UNIVERSITÀ DEGLI STUDI DI CATANIA

DIPARTIMENTO DI MEDICINA CLINICA E SPERIMENTALE

DOTTORATO DI RICERCA IN BIOMEDICINA TRASLAZIONALE:
XXXII Ciclo

COORDINATORE: PROFESSORE LORENZO MALATINO

Dott. Gioacchino Paolo Marceca

Tesi di dottorato

IMPLEMENTAZIONE DI DUE METODI PATHWAY TOPOLOGY-BASED PER L'ANALISI DELLA PERTURBAZIONE DI PATHWAY BIOLOGICI

Tutor
Prof.re Alfrdo Ferro

Co-Tutor
Dott. Salvatore Alaimo

Anno Accademico 2016-2017

Sommario

Riassunto

L'applicazione in campo biomedico delle odierne tecnologie di sequenziamento e di rilevamento massivo, in congiunzione all'elaborazione bioinformatica dei dati da esse prodotti, ha permesso l'identificazione di diversi biomarcatori diagnostici e prognostici, nonché la ricostruzione ad elevata risoluzione di pathway biologici e network di interazioni molecolari, le cui anomalie sottendono diverse malattie e disturbi fisiologici. In questo contesto, la biologia dei sistemi si pone come mezzo di facilitazione per la ricerca biomedica di base, aprendo nuove strade allo studio delle patologie umane e fornendo, allo stesso tempo, un valido supporto al disegno di trattamenti personalizzati, alla scoperta di nuovi farmaci e al riposizionamento dei farmaci già esistenti. In linea generale, ciò fa tendere il pensiero della comunità scientifica verso una nuova branca della medicina, definita come medicina dei sistemi, che offre al medico nuovi approcci multidisciplinari e fortemente basati sulla modellazione computazionale al fine di ottenere diagnosi più accurate e trattamenti più efficaci in tempi più celeri. Tra i metodi analitici propri della biologia dei sistemi, la pathway analysis svolge un ruolo centrale nella definizione e nella predizione del fenotipo a partire da dati genomici, trascrittomici o proteomici. Nel corso degli ultimi quindici anni, una buona parte della comunità scientifica del settore ha concentrato i propri sforzi sullo sviluppo di diversi metodi analitici per pathway, giungendo infine all'elaborazione dei cosiddetti metodi di terza generazione, o anche pathway topology-based, più efficienti e accurati rispetto ai metodi precedentemente implementati. Nel presente elaborato vengono descritti due metodi per pathway analysis di terza generazione, in grado di identificare pathway o subpathway significativi per le patologie in questione, nonché di operare analisi di arricchimento dei termini ricavati, mostrando buoni livelli di accuratezza, grazie anche all'integrazione dei pathway di KEGG con informazioni relative a interazioni microRNA-target e induzione dell'espressione di specifici microRNA da parte di specifici fattori della trascrizione.

1. Introduzione

La mole pressoché incalcolabile di dati raccolti negli ultimi trent'anni in campo biomedico, le loro elaborazioni statistiche, le comunicazioni scientifiche che ne sono derivate e i molteplici sviluppi tecnici hanno totalmente rivoluzionato il pensiero e le prospettive dell'intera comunità scientifica, definendo i confini di una nuova epoca della medicina [1]. Unitamente all'accrescimento delle conoscenze riguardanti i meccanismi genetici ed epigenetici che sottendono l'insorgenza di svariate malattie, si è assistito all'affermarsi dell'idea di una "medicina personalizzata", in cui il paziente è posto al centro di una équipe multidisciplinare formata da medici che operano in prima linea con il supporto di figure quali biologi molecolari, biostatistici e bioinformatici. L'idea centrale è quella di stabilire terapie mirate che migliorino l'esito clinico del trattamento rispetto agli approcci tradizionali [2]. Intrinseco all'affermarsi della medicina personalizzata è lo sviluppo della medicina traslazionale, branca interdisciplinare del campo biomedico che ha lo scopo di velocizzare la scoperta di nuovi trattamenti e strumenti diagnostici [3]. Molte iniziative sono state promosse a livello governativo nell'ultimo quinquennio, specialmente nel contesto Europeo e Statunitense, che mirano a ulteriori sviluppi delle componenti di base della medicina personalizzata e traslazionale. Si tratta di ampi programmi di ricerca che finanziano la nascita di nuovi approcci nel campo della medicina di precisione, rigorosamente validati e utilizzabili per la costruzione di nuove linee guida per la pratica clinica. In particolare, l'iniziativa americana prevede e promuove il raggiungimento di due obiettivi principali: il primo, a breve-medio termine, ha il suo focus nel trattamento personalizzato dei tumori; il secondo, a più lungo termine, mira a generare conoscenze e tecniche applicabili all'intero spettro delle malattie conosciute, nonché al campo della prevenzione. Ciò è reso possibile grazie ai recenti progressi della ricerca di base, che includono settori quali biologia molecolare, genomica, trascrittomica e bioinformatica [4]. Ad oggi, tuttavia, la medicina personalizzata si basa principalmente su tecniche di sequenziamento ad elevato rendimento ("high-throughput sequencing", HTS), capaci di generare una grande quantità di dati analizzabili (sezione 1.3), su tecniche bioinformatiche per l'identificazione di specifici biomarcatori connessi allo stato fisiopatologico del paziente, e su modelli predittivi dalla biologia dei sistemi. La combinazione dei metodi forniti da queste tre discipline permette un costante monitoraggio della risposta al trattamento [5,6] e la caratterizzazione di fenomeni biologici complessi, dando quindi la possibilità, in un secondo momento, di predirne l'andamento (sezione 1.5). In tal senso, i metodi predittivi della biologia dei sistemi forniscono un valido supporto al disegno di nuovi approcci terapeutici, potenzialmente più appropriati per specifiche classi di pazienti [7]. Un ulteriore contributo al settore della medicina personalizzata è stato apportato dalla scoperta della regolazione genica mediata da specifiche classi di RNA non-codificanti (ncRNAs) [8]. Tra le classi ad oggi maggiormente conosciute

e meglio caratterizzare di ncRNAs vi sono i microRNA (miRNAs), i.e. RNA regolatori di piccole dimensioni (sncRNAs) principalmente coinvolti in meccanismi di repressione genica a livello post-trascrizionale [9], seguiti poi dai long non-coding RNAs (lncRNAs), che invece hanno dimensioni equiparabili a quelle dei trascritti codificanti e sono coinvolti in più processi di regolazione genica e di segnalazione cellulare [10]. Proprio grazie alla loro ampia caratterizzazione strutturale e funzionale, i miRNAs (sezione 1.4) sono diventati protagonisti indiscussi nel settore della medicina innovativa, essendo ritenuti (i) biomarcatori di grande importanza in svariate patologie, ma anche (ii) potenziali agenti terapeutici, applicabili in terapie personalizzate [11–14]. Per quel che concerne quest'ultimo punto, tuttavia, sembra che si sia ancora lontani dal poter mettere in atto le potenzialità terapeutiche di queste molecole in trattamenti clinici. Infine, da un punto di vista sistemico, grande importanza ha avuto la caratterizzazione (seppur limitata) delle reti di interazioni tra classi di ncRNAs e RNA codificanti (mRNAs) in varie patologie [15,16], secondo quanto illustrato dalla teoria degli RNA endogeni competitivi (ceRNA) [17]. Questo ha dato un notevole apporto alla biologia dei sistemi, grazie all'incremento della complessità e della realisticità delle network di interazioni biologiche, permettendo a sua volta la costruzione di algoritmi predittivi più accurati, come verrà discusso nei prossimi paragrafi.

1.1 Bioinformatica e Biologia dei sistemi: approccio riduzionistico vs. approccio olistico

Come accennato nel paragrafo precedente, i settori della bioinformatica e della biologia dei sistemi hanno acquisito un indiscusso ruolo di supporto, non solo nella odierna ricerca di base, ma, in misura sempre crescente, anche nel campo medico operativo. In particolare, la combinazione di queste due discipline rende possibile l'analisi di datasets di varie dimensioni e tipologie. Sebbene strettamente connesse tra loro, bioinformatica e biologia dei sistemi presentano divergenze concettuali. In quanto settore scientifico multidisciplinare, la bioinformatica mira alla risoluzione di problemi biologico-molecolare attraverso l'utilizzo di strumenti informatici e computazionali integrati ad approcci statistici e matematici [18]. Obiettivo primario della bioinformatica è quello di fornire all'utente una descrizione quantitativa e/o qualitativa dei fenomeni biologici analizzati attraverso l'elaborazione di dati biologici/clinici grezzi. Ciò è reso possibile sia mediante l'utilizzo di pacchetti software dedicati, sia grazie all'esistenza di una vasta gamma di risorse (databases e strumenti analitici) online da cui poter ricavare tali dati e attraverso cui poter eseguire l'analisi di specifiche tipologie di dati [18-20]. Alla pari della biologia molecolare, la bioinformatica utilizza un approccio "riduzionistico" e statico, focalizzando la propria attenzione su singoli eventi molecolari

che avvengono all'interno di una popolazione cellulare in un preciso momento, e trova applicazioni in diversi settori della biologia molecolare: analisi di espressione genica; annotazione genica (mappatura) su sequenze di riferimento; analisi di espressione proteica; annotazione di sequenze amminoacidiche su sequenze di riferimento [19]; predizione di strutture proteiche [21,22]; analisi delle modifiche epigenetiche ed epitrascrittomiche [23].

Diversamente dalla bioinformatica, la biologia dei sistemi si pone piuttosto l'obiettivo di studiare e simulare non solo interazioni statiche, ma anche dinamiche tra le varie molecole presenti nel sistema cellulare utilizzando un approccio "olistico". La costruzione di modelli biologici complessi e tendenzialmente dinamici necessita anzitutto dell'integrazione di dati genomici, trascrittomici e proteomici, utilizzati per dedurre relazioni all'interno di pathway o reti di interazioni (sezione 1.3) – tipicamente vie di regolazione genica, vie metaboliche, vie di trasduzione del segnale, reti di interazione proteina-proteina e reti di interazione miRNA-mRNA [24]. Il fine ultimo è quello di creare modelli di sistemi biologici complessi sempre più completi, che descrivano correttamente il funzionamento dinamico dei sistemi biologici reali e a cui possano essere applicati modelli computazionali capaci di predire in maniera sempre più accurata gli effetti di specifiche variazioni molecolari sul fenotipo cellulare nel tempo [25,26] (sezione 1.6.3).

1.2 Pathway e network biologici: concetti di base

Un sistema complesso è composto da un gran numero di componenti interconnessi, le cui interazioni possono rappresentare una varietà di funzioni del sistema. Questi sistemi vengono rappresentati in forma di grafi, che, in generale, sono definiti come un insieme di punti, denominati nodi o vertici del grafo (i componenti), interconnessi tra loro mediante linee dette archi (relazioni tra i componenti). Da un punto di vista biologico-molecolare, i componenti dei grafi rappresenteranno solitamente molecole biologiche quali geni, trascritti, proteine o metaboliti, mentre gli archi rappresenteranno le loro reciproche interazioni funzionali o correlazioni statistiche. Più precisamente, i grafi che descrivono interazioni o correlazioni tra componenti del sistema cellulare possono essere rappresentati in forma di pathway (vie) o network (reti) biologici. I pathway biologici (vie di attivazione genica, vie di segnalazione e vie metaboliche) sono grafi orientati, i cui componenti (solitamente proteine e metaboliti) sono tutti diretti verso uno stesso evento molecolare o manifestazione fenotipica (endpoint biologici), e le cui interazioni possono essere attivanti o disattivanti [27] (Figura 1.1a). Dunque, un pathway descrive la sequenza ordinata di eventi molecolari all'interno di una certa via di attivazione genica, metabolica o di segnalazione. Ciò conduce a uno specifico endpoint biologico a valle del pathway, determinando la regolazione dello

stato cellulare globale. Diversamente dal caso dei pathway, i componenti delle reti biologiche non sono diretti verso un endpoint comune. Piuttosto, le reti hanno lo scopo di mostrare le relazioni funzionali (o le correlazioni statistiche) esistenti tra diverse componenti cellulari (senza alcuna limitazione a specifiche classi molecolari) [28], o anche tra queste e condizioni patologiche (Figura 1.1b). Esempi comuni di reti biologiche includono le reti di interazioni proteina-proteina (reti PPI), le reti di relazioni geni-fattori di trascrizione, e le reti di correlazione statistica tra componenti molecolari e fenotipo (o anche correlazioni genotipo-fenotipo).

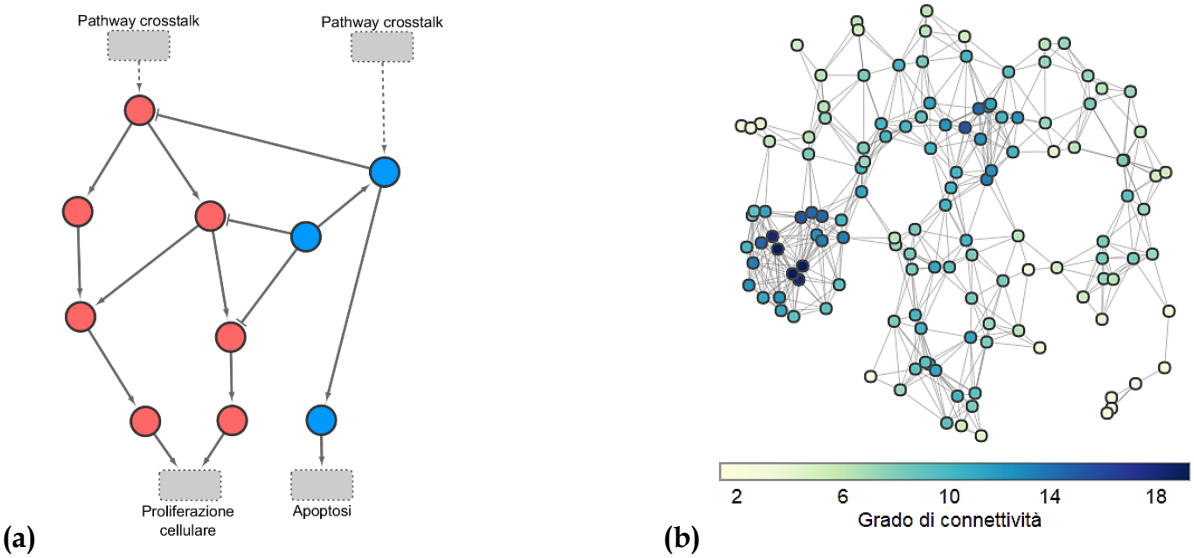


Figura 1.1 In figura è mostrato un esempio di grafo direzionato raffigurante un generico pathway biologico **(a)** e un grafo non direzionato raffigurante un generico network biologico **(b)**. Come spiegato nel testo, un pathway consiste in un grafo orientato, caratterizzato da interazioni attivanti e disattivanti e dove i componenti sono diretti verso uno specifico endpoint biologico. In questo caso, il grafo in (a) mostra un pathway di segnalazione, dove i nodi rossi (geni, RNA o proteine) si trovano a monte di processi di proliferazione cellulare (primo endpoint), mentre i nodi blu (geni, RNA o proteine) si trovano a monte di processi apoptotici (secondo endpoint). Diversamente dal caso del pathway, il grafo di un network biologico può essere o non essere orientato, mentre i suoi componenti non sono mai orientati verso un endpoint. In questo caso, il grafo in (b) mostra una generica rete di interazioni tra proteine (PPI), dove il grado di connettività di ogni nodo (proteina) viene mostrato attraverso l'intensità del colore: nodi a basso grado di connettività vengono indicati con colori più chiari; nodi ad alto grado di connettività vengono indicati con colori più scuri.

1.3 Tecniche HTS nella biologia dei sistemi e nella medicina personalizzata

Nel loro insieme, le ricostruzioni grafiche di pathway e network biologici rappresentano uno dei fondamenti della biologia dei sistemi (sezione 1.5). In tal senso, un importantissimo contributo

iniziale è stato dato dai metodi genetici e immunochimici a basso rendimento (low-throughput), che hanno permesso la ricostruzione di buona parte dei pathways biologici presenti nel sistema cellulare (e.g. [29,30]). Tuttavia, poiché l'obiettivo è il disegno di modelli sempre più accurati delle interazioni del sistema cellulare, nel corso degli anni si è reso di fondamentale importanza l'utilizzo di tecnologie sempre più avanzate, capaci di operare a livello sistemico e di consentire in breve tempo l'identificazione in parallelo di una grande quantità di interazioni molecolari insieme alla decifrazione delle loro funzioni. È proprio in questo contesto che le tecniche HTS – inerenti a genomica, trascrittomica e proteomica – hanno assunto un ruolo portante nei settori della biologia dei sistemi e della medicina personalizzata. In particolare, lo sviluppo di tecniche di sequenziamento Next-Generation [31,32] e di quantificazione dell'espressione genica hanno permesso il raggiungimento di importantissimi traguardi, che hanno segnato un punto di svolta nel campo biomedico. Tra questi, vanno certamente ricordati il sequenziamento del genoma umano, la sua caratterizzazione spaziale e l'individuazione di potenziali geni codificanti e non-codificanti non ancora funzionalmente caratterizzati [33–35].

