Aus dem Institut für Medizinische Biometrie
Universitätsklinik Heidelberg
Geschäftsführender Direktor: Prof. Dr. sc. hum. Meinhard Kieser

Identification and validation of circulating small non-coding RNAs associated with gallbladder cancer risk

Inauguraldissertation
zur Erlangung des
„Doctor scientiarum humanarum“

an der Medizinischen Fakultät Heidelberg
der Ruprecht-Karls-Universität

vorgelegt von
Alice Blandino
aus
Palermo, Italien

2024

Prodekan:	Herr Prof. Dr. Michael Boutros
Doktorvater:	Herr apl. Prof. Dr. Justo Lorenzo Bermejo

Contents

Abbreviations and symbols

β	Regression coefficient
ϱ	Spearman's rho coefficient
AIC	Akaike's information criterion
AJCC	American Joint Committee on Cancer
AS RNA	Antisense ribonucleic acid
ASR	Age-standardized rate
AUC	Area under the curve
BMI	Body-mass index
CANDELA	Consortium for the Analysis of the Diversity and Evolution of Latin America
cDNA	Complementary deoxyribonucleic acid
CI	Confidence interval
COPD	Chronic obstructive pulmonary disease
CRP	C-reactive protein
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
e.g.	For example (*exempli gratia*)
eQTL	Expression quantitative trait loci
eRNA	Enhancer ribonucleic acid
ESTHER	Early Detection and Optimised Therapy of Chronic Diseases in the Elderly Population

FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
FGFR	Fibroblast growth factor
GBC	Gallbladder cancer
GSA	Global screening array
GWAS	Genome-wide association study
HER2	Human epidermal growth factor receptor 2
HNR	Heinz Nixdorf Recall study
HUNT	Helseundersøkelsen i Nord-Trøndelag Health study
IBD	Identity by descent
i.e.	That is (*id est*)
IV	Instrumental variable
J-T test	Jonckheere-Terpstra test
LD	Linkage disequilibrium
log	Logarithm
Logit	Logit function
lincRNA	Long intervening ribonucleic acid
lncRNA	Long non-coding ribonucleic acid
MAD	Median absolute deviation
MAF	Minor allele frequency
MD	Mahalanobis distance
miRNA	Micro ribonucleic acid
ML	Machine learning
MR	Mendelian randomization
mRNA	Messenger ribonucleic acid
N	Total number of individuals
ncRNA	Non-coding ribonucleic acid

NGS	next-generation sequencing
OR	Odds ratio
PAT	Promoter-associated transcript
PCA	Principal component analysis
PCs	Principal components
P-value	Probability value
piRNA	Piwi-interacting ribonucleic acid
pri-miRNA	Primary micro ribonucleic acid
PRS	Polygenic risk score
RAIC	Robust Akaike's information criterion
RLC	RISC loading complex
RNA	Ribonucleic acid
RPKM	Reads per kilobase per million
rRNAs	Ribosomal ribonucleic acid
RR	Relative risk
siRNA	Small interfering ribonucleic acid
SIPA1L2	Signal induced proliferation associated 1 like 2 gene
SNP	Single nucleotide polymorphism
sncRNA	Small non-coding ribonucleic acid
snRNA	Small nuclear ribonucleic acid
snoRNA	Small nucleoar ribonucleic acid
tRNA	Transfer ribonucleic acid
x	Predictor variable
XGBoost	Extreme gradient boosting
y	Response variable

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Gallbladder cancer

1.1.1 Epidemiology

Gallbladder cancer (*GBC*; International Classification of Diseases, 10th Revision, diagnosis code C23) is an aggressive malignancy that accounts for approximately 89,000 deaths worldwide each year (Bray et al., 2024). This figure is projected to increase to 74% by 2045. In 2022, GLOBOCAN estimated that *GBC* ranks as the 22nd most frequent cancer globally. Despite a slight decline in incidence and mortality rates in recent years, the survival rate for this malignancy remains alarmingly low. This is largely attributable to late-stage diagnosis and limited treatment options, highlighting the critical need for improved prevention, early detection methods and also novel therapeutic strategies.

1.1.1.1 Incidence

According to the GLOBOCAN data projection, about 215,000 new cases of *GBC* are expected worldwide in 2045, a 57% increase over the registered cases in 2022 (Bray et al., 2024). Low- and middle-income countries are the most affected by *GBC*, with 83% of cases occurring in Asia and Latin America. In contrast, only 11% of cases are diagnosed in Europe and North America, where *GBC* is relatively uncommon. In particular, the highest age-standardized rates (*ASR*) of *GBC* per 100,000 person-years are observed in countries from Latin America

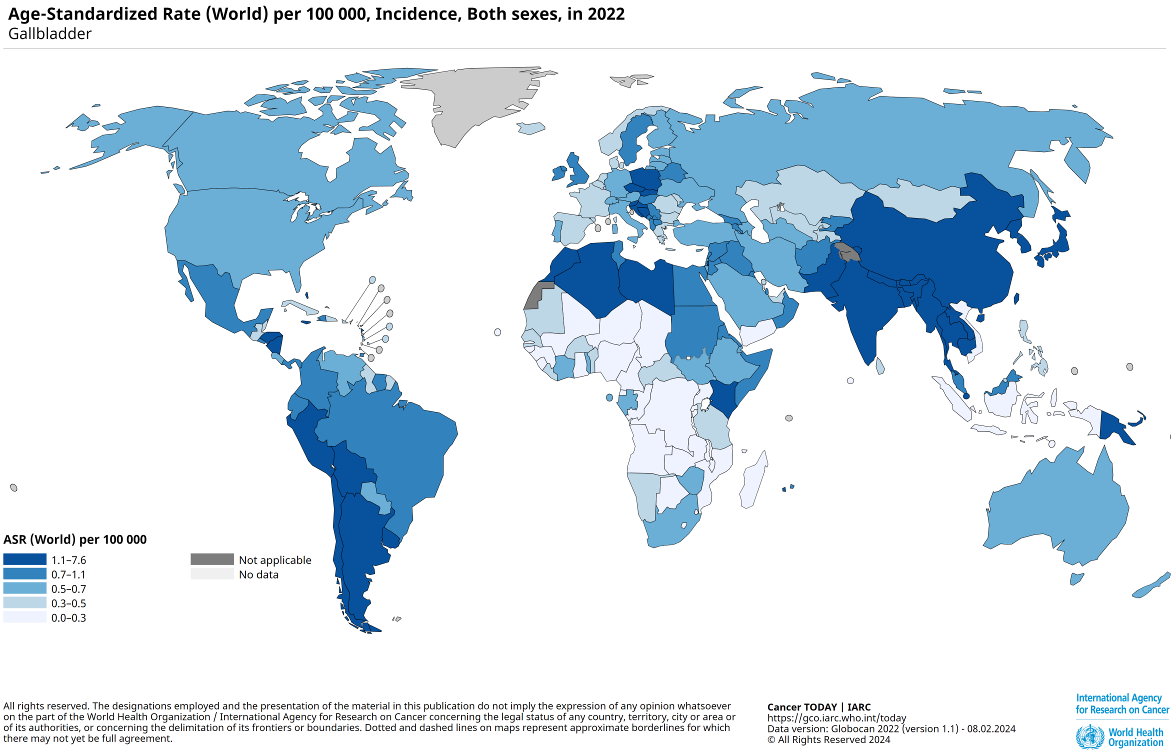
Age-Standardized Rate (World) per 100 000, Incidence, Both sexes, in 2022
Gallbladder
ASR (World) per 100 000
1.1–7.6
0.7–1.1
0.5–0.7
0.3–0.5
0.0–0.3
Not applicable
No data
All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.
Cancer TODAY | IARC
https://gco.iarc.who.int/today
Data version: Globocan 2022 (version 1.1) - 08.02.2024
© All Rights Reserved 2024
International Agency for Research on Cancer
World Health Organization

Figure 1.1: *Worldwide incidence of gallbladder cancer based on Bray et al. (2024). The top five countries with the highest incidence of gallbladder cancer are located in Latin America (Bolivia and Chile), East Asia (Bangladesh and Nepal), and North Africa (Algeria). Age-standardized incidence rate: countries in dark blue, very high incidences (ASR > 1.1/100,000 person-years). ASR: age-standardized rate.*

(Bolivia, ASR: 76/100,000 person-years; Chile, ASR: 57/100,000 person-years), East Asia (Bangladesh, ASR: 53/100,000 person-years; Nepal, ASR: 44/100,000 person-years), and North Africa (Algeria, ASR: 27/100,000 person-years). In Europe, the highest incidences occur in the Eastern countries, such as Croatia (ASR: 13/100,000 person-years), Albania (ASR: 12/100,000 person-years), and Bosnia Herzegovina (ASR: 12/100,000 person-years). As a result, GBC is rare in most parts of the world but poses a significant public health challenge in specific regions (Figure 1.1). As well as an unbalanced geographical distribution, GBC shows a marked sex difference, affecting women (ASR: 14/100,000 person-years) more often than men (ASR: 8.8/100,000 person-years) worldwide (Bray et al., 2024).

1.1.1.2 Prevention

Although cholecystectomy is strongly recommended following gallstone diagnosis especially in high-incidence regions, most patients are diagnosed too late, when surgery is not possible anymore (Kanthan et al., 2015). This late diagnosis often limits treatment options and adversely affects patients prognosis. Prevention, therefore, plays a crucial role. Cholecystectomy is recommended for individuals with symptomatic gallstones or GBC family history. An illustrative example of a GBC public health prevention policy is the initiative launched by the Chilean government in 2006, which integrated prophylactic cholecystectomy into the public health program for patients with gallstones aged 35 to 49 years as a strategy for GBC prevention (Koshiol et al., 2021). In 2010, the program was expanded to include asymptomatic women over 40 years of age who presented specific risk factors, including multiparity, and body-mass index (BMI) greater than 25, an educational level of 8 years or less, and at least one surname originating from the $Mapuche$ indigenous surname. In 2016, the program was further extended to encompass high-risk individuals, both women and men over 35 years old, based on the previously identified risk factors. Unfortunately, cholecystectomy remains an expensive and risky procedure, especially in older patients with comorbidities (Adamsen et al., 1997). GBC biomarkers, which can be used for population screening particularly in high incidence areas, could rectify some of the problems associated with cholecystectomies and promptly identify those individuals affected by early neoplastic lesions on the gallbladder. Non-coding RNA ($ncRNA$) research on GBC might help on this attempt.

1.1.1.3 Survival and mortality

In the early stages, when curative treatment is still possible, GBC is often asymptomatic or shows unspecific symptoms (Wistuba and Gazdar, 2004). Most GBC diagnoses occur at advanced stages, when the tumor has spread beyond the gallbladder, making curative surgery no longer possible (Kanthan et al., 2015). This, together with the very limited chemotherapy options, increases GBC mortality rates, particularly in countries where GBC incidence is high. Similarly to incidence, GBC death rates exhibit a clear both geographical and sex distribution. The five countries with the highest age-standardized mortality rates are Bolivia (mortality ASR: 6.3/100,000 person-years), Bangladesh (mortality ASR: 4.2/100,000 person-years), Chile (mortality ASR: 3.6/100,000 person-years), Nepal (mortality ASR:

3.3/100,000 person-years), and Republic of Korea (mortality *ASR*: 2.1/100,000 person-years) (Bray et al., 2024). Globally, *GBC* mortality rates are higher for women (global mortality *ASR*: 0.85/100,000 person-years), compared to men (global mortality *ASR*: 0.47/100,000 person-years). Most gallbladder tumors are diagnosed incidentally after routine cholecystectomy (Choi et al., 2015; Mantripragada et al., 2017). Due to the anatomical location of the gallbladder, the cancer rapidly spreads to nearby organs, including pancreas, liver, colon, and duodenum. According to the American Joint Committee on Cancer (*AJCC*), 8th edition, an overall 5-years survival rate from *GBC* of about 5-15% could be reached, if gallbladder resection is performed on time after cancer diagnosis (Madani et al., 2022). *GBC* survival rate strongly depends on the stage of the disease. Specifically, stage I *GBC* has a 5-year survival rate of approximately 50%, while stage IV *GBC* exhibits a markedly lower survival rate of only 3% (Roa et al., 2022).

1.1.1.4 Treatment

Particularly in patients with unresectable gallbladder, the implementation of molecular targeted therapies has provided greater hope and broader opportunities for the treatment of *GBC* (Zhou et al., 2023). The primary treatment options for *GBC* patients *gemcitabine* and platinum-based chemotherapies (Roa et al., 2022; Stein et al., 2015). In recent years, the advent of a new generation of sequencing technologies, has continuosly updated therapeutic strategies for *GBC*. Clinical application of targeted drugs include the *epidermal growth factor receptor* (*EGFR*), *fibroblast growth factor* (*FGFR*), and *human epidermal growth factor receptor* 2 (*HER*2) which have both been effectively employed as therapeutic targets in some clinical trials (Zhou et al., 2023). Thus, there is significant hope that differentiated therapy may enhance patients' survival, particularly in relation to specific molecular alterations that provide opportunities for new targeted therapeutics.

1.1.2 Risk factors

GBC is a multifactorial disease in which genetic variability, lifestyle, and environmental exposures contribute to an increased susceptibility to this malignancy. Some of the most common risk factors associated with *GBC* include female sex, advanced age, gallstone disease, high BMI, high Native American ancestry proportion, family history of GBC, and smoking (Kanthan et al., 2015).

1.1.2.1 Age, sex, and body mass index

Overall, the risk of GBC increases with age, with the median age reported in indexed literature being 67 years (Duffy et al., 2008). In Chile, GBC incidence rates increase from 1.3 per 100,000 for individuals aged 30 to 44 years, to 13 for those aged 45 to 59 years, 45.1 per 100,000 for ages 60 to 74 years, and 62.5 per 100,000 for those over 70 years (Bray et al., 2024). The mortality rates increase accordingly (Villanueva, 2016). Female sex is also considered a risk factor for GBC, with women accounting for three out of four diagnosed cases (Randi et al., 2009; Lai and Lau, 2008). One of the primary reasons for this difference may be women's greater exposure to estrogen (Randi et al., 2006). There is substantial evidence linking excess body weight to an increased risk of GBC (Campbell et al., 2017; Jackson et al., 2019; Li et al., 2016). When accounting for sex, the association appears to be significantly stronger for women: overweight women have a relative risk (RR) of 1.26 compared to obese women who present a RR of 1.67. In men, only obese subjects show a significantly higher GBC risk (RR: 1.42) (Tan et al., 2015). A study by Barahona Ponce et al. (2021) found a causal effect of BMI on GBC risk in Chileans through mendelian randomization (MR) analysis.

1.1.2.2 Gallstones

According to several case-control and cohort studies, gallstone disease is one of the most commonly reported risk factors for GBC across different populations (Lazcano-Ponce et al., 2001; Ryu et al., 2016; Villanueva, 2016). It is estimated that about 70-90% of GBC patients carry gallstones. Size, volume, and weight of the gallstones also seem to be correlated with the risk of developing GBC. A study conducted by Randi et al. (2006) found that individuals with a history of gallstone disease have a RR of 4.90 to develop GBC compared to those without gallstones evidence. The MR study from Barahona Ponce et al. (2021) also assessed that gallstone disease causally affects GBC risk in both Chileans and Europeans. This suggests that genetic and environmental factors play distinct roles in the pathogenesis of GBC. Similar to risk factors associated with GBC, gallstone biogenesis is influenced by both unmodifiable (such as female sex and increased age), and modifiable (such as high *BMI*) conditions (Di Ciaula et al., 2018).

1.1.2.3 Family history

According to Stinton and Shaffer (2012), the familial genetic background accounts for 25% of the total gallstone disease risk. As in most diseases, *GBC* familial risk may also be transmitted by intermediate conditions that act as cancer risk factors (Hemminki et al., 2022). For example, cholelithiasis, diabetes and obesity are strongly linked to family history and consequently increase the risk of *GBC*. Additionally, gene variations in *ABCB*1 and *ABCB*4 gene regions play a role in hepatobiliary phospholipid transporters, and have been recognized as possible risk factors for *GBC* (Mhatre et al., 2017). Low frequencies of the mismatch repair gene *MLH*1 have also been linked to biliary tract cancers, especially *GBC*.

1.1.2.4 Native American ancestry

As previously noted, *GBC* exhibits significant geographic variation, with particularly high prevalence in South American countries (Wistuba and Gazdar, 2004). Native Americans, including Chilean *Mapuche*, *Pima Indians*, and New Zealand *Maori* show higher rates of *GBC* incidence and gallstone prevalence (Hundal and Shaffer, 2014). Moreover, the higher prevalence of gallstones among Indigenous people, compared to the general population, suggests that variants associated with gallstone susceptibility may confer at least an indirect genetic predisposition to *GBC* in Native American populations (Carey and Paigen, 2002). *Mapuche*, the main indigenous people in Chile, show the highest ever reported *GBC* incidence and mortality, and are therefore the most studied subgroup in the context of *GBC*.

1.1.2.5 Cigarette smoking

A large Japanese prospective cohort study on biliary tract cancer found that current smokers exhibit a 1.35-fold increased cancer mortality risk compared to individuals who have never smoked (Lin et al., 2022). When differentiating *GBC* from other biliary tract cancers, the mortality risk associated with current smoking is even more marked, with a *RR* of 1.89. In men, the mortality risk associated with *GBC* shows a positive correlation with the number of cigarettes smoked daily. Another prospective cohort study on Korean adults revealed that current (*RR*: 1.12) and former (*RR*: 1.11) smokers are associated with an increased risk of *GBC* compared to non-smokers (Park et al., 2023). Stratified analyses based on the number of cigarette packages smoked per year showed that individuals who smoke between twenty and thirty packages annually have a 1.24 times higher risk of developing *GBC* compared to those who have never smoked. The highest risk was found among individuals who smoked more than twenty packages of cigarettes per year and also had diabetes, with a *RR* of 1.66 compared to non-smokers without diabetes.

1.1.3 Histopathology and pathogenesis

Microscopically, most gallbladder tumors (about 80-90%) are adenocarcinomas with cuboidal or columnar epithelial gland formation (Menon and Babiker, 2024). The remaining cases are mostly papillary, squamous cell, adenosquamous, undifferentiated, or small-cell carcinomas (Lai and Lau, 2008). *GBC* typically develops through a sequence of molecular and histological changes, starting with gallstone disease, then progressing to dysplasia, and ultimately leading to invasive cancer (Wistuba and Gazdar, 2004). Most of gallbladder carcinomas are associated with chronic inflammation by gallstone disease (chronic cholecystitis), while only a small proportion of *GBC* cases (less than 1%) result from changes in the bile due to the reflux of pancreatic juice into the common bile duct (Espinoza et al., 2016). This can be caused, as seen particularly in Asian countries, by an anomalous pancreaticobiliary ductal junction, or by polyps, which, if present for extended periods, lead to local inflammation (Bizama et al., 2015; Dutta, 2012; Kamisawa et al., 2017). The mutational profile of gallbladder adenocarcinoma predominantly features epigenetic mutations in $COX2$, *K-Ras*, $TP53$, $CDKN2a$, and *c-erb-b*2 (Nakamura et al., 2015). Furthermore, gene promoter hypermethylation has been progressively identified as a pathogenic contributor. The heterogeneity of genetic drivers further underscores the complex pathogenesis of *GBC* (Giraldo et al., 2022; Brägelmann et al., 2021).

1.2 Non-coding RNAs

The development and homeostasis of cells and tissues rely on gene expression and regulation, which are essential processes for all living organisms (Carthew, 2021). Francis Crick first conceptualized the relationship between genes and proteins (Crick, 1970). He stated, through his central dogma of molecular biology, that genetic information flows only in one direction: *deoxyribonucleic acid* (*DNA*) is transcribed into *messenger ribonucleic acid* (*mRNA*) and is translated into protein, or from *RNA* directly to protein. Consequently, for decades proteins were regarded as the primary functional products of genetic information, despite protein-coding genes represent less than 2% of the genome (Park et al., 2022). More recently, advances in sequencing technologies have led to the identification of other significant *RNAs* with no protein-coding prospect (Satam et al., 2023; Tripathi et al., 2017). At first, the role of this class of *RNAs* was not fully understood. However, an increasing number of *non-coding RNAs* (*ncRNAs*), which constitute nearly 60% of the transcriptional output in human cells, have demonstrated to have regulatory functions in multiple cellular biological pathways (Anastasiadou et al., 2018). By definition, *ncRNAs* are defined as an heterogeneous group of transcripts that are not translated into proteins (ENCODE Project Consortium, 2012). Since their discovery, the biological relevance of *ncRNAs* has increased more and more. Today, it is widely acknowledged that *ncRNAs* are not only simple intermediaries of protein synthesis towards *RNA*, but they play a crucial role as functional molecules in the regulation of gene expression and genome organization. Recent results from the GENCODE project show that the human genome is transcribed into more than 254,000 transcripts, of which only about 89,500 are protein coding (Frankish et al., 2019).

1.2.1 Classification

In recent decades, researchers have identified and extensively characterized many types of *ncRNAs*, according to their length, conformation and cellular function (Kaikkonen et al., 2011; Zhang et al., 2019). *ncRNAs* are mainly classified as either housekeeping or regulatory, depending on their role (Figure 1.2). Housekeeping *ncRNAs* are constitutively expressed in all cell types and serve as essential regulatory molecules in a variety of ribosomal and cellular activities. They include *ribosomal* (*rRNA*), *transfer* (*tRNA*), *small nuclear* (*snRNA*), and *small nucleolar RNAs* (*snoRNAs*). Regulatory *ncRNAs* are referred to as such because they are specifically engaged in regulatory processes. Based on their length, they consist of two main groups: *small* (*sncRNAs*, less than 200 nucleotides in length) and *long non-coding RNAs* (*lncRNAs*, more than 200 nucleotides in length). *sncRNAs* mainly comprise *microRNAs* (*miRNAs*), *small interfering RNAs* (*siRNAs*), and *piwi-*

interacting RNAs (*piRNAs*). The *lncRNAs* group includes *antisense RNAs* (*AS RNAs*) and *enhancer RNAs* (*eRNAs*). *eRNAs*, together with *promoter-associated transcripts* (*PATs*), and *circRNAs*, vary in length and can therefore be classified as both *sncRNAs* and *lncRNAs*. Most *ncRNAs* regulate the expression of nearby genes and are classified as cis-*ncRNAs* (López-Jiménez and Andrés-León, 2021; Elcheva and Spiegelman, 2020). Trans-acting *ncRNAs*, on the other hand, function at regions far from their transcription site, including the cytoplasm and other compartments of the cell.

This thesis particularly focuses on two types of *ncRNAs*: *lncRNAs* and *miRNAs*.

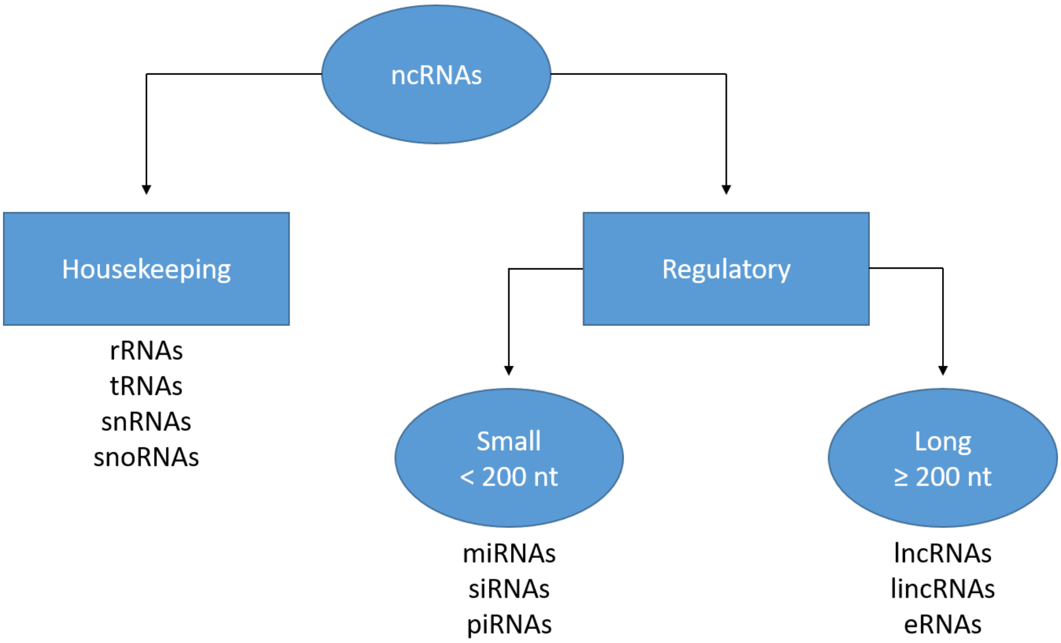

Figure 1.2: *Classification of non-coding RNAs. Housekeeping ncRNAs include: rRNAs, tRNAs, snRNAs, snoRNAs. Regulatory ncRNAs include: miRNAs, siRNAs, piRNAs, lncRNAs, lincRNAs, and eRNAs. nt: nucleotides; rRNAs: ribosomal RNAs; tRNAs: transfer RNAs; snRNAs: small nuclear RNAs; snoRNAs: small nucleolar RNAs; miRNAs: microRNAs; siRNAs: short interfering RNAs; piRNAs: piwi-interacting RNAs; lncRNAs: long non-coding RNAs; lincRNAs: long intervening RNAs; eRNAs: enhancer RNAs.*

1.2.1.1 Long non-coding RNAs: biogenesis and action mechanisms

lncRNAs are arbitrarily defined as non-coding transcripts that exceed 200 nucleotides in length, and constitute the majority of the non-protein-coding transcripts (Mattick et al., 2023; Mathy and Chen, 2017; Statello et al., 2021). Many *lncRNAs* share similar features with *mRNAs*, as on a molecular level they are also capped, spliced and polyadenylated, resulting in their characterization as "*mRNA*-like". In contrast to *mRNAs*, *lncRNAs* generally have fewer exons and typically exhibit lower expression levels. Furthermore, the open reading frame of *lncRNAs* is typically shorter than 300 nucleotides, which is considered indicative of their non-coding properties (Salido-Guadarrama et al., 2023). As a result, *lncRNAs* have not, or limited, translation properties. *lncRNAs* are classified into five groups, depending on their position with respect to protein-coding genes: *sense*, *antisense* (*AS*), *bidirectional*, *intronic*, and *intergenic* (Kaikkonen et al., 2011). The majority of *lncRNAs* are transcribed as complex networks of overlapping *sense* and *AS lncRNAs*. These latest are defined according to the nearest protein-coding gene position, and have no ability to be translated into proteins. The *lncRNA* biogenesis takes place in the nucleus and shows similarities to the synthesis of *mRNAs*: they are transcribed by *RNA* polymerase II and harbor a 5'methyl-cytosine cap and 3'-poly (A) tail (Liu et al., 2021). Nearly all *lncRNAs* exhibit canonical splice sites leading to at least two transcript isoforms, mainly composed by two exons. After their biogenesis and processing, several *lncRNAs* migrate to the cytoplasm, where they organize in thermodynamically stable structures. The most recent comprehensive integration of *lncRNAs* from existing databases includes 95,243 *lncRNA* genes and 323,950 transcripts in humans (Li et al., 2023).

1.2.1.2 MicroRNAs: biogenesis and action mechanisms

miRNAs, one of the most studied types of *sncRNAs*, are defined as small *RNA* molecules containing 18 to 28 nucleotides in length (Ratti et al., 2020). *miRNAs* are involved in *RNA* silencing, and influence protein production post-transcriptionally by binding *mRNAs* in a sequence-dependent manner. Canonically, *miRNAs* are encoded by introns of coding or non-coding transcripts, with very few being encoded by exonic regions (Lin et al., 2006). In human cells, *miRNAs* primarily act by destabilizing the *mRNA*. Due to their natural structure, *miRNAs* target up to thousands of transcripts, making them good regulators of several cell signaling pathways (Ha, 2011). Similarly to *lncRNAs*, *miRNA* genes are transcribed by *RNA* polymerase II, which initially yields a primary *miRNA* (*pri-miRNA*) (Lee et al., 2004). The *pri-miRNAs* are recognized and cleaved at the end of the hairpin structure by the double stranded *RNA* binding protein (*DGCR8*), which forms a nuclear *miRNA* processor complex with the *RNase* III enzyme Drosha (*pre-miRNAs*). *Pre-miRNAs* are then exported from the nucleus and transported into the cytoplasm, where the *RNA* is further elaborated by the RISC loading complex (*RLC*). The *RLC* retains the endoribonuclease *DICER1* which discards the loop of the *pre-miRNA* hairpin. The resulting mature *miRNA* is loaded onto the *RNA* induced silencing complex and the *miRNA* is released and degraded (Winter et al., 2009). According to the biological database for *microRNA* sequences and annotations miRBase, currently 2844 *miRNAs* are annotated in humans (Kozomara et al., 2019).

1.2.2 Non-coding RNAs in cancer

ncRNAs regulate key pathways involved in tumorigenesis, including apoptosis, cell cycle, migration, metastasis, angiogenesis and drug resistance (Zhang et al., 2022; Yang et al., 2023). Depending on their promoter or suppressor role, *ncRNAs* can act as either tumor suppressors or oncogenes. *RNA* dysregulation in cancer occurs through a variety of mechanisms, such as mutations in the *RNA* processing machinery, or alterations in *DNA* methylation affecting the transcription of the *pri-RNA* transcript. It has also been determined that *RNA* signatures can distinguish between normal and cancerous tissues, as well as differentiate between cancer subtypes (Bhattacharyya et al., 2015; Beg et al., 2022). In the last few years, several studies have investigated the role of *ncRNAs* in drug resistance as well as biomarkers for early diagnosis (Romano et al., 2017; Uppaluri et al., 2023).

1.2.2.1 Long non-coding RNAs in cancer

The majority of studies have investigated *lncRNA* expression in tissue samples, and their association with patient's prognosis. According to several studies on *ncRNAs*, a *lncRNA* with a major role most types of cancer is *HOTTIP*, derived from the *HOXA* gene (Ghafouri-Fard et al., 2020). Upregulation of *HOTTIP* increases cancer progression in patients with renal cell carcinoma, hepatocellular carcinoma, acute myeloid leukemia, and gastric cancer. Current research shows that also *LUCAT1* plays an oncogenic role, promoting cancer progression in gastrointestinal tract cancers and colorectal cancer (Xing et al., 2021; Wu et al., 2020). In osteosarcoma, *LUCAT1* is a promising target for cancer treatment, as its downregulation is associated to a reduced cell proliferation, migration, and invasion, representing a strategy to minimize drug resistance (Han and Shi, 2018). The tumor suppressive role of *Pvt1b*, a *p53*-dependent isoform of the *lncRNA*, has emerged in lung cancer and osteosarcoma (Olivero et al., 2020; Wang et al., 2023). In pharmacology, the deactivation of *Pvt1b* has also shown to promote drug resistance. Other *lncRNAs* associated with a better cancer prognosis are *DIRC3*, observed in melanoma and thyroid cancer patients, and *MALAT1*, a nuclear *lncRNA*, involved in breast, gastric, and gallbladder cancer (Coe et al., 2019; Xiao et al., 2023; Wysocki et al., 2023; Tsyganov and Ibragimova, 2023; Li et al., 2018).

1.2.2.2 MicroRNAs in cancer

In a recent German cohort study, Raut et al. (2024) derived and validated a serum-based *miRNA* risk score (miR-score) for colorectal cancer and other cancer types, such as breast, lung, and prostate cancer. This study particularly emphasized the potential of serum *miRNA* biomarkers for cancer-specific risk prediction, showing that the miR-score showed significant inverse associations with breast and lung cancer risk and a positive trend with prostate cancer. Other studies to date, as in the case of *lncRNAs*, have mainly focused on tissue-derived *miRNAs*. *miR-125b*, derived from the *MIR100HG lncRNA*, is one of the most studied *miRNAs* (Lu et al., 2017). Through targeting the *MALAT1 lncRNA*, *miR-125b* acts as either oncogene, or tumor suppressor, depending on the cancer type. *miR-125b* is well-known for being an oncogene in haematological malignancies, but serves a tumour suppressor in solid tumors, such as esophageal squamous cell carcinoma, bladder cancer, and hepatocellular carcinoma (Sun et al., 2013; Yang et al., 2021). The *let-7* and *miR-34* families are also rich in cancer-specific *miRNAs*, which mostly act as tumor suppressors, as they target many oncogenic genes including *E2F1*, *ARID3B*, *K-Ras* and *c-Myc* (Stahlhut and Slack, 2015). Studies on colon, lung, prostate, and pancreatic cancers highlighted that *let-7a*, *let-7b*, and *let-7c* are underexpressed in patients with cancer, compared to healthy controls (Ali et al., 2010; Ghanbari et al., 2015; Heegaard et al., 2012). In breast cancer, high *miR-34a* expression is associated with inhibition of the expansion of mammary gland stem cells through the suppression of *Wnt*/beta-catenin signaling (Bonetti et al., 2019). *miR-34b/c* also enhance cell attachment and suppress cell growth in lung cancer and hepatocellular carcinoma.

1.3 Non-coding RNAs for gallbladder cancer risk prediction

1.3.1 State of the art

Due to its heterogeneous nature, *GBC*'s molecular abnormalities underlying its pathogenesis are still not fully understood. Nonetheless, recent studies have succeeded in the attempt of identifying *ncRNAs* whose expression either promotes or inhibits *GBC* progression.

1.3.1.1 Long non-coding RNAs and gallbladder cancer

As in most human cancers, *p*53 overexpression is frequently observed in *GBC* (Yang et al., 2023). A study on Indian patients highlighted that overexpression of *p*53 is common in 56.25% of *GBC* cases compared to subjects with chronic cholecystitis or controls (Ghosh et al., 2013). In an old study on Spanish patients, 70.7% of gallbladder carcinomas exhibited overexpression of *p*53, with the expression increasing by tumor stage (Hidalgo Grau et al., 2004). The tumor suppressor *MEG*3 is another *lncRNA* contributing to the regulatory mechanisms of *GBC*. *MEG*3 provides a better *GBC* prognosis by modifying the activity of the *p*53 promoter, and through regulation of proliferation and apoptosis of *GBC* cells via induction of *NF-kB* signaling (Li et al., 2022). In *GBC* cell lines, *MEG*3 overexpression has shown to reduce the colony-forming ability of *GBC* cells and increase apoptosis rates by interacting with *p*53 (Xu et al., 2022). The oncogene *MALAT*1, responsible for tumor cell proliferation and metastasis, is overexpressed in *GBC* tissue samples by activating the *ERK/MAPK* signaling pathway. A higher expression of *CCAT*1 in *GBC* tissue is correlated with advanced tumor stages (T3 + T4) than early stages (T1 + T2) (Ma et al., 2015). *CCAT*1, known for its association with lymph node invasion in various cancers, has also been linked to metastasis in *GBC*, indicating that *CCAT*1 is a potential marker of poor *GBC* prognosis.

1.3.1.2 MicroRNAs and gallbladder cancer

Compared to *lncRNAs*, relatively few *miRNA-GBC* biomarkers have been identified over the past years. In 2013, a Japanese study determined that upregulation of *miR*-155 in *GBC* patients is associated with a poor prognosis, significantly increasing the risk of lymph node metastasis and vessel invasion (Kono et al., 2013). A Chinese study on *GBC* cell lines found similar properties for *miR*-144, which promotes migration and invasion of *GBC* cells by inhibiting the *RECK* gene (Zheng et al., 2020). Goeppert et al. (2019) suggested that *miR*-145-5*p* plays a functional role in biliary tract cancer by activating *STAT*1. Ishigami et al. (2018) demonstrated that *IL*-6/*STAT*-3 signaling pathway plays a crucial role in the growth of bile duct cancer cells and is associated with suppression of *miR*-31 expression. *miR*-125*b*, *miR*-136, and *miR*-30*a*-5*p* have been further identified as potential *GBC* suppressors (Yang et al., 2017; Niu et al., 2020; Ye et al., 2018). Another *miRNA*, *miR*-33*a*, has shown tumor-suppressive activity in *GBC* by inhibiting *IL*-6-mediated tumor progression through its interaction with *Twist*, a key regulator of cancer cell metastasis and invasion (Gao et al., 2020).

1.4 Recent advancements in non-coding RNA expression quantification

Nowadays, the identification and detection of an increasing number of *ncRNAs* have been facilitated by the advancement of next-generation sequencing (*NGS*) technologies. This process goes through sample preprocessing, library preparation, sequencing, and finally to bioinformatics analysis (Satam et al., 2023).

Below is a summary of the main sequencing techniques commonly used today, along with the public resources utilized in this thesis for analysis and prediction of *ncRNA* interactions with other biomolecules.

1.4.1 Microarrays

Microarray is a popular method used to perform global or parallel transcriptome expression analysis in different cell or tissue types (Yan et al., 2012). In brief, a large number of oligonucleotide probes are spotted on a solid surface. Then, sequences are hybridized from samples, and finally target sequences are fluorescently labeled. Despite its popularity, microarray holds some limitations, as it is only able to detect *RNAs* whose sequences are already known (Sun et al., 2020). Therefore, discovery of novel transcripts is not possible with such technique.

1.4.2 RNA sequencing

RNA sequencing (also called RNA-seq) is currently the most popular sequencing technology for *ncRNA* expression detection and discovery (Djebali et al., 2012; Wang et al., 2009). One more reason of this technique's popularity is that it can also identify single nucleotide polymorphisms (*SNPs*). RNA-seq is performed by converting *RNAs* into *complementary DNAs* (*cDNAs*) with either oligo (dT)-primers or random primers (Boone et al., 2018).

1.4.3 Non-coding RNA data preparation and exploration techniques

The aim of data pre-processing for large scale expression data, is to address systematic experimental bias and technical variation through preservation of biological variation (Nazer et al., 2023). Additionally, visually exploring *ncRNA* data is essential for gaining insights into the data characteristics. A comprehensive analysis of sequencing data facilitates the characterization of variation among replicates and helps determine whether the defined experimental groups exhibit significant differences.

