ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΙΑΤΡΙΚΗ ΣΧΟΛΗ
ΕΡΓΑΣΤΗΡΙΟ ΒΙΟΛΟΓΙΑΣ
ΚΑΙ
ΙΔΡΥΜΑ ΙΑΤΡΟΒΙΟΛΟΓΙΚΩΝ ΕΡΕΥΝΩΝ ΑΚΑΔΗΜΙΑΣ ΑΘΗΝΩΝ
ΤΟΜΕΑΣ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ
ΕΡΓΑΣΤΗΡΙΟ ΠΡΩΤΕΟΜΙΚΗΣ

Ανάλυση προωτεομικών δεδομένων απο φασματομετρία μάζας και ενσωμάτωσή τους με άλλα κλινικά και μοριακά δεδομένα σε κλινικά δείγματα και καρκινικές σειρές

ΡΑΦΑΗΛ ΣΤΡΟΓΓΥΛΟΣ
(ΠΤΥΧΙΟΥΧΟΣ ΒΙΟΛΟΓΙΑΣ)

ΑΘΗΝΑ 2023

ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΙΑΤΡΙΚΗ ΣΧΟΛΗ
ΕΡΓΑΣΤΗΡΙΟ ΒΙΟΛΟΓΙΑΣ
ΚΑΙ
ΙΔΡΥΜΑ ΙΑΤΡΟΒΙΟΛΟΓΙΚΩΝ ΕΡΕΥΝΩΝ ΑΚΑΔΗΜΙΑΣ ΑΘΗΝΩΝ
ΤΟΜΕΑΣ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ
ΕΡΓΑΣΤΗΡΙΟ ΠΡΩΤΕΟΜΙΚΗΣ

Analysis of mass spectrometry proteomics data and integration with publicly available transcriptomics and clinical data for the identification of molecular prognosticators in bladder cancer

ΡΑΦΑΗΛ ΣΤΡΟΓΓΥΛΟΣ
(ΠΤΥΧΙΟΥΧΟΣ ΒΙΟΛΟΓΙΑΣ)

ΑΘΗΝΑ 2023

ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΙΑΤΡΙΚΗ ΣΧΟΛΗ
ΕΡΓΑΣΤΗΡΙΟ ΒΙΟΛΟΓΙΑΣ
ΚΑΙ
ΙΔΡΥΜΑ ΙΑΤΡΟΒΙΟΛΟΓΙΚΩΝ ΕΡΕΥΝΩΝ ΑΚΑΔΗΜΙΑΣ ΑΘΗΝΩΝ
ΤΟΜΕΑΣ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ
ΕΡΓΑΣΤΗΡΙΟ ΠΡΩΤΕΟΜΙΚΗΣ

Analysis of mass spectrometry proteomics data and integration with publicly available transcriptomics and clinical data for the identification of molecular prognosticators in bladder cancer

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

- Ρουμπελάκη Μαρία: Αναπληρώτρια Καθηγήτρια, Εργαστήριο Βιολογίας, Ιατρική σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
- Αριστείδης Ηλιόπουλος: Καθηγητής, Εργαστήριο Βιολογίας, Ιατρική σχολή, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
- Καστρίτης Ευστάθιος: Καθηγητής, Βιοχημείας Ευκαρυωτικών Οργανισμών, Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΡΑΦΑΗΛ ΣΤΡΟΓΓΥΛΟΣ
(ΠΤΥΧΙΟΥΧΟΣ ΒΙΟΛΟΓΙΑΣ)

ΑΘΗΝΑ 2023

ΠΡΟΛΟΓΟΣ

Η παρούσα εργασία εκπονήθηκε στο εργαστήριο Πρωτεομικής του τομέα Βιοτεχνολογίας στο Ίδρυμα Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών (ΙΙΒΕΑΑ) σε συνεργασία με το εργαστήριο Βιολογίας της Ιατρικής Σχολής του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, στο πλαίσιο της διδακτορικής διατριβής μου.

Θα ήθελα να ευχαριστήσω την Αναπληρώτρια Καθηγήτρια Μαρία Ρουμπελάκη καθώς και την Δρ. Αντωνία Βλάχου για την ευκαιρία και τη δυνατότητα που μου έδωσαν να εκπονήσω τη διπλωματική μου εργασία στο διεπιστημονικό τομέα της βιοπληροφορικής και ανάλυσης πρωτεομικών δεδομένων. Θα ήθελα επίσης να ευχαριστήσω τους Καθηγητές Αριστείδη Ηλιόπουλο και Ευστάθιο Καστρίτη για τις διαφωτιστικές κατευθύνσεις που παρείχαν. Τους ευχαριστώ και τους δύο θερμά για τη καθοδήγηση, τις ιδέες, την επιστημονική βοήθεια και την άψογη συνεργασία.

Θα ήθελα να ευχαριστήσω τον Δρ. Ιερώνυμο Ζωιδάκη και τα υπόλοιπα μέλη του εργαστηρίου Πρωτεομικής για το αμείωτο ενδιαφέρον, την επιστημονική δράση και βοήθεια, τις ευκαιρίες, τις πολύτιμες συμβουλές, και για την καθοδήγηση που παρείχαν καθ' όλη τη διάρκεια εκπόνησης της διατριβής, καθώς για τις καρποφόρες συνεργασίες, για τη στήριξή τους και για τις όμορφες στιγμές που μοιραστήκαμε.

Θα ήθελα επίσης να ευχαριστήσω τους φίλους, Γκόλτσιο Θεόδωρο, Νάκο-Μπίμπο Μόδεστο, και Λούπη Παναγιώτη, και ιδιαίτερα την οικογένεια μου Μάρθα, Στέλιο. Ανδριάννα και Γεωργία για την ανεκτίμητη υλική και συναισθηματική βοήθεια που προσέφεραν, έτσι ώστε να περατώσω τις σπουδές με τους καλύτερους δυνατούς όρους.

Table of Contents

ΠΕΡΙΛΗΨΗ

Οι μοριακοί υπότυποι μιας ασθένειας συχνά συσχετίζονται με διαφορές ως προς την επιβίωση ή πρόοδο της νόσου και άλλοτε ως προς την απόκριση σε συγκεκριμένη θεραπεία. Την τελευταία δεκαετία, μελέτες μοριακής ταξινόμησης του ουροθηλιακού καρκίνου εστιάζουν κυρίως στον διηθητικό τύπο της ασθένειας (~20% των ασθένων στην αρχική διάγνωση) ο οποίος χαρακτηρίζεται από υψηλό κίνδυνο για μετάσταση και χαμηλά ποσοστά πενταετούς επιβίωσης. Οι παραπάνω μελέτες επέτρεψαν την ταυτοποιήση πολλαπλών γενομικών και μεταγραφικών υποτύπων οι οποίοι διαφέρουν ριζικά ως προς το μοριακό τους προφίλ, σχηματίζοντας δύο μεγάλες κατηγορίες: τους basal και τους luminal όγκους. Οι πρώτοι φαίνεται να σχετίζονται με πιο επιθετικούς καρκίνους εμπερικλείοντας όμως ένα σημαντικό ποσοστό ασθενών που ανταποκρίνονται στο βασικό χημειοθεραπευτικό σχήμα. Οι δέυτεροι (luminal) αρχικά προσδιορίστηκαν ως λιγότερο επιθετικοί, επόμενες μελέτες όμως αποκάλυψαν την σημαντική μοριακή ετερογένεια που τους χαρακτηρίζει και που αντανακλάται σε κλινικές παραμέτρους. Σήμερα, πιστέυεται ότι ο διηθητικός καρκίνος της ουροδόχου κύστης ταξινομείται σε 6 βασικούς υποτύπους, αλλά τα δεδομένα που υπάρχουν για να υποστηρίξουν την ένταξη των υποτύπων στην κλινική πράξη είναι ατελή και δεν συμφωνούν μεταξύ τους. Από την άλλη, ο μη διηθητικός τύπος της ασθενεις (~80% των περιπτώσεων στην αρχική διάγνωση) χαρακτηρίζεται από υψηλά ποσοστά υποτροπής και προόδου σε ανώτερο στάδιο καθώς και από σημαντικό δημόσιο οικονομικό κόστος εξαιτίας της αυξημένης συχνότητας παρακολούθησης που απαιτεί. Το μοριακό προφίλ του μη-διηθητικού καρκίνου έχει μελετηθεί σημαντικά λιγότερο από αυτό του διηθητικού, και μέχρι σήμερα υπάρχουν δύο μελέτες που επιχειρούν την ταξινόμησή του σε μοριακούς υποτύπους: η πρώτη στη βάση του μεταγραφώματος, η δέυτερη στη βάση της διακύμνασης αριθμού αντιγράφων. Το πρωτεομικό προφίλ όμως, τόσο του διηθητικού όσο και του μη-διηθητικού καρκίνου της ουροδόχου κύστης, μέχρι και σήμερα έχει μελετηθεί υποτυπωδώς. Σκοπός της παρούσας μελέτης είναι η διερεύνηση της ύπαρξης πρωτεομικών υποτύπων του μη διηθητικού ουροθηλιακού καρκίνου, ο μοριακός χαρακτηρισμός τους, η σχέση τους με προηγούμενα συστήματα ταξινόμησης, καθώς και η ταυτοποίηση απορυθμισμένων πρωτεϊνών και μονοπατιών με δυνητική προγνωστική αξία. Για την εξυπηρέτηση του παραπάνω σκοπού, 117 δείγματα καρκινικού ιστού από ασθενείς που πρωτοδιαγνώσθηκαν με ουροθηλιακό καρκίνο (98 μη-διηθητικό, 19 διηθητικό) συλλέχθησαν και το ολικό πρωτέομά τους απομονώθηκε και αρχικά ποσοτικοποιήθηκε

με τη μέθοδο Bradford. Κατόπιν διάσπασης με θρυψίνη, τα πεπτίδια διαχωρίστηκαν σε χρωματογραφική στήλη συνδεδεμένη με φασματογράφο μάζας τύπου Orbitrap. Οι φασματικές πληροφορίες για τα πεπτίδια αναλύθηκαν με το πρόγραμμα Proteome Discoverer θέτοντας FDR (False Discovery Rate) <0.01 και αντιστοιχήθηκαν σε πρωτεινικές ταυτότητες. Η πρωτεϊνική ποσοτικοποίηση έγινε με τη χρήση των τριών πιο άφθονων και μοναδικών πεπτιδίων ανά πρωτεΐνη, ενώ κατόπιν επεξεργασίας τα πρωτεομικά δεδομένα υποβλήθηκαν σε μια σειρά από υπολογιστικές αναλύσεις: μη επιτηρούμενη k-means συσταδοποίηση, ανάλυση κύριων συνιστωσών, ανάλυση για στατιστική σημαντικόντητα πρωτεϊνών, πρωτεϊνικών μονοπατιών, βιολογικών λειτουργιών και γονιδιακής έκφρασης καθώς και στην μοντελοιποίηση ενός μοριακού ταξινομητή Radnom Forest. Μέγιστη σταθερότητα συσταδοποίησης επιτεύχηκε για κ = 3 ομάδες, υποδηλώνοντας την ύπαρξη τριών πρωτεομικών υποτύπων στα δεδομένα. Η ομάδα 1 ήταν η μικρότερη σε μέγεθος (17/98), περιείχε κυρίως καρκίνους υψηλού σταδίου, αλλοίωσης και ρίσκου και παρουσίασε ένα μοριακό φαινότυπο ανοσοδιήθησης με υψηλά επιπέδα των μεταγραφικών παραγόντων STAT1, STAT3 και SND1, καθώς και πρωτεϊνων της αντιγονοπαρουσίασης, υποδηλώνοντας ενεργή ανταλλαγή πληροφοριών μεταξύ του ανοσοποιητικού και των καρκινικών κυττάρων. Παράλληλα, χαρακτηρίζονταν απο υψηλότερες ποσότητες πρωτεϊνών που συμμετέχουν στο κυτταρικό κύκλο, και στη μετάδοση στρεσογόνων σημάτων (αντίδραση μη αναδιπλωμένης πρωτεϊνης και επιδιόρθωση βλαβών του DNA). Η όμαδα 2 συγκέντρωσε ασθενείς με ποικίλα κλινικά χαρακτηριστικά που όμως έφεραν κοινώς, αυξημένες ποσότητες εξωκυττάριων πρωτεϊνών (στρώματος), και χαμηλά επιθηλιακά σήματα. Οι ασθενείς στην ομάδα 3 παρουσίασαν έναν πιο διαφοροποιημένο μοριακό φαινότυπο με υψηλότερα επίπεδα (UPKs και KRT20 κάθως και CDH1) που συμβαδίζει με τα κλινικά χαρακτηριστικά τους αφού οι περισσότεροι διαγιγνώσθηκαν με καρκίνους χαμηλού σταδίου και κινδύνου. Η ανάλυση για ενεργοποιημένα πρωτεϊνικά μονοπάτια έδειξε ότι οι ασθενείς της ομάδας 1 έιχαν ενεργή σηματοδότηση για βιοσυνθετικές διεργασίες, για ιντερφερόνη-γ, και αυξημένη δραστηριότητα των μεταγραφικών παραγόντων MYC και E2F, που ελέγχουν θετικά τον κυτταρικό κύκλο. Από την άλλη οι ασθνενείς της ομάδας 3 σχετίστηκαν με ενεργοποίηση μεταβολικών μονοπατιών όπως αυτό της αποτοξίνωσης μεσολαβούμενο από γλουταθειόνη καθώς και της γλυκογονόλυσης – γλυκόλυσης, αλλά και της απόπτωσης. Συγκρίνοντας το πρωτεομικό προφιλ των ασθένων με μη-διηθητικό καρκίνο με ασθενέις που είχαν διηθητικό καρκίνο χρησιμοποιώντας ανάλυση κύριων

συνιστωσών, αποκαλύφθηκε κοντινή σχέση της ομάδας 1 με ασθενείς που έφεραν διηθητικό ουροθηλιακό καρκίνο και αντίστροφα, μακρινή σχέση της ομάδας 3 με τους τελευταίους. Η ομάδα 2 εμφάνισε μεγάλη διασπορά επικαλύπτοντας περιοχές των προηγούμενων δύο ομάδων. Για την επικύρωση των πρωτεομικών αποτελεσμάτων, δεδομένα από μεταγραφικές έρευνες (UROMOL και LUND) αναλύθηκαν αναδρομικά. Στην UROMOL έρευνα επίσης ταυτοποιήθηκαν 3 υπότυποι ο ένας εκ των οποίων συγκέντρωσε τους περισσότερους ασθενείς με πρόδοο σε ανώτερο στάδιο (κακής πρόγνωσης υπότυπος). Συγκριτική ανάλυση μεταξύ των τριών πρωτεομικών ομάδων και των τριών υποτύπων της UROMOL έρευνας με το στατιστικό εργαλείο GSEA, έδειξε στατιστικώς σημαντικές φαινοτυπικές ομοιότητες μεταξύ της πρωτεομικής ομάδας 1 και του υποτύπου «κακής» πρόγνωσης της UROMOL καθώς και μεταξύ της πρωτεομικής ομάδας 3 και του υποτύπου «καλής πρόγνωσης». Χρησιμοποιώντας έναν μη επιτηρούμενο μοριακό ταξινομητή Random Forest, οι υψηλού κινδύνου και χαμηλού κινδύνου φαινότυποι των πρωτεομικών ομάδων 1 και 3, επιβεβαιώθηκαν ύστερα από την ταξινόμηση των ασθενών στους υποτύπους «κακής» και «καλής» πρόγνωσης αντίστοιχα, της UROMOL έρευνας. Στατιστικώς σημαντικες πρωτεΐνες που ξεχωρίζουν αυτές τις δυο ακραίες πρωτεομικές ομάδες αλλά και ταυτόχρονα τον διηθητικό από τον μη διηθητικό καρκίνο βρέθηκαν να διαφέρουν σημαντικά και στο επίπεδο του μεταγραφώματος μεταξύ των ομάδων «κακής» και «καλής» πρόγνωσης σε δύο ανεξάρτητες έρευνες (UROMOL και LUND). Τα παραπάνω μόρια συμμετέχουν σε βιολογικές λειτουργίες-κλειδιά για την ανάπτυξη του μη-διηθητικού καρκίνου, όπως στην επαγωγή αποκρίσεων πρωτεϊνικής σταθερότητας, στη σηματοδότηση κυτοκινών και ιντερφερονών, στην αντιγονοπαρουσίαση, στην επεξεργασία πρώιμων mRNAs, σε μετα-μεταφραστικές τροποποιήσεις αλλά και σε μονοπάτια κυτταρικής αύξησης. Συνολικά, η παρούσα μελέτη ταυτοποιεί τρεις πρωτεομικούς υποτύπους του μη διηθητικού καρκίνου και ακολουθώντας μια σύγκριτική ανάλυση με δύο ανεξάρτητες μεταγραφικές έρευνες, παρέχει ομάδες μορίων που μπορεί να οδηγούν τη πρόοδο του καρκίνου και που χρειάζονται επιπλέον επικύρωση στη κλινική πράξη.

ABSTRACT

DNA/RNA-based classification of Bladder Cancer (BC) supports the existence of multiple molecular subtypes, while investigations at the protein level are scarce. The

purpose of this study was to investigate if Non-Muscle Invasive Bladder Cancer (NMIBC) can be stratified to biologically meaningful proteomic groups, to establish associations between the proteomics subtypes and previous transcriptomics classification systems and to characterize the continuum of transcriptomics alterations observed in the different stages of the disease. Subsequently, tissue specimens from 117 patients at primary diagnosis (98 with NMIBC and 19 with MIBC), were processed for high resolution LC-MS/MS analysis. Protein quantification was conducted by utilizing the mean abundance of the top three most abundant unique peptides per protein. The proteomics output was subjected to unsupervised consensus clustering, principal component analysis (PCA), and investigation of subtype-specific features, pathways, and genesets, as well as for the construction and validation of a Random Forest based classifier. NMIBC patients were optimally stratified to 3 proteomic subtypes (classes), differing at size, clinico-pathological and molecular backgrounds: Class 1 (mostly high stage/grade/risk samples) was the smallest in size (17/98) and expressed an immune/inflammatory phenotype, along with features involved in cell proliferation, unfolded protein response and DNA damage response, whereas class 2 (mixed stage/grade/risk composition) presented with an infiltrated/mesenchymal profile. Class 3 was rich in luminal/differentiation markers, in line with its pathological composition (mostly low stage/grade/risk samples). PCA revealed a close proximity of class 1 and conversely, remoteness of class 3 to the proteome of MIBC. Samples from class 2 were distributed in a wider fashion at the rotated space. Comparative analysis with GSEA between the three proteomic classes and the three UROMOL subtypes indicated statistically significant associations between the proteomics class 1 and UROMOL subtype 2 (subtype with a bad prognosis) and also between the proteomics class 3 and UROMOL subtype 1 (subtype with the best prognosis). Utilizing a Random Forest based classifier, the predicted high- and low-risk phenotypes for the proteomic class 1 and class 3, were further supported by their classification into the “progressed” and “non-progressed” subtypes of the UROMOL study, respectively. Statistically significant proteins distinguishing these two extreme classes (1 and 3) and also MIBC from NMIBC samples were found to consistently differ at the mRNA levels between NMIBC “Progressors” and “Non-Progressors” groups of the UROMOL and LUND cohorts. Functional assessment of the observed molecular de-regulations suggested severe pathway alterations at unfolded protein response, cytokine and inferferone-γ signaling, antigen presentation, mRNA processing, post translational modifications and

in cell growth/division. Collectively, this study identifies three proteomic NMIBC subtypes and following a cross-omics analysis using transcriptomic data from two independent cohorts, shortlists molecular features potentially driving non-invasive carcinogenesis, meriting further validation in clinical trials.

1. BLADDER CANCER

1.1 Anatomy of the bladder

The bladder is a hollow organ located in the lower part of the abdomen (**Figure 1**). The anatomy of the bladder resembles a small balloon and bares a muscular wall that allows the adaptation of the shape of the organ depending on the volume of the urine produced by the kidney. The human body comprises of two kidneys, one on each side of the backbone, above the waist and tiny tubules in the kidneys filter the circulating blood. The urine passes from each kidney through the ureter into the bladder up until the urine passes through the urethra to be eliminated from the body.[1]

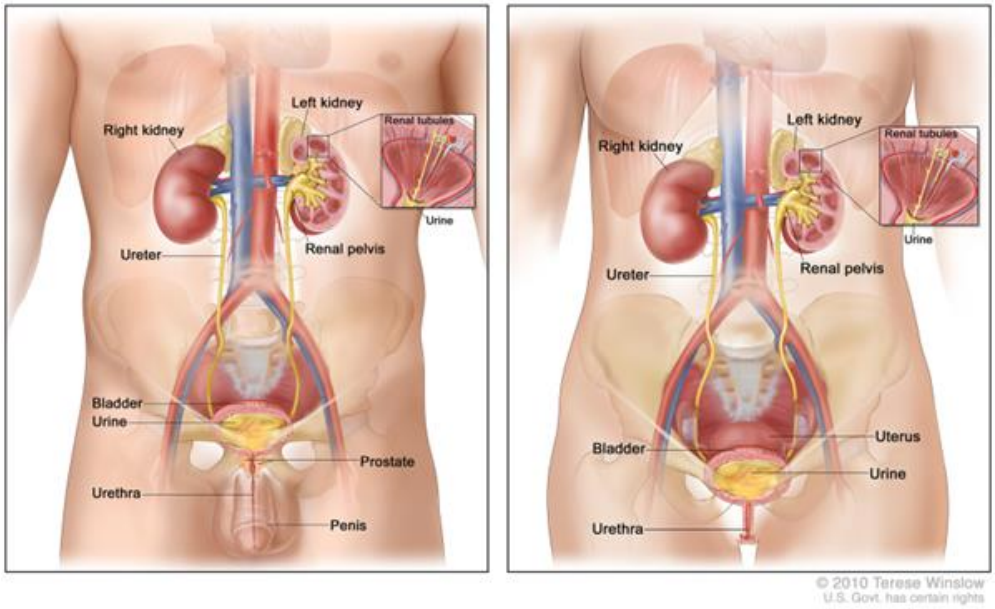


Figure 1: *Anatomy of the male urinary system (left panel) and female urinary system (right panel) showing the kidneys, ureters, bladder, and urethra. (source: National Cancer Institute - Visuals Online, Urinary System, Creator: Terese Winslow, 2010. https://visualsonline.cancer.gov/details.cfm?imageid=9098 https://visualsonline.cancer.gov/details.cfm?imageid=9050)*

The bladder wall consists of three different tissue types: epithelium, sub-epithelial loose connective tissue (lamina propria), and detrusor or muscularis propria (**Figure 2**). The epithelium of the bladder is stratified (multi-layered) and has evolved to withstand

mechanical (distention) and chemical insult. This type of epithelium is called transitional and is found also in the ureters and in urethra, hence, sometimes called urothelium. The layers of the urothelium can be stratified to three morphologically different zones: basal, intermediate and superficial. The basal zone, locates at the bottom of the epithelium and consists of a single layer of cuboidal cells, adhered to the basal (or basement) membrane, with the latter separating epithelium from lamina propria. In the normal urothelium, basal cells are $CD44^+$, $KRT5/6^+$ and $Ki67^+$ and are considered to be the most undifferentiated, harboring stem cell properties and also being susceptible to malignancy (1). Several layers of cells (3-6 based on the distention state) appearing with a spherical shape above the basal layer, comprise the intermediate zone. The intermediate zone has a high renewal capability since it includes stem cells and progenitors of other more differentiated cells. The upper part of the urothelium is the superficial zone. Here, the residing cells have a varying morphology that depends on the state of distention: when the tissue is relaxed superficial cells appear cuboidal, whereas upon expansion they take a squamous morphology (thin, flat plates). Superficial cells are well-differentiated, and have an extended golgi apparatus that allows for the synthesis of a thick impermeable keratin-based membrane. Their tight stratification across the epithelium is achieved via the establishment of tight junctions, rich in cadherin-1 (CDH1) and in catenins α-, β-, γ (2).

1.2 Bladder cancer epidemiology

According to the Global Burden of Disease study (3) there were 3.4 million cases of Bladder Cancer (BC) between 2005 and 2015, while in 2018, BC has led to an estimated of 17,200 and 200,000 deaths in the United States (4) and worldwide (5), respectively. Approximately three out of four BC patients present with non-muscle invasive disease (NMIBC), with the majority of them requiring lifelong monitoring and surveillance. BC is characterized by high prevalence, multiple recurrences and increased progression rates, making it the costliest type of cancer to treat (6). Considering also that epidemiological data predict a global increase on the incidence rates of the disease (7), there is an imperative need to radically improve the management of BC. Tobacco smoking, drinking of arsenic-contaminated water, industrial exposure to chemical carcinogens, infestations and irritations of the bladder, as well as familial history of concordant cancers have been linked to bladder carcinogenesis (7). Global

improvements in health care organizations, diagnostic tools and in therapeutics have certainly contributed to the observed decline in the mortality rates (7). However, the high molecular heterogeneity of the disease renders available treatment options non-effective for a number of patients, and the clinicopathologic parameters insufficient for predicting outcome. This is reflected at the high numbers of disease recurrence (~80%) and progression (~25%) for the NMIBC and also at the relatively low 5-year survival rates of the Muscle Invasive Bladder Cancer (MIBC) patients (46~65%) as well as of the metastasized cases (~15%) (8).

1.3 Bladder cancer staging

Transitional cell carcinoma cases were initially called superficial bladder cancer. However, malignant urothelial tumors confined to the bladder mucosa (urothelium and lamina propria compartments) are accurately termed non-muscle invasive (NMIBC) instead of being given the traditional "superficial" label. The traditional term suggested that all such tumors shared the relatively benign course of low grade papillary tumors. In contrast, patients with highly malignant lesions, including carcinoma in situ (CIS), actually have a worse prognosis if not recognized and treated successfully. For this reason, the staging system for bladder was updated in 2017 -American Joint Committee on Cancer/tumor, nodes, metastases (AJCC/TNM) staging system (9). Tumor spread in BC is determined according to the TNM Classification of Malignant Tumors (TNM) system (**Figure 2**). TNM is an acronym with T describing the size and depth of the tumor bulk through the bladder wall, N denotes affected nearby lymph nodes and M informs for occurrence of metastatic lesions at other parts of body.

Primary Tumor (T) T0 is used when primary tumor is not identified in the initial diagnosis in the biopsy or transurethral resection. Ta and Tis represent non-invasive papillary urothelial carcinoma and flat urothelial carcinoma in situ (CIS), respectively. T1 is used for invasion into the lamina propria and are usually high-grade tumors. Several features in the biopsy specimens are helpful for the grade determination of stromal invasion such as single cell infiltration, absence of basement membrane, finger-

like projections and stromal desmoplastic, or inflammatory reaction. Papillary stalk invasion of an exophytic lesion is considered lamina propria invasion.

T2 stage characterizes the invasion into the muscularis propria. T2 is further divided based on invasion into superficial (inner half) (T2a) or deep muscularis propria (outer half) (T2b). The biopsy and TUR specimen may be problematic in determining the depth of the invasion since the samples may not contain muscularis propria. Therefore, repeated procedures may be required to evaluate the extend of the invasion. If the invasion into the muscularis propria or muscularis mucosae is uncertain, it should be clearly stated in the examination of the biopsy.

T3 grade represents the invasion of the tumor into the perivesical fat. T3 is further divided according to microscopic perivesical fat invasion (T3a) or macroscopic invasion forming extravesical mass (T3b). Fat tissue can be found at all layers of urinary bladder wall, the biopsy specimens cannot distinguish whether the invasion has been into the perivesical fat.

T4 is used when primary tumor invades beyond urinary bladder. T4a characterizes primary tumors that exhibit invasion into prostatic stroma, seminal vesicles, uterus, or vagina in female patients, whilst T4b is used to define distant metastatic lesions.

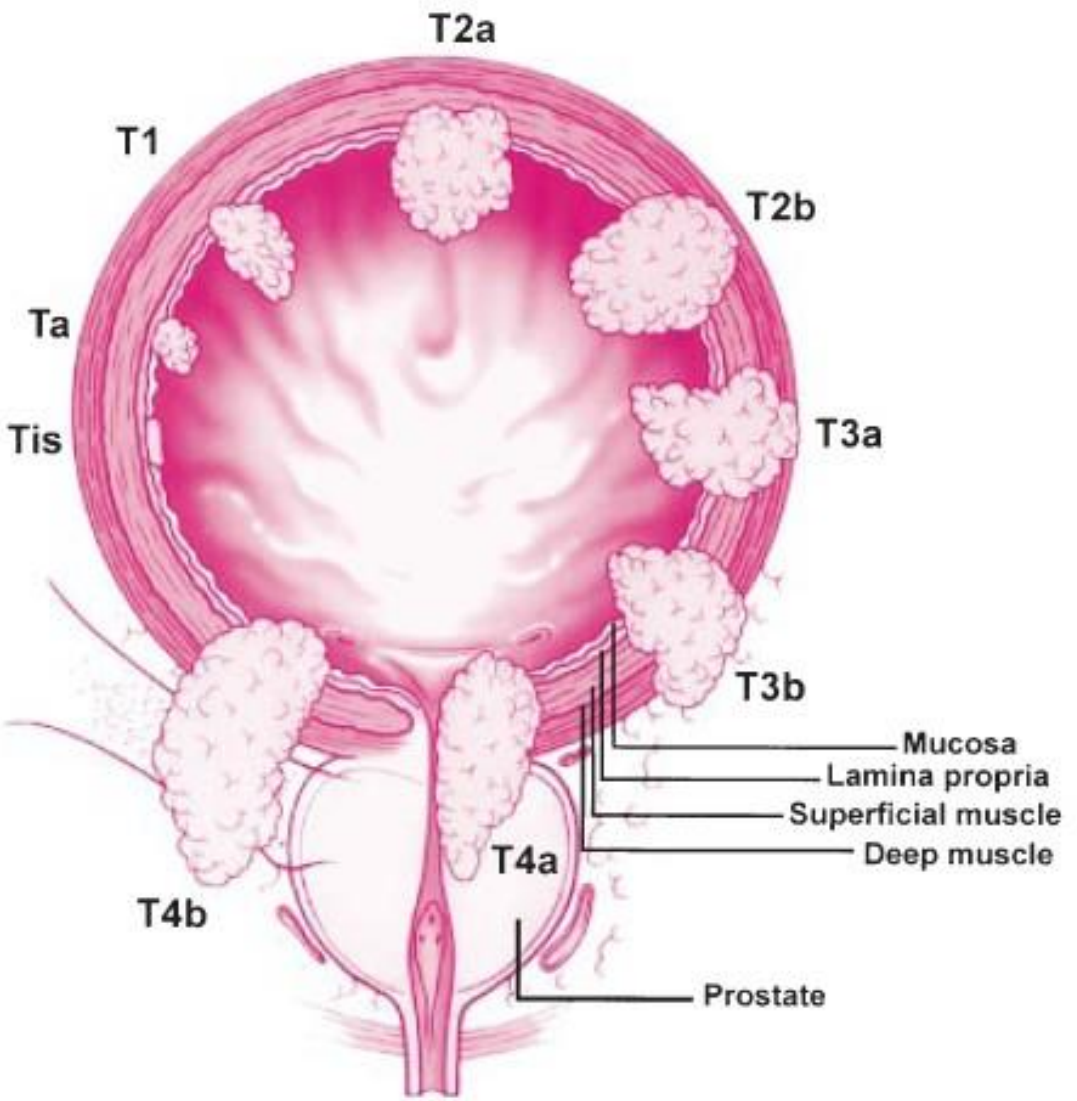


Figure 2: T categories of urothelial carcinoma of urinary bladder. *Ta: non-invasive papillary urothelial carcinoma, T1: invasive urothelial carcinoma with lamina propria invasion, T2: invasive urothelial carcinoma with muscularis propria invasion, T3: invasive urothelial carcinoma with perivesical fat invasion. T4: metastatic lesions. Tis (or CIS) is a high-grade, flat malignancy confined to the urothelium.*(*doi: 10.1594/ranzcr2011/R-0150*).

1.4 Histological features of bladder cancer

Based on the depth of invasion inside the bladder wall, non-metastatic BC is classified either as Mucle-Ivasive (MIBC; 20-30% of total cases) or Non-Muscle Invasive (NMIBC; 70-80% of total cases). MIBC includes tumors diagnosed at stage > T2 while NMIBC can be either Ta (spread of cancer cells is limited to the epithelium) or T1 (cancer cells have invaded lamina propria). Clinicians have long recognized that NMIBCs usually present with a papillary morphology extending from the epithelium to the bladder cavity, while conversely, MIBCs tend to lack it.

Bladder cancer can be classified histologically as urothelial (also called papillary/transitional cell carcinoma) or non-urothelial. Urothelial cancer has a propensity for divergent differentiation including, amongst others, squamous, glandular, micropapillary, nested, lymphepithelioma-like, plasmacytoid and sarcomatoid variants of urothelial cancer. Non-urothelial tumors are rare, more aggressive than urothelial (10, 11) and include the pure forms of squamous, sarcoma, adenocarcinoma, carcinosarcoma, paraganglioma, melanoma and lymphoma (12). Mostly, the current evidence suggests that urothelial cancer with divergent differentiation has a worse prognosis when compared with pure urothelial cancer (13), with genetic-based studies indicating that the histologic variants of urothelial cancer arise from a common clonal precursor (14, 15). Attempts to quantify the amount of divergent differentiation present, such as using the nonconventional differentiation number, have been made recently, which is anticipated to improve the ability to compare publications from different centres. The vast majority of BCs present with one or more of the following three epithelial cancer types:

Papillary/Transitional cell carcinoma (TCC): It refers to cancers initiating from the intermediate zone, acquiring a papillary conformation that grows inside the bladder cavity. This is the predominant histological type including even up to 90% of cases in Western Europe and the United States. Papillary cases often present with multi-focal tumors (~40%) and according to the guidelines of the World Health Organization grading system of 2004 (16), cytological findings can stratify TCC into low-grade or high-grade subcategories:

-Low-grade transitional cell carcinoma often recurs after treatment, but rarely spreads into the muscle layer of the bladder.

