UNIVERSIDAD AUTÓNOMA DE MADRID
FACULTAD DE MEDICINA
Departamento de Medicina Preventiva y Salud Pública

PhD Thesis

ESTUDIO PAN-GENÓMICO PRONÓSTICO EN EL CÁNCER DE VEJIGA

—

GENOME-WIDE PROGNOSTIC STUDY IN BLADDER CANCER

Antonio Clemente PICORNELL COMPANY

Madrid, 2013

La Dra. Núria Malats Riera, Jefa del Grupo de Epidemiología Genética y Molecular del Centro Nacional de Investigaciones Oncológicas (CNIO), como Directora.

CERTIFICA:

Que Don Antonio Clemente Picornell Company, Licenciado en Biología por la Universidad Complutense de Madrid, ha realizado la presente Tesis Doctoral **"Estudio pan-genómico pronóstico en el cáncer de vejiga"** y que a su juicio reúne plenamente todos los requisitos necesarios para optar al **Grado de Doctor**, a cuyos efectos será presentada en la Universidad Autónoma de Madrid, autorizando su presentación ante el Tribunal Calificador.

Y para que así conste se extiende el presente certificado,

Madrid, abril 2013.

Vº Bº de la Directora:

Dra. Núria Malats Riera

To Montse

To my parents and my brother

Acknowledgements

To my PhD. Thesis director, Núria Malats for her instruction and support during the last four years.

To all the patients from the Spanish Bladder Cancer (SBC)/EPICURO Study, the monitors that enrolled them and the MDs that took part in the different phases of the study.

To all the researchers and collaborators for trying to make a better world through our hard work.

To Gonzalo Gómez, David G. Pisano, Alfonso Valencia and Francisco Real for guiding me during my first steps into the field of cancer research at the CNIO.

To the postdocs Andre, Roger and Evangelina for sharing their astonishing knowledge and free time outside of the lab.

To my predoc colleages Gaëlle, Matt, Salman and Silvia for sharing knowledge, experiences, frustrations and happy moments.

To my PhD. tutor Fernando Artalejo, from the Department of Preventive Medicine and Public Health, for its invaluable help both in epidemiologic and bureaucratic issues.

To everybody that has ever helped me during my research.

To my parents and brother for their never-ending love and support.

To Montse for sharing our lives together.

CONTENTS

Figures

Tables

Summary

Introduction

Urothelial carcinoma of the bladder (UCB) is a public health problem in Spain because of its high incidence in the male population. Its chronic nature requires continuous monitoring of the patient for the rest of his life. This implies deterioration in the quality of life of patients and a major expense for the national health system.

The incidence of this tumor is between 3-7 times higher in men in comparison to women and increases with age, peaking between 50 and 70 years. In addition, the geographical distribution is heterogeneous, with a much higher incidence in developed countries, except in northeast Africa.

About 80% of cases are diagnosed in patients with non-muscle invasive bladder cancer (NMIBC) and the rest suffer the invasive subtype (MIBC). However, there is a marked heterogeneity between superficial tumors in terms of biology, pathological features - Ta/G1-G2 in low-risk, TaG3 + T1/G2-G3 for high risk - and prognosis. This stratification is further supported by the identification of genetic alterations in *FGFR3* and *PI3KCA* in low-risk NMIBC and *p53* and *Rb* in high-risk NMIBC and invasive bladder cancer.

Risk factors established for UCB are tobacco smoking and occupational exposures to aromatic amines or polycyclic aromatic hydrocarbons. Although, occasionally, it has been described familial aggregation for this tumor, no high penetrance gene has been identified. On the other hand, bladder cancer represents a paradigm regarding the participation in the development of low-penetrance variants: *GSMT1*-null and *NAT2*-slow, the latter in interaction with tobacco smoking. Massive genotyping initiatives (Genome Wide

Association Studies, GWAS) have recently identified 10 additional variants associated with the risk of developing the disease. Although some studies have suggested the role of genetic factors in the prognosis of bladder cancer, the results are still limited, inconclusive and to date there is no study at genome-wide level.

To establish the prognosis of the NMIBC, urologists consider mainly the extent and depth of tumor invasion, presence of multiple tumors and the size of the major mass. The TNM staging system is used for the MIBC assessment. However, these prognostic factors are insufficient to subclassify the patients and accurately predict the evolution of UCB.

Objectives

The overall objective was to prove that inherited genetic variants are also involved in cancer progression, specifically UCB. Therefore, the specific objectives were: 1) to identify SNPs independently associated with UCB (NMIBC and MIBC) outcomes through a genome-wide tiered association study; 2) To detect pairs of interacting SNPs associated with UCB outcomes, also at a genome-wide scale; 3) To identify biological functions altered by an overrepresentation of SNPs identified in the first phase and associated with each UCB outcome.

Methodology

We studied a cohort of 1,300 patients from the Spanish Bladder Cancer SBC/EPICURO Study with genetic information obtained from high-throughput genotyping by Illumina Infinium HumanHap 1M probe Beadchip platform. This study was carried out between 1998 and 2001 in 18 general hospitals in five Spanish regions. The SBC/EPICURO Study provides an invaluable source of information regarding epidemiological data, clinical

monitoring, molecular and genetic as well as being one of the largest studies on this topic worldwide.

The initial phases of the project focused on individualized analysis of SNPs regarding the assessment of recurrence, progression and death due to UCB. The analysis was performed using uni-/multivariate Cox regressions considering the traditional prognostic factors. The results were combined and analyzed using a meta-analysis with an independent cohort of bladder cancer patients from the MD Anderson Cancer Center Hospital, Houston, USA. Subsequently, we validated the top results in other European and American series with which collaboration was established.

The screening of all pairs of possible interactions between the genotyped SNPs was carried out using the algorithm BOOST. Survival analyses followed for those interactions described as statistically significant. Additionally, we studied the potential synergies that may arise between the SNPs associated with common biological functions by analyzing biological pathways. The algorithms called ALIGATOR, GeSBAP, *i*-Gsea*4*Gwas, GSA-SNP and ICSNPathway were used to assess the potentially altered biological functions related to cancer development.

Results

We initially identified a total of 57 SNPs whose survival adjusted models showed to be independently associated with NMIBC. After being validated in five independent cohorts, rs754799 and rs4246835 SNPs hold their significance for recurrence and progression, respectively. Other 7 SNPs showed associations with disease close to the statistical significance level. The survival analysis of the patients who develop MIBC showed 57 SNPs potentially associated with the outcomes. The SNPs rs16927851 and rs1015267 were

with the top SNPs associated with progression and cancer-related death in patients with MIBC. The validation of the latter findings in independent patient series will take place in the near future.

We also proceeded with the pairwise survival analysis of all the non-correlated SNPs contained in the genotyping platform. In this way, we found pairs of SNPs whose combined effect was associated with UCB prognosis

Finally, we studied whether some biological function could be altered due to an overrepresentation of SNPs in genes contained in particular biological pathways. After evaluating five different methods, specific results were obtained for non-invasive bladder cancer. These are associated to biological functions regarding the inflammatory system and the immune response. Moreover, in all the UCB outcomes we identified SNP enrichment in pathways in which the GTPases, the membrane transport systems, neuro/ axonogenesis and angiogenesis play important roles.

Conclusions

We identified SNPs whose main effects are associated with UCB prognosis. In addition, we observed synergistic actions of these markers by analyses based on both the effect of their interactions and their possible role in gene networks associated with biological functions.

Therefore, it was verified that germline common variants are associated with several UCB clinical outcomes. The identification of these genetic variants makes possible to use them as prognostic markers that may allow a better classification and treatment of the patients.

Resumen

Introducción

El carcinoma urotelial de vejiga (CUV) constituye un problema de salud pública en España debido a su alta incidencia en la población masculina y a su naturaleza crónica que requiere de un seguimiento continuado del paciente durante el resto de su vida. Ello conlleva un deterioro en la calidad de vida de los pacientes y un gasto muy importante para el sistema nacional de salud.

La incidencia de este tumor es entre 3-7 veces mayor en la población masculina y aumenta con la edad, con un pico entre los 50 y 70 años. Además, su distribución geográfica a escala mundial es heterogénea, con una incidencia muy superior en los países desarrollados, excepto en el noreste africano.

Alrededor del 80% de los casos diagnosticados corresponden a pacientes con tumores de vejiga no-invasivos del músculo (NMIBC) y el resto al subtipo invasivo (MIBC). Además, en los tumores no-invasivos, la heterogeneidad es acusada en cuanto a la biología, características patológicas - Ta/G1-G2 en los de bajo riesgo y TaG3+T1/G2-G3 para los de alto riesgo - y al pronóstico. Esta clasificación está sustentada en alteraciones genéticas en *FGFR3* y *PI3KCA* en los no- invasivos de músculo de bajo riesgo y en *p53* y *Rb* en los no-invasivos de alto riesgo e invasivos de músculo.

Los factores de riesgo establecidos para el CUV son el tabaco y las exposiciones ocupacionales a aminas aromáticas o hidrocarburos aromáticos policíclicos. Aunque se ha descrito ocasionalmente agregación familiar para este tumor, no se ha identificado ningún

gen de alta penetrancia. No obstante, el CUV representa un paradigma en cuanto a la participación en su desarrollo de variantes de baja penetrancia: *GSMT1*-null y *NAT2*-slow; este último en interacción con el tabaco. Iniciativas de genotipación masiva (Genome Wide Association Studies, GWAS) han permitido, recientemente, identificar 10 variantes genéticas adicionales asociadas a este tumor. Aunque algunos estudios han sugerido el papel de estos factores genéticos en el pronóstico del CUV, los resultados son aún escasos, poco concluyentes y, hasta la fecha, no hay ninguno a nivel pan-genómico.

Para establecer el pronóstico del cáncer de vejiga no-invasivo del músculo, los urólogos consideran, principalmente el grado de diferenciación y la profundidad de invasión tumoral de la pared vesical, la presencia de múltiples tumores y el tamaño de éstos. Por otra parte, se usa el sistema TNM en los subtipos invasivos de de la enfermedad. No obstante, estos factores pronósticos son insuficientes para subclasificar a los pacientes con precisión y predecir la evolución del CUV.

Objetivos

El objetivo principal era probar que las variantes genéticas de línea germinal están asocidas con la progresión del cáncer de vejiga, específicamente de CUV. Por ello, los objetivos específicos fueron: 1) la identificación de los SNPs independientemente asociados con los desenlaces del CUV (NMIBC y MIBC) mediante un estudio pan-genómico; 2) La detección de parejas de SNPs en interacción asociados con los desenlaces del CUV, también a escala pan-genómica; La identificación de las funciones biológicas alteradas por una sobrerrepresentación de SNPs identificados en la primera fase y asociados con cada uno de los desenlaces del CUV.

Metodología

Se estudió la cohorte de 1.300 pacientes del estudio EPICURO, que dispone de información genética obtenida mediante la genotipación masiva por la plataforma Illumina Infinium HumanHap de 1 millón de SNPs. Este estudio se llevó a cabo entre 1998 y 2001 en 18 hospitales generales de 5 regiones españolas. El estudio EPICURO proporciona una fuente de valor inestimable en cuanto a datos epidemiológicos, clínicos, de seguimiento, moleculares y genéticos, además de ser uno de los estudios más grandes sobre esta patología a escala mundial.

Las fases iniciales del proyecto se centraron en el análisis individualizado por SNP de la información procedente del genotipado de los pacientes en relación a la predicción de recurrencia, progresión y muerte por CUV. El análisis se llevó a cabo mediante regresiones de Cox uni-/multivariantes considerando los factores pronósticos clásicos en el estudio sobre la evolución de la enfermedad. Los resultados de estos análisis se analizaron mediante un meta-análisis en una serie independiente de pacientes de CUV procedente del Hospital MD Anderson Cancer Center, Houston, USA. Posteriormente se replicaron parte de los resultados en otras series europeas y americanas con las que se estableció una colaboración.

El cribado de todos los pares de interacciones posibles entre los SNPs genotipados se realizó mediante el algoritmo BOOST, seguido del análisis de supervivencia descrito para aquellas interacciones más significativas. Adicionalmente se estudiaron las posibles sinergias que puedan aparecer entre los SNPs asociados a funciones biológicas comunes mediante el análisis de grupos génicos. Se utilizaron los algoritmos ALIGATOR, GeSBAP,

i-Gsea*4*Gwas, GSA-SNP e ICSNPathway para evaluar el valor predictivo de las presuntas funciones biológicas alteradas en relación al desarrollo del cáncer.

Resultados

Inicialmente se identificaron un total de 57 SNPs cuyos modelos ajustados de supervivencia mostraban estar potencialmente asociados al cáncer de vejiga no-invasivo del músculo. Tras ser evaluados en cinco cohortes independientes, los SNPs rs754799 y rs4246835 se validaron para recurrencia y progresión, respectivamente. Otros 7 SNPs presentaron asociaciones con la enfermedad cercanas al nivel de significación estadística. Al analizar la supervivencia de los pacientes que desarrollan variantes invasivas del músculo, se obtuvieron 57 SNPs potencialmente asociados con los desenlaces de la enfermedad. Los SNPs rs16927851 y rs1015267 mostraron una potencial asociación con los eventos de progresión y muerte por CUV en los pacientes que presentaban neoplasias invasivas. La validación de estos resultados en series independientes de pacientes tendrá lugar en un futuro próximo.

También se procedió al análisis de supervivencia de todas las posibles interacciones entre los SNPs de la plataforma de genotipación utilizada. De este modo se detectaron aquellas parejas de SNPs cuyo efecto combinado estaba asociado con el pronóstico del CUV.

Finalmente, se estudió si alguna función biológica pudiera estar alterada debido a una sobrerrepresentación de SNPs con valor pronóstico asociados a los genes que la conforman. Tras evaluar cinco métodos distintos, se obtuvieron resultados específicos para el cáncer de vejiga no-invasivo del músculo. Éstos corresponden a funciones biológicas relativas al sistema inmune y a la respuesta inflamatoria. Por otra parte, se identificaron posibles

alteraciones en aquellas funciones biológicas en las que las GTPasas, los sistemas de transporte de membrana, la neuro/axonogénesis y la angiogénesis juegan un papel importante.

Conclusiones

Se identificaron SNPs cuyos efectos principales están asociados al pronóstico del CUV. Además, se estudiaron las acciones sinérgicas de estos marcadores mediante análisis basados tanto en el efecto de sus interacciones como en su posible papel desarrollado en redes génicas reguladoras de funciones biológicas.

Así, se verificó que las variantes genéticas de línea germinal están asociadas con diversos tipos de desenlace del CUV. La identificación de estas variantes genéticas hace posible su uso como marcadores pronósticos que podrían posibilitar una mejor clasificación y tratamiento de los pacientes.

Abbreviations

BC	Bladder cancer
BCG	Bacille Calmette-Guerin
BCRC	Bladder cancer research consortium
CI	Confidence interval
CNIO	Spanish national cancer research center
CNV	Copy-number variation
CP	Canonical pathways
CUETO	Club urológico español de tratamiento oncológico
DNA	Deoxyribonucleic acid
EORTC	European organization for research and treatment of cancer
FDR	False discovery rate
GO	Gene ontology
GSA	Gene set analysis
GWAS	Genome-wide association study
GWIA	Genome-wide interaction analysis
GWPS	Genome-wide prognostic study
HR	Hazard ratio
IBCC	International bladder cancer consortium
IMIM	Institut municipal d'investigació mèdica
ISUP	International society of urological pathology
kb	Kilo bases
LD	Linkage disequilibrium
LRR	Log R Ratio
mb-MDR	Model-based MDR
MDACC	MD Anderson cancer center
MDR	Multifactor dimensionality reduction
MIBC	Muscle invasive bladder cancer
MoI	Mode of inheritance
NCI	National cancer institute
NMIBC	Non-muscle invasive bladder cancer
PCA	Principal component analysis
OR	Odds ratio
PCR	Polymerase chain reaction
Q-Q	Quantile-quantile
SBC/EPICURO	Spanish bladder cancer/EPICURO
SNP	Single-nucleotide polymorphism
SWOG	South west oncology group
Tis	Carcinoma in situ
TUR	Transurethral resection

UCB	Urothelial cell carcinoma of the bladder
WHO	World health organization

Chapter 1: Introduction

Part I: Bladder cancer

1.1.1. An overview of bladder cancer

Cancer is the main cause of death in the economically development countries and the second cause of death in developing countries (Jemal, Bray et al. 2011). The present work is focused on bladder cancer, which represents one of the major types of cancer. According to the last available information in GLOBOCAN 2008 (Ferlay, Shin et al. 2010), the number of bladder cancer cases and deaths reach up to 382,660 and 150,282 respectively. The world age-standardized incidence rate is 5.3 per 100,000 but it shows higher rates in economically developed countries.

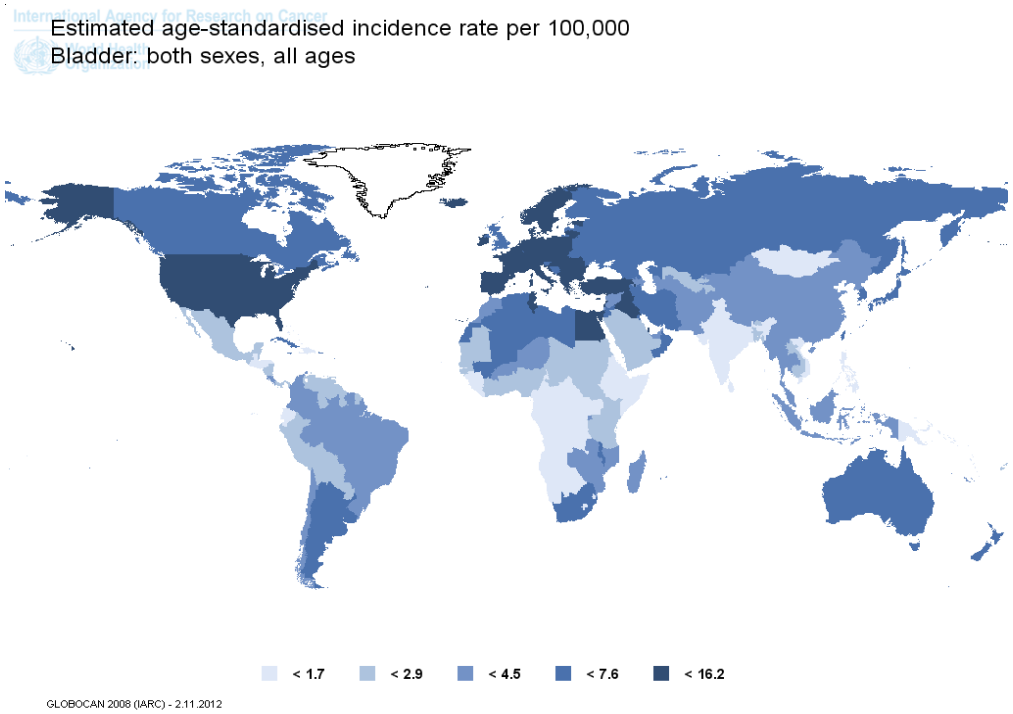


Figure 1. Bladder cancer world map. The age-standardized incidence rates for each country (GLOBOCAN 2008).

Bladder cancer represents the fifth most common cancer in Spain. However, the incidence is much higher among men (4th most frequent cancer) than among women (15th most frequent cancer). The Spanish population exhibits one of the highest incidence rates among men (27.7 per 100,000 person-year) and one of the lowest among women (3.2 per 100,000) worldwide, with a gender man:woman ratio of 7:1, in contrast with the 3:1 ratio in the other industrialized countries. The five-year prevalence is 3.4 times (592,663 vs. 174,815 cases) higher than the incidence in Westernized countries. This fact jointly with the need of clinical surveillance of patients with cystoscopies to avoid tumor progression results in an important cost for all heath care systems (Ferlay, Shin et al. 2010).

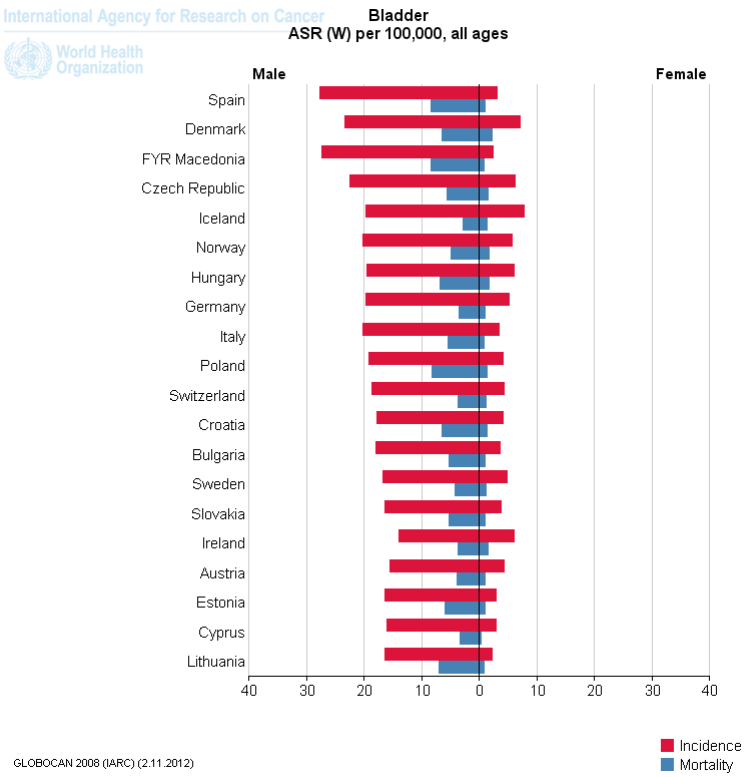


Figure 2. Age-standardized incidence and mortality rates of bladder cancer in the top European countries (GLOBOCAN 2008).

1.1.2. UCB symptoms, diagnosis and treatment

There are several types of bladder cancer (BC) according to the morphology/histology of the tumor. In the present work we focus on the of urothelial or transitional cell carcinoma of the bladder (UCB), which is the most common and it is diagnosed in about 90% of the bladder cancer patients (Silverman, Devesa et al. 2006). Patients with UCB usually present painless hematuria. However, the irritative symptoms (frequency, urgency or dysuria) can be present at early stages. There is usually a common delay in the diagnosis because of the similarity and the possible intermittence of these symptoms to other disorders or diseases (presence of renal calculi, cystitis, infection, or even prostatitis) that leads to a poorer prognosis. In advanced stages of the disease, metastases may be responsible of initial symptoms.

An initial physical examination must be performed in patients that are suspected to suffer UCB. It includes a digital rectal examination in men and a bimanual examination of the vagina and rectum in women. Nevertheless, this examination is normally unremarkable. If there is not a clear source of glomerular bleeding, a full urologic evaluation of the entire urinary tract is indicated. It consists of cystourethroscopy and urinary cytology followed by an evaluation of the upper tracts. Radiographic imaging of the upper tract and urinalysis may be also performed. The gold standard for the initial diagnosis and management of UCB is cystoscopy. Visible tumors are either biopsied to determine the histology and depth of invasion into the bladder submucosa and muscle layers. Additionally tumors could be resected by transurethral resection (TUR). Low-grade, non-invasive tumors are normally stalked and papillary. On the other hand high-grade, invasive tumors are frequently sessile. The less common carcinoma in situ (Tis) is a high-grade, non-invasive tumor that appears

as a flat lesion. Tumoral characteristics such as the size, stalk and location may predict the muscle invasion. Urine cytology is commonly used in combination with cystoscopy to assed the presence of Tis and to evaluated for possible upper tract malignancies. However, urine cytology has a relatively poor sensitivity, particularly for low-grade tumors (Lotan and Roehrborn 2003).

The patients with UCB are classified in two subgroups according to their prognosis and the tumor genetic alterations (Netto 2012): non-muscle invasive bladder cancer (NMIBC), representing 75-85% of UCB, and muscle invasive subtype (MIBC) (Babjuk, Oosterlinck et al. 2011). This classification is based on the tumor invasiveness or stage (T) and grade (G). Non-muscle invasive tumors are further classified based on the depth of the invasion into papillary carcinoma (Ta and T1, the latter invading subepithelial connectivity tissue) and Tis. Muscle-invasive tumors are classified as T2 when they only invade the muscle layer of the bladder, T3 when they invade perivesical tissue, and T4 when the tumor spreads to nearby organs. UCB is further subclassified according to the histological grading (G) of the tumors as well (G1), moderately (G2) or poorly differentiated (G3). By combining T and G, UCB are diagnosed as low (TaG1/G2) or high risk tumors (TaG3, T1 and Tis), a classification that is highly correlated with the probability to progress. The World Health Organization (WHO) and the International Society of Urological Pathology (ISUP) have proposed a new classification based on the histopathology of the tumor. It consists on a biopsy followed by a microscopic examination of its tissue slide, in which it could be observed two or three grades of cellular differentiation, ranging from poor to well differentiated, according to the distinct WHO/ISUP classifications: WHO1973 and WHO/ISUP 1999/2004 (Babjuk, Oosterlinck et al. 2011).

The treatment to be applied as first line therapy is determined by the kind of tumor. In NMIBC patients TUR is the standard treatment. Additionally in can be followed immunological therapy based on bacille Calmette-Guérin (BCG), chemotherapy or both. Tumor recurrence represents a clinical problem as far as, in 70% of the patients diagnosed with a NMIBC, the tumor reappears after a TUR (Kurth, Denis et al. 1992). Fortunately the additional BCG instillation reduces this event by 30% (Smith, Labasky et al. 1999). The benefits of using BCG over chemotherapy are well established for recurrence but its role in progression remains unclear (Gontero, Bohle et al. 2010). The treatments applied to MIBC patients are much aggressive compared with those mentioned before: radical cystectomy, multimodal therapy or cisplatin-containing chemotherapy (Stein, Lieskovsky et al. 2001).

After treatment, the patients need to be routinely monitored by cystoscopies to control the reappearance of tumoral growth. There is not any established period of time to perform this follow-up process but the European Association of Urology (Babjuk et al., 2008, 2009) gives some clues. Nevertheless, a cystoscopy after three months after the treatment has been shown to be predictive for the relapse of the disease. After that, the schedule of follow-ups is suggested to be different for low-risk and high-risk patients. If the results of the last mentioned cystoscopy are negative, it is recommended to perform a new one after 9 months and then annually for the next 5 years. The high-risk patients need a more exhaustive follow-up based on new cystoscopies every 3 months for the next 3 years, every 4 months in the third year, every six months until the fifth year and annually thereafter (Babjuk, Oosterlinck et al. 2011).

1.1.3. Evolution and prognosis of UCB

UCB can be classified in two groups: 75-80% of cases are NMIBC and the rest have muscle invasive or metastatic neoplasms. Just a few of the MIBC patients (3-15%) have a previous history of papillary tumors (Dinney, McConkey et al. 2004; van Rhijn, Burger et al. 2009). A patient with a primary non-muscle invasive tumor can be diagnosed as a new bladder neoplasm (called tumor relapse) during his follow-up. It may reappear in the form of a recurrence or a progression and lead to muscle invasive or metastatic forms of the disease. For primary diagnoses MIBC patients the progression is considered as the appearance of an advanced stage of the disease that may lead to death due to UCB or any other cause.

Clinical and pathological variables for UCB

Around 70% of NMIBC patients present as Ta stage, 20% as T1, and 10% as Tis. Recurrence occurs in 50–80% of the patients and is the main threat for patients with Ta tumors. On the other hand, progression, which appears in 10–30% of patients, is the main problem in T1 and Tis affected subjects (van Rhijn, Burger et al. 2009). The percentage of recurrence in NMIBC patients at 1 and 5 years ranges from 15–61% and 31–78%, respectively. The percentages for patients who suffered progression at 1 and 5 years range from <1–17% and 1–45%, in each case (Sylvester, van der Meijden et al. 2006). The wide range for recurrence and progression rates is due to the presence of tumor heterogeneity in these patients (Kurth, Denis et al. 1995; Allard, Bernard et al. 1998). High-risk non-muscle invasive tumors (TaG3 and T1G2/G3) tend to progress into muscle-invasive and metastatic forms of disease. Around 25% of T2 tumors, 50% of T3 tumors and 80% of T4 tumors

eventually evolve into metastasis. The MIBC 5-year survival is 67% for T2 tumors, 35% for T3 tumors and 27% for T4 tumors (Herr, Dotan et al. 2007).

The most accepted clinical prognostic factors for tumor recurrence in NMIBC are multiplicity (Parmar, Freedman et al. 1989; Shinka, Hirano et al. 1990; Kiemeney, Witjes et al. 1994; Millan-Rodriguez, Chechile-Toniolo et al. 2000), tumor stage (Shinka, Hirano et al. 1990; Kiemeney, Witjes et al. 1994), tumor grade (Shinka, Hirano et al. 1990; Kurth, Denis et al. 1995), and tumor size (Kurth, Denis et al. 1995; Millan-Rodriguez, Chechile-Toniolo et al. 2000). However, the result of the first cystoscopy performed after 3 months after TUR [9] can be considered the strongest prognostic factor (Parmar, Freedman et al. 1989). On the other hand, the most important variables for prediction of progression in NMIBC are the presence of Tis, a grade 3, and a stage T1 tumor (Sylvester 2006). Additional variables for this outcome have been suggested: a recurrence a first cystoscopy, stage, grade and prior tumor (Fernandez-Gomez, Solsona et al. 2008).

The most important predictor of MIBC outcome is radical cystectomy, followed by tumor stage (Stein, Lieskovsky et al. 2001). Lymph-node involvement is also one of the most important prognostic factors when progression and survival are evaluated (Margulis, Lotan et al. 2008). In addition, the presence of metastasis is a widely known prognostic factor incorporated in the TNM system. Resistance to neoadjuvant chemotherapy and divergent histology are suggested to take into account (Sternberg 2002; Jeon and Chang 2005; Turkolmez, Tokgoz et al. 2007; Bruins and Stein 2008). The roles of sex, MNA (mean nuclear area) and tumor grade remain uncertain in the prognosis of these patients (Jeon and Chang 2005).

There is a need to provide predictive tools to the physicians in order to assess the need for intravesical therapy and early cystectomy for NMIBC patients. On the other hand, MIBC patients also can take advantage of treatment assignment if the prediction of advanced disease response to primary treatment is available. The traditional way to estimate the risk of developing a particular outcome has been based in the TNM staging system.

Somatic genetic prognostic markers for UCB

The list of candidate biomarkers to improve the prediction of NMIBC events has been growing over the last decade. The earliest genetic alterations thought to have a potential diagnostic and prognostic value were those located in chromosome 9. After that, new chromosomal alterations located in 3q, 7p and 17q gains were suggested. Urine assays based on fluorescence *in situ* hybridization were shown to be useful to detect those alterations (Kruger, Mess et al. 2003; Kawauchi, Sakai et al. 2009). Genetic alterations on receptor protein kinase genes (RTK), such as *FGFR3, HRAS* and *PIK3CA*, have been associated to the pathogenesis of NMIBC. Mutated and/or altered expression of some of them (*FGFR3, EGFR, ERBB2* and *ERBB3*) is thought to have prognostic value for both NMI/MIBC. Multiplex PCR assays have been developed in order to detect these alterations (Lopez-Knowles, Hernandez et al. 2006; Kompier, Lurkin et al. 2010). Epigenetic analyses, which used quantitative methylation specific PCR (MSP-PCR), were applied on bladder tumor specimens and associated the disease progression to promoter hypermethylation of *RASSF1, DAPK, APC, CDH1* and *EDNRB (Yates, Rehman et al. 2007)*. Ploidy and S phase kinetics have also been suggested as potential prognostic factors for recurrence and progression in NMIBC using automated image or flow cytometry (Loughman, Lin et al. 2003). The most promising prognostic factors for NMIBC are tumor proliferation index

(calculated from either *KI67* or *MIB1*) and a molecular grade parameter based on combining *FGFR3* gene mutation status and *MIB1* labeling index (van Rhijn, Vis et al. 2003).

There is also an extensive list of potential biomarkers specific for MIBC. The alteration of tumor suppressor genes involved in the cell cycle regulation (*TP53*, *P16* and *RB1*) has been shown to be associated with the prognosis of MIBC (Sanchez-Carbayo, Socci et al. 2002; Mitra, Datar et al. 2006). The epigenetics also seems to play a significative role in the assessment of MIBC prognosis. The promoter hypermethylation of *RASSF1*, *CDH1* and *EDNRB* has been suggested (Yates, Rehman et al. 2007). The overexpression of RTKs also seems to be associated to MIBC and the alteration of ERBB2 seems to be important in assessing for the death due to UCB (Bolenz, Shariat et al. 2010). One particular protein emerges Recent evidence suggest the importance of mTOR pathway alterations in MIBC and it is being used as a potentially useful biomarker in targeted therapy in phase II trials (Tickoo, Milowsky et al. 2011).

There are evidences that suggest the potential prognostic value for some biomarkers in both NMIBC and MIBC. One of the most prominent is the angiogenesis marker overexpression (*VEGF*, *HIF1A* and *THBS1*). Protein Ras is supposed to promote survival and angiogenesis by upregulating the PI3K–AKT–mTOR pathway. Phase II clinical trials are being conducted using bevacizumab, a recombinant humanized monoclonal anti-VEGF antibody, in combination with gemcitabine/carboplatin chemotherapy in patients with metastatic UCB (Hahn, Stadler et al. 2011). Another phase II study is evaluating the same antibody with MVAC (methotrexate, vinblastine, adriamycin, and cisplatin) adjuvant chemotherapy (Elfiky and Rosenberg 2009). The loss of E-cadherin and gain of N-cadherin have been

studied and it was suggested that the latter is associated to UCB recurrence (Muramaki, Miyake et al. 2011; Muramaki, Miyake et al. 2012). On the other hand, N-cadherin-negative in muscle-invasive tumors is associated to poor prognosis (Jager, Becker et al. 2010).

Nomograms and scoring systems for UCB

The *International Bladder Cancer Consortium* (IBCC) developed a nomogram in which there were included more than 9,000 patients from 12 centers to predict the risk of tumor reappearance at 5 years after radical cystectomy. Its predictive accuracy was better than TNM staging (75% vs. 68%) (Bochner, Kattan et al. 2006). The *Bladder cancer Research Consortium* (BCRC) used a multi-center cohort of 731 patients, who went though radical cystectomy, in order to predict tumor reappearance and mortality (due to UCB or any other cause) at 2, 5 and 8 years after treatment. They obtained accuracies between 73-78% (Karakiewicz, Shariat et al. 2006; Shariat, Karakiewicz et al. 2006). The IBCC and BCRC nomograms have been validated in independent cohorts (Zaak, Burger et al. 2010).

A scoring system and risk tables for NMIBC has been developed by the *European Organization for Research and Treatment of Cancer* (EORTC) using individual data for 2,596 patients diagnosed with TaT1 tumors with neither a second TUR nor BCG maintenance therapy. This system relies on six clinical and pathological factors: T category, number of tumors/multiplicity, tumor grade, tumor size, presence of concomitant Tis, and prior recurrence rate. The scores are distributed into four categories that assess the probability of developing recurrence or progression at 1 and 5 years and define a classification system based on low-, intermediate- and high-risk groups (Sylvester, van der Meijden et al. 2006). Recently, the *Club Urológico Español de Tratamiento Oncológico*

(CUETO) has developed another scoring model based on BCG-treated patients. Some differences in the risk assessment arise when both methods are compared: The EORC tables offer lower risks for recurrence and progression in high-risk patients (Fernandez-Gomez, Madero et al. 2009). To our knowledge, there is a lack of reliable prognostic factors to predict the evolution of Tis tumors. There are only two reports regarding the poor prognosis of patients with concurrent Tis and T1 tumors compared with primary/extended Tis and the lack of response to the BCG instillation (Chade, Shariat et al. 2010; Babjuk, Oosterlinck et al. 2011). The use of molecular prognostic markers is not common yet though new tools came up recently such as a nomogram to predict NMIBC recurrence that makes use of increased concentration of nuclear matrix protein 22 (NMP22) in urine (Lotan, Capitanio et al. 2009).

The nomograms are believed to provide more predictive accuracy when compared to categorical models. Furthermore, they can be easily adapted in the clinical practice. Recently, a new generation of nomograms has incorporated biomarkers but we are still waiting for the addition of validated ones and new elements as modern imaging data. Nowadays, one of the weak points is the lack of standardized data collection methods that leads to the important heterogeneity within the variables used in this predictive tool.

Another important limitation in the assessment of UCB prognosis is the presence of tumor heterogeneity that leads to inefficient classification of the patients, clearly seen in the wide ranges of recurrence and progression mentioned before, that makes difficult to assign the correct treatment to the patient. Molecular diagnosis is routinely used in the clinical management of some cancers such as breast, lung and colon. However, its clinical use in UCB has been neglected. There is lack of validated biomarkers that may help the clinicians

to identify correctly the patients needing and early and aggressive kind of treatment; or the other way around, when patients are over-treated. The most studied UCB biomarkers are the somatic ones but the huge heterogeneity among them makes difficult to be validated. An extensive list of promising biomarkers is presented below.

Part II. The role of the inherited genetics

1.2.1. Germline genetic susceptibility markers for UCB

The assessment of UCB risk factors reveals the predominant effect of cigarette smoking, which triples the risk of developing UCB and accounts for 50-75% cases in men and 14–35% in women (Zeegers, Tan et al. 2000; Samanic, Kogevinas et al. 2006; Wu, Ros et al. 2008). Environmental risk factors and occupational exposures to potential carcinogens explain 10-20% of the cases (Silverman, Hartge et al. 1992; Vineis and Pirastu 1997). Other proposed non-genetic exposures still need to be confirmed, among them are chlorinated water, halogenated hydrocarbons, low arsenic levels, HPV, pioglitazone, nitrates/nitrites and second hand smoke (Kiriluk, Prasad et al. 2012). Genetic susceptibility to UCB etiology is suspected as far as just a few of the individuals with environmental exposures to potential risk factors will develop the disease.

Studies performed decades ago determined that UCB does not have a strong familial aggregation with no high-penetrance genetic variant being identified till present (Mueller, Caporaso et al. 2008). A large twin study conducted in Scandinavian population-based cohorts estimated that 31% of the total variance in UCB is explained by genetic factors, whereas non-shared environmental factors within the family would explain as much as the 67% (Lichtenstein, Holm et al. 2000). Studies on first-degree relatives of UCB patients reported an almost two-fold increased risk compared with general population (Aben, Witjes et al. 2002; Murta-Nascimento, Silverman et al. 2007). These studies present some limitation, among them the case-control study design, the reliability of past history of cancer in relatives, and the low number of high-risk families studied. As soon as the

beginning of 90's the role of *GSTM1*-null and *NAT2* in UCB susceptibility was described (Butler, Lang et al. 1992; Bell, Taylor et al. 1993). However at mid-2000's the common opinion was that "no major gene" model could explain the occurrence for sporadic UCB, even after conducting a large scale analysis of 1,193 families (Aben, Baglietto et al. 2006). All this background of knowledge suggested a role for common low-penetrance susceptibility loci.

The research of candidate genetic markers for susceptibility has been mainly based on the close study of a small number of genetic variants in genes and pathways whose involvement in UCB was already suspected (candidate markers). A considerable number of studies based on a candidate-approach pattern were conducted to assess the role of low penetrance genes (Malats 2008; Wu, Hildebrandt et al. 2009; Netto 2012). The studies focused mainly on polymorphisms in carcinogen metabolizing and DNA repair genes taking into account the role of confounding factors such as tobacco smoking and the occupational exposure to potential carcinogens. An astonishing number of potential genes that may play a role in the development of UCB have been reported (Grotenhuis, Vermeulen et al. 2010). Nevertheless, the vast majority of these results suggested false-positive findings due to underpowered studies and a worrying lack of interest or resources to reply them in larger and independent series (Wu, Hildebrandt et al. 2009; Grotenhuis, Vermeulen et al. 2010). Before GWAS (Genome-Wide Association Studies) were undertaken robust findings in susceptibility only existed for variants in *NAT2* that result in a slow acetylator phenotype and homozygous deletions of *GSTM1*. Both are involved in detoxification of carcinogens, environmental toxins and products of oxidative stress.

Hypothesis-driven candidate variant/gene/pathway approach is limited by the current knowledge of the disease and the number of biomarkers that are examined. The scarce reliable results using this approach made UCB a suitable target to apply a hypothesis-free approach based GWAS. This strategy has dramatically increased the number of susceptibility loci whose effects have been replicated in large sets of patients. Low-penetrance susceptibility loci have been identified in four independent studies (Kiemeney, Thorlacius et al. 2008; Wu, Ye et al. 2009; Rothman, Garcia-Closas et al. 2010; Garcia-Closas, Ye et al. 2011).

1.2.2. Germline genetic prognostic markers for UCB

All the germline-based studies published until now assessed for the prognostic potential of genes related to candidate pathways. Some of these studies reported statistically significant associations with different UCB outcomes (Grotenhuis, Vermeulen et al. 2010). These are the most commonly studied pathways: inflammation, DNA repair, xenobiotic metabolism, apoptosis, cell cycle control, cell adhesion, apoptosis, angiogenesis, stem cell biology and growth factor.

The inflammation pathway has been extensively studied searching for markers associated with recurrence in NMIBC. It was found that a variant in *IL6* is associated with longer survival times (Ahirwar, Kesarwani et al. 2008). Another popular set of genes to be evaluated for recurrence was the DNA repair pathway. Variants associated with shorter survival times in *ERCC6* (Gu, Zhao et al. 2005), *XRCC1* (Mittal, Singh et al. 2008) and *ERCC4* (Wang, Wang et al. 2010) were discovered. The cell-cycle control gene *TP53* was

also found to be associated with recurrence as far as a variant within the gene protects from developing this outcome (Horikawa, Nadaoka et al. 2008). Variants in the detoxification genes *hGPX1* (Zhao, Liang et al. 2005) and *NQO1* (Sanyal, Ryk et al. 2007) have been proven to be also associated with UCB recurrence, being associated to longer and shorter survival times, respectively. Moreover, the angiogenesis pathway was evaluated in high-grade NMIBC and two encoding variants, associated to high risk to develop recurrence were identified in *HIF1A* (Nadaoka, Horikawa et al. 2008). In addition, the role of cell adhesion genes in cancer was evaluated and an encoding variant with protective effect was discovered within *CDH1* (Lin, Dinney et al. 2006).

The variant in *IL6* found to be associated with UCB recurrence (rs1800795) is also significantly associated to NMIBC progression in an independent study (Leibovici, Grossman et al. 2005). This outcome was studied in Ta/T1 patients when the cell-cycle control genes were considered in two independent studies. A significant risk variant in *CDKN2A* was found in the first study (Leibovici, Grossman et al. 2005). The second one provided two variants in *CCND1* and *HRAS*, which were associated to shorter and longer progression times (Sanyal, Ryk et al. 2007). A haplotype-based study was conducted for the G-protein signaling pathway in NMIBC patients that develop progression and a risk haplotype was discovered in the first intron of *GNB4* (Riemann, Struwe et al. 2008).

The studies trying to discover genetic variants associated to UCB mortality have evaluated essentially the same pathway candidates mentioned before. The evaluation of the inflammation pathway offered two variants in *IL6* and *TNFA* associated to longer and sorter survival times, respectively (Leibovici, Grossman et al. 2005). Interestingly, the SNP mapping *IL6* (rs1800795) has a protective effect when recurrence and survival were

evaluated, but the opposite effect for NMIBC progression. Variants in the cell-cycle control pathway were also studied in MIBC patients after radical cystectomy and it was found that one variant in *TP53* is associated to shorter survival times (Horikawa, Nadaoka et al. 2008). The low sample sizes and alternative tumoral classification criteria make usual to consider altogether the high-grade NMIBC and MIBC cases. A study of that kind looked for prognostic variants in the angiogenesis pathway and found two of them in *HIF1A* associated with poor prognosis (Nadaoka, Horikawa et al. 2008). The mentioned haplotype in *GNB4* associated with progression was found to be also significantly associated with shorter survival times (Riemann, Struwe et al. 2008). The role of TGF-beta signaling has been studied in advanced stages of the disease. This is a dual-function cytokine that promotes tumor suppression and epithelial-mesenchymal transition that yields to cancer progression followed by metastasis (Wendt, Allington et al. 2009). In a population of 859 NMIBC and 246 MIBC patients, two SNPs associated with *TGFBR1* were found as significantly associated with disease-specific mortality when MIBC were evaluated (Castillejo, Rothman et al. 2009). A large-scale candidate gene study based on 400 cancer-related genes was conducted in a smaller population of 617 patients. The study successfully identified a list of SNPs with prognostic value for UCB survival. Longer survival was clearly associated with the detoxification gene *EPHX1*. Variants in several genes were found to be associated with shorter survival times: the surface antigen gene *CD80*, the apoptotic gene *BCL21*, the DNA repair gene *ERCC4*, the transcription factor *GATA3* and the inflammatory gene *CXCR2* (Andrew, Gui et al. 2009).

One of the most interesting gene sets to be analyzed is the inflammation pathway. The identification of SNPs with prognostic value in this pathway may give us some clues

regarding the success of BCG instillation treatments in NMIBC patients. The mentioned study of Leibovici *et al*. identified rs1800795 as a variant that increases the risk of recurrence and progression in patients receiving maintenance BCG (Leibovici, Grossman et al. 2005). However, the opposite effect is observed in other study (Ahirwar, Kesarwani et al. 2008). An additional assessment conducted by the same authors discovered two SNPs in *IFNG* and *TNFA* associated with shorter and longer times to recurrence, respectively (Ahirwar, Mandhani et al. 2009). The DNA repair pathway has been another popular approach in order to identify SNPs with prognostic value in patients with BCG instillation. Important risk associated with recurrence (HR between 3.07 and 6.80) was observed in studies discovering prognostic SNPs in *XRCC1* (Mittal, Singh et al. 2008), *ERCC2* (Gangawar, Ahirwar et al. 2010), *ERCC6* (Gu, Zhao et al. 2005) and *XPC* (Gangawar, Ahirwar et al. 2010). The last mentioned pathway has been extensively evaluated in MIBC patients that received aggressive treatments. Sakano *et al*. discovered two SNPs mapping *XRCC1* and *ERCC2* associated with good prognosis in T1G3 and MIBC patients that received chemoradiotherapy. Shinohara *et al.* conducted a study looking for prognostic variants in the cell-cycle control pathway. They found two SNPs mapping *MDAM2* and *TP53* associated with longer survival times in T1G3·and MIBC chemoradiotherapy-treated patients (Shinohara, Sakano et al. 2009).

As can be observed, the number of candidate-pathways studies is large but most of the studies are underpowered due to low sample sizes. This situation and the lack of replication in independent patient series may lead to contradictory results. An exception in this scenario is the prognostic evaluation of the Sonic hedgehog (Shh) pathway conducted by Chen *et al*. The *MD Anderson Cancer Center* (MDACC) recruited 419 patients treated with

TUR alone and discovered two SNPs associated with recurrence that could be replicated in an independent cohort, which consisted of 356 patients, from the *Spanish Bladder Cancer* SBC/EPICURO Study (Chen, Hildebrandt et al. 2010). The mentioned limitations have been tried to be solved using hypothesis-free approaches such as the genome-wide studies that now are going to be presented.

1.2.3. Fundamentals of genome-wide studies

Genome-wide approaches to assess the role of genetic biomarkers in human diseases date back to the early 80s. The first challenge was to select an appropriate approach to construct a linkage map of the human genome. The biotechnological knowledge in those days pointed out that restriction fragment length polymorphisms (RFLPs) may be the makers of choice (Botstein, White et al. 1980). The study of RFLPs, followed by Southern blot assays, exploits the presence of variations in homologous DNA sequences due to sequence changes through base deletions, substitutions or insertions that alter the common pattern of restriction enzyme recognition sites. Despite of the tediousness of the analysis of these biomarkers and the limitation to a few Mendelian diseases, in 1986 the first conclusions of this approach came out suggesting that most of the human traits and their associated diseases follow complex modes of inheritance (Lander and Botstein 1986). Based on previous knowledge on genetic animal models, an approach based on linkage disequilibrium (LD) mapping was proposed. The year after, the first genetic linkage map of the human genome was reported (Donis-Keller, Green et al. 1987) but the LD mapping in general human population was considered as impracticable because of the need of high

marker density. Thus, this approach was relegated to the study of populations with high founder effect.

The popularization and improvement of PCR (polymerase chain reaction) techniques in the late 80s and early 90s made possible to design PCR-based assays for microsatellite markers (Weber and May 1989; Weissenbach, Gyapay et al. 1992). The combination of family-linkage designs and LD to correct the gene locations became popular in genome-wide studies (Kerem, Rommens et al. 1989; Houwen, Baharloo et al. 1994; Puffenberger, Kauffman et al. 1994). Genetic susceptibility of many common diseases were studied and the hypothesis that few genes could explain the susceptibility to develop the disease was slowly rejected due to the evidences of a higher level of complexity based on the small effect of many loci acting altogether. This was a dead end because of the low ability to detect those loci in family-linkage studies.

In the mid-90s the efforts were focused on overcoming the mentioned limitations. A new approach based on association studies was suggested as a more powerful way to detect those common loci with small effects (Risch and Merikangas 1996). The family-linkage studies have a clear disadvantage when compared with the association studies because common alleles can be present in the family through multiple founders and show unclear inheritance patterns. In addition, it is much easier to recruit large numbers of unrelated individuals than relatives. On the other hand, the hypothesis of disease-common variant became the new cornerstone in genome-wide studies and the possible application of single nucleotide polymorphisms (SNPs) was pointed because their abundance, low mutation rate and ease of genotyping would made possible the construction of a dense map of polymorphisms for LD mapping (Lander 1996; Collins, Guyer et al. 1997). In 1999 the

U.K. Welcome Trust philanthropy and a group of the world's leading pharmaceutical companies[1] founded *The SNP Consortium* in order to create a LD map of the human genome using 300,000 common SNPs (Masood 1999). Early empirical observations highlighted a likely variation of the LD patterns across the genome and among different ethnical populations all over the world. These issues made necessary the construction of LD maps that had variable SNP densities in different locations and take into account the genetic background in several representative human populations. In addition, the initial suggested number of SNPs proved to be insufficient and it was extended to at least one million. All these observations and corrections contributed to the creation of the *International HapMap Project* in 2003 and the development of affordable genotyping platforms that would lead to the first GWAS just a few years later (Klein, Zeiss et al. 2005). From then, the number of GWAS has been increased every year with a total of 1533 GWAS which have reported 8699 significant loci have been published up March 2013 (Hindorff, Sethupathy et al. 2009).

[1] AP Biotech, AstraZeneca Group PLC, Aventis, Bayer Group AG, Bristol-Myers Squibb Co., F. Hoffmann-La Roche, Glaxo Wellcome PLC, IBM, Motorola, Novartis AG, Pfizer Inc., Searle, and SmithKline Beecham PLC.

1.2.4. Genome-wide association studies of UCB

Three independent UCB GWAS have been conducted at the *Radboud University Nijmegen Medical Center* (The Netherlands) in collaboration with *deCODE Genetics* (Iceland), the *MD Anderson Cancer Center* (Texas, USA) and the U.S. *National Cancer Institute* (USA) (Kiemeney, Thorlacius et al. 2008; Wu, Ye et al. 2009; Kiemeney, Sulem et al. 2010; Rothman, Garcia-Closas et al. 2010; Garcia-Closas, Ye et al. 2011). The published GWAS on UCB reveal twelve independent SNPs that modify the susceptibility to the disease.

The first published GWAS on bladder cancer pointed to an increase on susceptibility due to rs710521, which is located in a LD-block that affects *TP63*. It encodes P63, which regulates the cell-cycle arrest and the apoptotic process regarding the progression of UCB to the invasive subtype of the disease (Koga, Kawakami et al. 2003).

All UCB GWAS reported that chromosomal region 8q24 contains several loci that increase the cancer susceptibility. Although this region shows a scarce number of genes, several loci associated with breast, prostate, ovary, and colorectal cancer have been described to be located in this region (Ioannidis, Thomas et al. 2009). It has been hypothesized that they may play a role based on the regulation of *MYC* through epigenetic elements (Ahmadiyeh, Pomerantz et al. 2010). This gene encodes c-MYC, which is a nuclear phosphoprotein that regulates cell differentiation, apoptosis and growth regulation in regular situations. However, when it is deregulated it triggers malignant cell growth. The amplification of *MYC* gene has been observed in up to 30% of patients with UCB (Mahdy, Pan et al. 2001). In the same region a SNP (rs2294008) that alters the start codon of *PSCA* was located on its first exon. *PSCA* is a glycosylphosphatidylinositol-anchored cell surface protein that may

play a role in cell proliferation and migration. This gene is expressed at low levels in the transitional epithelium of normal bladder but has been shown to be overexpressed in most UCBs (Amara, Palapattu et al. 2001). In addition, *PSCA* may be a predictor for recurrence in NMIBC (Elsamman, Fukumori et al. 2006).

Following up on the results obtained in a GWAS initially performed on basal cell carcinoma and then extended to 16 additional cancer types, Rafnar *et al.* discovered two new SNPs (rs2736098 and rs401681) at 5p15.33 that increased the susceptibility for UCB. They are located in an LD-block that overlaps with *TERT* and *CLPTM1L*. The former encodes the catalytic subunit of the telomerase ribonucleoprotein complex, which adds telomeric repeat sequences at the end of the chromosomes. The latter is a predicted transmembrane protein expressed in a wide range of tissue carcinomas (Rafnar, Sulem et al. 2009).

Kiemeney *et al.* followed their first GWAS and extended it. A new SNP (rs798766) came up in 4p16.3 and was validated. This SNP is located in intro 5 of *TACC3*, which plays a role in the regulation of microtubules organization. The most interesting point is its close location at 70 kb from *FGFR3*. Activating mutations in this gene are the most common alterations in low-grade, NMIBC. A hypothetical link between these two genes has been considered in which germline variation in *TACC3* may lead to an overexpression of protein levels of *FGFR3*. It would make possible a higher rate of urothelial proliferation or an increased chance of mutation of this gene (Kiemeney, Sulem et al. 2010).

Rothman *et al*. performed the next GWAS of UCB in 2010. They identified four new susceptibility SNPs and confirmed four loci previously described on 3q28, 4p16.3, 8q24.21

and 8q24.3. The strongest signal came out for rs1014971 on 22q13.1 at 25 kb from *APOBEC3A* which plays a role on immunity, by restricting transmission of foreign DNA. However, no relation with carcinogenesis is known. The second SNP (rs8102137) resides into the genomic region of *CCNE1* on 19q12. Its transcript regulates the cell cycle control in the G1-S phase and its overexpression has been observed in many tumors, including UCB (Meyer 2004; Daly 2010). The third hit (rs11892031) is placed in an intronic *UGT1A* on 2q37.1 and plays a role in detoxification of endo- and xenobiotics through bile or urine by glucoronidation (Azzato, Pharoah et al. 2010; Sato, Yamamoto et al. 2011). Several gastrointestinal cancers and UCB have shown tissue-specific loss or reduced expression of *UGTs* (Huang, Heist et al. 2009; Penney, Pyne et al. 2010). This study also identified rs1495741 as a new locus mapping the well-known susceptibility gene *NAT2*.

Two simultaneous GWAS appeared on 2011 describing rs17674580 as a new susceptibility locus (Garcia-Closas, Ye et al. 2011; Rafnar, Vermeulen et al. 2011). It is located within *SLC14A1*, which is a urea transporter that regulates cellular osmotic pressure in kidney whereas it determines the Kidd blood groups in erythrocytes. In addition, Garcia-Closas *et al*. discovered another SNP perfectly correlated with the mentioned one (rs10853535) and a new independent one (rs7238033) located in the same gene.

The robustness of these susceptibility loci, reinforced with replicated results on different populations totaling thousands of individuals, contrast with the underpowered and usually non-replicated results in the candidate gene studies.

Table 1. UCB susceptibility loci reported after GWAS.

Study (year)	Cases (n): Controls (n)	SNPs (n)	Locus	Gene region	SNP ID	Risk allele	Allelic OR (95% CI)
Kimeney *et al.* (2008)	Discovery: 1803:34,336 Replication: 2165:3800	302,140 10	8q24.21	*MYC*	rs9642880	T	1.22 (1.15-1.29)
Kimeney *et al.* (2008)	Discovery: 1803:34,336 Replication: 2165:3800	302,140 10	3q28	*TP63*	rs710521	A	1.19 (1.12-1.27)
Rafnar *et al.* (2009)	Discovery: 4147:34,998 Replication: 3699:9076	1 1	5p15.33	*TERT*	rs2736098	A	1.16 (1.08-1.23)
Rafnar *et al.* (2009)	Discovery: 4147:34,998 Replication: 3699:9076	1 1	5p15.33	*CLPTM1L*	rs401681	C	1.12 (1.06-1.18)
Wu *et al.* (2009)	Discovery: 969:957 Replication I: 1713:3871 Replication II: 3985:34,762	556,429 50+10 1	8q24.3	*PSCA*	rs2294008	T	1.12 (1.06-1.18)
Kimeney *et al.* (2010)	Discovery: 1899:39,310 Replication: 2691:5959	304,073 12	4p16.3	*TACC3/FGFR3*	rs798766	T	1.22 (1.15-1.29)
Rothman *et al.* (2010)	Discovery: 3532:5120 Replication: 8381:48,275	589,299 100	22q13.1	*APOBEC3A*	rs1014971	C	0.88 (0.85-0.91)
Rothman *et al.* (2010)	Discovery: 3532:5120 Replication: 8381:48,275	589,299 100	19q12	*CCNE1*	rs8102137	C	1.13 (1.09-1.17)
Rothman *et al.* (2010)	Discovery: 3532:5120 Replication: 8381:48,275	589,299 100	2q37.1	*UGT1A*	rs11892031	C	0.84 (0.79-0.89)
Rothman *et al.* (2010)	Discovery: 3532:5120 Replication: 8381:48,275	589,299 100	8p22	*NAT2*	rs1495741	G	0.87 (0.83-0.91)
Garcia-Closas *et al.* (2011)	Discovery: 4501:6076 Replication I: 1382:2201	555,912 17	18q12.3	*SLC14A1*	rs17674580	T	1.16 (1.10-1.22)
Garcia-Closas *et al.* (2011)	Discovery: 4501:6076 Replication I: 1382:2201	555,912 17	18q12.3	*SLC14A1*	rs7238033	C	1.20 (1.13-1.28)

1.2.5. Genome-wide prognostic studies in the literature

The optimism regarding the susceptibility loci obtained using GWAS encouraged new analyses to discover novel loci that would assess the prognosis of different diseases. The pharmacogenomics pioneered the prognostic studies in the genome-wide field trying to identify loci that affect either drug toxicity or drug response (Meyer 2004; Daly 2010). As in the susceptibility studies, the candidate-gene approaches were the most common in the early days and they were mainly focused on drug metabolism, gene-coding drug targets and immune response. The availability of variability data on the human genome and its application using GWAS, made possible the assessment of differential treatment responses using genetic variations. The homogeneity of the patients and the expected large effects were considered as a basis to proceed with small sample sizes, thus one of the main weaknesses of the GWAS would be dodged. Because of this, a number of studies with moderate sample size were performed since 2007. However, problems usually arise when the drug responses cannot be quantified, when assessing an independent study to replicate the findings or when the effects are smaller that assumed. Despite of this, a bunch of pharmacogenomic GWAS have been published with statistically significant results and try to find their way into the clinical practice (Daly 2010; Pirmohamed 2011). Other studies tried to establish a link between the susceptibility and the prognostic value of replicated SNPs in GWAS. However, even in large studies that link remains unconfirmed (Fasching, Pharoah et al. 2012).

At the end of 2012 the number of genome-wide prognostic studies (GWPS) is almost anecdotic. Two recent reviews have mentioned the results obtained in four kinds of cancer (Gu and Wu 2011; Chang, Gu et al. 2012). An exhaustive search in Pubmed makes possible

to find at least five more studies[2]. In short, we are aware of nine studies assessing the prognosis of non-small-cell lung cancer (NSCLC) (Huang, Heist et al. 2009; Wu, Xu et al. 2010; Sato, Yamamoto et al. 2011; Wu, Ye et al. 2011), breast cancer (Meyer 2004), prostate cancer (Penney, Pyne et al. 2010), pancreatic cancer (Innocenti, Owzar et al. 2012), chronic lymphocytic leukemia (CLL) (Wade, Di Bernardo et al. 2011) and acute lymphoblastic leukemia (ALL) (Yang, Cheng et al. 2012). Prostate and breast cancer studies reported no SNP association at the genome-wide level after the replication phase. Nevertheless, the studies of NSCLC claim for genuine loci associated with overall survival but the borderline significance of most SNPs may suggest a critical evaluation of the results. The same claiming is done in the pancreatic cancer study for two SNPs that reach the genome-wide threshold of significance. However, the lack of replication in independent studies makes difficult to assume it as a genuine result. The study of ALL identifies an astonishing number of loci with prognostic value. However, the analysis design of this study based on an iterative "discovery vs replication" screening without any independent cohort to validate, may lead to an important amount of false positives. No genome-wide study designed to assess prognosis in UCB has been published until now.

The main limitations of these studies lay on the need of large cohorts with similar clinical characteristics and exhaustive patient information. Moreover, the common germ line variants with prognostic value may depend on the tumor subtype, the disease stage or the applied treatment. The lack of well-annotated patient data usually means small sample sizes, shortage of available information regarding the possible confounders and survival analyses whose only outcome is overall survival (Gu and Wu 2011).

[2] Pubmed Search: (genome-wide[Title/Abstract]) AND (cancer[Title/Abstract]) AND (cox[Title/Abstract] OR hazard[Title/Abstract]) AND (survival[Title/Abstract])

Table 2. Published genome-wide studies based on genetic variants assessing for prognosis in several cancers.

Study (year)	Cancer	Cases (N)	SNPs (n)	Loci* (N)	MoI	Outcomes
Huang *et al.* (2009)	NSC lung ǂ	Discovery: 100 Replication: 89	74,666 50	5	Additive	Overall survival
Azzato *et al.* (2010)	Breast	Discovery: 1145 Replication: 4335	262,264 10	-	Additive	Overall survival
Penney *et al.* (2010)	Prostate	Discovery: 637 Replication: 655	419,613 68	-	Additive	Overall survival
Wu *et al*. (2010)	NSC lung ǂ	Discovery: 245 Replication: 305	265,996 26	2	Additive	Overall survival
Sato *et al.* (2011)	NSC lung ǂ	Discovery: 105 Replication: -	109,365 -	3	Dominant Recessive Codominant	Overall survival
Wade *et al*. (2011)	CLL †	Discovery: 356 Replication: 380	346,831 10	3	Additive	PFS
Wu *et al*. (2011)	NSC lung ǂ	Discovery: 327 Replication I: 315 Replication II: 420	307,260 60 2	1	Dominant Recessive Additive	Overall survival
Innocenti *et al.* (2012)	Pancreatic	Discovery: 294 Replication: -	330,690 -	2	Additive	Overall survival
Yang *et al*. (2012)	ALL ‡	Discovery: 2532 Replication: 2532+2532	444,044 NA	134	Additive	Relapse

* Number of SNPs claimed to be statistically significant after multiple comparison adjustment.
ǂ Non-small cell lung cancer.
† Chronic lymphocytic leukemia.
‡ Acute lymphoblastic leukemia.
MoI, mode of inheritance; PFS, progression-free survival.

1.2.6. Gene set analysis in post-genome-wide results

Over the last decade, the performed GWAS have successfully identified many genetic variants associated their susceptibility (Hindorff, Sethupathy et al. 2009). However, the individual or even the combined effects of these variants explain just a small proportion of the risk associated to the disease (Manolio, Collins et al. 2009; Eichler, Flint et al. 2010). The sources of the "missing heritability" may be related (among others such as gene-gene interactions, gene-environment interactions, CNVs and rare variants) to the basic design of the GWAS: testing for the association between the disease phenotype and each SNP individually, even when small effects are expected (Hirschhorn and Daly 2005). It is also related with the fact that a large number of tests are performed in the GWAS and genuine, but weak, associations are missed after multiple comparison adjustments. Some proposals based on testing the joined effects of the SNPs have been suggested to overcome this limitation. Probably the most evident one is assessing for epistatic effects, but the number of SNP combinations in a genome-wide scenario and its mandatory adjustment for multiple comparisons makes the detection of real associations even more difficult. On the other hand, gene set analysis (GSA) analyses techniques applied in the assessment of gene-transcription studies are considered. The aim of the GSA is to increase the power by combining the signals from multiple SNPs that can hardly explain any risk on susceptibility or prognosis role independently, but explainable when grouped in biologically congruous groups (e.g. biological pathways, signaling pathways or protein-protein interaction networks).

The GSA workflow starts taking the SNP genotypes or its p-values obtained after a GWAS (or GWPS), assigning this information to the closest gene and running the statistical GSA

test itself. There are two families of GSA tests: the competitive tests, which compare disease associations for the genes in a gene set with the rest of the genes of the genome; and self-contained tests, which test the potential associations for the genes in a particular gene set. There is a great number of proposed algorithms for GSA but the statistical techniques underlying the most of the methods can be numbered in just a few: direct/modified Fisher's exact test, direct/modified Kolmogorov-Smirnov test, methods based on the Z-statistic or Z-score, the adaptive rank truncated product statistic, U-statistics and the SNP ratio test (Menashe, Maeder et al. 2010; Wang, Li et al. 2010; Fridley and Biernacka 2011; Wang, Jia et al. 2011).

Post-GWAS analyses using different GSA methods have been performed in a wide range of diseases that cover from several cancers to mental diseases (Holmans, Green et al. 2009; Medina, Montaner et al. 2009; Menashe, Maeder et al. 2010; Zhang, Cui et al. 2010). The application of these techniques to UCB is still in a preliminary phase in the assessment of susceptibility. Two gene set methodologies have been successfully applied in the assessment of the risk of developing UCB and show alterations in pathways regarding metabolic detoxification, clathrin-mediated vesicles and mitosis (Menashe, Figueroa et al. 2012). To our knowledge, no GSA has been conducted regarding prognosis.

Chapter 2: Hypothesis and objectives

General hypothesis:

Common germline variants are associated with clinical outcomes in UCB.

General objective:

To assess the role of genetic susceptibility in UCB evolution.

The objective will be accomplished through an agnostic GWPS and additional analyses based on the obtained results. In this kind of studies no SNP is thought, *a priori*, to have a higher probability of being associated with the evaluated clinical outcome than any other SNP.

Specific objectives:

1. To identify the independent SNPs associated with UCB clinical outcomes:
   1.1. SNPs independently associated with tumor recurrence, progression and relapse in NMIBC.
   1.2. SNPs independently associated with tumor progression, UCB-specific mortality, and overall survival in MIBC.

2. To identify the biological pathways associated with UCB outcomes.

3. To identify SNP-SNP interactions associated with UCB outcomes.

Chapter 3: Materials and Methods

A. Population and clinical & follow-up information

This work has considered different sources of UCB patients. The main sources of information are the Spanish Bladder Cancer (SBC)/EPICURO Study and the Texas Bladder Cancer (TXBC) Study. Both studies have are involved in the *International Consortium of Bladder Cancer*. They have recruited 1,150 and 1,542 patients with DNA samples, respectively. The use of the other sources of patients in validation stages is based upon collaborations that differ according to the group who leads the study. While the SBC/EPICURO Study leaded the inclusion of 5 retrospective cohorts in Europe and Canada (N=918), the TXBC provided a new prospective cohort of cases recruited at the same center (N=366). While we had full access to the information collected in the retrospective cohorts, only descriptive information and SNP prognostic estimates were shared between the two main studies. The full access to the retrospective cohorts' information made possible to perform pooled analyses. On the other hand, the restricted information shared with the TXBC Study only made possible to proceed with meta-analyses. Following is a description of each of the series.

A.1. Spanish Bladder Cancer (SBC)/EPICURO Study

Population

The Spanish Bladder Cancer (SBC)/EPICURO Study is a multicenter hospital-based case-control study which was conducted between 1997 and 2001. In order to study the prognosis of the disease, a nested cohort with the UCB patients (cases, N=1,356) within the

mentioned study was considered and followed-up at a yearly basis for more than 10 years. All incident UCB patients were treated in 18 hospitals located in 5 Spanish regions (Alicante, Asturias, Barcelona, Vallès Occidental/Bages and Tenerife). The participating hospitals were general or University-affiliated centers; none of them was a referral hospital specialized in urologic oncology. All patients gave written informed consent. The study was approved by the local Institutional Ethics Committees of each participating hospital and the Institutional Review Boards of the *Institut Municipal d'Investigació Mèdica* (IMIM) and the U.S. *National Cancer Institute* (NCI).

Information

All tumor-containing paraffin-embedded blocks produced at the time of patient's initial diagnosis were retrieved from the Department of Pathology of the participating hospitals and sent to the Coordinating Center (IMIM). From each block, a section was obtained and stained with H&E (hematoxylin and eosin). Diagnostic slides from each case were reviewed by a panel of expert study pathologists to confirm diagnosis and ensure uniformity of classification criteria across all cases. Tumors were staged and graded according to the criteria of the TNM classification and the WHO-ISUP (AJCC 1997; Mostofi, Davis et al. 1999). A panel of expert pathologists reviewed all paraffin-embedded slides of tumoral blocks in order to avoid heterogeneous classification based on the pathological assessment of tumors. They applied the same and most up-dated classification available at that time (WHO-ISUP 1999). Thereafter, information on stage, grade, and morphology of the tumor was correlated with that from the hospitals and the experts solved inconsistencies. Primary tumors from 68 (5.3%) cases could not been evaluated by the experts because, for a variety of reasons, it was not possible to obtain sufficient material for

pathological evaluation of tumors. As of the rest, 995 were finally classified as NMIBC and 283 as muscle invasive bladder cancer. Blood and/or saliva were collected in order to perform the genotyping. The details in this regard are explained below.

Clinical data gathering has been described in detail elsewhere (Puente, Malats et al. 2003). Briefly, information related to diagnostic procedures, stage, tumor characteristics, and first treatment was collected from medical records through reviews conducted by trained personnel using a structured questionnaire. Detailed macro- and microscopical tumor features were collected, as recorded in the hospital files, including number and location of masses, size, gross tumor appearance, mucosal appearance, tumor growth pattern, stage, grade, and histology of the largest mass. There was no attempt to treat patients uniformly at the various participating centers. Treatment management was categorized according to conventional criteria: transurethral resection (TUR) "alone", TUR+Bacillus Calmette Guerin (BCG), TUR+chemotherapy, TUR+BCG+chemotherapy, radical cystectomy, radiotherapy, and systemic chemotherapy. Treatment strategies different from those specified above were grouped under the "other" category.

The follow-up of the cases was conducted for more than 10 years. The information related to tumor recurrence and/or progression, change of management, and patient's vital status was collected annually from hospital records - using an *ad hoc* designed questionnaire - and through direct telephone interviews. Follow-up rate for NMIBC was 94%. Up to July 2007, mean follow-up period for the 995 patients with NMIBC who were "free of disease" was 82.7 months, ranging from 2.5 to 117.6 months, with a total of 13 (1.3%) deaths due to UCB recorded. According to hospital definition, 385 (38.7%) patients suffered at least 1 recurrence/progression of their tumors. MIBC patients were also followed up to July 2007;

mean follow-up period for the 235 patients with MIBC who were "free of disease" was 83.2 months, ranging from 50 to 105 months, with a total of 108 (46%) deaths due to UCB. At the end, 161 (68.5%) patients died due to UCB or any other cause.

Finally, we proceeded with 1,071 patients (N=836 NMIBC and N=235 MIBC) that had completed information for follow-up and genetic data. The relevant clinical and pathological variables for further analysis are shown in *Supplementary Table 1-3*.

A.2. Texas Bladder Cancer (TXBC) Study

Population

All cases were newly diagnosed, histologically confirmed, and previously untreated incident UCB cases recruited from The University of Texas MDACC and the Scott Department of Urology, *Baylor College of Medicine* from 1999 until present as previously described (Wu, Ye et al. 2009). Written informed consent was obtained from each participant before collection of epidemiological and clinical data and blood samples by trained MDACC staff interviewers. The response rate for cases was 92%. After diagnosis, all NMIBC patients were treated with TUR and followed with periodic cystoscopic examinations and intravesical treatment. This treatment consisted of either induction BCG (6 weekly instillations) or induction plus maintenance BCG according to *Southwest Oncology Group* (SOWG) protocol (induction BCG followed by instillations at 3, 6, and then every 6 months for 3 years). Approximately 90% of the patients in TXBC study were Caucasians. To limit the confounding effect of population substructure, we included only Caucasians in this study. The study protocols were approved by the Institutional Review Boards of MDACC and *Baylor College of Medicine*.

Information

Trained MDACC interviewers interviewed all cases of the TXBC study. Comprehensive epidemiological data on demographics, family history of cancer, and smoking status were collected. Blood sample was collected for DNA extraction at the end of the interview. Never smokers were patients who never smoked or had smoked less than 100 cigarettes in his or her lifetime. Ever smokers were patients who had smoked at least 100 cigarettes in their lifetimes. Former smokers were patients who had quit smoking at least 1 year prior to diagnosis. Current smokers were patients who were currently smoking or who had stopped <1 year prior to being diagnosed.

The clinical data for TXBC study was collected by trained MDACC chart reviewers on date of diagnosis, tumor size, tumor grade, tumor stage, tumor location, presence of Tis, number of tumor foci at diagnosis, intravesical therapy, dates of recurrence and progression events, systemic chemotherapy, radical cystectomy, pathologic findings at cystectomy, and mortality. The relevant clinical and pathological variables for further analysis are shown in *Supplementary Table 1-3*. All patients were followed-up with periodic cystoscopic examinations.

The TXBC study was divided in two subgroups. The data from the first one, known as TXBC-1, was used in the Discovery phase of the analyses. The second subgroup, known as TXBC-2, was used as a validation cohort in the prognostic assessment for independent SNPs performed in the NMIBC patients.