In this thesis, quantile normalization is employed as the primary data preparation technique, while principal component analysis (*PCA*) is utilized for data exploration. This section provides a brief overview of both methodologies.

1.4.3.1 Quantile normalization

The purpose of normalization is to eliminate or minimize technical variability. Dozens of normalization methods have been implemented in the last twenty years to account for experimental differences between arrays. Some examples are quantile normalization, the Reads Per Kilobase per Million mapped reads (*RPKM*), and the DESeq (Bolstad et al., 2003; Mortazavi et al., 2008; Love et al., 2014). Quantile normalization, initially designed for gene expression microarrays, has since been adapted for use across a wide range of high-dimensional omics platforms, including *RNA* sequencing (Zyprych-Walczak et al., 2015). Quantile normalization is designed to align the distribution of *RNA* counts across different runs. Its fundamental assumption is that all samples, regardless of their class or condition, exhibit a similar distribution of *ncRNA* expression levels. This helps reduce technical variation and enhance comparability across datasets. The quantile normalization process is straightforward: *RNAs* within each sample are ranked according to their expression values. For *RNAs* occupying the same rank across samples, their average value is calculated. This average is then assigned to all *RNAs* holding that particular rank. The final step involves reordering the *RNAs* in each sample back to their original positions, maintaining their relative ranks. In this thesis, a specialized form of quantile normalization, known as class-specific quantile normalization, is applied. This approach first separates the data based on phenotype classes, such as disease versus control groups, and then performs quantile normalization independently within each class. After normalization, the data from both classes are recombined into a single dataset. This method helps mitigate false positives or negatives that may arise when averaging out samples with different expression profiles, such as those from cancerous and normal tissues.

1.4.3.2 Genetic principal component analysis

PCA is a statistical technique that processes large datasets by reducing data dimensionality to a smaller set of linearly transformed dimensions, which capture the overall variation present in the dataset (Ringnér, 2008). *PCA* is often employed as a preliminary analysis for data exploration and description in population genetics research. Its applications are extensive: it can be used to assess the population structure among a group of individuals, exemplify ancestry and relatedness, analyze admixture, and detect outliers. One of the key advantages of *PCA* in population genetics is that the distances between clusters of individuals may correspond to the genetic and geographic distances between those groups. *PCA* results are typically illustrated as a two-dimensional plot, where the axes represent the principal components (*PC*s) that account for the variation within the dataset. The first principal component (*PC*1) captures the highest level of variation, followed by the second principal component (*PC*2), and so on.

1.4.4 Public resources

In recent years, multiple databases cataloging interactions between *ncRNAs* and genes or proteins have emerged (Rigden and Fernández, 2021). These advancements were driven by bioinformatics innovations, which enabled the development of databases and open-source tools offering summary statistics from genetic association studies (e.g., the ncRNA-eQTL database), pathway analysis (e.g., DIANA miRPath, MiEAA software), and experimentally derived data. These resources present substantial benefits, significantly reducing both costs and time in *ncRNA* research and functional annotation.

1.4.4.1 DIANA miRPath

DIANA-miRPath v3.0 offers an online platform designed to analyze the regulatory functions of *miRNAs* and identify the pathways they influence (Vlachos et al., 2015). The latest version supports functional annotation of single or multiple *miRNAs* through standard hypergeometric distributions, empirical distributions, and meta-analysis statistics. It includes comprehensive coverage of KEGG molecular pathways and various segments of Gene Ontology across seven species, including *Homo Sapiens*. The platform integrates over 600,000 experimentally validated *miRNA* targets from DIANA-TarBase, allowing users to supplement or replace in silico predictions with high-quality experimental data from DIANA-microT-CDS and TargetScan (Vergoulis et al., 2012). One of the advantages of using this tool is that it is open-source and freely accessible without the need for user registration.

1.4.4.2 MiEAA software

MiEAA is a web-based tool that offers a wide range of statistical tests, such as over representation analysis and *miRNA* set enrichment analysis (Aparicio-Puerta et al., 2023; Backes et al., 2016). In addition to its variety of statistical analyses, MiEAA provides extensive functionality in terms of *miRNA* classifications. The tool includes over 14,000 *miRNA* sets, covering areas like pathways, diseases, organs, and target genes. Notably, MiEAA is applicable to both *miRNA* precursors and mature *miRNAs*, enhancing its utility across different types of analyses. Like the DIANA miRPath software, MiEAA is open-source and freely accessible to users without requiring registration.

1.4.4.3 ncRNA-eQTL database

The ncRNA-eQTL database is an extensive resource focused on *ncRNA*-related expression quantitative trait loci (*eQTLs*), utilizing large cancer sample datasets to assess the impact of genetic variants on *ncRNA* expression (Li et al., 2020). This database includes cis- and trans-*eQTLs*, survival-*eQTLs*, and genome-wide association study (*GWAS*) *eQTLs*, and offers an intuitive interface for querying, browsing, and downloading relevant data. To the best of current knowledge, it is the first resource specifically designed to identify *ncRNA-eQTLs* across multiple cancer types, with the number of detected *eQTLs* increasing with sample size. While many previous *eQTL* studies analyzed fewer than 300 samples (Ongen et al., 2016), this database includes 12 cancer types with over 300 samples, making it one of the most comprehensive available *ncRNA-eQTL* resources.

1.5 Objectives

The primary objective of this thesis is to investigate the genetic and molecular mechanisms that contribute to the development of *GBC*, an aggressive and understudied malignancy. Specifically, it seeks to identify, validate, and functionally characterize circulating *ncRNA* biomarkers for early *GBC* detection and risk prediction before clinical onset. By examining the role of *ncRNAs* in two distinct populations, this thesis sheds light on their involvement in *GBC* development across Europeans and Latin Americans. The research presented here is structured around two major *ncRNA* types, *lncRNAs* and *miRNAs*, and focuses on two distinct populations: Chileans and Europeans, respectively. In two separate studies, *ncRNA* expression levels were evaluated in both tissue and serum samples.

Study 1: Identification of circulating *long non-coding RNAs* associated with gallbladder cancer risk:

- Preselect *lncRNAs* based on their expression changes along the sequence of gallstones, dysplasia, and *GBC* in gallbladder tissue samples.
- Identify and validate genetic variants (cis-*lncRNA-eQTLs*) associated with the expression of the preselected *lncRNAs* in serum samples.
- Predict *lncRNA* expression levels based on individual genotypes and assess their association with *GBC* risk in additional serum samples.

Study 2: Identification and validation of circulating *microRNAs* associated with gallbladder cancer risk in Europeans:

- Preselect *miRNAs* based on expression differences between normal and *GBC* tissue from German patients with *GBC* and gallstone disease.

- Screen *miRNA* expression differences in prospective serum samples from *GBC* cases and controls.
- Validate the *miRNA-GBC* risk associations in additional European prospective cohort serum samples.
- Investigate the interaction between identified *miRNAs* and their target genes through pathway analysis.
- Perform meta-analysis on validated *miRNAs*.

Through these analyses, this thesis aims to contribute to the understanding of *ncRNA* dysregulation in *GBC* and to develop potential non-invasive diagnostic tools for early detection and risk assessment.

Major parts of the content of this thesis have already been published (Blandino et al., 2022). All calculations were performed with the statistical software package R, version 4.2.2 (R Core Team, 2023). Codes to reproduce all the results are provided in Appendix B.

Chapter 2

Materials and methods

2.1 Study design, investigated patients and samples

Comment: Parts of the following Chapter have already been published in Cancers (Blandino et al., 2022). The original manuscript was written by myself, but also contains comments and corrections from the co-authors.

2.1.1 Identification of circulating long non-coding RNAs associated with gallbladder cancer risk

2.1.1.1 Study design

In the first study, *lncRNAs* linked with *GBC* progression are identified through a three-stage study design. This involves the screening of three distinct Chilean datasets, each one containing unique information on *lncRNA* expression profiles and individual genotypes.

lncRNAs exhibiting expression changes between gallstones, dysplasia, and *GBC* were first preselected on a dataset (*lncRNA* preselection dataset) comprising exclusively *lncRNA* expression data from gallbladder formalin-fixed paraffin-embedded (*FFPE*) tissue. *lncRNAs* were declared as preselected and passed to the next step only if they met the defined significance thresholds (adjusted *p-value* < 0.05), were measured in serum, were annotated as *lncRNAs*, and are not duplicated.

Then, *SNPs* in close proximity (located on the same chromosome as the *lncRNA*) to the preselected *lncRNAs* (cis-*lncRNA-eQTLs*) were identified through the ncRNA-eQTL database: http://ibi.hzau.edu.cn/ncRNA-eQTL/ (Li et al., 2020). These cis-*lncRNA-eQTLs* were subsequently validated in a second independent dataset, the *lncRNA-eQTL* validation dataset, which includes both *lncRNA* expression and individual genotypes.

Genetic associations from the previous step were exploited in a third independent data source containing only *SNP* information (the *lncRNA-GBC* association dataset) to predict the expression levels of circulating *lncRNAs* based on individual genotypes. The relationship between predicted *lncRNA* expression and *GBC* risk was finally evaluated and consistency with the preselection findings examined, as described in detail in the following sections.

2.1.1.2 Investigated patients and samples

For the *lncRNA* preselection dataset, 98 cholecystectomized Chilean patients diagnosed with gallstones (n = 31), dysplasia (n = 35), or *GBC* (n = 32) were invited to enroll to the study. With the exception of two patients with *GBC* who had missing information regarding gallstones, all *GBC* and dysplasia individuals in the study were confirmed to carry gallstones. Upon obtaining written informed consent, patients' tissue samples and clinical data were collected using standardized case report forms. Patients were recruited across seven hospitals throughout Chile. Exclusions were made for samples stored for over 5 years, and for patients with porcelain gallbladder, polyps, non-cholesterol stones, or abnormalities of the pancreatic or bile ducts. The study has been approved by the appropriate ethics committees in Chile. Additional details on the samples are described into more details in the following paper from Brägelmann et al (2021).

The dataset used for the identification and validation of cis-*lncRNA-eQTLs* comprises genome-wide data along with serum *lncRNA* expression data from 110 participants enrolled in Chilean studies on Chagas (n = 88) and chronic obstructive pulmonary disease (*COPD*, n = 22) (Díaz-Peña et al., 2022; Apt et al., 2021). *COPD* patients were recruited after providing written informed consent at the Hospital Regional de Talca located in Talca, south of Chile. Study participants have been previously described by Olloquequi et al. (2018). Ethics approvals were obtained from the Ethics Committees of Maulean Health Service and Universidad Autónoma de Chile. Patients with Chagas disease, i.e. individuals showing clear signs of chronic *T cruzi* infection, were invited to participate in the study upon medical written informed consent.

Prediction of serum *lncRNA* expression was performed using individual genotype data from 540 Chilean *GBC* patients and 2397 population-based controls. *GBC* subjects were recruited under informed consent between 2014 and 2020, with the majority (77%) diagnosed following cholecystectomy, except for a few cases diagnosed without surgical intervention. Ethics approvals were provided by the Medical Faculty of the Universidad de Chile (approval #123-2012), Southeast Health Service of the Santiago Metropolitan Region, Health Service of Concepcion Hospital (approval #16-11-97) and Central Santiago Metropolitan Health Service (approval #135). Controls were selected from the Chilean cohort of the Consortium for the Analysis of the Diversity and Evolution of Latin America (*CANDELA*), as well as from Chilean studies on *COPD* and Chagas disease (Barahona Ponce et al., 2021; Lorenzo Bermejo et al., 2017; Boekstegers et al., 2020). Recruitment of the *CANDELA* samples was performed upon written informed consent in Arica, in the northern part of Chile. Part of the collective has been previously described by Ruiz-Linares et al (2014). Ethics approvals for the controls were obtained from the Universidad de Tarapacá and the University College London.

The complete applied methodology and the main datasets' characteristics are represented in Figure 2.1.

2.1.2 Identification and validation of circulating microRNAs associated with gallbladder cancer risk

2.1.2.1 Study design

The design of the second study of this thesis is shown in Figure 2.2. This study follows a three-stage approach based on preselection, screening, and validation of differentially expressed *miRNAs* in European *GBC* individuals. Preselection relied on *miRNAs* exhibiting expression differences in *GBC FFPE* tissue samples compared to gallstone patients. Candidates which were not measured in serum as well as *miRNAs* which according to literature are potentially linked to confounders in serum (age, sex, smoking, *BMI* and physical activity), were excluded prior screening (Rounge et al., 2018). The preselected candidates were

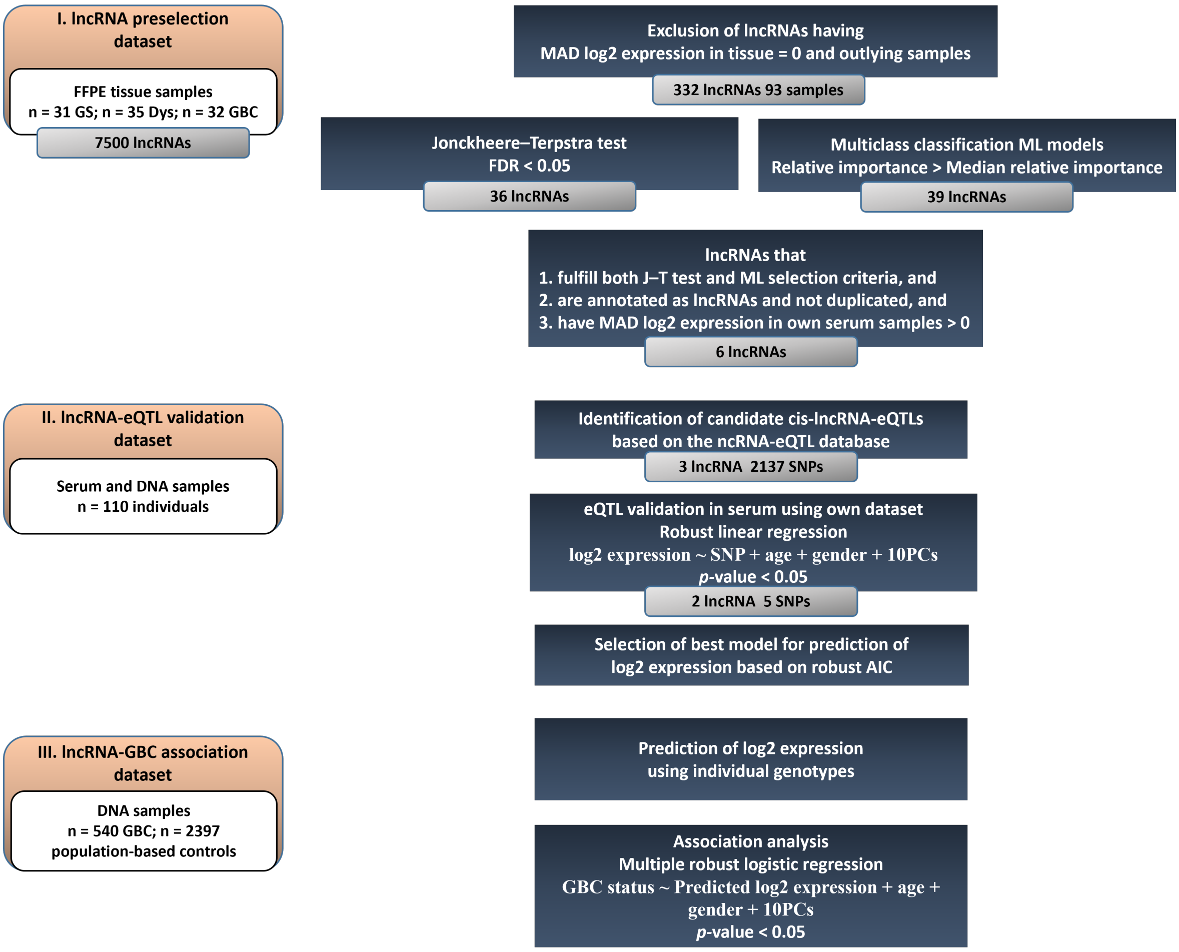
I. lncRNA preselection dataset
FFPE tissue samples
n = 31 GS; n = 35 Dys; n = 32 GBC
7500 lncRNAs
Exclusion of lncRNAs having
MAD log2 expression in tissue = 0 and outlying samples
332 lncRNAs 93 samples
Jonckheere–Terpstra test
FDR < 0.05
36 lncRNAs
Multiclass classification ML models
Relative importance > Median relative importance
39 lncRNAs
lncRNAs that
1. fulfill both J–T test and ML selection criteria, and
2. are annotated as lncRNAs and not duplicated, and
3. have MAD log2 expression in own serum samples > 0
6 lncRNAs
II. lncRNA-eQTL validation dataset
Serum and DNA samples
n = 110 individuals
Identification of candidate cis-lncRNA-eQTLs
based on the ncRNA-eQTL database
3 lncRNA 2137 SNPs
eQTL validation in serum using own dataset
Robust linear regression
log2 expression ~ SNP + age + gender + 10PCs
p-value < 0.05
2 lncRNA 5 SNPs
Selection of best model for prediction of
log2 expression based on robust AIC
III. lncRNA-GBC association dataset
DNA samples
n = 540 GBC; n = 2397
population-based controls
Prediction of log2 expression
using individual genotypes
Association analysis
Multiple robust logistic regression
GBC status ~ Predicted log2 expression + age +
gender + 10PCs
p-value < 0.05

Figure 2.1: *Flowchart of the long non-coding RNA study design. lncRNA: long non-coding RNA; FDR: false discovery rate; J-T test: Jonckheere-Terpstra test; SNP: single nucleotide polymorfism; PC: principal component; p-value: probability value; GS: gallstones; Dys: dysplasia; GBC: gallbladder cancer; AIC: Akaike's information criterion; MAD: median absolute deviation; FFPE: Formalin-fixed paraffin-embedded; eQTL: expression quantitative trait loci. (Adapted from Blandino et al. (2022))*

subsequently screened in prospective European serum samples, and only *miRNAs* displaying expression patterns consistent with those observed during preselection were chosen for further validation. Validation was carried out on additional serum samples and supported by meta-analysis. As sensitivity analysis, pathway analysis was conducted on the set of preselected *miRNAs* in *FFPE* tissue. Correlations between target genes from the significant pathways and the validated *miRNAs* were finally investigated in the pooled serum data.

It is important to note that following preselection and screening, and prior to *miRNA* sequencing for validation, this study and the *miRNA* validation protocol were officially registered at the *German Clinical Trials Register* (drks.de, March, 5th 2021) and the International Clinical Trials Registry Platform of the *World Health Organization* (*WHO*, https://trialsearch.who.int/Trial2.aspx?TrialID=DRKS00024573).

2.1.2.2 Investigated patients and samples

The preselection dataset used to identify *miRNAs* differentially expressed in *FFPE* tissue includes eight normal, non-neoplastic gallbladders and 40 *GBC* samples. Tissue samples from patients who underwent surgical removal of the gallbladder (cholecystectomy) were obtained by the tissue bank of the National Centre for Tumour Diseases (NCT Heidelberg, Germany). Cancer patients underwent cholecystectomy at the time of diagnosis and received no treatment prior to sampling. *GBC* cases were histologically confirmed by at least two specialized pathologists at the Institute of Pathology at Heidelberg University Hospital. Non-neoplastic gallbladder tissue samples were collected from cholecystectomized patients with gallstone disease and served as the reference group for normal tissue in this study. More information of this cohort can be found in the publication from Goeppert et al. (2019).

After *miRNA* preselection based on *FFPE* gallbladder tissue, 74 serum samples were investigated from three European prospective cohorts (n = 37 *GBC* case-control pairs, screening dataset). Data and samples were provided by the Norwegian *Janus Serum Bank* (n = 27 *GBC* case-control pairs), the German Early Detection and Optimised Therapy of Chronic Diseases in the Elderly Population (*ESTHER*) study (n = 9 *GBC* case-control pairs), and the German Heinz Nixdorf Recall (*HNR*) study (n = 1 *GBC* case-controls pair). The *Janus Serum Bank* is a population-based biobank for cancer research that contains pre-diagnostic biospecimens from 318,628 Norwegians (Langseth et al., 2017). Between 1972 and 2004, residual blood serum samples were collected in 17 Norwegian counties. The average age of study participants at enrollment was 41 years. Individuals were followed up from the date of first serum donation to the date of cancer diagnosis, emigration or death. Information on smoking, physical activity and *BMI* was available for 90% of participants. The *ESTHER* study is a cohort study conducted in Saarland, a federal state in south-west Germany (Raum et al., 2007). Between 2000 and 2002, 9,940 participants aged between 50 and 74 years were enrolled as part of routine medical check-ups. Cancer cases were determined on the basis of the cancer diagnoses reported by the participants themselves, which were also confirmed by physicians, and by record linkage with the Saarland Cancer Registry. The *HNR* study is a cohort study where study participants were selected at random from mandatory lists of places of residence (Stang et al., 2005). Between 2000 and 2003, 4,814 participants aged between 45 to 75 years were enrolled in the metropolitan Ruhr area in Germany and followed up for a median of 5 years. As only one case-control pair was available from the *HNR* study, this cohort was merged with the *ESTHER* study, both of which consist of German individuals. All controls were matched by age and sex with *GBC* cases.

The most promising *miRNAs* identified in the screening dataset were subsequently investigated in the validation dataset, which includes data and serum samples (n=36 *GBC* case-control pairs) from three large European prospective cohorts: the Norwegian Helseundersøkelsen i Nord-Trøndelag Health (*HUNT*) study (n = 15 *GBC* case-control pairs), the Finnish *FINRISK* cohort (n = 9 *GBC* case-control pairs), and the Swedish *TwinGene Registry* (n = 12 *GBC* case-control pairs). *HUNT* is a Norwegian population-based health study (Krokstad et al., 2013). Since 1984, more than 229,000 adults aged 20 years or older have joined the study. Biological samples were available from 95,000 study participants who were followed for nearly 40 years. The participation rate of those invited to join the study was high, ranging from 54% to 89%, making the cohort a good representation of the general Norwegian population. The Finnish population-based *FINRISK* study is part of the evaluation of the North Karelia project, a large community-based disease intervention started in 1972 (Borodulin et al., 2018). The target population of the *FINRISK* study was 25 to 74-year-old Finns who had lived in Finland for at least one year. To date, the *FINRISK* study has reached a total of 101,451 individuals from nine cross-sectional studies, who were followed up until 2014. The Swedish *TwinGene Registry* was established in the late 1950s to initially investigate the role of environmental factors such as smoking and alcohol on disease (Lichtenstein et al., 2002). In 2004, 22,000 twins among the older study participants were invited for blood collection for DNA and serum biobanking. The sample collection was completed in 2008 with an overall response rate of 56%. All controls included in the study were age- and sex-matched with the *GBC* cases.

In this study, *miRNA* expression levels of ten *GBC* cell lines (*G*-415, *GB*-*d*1, *Mz*-*Cha*-1, *NOZ*, *OCUG*-1, *OZ*, *SNU*308, *TGBC*1 (also known as *TGBC*1*TKB*), *TGBC*2 (also known as *TGBC*2*TKB*) and *YoMi* were also analyzed. Cell lines were tested for mycoplasma contamination using MycoAlert (Lonza, Basel, Switzerland) and authenticated by short tandem repeat analysis. More details on the cell-lines are available on the paper from Scherer et al. (2020).

All European samples analyzed in this study were collected upon ethical approval by the following institutions: Medical Faculty Heidelberg (Preselection dataset, *ESTHER*, #58/2000, *HNR*), the THL Biobank (*FINRISK*, #BB2016_32), the Regional Committee for Medical and Health Research Ethics (*Janus*, #2016/1290, *HUNT*, #2016/1222), and EPN (*TwinGene*, #2016/2:11). All participants provided written informed consent prior to participation.

2.2 Generation of small-RNA expression and genome-wide genotype data

2.2.1 RNA and DNA extraction

The protocol followed for *RNA* extraction, isolation, and profiling from *FFPE* gallbladder tissue has been described previously (Goeppert et al., 2019). Briefly, small-*RNA* samples were purified for microarray hybridization from microdissected *FFPE* material using the miRNeasy *FFPE* Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions. Agilent SurePrint Human *miRNA* microarrays (G4872A, miRBase Release 19.0, Agilent Technologies, Santa Clara, CA), which include 2006 human *miRNAs*, were used for

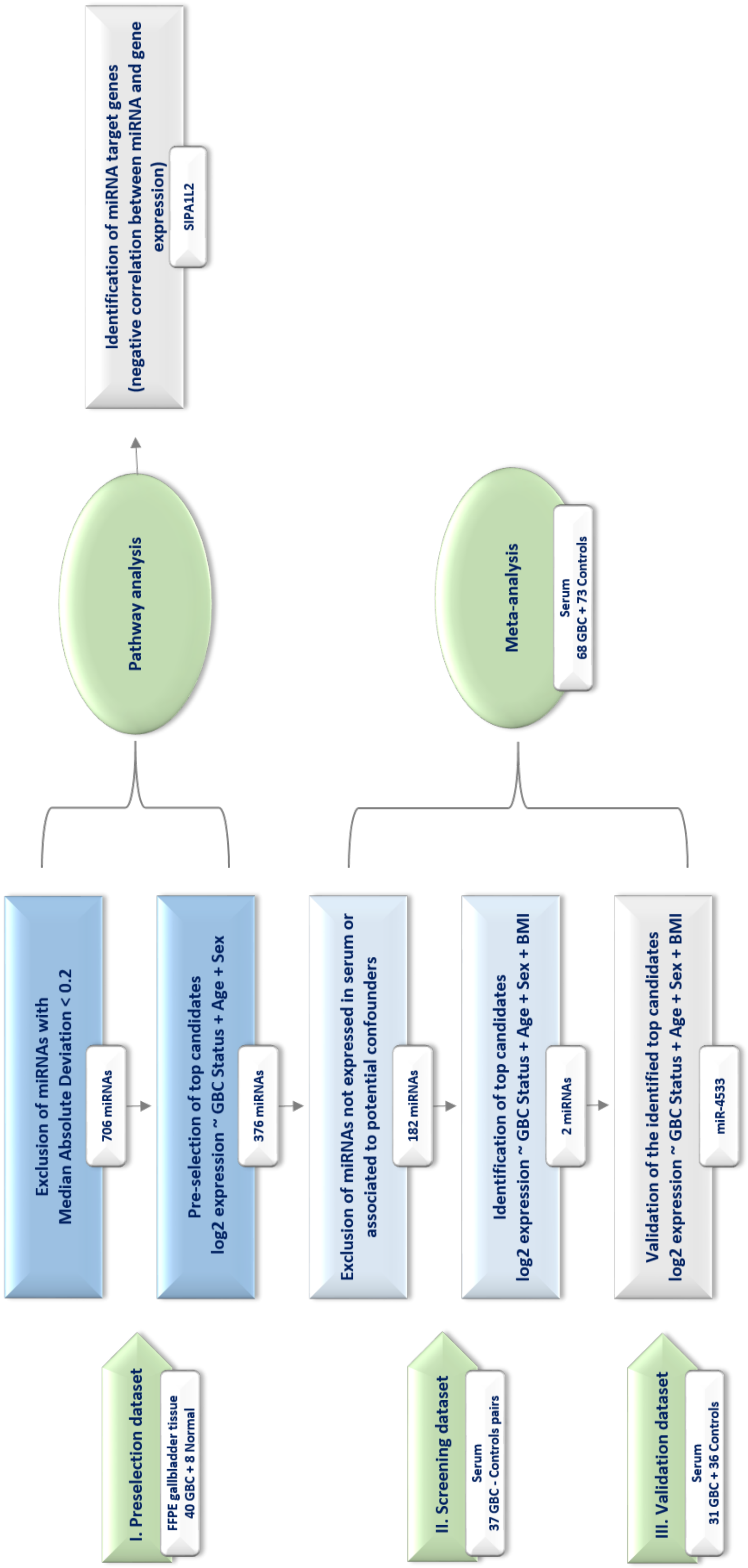
I. Preselection dataset
FFPE gallbladder tissue
40 GBC + 8 Normal
II. Screening dataset
Serum
37 GBC - Controls pairs
III. Validation dataset
Serum
31 GBC + 36 Controls
Exclusion of miRNAs with Median Absolute Deviation < 0.2
706 miRNAs
Pre-selection of top candidates log2 expression ~ GBC Status + Age + Sex
376 miRNAs
Exclusion of miRNAs not expressed in serum or associated to potential confounders
182 miRNAs
Identification of top candidates log2 expression ~ GBC Status + Age + Sex + BMI
2 miRNAs
Validation of the identified top candidates log2 expression ~ GBC Status + Age + Sex + BMI
miR-4533
Pathway analysis
Identification of miRNA target genes (negative correlation between miRNA and gene expression)
SIPA1L2
Meta-analysis
Serum
68 GBC + 73 Controls

Figure 2.2: *Flowchart of the microRNA study design. miRNA: microRNA; GBC: gallbladder cancer; BMI: Body-mass index; FFPE: Formalin-fixed, paraffin-embedded.*

miRNA profiling of normal gallbladder and *GBC* tumor samples. Labelling, hybridization and data processing were performed following the manufacturer's recommendations.

The protocol applied for small-*RNA* extraction and sequencing from serum samples has also been previously described (Umu et al., 2018; Rounge et al., 2018). Briefly, *RNA* was extracted from 2 x 200 μl (screening) and 1 x 200 μl of serum (validation) using phenolchloroform separation and the miRNeasy serum kit (Cat. no 1071073, Qiagen) on a QIAcube (Qiagen). During *RNA* extraction, G glycogen (Cat. no AM9510, Invitrogen) was used as carrier. Ampure beads XP (Agencourt) were used to concentrate the eluate.

Genomic *DNA* was extracted under standard laboratory procedure and standard commercial kits. As quality control measures, intraplate and interplate replicates and blinded duplicates were employed at 5%.

2.2.2 Small-RNA sequencing

The NEBNext Small-*RNA* kit was used to produce *RNA* sequencing libraries, which were sequenced on the *HiSeq* 2500 and 4000 (screening), and Novaseq 6000 (validation) platforms (Illumina, San Diego, CA, USA) for average depths of 18 M (screening) and 22 M reads per sample (validation), enabling to capture mapped *sncRNAs* fragments of up to 47 base pairs. *RNA* counts were calculated using the *sncRNA* pipeline (https://github.com/sinanugur/sncRNA-workflow/) (Umu et al., 2018). First, reads were adapter-trimmed (AdapterRemoval v2.1.7) (Schubert et al., 2016). Then, adapter-trimmed reads were mapped to the human genome (*hg*38) by Bowtie2 v2.2.9 aligner in end-to-end mode (Langmead and Salzberg, 2012). HTSeq was used to count reads mapped to *sncRNA* regions in *miRBase* (v22.1) and GENCODE v26 annotations (Anders et al., 2015).

2.2.3 Genotyping and data quality control

Genotyping of study participants was conducted using Illumina's OmniExpress and Global screening arrays (*GSA*). Both arrays included more than 700,000 genome-wide *SNPs*.

Genetic variants were filtered to exclude *SNPs* with a minor allele frequency (*MAF*) lower than 1% or a missing call rate above 5% . Also samples with a missing call rate over 5% were left out. Identity by descent (*IBD*) kinship coefficients were calculated to address for relatedness among individuals ($IBD > 0.1$). Within each related pair of individuals, the subject showing the lowest call rate was systematically excluded from the analysis. Following linkage disequilibrium (LD) pruning at $r^2 > 0.1$, 36,175 variants from the GSA array were utilized for subsequent genetic PCA, and $Mahalanobis$ distances (MD) were computed to account for samples with outlying genotypes, specifically targeting the 5% of individuals exhibiting the lowest statistical depth. Calculation of MAF and call rates was implemented using the R package available in the Bioconductor's repository `snpStats` (Solé et al., 2006). IBD kinship coefficients and LD pruning were performed using the R package `SNPRelate` (Zheng et al., 2012). PCA was carried out using the `eigenstrat` function available at: www.popgen.dk/software/index.php/Rscripts (Price et al., 2006).

2.3 Statistical analyses

2.3.1 Multiple imputation of missing genotype data

Missing genotypes were imputed with the TOPMed reference sample via the TOPMed imputation server, accessible at https://imputation.biodatacatalyst.nhlbi.nih.gov/ (Taliun et al., 2021).

2.3.2 Prediction of small-RNA expression based on individual genotypes

In the first study of this thesis, after obtaining the list of cis-$lncRNA$-$eQTLs$ associated with the preselected $lncRNAs$ from the ncRNA-eQTL database, robust linear regression models were fitted to validate the identified associations. Models were adjusted for confounders, as individual age, sex and the first ten genetic PCs:

$$log_2 Expression \sim SNP + Age + Sex + 10\,PCs \tag{2.1}$$

The investigated models included four types of penetrances: additive (number major alleles), three-genotype (genotype as a factor), dominant (affect allele against the other genotypes), recessive (other allele against the affect allele). After fitting single models for each genetic variant, model selection was performed including the different configurations of the identified cis-$lncRNA$-$eQTLs$. Also here, models were adjusted for age, sex, and the first ten PCs.

The selected model for prediction was the one with the lowest robust Akaike's information criterion (*RAIC*).

Individual genotype-based *lncRNA* expression in serum was predicted considering the summary statistics from the previous step (β_i) and the individual genotype (A_i) encoded based on the selected penetrance model:

$$\text{Predicted } log_2Expression = \sum_{i=1}^{k} \beta_i A_i \tag{2.2}$$

Ultimately, the association between genotype-based serum *lncRNA* expression and *GBC* risk was evaluated on the *lncRNA*-*GBC* association dataset through robust logistic regression models. The fitted models employed a tuning constant c of 1.2 in Huber's psi-function, while accounting for individual age, sex, and the first ten genetic PCs:

$$\text{GBC status} \sim \text{Predicted } log_2Expression + Age + Sex + 10\,PCs \tag{2.3}$$

The function `rlm` from the R package `MASS` was used to fit robust linear regression models (Venables and Ripley, 2002). Coefficients' *p-values* were calculated using the function `rob.pvals` from the R package `repmod` (Marin, 2021). *RAICs* were obtained using the function `AIC` in the R package `AICcmodavg` (Mazerolle, 2023).

2.3.3 Long non-coding RNA association analyses

lncRNA counts were log2-transformed and expression values with a median absolute deviation (*MAD*) equal to zero were left out from further statistical analyses. Counts were quantile normalized, first considering gallstone, dysplasia, and *GBC* samples separately, and then altogether. After normalization, global *lncRNA* expression profiles were examined through *PCA*. 5% of patients exhibiting outlying expression profiles, i.e. a low *MD*, were not included in the final dataset.

lncRNA preselection was carried out using both non-parametric and machine learning (*ML*) techniques. Monotonic increasing or decreasing changes from gallstones to *GBC* were firstly evaluated through two-sided *Jonckheere-Terpstra* (*J-T*) tests with 5000 permutations (Jonckheere, 1954). *P-values* were adjusted for multiplicity using false discovery rates (*FDRs*). The second method used to preselect differentially expressed *lncRNAs* is the extreme gradient boosting (*XGBoost*) algorithm, applied to train three-class classification *ML* models

(Chen and Guestrin, 2016). The preselection dataset was divided at random into training (n = 77) and test (n = 21) sets. The training dataset achieved class balance through upsampling, resulting in 27 samples each for gallstones, dysplasia, and *GBC*. Model's hyperparameters were tuned through five-fold cross validation using the training set only, and a random grid search approach was employed. Cross validation assessed the best model, i.e. the model with the lowest mean per class error. The model's performance was evaluated on the test set based on both mean per class error and weighted average area under the curve (*AUC*). Finally, *lncRNA* were sorted by relative importance.

The *J-T* tests were performed using the `JonckheereTerpstraTest` function available in the R package `DescTools`, and the *ML* algorithm was implemented using the `h2o` R package (Signorell, 2024; Fryda et al., 2014).

2.3.4 MicroRNA association analyses

miRNA read counts were log2-transformed and *miRNAs* with low *MAD* were excluded from subsequent analyses. In the preselection dataset, quantile normalization was first applied separately to *GBC* and normal samples, and then simultaneously to all samples. In the screening and validation datasets, quantile normalization was first applied to each cohort separately, then to *GBC* cases and controls, and finally to the complete dataset. *miRNA* expression profiles were examined through *PCA*. Outlying samples were subsequently excluded based on *MD*. The R package `stats` was used for *PCA* and statistical depth calculation (R Core Team, 2023).

Preselection, screening and validation of differentially expressed *miRNAs* were based on robust linear regression. The preselection regression models included *GBC* status, age categorized into quartiles, and sex. The screening and validation regression models additionally included *BMI* categorized into quartiles. *BMI* information was not available in the preselection dataset, and was therefore not considered as confounder in the model.

In the preselection stage, *p-values* from robust linear regression were adjusted for multiplicity using the *Bonferroni* method (for subsequent screening) and *FDR* (for pathway analysis), taking into account the number of *miRNAs* with *MAD* greater than zero. In the screening stage, *Bonferroni* and *FDR* adjustments for multiplicity considered the number of preselected *miRNAs* that were expressed in the serum samples, while in the validation stage multiplicity corrections were carried out according to the number of differentially expressed *miRNAs* identified in the screening stage. Robust linear regression models and each coefficients' *p-values* were evaluated, respectively, through the functions `rlm` in the R package `MASS`, and `rob.pvals` in the R package `repmod` (Venables and Ripley, 2002; Marin, 2021).