1.3.1 Cenni sulla tecnica del microarray

La quantificazione dell'espressione genica di specifici geni è uno strumento chiave nell'analisi dei processi fisiologici e patologici. Ad esempio, analizzando i livelli di espressione di un gene (codificante o non-codificante) in varie condizioni, ed eventualmente correlandoli con quelli di altri geni, è possibile decifrarne la funzione. Le informazioni ricavate possono inoltre essere utilizzate per l'identificazione dei biomarcatori di un certo stato fisiopatologico o dei geni drivers di specifiche patologie [36], oppure per monitorare l'andamento nel tempo di una malattia o di una terapia (e.g. [37–39]). Una delle tecniche più utilizzate per il rilevamento dell'espressione di RNA è stata certamente quella del microarray, che ha visto un vasto impiego in esperimenti di vario genere. Un microarray consiste in un insieme di piccole sonde – in questo caso oligonucleotidi di DNA (solitamente 10-20 nt) – fissate ad un supporto solido detto "chip". Più sonde aventi la stessa sequenza nucleotidica vengono disposte in una ben precisa locazione del chip. Ogni locazione è detta "spot". L'insieme di tutti gli spots disposti sul chip formando una matrice (array) (Figura 1.2a). Questo metodo permette di esaminare in parallelo i livelli di espressione genica di moltissimi dei geni dei pazienti in esame. Il principio di funzionamento del microarray è essenzialmente basato sull'interazione tra filamenti di DNA (ibridazione dei filamenti di DNA) a seguito della formazione di legami idrogeno tra le coppie di basi complementari dei due filamenti. Per sommi capi, il

protocollo prevede che siano anzitutto estratti gli RNA di interesse dal campione biologico in esame. Nella fase operativa seguente, i filamenti di RNA vengono convertiti in molecole di DNA complementare (cDNA) mediante l'uso dell'enzima trascrittasi inversa, e allo stesso momento marcati con una sonda fluorescente. A questo punto si procede verso l'ibridazione fra le sonde fissate sull'array e i corrispettivi cDNA target. Segue quindi una fase di lavaggio, che ha lo scopo di rimuovere dall'array tutte le interazioni sonda::cDNA deboli (e quindi con appaiamento imperfetto delle basi). I cDNA interagenti con le sonde vengono quindi identificati grazie al rilevamento della fluorescenza emessa, causata dalla sonda precedentemente incorporata nelle molecole di cDNA a seguito dall'ibridazione sonda::cDNA. Infine, l'intensità del colore emesso da ogni spot viene letta da un apposito macchinario, confrontata con un valore di riferimento, quindi normalizzata e trasdotta in valore numerico detto fold-change (inteso come variazione dell'espressione genica) (Figura 1.2b). Mettendo a confronto i valori di espressione ottenuti per una certa classe di pazienti (caso) con i valori ottenuti per individui sani (controllo) si ottiene, a seguito di opportuna elaborazione bioinformatica e statistica, un insieme di geni differenzialmente espressi (DEGs), cioè geni i cui valori di espressione sono in grado di distinguere tra stato sano e stato patologico [40]. Similmente, confronti tra espressioni geniche di pazienti appartenenti alla stessa classe (ad esempio leucemia) permettono di suddividere i pazienti in sottoclassi differenti, o anche di individuare pazienti ad uno stadio più avanzato della malattia, con un certo grado di accuratezza. A livello clinico, queste informazioni possono fornire delle linee guida per stabilire nuovi criteri di classificazione e stratificazione dei pazienti, operati sulla base di dati molecolari, e dunque potenzialmente più accurati rispetto ai criteri tradizionali.

Sebbene l'analisi microarray è stata classicamente utilizzata per il confronto dei livelli di espressione di set di geni codificanti [40], questa tecnologia ha trovato altri impieghi molto rilevanti. Ad esempio, l'analisi microarray caso/controllo di set di miRNA differenzialmente espressi [41] ha permesso l'individuazione di specifici subset di miRNA come biomarcatori diagnostici e prognostici in malattie oncologiche [42]. Inoltre, indagini sul profilo di espressione dei miRNA in diversi tessuti umani hanno rivelato l'esistenza di diversi cluster di miRNA sul genoma umano, laddove miRNA appartenenti allo stesso cluster vengono espressi simultaneamente, implicando una correlazione con specifici fattori della trascrizione [43]. Fissando in ogni spot dell'array specifiche classi di proteine o aptameri sintetici anziché sonde di DNA, è possibile ottenere protein microarray (proteinchip) di vario genere, utilizzati in biomedicina per rilevare la presenza e/o la quantità di specifiche proteine nei lisati cellulari dei campioni biologici in esame [44,45]. In particolare, i protein microarray vengono adoperati per l'identificazione di interazioni proteina-proteina, per l'identificazione dei substrati di proteine chinasi, o ancora per l'identificazione dei target di piccole molecole

biologicamente attive (Figura 1.3). Un ulteriore importante applicazione del microarray la si ritrova nel saggio di immunoprecipitazione della cromatina (chromatine immunoprecipitation assay, ChIP), adoperato nel campo dell'epigenetica per investigare riguardo alle interazioni proteine-DNA su scala locale [46].

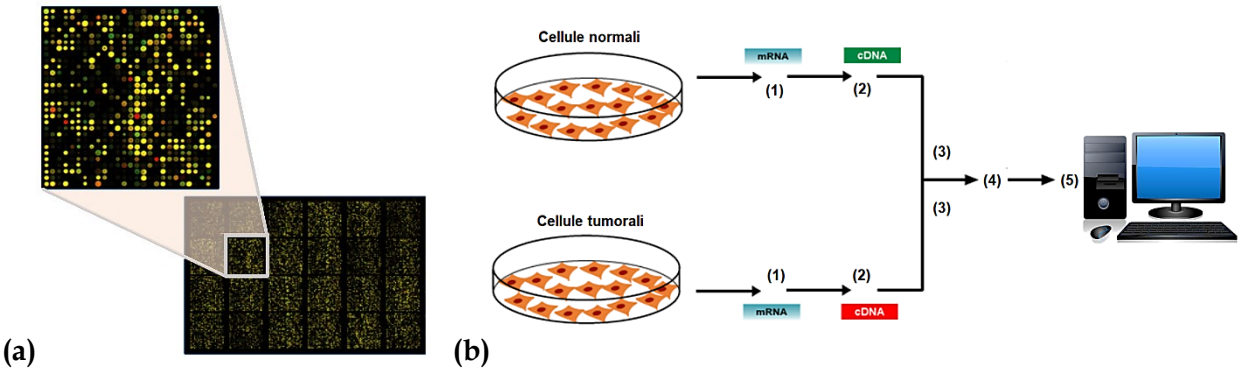


Figura 1.2 In **(a)** viene mostrata la struttura di un generico microarray. Come spiegato nel testo, le sonde vengono posizionate in clusters all'interno di ogni array. Ad ogni cluster viene assegnato uno spot ben preciso all'interno dell'array. Gli array sono quindi disposti in tandem su un dispositivo solido, detto chip. In **(b)** vengono illustrate le fasi salienti del protocollo di un'analisi di microarray. La fase (1) prevede l'estrazione di RNA (in questo caso mRNA) dal campione biologico; la fase (2) prevede la trascrizione inversa degli RNA estratti e la loro marcatura con molecole fluorescenti. (3) Le sonde vengono quindi ibridate con i cDNA marcati (4) e successivamente trattati con lavaggio per rimuovere eventuali appaiamenti deboli. Infine, (5) le ibridazioni vengono lette, trasmesse ad un sistema informatico e normalizzate sulla base del controllo al fine di identificare geni differenzialmente espressi nelle due condizioni.

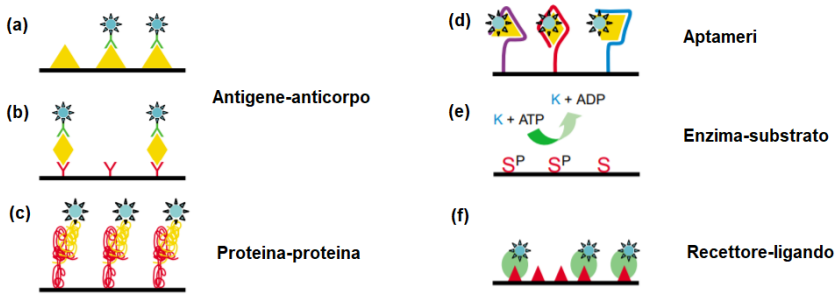


Figura 1.3 Per analizzare le interazioni proteiche vengono adoperate diverse classi di sonde nel protein microarray. In **(a)** è mostrato un saggio di interazioni antigene-anticorpo semplici, mentre in **(b)** lo stesso saggio viene eseguito con una tecnica a sandwich (sandwich immunoassay). In **(c)**, è mostrato un saggio di interazione proteina–proteina. In **(d)** viene illustrato un saggio di interazione acido nucleico-proteina, eseguito utilizzando come sonde degli aptameri, cioè acidi nucleici aventi la proprietà di legarsi a una o più proteine. In **(e)** è illustrato un saggio di interazione chinasi-substrato, dove i substrati (S) fungono da sonde. Infine, in **(f)** viene mostrato un saggio di interazione recettore-ligando. Immagine tratta da [44], modificata.

1.3.2 Cenni sul Next Generation Sequencing

Un risvolto storico si ebbe nel biennio 2005-2006 con la pubblicazione di due lavori che hanno rappresentato il preludio di una nuova era nel campo del sequenziamento degli acidi nucleici, grazie alla nascita delle tecnologie NGS, dette anche "tecnologie di sequenziamento di seconda generazione" [47,48]. Il concetto centrale qui consisteva nell'adoperare microsfere (~1 µm) o sottilissimi supporti planari di agarosio su cui far aderire i frammenti di DNA da amplificare e sequenziare in parallelo genomi interi [49] (Figura 1.4). Sebbene negli anni a seguire siano state sviluppate diverse tecnologie NGS, ognuna basata su principi differenti rispetto alle altre [50], è possibile stabilire per tutti questi metodi un framework generale comune [31,51,52]. Nel caso di uno studio qualitativo e quantitativo del trascrittoma di una certa popolazione cellulare o di un tessuto biologico, il primo step da seguire è, come nel caso dell'analisi microarray, l'estrazione dell'RNA dal campione d'interesse e la sua conversione in cDNA mediante trascrittasi inversa. Solitamente, a valle del protocollo di estrazione dell'RNA totale del campione, si esegue un impoverimento di RNA ribosomiale (rRNA) a favore di RNA codificante e/o non codificante. Chiaramente, lo specifico protocollo da seguire dipende dalla classe di RNA d'interesse. Nel caso si vogliano isolare specifiche classi di piccoli ncRNA, come ad esempio i miRNA, vengono aggiunti degli step di arricchimento nel protocollo di isolamento dell'RNA, che spesso fanno capo a tecniche differenti (ad esempio combinazioni di centrifugazioni differenziali ed elettroforesi). A questo punto, segue una fase di frammentazione dei filamenti a doppio strand di cDNA ottenuti nella fase di trascrizione inversa, in

quanto questo tipo di tecnologie non sono in grado di gestire acidi nucleici oltre una certa lunghezza. Poiché la frammentazione dei filamenti è un processo random, la stragrande maggioranza dei filamenti frammentati presenterà delle asimmetrie, dette overhangs, tra le estremità 3′ e 5′: ciò significa che una delle due estremità presenterà 1-3 nucleotidi in eccesso rispetto al filamento opposto, e quindi a singolo filamento. Dato che lo step successivo richiede la ligazione dei cDNA agli adattatori NGS – i.e. oligonucleotidi sintetizzati che hanno la funzione di ancorare i cDNA ai supporti NGS – è necessario rimuovere gli overhangs presenti nei cDNA frammentati e modificarne le estremità in modo da permetterne la ligazione. Ciò viene ottenuto con le fasi di riparazione delle estremità e adenilazione del 3′. Inoltre, è importante sottolineare che le procedure di ligazione e di eventuale emulsionamento fanno in modo che ogni frammento di cDNA vada a formare un cluster ben distinto rispetto agli altri (analogamente al concetto di spot nel microarray). Secondo lo schema operativo classico, una volta che i frammenti di cDNA sono stati ben saldati al supporto di base (microsfere o supporti planari) mediante gli adattatori, si può procedere con la fase di amplificazione mediante l'avviamento di diversi cicli di PCR. Questo ha lo scopo di migliorare il rilevamento del segnale nella fase successiva di sequenziamento, principalmente per una questione di detection limit dei rilevatori. Dunque, alla fine dei cicli di PCR si arriva ad avere, per ogni frammento di cDNA di partenza, molteplici copie di cDNA riunite in cluster (Figura 1.4). Terminata l'amplificazione, si procede con il sequenziamento. Qui vengono utilizzate coppie di primers che permettono alla DNA polimerasi di procedere con la sintesi di nucleotidi. In questa fase, per l'allungamento del filamento di nuova sintesi, solitamente non viene fatto uso di normali nucleotidi, bensì di nucleotidi chimicamente modificati capaci di emettere un segnale luminoso ogni volta che avviene la creazione di un nuovo legame fosfodiesterico. In alternativa a quanto appena descritto, poiché la fase di amplificazione mediante PCR introduce un certo bias a causa di errori che possono avvenire durante la procedura di duplicazione, sono stati studiati alcuni protocolli, detti "PCR-free", che permettono di bypassare questo step. Altre variazioni degne di nota nella procedura descritta riguardano il supporto di base su cui avviene la reazione di sequenziamento e la tecnica di rilevamento del segnale durante il sequenziamento. In particolare, alcune tecnologie fanno aderire al supporto di base unità di DNA polimerasi anziché oligonucleotidi adattatori. In questo caso lo step della ligazione viene bypassato, ma i vari frammenti di cDNA vengono catturati dall'enzima e sequenziati man mano che il sistema fornisce nucleotidi modificati, capaci di emettere fluorescenza [53]. Per quel che concerne la tecnica di rilevamento della formazione di legami fosfodiesterici, alcune tecnologie utilizzano la variazione di pH, anziché segnali luminosi, indotta dal rilascio di ioni idrogeno durante l'estensione del DNA [54].

I segnali captati durante la fase di sequenziamento vengono automaticamente tradotti in reads (in italiano "letture") grezze, contenuti in specifici files trattati mediante opportune suite di software bioinformatici. Ogni read consiste in una sequenza ordinata di lettere (A, T, G o C) che rappresenta, teoricamente, la sequenza nucleotidica di un certo frammento di cDNA. Ogni file A ognuna delle quattro lettere è assegnato un valore che esprime una stima dell'accuratezza del sequenziamento per il nucleotide in questione. I files contengono anche informazioni quantitative su ogni read. Proprio sulla base dei valori di accuratezza avvengono le prime due fasi del trattamento delle reads, ovvero "quality control" e "quality trimming", allo scopo di eliminare reads con stime troppo basse di accuratezza, di individuare eventuali sequenze chimeriche e di tagliare estremità di qualità scadente. Le reads ottenute alla fine di questo processo possono quindi essere utilizzate per vari tipi di analisi, che divergono sostanzialmente le une dalle altre. Tra queste, vi sono analisi che permettono di ricavare set di DEGs e altre che permettono di individuare mutazioni rispetto ad una sequenza di riferimento. Comunque, le procedure che portano a questi risultati esulano dagli scopi del presente elaborato e pertanto non verranno trattate.

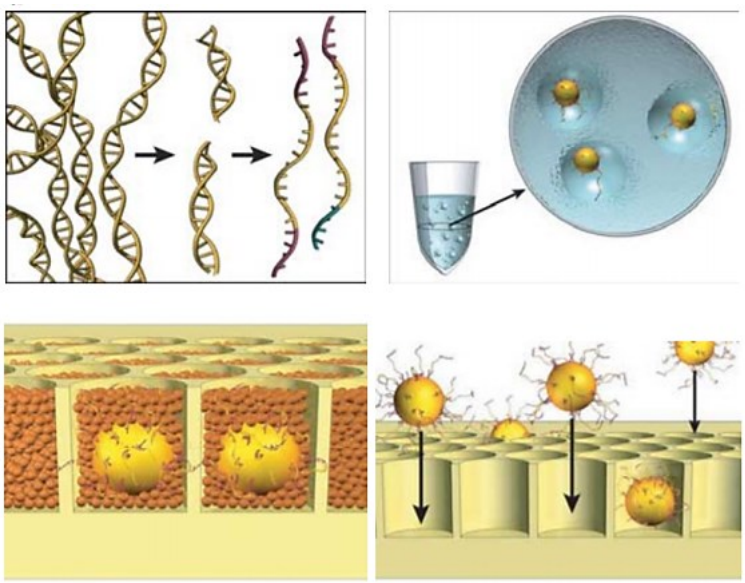


Figura 1.4 Immagine originale tratta da [47] in cui viene riassunto il design della prima tecnologia di sequenziamento NGS (metodo Roche 454) e il relativo schema di preparazione del campione da analizzare. **(a)** Il DNA Genomico viene anzitutto isolato, frammentato, ligato ad adattatori e separato in singoli filamenti (riquadro in alto a sinistra). Gli oligonucleotidi così ottenuti vengono fatti aderire a delle microsfere in modo tale da far sì che ad ogni microsfera si leghi solamente un frammento nucleotidico. Le microsfere vengono quindi catturate in microgocce di una soluzione per PCR emulsionata con olio. A questo punto viene avviata la reazione di amplificazione mediante PCR, che avrà luogo all'interno delle microgocce. Ciò risulterà in milioni di copie di uno stesso frammento per ogni microsfera (riquadro in alto a destra). Alla fine della reazione di PCR, le microsfere vengono rimosse dalla soluzione emulsionata, i frammenti adesi alle microsfere vengono denaturati in modo da mantenere il loro stato a filamento singolo, ed infine le microsfere vengono depositate in micropozzetti, in modo che ogni pozzetto contenga una sola sfera. In tal maniera, i frammenti vengono riuniti in clusters. I pozzetti sono posti in tandem su un supporto in fibra ottica, hanno forma esagonale, un diametro di ~44 µm, una profondità di ~55 µm, e sono distanziati gli uni dagli altri di circa 3 µm (riquadro in basso a destra). Microsfere ancora più piccole, a cui aderiscono gli enzimi necesari per la reazione di piro-sequenziamento, vengono depositate nei pozzetti (riquadro in basso a sinistra). In seguito, un rilevatore di fasci luminosi rileva la luce emessa ad ogni singolo step del piro-sequenziamento, trasmettendo quindi ogni segnale ad un sistema informatico capace di interpretare il segnale (non mostrato in figura).

1.4 MicroRNA: cenni su biogenesi, meccanismo d'azione e ruolo in vari processi fisiopatologici

Come anticipato nelle sezioni precedenti, i miRNA costituiscono una classe di sncRNA di notevole importanza in campo biomedico. Nella loro forma matura, i miRNA hanno dimensioni ~18-23 nt e sono coinvolti nella regolazione dell'espressione genica. È ormai noto che queste molecole svolgono la loro azione regolatrice prevalentemente a livello post-trascrizionale, sebbene siano noti anche meccanismi di regolazione a livello trascrizionale [55-57]. Prima di giungere allo stato maturo, i trascritti iniziali dei miRNA passano attraverso un processamento distinto in due fasi, di cui uno nucleare e l'altro citoplasmatico (Figura 1.5).