The validation set of TXBC (TXBC-2) consisted of 366 histologically confirmed NMIBC cases. These UCB cases were also obtained from MDACC, including additional cases from

our ongoing case-control study. We also included cases that were newly diagnosed (diagnosed within 1 year before referral to MDACC) and excluded from the ongoing case-control study because of previous treatment or recruited prior to the ongoing case-control study. The demographic and clinical data was collected as described above. The relevant clinical and pathological variables are displayed in and *Supplementary Table 2*.

A.3. International Series for NMIBC

Population

NMIBC cases from 5 international retrospective studies conducted in the Princess Margaret Hospital, Toronto, Canada; Aarhus University Hospital, Denmark; Hôpital Mondor, Créteil, France; Erasmus MC, Rotterdam; and the Human Genetics Foundation (HuGeF), Turin, Italy, were considered. Cases were recruited in each study according to different criteria. Only patients with available demographic and clinical data were included and an extensive review was performed to ensure the consistency of variables used in the analysis. Characteristics of each population are displayed in *Supplementary Table 2*. Heterogeneity reflects the differences through with patients are managed and treated in each centre.

Information

The data from the international series came from hospital-based studies. These studies have been conducted during long periods of time (from ~30 to 15 years of follow-up). The heterogeneity among them is very important because the main objectives to be accomplished for each one were different at the time of the design.

The cohorts recruited in *Hôpital Mondor* and *HuGeF* included only men, in contrast with the other series that included ~25% of women. The number of free-of-disease patients in

each center is also very different, suggesting heterogeneous kinds of patients. While the *Prince Margaret Hospital* in Toronto has 15% of free-of-disease patient, the other centers only have between 33 and 6.4% of these patients. When we explored the number and the kind of outcomes in each center, we observed a very low number of relapses in *HuGeF*, compared with all the other centers. This situation is remarkable because the tumoral pathology of these patients is not very different from the other centers. Regarding this point, we observed that the stage/grade and size and multiplicity of the tumors collected in *Hôpital Mondor* where usually higher than the ones collected in the other centers, suggesting a greater level of malignancy.

B. Genotyping

B.1. Genotyping in SBC/EPICURO GWAS Study

Blood was fractioned into serum, plasma, leukocytes, lymphocytes and erythrocytes. Leukocyte and saliva DNA were obtained as described elsewhere (Garcia-Closas, Malats et al. 2005). The initial genotype analysis was done on DNA derived from 2,191 blood samples (1,149 cases and 1,042 controls) and 185 buccal samples (42 cases and 143 controls). Genotyping was performed at the Core Genotyping Facility, National Cancer Institute, USA, using the Illumina HumanHap 1M probe BeadChip containing 1,072,820 markers (Rothman, Garcia-Closas et al. 2010).

Pre-genotyping quality control measures selected 1,149 blood samples and 42 buccal samples for genotyping. Completion rates >98% per individual study were required to estimate genotype clusters. SNP assays with locus call rates lower than 95% were excluded.

There were 178 duplicated samples (127 duplicates, 11 triplets and 3 quads) that yielded a concordance rate of 94%. The final delivered dataset (from both initial and repeat genotyping) included 2,424 results for 2,231 distinct individuals; 2,121 derived from blood samples, 105 from buccal samples and 5 from both blood and buccal samples.

We established the significance level threshold at $p < 1 \times 10^{-4}$ for a departure from Hardy-Weinberg equilibrium among controls to excluded SNPs for the prognostic study [R-package: **`HardyWeinberg`** (Weir 1996; Wigginton, Cutler et al. 2005)]. After quality control metrics, 998,347 SNPs were available for analysis in a 1,071-patient cohort (N=836 NMIBC and N=235 MIBC).

B.2. Genotyping in TXBC GWAS Study

Genomic DNA was extracted from peripheral blood lymphocytes by proteinase K digestion, followed by isopropanol extraction and ethanol precipitation and stored at −80°C (Wu, Gu et al. 2006). Genotyping for the TXBC data set was generated using the Illumina HumanHap610 chip containing 620,901 markers at MDACC (Wu, Ye et al. 2009). Detailed quality control measures were described previously (Wu, Ye et al. 2009). In short, all the subjects included in this study were Caucasians and had call rate >95%. Duplicated samples and population outliers were removed from the analysis. SNP call rate >95% criterion was applied. We further removed markers that deviated from Hardy-Weinberg equilibrium in the controls at $p < 1 \times 10^{-4}$. In the final analysis, we included 542,953 autosomal and mitochondrial SNPs for 496 NMIBC and 397 MIBC cases that passed strict quality control measures for the SNPs and subjects.

B.3. Genotyping in the Validation Phase of NMIBC

At MDACC, genotyping of the 57 SNPs identified by the discovery phase meta-analysis was conducted using TaqMan® SNP Genotyping assays (Applied Biosystems, Foster City, CA) in 366 patients. At CNIO, genotyping of the 47 SNPs for the international series was conducted using Taqman SNP Genotyping assays (Applied Biosystems, Foster City, CA), genotyping of 9 SNPs was conducted using Fluidigm Dynamic Array™ in 918 patients. The SNP rs4946483 failed both Taqman® and Fluidigm assay.

C. Endpoints of interest

After primary tumor diagnosis and treatment, patients may develop new bladder neoplasms (called tumor relapses) that can be considered as recurrences or progressions, and they may die because of cancer or other causes. For primary diagnosed NMIBC patients, we defined recurrence as the reappearance of a NMIBC following a previous negative follow-up cytoscopy while progression was defined as the development of a MIBC or metastatic disease. As indicated before, tumors relapse indicated the development of either recurrence or progression, whichever came first, after treatment of primary tumor and a period of improvement, corresponding to disease-free survival. For primary diagnosed MIBC patients, progression was defined as the reappearance of an advanced MIBC after a negative follow-up medical evaluation, this corresponding to progression-free survival; BC-specific mortality, when the event of death is caused exclusively by UCB, this corresponding to disease-free survival; and overall survival indicated the death by any cause.

Event-free and disease-free survival time for each endpoint was calculated from date of clinical intervention/diagnosis to the date of endpoint event or the date of last follow-up. In the SBC/EPICURO Study, it was assumed that any tumor event occurred between the date of the last medical visit in which the patient was free-of-disease and the date the event was detected. Hence, the intermediate point as interval censoring was used and the total time of follow-up from first clinical intervention was computed. Patients who did not present any event until the end of study, those lost to follow-up and those who died from other causes were censored either at last medical visit or at death.

At the end of follow-up (7/1/2007) we observed that 44% of the patients with NMIBC and 25% of patients with MIBC did not show any other event with a mean follow-up time of 117.6 and 83.2 months, respectively. The considered kinds of events were described before and their associated survival times were calculated from date of diagnosis to date of endpoint event or date of last follow-up.

The most intuitive model of UCB evolution would follow a linear sequence from the diagnosis of the primary tumor, to possible recurrence/s, followed by possible progression/s that may lead to death due to UCB. However, this hypothetical model was not always observed in our long-term follow-up where there are patients with NMIBC that recurred several times but did not progress and a small group of subjects that progressed so fast that no recurrence was reported. In this scenario the mean time to develop recurrences or progressions is rather similar, and it may reflect a non-described process of competing risk between these events. Patients free-of-disease at the end of study, those lost-to-follow-up, and those who died from other causes were censored either at last medical visit or at death.

Several survival times have been defined in order to characterize the kind of event. The time unit for these survival times was set to months:

- **Recurrence Free Survival (T_{RFS}) for NMIBC:** Time from primary diagnosis to first recurrence form.
- **Progression Free Survival (T_{PFS}) for NMIBC:** Time from primary diagnosis to first progression regardless of whether the tumor recurred before or not.
- **Disease/Event Free Survival (T_{EFS}) for NMIBC:** Time from primary diagnosis to any kind of event.
- **Progression Free Survival (T_{PFS}) for MIBC:** Time from primary diagnosis to first progression.
- **Disease Specific Survival (T_{DSS}) for MIBC:** Time from primary diagnosis to UCB death.
- **Overall Survival (T_{OS}) for MIBC:** Time from primary diagnosis to death due to any cause.

3.1. Independent SNPs associated with UCB clinical outcomes

3.1.1. Independent SNPs associated with NMIBC clinical outcomes

This analysis followed a design with a Discovery and a Replication phase. In the Discovery phase two independent large prospective cohorts of patients with NMIBC with long follow-up were included: the Texas Bladder Cancer (TXBC-1) and the SBC/EPICURO studies. Both studies prospectively followed patients up yearly and applied the same definitions as NMIBC patient outcomes.

Independent SNP analyses were conducted for each study using four genetic modes of inheritance (MoI: dominant, recessive, codominant, and additive) and unadjusted and adjusted Cox proportional hazard regression analysis (Cox 1972; Therneau and Grambsch 2000). Hazard ratios (HR), 95% confidence intervals (95% CI), and *p-values* were estimated. Distinct sets of adjusting covariates were used in TXBC-1, TXCB-2, SBC/EPICURO, partly due to the availability of these variables. We performed a Discovery meta-analysis to cross-validate results between TXBC-1 and SBC/EPICURO. Risk estimates (HRma, 95% CI, and *p-values*) were computed considering individual HR and their standard errors for 3,600 SNP models (150 most significant SNPs from each outcome x unadjusted/adjusted model x 4 MoI x 2 series) (Cooper and Hedges 1994). It yielded a number of 6,059 SNPs for the final combined SNP list for the cross-validation. The results of fixed or random effect models were reported as appropriate and displayed jointly with the coefficient of heterogeneity (I^2). A threshold of I^2 <30% was taken into account for the final selection of SNPs. Those with the most significant meta-results were subsequently selected for Validation.

In the Validation phase, a pooled analysis with the series from PMH, AUH, HM, EMC, and HuGeF was conducted using Cox regression with Firth's penalized likelihood for the outcomes of interest (Heinze and Dunkler 2008) and a validation meta-analysis was performed jointly with the TXBC-2. Finally, a combined meta-analysis resulting from both Discovery and Validation data was conducted to summarize all HR and *p-values*. The analysis workflow is summarized below.

The prognostic value for the resulting SNPs was displayed through Kaplan-Meier product limit method for each study (Kaplan and Meier 1958). The differences between categories of each variable were assessed using the log-rank test (Therneau and Grambsch 2000). Median follow-up time was calculated using reverse Kaplan-Meier estimator.

The concordance statistic (or c-statistic) was computed for each outcome without and with the SNPs for each participating cohort. In this way we can know how well the patients are classified in a binary prediction problem. The mentioned statistic is the most usual way to establish the discriminative ability of generalized linear regression models. When the outcome can be defined as binary variable, the c-statistic is equivalent to the area under the receiver operating characteristic (ROC) curve. In this kind of representation the sensitivity (true positive rate) is plotted against 1 – specificity (false positive rate). In all the described series we obtained the c-statistic through an initial step in which the Somers' D statistic with censored data is calculated (Harell 2001). The c-statistic overestimation in each model was controlled using 500 bootstrap samples in a process using the **`rms`** R-package.

Analyses done in TXBC Study applied STATA software (version 10.1, STATA Corporation, College Station, TX) and data manipulations were done using the PLINK

(version 1.03) and R (version 2.15.0). In the SBC/EPICURO Study, analysis was done using R (version 2.11).

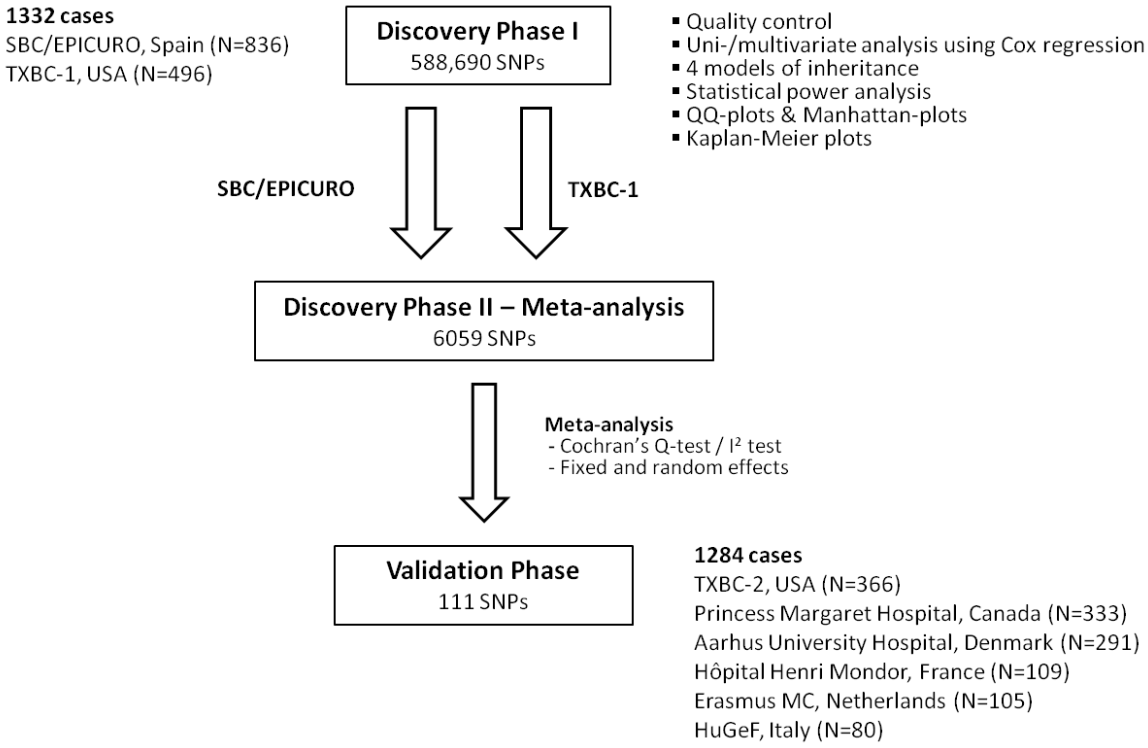


Figure 3. Genome-wide prognostic study analysis flowchart.

3.1.2. Independent SNPs associated with MIBC clinical outcomes

Similar to NMIBC, in the Discovery phase two independent large prospective cohorts of patients with MIBC with long follow-ups were included: the Texas Bladder Cancer (TXBC-1) and the SBC/EPICURO studies.

The statistical analyses were conducted using the same strategy described above for NMIBC patients. Overall, 4,412 SNPs remained in the final combined SNP list for the cross-validation in the Discovery phase. The results of fixed or random effect models were reported as appropriate and displayed jointly with the coefficient of heterogeneity (I^2). A threshold of I^2 <30% was taken into account for the final selection of SNPs. Finally 108

SNPs with the most significant meta-results were subsequently selected for a further the ongoing Validation phase that will be finished in the near future.

3.2. Biological pathways associated with UCB clinical outcomes

To detect the genetic pathways playing a role in the prognostic of UCB, we used the individual SNP results obtained from the SBC/EPICURO Study Cox regression analysis. The analyses were conducted for those NMIBC patients whose outcome was progression or recurrence; and the MIBC patients that shown progression or death due to UCB. We assigned to each SNP the minimal *p-value* among the 4 MoI obtained in the multivariate survival analysis performed in the SBC/EPICURO Study. Then, we selected a number of gene set analysis (GSA) methods that provided us a list of biological pathways significantly associated to the UCB outcomes were based on competitive tests. These tests take the list of SNPs and their associated *p-values* obtained in the GWPS and assess for a possible overrepresentation of significantly associated SNPs in the GWPS in the predefined gene sets (GS).

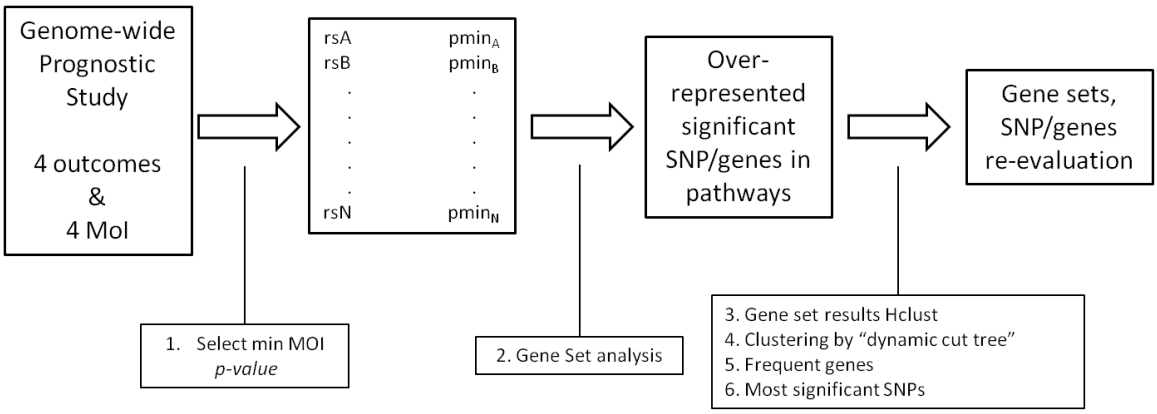


Figure 4. Gene set analysis flowchart.

The GSA methods are based on the comparison of the significantly associated SNPs' frequencies in a GS against the ones not present in the set. Several statistical algorithms

were studied: ALIGATOR (Holmans, Green et al. 2009) relying on a modified Fisher's exact test, GeSBAP (Medina, Montaner et al. 2009) using a segmentation test based on Fisher's (Al-Shahrour, Arbiza et al. 2007), GSA-SNP (Nam, Kim et al. 2010) based on the *Z*-statistic method, and *i*-Gsea*4*gwas (Zhang, Cui et al. 2010) that uses a weighted Kolmogorov-Smirnov running-test statistic (Fridley and Biernacka 2011). We also evaluated ICSNPathway (Zhang, Chang et al. 2011) because as far as we know, this is the only available method available to identify causal SNPs, genes and pathways in the genome-wide context. In the last three methods the GS information regarding Gene Ontology (GO) and Canonical Pathways (CP) was collected from MSigDB v3.0 (Subramanian, Tamayo et al. 2005). As a result, 1,454 GS in GO and 880 GS in CP were initially considered. In order to reduce the possible multiple test effect and to avoid uninformative narrow or wide broad GS, only those with 20-200 genes were used (Wang, Li et al. 2007). Applying this criterion 749 GO and 542 CP GS were finally selected. On the other hand, ALIGATOR used the predefined 6,723 GS, containing at least three genes in each category.

The complete list of SNPs and two linkage-disequilibrium-pruned (LD-pruned) SNP sets with stringent r^2 thresholds at 0.2 and 0.5 were analyzed. The pruning process was carried out taking groups of 1,000 consecutive SNPs in the genome and clustered applying the r^2 threshold of interest. In each cluster the most significant SNPs were selected. In the SNP-gene mapping we considered a scenario in which only the SNPs lying within a gene were kept and another one where those SNPs lying within 20 kb (5' or 3') of a gene were considered. The second criterion is based on the analysis of the location and the role of the eQTLs in the genome (Veyrieras, Kudaravalli et al. 2008).

ALIGATOR

The analysis was performed against GO for each of the outcomes. This method only maps the most significant SNPs into genes and tests for the putative enrichment within the predefined GO categories. In case that a SNP could be mapped within different genes, both would be included in the analysis. As far as this method only considers the most significant SNPs and the significance threshold is somewhat subjective *a priori*, three different thresholds were studied with *p-values* lower than 0.01, 0.001 and 0.0001. A fast screening analysis was performed with 5,000 gene lists resampling. Then, only the outcomes that showed significant overrepresented GO categories were reanalyzed and resampled with 50,000 gene lists.

GeSBAP

Three ranked lists of SNPs and their *p-values* were supplied for the mentioned LD thresholds. In each case, the algorithm only selects the SNPs that map into genes or closer than 5 kb to the nearest gene and selects the one with the lowest *p-value* as a proxy of the gene. An additional mapping process for these genes and the GS with 20-200 genes and 3-9 GO levels is carried out. Finally, a segmentation test based on sequential application of a Fisher's exact test is done in order to find an asymmetric distribution towards the extremes of the ranked list of genes generated in the intermediate step of this method (Al-Shahrour, Arbiza et al. 2007). The results are corrected by false discovery rate (FDR) (Benjamini and Hochberg 1995).

GSA-SNP

We have used an option based on the *Z*-statistic method. As far as it is based on PAGE (Parametric Analysis of Gene Set Enrichment), it uses the normal distribution to assess the statistical significance (Kim and Volsky 2005). Then, the *p-values* obtained for each GS are corrected using FDR. Only those GS with 20-200 genes were used in the analysis. In addition we used a three-step strategy analysis in order to maximize the false positives results removal. In the first step the SNP-gene mapping was performed considering the SNP with the second best *p-value* as the proxy of its gene; then the analysis was done. In the second step, the SNP with the best *p-value* acts as a proxy of the closest gene and the analysis is performed again. In the last step we kept only those common results in the previous two steps.

***i*-Gsea*4*Gwas**

This method is based on a variation of the Gene Set Enrichment Analysis (GSEA) adapted to the GWA studies of complex diseases (Wang, Li et al. 2007). In the first step, all the genes were ranked based on their significance. The second step tests whether the genes in a GS showed a higher significant rank when compared with the rest of the supplied genes. To reduce the possible multiple test effect and to avoid uninformative narrow or wide broad gene lists, only those GS with 20-200 genes were used (Wang, Li et al. 2007). The improved version of the method (*i*-GSEA) uses SNP label permutation and introduces the concept of significance proportion based enrichment score (SPES). The multiple test correction is achieved by applying FDR. The output interface for *i*-Gsea*4*Gwas offers the list of enriched GS whose FDR is lower than 0.25. The gene sets with FDR <0.05 are assumed as highly significant.

ICSNPathway

In the first stage of this method we selected the candidate causal SNPs by LD analysis of the most significant SNPs of the genome-wide analysis. We run the analyses using the default LD parameters for this method (r^2 = 0.8 in the HapMap CEU population) and two cut-offs for the most significant *p-values* at 10^{-5} and 10^{-4}. In the second stage the GS for the candidate causal SNPs are annotated by using the *i*-GSEA algorithm (Zhang, Cui et al. 2010). The significant gene sets were obtained after applying the FDR correction.

Post-analysis pathway evaluation

In order to have a more systematic view of the results, we checked the similarity or the enriched pathways identified. This task was performed through hierarchical clustering using the Euclidean distance and the Ward's minimum variance method (Ward 1963). In addition, we grouped the branches obtained in the clusters in order to interpret the results keeping into account the overrepresentation of some pathway subparts due to the lack of independence between the gene sets. We used the dynamic tree cut and the dynamic hybrid cut (Langfelder, Zhang et al. 2008) and the main assessment was done using the latter. The similarity was defined as the percentage of overlapping genes in the GS (Menashe, Maeder et al. 2010). Additionally, the GO similarities were assessed applying Lin's pairwise similarity and using the **`GOSim`** R-package when the GO terms were available (Lin 1998; Frohlich, Speer et al. 2007). We also checked whether the most frequent genes among the obtained pathways had been mapped to SNPs with significative main effects and the other way around; by checking whether the SNPs with the lowest *p-values* were located in fairly common genes in the enriched pathways.

3.3. SNP-SNP interactions associated with UCB clinical outcomes

The vast majority of the strategy and performance of the genome-wide interaction analysis (GWIA) was done by Jesús Herranz, the staff statistician of the Genetic and Molecular Epidemiology Group at the CNIO from 2010 to 2012. I played an active role in the design of the informatics platform to deal with the millions of interactions to be tested, the imputation of genotype missing values and the analysis of the results.

We assessed SNP-SNP interactions associated with UCB prognosis with the genetic and follow-up data from SBC/EPICURO Study. The analyses were conducted for those NMIBC patients whose outcome was progression or recurrence; and the MIBC patients that shown progression or death due to UCB. The Illumina HumanHap 1M platform was used in the genotyping tasks and those SNPs with a percentage of missing data >5% were removed from the analysis. Missing values of the remaining SNPs in the analysis were imputed by random forest algorithm, based in the values of the nearest SNPs. If the frequency of variant homozygous was <10 subjects for a given SNP, we considered the dominant model, pooling the heterozygous and variant homozygous. Still, if the frequency was <10 subjects after pooling the two categories, the SNP was removed from the analysis.

In the GWIA studies, the reduction of the number of genetic markers is a must for computational reasons. To this end, we selected the most representative SNPs for LD blocks taking a conservative threshold, r^2 >0.9, and prioritising the SNPs with less number of missing data. After having applied quality control filters and removing SNPs in LD, 585,220 and 552,220 SNPs were selected for the analysis for non-invasive and invasive UCB, respectively.

Table 3. Number of events and censored observations for each UCB outcome.

Outcome	Subphenotype	Sample size - N	Events - N(%)	Censored - N(%)
Recurrence	NMIBC	836	275 (32.9)	561 (67.1)
Progression	NMIBC	836	83 (9.9)	753 (90.1)
Progression	MIBC	235	129 (54.9)	106 (45.1)
BC-specific mortality	MIBC	235	108 (46.0)	127 (54.0)

Table 4. Number of SNPs included and interactions performance.

Subphenotype	No. SNPs	After remove by NA count	After remove by LD	No. interactions (millions)
NMIBC	998,349	682,741	585,220	171,241
MIBC	998,349	840,558	552,463	152,607

NA: non-available data
LD: linkage disequilibrium

The applied analytical strategy considered the following steps: Step 1- *Tuning and assessment*, aiming at defining the best way to analyze survival data with logistic regression and making a general assessment of the strategy; Step 2- *Screening* to select the most significant models fitted by logistic regression; Step 3- *Cox regression* models performed with those SNP-SNP interactions selected in step 2; Step 4- *Listing* the most relevant interactions by adjusting the Cox models for confounders, considering all modes of inheritance.

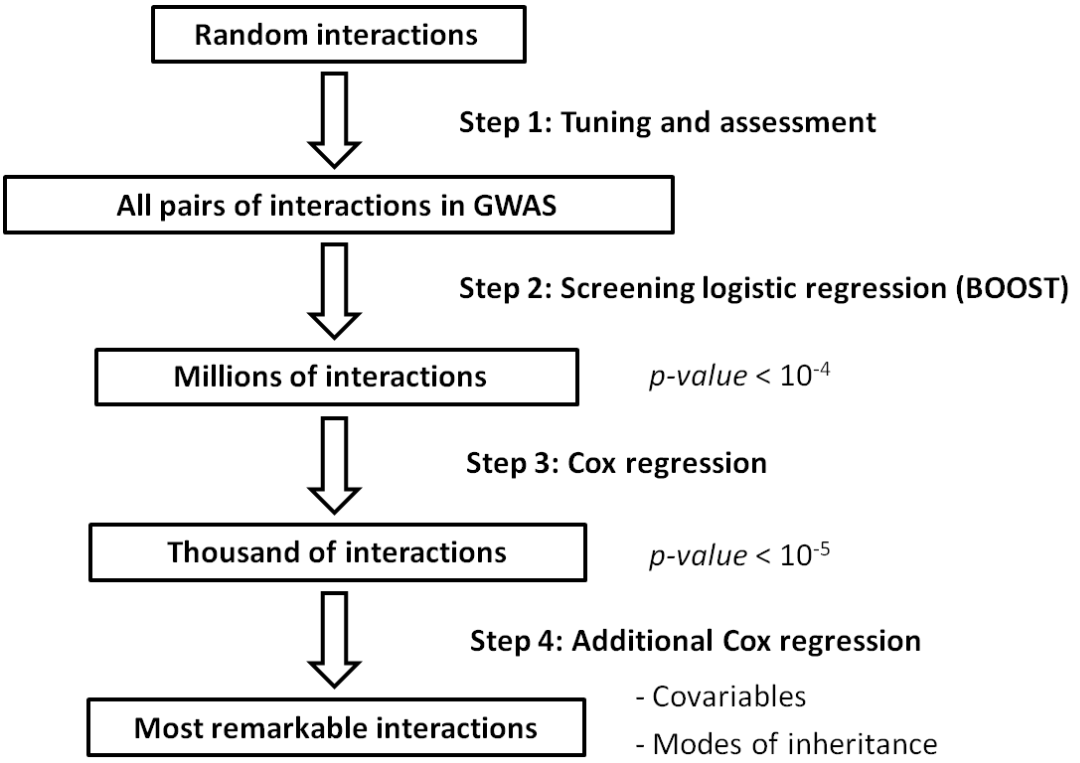


Figure 5. SNP-SNP interaction analysis flowchart.

Step 1 - Tuning and Assessment

It was capital to define the binary response variable (Y) in the logistic model using the follow-up time and to assess if the definition applied work reasonably well in our dataset. The simplest way to define the binary response variable is to assign 1 to the observations when the event occurred, and 0 if not. However, another more general definition of the binary response variable Y, depending on a survival time cut-off point was used in order to avoid situations in which the number of censored observations is small:

Y = 0 if status = 0

Y = 0 if status = 1 and the survival time > cut-off –point (Formula 1)

Y = 1 if status = 1 and the survival time ≤ cut-off point

In each survival dataset, different factors may affect the selection of the best cut-off point: sample size, percentage of censored observations and the distribution of the survival time. This study made possible to know if the approach based on logistic regression correlates with the results we would get with Cox regression.

In order to evaluate the interaction, we fitted two Cox regression models. The first model contained the main effects of the two SNPs, X_1 and X_2, and had the following formula:

$$\lambda(t/X) = \lambda_0(t) \cdot \exp(\beta_1 \cdot X_1 + \beta_2 \cdot X_2) \qquad \text{(Formula 2a)}$$

The second Cox regression model included both the main effects and the interaction terms as follows:

$$\lambda(t/X) = \lambda_0(t) \cdot \exp(\beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_{12} \cdot X_1 * X_2) \qquad \text{(Formula 2b)}$$

Let L_M and L_I be the log-partial-likelihoods of the main effect model and the interaction model, respectively. Then, we compared these models to evaluate the interaction, based on the difference L_I - L_M.

Similarly, two logistic regression models were proposed. First, the model with only the main effects of the two SNPs, X_1 and X_2:

$$\text{logit}(Y = 1/X) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 \qquad \text{(Formula 3a)}$$

The second logistic regression model included both the main effects and the interaction terms with the following formulation:

$$\text{logit}(Y = 1/X) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_{12} \cdot X_1 * X_2 \qquad \text{(Formula 3b)}$$

Let L_M and L_I be the log likelihoods of the main effect model and the interaction model, respectively. Again, we compared these two models to evaluate the interaction, based in the difference $L_I - L_M$.

We randomly selected a specific number of pairs of interactions from all the possible pairs and fitted the two Cox regression models and the two logistic regression models for each of these pairs. Then, we obtain two *p-values* from the two ratio likelihood tests, evaluating the interaction term with each regression technique. In order to compare the *p-values*, we changed them as $\log_{10}$ scale to check if the models with the lowest *p-values* from the two techniques were similar.

The relation between the *p-values*, as $-\log_{10}$ values, of the tests of the logistic and Cox regression was evaluated with the Pearson correlation coefficient. This analysis was repeated for different cut-off points of the survival time, defining different binary response variables in the logistic model. The cut-off point that had the highest correlation was then selected, since it is the best way to reproduce the results obtained by Cox regression with logistic regression. Finally, when the definition of the binary response variable was established, we performed a study selecting between up to 5 millions of random interactions.

Step 2: Screening with logistic regression

In the screening step, pair-wise interactions with all SNPs were analyzed by logistic regression applying the definition of the outcome binary variable as mentioned above. BOOST (Wan, Yang et al. 2010) was used to this end. BOOST uses a non-iterative method to approximate the likelihood ratio statistics (*Formula 3a* and *3b*) by evaluating all pairs of

SNPs. It selects those pairs that reach a specific threshold according to the likelihood ratio statistics, or equivalently, a threshold for the *p-value* of the tests. From a practical point of view, the choice of this threshold depends on the number of SNPs, and the computational availability in the posterior stages of the analysis. We could estimate roughly the number of selected models by logistic regression by multiplying the threshold by the total number of pair-wise interactions involved in our dataset. We suggest to select a small significant level to screen pair-wise interactions, for *p-values* between $<10^{-4}$ and $<10^{-6}$.

BOOST analyzes categorical variables coding in 2 or 3 categories meaning that it allows the analysis of either dominant, recessive or co-dominant mode of inheritance. We considered the latter to explore SNP-SNP interactions. Thus, SNPs were coded as 0, 1 and 2 for the common homozygous, the heterozygous, and the variant homozygous genotypes, respectively. It is worth noting that BOOST does not permit to include confounder variables in the analysis to adjust it for.

Step 3: Cox regression

It was used to evaluate the interactions identified in *Step 2* using the co-dominant mode of inheritance. The *p-value* for the partial log-likelihood ratio test was estimated to evaluate the interactions and was obtained from two Cox regression models, one with the main effects (*Formula 2a*) and the other including both the main effects and the interaction term (*Formula 2b*). We ended up with a list of the most important interactions using unadjusted Cox regression techniques to detect them.

In this step we also evaluated the potential interactions missed in the screening step (false negative results), those with significant interactions identified by Cox regression but not detected by logistic regression.

Step 4: Listing the most relevant interactions identified by Cox regression and confounder adjustment

We performed additional analyses assessing SNP-SNP interactions by adjusting for potential confounders/prognosticators and considering the recessive, dominant, additive and co-dominant modes of inheritance. It meant that both SNPs could be introduced in the model with any of the 4 modes of inheritance, making possible up to 16 combinations between a pair of SNPs.

Multiple Test

In order to avoid false positive results the level of statistical significance needed to be lowered. Among strategies allowing for multiple test correction, the Bonferroni correction assumes that the tests are independent. However, the presence of LD among the SNPs makes that many interactions tests may be highly correlated. In the context of GWAS with LD, this approach is accepted to be very conservative. The permutation test is an extended option to assess significance while allowing multiple test and correlations but we have discarded them because this approach is computationally prohibitive when analyzing several thousands of millions of interactions.

Only a few of strategies have been proposed to deal with the threshold of statistical significance in the GWIA studies. We referred to two of them, providing quite approximated results. The first one (Gao, Starmer et al. 2008) used the principal component

analysis (PCA) with a genotyped dataset to define the number of independent comparisons as the number of principal components explaining a large portion of the variance, usually 99.5%. This value is known as the "number of informative SNPs". The second one (Becker and Knapp 2004) suggested the use of the Bonferroni correction, adding a correction factor over the total number of interactions. Based in simulation studies in case-control studies for testing for allelic interaction, they propose a roughly approximation by dividing the original type I error obtained in the Bonferroni correction by a 0.4 factor. Becker indicated this correction factor is difficult to obtain when testing co-dominant models and he suggested that stronger correction factors could be appropriate.

The analysis of all modes of inheritance for the 2 SNPs introduces an additional correction in the multiple tests, because the tests analyzing different modes are not independent. Based on simulation studies, a 2.2 correction factor in the significant level has been suggested in order to calculate the effective number of tests (Gonzalez, Carrasco et al. 2008). Analyzing SNP-SNP interactions, the natural extension of this correction factor would be 4.84 (2.2×2.2) when all the combinations between the modes of inheritance for a pair of SNPs are explored.

Chapter 4: Results

4.1. Independent SNPs associated with UCB clinical outcomes

4.1.1. Independent SNPs associated with NMIBC clinical outcomes

A total of 2,616 patients with NMIBC were considered in this study. Characteristics of cases are summarized in *Supplementary Table 1* and *Supplementary Table 2*. Discovery population included 1,332 patients and Validation included 1,284 patients. In the TXBC, patients in the Discovery phase were recruited from 7/19/1999 to 3/6/2008 and the date of last follow-up was 10/24/2008. Out of 496 patients, 213 patients were free-of-disease and their median follow-up was 75.6 months. 57 patients were lost-to-follow-up (*Supplementary Table 1*). In the SBC/EPICURO Study, patients were recruited from 6/13/1998 to 6/28/2001 and the date of last follow-up was 7/1/2007. Out of 836 patients, 504 patients were free-of-disease and their median follow-up was 77.5 months. Only 9 patients were lost-to-follow-up (*Supplementary Table 1*). For the combined validation cohorts, patients were recruited from 1/1/1979 to 5/19/2010 and the date of last follow-up was 7/18/2011. Out of 1,284 patients, 486 were free-of-disease and the median follow-up was ranged from 26.3 to 113.0 months. The number of patients lost-to-follow-up was not available (*Supplementary Table 1* and *Supplementary Table 2*).

We observed similar distribution of main patient characteristics between TXBC-1 and TXBC-2 populations. Patients recruited in TXBC tend to have more aggressive disease than those from SBC/EPICURO because MDACC is a tertiary referral centre. Both studies had a low rate of progression ranging from 9.9%-17%. The international studies presented

heterogeneity in grade, multiplicity, tumor size, treatment variables, and rate of events (*Supplementary Table 1*). Among a total of 2,616 patients, 1,170 (44.7%) presented tumor recurrences, 380 (14.5%) tumor progressions, and 1,376 (52.6%) tumor relapses.

From now on, I am going to present and discuss the results obtained for the multivariate survival analyses. The univariate survival analysis will be considered only in the complementary assessment of SNPs using the Kaplan-Meier estimator. *Supplementary Figure 6* displays the Manhattan plots for the two Discovery studies for the three outcomes of interest with the *p-values* from the adjusted Cox regression models. In the Discovery phase, the meta-analysis of the combined TXBC-1 and SBC/EPICURO identified 57 SNPs significantly and independently associated with clinical outcome with meta *p-values* lower than 2.24×10^{-4} and no heterogeneity between studies (*Supplementary Table 4*). Among the 57 SNPs, 12 were associated with recurrence alone, 24 with progression alone, and 18 with relapse alone, 2 were associated with both recurrence and relapse, and one was associated with both progression and relapse. The distribution of these SNPs across the genome was uniform (*Figure 6*). These 57 SNPs were further followed-up in the Validation phase through genotyping 1,284 additional NMIBC cases from the TXBC-2 and 5 International series.

Risk of non-muscle invasive bladder cancer recurrence

Out of the 14 SNPs identified in the Discovery meta-analysis, SNP rs754799 (19p13.3, *Supplementary Figure 5a*) was replicated and significantly associated with recurrence in all the series (*Supplementary Table 4*). Patients with the homozygous genotype had a significantly increased risk of recurrence in the TXBC (HR= 3.00; 95% CI=1.51-5.93, $p=1.63\times10^{-3}$), SBC/EPICURO (HR= 2.68; 95% CI=1.63-4.42, $p=1.02\times10^{-4}$), and the

Validation series (HR=1.51; 95% CI=1.00-2.27, p=4.76×10^{-2}). The combined estimates with Discovery and Validation series provided a HR=2.06 (95% CI=1.55-2.75, p=7.45×10^{-7}) showing no significant heterogeneity among studies (*p-value* for heterogeneity = 0.21). The median recurrence-free survival time for individuals with the common allele was longer than for those carrying the rare homozygous genotype in TXBC (not computable versus vs. 4.41 months, logrank p=0.003) and in SBC/EPICURO (not computable vs. 17.1 months, logrank p=5.54×10^{-6}) (*Supplementary Figure 6*). The c-statistics without and with the SNPs were 0.64 and 0.65 in TXBC, 0.64 and 0.64 in the SBC/EPICURO, 0.68 and 0.68 in TXBC-2, and 0.67 and 0.66 in the International cohorts, respectively (*Supplementary Table 5*).

Risk of non-muscle invasive bladder cancer progression

SNP rs4246835 (*Supplementary Figure 5b*) was showed to be significantly associated with progression in all the series. Compared to the common homozygous genotype, patients carrying heterozygous genotype of rs4246835 had significantly reduced risk of progression: TXBC (HR=0.41, 95% CI=0.25-0.69, p=6.85×10^{-4}), SBC/EPICURO (HR=0.34, 95% CI=0.19-0.59, p=1.36×10^{-4}), and combined Validation data (HR=0.64, 95% CI=0.43-0.94, p=2.36×10^{-2}). The combined Discovery and Validation data yielded a HR=0.49 (95% CI=0.37-0.64, p=1.77×10^{-7}). Compared to the common homozygous genotype, subjects with the rare homozygous genotype had relative longer median progression-free survival time in TXBC and SBC/EPICURO (*Supplementary Figure 6*).

Five additional SNPs showed similar association with progression in the Discovery and in some of the Validation studies, although the Validation meta-analysis did not yield

significant results (*Table 6*). Of notice, all 6 SNPs had remarkable HR ($\geq$2 or $\leq$0.5) for the combined estimates from Discovery and Validation data and we observed no significant heterogeneity between studies for these SNPs in the combined estimates for Discovery and Validation data (*p* for heterogeineity >0.05). The c-statistics without and with these SNPs were 0.70 and 0.75 in TXBC, 0.77 and 0.84 for the SBC/EPICURO, 0.81 and 0.82 in TXBC-2, and 0.80 and 0.79 in the International cohorts (*Supplementary Table 5*).

Risk of non-muscle invasive bladder cancer relapse

SNP rs754799 was also replicated and significantly associated with relapse in the Validation data (*Supplementary Table 2*, *Supplementary Figure 5a*). Analyses of combined estimates from Discovery and Validation data indicated that patients harboring the rare homozygous genotype showed an increased risk of relapse compared to patients having the common allele (HR=1.89, 95% CI=1.43-2.49, $p=5.89\times10^{-6}$). rs754799 also showed the strongest association with recurrence and was the only SNP validated for both events. The median disease-free survival time was shorter for patients with the rare genotype in comparison to those harboring the common allele in the TXBC (4.41 months vs. 13.9 months, logrank $p=0.0293$) and in the SBC/EPICURO (17.1 months vs. not computable, logrank $p=1.13\times10^{-5}$) (*Supplementary Figure 6*).

Two additional SNPs, rs4946483 and rs11615759, also exhibited similar association with the risk of relapse in both Discovery and Validation data. The c-statistics without and with these SNPs were 0.63 and 0.64 in the TXBC, 0.62 and 0.63 in the SBC/EPICURO, 0.71 and 0.72 in the TXBC-2, and 0.67 and 0.67 in the International cohorts (*Supplementary Table 5*).

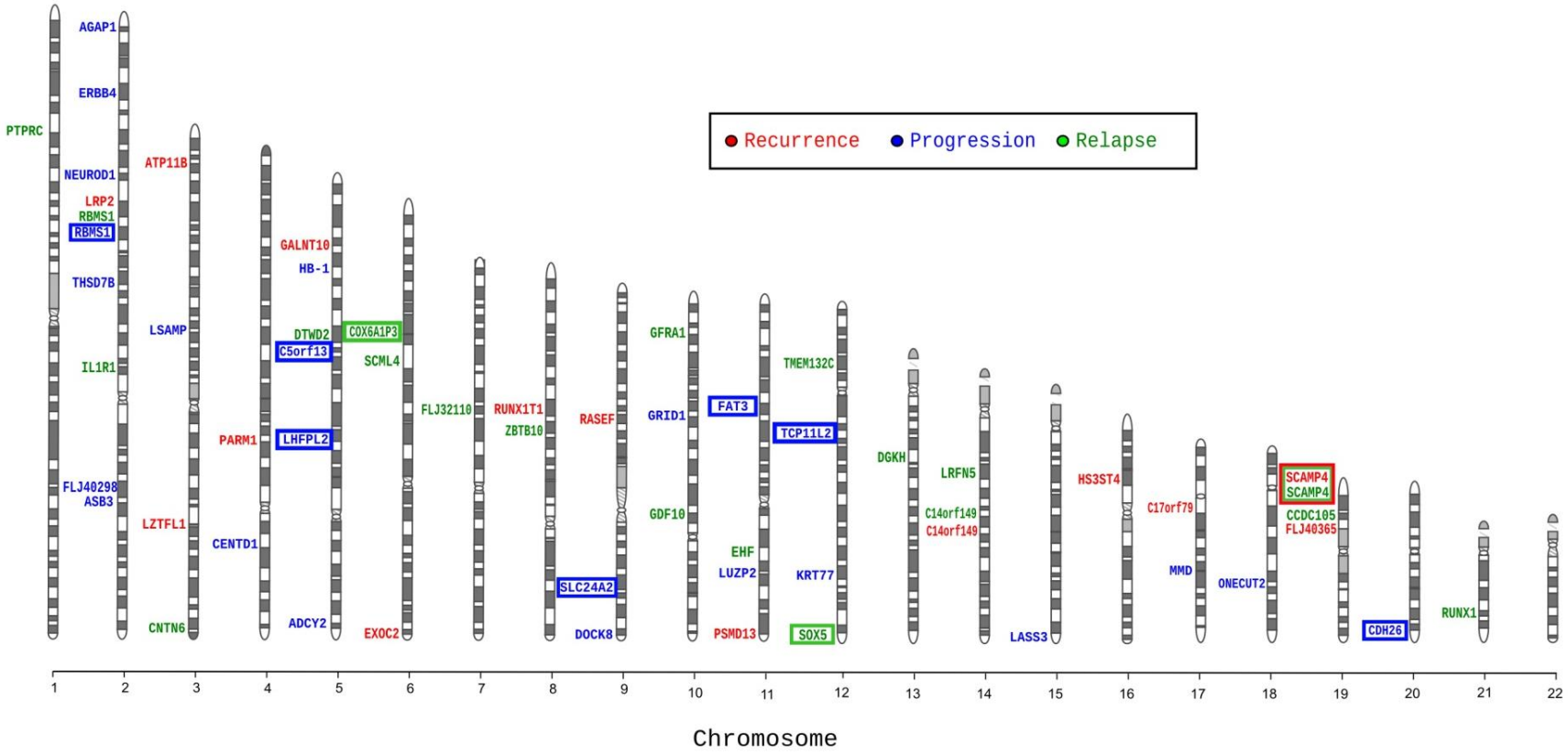


Figure 6. Chromosomal representation with the genomic location of the closest genes to the SNPs associated with non-muscle invasive UCB outcome (Recurrence, red; Progression, blue; and relapse, green) in Discovery analysis. Boxed are those SNPs that were replicated in Validation phase. The idiograms were plotted with the R/Biconductor package **`quantsmooth`** version 1.24.

Table 5. SNPs with lowest p-values associated with clinical outcome for non-muscle invasive UCB patients.

Marker [Alleles] Chr location gene*	Model	Study	Genotype Count	HR(95% CI)			P	P het
Recurrence								
rs754799[A/C]	recessive	TXBC	306/129/12	3.00	1.51	5.93	1.63×10^{-3}	
19p13.3		SBC/EPICURO	549/257/29	2.68	1.63	4.42	1.02×10^{-4}	
SCAMP4, ADAT3		Discovery	855/386/41	2.79	1.86	4.17	5.87×10^{-7}	0.798
		Validation	787/350/44	1.51	1.00	2.27	4.76×10^{-2}	0.840
		Combined	**1642/736/85**	**2.06**	**1.55**	**2.75**	$\mathbf{7.45\times10^{-7}}$	**0.213**
Progression								
rs4246835[G/A]	codom.het	TXBC	150/221/79	0.41	0.25	0.69	6.85×10^{-4}	
9p22-p13		SBC/EPICURO	312/372/152	0.34	0.19	0.59	1.36×10^{-4}	
SLC24A2, MLLT3		Discovery	462/593/231	0.38	0.26	0.55	3.72×10^{-7}	0.621
		Validation	441/538/196	0.64	0.43	0.94	2.36×10^{-2}	0.051
		Combined	**903/1131/427**	**0.49**	**0.37**	**0.64**	$\mathbf{1.77\times10^{-7}}$	**0.052**
rs6100810[A/G]	recessive	TXBC	303/134/13	2.95	1.03	8.48	4.46×10^{-2}	
20q13		SBC/EPICURO	554/252/30	5.65	2.62	12.2	9.73×10^{-6}	
C20orf197, CDH26		Discovery	857/386/43	4.51	2.43	8.40	1.95×10^{-6}	0.329
		Validation	843/305/37	1.55	0.70	3.47	2.83×10^{-1}	0.640
		Combined	**1700/691/80**	**3.03**	**1.85**	**4.95**	$\mathbf{9.76\times10^{-6}}$	**0.144**
rs7572970[G/A]	recessive	TXBC	221/195/35	3.78	1.98	7.23	5.65×10^{-5}	
2q24		SBC/EPICURO	434/334/67	2.28	1.18	4.38	1.35×10^{-2}	
RBMS1		Discovery	655/529/102	2.94	1.86	4.66	4.27×10^{-6}	0.279
		Validation	595/493/93	1.27	0.65	2.47	4.78×10^{-1}	0.689
		Combined	**1250/1022/195**	**2.24**	**1.54**	**3.27**	$\mathbf{2.88\times10^{-5}}$	**0.140**
rs12294567[A/G]	codom.hom	TXBC	355/88/8	4.08	1.08	15.4	3.83×10^{-2}	
11q14.3		SBC/EPICURO	632/183/21	6.14	2.68	14.1	1.79×10^{-5}	
FAT3		Discovery	987/271/29	5.48	2.71	11.1	2.17×10^{-6}	0.609
		Validation	907/259/19	0.56	0.10	3.07	5.06×10^{-1}	0.768
		Combined	**1894/530/48**	**3.92**	**2.05**	**7.52**	$\mathbf{3.76\times10^{-5}}$	**0.100**
rs17218455[G/A]	codom.hom	TXBC	325/110/16	5.03	2.30	11.0	5.25×10^{-5}	

12q32		SBC/EPICURO	572/243/21	2.91	1.17	7.22	2.11×10^{-2}	
TCP11L2		Discovery	897/353/37	3.98	2.20	7.20	4.89×10^{-6}	0.372
		Validation	818/326/40	1.32	0.57	3.04	5.19×10^{-1}	0.519
		Combined	**1715/679/77**	**2.75**	**1.70**	**4.47**	$\mathbf{4.08\times10^{-5}}$	**0.128**
rs3797725[A/G]	codom.hom	TXBC	305/131/15	2.77	1.02	7.54	4.57×10^{-2}	
5q22.1		SBC/EPICURO	497/294/45	3.38	1.82	6.28	1.16×10^{-4}	
C5orf13		Discovery	802/425/60	3.20	1.89	5.42	1.49×10^{-5}	0.741
		Validation	758/369/48	1.06	0.46	2.47	8.91×10^{-1}	0.329
		Combined	**1560/794/108**	**2.35**	**1.50**	**3.67**	$\mathbf{1.80\times10^{-4}}$	**0.122**
Relapse								
rs754799[A/C]	recessive	TXBC	307/130/12	2.31	1.17	4,54	1.56×10^{-2}	
19p13.3		SBC/EPICURO	549/257/29	2.43	1.52	3.88	2.15×10^{-4}	
SCAMP4, ADAT3		Discovery	856/387/41	2.39	1.62	3.51	9.92×10^{-6}	0.903
		Validation	787/350/44	1.48	1.00	2.20	4.91×10^{-2}	0.743
		Combined	**1643/737/85**	**1.89**	**1.43**	**2.49**	$\mathbf{5.89\times10^{-6}}$	**0.392**
rs4946483[G/A]	additive	TXBC	127/226/96	1.25	1.03	1.52	2.08×10^{-2}	
6q22		SBC/EPICURO	325/392/119	1.35	1.16	1.58	1.37×10^{-4}	
COX6A1P3		Discovery	452/618/215	1.31	1.16	1.48	1.00×10^{-5}	0.538
		Validation	102/122/65	1.16	0.91	1.49	2.29×10^{-1}	NA
		Combined	**554/740/280**	**1.28**	**1.15**	**1.43**	$\mathbf{6.89\times10^{-6}}$	**0.573**
rs11615759[A/G]	recessive	TXBC	331/110/8	2.51	1.13	5.56	2.32×10^{-2}	
12p12		SBC/EPICURO	589/219/28	2.66	1.64	4.33	7.47×10^{-5}	
SOX5		Discovery	920/329/36	262	1.73	3.96	5.02×10^{-6}	0.900
		Validation	887/284/13	1.06	0.52	2.15	8.77×10^{-1}	NA
		Combined	**1807/613/49**	**2.08**	**1.46**	**2.98**	$\mathbf{5.83\times10^{-5}}$	**0.095**

*Nearest gene to the SNP; NA: not available; Chr – Chromosome; P-het: *p-value* for test of heterogeneity; HR: hazard ratio; 95% CI: 95% confidence interval; Codom.hom: codominant.homozygote; Codom.het: codominant.heterozygote
TXBC: Texas Bladder Cancer Study; SBC/EPICURO: Spanish Bladder Cancer/EPICURO Study
Genotype Count: number of patients with homozygous common genotypes/heterozygous genotypes/homozygous rare genotypes
Discovery: combined analysis of Discovery populations from TXBC and SBC/EPICURO; Validation: combined analysis of Validation populations from TXBC-2, the Princess Margaret Hospital (PMH), Toronto; Aarhus University Hospital (AUH), Denmark; Hôpital Henri Mondor (HM), Créteil, France; Erasmus MC (EMC), Rotterdam; and the Human Genetics Foundation (HuGeF), Turin, Italy

4.1.2. Independent SNPs associated with MIBC clinical outcomes

A total of 632 patients with MIBC were considered in this study in the Discovery phase. The complementary analyses will be performed in the near future for the Validation phase. Characteristics of cases are summarized in *Supplementary Table 3*. In the TXBC, patients in the Discovery phase were recruited from 12/31/1997 to 4/25/2001; the date of last follow-up was 02/05/2009. Out of 397 patients, 200 patients were free-of-disease and their median follow-up was 43.7 months. 5 patients were lost-to-follow-up. In the SBC/EPICURO, patients were recruited from 4/25/1998 to 6/28/2001; the date of last follow-up was 7/1/2007. Out of 235 patients, 66 patients were free-of-disease and their median follow-up was 26 months. Only 3 patients were lost-to-follow-up.

The Manhattan plots for the two Discovery studies for the three outcomes of interest with the *p-values* from the adjusted Cox regression models are displayed in *Supplementary Figure 4*. In the Discovery phase, the meta-analysis of the combined TXBC-1 and SBC/EPICURO identified 57 SNPs significantly and independently associated with clinical outcome with meta *p-values* lower than 9.69×10^{-5} and no heterogeneity between studies (*Supplementary Table 7*). Among the 57 SNPs, 18 were associated with progression, 23 with death due to UCB, and 19 with overall survival; and 3 of them (rs2646727, rs2565721 and rs783145) were associated with both progression and BC-specific mortality. There was not any region with an overrepresentation of SNPs across the genome. These 57 SNPs are going to be validated in the same cohorts as the NMIBC part of the study. The details of the SNPs reaching or close to the Bonferroni multiple test correction threshold (*p-value* $<10^{-8}$) are described below.

Risk of muscle invasive bladder cancer progression

Out of the 18 SNPs identified in the Discovery phase meta-analysis, SNP rs16927851 (12p12.1) had the lowest *p-value* for the recessive MoI. The combined estimates in the Discovery series provided a HR=3.48 (95% CI=2.25-5.39, $p=2.08\times10^{-8}$). Patients with the homozygous genotype had a significantly increased risk of progression in the SBC/EPICURO (HR=2.79; 95% CI=1.58-4.94, $p=3.95\times10^{-4}$) and in the TXBC (HR= 4.75; 95% CI=2.41-9.37, $p=6.83\times10^{-6}$) (*Table 6*). No significant heterogeneity between the studies was detected (*p-value* for heterogeneity = 0.21). The median progression-free survival time for individuals with the common allele was longer than for those carrying the rare homozygous genotype in TXBC (not computable vs 30.8 months, logrank $p=4.92\times10^{-5}$) and in SBC/EPICURO (25.46 vs. 5.67 months, logrank $p=0.004$) (*Supplementary Figure 7*). The c-statistics without and with the SNPs were 0.67 and 0.71 in TXBC; and 0.69 and 0.76 in the SBC/EPICURO, respectively (*Supplementary Table 8*).

Risk of muscle invasive BC-specific mortality

Out of the 23 SNPs identified in the Discovery phase meta-analysis, SNP rs1015267 (11p14.2) had the lowest *p-value* for the codominant MoI. The combined estimates in the Discovery series provided a HR=3.96 (95% CI=2.51-6.23, $p=2.91\times10^{-9}$) showing no significant heterogeneity among studies (*p-value* for heterogeneity = 0.96). Patients with the homozygous genotype had a significantly increased risk of progression in the SBC/EPICURO (HR=3.91; 95% CI=2.03-7.54, $p=4.66\times10^{-5}$) and in the TXBC (HR= 4.00; 95% CI=2.13-7.51, $p=1.56\times10^{-5}$). The median survival time for individuals with the common allele was longer than for those carrying the rare homozygous genotype in TXBC

(not computable vs not computable vs 27.3 months, logrank p=0.013) and in SBC/EPICURO (not computable vs 33.9 vs. 23 months, logrank $p=3.6\times10^{-3}$) (*Supplementary Figure 7*). The same SNP appeared as significantly associated to BC-specific mortality for the recessive MoI; more information can be found in *Table 6* and in *Supplementary Figure 7*.

Four additional SNPs showed similar association with the risk of dying because of UCB in the Discovery studies, although the Discovery phase meta-analysis did not yield statistically significant results using the strict Bonferroni multiple test correction (*Table 6*). Of notice, all 4 SNPs had remarkable HR (between 2.50 and 3.86) for the combined estimates from Discovery data and no significant heterogeneity between studies for these SNPs was observed in the combined estimates for Discovery data (p for heterogeineity >0.05). The c-statistics without and with these SNPs were 0.73 and 0.76 in TXBC; and 0.76 and 0.77 for the SBC/EPICURO (*Supplementary Table 8*).

Risk of muscle invasive bladder cancer overall survival

Out of the 19 SNPs identified in the Discovery phase meta-analysis, none of them reached the stringent threshold to satisfy the multiple test correction. The most relevant SNPs were rs10437447 (10q26.2) and rs2565721 (6q26) for the dominant and the additive MoI, respectively. Nevertheless, we describe the details for these SNPs below.

The combined estimates for rs10437447 in the Discovery series provided a HR=1.88 (95% CI=1.47-2.41, $p=4.59\times10^{-7}$) (*Table 6*). No significant heterogeneity between the studies was detected (*p-value* for heterogeneity = 0.59). Patients with the homozygous genotype had a significantly increased risk of progression in the SBC/EPICURO (HR=2.02; 95%

CI=1.42-2.88, p=1.03×10^{-4}) and in the TXBC (HR= 1.76; 95% CI=1.25-2.48, p=1.11×10^{-3}). The median survival time for individuals with the common allele was longer than for those carrying the rare homozygous genotype in TXBC (60 vs 26.5 months, logrank p=5.67×10^{-3}) and in SBC/EPICURO (30.1 vs 22.9 months, logrank p=0.02). The combined estimates for rs2565721 in the Discovery series provided a HR=0.66 (95% CI=0.56-0.78, p=9.60×10^{-7}) (*Table 6*). No significant heterogeneity between the studies was detected (*p-value* for heterogeneity = 0.60). Patients with the homozygous genotype had a significantly increased risk of progression in the SBC/EPICURO (HR=0.69; 95% CI=0.54-0.89, p=3.57×10^{-3}) and in the TXBC (HR= 0.63; 95% CI=0.51-0.79, p=3.57×10^{-5}). In the *Supplementary Figure 7* we observe that the median survival time for individuals with the common allele was shorter than for those carrying the rare homozygous genotype in TXBC (20.1 vs 74.8 vs 85.1 months, logrank p=7.83×10^{-4}) and in SBC/EPICURO (18.4 vs 27.6 vs 36.9 months, logrank p=0.18).

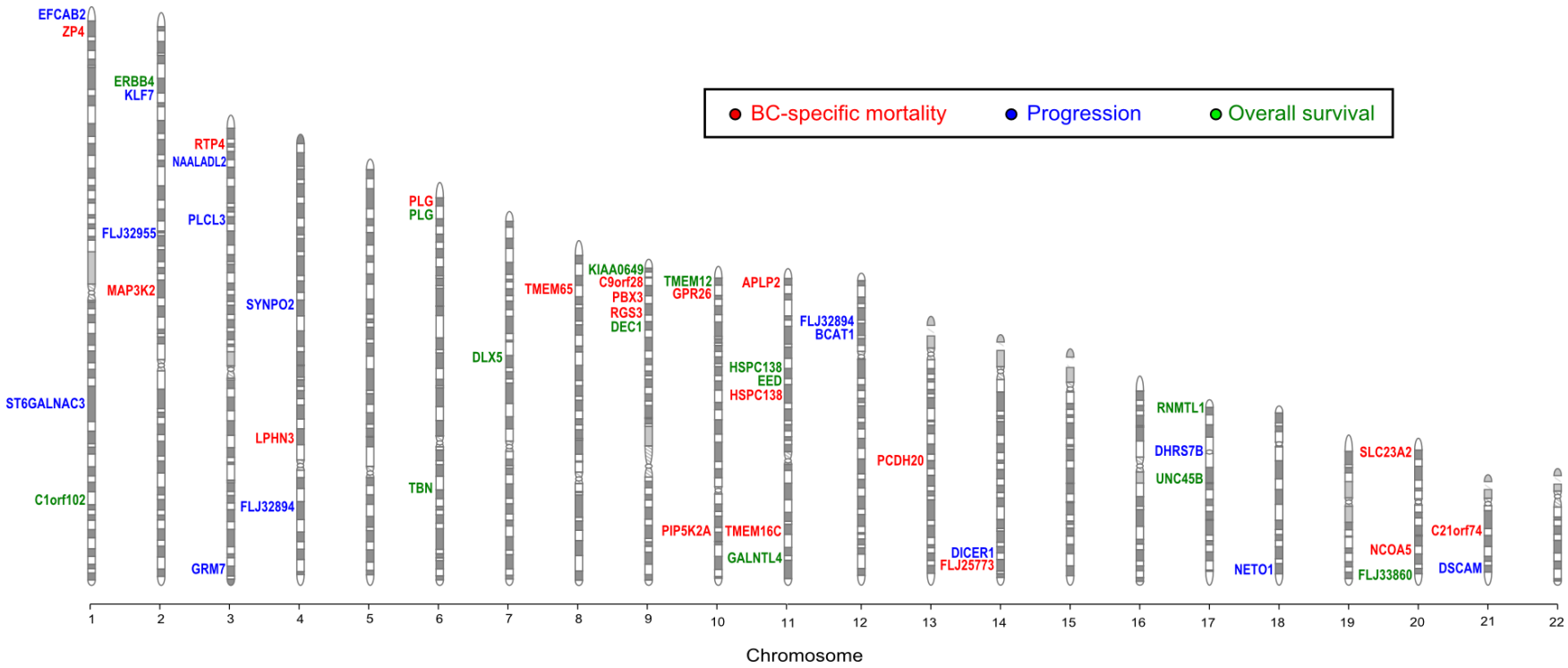


Figure 7. Chromosomal representation with the genomic location of the closest genes to the SNPs associated with muscle invasive bladder cancer outcome (BC-specific mortality, red; Progression, blue; and Overall survival, green) in Discovery analysis. The idiograms were plotted with the R/Biconductor package **`quantsmooth`** version 1.2.

Table 6. SNPs with lowest p-values associated with clinical outcome for muscle invasive bladder cancer patients.

Marker [Alleles] Chr location gene*	Model	Study	Genotype Count	HR(95% CI)			P	P het
Progression								
rs16927851 [A/G]	recessive	TXBC	231/131/22	4.75	2.41	9.37	6.83×10^{-6}	
12p12.1		SBC/EPICURO	125/87/22	2.79	1.58	4.94	3.95×10^{-4}	
BCAT1		**Discovery**	**356/218/44**	**3.48**	**2.25**	**5.39**	$\mathbf{2.08\times10^{-8}}$	**0.24**
rs9849682 [A/C]	recessive	TXBC	118/197/69	2.57	1.51	4.39	5.38×10^{-4}	
3q25.31		SBC/EPICURO	75/109/51	2.19	1.43	3.36	3.42×10^{-4}	
PLCL3		**Discovery**	**193/306/120**	**2.33**	**1.67**	**3.25**	$\mathbf{7.08\times10^{-7}}$	**0.64**
rs11732628 [A/G]	recessive	TXBC	141/182/60	2.19	1.21	3.97	9.91×10^{-3}	
4q26		SBC/EPICURO	93/110/32	2.92	1.78	4.79	2.04×10^{-5}	
SYNPO2		**Discovery**	**234/292/92**	**2.60**	**1.78**	**3.80**	$\mathbf{8.41\times10^{-7}}$	**0.46**
BC-specific mortality								
rs1015267 [A/G]	codom.hom	TXBC	182/169/32	4.00	2.13	7.51	1.56×10^{-5}	
11p14.2		SBC/EPICURO	96/114/24	3.91	2.03	7.54	4.66×10^{-5}	
TMEM16C		**Discovery**	**278/283/56**	**3.96**	**2.51**	**6.23**	$\mathbf{2.91\times10^{-9}}$	**0.96**
rs1015267 [A/G]	recessive	TXBC	182/169/32	3.66	2.04	6.59	1.41×10^{-5}	
11p14.2		SBC/EPICURO	96/114/24	3.16	1.73	5.78	1.76×10^{-4}	
TMEM16C		**Discovery**	**278/283/56**	**3.41**	**2.24**	**5.19**	$\mathbf{1.01\times10^{-8}}$	**0.73**
rs1008954 [A/G]	dominant	TXBC	344/38/1	2.78	1.60	4.83	2.79×10^{-4}	
11q24.3		SBC/EPICURO	201/33/1	2.72	1.62	4.59	1.65×10^{-4}	
APLP2		**Discovery**	**545/71/2**	**2.75**	**1.88**	**4.02**	$\mathbf{1.65\times10^{-7}}$	**0.96**
rs11221970 [A/G]	dominant	TXBC	346/36/1	2.77	1.57	4.87	4.18×10^{-4}	
11q24.3		SBC/EPICURO	203/31/1	2.78	1.64	4.71	1.54×10^{-4}	
APLP2		**Discovery**	**549/67/2**	**2.77**	**1.88**	**4.08**	$\mathbf{2.28\times10^{-7}}$	**0.99**
rs17603887 [G/A]	dominant	TXBC	313/63/6	2.09	1.28	3.40	3.14×10^{-5}	

10q26.13		SBC/EPICURO	194/38/3	3.02	1.83	4.98	1.54×10^{-5}	
GPR26		**Discovery**	**507/91/9**	**2.50**	**1.76**	**3.54**	$\mathbf{6.39\times10^{-7}}$	**0.30**
rs1537010 [A/G]	dominant	TXBC	365/18/0	3.52	1.70	7.30	7.21×10^{-4}	
9q33.3		SBC/EPICURO	217/16/1	4.25	2.04	8.87	1.17×10^{-4}	
C9orf28		**Discovery**	**582/34/1**	**3.86**	**2.30**	**6.49**	$\mathbf{3.15\times10^{-7}}$	**0.72**
Overall survival								
rs10437447 [G/A]	dominant	TXBC	284/97/3	1.76	1.25	2.48	1.11×10^{-3}	
10q26.2		SBC/EPICURO	168/61/5	2.02	1.42	2.88	1.03×10^{-4}	
TMEM12		**Discovery**	**452/158/8**	**1.88**	**1.47**	**2.41**	$\mathbf{4.59\times10^{-7}}$	**0.59**
rs2565721 [G/A]	additive	TXBC	106/189/89	0.63	0.51	0.79	3.57×10^{-5}	
6q26		SBC/EPICURO	74/108/53	0.69	0.54	0.89	3.57×10^{-3}	
PLG		**Discovery**	**180/297/142**	**0.66**	**0.56**	**0.78**	$\mathbf{9.60\times10^{-7}}$	**0.60**

*Nearest gene to the SNP; NA: not available; Chr – Chromosome; P het: *p-value* for test of heterogeneity; HR: hazard ratio; 95% CI: 95% confidence interval; Codom.hom: codominant homozygote; Codom.het: codominant.heterozygote
TXBC: Texas Bladder Cancer Study; SBC/EPICURO: Spanish Bladder Cancer/EPICURO Study
Genotype Count: number of patients with homozygous common genotypes/heterozygous genotypes/homozygous rare genotypes
Discovery: combined analysis of Discovery populations from TXBC and SBC/EPICURO

4.2. Biological pathways associated with UCB clinical outcomes

ALIGATOR results

In an initial fast screening process with 5,000 gene lists, three LD scenarios with two window sizes for the SNP-gene mapping and three *p-value* thresholds were considered. ALIGATOR offers the number of overrepresented GO categories reaching several levels of significance for overrepresentation, together with *p-values* that informs whether this number is greater than expected by chance. The lowest *p-values* were obtained in the NMIBC patients when the outcome was progression and the $p < 0.001$. For this outcome, we found significantly more enriched GO categories in the simulated gene lists. Thus, we focused on this outcome and the number of simulated gene lists was increased to 50,000 in order to get more accurate category-specific *p-values*. The enhanced results related the overrepresentation for this outcome is shown in the *Table 7*.

An excess of enrichment was observed in one of the most stringent scenarios by considering only the SNPs that lie within the genes with very low LD between them ($r^2 < 0.2$). In the *Table 7* is shown that in the mentioned scenario we obtained 54 categories with $p < 0.01$, based upon 569 autosomal genes with some SNP with independent predictive value with $p < 0.001$. We noticed that different LD thresholds cause a slight variation in the results. Nevertheless, a similar pattern of results was observed with more inclusive LD thresholds. However, the higher was the number of SNPs included in the analysis the lower was the observed significance. It would suggest that the performed analyses were at least mildly sensitive to random variation. The results were considered by joining all the significantly enriched categories, ignoring the fact that we dealt with different LD

thresholds. We got more putative enriched categories involved in the prognosis of the disease at the cost of including probable false positives.

We kept the enriched categories that had p <0.01 in these three different LD scenarios because the results obtained with less restrictive enriched *p-value* thresholds (e.g. p <0.05) showed a lack of statistical significance. The results obtained with LD r^2 <0.2 and LD r^2< 0.5 were very similar (the first one involved 54 categories, the second one 51 and they share 37) but when compared with the non-LD pruned list of SNPs (53 categories) there were 27 common enriched categories. When all the enriched categories obtained in the three LD situations were joined, we obtain 89 different sets. The most significant enriched categories revealed similarities between them after the clustering analysis. When those categories were analyzed in detail, a clear implication of several elements related to the inflammatory and the immunological response arose.

Table 7. ALIGATOR results for NMIBC patients with progression. The analyses were performed with 50,000 simulated gene lists for the SNPs that lie within the genes.

		Count	***p* < 0.05**		***p* <0.01**		***p* <0.001**	
LD	**GWPS *p***	**Significant genes**	**No. cat**	***p***	**No. cat**	***p***	**No. cat**	***p***
	0.001	175	135	0.82	34	0.10	2	0.45
LD 0.2	0.01	569	166	0.15	54	**0.03**	10	**0.03**
	0.1	2,232	183	0.26	30	0.38	2	0.54
	0.001	142	218	0.13	28	0.45	3	0.39
LD 0.5	0.01	483	162	0.12	51	**0.04**	7	**0.07**
	0.1	2,065	114	**0.10**	22	0.26	0	1.00
	0.001	211	152	**0.06**	33	0.14	5	0.15
Non-LD	0.01	741	190	0.15	53	**0.06**	3	0.39
	0.1	2,936	255	**0.06**	53	0.08	3	0.43

The most frequent genes among the obtained categories after joining the results of the different LD scenarios were *SYK*, *CD74*, *IL6* and *CD24*. *SYK* and *CD74* appeared in 35 and 30 enriched categories respectively. Meanwhile, *IL6* and *CD24* both appeared in 28 categories. None of them were related to SNPs with significant predictive values, therefore their putative role in the prognosis of the disease would be ignored in the preliminary SNP-independent genome-wide analysis. When all the SNPs in the obtained GS were evaluated, *ITGA4* came up with the lowest *p-value* (rs2305586 in the recessive MoI with $p=6.1\times10^{-11}$) and appeared in 5 categories associated with hematopoietic or lymphoid processes.

Additionally, we obtained enriched results close to the significance level in NMIBC patients when the outcome was recurrence when a LD r^2 <0.5 threshold was defined using a 20 kb window in the SNP-gene assignation. However, further analyses were not considered because the significance of the number of overrepresented categories was greatly decreased when different LD and SNP mapping criteria were evaluated.

GeSBAP results

The analyses were carried out using the three LD mentioned criteria for all the outcomes. The effect of the SNP mapped to a specific gene is not evaluable if the SNP is not within the gene or 5 kb apart from it. When the SNPs were evaluated without any LD restriction, we obtained a single enriched GS for "protein export" for the patients that die due to BC. With more restrictive LD values we observed enriched GS for almost all the possible outcomes. The most obvious observation is the lack of outcome specificity for the most significative enriched GO categories that were obtained. In most of the outcomes the role of small GTPase mediated signal transduction, calcium ion transport, cell adhesion molecules

and neuronal development seemed to be important. Despite of this, protein glycosylation categories in NMIBC appeared as distinctive GS for this outcome.

In NMIBC patients with progression, *TSC1* was the most frequent among all the significantly enriched GO categories, as far as it appeared in 6 of 9 GO categories. The most frequent genes in the other outcomes appeared in one third of the cases, at most. In this scenario we had a calcium channel subunit (*CACNA1B*) and a cholinergic receptor (*CHRNA7*) in MIBC, whose outcome was death, and *B4GALT1* for NMIBC with recurrences. None of the SNPs lying within these genes came up to statistical significance. When we looked for the SNPs with the lowest *p-values* in the GS that reached significance, a pair of SNPs for progression in MIBC and NMIBC came up. In the MIBC outcome were rs683004 (p=4.2×10^{-8} in the dominant MoI) and rs620508 (p = 6.4×10^{-8} in the dominant MoI), both mapped to *CDH6*. However, this gene appears only in one GS regarding cell junction organization. In the NMIBC outcome were rs2305586 (p=6.1×10^{-11} in the recessive MoI) and rs2036268 (p=2.5×10^{-8} in the recessive MoI), mapping to *ITGA4* and the RhoGEF and PH domain containing 4 (*FGD4*) respectively. The first one appeared only once in the CAMs pathway and the second one appeared in 4 GS associated with signal transduction in Rho/Ras GTPases.

