La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 "Dottorati Innovativi con caratterizzazione Industriale"

UNIVERSITÀ DEGLI STUDI DI MESSINA
Dipartimento di Medicina Clinica e Sperimentale

Dottorato di Ricerca
Biotecnologie Mediche e Chirurgiche

XXXIII ciclo
Coordinatore: Chiar.mo Prof. G. Squadrito

Sviluppo di una metodica *custom* di *NGS* e di una *pipeline* bioinformatica *ad hoc* per la caratterizzazione delle integrazioni del Virus dell'Epatite B nel Carcinoma Epatocellulare.

Tesi di Dottorato di:
Dott. Domenico Giosa

Tutor:
Chiar.ma Prof.ssa Teresa Pollicino

(SSD: MED/04)

Anno Accademico 2019/2020

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 "Dottorati Innovativi con caratterizzazione Industriale"

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università e della Ricerca

PON RICERCA E INNOVAZIONE

UNIVERSITÀ DEGLI STUDI DI MESSINA
Dipartimento di Medicina Clinica e Sperimentale

Dottorato di Ricerca
Biotecnologie Mediche e Chirurgiche

XXXIII ciclo
Coordinatore: Chiar.mo Prof. G. Squadrito

Sviluppo di una metodica *custom* di *NGS* e di una *pipeline* bioinformatica *ad hoc* per la caratterizzazione delle integrazioni del Virus dell'Epatite B nel Carcinoma Epatocellulare.

Tesi di Dottorato di:
Dott. Domenico Giosa

Tutor:
Chiar.ma Prof.ssa Teresa Pollicino

(SSD: MED/04)

Anno Accademico 2019/2020

Indice

1. Introduzione

L'infezione cronica da virus dell'Epatite B (Hepatitis B Virus - HBV) è uno dei principali fattori di rischio per lo sviluppo del carcinoma epatocellulare (HCC). Infatti, pazienti affetti da epatite cronica B (CHB) presentano un rischio, fino a 100 volte maggiore, rispetto agli individui sani, di sviluppare HCC. L'HCC rappresenta, attualmente, la quarta causa di morte correlata al cancro a livello mondiale, con circa 780.000 decessi per anno (El-Serag 2012, Bray et al. 2018). L'HBV, membro del genere degli *Orthohepadnavirus,* è un virus a DNA. Il suo genoma è circolare, a doppia elica incompleta, delle dimensioni di circa 3.200 nucleotidi, composto da 4 *Open Reading Frames* (ORF) parzialmente sovrapposte tra loro ed in grado di codificare per 7 differenti proteine (Schaefer et al. 2007, Tu et al. 2017). L'integrazione del DNA virale nel genoma umano è stata riscontrata in circa il 90% dei casi di HCC correlati ad infezione da HBV, ed il riscontro della sua presenza in tessuti epatici non cirrotici tumorali provenienti da bambini o da giovani pazienti supporta ulteriormente il potenziale coinvolgimento dell'integrazione virale nel processo di carcinogenesi epatica (Brechot et al. 1980, Edman et al. 1980, Sezaki et al. 2004). È noto, infatti, che le integrazioni del DNA virale possano condurre ad instabilità cromosomica, mutagenesi inserzionale, alterazione dell'espressione genica ed all'espressione di prodotti virali (in particolare, le proteine HBs e HBx) o di loro isoforme mutate, note per le loro proprietà oncogeniche (Levrero and Zucman-Rossi 2016, Tu et al. 2017). Diversi aspetti relativi al meccanismo di integrazione del virus nel genoma dell'ospite non sono ancora stati chiariti. Non è noto, ad esempio, quale sia la forma molecolare di HBV coinvolta nel processo di integrazione, se sia il DNA a singola elica o quello a doppia elica lineare ad inserirsi nelle interruzioni della contiguità cromosomica. Inoltre, è attualmente oggetto di dibattito se il meccanismo di integrazione avvenga mediante l'unione canonica, non omologa, indipendente dalla sequenza (*sequence-independent Non-Homologous End-Joining* - NHEJ) oppure preveda la presenza di meccanismi alternativi mediati da micro-omologia di sequenza fra la porzione virale e quella genomica target di integrazione

(*Microhomology-Mediated End-Joining* – MMEJ) (Mladenov et al. 2016, Zhao et al. 2016). È noto che gli eventi di integrazione virale nel genoma dell'ospite presentino una distribuzione casuale. Tuttavia, studi recenti hanno portato all'identificazione, nel tessuto tumorale, di geni interessati con maggiore frequenza da eventi di integrazione virale, tra cui i geni *CCNE1*, *MLL4* e *TERT* (Ding et al. 2012, Fujimoto et al. 2012, Fujimoto et al. 2016, Saigo et al. 2008). Va, tuttavia, sottolineato che tali siti di integrazione "ricorrenti" sono stati osservati in un numero molto limitato di casi di HCC, e che pertanto tali siti di integrazione sono da considerare eventi aneddotici se paragonati al numero totale dei siti di integrazione descritti sino ad oggi (circa 20.500) (https://bioinfo.uth.edu/VISDB) (Tang et al. 2019).

È stato, inoltre, evidenziato come non soltanto le regioni codificanti risultino interessate dagli eventi di integrazione virale, ma lo siano anche gli elementi ripetuti e/o strutturali del genoma, come gli elementi LINE (*Long Interspersed Nuclear Elements*), SINE (*Short Interspersed Nuclear Elements*), i centromeri e i telomeri (Bonilla Guerrero and Roberts 2005, Fujimoto et al. 2016, Zhao et al. 2016). Un importante studio ha recentemente riportato l'esistenza di trascritti chimerici, HBx-LINE1. Tali trascritti chimerici appaiono coinvolti nella promozione del processo di trasformazione epatocellulare, e sono stati riscontrati nel 23% degli HCC studiati, (Lau et al. 2014).

Per quanto concerne le regioni genomiche virali, recentemente è stato riportato che negli eventi di integrazione possono essere coinvolte tutte le regioni genomiche dell'HBV, anche se è stato osservato un maggior coinvolgimento delle regioni corrispondenti all'*HBx* ed all'*HBs* (Ruan et al. 2019, Podlaha et al. 2019).

I numerosissimi studi sull'integrazione di HBV pubblicati sino ad oggi, oltre ad utilizzare approcci tecnici assai diversi in termini di sensibilità e specificità, si sono spesso concentrati nell'analisi della sola regione virale dell'*HBx* (Tu et al. 2017, Tang et al. 2019). I risultati ottenuti sono stati, pertanto, estremamente eterogenei nell'identificazione e quantificazione dei siti d'integrazione virale.

Questo nostro studio si è, quindi, posto l'obiettivo di sviluppare sia una nuova metodica di sequenziamento per l'analisi dell'integrazione di HBV, basata sulla tecnologia di *Next Generation Sequencing* (NGS) (HBV *Integration Sequencing*, HBV-ISeq), sia di generare una specifica pipeline bioinformatica (HBV *Integration Finder*, HBVIF), al fine di identificare in maniera altamente sensibile e specifica gli eventi di integrazione virale nel genoma umano, e di stimare l'espansione clonale dei siti d'integrazione virale nell'HCC.

2. Materiali e Metodi

2.1 Tessuti epatici e linee cellulari

Sono stati studiati i tessuti tumorali di 7 pazienti HBsAg-positivi (6 uomini ed 1 donna; età media 66,1±8) affetti da HCC. Il relativo, tessuto non tumorale era disponibile da 6 dei 7 pazienti. Inoltre, sono stati esaminati tessuti epatici sani provenienti da 3 pazienti HBsAg-negativi sottoposti ad intervento chirurgico per metastasi epatica. Tutti i tessuti epatiti tumorali e non tumorali - ottenuti grazie alla collaborazione con la Chirurgia Oncologica dell'A.O.U. Policlinico "G. Martino" di Messina - sono risultati negativi sia all'infezione da virus dell'epatite Delta (HDV), che alle infezioni da virus dell'epatite C (HCV) e da virus dell'immunodeficienza acquisita (HIV), dopo specifica analisi molecolare. Ciascun tessuto epatico è stato crio-preservato in RNA later (Applied Biosystem/Ambion) a -80°C immediatamente dopo la resezione chirurgica. I dati demografici, le caratteristiche cliniche ed istologiche dei pazienti sono presentate in Tabella1.

Tabella 1. Caratteristiche demografiche, cliniche e virologiche dei pazienti in studio al momento della resezione chirurgica.

Patient	Age	Sex	Number of nodules	Dimension of the bigger nodule	Tumor grading	Months from antiviral treatment to surgery	Serum HBV DNA	Total liver HBV DNA*
0101	60	M	1	25 mm	G2	N.A.	4,00E+06 IU/ml	7,04E+03
0102	76	M	2	40 mm	G2	1	2,19E+02 IU/ml	3,46E+00
0103	78	F	1	30 mm	G1	40	N.D.	8,42E+00
0104	77	M	1	25 mm	G1	0	8,76E+07 UI/ml	N.D.
0105	55	M	2	20 mm	G2	90	N.D.	9,72E+00
0106	51	M	1	30 mm	G1	0	3,00E+05 IU/ml	2,90E+02
0107	66	M	1	35 mm	G2	9	N.D.	4,66E+00

Lo studio è stato approvato dal Comitato Etico dell'Università di Messina e tutti i pazienti analizzati hanno firmato il consenso informato.

Sono state, inoltre, analizzate 2 diverse linee cellulari stabili:

1) la linea PLC/PRF/5 ("Alexander"): cellule umane di epatocarcinoma contenenti diverse integrazioni di HBV (SIGMA *catalog number* 85061113; *Lot number*: 10D004), utilizzata come controllo positivo;

2) la linea Vero: cellule epiteliali renali del primate *Cercopithecus aethiops* (gentilmente fornita dalla Prof.ssa Maria Teresa Sciortino, Università di Messina), utilizzata come controllo negativo.

2.2 Estrazione del DNA e preparazione delle librerie

Per ciascun tessuto, il DNA è stato estratto mediante la classica metodica che prevede il trattamento con proteinasi K e l'estrazione con fenolo/cloroformio (Pearson and Stirling 2003). Successivamente, è stato utilizzato il kit LightCycler-Control Kit DNA (Roche Diagnostic) e lo strumento Light-Cycler (Roche Diagnostic) per quantificare la beta-globina nel DNA estratto e valutare, quindi, il numero di cellule presenti in ciascun tessuto epatico o linea cellulare analizzata.

Per l'identificazione dei siti di integrazione dell'HBV-DNA, è stato modificato il protocollo di sequenziamento dei siti di integrazione dell'HIV descritto da Cohn et al. nel 2015. Il DNA genomico isolato da circa 20 milioni di cellule è stato frammentato mediante sonicazione - utilizzando un omogenizzatore ad ultrasuoni (SONOPLUS, Bandelin) al 30% della potenza per 3 cicli (15min ON – 15min OFF) - per ottenere un intervallo di grandezza di frammenti di DNA pari a 100-1000bp. Le estremità dei frammenti sono state riparate mediante il kit End-It DNA Repair (Epicentre) e purificate utilizzando il kit MinElute Reaction Clean-up (Qiagen). Il DNA "blunted" è stato, quindi, adenosilato - mediante l'aggiunta di 1µl di dATP (10mM), 5µl di NEB buffer 2 (10X) e 2µl di Klenow *fragment* 3'->5' exo$^-$ (5000 U/ml) (New England Biolabs, Ipswich, MA) - e incubato per 1h a 37°C. I prodotti della reazione sono stati successivamente purificati mediante il kit MinElute Reaction Clean-up (Qiagen). I frammenti sono, quindi, stati ligati ad annealed-plinkers (200pmol di pLinkerTop + pLinkerBottom) - aggiungendo 4µl

di pLinkers, 5µl di NEB T4 DNA ligase buffer ed 1µl di T4 DNA ligase ($2x10^6$ U/ml) (New England Biolabs) ed incubati a 25°C per 1 ora - ed incubati *overnight* a 16°C. L'attività della ligasi è stata bloccata attraverso incubazione a 70°C per 20min. Successivamente sono state eseguite reazioni di semi-nested PCR utilizzando primers, *forward* o *reverse*, HBV-specifici. I frammenti di DNA ligati ai pLinkers arricchiti con i primer in *forward* o in *reverse* sono stati tenuti separati per il resto del protocollo. Il DNA arricchito è stato suddiviso in aliquote da 1µg ed a ciascuna aliquota sono stati aggiunti 20µl di Phusion HF buffer (5X), 3µl di dNTP (10mM), 1µl di primer (2,5µM) *forward* o *reverse* biotinilato (20mM), 1µl di Phusion Taq (2000U/ml) (New England Biolabs) ed H_2O fino al raggiungimento di 50µl.

La reazione di PCR è stata, quindi, eseguita utilizzando le seguenti condizioni:
1 ciclo di 98°C per 1min;

12 cicli di 98°C per 15sec - 65°C per 30sec - 72°C per 45sec;

1 ciclo di 72°C per 1min;

1 ciclo finale a 4°C ∞.

Dopo l'aggiunta di un' aliquota di pLinker (2,5µM) a ciascun prodotto di semi-nested PCR è stata eseguita un' ulteriore PCR, utilizzando le seguenti condizioni:
1 ciclo a 98°C per 1min;
35 cicli di 98°C per 15sec - 65°C per 30sec - 72°C per 45sec;

1 ciclo a 72°C per 5min;

1 ciclo a 4°C ∞.

In totale sono stati utilizzati 30 *primer* HBV-specifici, come mostrato, grazie al software Circos v 0.69-8 (Krzywinski et al. 2009), in (Figura 1).