La trascrizione dei geni miRNA è principalmente mediata dall'enzima RNA Polimerasi II (Pol II), e solo in pochi casi dell'enzima RNA polimerasi III (Pol III) [58]. Il trascritto risultante è detto "miRNA primario" (pri-miRNA) e può essere lungo anche diverse centinaia di nucleotidi. La stragrande maggioranza dei geni miRNA è localizzata in regioni non-codificanti del genoma umano. Di questi, ~46% si trovano in regioni intrageniche non-codificanti (introniche) e vengono escissi mediante azione enzimatica durante la fase di splicing del mRNA, mentre la parte rimanente è localizzata all'interno di regioni intergeniche e dispone di promotori propri [58–60]. Solo una ristretta minoranza di geni miRNA risiede in regioni codificanti del genoma [58–60]. Più della metà dei geni miRNA sono organizzati in cluster policistronici, il che implica la co-espressione dei miRNA che si trovano sullo stesso cluster [43,61]. Inoltre, diversi studi sperimentali hanno recentemente dimostrato una forte interdipendenza nel processo di biogenesi per miRNA appartenenti allo stesso cluster [61]. Una volta trascritto, il pri-miRNA si ripiega naturalmente a formare una struttura a forcina a doppio filamento, necessaria per la prima fase (nucleare) del processo di maturazione. Questa consiste nel taglio operato dal complesso proteico detto Microprocessore, le cui componenti fondamentali sono l'enzima RNasi di tipo III Drosha insieme al suo cofattore essenziale DGCR8 [58]. A seguito del taglio di Drosha, il pri-miRNA viene accorciato in un trascritto lungo ~60-70 nt, detto "precursore del miRNA" (pre-miRNA) e avente struttura secondaria di tipo stem-loop. Quest'ultimo viene poi esportato nel citoplasma, dove sarà ulteriormente tagliato ad opera della RNasi di tipo III Dicer, complessato con il cofattore TRBP. A seguito del taglio di Dicer, la regione loop del pre-miRNA viene rimossa, e il risultato finale è l'ottenimento di due piccoli filamenti complementari di RNA, di lunghezza media ~21 nt. Poiché questi due piccoli filamenti sono stati ottenuti dalle estremità 5' e 3' del pre-miRNA, vengono denominati, rispettivamente, con la dicitura miRNA-5p e miRNA-3p. Potenzialmente, entrambi i filamenti possono essere utilizzati come miRNA maturi regolatori. Tuttavia, solamente uno dei due filamenti verrà mantenuto alla fine del processo di biogenesi e sarà funzionale, mentre l'altro verrà degradato. A questo punto, il filamento

funzionale viene caricato su uno dei membri della famiglia delle proteine Argonauta (AGO1-4 negli esseri umani, di cui solo AGO2 è enzima funzionante). Il complesso miRNA-AGO costituisce l'unità core di un complesso ribonucleoproteico più grande detto "complesso di silenziamento indotto da miRNA" (in inglese "miRNA-induced silencing complex", miRISC), e rappresenta il reale macchinario regolatore della traduzione [58] (Figura 1.5). In realtà, il processo di biogenesi appena descritto rappresenta la via di maturazione canonica dei miRNA. Ad oggi, tuttavia, sono note altre tre vie di maturazione dei miRNA, non canoniche, indicate come (i) Drosha-DGCR8 indipendente, (ii) Dicer indipendente e (iii) TUTasi-dipendente [58]. Inoltre, è ormai noto che anche modifiche epitrascrittomiche dei pri-miRNA, come ad esempio metilazione e deamminazione dell'adenosina (A-to-I editing), hanno un forte impatto sui processi di maturazione di questi RNA regolatori, provocandone il blocco della biogenesi o l'alterazione della cinetica delle reazioni enzimatiche di taglio [62,63]. Tuttavia, una trattazione più dettagliata di questi argomenti esula dallo scopo del presente elaborato, e non verranno perciò approfonditi.

Senza alcun dubbio, il meccanismo di regolazione dominante svolto dai miRNA è rappresentato dal silenziamento genico operato a livello post-trascrizionale [64]. La condizione essenziale che sottende tale regolazione negativa del processo di traduzione consiste in un appaiamento termodinamicamente stabile tra (i) una regione ben definita del miRNA e (ii) una o più regioni complementari interne al trascritto codificante, dette "elementi di risposta ai miRNA" (in inglese "miRNA response elements", MREs). Notoriamente, le interazioni miRNA::mRNA più comuni, riconosciute come canoniche, coinvolgono la cosiddetta "seed region" del miRNA e una o più regioni MRE contenute all'interno della regione non tradotta al 3' (in inglese "3' untranslated region", 3'UTR) [65]. La seed region è localizzata all'estremità 5' di ogni miRNA, e include i nucleotidi 2-8 o 2-7. Tuttavia, diversi lavori scientifici hanno dimostrato che interazioni miRNA::mRNA non canoniche possono avvenire con una certa frequenza. Queste includono, ad esempio, interazioni tra la regione centrale di un miRNA (nt 4-16) e MREs del mRNA, oppure tra seed region del miRNA e MREs incluse all'interno degli esoni o della regione 5'UTR di un trascritto codificante [66-68].

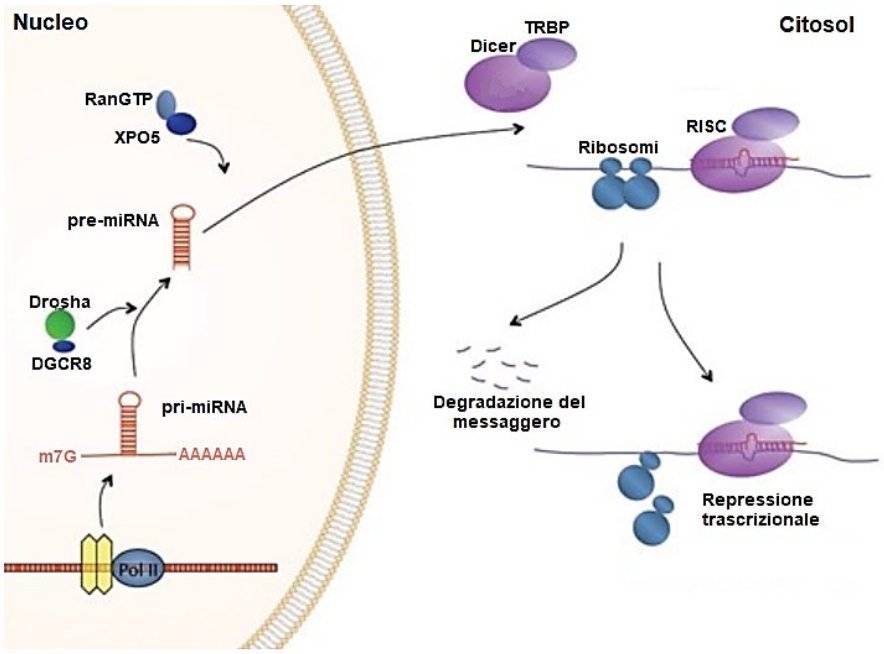


Figura 1.5 Modello canonico raffigurante le tappe principali della biogenesi dei miRNA e il meccanismo attraverso cui queste molecole modulano la traduzione dei trascritti codificanti. Figura modificata, tratta da (https://vetmed.tamu.edu/faculty/zhou-lab/research/microrna/).

In generale, il silenziamento genico mediato da miRNA può avvenire tramite degradazione diretta del trascritto codificante o tramite blocco della traduzione (Figura 1.5). Secondo alcuni studi, pare che la degradazione diretta del mRNA venga mediata dall'attività endonucleasica di AGO2, quanto meno con una certa frequenza. Secondo quanto riportato in letteratura, sembra che ciò accada a seguito dei rari casi di perfetta complementarietà tra seed region e MRE. Inoltre, si pensa che tale processo enzimatico porti, in ultimo, alla degradazione del miRNA funzionale stesso [69]. Comunque, nella maggior parte dei casi l'appaiamento miRNA::mRNA è parziale, presentando dei mismatch interni alla regione d'interesse, oppure avviene con appaiamenti termodinamicamente meno stabili, non di tipo Watson-Crick [65]. Ciò previene lo stimolo dell'attività endonucleasica di AGO2 e, allo stesso tempo, induce il reclutamento di diversi effettori citoplasmatici e proteine che legano l'RNA, provocando di conseguenza la repressione della traduzione e la degradazione del mRNA [64,70].

Alterazioni del normale profilo di espressione dei geni miRNA possono avere conseguenze negative sullo stato fisiologico cellulare, portando all'insorgenza di diverse patologie. La prima patologia umana ad essere stata associata ad alterazioni dell'espressione di geni miRNA è stata la leucemia linfatica cronica (LLC) [71]. In particolare, si scoprì che i due miRNA miR-15a e miR-16, localizzati sul cromosoma 14 (regione 13q) a formare il cluster miR-15a/16-1, erano deleti o fortemente sotto-

espressi nel ~68% dei casi esaminati di LLC a cellule B [71]. Le speculazioni derivanti da questo primo studio di correlazione aprirono la strada ad una vastissima ricerca in questo nuovo settore della biologia molecolare. In breve tempo, ulteriori studi di correlazione dimostrarono che un notevole numero di geni miRNA mappavano all'interno di regioni fragili del genoma o soggette ad alterazioni quali delezioni, inserzioni e duplicazioni associate all'insorgenza di tumori [72]. Inoltre, furono definiti diversi meccanismi molecolari attraverso cui ad alterazioni dei livelli intracellulari di miRNA maturi corrispondevano specifiche alterazioni in pathways biologici ben precisi – ad esempio, si scoprì che miR-15a e miR-16 targettavano regioni MRE interne al 3'UTR del gene BCL2, contribuendo perciò all'attivazione del pathway apoptotico [73]. Ciò cominciò a dare comprensione sul ruolo che questi acidi ribonucleici ricoprono nel mantenimento dell'omeostasi cellulare. Da allora, molti miRNA sono stai associati a varie tipologie di cancro. Più nello specifico, esattamente come alcuni geni codificanti vengono distinti tra proto-oncogeni (che stimolano la proliferazione cellulare) e oncosoppressori (che tendono a inibire la proliferazione), così anche anche i miRNA possono essere suddivisi in miRNA oncogeni (miRNA che targettano oncosoppressori) e miRNA oncosoppressori (miRNA che targettano proto-oncogeni) [14]. Comunque, oltre che all'insorgenza di vari tumori, scompensi nei livelli di espressione dei miRNA sono associati alla manifestazione o all'aggravamento di altre patologie, come ad esempio malattie cardiovascolari [74], autismo [75], disfunzioni del sistema immunitario e alterazioni metaboliche [76].

Oltre ad essere prodotti all'interno del sistema cellulare e ad esercitare la loro azione regolatrice a livello citoplasmatico nella cellula di origine, i miRNAs vengono normalmente secreti – in forma libera o inclusi in microvescicole – nella matrice intercellulare e nei fluidi corporei, e pertanto sono presenti in concentrazioni significative (talvolta particolarmente elevate) [77–79]. In questo contesto, i miRNA fungono da molecole di comunicazione tra cellule, oppure come elementi regolatori dell'espressione genica delle cellule riceventi [77,80,81]. Grazie all'azione protettrice delle vescicole fosfolipidiche che li contengono o delle proteine con le quali sono complessati (più frequentemente AGO2), i miRNA extracellulari mostrano una notevole stabilità nonostante la loro natura ribonucleica [82,83]. Alla stessa stregua dei miRNA operanti all'interno del sistema cellulare della cellula d'origine, i miRNA circolanti hanno assunto una grandissima importanza sia nella ricerca di base che in ambito clinico, in quanto contribuiscono fortemente alla progressione di diversi tumori, regolano processi di vascolarizzazione ed ematopoiesi, e rappresentano biomarcatori diagnostici e prognostici potenzialmente affidabili di diverse patologie [84].

Considerando il ruolo fondamentale che questi oligonucleotidi endogeni esercitano nella regolazione genica e la vastità delle loro interazioni con trascritti codificanti e non-codificanti all'interno del sistema cellulare, non sorprende che la loro introduzione all'interno di pathways e

networks biologici abbia portato al miglioramento, in termini di accuratezza, delle predizioni operate nel campo della biologia dei sistemi [85].

1.5 Biologia dei sistemi: schema operativo generale

Da un punto di vista operativo, la biologia dei sistemi, in combinazione con tecniche di bioinformatica, può utilizzare uno di due approcci sperimentali complementari tra loro per giungere ad ottenere dati di diversa tipologia e volume [86] (Figura 1.6). (i) Uno di questi due, detto "dall'alto verso il basso" (top-down) prevede anzitutto la raccolta di una grossa mole di dati, ottenuti con tecniche HTS (sezione 1.3), e la loro successiva elaborazione bioinformatica. In questa fase del flusso di lavoro, il ricercatore ottiene un quadro relativamente ampio dei cambiamenti che si verificano in risposta a una perturbazione ben definita. Ad esempio, la misurazione delle variazioni dell'espressione di migliaia di mRNA o miRNA mediante microarray o tecniche NGS [87] in campioni di tessuto tumorale è ormai una procedura sperimentale di routine. Ai dati HTS ottenuti vengono quindi applicate le opportune modellazioni statistiche, necessarie per rendere i dati accessibili alle indagini successive, ricavando informazioni sulla struttura di una rete di interazioni biologiche e identificando pattern biologici e determinando il coinvolgimento di specifici pathway e processi molecolari. Questo tipo di analisi permette di generare inferenze circa potenziali variazioni del sistema cellulare a partire da perturbazioni iniziali note, avvenute su singoli componenti della rete o del pathway, per esempio il knockdown o il knockout di un gene. (ii) L'altro approccio, detto "dal basso verso l'alto" (bottom-up), prevede lo studio di un minor numero di componenti, ma in maniera più approfondita, per comprenderne le reciproche relazioni quantitative e decifrare gli eventi conseguenti alle loro interazioni. Gli esperimenti svolti in questa prima fase misurano generalmente le variabili chiave del sistema in funzione del tempo, e, talvolta, anche in funzione dello spazio. Questi sistemi sono perciò analizzati utilizzando modelli dinamici in grado di simulare l'evoluzione del comportamento del sistema nel tempo. Più precisamente, quando le rappresentazioni della media del comportamento del sistema sono regolari (cioè, dati un particolare insieme di parametri e condizioni iniziali, questi genereranno sempre lo stesso output), i modelli dinamici applicati sono generalmente deterministici [88]. Se invece è necessario tenere in considerazione fluttuazioni di singole componenti (ad esempio singoli metaboliti o altre molecole), possono essere utilizzati modelli stocastici che tengono conto di tali eventi casuali [89]. Comunque, entrambe le tipologie di modelli dinamici possono generare predizioni quantitative che possono essere successivamente verificate sperimentalmente. Grazie a queste iterazioni tra simulazione ed esperimento, si possono dedurre meccanismi di regolazione che agiscono a livello di sistema, come

ad esempio cicli feedback o feed-forward. Dunque, le problematiche affrontate mediante modelli dinamici hanno il loro focus sui meccanismi che danno origine a proprietà emergenti a livello di sistema.

Dal punto di vista della biologia dei sistemi, l'insorgenza di una generica malattia è vista come il risultato di interazioni più o meno complesse tra elementi appartenenti a vie molecolari (pathways) perturbate, piuttosto che come "difetti" in singoli componenti o eventi molecolari – che invece costituiscono il focus della bioinformatica. L'approccio di tipo sistemico si è dimostrato particolarmente utile nello studio di malattie complesse che hanno molteplici fattori causali e manifestazioni cliniche, come ad esempio tumori, diabete mellito, malattie respiratorie e malattie cardiovascolari. [3]. A tal proposito, è interessante notare che un numero sempre maggiore di lavori scientifici ha applicato approcci sistemici [90] allo scopo di caratterizzare relazioni tra malattie complesse sulla base di parametri quali sintomatologia, prevalenza e comorbilità associate, senza considerare il meccanismo patologico in sé [91]. Approcci di tipo sistemico sono anche stati utilizzati come metodi predittivi per individuare elementi interni alle reti di interazioni biologiche, quali singoli nodi della rete o intere sottostrutture, potenzialmente rilevanti per il fenotipo della malattia. Grazie all'attuazione di queste strategie è stato possibile ricavare informazioni molto utili in campo biomedico. Ad esempio, Zhou e colleghi [91] hanno dimostrato che la variabilità dei sintomi di una particolare malattia è correlata alla densità delle interazioni proteina-proteina legate alla specifica patobiologia della malattia stessa. Ciò spiega la sovrapposizione di diversi sintomi in malattie che, sebbene differenti dal punto di vista eziologico, vedono il coinvolgimento di moduli comuni in termini di sotto-reti di interazioni proteina-proteina. In un altro studio sulle relazioni malattia-malattia, a seguito di un'analisi topologica della rete, Menche e colleghi [92] hanno mostrato che la collocazione topologica delle sotto-reti legate alla malattia in seno dell'intera rete di interazioni riflette la sua relazione patobiologica con altre malattie.

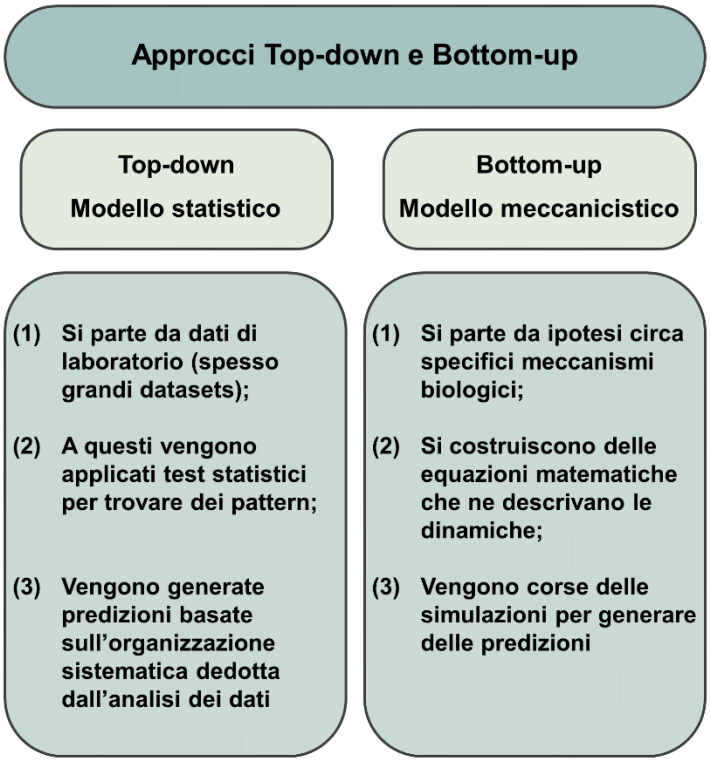


Figura 1.6 Approcci computazionali complementari utilizzati nella biologia dei sistemi. Per entrambe le categorie, il testo incluso nei box descrive, per sommi capi, la strategia logica perseguita.