GSA-SNP results

In this method we found remarkable the high number of statistically significant gene sets obtained after the analyses as well as the similarities among the outcomes. The role of the different LD (without restriction, r^2 <0.2 and 0.5) and the SNP-gene distance mapping (SNPs within the gene and those extended to 20 kb in the up-downstream) were evaluated. We obtained 34 common GS when the results for all the outcomes where compared. That

meant about 1/4 of the results for NMIBC with progression and 1/3 in MIBC and NMIBC with recurrence. The results trended towards the same kind of results obtained using the GeSBAP method. Only a few general differences arose: those GS regarding cardiomyopathies, the protein tyrosine kinase activity, the role of platelet activation, a more marked effect of neurogenesis and axonogenesis and the regulation of *GLP-1*, free fatty acids or acetylcholine.

The most frequent genes in all the outcomes were calcium channel, voltage-dependent subunits, especially *CACNA1C* and *CACNA1D*. They appeared in approximately 25% of the GS; except in NMBC cases with progressions where its presence dropped to 20%. Interestingly *KCNIP2* came out in the outcomes with progression with the same frequency as the other two mentioned genes. None of the SNPs located in those genes behaved as a putative allele risk.

The SNPs with significant main effects were associated with pathways regarding the MIBC and NMIBC patients that suffered progression. In the MIBC outcome came up the same 2 SNPs observed for this event in GeSBAP. On the other hand, 8 SNPs arose in the NMIBC scenario. The most significant one was rs2305586 mapped to *ITGA*4 and it was involved with 8 pathways regarding cell-cell interactions and cardiomyopathy (rs2305586, as described in GeSBAP). A similar putative risk was observed for rs34050907 and rs12761617 in the dominant MoI (p=7.1×10^{-10} and p=9.84×10^{-10} respectively) assigned to *RHOBTB1*. However, this gene was obtained just for the Rho GTPase cycle. Regarding *FGD4*, the same SNP and GS were obtained, when compared with the GeSBAP analysis. And an additional SNP (rs17036321, associated with *PPARG*, with p=8.7×10^{-8} in the

dominant MoI) suggests a putative association of nuclear receptor transcription with progression in NMIBC.

i-Gsea*4*Gwas results

We performed the analyses evaluating all the BC outcomes against the GO and CP gene sets. The effects of the LD (r^2 <0.2 and 0.5) and the SNP-gene distance mapping (SNPs within the gene and within 20 kb in the up-downstream of the closest gene) were also inspected in this method. After the SNP-gene assignation, we observed that only around 11k and 13k genes (LD r^2 <0.2 and 0.5, respectively) were considered in the different LD scenarios. That probably leads to an important loss of power, but the LD bias and the important background noise is minimized at this cost. When the GSA results were evaluated, we realized that this method may be quite sensitive to the considered LD thresholds and the SNP-gene distance association. The method offered high confidence results when we analyzed the data from the NMIBC and MIBC patients whose outcome was recurrence. They suggested possible alterations in the regulation of the lymphocytes and T cells, the biosynthesis of cytokines and effects on tumor suppressor genes such as *CDH1* and *p53*. Much milder effects were observed in NMIBC with progression regarding oncogenic pathways in which *FGFR* and *PI3K* are involved. The results seemed to go in another direction when the MIBC patients that show progression were evaluated. In this outcome the nuclear membrane transport seemed to be altered together with the *NOD* like receptor signaling pathway. However, these last results had far less statistical significance than the ones relevant to the NMIBC outcomes.

The most frequent element in the GS that show overrepresentation of risk alleles in NMIBC with recurrence was *IL12B*. It appeared in 11 of the 39 enriched GS associated to this

event; however none of the SNPs mapping this gene had high predictive values. Other genes related to the immune system such as *B7-H3*, *CD28* and *EBI3* appeared in 10 of those GS in the mentioned outcome and the lack of predictive values of their SNPs was also observed. In the MIBC with progression we saw that 21 genes appeared in 5 of 11 GS. These genes are part of the nucleoporin (NUP) family. None of the mentioned genes were associated with SNPs acting as putative risk alleles. The SNP with the lowest *p-value* mapping any gene involved in the enriched pathways was located in *ITGA4*. It was rs2305586 (p=6.1×10^{-11} for NMIBC in progression for the recessive MoI) and it appeared in the leishmania infection pathway.

In addition, this method makes possible to avoid the possible bias due to the major histocompatibility complex (MHC) inclusion in the analyses. A new round of analyses were performed masking the genes of this region (Horton, Wilming et al. 2004). When the results of with and without the MHC region were compared, we found out a decrease in the number of enriched GS (from 56 to 43) when the MHC region is masked. In most of the outcomes the enriched GS obtained in the MHC-masked analyses, were also obtained without that restriction. The only exception happens when this region is masked in the patients with MIBC whose outcome is progression, the LD r^2 threshold is 0.2 and the SNP-gene mapping distance is 20k using the GO categories. We obtained 4 low significant new GS related to intracellular transport and signal transduction when the MHC region was masked.

ICSNPathway results

Highly significant GS were obtained in patients with NMIBC and milder effects were observed in MIBC patients that die due to BC. When the different scenarios for LD,

genome-wide significance threshold, SNP-gene mapping and FDR levels of correction were considered we obtained 16 gene sets for the NMIBC patients in progression, 4 for those enduring recurrences and 10 for the deceased patients. Only two gene sets arose when the most restrictive threshold for genome-wide analysis and FDR were considered (*p-value* $<10^{-5}$ and FDR <0.05): the GTP binding and the vasopressin regulated water reabsorption in NMIBC patients that suffered progressions. When the threshold of genome-wide significance was loosen to $<10^{-4}$ and the FDR was kept at <0.05, a remarkable overrepresentation for GS regarding structural organization, tissue development and vasopressin regulated water reabsorption was obtained for patients with NMIBC that experienced progressions.

The most frequent gene in the obtained GS for the NMIBC patients that showed progression was *LIMA1*. It appears in 5 of the 28 enriched GS and it is present in GS regarding structural cell organization. None of the genotyped SNPs in this gene reached significance in the initial genome-wide analysis. Once we selected the most significant SNPs included in the obtained GS, we observed 2 SNPs that may act as risk alleles in those patients with progression in NMIBC. The first one (rs7113416 with $p=1.5\times10^{-13}$ in the recessive MoI) was mapped to *FANCF* in the DNA reparation pathway. The second one was associated to *ITGA4* (rs2305586 with $p=6.1\times10^{-11}$ in the recessive MoI), which is related to focal adhesion and the ECM receptor interaction.

Table 8. Summary of the enriched and most representative gene sets in the NMIBC cases developing recurrences (corrected *p-value* at 0.05). The gene sets were grouped in several modules according to the hierarchical cluster trees and the modules defined by dynamic tree cut method. ALIGATOR did not offer any significant result**.**

***i*-Gsea*4*Gwas**	**GeSBAP**	**GSA-SNP**		**ICSNPathway**
Module A - CDC20 phospho APC mediated degradation of cyclinA - Autodegradation of CDH1 by CDH1 APC **Module B** - Regulation of cytokine biosynthesis - Cytokine metabolic process **Module C** - Regulation of T cell activation - Regulation of lymphocyte activation - T-cell activation **Module D** - G protein coupled receptor binding	**Module A** - Di-, tri- valent inorganic cation transport - Calcium ion transport **Module B** - Response to UV - Response to light stimulus **Module C** - Protein amino acid N-linked glycosylation - Glycoprotein biosynthetic process **Module D** - Phosphatidylinositol signaling system - Calcium signaling pathway **Module E** - Regulation of small GTPase mediated signal transduction - Exocytosis - Cell-cell adhesion - Axon guidance - Long-term depression** - Neurotransmitter transport**	**Module A** - Calcium channel activity - Voltage gated potassium channel complex - Voltage gated channel activity - Voltage gated cation channel activity - Gated channel activity - Substrate specific channel activity - Ion channel activity - Cation channel activity - Cation transmembrane transporter - Metal ion transmembrane transporter activity **Module B** - Neurogenesis - Generation of neurons - Neuron differentiation - Axon guidance - Axonogenesis - Cellular morphogenesis during differentiation - Neurite development - Neuron development **Module C** - PTK activity - Transmembrane receptor PK activity - Transmembrane receptor PTK activity **Module D** - Hematopoietin interferon classD200 domain cytokine receptor activity - IL binding - Cytokine binding	**Module E** -Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell - Transmission across chemical synapses - Glutamate receptor activity - Synaptic transmission **Module F** - Enzyme linked receptor protein signaling - Transmembrane receptor PTK signaling **Module G** - NRAGE signals death through JNK - Rho GTPase cycle - Focal adhesion - ECM receptor interaction - Cell migration - Central nervous system development - Synapse - Axon guidance - Calcium signaling - Vascular smooth muscle contraction - Integrin binding - CAMs - Structural constituent of muscle - Muscle cell differentiation - Membrane organization and biogenesis **Module H** - Dilated cardiomyopathy - Hypertrophic cardiomyopathy - Arrhythmogenic right ventricular cardiomyopathy **Module I** - Axon guidance - NCAM1 interactions - NCAM signaling for neurite out growth	**Module A** - Carbohydrate binding

Table 9. Summary of the enriched and most representative gene sets in the NMIBC cases developing progression events (corrected *p-value* at 0.05). The gene sets were grouped in several modules according to the hierarchical cluster trees and the modules defined by dynamic tree cut method *i*-Gsea*4*GWAS did not offer any significant result.

ALIGATOR	GeSBAP	GSA-SNP		ICSNPathway
Module A	**Module A**	**Module A**	**Module H**	**Module A**
- Hemopoiesis	- Regulation of small GTPase mediated signal transduction	- Voltage gated channel	- Transmission of nerve impulse	- GTP binding
- Leukocyte differentiation	- Rho protein signal transduction	- Voltage gated cation channel	- Transmembrane receptor PTK signaling	**Module B**
Module B	- Regulation of Rho protein signal transduction	- Gated channel	- Muscle cell differentiation	- Actin filament organization
- Lymphocyte activation	- Regulation of Ras protein signal transduction	- Substrate specific channel	- Cytoskeletal protein binding	- Actin filament binding
- Lymphocyte proliferation	**Module B**	- Ion channel	- Sensory perception	**Module C**
- Regulation immune system	- Calcium ion transport	- Cation transmembrane transporter	- PTP activity	- Tissue development
Module C	- Calcium signaling pathway	- Cation channel	- ABC transporters	- Anatomical structure formation
- Unsaturated fatty acid metabolism	- Pl signaling system	- Metal ion transmembrane transporter	- O-glycan biosynthesis	**Module D**
- Leukotriene biosynthesis	**Module C**	**Module B**	- Nuclear receptor transcription	- Vasopressin regulated water reabsorption
- Icosanoid metabolism	- CAMs	- Monovalent inorganic cation transport	- Guanyl nucleotide exchange factor	**Module E**
Module D	- Regulation of neurotransmitter levels	- Metail ion transport	- Cell projection	- ECM receptor interaction
- Aspartic-type peptidase activity	- Cell recognition	- Cation transport	- Synapse	
- IL1 binding	- Cell-cell adhesion	- Ion transport	- Axon guidance	
- IP3 kinase activity		- Potassium ion transport	- Cell migration	
- Potassium channel activity		- Voltage gated potassium channel complex	- Brain development	
- Wnt-protein binding		- Voltage gated potassium channel activity	- Extracellular matrix part	
Module E		- Potassium channel activity	- Basement membrane	
- Histone H2A acetylation		**Module C**	- Focal adhesion	
- Alkene biosynthesis		- Neurogenesis	- ECM interactions	
- Glucan metabolism		- Neuron differentiation	- Integrin interactions	
- cAMP metabolism		- Axon guidance	- Tight junction	
- PLC activity by G-protein		- Axonogenesis	- Adherence junction	
- Telomerase regulation		- Cellular morphogenesis during differentiation	- Leukocyte transendothelial migration	
Module F		- Neurite development	**Module I**	
- Filopodium assembly		- Neuron development	- CREB phosphorilation of RAS	
- Collagen metabolism		**Module D**	- Post NMDA receptor activation events	
- Negative regulation of DNA replication		- NRAGE signals death JNK	- Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell	
- Negative regulation of inflammatory response		- Rho GTPase cycle	- Transmission across chemical synapses	
		- Signaling by NGF	**Module J**	
		- G-alpha 12 13 signaling events	- Platelet activation triggers	
		- Cell death via NRAGE NRIF NADE	- Formation of platelet plug	
		Module E	- Platelet activation	
		- PTK activity	**Module K**	
		- Transmembrane receptor PK	- Opioid signaling	
		- Transmembrane receptor PTK	- PLC beta mediated events	
		Module F	- CAMs	
		- TCR downstream signaling	- PLC gamma1 signaling	
		- PD1 signaling	- GAP junction	
		- Phosphorylation of CD3 and TCR zeta chains	- Vascular smooth muscle contraction	
		- CAMs	- GNRH signaling pathway	
		- Type I diabetes mellitus	- Calcium signaling pathway	
		Module G	- Long term potentiation	
		- Dilated cardiomyopathy	- Melanogenesis	
		- Arrythmogenic right ventricular cardiomyopathy	- ST myocyte AD pathway	
		- Sodium ion transport	- Regulation of insulin secretion by glucagon-like petide1	
		- Calcium channel activity	- Regulation of insulin secretion by free fatty acids	
			- NO1 pathway	
			- G-alpha S signaling events	

Table 10. Summary of the enriched and most representative gene sets in the MIBC cases developing progression events (corrected *p-value* at 0.05). The gene sets were grouped in several modules according to the hierarchical cluster trees and the modules defined by dynamic tree cut method. Results were obtained only when GeSBAP and GSA-SNP were applied.

GeSBAP

Module A
- Regulation of Rho protein signal transduction
- Rho protein signal transduction
- Regulation of Ras protein signal transduction
- Regulation of small GTPase mediated signal transduction

Module B
- Bone remodeling
- Tissue remodeling

Module C
- Di- , trivalent inorganic cation transport
- Calcium ion transport

Module D
- Regulation of neurotransmitter levels
- Neurotransmitter secretion

Module E
- Brain development
- Central nervous system development

Module F
- Cell junction organization
- Homophilic adhesion
- Cell-cell adhesion

Module G
- Anion transport
- Phosphate transmembrane transporter activity
- Axon guidance
- Transmembrane receptor PTK signaling
- Calcium signaling
- Muscle development
- Phospholipid transporter activity

GSA-SNP

Module A
- Voltage gated channel activity
- Voltage gated cation channel activity
- Gated channel activity
- Substrate specific channel activity
- Ion channel activity
- Cation transmembrane transporter activity
- Cation channel activity
- Metal ion transmembrane transporter activity

Module B
- Voltage gated potassium channel activity
- Potassium channel activity
- Metal ion transport
- Cation transport
- Ion transport
- Monovalent inorganic cation transport
- Potassium ion transport

Module C
- Neurogenesis
- Generation of neurons
- Neuron differentiation
- Axonogenesis
- Cellular morphogenesis during differentiation
- Neurite development
- Neuron development

Module D
- Collagen
- Extracellular matrix part
- Proteinaceous extracellular matrix
- Extracellular matrix

Module E
- PTK activity
- Transmembrane receptor PK activity
- Transmembrane receptor PTK activity

Module F
- Axon guidance
- NCAM signaling for neurite out growth
- NCAM1 interactions

Module G
- Calcium channel activity
- Arrhythmogenic right ventricular cardiomyopathy
- Hypertrophic cardiomyopathy
- Dilated cardiomyopathy

Module H
- Trafficking of AMPA receptors
- Transmission across chemical synapses
- Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell

Module I
- Formation of platelet plug
- Platelet activation
- Enzyme linked receptor protein signaling
- Brain development
- Central nervous system development
- GTPase regulator activity
- GTPase activator activity
- Focal adhesion
- ECM receptor interaction
- CAMs
- Type I diabetes mellitus
- Viral myocarditis
- Regulation of insulin secretion by glucagon like peptide1
- Glutamate receptor activity
- Phosphoric diester hydrolase activity
- Calcium signaling
- Cell-cell adhesion
- Cell junction organization
- Hematopoietin interferon classD200 domain cytokine receptor activity
- Regulation of neurotransmitter levels
- Embryonic development
- PTP activity
- Adherens junction
- Axon guidance
- Rho GTPase cycle

Table 11. Summary of the enriched and most representative gene sets in the in the MIBC cases dying due to BC (corrected *p-value* at 0.05). The gene sets were grouped in several modules according to the hierarchical cluster trees and the modules defined by dynamic tree cut method. Results were obtained only when GeSBAP and GSA-SNP were applied.

GeSBAP		GSA-SNP
Module A	**Module A**	**Module H**
- Rho protein signal transduction	- Neurogenesis	- Guanyl nucleotide exchange factor activity
- Regulation of Rho protein signal transduction	- Generation of neurons	- Signaling by NGF
- Regulation of Ras protein signal transduction	- Neuron differentiation	- Rho GTPase cycle
- Regulation of small GTPase mediated signal transduction	- Axonogenesis	- NRAGE signals death through JNK
	- Cellular morphogenesis during differentiation	- G-alpha 12 13 signaling events
Module B	- Neurite development	**Module I**
- Synaptic transmission	- Neuron development	- Adherens junction interactions
- Regulation of neurotransmistter levels	**Module B**	- Cell-cell adhesion system
- Neurotransmitter secretion	- Cation transport	- Cell junction organization
Module C	- Voltage gated channel activity	**Modules J**
- Di-, trivalent inorganic cation transport	- Voltage gated cation channel activity	- Collagen mediated activation cascade
- Calcium ion transport	- Gated cannel activity	- Reactome formation of platelet plug
Module D	- Ion channel activity	- Platelet activation
- Homophilic cell adhesion	- Cation Transmembrane transporter activity	**Modules K**
- Cell-cell adhesion	- Cation channel activity	- Enzyme linked receptor protein signaling
Module E	- Metal ion transmembrane transporter activity	- Transmembrane receptor PTK signaling
- Sodium ion transport	**Module C**	**Modules L**
- Phosphate transmembrate transporter activity	- Collagen	- Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell
- Transmembrane receptor PTK signaling	- Basement membrane	- Transmission across chemical synapses
- Adherens junction	- Extracellular matrix part	**Modules M**
- Phospholipid transporter activity	- Proteinaceous extracellular matrix	- Glutamate receptor activity
- Protein export	- Extracellular matrix	- Neurotransmitter levels
	Module D	- Synaptic transmission
	- PTK activity	**Modules N**
	- Transmembrane receptor PK activity	- Brain development
	- Transmembrane receptor PTK activity	- Central nervous system development
	Module E	**Module O**
	- Axon guidance	- Axon guidance
	- NCAM1 interactions	- Semaphorin interactions
	- NCAM signaling for neurite out growth	- Viral myocarditis
	Module F	- Cell-cell adhesion
	- Arrythmogenic right ventricular cardiomyopathy	- Adherens junction
	- Hypertrophic cardiomyopathy	- Tight junction
	- Dilated cardiomyopathy	- Type II diabetes mellitus
	Module G	- Calcium signaling pathway
	- Regulation of actin cytoskeleton	- Regulation of insulin secretion by glucagon like peptide1
	- Focal adhesion	- Synapse
	- ECM receptor interaction	- Hematopoietin interferon classD200 domain cytokine receptor activity
	- Integrin-cell surface interactions	- Amine compound SLC transporters

4.3. SNP-SNP interactions associated with UCB clinical outcomes

We studied prediction of tumour recurrence and progression in 836 UCB cases with NMIBC and progression and mortality in 235 MIBC. In the *Materials and Methods Section* was pointed the number of events and censored observations for each clinical outcome and the number of SNPs included in the analyses.

In the step of tuning and assessment of this new methodology, we explored different cut-off points of the survival time to define binary response variables: 30, 40, 50, 60, 70, and 80 months, considering to the *Formula 1*, the distributions of the survival and follow-up times for each event, and taking into account that the number of censored observations was quite high for the NMIBC clinical outcomes. We selected 10,000 random SNP pairs and estimated, for each pair, the partial likelihood ratio test based on the two fitted Cox regression models (*Formula 2a* and *2b*). For each one of these selected pairs and for each of the 6 possible definition of the binary response variable given by the different cut-off points in the survival time, we calculated the likelihood ratio test based on the two fitted logistic regression models (*Formulas 3a* and *3b*) and the Pearson correlation coefficient between the *p-values* of the Cox models and each of the 6 different logistic models, as $-\log_{10}$ values. These correlation coefficients are shown in the *Table 12*, where we can observe in all scenarios that the correlations increase with the cut-off point. The highest correlation was reached when the cut-off point coincided with the maximum survival time (i.e., 80 months).

Table 12. Correlation coefficients (r) between the p-values obtained by the logistic and the Cox regressions for 10,000 random SNP pairs.

	NMIBC				MIBC			
	Recurrence		Progression		Progression		Recurrence	
Cutpoints	**r**	**Censored n(%)**	**r**	**Censored n(%)**	**r**	**Censored n(%)**	**r**	**Censored n(%)**
30	0.816	607 (72.6)	0.669	782 (93.5)	0.79	116 (49.4)	0.721	147 (62.2)
40	0.869	586 (70.1)	0.743	776 (92.8)	0.807	112 (47.7)	0.793	136 (57.9)
50	0.913	573 (68.8)	0.637	768 (91.9)	0.829	107 (45.5)	0.826	132 (56.2)
60	0.932	566 (67.7)	0.891	762 (91.1)	0.836	106 (45.1)	0.842	129 (54.9)
70	0.936	564 (67.5)	0.954	756 (90.4)	0.836	106 (45.1)	0.848	127 (54.0)
80	0.946	561 (67.1)	0.995	753 (90.1)	0.836	106 (45.1)	0.848	127 (54.0)

In order to compare the results between logistic and Cox regression models and analyze the scenario for SNP interactions associated with low *p-values*, 5 million pairs were randomly selected and the four regression models run (*Formulas 2a*, *2b*, *3a* and *3b*). Then, we calculated the *p-values* from the two ratio likelihood tests evaluating the interaction term, and the Pearson correlation coefficient and showed a good relationship among them. However, we observed that the sample size influenced the correlation. In NMIBC patients (836 individuals), we obtained a very high correlation (r=0.944) for the analysis of recurrence, and a moderate correlation (r=0.944) for progressions. Analysing BC-specific mortality and progression in the MIBC (235 individuals), we obtained lower correlations, r = 0.858 and r = 0.840, respectively (*Supplementary Figure 10*). The number of models with a strong discrepancy between both kinds of regression is very low. Similar to the correlation coefficients, the number of discrepancies decreases with increasing the sample size.

In the second step of the analyses, we executed BOOST in order to detect the potential interactions between all the pairs of SNPs. We selected thr interactions that reached $P < 10^{-4}$ or lower. At last, we obtained around 22 millions of SNP-SNP interactions detected by logistic regression for each of the four outcomes considered. We also observed that the likelihood reported by BOOST was similar to that obtained directly fitting the two logistic regression models.

In the third step, two Cox regression models (*Formulas 2a* and *2b*) were fitted for all the pair-wise SNP-SNP interactions selected by BOOST and we estimated the *p-values* of the partial ratio likelihood tests. The models were not adjusted by potential confounding factors. Similarly to the logistic analysis performed using BOOST, we fitted unadjusted Cox regression using the co-dominant mode of inheritance. We set a conservative significance threshold at $P < 10^{-5}$ to select SNP-SNP interactions that may be of interest in the prognostic assessment of the disease. Using this criterion, around 1.5 million of interactions were select for further analyses.

In the last step we finally adjusted Cox regression models including confounder variables and combining the different inheritance models for the pairs of SNPs. We fitted the Cox regression models using the 16 possible combinations between the 4 modes of inheritance (additive, recessive, dominant and co-dominant) for each pair of SNPs.

Correction for multiple testing was needed at this stage and we applied the two approaches explained in the *Materials and Methods Section*. First, by applying the method based on the principal components and the informative SNPs, we set the threshold at 8.60×10^{-14} and 9.46×10^{-14} for the NMIBC and the MIBC subsample. The thresholds obtained by the

second of the methods, based on the 0.4 correction factor over the total number of interactions gave two similar thresholds, 1.06×10^{-13} and 1.17×10^{-13} for the NMIBC and MIBC subsamples, respectively. In the *Table 13*, we included the most significant models for each of our four outcomes, some of them with *p-values* smaller than the significant thresholds calculated by the 2 methods.

Table 13. SNP-SNP interactions with potential prognostic value in UCB.

Subphenotype	Outcome	SNP ID	Chr	Closest gene (SNP location)	MoI	SNP ID	Chr	Closest gene (SNP location)	MoI	P	Threshold Informative-SNPs	Threshold 0.4-Factor
NMIBC	**Recurrence**	rs7498329	16	*LAT* (flanking 3UTR)	A	rs909010	19	*TRPM4* (intron)	A	5.53×10^{-14}	Significative	Significative
NMIBC	**Recurrence**	rs941586	14	*CHGA* (intron)	R	rs6093059	20	*CDH4* (flanking 5UTR)	A	7.31×10^{-14}	Significative	Significative
NMIBC	**Recurrence**	rs668204	11	*NCAM1* (flanking 3UTR)	D	rs7141930	14	*SEL1L* (flanking 5UTR)	C	1.91×10^{-13}	Non-significative	Non-significative
NMIBC	**Recurrence**	rs2169685	10	*SLC16A9* (flanking 3UTR)	D	rs17004695	21	*C21orf29* (flanking 3UTR)	D	1.94×10^{-13}	Non-significative	Non-significative
NMIBC	**Progression**	rs253235	5	*FLJ46010* (intron)	R	rs10148938	14	*CDCA4* (flanking 5UTR)	D	8.43×10^{-14}	Non-significative	Non-significative
MIBC	**Progression**	rs3750272	2	*HTLF* (flanking 5UTR)	D	rs2126337	16	*SALL1* (flanking 5UTR)	R	1.66×10^{-13}	Non-significative	Non-significative
MIBC	**BC-related surv**	rs10110883	8	*MSR1* (flanking 5UTR)	D	rs10847791	12	*KIAA1944* (intron)	D	4.11×10^{-14}	Significative	Significative
MIBC	**BC-related surv**	rs11977984	7	*LHFPL3* (intron)	C	rs4394757	10	*IPMK* (flanking 3UTR)	C	7.24×10^{-14}	Significative	Significative
MIBC	**BC-related surv**	rs11588107	1	*PROX1* (flanking 5UTR)	D	rs6439470	3	*RYK* (flanking 5UTR)	D	1.56×10^{-13}	Non-significative	Non-significative
MIBC	**BC-related surv**	rs2066713	17	*SLC6A4* (intron)	C	rs2015823	22	*CARD10* (intron)	D	2.03×10^{-13}	Non-significative	Non-significative

MoI includes: A for additive; R for recessive; D for dominant; C for co-dominant.

Chapter 5: Discussion

The search of new and robust prognostic markers for UCB is, in fact, the driving force of this study that is unique in providing new clues on the involvement of inherited factors in UCB outcome, beyond the studies on cancer risk. It is highly unlikely that only one factor/marker will be able to discriminate those individuals presenting an outcome or not. Rather, it is highly probable that prognosis models need to consider several independent factors/markers. This reasoning is even more certain when considering genetic variables because of the relatively small effect of each of them individually.

5.1. Independent SNPs associated with UCB clinical outcomes

We have conducted a comprehensive genome-wide prognosis scan in patients with UCB and have identified distinct loci associated with the risk of tumor recurrence, progression, relapse and BC-specific/overall mortality. Our notable findings include: 1) the consistency of results between the two genome-wide studies; 2) the inclusion of independent populations for Validation when NMIBC is considered; and 3) the identification of several significant loci with relatively high estimates, especially when compared with those found in GWAS for cancer risk. The identified SNPs lie in genetic regions that have not been previously related to any UCB outcome. As in GWAS of cancer risk, most of the SNPs associated with prognosis are located in intergenic or intronic regions, suggesting that they may tag other SNPs directly associated with the outcomes of interest. Alternatively, they could influence gene expression, splicing, or other events at distance. A genetic enrichment analysis using Ingenuity® indicated that most of the genes harboring these significant SNPs

associated with the NMIBC outcomes were involved in cellular death and development, drug and lipid metabolism, and molecular transport. Regarding MIBC, enrichment was detected for infectious disease response, molecular transport and cell signaling, and hematological system development.

rs754799, showing the strongest association with risk of recurrence and of relapse localizes to 19p13.3 and lies less than 1kb 5' from *SCAMP4* and *ADAT3*, two genes that share common sequence but are alternatively transcribed. *SCAMP4* encodes a member of the secretory carrier membrane protein family (SCAMP) that has been implicated in membrane trafficking and vesicular transport and is ubiquitously expressed (Castle and Castle 2005). *ADAT3* encodes an adenosine deaminase that is involved in tRNA editing. The association of either of these genes with cancer has not been previously reported.

One SNP was significantly associated with progression in both the Discovery and Validation phases and 5 exhibited similar associations in both phases as well, three of which are located in introns. The strongest association was for rs4246835, which is located in a "gene desert" region on 9p22-p13, ~246kb from the nearest gene, *SLC24A2*. This gene encodes a sodium, potassium, and calcium ion exchanger and is implicated in retinal photoreceptor signaling (Sharon, Yamamoto et al. 2002). Centromeric to this SNP lies *MLL3* (also known as *AF9*, at ~312 Kb distance), a Drosophila Trithorax homolog gene that regulates cell differentiation, is often translocated in mixed lineage leukemia, and has been found to be somatically mutated in some tumors (COSMIC database, http://www.sanger.ac.uk) (Pina, May et al. 2008). It is possible that rs4246835 tags this gene in the variant's association with NMIBC progression, although further fine-mapping of this region is necessary to identify the causal SNP.

In addition to rs754799 described above, two SNPs showed similar association with relapse in both Discovery and Validation phases although results were not significant in the latter phase: rs4946483, a non-genic SNP on chromosome 6q22, and rs11615759, an intronic variant of *SOX5*. Chromosome 6q22 deletions in tumors correlate with malignant and metastatic progression in sporadic endocrine pancreatic tumors and are associated with shorter survival in primary central nervous system lymphoma patients, suggesting the presence of tumor suppressor gene(s) in this region (Barghorn, Speel et al. 2001; Nakamura, Kishi et al. 2003; McPhail, Law et al. 2011). *SOX5* is a member of the sex determining region Y (SRY)-related high mobility group (HMG)-box family of transcription factors involved in embryonic development and cell fate determination. *SOX5* amplification has been found in testicular seminomas (Zafarana, Gillis et al. 2002).

Some of the SNPs identified in the Discovery phase of the MIBC studies are statistically significant even with the most stringent scenarios of multiple test correction. The SNP rs1015267 shows the strongest association with risk of death due to UCB for the recessive and codominant modes of inheritance (*p-value* < 1×10^{-8}). It is located in the intronic region of *TMEM16C* in the chromosome region 11p14.2. This gene encodes a member of TMEM16 family of transmembrane proteins, which regulates calcium activated chloride channels (Hartzell, Yu et al. 2009). The members of this family are cell-surface proteins that are up-regulated in cancer (Galindo and Vacquier 2005). In addition, it seems that mutations described in one of the most relevant members of this family (*ANO1*) is not associated carcinogenesis, but cell proliferation or tumor progression (Miwa, Nakajima et al. 2008).

When the progression was evaluated in MIBC, another statistically significant recessive SNP arose (rs16927851). This SNP is located in the 3' flanking region of *BCAT1* (~ 100kb), which is located in 12p12.1. This gene encodes the cytosolic form of the enzyme branched-chain amino acid transaminase. It catalyzes the reversible transamination of branched-chain alpha-keto acids to branched-chain L-amino acids, essential for cell growth. Two different clinical disorders have been attributed to a defect of branched-chain amino acid transamination: hypervalinemia and hyperleucine-isoleucinemia. Within the *BCAT1* gene there is a functional *c-Myc* binding site located 3' of its transcription initiation site, and has been shown to be a direct target for *c-Myc* activity in both mice and humans (Benvenisty, Leder et al. 1992). It was pointed that *BCAT1* is expressed at significantly higher levels in tumor tissues with distant metastases, compared to those without them. It makes possible to suggest the use of this gene expression as a highly reliable predictive factor for distant metastasis in patients with advanced colorectal cancer (Yoshikawa, Yanagi et al. 2006). Additional studies on the association between cancer susceptibility and this gene overexpression suggest its involvement in epithelial ovarian cancer (Ju, Yoo et al. 2009), nasopharyngeal carcinoma (Zhou, Feng et al. 2007) and testicular germ cell tumors (Rodriguez, Jafer et al. 2003). On the other hand, there is also some evidence regarding the prognostic role of the mRNA levels of *BCAT1* in medulloblastoma patients with metastases compared with those without (de Bont, Kros et al. 2008).

Resequencing, searching for causal variants, and functional analyses should shed light on the mechanisms through which these SNPs associate with UCB outcomes. The application of high throughput genotyping to the identification of genetic variants associated with outcome poses major challenges. One of them is the heterogeneity of case series. In our

work, Discovery phase was performed with prospective studies and the Validation for NMIBC consisted mostly of retrospective studies having different aims and applying variable designs as well as classifications and therapeutic criteria. We made outstanding efforts to identify them and ensure data exhaustiveness and homogeneity. Their variability has likely resulted in the Validation of a smaller number of NMIBC SNPs through effect dilution, suggesting that the SNPs identified are robust to such variability and that additional prospective studies may identify new genetic markers. Even with this constraint, our study has identified several novel independent genetic markers associated with prognosis, most of them in proximity or within biologically plausible gene candidates. Several of these genes are implicated in cellular signaling or tissue/organ development, suggesting a link between these processes and tumor progression. A notable finding was the greater magnitude of the HR associated with the outcomes of interest in comparison to the risks observed in genetic risk association studies.

Importantly, in the NMIBC series there were no consistent significant associations of the SNPs with baseline tumor characteristics in the four series (*Supplementary Table 6*). None of the NMIBC SNP *p-values* reached significance after Bonferroni correction. In addition, the correlation coefficient was relatively small with the largest rho coefficient being 0.17. In the MIBC series no significant association was observed after correcting for multiple testing (*Supplementary Table 9*). Therefore, these SNPs were not highly correlated with baseline tumor stage, grade, or size and their association with the outcome was independent from the effects of other prognostic factors since those were included as covariates in the adjusted models.

UCB is a complex disease regarding both etiology and prognosis and the prognosticators that are used in standard practice discriminate poorly cases according to their tumor evolution. This complexity is shown when comparing the discrimination ability (c-statistics) of the identified SNPs with that of the used and most important prognosticators (stage-grade-TG, multiplicity, and size). For progression, the added value of the SNPs was 5%-6%, while neither multiplicity nor size adds any value yet they are generally used by urologists to prognosticate UCB progression. The c-statistics of the SNPs in discriminating recurrence is smaller, similar to most prognosticators and biomarkers for this outcome. We suspect the occurrence of recurrence is mainly driven by the skills of the surgeon rather than the biology of the tumor or the patient because there are no robust biological markers of associated with recurrence. Therefore, we believe that the study provides evidence of the potential clinical usefulness of inherited factors in UCB prognosis.

While two NMIBC subphenotypes are well recognized (low-risk and high-risk), this study aimed at identifying SNPs associated - as main effects - with outcome in overall UCB as evidence of the important role of the host factors in tumor evolution, similarly to what has been evidenced in cancer susceptibility assessments. Furthermore, the subphenotype analysis would require larger studies with a common pathological assessment to avoid misclassification. Furthermore, the study did not set out to assess the predictive value of the markers for treatment response, this requiring of specific study design and analysis.

This study has several major strengths, among them the agnostic comprehensive exploration of the genome, the large sample size and large number of events with several independent series for the Discovery and Validation (when available), the quality of the clinical and the prospective follow-up in Discovery studies, the long follow-up of all

patients, and the inclusion of all stages representative of the disease seen in a wide range of clinical settings. Although this project included patients from American and European cohorts, the analysis was restricted to Caucasian patients of European descent. In addition, genotyping for the Spanish series was performed at the Core Genotyping Facility, National Cancer Institute, USA. The stringent quality control measures applied to the NCI GWAS (Rothman, Garcia-Closas et al. 2010) did not show population substructure in the Spanish study with the other US studies. Therefore, our patient population is rather homogeneous: only the patients recruited in Tenerife shown slight population differences in a substructure-population test performed by principal component analysis. The genomic control inflation factor derived from the individual SNP analysis indicated that the effect of population substructure was minimal. The long follow-up (> 10 years) of the patients in almost all series made possible to have sufficient events in all the considered outcomes. Given the number of cases and events, we have 80% power to detect hazard ratios of about 1.5-1.6 or greater for homozygous rare genotype with frequencies of 0.10 and 0.08 for recurrence and relapse, respectively, at the unadjusted *p-value* of 1×10^{-4}. The results are robust for tumor progression in NMIBC patients when HR >1.8 for homozygous rare genotype with frequencies >0.20. When we considered the MIBC cases, the statistical power was diminished due to the lower sample size available for these patients. Nevertheless, we have 80% power to detect HR around 2 for homozygous genotypes with frequencies of >0.3 (*Supplementary Figure 8*). In addition, initial genome-wide scans were done in parallel in the Discovery phase and the main results were cross-validated in both series. The most significant results were further replicated in additional NMIBC series to minimize false discovery.

A further strength of the study is that it involved both a referral center and a wide range of non-referral centers in the discovery phase, indicating that the findings are robust and could be applied in a broad range of community-University hospital settings. While this study represents an important advance in the discovery of inherited susceptibility genetic factors involved in UCB outcome, the translation of such results into the clinics, both as prognostic and predictive markers in UCB, requires additional confirmation in homogeneously conducted follow-up studies.

In summary, we have identified genetic variants independently associated with outcome in patient with UCB through a cross-validation of two genome-wide scans performed in independent Caucasian populations and a further Validation for NMIBC patients with 6 independent series of different geographical location and urological practice. These findings highlight the important role of inherited genetic factors in tumor progression. The ease of analyzing germline SNPs and the robustness of the assays used are important assets of this strategy. Next steps would include efforts to extend these studies to larger series, combine independent SNPs, and dissect the biological underpinnings of inherited genetic factors in the clinical evolution of UCB. The lack of large clinical series with germline DNA, exhaustive and reliable clinical data, and enough follow-up is a major challenge for the Validation of genome-wide prognostic studies.

5.2. Biological pathways associated with UCB outcomes

In this study, we used an approach based on prior biological knowledge from public databases in order to detect gene sets that display an overrepresentation of SNPs

significantly associated with UCB outcomes. The main motivation of this approach is to increase the power to discover new genetic relations with a particular GS, when compared with the single-SNP strategy based in its main effects. We decided to proceed using different statistical approaches because it is known that even using the same set of data the results can vary substantially under different gene set analysis (GSA) scenarios (Wang, Li et al. 2007; Hong, Pawitan et al. 2009; Cantor, Lange et al. 2010).

The great amount of GS detected associated with the different outcomes made necessary to interpret the results using the clustering and similarity methods described in the *Materials and Methods Section*. According to the results obtained for NMIBC series – indistinctively for recurrence or progression – there is an important overrepresentation of GS associated with the signaling of *G protein, small GTPase and Rho protein*, which were obtained using *i*-Gsea*4*Gwas, ICSNPathway, GSA-SNP or GeSBAP. The association of alterations in these regulator GS with cancer development and their clinical importance have been discussed in multiple publications (Oxford and Theodorescu 2003; Gur, Kadowitz et al. 2011; Fujita, Shida et al. 2012; Smith, Baras et al. 2012). The alterations of key pieces in the *inflammatory response* seem to be also important because the regulation of T-cell/lymphocytes is also modified in recurrence and progression according to the results obtained in ALIGATOR/GSA-SNP and *i*-Gsea*4*Gwas respectively. There is plenty of literature associating the *T-cell/lymphocyte activity* with the prognosis after treatments based on BCG instillation (Hoffmann, Roumeguere et al. 2006; Takeuchi, Dejima et al. 2011); even its use as a prognostic factor was suggested in the past (Mizutani, Okada et al. 1996; Saint, Patard et al. 2002). Furthermore, GS regarding *membrane cell transport* based on cation transport or voltage-gated channel activity – obtained in GeSBAP and GSA-SNP

– also arose in the GSA. Interestingly, the more specific *calcium ion transport channel* appeared in both methods. Multiple GS associated to *neuronal/axon signaling and proliferation*; and *synaptic processes* appeared in recurrence and progression NMIBC series when GeSBAP and GSA-SNP were used. There are evidences regarding *neuron-associated protein expression* with the prognosis of NMIBC (Mhawech-Fauceglia, Ali et al. 2009).

On the other hand, we have observed that some groups of GS are more common in one of the NMIBC outcomes. Altered gene sets regarding *cell-cell junctions* (general CAMs, GAP junction, integrins and extracellular matrix) are very common in progression – when GeSBAP and GSA-SNP are used – but they are much scarce in recurrence and they are obtained only in GSA-SNP in last outcome. Probably these results were obtained because of the cell proliferation and reorganization in the context of cancer invasiveness (Zhong, Chen et al. 2010; Reis, Leite et al. 2012). In addition, alterations in the *actin filament organization* related pathways were detected in NMIBC progression when ALIGATOR and ICSNPathway were applied. These pathways, in addition to *small GTP-binding protein Rho*, have been described as a potential prognostic factor for UCB invasiveness (Kamai, Tsujii et al. 2003).

The results obtained for the MIBC series are much less rich than those obtained in the NMIBC series because of the modest statistical power available for this UCB subtype. Furthermore, we obtained significant results only when GSA-SNP and GesBAP were applied; and as it was pointed above, these methods have multiple important biases. Thus, the results and conclusions for the MIBC series should be accepted with caution. Probably the most significant aspect of the results is the extremely high concordance between the MIBC outcomes. However, there is one general difference: the presence of *structural*

tissue-development pathways in MIBC progression (e.g. bone remodeling, tissue remodeling and muscle development).

In the MIBC series we found again the GS regarding *small GTPase signalling, cell-membrane transporters and neuronal/axon*, which appeared also in the NMIBC series. It would suggest an action along the time that affects the prognosis of the disease from the reappearance of the tumor to the patient's death. It may be very interesting to evaluate the overlapping of the enriched GS between the progression outcomes in NMIBC and MIBC in order to detect some shared pathway that would lead us from NMIBC to the MIBC. Unfortunately we are limited by the modest statistical power in the MIBC series. Apart from the mentioned common GS found in all the analyzed outcomes, we could not find any significant block of similar pathways in the different progression outcomes. On the other hand, we observed that immune and inflammatory pathways such as the T-cell/lymphocyte regulation or the production of leukotrienes/icosanoids are specific of NMIBC progression.

A priori unexpected results aroused in some of the methods in all the analyzed outcomes, such as the groups of pathways related to *cardiomyopathy*. However, even in this case a possible link with UCB can be established through the action of some members of the S100 gene family, whose over-expression is associated with poor prognosis in UCB (Yao, Davidson et al. 2007).

The number of GSA procedures is increasing every few months trying to overcome the known limitations in these methodologies. However, the vast majority of these methods have been designed with the aim of being applied on studies based on a case-control strategy ant it leaves a rather reduced list of options for a prognostic study. Despite of this,

several GSA methods based on competitive tests (those which compare test statistics for SNP/genes in the pathway to a background defined by the remaining pathways) and the *p-values* obtained in the GWPS are available. It is difficult to give more credibility to one statistical approach because no much is known in terms of comparative power between the algorithms and even there are some contradictions establishing the most reliable method when some of them have been compared (Chen, Hutter et al. 2010; Jia, Wang et al. 2011; Kwon, Kim et al. 2012). Most of the previous studies showed lack of power in unfavorable scenarios such as small sample size and makers with low association with the disease.

One of the critical points in the SNP/gene strategies is the lack of common criteria in the reduction of the SNP information within each gene. In all the studied algorithms the SNP with the lowest p-value acts as a proxy of its closest gene. However this approach may lead to a loss of power because when multiple markers with risks of different directions map one single gene, their combined effect may be lost.

The presence of LD between the SNPs is a potential source of bias. It invalidates the assumption of independence between the SNPs and ignores that multiple signals from one single genuine marker may appear. Algorithms such as ALIGATOR claim to control this effect and we have seen that it is basically certain in our dataset; others based on *i*-GSEA request the use of LD-pruned before the GSA and those based on the Z-statistic method ignore the LD structure. Despite of this, the preservation of the LD patterns in the dataset is important but they are usually disrupted in SNP-label permutation procedures inherent in the assessment of the statistical significance. At the gene level we can establish potential biases regarding the gene length and their overlapping. The former means that larger genes will be more prone to have more disease-associated signal within them and inflate the

significance for those GS with many long genes. The latter implies that one single SNP may be mapped to several overlapping genes and will lead to an overrepresentation of its signal. The correction for gene size is accomplished in ALIGATOR and *i*-GSEA using resampling strategies, but no correction is performed neither in GSA-SNP nor GeSBAP. The authors of ALIGATOR and GSA-SNP mention that in case that a SNP could be mapped within different genes, both would be included in the analysis; nothing in this regard is mentioned in *i*-GSEA but the overlapping problem is issued (Jia, Wang et al. 2010; Wang, Jia et al. 2011).

The GSA methods have an inherent source of bias: the GS themselves. There is still an important gap in the knowledge of the biological function of all the genes in genome, so the genes with no functional evidence are lost. On the other hand, the well studied genes and biological functions are overrepresented in the GS. It may be an important issue in the competitive test strategies because an unrealistic background of analysis is constructed. An additional non-negligible factor emerges when the different GS sources are compared and reveal a lack of consensus in the common GS. Despite of this, considerable efforts are done in trying to compile different sources of information that overcome these problems by initiatives such as MSigDB (Subramanian, Tamayo et al. 2005). The different statistical methods underneath the GSA algorithms demand a collection of independent GS to be analyzed. Unfortunately this request is difficult to fulfill because a great number of genes play diverse roles in different pathways; not to mention the hierarchical structure of GO. In our study we used the provided GS lists in ALIGATOR and GeSBAP. On the other hand, we used the GS available in MSigDB in the other methods because it claims to keep only GS with a certain level of non-redundancy. Nonetheless, it may cause over-conservative

results because the higher is the number of GS, the more strict needs to be the multiple test correction. The differences in GS size can lead to biased results (in a similar way as we see in the gene-size issue) but it can be controlled by resampling (ALIGATOR) or permutations (*i*-GSEA) (Efron and Tibshirani 2007).

According to recent reviews on this issue (Cantor, Lange et al. 2010; Wang, Li et al. 2010; Fridley and Biernacka 2011; Wang, Jia et al. 2011), we selected four of the most popular and user-friendly tools to perform our analysis that were based on different statistical backgrounds. In addition, we added ICSNPathway as far as it claims to be the only tool able to obtain functional results. All these tools but GeSBAP rely on gene-based methodologies that need a two-step process from SNP to genes prior to the statistical analysis. It has been observed that this approach is more powerful than the direct SNP-based methods (Wang, Li et al. 2007; Yu, Li et al. 2009). The possibility of using directly the *p-values* obtained in GWPS makes feasible the study of prognostic data in a rather simple way. The main advantage of the *p-value-methods* choice relies on the possibility of reducing the genomic inflation and controlling the confounding factors. This task represents a challenge for the raw genotype-based methodologies but little has been done to deal with it.

In summary, an extensive list of gene sets has been found to be associated with specific UCB outcomes in the SBC/EPICURO Study. The obtained results highlight some of the most known biological pathways related to UCB prognosis. In addition, the great number of identified pathways and understanding their interactions is challenging because it suggests that a regulatory system underlying the genetic architecture of UCB needs to be accurately described. This kind of pathway-based approaches have provided a wider

biological perspective if compared with the single-marker analysis. The development of high-throughput sequencing techniques - and their functional interpretation - may provide in the close future a statistically powerful source of data suitable to be used with these pathway-based techniques and allow us to exploit the most of the GWAS data.

5.3. SNP-SNP interactions associated with UCB clinical outcomes

Restricted interaction assessments in case-control studies have been proposed, considering only the SNPs showing significant main effects and analyzing pair-wise interactions among them. While this approach could be easily extended to survival data, it would miss interesting interactions between SNPs without significant main effects. We went through this issue and shown that the SNPs participating in some of the most interesting interactions identified with the proposed approach did not show significant main effects.

From a practical point of view, the exhaustive search of genetic interactions requires large computational efforts, this being a limiting factor of this analysis. Therefore, at the genome-wide setting, only two-way interactions can be considered at present since a further search does not scale up to analyse higher-order interactions. Methodological papers about the available statistical and data mining techniques available for the gene-gene interactions assessment in case-control studies have been published (Cordell 2009), but again, only a few of these techniques have a natural extension to the time-to-event data, among them, classification trees (Breiman, Freidman et al. 1984), random forest (Ishwaran, Kogalur et al. 2008), and logic regression (Ruczinskie, Kooperberg et al. 2003) can be used to detect

high-order genetic interactions with time-to-event data. However, they are limited to manage a genome-wide size number of SNPs.

The key point of our strategy was determining the relationship between the *p-values* assessing the interactions of the logistic and Cox regression models. Genome-wide interaction analyses (GWIA) studies require testing thousands of millions of pairs, and the *p-values* for ensuring significance level need be extremely small. The multiple comparisons topic was also revised and two approaches were selected.

The use of simulated datasets could provide us a better understanding of the conditions in which this strategy could be applied and what is the best way to proceed. However, the high computational cost makes this process prohibitive because of the computational time needed to perform it. We preferred to illustrate the strategy with a particular case using the SBC/EPICURO Study and developing a methodology to assess the comparison between *p-values* in the two mentioned kinds of regression tests. In addition, we determined the limitations of the strategy, when survival data is analyzed with logistic regression.

we selected BOOST to analyse all the pairs of interactions for several reasons. From the technical point of view, BOOST is a very fast, flexible tool with freely available source code; and from the statistical point of view, it is based in the widely extended definition of interaction based in logistic regression models. Our first step of the strategy was to take advantage of the similarity between the fit of the logistic and Cox regression models. BOOST is very fast because uses an approximation to the logistic regression using contingency tables, based on the equivalence between logistic regression and log-linear model. Unfortunately, this approach based in contingency tables cannot be extended to Cox

regression with survival data because it is not possible to add survival time information to the contingency tables. Despite of the BOOST advantages, it has a modest statistical power because in situations with low minor allele frequencies, the contingency tables can be too sparse. It has some additional limitations such as the need of imputed missing values.

The SNP-SNP interactions with potential prognostic value in UCB are presented in the *Results Section*. SNP combinations were analyzed with STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) (Franceschini, Szklarczyk et al. 2013) and Ingenuity® in order to identify known pairwise associations. None of them have been described either in the existing interaction databases or in the Pubmed-based textmining analyses. In addition, we used all the single enrichment analysis tools contained in Babelomics (Medina, Carbonell et al. 2010) (FatiGO, Marmite and Snow) but no direct association between these pairs of SNPs were identified.

We also wanted to analyze the results from a broader point of view and we run the analyses using the complete list of SNP/genes altogether. When we used STRING, we observed that, in the NMIBC series, *NCAM1* and *CHGA* are mentioned in the same context in at least 50 published papers. However this occurrence is just because they are fairly common neuroendocrine biomarkers that are usually evaluated in all kinds of cancer-based studies. On the other hand, Snow provided us a potential MIBC association between *SALL1* and *PROX1*, using *SUMO1* as a link node between them. Despite of this, it seems that there is no published study associating these elements because they don't appear in the textmining tools we used (STRING and Marmite).

To date, very few GWAS have gone beyond the single-marker analysis and have incorporated the interaction testing. To our knowledge, the methodology and the results we have presented are the first ones to be applied to a real dataset in which the survival analyses have been taken into account in the assessment of cancer susceptibility. Novel pairwise interactions were identified in our dataset, even applying the most stringent criteria to avoid the multiple test effect. Despite of this, an independent study is needed to validate the reported results.

In the last two years the so-called genome-wide interaction-based association (GWIBA) studies have been developed and applied in order to identify novel susceptibility loci, whose effects were lost in the regular GWAS approach. There is a common effort trying to reduce de dimensionality of the analyses and a growing list of solutions is being established: methods based on Bayesian models (Zhang 2012), methods based on ranking scores calculated from the difference in marker dependencies between cases/control groups (Piriyapongsa, Ngamphiw et al. 2012), methods based on multidimensionality reduction (Oh, Lee et al. 2012) or methods based on clustering (Xie, Li et al. 2012). The important number of GWAS performed during the last decade and the naïve approach applied until now makes us think that evaluation the genome elements interactions will be the logical next step after the single-locus testing.

In summary, we confirmed that common germline variants are associated with clinical outcomes in UCB. In addition, we identified and validated independent genetic variants associated with NMIBC. Additional markers identified in MIBC patients were also identified and they are waiting to be validated in independent cohorts in the near future. A novel approach has been designed in order to identify pairwise SNP interactions. Technical

and statistical challenges have been solved in this regard and a list of potential SNP pairs has been identified for the first time in a real UCB dataset. The synergistic role of SNPs was also assessed using a complete list of GSA approaches and we have obtained a wide view of the biology underlying the disease. New technologies and challenges are emerging in the assessment of UCB prognosis and the study of germline variants will play an important role from now on.

5.4. Future plans and directions in the prognostic study of UCB

The data gathering from the MIBC patients Validation cohorts is being performed at the time of writing this Thesis. The genetic, pathological and clinical information will be used in order to validate the prognostic value of the SNPs obtained in the Discovery phase.

Once the most significant SNPs in the survival analyses have been identified it may be necessary to perform additional analyses in the chromosomal regions containing these variants. The reason to proceed in this way is because a particular SNP identified in the genome-wide analysis may just be one variant that is linkage disequilibrium with the causal maker. The imputation of the known SNPs contained in the region of interest (but not present in the genotyping platform) provides a source of information to study those regions in deeper detail. Custom genotyping platforms such as Fluidigm Dynamic Array™ would be useful to genotype few hundreds of SNPs restricted to areas of interest at an interesting cost-benefit scenario. However, the desirable option would be the use of Next Generation Sequencing (NGS) techniques and evaluate the presence of multiple genetic variants close to the studied *loci*. In this moment the SBC/EPICURO Study has gone into this field by

sequencing germline DNA from an initial group of 20 patients and another set of 72 patients are waiting to be sequenced in the future.

The assessment of copy-number variants (CNV) in genotyping platforms and its association with UCB susceptibility has been studied in detail in the SBC/EPICURO Study (Marenne, Rodriguez-Santiago et al. 2011; Marenne, Real et al. 2012). The extension to the assessment of UCB prognosis has been performed in exploratory univariate analyses taking into account the Log R Ratio (LRR) for each CNV but no statistically significant results were obtained. Our next efforts will be focused on assessing the potential CNV prognostic value in multivariate survival models.

New statistical approaches have been started to be applied in the study of complex diseases such as UCB. These techniques make possible to deal with the great number of variant candidates and create multifactorial models. Bayesian approaches in combination with shrinkage and selection methods for linear regressions, such as LASSO (Tibshirani 1996), have been started to be applied with promising results in the context of the UCB susceptibility in the Molecular and Epidemiology group at the CNIO. The great-scale study of epistatic biomarkers in cancer was limited because of the computational power needed to perform this task. However, the development of multifactor dimensionalityl reduction methods (MDR) has been around for more than one decade (Ritchie, Hahn et al. 2001) and it has been improved in order to run analyses based on particular statistical models. A MDR variant technique called mb-MDR (Calle, Urrea et al. 2008) has been developed to assess for the disease susceptibility of genetic biomarkers. Moreover, an alternative version that assess for survival is has been developed at the Molecular and Epidemiology group at the CNIO and is ready to be applied on the SBC/EPICURO dataset.

The effort to create genetic scores of germline biomarkers and combine them with the somatic ones into multifactorial models is on the way. In addition, the great amount of information available in the context of "-omics" studies (i.e., transcriptomics, genomics, epigenomics) makes interesting its integration into models that take into account the diverse biological aspects of the disease (Hamid, Hu et al. 2009; Bell, Pai et al. 2011).

Conclusions

I. Through a genome-wide prognostic scan of SNPs covering all autosomal chromosomes and by applying a tiered design (discovery cross-validation and validation phase) that tested each individual SNP, we found that common germline variants are associated with clinical outcomes of UCB.

II. The SNPs rs754799 and rs4246835, which were identified and validated in independent cohorts, were found to be independently associated with the risk of tumor recurrence and progression in NMIBC, respectively.

III. The SNPs rs16927851 and rs1015267, which were identified in independent cohorts, were found to be potentially independently associated with the risk of tumor progression and UCB-specific mortality in MIBC. The validation of these results will be assessed in independent cohorts in the near future.

IV. The application of five gene set analysis algorithms (ALIGATOR, GeSBAP, *i*-Gsea*4*Gwas, ICSNPathway and GSA-SNP), provided a set of biological pathways associated with UCB outcomes. Specific results were obtained for NMIBC regarding the inflammatory system and the immune response. Moreover, in all the UCB outcomes we identified SNP enrichment in pathways in which the GTPases, the membrane transport systems, neuro/axonogenesis and angiogenesis play important roles.

V. The pairwise survival analysis of the non-correlated SNPs genotyped through Illumina Infinium HumanHap 1M probe Beadchip in the SBC/EPICURO Study, let

us to identify pairs of SNPs whose combined effect was associated with the UCB prognosis. We identified four pairs of SNPs that reached the level of statistical significance among NMIBC patients who develop recurrence (rs7498329/rs909010 and rs941586/rs6093059) and MIBC patients who died due to the disease (rs10110883/rs10847791 and rs11977984/rs4394757).

Conclusiones

I. Mediante un análisis pronóstico pan-genómico de SNPs localizados en los cromosomas autosómicos y aplicando un diseño gradual (validación cruzada en la fase de cribado y validación posterior) que analizó cada SNP individualmente, identificamos variantes comunes de línea germinal asociadas con el pronóstico del carcinoma urotelial de vejiga.

II. Los SNPs rs754799 y rs4246835, identificados y validados en cohortes independientes, están independientemente asociados con el riesgo de padecer recurrencia y progresión tumoral de cáncer de vejiga no-invasivo del músculo, respectivamente.

III. Los SNPs rs16927851 y rs1015267, identificados en dos cohortes independientes de la fase de cribado, están potencialmente independientemente asociados con el riesgo de padecer progresión tumoral y a la supervivencia relacionada con el cáncer de vejiga invasivo del músculo. La validación de estos resultados se llevará a cabo en cohortes independientes en un futuro próximo.

IV. La aplicación de cinco algoritmos de análisis de enriquecimiento génico (ALIGATOR, GeSBAP, *i*-Gsea*4*Gwas, ICSNPathway y GSA-SNP) han proporcionado una lista de procesos biológicos potencialmente alterados en relación al pronóstico del carcinoma urotelial de vejiga. Se obtuvieron resultados específicos del cáncer de vejiga no-invasivo en relación al sistema inflamatorio y la respuesta inmune. Además, al estudiar los eventos propios de la enfermedad se observó una

sobrerrepresentación de SNPs en rutas biológicas en las que las GTPasas, los sistemas de transporte de membrana, la neuro/axonogenesis y la angiogénesis desempeñan papeles importantes.

V. El análisis de supervivencia de parejas de SNP no correlacionados genotipados mediante Illumina Infinium HumanHap 1M probe BeadChip en el estudio SBC/EPICURO identificó interacciones en dos parejas de SNPs que alcanzaron el nivel de significación estadística para su asociación con recurrencia: rs7498329/rs909010 y rs941586/rs6093059. Además, se identificaron las parejas rs10110883/rs10847791 y rs11977984/rs4394757 asociados a mortalidad cáncer-específica en pacientes con cáncer de vejiga invasivo del músculo.

Supplementary tables

Supplementary Table 1. Demographic, clinical and pathological characteristics, and follow-up of the NMIBC patients included in Discovery and Validation phases.

Characteristics	Discovery		Validation
	TXBC-1 N (%)	**SBC/EPICURO N (%)**	**All Series N (%)**
Total number	496	836	1284
Sex			
Men	415 (83.7)	732 (87.6)	1036(80.7)
Women	81 (16.3)	104 (12.4)	248(19.3)
Age			
Mean (sd) –years-	63.8 (11.4)	65.5 (10.21)	-
Mínimum - Maximum	19 - 89	22 - 80	18-97
Smoking status			
Non-smoker	142 (28.6)	146 (17.5)	104(8.1)
Former smoker	244 (49.2)	337 (40.3)	189(14.7)
Current smoker	110 (22.2)	346 (41.4)	65(5.1)
Missing	0 (0.0)	7 (0.8)	8(0.6)
Stage			
Ta	229 (46.2)	698 (83.5)	740(57.6)
T1	237 (47.8)	132 (15.8)	486(37.9)
Tis	28 (5.6)	6 (0.7)	52(4.0)
Missing	2 (0.4)	0 (0.0)	6(0.4)
Grade			
G1	20 (4.0)	352 (42.1)	342(26.6)
G2	168 (33.9)	286 (34.2)	317(24.7)
G3	289 (58.3)	198 (23.7)	610(47.6)
Missing	19 (3.8)	0 (0.0)	15(1.2)
Stage-Grade			
TaG1*	19 (3.8)	352 (42.1)	323(25.2)
TaG2	136 (27.4)	265 (31.7)	246(19.2)
TaG3**	72 (14.5)	81 (9.7)	170(13.2)
T1G1/T1Low			16(1.2)
T1G2	32 (6.5)	20 (2.4)	70(5.46)
T1G3***	201 (40.5)	112 (13.4)	399(31.1)
CIS	28 (5.6)	6 (0.7)	52(4.0)
Missing	8 (1.6)	0 (0.0)	8(0.6)
CIS			
Yes	161 (32.5)	-	114(8.9)
No	311 (62.7)	-	199(15.5)
Missing	24 (4.8)	-	53(4.1)
Number of bladder tumors			
1 tumor	191 (38.5)	554 (66.3)	697(54.3)

> 1 tumor	159 (32.1)	238 (28.5)	422(32.9)
Missing	146 (29.4)	44 (5.3)	165(12.9)
Tumor location			
1 localization	193 (38.9)	552 (66.0)	165(12.9)
Trigone	27 (5.4)	75 (9.0)	11(0.8)
Others	166 (33.5)	477 (57.1)	154(12.0)
>1 localization	190 (38.3)	261 (31.2)	118(9.2)
Missing	113 (22.8)	23 (2.8)	83(6.5)
Size of the largest tumor			
>=2cm	65 (13.1)	-	92(7.1)
>2-5cm	109 (22.0)	-	68(5.3)
>5cm	38 (7.7)	-	25(1.9)
< 3 cm	-	487 (58.3)	635(49.5)
≥ 3 cm	-	113 (13.5)	252(29.6)
Missing	284 (57.3)	236 (28.2)	31(2.4)
Treatment			
No BCG	248 (50.0)	566 (67.7)	-
iBCG	145 (29.2)	269 (32.2)	-
mBCG	103 (20.8)	0 (0.0)	-
TUR 'alone'	168 (33.9)	356 (42.6)	590(46.0)
TUR+BCG	207 (41.7)	248 (29.7)	436(34.0)
TUR+Chemother (Intravesical)	23 (4.6)	181 (21.7)	105(8.2)
TUR+BCG+ChemoT (Intravesical)	19 (3.8)	21 (2.5)	41(3.2)
TUR+BCG+IFN (Intravesical)	15 (3.0)	0 (0.0)	7(0.5)
Others	56 (11.3)	29 (3.5)	95(7.4)
Missing	8 (1.6)	1 (0.1)	10(0.8)
Follow-up			
Date 1st diagnosis	19/7/1999	6/13/1998	1/1/1979
Date last diagnosis	3/6/2008	6/28/2001	5/19/2010
Date last control	10/24/2008	7/1/2007	7/18/2011
Free-of-disease patients	213 (42.9)	504 (60.3)	486 (37.9)
Median follow-up (months)	75.6	77.5	26.3 – 113.0
Lost to follow-up	57 (11.5)	9 (2.7)	NA
Bladder cancer outcomes			
No. of patients with at least one tumor relapse	258 (52.0)	332 (39.7)	786(61.2)
No. of patients with at least one tumor recurrence	213 (42.9)	275 (32.9)	682(53.1)
No. of patients with tumor progression	85 (17.1)	83 (9.9)	212(16.5)

TXBC- Texas Bladder Cancer Study; SBC/EPICURO- Spanish Bladder Cancer Study; PMH - Princess Margaret Hospital; AUH - Aarhus University Hospital; HM - Hôpital Henri Mondor; EMC - Erasmus MC; HuGeF - Human Genetics Foundation.
*TaG1 includes TaLow and TaPUNLMP; **TaG3 includes TaHigh; ***T1G3 includes T1High; sd: standard deviation; NA: non-available; CIS: carcinoma in situ; iBCG: induction BCG; mBCG: maintanace BCG
Other treatments include: Cystectomy / Cystectomy+Radiotherapy, Cystectomy+Chemotherapy / Cystectomy+Chemotherapy+Radiotherapy / Chemotherapy+Radiotherapy, and Radiotherapy alone

Supplementary Table 2. Demographic, clinical, pathological, and follow-up characteristics of the NMIBC patients included in Stage II – Validation for each participating study.

Characteristics	Validation Studies						
		International Cohorts					
	TXBC-2 N (%)	POOL N (%)	PMH N (%)	AUH N (%)	HM N (%)	EMC N (%)	HuGeF N (%)
Total number	366	918	333	291	109	105	80
Sex							
Men	285(77.9)	751(81.8)	257(77.2)	227(78.0)	109(100.0)	78(74.3)	80(100.0)
Women	81(22.1)	167(18.2)	76(22.8)	64(22.0)	-	27(25.7)	-
Age							
Mean (sd) -years-	65.4(10.8)	67.6(11.1)	71.35(11.6)	66.9(9.9)	63.5(9.5)	64.6(12.2)	64.3(8.4)
Minimum - Maximum	18-93	22-97	22-97	27-86	28-79	23-88	40-75
Smoking status							
Non-smoker	104(28.4)	-	-	-	-	-	-
Former smoker	189(51.6)	-	-	-	-	-	-
Current smoker	65(17.8)	-	-	-	-	-	-
Missing	8(2.2)	-	-	-	-	-	-
Stage							
Ta	167(45.6)	573(62.4)	206(61.9)	178(61.2)	63(57.8)	72(68.6)	54(67.5)
T1	177(48.4)	309(33.7)	98(29.4)	107(36.8)	46(42.2)	33(31.4)	25(31.3)
Tis	16(4.4)	36(3.9)	29(8.7)	6(2.1)	-	-	1(1.3)
Missing	6(1.6)	-	-	-	-	-	-
Grade							
G1	14(3.8)	328(35.7)	133(39.9)	90(30.9)	15(13.8)	35(33.3)	55(68.8)
G2	117(32.0)	200(21.8)	36(10.8)	73(25.1)	48(44.0)	43(41.0)	-
G3	220(60.1)	390(42.5)	164(49.2)	128(44.0)	46(42.2)	27(25.7)	25(31.3)
Missing	15(4.1)	-	-	-	-	-	-
Stage-Grade							
TaG1*	14(3.8)	309(33.7)	129(38.7)	87(29.9)	15(13.8)	35(33.3)	43(53.8)
TaG2	92(25.1)	154(16.8)	30(9.0)	57(19.6)	38(34.9)	29(27.6)	-
TaG3**	60(16.4)	110(12.0)	47(14.1)	34(11.6)	10(9.2)	8(7.6)	11(13.8)
T1G1/T1Low		16(1.7)	4(1.2)	1(0.3)	-	-	11(13.8)
T1G2	23(6.3)	47(5.1)	6(1.8)	17(5.8)	10(9.2)	14(17.5)	-
T1G3***	153(41.8)	246(26.8)	88(26.4)	89(30.6)	36(36.0)	19(18.1)	14(17.5)
CIS	16(4.4)	36(3.9)	29(8.7)	6(2.0)	-	-	1(1.3)
Missing	8(2.2)	-	-	-	-	-	-
CIS							
Yes	114(31.1)	-	-	-	-	-	-
No	199(54.4)	-	-	-	-	-	-
Missing	53(14.5)	-	-	-	-	-	-
Number of bladder tumors							

1 tumor	121(33.1)	576(68.0)	180(54.1)	167(75.9)	84(77.1)	68(64.8)	77(96.3)
> 1 tumor	151(41.3)	271(32.0)	153(45.9)	53(24.1)	25(22.9)	37(35.2)	3(3.7)
Missing	94(25.7)	71(7.7)	-	71(24.4)	-	-	-
Tumor location							
1 localization	165(45.1)	-	-	-	-	-	-
Trigone	11(3.0)	-	-	-	-	-	-
Others	154(42.1)	-	-	-	-	-	-
>1 localization	118(32.2)	-	-	-	-	-	-
Missing	83(22.7)	-	-	-	-	-	-
Size of the largest tumor		-					
>=2cm	92(25.1)	-	-	-	-	-	-
>2-5cm	68(18.6)	-	-	-	-	-	-
>5cm	25(6.8)	-	-	-	-	-	-
< 3 cm.	-	635(71.6)	266(79.9)	187(71.1)	58(53.2)	68(66.7)	56(70.0)
≥ 3 cm.	-	252(28.4)	67(20.1)	76(28.9)	51(46.8)	34(33.3)	24(30.0)
Missing	181(49.5)	31(3.4)	-	28(9.6)	-	3(2.9)	-
Treatment							
No BCG	186(50.8)	585(63.7)	191(57.4)	232(79.7)	57(52.3)	59(56.2)	46(57.5)
iBCG	93(25.4)	333(36.3)	142(42.7)	59(20.3)	52(47.7)	46(43.8)	34(42.5)
mBCG	77(21.0)	0	0	0	0	0	0
TUR 'alone'	105(28.7)	485(52.8)	165(49.5)	226(77.7)	37(33.9)	50(47.6)	7(8.8)
TUR+BCG	128(35.0)	308(33.6)	138(41.4)	58(19.9)	52(47.7)	34(32.4)	26(32.5)
TUR+Chemother (Intravesical)	27(7.4)	78(8.5)	21(6.3)	1(0.3)	12(11.1)	9(8.6)	35(43.8)
TUR+BCG +ChemoT (Intravesical)	16(4.4)	25(2.7)	4(1.2)	1(0.3)	0(0.0)	12(11.4)	8(10.0)
TUR+BCG+IFN (Intravesical)	7(1.9)	-	-	-	-	-	-
Others	73(19.9)	22(2.4)	5(1.5)	5(1.7)	8(7.3)	-	4(5.0)
Missing	10(2.7)	-	-	-	-	-	-
Follow-up							
Date 1st diagnosis	25/2/1982	1/1/1979	2/21/1980	1/1/1979	5/3/1995	1/13/1983	10/4/1999
Date last diagnosis	12/29/2009	5/19/2010	5/19/2010	8/22/2007	8/21/2006	3/23/2007	6/4/2009
Date last control	6/11/2010	7/18/2011	3/28/2011	1/14/2009	7/18/2011	7/28/2010	3/25/2010
Free-of-disease patients	174(47.5)	312(34.0)	138(15.0)	30(3.3)	59(6.4)	32(3.5)	53(5.8)
Median follow-up (m)	43.7	58.8	55.1	26.3	69.5	113.0	36.3
Lost to follow-up	17 (4.6)	NA	NA	NA	NA	NA	NA
Bladder cancer outcomes							
No. of patients with at least one tumor relapse	180(49.2)	606(66.0)	195(58.9)	261(89.7)	50(45.9)	73(69.5)	27(33.7)
No. of patients with at least one tumor recurrence	142(38.8)	540(58.8)	180(54.5)	223(76.6)	45(41.2)	71(67.6)	21(26.2)
No. of patients with tumor progression	71(19.4)	141(15.3)	17(5.1)	90(30.9)	8(7.3)	18(17.1)	8(10.0)

TXBC-2- Texas Bladder Cancer Study for Validation; PMH - Princess Margaret Hospital; AUH - Aarhus University Hospital; HM - Hôpital Henri Mondor; EMC - Erasmus MC; HuGeF - Human Genetics Foundation.
*TaG1 includes TaLow and TaPUNLMP; **TaG3 includes TaHigh; ***T1G3 includes T1High; sd: standard deviation; NA: non-available; CIS: carcinoma in situ; iBCG: induction BCG; mBCG: maintanace BCG
Other treatments include: Cystectomy / Cystectomy+Radiotherapy, Cystectomy+Chemotherapy / Cystectomy+Chemotherapy+Radiotherapy / Chemotherapy+Radiotherapy, and Radiotherapy alone

Supplementary Table 3. Demographic, clinical and pathological characteristics, and follow-up of the MIBC patients included in Discovery phase.