2.3.5 Calculation of genetic gallstone disease risk score

Genotype information was available for some participants in the *ESTHER* (n = 18), *HUNT* (n = 29), *FINRISK* (n = 16), and *TwinGene* (n = 17) studies. Therefore, differences in *miRNA* expression were also investigated as a function of individual polygenic risk scores (*PRS*) for gallstone disease. The summary statistics used on this purpose relied on the association between genetic variants and gallstone disease from the UK Biobank (18,417 gallstone disease cases and 390,150 controls) for variants that were robustly (*p-value* < 5x10-8) associated with gallstone disease in the study by Ferkingstad et al. (2018). After excluding variants and samples with missing call rates of more than 5%, variants with a *MAF* of less than 1%, *LD* pruning ($r^2 > 0.1$), and harmonization of reference and alternative alleles in the UK Biobank and in the investigated prospective cohorts, *PRS* were calculated by multiplying the estimated additive genetic effects (β_i) by the individual allele counts (A_i).

$$PRS_j = \sum_{i=1}^{N} \beta_i A_i \tag{2.4}$$

2.3.6 Meta-analysis

After validation, meta-analysis was performed to combine the results from all serum prospective cohorts using the `rma` function in the `Metafor` package (Viechtbauer, 2010). The input values for the function were beta estimates with their corresponding standard errors from each cohort, and the cohort sample sizes as weights. Both fixed-effects and random-effects meta-analysis were taken into account, using the function `forest`, also from the `Metafor` package, to plot the results of the meta-analysis, and creating the remaining plots using the R package `ggplot2` (Wickham, 2016).

2.4 Pathway analyses

Based on the list of preselected *miRNAs* in *FFPE* gallbladder tissue, the web-based software DIANA-miRPath v3.0 was used (http://diana.imis.athena-innovation.gr) for *miRNA*-based pathway analysis (Vlachos et al., 2015). The over-represented pathways were then sorted by FDR-corrected *p-values*. In addition to *miRNA* expression, *mRNA* expression values based on small *RNA* sequencing were also available for the analyzed serum samples, and this information was used to investigate the relationship between *miRNA* and *mRNA* expression for the validated *miRNAs*. The total number of genes in the five pathways with the smallest *FDR*-corrected *p-values* was considered for *Bonferroni* adjustment of *p-values* from one-sided *Spearman* tests (negative *miRNA*-*mRNA* correlation), and possible differences in *mRNA* expression between *GBC* cases and controls were assessed by robust linear regression models adjusted for age, sex, and *BMI*. Finally, the *miRNA*-*mRNA* relationship was visually inspected, as well as differences between *GBC* cases and controls in *mRNA* expression, and *mRNA* expression in *GBC* cell lines using scatter and dot-and-box plots.

Chapter 3

Results

Comment: Parts of the following Chapter have already been published in Cancers (Blandino et al., 2022). The original manuscript was written by myself, but also contains comments and corrections from the co-authors.

3.1 Identification of circulating long non-coding RNAs associated with gallbladder cancer risk

3.1.1 Long non-coding RNA preselection in tissue

In the preselection dataset, a total of 7,500 *lncRNAs* was detected. Among these, 7,168 $lncRNAs$ exhibited a MAD of 0, leading to their exclusion from subsequent analyses. The $lncRNA$ expression profiles of the remaining 332 *lncRNAs* are depicted on the PCA plot in Figure 3.1, panel A. The expression profiles of patients with gallstones and dysplasia displayed notable similarities (represented by green and yellow dots), whereas GBC cases were predominantly located in the upper region of the graph. Furthermore, five outlying individuals were excluded from the analyses due to their lower statistical depth in comparison to the other global expression profiles (indicated by black empty dots). After exclusion, the preselection dataset consisted of 332 *lncRNAs*, and 93 samples, which included 28 patients with gallstones, 34 with dysplasia, and 31 diagnosed with GBC.

Two-sided *J-T* tests were used to assess the monotonic increase or decrease in expression from gallstones, to gallbladder dysplasia, and to GBC of 36 $lncRNAs$ ($FDR < 0.05$) (Figure 3.1, panel B, Table A.1). In contrast, the applied ML model ($AUC = 0.88$, mean per class error = 0.23) identified 39 $lncRNAs$ with relative importance greater than the median (Figure A.1). Eighteen $lncRNAs$ in total were selected both by *J-T* tests and ML, which were all annotated as $lncRNAs$ and not duplicated. Only the $\log_2$ expression of six of them exhibited a MAD greater than 0 in serum samples from the cis-$lncRNA$-$eQTL$ validation dataset. cis-$eQTL$ information from the $eQTL$-database was only available for $AC084082.3$, $LINC00662$, and $C22orf34$, which were the only ones to undergo $lncRNA$-$eQTL$ validation (Figure 3.1, panel C). The expression of $AC084082.3$ and $LINC00662$ was associated with an increased risk of GBC, while expression levels of $C22orf34$ decreased advancing malignancy. In Table 3.1 the stratified expression characteristics in gallstones, dysplasia, and GBC of the three preselected candidates are shown. On average, except for $LINC00662$, the expression differences were larger between gallstone and GBC patients, than between gallstone and dysplasia. Stratified analyses revealed that larger expression differences were solely observed in relation to age, with $LINC00662$ being overexpressed especially in younger GBC patients ($\text{Age} < 60$), and $C22orf34$ being downregulated in older ones ($\text{Age} \geqslant 60$).

3.1.2 Expression quantitative trait loci validation in serum

Data pre-processing determined the exclusion, in the validation dataset, of 460,632 $SNPs$ with MAF smaller than 0.01, four subjects due to low call rate, and eight related individuals (IBD Kinship coefficients > 0.1). In Figure 3.2, panel A the genetic profiles of all the included individuals are shown. The genetic PCA plot highlighted the presence of five outlying individuals with low statistical depth (represented by empty dots). These samples were therefore excluded from the final dataset, which included only 93 individuals.

Based on data from the ncRNA-eQTL database, 161 $SNPs$ were linked to the expression of $AC084082.3$. However, ten of these had low MAF or call rate and were, therefore, excluded from subsequent analyses. Furthermore, regardless of the four penetrance models examined, robust linear regression analyses did not validate any associations with $AC084082.3$.

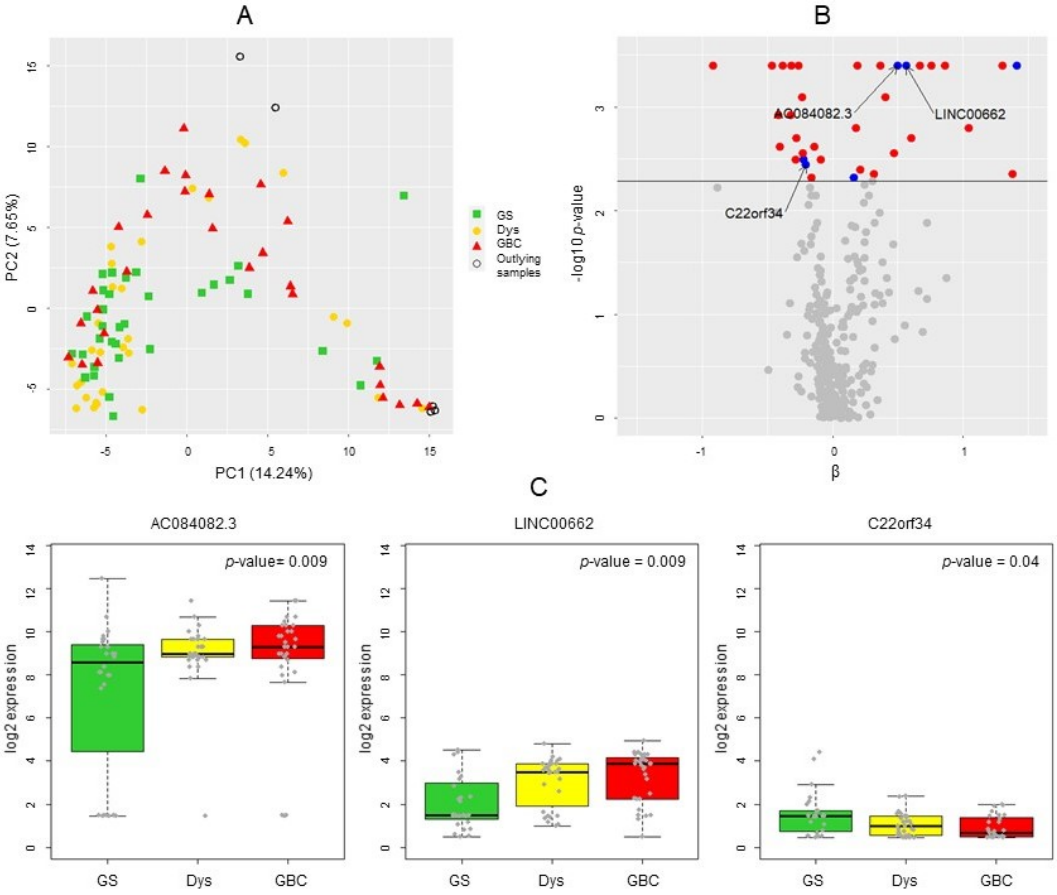

Figure 3.1: *Long non-coding RNA preselection in tissue. (A) Principal component analysis for the long non-coding RNA expression profiles in the preselection dataset. (B) Volcano plot for the preselection results. -log*$_{10}$ *p-values obtained by Jonckheere-Terpstra tests are represented on the y-axis. The applied significant threshold (FDR < 0.05) is represented by the black horizontal line. Preselected long non-coding RNAs with both non-parametric and machine learning techniques are depicted in red. Blue dots indicate long non-coding RNAs which, in addition, were also measured in serum samples. (C) Dot-and-box plots for the expression of* $AC084082.3$, $LINC00662$, *and* $C22orf34$ *in the preselection dataset. PC: principal component, p-value: probability value; GS: gallstones; Dys: dysplasia; GBC: gallbladder cancer. Adapted from Blandino et al. (2022)*

Table 3.1: *Expression of AC084082.3, LINC00662, and C22orf34 in the preselection dataset. Results are both global and stratified by patients' age and sex.*

Subgroup	**lncRNA**	**FDR**	**$\log_2$ expression**	**$\log_2$ expression difference**	**$\log_2$ expression difference**
			GS samples	**Dys vs GS**	**GBC vs GS**
			Median [5th;95th]	**Estimate [95%CI]**	**Estimate [95%CI]**
All	AC084082.3	0.009	8.23 [1.45-9.93]	0.51 [0.04;0.99]	0.76 [0.09;1.44]
n=28 GS; n=34 Dys;	LINC00662	0.009	1.48 [0.55-4.38]	1.09 [0.62;1.56]	0.86 [0.30;1.42]
n=31 GBC	C22orf34	0.04	1.44 [0.48-3.68]	-0.24 [-0.49;0.005]	-0.28 [-0.54;-0.01]
Women	AC084082.3	0.04	8.23 [1.45-9.78]	0.67 [0.18;1.15]	0.89 [0.15;1.63]
n=26 GS; n=20 Dys;	LINC00662	0.01	1.47 [0.54-4.07]	1.09 [0.61;1.56]	1.01 [0.45;1.57]
n=24 GBC	C22orf34	0.02	1.44 [0.48-3.80]	-0.30 [-0.57;-0.03]	-0.34 [-0.63;-0.04]
Men	AC084082.3	0.99	10.01	-0.52 [-1.02;-0.03]	-0.30 [-2.19;1.59]
n=1 GS; n=8 Dys;	LINC00662	0.99	4.53	-0.52 [-1.24;0.21]	-1.09 [-2.85;0.68]
n=6 GBC	C22orf34	0.99	0.49	0.43 [-0.66;1.53]	0.27 [-0.19;0.72]
Age $<$ 60	AC084082.3	0.43	8.23 [1.45-10.19]	0.73 [0.13;1.33]	0.64 [-0.22;1.50]
n=18 GS; n=11 Dys;	LINC00662	0.51	1.81 [0.58-4.33]	0.93 [0.30;1.55]	0.66 [-0.13;1.45]
n=9 GBC	C22orf34	0.58	1.43 [0.47-3.08]	-0.35 [-0.72;0.02]	-0.29 [-0.67;0.09]
Age $\geqslant$ 60	AC084082.3	0.17	8.96 [1.47-9.86]	0.29 [-0.33;0.90]	0.84 [-0.10;1.77]
n=9 GS; n=16 Dys;	LINC00662	0.05	1.46 [0.78-3.84]	1.24 [0.67;1.81]	1.06 [0.36;1.77]
n=18 GBC	C22orf34	0.17	1.46 [0.50-3.44]	-0.18 [-0.52;0.16]	-0.34 [-0.68;0.006]

lncRNA: long non-coding RNA; *FDR*: false discovery rate; *GS*: gallstones; *Dys*: dysplasia; *GBC*: gallbladder cancer; 5th;95th: 5th and 95th percentiles; *CI*: confidence interval. Adapted from Blandino et al. (2022)

Table 3.2: *Identification and validation of cis-long non-coding RNA-expression quantitative trait loci for AC084082.3, $LINC$00662, and $C22orf$34.*

lncRNA	**$\log_2$ expression**	**Location**	**# cis-eQTLs**	**# cis-eQTLs**	**# cis-eQTLs**	**Adjusted r^2**
	Median [5th;95th]	**(GRCh38)**	**(database)**	**(validated)**	**(predictors)**	**(best model)**
AC084082.3	6.59 [1.74;9.06]	chr8:66112667	161	-	-	-
LINC00662	3.40 [0.35;5.60]	chr19:27684580	1576	2	2	0.26
C22orf34	0.58 [0.03;2.65]	chr22:49414524	395	45	3	0.24

lncRNA: long non-coding RNA; 5th;95th: 5th and 95th percentiles; GRCh38: Genome Reference Consortium Human Build 38; *chr*: chromosome ; r^2: r-squared; eQTL: expression quantitative loci. Adapted from Blandino et al. (2022)

Among the 1,576 cis-$LINC$00662-$eQTLs$ identified by the ncRNA-eQTL database, 1,388 met the quality control criteria and were included in subsequent analyses. Two $SNPs$ were associated with the expression of $LINC$00662: rs11083486 (associated in all four penetrance models), and rs142521755 (associated in the dominant model). Both $SNPs$ were not in LD ($r^2 = 0.001$), and the best model for prediction was the one including rs11083486 additively, and rs142521755 with dominant penetrance ($RAIC = 357$).

According to the ncRNA-eQTL database, 396 cis-$lncRNA$-$eQTLs$ were associated with the expression of $C22orf$34. After selection criteria, 45 $SNPs$ were associated with $C22orf$34 in the validation dataset. Most of them (42 cis-$lncRNA$-$eQTLs$) were in LD, resulting in three selected $SNPs$ for prediction. rs5770650 and rs9628049 were selected from both additive and dominant models, while the association with rs6009824 emerged from the three-genotypes model. The best model used for prediction of $C22orf$34 included rs5770650 and rs9628049 additively, and rs6009824 as factor ($RAIC = 214.5$).

The comparisons between predicted and observed expressions of $LINC$00662 and $C22orf$34 are shown in Figure 3.2, panels B and C. More details on all identified and validated cis-$lncRNA$-$eQTLs$ are available in Table 3.2 and Table A.2.

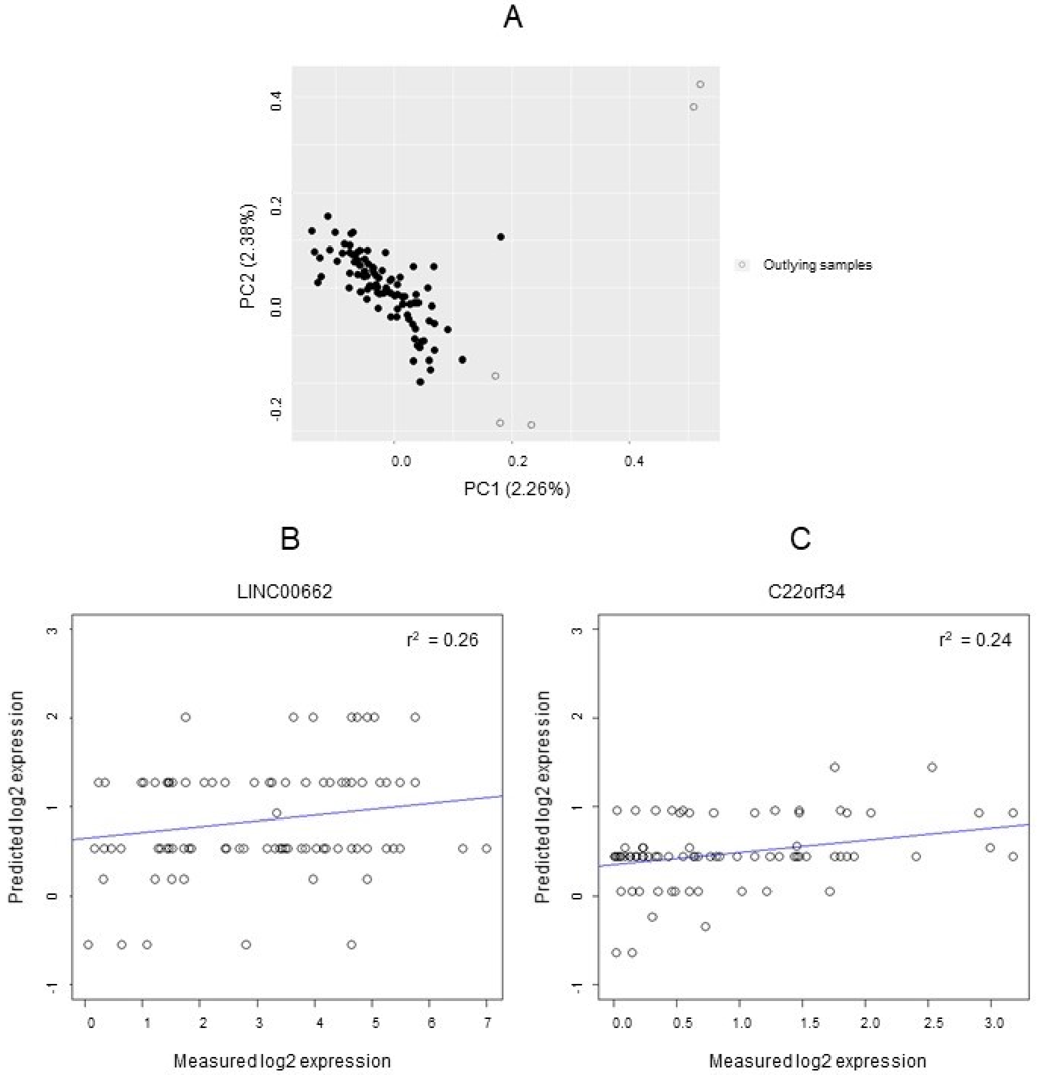

Figure 3.2: *Long non-coding RNA expression quantitative trait loci validation in serum. (A) Genetic principal component analysis in the long non-coding RNA expression quantitative trait loci validation dataset. (B,C) Predicted against measured long non-coding RNA expression for LINC00662 and C22orf34. PC: principal component, r^2: r-squared. Adapted from Blandino et al. (2022)*

3.1.3 Association between genotype-based long non-coding RNA expression and gallbladder cancer risk

The ultimate objective of this study was to identify circulating *lncRNAs* as potential *GBC*-risk biomarkers. Therefore, the final step was to assess the association between the genotype-based *lncRNA* expressions for $LINC00662$ and $C22orf34$ and *GBC* risk. The used dataset (*lncRNA-GBC* association dataset) was larger than the previous ones, being composed by 540 *GBC* cases and 2,397 population-based controls.

The predicted expression of $LINC00662$, consistently with the one observed in tissue, was higher in *GBC* cases compared to population-based controls (Figure 3.3). Most specifically, the risk related with the overexpression of this *lncRNA* was 25% higher in *GBC* cases than in controls ($OR = 1.25$, *p-value* $= 0.02$, Table 3.3). In contrast, although also the genotype-based expression levels of $C22orf34$ were coherent we the ones in tissue, the association with *GBC* risk was not statistically significant ($OR = 0.90$, *p-value* $= 0.59$, Figure A.2, Table 3.3).

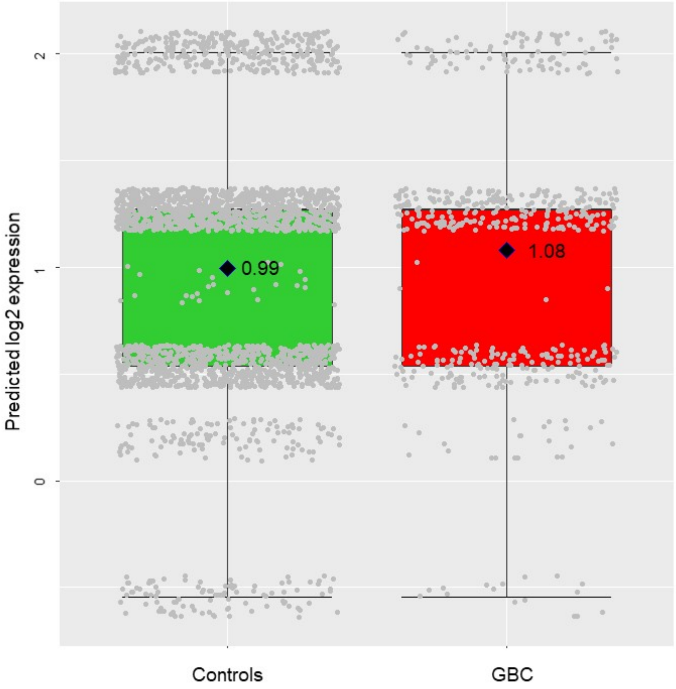

Figure 3.3: *Predicted expression of LINC00662 in the third long non-coding RNA dataset. The average predicted expressions for cases and controls are marked by rhombuses. GBC: gallbladder cancer. Adapted from Blandino et al. (2022)*

Table 3.3: *Predicted expression of LINC00662 and C22orf34, and association with risk of gallbladder cancer in the third long non-coding RNA dataset.*

lncRNA	**Median predicted $\log_2$ expression**	**OR (GBC)**	**95% CI**	**p-value**
LINC00662	1.27	1.25	1.04;1.52	0.02
C22orf34	0.39	0.90	0.61;1.32	0.59

lncRNA: long non-coding RNA; *OR*: odds ratio; *GBC*: gallbladder cancer; *CI*: confidence interval; *p-value*: probability value. Adapted from Blandino et al. (2022)

3.2 Identification and validation of circulating microRNAs associated with gallbladder cancer risk in Europeans

3.2.1 Cohort characteristics

Table 3.4 shows the main characteristics of the investigated datasets in this study. The preselection dataset (*FFPE* gallbladder tissue samples) contained more women (63%) than men (37%), while 44% of patients were older than 71 years. Information on *BMI* and smoking status was not available.

Women were also overrepresented in the screening dataset (*Janus*: 74%, *ESTHER* + *HNR*: 90%), and 50% of *Janus* participants were under 54 years, while 55% of participants in the German *ESTHER* and *HNR* studies were aged 64 to 71 years. Differences in *BMI* were also observed between the Norwegian and the German cohorts: the proportion of individuals with a *BMI* over 26.2 kg/m^2 was 38% in the *Janus* study, compared to 65% in the *ESTHER* and *HNR* cohorts. In terms of number of years between blood collection and *GBC* diagnosis, 63% of *Janus* participants were diagnosed 9 years after blood sampling, while all *ESTHER* and *HNR* participants were diagnosed within 9 years.

In the validation dataset, women were overrepresented in the *HUNT* cohort (85%), but not in *FINRISK* (44%) or *TwinGene* (50%). The proportion of individuals older than 71 years was 23% in *HUNT*, 41% in *FINRISK* and 45% in *TwinGene*. Percentages of participants with a *BMI* over 29.4 kg/m^2 were 28% in *HUNT*, 53% in *FINRISK* and

19% in *TwinGene*. Regarding the time between blood sampling and *GBC* diagnosis, the proportion of participants diagnosed at least 9 years after blood draw was 73% in *HUNT*, 26% in *FINRISK* and 0% in *TwinGene*.

Summing up, all the datasets investigated in this study were heterogeneous in terms of age, sex, *BMI* and time from blood collection to *GBC* diagnosis.

3.2.2 Preselection in tissue

Among the 2,006 *miRNAs* detected in *FFPE* gallbladder tissue, 1,300 showed low expression variability ($MAD < 0.2$) and were excluded from further analysis. A *PCA* plot based on the remaining 706 *miRNAs* revealed different global expression profiles in *GBC* and normal gallbladder tissue samples, with the first principal component explaining 19% of the overall variance in *miRNA* expression (Figure 3.4, panel A). *P-values* from robust linear regression adjusted for multiplicity using the *Bonferroni* method identified 376 *miRNAs* differentially expressed in *GBC* compared to normal gallbladder tissue (Figure 3.4, panel C, Table A.3). In particular, 215 *miRNAs* were overexpressed, and 161 *miRNAs* were underexpressed in *GBC* tissue.

3.2.3 Screening in serum samples

Figure 3.4, panel B shows the global expression profiles based on *MAD*-positive *miRNAs* in the screening dataset. In contrast to the preselection dataset, which included gallbladder tissue samples, *GBC* cases and controls showed similar global *miRNA* expression patterns in serum. Among the 376 preselected candidates, 186 *miRNAs* were also detectable in serum (Figure 3.4, panel D). Four *miRNAs* associated with potential confounders in previous research were excluded from further analysis (*miR*-320*d*, *miR*-4466, *miR*-4516, *miR*-4755-3*p*). After robust linear regression analysis and multiplicity correction, three *miRNAs* were associated with *GBC* risk. *miR*-3925-5*p* showed a protective effect, while *miR*-4533 and *miR*-671-5*p* were associated with an increased risk of *GBC*. However, only *miR*-4533 and *miR*-671-5*p* showed consistent expression differences in gallbladder tissue and serum samples. *miR*-3925-5*p* was underexpressed in *GBC* tissue but overexpressed in serum samples from *GBC* cases and was therefore excluded from further analyses.

Table 3.4: *Main patient characteristics in the investigated microRNA datasets. The preselection dataset consisted of formalin-fixed paraffin-embedded gallbladder tissue samples from gallbladder cancer and gallstone disease patients recruited in Germany. The screening dataset included serum samples from three European prospective cohorts (Janus in Norway, ESTHER and HNR in Germany). The validation dataset comprised serum samples from three prospective cohorts (HUNT in Norway, FINRISK in Finland, and TwinGene in Sweden).*

		Preselection		**Screening**				**Validation**					
Variable	**Level**	**Hei**		**Janus**		**ESTHER+HNR**		**HUNT**		**FINRISK**		**TwinGene**	
		n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Status	GBC	40	0.83	27	0.50	10	0.50	15	0.50	8	0.47	8	0.40
	Controls	8	0.17	27	0.50	10	0.50	15	0.50	9	0.53	12	0.60
Age	Q1: 25-54	8	0.17	27	0.50	1	0.05	5	0.17	3	0.18	0	0
	Q2: 54-64	10	0.21	21	0.39	6	0.30	6	0.20	2	0.12	0	0
	Q3: 64-71	9	0.19	1	0.02	11	0.55	12	0.40	5	0.29	11	0.55
	Q4: 71-89	21	0.44	5	0.09	2	0.10	7	0.23	7	0.41	9	0.45
Sex	Female	30	0.63	40	0.74	18	0.90	24	0.85	7	0.44	10	0.50
	Male	18	0.37	14	0.26	2	0.10	4	0.14	9	0.56	10	0.50
BMI	Q1: 18.1-23.3	-	-	18	0.35	4	0.20	8	0.28	1	0.06	3	0.19
	Q2: 23.3-26.2	-	-	14	0.27	3	0.15	7	0.24	3	0.18	6	0.38
	Q3: 26.2-29.4	-	-	12	0.23	7	0.35	6	0.21	4	0.24	4	0.25
	Q4: 29.4-45.9	-	-	8	0.15	6	0.30	8	0.28	9	0.53	3	0.19
Smoking	Never	-	-	16	0.31	8	0.57	11	0.42	6	0.38	-	-
	Former	-	-	15	0.28	4	0.29	9	0.34	7	0.44	-	-
	Current	-	-	22	0.41	2	0.14	6	0.23	3	0.18	-	-
Follow-up	Q1: 0-3.5	-	-	4	0.15	5	0.50	0	0	4	0.50	4	0.50
	Q2: 3.5-9	-	-	6	0.22	5	0.50	3	0.20	2	0.25	4	0.50
	Q3: 9-12.5	-	-	7	0.26	0	0	5	0.33	1	0.13	0	0
	Q4: 12.5-18	-	-	10	0.37	0	0	6	0.40	1	0.13	0	0

GBC: gallbladder cancer; *BMI*: body-mass index; $Q1$ - $Q4$: first to fourth quartiles; ESTHER: Early Detection and Optimised Therapy of Chronic Diseases in the Elderly Population; HNR: Heinz Nixdorf Recall study; HUNT: Helseundersøkelsen i Nord-Trøndelag Health study.

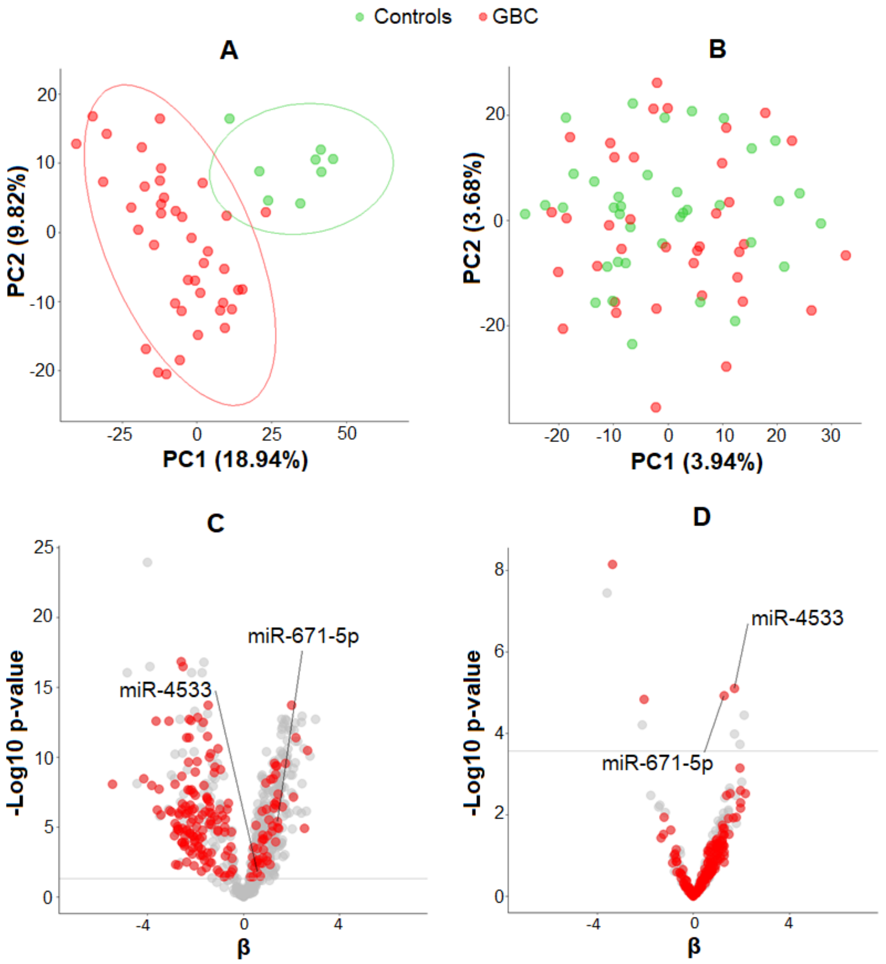

Figure 3.4: *Exploratory analysis of global microRNA expression profiles in the preselection and screening datasets. Principal component analysis plots of normalized log*$_2$ *expression counts for microRNAs in the preselection (A) and screening (B) datasets. The x-axis shows the first principal component and its explained variance in global microRNA expression; the y-axis shows the same information for the second principal component. Volcano plots for microRNAs in the preselection (C) and screening (D) datasets. The x-axis shows the estimated average expression difference, and the y-axis shows the* $-log10$ *probability value from robust linear regression. Red dots represent microRNAs expressed in both formalin-fixed paraffin-embedded gallbladder tissue and serum samples, and the grey horizontal lines show the statistical significance threshold (multiplicity-corrected Bonferroni probability value < 0.05). GBC: gallbladder cancer; p-value: probability value; PC: principal component.*

3.2.4 Validation in serum samples

Visual inspection of the global $miRNA$ expression profiles in the validation dataset revealed the presence of three outlying samples, which were excluded from further analyses based on statistical depth (Figure A.3), resulting in 31 GBC cases and 36 controls ultimately used for validation. Robust linear regression detected no association between the two $miRNAs$ identified in the screening dataset and GBC risk (Table A.4), but stratified analyses confirmed overexpression of miR-4533 in prospective serum samples from GBC cases in the $HUNT$ cohort, especially in individuals with a BMI below 26.2 kg/m^2, and with an increased genetic susceptibility to gallstones. miR-671-5p showed low overall expression in the validation dataset (Figure 3.6, panel C).

3.2.5 Meta-analysis

Both fixed-effect and random-effect meta-analyses suggested that miR-4533 expression is associated with an increased risk of GBC (Figure 3.6, panel B), but no association emerged for miR-671-5p (Figure 3.6, panel D). Table 3.5 shows the overall and stratified results from robust linear regression models for the two candidates considering simultaneously all prospective cohorts investigated. Results adjusted for age, sex and BMI confirmed the increased expression of miR-4533 in prospective serum samples of GBC patients, particularly in individuals younger than 63.5 years, or with a BMI below 26.2 kg/m^2.

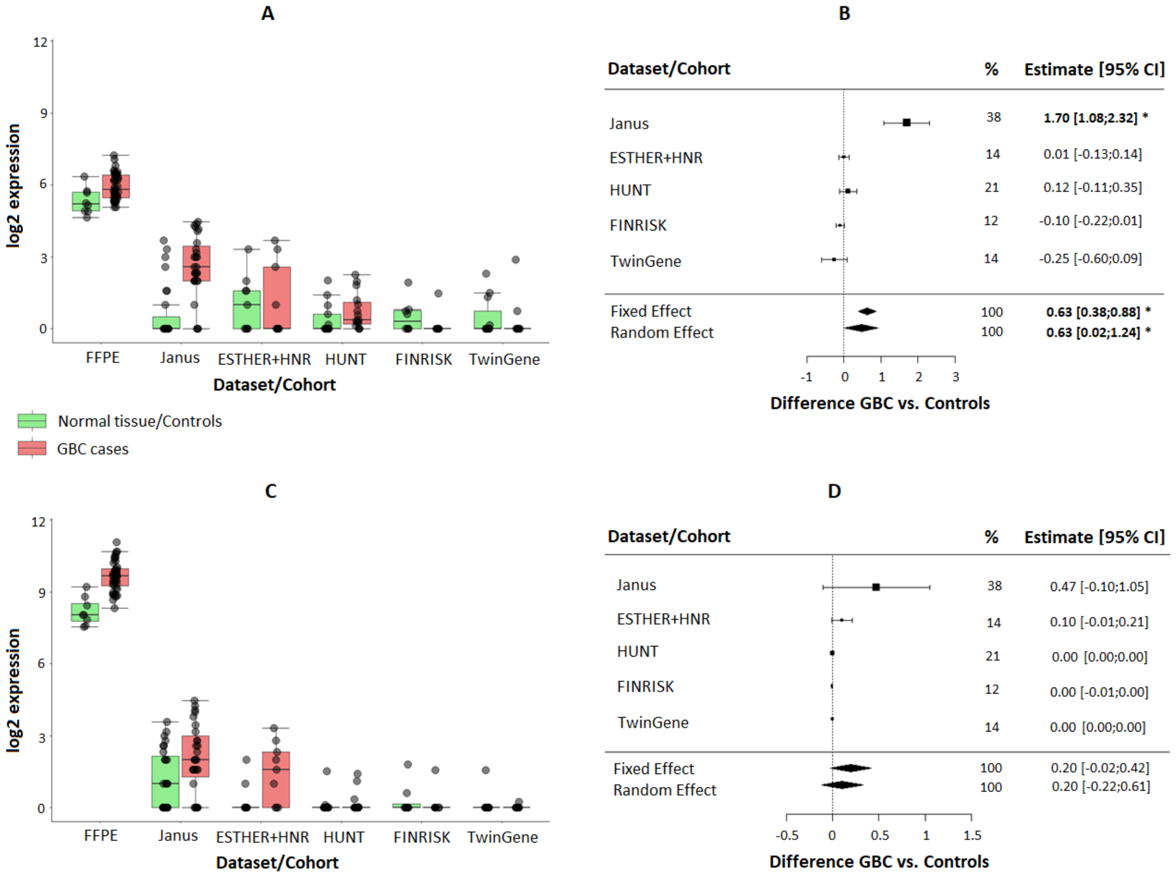

Figure 3.5: *Expression of miR-4533 and miR-671-5p in formalin-fixed paraffin-embedded gallbladder tissue and serum samples, and meta-analysis results. (A,C) Dot-and-box plots of $\log_2$ miR-4533 and miR-671-5p expression in the preselection dataset and in the five investigated prospective cohorts. (B,D) Forest plots and combined average differences in serum expression between gallbladder cancer cases and controls from fixed and random effects meta-analysis for miR-4533 and miR-671-5p. FFPE: formalin-fixed paraffin-embedded; CI: confidence interval; GBC: gallbladder cancer; ESTHER: Early Detection and Optimised Therapy of Chronic Diseases in the Elderly Population; HNR: Heinz Nixdorf Recall study; HUNT: Helseundersøkelsen i Nord-Trøndelag Health study.*

Table 3.5: *Overall and stratified differences in miR-4533 expression by age, sex, body-mass index and genetic susceptibility to gallstone disease between prospective gallbladder cancer cases and controls.*

Variable	**Level**	**$\log_2$ expression in controls**	**GBC Case-Control**
		Median [5th;95th]	**Difference [95% CI]**
All	-	0.00 [0.00; 2.21]	**0.43** [0.17; 0.69]
Age	$<$ 63.5 years	0.00 [0.00; 2.42]	**1.17** [0.63; 1.71]
	$\geqslant$ 63.5 years	0.00 [0.00; 1.76]	0.01 [-0.07; 0.09]
Sex	Female	0.00 [0.00; 1.99]	**0.42** [0.14; 0.70]
	Male	0.02 [0.00; 2.14]	0.32 [-0.21; 0.85]
BMI	$<$ 26.2 kg/m^2	0.00 [0.00; 1.89]	**0.83** [0.42; 1.24]
	$\geqslant$ 26.2 kg/m^2	0.00 [0.00; 1.97]	0.14 [-0.06; 0.34]
GSD-PRS	$<$ 2.88	0.00 [0.00; 1.36]	0.07 [-0.17; 0.31]
	$\geqslant$ 2.88	0.01 [0.00; 1.15]	-0.15 [-0.37; 0.05]

GBC: Gallbladder cancer; $5th; 95th$: 5th and 95th percentiles; CI: Confidence interval; BMI: body-mass index; *GSD-PRS*: Polygenic risk score for gallstone disease.
Bold type indicate that the 95% confidence interval does not include zero.

