5,TNC,FGF7,MYC,PDGFRB,VWF,ITGA1,COL6A3,ITGA7,ITGA5,CSF1R,LAMA4,THBS2,PPP2CB,FGFR1,CASC3,ICAM1,CCL2,STAT5B,CYBB,SELE,IL1B,BAX,PFKFB3,HMOX1,NOTCH2,PLN,PRKACA,RCAN1,ROCK1,ROCK2,SRF,MYLK,RGS2,MEF2D,PPP1R12A,IRAG1,GNAQ,MYL9,MMP9,HBEGF,STMN1,MAP2K3,DUSP5,DUSP3,IL1R1,CACNA1H,RAP1A,MAP3K20,CD14,FLNC,MAP4K4,GADD45B,GADD45A,NFKB2,ABL1,RAPGEF1,ITGB2,ENAH,PPP1R12B,ACTA2,SORBS1,RHOQ,CCN2,WWTR1,KLF2,APLNR,TNFAIP3,CXCL2,JUNB,DAB2IP,CFLAR,TNFRSF1B,BCL3,ANTXR2,ANTXR1,FOSL1,EGR3,LSP1,CXCL14,GRK2,CXCL12,UBE2I,PARP1,PLAU,RAB5B,RAB5C,ETS1,ETS2,RALBP1,PLA2G2A,RGL2,ARF6,DEGS1,LTBP1,DCN,GREM1,NBL1,SKP1,FBN1,CAB39,RBPJ,FHL1,STAT6,IL13RA1,IL6ST

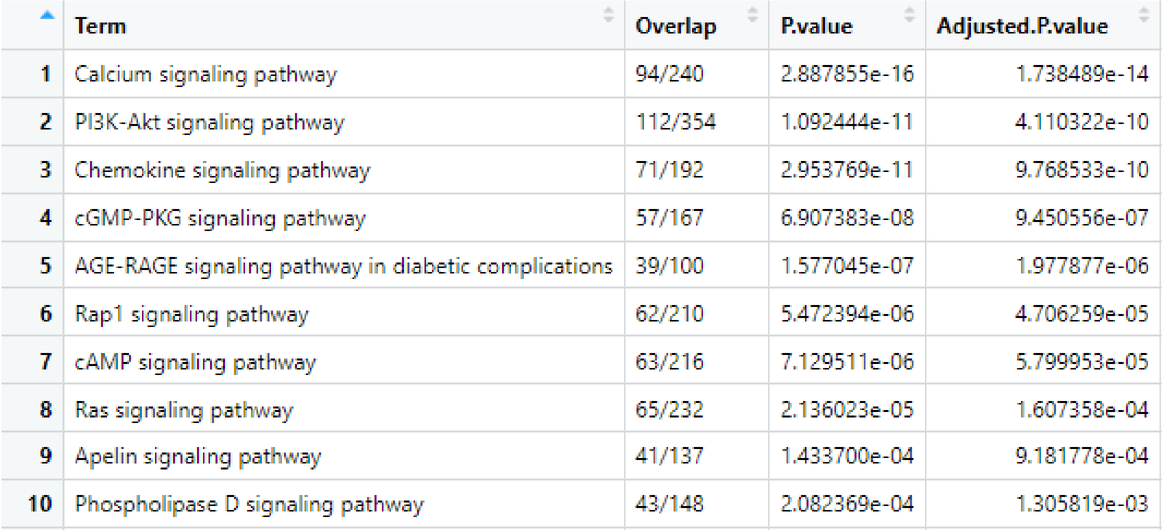
3.1.9 The certain signaling pathways were associated with each type of MIBC cluster

To identify the signaling pathways enriched in each cluster, the transcriptomic data was used with KEGG (Kyoto Encyclopedia of Genes and Genomes) term enrichment analysis. The metabolic pathways or signal transduction pathways associated with differentially expressed genes, comparing the whole genome background with the KEGG terms and padj < 0.05 are justified as significant enrichment. The top 10 significantly enriched terms in the KEGG enrichment analysis are displayed for each cluster (Fig. 30-32).



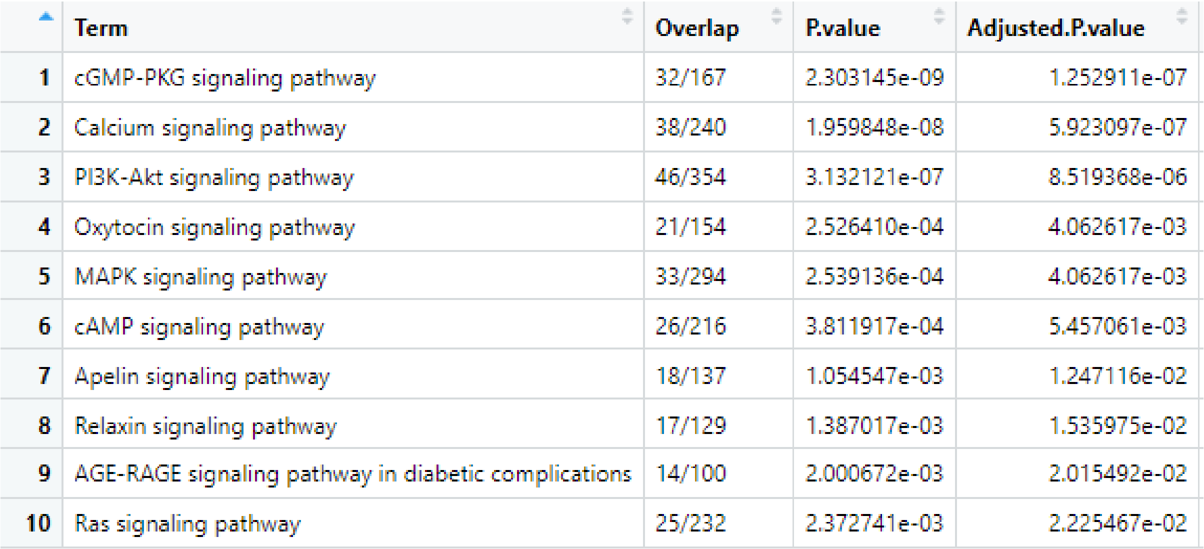
	Term	Overlap	P.value	Adjusted.P.value
1	Chemokine signaling pathway	44/192	2.243028e-10	9.164370e-09
2	Calcium signaling pathway	49/240	1.505157e-09	5.380938e-08
3	JAK-STAT signaling pathway	32/162	2.143350e-06	4.086653e-05
4	PI3K-Akt signaling pathway	55/354	2.452344e-06	4.383565e-05
5	B cell receptor signaling pathway	18/81	7.128145e-05	7.029825e-04
6	Ras signaling pathway	35/232	2.796262e-04	2.284946e-03
7	T cell receptor signaling pathway	19/104	6.665856e-04	4.766087e-03
8	cGMP-PKG signaling pathway	26/167	1.005570e-03	6.847454e-03
9	Rap1 signaling pathway	29/210	3.479738e-03	2.163489e-02
10	NF-kappa B signaling pathway	17/104	4.243612e-03	2.582283e-02

Figure 30. The top 10 significantly enriched terms in the KEGG enrichment analysis are displayed for cluster A



	Term	Overlap	P.value	Adjusted.P.value
1	Calcium signaling pathway	94/240	2.887855e-16	1.738489e-14
2	PI3K-Akt signaling pathway	112/354	1.092444e-11	4.110322e-10
3	Chemokine signaling pathway	71/192	2.953769e-11	9.768533e-10
4	cGMP-PKG signaling pathway	57/167	6.907383e-08	9.450556e-07
5	AGE-RAGE signaling pathway in diabetic complications	39/100	1.577045e-07	1.977877e-06
6	Rap1 signaling pathway	62/210	5.472394e-06	4.706259e-05
7	cAMP signaling pathway	63/216	7.129511e-06	5.799953e-05
8	Ras signaling pathway	65/232	2.136023e-05	1.607358e-04
9	Apelin signaling pathway	41/137	1.433700e-04	9.181778e-04
10	Phospholipase D signaling pathway	43/148	2.082369e-04	1.305819e-03

Figure 31. The top 10 significantly enriched terms in the KEGG enrichment analysis are displayed for cluster B



	Term	Overlap	P.value	Adjusted.P.value
1	cGMP-PKG signaling pathway	32/167	2.303145e-09	1.252911e-07
2	Calcium signaling pathway	38/240	1.959848e-08	5.923097e-07
3	PI3K-Akt signaling pathway	46/354	3.132121e-07	8.519368e-06
4	Oxytocin signaling pathway	21/154	2.526410e-04	4.062617e-03
5	MAPK signaling pathway	33/294	2.539136e-04	4.062617e-03
6	cAMP signaling pathway	26/216	3.811917e-04	5.457061e-03
7	Apelin signaling pathway	18/137	1.054547e-03	1.247116e-02
8	Relaxin signaling pathway	17/129	1.387017e-03	1.535975e-02
9	AGE-RAGE signaling pathway in diabetic complications	14/100	2.000672e-03	2.015492e-02
10	Ras signaling pathway	25/232	2.372741e-03	2.225467e-02

Figure 32. The top 10 significantly enriched terms in the KEGG enrichment analysis are displayed for cluster C

The calcium signaling pathway was found to be a general significant process in all clusters. However, the chemokine signaling pathway was observed significantly only in clusters 1 and 2. Interestingly, the immune signal transduction pathways, including JAK-STAT, B cell receptor signaling, and T cell receptor signaling pathways, were marked to be the key mechanism in cluster A of MIBC specifically, while AGE-RAGE and Rap1 signaling pathways were found as significantly enriched molecular processes in cluster B. For cluster C, cGMP-PKG, oxytocin, MAPK, and Relaxin signaling pathways were observed to be unique in this cluster. GO Enrichment analyses of the differentially expressed genes in each cluster (Fig. 33-35)

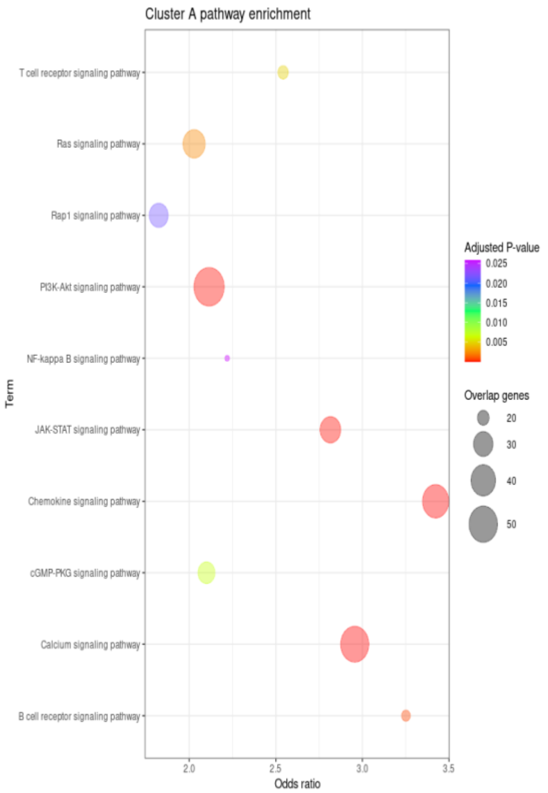


Figure 33. GO Enrichment analyses of the differentially expressed genes in cluster A

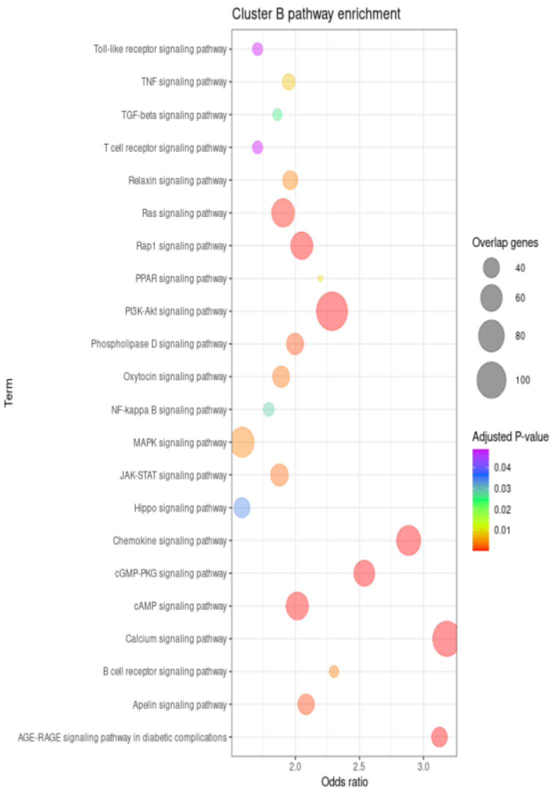


Figure 34. GO Enrichment analyses of the differentially expressed genes in cluster B

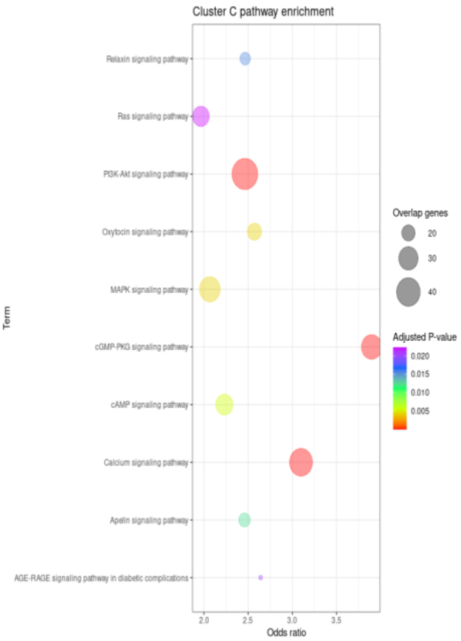


Figure 35. GO Enrichment analyses of the differentially expressed genes in cluster C

Number of differential gene expressions among different cluster after achieved by pathway enrichment were shown. (Fig. 36)

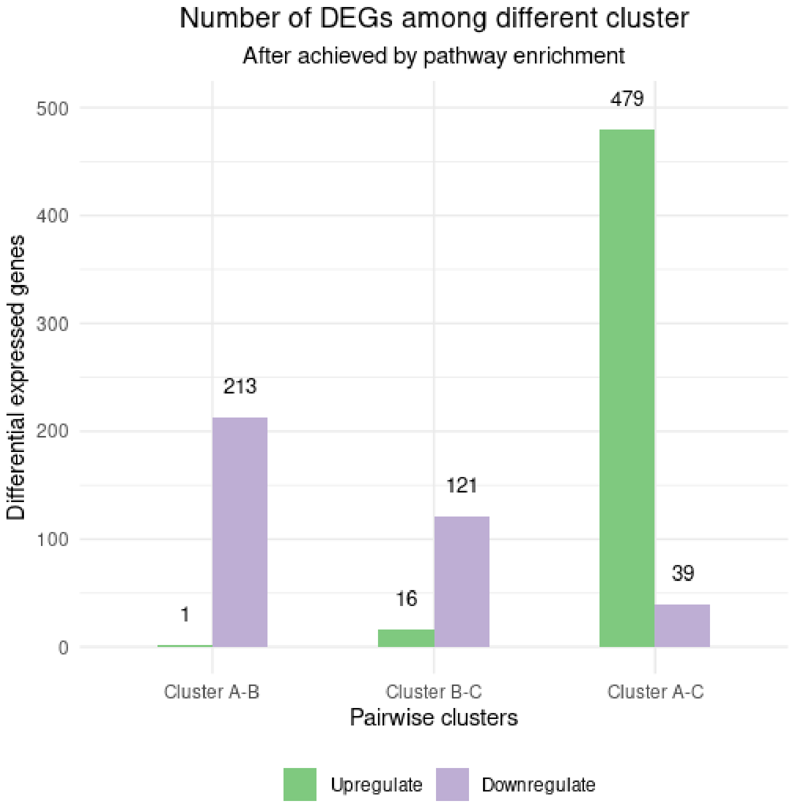


Figure 36. Number of differential gene expressions among different cluster after achieved by pathway enrichment

3.2 Results: Part II

3.2.1 Demographic and clinicopathological data

This study included 132 patients with MIBC who underwent radical cystectomy during the study period. Their mean age was 65.6 years, and the male to female ratio was 6.8:1. The demographic characteristics of the patients and immunoreactivity for each IHC marker are summarized in Table 5.

Table 5. The demographic characteristics of the patients and immunoreactivity for each IHC marker

Variable		Value (N;%)	5-year OS (95% CI)*	Log-rank p-value
Age (year)				
	Mean(SD)	65.6 (9.3)	-	
Sex				0.56
	Male	115 (87.1%)	25.7 (17.9-34.2)	
	Female	17 (12.9%)	35.3 (14.5-57.0)	
ECOG status				0.18
	0	29 (22.0%)	39.2 (21.6-56.5)	
	1	103 (78.0%)	23.7 (15.9-32.6)	
T stage				< 0.01
	T1/2	44 (33.3%)	57.4 (40.8-70.9)	
	T3	41 (31.1%)	18.0 (7.9-31.3)	
	T4	47 (35.6%)	6.4 (1.7-15.8)	
N stage				
	N0	90 (68.2%)	33.7 (23.8-43.8)	< 0.01
	N1	24 (18.2%)	12.2 (4.5-24.1)	
	N2	15 (11.4%)		
	N3	3 (2.3%)		
M stage				NA
	M0	128 (97.7%)	26.9 (19.4-34.9)	
	M1	3 (2.3%)	NA	
Tumor grade				0.04
	Low	7 (5.30%)	85.7 (33.4-97.9)	
	High	125 (94.7%)	23.7 (16.5-31.7)	
Chemotherapy				0.80
	No	106 (80.3%)	27.9 (19.5-37.0)	
	Yes	26 (19.7%)	23.1 (9.4-40.3)	
Diversion				0.03
	ileal conduit	123 (93.2%)	24.0 (16.6-32.1)	
	neobladder	9 (6.8%)	66.7 (28.2-87.8)	
LVI				< 0.01
	Negative	52 (39.4%)	40.0 (26.3-53.3)	

	Positive	80 (60.6%)	18.8 (11.1-28.2)	
CK20				0.45
	Negative	76 (57.6%)	25.3 (16.2-35.5)	
	Positive	56 (42.4%)	29.7 (18.0-42.4)	
CK5/6				0.02
	Negative	65 (49.2%)	16.2 (8.3-26.5)	
	Positive	67 (50.8%)	37.8 (26.2-49.3)	
CK14				0.63
	Negative	95 (72.0%)	26.0 (17.5-35.4)	
	Positive	37 (28.0%)	30.6 (16.6-45.7)	
GATA3				< 0.01
	Negative	26 (19.7%)	16.1 (5.9-30.9)	
	Positive	106 (80.3%)	30.5 (21.6-39.9)	

*: excluding 2 operative deaths; 5-Year OS (95% CI): 5-year overall survival (95% confidence interval); ECOG status: Eastern Cooperative Oncology Group performance status

Two patients who died at 3 and 7 days after surgery were considered to have operative mortality and were excluded from the survival analysis. As of January 2021, the median follow-up duration was 125 months (interquartile range 103–154 months). The median OS time was 12.2 months (interquartile range 4.7, 46.4 months), and the 5-year OS was 27.0% (95% CI 19.6%–35.0%). IHC showed positivity for GATA3, CK5/6, CK20, and CK14 with kappa value between 0.799–0.908 (93.2–96.2% agreement) (Table 6).

Table 6. Immunopositivity of the 4 markers studied and their correlation with clinic-pathological parameters

	All	GATA3	CK5/6	CK20	CK14
Positive staining (%)	132	101 (76.5%)	67 (50.8%)	56 (42.4%)	37 (28.0%)
Mean age in positive cases (SD)		64.7 (9.3)	64.7 (9.1)	64.8 (8.7)	65.7 (9.3)
Gender					
Male (%)	115 (87.1%)	91 (90.1%)	56 (83.6%)	50 (89.3%)	30 (81.1%)
Female (%)	17 (12.9%)	10 (9.9%)	11 (16.4%)	6 (10.7%)	7 (18.9%)
ECOG status, n (%)					
0	29 (22.0%)	25 (24.7%)	13 (19.4%)	13 (23.1%)	8 (21.6%)
1	103 (78.0%)	76 (75.3%)	54 (80.6%)	43(76.8%)	29 (78.3%)
T stage, n (%)					
pT1	21 (15.9%)	20 (19.8%)*	12 (17.9%)	14 (25.0%)*	0 (0.0%)*

pT2	23 (17.4%)	20 (19.8%)	14 (20.9%)	13 (23.2%)	8 (21.6%)
pT3	41 (31.1%)	28 (27.7%)	20 (29.9%)	10 (17.9%)	18 (48.7%)
pT4	47 (35.6%)	33 (32.6%)	21 (31.3%)	19 (33.9%)	11 (29.7%)
N stage, n (%)					
N0	90 (68.2%)	69 (68.3%)	47 (70.2%)	41 (73.2%)	27 (73.0%)
N1	24 (18.2%)	19 (18.8%)	12 (17.9%)	7 (12.5%)	7 (18.9%)
N2	15 (11.4%)	11 (10.9%)	7 (10.5%)	6 (10.7%)	2 (5.4%)
N3	3 (2.3%)	2 (2.0%)	1 (1.5%)	2 (3.6%)	1 (2.3%)
M stage, n (%)					
M0	128 (97.7%)	98 (98.0%)	65 (98.5%)	54 (98.2%)	35 (97.2%)
M1	3 (2.3%)	2 (2.0%)	1 (1.5%)	1 (1.8%)	1 (2.8%)
Tumor grade, n (%)					
Low	7 (5.30%)	6 (5.9%)	6 (9.0%)	3 (5.4%)	0 (0.0%)
High	125 (94.7%)	95 (94.1%)	61 (91.0%)	53 (94.6%)	37 (100.0%)
LVI, n (%)					
Neg	52 (39.4%)	41 (40.6%)	26 (38.8%)	25 (44.6%)	11 (29.7%)
Pos	80 (60.6%)	60 (59.4%)	41 (61.1%)	31 (55.4%)	26 (70.3%)

*p-value < 0.05 when comparing distribution between positive cases and all cases

The immunostains for GATA3, CK5/6, CK20, and CK14 showed positive results with 80.3%, 50.8%, 42.4%, and 28.0% of cases, respectively. GATA3 and CK5/6 immunopositivity was significantly associated with OS by log-rank analysis (Table 4). Twenty-six cases received a median of 3 cycles of adjuvant chemotherapy. GATA3 expression was significantly inversely correlated with pT stage progression. Some mixed and/or borderline cases, the positive immunostains were interpreted 2 times with the consensus of a pathologist found 5 of 132 cases (3.8%)

According to the Kaplan–Meier survival analysis, significant differences in outcomes with respect to OS were demonstrated among cases with positive GATA3 staining (*p* = 0.008) and in cases with positive CK5/6 staining (*p* = 0.038). The other markers did not show significant prognostication value for survival. The Kaplan–Meier survival curves for GATA3, CK5/6, CK14, and CK20 are depicted in Fig. 37

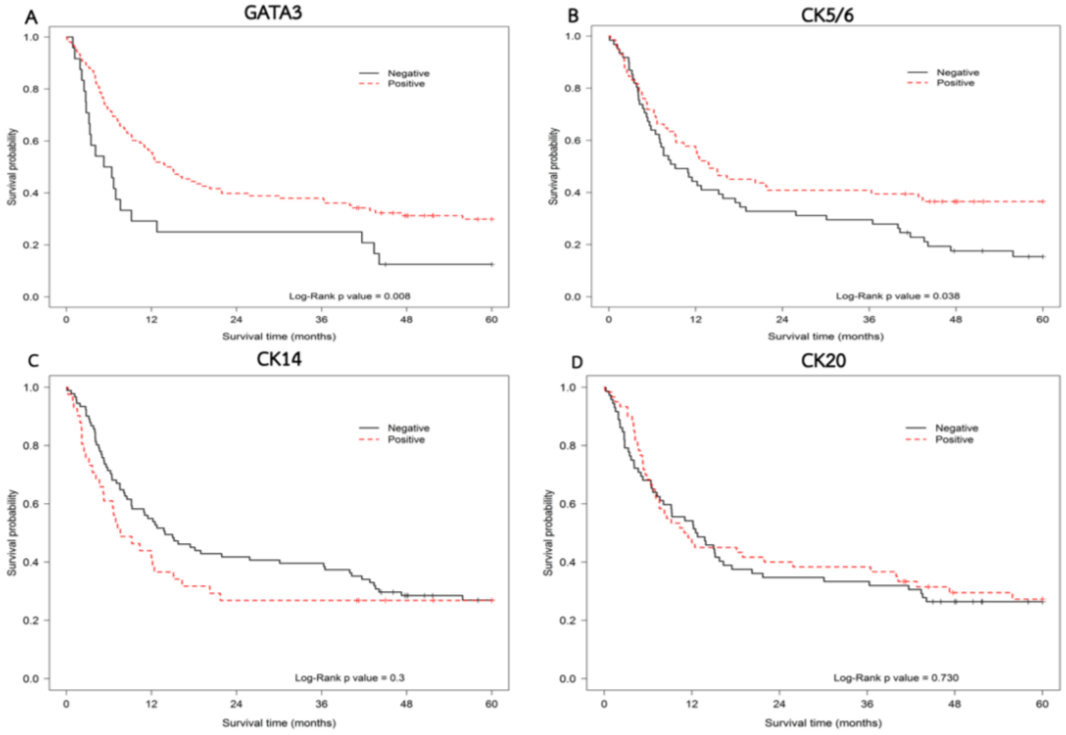


Figure 37. Kaplan-Meier curves demonstrating survival probability in 132 MIBC patients according to IHC markers expression; GATA3 (A), CK5/6 (B), CK14 (C) and CK 20 (D).

The correlation between each individual marker was evaluated by Pearson correlation test. The significant association of GATA3, CK5/6, and CK20 was found only in pathological stage 1 of patients. When the correlation between markers in the basal and luminal subtypes was assessed, moderate correlation was observed between GATA3 and CK20 expression, which indicated that the basal-like subtype was demonstrated by Pearson correlation at 0.46 (p = 0.022). The analysis showed small correlation between the luminal-like subtype markers, CK5/6 and CK14; 0.31 (p = 0.048) (Fig. 38).

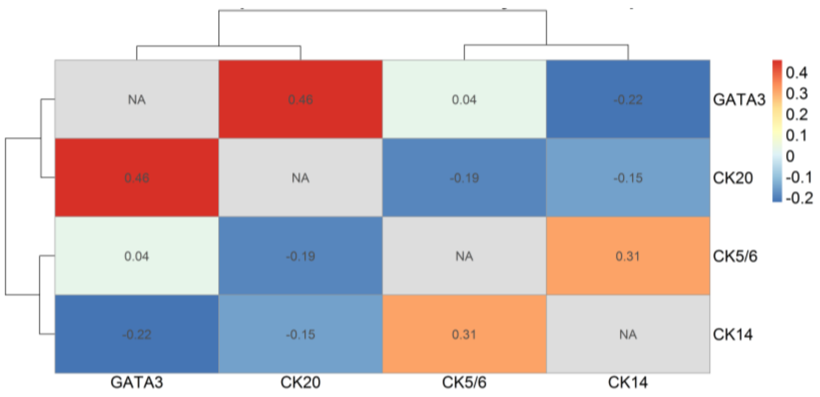


Figure 38. Correlation heatmap of immunohistochemistry markers expression among GATA3, CK20 (basal-like markers), CK5/6 and CK14 (luminal-like markers)

As GATA3 and CK5/6 were the only two markers representing different subtypes that were significantly associated with survival, we elected to categorize our cases into four groups according to those two markers: luminal-like (GATA3$^+$ and CK5/6$^-$), basal-like (GATA3$^-$ and CK5/6$^+$), mixed (GATA3$^+$ and CK5/6$^+$), and double-negative (GATA3$^-$ and CK5/6$^-$) subtypes. By this definition, the luminal-like, basal-like, mixed, and double-negative subtypes were observed in 38.6%, 12.9%, 37.9%, and 10.6% of cases, respectively. Associations between each subtype and clinicopathological factors including survival outcomes are displayed in Table 7. The double-negative subtype was significantly associated with higher incidence of pT4 disease.

Table 7. Patient's characteristic classified by IHC subtypes categorized by GATA3 and CK5/6

	Double-neg	Luminal-like	Basal-like	Mixed	p-value**
Total (%)	14(10.6)	51(38.6)	17(12.9)	50(37.9)	-
Mean age (SD)	70.2(6.0)	65.3(10.1)	66.6(10.7)	64.1(8.6)	0.18
Gender, n (%)					
Male	12 (85.7)	47 (92.2)	12 (70.6)	44 (88.0)	0.15
Female	2 (14.3)	4 (7.8)	5 (29.4)	6 (12.0)	
ECOG status, n (%)					0.29
0	1 (7.1)	15 (29.4)	3 (17.7)	10 (20.0)	

1	13 (92.9)	36 (70.6)	14 (82.3)	40 (80.0)	
Urinary Diversion type					0.98
Ileal conduit	13 (92.9)	47 (92.2)	16 (94.1)	47 (94.0)	
Neobladder	1 (7.1)	4 (7.8)	1 (5.9)	3 (6.0)	
T stage, n (%)					0.048
T1/2	0 (0.0)	18 (35.3)	4 (23.5)	22 (44.0)	
T3	6 (42.9)	15 (29.4)	7 (41.2)	13 (26.0)	
T4	8 (57.1)	18 (35.3)	6 (35.3)	15 (30.0)	
N stage, n (%)					0.71
N0	11 (78.6)	32 (62.8)	10 (58.8)	37 (74.0)	
N1	1 (7.1)	11 (21.6)	4 (23.5)	8 (16.0)	
N2	1 (7.1)	7 (13.7)	3 (17.7)	4 (8.0)	
N3	1 (7.1)	1 (2.0)	0 (0.0)	1 (2.0)	
M stage, n (%)					0.55
M0	13 (92.9)	50 (98.0)	17 (100.0)	48 (98.0)	
M1	1 (7.1)	1 (2.0)	0 (0.0)	1 (2.0)	
Tumor grade, n (%)					0.25
Low	0 (0.0)	1 (2.0)	1 (5.9)	5 (10.0)	
High	14 (100.0)	50 (98.0)	16 (94.1)	45 (90.0)	
Margin, n (%)					0.57
Neg	11 (78.6)	46 (90.2)	16 (94.1)	43 (86.0)	
Pos	3 (21.4)	5 (9.8)	1 (5.9)	7 (14.0)	
LVI, n (%)					0.97
Neg	5 (35.7)	21 (41.2)	6 (35.3)	20 (40.0)	
Pos	9 (64.3)	30 (58.8)	11 (64.7)	30 (60.0)	
CK20, n (%)					<0.01
Neg	12 (85.7)	17 (33.3)	16 (94.1)	31 (62.0)	
Pos	2 (14.3)	34 (66.7)	1 (5.9)	19 (38.0)	
CK14, n (%)					<0.01
Neg	12 (85.7)	48 (94.1)	4 (23.5)	31 (62.0)	
Pos	2 (14.3)	3 (5.9)	13 (76.5)	19 (38.0)	
5-year OS (%) (95% confidence interval)	7.14 (0.4-27.5)	18.9 (9.2-31.1)	23.5 (7.3-24.9)	42.8 (28.9-56.1)	<0.01

*: p-value by chi-square or Fisher's exact test, ECOG status: Eastern Cooperative Oncology Group performance status; LVI: Lymph Vascular Invasion; OS: overall survival

CK20 immunopositivity, a marker of the luminal molecular subtype, was significantly associated with the GATA3-defined luminal subtype ($p < 0.01$), whereas CK14 positivity was significantly associated with the CK5/6-defined basal subtype ($p < 0.01$). In the 50 mixed subtype (GATA3$^+$ and CK2/5$^+$) cases in this study, an equal number of cases with CK20 and CK5/6 positivity was found (Table 7).