Characteristics	Discovery	
	TXBC-1 N (%)	**SBC/EPICURO N (%)**
Total number	397	235
Sex		
Men	311 (78.3)	209 (88.9)
Women	86 (21.7)	26 (11.1)
Age		
Mean (sd) –years-	66.4 (10.7)	66.9 (8.9)
Mínimum - Maximum	39-89	36 - 80
Smoking status		
Non-smoker	94 (23.7)	30 (12.8)
Former smoker	202 (50.9)	78 (33.2)
Current smoker	101 (25.4)	127 (54.0)
Missing	0 (0.0)	0 (0.0)
Stage		
T1	4 (1.0)	0 (0.0)
T2	300 (75.6)	129 (54.9)
T3	57 (14.4)	54 (23.0)
T4	32 (8.1)	52 (22.1)
Missing	4 (1.0)	0 (0.0)
Grade		
G2	11 (2.8)	22 (9.4)
G3	375 (94.5)	213 (90.6)
Missing	11 (2.8)	0 (0.0)
Stage-Grade		
T1G2	0 (0.0)	0 (0.0)
T1G3	4 (1.0)	0 (0.0)
T2G2	10 (2.5)	10 (4.3)
T2G3	282 (71.0)	119 (50.6)
T3G2	0 (0.0)	6 (2.6)
T3G3	55 (13.9)	48 (20.4)
T4G2	1 (0.3)	6 (2.6)
T4G3	30 (7.6)	46 (19.6)
Missing	15 (3.8)	0 (0.0)
Number of bladder tumors		
1 tumor	220 (55.4)	163 (69.4)
> 1 tumor	79 (19.9)	53 (22.5)
Missing	98 (24.7)	19 (8.1)
Tumor location		
Trigone	34 (8.6)	28 (11.9)

Others	139 (35.0)	84 (35.7)
>1 localization	193 (48.6)	118 (50.2)
Missing	31 (7.8)	5 (2.1)
Size of the largest tumor		
>=2cm	15 (3.8)	-
>2-5cm	182 (45.8)	-
>5cm	62 (15.6)	-
< 3 cm	-	63 (26.8)
≥ 3 cm	-	68 (28.9)
Missing	138 (34.8)	104 (44.3)
Metastasis (M)		
M0	370 (93.2)	188 (80.0)
M1	20 (5.0)	27 (11.5)
Mx	7 (1.8)	20 (8.5)
Affected ganglia (N)		
N0	349 (87.9)	162 (68.9)
N1-N3	38 (9.6)	47 (20.0)
Nx	10 (2.5)	26 (11.1)
Treatment		
TUR+Cystectomy	116 (29.2)	82 (34.9)
TUR+Cystectomy+Chemotherapy	116 (29.2)	32 (13.6)
TUR+Chemothearpy (alone)	55 (13.9)	23 (9.8)
TUR+Radiotherapy+Chemotherapy	11 (2.8)	24 (10.2)
NMIBC treatment	42 (10.6)	18 (7.7)
Others	53 (13.4)	56 (23.8)
Missing	4 (1.0)	0 (0.0)
Follow-up		
Date 1st diagnosis	12/31/1997	6/28/1998
Date last diagnosis	9/24/2007	4/25/2001
Date last control	02/05/2009	7/1/2007
Free-of-disease patients	200(50.3)	66 (28.1)
Median follow-up (months)	43.7	26
Lost to follow-up	5	3
Bladder cancer outcomes		
No. of patients with tumor progression	74(18.6)	129 (54.9)
No. of patients dead by bladder cancer	97(24.4)	108 (46.0)
No. of patients dead by any cause	161(40.6)	161 (68.5)

TXBC- Texas Bladder Cancer Study; SBC/EPICURO- Spanish Bladder Cancer Study

Supplementary Table 4. NMIBC associations of SNPs identified in Stage I analysis with risk of recurrence, progression, and relapse: Meta-analysis for Stage I, Stage II, and All studies combined, Chromosome and gene location, minor allele, mode-of inheritance, and statistical model are also displayed.

SNP	Model			Discovery Meta-analysis results							
	Outcome	MoI	Alleles	MAF SBC/EPICURO	MAF TXBC-1	HR(95% CI) fixed	P fixed	HR(95% CI) random	P random	P het	I^2
rs754799	Recurrence	recessive	A/C	0.19	0.17	2.79(1.86-4.17)	5.87E-07	2.79(1.86-4.17)	5.87E-07	7.98E-01	0.00
rs12988804	Recurrence	codom.het	G/A	0.25	0.28	1.60(1.32-1.94)	1.45E-06	1.60(1.32-1.94)	1.45E-06	4.94E-01	0.00
rs489770	Recurrence	recessive	A/G	0.18	0.20	2.55(1.72-3.77)	2.68E-06	2.55(1.72-3.77)	2.68E-06	7.23E-01	0.00
rs980761	Recurrence	codom.het	A/G	0.38	0.37	1.62(1.32-2.00)	4.24E-06	1.62(1.32-2.00)	4.24E-06	5.07E-01	0.00
rs3736994	Recurrence	dominant	A/G	0.31	0.33	0.65(0.54-0.78)	4.61E-06	0.65(0.54-0.78)	4.61E-06	3.84E-01	0.00
rs7930624	Recurrence	additive	C/A	0.24	0.23	1.39(1.20-1.61)	9.67E-06	1.39(1.20-1.61)	9.67E-06	7.36E-01	0.00
rs12681007	Recurrence	recessive	G/A	0.40	0.46	1.65(1.31-2.07)	1.56E-05	1.65(1.31-2.07)	1.56E-05	3.23E-01	0.00
rs6443819	Recurrence	additive	A/G	0.27	0.23	1.37(1.19-1.59)	1.56E-05	1.37(1.19-1.59)	1.56E-05	3.85E-01	0.00
rs10867878	Recurrence	recessive	G/A	0.22	0.19	2.28(1.56-3.34)	2.09E-05	2.28(1.56-3.34)	2.09E-05	8.58E-01	0.00
rs9504361	Recurrence	codom.het	G/A	0.47	0.46	0.63(0.50-0.78)	2.92E-05	0.63(0.50-0.78)	2.92E-05	4.94E-01	0.00
rs6512003	Recurrence	recessive	C/A	0.42	0.40	1.63(1.29-2.04)	3.00E-05	1.63(1.29-2.04)	3.00E-05	5.46E-01	0.00
rs6505263	Recurrence	additive	G/A	0.27	0.25	0.71(0.61-0.84)	3.73E-05	0.71(0.61-0.84)	3.73E-05	5.87E-01	0.00
rs2943313	Recurrence	dominant	G/A	0.03	0.03	2.00(1.45-2.76)	2.37E-05	2.00(1.44-2.79)	3.76E-05	3.05E-01	0.05
rs2191031	Recurrence	dominant	G/A	0.16	0.18	0.63(0.51-0.78)	3.10E-05	0.63(0.50-0.80)	1.05E-04	2.85E-01	0.13
rs1421776	Progression	codom.hom	A/G	0.13	0.12	7.63(3.64-15.96)	6.96E-08	7.63(3.64-15.96)	6.96E-08	8.95E-01	0.00
rs4246835	Progression	codom.het	G/A	0.40	0.42	0.38(0.26-0.55)	3.72E-07	0.38(0.26-0.55)	3.72E-07	6.21E-01	0.00
rs10167220	Progression	dominant	G/A	0.11	0.11	2.34(1.68-3.28)	6.70E-07	2.34(1.68-3.28)	6.70E-07	6.58E-01	0.00
rs526509	Progression	additive	C/A	0.42	0.39	1.77(1.41-2.23)	7.61E-07	1.77(1.41-2.23)	7.61E-07	5.18E-01	0.00
rs6100810	Progression	recessive	A/G	0.19	0.18	4.51(2.43-8.40)	1.95E-06	4.51(2.43-8.40)	1.95E-06	3.29E-01	0.00
rs12294567	Progression	codom.hom	A/G	0.13	0.12	5.48(2.71-11.08)	2.17E-06	5.48(2.71-11.08)	2.17E-06	6.09E-01	0.00
rs4246835	Progression	dominant	G/A	0.40	0.42	0.46(0.33-0.63)	3.07E-06	0.46(0.33-0.63)	3.07E-06	6.84E-01	0.00
rs17218455	Progression	codom.hom	G/A	0.17	0.16	3.98(2.20-7.20)	4.89E-06	3.98(2.20-7.20)	4.89E-06	3.72E-01	0.00
rs6752816	Progression	dominant	G/A	0.06	0.07	2.52(1.69-3.75)	5.26E-06	2.52(1.69-3.75)	5.26E-06	4.90E-01	0.00
rs2950650	Progression	codom.hom	A/G	0.13	0.15	4.58(2.38-8.82)	5.35E-06	4.58(2.38-8.82)	5.35E-06	5.76E-01	0.00
rs9891348	Progression	codom.het	A/G	0.43	0.42	0.38(0.26-0.56)	6.72E-07	0.38(0.25-0.58)	5.42E-06	2.75E-01	0.16
rs5027573	Progression	dominant	G/A	0.05	0.03	2.72(1.77-4.19)	5.49E-06	2.72(1.77-4.19)	5.49E-06	4.76E-01	0.00
rs17831395	Progression	dominant	A/G	0.02	0.02	3.40(2.01-5.77)	5.50E-06	3.40(2.01-5.77)	5.50E-06	3.77E-01	0.00
rs1568519	Progression	dominant	A/C	0.06	0.07	2.57(1.71-3.87)	5.75E-06	2.57(1.71-3.87)	5.75E-06	4.10E-01	0.00
rs7721273	Progression	codom.hom	A/G	0.19	0.21	4.41(2.32-8.38)	5.76E-06	4.41(2.32-8.38)	5.76E-06	4.39E-01	0.00
rs7588481	Progression	dominant	A/C	0.18	0.14	2.13(1.52-2.97)	9.41E-06	2.13(1.52-2.97)	9.41E-06	4.83E-01	0.00
rs6531449	Progression	additive	A/G	0.31	0.27	1.71(1.34-2.17)	1.22E-05	1.71(1.34-2.17)	1.22E-05	7.60E-01	0.00

SNP	Model			Discovery Meta-analysis results							
	Outcome	MoI	Alleles	MAF SBC/EPICURO	MAF TXBC-1	HR(95% CI) fixed	P fixed	HR(95% CI) random	P random	P het	I^2
rs7936809	Progression	dominant	G/A	0.21	0.25	2.08(1.50-2.89)	1.30E-05	2.08(1.50-2.89)	1.30E-05	5.22E-01	0.00
rs3797725	Progression	codom.hom	A/G	0.23	0.18	3.20(1.89-5.42)	1.49E-05	3.20(1.89-5.42)	1.49E-05	7.41E-01	0.00
rs7572970	Progression	recessive	G/A	0.28	0.29	2.94(1.86-4.66)	4.27E-06	2.94(1.79-4.84)	2.16E-05	2.79E-01	0.15
rs10437619	Progression	additive	A/G	0.38	0.40	1.60(1.28-2.00)	3.31E-05	1.60(1.28-2.00)	3.31E-05	4.87E-01	0.00
rs1035856	Progression	dominant	G/A	0.02	0.02	3.36(1.96-5.76)	1.11E-05	3.35(1.88-5.97)	4.22E-05	2.84E-01	0.13
rs1847325	Progression	additive	G/A	0.47	0.48	0.59(0.45-0.76)	5.69E-05	0.59(0.45-0.76)	5.69E-05	3.68E-01	0.00
rs1432374	Progression	additive	G/A	0.36	0.33	1.56(1.25-1.95)	7.44E-05	1.56(1.25-1.95)	7.44E-05	3.21E-01	0.00
rs10783528	Progression	additive	G/A	0.38	0.33	1.62(1.29-2.04)	3.79E-05	1.62(1.27-2.06)	1.04E-04	2.88E-01	0.11
rs6757412	Progression	codom.hom	A/C	0.28	0.25	3.07(1.81-5.23)	3.45E-05	3.06(1.69-5.54)	2.24E-04	2.64E-01	0.20
rs3772337	Relapse	recessive	A/G	0.14	0.14	3.71(2.26-6.08)	2.01E-07	3.71(2.26-6.08)	2.01E-07	8.65E-01	0.00
rs11604069	Relapse	dominant	G/A	0.45	0.43	0.65(0.54-0.78)	1.67E-06	0.65(0.54-0.78)	1.67E-06	4.39E-01	0.00
rs446027	Relapse	codom.hom	A/G	0.36	0.32	1.84(1.43-2.36)	1.85E-06	1.84(1.43-2.36)	1.85E-06	6.52E-01	0.00
rs2694787	Relapse	codom.hom	G/A	0.19	0.20	2.53(1.71-3.75)	3.57E-06	2.53(1.71-3.75)	3.57E-06	7.27E-01	0.00
rs2834651	Relapse	codom.het	A/G	0.34	0.34	1.60(1.33-1.93)	4.36E-07	1.61(1.31-1.97)	3.74E-06	2.74E-01	0.16
rs11615759	Relapse	recessive	A/G	0.16	0.14	2.62(1.73-3.96)	5.02E-06	2.62(1.73-3.96)	5.02E-06	9.00E-01	0.00
rs3736994	Relapse	dominant	A/G	0.31	0.33	0.68(0.57-0.80)	5.79E-06	0.68(0.57-0.80)	5.79E-06	3.99E-01	0.00
rs4920993	Relapse	recessive	A/G	0.27	0.29	1.84(1.41-2.40)	6.41E-06	1.84(1.41-2.40)	6.41E-06	8.62E-01	0.00
rs8111608	Relapse	additive	G/A	0.34	0.33	0.74(0.65-0.84)	6.59E-06	0.74(0.65-0.84)	6.59E-06	8.10E-01	0.00
rs7572970	Relapse	recessive	G/A	0.28	0.30	1.85(1.41-2.43)	8.45E-06	1.85(1.41-2.43)	8.45E-06	4.68E-01	0.00
rs962312	Relapse	dominant	G/A	0.48	0.47	1.66(1.33-2.07)	8.56E-06	1.66(1.33-2.07)	8.56E-06	7.35E-01	0.00
rs754799	Relapse	recessive	A/C	0.19	0.17	2.39(1.62-3.51)	9.92E-06	2.39(1.62-3.51)	9.92E-06	9.03E-01	0.00
rs4946483	Relapse	additive	G/A	0.38	0.47	1.31(1.16-1.48)	1.00E-05	1.31(1.16-1.48)	1.00E-05	5.38E-01	0.00
rs9533040	Relapse	codom.hom	A/G	0.15	0.14	2.64(1.75-3.98)	4.10E-06	2.64(1.72-4.07)	1.03E-05	2.96E-01	0.09
rs846978	Relapse	dominant	G/A	0.11	0.08	1.58(1.29-1.94)	1.15E-05	1.58(1.29-1.94)	1.15E-05	7.97E-01	0.00
rs923435	Relapse	codom.het	A/C	0.28	0.29	0.66(0.55-0.80)	1.16E-05	0.66(0.55-0.80)	1.16E-05	5.22E-01	0.00
rs3917265	Relapse	dominant	G/A	0.46	0.47	1.55(1.27-1.89)	1.41E-05	1.55(1.27-1.89)	1.41E-05	4.44E-01	0.00
rs12435167	Relapse	dominant	G/A	0.29	0.31	0.69(0.58-0.81)	1.43E-05	0.69(0.58-0.82)	1.71E-05	3.13E-01	0.02
rs2028008	Relapse	dominant	A/G	0.29	0.30	1.44(1.22-1.71)	2.21E-05	1.44(1.22-1.71)	2.21E-05	4.33E-01	0.00
rs10888205	Relapse	recessive	G/A	0.23	0.23	1.95(1.43-2.67)	2.35E-05	1.95(1.43-2.67)	2.35E-05	3.66E-01	0.00
rs1036332	Relapse	additive	C/A	0.27	0.27	0.72(0.63-0.83)	6.29E-06	0.72(0.62-0.84)	5.01E-05	2.64E-01	0.20

SNP	Model			Validation Meta-analysis results					
	Outcome	MoI	Alleles	HR(95% CI) fixed	P fixed	HR(95% CI) random	P random	P het	I^2
rs754799	Recurrence	recessive	A/C	1.51(1.00-2.27)	4.76E-02	1.51(1.00-2.27)	4.76E-02	8.40E-01	0.0
rs12988804	Recurrence	codom.het	G/A	1.08(0.90-1.28)	4.08E-01	1.01(0.73-1.41)	9.41E-01	1.36E-01	55.0
rs489770	Recurrence	recessive	A/G	0.85(0.55-1.31)	4.68E-01	0.75(0.36-1.59)	4.52E-01	1.58E-01	49.8
rs980761	Recurrence	codom.het	A/G	1.06(0.89-1.27)	5.27E-01	1.00(0.73-1.37)	9.91E-01	1.52E-01	51.4
rs3736994	Recurrence	dominant	A/G	1.02(0.86-1.21)	8.30E-01	0.98(0.76-1.28)	8.96E-01	2.02E-01	38.5
rs7930624	Recurrence	additive	C/A	1.00(0.86-1.16)	9.97E-01	1.00(0.86-1.16)	9.97E-01	7.24E-01	0.0
rs12681007	Recurrence	recessive	G/A	1.15(0.94-1.40)	1.76E-01	1.18(0.89-1.57)	2.38E-01	2.28E-01	31.1
rs6443819	Recurrence	additive	A/G	1.06(0.92-1.21)	4.29E-01	1.14(0.82-1.58)	4.24E-01	6.42E-02	70.8
rs10867878	Recurrence	recessive	G/A	0.81(0.53-1.23)	3.21E-01	0.81(0.53-1.23)	3.21E-01	9.57E-01	0.0
rs9504361	Recurrence	codom.het	G/A	1.00(0.82-1.22)	9.94E-01	0.93(0.65-1.35)	7.18E-01	1.86E-01	42.8
rs6512003	Recurrence	recessive	C/A	0.99(0.80-1.24)	9.43E-01	0.94(0.66-1.34)	7.34E-01	2.04E-01	38.1
rs6505263	Recurrence	additive	G/A	0.99(0.86-1.13)	8.35E-01	0.96(0.79-1.17)	7.03E-01	2.27E-01	31.3
rs2943313	Recurrence	dominant	G/A	0.97(0.70-1.34)	8.47E-01	0.97(0.70-1.34)	8.47E-01	7.31E-01	0.0
rs2191031	Recurrence	dominant	G/A	0.94(0.79-1.13)	5.12E-01	0.88(0.62-1.24)	4.66E-01	1.28E-01	56.7
rs1421776	Progression	codom.hom	A/G	0.96(0.23-3.97)	9.57E-01	0.96(0.23-3.97)	9.57E-01	NA	NA
rs4246835	Progression	codom.het	G/A	0.64(0.43-0.94)	2.36E-02	0.57(0.25-1.28)	1.72E-01	5.14E-02	73.7
rs10167220	Progression	dominant	G/A	0.49(0.29-0.84)	8.86E-03	0.49(0.29-0.84)	8.86E-03	3.22E-01	0.0
rs526509	Progression	additive	C/A	1.09(0.86-1.38)	4.67E-01	1.09(0.86-1.38)	4.67E-01	6.36E-01	0.0
rs6100810	Progression	recessive	A/G	1.55(0.70-3.47)	2.83E-01	1.55(0.70-3.47)	2.83E-01	6.40E-01	0.0
rs12294567	Progression	codom.hom	A/G	0.56(0.10-3.07)	5.06E-01	0.56(0.10-3.07)	5.06E-01	7.68E-01	0.0
rs4246835	Progression	dominant	G/A	0.74(0.52-1.05)	8.71E-02	0.66(0.30-1.42)	2.85E-01	4.24E-02	75.7
rs17218455	Progression	codom.hom	G/A	1.32(0.57-3.04)	5.19E-01	1.32(0.57-3.04)	5.19E-01	5.19E-01	0.0
rs6752816	Progression	dominant	G/A	0.96(0.58-1.58)	8.58E-01	0.92(0.51-1.67)	7.94E-01	2.67E-01	19.0
rs2950650	Progression	codom.hom	A/G	0.76(0.10-5.52)	7.86E-01	0.76(0.10-5.52)	7.86E-01	NA	NA
rs9891348	Progression	codom.het	A/G	1.44(0.96-2.17)	7.70E-02	1.44(0.96-2.17)	7.70E-02	3.66E-01	0.0
rs5027573	Progression	dominant	G/A	0.64(0.29-1.40)	2.63E-01	0.64(0.29-1.40)	2.63E-01	5.32E-01	0.0
rs17831395	Progression	dominant	A/G	0.26(0.04-1.92)	1.88E-01	0.26(0.04-1.92)	1.88E-01	NA	NA
rs1568519	Progression	dominant	A/C	0.98(0.56-1.71)	9.33E-01	1.01(0.44-2.31)	9.76E-01	1.44E-01	53.2
rs7721273	Progression	codom.hom	A/G	0.86(0.34-2.20)	7.61E-01	0.86(0.34-2.20)	7.61E-01	8.96E-01	0.0
rs7588481	Progression	dominant	A/C	1.01(0.69-1.48)	9.40E-01	1.01(0.69-1.48)	9.40E-01	3.47E-01	0.0
rs6531449	Progression	additive	A/G	0.80(0.61-1.04)	9.55E-02	0.80(0.61-1.04)	9.55E-02	9.30E-01	0.0
rs7936809	Progression	dominant	G/A	0.82(0.58-1.15)	2.51E-01	0.91(0.41-2.05)	8.24E-01	2.55E-02	80.0
rs3797725	Progression	codom.hom	A/G	1.06(0.46-2.47)	8.91E-01	1.06(0.46-2.47)	8.91E-01	3.29E-01	0.0
rs7572970	Progression	recessive	G/A	1.27(0.65-2.47)	4.78E-01	1.27(0.65-2.47)	4.78E-01	6.89E-01	0.0

SNP	Model			Validation Meta-analysis results					
	Outcome	MoI	Alleles	HR(95% CI) fixed	P fixed	HR(95% CI) random	P random	P het	I^2
rs10437619	Progression	additive	A/G	0.92(0.73-1.17)	5.09E-01	0.88(0.59-1.32)	5.33E-01	1.11E-01	60.6
rs1035856	Progression	dominant	G/A	1.43(0.69-2.98)	3.36E-01	1.43(0.69-2.98)	3.36E-01	NA	NA
rs1847325	Progression	additive	G/A	1.16(0.91-1.48)	2.38E-01	1.16(0.91-1.48)	2.38E-01	6.87E-01	0.0
rs1432374	Progression	additive	G/A	0.75(0.58-0.98)	3.56E-02	0.75(0.58-0.98)	3.56E-02	4.02E-01	0.0
rs10783528	Progression	additive	G/A	0.94(0.73-1.21)	6.29E-01	0.91(0.64-1.30)	5.99E-01	1.84E-01	43.2
rs6757412	Progression	codom.hom	A/C	0.65(0.32-1.34)	2.47E-01	0.65(0.32-1.34)	2.47E-01	8.90E-01	0.0
rs3772337	Relapse	recessive	A/G	1.06(0.66-1.69)	8.18E-01	1.21(0.49-2.97)	6.85E-01	7.73E-02	67.9
rs11604069	Relapse	dominant	G/A	1.12(0.94-1.34)	2.16E-01	1.12(0.94-1.34)	2.16E-01	8.55E-01	0.0
rs446027	Relapse	codom.hom	A/G	0.93(0.73-1.19)	5.52E-01	0.93(0.73-1.19)	5.52E-01	5.23E-01	0.0
rs2694787	Relapse	codom.hom	G/A	0.95(0.59-1.53)	8.35E-01	0.95(0.59-1.53)	8.35E-01	6.29E-01	0.0
rs2834651	Relapse	codom.het	A/G	0.99(0.84-1.17)	8.92E-01	1.08(0.73-1.61)	6.99E-01	5.38E-02	73.1
rs11615759	Relapse	recessive	A/G	1.06(0.52-2.15)	8.77E-01	1.06(0.52-2.15)	8.77E-01	NA	NA
rs3736994	Relapse	dominant	A/G	1.07(0.91-1.26)	3.99E-01	1.07(0.91-1.26)	3.99E-01	7.25E-01	0.0
rs4920993	Relapse	recessive	A/G	1.18(0.87-1.60)	2.99E-01	0.97(0.45-2.07)	9.37E-01	7.52E-02	68.4
rs8111608	Relapse	additive	G/A	0.97(0.86-1.09)	5.75E-01	0.97(0.86-1.09)	5.75E-01	7.56E-01	0.0
rs7572970	Relapse	recessive	G/A	0.80(0.58-1.11)	1.87E-01	0.66(0.29-1.48)	3.09E-01	1.48E-01	52.3
rs962312	Relapse	dominant	G/A	1.08(0.91-1.29)	3.83E-01	1.08(0.91-1.29)	3.83E-01	4.37E-01	0.0
rs754799	Relapse	recessive	A/C	1.48(1.00-2.20)	4.91E-02	1.48(1.00-2.20)	4.91E-02	7.43E-01	0.0
rs4946483	Relapse	additive	G/A	1.16(0.91-1.49)	2.29E-01	1.16(0.91-1.49)	2.29E-01	NA	NA
rs9533040	Relapse	codom.hom	A/G	0.64(0.09-4.70)	6.61E-01	0.64(0.09-4.70)	6.61E-01	NA	NA
rs846978	Relapse	dominant	G/A	0.97(0.77-1.22)	8.02E-01	0.97(0.77-1.22)	8.02E-01	6.86E-01	0.0
rs923435	Relapse	codom.het	A/C	1.01(0.86-1.19)	8.93E-01	1.01(0.86-1.19)	8.93E-01	6.77E-01	0.0
rs3917265	Relapse	dominant	G/A	0.87(0.73-1.03)	9.79E-02	0.87(0.73-1.03)	9.79E-02	9.14E-01	0.0
rs12435167	Relapse	dominant	G/A	1.08(0.93-1.27)	3.19E-01	1.15(0.85-1.56)	3.75E-01	1.07E-01	61.4
rs2028008	Relapse	dominant	A/G	1.02(0.87-1.20)	7.92E-01	1.00(0.82-1.23)	9.62E-01	2.51E-01	24.0
rs10888205	Relapse	recessive	G/A	0.70(0.48-1.02)	6.02E-02	0.70(0.48-1.02)	6.02E-02	7.82E-01	0.0
rs1036332	Relapse	additive	C/A	1.00(0.88-1.13)	9.84E-01	1.00(0.88-1.13)	9.84E-01	8.63E-01	0.0

SNP	Model			All Combined Meta-analysis results					
	Outcome	MoI	Alleles	HR(95% CI) fixed	P fixed	HR(95% CI) random	P random	P het	I^2
rs754799	Recurrence	recessive	A/C	2.06(1.55- 2.75)	7.45E-07	2.10(1.46- 3.04)	7.71E-05	2.13E-01	33.2
rs12988804	Recurrence	codom.het	G/A	1.29(1.13- 1.47)	1.08E-04	1.27(0.96- 1.66)	8.90E-02	8.90E-03	74.1
rs489770	Recurrence	recessive	A/G	1.56(1.17- 2.09)	2.68E-03	1.39(0.69- 2.80)	3.50E-01	1.33E-03	80.8
rs980761	Recurrence	codom.het	A/G	1.27(1.11- 1.45)	5.05E-04	1.26(0.95- 1.68)	1.14E-01	7.55E-03	74.9
rs3736994	Recurrence	dominant	A/G	0.83(0.73- 0.94)	3.35E-03	0.79(0.58- 1.06)	1.10E-01	1.98E-03	79.7
rs7930624	Recurrence	additive	C/A	1.18(1.06- 1.30)	1.81E-03	1.20(0.98- 1.46)	8.05E-02	1.79E-02	70.2
rs12681007	Recurrence	recessive	G/A	1.34(1.16- 1.56)	1.14E-04	1.42(1.09- 1.84)	8.55E-03	4.54E-02	62.6
rs6443819	Recurrence	additive	A/G	1.20(1.08- 1.32)	3.89E-04	1.25(1.03- 1.53)	2.60E-02	1.24E-02	72.4
rs10867878	Recurrence	recessive	G/A	1.44(1.08- 1.91)	1.18E-02	1.42(0.75- 2.68)	2.86E-01	5.13E-03	76.5
rs9504361	Recurrence	codom.het	G/A	0.81(0.69- 0.94)	4.48E-03	0.75(0.55- 1.04)	8.34E-02	8.86E-03	74.2
rs6512003	Recurrence	recessive	C/A	1.26(1.08- 1.48)	4.33E-03	1.23(0.88- 1.71)	2.28E-01	1.04E-02	73.4
rs6505263	Recurrence	additive	G/A	0.86(0.78- 0.96)	5.44E-03	0.82(0.66- 1.02)	7.00E-02	1.14E-02	72.9
rs2943313	Recurrence	dominant	G/A	1.39(1.11- 1.75)	4.25E-03	1.41(0.89- 2.22)	1.42E-01	1.23E-02	72.5
rs2191031	Recurrence	dominant	G/A	0.80(0.70- 0.92)	1.61E-03	0.74(0.55- 0.99)	4.31E-02	1.02E-02	73.4
rs1421776	Progression	codom.hom	A/G	4.90(2.55- 9.44)	1.96E-06	4.18(1.25-13.90)	1.99E-02	3.95E-02	69.0
rs4246835	Progression	codom.het	G/A	0.49(0.37- 0.64)	1.77E-07	0.46(0.30- 0.72)	6.51E-04	5.20E-02	61.2
rs10167220	Progression	dominant	G/A	1.51(1.13- 2.00)	4.83E-03	1.16(0.49- 2.72)	7.41E-01	1.71E-05	87.9
rs526509	Progression	additive	C/A	1.41(1.19- 1.66)	4.71E-05	1.39(1.04- 1.85)	2.65E-02	2.85E-02	66.9
rs6100810	Progression	recessive	A/G	3.03(1.85- 4.95)	9.76E-06	2.78(1.41- 5.48)	3.13E-03	1.44E-01	44.5
rs12294567	Progression	codom.hom	A/G	3.92(2.05- 7.52)	3.76E-05	2.66(0.86- 8.18)	8.85E-02	1.00E-01	51.9
rs4246835	Progression	dominant	G/A	0.57(0.45- 0.72)	4.74E-06	0.55(0.37- 0.82)	3.47E-03	4.52E-02	62.7
rs17218455	Progression	codom.hom	G/A	2.75(1.70- 4.47)	4.08E-05	2.51(1.24- 5.07)	1.07E-02	1.28E-01	47.2
rs6752816	Progression	dominant	G/A	1.74(1.27- 2.37)	5.36E-04	1.54(0.84- 2.83)	1.63E-01	1.48E-02	71.4
rs2950650	Progression	codom.hom	A/G	3.84(2.06- 7.15)	2.28E-05	3.43(1.47- 7.98)	4.30E-03	2.07E-01	36.6
rs9891348	Progression	codom.het	A/G	0.71(0.54- 0.94)	1.50E-02	0.72(0.33- 1.59)	4.21E-01	2.60E-05	87.5
rs5027573	Progression	dominant	G/A	1.94(1.33- 2.82)	5.95E-04	1.53(0.70- 3.33)	2.87E-01	1.16E-02	72.8
rs17831395	Progression	dominant	A/G	2.87(1.72- 4.78)	5.05E-05	2.15(0.74- 6.22)	1.59E-01	3.44E-02	70.3
rs1568519	Progression	dominant	A/C	1.84(1.32- 2.55)	3.03E-04	1.68(0.90- 3.14)	1.06E-01	1.59E-02	71.0
rs7721273	Progression	codom.hom	A/G	2.62(1.54- 4.44)	3.65E-04	2.20(0.87- 5.54)	9.47E-02	3.56E-02	65.0
rs7588481	Progression	dominant	A/C	1.54(1.20- 1.98)	7.34E-04	1.52(0.96- 2.38)	7.20E-02	2.22E-02	68.8
rs6531449	Progression	additive	A/G	1.21(1.01- 1.44)	3.53E-02	1.18(0.76- 1.82)	4.69E-01	5.36E-04	82.9
rs7936809	Progression	dominant	G/A	1.33(1.05- 1.69)	1.87E-02	1.39(0.74- 2.59)	3.03E-01	1.54E-04	85.2
rs3797725	Progression	codom.hom	A/G	2.35(1.50- 3.67)	1.80E-04	2.10(1.07- 4.13)	3.18E-02	1.22E-01	48.2
rs7572970	Progression	recessive	G/A	2.24(1.54- 3.27)	2.88E-05	2.16(1.26- 3.70)	4.91E-03	1.40E-01	45.2

SNP	Model			All Combined Meta-analysis results					
	Outcome	MoI	Alleles	HR(95% CI) fixed	P fixed	HR(95% CI) random	P random	P het	I^2
rs10437619	Progression	additive	A/G	1.24(1.05- 1.45)	1.00E-02	1.19(0.83- 1.69)	3.49E-01	2.83E-03	78.7
rs1035856	Progression	dominant	G/A	2.49(1.61- 3.84)	3.99E-05	2.49(1.30- 4.79)	6.14E-03	1.05E-01	55.6
rs1847325	Progression	additive	G/A	0.84(0.71- 1.01)	5.77E-02	0.82(0.55- 1.23)	3.42E-01	1.85E-03	80.0
rs1432374	Progression	additive	G/A	1.16(0.98- 1.37)	9.06E-02	1.08(0.70- 1.67)	7.35E-01	2.82E-04	84.2
rs10783528	Progression	additive	G/A	1.26(1.07- 1.49)	6.83E-03	1.21(0.85- 1.73)	2.86E-01	5.13E-03	76.5
rs6757412	Progression	codom.hom	A/C	1.78(1.16- 2.73)	8.18E-03	1.48(0.60- 3.65)	3.90E-01	5.16E-03	76.5
rs3772337	Relapse	recessive	A/G	1.91(1.36- 2.69)	1.83E-04	2.14(0.94- 4.88)	6.89E-02	1.02E-03	81.5
rs11604069	Relapse	dominant	G/A	0.85(0.75- 0.96)	1.09E-02	0.84(0.61- 1.16)	2.92E-01	3.26E-04	83.9
rs446027	Relapse	codom.hom	A/G	1.30(1.09- 1.55)	3.53E-03	1.34(0.89- 2.02)	1.66E-01	1.65E-03	80.3
rs2694787	Relapse	codom.hom	G/A	1.71(1.26- 2.31)	5.63E-04	1.58(0.89- 2.80)	1.16E-01	1.87E-02	70.0
rs2834651	Relapse	codom.het	A/G	1.23(1.09- 1.39)	9.29E-04	1.34(0.96- 1.86)	8.56E-02	2.15E-04	84.6
rs11615759	Relapse	recessive	A/G	2.08(1.46- 2.98)	5.83E-05	1.97(1.10- 3.51)	2.20E-02	9.48E-02	57.6
rs3736994	Relapse	dominant	A/G	0.86(0.77- 0.97)	1.25E-02	0.84(0.64- 1.12)	2.37E-01	1.21E-03	81.1
rs4920993	Relapse	recessive	A/G	1.52(1.24- 1.86)	4.22E-05	1.43(1.01- 2.03)	4.21E-02	4.89E-02	61.9
rs8111608	Relapse	additive	G/A	0.86(0.78- 0.93)	5.58E-04	0.85(0.72- 1.00)	4.49E-02	3.11E-02	66.2
rs7572970	Relapse	recessive	G/A	1.32(1.07- 1.63)	9.16E-03	1.16(0.67- 2.04)	5.95E-01	5.84E-04	82.8
rs962312	Relapse	dominant	G/A	1.28(1.11- 1.46)	5.65E-04	1.36(1.04- 1.77)	2.53E-02	2.44E-02	68.1
rs754799	Relapse	recessive	A/C	1.89(1.43- 2.49)	5.89E-06	1.89(1.43- 2.49)	5.89E-06	3.92E-01	0.0
rs4946483	Relapse	additive	G/A	1.28(1.15- 1.43)	6.89E-06	1.28(1.15- 1.43)	6.89E-06	5.73E-01	0.0
rs9533040	Relapse	codom.hom	A/G	2.49(1.66- 3.73)	9.79E-06	2.42(1.41- 4.17)	1.44E-03	2.29E-01	32.2
rs846978	Relapse	dominant	G/A	1.27(1.09- 1.48)	2.00E-03	1.27(0.95- 1.69)	1.11E-01	1.88E-02	69.9
rs923435	Relapse	codom.het	A/C	0.84(0.74- 0.95)	4.73E-03	0.83(0.64- 1.07)	1.47E-01	7.93E-03	74.7
rs3917265	Relapse	dominant	G/A	1.11(0.97- 1.26)	1.15E-01	1.15(0.81- 1.63)	4.29E-01	1.95E-04	84.8
rs12435167	Relapse	dominant	G/A	0.88(0.78- 0.98)	2.64E-02	0.90(0.66- 1.21)	4.82E-01	3.46E-04	83.8
rs2028008	Relapse	dominant	A/G	1.20(1.07- 1.35)	1.95E-03	1.19(0.95- 1.49)	1.37E-01	1.54E-02	71.2
rs10888205	Relapse	recessive	G/A	1.28(1.01- 1.63)	4.03E-02	1.19(0.64- 2.20)	5.91E-01	4.18E-04	83.4
rs1036332	Relapse	additive	C/A	0.87(0.79- 0.95)	2.92E-03	0.85(0.69- 1.04)	1.06E-01	5.04E-03	76.6

SNP	Model			Source	SNP location					
	Outcome	MoI	Alleles		Chr	Chr position	HUGO (ensembl)	Nearest gene (Illumina)	NM (Illumina)	Location
rs754799	Recurrence	recessive	A/C	epicuro	19	1.855.612	ADAT3;SCAMP4	SCAMP4, ADAT3	NM_079834.2	flanking_5UTR
rs12988804	Recurrence	codom.het	G/A	epicuro	2	169.826.057	LRP2	LRP2	NM_004525.1	intron
rs489770	Recurrence	recessive	A/G	mdacc	5	153.619.501	GALNT10	GALNT10	NM_198321.2	intron
rs980761	Recurrence	codom.het	A/G	epicuro	4	76.300.144	NA	DKFZP564O0823	NM_015393.2	flanking_3UTR
rs3736994	Recurrence	dominant	A/G	epicuro	14	59.012.927	C14orf149	C14orf149	NM_144581.1	intron
rs7930624	Recurrence	additive	C/A	epicuro	11	243.841	PSMD13	PSMD13	NM_175932.1	flanking_3UTR
rs12681007	Recurrence	recessive	G/A	mdacc	8	92.883.202	NA	RUNX1T1	NM_175634.1	flanking_3UTR
rs6443819	Recurrence	additive	A/G	mdacc	3	183.857.095	NA	ATP11B	NM_014616.1	flanking_5UTR
rs10867878	Recurrence	recessive	G/A	epicuro	9	84.583.137	NA	RASEF	NM_152573.2	flanking_3UTR
rs9504361	Recurrence	codom.het	G/A	epicuro	6	522.820	EXOC2	EXOC2	NM_018303.4	intron
rs6512003	Recurrence	recessive	C/A	epicuro	19	14.996.741	CCDC105	FLJ40365,CCDC105	NM_173482.1	flanking_3UTR
rs6505263	Recurrence	additive	G/A	epicuro	17	27.177.993	NA	C17orf79	NM_018405.2	flanking_3UTR
rs2943313	Recurrence	dominant	G/A	mdacc	16	25.577.363	NA	HS3ST4	NM_001012981.2	flanking_5UTR
rs2191031	Recurrence	dominant	G/A	epicuro	3	45.885.874	NA	LZTFL1	NM_031200.1	flanking_5UTR
rs1421776	Progression	codom.hom	A/G	mdacc	5	143.230.072	NA	HB-1	NM_021182.1	flanking_3UTR
rs4246835	Progression	codom.het	G/A	epicuro	9	20.023.087	NA	SLC24A2	NM_020344.1	flanking_5UTR
rs10167220	Progression	dominant	G/A	epicuro	2	53.699.993	ASB3	ASB3	NM_145863.1	flanking_3UTR
rs526509	Progression	additive	C/A	epicuro	9	258.740	DOCK8	DOCK8	NM_203447.1	flanking_5UTR
rs6100810	Progression	recessive	A/G	epicuro	20	58.243.870	NA	CDH26	NM_001004305.1	flanking_5UTR
rs12294567	Progression	codom.hom	A/G	epicuro	11	91.265.310	NA	FAT3	NM_005959.3	flanking_5UTR
rs4246835	Progression	dominant	G/A	mdacc	9	20.023.087	NA	SLC24A2	NM_020344.1	flanking_5UTR
rs17218455	Progression	codom.hom	G/A	mdacc	12	105.240.571	TCP11L2	TCP11L2	NM_152772.1	intron
rs6752816	Progression	dominant	G/A	epicuro	2	54.485.107	NA	FLJ40298	NM_173486.1	flanking_3UTR
rs2950650	Progression	codom.hom	A/G	epicuro	2	182.273.888	NA	NEUROD1	NM_002500.1	flanking_5UTR
rs9891348	Progression	codom.het	A/G	epicuro	17	50.929.373	NA	MMD	NM_012329.2	flanking_5UTR
rs5027573	Progression	dominant	G/A	mdacc	10	87.881.252	GRID1	GRID1	NM_017551.1	intron
rs17831395	Progression	dominant	A/G	mdacc	18	53.269.990	ONECUT2	ONECUT2	NM_004852.1	intron
rs1568519	Progression	dominant	A/C	mdacc	2	212.714.637	ERBB4	ERBB4	NM_005235.1	intron
rs7721273	Progression	codom.hom	A/G	epicuro	5	7.577.548	ADCY2	ADCY2	NM_020546.1	intron
rs7588481	Progression	dominant	A/C	mdacc	2	137.365.673	THSD7B	CXCR4,THSD7B	NM_003467.2	flanking_5UTR
rs6531449	Progression	additive	A/G	epicuro	4	36.178.408	NA	CENTD1	NM_015230.2	flanking_5UTR
rs7936809	Progression	dominant	G/A	epicuro	11	25.335.857	NA	LUZP2	NM_001009909.2	flanking_3UTR
rs3797725	Progression	codom.hom	A/G	epicuro	5	111.103.627	C5orf13	C5orf13	NM_004772.1	intron
rs7572970	Progression	recessive	G/A	mdacc	2	160.844.902	RBMS1	RBMS1	NM_016839.2	intron

SNP	Model			Source	SNP location					
	Outcome	MoI	Alleles		Chr	Chr position	HUGO (ensembl)	Nearest gene (Illumina)	NM (Illumina)	Location
rs10437619	Progression	additive	A/G	epicuro	11	25.232.451	NA	LUZP2	NM_001009909.2	flanking_3UTR
rs1035856	Progression	dominant	G/A	mdacc	5	77.952.104	LHFPL2	LHFPL2	NM_005779.1	flanking_5UTR
rs1847325	Progression	additive	G/A	mdacc	15	98.885.065	LASS3	LASS3	NM_178842.2	intron
rs1432374	Progression	additive	G/A	epicuro	3	117.933.067	NA	LSAMP	NM_002338.2	flanking_5UTR
rs10783528	Progression	additive	G/A	epicuro	12	51.374.751	KRT77	KRT1B,KRT77	NM_175078.1	coding
rs6757412	Progression	codom.hom	A/C	epicuro	2	236.478.306	AGAP1	CENTG2, AGAP1	NM_014914.2	intron
rs3772337	Relapse	recessive	A/G	epicuro	3	1.268.647	CNTN6	CNTN6	NM_014461.2	intron
rs11604069	Relapse	dominant	G/A	mdacc	11	34.720.827	NA	EHF	NM_012153.3	flanking_3UTR
rs446027	Relapse	codom.hom	A/G	epicuro	8	81.614.888	NA	ZBTB10	NM_023929.2	flanking_3UTR
rs2694787	Relapse	codom.hom	G/A	epicuro	10	117.852.840	GFRA1	GFRA1	NM_005264.2	intron
rs2834651	Relapse	codom.het	A/G	mdacc	21	35.150.230	RUNX1	RUNX1	NM_001001890.1	intron
rs11615759	Relapse	recessive	A/G	epicuro	12	23.634.296	SOX5	SOX5	NM_178010.1	intron
rs3736994	Relapse	dominant	A/G	epicuro	14	59.012.927	C14orf149	C14orf149	NM_144581.1	intron
rs4920993	Relapse	recessive	A/G	epicuro	5	117.180.846	NA	DTWD2	NM_173666.1	flanking_3UTR
rs8111608	Relapse	additive	G/A	epicuro	19	15.004.453	NA	FLJ40365	NM_173482.1	flanking_3UTR
rs7572970	Relapse	recessive	G/A	epicuro	2	160.844.902	RBMS1	RBMS1	NM_016839.2	intron
rs962312	Relapse	dominant	G/A	epicuro	6	120.953.430	NA	COX6A1P3	NM_152730.3	flanking_3UTR
rs754799	Relapse	recessive	A/C	epicuro	19	1.855.612	ADAT3;SCAMP4	SCAMP4, ADAT3	NM_079834.2	flanking_5UTR
rs4946483	Relapse	additive	G/A	epicuro	6	120.792.408	NA	COX6A1P3	NM_152730.3	flanking_3UTR
rs9533040	Relapse	codom.hom	A/G	mdacc	13	41.694.721	DGKH	DGKH	NM_178009.2	intron
rs846978	Relapse	dominant	G/A	epicuro	6	108.149.027	SCML4	SCML4	NM_198081.1	intron
rs923435	Relapse	codom.het	A/C	epicuro	12	127.710.640	TMEM132C	SLC15A4, TMEM132C	NM_145648.1	flanking_3UTR
rs3917265	Relapse	dominant	G/A	epicuro	2	102.144.893	IL1R1	IL1R1	NM_000877.2	intron
rs12435167	Relapse	dominant	G/A	epicuro	14	40.206.466	NA	LRFN5	NM_152447.2	flanking_5UTR
rs2028008	Relapse	dominant	A/G	epicuro	7	88.949.497	NA	FLJ32110	NM_181646.2	flanking_3UTR
rs10888205	Relapse	recessive	G/A	epicuro	10	48.186.564	NA	GDF10	NM_004962.2	flanking_5UTR
rs1036332	Relapse	additive	C/A	mdacc	1	197.279.101	NA	PTPRC	NM_002838.2	flanking_3UTR

MOI - mode of inheritance
MAF SBC/EPICURO-minor allele frequency in the SBC/EPICURO Study; MAF MDACC - minor allele frequency in the MDACC Study
HR-hazard ratio, 95% CI - 95% confidence interval

Supplementary Table 5. C-statistics for (A) multivariate Cox regression for each NMIBC outcome including all co-variables except the SNP, (B) multivariate Cox regression including all co-variables and each SNP at a time, and (C) multivariate Cox regression including all co-variables and all significant SNP for each outcome. C-statistics were estimated for each original series and after bootstrapping

Study	SNP ID	Outcome	MoI	TXBC-1 C-stat original	TXBC-1 C-stat boots	SBC/EPICURO C-stat original	SBC/EPICURO C-stat boots	TXBC-2 C-stat original	TXBC-2 C-stat boots	Int'al Cohorts C-stat original	Int'al Cohorts C-stat boots
A											
Multivariate Cox without SNP information		**relapse**		0.63	0.61	0.62	0.60	0.71	0.69	0.67	0.66
		progression		0.70	0.66	0.77	0.73	0.81	0.78	0.80	0.78
		recurrence		0.64	0.62	0.64	0.61	0.69	0.66	0.68	0.67
B											
Multivariate Cox for each SNP	**rs11615759**	**relapse**	**recessive**	0.63	0.61	0.63	0.60	NA	NA	0.66	0.65
	rs754799	**relapse**	**recessive**	0.63	0.61	0.62	0.60	0.71	0.68	0.67	0.66
	rs4946483	**relapse**	**additive**	0.63	0.61	0.78	NA	0.72	0.69	0.73	NA
	rs4246835	**progression**	**codom.het**	0.72	0.67	0.80	0.76	0.81	0.78	0.80	0.77
	rs6100810	**progression**	**recessive**	0.71	0.67	0.78	0.74	0.81	0.77	0.79	0.77
	rs7572970	**progression**	**recessive**	0.72	0.68	0.78	0.73	0.81	0.77	0.79	0.77
	rs12294567	**progression**	**codom.hom**	0.71	0.66	0.79	0.74	0.81	0.77	0.80	0.77
	rs1035856	**progression**	**dominant**			0.78	0.73			0.80	0.77
	rs17218455	**progression**	**codom.hom**	0.72	0.67	0.79	0.74	0.81	0.77	0.79	0.77
	rs3797725	**progression**	**codom.hom**	0.71	0.66	0.78	0.73	0.81	0.77	0.80	0.78
	rs754799	**recurrence**	**recessive**	0.65	0.62	0.64	0.61	0.69	0.66	0.68	0.67
C											
Multivariate Cox joining SNPs		**relapse**		0.64	0.62	0.63	0.61	0.72	0.69	0.67	0.66
		progression		0.75	0.70	0.84	0.78	0.82	0.76	0.79	0.75
		recurrence		0.65	0.62	0.64	0.61	0.69	0.66	0.68	0.67

TXBC-1/2: Discovery/Validation phase subset of the Texas Bladder Cancer Study; SBC/EPICURO: Spanish Bladder Cancer/EPICURO Study
C-stat original: C-statistic obtained from the original model; C-stat boots: C-statistic obtained after 500 rounds of bootstrapping

Supplementary Table 6. Correlation between the significant SNPs identified in NMIBC series after Validation phase and the tumor baseline characteristics: stage, grade and tumor size.

SNP	SBC/EPICURO, Discovery									International series, Validation								
	Stage			Grade			Tumor size			Stage			Grade			Tumor size		
	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*
rs754799	835	-0.03	0.31	835	-0.03	0.46	600	-0.03	0.42	835	-0.03	0.31	835	-0.03	0.46	600	-0.03	0.42
rs4246835	836	-0.05	0.12	836	-0.05	0.16	600	-0.01	0.73	836	-0.05	0.12	836	-0.05	0.16	600	-0.01	0.73
rs7572970	835	-0.02	0.65	835	0.02	0.58	600	0.06	0.13	835	-0.02	0.65	835	0.02	0.58	600	0.06	0.13
rs3797725	836	-0.02	0.63	836	-0.01	0.72	600	-0.05	0.21	836	-0.02	0.63	836	-0.01	0.72	600	-0.05	0.21
rs4946483	836	-0.02	0.49	836	-0.04	0.20	600	0.07	0.10	836	-0.02	0.49	836	-0.04	0.20	600	0.07	0.10
rs12294567	836	0.02	0.56	836	0.04	0.30	600	-0.05	0.26	836	0.02	0.56	836	0.04	0.30	600	-0.05	0.26
rs11615759	836	-0.04	0.24	836	-0.02	0.63	600	-0.01	0.88	836	-0.04	0.24	836	-0.02	0.63	600	-0.01	0.88
rs17218455	836	-0.02	0.65	836	-0.02	0.63	600	-0.03	0.48	836	-0.02	0.65	836	-0.02	0.63	600	-0.03	0.48
rs6100810	836	0.02	0.62	836	0.07	0.06	600	0.00	0.95	836	0.02	0.62	836	0.07	0.06	600	0.00	0.95

SNP	TXBC-1, Discovery									TXBC-2, Validation								
	Stage			Grade			Tumor size			Stage			Grade			Tumor size		
	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*
rs754799	494	0.04	0.33	477	0.07	0.12	212	0.09	0.20	358	0.05	0.37	349	0.07	0.18	185	-0.02	0.84
rs4246835	493	0.04	0.33	476	0.05	0.32	211	0.07	0.31	345	-0.04	0.42	337	0.02	0.72	180	-0.06	0.39
rs7572970	494	0.08	0.07	477	-0.02	0.73	212	0.08	0.25	356	-0.10	0.05	347	0.03	0.63	183	-0.09	0.23
rs3797725	494	0.03	0.47	477	0.03	0.58	212	-0.01	0.86	352	0.03	0.51	344	0.14	0.01	182	-0.04	0.61
rs4946483	494	-0.10	0.03	477	-0.06	0.18	212	0.09	0.21	357	-0.04	0.42	349	0.04	0.40	185	-0.07	0.37
rs12294567	494	0.01	0.85	477	-0.07	0.10	212	-0.02	0.82	359	-0.01	0.88	350	-0.03	0.56	184	0.16	0.03
rs11615759	494	0.00	0.98	477	0.01	0.85	212	-0.08	0.22	356	-0.07	0.19	348	-0.01	0.90	184	-0.09	0.24
rs17218455	494	0.04	0.43	477	0.10	0.02	212	-0.08	0.27	358	0.01	0.93	349	0.00	0.99	184	0.03	0.67
rs6100810	493	0.08	0.09	476	0.02	0.66	212	-0.17	0.01	355	-0.07	0.18	346	-0.09	0.09	184	-0.01	0.95

Supplementary Table 7. MIBC associations of SNPs identified in Stage I analysis with risk of BC-specific mortality, overall survival, and progression: Meta-analysis for the Discovery phase. Chromosome and gene location, minor allele, and mode-of inheritance are also displayed.

SNP	Model			Discovery Meta-analysis results							
	Outcome	MoI	Alleles	MAF SBC/EPICURO	MAF TXBC-1	HR(95% CI) fixed	P fixed	HR(95% CI) random	P random	P het	I^2
rs1015267	BC-specific mort.	codom.hom	G/A	0.35	0.30	3.96(2.51-6.23)	2.91E-09	3.96(2.51-6.23)	2.91E-09	9.60E-01	0.00
rs1015267	BC-specific mort.	recessive	G/A	0.35	0.30	3.41(2.24-5.19)	1.01E-08	3.41(2.24-5.19)	1.01E-08	7.32E-01	0.00
rs1008954	BC-specific mort.	dominant	A/G	0.07	0.05	2.75(1.88-4.02)	1.65E-07	2.75(1.88-4.02)	1.65E-07	9.56E-01	0.00
rs11221970	BC-specific mort.	dominant	A/G	0.07	0.05	2.77(1.88-4.08)	2.28E-07	2.77(1.88-4.08)	2.28E-07	9.94E-01	0.00
rs17603887	BC-specific mort.	dominant	G/A	0.09	0.10	2.50(1.76-3.54)	2.85E-07	2.50(1.74-3.58)	6.39E-07	3.02E-01	0.06
rs1537010	BC-specific mort.	dominant	A/G	0.04	0.02	3.86(2.30-6.49)	3.15E-07	3.86(2.30-6.49)	3.15E-07	7.21E-01	0.00
rs7035632	BC-specific mort.	dominant	C/A	0.07	0.07	2.51(1.73-3.63)	1.08E-06	2.51(1.73-3.63)	1.08E-06	5.26E-01	0.00
rs2646727	BC-specific mort.	additive	A/G	0.40	0.43	1.70(1.37-2.11)	1.18E-06	1.70(1.37-2.11)	1.18E-06	4.01E-01	0.00
rs404678	BC-specific mort.	recessive	A/G	0.40	0.37	0.29(0.18-0.48)	1.57E-06	0.29(0.18-0.48)	1.57E-06	4.83E-01	0.00
rs2565721	BC-specific mort.	additive	G/A	0.46	0.48	0.60(0.49-0.74)	2.62E-06	0.60(0.48-0.75)	7.56E-06	2.95E-01	0.09
rs9323978	BC-specific mort.	dominant	C/A	0.20	0.20	2.06(1.52-2.78)	2.63E-06	2.06(1.52-2.78)	2.63E-06	5.31E-01	0.00
rs3102192	BC-specific mort.	additive	G/A	0.41	0.35	0.60(0.48-0.74)	4.59E-06	0.60(0.48-0.74)	4.59E-06	5.67E-01	0.00
rs4923350	BC-specific mort.	recessive	A/G	0.41	0.37	2.22(1.57-3.14)	6.17E-06	2.22(1.57-3.14)	6.17E-06	3.90E-01	0.00
rs725745	BC-specific mort.	codom.het	G/A	0.39	0.36	0.49(0.35-0.67)	7.44E-06	0.49(0.35-0.67)	7.44E-06	3.79E-01	0.00
rs6074012	BC-specific mort.	recessive	G/A	0.45	0.49	2.06(1.49-2.83)	9.78E-06	2.05(1.40-3.01)	2.25E-04	2.33E-01	0.30
rs4900384	BC-specific mort.	dominant	A/G	0.29	0.30	1.96(1.45-2.65)	1.27E-05	1.96(1.38-2.79)	1.69E-04	2.45E-01	0.26
rs2416996	BC-specific mort.	recessive	G/A	0.40	0.43	2.18(1.53-3.11)	1.54E-05	2.18(1.53-3.11)	1.54E-05	3.93E-01	0.00
rs6672666	BC-specific mort.	recessive	A/G	0.26	0.26	2.80(1.75-4.49)	1.71E-05	2.78(1.63-4.74)	1.67E-04	2.58E-01	0.22
rs783145	BC-specific mort.	additive	A/G	0.45	0.49	0.63(0.51-0.78)	2.32E-05	0.63(0.51-0.78)	2.32E-05	5.92E-01	0.00
rs1171509	BC-specific mort.	codom.hom	A/G	0.30	0.33	2.85(1.75-4.63)	2.45E-05	2.85(1.75-4.63)	2.45E-05	3.78E-01	0.00
rs6805542	BC-specific mort.	dominant	A/G	0.28	0.22	1.86(1.39-2.49)	3.23E-05	1.86(1.38-2.50)	3.76E-05	3.13E-01	0.02
rs4871475	BC-specific mort.	recessive	C/A	0.25	0.22	2.92(1.75-4.86)	3.90E-05	2.92(1.75-4.86)	3.90E-05	3.67E-01	0.00
rs335305	BC-specific mort.	additive	A/G	0.48	0.49	1.54(1.25-1.90)	4.45E-05	1.54(1.25-1.90)	4.45E-05	3.24E-01	0.00
rs2139142	BC-specific mort.	dominant	G/A	0.21	0.22	1.88(1.39-2.56)	5.23E-05	1.88(1.39-2.56)	5.23E-05	3.62E-01	0.00
rs10437447	Overall survival	dominant	G/A	0.15	0.13	1.88(1.47-2.41)	4.59E-07	1.88(1.47-2.41)	4.59E-07	5.89E-01	0.00
rs2565721	Overall survival	additive	G/A	0.46	0.48	0.66(0.56-0.78)	9.60E-07	0.66(0.56-0.78)	9.60E-07	5.98E-01	0.00
rs783145	Overall survival	additive	A/G	0.45	0.49	0.66(0.56-0.79)	1.63E-06	0.66(0.56-0.79)	1.63E-06	6.93E-01	0.00
rs2646727	Overall survival	additive	A/G	0.40	0.43	1.52(1.28-1.81)	1.92E-06	1.53(1.26-1.85)	1.25E-05	2.75E-01	0.16
rs9915388	Overall survival	dominant	G/A	0.10	0.10	1.97(1.48-2.61)	2.90E-06	1.97(1.48-2.61)	2.90E-06	6.36E-01	0.00

rs955387	Overall survival	dominant	G/A	0.14	0.11	1.86(1.43-2.41)	3.07E-06	1.86(1.43-2.41)	3.07E-06	3.26E-01	0.00
rs953517	Overall survival	additive	A/G	0.31	0.36	0.67(0.56-0.80)	6.85E-06	0.67(0.56-0.80)	6.85E-06	3.23E-01	0.00
rs4795064	Overall survival	dominant	A/G	0.42	0.43	0.57(0.45-0.73)	8.14E-06	0.57(0.45-0.73)	8.14E-06	3.62E-01	0.00
rs1621801	Overall survival	additive	C/A	0.45	0.48	0.68(0.58-0.81)	8.91E-06	0.68(0.58-0.81)	8.91E-06	6.72E-01	0.00
rs7793188	Overall survival	recessive	G/A	0.34	0.30	0.41(0.27-0.61)	1.00E-05	0.41(0.27-0.61)	1.00E-05	9.44E-01	0.00
rs7594456	Overall survival	codom.het	G/A	0.31	0.34	0.58(0.45-0.74)	1.58E-05	0.58(0.45-0.74)	1.58E-05	3.58E-01	0.00
rs2173281	Overall survival	additive	G/A	0.31	0.30	1.50(1.25-1.80)	1.60E-05	1.50(1.25-1.80)	1.60E-05	5.21E-01	0.00
rs6538017	Overall survival	dominant	G/A	0.23	0.22	1.67(1.32-2.12)	1.76E-05	1.67(1.32-2.12)	1.76E-05	4.74E-01	0.00
rs2989509	Overall survival	additive	G/A	0.31	0.29	1.41(1.20-1.66)	2.40E-05	1.41(1.20-1.66)	2.40E-05	3.83E-01	0.00
rs7103589	Overall survival	recessive	A/C	0.46	0.39	1.83(1.38-2.42)	2.43E-05	1.83(1.38-2.42)	2.43E-05	5.24E-01	0.00
rs2065411	Overall survival	recessive	A/C	0.37	0.35	2.01(1.45-2.78)	2.44E-05	2.01(1.45-2.78)	2.44E-05	4.80E-01	0.00
rs9471770	Overall survival	dominant	A/C	0.13	0.14	0.55(0.42-0.73)	2.87E-05	0.55(0.40-0.75)	1.67E-04	2.66E-01	0.19
rs6935921	Overall survival	additive	A/G	0.29	0.31	1.45(1.22-1.74)	4.24E-05	1.45(1.22-1.74)	4.24E-05	4.14E-01	0.00
rs11234582	Overall survival	additive	A/G	0.35	0.36	1.42(1.19-1.69)	9.69E-05	1.42(1.17-1.72)	3.19E-04	2.78E-01	0.15
rs16927851	Progression	recessive	A/G	0.28	0.23	3.48(2.25-5.39)	2.08E-08	3.53(2.11-5.90)	1.62E-06	2.41E-01	0.27
rs9849682	Progression	recessive	A/C	0.45	0.44	2.33(1.67-3.25)	7.08E-07	2.33(1.67-3.25)	7.08E-07	6.43E-01	0.00
rs11732628	Progression	recessive	A/G	0.37	0.39	2.60(1.78-3.80)	8.41E-07	2.60(1.78-3.80)	8.41E-07	4.63E-01	0.00
rs9668920	Progression	recessive	A/G	0.30	0.27	2.78(1.83-4.23)	1.73E-06	2.80(1.77-4.42)	1.02E-05	2.78E-01	0.15
rs16927851	Progression	additive	A/G	0.28	0.23	1.69(1.35-2.10)	3.21E-06	1.69(1.35-2.10)	3.21E-06	4.00E-01	0.00
rs4658680	Progression	recessive	G/A	0.31	0.29	2.58(1.73-3.86)	3.55E-06	2.58(1.73-3.86)	3.55E-06	6.06E-01	0.00
rs6774177	Progression	dominant	G/A	0.14	0.14	2.04(1.51-2.77)	3.88E-06	2.04(1.51-2.77)	3.88E-06	3.39E-01	0.00
rs4595635	Progression	additive	G/A	0.30	0.26	1.61(1.31-1.98)	5.55E-06	1.61(1.31-1.98)	5.55E-06	7.45E-01	0.00
rs2837472	Progression	recessive	G/A	0.30	0.27	3.14(1.90-5.20)	7.96E-06	3.14(1.90-5.20)	7.96E-06	6.00E-01	0.00
rs330579	Progression	dominant	G/A	0.39	0.37	0.51(0.38-0.69)	1.03E-05	0.51(0.38-0.69)	1.03E-05	5.76E-01	0.00
rs1263674	Progression	additive	A/G	0.33	0.31	1.60(1.30-1.98)	1.08E-05	1.61(1.28-2.01)	3.22E-05	2.90E-01	0.11
rs2052665	Progression	dominant	A/G	0.23	0.18	1.97(1.45-2.68)	1.39E-05	1.97(1.45-2.68)	1.39E-05	4.77E-01	0.00
rs899333	Progression	additive	A/G	0.50	0.48	0.63(0.51-0.78)	1.85E-05	0.63(0.51-0.78)	1.85E-05	4.18E-01	0.00
rs7973149	Progression	additive	G/A	0.33	0.30	1.55(1.26-1.90)	2.50E-05	1.55(1.26-1.90)	2.50E-05	6.62E-01	0.00
rs11624081	Progression	dominant	G/A	0.10	0.12	2.04(1.46-2.85)	2.73E-05	2.04(1.46-2.85)	2.73E-05	3.95E-01	0.00
rs4658683	Progression	recessive	A/G	0.26	0.21	2.78(1.71-4.52)	3.61E-05	2.78(1.71-4.52)	3.61E-05	4.25E-01	0.00
rs162713	Progression	codom.hom	G/A	0.40	0.36	2.43(1.59-3.70)	3.68E-05	2.43(1.59-3.70)	3.68E-05	5.24E-01	0.00
rs1656197	Progression	additive	C/A	0.49	0.46	0.65(0.53-0.80)	6.31E-05	0.65(0.53-0.80)	6.31E-05	8.15E-01	0.00
rs12032833	Progression	additive	A/G	0.40	0.45	1.56(1.25-1.55)	6.44E-05	1.56(1.25-1.94)	6.44E-05	8.37E-01	0.00

SNP	Model			Source	SNP location					
	Outcome	MoI	Alleles		Chr	Chr position	HUGO (ensemble)	Nearest gene (Illumina)	NM (Illumina)	Location
rs1015267	BC-specific mort.	codom.hom	G/A	epicuro	11	26317792	ANO3	TMEM16C	NM_031418.1	intron
rs1015267	BC-specific mort.	recessive	G/A	mdacc	11	26317792	ANO3	TMEM16C	NM_031418.1	intron
rs1008954	BC-specific mort.	dominant	A/G	mdacc	11	129479865	APLP2	APLP2	NM_001642.1	intron
rs11221970	BC-specific mort.	dominant	A/G	epicuro	11	129488492	APLP2,U4	APLP2	NM_001642.1	intron
rs17603887	BC-specific mort.	dominant	G/A	epicuro	10	125213543	NA	GPR26	NM_153442.1	flanking_5UTR
rs1537010	BC-specific mort.	dominant	A/G	epicuro	9	128033291	NA	C9orf28	NM_033446.1	flanking_5UTR
rs7035632	BC-specific mort.	dominant	C/A	epicuro	9	115447048	NA	RGS3	NM_021106.3	flanking_3UTR
rs2646727	BC-specific mort.	additive	A/G	epicuro	11	85716734	C11orf73,AP001148.4	HSPC138	NM_016401.2	intron
rs404678	BC-specific mort.	recessive	A/G	epicuro	20	4972237	NA	SLC23A2	NM_203327.1	flanking_5UTR
rs2565721	BC-specific mort.	additive	G/A	mdacc	6	161155088	NA	PLG	NM_000301.1	flanking_3UTR
rs9323978	BC-specific mort.	dominant	C/A	epicuro	14	97550986	NA	FLJ25773	NM_182560.1	flanking_5UTR
rs3102192	BC-specific mort.	additive	G/A	mdacc	13	60685194	NA	PCDH20	NM_022843.2	flanking_3UTR
rs4923350	BC-specific mort.	recessive	A/G	epicuro	11	26322855	ANO3	TMEM16C	NM_031418.1	intron
rs725745	BC-specific mort.	codom.het	G/A	epicuro	21	22763548	NA	C21orf74	XR_001010.1	flanking_3UTR
rs6074012	BC-specific mort.	recessive	G/A	epicuro	20	44133573	NCOA5	NCOA5	NM_020967.2	intron
rs4900384	BC-specific mort.	dominant	A/G	epicuro	14	97568704	NA	FLJ25773	NM_182560.1	flanking_5UTR
rs2416996	BC-specific mort.	recessive	G/A	epicuro	9	127883987	NA	PBX3	NM_006195.4	flanking_3UTR
rs6672666	BC-specific mort.	recessive	A/G	mdacc	1	236499507	RP11-136B18.1	ZP4	NM_021186.2	flanking_5UTR
rs783145	BC-specific mort.	additive	A/G	mdacc	6	161072439	PLG	PLG	NM_000301.1	intron
rs1171509	BC-specific mort.	codom.hom	A/G	mdacc	10	23000205	PIP5K2A	PIP5K2A	NM_005028.3	intron
rs6805542	BC-specific mort.	dominant	A/G	mdacc	3	188644930	NA	RTP4	NM_022147.2	flanking_3UTR
rs4871475	BC-specific mort.	recessive	C/A	epicuro	8	125334406	NA	TMEM65	NM_194291.1	flanking_3UTR
rs335305	BC-specific mort.	additive	A/G	epicuro	4	62076355	LPHN3	LPHN3	NM_015236.3	intron
rs2139142	BC-specific mort.	dominant	G/A	mdacc	2	127848232	MAP3K2	MAP3K2	NM_006609.2	flanking_5UTR
rs10437447	Overall survival	dominant	G/A	epicuro	10	129522059	NA	TMEM12	NM_152311.1	flanking_3UTR
rs2565721	Overall survival	additive	G/A	mdacc	6	161155088	NA	PLG	NM_000301.1	flanking_3UTR
rs783145	Overall survival	additive	A/G	mdacc	6	161072439	PLG	PLG	NM_000301.1	intron
rs2646727	Overall survival	additive	A/G	epicuro	11	85716734	C11orf73,AP001148.4	HSPC138	NM_016401.2	intron
rs9915388	Overall survival	dominant	G/A	mdacc	17	644926	NXN,RNMTL1	RNMTL1	NM_018146.2	flanking_3UTR
rs955387	Overall survival	dominant	G/A	mdacc	9	117031013	DEC1	DEC1	NM_017418.1	intron
rs953517	Overall survival	additive	A/G	mdacc	1	36682864	C1orf102	C1orf102	NM_145047.3	intron
rs4795064	Overall survival	dominant	A/G	mdacc	17	30527454	UNC45B	UNC45B	NM_001033576.1	intron

rs1621801	Overall survival	additive	C/A	mdacc	6	161116956	NA	PLG	NM_000301.1	flanking_3UTR
rs7793188	Overall survival	recessive	G/A	epicuro	7	96334989	NA	DLX5	NM_005221.4	flanking_3UTR
rs7594456	Overall survival	codom.het	G/A	epicuro	2	212111679	ERBB4,5S_rRNA	ERBB4	NM_005235.1	intron
rs2173281	Overall survival	additive	G/A	epicuro	11	85618223	NA	EED	NM_152991.1	flanking_5UTR
rs6538017	Overall survival	dominant	G/A	mdacc	9	137372128	C9orf62	KIAA0649	NM_014811.3	flanking_5UTR
rs2989509	Overall survival	additive	G/A	mdacc	9	117041491	DEC1	DEC1	NM_017418.1	intron
rs7103589	Overall survival	recessive	A/C	epicuro	11	11558113	GALNTL4	GALNTL4	NM_198516.1	intron
rs2065411	Overall survival	recessive	A/C	epicuro	20	58051064	NA	FLJ33860	NM_173644.1	flanking_5UTR
rs9471770	Overall survival	dominant	A/C	mdacc	6	42189620	AL512274.9	TBN	NM_138572.1	flanking_3UTR
rs6935921	Overall survival	additive	A/G	mdacc	6	161028526	NA	PLG	NM_000301.1	flanking_5UTR
rs11234582	Overall survival	additive	A/G	epicuro	11	85605600	NA	EED	NM_152991.1	flanking_5UTR
rs16927851	Progression	recessive	A/G	mdacc	12	24747765	SRP_euk_arch	BCAT1	NM_005504.4	flanking_3UTR
rs9849682	Progression	recessive	A/C	epicuro	3	156572679	PLCH1	PLCL3	NM_014996.1	flanking_3UTR
rs11732628	Progression	recessive	A/G	epicuro	4	120228595	NA	SYNPO2	NM_133477.1	flanking_3UTR
rs9668920	Progression	recessive	A/G	mdacc	12	24744641	NA	BCAT1	NM_005504.4	flanking_3UTR
rs16927851	Progression	additive	A/G	epicuro	12	24747765	SRP_euk_arch	BCAT1	NM_005504.4	flanking_3UTR
rs4658680	Progression	recessive	G/A	epicuro	1	243208879	EFCAB2	EFCAB2	NM_032328.1	intron
rs6774177	Progression	dominant	G/A	mdacc	3	176598901	NAALADL2,U6	NAALADL2	NM_207015.1	intron
rs4595635	Progression	additive	G/A	epicuro	12	24731875	NA	FLJ32894	NM_144667.1	flanking_5UTR
rs2837472	Progression	recessive	G/A	epicuro	21	40456382	DSCAM	DSCAM	NM_001389.3	intron
rs330579	Progression	dominant	G/A	epicuro	2	150633855	NA	FLJ32955	NM_153041.1	flanking_5UTR
rs1263674	Progression	additive	A/G	mdacc	2	207763968	NA	KLF7	NM_003709.1	flanking_5UTR
rs2052665	Progression	dominant	A/G	epicuro	18	68976818	NA	NETO1	NM_138966.2	flanking_5UTR
rs899333	Progression	additive	A/G	epicuro	17	21012955	DHRS7B	DHRS7B	NM_015510.3	intron
rs7973149	Progression	additive	G/A	epicuro	12	24730201	NA	FLJ32894	NM_144667.1	flanking_5UTR
rs11624081	Progression	dominant	G/A	epicuro	14	94651714	DICER1	DICER1	NM_030621.2	intron
rs4658683	Progression	recessive	A/G	epicuro	1	243223134	EFCAB2	EFCAB2	NM_032328.1	intron
rs162713	Progression	codom.hom	G/A	epicuro	3	7841171	NA	GRM7	NM_000844.2	flanking_3UTR
rs1656197	Progression	additive	C/A	epicuro	4	36920530	AC022463.5	FLJ11017	NM_018302.1	flanking_5UTR
rs12032833	Progression	additive	A/G	epicuro	1	76270941	NA	ST6GALNAC3	NM_152996.1	flanking_5UTR

MOI - mode of inheritance
MAF SBC/EPICURO-minor allele frequency in the SBC/EPICURO Study; MAF MDACC - minor allele frequency in the MDACC Study
HR-hazard ratio, 95% CI - 95% confidence interval

Supplementary Table 8. C-statistics for (A) multivariate Cox regression for each MIBC outcome including all co-variables except the SNP, (B) multivariate Cox regression including all co-variables and each SNP at a time, and (C) multivariate Cox regression including all co-variables and all significant SNP for each outcome. C-statistics were estimated for each original series and after bootstrapping

Study	SNP ID	Outcome	MoI	TXBC-1 C-stat original	TXBC-1 C-stat boots	SBC/EPICURO C-stat original	SBC/EPICURO C-stat boots
A							
Multivariate Cox without SNP information		**BC-specific mortality**		0.73	0.70	0.76	0.73
		Progression		0.67	0.61	0.69	0.67
		Overall survival		0.71	0.69	0.73	0.71
B							
Multivariate Cox for each SNP	**rs1015267**	**BC-specific mortality**	**codom.hom**	0.74	0.71	0.75	0.72
	rs1015267	**BC-specific mortality**	**recessive**	0.74	0.71	0.75	0.72
	rs1008954	**BC-specific mortality**	**dominant**	0.74	0.71	0.74	0.72
	rs11221970	**BC-specific mortality**	**dominant**	0.74	0.71	0.74	0.72
	rs17603887	**BC-specific mortality**	**dominant**	0.73	0.70	0.74	0.72
	rs1537010	**BC-specific mortality**	**dominant**	0.73	0.70	0.74	0.72
	rs16927851	**Progression**	**recessive**	0.68	0.63	0.73	0.70
	rs9849682	**Progression**	**recessive**	0.68	0.63	0.73	0.70
	rs11732628	**Progression**	**recessive**	0.68	0.63	0.73	0.70
	rs10437447	**Overall survival**	**dominant**	0.72	0.70	0.77	0.74
	rs2565721	**Overall survival**	**additive**	0.73	0.71	0.77	0.73
C							
Multivariate Cox joining SNPs		**BC-related survival**		0.76	0.73	0.77	0.73
		Progression		0.71	0.66	0.76	0.73
		Overall survival		0.74	0.72	0.76	0.74

TXBC-1: Discovery phase subset of the Texas Bladder Cancer Study; SBC/EPICURO: Spanish Bladder Cancer/EPICURO Study
C-stat original: C-statistic obtained from the original model; C-stat boots: C-statistic obtained after 500 rounds of bootstrapping

Supplementary Table 9. Correlation between the significant SNPs identified in MIBC series after Discovery phase and the tumor baseline characteristics: stage (T), affected ganglia (N) and metastasis (M).