3.2.6 Pathway analyses

Pathway analyses using the DIANA mirPath software indicated that *miR*-4533 is involved in the regulation of multiple cancer pathways. Sixty-five KEGG biological processes were significantly enriched (FDR-corrected *p-value* $<$ 0.05). The top five pathways involving *miR*-4533 were related to proteoglycans in cancer, renal cell carcinoma, glioma, *ErbB* signaling, and *Rap*1 signaling. These five pathways included 510 genes in total, but some of them belonged to several pathways and others were not expressed in our investigated serum samples, resulting in 308 genes examined in the $miRNA$-$mRNA$ correlation analyses.

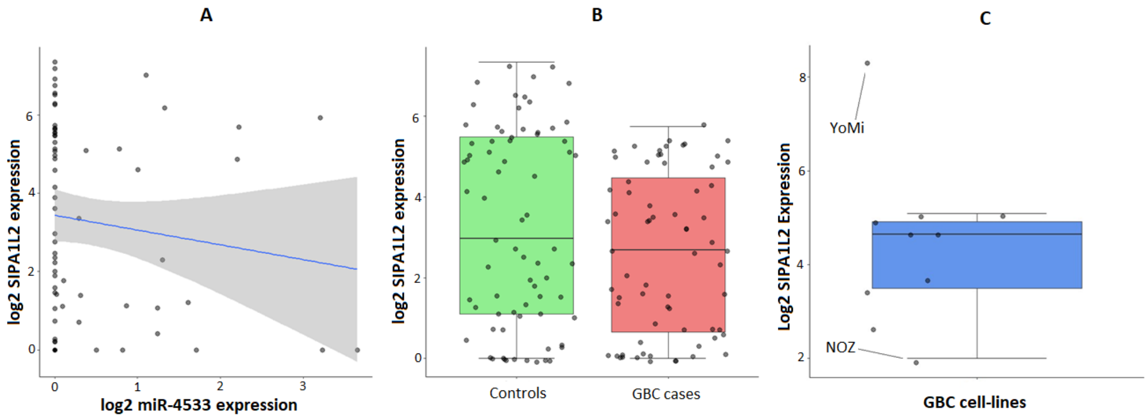

Figure 3.6: *miR-4533 and SIPA1L2 expression in serum samples and gallbladder cancer cell lines. (A) Scatterplot of* $\log_2$ *miR-4533 vs SIPA1L2 expression in serum samples from control subjects. (B)* $\log_2$ *SIPA1L2 expression in serum samples from control subjects (green) and gallbladder cancer cases (red) in the five prospective cohorts investigated. (C)* $\log_2$ *SIPA1L2 expression in ten gallbladder cancer cell-lines. NOZ and YoMi were the two cell-lines with the lowest and highest* $\log_2$ *SIPA1L2 expression, respectively.* $SIPA1L2$*: Signal Induced Proliferation Associated 1 Like 2 gene; GBC: gallbladder cancer.*

Table 3.6 shows the results for the ten genes most negatively and strongly correlated with miR-4533 expression. Among them, only $SIPA1L2$ (Signal Induced Proliferation Associated 1 Like 2 gene) and FAS (Fas Cell Surface Death Receptor) were associated with GBC risk. However, FAS was overexpressed in serum samples from prospective GBC cases, and the focus is therefore only on $SIPA1L2$ ($Spearman\ rho$ correlation -0.247, average GBC case-control expression difference -0.60). Figure 3.6, panel A depicts the negative relationship between miR-4533 and $SIPA1L2$ expression in the investigated prospective serum samples from control participants. Figure 3.6, panel B shows that $SIPA1L2$ is downregulated in serum samples from GBC cases compared to control. $SIPA1L2$ was expressed in ten GBC cell-lines, showing its highest expression in $YoMi$ (Figure 3.6, panel C).

3.2.7 Literature review

Without attempting an exhaustive review of the literature, the expression in serum samples of 34 $miRNAs$ previously associated with GBC was examined. Most studies (74%) were

Table 3.6: *List of the top ten genes with expression values most negatively correlated with miR-4533 expression in the 5 pathways with the smallest false-discovery-rate-corrected probability values.*

	Spearman rho correlation			*Difference Cases vs. controls*		
Gene	**Estimate**	**95% CI**	**p-value**	**Estimate**	**95% CI**	**p-value**
FLT4	-0.268	[-0.48; -0.05]	0.01	0.23	[-0.10; 0.56]	0.17
RAP1A	-0.262	[-0.51; -0.01]	0.01	-0.15	[-0.47; 0.17]	0.35
FGF7	-0.248	[-0.45; -0.03]	0.02	-0.01	[-0.20; 0.18]	0.93
SIPA1L2	-0.247	[-0.48; -0.02]	0.02	-0.60	[-1.18; -0.01]	0.04
ARNT2	-0.245	[-0.45; -0.02]	0.02	0.07	[-0.11; 0.27]	0.53
ITGAM	-0.189	[-0.39; 0.03]	0.05	0.00	[-0.16; 0.33]	0.97
MAPK9	-0.189	[-0.39; 0.02]	0.06	0.18	[-0.03; 0.67]	0.27
RAPGEF1	-0.187	[-0.41; 0.06]	0.06	-0.28	[-0.50; 0.19]	0.09
RAPGEF5	-0.187	[-0.42; 0.06]	0.06	0.12	[-0.10; 0.39]	0.41
FAS	-0.179	[-0.39; 0.06]	0.07	0.44	[0.16; 0.76]	0.01

GBC: gallbladder cancer; *CI*: confidence interval; *p-value*: probability value.

conducted in India, followed by China (24%), and all but one study investigated gallbladder tissue samples (Table A.7). Of the 34 *miRNAs*, eight showed an association between their serum expression levels and *GBC* risk (*miR*-145-5*p*, *miR*-144-5*p*, *miR*-196*a*-5*p*, *miR*-196*b*-5*p*, *miR*-32-5*p*, *miR*-3613-5*p*, *miR*-374*a*-5*p*, *miR*-378*c*). The expression of three *miRNAs* in serum (*miR*-144-5*p*, *miR*-145-5*p* in the Indian study (but not in the single European study) and *miR*-378*c*) was consistent with previous reports, where *miR*-144-5*p* and *miR*-145-5*p* were overexpressed in serum and gallbladder tissue of *GBC* patients, and *miR*-378*c* was downregulated in both types of samples.

Chapter 4

Discussion

Comment: Parts of the following Chapter have already been published in Cancers (Blandino et al., 2022). The original manuscript was written by myself, but also contains comments and corrections from the co-authors.

This chapter summarizes the contributions of this thesis to research. Additionally, the limitations of the current study are discussed, along with proposed directions for future research.

4.1 Contributions to research and limitations

GBC, the sixth most common gastrointestinal cancer globally, is one of the most prevalent forms of biliary tract cancer (Bray et al., 2024). *GBC* is highly aggressive and is usually diagnosed at advanced stages, making treatment strategies largely ineffective, and treatment options very limited (Wistuba and Gazdar, 2004). The geographic distribution of *GBC* varies significantly, with low prevalence in high-income countries, while low- and middle-income regions, particularly Latin America, experience much higher incidence rates (Bray et al., 2024). *GBC* is also strongly associated to both environmental and genetic factors. Modification of these determinants may offer great potential in preventing the development of this aggressive disease (Kanthan et al., 2015). Therefore, the prognosis of *GBC* patients could greatly improve with the adoption of primary and secondary prevention strategies, helping prevent tumor spread to adjacent organs.

A primary challenge in advancing effective management options for *GBC* patients has been the discovery of innovative diagnostic and prognostic biomarkers. *ncRNAs*, in particular, have demonstrated, through their regulatory role in many important biological processes, significant potential as biomarkers for cancer risk assessment and early detection (Anastasiadou et al., 2018). Additionally, while tissue biopsy provides direct insights into the local tumor micro environment, it is an invasive procedure (Armakolas et al., 2023). Serum circulating *ncRNAs*, on another hand, offer a less invasive method for evaluating cancer progression as they are easily accessible and very stable even under extreme temperatures and long-term storage (Glinge et al., 2017).

This thesis aims to identify circulating *ncRNAs* as potential biomarkers for *GBC*. The applied methodology focuses on the detection of *ncRNAs* that exhibit consistent expression levels in both tissue and serum, thereby enhancing their potential as reliable biomarkers. The first part of this thesis examines the role of *lncRNAs* in *GBC* progression among Chilean individuals, while the second part focuses on analyzing the expression patterns of *miRNAs* in European *GBC* cases.

The link between *ncRNA* expression and the development of *GBC* has been investigated to some extent, although findings remain inconsistent. This inconsistency can be attributed, partially, to the heterogeneity of the studied populations, but primarily to the low number of conducted studies and their limited sample sizes. Given the high prevalence of *GBC* in Asian regions, the majority of research on *ncRNAs* and *GBC* has been conducted in India, followed by China. In contrast, there has been limited research focusing on European or Latin American populations, especially considering the high incidence rate of *GBC* in Latin America. Additionally, the majority of existing studies have focused solely on analyzing the expression of specific *ncRNAs* in gallbladder tissue samples, a method that is both invasive and costly, as previously mentioned. A study by Saxena et al. (2023) involving five Indian patients identified 19 upregulated and 29 downregulated *miRNAs* in *GBC* tissue. Among the identified candidates, *miR*-145-5*p* exhibits an oncogenic role, which diverges with the findings from another study on 48 German patients, where *miR*-145-5*p* acts as tumor suppressor through the activation of *STAT*1 (Goeppert et al., 2019). The expression of *miR*-145-5*p* shows similar patterns as the ones observed in the aforementioned study also in the European prospective serum samples used for the thesis purpose. Other discrepancies between studies involving diverse populations have also been observed for *miR*-122. A study conducted in Chinese subjects demonstrated the antitumor effects of *miR*-122 in 20 *GBC* cell lines (Lu et al., 2015). Conversely, Li et al. (2015) found that elevated expression of *miR*-122 in *GBC* tissue is associated with a worse prognosis.

According to the results provided by this thesis, three specific *lncRNAs* show a progressive alteration in tissue expression from Chilean patients throughout the sequence from gallstones disease, to dysplasia, and ultimately to *GBC*: *AC*084082.3, *C*22*orf*34, and *LINC*00662. The expression levels of *AC*084082.3 and *LINC*00662 increase as malignancy progresses, while *C*22*orf*34 displays a decreasing trend in expression from gallstone disease to *GBC*. Additionally, the validation of *lncRNA*-*eQTLs* identified two cis-*eQTLs* linked to the expression of *LINC*00662 and three associated with *C*22*orf*34, which were then used for *lncRNA* expression prediction. Association analyses reveal that, consistent with tissue expression measurements, the genotype-based expression of *LINC*00662 is associated with a 25% increased risk of developing *GBC*.

Currently, there is insufficient evidence to determine the role of *AC*084082.3 in cancer. A study on endometriosis associated ovarian cancer indicates that this *lncRNA* is underexpressed in cancer patients compared to controls (Finall et al., 2023). In alignment with this thesis' findings, some studies have reported the potential tumor-suppressive relevance of *C*22*orf*34 in cancer biology. A recent study on renal cell carcinoma suggests that *C*22*orf*34 is under expressed in cancerous tissues compared to normal controls, with elevated levels of this *lncRNA* correlating with improved overall survival rates (Yang et al., 2024). Another study reveals that higher expression levels of *C*22*orf*34 are associated with a reduced risk of death in patients with cutaneous melanoma (Tang et al., 2022). Similar results to those presented in this thesis have also been observed in literature in relation to *LINC*00662. According to numerous studies on the respiratory, reproductive, nervous, and digestive systems, *LINC*00662 plays a significant oncogenic role by enhancing cell invasion (Xia et al., 2020; Gong et al., 2018; Lv et al., 2021). Research conducted in gastric cancer and hepatocellular carcinoma shows that the overexpression of *LINC*00662 is also closely associated with poor patient prognosis and reduced chemo sensitivity (He et al., 2021; Tian et al., 2020; Guo et al., 2020). Mechanistically, $LINC00662$ serves as $ceRNA$ for gene regulation and influences RNA metabolism, regulating $mRNA$ stability, and participating in numerous essential signaling pathways, including the $MAPK/ERK$ pathway.

This thesis also highlights that, in European subjects, miR-4533 and miR-671-5p show consistent expression differences in gallbladder tissue and serum samples, both playing an oncogenic role. However, only miR-4533 was validated through meta-analysis. Notably, miR-4533 overexpression is especially marked in individuals under the age of 63.5 years and those with a BMI below 26.2 kg/m^2. Pathway analysis revealed that miR-4533 is implicated in several cancer-related pathways, including proteoglycans in cancer, renal cell carcinoma, glioma, $ErbB$ and Rap1 signaling. Furthermore, a negative correlation between the expression of $SIPA1L2$ and miR-4533 suggests that $SIPA1L2$ serves as target gene of miR-4533, being underexpressed in serum samples from GBC cases.

There is little emerging evidence in the literature regarding the role of miR-4533 as a potential disease biomarker. One study demonstrated that miR-4533, through its interaction with $ABLIM1$, is involved in the regulation of intervertebral disc degeneration progression (Xie et al., 2024). A research investigating $miRNA$ in colorectal cancer, reported the overexpression of miR-4533 in the colorectal mucosa across individuals of various ancestries, including Hispanic, and Asian (Slattery et al., 2017). Additionally, elevated levels of miR-4533 have been observed in breast and prostate cancer, further supporting its potential role as an oncogenic biomarker (Lai et al., 2019). Mechanistically, miR-4533 may not be a canonical $miRNA$, since it does not have a hairpin loop and is probably dicer independent. Nevertheless, miR-4533 is listed in the miRbase database, and results from this thesis, combined with literature review, suggest that it is a potential serum biomarker for GBC (Griffiths-Jones et al., 2008). Dysregulation of miR-671-5p has also been observed in numerous cancers. miR-671-5p increases cell proliferation, invasion, and migration in hepatocellular carcinoma by targeting $ALDH2$ (Chen et al., 2022). Furthermore, the overexpression of miR-671-5p has also been associated with poor prognosis in colorectal cancer (Jin et al., 2019). Research on renal cell carcinoma unraveled the $HMGA1$-mediated role of miR-671-5p, which promotes metastasis through targeting of APC (Chi et al., 2020). The role of the $SIPA1L2$ gene in the tumor environment has been widely studied. According to The Human Protein Atlas,

$SIPA1L2$ is a biomarker for renal cancer (Uhlén et al., 2015). $SIPA1L2$ expression has also been associated with an unfavorable prognosis and poor survival in intestinal-type gastric and colorectal cancer patients (Zhang et al., 2018; Rahman et al., 2020). In contrast, a distinct pattern has been observed in hepatocellular carcinoma, characterized by the upregulation of $SIPA1L2$ expression (Ma et al., 2020). Among the key pathways involving the preselected *miRNAs* in this thesis, the *ErbB* signaling pathway emerges as the most frequently mutated in *GBC*, affecting 36.8% of *GBC* cases (Sicklick et al., 2016). The proteoglycans in cancer pathway plays a significant role in *GBC* progression too. A recent study demonstrated that this pathway is crucial for the progression of gallbladder inflammatory lesions to invasive cancer (Rawal et al., 2023). According to existing literature, several other pathways not identified in this thesis are also implicated in *GBC* pathogenesis, suggesting that the disease's molecular mechanisms may be more complex than those explored in this work. Some examples are the $PI3K/AKT/mTOR$ pathway, the hepatocyte growth factor, amphiregulin, and insulin-like growth factor 1 receptor (Sinkala, 2023; Cheng et al., 2022). Future follow-up research to this thesis could study these candidates in other populations and explore their potential involvement in other cancer-related pathways.

To address population and specimen bias, this thesis develops a multiple-stage approach for identifying circulating *ncRNAs* associated with *GBC* progression. The applied framework, which represents one of the novelties of this thesis, is based on the integration of data from diverse datasets, ancestry, and both tissue and serum biomarkers. Therefore, the identified biomarkers do not only rely on a single cohort or sample type. *ncRNAs* that do not demonstrate consistent directional expression across different datasets are excluded, thereby enhancing the selection of viable candidates and ensuring robustness of the results. The first study examined in this thesis on *lncRNAs* and *GBC* in Chileans, for example, consists of three independent datasets: the preselection dataset includes exclusively *lncRNA* expression from 98 *FFPE* tissue samples; the cis-*lncRNA*-*eQTL* validation dataset encompasses both *lncRNA* expression and genotype information over 110 serum samples; and the *lncRNA*-*GBC* association dataset contains genotype information alone for 540 *GBC* cases and 2397 population-based controls. The study design applied for the *miRNA*-*GBC* study in Europeans follows a similar structure, through *miRNA* preselection in 48 *FFPE* tissue samples, screening of *miRNA* expression differences in 72 prospective serum samples, and subsequent *miRNA* validation in additional 67 prospective serum samples. A limitation of the *ncRNA* identification framework applied in this thesis is the reduced number of identified *ncRNAs* compared to one-stage designs. In one-stage designs, the chances of of discovering new biomarkers are higher, since the selection of potential candidates does not undergo multiple filtering steps, which may decrease the chances of success. However, this streamlined approach raises concerns about the reliability of the findings, which, as discussed earlier in this chapter, may be limited to a specific population or specimen, potentially lacking broader applicability. Like the proposed design presented in this thesis, the development of adaptive study designs that optimize both cost-efficiency and time, while still effectively identifying biomarkers, presents a promising avenue for future research. These designs could strike a balance between comprehensive biomarker discovery and the need for external validation. One important strength of the applied design is the registration of the *miRNA* validation after preselection and screening on the German (drks.de, March, 5th 2021) and the International Clinical Trials Registry Platform of the World Health Organization (WHO, https://trialsearch.who.int/Trial2.aspx?TrialID=DRKS00024573).

Although the two studies conducted in this thesis are relatively large given the rarity of *GBC*, especially in Europe, the small sample sizes in the used datasets still represent a limitation in terms of the robustness and generalization of the findings. As a result, in the *lncRNA* study, only six *lncRNAs* out of the 332 screened were preselected. Although 2,137 instrumental variables were identified from the *eQTL*-database, only five associations could be validated in the *lncRNA*-*eQTL* validation dataset, which is likely due to the small sample size. As for the second study, even after combining data and samples from large European cohorts and conducting the largest prospective study to date, the sample size is still relatively small. Moreover, the heterogeneity of the prospective studied cohorts (diverging in terms of age, sex, *BMI*, and time from blood retrieval to *GBC* diagnosis) translates into a good representativeness of the results, but on the other hand, *miRNA* expression differences that are population-specific have been likely overlooked.

Sample size and power analyses are well-established methodologies in traditional biological studies, including *GWAS* and microarray gene expression studies (Uffelmann et al., 2021). These tools help ensure adequate statistical power to detect significant associations, guiding researchers in optimizing study design and interpretation of findings. Sample size and statistical power are heavily influenced by several key factors, including the number of comparisons (and methods used to account for multiple testing), biological variability within the data, data dispersion, the underlying distribution of the data, and the available budget. Careful consideration of these elements is essential for ensuring valid and reproducible results. To date, while tools like *RnaSeqSampleSize* are available to estimate the optimal sample size for differential gene expression analysis, there are no established tools specifically designed for sample size calculation in *ncRNA* studies (Zhao et al., 2018). This gap presents a significant challenge in ensuring adequate statistical power for *ncRNA* research. In an effort to determine the optimal sample size for further investigating the two *miRNAs* identified in this thesis, the R tool for sample size estimation, `pwr.t.test`, has been utilized (Bartlett and Charles, 2022). The obtained results indicate that to attain a statistical power of 0.80, 51 case-control pairs would be required to adequately detect *miR*-4533 (effect size: 0.62). Conversely, a considerably larger sample size of 531 case-control pairs is necessary to validate *miR*-671-5*p* (effect size: 0.19). These findings suggest that the sample sizes utilized in this study are likely sufficient for *miR*-4533, which shows clear expression differences between cases and controls, but are not enough for *miR*-671-5*p*. Therefore, further validation in additional cohorts is needed to confirm the utility and accuracy of the identified *lncRNAs* and *miRNAs* as serum biomarkers. Follow-up studies that include a larger number of study participants are necessary to identify and validate a higher number of *ncRNAs*, and more accurate estimates of individual *GBC* risk. Larger cohorts would provide more statistical power, improving the reliability and broader applicability of the identified biomarkers.

A review of the existing literature on *GBC* biomarker studies indicates that inadequate sample sizes are a widespread challenge, especially in studies which involve serum samples. For example, a study by Srivastava et al. (2023) examined only 34 paired serum samples, identifying five potential *miRNA* candidates. However, it is fundamental to report that their identification was based on *p-values* that were not adjusted for multiple comparisons. This implies that applying multiplicity corrections would have resulted in the exclusion of all *miRNAs* from the previous selection. A somewhat larger study included 85 *GBC* tissue samples alongside 11 normal gallbladder mucosas (Chang et al., 2013). A study conducted by Ma et al. (2015) on *lncRNAs* validated the oncogenic role of *CCAT*1, utilizing only 40 *GBC*-

control pairs in tissue samples. Similarly, Li et al. (2015) attempted to identify differentially expressed *miRNA* in blood, based on 40 peripheral blood samples. Additionally, Xue et al. (2019) incorporated only 58 tissue samples in total. Overall, these findings underscore that the two studies presented in this thesis show the largest sample sizes reported to date in *GBC* research, particularly in relation to serum analyses.

Regarding the unsuccessful validation of the cis-*eQTL* associations in the first study of this thesis, it is important to briefly discuss on the prevalence of genetic association studies on individuals of European descent. As of September 2023, most of the 6,574 publications and 552,954 associations included in the *GWAS* catalog are based on European studies (Sollis et al., 2023). The absence of *GWAS* data for populations outside of European ancestry is, therefore, a notable concern. This gap is particularly evident in African populations, whose unique haplotypic structures are well-suited to enable targeted genetic discoveries (The International HapMap Consortium, 2007). The situation is even more complicated for research on Latin Americans, as these populations are characterized by admixture primarily involving African, Native American, and European ancestries. Notably, only 1.3% of both discovery and replication studies have been conducted within these populations, and proportions of Native American ancestry are not taken into account (Bryc et al., 2015; Mills and Rahal, 2019). In this context, a Japanese study examined the association between prostate cancer and 23 *SNPs* that had been previously identified through *GWAS* on heterogeneous populations (Yamada et al., 2009). 16 *SNPs* emerged from studies on Europeans, two from Africans and five from diverse populations. The findings of this study revealed that only seven out of the 23 *SNPs* are linked with prostate cancer risk in the Japanese population, while the remaining 16 *SNPs* show no association or, as in five *SNPs*, opposite point estimates compared to what had been previously reported. Comparable considerations can be extended to type 2 diabetes, asthma, and cardiovascular diseases, conditions that are highly prevalent among Latin American populations, but which have been predominantly studied in European populations (Aguayo-Mazzucato et al., 2019; Maldonado et al., 2023). Therefore, catalogs should take this bias into account to ensure that population-specific variants are not overlooked.

A key strength and originality of this research lies not only in the investigated hypotheses, but in the approach used to identify disease effects in tissue by leveraging omics data. Based on current knowledge, the two studies presented in this thesis are the first to identify differentially expressed *ncRNAs* in *GBC* by combining both tissue samples and RNA sequencing data. While a moderate number of studies have explored the link between tissue and serum biomarkers, the existing literature on this topic in the context of *GBC* is still insufficient. In relation to breast cancer, a study by Karimi et al. (2020) underscored the importance of circulating biomarkers, demonstrating that key markers such as $CEA(O)$, $CK19$, ER, and *c-Myc* are detectable exclusively in blood samples and not in tissue samples. The concordance between tissue and plasma markers was also investigated in a study on lung adenocarcinoma, revealing that CA 19-9 and $CYFRA21$-1 exhibit same expression patterns in both tissue and serum samples (Jiao et al., 2021). Another study on non-alcoholic fatty liver disease, has also attempted to explore the connection between tissue and transcriptomics data using direct serum protein measurements to identify noninvasive biomarkers (Darci-Maher et al., 2023). A recent study on metastatic testicular cancer found that six *miRNAs* hold high sensitivity (96%) and specificity (78%) for cancer detection in serum samples, whereas their specificity in tissue is notably low (Ujfaludi et al., 2024). In the context of circulating metabolites, Cao et al. (2021) identified 17 metabolites that exhibited consistent expression alterations in pancreatic ductal adenocarcinoma compared to controls both in tissue and serum samples. In conclusion, the overexpression of *miR*-4533 and $LINC00662$ in both tissue and serum suggests their potential utility as diagnostic biomarkers for *GBC*. However, these encouraging findings require validation and further refinement in future studies, particularly concerning their applicability to other sample types (e.g., whole blood and plasma).

In this thesis, various statistical methods are tested and compared to identify differentially expressed *ncRNAs* in *GBC*. Preselection of *lncRNAs* exhibiting monotonically increasing or decreasing expression levels from gallstones to *GBC* relies on both the non-parametric two-sided *J-T* test, and *ML XGBoost* algorithm, used to train three-class classification *ML* models (Jonckheere, 1954; Chen and Guestrin, 2016). On the other hand, the preselection, screening and validation of differentially expressed *miRNAs* are conducted using robust linear regression, with validation further reinforced by metanalysis. The majority of existing literature indicates a general preference of research for methodologies such as the R package

DESeq2 or standard linear regression for the identification of differentially expressed *RNAs* (Love et al., 2014). Both methods hold certain advantages, but they are also accompanied by inherent limitations. Li et al. (2022) examined the performance of the DESeq2 package, specifically evaluating its propensity to generate false positives. Interestingly, their findings revealed that DESeq2 erroneously classifies 15.3% of cases as false positives. Soneson and Delorenzi (2013) reported that DESeq2 demonstrates effective performance, particularly with smaller sample sizes (Soneson and Delorenzi, 2013). However, they also noted that DESeq2 tends to yield an excess of large *p-values* and is associated with a lower number of true positives compared to other methodologies. Conversely, linear regression is a widely employed tool in clinical practice for assessing the relationship between disease status and the expression of specific molecular phenotypes, while effectively adjusting for potential confounders, thereby enhancing the accuracy of estimates and reducing bias. The efficacy of standard linear regression is compromised in the context of *RNA*-Seq data, where distributions are frequently skewed, and outlying observations are prevalent (Kvam et al., 2012). In fact, estimates from standard linear regression are heavily influenced by the presence of these divergent observations (Alanamu et al., 2023). This results in a loss of valuable information and a reduction in statistical power. Robust regression, in contrast, yields reliable coefficient estimates even in the presence of outliers by diminishing the influence of these outliers on the squared error loss, thereby minimizing their effect on the regression estimates (Yu et al., 2014). Similar considerations can be extended to non-parametric tests, which are free from assumptions and therefore more flexible (Sedgwick, 2015).

The methodology utilized in the first study to predict *lncRNA* expression based on individual genotypes represents another strength of this thesis. Prediction of *lncRNA* expression on *GBC* cases and controls with only genotype information available is carried out by exploiting the summary statistics from the association between cis-*eQTLs* and the expression of preselected *lncRNAs* candidates on a distinct cohort of 110 controls. The plausibility of the findings is strengthened by the positive association between the genotype-based expression of $LINC00662$ and *GBC* risk, which is consistent with the results from the preselection stage. Given the absence of existing softwares capable of predicting *ncRNA* expression for specific traits, the development of a methodology such as the one described here is essential for facilitating the assessment of cancer risk. Thousands of variants associated to complex disease have been identified since the advent of *GWAS*, with approximately 50% of these being *eQTLs* (Ding et al., 2024). In the last decade, a gene-based software known as PrediXcan has been implemented and is largely employed to predict tissue-specific gene expression from individual genotypes (Gamazon et al., 2015). However, research has demonstrated that the prediction accuracy of PrediXcan can be adversely affected by factors such as population stratification (Mikhaylova and Thornton, 2019).

A limitation of both studies presented in this thesis is the directionality of the associations, which specifically investigates whether *GBC* causes changes in either *lncRNA* or *miRNA* expression. While this type of information is particularly relevant for risk prediction and disease prevention, the reverse direction, "*lncRNA/miRNA* expression changes cause *GBC*" cannot be investigated using the approach outlined in this thesis. A future objective is to explore the causality of these associations through *MR*. To date, no studies have yet employed *MR* to investigate the causal relationship between *miRNA* expression and *GBC*. More broadly, only a limited number of studies have applied this technique to either *GBC* or *ncRNAs*. In recent years, *MR* studies have successfully established the causal link between *GBC* and type 2 diabetes, gallstones, *BMI*, and C-Reactive protein (Cheng et al., 2024; Barahona Ponce et al., 2021). Although no *MR* studies have been applied to investigate the role of *miRNAs* on biliary tract cancers, little evidence exists regarding the causal association between *miRNA* expression and other diseases, such as severe COVID-19, schizophrenia, Parkinson's disease, and lung cancer (Li et al., 2021; Mu et al., 2023; Shi et al., 2024; Huang et al., 2020). The lack of *MR* studies extends to research on *lncRNAs* as well, with only a few published studies exploring the relationship between *lncRNA* expression and type 2 diabetes (de Klerk et al., 2023; Pan et al., 2020).

The absence of data on gallstone disease in the *miRNA* study constitutes another limitation of this thesis. In contrast, this information was partially available in the *lncRNA* study, providing a more comprehensive analysis. The most important factor described for *GBC* development is individual history of gallstones, which are present in almost 85% of patients diagnosed with *GBC* in Chile (Randi et al., 2006). Gallstone disease incidence is higher in individuals with Native American ancestry compared to other populations, yet it remains a significant risk factor for *GBC* in European populations as well (Liebe et al., 2015). Chronic inflammation and irritation caused by gallstones increase the susceptibility of the gallbladder's mucosa to malignant transformation, thereby elevating the risk of developing cancer (Wistuba and Gazdar, 2004). In terms of gallstone size, those bigger than 3 cm in diameter are associated with a tenfold increase in the risk of *GBC* compared to smaller ones (Rawla et al., 2019). To address the lack of gallstone information in the *miRNA* study, a polygenic risk score was calculated using genetic variants robustly associated with gallstone disease. However, the analysis was constrained by the availability of genetic data for only 80 individuals, limiting the statistical power of the findings. Future studies following up this thesis should also address this lack of information.

4.2 Conclusions

In summary, *GBC* remains an under-researched malignancy that is relatively rare in high-income countries, yet it is poses a significant public health challenge in certain low-income regions, such as Chile, where mortality rates rank among the highest globally. Current research on molecular phenotypes, including *ncRNAs*, associated with *GBC* development is still limited. Moreover, the relationship between tissue and serum biomarkers, which are less invasive and easily accessible, has not been extensively studied in the context of *GBC*.

This thesis sought to address this gap by identifying circulating *lncRNAs* and *miRNAs* as potential biomarkers for the prevention and early diagnosis of *GBC*. Both studies presented in this thesis focused on preselecting biomarkers in tissue and validating them in serum, targeting two distinct populations: Chileans, where *GBC* is highly prevalent, and Europeans, where *GBC* is rare and no robust risk biomarkers have been established.

In Chileans, the *lncRNAs* *AC*084082.3 and *LINC*00662 demonstrated a progressive increase in expression across the spectrum of gallstones, dysplasia, and cancer, while a lower expression of *C*22*orf*34 in *GBC* patients was linked to poorer *GBC* outcome. Moreover, the genotype-based expression of *LINC*00662 showed also a positive association with *GBC* progression, confirming its potential as cancer risk biomarker.

In European prospective serum samples, *miR*-4533 and *miR*-671-5*p* showed elevated expression levels in *GBC* cases, but only *miR*-4533 was validated through meta-analysis. The overexpression of miR-4533 was particularly evident in younger individuals and those with a lower BMI. Pathway analyses also uncovered $SIPA1L2$ as a novel target gene, which was downregulated in GBC cases, shedding light on the molecular mechanisms underlying GBC pathogenesis.

In conclusion, this thesis represents a significant contribution to the understanding of $ncRNAs$ in GBC, highlighting few key $lncRNAs$ and $miRNAs$ as potential biomarkers for the disease. The findings provide a basis for future research aimed at improving the risk prediction and early diagnosis of GBC. Furthermore, these results offer a foundation for the development of non-invasive diagnostic tools, which could especially benefit regions with limited healthcare resources. Reducing unnecessary cholecystectomies while maintaining high sensitivity for GBC detection is of particular relevance in countries like Chile, where the healthcare burden of GBC is substantial. The findings also underscore the need for continued investigation into $ncRNA$ dysregulation in GBC, with the ultimate goal of developing novel prevention strategies and non-invasive screening tools, crucial for early detection and better clinical outcomes in this often asymptomatic disease. A better understanding of individual GBC risk could lead to more tailored surveillance strategies and inform decision-making regarding prophylactic cholecystectomy, particularly for high-risk individuals.

Chapter 5

Summary

This thesis focuses on the identification of circulating non-coding RNAs associated with the risk of developing gallbladder cancer, an aggressive disease with poor prognosis. Globally, gallbladder cancer exhibits high prevalence and mortality rates in specific geographic regions, such as Latin America, while remaining relatively rare in Europe. The molecular and genetic mechanisms underlying gallbladder cancer development have been partially explored, yet the precise contributions of specific biomarkers to its development remain inadequately understood. Non-coding RNAs play a central role in regulating abnormal cell processes, and hold promise as valuable biomarkers of early disease detection. Two different types of non-coding RNAs were investigated in this thesis: long non-coding RNAs and microRNAs. Long non-coding RNA expression levels were evaluated in the Chilean population, while microRNA regulation was investigated in individuals of European ancestry. Both studies relied on the combination of tissue and serum non-coding RNA expression data.

The first study integrated three datasets containing long non-coding RNA expression data alone (gallstone n = 31, dysplasia n = 35, gallbladder cancer n = 32), both long non-coding RNA expression and genotype data (controls n = 110), and genotype information exclusively (controls n = 2397, gallbladder cancer cases n = 540). On the first dataset, differentially expressed long non-coding RNAs along the progression from gallstones, to dysplasia and gallbladder cancer were preselected. In the second dataset, the associations between genetic variants (SNPs) and the serum expressions of the preselected long non-coding RNAs were

assessed, and the best models for prediction were selected. Finally, serum long non-coding RNA expressions were predicted based on individual genotypes, and the association with gallbladder cancer risk was estimated. AC084082.3 and LINC00662 exhibited increased expression levels (p-value = 0.009), while C22orf34 showed downregulation in progressing from gallstones to gallbladder cancer (p-value = 0.04). Two SNPs were identified and validated for LINC00662 ($r^2 = 0.26$) and three for C22orf34 ($r^2 = 0.24$). Only the predicted serum expression of LINC00662 was significantly associated with gallbladder cancer risk, and linked to a 25% higher risk of developing cancer (odds ratio = 1.25, p-value = 0.02).

In the second study, a three-step approach was applied to preselect microRNAs from German formalin-fixed paraffin-embedded tissue samples (gallstone n = 8, gallbladder cancer n = 40), screen microRNA expressions in serum prospective samples from three European cohorts (n = 37 gallbladder cancer case-control pairs), and validate the identified microRNAs in serum samples from three additional prospective cohorts (controls n = 36, gallbladder cancer cases n = 31). Statistical analyses also included pathway and meta-analysis, and examination of expression correlation between microRNAs and target genes. miR-4533 and miR-671-5p were overexpressed both in gallbladder cancer tissue and in the first set of serum samples. However, only the overexpression of miR-4533 was validated both in the second set of prospective serum samples, and through meta-analysis (p-value = 4.1x10^{-4}). miR-4533 was mostly upregulated in individuals under 63.5 years, and with a body-mass index below 26.2 kg/m^2. Pathway and correlation analyses revealed that miR-4533 targets SIPA1L2 in the Rap1 signaling pathway.

This thesis demonstrates the heterogeneous nature of gallbladder cancer molecular profiles. Results from the first study suggest that preselection of long non-coding RNAs based on tissue samples and exploitation of related genetic variants facilitates the identification of circulating long non-coding RNAs linked to cancer risk. The second study draws attention to the importance of integrating tissue and serum biomarkers for the preselection, screening and validation of differentially expressed microRNAs. Both studies highlight the need for international research collaborations to identify and validate biomarkers for secondary prevention of rare tumours such as gallbladder cancer. These results need to be validated and further refined in future studies, also with regard to their transferability to other sample types and populations.

Chapter 6

Zusammenfassung

Diese Doktorarbeit befasst sich mit der Charakterisierung von zirkulierenden nicht-kodierenden RNAs, die mit dem Risiko der Entwicklung von Gallenblasenkrebs, einer aggressiven Erkrankung mit schlechter Prognose, verbunden sind. Weltweit weist Gallenblasenkrebs in bestimmten geographischen Regionen wie Lateinamerika eine hohe Prävalenz und eine hohe Sterblichkeitsrate auf, während die Erkrankung in europäischen Ländern relativ selten vorkommt. Die molekularen und genetischen Mechanismen, die dem Gallenblasenkrebs zugrunde liegen, sind zum Teil erforscht, doch die genauen Beiträge spezifischer Biomarker sind noch unzureichend bekannt. Nichtcodierende RNAs spielen eine zentrale Rolle bei der Regulierung abnormaler Zellprozesse und versprechen wertvolle Biomarker für die Früherkennung von Krankheiten zu sein.