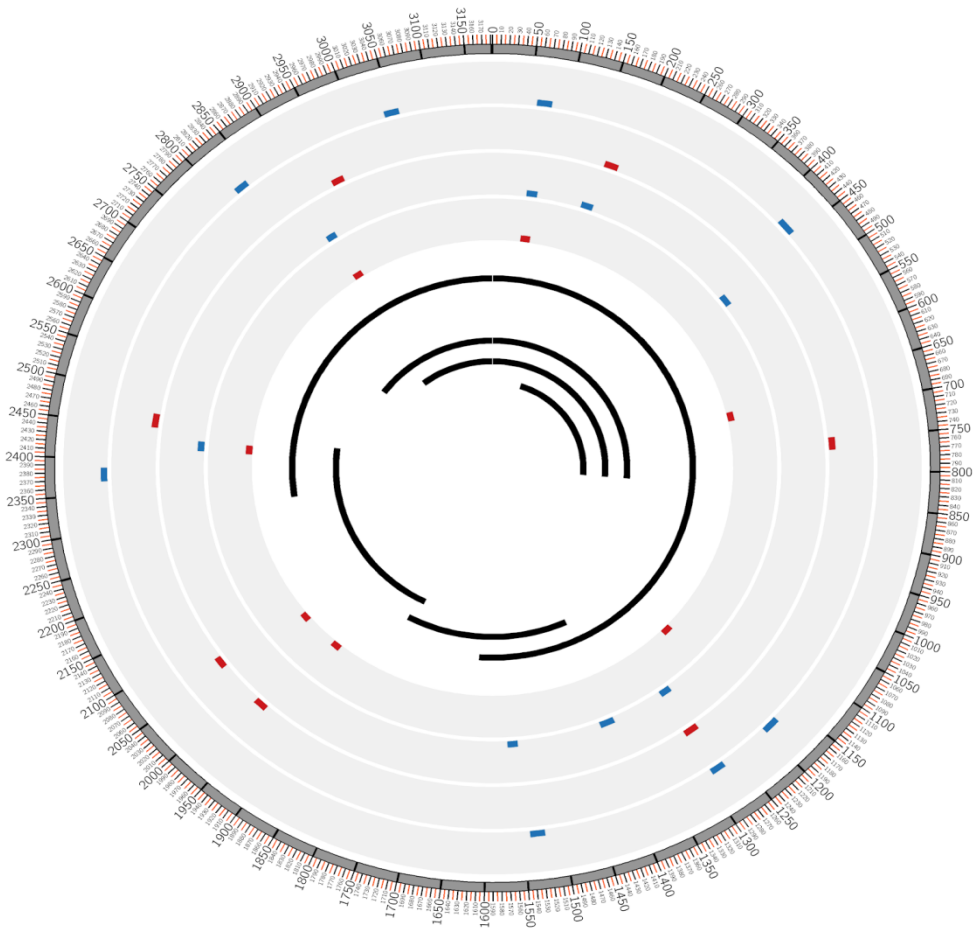

Figura 1. Rappresentazione grafica della localizzazione sul genoma virale dei primers (*forward* in blu; *reverse* in rosso) utilizzati per le reazioni di semi-nested HBV-specifiche.

Successivamente, i prodotti delle reazioni di PCR sono stati purificati, utilizzando il Qiaquick PCR purification kit (Qiagen), e separati mediante elettroforesi su gel di agarosio al 2%. I frammenti delle dimensioni di 300-1000bp sono stati purificati mediante estrazione con il Qiaquick gel extraction kit (Qiagen). Ai frammenti estratti da gel (prodotti di PCR sia in *forward* che in *reverse*) sono state aggiunte biglie magnetiche ligate a streptavidina (Invitrogen) ed è stata eseguita un'incubazione di 1h a temperatura ambiente sotto leggera agitazione. Le biglie sono state, quindi, isolate magneticamente e sottoposte a 3 successivi lavaggi in 500µl di B&W buffer 1x e 500µl di H_2O. Ad ogni aliquota da 25µl di biglie sono stati aggiunti 10µl di Phusion HF buffer (5X), 1,5µl di dNTP (10mM), 1µl di primer MiSeq-HBV (*forward* o *reverse*, 20µM), 1µl di MiSeq-pLinker (*forward* o *reverse*) (20µM) (tutti i primer denominati MiSeq contengono un adattatore per l'*annealing* alla superficie della *flow cell* Illumina), 0,5µl di Phusion Taq

(2000U/ml) ed 11µl di H_2O. Quindi ciascun campione è stato sottoposto alle seguenti condizioni di PCR:

1 ciclo a 98°C per 1min;

35 cicli di 98°C per 10sec - 65°C per 40sec - 72°C per 5min;

1 ciclo a 72°C per 5min;

1 ciclo a 4°C ∞.

I prodotti di PCR sono stati separati dalle biglie magneticamente e purificati a mezzo del Qiaquick PCR purification kit (Qiagen). Gli ampliconi ligati all'adattatore sono stati, quindi, indicizzati attraverso ulteriori 25 cicli di PCR utilizzando i primer Illumina *Index* 1 ed *Index* 2.

Le differenti librerie in *forward* e *reverse* di ciascun tessuto epatico sono state unite in quantità equimolari e sequenziate in *paired-end* su piattaforma Illumina MiSeq.

2.3 Estrazione dell'RNA e sequenziamento

L'RNA totale è stato estratto dai 3 tessuti epatici tumorali e non-tumorali ottenuti (T1, T2, T3 e NT1, NT2, NT3) e da 3 tessuti epatici sani (S1, S2, S3), utilizzando il reagente TRIzol (Invitrogen, Carlsbad, CA, USA), in accordo con le istruzioni del produttore. La qualità e l'integrità dell'RNA è stata valutata mediante lo strumento Agilent 2100 Bioanalyzer (Santa Clara, CA, USA). Il sequenziamento è stato eseguito secondo il protocollo TruSeq Stranded mRNA kit (Illumina) su piattaforma HiSeq 2500 (Illumina).

2.4 Validazione mediante PCR e Sequenziamento Sanger

Il DNA genomico di ciascun campione biologico è stato diluito serialmente e soggetto a nested-PCR utilizzando: (a) coppie di primer specifiche per le estremità delle chimere identificate (costituite, pertanto, da un primer specifico per la porzione genomica umana ed un primer specifico per la porzione genomica virale

della chimera), (b) la HotStart Taq Polymerase (Qiagen). Le condizioni utilizzate per la nested-PCR sono state le seguenti:

1 ciclo 98°C per 14min;

40 cicli di 98°C per 30sec - 55°C per 30sec - 72°C per 30sec;

1 ciclo a 72°C per 5min;

1 ciclo a 4°C ∞.

I prodotti di PCR sono stati quindi separati mediante elettroforesi su gel di agarosio (1,5%) ed estratti con il kit Qiaquick gel extraction (Qiagen). Ciascun prodotto di PCR è stato quindi sequenziato con la metodica del Sanger mediante lo strumento Applied Biosystem 3500 DNA analyzers (Applied Biosystem, Foster City, CA).

2.5 Analisi Bioinformatica

2.5.1 Identificazione delle integrazioni virali nel genoma dell'ospite

Al fine di rilevare tutti i siti di integrazione virale nel genoma dell'ospite, per ciascun campione è stata effettuata l'analisi di qualità delle *reads* mediante il *software* FastQC v0.11.8 (Andrews 2010). Le *reads* sono state, quindi, sottoposte a rimozione di adattatori e sequenze di bassa qualità mediante l'ausilio del software Trimmomaticv.0.39 (Bolger et at. 2014) utilizzando le seguenti opzioni: *sliding window* 5bp, *quality score* medio 16, lunghezza minima 35bp. Le rimanenti *reads* di buona qualità, siano esse rimaste *paired* o meno, sono state mappate con il *software* BWA v.0.7.17-r1188 (Li and Durbin 2009) contro il genoma ibrido umano-HBV (HG-HBV) ottenuto concatenando il genoma umano (*Genbank accession*: GCA_000001405.25; *genome assembly* GRCh38.p10) ed il genoma di HBV di genotipo D (*Genbank accession*: NC_003977.2). Dall'allineamento sono stati rimossi i duplicati ottici mediante il programma Picard tool v.2.22.0 (http://broadinstitute.github.io/picard), mentre il rimanente BAM è stato processato col *software* SAMtools v.1.9 (Li et al. 2009) per estrarre le chimere (*flag* utilizzato 2048, corrispondente al *supplementary alignment*

0x800). L'*utility* BEDtools v.2.28.0 (Quinlan 2014) è stata utilizzata per estrarre dagli allineamenti sia le coordinate derivanti dal *mapping* primario che dal secondario (utilizzando il *flag -cigar* nell'applicativo bamtobed), in modo da poter contare in modo corretto la profondità di sequenziamento per ciascun sito chimerico identificato. Le risultanti coordinate cromosomiche, sia umane che virali, sono state utilizzate per estrarre tutte le *reads* che mappavano in esse, al fine di ricostruire il *consensus* della sequenza chimerica mediante l'utilizzo di 2 iterazioni consecutive del *software* Cap3 (Huang and Madan 1999) ed una conclusiva con il *tool* cd-hit v.4.8.1 (Fu et al. 2012). I *consensus* chimerici ricostruiti sono dunque stati mappati nuovamente contro il genoma ibrido HG-HBV utilizzando l'algoritmo BLAST (Camacho et al. 2009) con le seguenti opzioni: *task*=blastn-short, *dust*=no, *soft_masking*=false, *word_size*=7, *penalty*=-3, *reward*=2, *gapopen*=5, *gapextend*=2. La presenza di micro-omologia (MH) è stata identificata con script *in house* basandosi sulle coordinate cromosomiche identificate con l'ultima iterazione di BLAST seguendo il principio che la stessa porzione della chimera (di almeno 1bp) risultasse mappare sia sul genoma umano che su quello virale. Infine le coordinate del genoma umano e quelle del genoma virale, corrispondenti alle 2 porzioni delle chimere, sono state confrontate con le rispettive annotazioni genomiche per poter dettagliare i *locus* genomici a livello dei quali è stato identificato ciascun evento di integrazione. Tali informazioni comprendono: (a) le coordinate genomiche del gene più vicino all'evento di integrazione (0, quando l'integrazione avviene all'interno del locus di quel dato gene); (b) la distanza, espressa in bp, dal gene più vicino; (c) la sovrapposizione del sito di integrazione con eventuali elementi ripetuti e/o trasponibili del genoma umano; (d) la porzione di genoma virale coinvolta nell'integrazione; (e) le sequenze di ciascuna chimera corrispondenti solo al genoma umano, solo al genoma virale o l'intera sequenza chimerica. Tutte i siti di integrazione identificati sono stati, infine, filtrati sulla base del numero minimo di *reads* chimeriche supportanti il sito di integrazione (≥3 *reads*) e allineate nella regione virale con un minimo di 32 nucleotidi. La pipeline (nominata HBV Integration Finder -

HBVIF), inoltre, produce diversi *file* in formato ".tab" visualizzabili su software quali Microsoft Excel o LibreOffice Calc, così da rendere più semplice ed immediata l'interpretazione dei risultati ottenuti. Infine, tale approccio bioinformatico è stato applicato sia sui dati ottenuti dall'HBV-ISeq del DNA dei pazienti sia su quelli provenienti dalle linee cellulari. Dai risultati sono state generate immagini mediante il pacchetto ggplot2 (Wickham 2016) del *software* R (R Core Team, 2019).

Uno schema rappresentativo della pipeline bioinformatica sviluppata è presentato in Figura 2.

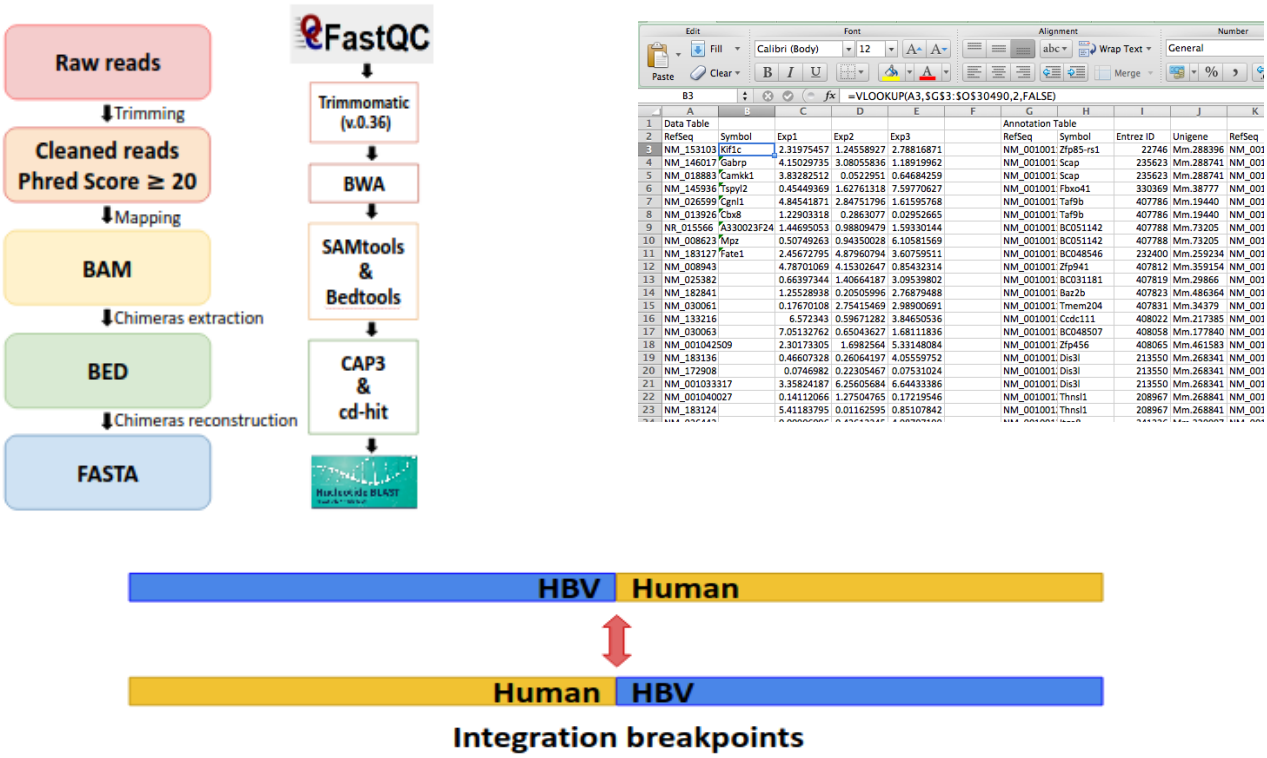

Figura 2. Rappresentazione schematica della pipeline bioinformatica e di un output visualizzabile su Microsoft Excel.

2.5.2 Analisi statistica per la caratterizzazione delle regioni umane e virali principalmente coinvolte negli eventi di integrazione genomica

I risultati ottenuti e suddivisi in differenti tipologie di confronto tra i tessuti tumorali e quelli non-tumorali sono stati comparati mediante differenti approcci statistici, tra cui il test del Chi-quadro ($\chi 2$), la correlazione punto biseriale (NPC) ed il test di Mann-Whitney, per determinare quali risultati mostrassero differenze statisticamente significative fra i due gruppi in oggetto, sulla base del numero di integrazioni virali trovate nel genoma dell'ospite.