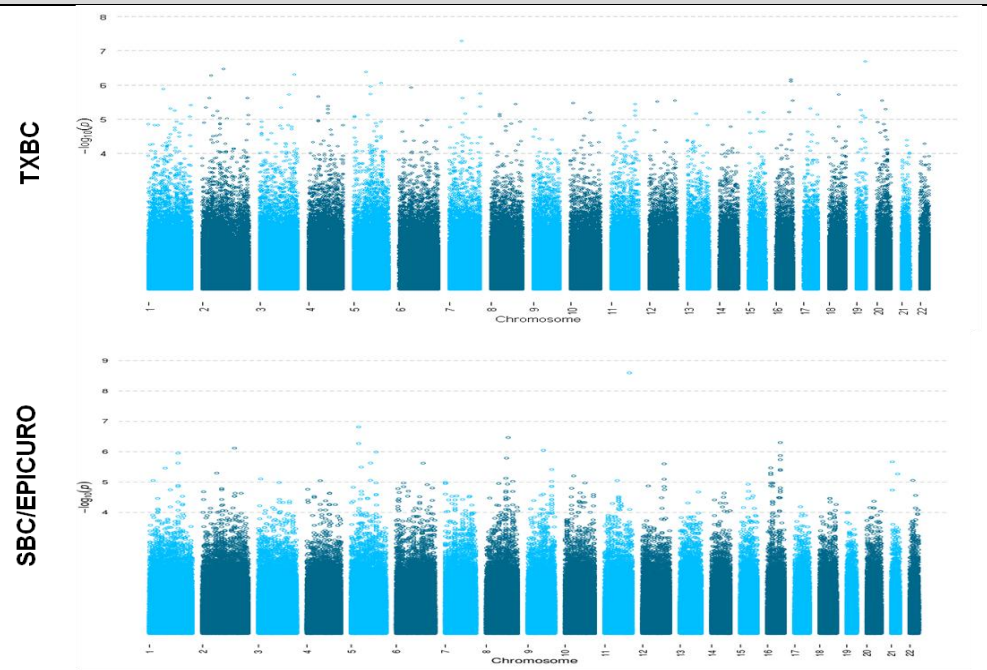
SNP	SBC/EPICURO, Discovery								
	T			N			M		
	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*
rs9849682	235	0.06	0.39	235	-0.12	0.07	235	-0.09	0.15
rs11732628	235	0.00	0.99	235	0.07	0.30	235	-0.08	0.23
rs2565721	235	0.08	0.22	235	-0.05	0.46	235	-0.06	0.35
rs1537010	234	0.12	0.07	234	-0.03	0.68	234	-0.03	0.63
rs17603887	235	-0.01	0.89	235	-0.16	0.01	235	-0.03	0.64
rs10437447	234	-0.07	0.26	234	-0.06	0.39	234	-0.04	0.49
rs1015267	234	0.08	0.22	234	-0.05	0.43	234	0.09	0.19
rs1008954	235	0.06	0.40	235	0.09	0.18	235	0.06	0.38
rs11221970	235	0.06	0.36	235	0.11	0.11	235	0.04	0.56

SNP	TXBC-1, Discovery								
	T			N			M		
	N	rho	*p-value*	N	rho	*p-value*	N	rho	*p-value*
rs9849682	393	-0.04	0.47	387	0.09	0.07	390	0.03	0.61
rs11732628	393	0.00	0.97	387	-0.04	0.48	390	0.02	0.74
rs2565721	393	-0.01	0.88	387	-0.08	0.12	390	0.00	0.99
rs1537010	393	0.04	0.38	387	0.08	0.10	390	0.06	0.21
rs17603887	393	0.03	0.54	387	0.02	0.67	390	0.08	0.10
rs10437447	393	0.00	0.94	387	0.05	0.33	390	0.01	0.80
rs1015267	393	-0.04	0.46	387	0.04	0.43	390	0.02	0.74
rs1008954	393	0.10	0.05	387	0.06	0.27	390	0.04	0.49
rs11221970	393	0.08	0.09	387	0.06	0.24	390	0.04	0.46

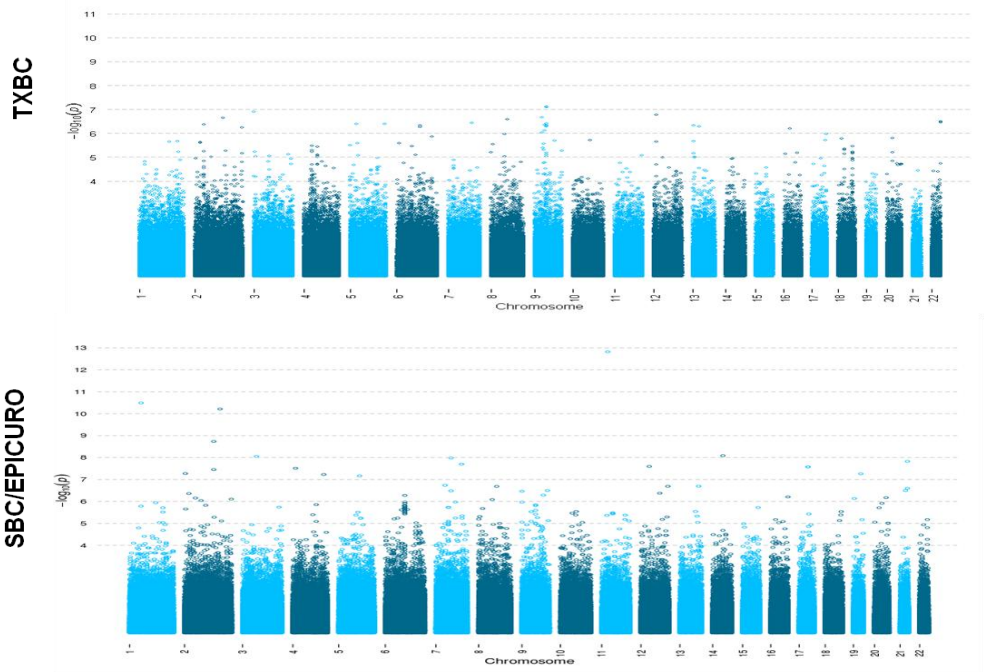
Supplementary figures

Supplementary Figure 1. Scatter (Manhattan) plot of chromosome position (x axis) against -$\log_{10}$ (*p-value*) for the 3 outcomes of interest (**a**: recurrence, **b**: progression, **c**: relapse) in the two Discovery studies

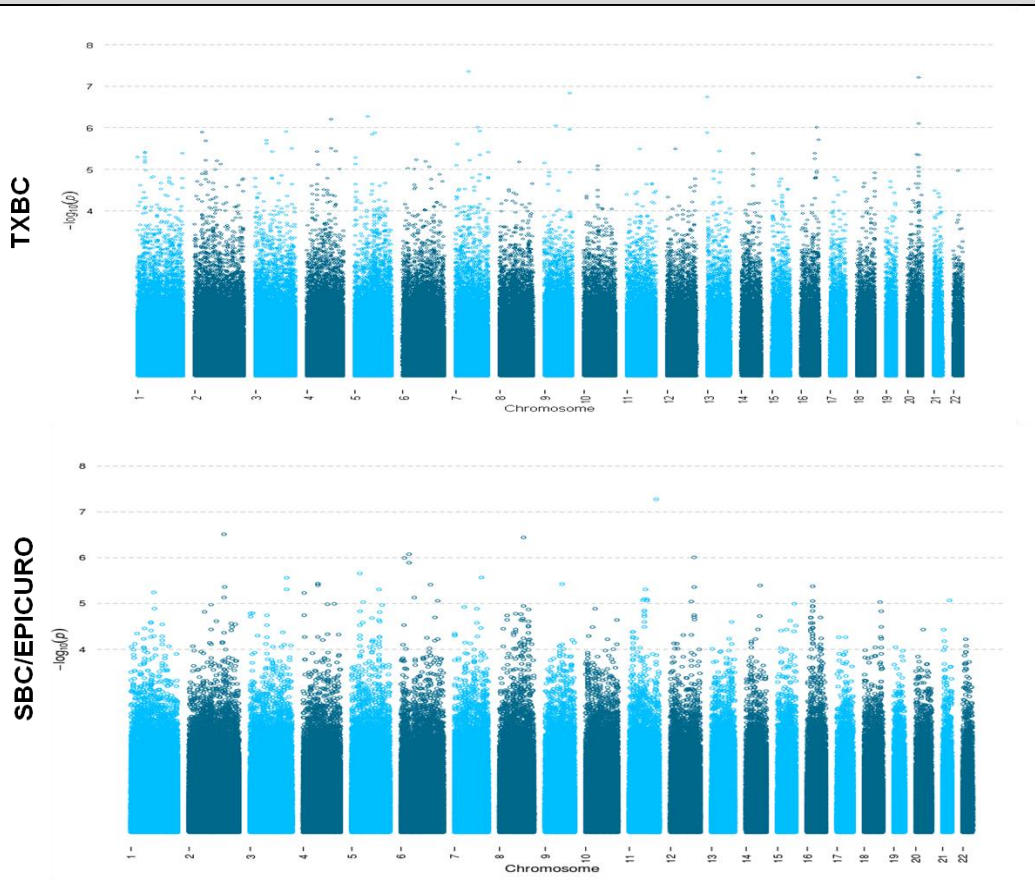
Supplementary Figure 1a. Recurrence.



Supplementary Figure 1b. Progression.

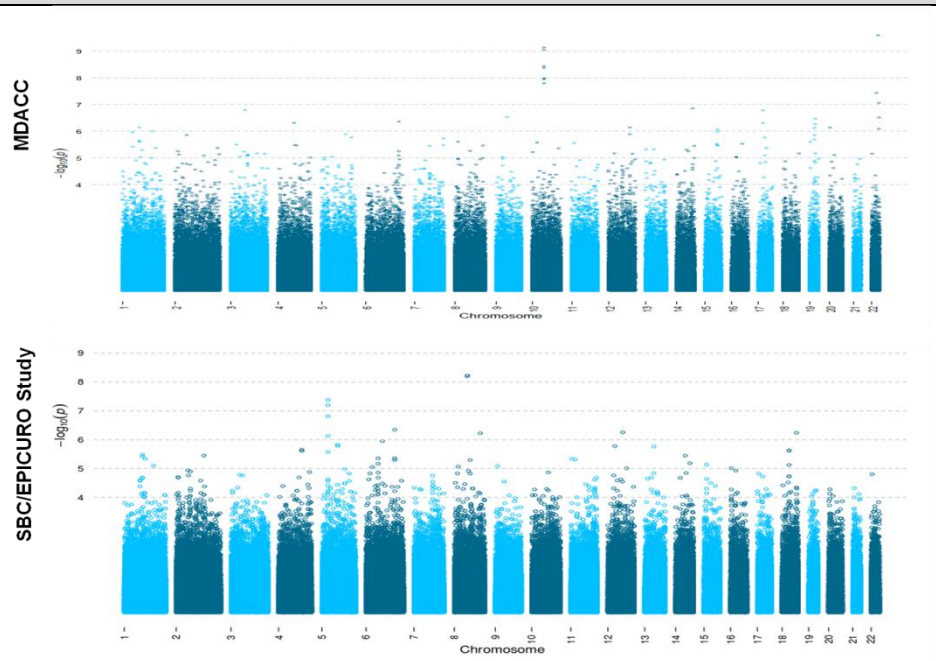


Supplementary Figure 1c. Relapse.

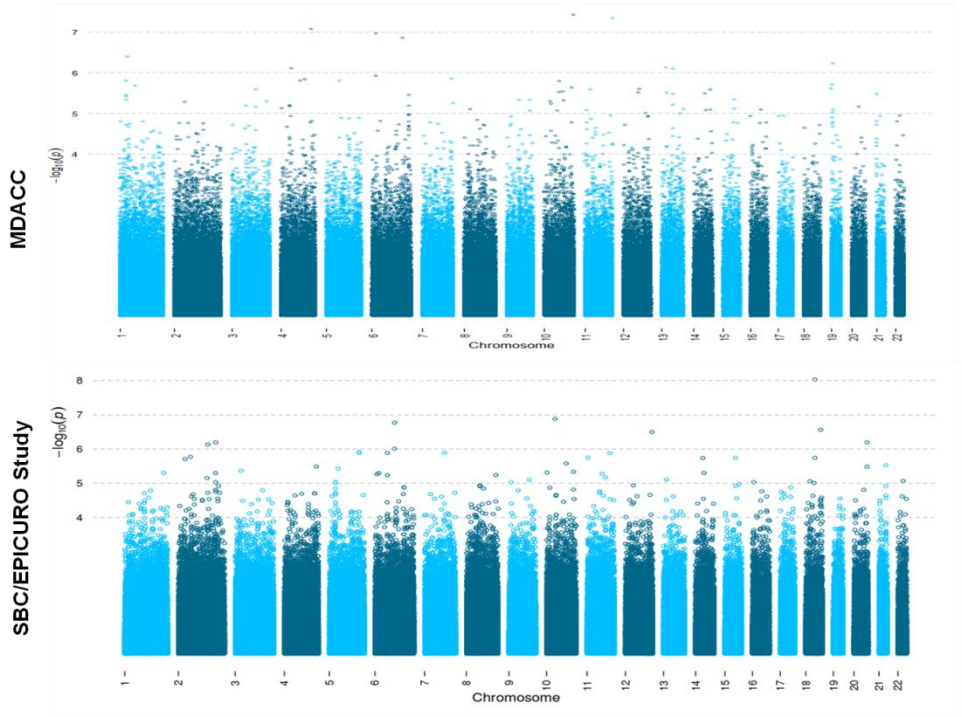


Supplementary Figure 2. Scatter (Manhattan) plot of chromosome position (x axis) against -log10 (p-value) for the 3 outcomes of interest (**a**: progression, **b**: BC-specific mortality, **c**: overall survival) in the two Discovery studies.

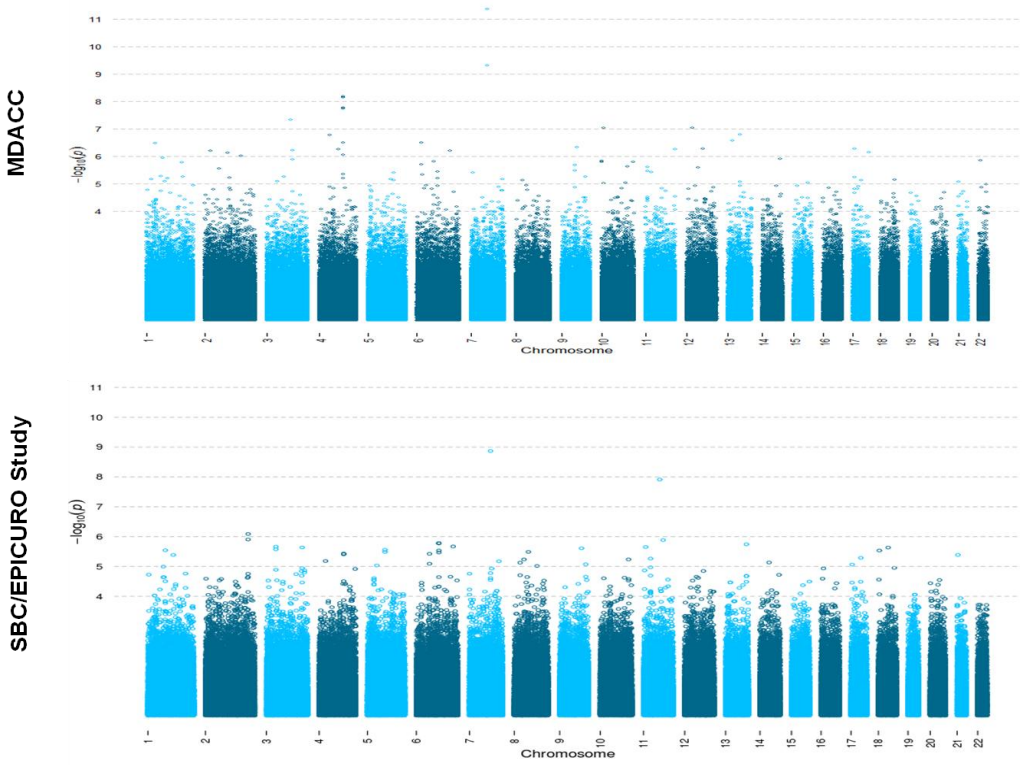
Supplementary Figure 2a. Progression.



Supplementary Figure 2b. BC-specific mortality.

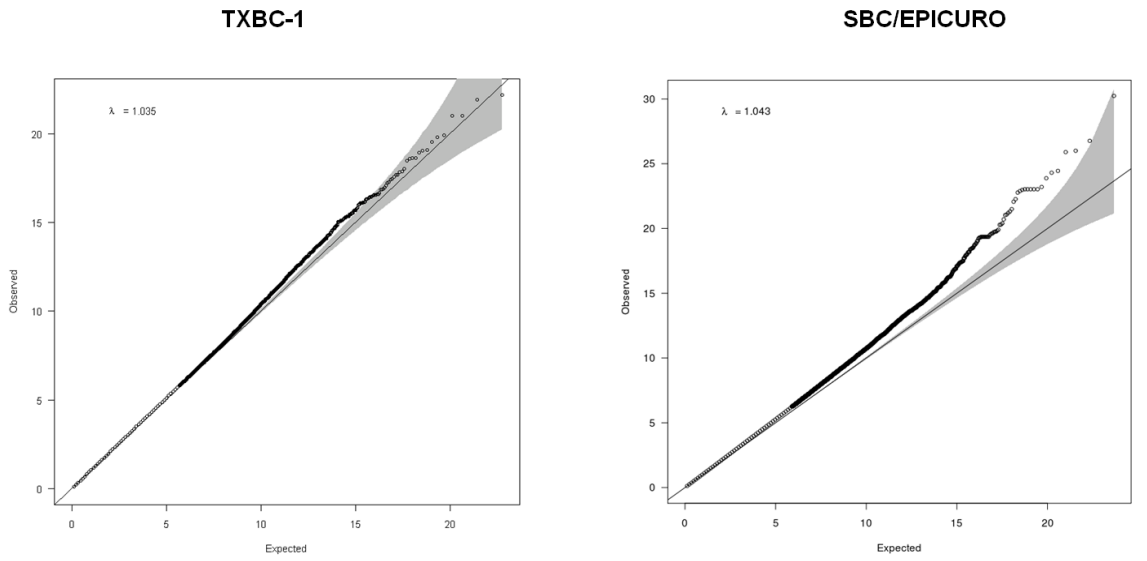


Supplementary Figure 2c. Overall survival.

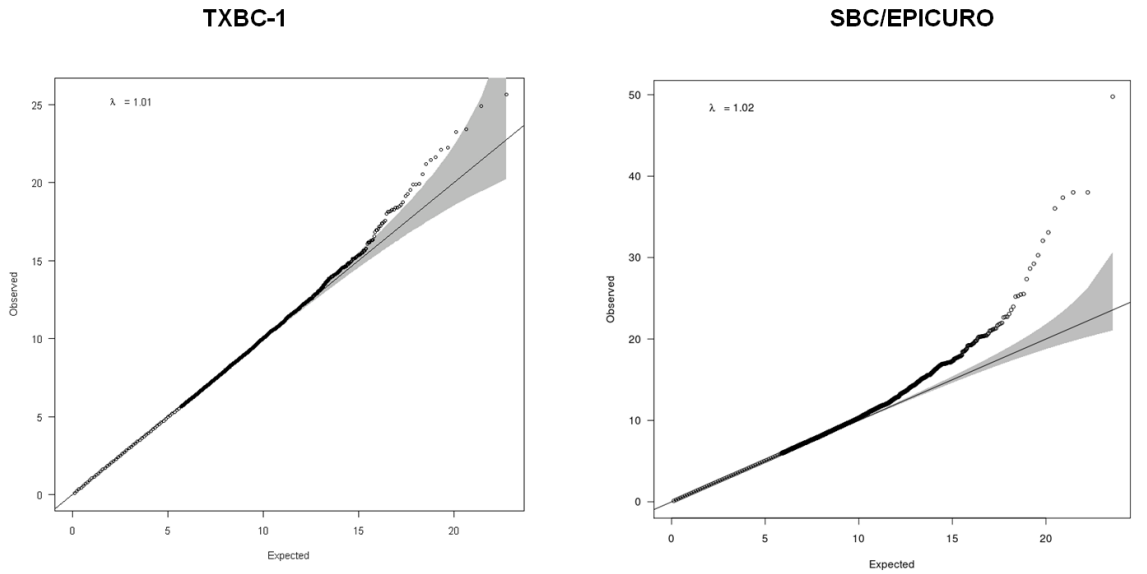


Supplementary Figure 3. Q-Q plot of observed against expected chi-square test in the dominant genetic model for the Discovery phase in NMIBC patients using adjusted analysis for tumor recurrence (**a**), progression (**b**) and relapse (**c**).

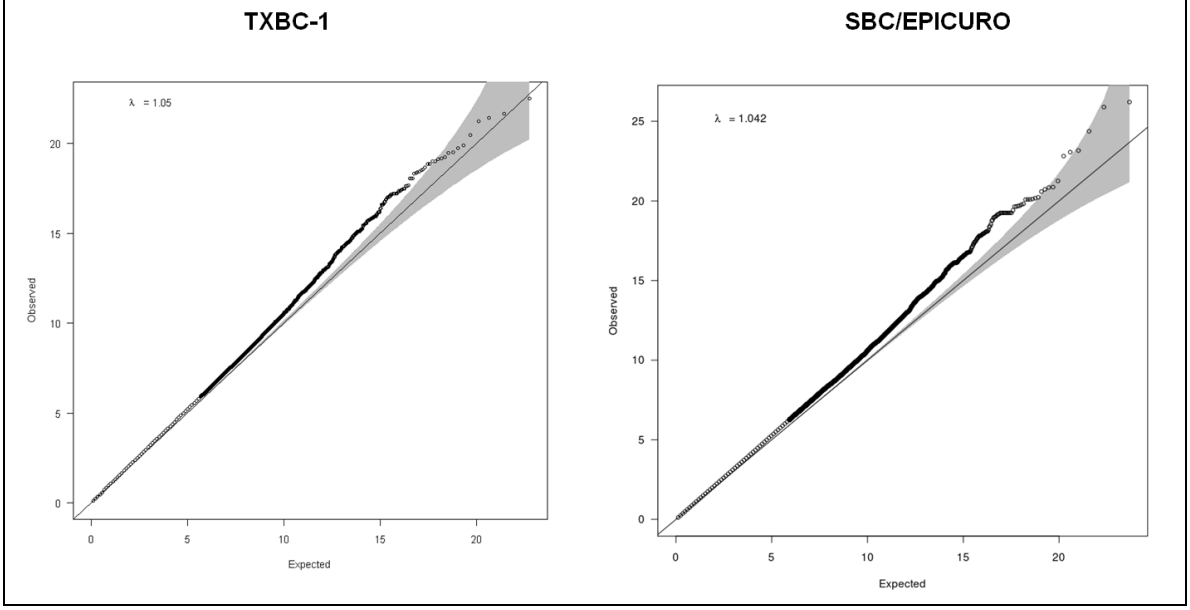
Supplementary Figure 3a. NMIBC Recurrence.



Supplementary Figure 3b. NMIBC Progression.

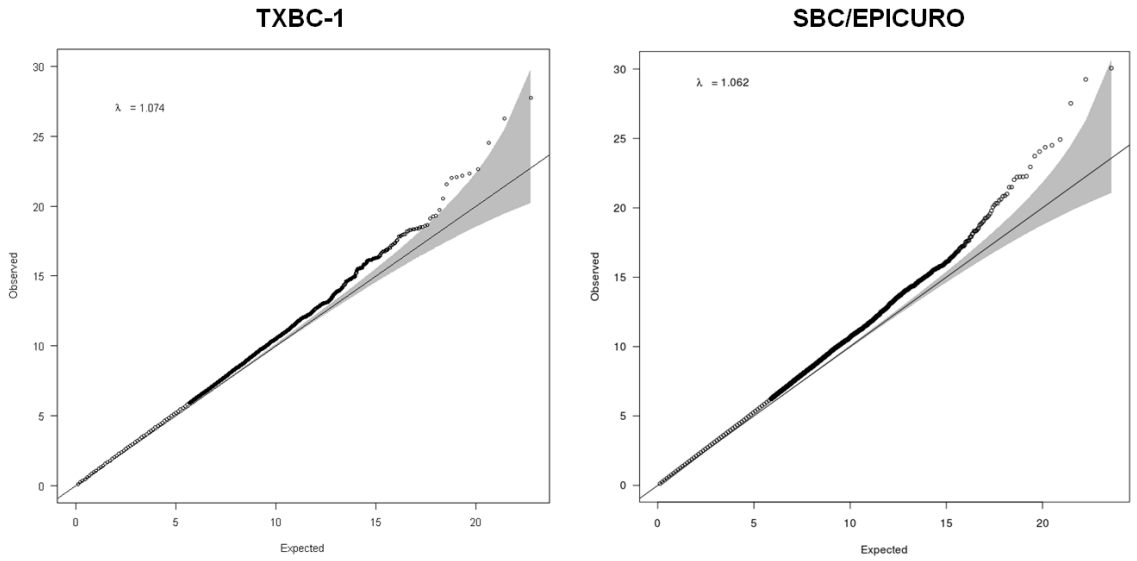


Supplementary Figure 3c. NMIBC Relapse.

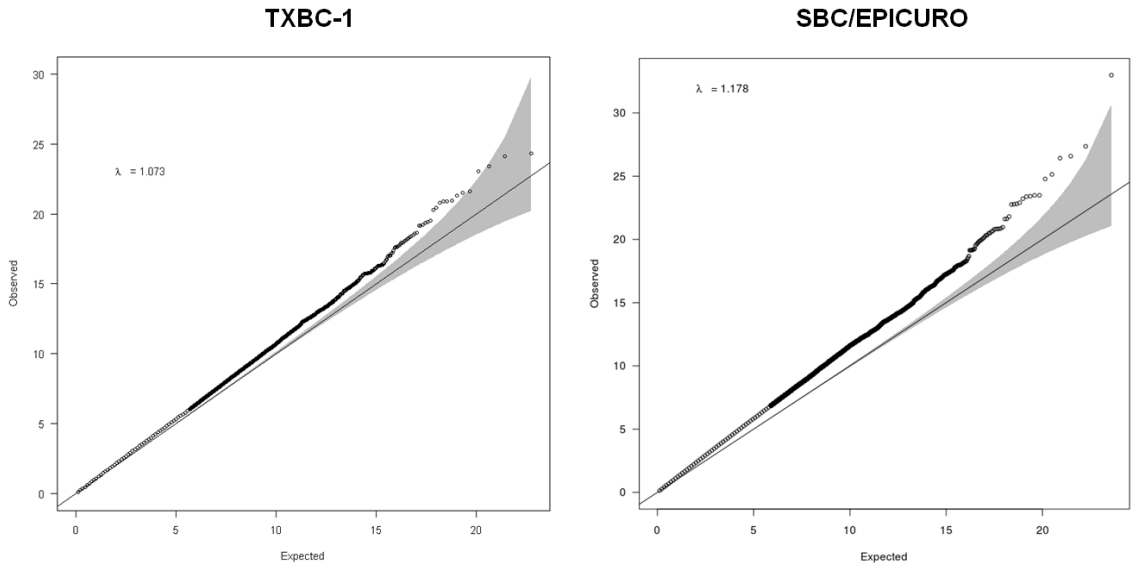


Supplementary Figure 4. Q-Q plot of observed against expected chi-square test in the dominant genetic model for the Discovery phase in MIBC patients using adjusted analysis for tumor progression (**a**), BC-specific mortality (**b**) and overall survival (**c**).

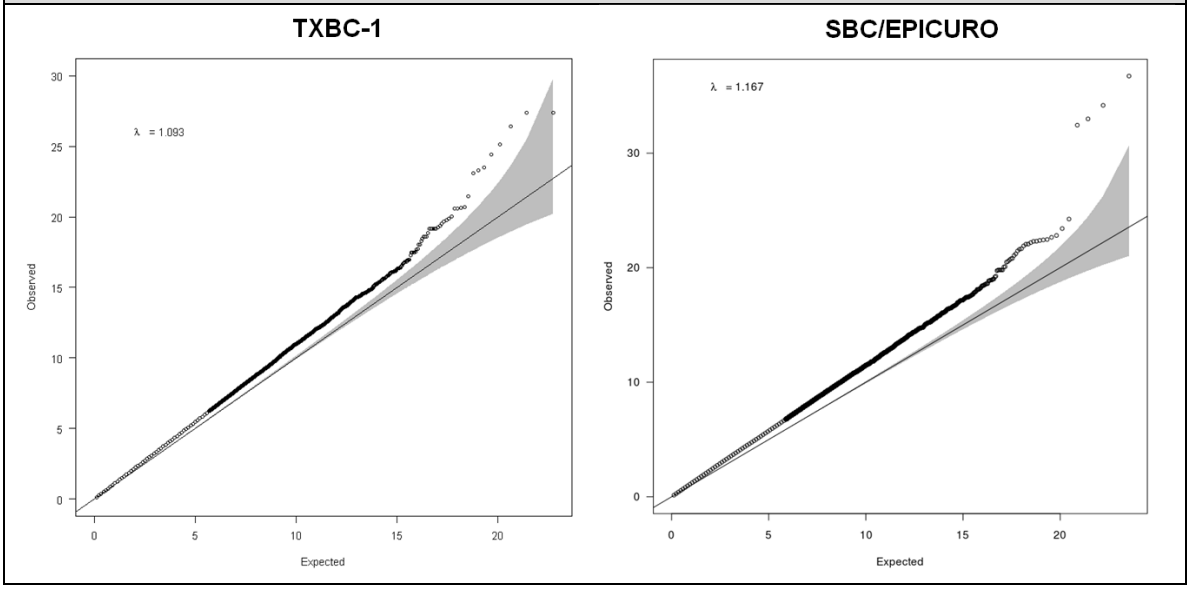
Supplementary Figure 4a. MIBC Progression.



Supplementary Figure 4b. MIBC BC-specific mortality.

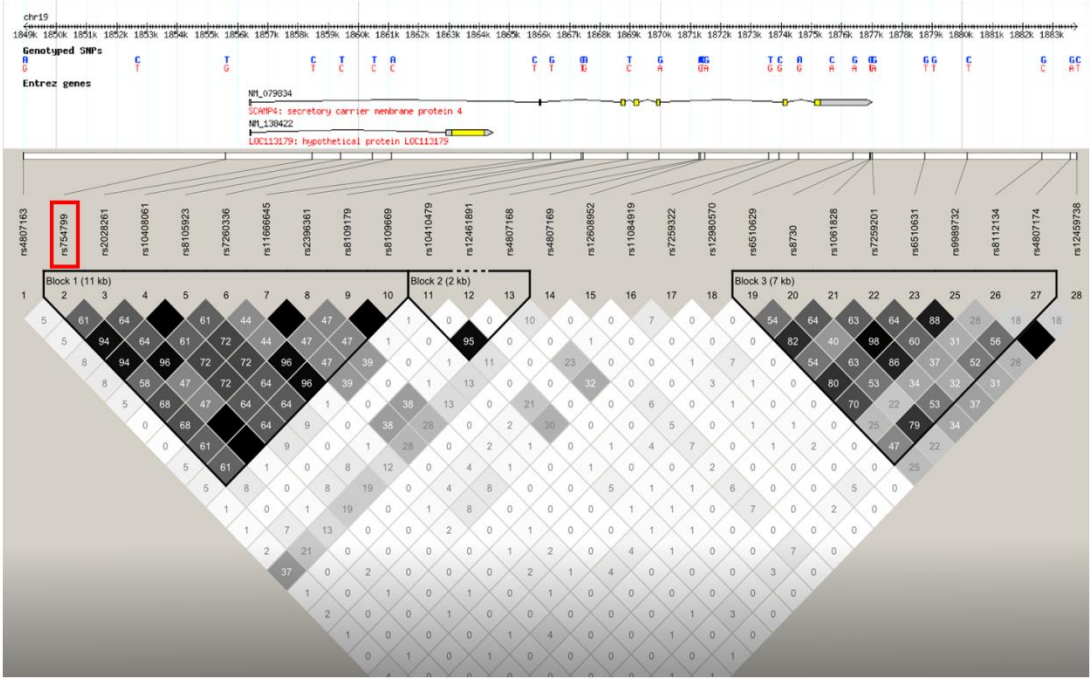


Supplementary Figure 4c. MIBC overall survival.

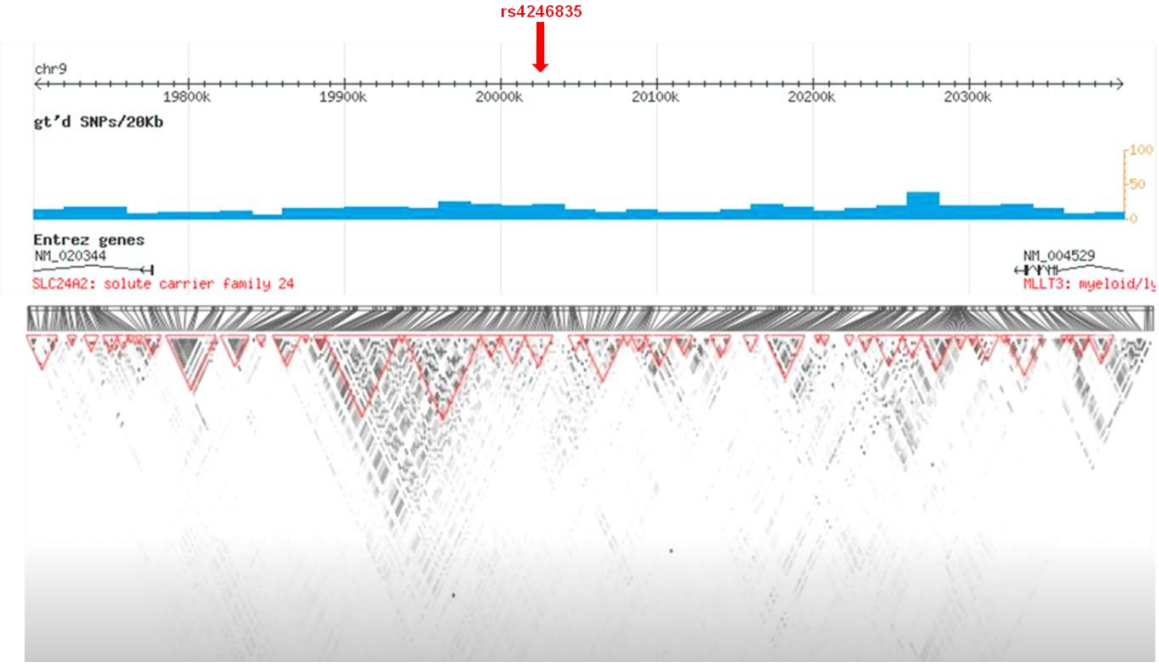


Supplementary Figure 5. Schematic view of the known genes, and neighborhood LD structure of the most significant SNPs associated with NMIBC recurrence, progression, and relapse. Location of the gene was obtained from UCSU genome browser. Pairwise LD structure by r^2 was derived from Haploview software (v4.1) using Hapmap ref 27, phase III, Feb09, NCBI 36, dbSNP 126, CEU dataset.

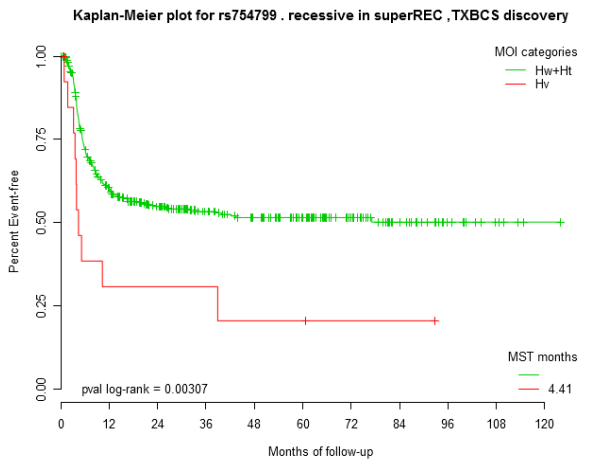
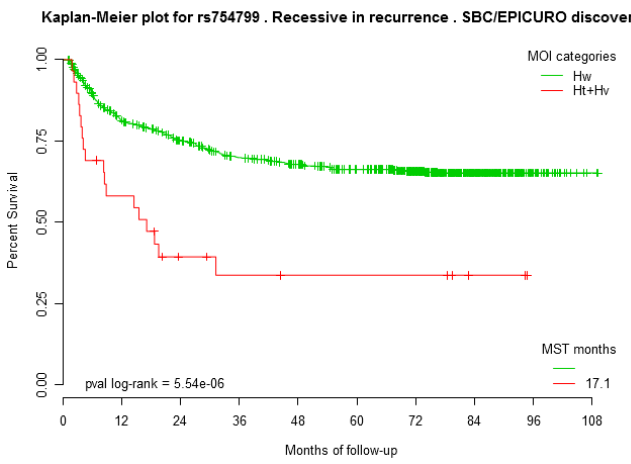
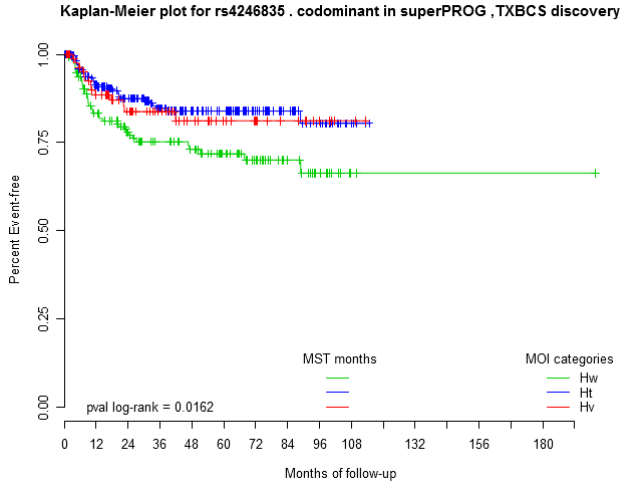
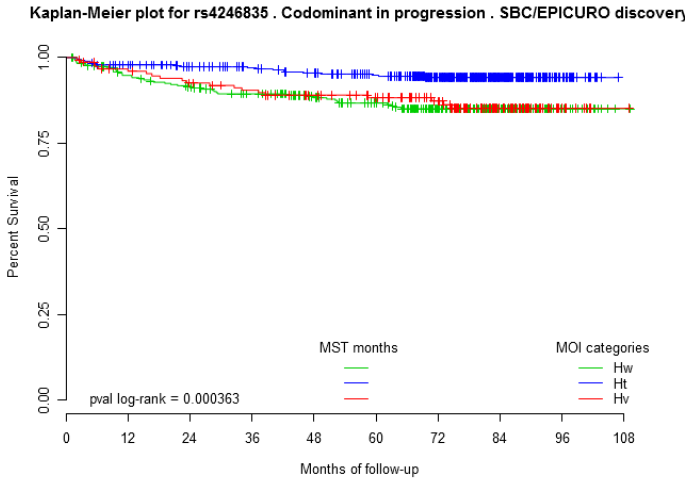
Supplementary Figure 2a. *rs754799*, chromosome 19, 1,848 – 1,925 Kb

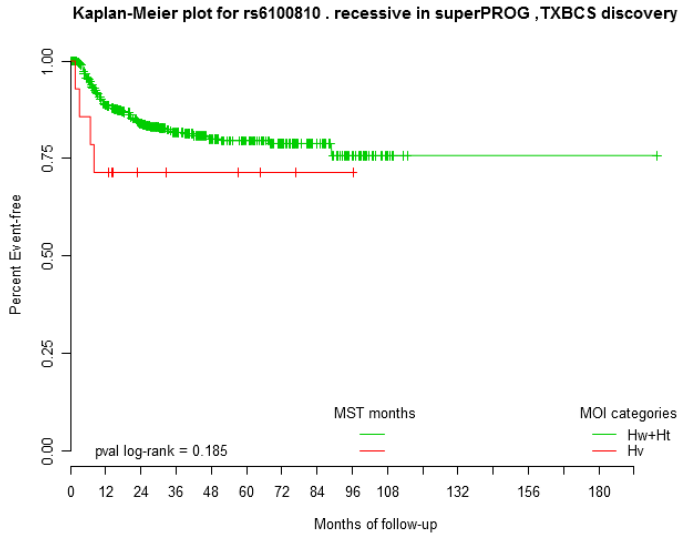
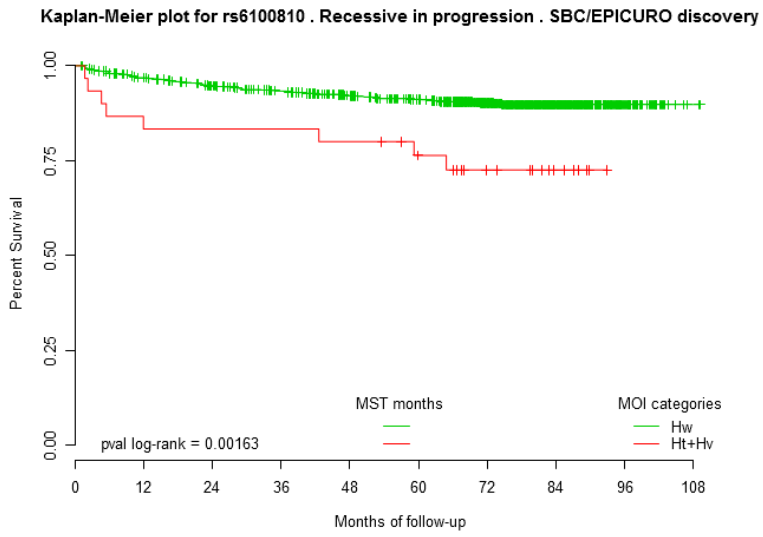
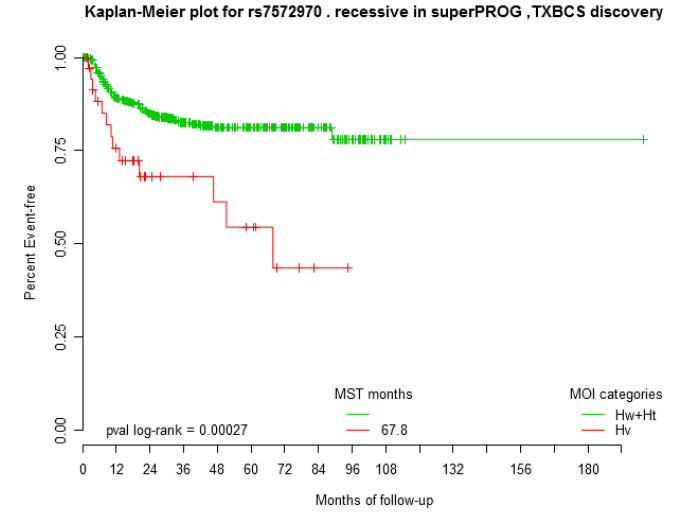
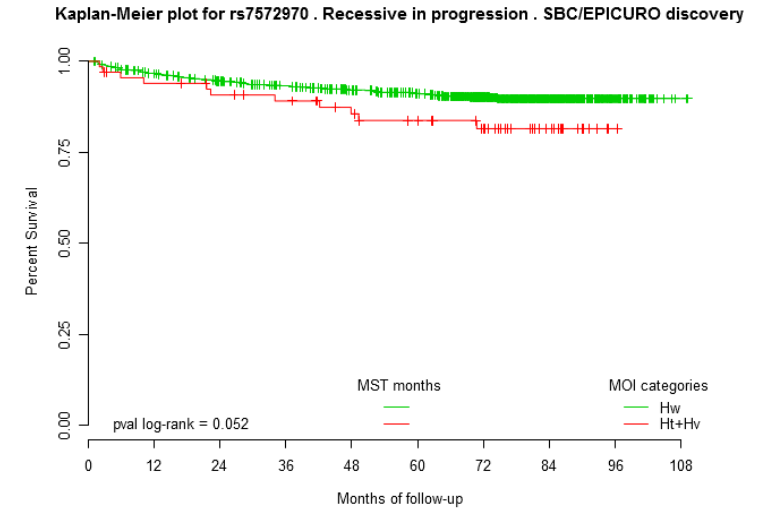


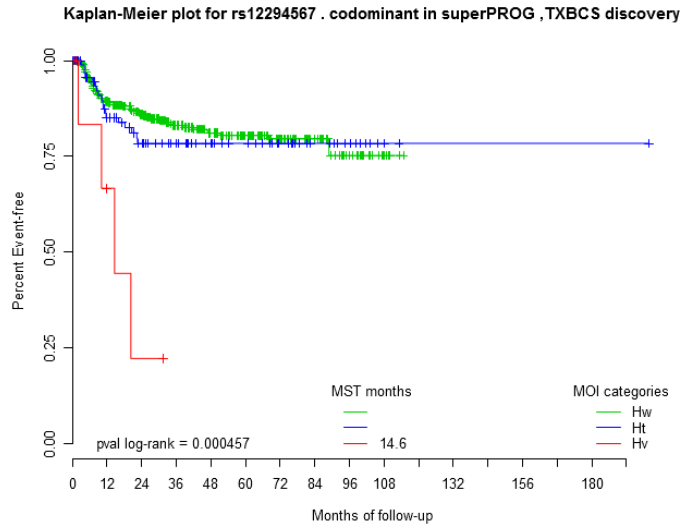
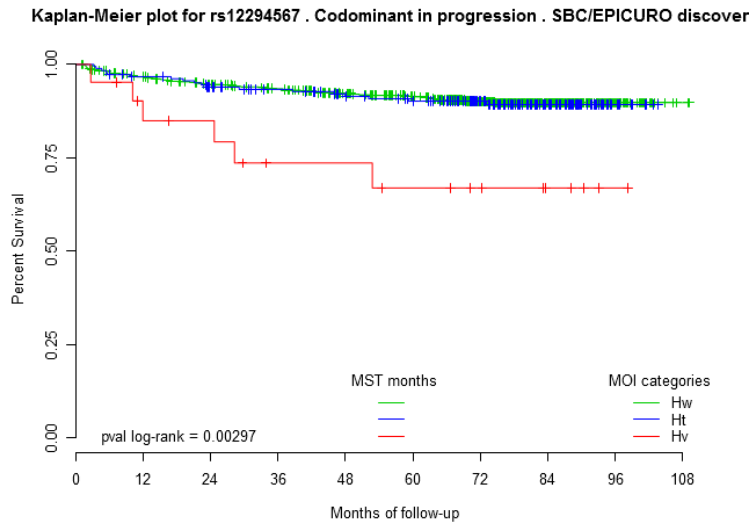
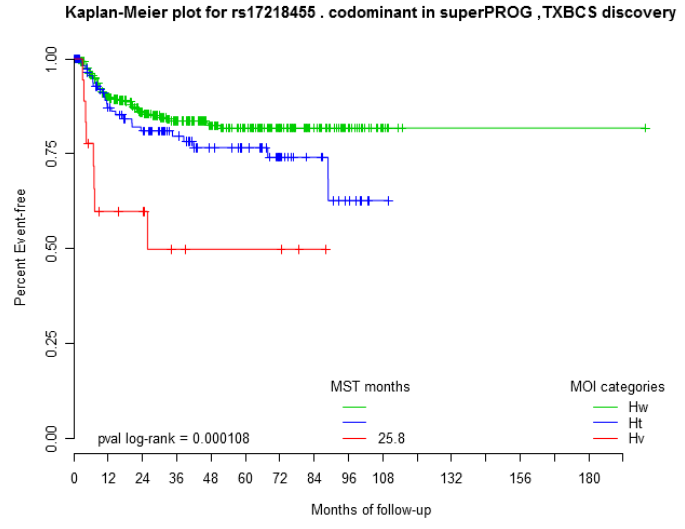
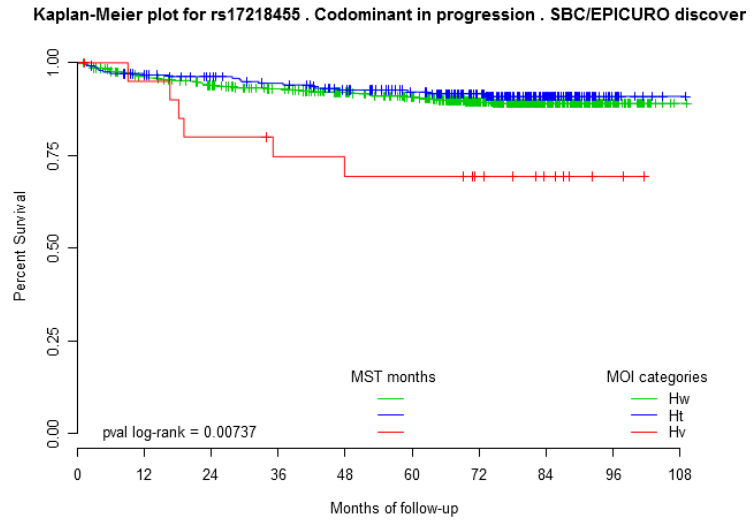
Supplementary Figure 2b. *rs4246835*, chromosome 9, 19,723 – 20,383 Kb

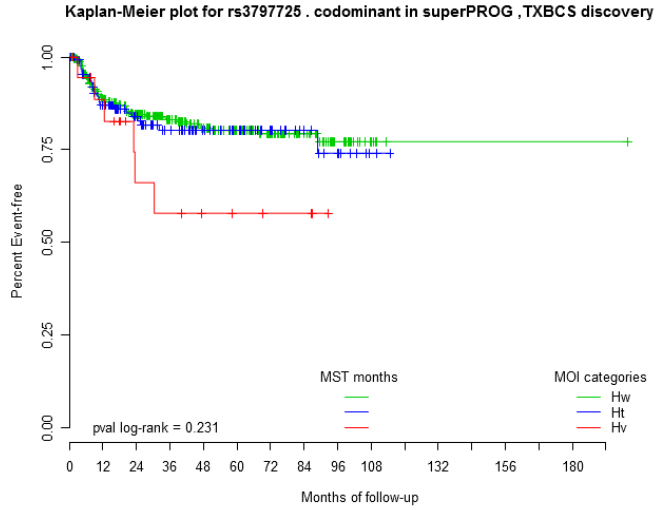
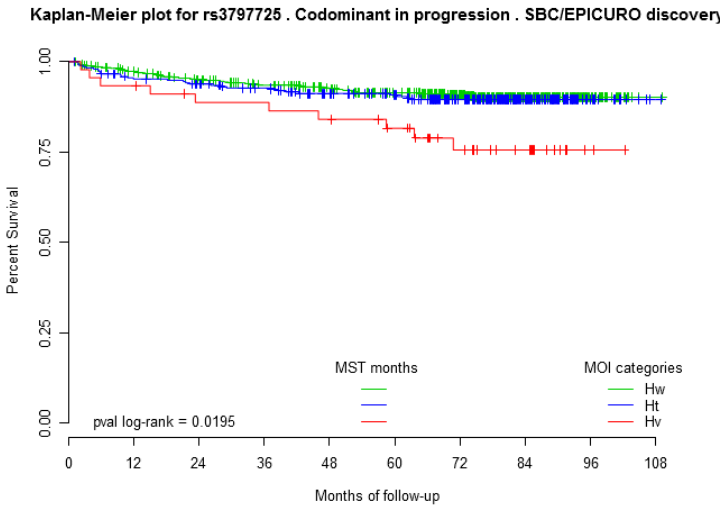
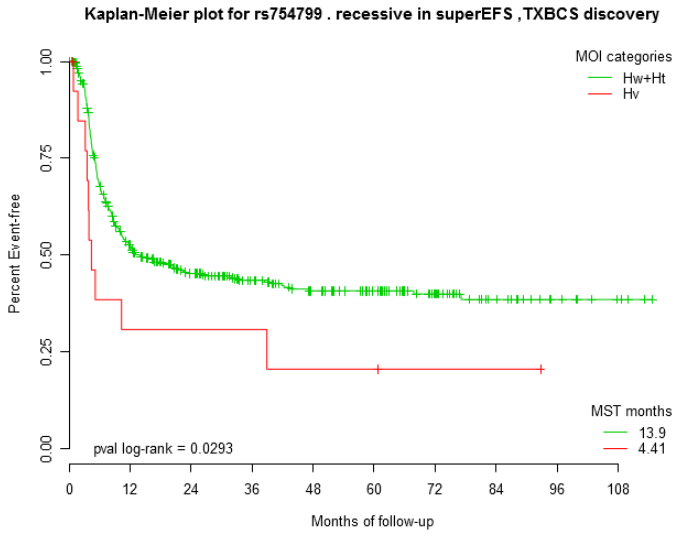
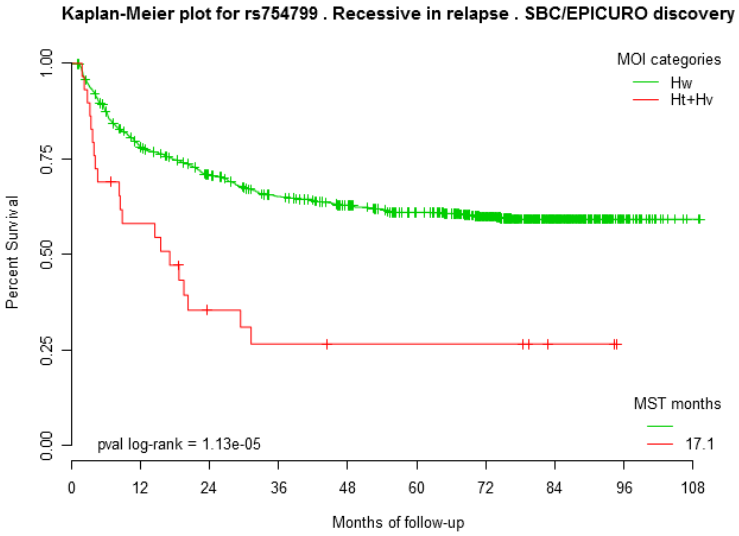


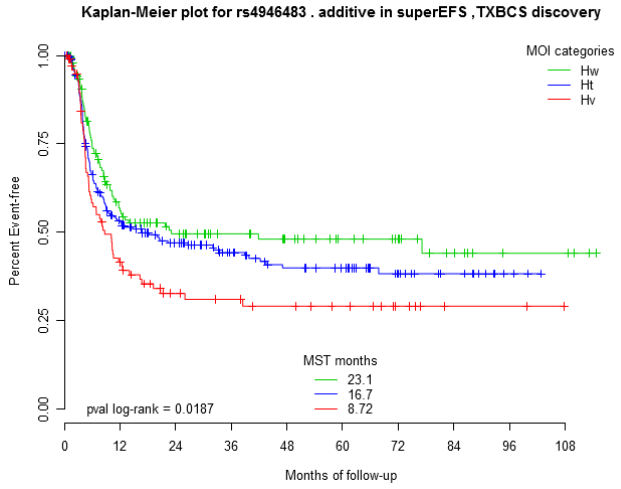
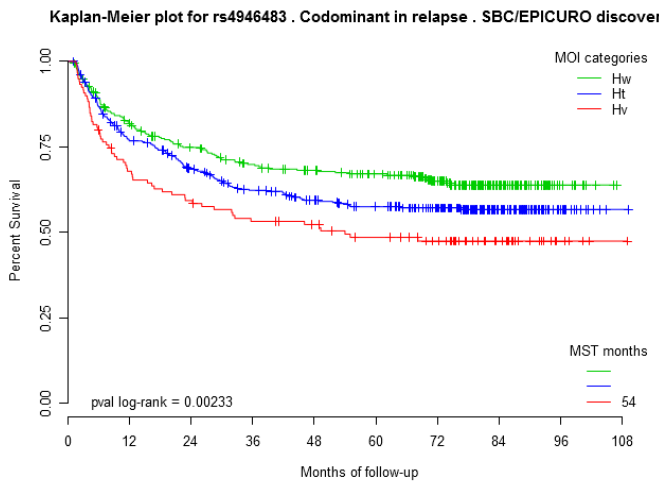
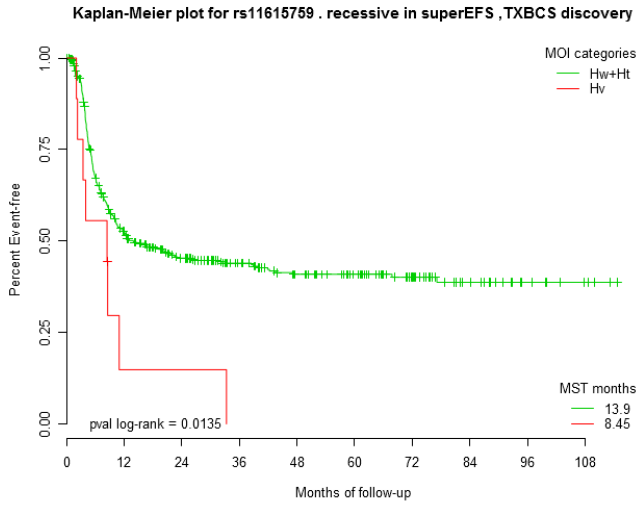
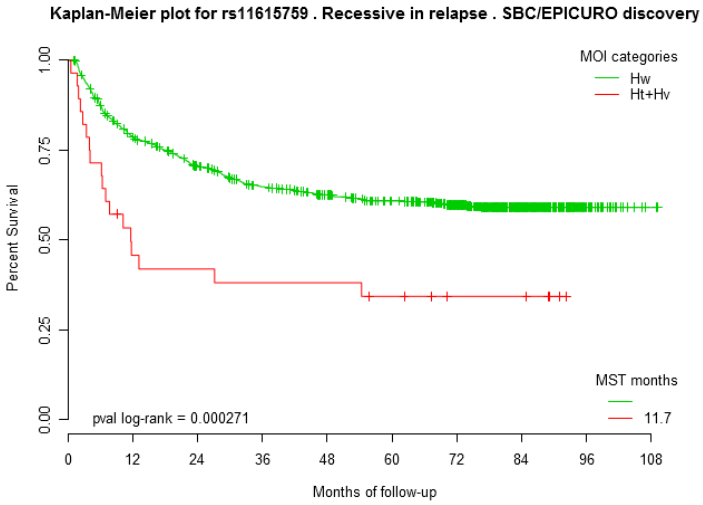
Supplementary Figure 6. Kaplan-Meier plots for NMIBC SNPs in *Table 5* for each Discovery population. Logrank test and median follow-up time using reverse Kaplan-Meier estimator.

TXBC-1	SBC/EPICURO
	
	

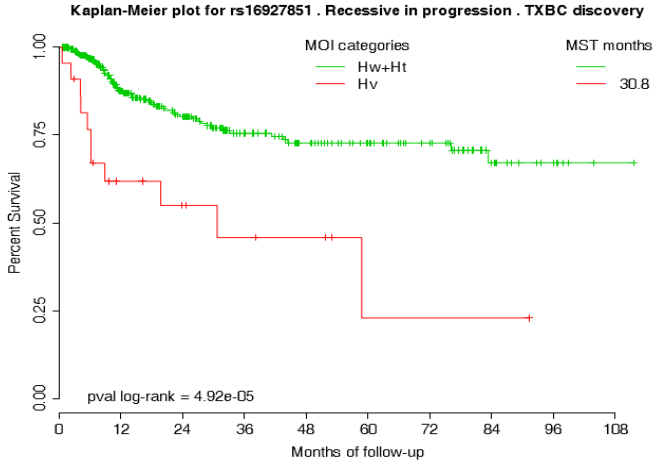
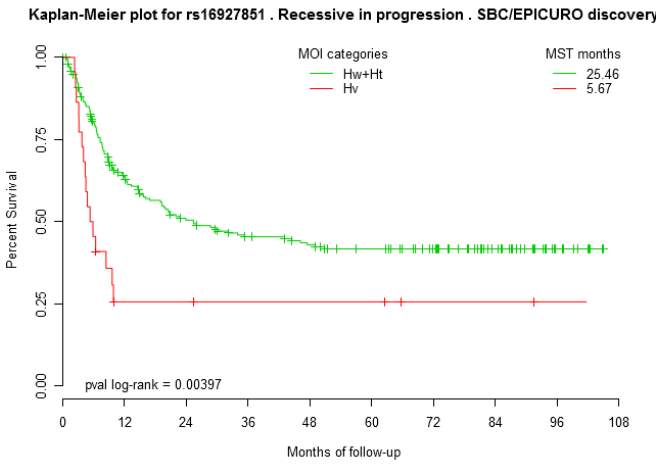
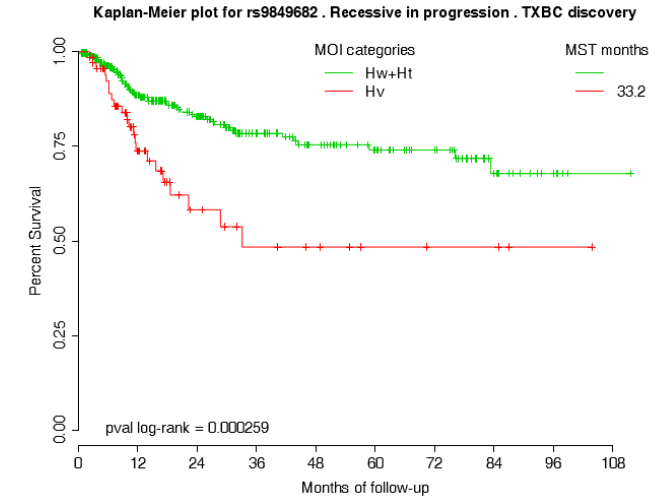
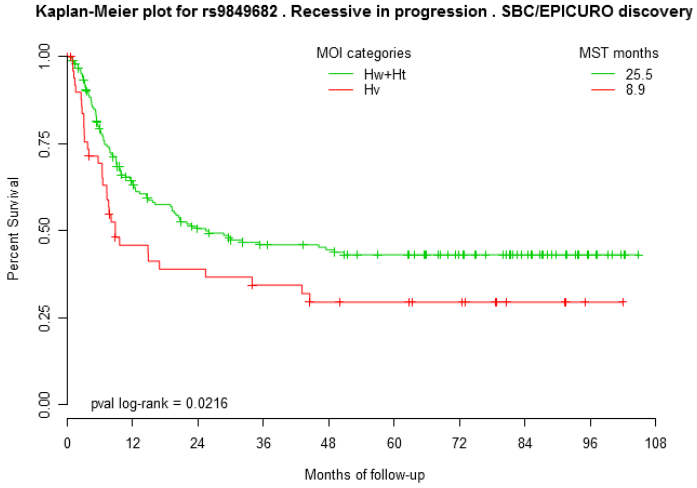
TXBC-1	SBC/EPICURO
	
	

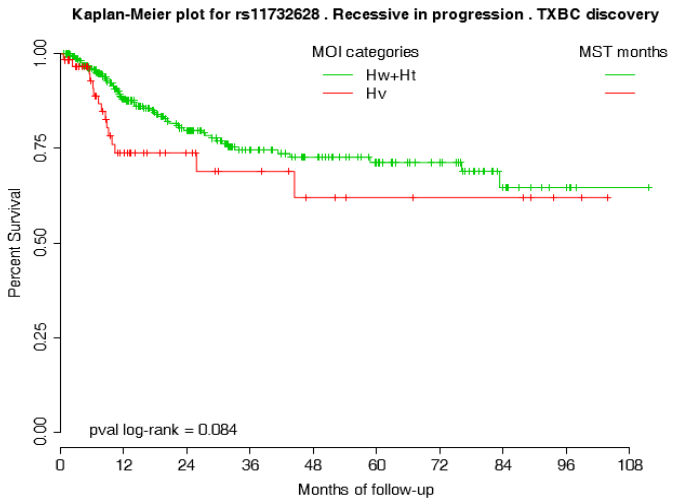
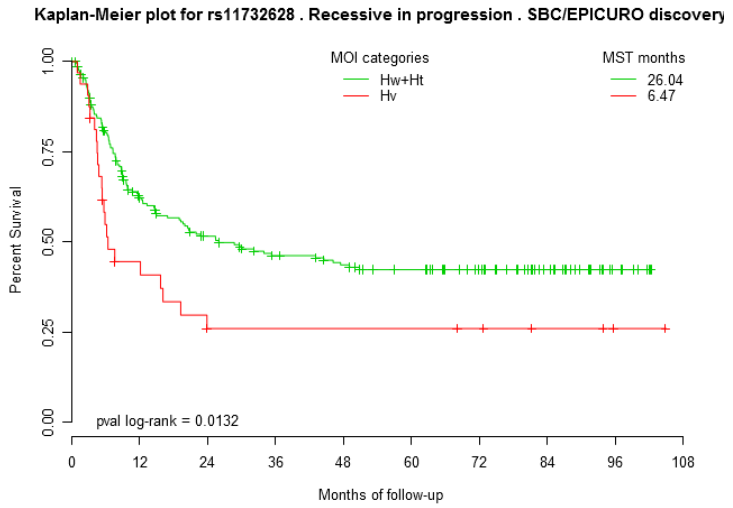
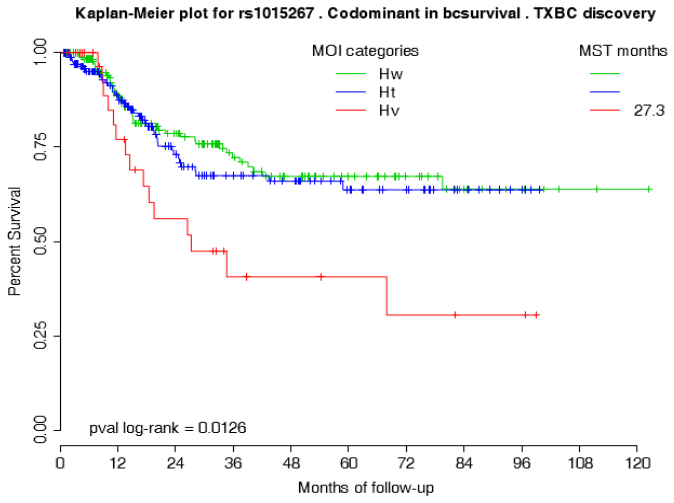
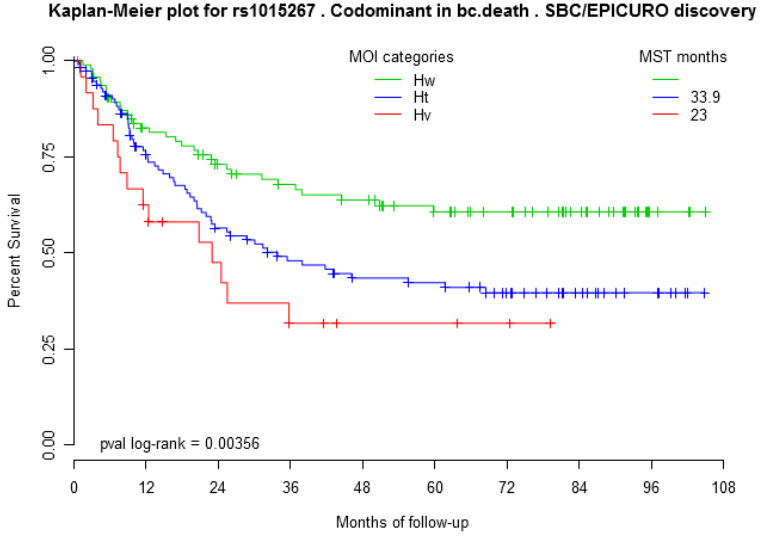
TXBC-1	SBC/EPICURO
	
	

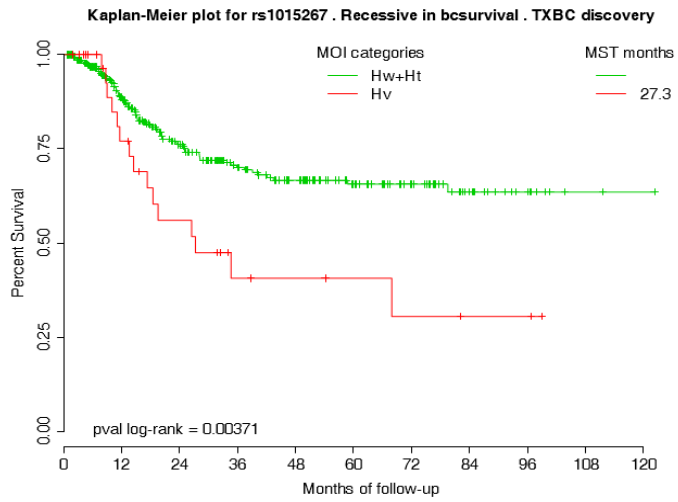
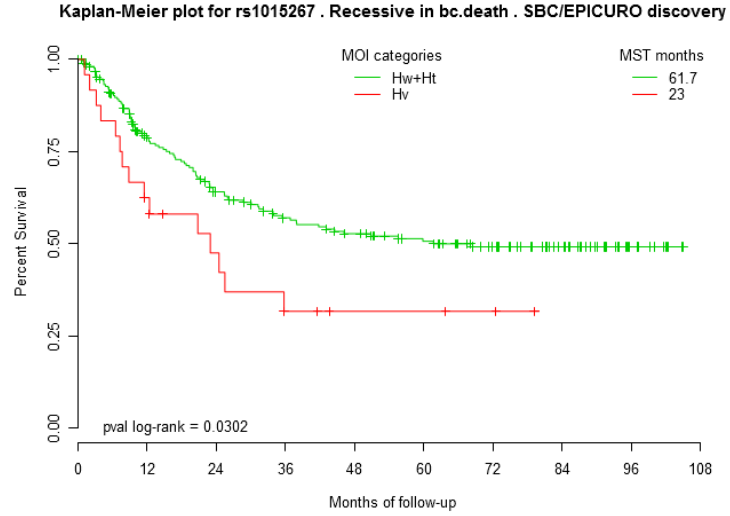
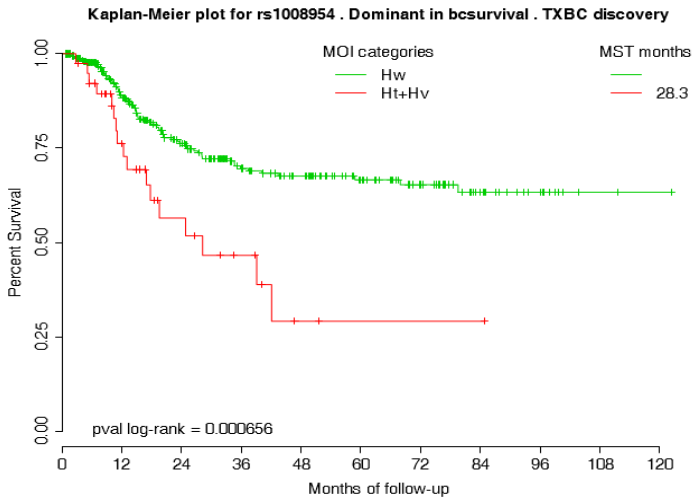
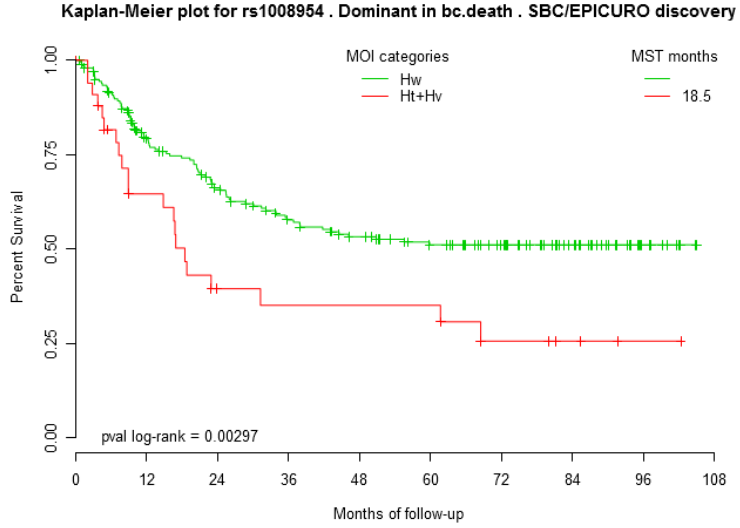
TXBC-1	SBC/EPICURO
	
	

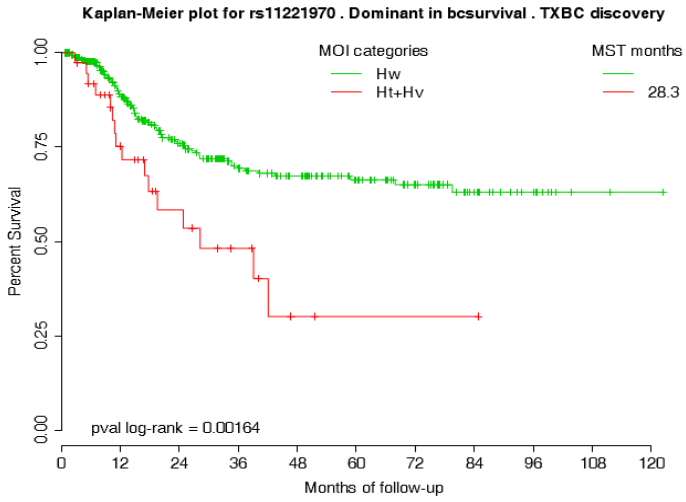
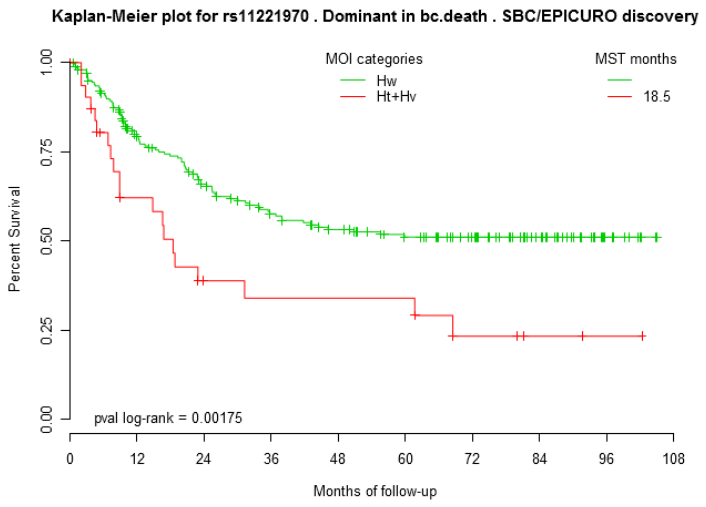
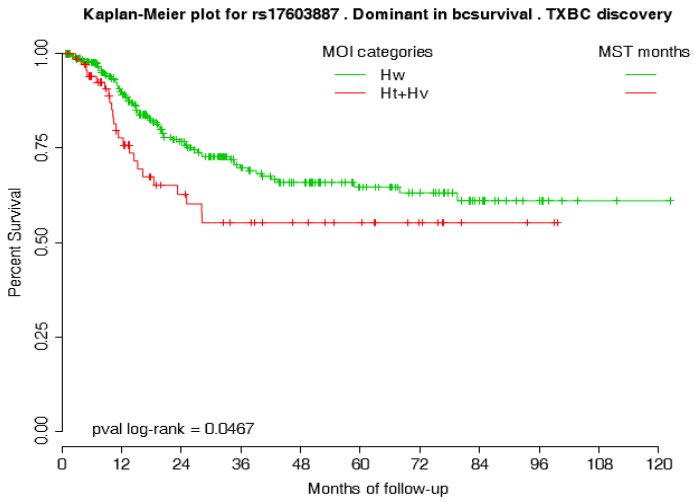
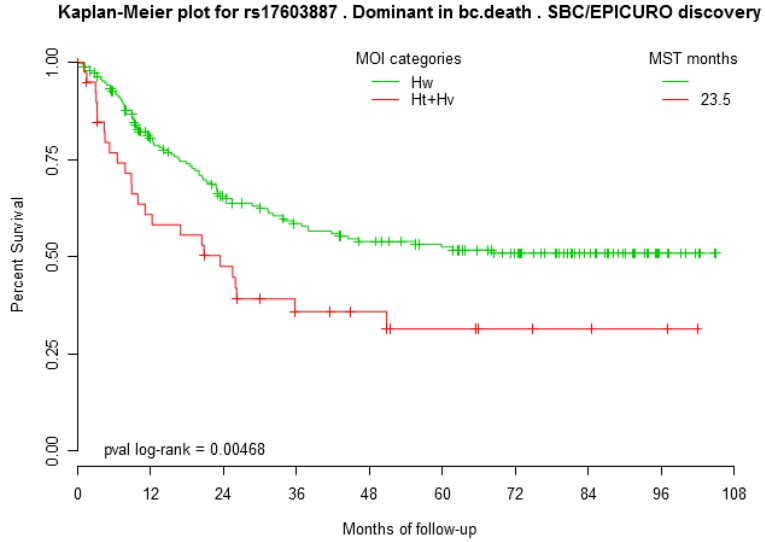
TXBC-1	SBC/EPICURO
	
	

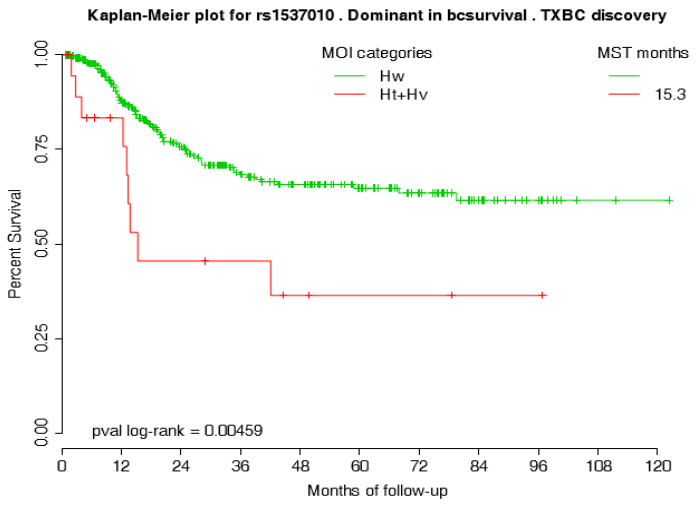
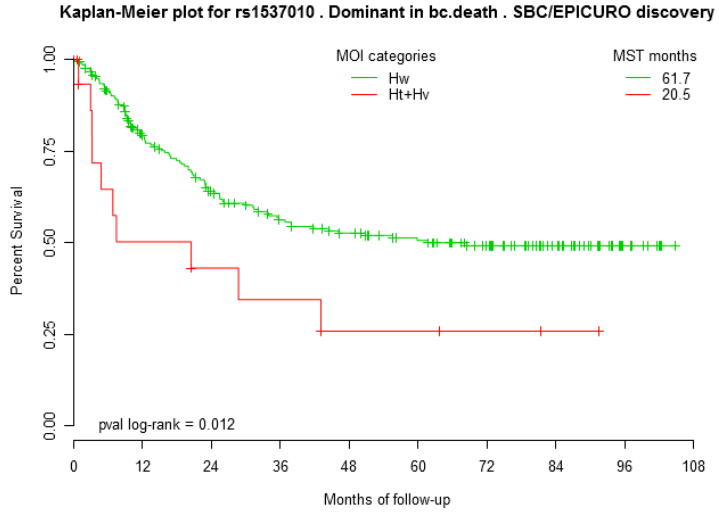
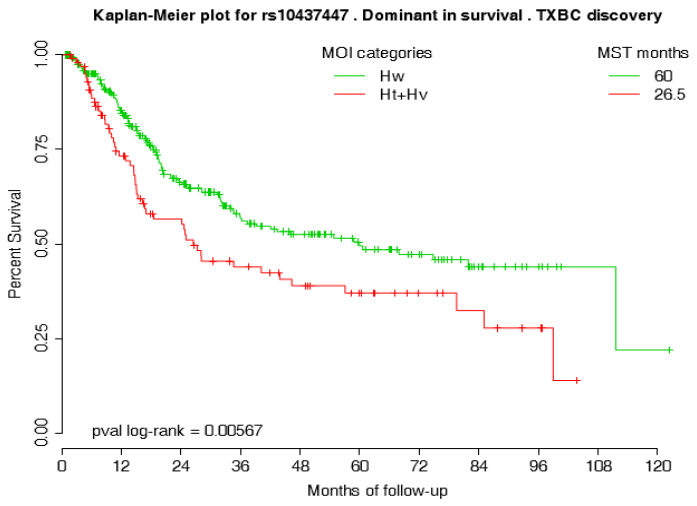
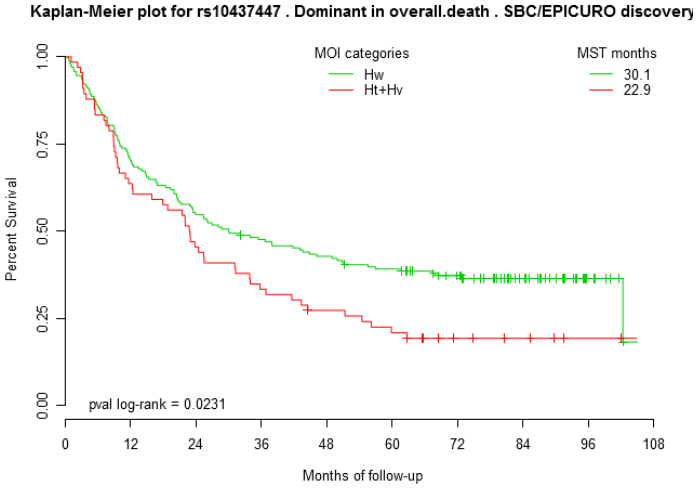
Supplementary Figure 7. Kaplan-Meier plots for MIBC SNPs in *Table 6* for each Discovery population. Logrank test and median follow-up time using reverse Kaplan-Meier estimator.

