UNIVERSITÀ
DI TORINO

Department of Medical Sciences
PhD in Biomedical Sciences and Oncology

Omics Approaches to Kidney and Urological Disease

Tutor:
Prof. Giuseppe Matullo
Prof. Simone Sanna-Cherchi

Candidate:
LIM Tze Yin

PhD Coordinator:
Prof. Andrea Graziani

XXXVIII cycle
Academic years: 2022-2025
Science-disciplinary sector: MED/03

Table of Contents

Abstract

A central challenge in human genetics is to understand how diverse molecular processes ranging from inherited genomic variation to dynamic regulatory states combine to shape disease susceptibility, heterogeneity, and progression. This thesis addresses two broad questions motivated by that challenge: **1) Can integrative multi-omic profiling of peripheral blood enable non-invasive detection of disease-relevant molecular programs in Bladder Cancer (BCa), despite the complexity and distance from the primary tumor tissue? 2) To what extent do structural variants contribute to the missing heritability of Idiopathic Nephrotic Syndrome (INS), a rare kidney disorder characterized by marked phenotypic and genetic heterogeneity?** Across these studies, the overarching goal is to determine how high-dimensional molecular data can be leveraged to uncover hidden biological signals underlying disease susceptibility, advancing our knowledge beyond what is currently known.

These questions address key gaps in existing BCa and INS research. For BCa, blood-based biomarkers remain difficult to identify because peripheral signals are often weak, noisy, or confounded. Existing blood-derived BCa studies also typically rely on isolated single-omic data type, limiting power to detect shared biological pathways. Multi-omic integration has the potential to overcome these barriers, but its application to blood-derived bladder cancer data has not been explored. In nephrotic syndrome, most genetic insights have centered on either monogenic forms or common variants identified through GWAS, especially in the pediatric and steroid resistant INS populations. However, the contribution of copy-number variation, a known source of large-effect pathogenic alleles has not been systematically characterized at scale, leaving a substantial fraction of heritability unexplained. Therefore, addressing these gaps is essential for both mechanistic insight and future translational applications.

This thesis makes two contributions. First, it presents the first blood-based multi-omics study of bladder cancer, integrating genome-wide genotyping, transcriptomics, and DNA methylation data with Multi-omics Factor Analysis (MOFA), using features that were guided by a genome-wide meta-analysis of BCa (18,559 cases and 1,075,650 controls) (**Chapter 1**). This approach predicted latent factors that discriminate cases from controls,

implicate biologically relevant pathways, and capture components of disease risk beyond germline genetic variation. These findings demonstrate that peripheral blood may encode meaningful information about BCa when high-dimensional signals are combined appropriately.

Second, the thesis reports, to our knowledge, the largest systematic characterization of copy number variation (CNV) in INS to date, spanning 3,144 INS cases across different ages of disease onset and response to therapy (**Chapter 2**). The study identifies Genomic Disorder CNVs (GD-CNV), rare CNV harboring INS-associated genes, and large but rare CNVs, explaining their contribution to unexplained heritability, and uses an analytical framework which has previously been described[1] for large-scale CNV discovery in this rare disorder. The analysis revealed that there is a relatively low contribution of CNV to the diagnostic rate of NS (~ 3%), but should not be overlooked since these variants add to the overall genetic diagnostic workup and have direct implication in clinical management for INS at the individual level.

This thesis is organized into three chapters. The first two chapters present the main studies in BCa (**Chapter 1**) and INS (**Chapter 2**), with each chapter containing its own introduction, methods, results, and discussion. The third chapter (**Chapter 3**) provides an integrative discussion that brings together findings from both studies and reflects on the broader lessons learned.

Acknowledgement

I thank my mentors Prof. Giuseppe Matullo and Prof. Simone Sanna-Cherchi for their invaluable guidance throughout my PhD studentship, and I am especially grateful for the support that has shaped my growth as a scientist.

To mentors and lecturers who once took a chance on me, Ms. Ong Chia Sui at Multimedia University, Mr. Lee Weng Wah at ACGT. Dr. Janeck Scott-Fordsmand at Aarhus University. Drs Ingo Hein, Eleanor Gilroy, Glenn Bryan, Linda Milne and David Marshall at the James Hutton Institute. Thank you.

I am sincerely grateful to everyone who contributed to the Turin Bladder Cancer Study and the Nephrotic Syndrome study. This work builds upon their invaluable efforts.

My gratitude also goes to colleagues at the Matullo laboratory. Alessandra Allione, Elisabetta Casalone, Carla Debernardi, Alessia Russo, Elton Herman Jalis, Khadija Sana-Hafeez, Cecilia Di Primio, Sara Devito, Federica Ragno, Ilaria Carelli, Rebecca Filomena, Chiara Catalano, Miriam Roselli and Angelo Savoca, for bringing an exciting dose of Italian research life into my time in Turin. I thank Giovanni Birolo who is a tremendous help with all things related to the Occam server, and to Prof. Francesco Soria from the Dept. of Urology, whose insights on Bladder Cancer helped shaped the direction of my analysis.

Lori, Enzo, Antoni and Simone from Café Movie, and Marissa from Artisti, whose coffee and kindness sustained me through many days in Turin.

To my colleagues at the Sanna-Cherchi Laboratory and beyond in New York, I owe a deep debt of gratitude. Yask Gupta (the bioinformatics guru), Juntao Ke, Elena Martinelli, Qingxue Liu, Massiel Baez-Belen, the year 2025 has been both tumultuous and rewarding, and your resilience and dedication to our shared goals are my inspiration. To Tanya Sezin, Nick Steers and Danielle Herb, thank you for the delightful conversations about science and life. To Atlas Khan, for the scientific banter. To Kathy McVeigh, for ensuring that all of this was workable.

Kavita, Calvin, Zarina, Miriam and Priya, thanks for being there.

Finally, I thank my family, my partner, Leo R., and his family, whose unwavering support, patience and encouragement have been the true foundation of this PhD journey. This thesis is as much a reflection of their steadfast belief in me as it is of my own efforts, and I am immensely grateful for the countless ways they have helped make this possible.

This page is intentionally set at 1.15 spacing to accommodate the many individuals I wish to thank.

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
accuracy R^2	Imputation info score R2
aCGH	Array Comparative Genomic Hybridization
ACMG	American College of Medical Genetics
ADME	Drug absorption, distribution, metabolism and excretion
ASR	Age-standardize rate
aSRNS	Adult SRNS
aSSNS	Adult SSNS
AUA	American Urological Association
AUA/SUO	American Urological Association/Society of Urologic Oncology
AUC	Atypical urothelial cells
AUC	Area under the curve
BAF	B allele frequency
BCa	Bladder cancer
BCG	Bacillus Calmette-Guérin
Beta	Beta effect size
BTA	Bladder Tumor Antigen assay
CAKUT	Congenital anomalies of the kidney and urinary tract
cfDNA	Cell-free DNA
CHR	Chromosome
CI	Confidence interval
c-index	Concordance index
CKD	Chronic kidney disease
CMA	Chromosomal microarray
CMT	Charcot-Marie-Tooth
CNF	Congenital nephrotic syndrome of the Finnish type
CNV	Copy number variations
CSV	Complex structural variants
CUETO	Spanish Urological Club for Oncological Treatment
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources
DEL	Deletion
DGE	Differential gene expression
DGV	Database of Genomic Variants
DUP	Duplication
EA	Effect allele
EAU	European Association of Urology
EORTC	European Organisation for Research and Treatment of Cancer
eQTL	Expression quantitative trait loci
ESKD	End stage kidney failure
FA	Factor analysis
FDA	U.S. Food and Drug Administration
FDR	False discovery rate
FISH	Fluorescence in situ hybridization

FN	False negative
FP	False positive
FPE	Foot process effacement
Freq	Allele frequency
FRNS	Frequent relapses nephrotic syndrome
FSGS	Focal segmental glomerulosclerosis
GBM	Glomerular basement membrane
GD-CNV	Genomics disorder CNVs
GSA	Global Screening Array
GWAS	Genome wide association studies
HetPVal	Heterogeneity P-value
HG	High-grade
HGUC	High-grade urothelial carcinoma
HGVSC	Human Genome Structural Variation Consortium
HMM	Hidden Markov model
HPRC	Pangenome graph reference
HRC	Haplotype Reference Consortium
HWE	Hardy weinberg equilibrium
I^2	Proportion of heterogeneity
ICGC	International Cancer Genome Consortium
ICI	Immune checkpoint inhibitor
IgAN	Immunoglobulin A nephropathy
INS	Idiopathic Nephrotic Syndrome
IQR	Interquartile range
IRB	Institutional Review Board
ISCA	International Standards for Cytogenomic Arrays
JDR	Joint Dimensionality Reduction
KM	Kaplan–Meier analysis
LG	Low grade urothelial carcinoma
LGUN	Low-grade urothelial neoplasm
LRR	Log R ratio
LRS	Long-read sequencing
LUTS	Lower urinary tract symptoms
MAF	Minor allele frequency
MCD	Minimal change disease
MEGA	Illumina Multi-Ethnic Global Array
MIBC	Muscle invasive bladder cancer
MLPA	Multiplex ligation-dependent probe amplification
MOFA	Multi-omics factor analysis
mQTL	Methylation quantitative trait loci
MVAC	Neoadjuvant methotrexate, vinblastine, doxorubicin plus cisplatin
NHGUC	Negative for high-grade urothelial carcinoma
NMIBC	Non-muscle invasive bladder cancer
NVH	Non-visible hematuria
OR	Odds Ratio
PC	Principal components

PFS	Progression-free survival
POS	Position
pQTL	Protein quantitative trait loci
PRS	Polygenic risk scores
pSRNS	Pediatric SRNS
pSSNS	Pediatric SSNS
PUNLMP	Papillary urothelial neoplasm of low malignant potential
qPCR	Quantitative PCR
RASi	Renin-Angiotensin System inhibitors
RFS	Recurrence-free survival
RHD	Renal Hypodysplasia
RMSE	Root Mean Squared Error
ROC	Receiver operating characteristic curves
ROHM	Risk of high-grade malignancy
SDNS	Steroid-dependent nephrotic syndrome
SE	Standard error
SMRT	Real-time long read sequencing platforms
SRNS	Steroid resistant nephrotic syndrome
SRS	Short-read sequencing
SSNS	Steroid sensitive nephrotic syndrome
SV	Structural variants
T1	Tumor invasion of the lamina propria
T2T	Telomere to telomere
Ta	Non-invasive papillary carcinoma
TBCS	Turin Bladder Cancer Study
TBCS-EPICOR	Case-controls derived from TBCS and an additional 290 non-cancer control from the EPICOR study
TCGA	The Cancer Genome Atlas
Tis/CIS	Carcinoma in situ
TMT	Trimodality therapy
TN	True negative
TNM	Tumor, Node, Metastasis
TOF	Tetralogy of Fallot
TP	True positive
TPS	Paris System for Reporting Urinary Cytology
TURBT	Transurethral resection
UC	Urothelial carcinoma
UICC	Union for International Cancer Control
UKBB	UK Biobank
uPCR	Urinary protein to creatinine ratio
VH	Visible hematuria
VUS	Variants of uncertain significance
WHO	World Health Organization
xQTL	Molecular quantitative trait loci

Chapter 1: Multi-omics approaches in Bladder Cancer

1.1. Introduction

1.1.1. Epidemiology and risk factors

Epidemiology

Bladder cancer (BCa) is the 9th highest malignancy worldwide with over 613,791 newly reported cases globally, and the 13th highest cancer mortality worldwide with 220,349 BCa- related deaths attributed in 2022[2]. BCa occurs four times more frequently in males than females with a 9.5 to 2.4 in 100,000 individuals age-standardize rate (ASR) [2,3]. Even though there is a higher prevalence in males, females BCa patients often have poorer survival outcomes, including higher disease recurrence and progression after adjusting for covariates like smoking and time to treatment[4–6]. In 2022, the International Agency for Research on Cancer reported the highest incidence ASR in Southern Europe (16.9 per 100,000) compared to the rest of the European continent, Northern Europe (12.7), Western Europe (12.0), Eastern Europe (8.7), followed by Northern America (11.0) and Northern Africa (9.5). Conversely, the lowest incidence rates were reported in Central and South America, Sub Saharan Africa and Southeast Asia. On the other hand, the highest mortality ASR was recorded in Northern Africa with 5.3 per 100,000 ASR and Southern Europe 3.4 ASR.

Risk factors

Several risk factors contributing to the differences in incidence and mortality ASR have been reported. These include tobacco smoking, e-cigarette smoking, environmental or occupational exposure to aromatic amines, arsenic levels in drinking water, parasitic infection by *Schistosoma haematobium*, age, gender, and genetic predisposition[7].

Tobacco smoking is the most common risk factor for BCa, whereby BCa risk was increased by threefold and contributes to 50% of BCa incidences and 40% of BC-related death [7–9]. The use of electronic cigarettes has also been associated with the development of urothelial hyperplasia[10]. Considering a lag-time of 20 to 30 years between tobacco

exposure and BCa diagnosis, the age-standardized incidence of BCa is higher in regions in which tobacco smoking was higher in the 1980s, such as in Spain and in Italy, with a reported 36.7 per 100,000 incidence rates in 2003 and 33.2 per 100,000 incidence rates in 2007[11,12]. Reduction of smoking prevalence in 27% of men and 38% of women, have potentially decreased the rates of diagnosis and deaths in these countries[13–15].

Occupational exposure also plays a major role: prolonged contact with aromatic amines such as benzidine, beta-naphthylamine common in industrial dye, rubber and paint can substantially increase risk for BCa, especially when combined with smoking[8]. In certain regions, the consumption of arsenic-contaminated food or water has been linked to an increased BCa risk. Studies[7,16] have reported such associations in Argentina, Chile and Bangladesh. Furthermore, prolonged contact with chemicals like diesel and gasoline engine exhaust or emission from stationary power plants may also contribute to BCa risk.

Parasitic infection with *schistosomiasis* is another notable risk factor, particularly contributing to the development of squamous cell carcinoma in parts of Northern Africa[17]. In Egypt, a large study involving 9,843 patients spanning the period of 1970 to 2007, reported a decline in *S. haematobium* infection, which paralleled a reduction in squamous cell carcinoma cases, alongside a rise in urothelial cell carcinoma incidence and an increase in the median age of patients[18,19].

The two main non-modifiable risk factors in the development of BCa are age and gender with a higher prevalence in men (male-to-female ratio at 4:1). It is also more commonly reported in older adults with the majority of cases occurring in patients who are > 55 years old and have an average age of diagnosis at 73 years[15].

Finally, genetic predisposition represents an important contribution. While the majority of BCa cases are sporadic, a subset arises in the context of hereditary cancer syndromes. For example, hereditary non-polyposis colon cancer (Lynch syndrome) accounts for approximately 5% of the development of upper tract urothelial carcinomas. This syndrome is characterized by pathogenic variants in mismatch repair genes such as *MLH1*, *MSH2*, *MSH6* and *PMS2*, leading to microsatellite instability[20]. A separate section describing somatic mutations identified in tumor samples is provided in section 1.1.3.

1.1.2. Tumor classification, staging and pathology

The urothelium encompasses three major cell layers: basal, intermediate and superficial cell layer, the latter serving as the crucial barrier that prevents urine reabsorption and protects from infections. Studies suggest that the human basal cells have the capacity to regenerate and, interestingly, both basal and intermediate cells have been implicated as the potential cells of origin for bladder tumors in mouse models. The peroxisome proliferator-activated receptor γ (PPARγ) is a transcription factor that regulates urothelial differentiation by controlling the expression of uroplakins, keratins and claudins, all of which are critical for maintaining the barrier function of the urothelium. PPARγ modulates transcription factors such as FOXA1, GATA2 and ELF3; in its absence, p63 maintains the basal identify of the cell[21].

BCa exhibits significant molecular heterogeneity and can be dichotomized into two categories: Non-muscle invasive bladder cancer (NMIBC) and muscle invasive bladder cancer (MIBC) depending on the depth of invasion of the carcinoma during transurethral resection (TURBT) or biopsy[22]. The American Urological Association (AUA) and European Association of Urology (EAU) primarily rely on the TNM (Tumor, Node, Metastasis) system for staging bladder cancer, with staging focused on the depth of tumor invasion in the bladder wall. This framework stratifies the disease according to tumor characteristic (T), lymphatic spread (N) and presence of distant metastasis (M), guiding the appropriate treatment decisions that ranged from TURBT for NMIBC to radical cystectomy for MIBC cases. NMIBC is classified into stages Ta (non-invasive papillary carcinoma), Tis (carcinoma in situ), and T1 (tumor invasion of the lamina propria). MIBC begins at stages T2 where the tumor invades the muscularis propria, further subdivided into T2a and T2b for inner and outer muscle invasion respectively. The more advanced stages include T3, defined as the tumor invasion to perivesical tissue and T4, where the tumor spreads beyond the bladder to adjacent tissues and organs. Within NMIBC, T1 tumors display molecular features that more closely resemble MIBC tumors, whereas Ta tumors generally follow a more benign disease course compared to Tis and T1[22,23].(Table 1)

T - Primary tumor	
TX	Primary tumor cannot be assessed
T0	No evidence of primary tumor
Ta	Non-invasive papillary carcinoma
Tis	Carcinoma in situ: ‘flat tumor’
T1	Tumor invades subepithelial connective tissue
T2	Tumor invades muscle
	T2a Tumor invades superficial muscle (inner half)
	T2b Tumor invades deep muscle (outer half)
T3	Tumor invades perivesical tissue
	T3a Microscopically
	T3b Macroscopically (extravesical mass)
T4	Tumor invades any of the following: prostate stroma, seminal vesicles, uterus, vagina pelvic wall, abdominal wall
	T4a Tumor invades prostate stroma, seminal vesicles, uterus or vagina
	T4b Tumor invades pelvic wall or abdominal wall
N – Regional lymph nodes	
NX	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastasis
N1	Metastasis in a single lymph node in the true pelvis (hypogastric, obturator, external iliac, or presacral)
N2	Metastasis in multiple regional lymph nodes in the true pelvis (hypogastric, obturator, external iliac, or presacral)
N3	Metastasis in common iliac lymph node(s)
M - Distant metastasis	
M0	No distant metastasis
	M1a Non-regional lymph nodes
	M1b Other distant metastases

Table 1: 2017 pTNM tumor staging system according to the Union for International Cancer Control (UICC)[23].

Histologically, bladder carcinoma can be categorized according to the World Health Organization (WHO) grading systems: WHO 1973, WHO 2004/2016 and WHO 2004/2022. The WHO 1973 histological classification introduced a three-grade system based primarily on cell type, cell sizes and overall tissue appearance[24] (Table 2).

Grade	Description
Grade 1	Tumors with the least degree of cellular anaplasia compatible with a diagnosis of malignancy
Grade 2	Histological features that lie between grades 1 and 3
Grade 3	Tumors with the most severe degrees of cellular anaplasia

Table 2: WHO 1973 grading system for bladder cancer. Table adapted from Mostofi et al., 1973[24].

This was later revised in 2004 to incorporate papillary urothelial neoplasm of low malignant potential (PUNLMP), non-invasive low grade urothelial carcinoma (LG) and high-grade urothelial carcinoma (HG). The latter was subsequently integrated into the WHO 2004/2016 and 2004/2022 classifications, to complement and update the original 1973 grades[25][Figure 1]. Despite these important revisions, histological grading remains challenging. Even when criteria are clearly defined, inter-observer reproducibility amongst pathologists is poor, leading to inconsistent prognosis. This difficulty reflects the inherent heterogeneity of tumor morphology which tends to fall along a spectrum rather than within discrete categories, making prognosis based solely on histological grading challenging[25–27].

Low grade carcinomas tend to be well-differentiated, prone to recurrence, and are generally associated with favorable patient outcomes. High-grade carcinomas may present as either NMIBC or MIBC, with NMIBC patients at increased risk for progression to muscle invasion and metastasis[28].

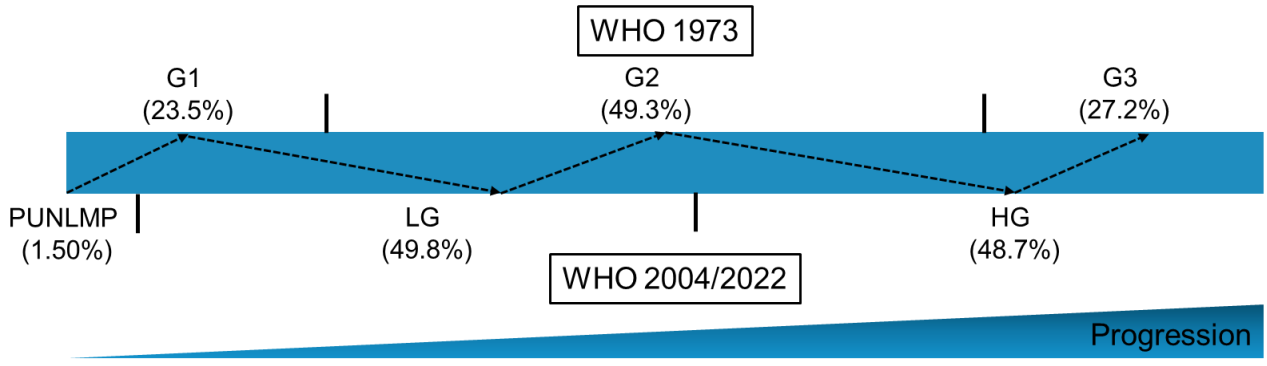


Figure 1: Graphical representation of tumor grading according to the WHO 1973 and WHO 2004/2022 classification. The dotted line and arrow indicate the grade shift in

Ta/T1 bladder tumors from the WHO 1973 (G1 to G3) to the WHO 2004/2022 (PUNMP, LG and HG) classification, with 5-year progression rates increasing from LG/G1, LG/G2, HG/G2 to HG/G3 patients. Percentages refer to the distribution of 5,145 primary Ta-T1 patients across the corresponding grades[29]. Image adapted from the EAU Guidelines on NMIBC (TaT1 and CIS)[30].

A multi-center study[29] involving 5,154 patients with primary Ta/T1 non-muscle invasive bladder cancer (NMIBC) found that both the WHO 1973 and WHO 2004/2016 grading systems were predictive for disease progression, but not recurrence. The WHO 1973 classification demonstrated stronger prognostic value for progression, with a concordance index (c-index) of 0.71 compared to 0.67 for WHO 2004/2016. Notably, combining both grading systems further improved prognostic accuracy, yielding a c-index of 0.73.

Pathologically, over 90% of BCa has the urothelial carcinoma (UC) histological subtype (or variant), arising from the urothelial lining of the blader. Less common histological types include squamous cell carcinoma (~3-5% cases), Adenocarcinoma (< 2%), small cell carcinoma and sarcoma (<1% each)[31–33]. The WHO Classification of Tumors of the Urinary System and Male Genital Organs subdivides urothelial carcinoma into various subtypes including micropapillary, plasmacytoid, nested, and lymphoepithelioma-like variants[28,34].

1.1.3. Molecular landscape

Genetic alterations are commonly observed in BCa, these include somatic mutations and gene amplifications in *FGFR3*, *TP53*, *TERT*, *PIK3CA*. Overall, MIBC exhibits a higher mutational burden than NMIBC[35,36] and is more often characterized by large structural variants. Somatic mutations driven by the Apolipoprotein B mRNA Editing Catalytic Polypeptide-like (APOBEC) family of cytidine deaminases account for over 60% of single-nucleotide mutations in these tumors. Chromosomal deletions on 9p and 9q occur in approximately 50% of both NMIBC and MIBC patients. Additionally, copy number variations (CNV) are subtype-specific. In NMIBC, 8-22% of patients carry copy number gains on 1q,5p,18q,20p and 20q, as well as copy number loss on 8p,11p,17p and 18q. In contrast, MIBC patients more commonly harbor amplifications in genomic regions encompassing *PPARG*, *E2F3*, *EGFR*, *ERBB2* and *CCNE1*[37–40]. The only recurrent

mutations consistently observed across all BCa tumors and stages involve the promoter region of the telomerase reverse transcriptase, *TERT*, ranging from 70% to 80% (Figure 2).

In NMIBC, point mutations activating the RAS-MAPK signalling (*FGFR3*, in 60% of patients), PI3K signalling (*PIK3CA,* in 30%) and *RAS* genes are common[41]. About 90% of Ta tumors carry either a *RAS* or *FGFR3* mutation, though the two are seldom observed together. Additionally, gain-of-function mutations activating PI3K signalling through *ERBB2* and *ERBB3* are also observed in 15% of stage T1 tumors. Inactivating mutations in the cohesion complex tumor suppressor gene, *STAG2*, and chromatin regulators such as *KDM6A*, *KMT2S*, *KMT2C*, *CREBBP*, *EP300* and *ARID1A* also characterize approximately 65% of NMIBC patients. *KDM6A* mutations are more prevalent in stage Ta, whereas ARID1A mutations are more common in stage T1 tumors[42] (Figure 3).

On the other hand, MIBC tumors are characterized by greater genetic instability, typically exhibiting complex copy number alterations and loss of cell cycle checkpoint controls due to the disruption of TP53, RB1 and ATM[43]. Genomic amplification frequently occurs in regions containing genes associated to oncogenic function such as *E2F3*, *MDM2* and *HER2*, while homozygous deletions are commonly found in chromosome 9p. Of particular importance is the chromosome 9p21.3 locus which harbors the cyclin-dependent kinase inhibitor 2A (*CDKN2A*) gene. This gene encodes two proteins, p16 and $p14^{ARF}$ that play central roles in tumor suppression by regulating cell cycle. Rebouissou et al. (2012)[44] reported that the inactivation of CDKN2A through homozygous deletion is associated to tumor progression when co-occuring with *FGFR3* mutations, regardless of stage and grade. Several other pathways implicated in MIBC include the DNA damage and DNA repair pathways involving *ERCC2* or *ATM*, MET signalling and NOTCH pathways.

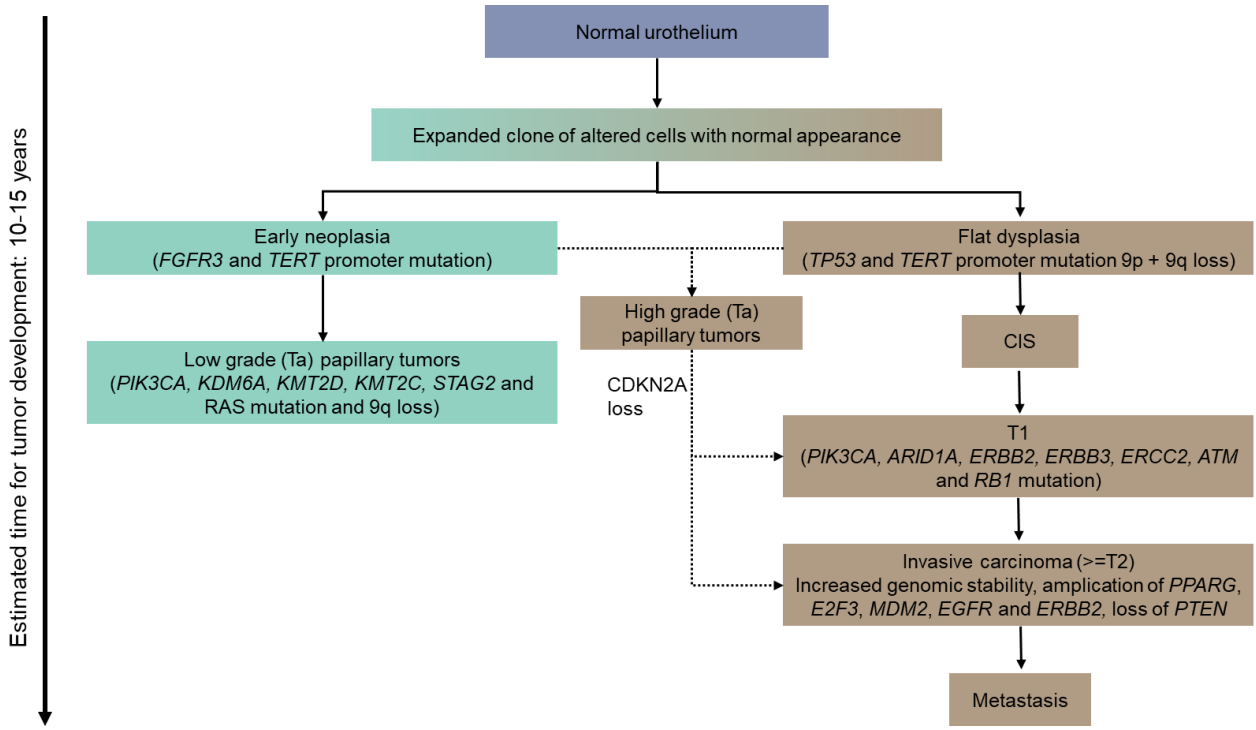
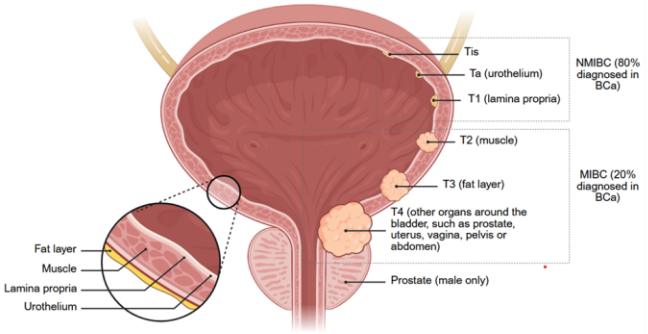


Figure 2: NMIBC vs MIBC pathogenesis pathways reproduced[41]. Solid arrow indicates pathways with histopathological and/or molecular evidence whereas dashed arrows indicate pathways with uncertainty.



NMIBC subtype	Signature	Mutations
Class 1 (20%, 97/476)	• PPARG+, UPK+ • Early cell cycle genes	*FGFR3*
Class 2 (52%, 249/476)	• Liminal-like differentiation • PPARG+, UPK+, KRT14+ • CIS positive • EMT transcription factors • Cancer stem cell activity • Late cell cycle genes • APOBEC+ signature	*TP53* *ERCC2*
Class 3 (27%, 130/476)	• Basal like undifferentiation • PPARG-, GATA3+, KRT5+, KRT14+, KRT15+, CD44+ • RNA-editing signature	*FGFR3*
MIBC subtype	**Signature**	**Mutations**
LumP (24%)	• PPARG+, FGFR3+, CDKN2A-	*FGFR3* (40%) *KDM6A* (38%)
LumNS (8%)	• PPARG+	*ELF3* (35%)
LumU (15%)	• PPARG+, E2F3+, ERBB2+ • Genomically unstable • Cell cycle positive • APOBEC+ • High TMB	*TP53* (76%) *ERCC2* (22%)
Stroma-rich (15%)	• Smooth muscle • Fibroblast • Myofiblast gene signatures	
Ba/Sq (35%)	• Squamous differentiation markers • Fibroblasts and myfibroblast gene signature • EGFR+	• *TP53* (61%) • *RB1* (25%)
NE-like (3%)	• Neuroendocrine differentiation marker • TP53-, RB1- • Cell cycle positive	• *TP53* (94%) • *RB1* (39%)

Figure 3: Bladder cancer staging and molecular subtypes of NMIBC and MIBC adapted from Tran et al. 2021[43], the proportion of MIBC subtypes were derived from 1,750 tumor samples[45]. The median overall survival years decrease from the luminal subtypes to the basal subtypes in MIBC with LumP (4 years), LumNS (1.8), LumU (2.9), Stroma-rich (3.8), Basal (1.2), and NE-like (1 year). Ba/Sq: basal/squamous; LumNS: luminal nonspecified; LumP: luminal papillary; LumU: luminal unstable; NE: neuroendocrine. Image created from BioRender.com.