-High-grade transitional cell carcinoma often recurs after treatment and may proceed to the muscle invasive type of bladder cancer. Metastasis to other parts the body and to lymph nodes can also occur. High-grade disease is the type responsible for the majority of deaths from bladder cancer. However, data comparing the prognostic potentiality of the older, three-tiered (Grade 1/2/3) grading system of WHO (1973) (17), against the newer, two-tiered (High/Low grade) (16) for NMIBC, show debatable results (18, 19), while clinicians and researches are allowed to use any of the two classifications.

Carcinoma in-situ (CIS): Tumors that grow longwise, inside the epithelium as flat dysplasia and are generally thought to progress to MIBC more often. Lesions of CIS can co-occur near the primary tumor and are usually resistant to neo-adjuvant chemotherapy.

Squamous cell carcinoma (SCC): Cancer that begins in squamous cells (thin, flat cells lining the inside of the bladder). Squamous cells can be distinguished by their hexagonal shape, by the typical inter-cellular bridges they form and also their positive staining for KRT14 and KRT5. SCC is more frequent in places with high prevalence of *Schistosoma haematobium* infection, as in East Africa and the Middle East (20).

1.5 Bladder cancer diagnosis

The diagnosis of bladder cancer is based on several procedures that facilitate the detection of malignant morphology in the tissue. Currently, the gold standard for the diagnosis of BC is cystoscopy which allows optical evaluation of the bladder structure. During the cystoscopy procedure, it is also feasible to obtain bladder biopsies for histological evaluation. This method is also called transurethral resection of bladder tumor (TURBT) and can be used for excising a tumor with papillary morphology (described below). Cystoscopy can be combined with urine cytology in which urine samples are examined under the microscope, and the clinician is seeking to identify atypical or malignant cells and the degree of their morphological alteration. The role of imaging in the diagnosis of cancer is undisputable. Therefore, in the bladder cancer setting several imaging test can be applied for the diagnosis of the disease such as computerized tomography (CT) (21), urogram or retrograde pyelogram (22), CT urogram allows a thorough evaluation of the urinary tract in order to detect any areas

that may be affected by the disease. In contrast, the retrograde pyelogram is use for the examination of the upper urinary tract.

1.6 Altered molecular pathways in bladder cancer

Studies between NMIBC and MIBC have revealed hallmark differences in their genomic backgrounds, establishing the dual track concept of bladder carcinogenesis (papillary and non-papillary; **Figure 3**). In the papillary route (characterizes NMIBC patients) tumor initiation is believed to take place in the intermediate zone. The most profound alterations in these cancers involve activating mutations in the FGFR3/HRAS pathway (~80%) which are considered to transform an early urothelial hyperplasia into a non-invasive papillary tumor (23, 24). The superficial zone of these tumors is typically rich in KRT20, UPK3 and CDH1, whereas KRT7/17, KRT8/18, p63 appear at highest levels in the intermediate zone (25). Intermediate cells may or may not be $KRT5^{+}$ and $CD44^{+}$(25). Progression of low- to high-grade malignancy is often accompanied by further alterations in the Akt/PIK3CA/mTOR pathway (**Figure 4**) which signals for cell growth, often by deletion of TSC1 (negative regulator of mTOR) and loss of function of the tumor suppressors STAG2 and CDKN2A (26, 27). On the other hand, the non-papillary route (characterizes MIBC patients), is thought to initiate from the basal layer and involves alterations mainly in the TP53/RB1 pathway that controls cell-cycle checkpoint, as well as deletion of PTEN which is an inhibitor of the Akt/PIK3CA/mTOR pathway (27). Up to 20% of MIBCs present with mutations in FGFR3, and CDKN2A (27), possibly suggesting origins from a low-grade hyperplasia that evolved to invasive disease. Clonal expansion in the non-papillary route spreads towards the muscle layer, and tumors are usually $N\text{-}Cad^{+}$, $KRT5/6^{+}$, $KRT14^{+}$, $CD44^{+}$, lack expression of KRT20 and CDH1 (25), while uniform staining patterns of KRT5 through the tumor parenchyma might accompany squamous differentiation (28). In general, both MIBC and NMIBC typically experience loss of multiple genetic loci at 9q, mutated TERT promoter and aberrant patterns of chromatin remodeling (29, 30). However, MIBC has higher mutational burden, more frequent copy number variations (CNVs), chromosomal translocations and heavier genomic instability (29).

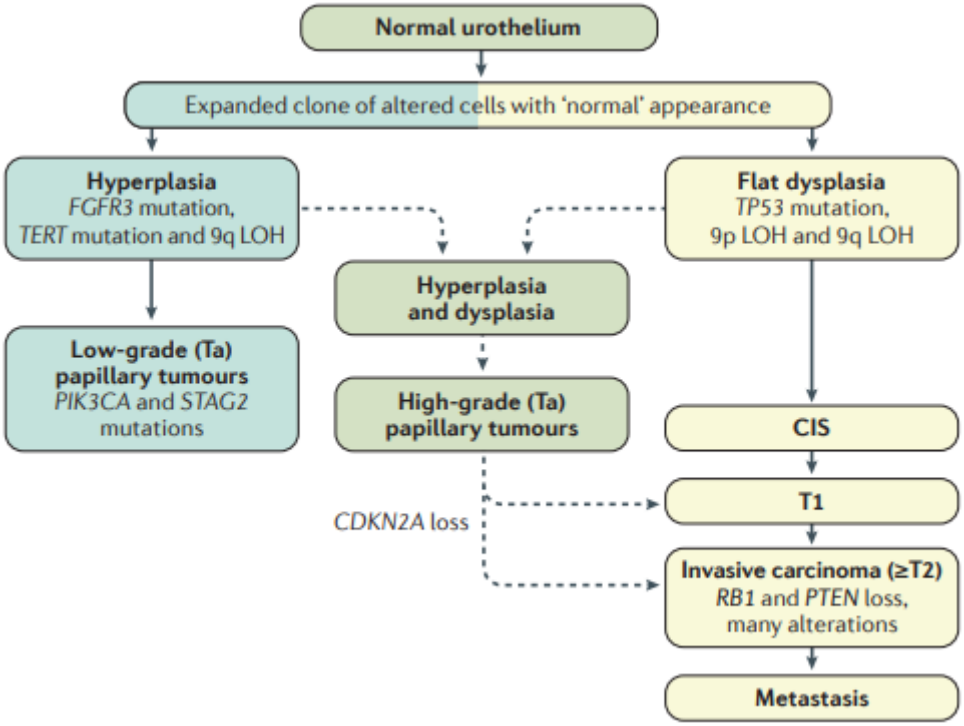


Figure 3: The dual track concept of bladder carcinogenesis. *The two distinct pathways of pathogenesis of papillary (leading to NMIBC) and solid (leading to MIBC) are shown. Low-grade papillary tumors can arise via simple hyperplasia and minimal dysplasia, and are characterized at the molecular level by loss of heterozygosity (LOH) of chromosome 9 and activating mutations of genes encoding fibroblast growth factor receptor 3 (FGFR3), telomerase reverse transcriptase (TERT), phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform (PIK3CA) and inactivating mutations of STAG2 (which encodes cohesin subunit SA-2). The aforementioned genes participate in biological processes such as cell proliferation, division and growth. MIBC is considered to arise via flat dysplasia and carcinoma in situ (CIS), which commonly harbors TP53 mutations in addition to LOH at chromosome 9, but fewer FGFR3 mutations. Low-grade papillary NMIBCs might progress to MIBCs as a result of CDKN2A (which encodes p16 and p14ARF) loss. Numerous potential differences in the molecular pathways to the major tumour types and their subtypes are known. Solid arrows indicate pathways for which there is histopathological and/or molecular evidence; uncertainty is indicated by dashed arrows.*(*source:* ref (23))

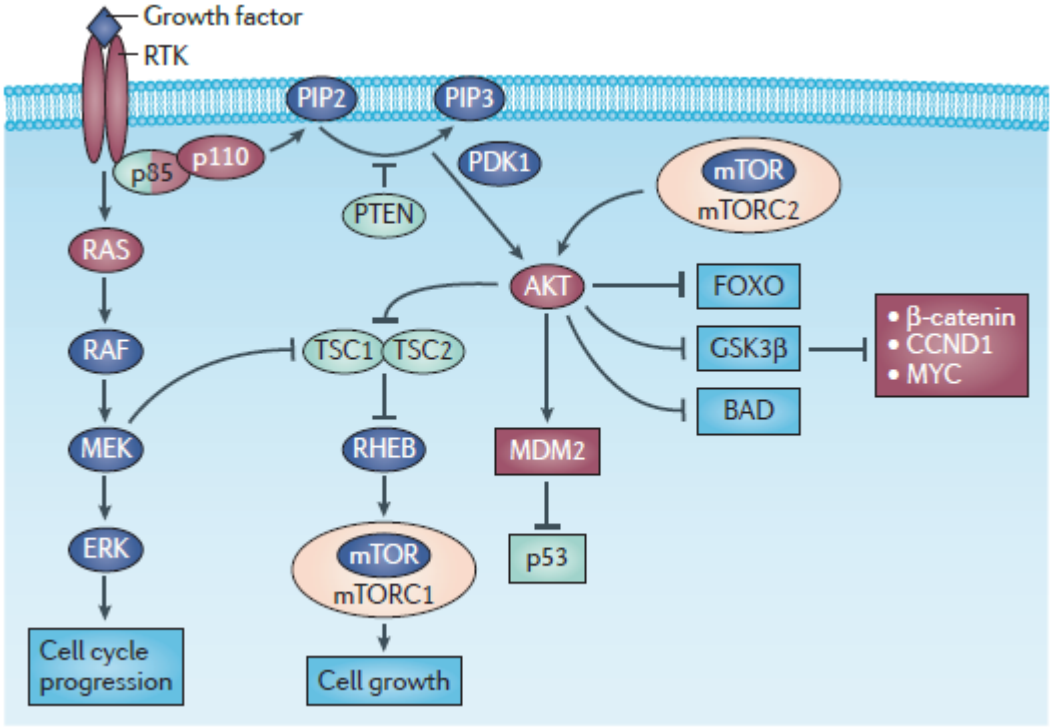


Figure 4: The Akt/PI3K/mTOR pathway in bladder cancer. *Growth factor-mediated signaling or mutational activation of both PI3K and MAPK pathways is common in bladder cancer. Receptor tyrosine kinases (RTKs), epidermal growth factor receptor (EGFR), ERBB2, ERBB3, fibroblast growth factor receptor 1 (FGFR1) and FGFR3 may be activated by ligand, overexpression and/or mutation in bladder cancer. Through adaptor proteins, these RTKs activate RAS. Signalling via the RAS–RAF–MEK–ERK cascade leads to phosphorylation of many substrates that can have multiple cellular effects depending on the intensity and duration of signalling. In many situations proliferation is induced. Activated RTKs bind to p85 (the regulatory subunit of PI3K) and recruit the enzyme to the membrane, where it phosphorylates phosphatidyinositol-4,5-bisphosphate (PIP2) to generate PIP3. Activated RAS can also directly activate PI3K. PIP3 recruits 3-phosphoinositide-dependent protein kinase 1 (PDK1; also known as PDPK1) and AKT, resulting in activation of AKT by phosphorylation, which leads to both positive and negative regulation of a wide range of target proteins (not all shown). Cyclin D1 (CCND1) and MDM2 are upregulated directly or indirectly, resulting in a positive stimulus via the RB or p53 pathways, respectively. AKT also phosphorylates and inactivates tuberous sclerosis 2 (TSC2), leading to activation of mTOR complex 1 (mTORC1), which controls protein synthesis. TSC1 forms an active complex with TSC2, and loss of function of either protein leads*

to dysregulated mTOR signalling. AKT phosphorylates and inactivates glycogen synthase kinase 3β (GSK3β), relieving its suppression of β-catenin, which is freed to enter the nucleus and activate gene expression. MYC expression is induced as a consequence of both ERK and AKT signalling. Key genes that are activated in bladder cancer are shown in dark red and those that are inactivated in green. BAD, BCL-2-associated agonist of cell death; FOXO, forkhead box O; RHEB, Ras homologue enriched in brain.(*source:* ref (27))

2. TRANSCRIPTOMICS AND PROTEOMICS

Over the last two decades, significant technological advancements in the molecular biology have enabled the study of molecules in large scales. With the advent of next generation sequencing and mass spectrometry analyzers, we can now quantify the abundances of thousands of cellular products, such as RNA, proteins, metabolites. Analysis of big data requires the usage of several advanced statistical and mathematical principles, along with computative tools and dedicated algorithms designed solely for this purpose. The subfield of molecular biology that leverages the above to investigate large scale molecular alterations, is called *systems biology*, or alternatively *omics*, and and depending on the cellular product at study, is referred to as genomics, methylomics, transcriptomics, proteomics, metabolomics, interactomics etc. Transciptomics, proteomics and metabolomics are the three quantitative omics, in a sense that not only they identify a particular molecule, but they also quantify its expression or abundance levels. The number of identified features per omic usually depends on the starting sample material, but generally in transcriptomics, it is at the scale of tens of thousands, in proteomics at single digit thousands, while for metabolomics is currently at tens or hundreds.

2.1 Transcriptomics

2.1.1 Microarrays

The technology of microarrays is based on the design of multiple DNA probes that are bound on a solid surface such as a glass slide. Microarray technologies have been widely used in research for measuring gene expression changes and elucidating the

relationship between genotypes and phenotypes. They are quite cost-effective for profiling gene expression when it comes to model organisms. Microarrays have also been used in clinical diagnostics. Some examples are the detection of copy number variants using SNP arrays such as the Cytogenetics Whole-Genome Array from Affymetrix or the HumanOmni1-Quad BeadChip and HumanCytoSNP-12 56 Integration of Omics Approaches and Systems Biology for Clinical Applications DNA Analysis BeadChip from Illumina. In general, microarrays can be used for general screenings, gene expression profiling, genotyping, and many other applications. However, like in PCR-based applications, the use of predefined oligonucleotides (probes) is based on previous knowledge availability. Thus, microarrays are used for quantification of known sequences and not for the discovery of new variants, transcripts, or other unexpected transcriptomics features (31). In order to fully illustrate the limitations of microarray technology, we should briefly present some basic concepts. Microarray detection is based on hybridization of sample DNA to nucleic acid probes, bound to the surface of a slide. The probes are oligonucleotides with a usual length of 25–120 nucleotides. To further measure the quantity of hybridization to each specific probe, the target sequence (DNA or cDNA) is labeled with fluorescent dyes. Then, after an image is taken and processed, signal intensities can be read and converted to normalized values in order to initiate the data analysis. Due to the nature of microarray probe design, the capabilities of this method are apparently restricted to known sequences and therefore do not allow detection of target sequences beyond the current knowledge. This factor can be a disadvantage for non-model organisms, but diagnostics of well-characterized organisms, such as humans, is feasible, although it relies on the quality of the available bioinformatics data at the moment the microarray was designed. Microarrays can be used for diagnostic transcriptome analysis. If properly designed, they will not only provide information on gene expression and expressed SNPs but also detect exon junctions and fusion genes (32). Normalization and processing of microarray data can involve quite complex bioinformatics methodologies and statistics. This is a consequence of the nature of the data produced by this technology that may become a limitation for someone not acquainted in the area. However, significant efforts were put into developing standardized procedures for microarray analysis. Some of these procedures as well as suggestions, guidelines, metrics, and thresholds, among other information, are publicly available under the MicroArray Quality Control (MAQC) website, together with the publications that

helped to reach consensus on these procedures. Refer to the MAQC project for further details [49].

2.1.2 Sequencing

The advances in DNA sequencing, and in particular the advances of NGS, have significantly improved the quantity and quality of genomic information that can be obtained from clinical samples. The reduced cost of NGS as well as the increase in throughput made whole- genome sequencing (WGS), as well as other NGS applications such as whole-exome sequencing (WES) or RNA-Seq, a possible and reliable approach for clinical diagnosis. However, there are still some challenges such as data storage, management, analysis, and interpretation that have to be considered for the proper use of this technology in clinical applications (33). Following the objectives of this chapter, the tools, applications, approaches, and examples presented here will mainly focus on the use of NGS for the analysis of the transcriptome in clinical applications. Many different platforms for massive parallel sequencing were developed. The first example, although currently obsolete, is the 454 Genome Sequencer from Roche Applied Sciences. Also outdated is the SOLiD platform from Life Technologies. The current and most widely used technology is the Solexa "Sequencing-by-Synthesis" technology that was acquired by Illumina in 2007. The strength of these technologies relies on a very high throughput at the expense of read accuracy and much shorter read length when compared with the well-known Sanger sequencing. However, the possibilities of use and applications of this technology led to significant scientific discoveries and diagnostic applications (33). Fortunately, some of the trade-offs are being reduced through continuous platform improvements and developments, which resulted in more advanced sequencer versions such as the Ion Torrent and Ion Proton from Life Technologies and the MiSeq and HiSeq from Illumina. In particular the HiSeq versions have greatly improved in accuracy and read length as well as in significantly higher throughput. Meanwhile, the run time has been decreasing, making it suitable for diagnostic use. Advances and ongoing efforts to improve these platforms even further have made the HiSeq platforms from Illumina the most widely used NGS sequencers. Depending on the sequencing platform of preference, many options are available for library preparations. The library preparation steps include all transformations the nucleic acids of interest may require prior to being completely ready for sequencing on the platform of choice. In general, NGS library preparations for transcriptomics consist

of cDNA synthesis and extension of the cDNA with specific ligated adapters for sequencing. Furthermore, it is quite common that a minimum quantity of RNA is required to ensure a minimal quality. For body fluids and tissues, approximately 10ng of RNA is often sufficient, while for samples containing degraded RNA, such as FFPE, a minimum of 100 ng is strongly recommended (34). In addition, many adaptations to library preparation protocols are reported in order to cover different aspects of the complexity of RNA processes and regulations such as posttranscriptional modifications, gene expression, isoforms, regulation, splicing, and degradation (35-38). For a better overview of published protocols, please refer to available collections of preparation methods such as the sequencing methods review published from Illumina Technology (39). The overwhelming quantity of data produced per sample requires advanced bioinformatics analysis to address the wide variety of possible questions. There are many tools and software packages available that can analyze these massive datasets, make inferences from the data, and offer biological interpretations. Despite their differences, there are some data analysis steps that are usually shared among the different approaches. Common steps include quality check of the sequencing data, sequence alignment to a reference genome or de novo assembly in some other cases, and the assessment of the specific experimental results in order to finally provide useful diagnostic information (33, 40). It is accepted as good practice to perform several quality checks at the different steps in the process of analyzing clinical samples. Several authors reviewed different quality measures and how to use them during the downstream analysis. A recent review by Li et al. exposed many sequencing quality checks specific for RNA-Seq experiments including checks assessing raw sequence quality, nucleotide composition, presence of rRNA or tRNA, and the presence of other contaminant nucleic acids (41). Another important step is the alignment of the sequenced reads to the reference genome, or transcriptome. The human genome is nowadays quite complete with the latest version 38 released on June 29, 2014, by the Genome Reference Consortium, patch 4 (GRCh38.p4) (www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human). In the alignment, the human genome is used as matching reference for the sequenced reads. RNA-Seq data alignments differ substantially from the DNA-Seq alignments. The nature of read sequences in RNA-Seq provides extra levels of complexity due to the fact that RNA molecules are the product of transcription and posttranscriptional processes such as splicing and RNA editing. The splicing removes part of the transcribed sequences (the introns) leaving the exons

present in the sequence. After the library preparation and its fragmentation step, which is an optional step and commonly performed by sonication, some of the shorter reads obtained may come from the region where two exons were joined. In this particular situation, the RNA-Seq aligners have to be flexible enough to be able to map part of the reads to one exon 58 Integration of Omics Approaches and Systems Biology for Clinical Applications and the other part to another exon, spanning an exon junction (42). There are many aligners available that can deal with RNA-Seq data, such as Bowtie2, GSNAP, STAR, and SpliceMap, among many others. Work has been done to review and report available alignment tools to help users through the, sometimes difficult, decision of selecting the best tools for applications in clinical diagnostics (34, 43). In general, all aligners offer the possibility to modify key parameters in order to adapt their algorithms according to the quality of available data and the question of relevance. Once a decent quality alignment is produced, the proper diagnosis is usually within reach. A common approach is to retrieve transcript abundance, as gene counts, for gene expression profiles or differential expression. However, prior to comparing two RNA-Seq datasets, the raw counts should be normalized to account for some differences introduced by handling during the library preparation steps. Due to this inherent variability, normalization of raw counts is required since these are not directly comparable between or within samples (44). There are many normalization methods, some correcting for gene length, GC content, and library size, as well as other bias adjustments. For better understanding of the available normalization procedures, Dillies et al. compared several normalization methods in order to clearly present their application in the context of RNA-Seq data. In summary, the available DESeq and TMM normalization methods showed to be able to maintain the power to detect differentially expressed genes while properly controlling the false positive rate (44). Another way of normalization to deal with extra biases found in cross-platform or interlaboratory comparisons relies on the inclusion of synthetic spike-in materials. In some cases these external RNA controls developed by the External RNA Controls Consortium (ERCC) became available for the evaluation of cross-platform performance according to GC content, transcript length, and sequencing accuracy (45). Extended information on RNA-Seq practices as well as some additional recommendations, benchmarking technology comparisons, reproducibility assessments, and evaluations of RNA-Seq for clinical applications was also published by the Sequencing Experiment Quality Control (SEQC) consortium. The SEQC project

is the third phase of the MAQC, and it involves 12 countries, 78 organizations, and 180 researchers (http://www.fda. gov/ScienceResearch/BioinformaticsTools). The wide range of available bioinformatics tools offers the possibility to answer various biological and diagnostic questions. However, bioinformatics analysis may not be able to overcome some limitations that we can still face with NGS data such as highly repetitive sequences, 3′ biases, and biased GC content. In general, the small loss of information due to these limitations is of low impact compared with the significant insights that NGS provides. Repetitive sequences in the human genome are well characterized, making it easier to handle problems related to polymorphic copy number variation in these regions. During the alignment steps, reads that map to many locations of the genome (not uniquely mapped) with equal quality are usually filtered. The enrichment of 3′ end sequences of genes, also known as 3′ bias, is a side effect of the fast degradation of mRNAs from the 5′ end of the transcript, which may be even more prominent when using poly-A enrichment methods during the library preparation. This effect can be widely avoided by using higher-quality RNA, which should be possible in a properly designed diagnostic setting. Additionally, 3′ biases may not affect the outcome of some analysis, such as gene expression measurement, since it is considered that all transcripts exhibit similar degradation and the same library preparation was performed within a particular well-controlled experiment. The last limitation, regarding some difficulties of sequencing high GC regions, is a problem that usually results from several causes. First, it is known that some polymerases may have increased difficulties to transcribe high GC content sequences. This, coupled with the inherent high repetitive nature of GC or AT enriched regions, makes these regions somehow tricky to analyze with higher levels of confidence. However, not all high GC are affected at the same level due to differences in GC percentages and other nucleic acid composition (41). Hansen et al. worked on an alternative normalization method to acquaint for the GC content as well as gene length of a particular gene using a conditional quartile normalization (46). However, their method did not outperform other less sophisticated normalization methods (44). Cancer is commonly regarded as an accumulation of genetic alterations such as single nucleotide variants (SNVs), altered DNA methylation patterns, and chromosomal abnormalities. As a consequence of DNA modifications, there may be dysfunctional genes leading to over- or underactivity and chimeric transcripts or gene fusions. These alterations may disrupt the proper function of the gene, which may become an oncogene, a malfunctioning tumor suppressor, or an

incorrect DNA repair gene. The occurrence of one or more of these genetic alterations may affect cellular growth and lead to tumor development. Since the landscape of cancer transcriptome is complex, RNA-Seq can be very useful for clinical diagnostic applications, offering a wider range of screening possibilities to check for the whole diversity of cancer-related alterations in a single run (40). Many studies have been carried out that contributed in the understanding of molecular determinants of tumor cell types. Cancer characterization is remarkably one of the research fields that has dedicated considerable efforts to The Use of Transcriptomics in Clinical Applications 59 adopt RNA-Seq for research purposes and to assess its potential in clinical applications (33, 46, 47). Since the accumulation of genetic alterations may be either inherited or somatically acquired, RNA-Seq becomes a strong complementary approach in screening and diagnostic applications.

2.2 Proteomics

One significant advantage of proteomics analysis is the capability to assess protein abundance. Since tissue is a site of disease initiation and progression, comparative analysis of the protein abundance between different physiological states provides a global “snapshot” on disease-associated changes. A main distinction in proteomics is the relative vs absolute quantification. In the relative quantification, thousands of proteins are identified and quantified across samples belonging to at least two experimental conditions, a set-up that enables a *relativistic* comparison of the protein levels. In the absolute quantification, the mass spectrometer is initially calibrated with a peptide of known concentration, allowing for an *absolute* quantification of the same peptide in other experimental samples. Absolute quiantification offers an accurate estimate of a peptide’s or a protein’s concentration levels but it suffers from low scaling capabilities, as the analysis is limited to 1 peptide/protein at a time. Instead, relative quantification, although not being able to accurately measure the real concentation, it offers information on the direction and the intensity of alterations happening at the abundance levels of thousands of proteins, between conditions. In this, the traditional quantification strategy included a separation of peptides using 2DE and application of dyes or fluorophores. Nowadays, peptide separation (prior to quantification) is conducted with liquid chromatography, while two main quantification strategies have been distinguished including label-based and label-free approaches (48). The former

method relies on introduction of isotope labels. Depending on the strategy of incorporation of the labels, several type of label-based quantification approaches were developed such as metabolic labeling (stable isotope labeling with amino acids in cell culture (SILAC), 15N), chemical labeling [isobaric tag for relative and absolute quantitation (iTRAQ), isotope-coded protein labeling (ICPL), isotope-coded affinity tag (iCAT), tandem mass tags (TMT)], or proteolytic labeling (18O) (49). Both chemical and proteolytic labeling have been utilized to quantify tissue proteomes, with the former being most commonly applicable. Even though metabolic labeling is typically limited to the analysis of cell line models, due to recent developments, SILAC can be also applied for the analysis of tumor tissue proteomes (called super-SILAC). Super-SILAC uses as a reference/ internal standard a mixture of different cancer cell lines labeled with SILAC reagent, which is added to tissue extracts in a fixed ratio (50). Additionally, a protocol combining super-SILAC with FACS sorting or LCM was developed for quantification of protein changes in cancer cell subpopulations derived from liquid and solid tumors, respectively. This method allows for identification of up to 8000 proteins from patient-derived samples using hybrid quadrupole-Orbitrap MS (51). An overview on recent developments and application of super-SILAC is provided by Shenoy et al. (52). On the contrary, label-free approach is easier to use, as it does not require additional labeling steps. Additionally, in the label-free approach there is no limitation with regard to the number of analyzed samples in comparison with label-based methods. However, each sample has to be analyzed individually, which may increase MS instrument use and variability. The accuracy and linearity of the label-free quantification can be affected particularly by the presence of other compounds in the samples, causing suppression effect. Irreproducibility in sample preparation is also a major concern. This might be remediated to some extent using labeled internal standards (53). Two quantification methods in label-free proteomics are spectral counting and intensity-based quantification (53). The first method relies on counting the number of MS/MS spectra for a specific protein. Therefore, more abundant proteins generate more abundant peptides, increasing the probability of ion selection for MS/MS analysis. However, differences in the physicochemical properties of peptides might affect detection of peptides by MS and thus may have an impact on quantification using spectral counting. These include peptide length, mass, amino acid sequence, solubility, net charge, and others. Therefore, to address this issue Lu et al. developed a novel method called absolute protein expression (APEX) measurements (54). In this method,

considering the physicochemical properties of individual peptides, probability of their detection is assessed by a supervised classification algorithm. In the intensity- based approach, the quantitation is performed at the MS1 level based on the area under the curve (AUC) from the extracted-ion chromatogram. Independent of the quantification strategies used, in an effort to accurately compare the quantification results between different samples, data normalization is required. By normalizing the data, an effect associated with differences in protein loading, ionization efficiency, carryover effect, and others can be taken into account. Up to now several normalization methods have been developed and are well described in the context of several manuscripts (55-57). Based on the aforementioned, numerous techniques are currently being applied to quantify the tissue proteome. It has been shown that both quantification methods were successfully applied either for the analysis of total tissue proteomes (58) or tissues subjected to LCM. Moreover, quantitative proteomics was used to analyze fresh-frozen as well as FFPE tissues. Comparative analysis of label-free and label-based methods has been broadly described in in vitro cultured cells (59, 60). It has been shown that both methods, enable achievement of high proteome coverage and apparently valid predictions in terms of protein differential expression (58). However, higher sequence coverage and higher number of differentially expressed proteins were demonstrated in the case of label-free approach. However, due to the limited number of analyzed samples, the risk for receiving false associations exists, indicating the need for the analysis of higher sample numbers and/or application of adjustment for multiple testing (58).

3. MOLECULAR SUBTYPING

3.1 Molecular subtyping in the era of -omics integration

Advancements in high throughput -omics technologies along with the implementation and improvisation of the available bioinformatics solutions, have together allowed the comprehensive analysis of large sample cohorts. Making use of the massive data, the emerging field of cancer subtyping aims to stratify the disease into homogeneous groups, so as, to enable the identification of novel molecular mechanisms, targets or biomarkers for response/outcome. As examples, some of the largest analyses (in terms of sample size) aiming to identify pan-cancer modules, have been conducted using over

10,000 tumor samples from The Cancer Genome Atlas (TCGA) (61-66). In these studies, the main objective is the application of machine learning in order to integrate and analyze data from different -omic sources. Such approaches offer the possibility to utilize information from different molecular levels (mutations, CNVs, promoter interactions, gene expression, protein abundance, post translational modifications). Although the last ten years were very productive in the establishment of new algorithms able to deal with the challenges that come with the big data, it appears that there is neither consensus agreement, nor a systematic comparison of the accuracy of the aforementioned tools. This is somewhat expected, since we are at the very onset of the -omics integration era.

In this section, some of the most used algorithms together with a brief description of their mathematical basis is provided. Such models are utilized for assessing the structure of the data (i.e. anomaly and batch effects detection), for extracting a feature subspace (i.e. feature selection), for allocating correlation patterns among datasets, for detecting variables with homogeneous characteristics (i.e. clustering) and for predicting continued or discrete values of new observations (i.e. regression, classification). Stratifying the available tools into categories according to the task or the mathematics they use is not a trivial task (67, 68). This is due to the fact that algorithms often use mixed methodologies and also, the established methodologies are constantly being improvised to deal with different biological questions. In a more general way, algorithms can be divided into unsupervised and supervised categories. The former seek to identify homogeneous data structures while the latter utilize the identified structures to predict labels (in classification) or values (in regression) for new variables. However, because this distinction overlooks common mathematical aspects among the different approaches, based on their overall algorithmic rational, here algorithms are stratified into the following six categories: Non-Negative Matrix Factorization, Non-Parametric Mixture models, Pathway based, Network based, Kernel function and Multi-step models.

2.1.1 Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) comprises a group of dimensionality reduction algorithms that are used to deliver a subset of highly correlated variables

among different datasets. They exploit the property of matrix multiplication: any non-negative matrix M can be the product of two significantly smaller non-negative factor matrices W and A

$$M = WA \qquad (1)$$

With M being the input dataset of dimensions m x p, W a latent subspace m x n and A the component matrix n x p, the reconstruction of M is based on finding the subset of features that their linear combinations in W weighted by the components in A could best approximate M. This is in fact controlled by minimizing the error function F:

$$\min_{W,A} = \|M-WA\|_F, \quad W \geq 0, \quad A \geq 0 \qquad (2)$$

Following the implementation of matrix factorization in -omics data integration (69, 70), NMF algorithms are now widely used in cancer subtyping and biomarker discovery (71-74) and also several other variants of NMF have been proposed. Yang and Michailidis introduced iNMF (75) were a penalty is applied at the latent matrix W to control for variance across the different datasets, while Lock and colleagues developed the Joint and Individual Variation Explained (JIVE) method (76, 77). In the latter, each -omics data-type M ($M1$, $M2$, … Mx) is decomposed into three portions: a low-rank pair of factor matrices W^S and A^S representing the approximation of the shared (joint) subspace among -omics datasets, a low-rank pair of factor matrices W^I and A^I representing the approximation of the individual variation, and residual noise E (3). The identification of a subset of well correlated joint variables is conducted in a permutative framework, inducing L1-sparsity while examining the output joint structure. This algorithm enables the integration of any type of expression data.

$$M = W^S A^S + W^I A^I + E \qquad (3)$$

Inspired by the splitting of the integrated data into shared and individual approximations, other solutions to the error minimization problem (3) adopt a similar rationale (78-80).