When clinicopathological parameters and IHC subtypes were analyzed against survival in a univariable Cox hazard model, tumor stage (pT and N), lymphovascular invasion, pathologic grade, loss of GATA3 immunoreactivity, and loss of CK5/6 immunoreactivity were significantly associated with poorer survival outcomes. Considering subtyping, while patients with the mixed subtype had the lowest risk, which was followed by patients with the luminal-like and basal-like subtypes, those with the double-negative subtype had the highest crude HR. In the multivariable analysis by stepwise Cox hazard regression, N stage (N > 0) and the double-negative subtype were significantly associated with higher risk (model $p = 0.0001$) (Table 8).

Table 8. Univariable and multivariable regression analysis of clinical outcomes in 132 MIBC patients

		Univariable analysis		Multivariable analysis	
Factor		crude HR (95%CI)	p-value	adj. HR (95%CI)	p-value
T-stage:	pT1	1.00 (reference)	< 0.01		
	pT2	1.61 (0.65, 3.95)			
	pT3	5.15 (2.36, 11.19)			
	pT4	6.50 (3.01, 14.02)			
N-stage:	N0	1.00 (reference)	< 0.01	1.00 (reference)	
	N1	1.78 (1.07, 2.96)		1.84 (1.09-3.13)	0.02
	N2	2.58 (1.43, 4.66)		2.63 (1.44-4.78)	<0.01
	N3	5.34 (1.63, 16.58)		4.45 (1.35-14.68)	0.01

LVI (positive)	1.94 (1.27, 2.96)	< 0.01		
Grading (high grade)	3.16 (1.00, 10.00)	0.02		
GATA3 (negative)	1.87 (1.20-2.90)	< 0.01		
CK5/6 (negative)	1.57 (1.06-1.35)	0.03		
CK20 (negative)	1.16 (0.78-1.75)	0.45		
CK14 (negative)	0.89 (0.58-1.39)	0.63		
Mixed subtype	0.52 (0.34-0.81)	< 0.01		
Basal subtype	1.39 (0.79-2.46)	0.25		
Luminal subtype	1.18 (0.79-1.76)	0.43		
Double negative	2.24 (1.27-3.96)	< 0.01		
Subtypes		< 0.01		
Mixed	1 (reference)		(reference)	1
Luminal-like	1.66 (1.03-2.68)		1.66 (0.86-3.21)	0.13
Basal-like	2.01 (1.06-3.81)		1.60 (0.99-2.60)	0.05
Double negative	3.12 (1.63-5.92)		3.29 (1.71-6.31)	< 0.01

crude HR: crude hazard ratio; adj.HR: adjusted hazard ratio; 95%CI: 95% confidence interval, LN: lymph node, LVI: lymphovascular invasion

3.2.2 Association between molecular subtypes and survival outcomes

Kaplan–Meier curves compare the survival probability of 132 patients with MIBC following radical surgery (Fig.39). The 5-year OS rates of patients with the mixed, basal, luminal, and double-negative subtypes were 42.5% (95% CI 28.9–56.1%), 23.5% (7.3–44.9%), 18.9% (9.2–31.1%), and 7.1% (0.4–27.5%), respectively.

In the 50 MIBCs of mixed subtype (GATA3^{+} and CK5/6^{+}), if CK20 and CK14 were added to the subcategorization criteria, which means that mixed subtype cases with positive CK20 were reclassified as luminal, whereas mixed subtype cases with positive CK14 were reclassified as basal. The univariable hazard model did not improve. Using the mixed subtype as a reference, the HRs (and 95% CIs) for the luminal, basal, and double-negative subtypes were 1.17 (95% CI 0.62–2.20), 1.54 (0.76–3.10), and 2.64 (1.22–5.73) respectively.

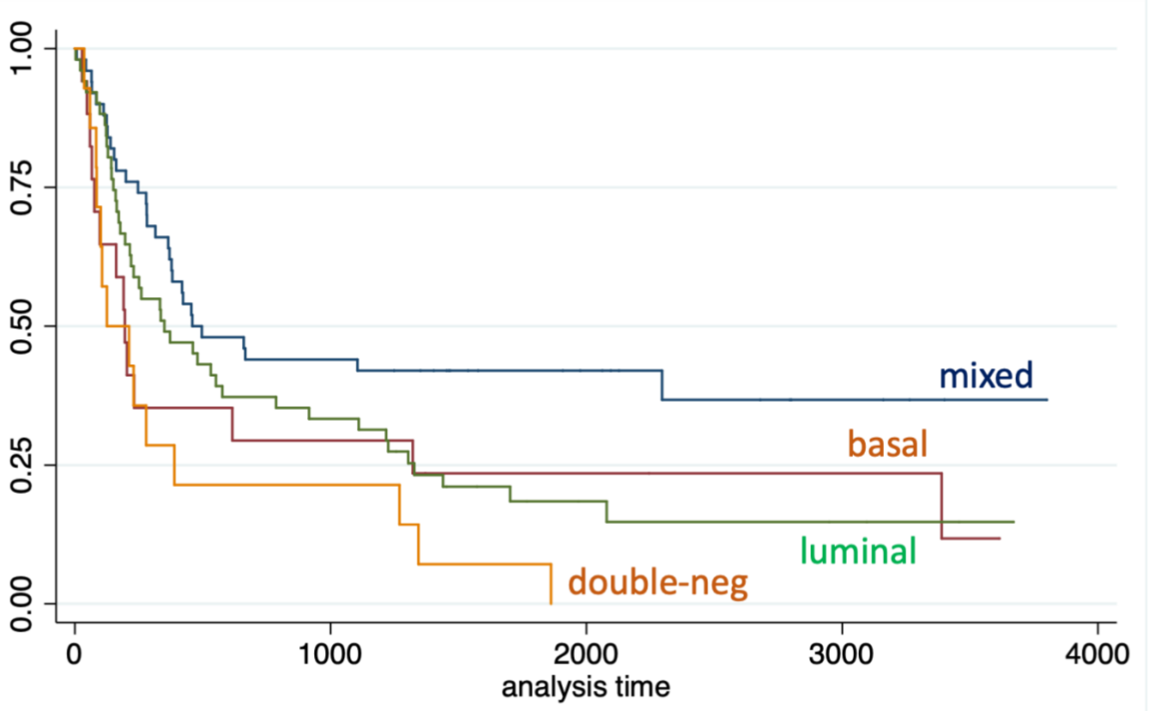


Figure 39. Kaplan-Meier survival curves of molecular subtypes of muscle invasive bladder cancer according to their immunohistological subtypes

3.3 Results: Part III

3.3.1 Integrative result of molecular subtyping from mRNA expressions using 4 markers from immunohistochemistry in our MIBC cohort

Each gene expressions were selected from IHC part (GATA3, CK 5/6, CK14 and CK 20) and compared between cluster from K-mean clustering in 3 cluster A-C. Furthermore, our interesting 4 markers mRNA expression were shown in violin plot. (Fig.40-43)

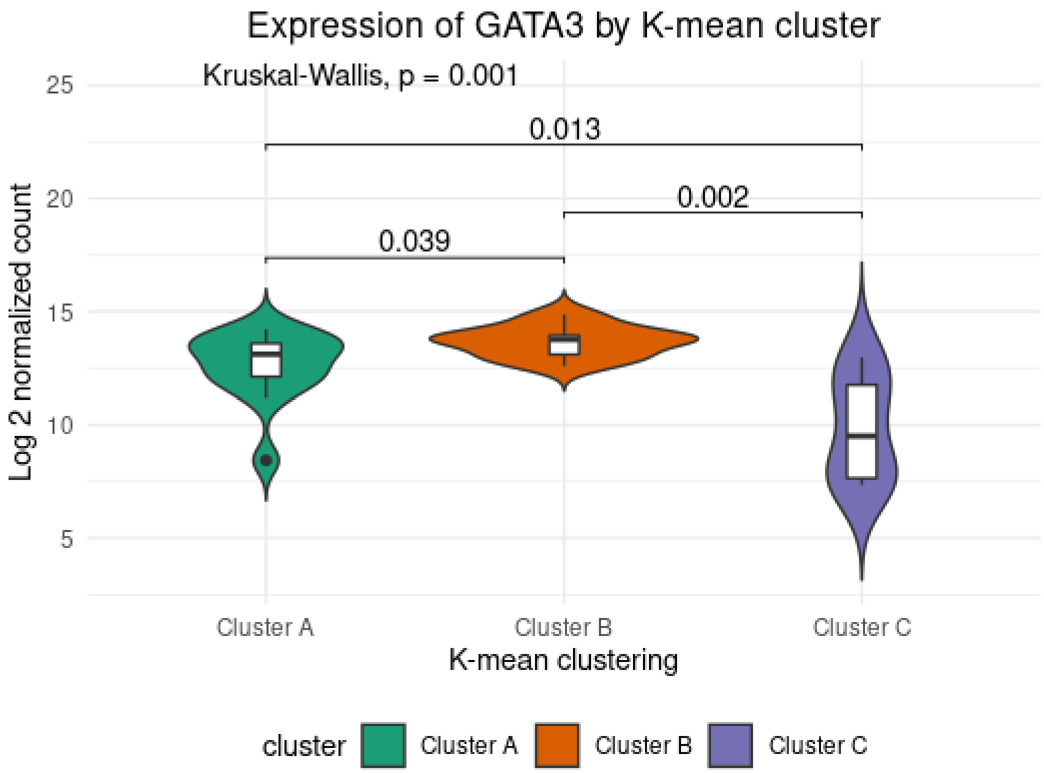


Figure 40. Violin plot of GATA3 expression in 3 clusters

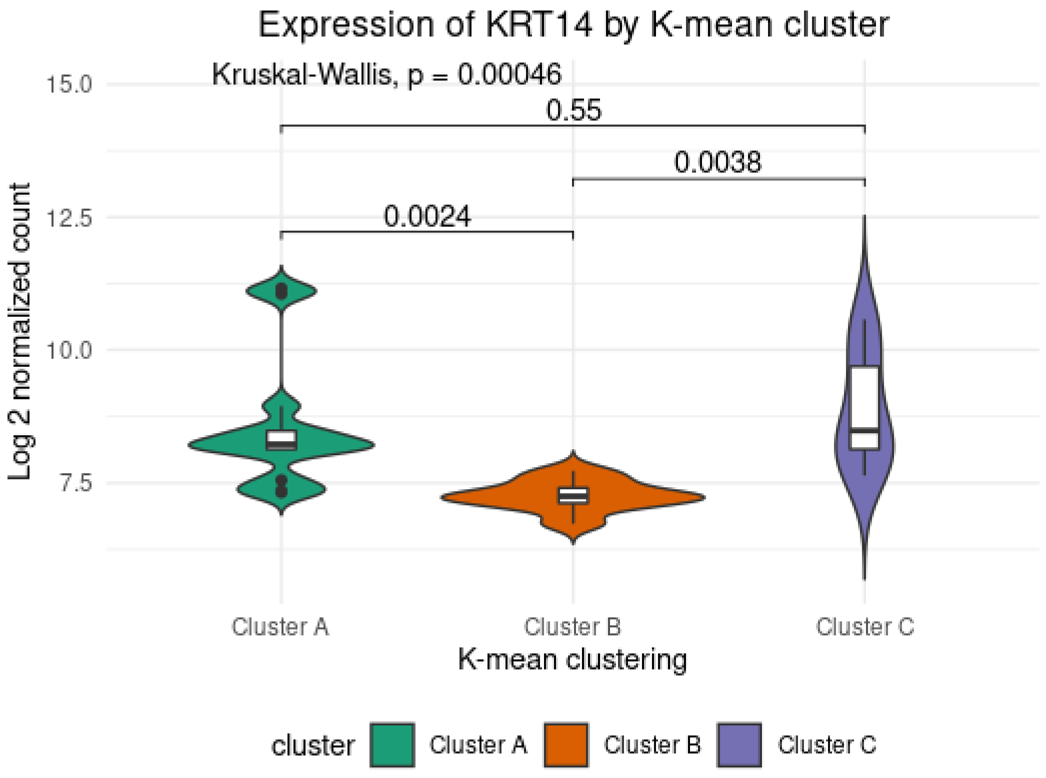


Figure 41. Violin plot of CK14 expression in 3 clusters

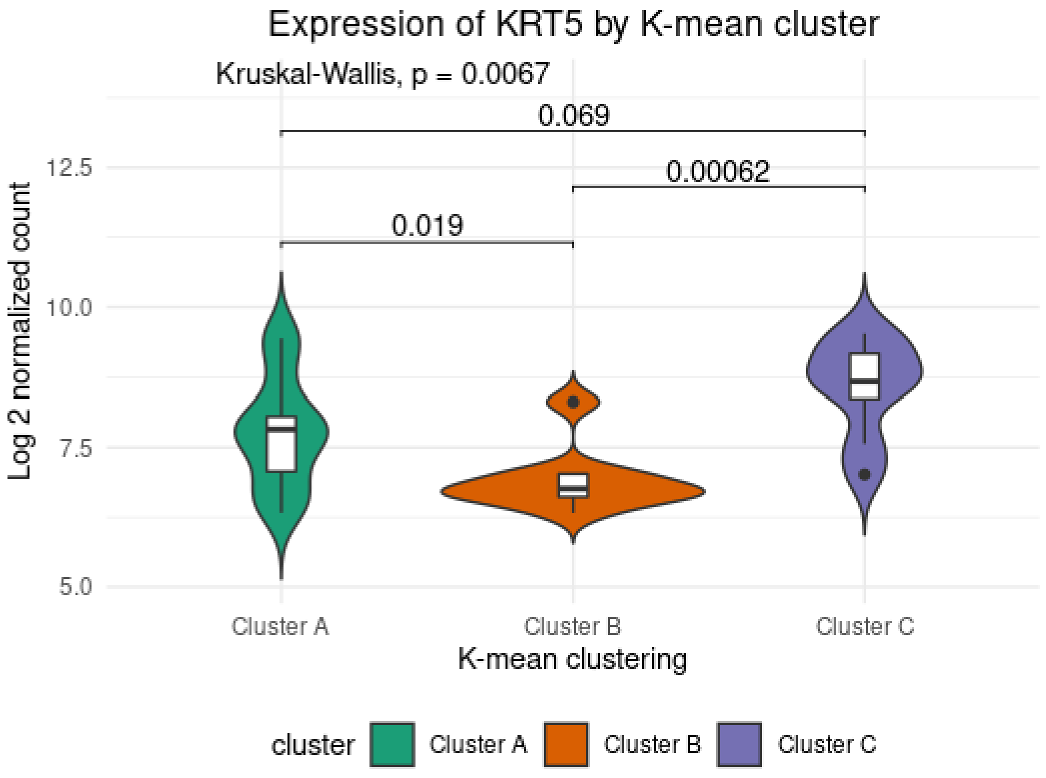


Figure 42. Violin plot of CK5/6 expression in 3 clusters

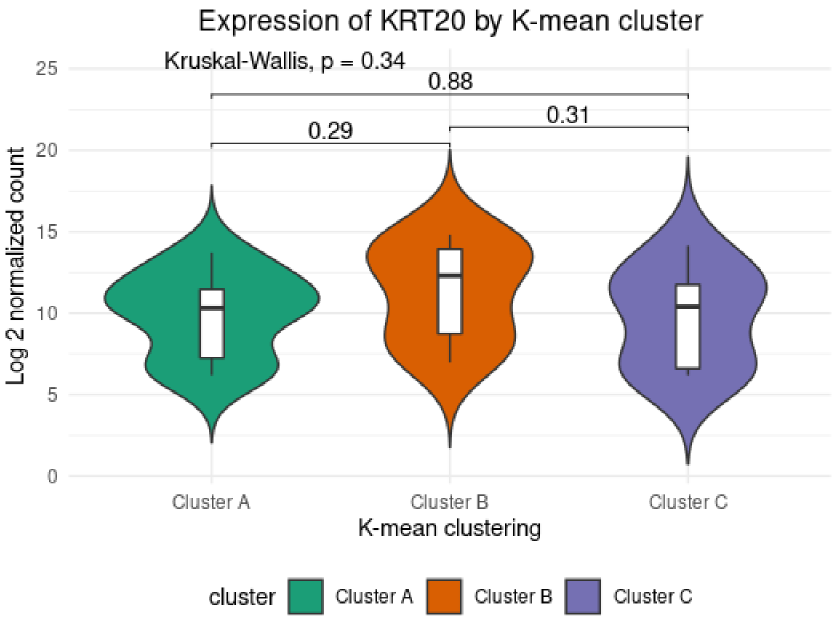


Figure 43. Violin plot of CK20 expression in 3 cluster

3.3.2 Validate each signature gene expression in TCGA data

Gene expression level of signature markers in our study were validated in TCGA data that exhibited the same fashion with our cohort (Fig. 44-47)

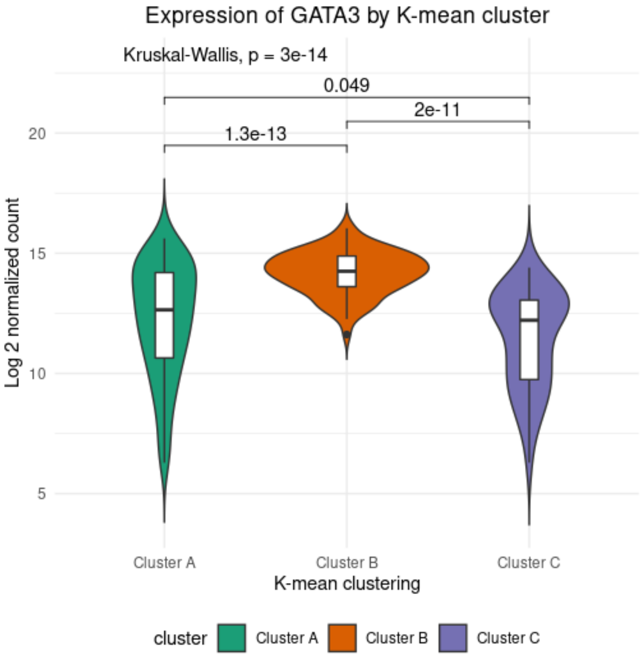


Figure 44. Violin plot of GATA3 expression in 3 cluster from TCGA data

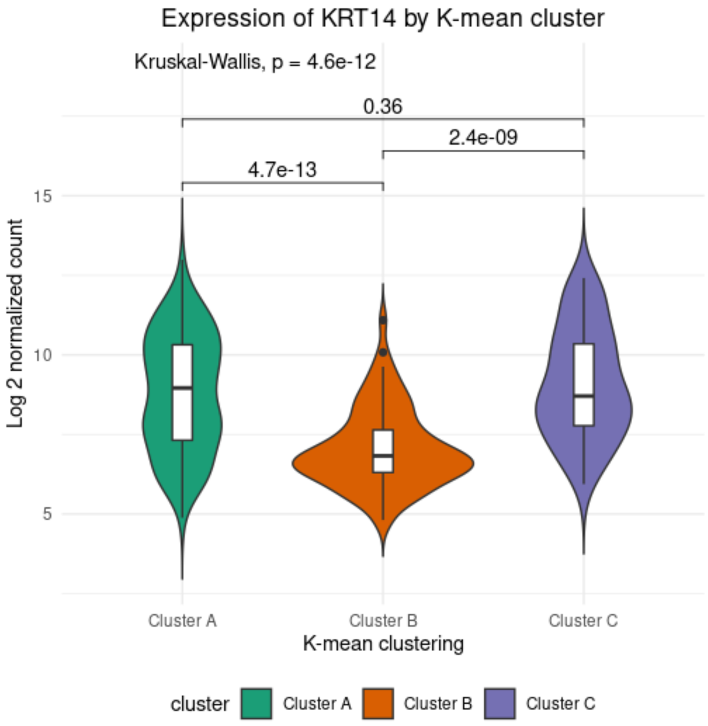


Figure 45. Violin plot of CK 14 expression in 3 cluster from TCGA data

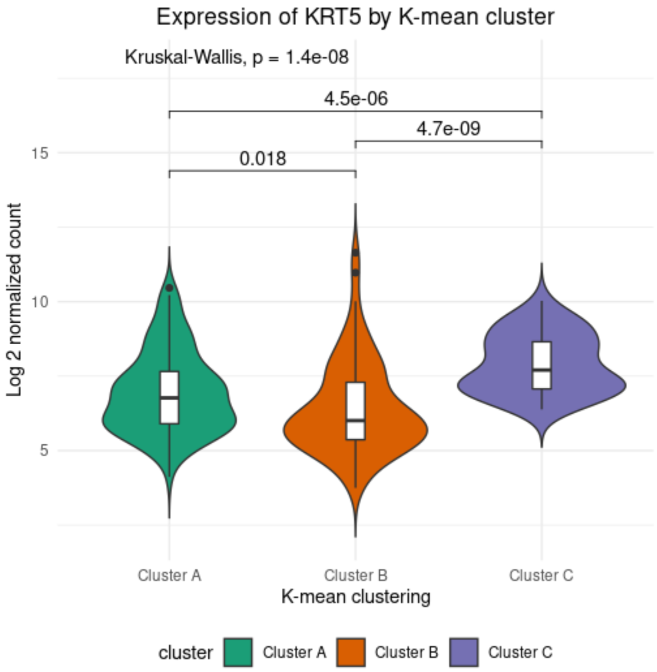


Figure 46. Violin plot of CK 5/6 expression in 3 cluster from TCGA data

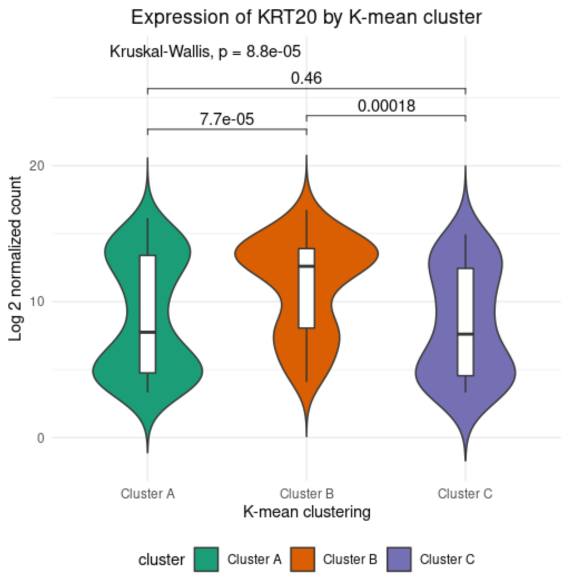


Figure 47. Violin plot of CK 20 expression in 3 cluster from TCGA data

3.3.3 AUC score of selected signature genes of molecular subtypes

We explored the significant genes that classified each clusters and AUC score were shown in Table 9 and sensitivity, specificity and cut point of selected signature genes. (Fig. 48)

Table 9 AUC score of selected signature genes classified clusters

	Cluster A-B	Cluster B-C	Cluster A-C
GATA3	**0.763**	**0.916**	**0.614**
CD274	**0.722**	**0.623**	**0.590**
SNCA	**0.640**	**0.608**	**0.547**
KRT20	**0.658**	**0.708**	**0.546**
KRT5	**0.627**	**0.813**	**0.723**
KRT14	**0.776**	**0.835**	**0.458**

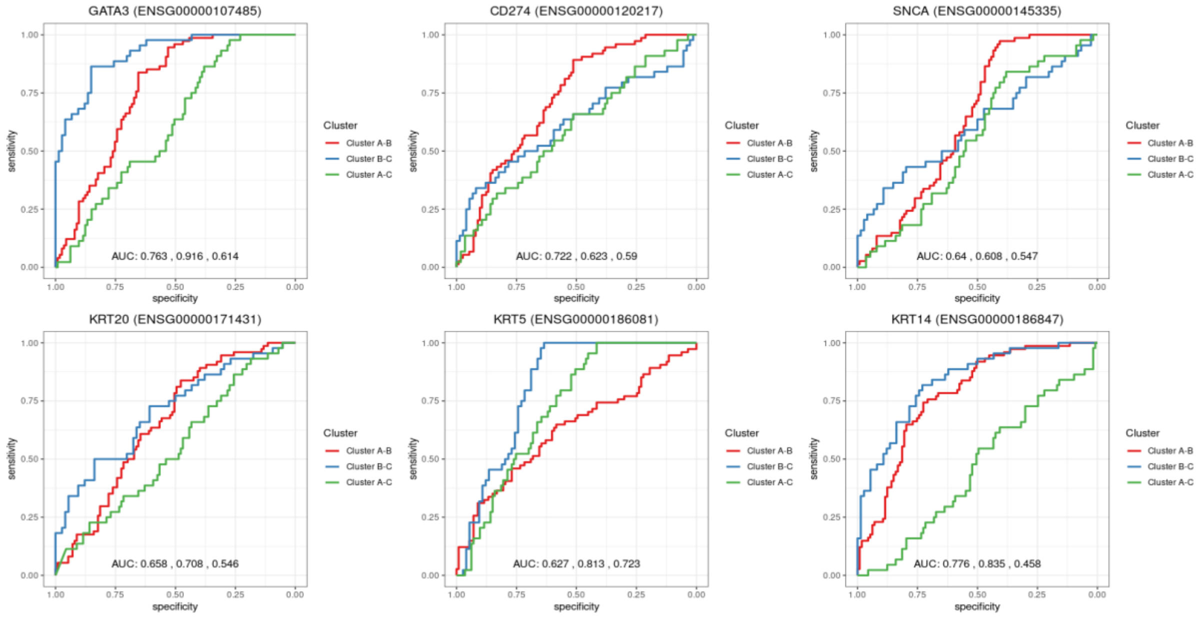


Figure 48. Sensitivity, specificity and cut point of selected signature genes

The most highest power of significant genes in cluster differentiation found in between cluster B-C were GATA 3, KRT 5 and KRT 14 (Table 10), in ROC curve of each gene are shown in Fig.