In dieser Doktorarbeit wurden zwei verschiedene Arten von nicht-kodierenden RNAs untersucht: lange nicht-kodierende RNAs und microRNAs. Die Expressionsniveaus von langen nichtkodierenden RNAs wurden in der chilenischen Bevölkerung bewertet, während die Regulierung von microRNAs bei Personen europäischer Abstammung untersucht wurde. Beide Studien stützten sich auf die Kombination von Daten von nicht codierenden RNAs aus Gewebe und Serum.

In der ersten Studie wurden drei separate Datensätze zusammengeführt: Der erste Datensatz enthielt ausschließlich Daten zur Expression langer nichtkodierender RNAs (Gallenstein n = 31, Dysplasie n = 35, Gallenblasenkrebs n = 32), der zweite Datensatz umfasste sowohl Daten zur Expression langer nichtkodierender RNAs als auch Genotypdaten (Kontrollen n = 110) und der letze Datensatz enthielt nur Genotypinformationen (Kontrollen n = 2397, Gallenblasenkrebsfälle n = 540). Zunächst wurden die unterschiedlich exprimierten langen nichtkodierenden RNAs entlang der Progression von Gallensteinen über Dysplasie bis hin zu Gallenblasenkrebs vorselektiert. Im zweiten Datensatz wurden danach die Assoziationen zwischen Einzelnukleotidpolymorphismen (SNPs) und der Serumexpression der vorselektierten langen nicht-kodierenden RNAs bewertet und die besten Modelle für die Prediktion ausgewählt. Schließlich wurden die Ausprägungen der langen nicht-kodierenden RNAs im Serum auf der Grundlage der einzelnen Genotypen vorhergesagt, und der Zusammenhang mit Gallenblasenkrebsrisiko wurde bestimmt. AC084082.3 und LINC00662 wiesen erhöhte Expressionswerte auf (p-Wert = 0.009), während C22orf34 bei der Entwicklung von Gallensteinen zu Gallenblasenkrebs herunterreguliert war (p-Wert = 0.04). Zwei SNPs wurden für LINC00662 (r^2 = 0.26) und drei für C22orf34 (r^2 = 0.24) identifiziert und validiert. Bemerkenswert ist, dass nur die vorhergesagte Serumexpression von LINC00662 signifikant mit dem Gallenblasenkrebsrisiko assoziiert und mit einem 25% höheren Krebsrisiko verbunden war (Odds Ratio = 1.25, p-Wert = 0.02).

In der zweiten Studie wurde ein dreistufiger Ansatz angewandt, um microRNAs aus deutschen formalinfixierten Gewebeproben (Gallenstein n = 8, Gallenblasenkrebs n = 40) vorzuselektieren, Screening der microRNA-Expressionsniveaus in prospektiven Serumproben aus drei europäischen Kohorten (n = 37 Gallenblasenkrebs-Fall-Kontroll-Paare) durchzuführen, und die identifizierten microRNA-Kandidaten in Serumproben aus drei weiteren prospektiven Kohorten (Kontrollen n = 36, Gallenblasenkrebs-Fälle n = 31) zu validieren. Die statistischen Analysen umfassten auch Pathway- und Meta-Analyse sowie eine Untersuchung der Expressionskorrelation zwischen mikroRNAs und Zielgenen. miR-4533 und miR-671-5p waren sowohl im Gallenblasenkrebsgewebe als auch in der ersten Gruppe von Serumproben überexprimiert. Allerdings wurde nur die Überexpression von miR-4533 sowohl im zweiten Satz prospektiver Serumproben als auch durch Meta-Analyse validiert (p-Wert = $4.1\text{x}10^{-4}$). miR-4533 war besonders bei Personen unter 63.5 Jahren und mit einem Body-Mass-Index unter 26.2 kg/m^2 hochreguliert. Pathway- und Korrelationsanalysen ergaben außerdem, dass miR-4533 auf SIPA1L2 im Rap1-Signalweg abzielt.

Die Ergebnisse dieser Dissertation zeigen, wie heterogen die Genetik von Gallenblasenkrebs ist. Die Ergebnisse der ersten Studie deuten darauf hin, dass die Vorauswahl langer nichtkodierender RNAs auf der Grundlage von Gewebeproben und die Nutzung verwandter genetischer Varianten die Identifizierung zirkulierender langer nichtkodierender RNAs, die mit dem Krebsrisiko verbunden sind, ermöglicht. Die zweite Studie weist auf die Bedeutung der Integration von Gewebe- und Serum-Biomarkern für die Vorauswahl, das Screening und die Validierung von unterschiedlich exprimierten microRNAs hin. Beide Studien unterstreichen die Notwendigkeit internationaler Forschungskooperationen zur Identifizierung und Validierung von Biomarkern für die Sekundärprävention von seltenen Tumorerkrankungen wie Gallenblasenkrebs. Diese vielversprechenden Ergebnisse müssen in künftigen Studien validiert und weiter verfeinert werden, auch im Hinblick auf ihre Übertragbarkeit auf andere Probenarten und Populationen.

Bibliography

Adamsen, S., Hansen, O. H., Funch-Jensen, P., Schulze, S., Stage, J. G., and Wara, P. (1997). Bile duct injury during laparoscopic cholecystectomy: a prospective nationwide series. *J. Am. Coll. Surg.*, 184(6):571–578.

Aguayo-Mazzucato, C., Andle, J., Lee, Jr, T. B., Midha, A., Talemal, L., Chipashvili, V., Hollister-Lock, J., van Deursen, J., Weir, G., and Bonner-Weir, S. (2019). Acceleration of β cell aging determines diabetes and senolysis improves disease outcomes. *Cell Metab.*, 30(1):129–142.e4.

Alanamu, Oyeyemi, Olaniran, and Adetunji (2023). Review of some robust estimators in multiple linear regressions in the presence of outlier(s). *African Journal of Mathematics and Statistics Studies*, 6(3):59–69.

Ali, S., Almhanna, K., Chen, W., Philip, P. A., and Sarkar, F. H. (2010). Differentially expressed miRNAs in the plasma may provide a molecular signature for aggressive pancreatic cancer. *Am. J. Transl. Res.*, 3(1):28–47.

Anastasiadou, E., Jacob, L. S., and Slack, F. J. (2018). Non-coding RNA networks in cancer. *Nat. Rev. Cancer*, 18(1):5–18.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Aparicio-Puerta, E., Hirsch, P., Schmartz, G. P., Kern, F., Fehlmann, T., and Keller, A. (2023).
miEAA 2023: updates, new functional microRNA sets and improved enrichment visualizations.
Nucleic Acids Res., 51(W1):W319–W325.

Apt, W., Llancaqueo, M., Zulantay, I., Canals, M., Kara, S., Arribada, A., Muñoz, G., and Martínez, G. (2021).
Clinical, electrocardiographic and echocardiographic evolution of chronic chagas disease treated with nifurtimox on prolonged follow-up in chile: observational study.
J. Glob. Antimicrob. Resist., 27:160–166.

Armakolas, A., Kotsari, M., and Koskinas, J. (2023).
Liquid biopsies, novel approaches and future directions.
Cancers (Basel), 15(5).

Backes, C., Khaleeq, Q. T., Meese, E., and Keller, A. (2016).
miEAA: microRNA enrichment analysis and annotation.
Nucleic Acids Res., 44(W1):W110–6.

Barahona Ponce, C., Scherer, D., Brinster, R., Boekstegers, F., Marcelain, K., Gárate-Calderón, V., Müller, B., de Toro, G., Retamales, J., Barajas, O., Ahumada, M., Morales, E., Rojas, A., Sanhueza, V., Loader, D., Rivera, M. T., Gutiérrez, L., Bernal, G., Ortega, A., Montalvo, D., Portiño, S., Bertrán, M. E., Gabler, F., Spencer, L., Olloquequi, J., Fischer, C., Jenab, M., Aleksandrova, K., Katzke, V., Weiderpass, E., Bonet, C., Moradi, T., Fischer, K., Bossers, W., Brenner, H., Hveem, K., Eklund, N., Völker, U., Waldenberger, M., Fuentes Guajardo, M., Gonzalez-Jose, R., Bedoya, G., Bortolini, M. C., Canizales-Quinteros, S., Gallo, C., Ruiz-Linares, A., Rothhammer, F., and Lorenzo Bermejo, J. (2021).
Gallstones, body mass index, c-reactive protein, and gallbladder cancer: Mendelian randomization analysis of chilean and european genotype data.
Hepatology, 73(5):1783–1796.

Bartlett, J. and Charles, S. (2022).

Power to the people: A beginner's tutorial to power analysis using jamovi.
Meta-Psychology, 6.

Beg, A., Parveen, R., Fouad, H., Yahia, M. E., and Hassanein, A. S. (2022).
Role of different non-coding RNAs as ovarian cancer biomarkers.
J. Ovarian Res., 15(1):72.

Bhattacharyya, M., Nath, J., and Bandyopadhyay, S. (2015).
MicroRNA signatures highlight new breast cancer subtypes.
Gene, 556(2):192–198.

Bizama, C., García, P., Espinoza, J. A., Weber, H., Leal, P., Nervi, B., and Roa, J. C. (2015).
Targeting specific molecular pathways holds promise for advanced gallbladder cancer therapy.
Cancer Treat. Rev., 41(3):222–234.

Blandino, A., Scherer, D., Rounge, T. B., Umu, S. U., Boekstegers, F., Barahona Ponce, C., Marcelain, K., Gárate-Calderón, V., Waldenberger, M., Morales, E., Rojas, A., Munoz, C., Retamales, J., de Toro, G., Barajas, O., Rivera, M. T., Cortés, A., Loader, D., Saavedra, J., Gutiérrez, L., Ortega, A., Bertrán, M. E., Gabler, F., Campos, M., Alvarado, J., Moisán, F., Spencer, L., Nervi, B., Carvajal-Hausdorf, D. E., Losada, H., Almau, M., Fernández, P., Gallegos, I., Olloquequi, J., Fuentes-Guajardo, M., Gonzalez-Jose, R., Bortolini, M. C., Gallo, C., Linares, A. R., Rothhammer, F., and Lorenzo Bermejo, J. (2022).
Identification of circulating lncRNAs associated with gallbladder cancer risk by tissue-based preselection, Cis-eQTL validation, and analysis of association with genotype-based expression.
Cancers (Basel), 14(3):634.

Boekstegers, F., Marcelain, K., Barahona Ponce, C., Baez Benavides, P. F., Müller, B., de Toro, G., Retamales, J., Barajas, O., Ahumada, M., Morales, E., Rojas, A., Sanhueza, V., Loader, D., Rivera, M. T., Gutiérrez, L., Bernal, G., Ortega, A., Montalvo, D., Portiño, S., Bertrán, M. E., Gabler, F., Spencer, L., Olloquequi, J., González Silos, R., Fischer, C., Scherer, D., Jenab, M., Aleksandrova, K., Katzke, V., Weiderpass, E., Moradi, T., Fischer, K., Bossers, W., Brenner, H., Hveem, K., Eklund, N., Völker, U., Waldenberger, M.,

Fuentes Guajardo, M., Gonzalez-Jose, R., Bedoya, G., Bortolini, M. C., Canizales, S., Gallo, C., Ruiz Linares, A., Rothhammer, F., and Lorenzo Bermejo, J. (2020).
ABCB1/4 gallbladder cancer risk variants identified in india also show strong effects in chileans.
Cancer Epidemiol., 65(101643):101643.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003).
A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.
Bioinformatics, 19(2):185–193.

Bonetti, P., Climent, M., Panebianco, F., Tordonato, C., Santoro, A., Marzi, M. J., Pelicci, P. G., Ventura, A., and Nicassio, F. (2019).
Dual role for mir-34a in the control of early progenitor proliferation and commitment in the mammary gland and in breast cancer.
Oncogene, 38(3):360–374.

Boone, M., De Koker, A., and Callewaert, N. (2018).
Capturing the 'ome': the expanding molecular toolbox for RNA and DNA library construction.
Nucleic Acids Res., 46(6):2701–2721.

Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., Kuulasmaa, K., Laatikainen, T., Männistö, S., Peltonen, M., Perola, M., Puska, P., Salomaa, V., Sundvall, J., Virtanen, S. M., and Vartiainen, E. (2018).
Cohort profile: The national FINRISK study.
Int. J. Epidemiol., 47(3):696–696i.

Brägelmann, J., Barahona Ponce, C., Marcelain, K., Roessler, S., Goeppert, B., Gallegos, I., Colombo, A., Sanhueza, V., Morales, E., Rivera, M. T., de Toro, G., Ortega, A., Müller, B., Gabler, F., Scherer, D., Waldenberger, M., Reischl, E., Boekstegers, F., Garate-Calderon, V., Umu, S. U., Rounge, T. B., Popanda, O., and Lorenzo Bermejo, J. (2021).
Epigenome-wide analysis of methylation changes in the sequence of gallstone disease, dysplasia, and gallbladder cancer.

Hepatology, 73(6):2293–2310.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024).
Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries.
CA: A Cancer Journal for Clinicians, 74(3):229–263.

Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D., and Mountain, J. L. (2015).
The genetic ancestry of african americans, latinos, and european americans across the united states.
Am. J. Hum. Genet., 96(1):37–53.

Campbell, P. T., Newton, C. C., Kitahara, C. M., Patel, A. V., Hartge, P., Koshiol, J., McGlynn, K. A., Adami, H.-O., Berrington de González, A., Beane Freeman, L. E., Bernstein, L., Buring, J. E., Freedman, N. D., Gao, Y.-T., Giles, G. G., Gunter, M. J., Jenab, M., Liao, L. M., Milne, R. L., Robien, K., Sandler, D. P., Schairer, C., Sesso, H. D., Shu, X.-O., Weiderpass, E., Wolk, A., Xiang, Y.-B., Zeleniuch-Jacquotte, A., Zheng, W., and Gapstur, S. M. (2017).
Body size indicators and risk of gallbladder cancer: Pooled analysis of individual-level data from 19 prospective cohort studies.
Cancer Epidemiol. Biomarkers Prev., 26(4):597–606.

Cao, Y., Zhao, R., Guo, K., Ren, S., Zhang, Y., Lu, Z., Tian, L., Li, T., Chen, X., and Wang, Z. (2021).
Potential metabolite biomarkers for early detection of stage-i pancreatic ductal adenocarcinoma.
Front. Oncol., 11:744667.

Carey, M. C. and Paigen, B. (2002).
Epidemiology of the american indians' burden and its likely genetic origins.
Hepatology, 36(4 Pt 1):781–791.

Carthew, R. W. (2021).

Gene regulation and cellular metabolism: An essential partnership.
Trends Genet., 37(4):389–400.

Chang, Y., Liu, C., Yang, J., Liu, G., Feng, F., Tang, J., Hu, L., Li, L., Jiang, F., Chen, C., Wang, R., Yang, Y., Jiang, X., Wu, M., Chen, L., and Wang, H. (2013).
MiR-20a triggers metastasis of gallbladder carcinoma.
J. Hepatol., 59(3):518–527.

Chen, T. and Guestrin, C. (2016).
XGBoost: A scalable tree boosting system.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chen, X., Luo, J., Gao, S., Jiang, J., Yang, B., and Zhang, Z. (2022).
MiR-671-5p promotes cell proliferation, invasion, and migration in hepatocellular carcinoma through targeting ALDH2.
Crit. Rev. Eukaryot. Gene Expr., 32(4):73–82.

Cheng, M., Zhou, X., Xue, Y., Zhou, E., Hu, J., Xu, J., Zhang, B., Shen, J., Zhang, J., Chen, Z., Wu, B., Peng, S., Wong, T.-W., Cao, J., and Chen, M. (2024).
Association between type 2 diabetes, alcohol intake frequency, age at menarche, and gallbladder cancer: a two-sample mendelian randomization study.
J. Gastrointest. Oncol., 15(3):1214–1223.

Cheng, W.-H., Kao, S.-Y., Chen, C.-L., Yuliani, F. S., Lin, L.-Y., Lin, C.-H., and Chen, B.-C. (2022).
Amphiregulin induces CCN2 and fibronectin expression by TGF-β through EGFR-dependent pathway in lung epithelial cells.
Respir. Res., 23(1):381.

Chi, X. G., Meng, X. X., Ding, D. L., Xuan, X. H., Chen, Y. Z., Cai, Q., and Wang, A. (2020).
HMGA1-mediated mir-671-5p targets APC to promote metastasis of clear cell renal cell carcinoma through wnt signaling.

Neoplasma, 67(1):46–53.

Choi, K. S., Choi, S. B., Park, P., Kim, W. B., and Choi, S. Y. (2015).
Clinical characteristics of incidental or unsuspected gallbladder cancers diagnosed during or after cholecystectomy: a systematic review and meta-analysis.
World J. Gastroenterol., 21(4):1315–1323.

Coe, E. A., Tan, J. Y., Shapiro, M., Louphrasitthiphol, P., Bassett, A. R., Marques, A. C., Goding, C. R., and Vance, K. W. (2019).
The MITF-SOX10 regulated long non-coding RNA DIRC3 is a melanoma tumour suppressor.
PLoS Genet., 15(12):e1008501.

Crick, F. (1970).
Central dogma of molecular biology.
Nature, 227(5258):561–563.

Darci-Maher, N., Alvarez, M., Arasu, U. T., Selvarajan, I., Lee, S. H. T., Pan, D. Z., Miao, Z., Das, S. S., Kaminska, D., Örd, T., Benhammou, J. N., Wabitsch, M., Pisegna, J. R., Männistö, V., Pietiläinen, K. H., Laakso, M., Sinsheimer, J. S., Kaikkonen, M. U., Pihlajamäki, J., and Pajukanta, P. (2023).
Cross-tissue omics analysis discovers ten adipose genes encoding secreted proteins in obesity-related non-alcoholic fatty liver disease.
EBioMedicine, 92(104620):104620.

de Klerk, J. A., Beulens, J. W. J., Mei, H., Bijkerk, R., van Zonneveld, A. J., Koivula, R. W., Elders, P. J. M., 't Hart, L. M., and Slieker, R. C. (2023).
Altered blood gene expression in the obesity-related type 2 diabetes cluster may be causally involved in lipid metabolism: a mendelian randomisation study.
Diabetologia, 66(6):1057–1070.

Di Ciaula, A., Wang, D. Q.-H., and Portincasa, P. (2018).
An update on the pathogenesis of cholesterol gallstone disease.
Curr. Opin. Gastroenterol., 34(2):71–80.

Díaz-Peña, R., Silva, R. S., Hosgood, 3rd, H. D., Agustí, À., and Olloquequi, J. (2022).
PERFIL ESPECÍFICO DE miRNAs EN ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA ASOCIADA a EXPOSICIÓN a HUMO DE BIOMASA.
Arch. Bronconeumol., 58(2):177–179.

Ding, R., Wang, Q., Gong, L., Zhang, T., Zou, X., Xiong, K., Liao, Q., Plass, M., and Li, L. (2024).
scQTLbase: an integrated human single-cell eQTL database.
Nucleic Acids Res., 52(D1):D1010–D1017.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., and Gingeras, T. R. (2012).
Landscape of transcription in human cells.
Nature, 489(7414):101–108.

Duffy, A., Capanu, M., Abou-Alfa, G. K., Huitzil, D., Jarnagin, W., Fong, Y., D'Angelica, M., Dematteo, R. P., Blumgart, L. H., and O'Reilly, E. M. (2008).
Gallbladder cancer (GBC): 10-year experience at memorial Sloan-Kettering cancer centre (MSKCC).
J. Surg. Oncol., 98(7):485–489.

Dutta, U. (2012).

Gallbladder cancer: can newer insights improve the outcome?
J. Gastroenterol. Hepatol., 27(4):642–653.

Elcheva, I. A. and Spiegelman, V. S. (2020).
The role of cis- and trans-acting RNA regulatory elements in leukemia.
Cancers (Basel), 12(12):3854.

ENCODE Project Consortium (2012).
An integrated encyclopedia of DNA elements in the human genome.
Nature, 489(7414):57–74.

Espinoza, J. A., Bizama, C., García, P., Ferreccio, C., Javle, M., Miquel, J. F., Koshiol, J., and Roa, J. C. (2016).
The inflammatory inception of gallbladder cancer.
Biochim. Biophys. Acta, 1865(2):245–254.

Ferkingstad, E., Oddsson, A., Gretarsdottir, S., Benonisdottir, S., Thorleifsson, G., Deaton, A. M., Jonsson, S., Stefansson, O. A., Norddahl, G. L., Zink, F., Arnadottir, G. A., Gunnarsson, B., Halldorsson, G. H., Helgadottir, A., Jensson, B. O., Kristjansson, R. P., Sveinbjornsson, G., Sverrisson, D. A., Masson, G., Olafsson, I., Eyjolfsson, G. I., Sigurdardottir, O., Holm, H., Jonsdottir, I., Olafsson, S., Steingrimsdottir, T., Rafnar, T., Bjornsson, E. S., Thorsteinsdottir, U., Gudbjartsson, D. F., Sulem, P., and Stefansson, K. (2018).
Genome-wide association meta-analysis yields 20 loci associated with gallstone disease.
Nat. Commun., 9(1):5101.

Finall, A., Davies, G., Jones, T., Emlyn, G., Huey, P., and Mullard, A. (2023).
Integration of rapid PCR testing as an adjunct to NGS in diagnostic pathology services within the UK: evidence from a case series of non-squamous, non-small cell lung cancer (NSCLC) patients with follow-up.
J. Clin. Pathol., 76(6):391–399.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala,

S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2019).
GENCODE reference annotation for the human and mouse genomes.
Nucleic Acids Res., 47(D1):D766–D773.

Fryda, T., LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava, M., Poirier, S., and Wong, W. (2014).
H2o: R interface for the 'H2O' scalable machine learning platform.
Title of the publication associated with this dataset: CRAN: Contributed Packages.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., GTEx Consortium, Nicolae, D. L., Cox, N. J., and Im, H. K. (2015).
A gene-based association method for mapping traits using reference transcriptome data.
Nat. Genet., 47(9):1091–1098.

Gao, C., Wei, J., Tang, T., and Huang, Z. (2020).
Role of microRNA-33a in malignant cells.
Oncol. Lett., 20(3):2537–2556.

Ghafouri-Fard, S., Dashti, S., and Taheri, M. (2020).
The HOTTIP (HOXA transcript at the distal tip) lncRNA: Review of oncogenic roles in human.
Biomed. Pharmacother., 127(110158):110158.

Ghanbari, R., Mosakhani, N., Sarhadi, V. K., Armengol, G., Nouraee, N., Mohammadkhani, A., Khorrami, S., Arefian, E., Paryan, M., Malekzadeh, R., and Knuutila, S. (2015).
Simultaneous underexpression of let-7a-5p and let-7f-5p microRNAs in plasma and stool samples from early stage colorectal carcinoma.

Biomark. Cancer, 7(Suppl 1):39–48.

Ghosh, M., Sakhuja, P., Singh, S., and Agarwal, A. K. (2013).
P53 and beta-catenin expression in gallbladder tissues and correlation with tumor progression in gallbladder cancer.
Saudi J. Gastroenterol., 19(1):34–39.

Giraldo, N. A., Drill, E., Satravada, B. A., Dika, I. E., Brannon, A. R., Dermawan, J., Mohanty, A., Ozcan, K., Chakravarty, D., Benayed, R., Vakiani, E., Abou-Alfa, G. K., Kundra, R., Schultz, N., Li, B. T., Berger, M. F., Harding, J. J., Ladanyi, M., O'Reilly, E. M., Jarnagin, W., Vanderbilt, C., Basturk, O., and Arcila, M. E. (2022).
Comprehensive molecular characterization of gallbladder carcinoma and potential targets for intervention.
Clin. Cancer Res., 28(24):5359–5367.

Glinge, C., Clauss, S., Boddum, K., Jabbari, R., Jabbari, J., Risgaard, B., Tomsits, P., Hildebrand, B., Kääb, S., Wakili, R., Jespersen, T., and Tfelt-Hansen, J. (2017).
Stability of circulating blood-based MicroRNAs - pre-analytic methodological considerations.
PLoS One, 12(2):e0167969.

Goeppert, B., Truckenmueller, F., Ori, A., Fritz, V., Albrecht, T., Fraas, A., Scherer, D., Silos, R. G., Sticht, C., Gretz, N., Mehrabi, A., Bewerunge-Hudler, M., Pusch, S., Bermejo, J. L., Dietrich, P., Schirmacher, P., Renner, M., and Roessler, S. (2019).
Profiling of gallbladder carcinoma reveals distinct miRNA profiles and activation of STAT1 by the tumor suppressive miRNA-145-5p.
Sci. Rep., 9(1):4796.

Gong, W., Su, Y., Liu, Y., Sun, P., and Wang, X. (2018).
Long non-coding RNA linc00662 promotes cell invasion and contributes to cancer stem cell-like phenotypes in lung cancer cells.
J. Biochem., 164(6):461–469.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008).

miRBase: tools for microRNA genomics.
Nucleic Acids Res., 36(Database issue):D154–8.

Guo, T., Gong, C., Wu, P., Battaglia-Hsu, S.-F., Feng, J., Liu, P., Wang, H., Guo, D., Yao, Y., Chen, B., Xiao, Y., Liu, Z., and Li, Z. (2020).
LINC00662 promotes hepatocellular carcinoma progression via altering genomic methylation profiles.
Cell Death Differ., 27(7):2191–2205.

Ha, T.-Y. (2011).
MicroRNAs in human diseases: From cancer to cardiovascular disease.
Immune Netw., 11(3):135–154.

Han, Z. and Shi, L. (2018).
Long non-coding RNA LUCAT1 modulates methotrexate resistance in osteosarcoma via miR-200c/ABCB1 axis.
Biochem. Biophys. Res. Commun., 495(1):947–953.

He, Y., Xu, Y., Yu, X., Sun, Z., and Guo, W. (2021).
The vital roles of LINC00662 in human cancers.
Front. Cell Dev. Biol., 9:711352.

Heegaard, N. H. H., Schetter, A. J., Welsh, J. A., Yoneda, M., Bowman, E. D., and Harris, C. C. (2012).
Circulating micro-RNA expression profiles in early stage nonsmall cell lung cancer.
Int. J. Cancer, 130(6):1378–1386.

Hemminki, K., Försti, A., Hemminki, O., Liska, V., and Hemminki, A. (2022).
Long-term incidence and survival trends in cancer of the gallbladder and extrahepatic bile ducts in denmark, finland, norway and sweden with etiological implications related to thorotrast.
Int. J. Cancer, 151(2):200–208.

Hidalgo Grau, L. A., Badia, J. M., Salvador, C. A., Monsó, T. S., Canaleta, J. F., Nogués, J. M. G., and Sala, J. S. (2004).

Gallbladder carcinoma: the role of p53 protein overexpression and ki-67 antigen expression as prognostic markers.
HPB (Oxford), 6(3):174–180.

Huang, R., Cho, W. C., Sun, Y., and Katie CHAN, K. H. (2020).
The lung cancer associated MicroRNAs and single nucleotides polymorphisms: A mendelian randomization analysis.
In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE.

Hundal, R. and Shaffer, E. A. (2014).
Gallbladder cancer: epidemiology and outcome.
Clin. Epidemiol., 6:99–109.

Ishigami, K., Nosho, K., Koide, H., Kanno, S., Mitsuhashi, K., Igarashi, H., Shitani, M., Motoya, M., Kimura, Y., Hasegawa, T., Kaneto, H., Takemasa, I., Suzuki, H., and Nakase, H. (2018).
MicroRNA-31 reflects IL-6 expression in cancer tissue and is related with poor prognosis in bile duct cancer.
Carcinogenesis, 39(9):1127–1134.

Jackson, S. S., Van Dyke, A. L., Zhu, B., Pfeiffer, R. M., Petrick, J. L., Adami, H.-O., Albanes, D., Andreotti, G., Beane Freeman, L. E., Berrington de González, A., Buring, J. E., Chan, A. T., Chen, Y., Fraser, G. E., Freedman, N. D., Gao, Y.-T., Gapstur, S. M., Gaziano, J. M., Giles, G. G., Grant, E. J., Grodstein, F., Hartge, P., Jenab, M., Kitahara, C. M., Knutsen, S. F., Koh, W.-P., Larsson, S. C., Lee, I.-M., Liao, L. M., Luo, J., McGee, E. E., Milne, R. L., Monroe, K. R., Neuhouser, M. L., O'Brien, K. M., Peters, U., Poynter, J. N., Purdue, M. P., Robien, K., Sandler, D. P., Sawada, N., Schairer, C., Sesso, H. D., Simon, T. G., Sinha, R., Stolzenberg-Solomon, R. Z., Tsugane, S., Wang, R., Weiderpass, E., Weinstein, S. J., White, E., Wolk, A., Yuan, J.-M., Zeleniuch-Jacquotte, A., Zhang, X., McGlynn, K. A., Campbell, P. T., and Koshiol, J. (2019).
Anthropometric risk factors for cancers of the biliary tract in the biliary tract cancers pooling project.

Cancer Res., 79(15):3973–3982.

Jiao, X.-D., Ding, L.-R., Zhang, C.-T., Qin, B.-D., Liu, K., Jiang, L.-P., Wang, X., Lv, L.-T., Ding, H., Li, D.-M., Yang, H., Chen, X.-Q., Zhu, W.-Y., Wu, Y., Ling, Y., He, X., Liu, J., Shao, L., Wang, H.-Z., Chen, Y., Zheng, J.-J., Inui, N., and Zang, Y.-S. (2021).
Serum tumor markers for the prediction of concordance between genomic profiles from liquid and tissue biopsy in patients with advanced lung adenocarcinoma.
Transl. Lung Cancer Res., 10(7):3236–3250.

Jin, W., Shi, J., and Liu, M. (2019).
Overexpression of mir-671-5p indicates a poor prognosis in colon cancer and accelerates proliferation, migration, and invasion of colon cancer cells.
Onco. Targets. Ther., 12:6865–6873.

Jonckheere, A. R. (1954).
A DISTRIBUTION-FREE k-SAMPLE TEST AGAINST ORDERED ALTERNATIVES.
Biometrika, 41(1-2):133–145.

Kaikkonen, M. U., Lam, M. T. Y., and Glass, C. K. (2011).
Non-coding RNAs as regulators of gene expression and epigenetics.
Cardiovasc. Res., 90(3):430–440.

Kamisawa, T., Kuruma, S., Chiba, K., Tabata, T., Koizumi, S., and Kikuyama, M. (2017).
Biliary carcinogenesis in pancreaticobiliary maljunction.
J. Gastroenterol., 52(2):158–163.

Kanthan, R., Senger, J.-L., Ahmed, S., and Kanthan, S. C. (2015).
Gallbladder cancer in the 21st century.
J. Oncol., 2015:967472.

Karimi, N., Oloomi, M., and Orafa, Z. (2020).
Circulating tumor cells detection in patients with early breast cancer using MACS immunomagnetic flow cytometry.
Avicenna J. Med. Biotechnol., 12(3):148–156.

Kono, H., Nakamura, M., Ohtsuka, T., Nagayoshi, Y., Mori, Y., Takahata, S., Aishima, S., and Tanaka, M. (2013).
High expression of microRNA-155 is associated with the aggressive malignant behavior of gallbladder carcinoma.
Oncol. Rep., 30(1):17–24.

Koshiol, J., Van De Wyngard, V., McGee, E. E., Cook, P., Pfeiffer, R. M., Mardones, N., Medina, K., Olivo, V., Pettit, K., Jackson, S. S., Paredes, F., Sanchez, R., Huidobro, A., Villaseca, M., Bellolio, E., Losada, H., Roa, J. C., Hildesheim, A., Araya, J. C., Ferreccio, C., and the Chile BiLS Study Group (2021).
The chile biliary longitudinal study: A gallstone cohort.
Am. J. Epidemiol., 190(2):196–206.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019).
miRBase: from microRNA sequences to function.
Nucleic Acids Res., 47(D1):D155–D162.

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., Bratberg, G., Heggland, J., and Holmen, J. (2013).
Cohort profile: The HUNT study, norway.
Int. J. Epidemiol., 42(4):968–977.

Kvam, V. M., Liu, P., and Si, Y. (2012).
A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data.
Am. J. Bot., 99(2):248–256.

Lai, C. H. E. and Lau, W. Y. (2008).
Gallbladder cancer–a comprehensive review.
Surgeon, 6(2):101–110.

Lai, J., Chen, B., Zhang, G., Wang, Y., Mok, H., Wen, L., Pan, Z., Su, F., and Liao, N. (2019).

Identification of a novel microRNA recurrence-related signature and risk stratification system in breast cancer.
Aging (Albany NY), 11(18):7525–7536.

Langmead, B. and Salzberg, S. L. (2012).
Fast gapped-read alignment with bowtie 2.
Nat. Methods, 9(4):357–359.

Langseth, H., Gislefoss, R. E., Martinsen, J. I., Dillner, J., and Ursin, G. (2017).
Cohort profile: The janus serum bank cohort in norway.
Int. J. Epidemiol., 46(2):403–404g.

Lazcano-Ponce, E. C., Miquel, J. F., Muñoz, N., Herrero, R., Ferrecio, C., Wistuba, I. I., Alonso de Ruiz, P., Aristi Urista, G., and Nervi, F. (2001).
Epidemiology and molecular pathology of gallbladder cancer.
CA Cancer J. Clin., 51(6):349–364.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., and Kim, V. N. (2004).
MicroRNA genes are transcribed by RNA polymerase II.
EMBO J., 23(20):4051–4060.

Li, C., Wu, A., Song, K., Gao, J., Huang, E., Bai, Y., and Liu, X. (2021).
Identifying putative causal links between MicroRNAs and severe COVID-19 using mendelian randomization.
Cells, 10(12):3504.

Li, J., Xue, Y., Amin, M. T., Yang, Y., Yang, J., Zhang, W., Yang, W., Niu, X., Zhang, H.-Y., and Gong, J. (2020).
ncRNA-eQTL: a database to systematically evaluate the effects of SNPs on non-coding RNA expression across cancer types.
Nucleic Acids Res., 48(D1):D956–D963.

Li, L., Gan, Y., Li, W., Wu, C., and Lu, Z. (2016).
Overweight, obesity and the risk of gallbladder and extrahepatic bile duct cancers: A meta-analysis of observational studies.

Obesity (Silver Spring), 24(8):1786–1802.

Li, S., Mei, Z., Hu, H.-B., and Zhang, X. (2018).
The lncRNA MALAT1 contributes to non-small cell lung cancer development via modulating miR-124/STAT3 axis.
J. Cell. Physiol., 233(9):6679–6688.

Li, Z., Gao, J., Sun, D., Jiao, Q., Ma, J., Cui, W., Lou, Y., Xu, F., Li, S., and Li, H. (2022).
LncRNA MEG3: Potential stock for precision treatment of cardiovascular diseases.
Front. Pharmacol., 13:1045501.

Li, Z., Liu, L., Feng, C., Qin, Y., Xiao, J., Zhang, Z., and Ma, L. (2023).
LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations.
Nucleic Acids Res., 51(D1):D186–D191.

Li, Z., Yu, X., Shen, J., Law, P. T. Y., Chan, M. T. V., and Wu, W. K. K. (2015).
MicroRNA expression and its implications for diagnosis and therapy of gallbladder cancer.
Oncotarget, 6(16):13914–13921.

Lichtenstein, P., De Faire, U., Floderus, B., Svartengren, M., Svedberg, P., and Pedersen, N. L. (2002).
The swedish twin registry: a unique resource for clinical, epidemiological and genetic studies.
J. Intern. Med., 252(3):184–205.

Liebe, R., Milkiewicz, P., Krawczyk, M., Bonfrate, L., Portincasa, P., and Krawczyk, M. (2015).
Modifiable factors and genetic predisposition associated with gallbladder cancer. a concise review.
J. Gastrointestin. Liver Dis., 24(3):339–348.

Lin, S.-L., Miller, J. D., and Ying, S.-Y. (2006).
Intronic microRNA (miRNA).
J. Biomed. Biotechnol., 2006(4):26818.

Lin, Y., Kawai, S., Sasakabe, T., Kurosawa, M., Tamakoshi, A., Kikuchi, S., and JACC Study Group (2022).
Associations between cigarette smoking and biliary tract cancer by anatomic subsite and sex: a prospective cohort study in japan.
Cancer Causes Control, 33(11):1335–1341.

Liu, S. J., Dang, H. X., Lim, D. A., Feng, F. Y., and Maher, C. A. (2021).
Long noncoding RNAs in cancer metastasis.
Nat. Rev. Cancer, 21(7):446–460.

López-Jiménez, E. and Andrés-León, E. (2021).
The implications of ncRNAs in the development of human diseases.
Noncoding RNA, 7(1):17.

Lorenzo Bermejo, J., Boekstegers, F., González Silos, R., Marcelain, K., Baez Benavides, P., Barahona Ponce, C., Müller, B., Ferreccio, C., Koshiol, J., Fischer, C., Peil, B., Sinsheimer, J., Fuentes Guajardo, M., Barajas, O., Gonzalez-Jose, R., Bedoya, G., Cátira Bortolini, M., Canizales-Quinteros, S., Gallo, C., Ruiz Linares, A., and Rothhammer, F. (2017).
Subtypes of native american ancestry and leading causes of death: Mapuche ancestry-specific associations with gallbladder cancer risk in chile.
PLoS Genet., 13(5):e1006756.

Love, M. I., Huber, W., and Anders, S. (2014).
Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.
Genome Biol., 15(12):550.

Lu, W., Zhang, Y., Zhou, L., Wang, X., Mu, J., Jiang, L., Hu, Y., Dong, P., and Liu, Y. (2015).
mir-122 inhibits cancer cell malignancy by targeting PKM2 in gallbladder carcinoma.
Tumour Biol., 37(12):15615–15625.