In order to infer cluster structure, Non-Negative Matrix Factorization can be combined with probabilistic models, such as the parametric Bayesian approach to the mixture models. Mixture models are utilized to describe the frequency/density of realizations (samples forming sub-populations) out of an overall population. They do not require learning of class labels but instead they assume that the data follow a statistical distribution with known properties. When combined to factorial models (1)(3), they draw a *prior* probability form a Bayesian distribution and calculate the *posterior* probability of a sample participating in a given class, loosely for the joint and individual variances across different data types (3). The Joint Bayes Factor model (81) instead of distinguishing between shared (A^S) and individual (A^I) component factors, utilizes a common component matrix A for all data-types which is further subjected to regularization using a beta-Bernoulli process (82, 83). The joint Bayes Factor algorithm has been used in correlating CNVs to gene expression, and together with all the aforementioned NMF models, they assume similar variable distributions (74) across the different data-types. In addition, NMF models require normalized data, non-negative values and exclusion of extreme observations. Unlike the aforementioned algorithm, iCluster and its update iClusterplus follow a joint matrix factorization based clustering approach that allows for negative values (84). The algorithm was designed for simultaneous clustering of various data-types including somatic mutations, CNVs and gene expression. iCluster builds a Gaussian model using sets of correlated latent joint-variables across data types and then performs k-means clustering on the factor scores. To ensure non-compromised selection of latent variables among the different data-types, a data-type specific L1-penalty is performed at the component matrix.

2.1.2 Non-Parametric Mixture models

Other more flexible clustering algorithms use the Dirichlet multinomial distribution to identify discrete hierarchical structure between randomly assigned data points (realizations). They exploit a non-parametric approach to mixture models (Dirichlet process), that induces a *prior* distribution (H) over partitions of the data, a distribution that is readily combined with a concentration parameter (α), controlling for local density around realizations. It does not require knowledge on the number of mixture components, and can be applied without factorial models. The Dirichlet process (**Figure 5**) assumes dependencies among different data-types, in a way that clustering

of a given -omics dataset have an impact on the cluster structure of the other(s). This category includes the algorithms Mulitple Dataset Integration (MDI) (85), Patient Specific Data Fusion (PSDF) (86), Transcriptional Modules Discovery (TMD) (87), and Clusternomics (88). MDI can take as input multiple types and numbers of omics datasets (even ChIP and protein-protein interactions dataset in a binary form), TMD takes as input two datasets, a ChIP and a gene expression dataset to discover transcriptional modules, PSDF also takes two datasets (ChIP or CNV and a gene expression dataset) to infer cancer subtypes, while Clusternomics can handle multiple types of datasets, including DNA methylation, gene expression and proteomics.

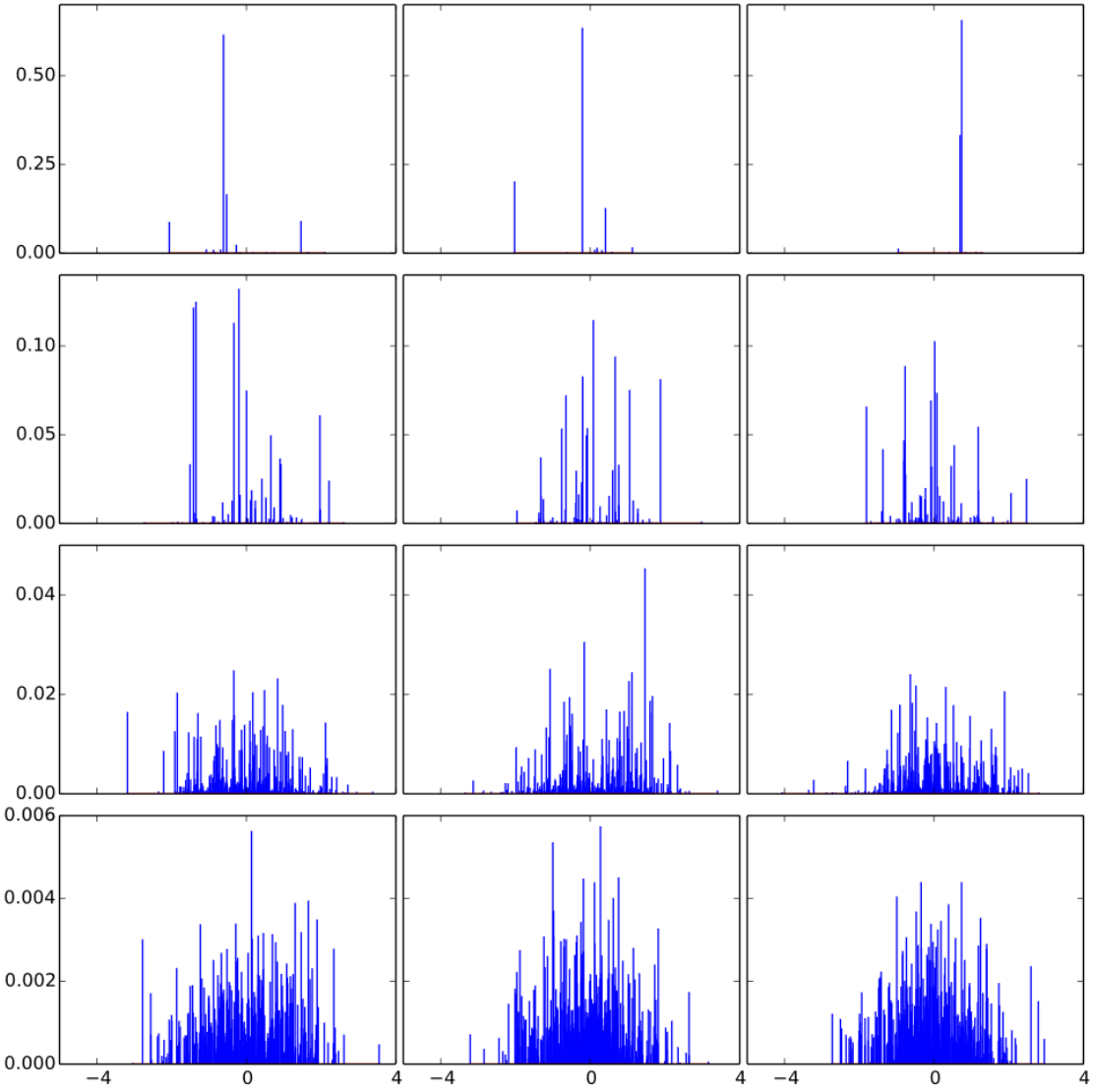


Figure 5: ***The Dirichlet process***. *The four rows use different concentration parameter α (top to bottom: 1, 10, 100, 1000) with each row illustrating 3 repetitions of the same experiment. For lower α, draws from the Dirichlet distribution (realizations) tend to be concentrated at a single value, while for higher α they become continuous. Clustered*

structure can be identified for intermediate values of α. For a given value of α, each realization is characterized by a posterior probability of participating in a cluster (*source:https://en.wikipedia.org/wiki/Dirichlet_process#/media/File:Dirichlet_process_draws.svg)*

2.1.3 Pathway based

Pathway based algorithms make use of libraries or databases with molecular pathways or Gene Ontology terms and assess cluster structure as a function of pathway-similarity between samples (89). Among the most popular is the PAthway Representation and Analysis by DIrect Reference on Graphical Models (PARADIGM) algorithm (89), which uses interactions between pathway entities from the National Cancer Institute (NCI) Pathway Interaction Database (PID) (90) to infer patient specific pathway activations in a probabilistic framework. For each patient and for a given pathway, based on CNV and gene expression datasets it calculates probability distribution over subsets of entities aiming to define a joint probability distribution of the pathway-factor graph. PARADIGM uses expectation-maximization (EM) to learn the parameters of the observation factors for each pathway and after averaging them, it calculates the *posterior* probability of pathway activation for each sample, individually. The results are summarized in the inferred pathway activation (IPA) matrix with values ranging from -1 (deactivated) to 1 (activated) which can be further submitted for clustering to derive groups of patients with homogeneous pathway activations. However, PARADIGM does not account for overlapping genes between pathways as it calculates pathway indexes individually. Integrative Genomics Robust iDentification of cancer subgroups (InGRiD) is a pathway based algorithm that was designed to deal with the issue of overlapping genes among pathways (91). It works in a semi-supervised manner, as it requires patient outcome data, and features two approaches for dealing with the overlapping issue. The algorithm uses Cox regression models at each step to assess the importance of each gene within each pathway and outputs pathway risk scores, which are further subjected to clustering yielding patient subgroups. A major limitation is that it currently works only with gene expression data, but the authors are planning to extend it in order to include multiple data-types

2.1.4 Network based

Several algorithms have been designed to integrate and cluster multiple omics datasets by exploiting graph and network properties. The Similarity Network Fusion (SNF) is a non-bayesian network-based tool used for integrative clustering (92). For each -omics data-type, SNF calculates correlation between samples and constructs a network of patients, with connections denoting the strenght of correlation. The algorithm fuses all data-type specific networks using k-nearest neighbour and graph diffusion, in an iterative process. The result is a global fused similarity matrix adjusted for local data-type effects, presenting sample class membership. SNF can take any type of omics dataset and it is among the most popular solutions. Similar to SNF, Affinity Network Fusion (ANF) constructs networks of patients but it uses a non-linear transformation of k-nearest neighbor graph together with a Gaussian kernel based network to infer similarity matrices (93). The latter are fused into an affinity matrix which is subjected to spectral clustering, while clustering performance as well as the optimal number of clusters is determined with an eigengap heuristic approach. The algorithm provides a semi-supervised classifier for predicting outcome by constructing a neural network model that can be fitted with the ANF output in a training/test set fashion. It can deal with gene expression (coding and miRNA data) and methylation datasets, but due to its increased complexity the training of the classifier may require case-specific optimization. COpy Number and EXpression In Cancer (CONEXIC) is a bayesian variation that aims to identify modules of concordantly deregulated CNVs and gene expression data (94). It uses net graph clustering approach to define clustering centroids and offers functional characterization of the clusters.

A different approach in data integration is conducted with Analysis Tool for Heritable and Environmental Network Associations (ATHENA) (95). It is a neural network that combines several omics data-types, but unlike the aforementioned network approaches it works in a predictive/supervised fashion. ATHENA combines selection of features associated with outcome with grammatical evolution neural networks (GENN) (96) to train individual classifiers from different data-types. The results are then summed up to an integrative model capable of predicting disease prognosis for the samples. Apart from its prognosticating features, the algorithm provides insights on the correlations between different data-types at a whole genomic scale, but a limitation is that it does not assume dependancies between data types. In the same context of supervised network learning, Mutual Information-based integrative Network Analysis (MINA)

integrates CNV, methylation and gene expression data to construct a network of gene-gene interactions and test the effect of each possible gene pair interaction on the outcome (97). It is an exploratory non-parametric approach aiming to discover gene regulation modules that could significantly affect outcome. MINA discretizes continuous values to infer a probabilistic model of interactions, where permutation at the outcome status of the patients is employed to distinguish between random and non-random gene-gene effects.

2.1.5 Kernel function

Data integration tasks can be conducted by implementing the powerful kernel function as well. These methods calculate the inner product of two data vectors into a higher dimensional space. Assuming two data vectors *x* and *y* that are located at some feature space $\mathbb{R}^m$ and can be mapped to another feature space $\mathbb{R}^n$ through φ (4), the kernel function *k* calculates the dot product of the vectors (*x*, *y*) in the projected feature space $\mathbb{R}^n$ (5).

$$\varphi : \mathbb{R}^m \rightarrow \mathbb{R}^n \quad (4)$$

$$k(x, y) = \varphi(x)\varphi(y) \quad (5)$$

The output of the kernel function can be seen as a similarity metric between pairs of samples. As such, in the data integration framework, they are utilized for selection of features that could optimally drive clustering and subtype identification either in a supervised or unsupervised way. Multiple Kernel Learning – Locality Preserving Projection (MKL-LPP) is an unsupervised method of feature selection where samples from different data-types are integrated in a concatenated feature subspace (98). In this, data-type specific multiple kernels are imputed and adjusted for optimal weights iteratively, using a regularization penalty to constrain overfitting. An optimal kernel is inferred for each data-type and all kernels are then fused into an integrative model. During feature selection, to maintain distances between samples defined with k-nearest neighbors, the algorithm uses the Locality Preserving Projection (LPP) method (99) and clustering of the integrated reduced subspace is conducted with k-means. MKL-LPP can integrate gene expression and methylation data. Multiple Kernel Density Clustering algorithm for Incomplete datasets (MKDCI) is a similar unsupervised method, developed to deal with the integrative clustering of incomplete datasets (100).

It incorporates the optimally trained kernel function into the t-Distributed Stochastic Neighbour Embedding (t-SNE), which is a method for creating a two-dimensional map of samples allowing thousands of features (101). MKDCI then, models the optimal sample locations in the feature space and clustering centroids are determined after implementing correction for outliers using Isolation Forests (102). The final output is characterized by a cluster number/clustering quality trade-off. MKDCI offers the possibility of integrating multiple categorical with multiple continuous data. In contrast to the above, Feature Selection Multiple Kernel Learning (FSMKL) is a supervised multiple kernel method for data integration (103). Feature selection is performed per data-type either with respect to a statistical criterion related with outcome of class labels, or based on pathway participation status. Taking into account CNV, gene expression, subsets of genes participating in a given pathway and various clinical data (such as subtype membership) the algorithm trains multiple classifiers and assess a confidence score for each sample. The classifier with the optimal decision function is then employed to predict survival outcome only for the high confidence samples, maximizing its accuracy efficiency.

2.1.6 Multi-step models

Finally, there are algorithms that do not fall into any of the aforementioned three categories. Those are typically characterized by multi-step statistical procedures. A recently developed algorithm, Similarity Regression Fusion (SRF) makes use of correlations between pairs of samples (104). For each data-type, SRF generates a data-type specific similarity matrix of Pearson correlation scores and subjects them to Fisher transformation (105). For each data-type specific similarity matrix the algorithm then calculates the corrected similarity scores, based on all the other data-types by training a generalized linear regression model in which parameters are learned with the maximum likelihood estimation (MLE) method. The regression model integrates the corrected similarity matrices and subtype membership is inferred by spectral clustering. SRF integrates gene expression with DNA methylation data and claims to provide comparable results to iCluster, SNF and ANF. CNAmet is another multi-step algorithm that integrates methylation, CNV and gene expression data to identify co-regulated modules (106). For a given gene, by assuming that hypomethylation and copy number gains result in over- and underexpression respectively, the algorithm conducts

integration in three steps: it first calculates the signal-to-noise ratio for the CNV and methylation scores relativistic to the corresponding gene expression, then, assigns weights to each gene according to the overall degree of its aberration and at the end infers statistical significance by permuting the weighted scores. Its flexibility and ease of use have made it a popular solution in the identification of co-regulated patterns among CNV, methylation and gene expression. Other multi-step algorithms in -omics integration include the In-Trans Process Associated and Cis-Correlated (iPAC) (107), the Multiple Concerted Disruption (MCD) (108), and the Anduril (109).

2.2 Molecular subtypes of bladder cancer

Efforts to subtype and stratify disease heterogeneity according to molecular profiles have resulted in the establishment of multiple classification themes and also at the realization that BC constitutes a wide family of cancers with highly flexible molecular backgrounds. Five classification schemes have been described (including mostly MIBC patients), based mainly on DNA/RNA analysis of respective patient cohorts (110-114). At the highest level, MIBC subtypes may be of luminal or basal phenotype, with the former enjoying better prognosis than the latter. However, due to the significant intrinsic heterogeneity, further divisions of the two general phenotypes have proven to reflect differences in molecular backgrounds and outcomes. For example, in the updated BC TCGA cohort (114), by using a Bayesian Non-Negative Matrix Factorization for the CNV and mutation datasets and also the Cluster Of Cluster Assignment (COCA) method (115, 116) for gene expression datasets (mRNA, miRNA and lncRNA), investigators identified three luminal, a basal and a rare neuronal subtype, with the latter two suffering worse prognosis. By combining gene expression data of all available classification systems, a recent BC meta-subtyping study suggests the existence of six major molecular phenotypes (BOLD subtypes; **Figure 6**) (117).

Molecular subtypes of the NMIBC are less studied. Classification of NMIBC based on RNA-sequencing (118) supports the existence of three subclasses, including a luminal/differentiated, a basal-like, and a luminal/CIS-like subtype, the latter associated with worse outcome in comparison to the two former (UROMOL study). The progressed subtype overexpressed mRNAs of the late cell-cycle, transcriptional activators of the epithelial-to-mesenchymal-transition (EMT), cancer stem cell markers and was enriched in the Gene Ontology biological process of immune infiltration and

vasculature development. Based on their mutation status, 52% of samples in this subtype had alterations in the DNA damage response (DDR), 35% in the MAPK/ERK, and 20% in the ERBB pathways. A second study by Hurst et al. (119) involved the analysis of low stage (Ta) and grade (1-2) tumors for Copy Number Variations (CNVs), predicting two subclasses differing in mTORC1 signaling, DDR, glycolysis, unfolded protein response (UPR) and cholesterol biosynthesis. Even though some common elements have been identified across previous studies (112-114, 117, 118, 120-123), e.g. associations of basal, luminal and neuroendocrine features to MIBC outcome, the diversity of published classification schemes reflects the disease complexity and indicates gaps in our understanding of disease biology.

With proteins being directly linked to phenotypes, protein-based molecular subtyping holds a promise to provide critical information on translating genome signals to cell function. Apart from a small set of 208 proteins analyzed by reverse phase-protein arrays (113, 114) and a proteomic analysis of a relative small set of MIBC (124), a comprehensive proteomics profiling of BC is largely missing.

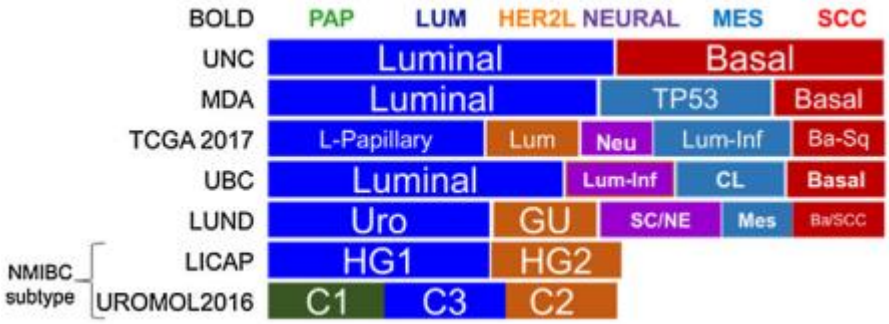


***Figure 6**: **Molecular subtypes of Bladder Cancer based on re-clustering analysis of all the published classification systems**. The scheme depicts the inter-relationship between the BOLD (metasubtypes) and published molecular subtypes. Color code: Purple = NEURAL; dark blue = LUM; green = PAP; orange = HER2L; red = SCC; light blue = MES. Ba/SCC = basal/squamous-cell carcinoma-like; Ba-Sq = basal-squamous; BOLD = bladder carcinoma subtypes of large meta-cohort database; CL,= claudinlow; diff. = differentiation; ECM = extracellular matrix; GU = genomic unstable; HER2L, HER2-like; LICAP = Leeds Institute of Cancer and Pathology; Lpapillary = luminal-papillary; Lum = luminal; LUM = luminal-like; Lum-inf = luminal infiltrated; LUND = Lund University; MDA = MD Anderson Cancer Center; Mes, mesenchymal; MES = mesenchymal-like; MIBC = muscle-invasive bladder*

carcinoma; Neu = neuronal; NEURAL = neural-like; NMIBC = nonmuscle-invasive bladder cancer; PAP = papillary-like; SC/NE = small cell/neuroendocrine; TCGA = The Cancer Genome Atlas Network; UBC = University of British Columbia, UNC = University of North Carolina; Uro = urobasal.(source: ref (117))

3. AIM OF THE STUDY

The aim of the thesis is to investigate if patients with Bladder Cancer can be stratified to biologically meaningful molecular subtypes based on a proteomics and transcriptomics analysis. These subtypes could offer novel diagnostic and therapeutic tools for improving patient management.

4. Chapter I: Proteomics

We previously had compiled a cohort of 117 primary Bladder Cancer patients in order to investigate the existence of proteomic subtypes in the Non-Muscle Invasive disease. The aim of this

4.1 Materials and methods

4.1.1 Patient samples

Fresh frozen bladder tissue specimens were collected from patients during transurethral resection of bladder tumor (TURBT), prior to any kind of treatment (chemotherapy, BCG, radiation) at the medical center Gennimatas General Hospital, in Athens (Greece). The study complied with the principles outlined in the Declaration of Helsinki and was approved by the respective local ethics committee (Gennimatas General Hospital, protocol number 4354 (18-2-2015)). All individuals gave written informed consent. Sections of cancer tissue corresponding to at least 95% tumor area, from a total of 121 BC patients were excised and prepared for the analysis. Four samples were excluded due to low protein concentration. Of the remaining 117 samples, 98 were NMIBC (Ta: n = 58, T1: n = 40; **Table 1**) and 19 were MIBC, all fully analyzed with proteomics (flowchart in **Figure 7**). Tumor stage was determined based on the TNM classification (125) and grading according to the World Health Organization (WHO) Grading System 1973 (17). Following harvesting, bladder tissue specimens were stored at -80°C until preparation for the proteomic analysis.

4.1.2 LC-MS sample preparation

Approximately 30-50 mg of BC tissue was homogenized in FASP lysis buffer using the bullet blender homogenizer (Next Advance, NY, USA). One scoop of stainless steel beads (0.9-2 mm diameter) was added to each sample and then samples were inserted into the homogenizer. The following homogenization settings were utilized: speed: 12; time: 5 min. One more homogenization step was performed reducing the speed to 10 and the time to 3 min. Samples were centrifuged at 16,000g for 10 min at room temperature and the supernatants were kept in clean tubes. Protein concentration was determined by Bradford assay. Protease inhibitors (Roche, Basel, Switzerland) were added at a final concentration of 3.6%. Protein extracts (200μg/sample) were processed using filter aided sample preparation (FASP) as described previously (126), with some minor modifications (127). Briefly, buffer exchange was performed in Amicon Ultra

Centrifugal filter devices (0.5 mL, 30 kDa MWCO; Merck) at 16,000 rcf for 15 min at room temperature. The protein extract was mixed with urea buffer (8M urea in 0.1M Tris-HCl pH 8.5) and centrifuged. The concentrate was diluted with urea buffer and centrifugation was repeated. Alkylation of proteins was performed with 0.05M iodoacetamide in urea buffer for 20 min in the dark followed by a centrifugation at 16,000 rcf for 10 min at room temperature. Additional series of washes were conducted with urea buffer (2 times) and ammonium bicarbonate buffer (50 mM NH4HCO3 pH 8.5, 2 times). Tryptic digestion was performed overnight at room temperature in the dark, using a trypsin to protein ratio of 1:100. Peptides were eluted by centrifugation, lyophilized, and stored at –80°C until further use.

4.1.3 LC-MS/MS quantification

Samples were injected into a Dionex Ultimate 3000 RSLS nano flow system (Dionex, Camberly, UK) configured with a Dionex 0.1 × 20 mm 5 μm C18 nano trap column. Mobile phase was 2% ACN: 0.1% FA with a flow rate of 5 μL / min. The analytical column was an Acclaim PepMap C18 nano column 75 μm × 50 cm, 2 μm 100 Å at a flow rate of 300 nL / min. The trap and nano-flow column were mainitained at 35°C. Samples were eluted with a gradient starting at 1% B for 5 min rising to 5% B at 10 min then to 25% B at 360 min and 65% B at 480 min. Mobile phase A constituted of 0.1% formic acid while mobile phase B of 80% CAN and 0.1% formic acid. The column was washed and re-equilibrated prior to each sample injection. The eluent was ionized using a Proxeon nano spray ESI source operating in positive ion mode. For mass spectrometry analysis, an Orbitrap LTQ Velos (Thermo Finnigan, Bremen, Germany) was operated in MS/MS mode, scanning from 380 to 2,000 m/z. Ionization voltage was 2.6 kV and the capillary temperature was 275 °C. The resolution of ions in MS1 was 60,000 and 7500 for higher-energy collisional dissociation (HCD) MS2. The top 20 multiply charged ions were selected from each scan for MS/MS analysis using HCD at 35% collision energy. Dynamic exclusion was enabled with a repeat count of 1, exclusion duration of 30 s.

4.1.4 Data processing and clustering analysis

Raw files were analyzed with Proteome Discoverer 1.4 software package (Thermo Finnigan), utilizing the Sequest search engine and the Uniprot human (Homo sapiens) reviewed database, downloaded on May 30, 2016. The search was performed using carbamidomethylation of cysteine as static and oxidation of methionine as dynamic modifications. Two missed cleavage sites, a precursor mass tolerance of 10 ppm and fragment mass tolerance of 0.05 Da were allowed. False discovery rate (FDR) was set to 0.01. The retrieved protein area files for each sample were merged by an in-house script in the R environment for statistical computing and graphs (version 3.4.4), according to their invasion status in the NMIBC and MIBC datasets, consisting of 1,309 and 1,515 protein entries, respectively. To investigate tissue intrinsic subtypes, abundant plasma proteins (n = 177) were excluded from the analysis. The two datasets were submitted for column (sample) normalization according to (6) yielding the processed protein matrices.

$$X' = \frac{X}{sum(Xi)} * 10^6 \qquad (6)$$

Consensus clustering of the processed NMIBC dataset was performed as described in Wilkerson et al. (128). To control for over-filtering while maintaining a maximum number of features, multiple subsets of the NMIBC dataset, differing at their frequency threshold (no threshold (0%), minimum frequency of 10%, 20%, 35%, 50%, 60%), were subjected to k-means clustering, forcing 95% random sample resampling across 1000 iterations. In every run, the curated protein intensities were median scaled and submitted for agglomerative hierarchical clustering. K-means clustering was performed using Pearson correlation as the distance metric, whereas neither weights nor other feature selection steps were applied prior to clustering. Outputs were inspected and compared against each other for the reproducibility of the classification across the different frequency thresholds. For k = 2, 3 and 4 cluster solutions, comparative analysis of the class assignments between the different frequency thresholds yielded a total of 1, 2 and 10 class switches respectively, all of which being allocated between 0%-10% and 10%-20% frequency thresholds, with 100% reproducibility of class assignments between the 20%, 35%, 50%, and 60% frequency threshold runs. Therefore, to maximize the proteome coverage, a 20% protein frequency threshold was selected for

the analysis and applied both to the NMIBC and MIBC datasets. Evaluation of the best k- clustering solution was conducted based on cluster size, on examinations of the Cumulative Distribution Function (CDF), delta Area Under Curve (AUC) (**Supplementary Figure 1**), and tracking plots as described in Wilkerson et al (128).

4.1.5 Molecular themes, features and signatures

Proteins investigated in this study (**Figure 8**) included basal and cancer stem cell markers (*CD44, CD47, ALDH1A1, MSN, MUC1, RPSA, COL18A1, TGM2, BAX*) (112, 114, 118, 129), previously investigated cell adhesion molecules (*FN1, VTN, LAMC1, LAMB2, LAMA4, LAMB1, CDH1, ITGA6, ITGB4, CTNNA1, CTNNB1, JUP, CTNND1*) (112, 129), cytokeratins (*KRT14, KRT6A, KRT16, KRT5, KRT7, KRT17, KRT8, KRT18, KRT20*) (112, 114, 118), markers of differentiation (*UPK2, UPK3BL1, UPK1B, GPX2, PDCD4, SRC, ADIRF, FBP1, FABP4*) (114, 118, 129, 130), a set of proteins potentially involved in EMT as curated from the Molecular Signatures Database (http://software.broadinstitute.org) (*VIM, COL1A1, COL1A2, TGFBI, CAV1, NID1, POSTN, FLNA*), proteins of the stromal compartment (*ACTC1, CNN1, MFAP4, ACTA2, DES, MYH11, MYL9, TAGLN, COL6A3, COL14A1*) (114, 131, 132), proteoglycans of the extracellular matrix (*HSPG2,DCN, LUM, BGN, OGN, VCAN, PRELP, SDC1*), cell-cycle progression molecules (*NASP, RCC2, CDC37, YWHAG, PAICS, NME2, GART, CDC42, BUB3*), markers of inflammation with functions varying from transcriptional activators to cytokine signal transduction and angiogenesis (*STAT1, STAT3, SND1, DHX9, HMGB1, HMGB2, HMGB3, PTGES3, RNF213, TYMP*), an antigen presentation signature (*HLA-A, SEC24C, CD74, TAP1, TAPBP, PSMB9*) (112), features of the DNA damage response (*RUVBL2, PCNA, PRKDC, PARP1, TOP2B, APEX1*) (114) and the unfolded protein response (*HSP90B1, HYOU1, SEC61A1, SRPRB, SSR1, HSP90AA1, TLN1*), enzymes of the glucolysis/glucogenolysis (*PYGL, PYGB, GALE, GNPDA1, PGAM1, IDH1, TPI1, PGK1, TXN, FBP2, GMPPA, TSTA3, GOT1*) as well as features from two CIS vs papillary gene signatures (upregulated in CIS: *S100A8, LYZ, CLIC4, RARRES1, AKR1B10, DPYSL2, TUBB*; downregulated in CIS: *TRIM29, IVL, ANXA10, BCAM, CTSE, LAD1*) (114, 133).

4.1.6 Statistical analysis of the subtypes

The non-parametric Mann-Whitney and Kruskal-Wallis tests were utilized for defining statistical significance for continuous variables. Fisher's exact and χ^2-tests were conducted for calculating significance of the categorical variables, while the likelihood ratio test was used for assessing changes in grade distribution between pairwise subtype comparisons. Visualizations of the protein and transcript abundances were constructed with the package ComplexHeatmap (v1.2) in R, and expression values shown are z-normalized (7), with μ being the mean and σ the standard deviation of each row. Principal component analysis was summoned in SPSS (version 23) and the input data included the $\log_2$ transformed intensities of the statistically significant proteins between the three classes (in **Figure 9a**, n = 626 proteins), and between MIBC and NMIBC subtypes (in **Figure 9b**, n = 618 proteins).

$$X' = \frac{X-\mu}{\sigma} \qquad (7)$$

Gene ontology and Reactome pathway analysis were conducted in the Cytoscape plug-in, ClueGO (134). Libraries were updated at May 19, 2018 and significance was defined by a two sided-hypergeometric test corrected with Benjamini – Hochberg $p < 0.05$. Enrichments for the Hallmark genesets (135) were predicted with the weighted Kolmogorov – Smirnov approach of the Gene Set Enrichment Analysis (GSEA) software (136). Signal2Noise was set as the ranking metric, and random enrichments were discarded by permuting for class labels (n = 1000 iterations). Significance was defined by FDR < 0.25 and nominal p-value < 0.05. Only proteins that reached statistical significance were used as input to GSEA.

4.1.7 Analysis for class specific pathways

For each class comparison, the statistically significant proteins were submitted to ClueGO in the form of two lists involving the up- and down-regulated proteins, respectively. The analysis predicted a total of 633 significantly deregulated Reactome pathways (BH $p < 0.05$) across the three class comparisons, which upon fusion based on presence of same parental node resulted in a final list of 186 pathways. "Class specific" pathways discriminating each class from all the rest, had to fulfill the following criteria: a) be enriched in the specific class in all pair-wise comparisons involving this class, and b) be not significant or absent in pair-wise comparisons not

involving this class. Following application of the aforementioned criteria, 68 pathways were shortlisted, while upon further omission of redundancies (e.g. combination of pathways of same involved molecules) a final list of 20 class specific pathways was generated [6 for class 1, 4 for class 2 and 10 for class 3 (**Table 2**)].

4.1.8 Validation of the proteomics classification

The validity of the proteomics classification was assessed in terms of its relation to the RNA-seq classification system of the UROMOL cohort (n = 476 samples). The three UROMOL's subtypes are cited in italics (e.g. *class 1/* Luminal; *class 2/*CIS-like; *class 3/*Basal-like), and are annotated as **P**rogressed (**P** for *class 2*) or **N**on-**P**rogressed (**NP** for *classes 1* and *3*).

Particularly, the Supplementary TableS3 of the UROMOL study that contains average FPKM intensities per *class*, statistical tests and regulation of transcripts across the three UROMOL *classes* was downloaded for the preparation of *class-specific* genesets, containing overexpressed transcripts per *class* and subsequent analysis with GSEA (**Figure 10**).

For the classification of the proteomics samples into the previously established UROMOL's classification system, the processed gene expression data of the 476 early stage UROMOL tumors (E-MTAB-4321) was download from https://www.ebi.ac.uk/arrayexpress/. First, a merged dataset (MD) containing the intersection of features and also the union of the samples between the proteomics NMIBC processed dataset at 20% threshold and UROMOL's processed data (E-MTAB-4321), was created (n = 1,275 features). Removal of batch effects between the two different -omics sources was conducted with the *ComBat* function of the R package, sva (137) (v3.20.2), and the concatenated feature space was inspected with PCA. Classifier was built in R, with the package randomForest. Training set: 476 early stage tumors from the UROMOL cohort, stratified to three class-labels. Test set: 98 early stage tumors of this cohort with protein quantifications. Out of bag error (=2.1) was minimized at *ntry* = 800 (number of trees) with an optimal of 70 and 37 features per tree for the training and test sets, respectively. Feature selection was conducted with the permutation and the Gini impurity tests, for the top 100, 200 and 300 most informative features and results (subtype assignments) were averaged (**Table 3**).