Table 10. Sensitivity, specificity and cut point of selected signature genes between cluster B-C

	Sensitivity	**Specificity**	**cut point**
GATA3	**0.8636**	**0.8514**	**13.2592**
KRT5	**0.8378**	**0.6591**	**8.2409**
KRT14	**0.6351**	**1**	**6.3685**

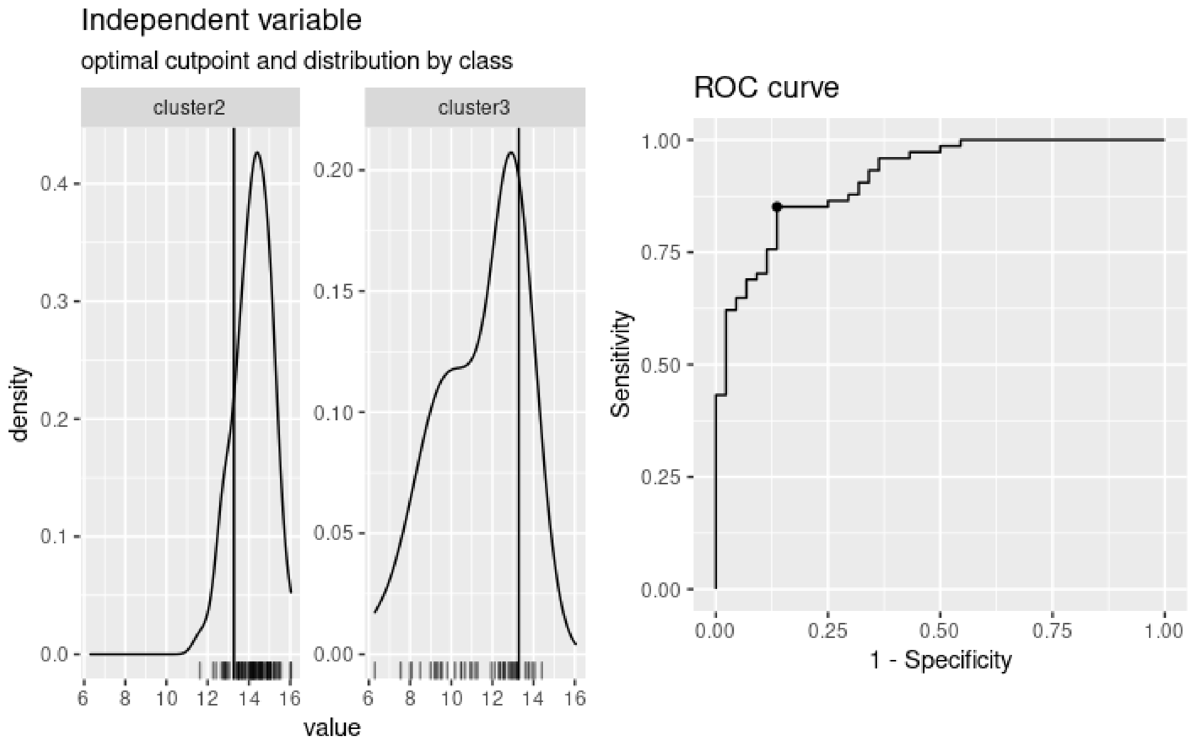


Figure 49. Cut point and ROC of GATA3 genes between cluster B-C

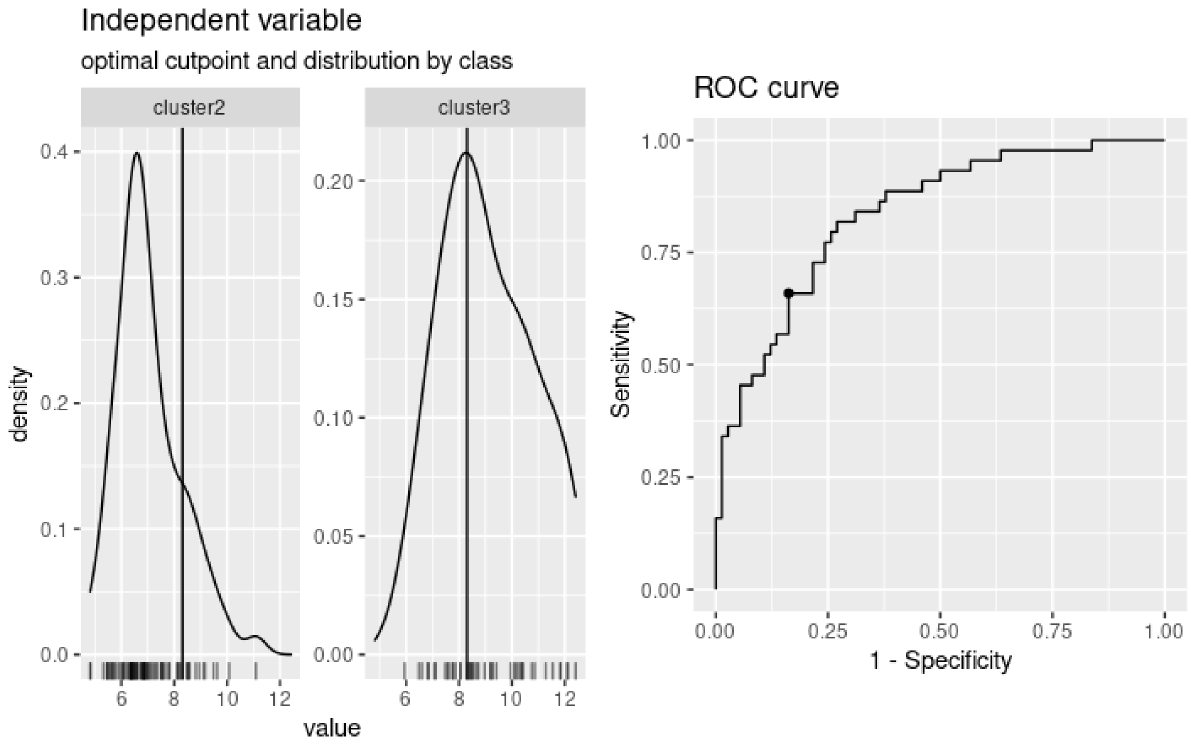


Figure 50. Cut point and ROC of KRT14(CK14) genes between cluster B-C

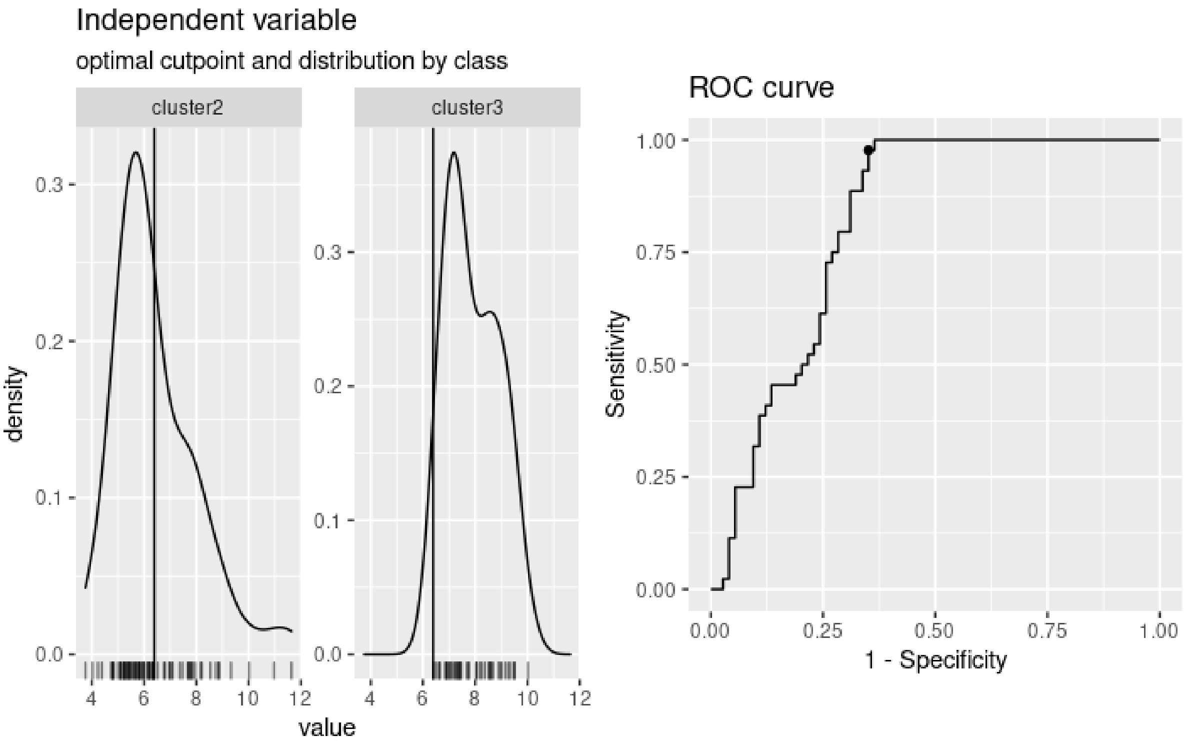


Figure 51. Cut point and ROC of KRT5 (CK5/6) genes between cluster B-C

3.3.4 Optimal cutpoint, accuracy, sensitivity, specificity and AUC of 37 difference gene in each cluster

We have concluded the limited gened of each cluster that have high accuracy, sensitivity, specificity and AUC and also optimal cutpoint if each gene. (Table 11)

Table 11 Optimal cutpoint, accuracy, sensitivity, specificity and AUC of 37 difference gene in each cluster

optimal_cutpoint	accuracy	sensitivity	specificity	AUC	pos_class	neg_class	gene.name	gene.id
15.13054314	0.949152542	0.909090909	0.972972973	0.988022113	cluster3	cluster2	COL6A2	ENSG00000142173
11.97505214	0.940677966	0.954545455	0.932432432	0.988022113	cluster3	cluster2	PDGFRB	ENSG00000113721
14.87228319	0.93220339	0.886363636	0.959459459	0.979115479	cluster3	cluster2	COL6A3	ENSG00000163359
16.60303405	0.93220339	0.909090909	0.945945946	0.967137592	cluster3	cluster2	COL1A2	ENSG00000164692
9.359841457	0.923728814	0.886363636	0.945945946	0.975737101	cluster3	cluster2	FGF7	ENSG00000140285
6.002827133	0.923728814	0.863636364	0.959459459	0.970515971	cluster3	cluster2	CAMK2A	ENSG00000070808
7.516355782	0.923728814	0.954545455	0.905405405	0.960380835	cluster3	cluster2	P2RX1	ENSG00000108405
8.392888585	0.923728814	1	0.878378378	0.956695332	cluster3	cluster2	KCNMB1	ENSG00000145936
7.611806926	0.915254237	0.931818182	0.905405405	0.947174447	cluster3	cluster2	FGF2	ENSG00000138685
11.93492692	0.915254237	0.977272727	0.878378378	0.955159705	cluster3	cluster2	MYLK	ENSG00000065534
9.786311303	0.906779661	0.931818182	0.891891892	0.941031941	cluster3	cluster2	CCL2	ENSG00000108691

9.394808887	0.906779661	0.863636364	0.932432432	0.950552826	cluster3	cluster2	EDNRA	ENSG00000151617
8.124055003	0.906779661	0.931818182	0.891891892	0.964373464	cluster3	cluster2	PRKG1	ENSG00000185532
7.600736771	0.898305085	0.886363636	0.905405405	0.944717445	cluster3	cluster2	HGF	ENSG00000019991
7.246230903	0.898305085	0.909090909	0.891891892	0.946253071	cluster3	cluster2	CREB5	ENSG00000146592
17.17557014	0.898305085	0.840909091	0.932432432	0.949938575	cluster3	cluster2	COL1A1	ENSG00000108821
8.864905314	0.898305085	0.886363636	0.905405405	0.953931204	cluster3	cluster2	KCNMA1	ENSG00000156113
11.65133272	0.889830508	0.931818182	0.864864865	0.946560197	cluster3	cluster2	TNC	ENSG00000041982
9.995064583	0.885350318	0.75	0.938053097	0.926588898	cluster3	cluster1	FGF7	ENSG00000140285
6.445600088	0.881355932	0.954545455	0.837837838	0.947174447	cluster3	cluster2	PTGFR	ENSG00000122420
8.278331858	0.881355932	0.772727273	0.945945946	0.915233415	cluster3	cluster2	ADRA2A	ENSG00000150594
12.83410809	0.878980892	0.795454545	0.911504425	0.913716814	cluster3	cluster1	MYLK	ENSG00000065534
7.320245712	0.872881356	0.863636364	0.878378378	0.938882064	cluster3	cluster2	GNAO1	ENSG00000087258
9.518218658	0.872611465	0.772727273	0.911504425	0.936041834	cluster3	cluster1	KCNMB1	ENSG00000145936
7.595931955	0.866242038	0.704545455	0.92920354	0.912107804	cluster3	cluster1	PTGFR	ENSG00000122420
7.26381242	0.866242038	0.659090909	0.946902655	0.879123089	cluster3	cluster1	IGF1	ENSG00000017427
7.00558066	0.86440678	0.886363636	0.851351351	0.950859951	cluster3	cluster2	PTGER3	ENSG00000050628
9.696381429	0.86440678	0.863636364	0.864864865	0.915233415	cluster3	cluster2	ITGA11	ENSG00000137809
6.923456429	0.86440678	0.818181818	0.891891892	0.889742015	cluster3	cluster2	ADORA1	ENSG00000163485
9.080415118	0.859872611	0.613636364	0.955752212	0.907079646	cluster3	cluster1	P2RX1	ENSG00000108405
15.97275163	0.859872611	0.659090909	0.938053097	0.885559131	cluster3	cluster1	COL6A2	ENSG00000142173
8.942791723	0.859872611	0.659090909	0.938053097	0.889179405	cluster3	cluster1	PRKG1	ENSG00000185532
9.614988721	0.853503185	0.704545455	0.911504425	0.879525342	cluster3	cluster1	KCNMA1	ENSG00000156113
7.970822028	0.847457627	0.704545455	0.932432432	0.915540541	cluster3	cluster2	PDE1A	ENSG00000115252
8.897935146	0.847457627	0.704545455	0.932432432	0.88544226	cluster3	cluster2	NPR 1.00	ENSG00000169418
10.89526475	0.840764331	0.613636364	0.92920354	0.859412711	cluster3	cluster1	CCL2	ENSG00000108691
8.529875041	0.840764331	0.636363636	0.920353982	0.885961384	cluster3	cluster1	FGF2	ENSG00000138685
7.89679706	0.840764331	0.75	0.876106195	0.859814964	cluster3	cluster1	PDE1A	ENSG00000115252
5.58740235	0.838983051	0.886363636	0.810810811	0.892813268	cluster3	cluster2	NGF	ENSG00000134259
10.08890118	0.834394904	0.568181818	0.938053097	0.833065165	cluster3	cluster1	EDNRA	ENSG00000151617
9.025286608	0.834394904	0.704545455	0.884955752	0.808125503	cluster3	cluster1	NPR 1.00	ENSG00000169418
8.795996736	0.834394904	0.681818182	0.89380531	0.845132743	cluster3	cluster1	ADRA2A	ENSG00000150594
6.622162229	0.830508475	0.772727273	0.864864865	0.872235872	cluster3	cluster2	BDKRB1	ENSG00000100739
9.741937824	0.830508475	0.636363636	0.945945946	0.875307125	cluster3	cluster2	TNXB	ENSG00000168477
12.97372033	0.828877005	0.92920354	0.675675676	0.863668979	cluster1	cluster2	COL6A2	ENSG00000142173
8.418409883	0.828025478	0.590909091	0.920353982	0.866854385	cluster3	cluster1	HGF	ENSG00000019991
6.144844467	0.821656051	0.818181818	0.82300885	0.862027353	cluster3	cluster1	CAMK2A	ENSG00000070808
15.51959587	0.818181818	0.902654867	0.689189189	0.868811289	cluster1	cluster2	COL1A1	ENSG00000108821
15.66478501	0.815286624	0.568181818	0.911504425	0.849356396	cluster3	cluster1	COL6A3	ENSG00000163359
10.3105166	0.815286624	0.522727273	0.92920354	0.814963797	cluster3	cluster1	TNXB	ENSG00000168477
8.918644275	0.815286624	0.5	0.938053097	0.850160901	cluster3	cluster1	GNAO1	ENSG00000087258
6.226280889	0.813559322	0.795454545	0.824324324	0.919226044	cluster3	cluster2	IGF1	ENSG00000017427
9.887559279	0.812834225	0.946902655	0.608108108	0.853384358	cluster1	cluster2	TNC	ENSG00000041982
14.67695898	0.812834225	0.911504425	0.662162162	0.856852428	cluster1	cluster2	COL1A2	ENSG00000164692
13.48829213	0.808917197	0.363636364	0.982300885	0.838495575	cluster3	cluster1	PDGFRB	ENSG00000113721
6.645688678	0.802547771	0.522727273	0.911504425	0.831255028	cluster3	cluster1	NTRK3	ENSG00000140538

9.636916223	0.802547771	0.477272727	0.92920354	0.799477072	cluster3	cluster1	ITGA8	ENSG00000077943
10.49857997	0.796791444	0.840707965	0.72972973	0.80399426	cluster1	cluster2	MYLK	ENSG00000065534
4.245436954	0.796791444	0.911504425	0.621621622	0.820318106	cluster1	cluster2	CAMK2A	ENSG00000070808
13.06476358	0.796791444	0.840707965	0.72972973	0.842621382	cluster1	cluster2	COL6A3	ENSG00000163359
6.208464543	0.796791444	0.849557522	0.716216216	0.820138723	cluster1	cluster2	CREB5	ENSG00000146592
7.616307314	0.789808917	0.295454545	0.982300885	0.734714401	cluster3	cluster1	RYR3	ENSG00000198838
8.342644928	0.789808917	0.568181818	0.876106195	0.784593725	cluster3	cluster1	COL4A4	ENSG00000081052
7.498288647	0.789808917	0.454545455	0.920353982	0.750804505	cluster3	cluster1	ATP1B2	ENSG00000129244
8.196355693	0.788135593	0.590909091	0.905405405	0.851658477	cluster3	cluster2	COL4A4	ENSG00000081052
18.59786455	0.78343949	0.454545455	0.911504425	0.735518906	cluster3	cluster1	COL1A1	ENSG00000108821
17.61894538	0.78343949	0.522727273	0.884955752	0.787208367	cluster3	cluster1	COL1A2	ENSG00000164692
5.785646207	0.780748663	0.920353982	0.567567568	0.789165271	cluster1	cluster2	FGF7	ENSG00000140285
6.755978322	0.780748663	0.902654867	0.594594595	0.798971538	cluster1	cluster2	KCNMA1	ENSG00000156113
6.295458069	0.779661017	0.590909091	0.891891892	0.828316953	cluster3	cluster2	NTRK3	ENSG00000140538
8.135940071	0.777070064	0.613636364	0.840707965	0.820997586	cluster3	cluster1	PTGER3	ENSG00000050628
9.027323188	0.77540107	0.752212389	0.810810811	0.824204736	cluster1	cluster2	ITGA11	ENSG00000137809
7.080592053	0.771186441	0.590909091	0.878378378	0.742628993	cluster3	cluster2	ATP1B2	ENSG00000129244
14.76613388	0.770700637	0.318181818	0.946902655	0.745776348	cluster3	cluster1	TNC	ENSG00000041982
7.344674719	0.770700637	0.613636364	0.831858407	0.762469831	cluster3	cluster1	ADORA1	ENSG00000163485
11.32377996	0.770053476	0.734513274	0.824324324	0.837598661	cluster1	cluster2	PDGFRB	ENSG00000113721
8.168522118	0.764705882	0.82300885	0.675675676	0.811169577	cluster1	cluster2	EDNRA	ENSG00000151617
8.312934396	0.757961783	0.386363636	0.902654867	0.727473854	cluster3	cluster1	BDKRB1	ENSG00000100739
7.077832763	0.754237288	0.863636364	0.689189189	0.774262899	cluster3	cluster2	AVPR1A	ENSG00000166148
6.559483816	0.754237288	0.931818182	0.648648649	0.825859951	cluster3	cluster2	ANGPT1	ENSG00000154188
8.217513355	0.754010695	0.89380531	0.540540541	0.768476441	cluster1	cluster2	CCL2	ENSG00000108691
5.566831178	0.754010695	0.82300885	0.648648649	0.770748625	cluster1	cluster2	P2RX1	ENSG00000108405
7.086393598	0.748663102	0.787610619	0.689189189	0.748505142	cluster1	cluster2	KCNMB1	ENSG00000145936
9.617669535	0.745762712	0.5	0.891891892	0.719594595	cluster3	cluster2	ITGA8	ENSG00000077943
6.876902256	0.74522293	0.340909091	0.902654867	0.783185841	cluster3	cluster1	NGF	ENSG00000134259
8.040385391	0.74522293	0.659090909	0.778761062	0.790426388	cluster3	cluster1	CREB5	ENSG00000146592
7.277550694	0.743315508	0.734513274	0.756756757	0.777445587	cluster1	cluster2	PRKG1	ENSG00000185532
5.15983813	0.737967914	0.814159292	0.621621622	0.723152356	cluster1	cluster2	PTGFR	ENSG00000122420
5.035757385	0.737967914	0.938053097	0.432432432	0.778880651	cluster1	cluster2	PTGER3	ENSG00000050628
5.932093438	0.737967914	0.805309735	0.635135135	0.761420713	cluster1	cluster2	ADORA1	ENSG00000163485
5.40067901	0.732620321	0.911504425	0.459459459	0.73056685	cluster1	cluster2	HGF	ENSG00000019991
10.53198408	0.732484076	0.045454545	1	0.686242961	cluster3	cluster1	ANGPT1	ENSG00000154188
11.65910541	0.732484076	0.272727273	0.911504425	0.669750603	cluster3	cluster1	ITGA11	ENSG00000137809
8.169702747	0.72611465	0.477272727	0.82300885	0.707562349	cluster3	cluster1	AVPR1A	ENSG00000166148
5.508098753	0.721925134	0.902654867	0.445945946	0.738698876	cluster1	cluster2	GNAO1	ENSG00000087258
6.639627253	0.703389831	0.568181818	0.783783784	0.721437346	cluster3	cluster2	RYR3	ENSG00000198838
4.589351842	0.700534759	0.82300885	0.513513514	0.688591246	cluster1	cluster2	IGF1	ENSG00000017427
5.95628405	0.689839572	0.82300885	0.486486486	0.681655106	cluster1	cluster2	ANGPT1	ENSG00000154188
5.696923613	0.684491979	0.805309735	0.5	0.703061469	cluster1	cluster2	BDKRB1	ENSG00000100739
4.764814732	0.684491979	0.752212389	0.581081081	0.665271466	cluster1	cluster2	NGF	ENSG00000134259
6.455683628	0.673796791	0.681415929	0.662162162	0.693853145	cluster1	cluster2	FGF2	ENSG00000138685

5.773886663	0.657754011	0.902654867	0.283783784	0.619588615	cluster1	cluster2	AVPR1A	ENSG00000166148
6.775991666	0.657754011	0.752212389	0.513513514	0.646496054	cluster1	cluster2	TNXB	ENSG00000168477
8.599869878	0.64171123	0.391891892	0.805309735	0.603204975	cluster2	cluster1	ITGA8	ENSG00000077943
5.891807142	0.636363636	0.796460177	0.391891892	0.617077254	cluster1	cluster2	PDE1A	ENSG00000115252
7.548315019	0.625668449	0.094594595	0.973451327	0.520090887	cluster2	cluster1	RYR3	ENSG00000198838
3.713780217	0.620320856	0.991150442	0.054054054	0.538387945	cluster1	cluster2	NTRK3	ENSG00000140538
7.497294766	0.614973262	0.646017699	0.567567568	0.628318584	cluster1	cluster2	NPR 1.00	ENSG00000169418
5.792872806	0.609625668	0.690265487	0.486486486	0.588854341	cluster1	cluster2	COL4A4	ENSG00000081052
9.094951437	0.609625668	0.013513514	1	0.542095193	cluster2	cluster1	ATP1B2	ENSG00000129244
4.230722549	0.604278075	1	0	0.582276967	cluster1	cluster2	ADRA2A	ENSG00000150594

CHAPTER 4

Discussion

This is the first report of our institutional MIBC cohort subtyping using an unsupervised clustering based on transcriptomic data in Thailand. The primary finding of this study is that the locations of MIBC cluster on the principal components identified from transcriptome data can be predicted from an understanding of the average coalescent differential genes for tissue samples. This analysis transformed the high-dimensional data into an orthogonal basis which represent the variant of mRNA expression profile in each sample. Unsupervised clustering revealed the three clusters of MIBC with the 37 genes expressed differently in all clusters. Interestingly, all signaling pathway was found to be increased in colon cancer (56), breast cancers (57), and liver cancer (58). The PI3K-Akt activation was also found in breast cancer (59), gastric cancer (60), and thyroid carcinoma (61). The ubiquitous signal transduction MAPK pathway also associated to cancer cell proliferation and survival, and inflammatory environment (62). Although all these signal transductions are in cancers, the dominant pathway in cancer cell is depend on the genetic background, the mutation status, or type of cancer (63) determining the aggressive behavior, the progression rate, and drug response of cancer.

Surprisingly, most of the genes are not related to the markers used for subtyping in the previous reports (6,7,9,10,22,45,64). This may be due to drug response study was not included for marker selection. Moreover, other studies performed the association of other biological factors such as lncRNA, miRNA, protein expression, or DNA methylation (6,7,9,10,22,45,64). We determined the sensitivity and specificity of the genes instead. ROC curve analysis revealed most of the genes showed high correlation of the sensitivity and specificity only for cluster B which may be used as expression markers for Thai MIBC patient. Our transcriptomic clustering provided three clusters of MIBC tissue which expressed the specific pattern of mRNA profiling. In addition to our MIBC transcriptomic study from patient tissue, we included the information from TCGA dataset for validation. However, PCA analysis with the comparison between two cohorts demonstrated the obviously different PC coordinates between data from our MIBC tissue samples and TCGA dataset. The variation of genetic background of the different population studied may be the factors that caused the difference of PCA data plot (65). By using the initial cluster centroids of the MIBC tissue data, we applied the distance of tissue PCA coordinated with mRNA expression profile to TCGA data for transcriptomic clustering of TCGA cohort. This gives us the influence of various expression scenarios on the relationships between MIBC patients identified from PCA and how to apply this PCA for data inference in other population. The three clusters obtained from this method were related to the significant difference of the overall survival of MIBC patients meaning that the classification based on transcriptomic data of MIBC tissue may be alternative way to predict the survival outcome.

The nature of disease heterogeneity is represented by bladder cancer molecular subtyping. MIBCs can be classified into at least three intrinsic subtypes, including luminal, basal, and double negative, according to previous gene expression analysis (34). High expression of terminal urothelial differentiation markers (GATA3, CK20, and uroplakin 2), often known as umbrella cells, is a feature of luminal malignancies (7). Because umbrella cells have a shorter lifespan than basal cells, they are less sensitive to genetic mutations, but their chromatin landscape changes more often. The tumor tissue of the basal-like bladder cancer expressed mesenchymal stem cell biomarkers (CK5/6 and CK14), as well as squamous and sarcomatous characteristics (23). According to recent research, GATA3 and CK5/6 can detect molecular subtypes in 80-90 percent of instances (30,34). GATA3 and CK5/6 were shown to be linked with survival result in 62 percent of cases in our analysis, and only these two markers could unambiguously divide cases into luminal, basal, or double-negative subtypes. Loss of expression in one of these two markers was linked to a lower chance of survival, and loss of both markers was a strong predictor of poor outcome. Adding CK20 and CK14 to the criteria to identify the subtypes did not seem to improve survival prediction, despite substantial associations between CK20 and GATA3 expression and CK14 and CK5/6 expression.

MIBCs that expressed GATA3 were found to be less aggressive and to have a higher chance of survival. GATA3, also known as GATA3 binding protein, is a transcription factor that controls the expression of genes involved in breast and urothelial epithelial luminal differentiation (68,69). GATA3 is also found in T-lymphocytes, the central nervous system, and erythrocytes (70). The triple negative subtype of breast cancer has been found to have lower GATA3 expression (71). Loss of GATA3 expression increased tumor cell motility and invasion in urothelial cell line models by upregulating oncogenes (72,73). GATA3 has been studied in bladder cancer in a number of clinical trials (74-76). GATA3 expression loss has been linked to high-grade malignancy (76). Furthermore, in most investigations, GATA3-negative bladder cancer patients had a poorer prognosis. (74,75,77). When those reports are combined with our findings, GATA3 appears to be a potential biomarker in MIBCs. CK20, another luminal marker investigated in this work, has been linked to greater tumor grade and stage in papillary urothelial carcinoma (78). However, no significant link was found between CK20 and any clinicopathological condition or survival result in our investigation.

CK5/6 is a cytokeratin that is expressed in the squamous epithelial lineage and is commonly utilized as a squamous differentiation marker that distinguishes the basal subtype (79). In multiple studies, CK5/6 expression in urothelial cancer was linked to a poorer prognosis (79,80). On the other hand, studies have shown that reduction of CK5/6 expression is linked to a lower survival rate in transitional cell carcinoma of the upper urinary tract (80,81).

Loss of CK5/6 expression was linked to a considerably lower survival outcome in our study. CK14 is another basal subtype marker whose expression has been shown to have a negative connection with MIBC survival.

Although molecular subtyping has been shown to be associated with disease progression and treatment outcomes in MIBCs, RNA expression profiling is not a routinely used technique. Several research have looked into the idea of evaluating an immunohistochemical panel to be utilized for intrinsic subtype categorization in MIBCs (74,80,82). Apart from GATA3, the predictive usefulness of other IHC markers remained a mystery. The gap in results could be explained in part by differences in staining techniques and interpretation. Because GATA3 and CK5/6 were found to have a substantial survival correlation in our investigation, these markers were combined into a simple subgroup categorization as luminal when the tumor had exclusive GATA3 expression and basal when the tumor had exclusive CK5/6 expression. The study discovered that the double-negative subtype, which means that both markers were negative, predicted the worst outcome. Other previously reported combinations, such as CK20 with CK5/6 or CK20 with CK14, were explored but yielded no intriguing results.

However, mRNA expression clustering that exhibited 3 clusters and KRT expressions in mRNA level as same as in IHC study found the important correlations of CK20, CK5/6, GATA3 and CK14 in each cluster.

The modest sample size of our investigation, as well as the lack of gene expression profiling to confirm concordance between molecular subtypes and IHC marker expression, were also limitations. Furthermore, only 20% of our patients underwent chemotherapy following a radical cystectomy since their physical condition prevented them from doing so. Because neoadjuvant chemotherapy is becoming more popular in the treatment of MIBC, the findings of our study may aid in identifying patients at high risk of treatment failure who should get chemotherapy before undergoing definitive surgery. The presented study's high core agreement, as well as earlier studies comparing TMA core expression to whole slides, suggests that core regions are typical of expression in the entire sample. Tissue microarrays, on the other hand, have clear limitations in terms of capturing tumor heterogeneity. To validate the efficacy of this methodology, full slide analysis comparing expression and subtype assignment of a core to complete sections will be necessary. Validation on complete slides is especially important since pathologists evaluate larger samples in real practice. Staining patterns would have to be assessable in ordinary workflow if subtyping was to be employed effectively in the clinical context.

- Small sample size
- Validating and refining subtype classification
- require prospective studies
- molecular subtyping of MIBC has mainly focused on stratifying global mRNA expression, which comprises less than 2% of total transcription, due to the majority of transcribed genes: ribosomal RNAs and non-coding RNAs

While IHC is a reliable and reasonable method for clinical subtyping and avoiding the difficulties of transcriptome profiling, it may be subject to artifacts of its own due to variances in antigen storage. The discovery of double negative subtypes and possible challenges with antigen preservation in this study emphasizes the importance of giving careful thought to these concerns about IHC staining repeatability. These concerns about antigen storage and staining intensities make it difficult to create a repeatable, therapeutically effective assay. Finally, the transcriptome profiles of these tumors should be evaluated in tandem with IHC in order to ensure that the two approaches produce identical results. In order to confirm if the same subtypes have been identified.

CHAPTER 5

Conclusion

Molecular subtyping classifications have provided insight into the biology of bladder tumors, especially regarding tumor heterogeneity. New genomic techniques provide insight into the marked genetic complexity of MIBC. Over the last decade, RNA- based molecular subtyping has identified distinct or partially overlapping molecular classifications of MIBC. Our studies show that molecular stratification of MIBC is of clinical importance into 3 clusters with validate these subtypes in TCGA dataset, suggesting that responses to chemotherapy and immunotherapy may be increased for specific MIBC subtypes. Further investigation is needed into the clinical applicability of molecular subtypes before their incorporation into the personalized care of MIBC patients. Moreover, GATA3, CK20, CK5/6, and CK14 staining was selected to be tested against clinical outcomes with respect to survival after a radical cystectomy. While subtype proportions and staining patterns differed by sample type, we believe this was primarily due to poor antigen preservation in cystectomy samples. In addition to the fact that TURBT samples are obtained at an earlier and perhaps more prognostically relevant timepoint prior to NACT treatment, this points to the utility of TURBT samples for MIBC subtyping, and potential limitations of IHC-based subtyping using cystectomy samples. The study evaluated 4 immunohistochemical markers that mark luminal subtype (GATA3 and CK20) and basal subtype (CK5/6 and CK14) in MIBC, focusing on their association with survival outcome after a radical cystectomy. GATA3 and CK5/6 were significantly associated with survival probability. When the 2 markers were combined, the double-negative subtype had the poorest prognosis.