Lu, Y., Zhao, X., Liu, Q., Li, C., Graves-Deal, R., Cao, Z., Singh, B., Franklin, J. L., Wang, J., Hu, H., Wei, T., Yang, M., Yeatman, T. J., Lee, E., Saito-Diaz, K., Hinger, S., Patton, J. G., Chung, C. H., Emmrich, S., Klusmann, J.-H., Fan, D., and Coffey, R. J. (2017).

lncRNA MIR100HG-derived mir-100 and mir-125b mediate cetuximab resistance via Wnt/β-catenin signaling.
Nat. Med., 23(11):1331–1341.

Lv, X., Lian, Y., Liu, Z., Xiao, J., Zhang, D., and Yin, X. (2021).
Exosomal long non-coding RNA LINC00662 promotes non-small cell lung cancer progression by miR-320d/E2F1 axis.
Aging (Albany NY), 13(4):6010–6024.

Ma, M.-Z., Chu, B.-F., Zhang, Y., Weng, M.-Z., Qin, Y.-Y., Gong, W., and Quan, Z.-W. (2015).
Long non-coding RNA CCAT1 promotes gallbladder cancer development via negative modulation of miRNA-218-5p.
Cell Death Dis., 6(1):e1583.

Ma, X., Zhou, L., and Zheng, S. (2020).
Transcriptome analysis revealed key prognostic genes and microRNAs in hepatocellular carcinoma.
PeerJ, 8(e8930):e8930.

Madani, A., Namazi, B., Altieri, M. S., Hashimoto, D. A., Rivera, A. M., Pucher, P. H., Navarrete-Welton, A., Sankaranarayanan, G., Brunt, L. M., Okrainec, A., and Alseidi, A. (2022).
Artificial intelligence for intraoperative guidance: Using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy.
Ann. Surg., 276(2):363–369.

Maldonado, B. L., Piqué, D. G., Kaplan, R. C., Claw, K. G., and Gignoux, C. R. (2023).
Genetic risk prediction in Hispanics/Latinos: milestones, challenges, and social-ethical considerations.
J. Community Genet., 14(6):543–553.

Mantripragada, K. C., Hamid, F., Shafqat, H., and Olszewski, A. J. (2017).

Adjuvant therapy for resected gallbladder cancer: Analysis of the national cancer data base.
J. Natl. Cancer Inst., 109(2):djw202.

Marin, D. H. (2021).
repmod: Create report table from different objects.
Title of the publication associated with this dataset: CRAN: Contributed Packages.

Mathy, N. W. and Chen, X.-M. (2017).
Long non-coding RNAs (lncRNAs) and their transcriptional control of inflammatory responses.
J. Biol. Chem., 292(30):12375–12382.

Mattick, J. S., Amaral, P. P., Carninci, P., Carpenter, S., Chang, H. Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M. E., Fitzgerald, K. A., Gingeras, T. R., Guttman, M., Hirose, T., Huarte, M., Johnson, R., Kanduri, C., Kapranov, P., Lawrence, J. B., Lee, J. T., Mendell, J. T., Mercer, T. R., Moore, K. J., Nakagawa, S., Rinn, J. L., Spector, D. L., Ulitsky, I., Wan, Y., Wilusz, J. E., and Wu, M. (2023).
Long non-coding RNAs: definitions, functions, challenges and recommendations.
Nat. Rev. Mol. Cell Biol., 24(6):430–447.

Mazerolle, M. J. (2023).
AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c).
R package version 2.3.3.

Menon, G. and Babiker, H. M. (2024).
Gallbladder carcinoma.
In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

Mhatre, S., Wang, Z., Nagrani, R., Badwe, R., Chiplunkar, S., Mittal, B., Yadav, S., Zhang, H., Chung, C. C., Patil, P., Chanock, S., Dikshit, R., Chatterjee, N., and Rajaraman, P. (2017).
Common genetic variation and risk of gallbladder cancer in india: a case-control genome-wide association study.

Lancet Oncol., 18(4):535–544.

Mikhaylova, A. V. and Thornton, T. A. (2019).
Accuracy of gene expression prediction from genotype data with PrediXcan varies across and within continental populations.
Front. Genet., 10:261.

Mills, M. C. and Rahal, C. (2019).
A scientometric review of genome-wide association studies.
Commun. Biol., 2(1):9.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008).
Mapping and quantifying mammalian transcriptomes by RNA-Seq.
Nat. Methods, 5(7):621–628.

Mu, C., Dang, X., and Luo, X.-J. (2023).
Mendelian randomization reveals the causal links between microRNA and schizophrenia.
J. Psychiatr. Res., 163:372–377.

Nakamura, H., Arai, Y., Totoki, Y., Shirota, T., Elzawahry, A., Kato, M., Hama, N., Hosoda, F., Urushidate, T., Ohashi, S., Hiraoka, N., Ojima, H., Shimada, K., Okusaka, T., Kosuge, T., Miyagawa, S., and Shibata, T. (2015).
Genomic spectra of biliary tract cancer.
Nat. Genet., 47(9):1003–1010.

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., and Mathur, P. (2023).
Bias in artificial intelligence algorithms and recommendations for mitigation.
PLOS Digit. Health, 2(6):e0000278.

Niu, J., Song, X., and Zhang, X. (2020).
Regulation of lncRNA PVT1 on mir-125 in metastasis of gastric cancer cells.
Oncol. Lett., 19(2):1261–1266.

Olivero, C. E., Martínez-Terroba, E., Zimmer, J., Liao, C., Tesfaye, E., Hooshdaran, N., Schofield, J. A., Bendor, J., Fang, D., Simon, M. D., Zamudio, J. R., and Dimitrova, N. (2020).
P53 activates the long noncoding RNA pvt1b to inhibit myc and suppress tumorigenesis.
Mol. Cell, 77(4):761–774.e8.

Olloquequi, J., Jaime, S., Parra, V., Cornejo-Córdova, E., Valdivia, G., Agustí, À., and Silva O, R. (2018).
Comparative analysis of COPD associated with tobacco smoking, biomass smoke exposure or both.
Respir. Res., 19(1):13.

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O. (2016).
Fast and efficient QTL mapper for thousands of molecular phenotypes.
Bioinformatics, 32(10):1479–1485.

Pan, W., Sun, W., Yang, S., Zhuang, H., Jiang, H., Ju, H., Wang, D., and Han, Y. (2020).
LDL-C plays a causal role on T2DM: a mendelian randomization analysis.
Aging (Albany NY), 12(3):2584–2594.

Park, E. G., Ha, H., Lee, D. H., Kim, W. R., Lee, Y. J., Bae, W. H., and Kim, H.-S. (2022).
Genomic analyses of non-coding RNAs overlapping transposable elements and its implication to human diseases.
Int. J. Mol. Sci., 23(16):8950.

Park, J.-H., Hong, J. Y., and Han, K. (2023).
Threshold dose-response association between smoking pack-years and the risk of gallbladder cancer: A nationwide cohort study.
Eur. J. Cancer, 180:99–107.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006).
Principal components analysis corrects for stratification in genome-wide association studies.
Nat. Genet., 38(8):904–909.

R Core Team (2023).
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria.

Rahman, F., Mahmud, P., Karim, R., Hossain, T., and Islam, F. (2020).
Determination of novel biomarkers and pathways shared by colorectal cancer and endometrial cancer via comprehensive bioinformatics analysis.
Inform. Med. Unlocked, 20(100376):100376.

Randi, G., Franceschi, S., and La Vecchia, C. (2006).
Gallbladder cancer worldwide: geographical distribution and risk factors.
Int. J. Cancer, 118(7):1591–1602.

Randi, G., Malvezzi, M., Levi, F., Ferlay, J., Negri, E., Franceschi, S., and La Vecchia, C. (2009).
Epidemiology of biliary tract cancers: an update.
Ann. Oncol., 20(1):146–159.

Ratti, M., Lampis, A., Ghidini, M., Salati, M., Mirchev, M. B., Valeri, N., and Hahne, J. C. (2020).
MicroRNAs (miRNAs) and long non-coding RNAs (lncRNAs) as new tools for cancer therapy: First steps from bench to bedside.
Target. Oncol., 15(3):261–278.

Raum, E., Rothenbacher, D., Löw, M., Stegmaier, C., Ziegler, H., and Brenner, H. (2007).
Changes of cardiovascular risk factors and their implications in subsequent birth cohorts of older adults in germany: a life course approach.
Eur. J. Cardiovasc. Prev. Rehabil., 14(6):809–814.

Raut, J. R., Bhardwaj, M., Schöttker, B., Holleczek, B., Schrotz-King, P., and Brenner, H. (2024).
Cancer-specific risk prediction with a serum microRNA signature.
Cancer Sci., 115(6):2049–2058.

Rawal, N., Awasthi, S., Dash, N. R., Kumar, S., Das, P., Ranjan, A., Chopra, A., Khan, M. A., Saluja, S., Hussain, S., and Tanwar, P. (2023).
Prognostic relevance of PDL1 and CA19-9 expression in gallbladder cancer vs. inflammatory lesions.
Curr. Oncol., 30(2):1571–1584.

Rawla, P., Sunkara, T., Thandra, K. C., and Barsouk, A. (2019).
Epidemiology of gallbladder cancer.
Clin. Exp. Hepatol., 5(2):93–102.

Rigden, D. J. and Fernández, X. M. (2021).
The 2021 nucleic acids research database issue and the online molecular biology database collection.
Nucleic Acids Res., 49(D1):D1–D9.

Ringnér, M. (2008).
What is principal component analysis?
Nat. Biotechnol., 26(3):303–304.

Roa, J. C., García, P., Kapoor, V. K., Maithel, S. K., Javle, M., and Koshiol, J. (2022).
Gallbladder cancer.
Nat. Rev. Dis. Primers, 8(1):69.

Romano, G., Veneziano, D., Acunzo, M., and Croce, C. M. (2017).
Small non-coding RNA and cancer.
Carcinogenesis, 38(5):485–491.

Rounge, T. B., Umu, S. U., Keller, A., Meese, E., Ursin, G., Tretli, S., Lyle, R., and Langseth, H. (2018).
Circulating small non-coding RNAs associated with age, sex, smoking, body mass and physical activity.
Sci. Rep., 8(1):17650.

Ryu, S., Chang, Y., Yun, K. E., Jung, H.-S., Shin, J. H., and Shin, H. (2016).
Gallstones and the risk of gallbladder cancer mortality: A cohort study.

Am. J. Gastroenterol., 111(10):1476–1487.

Salido-Guadarrama, I., Romero-Cordoba, S. L., and Rueda-Zarazua, B. (2023).
Multi-omics mining of lncRNAs with biological and clinical relevance in cancer.
Int. J. Mol. Sci., 24(23).

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., and Malonia, S. K. (2023).
Next-generation sequencing technology: Current trends and advancements.
Biology (Basel), 12(7).

Saxena, R., Chakrapani, B., Sarath Krishnan, M. P., Gupta, A., Gupta, S., Das, J., Gupta, S. C., Mirza, A. A., Rao, S., and Goyal, B. (2023).
Next generation sequencing uncovers multiple miRNAs associated molecular targets in gallbladder cancer patients.
Sci. Rep., 13(1):19101.

Scherer, D., Dávila López, M. D., Goeppert, B., Abrahamsson, S., González Silos, R. G., Nova, I., Marcelain, K., Roa, J. C., Ibberson, D., Umu, S. U., Rounge, T. B., Roessler, S., and Bermejo, J. L. (2020).
RNA sequencing of hepatobiliary cancer cell lines: Data and applications to mutational and transcriptomic profiling.
Cancers (Basel), 12(9):2510.

Schubert, M., Lindgreen, S., and Orlando, L. (2016).
AdapterRemoval v2: rapid adapter trimming, identification, and read merging.
BMC Res. Notes, 9(1):88.

Sedgwick, P. (2015).
A comparison of parametric and non-parametric statistical tests.
BMJ, 350(apr17 1):h2053.

Shi, G., Wu, T., Li, X., Zhao, D., Yin, Q., and Zhu, L. (2024).
Systematic genome-wide mendelian randomization reveals the causal links between miRNAs and parkinson's disease.

Front. Neurosci., 18:1385675.

Sicklick, J. K., Fanta, P. T., Shimabukuro, K., and Kurzrock, R. (2016).
Genomics of gallbladder cancer: the case for biomarker-driven clinical trial design.
Cancer Metastasis Rev., 35(2):263–275.

Signorell, A. (2024).
DescTools: Tools for Descriptive Statistics.
R package version 0.99.57.

Sinkala, M. (2023).
Mutational landscape of cancer-driver genes across human cancers.
Sci. Rep., 13(1):12742.

Slattery, M. L., Lee, F. Y., Pellatt, A. J., Mullany, L. E., Stevens, J. R., Samowitz, W. S., Wolff, R. K., and Herrick, J. S. (2017).
Infrequently expressed miRNAs in colorectal cancer tissue and tumor molecular phenotype.
Mod. Pathol., 30(8):1152–1169.

Solé, X., Guinó, E., Valls, J., Iniesta, R., and Moreno, V. (2006).
SNPStats: a web tool for the analysis of association studies.
Bioinformatics, 22(15):1928–1929.

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S. A., Inouye, M., Parkinson, H., and Harris, L. W. (2023).
The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource.
Nucleic Acids Res., 51(D1):D977–D985.

Soneson, C. and Delorenzi, M. (2013).
A comparison of methods for differential expression analysis of RNA-seq data.
BMC Bioinformatics, 14(1):91.

Srivastava, P., Mishra, S., Agarwal, A., Pandey, A., and Husain, N. (2023).
Circulating microRNAs in gallbladder cancer: Is serum assay of diagnostic value?
Pathol. Res. Pract., 242(154320):154320.

Stahlhut, C. and Slack, F. J. (2015).
Combinatorial action of MicroRNAs let-7 and mir-34 effectively synergizes with erlotinib to suppress non-small cell lung cancer cell proliferation.
Cell Cycle, 14(13):2171–2180.

Stang, A., Moebus, S., Dragano, N., Beck, E. M., Möhlenkamp, S., Schmermund, A., Siegrist, J., Erbel, R., Jöckel, K. H., and Heinz Nixdorf Recall Study Investigation Group (2005).
Baseline recruitment and analyses of nonresponse of the heinz nixdorf recall study: identifiability of phone numbers as the major determinant of response.
Eur. J. Epidemiol., 20(6):489–496.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021).
Gene regulation by long non-coding RNAs and its biological functions.
Nat. Rev. Mol. Cell Biol., 22(2):96–118.

Stein, A., Arnold, D., Bridgewater, J., Goldstein, D., Jensen, L. H., Klümpen, H.-J., Lohse, A. W., Nashan, B., Primrose, J., Schrum, S., Shannon, J., Vettorazzi, E., and Wege, H. (2015).
Adjuvant chemotherapy with gemcitabine and cisplatin compared to observation after curative intent resection of cholangiocarcinoma and muscle invasive gallbladder carcinoma (ACTICCA-1 trial) - a randomized, multidisciplinary, multinational phase III trial.
BMC Cancer, 15(1):564.

Stinton, L. M. and Shaffer, E. A. (2012).
Epidemiology of gallbladder disease: cholelithiasis and cancer.
Gut Liver, 6(2):172–187.

Sun, Y., Wang, W., Tang, Y., Wang, D., Li, L., Na, M., Jiang, G., Li, Q., Chen, S., and Zhou, J. (2020).

Microarray profiling and functional analysis of differentially expressed plasma exosomal circular RNAs in graves' disease.
Biol. Res., 53(1):32.

Sun, Y.-M., Lin, K.-Y., and Chen, Y.-Q. (2013).
Diverse functions of mir-125 family in different cell contexts.
J. Hematol. Oncol., 6(1):6.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-B., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., Shetty, A. C., Blackwell, T. W., Smith, A. V., Wong, Q., Liu, X., Conomos, M. P., Bobo, D. M., Aguet, F., Albert, C., Alonso, A., Ardlie, K. G., Arking, D. E., Aslibekyan, S., Auer, P. L., Barnard, J., Barr, R. G., Barwick, L., Becker, L. C., Beer, R. L., Benjamin, E. J., Bielak, L. F., Blangero, J., Boehnke, M., Bowden, D. W., Brody, J. A., Burchard, E. G., Cade, B. E., Casella, J. F., Chalazan, B., Chasman, D. I., Chen, Y.-D. I., Cho, M. H., Choi, S. H., Chung, M. K., Clish, C. B., Correa, A., Curran, J. E., Custer, B., Darbar, D., Daya, M., de Andrade, M., DeMeo, D. L., Dutcher, S. K., Ellinor, P. T., Emery, L. S., Eng, C., Fatkin, D., Fingerlin, T., Forer, L., Fornage, M., Franceschini, N., Fuchsberger, C., Fullerton, S. M., Germer, S., Gladwin, M. T., Gottlieb, D. J., Guo, X., Hall, M. E., He, J., Heard-Costa, N. L., Heckbert, S. R., Irvin, M. R., Johnsen, J. M., Johnson, A. D., Kaplan, R., Kardia, S. L. R., Kelly, T., Kelly, S., Kenny, E. E., Kiel, D. P., Klemmer, R., Konkle, B. A., Kooperberg, C., Köttgen, A., Lange, L. A., Lasky-Su, J., Levy, D., Lin, X., Lin, K.-H., Liu, C., Loos, R. J. F., Garman, L., Gerszten, R., Lubitz, S. A., Lunetta, K. L., Mak, A. C. Y., Manichaikul, A., Manning, A. K., Mathias, R. A., McManus, D. D., McGarvey, S. T., Meigs, J. B., Meyers, D. A., Mikulla, J. L., Minear, M. A., Mitchell, B. D., Mohanty, S., Montasser, M. E., Montgomery, C., Morrison, A. C., Murabito, J. M., Natale, A., Natarajan, P., Nelson, S. C., North, K. E., O'Connell, J. R., Palmer, N. D., Pankratz, N., Peloso, G. M., Peyser, P. A., Pleiness, J., Post, W. S., Psaty, B. M., Rao, D. C., Redline, S., Reiner, A. P., Roden, D., Rotter, J. I., Ruczinski, I., Sarnowski, C., Schoenherr, S., Schwartz, D. A., Seo, J.-S., Seshadri, S., Sheehan, V. A., Sheu, W. H., Shoemaker, M. B., Smith, N. L., Smith, J. A., Sotoodehnia, N., Stilp, A. M., Tang, W.,

Taylor, K. D., Telen, M., Thornton, T. A., Tracy, R. P., Van Den Berg, D. J., Vasan, R. S., Viaud-Martinez, K. A., Vrieze, S., Weeks, D. E., Weir, B. S., Weiss, S. T., Weng, L.-C., Willer, C. J., Zhang, Y., Zhao, X., Arnett, D. K., Ashley-Koch, A. E., Barnes, K. C., Boerwinkle, E., Gabriel, S., Gibbs, R., Rice, K. M., Rich, S. S., Silverman, E. K., Qasba, P., Gan, W., NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Papanicolaou, G. J., Nickerson, D. A., Browning, S. R., Zody, M. C., Zöllner, S., Wilson, J. G., Cupples, L. A., Laurie, C. C., Jaquish, C. E., Hernandez, R. D., O'Connor, T. D., and Abecasis, G. R. (2021).
Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program.
Nature, 590(7845):290–299.

Tan, W., Gao, M., Liu, N., Zhang, G., Xu, T., and Cui, W. (2015).
Body mass index and risk of gallbladder cancer: Systematic review and meta-analysis of observational studies.
Nutrients, 7(10):8321–8334.

Tang, Y., Feng, H., Zhang, L., Qu, C., Li, J., Deng, X., Zhong, S., Yang, J., Deng, X., Zeng, X., Wang, Y., and Peng, X. (2022).
A novel prognostic model for cutaneous melanoma based on an immune-related gene signature and clinical variables.
Sci. Rep., 12(1):20374.

The International HapMap Consortium (2007).
A second generation human haplotype map of over 3.1 million SNPs.
Nature, 449(7164):851–861.

Tian, X., Wu, Y., Yang, Y., Wang, J., Niu, M., Gao, S., Qin, T., and Bao, D. (2020).
Long noncoding RNA LINC00662 promotes M2 macrophage polarization and hepatocellular carcinoma progression via activating Wnt/β-catenin signaling.
Mol. Oncol., 14(2):462–483.

Tripathi, R., Chakraborty, P., and Varadwaj, P. K. (2017).
Unraveling long non-coding RNAs through analysis of high-throughput RNA-sequencing data.

Noncoding RNA Res., 2(2):111–118.

Tsyganov, M. M. and Ibragimova, M. K. (2023).
MALAT1 long non-coding RNA and its role in breast carcinogenesis.
Acta Naturae, 15(2):32–41.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021).
Genome-wide association studies.
Nat. Rev. Methods Primers, 1(1).

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015).
Proteomics. tissue-based map of the human proteome.
Science, 347(6220):1260419.

Ujfaludi, Z., Fazekas, F., Biró, K., Oláh-Németh, O., Buzogany, I., Sükösd, F., Beöthe, T., and Pankotai, T. (2024).
mir-21, mir-29a, and mir-106b: serum and tissue biomarkers with diagnostic potential in metastatic testicular cancer.
Sci. Rep., 14(1):20151.

Umu, S. U., Langseth, H., Bucher-Johannessen, C., Fromm, B., Keller, A., Meese, E., Lauritzen, M., Leithaug, M., Lyle, R., and Rounge, T. B. (2018).
A comprehensive profile of circulating RNAs in human serum.
RNA Biol., 15(2):242–250.

Uppaluri, K. R., Challa, H. J., Gaur, A., Jain, R., Krishna Vardhani, K., Geddam, A., Natya, K., Aswini, K., Palasamudram, K., and K, S. M. (2023).
Unlocking the potential of non-coding RNAs in cancer research and therapy.

Transl. Oncol., 35(101730):101730.

Venables, W. N. and Ripley, B. D. (2002).
Modern Applied Statistics with S.
Springer, New York, fourth edition.
ISBN 0-387-95457-0.

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A. G. (2012).
TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support.
Nucleic Acids Res., 40(Database issue):D222–9.

Viechtbauer, W. (2010).
Conducting meta-analyses in R with the metafor package.
Journal of Statistical Software, 36(3):1–48.

Villanueva, L. (2016).
Cancer of the gallbladder-chilean statistics.
Ecancermedicalscience, 10:704.

Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., Dalamagas, T., and Hatzigeorgiou, A. G. (2015).
DIANA-miRPath v3.0: deciphering microRNA function with experimental support.
Nucleic Acids Res., 43(W1):W460–6.

Wang, H., Guo, M., Wei, H., and Chen, Y. (2023).
Targeting p53 pathways: mechanisms, structures, and advances in therapy.
Signal Transduct. Target. Ther., 8(1):92.

Wang, Z., Gerstein, M., and Snyder, M. (2009).
RNA-Seq: a revolutionary tool for transcriptomics.
Nat. Rev. Genet., 10(1):57–63.

Wickham, H. (2016).

ggplot2: Elegant Graphics for Data Analysis.
Springer-Verlag New York.

Winter, J., Jung, S., Keller, S., Gregory, R. I., and Diederichs, S. (2009).
Many roads to maturity: microRNA biogenesis pathways and their regulation.
Nat. Cell Biol., 11(3):228–234.

Wistuba, I. I. and Gazdar, A. F. (2004).
Gallbladder cancer: lessons from a rare tumour.
Nat. Rev. Cancer, 4(9):695–706.

Wu, R., Li, L., Bai, Y., Yu, B., Xie, C., Wu, H., Zhang, Y., Huang, L., Yan, Y., Li, X., and Lin, C. (2020).
The long noncoding RNA LUCAT1 promotes colorectal cancer cell proliferation by antagonizing nucleolin to regulate MYC expression.
Cell Death Dis., 11(10):908.

Wysocki, P. T., Czubak, K., Marusiak, A. A., Kolanowska, M., and Nowis, D. (2023).
lncRNA DIRC3 regulates invasiveness and insulin-like growth factor signaling in thyroid cancer cells.
Endocr. Relat. Cancer, 30(8).

Xia, X.-Q., Lu, W.-L., Ye, Y.-Y., and Chen, J. (2020).
LINC00662 promotes cell proliferation, migration and invasion of melanoma by sponging mir-890 to upregulate ELK3.
Eur. Rev. Med. Pharmacol. Sci., 24(16):8429–8438.

Xiao, Y., Xia, Y., Wang, Y., and Xue, C. (2023).
Pathogenic roles of long noncoding RNAs in melanoma: Implications in diagnosis and therapies.
Genes Dis., 10(1):113–125.

Xie, W., Wang, Z., Wang, J., Wang, X., and Guan, H. (2024).
Investigating the molecular mechanisms of microRNAâ€‘409â€‘3p in tumor progression: Towards targeted therapeutics (review).

Int. J. Oncol., 65(1).

Xing, C., Sun, S.-G., Yue, Z.-Q., and Bai, F. (2021).
Role of lncRNA LUCAT1 in cancer.
Biomed. Pharmacother., 134(111158):111158.

Xu, J., Wang, X., Zhu, C., and Wang, K. (2022).
A review of current evidence about lncRNA MEG3: A tumor suppressor in multiple cancers.
Front. Cell Dev. Biol., 10:997633.

Xue, L., Guo, C., Zhang, K., Jiang, H., Pang, F., Dou, Y., Liu, X., Lin, H., Dong, X., Zhao, S., Yao, M., Wang, K., Feng, Y., and Gu, W. (2019).
Comprehensive molecular profiling of extrahepatic cholangiocarcinoma in chinese population and potential targets for clinical practice.
Hepatobiliary Surg. Nutr., 8(6):615–622.

Yamada, H., Penney, K. L., Takahashi, H., Katoh, T., Yamano, Y., Yamakado, M., Kimura, T., Kuruma, H., Kamata, Y., Egawa, S., and Freedman, M. L. (2009).
Replication of prostate cancer risk loci in a japanese case-control association study.
J. Natl. Cancer Inst., 101(19):1330–1336.

Yan, D., Wang, P., Knudsen, B. S., Linden, M., and Randolph, T. W. (2012).
Statistical methods for tissue array images - algorithmic scoring and co-training.
Ann. Appl. Stat., 6(3):1280–1305.

Yang, D., Zhan, M., Chen, T., Chen, W., Zhang, Y., Xu, S., Yan, J., Huang, Q., and Wang, J. (2017).
mir-125b-5p enhances chemotherapy sensitivity to cisplatin by down-regulating bcl2 in gallbladder cancer.
Sci. Rep., 7:43109.

Yang, M., Wei, S., Zhao, H., Zhou, D., and Cui, X. (2021).
The role of miRNA125b in the progression of hepatocellular carcinoma.
Clin. Res. Hepatol. Gastroenterol., 45(5):101712.

Yang, S., Wang, X., Zhou, X., Hou, L., Wu, J., Zhang, W., Li, H., Gao, C., and Sun, C. (2023).
ncRNA-mediated ceRNA regulatory network: Transcriptomic insights into breast cancer progression and treatment strategies.
Biomed. Pharmacother., 162(114698):114698.

Yang, T., Li, Y., Zheng, Z., Qu, P., Shao, Z., Wang, J., Ding, N., and Wang, W. (2024).
Comprehensive analysis of lncRNA-mediated ceRNA network in renal cell carcinoma based on GEO database.
Medicine (Baltimore), 103(35):e39424.

Ye, Y.-Y., Mei, J.-W., Xiang, S.-S., Li, H.-F., Ma, Q., Song, X.-L., Wang, Z., Zhang, Y.-C., Liu, Y.-C., Jin, Y.-P., Hu, Y.-P., Jiang, L., Liu, F.-T., Zhang, Y.-J., Hao, Y.-J., and Liu, Y.-B. (2018).
MicroRNA-30a-5p inhibits gallbladder cancer cell proliferation, migration and metastasis by targeting E2F7.
Cell Death Dis., 9(3):410.

Yu, C., Yao, W., and Bai, X. (2014).
Robust linear regression: A review and comparison.

Zhang, J., Liu, X., Yu, G., Liu, L., Wang, J., Chen, X., Bian, Y., Ji, Y., Zhou, X., Chen, Y., Ji, J., Xiang, Z., Guo, L., Fang, J., Sun, Y., Cao, H., Zhu, Z., and Yu, Y. (2018).
UBE2C is a potential biomarker of intestinal-type gastric cancer with chromosomal instability.
Front. Pharmacol., 9:847.

Zhang, S., Wang, Y., Jia, L., Wen, X., Du, Z., Wang, C., Hao, Y., Yu, D., Zhou, L., Chen, N., Chen, J., Chen, H., Zhang, H., Celik, I., Gülsoy, G., Luo, J., Qin, B., Cui, X., Liu, Z., Zhang, S., Esteban, M. A., Ay, F., Xu, W., Chen, R., Li, W., Hoffman, A. R., Hu, J.-F., and Cui, J. (2019).
Profiling the long noncoding RNA interaction network in the regulatory elements of target genes by chromatin in situ reverse transcription sequencing.
Genome Res., 29(9):1521–1532.

Zhang, Y., Zhou, Y., Zhu, W., Liu, J., and Cheng, F. (2022).
Non-coding RNAs fine-tune the balance between plant growth and abiotic stress tolerance.
Front. Plant Sci., 13:965745.

Zhao, S., Li, C.-I., Guo, Y., Sheng, Q., and Shyr, Y. (2018).
RnaSeqSampleSize: real data based sample size estimation for RNA sequencing.
BMC Bioinformatics, 19(1).

Zheng, Q., Zhu, Q., Li, C., Hao, S., Li, J., Yu, X., Qi, D., and Pan, Y. (2020).
microRNA-144 functions as a diagnostic and prognostic marker for retinoblastoma.
Clinics (Sao Paulo), 75(e1804):e1804.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012).
A high-performance computing toolset for relatedness and principal component analysis of SNP data.
Bioinformatics, 28(24):3326–3328.

Zhou, Y., Yuan, K., Yang, Y., Ji, Z., Zhou, D., Ouyang, J., Wang, Z., Wang, F., Liu, C., Li, Q., Zhang, Q., Li, Q., Shan, X., and Zhou, J. (2023).
Gallbladder cancer: current and future treatment options.
Front. Pharmacol., 14:1183619.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M., and Siatkowski, I. (2015).
The impact of normalization methods on RNA-seq data analysis.
Biomed Res. Int., 2015:621690.

Appendix A: Additional Tables and Figures

A.1 Identification of circulating long non-coding RNAs associated with gallbladder cancer risk

Comment: Parts of the following Chapter have already been published in Cancers (Blandino et al., 2022). The original manuscript was written by myself, but also contains comments and corrections from the co-authors.

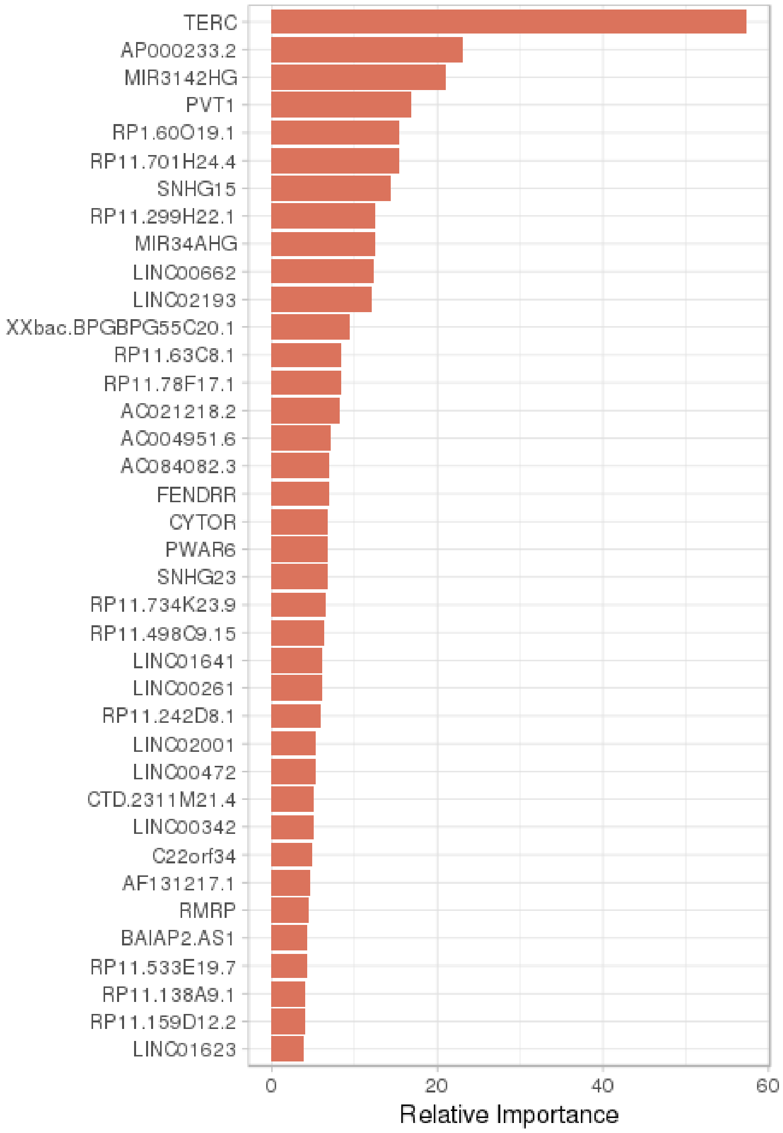

Figure A.1: *39 high-quality preselected long non-coding RNA candidates using machine learning, ordered by relative importance.*

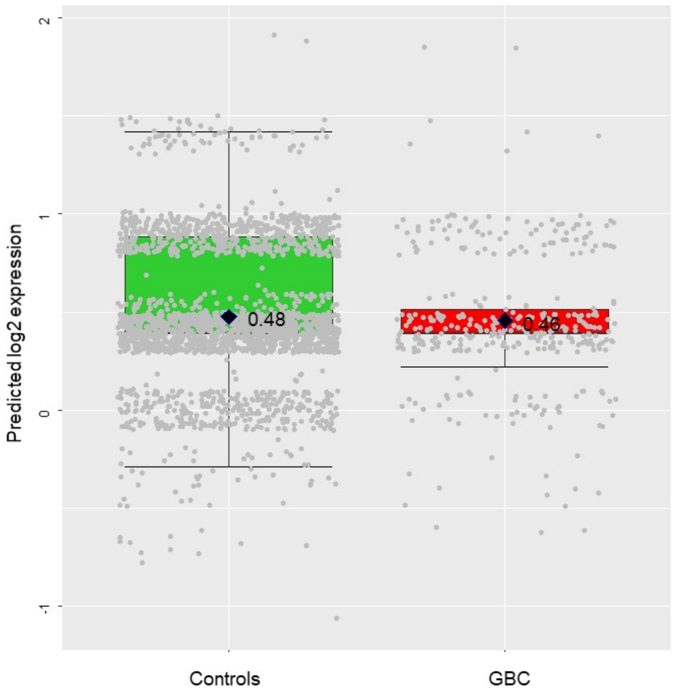

Figure A.2: *Genotype-based expression of C22orf34 in the long non-coding RNA-gallbladder cancer association dataset. GBC: gallbladder cancer. Rhombuses indicate the average log2 expression in cases and controls, respectively.*

Table A.1: *Preselected long non-coding RNAs based on Jonckheere-Terpstra tests.*

lncRNA	**p-value**	**Log2 expression GS**	**OR Dys**	**OR GBC**
		Median [5th;95th]	**Estimate [95% CI]**	**Estimate [95% CI]**
AC084082.3	0.009	8.23 [1.45 - 9.93]	2.10 [0.86 - 5.11]	1.39 [1.04 - 1.85]
FAM95B1	0.009	1.44 [0.48 - 2.49]	0.15 [0.03 - 0.78]	0.13 [0.03 - 0.64]
HCG11	0.009	1.50 [0.65 - 2.78]	3.01 [1.14 - 7.99]	2.99 [1.12 - 7.96]
LINC00472	0.009	1.48 [0.63 - 2.53]	0.89 [0.43 - 1.88]	0.11 [0.02 - 0.64]
LINC00662	0.009	1.48 [0.55 - 4.38]	2.73 [1.41 - 5.30]	2.00 [1.12 - 3.58]
LINC00869	0.009	2.62 [0.92 - 3.97]	2.41 [1.19 - 4.85]	3.35 [1.48 - 7.56]
MIR155HG	0.009	7.66 [1.47 - 9.73]	1.55 [1.03 - 2.31]	2.33 [1.29 - 4.18]
MIR3142HG	0.009	10.56 [3.42 - 13.29]	1.31 [0.94 - 1.84]	3.14 [1.30 - 7.59]
PVT1	0.009	1.02 [0.45 - 1.75]	0.43 [0.10 - 1.86]	4.36 [0.87 - 21.86]
PWAR6	0.009	1.65 [0.85 - 3.37]	0.70 [0.33 - 1.47]	0.26 [0.08 - 0.80]
RP1.60O19.1	0.009	3.02 [0.92 - 5.32]	1.04 [0.67 - 1.64]	0.58 [0.34 â€“ 1.00]
RP11.701H24.4	0.009	11.03 [1.47 - 12.47]	1.07 [0.87 - 1.32]	0.31 [0.11 - 0.84]
RP4.561L24.3	0.009	6.76 [1.44 - 8.87]	2.23 [1.24 - 4.03]	2.63 [1.30 - 5.34]
TERC	0.009	1.50 [0.73 - 2.85]	2.60 [1.17 - 5.78]	3.61 [1.53 - 8.55]
LL0XNC01.237H1.2	0.02	1.02 [0.45 - 1.96]	2.11 [0.90 - 4.93]	3.14 [1.13 - 8.73]
RP11.78F17.1	0.02	1.20 [0.50 - 1.82]	0.20 [0.04 - 0.98]	0.09 [0.02 - 0.52]
FENDRR	0.02	1.49 [0.82 - 2.88]	1.99 [0.75 - 5.26]	0.13 [0.02 - 0.71]
LINC00261	0.02	2.07 [0.54 - 4.41]	1.04 [0.64 - 1.67]	0.45 [0.22 - 0.90]
LINC02001	0.03	4.30 [1.20 - 6.60]	1.86 [1.21 - 2.86]	1.68 [1.12 - 2.50]
RP11.498C9.15	0.03	0.98 [0.46 - 1.59]	1.60 [0.58 - 4.45]	2.29 [0.73 - 7.15]
RP11.170M17.1	0.03	1.44 [0.45 - 4.27]	0.70 [0.38 - 1.29]	0.14 [0.02 - 0.77]
SNHG9	0.03	2.55 [1.09 - 4.33]	2.10 [1.07 - 4.13]	3.50 [1.43 - 8.60]
MEG3	0.03	3.69 [1.44 - 6.23]	0.95 [0.60 - 1.50]	0.39 [0.18 - 0.83]
RP6.74O6.2	0.03	1.46 [0.50 - 2.79]	0.77 [0.36 - 1.62]	0.51 [0.21 - 1.26]
RP1.140K8.5	0.04	1.49 [0.59 - 3.05]	1.02 [0.55 - 1.90]	0.34 [0.11 - 1.02]
RP11.304L19.13	0.04	1.44 [0.52 - 3.16]	1.43 [0.69 - 2.96]	2.77 [1.17 - 6.56]
CTD.2311M21.4	0.04	1.42 [0.45 - 2.79]	0.00 [0.00 - 0.15]	0.25 [0.06 - 1.02]
CTD.2626G11.2	0.04	1.44 [0.50 - 2.20]	0.17 [0.04 - 0.86]	0.42 [0.10 - 1.75]
OLMALINC	0.04	1.48 [0.51 - 2.91]	0.76 [0.35 - 1.64]	0.29 [0.09 - 0.95]
C22orf34	0.04	1.44 [0.48 - 3.68]	0.28 [0.08 - 1.07]	0.36 [0.10 - 1.28]
CTD.2210P24.2	0.04	1.46 [0.61 - 4.85]	0.85 [0.45 - 1.63]	2.52 [1.10 - 5.77]
MIR34AHG	0.04	6.35 [1.44 - 9.78]	1.60 [1.13 - 2.28]	2.02 [1.23 - 3.34]
CYTOR	0.04	1.44 [0.48 - 2.18]	0.85 [0.33 - 2.16]	2.27 [0.59 - 8.70]
RP11.714M23.2	0.04	1.44 [0.51 - 2.26]	0.91 [0.46 - 1.80]	0.36 [0.10 - 1.35]

lncRNA: long non-coding RNA; *p-value*: probability value; $5th$;$95th$: 5th and 95th percentiles; *GS*: gallstones; *OR*: odds ratio; *GBC*: gallbladder cancer; *CI*: confidence interval.