1.6 Approcci computazionali top-down per l'analisi dei pathway

Come si può facilmente intuire dalle informazioni e dai concetti di base riportati nelle sezioni precedenti, lo sviluppo di sistemi *in silico* di supporto per terapie personalizzate è un processo notevolmente complesso, che richiede l'utilizzo di algoritmi di diversa tipologia. Tra questi, gli algoritmi di analisi dei pathway rappresentano uno strumento chiave. Sulla base di quanto esposto nella sezione 1.5, è possibile suddividere tutte le metodologie di pathway analysis sviluppate fino ad oggi in due grosse categorie: (i) metodi top-down che sfruttano le conoscenze *a priori* sui pathway biologici fornite da database pubblici – come ad esempio Gene Ontology (GO) [93], Molecular Signatures Database (MSigDB) [94], Reactome [95,96], Panther Database (PantherDB) [97] o Kyoto Encyclopedia of Genes and Genomes (KEGG) [98,99] – e (ii) metodi bottom-up che, invece, deducono la struttura di un pathway biologico (o di un suo modulo) a partire da misurazioni molecolari – ad esempio di espressione genica. Poiché il presente elaborato descrive lo sviluppo di due metodi top-down, e data la complessità matematica e statistica dei metodi bottom-up, solamente i metodi appartenenti al primo raggruppamento verranno brevemente discussi di seguito.

Il fine ultimo delle metodologie top-down basate sulle conoscenze *a priori* è quello di identificare pathway potenzialmente alterati in una certa condizione correlando dati di espressione genica (o attività proteica) con le informazioni contenute nel database consultato. Il risultato sarà la predizione dell'alterazione dell'espressione genica (o dell'attività proteica) di un insieme di geni (o proteine), e dunque la conseguente alterazione dei pathway o dei sotto-pathway – in termini di sovra-regolazione o sotto-regolazione – che includono quei geni. Ad ogni pathway predetto come alterato viene assegnato un p-value – cioè la probabilità di ottenere un risultato uguale o "più estremo" di quello osservato, supposta vera l'ipotesi nulla (ad esempio cfr. sezione 3.1.2) – che conferisce una significatività statistica alla predizione. I metodi top-down per analisi dei pathway possono a loro volta essere raggruppati in tre principali categorie: (i) Over-Representation Analysis ("analisi di sovra-rappresentazione", ORA), (ii) Functional Class Scoring ("valutazione della classe funzionale", FCS) e (iii) Pathway Topology-Based ("basati sulla topologia del pathway", PTB) [100] (Figura 1.7). Per semplificare la descrizione dei metodi, di seguito si farà cenno esclusivamente a studi di espressione genica, senza esplicitare casi di studi dell'espressione genica o dell'alterazione dell'attività proteica, che tuttavia seguono la stessa logica.

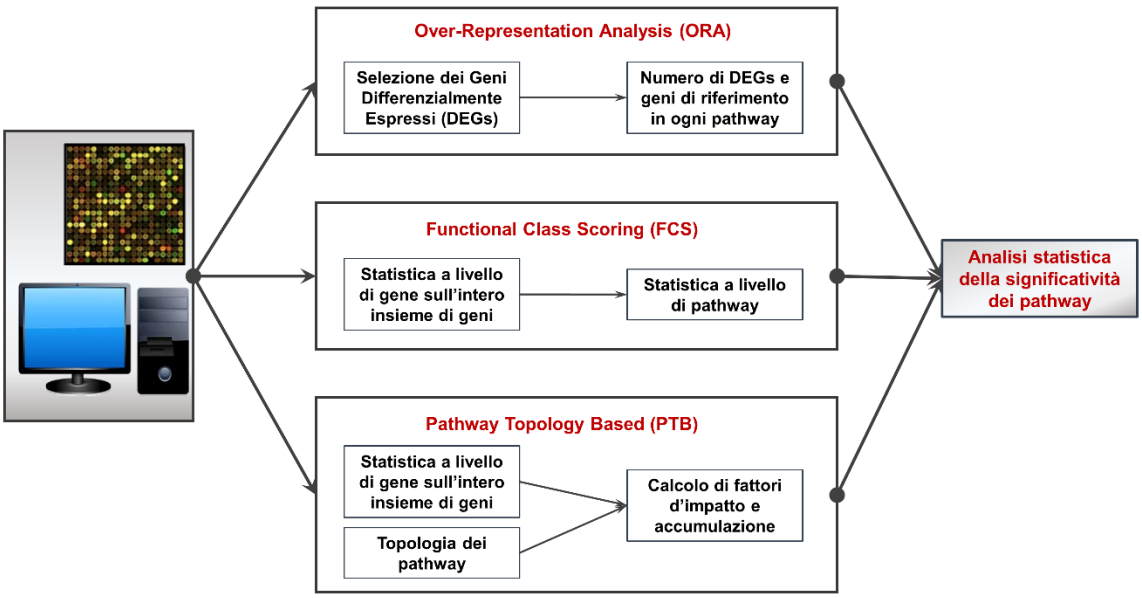


Figura 1.7 Panoramica dei possibili approcci all'analisi dei pathway. I dati generati da tecnologie di sequenziamento ad elevate prestazioni (HTS), insieme a dataset e annotazioni funzionali storate in database pubblici, rappresentano i dati in input utilizzabili per l'analisi dei pathway biologici. L'analisi di prima generazione ORA richiede in input la lista di DEGs, a prescindere dal loro fold-change, e ricava la significatività statistica dal numero di DEGs e di geni di riferimento inclusi in ogni pathway. I metodi di seconda generazione FCS prendono in input l'intero set di geni esaminati, tenendo conto dei valori di fold-change di tutti i geni. Sulla base di questi ultimi, viene anzitutto applicato un test statistico a livello di gene, mentre in una seconda fase viene applicate una statistica a livello di pathway, In aggiunta a quanto detto per I metodi FCS, i metodi PTB sfruttano le conoscenze relative alla topologia dei pathway biologici per incrementare l'accuratezza delle statistiche a livello di pathway. Per tutte e tre le categorie di metodi, l'output è una lista di pathway statisticamente significtivi per la condizione fisiopatologica esaminata. Si noti che la procedura adoperata da questi metodi è estendibile anche a studi di proteomica e metabolomica, oltre che a studi di trascrittomica.

1.6.1 Approcci ORA

Gli approcci ORA, i primi ad essere stati implementati e pertanto definiti come approcci "di prima generazione", prendono in input dati microarray e basano le proprie predizioni esclusivamente sulle informazioni contenute in GO. In linea generale, lo schema operativo degli approcci ORA si compone comunemente delle seguenti fasi: (1) i dati di espressione presi in input vengono anzitutto filtrati attraverso filtri settati dall'utente (ad esempio un valore soglia minimo di log fold-change oppure uno specifico valore di false discovery rate). (2) Nella seconda fase, per ogni pathway di GO vengono contati i geni presi in input che fanno parte di quel pathway, e vengono presi in considerazione per lo step successivo tutti i pathway che includono un certo numero di elementi in input (ogni metodo stabilisce una soglia minima). Chiaramente, poiché i pathway biologici

presentano un certo grado di ridondanza dei loro elementi, è possibile che uno stesso gene o uno stesso sotto-gruppo di geni vengano ritrovati, contemporaneamente, in più pathway. (3) Infine, ogni pathway identificato durante la fase (2) viene valutato per la significatività statistica mediante applicazione di opportuni test statistici (solitamente distribuzione ipergeometrica, distribuzione chi-quadrato o distribuzione binomiale). Lo schema operativo degli approcci ORA presenta alcune limitazioni. Anzitutto, la selezione iniziale basata sui filtri impostati dall'utente causa la perdita di informazioni circa i geni ritenuti meno significativi, potenzialmente importanti ai fini dell'analisi. In secondo luogo, i test statistici utilizzati in fase (3) tengono in considerazione esclusivamente il numero di geni facenti parte di ogni pathway di GO, ma non tengono in considerazione delle variazioni di altri importanti parametri associati ad ogni singolo gene, ad esempio il valore di fold-change rilevato dall'analisi microarray. Pertanto, i metodi ORA trattano egualmente ogni gene in input. Come conseguenza, ogni gene viene trattato come una variabile indipendente rispetto agli altri geni, ignorando qualsiasi tipo di interdipendenza tra di essi. Similmente, anche i pathway biologici vengono erroneamente trattati come variabili indipendenti in questo tipo di analisi predittiva [100].

1.6.2 Approcci FCS

Per sopperire alle limitazioni degli approcci ORA, gli approcci FCS, detti "di seconda generazione", includono nella loro analisi anche variazioni dell'espressione genica di minore entità, che posso avere importanti effetti sulla perturbazione dei pathway. (1) Questo viene reso possibile grazie all'applicazione, nella prima fase dello schema operativo generale, di test statistici bi- o multivariati, quali correlazione variazione molecolare-fenotipo, rapporto segnale-rumore, ANOVA, t-test o Z-score, a livello dei dati di espressione differenziale ottenuti dall'analisi high-throughput. Ciò conferisce diversi vantaggi: anzitutto, questa procedura permette di mantenere l'informazione originale (infatti non fa uso di filtri che discriminano tra geni significativi e non significativi). (2) In più, aggregando le statistiche ottenute a livello dei dati di espressione in un'unica statistica a livello di pathway permette, nella seconda fase dello schema operativo, l'applicazione di test multivariati, grazie ai quali vengono estrapolate informazioni circa l'interdipendenza tra le espressioni geniche o proteiche. Ciò permette di predire con maggiore accuratezza variazioni nell'attività dei pathway. In alternativa, possono essere utilizzati anche test univariati che non tengono in considerazioni le interdipendenze tra geni. Comunque, a prescindere che si scelga di utilizzare un test multivariato o univariato, la potenza del test a livello di pathway dipenderà da parametri quali la proporzione dei DEGs all'interno di un pathway, le dimensioni del pathway stesso, la forza della correlazione tra

geni appartenenti allo stesso pathway. (3) Nella terza fase dello schema operativo dei metodi FCS si procede verso la valutazione della significatività statistica dei risultati ottenuti al termine della fase precedente. A tal fine, la procedura farà riferimento ad una di due possibili ipotesi nulle, ognuna facente capo ad un criterio differente. Una delle due, detta ipotesi nulla "self-contained", afferma che nessun gene incluso in un dato pathway è differenzialmente espresso; dunque, un test basato su questo tipo di ipotesi nulla confronterà l'insieme dei geni di un certo pathway con sé stesso, non tenendo in considerazione i geni che non sono parte di quel dato pathway. L'altra, detta ipotesi nulla "competitiva", afferma che i geni di un dato pathway vengono espressi differenzialmente al massimo con la stessa probabilità con cui geni non presenti nello stesso pathway vengono espressi differenzialmente; test basati su questo tipo di ipotesi nulla confrontano l'insieme di geni di un dato pathway con un insieme di geni che non appartengono allo stesso pathway. Come si può intuire, le ipotesi nulle "self-contained" sono più restrittive delle competitive, e dunque hanno potenza statistica maggiore. Sebbene gli approcci FCS sono stati strutturati in modo da superare alcune limitazioni proprie dei metodi ORA, anch'essi presentano limitazioni. Anzitutto, similmente a quanto visto per gli approcci ORA, anche in questo l'analisi statistica tratta erroneamente ogni pathway indipendentemente rispetto agli altri. Inoltre, molti metodi FCS eseguono il ranking dei geni di un dato pathway utilizzando i valori delle variazioni dei livelli di espressione, senza però eseguire un ranking complessivo di tutti i geni. Ad esempio, supponendo di avere due geni X e Y, inclusi in due pathway distinti e con variazione del livello di espressione di 2-fold e 10-fold rispettivamente. Se durante la fase di ranking questi due geni verranno classificati in egual maniera all'interno dei rispettivi pathway, il metodo li tratterà alla stessa maniera [100]. Un'ulteriore limitazione degli approcci ORA e FCS risiede nel fatto che questi identificano pathway significativamente perturbate considerando solamente il numero di geni contenuti in un pathway o la co-espressione genica, mentre non tengono in considerazione le connessioni tra geni-proteine nel sistema cellulare. Di conseguenza, anche nel caso in cui le relazioni tra geni di un pathway dovessero essere ridisegnate o modificate (modifiche topologiche) senza però modificare l'insieme di geni del dato pathway, questi approcci continuerebbero a dare sempre gli stessi risultati. Questo limite viene superato dagli approcci PTB (detti "di terza generazione"), che sono stati implementati per tenere conto, durante l'analisi, delle informazioni aggiuntive fornite da knowledgebase quali KEGG, Reactome e PantherDB [100,101].

1.6.3 Approcci PTB

Riguardo allo schema operativo generale, gli approcci PTB seguono essenzialmente la stessa procedura illustrata per gli approcci FCS. La differenza principale tra questi approcci risiede nell'utilizzo della topologia dei pathway per calcolare le statistiche a livello dei dati di espressione in input in fase (1). Ad esempio, analogamente a quanto avviene per i metodi FCS, il metodo SorePAGE [102] calcola in fase (1) la similarità tra i valori di espressione per ogni coppia di geni inclusi in un dato pathway. Tuttavia, a differenza dei metodi FCS, qui la similarità è ottenuta applicando un metodo adattativo (i.e. che seleziona la misura che più si adatta ai dati da trattare) che tiene conto di quattro diverse misure: correlazione, covarianza, distanza del coseno e prodotto di punti. Inoltre, anziché assegnare ugual peso ai geni che hanno ranking uguale, ScorePAGE divide i valori di similarità ottenuti per ogni coppia di geni per la loro distanza all'interno del pathway. In fase (2) i valori di similarità vengono quindi combinati tra loro per ottenere un punteggio a livello di pathway, e in fase (3) l'algoritmo stabilisce la significatività di ciascun punteggio a livello dei pathway, come avviene per i metodi FCS. Un limite del metodo ScorePAGE consiste nel fatto che quest'ultimo è in grado di analizzare esclusivamente vie metaboliche. Sviluppato successivamente rispetto a ScorePAGE, nell'algoritmo SPIA [103] viene introdotto il calcolo di un fattore di perturbazione (PF) per ogni gene e di un fattore d'impatto (IF) per ogni pathway [104,105] (sezione 3.1.2). Il calcolo del PF corrisponde alla fase di computazione della statistica a livello dei dati di espressione eseguita dai metodi FCS. Il PF rappresenta la somma delle variazioni dell'espressione di ogni gene, nonché una funzione lineare dei PF di tutti i geni inclusi in un dato pathway. Poiché il PF di ogni gene è definito da una funzione lineare, l'intero pathway sarà definito come un sistema lineare. Ciò ha il grande vantaggio di permettere l'individuazione di eventuali loop all'interno di un pathway [105]. Come i metodi FCS le statistiche ottenute a livello dei dati di espressione per ogni gene vengono aggregate in un'unica statistica a livello di pathway, alla stessa maniera la somma dei PF dei singoli geni di un pathway definiscono l'IF dell'intero pathway. Così facendo, l'IF tiene in considerazione la struttura di ogni pathway per cui viene calcolato, poiché tiene conto di fattori biologici come variazioni dell'espressione genica, tipo di interazione tra i geni e localizzazione dei geni all'interno del pathway. L'analisi dell'IF presuppone che ogni pathway venga modellato in forma di grafo i cui i nodi rappresentano i geni del pathway, mentre gli archi rappresentano le interazioni tra geni. Sviluppato in parallelo al metodo SPIA, l'algoritmo NetGSA [106] è stato strutturato in modo da tener conto, in fase (1), dei cambiamenti nella correlazione dei valori di espressione tra coppie di geni, e dunque nella struttura della rete di interazioni, al cambiare delle condizioni sperimentali. Sebbene anche questo approccio, come SPIA, rappresenta l'espressione genica come una funzione lineare delle espressioni degli altri geni inclusi nella rete di interazioni,

esso ne differisce per due aspetti. Anzitutto, NetGSA tiene in considerazione l'espressione basale di ogni gene in questione rappresentandola in forma di variabile latente nel modello computazionale. In secondo luogo, in fase (2) l'algoritmo richiede che i pathway vengano rappresentati come grafi diretti aciclici (i.e. che non presentano loop al loro interno). Nel caso in cui un pathway contenga uno o più loop, l'algoritmo richiede l'aggiunta di ulteriori variabili latenti nel modello computazionale, che però pregiudicherebbero la computazione delle statistiche per i geni inclusi nel loop. Questa limitazione non si presenta nel caso dell'IF del metodo SPIA. Ulteriori sviluppi sono stati ottenuti dal metodo PARADIGM [107], in cui i dati di ogni singolo gene ricavati dall'esame di ogni singolo campione sono integrati con un insieme di quattro parametri biologici differenti che ne descrivono il numero di copie sul DNA, i livelli del relativo mRNA e proteina, e i livelli di attività della proteina in questione, oltre che le informazioni sulle interazioni tra geni-proteine contenute nel grafo del pathway. PARADIGM è capace di predire il grado di alterazione dell'attività di un dato pathway grazie all'impiego di un algoritmo probabilistico basato su modello grafico (PGM, i.e. un modello probabilistico per il quale un grafo esprime la struttura della dipendenza condizionale tra variabili). Infine, tra i metodi PTB più recentemente sviluppati, alcuni tengono in considerazione importanti eventi regolatori post-trascrizionali causati da specifiche classi di ncRNA, in particolare miRNA. Tra questi metodi vanno certamente annoverati Micrographite [108] e MITHrIL [109]. In un'analisi comparativa, quest'ultimo ha restituito risultati maggiormente accurati rispetto a quelli restituiti dai metodi SPIA, PARADIGM e Micrographite, ed è stato estesamente descritto nella sezione 3.1.2.

Negli ultimi anni, diversi lavori scientifici hanno posto il loro focus sullo studio di metodi che permettono l'identificazione di sottostrutture sovrarappresentate in una certa condizione all'interno dei pathway. Queste sottostrutture (moduli) dei pathway sono dette subpathway (e dunque i metodi sono sub-pathway topology based, SPTB), e rappresentano sotto-grafi interni al pathway formati da almeno due nodi (geni) connessi tra loro. L'idea centrale di queste metodologie è che, poiché i pathway biologici presentano un'elevata ridondanza di diversi moduli e poiché la risposta di uno o più moduli al variare dello stato fisiopatologico induce la risposta del pathway che li contiene, un'analisi SPTB basata sull'identificazione di sottostrutture sovrarappresentate restituirà risultati potenzialmente più accurati rispetto a quelli ottenuti con metodi PTB. Lo schema operativo generale di alcuni metodi di estrazione di subpathway sovrarappresentati in campioni biologici è stato brevemente descritto nel capitolo 2.

Sebbene gli approcci PTB si differenziano in più parti del loro schema operativo, essi presentano alcune limitazioni comuni. Una di queste riguarda il fatto che, nella realtà dei fatti, la topologia dei pathway varia a seconda dello specifico contesto cellulare e dello stato di differenziamento cellulare.