1.1.4. Diagnosis and treatment

Hematuria, the presence of blood in urine, is a well-recognized indicator of urinary tract disease and is observed in approximately 75% of patients with suspected bladder cancer. It is classified as either visible hematuria (VH), also referred to as “macroscopic” or “gross” or non-visible hematuria (NVH), also termed “microscopic”. The IDENTIFY study, a large prospective study of over 10,000 patients across 26 countries found that VH, along with other risk factors like age and smoking status (particularly in former and current smokers) are strongly associated with BCa. Among patients presenting with VH, the prevalence of BCa was 22.4%, compared with 5.23% in those presenting with NVH. Age further influenced risk in patients with VH, as BCa was detected in 4.7% of patients younger than 35 years, while in 30.6% of those older than 75. By contrast, among patients with NVH, no BCa cases was reported in patients under 35, while 10.6% of those over 75 were diagnosed with BCa[46]**.** Moreover, in newly diagnosed BCa patients, presentation with VH at diagnosis has been associated with more advanced pathological stage, whereas presentation with NVH (microscopic hematuria) tend to correspond to lower disease stage. In an observational study of 1,284 newly diagnosed adult BCa patients with hematuria, Ramirez et al.[47] reported that high-grade disease was present in 64% of BCa patients with VH and 57.1% of those with NVH. The stage distribution suggested that NVH presentations were more often associated with earlier-stage disease. Ta/CIS tumors were observed in 68.8% of NVH cases and 55.9% of VH cases; T1 tumors in 19.6% for both groups. Conversely, advanced tumors (>=T2) are more frequently seen in VH (17.9%) than NVH (11.6%).

Unlike adults, the incidence of BCa is rare in pediatric cases (less than 20 years old), accounting for only 0.1 to 0.4% of all BCa diagnoses[48]. A systematic review involving 243 pediatric cases reported that hematuria. VH was the predominant presenting symptom (75.6%), followed by a combination of hematuria and lower urinary tract symptoms (LUTS, 8.6%), and abdominal pain (3.4%). Most pediatric patients presented with NMIBC, with Ta tumors accounting for the majority of the cases, 86.4%, whereas advance-stage (>= T2) tumors are less common, occurring in 4.1%[49].

The goal standard for diagnosing BCa is through cystoscopy. If a lesion is detected, TURBT will be performed to obtain tissues for histopathological grading and TNM staging to inform subsequent treatment decisions[50]. Orthogonal tools such as urine cytology and biomarker-based assays are increasingly used as non-invasive test to improve diagnostic accuracy. Among these, urine cytology, which refers to the procedure of inspecting cells in the urine, remains the gold standard due to its cost-effectiveness and high specificity in detecting high grade BCa[51,52]. Cytology specimens are classified based on the Paris System for Reporting Urinary Cytology (TPS)[51], which defines seven diagnostic categories: nondiagnostic, negative for high-grade urothelial carcinoma (NHGUC), atypical urothelial cells (AUC), suspicious for high-grade urothelial carcinoma (SHGUC), high-grade urothelial carcinoma (HGUC), low-grade urothelial neoplasm (LGUN), and other malignancies. In a meta-analysis of 28 studies, Nikas et al.[52] evaluated diagnostic accuracy using the pooled risk of high-grade malignancy (ROHM) across TPS diagnostic categories. Their findings showed that urine cytology performed particularly well in the HGUC category, with ROHM exceeding 90%. Overall, pooled performance metrics demonstrated a sensitivity of 66.9%, specificity of 89.9% and an AUC of 0.849, indicating robustness for high-grade disease. As evidenced in multiple studies, urine cytology yields notably lower sensitivity for low grade urothelial carcinoma primarily due to the challenges in differentiating these lesions when they exhibit patterns and cellular configurations that are similar to normal cells[53].

Extensive efforts have focused on developing more protein-based and urine-based marker tests to improve bladder cancer diagnosis. These have led to several FDA-approved tests including the Bladder Tumor Antigen assay (BTA)[54], which detects

elevated levels of human complement factor H-related protein in urine, Nuclear Matrix Protein 22 (NMP22) assay[55], which measures the nuclear matrix protein released into the urine from dying tumor cells, cell-based test such as the UroVision FISH[56] uses a multicolor fluorescence in situ hybridization (FISH) assay that identifies chromosomal aneuploidy in chromosomes 3, 7, 17 and 9p21 from exfoliated urothelial cells, ImmunoCyt an immunofluorescent assay that uses antibodies to detect tumor-associated mucin and carcinoembryonic antigens from urine cells. More recent approaches include cell-free DNA test such as Bladder CARE[57], which uses quantitative PCR to assess DNA methylation biomarkers, and mRNA-based assays such as CxBladder[58] and Xpert Bladder Cancer Monitor[59], which use qPCR to measure multiple mRNA markers in urinary cells. Collectively, these assays demonstrate higher sensitivity for BCa detection compared to cytology but at the expense of lower specificity, leading to false positives (Table 3).

Multiplex urinary marker	Target	Overall sensitivity*	HG	Overall specificity*	HG	N studies/patients
XPERT BC MONITOR	5 mRNAs	52-91	79-100	41-91	76-91	11 studies, 2800 patients
EpiCheck	15 DNA methylations	62-90	78-95	82-88	NA	6 studies, 2236 patients
CX BLADDER	5 mRNAs	93	95	61	NA	1 study, 763 patients
UROMONITOR	DNA mutations (FGFR3 + TERT + Kras)	49-93	NA	86-99	NA	5 studies, 1190 patients
Galeas Bladder	Multiple DNA mutations (N = 443 in 23 genes)	86	100	63	NA	1 study, 293 patients

Table 3: Performance of multiplex urine markers (those that measured multiple genetic changes) in disease surveillance. Reproduced from https://uroweb.org/guidelines/non-muscle-invasive-bladder-cancer/chapter/followup-of-patients-with-nmibc, accessed on 10 September 2025. *referred to the minimal and maximum range calculated across studies.

Treatment options for BCa vary depending on tumor stage, symptom severity and patient's overall health status. The main modalities include TURBT, intravesical therapy,

radical cystectomy, trimodality therapy (TMT), and perioperative systemic therapy. The standard initial management of BCa typically involves TURBT, which aims to achieve complete removal of the tumor in the case of NMIBC and proper local staging to optimize subsequent treatment like radical cystectomy or trimodality therapy in the case of MIBC.

Following TURBT, intravesical chemotherapy, particularly the Bacillus Calmette-Guérin (BCG) therapy, is widely used to reduce the risk of recurrence and progression in patients with NMIBC[60]. Adverse effects of BCG have been reported and include inflammation, infection of bladder, malaise, fever and BCG sepsis. A trial conducted by the European Organization for Research and Treatment of Cancer (EORTC) found that the overall rate of adverse events in patients treated with BCG is approximately 70%, with 8% of patients discontinuing treatment due to toxicity[61]. Since 2013, intermittent shortages of BCG have posed a global challenge, prompting an urgent need for alternative treatment strategies such as the use of intravesical maintenance chemotherapy[62]. Alternative modes of intravesical treatment include gene therapy with nadofaragene firadenovec[63] and systemic immune checkpoint inhibitor (ICI) therapy with pembrolizumab[64], which have been approved by the FDA for patients with BCG-unresponsive NMIBC with CIS.

Radical cystectomy is a standard of care for patients with localized MIBC as well as for those with NMIBC who are unresponsive to BCG. The surgical procedure consists of three key components involving cystectomy, pelvic lymph node dissection and urinary diversion. The surgical approach for cystectomy differs by sex. In men, radical cystectomy entails removal of the bladder, prostate, seminal vesicles and distal ureters. In women, it involves removal of the bladder, entire urethra, anterior vaginal wall, uterus and distal ureters.

In certain circumstances, MIBC patients who are medically-unfit for radical cystectomy may be offered an option for trimodality bladder-sparing treatment. This treatment is most favorable in patients who meet certain criteria, which include a predominant urothelial carcinoma histology at stages T2 to T3a, a unifocal tumor measuring less than 7cm, and a visibly complete TURBT. Additional criteria include the absence of extensive CIS, no

presentations of hydronephrosis and the preservation of good bladder function[65,66]. This treatment involves a maximal TURBT followed by concurrent radio-sensitizing chemotherapy and radiotherapy, otherwise known as chemoradiotherapy. The most common regiments are cisplatin-based, followed by carboplatin, gemcitabine, paclitaxel and 5-FU+mitomycin[67]. After treatment, lifelong surveillance is necessary to detect recurrences or second primary tumors. Approximately 10 to 15% of patients may require salvage cystectomy, which carries a higher risk of overall and major late complication compared to primary cystectomy[68].

Additionally, perioperative systemic therapy is recommended for locally advanced MIBC patients at higher risk of metastatic recurrence. Two pivotal randomized trials have shown the benefit of neoadjuvant cisplatin-based chemotherapy in patients with MIBC, specifically in those with clinical stage T2-T4aN0M0 disease. Neoadjuvant therapy refers to the treatment given before the primary intervention. In this context, it involves chemotherapy administered prior to cystectomy. The BA06 30894 trial, the largest study to date including 976 MIBC patients, evaluated cisplatin, methotrexate, and vinblastine before definitive local therapy and reported a significantly improved survival[69]. Similarly, the SWOG 8710 trial involving 307 patients demonstrated that neoadjuvant methotrexate, vinblastine, doxorubicin plus cisplatin (MVAC) prior to cystectomy improved overall survival compared with surgery alone[70]. More recently in 2024, the phase III NIAGARA trial demonstrated that adding perioperative durvalumab to neoadjuvant gemcitabine-cisplatin significantly improved event-free and overall survival at 24 months compared with neoadjuvant chemotherapy alone at 82.2% and 75.2%, respectively[71].

In cases of advanced metastatic MIBC disease, or when patients are ineligible for cisplatin, systemic therapy is employed as the primary treatment approach. Although randomized clinical trials have demonstrated a survival benefit with using MVAC, this cisplatin-based chemotherapeutic combination is associated with significant toxicity including nephrotoxicity, nausea, vomiting, myelosuppression and neurotoxicity[72]. These have led to the adoption of alternative regimens such as the gemcitabine-cisplatin combination which have yielded similar efficacy but have less reported toxicity[73]. In 2015,

immune checkpoint inhibitors (ICI) targeting PD-1 (Programmed cell death protein 1) and PD-L1 (Programmed death- ligand 1), such as pembrolizumab and atezolizumab, demonstrated objective responses in approximately 20-25% patients with metastatic MIBC. These were subsequently approved for use in patients who experienced disease progression even after platinum-based chemotherapy[74] or as a first-line treatment option for those who are ineligible for cisplatin[64,75].

1.1.5. Risk stratification based on EAU guidelines on NMIBC

Risk stratification models based on clinical variables

NMIBC accounts for over 70% of newly diagnosed BCa cases. These tumors are heterogenous, often leading to variable clinical outcomes, which necessitates the development of risk stratification tools to guide management. Typically, patients are stratified into low-, intermediate-, and high-risk groups, reflecting good, intermediate and poor prognosis. Following TURBT and a single immediate instillation of chemotherapy, subsequent treatment is tailored based on prognosis whereby patients with a good prognosis may receive no further instillation or additional intravesical chemotherapy, whereas those with a poor prognosis are generally treated with BCG, maintenance therapy or cystectomy[76]. The most prominent models include those from the EORTC[76], Spanish Urological Club for Oncological Treatment (CUETO)[77], EAU[78] and the American Urological Association/Society of Urologic Oncology (AUA/SUO) models[79](Table 4). The EORTC and CUETO models are scoring systems developed based on clinical trial data that account for patient and tumor characteristics to estimate the probability of two endpoints, specifically recurrence and progression at one and five years. Model accuracy can be evaluated using Harrell's bias-corrected C-index, which assesses a risk stratification tool's ability to distinguish patients who experience a particular outcome, such as recurrence and progression, from those who do not. The C-index ranges from zero to one, with 0.5 implying a poor model with no predictive ability beyond random chance and 1.0 representing perfect model discrimination[80].

Risk Model	Variables
EORTC	Number of tumors, Tumor size, T stage, Tumor grade (WHO 1973), Prior recurrence rate, Presence of concurrent carcinoma in situ.
CUETO	Age, Gender, Number of tumors, Recurrent tumor, T stage, Tumor grade (WHO 1973), Presence of concurrent carcinoma in situ.
EAU	Age, Number of tumors, Tumor grade (WHO 1973 or 2004/2016), T stage, Recurrent tumor, Tumor size.
AUA/SUO	Tumor size, Number of tumors, Tumor grade (WHO 2004/2016), and T stage, Lymphovascular invasion, High grade prostatic urethral involvement, Variant histology, BCG failure in high-grade tumors, Recurrent tumor.

Table 4: Current risk stratification models and variables used , table reproduced from https://auanews.net/issues/articles/2021/july-2021/risk-stratification-in-nonmuscle-invasive-bladder-cancer-a-review-and-glimpse-into-the-future , accessed on September 10 2025.

The original EORTC scoring model, published in 2006, combined data from seven trials involving 2,596 patients with stage Ta and T1 BCa. Key prognostic variables for recurrence included prior recurrence rate, number of tumors, tumor size, while progression was best predicted using the T category, WHO 1973 grade, and presence of CIS variables. Variables such as prior treatment, age and gender were excluded[76]. Based on the coefficient of these variables, each patient received a recurrence score from 0 indicating best prognosis to 17 for worst prognosis, and a progression score from 0 to 23. Patients were then grouped into four risk categories reflecting increasing probabilities of recurrence and progression. The reported C-index for recurrence and progression were 0.66 and 0.75, respectively.

The CUETO model[77] incorporates similar prognostic variables but is uniquely tailored for patients treated with BCG therapy. In addition, CUETO includes patient variables like age and gender. Based on data from 1,062 BCG-treated patients, the model generally predicted lower risk of recurrence and progression compared to the EORTC model. In 2016, EORTC introduced an updated risk group analysis to predict recurrence, progression, disease-specific and overall survival in 1,812 patients with intermediate and

high risk NMIBC treated with induction BCG and 1 to 3 years of maintenance BCG, in line with EAU recommendations[78]. Reported C-indices were 0.56-0.59 for late recurrence, 0.64-0.72 for disease progression, 0.71-0.72 for BCa-specific death and 0.68 for overall survival. Notably, when the CUETO model was applied to this cohort, its performance declined substantially, with C-indexes decreasing from 0.64 to 0.48 for recurrence and from 0.69 to 0.53 for progression. Several studies have been performed to evaluate the ability of the EORTC and CUETO models to predict recurrence and progression in external populations and have yielded variable c-index ranging from 0.52 to 0.66 for recurrence, and 0.62 to 0.81 for progression[81–85].

Up to this point, all risk models (EORTC2006, CUETO, EORTC2016) have relied on the 1973 World Health Organization (WHO) grading system as a prognostic variable. In contrast, the updated EAU2021[27] risk stratifications incorporates both WHO1973 and the more recent WHO2004/2016 grading system. This study included 3,401 patients treated with TURBT with and without intravesical chemotherapy, demonstrated strong predictive performance with a C-index of 0.8 for 5-year progression and 0.79 for 10-year progression. Notably, this study introduced four progression risk categories: Low, Intermediate, High and Very High (Table 5). Patients in very high-risk group accounted for only 2% to 3% of patients studied. Importantly, the probabilities of progression at 5 years increased substantially from 12% in the current EAU high-risk group to 40% and 44% in the newly-defined “Very High” risk group under the WHO 2004/2016 and 1973 grading systems, respectively.

Future research is likely to explore the possibility of integrating current prognostic models with molecular signatures associated with cell cycle, MAPK pathways, apoptosis, tumor microenvironment, chromatin instability and DNA-damage response to improve the predictive capabilities of the existing clinical models[86].

NMIBC risk group	Description
Low	• A primary, single, Ta LG/G1 tumor ≤3 cm in diameter without CIS in a patient ≤70 yr • A primary LG/G1 tumor with at most ONE of the following additional clinical risk factors: • Age > 70 years, Multiple tumors, tumor diameter >=3cm or stage T1
Intermediate	• Patients without CIS who are not included in either the low, high or very high risk groups
High	• All T1 HG/G3 without CIS, except those included in the very high risk group • All CIS patients, except those included in the very high risk group • Stage, grade with additional clinical risk factors* 1. Ta LG/G2 or T1 G1, no CIS with all 3 risk factors 2. Ta HG/G3 or T1 LG, no CIS with at least 2 risk factors 3. T1 G2 no CIS with at least 1 risk factor
Very high risk	• Stage, grade with additional clinical risk factor* 1. Ta HG/G3 and CIS with all 3 risk factors 2. T1 G2 and CIS with at least 2 risk factors 3. T1 HG/G3 and CIS with at least 1 risk factor 4. T1 HG/G3 no CIS with all 3 risk factors

Table 5: The new European Association of Urology NMIBC prognostic factor risk groups based on WHO 2004/2016 or WHO 1973 grading classification systems reproduced from Sylvester et al. 202127. “*” indicates additional clinical risk factors which included age > 70 years, multiple tumors, and a tumor diameter >= 3cm. Abbreviations: CIS: carcinoma in situ; HG: high grade; LG: low grade; TURBT: transurethral resection of bladder tumor.

Genome wide association studies (GWAS) and polygenic risk scores (PRS) in BCa

Several studies have used GWASs to systematically identify variants that influence BCa risk[87–107]. GWAS susceptibility loci account for approximately 12% of familial risk in BCa[87,108]. Recent work have highlighted the potential use of these GWAS findings in establishing PRS to estimate lifetime BCa risk. These PRSs leverage the cumulative effect size of multiple associated risk variants, thereby facilitating disease stratification and enabling opportunities for earlier disease detection. A comprehensive meta-analysis of 32 BCa GWAS involving 13,790 cases and 343,502 controls identified 24 independent risk loci, including six novel associations (6p22.3, 7q36,3, 8q21.13, 9q21.3, 10q221.1 and 19q13.33), underscoring the value of large-scale meta-analysis in expanding the polygenic landscape of BCa[107]. Notably, a significant effect modification was observed in a sex-stratified analysis, with rs2896518 at the 4p16.3 (FGFR3/TACC3) locus conferring a stronger effect in women (OR = 1.34, 95% CI = 1.22-1.47, $P = 1.93 \times 10^{-9}$) than in men (OR 1.12, 95% CI=1.06-1.18, $P = 4.20 \times 10^{-5}$), with a significant multiplicative interaction ($P_{interaction} = 0.002$). Gene-environment interactions were also evident, as carriers of susceptibility variants at 8p22 (*NAT2*), 8q21.13 (*PAG1*) and 9p21.3 (*LOC107987026*/*MTAP*/*CDKN2A*) exhibited increased BCa risk in the context of smoking. Building on these findings, Koutros et al.[107] constructed a PRS from 24 independent genome-wide significant loci, which when combined with smoking history and other risk factors, enabled the stratification of lifetime BCa risk disease (OR = 1.49, CI = 1.44-1.53), revealing a fourfold difference in susceptibility among smokers and non-smokers (Table 6).

SNP	Chr	Band	Position	Gene Region	EA	OR	EAF
GSTM1 Composite*	1	1p13.3	110229772	*GSTM1*	(-), G	1.20	0.50
rs17863783*	2	2q37.1	234602277	*UGT1A* cluster	G	1.75	0.95
rs10936599	3	3q26.2	169492101	*MYNN, TERC*	C	1.10	0.75
rs710521	3	3q28	189645933	*TP63*	T	1.15	0.74
rs2896518	4	4p16.3	1757559	*TACC3, FGFR3*	A	1.17	0.21
rs2242652	5	5p15.33	1280028	*CLPTM1L, TERT*	G	1.18	0.8
rs6910215	6	6p22.3	20783394	*CDKAL1*	C	1.10	0.57
rs72826305	6	6p22.3	21826729	*CASC15/LOC105374970*	C	1.12	0.34

rs2125484	7	7q36.3	155759638	*LOC389602*	G	1.11	0.58
rs1495741*	8	8p22	18272881	*NAT2*	A	1.16	0.60
rs5003154	8	8q21.13	81986953	*PAG1*	C	1.11	0.51
rs10094872	8	8q24.21	128719884	*CASC11, MYC*	T	1.24	0.38
rs2294008	8	8q24.3	143761931	*PSCA*	T	1.14	0.45
rs1414253	9	9p21.3	21755630	*LOC107987026, MTAP/CDKN2A*	A	1.08	0.42
rs4743687	9	9q31.1	106856910	*SMC2*	C	1.10	0.44
rs7076867	10	10q22.1	71582996	*COL13A1*	C	1.31	0.95
rs907611	11	11p15.5	1874072	*TNNT3, LSP1*	A	1.11	0.32
rs7937265	11	11p15.5	1947800	*TNNT3, LSP1*	G	1.13	0.19
rs4907479	13	13q34	113659108	*MCF2L*	A	1.10	0.27
rs10853535	18	18q12.3	43317547	*SLC14A1*	C	1.14	0.44
rs8102137	19	19q12	30296853	*CCNE1*	C	1.12	0.33
rs411482	19	19q13.33	49103447	*SULT2B1-FAM83E*	C	1.13	0.61
rs62185668	20	20p12.2	10961935	gene desert	A	1.11	0.25
rs1014971	22	22q13.1	39332623	*APOBEC3A*	T	1.12	0.65

Table 6: Common risk variants associated with BCa from Koutros et al. 2023. *GSTM1* composite allele refers to the deletion and is tagged with a proxy marker chr1:110229772, *UGT1A* marker frequency refers to the GG genotype; whereas *NAT2* frequency refers to the AA genotype. EA: Effect allele; EAF: Effect allele frequency.

1.1.6. Multi-omics analysis and performance metrics

Multi-omics is an integrative approach that combines multiple layers of biological data such as genomics, transcriptomics, epigenomics, proteomics and metabolomics to provide a comprehensive view of complex systems biology and disease mechanisms[109].

In cancer research, multi-omics integration has been greatly advanced by initiatives from the International Cancer Genome Consortium (ICGC), The Cancer Genome Atlas (TCGA), and European Organisation for Research and Treatment of Cancer (EORTC). These consortia have systematically generated and curated diverse omics datasets across thousands of tumors spanning multiple cancer types. Several data repositories for multi-omics data are available (Table 7).

Repository	Website	Disease	Multi-omics data types available
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/ https://www.cbioportal.org/	Cancer	RNA-Seq, DNA-Seq, miRNA-Seq, SNV,CNV, DNA methylation, and RPPA
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://pdc.cancer.gov/pdc/browse	Cancer	Proteomics data from TCGA cohorts
International Cancer Genomics Consortium (ICGC)	https://dcc.icgc.org/	Cancer	WGS, genomic variations (somatic and germline mutation)
Cancer Cell Line Encyclopedia (CCLE)	https://depmap.org/portal/data_page/?tab=allData	Cancer cell line	RNA-Seq, CNV, WES, WGS, pharmacological profiles of 24 anticancer drugs, Genome-wide knockout screen data
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	https://www.cbioportal.org/study/summary?id=brca_metabric	Breast cancer	Clinical traits, gene expression, SNP, CNV, methylation
TARGET	https://www.cancer.gov/ccg/research/genome-sequencing/target	Pediatric cancer	RNAseq, miRNA, CNV, Sequencing data
Omics Discovery Index	https://www.omicsdi.org/	Cancer and non-cancer	Genomics, transcriptomics, proteomics, metabolomics
NCI GDC Data Portal	https://portal.gdc.cancer.gov/	Cancer	Consolidated data from TCGA, CPTAC, FMI, HCMI

Table 7: List of multi-omics data repositories adapter from Subramanian et al. 2020.[109]Abbreviations: CNV: Copy Number Variations, FMI: Foundation Medicine Inc.,

HCMI: Human Cancer Models Initiatives, miRNA: microRNA, RPPA: reverse phase protein array; SNV: Single Nucleotide Variants, WES: Whole Exome Sequencing, WGS: Whole Genome Sequencing. Links to all websites are current and active, accessed on 18/9/2025.

The integration of omics layers can be described in two ways. The first is by stages, which refers to when the data are combined during the analysis pipeline: early, intermediate or late. In early integration, raw features from multiple omics are merged into a single set before integration, which allows joint modelling but introduces noise. Intermediate integration extracts key features from each omic separately and then combines them, while late integration analyzes each omic independently and merges the results only at the interpretation stage, emphasizing complementarity over direct data merging[107]. The second perspective is by structure: vertical integration combines different omics types (i.e. genomics, transcriptomics, epigenomics) from the same samples to provide a systems-level view, whereas horizontal integration pools the same omics type across multiple sources studies to increase statistical power and generalizability[110].

Several tools have been developed for multi-omics integration and can be broadly grouped by the following methods: joint Dimensionality Reduction (jDR), Correlation and Covariance-based jDR (COR), Factor analysis (FA), Probabilistic/Bayesian Models (PR), Similarity (Kernel) based (KB), Network-based integration (NB), Regression-based (RB) and Deep Learning (DL). Examples include MIXOmixs (COR-based)[111], iClusterPlus[112] and Multi Omics Factor Analysis[113] (MOFA, both jDR based) and Similarity network fusion (SNF, KB-based) which are used to cluster multi-omics data and identify shared patterns to define tumor subtypes (Figure 4). The most frequently used bulk multi-omics data types in cancer research include transcriptomics, epigenomics and genomics[114]. Ultimately, the aims of a multi-omics study shape its analytical direction, as the pursuit of different biological questions such as the identification of disease subtypes, classifying samples disease risk, detection of molecular patterns associated with the disease, or prediction of respond to treatment guides the choice of integration strategies and tools[110,114].

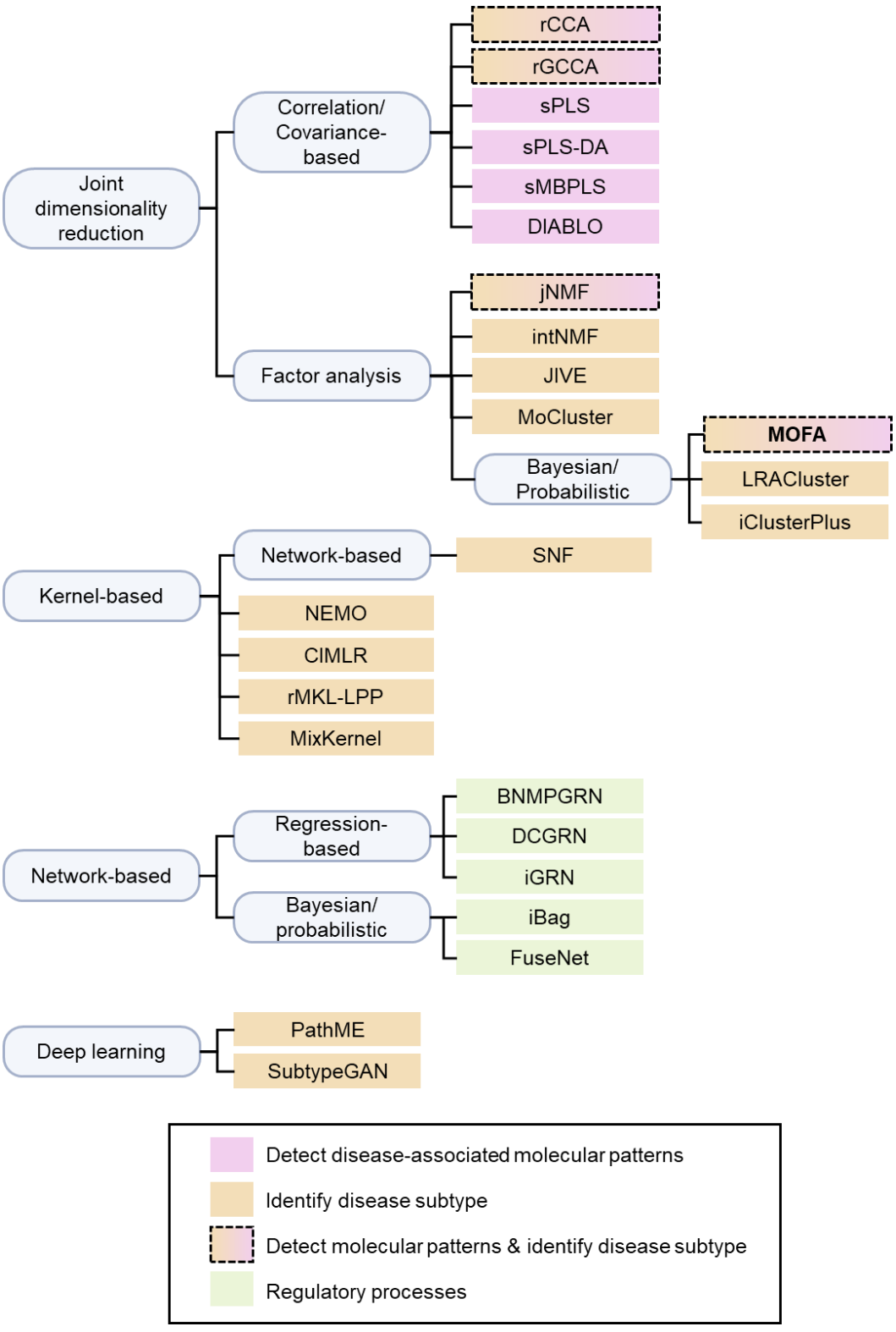


Figure 4: Multi-omics tools grouped by data integration method adapted from Athieniti and Spyrou, 2023[114]. Tools should be selected based on scientific objectives, tools

which can be used to detect disease-associated molecular patterns and identify disease subtypes are outlined in dashed line.

This thesis explores the application of Multi-omics Factor Analysis (MOFA)[113], an unsupervised framework based on factor analysis for integrating multiple data modalities. MOFA takes several data matrices corresponding to the same or partially overlapping sets of samples as input, $M_{1..n}$, and decomposes them into a set of latent factors, Z, that captures major sources of variation. For each data modality, MOFA also estimates a corresponding weight matrix, W, enabling the interpretation of how each omic layer contributes to the latent factors. These factors provide a low-dimensional representation of the data, summarizing shared and omic-specific variation, and can be used for downstream visualization, clustering or classification in a manner analogous to principal component analysis. Important strengths of MOFA include its ability to efficiently handle missing values and integrate heterogenous data types including binary, counts or continuous measurements. The algorithm is available as an R package, https://github.com/bioFAM/MOFA, accessed 19 September 2025. Over the years, MOFA has been applied on various diseases for disease subtyping and survival prognosis[113,115–118] and has been updated to integrate single cell multi-omics data[119].

Multi-omics model evaluation criteria

Multi-omics models can be assessed for its generalizability to unseen data and identify potential biases. One common approach is the use of confusion matrix, which categorizes test samples by comparing the predicted and true values. The outcomes include true positive (TP) where a model correctly predicts a positive instance; true negative (TN) where it correctly predicts a negative instance; false positive (FP), where a negative instance is misclassified as positive, and false negative (FN) where a positive instance is misclassified as negative[120]. Table 8 below summarizes the evaluation metrics used depending on the downstream use case.

Use case	Evaluation metrics
Prognosis prediction, survival analysis	C-index
Classification	Accuracy F1 score Area under the curve (AUC) Precision and Recall Sensitivity and Specificity
Regression	Root Mean Squared Error (RMSE) Pearson's correlation coefficient Coefficient of determination, R^2

Table 8: Multi-omics model evaluation metrics adapted from Abdelaziz et al. 2024[121]

The molecular landscape of BCa is heterogenous, making multi-omics approaches essential for understanding tumor biology and supporting more personalized treatment strategies. By integrating CNV, transcriptomics, epigenomics and proteomics data from 862 NMIBC tumor samples (N = 834 patients), Lindskorg et al.[122] identified four molecular classes of NMIBC (1, 2a, 2b and 3) using ConsensusClusterPlus[123]. These classes reflected distinct molecular features and clinical outcomes, with partial overlap with the previously defined UROMOL2016 classes (1-3)[124]. Importantly, the classes showed significantly different progression-free survival (PFS log rank test, $P = 6.6 \times 10^{-5}$), and recurrence-free survival (RFS log rank test, $P = 0.025$). The NMIBC classifier is publicly available at https://github.com/sialindskrog/classifyNMIBC, (accessed 18 September 2025)[122]. In MIBC, Mo et al. applied iClusterBayes[112] to 388 samples with complete sets of somatic mutation, CNV, methylation, and RNA-seq, identifying two subtypes: integrative basal (iBasal) and integrated luminal/differentiated (or iLuminal). These were defined by combinations of CNV and gene expression profile differences. Separately, Zhang et al.[125] employed deep learning and machine learning approaches to improve MIBC prognostic prediction. The authors used Autoencoders (https://github.com/keras-team/keras) to integrate gene expression, CNV, miRNA and DNA methylation data from TCGA-MIBC. Subsequently, patients were then stratified into high or low-risk subtypes with machine learning methods such as random forest, Naïve Bayes, k-NN and Adaboost. High risk subtypes displayed distinct immune compositions including macrophages,

resting NK cells, regulatory T, plasma and naïve B cells and were enriched in regulatory pathways such as activated IL-6/JAK/STAT3 signaling, interferon-alpha response, reactive oxygen species pathway, and unfolded protein response. Notably, KRT7 was identified as a key biomarker of MIBC risk and was supported by both RNA and protein levels through immunohistochemistry analysis. Further, Chu et al. 2023 combined mRNA, long non-coding RNA (lncRNA), miRNA, genomics mutations and DNA methylation data to construct an integrated consensus subtype of advance MIBC using 10 multi-omics clustering algorithms. The authors identified 32 stable prognosis-related genes which could divide advance MIBC patients into three subtypes with significant prognostic value in predicting response to immunotherapy and other drug-based therapies[126].