2.5.3 Ricerca di *Hot-spot* genomici e motivi genomici sovra-rappresentati

Si definisce "*hot-spot"* quella regione genomica umana a livello della quale viene osservato un numero di integrazioni virali con una frequenza statisticamente superiore rispetto ad una frequenza attesa di *background*, generata casualmente. A tal fine è stato utilizzato il *tool* shuffle contenuto nella *suite* BEDtools v.2.28.0 (Quinlan 2014) per generare, a partire da ciascun sito di integrazione identificato, 3 *dataset* randomici corrispondenti a 100, 1.000 e 10.000 volte il numero dei siti di integrazione. Tali *dataset* sono stati utilizzati come *background* per poter effettuare il test statistico Monte Carlo (Sawilowsky 2003) e successivamente sono stati comparati, per ciascun cromosoma umano, il numero di eventi di integrazione osservati contro i differenti *dataset* simulati computazionalmente, attraverso l'utilizzo del $\chi 2$ test.

Inoltre, per la ricerca di motivi ricorrenti - intesi come sequenze genomiche ricorrenti nel totale delle sequenze chimeriche identificate - effettuata con il *software* HOMER (http://homer.ucsd.edu/homer), sono state valutate, a partire dai *consensus* chimerici, solo le sequenze virali, solo quelle umane o l'intera chimera. I tre *dataset* sono stati analizzati separatamente e per ciascuno di essi sono stati utilizzati 3 differenti *background*, ovvero l'intero genoma umano, l'intero genoma di HBV ed il genoma ibrido HG-HBV.

2.5.4 Determinazione della clonalità dei siti di integrazione

Al fine di valutare se le integrazioni identificate fossero clonali o meno, per ciascun sito di integrazione sono state estratte le *reads* corrispondenti ed allineate con il programma MAFFT v7.310 (Katoh and Standley 2013) utilizzando le seguenti opzioni: --localpair, --reorder, --maxiterate 1000. Gli allineamenti sono stati visualizzati, ed ove necessario corretti manualmente, attraverso il *tool* AliView v.2019 (Larsson 2014). Affinché un sito di integrazione fosse considerato clonale, almeno il 92% del totale delle *reads* a supporto dello stesso doveva possedere la medesima sequenza nucleotidica a cavallo del punto di integrazione (*breakpoint*), con un numero massimo di *mismatch* pari ad una base. Per le due tipologie di tessuto epatico analizzato (tumore o non-tumore), è stato confrontato il numero di siti di integrazione, espansi clonalmente, mediante il $\chi 2$ test (*P*-value threshold$\leq$0,05). Inoltre, lo stesso test è stato utilizzato per valutare le differenze - in termini di abbondanza relativa - fra il numero di integrazioni identificate in ciascun cromosoma umano. Infine è stato utilizzato il test di combinazione non parametrica (NPC) per calcolare il valore probabilistico del verificarsi delle integrazioni di HBV in ciascun cromosoma.

2.5.5 Determinazione dei geni interessati da integrazioni virali condivisi da diversi campioni

Al fine di determinare se differenti (almeno 2) tessuti epatici condividessero geni coinvolti da eventi di integrazione virale, sono state estrapolate le liste dei geni *target* di integrazione virale presenti in ciascun campione. La condivisione di tali geni è stata determinata mediante il pacchetto lapply v.3.6.0 del *software* R (R Core Team, 2019).

2.5.6 Analisi di espressione genica differenziale

Le *reads* provenienti dal sequenziamento dell'RNA sono state dapprima ispezionare con il *tool* FastQC v0.11.8 (Andrews 2010), quindi pulite con il *software* Trimmomatic v.0.39 (Bolger et at. 2014) utilizzando le seguenti opzioni*:* *sliding window* 4bp, *quality score* medio 25, *leading* 25, *trailing* 25, lunghezza minima 35bp. Le *reads* sono state mappate contro il genoma ibrido HG-HBV utilizzando il software STAR v.2.7.1a (Dobin et al. 2013). È stato, quindi, confermato che il sequenziamento fosse *strand-specific* mediante lo script in python infer_experiment.py contenuto nel programma RseQC v.2.6.4 (Wang et al. 2012). I livelli di espressione genica, per ciascun campione, sono stati valutati utilizzando l'*utility* featureCounts v2.0.0 contenuta dell'algoritmo Subread (Liao et al. 2016). Le matrici di *counts,* ovvero il numero di *reads* associate a ciascun gene in ciascun campione, sono quindi state importate nell'ambiente Rstudio (R Core Team, 2019) e normalizzate con il metodo del "Trimmed Mean of M-value" (TMM) utilizzando il pacchetto HTSfilter (Rau et al. 2013). L'analisi della componente principale (Principal Component Analysis; PCA) è stata effettuata utilizzando il pacchetto stats v.3.6.0 (R Core Team, 2019). I campioni sono stati separati in 3 classi, ovvero tumori (T), non tumori (NT) e sani (S), e successivamente è stato utilizzato il pacchetto edgeR v.3.28.0 (Robinson et al. 2010) per effettuare l'analisi di espressione genica differenziale, utilizzando un valore di *False Discovery Rate* (FDR) ≤0,05 per considerare i geni differenzialmente espressi. Tutti i risultati sono stati espessi in grafico mediante il pacchetto ggplot2 (Wickham 2016). I geni differenzialmente espressi sono stati sottoposti a *Gene Ontology Enrichment Analysis* (GOEA) utilizzando uno script *in-house* concepito in accordo con quello utilizzato da argiGO v.2.0 (Tian et al. 2017).

3 Risultati

3.1 Identificazione dei siti d'integrazione di HBV

Il sequenziamento ottenuto dalla metodica HBV-ISeq ha prodotto un totale di 134.122.676 *reads*, di cui 101.674.766 (75,8%) relative alle resezioni epatiche [(il 40,7% relative ai Tumori (T), il 25,9% ai non-tumori (NT), il 9,2% ai tessuti sani (S)] e le rimanenti 32.447.910 provenienti dal sequenziamento delle linee cellulari. Dopo la pulizia, è rimasto di buona qualità (*paired* o *unpair*) un totale di 36.050.653 (~27%) *reads*, successivamente processato con la pipeline HBVIF. L'accoppiata HBV-ISeq/HBVIF relativa al DNA estratto dai tessuti dei pazienti HBsAg-negativi e dalle linee cellulari VERO non ha identificato alcuna integrazione virale nel genoma umano, confermando che l'HBV-ISeq non produce falsi positivi in campioni non infettati dall'HBV. Per quanto concerne la linea cellulare PLC/PRF/5 sono stati identificati 10 eventi di integrazione unici, confermando dati già riportati in letteratura. In particolare, sono stati confermati sia il sito di integrazione situato nella regione *promoter* del gene *TERT* che i siti di integrazione a livello dei geni *MVK*, *CCDC57* e *UNC5D* (Graef et al. 1994, Watanabe et al. 2015, Ishii et al. 2020). Di queste integrazioni, quella a livello della regione *promotor* di *TERT* coinvolgeva la porzione virale ENH1/X-promoter, quelle dei geni *MVK* e *CCDC57* coinvolgevano due differenti porzioni del gene *S*; infine, quella a livello del gene *UNC5D* coinvolgeva una porzione del gene *Core*. Delle altre integrazioni virali, una è stata riscontrata a livello di un lncRNA *LOC105375660* e coinvolgeva una porzione del gene *Core*, una è stata osservata in prossimità dello pseudogene ribosomiale *RPS23P4* e mostrava il coinvolgimento del gene virale *X*, mentre le rimanenti 4 integrazioni sono state riscontrate a livello di regioni intergeniche e coinvolgevano, in 3 casi, una porzione della regione genica virale *S* e, nell'altro caso, la regione ENH1/X promoter di HBV. Le integrazioni virali a livello del promotore di *TERT* e del gene *CCDC57* sono state confermate mediante PCR e sequenziamento in Sanger.

La metodica di sequenziamento dei siti d'integrazione di HBV e la successiva analisi bioinformatica, applicata ai 7 tessuti tumorali ed ai 6 tessuti non tumorali, si è rivelata estremamente sensibile, ed ha portato all'identificazione di un totale di 2.671 eventi unici di integrazione virale nel genoma umano. In particolare, 2.330 siti d'integrazione sono stati rivelati nei T e 341 nei NT, mostrando una differenza statisticamente significativa (*P*=0,027, *t-student test*). In tutti i tessuti analizzati sono state osservate integrazioni virali. Tuttavia, soltanto 24 *breakpoints* d'integrazione erano condivisi tra il tessuto tumorale e l'adiacente tessuto non tumorale. Dei 24 *breakpoint*, 13 erano condivisi tra T e NT del paziente 0104, 5 tra T e NT del paziente 0103, 3 tra T e NT del paziente 0102, 2 tra T e NT del paziente 0105 ed uno tra T e NT del paziente 0101.

3.2 Caratterizzazione delle regioni umane e virali coinvolte nell'integrazione genomica

Il confronto fra le coordinate delle integrazioni virali, l'annotazione umana e quella virale ha consentito di ottenere le frequenze di integrazione nelle differenti regioni del genoma umano, e ciò sia considerando ogni singolo paziente che il totale dei tessuti tumorali o dei non-tumorali. Tale analisi ha consentito di evidenziare come - a seconda della tipologia di *feature* genomica presa in considerazione - si potessero rilevare, frequentemente, differenze statisticamente significative nel confronto tra T e NT. Nonostante, sia nei T che nei NT, le regioni del genoma umano principalmente coinvolte dall'integrazione fossero quelle geniche - rappresentate dal 59,8% degli eventi totali nei tessuti non tumorali (204/341) e dal 53,4% (1.244/2.330) nei tessuti tumorali - tali eventi mostravano una differenza statisticamente significativa tra NT e T (*P*=0,025, mediante $\chi 2$ test), in accordo con lo studio di Yang e collaboratori (Yang et al. 2017). Tra queste regioni geniche è emersa un'ulteriore differenza significativa tra T e NT nel numero di siti integrazioni virali a livello degli introni, riscontrati in 830 dei 2.330 (~35,6%) *breakpoints* tumorali e in 142 dei 341 (~41,5%) *breakpoints* non-tumorali, con un *P*-value pari a 0,03 ($\chi 2$ test). Il numero delle integrazioni a livello

degli esoni e di regioni promotrici non ha mostrato differenze significative ($\chi 2$ test). Significativamente differenti sono, invece, risultate anche le integrazioni a livello intergenico (P<0,0001, $\chi 2$ test), rappresentando il ~38,0% del numero totale di integrazioni a livello tumorale (886/2.330) ed il ~27,7% (94/341) a livello non tumorale. Inoltre, il numero medio di integrazioni di HBV a livello dei lncRNA era maggiore nei T rispetto ai NT, in maniera statisticamente significativa (44,1±36,9 versus 9±7; Mann-Whitney test P=0,045). La distribuzione delle integrazioni relative a tali elementi genomici è mostrata in Figura 3.

Figura3. Integrazioni virali a livello dei differenti elementi genomici nei tessuti tumorali (rosso) e non tumorali (blu).

È stato, inoltre, osservato che l'80,1% ed il 58,9% e del totale delle integrazioni virali a livello tumorale e non tumorale, rispettivamente, erano localizzate in elementi ripetuti e/o trasponibili del genoma umano, con una importante differenza statistica (P<0,0001, $\chi 2$ test). Tra questi elementi, quelli in proporzione più abbondanti nei tessuti non tumorali rispetto ai tumorali erano le *Simple Repeat* (NT 94/341 vs T 190/2.330; $\chi 2$ P<0,0001) e le ripetizioni semplici (*Low complexity*) (NT 6/341 vs T 9/2.330; $\chi 2$ P=0,0015). Invece, nei tessuti tumorali rispetto ai tessuti non-tumorali è stata osservata una differenza statisticamente

significativa per i seguenti elementi genomici: *Short Interspersed Nuclear Elements* (SINE) (T 310/2.330 vs NT 23/341; *P*=0,0006, χ2 test); DNA Satellite (T 333/2.330 vs NT 10/341; *P*<0,0001, χ2 test); *Long Terminal Repeats* (LTR) (T 292/2.330 vs NT 26/341; *P*=0,009, χ2 test) e regioni centromeriche (T 331/2.330 vs NT 2/341; χ2 test, *P*<0,0001). Il numero di integrazioni a livello degli elementi LINE (*Long Interspersed Nuclear Elements*) è risultato pari al ~11% del totale (298/2.671), ma non sono state rilevate differenze statisticamente significative fra T e NT (T 264/2.330; NT 34/341; *proportion test P*=0,94). I risultati relativi alle integrazioni virali negli elementi ripetuti e/o trasponibili del genoma, nei tessuti tumorali e non tumorali sono riportati in Figura 4.

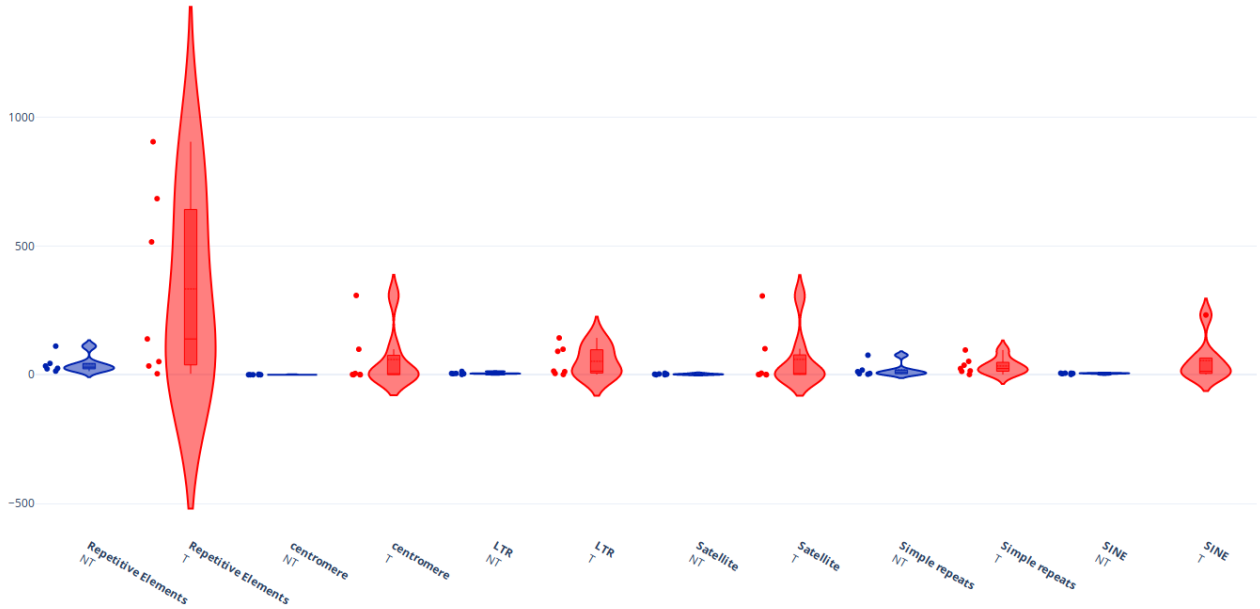

Figura4. Integrazioni virali a livello dei differenti elementi ripetuti/trasponibili in tessuti tumorali (rosso) e non tumorali (blu).