Queste informazioni raramente sono rese disponibili nei knowledgebase ad oggi esistenti, ed anche nei casi in cui vengono riportate esse risultano alquanto frammentarie [110]. Un'altra limitazione include l'incapacità di considerare il cross-talk tra pathway a causa della scarsità di informazioni contenuta negli knowledgebase circa le connessioni tra pathway. Tuttavia, quest'ultima limitazione viene parzialmente superata dagli approcci basati sull'analisi di subpathway, che mettono in correlazione tutti i pathway contenenti la sottostruttura sovrarappresentata in questione.

2. Obiettivi

Lo scopo del presente elaborato è la descrizione di due metodi top-down, PTB, per analisi dei pathway, sviluppati dal gruppo di ricerca guidato dal Professore A. Ferro, Università degli Studi di Catania. I metodi sono stati nominati rispettivamente SPECifIC (SubPathway ExtraCtor and enrICher) e PHENSIM (PHENotype SIMulator).

SPECifIC [111] rappresenta un'estensione dell'algoritmo MITHrIL (Mirna enrIched paTHway Impact anaLysis), precedentemente sviluppato dallo stesso gruppo di ricerca e costruito sulla base del metodo elaborato dal gruppo di ricerca di S. Draghici [103–105]. Il metodo SPECifIC rientra nella sottocategoria di approcci SPTB e implica l'esecuzione della computazione attraverso sei fasi distinte: (i) costruzione del meta-pathway; (ii) calcolo della perturbazione tramite l'esecuzione dell'algoritmo MITHrIL; (iii) calcolo del p-value per l'individuazione di sottostrutture (subpathway) statisticamente significative; (iv) selezione semi-automatizzata dei nodi d'interesse; (v) estrazione delle sottostrutture (sub-pathways), e (vi) enrichment analysis. Ognuna di queste fasi verrà descritta nella sezione 3.1.

Al fine di valutarne le performance, SPECifIC è stato confrontato con tre altre metodiche SPTB: Subpathway-GM, Subpathway-GMir, e DEsubs. Le metodiche erano state precedentemente sviluppate e pubblicate da altri gruppi di ricerca, e il loro schema operativo è stato brevemente descritto di seguito. Subpathway-GM [112] sfrutta la similarità strutturale (topologica) interna ai pathway per identificare ed estrarre subpathway metabolici d'interesse a partire dalle informazioni fornite dall'utente su geni e/o metaboliti, che dunque rappresenteranno i nodi di partenza. Una volta inseriti questi dati iniziali, l'algoritmo di Subpathway-GM opera passando attraverso tre fasi operative: (i) mappatura dei geni e dei metaboliti d'interesse su grafi rappresentanti vie metaboliche; (ii) localizzazione dei subpathway contenuti nelle vie metaboliche sulla base dei nodi di partenza; (iii) valutazione della significatività statistica di ogni sottostruttura identificata mediante l'applicazione di un test ipergeometrico. Similmente a quanto detto per Subpathway-GM, Subpathway-GMir [113] ha lo scopo di individuare alterazioni significative su subpathway metabolici basandosi sulla similarità topologica dei subpathway. Rispetto a Subpathway-GM, qui la principale innovazione consiste nel fatto che l'algoritmo tiene conto della regolazione dell'espressione genica mediata dai miRNA. La procedura seguita dall'algoritmo è simile a quella descritta per Subpathway-GM, e consiste in quattro principali fasi operative: (i) ricostruzione delle vie metaboliche messe a disposizione da KEGG e annotazione di ogni via con le interazioni miRNA-gene (solo per interazioni validate con metodiche low-throughput); (ii) mappatura dei geni e dei miRNA d'interesse sulle vie metaboliche precedentemente ricostruite; (iii) identificazione dei subpathway, annotati con i relativi miRNA, contenuti nelle vie metaboliche ricostruite sulla base dei

nodi d'interesse; (iv) valutazione della significatività statistica di ogni sottostruttura identificata mediante l'applicazione di un test ipergeometrico. Infine, DEsubs [114] include un framework esteso e personalizzabile grazie a un'ampia gamma di modalità operative, permettendo all'utente di procedere con un approccio adattativo caso-specifico. In linea generale, comunque, il metodo si sviluppa in quattro fasi differenti: (i) conversione dei pathway di KEGG in una rete di pathway che conserva la topologia e il flusso d'informazione originale; (ii) mappatura dei dati processati di RNA-seq (in input) sui nodi della rete e filtraggio dei nodi mediante due procedure di pruning basate, rispettivamente, su significatività statistica e su evidenze biologiche note; (iii) estrazione dei subpathway sulla base di proprietà topologiche specificate dall'utente; (iv) utilizzo di un test ipergeometrico per stimare l'associazione dei subpathway estratti con vari termini biologici e farmacologici.

Questi quattro metodi sono stati testati su dati estratti da campioni di tumore al seno (BRCA) e al colon (COAD) forniti dal database TCGA (The Cancer Genome Atlas) [115]. Le due casistiche (BRCA e COAD) sono state scelte sulla base di alcuni criteri logici. Anzitutto, BRCA e COAD rientrano nella categoria dei tumori più diffusi a livello mondiale e causano la morte di decine-centinaia di milioni di pazienti oncologici l'anno [116,117]. In secondo luogo, i due tumori in questione presentano diverso grado di eterogeneità. ed elevata omogeneità dal punto di vista molecolare. BRCA è una patologia oncologica ad elevata eterogeneità, classificata in più sottotipi, dove ogni sottotipo coinvolge uno specifico set di geni implicati in processi molecolari ben precisi. Secondo le linee guida più consolidate, la definizione del sottotipo di BRCA fa primariamente affidamento sull'espressione del recettore degli estrogeni (ER), del recettore del progesterone (PR), del recettore di tipo 2 del fattore di crescita epiteliale (ERBB2), e della proteina citocheratina (CK) [118]. Tuttavia, è noto che prendendo in considerazione il carico mutazionale di geni oncosoppressori coinvolti in vie di segnalazione, come ad esempio le vie della proteina p53, delle chiansi MAP (MAPK), della fosfatidilinositolo-3-chinasi (PI3K) e della proteina del retinoblastoma (pRB), è possibile ottenere una suddivisione più accurata delle tipologie tumorali di BRCA [119]. A differenza del BRCA, COAD presenta un notevole grado di omogeneità. Inoltre, è noto che nonostante il COAD presenti differenze epidemiologiche e istologiche rispetto all'adenocarcinoma al retto (READ) [120], la caratterizzazione molecolare di questi due tumori in stadio non avanzato ha rivelato differenze non significative sia nei pattern di numeri di copie di specifici geni sia nei livelli di espressione di set ben definiti di mRNA e miRNA (dunque nessuna distinzione significativa tra COAD e READ) [121]. Dunque, ciò che ci si aspetta di ottenere al termine dell'analisi SPTB è: (i) l'identificazione di un numero relativamente elevato di pathway significativamente coinvolte nella patologia del BRCA, di cui almeno una frazione dovrà corrispondere con quelle precedentemente identificate tramite studi

molecolari, e (ii) un numero più ristretto di pathway nel caso del COAD, dove, anche qui, almeno una frazione dei pathway restituiti dal tool dovrà coincidere con quelli identificati da studi molecolari.

A differenza di SPECifIC, l'algoritmo di PHENSIM si discosta da quello di MITHrIL e rientra nella categoria degli approcci PTB. Il metodo PHENSIM elabora la propria analisi attraverso l'esecuzione di quattro fasi distinte: (i) costruzione del meta-pathway; (ii) calcolo delle probabilità di attivazione o inattivazione per ogni nodo del meta-pathway; (iv) calcolo dei rapporti delle probabilità rispetto al modello nullo e degli activity score; (v) determinazione dei p-value a livello di pathway. Ognuna di queste fasi verrà descritta nella sezione 3.2. Comunque, va sottolineato che tool bioinformatico PHENSIM è attualmente in fase di rifinitura.
Al fine di valutarne le performance, PHENSIM è stato testato sulla base di quattro casistiche ricavate dalla letteratura, ovvero: (i) trattamento con metformina [122]; (ii) trattamento di cellule del tessuto mammario con everolimus [123,124]; (iii) impatto dei miRNA esosomiali derivati da cellule di leucemia mieloide acuta (LMA) nelle cellule riceventi [125–127]; (iv) efficacia di un modello di letalità sintetica in sei linee cellulari tumorali [128].

(i) La metformina, un derivato del biguanide, è un farmaco ampiamente utilizzato per trattare il diabete mellito di tipo 2. Ad oggi è noto che questo farmaco presenta due target molecolari principali, di cui uno diretto e uno indiretto. Il target molecolare diretto è rappresentato dal complesso 1 della catena di trasporto degli elettroni nel sistema della fosforilazione ossidativa. Qui la metformina causa il disaccoppiamento della catena di trasporto degli elettroni, causando un decremento della produzione di ATP a favore di un incremento dei livelli di AMP citoplasmatici. Come conseguenza, la chiansi LKB1 (STK11) viene attivata. L'azione congiunta dell'incremento del rapporto AMP/ATP e dell'attivazione di LKB1 causano l'attivazione del driver molecolare AMPK (proteina chinasi attivata da AMP), che rappresenta uno dei principali regolatori del metabolismo energetico cellulare. Tra i ruoli regolatori svolti nel sistema cellulare, AMPK è un regolatore positivo (diretto) del complesso TBC1D7-TSC1-TSC2 e un regolatore negativo (diretto e indiretto) del macchinario molecolare mTORC1 [122]. Il target molecolare indiretto della metformina è invece rappresentato dall'insulina (Ins) e dal fattore simile all'insulina 1 (IGF-1). A causa di un meccanismo molecolare ancora ignoto, la somministrazione di insulina causa una sottoespressione di questi due peptidi, nonché dei loro rispettivi recettori di membrana. Notoriamente, questi due target molecolari si trovano a monte di pathway di segnalazione molto importanti, capaci di regolare stato

energetico, trascrizione genica e proliferazione/apoptosi cellulare. Tra questi, degni di nota sono i pathway Ras/Raf/MAPK/MEK/ERK e PI3K/Akt/mTOR [122] (Figura 2.1). Grazie a queste caratteristiche, la metformina ha mostrato di avere proprietà anti-tumorali.

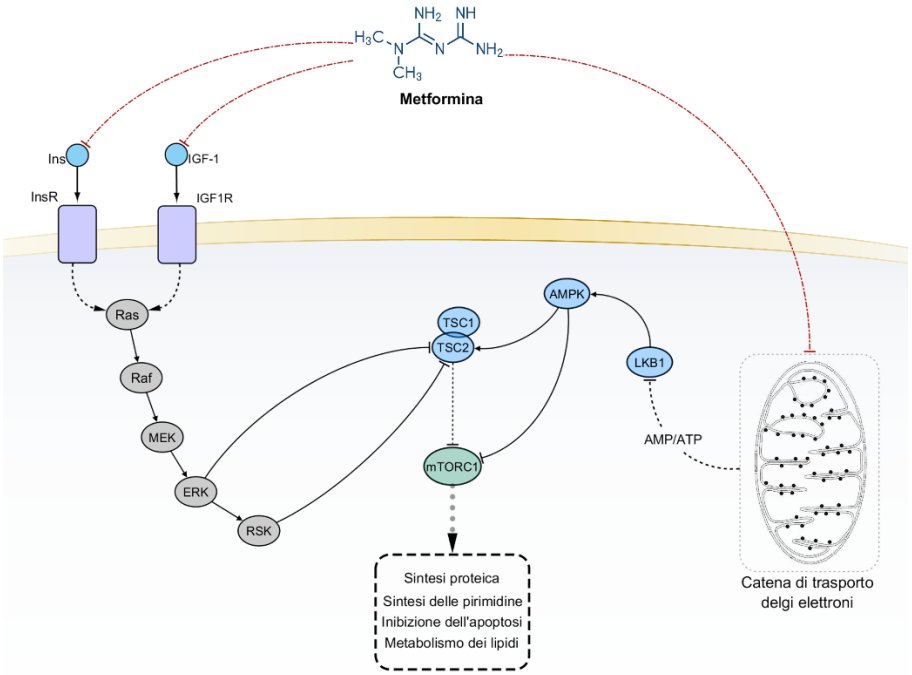


Figura 2.1 Meccanismi molecolari innescati dall'uptake cellulare del biguanide metformina (maccanismo diretto: disaccoppiamento della catena di trasporto degli elettroni; meccanismo indiretto: sotto-espressione di insulina, IGF-1 e dei relativi recettori). La figura mostra i pathway a valle di tali meccanismi, notoriamente coinvolti in processi tumorali.

(ii) Everolimus (RAD001, Afinitor®), un analogo della rapamicina, è un inibitore allosterico del macchinario molecolare mTORC1, che rappresenta uno dei principali regolatori di diverse vie anaboliche cellulari, della trascrizione di vari geni lipogenici e della proliferazione cellulare. mTORC1 è iperattivato in una elevata percentuale di tumori, e ricopre un ruolo di grande rilevanza nella loro progressione [129–131] (Figura 2.2). Per questo motivo, l'ente americano FDA (Food and Drug Administration) ha approvato l'utilizzo di questo farmaco per il trattamento di diverse tipologie tumorali, spesso in combinazione con altri chemioterapici (si veda https://www.fda.gov/drugs/informationondrugs/approveddrugs/ucm488028.htm). Tra i tumori trattati, rientrano anche diversi casi di cancro alla mammella. Alcuni interessanti studi sperimentali *in vitro* svolti recentemente [123,124] hanno riportato

diversi effetti sull'espressione genica e sull'attività chinasica intracellulare causati dalla somministrazione di everolimus in diverse linee cellulari continue di tumore al seno. Tra i risultati più importanti segnalati da questi lavori, vi sono certamente la sottoregolazione dell'attività chinasica di S6K, la repressione dei fattori di traduzione e la sottoregolazione dell'espressione del recettore degli estrogeni ERBb2.

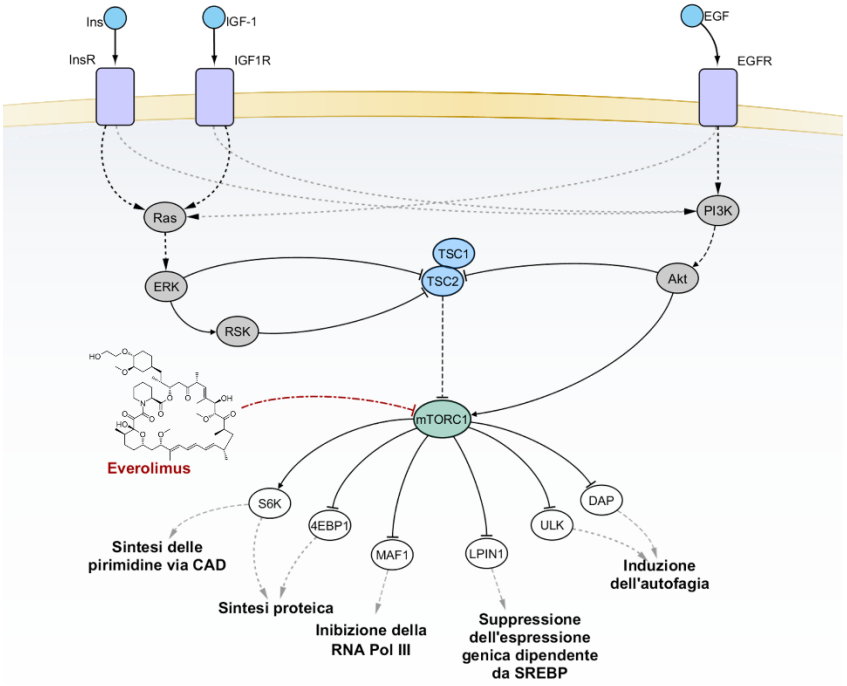


Figura 2.2 Blocco del macchinario molecolare mTORC1 a seguito dell'uptake cellulare di everolimus (RAD001), analogo della rapamicina. In figura sono mostrate le maggiori vie di segnalazione a monte e a valle rispetto a mTORC1.

(iii) Gli esosomi secreti da cellule di LAM trasportano diverse proteine e acidi ribonucleici prodotti nella cellula d'origine. Attraverso il meccanismo di comunicazione paracrina mediato da esosomi, le cellule di LAM modulano le proprietà del microambiente del midollo osseo, alterando i processi di differenziazione cellulare e l'egresso delle cellule progenitrici delle cellule ematopoietiche (HSPCs) [127,132]. Alcuni studi analitici condotti in vitro e in vivo hanno rivelato la presenza di specifici pattern di miRNA all'interno di cellule di LAM [125,126,133]. In particolare, miR-150-5p e miR-155-5p sono stati trovati in elevate concentrazioni per tutte le linee cellulari esaminate. Una più completa caratterizzazione del carico esosomiale di miRNA, comunque, è stata effettuata solo per la linea cellulare Molm-14, che si è mostrata più aggressiva rispetto alle altre [125]. Tuttavia, gli autori non hanno focalizzato l'attenzione sui potenziali effetti a valle

causati da questa comunicazione paracrina, che invece sono stati analizzati più recentemente su modello murino da un altro gruppo di ricerca [127]. Questi consistono, sostanzialmente, in sovra- o sotto-regolazione di fattori trascrizionali e peptidi appartenenti alla superfamiglia delle citochine (Figura 2.3).

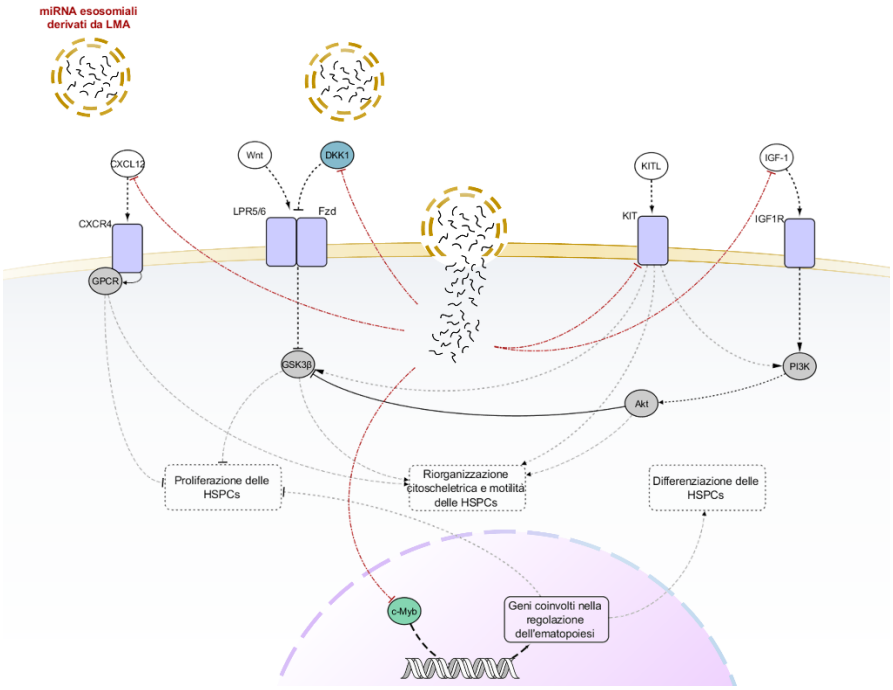


Figura 2.3 In figura sono mostrati alcuni dei meccanismi molecolari coinvolti nella differenziazione e nella regolazione dello stato di quiescenza delle cellule staminali-progenitrici delle cellule ematopoietiche (HSPCs). Dati sperimentali hanno mostrato che il carico esosomiale derivante da cellule di leucemia mieloide acuta (LMA), e in modo particolare i miRNA esosomiali, alterano le espressioni geniche di diverse proteine coinvolte in tali meccanismi.

