UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DOTTORATO DI RICERCA IN PRECISION MEDICINE

Settore Scientifico Disciplinare - MED 0/1

XXXV ciclo

PIERCE: Pipeline to InfER Cancer Evolution through circulating tumor cells in cholangiocarcinoma patients.

DOTTORANDO
Marco Silvestri
87444

SUPERVISORE
Stefano Calza

ABSTRACT

Il colangiocarcinoma (CC) è una malattia rara e aggressiva con opzioni terapeutiche limitate e prognosi infausta nella maggior parte dei casi. Inoltre, il CC è difficile da rimuovere completamente tramite chirurgia a causa della sua posizione anatomica e della sua diffusione lungo il sangue o i dotti biliari, portando al problema di ottenere campioni seriali di tumore per monitorare la risposta al trattamento e quindi stabilire possibili terapie alternative. In questo contesto, una pipeline per monitorare l'evoluzione del cancro attraverso l'inferenza di firme molecolari specifiche relative al tumore primario sulle cellule tumorali circolanti (CTC) rappresenta un approccio non invasivo in grado di studiare la progressione e l'eterogeneità del cancro attraverso campioni di sangue longitudinali.

Noi ipotizziamo che la caratterizzazione molecolare delle CTC sia la chiave per monitorare l'evoluzione del cancro all'interno dei pazienti. Riconoscendo l'importanza delle CTC come surrogato tissutale adatto alla gestione clinica, l'inferenza della specifica firma molecolare associata al tumore primario sulle CTCs offre una reale opportunità per monitorare l'evoluzione del tumore sotto pressione terapeutica dando vita ad uno strumento predittivo che manca ancora non solo nel contesto di colangiocarcinoma ma anche in altri tipi di cancro.

Sono stati raccolti dataset genomici e trascrittomici (disponibili pubblicamente) di campioni normali e tumorali da serie ben annotate di pazienti CC: sono state estratte, integrate e convalidate firme molecolari specifiche del tumore significativamente diverse dalla controparte normale utilizzando approcci di machine learning per definire sottogruppi biologici distinti e creare uno strumento predittivo sulla base dei dati CNA. Allo stesso tempo, le CTC da prelievi di sangue di pazienti CC sono state recuperate e analizzate attraverso saggi genomici. Infine, lo strumento predittivo costruito dalla raccolta dei tessuti è stato utilizzato per predire l'associazione delle CTC a specifici sottogruppi biologici di CC consentendo il monitoraggio dell'evoluzione del cancro sotto pressione terapeutica.

Cholangiocarcinoma (CC) is a rare and aggressive disease with limited therapeutic options and dismal prognosis in most of the cases. Moreover, CC is difficult to completely resect by surgery because of its anatomical location and spread along the blood or bile ducts, leading to the problem of getting serial tumor samples for monitoring treatment response and then establishing possible alternative therapies. In this context, a pipeline to monitor cancer evolution through the inference of specific molecular signatures related to primary tumor on circulating tumor cells (CTCs) represents a noninvasive approach able to reflect cancer progression and heterogeneity through longitudinal blood samples.

We hypothesize that CTCs molecular characterization is key for monitoring cancer evolution within patients. Acknowledging the importance of CTCs as tissue surrogate suitable for clinical management, the inference of specific molecular signature associated to primary tumor on CTCs offers a real opportunity for monitoring tumor evolution under therapy pressure giving rise to a predictive tool that is still lacking not only in the context of CC but also in other cancer types.

Publicly available genomic and transcriptomic datasets of normal and tumor samples from well annotated series of CC patients were collected: tumor specific molecular signatures significantly different from normal counterpart were extracted, integrated and validated using machine learning approaches to define biological distinct subgroups and create a predictive tool based on copy number alteration (CNA) data. At the same time, CTCs from blood draws of CC patients were recovered and analyzed through genomic assays. Finally, the predictive tool built from the tissue collection were used to infer the CTCs association to a specific CC biological subgroups allowing the monitoring of cancer evolution under treatment pressure.

This study was conducted in collaboration with Drs. Vera Cappelletti, Dr. Federico Nichetti, Drs. Monica Niger at Fondazione IRCCS Istituto Nazionale dei Tumori di Milano (Milan, Italy) and of Prof. Yudi Pawitan and Dr. Nghia Vu Trung at Department of Medical Epidemiology and Biostatistics at Karolinska Institutet.

The current project is funded by Italian Association for Cancer Research as a fellowship (recipient: Marco Silvestri, ID 25430) for the period April 2021 – April 2023.

TABLE OF CONTENTS

ABBREVIATIONS

BL	Baseline
CCA	Cholangiocarcinoma
CI	Confidence interval
cis/gem	Cisplatin plus gemcitabine
CNA	Copy number alteration
CTC	Circulating tumor cell
DT	During treatment
ECC	Extrahepatic cholangiocarcinoma
EOT	End of treatment
FF	Fresh frozen
FFPE	Formalin-fixed Paraffine-embedded
FOLFOX	Oxaliplatin, L-folinic acid and 5-fluorouracil
FU	Follow-up
GBC	Gallbladder cancer
GII	Genome integrity index
HR	Hazard ratio
ICC	Intrahepatic Cholangiocarcinoma
INDELs	Small insertions and deletions
KNN	K-nearest neighbour
LM-PCR	Ligation-mediated PCR
LM	Linear model
lpWGS	Low-pass whole-genome sequencing
LST	Large-scale state transition
NGS	Next generation sequencing
NN	Neural network
OS	Overall survival
PCA	Principal component analysis
PD	Progression disease
PR	Partial response
QC	Quality control
RNAseq	RNA sequencing
RF	fast unified random forest
RFS	Relapse-free survival
SD	Stable disease
SNV	Single nucleotide variation
SVM	Support vector machine
Target-seq	Target sequencing
WBC	White blood cell
WGA	Whole-genome amplification
WGS	Whole genome sequencing
WES	Whole exome sequencing

1.INTRODUCTION

1.1. Cholangiocarcinoma

Cholangiocarcinoma (CC) consists of different epithelial malignancies arising in any part of the biliary tree and includes gallbladder cancer (GBC) and ampulla of Vater cancer. According to the location, cholangiocarcinomas are subdivided into intrahepatic (ICC) and extrahepatic (ECC) CC (located in the intrahepatic and extrahepatic bile ducts, respectively) and the latter can be further divided into distal and hilar CC (Figure 1) [1].

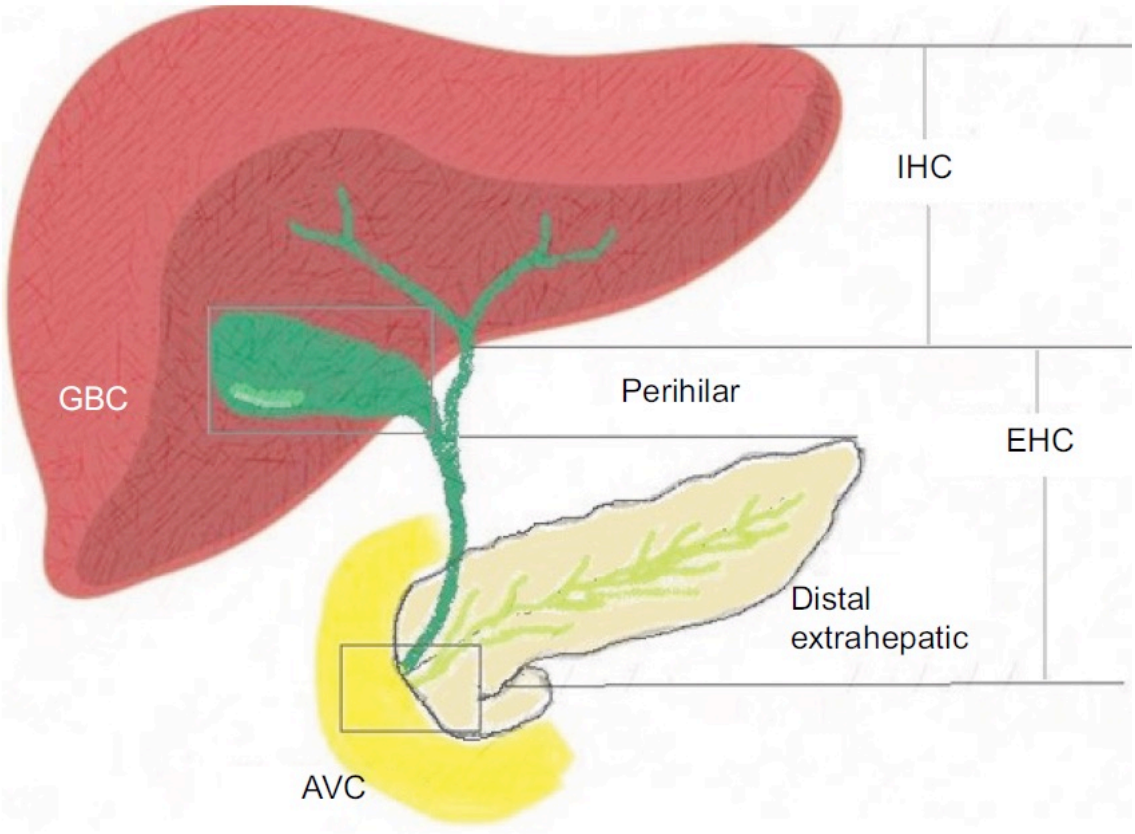


Figure 1. Anatomical sub-variants of CC. According to the location of the tumor, CCs are subdivided into gallbladder cancer (GBC), ampulla of Vater cancer (AVC), intrahepatic cholangiocarcinoma (IHC) and extrahepatic cholangiocarcinoma (EHC), further subdivided into perihilar and distal extrahepatic cholangiocarcinoma. (Adapted from Tariq N. et al.,2019)

CC accounts for approximately 3% of all gastrointestinal malignancies [2] and is the second most common hepatobiliary cancer, after hepatocellular carcinoma [3]. Although in most countries it is a rare disease (< 6 cases per 100,000 people) [4], its incidence is exceptionally high in some Asian countries (up to 85 cases per 100,000 people for northeast Thailand) due to different geographical

risk factors and genetic determinants [5]. In 2017, the global CC incidence was 211,000 cases, with 174,000 deaths [Global Burden of Disease Collaboration, 2019] and, over the past decades, both its incidence and its mortality have increased worldwide, in particular with regards to intrahepatic CC [6]–[8].

CC are aggressive diseases characterized by a poor prognosis (5-years survival rate = 5-15%, considering all stages) [9]. Moreover, since they are generally asymptomatic in early stages, most CCs are diagnosed at metastatic stage, when the 5-year survival rate is only 2% [1].

1.1.1. Clinical management of CC

Currently, the treatment of CC is not based on the anatomical subtypes, but solely on the stage of the disease and it essentially consists of surgery and systemic chemotherapy [10]. The only potentially curative therapy for CC is radical surgical resection, with a 5-year survival rate of 18% [11]. Unfortunately, only approximately 20% of patients present an early stage disease at diagnosis and are therefore eligible for surgery [12]. Moreover, the majority of patients undergoing surgical resection will relapse, predominantly developing liver metastasis [13]. In the advanced setting, the combination of cisplatin and gemcitabine has long represented the standard of care with a median progression-free survival (PFS) of 8 months [14]. A new option is adding immunotherapy, i.e., durvalumab, which further reduces the risk of progression by 25%, and increases survival by 20%. Yet more than half of patients with inoperable CC continue to die within a year from diagnosis. [15].

1.1.2. Molecular characterization as a future challenge for clinical management of CC

Considering the modest therapeutic efficacy of chemotherapy in CC, new therapies are urgently needed. For such purpose, a molecular characterization of CC is now helping revealing the complex mechanism the disease, opening to the possibility of a new clinical management.

In the last years, the widespread use of next generation sequencing (NGS) technologies revealed a complex genomic, epigenomic and transcriptomic landscape of CC, leading to the identification of

distinct molecular subtypes and new targets for molecularly informed treatments. Among these, clinical trials demonstrated significant activity for drugs targeting *IDH1* [16], *FGFR2* [17]–[19], *BRAF* [20] and *HER2* [21], which are now entering clinical practice. Nonetheless, the majority of CC does not harbor alterations in these genes [22], [23], therefore, the identification of new molecular biomarkers is an urgent unmet need. Considering the different anatomical subtype, alterations in *IDH1/2, EPHA1, BAP1* and *FGFR2* were more frequently found in intrahepatic CC, whereas gene fusions involving *PRKACA* or *PRKACB* and genetic aberration in *ARID1A, PI3KCA* and the ERBB family were detected in extrahepatic CCA (Figure 2); GBC was instead characterized by *ERBB3* and *EGFR* mutations [24].

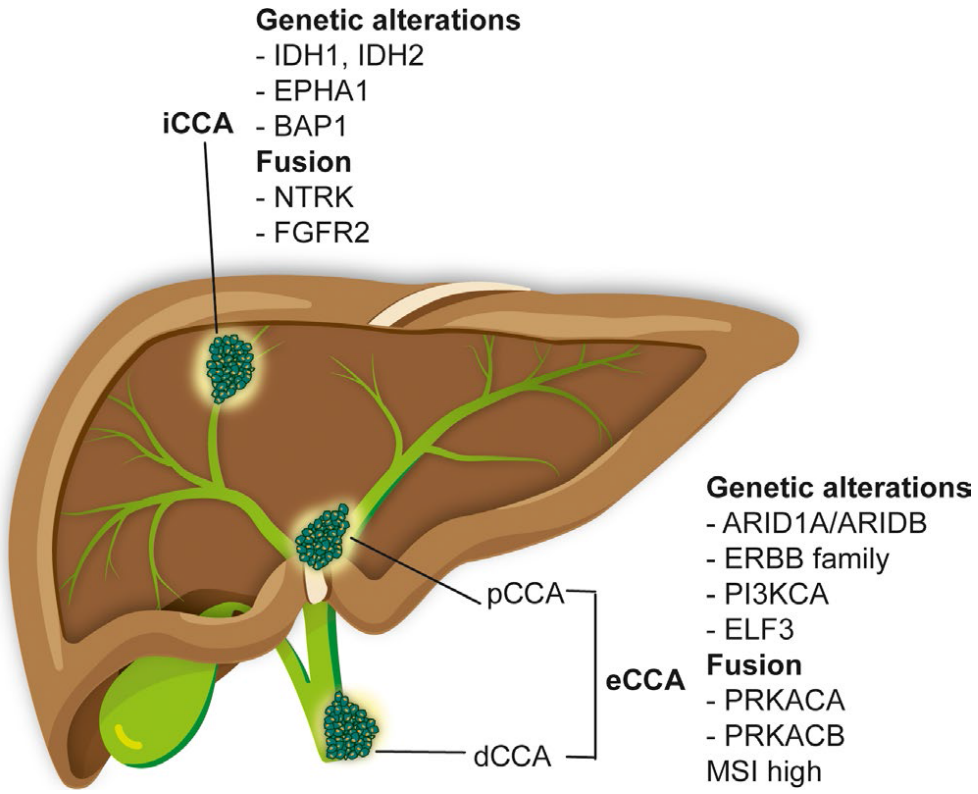


Figure 2. Molecular spectrum of intrahepatic and extrahepatic CC. The most unique and prevalent genetic alterations found in different anatomical locations are reported. ICC, intrahepatic cholangiocarcinoma; pCC, perihilar cholangiocarcinoma; dCCA, distal cholangiocarcinoma; ECC, extrahepatic cholangiocarcinoma. (Adapted from Braconi C. et al., 2019)

These results highlighted the high heterogeneity of CC and the need of including molecular profiling for clinical decisions. In this light, a series of transcriptomics and genomics studies based on NGS technologies on primary CC tumors attempted to elucidate the mechanistic insights of CC and to identify transcriptomic subtypes with a predictive and prognostic relevance. Considering 149 ICC cases, Sia *et. al.* demonstrated the presence of two classes, named as proliferation- and inflammation-related, associated with up-regulation of EGF, RAS, AKT, MET signaling and immune response-related pathways, respectively [25]. In another study, Andersen *et. al.* identified two prognostic subtypes and demonstrated the therapeutic potential of tyrosine kinase inhibition in CC cell lines with activated EGFR and HER2 signaling pathways [26]. Regarding ECC tumors, Montal *et. al.* identified in 189 patients the presence of 4 classes characterized by different transcriptomic and genomic patterns related to metabolic, proliferation, mesenchymal and immunological processes with comparable prognosis in terms of overall survival [27]. Finally, Nakamura *et. al.*, with a comprehensive analysis of ICC, ECC and including also gallbladder tumors, demonstrated the presence of 4 subgroups defined by specific gene expression and correlated genomic profile, associated with clinical outcome [28].

Despite these results represented a significant improvement in the biological understanding of CC, their clinical implications are still limited. The first reason is that all these classifications were generated from different and small cohorts and have not been compared and unified, thus limiting their applicability. Secondly, albeit these results represent a significant improvement for CC treatment, the clinical application is widely limited by the difficulty to obtain longitudinal tumor samples for monitoring treatment response and then establishing possible alternative therapies.

In this context, liquid biopsy represents a noninvasive approach able to detect function-related biomarkers reflecting tumor progression and treatment resistance and amenable for longitudinal analysis of cancer molecular features. In particular, circulating tumor cells (CTCs) released in blood by primary and metastatic lesions could replace invasive surgical biopsies, by informing on

biomolecular tumor features and anticipating the detection of progression and therapy–induced molecular changes [29].

1.2. Machine Learning techniques in Precision Medicine

Precision Medicine is an emerging approach to clinical research and patient care that focuses on understanding and treating disease by integrating multi-omics data from an individual to make patient-tailored decisions. With the large and complex datasets generated using diagnostic approaches like as NGS, novel techniques to process and understand these complex data were needed. At the same time, computer science has progressed rapidly to develop techniques that enable the storage, processing, and analysis of these complex datasets. Machine Learning is a collection of computer science methodologies that aims to identify complex patterns in data that can be used to train an automatic system in order to make predictions on new unseen data analysis with minimal or no further human intervention. The application of Machine Learning algorithms in the context of Precision Medicine data allows for broad analysis of large datasets and ultimately a greater understanding of human health and disease.

Overall, Machine Learning algorithms can be divided into row approaches: supervised and unsupervised learning. Unsupervised learning aims to uncover patterns in unlabeled data, identifying clusters of similar cases within a dataset. Popular unsupervised learning models include principal component analysis (PCA), hierarchical clustering, or variational autoencoders (an unsupervised deep learning architecture). On the contrary, supervised learning techniques aim to identify patterns in multi-dimensional dataset based on labelled data (e.g., healthy vs. disease or outcome scores). In particular, a training dataset with ground truth labels is typically used to build a model and to optimize the performance for the desired outcome [30][31]. The uncovered (learnt) patterns can then be used to classify new datasets or make data-driven, patient-individual predictions. Popular supervised statistical and machine learning techniques include, for example, Support Vector Machine (SVM), Random Forests (RF), Generalized Linear Models (GLM), and deep Neural Networks (NN).

1.3. Circulating tumor cells

In patients with solid tumors, CTCs are released from both the primary tumor and the metastatic lesions into the bloodstream during the course of the disease. Different technologies allow the detection and the characterization of CTCs, which are therefore considered a real-time liquid biopsy of tumors [32]. The term liquid biopsy also includes the analysis of other tumor-derived elements circulating in the blood, such as circulating tumor DNA (ctDNA), tumor-derived exosomes and microvesicles, tumor-educated platelets, circulating tumor microRNA, mRNA and non-coding RNA [33], each of which can provide different and complementary information. CTCs, in particular, being intact and viable cells, offer the possibility of performing a multilevel analysis of genotype (DNA) and phenotype (RNA and proteins). Moreover, they are a highly selected subpopulation of tumor cells, able to leave the primary tumor and survive in the bloodstream (the majority of CTCs die soon after entering the blood vessels, due to anoikis, attack by cells of the immune system and fluid shear stress, suggesting that CTCs could be representative of the most aggressive clones of the tumor [34]. Overall, CTC analysis can potentially be used for *i)* early detection of cancer, *ii)* prognostic stratification of patients, *iii)* identification of therapeutic targets, *iv)* prediction of response to targeted treatments, *v)* treatment monitoring and *vi)* identification of resistance mechanisms [35].