La valutazione delle regioni virali principalmente coinvolte negli eventi di integrazione ha rivelato che le integrazioni contenenti porzioni del gene *S* rappresentavano circa il 52% (178/341) del totale degli eventi osservati nei tessuti non tumorali, e circa il 24% (568/2.330; χ2 test *P*<0,0001) degli eventi osservati nei tessuti tumorali. Anche la regione immunodominante "A-determinant" del gene *S* è risultata arricchita nei tessuti non tumorali (27,5%) rispetto ai tumorali (~7%) (T 166/2.330 vs NT 94/341; *P*<0,0001, χ2 test). Di contro la regione virale

principalmente integrata è risultata essere quella del gene *X*, con 1.401/2.330 (~60%) eventi osservati nei T rispetto ai 134/341 (39%) eventi nei NT (P<0,0001, $\chi 2$ test). Anche la regione del *PreCore/Core* è risultata significativamente più arricchita nei tessuti tumorali (488/2.330; ~21%) rispetto ai non tumorali (28/341; ~8%) (P<0,0001, $\chi 2$ test). Le regioni virali coinvolte nei siti di integrazione, con il relativo *coverage*, sono mostrate in Figura 5.

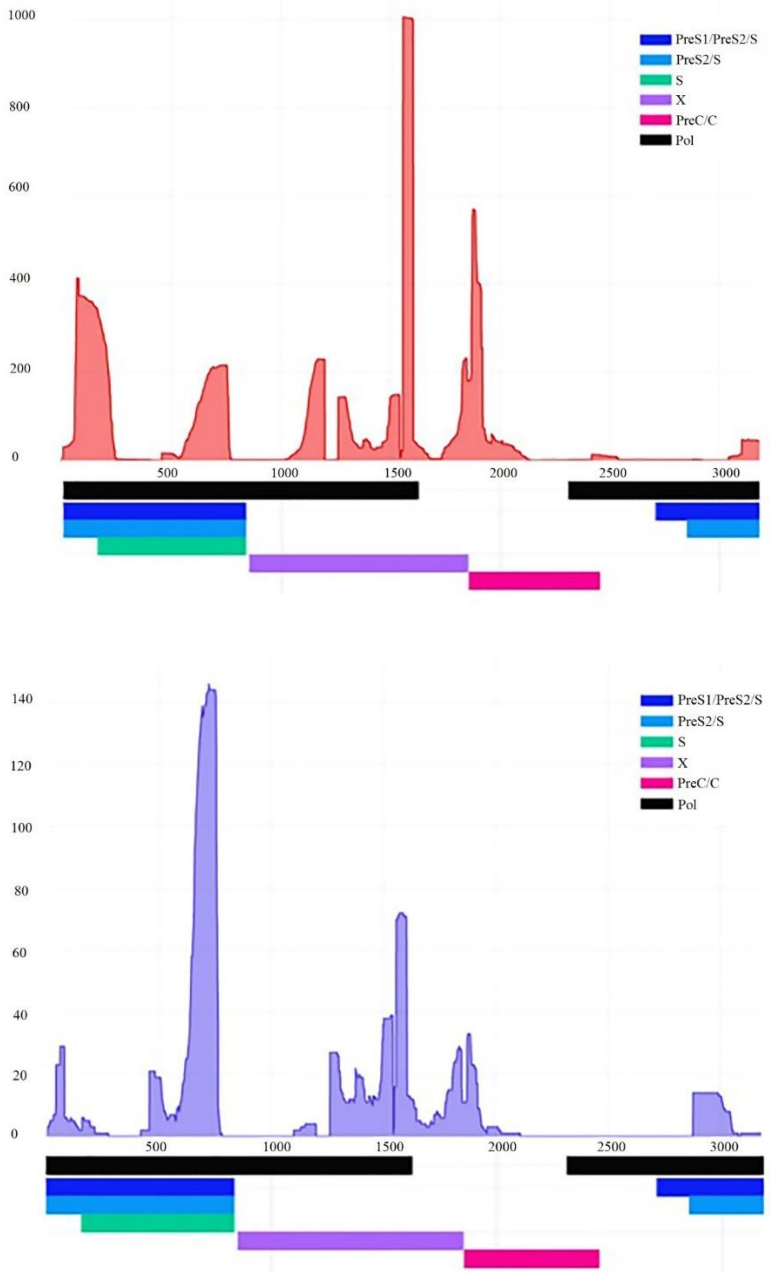

Figura 5. *Coverage* delle *reads* coinvolte negli eventi di integrazione lungo il genoma di HBV identificate nei tessuti tumorali (alto) e non tumorali (basso).

È stato anche valutato se la micro-omologia (MH) tra le estremità della sequenza genomica umana e le estremità delle sequenze virali a livello del *breakpoint* d'integrazione potesse aver giocato un ruolo nel favorire l'evento d'integrazione virale. A tal fine è stato confrontato il numero di eventi di integrazione che presentassero MH rispetto al totale degli eventi, ed è stato applicato il test del $\chi 2$ per rilevare differenze statisticamente significative tra tessuti tumorali e non tumorali. Le diverse dimensioni, in termini di basi pari (bp), delle MH prese in considerazione sono state: 3bp, 5bp, 7bp, 9bp, 11bp, 13bp, 15bp (Tabella 2).

Tabella 2. Abbondanza di siti in cui era presente MH tra genoma umano e virale nel sito di integrazione. * indica la non significatività statistica.

	Non Tumori			Tumori		
	Abbondanza	% sul totale	*P*-value	Abbondanza	% sul totale	*P*-value
MH ≥3bp	258	75,6	<0,0001	1.895	81,3	<0,0001
MH ≥5bp	236	69,2	<0,0001	1.734	74,4	<0,0001
MH ≥7bp	210	61,5	<0,0001	1.544	66,2	<0,0001
MH ≥9bp	166	48,6	<0,0001	1.066	45,7	<0,0001
MH ≥11bp	109	31,9	<0,0001	667	28,6	<0,0001
MH ≥13bp	84	24,6	0,024	514	22,0	<0,0001
MH ≥15bp	56	16,4	0,46*	393	16,8	0,0001

3.3 *Hot-spot* genomici e motivi genomici ricorrenti

Per poter analizzare la presenza di potenziali *hot-spot* genomici di integrazione, sono stati generati 3 differenti *dataset* randomici a partire dal totale delle 2.671 integrazioni identificate. In particolare, un *dataset* era costituito da 267.100 siti, un altro da 2.671.000 siti ed un altro ancora da 26.710.000 siti. Questi *dataset* sono stati utilizzati come *background* casuale di integrazione per poter effettuare il test di Monte Carlo. L'applicazione di tale test non ha identificato l'arricchimento di alcun sito nel *dataset* reale, rispetto a quelli simulati, come mostrato in Figura 6.

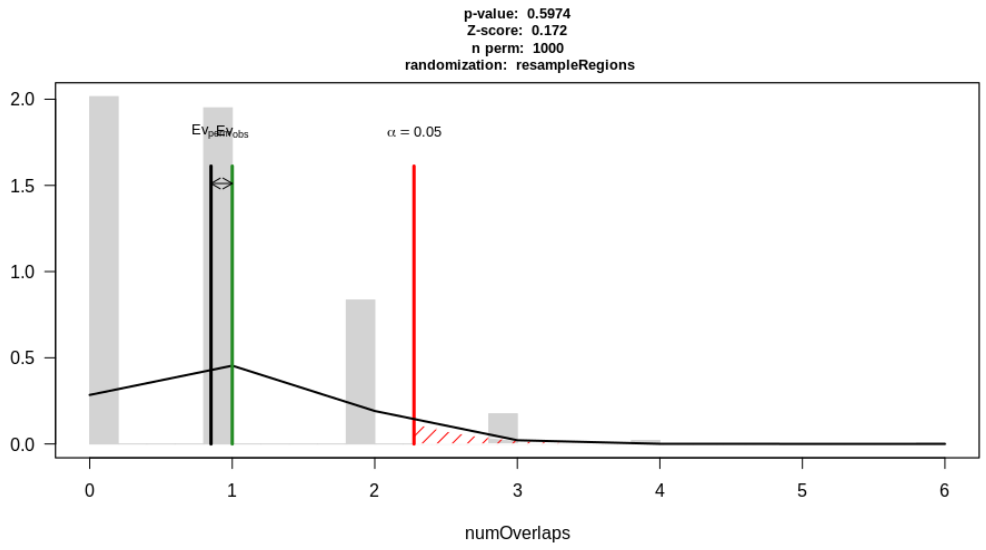

Figura 6. Assenza di significatività statistica fra i siti identificati ed un *dataset* 1000 volte più numeroso. Valutazione effettuata attraverso il test di Monte Carlo.

Applicando tuttavia un confronto fra integrazioni identificate in ciascun cromosoma e normalizzando per lunghezza cromosomica, è stato osservato un arricchimento di integrazioni virali a livello del cromosoma 20 (*P*=0,018, NPC test). Inoltre, si è osservato che tale differenza era dovuta ad un arricchimento di eventi di integrazione a livello centromerico presenti in numero statisticamente maggiore (*P*<0,0001, $\chi 2$ test) nei tessuti tumorali (331/2.330) rispetto a quelli non tumorali (2/341). In tutti gli altri cromosomi, sia i tessuti tumorali che quelli non tumorali presentavano una distribuzione di integrazione virale casuale ed omogeneamente distribuita. La distribuzione delle integrazioni per ciascun cromosoma normalizzate per lunghezza cromosomica sono mostrate in Figura 7.

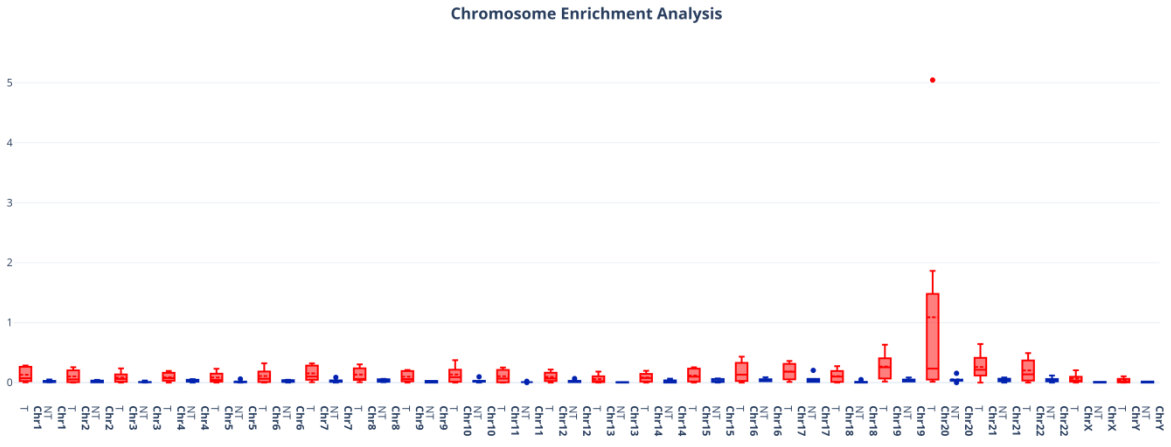

Figura 7. Distribuzione cromosomica delle integrazioni virali nei tessuti tumorali (rosso) e non tumorali (blu).

La ricerca di motivi ricorrenti nelle sequenze chimeriche, utilizzando come background il solo genoma umano, il solo genoma virale ed il genoma ibrido HG-HBV, non ha evidenziato nessuna sequenza sovra-rappresentata all'interno del totale delle sequenze chimeriche identificate e ricostruite (Figura 8).

* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1 *		1e0	-9.180e-01	50.26%	43.62%	77.1bp (5459637.7bp)
2 *		1e0	-9.628e-02	32.34%	43.00%	58.4bp (5217124.0bp)
3 *		1e0	-5.923e-03	0.03%	0.00%	0.0bp (0.0bp)
4 *		1e0	-2.000e-06	33.73%	81.42%	64.3bp (5088491.1bp)
5 *		1e0	0.000e+00	2.21%	98.48%	74.1bp (4524835.2bp)
6 *		1e0	0.000e+00	4.49%	85.56%	76.6bp (4162242.1bp)
7 *		1e0	0.000e+00	0.10%	85.56%	28.9bp (4418977.6bp)
8 *		1e0	0.000e+00	0.03%	85.56%	0.0bp (5093105.4bp)
9 *		1e0	0.000e+00	1.32%	85.56%	111.3bp (4300592.7bp)
10 *		1e0	0.000e+00	12.87%	85.56%	43.8bp (3781648.4bp)

Figura 8. Risultati della ricerca di motivi genomici ricorrenti effettuata mediante il software HOMER.

3.4 Clonalità dei siti di integrazione

Analizzando gli allineamenti delle *reads* - corrispondenti a ciascun sito di integrazione (identificato con la pipeline HBVIF) - a livello di ciascuna sequenza chimerica ricostruita, è stato possibile determinare la percentuale di eventi di integrazione clonale. Valutando quali integrazioni espanse clonalmente solo quelle supportate da *reads* con identica sequenza a cavallo di ciascun *breakpoint* di integrazione (almeno il 92%), è emerso che il 63,1% del totale degli eventi di integrazione (1.686/2.671) era clonale, di cui l'86,2% (1.454/2.107) appartenente

ai T ed il 13,8% (232/2,107) appartenete ai NT. Ciò ha evidenziato come i tessuti tumorali e quelli non-tumorali presentassero una differenza statisticamente significativa dell'espansione clonale dei *breakpoint* d'integrazione (*P*=0,02, *proportion test*). Tra le diverse *features* genomiche, soltanto le regioni intergeniche hanno mostrato differenze statisticamente significative tra T e NT (510/1.686, 35% vs 66/232, 28,2%; *P*=0,046, $\chi 2$ test), mentre tutte le altre regioni correlate ad elementi codificanti non risultavano arricchite né nel tessuto tumorale né in quello non tumorale (Figura 9).