4.1.9 Post-machine learning analysis for features of prognostic potential

To evaluate the relevance of the proteomic output, previously published LC-MS/MS proteomic bladder cancer datasets (138) were screened for overlaps with the presented data. To investigate features potentially involved in tumor aggressiveness, transcriptomic data of two previous cohorts, LUND (112) and UROMOL (118) were employed (**Figure 11**). From the LUND taxonomy (112), the processed data, as deposited in https://www.ebi.ac.uk/arrayexpress/ (E-GEOD-32894), were downloaded. After selecting the NMIBC subset (n = 213 samples), these samples were further grouped based on disease progression status (Progressors *vs* Non-Progressors). For the Non-Progressors group, a follow-up history of at least 12 months was set as a requirement to confirm lack of progression. This resulted in a final set of 161 patients corresponding to Progressors (n = 17) and Non-Progressors (n = 144), and differentially expressed mRNAs were identified (n = 1,817; Mann-Whitney $p < 0.05$). From the UROMOL study (118) the Supplementary TableS3 was downloaded, which contains the statistically significant transcripts (q-adjusted < 0.05) and their regulation across the three UROMOL class comparisons (mean FPKM values and fold changes). The UROMOL's subtypes are cited in italics (e.g. *class 1, class 2, class 3*) with *class 2* having significantly worse progression free survival rates when compared to the other two (118). Pair-wise comparisons of the Progressed (P) *class 2* versus the Non-Progressed (NP) *classes* (*1* and *3*) were performed. For the visualization of the mean FPKM values of the UROMOL (P) and (NP) subtypes (**Figure 12b**), data were expressed as row percentages according to (8).

$$Xi' = \frac{Xi}{sum\ (X)} \qquad (8)$$

4.2 Proteomic subtyping results

4.2.1 Proteomics data collection and evaluation

Tissue specimens from 121 BC patients (samples received during TURBT, prior to any other treatment, as described in Methods) were processed for proteomics analysis by high resolution LC-MS/MS (**Figure 7**). Four samples were excluded due to low protein

concentration, resulting in 117 analyzed cases of which 98 presented with NMIBC (Ta: n=58, T1: n=40; **Table 1**) and 19 with MIBC. To increase reliability of the proteomic output, only high-confident peptides (FDR < 0.01) and proteins present in ≥ 20% of samples were considered. To facilitate the detection of intrinsic tumor characteristics, abundant plasma proteins were excluded from further analysis. Using these criteria, the NMIBC and MIBC datasets consisted of 1,309 and 1,515 proteins respectively (at 20% threshold). Among the most abundant features were cytokeratins (KRT7, KRT8, KRT19), actin isoforms (ACTA2, ACTB, ACTC1) and nucleosome components (HIST2H2AB, HIST1H2AH, HIST1H4A, H2AFZ), in line with previous BC proteomics datasets (138). Gene Ontology (GO) analysis of the detected proteins supported their involvement into biological processes highly relevant to BC, such as stem cell differentiation, viral infection, DNA damage response and p53 checkpoint, protein synthesis, chromatin remodeling, regulation of cell cycle, and detoxification (BH p < 0.05; **Supplementary Table 1**). These results collectively supported the biological relevance of the proteomics output, prompting further investigations.

4.2.2 Identification of three NMIBC molecular subtypes of distinct pathological phenotypes

To determine whether discrete molecular subtypes exist in the tissue proteome of NMIBC, consensus clustering (128) was performed at the respective proteomic dataset (n = 98 NMIBC patients). For a range of 2-10 possible k-solutions, highest clustering stability (as described in methods) at k = three clusters was detected (**Supplementary Figure 1**). Of note, these NMIBC classes remained largely stable when different protein frequency thresholds were tested (35%, 50%, 60%, data not shown). The three classes differed in size, with class 1 being the smallest of all (**Table 1).** Statistically significant differences were observed in stage, grade, and EORTC risk composition, with class 1 harboring mostly T1-grade 3, high risk tumors, class 3 conversely, mostly Ta-grade 1-low risk tumors and class 2 representing a more heterogeneous group (**Table 1**). Patients with squamous differentiation (n=21) tended to be classified either as class 1 (4/17) or class 2 (13/42), while out of the 39 patients classified as class 3, only 4 presented with squamous histology (**Table 1**).

	Overall N	Class 1 n (%)	Class 2 n (%)	Class 3 n (%)	Class 1 vs Class 2 p-value	Class 1 vs Class 3 p-value	Class 2 vs Class 3 p-value	Class 1vs2vs3 p-value
Age, *mean ± SD*	70.0± *12.1*	71.8 ± *10.9*	71.6 ± *10.5*	67.4 ± *14.0*	0.953[a]	0.256 [a]	0.128 [a]	0.309 [a]
Gender, *n (%)*					0.662 [b]	0.250 [b]	0.292 [c]	0.347[b]
Male	84	16 (19.0)	37 (44.0)	31 (36.9)				
Female	14	1 (7.1)	5 (35.7)	8 (57.1)				
Smoking history					0.704[b]	0.626[b]	0.891[b]	0.853[b]
Current	45	6 (13.3)	21 (46.7)	18 (40.0)				
Former	14	3 (21.4)	7 (50.0)	4 (28.6)				
Never	24	5 (20.8)	10 (41.7)	9 (37.5)				
Missing	15	3 (20.0)	4 (26.7)	8 (53.3)				
Tumor stage					***0.043*** [b]	***<0.001*** [c]	***0.018*** [c]	***0.003*** [b]
Ta	58	4 (6.9)	23 (39.7)	31 (53.4)				
T1	40	13 (32.5)	19 (47.5)	8 (20.0)				
Tumor grade (WHO 1973)					***0.030*** [d]	***<0.001*** [d]	***0.002*** [d]	***<0.001*** [b]
Grade 1	43	2 (4.6)	14 (32.6)	27 (62.8)				
Grade 2	24	3 (12.5)	12 (50.0)	9 (37.5)				
Grade 3	31	12 (38.7)	16 (51.6)	3 (9.7)				
CIS					0.733 [b]	NC	0.550 [b]	0.999 [b]
No	71	12 (16.9)	32 (45.1)	27 (38.0)				
Yes	1	0 (0.0)	1 (100.0)	0 (0.0)				
Missing	26	5 (19.2)	9 (34.6)	12 (46.2)				
Tumor size					0.095 [c]	0.109 [c]	0.994 [c]	0.263[b]
<3cm	56	7 (12.5)	27 (48.2)	22 (39.3)				
≥ 3cm	29	8 (27.6)	12 (41.4)	9 (31.0)				
Missing	13	2 (15.4)	3 (23.1)	8 (61.5)				
Tumor multiplicity					0.601 [c]	0.872 [c]	0.639 [c]	0.844[b]
No	65	12 (18.5)	26 (40.0)	27 (41.5)				
Yes	32	5 (15.6)	15 (46.9)	12 (37.5)				
Missing	1	0 (0.0)	1 (100)	(0.0)				
Squamous cell differentiation					0.753[b]	0.228[b]	***0.029***[b]	0.070[b]
No	77	13 (16.9)	29 (37.6)	35 (45.5)				
Yes	21	4 (19.0)	13 (61.9)	4 (19.0)				
EORTC risk for NMIBC					0.403 [b]	***0.001*** [b]	***0.002*** [b]	***<0.001*** [b]
High	30	9 (30.0)	17 (56.7)	4 (13.3)				
Low	68	8 (11.8)	25 (36.8)	35 (51.4)				

Table 1. Distribution of the clinical and histopathological features across the three NMIBC proteomic subtypes. *Statistical tests were conducted for all the possible comparisons. All percentages are expressed as row proportions. NC = Non-calculable,* [a]*Kruskal – Wallis H-test;* [b]*Fisher's exact test;* [c]χ^2*-test;* [d]*Likelihood ratio.*

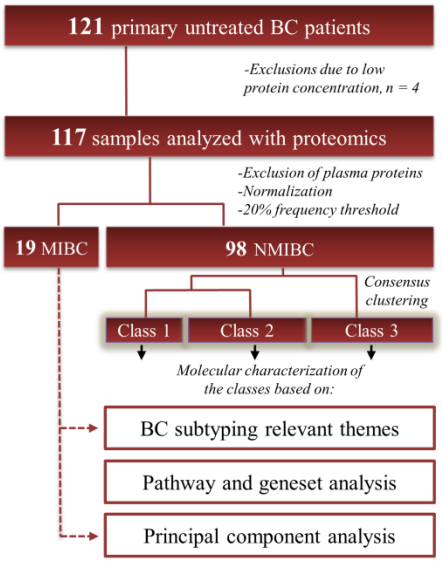


***Figure 7**: Flowchart for the classification analysis of the NMIBC patients. The proteomic MIBC dataset (n = 19) was utilized as a reference group to aid with the molecular characterization of the three NMIBC subtypes.*

4.2.3 Proteomics profiling of the NMIBC subtypes

To comprehensively delineate the proteomic phenotypes of the three classes, pairwise comparisons were initially performed. Among the differentially expressed proteins in each case (ranging from 370-520 proteins), multiple previously described BC subtype markers were found (described in details in the Methods section). As shown in **Figure 8**, some basal and cancer stem cell markers, including MSN, COL18A1, RPSA, ALDH1A1, TGM2, and BAX were of higher abundance in classes 1-2, with the basal-layer antigen CD47 being over represented in class 1. Molecules that guide cell adhesion during wound healing (FN1, VTN) and several lamina propria laminins, were found up-regulated in classes 1 and 2. In contrast, the hemidesmosomal complex ITGA6/ITGB4, which facilitates cell anchorage to the basal membrane as well as proteins mediating stratified epithelial cell-cell attachment and communication (E-cadherin, CTNNA, CTNNB, JUP, and CTNND) were at increased abundance in class 3. This class also showed increased abundance of several luminal proteins such as KRT20 and epithelial cytokeratins expressed by semi- or terminally differentiated cells (KRT8/KRT18, KRT7/KRT19) (112), uroplakins (UPK2, UPK1B, and UPK3BL1),

and other luminal proteins (GPX2, SRC, ADIRF), earlier reported to be at increased abundance in low risk tumors (114, 117, 121).

Tumors with high levels of proteins potentially involved in Epithelial-to-Mesenchymal Transition (EMT) (VIM, COL1A, COL1A2, TGFBI, CAV1), as well as markers of stromal infiltration (ACTA2, DES, TAGLN, COL6A3), and structural components of the extracellular matrix, like proteoglycans (LUM, DCN, BGN, PRELP) co-clustered in class 2, suggesting that this cluster may be driven by stromal elements. Features of inflammation (S100A8, S100A9, SND1, RNF213), cytokine signal transduction (STAT1, STAT3, DHX9), and angiogenesis (HMGB1, HMGB2, TYMP) were all at highest levels in class 1. Moreover, class 1 encompassed tumors positive for the antigen presentation signature (HLA-A, SEC24C, CD74, TAP1, TAPBP, PSMB9) (112), molecular chaperones that positively regulate proliferation (NASP, CDC37), cell cycle proteins (RCC2, YWHAG, PAICS, NME2, GART) as well as proteins related to chromosomal rearrangements and telomere maintenance (RUVBL2, PCNA, PRKDC), features of DNA repair (PARP1, TOP2B, APEX1) and UPR (HSP90B1, HYOU1, SEC61A1, SRPRB, SSR1, HSP90AA1, TLN1). Several enzymes involved in the glucolytic/glucogenolytic pathway were at higher levels in the low grade-differentiated class 3, which also was selectively positive for the *downregulated* CIS gene signature (114), reflecting the papillary origin of these tumors.

Pathway analysis predicted selectively for class 1 up-regulation of transcription (tRNA-aminoacylation, transcriptional regulation of pluripotent stem cells), anabolism, and heat shock response (**Table 2**). Pathways enriched selectively in class 2 involved alterations of the ECM and signaling through G-protein coupled receptors and RHO GTPases (**Table 2**). Consistent with a more differentiated phenotype, class 3 presented with increased enrichments for tight and adherens junctions, xenobiotic metabolism and apoptosis (**Table 2**). Gene set enrichment analysis between classes 1 and 3 predicted, among others, enrichments in IFN type-I signaling, in MYC, and in E2F transcriptional targets for class 1 (**Supplementary Table 2**).

Class	Reactome pathway	P - value	Proteins
Class 1	Cytosolic tRNA aminoacylation	2.11E-07	[AARS, AIMP2, EPRS, GARS, IARS2, MARS, QARS, RARS, TARS, WARS, YARS]
	Transcriptional regulation of pluripotent stem cells	6.57E-06	[ACTL6A, HNRNPA1, HNRNPF, HNRNPM, HSP90AA1, PARP1, PCNA, PRPF8, PTBP1, RAD23B, RAN, RUVBL1, SF3B1, SF3B2, SF3B4, SNRPB, SNRPD1, SNRPD3, SNU13, SRSF1, SRSF7,STAT3, SUMO2, TIA1, TIAL1, USP7]
	Pentose phosphate pathway (hexose monophosphate shunt)	4.27E-04	[G6PD, PGD, TALDO1, TKT]
	Regulation of HSF1-mediated heat shock response	2.93E-03	[CCAR2, FKBP4, H2AFZ, HIST1H2BJ, HIST1H2BK, HIST1H3A, HIST1H4A, HSP90AA1, HSP90AB1, HSPA9, KPNB1, PTGES3, RAB2A, RAN, SET, SLC25A5, SLC25A6, USO1, YWHAE]
	DEx/H-box helicases activate type I IFN and inflammatory cytokines production	7.91E-03	[DDOST, DHX9, HMGB1, PRKCSH]
	Mitochondrial protein import	2.34E-02	[CHCHD3, HSPA9, HSPD1, SLC25A6, TOMM22, TOMM70]
Class 2	RHO GTPases activate CIT	6.07E-04	[MYH11, MYH9, MYL9,MYL12A, MYL6]
	Scavenging by Class A Receptors	5.48E-03	[COL1A1, COL1A2, FLNA]
	Cell-extracellular matrix interactions	1.16E-02	[ACTB, ACTN1, FLNA]
	Keratan sulfate degradation	1.24E-02	[LUM, OGN, PRELP]
Class 3	Tight junction interactions	4.82E-05	[CDH1, CTNNA1, CTNND1, F11R, ITGA6, ITGB4, JUP, KRT5, PLEC, VASP]
	Caspase-mediated cleavage of cytoskeletal proteins	9.90E-05	[ADD1, CDH1, DBNL, DSG2, PLEC, SPTAN1, TJP2]
	Formation of annular gap junctions	1.11E-04	[CLTA, CLTB, DNM2, MYO6]
	Glucose metabolism	3.12E-04	[CALM1, FBP1, FBP2, GBE1, GNPDA1, GOT1, GOT2, HK1, PFKL, PGAM1, PGK1, PGM2, PYGL, PYGB, TPI1]
	Type I hemidesmosome assembly	1.04E-03	[ITGA6, ITGB4, KRT5, PLEC]
	Glutathione synthesis and recycling	3.27E-03	[CNDP2, GSTO1, GSTP1, MGST1, MGST2, MGST3]
	Adherens junctions interactions	4.17E-03	[CDH1, CTNNA1, CTNND1, JUP]
	Phase II conjugation	1.73E-02	[ABHD14B, BPNT1, CNDP2, COMT, GSTO1, GSTP1, MGST1, MGST2, MGST3, UGT1A6]
	Aryl hydrocarbon receptor signalling	2.01E-02	[ABHD14B, ACY1, BPNT1, CNDP2, COMT, GSTO1, GSTP1, MGST1, MGST2, MGST3]
	Clathrin-mediated endocytosis	2.34E-02	[ACTR2, ACTR3, ARF6, ARPC1A, ARPC2,ARPC4, CLTB, CLTA, CLTB, CTTN, DNM2, HSPA8, VAMP3]

Table 2. Class-specific pathway enrichments of the three proteomic subtypes. *The analysis was conducted in the Cytoscape plug-in, ClueGO, with a two-sided hypergeometric test. P - value corresponds to the Benjamini - Hochberg correction.*

Collectively, the results, based on both pathology and molecular characteristics, indicate an aggressive profile for class 1, a heterogeneous-mesenchymal phenotype for class 2 and a luminal, more differentiated, and less aggressive phenotype for class 3. To further evaluate the validity of this observation, comparison of the proteomics profile of the three NMIBC classes to that of the 19 MIBCs was performed. As shown in **Figure 8**, MIBC exhibited protein expression patterns highly similar to class 1. This was also verified with Principal Component Analysis (**Figures 9a and 9b**), where MIBC appeared more proximal to class 1 tumors and conversely, more distant to class 3 tumors (**Figure 9b**).

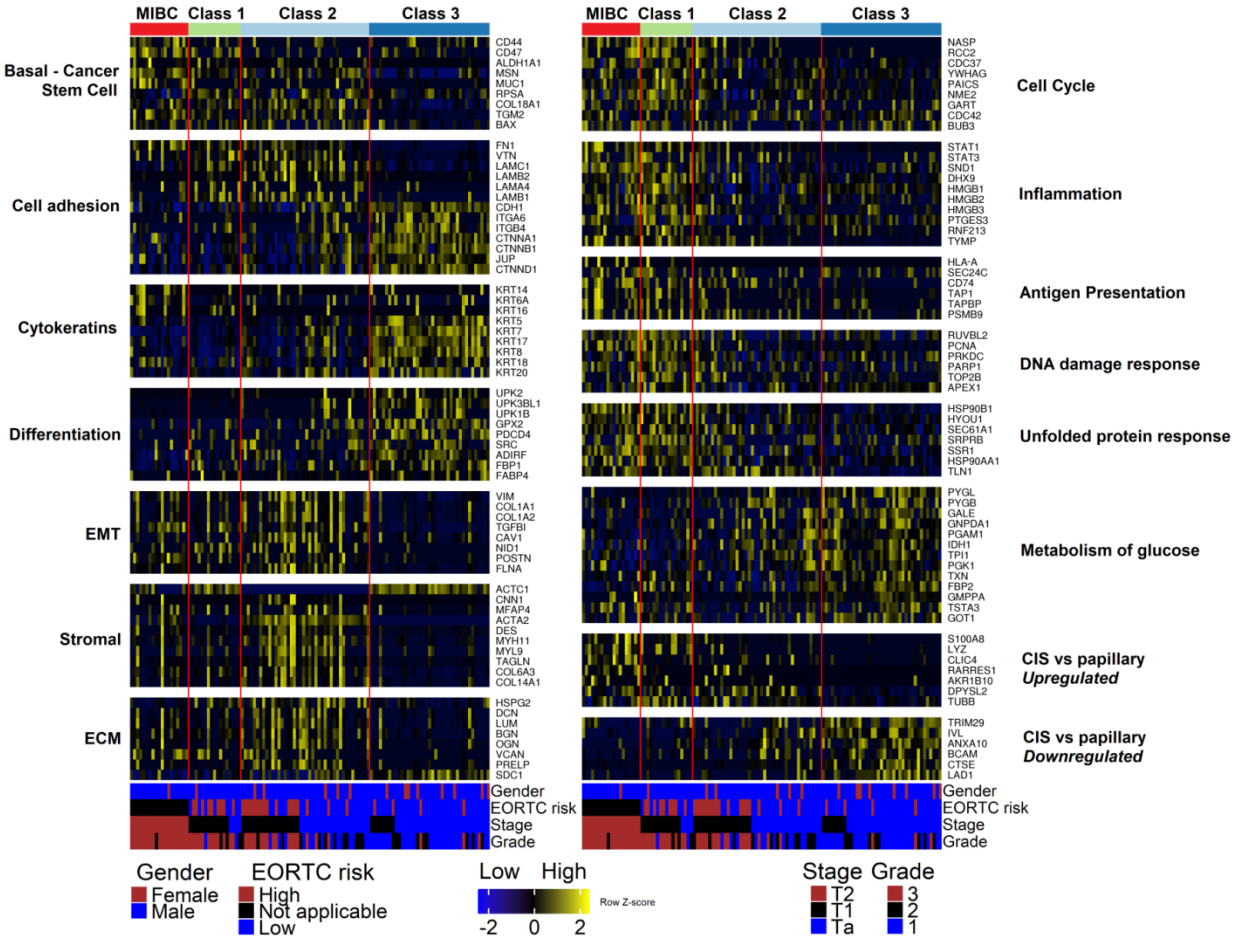


Figure 8: Heatmap showing the expression pattern of proteins across the three identified NMIBC subtypes and MIBC samples (columns). *Proteins (rows) are organized in molecular themes related to previous bladder cancer subtyping studies. EMT = Epithelial-Mesenchymal Transition, ECM = Extracellular Matrix. CIS = Carcinoma in-situ.*

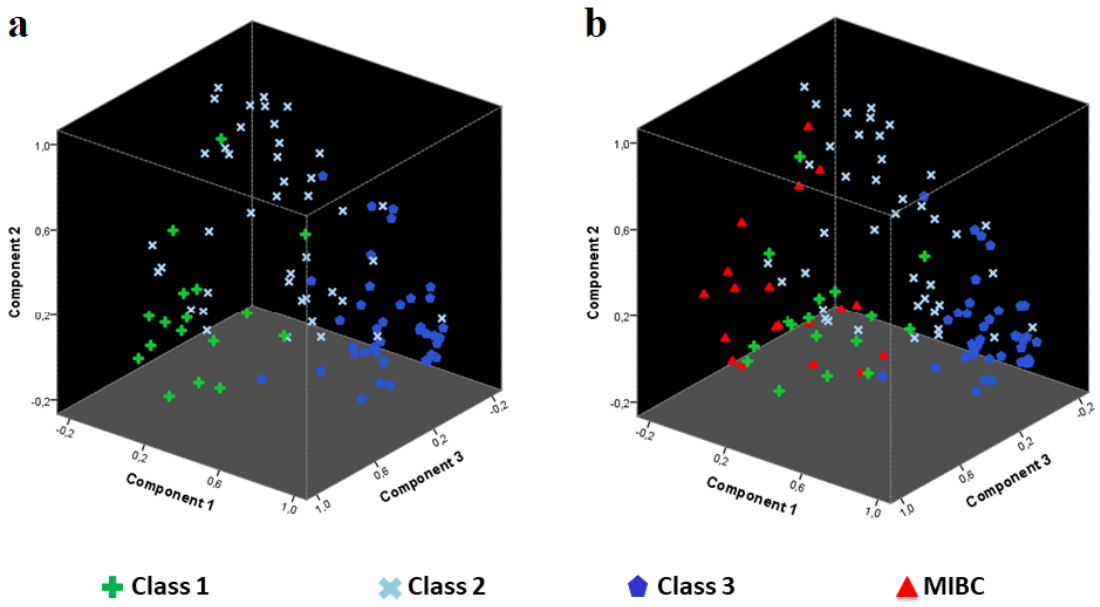


Figure 9. Scatter plot visualizations of the Principal Component Analysis (PCA) output*. Plots illustrate the distribution of the 117 analyzed BC samples in the rotated space, as a function of their segregation to subtypes. Distance between the plotted samples reflects their phenotypic relationships.* ***(a)*** *PCA applied only at the 98 NMIBC dataset, using as input the 626 significantly different proteins among the three classes (variance explained by the first three components = 62.5%).* ***(b)*** *PCA applied simultaneously at the 98 NMIBC and the 19 MIBC patients. Tumors from class 1 (green crosses) exhibit close proximity to the MIBC group (red triangles), whereas the inverse is observed for class 3 samples (blue pentagons). Patients from class 2 (light blue x marks) are allocated more heterogeneously (input = 618 significant proteins between MIBC and the three NMIBC classes; variance explained by the first three components = 65.4%).*

4.2.4 Validation of the proteomics classification

The validity of the proteomics classification was assessed in terms of its relation to the RNA-seq classification system of the UROMOL cohort (n = 476 samples). The UROMOL subtypes are cited in italics (e.g. *class 1*, *class 2*, *class 3*) with *class 2* suffering significantly worse progression free survival rates when compared to the other two (118). To assess the relationships between gene expression and protein abundance among the two studies, genesets overexpressed in each UROMOL subtype were generated (n = 3 genesets, namely, Luminal for *class 1*, CIS-like for *class 2* and Basal-

like for *class 3*). The three genesets were screened against the three proteomic class comparisons, and statistically significant enrichments were calculated with GSEA. Significant enrichments were identified for the Luminal and the CIS-like genesets only in the proteomics comparison class 1 vs class 3, with the Luminal geneset being expectedly overrepresented in the proteomics class 3 and the CIS-like geneset in the proteomics class 1 (**Figure 10**).

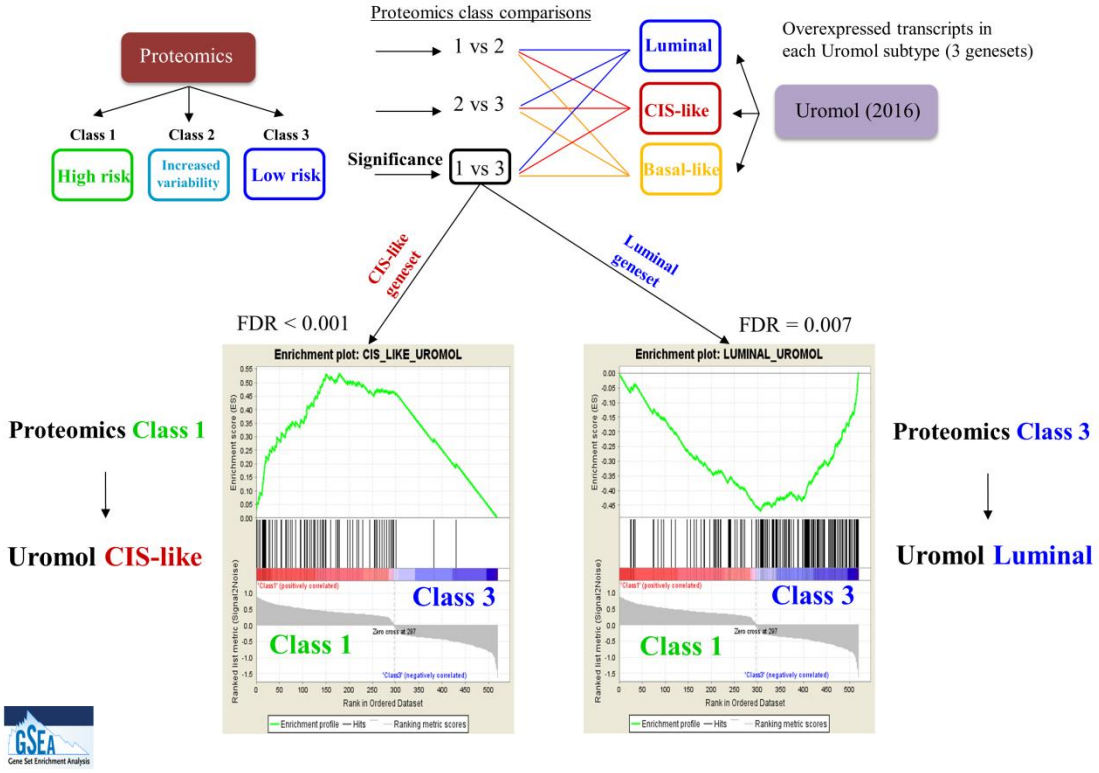


Figure 10: GSEA analysis of the proteomics subtypes against the three UROMOL derived genesets. *Three genesets (Luminal, CIS-like, Basal-like) containing unique overexpressions for each UROMOL subtype (class 1, class 2, class 3, respectively) were prepared and analyzed against the three proteomics comparisons, with GSEA. Statistically significant results were obtained only for the proteomics comparison class 1 vs class 3, in which the CIS-like geneset was associated with the proteomics class 1 while the Luminal with class 3 tumors.*

Given the phenotypic associations between luminal and aggressive subtypes of the UROMOL and the proteomics classifications, we then investigated if the proteomics

samples could be meaningfully classified into the three UROMOL RNA-seq subtypes. Consequently, we chose to train a classifier on the UROMOL data and test it on our proteomics samples. A Random Forest algorithm was selected for this purpose, due to their i) powerful accuracy as they offer a double assessment of the important features (permutation and Gini impurity tests), ii) built-in cross-validation system that performs bootstrapping and out of bag (OOB) error estimation independently for each tree iii) simplicity in tuning the parameters of the classifier (139). The results showed that 87.5% of the proteomics class 1 samples were classified as UROMOL's *class 2* (Progressed subtype), whereas 91.7% of the proteomics class 3 were classified into the Non-Progressed subtypes of the UROMOL (25% as *class 1* and 66.7% as *class 3*). Interestingly, the intrinsic molecular heterogeneity of the proteomics class 2 was also reflected at the results, as 35.5% of the samples joined the progressed *class 2* and the remaining were stratified into the non-progressed UROMOL subtypes *class 1* and *class 3* (**Table 3**). The results validate the high risk/ low risk states of the proteomics classes 1 and 3, respectively.

Proteomic subtypes	**UROMOL subtypes**		
	Luminal *class 1*	**CIS-like** *class 2*	**Basal-like** *class 3*
Class 1	12.5%	**87.5%**	6.3%
Class 2	5.8%	35.3%	**58.8%**
Class 3	25%	8.3%	**66.7%**

Table 3: Results from the Random Forest classifier, trained on UROMOL's subtypes. *The table depicts class assignments of the proteomics samples (as percentages) into the UROMOL's RNA-seq subtypes, averaged between the two feature selection approaches (permutation and Gini impurity).*

4.2.5 Post –machine shortlisting of potential prognosticators for NMIBC aggressiveness

Given the stratification of NMIBC patients to groups of cancers that exhibit apparent high-risk (class 1) and low-risk (class 3) molecular and pathological features, the proteomics data were then investigated for molecules potentially reflecting disease aggressiveness. Towards that end, proteins overexpressed uniquely in class 1 or class 3 and also exhibiting concordant regulation at the comparison "MIBC vs NMIBC" were investigated (workflow shown in **Figure 11**, respective results in **Figure 12**).

Specifically, 73 proteins were at high abundance solely in class 1 (compared to the other two) and in the MIBC dataset compared to NMIBC. Interestingly, these were implicated in highly relevant GO Biological Processes e.g. negative regulation of cell cycle arrest, response to type I interferon and ERBB2 signaling pathway (**Supplementary Table 3**). Along the same lines, 82 proteins were at higher abundances in class 3 in comparison to all other classes and in NMIBC compared to MIBC. These 82 proteins were found to be involved in metabolic pathways, such as hexose catabolic process, cellular oxidant detoxification, and glycerolipid catabolic process (**Supplementary Table 4**).

The validity of these 155 proteins (73 overexpressed uniquely in class 1 and in MIBC; 82 uniquely overexpressed in class 3 and in NMIBC), as potentially differing between NMIBC patients with higher/lower risk for progression was investigated at the mRNA level in UROMOL (118) and LUND (112) datasets (**Figure 11**). This approach was chosen since follow-up data are not available for the samples investigated in this study.

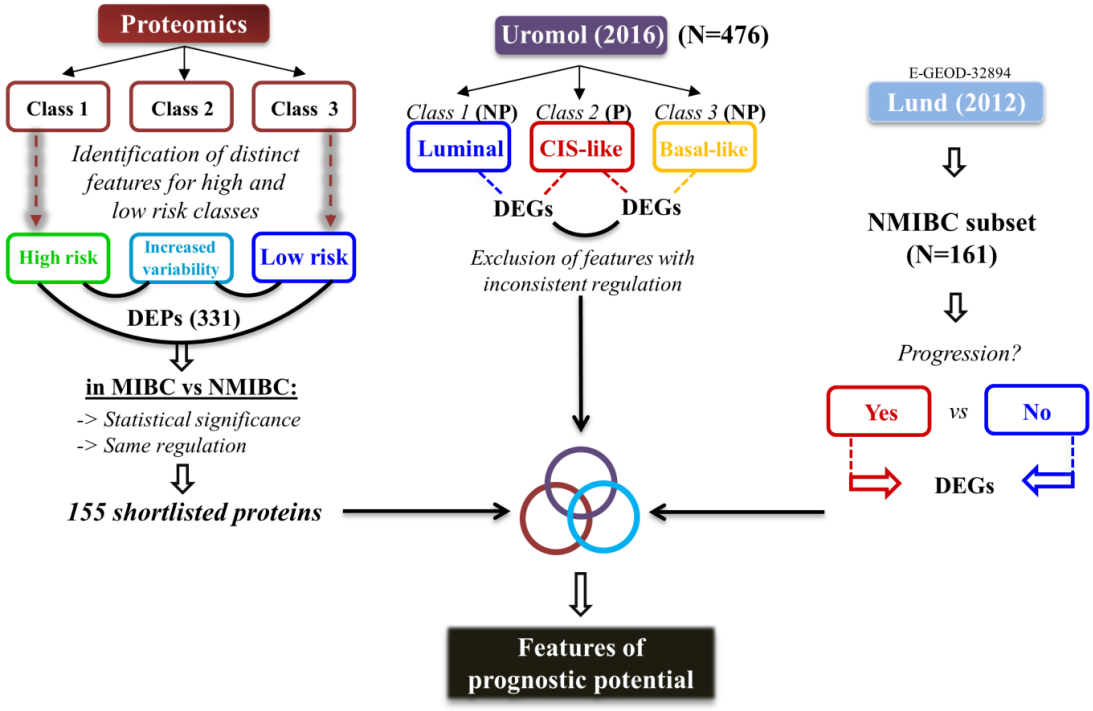


Figure 11: Workflow for the cross-omics analysis for the identification of molecular features potentially marking disease progression. *In brief, 155 proteins distinguishing the two extreme proteomic classes 1 and 3, were found to consistently differ between*

MIBC and NMIBC patients and were further screened for overlaps against the progressed vs non-progressed subtypes of the UROMOL and against the progressors vs non-progressor comparison of the LUND study. DEPs = Differentially Expressed Proteins, DEGs = Differentially Expressed Genes, NP = Non-Progressed subtype, P = Progressed subtype

Overlaps between the 155 shortlisted proteins and statistically significant transcripts differing in UROMOL *class 2* (Progressed subtype) versus *classes 1* and *3* (Non-Progressed) were defined (as described in Methods). These corresponded to 96 overlapping features with consistent regulation at mRNA and protein levels (**Figure 12**). Of these, using interaction analysis (string-db.org), features overexpressed in high risk groups ($n_{features}$ = 54) were found to be part of a network (p = 2.5E-14), that harbored four signaling hubs each consisting of at least four nodes. The four signaling hubs represented the processes of unfolded protein response (HSP90AB1, HSP90B1, HYOU1, HSPD1, PDIA4, TXNDC5), pre-mRNA splicing (EIF4A3, EIF4G1, SF3B2, SNRPA, RALY), antigen presentation (HLA-A, HLA-DRA, TAP1, RAB7A), and post-translational modifications associated with the Oligosaccharyltransferase (OST) complex (RPN1, RPN2, SSR1, DDOST), potentially marking the significance of the above mechanisms and molecules in NMIBC progression. In the case of the LUND cohort, comparative analysis between Progressors (n = 144) and Non-Progressors (n = 17) NMIBC cases, highlighted 28 mRNAs overlapping with the 155 shortlisted proteins (**Figure 12**). Among them, features overexpressed in the Progressors group were similarly, associated with RNA processing (RALY, SNRPA, EIF4A3, SF3B2, YARS), inflammation (S100A8, S100A9, HMGB2), but also tumor growth (USP7, NASP, PDIA4, SND1). Features found at high levels in the Non-Progressors group were involved in detoxification (CYB5R1, MGST2, PRDX5), in protection from cell-senescence (ASAH1, PYGL), and in actin dynamics and tissue morphogenesis (PDLIM1, ARF6, ACTR3, CAST, KRT13). In addition, Non-Progressors exhibited high levels of the H2AFY, which is a histone variant localized at regions of heterochromatin, thought to play a positive role in the suppression of cell proliferation.(140, 141) Considering all three studies, the overall intersecting area was 18 features, of which 12 were at high levels in high risk tumors and were implicated mostly in RNA processing, inflammation, and growth signaling **(Figure 12, Table 4)**.