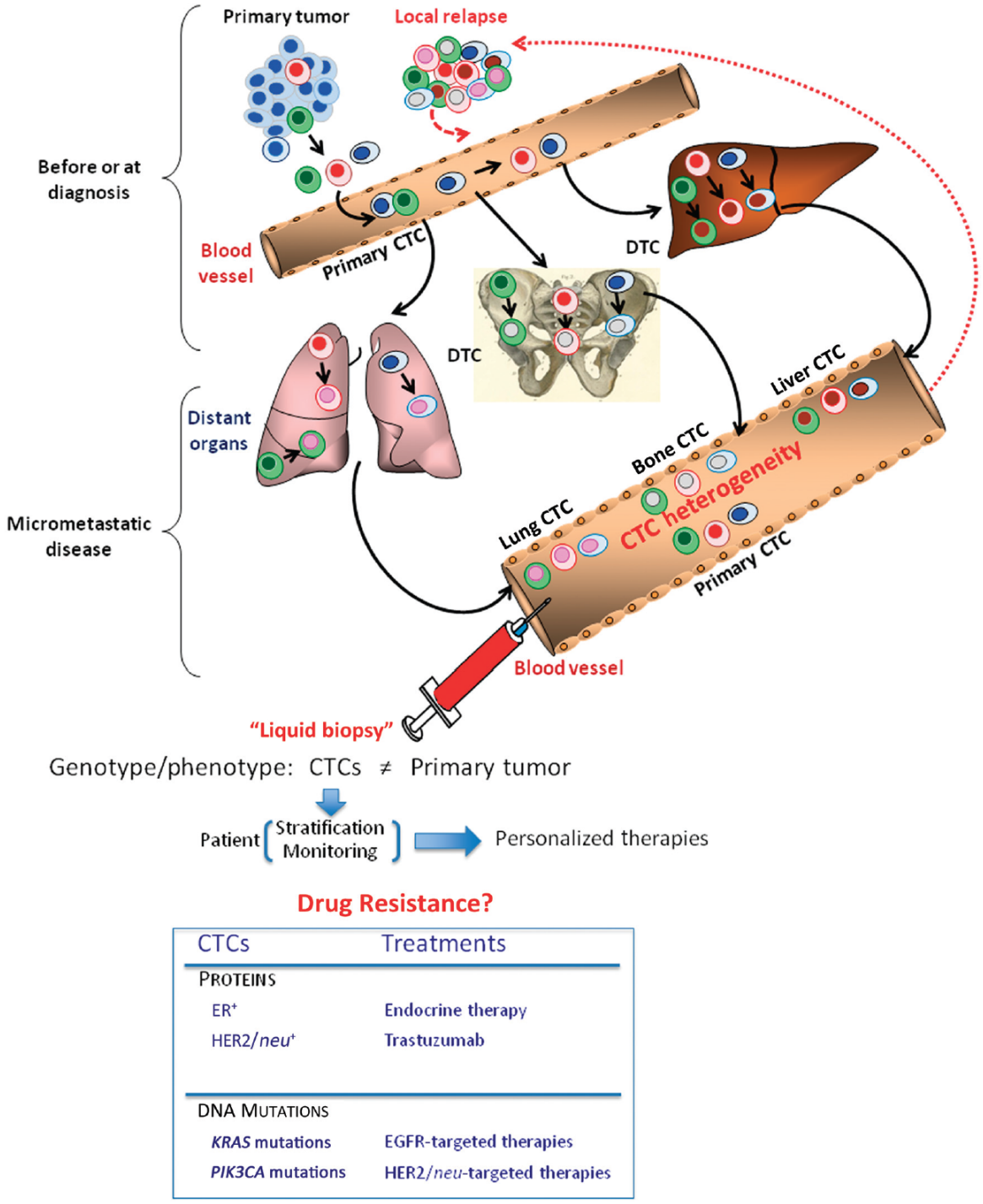


Figure 3. CTCs as a real-time liquid biopsy. CTCs can be derived from primary tumor or organs of metastasis. CTCs serve as a liquid biopsy of cancer and reveal important information on therapeutic targets and/or resistance mechanisms, which might be used in the future to stratify patients for such targeted therapies as inhibition of EGFR/HER2 or endocrine therapy and to monitor the efficacy of treatment and the development of resistance in real-time. ER+, estrogen receptor positive. (Adapted from Alix-Panabieres C. and Pantel K., 2013)

1.3.1. CTCs detection and molecular characterization

With regards to characterization approaches, recent technological advances (such as instruments for single-cell isolation and NGS) have allowed the development of methods for the genomic analysis of CTCs at the single-cell level [36]. These methods include, after the enrichment, a step for the isolation of single cells. Cells can be individually isolated by laser capture microdissection and fluorescence-activated cell sorting (FACS), or by using specific instruments such as the DEPArray™ which isolates single cells by exploiting dielectrophoresis [37], microscopic manipulators as the CellCelector™ [38] and microfluidic devices.

Independently from the type of isolation method used, all isolated cells will undergo the whole-genome amplification (WGA) in order to be analyzed, that is based on two steps: PCR and multiple displacement amplification (MDA). Therefore, after WGA, quality control (QC) assays are performed to assess the DNA yield and the length of amplified fragments. Good quality samples can undergo any type of sequencing analysis (including Sanger sequencing, array comparative genomic hybridization (aCGH) platforms or genome-wide NGS) for the detection of a variety of genomic alterations including small-scale alterations (single nucleotide variants (SNVs), INDELs and microsatellite instability), and large-scale alterations (copy number variations (CNVs), chromosomal breakpoints or large-scale state transitions (LSTs), and chromosomal rearrangements) (Figure 4).

The methods for CTC characterization at the single-cell level opened a new chapter of liquid biopsy research, aimed at characterizing and monitoring changes in tumor heterogeneity in individual patients to further understand the biology of tumor evolution.

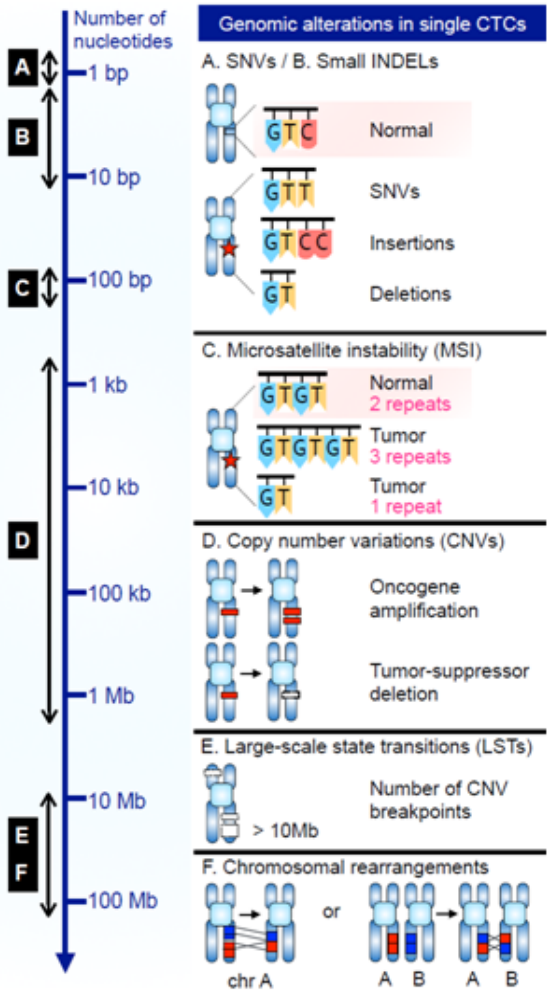


Figure 4. Genomic alterations in single CTCs. List and characteristics of the type of genomic alterations that can be detected in CTCs by single-cell sequencing. (Adapted from Lim S.B. et al., 2019)

2.SCOPE OF THE THESIS

We hypothesize that CTCs molecular characterization could represent a fundamental for monitoring cancer evolution within patients. Acknowledging the importance of CTCs as tissue surrogate suitable for clinical management, the inference of specific molecular signature associated to primary tumor on CTCs offers a real opportunity for monitoring tumor evolution under therapy pressure giving rise to a predictive tool that is still lacking not only in the context of CC but also in other cancer types.

The aims of the current project were: (1) to collect publicly available CC datasets of normal and primary tumor tissues for the extraction of tumor specific molecular signatures; (2) to integrate and validate different molecular signatures to define biological distinct subgroups and create a predictive tool for their identification; (3) to collect and analyze by genomic approaches CTCs from CC patients; (4) to apply the predictive tool on CTCs data for monitoring cancer evolution under treatment pressure.

With this project, we expected to: (1) infer specific molecular signature related to CC biological distinct subgroups on CTCs collected along treatment to monitor cancer evolution within single patient; (2) improve the understanding of tumor heterogeneity based on CTCs biological subgroups and progression within individual patients; (3) establish a frame for predictive tools informative in CC and in other tumor types.

3.MATERIAL AND METHODS

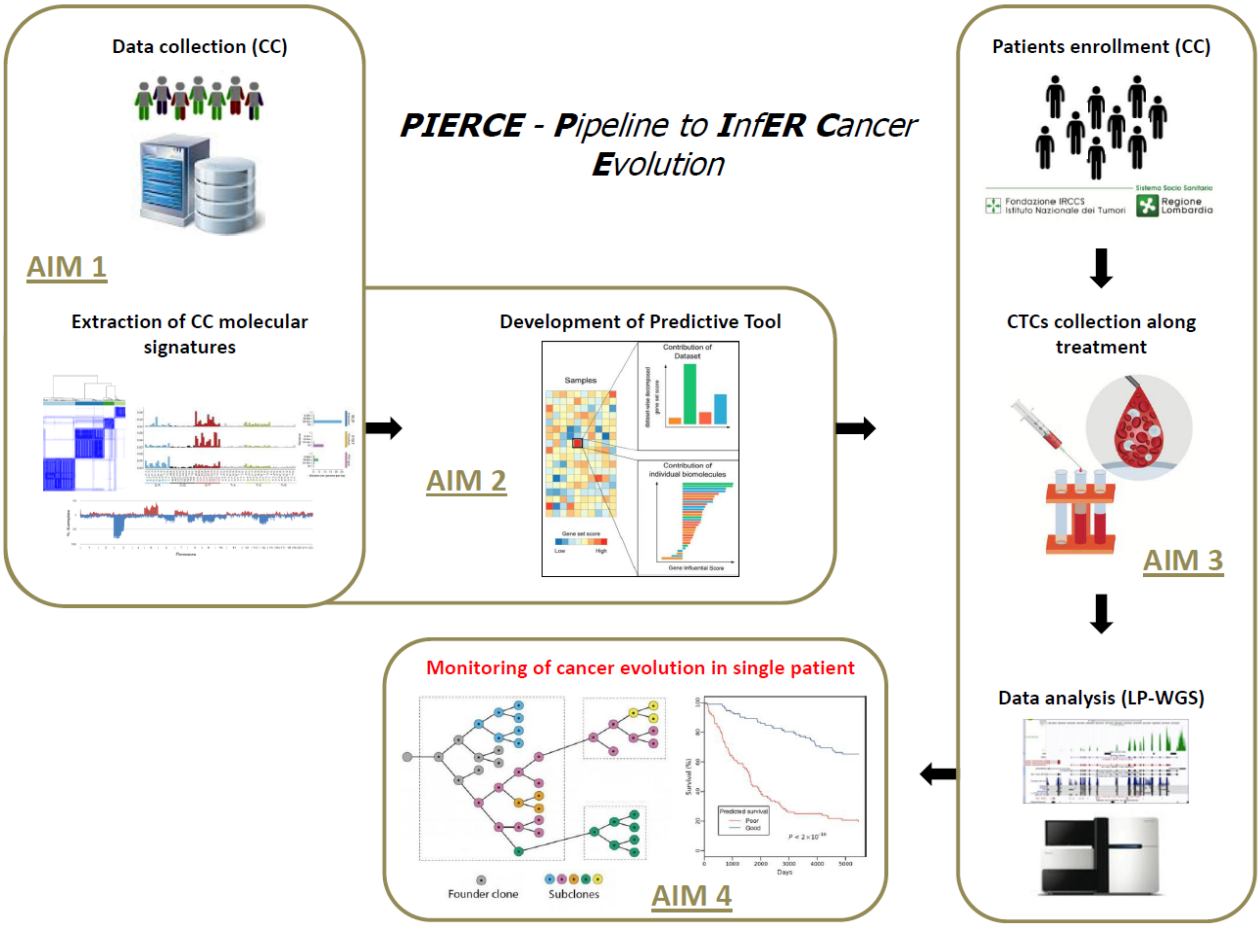


Figure 5. PIERCE pipeline. The workflow summarizes all the steps of the current project.

3.1. Retrieval of public ICC and ECC datasets and definition of discovery and validation sets

Gene expression and genomic profiles were collected from eight public datasets, including 1000 patients, i.e., 715 (71.5 %) ICC and 285 (28,5%) ECC with no previous history of hepatitis or fluke infection, and generated by microarray (6) or RNA sequencing (RNAseq, 2) or Whole genome sequencing (WGS, 1) or Whole exome sequencing (WES, 2) or Target sequencing (Target-seq, 3). Overall, 126 of 664 (19%) were from fresh frozen (FF) tumor tissues, while 538 (81%) were from formalin-fixed paraffin embedded (FFPE) specimens. For all the downstream analyses, seven datasets were used for discovery; whereas the EGAD00001001693 dataset was used for validation given that it was the only one with exhaustive information on patient survival (**Table 1**).

Table 1. Collection of gene expression and genomic dataset considered for ICC and ECC discovery and validation sets.

Histotype	Dataset ID	Specimen	Reference	No. of samples	Platform	Discovery (D) Validation (V)
ECC	GSE132305	FFPE	*Montal R., et. al., Journal of Hepatology 2020*	182	Microarray	D
ICC	GSE32225	FFPE	*Sia D. et. al., Gastroenterology 2013*	141	Microarray	D
ICC, ECC	GSE89749, EGAD00001001988	FFPE	*Jusakul A. et. al., Cancer Discovery 2017*	139	Microarray, WGS, Target-seq	D
ICC, ECC	MSK2018	FFPE	*Lowery M. A. et. al., Clinical Cancer Research 2018*	122	Target-seq	D
ICC	MSK2021	FFPE	*Boerner T. et. al., Hepatology 2021*	123	Target-seq	D
ICC	GSE26566	FF	*Andersen J.B. et. al., Gastroenterology 2012*	103	Microarray	D
ICC, ECC	TCGA-CHOL	FFPE	*Farshidfar F. et. al., Cell Reports 2017*	36	RNAseq, SNP6, WES	D
ICC	GSE32879	FF	*Oishi N. et. al., Hepatology 2021*	16	Microarray	D
ICC	GSE57555	FF	*Murakami Y. et. al., Scientific Reports 2015*	7	Microarray	D
ICC, ECC	EGAD00001001693	FFPE	*Nakamura H. et. al., Nature Genetics 2015*	131	RNAseq	V

3.2. Data processing

3.2.1. Gene expression data

For the training set, we used the transcript quantification data of the TCGA-CHOL cohort (level 3) obtained using RSEM (level 3) downloaded from Firebrowse database (accession date: January 2022). TPM values followed by quantile normalization and log2 transformation were considered. All the microarray data were retrieved from GEO using a bespoke R pipeline. Considering Illumina microarray dataset, robust-spline normalization followed by log2 scaling were used to normalize gene expression values. Affymetrix and Agilent datasets did not require any additional normalization after download from GEO database.

Regarding the validation set, RNAseq data were downloaded from the EGA repository after the accession permission released by ICGC consortium (application number: DACO-6992). Raw sequencing data (fastq format) were trimmed to remove low-quality bases and adapters using trimmomatic (version 0.39) [39], quality checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and aligned to human reference genome

(hg19) using STAR (version 2.7.9a) [40]. After alignment quality control by bedtools (version 2.25.0) [41] and qualimap (version 2.2.2-dev) [42], counting of reads aligned over exonic features for gene expression quantification was performed by RSEM (1.3.1) [43]. TPM values related to each gene were considered and submitted to quantile normalization and log2 scaling. All the processing (discovery set) and post-processing analysis (validation set) were performed with R software (https://www.R-project.org/, version 4.1.1, see section "list of R packages").

3.2.2. Genomic data

For the training set, mutational and copy number alteration data of the TCGA-CHOL cohort obtained using SNP6 array and WES platforms were downloaded from Firebrowse database (accession date: January 2022). Target-seq data related to MSK-2018 and MSK-2021 were obtained from cBioportal (accession date: January 2022) and all samples with tumor purity score < 20 were excluded from the analysis. Mutation and copy number alteration data of EGAD00001001988 dataset were downloaded from cBioportal (accession date: Januray 2022) and from EGA repository after the accession permission released by ICGC consortium (application number: DACO-6992).

Regarding the validation set, WES data were downloaded from the EGA repository after the accession permission released by ICGC consortium (application number: DACO-6992). Raw sequencing data (fastq format) were trimmed to remove low-quality bases and adapters using BBDuck tool included in BBTools package (https://sourceforge.net/projects/bbmap, version 38.98), quality checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and aligned to human reference genome (hg19) using mrsfast (version 3.3.9) [44]. After alignment quality control by bedtools (version 2.25.0) [41] and qualimap (version 2.2.2-dev) [42], CNA analysis were performed using ichorCNA R package [45]. For mutation, the data were retrieved from cBioportal (accession date: January 2022) and all the variants showing at least one of the following parameters were escluded: depth of coverage < 20; non coding region; FFPE artifact (vaf ≤ 10% & C>T or vaf ≤ 10% & G>A); vaf ≤ 5%; vaf > 75%; gnomAD_EAS_AF ≥ 1%; gnomAD_AF ≥ 1%. All the

processing (discovery set) and post-processing analysis (validation set) were performed with R software (https://www.R-project.org/, version 4.1.1, see section "list of R packages").

Figure 5 reports in detail all the steps of the pipeline adopted in the present work.

3.3. Data integration and unsupervised clustering analysis of ICC and ECC discovery sets

3.3.1. Gene expression data

Normalized gene expression data of each dataset included in the ICC and ECC discovery sets were merged and only common transcripts (n=13,228) were considered. Then, quantile normalization followed by batch adjustment based on empirical Bayes method [46] were performed to make data comparable and remove batch effect associated with the different samples source (FF, FFPE) and platforms to obtain gene expression profiles. GTEx database (accession date: February 2022) was interrogated to remove 386 liver-specific transcripts associated to normal tissues. A filtering step was performed using a custom made pipeline to reduce the number of features in the ICC (n=1358) and ECC (n=676) cohorts, respectively. In particular, the method allows to identify genes that drive biological heterogeneity in a dataset decomposing the total variance of each gene into its biological and technical components by fitting a trend to the endogenous variances [47]. For each gene, the fitted value of the trend is an estimate of the technical component while the biological component is retrieved by subtracting the technical component from the total variance. For the ICC training set, genes with significant biological component were selected using a criteria based on a FDR < 0.05. Due to the small number of datasets in the ECC training set, the filtering step was performed separately for each dataset (block) and only genes with FDR ≤ 0.1 in each block were selected.