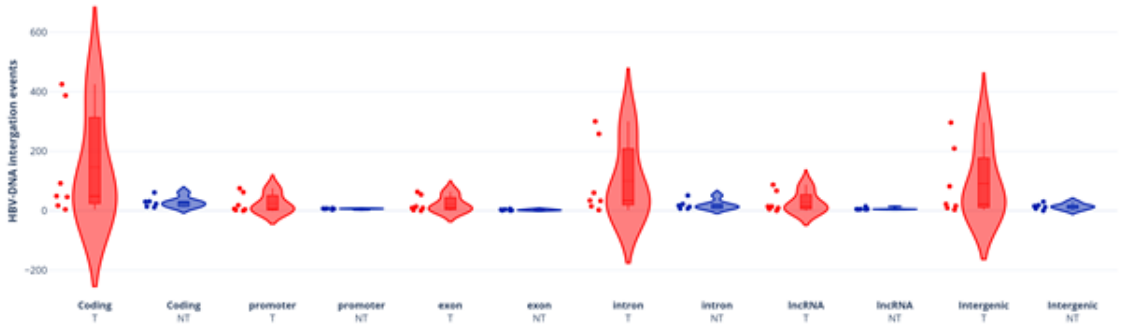

Figura 9. Distribuzione delle integrazioni virali a livello di elementi codificanti identificate nei tessuti tumorali (rosso) e non tumorali (blu).

Spostando, invece, l'attenzione sugli elementi ripetuti e/o trasponibili del genoma si è osservato un arricchimento di integrazioni clonali a livello di regioni *Low complexity* come le SINE, i LTR, il DNA satellite e gli elementi centromerici (*P*<0,0001, $\chi 2$ test) nei tessuti tumorali, ed un arricchimento a livello delle *Simple Repeats* nei tessuti non tumorali (*P*<0,0001, $\chi 2$ test), come mostrato in Figura 10.

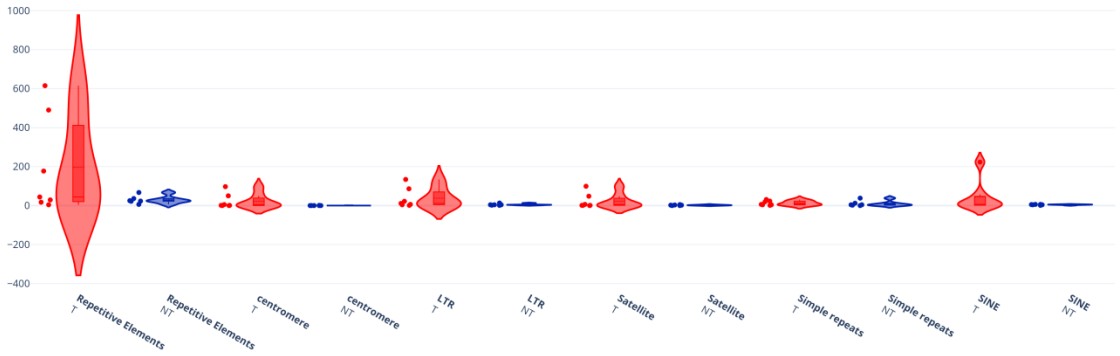

Figura 10. Distribuzione delle integrazioni virali a livello di elementi ripetuti e/o trasponibili, identificate nei tessuti tumorali (rosso) e non tumorali (blu).

Infine, le regioni virali principalmente coinvolte nei siti d'integrazione clonali, seppur ridotte in numero assoluto, hanno mantenuto le proporzioni osservate nel totale degli eventi d'integrazione identificati, mostrando quindi una preferenza di integrazioni di porzioni del gene *S* nei NT, rappresentati da ~50,7% (118/232), rispetto al ~25,3% (368/1.686) degli eventi tumorali (P<0,0001, $\chi 2$ test). La regione virale coinvolta nel circa 70% (1.018/1.686) delle integrazioni clonali rilevate nei tessuti tumorali, era quella del gene *X*, presente in ~40% (94/232) delle integrazioni clonali dei tessuti non-tumorali (P<0,0001, $\chi 2$ test). Sia nei T che nei NT, porzioni del gene *Core* sono state riscontrate in circa il 7% dei siti clonali. A differenza dell'arricchimento in integrazioni virali osservato a livello del cromosoma 20 nel totale degli eventi osservati, non è stata osservata alcuna differenza statisticamente significativa (NPC test) in termini di frequenza di integrazioni clonali nei diversi cromosomi (Figura 11).

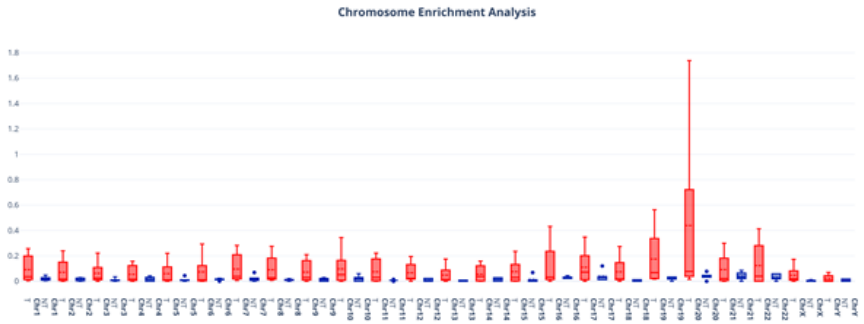

Figura 11. Distribuzione delle integrazioni virali in ciascun cromosoma umano, dopo normalizzazione per lunghezza cromosomica.

3.5 Risultati dei geni interessati da integrazioni virali condivisi da diversi campioni

Sono stati identificati un totale di 308 geni umani affetti da eventi di integrazione virale condivisi da almeno 2 campioni biologici. I geni maggiormente condivisi sono presentati in Tabella 3.

Tabella 3. Principali geni coinvolti nell'integrazione virale e condivisi da almeno 2 campioni biologici.

Gene	Occurences	Gene	Occurences	Gene	Occurences	Gene	Occurences
WNT3A	9	ABCB8	4	CCT6P3	3	PNOC	3
TECR	8	ACTR3C	4	CDC27P9	3	POR	3
WDR66	8	AOAH	4	CLK3	3	RNU6-221P	3
CPEB1-AS1	7	BRD9	4	DLGAP1	3	RORC	3
KRT14	7	C4orf50	4	EXOC4	3	RUNX1	3
LINC01237	7	CDC27P10	4	FSTL4	3	SEMA4C	3
MAL2	7	CDH4	4	H6PD	3	SORCS2	3
MED26	7	DPP6	4	HIVEP3	3	TCEAL7	3
NAPA-AS1	7	FCN2	4	IQSEC1	3	TMEM71	3
SLC22A7	7	GUS2	4	KCNQ3	3	TRAPPC9	3
CETP	6	HPCAL1	4	LOC105372880	3	TRNP	3
CNTFR	6	INPP5A	4	LOC105375594	3	UNC13C	3
IGH	6	JARID2	4	LOC105379506	3	UNK	3
MPG	6	KCNQ2	4	LOC105379508	3	ZC3H3	3
PFKL	6	LMF1	4	LOC107987293	3	ZNF536	3
ZHX2	6	LOC100287402	4	LOXL4	3	ABCG8	2
ADAMTS12	5	LOC105376791	4	LRFN2	3	ADCY5	2
BAHCC1	5	LOC105378653	4	LYPD8	3	AEBP1	2
C22orf34	5	LOC105379385	4	MACROD1	3	ANAPC15	2
KPNA7	5	MLPH	4	MYH9	3	ANKH	2
LINC01270	5	SEC1P	4	NCOR2	3	ANKS1B	2
LOC100128325	5	TNS1	4	NFAT5	3	ANPEP	2
LOC101929268	5	TUB	4	NOTCH1	3	ARHGAP39	2
LOC107987294	5	ZMIZ1	4	PCSK6	3	ARHGEF11	2
MROH5	5	ZNF664-FAM101A	4	PDE10A	3	ASB11	2
PTPRN2	5	ADGRB1	3	PHTF2	3	ATP11A	2
SPATS1	5	CAMTA1	3	PLD5	3	AUTS2	2

In Figura 12 sono mostrati il numero di geni condivisi da coppie di tessuti epatici.

	0102_NT	0103_NT	0104_NT	0105_NT	0106_NT	0101_T	0102_T	0103_T	0104_T	0105_T	0106_T	0107_T
0101_NT	0	0	0	0	0	1	0	1	0	0	0	0
0102_NT		5	21	1	9	20	17	42	12	15	38	1
0103_NT			5	2	2	4	7	10	5	1	1	1
0104_NT				3	7	20	18	43	41	10	25	0
0105_NT					1	2	0	1	0	1	1	0
0106_NT						8	9	9	3	10	15	1
0101_T							12	46	11	10	39	1
0102_T								25	20	13	20	1
0103_T									29	19	190	1
0104_T										6	15	1
0105_T											24	0
0106_T												1

Figura 12. Matrice rappresentante il numero di geni condivisi da ciascuna coppia di campioni

3.6 Analisi di espressione genica differenziale

Il sequenziamento dell'RNA estratto da 9 tessuti epatici, 3 tumorali (T), 3 non tumorali (NT) e 3 controllo (S), condotto su piattaforma Illumina HiSeq 2500, ha prodotto un totale di 755.120.686 *paired-end reads*, di cui 294.744.328 relative ai T, 219.038.892 relative ai NT e 241.337.466 relative ai S. Dopo la pulizia sono rimaste di buona qualità ~87% (256.457.700 relative ai T), ~92% (201.019.634 relative ai NT) e ~94% (228.227.474 relative ai S) delle *reads*. Tali *reads*, sono quindi state utilizzate per l'analisi di espressione differenziale. Dopo il *mapping* contro il genoma ibrido HG-HBV, utilizzando l'annotazione ibrida, è stata confermato il protocollo di sequenziamento *stranded* mediante lo script in python infer_experiment.py, come evidenziato dal risultato di seguito:

"This is PairEnd Data

Fraction of reads failed to determine: 0.0056

Fraction of reads explained by "1++,1--,2+-,2-+": 0.0726

Fraction of reads explained by "1+-,1-+,2++,2--": 0.9218"

Dai BAM file sono state calcolate le matrici di *counts* ed importate in R per effettuare l'analisi di espressione differenziale.

A seguito della normalizzazione TMM, è stata condotta la PCA, che ha mostrato una buona separazione, secondo la PC1 (~75%), dei campioni in funzione della condizione biologica di provenienza ed una discreta variabilità biologica fra componenti della stessa tipologia secondo la PC2 (~15%), come mostrato in Figura 13.

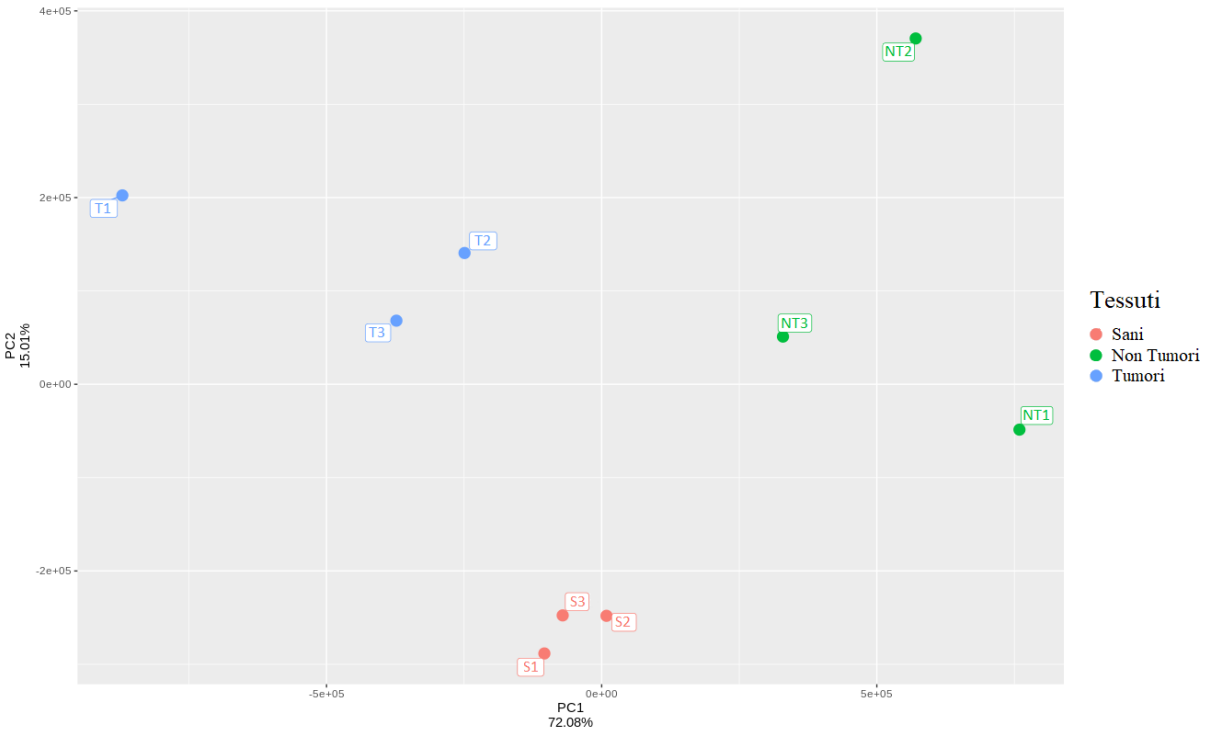

Figura 13. Principal Component Analysis (PCA) dei campioni di RNA-seq. Si noti la buona separazione degli stessi in funzione della PC1, supportata da una discreta variabilità biologica dei tessuti analizzati (PC2).

L'indice di Jaccard calcolato dal pacchetto HTSfilter (Rau et al. 2013) ha permesso di identificare, pari a ~39, il *coverage* minimo normalizzato di espressione genica in grado di massimizzare la similarità fra i tessuti biologici (T, NT e S) ascritti alla stessa condizione. Pertanto, tutti i geni il cui valore di espressione normalizzato era al di sotto di 39, sono stati scartati dalla successiva analisi. La curva relativa all'ottimizzazione dell'indice di Jaccard è mostrato in Figura 14.