We believe that altogether, this work demonstrates that a much simpler, IHC-based assay for subtyping retains key biologic and clinical associations seen previously with more complex profiling methods. Future work will validate the prognostic associations in larger cohorts of patients, as well as investigate the predictive utility of IHC-based subtypes and also new molecular subtyping of our study.

- RNA expression-based subtypes in muscle-invasive bladder cancer: unique and offer insights to biology and subtype specific treatment
- novel molecular subtypes of MIBC: 3 clusters
- Neoadjuvant chemotherapy and immunotherapy response: associated with each subtype and may provide insights into the mechanisms of treatment response >> further evaluation
- Clinical trials validating predictive biomarkers: essential for precision medicine
- a simplified four-gene signature: a practical, cost-effective platform to translational research
- identifying 4 molecular subtypes (luminal, basal, mixed and double negative)
- double negative molecular subtypes: worse bladder cancer-related mortality

Bibliography

1. Lenis AT, Lec PM, Chamie K, Mshs MD. Bladder Cancer: A Review. JAMA. 2020 Nov 17;324(19):1980-1991.
2. Richters A, Aben KKH, Kiemeney LALM. The global burden of urinary bladder cancer: an update. World J Urol. 2020 Aug;38(8):1895-1904.
3. Ru Y, Dancik GM, Theodorescu D. Biomarkers for prognosis and treatment selection in advanced bladder cancer patients. Curr Opin Urol. 2011 Sep;21(5):420-7.
4. Bejrananda T, Pripatnanont C, Tanthanuch M, Karnjanawanichkul W. Oncological Outcomes of Radical Cystectomy for Transitional Cell Carcinoma of Bladder. J Med Assoc Thai. 2017 Jan;100(1):24-32.
5. Sjödahl G, Lauss M, Lövgren K, Chebil G, Gudjonsson S, Veerla S, Patschan O, Aine M, Fernö M, Ringnér M, Månsson W, Liedberg F, Lindgren D, Höglund M. A molecular taxonomy for urothelial carcinoma. Clin Cancer Res. 2012 Jun 15;18(12):3377-86.
6. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer Cell. 2014 Feb 10;25(2):152-65.
7. Damrauer JS, Hoadley KA, Chism DD, Fan C, Tiganelli CJ, Wobker SE, et al. Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. Proc Natl Acad Sci U S A. 2014 Feb 25;111(8):3110-5.
8. Hussain SA, Palmer DH, Syn WK, Sacco JJ, Greensmith RMD, Elmetwali T, Aachi V, Lloyd BH, Jithesh PV, Arrand J, Barton D, Ansari J, Sibson DR, James ND. Gene expression profiling in bladder cancer identifies potential therapeutic targets. Int J Oncol. 2017 Apr;50(4):1147-1159.
9. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. TCGA Research Network, Weinstein JN, Kwiatkowski DJ, Lerner SP. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017 Oct 19;171(3):540-556.
10. Kamoun A, de Reyniès A, Allory Y, Sjödahl G, Robertson AG, Seiler R, et al. Bladder Cancer Molecular Taxonomy Group. A Consensus Molecular Classification of Muscle-invasive Bladder Cancer. Eur Urol. 2020 Apr;77(4):420-433.

11. Mollica V, Rizzo A, Montironi R, Cheng L, Giunchi F, Schiavina R, et al. Current Strategies and Novel Therapeutic Approaches for Metastatic Urothelial Carcinoma. Cancers (Basel). 2020 Jun 2;12(6):1449.
12. Necchi A, de Jong JJ, Raggi D, Briganti A, Marandino L, Gallina A, et al. Molecular Characterization of Residual Bladder Cancer after Neoadjuvant Pembrolizumab. Eur Urol. 2021 Aug;80(2):149-159.
13. Necchi A, Raggi D, Gallina A, Madison R, Colecchia M, Lucianò R, et al. Updated Results of PURE-01 with Preliminary Activity of Neoadjuvant Pembrolizumab in Patients with Muscle-invasive Bladder Carcinoma with Variant Histologies. Eur Urol. 2020 Apr;77(4):439-446.
14. Necchi A, Raggi D, Gallina A, Ross JS, Farè E, Giannatempo P, et al. Impact of Molecular Subtyping and Immune Infiltration on Pathological Response and Outcome Following Neoadjuvant Pembrolizumab in Muscle-invasive Bladder Cancer. Eur Urol. 2020 Jun;77(6):701-710.
15. Lopez-Beltran A, Cimadamore A, Montironi R, Cheng L. Molecular pathology of urothelial carcinoma. Hum Pathol. 2021 Jul;113:67-83.
16. Lopez-Beltran A, Henriques V, Montironi R, Cimadamore A, Raspollini MR, Cheng L. Variants and new entities of bladder cancer. Histopathology. 2019 Jan;74(1):77-96.
17. Lopez-Beltran A, López-Rios F, Montironi R, Wildsmith S, Eckstein M. Immune Checkpoint Inhibitors in Urothelial Carcinoma: Recommendations for Practical Approaches to PD-L1 and Other Potential Predictive Biomarker Testing. Cancers (Basel). 2021 Mar 20;13(6):1424.
18. Al-Obaidy KI, Cheng L. Fibroblast growth factor receptor (*FGFR*) gene: pathogenesis and treatment implications in urothelial carcinoma of the bladder. J Clin Pathol. 2021 Aug;74(8):491-495.
19. Yoshida T, Kates M, Fujita K, Bivalacqua TJ, McConkey DJ. Predictive biomarkers for drug response in bladder cancer. Int J Urol. 2019 Nov;26(11):1044-1053.
20. Rentsch CA, Müller DC, Ruiz C, Bubendorf L. Comprehensive Molecular Characterization of Urothelial Bladder Carcinoma: A Step Closer to Clinical Translation? Eur Urol. 2017 Dec;72(6):960-961.
21. Al-Ahmadie H, Netto GJ. Updates on the Genomics of Bladder Cancer and Novel Molecular Taxonomy. Adv Anat Pathol. 2020 Jan;27(1):36-43.

22. McConkey DJ, Choi W, Shen Y, Lee IL, Porten S, Matin SF, et al. A Prognostic Gene Expression Signature in the Molecular Classification of Chemotherapy-naïve Urothelial Cancer is Predictive of Clinical Outcomes from Neoadjuvant Chemotherapy: A Phase 2 Trial of Dose-dense Methotrexate, Vinblastine, Doxorubicin, and Cisplatin with Bevacizumab in Urothelial Cancer. Eur Urol. 2016 May;69(5):855-62.
23. Weyerer V, Stoehr R, Bertz S, Lange F, Geppert CI, Wach S, Taubert H, Sikic D, Wullich B, Hartmann A, Eckstein M. Prognostic impact of molecular muscle-invasive bladder cancer subtyping approaches and correlations with variant histology in a population-based mono-institutional cystectomy cohort. World J Urol. 2021 Nov;39(11):4011-4019.
24. Sjödahl G, Eriksson P, Liedberg F, Höglund M. Molecular classification of urothelial carcinoma: global mRNA classification versus tumour-cell phenotype classification. J Pathol. 2017 May;242(1):113-125.
25. Sjödahl G, Jackson CL, Bartlett JM, Siemens DR, Berman DM. Molecular profiling in muscle-invasive bladder cancer: more than the sum of its parts. J Pathol. 2019 Apr;247(5):563-573.
26. Seiler R, Ashab HAD, Erho N, van Rhijn BWG, Winters B, et al. Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy. Eur Urol. 2017 Oct;72(4):544-554.
27. Satyal, U.; Sikder, R.K.; Mcconkey, D.; Plimack, E.R.; Abbosh, P.H. Clinical implications of molecular subtyping in bladder cancer. *Curr. Opin. Urol.* 2019, *29*, 350–356.
28. Rodriguez Pena, M.D.C.; Chaux, A.; Eich, M.L.; Tregnago, A.C.; Taheri, D.; Borhan, W.; Sharma, R.; Rezaei, M.K.; Netto, G.J. Immunohistochemical assessment of basal and luminal markers in non-muscle invasive urothelial carcinoma of bladder. *Virchows Arch.* 2019, *475*, 349–356.
29. Warrick, J.I.; Knowles, M.A.; Yves, A.; Van Der Kwast, T.; Grignon, D.J.; Kristiansen, G.; Egevad, L.; Hartmann, A.; Cheng, L. Report from the International Society of Urological Pathology (ISUP) Consultation Conference on Molecular Pathology of Urogenital Cancers. II. Molecular Pathology of Bladder Cancer: Progress and Challenges. *Am. J. Surg. Pathol.* 2020, *44*, E30–E46.
30. Dadhania, V.; Zhang, M.; Zhang, L.; Bondaruk, J.; Majewski, T.; Siefker-Radtke, A.; Guo, C.C.; Dinney, C.; Cogdell, D.E.; Zhang, S.; et al. Meta-Analysis of the Luminal

and Basal Subtypes of Bladder Cancer and the Identification of Signature Immunohistochemical Markers for Clinical Use. *EBioMedicine* 2016, *12*, 105–117.

31. Kim, J.; Kwiatkowski, D.; McConkey, D.J.; Meeks, J.J.; Freeman, S.S.; Bellmunt, J.; Getz, G.; Lerner, S.P. The Cancer Genome Atlas Expression Subtypes Stratify Response to Checkpoint Inhibition in Advanced Urothelial Cancer and Identify a Subset of Patients with High Survival Probability. *Eur. Urol.* 2019, *75*, 961–964.
32. Kim, B.; Jang, I.; Kim, K.; Jung, M.; Lee, C.; Park, J.H.; Kim, Y.A.; Moon, K.C. Comprehensive gene expression analyses of immunohistochemically defined subgroups of muscle-invasive urinary bladder urothelial carcinoma. *Int. J. Mol. Sci.* 2021, *22*, 628.
33. Kardos, J.; Rose, T.L.; Manocha, U.; Wobker, S.E.; Damrauer, J.S.; Bivalaqua, T.J.; Kates, M.; Moore, K.J.; Parker, J.S.; Kim, W.Y. Development and validation of a NanoString BASE47 bladder cancer gene classifier. *PLoS ONE* 2020, *15*, e0243935.
34. Guo, C.C.; Bondaruk, J.; Yao, H.; Wang, Z.; Zhang, L.; Lee, S.; Lee, J.G.; Cogdell, D.; Zhang, M.; Yang, G.; et al. Assessment of Luminal and Basal Phenotypes in Bladder Cancer. *Sci. Rep.* 2020, *10*, 9743.
35. Damrauer, J.S.; Roell, K.R.; Smith, M.A.; Sun, X.; Kirk, E.L.; Hoadley, K.A.; Benefield, H.C.; Iyer, G.; Solit, D.B.; Milowsky, M.I.; et al. Identification of a novel inflamed tumor microenvironment signature as a predictive biomarker of bacillus Calmette-Guérin immunotherapy in non-muscle-invasive bladder cancer. *Clin. Cancer Res.* 2021, *27*, 4599–4609.
36. Hodgson, A.; Liu, S.K.; Vesprini, D.; Xu, B.; Downes, M.R. Basal-subtype bladder tumours show a 'hot' immunophenotype. *Histopathology* 2018, *73*, 748–757.
37. Hurst, C.D.; Alder, O.; Platt, F.M.; Droop, A.; Stead, L.F.; Burns, J.E.; Burghel, G.J.; Jain, S.; Klimczak, L.J.; Lindsay, H.; et al. Genomic Subtypes of Non-invasive Bladder Cancer with Distinct Metabolic Profile and Female Gender Bias in KDM6A Mutation Frequency. *Cancer Cell* 2017, *32*, 701–715.e7.
38. Ikeda, J.; Ohe, C.; Yoshida, T.; Kuroda, N.; Saito, R.; Kinoshita, H.; Tsuta, K.; Matsuda, T. Comprehensive pathological assessment of histological subtypes, molecular subtypes based on immunohistochemistry, and tumor-associated immune cell status in muscle-invasive bladder cancer. *Pathol. Int.* 2021, *71*, 173–182.
39. Jalanko, T.; de Jong, J.J.; Gibb, E.A.; Seiler, R.; Black, P.C. Genomic Subtyping in Bladder Cancer. *Curr. Urol. Rep.* 2020, *21*.

40. Olkhov-Mitsel, E. Hodgson, A., Liu, S.K., et al. Immune gene expression profiles in high-grade urothelial carcinoma of the bladder: A NanoString study. *J. Clin. Pathol.* **2021**, *74*, 53–57.

41. Eriksson P. Molecular Characterization of Bladder Cancer Subtypes. Lund University, Faculty of Medicine, 2018.

42. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med. 2009 Feb 19;360(8):790–800.

43. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast Cancer. Nat Rev Dis Prim. 2019 Sep 23;5(1):66.

44. Aine M, Eriksson P, Liedberg F, Sjödahl G, Höglund M. Biological determinants of bladder cancer gene expression subtypes. Sci Rep. 2015 Jun 8;5:10957.

45. TCGA Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315-322.

46. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415-21.

47. Lindgren D, Sjödahl G, Lauss M, Staaf J, Chebil G, Lövgren K, Gudjonsson S, Liedberg F, Patschan O, Månsson W, Fernö M, Höglund M. Integrated genomic and gene expression profiling identifies two major genomic circuits in urothelial carcinoma. PLoS One. 2012;7(6):e38863.

48. van Rhijn BW, Vis AN, van der Kwast TH, Kirkels WJ, Radvanyi F, Ooms EC, et al. Molecular grading of urothelial cell carcinoma with fibroblast growth factor receptor 3 and MIB-1 is superior to pathologic grade for the prediction of clinical outcome. J Clin Oncol. 2003 May 15;21(10):1912–2.

49. Droller MJ. Bladder cancer. Current problems in surgery. 1981;18(4):205–279.

50. Lotan Y, Boorjian SA, Zhang J, Bivalacqua TJ, Porten SP, Wheeler T, et al. Molecular Subtyping of Clinically Localized Urothelial Carcinoma Reveals Lower Rates of Pathological Upstaging at Radical Cystectomy Among Luminal Tumors. Eur Urol. 2019 Aug;76(2):200-206.

51. Taber A, Christensen E, Lamy P, Nordentoft I, Prip F, Lindskrog SV, Birkenkamp-Demtröder K, Okholm TLH, Knudsen M, Pedersen JS, Steiniche T, Agerbæk M, Jensen JB, Dyrskjøt L. Molecular correlates of cisplatin-based chemotherapy response in muscle invasive bladder cancer by integrated multi-omics analysis. Nat Commun. 2020 Sep 25;11(1):4858.

52. Mariathasan S, Turley SJ, Nickles D, et al. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature. 2018 Feb 22;554(7693):544-548.
53. Bernardo C, Eriksson P, Marzouka N, Liedberg F, Sjödahl G, Höglund M. Molecular pathology of the luminal class of urothelial tumours. J Pathol. 2019 Nov;249(3):308–318.
54. Font A, Domènech M, Benítez R, Rava M, Marqués M, Ramírez JL, et al. Immunohistochemistry-Based Taxonomical Classification of Bladder Cancer Predicts Response to Neoadjuvant Chemotherapy. Cancers (Basel). 2020 Jul 3;12(7):1784.
55. Rebola J, Aguiar P, Blanca A, Montironi R, Cimadamore A, Cheng L, Henriques V, Lobato-Faria P, Lopez-Beltran A. Predicting outcomes in non-muscle invasive (Ta/T1) bladder cancer: the role of molecular grade based on luminal/basal phenotype. Virchows Arch. 2019 Oct;475(4):445-455.
56. Cs A, Wa K, P P, Sj R-T, Gr M. Plasma membrane Ca2+-ATPase expression during colon cancer cell line differentiation. Biochemical and biophysical research communications [Internet]. 2007 Apr 20;355(4).
57. Feng M, Grice DM, Faddy HM, Nguyen N, Leitch S, Wang Y, et al. Store-independent activation of Orai1 by SPCA2 in mammary tumors. Cell. 2010 Oct 1;143(1):84–98.
58. Maynard JP, Lee J-S, Sohn BH, Yu X, Lopez-Terrada D, Finegold MJ, et al. P2X3 purinergic receptor overexpression is associated with poor recurrence-free survival in hepatocellular carcinoma patients. Oncotarget. 2015 Dec 1;6(38):41162–79.
59. Bellacosa A, Kumar CC, Cristofano AD, Testa JR. Activation of AKT Kinases in Cancer: Implications for Therapeutic Targeting. In: Advances in Cancer Research [Internet]. Academic Press; 2005 [cited 2021 Dec 29]. p. 29–86.
60. Kobayashi I, Semba S, Matsuda Y, Kuroda Y, Yokozaki H. Significance of Akt phosphorylation on tumor growth and vascular endothelial growth factor expression in human gastric carcinoma. Pathobiology. 2006;73(1):8-17.
61. Altomare DA, Testa JR. Perturbations of the AKT signaling pathway in human cancer. Oncogene. 2005 Nov 14;24(50):7455-64.
62. Lee S, Rauch J, Kolch W. Targeting MAPK Signaling in Cancer: Mechanisms of Drug Resistance and Sensitivity. Int J Mol Sci. 2020 Feb 7;21(3):1102.
63. Sever R, Brugge JS. Signal Transduction in Cancer. Cold Spring Harb Perspect Med. 2015 Apr;5(4):a006098.

64. Brown BC, Bray NL, Pachter L. Expression reflects population structure. PLoS Genet. 2018 Dec 19;14(12):e1007841.
65. Marzouka NA, Eriksson P, Rovira C, Liedberg F, Sjodahl G, Hoglund M. A validation and extended description of the Lund taxonomy for urothelial carcinoma using the TCGA cohort. Scientific reports. 2018;8(1):3737.
66. Ochoa AE, Choi W, Su X, et al. Specific micro-RNA expression patterns distinguish the basal and luminal subtypes of muscle-invasive bladder cancer. Oncotarget. 2016;7(49):80164–80174.
67. Rinaldetti S, Rempel E, Worst TS, et al. Subclassification, survival prediction and drug target analyses of chemotherapy-naive muscle-invasive bladder cancer with a molecular screening. Oncotarget. 2018;9(40):25935–25945.
68. Miettinen M, McCue PA, Sarlomo-Rikala M, Rys J, Czapiewski P, Wazny K, et al. GATA3: a multispecific but potentially useful marker in surgical pathology: a systematic analysis of 2500 epithelial and nonepithelial tumors. Am J Surg Pathol. 2014;38(1):13-22.
69. Liu H, Shi J, Wilkerson ML, Lin F. Immunohistochemical evaluation of GATA3 expression in tumors and normal tissues: a useful immunomarker for breast and urothelial carcinomas. Am J Clin Pathol. 2012;138(1):57-64.
70. Lentjes MH, Niessen HE, Akiyama Y, de Bruine AP, Melotte V, van Engeland M. The emerging role of GATA transcription factors in development and disease. Expert Rev Mol Med. 2016;18:e3.
71. Cimino-Mathews A, Subhawong AP, Illei PB, Sharma R, Halushka MK, Vang R, et al. GATA3 expression in breast carcinoma: utility in triple-negative, sarcomatoid, and metastatic carcinomas. Hum Pathol. 2013;44(7):1341-9.
72. Li Y, Ishiguro H, Kawahara T, Kashiwagi E, Izumi K, Miyamoto H. Loss of GATA3 in bladder cancer promotes cell migration and invasion. Cancer Biol Ther. 2014;15(4):428-35.
73. Li Y, Ishiguro H, Kawahara T, Miyamoto Y, Izumi K, Miyamoto H. GATA3 in the urinary bladder: suppression of neoplastic transformation and down-regulation by androgens. Am J Cancer Res. 2014;4(5):461-73.
74. Wang CC, Tsai YC, Jeng YM. Biological significance of GATA3, cytokeratin 20, cytokeratin 5/6 and p53 expression in muscle-invasive bladder cancer. PLoS One. 2019;14(8):e0221785.

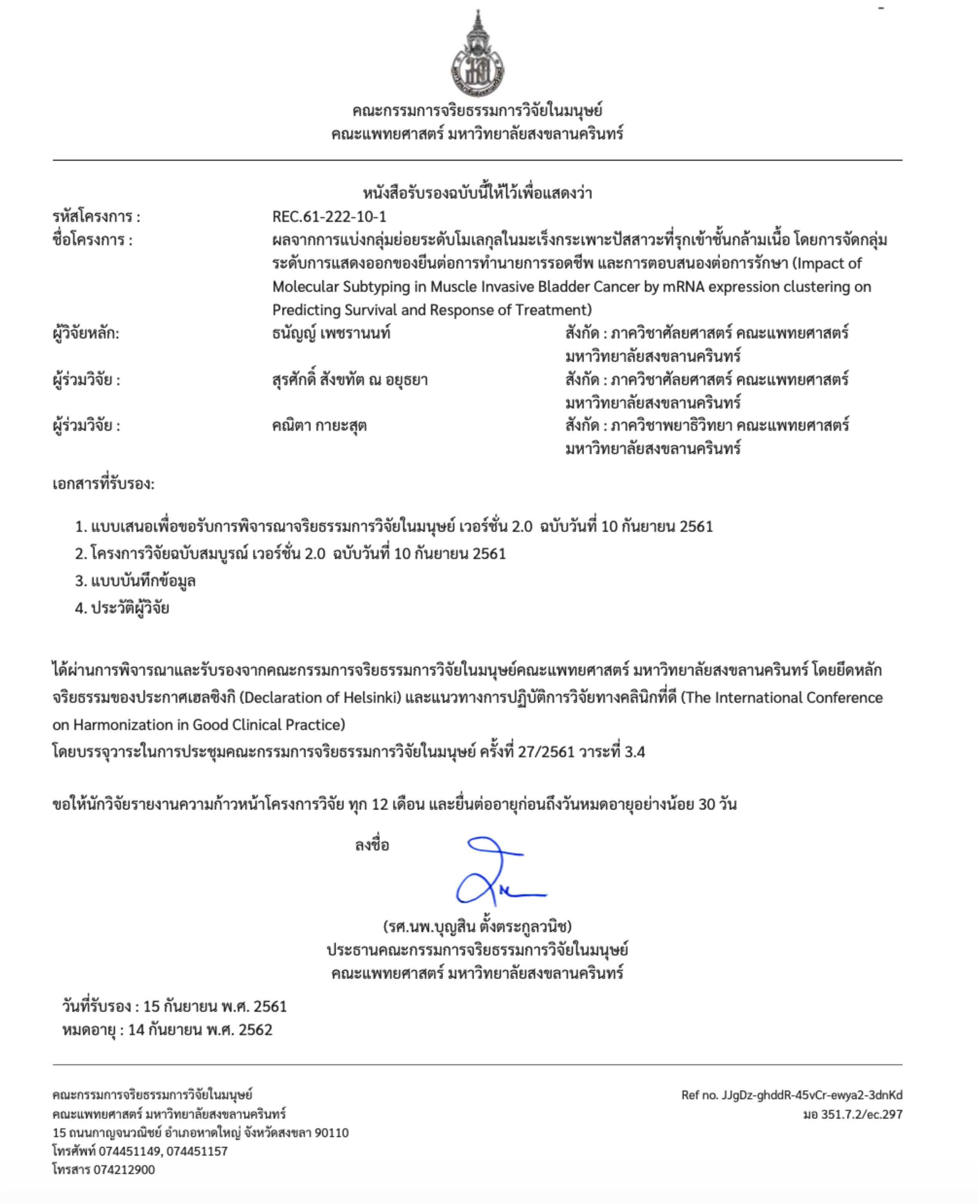
75. Jangir H, Nambirajan A, Seth A, Sahoo RK, Dinda AK, Nayak B, et al. Prognostic stratification of muscle invasive urothelial carcinomas using limited immunohistochemical panel of Gata3 and cytokeratins 5/6, 14 and 20. Ann Diagn Pathol. 2019;43:151397.
76. Naik M, Rao BV, Fonseca D, Murthy SS, Giridhar A, Sharma R, et al. GATA-3 Expression in all Grades and Different Variants of Primary and Metastatic Urothelial Carcinoma. Indian J Surg Oncol. 2021;12(Suppl 1):72-8.
77. Kamel NA, Abdelzaher E, Elgebaly O, Ibrahim SA. Reduced expression of GATA3 predicts progression in non-muscle invasive urothelial carcinoma of the urinary bladder. J Histotechnol. 2020;43(1):21-8.
78. Desai S, Lim SD, Jimenez RE, Chun T, Keane TE, McKenney JK, et al. Relationship of cytokeratin 20 and CD44 protein expression with WHO/ISUP grade in pTa and pT1 papillary urothelial neoplasia. Mod Pathol. 2000;13(12):1315-23.
79. Hashmi AA, Hussain ZF, Irfan M, Edhi MM, Kanwal S, Faridi N, et al. Cytokeratin 5/6 expression in bladder cancer: association with clinicopathologic parameters and prognosis. BMC Res Notes. 2018;11(1):207.
80. Calvete J, Larrinaga G, Errarte P, Martin AM, Dotor A, Esquinas C, et al. The coexpression of fibroblast activation protein (FAP) and basal-type markers (CK 5/6 and CD44) predicts prognosis in high-grade invasive urothelial carcinoma of the bladder. Hum Pathol. 2019;91:61-8.
81. Langner C, Wegscheider BJ, Rehak P, Ratschek M, Zigeuner R. Prognostic value of keratin subtyping in transitional cell carcinoma of the upper urinary tract. Virchows Arch. 2004;445(5):442-8.
82. Akhtar M, Rashid S, Gashir MB, Taha NM, Al Bozom I. CK20 and CK5/6 Immunohistochemical Staining of Urothelial Neoplasms: A Perspective. Adv Urol. 2020 Nov 4;2020:4920236.
83. Jiménez-Jacinto V, Sanchez-Flores A, Vega-Alvarado L. Integrative Differential Expression Analysis for Multiple EXperiments (IDEAMEX): A Web Server Tool for Integrated RNA-Seq Data Analysis. Front Genet. 2019 Mar 29;10:279.