Detection of distinct subgroups within the ICC and ECC training set was performed by hierarchical clustering analysis with number of clusters selection based on NbClust tool [48]. In particular, 19 out 30 indices were considered to validate the number of the clusters (n min=2, n max=6) along with Euclidean distance and Ward's linkage method [49].

3.3.2. Genomic data

For mutational data, cleaned gene variants of each dataset included in the ICC and ECC discovery sets were merged and only common genes (n=305) among the platforms (WES, Target-seq) were considered. Only fequently altered genes (frequency of mutation > mean of the number of genes variants) were considered for unsupervised clustering analysis in ICC (n=298) and ECC (n=77) cohorts, respectively.

For CNA data, alterations at gene level of each dataset included in the ICC and ECC discovery sets were merged and only common genes (n=474) among the platforms (WES, Target-seq) were considered. Only fequently altered genes (gene copy number gain and loss altered in at least two different datasets) were considered for unsupervised clustering analysis in ICC (n=69) and ECC (n=8) cohorts, respectively.

Detection of distinct subgroups within the ICC and ECC training set was performed by hierarchical clustering analysis with cluster number selection performed using NbClust tool for both mutational and CNA data[48]. In particular, 19 out 30 indices were considered to validate the number of the clusters (n min=2, n max=6) along with Ward's linkage method [49] and with Jaccard distance computed over variant classification type (SNV, INDEL, mixed) associated to genes in each samples.

3.4. Biological and immunological characterization of ICC and ECC subgroups

Considering gene expression data ICC and ECC discovery sets, differential expression analysis between subgroups were performed using gene level linear models with moderated t-test (LIMMA) [50]. T-statistic values were used to rank gene list for the gene set enrichment analysis by GSEA [51] considering HALLMARK and C2 canonical pathway database [52]. For each cluster comparison, significant up-regulated/down-regulated pathways were selected according to thresholds of p-value < 0.05 and p-value < 0.01 for HALLMARK and C2 canonical pathway gene sets, respectively. Singscore tool [53], a single sample scoring method, was used to evaluate the enrichment of specific cholangiocarcinoma pathways described by Banales *et. al.* [54] within each ICC and ECC subgroups.

T-test statistic was applied to detect significant differences in terms of enrichment score. ImmuCellAI [55] was used to evaluate immune cell component within each ICC and ECC subgroups based on the deconvolution of gene expression profiles of 24 immune cell. A t-test statistics was applied to detect significant difference between subgroups in terms of immunological infiltration score.

Considering genomic data, biological characterization was performed considering canonical oncogenic signaling pathways defined by Vega and colleagues [56] through ad hoc function implemented in maftools R package [57]. Moreover, information about the presence/absence of druggable alterations (mutation and CNA) at gene level were retrieved from "The drug gene interaction database" considering clinically actionable level [58]. Based on previously reported by Nhgia Vu *et. al.* [59], a two-test statistic approach (odds ratio plus + chi-squared tests) was applied to evaluate the subgroups specificity of each druggable gene. The level of significance for each test was consider at p-value < 0.05.

3.5. Building of ICC and ECC predictors

For gene expression data, leveraging on differential expression analysis by LIMMA, two different approaches were adopted to define the genes to be considered for the predictor establishment in ICC and ECC discovery set. For ICC cohort, genes with a log2 fold change (log2FC) < -1 and > 1 and FDR < 0.1 values within each comparison were selected (n=19). Considering the ECC cohort, among the top 60 differentially expressed genes for each group based on t-statistic values, only the transcripts with FDR < 0.25 in all subgroups were selected (n=21).

For ICC mutational data, only druggable genes significantly associated to each subgroup base on the application of two-test statistic approach (see section 3.4 of materials and methods) were considered (n=17). Moreover, to reduce the misclassification rate inside predictors, all the combinations among the 17 genes were tested, resulting in the selection of 13 genes. Within ECC cohort, shared mutated genes between discovery and validation set were considered for building the predictors.

Considering CNA data, only shared alteration between discovery and validation set were considered for the predictor establishment.

The SMOTE (Synthetic minority over-sampling technique) algorithm [60], was applied to account for sample imbalance between the subgroups within ICC and ECC cohort. Three different Machine Learning algorithms were considered in order to build the predictors: k-Nearest Neighbors (KNN) [61], Support Vector Machine (SVM) [62] and Random Forest (RF) [63]. In order to evaluate the KNN and SVM methods, training and testing-sets were created sampling the 60% and 40% of ICC and ECC discovery sets, respectively. Moreover, a 10-fold cross validation to test the performance of the models was applied. For RF algorithm, bootstrap without replacement was considered as resampling method to derive the estimates of standard errors and confidence intervals. ROC curves and AUC were used to visualize and evaluate the performance of the classifiers for each ICC and ECC subgroup.

3.6. Prediction and evaluation of biological subgroups in ICC and ECC validation sets

Starting from the same genes considered for the unsupervised clustering analysis in discovery set, NbClust was applied to validation sets. In particular, 19 out 30 indices were considered to validate the number of the clusters along with Ward's linkage method and euclidean and jaccard distance for gene expression and genomic data, respectively [49]. For ICC validation set the minimum and maximum number of groups admitted were 2 and 6 while for ECC, due to the low number of samples (n=29), the range was set to 2 and 5. At the same time, predictors based on RF algorithm were applied to predict the presence of the 4 biological subgroups in ICC and ECC validation sets. Diagonal dominant matrix [64] approach along was used to understand the concordance between NbClust-based and predictors-based classification. For gene expression data only, unsupervised clustering analysis based on the median expression of the 19 and 21 genes of each biological sublclass identified by the two classification methods were applied in ICC and ECC validation sets. Euclidean distance and Ward.D linkage were considered for the analysis.

3.7. Survival analysis

Survival analysis methods were used to analyze overall survival (OS) and relapse free survival (RFS). OS was calculated from the date of disease diagnosis to death or last follow-up, while RFS was calculated from the date of disease diagnosis to the first event (i.e. disease relapse or death). Patients were divided into biological subgroups returned by RF predictor and unadjusted *p*-values were calculated using log-rank test considering p-value < 0.05 as threshold for statistical significance. For the ICC cohort, 3 patients were excluded from the survival analysis due to missing OS and RFS information.

3.8. Patient information and clinical sample collection

This was a prospective, monocentric, observational study conducted at Fondazione IRCCS Istituto Nazionale dei Tumori (Milan, Italy). For this study, 24 patients with a confirmed diagnosis of metastatic/unresectable CC were consecutively recruited between January 2015 and March 2017. The number of enrolled patients was consistent, with the entropy-based approach to sample size in translational clinical trials as proposed by Piantadosi and colleagues [65].

Patients have been treated and followed up as per clinical practice, with frequent clinical evaluations and tumor assessment with chest/abdomen CT scans and/or MRIs performed every 2-3 months. The treatment efficacy was assessed according to RECIST v1.1. Clinical information was collected from medical records and included demographic data, tumor anatomical location, tumor extension, and treatment history. The patients' vital status was updated at the end of June 2018.

All CTC evaluations were carried out without the knowledge of the patient's clinical status.

Samples of peripheral venous whole blood (10 mL) were drawn in EDTA tubes (K_2EDTA BD Vacutainer®, Becton Dickinson, Franklin Lakes, NJ, USA), stored at 4 °C protected from light and processed within 1 hour for CTC enrichment (the first mL of blood was discarded to avoid skin epithelial cell contamination). Blood samples were longitudinally collected at times corresponding to

baseline (BL), *i.e.* before initiation of a new treatment line, during treatment (DT) close to clinical and imaging evaluations, at the end of treatment (EOT) and at subsequent follow-up (FU) or new treatment lines.

All subjects have signed a written informed consent form accepting participation in this study, which was approved by the local ethical board in November 2014 (INT 177/14) and subsequently reconfirmed in January 2018.

3.9. CTCs processing

3.9.1. Collection and processing

Blood samples (10 mL) collected in K2EDTA tubes were subjected to CTC enrichment with Parsortix™ (Angle plc, Guildford, UK) within 1 h from blood draw. Enriched cells were harvested according to manufacturer's instructions and fixed for 20 min at room temperature (RT) with 2% paraformaldehyde. Fixed samples were stained immediately or within 24 h from enrichments.

Fixed samples were fluorescently stained with phycoerythrin (PE)-labeled antibodies against epithelial markers EpCAM (clone HEA-125, Miltenyi Biotec, Bergisch Gladbach, Germany, working dilution 1:11 for 10 min at 4° C), cytokeratins (pan cytokeratin clone C11, Abcam, San Francisco, CA, USA, and pan cytokeratin clone AE1/AE3, NSJ Bioreagents, San Diego, CA, USA, working dilution 1:10 for 10 min at RT) and EGFR (clone 423103, Santa Cruz Biotechnology, Dallas, TX, USA, working dilution 1:11 for 10 min at 4° C), and with allophycocyanin (APC)-labeled antibodies recognizing leukocytes and monocytes: CD45 (clone 5B1, Miltenyi Biotec, working dilution 1:11 for 10 min at 4° C), CD14 (clone M5E2, BD Biosciences Pharmigen, San Diego, CA, USA, working dilution 1:20 for 10 min at 4° C), and CD16 (clone 3G8, BD Biosciences Pharmigen, San Diego, CA, USA, working dilution 1:20 for 10 min at 4° C). Nuclei were stained with 1 g/mL Hoechst 33342 (Sigma-Aldrich, Saint Louis, MI, USA) for 5 min at RT. Labeled cells were analyzed using the DEPArray™ - (Menarini Silicon Biosystems, Bologna, Italy) within 2 days from staining to visualize and recover single cells manually selected based on fluorescence labeling and morphology.

Selected single epithelial or double-negative (PE-ve/APC-ve) cells were recovered for downstream molecular analyses.

3.9.2. Molecular characterization

Recovered single cells were subjected to whole genome amplification employing the Ampli1™Low Pass kit for Ion Torrent (Silicon Biosystems), pooling 16 or 24 samples depending on the amplified DNA quality or the Ampli1™Low Pass kit for Illumina (Silicon Biosystems, pools of 96 samples). Libraries were subjected to sequencing with the IonTorrent Ion S5™ system (Thermo Fisher) using the Ion530 chip as for manufacturer's instructions, or with HiSeq system (Illumina).

Considering Ion Torrent sequencing, samples with failed QC returned by Ion Reporter software were excluded from the analysis. For Illumina sequencing, "Per base sequence content", "Per sequence GC content", "Sequence length distribution" and "Overrepresented sequences" were considered as parameters during the QC analysis in fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc, accessed on 15 January 2021). Samples with > 1 "Fail" returned by software and with aligned read counts lower than 400,000 were excluded from the analysis.

3.9.3. Sequencing data analysis and evaluation of CTCs phylogenie

WGS sequences were aligned to the Human Reference Genome (hg19) using tmap (Torrent_Suite 5.10.0) and bwa aligner tool for Ion Torrent and Illumina samples respectively. CNAs were predicted by using QDNAseq R package [66].

Considering the evaluation of CNA profile, chr19 was not considered due to its biased deletion associated with the high CG base percentage. The discrimination between aberrant (CTCs) or normal cells (WBC) was based on previously published criteria related to the amount and distribution of genomic aberration [67]. Segmented copy number data of each sample were extracted starting from logRatio value.

TraslationalOncology (TRONCO) pipeline, an assortment of algorithms to infer progression models via the approach of Suppes-Bayes Causal Network were used to perform phylogentic analysis on

CTCs patients collected in our Institute [68]. In particular, the CNAs prioritization returned by CAPRESE, an algorithm that use a shrinkage-alike estimator combining correlation and probability raising among pair of events, was used to identify and to map clonal relationship along time. All the steps related to the optimization of phylogenetic analysis are currently ongoing.

3.10. List of R packages

WGCNA, UBL, tidyr, sva singscore, RColorBrewer, randomForestSRC, plyr, org.Hs.eg.db, NMF, NbClust, multiROC, lumi, limma, lattice, irr, GSA, gplots, ggthemes, ggpubr, ggplot2, ggfortify, GEOquery, FunCluster, fgsea, factoextra, edgeR, dplyr, DBI, data.table, ComplexHeatmap, circlize, caret, Biobase, ichorCNA, QDNAseq and maftools.

4.RESULTS

4.1. Aim1: Collection of publicly available CC datasets of primary tumor tissues and molecular signatures extraction

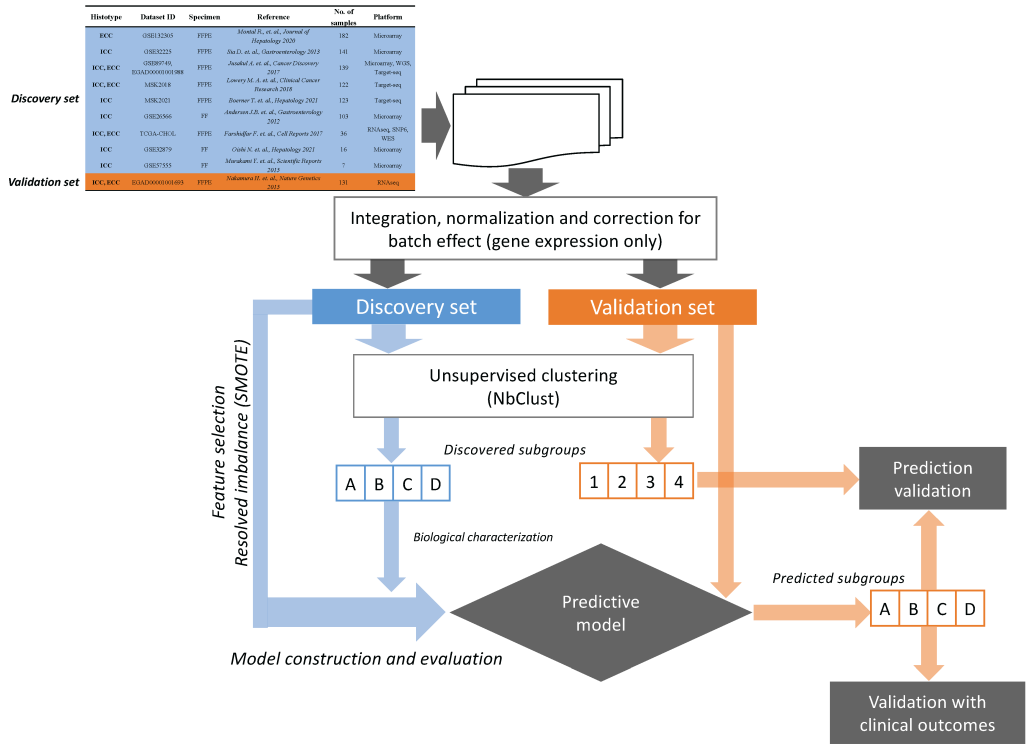


Figure 6. Workflow of aim1. The pipeline reports all the steps used in the aim 1 of the present works, from data collection to the clinical evaluation of the defined ICC and ECC subgroups in validation set.

4.1.1. Gene expression data

4.1.1.1. Integration of data from different platforms for the establishment of discovery set

A total of 340 ICC and 203 ECC patients profiled for gene expression data with no previous history of hepatitis or fluke infection were collected from seven different datasets and considered as discovery dataset. Overall, the intersection of the different platforms used to characterize each sample allowed to obtain the expression profile of 13,228 genes. The application of preliminary steps for sample normalization and batch effect adjustment lead to a higher correlation between each sample in the cohorts (range: 0.5-1 for ICC; range: 0.6-1 for ECC) and to efficient batch correction.

Liver-specific genes (N=386) obtained from GTEx database were removed to avoid contamination given by transcripts non biologically associated to CC [69]. Moreover, the additional filtering step

performed based on the mean expression and variance of each genes allowed to select 1,358 and 676 genes in ICC and ECC cohort, respectively.

4.1.1.2. Unsupervised clustering analysis identifies distinct biological subgroups in ICC and ECC cohorts

Unsupervised clustering analysis using NbClust led to the identification of 4 subgroups within both the ICC (N = 340) and ECC (N = 203) cohort (Figure 7A-B), with no evidence of batch effect related to the different datasets forming each cohort.

To investigate the presence of distinct biological and immunological features of the newly defined subgroups, we collected the differentially expressed genes subgroups of each group. Results highlighted significant differences between each subgroup, with the distribution of up- and down-regulated genes that varied among the ICC and ECC cohort. To further understand the unique biological traits of the subgroups, enrichment analysis along with the evaluation of immunological components were performed.

Regarding the ICC cohort, up-regulated pathways/signatures characterized distinct biological features of each subgroup, mainly related to apoptosis and progression (subgroup A), metabolism and TGF-ß (subgroup B), cell cycle and DNA replication (subgroup C), DNA repair and KRAS down-regulation (subgroup D). Interestingly, the evaluation of the immunological components revealed the presence of immune cell infiltration in all samples, with subgroup A being enriched only with exhausted T cells (Figure 7A). The enrichment analysis on samples of the ECC cohort also showed distinct biological traits associated to each subgroup, related to EGFR and MYC target (subgroup A), ECM and cell adhesion/interaction/motility (subgroup B), immune pathways and cell-cycle (subgroup C), and metabolic pathways (subgroup D). Moreover, all the subgroups presented different immunological components, with subgroup C being characterized by the highest rate of immune cells infiltration (see Figure 7B).

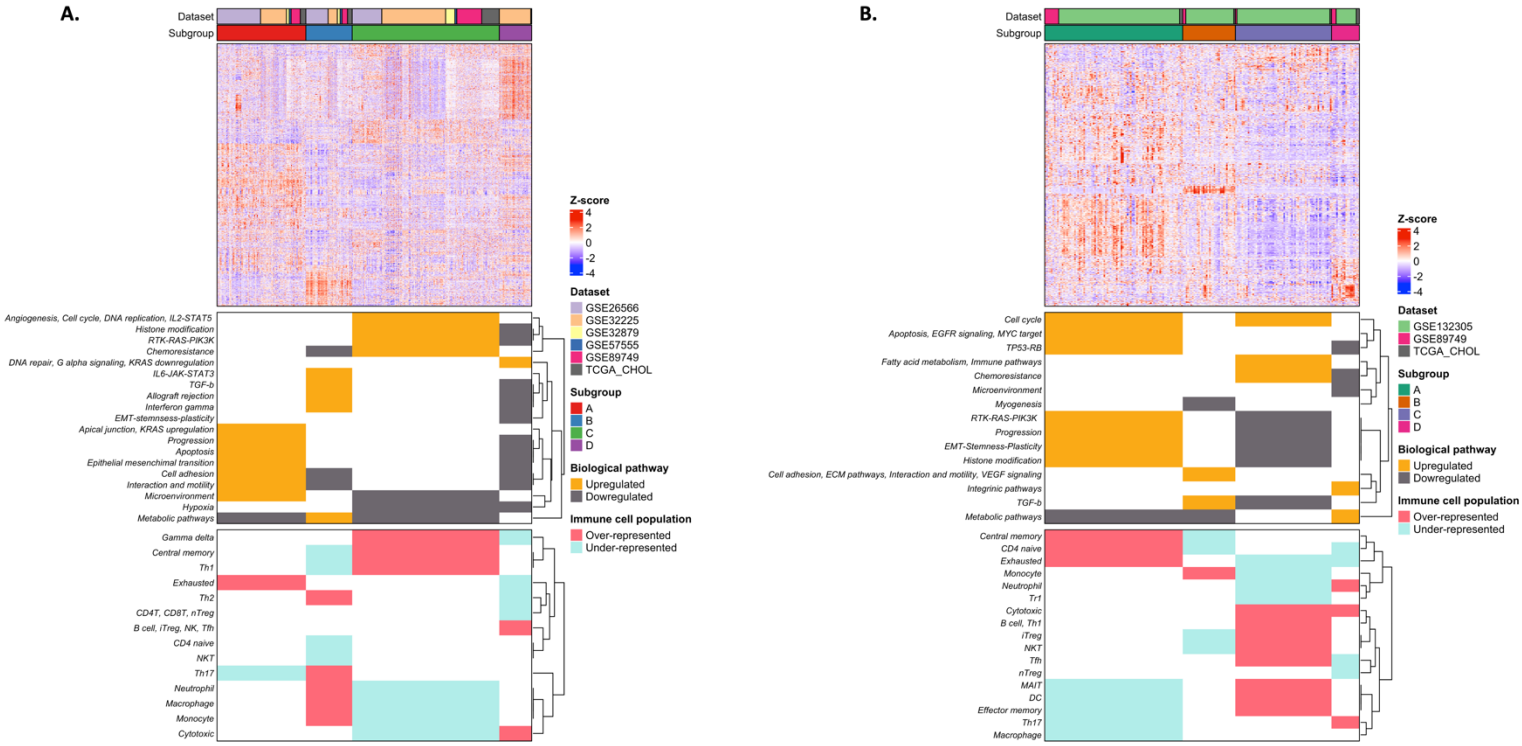


Figure 7. Identification of ICC and ECC distinct biological subgroups. Unsupervised clustering analysis using NbClust identified four subgroups in both ICC (A) and ECC (B) discovery set. The heatmaps reports samples on the column and genes on the rows. For each sample, dataset of origin and the cluster group membership are reported as color bars on the top of the heatmap. In order to evaluate the biological characteristics associated to each subgroup, GSEA, Singscore and ImmuCellAI were applied on ICC and ECC cohort. Upregulated/downregulated pathways and Over-Under represented immune cell populations (rows) associated to each ICC and ECC subclass (columns) were reported as annotation below the heatmaps.