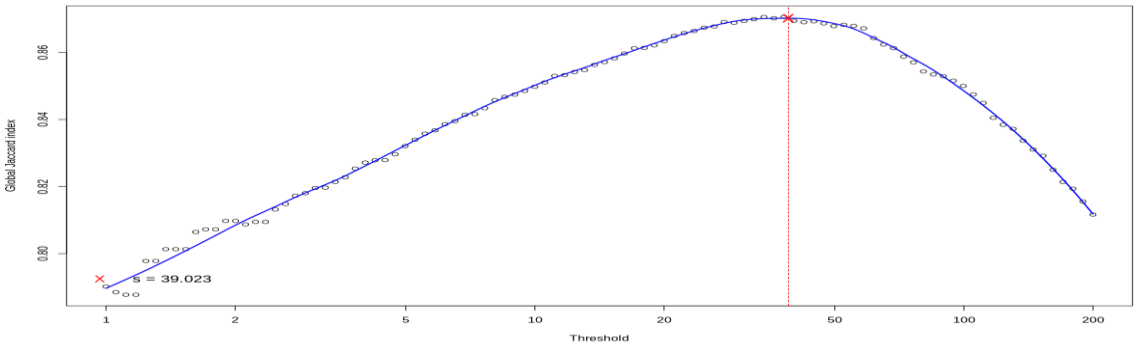

Figura 14. L'indice di Jaccard ha permesso di ottenere una similarità fra i campioni della stessa condizione del circa 86%.

L'analisi con edgeR ha permesso l'identificazione di 724 geni differenzialmente espressi nei tessuti tumorali rispetto ai tessuti controllo. Dei 724 geni, 340 sono risultati up-regolati e 384 down-regolati. In Figura 15, la rappresentazione grafica di geni differentemente espressi.

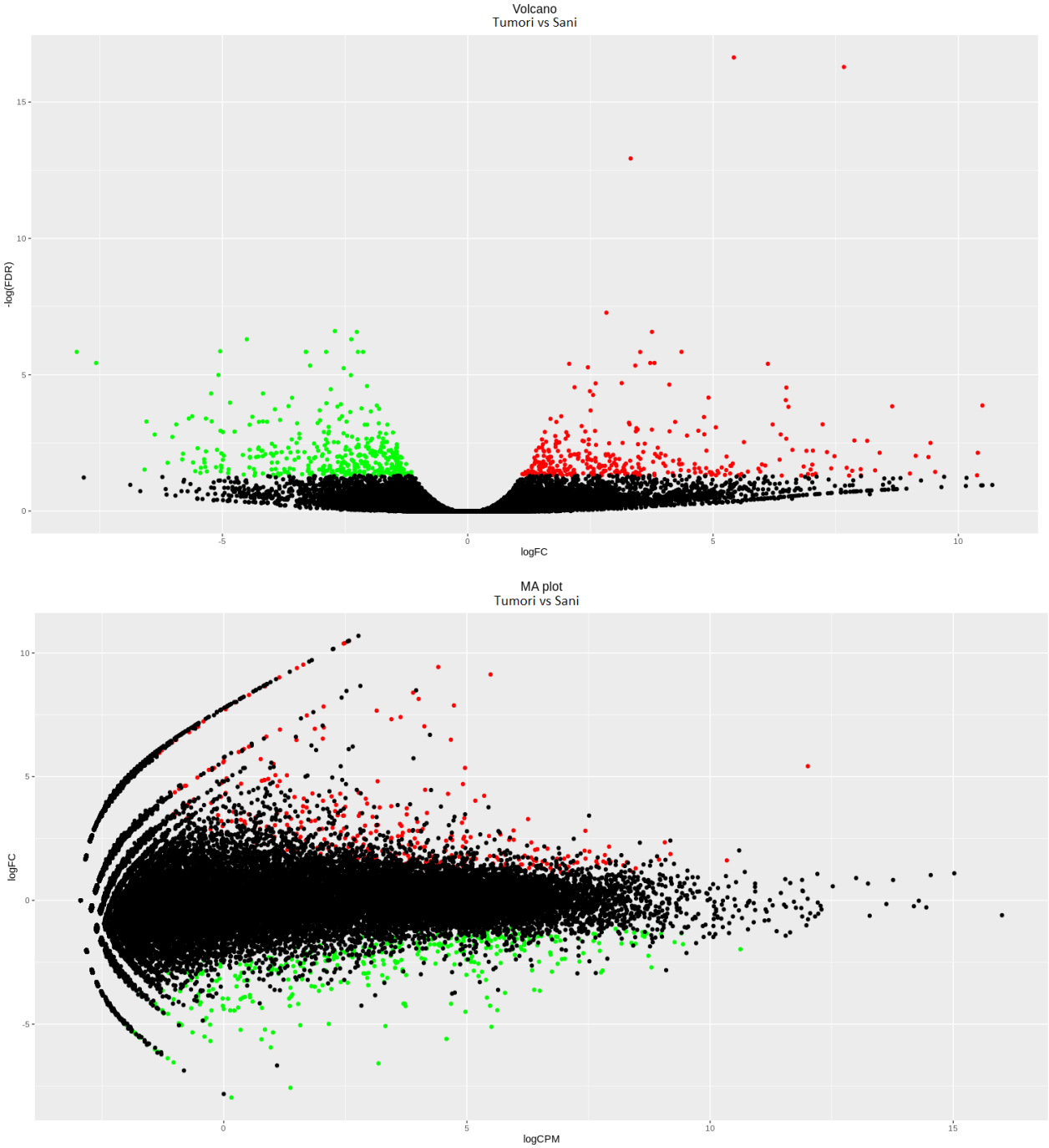

Figura 15. Volcano plot (in alto) e MA plot (in basso) rappresentanti i geni differenzialmente espressi, UP (rosso) e DOWN (verde) nei tessuti tumorali rispetto ai tessuti sani. Il Volcano plot mostra in ascisse il log2FC ed in ordinata il -log(FDR), mentre il MA plot mostra in ascissa il logCPM ed in ordinata il log2FC.

L'analisi per l'identificazione dei processi ontologici arricchiti nei geni differenzialmente espressi (GOEA) ha mostrato la presenza di svariati processi biologici arricchiti nel tumore rispetto ai tessuti sani, e viceversa, come mostrato nella Figura 16, che evidenzia a sinistra i processi biologici arricchiti nei tessuti tumorali, ed a destra quelli arricchiti nei tessuti sani. I geni up-regolati nei tessuti tumorali determinerebbero un arricchimento dei processi biologici coinvolti nella trascrizione, traduzione e maturazione proteica, nonché dei processi di trascrizione virale, di risposta a stress redox, coinvolti nel mantenimento della chemostasi, ed anche nell'attivazione dell'apoptosi e nell'aumento di trasportatori di membrana dello ione Ca^{++}. Di contro sono stati osservati differenti processi biologici arricchiti dalla down-regolazione genica, come la glicolisi, la regolazione negativa dei pathway di segnalazione correlati a TOR, processi redox, sintesi di colesterolo, etc.

Figura 16. GOEA relativa ai geni UP-regolati (sinistra) e DOWN-regolati (destra) nei tessuti tumorali rispetto a quelli sani.

L'analisi di espressione genica differenziale condotta tra i campioni non tumorali e quelli sani ha evidenziato la presenza di 101 geni espressi a livelli differenti e

statisticamente significativi, di cui 67 sovra-espressi e 34 sotto-espressi nei tessuti non tumorali rispetto ai sani (Figura 17).

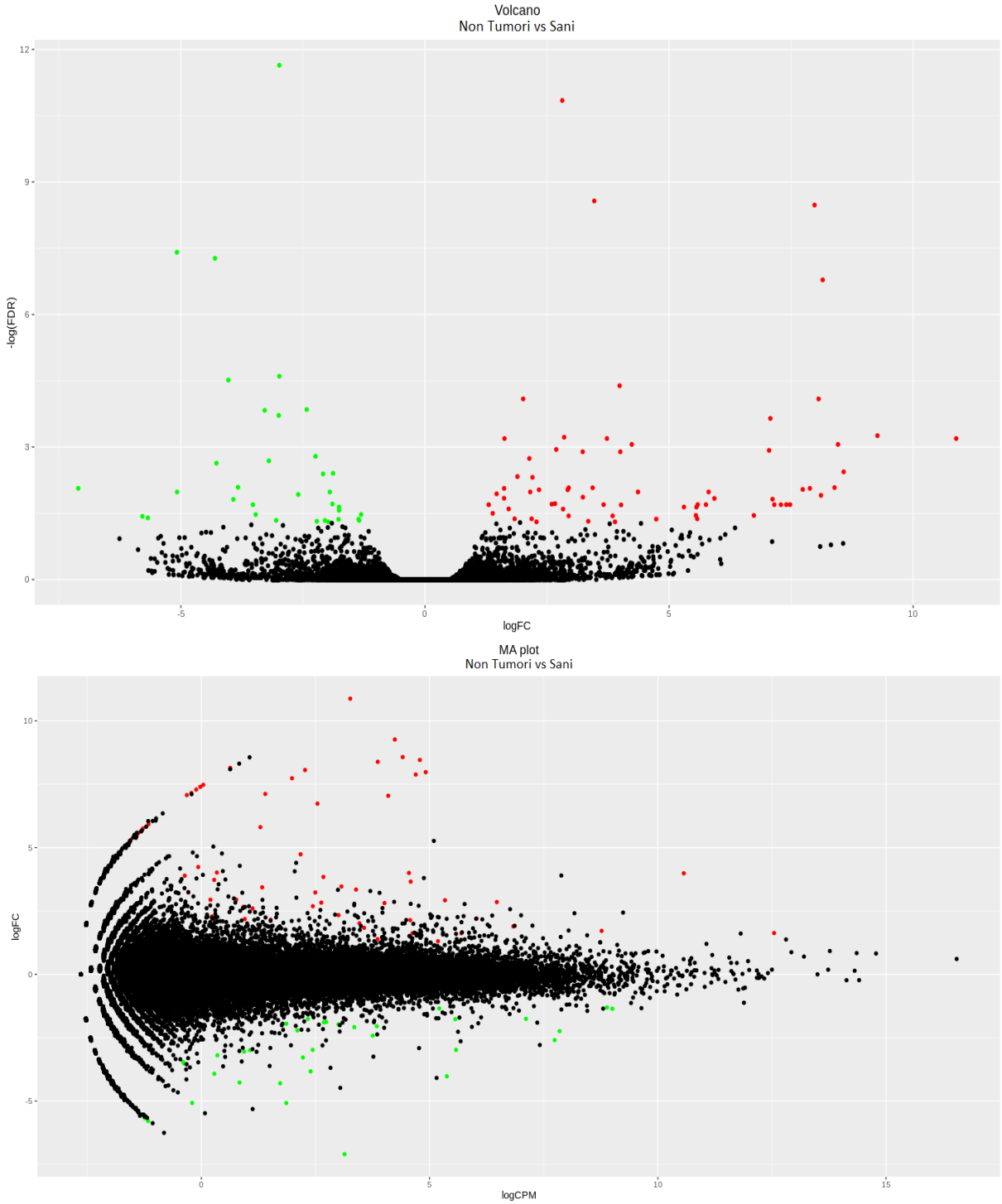

Figura 17. Volcano plot (sopra) e MA plot (sotto) rappresentanti i geni differenzialmente espressi nel confronto fra tessuti non tumorali e sani.

Nella ricerca di processi ontologici arricchiti, a partire dai geni UP-regolati nei tessuti non tumorali rispetto ai sani, sono risultati il pathway di segnalazione

dell'acido ɣ-amminobutirrico, delle vie di segnalazione di ormoni steroidei, di meccanismi trasporto di ioni attraverso la membrana cellulare, della regolazione negativa del pathway di segnalazione dell'insulina e della regolazione positiva della proliferazione dei fibroblasti.

Di contro sono stati osservati diversi processi biologici impoveriti nei tessuti non tumorali rispetto a quelli sani, tra cui i principali sono i pathway dei recettori accoppiati alle proteine-G, del metabolismo (ossidazione e beta-ossidazione) degli acidi grassi, dei processi metabolici che coinvolgono l'S-adenosil metionina, e regolazione positiva del trasporto di ioni Ca^{++} a livello citosolico. Le categorie ontologiche arricchite sono mostrate in Figura 18.

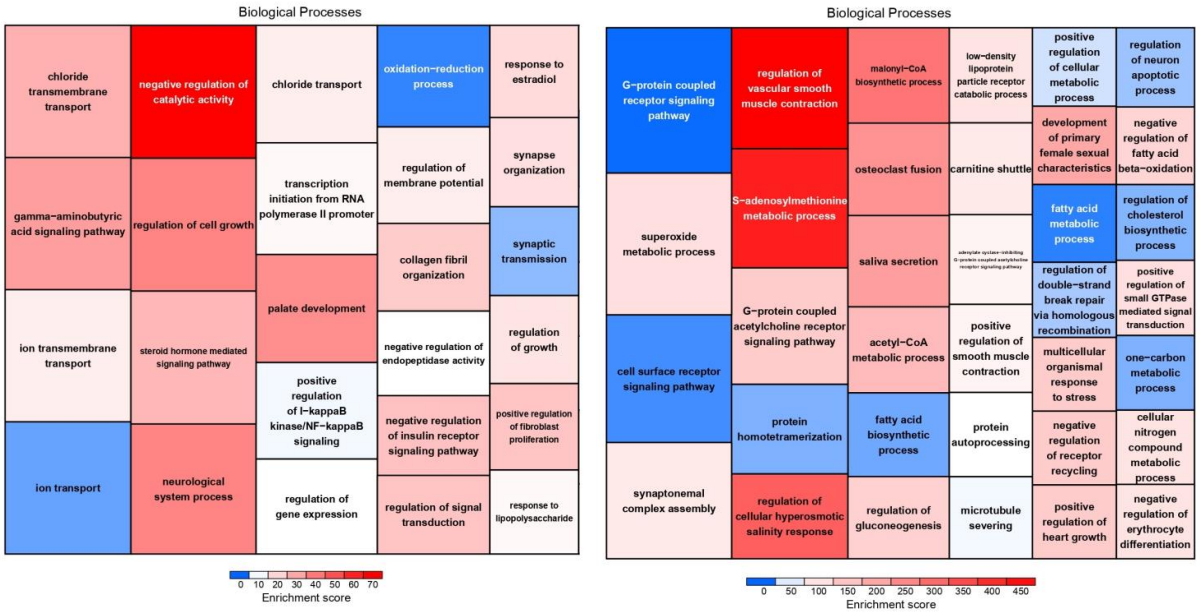

Figura 18. GOEA relativa ai geni UP-regolati (sinistra) e DOWN-regolati (destra) nei tessuti non tumorali rispetto a quelli sani.