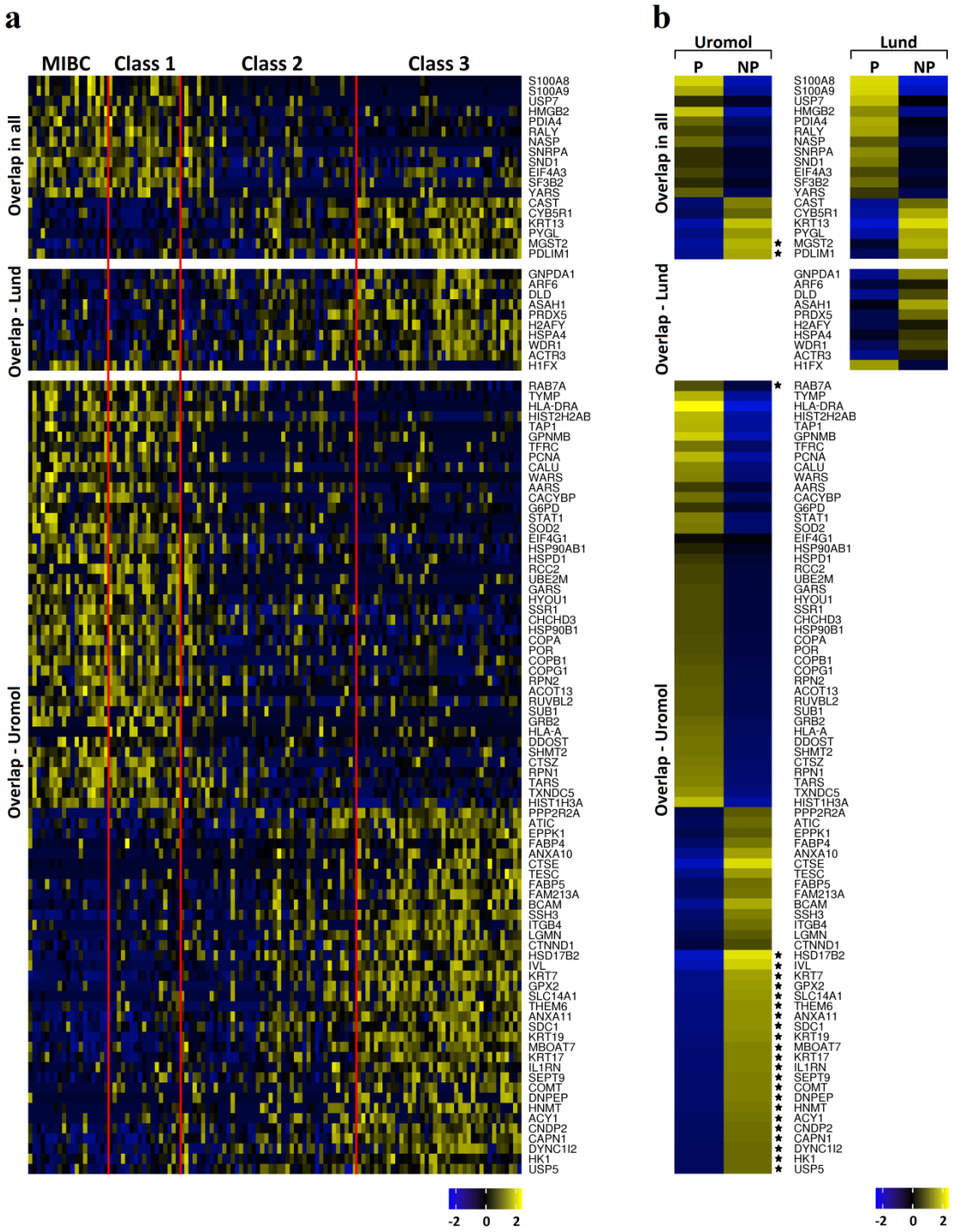


Figure 12: Heatmaps showing protein and transcript abundances of features concordantly deregulated between our study, UROMOL, and LUND cohorts. (a) *Abundance of proteins uniquely overexpressed in the proteomic classes 1 and 3, filtered for significance and same regulation at the comparison MIBC vs NMIBC.* ***(b)*** *Mean abundances of the respective mRNAs in the Progressors (P) and Non-Progressors (NP)*

groups from the UROMOL and LUND cohorts. In the former, the mean abundance of the Progressors (P) and Non-Progressors (NP) groups corresponds to UROMOL subtypes class 2 and class 3, respectively. Asterisk (*) marks those features that reached statistical significance at the UROMOL comparison class 2 vs class 1 and hence, for these transcripts, the depicted mean abundance of the NP group corresponds to UROMOL class 1.

			Proteomics	**Proteomics**	**UROMOL 2016**	**LUND 2012**
	Gene name	**Protein name**	**Log2 FC MIBC / NMIBC**	**Log2 FC class1 / class3**	**Log2 FC *class2* / *class3***	**FC Progressors – NonProgressors**
1.	S100A8	Protein S100-A8, Calprotectin L1L subunit	2.42	4.87	1.81	1.75
2.	S100A9	Protein S100-A9, Calprotectin L1H subunit	1.94	4.40	1.40	1.31
3.	HMGB2	High mobility group protein B2	1.13	1.14	1.65	0.40
4.	NASP	Nuclear autoantigenic sperm protein	1.54	Only in class1	0.84	0.27
5.	USP7	Ubiquitin carboxyl-terminal hydrolase 7	1.27	2.93	0.45	0.45
6.	PDIA4	Protein disulfide-isomerase A4	0.58	0.76	0.84	0.32
7.	RALY	RNA-binding protein Raly	0.63	1.02	0.58	0.31
8.	SNRPA	U1 small nuclear ribonucleoprotein A	0.87	1.38	0.47	0.24
9.	SND1	Staphylococcal nuclease domain-containing protein 1	0.85	0.59	0.48	0.21
10.	EIF4A3	Eukaryotic initiation factor 4A-III	0.51	1.14	0.58	0.21
11.	SF3B2	Splicing factor 3B subunit 2	0.78	0.98	0.44	0.20
12.	YARS	Tyrosine--tRNA ligase, cytoplasmic	1.54	2.53	0.78	0.18
13.	KRT13	Keratin, type I cytoskeletal 13	-1.19	-2.58	-1.13	-1.58
14.	CYB5R1	NADH-cytochrome b5 reductase 1	-1.21	-2.18	-0.39	-0.71
15.	PYGL	Glycogen phosphorylase, liver form	-1.11	-2.34	-0.77	-0.69
16.	CAST	Calpastatin	-1.40	-1.82	-0.59	-0.45
17.	MGST2	Microsomal glutathione S-transferase 2	-0.80	-0.74	-1.01*	-0.40
18.	PDLIM1	PDZ and LIM domain protein 1	-0.69	-0.51	-0.88*	-0.36

***Table 4. Eighteen features intersecting between our study, UROMOL, and LUND cohorts**. The depicted 18 features could significantly discriminate between the "extreme" proteomic classes 1 and 3 as well as between the proteomic MIBC and NMIBC datasets, and were also identified as consistently deregulated at the mRNA levels when comparing Progressors versus Non-Progressors in both UROMOL and LUND studies. Red color indicates up-regulation and blue down-regulation. The fold changes of these features at the protein level when comparing MIBC vs NMIBC is also provided. Data from the UROMOL study are depicted as found in the Supplementary*

TableS3 of the UROMOL published article(118). Asterisk (*) marks features that reached statistical significance at the UROMOL comparison class 2 vs class 1, and thus the regulation shown corresponds to the latter comparison.

5. Chapter II: Transcriptomics

We compiled a discovery meta-cohort of 1,135 Bladder Cancer (BLCA) microarray transcriptomes along with two RNA-seq validation sets, and addressed the disease as a molecular continuum of alterations. Using the stage as a checkpoint variable that reflects the cumulative processes of tumor progression, we investigated how molecular processes and gene expression levels change, starting from non-malignant adjacent urothelium (NAU) and continuing through the disease stages Ta, T1, T2, T3, and T4. The analysis aimed to shed light on previously unknown aspects of the molecular pathophysiology of BLCA, highlighting pathways whose activation progressively increases or diminishes with cancer growth, while also reporting for the first time on the gene co-expression profiles of the disease stages. Based on the analysis for the monotonal traits, and towards better patient monitoring, we propose an 8-gene signature capable of prognosing 5-year survival for patients with BLCA.

5.1 Materials and Methods

5.1.1 Dataset search strategy

To perform a comprehensive investigation of available molecular data, we searched for Bladder Cancer (BLCA) omics studies in public repositories. All genomic urothelial cancer data from cBioportal (including The Cancer Genome Atlas) were downloaded. Gene Expression Omnibus (GEO) was queried for transcriptomics, additional genomics or protein array datasets using the search terms “bladder cancer” and “urothelial carcinoma”. We also queried ArrayExpress using the special filter “Array express data only” to obtain any additional datasets missing from GEO. All cohort data published or updated between 2010 and 2020, annotated as *Homo sapiens*, coming from tissue samples with sample size >10, were retrieved. All used datasets were published and downloaded anonymized. This resulted in the collection of 105 datasets comprising more than 8,000 individuals, encompassing genomics, methylomics, transcriptomics and proteomics data, derived from a variety of technologies used in the field.

5.1.2 Inclusion and exclusion criteria

We analyzed tumor transcriptomes and selected studies having at least clinical or pathological stage information per subject. Samples or datasets collected after administration of (neo)adjuvant chemotherapy, as well as secondary or recurrent bladder cancers were excluded, in order to minimize drug-induced variation in gene expression. To compile a microarray discovery set while preserving as high integrity as possible, we chose microarray data quantified by the most frequently used single-color channel vendors (Affymetrix and Illumina). The overall workflow is summarized in **Figure 13**. This resulted in a final dataset of 1,135 patients coming from 12 different studies (Table A1). We used the TCGA-BLCA-2017 and the IMvigor210 studies for the validation purposes, and particularly analyzed primary BLCA samples collected prior to administration of (neo)adjuvant chemotherapy (n samples: TCGA = 188, IMvigor = 132).

5.1.3 Description of the discovery meta-cohort

The discovery cohort included the following microarray datasets: GSE121711, GSE93527, E-MTAB-1940, GSE31684, GSE104922, GSE128959, GSE83586, GSE48276, GSE52219, GSE69795, GSE13507, GSE48075. These data (summarized in Table 1), comprised of 1,054 primary bladder cancer tumor transcriptomes of treatment-naïve patients without any prior cancer history, along with profiles from 81 non-malignant urothelium tissue adjacent to the tumor site (NAU); correspond to a total of 1,135 gene expression profiles. Stage distribution among the utilized datasets is shown in Table A1. Table 1 shows sample allocation to clinical variables both for the discovery and validation sets. In the discovery set, the ratio of men : women was 3.5 : 1, with equal distribution among NMI and MI disease ($p = 0.99$) and similar mean age at baseline diagnosis (68 years, $p = 0.81$). Percentages of NAU, NMI, and MI in the dataset were 7.1%, 43.5%, and 49.4%, respectively, with the grade distribution being as follows: 16.5% low grade, 48.8% high grade disease, with the remaining samples lacking available grade information. Detailed histological records were missing for 71.5% of the cohort, with the most frequently reported histology among the available records being urothelial/papillary (23.3%), and squamous differentiation being the most frequent variant (1.3%).

5.1.4 Assessing the right method for batch effect removal

Gene expression distribution is tighly linked with the experimental conditions in which it is being tested. As a consequence, expression distributions between samples being processed in different batches, dates, or in different labs, differs significantly (142). This phenomenon in molecular biology has been documented as *batch effects*, and is defined as the change in the data distribution caused by non-biological factors affecting the experiment (142). Various tools have been developed to correct for batch effects. The first ones [surrogate variable analysis (SVA) and limma] estimate a set of inferred variables (eigenvectors), which are then used to apply a linear correction (factor analysis, singular value decomposition, or regression) to the data prior to statistical testing (143, 144). ComBat (145), is a Bayesian method in which every gene undergoes an independent scale adjustment based on location parameters calculated either parametrically or non-parametrically. It is the most widely used tool, while recently an extension (ComBat_seq) has been developed for RNA-seq data (146). Other batch correction methods have been designed for situtations when the number of batch variables is not known, but rather, is in question. CONFETI (147) adjusts distributions in expression quantitative trait loci analysis, by inititally identifying and removing out genetic effects, utilizing principal component analysis. RUV-2 assesses batch variation with respect to a set of (user determined) negative control genes, which are defined as those genes whose regulation is known to be stable across the tested experimental conditions (148). An extension (RUV_seq) for count RNA-seq data has been also available (149).

Here we test batch correction in the discovery meta-cohort using three methodologically different approaches, namely the limma (function removeBatchEffects), ComBat, and RUVnormalize, with the aim of selecting the most optimal one. We define the 4 following criteria based on which we evaluate quality of the adjusted (batch-free) data:

i) Distribution of gene expression among samples should exhibit a limited variance in the Relative Epxression Plots
ii) Sample allocation in the Principal Component Anaysis 2D plots should not associate with dataset of origin
iii) Housekeeping genes are usually expressed in higher levels and their expression levels show decreased variability compared to other genes (150,

151). Variation-to-mean expression plots should thus, illustrate the degree of preservation of these properties for the three batch correction methods.

iv) There are genes whose expression is known across non-malignant adjacent urothelium (NAU), NMI and MI bladder cancer, or across NAU, low-grade and high-grade disease. We use 12 known genes as positive control reference set and contrast their expected differential expression against the three batch correction methods.

Three batch corrected datasets were produced using the functions removeBatchEffect (limma package), ComBat (sva package), and naiveRandRUV (package RUVnormalize). For the removeBatchEffect and ComBat methods, adjustments were parametric and no covariate matrix was supplied. In the ComBat function, batch effect was scaled after adjusting for mean (parameter: mean.only = FALSE). In the naiveRandRUV correction, a set of negative control genes whose regulation remains unchanged across conditions had to be initially defined. This is a property intrinsic to the housekeeping genes, so we used a set found in Eisenberg and Levanon (151), which derives 3,804 genes expressed uniformly across a panel of tissues. Out of these 3,804 genes, in order to provide only the most stably expressed subset in the adjustment, we firstly calculated their median absolute deviation (of gene expression) individually in each of the 12 datasets and extracted the top 700 least variably expressed genes per dataset. The 12 lists, each one containg 700 gene names, were then parsed into a Cross-Entropy Monte-Carlo rank estimation algorithm (152), which returns a unified aggregated rank ordered list of the most important ones. The algorithm provides the option of defining an arbital number of most important genes, and we set this number at 300, as this has been shown to work efficiently in a previous naiveRandRUV microarray correction (153). The other parameters of the rank aggregation method were: weights=NULL, method="CE", distance="Spearman", seed=42, maxIter = 1000, convIn=7, rho=0.01, weight=.25, v1=NULL, N = 1000, standardizeWeights = TRUE. Runing this, a ranked list of 300 genes was optimally determined in the 120^{th} iteration at a value of 63,282. We then performed the naiveRandRUV adjustment with a regularization coefficient of 0.01 (parameter nu.coeff), and arranged the desired rank of the estimated unwanted variation factors k to 12, reflecting the number of datasets to be adjusted.

5.1.5 Monotonicity in pathway activation and de-activation across BLCA stages

To identify genes that form a continuum of changes across BLCA stages, each of the 5 disease stages was initially compared against non-malignant adjacent urothelium (NAU). A total of 3,018 genes differed significantly (Mann-Whitney $p < 0.05$) while having the same orientation of change in all comparisons. We refer to this set of genes (n=3,108) as Concordantly Differentially Expressed Genes (CDEGs). CDEGs were utilized to infer pathway activation scores and to create stage co-expression networks. Pathway activation scores per sample were calculated with the ssGSEA-GSVA method [21], using the Molecular Signature Database libraries of Hallmark, Canonical Pathways (Reactome subset), C3 (GO biological processes subset) and C5 (GTRD subset of transcription factor targets). Dorothea (https://github.com/saezlab/dorothea) [22] was utilized to assess regulon activity. To further identify the subset of pathways whose activation had a monotonal trait across non-malignant adjacent urothelium (NAU) and disease stages, each pathway's activation scores across stages were compared to NAU with Mann-Whitney tests, and direction of change was defined based on fold change (= Mean of stage – Mean of NAU). Monotonicity for a pathway was defined as being significantly different in all stage comparisons to NAU, and also having a continuously larger/smaller fold change with increasing stage. Stromal infiltration scores were imputed with the ESTIMATE algorithm [23].

5.1.6 Construction and analysis of stage specific coexpression networks

Gene-pair co-expression weights among the 3,108 CEDEGs were approximated with ensemble learning, using GENIE3 [24], while direction of co-expression (positive/negative) was determined by the Spearman's coefficient. Out of the 3,108x3,108 = 9,659,664 gene-pair weights calculated individually per condition (i.e. NAU and 5 BLCA stages), gene-pairs with the highest GENIE3 weights, being also positively correlated based on the Spearman's coefficient, were used to construct networks. The cut-off for this selection was determined based on the gene-size of the resulting networks: to avoid network saturation we opted to keep their gene sizes close to half the number of CDEGS (= 1,554 genes), which resulted in setting the cut-off to the top 5,600 gene-pairs. Networks were constructed with igraph and were analysed with Louvain clustering [25] to identify local modules of co-expression relationships (communities) per condition. The top five in size (= number of genes) communities of co-expressed genes per condition were analysed for Gene Ontology Biological Processes with clusterProfiler [26]. Potential drug targets in the co-expression networks

were defined based on the betweenness centrality metric [27] using default cut-offs (computed with igraph).

5.1.7 Monotonicity in individual gene expression and development of a prognostic signature

We utilized CDEGs to extract genes whose expression levels was monotonically increasing or decreasing with higher stage. Monotonicity for a gene was defined as being a CDEG and additionally having a continuously larger/smaller fold change with increasing stage (fold change, as defined in each disease stage versus NAU). Functional annotation and enrichment were performed using PubMed and the online tool GeneCards (https://ga.genecards.org/), respectively. Out of the monotonal subset, 43 genes were found of prognostic value (Cox univariate association to 5-year outcome). Eight of these genes, validated in the TCGA-BLCA dataset, were further utilized to construct a sample-wise scoring system, the 8-gene prognostic signature, by summing the expression values of upregulated genes (n = 4) while subtracting the downregulated genes (n = 4). Before calculations, each of the gene expression values per sample were divided by the gene's variation across the dataset, in order to minimize the effect of individual gene variability on the final signature score:

$$Si = Ai/VarA + Bi/VarB + Ci/VarC + Di/VarD - Ei/VarE - Fi/VarF - Gi/VarG - Hi/VarH \quad (9)$$

Where Si denotes the sample-wise derived signature score, Ai, Bi, Ci, Di, and Ei, Fi, Gi, Hi, denote the sample-wise gene expression of the 4 upregulated and 4 downregulated genes, respectively, while Var denotes the gene expression variance across the entire set of samples. The 8-gene signature along with the disease stage were further used as input in a multivariate Cox regression model, to identify if there is independent prognostic value. This procedure was applied both on the discovery meta-cohort and on the TCGA validation data.

Since our procedure for the detection of monotonal traits in genes and pathway activities involved multiple filtering criteria, to avoid over-elimination due to Type-II error, significance was defined at unadjusted Mann-Whitney $p < 0.05$. In contrast,

significance for pathway over-representation (clusterProfiler output) was determined by FDR correction at $p < 0.05$. Categorical variables were investigated for significance with the Pearson's chi-squared test and were adjusted for multiple hypotheses (package RVAideMemoire). All reported correlation scores correspond to the Spearman's Rank coefficient. Cox proportional hazards regression was performed with the packages survminer and survival, and statistical significance was determined with the log-rank method. CIBERSORT analysis was conducted in the web platform https://cibersort.stanford.edu/, and only samples with successful deconvolution ($p < 0.05$, $n = 350$ samples) were further used for the statistical comparisons of relative immune populations among stages. Read counts from the IMvigor data were acquired from the IMvigor210CoreBiologies and were normalized with the variance stabilization transformation [28]. Unless stated otherwise, all processing, analyses and visualizations were conducted in the R statistical environment (version 4.0.2).

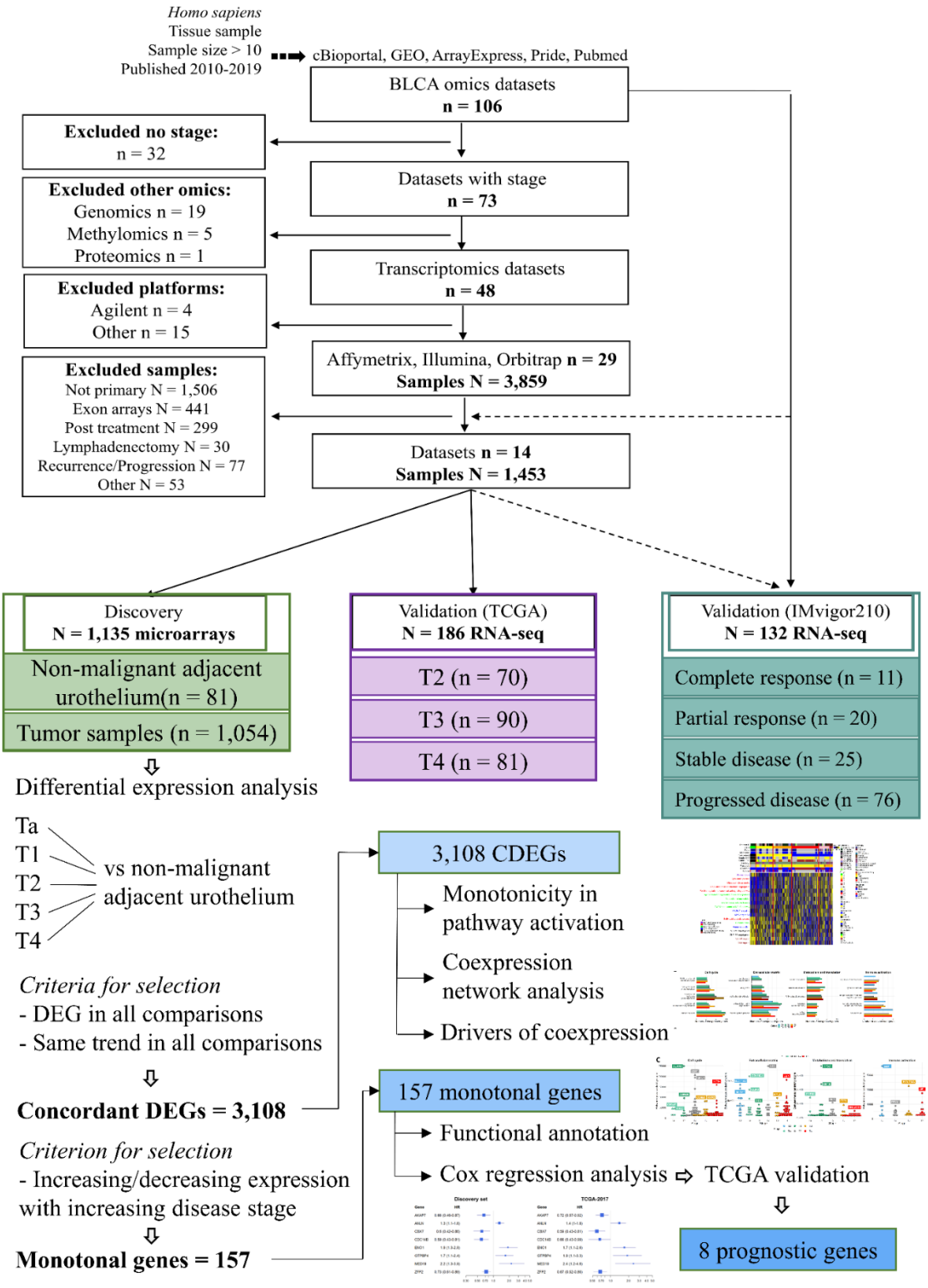


Figure 13*. Study design and workflow for the analysis of the selected primary BLCA transcriptomes.*

5.2 Results

5.2.1 Assessment of batch correction methods

We tested the effect of batch correction of three methods and assessed their effriciency with a number of criteria (described in Methods). All three methods produced efficient relative log expression and principal component plots, with the naiveRandRUV showing a larger inter-sample variation and standard deviation compared to the other two. Housekeeping genes wh

5.2.1 Increasing activation levels of the Wnt, mTORC1, and MYC pathways associate with Bladder Cancer development and growth

For initial assessment of the gene expression relations between non-malignant adjacent urothelium (referred to as NAU) and cancerous samples, we performed differential expression analysis of NAU versus NMI and NAU versus MI samples. To investigate transcriptional changes associated with increasing malignancy, we compiled genes being differentially expressed in all stage comparisons to NAU, showing also a persistent change, either up or down, and we further denoted them as Concordantly Differentially Expressed Genes (defined in Methods; CDEGs, n = 3,108). Due to their consistent regulation compared to NAU, CDEGs likely reflect fundamental alterations occurring during bladder carcinogenesis, and thus we focused the analysis on this particular set of genes. We initially performed a GSVA-ssGSEA analysis using CDEGs, aiming to identify pathways and biological processes whose activation is continually enhanced or diminished through disease stages. Towards increasing disease stage, results indicated gradually stronger activations of several mitotic processes, positive regulation of the canonical Wnt pathway, mTORC1 signaling, expression of MYC targets, degradation of anaphase inhibitors, metabolism of nucleotides, mobility of formins, and the TNFR2/non-canonical NF-kB pathway (**Figure 14**). Conversely, diminished activity was recorded for the lipid and fatty acid catabolic processes, for the metabolism of heme, and interestingly for the circadian clock process (**Figure 14**). Regulon activity per sample was additionally estimated and respective scores between disease stages and normal tissue were compared. This analysis highlighted GATA3 and GLI2 regulons whose activity was significantly diminished with increasing malignancy (**Figure 14**).

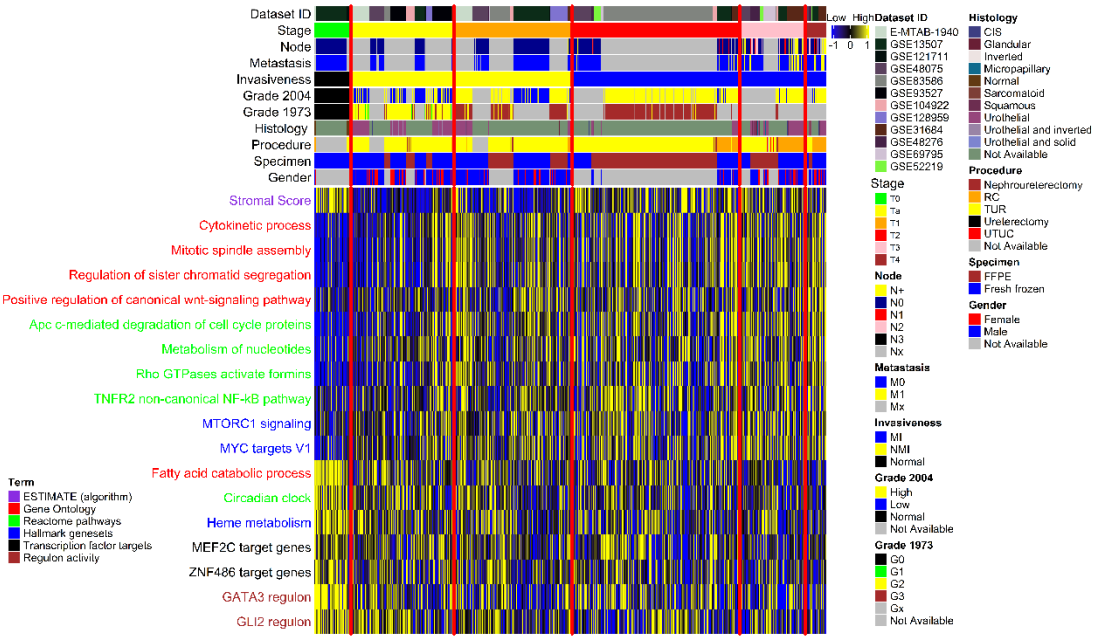


Figure 14. Excerpt of the pathways showing a monotonal increase or decline in their activation scores with higher stage. *Pathway activation scores have been z-scaled across samples for visualization. Pathways are colored based on their database of origin. The top side of the heatmap presents dataset and clinical information. Samples (columns) have been ordered based on the stage variable; from left to right: non-malignant adjacent urothelium (NAU), Ta, T1, T2, T3, and T4. Red lines indicate boundaries between adjacent stages.*

5.2.2 Stage specific coexpression reveals variable and stable subnetworks with BLCA development and growth

To further investigate co-expression alterations occurring in the disease stages, an integrated network-pathway analysis was performed. Stage specific networks were constructed and clustered to identify communities (sub-networks) of co-expressed genes (described in methods). We analyzed the five largest communities (based on number of genes) per disease stage and NAU, and used the Gene Ontology – Biological Process (GO-BP) library to identify affected molecular processes (**Figure 15**). The analysis revealed large differences in gene co-expression between NAU and disease stages with four out of the five largest communities associating clearly with specific biological processes.

Three out of the top five communities were consistently detected in all BLCA stage networks. Based on examination of their enriched processes, these were labeled as 1) the cell-cycle community, 2) the ECM and developmental community 3) the metabolic and translational community (**Figure 15A, 15B**).

The cell cycle communities of the different tumor stages involved a total of 288 genes, 178 of which had a proliferation related GO-BP annotation. Hypergeometric tests for each of the stage networks indicated highly similar cell cycle BPs being over-represented across stage. An excerpt of the statistically most significant ones along with the number of implicated genes is presented in Figure 15B. Out of the 178 cell-cycle genes, 80 were co-expressed consistently in all stage networks. These were also upregulated in tumor compared to NAU, possibly forming the backbone of cell proliferation in BLCA (**Supplementary Table 5**). The gene size of this community increased towards higher disease stage (Ta n = 118, T1 n = 148 and MIBC n = 168-170 genes). The communities also included genes lacking cell-cycle GO annotation (11.9% for Ta, 17.6% for T1, 21.9% for T2, 24.9% for T3, and 29.8% for T4). An over-representation test of the 110 genes lacking GO-BP annotation across the stages revealed that 20 of them participate in the metabolism of nucleotides (unadjusted p = 0.03), likely suggesting a rewiring component controlling both the regulation of proliferation and the processing of nucleotides. To detect the most relevant potential drug targets within the cell cycle communities, we determined their betweenness centrality scores. The most prominent ones included CDC5, KIF2C, FOXM1, AURKB, CDT1, SMC4, CCNB1, RRM2, and KIF14 (**Figure 15C**).

The community of ECM and developmental processes encompassed a total of 291 genes and was enriched in cell-cell communication and cell-matrix interaction processes, in responses to microenvironmental stress, as well as in differentiation programs of epithelial, mesenchymal and stem cells (**Figure 15B**). This GO-BP composition suggests that these co-expression signals originate either from tumoral or non-tumoral cells, or might be the product of their interaction. For example, the process of extracellular matrix organization included co-expressions of 15-36 genes (depending on the stage network) of which COL13A1, FGFR4, FOXF2 and SCUBE1 were co-expressed only in the NAU samples compared to disease stages. Contrarily, 26 genes, including mediators of epithelial to mesenchymal transition (COL6A1/A2, COL16A1, MFAP5, MMP11) were co-expressed in tumor tissue but not in NAU. In line with

recent observations [30], we noticed that NAU presented with an active ECM remodeling profile. Sixteen of the ECM associated genes were co-expressed both in NAU and in the NMIBC stages, including genes promoting basolateral tumor cell migration (MMP2, CTSK, PDPN), fibrotic collagens (COL1A2, COL6A3, COL14A1, COL15A1), and pro-angiogenic factors (PDGFRA, RECK), suggesting a pro-tumorigenic potential in the NAU. However, expression in the NAU was predicted to be driven by ALDH1A2 and MFAP4 (**Figure 15C**), genes which are both notoriously down-regulated in other genitourinary malignancies compared to normal samples [31, 32], likely suggesting tumor suppressive roles. In contrast, co-expression in the Ta stage (confined to the internal lining of the bladder), was predicted to be regulated by the hub genes COL16A1 and CLIP3. CLIP3 interacts with both AKT1 and AKT2 [33], and may therefore have an important role in the early AKT/PI3K/mTOR axis of hyperplastic carcinogenesis.