4.1.1.3. Supervised classifier to identify the proposed subgroups using gene expression

The presence of distinct ICC and ECC biological subgroups within the discovery set opened the possibility to build and compare specific predictors to validate our findings in an independent dataset (validation set). For such a purpose, feature selections were performed specifically for ICC and ECC cohorts starting from the differential expression analysis between each subclass. A total of 19 and 21 genes were selected for ICC and ECC samples, respectively and used to build the predictor from training set. Due to imbalance between subgroups, we adopted the SMOTE algorithm in order to oversample the minority class prior to classifier definition on the discovery set. Among the 3 machine learning algorithms tested, RF showed the best performance with an overall misclassification rate of 0.08 and 0.07 for ICC and ECC, respectively. Moreover, RF predictor showed an AUC value of 0.98 (KNN=0.84; SVM=0.9) and of 0.99 (KNN=0.95; SVM=0.93) for ICC and ECC cohort, respectively.

In particular, the classification error associated to each subclass were similar except for group A in the ECC cohort, where a value of 0.17 was obtained (Figure 8).

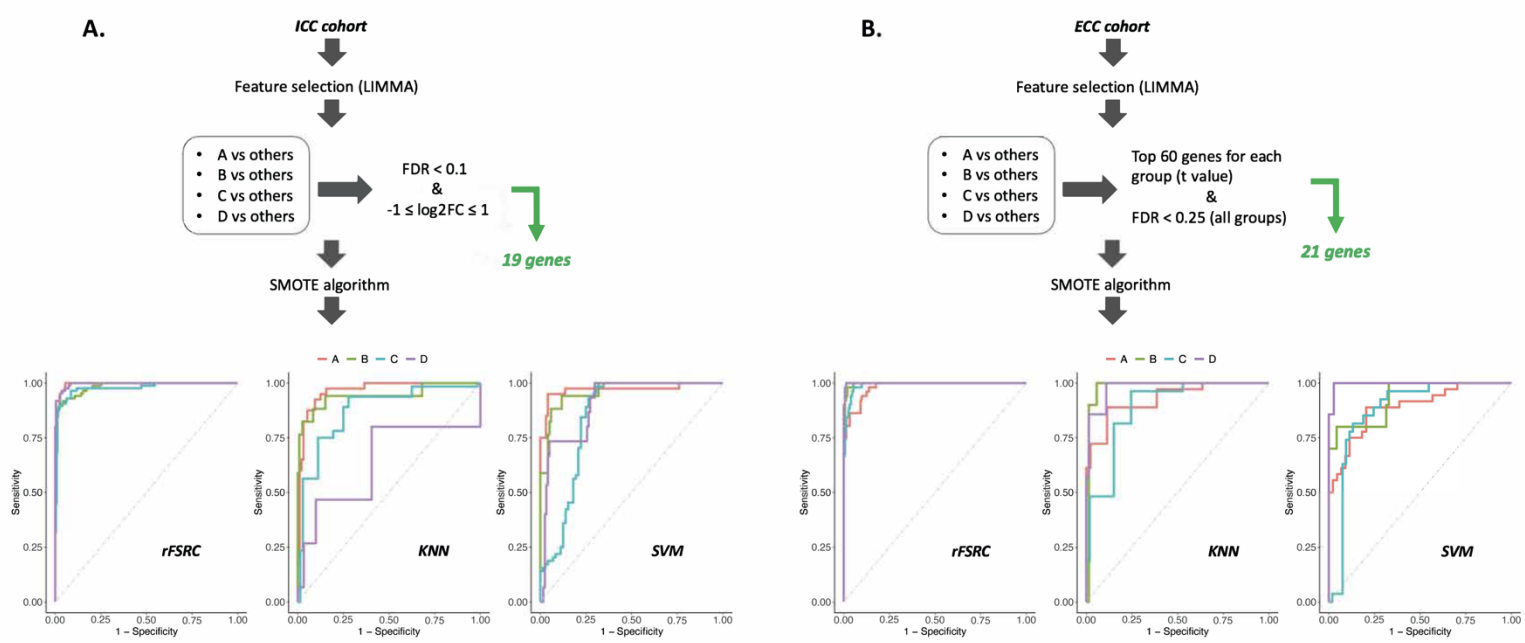


Figure 8. Building of predictors related to the distinct biological subgroups in ICC and ECC cohort. For ICC, class comparison results obtained from LIMMA allowed to select 19 genes specifically associated to each subgroup (FDR < 0.1 & -1 ≤ log2FC ≤ 1) (A). Within ECC, class comparison results obtained from LIMMA allowed to select the top 60 differentially expressed genes based on t value. Then, the filtering based on FDR value (FDR < 0.25 in all groups) allowed to identify 21 genes specifically associated to each subgroup (B). For both ICC and ECC cohort, SMOTE algorithm was applied to balance the distribution of the samples within subgroups and three different machine learning methods were evaluated for the establishment of the predictor. The ROC curves report specificity and sensitivity values obtained from RF, KNN and SVM algorithms.

4.1.1.4. Identification of ICC and ECC biological subgroups in validation set

To validate and establish the clinical significance of our predictors, an independent dataset composed of 131 patients (102 ICC, 29 ECC) profiled by RNAseq was considered as ICC and ECC validation sets (Table 2). Similarly to the discovery set, all of the patients did not present a previous history of hepatitis or fluke infection. After transcripts quantification, normalization and gene filtering steps, we used a RF-based classifier to predict the presence of the 4 specific biological subgroups in both ICC and ECC of the validation set. In the ICC cohort, the subgroups A (n=43) and D (n=1) were characterized by the highest and lowest number of patients, respectively, consistently with the discovery set composition. For the ECC cohort, most of the patients were identified in subgroups A (n=24) and only a few samples were assigned to the B (n=2), C (n=2) and D (n=1) subgroups.

To validate the subgroups predicted by the supervised classifier, we compared them to the subgroups discovered by unsupervised clustering analysis on the validation sets, which were independent from the discovery sets. The method was applied on the same genes (N=1,358 for ICC; N= 676 for ECC) used to detect the subgroups in the discovery dataset. In order to match the group labels provided by the unsupervised hierarchical clustering to the predicted subgroup classes, we tabulated the unsupervised labels versus the predictor-based groups is a 4x4 matrix and searched for the optimal matching based on dominant diagonal matrix (the matrix that maximize the numbers along the main diagonal) using column permutation and we evaluated the median expression of genes within each specific subgroup. Considering the ICC cohort, the column permutation showed specific concordance between the subgroups identified by the two methods, as strongly demonstrated by the association between VS_1 - C and VS_2 – B classes. Interestingly, the 19 genes considered for the classification presented peculiar patterns of up-regulation (e.g., VTN, ADH1C, ALDOB, FABP1 in VS_2 - B) and down-regulation (e.g., TCN1, OLFM4, VTN, ADH1C in VS_3 - D) demonstrating the presence of biological differences between the subgroups (Figure 9A). Within ECC cohort, the column permutation highlighted specific association between the subgroups from predictor-based and hierarchical clustering-based approaches except for groups B and VS_3, where the low number of patients in validation set affects the degree of concordance. Similar to ICC, the comparison of median expression showed patterns of up-regulation and down-regulation for the 21 genes, also for the B-VS_3 groups (e.g., GREM1, PPARGC1A, GHR and CHGB) (Figure 9B).

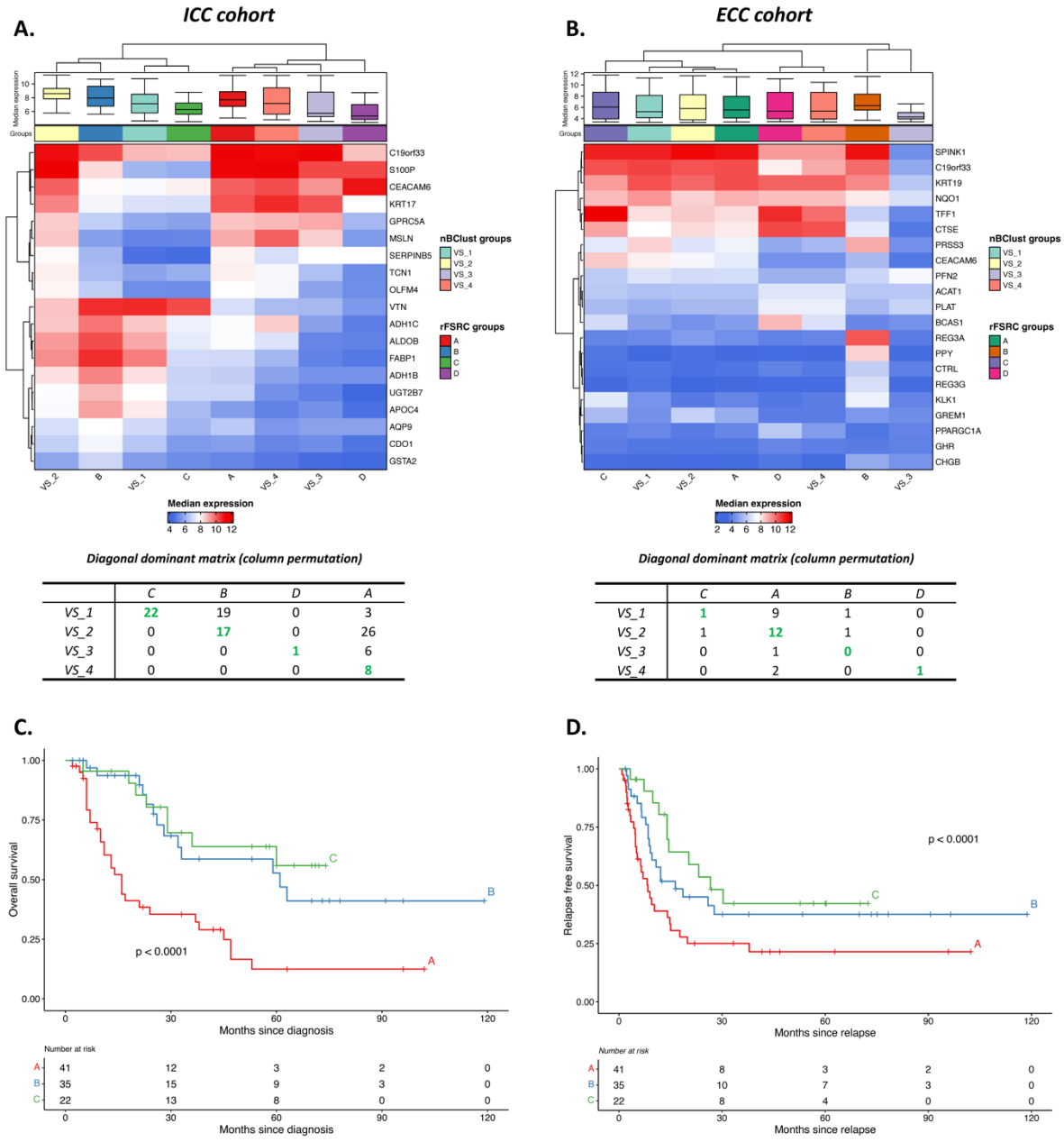
Table 2. Clinical characteristics of patients in validation set.

Characteristic	ICC N (%)	ECC N (%)
Age		
< 67 years	47 (46.1%)	9 (69%)
≥ 67 years	52 (51%)	20 (31%)
Missing	3 (2.9%)	/
Clinical Tumor Size		
cT1	4 (39.2%)	7 (24.1%)
cT2	47 (46.1%)	10 (34.5%)
cT3	17 (16.6%)	10 (34.5%)
cT4	31 (30.4%)	2 (6.9%)
Missing	3 (2.9%)	/
Clinical Nodal Status		
cN0	67 (65.7%)	22 (75.9%)
cN1	32 (31.4%)	7 0 (24.1%)
Missing	3 (2.9%)	/
Stage		
1	4 (39.2%)	10 (34.5%)
2	34 (33.3%)	15 (51.7%)
3	12 (11.8%)	2 (6.9%)
4	49 (48%)	2 (6.9%)
Missing	3 (2.9%)	/
Relapse		
Yes	61 (59.8%)	10 (34.5%)
No	38 (37.3%)	19 (65.5%)
Missing	3 (2.9%)	/

4.1.1.5. Identified ICC biological subgroups are associated with clinical outcome

In order to understand the clinical relevance of the expression subgroups, OS and RFS were evaluated in the ICC validation cohort. Considering OS, subgroup A showed a significantly worse prognosis compared to subgroups B and C (p-value < 0.0001, Figure 9C). Similarly, the analysis of RFS demonstrated that subgroup A was characterized by patients with the shorter RFS (Figure 9D). These results support a link between biological characteristics of the primary tumors and the prognosis. Indeed, patients within subgroup A had the worst prognosis and RFS with tumors characterized by KRAS up-regulation, epithelial-to mesenchymal-transition and apoptosis and by the absence of immune cell infiltration. On the contrary, despite the presence of several cancer associated up-regulated pathways such as RTK-RAS-PIK3K and TGF-ß, tumors belonging to subgroups B and C were characterized by a strong immunological component, improving patients' prognosis both in

terms of OS and RFS (Figure 9C-D). Unfortunately, the low number of patients prevented a similar survival analysis for ECC subgroups.



	C	B	D	A
VS_1	22	19	0	3
VS_2	0	17	0	26
VS_3	0	0	1	6
VS_4	0	0	0	8

	C	A	B	D
VS_1	1	9	1	0
VS_2	1	12	1	0
VS_3	0	1	0	0
VS_4	0	2	0	1

Figure 9. Comparison of predictor-based and NbClust-based classification in ICC and ECC validation set and evaluation of the clinical outcome. To evaluate the validity of classification returned by the ICC and ECC predictors, a comparison with subgroups returned by NbClust was performed using diagonal dominant matrix approach and comparing the median expression values associated to each group. For both ICC (A) and ECC (B) cohort, heatmap shows subgroups on the columns and genes on the rows. The subgroup membership and the associated gene expression levels are reported at the top of the heatmap as color bar and box plot, respectively. Diagonal dominant matrix results are shown at the bottom of the heatmap. Overall survival and relapse free survival analysis in ICC cohort are represented using Kaplan-Meyer method. Due to the low number of patients, ECC cohort was not considered for survival analysis.

4.1.2. Mutational data

4.1.2.1. Integration of data from different platforms for the establishment of discovery set

A total of 361 ICC and 49 ECC patients with available mutational data with no previous history of hepatitis or fluke infection were collected from four different datasets and considered as discovery dataset. Overall, the intersection of the different platforms used to characterize each sample along with the application of all the filtering steps to remove artifacts and not interested variants (see section 3.2.2 of materials and methods) allowed to select 298 and 77 mutated genes in ICC and ECC cohort, respectively. In particular, the most frequent variant types identified in both of the cohorts were Missense mutations (61% in ICC; 55% in ECC) followed by INDEL (26% in ICC; 30% in ECC), with a prevalence of C > T base change and a median number of muatated genes per samples of 3 and 2 for ICC and ECC, respectively.

4.1.2.1. Unsupervised clustering analysis identifies distinct biological subgroups in ICC and ECC cohorts

Unsupervised clustering analysis using NbClust led to the identification of 3 and 2 subgroups within ICC (N = 361) and ECC (N = 49) cohort, respectively, with no biased distribution of samples derived from the different collected datasets (Figure 10A-B).

To investigate the presence of distinct biological traits of the newly defined subgroups, we performed enrichment analysis interrogating canonical oncogenic signaling pathways defined by Vega and colleagues [56] and The drug gene interaction database [58].