(iv) La citochina TNFα, appartenente alla superfamiglia dei fattori tumorali della necrosi, ricopre un ruolo importante nella risposta immunitaria innata, essendo adoperato nei processi di infiammazione acuta. Questo peptide può anche essere utilizzato nell'induzione dei processi apoptotici; tuttavia, diversi tipi cellulari manifestano meccanismi di resistenza contro l'apoptosi indotta da TNFα. Uno studio recentissimo (PNAS in press) ha dimostrato che uno dei checkpoint del blocco dell'apoptosi indotta da TNFα è la chinasi MAP3K8, detta anche TPL2, che ha un ruolo importante nei processi infiammatori e nell'oncogenesi. È stato altresì dimostrato nello stesso studio che il knockdown (KD) di TPL2 ha come conseguenza l'inibizione dell'espressione di miR-21-5p e la susseguente sovraregolazione di caspasi-8 (Casp8), target validato di miR-21-5p. A seguito di quanto detto, si è visto che in cellule tumorali KD per TPL2 trattate con

TNFα si ha una forte sovraregolazione di Casp8 attiva, e quindi linduzione forzata dell'apoptosi caspasi-dipendente. La marcata attivazione di Casp8 non è dovuta solamente all'inibizione dell'espressione di miR-21-5p, ma anche alla sottoregolazione di cFLIP, inibitore del taglio proteolitico che comporta l'attivazione di Casp8 – precisamente che comporta il passaggio pre-caspasi-8 → caspasi-8 enzimaticamente attiva (Figura 2.4). L'attivazione di Casp8 promuove anche l'attivazione del pathway di apoptosi mitocondriale, sebbene alcune molecole coinvolte nell'attivazione di tale pathway mostrano un pattern di espressione genica inverso rispetto a quello atteso. Questo modello di letalità sintetica è stato testato e validato mediante esperimenti *in vitro* su più linee cellulari tumorali continue, appartenenti a modelli tumorali indipendenti, ed è stato nominato "letalità indotta da TNFα/siTPL2". Tuttavia, è da sottolineare che questo modello di letalità indotta non mostra uguale efficacia in tutte le linee cellulari, ma alcune linee cellulari sono state identificate come resistenti al trattamento [128].

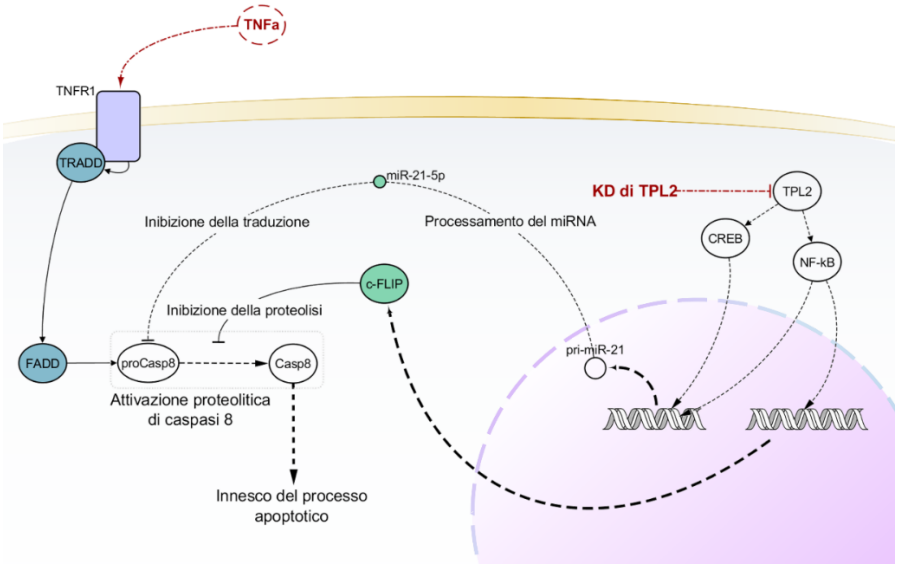


Figura 2.4 Schema riassuntivo dei meccanismi molecolari coinvolti nella letalità sintetica data dal knockdown di TPL2 e dalla somministrazione di TNFα.

3. Descrizione dei Metodi:

3.1 SPECifIC

Come accennato al capitolo 2, l'algoritmo di SPECifIC si compone di sei fasi operative, descritte qui di seguito.

3.1.1 Costruzione del meta-pathway

SPECifIC basa l'analisi predittiva sui pathway biologici forniti dal database KEGG. Con l'obiettivo di ottenere una maggiore completezza dei pathways ed incrementare le potenzialità predittive del tool nonché l'accuratezza delle predizioni stesse, il metodo tiene conto (a) delle interazioni miRNA::mRNA e (b) della dipendenza dell'espressione dei miRNA da specifici fattori della trascrizione (TFs). Le interazioni miRNA::mRNA sono state ricavate dai database miRTarBase (release 2.5) [134] e miRecords (aggiornato al 27 Aprile, 2013) [135], mentre le relazioni TFs-miRNA sono state ricavate dal database TransmiR (release 1.0) [136]. Ai fini della computazione negli step successivi, si è resa necessaria la standardizzazione degli ID degli elementi presi in considerazione. Nel caso dei miRNA, la mappatura dei rispettivi ID è stata eseguita utilizzando come database di riferimento miRBase (release 20) [137,138]. Nel caso dei geni target e dei TFs, la mappatura è stata eseguita in due fasi: (i) l'ID di ogni gene è stato anzitutto convertito nel corrispettivo Entrez ID; (ii) poi, mediante l'intervento dell'Application Programming Interface in REST-style di KEGG (KEGG REST-style API), ogni Entrez ID è stato convertito nel corrispettivo KEGG ID. La lista ottenuta al termine della procedura è stata poi filtrata al fine di rimuovere i duplicati. Il meta-pathway di KEGG ottenuto tramite rimozione dei duplicati è stato quindi annotato con i miRNA tenendo in considerazione le interazioni riportate dai suddetti database. I miRNA ricavati da tale procedura sono quindi stati impostati come nuovi nodi all'interno del network biologico di KEGG e connessi ai nodi circostanti mediante archi direzionali del tipo: (a) miRNA-gene (inibitorio); (b) TF-miRNA (attivante). La procedura appena descritta ha permesso la costruzione di un knowledgebase contenente 10,537 interazioni sperimentalmente validate tra 385 miRNA e 3,080 geni [109]. Infine, è stato applicato un depth-first search algorithm (DFS, in italiano "ricerca in profondità") [139] che permette di acquisire informazioni circa gli endpoints putativi contenuti in ogni pathway di KEGG, marcandone automaticamente i nodi (geni) posti alla fine di ogni catena di interazioni. Per effettuare la ricerca di questi nodi, l'algoritmo DFS parte da un nodo random all'interno del grafo, per poi percorrere il cammino dettato dagli archi direzionati (interazioni) tra nodi, fino ad arrivare a nodi

dai quali nessun altro nodo a valle può essere raggiunto. L'algoritmo DFS viene iterato finché tutti i nodi siano stati analizzati.

3.1.2 Esecuzione dell'algoritmo MITHrIL: calcolo della perturbazione e dell'impatto

Partendo dai concetti relativi al modello predittivo presentato dal gruppo di Draghici in [103], l'analisi della perturbazione dei pathway combina due evidenze emergenti dall'analisi dei microarray (o da analisi NGS) nei confronti casi/controllo: (i) la sovrarappresentazione di geni differenzialmente espressi (DEGs) in un dato pathway e (ii) la perturbazione anomala dello stesso pathway in termini di propagazione dell'alterazione dei tassi di espressione dei suoi geni (misurati tramite i valori di fold-change). Questi due aspetti vengono descritti dai valori di due probabilità indipendenti, $\mathbf{p_{nde}(P_i)}$ e $\mathbf{p_{pert}}$.

In termini statistici, il valore di $\mathbf{p_{nde}(P_i)}$ rappresenta la probabilità di ottenere un numero di DEGs, in un dato pathway P_i, almeno uguale a quello osservato, cioè:

$$p_{nde}(P_i) = p(X \geq N_{de} \mid H_0). \tag{1.1}$$

Qui, N_{de} rappresenta il numero di DEGs osservato, e H_0 rappresenta l'ipotesi nulla, secondo la quale i geni segnalati come DEGs in P_i sono risultati tali per caso. I valori di $p_{nde}(P_i)$ vengono calcolati assumendo che N_{de} segua una distribuzione ipergeometrica a tre parametri x, y e z, dove: (x) è il numero di tutti i geni $g \in P_i$ presenti sull'array; (y) è il numero di tutti i geni $g \notin P_i$ presenti sull'array; (z) è il numero totale di DEGs presenti nell'array. Se H_0 risultasse vera, ciò implicherebbe che il pathway biologico P_i non è rilevante per la condizione fenotipica presa in esame. In altre parole, $p_{nde}(P_i)$ ha lo scopo di descrivere la significatività statistica della sovrarappresentazione di P_i data l'espressione differenziale dei suoi geni.

$\mathbf{p_{pert}}$ è invece calcolata tenendo in considerazione l'ammontare delle perturbazioni misurate in ogni pathway biologico. Dunque, per ogni nodo (gene) incluso nel generico pathway P_i, l'algoritmo calcola un *Perturbation factor* (**PF**). Quest'ultimo rappresenta una stima di quanto l'attività del gene/proteina in questione è alterata in relazione (a) al suo grado di espressione/attività e (b) all'attività dei nodi adiacenti a monte. Un valore positivo (o negativo) del PF indica che il gene corrispondente sarà sovra-regolato (o sotto-regolato) con una certa probabilità. Più precisamente, sia g un nodo qualsiasi del pathway P_i. Matematicamente, il PF è definito come:

$$PF(g_i, P_i) = \Delta E(g_i) + \sum_{u \in U(g_i, P_i)} \beta(u, g) \cdot \frac{PF(u, P_i)}{\sum_{d \in D(u, P_i)} |\beta(u, d)|}$$

(1.2)

dove ΔE(g) rappresenta la misura (normalizzata) dell'alterazione dell'espressione genica del gene $g \in P_i$, espressa come log fold-change; U(g, Pi) e D(g, Pi) rappresentano l'insieme dei geni $u_i \in P_i$ e $d_i \in P_i$, a monte e a valle di n, rispettivamente. β(u,g) è una funzione che indica la forza e il tipo di interazione tra i nodi $u \in U(n,P_i)$ e $g \in P_i$. In particolare, il valore assoluto di β è la quantificazione della forza dell'interazione tra $u \in U(n,P_i)$ e $g \in P_i$, mentre il segno di β indica il tipo di interazione tra u_i e g. Un valore negativo di β descrive un'interazione inibitoria, mentre un valore positivo descrive un'interazione attivante. Nella sua totalità, il secondo termine dell'equazione (1) rappresenta la somma dei PF dei geni adiacenti a monte del gene g, normalizzati per la sommatoria della forza delle interazioni con i geni a valle di ognuno dei geni u_i considerati. Essenzialmente, quindi, l'equazione (1.2) descrive il PF per il generico gene $g \in P_i$ come una funzione lineare dei fattori di perturbazione di tutti i $g_i \in P_i$.

Dopo aver calcolato tutti i PF dei nodi di un pathway, l'algoritmo calcola, per ogni pathway, due valori noti come *Impact Factor* (**IF**) e *Accumulator* (**Acc**).

L'**IF** di un pathway riflette l'importanza delle alterazioni riportate/predette in seno al pathway stesso – maggiore è il valore del IF, maggiore sarà l'importanza (cioè l'impatto) delle alterazioni dei nodi del pathway. L'IF viene calcolato dall'algoritmo secondo la formula:

$$IF(P_i) = \log \frac{1}{p_{nde}(P_i)} + \frac{\sum_{g_i \in P_i} |PF(g_i, P_i)|}{\overline{|\Delta E|} \cdot N_{de}(P_i)}$$

(1.3)

dove $p_{nde}(P_i)$ è la probabilità di ottenere un numero di DEGs almeno uguale a quelli osservati in P_i (come accennato sopra); $\overline{|\Delta E|}$ è la media del log fold-change P_i.

L'**Acc** indica il livello totale della perturbazione di un pathway e mostra la tendenza generale delle perturbazioni dei geni inclusi nel pathway in questione – Acc positivo indica una sovra-regolazione generalizzata; Acc negativo indica sotto-regolazione generalizzata. Il valore di Acc sarà anche accompagnato da un p-value aggiustato per la moltiplicità dei test, e da un False Discovery Rate (FDR) [140]. L'Acc viene invece calcolato secondo le due formule:

$$Acc_{gene}(P_i) = \sum_{g_i \in P_i^g} [PF(g_i, P_i) - \Delta E\ (g_i)]$$

(1.4)

$$Acc_{mir}(P_i) = \sum_{m_i \in P_i^m} [PF(m_i, P_i) - \Delta E\ (m_i)]$$

(1.5)

dove P_i^g e P_i^m rappresentano, rispettivamente, l'insieme dei geni e dei miRNA inclusi nel pathway P_i, e $\Delta E(g_i)$ e $\Delta E(m_i)$ rappresentano, rispettivamente, il log fold-change del generico gene $g_i \in P_i$ e il log fold-change del generico miRNA $m_i \in P_i$). In (1.4) e (1.5) la sottrazione dei $\Delta E(g_i)$ ($\Delta E(m_i)$) dai valori di $PF(g_i,P_i)$ ($PF(m_i,P_i)$) ha lo scopo di far emergere l'effetto dominante (cumulativo) delle variazioni dell'espressione dei DEGs. Dunque, applicando queste due formule, vengono sommate le perturbazioni di tutti i geni e i miRNA apparteneti al pathway P_i. Dopo aver eseguito il calcolo di queste due misure, l'algoritmo calcola l'**Acc totale**, che dà una stima riguardo allo stato di attività del pathway P_i (sovraregolazione o sottoregolazione). La principale innovazione dell'algoritmo risiede, appunto, nell'inclusione degli effetti inibitori dei miRNA sull'espressione genica durante la computazione. Ciò incrementa l'accuratezza della simulazione. Dunque si avrà:

$$Acc_{tot}(P_i) = Acc_{gene}(P_i) - Acc_{mir}(P_i) - E[Acc_{tot}(P_i)]$$

(1.6)

dove $E[Acc_{tot}(P_i)]$ è una stima del valore atteso della distribuzione di tutti gli Acc calcolati per il pathway P_i.

A questo punto, viene calcolata la $\mathbf{p_{pert}}$ come la probabilità di osservare una perturbazione accumulata totale maggiore o uguale, per caso (Acc_{rand}), rispetto a quella calcolata ($Acc_{tot}(P_i)$). In termini matematici, è:

$$p_{pert} = p(Acc_{rand} \geq Acc_{tot} \mid H_0)$$

(1.7)

p_{pert} viene calcolata facendo uso del metodo di bootstrap. Durante questa procedura, viene assegnata una posizione random, all'interno di P_i, ad un numero di DEGs$\in(P_i)$ selezionati casualmente, pari a $N_{de}(P_i)$. A tali geni vengono quindi assegnati in maniera casuale dei valori di log fold-change inclusi

nel range dei valori osservati per i geni rilevati dall'array come differenzialmente espressi. Ciò permette la determinazione empirica della distribuzione nulla dei valori di Acc_{rand}.

3.1.3 Calcolo del p-value per l'individuazione di sottostrutture statisticamente significative

Dopo aver eseguito le suddette computazioni mediante l'algoritmo MITHrIL, il metodo SPECifIC procede verso la ricerca e l'estrazione di sottostrutture (in termini di sotto-grafi) significativamente perturbate all'interno nel meta-pathway. Questa procedura si basa sulla significatività statistica calcolata per ogni nodo del meta-pathway, che consiste nella probabilità (p-value) di osservare un PF minore rispetto a quello calcolato per il modello nullo, come spiegato di seguito. Sia n un generico nodo del meta-pathway e sia PF(n) il suo fattore di perturbazione. Similmente a quanto mostrato per l'equazione (1.7), i valori di log fold-change vengono ridistribuiti in maniera random ai nodi del meta-pathway, al fine di ottenere un insieme delle perturbazioni random $PF^R(n) = \{PF^R1(n), PF^R2(n), PF^R3(n),\dots PF^Rk(n)\}$. Il p-value per il nodo n sarà quindi calcolato secondo la formula:

$$p(n) = \frac{|\{x \in PF^R(n) | \; x \; \geq PF(n)\}|}{k} \tag{1.8}$$

I valori delle probabilità così ricavate vengono quindi aggiustati per la molteplicità dei test e tenuti in considerazione dall'algoritmo per le computazioni degli step successivi.

3.1.4 Selezione dei nodi d'interesse (NoIs)

La selezione dei NoIs è di fondamentale importanza per il corretto svolgimento della procedura di estrazione delle sottostrutture. Per questa ragione, il tool è stato provvisto di un sistema semi-automatizzato di supporto per la selezione unbiased dei NoIs. In questo caso, all'utente verrà lasciato l'incarico di stabilire le threshold di significatività dei pathways (cioè dei singoli pathways costituenti il meta-pathway) e dei NoIs, inserendole negli appositi campi dell'interfaccia web (sezione 3.1.7). Avendo come punto di riferimento questi due valori di threshold, l'algoritmo procede anzitutto verso la ricerca di tutti i pathways che soddisfano la prima condizione (pathways max p-value). Poi, tutti i nodi costituenti i pathways selezionati dal tool verranno analizzati sulla base dei loro p-values, e tutti i nodi con p-values non significativo vengono scartati. I pathways

filtrati vengono quindi integrati tra loro, e i duplicati vengono rimossi – ugualmente alla procedura eseguita per la generazione del meta-pathway. Viene così generata una lista preliminare di potenziali NoIs. Questa viene ulteriormente filtrata sulla base del secondo valore di threshold precedentemente specificato dall'utente (NoIs max p-value). Alla lista risultante di nodi viene quindi applicata la procedura di estrazione delle sottostrutture esposta nella sezione seguente. I sotto-grafi ottenuti a valle della procedura vengono poi analizzati al fine di identificare eventuali sovrapposizioni. I duplicati vengono così eliminati.