APPENDICES

Appendix A

The ethics committee approval

The ethics committee approval



คณะกรรมการจริยธรรมการวิจัยในมนุษย์
คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์

หนังสือรับรองฉบับนี้ให้ไว้เพื่อแสดงว่า

รหัสโครงการ :	REC.61-222-10-1	
ชื่อโครงการ :	ผลจากการแบ่งกลุ่มย่อยระดับโมเลกุลในมะเร็งกระเพาะปัสสาวะที่รุกเข้าชั้นกล้ามเนื้อ โดยการจัดกลุ่มระดับการแสดงออกของยืนต่อการทำนายการรอดชีพ และการตอบสนองต่อการรักษา (Impact of Molecular Subtyping in Muscle Invasive Bladder Cancer by mRNA expression clustering on Predicting Survival and Response of Treatment)	
ผู้วิจัยหลัก:	ธนัญญ์ เพชรานนท์	สังกัด : ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์
ผู้ร่วมวิจัย :	สุรศักดิ์ สังขทัต ณ อยุธยา	สังกัด : ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์
ผู้ร่วมวิจัย :	คณิตา กายะสุต	สังกัด : ภาควิชาพยาธิวิทยา คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์

เอกสารที่รับรอง:

1. แบบเสนอเพื่อขอรับการพิจารณาจริยธรรมการวิจัยในมนุษย์ เวอร์ชั่น 2.0 ฉบับวันที่ 10 กันยายน 2561
2. โครงการวิจัยฉบับสมบูรณ์ เวอร์ชั่น 2.0 ฉบับวันที่ 10 กันยายน 2561
3. แบบบันทึกข้อมูล
4. ประวัติผู้วิจัย

ได้ผ่านการพิจารณาและรับรองจากคณะกรรมการจริยธรรมการวิจัยในมนุษย์คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์ โดยยึดหลักจริยธรรมของประกาศเฮลซิงกิ (Declaration of Helsinki) และแนวทางการปฏิบัติการวิจัยทางคลินิกที่ดี (The International Conference on Harmonization in Good Clinical Practice)
โดยบรรจุวาระในการประชุมคณะกรรมการจริยธรรมการวิจัยในมนุษย์ ครั้งที่ 27/2561 วาระที่ 3.4

ขอให้นักวิจัยรายงานความก้าวหน้าโครงการวิจัย ทุก 12 เดือน และยื่นต่ออายุก่อนถึงวันหมดอายุอย่างน้อย 30 วัน

ลงชื่อ

(รศ.นพ.บุญสิน ตั้งตระกูลวนิช)
ประธานคณะกรรมการจริยธรรมการวิจัยในมนุษย์
คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์

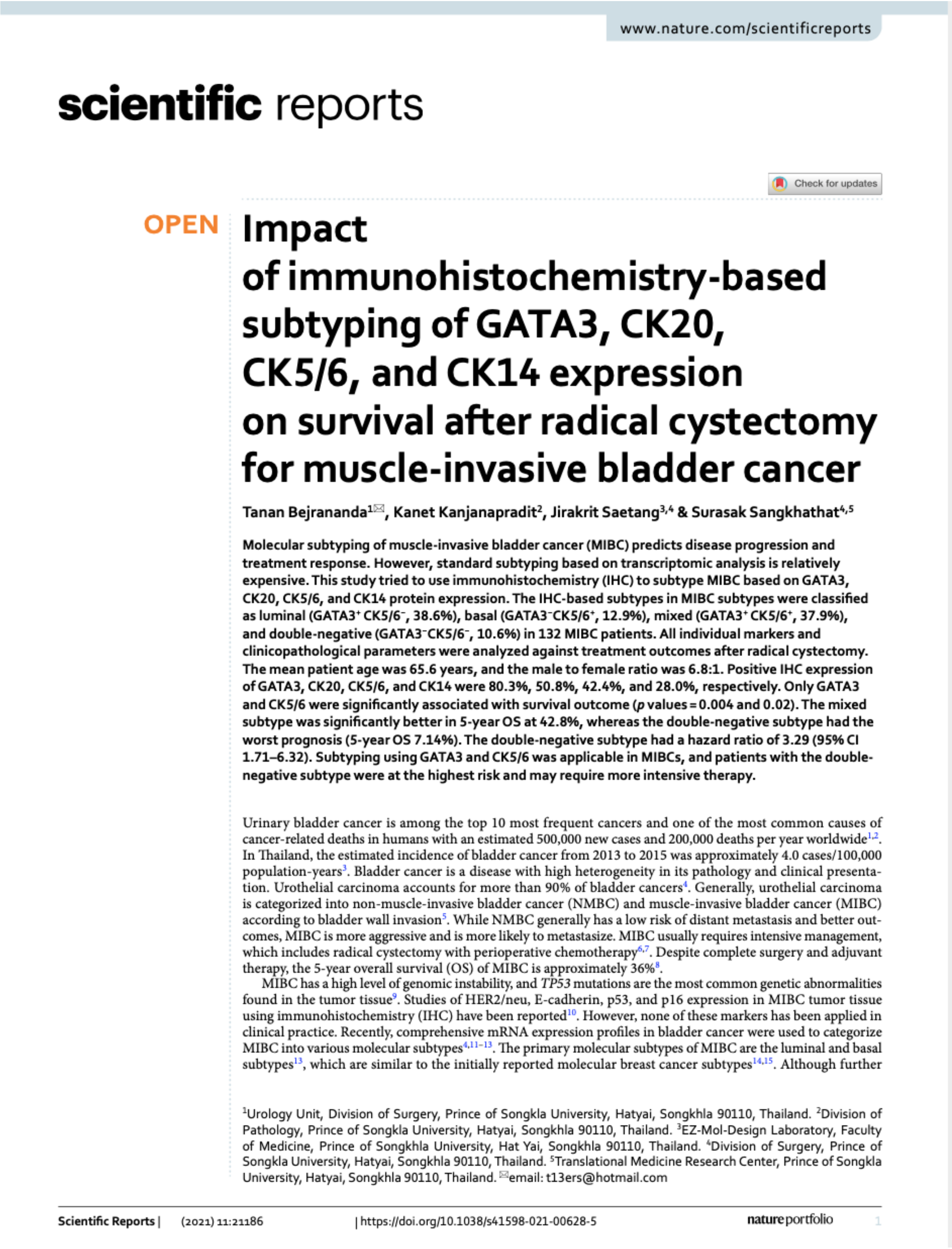
วันที่รับรอง : 15 กันยายน พ.ศ. 2561
หมดอายุ : 14 กันยายน พ.ศ. 2562

คณะกรรมการจริยธรรมการวิจัยในมนุษย์
คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์
15 ถนนกาญจนวณิชย์ อำเภอหาดใหญ่ จังหวัดสงขลา 90110
โทรศัพท์ 074451149, 074451157
โทรสาร 074212900

Ref no. JJgDz-ghddR-45vCr-ewya2-3dnKd
มอ 351.7.2/ec.297

Appendix B

Publication



www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Impact of immunohistochemistry-based subtyping of GATA3, CK20, CK5/6, and CK14 expression on survival after radical cystectomy for muscle-invasive bladder cancer

Tanan Bejrananda[1]✉, Kanet Kanjanapradit[2], Jirakrit Saetang[3,4] & Surasak Sangkhathat[4,5]

Molecular subtyping of muscle-invasive bladder cancer (MIBC) predicts disease progression and treatment response. However, standard subtyping based on transcriptomic analysis is relatively expensive. This study tried to use immunohistochemistry (IHC) to subtype MIBC based on GATA3, CK20, CK5/6, and CK14 protein expression. The IHC-based subtypes in MIBC subtypes were classified as luminal ($GATA3^+$ $CK5/6^-$, 38.6%), basal ($GATA3^-CK5/6^+$, 12.9%), mixed ($GATA3^+$ $CK5/6^+$, 37.9%), and double-negative ($GATA3^-CK5/6^-$, 10.6%) in 132 MIBC patients. All individual markers and clinicopathological parameters were analyzed against treatment outcomes after radical cystectomy. The mean patient age was 65.6 years, and the male to female ratio was 6.8:1. Positive IHC expression of GATA3, CK20, CK5/6, and CK14 were 80.3%, 50.8%, 42.4%, and 28.0%, respectively. Only GATA3 and CK5/6 were significantly associated with survival outcome (*p* values = 0.004 and 0.02). The mixed subtype was significantly better in 5-year OS at 42.8%, whereas the double-negative subtype had the worst prognosis (5-year OS 7.14%). The double-negative subtype had a hazard ratio of 3.29 (95% CI 1.71–6.32). Subtyping using GATA3 and CK5/6 was applicable in MIBCs, and patients with the double-negative subtype were at the highest risk and may require more intensive therapy.

Urinary bladder cancer is among the top 10 most frequent cancers and one of the most common causes of cancer-related deaths in humans with an estimated 500,000 new cases and 200,000 deaths per year worldwide[1,2]. In Thailand, the estimated incidence of bladder cancer from 2013 to 2015 was approximately 4.0 cases/100,000 population-years[3]. Bladder cancer is a disease with high heterogeneity in its pathology and clinical presentation. Urothelial carcinoma accounts for more than 90% of bladder cancers[4]. Generally, urothelial carcinoma is categorized into non-muscle-invasive bladder cancer (NMBC) and muscle-invasive bladder cancer (MIBC) according to bladder wall invasion[5]. While NMBC generally has a low risk of distant metastasis and better outcomes, MIBC is more aggressive and is more likely to metastasize. MIBC usually requires intensive management, which includes radical cystectomy with perioperative chemotherapy[6,7]. Despite complete surgery and adjuvant therapy, the 5-year overall survival (OS) of MIBC is approximately 36%[8].

MIBC has a high level of genomic instability, and *TP53* mutations are the most common genetic abnormalities found in the tumor tissue[9]. Studies of HER2/neu, E-cadherin, p53, and p16 expression in MIBC tumor tissue using immunohistochemistry (IHC) have been reported[10]. However, none of these markers has been applied in clinical practice. Recently, comprehensive mRNA expression profiles in bladder cancer were used to categorize MIBC into various molecular subtypes[4,11–13]. The primary molecular subtypes of MIBC are the luminal and basal subtypes[13], which are similar to the initially reported molecular breast cancer subtypes[14,15]. Although further

[1]Urology Unit, Division of Surgery, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand. [2]Division of Pathology, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand. [3]EZ-Mol-Design Laboratory, Faculty of Medicine, Prince of Songkhla University, Hat Yai, Songkhla 90110, Thailand. [4]Division of Surgery, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand. [5]Translational Medicine Research Center, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand. ✉email: t13ers@hotmail.com

studies have extended the molecular classification of MIBC into 5 or 6 subtypes[16,17], the luminal and basal subtypes remain the fundamental types.

Intrinsic subtypes of bladder cancer have demonstrated increased utility in predicting treatment outcomes, especially in patients with MIBC who undergo radical cystectomy followed by adjuvant chemotherapy[18]. The established molecular subtypes proposed in previous studies were primarily established by high-throughput molecular technology, especially transcriptomic analysis. Traditional IHC techniques are not technology-dependent, and hence it is more feasible to classify subtypes at the protein level using IHC. Discovery of potential IHC markers that can stratify molecular subtypes of MIBC may be useful in the prediction of disease progression. Previous studies have reported a correlation between mRNA expression profiles and IHC staining results in luminal (CK20 expression) and basal (CK5/6 expression) subtypes[11] and also confirmed that GATA3 and CK5/6 expression by IHC may also identify these two subtypes with greater than 90% accuracy according to a meta-analysis[19]. Another study revealed that the basal/squamous-like subtype was correlated with poor clinical outcome[20], as was decreased GATA3 expression[21]. In this study, GATA3, CK20, CK5/6, and CK14 staining was selected to be tested against clinical outcomes with respect to survival after a radical cystectomy. This study aimed to determine IHC markers or patterns that may predict prognosis in patients with MIBC.

Results

Demographic and clinicopathological data.

This study included 132 patients with MIBC who underwent radical cystectomy during the study period. Their mean age was 65.6 years, and the male to female ratio was 6.8:1. The demographic characteristics of the patients and immunoreactivity for each IHC marker are summarized in Table 1. Two patients who died at 3 and 7 days after surgery were considered to have operative mortality and were excluded from the survival analysis. As of January 2021, the median follow-up duration was 125 months (interquartile range 103–154 months). The median OS time was 12.2 months (interquartile range 4.7, 46.4 months), and the 5-year OS was 27.0% (95% CI 19.6%–35.0%). IHC showed positivity for GATA3, CK5/6, CK20, and CK14 with kappa value between 0.799–0.908 (93.2–96.2% agreement) (Table 2). The immunostains for GATA3, CK5/6, CK20, and CK14 showed positive results with 80.3%, 50.8%, 42.4%, and 28.0% of cases, respectively (Fig. 1). GATA3 and CK5/6 immunopositivity was significantly associated with OS by log-rank analysis (Table 1). Twenty-six cases received a median of 3 cycles of adjuvant chemotherapy. GATA3 expression was significantly inversely correlated with pT stage progression.

According to the Kaplan–Meier survival analysis, significant differences in outcomes with respect to OS were demonstrated among cases with positive GATA3 staining ($p = 0.008$) and in cases with positive CK5/6 staining ($p = 0.038$). The other markers did not show significant prognostication value for survival. The Kaplan–Meier survival curves for GATA3, CK5/6, CK14, and CK20 are depicted in Fig. 2.

The correlation between each individual marker was evaluated by Pearson correlation test. As showed in Table 2, the significant association of GATA3, CK5/6, and CK20 was found only in pathological stage 1 of patients. When the correlation between markers in the basal and luminal subtypes was assessed, moderate correlation was observed between GATA3 and CK20 expression, which indicated that the basal-like subtype was demonstrated by Pearson correlation at 0.46 ($p = 0.022$). The analysis showed small correlation between the luminal-like subtype markers, CK5/6 and CK14; 0.31 ($p = 0.048$) (Fig. 3).

As GATA3 and CK5/6 were the only two markers representing different subtypes that were significantly associated with survival, we elected to categorize our cases into four groups according to those two markers: luminal-like (GATA3$^+$ and CK5/6$^-$), basal-like (GATA3$^-$ and CK5/6$^+$), mixed (GATA3$^+$ and CK5/6$^+$), and double-negative (GATA3$^-$ and CK5/6$^-$) subtypes. By this definition, the luminal-like, basal-like, mixed, and double-negative subtypes were observed in 38.6%, 12.9%, 37.9%, and 10.6% of cases, respectively. Associations between each subtype and clinicopathological factors including survival outcomes are displayed in Table 3. The double-negative subtype was significantly associated with higher incidence of pT4 disease.

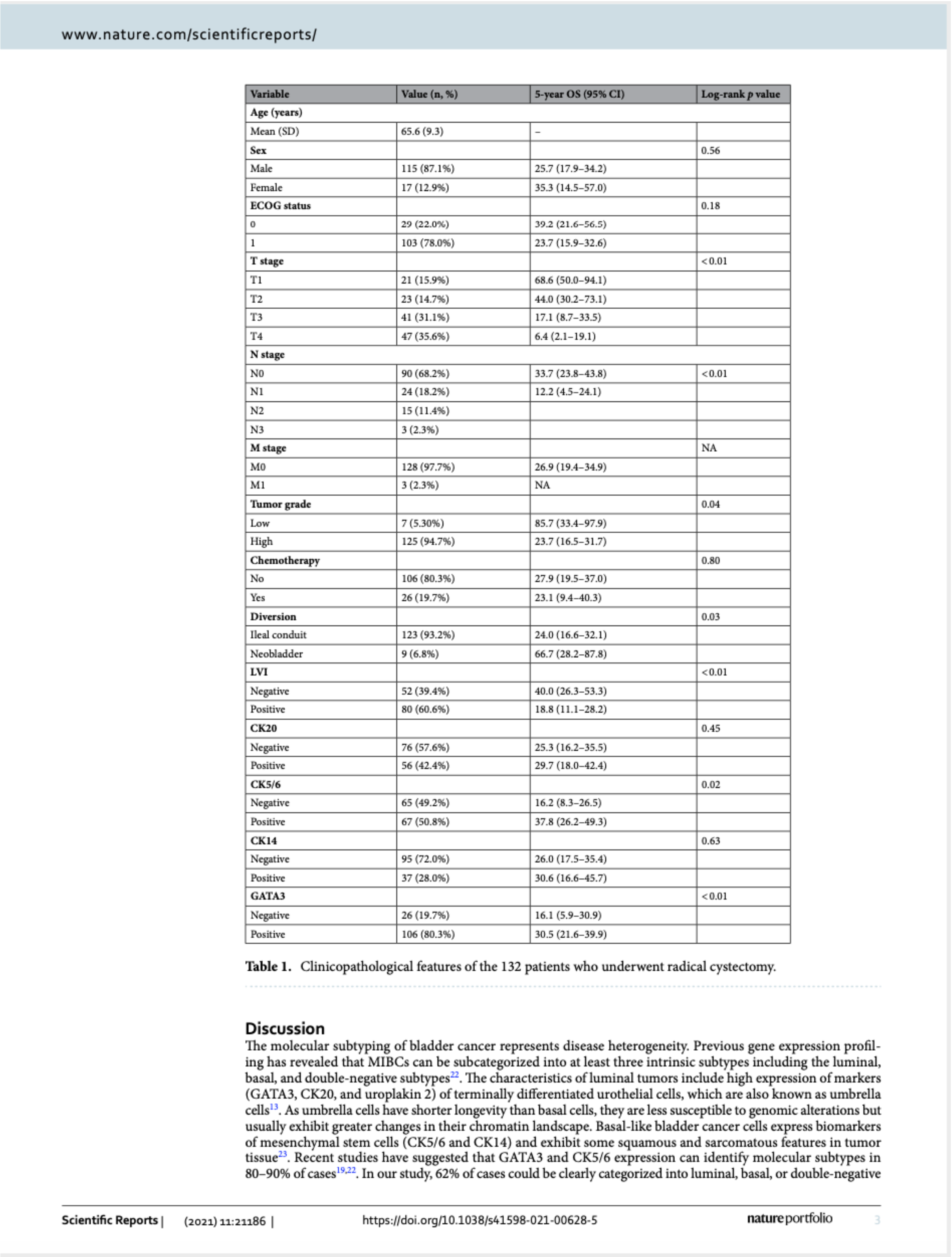
CK20 immunopositivity, a marker of the luminal molecular subtype, was significantly associated with the GATA3-defined luminal subtype ($p < 0.01$), whereas CK14 positivity was significantly associated with the CK5/6-defined basal subtype ($p < 0.01$). In the 50 mixed subtype (GATA3$^+$ and CK2/5$^+$) cases in this study, an equal number of cases with CK20 and CK5/6 positivity was found (Table 3).

When clinicopathological parameters and IHC subtypes were analyzed against survival in a univariable Cox hazard model, tumor stage (pT and N), lymphovascular invasion, pathologic grade, loss of GATA3 immunoreactivity, and loss of CK5/6 immunoreactivity were significantly associated with poorer survival outcomes. Considering subtyping, while patients with the mixed subtype had the lowest risk, which was followed by patients with the luminal-like and basal-like subtypes, those with the double-negative subtype had the highest crude HR. In the multivariable analysis by stepwise Cox hazard regression, N stage (N > 0) and the double-negative subtype were significantly associated with higher risk (model $p = 0.0001$) (Table 4).

Association between molecular subtypes and survival outcomes.

Kaplan–Meier curves compare the survival probability of 132 patients with MIBC following radical surgery (Fig. 4). The 5-year OS rates of patients with the mixed, basal, luminal, and double-negative subtypes were 42.5% (95% CI 28.9–56.1%), 23.5% (7.3–44.9%), 18.9% (9.2–31.1%), and 7.1% (0.4–27.5%), respectively.

In the 50 MIBCs of mixed subtype (GATA3$^+$ and CK5/6$^+$), if CK20 and CK14 were added to the subcategorization criteria, which means that mixed subtype cases with positive CK20 were reclassified as luminal, whereas mixed subtype cases with positive CK14 were reclassified as basal. The univariable hazard model did not improve. Using the mixed subtype as a reference, the HRs (and 95% CIs) for the luminal, basal, and double-negative subtypes were 1.17 (95% CI 0.62–2.20), 1.54 (0.76–3.10), and 2.64 (1.22–5.73) respectively.



Variable	Value (n, %)	5-year OS (95% CI)	Log-rank *p* value
Age (years)			
Mean (SD)	65.6 (9.3)	–	
Sex			0.56
Male	115 (87.1%)	25.7 (17.9–34.2)	
Female	17 (12.9%)	35.3 (14.5–57.0)	
ECOG status			0.18
0	29 (22.0%)	39.2 (21.6–56.5)	
1	103 (78.0%)	23.7 (15.9–32.6)	
T stage			< 0.01
T1	21 (15.9%)	68.6 (50.0–94.1)	
T2	23 (14.7%)	44.0 (30.2–73.1)	
T3	41 (31.1%)	17.1 (8.7–33.5)	
T4	47 (35.6%)	6.4 (2.1–19.1)	
N stage			
N0	90 (68.2%)	33.7 (23.8–43.8)	< 0.01
N1	24 (18.2%)	12.2 (4.5–24.1)	
N2	15 (11.4%)		
N3	3 (2.3%)		
M stage			NA
M0	128 (97.7%)	26.9 (19.4–34.9)	
M1	3 (2.3%)	NA	
Tumor grade			0.04
Low	7 (5.30%)	85.7 (33.4–97.9)	
High	125 (94.7%)	23.7 (16.5–31.7)	
Chemotherapy			0.80
No	106 (80.3%)	27.9 (19.5–37.0)	
Yes	26 (19.7%)	23.1 (9.4–40.3)	
Diversion			0.03
Ileal conduit	123 (93.2%)	24.0 (16.6–32.1)	
Neobladder	9 (6.8%)	66.7 (28.2–87.8)	
LVI			< 0.01
Negative	52 (39.4%)	40.0 (26.3–53.3)	
Positive	80 (60.6%)	18.8 (11.1–28.2)	
CK20			0.45
Negative	76 (57.6%)	25.3 (16.2–35.5)	
Positive	56 (42.4%)	29.7 (18.0–42.4)	
CK5/6			0.02
Negative	65 (49.2%)	16.2 (8.3–26.5)	
Positive	67 (50.8%)	37.8 (26.2–49.3)	
CK14			0.63
Negative	95 (72.0%)	26.0 (17.5–35.4)	
Positive	37 (28.0%)	30.6 (16.6–45.7)	
GATA3			< 0.01
Negative	26 (19.7%)	16.1 (5.9–30.9)	
Positive	106 (80.3%)	30.5 (21.6–39.9)	

Table 1. Clinicopathological features of the 132 patients who underwent radical cystectomy.

Discussion

The molecular subtyping of bladder cancer represents disease heterogeneity. Previous gene expression profiling has revealed that MIBCs can be subcategorized into at least three intrinsic subtypes including the luminal, basal, and double-negative subtypes[22]. The characteristics of luminal tumors include high expression of markers (GATA3, CK20, and uroplakin 2) of terminally differentiated urothelial cells, which are also known as umbrella cells[13]. As umbrella cells have shorter longevity than basal cells, they are less susceptible to genomic alterations but usually exhibit greater changes in their chromatin landscape. Basal-like bladder cancer cells express biomarkers of mesenchymal stem cells (CK5/6 and CK14) and exhibit some squamous and sarcomatous features in tumor tissue[23]. Recent studies have suggested that GATA3 and CK5/6 expression can identify molecular subtypes in 80–90% of cases[19,22]. In our study, 62% of cases could be clearly categorized into luminal, basal, or double-negative

	All	GATA3	CK5/6	CK20	CK14
Positive staining (%)	132	101 (76.5%)	67 (50.8%)	56 (42.4%)	37 (28.0%)
Mean age of positive cases (SD)		64.7 (9.3)	64.7 (9.1)	64.8 (8.7)	65.7 (9.3)
Gender					
Male (%)	115 (87.1%)	91 (90.1%)	56 (83.6%)	50 (89.3%)	30 (81.1%)
Female (%)	17 (12.9%)	10 (9.9%)	11 (16.4%)	6 (10.7%)	7 (18.9%)
ECOG status, n (%)					
0	29 (22.0%)	25 (24.7%)	13 (19.4%)	13 (23.1%)	8 (21.6%)
1	103 (78.0%)	76 (75.3%)	54 (80.6%)	43 (76.8%)	29 (78.3%)
T stage, n (%)					
pT1	21 (15.9%)	20 (19.8%)*	12 (17.9%)	14 (25.0%)*	0 (0.0%)*
pT2	23 (17.4%)	20 (19.8%)	14 (20.9%)	13 (23.2%)	8 (21.6%)
pT3	41 (31.1%)	28 (27.7%)	20 (29.9%)	10 (17.9%)	18 (48.7%)
pT4	47 (35.6%)	33 (32.6%)	21 (31.3%)	19 (33.9%)	11 (29.7%)
N stage, n (%)					
N0	90 (68.2%)	69 (68.3%)	47 (70.2%)	41 (73.2%)	27 (73.0%)
N1	24 (18.2%)	19 (18.8%)	12 (17.9%)	7 (12.5%)	7 (18.9%)
N2	15 (11.4%)	11 (10.9%)	7 (10.5%)	6 (10.7%)	2 (5.4%)
N3	3 (2.3%)	2 (2.0%)	1 (1.5%)	2 (3.6%)	1 (2.3%)
M stage, n (%)					
M0	128 (97.7%)	98 (98.0%)	65 (98.5%)	54 (98.2%)	35 (97.2%)
M1	3 (2.3%)	2 (2.0%)	1 (1.5%)	1 (1.8%)	1 (2.8%)
Tumor grade, n (%)					
Low	7 (5.30%)	6 (5.9%)	6 (9.0%)	3 (5.4%)	0 (0.0%)
High	125 (94.7%)	95 (94.1%)	61 (91.0%)	53 (94.6%)	37 (100.0%)
LVI, n (%)					
Negative	52 (39.4%)	41 (40.6%)	26 (38.8%)	25 (44.6%)	11 (29.7%)
Positive	80 (60.6%)	60 (59.4%)	41 (61.1%)	31 (55.4%)	26 (70.3%)

Table 2. Immunopositivity of the four markers analyzed and their correlation with clinicopathological parameters. *p value < 0.05 when distribution between positive cases and among all cases was compared; LVI: lymphovascular invasion.

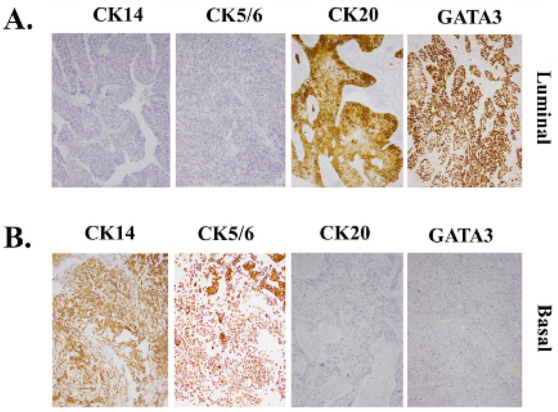


Figure 1. Immunohistochemical staining of MIBC tissues for GATA3, CK20, CK5/6, and CK14. (**A**) Luminal type, (**B**) Basal type.

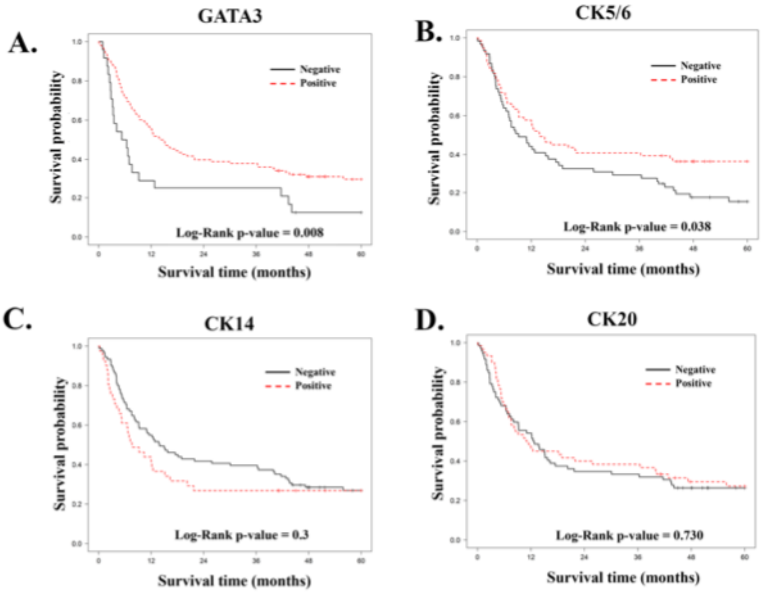


Figure 2. (**A**–**D**) Kaplan–Meier curves demonstrate the survival probability in 132 patients with MIBC according to marker expression by IHC; GATA3 (**A**), CK5/6 (**B**), CK14 (**C**), and CK20 (**D**).

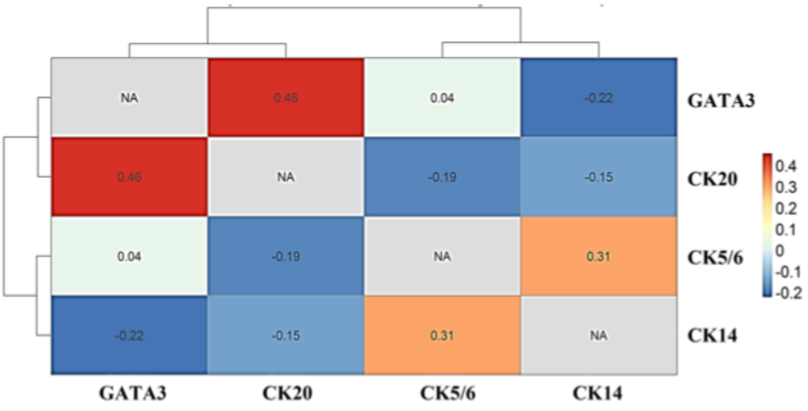


Figure 3. Correlation heatmap of GATA3, CK20 (basal-like markers), CK5/6, and CK14 (luminal-like markers) expression by IHC.

subtypes based on GATA3 and CK5/6 expression, and only these two markers were associated with survival outcome. Loss of expression of either GATA3 or CK5/6 was associated with poorer survival probability, whereas loss of expression of both markers was a strong predictor of poor outcome. Although a significant association was observed between CK20 and GATA3 expression and between CK14 and CK5/6 expression, the addition of CK20 and CK14 to the criteria to categorize the subtypes did not appear to improve survival prediction.