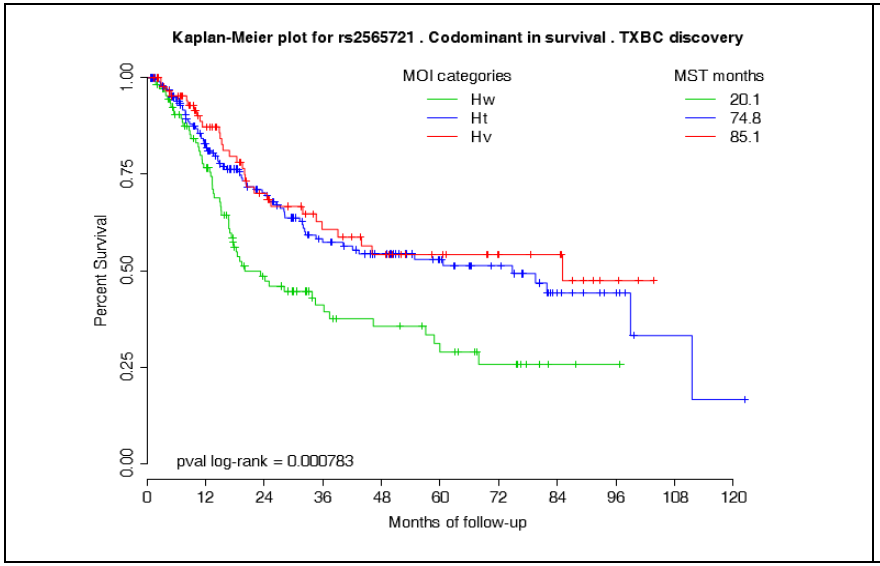
TXBC-1	SBC/EPICURO
	
	

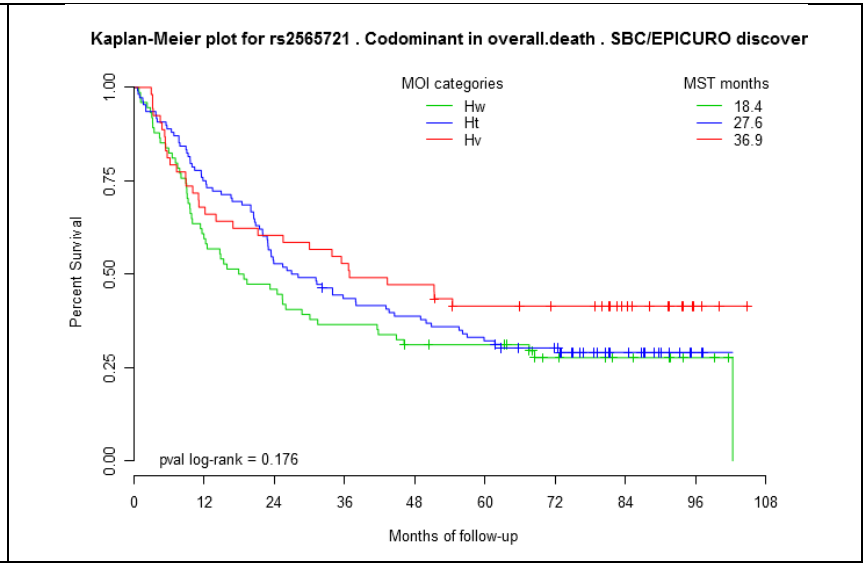
TXBC-1	SBC/EPICURO
	
	

TXBC-1	SBC/EPICURO
	
	

TXBC-1	SBC/EPICURO
	
	

TXBC-1	SBC/EPICURO
	
	


Kaplan-Meier plot for rs2565721 . Codominant in survival . TXBC discovery
MOI categories
Hw
Ht
Hv
MST months
20.1
74.8
85.1
Percent Survival
0.00
0.25
0.50
0.75
1.00
pval log-rank = 0.000783
0
12
24
36
48
60
72
84
96
108
120
Months of follow-up


Kaplan-Meier plot for rs2565721 . Codominant in overall.death . SBC/EPICURO discover
MOI categories
Hw
Ht
Hv
MST months
18.4
27.6
36.9
Percent Survival
0.00
0.25
0.50
0.75
1.00
pval log-rank = 0.176
0
12
24
36
48
60
72
84
96
108
Months of follow-up

Supplementary Figure 8. Statistical power (Y-axis) for each NMIBC (a: recurrence, b: progression and c: relapse) and each MIBC (d: progression, e: BC-specific mortality and f: overall survival) outcome under the participating study's conditions of sample size, HR, proportion of variants, proportion of events and the type I error.

Figure 8a. NMIBC Recurrence

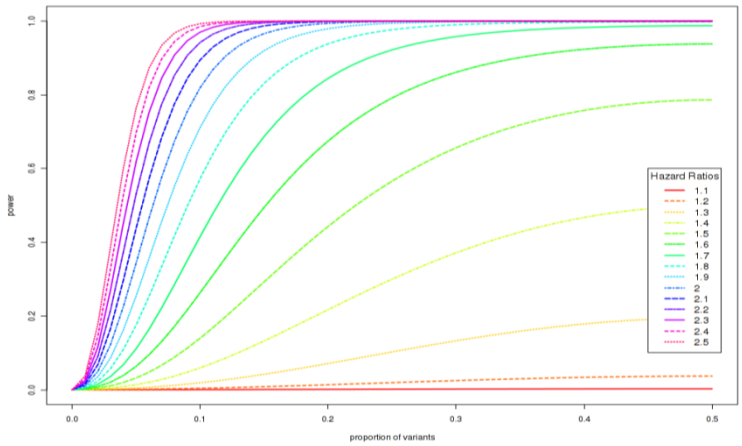


Figure 8b. NMIBC Progression

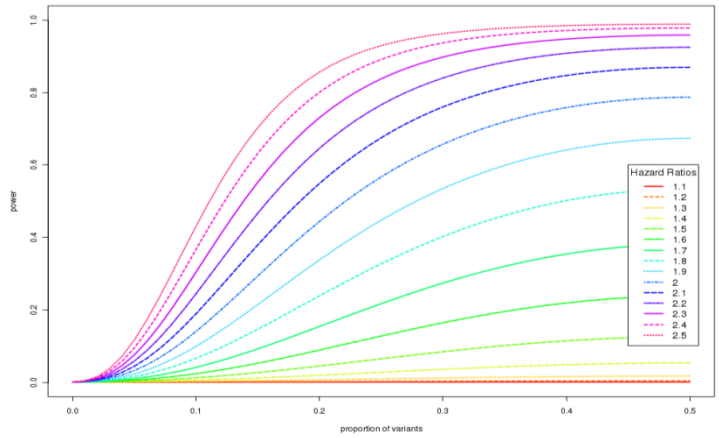


Figure 8c. NMIBC Relapse

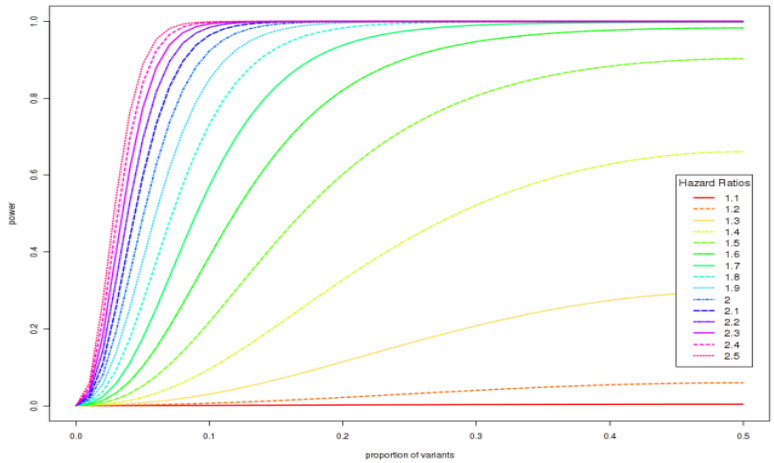


Figure 8d. MIBC Progression

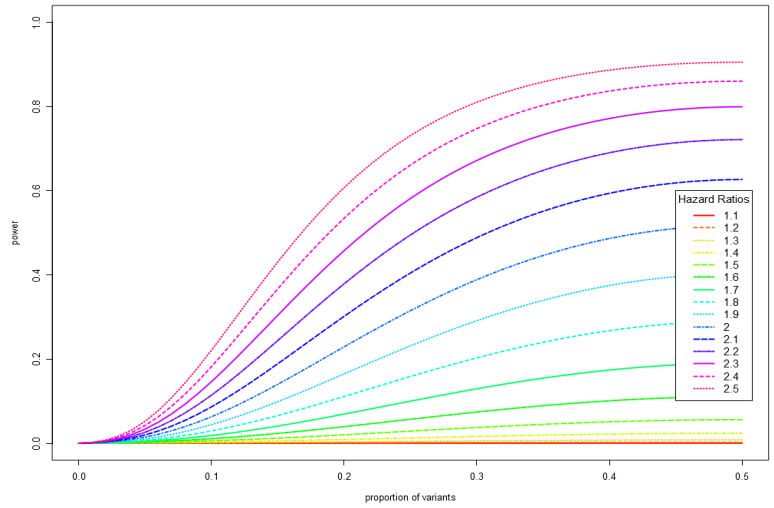


Figure 8e. MIBC BC-specific mortality

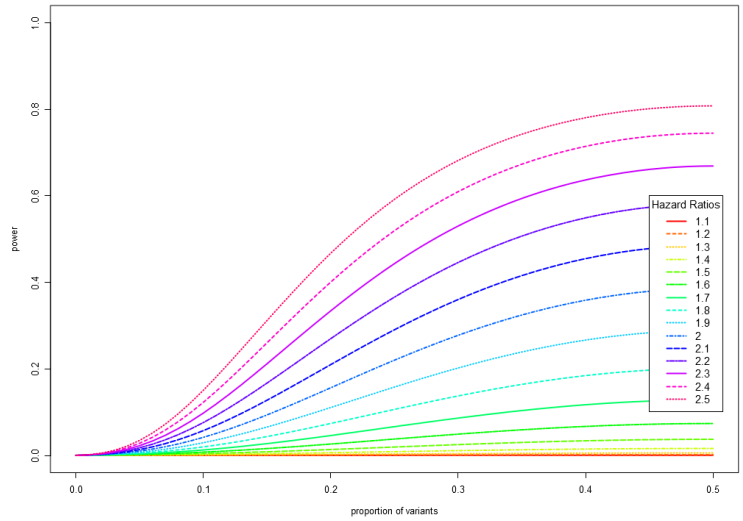
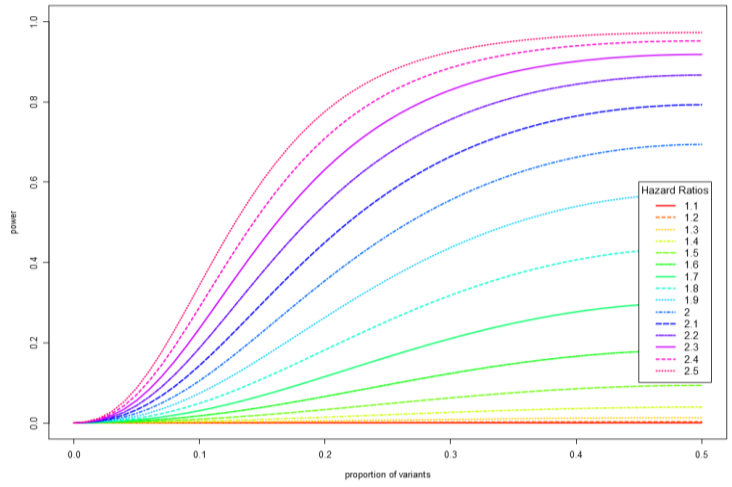


Figure 8f. MIBC Overall survival



Supplementary Figure 9. Analysis of similarity of the enriched pathways obtained after the gene set analyses (GSA). Hierarchical clusters followed by dynamic tree cut and dynamic hybrid cut were performed in order to merge the similar cluster branches in modules (A-Z).

Figure 9a. GO categories associated with progression in NMIBC using ALIGATOR

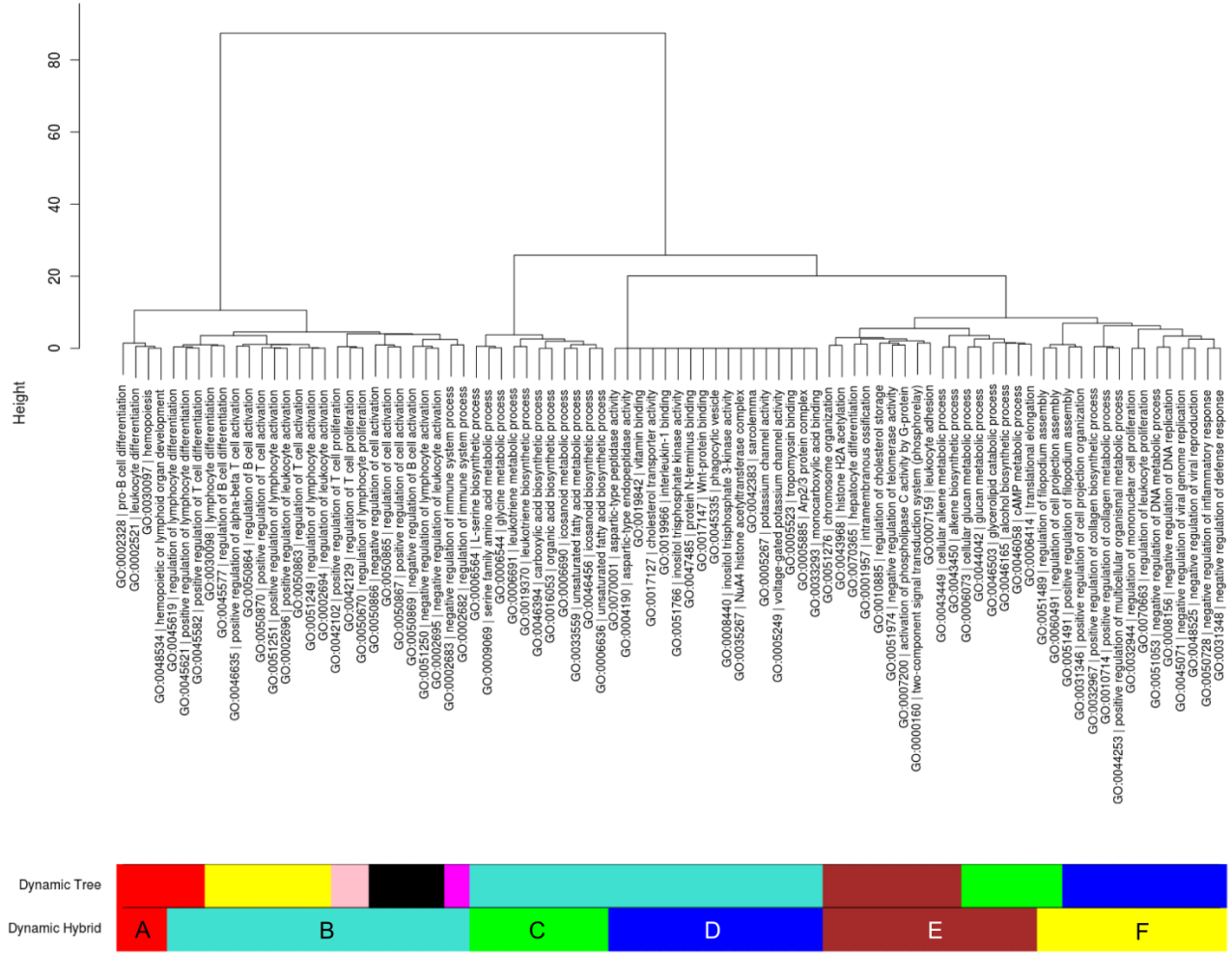


Figure 9b. GO categories and canonical pathways associated with BC-specific mortality in MIBC using GeSBAP

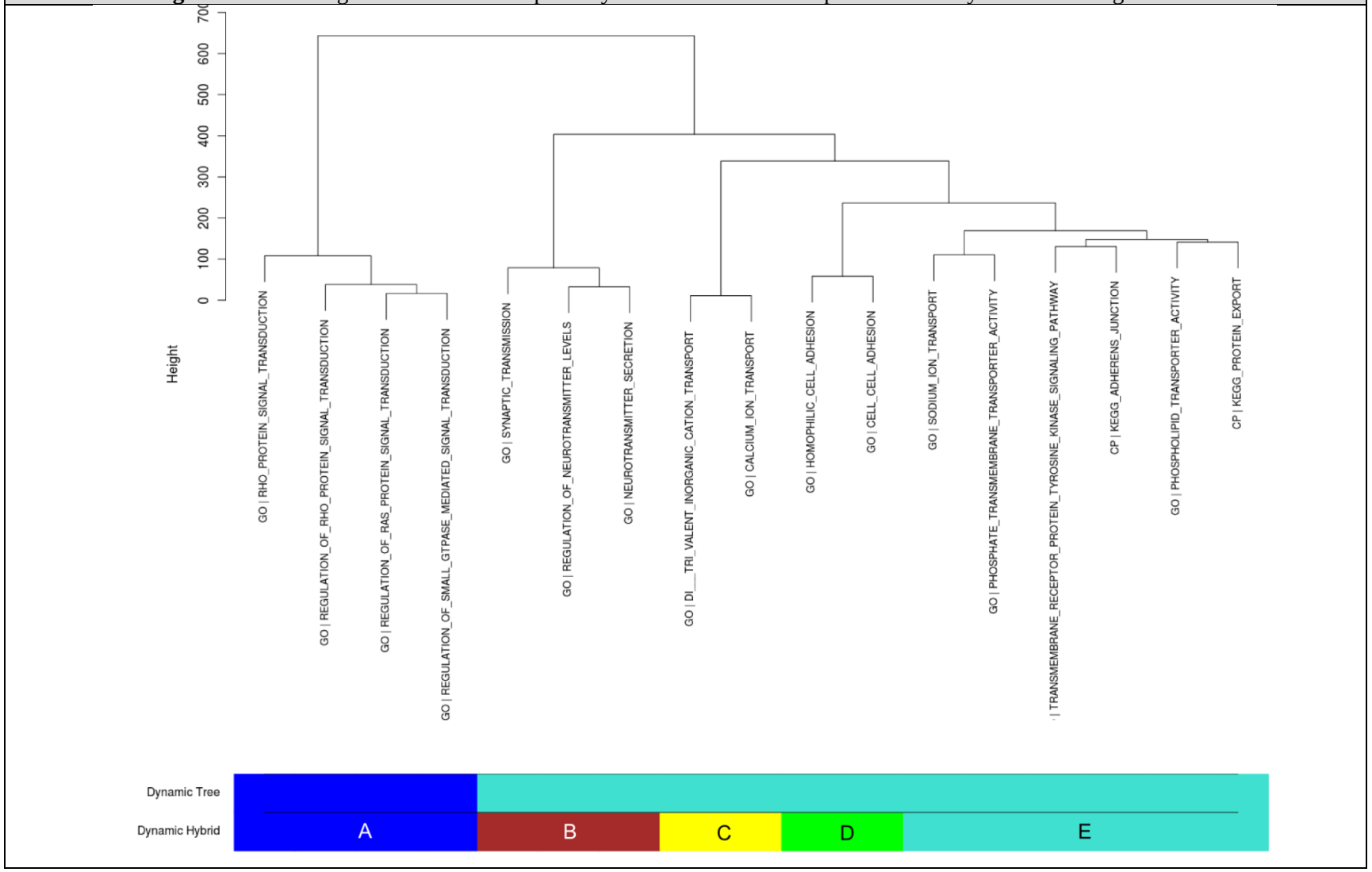


Figure 9c. GO categories and canonical pathways associated with progression in MIBC using GeSBAP

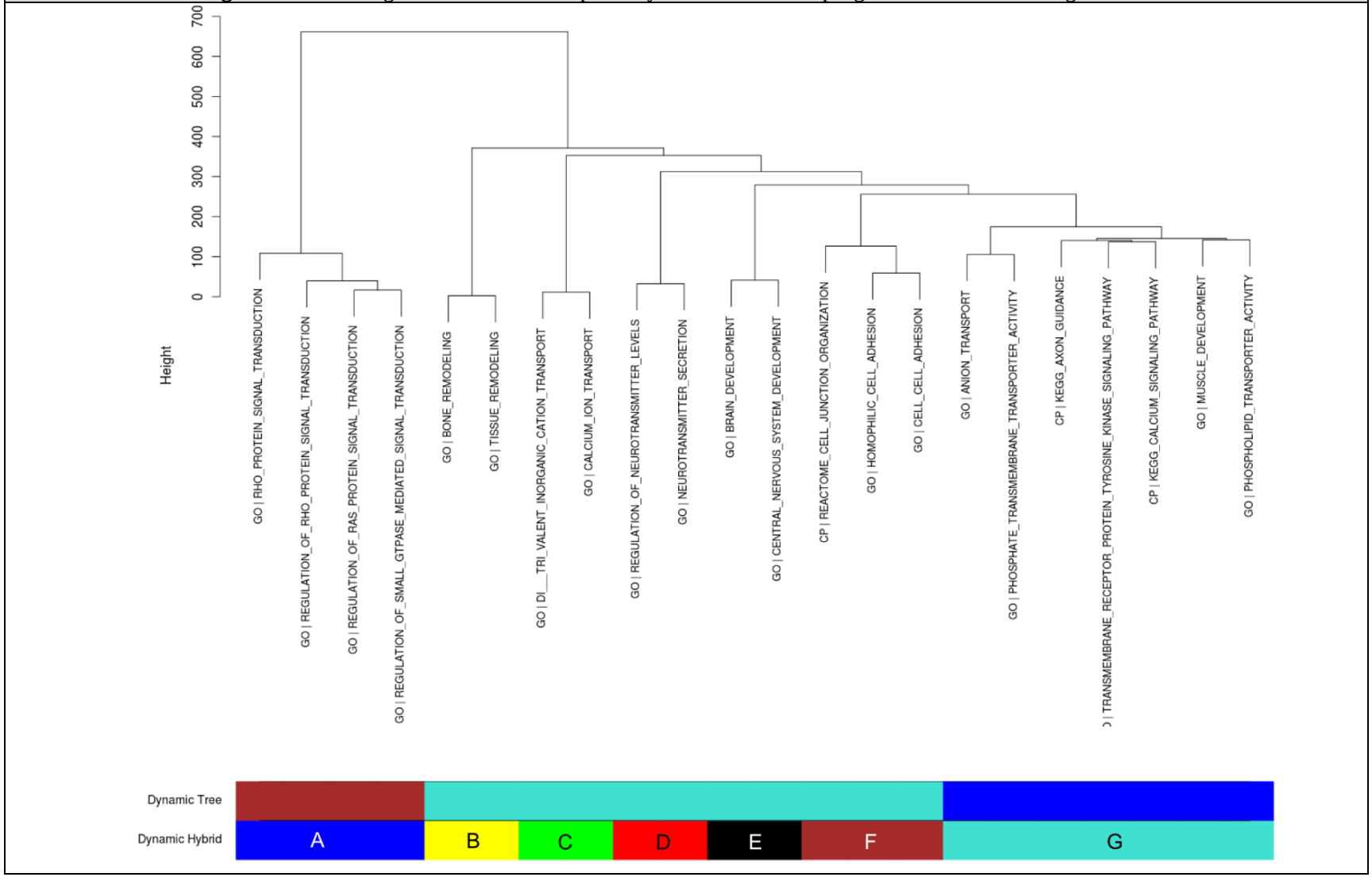


Figure 9d. GO categories and canonical pathways associated with progression in NMIBC using GeSBAP

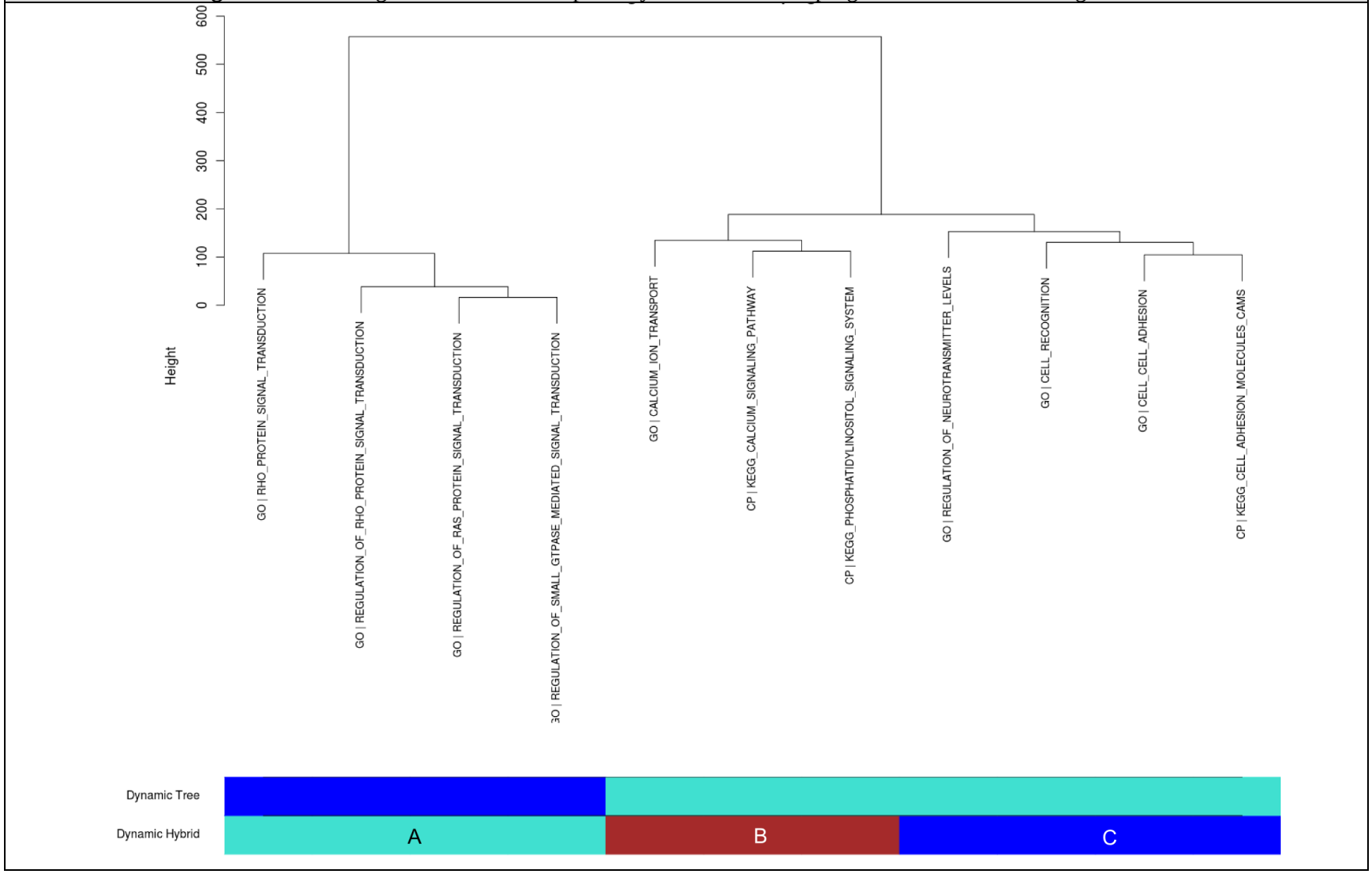


Figure 9e. GO categories and canonical pathways associated with recurrence in NMIBC using GeSBAP

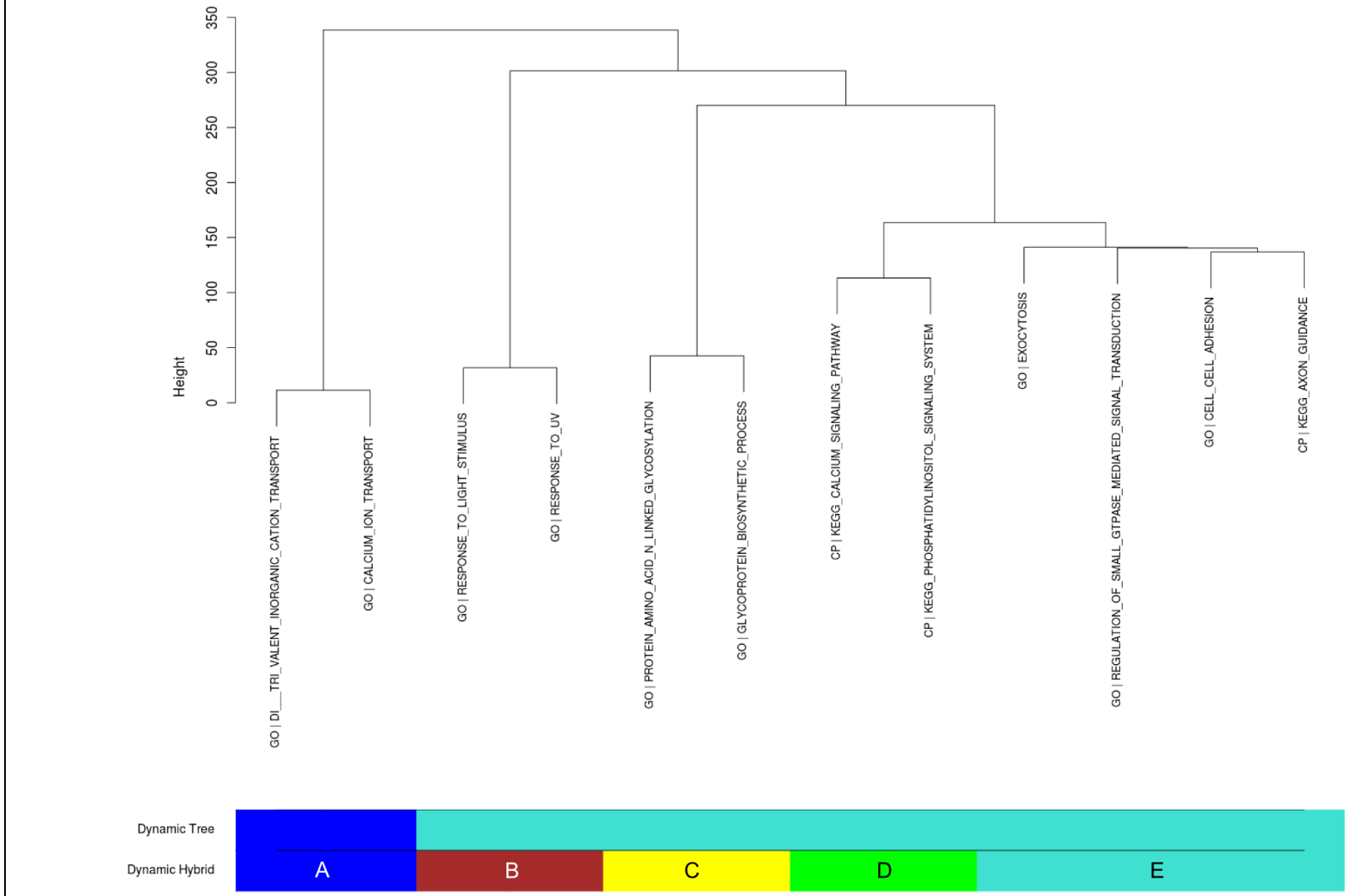


Figure 9f. GO categories and canonical pathways associated with BC-specific mortality in MIBC using GSA-SNP

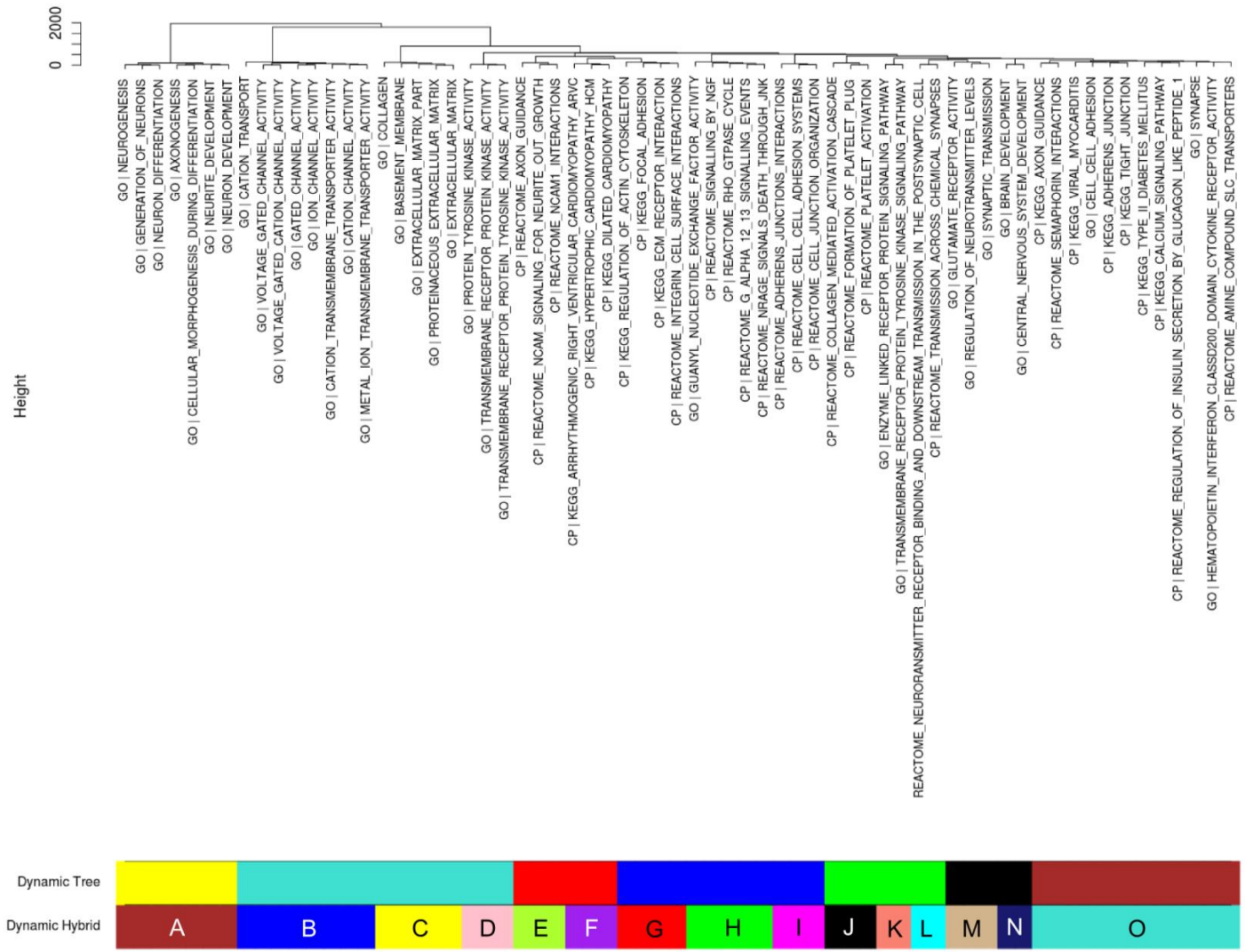


Figure 9g. GO categories and canonical pathways associated with progression in MIBC using GSA-SNP

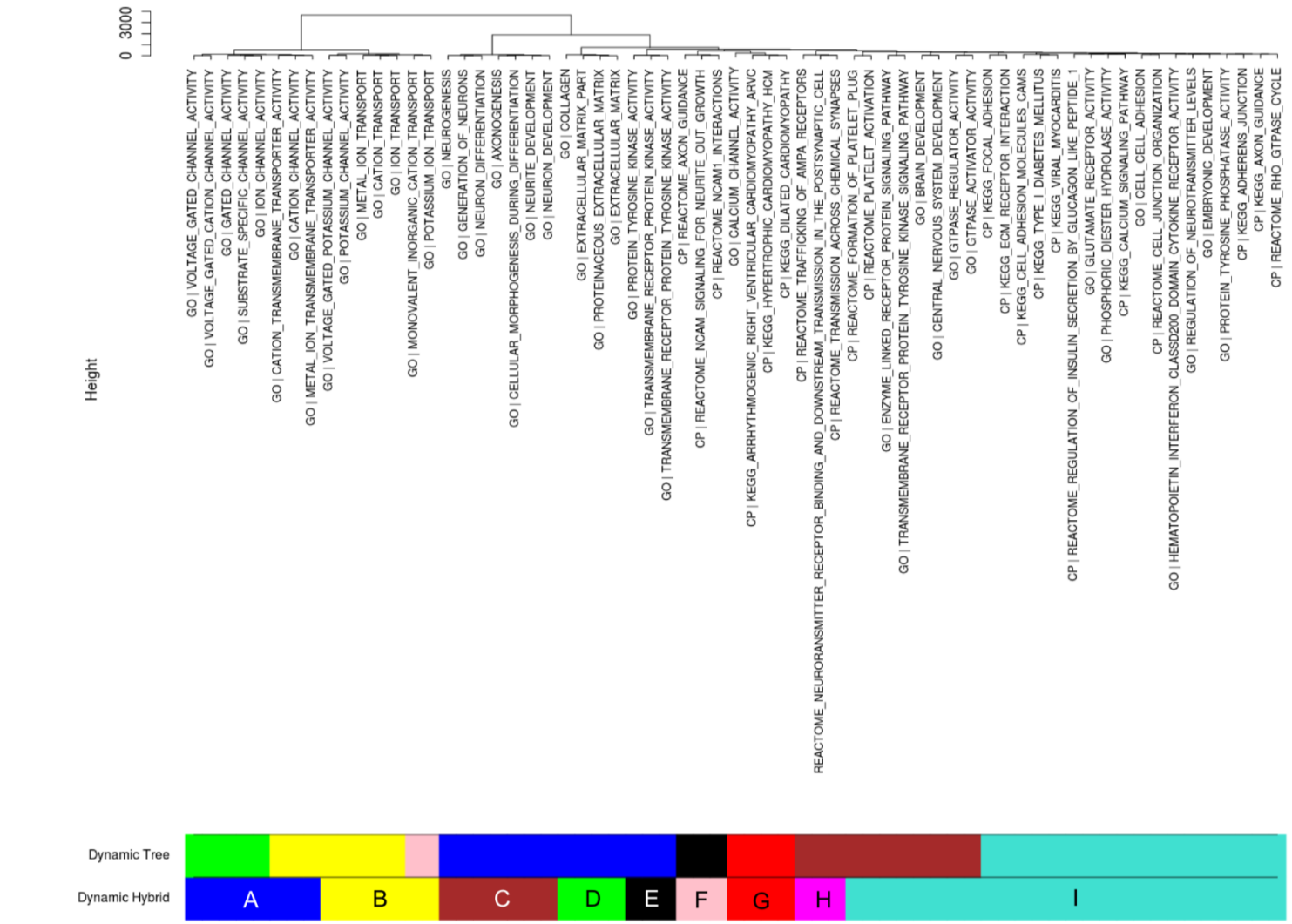


Figure 9h. GO categories and canonical pathways associated with progression in NMIBC using GSA-SNP

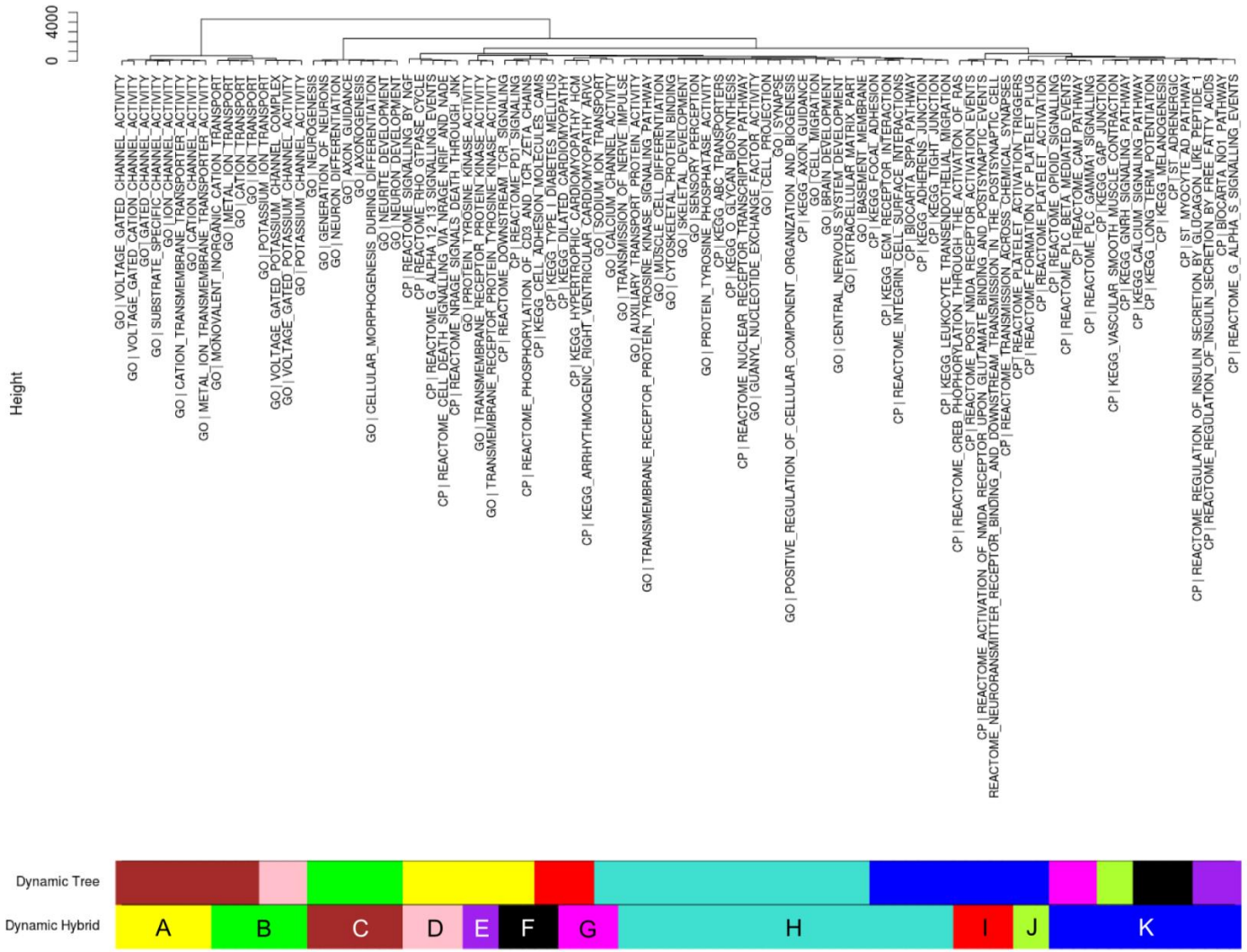


Figure 9i. GO categories and canonical pathways associated with recurrence in NMIBC using GSA-SNP

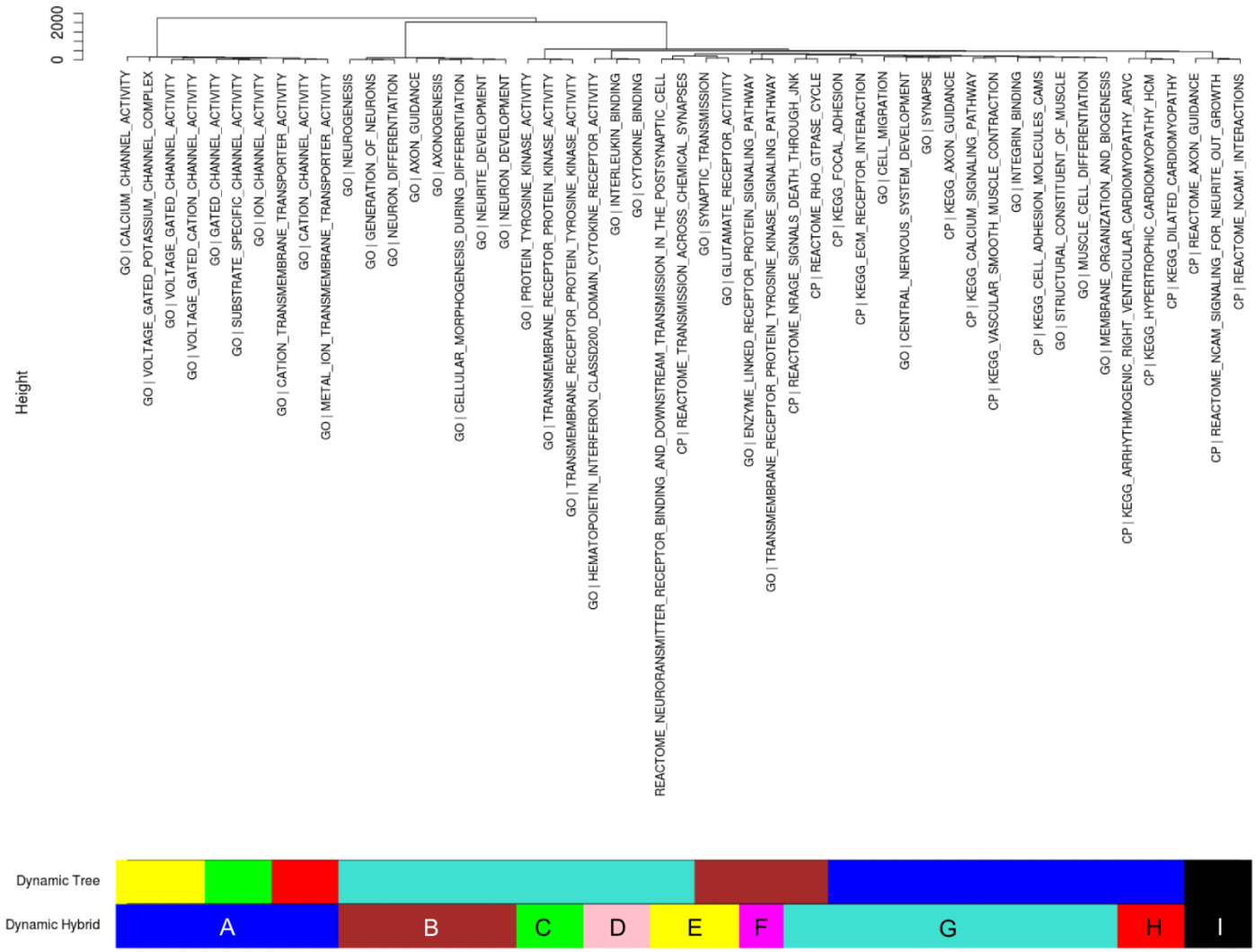


Figure 9j. GO categories and canonical pathways associated with BC-specific mortality in MIBC using ICSNPathway

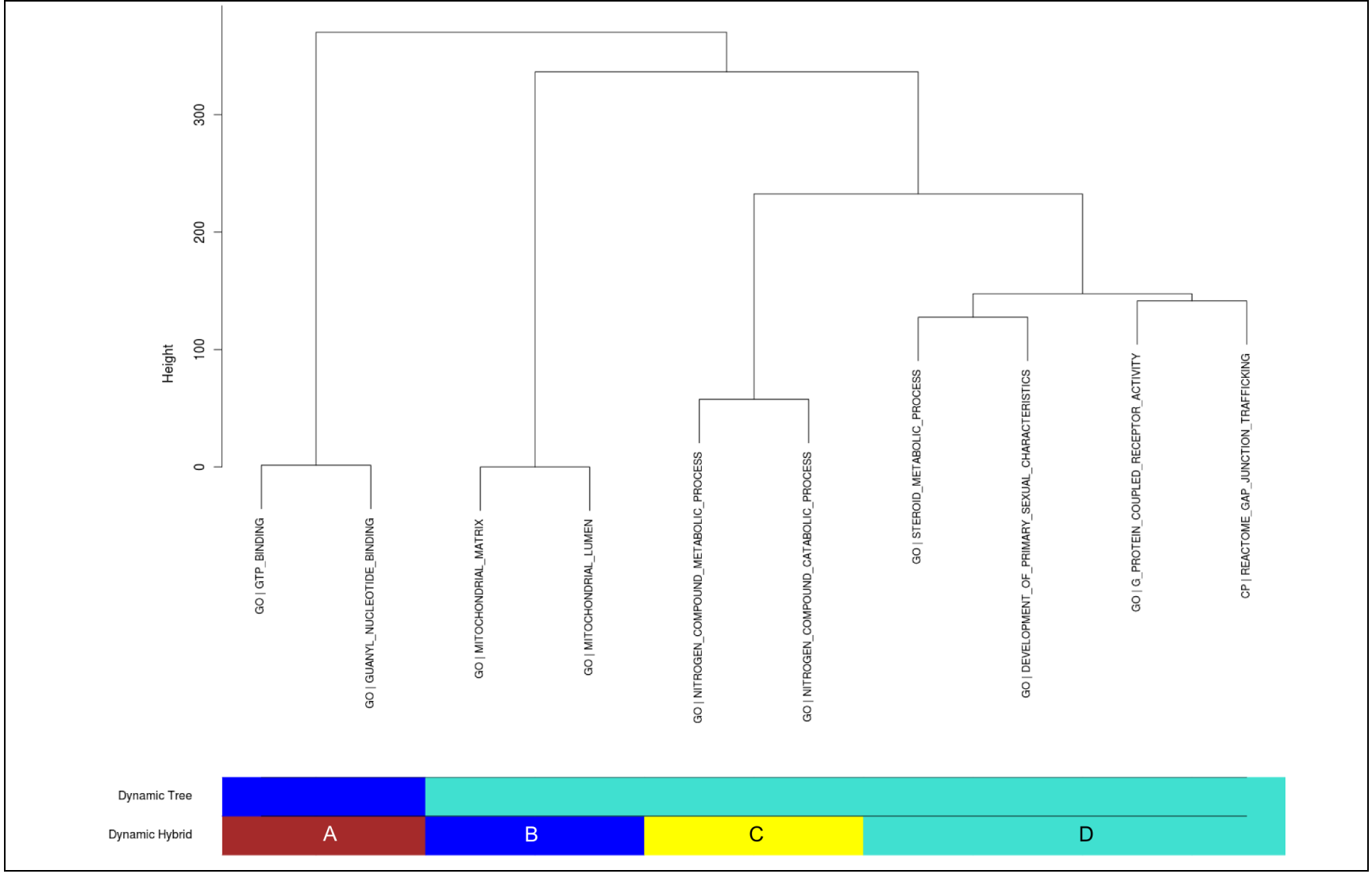


Figure 9k. GO categories and canonical pathways associated with progression in NMIBC using ICSNPathway

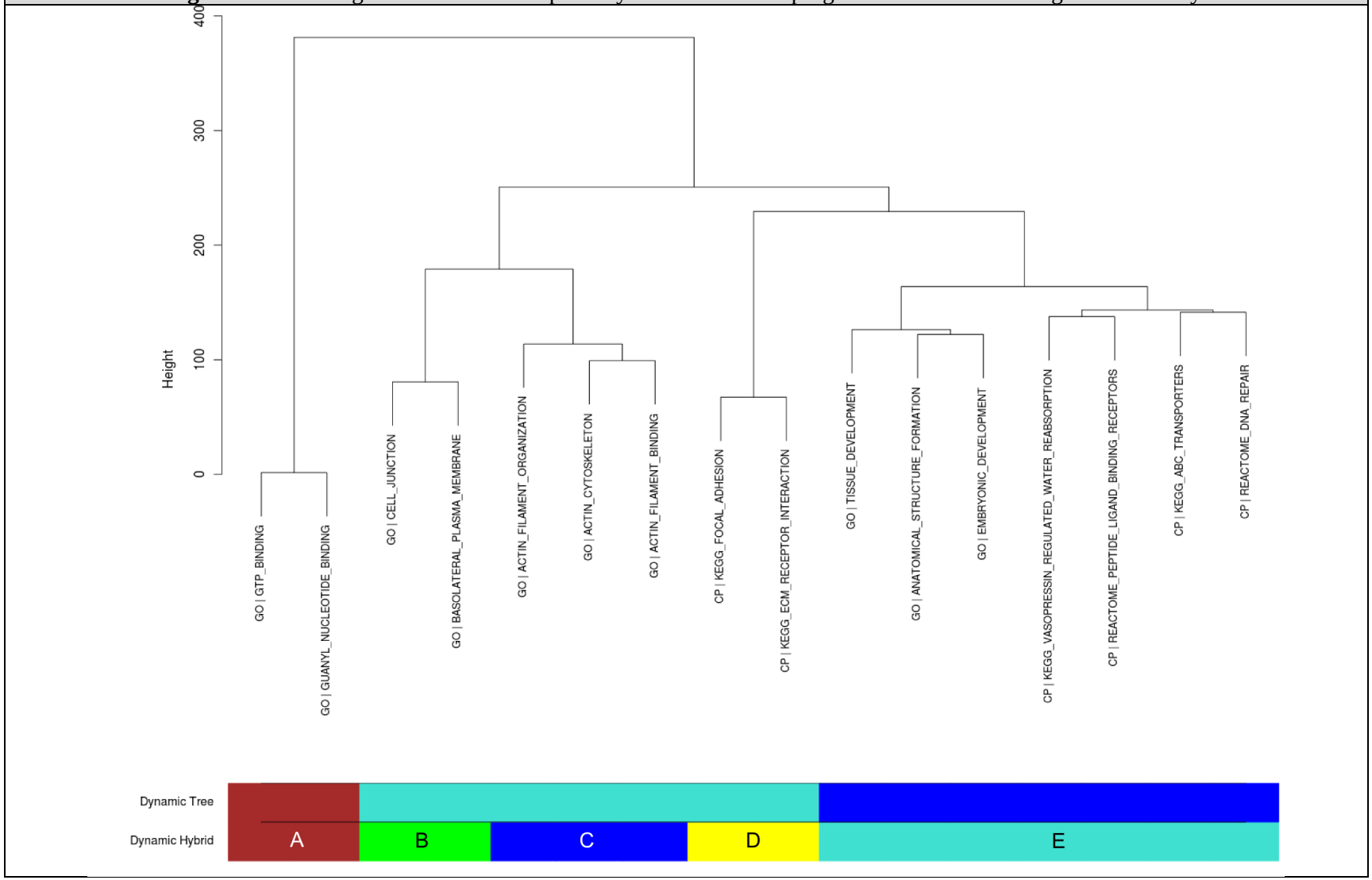


Figure 9l. GO categories and canonical pathways associated with recurrence in NMIBC using ICSNPathway

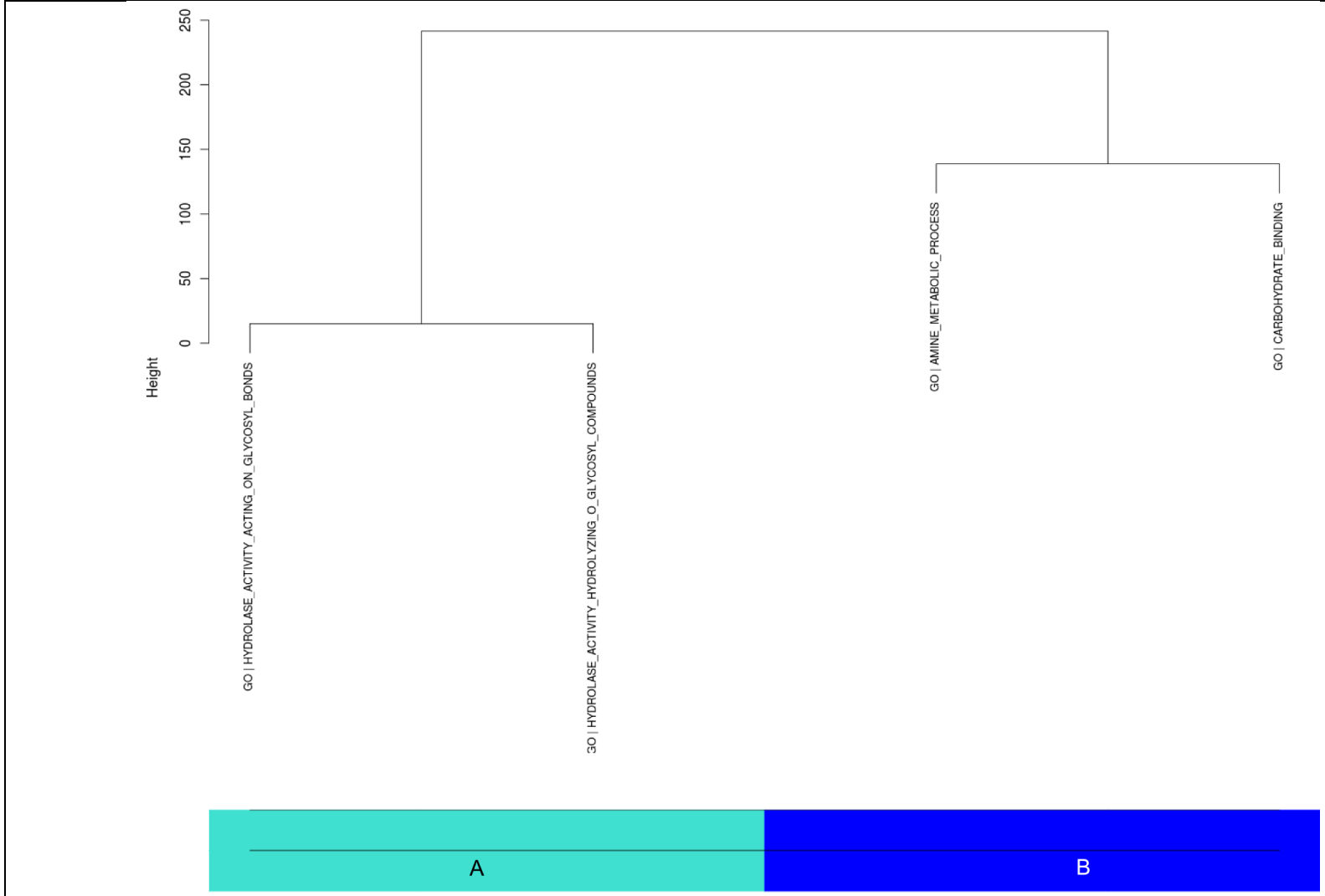


Figure 9m. GO categories and canonical pathways associated with progression in MIBC using *i*-Gsea*4*Gwas

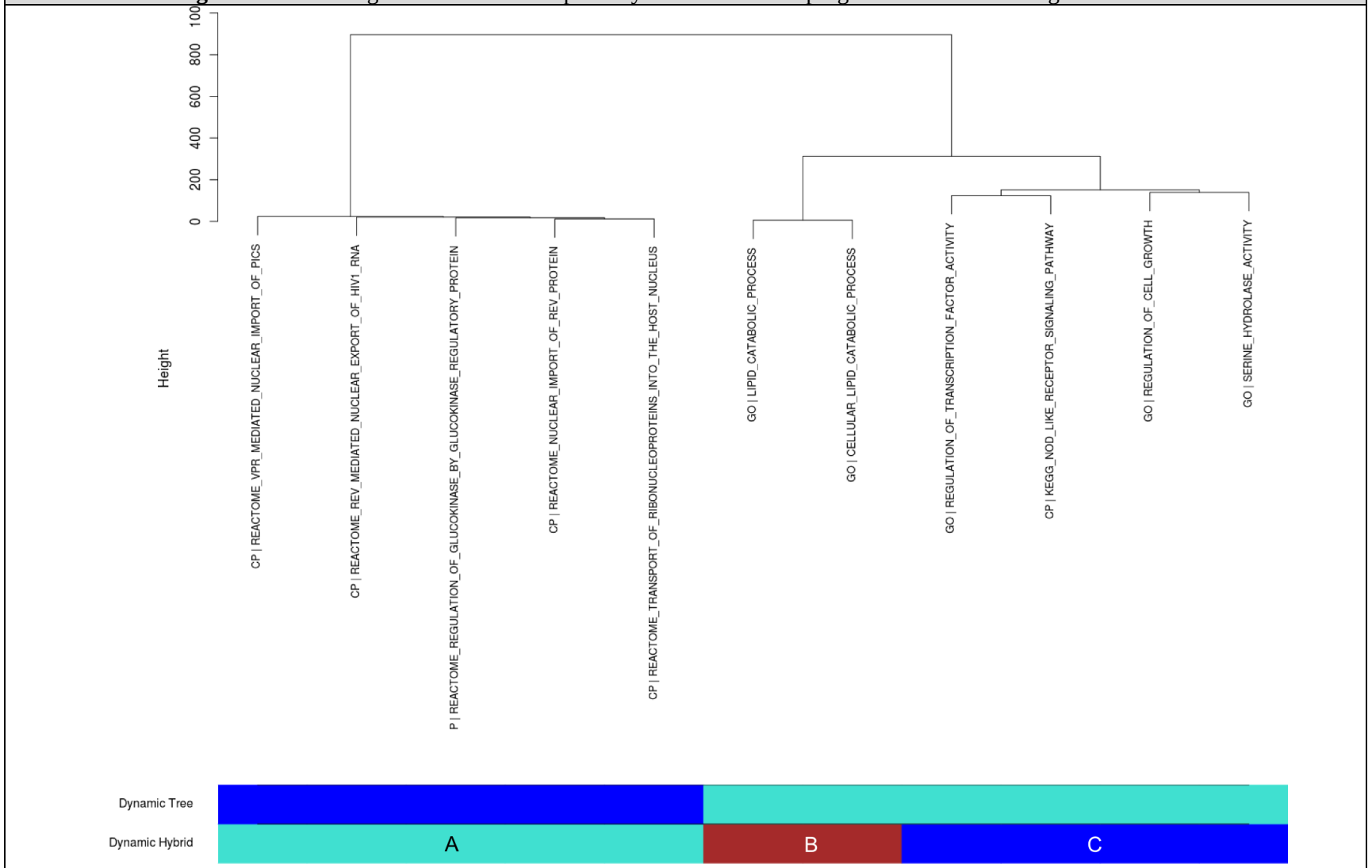


Figure 9n. GO categories and canonical pathways associated with progression in NMIBC using *i*-Gsea*4*Gwas

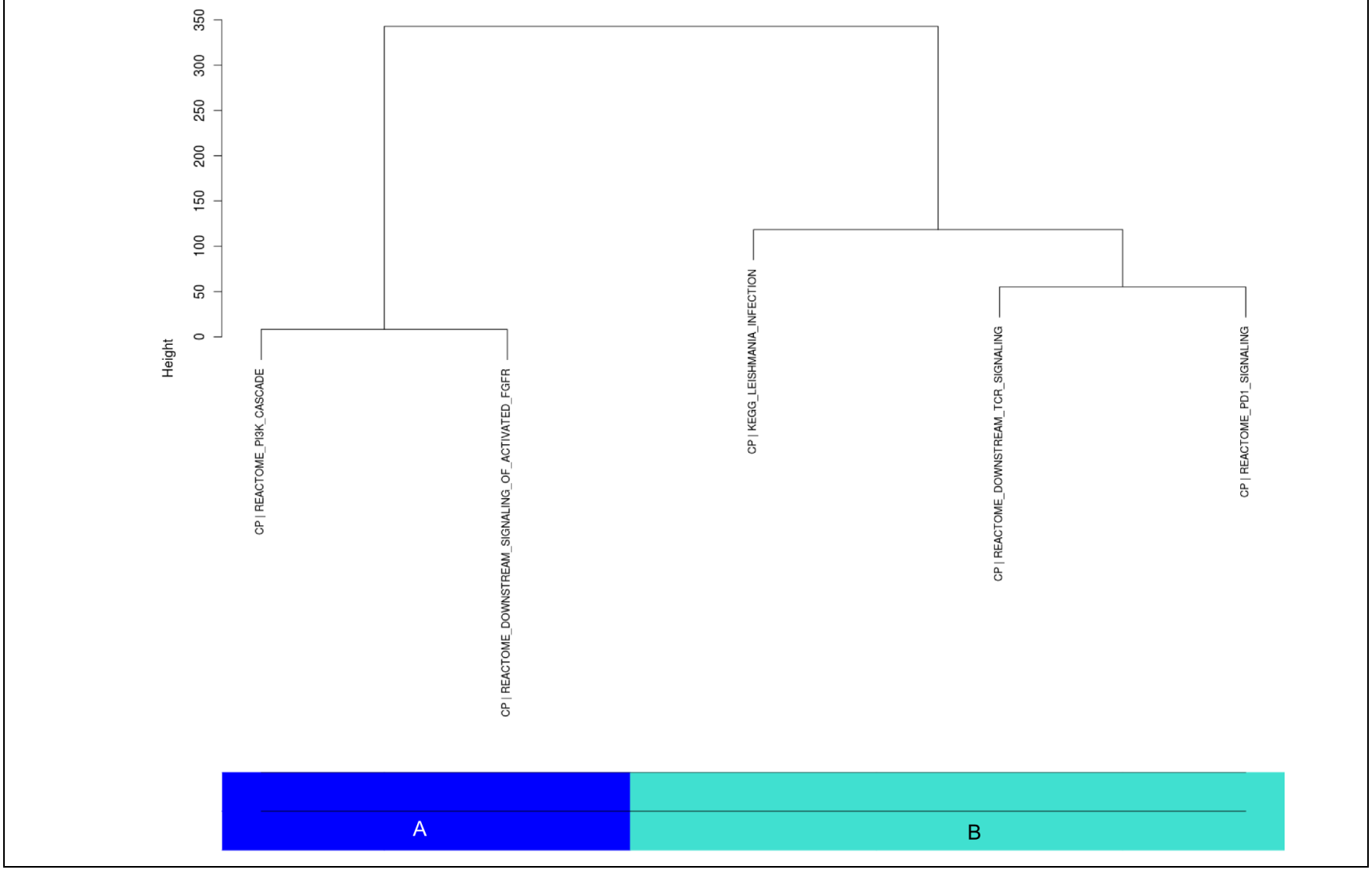
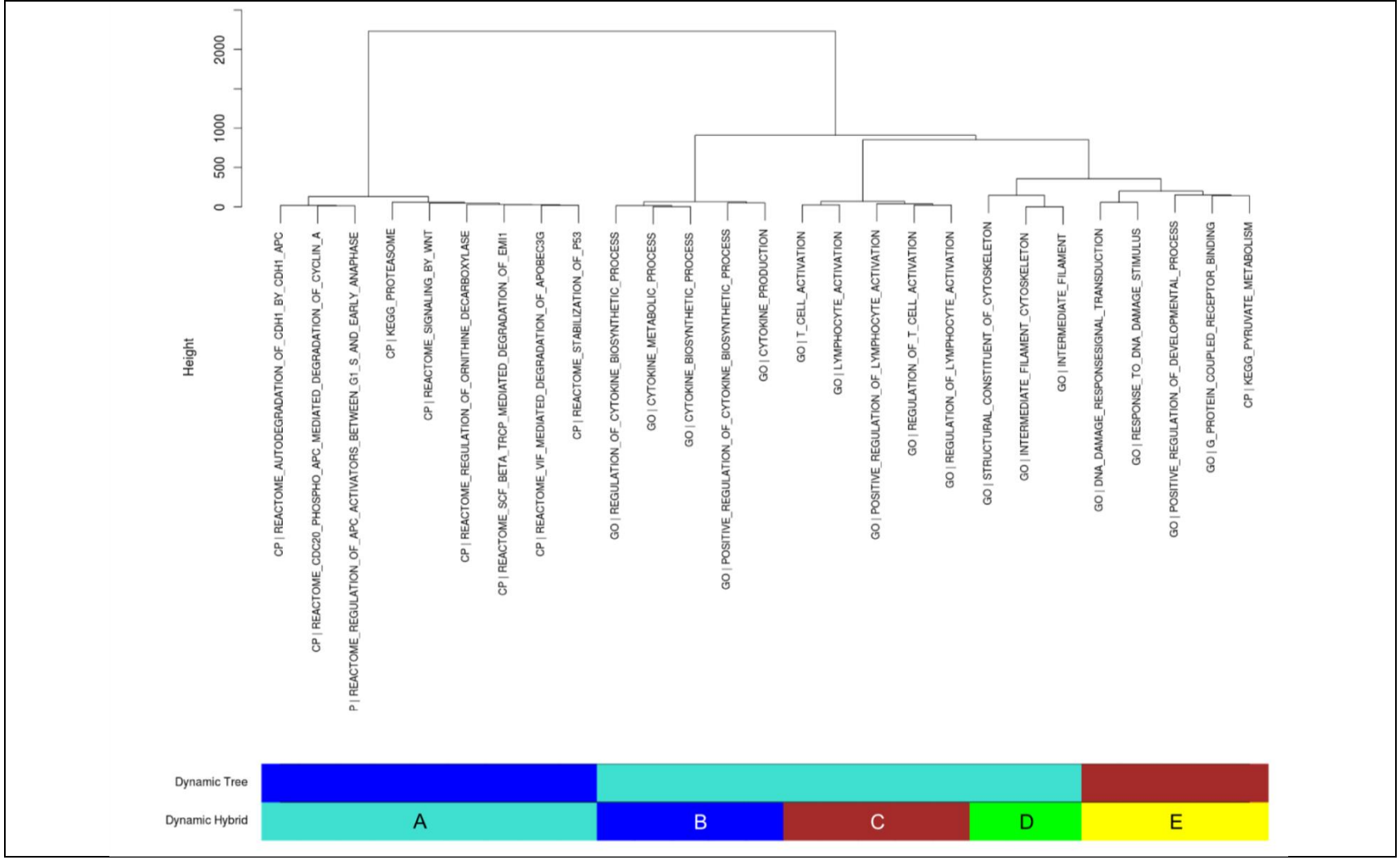


Figure 9o. GO categories and canonical pathways associated with recurrence in NMIBC using *i*-Gsea*4*Gwas



Supplementary Figure 10. Scatter plots comparing the -$\log_{10}$(*p-values*) obtained in the logistic and Cox regressions in NMIBC and MIBC clinical outcomes for 5M randomly selected SNP pairs.