The community of metabolism and translation encompassed mitochondrial, translational and multiple metabolic processes being activated during carcinogenesis, and was more profound in the T1 and more advanced tumors. Cellular respiration, translational initiation, mRNA catabolic process, nonsense mediated decay and protein targeting to ER were consistently enriched in most BLCA stages. The results highlighted a set of 12 genes commonly co-expressed across stages for these processes, including COX7B, DLD, NDUFS4, UQCRFS1, PAIP2, RPL15, RPL30, RPL7, RPS23, RPS27, RPS27A, RPS4X.

Besides the abovementioned consistently detected communities in BLCA, a community enriched in processes of immune cell differentiation, cytokine secretion, and GPCR activity was identified in the NAU and the MIBC stages and involved both innate and adaptive responses, as well as processes of immune cell adhesion and migration. This immune associated community presented low variation in the composition of genes participating in the co-expression networks between NAU and MIBC stages. Out of the 17 genes of the process of T-cell activation that were commonly co-expressed at the MIBC stages, 15 were also co-expressed in the NAU samples. To further investigate these observations, the transcriptome data per sample were deconvoluted into relative abundances of immune cell populations using CIBERSORT, and cell fractions between disease stages were compared. Significant results were obtained for the following populations: CD8+, activated CD4+, activated NK, Monocytes, Macrophages M2 and

activated Dendritic cells (**Figure S6**). Results indicated differential commitment of immune cells to NAU and BLCA stages. NAU samples (n = 37) were significantly more infiltrated with CD8+ (p = 0.046) and monocytes (p = 4.6E-4) than tumor samples (n = 313), consistent with an over-representation of the excluded over the inflamed phenotype, previously seen in the IMvigor210 trial data [34]. However, compared to tumor, NAU samples had significantly less abundance of activated CD4+ cells (p = 5.28E-3), of macrophages (p = 16.8E-5), of activated dendritic cells (p = 0.0002) and of activated NK cells (p = 0.015). Generally, NMIBC had lower immune infiltration than MIBC. Activated dendritic cells were significantly increased in Ta tumors (n = 34) compared to other BLCA stages (p = 0.024). Abundance of CD8+, of activated NK cells, and of M2 macrophages increased linearly with higher malignancy. Interestingly, AIF1 a gene that promotes macrophage survival and M2 polarization [35], was predicted to be a driver of immune co-expression in the T4 tumors. Along with the Cibersort results which indicate a higher abundance of M2 macrophages in the T4 samples, we hypothesize that AIF1 is actively involved in the immune suppression, and thus, its expression levels might indicate putative candidates for immune checkpoint inhibition.

A

Communities of coexpressed genes	NAU	Ta	T1	T2	T3	T4
Cell Cycle	-	+ (105/118)	+ (124/148)	+ (134/170)	+ (130/170)	+ (118/168)
Extracellular matrix	+ (105/121)	+ (119/170)	+ (145/192)	+ (122/179)	+ (84/141)	+ (144/201)
Metabolism and translation	-	+ (67/134)	+ (119/220)	+ (65/123)	+ (44/165)	+ (95/221)
Immune activation	+ (87/108)	-	-	+ (95/124)	+ (108/138)	+ (97/123)

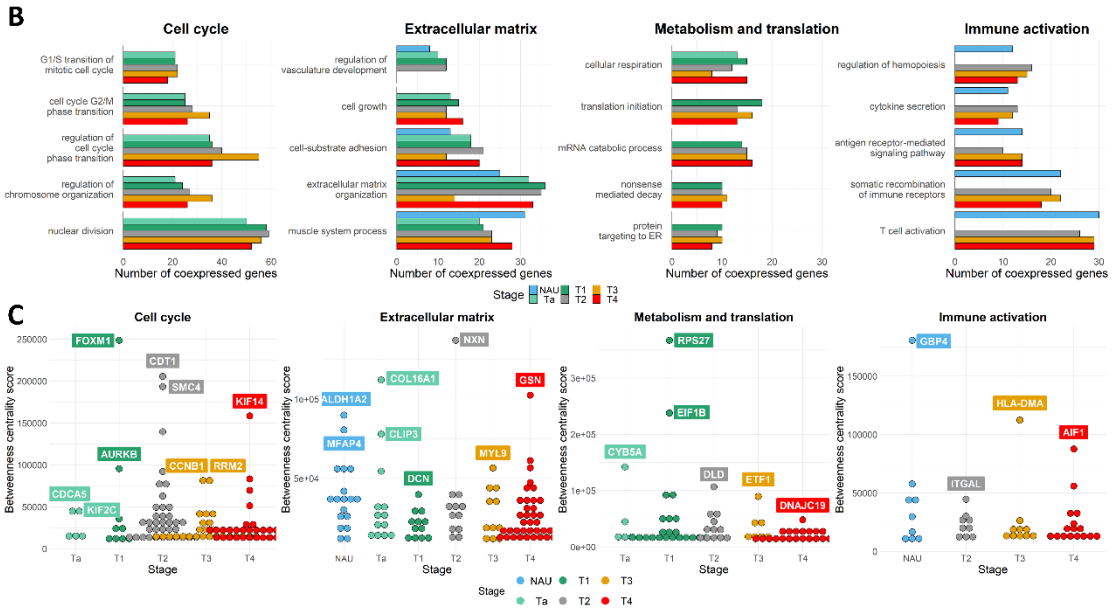


Figure 15. Biological process analysis of the largest in size co-expressed communities identified in each BLCA stage network. *(A) Coherent communities identified and characterized across non-malignant adjacent urothelium (NAU) and disease stages. Presence of a community is indicated by the + symbol. Numbers in parentheses show the fraction of genes with Biological Process annotation relevant to the community, with respect to the total number of genes found to be co-expressed in the community. (B) Barplots of the most significantly enriched biological processes per community depicting number of co-expressed genes for each. (C) Hub genes identified across the studied conditions based on the betweenness centrality scores (y axis).*

3.2. Monotonicity in individual genes, prognostic signature and validation

Using the 3,108 CDEGs, we extracted genes having a monotonal (i.e. continuously increasing or decreasing) change in expression in the spectrum NAU-Ta-T1-T2-T3-T4. A total of 157 genes were identified having the trait of monotonicity, of which 118 were up- and 39 were downregulated with increasing stage **(Figure 16, Supplementary**

Table 6). Functional analysis revealed that for 46 of these genes, experimental evidence on mediating cell cycle progression exists. Upregulated cell-cycle associated genes (n = 44) were not phase specific and included cyclins, DNA polymerases, regulators of the cohesin complex and kinetochore components. The list also included 23 genes involved in signal transduction, 6 of which (ARHGAP11A, AURKA, CDKN3, PBK, PLK1, RRM2), promote cell-cycle progression and were all upregulated with increased stage. The data also indicated an overactivation of the Wnt pathway with increasing disease stage, with its upstream inhibitor APCDD1 being downregulated and its activating ligand WNT2 upregulated. Fourteen of the 157 genes were transcriptional or translational regulators, including genes with known upregulation in bladder cancer (transcription factors E2F1, DEPDC1 [36, 37]). Based on the monotonal changes with higher stage, increased androgen receptor activity may be predicted, as both its translational enhancer BUD31 [38] and its downstream transcription factor ELK1 [39] were upregulated. Four of the 157 genes (HTR2C, LRP8, NENF, NMU) are involved in neurotransmission or neuronal development, all upregulated. Among the 157 genes, 21 were of not well described or unknown function, including the oncogenic factor TRIM65 [40] found upregulated with increasing bladder cancer stage. Further functional enrichment using GeneCards for the 157 genes verified their involvement in cell-cycle pathways, with top hits being related to the regulation of the Anaphase promoting (APC) complex (score = 31.53), to PLK1 (score = 24.47) and Aurora B (score = 20. 95) signaling, as well as to TP53 (score = 19.06) and RB1 (score = 17.85) cell cycle checkpoint control (**Figure 16C**). Univariate cox regression analysis indicated 48 genes with potentially prognostic impact at $p < 0.01$.

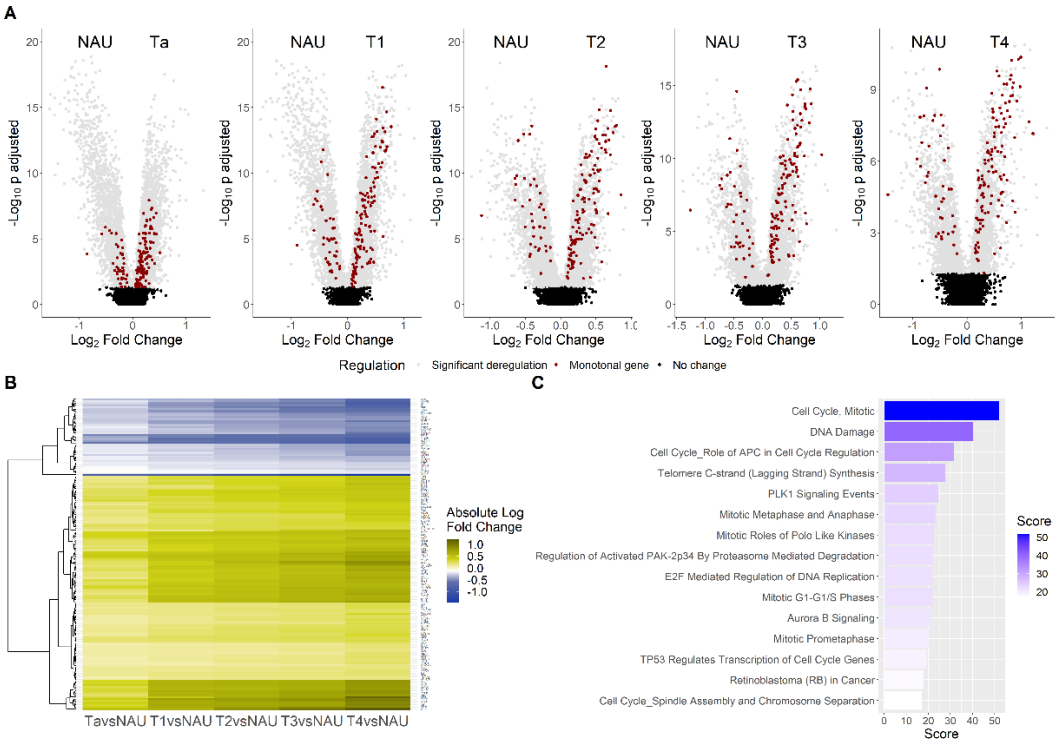


Figure 16. Differential expression analysis between non-malignant adjacent urothelium (NAU) and BLCA stages. *(A) Volcano plots of the five stage comparisons to NAU, with red color indicating the 157 genes showing a monotonal trend of expression across stages. (B) Heatmap of the absolute fold changes of the 157 monotonal genes, being either continuously up- (yellow color) or downregu-lated (blue color), in the comparisons between disease stages and NAU. (C) Top 15 pathways of the 157 monotonal genes, sorted by the GeneCards enrichment score.*

In lack of an RNA-seq dataset comprising all the disease stage spectrum of BLCA incidents, the observed stage alterations in the discovery set were investigated for their reproducibility in the TCGA-BLCA RNA seq data [42]. To align the validation samples to the discovery set, patients with unknown history of prior treatment for non-muscle invasive bladder cancer, as well as patients with history of other malignancies were excluded. Differential expression analysis among the available stage comparisons (T3 vs T2 and T4 vs T3) in the TCGA data validated 43 of the 157 monotonal genes. Cox regression analysis in the TCGA data validated 8 out of the 48 monotonal genes that were found to be of prognostic value in the discovery set (**Supplementary Table 7**), including MED19, ENO1, ANLN, GTPBP4, higher levels of which associating with

worse survival and CBX7, ZFP2, AKAP7, CDC14B higher levels of which associating with better survival probability (**Figure 17A**). We utilized these 8 genes to construct a combined score that characterizes each individual sample (see methods). Along with the disease stage, the 8-gene signature had independent prognostic value, both in the discovery and in the validation set (**Figure 17B**). Specifically, we constructed a survival model to compare 5-year survival rates between those with high and low 8-gene signature scores (defined by a median cut-off). Patients with a high 8-gene signature score had a worse 5-year survival probability in the discovery set and this find-ing was validated in the TCGA data (**Figure 17C**). The 8-gene signature did not differ significantly between males and females ($p = 0.36$), and was weakly associated with age and stage (Figure S7), suggesting its independent value with respect to other clinical variables. In an attempt to verify the co-expression analysis findings, stage specific co-expression networks were also created using the TCGA data, and were clustered with the Louvain algorithm. GO-Biological process analysis of the communities validated the differential segregation of the cell-cycle, extracellular matrix and immune activation processes to distinct communities (**Figure S8**). In order to validate the value of the AIF1 as a candidate biomarker for response to immunotherapy, we analyzed RNAseq data from the IMvigor210 study, a trial investigating response to atezolizumab immunotherapy in patients with metastatic BLCA. High AIF1 expression in the IMvigor data associated with a complete response to atezolizumab **(Figure 17D**).

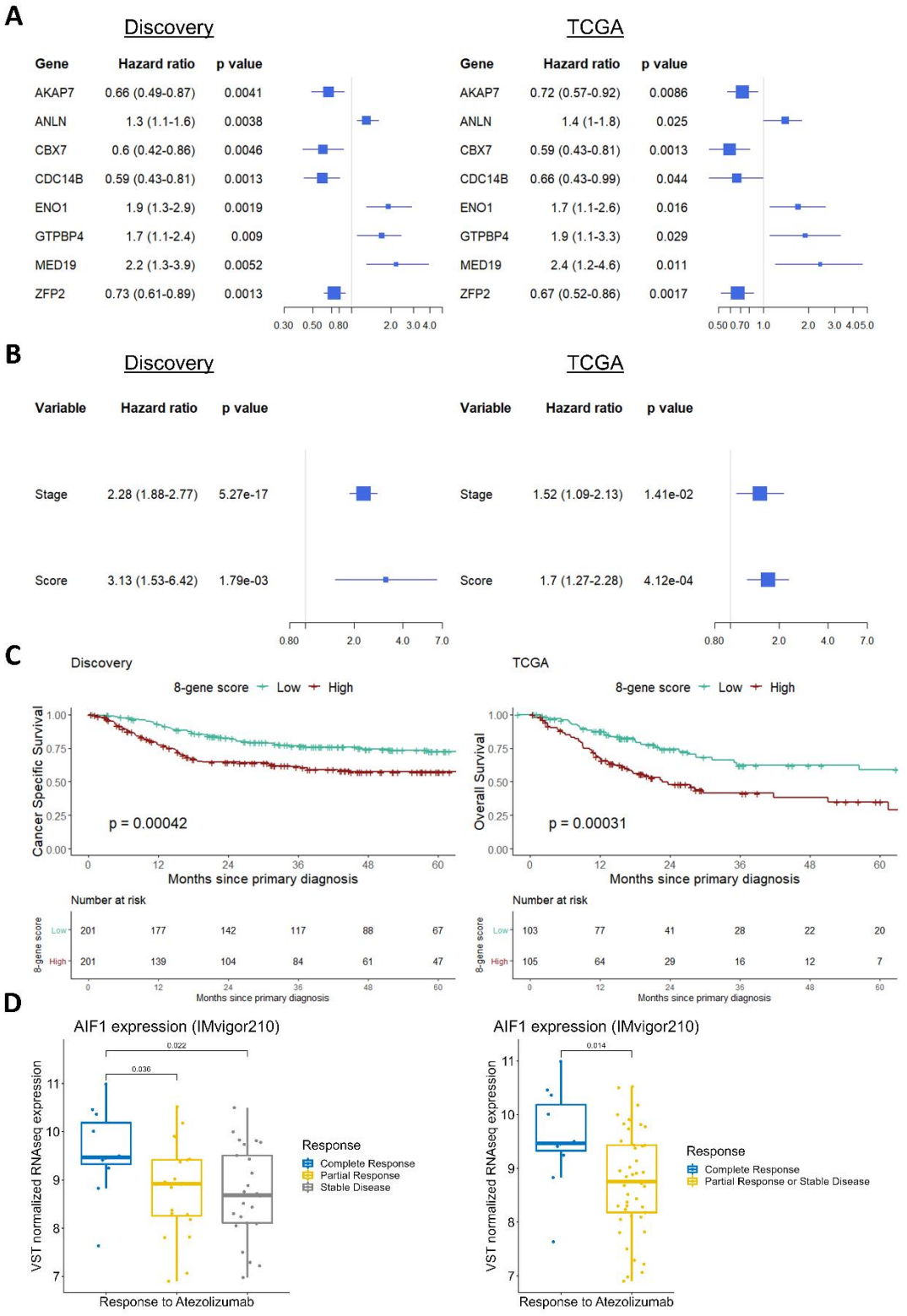


Figure 17. Validation of key findings in the TCGA-2017 and the IMvigor210 cohorts. *(A) Forest plots showing hazard ratios (HR) for the 8 monotonal genes having univariate prognostic value both in the discovery and the TCGA validation datasets. (B) Multivariate analysis of stage and the 8-gene signature score in the discovery and*

the TCGA validation sets. (C) 5-year survival analysis between patients with high and low 8-gene signature score, in the discovery and the TCGA data. (D) Data from the IMvigor210 trial illustrating AIF1 expression across response to atezolizumab immunotherapy groups.

6. DISCUSSION

This study reports on the first proteomics classification of NMIBC based on an unbiased comprehensive LC-MS/MS approach. Three proteomic subtypes were identified and their molecular profiles were characterized based on existing subtype-specific features. Classes 1 and 2 shared some basal features, whereas class 3 tumors presented with a more differentiated/luminal phenotype. This is in contradiction to the UROMOL study where most of the NMIBC samples were characterized as luminal; this difference may be attributed to sample and study design differences but also to divergence of gene expression and tumor cell phenotype, originally reported in MIBC subtyping (122). Molecular subtypes of MIBC have been shown to transcend pathological staging, but this has not been confirmed for NMIBC. Instead, NMIBC subtypes appear to not follow (118) the aforementioned observation, something that is also reproduced in this study.

Class 1, contained mostly (13/17) T1 and Grade 3 (12/17) samples, was rich in some basal markers (CD47, TGM2, BAX, COL18A1, MSN), in lamina propria components, and presented with low levels of proteins that facilitate cell-basal membrane and cell-cell attachment, possibly indicating increased cell motility. Additionally, these tumors expressed at high levels proteins of the cell cycle progression, MYC and E2F transcriptional targets, all being features of aggressive BC (122, 123), also reflected at their higher proximity to the MIBC proteome (**Figure 8**). Considering also its size as captured in our cohort (17/98), these findings are in line with observations from a recent transcriptomics meta-subtyping of BC, supporting that approximately 20% of NMIBC resemble MIBC at the molecular level (117). The high risk nature of these tumors is also supported by their classification predominantly to the progressed UROMOL subtype *class 2* (**Table 3**). Moreover, the proteomics class 1 expressed the CIS-like UROMOL geneset (**Figure 10**), features of antigen presentation, inflammation, and was also enriched in IFN-γ response pathway, in concordance with RNA-based observations that high grade, aggressive basal subtypes both in MIBC and in NMIBC are presenting with an immune-infiltrated phenotype (154, 155). Proteins involved in

the DDR and the UPR were also solely overexpressed in class 1. Consequently, we suspect that class 1 tumors may likely respond to immune checkpoint, PARP or HSP inhibitors.

Class 2 patients shared similar basal and cell adhesion protein expression patterns with class 1 but lacked expression of the basal surface antigen CD47. While being heterogeneous both at clinicopathological (**Table 1**) and molecular level (**Figure 9**), tumors from class 2 were commonly dominated by the overexpression of several ECM/mesenchymal proteins (**Figure 8**). The latter may have impeded the detection of (epithelial) tumor cell signals, however, high presence of stromal elements around the tumor cells might be a sign of "reactive" responses against tumor spread [5]. Intriguingly, a subset of class 2 tumors (mainly of Ta stage) expressed at high levels both basal and luminal features (**Figure 8**). The intrinsic molecular heterogeneity of class 2 was also reflected at the results from the classifier (**Table 3**). The results here collectively suggest potentially increased variability with regards to outcome for class 2 and also, the need for larger sample sizes for the identification of its biologically relevant subgroups. Since pathway predictions indicate activation of Rho-GTPases, responders to inhibitors of these proteins or their downstream effectors are likely to segregate in this class.

Class 3 tumors were characterized by increased abundance of KRT20, CDH1, and UPKs denoting high levels of differentiation, and were also $KRT5^{+}$, $ITGA6/ITGB4^{+}$, resembling UrobasalA tumors from the LUND taxonomy (112) and *classes 1* and *3* from UROMOL (118). The low risk nature of these tumors was validated in the UROMOL cohort were the majority of class 3 patients (91,7%) were classified at the non-progressed UROMOL subtypes (*classes 1* and *3*; **Table 3**). Therefore, patients in this class may not need tight surveillance. Analysis for class specific pathways indicated that class 3 was selectively enriched in detoxification activity by glutathione, which is considered to inactivate cisplatin, offering chemo-resistant properties to tumor cells (156). This could partially explain why luminal variants of MIBC appear to be insensitive to standard chemotherapy or chemo-radiation (111, 157, 158).

Shortlisting of proteins concordantly overexpressed in the proteomics class 1 and in MIBC (n = 73) as well as in the proteomics class 3 and in NMIBC (n = 82) followed by subsequent investigations for their regulation in the mRNA data between

"Progressors" and "Non-Progressors" groups from the UROMOL (118) and LUND (112) cohorts, indicated a set of 96 and 28 consistently deregulated features, respectively (**Figure 12**). Approximately, half of these molecules were overexpressed in the "Progressors" groups (i.e. UROMOL's *class 2* or LUND's Progressors) and based on their molecular function, were found to represent four main processes (described below).

i) Unfolded protein response (UPS): differences at the mRNA levels of features involved in UPS have been previously reported among low grade and stage NMIBC subtypes (119). Accordingly, in our analysis, aggressive subtypes (proteomics class 1 and UROMOL *class 2*) also overexpressed features of the UPS (HSP90AB1, HSP90B1, HYOU1, HSPD1, PDIA4, TXNDC5), suggesting an association between protein stability and the development of NMIBC.

ii) Inflammation and immune recognition: proteins belonging to damage associated molecular patterns (such as S100A8/A9 and HMGB2), reported to increase with BC stage (159, 160), were found at high abundance in the proteomics class 1 and in the Progressors groups from the UROMOL and LUND cohorts. At the same time, high levels of STAT1, EIF4A3, and EIF4G1 in the proteomics class 1 as well as in the UROMOL *class 2*, which are potentially controlling PD-L1 regulation based on evidence from melanoma (161), were detected. Moreover, four molecules (HLA-DRA, HLA-A, TAP1, and RAB7A) involved in antigen presentation were identified as overexpressed in the proteomics class 1 and in UROMOL progressed *class 2*. Interestingly, this is also in line with the Lund taxonomy where high mRNA levels of antigen presentation molecules characterized the most aggressive BC subtypes (112).

iii) RNA processing and post translational modifications: Previously, *in silico* comparisons between the tissue proteome of MIBC and NMIBC by our team (138), suggested a significant up-regulation of proteins involved in the transcriptional-translational machinery of the former (MIBC). Here, a group of features involved in two types of RNA processing, pre-mRNA splicing (SNRPA, SF3B2, RALY, EIF4A3) and tRNA-aminoacylation (AARS, GARS, TARS, WARS, YARS), was detected as overexpressed in the proteomics class 1 and in UROMOL progressed *class 2*, while a subset of them (SNRPA, SF3B2, RALY, EIF4A3, YARS) was also significantly overexpressed in Progressors from the LUND cohort. This may be reflective of the

increased biosynthetic, translation, and turnover rates required to maintain fast proliferation (162, 163). In addition, as a novel finding, up-regulation of the components of the Oligosaccharyltransferase (OST) complex (RPN1, RPN2, DDOST, SSR1) was detected in the proteomics class 1 and UROMOL progressed *class 2*. The OST complex catalyzes glycosylation of nascent peptides at asparagine residues, a modification which has been found to be critical for surface localization of epidermal growth factor and whose inhibition induces senescence in receptor-tyrosine-kinase-dependent tumors (164).

iv) Oncogene signaling: Among the overlaps between the proteomics data and UROMOL or LUND cohorts overexpressed in the aggressive-progressive groups, were features of attributed oncogenic nature, such as USP7, NASP, SND1, GRB2, and PCNA. As examples, USP7 (shortlisted from all 3 cohorts) is a hydrolase containing a Ubiquitin-like domain and functions as a de-ubiquitinylation enzyme targeting regulators of cell-cycle, eventually protecting them from ubiquitinylation and proteasomal degradation. *In vivo* knockout studies has demonstrated that depletion of USP7 results in destabilization of MDM2 and inhibition of proliferation (165). In the same context, NASP (deregulated in all 3 cohorts) is chaperone facilitating transportation of histones into the nucleus (166); its inactivation by microRNA-29c resulted in cell-cycle arrest in gastric cancer (167). Similar anti-tumor effects upon microRNA-29c overexpression, have been also observed in BC (168). SND1 (deregulated in all cohorts) has attributed pleiotropic functions, in breast cancer being involved in the TGF-β1 pathway, where it acts as an essential transcription activator of the Smad proteins (169), but also, found to interact with members of the STAT family, serving as co-activator of downstream genes (170). GRB2 links activated surface receptors with intracellular signal transducers and has been shown to be critical for cell-cycle progression and actin reorganization, contributing to tumor metastasis (171). In BC cell lines, GRB2 was found to be overexpressed in the lack of EGFR overexpression or H-Ras mutations (172). PCNA is associated with genomic stability, as it is involved in processes such as DNA repair, cell-cycle progression and chromatin remodeling (173), and in the case of BC for its increase in immunostaining levels with increasing cancer stage and grade (174).

Integration of BLCA molecular data has been previously performed in the context of characterizing molecular subtypes [43], or validating results of either (single cell)

scRNA-seq [44] or RNA-seq re-analysis [45]. In this study, we performed an integration meta-analysis of datasets from non-tumor-bearing adjacent urothelium and BLCA stages, aiming to identify continuous, as well as concerted gene expression alterations with increasing malignancy. To our knowledge, this is the first attempt to associate molecular alterations with clinical classification based on the analysis of more than one thousand well-characterized, primary tumor datasets. Instead of focusing on molecular subtypes, we increased power and addressed the disease as a continuum under the assumption that individual samples reflect different snapshots of the whole process. Our novel design based on the hypothesis of continuous evolution through stages has been successfully applied here and resulted in novel findings on gene regulation associated with cancer pro-gression.

Starting from a normalized expression dataset comprising 12 microarray cohorts, we identified genes being differentially expressed between disease stages and NAU (CDEGs), and further analyzed them for pathways/processes being progressively altered with higher stage, for changes in the co-expression profiles of stages, as well as for genes showing a monotonal change in expression with higher stage. Our analysis highlighted expected landmark pathways, such as mTORC1 pathway [46] and MYC targets [47] which were upregulated, but also novel downregulated pathways such us the circadian clock and the metabolism of heme. These results associate for the first time BLCA progression with the disruption of the circadian homeostasis and to iron metabolism deficiencies, events that are thought to be tumorigenic [48, 49], but their exact mechanism of action is not well understood. In addition to the GATA3 regulon, a known driver of luminal biology, we found a novel progressive downregulation of the GLI2 regulon. GLI proteins are transcription factors of the Sonic hedgehog (Shh) pathway and although GLI2 expression levels positively correlate with more invasive BLCA cell lines, Shh genes do not behave accordingly [50]. Our results validate these observations, as the entire regulon of the GLI2 TF was inactivated with increasing BLCA malignancy, suggesting no potential therapeutic effect in its inhibition.

Tumor initiation in the bladder is thought to occur within the basal layers of the urothelium, when the accumulated burden of mutations dysregulates cell's homeostatic mechanisms, favoring uncontrolled proliferation over apoptosis [18]. The process of initi-ation is lengthy in time, and affects the entire neighborhood of the adjacent cells, which are continuously exposed to a pro-tumorigenic environment. Here we find that

most of the alterations in the non-tumor-bearing adjacent cells involve genes operating during embryogenesis or during ECM remodeling (Community 3, Table S2). These are likely among the first to acquire an organized pattern of co-expression. Interestingly, co-expression in NAU was driven mostly by ALDH1A2 (and partly by MFAP4), which catalyzes the for-mation of retinoic acid (RA). In the progenitor cells, during embryonic development, receptors of the RA form complexes with chromatin modifiers, leading to the activation of self-renewal and differentiation programs [51]. These data give rise to the hypothesis that the cells in the NAU may be expressing and maintaining parts of a stem cell-like RA related program. Indeed, the biological process of response to RA appears enriched in the ECM communities of NAU ($p = 4.57e-05$), and T2 ($p = 1.12e-02$) stage. T2 tumors are far more dedifferentiated in comparison to both NAU and NMBIC, and can host multiple differentiation genetic components [42].

The immune activation community was present in both the NAU and MIBC samples. Results of the CIBERSORT analysis showed that inside the bladder tumor and with increasing stage, most monocytes preferentially differentiate into macrophages with M2 polarization. This is in line with the findings from Chen et al. [44] in which authors analyzed scRNA-seq data of BLCA patients and observed a similar pattern of differentiation for monocytes. In our data, T4 samples had the highest proportion of M2-macrophages, which could partially explain their immunosuppressive state. A novel finding here is the observation that AIF1 appears to drive co-expression in the immune cells of T4 tumors. Interestingly, high AIF1 expression associated with complete response to atezolizumab (**Figure 17D**), a finding which might have implications on patient selection for immunotherapy. Together with the observation that AIF1 is highly expressed in macrophages responding to a M2 stimulation [52], the data suggest that AIF1 expression in the M2-macrophages could potentially trigger a PD-L1 signature in the tumor and the surrounding immune cells, leading to immune suppression [53], but further work is required to validate these preliminary observations.

We specifically searched for genes showing a monotonal trend in their expression level with increasing disease stage, as this property may mark those genes whose quantification could offer additive prognostic value. Out of the 157 identified monotonal genes, almost half of them were components of the cell cycle machinery, or kinases signaling positively for it, or transcription factors responsible for the expression of cell

cycle genes. Eight out of the 48 monotonal genes with prognostic value in the discovery cohort were validated in independent RNA-seq data, and were utilized to develop a sample-wise gene signature. Of these, only ENO1, (higher levels of which had been previously linked with worse BLCA outcome [54]) and CBX7, (downregulation of which was associated with worse survival [55]) were described previously all the remaining 6 associations to survival are novel. MED19, a component of the mediator complex that regulates the transcription of RNA-polymerases, was found overexpressed by IHC in human BLCA compared to normal tissues, and its knockdown in the T24 and 5637 bladder cell lines resulted in cell-cycle arrest at the G0/G1 checkpoint and attenuation of cell growth [56]. The involvement of GTPBP4 in BLCA development has not been characterized, but oncogenic properties have been attributed to this gene in hepatocellular carcinoma [57]. ANLN, AKAP7, and CDC14B are thought to regulate bladder cell growth and apoptosis in a TP53 independent manner [58]. ICA1L is naturally expressed in sperm cells. Its role in BLCA has not been described yet. The CDC14B gene is located on the 9q chromosome, a region that is often deleted in BLCA. This might also explain its overall downregulation in malignancy in comparison to NAU, with additional mechanisms (such as the increasing number of tumor cells), resulting in the observed further downregulation with increasing stage, as observed in the discovery set. CDC14B is believed to dephosphorylate TP53 [59], but the functional consequence on the mitotic or DNA damage repair pathways is not well clarified yet [60]. CBX7 is a component of the chromatin modifier PRC1-complex and is required for the propagation of the transcriptionally repressive state of multiple genes through cell-division, during embryonic development [61], including Hox genes [62]. Expectedly, while CBX7 levels were monotonically decreasing, we noticed that HOXC6 and HOXC9 were both monotonically upregulated with increasing malignancy (Table S1). ZFP2 is a probable transcription factor and evidence suggest an epigenetic role as well [63]. High load of mutations in ZFP36, another member of the ZFP family, were associated with upper tract urothelial carcinoma [64].

Our study has its limitations, including the retrospective nature of the analysis, and restrictions in the validation set imposed by lack of samples from all disease stages, which did not allow validating the observations particularly at the NMIBC. Clinical stage assignment is known to have varying rates of error. However, the high number of

samples used in each of the stage categories is expected to balance out to some extent error while increasing power of the received results. It should also be noted that our scope was to identify, if existent, common 'core' molecular themes during BLCA evolution with high statistical power. This does not rule out the existence of intra-tumoral heterogeneity which still has to be considered (together with the observed molecular changes, as defined in our study) when predicting therapeutic response.