Table A.2: *Identified and validated cis-expression quantitative trait loci for LINC00662 and C22orf34.*

LINC00662

SNP ID	**Location**	**MAF**	**Model**	$\beta1$	**p-value1**	$\beta2$	**p-value2**
rs11083486	chr19:28407449:G:T	0.31	Additive	-0.74	0.01	-	-
rs11083486	chr19:28407449:G:T	0.31	Three-Geno	-0.96	0.03	-1.57	0.01
rs11083486	chr19:28407449:G:T	0.31	Dominant	-0.86	0.03	-	-
rs11083486	chr19:28407449:G:T	0.31	Recessive	1.29	0.03	-	-
rs142521755	chr19:27284894:T:A	0.07	Dominant	1.08	0.04	-	-

C22orf34

SNP ID	**Location**	**MAF**	**Model**	$\beta1$	**p-value1**	$\beta2$	**p-value2**
rs5770650	chr22:49683714:A:C	0.13	Additive	0.48	0.01	-	-
rs9628049	chr22:49551343:C:T	0.06	Additive	-0.60	0.02	-	-
rs5770650	chr22:49683714:A:C	0.13	Dominant	0.52	0.01	-	-
rs9628049	chr22:49551343:C:T	0.06	Dominant	-0.60	0.02	-	-
rs80641	chr22:49548950:G:T	0.11	Three-Geno	-2.19	0.006	-1.99	0.01
rs135786	chr22:49550809:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135787	chr22:49550871:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135788	chr22:49551103:T:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135789	chr22:49551309:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135791	chr22:49552575:C:T	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135792	chr22:49553166:G:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135793	chr22:49553257:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135794	chr22:49553508:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135796	chr22:49554141:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135797	chr22:49554220:G:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135798	chr22:49554437:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135799	chr22:49554674:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135800	chr22:49555086:C:T	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135801	chr22:49555128:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs8140696	chr22:49555464:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs8140728	chr22:49555542:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs8140866	chr22:49555658:A:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs1054180151	chr22:49555702:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135803	chr22:49555956:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009

rs135804	chr22:49556003:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135805	chr22:49556247:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135806	chr22:49556251:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135807	chr22:49556406:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135810	chr22:49557021:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135811	chr22:49557199:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135812	chr22:49557423:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135813	chr22:49557486:A:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135814	chr22:49557526:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs9627745	chr22:49557770:C:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs35356406	chr22:49558924:G:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135815	chr22:49559001:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135816	chr22:49559524:C:T	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135817	chr22:49560766:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135821	chr22:49562360:T:G	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs13055340	chr22:49562667:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs1661563636	chr22:49562872:C:T	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs112515352	chr22:49563159:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135822	chr22:49563851:T:C	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135823	chr22:49564023:G:A	0.12	Three-Geno	-2.25	0.004	-1.98	0.009
rs135826	chr22:49565810:G:A	0.11	Three-Geno	-2.25	0.004	-1.98	0.009
rs6009823	chr22:49692686:C:T	0.12	Three-Geno	1.58	0.004	0.96	0.04
rs6009824	chr22:49692725:G:A	0.12	Three-Geno	1.58	0.004	0.96	0.04

SNP: single nucleotide polymorfism; *MAF*: minor allele frequency; *p-value*: probability value; *chr*: chromosome.

A.2 Identification and validation of circulating microRNAs associated with gallbladder cancer risk

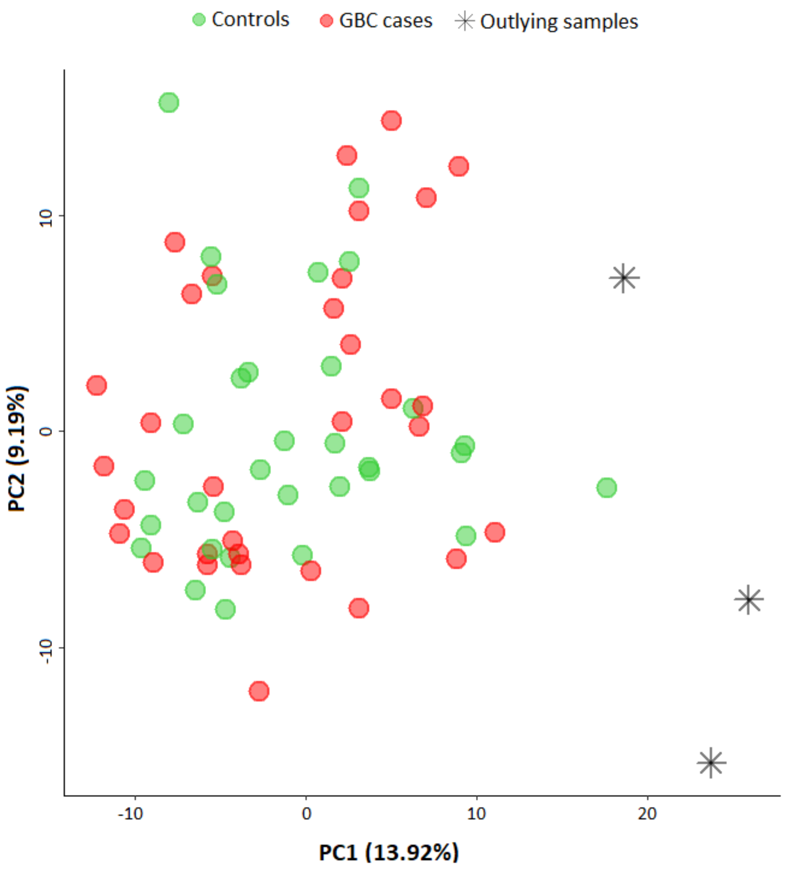

Figure A.3: *Global microRNA expression profiles in the validation dataset. GBC: gallbladder cancer; PC: principal component.*

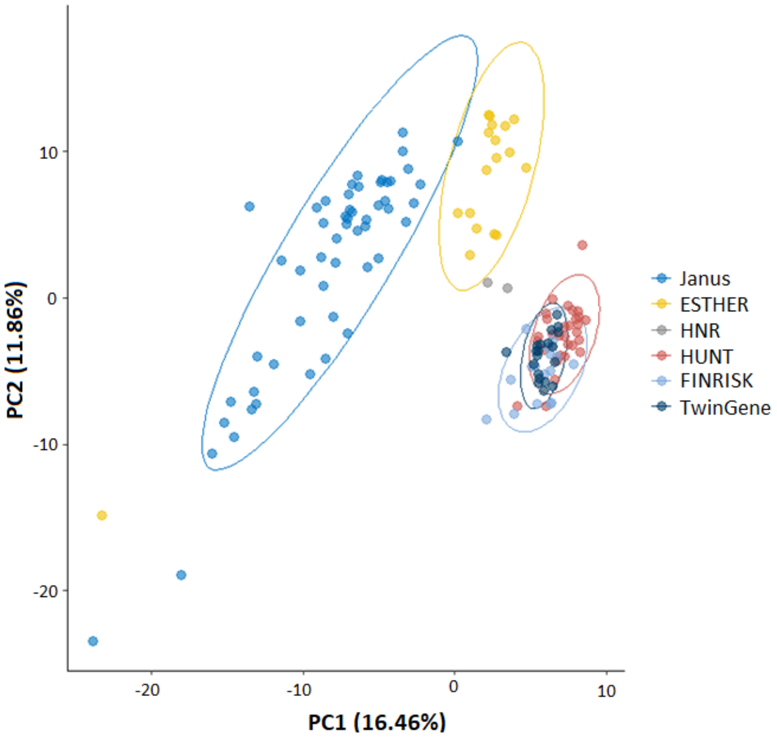

Figure A.4: *Global microRNA expression profiles in all the investigated cohorts. PC: principal component; ESTHER: Early detection and optimised therapy of chronic diseases in the elderly population; HNR: Heinz Nixdorf recall study; HUNT: Nord-Trøndelag Health study.*

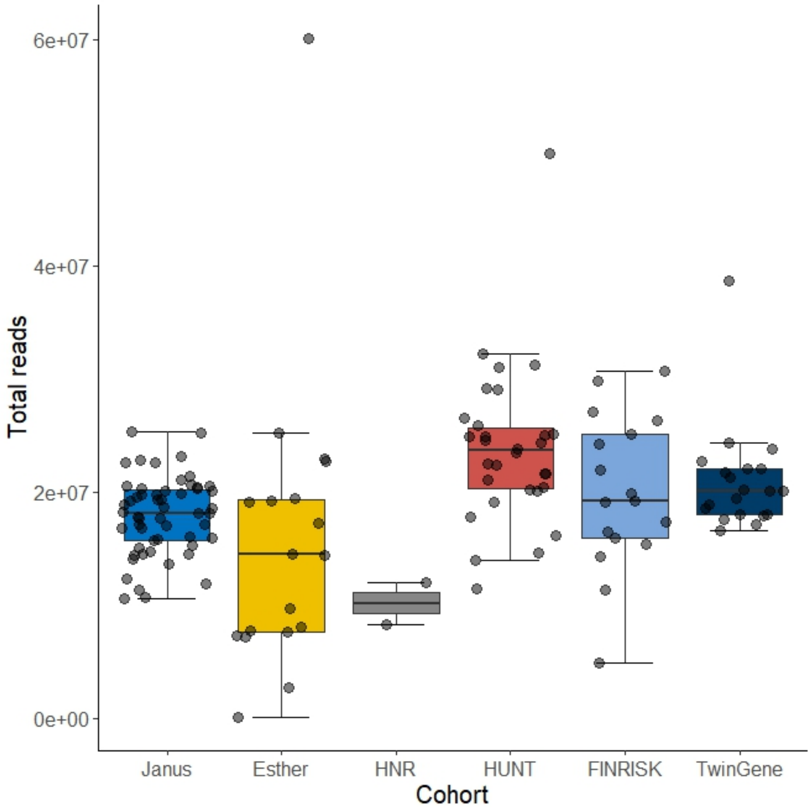

Figure A.5: *Boxplots for the total number of reads and the number of microRNAs for the investigated cohorts. ESTHER: Early detection and optimised therapy of chronic diseases in the elderly population; HNR: Heinz Nixdorf recall study; HUNT: Nord-Trøndelag Health study.*

Table A.3: *List of preselected microRNAs based on formalin-fixed paraffin-embedded tissue samples also expressed in serum samples from the screening dataset.*

	Preselection		***Screening***	
miRNA	**Expression in controls Median [5th;95th]**	**Case-control Difference [95% CI]**	**Expression in controls Median [5th;95th]**	**Case-control Difference [95% CI]**
miR-204-5p	7.49 [6.59;8.74]	-2.61 [-2.87;-2.35]	7.06 [5.16;7.95]	0.6 [-0.2;1.41]
miR-495-3p	7.64 [6.7;8.39]	-2.53 [-2.84;-2.23]	9.91 [8.82;10.78]	1.65 [0.41;2.9]
miR-379-5p	6.71 [6.26;7.19]	-1.48 [-1.7;-1.26]	9.72 [9.14;9.85]	1.14 [-0.27;2.55]
miR-1224-5p	8.73 [8.35;9.57]	2.01 [1.71;2.3]	9.75 [8.79;10.66]	0.37 [-0.62;1.35]
miR-136-3p	6.89 [5.95;7.64]	-1.92 [-2.2;-1.63]	5.51 [5.2;5.67]	0.13 [-0.87;1.12]
miR-29c-5p	8.13 [7.62;9.06]	-2.28 [-2.65;-1.9]	6.62 [5.93;7.78]	0.61 [-0.23;1.46]
miR-381-3p	7.9 [7.27;8.49]	-2.2 [-2.57;-1.83]	9.98 [9.45;10.66]	1.22 [0.01;2.44]
miR-145-3p	8.85 [8.24;9.41]	-3.11 [-3.64;-2.58]	9.82 [8.08;10.09]	0.14 [-1.37;1.65]
miR-144-3p	8.98 [7.8;10.45]	-3.65 [-4.26;-3.04]	5.92 [5.22;6.47]	0.01 [-1.31;1.33]
miR-411-5p	6.52 [5.85;7.24]	-1.67 [-1.94;-1.41]	5.94 [5.41;6.23]	0.49 [-0.57;1.54]
miR-654-3p	6.99 [6.41;7.59]	-1.49 [-1.76;-1.22]	5.83 [5.52;6.12]	1.03 [-0.29;2.34]
miR-126-5p	8.29 [7.73;8.97]	-2.38 [-2.83;-1.94]	7.8 [7.41;8.2]	0.4 [-0.56;1.36]
miR-136-5p	8 [6.94;8.73]	-2.28 [-2.7;-1.85]	6.25 [5.73;6.65]	1.24 [0.22;2.27]
miR-4497	9.27 [9.02;10.32]	2.19 [1.78;2.61]	6.53 [6.21;7.66]	-0.68 [-1.57;0.21]
miR-493-5p	6.14 [5.68;6.55]	-1.06 [-1.27;-0.85]	5.02 [4.88;5.35]	1.48 [0.18;2.78]
miR-4443	9.79 [8.9;12.61]	2.68 [2.16;3.21]	5.67 [5.19;5.81]	-0.09 [-1.11;0.93]
miR-32-5p	6.28 [5.22;7.4]	-1.37 [-1.63;-1.11]	6.92 [6.05;7.55]	1.23 [-0.03;2.49]
miR-382-5p	6.97 [6.5;7.66]	-1.42 [-1.72;-1.12]	5.79 [5.49;6.4]	1.29 [0.14;2.44]
miR-30e-3p	7.66 [6.54;8.23]	-1.95 [-2.37;-1.52]	6.09 [5.52;6.83]	0.66 [-0.35;1.68]
miR-30a-3p	8.14 [7;8.82]	-2.28 [-2.78;-1.79]	5.7 [5.14;6.49]	-0.21 [-0.97;0.55]
miR-10b-3p	6.19 [5.09;6.86]	1.74 [1.37;2.11]	8.73 [8.35;9.57]	0.26 [-0.57;1.09]
miR-3679-5p	9.66 [9.14;10.58]	1.3 [1.02;1.58]	7.77 [7.19;8.11]	0.17 [-0.75;1.1]
miR-6126	8.64 [8.07;9]	1.34 [1.04;1.65]	8.44 [7.85;8.93]	-0.7 [-1.45;0.04]
miR-99b-3p	6.05 [5.7;6.95]	1.4 [1.09;1.71]	8.06 [7.74;8.67]	0.72 [-0.16;1.61]
miR-320b	9.95 [9.52;10.21]	-1.19 [-1.46;-0.92]	11.45 [10.85;11.96]	0.61 [-0.09;1.31]
miR-505-3p	6.43 [6.18;6.77]	-0.98 [-1.2;-0.75]	8.15 [7.68;8.93]	1.54 [0.57;2.52]
miR-361-3p	7.99 [6.68;8.39]	-2 [-2.47;-1.53]	5.19 [4.95;5.55]	0.13 [-0.52;0.79]
miR-484	7.45 [6.91;7.7]	-1.22 [-1.5;-0.93]	6.36 [5.58;7.44]	1.01 [-0.24;2.27]
miR-127-3p	9.12 [8.45;9.56]	-2.58 [-3.21;-1.95]	14.42 [13.85;14.74]	1.38 [0.49;2.27]

miR-4508	5.65 [5.09;5.97]	1.24 [0.94;1.54]	7 [6.61;7.81]	0.08 [-0.84;1.01]
miR-99a-5p	12.4 [11.85;12.68]	-4.16 [-5.19;-3.13]	6.04 [5.57;6.27]	0.74 [0;1.48]
miR-877-5p	5.79 [5.5;6.18]	1.13 [0.85;1.41]	7.52 [6.87;8.1]	0.18 [-0.83;1.19]
miR-338-5p	5.94 [5.41;6.23]	0.93 [0.7;1.16]	8.17 [7.47;9.17]	0.58 [-0.68;1.84]
miR-150-5p	10.27 [9.24;10.8]	-2.86 [-3.61;-2.12]	11.48 [9.35;12.9]	0.69 [-0.19;1.56]
miR-451a	13.94 [12.34;15.23]	-5.46 [-6.86;-4.07]	9.61 [8.68;10.01]	1.27 [-0.42;2.96]
miR-143-3p	11.21 [10.32;11.97]	-3.84 [-4.85;-2.84]	7.84 [6.92;8.36]	-0.27 [-1.31;0.76]
miR-340-5p	7.22 [6.59;8.29]	-1.65 [-2.08;-1.22]	8.25 [7.79;8.92]	0.6 [-0.59;1.8]
miR-100-5p	12.21 [11.99;12.51]	-3.54 [-4.52;-2.55]	11.22 [10.31;11.6]	0.32 [-0.45;1.09]
miR-140-5p	8.47 [7.85;9.31]	-2.33 [-2.95;-1.7]	6.26 [5.41;7.17]	-0.08 [-1.2;1.03]
miR-342-3p	10.1 [9.49;10.37]	-2.17 [-2.77;-1.58]	5.62 [5.12;6.42]	1.2 [0.15;2.25]
miR-140-3p	9.1 [8.36;9.66]	-2.25 [-2.86;-1.65]	6.09 [5.97;6.92]	-0.55 [-1.49;0.4]
miR-660-5p	8.2 [7.03;9.18]	-2.09 [-2.66;-1.52]	7.49 [6.93;8.04]	-0.01 [-1.15;1.12]
miR-1268a	10.2 [9.42;10.6]	1.45 [1.05;1.86]	10.46 [9.69;10.53]	-0.75 [-1.61;0.11]
miR-142-5p	8.39 [7.79;9.7]	-2.01 [-2.58;-1.44]	6.7 [6.02;7.32]	0.89 [-0.24;2.03]
miR-3925-5p	6.08 [5.16;7.36]	2.07 [1.5;2.65]	7.27 [6.16;7.69]	-3.35 [-4.33;-2.37]
miR-186-5p	8.12 [7.85;8.53]	-1.53 [-1.98;-1.09]	9.68 [9.06;10.73]	0.71 [-0.25;1.67]
miR-185-5p	7.55 [7.14;8.63]	-1.49 [-1.92;-1.06]	8.59 [7.99;9.65]	0.67 [-0.57;1.9]
miR-320e	9.72 [9.14;9.85]	-1.43 [-1.85;-1.01]	9.93 [9.51;10.38]	0.48 [-0.27;1.24]
miR-345-5p	6.74 [5.95;7.55]	1.38 [0.98;1.78]	11.26 [10.92;12.3]	0.61 [-0.52;1.75]
miR-769-5p	5.93 [5.74;6.21]	-0.63 [-0.82;-0.44]	7.03 [6.56;7.77]	-0.22 [-1.05;0.62]
miR-150-3p	8.15 [7.68;8.61]	1.29 [0.9;1.67]	8.47 [7.85;9.31]	-0.49 [-1.38;0.41]
miR-30a-5p	11.45 [10.85;11.96]	-2.56 [-3.34;-1.78]	8.25 [7.17;8.76]	-0.43 [-1.11;0.25]
miR-4488	6 [5.71;7.04]	1.35 [0.94;1.76]	5.78 [5.23;6.28]	-1.24 [-2.33;-0.15]
miR-125a-5p	10.46 [9.69;10.53]	-2.21 [-2.89;-1.54]	6.57 [6.21;6.91]	-0.35 [-1.12;0.43]
miR-374a-5p	9.91 [8.82;10.78]	-2.42 [-3.17;-1.68]	10.3 [9.5;10.79]	1.8 [0.44;3.16]
miR-409-3p	7.16 [6.91;7.99]	-1.39 [-1.82;-0.97]	8.65 [6.9;9.64]	1.01 [-0.03;2.04]
miR-1307-5p	6.26 [5.41;7.17]	1.62 [1.12;2.11]	6.59 [5.99;7.48]	-0.86 [-2.03;0.31]
miR-4538	7.31 [6.57;7.74]	1.27 [0.88;1.65]	10.42 [9.61;11.01]	0.29 [-0.55;1.14]
miR-30e-5p	10.15 [9.47;10.95]	-2.06 [-2.71;-1.41]	12.78 [11.75;13.11]	0.25 [-0.8;1.3]
miR-4535	6.84 [6.16;7.39]	0.96 [0.67;1.24]	5.83 [5.52;6.03]	-1.21 [-2.13;-0.3]
miR-502-3p	6.47 [5.91;7]	-0.98 [-1.29;-0.68]	9.62 [8.85;9.99]	0.93 [-0.41;2.28]
miR-744-5p	5.67 [5.16;6.06]	-0.69 [-0.9;-0.47]	8.89 [8.11;10.21]	0.02 [-1.03;1.07]
miR-10b-5p	9.62 [8.93;10.15]	-2.58 [-3.39;-1.76]	6.08 [5.87;6.31]	-0.18 [-0.91;0.54]
miR-28-5p	8.35 [7.09;8.86]	-2.14 [-2.81;-1.47]	6.89 [6.46;7.46]	0.97 [0.07;1.87]

miR-125b-5p	14.42 [13.85;14.74]	-3.61 [-4.77;-2.45]	9.72 [9.02;10.53]	-0.19 [-0.89;0.5]
miR-126-3p	12.28 [11.99;12.68]	-3.1 [-4.11;-2.1]	6.54 [6;7.03]	0.82 [0;1.64]
miR-374b-5p	9.22 [7.93;9.85]	-2.29 [-3.01;-1.56]	5.59 [5.15;6.8]	1.28 [-0.02;2.58]
miR-4470	6.72 [5.98;7.37]	0.78 [0.55;1.01]	6.82 [6.55;8.04]	0.33 [-0.45;1.11]
miR-532-5p	7.57 [7.15;8.33]	-1.6 [-2.12;-1.09]	7.9 [7.27;8.49]	-0.42 [-1.53;0.7]
miR-29b-3p	12.23 [10.77;13.19]	-3 [-3.98;-2.03]	5.72 [5.33;6.52]	0.97 [-0.14;2.08]
miR-335-5p	7.27 [6.16;7.69]	-1.43 [-1.89;-0.97]	10.62 [9.94;11.47]	0.86 [-0.39;2.12]
miR-30c-5p	10.6 [9.09;11.01]	-2.71 [-3.59;-1.83]	6.37 [5.66;7.06]	0.54 [-0.36;1.43]
miR-363-3p	6.53 [6.21;7.66]	-1.34 [-1.76;-0.91]	5.33 [4.92;5.64]	1.13 [-0.21;2.46]
miR-223-3p	10.26 [9.93;10.99]	-2.36 [-3.15;-1.56]	6.69 [5.8;7.21]	1.96 [0.59;3.33]
miR-29a-3p	12.75 [11.84;13.48]	-2.41 [-3.22;-1.61]	8.22 [7.13;8.9]	0.48 [-0.47;1.43]
miR-99b-5p	9.02 [8.5;9.33]	-1.72 [-2.3;-1.15]	9.61 [8.51;10.37]	-0.29 [-0.99;0.41]
miR-29c-3p	13.27 [11.7;14.36]	-3.45 [-4.6;-2.29]	11.65 [11.04;12.47]	0.11 [-0.77;1]
miR-101-3p	10.12 [8.84;11.01]	-2.68 [-3.58;-1.78]	9.81 [8.97;10.6]	-0.35 [-1.45;0.76]
miR-486-5p	7.61 [6.84;8.72]	-0.89 [-1.16;-0.62]	7.67 [6.85;8.3]	0.09 [-0.89;1.07]
miR-22-5p	6.62 [5.93;7.78]	-1.21 [-1.61;-0.81]	7.02 [6.38;7.82]	1.29 [-0.13;2.71]
miR-361-5p	8.75 [7.73;9.09]	-1.87 [-2.51;-1.23]	5.66 [5.32;6.28]	0.73 [-0.49;1.96]
miR-142-3p	10.76 [9.93;12]	-2.11 [-2.85;-1.37]	11.3 [10.16;11.41]	0.94 [0.01;1.87]
miR-500a-3p	6.47 [5.98;6.84]	-0.75 [-1.01;-0.49]	6.44 [5.71;7.04]	0.68 [-0.61;1.97]
miR-422a	5.68 [5.09;6.12]	1.07 [0.7;1.44]	7.22 [6.59;8.29]	-0.21 [-1.16;0.74]
miR-652-3p	6.44 [5.56;7.03]	-1.1 [-1.48;-0.72]	10.42 [9.78;11.48]	0.35 [-0.94;1.64]
miR-30d-5p	9.93 [9.51;10.38]	-1.38 [-1.88;-0.89]	12.77 [12.24;12.98]	0.43 [-0.51;1.36]
miR-671-5p	8.05 [7.57;9.06]	1.42 [0.92;1.92]	6.54 [6.16;7.43]	1.28 [0.75;1.81]
miR-22-3p	11.65 [11.04;12.47]	-2.19 [-2.98;-1.39]	5.06 [4.88;5.2]	-0.47 [-1.34;0.4]
miR-130a-3p	11.3 [10.16;11.41]	-2.86 [-3.91;-1.81]	6.33 [5.89;6.85]	0.95 [-0.17;2.07]
miR-92a-3p	10.15 [9.72;10.58]	-1.42 [-1.94;-0.89]	7.86 [7.59;9.22]	0.09 [-1.02;1.21]
miR-148b-3p	7.3 [6.39;7.79]	-1.36 [-1.85;-0.86]	6.01 [5.66;6.54]	-0.51 [-1.6;0.58]
miR-10a-3p	5.28 [5.08;5.5]	0.53 [0.33;0.73]	11.27 [10.53;12.41]	0.1 [-0.67;0.86]
miR-1268b	10.34 [9.39;11.02]	1.21 [0.76;1.65]	5.11 [4.7;5.58]	-0.71 [-1.58;0.16]
miR-224-5p	8.35 [6.12;8.92]	-1.96 [-2.68;-1.25]	12.01 [11.62;12.33]	0.63 [-0.48;1.75]
miR-30b-5p	11.51 [9.85;12.09]	-2.75 [-3.77;-1.72]	6.29 [5.28;6.8]	1.96 [0.73;3.18]
miR-221-5p	5.7 [5.14;6.49]	-0.78 [-1.06;-0.5]	8.98 [8.26;9.54]	0.72 [-0.33;1.78]
miR-4665-5p	6.66 [6.03;7.57]	1.42 [0.89;1.95]	8.16 [7.86;8.94]	0.4 [-0.61;1.42]
miR-19b-3p	11.63 [10.35;12.29]	-2.54 [-3.51;-1.57]	6.15 [5.45;6.93]	-0.1 [-1.38;1.19]
miR-642a-3p	11.43 [9.74;13.73]	2.54 [1.59;3.49]	8.74 [7.45;9]	-0.29 [-0.97;0.39]

miR-125a-3p	8.74 [8.23;9.61]	1.46 [0.91;2.01]	5.01 [4.72;5.28]	0.12 [-1.18;1.43]
miR-23b-3p	12.78 [11.75;13.11]	-2.74 [-3.8;-1.69]	11.7 [8.87;12.52]	1.27 [0.15;2.4]
miR-19a-3p	9.68 [8.4;10.69]	-2.38 [-3.29;-1.46]	9.06 [8.35;10.8]	0.36 [-0.93;1.65]
let-7g-5p	12.64 [11.37;13.24]	-2.75 [-3.82;-1.68]	10.01 [8.47;10.7]	0.66 [-0.44;1.77]
miR-574-3p	9.29 [8.68;9.69]	-1.07 [-1.49;-0.65]	8.41 [7.34;9.22]	1.46 [0.35;2.57]
miR-151b	8.86 [7.68;9.33]	-1.7 [-2.36;-1.05]	11.21 [10.32;11.97]	-0.16 [-1.15;0.84]
miR-199b-5p	10.21 [9.87;10.82]	-2.67 [-3.74;-1.6]	5.76 [5.29;6.25]	0.66 [-0.37;1.7]
miR-27b-3p	12.27 [11.18;12.93]	-2.36 [-3.3;-1.43]	9.06 [7.21;10.66]	0.55 [-0.38;1.48]
miR-3180-3p	5.54 [4.96;6.14]	1.18 [0.71;1.65]	12.75 [11.84;13.48]	-0.19 [-0.94;0.56]
miR-151a-3p	7.92 [7.67;8.71]	-1.06 [-1.49;-0.63]	10.76 [9.93;12]	0.42 [-0.46;1.29]
miR-501-3p	6.12 [5.88;6.53]	-0.51 [-0.71;-0.31]	6.14 [6.05;6.77]	0.79 [-0.56;2.14]
miR-139-5p	6.01 [5.66;6.54]	-0.79 [-1.1;-0.48]	9.47 [8.77;10.48]	0.71 [-0.02;1.43]
miR-16-5p	12.6 [12.2;13.08]	-2.21 [-3.12;-1.31]	8.64 [7.79;9.6]	1.28 [-0.22;2.78]
miR-151a-5p	9.96 [8.47;10.41]	-2.24 [-3.15;-1.34]	8.39 [7.79;9.7]	-0.06 [-1.32;1.2]
miR-3196	10.3 [9.5;10.79]	0.81 [0.48;1.14]	8.13 [7.62;9.06]	-0.78 [-1.69;0.12]
miR-4738-3p	6.87 [6.2;7.91]	1 [0.61;1.4]	4.85 [4.67;5.21]	0.15 [-0.71;1.01]
miR-98-5p	8.33 [7.34;9.15]	-1.94 [-2.73;-1.14]	9.51 [7.99;10.44]	1.96 [0.64;3.28]
miR-26b-5p	12.3 [10.74;12.75]	-2.9 [-4.09;-1.7]	9.12 [7.76;9.76]	0.64 [-0.36;1.63]
let-7i-5p	12.25 [11.9;12.87]	-2.12 [-3.02;-1.21]	12.21 [11.99;12.51]	1.02 [0.02;2.01]
let-7d-5p	11.31 [9.94;11.47]	-2.44 [-3.45;-1.42]	12.64 [11.37;13.24]	1.06 [-0.08;2.21]
let-7b-5p	14.19 [13.63;14.46]	-2.09 [-2.96;-1.21]	11.31 [9.94;11.47]	-0.06 [-1.05;0.93]
miR-425-5p	8.41 [7.34;9.22]	-1.73 [-2.45;-1.01]	11.61 [10.34;12.64]	1.12 [0.05;2.18]
miR-4429	5.89 [5.52;6.59]	0.81 [0.47;1.16]	7.99 [6.68;8.39]	0.61 [-0.06;1.27]
miR-146b-5p	8.64 [7.79;9.6]	-1.83 [-2.61;-1.04]	6.89 [5.95;7.64]	0.31 [-0.47;1.08]
miR-939-5p	9.85 [9.13;10.55]	0.86 [0.5;1.23]	6.72 [5.98;7.37]	0.15 [-0.77;1.07]
miR-15a-5p	9.68 [9.06;10.73]	-2.03 [-2.93;-1.13]	8.85 [8.24;9.41]	1.29 [0.22;2.35]
miR-3620-5p	6.82 [6.55;8.04]	1.34 [0.75;1.94]	4.93 [4.65;5.21]	0.11 [-0.61;0.84]
miR-24-3p	12.32 [11.05;12.43]	-2.03 [-2.94;-1.12]	11.56 [9.23;12.29]	0.2 [-0.8;1.21]
let-7f-5p	13.79 [12.93;14.44]	-2.57 [-3.74;-1.41]	12.25 [11.9;12.87]	0.75 [-0.54;2.04]
miR-199a-5p	12.01 [11.62;12.33]	-2.45 [-3.57;-1.33]	7.55 [7.14;8.63]	0.78 [-0.33;1.9]
let-7a-5p	14.46 [13.58;14.88]	-2.56 [-3.73;-1.39]	14.19 [13.63;14.46]	0.71 [-0.49;1.92]
miR-26a-5p	13.17 [12.12;13.64]	-2.48 [-3.62;-1.33]	10.78 [9.38;11.46]	0.97 [-0.01;1.96]
miR-20b-5p	9.12 [7.76;9.76]	-1.89 [-2.77;-1.01]	7.55 [7.07;8.33]	0.34 [-0.83;1.5]
miR-199a-3p	13 [12.6;13.4]	-2.48 [-3.66;-1.29]	4.98 [4.77;5.23]	1.22 [-0.08;2.52]
miR-4306	8.41 [8.01;8.94]	-0.54 [-0.79;-0.29]	4.92 [4.72;5.04]	0.34 [-0.31;0.99]

miR-4492	4.69 [4.54;4.87]	0.43 [0.22;0.63]	5.47 [4.99;5.98]	-0.03 [-0.78;0.71]
miR-103a-3p	11.22 [10.31;11.6]	-1.89 [-2.8;-0.98]	6.19 [5.09;6.86]	0.62 [-0.51;1.75]
miR-107	11.03 [9.78;11.34]	-2.11 [-3.15;-1.08]	11.61 [11.04;13.46]	0.65 [-0.61;1.91]
miR-1299	6.25 [5.73;6.65]	0.57 [0.3;0.83]	7.18 [5.73;8.66]	0.9 [-0.5;2.31]
miR-23a-3p	12.77 [12.24;12.98]	-1.76 [-2.63;-0.89]	9.68 [8.4;10.69]	1.93 [0.88;2.99]
miR-3141	8.67 [8.22;9.32]	0.84 [0.43;1.25]	11.59 [11.01;12.27]	-0.03 [-0.7;0.63]
miR-452-5p	6.68 [5.95;7.24]	-0.65 [-0.97;-0.33]	6.78 [5.86;7.48]	0.54 [-0.53;1.61]
let-7e-5p	10.79 [9.93;11.17]	-2.23 [-3.37;-1.1]	5.63 [5.37;6.39]	0.2 [-0.68;1.08]
miR-1260b	10.52 [9.73;11.77]	-1.5 [-2.28;-0.73]	8.74 [8.23;9.61]	1.13 [-0.02;2.28]
miR-27a-3p	11.59 [11.01;12.27]	-1.35 [-2.05;-0.64]	8.41 [7.42;9.68]	1.22 [0.25;2.2]
miR-20a-5p	10.78 [9.38;11.46]	-2.14 [-3.28;-1]	9.67 [8.32;10.11]	0.51 [-0.87;1.9]
miR-148a-3p	9.82 [9.1;11.08]	-1.73 [-2.67;-0.79]	8 [6.94;8.73]	0.82 [-0.3;1.93]
miR-17-5p	9.65 [8.2;10.22]	-1.81 [-2.77;-0.84]	7.89 [7.64;9.29]	0.2 [-1.04;1.44]
miR-324-3p	9.98 [9.45;10.66]	-0.75 [-1.15;-0.34]	4.81 [4.65;6.03]	-0.37 [-1.41;0.66]
miR-15b-5p	10.06 [9.31;10.67]	-1.67 [-2.59;-0.75]	6.03 [5.3;6.89]	2.18 [0.79;3.56]
miR-1290	7.21 [6.86;7.94]	0.98 [0.44;1.52]	9.12 [8.45;9.56]	-1.76 [-2.77;-0.75]
miR-378c	5.55 [5.18;5.85]	0.37 [0.17;0.56]	5.12 [4.88;5.47]	-0.33 [-1.4;0.75]
miR-424-3p	6.57 [5.77;6.96]	0.79 [0.36;1.23]	6.74 [5.95;7.55]	-0.33 [-1.54;0.89]
miR-181a-3p	5.53 [5.29;6.29]	-0.46 [-0.72;-0.21]	7.3 [6.39;7.79]	-0.7 [-1.63;0.23]
miR-5585-3p	9.16 [8.37;9.77]	0.62 [0.27;0.97]	6.57 [5.77;6.96]	-1.33 [-2.56;-0.1]
miR-1260a	11.42 [9.82;12.51]	-1.62 [-2.56;-0.67]	5.24 [4.93;5.71]	1.03 [-0.13;2.2]
miR-4448	5.51 [5.05;5.99]	0.57 [0.23;0.91]	7 [6.07;8.18]	0.39 [-0.57;1.35]
miR-1246	9.72 [9.02;10.53]	0.99 [0.4;1.57]	11.7 [10.99;12.83]	-0.67 [-1.81;0.47]
miR-200a-3p	11.7 [8.87;12.52]	-2.37 [-3.79;-0.95]	5.24 [5.03;5.5]	0.8 [-0.21;1.81]
miR-106b-5p	9.81 [8.97;10.6]	-1.41 [-2.28;-0.54]	6.42 [6.03;6.94]	0.25 [-1;1.49]
miR-654-5p	5.63 [5.46;5.94]	0.5 [0.19;0.81]	8.41 [8.01;8.94]	0.46 [-0.63;1.55]
miR-25-3p	9.65 [9.4;10.43]	-1.34 [-2.19;-0.5]	6.08 [5.06;6.98]	0.77 [-0.47;2]
miR-4646-5p	8.05 [7.47;8.68]	0.56 [0.22;0.9]	5.51 [5.19;6.08]	0.04 [-0.62;0.69]
miR-3960	14.24 [13.19;15]	0.79 [0.31;1.27]	7.34 [6.51;7.94]	-0.94 [-1.74;-0.15]
miR-196b-5p	5.06 [4.88;5.2]	0.7 [0.26;1.15]	7.25 [6.83;7.41]	0.96 [0.05;1.87]
miR-431-5p	5.21 [4.94;6.02]	1.11 [0.41;1.81]	6.63 [5.96;6.87]	0.49 [-0.43;1.41]
miR-194-5p	12.74 [9.24;13.34]	-2.84 [-4.63;-1.04]	8.42 [7.86;8.87]	0.8 [-0.46;2.06]
miR-192-5p	12.33 [9.43;13.3]	-2.74 [-4.5;-0.97]	9.65 [8.2;10.22]	-0.09 [-0.91;0.74]
miR-200c-3p	11.56 [9.23;12.29]	-2.08 [-3.42;-0.73]	5.12 [5.01;5.6]	0.93 [-0.1;1.97]
miR-129-5p	5.39 [5.04;5.77]	0.37 [0.13;0.61]	10.34 [9.39;11.02]	0.3 [-0.7;1.3]

miR-10a-5p	10.04 [9.55;10.42]	-1.54 [-2.59;-0.5]	5.84 [5.46;6.1]	-0.2 [-0.98;0.59]
miR-155-5p	7.79 [7.19;8.52]	0.46 [0.15;0.77]	8.98 [7.8;10.45]	0.71 [-0.07;1.49]
miR-450a-5p	5.57 [5.3;6.48]	-0.45 [-0.77;-0.13]	7.42 [6.2;8.01]	0.17 [-1;1.34]
miR-146a-5p	7.44 [6.96;9.24]	-1.1 [-1.89;-0.31]	8.89 [6.7;10.33]	0.47 [-0.63;1.57]
miR-200b-3p	12.02 [9.63;12.76]	-1.76 [-3.06;-0.46]	5.93 [5.46;6.51]	0.73 [-0.24;1.7]
miR-197-3p	7.02 [6.38;7.82]	-0.51 [-0.89;-0.13]	6.15 [5.76;7.06]	1.18 [0.05;2.31]
miR-4533	5.2 [4.73;6.14]	0.58 [0.14;1.02]	10.19 [9.22;11.1]	1.71 [1.02;2.39]
miR-93-5p	8.92 [8.24;9.4]	-1.02 [-1.81;-0.23]	4.63 [4.49;4.76]	-0.22 [-1.6;1.16]
miR-181b-5p	8.42 [7.86;8.87]	-0.59 [-1.07;-0.1]	8.15 [7.68;8.61]	0.07 [-0.82;0.95]
miR-221-3p	8.12 [7.57;8.93]	-0.77 [-1.41;-0.12]	11.06 [10.37;12.36]	1.2 [0.06;2.33]
miR-3614-5p	5.98 [5.65;6.74]	0.69 [0.1;1.28]	5.37 [4.96;5.56]	0.95 [-0.04;1.93]
miR-181a-5p	10.05 [9.65;10.73]	-0.81 [-1.52;-0.11]	5.95 [5.29;6.57]	0.1 [-0.8;1.01]
miR-3605-5p	6.63 [5.96;6.87]	0.28 [0.04;0.53]	5.04 [4.83;5.32]	0.09 [-0.92;1.11]

miRNA: microRNA; $5th$;$95th$: 5th and 95th percentiles; *CI*: confidence interval.