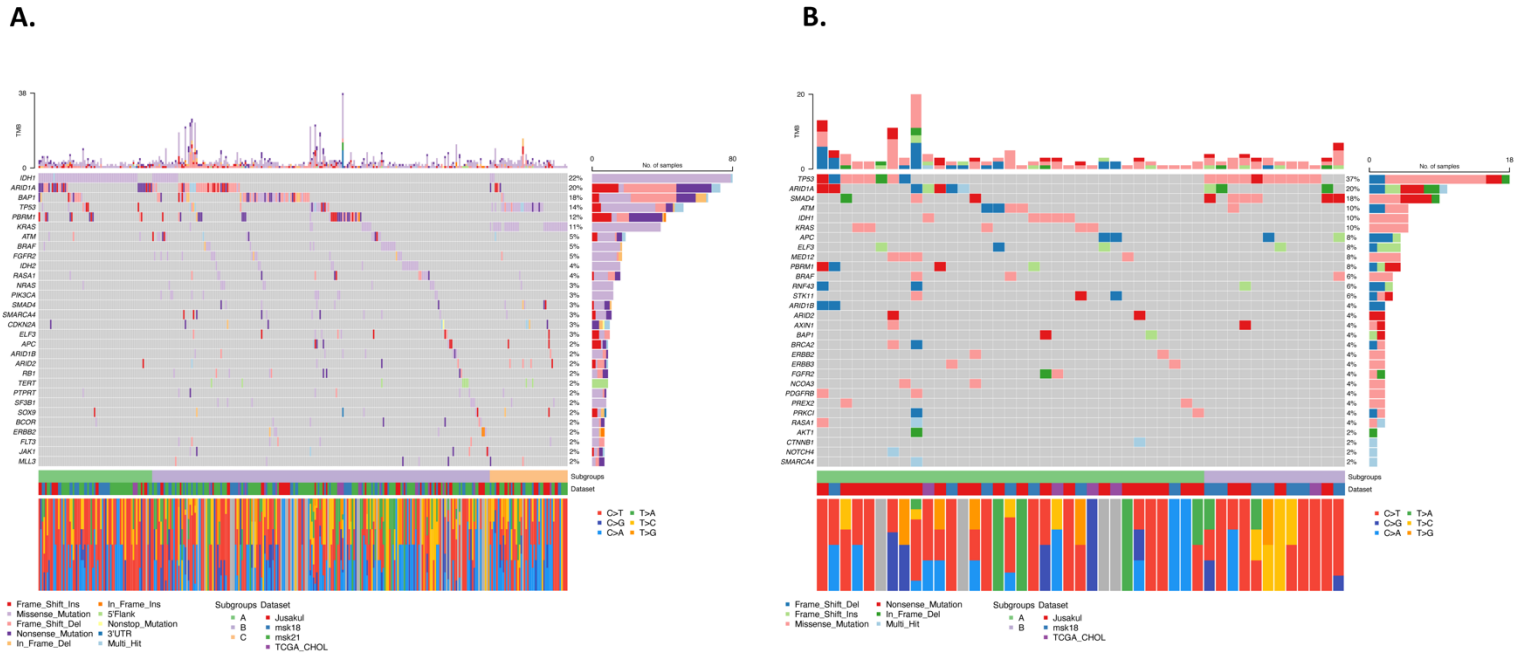
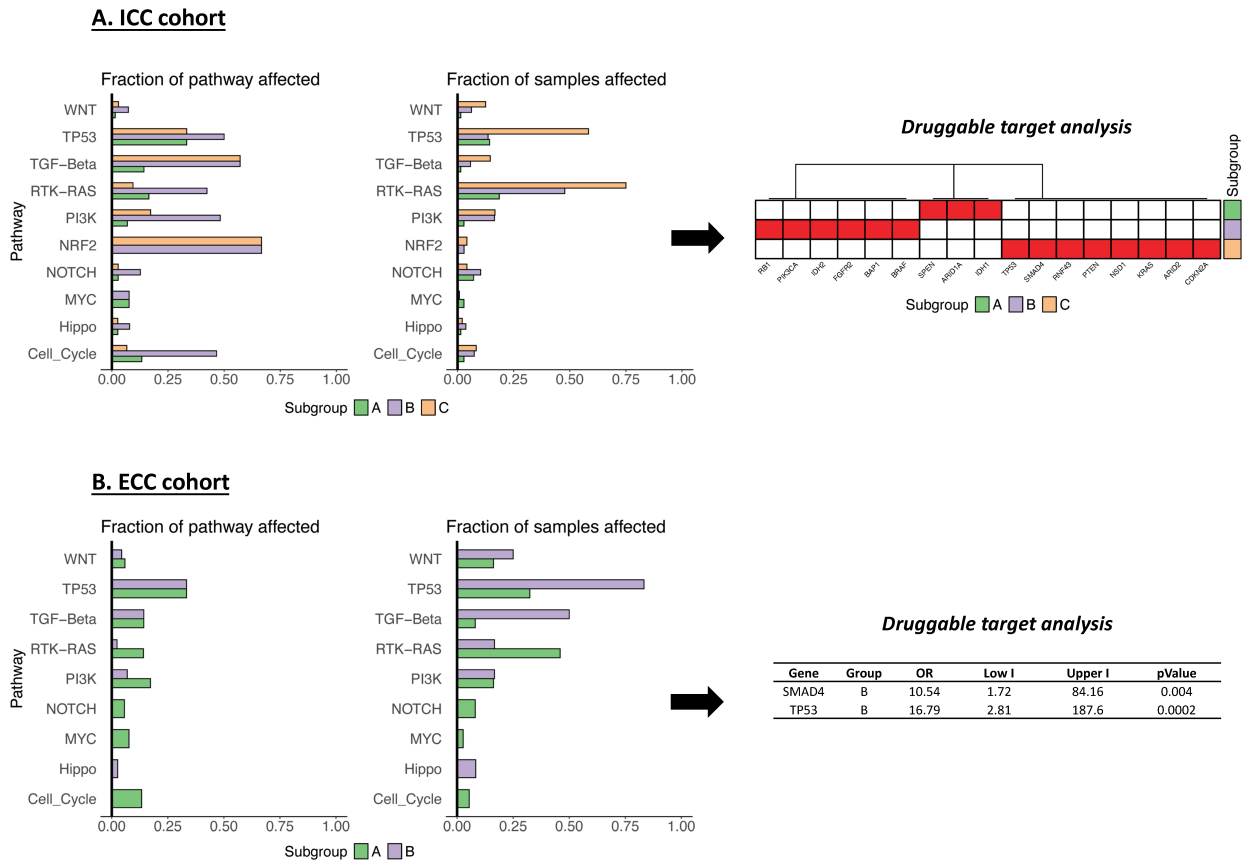


Figure 10. Identification of ICC and ECC distinct subgroups considering mutational data. Unsupervised clustering analysis using NbClust identified 3 subgroups in ICC (A) and 2 in ECC (B) discovery set. The heatmaps reports samples on the column and the top 30 mutated genes on the rows. For each sample, dataset of origin and the subgroup membership are reported as color bars on the bottom of the heatmap.

Regarding the ICC cohort, oncogenic pathways characterized distinct biological features of each subgroup. The subgroup A showed a low fraction of affected pathways compared to B and C that they were mainly characterized by and overrepresentation of TP53, RTK-RAS, TGF-Beta, PI3K and NRF2 signaling. Interestingly, the evaluation of the clinically actionable genes through the application of two-test statistic (see section 3.4 of materials and methods) revealed the presence of druggable targets (n=17) specifically associated to each subgroup, like as SPEN-ARID1A (A), TP53-SMAD4 (B) and RB1-PIK3CA (C) (Figure 7A). The pathways analysis on samples of the ECC cohort also showed distinct biological traits associated to each subgroup, related to Cell cycle, NOTCH and MYC target (subgroup A), TP53 and TGF-Beta (subgroup B). Moreover, only subgroup B is characterized by the presence of clinically actionable genes, corresponding to TP53 and SMAD4 (Figure 11B)



Gene	Group	OR	Low I	Upper I	pValue
SMAD4	B	10.54	1.72	84.16	0.004
TP53	B	16.79	2.81	187.6	0.0002

Figure 11. Biological characterization of ICC and ECC subgroups considering mutational data. Canonical oncogenic signaling pathways and The drug gene interaction database were interrogated to detect the presence of specific mutated genes in ICC (A) and ECC (B) discovery set. The barplots on the left report the fraction of affected pathways and samples within each cohort while the heatmap/table on the right showed the druggable genes associated to each subgroup resulting from the two-test statistic approach (see section 3.4 of materials and methods). Red and white color within the heatmap refers to the presence/absence of mutated genes in the subgroups.

4.1.2.2. Supervised classifier to identify the proposed subgroups using mutations

Similarly to gene expression data, the presence of distinct ICC and ECC biological subgroups within the discovery set allowed to build and compare specific predictors to validate our findings in validation set. For such a purpose, specific clinically actionable genes were extracted for ICC cohort while, for ECC, only common mutated genes between discovery and validation set were considered. A total of 13 and 33 genes were selected for ICC and ECC samples, respectively and used to build the predictor from training set. Due to imbalance between subgroups, we adopted the SMOTE algorithm in order to oversample the minority class prior to classifier definition on the discovery set. Among the 3 machine learning algorithms tested, RF showed the best performance with an overall misclassification rate of 0.05 and 0.08 for ICC and ECC, respectively. Moreover, RF predictor

showed an AUC value of 0.99 (KNN=0.95; SVM=0.94) and of 0.97 (KNN=0.72; SVM=0.35) for ICC and ECC cohort, respectively (Figure 12).

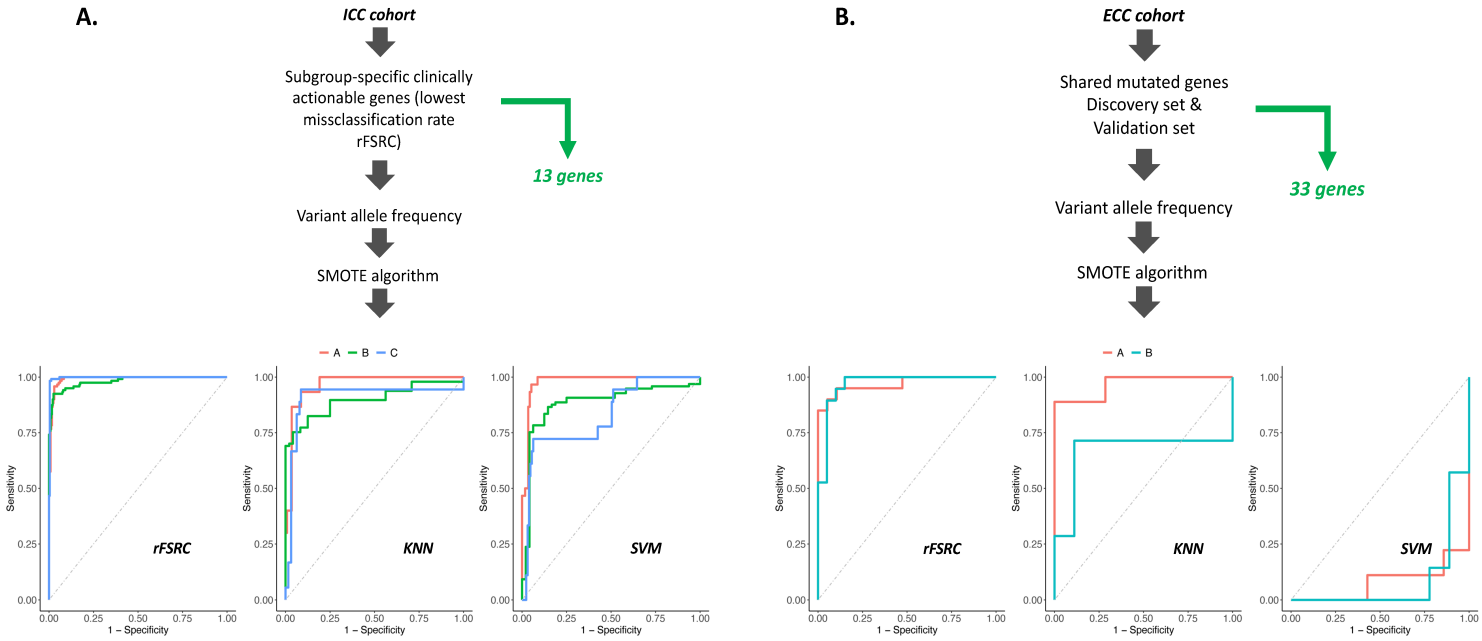


Figure 12. Building of predictors related to the distinct biological subgroups in ICC and ECC cohort. For ICC, subgroup-specific clinically actionable genes (n=13) were considered (A). Within ECC, only shared mutated genes between discovery and validation set were selected (B). For both ICC and ECC cohort, SMOTE algorithm was applied to balance the distribution of the samples within subgroups and three different machine learning methods were evaluated for the establishment of the predictor. The ROC curves report specificity and sensitivity values obtained from RF, KNN and SVM algorithms.

4.1.2.3. Identification of ICC and ECC biological subgroups in validation set

To validate and establish the clinical significance of our predictors, the same independent dataset composed of 131 patients (102 ICC, 29 ECC) used for gene expression data was considered as ICC and ECC validation sets (Table 2). After the application of the gene variants filtering adopted for discovery set, we used the Random Forest-based classifier to predict the presence of the specific biological subgroups in both ICC (n=3) and ECC (n=2) of the validation set. In the ICC cohort, the subgroups A (n=13) and B (n=58) were characterized by the lowest and highest number of patients, respectively, consistently with the discovery set composition. For the ECC cohort, the low number of patients affected the prediction with most of the patients identified in subgroups A (n=22) and only 7 assigned to subgroup B.

To validate the subgroups predicted by the supervised classifier, we compared them to the subgroups discovered by unsupervised clustering analysis on the validation sets, which were independent from the discovery sets. The method was applied on shared mutated genes between the discovery and validation set. In order to match the group labels provided by the unsupervised clustering to the predicted subgroup classes, we tabulated the unsupervised labels versus the predictor-based groups is a 3x3 (ICC) or 2x2 (ECC) matrix and searched for the optimal matching based on the dominant diagonal matrix (the matrix that maximize the numbers along the main diagonal) using column permutation. Considering the ICC cohort, the column permutation showed specific concordance between the subgroups identified by the two methods except for the match VS2-A. For these subgroups, the absence of the concordance between the two classification methods could be related to the different rate of mutations between discovery and testing set, reflecting the discrepancy between the two cohort (Table 3).

Within ECC patients, the column permutation highlighted specific association between the subgroups from predictor-based and NbClust-based approaches (Table 3).

Table 3. Diagonal dominant matrix (column permutation) results in ICC and ECC cohort.

ICC cohort	**A**	**B**	**C**
VS_1	0	13	0
VS_2	12	46	0
VS_3	0	9	19
ECC cohort	**A**	**B**	
VS_1	9	0	/
VS_2	13	7	/

4.1.2.4. Identified ICC biological subgroups are associated with clinical outcome

In order to understand the clinical relevance of the expression subgroups, OS and RFS were evaluated in the ICC validation cohort. Considering OS, subgroup C showed a significantly worse prognosis compared to subgroups A and B (p-value < 0.0001, Figure 13A). Similarly, the analysis of RFS demonstrated that subgroup C was characterized by patients with the shorter RFS (Figure 9B). These results support a link between biological characteristics of the primary tumors and the prognosis. Indeed, patients within subgroup C had the worst prognosis and RFS with tumors characterized by mutations affected TP53 and KRAS signaling pathways and by an higher number of mutated druggable targets (n=8) compared to the other subgroups. On the contrary, patients belonging to subgroups A were characterized by the lowest number of mutations affected canonical oncogenic pathways and by a small number of altered clinically actionable genes (n=3), improving patients' prognosis both in terms of OS and RFS (Figure 13A-B). Unfortunately, the low number of patients prevented a similar survival analysis for ECC subgroups.

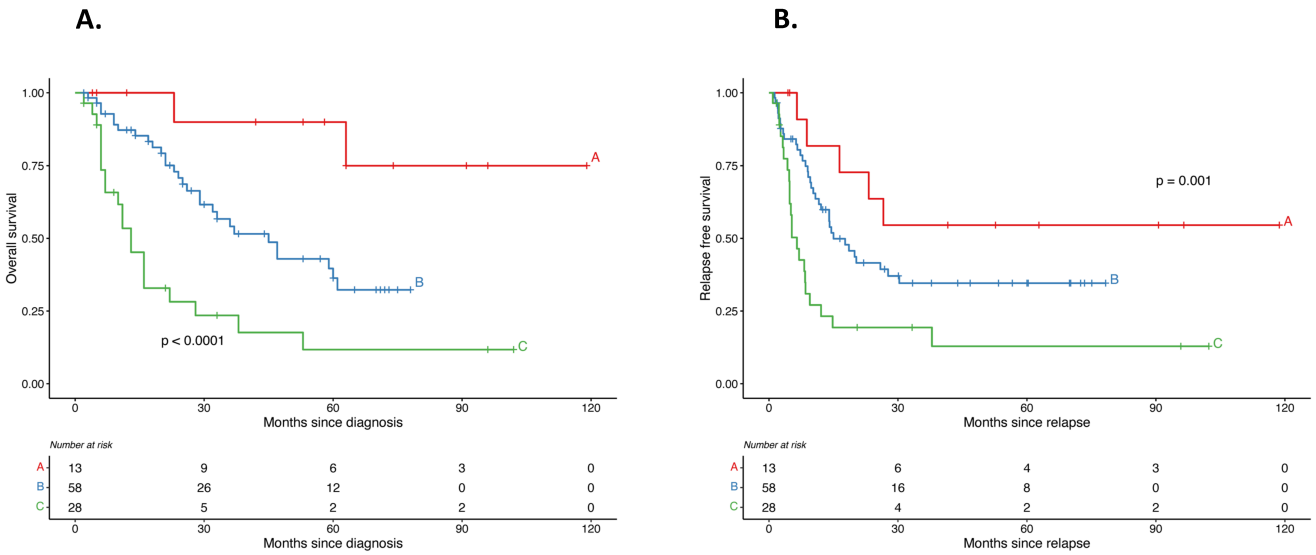


Figure 13. Evaluation of clinical outcome in ICC validation set. Overall survival (A) and relapse free survival (B) analysis in ICC cohort are represented using Kaplan-Meyer method. Due to the low number of patients, ECC cohort was not considered for survival analysis.

4.1.3. Copy number data

4.1.3.1. Integration of data from different platforms for the establishment of discovery set

A total of 298 ICC and 30 ECC patients with available copy number alteration with no previous history of hepatitis or fluke infection were collected from four different datasets and considered as discovery dataset. Overall, the intersection of the different platforms used to characterize each sample (see section 3.2.2 of materials and methods) allowed to select 319 and 81 altered genes in ICC and ECC cohort, respectively. In contrast to gene expression and mutational data, no filtering step on detected CNA were not necessary. The most frequent CNA types identified in both of the cohorts were Gain (63% in ICC; 77% in ECC) with only a low number of loss detected (37% in ICC; 23% in ECC). Moreover, the top altered genes were represented by CDKN2A and CDKN2B in both of the cohort (loss), while the most altered chromosomal arms represented by 1q, 11q, 9p and 1q, 1p, 9p for ICC and ECC, respectively.

4.1.3.2. Unsupervised clustering analysis identifies distinct biological subgroups in ICC and ECC cohorts

Unsupervised clustering analysis using NbClust led to the identification of 2 subgroups in both ICC (N = 361) and ECC (N = 49) cohort. Due to the low number of patients, a biased distribution of tumors derived from the different collected datasets were detected in ECC cohort (Figure 14A-B). Similarly to mutational data, we investigated the presence of distinct biological traits of the newly defined subgroups through enrichment analysis interrogating canonical oncogenic signaling pathways defined by Vega and colleagues [56] and The drug gene interaction database [58].

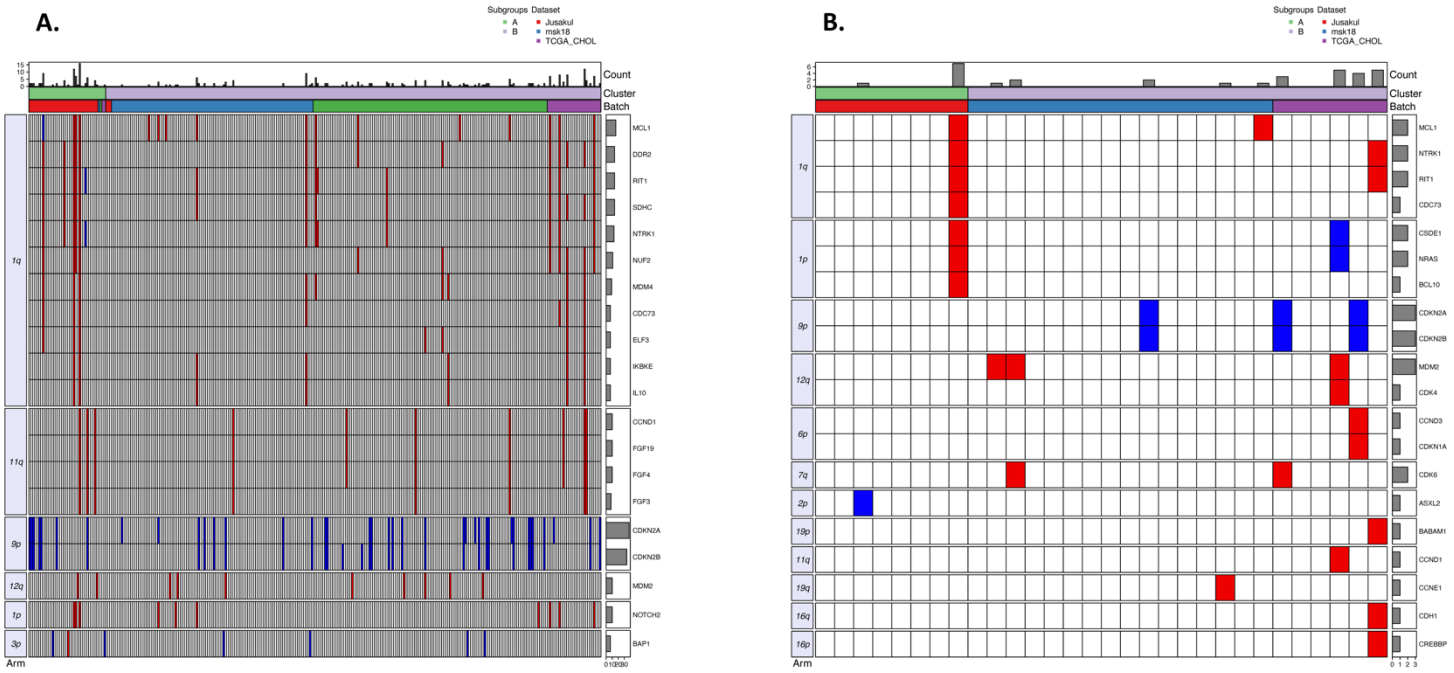


Figure 14. Identification of ICC and ECC distinct subgroups considering CNA data. Unsupervised clustering analysis using NbClust identified 2 subgroups in ICC (A) and ECC (B) discovery set. The heatmaps reports samples on the column and the top 20 altered genes on the rows. For each sample, dataset of origin and the subgroup membership are reported as color bars at the top of the heatmap.