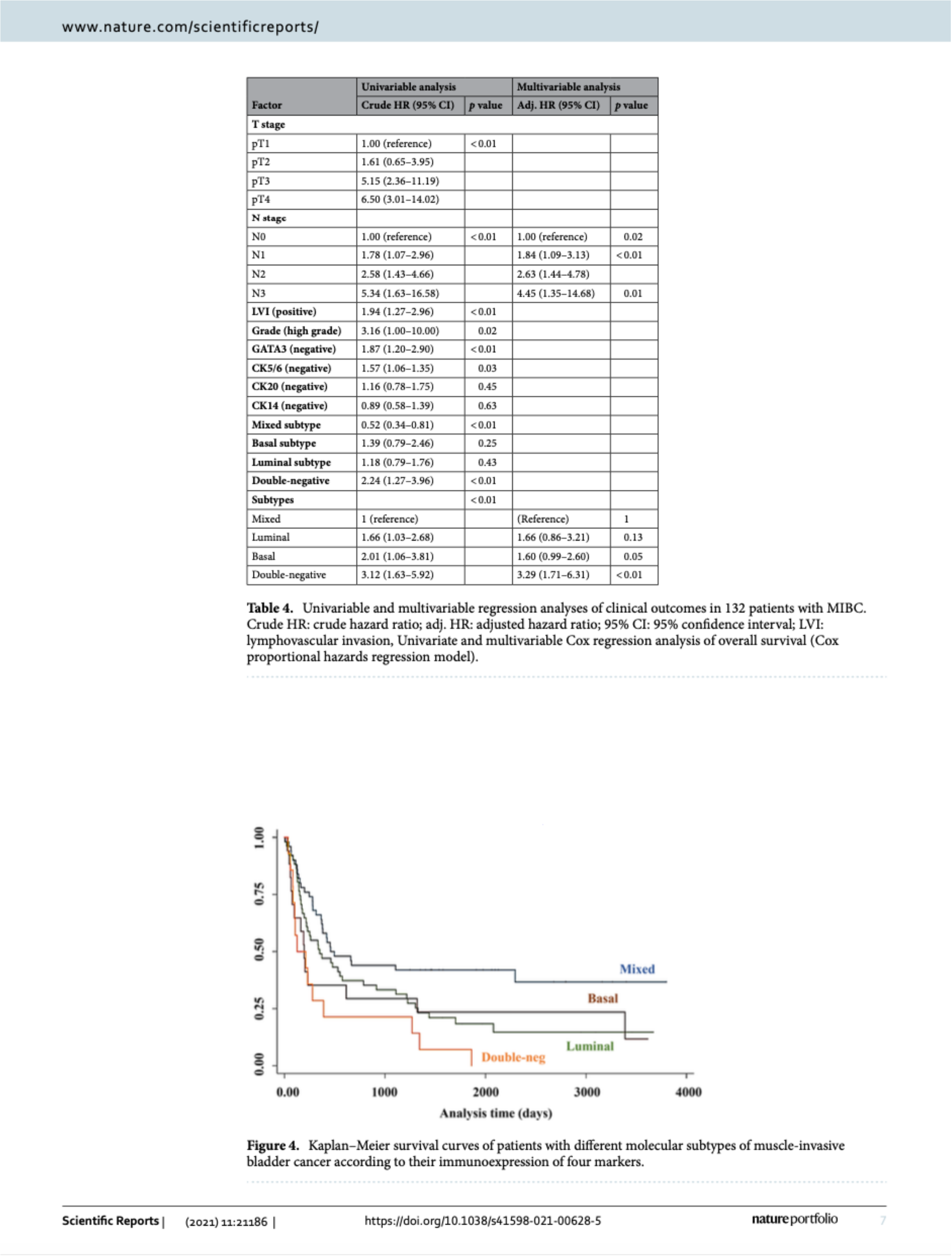
MIBCs expressing GATA3 exhibited less aggressive characteristics and were associated with significantly better survival. GATA3, also known as GATA3 binding protein, is a transcription factor that regulates the expression of genes that function in the luminal differentiation of breast and urothelial epithelium[24,25]. In addition, GATA3 is expressed in T-lymphocytes, the central nervous system, and erythrocytes[26]. In breast cancers, reduced GATA3

	Double-neg	Luminal-like	Basal-like	Mixed	*p* value**
Total (%)	14 (10.6)	51 (38.6)	17 (12.9)	50 (37.9)	–
Mean age (SD)	70.2 (6.0)	65.3 (10.1)	66.6 (10.7)	64.1 (8.6)	0.18
Gender, n (%)					
Male	12 (85.7)	47 (92.2)	12 (70.6)	44 (88.0)	0.15
Female	2 (14.3)	4 (7.8)	5 (29.4)	6 (12.0)	
ECOG status, n (%)					0.29
0	1 (7.1)	15 (29.4)	3 (17.7)	10 (20.0)	
1	13 (92.9)	36 (70.6)	14 (82.3)	40 (80.0)	
Urinary diversion type					0.98
Ileal conduit	13 (92.9)	47 (92.2)	16 (94.1)	47 (94.0)	
Neobladder	1 (7.1)	4 (7.8)	1 (5.9)	3 (6.0)	
T stage, n (%)					0.24
T1	0 (0.0)	9 (17.6)	1 (5.9)	11 (22.0)	
T2	0 (0.0)	9 (17.6)	3 (17.6)	11 (22.0)	
T3	6 (42.9)	15 (29.4)	7 (41.2)	13 (26.0)	
T4	8 (57.1)	18 (35.3)	6 (35.3)	15 (30.0)	
N stage, n (%)					0.71
N0	11 (78.6)	32 (62.8)	10 (58.8)	37 (74.0)	
N1	1 (7.1)	11 (21.6)	4 (23.5)	8 (16.0)	
N2	1 (7.1)	7 (13.7)	3 (17.7)	4 (8.0)	
N3	1 (7.1)	1 (2.0)	0 (0.0)	1 (2.0)	
M stage, n (%)					0.55
M0	13 (92.9)	50 (98.0)	17 (100.0)	48 (98.0)	
M1	1 (7.1)	1 (2.0)	0 (0.0)	1 (2.0)	
Tumor grade, n (%)					0.25
Low	0 (0.0)	1 (2.0)	1 (5.9)	5 (10.0)	
High	14 (100.0)	50 (98.0)	16 (94.1)	45 (90.0)	
Margin, n (%)					0.57
Negative	11 (78.6)	46 (90.2)	16 (94.1)	43 (86.0)	
Positive	3 (21.4)	5 (9.8)	1 (5.9)	7 (14.0)	
LVI, n (%)					0.97
Negative	5 (35.7)	21 (41.2)	6 (35.3)	20 (40.0)	
Positive	9 (64.3)	30 (58.8)	11 (64.7)	30 (60.0)	
CK20, n (%)					<0.01
Negative	12 (85.7)	17 (33.3)	16 (94.1)	31 (62.0)	
Positive	2 (14.3)	34 (66.7)	1 (5.9)	19 (38.0)	
CK14, n (%)					<0.01
Negative	12 (85.7)	48 (94.1)	4 (23.5)	31 (62.0)	
Positive	2 (14.3)	3 (5.9)	13 (76.5)	19 (38.0)	
5-Year OS (%) (95% confidence interval)	7.14 (0.4–27.5)	18.9 (9.2–31.1)	23.5 (7.3–24.9)	42.8 (28.9–56.1)	<0.01

Table 3. Patient characteristics and classification by IHC subtype according to GATA3 and CK5/6 expression. **p* value by Chi-square or Fisher's exact test; ECOG status: Eastern Cooperative Oncology Group performance status; LVI: lymphovascular invasion; OS: overall survival.

expression was reported in the triple-negative subtype[27]. In urothelial cell line models, the loss of GATA3 expression promoted tumor cell migration and invasion via upregulation of oncogenes[28,29]. Several clinical studies of GATA3 in bladder cancer have been conducted[30–32]. Loss of GATA3 expression was associated with high-grade cancer[32], and patients with GATA3-negative bladder cancer had poorer survival outcomes in most studies[30,31,33]. Taken together, those reports and our data demonstrate that GATA3 is a promising biomarker of MIBC. Another luminal marker evaluated in this study, CK20, has been reported to be correlated with higher tumor grade and stage in papillary urothelial carcinoma[34]. However, our study did not reveal a significant association between CK20 and any clinicopathological factor or survival outcome.

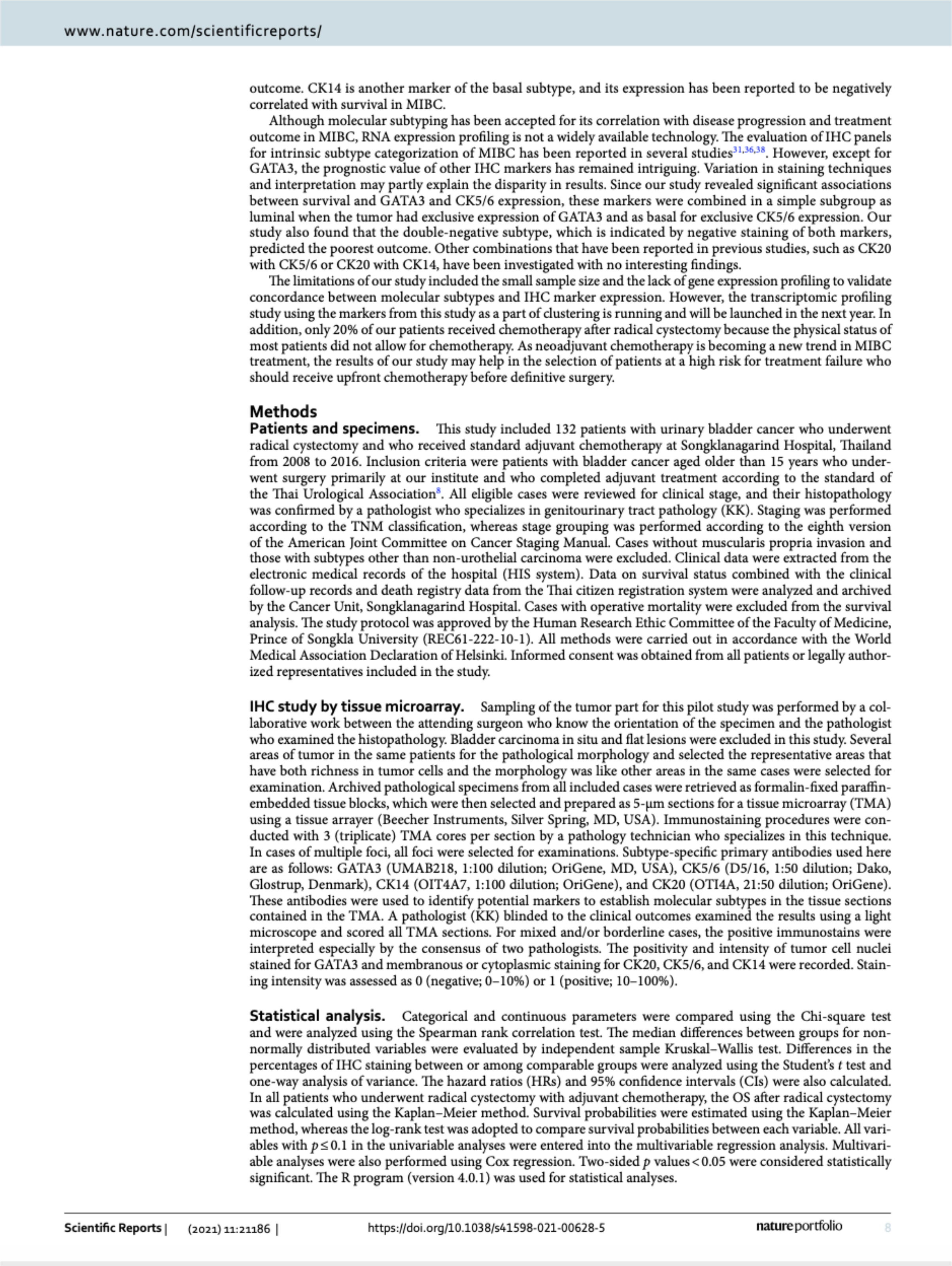
CK5/6 is a cytokeratin expressed in a squamous epithelial lineage and is generally used as a marker of squamous differentiation, which indicates the basal subtype[35]. Expression of CK5/6 in urothelial carcinoma was associated with poorer survival in several reports[35,36]. In contrast, some reports also demonstrated that loss of CK5/6 expression was associated with decreased survival probability in patients with transitional cell carcinoma of the upper urinary tract[36,37]. In our study, loss of CK5/6 expression was associated with significantly worse survival



Factor	Univariable analysis Crude HR (95% CI)	 *p* value	Multivariable analysis Adj. HR (95% CI)	 *p* value
T stage				
pT1	1.00 (reference)	< 0.01		
pT2	1.61 (0.65–3.95)			
pT3	5.15 (2.36–11.19)			
pT4	6.50 (3.01–14.02)			
N stage				
N0	1.00 (reference)	< 0.01	1.00 (reference)	0.02
N1	1.78 (1.07–2.96)		1.84 (1.09–3.13)	< 0.01
N2	2.58 (1.43–4.66)		2.63 (1.44–4.78)	
N3	5.34 (1.63–16.58)		4.45 (1.35–14.68)	0.01
LVI (positive)	1.94 (1.27–2.96)	< 0.01		
Grade (high grade)	3.16 (1.00–10.00)	0.02		
GATA3 (negative)	1.87 (1.20–2.90)	< 0.01		
CK5/6 (negative)	1.57 (1.06–1.35)	0.03		
CK20 (negative)	1.16 (0.78–1.75)	0.45		
CK14 (negative)	0.89 (0.58–1.39)	0.63		
Mixed subtype	0.52 (0.34–0.81)	< 0.01		
Basal subtype	1.39 (0.79–2.46)	0.25		
Luminal subtype	1.18 (0.79–1.76)	0.43		
Double-negative	2.24 (1.27–3.96)	< 0.01		
Subtypes		< 0.01		
Mixed	1 (reference)		(Reference)	1
Luminal	1.66 (1.03–2.68)		1.66 (0.86–3.21)	0.13
Basal	2.01 (1.06–3.81)		1.60 (0.99–2.60)	0.05
Double-negative	3.12 (1.63–5.92)		3.29 (1.71–6.31)	< 0.01

Table 4. Univariable and multivariable regression analyses of clinical outcomes in 132 patients with MIBC. Crude HR: crude hazard ratio; adj. HR: adjusted hazard ratio; 95% CI: 95% confidence interval; LVI: lymphovascular invasion, Univariate and multivariable Cox regression analysis of overall survival (Cox proportional hazards regression model).

Figure 4. Kaplan–Meier survival curves of patients with different molecular subtypes of muscle-invasive bladder cancer according to their immunoexpression of four markers.



outcome. CK14 is another marker of the basal subtype, and its expression has been reported to be negatively correlated with survival in MIBC.

Although molecular subtyping has been accepted for its correlation with disease progression and treatment outcome in MIBC, RNA expression profiling is not a widely available technology. The evaluation of IHC panels for intrinsic subtype categorization of MIBC has been reported in several studies[31,36,38]. However, except for GATA3, the prognostic value of other IHC markers has remained intriguing. Variation in staining techniques and interpretation may partly explain the disparity in results. Since our study revealed significant associations between survival and GATA3 and CK5/6 expression, these markers were combined in a simple subgroup as luminal when the tumor had exclusive expression of GATA3 and as basal for exclusive CK5/6 expression. Our study also found that the double-negative subtype, which is indicated by negative staining of both markers, predicted the poorest outcome. Other combinations that have been reported in previous studies, such as CK20 with CK5/6 or CK20 with CK14, have been investigated with no interesting findings.

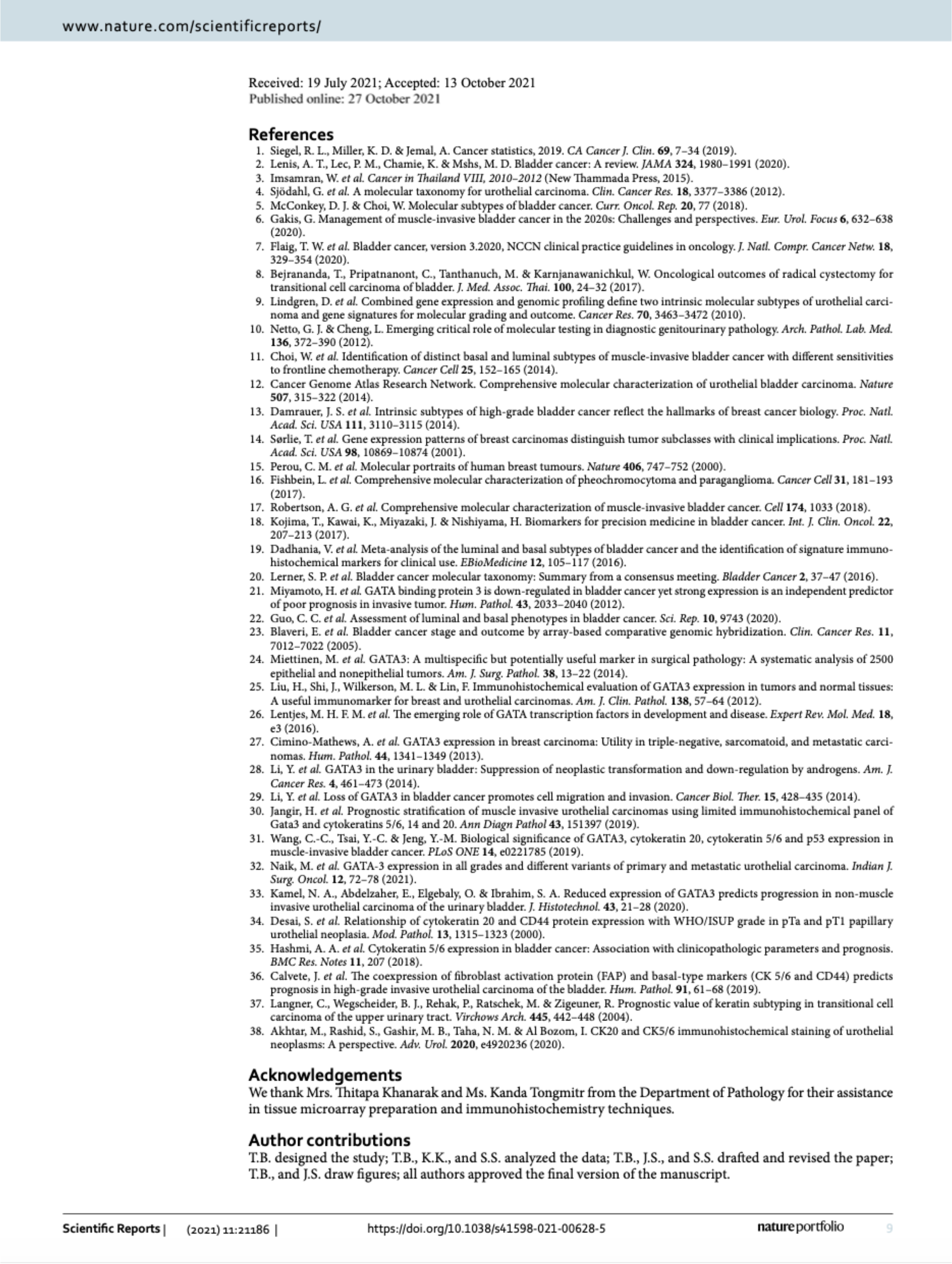
The limitations of our study included the small sample size and the lack of gene expression profiling to validate concordance between molecular subtypes and IHC marker expression. However, the transcriptomic profiling study using the markers from this study as a part of clustering is running and will be launched in the next year. In addition, only 20% of our patients received chemotherapy after radical cystectomy because the physical status of most patients did not allow for chemotherapy. As neoadjuvant chemotherapy is becoming a new trend in MIBC treatment, the results of our study may help in the selection of patients at a high risk for treatment failure who should receive upfront chemotherapy before definitive surgery.

Methods

Patients and specimens. This study included 132 patients with urinary bladder cancer who underwent radical cystectomy and who received standard adjuvant chemotherapy at Songklanagarind Hospital, Thailand from 2008 to 2016. Inclusion criteria were patients with bladder cancer aged older than 15 years who underwent surgery primarily at our institute and who completed adjuvant treatment according to the standard of the Thai Urological Association[8]. All eligible cases were reviewed for clinical stage, and their histopathology was confirmed by a pathologist who specializes in genitourinary tract pathology (KK). Staging was performed according to the TNM classification, whereas stage grouping was performed according to the eighth version of the American Joint Committee on Cancer Staging Manual. Cases without muscularis propria invasion and those with subtypes other than non-urothelial carcinoma were excluded. Clinical data were extracted from the electronic medical records of the hospital (HIS system). Data on survival status combined with the clinical follow-up records and death registry data from the Thai citizen registration system were analyzed and archived by the Cancer Unit, Songklanagarind Hospital. Cases with operative mortality were excluded from the survival analysis. The study protocol was approved by the Human Research Ethic Committee of the Faculty of Medicine, Prince of Songkla University (REC61-222-10-1). All methods were carried out in accordance with the World Medical Association Declaration of Helsinki. Informed consent was obtained from all patients or legally authorized representatives included in the study.

IHC study by tissue microarray. Sampling of the tumor part for this pilot study was performed by a collaborative work between the attending surgeon who know the orientation of the specimen and the pathologist who examined the histopathology. Bladder carcinoma in situ and flat lesions were excluded in this study. Several areas of tumor in the same patients for the pathological morphology and selected the representative areas that have both richness in tumor cells and the morphology was like other areas in the same cases were selected for examination. Archived pathological specimens from all included cases were retrieved as formalin-fixed paraffin-embedded tissue blocks, which were then selected and prepared as 5-μm sections for a tissue microarray (TMA) using a tissue arrayer (Beecher Instruments, Silver Spring, MD, USA). Immunostaining procedures were conducted with 3 (triplicate) TMA cores per section by a pathology technician who specializes in this technique. In cases of multiple foci, all foci were selected for examinations. Subtype-specific primary antibodies used here are as follows: GATA3 (UMAB218, 1:100 dilution; OriGene, MD, USA), CK5/6 (D5/16, 1:50 dilution; Dako, Glostrup, Denmark), CK14 (OIT4A7, 1:100 dilution; OriGene), and CK20 (OTI4A, 21:50 dilution; OriGene). These antibodies were used to identify potential markers to establish molecular subtypes in the tissue sections contained in the TMA. A pathologist (KK) blinded to the clinical outcomes examined the results using a light microscope and scored all TMA sections. For mixed and/or borderline cases, the positive immunostains were interpreted especially by the consensus of two pathologists. The positivity and intensity of tumor cell nuclei stained for GATA3 and membranous or cytoplasmic staining for CK20, CK5/6, and CK14 were recorded. Staining intensity was assessed as 0 (negative; 0–10%) or 1 (positive; 10–100%).

Statistical analysis. Categorical and continuous parameters were compared using the Chi-square test and were analyzed using the Spearman rank correlation test. The median differences between groups for non-normally distributed variables were evaluated by independent sample Kruskal–Wallis test. Differences in the percentages of IHC staining between or among comparable groups were analyzed using the Student's *t* test and one-way analysis of variance. The hazard ratios (HRs) and 95% confidence intervals (CIs) were also calculated. In all patients who underwent radical cystectomy with adjuvant chemotherapy, the OS after radical cystectomy was calculated using the Kaplan–Meier method. Survival probabilities were estimated using the Kaplan–Meier method, whereas the log-rank test was adopted to compare survival probabilities between each variable. All variables with $p \leq 0.1$ in the univariable analyses were entered into the multivariable regression analysis. Multivariable analyses were also performed using Cox regression. Two-sided p values < 0.05 were considered statistically significant. The R program (version 4.0.1) was used for statistical analyses.



Received: 19 July 2021; Accepted: 13 October 2021
Published online: 27 October 2021

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.* **69**, 7–34 (2019).
2. Lenis, A. T., Lec, P. M., Chamie, K. & Mshs, M. D. Bladder cancer: A review. *JAMA* **324**, 1980–1991 (2020).
3. Imsamran, W. *et al. Cancer in Thailand VIII, 2010–2012* (New Thammada Press, 2015).
4. Sjödahl, G. *et al.* A molecular taxonomy for urothelial carcinoma. *Clin. Cancer Res.* **18**, 3377–3386 (2012).
5. McConkey, D. J. & Choi, W. Molecular subtypes of bladder cancer. *Curr. Oncol. Rep.* **20**, 77 (2018).
6. Gakis, G. Management of muscle-invasive bladder cancer in the 2020s: Challenges and perspectives. *Eur. Urol. Focus* **6**, 632–638 (2020).
7. Flaig, T. W. *et al.* Bladder cancer, version 3.2020, NCCN clinical practice guidelines in oncology. *J. Natl. Compr. Cancer Netw.* **18**, 329–354 (2020).
8. Bejrananda, T., Pripatnanont, C., Tanthanuch, M. & Karnjanawanichkul, W. Oncological outcomes of radical cystectomy for transitional cell carcinoma of bladder. *J. Med. Assoc. Thai.* **100**, 24–32 (2017).
9. Lindgren, D. *et al.* Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome. *Cancer Res.* **70**, 3463–3472 (2010).
10. Netto, G. J. & Cheng, L. Emerging critical role of molecular testing in diagnostic genitourinary pathology. *Arch. Pathol. Lab. Med.* **136**, 372–390 (2012).
11. Choi, W. *et al.* Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell* **25**, 152–165 (2014).
12. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
13. Damrauer, J. S. *et al.* Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci. USA* **111**, 3110–3115 (2014).
14. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869–10874 (2001).
15. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
16. Fishbein, L. *et al.* Comprehensive molecular characterization of pheochromocytoma and paraganglioma. *Cancer Cell* **31**, 181–193 (2017).
17. Robertson, A. G. *et al.* Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* **174**, 1033 (2018).
18. Kojima, T., Kawai, K., Miyazaki, J. & Nishiyama, H. Biomarkers for precision medicine in bladder cancer. *Int. J. Clin. Oncol.* **22**, 207–213 (2017).
19. Dadhania, V. *et al.* Meta-analysis of the luminal and basal subtypes of bladder cancer and the identification of signature immunohistochemical markers for clinical use. *EBioMedicine* **12**, 105–117 (2016).
20. Lerner, S. P. *et al.* Bladder cancer molecular taxonomy: Summary from a consensus meeting. *Bladder Cancer* **2**, 37–47 (2016).
21. Miyamoto, H. *et al.* GATA binding protein 3 is down-regulated in bladder cancer yet strong expression is an independent predictor of poor prognosis in invasive tumor. *Hum. Pathol.* **43**, 2033–2040 (2012).
22. Guo, C. C. *et al.* Assessment of luminal and basal phenotypes in bladder cancer. *Sci. Rep.* **10**, 9743 (2020).
23. Blaveri, E. *et al.* Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clin. Cancer Res.* **11**, 7012–7022 (2005).
24. Miettinen, M. *et al.* GATA3: A multispecific but potentially useful marker in surgical pathology: A systematic analysis of 2500 epithelial and nonepithelial tumors. *Am. J. Surg. Pathol.* **38**, 13–22 (2014).
25. Liu, H., Shi, J., Wilkerson, M. L. & Lin, F. Immunohistochemical evaluation of GATA3 expression in tumors and normal tissues: A useful immunomarker for breast and urothelial carcinomas. *Am. J. Clin. Pathol.* **138**, 57–64 (2012).
26. Lentjes, M. H. F. M. *et al.* The emerging role of GATA transcription factors in development and disease. *Expert Rev. Mol. Med.* **18**, e3 (2016).
27. Cimino-Mathews, A. *et al.* GATA3 expression in breast carcinoma: Utility in triple-negative, sarcomatoid, and metastatic carcinomas. *Hum. Pathol.* **44**, 1341–1349 (2013).
28. Li, Y. *et al.* GATA3 in the urinary bladder: Suppression of neoplastic transformation and down-regulation by androgens. *Am. J. Cancer Res.* **4**, 461–473 (2014).
29. Li, Y. *et al.* Loss of GATA3 in bladder cancer promotes cell migration and invasion. *Cancer Biol. Ther.* **15**, 428–435 (2014).
30. Jangir, H. *et al.* Prognostic stratification of muscle invasive urothelial carcinomas using limited immunohistochemical panel of Gata3 and cytokeratins 5/6, 14 and 20. *Ann Diagn Pathol* **43**, 151397 (2019).
31. Wang, C.-C., Tsai, Y.-C. & Jeng, Y.-M. Biological significance of GATA3, cytokeratin 20, cytokeratin 5/6 and p53 expression in muscle-invasive bladder cancer. *PLoS ONE* **14**, e0221785 (2019).
32. Naik, M. *et al.* GATA-3 expression in all grades and different variants of primary and metastatic urothelial carcinoma. *Indian J. Surg. Oncol.* **12**, 72–78 (2021).
33. Kamel, N. A., Abdelzaher, E., Elgebaly, O. & Ibrahim, S. A. Reduced expression of GATA3 predicts progression in non-muscle invasive urothelial carcinoma of the urinary bladder. *J. Histotechnol.* **43**, 21–28 (2020).
34. Desai, S. *et al.* Relationship of cytokeratin 20 and CD44 protein expression with WHO/ISUP grade in pTa and pT1 papillary urothelial neoplasia. *Mod. Pathol.* **13**, 1315–1323 (2000).
35. Hashmi, A. A. *et al.* Cytokeratin 5/6 expression in bladder cancer: Association with clinicopathologic parameters and prognosis. *BMC Res. Notes* **11**, 207 (2018).
36. Calvete, J. *et al.* The coexpression of fibroblast activation protein (FAP) and basal-type markers (CK 5/6 and CD44) predicts prognosis in high-grade invasive urothelial carcinoma of the bladder. *Hum. Pathol.* **91**, 61–68 (2019).
37. Langner, C., Wegscheider, B. J., Rehak, P., Ratschek, M. & Zigeuner, R. Prognostic value of keratin subtyping in transitional cell carcinoma of the upper urinary tract. *Virchows Arch.* **445**, 442–448 (2004).
38. Akhtar, M., Rashid, S., Gashir, M. B., Taha, N. M. & Al Bozom, I. CK20 and CK5/6 immunohistochemical staining of urothelial neoplasms: A perspective. *Adv. Urol.* **2020**, e4920236 (2020).

Acknowledgements

We thank Mrs. Thitapa Khanarak and Ms. Kanda Tongmitr from the Department of Pathology for their assistance in tissue microarray preparation and immunohistochemistry techniques.

Author contributions

T.B. designed the study; T.B., K.K., and S.S. analyzed the data; T.B., J.S., and S.S. drafted and revised the paper; T.B., and J.S. draw figures; all authors approved the final version of the manuscript.