Population genetics studies like GWASs typically adopt a horizontal integration strategy whereby an independent GWAS analysis is first performed to identify disease-associated risk loci. Subsequently, these are then fine mapped on to molecular quantitative trait loci (xQTL) such as eQTL, mQTL or pQTL data to identify regulatory variants affecting gene expression, DNA methylation or protein abundances respectively. This strategy leverages well-powered, specialized dataset, reduces computational complexity and allows prioritization of variants most likely to mediate disease through molecular mechanisms. Thus far, these sequential combinations of single omics analysis are most frequently performed[127].

1.2. Hypothesis and aims

BCa exhibits substantial biological heterogeneity, complicating reliable disease detection using current clinical tools, which rely on invasive procedures or non-invasive assays with limited sensitivity, particularly at early disease stages[2,53]. Despite advances in molecular profiling, blood-based biomarker has yet achieved sufficient performance for robust BCa detection largely due to inter-individual variability and the challenge of resolving true disease-associated signals from background molecular noise[128]. While germline genomic variation defines inherited risk, it does not capture context-dependent systemic host response to cancer and treatment. We therefore hypothesize that integrative multi-omics[109] profiling of peripheral blood can capture complementary sources of biological variation, ranging from static germline susceptibility encoded in the genome, to dynamic changes reflected in the transcriptome and epigenome, thereby enabling the detection of BCa-relevant biological signatures beyond what is achievable with single-omic approaches.

In this study, we sought to characterize the genetic and molecular landscape of bladder cancer by leveraging the Turin Bladder Cancer Study (TBCS), a moderately sized but homogeneous and deeply phenotyped cohort. We performed integrative multiomic analysis using Multi-Omics Factor Analysis (MOFA2)[113,119] on genome-wide genotyping data, blood-derived DNA methylation profiles, and transcriptomic data. We then evaluated the ability of MOFA-derived latent factors to discriminate bladder cancer cases from controls and compared MOFA scores to established bladder cancer polygenic risk scores (PRS), offering insights into their relative utility. Finally, we assessed MOFA latent factor's predictive value for recurrence-free survival in Non-Muscle Invasive Bladder Cancer (NMIBC). Our findings provide a framework for the potential of unsupervised multi-omic integration to uncover biologically relevant axes of variation and support the development of minimally invasive, blood-based biomarkers for risk stratification in bladder cancer.

1.3. Methods and analysis workflow

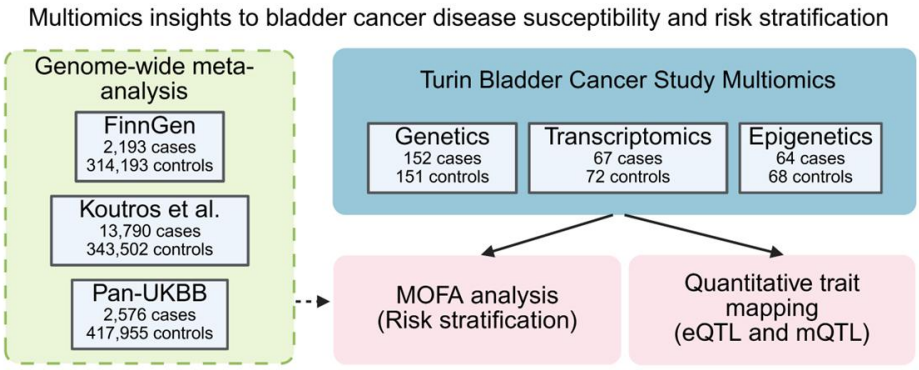


Figure 5: Multi-omics integration in bladder cancer. We performed a GWAS meta-analysis on published BCa summary statistics (18,559 cases and 1,075,650 controls). Separately, MOFA, an unsupervised multi-omics factor analysis, was used to integrate genomic, transcriptomic, and epigenomic profiles from 120 TBCS samples with paired omics data. Features guided by overlapping SNPs derived from the GWAS meta-analysis, transcriptomics analysis (67 cases, 72 controls) and methylation analysis (64 cases, 68 controls) were used as input for MOFA. In parallel, we performed quantitative trait mapping to generate a resource of expression (eQTL) and methylation (mQTL) quantitative trait loci within this dataset.

1.3.1. Turin Bladder Cancer Study population

TBCS consists of histologically-confirmed adult male BCa patients and non-cancer controls who were enrolled between the years 1994 and 2008 from the Turin metropolitan area[129–133]. Blood samples were obtained prior to treatment, and a subset of samples were subjected to genomics, transcriptomics, and methylation profiling. All biological material and data collection of this cohort was completed in 2015. The exclusion criteria involved patients with potentially confounding diagnoses, specifically those with a history of cancer, chronic liver, kidney or smoking-related illnesses. Male controls included patients with non-neoplastic diseases such as prostatic hyperplasia, cystitis and patients from medical and surgical departments with hernias, vasculopathies, diabetes, heart failure, asthma, or other benign diseases. The study was approved by the Ethical Committee Board of the Città della Salute Hospital, Turin. Additional 290 non-cancer

controls were obtained from EPICOR (European Prospective Investigation into Cancer and Nutrition Italian cohort) study[134,135], which is a case-cohort study nested within the EPIC-Italy prospective cohort. Signed informed consent was obtained according to the Helsinki Declaration in both studies.

1.3.2. TBCS-EPICOR exploratory GWAS

Genome-wide genotyping was performed with the Infinium OmniExpressExome-8 kit on a cohort of 594 individuals, which included 304 male individuals (153 cases, 151 controls) recruited as part of TBCS and an additional 290 non-cancer controls of all genders pooled from EPICOR, hereafter referred to as the TBCS-EPICOR GWAS dataset. For all 593 samples, we extracted a set of 125,424 common, informative and independent SNPs (genotyping rate 95%, MAF 0.01, hwe $p > 1x10^{-6}$, --indep-pairwise 50 5 0.2, excluding genomic ranges of high linkage-disequilibrium regions in Hg19) to calculate principal components using Flashpca. We computed Z-scores of PC1 and PC2 of each sample and any sample with a Z-score deviation greater than 6 was considered an outlier and removed. The resulting dataset of 582 individuals (147 cases, 435 controls) were imputed with the Haplotype Reference Consortium (HRC) reference panel in Michigan Imputation Server[136,137]. Subsequently, we retained 5,778,682 imputed SNPs with an imputation quality score $R^2 > 0.3$, call rate > 0.95, MAF > 0.01, MAC >= 20 , hwe $p > 1x\ 10^{-6}$, --max-alleles 2, --min-alleles 2 for downstream logistic regression analysis with PLINK2[138]. We applied an additive model and adjusted for age, sex (when required), smoking statuses and 2 principal components.

1.3.3. External BCa GWAS meta-analysis

We performed a fixed effect meta-analysis using METAL[139] under the standard error (STDERR) scheme to combine the summary statistics from three publicly available large BCa GWASs (18,559 cases and 1,075,650 controls): FinnGen (finngen_R10_C3_BLADDER_EXALLC with 2,193 cases, 314,193 controls), Koutros et al., 2023 (13,790 cases and 343,502 controls) and Pan-UKBB (icd10-C67-both_sexes.tsv.bgz with 2,576 cases, 417,955 controls). FinnGen summary statistics that

were originally mapped to the GRCh38 coordinates were lifted over to the GRCh37 (Hg19) build using UCSC LiftOver[140]. Lambda inflation factor was estimated with the estlambda function within the GenABEL R package[141] with default parameters. Associated SNP heterogeneity were assessed based on I^2 and Cochran's Q-test parameters from METAL.

To compare the per-allele effect size between the TBCS-EPICOR GWAS and Koutros et al. 2023[107] (13,790 cases and 343,502 controls), we first merged the two respective summary statistics by chromosome and position and aligned the effect alleles and beta based on the latter in R. We then identified overlapping SNPs that reached genome-wide significance in the meta-analysis ($P < 5 \times 10^{-8}$), and assessed the correlation of effect sizes across the two studies using Pearson's product moment correlation coefficient, cor.test.

1.3.4. PRS validation

We performed PRS calculations on all 582 individuals that passed QC in the TBCS-EPICOR GWAS based on 16 BCa PRS from the PGS catalog[142] and Koutros et al. To verify the relationship between these PRS scores and disease risk. We extracted imputed SNPs with an accuracy $R^2 >= 0.3$, MAF 0.01 and call rate 0.90, aligned the effect allele, flipped strands, and removed mismatched or ambiguous SNPs to harmonize the genotypes from the TBCS against established PRS summary statistics according to the protocols described in Choi et al. [143,144]. We used PRSice-2 (P+T method) to calculate scores with the following parameters: PRSice.R --prsice PRSice_linux --stat BETA --beta --no-clump --score sum --bar-levels 1 --binary-target T". Subsequently, we standardize each PRS based on its mean and standard deviation, defining the effect of each standardized PRS as odds ratios (OR) per standard deviation increase of the PRS. We evaluated the performance of the standardized scores in predicting case-control statuses using a logistic regression model adjusting for age of recruitment, smoking status, sex (when required) and first 2 PCs with the glm function. We assessed PRS performance by calculating AUC using the pROC R package v1.18.5 in R 4.4.2.

1.3.5. Transcriptomic analysis

Blood transcriptomics data from the TBCS cohort was generated and analyzed previously by our group[132]. Briefly, we used the Illumina HT-12 v4 Beadchip to perform mRNA expression profiling of peripheral blood samples in three batches. The resulting 139 raw IDAT files (67 cases and 72 controls) were re-analyzed in R (v4.0.5) using limma and illuminaHumanv4.db. We imported probe-level expression data using read.idat, and applied background correction and quantile normalization with the neqc function, yielding expression values for 47,323 probes. To restrict the analysis to high-confidence probes, we excluded 12,847 probes including 12,379 that aligned to repeat sequences, intergenic or intronic regions, or lacking transcript specificity, and 468 that failed to map to any genomic region based on illuminaHumanv4PROBEQUALITY annotations (illuminaHumanv4.db_1.26.0). We then used removeBatchEffect on the remaining 34,476 probes to correct for batch effects associated with mRNA profiling batch IDs. Subsequently, we excluded 1,721 non-RefSeq probes. For 6,362 genes targeted by multiple probes, we selected the maximum expression level for each gene using the aggregate function, resulting in 23,611 gene expressions values. We stratified the cohort based on affection status (case vs control) and constructed a linear model for differential gene expression analysis, adjusting for age and smoking statuses using lmFit. We extracted a table of top genes associated with case-control status using topTable, correcting for multiple testing with false discovery rate (FDR).

1.3.6. Methylation analysis and immune cell deconvolution

Methylation profiling was performed on 132 post-diagnostic blood samples from the TBCS using the Illumina HumanMethylation450K (64 cases and 68 controls). We used the ChAMP pipeline[145] in R(v4.0.5) to perform methylation probe-level QC, removing probes with a detection $P > 0.01$ (29,742), probes with a beadcount < 3 in at least 5% of samples (115), non-CpG probes(2,270), probes excluded from Zhou et al.[146] (54,885), probes ambiguously aligning to multiple locations from Nordlund et al.[147] (11), sex chromosome probes (8,873), resulting in 389,616 remaining probes. Combat[148] was used to correct for batch effect on the remaining probes. Simultaneously, MINFI[149] was used to derive

RGChannelSet from raw idat files to estimate immune cell type proportions according to the FlowSorted.Blood.450k panel[150]. Limma[151] was used for differential methylation analysis. Briefly, we constructed a design matrix to adjust for potential confounders, including age, smoking status and the estimated cell proportions of CD8T, CD4T, NK, B cell, monocyte, granulocytes and disease statuses. Next, we fitted the linear model to the normalized, batch-corrected, methylation matrix consisting of ChAMp-filtered probes with lmFit(). eBayes() was used to obtain variance estimates and t-statistics for each probe. All probes were annotated according to the IlluminaHumanMethylation450kanno.ilmn12.hg19 package.

1.3.7. MOFA analysis and input feature selection

We performed multi-omics integration with MOFA[113,119] in 120 TBCS blood samples (58 cases, 62 controls) with paired DNA genotyping, mRNA and methylation profiles and have fully matched data matrices(Figure 6). Briefly, MOFA implements an unsupervised factor analysis to decompose sources of variation in each data modality into a matrix of latent factors (or MOFA factors) for each sample and weight matrices for each omic feature, enabling downstream analyses such as clustering and pathway enrichment analysis. We structured our MOFA input as data matrices (M), with N x D_m dimensions, whereby N represented the total number of sample and D_m were the omic features from each data modality (known as view).

We used data from single-omic analysis to select highly variable features for model training. For genotyping data, the TBCS-EPICOR discovery GWAS was underpowered to identify genome-wide significant loci. Therefore, we used the BCa GWAS meta-analysis results to guide SNP selection. From this meta-analysis, we selected 104 independent and robust SNPs ($P_{heterogeneity}$ < 0.05) reaching at least suggestive significance ($P < 1 \times 10^{-5}$) by performing LD clumping with parameters: clump-p1 1e-05 --clump-r2 0.1 --clump-kb 500 using the 1000 Genomes Phase 3, unrelated European data as LD reference panel. The resulting variants were intersected with SNPs present in the TBCS dataset and their dosages (coded 0, 1, or 2) were used as genetic features for MOFA. For transcriptomics, we defined genes-of-interest as those that were nominally

significant ($P < 0.01$) and had an absolute log2 fold change >=0.2 when comparing cases against controls. We constructed a matrix of residuals from functional-normalized gene expression values after regressing out age and smoking status with the residuals function in R. For methylation data, residuals from M-values of methylated CpGs (nominal *P*-value < 0.01, absolute logFC >=0.2) were used after regressing age, smoking status and immune cell proportions. We trained the MOFA models with default settings on 120 TBCS samples using 15 latent factors to determine if these latent factors could be used to distinguish cases from controls. To further test MOFA's prognostic utility in predicting recurrence, we then constructed a multivariate Cox proportional-hazards model on 47 NMIBC cases adjusting for age and smoking statuses. We calculated the Kaplan-Meier (KM) curve and used the log-rank test to test the difference in the time-to-event between stratified latent factor groups. We performed all analyses in R (version 4.0.5).

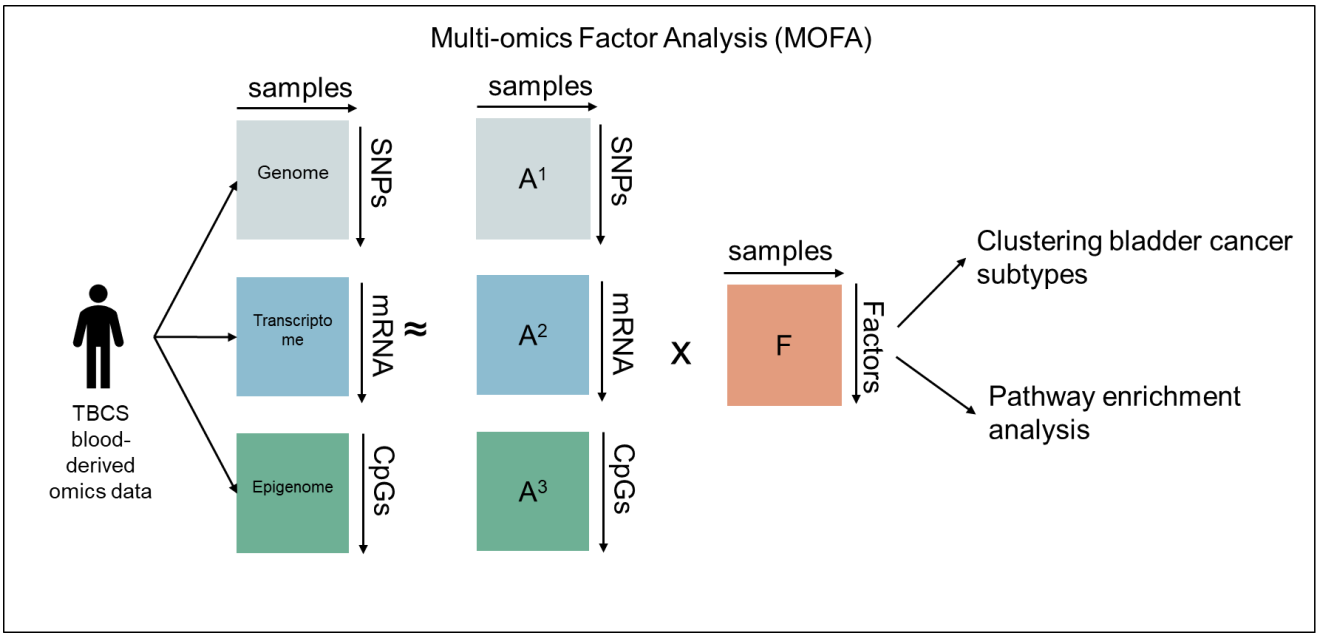


Figure 6: Unsupervised MOFA analysis pipeline integrating genomics, transcriptomics and methylation data from 120 samples derived from TBCS (58 BCa patients, 62 controls).

1.3.8. Quantitative traits mapping (QTM)

We performed QTM to identify expression quantitative trait loci (eQTL) and methylation quantitative trait loci (mQTL) using the R package Matrix eQTL. Paired genotype-expression and genotype-methylation data available within the TBCS was used. A linear model was used to test the association between each SNP dosages and gene expression levels (eQTL) and SNP dosages and methylation levels (mQTL) while adjusting for age and smoking statuses. In both cases, we defined a significant cis-QTL association as SNP-gene pairs or SNP-CpG pairs within 1 Mb of each other and an adjusted *p*-value threshold of 0.05.

1.4. Results

1.4.1. Study characteristics

The TBCS cohort included 726 participants, of whom 306 individuals (154 adult, male BCa cases and 152 male non-cancer controls) were subjected to genome-wide genotyping, with cases and controls matched by age, smoking status. Whole blood-derived RNA and DNA methylation data were available for 139 and 132 individuals, respectively.120 individuals have fully overlapping multi-omics data. The median age of cases and controls was 64 years. Within TBCS cases, 138 (90.2%) were classified as NMIBC (Tis, Ta, T1), and the remaining 16 (10.5%, >T2) as MIBC (Table 9). Seventy-eight patients were classified as high-grade (HG) BCa, and the remaining 76 were low-grade (LG) according to the WHO 2004/2016 grading classification. Over a median follow-up of 54 months (IQR 33-93), 28 (18.3%) patients underwent cystectomy, 24 (15.7%) patients were deceased, of which 11 were BCa-specific mortality. Among the 153 BCa patients, 55 (35.9%) experienced disease recurrence while 99 (64.7%) remained recurrence-free at last follow-up. NMIBC cases were stratified into 58 (42.03%) high, 44 (31.88%) intermediate and 36 (26.09%) low risk classes.

Bladder cancer	Case (n=154)	Control (n=152)	MOFA Case (n=58)	MOFA Control (n=62)
Median age	64.5	64.6	62.3	61.52
Smoking				
Non-smoker	25	24	8	9
Former smoker	93	94	37	36
Current smoker	36	34	13	17
Tumor type				
Non muscular invasive bladder cancer (NMIBC)	138		49	
Muscular invasive bladder cancer (MIBC)	16		9	
WHO 2004 Grade				
High grade (HG)	78		34	
Low grade (LG)	76		24	
NMIBC Risk class				
High risk	58		24	
Intermediate risk	44		13	
Low risk	36		12	
Cystectomy				
No	126		45	
Yes	28		13	
Recurrence				
No	99		35	
Yes	55		23	
Mortality				
Alive	130		48	
Deceased	24		10	

Table 9: TBCS is a nested case-control cohort with 726 participants, of whom 306 individuals (154 adult, male BCa cases and 152 male non-cancer controls) were initially subjected to genome-wide genotyping. 120 samples (58 cases, 62 controls) were simultaneously subjected to genotyping, transcriptomic and methylation profiling and were used for downstream MOFA integration.

1.4.2. External BCa meta-analysis yielded 4 additional susceptibility loci.

GWAS meta-analysis of BCa summary statistics from three studies totalling 18,559 cases and 1,075,650 controls (Koutros et al., Pan-UKBB, FinnGen) was performed on

19,073,763 common SNPs (λ = 1.055). Lead variants are defined as the most significant variant within a 1Mb window (Figure 7, Table 10). The meta-analysis yielded 29 genome-wide significant signals encompassing 25 previously reported associations[107,152] and revealed 4 additional signals, including loci at chr5q32 (rs6580593; OR =1.083; *P* = 2.75 x10^{-9}; CTC-529P8.1/SH3TC2; with significant cis-eQTL association to *SH3TC2* in eQTLGen phase 1), chr7q36.3 (rs6459970; OR = 1.098; *P* = 4.72 x 10^{-11}; non-coding region), chr9q34.13 (rs67080241; OR=1.136 ; *P* = 3.81 x 10^{-9} ; with significant blood cis-eQTL association to *RAPGEF1* in eQTLGen phase 1), and chr11q22.1(rs533571, OR= 0.919 , *P* =7.97 x 10^{-9}; *ARHGAP42*). No association or colocalization was found between the susceptibility variant and the TBCS-derived eQTL and mQTL data (data not shown).

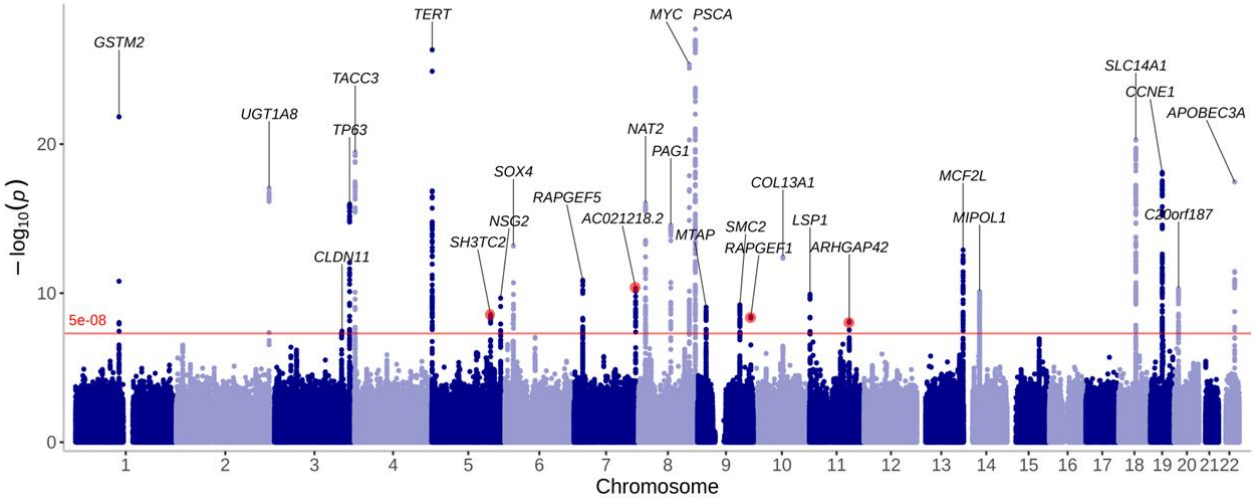


Figure 7: Manhattan plot depicting the genome-wide meta-analysis of external BCa GWAS excluding the TBCS-EPICOR exploratory GWAS cohort (Koutros et al., 2023; FinnGen; Pan-UKBB; 18,559 cases and 1,075,650 controls) used for SNP feature selection in MOFA. Points highlighted in Red indicated candidate novel loci in this meta-analysis.

rsID	Nearest Gene	CHR	POS	EA	Freq	Beta	SE	*P*	Direction	I^2	HetPVal
Previously reported loci											
rs36209093	*GSTM2*	1	110229787	T	0.66	0.18	0.02	$1.47x10^{-22}$	+?+	48.6	0.16
rs17868316	*UGT1A8*	2	234563646	A	0.99	0.54	0.06	$9.26x10^{-18}$	+++	0	0.38
rs1317082	*MYNN*	3	169497585	A	0.75	0.09	0.02	$4.86x10^{-08}$	+++	0	0.71
rs34666239	*TP63*	3	189633513	A	0.25	-0.13	0.01	$1.07x10^{-16}$	---	9.1	0.33
rs28699509	*TACC3*	4	1758949	T	0.79	-0.15	0.02	$3.62x10^{-20}$	---	36.1	0.21
rs2242652	*TERT*	5	1280028	A	0.20	-0.18	0.02	$4.76x10^{-27}$	---	22.3	0.28
rs4446484	*RP11-267A15.1*	5	173853681	A	0.46	-0.09	0.01	$2.18x10^{-10}$	---	0	0.40
rs7765413	*CDKAL1*	6	20751241	C	0.56	0.08	0.01	$1.06x10^{-08}$	+?+	0	0.39
rs72826305	*CASC15*	6	21826729	T	0.65	-0.11	0.01	$6.83x10^{-14}$	---	0	0.88
rs35675999	*RAPGEF5*	7	22397515	A	0.20	-0.12	0.02	$1.38x10^{-11}$	---	0	0.81
rs1495743	*NAT2*	8	18273300	C	0.77	0.13	0.02	$9.21x10^{-17}$	+++	0	0.61
rs4739782	*PAG1*	8	81987666	A	0.53	0.11	0.01	$2.86x10^{-15}$	+++	0	0.59
rs10094872	*CASC11*	8	128719884	A	0.59	-0.23	0.02	$4.45x10^{-26}$	?--	0	0.85
rs2976398	*PSCA*	8	143764879	C	0.46	0.15	0.01	$1.93x10^{-28}$	+++	31.1	0.23
rs13297760	*MTAP*	9	21823952	T	0.58	-0.08	0.01	$8.89x10^{-10}$	---	0	0.46
rs4742905	*SMC2*	9	106856972	C	0.57	-0.08	0.01	$6.15x10^{-10}$	---	0	0.70
rs7076867	*COL13A1*	10	71582996	T	0.06	-0.21	0.03	$3.72x10^{-13}$	---	63.2	0.06
rs55840650	*LSP1*	11	1879326	T	0.32	0.09	0.01	$1.22x10^{-10}$	+++	32	0.23
rs4907572	*MCF2L*	13	113647321	A	0.27	0.11	0.01	$1.28x10^{-13}$	+++	0	0.89
rs7145592	*MIPOL1*	14	37929743	T	0.28	0.10	0.01	$8.14x10^{-11}$	+++	0	0.43
rs17674580	*SLC14A1*	18	43309911	T	0.35	0.13	0.01	$5.21x10^{-21}$	+++	0	0.89
rs62104473	*CCNE1*	19	30289779	T	0.33	0.12	0.01	$7.99x10^{-19}$	+++	21.4	0.28
rs411482	*FAM83E*	19	49103447	T	0.38	-0.09	0.01	$2.92x10^{-10}$	---	80.3	0.01
rs7353743	*RP11-103J8.1*	20	10958946	C	0.75	-0.10	0.01	$5.07x10^{-11}$	---	0	0.63
rs5750711	*APOBEC3A*	22	39341957	T	0.63	0.12	0.01	$3.47x10^{-18}$	+++	0	0.87
Candidate novel loci											
rs6580593	*CTC-529P8.1/ SH3TC2*	5	148463601	C	0.56	0.08	0.01	$2.75x10^{-09}$	+++	0	0.77

rs6459970	*AC021218.2*	7	155742321	A	0.67	0.09	0.01	$4.72x10^{-11}$	+++	0	0.76
rs67080241	*RAPGEF1*	9	134630267	T	0.11	0.13	0.02	$3.81x10^{-09}$	+++	28	0.23
rs533571	*ARHGAP42*	11	100850202	A	0.29	-0.08	0.01	$7.97x10^{-09}$	---	56.8	0.10

Table 10: Genome-wide significant signals from the BCa meta-analysis. Previously reported loci were obtained from Koutros et al. 2023, Verma et al. 2024, Larrson et al. 2025. CHR: chromosome; POS: position; EA: Effect allele; Freq: Frequency of effect allele; Beta: Effect size of the effect allele; SE: Standard error; I^2: Proportion of heterogeneity in the meta-analysis on a scale of 1-100%, HetPVal: Heterogeneity P-value.

GWAS was similarly performed within the TBCS-EPICOR cohort using PLINK2 under an additive model, adjusting for age, sex, smoking statuses and 2 PCs (after QC: 147 cases, 435 controls) as an exploratory data analysis. Given the limited cohort size, this GWAS is inherently under-powered (~0.25%) to detect genome-wide associations under the following additive model assumptions: N=582, case rate=0.25, MAF=0.25 and a moderately large effect size OR=1.5 with genpwr (v1.0.4)[153]. This analysis should be interpreted as exploratory, providing contextual insight to determine whether the observed genetic architecture was consistent with signals reported in larger studies. Nonetheless, three loci showed suggestive associations ($P < 5 \times 10^{-6}$): chr15q21.3 (rs28512496; OR = 2.13; P = 1.79 x 10^{-6}; nearest gene *FAM214A*), chr10p15.3 (rs61311809, OR=3.13, P = 4.29 x 10^{-6}; closest gene *DIP2C*), and chr5p13.3 (rs7717852; OR = 2.13; P = 4.47x10^{-6}; *ADAMTS12*) (Figure 8). These suggestive loci were not replicated in meta-analysis, indicating that they likely represent cohort-specific or non-significant associations. A comparison of effect sizes in 28 genome-wide significant loci from the meta-analysis and TBCS-EPICOR GWAS demonstrated strong concordance in both direction and magnitude (Pearson r = 0.7389, P = 7.12 x 10^{-6}) (Figure 9).

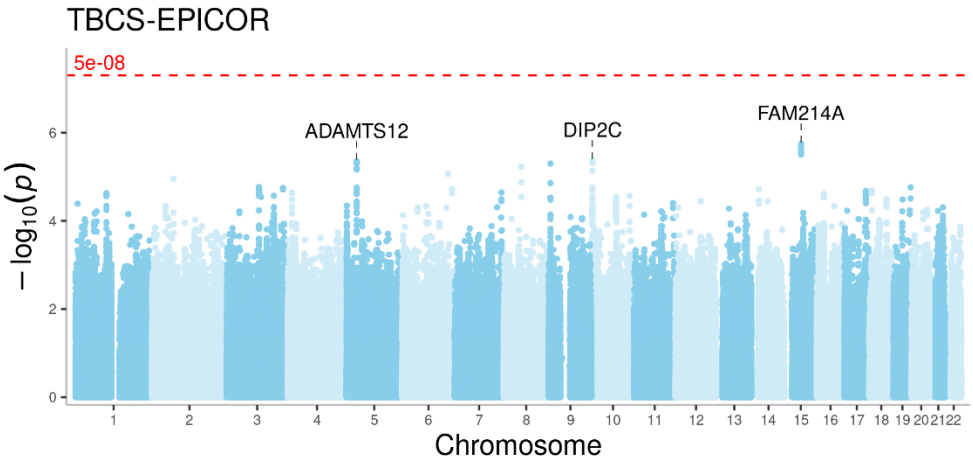


Figure 8: Manhattan plot from the TBCS-EPICOR discovery GWASs cohort (582 individuals: 147 male cases, 435 male and female controls) identified three loci near *ADAMTS12*, *DIP2C* and *FAM214A* with suggestive significance ($P < 5 \times 10^{-6}$).