Regarding the ICC cohort, oncogenic pathways characterized distinct biological features of each subgroup. The subgroup A is mainly characterized by CNA TGF-Beta (Gain), cell cycle (Gain), TP53 (Loss) and NRF2 (Loss) signaling while subgroup B showed alterations involved in NRF2 (Gain), MYC (Gain) and TGF-Beta (Loss). Notably, the evaluation of the clinically actionable genes through the application of two-test statistic (see section 3.4 of materials and methods) revealed the presence of druggable targets (n=42) associated to subgroup A only (Figure 11A). The pathways analysis on samples of the ECC cohort also showed distinct biological traits associated to subgroup B mostly, as expected by the distribution of the samples among the subgroups. Moreover, no subgroup-specific druggable targets were detected (Figure 15B).

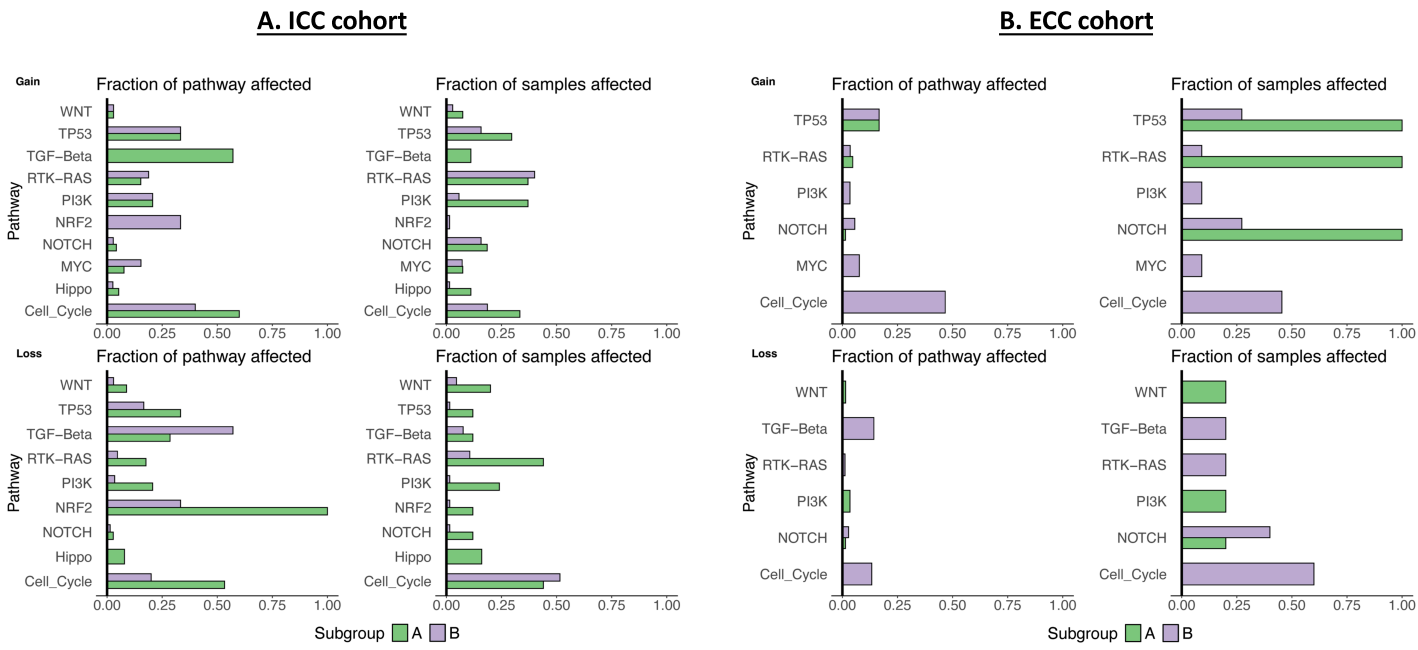


Figure 15. Biological characterization of ICC and ECC subgroups considering CNA data. Canonical oncogenic signaling pathways and The drug gene interaction database were interrogated to detect the presence of specific altered genes in ICC (A) and ECC (B) discovery set. The barplots are divided by alterations type and report the fraction of affected pathways and samples within each cohort. Druggable targets were evaluated using the same approach adopted for mutational data (see section 3.4 of materials and methods).

4.1.3.1. Supervised classifier and identification of ICC and ECC biological subgroups in validation set

Based on the same methods used for gene expression and mutational data, the steps related to the building of supervised classifier and to the identification of ICC and ECC biological subgroup in validation set is currently ongoing, as a conclusion of step1. Finally, evaluation of clinical outcome associated to each identified CC subgroups will be performed.

4.2. AIM2: Integration and validation of different molecular signatures to define biological distinct subgroups and create predictive tool

The extraction of different molecular signatures along with biological characterization performed in aim1 will imply the prediction of multiple data type (e.g., CNA, somatic mutations and gene expression) on validation set, in order to obtain a complete characterization of ICC and ECC cohort. Indeed, due to the presence of matched transcriptomics and genomics data for each patient, an approach based on consensus clustering will redefine the biological subgroups identified in discovery set allowing a comprehensive characterization of CC tumors. Then, a predictive tool based on CNA data will be created to allow inference on CTCs derived from CC patients profiled by lpWGS (see section 3.9.2 of Materials and Methods). The present aim2 is currently ongoing (Figure 16).

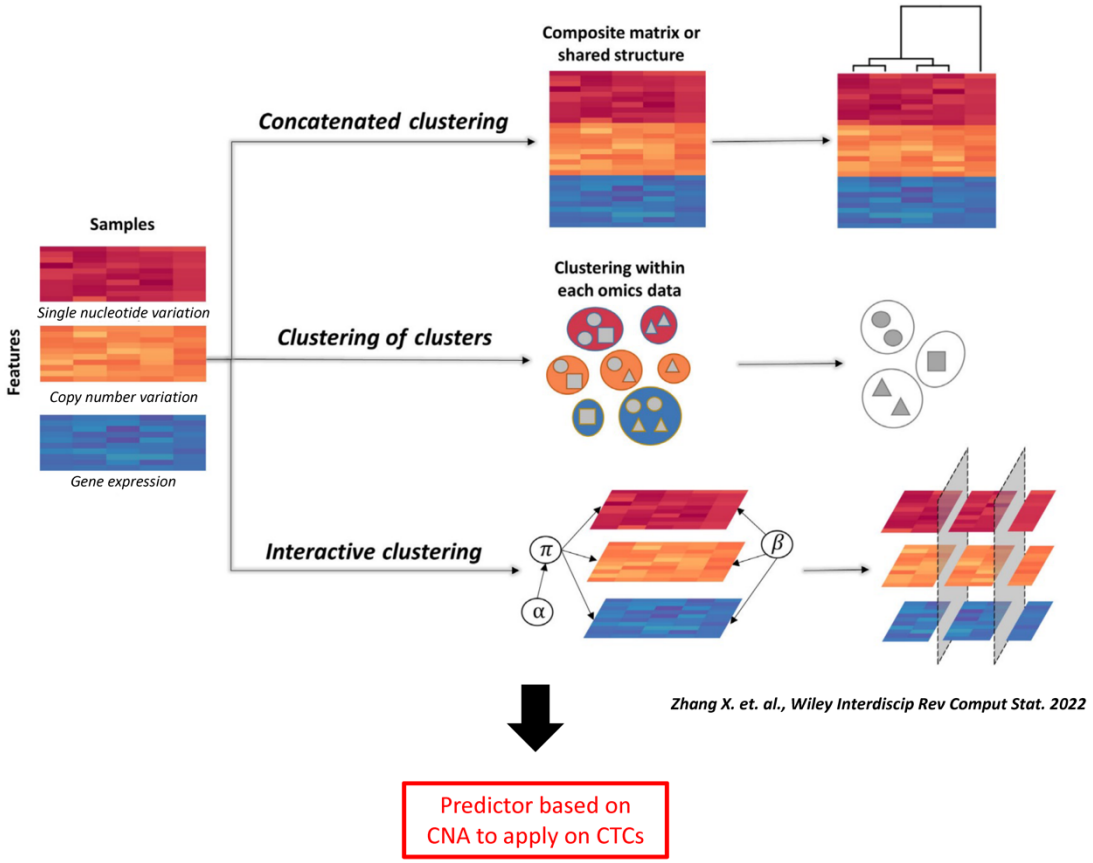


Figure 16. Framework of aim2. The pipeline reports all the steps that will be performed in aim2 starting from different molecular signatures extracted from aim1. [Adapted from Zhang and colleagues in 2022]

4.3. AIM3: Collection and molecular characterization of CTCs from CC patients

4.3.1. Genomic characterization of CTCs depicted high level of heterogeneity intra and extra patient

A total of 88 CTCs were collected from 28 CC patients enrolled at Fondazione IRCCS Istituto Nazionale dei Tumori di Milano and profiled using lpWGS. Through this technique, it is possible to obtain CNA profiling of single cells at a lower cost than microarrays and high-throughput WGS, thus allowing the analysis of a higher number of samples. Compared with deep sequencing strategies, lpWGS produces only a fraction of the data per sample and relies on computational methods to fill in the missing information. So, while it does not provide information on small-scale alterations (INDELs), it instead gives an overview of the macro-aberrations present in the entire genome of a cell, thus indicating whether a cell is normal (flat, diploid profile) or cancerous (aberrant profile).

Globally, 63 (72%) and 25 (28%) CTCs were obtained from ICC and ECC patients, respectively, at Baseline (BL; n. ICC=30, n. ECC=10), during treatment (DT; n. ICC=12, n. ECC=2), end of treatment (EOT; n. ICC=11, n. ECC=11), post end of treatment (postEOT; n. ICC=0, n. ECC=2) and at follow-up (FU; n. ICC=10, n. ECC=0).

Considering the number of CNAs detected, the most altered chromosomal regions were 1p (6.1%), 1q (5.8%) and 2q (4.7%) in ICC while, within ECC cohort, 11q (5.7%), 12q (5%) and 15q (4.8%) were the most affected arms. The difference between the two cohorts was confirmed also at gene levels, ARID1B and BMP6 in ICC and MGMT and EBF3 in ECC patients harbored the higher number of alterations.

Overall, the CNA profiles associated to each CTC showed a high level of heterogeneity intra and extra patient in both ICC and ECC. Notably, this characteristic was evident not only considering all the collected CTCs within each patient along follow-up, but also focusing on single timing, suggesting the presence of different selected tumor clones over the course of the disease (Figure 17).

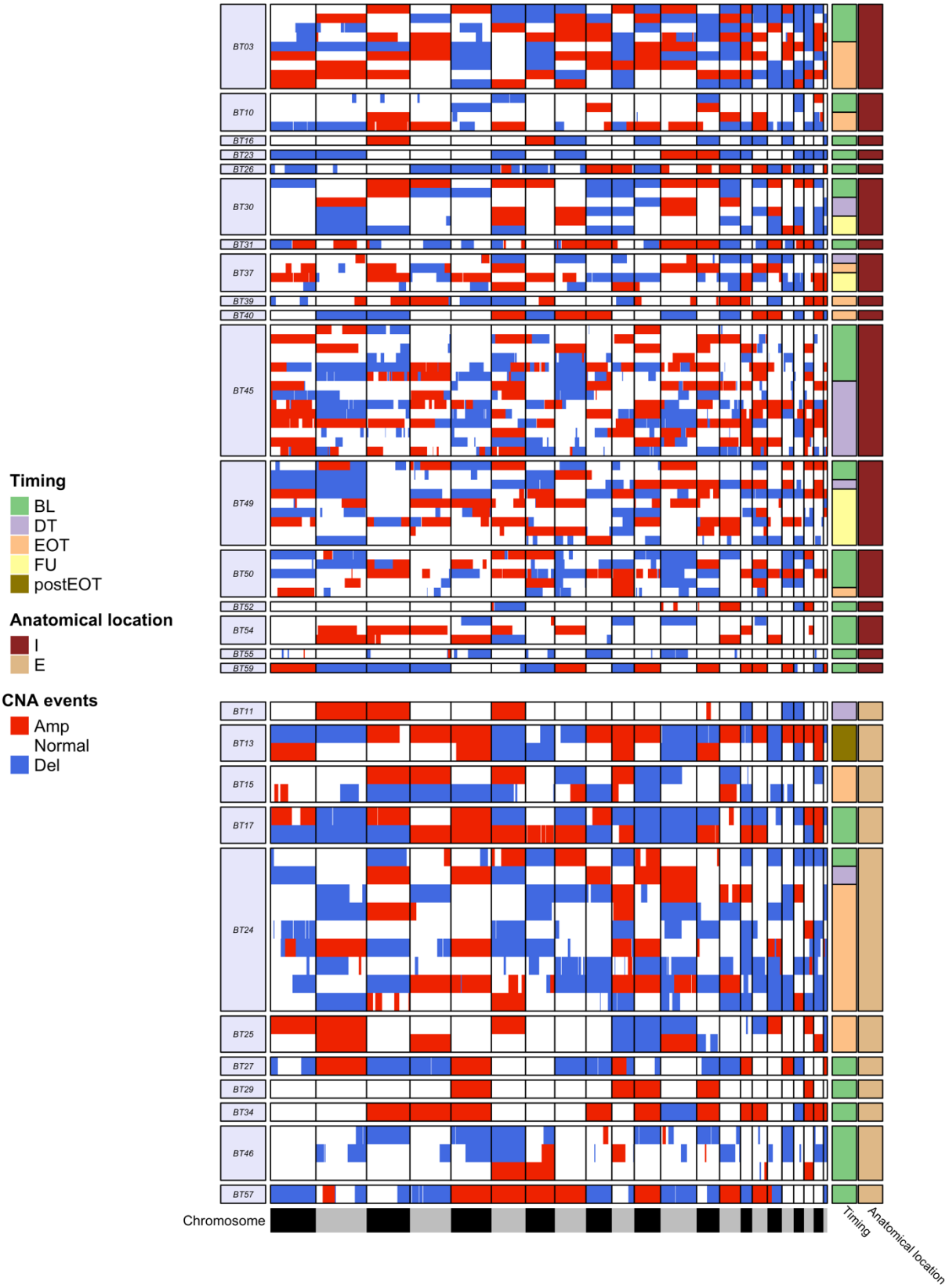


Figure 17. CNAs profiles of CTCs collected in ICC and ECC patients. The heatmaps show the genomic profiles in terms of CNAs of CTCs collected from patients enrolled in Fondazione IRCCS Istituto Nazionale dei Tumori di Milano. Columns report genomic regions (chr1 to chr22) from left to right while collected CTCs, divided for each patient, were reported on the rows. Annotations on the right report the timing of withdrawal of each CTC and the type of CC cohort (ICC, ECC). Red and blue colors refer to amplification and deletion.

4.4. AIM4: Monitoring of cancer evolution under treatment pressure through phylogenetic analysis of CTCs.

4.4.1. Phylogenetic analysis of CTCs allows to evaluate tumor evolution within single patient.

The amount of information retrieved analyzing transcriptomic and genomic data in aim1, allowed to obtain a comprehensive characterization of CC tumor and specifically associated to ICC and ECC subtype. Since aim2, that will be based on the integration and validation of different molecular signatures to define biological distinct subgroups and create predictive tool to apply on CTCs collected in aim3, is still ongoing, we used already published gene signatures to perform phylogenetic analysis within single patients [29]. In particular, Reduzzi and colleagues identified genes (n=10) associated to genomic region 3p11.1 as markers able to distinguish CTCs of responding from non-responding patients to the treatment line. Leveraging on this gene signature, we mapped the CTCs evolutionary relationship using the phylogenetic analysis performed by the TRONCO pipeline [68] Patient BT24, the one considered as first test in our cohort, had 9 CTCs collected at baseline (n=1), during (n=1) and after treatment (n=1) of cis/gem therapy and, after a change of treatment due to progression, at the end of immunotherapy (n=6). Based on the processes regulated the carcinogenesis, WBC was considered as hypothetical root of the phylogeny. During the first line of therapy (cis/gem), clones A (n=1), B (n=1) and C (n=1) characterized by different CNAs features were identified at baseline, during and after treatment highlighting the modulation performed by therapy. Despite the response to cis/gem at the end of treatment, patient went into progression and, after immunotherapy, 6 CTCs were collected. Interestingly, they belong to the same three clones A (n=2), B (n=1) and C (n=3) detected before the progression, suggesting their role as tumor-driver event in the evolution of the disease (Figure 18).

Overall, these results revealed that CTCs are able to describe the complex biology of the tumor within single patients, showing how the level of heterogeneity along time is shaped by resistant and non-resistant clones arising under therapy pressure.