Figure 10a. Recurrence in NMIBC

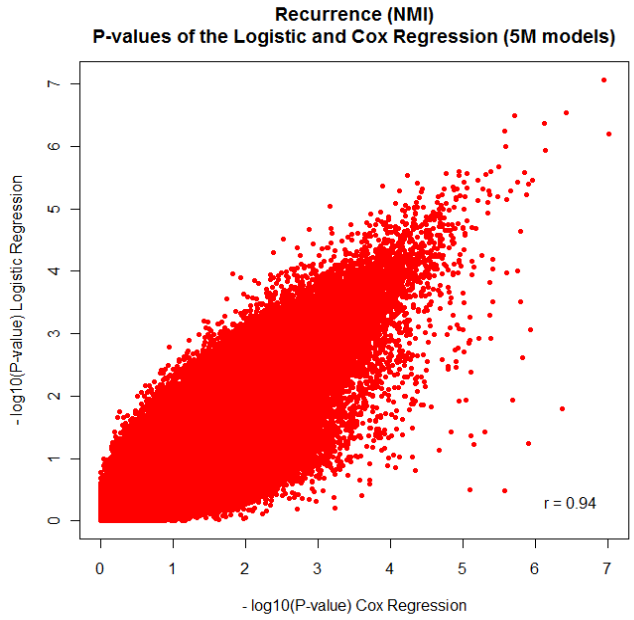


Figure 10b. Progression in NMIBC

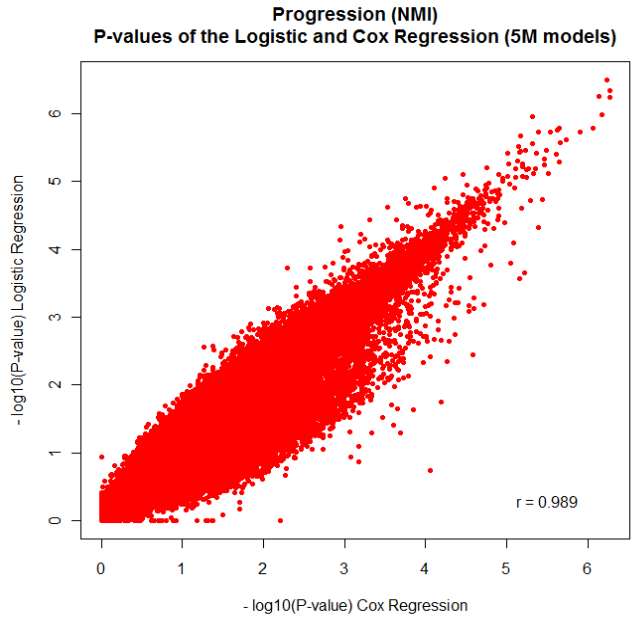


Figure 10c. Progression in MIBC

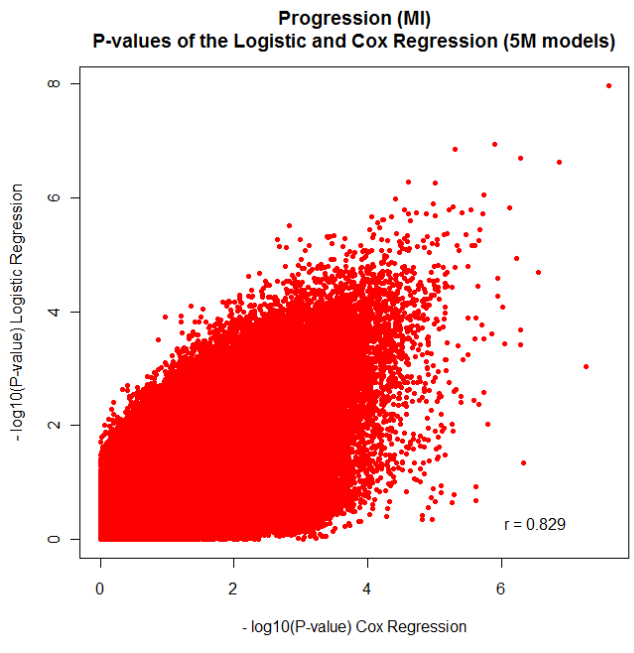
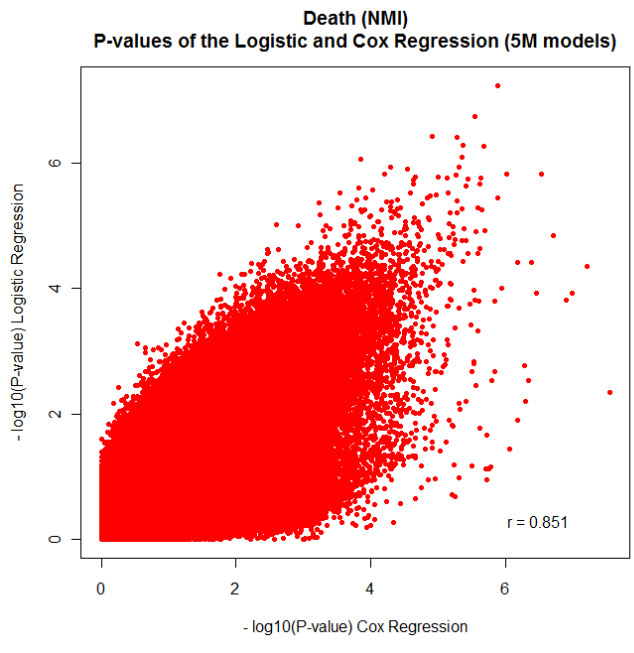


Figure 10d. BC-specific mortality in MIBC



References

Aben, K. K., L. Baglietto, et al. (2006). "Segregation analysis of urothelial cell carcinoma." Eur J Cancer **42**(10): 1428-33.

Aben, K. K., J. A. Witjes, et al. (2002). "Familial aggregation of urothelial cell carcinoma." Int J Cancer **98**(2): 274-8.

Ahirwar, D., P. Kesarwani, et al. (2008). "Anti- and proinflammatory cytokine gene polymorphism and genetic predisposition: association with smoking, tumor stage and grade, and bacillus Calmette-Guerin immunotherapy in bladder cancer." Cancer Genet Cytogenet **184**(1): 1-8.

Ahirwar, D. K., A. Mandhani, et al. (2009). "Association of tumour necrosis factor-alpha gene (T-1031C, C-863A, and C-857T) polymorphisms with bladder cancer susceptibility and outcome after bacille Calmette-Guerin immunotherapy." BJU Int **104**(6): 867-73.

Ahmadiyeh, N., M. M. Pomerantz, et al. (2010). "8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC." Proc Natl Acad Sci U S A **107**(21): 9742-6.

AJCC (1997). Urinary bladder: AJCC Cancer Staging Manual. Philadelphia, Lippincott-Raven.

Al-Shahrour, F., L. Arbiza, et al. (2007). "From genes to functional classes in the study of biological systems." BMC Bioinformatics **8**: 114.

Allard, P., P. Bernard, et al. (1998). "The early clinical course of primary Ta and T1 bladder cancer: a proposed prognostic index." Br J Urol **81**(5): 692-8.

Amara, N., G. S. Palapattu, et al. (2001). "Prostate stem cell antigen is overexpressed in human transitional cell carcinoma." Cancer Res **61**(12): 4660-5.

Andrew, A. S., J. Gui, et al. (2009). "Bladder cancer SNP panel predicts susceptibility and survival." Hum Genet **125**(5-6): 527-39.

Azzato, E. M., P. D. Pharoah, et al. (2010). "A genome-wide association study of prognosis in breast cancer." Cancer Epidemiol Biomarkers Prev **19**(4): 1140-3.

Babjuk, M., W. Oosterlinck, et al. (2011). "EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder, the 2011 update." Eur Urol **59**(6): 997-1008.

Barghorn, A., E. J. Speel, et al. (2001). "Putative tumor suppressor loci at 6q22 and 6q23-q24 are involved in the malignant progression of sporadic endocrine pancreatic tumors." Am J Pathol **158**(6): 1903-11.

Becker, T. and M. Knapp (2004). "A powerful strategy to account for multiple testing in the context of haplotype analysis." Am J Hum Genet **75**(4): 561-70.

Bell, D. A., J. A. Taylor, et al. (1993). "Genetic risk and carcinogen exposure: a common inherited defect of the carcinogen-metabolism gene glutathione S-transferase M1 (GSTM1) that increases susceptibility to bladder cancer." J Natl Cancer Inst **85**(14): 1159-64.

Bell, J. T., A. A. Pai, et al. (2011). "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines." Genome Biol **12**(1): R10.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological) **57**(1): 289-300.

Benvenisty, N., A. Leder, et al. (1992). "An embryonically expressed gene is a target for c-Myc regulation via the c-Myc-binding sequence." Genes Dev **6**(12B): 2513-23.

Bochner, B. H., M. W. Kattan, et al. (2006). "Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer." J Clin Oncol **24**(24): 3967-72.

Bolenz, C., S. F. Shariat, et al. (2010). "Human epidermal growth factor receptor 2 expression status provides independent prognostic information in patients with urothelial carcinoma of the urinary bladder." BJU Int **106**(8): 1216-22.

Botstein, D., R. L. White, et al. (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." Am J Hum Genet **32**(3): 314-31.

Breiman, L., J. H. Freidman, et al. (1984). Classification and regression trees. Belmont, California, Wadsworth.

Bruins, H. M. and J. P. Stein (2008). "Risk factors and clinical outcomes of patients with node-positive muscle-invasive bladder cancer." Expert Rev Anticancer Ther **8**(7): 1091-101.

Butler, M. A., N. P. Lang, et al. (1992). "Determination of CYP1A2 and NAT2 phenotypes in human populations by analysis of caffeine urinary metabolites." Pharmacogenetics **2**(3): 116-27.

Calle, M. L., V. Urrea, et al. (2008). "Improving strategies for detecting genetic patterns of disease susceptibility in association studies." Stat Med **27**(30): 6532-46.

Cantor, R. M., K. Lange, et al. (2010). "Prioritizing GWAS results: A review of statistical methods and recommendations for their application." Am J Hum Genet **86**(1): 6-22.

Castillejo, A., N. Rothman, et al. (2009). "TGFB1 and TGFBR1 polymorphic variants in relationship to bladder cancer risk and prognosis." Int J Cancer **124**(3): 608-13.

Castle, A. and D. Castle (2005). "Ubiquitously expressed secretory carrier membrane proteins (SCAMPs) 1-4 mark different pathways and exhibit limited constitutive trafficking to and from the cell surface." J Cell Sci **118**(Pt 16): 3769-80.

Collins, F. S., M. S. Guyer, et al. (1997). "Variations on a theme: cataloging human DNA sequence variation." Science **278**(5343): 1580-1.

Cooper, H. and L. V. Hedges (1994). The Handbook of Research Synthesis. Newbury Park, CA, Russell Sage Foundation.

Cordell, H. J. (2009). "Detecting gene-gene interactions that underlie human diseases." Nat Rev Genet **10**(6): 392-404.

Cox, D. R. (1972). "Regression Models and Life Tables." Journal of the Royal Statistical Society. Series B (*Methodological*) **34**: 187–220.

Chade, D. C., S. F. Shariat, et al. (2010). "Clinical outcomes of primary bladder carcinoma in situ in a contemporary series." J Urol **184**(1): 74-80.

Chang, D. W., J. Gu, et al. (2012). "Germline prognostic markers for urinary bladder cancer: obstacles and opportunities." Urol Oncol **30**(4): 524-32.

Chen, L. S., C. M. Hutter, et al. (2010). "Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data." Am J Hum Genet **86**(6): 860-71.

Chen, M., M. A. Hildebrandt, et al. (2010). "Genetic variations in the sonic hedgehog pathway affect clinical outcomes in non-muscle-invasive bladder cancer." Cancer Prev Res (Phila) **3**(10): 1235-45.

Daly, A. K. (2010). "Genome-wide association studies in pharmacogenomics." Nat Rev Genet **11**(4): 241-6.

de Bont, J. M., J. M. Kros, et al. (2008). "Differential expression and prognostic significance of SOX genes in pediatric medulloblastoma and ependymoma identified by microarray analysis." Neuro Oncol **10**(5): 648-60.

Dinney, C. P., D. J. McConkey, et al. (2004). "Focus on bladder cancer." Cancer Cell **6**(2): 111-6.

Donis-Keller, H., P. Green, et al. (1987). "A genetic linkage map of the human genome." Cell **51**(2): 319-37.

Efron, G. and R. Tibshirani (2007). "On testing the significance of sets of genes." Ann Appl Stat **1**: 107.

Eichler, E. E., J. Flint, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nat Rev Genet **11**(6): 446-50.

Elfiky, A. A. and J. E. Rosenberg (2009). "Targeting angiogenesis in bladder cancer." Curr Oncol Rep **11**(3): 244-9.

Elsamman, E., T. Fukumori, et al. (2006). "Prostate stem cell antigen predicts tumour recurrence in superficial transitional cell carcinoma of the urinary bladder." BJU Int **97**(6): 1202-7.

Fasching, P. A., P. D. Pharoah, et al. (2012). "The role of genetic breast cancer susceptibility variants as prognostic factors." Hum Mol Genet **21**(17): 3926-39.

Ferlay, J., H. R. Shin, et al. (2010). "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008." Int J Cancer **127**(12): 2893-917.

Fernandez-Gomez, J., R. Madero, et al. (2009). "Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model." J Urol **182**(5): 2195-203.

Fernandez-Gomez, J., E. Solsona, et al. (2008). "Prognostic factors in patients with non-muscle-invasive bladder cancer treated with bacillus Calmette-Guerin: multivariate analysis of data from four randomized CUETO trials." Eur Urol **53**(5): 992-1001.

Franceschini, A., D. Szklarczyk, et al. (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." Nucleic Acids Res **41**(Database issue): D808-15.

Fridley, B. L. and J. M. Biernacka (2011). "Gene set analysis of SNP data: benefits, challenges, and future directions." Eur J Hum Genet **19**(8): 837-43.

Fridley, B. L. and J. M. Biernacka (2011). "Gene set analysis of SNP data: benefits, challenges, and future directions." Eur J Hum Genet.

Frohlich, H., N. Speer, et al. (2007). "GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products." BMC Bioinformatics **8**: 166.

Fujita, A., A. Shida, et al. (2012). "Clinical significance of Rho GDP dissociation inhibitor 2 in colorectal carcinoma." Int J Clin Oncol **17**(2): 137-42.

Galindo, B. E. and V. D. Vacquier (2005). "Phylogeny of the TMEM16 protein family: some members are overexpressed in cancer." Int J Mol Med **16**(5): 919-24.

Gangawar, R., D. Ahirwar, et al. (2010). "Impact of nucleotide excision repair ERCC2 and base excision repair APEX1 genes polymorphism and its association with recurrence after adjuvant BCG immunotherapy in bladder cancer patients of North India." Med Oncol **27**(2): 159-66.

Gao, X., J. Starmer, et al. (2008). "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms." Genet Epidemiol **32**(4): 361-9.

Garcia-Closas, M., N. Malats, et al. (2005). "NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses." Lancet **366**(9486): 649-59.

Garcia-Closas, M., Y. Ye, et al. (2011). "A genome-wide association study of bladder cancer identifies a new susceptibility locus within SLC14A1, a urea transporter gene on chromosome 18q12.3." Hum Mol Genet **20**(21): 4282-9.

Gontero, P., A. Bohle, et al. (2010). "The role of bacillus Calmette-Guerin in the treatment of non-muscle-invasive bladder cancer." Eur Urol **57**(3): 410-29.

Gonzalez, J. R., J. L. Carrasco, et al. (2008). "Maximizing association statistics over genetic models." Genet Epidemiol **32**(3): 246-54.

Grotenhuis, A. J., S. H. Vermeulen, et al. (2010). "Germline genetic markers for urinary bladder cancer risk, prognosis and treatment response." Future Oncol **6**(9): 1433-60.

Gu, J. and X. Wu (2011). "Genetic susceptibility to bladder cancer risk and outcome." Per Med **8**(3): 365-374.

Gu, J., H. Zhao, et al. (2005). "Nucleotide excision repair gene polymorphisms and recurrence after treatment for superficial bladder cancer." Clin Cancer Res **11**(4): 1408-15.

Gur, S., P. J. Kadowitz, et al. (2011). "RhoA/Rho-kinase as a therapeutic target for the male urogenital tract." J Sex Med **8**(3): 675-87.

Hahn, N. M., W. M. Stadler, et al. (2011). "Phase II trial of cisplatin, gemcitabine, and bevacizumab as first-line therapy for metastatic urothelial carcinoma: Hoosier Oncology Group GU 04-75." J Clin Oncol **29**(12): 1525-30.

Hamid, J. S., P. Hu, et al. (2009). "Data integration in genetics and genomics: methods and challenges." Hum Genomics Proteomics **2009**.

Harell, F. E. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, Springer.

Hartzell, H. C., K. Yu, et al. (2009). "Anoctamin/TMEM16 family members are Ca2+-activated Cl- channels." J Physiol **587**(Pt 10): 2127-39.

Heinze, G. and D. Dunkler (2008). "Avoiding infinite estimates of time-dependent effects in small-sample survival studies." Stat Med **27**(30): 6455-69.

Herr, H. W., Z. Dotan, et al. (2007). "Defining optimal therapy for muscle invasive bladder cancer." J Urol **177**(2): 437-43.

Hindorff, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-7.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.

Hoffmann, P., T. Roumeguere, et al. (2006). "Use of statins and outcome of BCG treatment for bladder cancer." N Engl J Med **355**(25): 2705-7.

Holmans, P., E. K. Green, et al. (2009). "Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder." Am J Hum Genet **85**(1): 13-24.

Hong, M. G., Y. Pawitan, et al. (2009). "Strategies and issues in the detection of pathway enrichment in genome-wide association studies." Hum Genet **126**(2): 289-301.

Horikawa, Y., J. Nadaoka, et al. (2008). "Clinical implications of the MDM2 SNP309 and p53 Arg72Pro polymorphisms in transitional cell carcinoma of the bladder." Oncol Rep **20**(1): 49-55.

Horton, R., L. Wilming, et al. (2004). "Gene map of the extended human MHC." Nat Rev Genet **5**(12): 889-99.

Houwen, R. H., S. Baharloo, et al. (1994). "Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis." Nat Genet **8**(4): 380-6.

Huang, Y. T., R. S. Heist, et al. (2009). "Genome-wide analysis of survival in early-stage non-small-cell lung cancer." J Clin Oncol **27**(16): 2660-7.

Innocenti, F., K. Owzar, et al. (2012). "A genome-wide association study of overall survival in pancreatic cancer patients treated with gemcitabine in CALGB 80303." Clin Cancer Res **18**(2): 577-84.

Ioannidis, J. P., G. Thomas, et al. (2009). "Validating, augmenting and refining genome-wide association signals." Nat Rev Genet **10**(5): 318-29.

Ishwaran, H., U. B. Kogalur, et al. (2008). "Random survival forests." Ann. App. Statist. **2**: 841-860.

Jager, T., M. Becker, et al. (2010). "The prognostic value of cadherin switch in bladder cancer." Oncol Rep **23**(4): 1125-32.

Jemal, A., F. Bray, et al. (2011). "Global cancer statistics." CA Cancer J Clin **61**(2): 69-90.

Jeon, S. H. and S. G. Chang (2005). "Clinical prognostic factors for radical cystectomy in bladder cancer." Cancer Res Treat **37**(1): 48-53.

Jia, P., L. Wang, et al. (2010). "Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data." Schizophr Res **122**(1-3): 38-42.

Jia, P., L. Wang, et al. (2011). "Pathway-based analysis of GWAS datasets: effective but caution required." Int J Neuropsychopharmacol **14**(4): 567-72.

Ju, W., B. C. Yoo, et al. (2009). "Identification of genes with differential expression in chemoresistant epithelial ovarian cancer using high-density oligonucleotide microarrays." Oncol Res **18**(2-3): 47-56.

Kamai, T., T. Tsujii, et al. (2003). "Significant association of Rho/ROCK pathway with invasion and metastasis of bladder cancer." Clin Cancer Res **9**(7): 2632-41.

Kaplan, E. L. and P. Meier (1958). "Non parametric estimation from incomplete." J ournal of the American Statistical Association **53**: 457-481.

Karakiewicz, P. I., S. F. Shariat, et al. (2006). "Nomogram for predicting disease recurrence after radical cystectomy for transitional cell carcinoma of the bladder." J Urol **176**(4 Pt 1): 1354-61; discussion 1361-2.

Kawauchi, S., H. Sakai, et al. (2009). "9p21 index as estimated by dual-color fluorescence in situ hybridization is useful to predict urothelial carcinoma recurrence in bladder washing cytology." Hum Pathol **40**(12): 1783-9.

Kerem, B., J. M. Rommens, et al. (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-80.

Kiemeney, L. A., P. Sulem, et al. (2010). "A sequence variant at 4p16.3 confers susceptibility to urinary bladder cancer." Nat Genet **42**(5): 415-9.

Kiemeney, L. A., S. Thorlacius, et al. (2008). "Sequence variant on 8q24 confers susceptibility to urinary bladder cancer." Nat Genet **40**(11): 1307-12.

Kiemeney, L. A., J. A. Witjes, et al. (1994). "Dysplasia in normal-looking urothelium increases the risk of tumour progression in primary superficial bladder cancer." Eur J Cancer **30A**(11): 1621-5.

Kim, S. Y. and D. J. Volsky (2005). "PAGE: parametric analysis of gene set enrichment." BMC Bioinformatics **6**: 144.

Kiriluk, K. J., S. M. Prasad, et al. (2012). "Bladder cancer risk from occupational and environmental exposures." Urol Oncol **30**(2): 199-211.

Klein, R. J., C. Zeiss, et al. (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-9.

Koga, F., S. Kawakami, et al. (2003). "Impaired p63 expression associates with poor prognosis and uroplakin III expression in invasive urothelial carcinoma of the bladder." Clin Cancer Res **9**(15): 5501-7.

Kompier, L. C., I. Lurkin, et al. (2010). "FGFR3, HRAS, KRAS, NRAS and PIK3CA mutations in bladder cancer and their potential as biomarkers for surveillance and therapy." PLoS One **5**(11): e13821.

Kruger, S., F. Mess, et al. (2003). "Numerical aberrations of chromosome 17 and the 9p21 locus are independent predictors of tumor recurrence in non-invasive transitional cell carcinoma of the urinary bladder." Int J Oncol **23**(1): 41-8.

Kurth, K., L. Denis, et al. (1992). "The natural history and the prognosis of treated superficial bladder cancer. EORTC GU Group." Prog Clin Biol Res **378**: 1-7.

Kurth, K. H., L. Denis, et al. (1995). "Factors affecting recurrence and progression in superficial bladder tumours." Eur J Cancer **31A**(11): 1840-6.

Kwon, J. S., J. Kim, et al. (2012). "Performance Comparison of Two Gene Set Analysis Methods for Genome-wide Association Study Results: GSA-SNP vs i-GSEA4GWAS." Genomics Inform **10**(2): 123-7.

Lander, E. S. (1996). "The new genomics: global views of biology." Science **274**(5287): 536-9.

Lander, E. S. and D. Botstein (1986). "Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map." Cold Spring Harb Symp Quant Biol **51 Pt 1**: 49-62.

Langfelder, P., B. Zhang, et al. (2008). "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R." Bioinformatics **24**(5): 719-20.

Leibovici, D., H. B. Grossman, et al. (2005). "Polymorphisms in inflammation genes and bladder cancer: from initiation to recurrence, progression, and survival." J Clin Oncol **23**(24): 5746-56.

Lichtenstein, P., N. V. Holm, et al. (2000). "Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland." N Engl J Med **343**(2): 78-85.

Lin, D. (1998). An information-theoretic denition of similarity. Proceedings of the 15th International Conference on Machine Learning. M. Kaufmann. San Francisco, CA. **1:** 296-304.

Lin, J., C. P. Dinney, et al. (2006). "E-cadherin promoter polymorphism (C-160A) and risk of recurrence in patients with superficial bladder cancer." Clin Genet **70**(3): 240-5.

Lopez-Knowles, E., S. Hernandez, et al. (2006). "PIK3CA mutations are an early genetic alteration associated with FGFR3 mutations in superficial papillary bladder tumors." Cancer Res **66**(15): 7401-4.

Lotan, Y., U. Capitanio, et al. (2009). "Impact of clinical factors, including a point-of-care nuclear matrix protein-22 assay and cytology, on bladder cancer detection." BJU Int **103**(10): 1368-74.

Lotan, Y. and C. G. Roehrborn (2003). "Sensitivity and specificity of commonly available bladder tumor markers versus cytology: results of a comprehensive literature review and meta-analyses." Urology **61**(1): 109-18; discussion 118.

Loughman, N. T., B. P. Lin, et al. (2003). "DNA ploidy of bladder cancer using bladder biopsy supernate specimens." Anal Quant Cytol Histol **25**(3): 146-58.

Mahdy, E., Y. Pan, et al. (2001). "Chromosome 8 numerical aberration and C-MYC copy number gain in bladder cancer are linked to stage and grade." Anticancer Res **21**(5): 3167-73.

Malats, N. (2008). "Genetic epidemiology of bladder cancer: scaling up in the identification of low-penetrance genetic markers of bladder cancer risk and progression." Scand J Urol Nephrol Suppl(218): 131-40.

Manolio, T. A., F. S. Collins, et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-53.

Marenne, G., F. X. Real, et al. (2012). "Genome-wide CNV analysis replicates the association between GSTM1 deletion and bladder cancer: a support for using continuous measurement from SNP-array data." BMC Genomics **13**: 326.

Marenne, G., B. Rodriguez-Santiago, et al. (2011). "Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study." Hum Mutat **32**(2): 240-8.

Margulis, V., Y. Lotan, et al. (2008). "Predicting survival after radical cystectomy for bladder cancer." BJU Int **102**(1): 15-22.

Masood, E. (1999). "As consortium plans free SNP map of human genome." Nature **398**(6728): 545-6.

McPhail, E. R., M. E. Law, et al. (2011). "Influence of 6q22-23 on overall survival in primary central nervous system lymphoma. Analysis of North Central Cancer Treatment Group trials 86 72 52, 93 73 51 and 96 73 51." Br J Haematol **154**(1): 146-50.

Medina, I., J. Carbonell, et al. (2010). "Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling." Nucleic Acids Res **38**(Web Server issue): W210-3.

Medina, I., D. Montaner, et al. (2009). "Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies." Nucleic Acids Res **37**(Web Server issue): W340-4.

Menashe, I., J. D. Figueroa, et al. (2012). "Large-scale pathway-based analysis of bladder cancer genome-wide association data from five studies of European background." PLoS One **7**(1): e29396.

Menashe, I., D. Maeder, et al. (2010). "Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade." Cancer Res **70**(11): 4453-9.

Meyer, U. A. (2004). "Pharmacogenetics - five decades of therapeutic lessons from genetic diversity." Nat Rev Genet **5**(9): 669-76.

Mhawech-Fauceglia, P., L. Ali, et al. (2009). "Prognostic significance of neuron-associated protein expression in non-muscle-invasive urothelial bladder cancer." J Clin Pathol **62**(8): 710-4.

Millan-Rodriguez, F., G. Chechile-Toniolo, et al. (2000). "Multivariate analysis of the prognostic factors of primary superficial bladder cancer." J Urol **163**(1): 73-8.

Mitra, A. P., R. H. Datar, et al. (2006). "Molecular pathways in invasive bladder cancer: new insights into mechanisms, progression, and target identification." J Clin Oncol **24**(35): 5552-64.

Mittal, R. D., R. Singh, et al. (2008). "XRCC1 codon 399 mutant allele: a risk factor for recurrence of urothelial bladder carcinoma in patients on BCG immunotherapy." Cancer Biol Ther **7**(5): 645-50.

Miwa, S., T. Nakajima, et al. (2008). "Mutation assay of the novel gene DOG1 in gastrointestinal stromal tumors (GISTs)." J Gastroenterol **43**(7): 531-7.

Mizutani, Y., Y. Okada, et al. (1996). "Prognostic significance of circulating cytotoxic lymphocytes against autologous tumors in patients with bladder cancer." J Urol **155**(3): 888-92; discussion 892-4.

Mostofi, F. K., C. J. Davis, et al. (1999). Histological Typing of Urinary Bladder Tumours. World Health Organization International Classification of Histological Tumours. Berlin, Springer Verlag.

Mueller, C. M., N. Caporaso, et al. (2008). "Familial and genetic risk of transitional cell carcinoma of the urinary tract." Urol Oncol **26**(5): 451-64.

Muramaki, M., H. Miyake, et al. (2012). "Expression profile of E-cadherin and N-cadherin in non-muscle-invasive bladder cancer as a novel predictor of intravesical recurrence following transurethral resection." Urol Oncol **30**(2): 161-6.

Muramaki, M., H. Miyake, et al. (2011). "Expression profile of E-cadherin and N-cadherin in urothelial carcinoma of the upper urinary tract is associated with disease recurrence in patients undergoing nephroureterectomy." Urology **78**(6): 1443 e7-12.

Murta-Nascimento, C., D. T. Silverman, et al. (2007). "Risk of bladder cancer associated with family history of cancer: do low-penetrance polymorphisms account for the increase in risk?" Cancer Epidemiol Biomarkers Prev **16**(8): 1595-600.

Nadaoka, J., Y. Horikawa, et al. (2008). "Prognostic significance of HIF-1 alpha polymorphisms in transitional cell carcinoma of the bladder." Int J Cancer **122**(6): 1297-302.

Nakamura, M., M. Kishi, et al. (2003). "Novel tumor suppressor loci on 6q22-23 in primary central nervous system lymphomas." Cancer Res **63**(4): 737-41.

Nam, D., J. Kim, et al. (2010). "GSA-SNP: a general approach for gene set analysis of polymorphisms." Nucleic Acids Res **38**(Web Server issue): W749-54.

Netto, G. J. (2012). "Molecular biomarkers in urothelial carcinoma of the bladder: are we there yet?" Nat Rev Urol **9**(1): 41-51.

Oh, S., J. Lee, et al. (2012). "A novel method to identify high order gene-gene interactions in genome-wide association studies: gene-based MDR." BMC Bioinformatics **13 Suppl 9**: S5.

Oxford, G. and D. Theodorescu (2003). "The role of Ras superfamily proteins in bladder cancer progression." J Urol **170**(5): 1987-93.

Parmar, M. K., L. S. Freedman, et al. (1989). "Prognostic factors for recurrence and followup policies in the treatment of superficial bladder cancer: report from the

British Medical Research Council Subgroup on Superficial Bladder Cancer (Urological Cancer Working Party)." J Urol **142**(2 Pt 1): 284-8.

Penney, K. L., S. Pyne, et al. (2010). "Genome-wide association study of prostate cancer mortality." Cancer Epidemiol Biomarkers Prev **19**(11): 2869-76.

Pina, C., G. May, et al. (2008). "MLLT3 regulates early human erythroid and megakaryocytic cell fate." Cell Stem Cell **2**(3): 264-73.

Piriyapongsa, J., C. Ngamphiw, et al. (2012). "iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies." BMC Genomics **13 Suppl 7**: S2.

Pirmohamed, M. (2011). "Pharmacogenetics: past, present and future." Drug Discov Today **16**(19-20): 852-61.

Puente, D., N. Malats, et al. (2003). "Gender-related differences in clinical and pathological characteristics and therapy of bladder cancer." Eur Urol **43**(1): 53-62.

Puffenberger, E. G., E. R. Kauffman, et al. (1994). "Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22." Hum Mol Genet **3**(8): 1217-25.

Rafnar, T., P. Sulem, et al. (2009). "Sequence variants at the TERT-CLPTM1L locus associate with many cancer types." Nat Genet **41**(2): 221-7.

Rafnar, T., S. H. Vermeulen, et al. (2011). "European genome-wide association study identifies SLC14A1 as a new urinary bladder cancer susceptibility gene." Hum Mol Genet **20**(21): 4268-81.

Reis, S. T., K. R. Leite, et al. (2012). "Increased expression of MMP-9 and IL-8 are correlated with poor prognosis of Bladder Cancer." BMC Urol **12**: 18.

Riemann, K., H. Struwe, et al. (2008). "Characterization of intron-1 haplotypes of the G protein beta 4 subunit gene--association with survival and progression in patients with urothelial bladder carcinoma." Pharmacogenet Genomics **18**(11): 999-1008.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.

Ritchie, M. D., L. W. Hahn, et al. (2001). "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer." Am J Hum Genet **69**(1): 138-47.

Rodriguez, S., O. Jafer, et al. (2003). "Expression profile of genes from 12p in testicular germ cell tumors of adolescents and adults associated with i(12p) and amplification at 12p11.2-p12.1." Oncogene **22**(12): 1880-91.

Rothman, N., M. Garcia-Closas, et al. (2010). "A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci." Nat Genet **42**(11): 978-84.

Ruczinskie, I., C. Kooperberg, et al. (2003). "Logic regression." Journal of Computational and Graphical Statistics **12**: 475-511.

Saint, F., J. J. Patard, et al. (2002). "Prognostic value of a T helper 1 urinary cytokine response after intravesical bacillus Calmette-Guerin treatment for superficial bladder cancer." J Urol **167**(1): 364-7.

Samanic, C., M. Kogevinas, et al. (2006). "Smoking and bladder cancer in Spain: effects of tobacco type, timing, environmental tobacco smoke, and gender." Cancer Epidemiol Biomarkers Prev **15**(7): 1348-54.

Sanchez-Carbayo, M., N. D. Socci, et al. (2002). "Molecular profiling of bladder cancer using cDNA microarrays: defining histogenesis and biological phenotypes." Cancer Res **62**(23): 6973-80.

Sanyal, S., C. Ryk, et al. (2007). "Polymorphisms in NQO1 and the clinical course of urinary bladder neoplasms." Scand J Urol Nephrol **41**(3): 182-90.

Sato, Y., N. Yamamoto, et al. (2011). "Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel." J Thorac Oncol **6**(1): 132-8.

Shariat, S. F., P. I. Karakiewicz, et al. (2006). "Nomograms provide improved accuracy for predicting survival after radical cystectomy." Clin Cancer Res **12**(22): 6663-76.

Sharon, D., H. Yamamoto, et al. (2002). "Mutated alleles of the rod and cone Na-Ca+K-exchanger genes in patients with retinal diseases." Invest Ophthalmol Vis Sci **43**(6): 1971-9.

Shinka, T., A. Hirano, et al. (1990). "Clinical study of prognostic factors of superficial bladder cancer treated with intravesical bacillus Calmette-Guerin." Br J Urol **66**(1): 35-9.

Shinohara, A., S. Sakano, et al. (2009). "Association of TP53 and MDM2 polymorphisms with survival in bladder cancer patients treated with chemoradiotherapy." Cancer Sci **100**(12): 2376-82.

Silverman, D., S. Devesa, et al. (2006). Bladder cancer. Cancer epidemiology and prevention. D. Schottenfeld and J. J. Fraumeni. New York, Oxford University Press**:** 1101-27.

Silverman, D. T., P. Hartge, et al. (1992). "Epidemiology of bladder cancer." Hematol Oncol Clin North Am **6**(1): 1-30.

Smith, J. A., Jr., R. F. Labasky, et al. (1999). "Bladder cancer clinical guidelines panel summary report on the management of nonmuscle invasive bladder cancer (stages Ta, T1 and TIS). The American Urological Association." J Urol **162**(5): 1697-701.

Smith, S. C., A. S. Baras, et al. (2012). "Transcriptional signatures of Ral GTPase are associated with aggressive clinicopathologic characteristics in human cancer." Cancer Res **72**(14): 3480-91.

Stein, J. P., G. Lieskovsky, et al. (2001). "Radical cystectomy in the treatment of invasive bladder cancer: long-term results in 1,054 patients." J Clin Oncol **19**(3): 666-75.

Sternberg, C. N. (2002). "Current perspectives in muscle invasive bladder cancer." Eur J Cancer **38**(4): 460-7.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-50.

Sylvester, R. J. (2006). "Natural history, recurrence, and progression in superficial bladder cancer." ScientificWorldJournal **6**: 2617-25.

Sylvester, R. J., A. P. van der Meijden, et al. (2006). "Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials." Eur Urol **49**(3): 466-5; discussion 475-7.

Takeuchi, A., T. Dejima, et al. (2011). "IL-17 production by gammadelta T cells is important for the antitumor effect of Mycobacterium bovis bacillus Calmette-Guerin treatment against bladder cancer." Eur J Immunol **41**(1): 246-51.

Therneau, T. and P. Grambsch (2000). Modeling Survival Data: Extending the Cox Model, Springer-Verlag.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso. ." J. Royal. Statist. Soc B. **58**(1): 267-288.

Tickoo, S. K., M. I. Milowsky, et al. (2011). "Hypoxia-inducible factor and mammalian target of rapamycin pathway markers in urothelial carcinoma of the bladder: possible therapeutic implications." BJU Int **107**(5): 844-9.

Turkolmez, K., H. Tokgoz, et al. (2007). "Muscle-invasive bladder cancer: predictive factors and prognostic difference between primary and progressive tumors." Urology **70**(3): 477-81.

van Rhijn, B. W., M. Burger, et al. (2009). "Recurrence and progression of disease in non-muscle-invasive bladder cancer: from epidemiology to treatment strategy." Eur Urol **56**(3): 430-42.

van Rhijn, B. W., A. N. Vis, et al. (2003). "Molecular grading of urothelial cell carcinoma with fibroblast growth factor receptor 3 and MIB-1 is superior to pathologic grade for the prediction of clinical outcome." J Clin Oncol **21**(10): 1912-21.

Veyrieras, J. B., S. Kudaravalli, et al. (2008). "High-resolution mapping of expression-QTLs yields insight into human gene regulation." PLoS Genet **4**(10): e1000214.

Vineis, P. and R. Pirastu (1997). "Aromatic amines and cancer." Cancer Causes Control **8**(3): 346-55.

Wade, R., M. C. Di Bernardo, et al. (2011). "Association between single nucleotide polymorphism-genotype and outcome of patients with chronic lymphocytic leukemia in a randomized chemotherapy trial." Haematologica **96**(10): 1496-503.

Wan, X., C. Yang, et al. (2010). "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies." Am J Hum Genet **87**(3): 325-40.

Wang, K., M. Li, et al. (2007). "Pathway-based approaches for analysis of genomewide association studies." Am J Hum Genet **81**(6): 1278-83.

Wang, K., M. Li, et al. (2010). "Analysing biological pathways in genome-wide association studies." Nat Rev Genet **11**(12): 843-54.

Wang, L., P. Jia, et al. (2011). "Gene set analysis of genome-wide association studies: methodological issues and perspectives." Genomics **98**(1): 1-8.

Wang, M., M. Wang, et al. (2010). "A novel XPF -357A>C polymorphism predicts risk and recurrence of bladder cancer." Oncogene **29**(13): 1920-8.

Ward, J. (1963). "Hierarchical grouping to optimize an object function." Journal of the American Statistical Association **58**: 236-244.

Weber, J. L. and P. E. May (1989). "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction." Am J Hum Genet **44**(3): 388-96.

Weir, B. S. (1996). Genetic data analysis II. Massachusetts, Sinauer Associates.

Weissenbach, J., G. Gyapay, et al. (1992). "A second-generation linkage map of the human genome." Nature **359**(6398): 794-801.

Wendt, M. K., T. M. Allington, et al. (2009). "Mechanisms of the epithelial-mesenchymal transition by TGF-beta." Future Oncol **5**(8): 1145-68.

Wigginton, J. E., D. J. Cutler, et al. (2005). "A note on exact tests of Hardy-Weinberg equilibrium." Am J Hum Genet **76**(5): 887-93.

Wu, C., B. Xu, et al. (2010). "Genome-wide interrogation identifies YAP1 variants associated with survival of small-cell lung cancer patients." Cancer Res **70**(23): 9721-9.

Wu, X., J. Gu, et al. (2006). "Bladder cancer predisposition: a multigenic approach to DNA-repair and cell-cycle-control genes." Am J Hum Genet **78**(3): 464-79.

Wu, X., M. A. Hildebrandt, et al. (2009). "Genome-wide association studies of bladder cancer risk: a field synopsis of progress and potential applications." Cancer Metastasis Rev **28**(3-4): 269-80.

Wu, X., M. M. Ros, et al. (2008). "Epidemiology and genetic susceptibility to bladder cancer." BJU Int **102**(9 Pt B): 1207-15.

Wu, X., Y. Ye, et al. (2009). "Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer." Nat Genet **41**(9): 991-5.

Wu, X., Y. Ye, et al. (2011). "Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy." J Natl Cancer Inst **103**(10): 817-25.

Xie, M., J. Li, et al. (2012). "Detecting genome-wide epistases based on the clustering of relatively frequent items." Bioinformatics **28**(1): 5-12.

Yang, J. J., C. Cheng, et al. (2012). "Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia." Blood.

Yao, R., D. D. Davidson, et al. (2007). "The S100 proteins for screening and prognostic grading of bladder cancer." Histol Histopathol **22**(9): 1025-32.

Yates, D. R., I. Rehman, et al. (2007). "Promoter hypermethylation identifies progression risk in bladder cancer." Clin Cancer Res **13**(7): 2046-53.

Yoshikawa, R., H. Yanagi, et al. (2006). "ECA39 is a novel distant metastasis-related biomarker in colorectal cancer." World J Gastroenterol **12**(36): 5884-9.

Yu, K., Q. Li, et al. (2009). "Pathway analysis by adaptive combination of P-values." Genet Epidemiol **33**(8): 700-9.

Zaak, D., M. Burger, et al. (2010). "Predicting individual outcomes after radical cystectomy: an external validation of current nomograms." BJU Int **106**(3): 342-8.

Zafarana, G., A. J. Gillis, et al. (2002). "Coamplification of DAD-R, SOX5, and EKI1 in human testicular seminomas, with specific overexpression of DAD-R, correlates with reduced levels of apoptosis and earlier clinical manifestation." Cancer Res **62**(6): 1822-31.

Zeegers, M. P., F. E. Tan, et al. (2000). "The impact of characteristics of cigarette smoking on urinary tract cancer risk: a meta-analysis of epidemiologic studies." Cancer **89**(3): 630-9.

Zhang, K., S. Cui, et al. (2010). "i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study." Nucleic Acids Res **38**(Web Server issue): W90-5.

Zhang, K., S. Chang, et al. (2011). "ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework." Nucleic Acids Res **39**(Web Server issue): W437-43.

Zhang, Y. (2012). "A novel bayesian graphical model for genome-wide multi-SNP association mapping." Genet Epidemiol **36**(1): 36-47.

Zhao, H., D. Liang, et al. (2005). "Glutathione peroxidase 1 gene polymorphism and risk of recurrence in patients with superficial bladder cancer." Urology **66**(4): 769-74.

Zhong, W. D., Q. B. Chen, et al. (2010). "Extracellular matrix metalloproteinase inducer expression has an impact on survival in human bladder cancer." Cancer Epidemiol **34**(4): 478-82.

Zhou, W., X. Feng, et al. (2007). "Functional evidence for a nasopharyngeal carcinoma-related gene BCAT1 located at 12p12." Oncol Res **16**(9): 405-13.

Published papers

Research Article

Cancer Prevention Research

Genetic Variations in the Sonic Hedgehog Pathway Affect Clinical Outcomes in Non–Muscle-Invasive Bladder Cancer

Meng Chen[1], Michelle A.T. Hildebrandt[1], Jessica Clague[1], Ashish M. Kamat[2], Antoni Picornell[4], Joshua Chang[1], Xiaofan Zhang[1], Julie Izzo[3], Hushan Yang[1], Jie Lin[1], Jian Gu[1], Stephen Chanock[5], Manolis Kogevinas[6,7,8,9], Nathaniel Rothman[5], Debra T. Silverman[5], Montserrat Garcia-Closas[5], H. Barton Grossman[2], Colin P. Dinney[2], Núria Malats[4], and Xifeng Wu[1]

Abstract

Sonic hedgehog (Shh) pathway genetic variations may affect bladder cancer risk and clinical outcomes. Therefore, we genotyped 177 single-nucleotide polymorphisms (SNP) in 11 Shh pathway genes in a study including 803 bladder cancer cases and 803 controls. We assessed SNP associations with cancer risk and clinical outcomes in 419 cases of non–muscle-invasive bladder cancer (NMIBC) and 318 cases of muscle-invasive and metastatic bladder cancer (MiMBC). Only three SNPs (*GLI3* rs3823720, rs3735361, and rs10951671) reached nominal significance in association with risk ($P \leq 0.05$), which became nonsignificant after adjusting for multiple comparisons. Nine SNPs reached a nominally significant individual association with recurrence of NMIBC in patients who received transurethral resection (TUR) only ($P \leq 0.05$), of which two (*SHH* rs1233560 and *GLI2* rs11685068) were replicated independently in 356 TUR-only NMIBC patients, with P values of 1.0×10^{-3} (*SHH* rs1233560) and 1.3×10^{-3} (*GLI2* rs11685068). Nine SNPs also reached a nominally significant individual association with clinical outcome of NMIBC patients who received Bacillus Calmette-Guérin (BCG; $P \leq 0.05$), of which two, the independent *GLI3* variants rs6463089 and rs3801192, remained significant after adjusting for multiple comparisons ($P = 2 \times 10^{-4}$ and 9×10^{-4}, respectively). The wild-type genotype of either of these SNPs was associated with a lower recurrence rate and longer recurrence-free survival (versus the variants). Although three SNPs (*GLI2* rs735557, *GLI2* rs4848632, and *SHH* rs208684) showed nominal significance in association with overall survival in MiMBC patients ($P \leq 0.05$), none remained significant after multiple-comparison adjustments. Germ-line genetic variations in the Shh pathway predicted clinical outcomes of TUR and BCG for NMIBC patients. *Cancer Prev Res; 3(10); 1235–45. ©2010 AACR.*

Introduction

Malignant tumors of the bladder account for approximately 5% of all new primary cancers diagnosed in the United States, with an estimated 70,980 new cases in 2009 (1). Cigarette smoking is the most important etiologic factor for bladder cancer (2), but this disease is multifactorial and involves several environmental and genetic factors. Genetic polymorphisms in pathways controlling essential cellular activities may also play a role in bladder cancer etiology (3–5). Seventy percent to 80% of bladder cancers are non–muscle-invasive bladder cancer (NMIBC; ref. 6). Although NMIBC has an excellent prognosis and has a >80% overall 5-year survival rate, tumor recurrence is a major clinical problem and occurs in up to 70% of NMIBC patients after transurethral resection (TUR; ref. 7). Furthermore, 10% to 20% of such recurrences progress to invasive disease (7).

To reduce the recurrence and progression of NMIBC, intravesical instillation of agents is often administered after

Authors' Affiliations: Departments of [1]Epidemiology, [2]Urology, and [3]Experimental Therapeutics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas; [4]Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Center, Madrid, Spain; [5]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland; [6]Centre for Research in Environmental Epidemiology (CREAL); [7]Municipal Institute of Medical Research; [8]CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain; and [9]National School of Public Health, Athens, Greece

Note: Supplementary data for this article are available at Cancer Prevention Research Online (http://cancerprevres.aacrjournals.org/).

M. Chen and M. Hildebrandt are co-first authors and contributed equally to this work. N. Malats and X. Wu are co-senior authors and contributed equally to this work.

Corresponding Author: Xifeng Wu, Department of Epidemiology, Unit 1340, The University of Texas M.D. Anderson Cancer Center, 1155 Pressler Boulevard, Houston, TX 77030. Phone: 713-745-2485; Fax: 713-792-4657; E-mail: xwu@mdanderson.org.

doi: 10.1158/1940-6207.CAPR-10-0035

TUR (8). Bacillus Calmette-Guérin (BCG) is the major choice for intravesical therapy, which includes induction BCG (iBCG) in a 6-week cycle and maintenance BCG (mBCG) in 3-week cycles given 3, 6, 12, 18, 24, and 30 months after iBCG (9). Although the NMIBC rate of response to BCG is 60% to 70%, as many as one third of the patients who initially respond will still develop recurrence and progression (10). Furthermore, more than 90% of patients receiving BCG experience side effects such as fever, leukocyturia, and cystitis symptoms, and approximately 5% suffer severe toxicities including sepsis and even death (11). These toxic effects cause a large number of patients to discontinue treatment, especially mBCG (9). In addition, there is evidence that progression and death are less common in association with initial radical cystectomy than with radical cystectomy following failed BCG treatment (12). Therefore, early identification of patients who will fail BCG treatment or experience adverse side effects will not only reduce unnecessary impairment of their quality of life but will also aid physicians in selecting optimal, or personalizing, therapy. Traditional clinical variables have less prognostic value in patients treated with BCG than in patients receiving TUR only (13); therefore, biomarkers that can better predict response to BCG therapy are highly desired.

Cancer stem cells, or tumor-initiating cells, are the putative origin of cancer developing from normal stem or progenitor cells (14, 15). Cancer stem cells play important roles in driving recurrence or metastasis (16–19) and affecting treatment response (20–24). The sonic hedgehog (Shh) pathway is one of the major signaling pathways that regulate cancer stem cells; it also controls cell proliferation, differentiation, and tissue patterning during organ development. Normally inactivated in adult tissues, the Shh pathway is reactivated in a wide range of cancers (25). On activation of the pathway, the secreted ligand sonic hedgehog (SHH) binds to the Patched (PTCH) receptor and activates the transmembrane protein Smoothened (SMO). The activation of SMO initiates a downstream cascade that releases three transcriptions factors (GLI1, GLI2, and GLI3) from the cytoplasm to enter the nucleus to activate specific target genes (Fig. 1; refs. 26, 27).

Uncontrolled activation of the Shh pathway occurs in bladder and many other cancers (28, 29), and Shh signaling is involved in tumor growth, recurrence, metastasis, and stem cell survival and expansion (18). *PTCH1* has been investigated as a potential tumor suppressor in bladder cancer (30–32). The role of *PTCH1* and Shh signaling in bladder cancer risk, however, is still being debated (33, 34). To our knowledge, no previous studies have addressed the association of genetic variations in the Shh pathway with bladder cancer susceptibility and outcome.

In the current study, we determined whether genetic variations, or single-nucleotide polymorphisms (SNP), in core functional components of the Shh pathway were associated with bladder cancer risk. We also evaluated the role of these SNPs in modulating recurrence and the risk of progression in patients receiving or not receiving BCG in our study's NMIBC subpopulation of patients and survival in muscle-invasive and metastatic bladder cancer (MiMBC) patients.

Materials and Methods

Study subjects

The Texas Bladder Cancer Study (TXBCS) recruited bladder cancer cases from The University of Texas M.D.

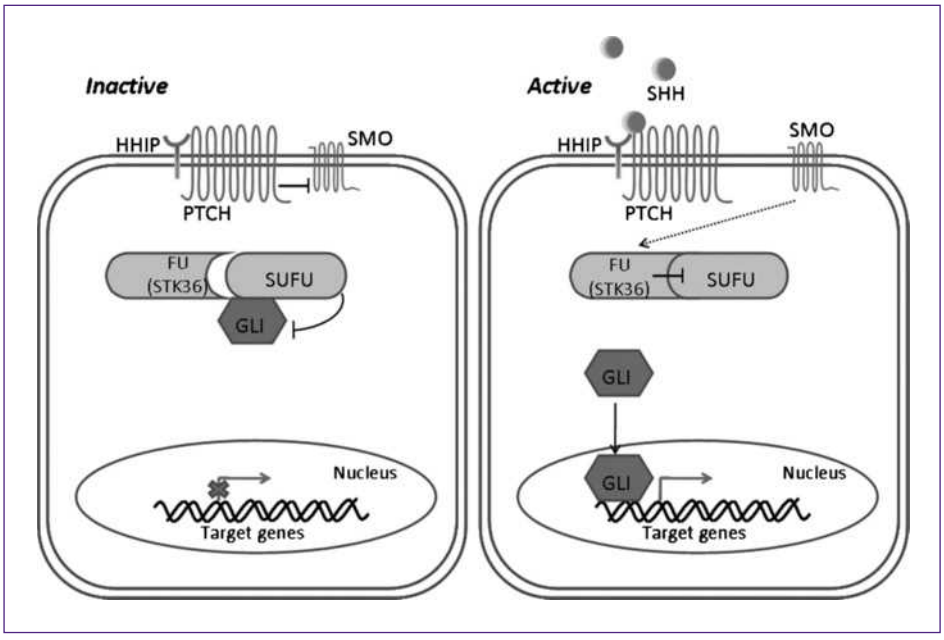


Fig. 1. Sonic hedgehog signaling pathway.

Anderson Cancer Center and Baylor College of Medicine through a daily review of computerized appointment schedules as a part of an ongoing project since 1995. Cases were all newly diagnosed within 1 year before recruitment, histologically confirmed, and previously untreated with chemotherapy or radiotherapy. Control subjects with no prior diagnosis of any type of cancer, except nonmelanoma skin cancer, were recruited from Kelsey Seybold, the largest private multispecialty physician group in Houston (35). These participants were matched 1:1 to the cases based on sex, age (±5 years), and ethnicity to evaluate the main effect of the genotype. There were no age, gender, or stage restrictions on recruitment. Because more than 90% of our recruited cases were pure transitional cell carcinoma and the etiology of transitional cell carcinoma differs from that of squamous cell carcinoma, we included patients with NMIBC and MiMBC in this study (Supplementary Table S1). In addition, because 90.6% of the patients in our capture population were Caucasians, we included only Caucasians in this study so as to limit the confounding effect of population structure. Individuals who never smoked or had smoked less than 100 cigarettes in his or her lifetime were defined as never smokers. Cases who had quit smoking at least 1 year before diagnosis and controls who had quit smoking at least 1 year before the interview were defined as former smokers. Individuals who were currently smoking or who had stopped <1 year before being diagnosed (cases) or before interview (controls) were defined as current smokers. Current and former smokers were defined as ever smokers.

An independent validation set for the TXBCS NMIBC patient data was obtained from the Spanish Bladder Cancer (SBC)/Epidemiology of Cancer of the Urothelium (EPICURO) study. All incident NMIBC patients were treated during 1998-2001 in 18 general or university-affiliated hospitals located in five geographic areas of Spain. The replication study included NMIBC patients who received TUR only and excluded NMIBC patients who received BCG mainly because of substantial differences in BCG regimens between the TXBCS and SBC/EPICURO study.

Epidemiologic and clinical data collection

Epidemiologic data of the TXBCS were collected by M.D. Anderson interviewers in a 45-minute interview on demographics, family history, and smoking status. Immediately after the interview, a blood sample was collected for DNA extraction. The clinical data for TXBCS such as tumor size, grade, stage, presence of carcinoma *in situ*, number of tumor foci at diagnosis, intravesical therapy, dates of recurrence and progression events, systemic chemotherapy, radical cystectomy, pathologic findings at cystectomy, and mortality were collected by trained chart reviewers. All patients were followed up with periodic cystoscopic examinations. The end points of outcome assessment in this study included recurrence, defined as a newly found bladder tumor following a previous negative follow-up cystoscopy; progression, defined as the transition from non–muscle-invasive to invasive or metastatic tumors; and overall survival, which was calculated from the date of diagnosis to the date of death or last follow-up, whichever came first. All of the human participation procedures were approved by the University of Texas M.D. Anderson Cancer Center, Baylor College of Medicine, and Kelsey Seybold institutional review boards. Written informed consent was obtained from all patients before interview. Clinical data collection in the SBC/EPICURO study has been described in detail previously (36). Written informed consent was obtained from all participants, and the study was approved by the local institutional ethics committee of each participating hospital and by the institutional review boards of the Institut Municipal d'Investigació Mèdica and U.S. National Cancer Institute.

Genotyping

Genotyping for the TXBCS was done at M.D. Anderson Cancer Center. Laboratory personnel were blinded to case and control status. Genomic DNA was isolated from peripheral blood using the QIAamp DNA Blood Maxi Kit (Qiagen) according to the manufacturer's protocol. We combined literature search and database mining to select candidate genes in the Shh pathway following a procedure as previously described (4). A total of 177 haplotype-tagging SNPs from 11 Shh pathway genes, including *GLI1*, *GLI2*, *GLI3*, *GLI4*, *HHIP* (*Hedgehog-interacting protein*), *STK36*, *SUFU*, *SHH*, *SMO*, *PTCH*, and *PTCH2*, were selected for genotyping. The genotyping of 150 Shh SNPs in *GLI2*, *GLI3*, *GLI4*, *SUFU*, *PTCH*, *PTCH2*, *SMO*, and *SHH* was done using the Illumina iSelect custom SNP array platform, and 27 Shh SNPs in *GLI1*, *STK36*, and *HHIP* were obtained from our published genome-wide association study using the Illumina Human-Hap610 BeadChips (37), according to the manufacturer's Infinium II assay protocol (Illumina), with 750 ng of input DNA for each sample. All the genotyping data were analyzed and exported using BeadStudio software (Illumina). The average call rate for the SNP array was 99.7%. SNPs selected for replication were genotyped with the Infinium Illumina Human 1M probe BeadChip in SBC/EPICURO patients (38).

Statistical analysis

Most statistical analyses were done using the Intercooled Stata 10 statistical software package (Stata). Pearson's χ^2 test or Fisher's exact test was used to compare the difference in distribution of categorical variables, and Wilcoxon rank sum test or Student's *t* test was used for continuous variables where appropriate. Hardy-Weinberg equilibrium was tested using a goodness-of-fit χ^2 analysis. The effects of genotypes of SNPs on bladder cancer risks were estimated as odds ratios and 95% confidence intervals (95% CI) using unconditional multivariate logistic regression under the dominant, recessive, and additive models of inheritance adjusted for age, gender, and smoking status, where appropriate. For clinical outcome analyses, the main effect of individual SNPs on time to the event of each end point, hazard ratios (HR), and 95% CIs were estimated by multivariate Cox proportional hazard regression, adjusting

for age, gender, smoking status, tumor grade, tumor stage, and treatments. The patients who were lost to follow-up or died before the end point were censored. Because many SNPs and tests were done in the analysis, the *Q* value (a false discovery rate adjusted *P* value) was used to adjust the significance level for individual SNPs (39–41). We calculated the *Q* value by the Q value package implemented in the R software. We applied a bootstrap resampling method to internally validate the results. We generated 100 bootstrapped samples for SNPs that remained significant after multiple comparison. Each bootstrap sample was drawn from the original data set, and a *P* value was obtained for each SNP in the dominant, recessive, and additive models. Stratified analysis was used to compare the effects of individual genotypes on different treatment subgroups. In the replication study in SBC/EPICURO patients, HRs and 95% CIs were estimated by a multivariate Cox proportional hazard model, with adjustments for area, sex, stage, T-stage and grade, multiplicity, tumor size, and treatment. The individual effects of all SNPs on recurrence in TUR-only NMIBC patients in the combined TXBCS and SBC/EPICURO were summarized in a meta-analysis. All statistical analyses were two-sided. Kaplan-Meier plots and log-rank tests were applied to compare the difference between the recurrence-free survival time of homozygous wild-type and variant genotypes, which was calculated from the diagnosis date to the end of the follow-up or recurrence.

Results

Subject characteristics

A total of 803 Caucasian patients with transitional cell carcinoma of bladder cancer and 803 Caucasian controls were included in this study (Supplementary Table S1). Cases and controls were perfectly matched on sex (P = 1.00) and no significant difference was observed for cases (63.8 ± 10.9 years) and controls (64.7 ± 11.1 years) on age (P = 0.10). As we predicted, cases were more likely to be current smokers (23.3%) than controls (8.3%, $P < 0.01$), and among ever smoking participants, cases had a significantly higher mean pack-years (43.0 ± 30.7) than did the controls (29.9 ± 27.9; $P < 0.01$).

There were 419 NMIBC patients and 318 MiMBC patients with full follow-up data among the 803 cases from the case-control study. Of 419 NMIBC patients, 228 cases developed a recurrence. Table 1 shows the distribution of demographic and clinical variables in TXBCS study. The percentage of male patients with recurrence (56.8%) was significantly higher than that of females (43.4%; P = 0.03). There were no statistical differences between the recurrence and nonrecurrence groups in smoking status and clinical factors (tumor stage and grade) except for treatment. We categorized the 419 NMIBC patients into the four following treatment subgroups: TUR only, iBCG (received after TUR), mBCG (received after the TUR and iBCG), and others (such as intravesical chemotherapy but no BCG). Patients receiving mBCG were less likely to develop recurrence than those without mBCG ($P < 0.01$). Among these patients, 71 had progression. Factors associated with progression included sex, age, stage, grade, and treatment. Male patients (18.7%) were more likely to progress than women (9.2%; P = 0.05) and patients who had progression were significantly older at diagnosis (mean age, 66.2 years) than patients without progression (mean age, 62.7 years; P = 0.02). Higher stage and grade are significant risk factors for progression. Patients receiving mBCG treatment were less likely to progress ($P < 0.01$). Because BCG is primarily administered to those with higher risk of recurrence, we compared the stages and grades of NMIBC patients who received TUR only or any type of BCG (iBCG and mBCG). As expected, patients receiving BCG had higher stage and grade than the TUR-only subgroup ($P < 0.001$; data not shown). In the 204 patients who received BCG treatment, there were 65 (32%) stage Ta (4G1, 26G2, 34G3, and one unknown grade), 120 (59%) stage T1 (14G2, 104G3, and two unknown grade), and 18 (9%) stage Tis (11G3 and seven unknown grade; all data not shown). Of the 318 MiMBC patients, 184 (58%) were alive at the end of our study period. There was a significant difference between deceased and alive patients in terms of age, gender, stage, and treatment regimen ($P \leq 0.05$; Supplementary Table S2).

The characteristics of TUR-only NMIBC patients of the TXBCS and SBC/EPICURO study are listed in Table 2. There were 146 such patients in TXBCS, 97 of whom had recurrence. There were no significant differences between TXBCS patients who did and did not have recurrence in gender, age, smoking status, stage, or grade (Table 2). There were 356 NMIBC patients in the SBC/EPICURO study, among whom 133 showed recurrence. There were no significant differences between SBC/EPICURO patients with and without recurrence in gender, age, smoking status, or stage, although grade was higher in patients with recurrence.

Associations between SNPs and bladder cancer risk

Among the 177 individual SNPs we analyzed in relation to cancer risk (Supplementary Table S3), three SNPs on the *GLI3* gene, rs3735361, rs3823720, and rs10951671, reached nominal significance ($P < 0.05$; Table 3); however, none of these associations remained significant after adjusting for multiple testing (data not shown). These same three *GLI3* SNPs were consistently at the top of the list of SNP associations with risk in an analysis restricted to the 419 NMIBC patients (versus 803 controls; Supplementary Table S4) or, albeit nonsignificantly, in an analysis restricted to the 318 MiMBC patients (versus 803 controls; Supplementary Table S5).

Recurrence predictors in TUR-only patients

Nine SNPs had a nominally significant individual association with recurrence in patients receiving TUR only ($P \leq 0.05$) in the TXBCS (Supplementary Table S6). Six of these nine SNPs (i.e., rs17172001, rs1017024, rs1233560, rs2718107, rs2310897, and rs11594179) and rs11677381, which is in strong linkage with *GLI2*

Table 1. Demographic and clinical variables for NMIBC patients of the TXBCS

	NMIBC (*n* = 419)					
	Recurrence, *n* (%)			Progression, *n* (%)		
	Yes (*n* = 228)	No (*n* = 191)	*P**	Yes (*n* = 71)	No (*n* = 347)	*P**
Sex			**0.03**			**0.05**
Male	195 (56.8)	148 (43.2)		64 (18.7)	278 (81.3)	
Female	33 (43.4)	43 (56.6)		7 (9.2)	69 (90.8)	
Age (y)						
Mean (SD) y	63.0 (11.2)	63.6 (11.4)	0.63	66.2 (9.9)	62.7 (11.5)	**0.02**
Smoking status†			0.63			0.40
Never	64 (52.5)	58 (47.5)		16 (13.2)	105 (86.8)	
Former	117 (56.8)	89 (43.2)		37 (18.0)	169 (82.0)	
Current	47 (51.6)	44 (48.4)		18 (19.8)	73 (80.2)	
Stage			0.28			**<0.01**
Ta	105 (54.4)	88 (45.6)		21 (10.9)	172 (89.1)	
Tis	16 (69.6)	7 (30.4)		7 (30.4)	16 (69.6)	
T1	104 (52.0)	96 (48.0)		42 (21.1)	157 (78.9)	
Grade (G)			0.20			**<0.01**
G1	5 (31.2)	11 (68.8)		1 (6.2)	15 (93.8)	
G2	81 (54.0)	69 (46.0)		12 (8.0)	138 (92.0)	
G3	128 (54.0)	109 (46.0)		53 (22.5)	183 (77.5)	
Treatment‡			**<0.01**			**<0.01**
TUR	97 (66.4)	49 (33.6)		16 (11.0)	129 (89.0)	
iBCG	91 (74.6)	31 (25.4)		37 (30.3)	85 (69.7)	
mBCG	30 (36.6)	52 (63.4)		15 (18.3)	67 (81.7)	
Others	10 (16.7)	50 (83.3)		3 (5.0)	57 (95.0)	

NOTE: The numbers of each variable may not add up to 419 due to missing data.

*P values were derived from Pearson's χ^2 test or Fisher's exact test for categorical variables, and Student's t test for continuous variables.

†Smoking status: individuals who had smoked more than 100 cigarettes in their lifetime were defined as ever smokers; others were never smokers. Smokers included current smokers and former smokers. Individuals who had quit smoking at least 1 y before diagnosis were categorized as former smokers.

‡Treatment: TUR, subgroup who had no further therapy after TUR; iBCG, subgroup who received iBCG after TUR; mBCG, subgroup who further received mBCG after the TUR and iBCG treatment; and others, which included those who received intravesical chemotherapy but no BCG.

rs11685068 ($R^2 = 1$), were genotyped in the validation data set from the 356 NMIBC patients of SBC/EPICURO (Supplementary Table S6). The SNPs rs1233560 (of *SHH*) and rs11685068 (*GLI2*) were significantly associated with recurrence in both the TXBCS and SBC/EPICURO study (Table 4). The recurrence HR was 2.07 (95% CI, 1.33-3.21; $P = 1.3 \times 10^{-3}$) for *GLI2* rs11685068 and 1.39 (95% CI, 1.14-1.70; $P = 1.0 \times 10^{-3}$) for *SHH* rs1233560 in a meta-analysis of the combined TXBCS and SBC/EPICURO data.

Recurrence predictors in BCG patients

In 204 patients receiving BCG treatment (including 122 patients in iBCG subgroup and 82 patients in mBCG subgroup), nine SNPs located on *GLI3*, *GLI2*, and *HHIP* were associated individually with time to recurrence at $P < 0.05$. After adjustment of multiple testing, two variant genotypes of *GLI3*, rs6463089 and rs3801192, remained significant and associated with a 2.40-fold (95% CI, 1.50-3.84) and 2.54-fold (95% CI, 1.47-4.39) increased recurrence risk, respectively, compared with their corresponding homozygous wild-type genotype (Table 5). Interestingly, variant genotypes of *GLI3* rs6463089 and rs3801192 showed a protective effect on recurrence in the TUR-only subgroup, with HRs of 0.74 (95% CI, 0.42-1.33, $P = 0.32$) and 0.43 (95% CI, 0.21-0.88, $P = 0.02$), respectively (Table 5). In the Kaplan-Meier estimates of recurrence-free survival in the BCG treatment group, compared with patients with the homozygous wild-type genotype of rs6463089 (recurrence-free median survival time [MST], 16.3 months; $P_{\text{log-rank}} < 0.01$) and rs3801192 (13.1 months, $P_{\text{log-rank}} = 0.02$), those with at least one variant allele at either of these two SNPs showed a shorter recurrence-free MST of 5.5 months (Fig. 2). Conversely, in patients receiving TUR only, compared with the

Table 2. Host characteristics for NMIBC patients receiving TUR only in the TXBCS and SBC/EPICURO study

	TXBCS			SBC/EPICURO study		
	Recurrence, *n* (%)			Recurrence, *n* (%)		
	Yes (*n* = 97)	No (*n* = 49)	*P**	Yes (*n* = 133)	No (*n* = 223)	*P**
Sex			0.18			0.59
Male	77 (69.4)	34 (30.6)		118 (36.9)	202 (63.1)	
Female	20 (57.1)	15 (42.9)		15 (41.7)	21 (58.3)	
Age (y)						
Mean (SD) y	62.5 (12.2)	60.6 (13.7)	0.39	65.69 (10.07)	66.14 (10.39)	0.69
Smoking status†			0.46			0.17
Never	28 (60.9)	18 (39.1)		17 (43.6)	22 (56.4)	
Former	48 (71.6)	19 (28.4)		52 (32.7)	107 (67.3)	
Current	21 (63.6)	12 (36.4)		64 (42.1)	88 (57.9)	
Stage			0.50			0.85
Ta	62 (63.9)	35 (36.1)		123 (37.7)	203 (62.3)	
Tis	31 (68.9)	14 (31.1)		1 (50)	1 (50)	
T1	2 (100)	0 (0.0)		9 (32.1)	19 (67.9)	
Grade (G)			0.55			**<0.01**
G1	4 (50)	4 (50)		57 (29.8)	134 (70.2)	
G2	55 (64.7)	30 (35.3)		57 (48.3)	61 (51.7)	
G3	34 (69.4)	15 (30.6)		19 (40.4)	28 (59.6)	

*P values were derived from Pearson's χ^2 test or Fisher's exact test for categorical variables, and Student's t test for continuous variables.