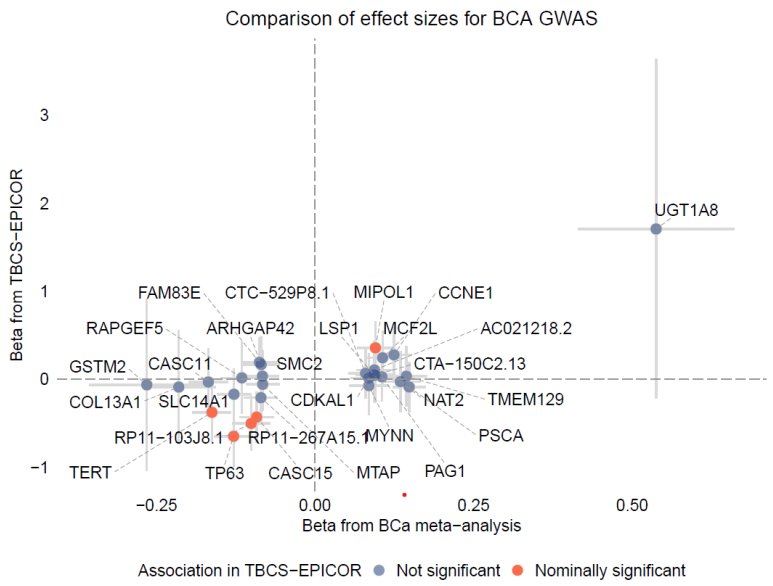


Figure 9: Comparison of per-allele effect sizes between the external BCa meta-analysis (section 1.4.2.) and the TBCS-EPICOR discovery GWAS showed consistent directionality with positive correlation (r = 0.73, P = 7.119 x 10^{-6}); points represent lead variants, annotated by nearest gene (hg19) and colored by TBCS-EPICOR GWAS significance, blue points are loci that were not significant (P > 0.05) in the TBCS-EPICOR GWAS, whereas orange points were nominally significant (P < 0.05) in the TBCS-EPICOR GWAS.

1.4.3. Differential gene expression (DGE) and differential methylation analysis.

Whole blood mRNA DGE analysis from 67 TBCS cases and 72 controls identified 8 differentially expressed genes (adjusted P < 0.05): 7 were significantly down-regulated: *MMP23A* (a pseudogene of *MMP23B*, logFC =-0.045, P = 1.7x10^{-3}), *CLDND2* (logFC= -0.450 , P = 6.9 x 10^{-3}), LOC643035 (logFC= -0.324, P = 1.2x10^{-2}), *FEZ1* (logFC=-0.542, P = 3.08x10^{-2}), *SPON2* (logFC=-0.570, P = 3.6x10^{-2}), *SLC1A7* (logFC=-0.471, P = 4.3x10^{-2}), and *ASCL2* (logFC= -0.448, P = 4.51x10^{-2}). One gene was up-regulated C6orf105/*ADTRP* (logFC=0.653, P = 5.5x10^{-3}). No differential methylation CpG sites were detected in the differential methylation analysis of 64 cases and 68 controls after adjustment with age, smoking statuses, and immune cell proportions.

1.4.4. MOFA factors detect variability within BCa cases and controls.

We performed multi-omics analysis with MOFA to look for sources of variation that could distinguish BCa cases and controls into subgroups. Prior to data integration, the model requires a pre-filtered set of features. To this end, we selected 64 pruned SNPs from the GWAS, the residuals of 951 nominally significant differentially methylation CpGs, and the residuals of 144 nominally significant differentially expressed genes, as features in the MOFA analysis (Figure 10). The model was trained on 120 samples with fully overlapping multi-omic data (58 cases, 62 controls), all labelled as Group 1. MOFA identified 15 factors which could explain variability in the data. This analysis revealed that both gene expression and methylation data explained more than 40% variance in the data, whereas SNP dosages explained less than 10% of variance. (Figure 11A). Transcriptomics features explained the largest variability in Factor 1 (Figure 11B). When correlating individual factor to disease covariates, factor 1 was most significantly correlated with disease status (case/control), WHO 2004 and WHO 1973 tumor grade and number of recurrence(s) (Figure 11C). Subsequently, we applied a non-parametric Two-sample Kolmogorov-Smirnov test to compare the mean distribution of latent factor 1 across disease status, WHO 2004 tumor grade classification and tumor recurrences (Figure 11D). The distributions of latent factor 1 were significantly different between cases and controls (P = 1.11 x 10^{-3}), non-high grade tumor and controls (P = 3.77 x 10^{-2}), high grade tumor and control (P = 1.44 x 10^{-3}), no recurrence and control (P = 1.36 x 10-3), and patient with recurrence and control (P = 1.20 x 10^{-2}). However, no significant differences were observed within cases. Reactome pathway enrichment analysis showed that factor 1 genes were involved in pathways related to Phase II metabolic enzymes (conjugation of compounds), GRB7 events in ERBB2 signaling, glutathione conjugation and drug absorption, distribution, metabolism and excretion (ADME) (Table 11).

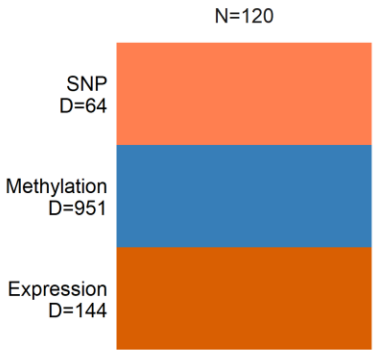


Figure 10: MOFA was trained on 120 samples with fully paired genotyping, blood-derived DNA methylation profiling and mRNA profiling data. Input features including 64 SNPs, residuals from 951 differentially methylated CpGs (nominal P < 0.01, |logFC| >= 0.2), and residuals from 144 differentially expressed genes (nominal P < 0.01, |logFC| >= 0.2) were used in the training model. No values were imputed.

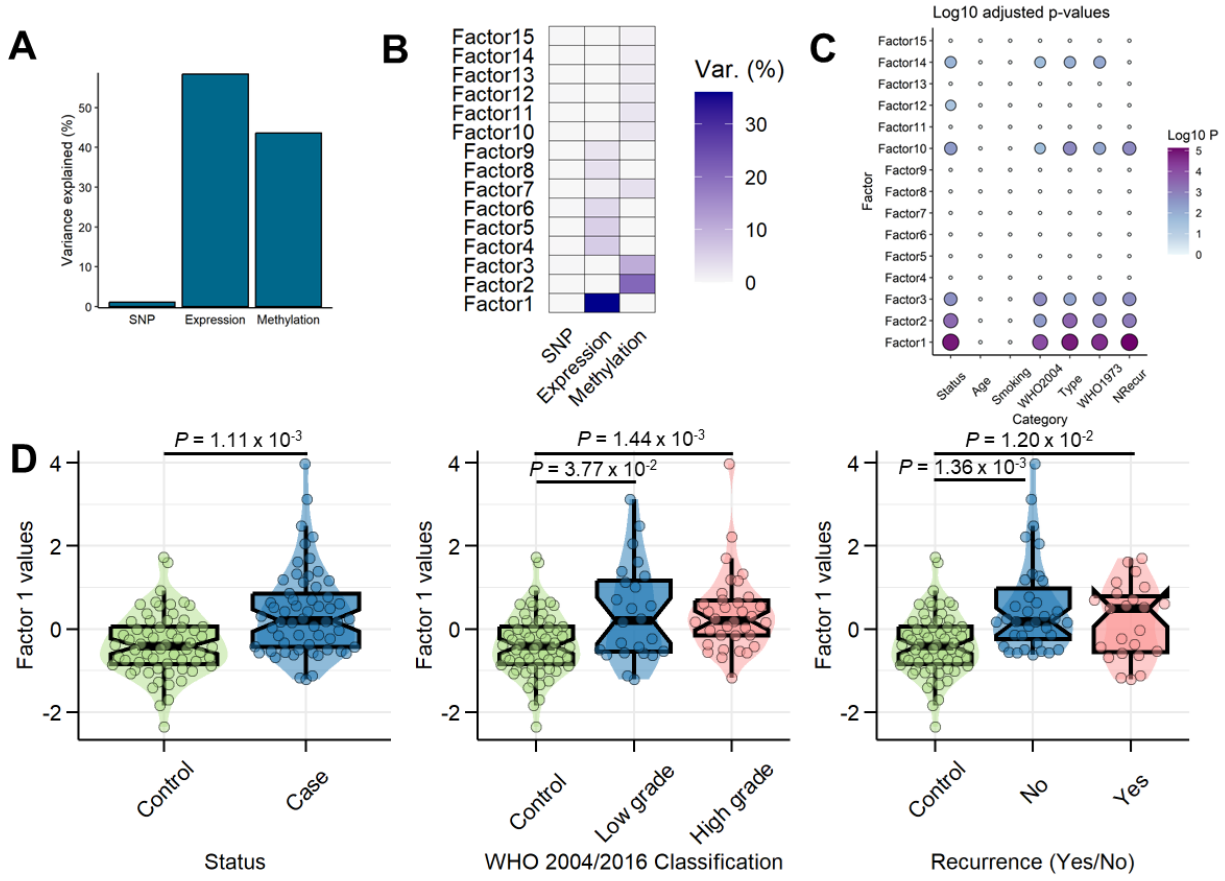


Figure 11: MOFA was trained on 120 samples (58 cases and 62 controls) without group labels. (A) Total percentage of variance explained by each data modality. (B) Variance explained by each modality across 15 latent factors. (C) Associations between disease

covariates and MOFA-derived latent factors. (D) Distribution of MOFA-predicted Factor 1 across disease status, tumor grade (LG or HG), and recurrence. Each point represents an individual sample (N = 120). Factor 1 values were significantly different between controls and disease groups, but not among disease categories. P-values calculated with two-sample Kolmogorov–Smirnov test.

Enrichment term	*P*	*Q*
Phase II - Conjugation of Compounds	$4.37x10^{-5}$	0.01
GRB7 Events in ERBB2 Signaling	$1.20x10^{-4}$	0.02
Glutathione Conjugation	$2.92x10^{-4}$	0.03
Drug ADME	$5.80x10^{-4}$	0.04
Downregulation of ERBB2 ERBB3 Signaling	$9.19x10^{-4}$	0.04
ERBB2 Activates PTK6 Signaling	$9.19x10^{-4}$	0.04
Biological Oxidations	$1.08x10^{-3}$	0.04
ERBB2 Regulates Cell Motility	$1.23x10^{-3}$	0.04
GRB2 Events in ERBB2 Signaling	$1.40x10^{-3}$	0.04
PI3K Events in ERBB2 Signaling	$1.40x10^{-3}$	0.04

Table 11: Reactome pathways enriched terms for the top features that carried the most weights in MOFA factor 1. *Q* values referred to false discovery rate (FDR).

1.4.5. Performance comparison between established genetic risk score and MOFA-derived factor 1.

The discriminative performance of the MOFA latent factor 1 score was compared against published polygenic risk scores (PGS) for distinguishing BCa cases from controls (Figure 12). In the TBCS cohort (N = 120), MOFA scores revealed a stronger association (OR = 2.79, $P = 1.12 \times 10^{-4}$) whereas the PRS models yielded an average OR of 1.30. Only two PRSs reached statistical significance within this cohort: PGS000608 (OR = 1.78, $P = 4.85x10^{-3}$) and PGS000607 (OR = 1.45, $P = 5.27x10^{-2}$). MOFA latent factor 1 yielded an area under the curve (AUC) of 0.71 in the TBCS cohort. By comparison, the highest-performing PRS (PGS000608) in this dataset achieved an AUC of 0.67, and the average AUC across all PRSs was 0.58. To assess the impact of sample size on PRS stability, performance was subsequently evaluated in the combined TBCS-EPICOR GWAS

samples (N = 582) (Table 12). Interestingly, PRS effect estimates were found to be more consistent, with 11 PRSs achieving statistically significant associations with disease risk.

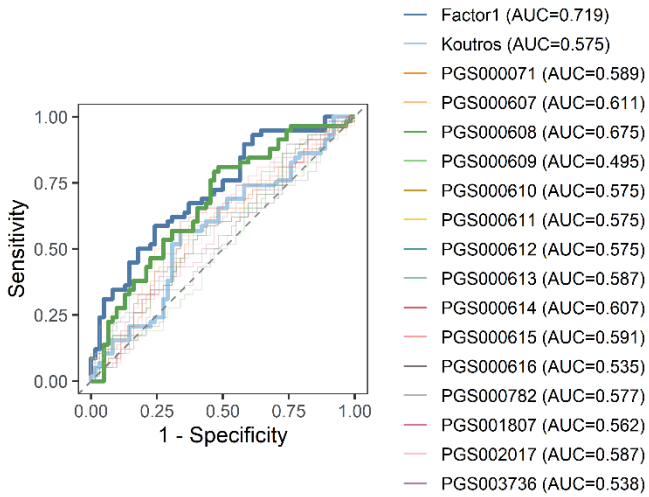


Figure 12: Receiver operating characteristic (ROC) performance comparison between MOFA-predicted Factor 1 and published BCa polygenic scores (PGSs) in distinguishing cases from controls. MOFA achieved an AUC of 0.71 (calculated using the pROC R package).

PGS	Variants	Full GWAS dataset (N=582)				MOFA-only dataset (N=120)			
		OR	CI (lower)	CI (upper)	*P*	OR	CI (lower)	CI (upper)	*P*
PGS000071	15	1.48	1.22	1.80	8.04×10^{-5}	1.38	0.95	2.00	0.09
PGS000613	15	1.43	1.18	1.74	2.72×10^{-4}	1.40	0.96	2.03	0.08
PGS000782	15	1.43	1.18	1.74	2.98×10^{-4}	1.33	0.92	1.93	0.13
PGS002017	510,453	1.40	1.15	1.69	7.32×10^{-4}	1.37	0.94	1.98	0.10
PGS000610	13	1.39	1.15	1.69	7.33×10^{-4}	1.36	0.94	1.97	0.11
PGS000611	13	1.39	1.15	1.69	7.33×10^{-4}	1.36	0.94	1.97	0.11
PGS000612	13	1.39	1.15	1.69	7.33×10^{-4}	1.36	0.94	1.97	0.11
Koutros et al. 2023	24	1.36	1.11	1.65	2.36×10^{-3}	1.33	0.92	1.92	0.14
PGS003736	13	1.30	1.07	1.57	7.08×10^{-3}	1.24	0.86	1.79	0.25
PGS000615	106	1.28	1.06	1.54	1.06×10^{-2}	1.38	0.95	2.00	0.09
PGS001807	291	1.28	1.06	1.55	1.11×10^{-2}	1.29	0.90	1.86	0.17
PGS000614	1,119,238	1.15	0.95	1.38	0.15	1.41	0.97	2.06	0.07
PGS000616	24,359	1.14	0.95	1.38	0.16	1.15	0.80	1.65	0.45
PGS000608	1,097,063	1.13	0.94	1.36	0.21	1.79	1.19	2.68	4.86×10^{-3}
PGS000607	1,095,241	1.05	0.87	1.27	0.58	1.45	1.00	2.11	0.05
PGS000609	1,130	0.98	0.82	1.19	0.87	1.00	0.70	1.43	1.00

Table 12: Published BCa PGSs available in the PGS catalog and from Koutros et al., 2023 were tested on the full case/control (N=582) and MOFA-only (N=120) TBCS samples with PRSice-2. All PRSs were standardized, and odds ratio (OR) were expressed as OR per standard deviation increase. All models were adjusted with age, sex (when required), smoking statuses and the first 2 principal components.

1.4.6. MOFA application in NMIBC and recurrence-free survival.

Multivariable Cox proportional-hazards model to evaluate the association between clinical covariates and MOFA latent factors with recurrence-free survival in 47 primary NMIBC patients (2 Tis, 28 Ta, and 17 T1). Longitudinal data were available for a median follow-up duration of 125 months (IQR 83-183), during which 27 patients experienced disease recurrence and 20 remained recurrence-free. We focused our analysis on recurrence prediction, as progression events were recorded in only two patients. A Cox model incorporating age, smoking status (3 never, 2 former, and 9 current smokers), multifocal disease (9) and WHO 1973 grades (10 G1, 21 G2 and 16 G3) achieved a concordance index (c-index) of 0.62 (95% CI= 0.49 - 0.76). Age, smoking status, nor lesion categories were not independently associated with recurrence risk. When the WHO 2004 grading system (17 low grade, 30 high grade) was used, the model's c-index declined to 0.55 (95% CI =0.41-0.68), indicating that the WHO 1973 grades provided better prognostic discrimination for recurrence within the TBCS cohort. Among the MOFA-predicted latent factors, Factor 5 was significantly associated with a reduced risk of recurrence (HR = 0.17, CI = 0.04 - 4.30, P = 1.63 x 10^{-2}). Incorporating age, smoking status, lesion categories and Factor 5 values yielded a c-index of 0.67 (CI = 0.57-0.77). No other latent factors were significantly associated with recurrence risk. We also performed a KM analysis in the 47 NMIBC cases, dichotomized into low (high factor 5, n = 24) and high (low factor 5, n = 23) risk of recurrence using the median value as cutoff (-0.25). The difference in recurrence-free survival between the two groups did not reach statistical significance (log-rank P = 0.21) (Figure 13). Among all omics layers, the transcriptomic data explained 5.48% of total variance in the data (SNP = 0.05% and methylation = 0.06%), with genes involved in natural killer cell-mediated immunity (*HCST*, *KLRD1*), positive regulation of intrinsic apoptotic signaling pathway by P53 class mediator (*MYC*), inflammatory

response (*PTGDR*), TRAIL-activated Apoptotic Signaling Pathway (*TNFRSF10A*), SCF-dependent Proteasomal Ubiquitin-Dependent Protein Catabolic Process (*FBXW5*), *SYTL2* and *SAMD3* carrying the most weights.

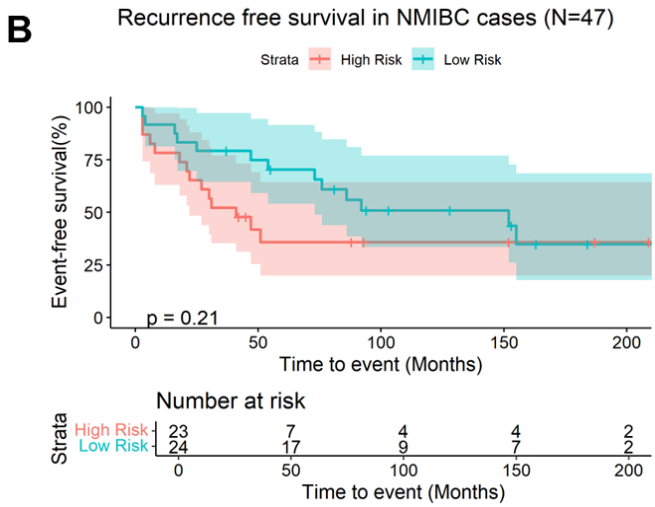


Figure 13: Kaplan–Meier (KM) analysis was performed to dichotomized NMIBC cases by risk of recurrence groups (high vs. low) based on the median value of factor 5. The analysis did not yield statistically significant association to recurrence (log-rank *P* = 0.21).

1.5. Discussion

BCa is known for its high recurrence rate, heterogeneous progression, and the need for lifelong surveillance[2]. However, the current clinical approach is heavily reliant on invasive procedures such as cystoscopy or non-invasive but insensitive tests, such as urine cytology[50,51].

Because of these limitations, there is an urgent need for accurate, reproducible, and non-invasive biomarkers that can detect disease early, predict progression, monitor for recurrence, guide therapy selection, and assess treatment response.

Despite advances in molecular profiling and genomics, there is still no widely adopted non-invasive biomarker that matches the sensitivity and specificity that are needed for routine clinical use. The challenges in identifying such biomarkers are many and include the biological variability of tumors (e.g., different grades, molecular subtypes, inter-individual variability), the lack of validation of candidate biomarkers in large and diverse cohorts[154], and, particularly relevant to this study, the difficulty in distinguishing tumor-related signals from the background confounding noise that characterizes the molecular profiling of readily accessible urine or blood, which are distantly related to the site of the cancer[128].

In this study, we sought to develop and test a framework for the generation of BCa biomarkers from blood, using a multi-omics approach that integrates genetics, transcriptomics and epigenomics. We conducted comprehensive multi-omics profiling and data integration with MOFA on TBCS-derived data, a very uniform and deeply phenotyped dataset from BCa patients and non-cancer controls enrolled from the Turin metropolitan area, in Italy[129,131–133,155,156]. We first conducted an external GWAS meta-analysis of three large, published studies, totaling 18,559 cases and 1,075,650 controls. This GWAS yielded 29 loci, of which four were novel. These novel associations pointed to variants nearest to *SH3TC2*, chr7q36.3, *RAPGEF1*, and *ARHGAP42* as susceptibility loci for BCa, expanding our current genetic understanding of disease and will require follow-up genetic and functional studies. We next conducted a GWAS in the TBCS-EPICOR dataset, which, not unexpectedly given the smaller sample size, did not reveal any genome-wide significant associations, but showed strong correlation of effect size at

significant loci from the meta-analysis, supporting the validity of this cohort for genetic data integration with transcriptomic and methylation data using MOFA.

Differential RNA expression showed 8 differentially expressed genes. Notably, the down regulation of *FEZ1* was previously identified in BCa cell lines and primary tumors, and its decreased expression was correlated with tumor grade and invasiveness[157].

Given that the non-dynamic nature of inherited genetic variation cannot provide information on cancer biology or behavior, we conducted multi-omics analysis using MOFA, leveraging on multidimensional data integration to capture the dynamic nature of disease presence[158]. MOFA analysis showed that multi-omics integration was superior to single-omics analysis and could potentially help in broadly distinguishing BCa cases from controls, with AUC superior to published PRS.

This supported the notion that systemic molecular signatures in the blood can potentially be applied to detect disease presence[159], highlighting the possible utility of blood-based profiling as a minimally invasive screening tool. Most previous multi-omic studies in bladder cancer have focused on tumor tissue[122], precluding direct comparison with our blood-based results. We identified several key molecular pathways through features carrying the most weight in MOFA. Among these, pathways associated with Phase II metabolic enzymes, specifically those involved in the conjugation of compounds are of critical interest. Notably, the enzyme glutathione S-transferases (GSTs) were involved in cellular detoxification processes by catalyzing conjugation reactions of carcinogens to glutathione. Consistently, GSTM1-null genotypes were previously associated with BCa susceptibility[160]. Furthermore, the dysregulation in this pathway involving glutathione conjugation is linked to drug metabolism and chemoresistance[161]. A single MOFA factor (factor 5) showed modest association (c-index 0.674) in predicting recurrence in the Cox model, which will need to be validated in external and larger datasets as “time-to-recurrence” may differ as our study used the date of first TURB, others may use the completion of BCG or re-resection as the initial time point.

Our study also emphasizes the importance of a replication cohort in both genetic association studies and multi-omics studies. In the GWAS, the significant correlation ($r = 0.7389$, $P = 7.12 \times 10^{-6}$) in the effect estimates at genome-wide significant loci between

BCa meta-analysis and TBCS-EPICOR GWAS supported the presence of shared genetic signals across studies.

While this study describes a detailed framework for conducting multi-omics analyses using GWAS and blood-derived transcriptomics and methylation data to potentially improve diagnosis and predict outcome in BCa, it represents, by design, a proof of concept with analytical and not clinical relevance and presents several limitations. First, this cohort underscored the challenges and variability inherent to genetic association studies in cohorts of small sample-size. Second, MOFA factors did not significantly separate cases based on tumor stage or recurrence, although this observation is not unexpected given that the input features were preselected based on case/control differences and not on tumor biological differences or longitudinal sampling. Additionally, even though the MOFA model identified latent factors that differentiated cases from controls, and at least one factor showed association with outcome, the results should be interpreted cautiously given the relatively small case cohort that may reduce the robustness of the derived factors in the training model. Finally, the absence of an independent cohort with matched multi-omics data precluded replication of the MOFA-derived features, limiting our ability to assess their generalizability across bladder cancer studies.

In conclusion, our study provides support to the notion that blood-based molecular signatures are biologically meaningful readouts for BCa and may potentially serve as early biomarkers of disease. Larger studies and integration of additional layers of omics data, including cell-free DNA (cfDNA), that have shown promising results in predicting recurrence, metastasis, and treatment response[162], will improve the generation of robust, sensitive, and specific biomarkers of disease. As such, the integration of cfDNA and whole-blood multi-omics may enhance precision oncology efforts, enabling non-invasive monitoring and risk stratification in both early- and late-stage BCa.

Chapter 2: Characterizing the CNV landscape in Idiopathic Nephrotic Syndrome (INS)

2.1. Introduction to Idiopathic Nephrotic Syndrome

2.1.1. Epidemiology and clinical presentation

Nephrotic syndrome (NS) is characterized by the presence of heavy proteinuria, hypoalbuminemia and peripheral edema resulting from podocyte injury leading to changes in the glomerular filtration barrier[163]. Complications within NS include venous thrombosis, hyperlipidemia, infection and acute kidney injury.

The incidence of NS varies between age groups and gender. The annual incidence of NS is 3 in 100,000 in adults, and 2-16.9 cases per 100,000 in children, depending on geographical origins[164,165]. Amongst children, incidence is highest in Asian children (~7 in 100,000), followed by Black (3.53 in 100,000), Hispanic (2.13 in 100,000) and White (1.83 in 100,000). Approximately 80 to 90% of NS cases are idiopathic, without identifiable underlying systemic, genetic or secondary forms of the disease[166,167]. These cases were traditionally classified based on histological patterns of kidney injury observed on kidney biopsy or by their response to immunosuppressive therapy. However, there has been a recent paradigm shift toward considering these disorders collectively as “podocytopathies”. Current understanding of monogenic causes of NS revealed that pathological lesions such as focal segmental glomerulosclerosis (FSGS) and minimal change disease (MCD) are not disease specific[168–170]. Instead, they represent overlapping patterns of podocyte lesions rather than distinct diagnostic entities. The same genetic defect may give rise to multiple pathological patterns, on the other hand, a histological pattern may be associated with diverse genetic etiologies and variable treatment response[163].

In terms of histological subtypes, membranous nephropathy has a higher prevalence in whites and FSGS were more often observed in blacks. Together they account for 30% to 35% of NS cases within adults. MCD and immunoglobulin A nephropathy (IgAN) occur in approximately 15% adult cases[171]. Conversely, childhood INS are most often attributed to MCD, accounting for approximately 74% to 85% of INS cases in patients under age 5[165,172,173]. FSGS develops later at a median age of 6 years. INS also has a male predominance and occur at a ratio of 3:1 male-to-female ratio in children[174].

Secondary causes of nephrotic syndrome can arise from a variety of systemic and hereditary conditions. Common acquired causes include diabetic nephropathy, systemic lupus erythematosus and amyloidosis, malignancies, drug exposure and viral infections. In addition, several congenital disorders such as Alport's syndrome, Congenital Nephrotic Syndrome of the Finnish type, Pierson's syndrome, Nail-patella syndrome and Denys Drash syndrome are also recognized contributors to secondary forms of INS[167].

Patients presenting with heavy proteinuria often undergo a series of serologic testing and kidney biopsy. Corticosteroids immunosuppressive therapy is the first line of treatment for INS patients if there were no directly treatable underlying cause. Patients who are responsive to steroid (steroid sensitive nephrotic syndrome, SSNS) generally have a good long-term prognosis and are classified as having primary INS. By contrast, patients who are non-responsive (steroid resistant nephrotic syndrome, SRNS) often progress to end stage kidney failure (ESKD) typically within 5 years in monogenic forms and within 10 years in non-genetic forms of disease[175](Figure 14).

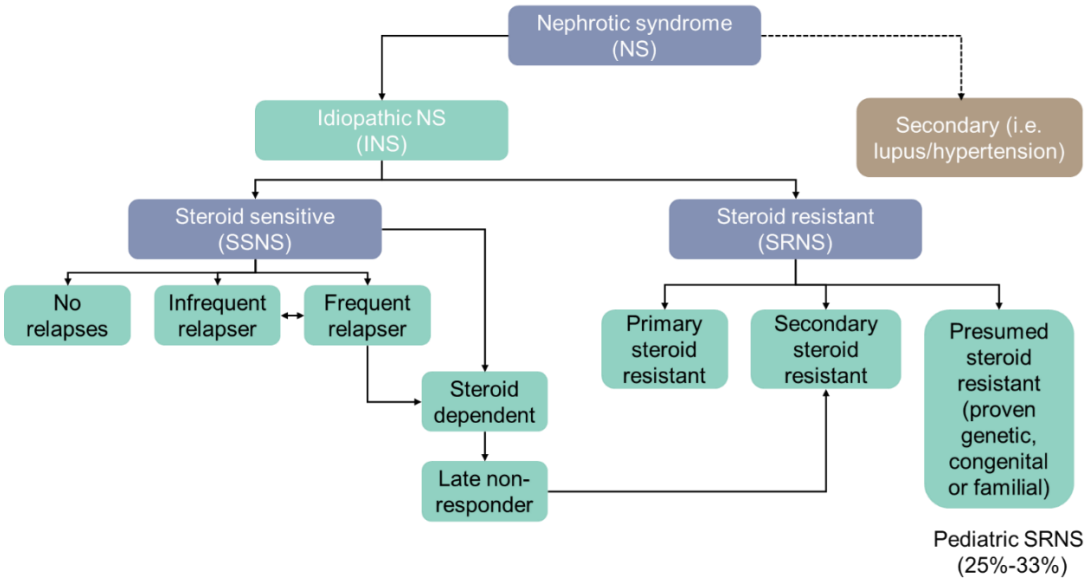


Figure 14: Current molecular stratification of INS reproduced from Saleem, 2025[176].

2.1.2. Diagnosis and treatment

Urinary dipstick

Initial assessment of NS differs between children and adults. In children, NS frequently presents with edema due to hypoalbuminemia and excess saline. Proteinuria is defined as urinary protein to creatinine ratio (uPCR) > 0.2 or proteinuria > 100 mg/m^2 per day. A urinary dipstick test can be used, however reports found that a positive dipstick for protein will occur in 5 to 10% of school-aged children and adolescents[177], persistent proteinuria decreases to 0.1% after repetitive testing (2 positive urine dipstick detections over the period of more than 3 months)[178]. Patients manifesting syndromic features should undergo further phenotyping and checked for hearing and visual loss to rule out syndromic podocytopathy or other specific syndromes[179–181]. In adults, nephrotic and sub-nephrotic ranged proteinuria are most frequently identified incidentally during routine urinalysis. In this case, nephrotic-range proteinuria is characterized by a uPCR >= 300 mg/mmol or a 24-hour protein excretion of >= 3.5g in the presence of normal serum albumin. By contrast, sub-nephrotic proteinuria corresponds to a uPCR < 300mg/mmol or a daily protein excretion of 300-3,400 mg per day. Differential diagnosis should be performed to rule out other potential causes including syndromic features, exposure to

drugs, toxins, viral or bacterial infection, autoimmune diseases, cancer and increased body mass index. Nonetheless, adults presenting with marked edema, venous thrombosis or infections warrants an investigation for a podocytopathy of immunological or genetic cause[167,171].

Kidney biopsy

Most children with isolated NS do not typically undergo a kidney biopsy as the majority would present with MCD and achieve complete remission after standard steroid therapy[174,182]. Moreover, steroid responsivity serves as a more reliable long-term prognostic indicator than kidney biopsy, since 10 to 20% of children with INS are diagnosed with either MCD or FSGS. A retrospective analysis performed by Hama et al. 2012 revealed that a uPCR cut-off of >= 0.5 g/g in children with asymptomatic, isolated persistent proteinuria increases the likelihood of FSGS lesion during biopsy. Using this criterion, kidney biopsy could be sparred in children with lower uPCR at < 0.5 g/g as the risk of FSGS lesion is low[183]. Kidney biopsy may be required if the patients show signs which make MCD less likely including hematuria, hypertension, reduced renal function and low serum complement levels. On the other hand, kidney biopsy is generally performed in adults with nephrotic-range proteinuria, with the most common findings being MCD, FSGS and collapsing glomerulopathy. Podocytopathies can be differentiated from other glomerular diseases using light microscopy and immunoglobulin staining. Additionally, higher resolution electron microscopy can also be used to inform the extent of foot process effacement (FPE) and abnormalities of the glomerular basement membrane (GBM).