7. CONCLUSIONS

Bladder Cancer subtyping studies at the protein level are scarce and are needed to enhance present findings and fill existing gaps. This study reports on the first proteomics classification of Non-Muscle Invasive Bladder Cancer (NMIBC) based on an unbiased comprehensive LC-MS/MS approach, investigating how existing pathological and molecular subtypes are reproduced in the tissue proteome. Two of the three identified and characterized proteomic subtypes appear to share concordant molecular profiles with the UROMOL study, despite differences in clinicopathologic parameters. The added value of the observed molecular changes to the EORTC risk predictions, remain to be further investigated as, even though classes 1 and 3 segregated in their majority EORTC high and low risk tumors respectively, still in both classes, a number of samples deviated from this general pattern. Cross-omics analysis for features potentially involved in aggressiveness highlighted molecular processes that could likely drive NMIBC subtypes. The derived subtype-specific signatures remain to be validated in the clinical setting. Our study has its limitations. These include the lack of follow-up information and the lack of a side by side analysis at the mRNA level, with the latter having been attempted, but with the mRNA quality found inadequate for a comprehensive study. These shortcomings, collectively, do not allow reaching conclusive statements with respect to the association of the observed protein changes with disease progression; nevertheless, the consistency between the protein and mRNA level changes at two independent cohorts, strongly point in this direction, forming the basis for further validation in prospective cohorts.

8. SUPPLEMENTARY MATERIAL

Supplementary Table 1: Statistically significant Biological Processes of the most abundant proteins (top 200 proteins).

GO ID	GO Term	Benjamini -Hochberg p-value	% Associated genes	Associated Genes Found
GO:0051261	protein depolymerization	0,01	4,95	[CAPG, CFL1, GSN, HSPA8, WDR1]
GO:0061572	actin filament bundle organization	0,00	5,56	[ACTN1, ACTN4, EZR, HSP90B1, MARCKS, PFN1, PFN2, TPM1]
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	0,00	5,43	[RPL13, RPL19, RPL22, RPL29, RPS14, RPS3, RPSA]
GO:0032205	negative regulation of telomere maintenance	0,01	7,69	[HNRNPA1, HNRNPC, HNRNPU]
GO:0051261	protein depolymerization	0,01	4,95	[CAPG, CFL1, GSN, HSPA8, WDR1]
GO:0002274	myeloid leukocyte activation	0,00	5,29	[ALDOA, ALDOC, ANXA2, CAP1, CAPN1, CSTB, CTSD, DHRS2, EEF1A1, EEF2, GPI, GSN, GSTP1, HMGB1, HP, HSP90AA1, HSP90AB1, HSPA1A, HSPA8, HSPD1, IDH1, NME2, PGAM1, PKM, PPIA, PRDX5, PRDX6, RAB10, RAB14, RPSA, S100A11, S100P, TUBB, TUBB4B, VCL, VCP]
GO:0051410	detoxification of nitrogen compound	0,00	60,00	[GSTM1, GSTM2, GSTM3]
GO:0034614	cellular response to reactive oxygen species	0,00	5,39	[AKR1C3, ANXA1, HNRNPD, PRDX1, PRDX2, PRDX5, RPS3, TRAP1, TXN]
GO:1903749	positive regulation of establishment of protein localization to mitochondrion	0,00	5,60	[SFN, YWHAB, YWHAE, YWHAG, YWHAH, YWHAQ, YWHAZ]
GO:0010523	negative regulation of calcium ion transport into cytosol	0,00	13,64	[CALM1, GSTM2, SRI]
GO:0051290	protein heterotetramerization	0,00	11,54	[ANXA2, HIST1H3A, HIST1H4A]
GO:0030968	endoplasmic reticulum unfolded protein response	0,00	6,72	[AGR2, CALR, CASP12, HSP90B1, HSPA1A, HSPA5, LMNA, PDIA6, VCP]
GO:0006984	ER-nucleus signaling pathway	0,00	10,64	[AGR2, CALR, HSP90B1, HSPA5, LMNA]
GO:0035966	response to topologically incorrect protein	0,00	7,58	[AGR2, CALR, CASP12, HSP90AA1, HSP90AB1, HSP90B1, HSPA1A, HSPA5, HSPA8, HSPB1, HSPD1, HSPE1, LMNA, PDIA6, VCP]

GO:0032204	regulation of telomere maintenance	0,00	6,98	[HNRNPA1, HNRNPA2B1, HNRNPC, HNRNPD, HNRNPU, LMNA]
GO:0031424	keratinization	0,00	5,98	[CAPN1, KRT10, KRT13, KRT17, KRT18, KRT19, KRT20, KRT5, KRT6A, KRT7, KRT75, KRT79, KRT8, SFN]
GO:0061077	chaperone-mediated protein folding	0,00	10,14	[CALR, HSPA8, HSPB1, HSPD1, HSPE1, PPIB, TRAP1]
GO:0006413	translational initiation	0,00	4,41	[HSPB1, NPM1, RPL13, RPL19, RPL22, RPL29, RPS14, RPS3, RPSA]
GO:0060048	cardiac muscle contraction	0,01	4,41	[ACTC1, CALM1, FLNA, GSTM2, SRI, TPM1]
GO:0032781	positive regulation of ATPase activity	0,00	7,41	[HNRNPU, PFN1, PFN2, TPM1]
GO:0000380	alternative mRNA splicing, via spliceosome	0,03	5,45	[HNRNPA1, HNRNPL, HNRNPU]
GO:0051290	protein heterotetramerization	0,00	11,54	[ANXA2, HIST1H3A, HIST1H4A]
GO:0001895	retina homeostasis	0,00	6,41	[ACTB, HSPB1, POTEE, POTEF, PRDX1]
GO:0034109	homotypic cell-cell adhesion	0,00	11,39	[ACTB, CD9, CLIC1, FLNA, HSPB1, MYH9, MYL12A, POTEF, VCL]
GO:0006735	NADH regeneration	0,00	25,00	[ALDOA, ALDOC, ENO1, GAPDH, GPI, PGAM1, PGK1, PKM, TPI1]
GO:0007584	response to nutrient	0,00	5,13	[AKR1C3, COL1A1, EEF2, GNAI2, GSN, GSTP1, HNRNPC, LDHA, PKM, VDAC2]
GO:0009168	purine ribonucleoside monophosphate biosynthetic process	0,00	7,89	[ALDOA, ATP5A1, ATP5B, ENO1, PKM, VCP]
GO:0005996	monosaccharide metabolic process	0,00	5,47	[ALDOA, ALDOC, CYB5A, ENO1, GAPDH, GOT2, GPI, HMGB1, KRT17, MDH1, MDH2, PGAM1, PGK1, PKM, TALDO1, TKT, TPI1]
GO:0031532	actin cytoskeleton reorganization	0,00	6,32	[ANXA1, ARHGDIA, EZR, FLNA, GSN, MYH9]
GO:0010954	positive regulation of protein processing	0,00	13,04	[ENO1, GSN, MYH9]
GO:0031532	actin cytoskeleton reorganization	0,00	6,32	[ANXA1, ARHGDIA, EZR, FLNA, GSN, MYH9]
GO:0045104	intermediate filament cytoskeleton organization	0,00	13,33	[DES, KRT17, KRT18, KRT20, KRT6A, VIM]
GO:0008637	apoptotic mitochondrial changes	0,00	8,21	[HSPA1A, HSPD1, LMNA, SFN, SLC25A5, YWHAB, YWHAE, YWHAG, YWHAH, YWHAQ, YWHAZ]
GO:0032612	interleukin-1 production	0,02	4,55	[ANXA1, GSTP1, HMGB1, HSPB1]

GO:0071103	DNA conformation change	0,00	4,74	[ANXA1, H2AFY, H3F3A, HIST1H1B, HIST1H1C, HIST1H2BJ, HIST1H2BK, HIST1H3A, HIST1H4A, HMGB1, HNRNPA2B1, NPM1, UBC]
GO:0060048	cardiac muscle contraction	0,01	4,41	[ACTC1, CALM1, FLNA, GSTM2, SRI, TPM1]
GO:0061077	chaperone-mediated protein folding	0,00	10,14	[CALR, HSPA8, HSPB1, HSPD1, HSPE1, PPIB, TRAP1]
GO:0006103	2-oxoglutarate metabolic process	0,00	13,64	[GOT2, IDH1, IDH2]
GO:0072593	reactive oxygen species metabolic process	0,00	4,35	[AKR1C3, GSTP1, HP, HSP90AA1, HSP90AB1, PRDX1, PRDX2, PRDX5, PRDX6, TRAP1, VDAC1, VDAC2]
GO:0090307	mitotic spindle assembly	0,04	4,76	[FLNA, HNRNPU, HSPA1A]
GO:0032612	interleukin-1 production	0,02	4,55	[ANXA1, GSTP1, HMGB1, HSPB1]
GO:0010927	cellular component assembly involved in morphogenesis	0,00	6,86	[ACTC1, CD9, KRT19, KRT8, MYH11, TPM1, WDR1]
GO:0001649	osteoblast differentiation	0,00	4,67	[ATP5B, CLIC1, CLTC, COL1A1, COL6A1, H3F3A, HNRNPC, HNRNPU, HSPE1, TPM4]
GO:0032637	interleukin-8 production	0,00	6,10	[ANXA1, ANXA4, HSPA1A, RAB1A, RPSA]
GO:0051764	actin crosslink formation	0,00	23,08	[ACTN1, FLNA, MARCKS]
GO:0036500	ATF6-mediated unfolded protein response	0,00	27,27	[CALR, HSP90B1, HSPA5]
GO:1902749	regulation of cell cycle G2/M phase transition	0,00	4,69	[H2AFY, HSP90AA1, NPM1, PSMA1, TUBA4A, TUBB, TUBB4B, UBC, YWHAE, YWHAG]
GO:0036344	platelet morphogenesis	0,00	15,00	[ACTN1, MYH9, WDR1]
GO:0006479	protein methylation	0,00	4,30	[CALM1, EEF1A1, EEF2, H2AFY, HIST1H1B, HIST1H1C, HSPA8, VCP]
GO:0045471	response to ethanol	0,00	4,67	[ACTC1, EEF2, GOT2, GSN, GSTP1, HPGD, TUFM]
GO:0000060	protein import into nucleus, translocation	0,00	8,00	[AKR1C3, HSP90AB1, RAN, TXN]
GO:0032612	interleukin-1 production	0,02	4,55	[ANXA1, GSTP1, HMGB1, HSPB1]
GO:0010803	regulation of tumor necrosis factor-mediated signaling pathway	0,01	6,35	[GSTP1, HIST1H2BJ, HSPA1A, UBC]
GO:0001738	morphogenesis of a polarized epithelium	0,01	4,11	[CLTC, PFN1, PSMA1, RAB10, UBC, WDR1]
GO:0030865	cortical cytoskeleton organization	0,01	8,82	[CALR, EZR, WDR1]

GO:0043277	apoptotic cell clearance	0,01	8,11	[HMGB1, HNRNPC, PDIA6]
GO:0070670	response to interleukin-4	0,01	8,11	[HSP90AB1, HSPA5, TUBA1B]
GO:1901264	carbohydrate derivative transport	0,01	5,41	[AGR2, RPSA, SLC25A5, SLC25A6]
GO:1904019	epithelial cell apoptotic process	0,03	4,12	[AKR1C3, GSN, KRT18, KRT8]
GO:0034381	plasma lipoprotein particle clearance	0,03	5,26	[ANXA2, CLTC, HNRNPK]
GO:0060306	regulation of membrane repolarization	0,01	9,68	[FLNA, WDR1, YWHAE]
GO:0070849	response to epidermal growth factor	0,02	6,38	[COL1A1, EEF1A1, GSTP1]
GO:0007566	embryo implantation	0,03	5,66	[CALR, H3F3A, RPL29]
GO:0033574	response to testosterone	0,03	5,45	[CALR, GPI, NME1]
GO:0034394	protein localization to cell surface	0,03	5,08	[FLNA, HSP90AB1, VCL]
GO:0050819	negative regulation of coagulation	0,03	5,00	[ANXA2, ANXA5, CD9]
GO:2000378	negative regulation of reactive oxygen species metabolic process	0,03	5,00	[HP, TRAP1, VDAC1]
GO:1902305	regulation of sodium ion transmembrane transport	0,04	4,69	[ACTN4, FXYD3, YWHAH]
GO:0006892	post-Golgi vesicle-mediated transport	0,03	4,08	[KRT18, LYPLA1, RAB10, RAB14]
GO:1901616	organic hydroxy compound catabolic process	0,04	4,48	[AKR1C3, MAOA, TPI1]
GO:0045606	positive regulation of epidermal cell differentiation	0,00	12,00	[H2AFY, NME2, SFN]
GO:1902402	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0,05	4,41	[NPM1, SFN, UBC]

Supplementary Table 1: *Statistically significant Biological Processes of the most abundant proteins (top 200 proteins).*

Supplementary Table 2: Gene Set Enrichment Analysis results for the comparison class 1 versus class 3

GSEA Report for Class 1					
NAME	**Size**	**ES**	**NES**	**Nom p-val**	**FDR q-val**
HALLMARK_E2F_TARGETS	19	0,52	1,20	0,00	0,01

HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	24	0,41	1,07	0,10	0,04
HALLMARK_INTERFERON_GAMMA_RESPONSE	16	0,37	1,03	0,02	0,05
HALLMARK_MYC_TARGETS_V1	39	0,41	0,96	0,02	0,08
HALLMARK_G2M_CHECKPOINT	15	0,50	1,02	0,50	0,45
HALLMARK_COMPLEMENT	18	0,33	0,87	0,50	0,78
HALLMARK_MTORC1_SIGNALING	34	0,19	0,59	1,00	1,00
HALLMARK_ADIPOGENESIS	21	0,20	0,56	1,00	1,00
HALLMARK_FATTY_ACID_METABOLISM	22	0,21	0,52	1,00	1,00
HALLMARK_HEME_METABOLISM	16	0,21	0,49	1,00	1,00
HALLMARK_XENOBIOTIC_METABOLISM	28	0,17	0,39	1,00	1,00
GSEA Report for Class 3					
NAME	**SIZE**	**ES**	**NES**	**NOM p-val**	**FDR q-val**
HALLMARK_GLYCOLYSIS	21	-0,35	-1,12	0,03	0,11
HALLMARK_ESTROGEN_RESPONSE_LATE	17	-0,33	-1,10	0,33	1,00
HALLMARK_APOPTOSIS	17	-0,32	-0,92	0,50	1,00
HALLMARK_P53_PATHWAY	15	-0,35	-0,87	0,57	1,00
HALLMARK_APICAL_JUNCTION	16	-0,32	-0,75	0,67	1,00
HALLMARK_OXIDATIVE_PHOSPHORYLATION	25	-0,32	-0,75	1,00	1,00
HALLMARK_MYOGENESIS	15	-0,22	-0,63	1,00	0,96

Supplementary Table 2: *Gene Set Enrichment Analysis results for the comparison class 1 versus class 3 using a robust collection of cancer genesets, the Hallmark Genesets (as found in the MSigDB databse). Significance is defined at nom p <0.05 and FDR <0.25. ES=Enrichment Score, NES=Normalized Enrichment Score, NOM=Nominal, FDR=False Discovery Rate*

Supplementary Table 3: Statistically significant Biological Processes of the 73 upregulated features from the 155 shortlisted proteins

GO ID	GO Term	Benjamini-Hochberg p-value	% Associated Genes	Associated Genes Found
GO:0002181	cytoplasmic translation	0,00	4,55	[EEF2, EIF4G1, RPL13A]

GO:0038128	ERBB2 signaling pathway	0,00	7,50	[CDC37, GRB2, HSP90AA1]
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	0,00	4,44	[ARF5, COPA, COPB1, COPG1]
GO:0042026	protein refolding	0,00	12,50	[HSP90AA1, HSPD1, SNRNP70]
GO:0051131	chaperone-mediated protein complex assembly	0,00	16,67	[HSP90AA1, HSP90AB1, HSPD1]
GO:0006487	protein N-linked glycosylation	0,00	4,00	[DDOST, RPN1, RPN2]
GO:0018196	peptidyl-asparagine modification	0,00	6,25	[DDOST, RPN1, RPN2]
GO:0018279	protein N-linked glycosylation via asparagine	0,00	6,38	[DDOST, RPN1, RPN2]
GO:0043038	amino acid activation	0,00	9,09	[AARS, GARS, TARS, WARS, YARS]
GO:0043039	tRNA aminoacylation	0,00	9,26	[AARS, GARS, TARS, WARS, YARS]
GO:0006418	tRNA aminoacylation for protein translation	0,00	9,80	[AARS, GARS, TARS, WARS, YARS]
GO:0034340	response to type I interferon	0,00	5,49	[CDC37, HLA-A, HSP90AB1, SHMT2, STAT1]
GO:0060330	regulation of response to interferon-gamma	0,00	11,11	[CDC37, HSP90AB1, STAT1]
GO:0071357	cellular response to type I interferon	0,00	4,60	[CDC37, HLA-A, HSP90AB1, STAT1]
GO:0060333	interferon-gamma-mediated signaling pathway	0,00	5,15	[CDC37, HLA-A, HLA-DRA, HSP90AB1, STAT1]
GO:0060337	type I interferon signaling pathway	0,00	4,60	[CDC37, HLA-A, HSP90AB1, STAT1]
GO:0060334	regulation of interferon-gamma-mediated signaling pathway	0,00	11,11	[CDC37, HSP90AB1, STAT1]
GO:0060338	regulation of type I interferon-mediated signaling pathway	0,00	6,82	[CDC37, HSP90AB1, STAT1]
GO:0002200	somatic diversification of immune receptors	0,00	4,00	[HMGB2, HSPD1, TFRC]
GO:0002833	positive regulation of response to biotic stimulus	0,00	6,38	[HSPD1, S100A8, S100A9]
GO:0002702	positive regulation of production of molecular mediator of immune response	0,00	4,40	[HLA-A, S100A8, S100A9, TFRC]
GO:0042116	macrophage activation	0,00	4,23	[HSPD1, S100A8, S100A9]
GO:0050918	positive chemotaxis	0,00	5,56	[GPNMB, HMGB2, S100A8, S100A9]
GO:0043030	regulation of macrophage activation	0,00	6,67	[HSPD1, S100A8, S100A9]
GO:0070671	response to interleukin-12	0,00	5,97	[S100A8, S100A9, SOD2, STAT1]
GO:1901571	fatty acid derivative transport	0,00	5,45	[GPNMB, S100A8, S100A9]
GO:0002720	positive regulation of cytokine production involved in immune response	0,00	6,38	[HLA-A, S100A8, S100A9]
GO:0042100	B cell proliferation	0,00	4,04	[HSPD1, S100A8, S100A9, TFRC]

GO:0035722	interleukin-12-mediated signaling pathway	0,00	6,06	[S100A8, S100A9, SOD2, STAT1]
GO:0071157	negative regulation of cell cycle arrest	0,00	10,34	[HSP90AB1, S100A8, S100A9]
GO:0071349	cellular response to interleukin-12	0,00	6,06	[S100A8, S100A9, SOD2, STAT1]
GO:0002562	somatic diversification of immune receptors via germline recombination within a single locus	0,00	4,84	[HMGB2, HSPD1, TFRC]
GO:0016444	somatic cell DNA recombination	0,00	4,84	[HMGB2, HSPD1, TFRC]
GO:0030888	regulation of B cell proliferation	0,00	4,55	[S100A8, S100A9, TFRC]
GO:0071715	icosanoid transport	0,00	5,45	[GPNMB, S100A8, S100A9]
GO:0030890	positive regulation of B cell proliferation	0,00	6,52	[S100A8, S100A9, TFRC]
GO:0032309	icosanoid secretion	0,00	5,77	[GPNMB, S100A8, S100A9]
GO:0015909	long-chain fatty acid transport	0,00	4,35	[GPNMB, S100A8, S100A9]
GO:1903963	arachidonate transport	0,00	8,57	[GPNMB, S100A8, S100A9]
GO:0050482	arachidonic acid secretion	0,00	8,57	[GPNMB, S100A8, S100A9]

Supplementary Table 3: *Statistically significant Biological Processes of the 73 upregulated features from the 155 shortlisted proteins. The inputted 73 proteins were found at increased abundance in class 1 when compared to both classes 2 & 3 and also at high levels in the MIBC samples when compared to the NMIBC.*

Supplementary Table 4: Statistically significant Biological Processes of the downregulated 82 features from the 155 shortlisted proteins

GO ID	GO Term	Benjamini-Hochberg p-value	% Associated Genes	Associated Genes Found
GO:0042762	regulation of sulfur metabolic process	0,00	11,11	[COMT, CTNNB1, DLD]
GO:0046503	glycerolipid catabolic process	0,00	4,84	[FABP4, FABP5, PRDX5]
GO:0070268	cornification	0,00	7,69	[CAPN1, CAPNS1, IVL, KRT13, KRT17, KRT19, KRT7, KRT75, KRT8]
GO:0030239	myofibril assembly	0,00	4,69	[KRT19, KRT8, WDR1]
GO:0045214	sarcomere organization	0,00	6,82	[KRT19, KRT8, WDR1]

GO:0045682	regulation of epidermis development	0,00	4,76	[CTNNB1, H2AFY, H2AFY2, KRT17]
GO:0045684	positive regulation of epidermis development	0,00	7,89	[H2AFY, H2AFY2, KRT17]
GO:0030858	positive regulation of epithelial cell differentiation	0,00	5,00	[CTNNB1, H2AFY, H2AFY2]
GO:0097306	cellular response to alcohol	0,00	5,00	[CDH1, CTNNB1, GNAS]
GO:0035635	entry of bacterium into host cell	0,00	14,29	[CDH1, CTNNB1, CTNND1]
GO:0045670	regulation of osteoclast differentiation	0,00	4,62	[CTNNB1, FAM213A, GNAS]
GO:0098754	detoxification	0,00	4,63	[FAM213A, GPX2, GSTP1, MGST2, PRDX5]
GO:1990748	cellular detoxification	0,00	4,81	[FAM213A, GPX2, GSTP1, MGST2, PRDX5]
GO:0097237	cellular response to toxic substance	0,00	4,42	[FAM213A, GPX2, GSTP1, MGST2, PRDX5]
GO:0098869	cellular oxidant detoxification	0,00	5,00	[FAM213A, GPX2, GSTP1, MGST2, PRDX5]
GO:0006749	glutathione metabolic process	0,00	5,00	[CNDP2, GSTP1, MGST2]
GO:0042398	cellular modified amino acid biosynthetic process	0,00	7,50	[ATIC, CNDP2, MGST2]
GO:0061621	canonical glycolysis	0,00	13,89	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0061620	glycolytic process through glucose-6-phosphate	0,00	13,51	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0016052	carbohydrate catabolic process	0,00	4,04	[GALE, GNPDA1, HK1, PFKL, PGM2, PKM, PYGL, TPI1]
GO:0044275	cellular carbohydrate catabolic process	0,00	6,25	[PGM2, PYGL, TPI1]
GO:0046365	monosaccharide catabolic process	0,00	9,59	[GALE, GNPDA1, HK1, PFKL, PGM2, PKM, TPI1]
GO:0051156	glucose 6-phosphate metabolic process	0,00	9,76	[GNPDA1, HK1, PGM2, TPI1]
GO:0019320	hexose catabolic process	0,00	11,29	[GALE, GNPDA1, HK1, PFKL, PGM2, PKM, TPI1]
GO:0006007	glucose catabolic process	0,00	11,90	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0019362	pyridine nucleotide metabolic process	0,00	4,06	[DLD, GNPDA1, HK1, PFKL, PGM2, PKM, PRDX5, TPI1]
GO:0006090	pyruvate metabolic process	0,00	4,29	[DLD, GNPDA1, HK1, KRT17, PFKL, PKM, TPI1]
GO:0046496	nicotinamide nucleotide metabolic process	0,00	4,06	[DLD, GNPDA1, HK1, PFKL, PGM2, PKM, PRDX5, TPI1]

GO:0061718	glucose catabolic process to pyruvate	0,00	13,89	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0006757	ATP generation from ADP	0,00	4,31	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0006739	NADP metabolic process	0,00	6,52	[PGM2, PRDX5, TPI1]
GO:0019674	NAD metabolic process	0,00	6,10	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0046031	ADP metabolic process	0,00	4,13	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0006096	glycolytic process	0,00	4,35	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0006735	NADH regeneration	0,00	13,89	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0006734	NADH metabolic process	0,00	10,42	[GNPDA1, HK1, PFKL, PKM, TPI1]
GO:0061615	glycolytic process through fructose-6-phosphate	0,00	13,51	[GNPDA1, HK1, PFKL, PKM, TPI1]

Supplementary Table 4: *Statistically significant Biological Processes of the downregulated 82 features from the 155 shortlisted proteins. The inputted 82 proteins were found at increased abundance in class 3 when compared to both classes 1 & 2 and also at high levels in the NMIBC samples when compared to the MIBC.*

Supplementary Table 5: The 80 genes found in the cell-cycle coexpression networks in all BLCA stages.

Gene	scoreTa	scoreT1	scoreT2	scoreT3	scoreT4	is monotonal
ANLN	481	436	714	18041	2	TRUE
AURKA	5404	6800	13240	8938	605	TRUE
CCNA2	2909	4486	11632	41245	16543	TRUE
CCNB1	707	108	942	84561	18737	TRUE
CDC20	1501	7897	9	17365	25679	TRUE
CDCA5	44873	11804	49730	6700	13955	TRUE
CDCA8	8870	12299	3542	18378	11613	TRUE
CDKN3	24	992	13434	10727	7293	TRUE
CENPA	2166	1644	930	8645	9702	TRUE
CENPN	18086	8239	3048	1671	554	TRUE
CEP55	681	11627	2791	2962	11710	TRUE
DTL	26	10	13590	42	1957	TRUE
EXO1	1587	814	1485	3586	9030	TRUE
FBXO5	0	39	4245	13659	5116	TRUE
GINS2	83	0	77542	1333	6469	TRUE
KIF11	698	1216	1523	2433	20747	TRUE
KIF14	520	1523	2570	14128	158551	TRUE

KIF23	64	2203	6453	7830	3980	TRUE
MCM10	1233	854	11193	13	7153	TRUE
MELK	3408	2009	25349	10979	8132	TRUE
MKI67	523	892	415	78	25	TRUE
MND1	0	1	17	318	994	TRUE
NDC80	100	92	1499	4424	1087	TRUE
OIP5	10	477	1	926	19138	TRUE
PBK	34	1	112	8507	5366	TRUE
POLE2	4	16	0	2	120	TRUE
PRC1	1750	12369	38237	5498	1375	TRUE
PTTG1	2595	716	13893	5097	288	TRUE
RAD51AP1	32	8881	31700	34174	1528	TRUE
RRM2	19	93	242	78192	69546	TRUE
UBE2T	14	0	1777	228	0	TRUE
UHRF1	1271	2263	33179	1495	975	TRUE
E2F7	0	0	3192	729	1510	TRUE
MCM4	0	0	10	5	19167	TRUE
PLK1	1253	20	943	1059	1501	TRUE
ASF1B	4864	350	10948	40	1722	FALSE
ASPM	920	1611	425	13224	2009	FALSE
AURKB	2035	95451	51815	1602	8950	FALSE
BIRC5	182	568	625	5479	12397	FALSE
BLM	14	6	10324	0	3	FALSE
BRCA1	45	26	3809	867	13252	FALSE
BUB1	240	9179	7065	4193	8712	FALSE
BUB1B	40	276	76617	704	3140	FALSE
CCNB2	1474	2930	8846	2453	21761	FALSE
CCNE2	95	1950	4682	59	22	FALSE
CDC25C	3	0	2	1625	10052	FALSE
CDCA2	21	0	4	1722	3658	FALSE
CENPE	66	1967	1035	1223	917	FALSE
CENPF	978	347	47080	12836	50	FALSE
CENPK	9	9	884	1391	8	FALSE
CHAF1B	1081	4358	3468	1153	19537	FALSE
CHEK1	733	1492	29879	73	81	FALSE
EZH2	11	3630	12341	4918	20668	FALSE
FANCD2	3558	1118	62875	8	6881	FALSE
FOXM1	2013	248692	4913	5527	83315	FALSE
HMMR	7	15	0	11836	3381	FALSE
KIF15	46	206	6581	13537	19776	FALSE
KIF20A	12487	8816	300	1145	807	FALSE
KIF2C	44734	6320	3475	5554	8315	FALSE
KIF4A	1673	6304	2669	1660	27717	FALSE
KNTC1	5467	777	35820	0	7	FALSE

MCM2	5158	20158	139586	544	51316	FALSE
NCAPG	410	502	29290	27683	24269	FALSE
NEK2	132	870	0	6945	4161	FALSE
NUF2	17	151	34939	3446	19288	FALSE
NUSAP1	532	3679	5554	14397	5418	FALSE
PLK4	541	769	867	248	1	FALSE
POLQ	579	434	0	64	73	FALSE
RAD54L	1253	2823	5983	7412	14193	FALSE
STIL	3218	461	4818	3843	25471	FALSE
TACC3	1406	35712	0	0	4127	FALSE
TOP2A	1881	1241	27686	2363	3193	FALSE
TPX2	8357	6812	92143	26014	7312	FALSE
TRIP13	2505	1366	9669	2698	4488	FALSE
TTK	4892	455	4258	5663	2694	FALSE
UBE2C	16280	1191	668	1260	1805	FALSE
ZWINT	2436	7481	1589	145	5402	FALSE
CENPL	0	0	16	2812	7136	FALSE
NCAPD2	0	0	0	0	7160	FALSE
RACGAP1	941	37	2199	19570	3598	FALSE

Supplementary Table 5: *The 80 genes found in the cell-cycle coexpression networks in all BLCA stages, along with their stage specific betweenness centrality scores and an indication on whether their expression follows a monotonal increase with increasing BLCA stage.*

Supplementary Table 6: The 157 genes whose expression is monotonically increasing or decreasing with increasing BLCA stage.

Gene symbol	Regulation with increasing stage	Gene symbol	Regulation with increasing stage
ADHFE1	Down	KIT	Down
AEBP2	Down	LMBRD1	Down
AKAP7	Down	LMNB2	Up
ALDH7A1	Down	LONRF1	Down
ANLN	Up	LRP8	Up
APCDD1	Down	LYAR	Up
ARF5	Up	MCM10	Up
ARHGAP11A	Up	MCM4	Up
ARHGDIA	Up	MED19	Up
ARPC5L	Up	MELK	Up
ASS1	Down	MGST1	Down
AURKA	Up	MKI67	Up

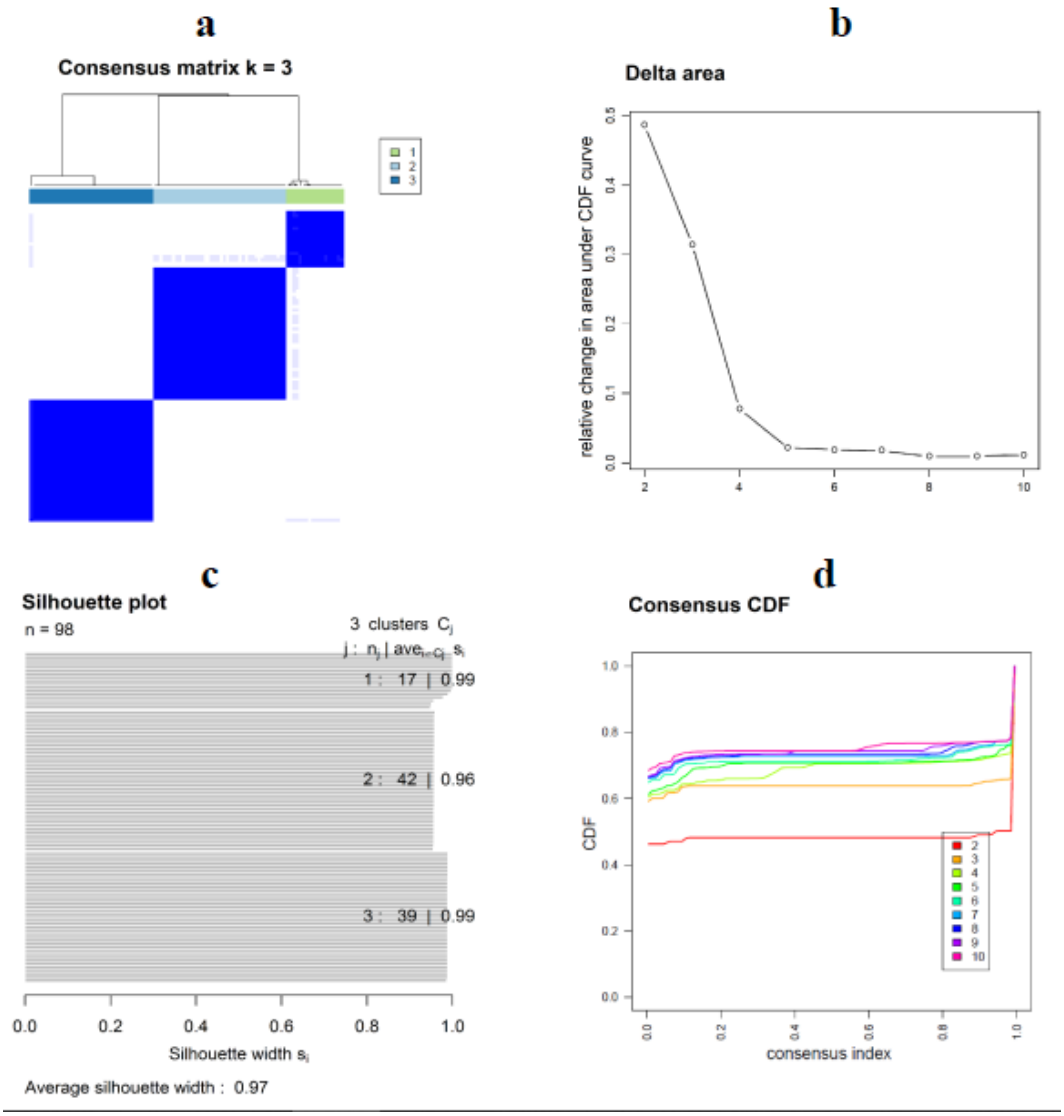
BOP1	Up	MLF2	Up
BTG2	Down	MND1	Up
BUD31	Up	MRPL37	Up
CAD	Up	MTMR9	Down
CAT	Down	MYO10	Up
CBX7	Down	NDC80	Up
CCDC124	Up	NEIL3	Up
CCNA2	Up	NENF	Up
CCNB1	Up	NHS	Down
CCT5	Up	NIP7	Up
CDC14B	Down	NMU	Up
CDC20	Up	NNAT	Down
CDCA3	Up	NOX4	Up
CDCA5	Up	NTHL1	Up
CDCA8	Up	NUDT1	Up
CDK2AP2	Up	NXT1	Up
CDKN3	Up	OIP5	Up
CENPA	Up	PACSIN3	Up
CENPN	Up	PAQR4	Up
CEP55	Up	PBK	Up
CEP72	Up	PIGR	Down
CHST9	Down	PLK1	Up
CIRBP	Down	POLD1	Up
CLPTM1L	Up	POLE2	Up
CTSA	Up	POLR2D	Up
CTSO	Down	POP7	Up
CYP4V2	Down	PPP1R14C	Up
DAGLA	Up	PPP1R1B	Down
DAXX	Up	PRC1	Up
DEPDC1	Up	PSMB3	Up
DET1	Down	PSMC1	Up
DNAJC9	Up	PTCH1	Down
DTL	Up	PTTG1	Up
E2F1	Up	RAD51AP1	Up
E2F7	Up	RAE1	Up
EIF4EBP1	Up	RBM28	Up
ELK1	Up	RECQL4	Up
ENO1	Up	RP9	Up
EXO1	Up	RRM2	Up
FAM50A	Up	RSL1D1	Down
FAM91A1	Up	RUVBL2	Up
FBXO5	Up	RWDD3	Down
FBXW2	Up	SAC3D1	Up
FOXA3	Up	SBSN	Up

FYCO1	Down	SLC22A15	Up
GALNT12	Down	SLC25A13	Up
GARNL3	Down	SLC25A27	Down
GBP6	Up	SLC26A6	Up
GINS2	Up	SMC4	Up
GP6	Up	SNX8	Up
GTPBP4	Up	SPP1	Up
HOXC6	Up	SRD5A1	Up
HOXC9	Up	SRPRB	Up
HTR2C	Up	TAPT1	Down
ICA1L	Down	TEAD4	Up
IFNAR1	Down	TRAF2	Up
IGFBP2	Down	TRIM65	Up
IGFL2	Up	TUBB	Up
IMPDH1	Up	UBE2T	Up
ISG15	Up	UHRF1	Up
ITM2C	Down	UST	Down
ITPR3	Up	WNT2	Up
KIAA2013	Up	XPO5	Up
KIF11	Up	YIF1A	Up
KIF14	Up	ZFP2	Down
KIF23	Up	ZNF181	Down
KIFC1	Up		

Supplementary Table 6*: The 157 genes whose expression is monotonically increasing or decreasing with increasing BLCA stage.*

Supplementary Table 7: Genes with prognostic value validated in TCGA data.