3.1.5 Estrazione dei subpathway

La procedura di estrazione delle sottostrutture specifiche della condizione patologica viene eseguita a partire dal meta-pathway di KEGG annotato con i PF e i p-values calcolati nelle fasi precedenti. Tale procedura è in grado di estrarre cinque tipologie differenti di sottostrutture a partire dai nodi d'interesse (NoIs) selezionati dall'utente o per mezzo della procedura semi-automatizzata: (i) paths (cammini); (ii) trees (alberi); (iii) neighborhoods (vicinato); (iv) sotto-grafi indotti e (v) comunità di sotto-grafi indotti. Cammini, alberi e vicinato vengono ottenuti facendo correre sul meta-pathway una versione modificata dell'algoritmo di ricerca in ampiezza (in inglese breadth-first search algorithm, BFS) [141]. BFS è un algoritmo di ricerca per grafi che partendo da un nodo, detto "sorgente", permette di cercare il cammino fino ad un altro nodo scelto e connesso al nodo sorgente (nel caso in cui i nodi scelti siano più di uno, permette di trovarne i cammini che li collegano). L'algoritmo è stato modificato in modo da tener conto dei p-value assegnati ai nodi: ogni cammino a partire dal nodo sorgente viene esteso con un nuovo nodo se, e solo se, il p-value assegnato a quest'ultimo è maggiore rispetto ad un valore soglia (threshold) stabilito dall'utente. I cammini identificati tramite l'algoritmo BFS formano così un albero la cui radice è il nodo sorgente. Le topologie dei cammini e degli alberi vengono quindi accorpati a formare componenti connesse tra loro, ottenendo così sotto-grafi indotti e comunità. Per ogni sotto-pathway identificato, l'algoritmo calcola un **p-value combinato ψ**, ottenuto aggregando i p-values dei singoli nodi inclusi nella sottostruttura. Il calcolo viene eseguito adoperando il Metodo Empirico di Brown [142] , un adattamento empirico del metodo di Brown [143] che ne permette l'applicazione a dataset di grandi dimensioni e contenenti dati correlati tra loro, tipicamente ottenuti con l'utilizzo di metodi high-throughput della biologia molecolare. In breve, il metodo di Brown rappresenta un'estensione del metodo di Fisher [144] finalizzata a combinare tra loro p-values dipendenti, ammesso che sia nota la covarianza. Secondo il metodo Fisher, dato un insieme di k p-values tra loro indipendenti, con p_i che rappresenta l'i-esimo p-value, si ha che:

$$\psi = -2 \sum_{i=1}^{k} \ln(p_i) \sim \chi^2_{2k}$$

(1.9)

In altri termini, Fisher mostrò che in questo caso ψ segue una distribuzione χ^2 con 2k gradi di libertà. Brown estese (1.9) a casi di p-values dipendenti tra loro ricalcolando la distribuzione χ^2 secondo la nozione:

$$\psi = -2 \sum_{i=1}^{k} \ln(p_i) \sim c\chi^2_{2f}$$

(1.10)

Qui, f rappresenta il numero ricalcolato dei gradi di libertà, ed è ottenuto dividendo il valore atteso di ψ al quadrato ($E[\psi]^2$) per la varianza di ψ (var[ψ]); c rappresenta un fattore di scala ottenuto dal rapporto tra i gradi di libertà del metodo di Fisher e i gradi di libertà del metodo di Brown. In termini matematici è:

$$f = \frac{E[\psi]^2}{var[\psi]} \quad e \quad c = \frac{var[\psi]}{2E[\psi]} = \frac{k}{f}$$

(1.11)

Dunque, adoperando una distribuzione χ^2 ricalcolata con 2f gradi di libertà, SPECifIC calcola il p-value combinato ψ per un generico sotto-pathway estratto dal meta-pathway di KEGG e formato da k nodi. Infine, i p-values ψ ottenuti mediante il Metodo Empirico di Brown vengono aggiustati per la molteplicità dei test mediante la correzione di Holm-Bonferroni [145], al fine di tenere sotto controllo il cosiddetto "family-wise error rate" (FWER).

Alla fine di questo step, tutti i risultati vengono filtrati rimuovendo i sotto-grafi aventi numero di nodi inferiore rispetto alla soglia minima indicata dall'utente. Oltre alla threshold del p-value dei nodi inclusi nel sotto-grafo e al numero minimo di nodi dei sotto-grafi, l'utente può impostare altri valori soglia sulla base dei quali i risultati preliminari verranno filtrati, quali ad esempio il valore soglia di ψ, il valore soglia dei NoIs, e il valore soglia di p_{pert}.

3.1.6 Enrichment analysis

Nell'ultima fase dell'intera procedura del metodo, tutti i sotto-pathway estratti e filtrati secondo i criteri precedentemente esposti vengono esaminati per mezzo di una enrichment analysis. SPECifIC è in grado di tenere in considerazione associazioni con (i) altri pathways, (ii) termini di gene ontology (GO), (iii) patologie e (iv) farmaci. L'enrichment analysis di SPECifIC viene eseguita applicando la metodologia descritta da Li et al. [112]. Dunque, dato un sotto-grafo S del meta-pathway, il p-value **p(S, t)** per un termine t (ad esempio un farmaco) viene calcolato secondo la formula:

$$p(S,t) = 1 - \sum_{i=1}^{k} \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}} \tag{1.12}$$

dove k è il numero di nodi in S annotati con t; M è il numero di nodi del meta-pathway annotati con t; N è il numero di nodi annotati nel meta-pathway; n è il numero di nodi annotati in S. Al termine dei calcoli, tutti i p-values ottenuti vengono aggiustati per la molteplicità dei test utilizzando il metodo Benjamini–Hochberg [146] per tenere sotto controllo il FDR, e tutti i termini aventi p-value maggiore rispetto alla soglia specificata dall'utente vengono esclusi dall'analisi. Un'importante considerazione da tenere a mente è che alcuni termini potrebbero annotare pochi nodi rispetto al totale dei nodi inclusi nella sottostruttura, risultando comunque significativi. Al fine di evitare una tale circostanza, l'algoritmo esclude dall'analisi tutte le annotazioni trovate in meno di un terzo dei nodi inclusi in un subpathway.

3.1.7 Sorgente dei dati di espressione e interfaccia web

La piattaforma di SPECifIC è stata progettata per eseguire analisi specifiche per patologie ben precise sfruttando i dati di espressione forniti da database TCGA. In particolare, dai profili di espressione genica di tutti i pazienti sono stati rimossi tutti quelli per cui nessun controllo era disponibile. Dal profilo di espressione genica di ogni paziente sono stati estrapolati solamente dati relativi a geni codificanti e miRNAs, ottenuti, rispettivamente, con RNASeqV2 e miRNASeq eseguiti su piattaforme Illumina Genome Analyzer and Illumina HiSeq. Alla fine è stato creato un dataset che include 10 patologie oncologiche, suddivise per tipologia e stadio (si veda la Tabella 3.1).

Tabella 3.1 Lista delle tipologie tumorali estratte dal database TCGA. In tabella sono mostrati i codici identificativi per ogni tipologia tumorale, il numero di casi e controlli (campioni tissutali), e le relative sottocategorie tumorali per ogni tumore.

Code	Cancer Type	Control Samples	Case Samples	Case Samples Categories
BLCA	Bladder Urothelial Carcinoma	19	193	Stage I, II, III, IV
BRCA	Breast invasive carcinoma	86	642	Stage I, II, III, IV, X
COAD	Colon adenocarcinoma	8	389	Stage I, II, III, IV
KICH	Kidney Chromophobe	25	66	Stage I, II, III, IV
KIRC	Kidney renal clear cell carcinoma	71	224	Stage I, II, III, IV
LUAD	Lung adenocarcinoma	19	388	Stage I, II, III, IV
LUSC	Lung squamous cell carcinoma	37	247	Stage I, II, III, IV
PRAD	Prostate adenocarcinoma	50	191	Category 6, 7, 8, 9, 10
READ	Rectum adenocarcinoma	3	150	Stage I, II, III, IV
UCEC	Uterine Corpus Endometrial Carcinoma	14	231	Stage I, II, III, IV
	All Samples	332	2721	

Per garantire un'ampia accessibilità e fruibilità del tool, l'algoritmo di SPECifIC è stato implementato in un'applicazione web disponibile all'indirizzo https://alpha.dmi.unict.it/specific/. Il servizio è stato sviluppato facendo uso di PHP (Hypertext Preprocessor – Personal Home Page), HTML (HyperText Markup Language) e JavaScript, mediante l'aiuto fornito dal framework open source di Laravel (disponibile online all'indirizzo https://laravel.com). Per incrementarne le performance, la componente computazionale relativa ai grafi è stata sviluppata in Java, mentre l'enrichment analysis viene implementata utilizzando l'ambiente statistico di R [147]. La visualizzazione grafica dei sotto-grafi viene ottenuta grazie all'utilizzo della piattaforma open source Cytoscape [148]. Inoltre, per minimizzare il tempo d'attesa, tutte le computazioni sono state memorizzate per poter essere riutilizzate. Ciò implica che i dati degli utenti non vengono conservati in memoria. In Tabella 3.2 sono riportate tutte le sorgenti impiegate per l'enrichment analysis.

Tabella 3.2 Lista dei termini utilizzati per l'analisi di arricchimento (enrichment analysis) e relative sorgenti informatiche. Le sorgenti sono state raggruppate per categoria. Per ogni sorgente sono stati riportati nome, numero di termini, e numero di nodi arricchiti.

Category	Source	# Terms	# Nodes
Diseases			
	DisGeNET	7607	2978
	GAD	403	1519
	KEGG Diseases	1278	1234
	OMIM	89	518
Drugs			
	Drugbank Carriers	247	7
	Drugbank Enzymes	797	180
	Drugbank Targets	4815	1494
	Drugbank Transporters	560	18
	KEGG Drugs	3793	706
Gene Ontology			
	GO Biological Processes	11,386	4850
	GO Cellular Component	1545	4852
	GO Molecular Function	4146	4832
Pathways			
	KEGG Pathways	310	4904

Accedendo alla pagina web sopraindicata, all'utente viene richiesto di selezionare una delle patologie messe a disposizione dal sistema (Figura 3.1a). A seguire, uno o più NoIs potranno eventualmente essere selezionati dall'utente accedendo alla seconda sezione (Figura 3.1b). In tal caso, verrà chiesto all'utente di inserire, nella terza sezione dell'interfaccia web, un valore di p-value per ogni nodo d'interesse selezionato (Figura 3.1c). Il p-value da inserire dovrà essere inferiore rispetto alla threshold preimpostata (valore di default o settato dall'utente); in caso contrario, il/i NoI/s selezionato/i dall'utente verranno scartati durante la procedura computazionale, secondo quanto esposto prima. Nella terza sezione, all'utente è data la possibilità di impostare altri valori threshold (Figura 3.1c), come specificato nella sottosezione "Estrazione dei sub-pathways". Da questa schermata l'utente può anche scegliere se far procedere la ricerca delle sottostrutture significative a ritroso o a valle rispetto ai NoIs selezionati dall'utente o dal sistema automatizzato incluso nel tool. Ciò è reso possibile selezionando o deselezionando la funzione "Backward visit",

visualizzato come bottone in basso (Figura 3.1c). Una volta settati tutti i parametri necessari, l'utente può avviare l'analisi predittiva cliccando sul bottone "Submit".

Al termine delle computazioni, verrà generata una tabella contenente una lista delle sottostrutture statisticamente significative (Figura 3.1d). In particolare, per ogni sottostruttura la tabella riporta (i) il/i nodo/i di partenza, (ii) il tipo di sottostruttura (cammino, albero, vicinato, sotto-grafo indotto o comunità di sotto-grafi), (iii) numero di nodi inclusi nella sottostruttura, (iv) valore della perturbazione, e (v) p-value aggiustato per la molteplicità dei test mediante metodo Holm–Bonferroni. Inoltre, sulla destra, il pulsante verde permette di visualizzare i risultati dell'enrichment analysis (Figura 3.2). Questi ultimi vengono mostrati in una pagina a parte contenente (i) una visualizzazione grafica della sottostruttura generata mediante Cytoscape, in alto nella pagina, e una tabella, in basso nella pagina, in cui vengono elencati (ii) i termini (ad esempio farmaci) con cui sono stati annotati i nodi della sottostruttura, (iii) la sorgente da cui è stata tratta l'associazione termine-nodo (ad esempio KEGG Drugs o DrugBank), (iv) l'ID di ogni termine riportato nella rispettiva sorgente, (v) il numero di nodi della sottostruttura annotati con il termine in questione, (vi) p-values e (vii) p-values aggiustati per la molteplicità dei test con il metodo Benjamini–Hochberg (Figura 3.2). In fondo alla pagina relativa all'enrichment analysis, l'utente troverà due pulsanti, "Download Network" e "Download Enrichment Terms", che permettono, rispettivamente, il download di un file di grafica XGMML contenente la sottostruttura e un file di testo separato da tabulazioni contenente tutte le informazioni visualizzate nella tabella dei termini.

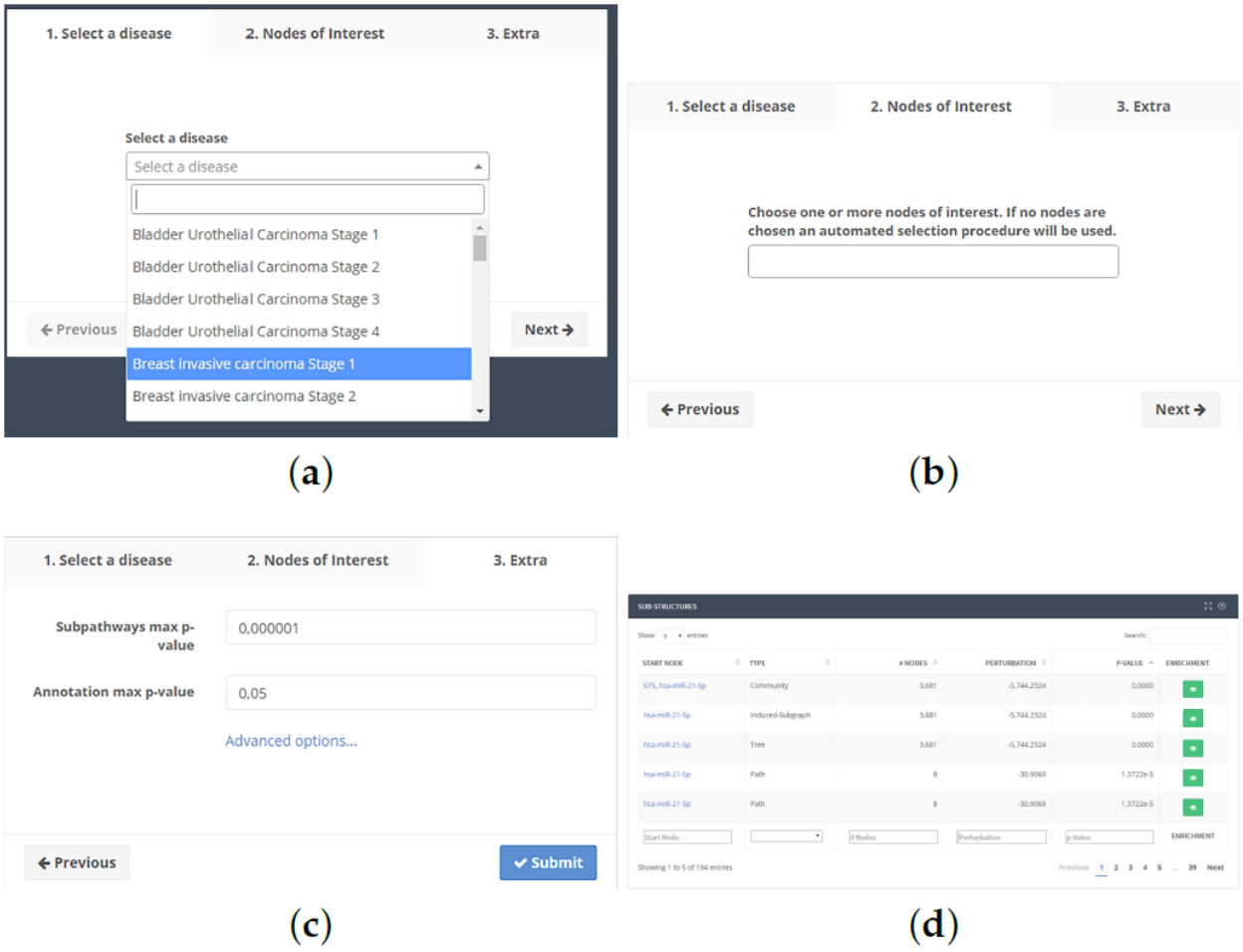


Figura 3.1 Workflow di SPECifIC. **(a)** Dopo aver selezionato la tipologia tumorale e lo stadio del tumore tra quelli elencati nel box, **(b)** l'utente può eventualmente scegliere uno o più nodi d'interesse (NoIs), **(c)** e, dopo aver modificato i parametri opzionali, l'analisi può essere sottoposta al processamento del sistema. Al termine della procedura analitica, **(d)** il Sistema restituirà una tabella riassuntiva, in cui vengono elentati pathway e subpathway statisticamente significativi. Su questa lista di termini, l'utente potrà quindi eseguire l'annotazione funzionale descritta nel testo. Figura tratta da [111].