Genetic testing

Children and young adults (<30 years old) who are non-responsive to glucocorticoid are usually referred for genetic testing, as the possibility of a positive molecular finding for genetic disorder is approximately 30% depending on age stratifications[179,184,185]. In adults with SRNS, molecular findings are reported in approximately 14% to 21%, with the higher estimate largely driven by the inclusion of high-risk alleles carriers in the *APOL1* gene[185–187]. Genetic variants associated with FSGS and SRNS will be further discussed in section 2.1.3.

Steroids as first-line of therapy for NS management in children and adults

The response of NS patients to glucocorticoid therapy strongly influences prognosis[188], with the majority of those with MCD (75–92%) and a subset of other patients with FSGS (47–66%) achieving remission and not progressing to ESKD[189]. SRNS is a major predictor for kidney function decline, with kidney survival rates falling to 30% within 10 years from diagnosis. The prognosis for patients with genetic podocytopathies is poor with an estimated 50% developing ESKD within 5 years of diagnosis[179]. Patients with non-nephrotic proteinuria are generally managed with Renin-Angiotensin System inhibitors (RASi) and dietary salt restriction[190]. In children and adolescents presenting with newly onset isolated NS who have not undergone kidney biopsy, initial treatment with oral steroids for 2 to 3 months is recommended[191]. Among SSNS children, complete remission is typically achieved in 80-90% within 4 weeks of therapy initiation. However, only 30% of these children maintain long-term remission. Approximately 10-20% experience infrequent relapses (<4 episodes), while the remainder suffer from frequent relapses (FRNS). Given this heterogeneity in clinical course and steroid responsiveness, treatment regiments should be personalized in dose and duration[192]. By contrast, SRNS is defined by the absence of complete remission or a modest partial remission, at best. Once SRNS is ascertained, patients should be referred for biopsy and genetic testing. Similarly, glucocorticoids are the first line of treatment for NS in adults. However, adults typically require a longer course of treatment due to a slower response to therapy prior to defining treatment failure. Immunosuppressants such as Mycophenolate mofetil can be administered in conjunction with low-dose steroids to induce remission comparable with standard therapy[193,194].

Most children and adult SSNS patients experience relapse(s)[163]. Over time, many progress to steroid dependent (SDNS) or develop frequent relapses (FRNS) after stopping treatment, with the risk of relapsing being highest in children younger than 5 years at onset. This often requires a dose adjustment above the individual threshold and can be managed in two ways: 1) alternate-day steroids dosing in children, 2) the use of steroid-sparing agents including immunosuppressants such as ciclosporin or tacrolimus (both calcineurin inhibitors), rituximab, levamisole and cyclophosphamide.

Steroids are rarely effective in SRNS patients with a genetic cause. Therefore, immunosuppressive treatment can be stopped when a genetic cause has been identified, and anti-proteinuric treatments such as RASi can be initiated to slow CKD progression. An estimated 60% of SRNS patients, typically those without identifiable genetic causes, respond to calcineurin inhibitors by showing reduced proteinuria and slowed CKD progression. Thus far, there are no drugs that could maintain remission in these individuals. Certain cases warrant avoidance of unnecessary immunosuppressive therapy, particularly in patients with genetic form of the disease for which alternative treatment options are available. For instance, patients with pathogenic variants in the coenzyme Q_{10} biosynthesis genes (*COQ2*, *COQ6*, and *ADCK4*) may respond to oral supplementation of coenzyme Q10. Additionally, patients with pathogenic variants in collagen genes will benefit from early diagnosis by avoiding calcineurin inhibitors altogether and by administration of RASi. Patients with *WT1* pathogenic mutations positioned in exon 8 and 9 will benefit from screening for Wilms tumor malignancies and may opt for prophylactic nephrectomy.

2.1.3. Genetic determinants of INS

Rare variants

SRNS often has an unclear etiology, but many cases are increasingly recognized to have a genetic basis. Monogenic mutations account for a significant proportion of SRNS cases, especially in children. To date, more than 60 monogenic causes of SRNS have been identified, most of which follow a recessive inheritance pattern. These genes predominantly affect the structure and function of the glomerular podocyte and slit membrane when mutated (Figure 15)[195]. Overall, pathogenic variants in NS-associated genes account for up to 25%-33% of childhood cases. Of these genes, mutations were commonly found in Nephrin (*NPHS1*), Podocin (*NPHS2*) and Wilms tumor suppressor-1 gene (*WT1*)[175,185,192,196,197]. The genetic landscape in adults differ, with most pathogenic variants in SRNS identified in *COL4A3-5* genes (38% with familial FSGS, 3% with sporadic FSGS, and most mutations occurring in *COL4A5*)[198].

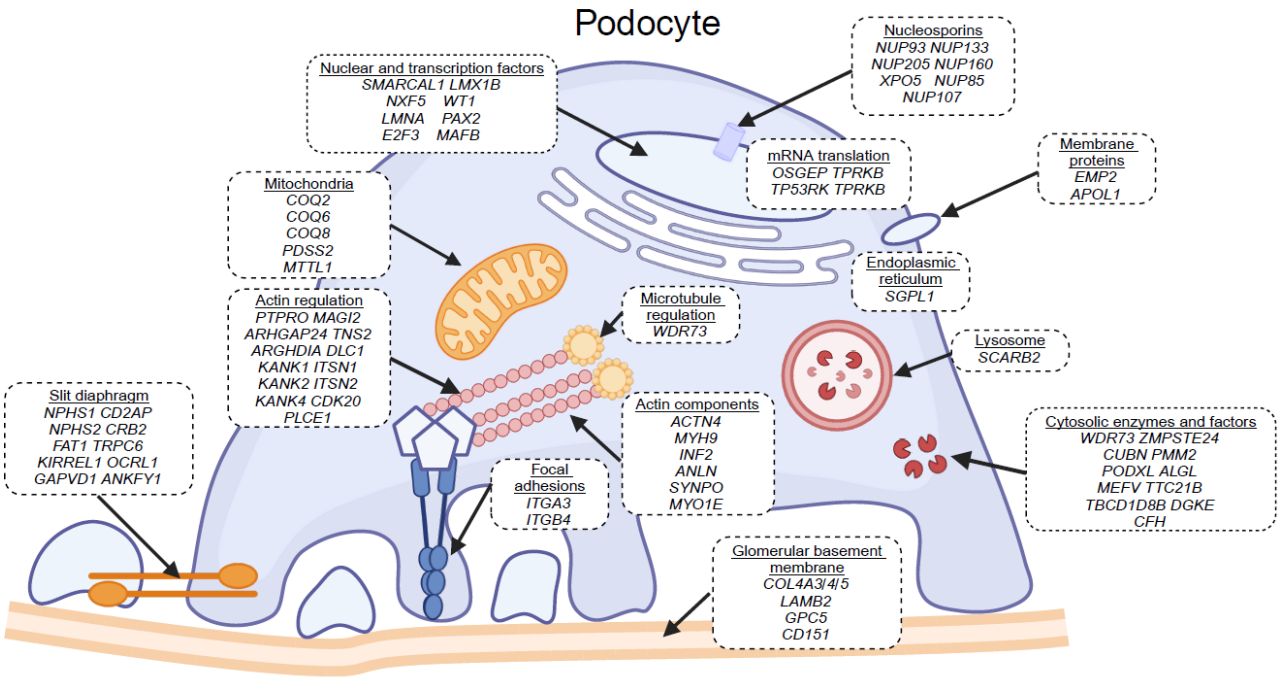


Figure 15: Over 60 gene mutations have been linked to the development of INS or podocytopathies. Figure was adapted from Meliambro, He, and Campbell, 2024[199]. Created in https://BioRender.com.

NPHS1 encodes Nephrin, a transmembrane protein of the immunoglobulin superfamily that is essential for slit diaphragm formation. Biallelic pathogenic variants in *NPHS1* result in absent or dysfunctional slit diaphragms, leading to massive proteinuria and kidney failure, the hallmark features of congenital nephrotic syndrome of the Finnish type (CNF)[200]. Similarly, disruption of *NPHS2,* encoding podocin, a key component of the slit diaphragm, critically impairs the structural and functional integrity of the glomerular filtration barrier[201] causing early onset nephrotic syndrome which is resistant to immunosuppressive therapy. The Wilms tumor suppressor-1 gene (*WT1*) encodes a zinc-finger transcription factor that is necessary for kidney and gonad development. Heterozygous mutations in WT1 cause distinct syndromes characterized by NS associated with varying risk of Wilms tumor and gonadal abnormalities such as Denys-Drash syndrome and Frasier syndrome[202]. WT1 mutations occurring in exons 8 and 9 are associated to isolated SRNS[182].

Common variants

GWAS have identified numerous susceptibility loci for INS, particularly in pediatric SSNS (pSSNS)[203–209]. To date, no monogenic forms of pSSNS have been confirmed, but studies suggest that genetically conferred risk of immune dysregulation is a major contributor with established associations in the HLA class II region[210,211]. A GWAS meta-analysis from our groups, totaling 2,440 pSSNS cases and over 36,000 controls from 12 studies of mixed ancestry populations, identified 12 genome-wide significant associations, including two independent signals fine mapped to *HLA-DQB1* and MHC Class I Chain-related Gene A (*MICA*). Additional loci including Calcium Homeostasis Modulator Family Member 6 (*CALHM6*), TNF Superfamily Member 15 (*TNFSF15*), Abelson Helper Integration Site 1 (*AHI1*), betacellulin (*BTC),* CD28 molecule (*CD28*), C-type Lectin Domain Containing 16 A (*CLEC16A*)[212] point to roles in immune response regulation. Interestingly, a non-coding pSSNS risk variant near *NPHS1*, a monogenic kidney gene, may act by modulating Nephrin expression in kidney cells. Several associations were ancestry-specific: variants at the *CALHM6* locus conferred increased risk in Europeans, whereas those near *TNFSF15* and *NPHS1* were associated with increased risk in East Asians.

APOL1

One of the most extensively studied associations with chronic kidney disease (CKD) and FSGS involves *APOL1*. Two common protein-altering mutually exclusive haplotypes, termed “G1” and “G2”, when co-inherited in a recessive fashion, exert strong effects (OR = 10.5-29 for FSGS; OR ~4 for non-diabetic CKD) despite their high population frequency[213–215]. These variants largely explain the increased risk of FSGS (17 to 30-fold) lesions and CKD (3 to 5-fold) observed in individuals of West African ancestry compared with those of European ancestry. Among African Americans, about 13% carry the recessive high-risk *APOL1* genotypes (G1/G1, G1/G2 or G2/G2). G1 is generally more common than G2 in this population with reported allele frequencies of 23% and 13%, respectively[216]. ApoL1 is a serum factor that lyses trypanosomes. In vitro studies have demonstrated that the kidney disease-associated *APOL1* variants confer resistance against African trypanosomes, providing an evolutionary advantage against sleeping

sickness among heterozygous carriers. Specifically, the *APOL1* G2 allele is strongly associated with resistance to *T. brucei rhodesiense* whereby carriers of a single copy of G2 have a 5-7 fold reduction in susceptibility to the disease[217,218]. Conversely, the G1 allele protects against T. brucei gambiense and is more common in West African population[217,219]. While these variants provide protection against parasitic infection, individuals carrying the two risk alleles face a markedly increased risk of kidney disease, and regions with high *APOL1* risk alleles frequencies reports CKD prevalence as high as 16%[220]. Transgenic animal studies revealed that the expression of either APOL1 risk variants is sufficient to induce FSGS, global glomerulosclerosis and CKD whereby disease severity is highly correlated with increased APOL1 expression levels, while the G0 allele, even when overexpressed in mice (which do not have a Apol1 gene), is not able to induce disease[221].

2.1.4. Copy Number Variation (CNV) detection and genomic advances

Structural variants (SV) (Figure 16) is an umbrella term representing large (>50bp) structural alterations and rearrangements in the genome involving deletion, duplication, inversions, translocation and complex rearrangements[222–224]. Copy number variants (CNV) are a subset of SV involving the gain of copies (duplication) or loss (deletion) of DNA segments larger than 1 kilobases (Kb), deviating from the normal diploid state[222]. These variants can influence disease risk by changing gene structures, dosage sensitivity and regulatory effects[225].

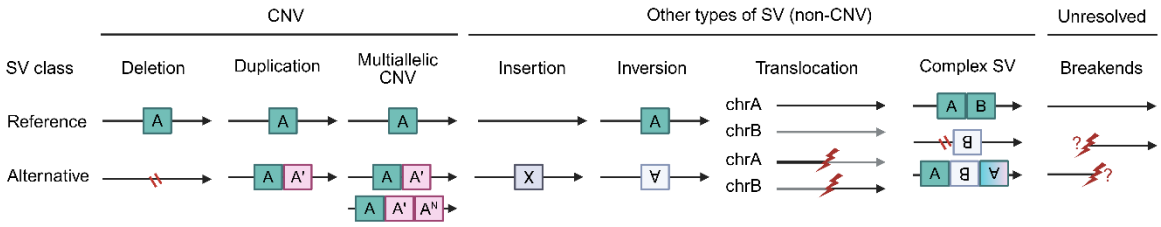


Figure 16: Types of SV classes detected in population studies adapted from Collins et al. 2020[226]. SV can be grouped into different classes which included gains or losses of DNA

observed in CNVs and multiallelic CNVs, and balance rearrangements that do not alter dosages such as inversion and translocation. The advent of long read sequencing enabled the detection of complex SVs involving at least two or more SV classes in a single mutational event, ranging from CNV-flanked inversions to chromothripsis, a rare event defined by chromosome shattering involving extensive rearrangements of one or more chromosomes[227]. Created in https://BioRender.com

Overall, an average genome carries 3 to 7 rare CNVs, 5% to 10% of individuals carry CNVs larger than 500Kb, whereas a smaller proportion of 1% to 2% of individuals carry CNVs larger than 1Mb. The size and gene density of a CNV are inversely correlated with allele frequency[228]. Approximately 10% of a human genome have recurrent CNV events, known as “CNV hotspots”[229]. Understanding the cause of the pathogenicity of a CNV remains challenging. Some genes can be completely deleted without causing an apparent effect. Zarrei et al. 2015[223] constructed an updated CNV map by analyzing published CNV data from 2,647 healthy individuals across 23 studies in the Database of Genomic Variants (DGV, https://dgv.tcag.ca/dgv/app/home, accessed on 26 September 2025). They identified CNVs that are homozygously deleted, referred to as null CNVs. These null CNVs are modestly enriched in genes with paralogues and encompassed approximately 107 protein coding genes in which at least 85% of the exons are deleted[223]. Benign CNVs are typically small, intergenic or contains genes that can tolerate copy number alterations. By contrast, pathogenic CNVs are enriched for genes involved in development and for genes that are evolutionary constrained for copy number gains and losses[230]. The American College of Medical Genetics (ACMG) standards and guidelines for the interpretation and reporting of postnatal constitutional CNVs[231], and for broader CNV interpretations[232] were developed to help evaluate and promote consistent reporting of CNVs. These consider variables such as genomic disorders, size, genomic content, comparison of CNV frequencies, dosage, reported pathogenicity with internal and external databases, and whether the CNV is inherited or de novo in variants of uncertain significance (VUS).

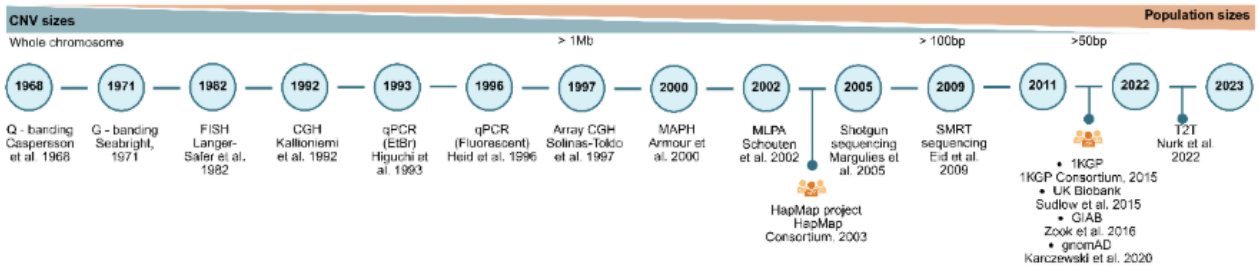


Figure 17: SV detection methods over the past century have enabled the detection and characterization of diverse SV classes at higher resolution in large cohort studies, figure adapted from Lim et al. 2023[233]. Created in https://BioRender.com

Large but rare CNVs (<1% in population) have been implicated for neurological and neurocognitive disorders[234–238]. Furthermore, multicopy gene families that are frequently subjected to copy number changes can play a role in disease susceptibility, particularly in immune diseases[239,240]. For example, a low copy number in FCGR3B was associated with glomerulonephritis in the autoimmune-related systemic lupus erythematosus[239]. Therefore, a comprehensive characterization of SV is crucial to fully understand their contribution to diverse phenotypic traits. Over the past few decades, SV detection and analysis have advanced considerably, evolving from karyotype banding for the identification of large SVs and chromosomal abnormalities with sizes ranging from 5-10 Mb, to array Comparative Genomic Hybridization (aCGH) and chromosomal microarray (CMA)-based technologies for CNVs, to short-read sequencing (SRS) for SNV detection, and more recently, to long-read sequencing (LRS) for the detection of complex structural variants (CSV). Collectively, these advancements have expanded our understanding of human genetic variation by capturing different classes of variants. Genome-wide approaches now enable the detection of SVs at a higher resolution, including the phasing and characterization of CSV (Figure 17). Currently available real-time long read sequencing platforms (SMRT) such as Pacific BioSciences (PacBio) and Oxford Nanopore produce an average read length that exceeds 20 kb with uniform coverage and has shown superior performance over SRS for SV discovery (average read length in SRS 100 bases)[241,242]. These enable complex rearrangements, especially those in repetitive regions to be solved in rare diseases[243–247]. Chaisson et al reports a six-fold

increase of SV discovery using SMRT leading to 26,079 euchromatic insertion/deletion to be resolved[248], another work by the Human Genome Structural Variation Consortium (HGVSC) demonstrated a three to seven-fold increase in SV detection using multiple SMRT platforms[249]. Additionally, these technologies were also crucial to aid the completion of the first human genome termed “T2T-CHM13”, uncovering 200 million base pairs of sequences with 1956 gene prediction and unlocking new complex regions of the genomes for further variation and functional studies[250]. To investigate the prevalence of complex structural variants (CSV), the HGVSC constructed a pangenome reference by sequencing 65 diverse human genomes. This effort produced 130 haplotype-resolved assemblies, closed 92% of gaps present in earlier draft human pangenome assemblies[251,252], and achieved telomere-to-telomere completeness for 39% of chromosomes. On average, each individual was found to harbor 26,115 SV, and 72 CSVs per genome[247]. Comprehensive population-scale SVs[253] generated from LRS are now available as a resource for benchmarking data and variant prioritization (https://www.internationalgenome.org/data-portal/data-collection/hgsvc3 , accessed on 24 September 2025). SV detection tools that have been adapted for LRS include Sniffles[254],DELLY[255] for linear genomes such as GRCh37, GRCh38 and CHM13, SVarp[256] for Pangenome graph reference (HPRC).

Array-based technologies

Together, these advances show how LRS and complete assemblies such as the T2T-CHM13 and pangenome assemblies are reshaping the discovery and characterization of SV[257,258]. Nevertheless, CMA-based approaches for CNV detection continue to be used in both clinical and research settings[259–261], owing to their affordability and mature analytical pipelines[262]. By 2010, CMA became the frontline assay to replace cytogenic approaches to detect CNV as a diagnostic test[263]. In this thesis, CNV detection was performed using CMA approaches, particularly with the Illumina genotyping chips. These methods typically capture CNVs ranging from 50kb to over 1 Mb in size, predominantly located in non-repetitive genomic regions. The ability to accurately genotype CNV breakpoint regions remains a major challenge. Smaller CNVs (<100Kb) have lower

sensitivity and specificity due to probe density and technical noise[264,265]. Although originally designed to genotype SNPs for GWASs[266], CMA platforms have also been repurposed for rare and de novo CNV detection since probe intensities can reveal dosage change. This adaptation enabled the discovery and characterization of genomic disorders (GD) associated to neurodevelopmental disorders[235,267] and congenital anomalies[268–270]. Importantly, these studies underscored the role of ultra-rare, large-effect CNVs in human diseases, with a pronounced impact in pediatric cases. An estimated 14.2% of children with developmental delay carry ultra-rare large, de novo CNVs (>400Kb) that were seldom observed in unaffected individuals and are often regarded as the primary cause of the disorder[271]. In a trios study of 114 families with sporadic Tetralogy of Fallot (TOF), a severe congenital heart malformation without additional anomaly, rare de novo CNVs were identified in 11 probands, accounting for ~10% of cases. These CNVs were either absent or observed in less than 0.1% of population controls. Additional recurrent CNVs were also observed at 3p25.1, 7p21.3 and 22q11.2. Notably, one TOF patient carried CNVs at 6 loci, two of which disrupted established disease-associated genes such as *NOTCH1* and J*AG1*, highlighting the important role of dosage-altering mutations in genes essential for cardiac development[268].

Most pathogenic CNVs discovered to this day are still relatively large and span across multiple genes, although single genes have been implicated for some GD such as the 17q21.31 and KANSL1 gene[272], and in some cases, the chromosome 16p11.2 microdeletion associated with autism were attributed to the imbalance of multiple genes contributing to the phenotypic features[273]. Considerable amount of variable expressivity is commonly observed with the same CNV giving rise to markedly different disease outcomes. Eventhough the associated disease risk is established, the specific phenotypic consequences of most large CNVs remain poorly defined, and their effect-modifying loci have yet to be fine-mapped[271].

2.1.5. CNVs studied in kidney diseases

Copy number variants (CNVs), defined as genomic deletions or duplications larger than 50 base pairs[274], are increasingly recognized as important contributors to the genetic

architecture of kidney diseases. Large cohort studies have established their role across several monogenic kidney diseases including congenital anomalies of the kidney and urinary tract (CAKUT)[1,275–277], Nephronophthisis[278], and CKD[279].

Within CAKUT, 4% to 10.5% of cases have been attributed to pathogenic genomics disorder CNVs (GD-CNV), which are large structural rearrangements ranging from 100Kb to 155.3Mb[1]. In a cohort of 192 patients with Renal Hypodysplasia (RHD), GD-CNVs with potential gene-disrupting effects were identified in 10.5% of cases, compared with only 0.2% of matched controls[276]. Deletions involving loci such as HNF1B (17q11–12) and the DiGeorge/velocardiofacial region (22q11.21) were frequently observed. Similarly, the KIMONO-GENE cohort of 80 CAKUT patients reported GD-CNV in 6% of cases[275]. A larger study of 2,824 patients further confirmed the enrichment of CNVs in CAKUT, identifying GD-CNVs in 4% of cases and additional 1.7% carrying large, rare diagnostic CNVs[1]. The CNV yield and the deletion versus duplication contribution varied among CAKUT subcategories, with rare large deletion being more frequent in kidney anomalies, while duplications more frequent in lower urinary tract defects.

Beyond CAKUT, CNVs have also been implicated in CKD. A recent study involving 2,432 patients with general kidney disease at the University Medical Centre Utrecht reported that CNV alone accounted for 2.4% diagnostic yield[279].

Compared with CAKUT, the contribution of CNVs to nephrotic syndrome (NS) has been less explored. Based on few studies of very limited sample size, the yield was approximately ~1.5% in steroid-resistant NS (SRNS)[1,277]. Nevertheless, many studies have highlighted that CNV analysis remains a valuable complementary tool to SNV detection and can reveal actionable diagnoses, particularly in the context of NS[280]. Nagano et al.[281] detected CNVs in seven cases with inherited kidney diseases, including two with Alport syndrome carrying deletions encompassing the X-linked *COL4A5* phenocopy gene. Nakanishi et al.[280] described an infantile SRNS case in which sequencing initially identified a single pathogenic missense mutation in exon 7 of the autosomal recessive *COQ6*, gene; subsequent array comparative genomic hydbridization (aCGH) paired CNV analysis revealed a deletion spanning exon 1-2 of the same gene. This combined approach established a compound heterozygous cause, enabling early

coenzyme Q10 supplementation leading to complete remission. More recently, we contributed to a genome-wide microarray CNV screening in 138 “genetically-unresolved” SRNS families[282] in which a SNV-only genetic cause could not be determined through a survey for pathogenic or likely pathogenic single nucleotide variants in 60 known NS-associated genes and in 13 phenocopy genes. Novel CNVs were identified in 1.5% cases (2 of 138 families) which encompassed homozygous deletions *PLCE1* and *NPHS2*, two genes with reported autosomal recessive mode of inheritance.

2.1.6. APOL1 genotypes and CNV

Ruchi et al. 2015[283] analyzed sequences from the 1000 genomes project phase 3 and exome data from African American kidney disease cases to investigate if alteration in SV encompassing *APOL1* affect phenotypic variability. This spans a ~100Kb region which includes *APOL2*, *APOL1* and segments of *MYH9*. This work identified rare duplications that were significantly enriched among kidney-disease cases compared to controls within individuals genotyped as apparent G0/G1 heterozygotes (4.06% vs 0.78%). These findings raise the possibility that some individuals classified as single-risk-allele carriers may actually harbor additional APOL1 copies (such as G1/G1G1), effectively carrying multiple risk alleles and potentially explaining the increased risk observed in nominal heterozygotes. However, CNV duplications were not commonly observed in kidney disease patients in Chinese FSGS patients[284].

2.2. Hypothesis and aims

Idiopathic Nephrotic Syndrome (INS) is a heterogenous trait characterized by heavy protein loss in the urine, low serum albumin, edema, and is a frequent cause of kidney failure. The utility of genetic diagnostics for rare variants in Mendelian INS-associated genes has been reported in patients with various kidney diseases. However, genetic studies have overwhelmingly focused on the contribution of single nucleotide variants, leaving the role of rare Copy Number Variations (CNV) in INS largely unexplored. We hypothesize that rare CNVs in INS-associated genes may contribute to the diagnostic yield of INS, particularly in subset of patients who are genetically “unsolved” and may reveal ancestry or phenotype-specific patterns of genetic risk.

DNA chromosomal microarray (CMA) genotyping was conducted on a multi-ancestry cohort of 3,144 INS patients across ages of disease onset and response to therapy to define the burden of CNVs in INS and a set of established 21,498 population controls[1]. We sought to identify gene and locus specific CNVs of diagnostic relevance by characterizing CNVs intersecting known GD-CNV and INS-associated genes, and evaluate their contribution across the clinical subgroups.

2.3. Methods

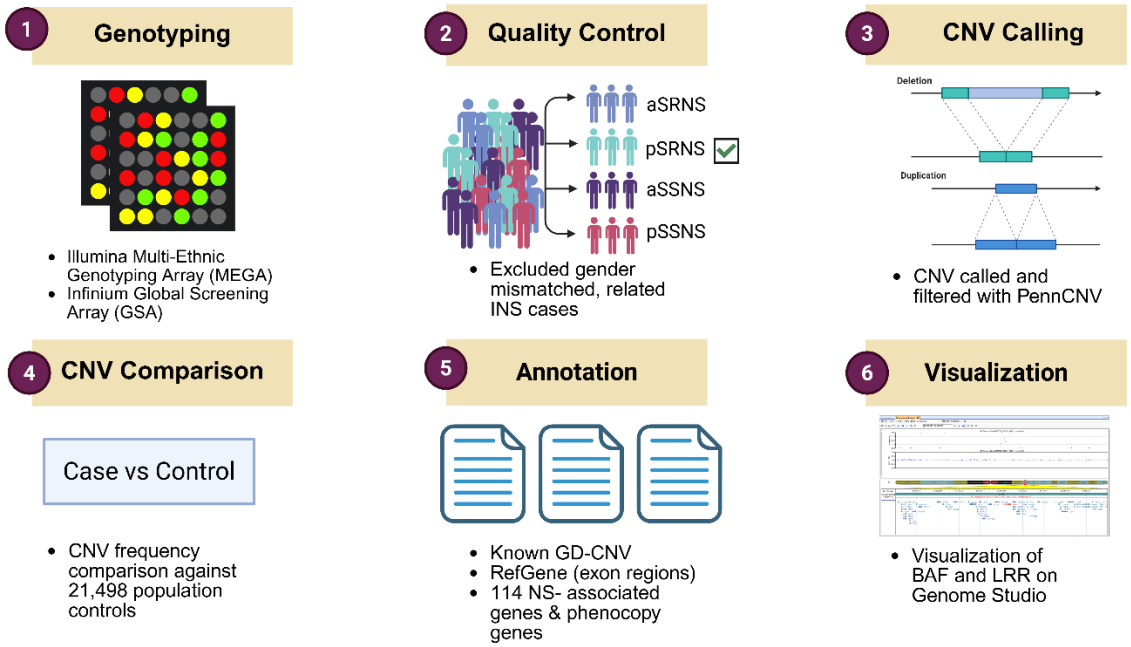


Figure 18: Analysis workflow to systematically characterize CNVs within INS across ages of disease onset and response to therapy. INS cases were genotyped with Illumina MEGA and Infinium GSA arrays. Subsequently, CNVs were called with PennCNV using previously established workflow from Verbitsky et al. 2019[1]. We annotated the CNV start and stop boundaries to 221 known Genomic Disorder regions (GD-CNV), RefGene, and a list of 114 genes that when mutated are known to cause INS of a phenocopy of it, accounting for mode of inheritance. Image created with BioRender.com.

2.3.1. Ethics approval and INS cohort stratifications

All participants recruited at Columbia University and collaborating institutions provided written informed consent under Columbia University Institutional Review Board (IRB) Protocol AAAC7385. International participants provided consent in accordance with the Declaration of Helsinki, with approval from local ethics committees under the parent IRB protocol at Columbia University.

CMA genotyping was performed using multiple iterations of the Illumina Multi-Ethnic Global Array (MEGA) including MEGA 1.0, MEGA 1.1 and MEGA^{EX}, and the Infinium Global Screening Array (GSA). The INS cases consisted of 3,144 individuals from multiple

genetic ancestries, spanning different categories of ages at disease onset (pediatric or adult, with a cutoff of 21 years old) and response to corticosteroid therapy (resistant or sensitive). This stratification resulted in five analytical subgroups: adult steroid sensitive nephrotic syndrome (aSSNS), adult steroid resistant nephrotic syndrome (aSRNS), pediatric steroid sensitive nephrotic syndrome (pSSNS), pediatric steroid resistant nephrotic syndrome (pSRNS), and “Other” for persons with unknown age of disease onset and/or response to therapy. Case inclusion criteria comprised individuals with biopsy-proven FSGS or MCD. For individuals without biopsy, inclusion required clinical features consistent with INS, including steroid-sensitive (SSNS) or steroid-resistant (SRNS) disease across all ages of onset. Exclusion criteria included cases with membranous nephropathy, IgA nephropathy, diabetic nephropathy, or any secondary form of proteinuria/NS, such as obesity and prematurity. Population controls comprised 21,498 individuals without reported association to nephropathy or developmental disorders as previously described in Verbitsky 2019[1].