	Discovery		TCGA	
Gene	HR (95% CI for HR)	p.value	HR (95% CI for HR)	p.value
CBX7	0.6 (0.42-0.86)	0.0046	0.59 (0.43-0.81)	0.0013
ZFP2	0.73 (0.61-0.89)	0.0013	0.67 (0.52-0.86)	0.0017
AKAP7	0.66 (0.49-0.87)	0.0041	0.72 (0.57-0.92)	0.0086
MED19	2.2 (1.3-3.9)	0.0052	2.4 (1.2-4.6)	0.011
ENO1	1.9 (1.3-2.9)	0.0019	1.7 (1.1-2.6)	0.016
ANLN	1.3 (1.1-1.6)	0.0038	1.4 (1-1.8)	0.025
GTPBP4	1.7 (1.1-2.4)	0.009	1.9 (1.1-3.3)	0.029
CDC14B	0.59 (0.43-0.81)	0.0013	0.66 (0.43-0.99)	0.044



Supplementary Figure 1: Clustering performance of the 20% protein frequency NMIBC dataset, for k = 3 clusters.

(a) Consensus matrix heatmap of the three clusters showing size, boundaries and their classification tree. Values range from 0 = no correlation (white), to 1 = perfect correlation (blue).

(b) Delta plot showing the relative change in the area under the CDF curve. More pronounced change is observed from k = 3 to 4 classes, indicating the existence of three clusters in the proteomics dataset.

(c) Silhouette plot displaying the within cluster consistency of the output. Silhouette width informs for the relatedness of a sample to its own cluster. Values near 1 indicate a well matched case, whereas those near 0 indicate the opposite.

(d) Cumulative distribution function (CDF) plot showing the repeatability of item co-clustering across iterations (consensus index 1.0 = co-clustered in 100% of the iterations) for the different k- solutions. Curves with gains near 0 and 1, that also deliver greater area under them, exhibit the most stable clustering performance

9. REFERENCES

1. Yang YM, Chang JW. Bladder cancer initiating cells (BCICs) are among EMA-CD44v6+ subset: novel methods for isolating undetermined cancer stem (initiating) cells. Cancer investigation. 2008;26(7):725-33. Epub 2008/07/09.

2. Bryan RT. Cell adhesion and urothelial bladder cancer: the role of cadherin switching and related phenomena. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2015;370(1661):20140042. Epub 2014/12/24.

3. Collaborators GBDRF, Forouzanfar MH, Alexander L, Anderson HR, Bachman VF, Biryukov S, Brauer M, Burnett R, Casey D, Coates MM, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;386(10010):2287-323.

4. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: a cancer journal for clinicians. 2018;68(1):7-30.

5. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.

6. Leal J, Luengo-Fernandez R, Sullivan R, Witjes JA. Economic Burden of Bladder Cancer Across the European Union. European urology. 2016;69(3):438-47.

7. Wong MCS, Fung FDH, Leung C, Cheung WWL, Goggins WB, Ng CF. The global epidemiology of bladder cancer: a joinpoint regression analysis of its incidence and mortality trends and projection. Scientific reports. 2018;8(1):1129.

8. Breau RH, Karnes RJ, Farmer SA, Thapa P, Cagiannos I, Morash C, Frank I. Progression to detrusor muscle invasion during urothelial carcinoma surveillance is associated with poor prognosis. BJU international. 2014;113(6):900-6.

9. Park J, Moon K. Tumor, Nodes, Metastases (TNM) Classification System for Bladder Cancer. Ku J, editor: Academic Press; 2018.

10. Kassouf W, Spiess PE, Siefker-Radtke A, Swanson D, Grossman HB, Kamat AM, Munsell MF, Guo CC, Czerniak BA, Dinney CP. Outcome and patterns of recurrence of nonbilharzial pure squamous cell carcinoma of the bladder: a contemporary review of The University of Texas M D Anderson Cancer Center experience. Cancer. 2007;110(4):764-9.

11. Spiess PE, Kassouf W, Steinberg JR, Tuziak T, Hernandez M, Tibbs RF, Czerniak B, Kamat AM, Dinney CP, Grossman HB. Review of the M.D. Anderson experience in the treatment of bladder sarcoma. Urologic oncology. 2007;25(1):38-45.

12. Helpap B. Nonepithelial neoplasms of the urinary bladder. Virchows Archiv : an international journal of pathology. 2001;439(4):497-503.

13. Klaile Y, Schlack K, Boegemann M, Steinestel J, Schrader AJ, Krabbe LM. Variant histology in bladder cancer: how it should change the management in non-muscle invasive and muscle invasive disease? Translational andrology and urology. 2016;5(5):692-701.

14. Sung MT, Wang M, MacLennan GT, Eble JN, Tan PH, Lopez-Beltran A, Montironi R, Harris JJ, Kuhar M, Cheng L. Histogenesis of sarcomatoid urothelial carcinoma of the urinary bladder: evidence for a common clonal origin with divergent differentiation. The Journal of pathology. 2007;211(4):420-30.

15. Armstrong AB, Wang M, Eble JN, MacLennan GT, Montironi R, Tan PH, Lopez-Beltran A, Zhang S, Baldridge LA, Spartz H, et al. TP53 mutational analysis supports monoclonal origin of biphasic sarcomatoid urothelial carcinoma (carcinosarcoma) of the urinary bladder. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc. 2009;22(1):113-8.

16. Sauter G, Algaba F, Amin MB, Busch C, Cheville J, Gasser T, Grignon DJ, Hofstädter F, Lopez-Beltran A, Epstein JI. Non-invasive urothelial tumours. In: Eble JN, Sauter G, Epstein JI, Sesterhenn IA, editors. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs. Lyon: IARCPress; 2004.

17. Mostofi F, Sobin L, Torloni H. Histological Typing of Urinary Bladder Tumours. International Histological Classification of Tumours. Geneva: World Health Organisation; 1973.

18. Cao D, Vollmer RT, Luly J, Jain S, Roytman TM, Ferris CW, Hudson MA. Comparison of 2004 and 1973 World Health Organization grading systems and their relationship to pathologic staging for predicting long-term prognosis in patients with urothelial carcinoma. Urology. 2010;76(3):593-9.

19. Chen Z, Ding W, Xu K, Tan J, Sun C, Gou Y, Tong S, Xia G, Fang Z, Ding Q. The 1973 WHO Classification is more suitable than the 2004 WHO Classification for predicting prognosis in non-muscle-invasive bladder cancer. PloS one. 2012;7(10):e47199.

20. Mostafa MH, Sheweita SA, O'Connor PJ. Relationship between schistosomiasis and bladder cancer. Clinical microbiology reviews. 1999;12(1):97-111.

21. Sadow CA, Silverman SG, O'Leary MP, Signorovitch JE. Bladder cancer detection with CT urography in an Academic Medical Center. Radiology. 2008;249(1):195-202.

22. Chen GL, El-Gabry EA, Bagley DH. Surveillance of upper urinary tract transitional cell carcinoma: the role of ureteroscopy, retrograde pyelography, cytology and urinalysis. The Journal of urology. 2000;164(6):1901-4.

23. Sanli O, Dobruch J, Knowles MA, Burger M, Alemozaffar M, Nielsen ME, Lotan Y. Bladder cancer. Nature reviews Disease primers. 2017;3:17022.

24. Seront E, Machiels JP. Molecular biology and targeted therapies for urothelial carcinoma. Cancer treatment reviews. 2015;41(4):341-53.

25. Ohishi T, Koga F, Migita T. Bladder Cancer Stem-Like Cells: Their Origin and Therapeutic Perspectives. International journal of molecular sciences. 2015;17(1).

26. Amorim GL, Veloso DF, Vieira JC, Alves PR. Molecular aspects of bladder cancer. Einstein. 2011;9(1):95-9.

27. Knowles MA, Hurst CD. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. Nature reviews Cancer. 2015;15(1):25-41.

28. Sjodahl G, Lovgren K, Lauss M, Patschan O, Gudjonsson S, Chebil G, Aine M, Eriksson P, Mansson W, Lindgren D, et al. Toward a molecular pathologic classification of urothelial carcinoma. The American journal of pathology. 2013;183(3):681-91.

29. Hurst CD, Platt FM, Taylor CF, Knowles MA. Novel tumor subgroups of urothelial carcinoma of the bladder defined by integrated genomic analysis. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012;18(21):5865-77.

30. Allory Y, Beukers W, Sagrera A, Flandez M, Marques M, Marquez M, van der Keur KA, Dyrskjot L, Lurkin I, Vermeij M, et al. Telomerase reverse transcriptase promoter mutations in bladder cancer: high frequency across stages, detection in urine, and lack of association with outcome. European urology. 2014;65(2):360-6.

31. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008;5(7):621-8. Epub 2008/06/03.

32. Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, Stefano GB. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. Medical science monitor basic research. 2014;20:138-42. Epub 2014/08/26.

33. Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L. Next-generation sequencing and its applications in molecular diagnostics. Expert review of molecular diagnostics. 2011;11(3):333-43. Epub 2011/04/06.

34. Ergin S, Kherad N, Alagoz M. RNA sequencing and its applications in cancer and rare diseases. Molecular biology reports. 2022;49(3):2325-33. Epub 2022/01/07.

35. Feng H, Qin Z, Zhang X. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. Cancer letters. 2013;340(2):179-91. Epub 2012/12/01.

36. McGettigan PA. Transcriptomics in the RNA-seq era. Current opinion in chemical biology. 2013;17(1):4-11. Epub 2013/01/08.

37. Dominissini D, Moshitch-Moshkovitz S, Amariglio N, Rechavi G. Adenosine-to-inosine RNA editing meets cancer. Carcinogenesis. 2011;32(11):1569-77. Epub 2011/07/01.

38. Sanchez-Pla A, Reverter F, Ruiz de Villa MC, Comabella M. Transcriptomics: mRNA and alternative splicing. Journal of neuroimmunology. 2012;248(1-2):23-31. Epub 2012/05/26.

39. Christen R. Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. Microbes and environments. 2008;23(4):253-68. Epub 2008/01/01.

40. Shyr D, Liu Q. Next generation sequencing in cancer research and clinical application. Biological procedures online. 2013;15(1):4. Epub 2013/02/15.

41. Li X, Nair A, Wang S, Wang L. Quality control of RNA-seq experiments. Methods Mol Biol. 2015;1269:137-46. Epub 2015/01/13.

42. Au KF. Accurate mapping of RNA-Seq data. Methods Mol Biol. 2015;1269:147-61. Epub 2015/01/13.

43. Petric RC, Pop LA, Jurj A, Raduly L, Dumitrascu D, Dragos N, Neagoe IB. Next generation sequencing applications for breast cancer research. Clujul Med. 2015;88(3):278-87. Epub 2015/11/27.

44. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics. 2013;14(6):671-83. Epub 2012/09/19.

45. Devonshire AS, Sanders R, Wilkes TM, Taylor MS, Foy CA, Huggett JF. Application of next generation qPCR and sequencing platforms to mRNA biomarker analysis. Methods. 2013;59(1):89-100. Epub 2012/07/31.

46. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13(2):204-16. Epub 2012/01/31.

47. Chaussabel D. Assessment of immune status using blood transcriptomics and potential implications for global health. Seminars in immunology. 2015;27(1):58-66. Epub 2015/04/01.

48. Pappireddi N, Martin L, Wuhr M. A Review on Quantitative Multiplexed Proteomics. Chembiochem : a European journal of chemical biology. 2019;20(10):1210-24. Epub 2019/01/05.

49. Anand S, Samuel M, Ang CS, Keerthikumar S, Mathivanan S. Label-Based and Label-Free Strategies for Protein Quantitation. Methods Mol Biol. 2017;1549:31-43. Epub 2016/12/16.

50. Chen X, Wei S, Ji Y, Guo X, Yang F. Quantitative proteomics using SILAC: Principles, applications, and developments. Proteomics. 2015;15(18):3175-92. Epub 2015/06/23.

51. Bohnenberger H, Strobel P, Mohr S, Corso J, Berg T, Urlaub H, Lenz C, Serve H, Oellerich T. Quantitative mass spectrometric profiling of cancer-cell proteomes derived from liquid and solid tumors. Journal of visualized experiments : JoVE. 2015(96):e52435. Epub 2015/04/14.

52. Shenoy A, Geiger T. Super-SILAC: current trends and future perspectives. Expert review of proteomics. 2015;12(1):13-9. Epub 2014/11/19.

53. Brown KA, Melby JA, Roberts DS, Ge Y. Top-down proteomics: challenges, innovations, and applications in basic and clinical research. Expert review of proteomics. 2020;17(10):719-33. Epub 2020/11/25.

54. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nature biotechnology. 2007;25(1):117-24. Epub 2006/12/26.

55. Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. BMC bioinformatics. 2012;13 Suppl 16:S5. Epub 2012/11/28.

56. Valikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. Briefings in bioinformatics. 2018;19(1):1-11. Epub 2016/10/04.

57. Dubois E, Galindo AN, Dayon L, Cominetti O. Assessing normalization methods in mass spectrometry-based proteome profiling of clinical samples. Bio Systems. 2022;215-216:104661. Epub 2022/03/06.

58. Latosinska A, Vougas K, Makridakis M, Klein J, Mullen W, Abbas M, Stravodimos K, Katafigiotis I, Merseburger AS, Zoidakis J, et al. Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis. PloS one. 2015;10(9):e0137048. Epub 2015/09/04.

59. Sjodin MO, Wetterhall M, Kultima K, Artemenko K. Comparative study of label and label-free techniques using shotgun proteomics for relative protein quantification. Journal of chromatography B, Analytical technologies in the biomedical and life sciences. 2013;928:83-92. Epub 2013/04/24.

60. Trinh HV, Grossmann J, Gehrig P, Roschitzki B, Schlapbach R, Greber UF, Hemmi S. iTRAQ-Based and Label-Free Proteomics Approaches for Studies of Human Adenovirus Infections. International journal of proteomics. 2013;2013:581862. Epub 2013/04/05.

61. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173(2):291-304 e6.

62. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell. 2018;173(2):305-20 e10.

63. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321-37 e10.

64. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al. The Immune Landscape of Cancer. Immunity. 2018;48(4):812-30 e14.

65. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell. 2018;173(2):355-70 e14.

66. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;174(4):1034-5.

67. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC bioinformatics. 2016;17 Suppl 2:15.

68. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in genetics. 2017;8:84.

69. Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics. 2011;27(13):i401-9.

70. Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic acids research. 2012;40(19):9379-91.

71. Fujita N, Mizuarai S, Murakami K, Nakai K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Scientific reports. 2018;8(1):9743.

72. Ray B, Liu W, Fenyo D. Adaptive Multiview Nonnegative Matrix Factorization Algorithm for Integration of Multimodal Biomedical Data. Cancer informatics. 2017;16:1176935117725727.

73. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PloS one. 2017;12(5):e0176278.

74. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(11):4245-50.

75. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2016;32(1):1-8.

76. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and Individual Variation Explained (Jive) for Integrated Analysis of Multiple Data Types. The annals of applied statistics. 2013;7(1):523-42.

77. O'Connell MJ, Lock EF. R.JIVE for exploration of multi-source molecular data. Bioinformatics. 2016;32(18):2877-9.

78. De Roover K, Timmerman ME, Mesquita B, Ceulemans E. Common and cluster-specific simultaneous component analysis. PloS one. 2013;8(5):e62280.

79. Zhou G, Cichocki A, Zhang Y, Mandic DP. Group Component Analysis for Multiblock Data: Common and Individual Feature Extraction. IEEE transactions on neural networks and learning systems. 2016;27(11):2426-39.

80. Schouteden M, Van Deun K, Wilderjans TF, Van Mechelen I. Performing DISCO-SCA to search for distinctive and common information in linked data. Behavior research methods. 2014;46(2):576-87.

81. Ray P, Zheng L, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. Bioinformatics. 2014;30(10):1370-6.

82. Ghahramani Z, Griffiths TL. Infinite latent feature models and the Indian buffet process. Vancouver2005.

83. Thibaux R, Jordan MI. Hierarchical beta processes and the indian buffet process. San Juan: AISTATS; 2006.

84. Shen R, Wang S, Mo Q. Sparse Integrative Clustering of Multiple Omics Data Sets. The annals of applied statistics. 2013;7(1):269-94.

85. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28(24):3290-7.

86. Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. PLoS computational biology. 2011;7(10):e1002227.

87. Savage RS, Ghahramani Z, Griffin JE, de la Cruz BJ, Wild DL. Discovering transcriptional modules by Bayesian data integration. Bioinformatics. 2010;26(12):i158-67.

88. Gabasova E, Reid J, Wernisch L. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLoS computational biology. 2017;13(10):e1005781.

89. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237-45.

90. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic acids research. 2009;37(Database issue):D674-9.

91. Wei W, Sun Z, da Silveira WA, Yu Z, Lawson A, Hardiman G, Kelemen LE, Chung D. Semi-supervised identification of cancer subgroups using survival outcomes and overlapping grouping information. Statistical methods in medical research. 2018:962280217752980.

92. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nature methods. 2014;11(3):333-7.

93. Ma T, Zhang A. Affinity network fusion and semi-supervised learning for cancer patient clustering. Methods. 2018;145:16-24.

94. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. An integrated approach to uncover drivers of cancer. Cell. 2010;143(6):1005-17.

95. Kim D, Li R, Dudek SM, Ritchie MD. ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. BioData mining. 2013;6(1):23.

96. Ritchie MD, Motsinger AA, Bush WS, Coffey CS, Moore JH. Genetic Programming Neural Networks: A Powerful Bioinformatics Tool for Human Genetics. Applied soft computing. 2007;7(1):471-9.

97. Jeong HH, Leem S, Wee K, Sohn KA. Integrative network analysis for survival-associated gene-gene interactions across multiple genomic profiles in ovarian cancer. Journal of ovarian research. 2015;8:42.

98. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics. 2015;31(12):i268-75.

99. He X, Niyogi P. Locality preserving projections. Thrun S, editor. Cambridge: MIT Press; 2004. 153–60 p.

100. Liao L, Li K, Li K, Yang C, Tian Q. A multiple kernel density clustering algorithm for incomplete datasets in bioinformatics. BMC systems biology. 2018;12(Suppl 6):111.

101. E. Hinton G. Visualizing High-Dimensional Data Using t-SNE2008. 2579-605 p.

102. Liu FT, Ting KM, Zhou Z-H. Isolation-Based Anomaly Detection. ACM Trans Knowl Discov Data. 2012;6(1):1-39.

103. Seoane JA, Day IN, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2014;30(6):838-45.

104. Guo Y, Zheng J, Shang X, Li Z. A Similarity Regression Fusion Model for Integrating Multi-Omics Data to Identify Cancer Subtypes. Genes. 2018;9(7).

105. Hu R, Qiu X, Glazko G, Klebanov L, Yakovlev A. Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. BMC bioinformatics. 2009;10:20.

106. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics. 2011;27(6):887-8.

107. Aure MR, Steinfeld I, Baumbusch LO, Liestol K, Lipson D, Nyberg S, Naume B, Sahlberg KK, Kristensen VN, Borresen-Dale AL, et al. Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. PloS one. 2013;8(1):e53014.

108. Chari R, Coe BP, Vucic EA, Lockwood WW, Lam WL. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. BMC systems biology. 2010;4:67.

109. Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomaki V, Valo E, Nunez-Fontarnau J, Rantanen V, Karinen S, et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome medicine. 2010;2(9):65.

110. Damrauer JS, Hoadley KA, Chism DD, Fan C, Tiganelli CJ, Wobker SE, Yeh JJ, Milowsky MI, Iyer G, Parker JS, et al. Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(8):3110-5.

111. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, Roth B, Cheng T, Tran M, Lee IL, et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. Cancer cell. 2014;25(2):152-65.

112. Sjodahl G, Lauss M, Lovgren K, Chebil G, Gudjonsson S, Veerla S, Patschan O, Aine M, Ferno M, Ringner M, et al. A molecular taxonomy for urothelial carcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012;18(12):3377-86.

113. Cancer Genome Atlas Research N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315-22.

114. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, Hinoue T, Laird PW, Hoadley KA, Akbani R, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell. 2017;171(3):540-56 e25.

115. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.

116. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929-44.

117. Tan TZ, Rouanne M, Tan KT, Huang RY, Thiery JP. Molecular Subtypes of Urothelial Bladder Cancer: Results from a Meta-cohort Analysis of 2411 Tumors. European urology. 2018.

118. Hedegaard J, Lamy P, Nordentoft I, Algaba F, Hoyer S, Ulhoi BP, Vang S, Reinert T, Hermann GG, Mogensen K, et al. Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. Cancer cell. 2016;30(1):27-42.

119. Hurst CD, Alder O, Platt FM, Droop A, Stead LF, Burns JE, Burghel GJ, Jain S, Klimczak LJ, 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(5):701-15 e7.

120. Aine M, Eriksson P, Liedberg F, Sjodahl G, Hoglund M. Biological determinants of bladder cancer gene expression subtypes. Scientific reports. 2015;5:10957-69.

121. Lerner SP, McConkey DJ, Hoadley KA, Chan KS, Kim WY, Radvanyi F, Hoglund M, Real FX. Bladder Cancer Molecular Taxonomy: Summary from a Consensus Meeting. Bladder cancer. 2016;2(1):37-47.

122. Sjodahl G, Eriksson P, Liedberg F, Hoglund M. Molecular classification of urothelial carcinoma: global mRNA classification versus tumour-cell phenotype classification. The Journal of pathology. 2017;242(1):113-25.

123. 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):3737-48.

124. de Velasco G, Trilla-Fuertes L, Gamez-Pozo A, Urbanowicz M, Ruiz-Ares G, Sepulveda JM, Prado-Vazquez G, Arevalillo JM, Zapater-Moros A, Navarro H, et al. Urothelial cancer proteomics provides both prognostic and functional information. Scientific reports. 2017;7(1):15819.

125. Sobin LH, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours. 7th ed: Wiley-Blackwell; 2009.

126. Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nature methods. 2009;6(5):359-62.

127. Latosinska A, Vougas K, Makridakis M, Klein J, Mullen W, Abbas M, Stravodimos K, Katafigiotis I, Merseburger AS, Zoidakis J, et al. Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis. PloS one. 2015;10(9):e0137048-72.

128. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572-3.

129. Dyrskjot L, Reinert T, Algaba F, Christensen E, Nieboer D, Hermann GG, Mogensen K, Beukers W, Marquez M, Segersten U, et al. Prognostic Impact of a 12-gene Progression Score in Non-muscle-invasive Bladder Cancer: A Prospective Multicentre Validation Study. European urology. 2017;72(3):461-9.

130. Dadhania V, Zhang M, Zhang L, Bondaruk J, Majewski T, Siefker-Radtke A, Guo CC, Dinney C, Cogdell DE, 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;12:105-17.

131. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Perez C, Lopez-Bigas N, Kamoun A, Neuzillet Y, Gestraud P, et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell reports. 2014;9(4):1235-45.

132. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications. 2013;4:2612-22.

133. Dyrskjot L, Kruhoffer M, Thykjaer T, Marcussen N, Jensen JL, Moller K, Orntoft TF. Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. Cancer research. 2004;64(11):4040-8.

134. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091-3.

135. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems. 2015;1(6):417-25.

136. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545-50.

137. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.

138. Latosinska A, Mokou M, Makridakis M, Mullen W, Zoidakis J, Lygirou V, Frantzi M, Katafigiotis I, Stravodimos K, Hupe MC, et al. Proteomics analysis of bladder cancer invasion: Targeting EIF3D for therapeutic intervention. Oncotarget. 2017;8(41):69435-55.

139. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings in bioinformatics. 2013;14(3):315-26.

140. Angelov D, Molla A, Perche PY, Hans F, Cote J, Khochbin S, Bouvet P, Dimitrov S. The histone variant macroH2A interferes with transcription factor binding and SWI/SNF nucleosome remodeling. Molecular cell. 2003;11(4):1033-41.

141. Doyen CM, An W, Angelov D, Bondarenko V, Mietton F, Studitsky VM, Hamiche A, Roeder RG, Bouvet P, Dimitrov S. Mechanism of polymerase II transcription repression by the histone variant macroH2A. Molecular and cellular biology. 2006;26(3):1156-64.

142. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature reviews Genetics. 2010;11(10):733-9. Epub 2010/09/15.

143. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences of the United States of America. 2000;97(18):10101-6. Epub 2000/08/30.

144. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS. Adjustment of systematic microarray data biases. Bioinformatics. 2004;20(1):105-14. Epub 2003/12/25.

145. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27. Epub 2006/04/25.

146. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics. 2020;2(3):lqaa078. Epub 2020/10/06.

147. Ju JH, Shenoy SA, Crystal RG, Mezey JG. An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci. PLoS computational biology. 2017;13(5):e1005537. Epub 2017/05/16.

148. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539-52. Epub 2011/11/22.

149. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology. 2014;32(9):896-902. Epub 2014/08/26.

150. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends in genetics : TIG. 2003;19(7):362-5. Epub 2003/07/10.

151. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends in genetics : TIG. 2013;29(10):569-74. Epub 2013/07/03.

152. Pihur V, Datta S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. Bioinformatics. 2007;23(13):1607-15. Epub 2007/05/08.

153. Weishaupt H, Johansson P, Sundstrom A, Lubovac-Pilav Z, Olsson B, Nelander S, Swartling FJ. Batch-normalization of cerebellar and medulloblastoma gene expression datasets utilizing empirically defined negative control genes. Bioinformatics. 2019;35(18):3357-64. Epub 2019/02/05.

154. Hodgson A, Liu SK, Vesprini D, Xu B, Downes MR. Basal-subtype bladder tumours show a 'hot' immunophenotype. Histopathology. 2018;73(5):748-57.

155. Mo Q, Nikolos F, Chen F, Tramel Z, Lee YC, Hayashi K, Xiao J, Shen J, Chan KS. Prognostic Power of a Tumor Differentiation Gene Signature for Bladder Urothelial Carcinomas. Journal of the National Cancer Institute. 2018;110(5):448-59.

156. Shen DW, Pouliot LM, Hall MD, Gottesman MM. Cisplatin resistance: a cellular self-defense mechanism resulting from multiple epigenetic and genetic changes. Pharmacological reviews. 2012;64(3):706-21.

157. Seiler R, Ashab HAD, Erho N, van Rhijn BWG, Winters B, Douglas J, Van Kessel KE, Fransen van de Putte EE, Sommerlad M, Wang NQ, et al. Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy. European urology. 2017;72(4):544-54.

158. Tanaka H, Yoshida S, Koga F, Toda K, Yoshimura R, Nakajima Y, Sugawara E, Akashi T, Waseda Y, Inoue M, et al. Impact of Immunohistochemistry-Based Subtypes in Muscle-

Invasive Bladder Cancer on Response to Chemoradiation Therapy. International journal of radiation oncology, biology, physics. 2018.

159. Yasar O, Akcay T, Obek C, Turegun FA. Significance of S100A8, S100A9 and calprotectin levels in bladder cancer. Scandinavian journal of clinical and laboratory investigation. 2017;77(6):437-41.

160. Wang W, Jiang H, Zhu H, Zhang H, Gong J, Zhang L, Ding Q. Overexpression of high mobility group box 1 and 2 is associated with the progression and angiogenesis of human bladder carcinoma. Oncology letters. 2013;5(3):884-8.

161. Cerezo M, Guemiri R, Druillennec S, Girault I, Malka-Mahieu H, Shen S, Allard D, Martineau S, Welsch C, Agoussi S, et al. Translational control of tumor immune escape via the eIF4F-STAT1-PD-L1 axis in melanoma. Nature medicine. 2018;24(12):1877-86.

162. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes & development. 2010;24(21):2343-64.

163. Kim D, Kwon NH, Kim S. Association of aminoacyl-tRNA synthetases with cancer. Topics in current chemistry. 2014;344:207-45.

164. Lopez-Sambrooks C, Shrimal S, Khodier C, Flaherty DP, Rinis N, Charest JC, Gao N, Zhao P, Wells L, Lewis TA, et al. Oligosaccharyltransferase inhibition induces senescence in RTK-driven tumor cells. Nature chemical biology. 2016;12(12):1023-30.

165. Bhattacharya S, Chakraborty D, Basu M, Ghosh MK. Emerging insights into HAUSP (USP7) in physiology, cancer and other diseases. Signal transduction and targeted therapy. 2018;3:17.

166. Richardson RT, Bencic DC, O'Rand MG. Comparison of mouse and human NASP genes and expression in human transformed and tumor cell lines. Gene. 2001;274(1-2):67-75.

167. Yu B, Chen X, Li J, Gu Q, Zhu Z, Li C, Su L, Liu B. microRNA-29c inhibits cell proliferation by targeting NASP in human gastric cancer. BMC cancer. 2017;17(1):109-19.

168. Zhao X, Li J, Huang S, Wan X, Luo H, Wu D. MiRNA-29c regulates cell growth and invasion by targeting CDK6 in bladder cancer. American journal of translational research. 2015;7(8):1382-89.

169. Yu L, Di Y, Xin L, Ren Y, Liu X, Sun X, Zhang W, Yao Z, Yang J. SND1 acts as a novel gene transcription activator recognizing the conserved Motif domains of Smad promoters, inducing TGFbeta1 response and breast cancer metastasis. Oncogene. 2017;36(27):3903-14.

170. Jariwala N, Rajasekaran D, Srivastava J, Gredler R, Akiel MA, Robertson CL, Emdad L, Fisher PB, Sarkar D. Role of the staphylococcal nuclease and tudor domain containing 1 in oncogenesis (review). International journal of oncology. 2015;46(2):465-73.

171. Giubellino A, Burke TR, Jr., Bottaro DP. Grb2 signaling in cell motility and cancer. Expert opinion on therapeutic targets. 2008;12(8):1021-33.

172. Watanabe T, Shinohara N, Moriya K, Sazawa A, Kobayashi Y, Ogiso Y, Takiguchi M, Yasuda J, Koyanagi T, Kuzumaki N, et al. Significance of the Grb2 and son of sevenless (Sos) proteins in human bladder cancer cell lines. IUBMB life. 2000;49(4):317-20.

173. Stoimenov I, Helleday T. PCNA on the crossroad of cancer. Biochemical Society transactions. 2009;37(Pt 3):605-13.

174. Yildirim A, Kosem M, Sayar I, Gelincik I, Yavuz A, Bozkurt A, Erkorkmaz U, Bayram I. Relationship of PCNA, C-erbB2 and CD44s expression with tumor grade and stage in

urothelial carcinomas of the bladder. International journal of clinical and experimental medicine. 2014;7(6):1516-23.