Funding

The study was supported by a grant-in-aid from the Health Systems Research Institute and the Faculty of Medicine, Prince of Songkla University.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix C

Shell command: Alignment and transcript count

Appendix C

Shell command: Alignment and transcript count

```
## STAR 1 PASS ##
/home/user/software/STAR-2.7.8a/source/STAR \
--genomeDir ${reference_tcga}/human_grch38_star2.7.8 \
--readFilesIn ${out_trim}/${sample_name}_1_pair.fastq.gz
${out_trim}/${sample_name}_2_pair.fastq.gz \
--runThreadN $NSLOTS \
--outFilterMultimapScoreRange 1 \
--outFilterMultimapNmax 20 \
--outFilterMismatchNmax 10 \
--alignIntronMax 500000 \
--alignMatesGapMax 1000000 \
--sjdbScore 2 \
--alignSJDBoverhangMin 1 \
--genomeLoad NoSharedMemory \
--readFilesCommand zcat \
--outFilterMatchNminOverLread 0.33 \
--outFilterScoreMinOverLread 0.33 \
--sjdbOverhang 100 \
--outSAMstrandField intronMotif \
--outSAMtype None \
--outSAMmode None \
--outFileNamePrefix ${out_align}/${sample_name}.
```

```
## STAR GENOME GENERATE
/home/user/software/STAR-2.7.8a/source/STAR \
--runMode genomeGenerate \
--genomeDir ${out_align}/genome \
--genomeFastaFiles ${reference_tcga}/GRCh38.d1.vd1.fa \
--sjdbOverhang 100 \
--runThreadN $NSLOTS \
--sjdbFileChrStartEnd ${out_align}/${sample_name}.SJ.out.tab \
--outFileNamePrefix ${out_align}/${sample_name}.
```

```
## STAR 2 PASS
/home/user/software/STAR-2.7.8a/source/STAR \
--genomeDir ${out_align}/genome \
--readFilesIn ${out_trim}/${sample_name}_1_pair.fastq.gz
${out_trim}/${sample_name}_2_pair.fastq.gz \
--runThreadN $NSLOTS \
--outFilterMultimapScoreRange 1 \
--outFilterMultimapNmax 20 \
--outFilterMismatchNmax 10 \
--alignIntronMax 500000 \
--alignMatesGapMax 1000000 \
--sjdbScore 2 \
--alignSJDBoverhangMin 1 \
--genomeLoad NoSharedMemory \
--limitBAMsortRAM 0 \
--readFilesCommand zcat \
--outFilterMatchNminOverLread 0.33 \
--outFilterScoreMinOverLread 0.33 \
--sjdbOverhang 100 \
--outSAMstrandField intronMotif \
--outSAMattributes NH HI NM MD AS XS \
--outSAMunmapped Within \
--outSAMtype BAM SortedByCoordinate \
--outSAMheaderHD @HD VN:1.4 \
--outSAMattrRGline ID:MIBC \
--outFileNamePrefix ${out_align}/${sample_name}.

## HTseq count
htseq-count \
-f bam \
-r name \
-s no \
-a 10 \
-t exon \
-i gene_id \
-m intersection-nonempty \
${out_align}/${sample_name}.Aligned.sortedByCoord.out.bam \
${reference_tcga}/gencode.v22.annotation.gtf > \
${out_count}/${sample_name}_htseq.count
```

Appendix D

R code for analysis

Appendix D

R code for analysis

```
##Load library
library(data.table)
library(DESeq2)

library(ggplot2)

library(dplyr)
library(tidyverse)

library(biomaRt)
library(httr)

library(ggvenn)

library(enrichR)

library(NMF)
library(grid)
library(gridExtra)
library(ggrepel)

library(ClusterR)

library(survival)
library(survminer)

library(TCGAbiolinks)
library(DT)
library(TCGAutils)
#BiocManager::install("")
#install.packages("")
library(pROC)
####################################################################################
################################################
##Set directory
setwd("~/bladder/DEG_2pass/use")
##Load data
list.files <- list.files(path = ".")
##Get file name
file_name <- NULL
for (i in list.files) {
  file_name <- c(file_name,gsub('*_htseq.count','\\1',i))
}
rm(i)
##Loop create data frame
```

```
i <- 1
for (j in list.files ) {
  if (i == 1) {
    file <- fread(j,header = FALSE)
    file <- file[-c(60484,60485,60486,60487,60488),]
    gsub('*_htseq.count','\\1',j)
    ## Add first sample
    #read.counts <- data.frame(file$expected_count,row.names = file$gene_id)
    read.counts <- data.frame(file$V2,row.names = file$V1)
    colnames(read.counts) <- gsub('*_htseq.count','\\1',j)
    ## Set condition
    #sample.info <- data.frame(gsub('*_htseq.count','\\1',j),"normal")
    sample.info <- data.frame(gsub('*_htseq.count','\\1',j),"cancer")
    colnames(sample.info) <- c("name","condition")
  } else {
    ## Add another sample ##
    file <- fread(j, header = FALSE)
    file <- file[-c(60484,60485,60486,60487,60488),]
    ## Set sample name ##
    sample <- gsub('*_htseq.count','\\1',j)
    ## Set condition
    condition <- "cancer"
    ## Insert sample to sample.info and read.counts ##
    count <- data.frame(file$V2)
    colnames(count) <- sample
    info <- data.frame(sample,condition)
    colnames(info) <- c("name","condition")
    read.counts <- cbind(read.counts,count)
    sample.info <- rbind(sample.info,info)
  }
  i <- i + 1
}
rm(count)
rm(sample)
rm(info)
rm(condition)
rm(i)
rm(j)
rm(file)
###########################################################################
###########################################
read.counts$gene_id <- rownames(read.counts)
## Annotate with Ensemble gene symbol
list <- NULL
for (i in 1:nrow(read.counts) ) {
  list <- c(list, unlist(strsplit(read.counts[i,"gene_id"], split = "[.]"))[1])
}
read.counts$gene_id_new <- list
##Annotate gene
httr::set_config(httr::config(ssl_cipher_list = "DEFAULT@SECLEVEL=1"))
```

```
httr::set_config(httr::config(ssl_verifypeer = FALSE))
## Annotate gene name ##
ensembl = useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl",
mirror="uswest")
genemap <- getBM( attributes = c("ensembl_gene_id", "entrezgene_id",
                          "hgnc_symbol", "external_gene_name",
                          "description", "chromosome_name",
                          "strand"),
               filters = "ensembl_gene_id",
               values = read.counts$gene_id_new,
               mart = ensembl)
##Create gene filter data frame
gene_filter <- genemap %>% filter(str_detect(genemap$description, "pseudogene") |
                            str_detect(genemap$description, "antisense")  |
                            str_detect(genemap$description, "long intergenic") )
##Remove column
read.counts <- read.counts %>% dplyr::select(-c(gene_id,gene_id_new))
################################################################################
##############################################
##FILTER none mRNA gene
read.counts.test <- read.counts
colnames(read.counts.test)
read.counts.test$gene_id <- rownames(read.counts.test)
list <- NULL
for (i in 1:nrow(read.counts.test) ) {
  list <- c(list, unlist(strsplit(read.counts.test[i,"gene_id"], split = "[.]"))[1])
}
read.counts.test$gene_id_new <- list
genemap.test <- getBM( attributes = c("ensembl_gene_id", "hgnc_symbol"),
                  filters = "ensembl_gene_id",
                  values = read.counts.test$gene_id_new,
                  mart = ensembl)
read.counts.test <- read.counts.test %>% left_join(genemap.test, by = c("gene_id_new" =
"ensembl_gene_id"))
read.counts.test <- read.counts.test %>% filter( !(read.counts.test$gene_id_new %in%
gene_filter$ensembl_gene_id) )
read.counts.test <- read.counts.test[!(is.na(read.counts.test$hgnc_symbol) |
read.counts.test$hgnc_symbol==""), ]
read.counts.test <- distinct_at(read.counts.test, vars(gene_id), .keep_all = TRUE)
rownames(read.counts.test) <- read.counts.test$gene_id
read.counts.test <- read.counts.test %>% dplyr::select(-
c(gene_id,gene_id_new,hgnc_symbol))
################################################################################
##############################################
## UNSUPERVISE CLUSTERING ##
# With no condition #
DESeq.ds.test <- DESeqDataSetFromMatrix(read.counts.test ,colData = sample.info,~ 1)
## Include gene which counts more than 0 ##
DESeq.ds.test <- DESeq.ds.test[ rowSums(counts(DESeq.ds.test)) > 10, ] #10
## Calculate size factor of normalization counts ##
```

```
DESeq.ds.test <- estimateSizeFactors(DESeq.ds.test)
## Normalization and Log2 transform by DESeq2 ##
DESeq.rlog.test  <- vst(DESeq.ds.test , blind = TRUE)
## Summarized object data to matrix ##
rlog.norm.counts.test  <- assay(DESeq.rlog.test)
rlog.norm.counts.test <- data.frame(rlog.norm.counts.test)

rlog.norm.counts.test$gene_id <- rownames(rlog.norm.counts.test)
## Annotate with Ensemble gene symbol
list <- NULL
for (i in 1:nrow(rlog.norm.counts.test) ) {
  list <- c(list, unlist(strsplit(rlog.norm.counts.test[i,"gene_id"], split = "[.]"))[1])
}
rlog.norm.counts.test$gene_id_new <- list
## Connect http
httr::set_config(httr::config(ssl_cipher_list = "DEFAULT@SECLEVEL=1"))
httr::set_config(httr::config(ssl_verifypeer = FALSE))
## Annotate gene name ##
ensembl = useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl",
mirror="uswest")
genemap <- getBM( attributes = c("ensembl_gene_id", "hgnc_symbol"),
                  filters = "ensembl_gene_id",
                  values = rlog.norm.counts.test$gene_id_new,
                  mart = ensembl)
rlog.norm.counts.test <- rlog.norm.counts.test %>% left_join(genemap, by =
c("gene_id_new" = "ensembl_gene_id"))
gene_id.test <- rlog.norm.counts.test$gene_id
rlog.norm.counts.test  <- assay(DESeq.rlog.test)
rlog.norm.counts.test <- data.frame(rlog.norm.counts.test)
## ggplot2 box plot ##
df_norm.counts.test <- list(counts = as.numeric(unlist(rlog.norm.counts.test)), group =
sample.info$name)
df_norm.counts.test <- data.frame(df_norm.counts.test)
# Plot
p <-  ggplot(df_norm.counts.test, aes(x = group, y = counts)) +
  geom_boxplot(varwidth = FALSE, outlier.colour = "black", outlier.size = 1.5, outlier.stroke
= 1,
               outlier.shape = 1, notch = FALSE,
               color="black", outlier.alpha = 1) +
  theme(axis.title.x=element_blank(), axis.title.y=element_blank() ) +
  theme_bw() + ggtitle("Normalized read counts") + ylab("log 2 read counts") +
xlab("Sample") +
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust=0.5))+
  theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) )
p

rlog.norm.counts.test[which(rlog.norm.counts.test$hgnc_symbol == "GATA3"),]

#Silhouette analysis for identifying optimal cluster number
```

```
#repeat k-means for 1:20 and extract silhouette:
sil <- rep(0, 20)
for(i in 2:20){
  cluster <- kmeans(rlog.norm.counts.test, centers = i, nstart = 20, iter.max = 200)
  ss <- silhouette(cluster$cluster, dist(rlog.norm.counts.test))
  sil[i] <- mean(ss[, 3])

}
# Plot the  average silhouette width
plot(1:20, sil, type = "b", pch = 19, xlab = "Number of clusters k", ylab="Average silhouette
width")
abline(v = which.max(sil), lty = 2)

#ev <- c()
#for (i in 1:20) {
#  km <- kmeans(rlog.norm.counts.test, centers = i, nstart = 20, iter.max = 200)
#  ev[i] <- sum(km$betweenss)/km$totss
#}
#plot(1:20, ev, col="red", lwd=2, type = "l", xlab = "Number of Clusters", ylab = "Explained
Variance")

## Elbow method
set.seed(123)
# function to compute total within-cluster sum of square
wss <- function(k) {
  kmeans(rlog.norm.counts.test, k, nstart = 20 )$tot.withinss
}
# Compute and plot wss for k = 1 to k = 15
k.values <- 1:20
# extract wss for 2-20 clusters
wss_values <- map_dbl(k.values, wss)
plot(k.values, wss_values,
     type="b", pch = 19,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")

## Clustering
#kmean1 <- KMeans_rcpp(data = t(rlog.norm.counts.test), clusters = 3, num_init = 200,
max_iters = 100, seed = 590,
#                    tol_optimal_init=1,initializer = "kmeans++")
kmean <- KMeans_arma(data = t(rlog.norm.counts.test), clusters = 3, n_iter = 200,
seed_mode = "random_subset",
                     verbose = T, CENTROIDS = NULL, seed = 900) #900-950
pr <- predict_KMeans(data = t(rlog.norm.counts.test), kmean)
pr

#list <- kmean$clusters
#list
#sample.info$cluster <- paste("cluster",kmean$clusters,sep = "")
sample.info$cluster <- paste("cluster",pr,sep = "")
```

```
## Create PCA by ggplot2 ##
df_pca <- prcomp(t(rlog.norm.counts.test)) #Transform data row->col
df_out <- as.data.frame(df_pca$x) #Transform to data frame
df_out$Cluster <- as.character(sample.info$cluster) #Add group
pca <- ggplot(df_out, aes(x=PC1,y=PC2, color= Cluster,fill = Cluster,
label=row.names(df_out)  )) +
  geom_point(size = 2) +
  theme_bw() +
  ggtitle("Principle Component Analysis of MIBC mRNA expression classified by K-mean
Clustering") +
  geom_label_repel(
    aes(label = rownames(df_out),fill = Cluster),
    color = 'white',
    size = 3.5,
    max.overlaps = Inf,
    segment.color = "grey50"
  ) +
  guides(fill = guide_legend(override.aes = aes(label = "")),color= "none") +
  theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) )+
  scale_fill_discrete(name = "K-mean Clustering", labels = c("Cluster A", "Cluster
B","Cluster C"))
pca
###################################################################################
################################################
## DIFFERENTIAL EXPRESSION BETWEEN CLUSTERS ##
cluster_list <- list(c("cluster1","cluster2"),c("cluster2","cluster3"),c("cluster3","cluster1"))

for (i in cluster_list) {
  cluster_compare <- i[1]
  cluster_base <- i[2]
  #compare cluster
  sample.compare<- subset(sample.info, cluster==cluster_compare|cluster==cluster_base)
#####Edit cluster here
  read.counts.compare <- read.counts.test[sample.compare$name]
  DESeq.ds.compare <- DESeqDataSetFromMatrix(read.counts.compare ,colData =
sample.compare,~ cluster)
  DESeq.ds.compare <- DESeq.ds.compare[ rowSums(counts(DESeq.ds.compare)) > 10, ]
  DESeq.ds.compare <- estimateSizeFactors(DESeq.ds.compare)
  str(colData(DESeq.ds.compare)$cluster)
  colData(DESeq.ds.compare)$cluster  <- relevel(colData(DESeq.ds.compare)$cluster ,
cluster_base) #####Edit cluster here
  DESeq.ds.compare <- DESeq(DESeq.ds.compare)
  ## Extract result ##
  DGE.results.compare <- results(DESeq.ds.compare , independentFiltering = TRUE , alpha
= 0.01)
  DGE.results.compare <- data.frame(DGE.results.compare)
  DGE.results.compare <- na.omit(DGE.results.compare)
  DGE.results.compare$gene_id <- rownames(DGE.results.compare)
  ## Annotate with Ensemble gene symbol
  list <- NULL
  for (i in 1:nrow(DGE.results.compare) ) {
    list <- c(list, unlist(strsplit(DGE.results.compare[i,"gene_id"], split = "[.]"))[1])
  }
  DGE.results.compare$gene_id_new <- list
  ## Connect http
  httr::set_config(httr::config(ssl_cipher_list = "DEFAULT@SECLEVEL=1"))
  httr::set_config(httr::config(ssl_verifypeer = FALSE))
  ## Annotate gene name ##
  ensembl = useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl",
mirror="uswest")
  genemap <- getBM( attributes = c("ensembl_gene_id", "hgnc_symbol"),
                    filters = "ensembl_gene_id",
                    values = DGE.results.compare$gene_id_new,
                    mart = ensembl)
  DGE.results.compare <- DGE.results.compare %>% left_join(genemap, by =
c("gene_id_new" = "ensembl_gene_id"))
  DGE.results.compare <- DGE.results.compare[!(is.na(DGE.results.compare$hgnc_symbol)
| DGE.results.compare$hgnc_symbol==""), ]

  significant <- DGE.results.compare %>% mutate(
    Expression = case_when(
      DGE.results.compare$log2FoldChange > 2 & DGE.results.compare$padj < 0.05 ~
"upregulate",
      DGE.results.compare$log2FoldChange < -2 & DGE.results.compare$padj < 0.05 ~
"downregulate",
      (DGE.results.compare$log2FoldChange <= 2 & DGE.results.compare$log2FoldChange
>= -2 ) | DGE.results.compare$padj >= 0.05 ~ "non-significant",
      is.na(DGE.results.compare$padj) ~ "non-significant" ) )

  object_name <- paste("plot",cluster_compare,sep = "_")
  assign(object_name,significant)

  object_name <- paste("significant",cluster_compare,sep = "_")
  temp_df <- significant %>% filter(Expression != "non-significant")
  assign(object_name,temp_df) #####Edit cluster here
}
significant_cluster1.sorted <- significant_cluster1[order(significant_cluster1$padj), ]
significant_cluster2.sorted <- significant_cluster2[order(significant_cluster2$padj), ]
significant_cluster3.sorted <- significant_cluster3[order(significant_cluster3$padj), ]
DGEgenes.cluster1 <- subset(significant_cluster1.sorted[1:30,],)
DGEgenes.cluster2 <- subset(significant_cluster2.sorted[1:30,],)
DGEgenes.cluster3 <- subset(significant_cluster3.sorted[1:30,],)
#############################################################
# Plot
MA <- ggplot(data = plot_cluster1, aes(x = baseMean, y = log2FoldChange, col=Expression)
)+
  geom_point(aes(color=as.factor(Expression), fill=Expression), alpha=0.8, size=1.3) +
  scale_x_log10() +
  ggtitle("MA plot of mRNA Expression of derived from MIBC") +
  theme_bw() +
  geom_hline(yintercept=c(-2.0, 2.0), col=" red",linetype = "dashed", size = 0.4) +
  geom_label_repel(
    data = plot_cluster1 %>% filter( hgnc_symbol %in% cluster_all_gene$hgnc_symbol ),
    aes(x=baseMean, y=log2FoldChange,label = hgnc_symbol,fill=as.factor(Expression)),
    color = 'white',
    max.overlaps = Inf,
    segment.color = "grey70",
    size = 3,
    box.padding = unit(0.35, "lines"),
    point.padding = unit(0.3, "lines")
  ) +
  ylim(-15,15) +
  guides(fill = guide_legend(override.aes = aes(label = "")),color="none") +
  theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) ) +
  scale_fill_discrete(name = "Significant Expression", labels = c("Downregulate", "Non-
Significant", "Upregulate"))
MA

plot.volcano <- na.omit(plot_cluster3)

options(ggrepel.max.overlaps = Inf) ##Change ggrepel max overlap, it may notice when we
have more data.

volcano <-  ggplot(plot.volcano) +
  geom_point(aes(x=log2FoldChange, y=-log10(padj), fill=Expression, color = Expression))
+
  ggtitle("Volcano plot of mRNA expression derived from MIBC") +
  xlab("log2 fold change") +
  ylab("-log10 adjusted p-value") +
  theme_bw() +
  geom_vline(xintercept=c(-2.0, 2.0), col=" red",linetype = "dashed", size = 0.4) +
  geom_hline(yintercept=-log10(0.05), col=" red",linetype = "dashed", size = 0.4) +
  geom_label_repel(
    data = plot.volcano %>% filter( hgnc_symbol %in% cluster_all_gene$hgnc_symbol ),
    aes(x=log2FoldChange, y=-log10(padj),label = hgnc_symbol,fill=Expression),
    color = 'white',
    max.overlaps = Inf,
    segment.color = "grey70",
    size = 3,
    box.padding = unit(0.35, "lines"),
    point.padding = unit(0.3, "lines")
  ) +
  guides(fill = guide_legend(override.aes = aes(label = "")),color="none") +
  theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) ) +
  scale_fill_discrete(name = "Significant Expression", labels = c("Downregulate", "Non-
Significant", "Upregulate")) +
  xlim(-10,10)
volcano
#####################################################################################
#################################################
#Subtract Gene
#clusterA <- significant_cluster1$hgnc_symbol[!(significant_cluster1$hgnc_symbol
#                                  %in% c(significant_cluster2$hgnc_symbol,
significant_cluster3$hgnc_symbol))]
#clusterB <- significant_cluster2$hgnc_symbol[!(significant_cluster2$hgnc_symbol
#                                  %in% c(significant_cluster1$hgnc_symbol,
significant_cluster3$hgnc_symbol))]
#clusterC <- significant_cluster3$hgnc_symbol[!(significant_cluster3$hgnc_symbol
#                                  %in% c(significant_cluster1$hgnc_symbol,
significant_cluster2$hgnc_symbol))]

clusterA <- significant_cluster1$hgnc_symbol
clusterB <- significant_cluster2$hgnc_symbol
clusterC <- significant_cluster3$hgnc_symbol
## Enrichment Analysis ##
dbs <- c("GO_Molecular_Function_2021", "GO_Cellular_Component_2021",
"GO_Biological_Process_2021",

"KEGG_2021_Human","Reactome_2016","Panther_2016","WikiPathway_2021_Human","B
ioCarta_2016")

dbs <- "KEGG_2021_Human"
for (i in c("clusterA","clusterB","clusterC") ) {
  for (j in dbs) {
    enriched <- enrichr(eval(as.symbol(i)), dbs)
    object_name <- paste(i,"enrich",sep = "_")
    temp_df <- enriched[[1]]
    assign(object_name,temp_df)
  }
}

for ( i in c("clusterA_enrich","clusterB_enrich","clusterC_enrich") ) {
  column <- "Adjusted.P.value"
  temp_df <- eval(as.symbol(i))[which( eval(as.symbol(i))[[column]] < 0.05  ),]
  temp_df <- data.frame(temp_df)
  temp_df <- temp_df %>% filter(str_detect(temp_df$Term, "pathway") )
  object_name <- paste(i,"pathway",sep = "_")
  assign(object_name,temp_df)
  temp_list <- unique(unlist(str_split(temp_df[["Genes"]], ";")))
  object_name <- paste(i,"gene",sep = "_")
  assign(object_name,temp_list)
}

# Store value
#x <- list(
#  ClusterA = significant_cluster1$hgnc_symbol,
#  ClusterB = significant_cluster2$hgnc_symbol,
#  ClusterC = significant_cluster3$hgnc_symbol
#)
x <- list("ClusterA-B" = clusterA_enrich_gene,
          "ClusterB-C" = clusterB_enrich_gene,
          "ClusterA-C" = clusterC_enrich_gene)
## Venn diagram
ggvenn(
  x,
  show_elements = FALSE,
  fill_color = c("#E69F00", "#56B4E9", "#009E73"),
  fill_alpha = 0.3,
  stroke_color = "black",
  stroke_size = 0,
  set_name_size = 5,
  text_size = 4
)
## Plot p value
#_________ installing Packages
#install.packages("ggplot2", dependencies = TRUE)
#install.packages("gridExtra", dependencies = TRUE)

#--------- loading lib
library("ggplot2")
library("gridExtra")

#Saving in png
#png("ggplot2sizing.png",height=400,width=850)

#df=data.frame(dOut_x,dOut_y,d_pvalue)

head(clusterA_enrich_pathway)

colnames(clusterA_enrich_pathway)

clusterA_enrich_pathway$Overlap.gene <- clusterA_enrich_pathway$Overlap %>%
str_match_all("[0-9]+") %>% data.frame() %>% .[1,] %>% unlist  %>% as.numeric
clusterB_enrich_pathway$Overlap.gene <- clusterB_enrich_pathway$Overlap %>%
str_match_all("[0-9]+") %>% data.frame() %>% .[1,] %>% unlist  %>% as.numeric
clusterC_enrich_pathway$Overlap.gene <- clusterC_enrich_pathway$Overlap %>%
str_match_all("[0-9]+") %>% data.frame() %>% .[1,] %>% unlist  %>% as.numeric

#Graph3 with scale_color_gradien :: log10
ggplot(data=clusterA_enrich_pathway, aes(x=Odds.Ratio,y=Term, size=Overlap.gene,
color=Adjusted.P.value))+
  geom_point(alpha=0.4)+scale_colour_gradientn(colours=rainbow(5))+
  scale_size(range = c(2, 15), name="Overlap genes") +
  theme_bw() +
  labs(title = "Cluster A pathway enrichment", x= "Odds ratio", y= "Term", color = "Adjusted
P-value",plot.title = element_text(hjust = 0.5))
```

```
ggplot(data=clusterB_enrich_pathway, aes(x=Odds.Ratio,y=Term, size=Overlap.gene,
color=Adjusted.P.value))+
  geom_point(alpha=0.4)+scale_colour_gradientn(colours=rainbow(5))+
  scale_size(range = c(2, 15), name="Overlap genes") +
  theme_bw() +
  labs(title = "Cluster B pathway enrichment", x= "Odds ratio", y= "Term", color = "Adjusted
P-value",plot.title = element_text(hjust = 0.5))

ggplot(data=clusterC_enrich_pathway, aes(x=Odds.Ratio,y=Term, size=Overlap.gene,
color=Adjusted.P.value))+
  geom_point(alpha=0.4)+scale_colour_gradientn(colours=rainbow(5))+
  scale_size(range = c(2, 15), name="Overlap genes") +
  theme_bw() +
  labs(title = "Cluster C pathway enrichment", x= "Odds ratio", y= "Term", color = "Adjusted
P-value",plot.title = element_text(hjust = 0.5))

###########################################################################
###########################################
## REVALIDATE ##
#read.counts.filter <- read.counts.test
#colnames(read.counts.filter)
#read.counts.filter$gene_id <- rownames(read.counts.filter)
#list <- NULL
#for (i in 1:nrow(read.counts.filter) ) {
#  list <- c(list, unlist(strsplit(read.counts.filter[i,"gene_id"], split = "[.]"))[1])
#}
#read.counts.filter$gene_id_new <- list

#genemap.filter <- getBM( attributes = c("ensembl_gene_id", "hgnc_symbol"),
#                        filters = "ensembl_gene_id",
#                        values = read.counts.filter$gene_id_new,
#                        mart = ensembl)
#read.counts.filter <- read.counts.filter %>% left_join(genemap.filter, by = c("gene_id_new"
= "ensembl_gene_id"))
#read.counts.filter <- read.counts.filter %>% filter( !(read.counts.filter$gene_id_new %in%
gene_filter$ensembl_gene_id) )
#read.counts.filter <- read.counts.filter %>% filter( hgnc_symbol %in%
c(clusterA_enrich_gene,clusterA_enrich_gene,clusterB_enrich_gene ) )
#read.counts.filter <- read.counts.filter[!(is.na(read.counts.filter$hgnc_symbol) |
read.counts.filter$hgnc_symbol==""), ]
#read.counts.filter <- distinct_at(read.counts.filter, vars(gene_id), .keep_all = TRUE)
#rownames(read.counts.filter) <- read.counts.filter$gene_id
#read.counts.filter <- read.counts.filter %>% dplyr::select(-
c(gene_id,gene_id_new,hgnc_symbol))
########################################################
# import
#DESeq.ds.filter <- DESeqDataSetFromMatrix(read.counts.filter ,colData = sample.info,~ 1)
## Include gene which counts more than 0 ##
#DESeq.ds.filter <- DESeq.ds.filter[ rowSums(counts(DESeq.ds.filter)) > 10, ]
```

```
## Calculate size factor of normalization counts ##
#DESeq.ds.filter <- estimateSizeFactors(DESeq.ds.filter)
## Normalization and Log2 transform by DESeq2 ##
#DESeq.rlog.filter  <- varianceStabilizingTransformation(DESeq.ds.filter , blind = TRUE)
## Summarized object data to matrix ##
#rlog.norm.counts.filter  <- assay(DESeq.rlog.filter)
#rlog.norm.counts.filter <- data.frame(rlog.norm.counts.filter)


# Do K-mean Clustering
#kmean <- KMeans_rcpp(t(rlog.norm.counts.filter), clusters = 2, num_init = 5, max_iters =
100, initializer = "random")
#list <- kmean$clusters
#list
sample.info$cluster_filter <- paste("cluster",kmean$clusters,sep = "")

## Create PCA by ggplot2 ##
#df_pca <- prcomp(t(rlog.norm.counts.filter)) #Transform data row->col
#df_out <- as.data.frame(df_pca$x) #Transform to data frame
#df_out$Cluster <- as.character(sample.info$condition) #Add group
#df_out$Cluster <- as.character(sample.info$cluster_filter) #Add group

#pca <- ggplot(df_out, aes(x=PC1,y=PC2, color= Cluster,fill = Cluster,
label=row.names(df_out)  )) +
#  geom_point(size = 2) +
#  theme_bw() +
#  ggtitle("Principle Component Analysis of MIBC mRNA expression classified by K-mean
Clustering") +
#  geom_label_repel(
#    aes(label = rownames(df_out),fill = Cluster),
#    color = 'white',
#    size = 3.5,
#    max.overlaps = Inf,
#    segment.color = "grey50"
#  ) +
#  guides(fill = guide_legend(override.aes = aes(label = "")),color= "none") +
#  theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) )+
#  scale_fill_discrete(name = "K-mean Clustering", labels = c("Cluster A", "Cluster
B","Cluster C"))
#pca
##############################################################################
#############################################
## GET TCGA DATA
setwd("~/bladder/tcga")
# RAW
query <- GDCquery(
  project = "TCGA-BLCA",
  data.category = "Transcriptome Profiling",
  data.type = "Gene Expression Quantification",
  workflow.type = "HTSeq - Counts"
)
# FPKM
#query <- GDCquery(
#  project = "TCGA-BLCA",
#  data.category = "Transcriptome Profiling",
#  data.type = "Gene Expression Quantification",
#  workflow.type = "HTSeq - FPKM"
#)
GDCdownload(query, method = "api", files.per.chunk = 5)
data <- GDCprepare(query)
metadata <- getManifest(query)
sample <- as.data.frame(colData(data))
colnames(metadata)
head(sample)
sample[1,"treatments"]
# column paper_mRNA.cluster paper_Histologic.subtype paper_Histologic.grade
# paper_AJCC.Tumor.category paper_Lymphovascular.invasion paper_AJCC.LN.category
paper_Number.of.LNs.examined paper_AJCC.metastasis.category
# paper_Tumor.category.12.vs..34 paper_LN.negative.vs..positive
paper_Combined.T.and.LN.category

# Find barcode and file name
barcode <-UUIDtoBarcode(metadata$id, from_type = "file_id") ##
metadata$barcode <- barcode$associated_entities.entity_submitter_id ##
# Subset TCGA data
sample.df <- data.frame(sample$barcode, sample$ajcc_pathologic_stage,
sample$ajcc_pathologic_t, sample$ajcc_pathologic_n ,
                        sample$ajcc_pathologic_m , sample$days_to_last_follow_up,
                        sample$gender, sample$age_at_diagnosis,sample$vital_status )

colnames(sample.df) <-
c("barcode","ajcc_pathologic_stage","ajcc_pathologic_t","ajcc_pathologic_n","ajcc_patholog
ic_m",
                        "days_to_last_follow_up","gender","age_at_diagnosis","vital_status")
metadata.new <- metadata %>% dplyr::select(barcode,filename)