2.3.2. Genotyping data processing, ancestry inference and CNV calling

Raw genotyping data were processed in Illumina GenomeStudio v2011 (https://www.illumina.com/products/by-type/informatics-products/microarray-software/genomestudio.html, accessed 10 September 2025) to generate probe-level chromosome, position, log R ratio (LRR) and B allele frequency (BAF) values for CNV detection. LRR represents the normalized signal intensity of each SNP in a SNP array, calculated by obtaining the log2 ratio between the observed and expected signals of a SNP in the genome. Post-normalization LRR values for diploid SNPs are usually clustered around zero, with higher values (> 0) indicating a duplication event and lower values (< 0) suggesting a deletion[285]. On the other hand, B Allele Frequency (BAF) quantifies the relative proportion of the B allele compared to the total signal intensity (A and B alleles) in the SNP array. BAF values range from 0 for homozygous A/A alleles, 0.5 for diploid heterozygous alleles (A/B), and 1.0 for homozygous B/B alleles. In the presence of copy number alterations, BAFs deviates from these expected values: for

single deletions (copy number =1), values cluster near 0 and 1, while for single duplications (copy number=3), values typically cluster near 0, 0.33, 0.67 and 1[286]. Gender estimates and PLINK-compatible map and ped files were also obtained from the genotyping module PLINK Input Report Plug-in v2.1.4(https://support.illumina.com/array/array_software/genomestudio/downloads.html, accessed 10 September 2025) for further data quality controls. Samples with discrepancies between self-declared and estimated gender were excluded. Relatedness was assessed using common SNPs, and individuals with kinship coefficients >= 0.0844 (second-degree relatives or closer) were removed. Ancestry inference was performed with KING[287] using the 1000 genome reference panel[288]. The *APOL1* G1 (rs73885319) and G2 (rs71785313) were directly genotyped in the Mega arrays, and imputed (TopMedv1) in the GSA arrays. We defined “High Risk” APOL1 persons based on the G1/G1,G1/G2 and G2/G2 genotypes and “Low Risk” *APOL1* based on the G0/G0, G0/G1 and G0/G2 genotypes.

CNV calling was performed with the PennCNV software (version 2011-05-03) which uses a hidden Markov model (HMM) model to detect kilobase-resolution large CNVs from high density SNP genotyping data[289] by incorporating information such as total signal intensity, allelic intensity ratio of each SNP, distance between SNPs and allele frequencies of SNPs. All CNVs were initially called in the Hg18 genome assembly coordinates with default quality control metrics for LRR standard deviation, BAF drift and waviness factor. CNVs with a minimum confidence score of 30 were retained for downstream analysis. CNV start and stop coordinates were then converted to Hg19 using the UCSC liftover tool[140] for annotations.

2.3.3. Generation of the Nephrotic Syndrome genes list

We curated a list of 114 genes known to cause glomerular disease based on a literature review of previously published studies[179,197]. These included “podocytopathy” genes that were involved in podocyte structure or function, and “phenocopy” genes that are typically not known to be associated with INS but mimics similar syndromic presentations.

2.3.4. CNV annotations

Predicted CNV start and stop coordinates were annotated with: RefGenes (N=56,263 transcripts), 221 known genomic disorder CNVs (GD-CNV) curated from the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER)[290,291] and International Standards for Cytogenomic Arrays (ISCA) databases[263]; a panel of known kidney disease genes (N=642); and the previously described 114 nephrotic syndrome gene panel. CNV were prioritized according to established criteria[1,277] using a stepwise approach. Briefly, a stepwise filtering was performed. First, CNVs overlapping at least 70% of a known GD-CNV were classified as pathogenic GD-CNVs. Second, CNVs intersecting an exon of any of the 114 INS-associated genes, irrespective of CNV sizes, were classified as either likely pathogenic or variant of unknown significance (VUS), depending on the gene's known dosage sensitivity and its compatibility with the observed CNV (any size within NS-associated genes). Third, a candidate likely pathogenic was defined as a rare, large CNV (>=100Kb) intersecting at least one exon of a RefGene, present in less than 0.02% of the 21,498 population control, and were not overlapping any clinically interpreted benign or likely benign CNVs in the ISCA databases. These candidate variants were further prioritized based on presence/absence of haploinsufficient genes, expression in kidney (podocyte), established mouse models.

2.4. Results

2.4.1. Overall study characteristics

The study cohort comprised 3,144 individuals with INS (Table 14), including 1,809 males and 1,335 females. Age at disease onset was well distributed, with 1,488 adults (47.33%), 1,483 pediatric cases (47.17%), and 173 individuals with unreported onset age. KING ancestry inference identified representation across five major superpopulations: European (2,218, 70.6%), African (365, 11.6%), Admixed American (267, 8.5%), South Asian (138, 4.4%), East Asian (75, 2.4%), and multiple ancestries (81, 2.6%). Most participants were resistant to corticosteroids (1,876, 59.7%), while 1,162 were steroid-sensitive (36.9%); steroid responsiveness was not available for 106 patients (3.4%). Stratification by age at disease onset revealed 1,014 adult SRNS (32.3%), 746 pediatric SRNS (23.7%), 449 adult SSNS (14.3%), 695 pediatric SSNS (22.1%), and 240 cases with other or unclassified phenotypes (7.6%). *APOL1* genotyping identified 239 high-risk and 2,595 low-risk individuals.

Characteristics	N (3144)	%
Gender		
Female	1335	42.46
Male	1809	57.54
Age of onset		
Adult	1488	47.33
Pediatric	1483	47.17
Unknown	173	5.503
Inferred ancestry		
EUR (European)	2218	70.55
AFR (African)	365	11.61
AMR (Admixed American)	267	8.492
SAS (South Asian)	138	4.389
EAS (East Asian)	75	2.385
More than one ancestry	81	2.576
Steroid responsiveness		
Resistant	1876	59.67
Sensitive	1162	36.96
Unknown	106	3.372
Steroid responsiveness based on age of onset		
Adult Steroid Resistant Nephrotic Syndrome (aSRNS)	1014	32.25
Pediatric Steroid Resistant Nephrotic Syndrome (pSRNS)	746	23.73
Adult Steroid Sensitive Nephrotic Syndrome (aSSNS)	449	14.28

Pediatric Steroid Sensitive Nephrotic Syndrome (pSSNS)	695	22.11
Other	240	7.634
APOL1 Genotypes		
High risk	239	7.602
Low risk	2595	82.54
Variant not imputed	310	9.86

Table 14: Overall INS case characteristics across ages of disease onset, genetic ancestry and therapy response. Genetic ancestries were inferred using superpopulation labels from the 1000 genomes reference panel. Patients 21 years old or younger were categorized as pediatric, those with an unknown ages of disease onset (n = 173) and unknown response to steroids (n=106) were labeled as “Other” for downstream analysis. “High Risk” APOL1 persons were defined based on the G1/G1,G1/G2 and G2/G2 genotypes and “Low Risk” APOL1 were defined based on the G0/G0, G0/G1 and G0/G2 genotypes.

2.4.2. General CNV characteristics

PennCNV identified 30,498 high confidence CNVs across the cohort. Each individual carried a median of 8 CNVs with a median size of 63.7 kilobases (Kb). Of these, 14,328 were deletions and 16,170 duplications. Duplications tended to be larger (100-500Kb) whereas deletions were shorter (10-50Kb) (Figure 19). Most CNVs did not overlap coding regions. Moreover, large duplications were more likely to intersect multiple genes, with 30% spanning four or more genes compared to 15% of deletions (Figure 20).

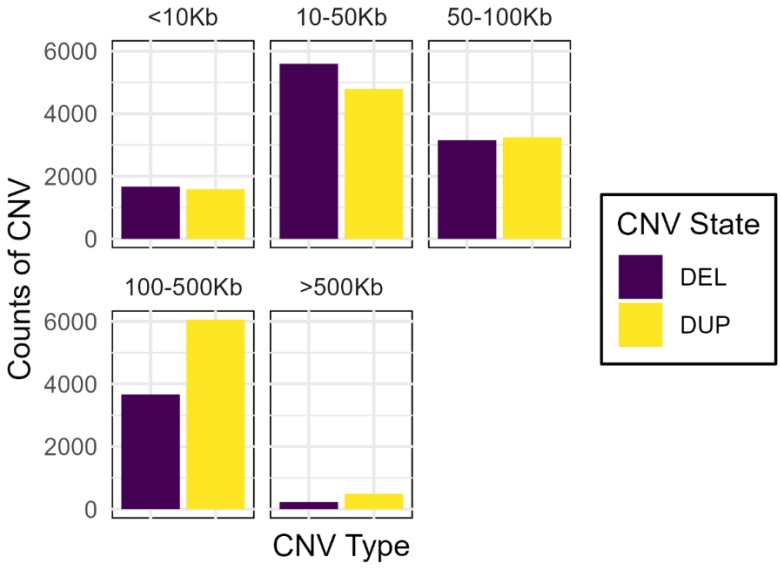


Figure 19: CNV size distribution stratified by CNV states. Deletions (DEL) included deletion of one or two copies of the region, whereas duplications (DUP) included the single or double copy duplications called by PennCNV.

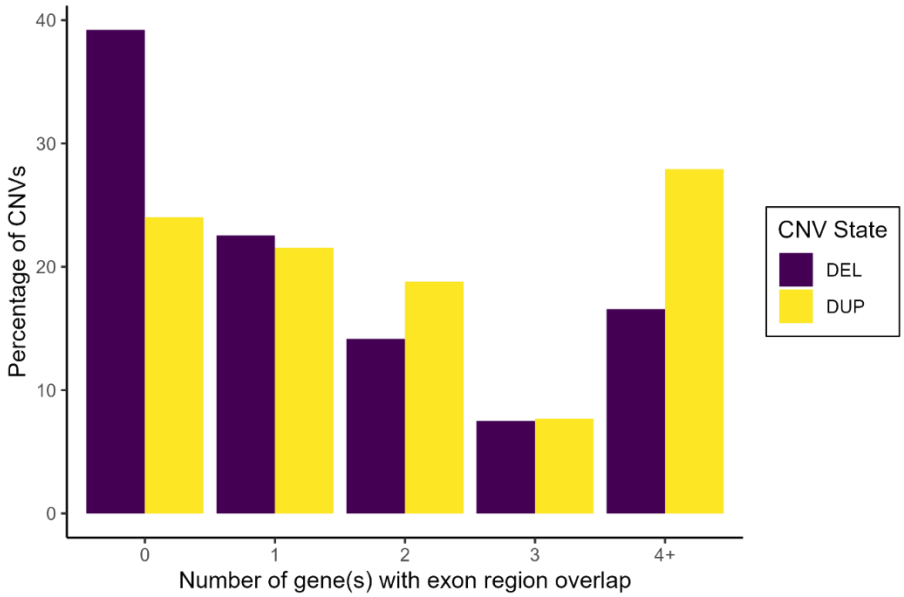
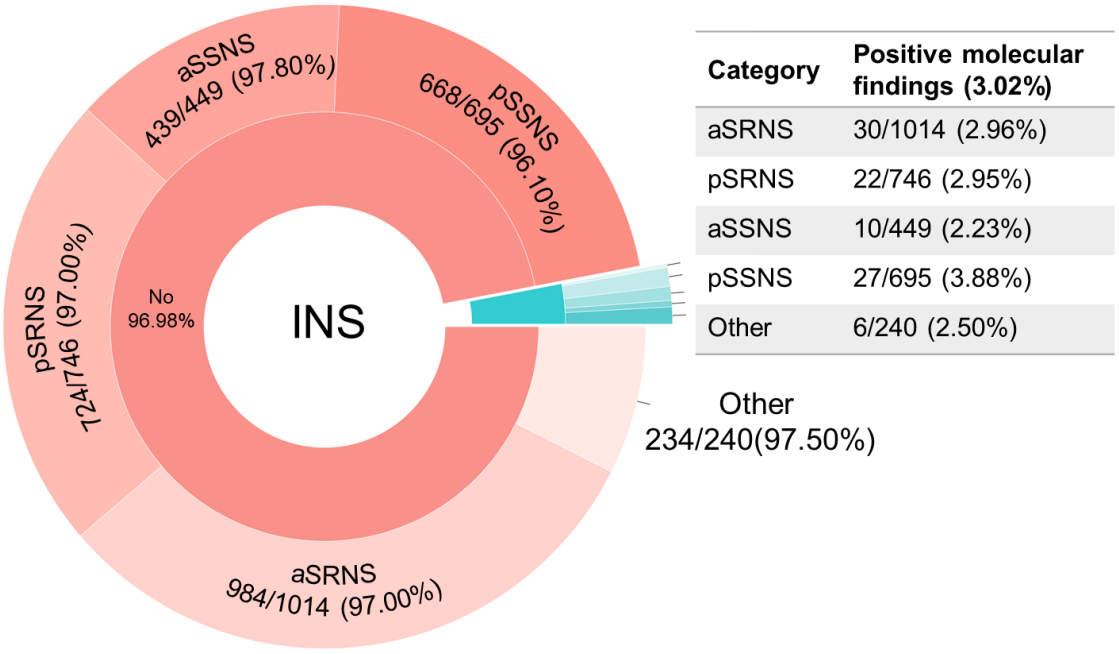


Figure 20: Percentage of CNVs overlapping an exon-region of gene(s). CNVs without gene overlaps (Deletion = 39.2% of 14,328; Duplication = 24% of 16,170) were either in intronic regions or did not have an annotated gene nearby.

2.4.3. Molecular findings of CNVs

Overall, positive molecular findings were identified in 3.02% INS individuals (95 of 3,144; Figure 21). These comprised CNVs intersecting known GD-CNVs (n=27) all classified pathogenic; CNVs overlapping genes implicated in podocytopathies or a phenocopy of if (n=22), of which 7 were classified likely pathogenic (LP) and 15 as VUS; and candidate CNVs that were rare, large and overlapped at least one RefGene exon (n=46), all classified as VUS. Pathogenic and LP CNVs together accounted for approximately 1% (34 of 3,144) of the INS patients. Surprisingly, pediatric patients who respond to steroids therapy (pSSNS) showed the highest rate of CNV-based positive molecular finding at 3.88% (*P*= 0.38, OR = 1.34, when compared to pSRNS). Within pSSNS, 10 individuals (1.44% of 695) carried GD-CNV, 5 (0.72%) harbored rare CNVs intersecting key INS/phenocopy genes, and 12 (1.73%) carried rare, large (>500Kb) CNVs. However, this

likely represents an overestimate in the pSSNS subgroup, since a significant proportion of SRNS are instead explained by rare Mendelian SNVs, thereby reducing the relative contributions of CNVs in resistant diseases.



Category	Positive molecular findings (3.02%)
aSRNS	30/1014 (2.96%)
pSRNS	22/746 (2.95%)
aSSNS	10/449 (2.23%)
pSSNS	27/695 (3.88%)
Other	6/240 (2.50%)

Figure 21: Distribution of molecular findings across the INS case group. The pie-donut chart depicted key percentages of diagnostic yield (or molecular findings) across INS disease subtypes: aSRNS (n=1,014), pSRNS (n=746), aSSNS (n=449) and pSSNS (n=695). Inner ring segments represent the percentages of CNV molecular findings in the overall INS case cohort (n=3,144), with positive (Teal) and negative CNV molecular findings (Pink) observed in 3.02% and 96.98% of INS, respectively. Outer segments denote the relative percentages of molecular findings across INS disease subtypes. Positive molecular findings included CNVs intersecting a known GD-CNV (n=27), CNVs intersecting genes that have either been reported to cause podocytopathies or a phenocopy of it (n=22) and CNVs which were rare, occurring in less than 0.02% of population controls, large, and intersected at least one exon of a RefGene (n= 46).

2.4.4. Diagnostic yield contributed by GD-CNV

GD-CNVs were detected in 27 of 3,144 INS cases (0.86%), a rate comparable to controls (0.68%). Most of the identified GD-CNVs likely reflect the background mutational burden of GD-CNVs with low penetrance and variable expressivity (such as the 16p13.11 duplication, or the Triple X duplication) and their direct involvement in the pathobiology of NS remains elusive (Table 15). On the other hand, we identified two events of clear diagnostic relevance: the chr17q12 deletion causing RCAD syndrome and the chr2q13 homozygous deletion in NPHP1, a known cause of nephronophthisis. These CNVs are associated to Mendelian forms of pediatric kidney structural defects that can phenocopy FSGS[292,293].

Across both SRNS and SSNS groups, GD-CNVs were more frequent in pediatric patients (1.44% pSSNS, 0.8% pSRNS) compared with adults (0.45% aSSNS, 0.69% aSRN), supporting the notion of a higher genetic load in children as compared to adults (Figure 22).

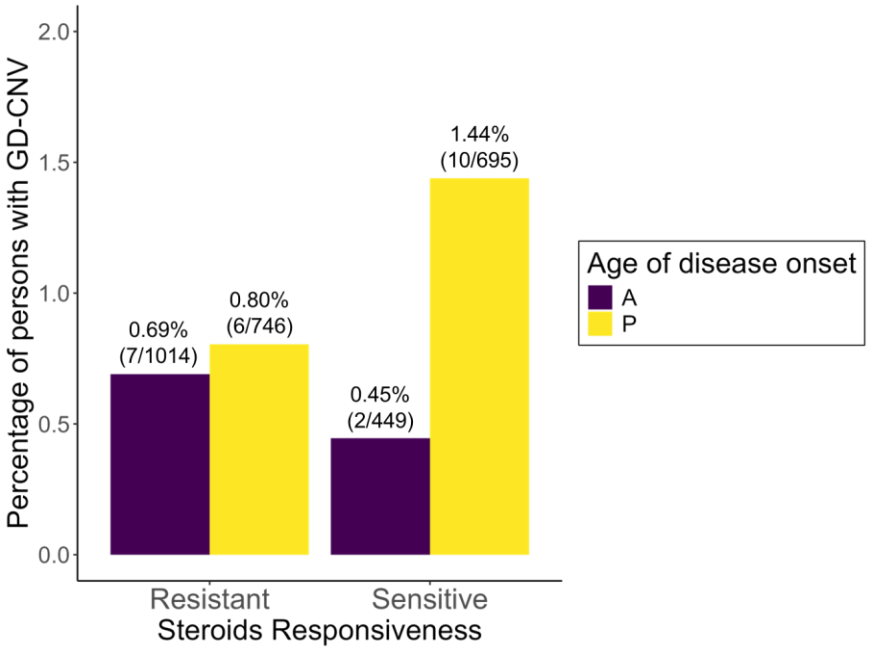


Figure 22: Distribution of persons with GD-CNV stratified by steroids responsiveness and age of disease onset. 2 persons carrying a GD-CNV have either an unknown age of disease onset or steroid responsiveness, data not shown. A: Adult, P: Pediatric, U: Unknown.

GD-CNV	Type	Cases (n=3,144)	Controls (n=21,498)	OR	Fisher *P*
1q21.1 susceptibility locus for Thrombocytopenia-Absent Radius (TAR) syndrome	Del	1	7	0.977	1.000
1q21.1 TAR Syndrome region duplication	Dup	3	8	2.565	0.157
1q21.1 recurrent microdeletion	Del	1	4	1.710	0.495
1q21.1 recurrent microduplication	Dup	1	6	1.140	1.000
1q43-1q44 deletion	Del	1	0	Inf	0.128
2q13 homozygous deletion Nephronophtisis 1	Del (homo)	1	0	Inf	0.128
3q29 microduplication syndrome Intellectual disability schizophrenia	Dup	1	1	6.838	0.239
Cri du Chat Syndrome	Del	1	0	Inf	0.128
7p interstitial duplication	Dup	1	1	6.838	0.239
Williams Beuren Syndrome	Del	1	0	Inf	0.128
8p23.1 duplication	Dup	1	0	Inf	0.128
9q34 deletion,9q subtelomeric deletion syndrome	Del	1	0	Inf	0.128
16p13.11 duplication,16p13.11 recurrent microduplication (neurocognitive disorder susceptibility locus)	Dup	3	29	0.707	0.791
16p13.11 recurrent microdeletion (neurocognitive disorder susceptibility locus)	Del	2	3	4.561	0.125
Charcot-Marie-Tooth syndrome type 1A (CMT1A)	Dup	1	8	0.855	1.000
Hereditary Liability to Pressure Palsies (HNPP)	Del	1	12	0.570	1.000
RCAD deletion	Del	2	0	Inf	0.016
Early-onset Alzheimer disease with cerebral amyloid angiopathy	Dup	1	0	Inf	0.128
TripleX	Dup	2	0	Inf	0.016
Steroid sulphatase deficiency (STS) cn0	Del(homo)	1	3	2.280	0.421

Table 15: List of GD-CNVs carried by the 3,144 INS cases and 21,498 controls. Del: deletion, Dup: duplication.

2.4.5. Diagnostic yield contributed by CNV in INS-associated genes

Rare CNVs intersecting known NS-associated genes were observed in 22 of 3,144 INS cases (0.69% of cases, and 0.49% of 21,498 controls) (Table 16). Notably, 12 of 22 INS cases carried heterozygous deletions encompassing *INF2*, with CNV sizes ranging from 30Kb to 120Kb. CNV carriers were observed across all INS subtypes with no significant associations with either ages of disease onset or therapy response. Deletions spanning *INF2* were rare in the 21,498 controls (n=86; 0.40%; OR= 0.95(0.47-1.76); not significant), although their frequencies varied across the different *INF2*-deletion CNV size range. These were also absent in both the 1000 genomes project and gnomAD. In contrast, the largest 120Kb genomic interval harbored 12 duplications in gnomAD CNVs v4.1.0 (chr14:104687768-104805499 in GRCh38; median size 97.09Kb) with site frequencies ranging from 2.15×10^{-6} to 1.08×10^{-5}. *INF2* (Inverted formin 2) is a well-established causal gene for autosomal dominant FSGS[294,295] where pathogenicity is typically driven by gain-of-function missense mutations clustering in the diaphanous inhibitory domain (DID) domain. A heterozygous deletion, representing potential haploinsufficiency is not a recognized disease mechanism for this gene, consistent with its moderate evidence of intolerance to loss-of-function variation (pLI = 0.33; haploinsufficiency score of 0.695). Current evidence does not support classifying these CNV as pathogenic by analogy with more common missense variants reported in literature[296]. Therefore, additional studies will be needed to establish or refute causality for these variants. Beyond *INF2* deletions, the next most frequent CNVs involved duplications spanning the *APOL1* gene. With the exception of one large 15.77Mb duplication in a European SSNS case (G0/G0), which covered 259 genes (among these and beyond *APOL1*, *MYH9* and *XPNPEP3* are known to be involved in Mendelian nephropathies), all 84Kb-94Kb APOL1 duplications were observed in 4 SRNS cases of African ancestry, aligning with the population-specific contribution of *APOL1* to kidney disease risk[283,297]. These duplications were observed in 3 cases with apparent G1 heterozygosity (G0/G1), whereas only one case carried the high risk G1/G2 genotype. Given that transgenic mouse models indicate that overexpression of *APOL1* risk variants (G1 and G2) increases risk of proteinuric kidney disease[298], it is conceivable that these *APOL1* duplication are linked to NS, even in the

heterozygous risk carriers. It is intriguing that recent data from West Africa point to a possible heterozygous effect for G1 and G2 in imparting risk for CKD and FSGS in individuals from Ghana and Nigeria[299], raising the possibility that these heterozygous duplications contribute to this signal

Furthermore, we identified a pSRNS case with a homozygous deletion in *NPHS1*. This genotype is classified as pathogenic, as *NPHS1* bi-allelic loss-of-function mutations are the cause of congenital nephrotic syndrome of the Finnish type. Lastly, likely pathogenic CNVs encompassing glomerulopathy-associated genes such as *COPA*, *CFH* and *ARHGAP24* were also observed in single cases.

ID	Position (Mb, Hg19)	Size (Mb)	CNV	Genes	Inh.	Cat.	Sex	ESKD	TX	P/LP	Anc	APOL1	pHaplo	pTriplo	Class
S01534	1:159.25-160.70	1.451	Dup	*COPA*	AD	aSRNS	F	N	N	*COL4A3*	EUR	G0/G0	0.91	0.99	LP
S00833	1:196.61-196.65	0.04	Del	*CFH*	AD/AR	aSRNS	F	Y	Y	U	EUR	G0/G0	0.68	0.37	VUS
S02699	4:86.55-86.85	0.302	Del	*ARHGAP24*	AD	pSRNS	F	U	U	U	EUR	G0/G0	0.10	0.17	VUS
S01814	12:0.19-133.78	132.58	Dup	*CEP83*, *LYZ*, *NUP107*, *PTPRO*, *TNFRSF1A*, *TNS2*	AR/AD	aSRNS	M	U	U	U	AFR	G1/G1	0.53	0.54	VUS
S01244	14:105.13-105.20	0.071	Del	*INF2*	AD	aSRNS	F	Y	Y	U	EUR	G0/G0	0.70	0.67	VUS
S00960	14:105.14-105.20	0.062	Del	*INF2*	AD	aSRNS	F	Y	Y	U	EUR	G0/G0	0.70	0.67	VUS
S02476	14:105.15-105.20	0.053	Del	*INF2*	AD	pSRNS	M	Y	Y	U	EUR	G0/G0	0.70	0.67	VUS
S02648	14:105.15-105.19	0.038	Del	*INF2*	AD	aSRNS	F	U	U	U	EUR	G0/G0	0.70	0.67	VUS
S00746	14:105.15-105.20	0.048	Del	*INF2*	AD	Other	F	U	U	U	EUR	G0/G0	0.70	0.67	VUS
S00852	14:105.15-105.20	0.048	Del	*INF2*	AD	pSSNS	F	U	U	U	EUR	G0/G0	0.70	0.67	VUS
S00880	14:105.15-105.20	0.048	Del	*INF2*	AD	pSRNS	M	U	U	U	EUR	G0/G0	0.70	0.67	VUS
S03052	14:105.15-105.27	0.118	Del	*INF2*	AD	pSRNS	M	U	U	U	EAS; EUR	G0/G0	0.70	0.67	VUS
S00741	14:105.16-105.20	0.045	Del	*INF2*	AD	pSSNS	M	U	U	U	EUR	G0/G0	0.70	0.67	VUS
S02483	14:105.16-105.20	0.045	Del	*INF2*	AD	aSRNS	M	Y	Y	U	EUR	G0/G0	0.70	0.67	VUS
S00773	14:105.16-105.20	0.038	Del	*INF2*	AD	pSSNS	F	U	U	U	EUR	G0/G0	0.70	0.67	VUS
S00788	14:105.17-105.20	0.033	Del	*INF2*	AD	pSSNS	M	U	U	U	EAS;EUR	G0/G0	0.70	0.67	VUS

S00735	19:36.30-36.32	0.025	Del (hom)	*NPHS1*	AR	pSRNS	M	U	U	U	EUR	G0/G0	0.52	0.29	LP
S00881	22:34.27-50.04	15.77	Dup	*APOL1*, *MYH9*, *XPNPEP3*	AR/AD	pSSNS	F	U	U	U	EUR	G0/G0	0.99	0.99	LP
S02575	22:36.62-36.71	0.095	Dup	*APOL1*, *MYH9*	AR/AD	aSRNS	F	N	N	U	AFR	G1/G0	0.99	0.99	LP
S01098	22:36.62-36.71	0.094	Dup	*APOL1*, *MYH9*	AR/AD	pSRNS	F	U	U	U	AFR	G1/G0	0.99	0.99	LP
S00008	22:36.62-36.72	0.095	Dup	*APOL1*, *MYH9*	AR/AD	pSRNS	M	Y	Y	U	AFR	G1/G2	0.99	0.99	LP
S00165	22:36.63-36.71	0.084	Dup	*APOL1*, *MYH9*	AR/AD	pSRNS	F	Y	Y	*COL4A5*	AFR	G1/G0	0.99	0.99	LP

Table 16: Distribution of rare likely-pathogenic CNVs intersecting known NS-associated/phenocopy genes regardless of CNV sizes. CNV types dup: duplication, del: deletion. Genes refer to the INS-associated or phenocopy genes in which a CNV covers. Gene mode of inheritance was derived from OMIM and literature review. P/LP indicates is a person carry a pathogenic or likely pathogenic SNV with genes shown. pHaplo and pTriplo referred to probabilities of dosage sensitivities[300]. Anc: 1KGP inferred ancestry; Cat.: INS categories; Class: CNV pathogenicity classification; ESKD: End stage kidney disease; F: Female; hom: homozygous; Inh.: Mode of inheritance; LP: Likely pathogenic; M: Male; Y: Yes; N: No; U: Unknown; Tx: transplant;.VUS: Variants of unknown significance. All CNVs are heterozygous unless otherwise indicated.

2.5. Discussion

Copy-number variation remains an important but insufficiently understood contributor to the genetic architecture of idiopathic nephrotic syndrome (INS), with prior work largely confined to pediatric and steroid-resistant populations. By extending CNV characterization across ages of onset and treatment responses in the largest INS cohort analyzed to date, this study presents a comprehensive evaluation of the CNV landscape in INS and broadens the scope of genetic insights beyond previously studied subgroups. We observed that approximately 3% of individuals carried pathogenic, likely pathogenic, and VUS CNVs, notably lower than in CAKUT, where CNVs account for 3-8% of etiologic burden[1,277]. No specific GD-CNV or likely pathogenic rare CNV was enriched across the four 4 INS subtypes, consistent with a model in which structural variation influence disease through variable expressivity rather than subtype-defining mechanisms[301,302].

Although the overall CNV diagnostic yield is modest (3% overall positive molecular findings, of which 1% comprised pathogenic or LP CNVs), several molecular findings may still hold potential clinical relevance. CNVs involving the RCAD region affecting *HNF1B* have established biological plausibility and warrant ongoing surveillance for maturity-onset diabetes of the young (MODY) type 5[303]. Similarly, CNVs affecting *NPHP1* underscore the importance of considering phenotypic overlap with ciliopathies[304]. During this analysis, two observations emerged that warrant closer consideration. One individual with INS carried the canonical CMT1A duplication at 17p11.2, a dosage-increasing alteration of *PMP22* that is classically associated with peripheral neuropathy rather than INS[305]. In addition, rare deletions spanning *INF2* were identified in 12 INS patients and classified as VUS. CNVs involving *INF2* are not well characterized, as isolated FSGS or the syndromic form of CMT-neuropathy/FSGS at this locus is usually driven by gain-of-function missense variants in the armadillo repeat domains rather than dosage perturbation[306]. These raised broader questions about how gene–phenotype relationships are interpreted, particularly for SVs known to exhibit incomplete penetrance and variable expressivity, and further underscore the challenges of assigning pathogenic

relevance based on established gene–phenotype correlation alone and highlight the need for more comprehensive, cross-system phenotyping[307].

Another locus of interest in this INS cohort involves structural variation spanning *APOL1*. The *APOL1* G1 and G2 risk variants remain the most compelling and reproducible genetic contributors to kidney disease among individuals with recent West African ancestry. In CKD and FSGS, carriers of two *APOL1* risk alleles confer 1.25 and 1.84 increased risk of progressive kidney dysfunction, respectively[299]. These associations are biologically plausible, given the reported gain-of-function cytotoxicity of the G1/G2 variants[308] and the strong selective pressures that have shaped the *APOL1* gene cluster in Africa. In light of this, structural variation at the *APOL1* locus has emerged as a potentially important, but poorly understood, layer of genetic complexity[283].

In this study, duplications of approximately 94-95 kb (chr22:36.62 to 36.71Mb) spanning *APOL1* and *MYH9* were identified in four individuals of African ancestry. All four carriers possessed at least one *APOL1* renal risk allele (three with G0/G1 and one with G1/G2 genotypes). In addition, we observed a separate non-African case with the non-risk G0/G0 genotype who carried a larger 15.5 Mb CNV encompassing *APOL1*, *MYH9*, and *XPNPEP3*. The copy number architecture surrounding *APOL1* is highly polymorphic, shaped by evolutionary pressures related to trypanosome resistance and by segmental duplications across the *APOL1-APOL4* gene cluster[309]. While G1 and G2 are coding variants, *APOL1* resides in a region enriched for structural rearrangement, including duplications that may modulate gene dosage or transcriptional regulation. A duplication that includes one of the canonical risk alleles could, in principle, increase the effective number of risk copies within an individual[308]. Even if present on only one haplotype, additional copies may enhance the net expression of the cytotoxic *APOL1* isoform, lowering the threshold required to trigger podocyte injury, analogous to a gene-dosage effect. Under this model, a duplication containing a risk allele could function as a “third hit”, compounding the risk architecture traditionally defined by G1/G2 genotype[283,302].