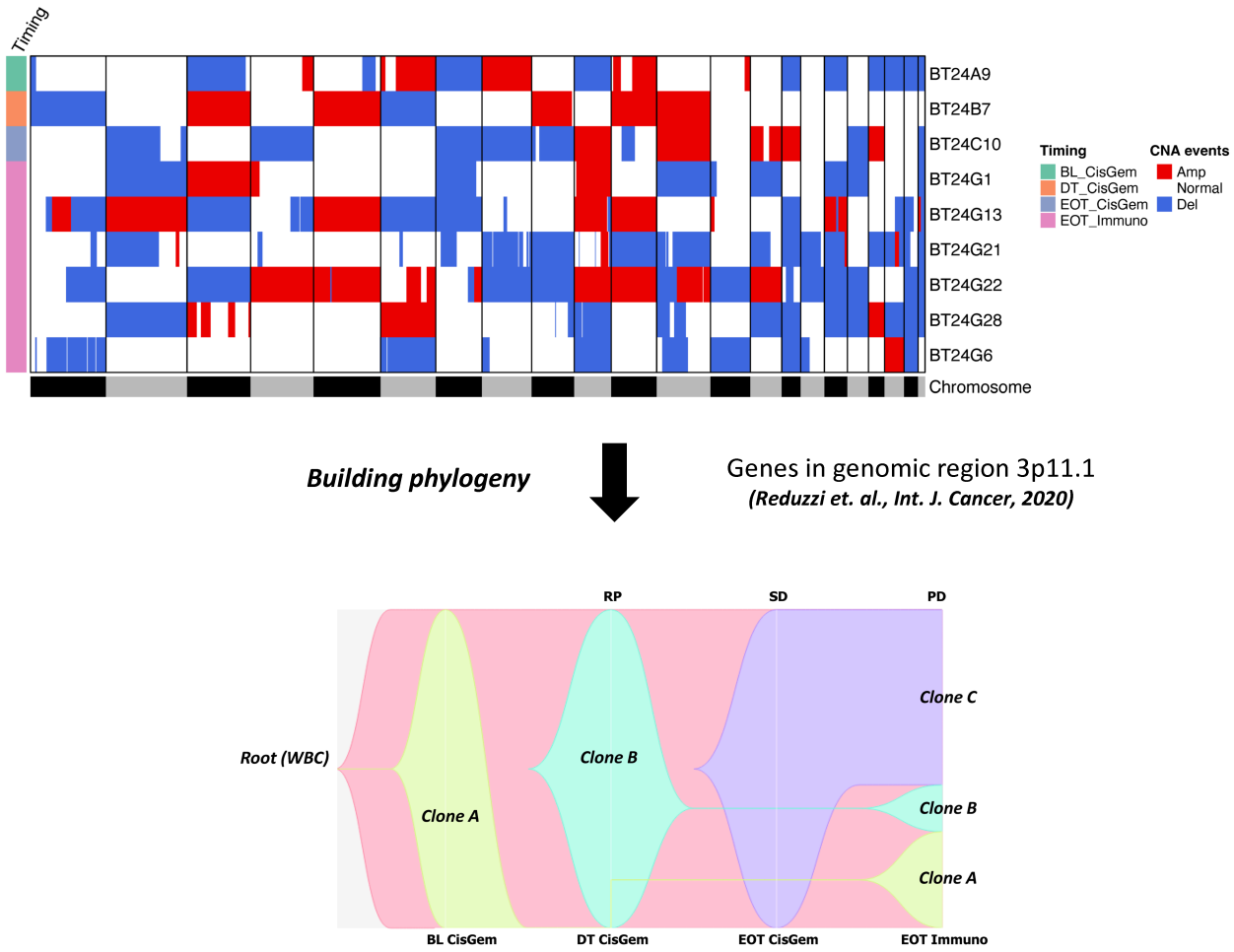


Figure 18. Phylogenetic analysis of Patient BT24. Phylogenetic analysis was performed on CTCs collected at baseline, during and after treatment in patient BT24. Heatmap on the top shows CNAs profile of CTCs: genomic region, from left to right, are reported on the column (chr1 to chr22) while single CTCs are reported on the rows. Red and blue colors refer to amplification and deletion, respectively. On the bottom, Timescape plot reports on Y axis the prevalence of each clone (defined by different colors) detected along the time in patient BT24. Therapy line (bottom) and disease status (RP = Partial response; SD = Stable disease; PD = Progression disease; top) are reported on X axis, respectively.

5.DISCUSSION

CC is a deadly disease with limited therapeutic options. Although it is possible to perform surgery with curative intent in the early-stage disease, most patients are diagnosed with advanced CC, when the only option is systemic chemotherapy. The standard first-line treatment for advanced CC is cis/gem, which has limited efficacy, and there is no standard second-line treatment.

Molecular profiling has not only revealed that CC carry multiple potentially actionable mutations, but has also shown a great heterogeneity between different anatomical subtypes (ICC, ECC). These results support the possibility of personalized therapeutic approaches in CC patients, but also highlight the need of performing the molecular analysis of the tumors, which is not feasible for many patients due to tumor tissue inaccessibility. Moreover, the observed intra-tumor heterogeneity can represent another obstacle in the implementation of personalized medicine in CC, since a single tissue biopsy may not recapitulate the overall tumor heterogeneity.

In the present study, we report a pipeline to infer cancer evolution within single patients through the use of circulating tumor cells evaluated on transcriptomics and genomics signatures extracted from the collection of publicly available CC datasets. Beside to the definition of distinct biological subgroups, when possible, these signatures were associated to a prognostic value.

Considering gene expression data, the identified subgroups are characterized by the enrichment of specific molecular pathways such as EMT, IL6, RTK-RAS-PIK3K and DNA repair in ICC tumors and EGFR, VEGF signaling immune and integrine-pathways in ECC. Immune cell infiltrates are peculiar and distinct among subgroups supporting their unique biological profiles and possible treatment implications. The analysis of genomic data, mutations and CNAs, revealed a distinct samples distribution within discovery set compared to transcriptomic landscape, showing the presence of peculiar oncogenic pathways like as TP53, TGF-Beta, NFR2 and NOTCH. Moreover, druggable targets associated to specific subgroups mutations were identified, demonstrating the potential role of these classification in clinical management.

Finally, for both mutational and gene expression data the analysis of the overall and relapse-free survival highlights a significant association between biological subgroups and clinical outcome in the

ICC cohort, identifying one specific subclass characterized by poor prognosis. The clinical evaluation of CNAs subgroups is currently ongoing.

Several studies aimed at characterizing cholangiocarcinoma using integrative molecular analyses [25], [27], [28], [70]. However, a comprehensive analysis of ICC and ECC derived from different patient cohorts for obtaining biological insights and prognostic information was still lacking. Herein, at variance with previous studies, we have collected and combined for the first time all public transcriptomic and genomic datasets available at date in a unique cohort split into discovery and validation sets.

When collecting large numbers of samples derived from different studies the careful integration of the gene expression, mutation and CNAs datasets derived from multiple high throughput platforms is mandatory. The choice of the optimal method for integration of these data represents a critical step impacting on the final relevance of findings. Herein, we did not only face the need of integrating heterogeneous sample sources (FF, FFPE), but also data derived from different platforms ranging from microarrays to RNAseq (gene expression), and from Target/Whole/Exome-seq to SNP6 array (genomic). The use of cross-platform analysis was a crucial step that allowed improving the classification and identification of tumor phenotypes when considering not only different type of microarrays [71] and RNAseq technology [72]–[74] but also several genomic platforms [75], [76].

The main issue for the building of ICC and ECC discovery sets was related to gene expression data derived by distinct datasets, that was made possible by resorting to empirical Bayesan methodology (ComBat) to remove batch effects associated with each dataset, a mandatory step as already reported [77].

In a data-driven approach, such as ours, the choice and careful evaluation of clustering methods is critical. In the current study, the application of NbClust was instrumental for the robust definition of the optimal number of clusters within ICC and ECC of the discovery sets.

In fact, the robustness of the identified subgroups was confirmed by statistical considerations, but was also fully supported by the biologically distinct traits of each identified cluster. For digging

deeply into the biology of subgroups, the gene set enrichment, single sample scoring and oncogenic pathways analyses were complemented by the evaluation of immune cell infiltration and druggable target in gene expression and genomic data, respectively. The biological singularity of each cluster identified by unsupervised classification was indeed further supported by the different profiles of immune infiltrate or the presence of different clinically actionable genes associated with each subgroup.

Beside to the identification of druggable targets through the analysis of genomic data, interesting results in terms of clinically application was obtained to the evaluation of gene expression profiles. In particular, our results showed not only the presence of distinct canonical CC gene signaling such as TGF-b, RTK-RAS-PIK3K [54], cell cycle and the up-regulation/down-regulation of biological pathways related to metabolic processes but also different immune cell populations within subgroups identifying in both ICC and ECC a subgroup (subgroup A) of tumors characterized by an immune exhausted tumor microenvironment. In ICC, subgroup A displayed a gene expression profile suggesting epithelial-mesenchymal transition with modifications of cell-cell interaction, adhesion and motility accompanied by KRAS up-regulation, whereas in ECC immune exhaustion was associated with proliferation and EGFR signaling. For ECC tumors it is interesting to note the strong immune infiltration detected in subgroup C, which was characterized by the presence of 9 different immune cell populations. Indeed, in this latter group GO identified up-regulation of immune pathways, but also pathways linked to chemoresistance.

Having identified interesting subgroups of ICC and ECC tumors endeavored with possible treatment implications we set to find a way to apply those results to the real- world daily practice.

The first step was the building and the validation of a classifier with a reduced number of genes. To such a purpose, short transcriptomic and genomic gene signatures were extracted from differential expression analysis between subgroups taking advantage from the SMOTE algorithm for adjusting sample number unbalances. These steps were necessary to obtain a more accurate establishment of the ICC and ECC predictors after the test of three different machine learning algorithms. RF method

showed the best performance in both of the cohort considering all the data types, given the low error rate in classification.

As first step in the validation process, we compared the predictor-based and the NbClust-based classifications by using diagonal dominant matrix and, only for gene expression data, evaluating the median gene expression levels. Considering the transcriptomic profiles, the two classification methods correlate well despite the hugely different number of genes considered, demonstrating the presence of the four biological subgroups in both cohorts. For mutational data, the diagonal dominant matrix approach showed concordance between NbClust-based and predictor-based methods except for 1 association (VS2-A). For these subgroups, the absence of correlation could be related to the different rate of mutations between discovery and testing set, reflecting the biological discrepancy between the two cohort.

The use of gene expression and mutation classifications questions the use of the same type of therapeutic approach for all CC since both among ECC and ICC clear differences on activation of pathways have been observed. Indeed, survival analyses performed in the ICC validation set disclosed an association between predicted subclass and the patient's clinical outcome, with patients whose tumors fell in subgroup A (gene expression-based classification) and C (mutation-based classification) showing the worst prognosis in terms of overall and relapse free survival. Due to the low number of samples, it was not possible to perform the analysis with a statistical significance in ECC cohort.

ICC patients with tumors classified as subgroup A based on gene expression predictor represent 43% and their poor prognosis is supported by up-regulation of pathways linked to intracellular signaling and epithelial mesenchymal transition and by an enrichment of exhausted T-cells in their infiltrate. These patients might therefore respond to checkpoint inhibitors provided that appropriate biomarkers are used since their poor prognosis is strongly determined by such peculiar type of immune infiltrate. ICC patients classified as subgroup C by mutation-based predictor represent 28% of the cohort and they were characterized by a median overall survival time < 30 months. These patients were defined

by the highest number of mutations on clinically actionable genes like as TP53, KRAS, PTEN, ARID2 and SMAD4 and, consequently, to oncogenic signaling pathways defined by these targets.

The integration of different datasets and the building of robust classifiers represent main strengths of our study broadening both biological and clinical implications. Indeed, given the rarity of this tumor entity, here we managed to assemble the largest cohort of CC cases profiled with transcriptomic and genomic platform, providing a comprehensive and reproducible tool to define CC subgroups. This effort will lay, once the aim2 will be completed, the groundwork for the evaluation of tumor evolution within single patients through the analysis of circulating tumor cells.

The molecular analysis of CTCs represents a valuable option in the study of CC, offering an alternative and easy-to-get source of tumor material, and allowing for a better understanding of intra-tumor subclonal composition. Moreover, since CTCs can be repeatedly assessed over time through simple blood draws, they can be used to monitor disease evolution in response to treatment. Leveraging on their potential role in treatment monitoring in CC [29], [78], we collected and profiled with lpWGS 88 CTCs from 28 patients enrolled in our institute to map evolutionary trajectory using signatures extracted from the previous aims of the study. Overall, CTCs showed a high level of heterogeneity intra- and extra-patients in both ICC and ECC cohort in terms of CNAs profile, similarly to what reported in a study on 14 patients with metastatic castration-resistant prostate cancer where CNA profiles of 185 single CTCs were analyzed [79].

The presence of CTCs collected along follow-up within single patients, opened the possibility to apply phylogenetic algorithm to map clonal architecture of the tumor and clarify its evolution during time. For such a purpose, since the definition of predictive tool based on transcriptomic and genomic signatures to apply on CTCs is currently ongoing (aim2), we considered genes in chromosomal region distinguishing responder to non-responder patients in CC as marker for the phylogenetic analysis [29]. The test performed on patient BT24 showed how CTCs were able to track tumor heterogeneity revealing the existence of clones harboring different CNAs that specifically characterize each time point of the disease. Taking together, this first attempt of phylogenetic analysis demonstrated the

utility to characterize CTCs to understand the tumor evolution and to predict, ideally, the course of the disease under a specific therapy pressure within single patient.

The current work is still ongoing and has limitations. A major limitation deals with the low number of patients with ECC tumors. This did not hinder identification and validation of ECC subgroups, but definitely interfered with evaluation of their prognostic relevance. Nonetheless, the interesting results obtained in ICC, imply that our approach levering public datasets is valuable and can be further implemented in the future when more ECC dataset will be available. Finally, the phylogenetic analysis was based on the application of standard algorithm already reported in literature: with the collection of a higher number of CTCs the future step will be represented by the implementation of new method specifically related to our context.

This study will offer a concrete opportunity to generate a pipeline for monitoring cancer evolution under treatment pressure providing new insights for the comprehension of tumor heterogeneity and progression within individual patients. Leveraging on this effort, the expectation is to generate a useful framework not only for CC but also for other tumor types.

6.REFERENCES

[1] N. ul A. Tariq, M. G. McNamara, and J. W. Valle, " Biliary tract cancers: current knowledge, clinical candidates and future challenges ," *Cancer Manag Res*, vol. 11, pp. 2623–2642, Mar. 2019, doi: 10.2147/CMAR.S157092.

[2] H. Charbel and F. H. Al-Kawas, "Cholangiocarcinoma: Epidemiology, Risk Factors, Pathogenesis, and Diagnosis," *Current Gastroenterology Reports 2011 13:2*, vol. 13, no. 2, pp. 182–187, Jan. 2011, doi: 10.1007/S11894-011-0178-8.

[3] T. P. Hennedige, W. T. Neo, and S. K. Venkatesh, "Imaging of malignancies of the biliary tract- an update," *Cancer Imaging 2014 14:1*, vol. 14, no. 1, pp. 1–21, Apr. 2014, doi: 10.1186/1470-7330-14-14.

[4] J. M. Banales *et al.*, "Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the European Network for the Study of Cholangiocarcinoma (ENS-CCA)," *Nature Reviews Gastroenterology & Hepatology 2016 13:5*, vol. 13, no. 5, pp. 261–280, Apr. 2016, doi: 10.1038/nrgastro.2016.51.

[5] S. A. Khan, S. Tavolari, and G. Brandi, "Cholangiocarcinoma: Epidemiology and risk factors," *Liver International*, vol. 39, no. S1, pp. 19–31, May 2019, doi: 10.1111/LIV.14095.

[6] T. Patel, "Worldwide trends in mortality from biliary tract malignancies," *BMC Cancer*, vol. 2, no. 1, pp. 1–5, May 2002, doi: 10.1186/1471-2407-2-10/FIGURES/1.

[7] P. Bertuccio, C. Bosetti, F. Levi, A. Decarli, E. Negri, and C. la Vecchia, "A comparison of trends in mortality from primary liver cancer and intrahepatic cholangiocarcinoma in Europe," *Annals of Oncology*, vol. 24, no. 6, pp. 1667–1674, Jun. 2013, doi: 10.1093/annonc/mds652.

[8] S. D. Taylor-Robinson *et al.*, "Increase in mortality rates from intrahepatic cholangiocarcinoma in England and Wales 1968–1998," *Gut*, vol. 48, no. 6, pp. 816–820, Jun. 2001, doi: 10.1136/GUT.48.6.816.

[9] C. Anderson and R. Kim, "Adjuvant therapy for resected extrahepatic cholangiocarcinoma: A review of the literature and future directions," *Cancer Treat Rev*, vol. 35, no. 4, pp. 322–327, Jun. 2009, doi: 10.1016/j.ctrv.2008.11.009.

[10] J. W. Valle *et al.*, “Biliary cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up,” *Annals of Oncology*, vol. 27, pp. v28–v37, Sep. 2016, doi: 10.1093/annonc/mdw324.

[11] H. Nathan, T. M. Pawlik, C. L. Wolfgang, M. A. Choti, J. L. Cameron, and R. D. Schulick, “Trends in Survival after Surgery for Cholangiocarcinoma: A 30-Year Population-Based SEER Database Analysis,” *Journal of Gastrointestinal Surgery 2007 11:11*, vol. 11, no. 11, pp. 1488–1497, Sep. 2007, doi: 10.1007/S11605-007-0282-0.

[12] J. Bridgewater *et al.*, “Guidelines for the diagnosis and management of intrahepatic cholangiocarcinoma,” *J Hepatol*, vol. 60, no. 6, pp. 1268–1289, Jun. 2014, doi: 10.1016/J.JHEP.2014.01.021.

[13] M. Miyazaki *et al.*, “Clinical implication of surgical resection for recurrent biliary tract cancer: Does it work or not?,” *Ann Gastroenterol Surg*, vol. 1, no. 3, pp. 164–170, Sep. 2017, doi: 10.1002/AGS3.12036.

[14] J. Valle *et al.*, “Cisplatin plus Gemcitabine versus Gemcitabine for Biliary Tract Cancer,” *New England Journal of Medicine*, vol. 362, no. 14, pp. 1273–1281, Apr. 2010, doi: 10.1056/NEJMOA0908721/SUPPL_FILE/NEJMOA0908721_DISCLOSURES.PDF.

[15] D.-Y. Oh *et al.*, “Durvalumab plus Gemcitabine and Cisplatin in Advanced Biliary Tract Cancer,” *NEJM Evidence*, Jun. 2022, doi: 10.1056/EVIDOA2200015.

[16] G. K. Abou-Alfa *et al.*, “Ivosidenib in IDH1-mutant, chemotherapy-refractory cholangiocarcinoma (ClarIDHy): a multicentre, randomised, double-blind, placebo-controlled, phase 3 study,” *Lancet Oncol*, vol. 21, no. 6, pp. 796–807, Jun. 2020, doi: 10.1016/S1470-2045(20)30157-1.

[17] G. K. Abou-Alfa *et al.*, “Pemigatinib for previously treated, locally advanced or metastatic cholangiocarcinoma: a multicentre, open-label, phase 2 study,” *Lancet Oncol*, vol. 21, no. 5, pp. 671–684, May 2020, doi: 10.1016/S1470-2045(20)30109-1.

[18] F. Meric-Bernstam *et al.*, "Futibatinib, an Irreversible FGFR1–4 Inhibitor, in Patients with Advanced Solid Tumors Harboring FGF/FGFR Aberrations: A Phase I Dose-Expansion Study," *Cancer Discov*, vol. 12, no. 2, pp. 402–415, Feb. 2022, doi: 10.1158/2159-8290.CD-21-0697/673838/AM/FUTIBATINIB-AN-IRREVERSIBLE-FGFR1-4-INHIBITOR-IN.

[19] V. Mazzaferro *et al.*, "Derazantinib (ARQ 087) in advanced or inoperable FGFR2 gene fusion-positive intrahepatic cholangiocarcinoma," *British Journal of Cancer 2018 120:2*, vol. 120, no. 2, pp. 165–171, Nov. 2018, doi: 10.1038/s41416-018-0334-0.

[20] V. Subbiah *et al.*, "Dabrafenib plus trametinib in patients with BRAFV600E-mutated biliary tract cancer (ROAR): a phase 2, open-label, single-arm, multicentre basket trial," *Lancet Oncol*, vol. 21, no. 9, pp. 1234–1243, Sep. 2020, doi: 10.1016/S1470-2045(20)30321-1.

[21] M. Javle *et al.*, "Pertuzumab and trastuzumab for HER2-positive, metastatic biliary tract cancer (MyPathway): a multicentre, open-label, phase 2a, multiple basket study," *Lancet Oncol*, vol. 22, no. 9, pp. 1290–1300, Sep. 2021, doi: 10.1016/S1470-2045(21)00336-3.

[22] A. Lamarca *et al.*, "Second-line FOLFOX chemotherapy versus active symptom control for advanced biliary tract cancer (ABC-06): a phase 3, open-label, randomised, controlled trial," *Lancet Oncol*, vol. 22, no. 5, pp. 690–701, May 2021, doi: 10.1016/S1470-2045(21)00027-9.

[23] J. Valle *et al.*, "Cisplatin plus Gemcitabine versus Gemcitabine for Biliary Tract Cancer," *New England Journal of Medicine*, vol. 362, no. 14, pp. 1273–1281, Apr. 2010, doi: 10.1056/NEJMOA0908721/SUPPL_FILE/NEJMOA0908721_DISCLOSURES.PDF.