sample.df <- sample.df %>% left_join(metadata.new, by = c("barcode" = "barcode"))
#select only MIBC

mibc <- sample.df[which( sample$primary_diagnosis == "Transitional cell carcinoma" &
                          ( sample$ajcc_pathologic_t == "T2" | sample$ajcc_pathologic_t == "T2a" |
sample$ajcc_pathologic_t == "T2b" |
                            sample$ajcc_pathologic_t == "T3" | sample$ajcc_pathologic_t == "T3a" |
sample$ajcc_pathologic_t == "T3b" |
                            sample$ajcc_pathologic_t == "T4" | sample$ajcc_pathologic_t == "T4a" |
sample$ajcc_pathologic_t == "T4b" ) ),]
mibc$cluster
mibc <- na.omit(mibc)

## make temp df
```

```
temp.sample <- data.frame(sample$barcode, sample$paper_mRNA.cluster,
sample$paper_Histologic.subtype,
                          sample$paper_Histologic.grade, sample$paper_Lymphovascular.invasion)
colnames(temp.sample) <- c("barcode","tcga_cluster","subtype","grade","invasion")
temp.sample$tcga_cluster

## merge some data
mibc <- mibc %>% left_join(temp.sample, by = c("barcode" = "barcode"))

mibc <- mibc %>% mutate(tcga_cluster_modify = case_when(mibc$tcga_cluster ==
"Basal_squamous" ~"Basal",
                                                        mibc$tcga_cluster == "Luminal_infiltrated" |
mibc$tcga_cluster == "Luminal_papillary" |  mibc$tcga_cluster == "Luminal" ~"Luminal",
                                                        mibc$tcga_cluster == "Neuronal" ~"Neuronal"))

mibc <- mibc %>% mutate(cluster_filter_modify = case_when(mibc$cluster_filter ==
"cluster1" ~"Cluster A",
                                                          mibc$cluster_filter == "cluster2" ~"Cluster B",
                                                          mibc$cluster_filter == "cluster3" ~"Cluster C"))

mibc <- mibc %>% mutate(ajcc_pathologic_t_modify = case_when(mibc$ajcc_pathologic_t
== "T2"|mibc$ajcc_pathologic_t == "T2a"|mibc$ajcc_pathologic_t == "T2b" ~"T2",
                                                             mibc$ajcc_pathologic_t ==
"T3"|mibc$ajcc_pathologic_t == "T3a"|mibc$ajcc_pathologic_t == "T3b" ~"T3",
                                                             mibc$ajcc_pathologic_t ==
"T4"|mibc$ajcc_pathologic_t == "T4a"|mibc$ajcc_pathologic_t == "T4b" ~"T4"))

#mibc <- mibc %>% dplyr::select(-mibc)

mibc$filename <- gsub('*.gz','\\1',mibc$filename)
mibc <- mibc %>% mutate(status = case_when(mibc$vital_status == "Alive" ~0,
mibc$vital_status == "Dead" ~1,mibc$vital_status == "Not Reported" ~1))
################################################################################
############################################
# Clinical graph
colnames(mibc)
#> colnames(mibc)
#[1] "barcode"              "ajcc_pathologic_stage"  "ajcc_pathologic_t"
"ajcc_pathologic_n"      "ajcc_pathologic_m"      "days_to_last_follow_up"
#[7] "gender"               "age_at_diagnosis"       "vital_status"          "filename"
"cluster_filter"         "tcga_cluster"
#[13] "subtype"              "grade"                  "invasion"              "tcga_cluster_modify"
"cluster_filter_modify"

#summ(mibc$days_to_last_follow_up/365, by=mibc$cluster_filter_modify)

#tabpct(mibc$invasion ,mibc$cluster_filter_modify)


## Cluster VS Cluster
```

```
cluster_tcga <- na.omit(mibc) %>%
  group_by(cluster_filter_modify, tcga_cluster_modify) %>%
  dplyr::summarize(n = n()) %>%
  mutate(pct = n/sum(n),
         lbl = scales::percent(pct))
cluster_tcga

ggplot(cluster_tcga,
       aes(x = cluster_filter_modify,
           fill = tcga_cluster_modify,
           y = pct
           ))  +
  geom_bar(stat = "identity",
           position = "fill") +
  scale_fill_brewer(palette = "Set2") +
  scale_y_continuous(breaks = seq(0, 1, .2),labels = scales::percent) +
  geom_text(aes(label = lbl),
            size = 3,
            position = position_stack(vjust = 0.5)) +
  labs(y = "Percent",
       fill = "TCGA classfication",
       x = "K-mean clustering",
       title = "Cumulative MIBC cases by K-mean clustering",
       subtitle = "Classified by mRNA TCGA classification") +
  theme_minimal()

## ajcc_pathologic_T_stage
ajcc_pathologic_stage <- na.omit(mibc) %>%
  group_by(cluster_filter_modify, ajcc_pathologic_stage) %>%
  dplyr::summarize(n = n()) %>%
  mutate(pct = n/sum(n),
         lbl = scales::percent(pct))
ajcc_pathologic_stage

ggplot(ajcc_pathologic_stage,
       aes(x = cluster_filter_modify,
           fill = ajcc_pathologic_stage,
           y = pct
       ))  +
  geom_bar(stat = "identity",
           position = "fill") +
  scale_fill_brewer(palette = "Accent") +
  scale_y_continuous(breaks = seq(0, 1, .2),labels = scales::percent) +
  geom_text(aes(label = lbl),
            size = 3,
            position = position_stack(vjust = 0.5)) +
  labs(y = "Percent",
       fill = "AJCC pathologic T stage",
       x = "K-mean clustering",
       title = "Cumulative MIBC cases by K-mean clustering",
       subtitle = "Classified by AJCC pathologic T stage") +
  theme_minimal()

## ajcc_pathologic_n_stage
ajcc_pathologic_n <- na.omit(mibc) %>%
  group_by(cluster_filter_modify, ajcc_pathologic_n) %>%
  dplyr::summarize(n = n()) %>%
  mutate(pct = n/sum(n),
         lbl = scales::percent(pct))
ajcc_pathologic_n

ggplot(ajcc_pathologic_n,
       aes(x = cluster_filter_modify,
           fill = ajcc_pathologic_n,
           y = pct
       ))  +
  geom_bar(stat = "identity",
           position = "fill") +
  scale_fill_brewer(palette = "Accent") +
  scale_y_continuous(breaks = seq(0, 1, .2),labels = scales::percent) +
  geom_text(aes(label = lbl),
            size = 3,
            position = position_stack(vjust = 0.5)) +
  labs(y = "Percent",
       fill = "AJCC pathologic N stage",
       x = "K-mean clustering",
       title = "Cumulative MIBC cases by K-mean clustering",
       subtitle = "Classified by AJCC pathologic N stage") +
  theme_minimal()

## ajcc_pathologic_m
ajcc_pathologic_m <- na.omit(mibc) %>%
  group_by(cluster_filter_modify, ajcc_pathologic_m) %>%
  dplyr::summarize(n = n()) %>%
  mutate(pct = n/sum(n),
         lbl = scales::percent(pct))
ajcc_pathologic_m

ggplot(ajcc_pathologic_m,
       aes(x = cluster_filter_modify,
           fill = ajcc_pathologic_m,
           y = pct
       ))  +
  geom_bar(stat = "identity",
           position = "fill") +
  scale_fill_brewer(palette = "Accent") +
  scale_y_continuous(breaks = seq(0, 1, .2),labels = scales::percent) +
  geom_text(aes(label = lbl),
            size = 3,
            position = position_stack(vjust = 0.5)) +
  labs(y = "Percent",
       fill = "AJCC pathologic M stage",
       x = "K-mean clustering",
       title = "Cumulative MIBC cases by K-mean clustering",
       subtitle = "Classified by AJCC pathologic M stage") +
  theme_minimal()


## invasion
invasion <- na.omit(mibc) %>%
  group_by(cluster_filter_modify, invasion) %>%
  dplyr::summarize(n = n()) %>%
  mutate(pct = n/sum(n),
         lbl = scales::percent(pct))
invasion

ggplot(invasion,
       aes(x = cluster_filter_modify,
           fill = invasion,
           y = pct
       ))  +
  geom_bar(stat = "identity",
           position = "fill") +
  scale_fill_brewer(palette = "Set1") +
  scale_y_continuous(breaks = seq(0, 1, .2),labels = scales::percent) +
  geom_text(aes(label = lbl),
            size = 3,
            position = position_stack(vjust = 0.5)) +
  labs(y = "Percent",
       fill = "Lymphovascular invasion",
       x = "K-mean clustering",
       title = "Cumulative MIBC cases by K-mean clustering",
       subtitle = "Classified by lymphovascular invasion")+
  theme_minimal()


ggplot(na.omit(mibc),
       aes(x = cluster_filter_modify,
           fill = gender)) +
  geom_bar(position = "fill") +
  labs(y = "Proportion")

################################################################################
#############################################
# Get expression from TCGA file
list_gene <- row.names(read.counts)
#file <- fread("0a1b146d-7b9f-4382-9cfc-c187fafbb47c.htseq.counts")
#file <- subset(file, file$V1 %in% list_gene)
##Loop create data frame
setwd("~/bladder/tcga/GDCdata/TCGA-BLCA")
```

```
i <- 1
for (j in mibc$filename ) {
  #j <- "74fa890a-d4e6-4718-959a-2b2c125c892e.FPKM.txt"
  if (i == 1) {
    file <- fread(j)
    file <- subset(file, file$V1 %in% list_gene)
    #gsub('*.genes.results','\\1',j)
    ## Add first sample
    read.counts.tcga <- data.frame(as.numeric(format(file$V2,scientific = FALSE),row.names
= file$V1) )
    colnames(read.counts.tcga) <- mibc$barcode[which(mibc$filename==j)]
    ## Set condition
    #sample.info <- data.frame(gsub('*.genes.results','\\1',j),"normal")
    sample.info.tcga <- data.frame(mibc$barcode[which(mibc$filename==j)],"cancer")
    colnames(sample.info.tcga) <- c("name","condition")
    row.names(read.counts.tcga) <- file$V1
  } else {
    ## Add another sample ##
    file <- fread(j)
    file <- subset(file, file$V1 %in% list_gene)
    ## Set sample name ##
    sample <- mibc$barcode[which(mibc$filename==j)]
    ## Set condition
    if (i >= 2) {
      condition <- "cancer"
    } else {
      condition <- "cancer"
    }
    ## Insert sample to sample.info and read.counts ##
    count <- data.frame(as.numeric(format(file$V2,scientific = FALSE) ) )
    colnames(count) <- sample
    info <- data.frame(sample,condition)
    colnames(info) <- c("name","condition")
    read.counts.tcga <- cbind(read.counts.tcga,count)
    sample.info.tcga <- rbind(sample.info.tcga,info)
  }
  i <- i + 1
}
rm(count)
rm(sample)
rm(info)
rm(condition)
rm(i)
rm(j)
rm(file)
################################################################################
#############################################
# Discard non mRNA gene
read.counts.tcga.test <- read.counts.tcga
colnames(read.counts.tcga.test)
```

```
read.counts.tcga.test$gene_id <- rownames(read.counts.tcga.test)
list <- NULL
for (i in 1:nrow(read.counts.tcga.test) ) {
  list <- c(list, unlist(strsplit(read.counts.tcga.test[i,"gene_id"], split = "[.]"))[1])
}
read.counts.tcga.test$gene_id_new <- list

genemap.test <- getBM( attributes = c("ensembl_gene_id", "hgnc_symbol"),
                    filters = "ensembl_gene_id",
                    values = read.counts.tcga.test$gene_id_new,
                    mart = ensembl)
read.counts.tcga.test <- read.counts.tcga.test %>% left_join(genemap.test, by =
c("gene_id_new" = "ensembl_gene_id"))
read.counts.tcga.test <- read.counts.tcga.test %>% filter( !(read.counts.tcga.test$gene_id_new
%in% gene_filter$ensembl_gene_id) )
read.counts.tcga.test <- read.counts.tcga.test[!(is.na(read.counts.tcga.test$hgnc_symbol) |
read.counts.tcga.test$hgnc_symbol==""), ]
read.counts.tcga.test <- distinct_at(read.counts.tcga.test, vars(gene_id), .keep_all = TRUE)
rownames(read.counts.tcga.test) <- read.counts.tcga.test$gene_id
read.counts.tcga.test <- read.counts.tcga.test %>% dplyr::select(-
c(gene_id,gene_id_new,hgnc_symbol))
###############################################################################
#############################################
## VALIDATE IN TCGA ##
read.counts.tcga.test <- read.counts.tcga
colnames(read.counts.tcga.test)
read.counts.tcga.test$gene_id <- rownames(read.counts.tcga.test)
list <- NULL
for (i in 1:nrow(read.counts.tcga.test) ) {
  list <- c(list, unlist(strsplit(read.counts.tcga.test[i,"gene_id"], split = "[.]"))[1])
}
read.counts.tcga.test$gene_id_new <- list

genemap.test <- getBM( attributes = c("ensembl_gene_id", "hgnc_symbol"),
                    filters = "ensembl_gene_id",
                    values = read.counts.tcga.test$gene_id_new,
                    mart = ensembl)
read.counts.tcga.test <- read.counts.tcga.test %>% left_join(genemap.test, by =
c("gene_id_new" = "ensembl_gene_id"))
read.counts.tcga.test <- read.counts.tcga.test %>% filter( !(read.counts.tcga.test$gene_id_new
%in% gene_filter$ensembl_gene_id) )
read.counts.tcga.test <- read.counts.tcga.test %>% filter( read.counts.tcga.test$gene_id %in%
gene_id.test )
read.counts.tcga.test <- read.counts.tcga.test[!(is.na(read.counts.tcga.test$hgnc_symbol) |
read.counts.tcga.test$hgnc_symbol==""), ]
read.counts.tcga.test <- distinct_at(read.counts.tcga.test, vars(gene_id), .keep_all = TRUE)
rownames(read.counts.tcga.test) <- read.counts.tcga.test$gene_id
read.count.tcga.gene <- read.counts.tcga.test %>%
dplyr::select(gene_id,gene_id_new,hgnc_symbol)
```

```
read.counts.tcga.test <- read.counts.tcga.test %>% dplyr::select(-
c(gene_id,gene_id_new,hgnc_symbol))
# import
DESeq.ds.test.tcga <- DESeqDataSetFromMatrix(read.counts.tcga.test ,colData =
sample.info.tcga,~ 1)
## Include gene which counts more than 0 ##
#DESeq.ds.test.tcga <- DESeq.ds.test.tcga[ rowSums(counts(DESeq.ds.test.tcga)) > 0, ]
## Calculate size factor of normalization counts ##
DESeq.ds.test.tcga <- estimateSizeFactors(DESeq.ds.test.tcga)
## Normalization and Log2 transform by DESeq2 ##
DESeq.rlog.test.tcga  <- varianceStabilizingTransformation(DESeq.ds.test.tcga , blind =
TRUE)
## Summarized object data to matrix ##
rlog.norm.counts.test.tcga  <- assay(DESeq.rlog.test.tcga)
rlog.norm.counts.test.tcga <- data.frame(rlog.norm.counts.test.tcga)

head(rlog.norm.counts.test.tcga)

## violin plot expression by gene
rlog.norm.counts.test.tcga$gene_id <- rownames(rlog.norm.counts.test.tcga)
list <- NULL
for (i in 1:nrow(rlog.norm.counts.test.tcga) ) {
  list <- c(list, unlist(strsplit(rlog.norm.counts.test.tcga[i,"gene_id"], split = "[.]"))[1])
}
rlog.norm.counts.test.tcga$gene_id_new <- list

rlog.norm.counts.test.tcga <- rlog.norm.counts.test.tcga %>% left_join(genemap.test, by =
c("gene_id_new" = "ensembl_gene_id"))

rlog.norm.counts.test.tcga[which(rlog.norm.counts.test.tcga$hgnc_symbol == "KRT56"),]

tcga_ihc <-
data.frame(t(rlog.norm.counts.test.tcga[which(rlog.norm.counts.test.tcga$hgnc_symbol ==
"GATA3"|
                                              rlog.norm.counts.test.tcga$hgnc_symbol == "KRT14"|
                                              rlog.norm.counts.test.tcga$hgnc_symbol == "KRT5"|
                                              rlog.norm.counts.test.tcga$hgnc_symbol == "KRT20"|
                                              rlog.norm.counts.test.tcga$hgnc_symbol == "SNCA"|
                                              rlog.norm.counts.test.tcga$hgnc_symbol == "CD274"
),1:231]))
colnames(tcga_ihc) <- c("GATA3","KRT14","KRT5","KRT20","SNCA","CD274")

tcga_ihc$cluster <- as.factor(mibc$cluster_filter_modify)

for( i in 1:6) {
  #i <- 2
  plot_gene <- colnames(tcga_ihc[i])
  #plot_gene <- "GATA3"
  plot_name <- paste0("Expression of ",plot_gene," by K-mean cluster")
  plot_temp_df <- data.frame(tcga_ihc[[i]],tcga_ihc$cluster)
  colnames(plot_temp_df) <- c("rlog","cluster")
  plot_temp <- ggplot(plot_temp_df, aes(x=cluster, y=rlog, fill=cluster)) +
    geom_violin(trim=FALSE)+
    geom_boxplot(width=0.1, fill="white")+
    labs(title=plot_name,x="K-mean clustering", y = "Log 2 normalized count") +
    scale_fill_brewer(palette="Dark2") +
    theme_minimal() +
    theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) )
  assign(plot_gene,plot_temp)
}

#pr <- predict_KMeans(t(rlog.norm.counts.test.tcga), kmean$centroids)

pr = predict_KMeans(data = t(rlog.norm.counts.test.tcga), kmean)

#kmean <- KMeans_rcpp(data = t(rlog.norm.counts.test.tcga), clusters = 3, num_init = 100,
max_iters = 100, seed = 100,
#                    tol_optimal_init=0.8,initializer = "kmeans++")

## Create PCA by ggplot2 ##
df_pca <- prcomp(t(rlog.norm.counts.test.tcga)) #Transform data row->col
df_out <- as.data.frame(df_pca$x) #Transform to data frame
df_out$Cluster <- paste("cluster",pr,sep = "") #Add group
#df_out$Cluster <- paste("cluster",kmean$clusters,sep = "")

pca <- ggplot(df_out, aes(x=PC1,y=PC2, color= Cluster,fill = Cluster,
label=row.names(df_out)  )) +
  geom_point(size = 2) +
  theme_bw() +
  ggtitle("Principle Component Analysis of TCGA MIBC mRNA expression classified by K-
mean Clustering") +
  theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) )
pca
################################################################################
#############################################
read.counts.filter.tcga <- read.counts.tcga
colnames(read.counts.filter.tcga)
read.counts.filter.tcga$gene_id <- rownames(read.counts.filter.tcga)
list <- NULL
for (i in 1:nrow(read.counts.filter.tcga) ) {
  list <- c(list, unlist(strsplit(read.counts.filter.tcga[i,"gene_id"], split = "[.]"))[1])
}
read.counts.filter.tcga$gene_id_new <- list

genemap.filter.tcga <- getBM( attributes = c("ensembl_gene_id", "hgnc_symbol"),
                        filters = "ensembl_gene_id",
                        values = read.counts.filter.tcga$gene_id_new,
                        mart = ensembl)
```

```
read.counts.filter.tcga <- read.counts.filter.tcga %>% left_join(genemap.filter.tcga, by =
c("gene_id_new" = "ensembl_gene_id"))
read.counts.filter.tcga <- read.counts.filter.tcga %>% filter(
!(read.counts.filter.tcga$gene_id_new %in% gene_filter$ensembl_gene_id) )
read.counts.filter.tcga <- read.counts.filter.tcga %>% filter( hgnc_symbol %in%
c(clusterA_enrich_gene,clusterA_enrich_gene,clusterB_enrich_gene ) )
read.counts.filter.tcga <- read.counts.filter.tcga[!(is.na(read.counts.filter.tcga$hgnc_symbol) |
read.counts.filter.tcga$hgnc_symbol==""), ]
read.counts.filter.tcga <- distinct_at(read.counts.filter.tcga, vars(gene_id), .keep_all = TRUE)
rownames(read.counts.filter.tcga) <- read.counts.filter.tcga$gene_id
read.counts.filter.tcga <- read.counts.filter.tcga %>% dplyr::select(-
c(gene_id,gene_id_new,hgnc_symbol))
# import
DESeq.ds.filter.tcga <- DESeqDataSetFromMatrix(read.counts.filter.tcga ,colData =
sample.info.tcga,~ 1)
## Include gene which counts more than 0 ##
DESeq.ds.filter.tcga <- DESeq.ds.filter.tcga[ rowSums(counts(DESeq.ds.filter.tcga)) > 100, ]
## Calculate size factor of normalization counts ##
DESeq.ds.filter.tcga <- estimateSizeFactors(DESeq.ds.filter.tcga)
## Normalization and Log2 transform by DESeq2 ##
DESeq.rlog.filter.tcga  <- varianceStabilizingTransformation(DESeq.ds.filter.tcga , blind =
TRUE)
## Summarized object data to matrix ##
rlog.norm.counts.filter.tcga  <- assay(DESeq.rlog.filter.tcga)
rlog.norm.counts.filter.tcga <- data.frame(rlog.norm.counts.filter.tcga)

pr <- predict_KMeans(t(rlog.norm.counts.filter.tcga), kmean$centroids)
sample.info.tcga$cluster <- paste("cluster",pr,sep = "")




## Create PCA by ggplot2 ##
df_pca <- prcomp(t(rlog.norm.counts.filter.tcga)) #Transform data row->col
df_out <- as.data.frame(df_pca$x) #Transform to data frame
df_out$Cluster <- as.character(pr) #Add group

pca <- ggplot(df_out, aes(x=PC1,y=PC2, color= Cluster,fill = Cluster,
label=row.names(df_out)  )) +
  geom_point(size = 2) +
  theme_bw() +
  ggtitle("Principle Component Analysis of MIBC mRNA expression classified by K-mean
Clustering") +
  theme(legend.position="bottom", legend.box = "horizontal", plot.title = element_text(hjust
= 0.5) )
pca
################################################################################
############################################
## Survival analysis
colnames(mibc)
```

```
mibc$cluster_filter <- sample.info.tcga$cluster
#mibc$cluster_filter <- paste("cluster",kmean$clusters,sep = "")

mibc <- mibc %>% mutate(duration5 = ifelse(days_to_last_follow_up <=1825,
days_to_last_follow_up, 1825))
mibc <- mibc %>% mutate(status5 = ifelse(days_to_last_follow_up <=1825, status, 0))

#surv_object <- Surv(time = mibc$days_to_last_follow_up, event = mibc$status)
surv_object <- Surv(time = mibc$days_to_last_follow_up, event = mibc$status)
fit2 <- survfit(surv_object ~cluster_filter , data = mibc)
ggsurvplot(fit2, data = mibc, pval = TRUE,legend.title="")

survplot <- ggsurvplot(
  fit2,
  data = mibc,
  risk.table = TRUE,
  pval = TRUE,
  pval.method=TRUE,
  pval.coord = c(750, 0.075),
  pval.method.coord = c(750, 0.15),
  conf.int = T,
  risk.table.y.text.col = TRUE,
  legend.labs=c("Cluster A", "Cluster B", "Cluster C"),
  size=0.7,
  xlim = c(0,1825), #
  #alpha=c(0.4),
  conf.int.alpha=c(0.1),
  break.x.by = 300,
  xlab="Days of follow-up",
  ylab="Probability of native liver survival",
  surv.median.line = "hv",
  ylim=c(0,1),
  surv.scale="percent",
  tables.col="strata",
  risk.table.col = "strata",
  risk.table.y.text = FALSE,
  tables.y.text = FALSE,
  legend.title="K-mean Cluster",
  palette = "Dark2")
survplot$plot <- survplot$plot + labs(
  title    = "Kaplan-Meier of MIBC derived from TCGA",
  subtitle = "Classified by K-mean clustering"
) + theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 0.5))
survplot
###########################################################################
###########################################
## ROC gene expression with cluster
DGEgenes.cluster1
DGEgenes.cluster2
DGEgenes.cluster3
```

```
clusterA_enrich_gene

clusterA_unique_gene <- clusterA_enrich_gene[!(clusterA_enrich_gene %in%
c(clusterB_enrich_gene,clusterC_enrich_gene))]
clusterB_unique_gene <- clusterB_enrich_gene[!(clusterB_enrich_gene %in%
c(clusterA_enrich_gene,clusterC_enrich_gene))]
clusterC_unique_gene <- clusterC_enrich_gene[!(clusterC_enrich_gene %in%
c(clusterA_enrich_gene,clusterB_enrich_gene))]

cluster_all_gene <- clusterA_enrich_gene[which( (clusterA_enrich_gene %in%
clusterB_enrich_gene) &
                                         (clusterA_enrich_gene %in% clusterC_enrich_gene) )]

cluster_all_gene <- data.frame(cluster_all_gene)
colnames(cluster_all_gene) <- "hgnc_symbol"
cluster_all_gene <- cluster_all_gene %>% left_join(read.count.tcga.gene, by =
c("hgnc_symbol" = "hgnc_symbol") )
cluster_all_gene$hgnc_symbol

clusterA_unique_gene <- data.frame(clusterA_unique_gene)
colnames(clusterA_unique_gene) <- "hgnc_symbol"
clusterA_unique_gene <- clusterA_unique_gene %>% left_join(read.count.tcga.gene, by =
c("hgnc_symbol" = "hgnc_symbol") )
clusterA_unique_gene$hgnc_symbol

clusterB_unique_gene <- data.frame(clusterB_unique_gene)
colnames(clusterB_unique_gene) <- "hgnc_symbol"
clusterB_unique_gene <- clusterB_unique_gene %>% left_join(read.count.tcga.gene, by =
c("hgnc_symbol" = "hgnc_symbol") )
clusterB_unique_gene$hgnc_symbol

clusterC_unique_gene <- data.frame(clusterC_unique_gene)
colnames(clusterC_unique_gene) <- "hgnc_symbol"
clusterC_unique_gene <- clusterC_unique_gene %>% left_join(read.count.tcga.gene, by =
c("hgnc_symbol" = "hgnc_symbol") )
clusterC_unique_gene$hgnc_symbol

cluster_list <- list(c("cluster1","cluster2"),c("cluster2","cluster3"),c("cluster3","cluster1"))

plot_list <- list()
auc_score <- list()
temp_df <- cluster_all_gene ## change here
for (i in temp_df$gene_id) {
  rocobj <- list()
  auc <- list()
  for (j in cluster_list) {
    cluster_compare <- j[1]
    cluster_base <- j[2]
    gene.name <- temp_df$hgnc_symbol[which(temp_df$gene_id == i)]
    gene.id <- temp_df$gene_id_new[which(temp_df$gene_id == i)]
    roc.gene <- i
    roc.expression <- data.frame(t(rlog.norm.counts.test.tcga[roc.gene,]))
    roc.expression$cluster <- sample.info.tcga$cluster
    colnames(roc.expression) <- c("gene","cluster")
    roc.expression <- roc.expression %>% filter(cluster == cluster_base | cluster ==
cluster_compare)

    rocobj[[cluster_compare]] <- roc(roc.expression$cluster,roc.expression$gene)
    auc[[cluster_compare]] <- round(auc(roc.expression$cluster,roc.expression$gene),3)
  }
  #create ROC plot
  plot_list[[i]] <- ggroc(rocobj, size = 1) +
    ggtitle(paste0(gene.name,' (',gene.id, ')')) +
    theme_bw() +
    labs(color='Cluster') +
    theme(plot.title = element_text(hjust = 0.5) ) +
    annotate("text", x=0.5, y=0.05, label=paste( "AUC:",auc[1],",",auc[2],",",auc[3] ) )
  auc_score[[i]] <- auc
}
auc_score <- data.frame(matrix(unlist(auc_score), nrow=length(auc_score), byrow=TRUE),
                        row.names = temp_df$hgnc_symbol)
colnames(auc_score) <- c("cluster1","cluster2","cluster3")

grid.arrange(grobs=plot_list[1:16],ncol=4, newpage = TRUE)
grid.arrange(grobs=plot_list[17:32],ncol=4, newpage = TRUE)
grid.arrange(grobs=plot_list[33:37],ncol=4, newpage = TRUE)

grid.arrange(grobs=plot_list,ncol=4, newpage = TRUE)

##Check pathway
dbs <- "KEGG_2021_Human"
enriched <- enrichr(cluster_all_gene$hgnc_symbol, dbs)
temp_df <- enriched[[1]]
temp_df <- temp_df %>% filter(str_detect(temp_df$Term, "pathway") )
####################################################################################
################################################
## ROC gene expression with clinical data
plot_list <- list()
auc_score <- list()
temp_df <- cluster_all_gene ## change here
mibc <- mibc %>% mutate(adjust_pN = case_when(mibc$ajcc_pathologic_n == "N0" ~ 0,
                                              mibc$ajcc_pathologic_n != "N0" ~1))
mibc <- mibc %>% mutate(adjust_pT = case_when(mibc$ajcc_pathologic_t == "T2" |
mibc$ajcc_pathologic_t == "T2a" | mibc$ajcc_pathologic_t == "T2b" ~ 0,
                                              mibc$ajcc_pathologic_t != "T2" & mibc$ajcc_pathologic_t !=
"T2a" & mibc$ajcc_pathologic_t != "T2b" ~ 1 ))
factor_base <- "T2"
factor_compare <- "T4"
for (i in temp_df$gene_id) {
  #i <- "ENSG00000087258.12"
  rocobj <- list()
  auc <- list()
  gene.name <- temp_df$hgnc_symbol[which(temp_df$gene_id == i)]
  gene.id <- temp_df$gene_id_new[which(temp_df$gene_id == i)]
  roc.gene <- i
  roc.expression <- data.frame(t(rlog.norm.counts.test.tcga[roc.gene,]))
  roc.expression$factor <- mibc$ajcc_pathologic_t
  colnames(roc.expression) <- c("gene","factor")
  roc.expression <- roc.expression %>% filter(factor %like% factor_base | factor %like%
factor_compare)
  roc.expression <- roc.expression %>% mutate(factor = case_when(roc.expression$factor
%like% factor_base ~ 0,
                                                                roc.expression$factor %like% factor_compare ~1))
  rocobj <- roc(roc.expression$factor,roc.expression$gene)
  auc <- round(auc(roc.expression$factor,roc.expression$gene),3)

  #create ROC plot
  plot_list[[i]] <- ggroc(rocobj, size = 1) +
    ggtitle(paste0(gene.name,' (',gene.id, ')')) +
    theme_bw() +
    labs(color='factor') +
    theme(plot.title = element_text(hjust = 0.5) ) +
    annotate("text", x=0.5, y=0.05, label=paste( "AUC:",auc ) )
  auc_score[[i]] <- auc
}
grid.arrange(grobs=plot_list[1:16],ncol=4, newpage = TRUE)
grid.arrange(grobs=plot_list[17:32],ncol=4, newpage = TRUE)
grid.arrange(grobs=plot_list[33:37],ncol=4, newpage = TRUE)

auc_score <- data.frame(matrix(unlist(auc_score), nrow=length(auc_score), byrow=TRUE),
                        row.names = temp_df$hgnc_symbol)

colnames(auc_score) <- c("cluster1","cluster2","cluster3")
```

VITAE

Name Mr.Tanan Bejrananda
Student ID 5910330031
Educational Attainment

Degree	Name of Institution	Year of Graduation
Bachelor of Medicine	Prince of Songkla University	2008
Diploma of Thai Board of Surgery	Prince of Songkla University	2014
Diploma of Thai Board of Urology	Prince of Songkla University	2016

Scholarship Awards during Enrolment

- Certificate of Presentation Award, an outstanding oral presentation award entitled "Impact of molecular subtyping in muscle invasive cancer by molecular expression clustering on predicting survival and response of treatment" at the 7th Joint Symposium BMS-BME-EU-HS: Post-graduate Health Science and Technology Conference. 2-3 September 2020

Work – Position and Address

- Doctor at Division of Urology, Department of Surgery, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand

List of Publication and Proceeding

- Bejrananda T, Kanjanapradit K, Saetang J, Sangkhathat S. Impact of immunohistochemistry-based subtyping of GATA3, CK20, CK5/6, and CK14 expression on survival after radical cystectomy for muscle-invasive bladder cancer. Sci Rep. 2021 Oct 27;11(1):21186. doi: 10.1038/s41598-021-00628-5. PMID: 34707176; PMCID: PMC8551252.