†Smoking status: individuals who had smoked more than 100 cigarettes in their lifetime were defined as ever smokers; others were never smokers. Smokers included current smokers and former smokers. Individuals who had quit smoking at least 1 y before diagnosis were categorized as former smokers. Six patients without recurrence had missing data for smoking status variable in the SBC/EPICURO study.

homozygous wild-type genotype (rs6463089 recurrence-free MST, 6.4 months; rs3801192 recurrence-free MST, 6.2 months), those with the variant alleles at either of these two SNPs had longer recurrence-free MST (for rs6463089, 10.6 months, $P_{\text{log-rank}} = 0.22$; for rs3801192, >109.9 months, $P_{\text{log-rank}} = 0.01$; Fig. 2). Although we did not conduct validation assessments of SNP association in BCG patients because of BCG differences between the SBC/EPICURO study and TXBCS (also stated in Materials and Methods), we performed bootstrap sampling to internally validate associations in the primary analysis. The overall HRs and 95% CIs generated by bootstrapping were consistent with our initial results. Table 5 lists the number of times that the bootstrap-generated P value was 0.05, 0.001, or 0.0001 for each SNP. The significant results of *GLI3* rs6463089 and rs3801192 in the BCG group reached significance at $P = 0.05$ in >90% of 100 bootstrap samplings. The bootstrap findings indicate that the results for these SNPs in the primary analysis were unlikely due to chance alone.

Table 3. Association between selected Shh pathway–related SNPs and bladder cancer risk

Gene	SNP	Genotype	Case/control	OR (95% CI)*	*P*
GLI3	rs3735361 (G>A)	GG/GA	708/733	Reference	
	Flanking 3′ UTR	AA	95/70	1.42 (1.01-1.99)	0.04
	rs3823720 (G>A)	GG/GA	704/738	Reference	
	3′ UTR	AA	99/65	1.57 (1.11-2.20)	0.01
	rs10951671 (G>A)	GG/GA	748/769	Reference	
	Intron	AA	55/34	1.75 (1.11-2.74)	0.02

Abbreviations: OR, odds ratio; UTR, untranslated region.

*Adjusted by age, gender, and smoking status.

Table 4. Significant SNP associations with recurrence in TUR-only NMIBC patients in the TXBCS and SBC/EPICURO study

	rs11685068*			rs1233560		
	Genotype count	HR (95% CI)	*P*	Genotype count	HR (95% CI)	*P*
Gene		*GLI2*			*SHH*	
Allele		G>A			A>G	
Best model†		DOM			ADD	
TXBCS	124/15/2	2.19 (1.22-3.93)	0.01	47/70/24	1.49 (1.07-2.07)	0.02
SBC/EPICURO study	335/21/0	1.91 (0.97-3.76)	0.06	105/175/76	1.34 (1.05-1.71)	0.02
Combined	459/36/2	2.07 (1.33-3.21)	1.3×10^{-3}	152/245/100	1.39 (1.14-1.70)	1.0×10^{-3}

Abbreviations: DOM, dominant; ADD, additive.

**GLI2* rs11685068 is not genotyped in the Spanish study, but in strong linkage with rs11677381 with $R^2 = 1.0$. The validation result of rs11685068 was derived from the result of rs11677381.

†Best model: the model with smallest *P* value.

Progression in NMIBC patients

We also assessed the association of SNPs with progression of NMIBC; however, we did not observe any SNPs significantly associated with bladder cancer progression (data not shown). Compared with recurrence, the progression rate in NMIBC patients is relatively low; therefore, we did not have adequate sample size to do the stratification analysis by BCG treatment status.

Overall survival in MiMBC patients

Two SNPs on *GLI2*, rs735557 and rs4848632, and the *SHH* SNP rs208684 showed individual associations with overall survival in MiMBC patients ($P < 0.05$; Supplementary Table S7). All three associations became nonsignificant, however, after adjusting for multiple comparisons (data not shown).

Discussion

The present results have important implications for the clinical management of NMIBC. Nine SNPs were significantly associated with recurrence of NMIBC patients in the TXBCS who received TUR treatment only; two of these SNPs, *SHH* rs1233560 and *GLI2* rs11685068, were validated independently in the SBC/EPICURO cohort. With initial TXBCS and replicated SBC/EPICURO results, *SHH* rs1233560 and *GLI2* rs11685068 have potential real-time clinical utility for predicting recurrence in NMIBC patients receiving TUR only. In NMIBC patients who received BCG, two variant genotypes of *GLI3* (rs6463089 and rs3801192) were significantly associated with an increased risk of recurrence and a shorter recurrence-free survival (~2.5-fold changes in each) after adjusting for multiple comparisons, and the associations were internally validated by bootstrap analysis. These results suggest that NMIBC patients with wild-type genotypes of these two *GLI3* SNPs are good candidates for BCG therapy, whereas those with the variant genotypes of these two SNPs should be spared BCG therapy.

The homozygous variant genotypes of three SNPs in *GLI3* showed significant associations with increased overall bladder cancer risk in our 803 bladder cancer cases and 803 controls. These associations remained at the top of the list of risk-associated SNPs in the case-control study restricted to the 419 NMIBC cases or 318 MiMBC cases (versus 803 controls). These associations became nonsignificant, however, after adjusting for multiple comparisons in the overall analysis and the analysis stratified for NMIBC or MiMBC. Further validation studies in independent populations are warranted to clarify the associations of these SNPs with bladder cancer risk. Two *GLI2* SNPs and an *SHH* SNP were significantly associated with overall survival in MiMBC patients, but these associations became nonsignificant after adjusting for multiple comparisons.

It is suggested that Shh signaling plays an important role in the development and prognosis of bladder cancer. For example, the most frequent loss-of-heterozygosity region in NMIBC is the locus of the Shh pathway gene *PTCH* on 9q22 (42). Changes in the level of Shh pathway gene expressions also might predict overall survival in bladder cancer patients (43). All previous studies were conducted in tumor tissues, where they showed that somatic changes in Shh genes (such as gene expression level and loss of heterozygosity) may be involved in cancer development. Our study was the first, however, to examine germ-line genetic variations in Shh signaling as cancer susceptibility factors and predictors of outcome.

We found that certain Shh pathway SNPs affected recurrence in patients receiving BCG. Although its antitumor mechanism is not fully understood, BCG can trigger a strong local immune response that leads to the expression of many cytokines at the tumor site and to an influx of granulocytes and mononuclear cells into the bladder wall (44, 45). It is hypothesized that intravesical BCG treatment can directly inhibit urothelial tumor cell growth

Table 5. Recurrence in NMIBC patients receiving BCG treatment versus those receiving TUR only and internal bootstrap validation

BCG vs TUR only

SNP	Gene	Genotype	Best model*	BCG subgroup recurrence				TUR-only subgroup recurrence			
				Yes/No		HR (95% CI)†	*P*	Yes/No		HR (95% CI)†	*P*
				ww	wv + vv			ww	wv + vv		
rs6463089	***GLI3***	G>A	**DOM**	**92/78**	**26/5**	**2.40 (1.50-3.84)**	**2×10^{-4}**	78/38	14/11	0.74 (0.42-1.33)	0.32
rs3801192	***GLI3***	G>A	**DOM**	**100/77**	**17/6**	**2.54 (1.47-4.39)**	**9×10^{-4}**	83/37	9/12	0.43 (0.21-0.88)	0.02
rs277534	*GLI2*	A>G	DOM	69/46	49/37	0.60 (0.40-0.90)	0.01	63/31	29/18	1.02 (0.64-1.61)	0.95
rs3801210	*GLI3*	G>A	ADD	52/37	65/46	1.43 (1.08-1.89)	0.01	38/14	54/35	0.86 (0.62-1.20)	0.39
rs6974655	*GLI3*	C>A	DOM	67/42	51/41	0.62 (0.42-0.93)	0.02	50/21	42/28	0.72 (0.47-1.10)	0.13
rs2237425	*GLI3*	G>C	ADD	76/60	42/20	1.45 (1.04-2.03)	0.03	70/38	22/11	1.08 (0.71-1.63)	0.73
rs2286294	*GLI3*	A>G	DOM	28/29	89/51	1.63 (1.04-2.55)	0.03	31/11	61/37	0.97 (0.62-1.52)	0.90
rs2306924	*HHIP*	A>G	DOM	25/15	93/68	0.61 (0.38-0.99)	0.04	18/14	74/35	1.44 (0.83-2.50)	0.19
rs7785287	*GLI3*	G>A	ADD	73/58	44/25	1.39 (1.00-1.93)	0.05	56/36	36/13	1.30 (0.89-1.89)	0.18

Internal bootstrap validation of significant results in patients receiving BCG treatment

SNP	Gene	Genotype	Best model*	Bootstrap‡ BCG subgroup recurrence			
				HR (95% CI)†	*P* < 0.0001	*P* < 0.001	*P* < 0.05
rs6463089	***GLI3***	G>A	**DOM**	**2.40 (1.55-3.71)**	**46**	**68**	**92**
rs3801192	***GLI3***	G>A	**DOM**	**2.54 (1.46-4.40)**	**34**	**56**	**93**
rs277534	*GLI2*	A>G	DOM	0.60 (0.38-0.94)	8	18	66
rs3801210	*GLI3*	G>A	ADD	1.43 (1.06-1.91)	9	23	69
rs6974655	*GLI3*	C>A	DOM	0.62 (0.38-1.02)	3	16	58
rs2237425	*GLI3*	G>C	ADD	1.45 (1.11-1.91)	3	6	55
rs2286294	*GLI3*	A>G	DOM	1.63 (1.01-2.62)	3	11	53
rs2306924	*HHIP*	A>G	DOM	0.61 (0.36-1.02)	4	16	61
rs7785287	*GLI3*	G>A	ADD	1.39 (1.02-1.89)	1	10	46

NOTE: Significant SNPs after correcting for multiple comparisons by *Q* value with a false discovery rate of ≤10% are in boldface. Abbreviations: w-wild type allele; v-variant allele; ww, homozygous wildtype genotype; wv, heterozygous variant genotype; vv, homozygous variant genotype.

*Best model: the model with smallest *P* value.

†Adjusted by age, sex, smoking status, tumor stage, and tumor grade using Cox proportional hazard regression where appropriate.

‡We did internal validation of the results choosing from the best genetic model using bootstrap 100 times.

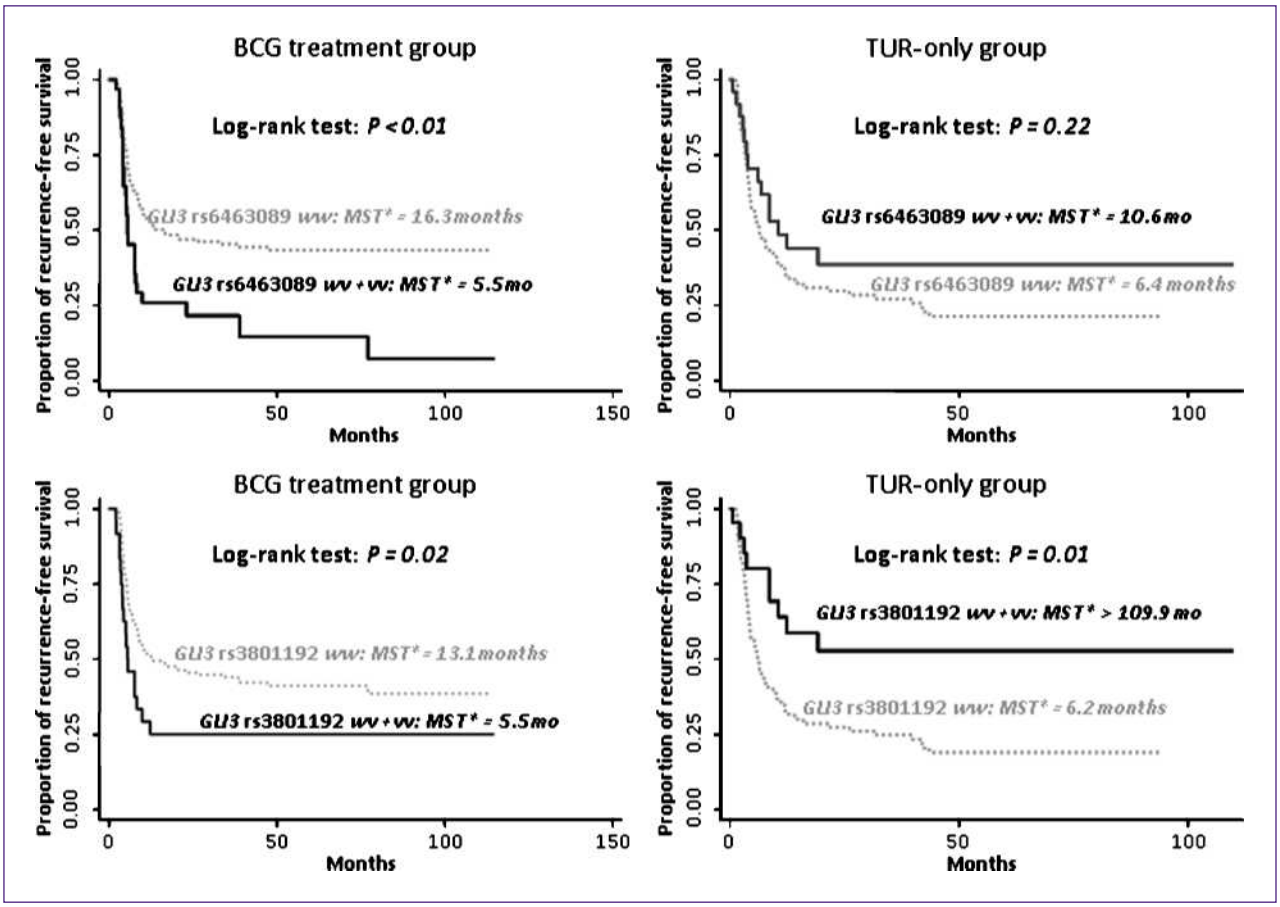


Fig. 2. Kaplan-Meier curve of recurrence-free survival in NMIBC patients receiving BCG treatment versus those receiving TUR only. w-wild type allele; v-variant allele; ww, homozygous wildtype genotype; wv, heterozygous variant genotype; vv, homozygous variant genotype.

and exerts its antitumor activity through cell immunity mediated by the T-helper type 1 cytokines (46). Previous studies have suggested that genetic variations in inflammation genes and DNA repair genes have regulatory effects on these cytokines and, hence, on BCG response (47, 48). Recent data indicate that the activation of Shh signaling is mediated by NF-κB, which is the hallmark of the inflammatory response (49, 50). Shh signaling also may influence T-cell activation (51). Therefore, it is plausible that genetic variations in the Shh pathway may affect response to BCG through their effects in regulating the inflammatory response to BCG treatment. Alternatively to this direct mediation, these genetic variations may function as a surrogate biomarker of response to BCG.

Our present findings that the SNPs most significantly associated with recurrence in NMIBC patients were *GLI2* rs11685068 (Table 4) in TUR-only patients and the independent (i.e., not or only weakly linked to one another) *GLI3* SNPs rs6463089 and rs3801192 in BCG (Table 5), induction plus or minus maintenance (Supplementary Table S8), patients showed the importance of *GLI*-family genes in bladder cancer. The *GLI* family encodes zinc-finger proteins that are amplified in malignant glioma (52). These GLI proteins are transcription factors with distinct functions in a context-dependent manner (53). The effect of Shh signaling finally depends on the flipping of GLI in a combinatorial and cooperative manner. In the absence of Shh, GLI1 generally is transcriptionally repressed, whereas GLI2 and GLI3 are cleaved by the proteasome and act as repressors of transcription. On activation by Shh, the full-length GLI becomes a transcriptional activator. GLI2 and GLI3 are weaker than GLI1 in activating transcription (54).

Strong evidence suggests that altered *GLI2* and *GLI3* play a role in cancer development and progression. *GLI2* is less studied in human diseases, but several studies in mice indicate that GLI2 overexpression or mutation is associated with basal cell carcinomas and skeletal defects and disorders (55). *GLI3* translocation, deletion, and mutations have been implicated in several types of birth defects (55). It is interesting to note that genetic variations in *GLI3* seem to modulate both bladder cancer recurrence in patients treated with BCG and bladder cancer risk. A recent *in vitro* study showed that arsenic treatment activated Shh signaling in bladder cancer cells by decreasing the stability of the repressor form of GLI3; furthermore, high levels of arsenic exposure were associated with high levels of SHH activity in

tumor samples from a cohort of bladder cancer patients (56). These results suggest that GLI3-mediated activation of the Shh pathway plays an important role at least in bladder cancer induced by the environmental toxin arsenic. It is likely that smoking, BCG, and other exposures also may affect GLI3 and lead to the activation of the Shh pathway, which may explain the risk and recurrence associations in our study.

SHH is the secreted ligand that activates the PTCH receptor and thus initiates the Shh signaling pathway. Endogenous overexpression of SHH can drive abnormal activation of the Shh signaling pathway and cell growth in solid tumors (57). SHH has been reported to be an early and late mediator of pancreatic tumorigenesis (58). We found that the SNP *SHH* rs1233560 was associated with increased recurrence in TUR-only NMIBC patients in both the TXBCS and SBC/EPICURO cohorts. Because this SNP is located in the flanking 3′ untranslated region, we hypothesized that it may affect SHH expression. Because all of the SNPs in *SHH*, *GLI2*, and *GLI3* with significant associations are haplotype-tagging SNPs located either in intergenic regions or introns without a clear functional indication, they likely are not the causal variants but are in strong linkage disequilibrium with the causal SNPs proximate to these tagging SNPs. Future studies to accurately map the causal SNPs and identify the biological mechanisms underlying the associations of *SHH*, *GLI2*, and *GLI3* with BCG-associated recurrence are needed.

The major strength of our present study is that our recurrence results in TUR-only NMIBC patients of the TXBCS were externally validated in the SBC/EPICURO cohort, which reduced the possibility of chance findings. Although it was not feasible to externally validate recurrence results in BCG-treated patients in the SBC/EPICURO population (because of substantial differences in BCG regimens between the TXBCS and SBC/EPICURO study, as mentioned earlier), we controlled for the false discovery rate of these results by adjusting for multiple testing, and we performed an internal validation bootstrap analysis. Another study strength is that the TXBCS population of NMIBC patients was large and ethnically homogeneous, which limited potential confounding by population substructure. Furthermore, all NMIBC patients in the TXBCS had complete clinical data and a long median follow-up of 48.2 months (59), which allowed for a more accurate assessment of recurrence outcomes in TUR-only and BCG patients.

In conclusion, this study indicates that Shh signaling may predict the outcome of TUR only or BCG in NMIBC patients and thus could help in future approaches to the clinical management of these patients and reducing the burden of this morbid disease.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

NIH grants U01 CA 127615, R01 CA 74880, P50 CA 91846, and R01 CA 131335.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 02/16/2010; revised 08/12/2010; accepted 08/13/2010; published OnlineFirst 09/21/2010.

References

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clin 2009;59:225–49.
2. Johansson SL, Cohen SM. Epidemiology and etiology of bladder cancer. Semin Surg Oncol 1997;13:291–8.
3. Wu X, Gu J, Grossman HB, et al. Bladder cancer predisposition: a multigenic approach to DNA-repair and cell-cycle-control genes. Am J Hum Genet 2006;78:464–79.
4. Yang H, Gu J, Lin X, et al. Profiling of genetic variations in inflammation pathway genes in relation to bladder cancer predisposition. Clin Cancer Res 2008;14:2236–44.
5. Yang H, Dinney CP, Ye Y, Zhu Y, Grossman HB, Wu X. Evaluation of genetic variants in microRNA-related genes and risk of bladder cancer. Cancer Res 2008;68:2530–7.
6. Dalbagni G. The management of superficial bladder cancer. Nat Clin Pract Urol 2007;4:254–60.
7. Sylvester RJ. The use of intravesical chemotherapy and possibilities for improving its efficacy. Eur Urol 2006;50:233.
8. Babjuk M, Oosterlinck W, Sylvester R, Kaasinen E, Bohle A, Palou-Redorta J. EAU guidelines on non-muscle-carcinoma of the bladder. Actas Urol Esp 2009;33:361–71.
9. Lamm DL, Blumenstein BA, Crissman JD, et al. Maintenance bacillus Calmette-Guerin immunotherapy for recurrent TA, T1 and carcinoma *in situ* transitional cell carcinoma of the bladder: a randomized Southwest Oncology Group Study. J Urol 2000;163:1124–9.
10. Merz VW, Marth D, Kraft R, Ackermann DK, Zingg EJ, Studer UE. Analysis of early failures after intravesical instillation therapy with bacille Calmette-Guerin for carcinoma *in situ* of the bladder. Br J Urol 1995;75:180–4.
11. Brandau S, Suttmann H. Thirty years of BCG immunotherapy for non-muscle invasive bladder cancer: a success story with room for improvement. Biomed Pharmacother 2007;61:299–305.
12. Schrier BP, Hollander MP, van Rhijn BW, Kiemeney LA, Witjes JA. Prognosis of muscle-invasive bladder cancer: difference between primary and progressive tumours and implications for therapy. Eur Urol 2004;45:292–6.
13. Saint F, Salomon L, Quintela R, et al. Do prognostic parameters of remission versus relapse after Bacillus Calmette-Guerin (BCG) immunotherapy exist? Analysis of a quarter century of literature. Eur Urol 2003;43:351–60, discussion 60–1.
14. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. Nature 2001;414:105–11.
15. Pardal R, Clarke MF, Morrison SJ. Applying the principles of stem-cell biology to cancer. Nat Rev Cancer 2003;3:895–902.
16. Dittmar T, Nagler C, Schwitalla S, Reith G, Niggemann B, Zanker KS. Recurrence cancer stem cells—made by cell fusion? Med Hypotheses 2009;73:542–7.
17. Salama P, Platell C. Colorectal cancer stem cells. ANZ J Surg 2009; 79:697–702.
18. Varnat F, Duquet A, Malerba M, et al. Human colon cancer epithelial cells harbour active HEDGEHOG-GLI signalling that is essential for tumour growth, recurrence, metastasis and stem cell survival and expansion. EMBO Mol Med 2009;1:338–51.

19. Xu X, Xing B, Han H, et al. The properties of tumor-initiating cells from a hepatocellular carcinoma patient's primary and recurrent tumor. Carcinogenesis 2009;31:167–74.
20. Liu G, Yuan X, Zeng Z, et al. Analysis of gene expression and chemoresistance of CD133$^+$ cancer stem cells in glioblastoma. Mol Cancer 2006;5:67.
21. Rich JN. Cancer stem cells in radiation resistance. Cancer Res 2007; 67:8980–4.
22. Ma S, Lee TK, Zheng BJ, Chan KW, Guan XY. CD133$^+$ HCC cancer stem cells confer chemoresistance by preferential expression of the Akt/PKB survival pathway. Oncogene 2008;27:1749–58.
23. Shafee N, Smith CR, Wei S, et al. Cancer stem cells contribute to cisplatin resistance in Brca1/p53-mediated mouse mammary tumors. Cancer Res 2008;68:3243–50.
24. Shervington A, Lu C. Expression of multidrug resistance genes in normal and cancer stem cells. Cancer Invest 2008;26:535–42.
25. Scales SJ, de Sauvage FJ. Mechanisms of Hedgehog pathway activation in cancer and implications for therapy. Trends Pharmacol Sci 2009;30:303–12.
26. Kelleher FC, Fennelly D, Rafferty M. Common critical pathways in embryogenesis and cancer. Acta Oncol 2006;45:375–88.
27. Rubin LL, de Sauvage FJ. Targeting the Hedgehog pathway in cancer. Nat Rev Drug Discov 2006;5:1026–33.
28. Yang L, Xie G, Fan Q, Xie J. Activation of the hedgehog-signaling pathway in human cancer and the clinical implications. Oncogene 2010;29:469–81.
29. Pasca di Magliano M, Hebrok M. Hedgehog signalling in cancer formation and maintenance. Nat Rev Cancer 2003;3:903–11.
30. Habuchi T, Devlin J, Elder PA, Knowles MA. Detailed deletion mapping of chromosome 9q in bladder cancer: evidence for two tumour suppressor loci. Oncogene 1995;11:1671–4.
31. Linnenbach AJ, Pressler LB, Seng BA, Kimmel BS, Tomaszewski JE, Malkowicz SB. Characterization of chromosome 9 deletions in transitional cell carcinoma by microsatellite assay. Hum Mol Genet 1993; 2:1407–11.
32. Spruck CH III, Ohneseit PF, Gonzalez-Zulueta M, et al. Two molecular pathways to transitional cell carcinoma of the bladder. Cancer Res 1994;54:784–8.
33. LaRue H, Simoneau M, Aboulkassim TO, et al. The PATCHED/Sonic Hedgehog signalling pathway in superficial bladder cancer. Med Sci (Paris) 2003;19:920–5.
34. McGarvey TW, Maruta Y, Tomaszewski JE, Linnenbach AJ, Malkowicz SB. PTCH gene mutations in invasive transitional cell carcinoma of the bladder. Oncogene 1998;17:1167–72.
35. Hudmon KS, Honn SE, Jiang H, et al. Identifying and recruiting healthy control subjects from a managed care organization: a methodology for molecular epidemiological case-control studies of cancer. Cancer Epidemiol Biomarkers Prev 1997;6:565–71.
36. Puente D, Malats N, Cecchini L, et al. Gender-related differences in clinical and pathological characteristics and therapy of bladder cancer. Eur Urol 2003;43:53–62.
37. Wu X, Ye Y, Kiemeney LA, et al. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat Genet 2009;41:991–5.
38. Rothman N, Garcia-Closas M, Chatterjee N, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet. In revision.
39. Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B-Stat Method 2002;64:479–98.
40. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B-Stat Method 2004;66:187–205.
41. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100:9440–5.
42. Aboulkassim TO, LaRue H, Lemieux P, Rousseau F, Fradet Y. Alteration of the PATCHED locus in superficial bladder cancer. Oncogene 2003;22:2967–71.
43. Quint K, Stintzing S, Alinger B, et al. The expression pattern of PDX-1, SHH, Patched and Gli-1 is associated with pathological and clinical features in human pancreatic cancer. Pancreatology 2009;9: 116–26.
44. Alexandroff AB, Jackson AM, O'Donnell MA, James K. BCG immunotherapy of bladder cancer: 20 years on. Lancet 1999;353: 1689–94.
45. Bohle A, Brandau S. Immune mechanisms in bacillus Calmette-Guerin immunotherapy for superficial bladder cancer. J Urol 2003; 170:964–9.
46. Luo Y, Chen X, O'Donnell MA. Role of Th1 and Th2 cytokines in BCG-induced IFN-γ production: cytokine promotion and simulation of BCG effect. Cytokine 2003;21:17–26.
47. Gu J, Zhao H, Dinney CP, et al. Nucleotide excision repair gene polymorphisms and recurrence after treatment for superficial bladder cancer. Clin Cancer Res 2005;11:1408–15.
48. Leibovici D, Grossman HB, Dinney CP, et al. Polymorphisms in inflammation genes and bladder cancer: from initiation to recurrence, progression, and survival. J Clin Oncol 2005;23:5746–56.
49. Nakashima H, Nakamura M, Yamaguchi H, et al. Nuclear factor-κB contributes to hedgehog signaling pathway activation through sonic hedgehog induction in pancreatic cancer. Cancer Res 2006;66: 7041–9.
50. Yamasaki A, Kameda C, Xu R, et al. Nuclear factor κB-activated monocytes contribute to pancreatic cancer progression through the production of Shh. Cancer Immunol Immunother 2010;59: 675–86.
51. Crompton T, Outram SV, Hager-Theodorides AL. Sonic hedgehog signalling in T-cell development and activation. Nat Rev Immunol 2007;7:726–35.
52. Kinzler KW, Bigner SH, Bigner DD, et al. Identification of an amplified, highly expressed gene in a human glioma. Science 1987;236:70–3.
53. Ruiz i Altaba A, Sanchez P, Dahmane N. Gli and hedgehog in cancer: tumours, embryos and stem cells. Nat Rev Cancer 2002; 2:361–72.
54. Stecca B, Ruiz IAA. Context-dependent regulation of the GLI code in cancer by HEDGEHOG and non-HEDGEHOG signals. J Mol Cell Biol 2010;2:84–95.
55. Villavicencio EH, Walterhouse DO, Iannaccone PM. The sonic hedgehog-patched-gli pathway in human development and disease. Am J Hum Genet 2000;67:1047–54.
56. Fei DL, Li H, Kozul CD, et al. Activation of Hedgehog Signaling by the Environmental Toxicant Arsenic May Contribute to the Etiology of Arsenic-Induced Tumors. Cancer Res 70:1981–8.
57. Berman DM, Karhadkar SS, Maitra A, et al. Widespread requirement for Hedgehog ligand stimulation in growth of digestive tract tumours. Nature 2003;425:846–51.
58. Thayer SP, di Magliano MP, Heiser PW, et al. Hedgehog is an early and late mediator of pancreatic cancer tumorigenesis. Nature 2003; 425:851–6.
59. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials 1996;17:343–6.

OPEN ACCESS Freely available online

PLoS one

Large-Scale Pathway-Based Analysis of Bladder Cancer Genome-Wide Association Data from Five Studies of European Background

Idan Menashe[1*☯], Jonine D. Figueroa[1☯], Montserrat Garcia-Closas[2], Nilanjan Chatterjee[1], Nuria Malats[3], Antoni Picornell[3], Dennis Maeder[1], Qi Yang[1], Ludmila Prokunina-Olsson[1], Zhaoming Wang[4], Francisco X. Real[3,5], Kevin B. Jacobs[4], Dalsu Baris[1], Michael Thun[6], Demetrius Albanes[1], Mark P. Purdue[1], Manolis Kogevinas[7,8,9,10], Amy Hutchinson[4], Yi-Ping Fu[1], Wei Tang[1], Laurie Burdette[4], Adonina Tardón[9], Consol Serra[9,11], Alfredo Carrato[12], Reina García-Closas[13], Josep Lloreta[14], Alison Johnson[15], Molly Schwenn[16], Alan Schned[17], Gerald Andriole Jr.[18], Amanda Black[1], Eric J. Jacobs[6], Ryan W. Diver[6], Susan M. Gapstur[6], Stephanie J. Weinstein[1], Jarmo Virtamo[19], Neil E. Caporaso[1], Maria Teresa Landi[1], Joseph F. Fraumeni Jr.[1], Stephen J. Chanock[1], Debra T. Silverman[1], Nathaniel Rothman[1]

1 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America, **2** Institute for Cancer Research, Surrey, United Kingdom, **3** Spanish National Cancer Research Centre, Madrid, Spain, **4** Core Genotype Facility, SAIC-Frederick, Inc., National Cancer Institute-Frederick, Frederick, Maryland, United States of America, **5** Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain, **6** Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, United States of America, **7** Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain, **8** Municipal Institute of Medical Research, Barcelona, Spain, **9** CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain, **10** National School of Public Health, Athens, Greece, **11** Universitat Pompeu Fabra, Barcelona, Spain, **12** Ramón y Cajal University Hospital, Madrid, Spain, **13** Unidad de Investigación, Hospital Universitario de Canarias, La Laguna, Spain, **14** Hospital del Mar-Institut Municipal d'Investigació Mèdica (IMIM), Universitat Pompeu Fabra, Barcelona, Spain, **15** Vermont Cancer Registry, Burlington, Vermont, United States of America, **16** Maine Cancer Registry, Augusta, Maine, United States of America, **17** Dartmouth Medical School, Hanover, New Hampshire, United States of America, **18** Department of Urology, Washington University School of Medicine, St. Louis, Missouri, United States of America, **19** National Institute for Health and Welfare, Helsinki, Finland

Abstract

Pathway analysis of genome-wide association studies (GWAS) offer a unique opportunity to collectively evaluate genetic variants with effects that are too small to be detected individually. We applied a pathway analysis to a bladder cancer GWAS containing data from 3,532 cases and 5,120 controls of European background (n = 5 studies). Thirteen hundred and ninety-nine pathways were drawn from five publicly available resources (Biocarta, Kegg, NCI-PID, HumanCyc, and Reactome), and we constructed 22 additional candidate pathways previously hypothesized to be related to bladder cancer. In total, 1421 pathways, 5647 genes and ~90,000 SNPs were included in our study. Logistic regression model adjusting for age, sex, study, DNA source, and smoking status was used to assess the marginal trend effect of SNPs on bladder cancer risk. Two complementary pathway-based methods (gene-set enrichment analysis [GSEA], and adapted rank-truncated product [ARTP]) were used to assess the enrichment of association signals within each pathway. Eighteen pathways were detected by either GSEA or ARTP at $P \leq 0.01$. To minimize false positives, we used the I^2 statistic to identify SNPs displaying heterogeneous effects across the five studies. After removing these SNPs, seven pathways ('Aromatic amine metabolism' [$P_{GSEA} = 0.0100$, $P_{ARTP} = 0.0020$], 'NAD biosynthesis' [$P_{GSEA} = 0.0018$, $P_{ARTP} = 0.0086$], 'NAD salvage' [$P_{ARTP} = 0.0068$], 'Clathrin derived vesicle budding' [$P_{ARTP} = 0.0018$], 'Lysosome vesicle biogenesis' [$P_{GSEA} = 0.0023$, $P_{ARTP} < 0.00012$], 'Retrograde neurotrophin signaling' [$P_{GSEA} = 0.00840$], and 'Mitotic metaphase/anaphase transition' [$P_{GSEA} = 0.0040$]) remained. These pathways seem to belong to three fundamental cellular processes (metabolic detoxification, mitosis, and clathrin-mediated vesicles). Identification of the aromatic amine metabolism pathway provides support for the ability of this approach to identify pathways with established relevance to bladder carcinogenesis.

Citation: Menashe I, Figueroa JD, Garcia-Closas M, Chatterjee N, Malats N, et al. (2012) Large-Scale Pathway-Based Analysis of Bladder Cancer Genome-Wide Association Data from Five Studies of European Background. PLoS ONE 7(1): e29396. doi:10.1371/journal.pone.0029396

Editor: Zhongming Zhao, Vanderbilt University Medical Center, United States of America

Received September 14, 2011; **Accepted** November 28, 2011; **Published** January 4, 2012

Funding: This project has been funded in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Support for individual studies that participated in the effort is as follows: SBCS (Dr. Silverman) - Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics and intramural contract number NCI N02-CP-11015. FIS/Spain 98/1274, FIS/Spain 00/0745, PI061614, and G03/174, Fundació Marató TV3, Red Temática Investigación Cooperativa en Cáncer (RTICC), Consolíder ONCOBIO, EU-FP7-201663; and RO1- CA089715 and CA34627. NEBCS (Dr. Silverman) - Intramural research program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics and intramural contract number NCI N02-CP-01037 PLCO (Dr. Purdue) - The National Institutes of Health (NIH) Genes, Environment and Health Initiative (GEI) partly funded DNA extraction and statistical analyses (HG-06-033-NCI-01 and RO1HL091172-01), genotyping at the Johns Hopkins University Center for Inherited Disease Research (U01HG004438 and NIH HHSN268200782096C), and study coordination at the GENEVA (Dr. Caporaso)- The NIH Genes, Environment and Health Initiative [GEI] partly funded DNA extraction and statistical analyses (HG-06-033-NCI-01 and RO1HL091172-01), genotyping at the Johns Hopkins University Center for Inherited Disease Research (U01HG004438 and NIH HHSN268200782096C) and study coordination at the GENEVA Coordination Center (U01 HG004446) for EAGLE and part of PLCO studies. Genotyping for the remaining part of PLCO and all ATBC and CPS-II samples were supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The PLCO is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, National Institutes of Health. ATBC (Dr. Albanes) - This research was supported in part by the Intramural Research Program of the NIH and the National Cancer Institute. Additionally, this research was supported by U.S. Public Health Service contracts N01-CN-45165, N01-RC-45035, and N01-RC-37004 from the National Cancer Institute, Department of Health and Human Services. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: menashei@mail.nih.gov

☯ These authors contributed equally to this work.

Introduction

Genome-wide association studies (GWAS) have served as a useful tool to identify common genetic variants associated with various complex traits [1]. As expected, each variant explains a tiny portion of the heritable component of their associated phenotypes [2,3]. Recently, Park and colleagues estimated that some proportion of the 'missing heritability' may reside in additional common low-penetrance susceptibility variants that can be discovered in larger GWAS [4]. In principle, other methods could complement the primary single-locus tests of GWAS in identifying additional susceptibility loci. One such approach is pathway (gene-set) analysis [5,6], which examines whether association signals of a collection of functionally related loci (typically genes) consistently deviate from what is expected by chance. This approach may suggest new candidate susceptibility loci and possibly provide insights into the mechanisms underlying complex traits. Pathway-based analyses have been applied to GWAS of complex diseases, including multiple sclerosis [7], type-2 diabetes [8,9], Crohn's disease [10,11], Parkinson's disease [12,13], colon [14] and breast [15] cancers.

Bladder cancer is the fourth most common malignancy among men in the western world [16]. Epidemiological studies have shown that exposure to aromatic amines (AAs) from tobacco smoking or occupation is strongly associated with bladder cancer risk [16,17,18,19]. Additionally, genetic studies have demonstrated that functional polymorphisms in two genes involved in carcinogen metabolism (N-acetyltransferase 2 [*NAT2*] and glutathione S-transferase M1 [GSTM1]) are associated with bladder cancer risk [20,21]. Notably, the risk of bladder cancer associated with NAT2 slow acetylation genotype is restricted to smokers [20,22]. Recently, a series of GWAS have identified previously unknown susceptibility loci for bladder cancer, with the prospects of more to be discovered [22,23,24,25]. To identify additional regions that harbor plausible candidate genes and shed further light on genetic basis of this disease, we applied pathway analysis to the first stage of the NCI's CGEMS bladder cancer GWAS containing 3,532 cases and 5,120 controls [22]. We report here seven pathways implicated in diverse carcinogenic processes to be enriched with bladder cancer susceptibility loci.

Materials and Methods

Study population

We applied our analyses to primary scan data of 591,637 SNPs from NCI's bladder cancer GWAS containing 3,532 cases and 5,120 controls of European ancestry from five studies (Spanish Bladder Cancer Study [SBCS], New England, Maine and Vermont Bladder Cancer Study [NEBCS-ME/VT], Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study [ATBC], the American Cancer Society Cancer Prevention Study II Nutrition Cohort [CPS-II], and the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial [PLCO]) [22].

Pathway data construction

We collected gene-sets from five publicly available pathway resources: BioCarta [26], Kyoto Encyclopedia of Genes and Genomes (KEGG) [27], NCI's Pathway Interaction Database (PID) [28], Reactome [29], and Encyclopedia of Homo sapiens Genes and Metabolism (HumanCyc) [30]. Inclusion criteria of pathways for analysis were those containing 5–100 genes to avoid testing too narrowly- or too broadly- defined functional categories. In addition, we constructed 22 candidate pathways (Table S2) based on known bladder cancer risk factors and general carcinogenic processes [31,32,33] which were not represented in the public databases above. Specifically, selection of genes was determined through 1) biochemical data for the detoxification of aromatic amines [34,35]; 2) Ingenuity pathway lists [36]; and 3) Gene ontology lists [37].

To explore the similarity between pathways in our database, we assessed the percentage of overlapping genes between each two pathways (A and B) as:

$$Overlap(\%) = \frac{\left(\frac{N_{[A\cap B]}}{N_{[A\cup B]}} + \frac{N_{[A\cap B]}}{\min\{N_A, N_B\}}\right)}{2} \times 100\% \qquad (1)$$

where N_A and N_B are the number of genes within pathways A and B.

SNPs from the first stage of the NCI bladder cancer GWAS [22] were mapped to genes in these pathways if they were located in a region encompassing 20 kb 5′ upstream and 10 kb 3′

downstream from the genes' coding regions (NCBI's human genome build 36.3). These gene's boundaries were selected attempting to capture most of the gene's coding and regulatory variants [38] as well as minimizing the overlap between genes. Overall, 1,422 pathways containing 5,647genes (24.3±21.7 [mean ± SD] genes per pathway) and ~92,000 SNPs were included in our database. A complete list of the studied pathways is available in Table S1.

Statistical analysis

SNPs with MAF<1% among controls were excluded from the analysis. We fitted logistic regression models adjusted for age, sex, study center, DNA source (buccal/blood), and smoking status (current/former/never/occasional), to assess the marginal effect of each SNP (1 degree of freedom trend test) on the risk of bladder cancer, as previously described [22]. For each gene G_j ($j = 1, \ldots, N$, where N is the total number of genes in our dataset), the SNP with the lowest p-value among all SNPs that were mapped to its region was selected to represent the gene in the pathway analysis. We used two approaches to test for overrepresentation of association signals within pathways in our database:

A. Gene-set enrichment analysis (GSEA; [12]): In this approach, the −log10 of the p-value of each gene's best SNP was used as the gene's test statistics ($r_j = -\log10(p_j)$. Then, a weighted Kolmogorov-Smirnov procedure was used to assess for overrepresentation of gene's statistics Enrichment Score (*ES*) within each pathway (*S*) [15].

$$ES_s = \max_{1 \le j \le N} \{ \sum_{G_{j*} \in S, j* \le j} \frac{|r_{j*}|}{W_s} - \sum_{G_{j*} \notin S, j* \le j} \frac{1}{N - N_H} \} \quad (2)$$

where, $W_S = \sum_{G_{j*} \in S} |r_{j*}|$ and N_H is the number of genes in a pathway.

The statistical significance of ES_S was empirically evaluated using 10,000 permutations (permuting the genotype data between individuals and keeping the LD between SNPs intact).

B. Adaptive Rank-Truncated Product (ARTP; [39]): In this approach the genes' best SNP p-values (p_j) in each pathway were ordered from lowest to highest. Then, the mathematical product was computed for all possible sets of $p_{(j)}$ such that

$$W(K) = \prod_{j=1}^{K} (p_{(j)}) \quad (3)$$

with K, $1 \le K \le L$, being all possible integers (the truncation points) between 1 and L, with L being the number of genes in a pathway. In words, $W(K)$ is simply the product of the K smallest *P*-values in a pathway. Next, we used the *minP* statistics [40,41] to evaluated what is the K truncation point where the $W(K)$ get the most statistically significant value.

$$\min P = \min_{1 \le j \le J} \breve{s}(K_j) \quad (4)$$

where $\breve{s}(K_j)$ be the estimated P-value for $W(K_j)$, $K_1 \le \ldots \le K$. We then used two-level permutation procedure (10, 000 permutations, permuting the genotype data between individuals and keeping the LD structure between SNPs intact) to estimate $\breve{s}(K_j)$, and to adjust for multiple testing over the different truncation points used.

Using both the GSEA and ARTP methods that employ different approaches to assess the enrichment of gene-based signals within predefined gene-sets may facilitate capturing a broader range of candidate pathways for bladder cancer susceptibility.

Finally, we calculated a false discovery rate (*FDR*) to assess the proportion of expected false positive findings in the GSEA and ARTP analyses. In short, we normalized the GSEA and ARTP statistics for each pathway ($NS_{S(GSEA)}$ and $NS_{S(ARTP)}$ respectively) based on the mean and standard deviation of the corresponding permutation data [12]. This procedure allows a direct comparison of pathways with different sizes and gene compositions. Then, we used these normalized statistics to calculate the FDR as:

$$FDR = \frac{\sum_S^{per} NS_S^{per} \ge NS_S^*}{\sum_S^{per} NS_S^{per}} \Big/ \frac{\sum_S NS_S \ge NS_S^*}{\sum_S NS_S} \quad (5)$$

Genetic heterogeneity analysis

To minimize false positives, we estimated the I-squared statistic (I^2) [42] to identify SNPs displaying heterogeneous effects across the five studies [ATBC, CPSII, NEBCS (ME, VT), PLCO, and SBCS]. I^2describes the proportion of total variation in study estimates that is due to heterogeneity. In short, a meta-analysis was applied to every SNP belonging to one of the top pathways using the genotype frequency counts of cases and controls to estimate per-allele OR and CI's. SNPs with I^2 *P*-values<0.2 were removed from further analyses. We evaluated the OR, CI and p values for both the meta-analysis and they were similar in both models, and did not change the interpretation of the data. These analyses were done using STATA (Version 11, STATA Corporation, College Station, TX).

Results

Overall, there was good correlation between the results of the GSEA and the ARTP methods (r = 0.74, P<0.0001). A detailed examination of the results revealed that, on average, GSEA performed better in detecting pathways enriched with multiple weak association signals while ARTP appeared to be more powerful in detecting pathways where only few genes with relatively strong signals are dominating. Notably, the AA metabolism pathway, which contains several known bladder cancer susceptibility loci, was detected by both GSEA and ARTP methods ($P_{GSEA} = 0.0100$, $P_{ARTP} = 0.0020$). Therefore, we used its significance level as a reference for highlighting additional candidate susceptibility pathways. Of the 1421 pathways examined, 18 were significantly enriched with association signals at the $P<0.01$ level (Table 1). Of these, seven pathways were detected by both GSEA and ARTP, four pathways were detected only by GSEA, and seven were detected only by ARTP. After removing SNPs with heterogeneous effects across the five studies (I^2 *P*-value<0.2), the enrichment signals remained significant ($P<0.01$) in seven pathways belonging to four cellular processes ("aromatic amine [AA] metabolism", "Nicotinamide adenine dinucleotide [NAD] metabolism", "Clathrin-mediated vesicles", and "Mitosis"). For clarity, from this point forward, we will refer only to the results from the post heterogeneity analysis.

Table 1. Pathways enriched with bladder cancer susceptibility loci at a $P \leq 0.01$ level using GSEA and ARTP.

			GSEA			ARTP			Gene overlap (%)
Pathway	**source**	**# genes[1]**	**# genes[2]**	**p-value[3]**	**FDR[4]**	**# genes[2]**	**p-value[3]**	**FDR[4]**	
Aromatic amine metabolism	Self	11	(5); 1	**(0.0059); 0.0100**	(>0.5)	(9); 1	**(0.0012); 0.0020**	(0.28)	NA
NAD biosynthesis I (from aspartate)	HumanCyc	5	(4); 4	**(0.0021); 0.0018**	(>0.5)	(4); 4	**(0.0086); 0.0086**	(0.36)	44%
NAD salvage pathway II	HumanCyc	9	(5); 6	(0.0150); 0.0583	(>0.5)	(7); 8	**(0.0033); 0.0068**	(0.32)	
Clathrin derived vesicle budding	Reactome	15	(6); 6	(0.0210); 0.0189	(>0.5)	(9); 9	**(0.0018) 0.0018**	(0.35)	
Lysosome Vesicle Biogenesis	Reactome	10	(6); 7	**(0.0031); 0.0023**	(>0.5)	(7); 7	**(<0.0001); <0.0001**	(0.16)	49%
Retrograde neurotrophin signaling	Reactome	9	(4); 4	**(0.0092); 0.0084**	(>0.5)	(4) ;4	(0.0192); 0.0192	(0.41)	
Mitotic Metaphase/Anaphase Transition	Reactome	8	(3); 3	**(0.0043); 0.0040**	(>0.5)	(3); 3	(0.0187); 0.0187	(0.43)	55%
Mitotic Prometaphase	Reactome	80	(12); 12	(0.0955); 0.2567	(>0.5)	(13); 12	**(0.0095)**; 0.0346	(0.37)	
Control of skeletal myogenesis by hdac and calcium/calmodulin-dependent kinase (camk)	BioCarta	21	(11); 10	(0.1216); 0.2322	(>0.5)	(7); 3	**(0.0040)**; 0.0617	(0.29)	12%
B cell receptor signaling pathway	KEGG	75	(29); 28	(0.1121); 0.1931	(>0.5)	(10); 9	**(0.0059)**; 0.0244	(0.38)	
Syndecan-1-mediated signaling events	PID	15	(12); 9	**(0.0014)**; 0.0388	(>0.5)	(12); 11	**(0.0092)**; 0.1666	(0.43)	18%
Syndecan-2-mediated signaling events	PID	31	(19); 16	**(0.0048)**; 0.0559	(>0.5)	(31); 31	**(0.0078)**; 0.1404	(0.42)	
TGF-beta signaling pathway	KEGG	85	(41); 36	**(0.0090)**; 0.0988	(>0.5)	(57); 57	(0.0251); 0.2196	(>0.5)	NA
Activated AMPK stimulates fatty-acid oxidation in muscle	Reactome	8	(4); 3	(0.0434); 0.2470	(>0.5)	(8); 8	**(0.0017)**; 0.0454	(0.41)	
AMPK inhibits chREBP transcriptional activity	Reactome	5	(3); 2	**(0.0010)**; 0.0411	(>0.5)	(3); 2	**(0.0014)**; 0.0465	(0.33)	39%
Reversal of insulin resistance by leptin	BioCarta	10	(5); 7	(0.0170); 0.6432	(>0.5)	(10); 2	**(0.0028)**; 0.1635	(0.37)	
Maturity onset diabetes of the young	KEGG	25	(12); 11	**(0.0067)**; 0.0308	(>0.5)	(12); 16	(0.0390); 0.1908	(>0.5)	NA
Metabolism of polyamines	Reactome	12	(6); 4	**(0.0055)**; 0.0460	(>0.5)	(7); 5	**(0.0040)**; 0.0657	(0.32)	NA

Results of the top ranked pathways ($P<0.01$) using GSEA and ARTP. In parenthesis are results prior of removal SNPs displaying heterogeneous signals.
[1]The number of genes in the pathway.
[2]The number of genes underlying the enrichment signal in the pathway.
[3]*P*-value of the enrichment score based on 10,000 permutations.
[4]False-discovery rate calculated based on the normalized statistics of the permutation data to account for the variable sizes of genes and pathways.
doi:10.1371/journal.pone.0029396.t001

Aromatic amine [AA] metabolism

Table 2 displays the results for the genes in the AA pathway. The enrichment signals in this pathway were mainly driven by SNPs in the *UGT1A9* and *NAT2* genes. SNPs in these genes were identified in the primary analysis of this GWAS [22]. Removing these two genes from the pathway analyses reduced the enrichment signal in the AA metabolism pathway in both methods but still ranked it relatively high using the GSEA ($P_{GSEA} = 0.0130$, $P_{ARTP} = 0.1217$). Apart from *UGT1A9* and *NAT2*, five additional genes in this pathway had SNPs with significant genetic effect ($P_{\text{trend}}<0.05$). These included *NAT1*, *UGT1A4*, *UGT1A6*, *NQO1* and *CYP1B1*.

Some of the genes in the AA metabolism pathway (i.e. *CYP1A1* and *CYP1A2*; *UGT1A4*, *UGT1A6* and *UGT1A9*; *SULT1A1* and *SULT1A2*) occur on the same chromosomal locus and consequently share similar tagging SNPs. To assess the effect of this redundancy on the pathway enrichment signal, we pooled together genes with overlapping SNPs and treated them as a single genetic unit in our pathway analyses. Consequently, the number of loci included in the AA metabolism pathway was reduced to seven, (Table S2) and the corresponding enrichment signals were strengthened ($P_{GSEA} = 0.0046$, $P_{ARTP} = 0.0001$). Even when removing the *NAT2* and *UGT1A* regions from this gene-set, its corresponding enrichment signal remains relatively high ($P_{GSEA} = 0.024$, $P_{ARTP} = 0.0921$).

NAD metabolism

Two nicotinamide adenine dinucleotide (NAD) metabolism pathways were detected in this analysis. The "NAD biogenesis I" pathway (HumanCyc) was detected by both GSEA and ARTP ($P_{GSEA} = 0.0018$, $P_{ARTP} = 0.0086$), and the "NAD salvage II" pathway (HumanCyc) was detected only by the ARTP method ($P_{ARTP} = 0.0068$). Table 3 presents the results for the genes in these pathways. The three NMNAT genes (*NMNAT1*, *NMNAT2*, and *NMNAT3*) that are shared by both of these two pathways harbor SNPs with significant genetic effect ($P_{trend}<0.05$) and therefore likely to dominate the significant enrichment signals in these pathways. Other genes displaying significant bladder cancer risk are *QPRT* in the "NAD I" pathway, and *ACP6*, *ITGB1BP3*, *ACPL2* in the "NAD II" pathway.

Vesicle biogenesis and budding

Three pathways involved in clathrin-dependent vesicle biogenesis and budding were detected in this analysis. The "Lysosome Vesicle Biogenesis" pathway (Reactome) showed the strongest enrichment signal among all pathways in this study, and was detected by both GSEA and ARTP ($P_{GSEA} = 0.0023$, $P_{ARTP}<0.0001$). The "Clathrin derived vesicle budding" pathway (Reactome) was detected only by ARTP ($P_{ARTP} = 0.0018$), while the "Retrograde neurotrophin signaling" pathway (Reactome) was detected only by GSEA ($P_{GSEA} = 0.0084$). Table 4 displays the

Table 2. Summary of genes in the aromatic amine metabolism pathway used for pathway-based analysis of multi-study bladder cancer GWAS.

Gene	# SNPs[1]	SNP[2]	SNP[3] rank	MAF[4]	Allelic OR (95% CI)[5]	*P*-value[6]		
UGT1A9	72	rs11892031	1	0.08	0.77	0.68	0.87	3.6×10^{-5}
NAT2	15	rs4646249	1	0.28	0.89	0.83	0.95	0.0013
NAT1	11	rs9650592	1	0.11	0.86	0.78	0.96	0.0054
UGT1A4	41	rs4148328	1	0.38	0.91	0.85	0.98	0.0086
UGT1A6	62	rs4148328	1	0.38	0.91	0.85	0.98	0.0086
NQO1	6	rs1437135	1	0.20	0.91	0.84	0.99	0.0275
CYP1B1	13	rs2855658	1	0.43	0.94	0.88	1	0.0477
CYP1A1	4	rs2472297	2	0.22	1.03	0.95	1.11	0.4758
CYP1A2	5	rs2472297	4	0.22	1.03	0.95	1.11	0.4758
SULT1A1	1	rs1968752	1	0.37	1.01	0.95	1.08	0.7321
SULT1A2	1	rs4788073	1	0.37	0.99	0.93	1.06	0.8344

[1]Number of SNPs genotyped in the gene region (20 kb 5′ upstream and 10 kb 3′ downstream from the gene's coding region).
[2]The SNP representing the gene in the pathway analysis after the removal of SNPs with heterogeneous effects.
[3]The rank of the SNP among all SNPs in the gene's region based on their p-values.
[4]Minor allele frequency among controls.
[5]Per allele odds ratios +95% confidence intervals from logistic regression models adjusting for age, sex, study center, DNA source , and smoking.
[6]1 d.f. trend test.
doi:10.1371/journal.pone.0029396.t002

results for the genes in these pathways. Three genes are shared by the three pathways: *CLTA* and *CLTC*, which encode for the light and heavy chains of clathrin respectively, and *SH3GL2* which is associated with clathrin-mediated endocytosis. The association of SNPs in these three genes with bladder cancer risk ranked them among the top four genes in these pathways.

Mitosis

The ''Mitotic metaphase/anaphase transition'' (Reactome) was detected by the GSEA method ($P_{GSAE} = 0.0040$) and was marginally significant using ARTP ($P_{ARTP} = 0.0187$). Interestingly, all eight genes in this pathway are included in the more comprehensive ''Mitotic prometaphase'' pathway that was detected in the initial pathway screening, but had a less significant signal after removing SNPs with heterogeneous signals (Table 1). Results for the eight genes included in the ''Mitotic metaphase/anaphase transition'' pathway are presented in Table 5. Three SNPs in three genes (*FBXO5*, *SMC3* and *SPC24*) were associated with significant protective effect on bladder cancer ($P_{trend} < 0.05$).

Table 3. Summary of genes in the NAD metabolism pathways used for pathway-based analysis of multi-study bladder cancer GWAS.

Pathway	Gene	# SNPs[1]	SNP[2]	SNP[3] rank	MAF[4]	Allelic OR (95% CI)[5]			*P*-value[6]
NAD1/NAD2	NMNAT3	36	rs7636269	1	0.48	1.12	1.05	1.20	0.0004
NAD2	ACP6	16	rs1344	1	0.41	1.11	1.04	1.18	0.0017
NAD1	QPRT	7	rs3862476	1	0.07	1.19	1.04	1.35	0.0087
NAD1/NAD2	NMNAT2	36	rs4652795	1	0.38	0.92	0.86	0.98	0.0099
NAD1/NAD2	NMNAT1	8	rs1220398	1	0.14	0.89	0.81	0.98	0.0169
NAD2	ITGB1BP3	8	rs2304191	1	0.11	1.11	1.01	1.23	0.0355
NAD2	ACPL2	31	rs3210458	2	0.09	1.12	1.00	1.25	0.0421
NAD2	NUDT12	5	rs371315	1	0.28	1.07	1.00	1.15	0.0686
NAD2	NT5C3L	6	rs9907244	1	0.43	0.95	0.89	1.01	0.1094
NAD1	NADSYN1	17	rs4945007	1	0.06	1.10	0.96	1.25	0.1555
NAD2	C9orf95	19	rs7021664	1	0.08	0.94	0.83	1.06	0.3193

[1]Number of SNPs genotyped in the gene region (20 kb 5′ upstream and 10 kb 3′ downstream from the gene's coding region).
[2]The SNP representing the gene in the pathway analysis after the removal of SNPs with heterogeneous effects.
[3]The rank of the SNP among all SNPs in the gene's region based on their p-values.
[4]Minor allele frequency among controls.
[5]Per allele odds ratios +95% confidence intervals from logistic regression models adjusting for age, sex, study center, DNA source , and smoking.
[6]1 d.f. trend test.
doi:10.1371/journal.pone.0029396.t003

Table 4. Summary of genes in the Clathrin-mediated vesicle pathways used for pathway-based analysis of multi-study bladder cancer GWAS.

Pathway	Gene	# SNPs[1]	SNP[2]	SNP[3] rank	MAF[4]	Allelic OR (95% CI)[5]			*P*-value[6]
Clathrin/Lysosome/Retrograde	CLTA	10	rs10972786	1	0.06	1.27	1.11	1.45	0.0004
Clathrin/Lysosome	ARRB1	29	rs667791	1	0.39	1.11	1.04	1.19	0.0014
Clathrin/Lysosome/Retrograde	SH3GL2	92	rs2209426	1	0.17	0.87	0.80	0.95	0.0020
Clathrin/Lysosome/Retrograde	CLTC	10	rs7224631	1	0.09	1.19	1.06	1.32	0.0023
Clathrin/Lysosome	DNAJC6	38	rs1325607	1	0.21	1.12	1.03	1.21	0.0057
Clathrin/Lysosome	HSPA8	8	rs11218950	1	0.05	0.80	0.68	0.95	0.0087
Retrograde	NGF	45	rs12760036	1	0.10	0.85	0.76	0.96	0.0096
Clathrin/Lysosome	AP1G1	7	rs9932707	1	0.45	1.07	1.00	1.14	0.0353
Clathrin	VAMP2	3	rs3202848	1	0.37	0.93	0.86	1.00	0.0572
Clathrin	VAMP8	9	rs719023	1	0.39	0.94	0.88	1.00	0.0631
Retrograde	DNAL4	7	rs738141	1	0.17	1.08	1.00	1.18	0.0645
Clathrin	SNAP23	3	rs4924682	1	0.01	1.27	0.95	1.70	0.1087
Clathrin/Lysosome	DNM2	16	rs4804528	1	0.43	0.95	0.89	1.02	0.1437
Retrograde	DNM1	13	rs13285411	1	0.12	0.93	0.84	1.03	0.1463
Clathrin/Lysosome	AP1B1	14	rs5763140	1	0.11	1.08	0.97	1.19	0.1500
Clathrin/Lysosome	ARF1	4	rs3768331	1	0.38	1.05	0.98	1.12	0.1536
Clathrin	GBF1	15	rs1057050	1	0.06	0.90	0.78	1.04	0.1673
Retrograde	NTRK1	13	rs1888861	1	0.23	0.95	0.88	1.03	0.2275
Retrograde	AP2A2	12	rs7483870	1	0.23	0.96	0.89	1.04	0.3014
Retrograde	AP2A1	9	rs2286948	1	0.36	1.03	0.96	1.10	0.3694
Clathrin	STX4	1	rs10871454	1	0.39	1.00	0.94	1.07	0.9722

[1]Number of SNPs genotyped in the gene region (20 kb 5′ upstream and 10 kb 3′ downstream from the gene's coding region).
[2]The SNP representing the gene in the pathway analysis after the removal of SNPs with heterogeneous effects.
[3]The rank of the SNP among all SNPs in the gene's region based on their p-values.
[4]Minor allele frequency among controls.
[5]Per allele odds ratios +95% confidence intervals from logistic regression models adjusting for age, sex, study center, DNA source, and smoking.
[6]1 d.f. trend test.
doi:10.1371/journal.pone.0029396.t004

Discussion

Our pathway-based analysis of a large bladder cancer GWAS using two complementary pathway-based methods (GSEA and ARTP) identified an overrepresentation of association signals in seven pathways ('Aromatic amine metabolism', 'NAD biosynthesis', 'NAD salvage', 'Clathrin derived vesicle budding', 'Lysosome vesicle biogenesis', 'Retrograde neurotrophin signaling', and 'Mitotic metaphase/anaphase transition') and suggest involvement in at least three cellular processes (metabolic detoxification, mitosis, and clathrin-mediated vesicles).

The identification of the AA metabolism pathway in this study by both GSEA and ARTP could be considered a good indication for the utility of this approach, since AA metabolism has established relevance to bladder cancer susceptibility. Interestingly, the enrichment signal in this pathway is driven by variations in the *UGT1A* gene cluster and the *NAT1*, *NAT2*, and *NQO1* genes (Table 1) that are involved in detoxification processes in the AA pathway [34,35]. The strong enrichment signal left in this pathway even after the removal of the *UGT1A* and *NAT2* genes from the analysis indicates that other genetic variations affecting aromatic amines detoxification may contribute to bladder cancer susceptibility.

The detection of the NAD metabolism pathway may be relevant to bladder cancer susceptibility through several carcinogenic mechanisms. First, NAD homeostasis has been shown to play a role in various redox reactions that may lead to irreversible cellular damage and consequently to the initiation of malignant tumor [43]. In addition, NAD has been shown to be involved in DNA repair and telomere maintenances [44] as well as in energy production both of which are important processes in cancer development. Interestingly, NAD metabolism pathway has been implicated in a recent pathway-based analysis of colon cancer GWAS [14]. Colon and bladder cancers have been associated with NAT2 acetylation status. For bladder cancer, in which N-acetylation is a detoxification step, NAT2 slow acetylator phenotype presents a higher risk. In contrast, for heterocyclic amine-related colon cancer in which N-acetylation is negligible and O-acetylation is a carcinogen-activation step, NAT2 rapid acetylator phenotype presents a higher risk [45]. Thus, similar metabolic pathways could play diverse roles in the etiology of these two cancers.

Three clathrin-mediated vesicle pathways are also highlighted in this study. Clathrin-coated vesicles play essential role in intracellular trafficking, endocytosis, and exocytosis [46]. In this realm, it has been shown that clathrin-mediated vesicle pathways regulate the signaling and cellular localization of several growth factors [47] that are known to play a role in cancer susceptibility. Interestingly, clathrin may be also relevant to the Mitotic Metaphase/Anaphase transition pathway that was also implicated in this study. During mitosis, clathrin helps stabilizing the

Table 5. Summary of genes in the Mitotic Metaphase/Anaphase Transition pathway used for pathway-based analysis of multi-study bladder cancer GWAS.

Gene	# SNPs[1]	SNP[2]	SNP[3] rank	MAF[4]	Allelic OR (95% CI)[5]			*P*-value[6]
FBXO5	11	rs9479476	1	0.11	0.83	0.75	0.93	0.0010
SMC3	8	rs7918064	1	0.27	0.90	0.84	0.97	0.0073
SPC24	18	rs4804149	2	0.28	0.92	0.85	0.99	0.0202
CENPQ	7	rs4267943	1	0.36	0.94	0.87	1.01	0.0706
NDC80	15	rs13381300	1	0.07	0.91	0.80	1.04	0.1673
NUP107	7	rs11177325	1	0.31	0.95	0.89	1.02	0.1951
CENPA	4	rs2060390	1	0.26	0.98	0.91	1.06	0.6106
SMC1A	2	rs1264013	1	0.42	1.00	0.95	1.05	0.9876

[1]Number of SNPs genotyped in the gene region (20 kb 5′ upstream and 10 kb 3′ downstream from the gene's coding region).
[2]The SNP representing the gene in the pathway analysis after the removal of SNPs with heterogeneous effects.
[3]The rank of the SNP among all SNPs in the gene's region based on their p-values.
[4]Minor allele frequency among controls.
[5]Per allele odds ratios +95% confidence intervals from logistic regression models adjusting for age, sex, study center, DNA source , and smoking.
[6]1 d.f. trend test.
doi:10.1371/journal.pone.0029396.t005

kinetochore fibers which are required for the proper function of the mitotic spindle [48]. Thus, the overrepresentation of association signals in two distinct pathways associated with mitosis suggest that perturbations in the mitotic process, and particularly those related to the metaphase/anaphase transition, may modify the risk of human bladder cancer.

Strengths of our study are the large sample size; the use of primary scan data from five independent studies allowing us to address consistency of effects across the different populations; and the use of two complementary pathway-based methods. A limitation of our study is the lack of pathway-based signals to reach a noteworthy FDR significance level, with only one pathway (Lysosome Vesicle Biogenesis) having an FDR value <0.2. This could be partially due to the inherent limits of the methods used, the inadequate annotation of relevant pathways in public databases, or due to weak association signals in our data. Recent analysis of bladder cancers using RNA expression data, have also highlighted enrichment of genes with similar processes as we identified in our genomic data here, including metabolic processes, which provide further plausibility that the pathways identified may be relevant to bladder cancer susceptibility [49]. Furthermore, the high rank of the AA metabolism pathway in both GSEA and ARTP support the power of these methods to highlight pathways with established relevance to bladder cancer susceptibility and may therefore similarly suggest the involvement of metabolic detoxification, mitosis and clathrin-mediated pathways in bladder carcinogenesis.

Supporting Information

Table S1 Details and results for all 1423 pathways included in this study.
(XLS)

Table S2 List of genes included in the 22 self-constructed candidate pathways.
(XLS)

Acknowledgments

We would like to thank Leslie Carroll (Information Management Services, Silver Spring, MD, USA), Gemma Castaño-Vinyals (Institut Municipal d'Investigació Mèdica, Barcelona, Spain), Fernando Fernández (Institut Municipal d'Investigació Mèdica, Barcelona, Spain), Paul Hurwitz (Westat, Inc., Rockville, MD, USA)

Charles Lawrence (Westat, Inc., Rockville, MD, USA), Marta Lopez-Brea (Marqués de Valdecilla University Hospital, Santander, Cantabria, Spain), Anna McIntosh (Westat, Inc., Rockville, MD, USA)

Angeles Panadero (Hospital Ciudad de Coria, Coria (Cáceres), Spain), Fernando Rivera (Marqués de Valdecilla University Hospital, Santander, Cantabria, Spain), Robert Saal (Westat, Rockville, MD, USA)

Maria Sala (Institut Municipal d'Investigació Mèdica, Barcelona, Spain), Kirk Snyder (Information Management Services, Inc., Silver Spring, MD), Anne Taylor (Information Management Services, Inc., Silver Spring, MD), Montserrat Torà (Institut Municipal d'Investigació Mèdica, Barcelona, Spain), Jane Wang (Information Management Services, Silver Spring, MD, USA)

Author Contributions

Conceived and designed the experiments: IM JDF MG NC SJC DTS NR. Performed the experiments: ZW KBJ AH LB. Analyzed the data: IM JF QY DM. Contributed reagents/materials/analysis tools: NM AP LP FXR MT DA MPP MK YF WT AT CS AC RG JL AJ MS AS GA AB AJJ RWD SMG SJW JV NEC MTL JFF. Wrote the paper: IM JF.

References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
3. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446–450.
4. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, et al. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42: 570–575.
5. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, et al. (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol.
6. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11: 843–854.
7. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, et al. (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet.
8. Perry JR, McCarthy MI, Hattersley AT, Zeggini E, Weedon MN, et al. (2009) Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach. Diabetes.
9. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE (2010) Integrating pathway analysis and genetics of gene expression for genome-wide association studies. Am J Hum Genet 86: 581–591.
10. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, et al. (2009) Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet 84: 399–405.
11. Chen X, Wang L, Hu B, Guo M, Barnard J, et al. (2010) Pathway-based analysis for genome-wide association studies using supervised principal components. Genet Epidemiol 34: 716–724.
12. Wang K, Li M, Bucan M (2007) Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet 81: 1278–1283.
13. Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, et al. (2007) A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet 3: e98.
14. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, et al. (2010) Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet 86: 860–871.
15. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, et al. (2010) Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res 70: 4453–4459.
16. Silverman DT, Devesa SS, Moore LE, Rothman N (2006) Bladder cancer. In: Schottenfeld D, Fraumeni JF, Jr., eds. Cancer Epidemiology and Prevention. 3 ed. New York: Oxford University Press. pp 1101–1127.

17. Silverman DT, Hartge P, Morrison AS, Devesa SS (1992) Epidemiology of bladder cancer. Hematol Oncol Clin North Am 6: 1–30.
18. Vineis P, Pirastu R (1997) Aromatic amines and cancer. Cancer Causes Control 8: 346–355.
19. Talaska G (2003) Aromatic amines and human urinary bladder cancer: exposure sources and epidemiology. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 21: 29–43.
20. Garcia-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, et al. (2005) NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. Lancet 366: 649–659.
21. Moore LE, Baris DR, Figueroa JD, Garcia-Closas M, Karagas MR, et al. (2011) GSTM1 null and NAT2 slow acetylation genotypes, smoking intensity and bladder cancer risk: results from the New England bladder cancer study and NAT2 meta-analysis. Carcinogenesis 32: 182–189.
22. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, et al. (2010) A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet 42: 978–984.
23. Kiemeney LA, Sulem P, Besenbacher S, Vermeulen SH, Sigurdsson A, et al. (2010) A sequence variant at 4p16.3 confers susceptibility to urinary bladder cancer. Nat Genet 42: 415–419.
24. Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, et al. (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet 40: 1307–1312.
25. Wu X, Ye Y, Kiemeney LA, Sulem P, Rafnar T, et al. (2009) Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat Genet 41: 991–995.
26. BioCarta website. Available: http://www.biocarta.com/genes/allpathways.asp. Accessed 2011 Mar 10.
27. KEGG website. Available: http://www.genome.jp/kegg/pathway.html. Accessed 2011 Mar 10.
28. Pathway Interaction Database. Available: http://pid.nci.nih.gov. Accessed 2011 Mar 10.
29. Reactome website. Available: http://www.reactome.org. Accessed 2011 Mar 10.
30. HumanCyc website. Available: http://humancyc.org. Accessed 2011 Mar 10.
31. Figueroa JD, Malats N, Rothman N, Real FX, Silverman D, et al. (2007) Evaluation of genetic variation in the double-strand break repair pathway and bladder cancer risk. Carcinogenesis 28: 1788–1793.
32. Figueroa JD, Malats N, Real FX, Silverman D, Kogevinas M, et al. (2007) Genetic variation in the base excision repair pathway and bladder cancer risk. Hum Genet 121: 233–242.
33. Figueroa JD, Garcia-Closas M, Rothman N (2010) Case studies: Cumulative assessment of the role of human genome variation in specific diseases - bladder cancer. In: Khoury M, Bedrosian S, Gwinn M, Higgins J, Ioannidis J, et al. eds. Human Genome Epidemiology, 2nd Edition Building the evidence for using genetic information to improve health and prevent disease Oxford University Press.
34. Skipper PL, Tannenbaum SR (1994) Molecular dosimetry of aromatic amines in human populations. Environ Health Perspect 102 Suppl 6: 17–21.
35. Skipper PL, Kim MY, Sun HL, Wogan GN, Tannenbaum SR (2010) Monocyclic aromatic amines as potential human carcinogens: old is new again. Carcinogenesis 31: 50–58.
36. Ingenuity. website. Available: http://www.ingenuity.com/. Accessed 2010 Feb 24.
37. The Gene Ontology website. Available: http://www.geneontology.org/. Accessed 2010 Feb 24.
38. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4: e1000214.
39. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, et al. (2009) Pathway analysis by adaptive combination of P-values. Genet Epidemiol 33: 700–709.
40. Dudbridge F, Koeleman BP (2003) Rank truncated product of P-values, with application to genomewide association scans. Genet Epidemiol 25: 360–366.
41. Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 11: 2115–2119.
42. Higgins JP, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Stat Med 21: 1539–1558.
43. Magni G, Orsomando G, Raffelli N, Ruggieri S (2008) Enzymology of mammalian NAD metabolism in health and disease. Front Biosci 13: 6135–6154.
44. Burkle A (2005) Poly(ADP-ribose). The most elaborate metabolite of NAD+. FEBS J 272: 4576–4589.
45. Hein DW (2002) Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. Mutat Res 506–507: 65–77.
46. Royle SJ (2006) The cellular functions of clathrin. Cell Mol Life Sci 63: 1823–1832.
47. Kirisits A, Pils D, Krainer M (2007) Epidermal growth factor receptor degradation: an alternative view of oncogenic pathways. Int J Biochem Cell Biol 39: 2173–2182.
48. Royle SJ, Bright NA, Lagnado L (2005) Clathrin is required for the function of the mitotic spindle. Nature 434: 1152–1157.
49. Li X, Chen J, Hu X, Huang Y, Li Z, et al. (2011) Comparative mRNA and microRNA expression profiling of three genitourinary cancers reveals common hallmarks and cancer-specific molecular events. PLoS One 6: e22570.