Non sono stati identificati geni differenzialmente espressi nel confronto fra tessuti tumorali e non tumorali.

4 Discussione

L'infezione cronica da virus dell'epatite B (HBV) è uno dei principali fattori di rischio del carcinoma epatocellulare (HCC), che rappresenta la quarta causa di morte correlata al cancro a livello globale (Bray et al. 2018). L'integrazione virale nel genoma degli epatociti infettati è stata osservata in oltre il 90% dei casi di HCC correlati all'infezione da HBV; ciò supporta fortemente il coinvolgimento del virus nel processo di carcinogenesi epatica (Brechot et al. 1980, Edman et al. 1980, Sezaki et al. 2004). Vi sono, infatti, numerosissime evidenze che dimostrano che l'integrazione di HBV possa portare sia all'alterazione dell'espressione di importanti geni cellulari che alla produzione di proteine virali (HBx ed HBs) mutate dotate di proprietà oncogeniche (Levrero and Zucman-Rossi 2016, Tu et al. 2017). Resta, tuttavia, ancora del tutto sconosciuto il meccanismo attraverso cui il virus sia in grado di integrarsi nel genoma umano, ed è ancora dibattuto se ad essere coinvolti nel processo di integrazione siano i meccanismi non mediati da omologia di sequenza (NHEJ) oppure quelli mediati da micro-omologia tra sequenza virale e target umano (MMEJ) (Zhao et al. 2016, Tu et al. 2017). Ad oggi sono stati individuati oltre 20.000 siti di integrazione virale nel genoma umano, e ciò attraverso l'impiego di metodiche assai diverse in termini di sensibilità e specificità. Si è infatti passati dall'impiego del Southern blotting a metodiche basate sull'utilizzo della PCR (es. Alu-PCR, inverse-PCR) e del sequenziamento di Sanger, fino al più recente impiego di diversi approcci di NGS (Tang et al. 2019). Sino ad oggi, tuttavia, non è ancora stata messa a punto una metodica di sequenziamento che presenti elevata sensibilità ed accuratezza nell'identificazione e caratterizzazione dei *breakpoint* d'integrazione. In questo studio abbiamo sviluppato un nuovo approccio di sequenziamento HBV-specifico, ad alta produttività, basato sulla tecnologia NGS - HBV-ISeq - ed una pipeline bioinformatica *ad hoc* - HBVIF - per provare a rispondere a questa necessità.

Tutti i controlli negativi (linee cellulari HBV negative e tessuti epatici provenienti da pazienti HBsAg-negativi senza malattia di fegato) sono risultati privi di integrazioni virali, confermando la specificità della metodica HBV-ISeq. Nel controllo positivo, la linea cellulare PLC/PRF/5, sono state identificate 10 integrazioni virali, di cui quelle riscontrate a livello del promotore del gene *TERT* o a livello dei geni *MVK*, *CCDC57* e *UNC5D* erano state già riportate in letteratura (Graef et al. 1994, Watanabe et al. 2015, Zhao et al. 2016, Ishii et al. 2020). Le integrazioni a livello del gene *CCDC57* e del promotore di *TERT* sono state confermate attraverso il sequenziamento di Sanger, evidenziando la precisione e sensibilità sia della metodica di sequenziamento che dell'approccio bioinformatico. Le rimanenti integrazioni interessavano un lncRNA, *LOC105375660*, attualmente ancora non caratterizzato, ed uno pseudogene ribosomiale *RPS23P4*, la cui espressione – dalla valutazione del database GeneCards (https://www.genecards.org/) – non è stata riportata, ad oggi, in alcuno studio che abbia effettuato analisi trascrittomica su tessuto epatico. Le rimanenti integrazioni identificate in questa linea cellulare erano localizzate in regioni intergeniche.

L'utilizzo combinato dell'HBV-ISeq e dell'HBVIF, identificando oltre 2.500 eventi di integrazione virale in 13 tessuti epatici (7 T e 6 NT) da pazienti HBsAg-positivi ha rivelato l'altissima sensibilità delle due metodiche. Inoltre, nei tessuti tumorali è stato rilevato un numero di siti di integrazione (pari a 2.330) significativamente più alto che nei tessuti non tumorali (pari a 341). Tale dato conferma i risultati di un precedente studio (Zhao et al. 2016). Le regioni genomiche di HBV riscontrate con maggiore frequenza negli eventi di integrazione sono state: la regione genomica dell'*X* - incluse le regioni regolatrici ENHI/X promoter ed ENHII/Basal Core Promoter - e la regione genomica dell'*S*. Tali dati sono in accordo con quanto già riportato in letteratura (Sung et al. 2012, Zhao et al. 2016, Ruan et al. 2019, Podlaha et al. 2019). Inoltre, è stato riscontrato che, tra gli elementi del genoma umano, le regioni codificanti rappresentano il principale target degli eventi d'integrazione virale, essendo le regioni geniche

coinvolte, rispettivamente, nel 59,8% e nel 53,4% del totale degli eventi osservati nei tessuti tumorali e non tumorali. In particolare, nei tessuti non tumorali è stato rilevato un maggior interessamento delle regioni introniche. Di contro, nei tessuti tumorali, è stato osservato un maggior numero di eventi di integrazione nelle regioni genomiche intergeniche, e ciò in accordo con quanto già riportato in studi precedenti (Li et al. 2014, Zhao et al. 2016). Numerose integrazioni sono state rilevate, anche, a livello degli elementi trasponibili e/o ripetuti del genoma umano, con preferenza per gli elementi ripetuti a bassa complessità e *Simple Repeats* nei tessuti non tumorali. Nei tessuti tumorali, invece, sono le *SINE*, gli *LTR* ed i DNA *elements* a risultare target frequenti di integrazione. Bersaglio di integrazione virale sono risultate anche le regioni *LINE*, ma senza che vi fosse una differenza statisticamente significativa tra tessuti tumorali e non tumorali. Dai dati ottenuti è, anche, emerso che i meccanismi di ricombinazione basata sulla micro-omologia (MMEJ), che nel nostro studio ha rappresentato oltre il 75% del totale dei siti di integrazione sia nei T che nei NT, possano rappresentare un importante step del processo di integrazione virale. Questi dati ci portano ad ipotizzare come le integrazioni in siti fragili del genoma umano possano determinare instabilità cromosomica e possano fungere da meccanismi che conferiscono alle cellule un vantaggio proliferativo, sì da espandersi clonalmente ed andare incontro a trasformazione neoplastica. In questo studio è stata, infatti, osservata un'espansione clonale in circa il 63% delle integrazioni, il che porta ad ipotizzare che non tutte le integrazioni clonali possano essere frutto di espansione selettiva, ma piuttosto il risultato di un'espansione clonale "benigna", e che solo pochi cloni - in particolare, quelli le cui cellule contengono integrazioni virali in geni "driver" - possano andare incontro a trasformazione neoplastica. L'assenza di motivi ricorrenti all'interno delle sequenze chimeriche, identificate in questo studio, ben supporta il dato ottenuto dell'assenza di *hot-spot* e della distribuzione casuale ed omogenea dell'integrazione virale nel genoma umano. Rispetto a precedenti studi, l'applicazione dell'HBV-ISeq e dell'HBVIF ha consentito di rilevare che un altissimo numero di geni è target di integrazione virale nel corso d'infezione da

HBV. Tali dati sono stati peraltro validati dai dati ottenuti dall'analisi trascrittomica, che ha evidenziato un'alterazione dell'espressione di numerosi geni e di intere vie metaboliche, sia nei tessuti tumorali che non tumorali rispetto ai tessuti sani. Sono stati, infatti, identificati oltre 700 geni differenzialmente espressi fra i tessuti tumorali e quelli sani, con significativo arricchimento dei processi di trascrizione virale, proprio nei tessuti in cui il virus è risultato in fase di attiva replicazione e trascrizione genica. I dati ottenuti attraverso la *Gene Ontology Enrichment Analysis* (GOEA) hanno, inoltre, dimostrato che i tessuti tumorali presentano un arricchimento generale dei processi trascrizionali, traduzionali e post-traduzionali, e che ciò si associa alla presenza di infezione virale in fase di attiva replicazione.

La deregolazione dell'espressione del gene *WNT3A* è stata associata da numerosi studi all'HCC (Sakurai et al. 2017, Modica et al. 2019, Wang et al. 2014, Li uet al. 2016). In particolare l'alterata espressione di *WNT3A* correla con gli *enrichment* - osservati anche nel nostro studio - sia delle vie di trasporto dello ione Ca^{++} che del pathway di mTOR e dei pathway di stress cellulare, come quello ossidativo (Sakurai et al. 2017, Modica et al. 2019, Wang et al. 2014, Li uet al. 2016). É di notevole rilievo il ritrovamento di integrazione virale a livello del gene *WNT3A* in 9 dei 13 tessuti studiati. La GOEA effettuata sui circa 100 geni differenzialmente espressi nei tessuti non tumorali rispetto ai tessuti sani ha messo in evidenza l'*enrichment* del pathway dell'acido ɣ-amminobutirrico e dei pathway di segnalazione mediati da ormoni steroidei, entrambi coinvolti nell'omeostasi del tessuto epatico, consentendo di preservare l'integrità mitocondriale, e di incrementare la quantità di proteine antiossidanti ed anti-apoptotiche (Hata et al. 2019), e di conseguenza di prevenire lo sviluppo dell'HCC (Sukocheva 2018, Charni-Natan et al. 2019). In aggiunta, la GOEA ha evidenziato, nei tessuti non tumorali, una riduzione dei processi metabolici che coinvolgono l'S-adenosil metionina, generalmente associati allo stress ossidativo, all'attivazione della proliferazione cellulare ed allo sviluppo di HCC (Lu and Mato, 2012). La riduzione dei processi di beta-ossidazione degli acidi grassi osservata nei tessuti

non tumorali rispetto ai tessuti sani è in linea con le osservazioni di Nakagawa e collaboratori (Nakagawa et al. 2018), che hanno evidenziato come elevati livelli di acidi grassi ed alterazioni metaboliche correlano con lo sviluppo di HCC. In questo studio si è, inoltre, osservata una riduzione dei pathway associati alle proteine-G. La ridotta attivazione di questi pathway è stata correlata, anch'essa, con l'attivazione della proliferazione cellulare, con stress ossidativo e sviluppo di HCC, ed assieme ai sopramenzionati ed alterati processi biologici, comporta anche lo sviluppo e proliferazione dei fibroblasti (Li et al. 2016, Peng et al. 2018).

I dati ottenuti in questo studio hanno, quindi, messo in luce che - nell'HCC relato all'infezione da HBV - l'altissimo numero di eventi di integrazione virale è accompagnato dalla deregolazione di importanti geni e vie di segnalazione associati alla trasformazione cellulare. Di contro, nei tessuti non tumorali, si è osservato un *enrichment* in dei pathway coinvolti nell'omeostasi epatocitaria, nella risposta al danno ossidativo, indicando quali siano i processi biologici messi in atto per contrastare il danno e la trasformazione cellulare.

5 Conclusioni

Grazie allo sviluppo di una metodica di sequenziamento basata su NGS (HBV-ISeq) e di una specifica pipeline bioinformatica (HBVIF), È stato possibile identificare un altissimo numero di eventi di integrazione del DNA dell'HBV nel genoma umano. Inoltre, l'utilizzo dell'HBV-ISeq/HBVIF in combinazione con l'RNA-seq ha consentito evidenziare eventi molecolari e genetici che possono giocare un ruolo potenzialmente importante nello sviluppo dell'HCC associato all'infezione da HBV.

Bibliografia

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bayard Q, Meunier L, Peneau C, Renault V, Shinde J, Nault JC, Mami I, Couchy G, Amaddeo G, Tubacher E, Bacq D, Meyer V, La Bella T, Debaillon-Vesque A, Bioulac-Sage P, Seror O, Blanc JF, Calderaro J, Deleuze JF, Imbeaud S, Zucman-Rossi J, Letouzé E. Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress. Nat Commun. 2018 Dec 7;9(1):5235. doi: 10.1038/s41467-018-07552-9. PMID: 30531861; PMCID: PMC6286353.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

Bonilla Guerrero, R. and L. R. Roberts (2005). "The role of hepatitis B virus integrations in the pathogenesis of human hepatocellular carcinoma." J Hepatol42(5): 760-777.

Brechot, C.; Pourcel, C.; Louise, A.; Rain, B.; Tiollais, P. Presence of integrated hepatitis B virus DNA sequences in cellular DNA of human hepatocellular carcinoma. Nature 1980, 286, 533–535.

Bray, F., J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal (2018). "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA Cancer J Clin 68(6): 394-424.

Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. Published 2009 Dec 15. doi:10.1186/1471-2105-10-421.

Charni-Natan, M., Aloni-Grinstein, R., Osher, E., & Rotter, V. (2019). Liver and Steroid Hormones-Can a Touch of p53 Make a Difference?. Frontiers in endocrinology, 10, 374. https://doi.org/10.3389/fendo.2019.00374.

Cohn, L. B., I. T. Silva, T. Y. Oliveira, R. A. Rosales, E. H. Parrish, G. H. Learn, B. H. Hahn, J. L. Czartoski, M. J. McElrath, C. Lehmann, F. Klein, M. Caskey, B. D. Walker, J. D. Siliciano, R. F. Siliciano, M. Jankovic and M. C. Nussenzweig (2015). "HIV-1 integration landscape during latent and active infection." Cell160(3): 420-432

Ding, D., X. Lou, D. Hua, W. Yu, L. Li, J. Wang, F. Gao, N. Zhao, G. Ren, L. Li and B. Lin (2012). "Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach." PLoS Genet8(12): e1003065.

Ding, X. X., Zhu, Q. G., Zhang, S. M., Guan, L., Li, T., Zhang, L., Wang, S. Y., Ren, W. L., Chen, X. M., Zhao, J., Lin, S., Liu, Z. Z., Bai, Y. X., He, B., & Zhang, H. Q. (2017). Precision medicine for hepatocellular carcinoma: driver mutations and targeted therapy. Oncotarget, 8(33), 55715–55730. https://doi.org/10.18632/oncotarget.18382.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

Edman, J.C.; Gray, P.; Valenzuela, P.; Rall, L.B.; Rutter,W.J. Integration of hepatitis B virus sequences and their expression in a human hepatoma cell. Nature 1980, 286, 535–538.

El-Serag, H. B. (2012). "Epidemiology of viral hepatitis and hepatocellular carcinoma." Gastroenterology 142(6): 1264-1273 e1261.