Table A.4: *Overall and stratified differences in microRNA expression by age, sex, body-mass index and genetic susceptibility to gallstone disease between prospective gallbladder cancer cases and controls, by cohort.*

								miR-4533								
		Janus			**ESTHER+HNR**			**HUNT**			**FINRISK**			**TwinGene**		
	Median	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**
Age	<63.5	1.69	[1.17;2.20]	0.002	-0.02	[-0.22;0.17]	0.81	0.13	[-0.21;0.47]	0.47	-0.23	[-0.32;-0.14]	0.001	-0.61	[-0.81;-0.23]	0.004
	>63.5	1.34	[-0.07;2.75]	0.07	0.06	[-0.15;0.26]	0.62	0.05	[-0.15;0.25]	0.61	0.001	[-0.01;0.02]	0.52	0.00	[0.002;0.005]	0.002
Sex	F	1.58	[0.84;2.33]	0.002	-0.001	[-0.14;0.14]	0.99	0.14	[-0.04;0.33]	0.14	0.00	[-0.01;0.03]	0.44	-0.22	[-0.51;0.08]	0.19
	M	2.00	[0.89;3.11]	0.004	0.00	[0.00;0.00]	0.99	-0.12	[-0.45;0.22]	0.56	-0.27	[-0.27;0.30]	0.75	-0.03	[-0.12;0.06]	0.46
BMI	<26.2	2.13	[1.47;2.79]	0.002	-0.08	[-0.70;0.53]	0.79	0.32	[0.16;0.47]	0.002	-0.27	[-0.27;-0.23]	0.002	-0.29	[-0.72;0.14]	0.22
	>26.2	1.15	[0.03;2.27]	0.05	0.02	[-0.05;0.09]	0.66	-0.13	[-0.53;0.27]	0.54	0.00	[-0.03;0.03]	0.99	-0.03	[-0.12;0.05]	0.46
PRS-GS	<2.88	-	-	-	0.01	[-0.01;0.03]	0.41	0.02	[-0.39;0.44]	0.91	0.15	[-0.59;0.89]	0.70	0.00	[-0.02;0.02]	0.75
	>2.88	-	-	-	0.13	[-0.41;0.66]	0.66	0.17	[0.04;0.31]	0.04	-0.15	[-0.36;0.06]	0.21	-0.19	[-0.47;0.10]	0.24

								miR-671-5p								
		Janus			**ESTHER+HNR**			**HUNT**			**FINRISK**			**TwinGene**		
	Median	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**	β	**95% CI**	**p-value**
Age	<63.5	0.71	[-0.06;1.48]	0.08	-0.11	[-0.20;-0.01]	0.05	0.00	[-0.00;0.00]	0.72	-0.02	[-0.22;0.17]	0.83	-0.02	[-0.02;0.05]	0.26
	>63.5	0.30	[-0.57;1.17]	0.51	-0.69	[-2.29;0.91]	0.46	0.07	[-0.16;0.31]	0.55	0.00	[0.00;0.00]	0.99	0.00	[0.00;0.00]	0.99
Sex	F	0.63	[0.009;1.25]	0.05	-0.15	[-0.22;-0.09]	0.001	-0.01	[-0.01;0.04]	0.68	-0.62	[-1.10;-0.13]	0.06	0.00	[0.00;0.00]	0.99
	M	0.17	[-1.08;1.42]	0.79	0.00	[0.00;0.00]	0.99	0.20	[-1.11;1.51]	0.79	0.00	[-0.02;0.02]	0.89	0.00	[-0.01;0.01]	0.57
BMI	<26.2	0.61	[-0.14;1.37]	0.12	-0.20	[-0.21;-0.19]	0.01	0.01	[-0.004;0.02]	0.23	0.00	[0.00;0.00]	0.99	0.00	[0.00;0.00]	0.99
	>26.2	0.38	[-0.48;1.24]	0.39	-0.01	[-0.04;0.03]	0.65	-0.01	[-0.04;0.002]	0.52	-0.67	[-0.99;-0.33]	0.01	0.00	[0.00;0.00]	0.99
PRS-GS	<2.88	-	-	-	-0.10	[-0.11;-0.09]	0.003	0.00	[-0.02;-0.01]	0.03	-0.002	[-0.001;0.04]	0.62	0.00	[-0.01;0.01]	0.75
	>2.88	-	-	-	-0.14	[-0.41;0.14]	0.33	0.01	[0.01;0.80]	0.001	-0.10	[-0.36;0.16]	0.46	0.00	[0.00;0.00]	0.99

BMI: body-mass index; *PRS*: polygenic risk score; *GS*: gallstones ; *CI*: confidence interval; *p-value*: probability value; ESTHER: Early detection and optimised therapy of chronic diseases in the elderly population; HNR: Heinz Nixdorf recall study; HUNT: Nord-Trøndelag Health study.

Table A.5: *Results from pathway analysis for the preselected microRNAs.*

KEGG pathway	**p-value**	**# genes**	**# miRNAs**
Proteoglycans in cancer	0.000000002	156	68
Renal cell carcinoma	0.000001	60	55
Glioma	0.000001	55	55
ErbB signaling pathway	0.000001	75	63
Rap1 signaling pathway	0.00001	164	73
Hippo signaling pathway	0.00003	115	65
Amphetamine addiction	0.00003	53	62
Axon guidance	0.00003	98	63
Sphingolipid signaling pathway	0.00003	92	67
Ras signaling pathway	0.00003	168	72
Pancreatic cancer	0.00004	54	55
Choline metabolism in cancer	0.0001	83	64
Adherens junction	0.0001	62	60
cAMP signaling pathway	0.0001	154	71
FoxO signaling pathway	0.0001	105	65
mTOR signaling pathway	0.0002	52	54
Signaling pathways of stem cells	0.0002	107	68
TGF-beta signaling pathway	0.0002	62	61
Colorectal cancer	0.0002	54	57
Focal adhesion	0.0002	157	69
N-Glycan biosynthesis	0.0002	39	51
Oxytocin signaling pathway	0.0002	121	69
Pathways in cancer	0.0002	288	76
MAPK signaling pathway	0.0002	187	77
Cocaine addiction	0.0002	38	56
Prostate cancer	0.0003	72	63
Thyroid hormone signaling pathway	0.0004	90	64
AMPK signaling pathway	0.0005	96	68
Long-term depression	0.0005	46	53
Endocytosis	0.0005	156	73
Adrenergic signaling in cardiomyocytes	0.0005	108	75
Circadian rhythm	0.0008	28	47
Glutamatergic synapse	0.002	86	62

Endometrial cancer	0.002	43	53
Chronic myeloid leukemia	0.002	58	54
Neurotrophin signaling pathway	0.002	93	64
Acute myeloid leukemia	0.002	47	52
Platelet activation	0.002	97	64
Melanoma	0.002	58	60
Ubiquitin mediated proteolysis	0.003	102	61
Wnt signaling pathway	0.004	108	68
Transcriptional misregulation in cancer	0.004	122	73
Prolactin signaling pathway	0.004	53	54
Non-small cell lung cancer	0.005	44	55
Dopaminergic synapse	0.005	97	71
PI3K-Akt signaling pathway	0.01	238	73
TNF signaling pathway	0.01	83	61
Estrogen signaling pathway	0.01	72	66
mRNA surveillance pathway	0.01	70	63
Hepatitis B	0.01	100	68
cGMP-PKG signaling pathway	0.01	121	73
Phosphatidylinositol signaling system	0.01	58	54
Prion diseases	0.01	20	34
Insulin signaling pathway	0.02	103	64
Small cell lung cancer	0.02	65	56
Regulation of TRP channels	0.02	71	59
Regulation of actin cytoskeleton	0.02	152	69
Long-term potentiation	0.02	52	60
ARVC	0.02	53	52
Type II diabetes mellitus	0.02	38	49
Aldosterone-regulated sodium reabsorption	0.02	32	48
Lysine degradation	0.03	35	60
Dorso-ventral axis formation	0.03	23	40
Bacterial invasion of epithelial cells	0.04	57	61
Cholinergic synapse	0.05	82	64

p-value: probability value; *miRNAs*: microRNAs.

Table A.6: *List of genes negatively correlated with the expression of miR-4533 in the five most significant pathways.*

Gene	**Spearman Rho**	**95% CI**	**p-value**
FLT4	-0.268	[-0.47;-0.04]	0.011
RAP1A	-0.262	[-0.49;0.01]	0.013
FGF7	-0.248	[-0.44;-0.02]	0.018
SIPA1L2	-0.247	[-0.48;-0.02]	0.018
ARNT2	-0.245	[-0.44;-0.01]	0.019
ITGAM	-0.19	[-0.39;0.03]	0.055
MAPK9	-0.189	[-0.39;0.04]	0.055
RAPGEF1	-0.187	[-0.42;0.05]	0.057
RAPGEF5	-0.187	[-0.42;0.07]	0.058
FAS	-0.179	[-0.39;0.05]	0.066
CAMK4	-0.176	[-0.4;0.07]	0.069
FLNB	-0.176	[-0.4;0.06]	0.069
RAPGEF4	-0.17	[-0.38;0.05]	0.077
EGLN1	-0.167	[-0.38;0.06]	0.08
IQGAP1	-0.163	[-0.38;0.07]	0.086
MAPK8	-0.163	[-0.38;0.08]	0.086
PIK3R2	-0.157	[-0.37;0.08]	0.093
SHH	-0.157	[-0.36;0.07]	0.093
VAV1	-0.158	[-0.36;0.07]	0.093
RAC1	-0.157	[-0.37;0.07]	0.095
E2F2	-0.15	[-0.35;0.09]	0.105
FGF10	-0.145	[-0.36;0.08]	0.113
AKT2	-0.144	[-0.38;0.11]	0.114
INSR	-0.142	[-0.37;0.09]	0.117
ANK3	-0.141	[-0.35;0.09]	0.118
E2F1	-0.14	[-0.36;0.1]	0.12
PRKACB	-0.133	[-0.35;0.1]	0.132
MAP2K4	-0.132	[-0.35;0.11]	0.134
RASGRP3	-0.13	[-0.35;0.13]	0.138
HGF	-0.128	[-0.34;0.11]	0.142

p-value: probability value; *CI*: confidence interval.

Table A.7: *Serum microRNA expression in controls, and expression differences between prospective gallbladder cancer cases and controls for 34 microRNAs previously linked with gallbladder cancer in literature.*

						log2 expression	**Case-Control**	
miRNA	**PMID**	**Pop**	**Sample**	**N**	**Regulation**	**in controls**	**Difference**	**Same**
						Median [5th;95th]	**[95% CI]**	
miR-133a-3p	27904763	Chinese	Tissue	23	down	1.19 [0.00; 5.03]	0.13 [-0.31; 0.56]	No
miR-145-5p	30886199	European	Tissue	48	down	0.00 [0.00; 2.04]	0.28 [0.07; 0.49]	No
miR-146b-5p	25760482	Chinese	Tissue	92	down	10.09 [9.62; 11.55]	-0.03 [-0.16; 0.10]	Yes
miR-26b-5p	31570091	Chinese	Tissue	35	down	7.26 [4.87; 9.93]	0.09 [-0.12; 0.29]	No
miR-122-5p	37925508	Indian	Tissue	5	up	15.41 [13.22;16.61]	-0.02 [-0.26; 0.22]	No
miR-127-5p	37925508	Indian	Tissue	5	up	0.01 [0.00;3.53]	-0.10 [-0.34; 0.14]	No
miR-1284	37925508	Indian	Tissue	5	down	0.00 [0.00;2.29]	0.03 [-0.07; 0.14]	No
miR-144-5p	37925508	Indian	Tissue	5	up	2.91 [0;5.57]	0.75 [0.34; 1.17]	Yes
miR-145-5p	37925508	Indian	Tissue	5	up	0.00 [0.00;2.04]	0.28 [0.07; 0.49]	Yes
miR-196a-5p	37925508	Indian	Tissue	5	down	0.00 [0.00;2.19]	0.17 [0.03; 0.31]	No
miR-196b-5p	37925508	Indian	Tissue	5	down	0.17 [0.00;2.90]	0.55 [0.30; 0.81]	No
miR-21-5p	37925508	Indian	Tissue	5	down	12.48 [10.74;15.15]	-0.11 [-0.25; 0.03]	Yes
miR-214-5p	37925508	Indian	Tissue	5	up	0.00 [0.00;2.25]	0.00 [-0.05; 0.05]	No
miR-23a-5p	37925508	Indian	Tissue	5	up	2.55 [0.00;4.57]	0.32 [-0.05; 0.68]	Yes
miR-32-5p	37925508	Indian	Tissue	5	down	2.76 [0.00;4.97]	0.77 [0.40; 1.14]	No
miR-3613-5p	37925508	Indian	Tissue	5	down	0.36 [0.00;3.48]	0.38 [0.04; 0.73]	No
miR-374a-5p	37925508	Indian	Tissue	5	down	1.07 [0.00;5.46]	0.49 [0.13; 0.84]	No
miR-378c	37925508	Indian	Tissue	5	down	5.52 [0.00;7.25]	-0.26 [-0.50; -0.02]	Yes
miR-382-5p	37925508	Indian	Tissue	5	up	6.70 [3.64;8.73]	0.03 [-0.28; 0.34]	Yes
miR-432-5p	37925508	Indian	Tissue	5	up	6.52 [0.00;7.91]	-0.09 [-0.43; 0.24]	No
miR-452-5p	37925508	Indian	Tissue	5	up	2.99 [0.00;4.91]	-0.12 [-0.55; 0.31]	No
miR-4732-5p	37925508	Indian	Tissue	5	up	4.57 [0.00;6.45]	0.19 [-0.24; 0.61]	Yes
miR-486-5p	37925508	Indian	Tissue	5	up	14.97 [12.87;16.61]	0.05 [-0.12; 0.23]	Yes
miR-493-5p	37925508	Indian	Tissue	5	up	3.08 [0.00;5.22]	0.51 [-0.02; 1.03]	Yes
miR-499a-5p	37925508	Indian	Tissue	5	down	2.86 [0.00;5.42]	0.07 [-0.27; 0.42]	No
miR-6852-5p	37925508	Indian	Tissue	5	down	2.94[0.00;5.08]	-0.13 [-0.50; 0.24]	Yes
miR-766-5p	37925508	Indian	Tissue	5	up	3.57[0.00;5.76]	-0.07 [-0.38; 0.23]	No
miR-9-5p	37925508	Indian	Tissue	5	down	1.90 [0.00;4.85]	0.01 [-0.43; 0.45]	No

miR-96-5p	37925508	Indian	Tissue	5	down	0.48[0.00;3.67]	0.14 [-0.13; 0.41]	No
miR-218-5p	25569100	Chinese	Tissue	80	down	0.45 [0.00; 4.74]	0.06 [-0.24; 0.37]	No
miR-30d-5p	29569755	Chinese	Tissue	80	down	11.80 [10.90; 13.24]	0.05 [-0.06; 0.17]	No
miR-143-3p	29416013	Chinese	Tissue	98	down	10.16 [8.89; 11.78]	-0.09 [-0.29; 0.10]	Yes
miR-29c-5p	28060377	Chinese	Tissue	80	down	0.66 [0.00; 3.56]	0.01 [-0.28; 0.30]	No
miR-92b-3p	32514152	Chinese	Serum	243	up	3.62 [0.00; 4.90]	-0.11 [-0.45; 0.23]	No

miRNA: microRNA; *PMID*: PubMed study ID; *Pop*: study population; $5th; 95th$: 5th and 95th percentiles; *CI*: confidence interval.

Appendix B

Appendix B: Implementations in R

Comment: Parts of the following Chapter have already been published in Cancers (Blandino et al., 2022). The original manuscript was written by myself, but also contains comments and corrections from the co-authors.

B.1 R Code: Identification of circulating long non-coding RNAs associated with gallbladder cancer risk

The following R codes describe the preselection of differentially expressed lncRNAs, model selection for prediction, and the prediction of the genotype-based lncRNA expression.

Preselection of differentially expressed lncRNAs - Jonckheere-Terpstra Test

```
########################################################
#
# program name:    01_LINC00662_preselection.R
# program title:      Preselection of differentially expressed lncRNAs
   along the sequence GS -> Dys -> GBC
# author:             Alice Blandino
# version:            1.0
# description:     Calculation of two-sided Jonckheere-Terpstra test

# input files:     01_data_LINC00662_preselection.txt
# Available at        www.biometrie.uni-heidelberg.de/
#                     StatisticalGenetics/Software_and_Data
#
########################################################
# "01_data_LINC00662_preselection.txt"
# A text file with a header line, and then one line per participant
# with the following two fields:
# LINC00662    expression of LINC00662 in FFPE tissue
# group           patients’ status (gallstones,dysplasia,GBC)

# install and activate package to run two-sided J-T test
install.packages("DescTools", dependencies = TRUE)
library(DescTools)
# load data of study participants
setwd("*Path:\*")
data_preselection <- read.table("01_data_LINC00662_preselection.txt",
   header=T)
# order the group variable
data_preselection$group <- factor(data_preselection$group,
                                  levels=c("GBC", "dysplasia", "
                                     gallstones"),
                                  ordered=TRUE)
# perform J-T test
jt.test<-JonckheereTerpstraTest(data_preselection$LINC00662,
                                data_preselection$group,
                                alternative = "two.sided",nperm =
                                   5000)
```

Selection of the best model for prediction based on robust AIC from robust linear regression models

```
#######################################################
#
# program name:   02_LINC00662_validation.R
# program title:     Selection of best model for prediction
# author:      Alice Blandino
# version:     1.0
# description:   Model selection based on robust AIC from robust
   linear regression models
# input files:       02_data_LINC00662_validation.txt
# Available at       www.biometrie.uni-heidelberg.de/
#                    StatisticalGenetics/Software_and_Data
#
#######################################################
# "02_data_LINC00662_validation.txt"
#
# A text file with a header line, and then one line per participant
# with the following fields:
#
# LINC00662    LINC00662 expression in serum
# rs11083486            genotype for rs11083486 (0=G/G ;1=G/T ;2=T/T)
# rs142521755        genotype for rs142521755 (0=A/A ;1=A/T ;2=T/T)
# age             study participants’ age
# sex       study participants’ sex
# PC1-PC10        first 10 PCs

# install and activate package to add variables to dataframe
```

```
install.packages("dplyr", dependencies = TRUE)
library(dplyr)

setwd("*Path:\*")
data_validation <- read.table("02_data_LINC00662_validation.txt",
   header=T)

# add new variables where:
# rs11083486 is once encoded dominantly (0+1 vs. 2), once encoded
   recessively (0 vs. 1+2)
# rs142521755 is encoded dominantly (0+1 vs. 2)
data_validation_new<-data_validation%>%
                     mutate(rs11083486.dominant=ifelse(rs11083486=="0"
                        ,1,rs11083486),
                           rs11083486.recessive=ifelse(rs11083486=="2
                              ",1,rs11083486),
                           rs142521755.dominant=ifelse(rs142521755=="
                              0",1,rs142521755))

# model selection
# install and activate package to run robust linear regression models
install.packages(c("MASS","repmod","AICcmodavg"), dependencies = TRUE)
library(MASS)
library(repmod)
library(AICcmodavg)
# 1.
# MODELS WITH rs11083486 ONLY
# additive
model.rs11083486.additive<-rlm(LINC00662~rs11083486+age+sex+PC1+PC2+
   PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_validation_new)
# three-genotypes
model.rs11083486.three<-rlm(LINC00662~as.factor(rs11083486)+age+sex+
   PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_validation_new)
# dominant
model.rs11083486.dom<-rlm(LINC00662~rs11083486.dominant+age+sex+PC1+
   PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_validation_new)
# recessive
model.rs11083486.rec<-rlm(LINC00662~rs11083486.recessive+age+sex+PC1+
   PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_validation_new)

# 2.
# MODEL WITH rs142521755 ONLY
# rs142521755 dominant
model.rs142521755.dom<-rlm(LINC00662~rs142521755.dominant+age+sex+PC1+
   PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_validation_new)

# 3.
# MODELS WITH BOTH rs11083486 AND rs142521755
# rs11083486 additive & rs142521755 dominant
model.add.dom<-rlm(LINC00662~rs11083486+rs142521755.dominant+age+sex+
   PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_validation_new)
# rs11083486 three-genotypes & rs142521755 dominant
```

```
model.three.dom<-rlm(LINC00662~as.factor(rs11083486)+rs142521755.
   dominant+age+sex+PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=
   data_validation_new)
# rs11083486 dominant & rs142521755 dominant
model.dom.dom<-rlm(LINC00662~rs11083486.dominant+rs142521755.dominant+
   age+sex+PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_
   validation_new)
# rs11083486 recessive & rs142521755 dominant
model.rec.dom<-rlm(LINC00662~rs11083486.recessive+rs142521755.dominant
   +age+sex+PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,data=data_
   validation_new)

# create a dataframe with each model's name and its AIC
# vector with AICs:
AICs<-c(AIC(model.rs11083486.additive),AIC(model.rs11083486.three),AIC
   (model.rs11083486.dom),AIC(model.rs11083486.rec),
        AIC(model.rs142521755.dom),AIC(model.add.dom),AIC(model.three.
           dom),AIC(model.dom.dom),AIC(model.rec.dom))
# vector with models' characteristics
models<-c("rs11083486.additive","rs11083486.three","rs11083486.
   dominant","rs11083486.recessive",
          "rs142521755.dominant","additive+dominant","three+dominant",
             "dominant+dominant","recessive+dominant")
# dataframe with both AIC and models' characteristics
summary.AIC<-data.frame(AICs,models)
# find which model has the lowest RAIC
summary.AIC[order(summary.AIC$AICs),,drop=FALSE] [1,]
```

Prediction of lncRNA expression based on individual genotype data and quantification of GBC risk

```
########################################################
# program name:   03_LINC00662_prediction.R
# program title:     genotype-based lncRNA expression prediction
# author:            Alice Blandino
# version:     1.0
# description:    prediction of lncRNA based on individual genotypes
   and GBC risk quantification
# input files:       03_data_LINC00662_prediction.txt
# Available at       www.biometrie.uni-heidelberg.de/
#                    StatisticalGenetics/Software_and_Data
#
#######################################################
# "03_data_LINC00662_prediction.txt"
#
# A text file with a header line, and then one line per participant
# with the following fields:
#
# rs11083486            genotype for rs11083486 (0=T/T ;1=G/T ;2=G/G)
# rs142521755        genotype for rs142521755 (0=A/A ;1=A/T ;2=T/T)
# pheno        patients' status (Control, Case)
# age             study participants' age
# sex        study participants’ sex
# PC1-PC10          first 10 PCs

# install and activate package to add variables to dataframe
install.packages(c("robustbase","dplyr"), dependencies = TRUE)
library(robustbase)
library(dplyr)

setwd("*Path:\*")
data_prediction <- read.table("03_data_LINC00662_prediction.txt",
    header=T)

# calculate the SNP-based expression
data_prediction_calculation<-data_prediction%>%
  mutate(rs11083486.coeff=ifelse(rs11083486=="0",-0.7352*0,ifelse(
      rs11083486=="1",-0.7352*1,-0.7352*2)),
          rs142521755.coeff=ifelse(rs142521755=="0",1.0797*0,ifelse(
              rs142521755=="1",1.0797*0,1.0797)),
          predicted.LINC00662=0.9267+rs11083486.coeff+rs142521755.coeff
              )

# association analysis fitting robust logistic regression model

# set controls as baseline category
data_prediction_calculation$pheno<-ordered(data_prediction_calculation
    $pheno, levels = c("Control", "Case"))

# model fitting
mod<-glmrob(as.factor(data_prediction_calculation$pheno)~predicted.
    LINC00662+age+sex+PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10,family
    = binomial, method= "Mqle",control= glmrobMqle.control(tcc=1.2),
    data=data_prediction_calculation)
summary(mod)

# extract Oddsratio for Cases
exp(summary(mod)$coefficients[2])

# extract lower and upper limits for confidence intervals
exp(summary(mod)$coefficients[2] + qnorm(c(0.5,0.025,0.975)) * summary
    (mod)$coefficients[2,2])[2]
exp(summary(mod)$coefficients[2] + qnorm(c(0.5,0.025,0.975)) * summary
    (mod)$coefficients[2,2])[3]
```

Own contribution to data collection and personal publications

I had no part in the patient recruitment and data acquisition. Patient recruitment, sample genotyping, and RNA sequencing were part of the following collaborative studies: *"Identification and validation of circulating sncRNAs causally associated with gallbladder cancer and development of a multifactorial risk prediction score"* supported by the German Research Foundation (DFG) (grant LO 1928/11-1, project number 424112940); *"European-Latin American Research Consortium towards Eradication of Preventable Gallbladder Cancer-EULAT Eradicate GBC"* founded by the European Union's Horizon 2020 research and innovation program (grant 825741); *"Identification of biomarkers for gallbladder cancer risk prediction-Towards personalized prevention of an orphan disease"* from The European Union's project (FP7 Research infrastructures: The european-wide Biobanking and Biomolecular Resources Research Infrastructure-Large Prospective Cohorts project (BBMRI-LPC); GA no. 313010). My research was also funded by the German Academic Exchange Service (DAAD) (grant 91778799).

The results described in my thesis are part of the following manuscripts:

Blandino A., Scherer D., Rounge T.B., Umu S.U., Boekstegers F., Barahona Ponce C., Marcelain K., Gárate-Calderón V., Waldenberger M., Morales E., Rojas A., Munoz C., Retamales J., de Toro G., Barajas O., Rivera M.T., Cortés A., Loader D., Saavedra J., Gutiérrez L., Ortega A., Bertrán M.E., Gabler F., Campos M., Alvarado J., Moisán F., Spencer L., Nervi B., Carvajal-Hausdorf D.E., Losada H., Almau M., Fernández P., Gallegos I., Olloquequi J., Fuentes-Guajardo M., Gonzalez-Jose R., Bortolini M.C., Gallo C., Linares A.R., Rothhammer F., Lorenzo Bermejo J. **Identification of Circulating lncRNAs Associated with Gallbladder Cancer Risk by Tissue-Based Preselection, Cis-eQTL Validation, and Analysis of Association with Genotype-Based Expression.** *Cancers (Basel). 2022 Jan 27;14(3):634. doi: 10.3390/cancers14030634. PMID: 35158906; PMCID: PMC8833674.*

Blandino A., Scherer D., Boekstegers F., Rounge T.B., Langseth H., Roessler S., Hveem K., Brenner H., Pechlivanis S., Waldenberger M., Lorenzo Bermejo J. **Small-RNA sequencing reveals potential serum biomarkers for gallbladder cancer: Results from a three-stage collaborative study of large European prospective cohorts.** *European Journal of Cancer* (under review).

Further own publications include:

Sciandra, M. & **Blandino, A. Un mese di Covid-19 in Italia: una guida alla lettura dei dati per bloccare la disinformazione.** *Societá Italiana Statistica (SIS). Anno IX EDIZIONE SPECIALE COVID-19. Url: http://www.rivista.sis-statistica.org/cms/?p=1170*

Conference contributions

Blandino A., Scherer, D., Lorenzo Bermejo, J. cis-eQTL-based identification of circulating lncRNAs associated with gallbladder cancer risk. *3rd International Conference on Cancer Prevention.* November 2022, Heidelberg, Germany.

Blandino A., Scherer, D., Lorenzo Bermejo, J. Robust linear regression for prediction of circulating long non-coding RNA expression based on individual genotypes. *European Mathematical Genetics Meeting (EMGM).* April 2022, Cambridge, United Kingdom.

Blandino A., Scherer, D., Lorenzo Bermejo, J. Genotype-based microRNA expression and gallbladder cancer (GBC) risk. *30th Annual Meeting of the International Genetic Epidemiology Society (IGES).* October 2021, Online.

Curriculum Vitae

Alice Blandino

born 27 June 1995 in Palermo, Italy

Nationality: Italian

Education

University of Heidelberg	since 05/2020
Doctoral student (Dr. sc. hum.)	
University of Palermo, Italy	10/2017 – 03/2020
Master of Statistical Sciences (M.Sc.)	07/03/2020
110/110 cum laude	
University of Applied Sciences Stuttgart	02/2018 – 08/2018
Study Semester	
University of Palermo, Italy	09/2014 – 09/2017
Bachelor of Statistics and Data Science (B.Sc.)	19/10/2017
109/110	
Ruggero Settimo Linguistic High school	09/2009 – 07/2014
98/100 (Diploma)	03/07/2014
Filippo Cordova Secondary School	09/2006 – 07/2009
Secondary Education (Licenza Media)	01/07/2009
Primary School Ferdinando I	09/2001 – 06/2006

Professional experience

University of Heidelberg Research assistant at the Statistical Genetics Research Group, Institute of Medical Biometry	since 05/2020
University of Heidelberg Intern at the Unit of Epidemiology and Biostatistics, Heidelberg Institute of Global Health	02/2019 – 06/2019
Palermo City Hall Intern at the Statistical Staff Unit, Strategic Development Sector	07/2017 - 09/2017

Acknowledgments

First of all, I would like to thank my doctoral supervisor Prof. Dr. Justo Lorenzo Bermejo for giving me the opportunity to join his research group. I really appreciate the constructive mentoring and support that contributed to the development of this thesis and to the maturing of my professional profile.

A great thanks goes to Dr. Dominique Scherer and Dr. Carol Barahona Ponce not only for their essential professional guidance throughout these years but, most importantly, for the emotional and moral support given me, especially in the last part of my PhD.

I would also like to acknowledge my colleagues and friends Valentina Gárate Calderón and Linda Zollner for the helpful and thorough revision of this thesis, as well as all the members of the "best statistical genetics research group". Thank you for these fun PhD years together, the inspiring never-ending conversations, and for making our coffee and lunch breaks so memorable.

A big thank you goes also to my Heidelberg family: Dr. Khatia Antia, Laura Batini, Manuel Jahn, Nikolas Liebster, Andrea Miola, Miriam Pasinato, Dr. Lisa Ringena, and Carlo Tombolini. Thank you for contributing in making Heidelberg home. I will always be grateful for your kind words of support, for cheering my successes and for holding my hand through the hard times.

To my sister and partner in crime Dr. Daniela Carraturo, thank you for showing me the value of true friendship. Words cannot express how grateful I am to have had you by my side since day one of this long and challenging journey. It has been an honor for me to grow into scientists together.

A special thank you goes to Alexander Stadler. Thank you for your unconditional love and for always listening and supporting me throughout these last months. We met during one of the most difficult moments of my PhD, but you were always there for me when I needed you the most.

To my sister. I know it may not look like it from my childhood drawings, but I am deeply grateful to have always shared every accomplishment of my life with you right next to me. We may not be physically together every day anymore, but you are always in my heart.

To my parents. Thank you for teaching me the importance of perseverance, for supporting my dreams, even when it meant being thousands of miles apart. You have always been my number one fans, and I could not be more grateful.

Finally, I would like to say to my old self, who started this long journey afraid of the unknown, that we finally made it.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema
   Identification and validation of circulating small non-coding RNAs associated with gallbladder cancer risk
   handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift

Angaben zu verwendeten KI-basierter elektronischer Hilfsmittel

Zur Dokumentation der verwendeten Hilfsmittel ist der schriftlichen Ausarbeitung ein besonderer Anhang hinzugefügt, der eine Liste und Beschreibung aller verwendeter KI-basierter Hilfsmittel enthält. Der besondere Anhang zur Dokumentation der verwendeten Hilfsmittel erfüllt folgende Kriterien:

1. Auflistung der Ziele, für die die KI-basierten Hilfsmittel in der vorliegenden Arbeit eingesetzt wurden.
2. Dokumentation der Verwendungsweise der KI-basierten Hilfsmittel
3. Nennung der Kapitel und Abschnitte der vorliegenden Arbeit, in denen die KI-basierten Hilfsmittel eingesetzt wurden, um Inhalte zu erzeugen.

Der Gebrauch dieser Hilfsmittel inklusive Art, Ziel und Umfang des Gebrauchs wurde mit meinem offiziellen Betreuer Herr apl Prof. Dr. Justo Lorenzo Bermejo abgesprochen.

Mir ist bewusst, dass insbesondere der Versuch einer nicht dokumentierten Nutzung KI-basierter Hilfsmittel als Täuschungsversuch zu werten ist:

Gem. § 16 Abs. 2 der Promotionsordnung "Dr. med./dent.":

"Ergibt sich vor Aushändigung der Promotionsurkunde, dass der Kandidat/die Kandidatin bei einer Promotionsleistung getäuscht hat, so können einzelne oder alle Promotionsleistungen für ungültig erklärt werden. In schweren Fällen kann die Zulassung zum Promotionsverfahren zurückgenommen werden."

Und § 16 Abs. 2 der Promotionsordnung "Dr. sc. hum.":

"Ergibt sich vor Aushändigung der Promotionsurkunde, dass der Doktorand / Doktorandin bei einer Promotionsleistung getäuscht hat, so kann der Promotionsausschuss diese Promotionsleistung oder alle bisher erbrachten Promotionsleistungen für ungültig erklären. In besonders schweren Fällen kann der Promotionsausschuss die Annahme als Doktorand / Doktorandin endgültig widerrufen."

Ort und Datum Unterschrift