Interpretation of these CNV findings must also account for several methodological constraints. As a primary consideration, the CMA-based CNV detection method is inherently constrained by array design, probe density, and genomic coverage[310]. It is

restricted to identifying larger deletions and duplications (>10 kb) and cannot detect copy-neutral or complex structural variants such as inversions, translocations, or multi-allelic events[311]. As a result, the full spectrum of disease-relevant structural variation is not captured. Detection power also varies considerably across array platforms, leading to differences in the number, size, and resolution of CNV calls. Thus far, this analysis has been focused on CNVs that intersected known GD-CNV (n=27), INS-associated or phenocopy (n=22) and protein-coding gene (n=46). The functional relevance of CNVs that predominantly fall within non-coding regions remained unclear. More definitive interpretation will require an annotation framework[312,313] that incorporates topological associated domains (TAD), which partitions the 3D spatial genome into distinct regulatory compartments covering gene-enhancer interacting regions, and when disrupted led to alterations of transcriptional regulation[314–316]. Furthermore, CNV calling was initially performed with an older hg18 genome build which may limit breakpoint resolution. Although most contemporary analyses now rely on GRCh38, this assembly still contains missing sequences and residual assembly errors that are resolved only in the telomere-to-telomere (T2T-CHM13) reference. At present, however, few databases are built directly from T2T-CHM13, and many annotations continue to rely on liftovers from older genome versions. These inherited limitations including a reduced coverage in centromeric regions and the short arms of the acrocentric chromosomes 13, 14, 15, 21, and 22 may influence variant detection and interpretation[317,318]. Lastly, while LRR and BAF plots were used to inspect the CNVs, validation using orthogonal methods such as Multiplex ligation-dependent probe amplification (MLPA)[319] or quantitative PCR (qPCR)[320] were not available for confirmation.

Collectively, these findings indicate that CNVs contribute to INS in a minority of cases. Future studies integrating long-read sequencing, T2T or Pangenome-aligned SV discovery, and consent frameworks enabling recontact will be essential for fully resolving the contribution of structural variation to INS pathogenesis and for establishing the clinical relevance of rare CNVs identified in this population.

Chapter 3: Integrative discussion

3.1. Blood-based genomic and multi-omics approaches in bladder cancer and INS

The management of kidney (INS) and urological diseases (BCa) shares a unique space at the intersection of high disease burden, invasive diagnostic methods, and limited precision diagnostic and prognostic tools. Despite significant advances in genomics and biomarker discovery, clinicians still heavily rely on resource-intensive and invasive biopsy procedures for diagnosis and disease monitoring. This creates a two-part problem: patients undergoing the procedure face a compromise in the quality of life, while healthcare systems suffer from an escalating cost from disease surveillance and often inefficacious treatments. This highlights the need for non-invasive biomarkers that improve patient stratification and diminish the need for repeated invasive procedures.

With this rationale, this thesis addressed two central questions.

1) Can integrative multi-omics profiling of peripheral blood enable non-invasive detection of disease-relevant molecular programs in bladder cancer, despite the complexity and distance from the primary tumor? This question reflects the broader hypothesis that circulating molecular signals may capture systemic responses such as immune activation, inflammatory states, or genetically encoded regulatory variation that complement tumor-intrinsic biology and provide prognostic insight[321].
2) To what extent do structural variants contribute to the missing heritability of INS, a disease defined by marked phenotypic and genetic heterogeneity? INS exhibits clinical and genetic diversity, and structural variants, which perturb dosage-sensitive pathways contributing to phenotypic variation in many diseases, represent a plausible yet incompletely characterized source of its unexplained genetic risk. This thesis systematically characterizes CNVs across the INS spectrum, expanding existing study across treatment response, and age of onset.

To summarize, this thesis advances the study in BCa and INS through two complementary studies. The first demonstrates that blood-derived multi-omics profiles integrating genotyping, transcriptomics, and DNA methylation can reveal features relevant to BCa biology and prognosis, suggesting that peripheral blood carries meaningful systemic signals when analyzed in a unified framework. The second provides the largest systematic assessment of CNV in INS to date, identifying genomic disorder CNVs and rare structural variants that contribute modestly but clinically importantly to disease risk and genetic diagnosis. Together, these studies highlight the potential of blood-based genomic and multi-omics approaches to illuminate disease mechanisms and inform precision medicine across both urological cancers and renal disorders.

3.2. Lessons learned from Chapter 1: Blood-based multi-omics in Bladder Cancer

While this study establishes a framework for integrating inherited genotyping data with blood-derived transcriptomic and methylation profiles to infer disease-relevant biology in bladder cancer, several important limitations constrain its interpretability and potential for clinical translation.

Circulating blood is a complex mixture of immune and stromal cells, and only a very small fraction consists of circulating tumor material. To better understand the interplay between tumor-intrinsic and systemic immunity in BCa, a recent study showed that immune “hot” and “cold” tumors exhibited distinct immune cell compositions within the tumor microenvironment, but these differences were not reflected in peripheral blood. A follow-up analysis through consensus clustering revealed two distinct immune cell composition in both circulating blood and tumor[1]. Within the broader context of host systemic immunity, however, it remains impossible to disentangle which circulating signals originate from the tumor and which are host-driven, underscoring the challenges of interpreting blood-based immune signatures in bladder cancer[322].

Furthermore, the study cohort consisted entirely of European male participants, limiting the generalizability of the findings. The inclusion of women and ancestrally diverse

populations will be essential for determining whether the molecular programs identified here extend across different genetic backgrounds and clinical contexts. Moreover, due to the small sample size and the absence of a blood-based replication cohort, we were unable to assess the transferability of our findings.

Future studies will benefit from more diverse cohorts, longitudinal progression data, and comparative analyses across blood, urine, and tumor tissues. Such efforts will be essential for determining the true translational potential of blood-based omics in bladder cancer.

3.3. Lessons learned from Chapter 2: CNV landscape in INS

In INS, germline variants derived from blood play a more direct and causally interpretable role than in bladder cancer. CNVs representing large structural changes in the genomes have emerged as important contributors to rare kidney disease such as CAKUT[1,277]. Unlike monogenic disorders in which specific GD-CNV or pathogenic SNV define clear diagnostic categories, the CNV landscape within this INS population was heterogeneous with no enrichment in any disease subgroup.

Surprisingly, we observed similar rates of positive molecular findings in patients with pSRNS (3.88%) and pSSNS (2.95%), which contrast with existing literature in which rare, large-effect variants are reported predominantly in pediatric SRNS cohorts[282,283]. If substantiated, these findings suggest that earlier CNV studies, centered primarily on pediatric SRNS cases, may have captured only a narrow segment of the broader etiologic diversity present across INS.

One plausible explanation for this divergence from existing literature is the way steroid-response phenotypes were ascertained in our cohort. Steroid response was recorded only at the time of recruitment rather than through longitudinal follow-up, preventing the identification of individuals who may have initially responded to steroids but later progressed to resistance[323]. As a result, some patients categorized as steroid-sensitive at baseline may, in fact, represent evolving or mixed-response phenotypes, thereby blurring distinctions between SSNS and SRNS. Nevertheless, while this is a possibility,

INS cases that show early response and late resistance to immunosuppression have traditionally been considered as the classic primary/immune FSGS, in which Mendelian NS genes variants usually do not play a role.

In conclusion, our analysis revealed a modest, yet clinically-meaningful contribution of CNVs to INS, underscoring their potential causal role, particularly in GD-CNV and CNVs which has been predicted to span haploinsufficent or triplosensitive INS-associated/phenocopy genes, despite the current challenges in delineating their precise biological mechanisms.

3.4. Outlook

Future work in BCa will benefit from expanding beyond bulk blood-derived omics to incorporate higher-resolution molecular profiling technologies. Single-cell transcriptomics, immune profiling, and ctDNA analysis have the potential to reveal finer-grained interactions between tumor biology and systemic immune responses than bulk blood measurements can capture. Given the growing recognition that genetic determinants of disease susceptibility often differ from those shaping progression, dedicated progression-specific GWAS and predictive models will be essential for understanding recurrence and treatment response more precisely[324]. Integrating genomic data with longitudinal electronic health records may ultimately allow for individualized risk trajectories and more targeted surveillance strategies in bladder cancer[325].

For INS, the next major gains in understanding structural variation will come from improved genomic technologies and deeper phenotyping. Long-read sequencing and gapless T2T genome assemblies will substantially enhance the detection of complex structural variants that remain invisible to array-based or short-read platforms[311]. Harmonized longitudinal phenotyping will be critical for refining CNV genotype-phenotype correlation. Beyond SVs, the adoption of artificial intelligence (AI) and foundation models integrating kidney pathology, bulk and single-cell omics data will be key to accelerating the discovery of disease mechanisms[326,327].

Bibliography

1. Verbitsky, M. *et al.* The copy number variation landscape of congenital anomalies of the kidney and urinary tract. *Nat Genet* **51**, 117–127 (2019).

2. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229–263 (2024).

3. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).

4. Wang, P.-F., Song, H.-F., Zhang, Q. & Yan, C.-X. Pan-cancer immunogenomic analyses reveal sex disparity in the efficacy of cancer immunotherapy. *Eur. J. Cancer Oxf. Engl. 1990* **126**, 136–138 (2020).

5. Uhlig, A. *et al.* Gender-specific Differences in Recurrence of Non-muscle-invasive Bladder Cancer: A Systematic Review and Meta-analysis. *Eur. Urol. Focus* **4**, 924–936 (2018).

6. Kluth, L. A. *et al.* Female gender is associated with higher risk of disease recurrence in patients with primary T1 high-grade urothelial carcinoma of the bladder. *World J. Urol.* **31**, 1029–1036 (2013).

7. Burger, M. *et al.* Epidemiology and risk factors of urothelial bladder cancer. *Eur. Urol.* **63**, 234–241 (2013).

8. Freedman, N. D., Silverman, D. T., Hollenbeck, A. R., Schatzkin, A. & Abnet, C. C. Association Between Smoking and Risk of Bladder Cancer Among Men and Women (vol 306, pg 737, 2011). *Jama-J. Am. Med. Assoc.* **306**, 2220–2220 (2011).

9. Wilhelm-Benartzi, C. S. *et al.* Association of secondhand smoke exposures with DNA methylation in bladder carcinomas. *Cancer Causes Control* **22**, 1205–1213 (2011).

10. Tang, M.-S. *et al.* Electronic-cigarette smoke induces lung adenocarcinoma and bladder urothelial hyperplasia in mice. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21727–21731 (2019).

11. Ng, M. *et al.* Smoking prevalence and cigarette consumption in 187 countries, 1980-2012. *JAMA* **311**, 183–192 (2014).

12. Antoni, S. *et al.* Bladder Cancer Incidence and Mortality: A Global Overview and Recent Trends. *Eur. Urol.* **71**, 96–108 (2017).

13. Dai, X., Gakidou, E. & Lopez, A. D. Evolution of the global smoking epidemic over the past half century: strengthening the evidence base for policy action. *Tob Control* **31**, 129–137 (2022).

14. Flor, L. S., Reitsma, M. B., Gupta, V., Ng, M. & Gakidou, E. The effects of tobacco control policies on global smoking prevalence. *Nat Med* **27**, 239–243 (2021).

15. Safiri, S., Kolahi, A. A., Naghavi, M. & Global Burden of Disease Bladder Cancer, C. Global, regional and national burden of bladder cancer and its attributable risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease study 2019. *BMJ Glob Health* **6**, (2021).

16. Hashim, D. & Boffetta, P. Occupational and environmental exposures and cancers in developing countries. *Ann. Glob. Health* **80**, 393–411 (2014).

17. Zaghloul, M. S., Zaghloul, T. M., Bishr, M. K. & Baumann, B. C. Urinary schistosomiasis and the associated bladder cancer: update. *J. Egypt. Natl. Cancer Inst.* **32**, 44 (2020).

18. Gouda, I., Mokhtar, N., Bilal, D., El-Bolkainy, T. & El-Bolkainy, N. M. Bilharziasis and bladder cancer: a time trend analysis of 9843 patients. *J. Egypt. Natl. Cancer Inst.* **19**, 158–162 (2007).

19. Ishida, K. & Hsieh, M. H. Understanding Urogenital Schistosomiasis-Related Bladder Cancer: An Update. *Front. Med.* **5**, 223 (2018).

20. van der Post, R. S. *et al.* Risk of urothelial bladder cancer in Lynch syndrome is increased, in particular among MSH2 mutation carriers. *J. Med. Genet.* **47**, 464–470 (2010).

21. Fishwick, C. *et al.* Heterarchy of transcription factors driving basal and luminal cell phenotypes in human urothelium. *Cell Death Differ.* **24**, 809–818 (2017).

22. Magers, M. J. *et al.* Staging of bladder cancer. *Histopathology* **74**, 112–134 (2019).

23. Paner, G. P. *et al.* Updates in the Eighth Edition of the Tumor-Node-Metastasis Staging Classification for Urologic Cancers. *Eur. Urol.* **73**, 560–569 (2018).

24. Mostofi, F. K., Sobin, L. H., Torloni, H. & Organization, W. H. *Histological Typing of Urinary Bladder Tumours*. (World Health Organization, 1973).

25. Jones, T. D. & Cheng, L. Histologic Grading of Bladder Tumors: Using Both the 1973 and 2004/2016 World Health Organization Systems in Combination Provides Valuable Information for Establishing Prognostic Risk Groups. *Eur. Urol.* **79**, 489–491 (2021).

26. Bosschieter, J. *et al.* Reproducibility and Prognostic Performance of the 1973 and 2004 World Health Organization Classifications for Grade in Non-muscle-invasive Bladder Cancer: A Multicenter Study in 328 Bladder Tumors. *Clin. Genitourin. Cancer* **16**, e985–e992 (2018).

27. Sylvester, R. J. *et al.* European Association of Urology (EAU) Prognostic Factor Risk Groups for Non-muscle-invasive Bladder Cancer (NMIBC) Incorporating the WHO 2004/2016 and WHO 1973 Classification Systems for Grade: An Update from the EAU NMIBC Guidelines Panel. *Eur. Urol.* **79**, 480–488 (2021).

28. Moch, H. *et al.* The 2022 World Health Organization Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours. *Eur. Urol.* **82**, 458–468 (2022).

29. van Rhijn, B. W. G. *et al.* Prognostic Value of the WHO1973 and WHO2004/2016 Classification Systems for Grade in Primary Ta/T1 Non-muscle-invasive Bladder Cancer: A Multicenter European Association of Urology Non-muscle-invasive Bladder Cancer Guidelines Panel Study. *Eur. Urol. Oncol.* **4**, 182–191 (2021).

30. Gontero, P. *et al.* European Association of Urology Guidelines on Non-muscle-invasive Bladder Cancer (TaT1 and Carcinoma In Situ)-A Summary of the 2024 Guidelines Update. *Eur. Urol.* **86**, 531–549 (2024).

31. Catto, J. W. F. *et al.* Radical Cystectomy Against Intravesical BCG for High-Risk High-Grade Nonmuscle Invasive Bladder Cancer: Results From the Randomized Controlled BRAVO-Feasibility Study. *J. Clin. Oncol.* **39**, 202–214 (2021).

32. Jubber, I. *et al.* Epidemiology of Bladder Cancer in 2023: A Systematic Review of Risk Factors. *Eur. Urol.* **84**, 176–190 (2023).

33. Godlewski, D., Bartusik-Aebisher, D., Czech, S., Szpara, J. & Aebisher, D. Bladder cancer biomarkers. *Explor. Target. Anti-Tumor Ther.* **6**, 1002301 (2025).

34. Netto, G. J. *et al.* The 2022 World Health Organization Classification of Tumors of the Urinary System and Male Genital Organs-Part B: Prostate and Urinary Tract Tumors. *Eur. Urol.* **82**, 469–482 (2022).

35. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

36. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

37. Agarwal, N. *et al.* TRIM28 is a transcriptional activator of the mutant TERT promoter in human bladder cancer. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2102423118 (2021).

38. Borah, S. *et al.* Cancer. TERT promoter mutations and telomerase reactivation in urothelial cancer. *Science* **347**, 1006–1010 (2015).

39. Nickerson, M. L. *et al.* Molecular analysis of urothelial cancer cell lines for modeling tumor biology and drug response. *Oncogene* **36**, 35–46 (2017).

40. Nickerson, M. L. *et al.* Concurrent alterations in TERT, KDM6A, and the BRCA pathway in bladder cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **20**, 4935–4948 (2014).

41. Dyrskjøt, L. *et al.* Bladder cancer. *Nat. Rev. Dis. Primer* **9**, 58 (2023).

42. Hurst, C. D. *et al.* Genomic Subtypes of Non-invasive Bladder Cancer with Distinct Metabolic Profile and Female Gender Bias in KDM6A Mutation Frequency. *Cancer Cell* **32**, 701-715.e7 (2017).

43. Tran, L., Xiao, J.-F., Agarwal, N., Duex, J. E. & Theodorescu, D. Advances in bladder cancer biology and therapy. *Nat. Rev. Cancer* **21**, 104–121 (2021).

44. Rebouissou, S. *et al.* CDKN2A homozygous deletion is associated with muscle invasion in FGFR3-mutated urothelial bladder carcinoma. *J. Pathol.* **227**, 315–324 (2012).

45. Kamoun, A. *et al.* A Consensus Molecular Classification of Muscle-invasive Bladder Cancer. *Eur. Urol.* **77**, 420–433 (2020).

46. Khadhouri, S. *et al.* The IDENTIFY study: the investigation and detection of urological neoplasia in patients referred with suspected urinary tract cancer - a multicentre observational study. *BJU Int.* **128**, 440–450 (2021).

47. Ramirez, D. *et al.* Microscopic haematuria at time of diagnosis is associated with lower disease stage in patients with newly diagnosed bladder cancer. *BJU Int.* **117**, 783–786 (2016).

48. Karatzas, A. & Tzortzis, V. Lower urinary tract symptoms and bladder cancer in children: The hidden scenario. *Urol. Ann.* **11**, 102–104 (2019).

49. Rezaee, M. E., Dunaway, C. M., Baker, M. L., Penna, F. J. & Chavez, D. R. Urothelial cell carcinoma of the bladder in pediatric patients: a systematic review and data analysis of the world literature. *J. Pediatr. Urol.* **15**, 309–314 (2019).

50. Babjuk, M. *et al.* European Association of Urology Guidelines on Non-muscle-invasive Bladder Cancer (Ta, T1, and Carcinoma in Situ). *Eur. Urol.* **81**, 75–94 (2022).

51. Barkan, G. A. *et al.* The Paris System for Reporting Urinary Cytology: the quest to develop a standardized terminology. *J. Am. Soc. Cytopathol.* **5**, 177–188 (2016).

52. Nikas, I. P. *et al.* The Paris System for Reporting Urinary Cytology: A Meta-Analysis. *J. Pers. Med.* **12**, 170 (2022).

53. Luo, Y., She, D.-L., Xiong, H., Yang, L. & Fu, S.-J. Diagnostic Value of Liquid-Based Cytology in Urothelial Carcinoma Diagnosis: A Systematic Review and Meta-Analysis. *PloS One* **10**, e0134940 (2015).

54. Guo, A. *et al.* Bladder tumour antigen (BTA stat) test compared to the urine cytology in the diagnosis of bladder cancer: A meta-analysis. *Can. Urol. Assoc. J. J. Assoc. Urol. Can.* **8**, E347-352 (2014).

55. Zippe, C. *et al.* NMP22: a sensitive, cost-effective test in patients at risk for bladder cancer. *Anticancer Res.* **19**, 2621–2623 (1999).

56. Dimashkieh, H. *et al.* Evaluation of urovysion and cytology for bladder cancer detection: a study of 1835 paired urine samples with clinical and histologic correlation. *Cancer Cytopathol.* **121**, 591–597 (2013).

57. Piatti, P. *et al.* Clinical evaluation of Bladder CARE, a new epigenetic test for bladder cancer detection in urine samples. *Clin. Epigenetics* **13**, 84 (2021).

58. Barocas, D. A. *et al.* Updates to Microhematuria: AUA/SUFU Guideline (2025). *J. Urol.* **213**, 547–557 (2025).

59. Wallace, E. *et al.* Development of a 90-Minute Integrated Noninvasive Urinary Assay for Bladder Cancer Detection. *J. Urol.* **199**, 655–662 (2018).

60. Morales, A., Eidinger, D. & Bruce, A. W. Intracavitary Bacillus Calmette-Guerin in the treatment of superficial bladder tumors. *J. Urol.* **116**, 180–183 (1976).

61. Brausi, M. *et al.* Side effects of Bacillus Calmette-Guérin (BCG) in the treatment of intermediate- and high-risk Ta, T1 papillary carcinoma of the bladder: results of the EORTC genito-urinary cancers group randomised phase 3 study comparing one-third dose with full dose and 1 year with 3 years of maintenance BCG. *Eur. Urol.* **65**, 69–76 (2014).

62. Liu, Y., Lu, J., Huang, Y. & Ma, L. Clinical Spectrum of Complications Induced by Intravesical Immunotherapy of Bacillus Calmette-Guérin for Bladder Cancer. *J. Oncol.* **2019**, 6230409 (2019).

63. Boorjian, S. A. *et al.* Intravesical nadofaragene firadenovec gene therapy for BCG-unresponsive non-muscle-invasive bladder cancer: a single-arm, open-label, repeat-dose clinical trial. *Lancet Oncol.* **22**, 107–117 (2021).

64. Balar, A. V. *et al.* Pembrolizumab monotherapy for the treatment of high-risk non-muscle-invasive bladder cancer unresponsive to BCG (KEYNOTE-057): an open-label, single-arm, multicentre, phase 2 study. *Lancet Oncol.* **22**, 919–930 (2021).

65. Rödel, C. *et al.* Combined-modality treatment and selective organ preservation in invasive bladder cancer: long-term results. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **20**, 3061–3071 (2002).

66. Ploussard, G. *et al.* Critical analysis of bladder sparing with trimodal therapy in muscle-invasive bladder cancer: a systematic review. *Eur. Urol.* **66**, 120–137 (2014).

67. Rose, T. L. *et al.* Patterns of Bladder Preservation Therapy Utilization for Muscle-Invasive Bladder Cancer. *Bladder Cancer Amst. Neth.* **2**, 405–413 (2016).

68. Pieretti, A. *et al.* Complications and Outcomes of Salvage Cystectomy after Trimodality Therapy. *J. Urol.* **206**, 29–36 (2021).

69. International Collaboration of Trialists *et al.* International phase III trial assessing neoadjuvant cisplatin, methotrexate, and vinblastine chemotherapy for muscle-invasive bladder cancer: long-term results of the BA06 30894 trial. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **29**, 2171–2177 (2011).

70. Grossman, H. B. *et al.* Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer. *N. Engl. J. Med.* **349**, 859–866 (2003).

71. Powles, T. *et al.* Perioperative Durvalumab with Neoadjuvant Chemotherapy in Operable Bladder Cancer. *N. Engl. J. Med.* **391**, 1773–1786 (2024).

72. Iwasaki, K. *et al.* Neoadjuvant gemcitabine plus carboplatin for locally advanced bladder cancer. *Jpn. J. Clin. Oncol.* **43**, 193–199 (2013).

73. von der Maase, H. *et al.* Long-term survival results of a randomized trial comparing gemcitabine plus cisplatin, with methotrexate, vinblastine, doxorubicin, plus cisplatin in patients with bladder cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23**, 4602–4608 (2005).

74. Bellmunt, J. *et al.* Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. *N. Engl. J. Med.* **376**, 1015–1026 (2017).

75. Balar, A. V. *et al.* Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *Lancet Lond. Engl.* **389**, 67–76 (2017).

76. Sylvester, R. J. *et al.* 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**, 466–465; discussion 475-477 (2006).

77. Fernandez-Gomez, J. *et al.* Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model. *J. Urol.* **182**, 2195–2203 (2009).

78. Cambier, S. *et al.* EORTC Nomograms and Risk Groups for Predicting Recurrence, Progression, and Disease-specific and Overall Survival in Non-Muscle-invasive Stage Ta-T1 Urothelial Bladder Cancer Patients Treated with 1-3 Years of Maintenance Bacillus Calmette-Guérin. *Eur. Urol.* **69**, 60–69 (2016).

79. Holzbeierlein, J. M. *et al.* Diagnosis and Treatment of Non-Muscle Invasive Bladder Cancer: AUA/SUO Guideline: 2024 Amendment. *J. Urol.* **211**, 533–538 (2024).

80. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).

81. Fernandez-Gomez, J. *et al.* The EORTC tables overestimate the risk of recurrence and progression in patients with non-muscle-invasive bladder cancer treated with bacillus Calmette-Guérin: external validation of the EORTC risk tables. *Eur. Urol.* **60**, 423–430 (2011).

82. Krajewski, W. *et al.* Accuracy of the CUETO, EORTC 2016 and EAU 2021 scoring models and risk stratification tables to predict outcomes in high-grade non-muscle-invasive urothelial bladder cancer. *Urol. Oncol.* **40**, 491.e11-491.e19 (2022).

83. Vedder, M. M. *et al.* Risk prediction scores for recurrence and progression of non-muscle invasive bladder cancer: an international validation in primary tumours. *PloS One* **9**, e96849 (2014).

84. Altieri, V. M. *et al.* Recurrence and progression in non-muscle-invasive bladder cancer using EORTC risk tables. *Urol. Int.* **89**, 61–66 (2012).

85. Rieken, M. *et al.* Comparison of the EORTC tables and the EAU categories for risk stratification of patients with nonmuscle-invasive bladder cancer. *Urol. Oncol.* **36**, 8.e17-18.e24 (2018).

86. Olislagers, M. *et al.* Molecular biomarkers of progression in non-muscle-invasive bladder cancer - beyond conventional risk stratification. *Nat. Rev. Urol.* **22**, 75–91 (2025).

87. Figueroa, J. D. *et al.* Identification of a novel susceptibility locus at 13q34 and refinement of the 20p12.2 region as a multi-signal locus associated with bladder cancer risk in individuals of European ancestry. *Hum. Mol. Genet.* **25**, 1203–1214 (2016).

88. Figueroa, J. D. *et al.* Genome-wide association study identifies multiple loci associated with bladder cancer risk. *Hum. Mol. Genet.* **23**, 1387–1398 (2014).

89. García-Closas, M. *et al.* A single nucleotide polymorphism tags variation in the arylamine N-acetyltransferase 2 phenotype in populations of European background. *Pharmacogenet. Genomics* **21**, 231–236 (2011).

90. Garcia-Closas, M. *et al.* 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**, 4282–4289 (2011).

91. Rafnar, T. *et al.* Genome-wide association study yields variants at 20p12.2 that associate with urinary bladder cancer. *Hum Mol Genet* **23**, 5545–57 (2014).

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

93. Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat. Genet.* **42**, 978–984 (2010).

94. Kiemeney, L. A. *et al.* Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat. Genet.* **40**, 1307–1312 (2008).

95. Wu, X. *et al.* Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat. Genet.* **41**, 991–995 (2009).

96. Kiemeney, L. A. *et al.* A sequence variant at 4p16.3 confers susceptibility to urinary bladder cancer. *Nat. Genet.* **42**, 415–419 (2010).

97. Matsuda, K. *et al.* Genome-wide association study identified SNP on 15q24 associated with bladder cancer risk in Japanese population. *Hum. Mol. Genet.* **24**, 1177–1184 (2015).

98. Wang, M. *et al.* Genome-Wide Association Study of Bladder Cancer in a Chinese Cohort Reveals a New Susceptibility Locus at 5q12.3. *Cancer Res.* **76**, 3277–3284 (2016).

99. Ma, Z. *et al.* Systematic evaluation of bladder cancer risk-associated single-nucleotide polymorphisms in a Chinese population. *Mol. Carcinog.* **52**, 916–921 (2013).

100. Wu, J. *et al.* The Rare Variant rs35356162 in UHRF1BP1 Increases Bladder Cancer Risk in Han Chinese Population. *Front. Oncol.* **10**, 134 (2020).

101. Fu, Y.-P. *et al.* Common genetic variants in the PSCA gene influence gene expression and bladder cancer risk. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 4974–4979 (2012).

102. Kohaar, I. *et al.* Genetic variant as a selection marker for anti-prostate stem cell antigen immunotherapy of bladder cancer. *J. Natl. Cancer Inst.* **105**, 69–73 (2013).

103. Tang, W. *et al.* Mapping of the UGT1A locus identifies an uncommon coding variant that affects mRNA expression and protects from bladder cancer. *Hum. Mol. Genet.* **21**, 1918–1930 (2012).

104. Fu, Y.-P. *et al.* The 19q12 bladder cancer GWAS signal: association with cyclin E function and aggressive disease. *Cancer Res.* **74**, 5808–5818 (2014).

105. Koutros, S. *et al.* Differential urinary specific gravity as a molecular phenotype of the bladder cancer genetic association in the urea transporter gene, SLC14A1. *Int. J. Cancer* **133**, 3008–3013 (2013).

106. Middlebrooks, C. D. *et al.* Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.* **48**, 1330–1338 (2016).

107. Koutros, S. *et al.* Genome-wide Association Study of Bladder Cancer Reveals New Biological and Translational Insights. *Eur Urol* **84**, 127–137 (2023).

108. Sampson, J. N. *et al.* Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *J. Natl. Cancer Inst.* **107**, djv279 (2015).

109. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinforma. Biol. Insights* **14**, 1177932219899051 (2020).

110. Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).

111. Welham, Z., Déjean, S. & Lê Cao, K.-A. Multivariate Analysis with the R Package mixOmics. *Methods Mol. Biol. Clifton NJ* **2426**, 333–359 (2023).

112. Mo, Q. *et al.* A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostat. Oxf. Engl.* **19**, 71–86 (2018).

113. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, e8124 (2018).

114. Athieniti, E. & Spyrou, G. M. A guide to multi-omics data collection and integration for translational medicine. *Comput. Struct. Biotechnol. J.* **21**, 134–149 (2023).

115. Alcala, N. *et al.* Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. *Nat. Commun.* **10**, 3407 (2019).

116. Pekayvaz, K. *et al.* Multiomic analyses uncover immunological signatures in acute and chronic coronary syndromes. *Nat. Med.* **30**, 1696–1710 (2024).

117. Sharma, A. *et al.* Comprehensive multi-omics analysis of breast cancer reveals distinct long-term prognostic subtypes. *Oncogenesis* **13**, 22 (2024).

118. Omran, M. M. *et al.* Comparative analysis of statistical and deep learning-based multi-omics integration for breast cancer subtype classification. *J. Transl. Med.* **23**, 709 (2025).

119. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

120. Kiruba, B. *et al.* Intervention of machine learning in bladder cancer research using multi-omics datasets: systematic review on biomarker identification. *Discov. Oncol.* **16**, 1010 (2025).

121. Abdelaziz, E. H., Ismail, R., Mabrouk, M. S. & Amin, E. Multi-omics data integration and analysis pipeline for precision medicine: Systematic review. *Comput. Biol. Chem.* **113**, 108254 (2024).

122. Lindskrog, S. V. *et al.* An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. *Nat. Commun.* **12**, 2301 (2021).

123. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinforma. Oxf. Engl.* **26**, 1572–1573 (2010).

124. Hedegaard, J. *et al.* Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. *Cancer Cell* **30**, 27–42 (2016).

125. Zhang, X. *et al.* Robust Prognostic Subtyping of Muscle-Invasive Bladder Cancer Revealed by Deep Learning-Based Multi-Omics Data Integration. *Front. Oncol.* **11**, 689626 (2021).

126. Chu, G., Ji, X., Wang, Y. & Niu, H. Integrated multiomics analysis and machine learning refine molecular subtypes and prognosis for muscle-invasive urothelial cancer. *Mol. Ther. Nucleic Acids* **33**, 110–126 (2023).

127. Ye, Y., Zhang, Z., Liu, Y., Diao, L. & Han, L. A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. *Trends Genet. TIG* **36**, 318–336 (2020).

128. Goodison, S., Rosser, C. J. & Urquidi, V. Bladder Cancer Detection and Monitoring: Assessment of Urine- and Blood-Based Marker Tests. *Mol. Diagn. Ther.* **17**, 71–84 (2013).

129. Sacerdote, C. *et al.* Polymorphisms in the XRCC1 gene modify survival of bladder cancer patients treated with chemotherapy. *Int J Cancer* **133**, 2004–9 (2013).

130. Ricceri, F. *et al.* ERCC1 haplotypes modify bladder cancer risk: a case-control study. *DNA Repair* **9**, 191–200 (2010).