[24] C. Braconi, S. Roessler, B. Kruk, F. Lammert, M. Krawczyk, and J. B. Andersen, "Molecular perturbations in cholangiocarcinoma: Is it time for precision medicine?," *Liver International*, vol. 39, no. S1, pp. 32–42, May 2019, doi: 10.1111/LIV.14085.

[25] D. Sia *et al.*, "Integrative molecular analysis of intrahepatic cholangiocarcinoma reveals 2 classes that have different outcomes.," *Gastroenterology*, vol. 144, no. 4, pp. 829–40, Apr. 2013, doi: 10.1053/j.gastro.2013.01.001.

[26] J. B. Andersen *et al.*, “Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors.,” *Gastroenterology*, vol. 142, no. 4, pp. 1021-1031.e15, Apr. 2012, doi: 10.1053/j.gastro.2011.12.005.

[27] R. Montal *et al.*, “Molecular classification and therapeutic targets in extrahepatic cholangiocarcinoma,” *J Hepatol*, vol. 73, no. 2, pp. 315–327, Aug. 2020, doi: 10.1016/J.JHEP.2020.03.008.

[28] H. Nakamura *et al.*, “Genomic spectra of biliary tract cancer,” *Nature Genetics 2015 47:9*, vol. 47, no. 9, pp. 1003–1010, Aug. 2015, doi: 10.1038/ng.3375.

[29] C. Reduzzi *et al.*, “A novel circulating tumor cell subpopulation for treatment monitoring and molecular characterization in biliary tract cancer,” *Int J Cancer*, vol. 146, no. 12, pp. 3495–3503, Jun. 2020, doi: 10.1002/IJC.32822.

[30] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, “A primer on deep learning in genomics,” *Nature Genetics 2018 51:1*, vol. 51, no. 1, pp. 12–18, Nov. 2018, doi: 10.1038/s41588-018-0295-5.

[31] H. Husi, “Computational Biology,” Nov. 2019, doi: 10.15586/COMPUTATIONALBIOLOGY.2019.

[32] K. Pantel and C. Alix-Panabières, “Circulating tumour cells in cancer patients: Challenges and perspectives,” *Trends Mol Med*, vol. 16, no. 9, pp. 398–406, Sep. 2010, doi: 10.1016/j.molmed.2010.07.001.

[33] L. Keller and K. Pantel, “Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells,” *Nature Reviews Cancer 2019 19:10*, vol. 19, no. 10, pp. 553–567, Aug. 2019, doi: 10.1038/s41568-019-0180-2.

[34] K. Pantel and C. Alix-Panabières, “Real-time liquid biopsy in cancer patients: Fact or fiction?,” *Cancer Res*, vol. 73, no. 21, pp. 6384–6388, Nov. 2013, doi: 10.1158/0008-5472.CAN-13-2030/659108/P/REAL-TIME-LIQUID-BIOPSY-IN-CANCER-PATIENTS-FACT-OR.

[35] C. Alix-Panabières and K. Pantel, “Clinical applications of circulating tumor cells and circulating tumor DNA as liquid biopsy,” *Cancer Discov*, vol. 6, no. 5, pp. 479–491, May 2016, doi: 10.1158/2159-8290.CD-15-1483/43226/P/CLINICAL-APPLICATIONS-OF-CIRCULATING-TUMOR-CELLS.

[36] S. bin Lim, C. T. Lim, and W. T. Lim, “Single-Cell Analysis of Circulating Tumor Cells: Why Heterogeneity Matters,” *Cancers 2019, Vol. 11, Page 1595*, vol. 11, no. 10, p. 1595, Oct. 2019, doi: 10.3390/CANCERS11101595.

[37] D. J. E. Peeters *et al.*, “Semiautomated isolation and molecular characterisation of single or highly purified tumour cells from CellSearch enriched blood samples using dielectrophoretic cell sorting,” *British Journal of Cancer 2013 108:6*, vol. 108, no. 6, pp. 1358–1367, Mar. 2013, doi: 10.1038/bjc.2013.92.

[38] R. Lampignano *et al.*, “A Novel Workflow to Enrich and Isolate Patient-Matched EpCAMhigh and EpCAMlow/negative CTCs Enables the Comparative Characterization of the PIK3CA Status in Metastatic Breast Cancer,” *International Journal of Molecular Sciences 2017, Vol. 18, Page 1885*, vol. 18, no. 9, p. 1885, Aug. 2017, doi: 10.3390/IJMS18091885.

[39] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/BIOINFORMATICS/BTU170.

[40] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, p. 15, Jan. 2013, doi: 10.1093/BIOINFORMATICS/BTS635.

[41] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/BIOINFORMATICS/BTQ033.

[42] F. García-Alcalde *et al.*, “Qualimap: evaluating next-generation sequencing alignment data,” *Bioinformatics*, vol. 28, no. 20, pp. 2678–2679, Oct. 2012, doi: 10.1093/BIOINFORMATICS/BTS503.

[43] B. Li and C. N. Dewey, “RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome,” *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–16, Aug. 2011, doi: 10.1186/1471-2105-12-323.

[44] F. Hach *et al.*, “mrsFAST: a cache-oblivious algorithm for short-read mapping,” *Nature Methods 2010 7:8*, vol. 7, no. 8, pp. 576–577, Aug. 2010, doi: 10.1038/nmeth0810-576.

[45] V. A. Adalsteinsson *et al.*, “Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors,” *Nature Communications 2017 8:1*, vol. 8, no. 1, pp. 1–13, Nov. 2017, doi: 10.1038/s41467-017-00965-y.

[46] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, doi: 10.1093/BIOSTATISTICS/KXJ037.

[47] A. T. L. Lun, D. J. McCarthy, and J. C. Marioni, “A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor,” *F1000Res*, vol. 5, 2016, doi: 10.12688/F1000RESEARCH.9501.2.

[48] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set,” *J Stat Softw*, vol. 61, no. 6, pp. 1–36, Nov. 2014, doi: 10.18637/JSS.V061.I06.

[49] F. Murtagh and P. Legendre, “Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion?,” *Journal of Classification 2014 31:3*, vol. 31, no. 3, pp. 274–295, Oct. 2014, doi: 10.1007/S00357-014-9161-Z.

[50] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res*, vol. 43, no. 7, pp. e47–e47, Apr. 2015, doi: 10.1093/NAR/GKV007.

[51] A. Subramanian *et al.*, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG.

[52] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The Molecular Signatures Database (MSigDB) hallmark gene set collection," *Cell Syst*, vol. 1, no. 6, p. 417, Dec. 2015, doi: 10.1016/J.CELS.2015.12.004.

[53] M. Foroutan, D. D. Bhuva, R. Lyu, K. Horan, J. Cursons, and M. J. Davis, "Single sample scoring of molecular phenotypes," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–10, Nov. 2018, doi: 10.1186/S12859-018-2435-4/FIGURES/2.

[54] J. M. Banales *et al.*, "Cholangiocarcinoma 2020: the next horizon in mechanisms and management," *Nature Reviews Gastroenterology & Hepatology 2020 17:9*, vol. 17, no. 9, pp. 557–588, Jun. 2020, doi: 10.1038/s41575-020-0310-z.

[55] Y. R. Miao *et al.*, "ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy," *Advanced Science*, vol. 7, no. 7, Apr. 2020, doi: 10.1002/ADVS.201902880.

[56] F. Sanchez-Vega *et al.*, "Oncogenic Signaling Pathways in The Cancer Genome Atlas," *Cell*, vol. 173, no. 2, pp. 321-337.e10, Apr. 2018, doi: 10.1016/J.CELL.2018.03.035.

[57] A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, and H. P. Koeffler, "Maftools: efficient and comprehensive analysis of somatic variants in cancer," *Genome Res*, vol. 28, no. 11, pp. 1747–1756, Nov. 2018, doi: 10.1101/GR.239244.118.

[58] S. L. Freshour *et al.*, "Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts," *Nucleic Acids Res*, vol. 49, no. D1, pp. D1144–D1151, Jan. 2021, doi: 10.1093/NAR/GKAA1084.

[59] T. N. Vu *et al.*, "Comprehensive landscape of subtype-specific coding and non-coding RNA transcripts in breast cancer," *Oncotarget*, vol. 7, no. 42, pp. 68851–68863, Sep. 2016, doi: 10.18632/ONCOTARGET.11998.

[60] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.

[61] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "-Nearest Neighbor Classification," pp. 83–106, 2009, doi: 10.1007/978-0-387-88615-2_4.

[62] C. Cortes, V. Vapnik, and L. Saitta, "Support-vector networks," *Machine Learning 1995 20:3*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[63] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *https://doi.org/10.1214/08-AOAS169*, vol. 2, no. 3, pp. 841–860, Sep. 2008, doi: 10.1214/08-AOAS169.

[64] R. A. Horn and C. R. Johnson, "Matrix Analysis," *Matrix Analysis*, Dec. 1985, doi: 10.1017/CBO9780511810817.

[65] S. Piantadosi, "Clinical Trials: A Methodologic Perspective: Second Edition," *Clinical Trials: A Methodologic Perspective: Second Edition*, pp. 1–687, Jul. 2005, doi: 10.1002/0471740136.

[66] I. Scheinin *et al.*, "DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly," *Genome Res*, vol. 24, no. 12, pp. 2022–2032, Dec. 2014, doi: 10.1101/GR.175141.114.

[67] M. Vismara *et al.*, "Single-Cell Phenotypic and Molecular Characterization of Circulating Tumor Cells Isolated from Cryopreserved Peripheral Blood Mononuclear Cells of Patients with Lung Cancer and Sarcoma," *Clin Chem*, vol. 68, no. 5, pp. 691–701, May 2022, doi: 10.1093/CLINCHEM/HVAC019.

[68] L. de Sano *et al.*, "TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data," *Bioinformatics*, vol. 32, no. 12, pp. 1911–1913, Jun. 2016, doi: 10.1093/BIOINFORMATICS/BTW035.

[69] R. Montal *et al.*, "Molecular classification and therapeutic targets in extrahepatic cholangiocarcinoma," *J Hepatol*, vol. 73, no. 2, pp. 315–327, Aug. 2020, doi: 10.1016/j.jhep.2020.03.008.

[70] A. Jusakul *et al.*, "Whole-genome and epigenomic landscapes of etiologically distinct subtypes of cholangiocarcinoma," *Cancer Discov*, vol. 7, no. 10, pp. 1116–1135, Oct. 2017, doi: 10.1158/2159-8290.CD-17-0368/333261/AM/WHOLE-GENOME-AND-EPIGENOMIC-LANDSCAPES-OF.

[71] P. Warnat, R. Eils, and B. Brors, "Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes," *BMC Bioinformatics*, vol. 6, no. 1, pp. 1–15, Nov. 2005, doi: 10.1186/1471-2105-6-265/FIGURES/6.

[72] D. Castillo, J. M. Gálvez, L. J. Herrera, B. S. Román, F. Rojas, and I. Rojas, "Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–15, Nov. 2017, doi: 10.1186/S12859-017-1925-0/FIGURES/11.

[73] C. Meng, B. Kuster, A. C. Culhane, and A. M. Gholami, "A multivariate approach to the integration of multi-omics datasets," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–13, May 2014, doi: 10.1186/1471-2105-15-162/FIGURES/4.

[74] T. Ma, F. Liang, S. Oesterreich, and G. C. Tseng, "A Joint Bayesian Model for Integrating Microarray and RNA Sequencing Transcriptomic Data," *Journal of Computational Biology*, vol. 24, no. 7, pp. 647–662, Jul. 2017, doi: 10.1089/CMB.2017.0056/ASSET/IMAGES/LARGE/FIGURE6.JPEG.

[75] Y. T. Huang, "Integrative modeling of multi-platform genomic data under the framework of mediation analysis," *Stat Med*, vol. 34, no. 1, pp. 162–178, Jan. 2015, doi: 10.1002/SIM.6326.

[76] R. G. W. Verhaak *et al.*, "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, Jan. 2010, doi: 10.1016/j.ccr.2009.12.020.

[77] A. Irigoyen *et al.*, "Integrative multi-platform meta-analysis of gene expression profiles in pancreatic ductal adenocarcinoma patients for identifying novel diagnostic biomarkers," *PLoS One*, vol. 13, no. 4, Apr. 2018, doi: 10.1371/JOURNAL.PONE.0194844.

[78] A. C. Backen *et al.*, “Circulating biomarkers during treatment in patients with advanced biliary tract cancer receiving cediranib in the UK ABC-03 trial,” *British Journal of Cancer 2018 119:1*, vol. 119, no. 1, pp. 27–35, Jun. 2018, doi: 10.1038/s41416-018-0132-8.

[79] M. B. Lambros *et al.*, “Single-cell analyses of prostate cancer liquid biopsies acquired by apheresis,” *Clinical Cancer Research*, vol. 24, no. 22, pp. 5635–5644, Nov. 2018, doi: 10.1158/1078-0432.CCR-18-0862/255700/AM/SINGLE-CELL-ANALYSES-OF-PROSTATE-CANCER-LIQUID.

7.PUBLICATIONS

The following section contains all the articles/conference-articles published during the three years of the PhD course in Precision Medicine at University of Brescia (XXXV cycle).
"*****" highlights articles where the candidate served as first name author.

***1.** *Silvestri M, Nghia Vu T, Nichetti F, et al. Comprehensive transcriptomic analysis to identify biological and clinical differences in cholangiocarcinoma. Cancer Medicine 2023, under revision* (Article)

***2.** *Silvestri M, Dugo M, Vismara M, et al. Copy number alterations analysis of primary tumor tissue and circulating tumor cells from patients with early-stage triple negative breast cancer. Scientific Reports 2022 12:1. 2022;12(1):1-7. doi:10.1038/s41598-022-05502-6.* (Article)

***3.** *Silvestri M, Reduzzi C, Feliciello G, et al. Detection of Genomically Aberrant Cells within Circulating Tumor Microemboli (CTMs) Isolated from Early-Stage Breast Cancer Patients. Cancers 2021, Vol 13, Page 1409. 2021;13(6):1409. doi:10.3390/CANCERS13061409.* (Article)

***4.** *Silvestri M, Vingiani A, de Cecco L, et al. 12P The RODILIA pilot study for molecular screening of patients with metaplastic breast cancer. Annals of Oncology. 2021;32:S25-S26. doi:10.1016/j.annonc.2021.03.026.* (Conference-article)

***5.** *La Rocca E, de Santis MC, Silvestri M, et al. Early stage breast cancer follow-up in real-world clinical practice: the added value of cell free circulating tumor DNA. J Cancer Res Clin Oncol. 2022;148(6):1543-1550. doi:10.1007/S00432-022-03990-7.* (Article)

***6.** *Di Cosimo S, Appierto V, Silvestri M, et al. Targeted-Gene Sequencing to Catch Triple Negative Breast Cancer Heterogeneity before and after Neoadjuvant Chemotherapy. Cancers (Basel). 2019;11(11):1753. doi:10.3390/cancers11111753.* (Article)

7. De Mattia E, Silvestri M, Polesel J, et al. Rare genetic variant burden in DPYD predicts severe fluoropyrimidine-related toxicity risk. Biomedicine & Pharmacotherapy. 2022;154:113644. doi:10.1016/J.BIOPHA.2022.113644. (Article)

8. Reduzzi C, Gerratana L, Zhang Y, et al. CK+/CD45+ (dual-positive) circulating cells are associated with prognosis in patients with advanced breast cancer.

https://doi.org/101200/JCO20224016_suppl1093.2022;40(16_suppl):1093-1093. doi:10.1200/JCO.2022.40.16_SUPPL.1093. (Conference-article)

9. Vismara M, Reduzzi C, Silvestri M, et al. Single-Cell Phenotypic and Molecular Characterization of Circulating Tumor Cells Isolated from Cryopreserved Peripheral Blood Mononuclear Cells of Patients with Lung Cancer and Sarcoma. Clin Chem. 2022;68(5):691-701. doi:10.1093/CLINCHEM/HVAC019. (Article)

10. Di Cosimo S, Appierto V, Ortolan E, et al. Circulating tumor DNA and disease recurrence in early stage breast cancer: From a case-control study to a prospective longitudinal trial. Annals of Oncology. 2019;30:iii28-iii29. doi:10.1093/ANNONC/MDZ096.005. (Article)

11. Reduzzi C, Vismara M, Silvestri M, et al. Abstract 1390: Molecular characterization of circulating tumor cells in cholangiocarcinoma patients: A new tool for treatment management. Cancer Res. 2019;79(13 Supplement):1390-1390. doi:10.1158/1538-7445.AM2019-1390. (Conference-article)

12. Di Cosimo S, Silvestri M, Dugo M, et al. 59P Primary tumour and circulating tumour cell (CTC) copy number alterations (CNAs) in triple negative breast cancer (TNBC) patients (pts) treated with neoadjuvant chemotherapy (NAC). Annals of Oncology. 2020;31:S35-S36. doi:10.1016/J.ANNONC.2020.03.193. (Conference-article)

13. Reduzzi C, Vismara M, Silvestri M, et al. A novel subpopulation of circulating tumor cells in patients with cholangiocarcinoma. https://doi.org/101200/JCO20193715_suppl.e15637. 2019;37(15_suppl):e15637-e15637. doi:10.1200/JCO.2019.37.15_SUPPL.E15637. (Conference-article)

14. Romani C, Capoferri D, Grillo E, et al. The Claudin-Low Subtype of High-Grade Serous Ovarian Carcinoma Exhibits Stem Cell Features. Cancers 2021, Vol 13, Page 906. 2021;13(4):906. doi:10.3390/CANCERS13040906. (Article)

15. Ortolan E, Appierto V, Silvestri M, et al. Blood-based genomics of triple-negative breast cancer progression in patients treated with neoadjuvant chemotherapy. ESMO Open. 2021;6(2):100086. doi:10.1016/J.ESMOOP.2021.100086. (Article)

16. Romani C, Zizioli V, Silvestri M, et al. Low Expression of Claudin-7 as Potential Predictor of Distant Metastases in High-Grade Serous Ovarian Carcinoma Patients. Front Oncol. 2020;10:1287. doi:10.3389/fonc.2020.01287. (Article)

17. Di Cosimo, Appierto V, Pizzamiglio S, et al. Early Modulation of Circulating MicroRNAs Levels in HER2-Positive Breast Cancer Patients Treated with Trastuzumab-Based Neoadjuvant Therapy. Int J Mol Sci. 2020;21(4):1386. doi:10.3390/ijms21041386. (Article)

18. Di Cosimo S, Appierto V, Silvestri M, et al. Primary tumor somatic mutations in the blood of women with ductal carcinoma in situ of the breast. Annals of Oncology. 2020;31(3):435-437. doi:10.1016/j.annonc.2019.11.022. (Article)

19. Reduzzi C, Vismara M, Silvestri M, et al. A novel circulating tumor cell subpopulation for treatment monitoring and molecular characterization in biliary tract cancer. Int J Cancer. 2020;146(12):3495-3503. doi:10.1002/ijc.32822. (Article)

20. Cappelletti V, Verzoni E, Ratta R, et al. Analysis of Single Circulating Tumor Cells in Renal Cell Carcinoma Reveals Phenotypic Heterogeneity and Genomic Alterations Related to Progression. Int J Mol Sci. 2020;21(4):1475. doi:10.3390/ijms21041475. (Article)

21. Reduzzi C, Vismara M, Gerratana L, et al. The curious phenomenon of dual-positive circulating cells: Longtime overlooked tumor cells. Semin Cancer Biol. Published online October 15, 2019. doi:10.1016/J.SEMCANCER.2019.10.008. (Article)