Fujimoto, A., M. Furuta, Y. Totoki, T. Tsunoda, M. Kato, Y. Shiraishi, H. Tanaka, H. Taniguchi, Y. Kawakami, M. Ueno, K. Gotoh, S. Ariizumi, C. P. Wardell, S.

Hayami, T. Nakamura, H. Aikata, K. Arihiro, K. A. Boroevich, T. Abe, K. Nakano, K. Maejima, A. Sasaki-Oku, A. Ohsawa, T. Shibuya, H. Nakamura, N. Hama, F. Hosoda, Y. Arai, S. Ohashi, T. Urushidate, G. Nagae, S. Yamamoto, H. Ueda, K. Tatsuno, H. Ojima, N. Hiraoka, T. Okusaka, M. Kubo, S. Marubashi, T. Yamada, S. Hirano, M. Yamamoto, H. Ohdan, K. Shimada, O. Ishikawa, H. Yamaue, K. Chayama, S. Miyano, H. Aburatani, T. Shibata and H. Nakagawa (2016). "Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer." Nat Genet48(5): 500-509.

Fujimoto, A., Y. Totoki, T. Abe, K. A. Boroevich, F. Hosoda, H. H. Nguyen, M. Aoki, N. Hosono, M. Kubo, F. Miya, Y. Arai, H. Takahashi, T. Shirakihara, M. Nagasaki, T. Shibuya, K. Nakano, K. Watanabe-Makino, H. Tanaka, H. Nakamura, J. Kusuda, H. Ojima, K. Shimada, T. Okusaka, M. Ueno, Y. Shigekawa, Y. Kawakami, K. Arihiro, H. Ohdan, K. Gotoh, O. Ishikawa, S. Ariizumi, M. Yamamoto, T. Yamada, K. Chayama, T. Kosuge, H. Yamaue, N. Kamatani, S. Miyano, H. Nakagama, Y. Nakamura, T. Tsunoda, T. Shibata and H. Nakagawa (2012). "Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators." Nat Genet44(7): 760-764.

Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150–3152. doi:10.1093/bioinformatics/bts565.

Furuta M, Tanaka H, Shiraishi Y, Unida T, Imamura M, Fujimoto A, Fujita M, Sasaki-Oku A, Maejima K, Nakano K, Kawakami Y, Arihiro K, Aikata H, Ueno M, Hayami S, Ariizumi SI, Yamamoto M, Gotoh K, Ohdan H, Yamaue H, Miyano S, Chayama K, Nakagawa H. Characterization of HBV integration patterns and timing in liver cancer and HBV-infected livers. Oncotarget. 2018 May 18;9(38):25075-25088. doi: 10.18632/oncotarget.25308. Erratum in: Oncotarget. 2018 Aug 3;9(60):31789. PMID: 29861854; PMCID: PMC5982772.

Graef E, Caselmann WH, Wells J, Koshy R. Insertional activation of mevalonate kinase by hepatitis B virus DNA in a human hepatoma cell line. Oncogene. 1994 Jan;9(1):81-7. PMID: 8302606.

Hai, H., Tamori, A., & Kawada, N. (2014). Role of hepatitis B virus DNA integration in human hepatocarcinogenesis. *World journal of gastroenterology*, *20*(20), 6236–6243. https://doi.org/10.3748/wjg.v20.i20.6236.

Hata, T., Rehman, F., Hori, T., & Nguyen, J. H. (2019). GABA, γ-Aminobutyric Acid, Protects Against Severe Liver Injury. The Journal of surgical research, 236, 172–183. https://doi.org/10.1016/j.jss.2018.11.047.

Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res*. 1999;9(9):868–877. doi:10.1101/gr.9.9.868.

Ishii, T., Tamura, A., Shibata, T., Kuroda, K., Kanda, T., Sugiyama, M., Mizokami, M., & Moriyama, M. (2020). Analysis of HBV Genomes Integrated into the Genomes of Human Hepatoma PLC/PRF/5 Cells by HBV Sequence Capture-Based Next-Generation Sequencing. *Genes*, *11*(6), 661. https://doi.org/10.3390/genes11060661.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *MolBiolEvol*. 2013;30(4):772–780. doi:10.1093/molbev/mst010.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109.

Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*. 2014;30(22):3276-3278. doi:10.1093/bioinformatics/btu531.

Lau, C. C., T. Sun, A. K. Ching, M. He, J. W. Li, A. M. Wong, N. N. Co, A. W. Chan, P. S. Li, R. W. Lung, J. H. Tong, P. B. Lai, H. L. Chan, K. F. To, T. F. Chan

and N. Wong (2014). "Viral-human chimeric transcript predisposes risk to liver cancer development and progression." Cancer Cell25(3): 335-349.

Levrero, M. and J. Zucman-Rossi (2016). "Mechanisms of HBV-induced hepatocellular carcinoma." J Hepatol 64(1 Suppl): S84-S101.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools.*Bioinformatics*. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Li, X., Zhang, J., Yang, Z., Kang, J., Jiang, S., Zhang, T., Chen, T., Li, M., Lv, Q., Chen, X., McCrae, M. A., Zhuang, H., & Lu, F. (2014). The function of targeted host genes determines the oncogenicity of HBV integration in hepatocellular carcinoma. Journal of hepatology, 60(5), 975–984. https://doi.org/10.1016/j.jhep.2013.12.014.

Li, Y., Zhang, W., Doughtie, A., Cui, G., Li, X., Pandit, H., Yang, Y., Li, S., & Martin, R. (2016). Up-regulation of fibroblast growth factor 19 and its receptor associates with progression from fatty liver to hepatocellular carcinoma. Oncotarget, 7(32), 52329–52339. https://doi.org/10.18632/oncotarget.10750.

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13. PMID: 24227677.

Liu LJ, Xie SX, Chen YT, Xue JL, Zhang CJ, Zhu F. Aberrant regulation of Wnt signaling in hepatocellular carcinoma. *World J Gastroenterol*. 2016;22(33):7486-7499. doi:10.3748/wjg.v22.i33.7486.

Lu, Shelly C, and José M Mato. “S-adenosylmethionine in liver health, injury, and cancer.” Physiological reviews vol. 92,4 (2012): 1515-42. doi:10.1152/physrev.00047.2011.

Mladenov, E.; Magin, S.; Soni, A.; Iliakis, G. DNA double-strand-break repair in higher eukaryotes and its role in genomic instability and cancer: Cell cycle and proliferation-dependent regulation. Semin. Cancer Biol. 2016, 37–38, 51–64.

Modica, T., Dituri, F., Mancarella, S., Pisano, C., Fabregat, I., & Giannelli, G. (2019). Calcium Regulates HCC Proliferation as well as EGFR Recycling/Degradation and Could Be a New Therapeutic Target in HCC. *Cancers*, *11*(10), 1588. https://doi.org/10.3390/cancers11101588.

Nakagawa, H., Hayata, Y., Kawamura, S., Yamada, T., Fujiwara, N., & Koike, K. (2018). Lipid Metabolic Reprogramming in Hepatocellular Carcinoma. *Cancers*, *10*(11), 447. https://doi.org/10.3390/cancers10110447.

Pearson H., Stirling D. (2003) DNA Extraction from Tissue. In: Bartlett J.M.S., Stirling D. (eds) PCR Protocols. Methods in Molecular Biology™, vol 226. Humana Press. https://doi.org/10.1385/1-59259-384-4:33.

Peng, W. T., Sun, W. Y., Li, X. R., Sun, J. C., Du, J. J., & Wei, W. (2018). Emerging Roles of G Protein-Coupled Receptors in Hepatocellular Carcinoma. International journal of molecular sciences, 19(5), 1366. https://doi.org/10.3390/ijms19051366.

Podlaha O, Wu G, Downie B, Ramamurthy R, Gaggar A, Subramanian M, et al. (2019) Genomic modeling of hepatitis B virus integration frequency in the human genome. PLoS ONE 14(7): e0220376. https://doi.org/10.1371/journal.pone.0220376.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rau A, Gallopin M, Celeux G, Jaffrezic F (2013). “Data-based filtering for replicated high-throughput transcriptome sequencing experiments.” *Bioinformatics*, **29**(17), 2146-2152.

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.

Ruan, P., Dai, X., Sun, J., He, C., Huang, C., Zhou, R., Cao, Z., & Ye, L. (2019). Different types of viral-host junction found in HBV integration breakpoints in HBV-infected patients. Molecular medicine reports, 19(2), 1410–1416. https://doi.org/10.3892/mmr.2018.9709

Saigo, K.; Yoshida, K.; Ikeda, R.; Sakamoto, Y.; Murakami, Y.; Urashima, T.; Asano, T.; Kenmochi, T.; Inoue, I. Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (MLL4) gene and rearrangements of MLL4 in human hepatocellular carcinoma. Hum. Mutat. 2008, 29, 703–708.

Sakurai, Y., Kubota, N., Takamoto, I., Obata, A., Iwamoto, M., Hayashi, T., Aihara, M., Kubota, T., Nishihara, H., & Kadowaki, T. (2017). Role of insulin receptor substrates in the progression of hepatocellular carcinoma. *Scientific reports*, *7*(1), 5387. https://doi.org/10.1038/s41598-017-03299-3.

Schaefer S. Hepatitis B virus taxonomy and hepatitis B virus genotypes. World J Gastroenterol. 2007;13(1):14-21. doi:10.3748/wjg.v13.i1.14

Sezaki, H., M. Kobayashi, T. Hosaka, T. Someya, N. Akuta, F. Suzuki, A. Tsubota, Y. Suzuki, S. Saitoh, Y. Arase, K. Ikeda, M. Kobayashi, M. Matsuda, K. Takagi, J. Sato and H. Kumada (2004). "Hepatocellular carcinoma in noncirrhotic young adult patients with chronic hepatitis B viral infection." J Gastroenterol 39(6): 550-556.

Sawilowsky, S. S. (2003). You think you’ve got trivials? Journal of Modern Applied Statistical Methods, 2(1), 218-225.

Sukocheva O. A. (2018). Estrogen, estrogen receptors, and hepatocellular carcinoma: Are we there yet?. World journal of gastroenterology, 24(1), 1–4. https://doi.org/10.3748/wjg.v24.i1.1.

Sung, W. K., H. Zheng, S. Li, R. Chen, X. Liu, Y. Li, N. P. Lee, W. H. Lee, P. N. Ariyaratne, C. Tennakoon, F. H. Mulawadi, K. F. Wong, A. M. Liu, R. T. Poon, S. T. Fan, K. L. Chan, Z. Gong, Y. Hu, Z. Lin, G. Wang, Q. Zhang, T. D. Barber, W. C. Chou, A. Aggarwal, K. Hao, W. Zhou, C. Zhang, J. Hardwick, C. Buser, J. Xu, Z. Kan, H. Dai, M. Mao, C. Reinhard, J. Wang and J. M. Luk (2012). "Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma." Nat Genet 44(7): 765-769.

Deyou Tang, Bingrui Li, Tianyi Xu, Ruifeng Hu, Daqiang Tan, Xiaofeng Song, Peilin Jia, Zhongming Zhao, VISDB: a manually curated database of viral integration sites in the human genome, *Nucleic Acids Research*, Volume 48, Issue D1, 08 January 2020, Pages D633–D641, https://doi.org/10.1093/nar/gkz867.

Tatsuno, K., Midorikawa, Y., Takayama, T., Yamamoto, S., Nagae, G., Moriyama, M., Nakagawa, H., Koike, K., Moriya, K., & Aburatani, H. (2019). Impact of AAV2 and Hepatitis B Virus Integration Into Genome on Development of Hepatocellular Carcinoma in Patients with Prior Hepatitis B Virus Infection. Clinical cancer research : an official journal of the American Association for Cancer Research, 25(20), 6217–6227. https://doi.org/10.1158/1078-0432.CCR-18-4041.

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017 Jul 3;45(W1):W122-W129. doi: 10.1093/nar/gkx382. PMID: 28472432; PMCID: PMC5793732.

Tu, T.; Budzinska, M.A.; Shackel, N.A.; Urban, S. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. *Viruses* 2017, *9*, 75.

Liguo Wang, Shengqin Wang, Wei Li, RSeQC: quality control of RNA-seq experiments, *Bioinformatics*, Volume 28, Issue 16, 15 August 2012, Pages 2184–2185, https://doi.org/10.1093/bioinformatics/bts356.

Wang, Z., Jin, W., Jin, H., & Wang, X. (2014). mTOR in viral hepatitis and hepatocellular carcinoma: function and treatment. BioMed research international, 2014, 735672. https://doi.org/10.1155/2014/735672.

Watanabe, Y., Yamamoto, H., Oikawa, R., Toyota, M., Yamamoto, M., Kokudo, N., Tanaka, S., Arii, S., Yotsuyanagi, H., Koike, K., & Itoh, F. (2015). DNA methylation at hepatitis B viral integrants is associated with methylation at flanking human genomic sequences. *Genome research*, *25*(3), 328–337. https://doi.org/10.1101/gr.175240.114

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, http://ggplot2.org.

Yang, X., Wu, L., Lin, J., Wang, A., Wan, X., Wu, Y., Robson, S.C., Sang, X. and Zhao, H. (2017), Distinct hepatitis B virus integration patterns in hepatocellular carcinoma and adjacent normal liver tissue. Int. J. Cancer, 140: 1324-1330. doi:10.1002/ijc.30547.

Zhang, H. Wu, S. Huang, M. D. Wang, L. Tang, H. Z. Cao, L. Wang, T. L. Lee, H. Jiang, Y. X. Tan, S. X. Yuan, G. J. Hou, Q. F. Tao, Q. G. Xu, X. Q. Zhang, M. C. Wu, X. Xu, J. Wang, H. M. Yang, W. P. Zhou and H. Y. Wang (2016). "Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma." Nat Commun 7: 12992.

Zhao, L. H., X. Liu, H. X. Yan, W. Y. Li, X. Zeng, Y. Yang, J. Zhao, S. P. Liu, X. H. Zhuang, C. Lin, C. J. Qin, Y. Zhao, Z. Y. Pan, G. Huang, H. Liu, J. Zhang, R. Y. Wang, Y. Yang, W. Wen, G. S. Lv, H. L. Zhang, H. Wu, S. Huang, M. D. Wang, L. Tang, H. Z. Cao, L. Wang, T. L. Lee, H. Jiang, Y. X. Tan, S. X. Yuan, G. J. Hou, Q. F. Tao, Q. G. Xu, X. Q. Zhang, M. C. Wu, X. Xu, J. Wang, H. M. Yang, W. P. Zhou and H. Y. Wang (2016). "Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma." Nat Commun 7: 12992.