131. Turinetto, V. *et al.* H2AX phosphorylation level in peripheral blood mononuclear cells as an event-free survival predictor for bladder cancer. *Mol Carcinog* **55**, 1833–1842 (2016).

132. Allione, A. *et al.* MMP23B expression and protein levels in blood and urine are associated with bladder cancer. *Carcinogenesis* **39**, 1254–1263 (2018).

133. Russo, A. *et al.* Shorter leukocyte telomere length is independently associated with poor survival in patients with bladder cancer. *Cancer Epidemiol Biomark. Prev* **23**, 2439–46 (2014).

134. Palli, D. *et al.* A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. *Tumori* **89**, 586–93 (2003).

135. Guarrera, S. *et al.* Gene-specific DNA methylation profiles and LINE-1 hypomethylation are associated with myocardial infarction risk. *Clin. Epigenetics* **7**, 133 (2015).

136. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–83 (2016).

137. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).

138. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

139. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–1 (2010).

140. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590-8 (2006).

141. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinforma. Oxf. Engl.* **23**, 1294–1296 (2007).

142. Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* **53**, 420–425 (2021).

143. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**, (2019).

144. Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759–2772 (2020).

145. Morris, T. J. *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428–30 (2014).

146. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* **45**, e22 (2017).

147. Nordlund, J. *et al.* Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol* **14**, r105 (2013).

148. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. Oxf. Engl.* **8**, 118–127 (2007).

149. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

150. Jaffe, A. E. FlowSorted.Blood.450k. Bioconductor https://doi.org/10.18129/B9.BIOC.FLOWSORTED.BLOOD.450K (2017).

151. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

152. Larsson, S. C., Chen, J., Ruan, X., Li, X. & Yuan, S. Genome-wide association study and Mendelian randomization analyses reveal insights into bladder cancer etiology. *JNCI Cancer Spectr.* **9**, pkaf014 (2025).

153. Moore, C. M., Jacobson, S. A. & Fingerlin, T. E. Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Hum. Hered.* **84**, 256–271 (2019).

154. Vlachostergios, P. J. & Faltas, B. M. The molecular limitations of biomarker research in bladder cancer. *World J. Urol.* **37**, 837–848 (2019).

155. Ricceri, F. *et al.* ERCC1 haplotypes modify bladder cancer risk: a case-control study. *DNA Repair Amst* **9**, 191–200 (2010).

156. Pardini, B. *et al.* microRNA profiles in urine by next-generation sequencing can stratify bladder cancer subtypes. *Oncotarget* **9**, 20658–20669 (2018).

157. Vecchione, A. *et al.* FEZ1/LZTS1 is down-regulated in high-grade bladder cancer, and its restoration suppresses tumorigenicity in transitional cell carcinoma cells. *Am. J. Pathol.* **160**, 1345–1352 (2002).

158. Akand, M. *et al.* Deciphering the molecular heterogeneity of intermediate- and (very-)high-risk non–muscle-invasive bladder cancer using multi-layered –omics studies. *Front. Oncol.* **14**, (2024).

159. McAllister, S. S. & Weinberg, R. A. The tumour-induced systemic environment as a critical regulator of cancer progression and metastasis. *Nat Cell Biol* **16**, 717–27 (2014).

160. Simic, T., Savic-Radojevic, A., Pljesa-Ercegovac, M., Matic, M. & Mimic-Oka, J. Glutathione S-transferases in kidney and urinary bladder tumors. *Nat Rev Urol* **6**, 281–9 (2009).

161. Townsend, D. M. & Tew, K. D. The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* **22**, 7369–75 (2003).

162. Patel, K. R., Rais-Bahrami, S. & Basu, A. High sensitivity ctDNA assays in genitourinary malignancies: current evidence and future directions. *The Oncologist* **29**, 731–737 (2024).

163. Kopp, J. B. *et al.* Podocytopathies. *Nat. Rev. Dis. Primer* **6**, 68 (2020).

164. Londeree, J. *et al.* Estimation of childhood nephrotic syndrome incidence: data from the atlanta metropolitan statistical area and meta-analysis of worldwide cases. *J. Nephrol.* **35**, 575–583 (2022).

165. Chanchlani, R. & Parekh, R. S. Ethnic Differences in Childhood Nephrotic Syndrome. *Front. Pediatr.* **4**, 39 (2016).

166. Trautmann, A. *et al.* IPNA clinical practice recommendations for the diagnosis and management of children with steroid-sensitive nephrotic syndrome. *Pediatr. Nephrol. Berl. Ger.* **38**, 877–919 (2023).

167. Kodner, C. Diagnosis and Management of Nephrotic Syndrome in Adults. *Am. Fam. Physician* **93**, 479–485 (2016).

168. Sethi, S., Glassock, R. J. & Fervenza, F. C. Focal segmental glomerulosclerosis: towards a better understanding for the practicing nephrologist. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **30**, 375–384 (2015).

169. Rosenberg, A. Z. & Kopp, J. B. Focal Segmental Glomerulosclerosis. *Clin. J. Am. Soc. Nephrol. CJASN* **12**, 502–517 (2017).

170. Müller-Deile, J., Schenk, H. & Schiffer, M. [Minimal change disease and focal segmental glomerulosclerosis]. *Internist* **60**, 450–457 (2019).

171. Hull, R. P. & Goldsmith, D. J. A. Nephrotic syndrome in adults. *BMJ* **336**, 1185–1189 (2008).

172. Adedoyin, O. T. *et al.* Histopathologic Characteristics of Childhood Nephrotic Syndrome in a Tertiary Health Facility in Nigeria. *West Afr. J. Med.* **41**, 493–498 (2024).

173. Eddy, A. A. & Symons, J. M. Nephrotic syndrome in childhood. *Lancet Lond. Engl.* **362**, 629–639 (2003).

174. Vivarelli, M., Massella, L., Ruggiero, B. & Emma, F. Minimal Change Disease. *Clin. J. Am. Soc. Nephrol. CJASN* **12**, 332–345 (2017).

175. Bierzynska, A. *et al.* Genomic and clinical profiling of a national nephrotic syndrome cohort advocates a precision medicine approach to disease management. *Kidney Int.* **91**, 937–947 (2017).

176. Saleem, M. A. Molecular stratification of idiopathic nephrotic syndrome. *Nat. Rev. Nephrol.* **15**, 750–765 (2019).

177. Gattineni, J. Highlights for the management of a child with proteinuria and hematuria. *Int. J. Pediatr.* **2012**, 768142 (2012).

178. Leung, A. K. C., Wong, A. H. C. & Barg, S. S. N. Proteinuria in Children: Evaluation and Differential Diagnosis. *Am. Fam. Physician* **95**, 248–254 (2017).

179. Landini, S. *et al.* Reverse Phenotyping after Whole-Exome Sequencing in Steroid-Resistant Nephrotic Syndrome. *Clin. J. Am. Soc. Nephrol.* **15**, 89 (2020).

180. Wong, L., Huang, L. L., Nedeljkovic, M., Irish, A. & McMahon, L. P. Nephritis and Hearing Loss-Not All Roads Lead to Alport Syndrome. *Kidney Int. Rep.* **6**, 2922–2925 (2021).

181. Anvesh, G., Raju, S. B., Prasad, K., Sharma, A. & Surendra, M. Rare Association of Waardenburg Syndrome with Minimal Change Disease. *Indian J. Nephrol.* **28**, 226–228 (2018).

182. Nephrotic syndrome in children: prediction of histopathology from clinical and laboratory characteristics at time of diagnosis. A report of the International Study of Kidney Disease in Children. *Kidney Int.* **13**, 159–165 (1978).

183. Hama, T. *et al.* Renal biopsy criterion in children with asymptomatic constant isolated proteinuria. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **27**, 3186–3190 (2012).

184. Giglio, S. *et al.* Heterogeneous genetic alterations in sporadic nephrotic syndrome associate with resistance to immunosuppression. *J. Am. Soc. Nephrol. JASN* **26**, 230–236 (2015).

185. Sadowski, C. E. *et al.* A single-gene cause in 29.5% of cases of steroid-resistant nephrotic syndrome. *J. Am. Soc. Nephrol. JASN* **26**, 1279–1289 (2015).

186. Santín, S. *et al.* Clinical utility of genetic testing in children and adults with steroid-resistant nephrotic syndrome. *Clin. J. Am. Soc. Nephrol. CJASN* **6**, 1139–1148 (2011).

187. Gribouval, O. *et al.* Identification of genetic causes for sporadic steroid-resistant nephrotic syndrome in adults. *Kidney Int.* **94**, 1013–1022 (2018).

188. Zhao, J. & Liu, Z. Treatment of nephrotic syndrome: going beyond immunosuppressive therapy. *Pediatr. Nephrol. Berl. Ger.* **35**, 569–579 (2020).

189. Troyanov, S. *et al.* Focal and segmental glomerulosclerosis: definition and relevance of a partial remission. *J. Am. Soc. Nephrol. JASN* **16**, 1061–1068 (2005).

190. Radhakrishnan, J. & Cattran, D. C. The KDIGO practice guideline on glomerulonephritis: reading between the (guide)lines--application to the individual patient. *Kidney Int.* **82**, 840–856 (2012).

191. Tune, B. M. & Mendoza, S. A. Treatment of the idiopathic nephrotic syndrome: regimens and outcomes in children and adults. *J. Am. Soc. Nephrol. JASN* **8**, 824–832 (1997).

192. Trautmann, A. *et al.* Long-Term Outcome of Steroid-Resistant Nephrotic Syndrome in Children. *J. Am. Soc. Nephrol. JASN* **28**, 3055–3065 (2017).

193. Rémy, P. *et al.* An open-label randomized controlled trial of low-dose corticosteroid plus enteric-coated mycophenolate sodium versus standard corticosteroid treatment for minimal change nephrotic syndrome in adults (MSN Study). *Kidney Int.* **94**, 1217–1226 (2018).

194. Senthil Nayagam, L. *et al.* Mycophenolate mofetil or standard therapy for membranous nephropathy and focal segmental glomerulosclerosis: a pilot study. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **23**, 1926–1930 (2008).

195. Connaughton, D. M. *et al.* Monogenic causes of chronic kidney disease in adults. *Kidney Int.* **95**, 914–928 (2019).

196. Büscher, A. K. *et al.* Rapid Response to Cyclosporin A and Favorable Renal Outcome in Nongenetic Versus Genetic Steroid-Resistant Nephrotic Syndrome. *Clin. J. Am. Soc. Nephrol. CJASN* **11**, 245–253 (2016).

197. Warejko, J. K. *et al.* Whole Exome Sequencing of Patients with Steroid-Resistant Nephrotic Syndrome. *Clin. J. Am. Soc. Nephrol. CJASN* **13**, 53–62 (2018).

198. Gast, C. *et al.* Collagen (COL4A) mutations are the most frequent mutations underlying adult focal segmental glomerulosclerosis. *Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc.* **31**, 961–970 (2016).

199. Meliambro, K., He, J. C. & Campbell, K. N. Podocyte-targeted therapies - progress and future directions. *Nat. Rev. Nephrol.* **20**, 643–658 (2024).

200. Tryggvason, K., Ruotsalainen, V. & Wartiovaara, J. Discovery of the congenital nephrotic syndrome gene discloses the structure of the mysterious molecular sieve of the kidney. *Int. J. Dev. Biol.* **43**, 445–451 (1999).

201. Boute, N. *et al.* NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome. *Nat. Genet.* **24**, 349–354 (2000).

202. Arroyo-Parejo Drayer, P. *et al.* Spectrum of Clinical Manifestations in Children With WT1 Mutation: Case Series and Literature Review. *Front. Pediatr.* **10**, 847295 (2022).

203. Qiu, C. *et al.* Renal compartment-specific genetic variation analyses identify new pathways in chronic kidney disease. *Nat. Med.* **24**, 1721–1731 (2018).

204. Gbadegesin, R. A. *et al.* HLA-DQA1 and PLCG2 Are Candidate Risk Loci for Childhood-Onset Steroid-Sensitive Nephrotic Syndrome. *J. Am. Soc. Nephrol. JASN* **26**, 1701–1710 (2015).

205. Debiec, H. *et al.* Transethnic, Genome-Wide Analysis Reveals Immune-Related Risk Alleles and Phenotypic Correlates in Pediatric Steroid-Sensitive Nephrotic Syndrome. *J. Am. Soc. Nephrol. JASN* **29**, 2000–2013 (2018).

206. Dufek, S. *et al.* Genetic Identification of Two Novel Loci Associated with Steroid-Sensitive Nephrotic Syndrome. *J. Am. Soc. Nephrol. JASN* **30**, 1375–1384 (2019).

207. Jia, X. *et al.* Strong Association of the HLA-DR/DQ Locus with Childhood Steroid-Sensitive Nephrotic Syndrome in the Japanese Population. *J. Am. Soc. Nephrol. JASN* **29**, 2189–2199 (2018).

208. Downie, M. L. *et al.* Common Risk Variants in AHI1 Are Associated With Childhood Steroid Sensitive Nephrotic Syndrome. *Kidney Int. Rep.* **8**, 1562–1574 (2023).

209. Barry, A. *et al.* Multi-population genome-wide association study implicates immune and non-immune factors in pediatric steroid-sensitive nephrotic syndrome. *Nat Commun* **14**, 2481 (2023).

210. Shalhoub, R. J. Pathogenesis of lipoid nephrosis: a disorder of T-cell function. *Lancet Lond. Engl.* **2**, 556–560 (1974).

211. Iijima, K., Sako, M., Kamei, K. & Nozu, K. Rituximab in steroid-sensitive nephrotic syndrome: lessons from clinical trials. *Pediatr. Nephrol. Berl. Ger.* **33**, 1449–1455 (2018).

212. Schuster, C. *et al.* The Autoimmunity-Associated Gene CLEC16A Modulates Thymic Epithelial Cell Autophagy and Alters T Cell Selection. *Immunity* **42**, 942–952 (2015).

213. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).

214. Tzur, S. *et al.* Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the MYH9 gene. *Hum. Genet.* **128**, 345–350 (2010).

215. Friedman, D. J. & Pollak, M. R. APOL1 Nephropathy: From Genetics to Clinical Applications. *Clin. J. Am. Soc. Nephrol. CJASN* **16**, 294–303 (2021).

216. Kopp, J. B. *et al.* APOL1 Genetic Variants in Focal Segmental Glomerulosclerosis and HIV-Associated Nephropathy. *J. Am. Soc. Nephrol. JASN* **22**, 2129–2137 (2011).

217. Cooper, A. *et al.* APOL1 renal risk variants have contrasting resistance and susceptibility associations with African trypanosomiasis. *eLife* **6**, e25461 (2017).

218. Kamoto, K. *et al.* Association of APOL1 renal disease risk alleles with Trypanosoma brucei rhodesiense infection outcomes in the northern part of Malawi. *PLoS Negl. Trop. Dis.* **13**, e0007603 (2019).

219. Kaboré, J. W. *et al.* Candidate gene polymorphisms study between human African trypanosomiasis clinical phenotypes in Guinea. *PLoS Negl. Trop. Dis.* **11**, e0005833 (2017).

220. Abd ElHafeez, S. *et al.* Prevalence and burden of chronic kidney disease among the general population and high-risk groups in Africa: a systematic review. *BMJ Open* **8**, e015069 (2018).

221. Beckerman, P. *et al.* Transgenic expression of human APOL1 risk variants in podocytes induces kidney disease in mice. *Nat. Med.* **23**, 429–438 (2017).

222. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).

223. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).

224. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).

225. Gamazon, E. R. & Stranger, B. E. The impact of human copy number variation on gene expression. *Brief. Funct. Genomics* **14**, 352–357 (2015).

226. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).

227. Chiang, C. *et al.* Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.* **44**, 390–397, S1 (2012).

228. Itsara, A. *et al.* Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).

229. Mefford, H. C. & Eichler, E. E. Duplication hotspots, rare genomic disorders, and common disease. *Curr. Opin. Genet. Dev.* **19**, 196–204 (2009).

230. Rice, A. M. & McLysaght, A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat. Commun.* **8**, 14366 (2017).

231. Kearney, H. M. *et al.* American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **13**, 680–685 (2011).

232. Riggs, E. R. *et al.* Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med. Off. J. Am. Coll. Med. Genet.* **22**, 245–257 (2020).

233. Lim, T. Y., Verbitsky, M. & Sanna-Cherchi, S. ParseCNV2: a versatile and integrated tool for copy number variation association studies. *Eur. J. Hum. Genet.* **31**, 275–277 (2023).

234. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).

235. Sharp, A. J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).

236. de Vries, B. B. A. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).

237. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet. TIG* **18**, 74–82 (2002).

238. Goh, S., Thiyagarajan, L., Dudding-Byth, T., Pinese, M. & Kirk, E. P. A systematic review and pooled analysis of penetrance estimates of copy-number variants associated with neurodevelopment. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **27**, 101227 (2025).

239. Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).

240. Fellermann, K. *et al.* A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.* **79**, 439–448 (2006).

241. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).

242. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).

243. Sone, J. *et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat. Genet.* **51**, 1215–1221 (2019).

244. Mizuguchi, T. *et al.* A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J. Hum. Genet.* **64**, 359–368 (2019).

245. Sanchis-Juan, A. *et al.* Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.* **10**, 95 (2018).

246. Garg, S. *et al.* Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).

247. Logsdon, G. A. *et al.* Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430–441 (2025).

248. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).

249. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).

250. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

251. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

252. Porubsky, D. *et al.* Gaps and complex structurally variant loci in phased genome assemblies. *Genome Res.* **33**, 496–510 (2023).

253. Schloissnig, S. *et al.* Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442–452 (2025).

254. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

255. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinforma. Oxf. Engl.* **28**, i333–i339 (2012).

256. Söylev, A. *et al.* SVarp: pangenome-based structural variant discovery. 2024.02.18.580171 Preprint at https://doi.org/10.1101/2024.02.18.580171 (2024).

257. Porubsky, D. & Eichler, E. E. A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell* **187**, 1024–1037 (2024).

258. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).

259. Chau, M. H. K. *et al.* Detection of Clinically Relevant Monogenic Copy-Number Variants by a Comprehensive Genome-Wide Microarray with Exonic Coverage. *Clin. Chem.* **71**, 141–154 (2025).

260. Hahn, E. *et al.* Copy number variant analysis improves diagnostic yield in a diverse pediatric exome sequencing cohort. *NPJ Genomic Med.* **10**, 16 (2025).

261. Kim, Y. *et al.* High-resolution Chromosomal Microarray with Diagnostic Potential for Detecting Exon-level Copy Number Variations Using Targeted and Non-targeted Approaches. *Ann. Lab. Med.* https://doi.org/10.3343/alm.2025.0123 (2025) doi:10.3343/alm.2025.0123.

262. Li, Y. R. *et al.* Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nat. Commun.* **11**, 255 (2020).

263. Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).

264. Haraksingh, R. R., Abyzov, A. & Urban, A. E. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics* **18**, 321 (2017).

265. Peters, G. B. & Pertile, M. D. Chromosome microarrays in diagnostic testing: interpreting the genomic data. *Methods Mol. Biol. Clifton NJ* **1168**, 117–155 (2014).

266. Craig, J. E. *et al.* Rapid inexpensive genome-wide association using pooled whole blood. *Genome Res.* **19**, 2075–2080 (2009).

267. Vissers, L. E. L. M. *et al.* Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am. J. Hum. Genet.* **73**, 1261–1270 (2003).

268. Greenway, S. C. *et al.* De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat. Genet.* **41**, 931–935 (2009).

269. Wojcik, M. H. & Agrawal, P. B. Deciphering congenital anomalies for the next generation. *Cold Spring Harb. Mol. Case Stud.* **6**, a005504 (2020).

270. Wapner, R. J. *et al.* Chromosomal microarray versus karyotyping for prenatal diagnosis. *N. Engl. J. Med.* **367**, 2175–2184 (2012).

271. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).

272. Koolen, D. A. *et al.* Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).

273. Taylor, C. M. *et al.* 16p11.2 Recurrent Deletion. in *GeneReviews*® (eds Adam, M. P. et al.) (University of Washington, Seattle, Seattle (WA), 1993).

274. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

275. Westland, R. *et al.* Copy number variation analysis identifies novel CAKUT candidate genes in children with a solitary functioning kidney. *Kidney Int* **88**, 1402–1410 (2015).

276. Sanna-Cherchi, S. *et al.* Copy-number disorders are a common cause of congenital kidney malformations. *Am. J. Hum. Genet.* **91**, 987–997 (2012).

277. Wu, C.-H. W. *et al.* Copy Number Variation Analysis Facilitates Identification of Genetic Causation in Patients with Congenital Anomalies of the Kidney and Urinary Tract. *Eur. Urol. Open Sci.* **44**, 106–112 (2022).

278. Snoek, R. *et al.* NPHP1 (Nephrocystin-1) Gene Deletions Cause Adult-Onset ESRD. *J. Am. Soc. Nephrol. JASN* **29**, 1772–1779 (2018).

279. Claus, L. R. *et al.* The Importance of Copy Number Variant Analysis in Patients with Monogenic Kidney Disease. *Kidney Int. Rep.* **9**, 2695–2704 (2024).

280. Nakanishi, K. *et al.* Pair analysis and custom array CGH can detect a small copy number variation in COQ6 gene. *Clin. Exp. Nephrol.* **23**, 669–675 (2019).

281. Nagano, C. *et al.* Detection of copy number variations by pair analysis using next-generation sequencing data in inherited kidney diseases. *Clin. Exp. Nephrol.* **22**, 881–888 (2018).

282. Pantel, D. *et al.* Copy number variation analysis in 138 families with steroid-resistant nephrotic syndrome identifies causal homozygous deletions in PLCE1 and NPHS2 in two families. *Pediatr. Nephrol. Berl. Ger.* **39**, 455–461 (2024).

283. Ruchi, R. *et al.* Copy Number Variation at the APOL1 Locus. *PloS One* **10**, e0125410 (2015).

284. Peng, T., Li, G., Zhong, X. & Wang, L. Does copy number variation of APOL1 gene affect the susceptibility to focal segmental glomerulosclerosis? *Ren. Fail.* **39**, 500–504 (2017).

285. de Araújo Lima, L. & Wang, K. PennCNV in whole-genome sequencing data. *BMC Bioinformatics* **18**, 383 (2017).

286. Lin, C.-F., Naj, A. C. & Wang, L.-S. Analyzing Copy Number Variation using SNP Array Data: Protocols for Calling CNV and Association Tests. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **79**, Unit-1.27. (2013).

287. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinforma. Oxf. Engl.* **26**, 2867–2873 (2010).

288. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

289. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).

290. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).

291. Swaminathan, G. J. *et al.* DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum. Mol. Genet.* **21**, R37-44 (2012).

292. Braun, D. A. & Hildebrandt, F. Ciliopathies. *Cold Spring Harb. Perspect. Biol.* **9**, a028191 (2017).

293. Sambharia, M., Rastogi, P. & Thomas, C. P. Monogenic focal segmental glomerulosclerosis: A conceptual framework for identification and management of a heterogeneous disease. *Am. J. Med. Genet. C Semin. Med. Genet.* **190**, 377–398 (2022).

294. Boyer, O. *et al.* Mutations in INF2 Are a Major Cause of Autosomal Dominant Focal Segmental Glomerulosclerosis. *J. Am. Soc. Nephrol. JASN* **22**, 239–245 (2011).

295. Subramanian, B. *et al.* INF2 mutations cause kidney disease through a gain-of-function mechanism. *Sci. Adv.* **10**, eadr1017 (2024).

296. Labat-de-Hoz, L. & Alonso, M. A. The formin INF2 in disease: progress from 10 years of research. *Cell. Mol. Life Sci. CMLS* **77**, 4581–4600 (2020).

297. Daneshpajouhnejad, P., Kopp, J. B., Winkler, C. A. & Rosenberg, A. Z. The evolving story of apolipoprotein L1 nephropathy: the end of the beginning. *Nat. Rev. Nephrol.* **18**, 307–320 (2022).

298. McCarthy, G. M. *et al.* Recessive, gain-of-function toxicity in an APOL1 BAC transgenic mouse model mirrors human APOL1 kidney disease. *Dis. Model. Mech.* **14**, dmm048952 (2021).

299. Gbadegesin, R. A. *et al.* APOL1 Bi- and Monoallelic Variants and Chronic Kidney Disease in West Africans. *N. Engl. J. Med.* **392**, 228–238 (2025).

300. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041-3055.e25 (2022).

301. Girirajan, S. & Eichler, E. E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum. Mol. Genet.* **19**, R176-187 (2010).

302. Jensen, M. *et al.* Genetic modifiers and ascertainment drive variable expressivity of complex disorders. *Cell* S0092-8674(25)01080–3 (2025) doi:10.1016/j.cell.2025.09.012.

303. Faguer, S. *et al.* Diagnosis, management, and prognosis of HNF1B nephropathy in adulthood. *Kidney Int.* **80**, 768–776 (2011).

304. Lindstrand, A. *et al.* Recurrent CNVs and SNVs at the NPHP1 locus contribute pathogenic alleles to Bardet-Biedl syndrome. *Am. J. Hum. Genet.* **94**, 745–754 (2014).

305. Chance, P. F. *et al.* Trisomy 17p associated with Charcot-Marie-Tooth neuropathy type 1A phenotype: evidence for gene dosage as a mechanism in CMT1A. *Neurology* **42**, 2295–2299 (1992).

306. Boyer, O. *et al.* INF2 mutations in Charcot-Marie-Tooth disease with glomerulopathy. *N. Engl. J. Med.* **365**, 2377–2388 (2011).

307. De Rechter, S., De Waele, L., Levtchenko, E. & Mekahli, D. Charcot-Marie-Tooth: are you testing for proteinuria? *Eur. J. Paediatr. Neurol. EJPN Off. J. Eur. Paediatr. Neurol. Soc.* **19**, 1–5 (2015).

308. Datta, S. *et al.* Kidney Disease-Associated APOL1 Variants Have Dose-Dependent, Dominant Toxic Gain-of-Function. *J. Am. Soc. Nephrol. JASN* **31**, 2083–2096 (2020).

309. Monajemi, H., Fontijn, R. D., Pannekoek, H. & Horrevoets, A. J. G. The apolipoprotein L gene cluster has emerged recently in evolution and is expressed in human vascular tissue. *Genomics* **79**, 539–546 (2002).

310. Haraksingh, R. R., Abyzov, A. & Urban, A. E. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics* **18**, 321 (2017).

311. Schmitz, D., Ameur, A. & Johansson, Å. T2T-CHM13 improves read mapping and detection of clinically relevant genetic variation in the Swedish population. *Genome Res.* **35**, 2377–2388 (2025).

312. Spector, J. D. & Wiita, A. P. ClinTAD: a tool for copy number variant interpretation in the context of topologically associated domains. *J. Hum. Genet.* **64**, 437–443 (2019).

313. Hertzberg, J., Mundlos, S., Vingron, M. & Gallone, G. TADA-a machine learning tool for functional annotation-based prioritisation of pathogenic CNVs. *Genome Biol.* **23**, 67 (2022).

314. Ibn-Salem, J. *et al.* Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol.* **15**, 423 (2014).

315. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

316. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

317. Schmitz, D., Ameur, A. & Johansson, Å. T2T-CHM13 improves read mapping and detection of clinically relevant genetic variation in the Swedish population. *Genome Res.* **35**, 2377–2388 (2025).

318. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

319. Shen, Y. & Wu, B.-L. Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome. *J. Genet. Genomics Yi Chuan Xue Bao* **36**, 257–265 (2009).

320. Ma, L. & Chung, W. K. Quantitative analysis of copy number variants based on real-time LightCycler PCR. *Curr. Protoc. Hum. Genet.* **80**, 7.21.1-7.21.8 (2014).

321. Hiam-Galvez, K. J., Allen, B. M. & Spitzer, M. H. Systemic immunity in cancer. *Nat. Rev. Cancer* **21**, 345–359 (2021).

322. Chen, J.-Q. *et al.* Matched analysis of detailed peripheral blood and tumor immune microenvironment profiles in bladder cancer. *Epigenomics* **16**, 41–56 (2024).

323. Kim, J. S. *et al.* High incidence of initial and late steroid resistance in childhood nephrotic syndrome. *Kidney Int.* **68**, 1275–1281 (2005).

324. Yang, Z. *et al.* Limited overlap between genetic effects on disease susceptibility and disease survival. *Nat. Genet.* **57**, 2418–2426 (2025).

325. Wang, C. *et al.* Integrating electronic health records and GWAS summary statistics to predict the progression of autoimmune diseases from preclinical stages. *Nat. Commun.* **16**, 180 (2025).

326. Nephrobase Cell+: Multimodal Single-Cell Foundation Model for Decoding Kidney Biology | bioRxiv. https://www.biorxiv.org/content/10.1101/2025.09.30.679471v1.

327. Yan, F. *et al.* PathOrchestra: a comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks. *NPJ Digit. Med.* **8**, 695 (2025).

Manuscripts

- Tze Y. Lim et al. “Insights from a multi-omics study on bladder cancer”. Manuscript submitted.
- Tze Y. Lim et al. “Small but Clinically-Relevant Contribution of CNVs to Idiopathic Nephrotic Syndrome”. Manuscript in preparation.

Co-authorships out with the scope of this PhD:-

Wooden, Benjamin et al. “Natural History and Clinicopathological Associations of TRPC6-Associated Podocytopathy.” Journal of the American Society of Nephrology : JASN vol. 36,2 (2025): 274-289. doi:10.1681/ASN.0000000501

Milo Rasouly, Hila et al. “Exome analysis links kidney malformations to developmental disorders and reveals causal genes.” Nature communications vol. 16,1 7290. 7 Aug. 2025, doi:10.1038/s41467-025-62319-3

Batal, Ibrahim et al. “Association of Collapsing Glomerulopathy with Donor Apolipoprotein L1 Risk Variants in Kidney Allografts from Black Donors.” Clinical journal of the American Society of Nephrology : CJASN, 10.2215/CJN.0000000778. 7 Jul. 2025, doi:10.2215/CJN.0000000778

Riedhammer, Korbinian M et al. “Implication of transcription factor FOXD2 dysfunction in syndromic congenital anomalies of the kidney and urinary tract (CAKUT).” Kidney international vol. 105,4 (2024): 844-864. doi:10.1016/j.kint.2023.11.032

Pantel, Dalia et al. “Copy number variation analysis in 138 families with steroid-resistant nephrotic syndrome identifies causal homozygous deletions in PLCE1 and NPHS2 in two families.” Pediatric nephrology (Berlin, Germany) vol. 39,2 (2024): 455-461. doi:10.1007/s00467-023-06134-2

Khan, Atlas et al. “Mendelian Randomization Unveils Drug Targets for IgA Nephropathy.” Journal of the American Society of Nephrology : JASN vol. 35,8 (2024): 988-991. doi:10.1681/ASN.0000000000000434

Gupta, Y., Friedman, D.J., McNulty, M.T. et al. Strong protective effect of the APOL1 p.N264K variant against G2-associated focal segmental glomerulosclerosis and kidney disease. Nat Commun 14, 7836 (2023). https://doi.org/10.1038/s41467-023-43020-9

Martino, Jeremiah et al. “Mouse and human studies support DSTYK loss of function as a low-penetrance and variable expressivity risk factor for congenital urinary tract anomalies.” Genetics in medicine : official journal of the American College of Medical Genetics vol. 25,12 (2023): 100983. doi:10.1016/j.gim.2023.100983

Gehin C, Lone MA, Lee W, et al. CERT1 mutations perturb human development by disrupting sphingolipid homeostasis. *J Clin Invest*. 2023;133(10):e165019. Published 2023 May 15. doi:10.1172/JCI165019

Barry A, McNulty MT, Jia X, et al. Multi-population genome-wide association study implicates immune and non-immune factors in pediatric steroid-sensitive nephrotic syndrome. *Nat Commun*. 2023;14(1):2481. Published 2023 Apr 29. doi:10.1038/s41467-023-37985-w

Ahram DF, **Lim TY**, Ke J, et al. Rare Single Nucleotide and Copy Number Variants and the Etiology of Congenital Obstructive Uropathy: Implications for Genetic Diagnosis. *J Am Soc Nephrol*. 2023;34(6):1105-1119. doi:10.1681/ASN.0000000000000132