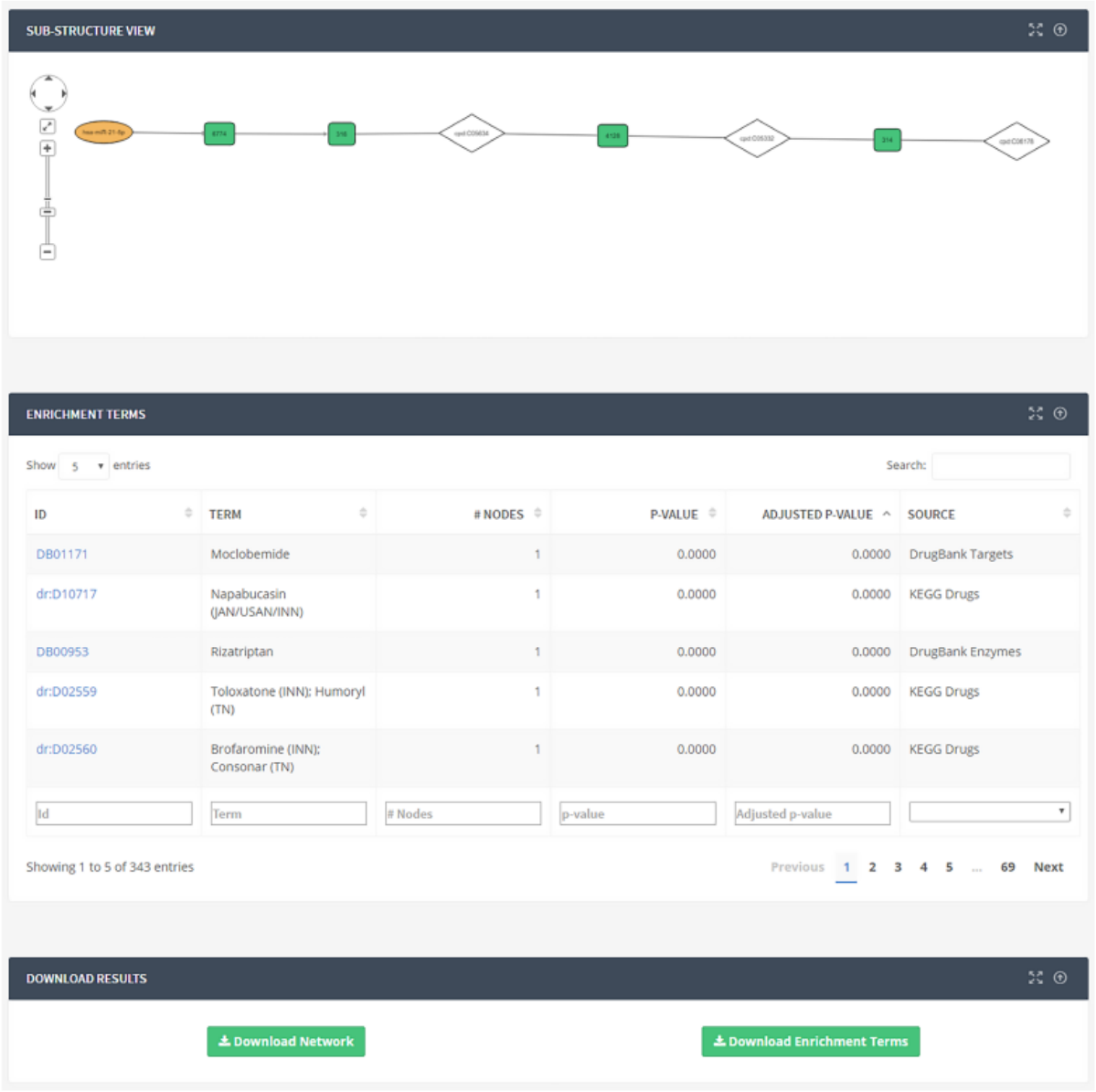


Figura 3.2 Lo screenshoot mostra i risultati dell'annotazione funzionale dei subpathway. I risultati vengono mostrati in una tabella (riquadro inferiore) insieme a grafi che mostrano le singole sottostrutture (riquadro superiore). In tabella vengono riportati gli identificativi (ID) dei termini, I loro rispettivi nomi ufficiali, il numero di nodi annotati per ogni singolo termine, il p-value, il p-value aggiustati per la molteplicità dei test, e infine la sorgente (database) di ogni termine. Cliccando sugli ID, verranno evidenziati nel grafo sovrastante i nodi annotati con il termine in questione. Inoltre, l'utente ha la possibilità di scaricare i risultati presenti in tabella in forma di file di testo spaziato da tabulazioni, e il grafo in un file formato XGMML. Figura tratta da [111].

3.2 PHENSIM

Come accennato al capitolo 2, l'algoritmo di PHENSIM si compone di cinque fasi operative, descritte qui di seguito.

3.2.1 Costruzione del meta-pathway

La procedura seguita dall'algoritmo di PHENSIM per costruire il meta-pathway di KEGG è uguale a quella già descritta nella sezione 3.1.1 (prima fase operativa dell'algoritmo di SPECifIC). Questa è l'unica fase operativa che i due algoritmi hanno in comune.

3.2.2 calcolo delle probabilità di attivazione o inattivazione dei nodi

Per calcolare l'attività dei nodi del meta-pathway a seguito di una deregolazione, l'algoritmo di PHENSIM considera ogni nodo (X_i) come una variabile random discreta che può assumere uno di tre possibili valori: +1 (attivato), -1 (inibito), 0 (nessuna variazione). Poiché la distribuzione delle probabilità per ogni variabile non è nota, PHENSIM ne dà una stima simulando combinazioni di valori random dei nodi in input e calcolando lo stato probabile di ogni nodo eseguendo un'analisi topologica della perturbazione dei pahway.

Più precisamente, sia $E = \{X_{j1} = x_{j1}, \dots X_{jn} = x_{jn}\}$ l'insieme dei valori dati in input dall'utente, dove X_{jk} rappresentano i nodi perturbati iniziali, con $1 \leq k \leq n$, mentre $x_{jk} \in \{-1, 1\}$ rappresentano i valori di deregolazione assegnati a X_{jk} dall'utente, con $1 \leq k \leq n$. Per calcolare lo stato di ogni nodo X_i del meta-pathway, PHENSIM svolge la suddetta analisi topologica della perturbazione dei pathway al fine di stimare la probabilità $P_r (X_i = x_i \mid E)$ della perturbazione di ogni X_i, dove $x_i \in \{-1, 0, 1\}$. Ciò viene ottenuto campionando lo spazio dei possibili valori di log fold-change (LFC) per ogni X_{jk} e stimandone la perturbazione $P_E (X_i, t)$ mediante l'algoritmo di MITHrIL. Dunque, dato un nodo in input $X_{jk} \in E$, il suo LFC al t-esimo step della simulazione viene campionato nel modo seguente:

$$LFC_E(X_{jk}, t) \sim \begin{cases} U(R^+) \; se \; X_{jk} = 1 \\ U(R^-) \; se \; X_{jk} = -1 \end{cases} \tag{2.1}$$

dove U rappresenta la distribuzione uniforme. Successivamente, il valore della perturbazione $P_E (X_i, t)$ viene stimata come:

$$P_E\ (X_i, t) = LFC_E(X_i, t) + \sum_{X_u \in U(X_i)} \frac{w(X_u, X_i)}{\sum_{d \in D(X_u)} w(X_u, X_d)}\ P_E\ (X_u, t)$$

(2.2)

dove $U(X_i)$ e $D(X_u)$ rappresentano, rispettivamente, il set dei nodi a monte e a valle di X_i, mentre w rappresenta il peso stabilito per il tipo di interazione tra il nodo adiacente a monte (X_u) e il nodo X_i. Dunque, lo stato S_E del nodo X_i allo step t-esimo della simulazione può essere stimato come:

$$S_E(X_i, t) \begin{cases} 1\ se\ P_E(X_i) > \varepsilon \\ -1\ se\ P_E(X_i) < -\varepsilon \\ 0\ altrimenti \end{cases}$$

(2.3)

dove ε rappresenta una threshold di inattività definita dall'utente. Iterando la procedura appena descritta T volte, la probabilità $P_r\ (X_i = x_i \mid E)$ viene stimata come:

$$P_r(X_i = x_i|E) = \frac{\#S_E(X_i, t) = x_i}{T}$$

(2.4)

Infine, iterando la procedura descritta dalle equazioni 2.2, 2.3 e 2.4 per le varianti randomizzate di E – considerando l'insieme delle randomizzazioni dei nodi in input $E_r = \{X_{r1} = x_{ji} \ldots, X_{rn} = x_{jn}\}$ – è possibile stimare la probabilità $P_r\ (X_i = x_i \mid E)$ delle alterazioni dei nodi nel modello nullo.

3.2.3 Calcolo dei rapporti delle probabilità rispetto al modello nullo e degli activity score

Dato il suddetto insieme dei valori in input E, PHENSIM riassume l'attività del generico nodo X_i attraverso un valore numerico detto "activity score" ($\mathbf{Ac_E(X_i)}$), che può assumere valori positivi o negativi. Tale valore ha una duplice funzione: il segno dell'$Ac_E(X_i)$ indica l'attività predetta del gene/proteina in questione – il segno positivo ne indica la presunta attivazione; il segno negativo ne indica la presunta inibizione – mentre il valore in sé rappresenta la probabilità, espressa come log-likelihood, di osservare la manifestazione fenotipica x_i predetta per il generico nodo X_i rispetto al modello nullo. Al fine di ottenere l'$Ac_E(X_i)$, PHENSIM calcola anzitutto un rapporto log-likelihood $L(E \mid X_i = x_i)$ per ogni possibile esito x_i. Siano $P_r\ (X_i = x_i \mid E)$ e $P_r\ (X_i = x_i \mid E_r)$ le probabilità che lo stato

di X_i sia x_i in input e nel modello nullo, rispettivamente. Il rapporto log-likelihood $L(E \mid X_i = x_i)$ può essere determinato secondo l'equazione:

$$L(E|X_i = x_i) = log \frac{P_r(X_i = x_i|E)}{P_r(X_i = x_i|E_r)} \tag{2.5}$$

Di conseguenza, l'Activity Score del nodo X_i può essere determinato come segue:

$$Ac_E(X_i) = \begin{cases} +L(E|X_i = 1) \; se \; L(E|X_i = 1) > L(E|X_i = -1) \\ -L(E|X_i = 1) \; se \; L(E|X_i = 1) < L(E|X_i = -1) \\ 0 \quad altrimenti \end{cases} \tag{2.6}$$

3.2.4 Determinazione dei p-value dei pathway

L'ultima fase dell'algoritmo di PHENSIM permette di filtrare i risultati ottenuti durante il calcolo degli $Ac_E(X_i)$ determinando la significatività statistica di questi ultimi. Di fatto, l'$Ac_E(Xi)$ può essere considerato rilevante se il rapporto del log-likelihood tra il valore in input dato dall'utente e il modello nullo risulta essere "sufficientemente" elevato. Per stabilire ciò, è possibile calcolare un p-value che permette di confrontare gli stati di distribuzione delle probabilità per ogni nodo X_i in entrambi i modelli. Dunque, siano $Ex[S_E(X_i, t)]$ e $Va[S_E(X_i, t)]$ rispettivamente il valore atteso e la varianza degli stati di distribuzione calcolati per il nodo X_i. Questi valori possono essere calcolati utilizzando i valori precedentemente calcolati mediante l'equazione 2.4 secondo le seguenti formule:

$$Ex[S_E(X_i, t)] = \sum_{x_i \in \{-1,0,1\}} P_r(X_i = x_i|E) \cdot x_i \tag{2.7}$$

$$Va[S_E(X_i, t)] = \sum_{x_i \in \{-1,0,1\}} P_r(X_i = x_i|E) \cdot (x_i - Ex[S_E(X_i, t)])^2 \tag{2.8}$$

Questi due valori possono essere corretti utilizzando il metodo di Welford per il valore atteso e la correzione di Bessel per la varianza, secondo le formule:

$$Ex_u[S_E(X_i,t)] = Ex[S_E(X_i,t)] + \frac{1}{T}\cdot \sum_{x_i\in\{-1,0,1\}} (x_i - Ex[S_E(X_i,t)])$$

(2.9)

$$Va_u[S_E(X_i,t)] = \frac{T}{T-1}\cdot Va[S_E(X_i,t)]$$

(2.10)

dove Ex_u e Va_u rappresentano, rispettivamente, valore atteso unbiased e varianza unbiased. Lo stesso calcolo viene eseguito per il modello nullo. Infine, a valle di questa procedura, le distribuzioni dei due modelli vengono confrontate utilizzando un unpaired two-samples T-test eteroschedastico.

3.2.5 Interfaccia web e risoluzione del problema della dipendenza dei nodi

Per garantire un'ampia accessibilità e fruibilità del tool, l'algoritmo di PHENSIM è stato implementato in un'applicazione web prossimamente resa disponibile. Il servizio è stato sviluppato facendo uso di PHP, HTML e JavaScript, mediante l'aiuto fornito dal framework open source di Laravel (https://laravel.com).

Accedendo allo "User Panel", l'utente può scegliere di correre l'analisi predittiva mediante due possibili modalità: "Simple mode" o "Advanced mode". Cliccando sul pulsante "Simple mode", l'utente accede ad una schermata in cui vengono mostrati dei pannelli in sequenza (Figura 3.3a). Ogni pannello contiene dei campi in cui andranno inserite le informazioni riguardanti l'analisi predittiva che si intende lanciare: nome del lavoro, specie su cui si vuole condurre la simulazione (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*), Entrez ID dei nodi deregolati in input (è necessario inserire almeno un nodo deregolato per lanciare un'analisi), eventuale lista di geni non espressi in uno specifico tipo cellulare o tessuto, e valore epsilon. È importante sottolineare che per ottenere risultati unbiased in Simple mode è necessario che i nodi dati in input dall'utente siano indipendenti tra loro – ovvero che questi non siano connessi da un cammino diretto all'interno del meta-pathway. Una volta inserite tutte le informazioni richieste, l'utente potrà sottomettere l'analisi predittiva al sistema cliccando sul pulsante "Submit". Cliccando invece sul pulsante "Advanced mode", l'utente accede ad un unico pannello che include 11 diversi campi (Figura 3.3b). Qui, la maggior parte dei campi è provvista di pulsanti che permettono di caricare nel sistema files di testo spaziati da tabulazione opportunamente strutturati, secondo le istruzioni riportate nella sezione "User

Manual". Questa modalità è stata principalmente sviluppata per consentire di effettuare le analisi predittive anche nei casi in cui i nodi di partenza sono nodi dipendenti. Infatti, l'equazione 2.1 precedentemente discussa

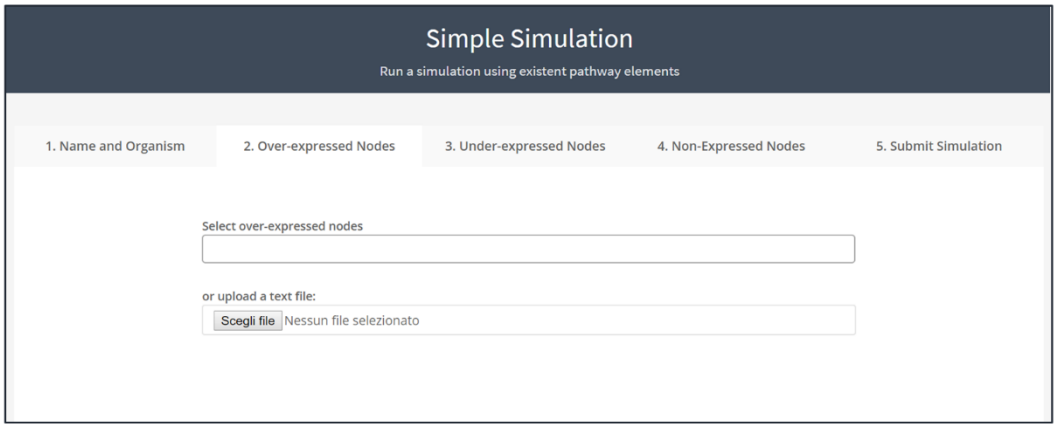
$$LFC_E(X_{jk}, t) \sim \begin{cases} U(R^+) \; se \; X_{jk} = 1 \\ U(R^-) \; se \; X_{jk} = -1 \end{cases}$$

richiede che tutti i nodi in input siano indipendenti gli uni dagli altri, permettendo di assegnare uniformemente i valori nel range dei possibili di log fold-change. Al contrario, nel caso in cui si dessero in input due o più nodi dipendenti, i valori di LFC dei nodi in input a monte andrebbero ad influenzare i valori dei nodi in input a valle, e dunque una distribuzione uniforme dei valori di LFC per questi nodi non sarebbe più approrpiata. Per superare questa limitazione, PHENSIM applica una strategia opportunamente studiata. Sia E l'insieme dei valori dati in input $\{X_{j1} = x_{j1}, \ldots X_{jn} = x_{jn}\}$, e sia E'= $\{X_{j1} = x_{j1}, \ldots X_{jt} = x_{jt}\}$ un sottoinsieme di E i cui componenti sono nodi dipendenti. È possibile creare un nuovo nodo X* direttamente connesso a E'. Per ogni arco $X^* \rightarrow X_{jk}$, il suo peso (che stabilisce il tipo di interazione) può essere assegnato come $W(X^*, X_{jk}) = x_{ik}$, dove x_{jk} rappresenta la direzione della deregolazione (up o down) che l'utente vuole simulare. Sulla base di questo ragionamento, è possibile costruire un nuovo insieme di nodi in input E*, i cui nodi sono tutti indipendenti, secondo la logica $E^* = \{X^* = 1\} \cup E \setminus E'$. Questo nuovo insieme di elementi in input può essere utilizzato per mantenere una distribuzione uniforme dei valori di LFC, senza la necessità di valutare come la dipendenza dei nodi possa alterare la distribuzione dei valori di espressione. Oltre a quanto appena descritto, le analisi lanciate in Advanced mode permettono all'utente di creare nodi fittizi che connettono elementi non presenti nei pathway di KEGG (ad esempio farmaci) con nodi inclusi nel meta-pathway. In tal caso, la categoria molecolare a cui appartiene l'elemento non presente nel database KEGG verrà specificato in base ai codici forniti nella sezione User Manual. Alla stessa maniera, anche il tipo di interazione (es. inibitoria o attivante) dovrà essere specificata sulla base dei codici forniti dal sistema.

Al termine delle computazioni, il sistema genera una tabella contenente una lista dei pathway di KEGG, ognuno annotato con il proprio activity score. Inoltre, per ogni activity score viene riportata la significatività statistica (Figura 3.4). Cliccando sulle icone verdi presenti sulla destra della tabella, è possibile accedere ad una tabella che riporta gli activity score calcolati per i nodi del pathway in questione. Infine, da quest'ultima schermata è possibile accedere ad una visualizzazione grafica delle simulazioni, dove i nodi colorati in rosso rappresentano geni o proteine significativamente sovraregolate, mentre quelli colorati in blu rappresentano geni o proteine significativamente

sottoregolate. I nodi colorati in verde (colore di default per i nodi KEGG) rappresentano geni o proteine per cui l'algoritmo di PHENSIM non ha calcolato alcuna perturbazione significativa.

(a)



(b)

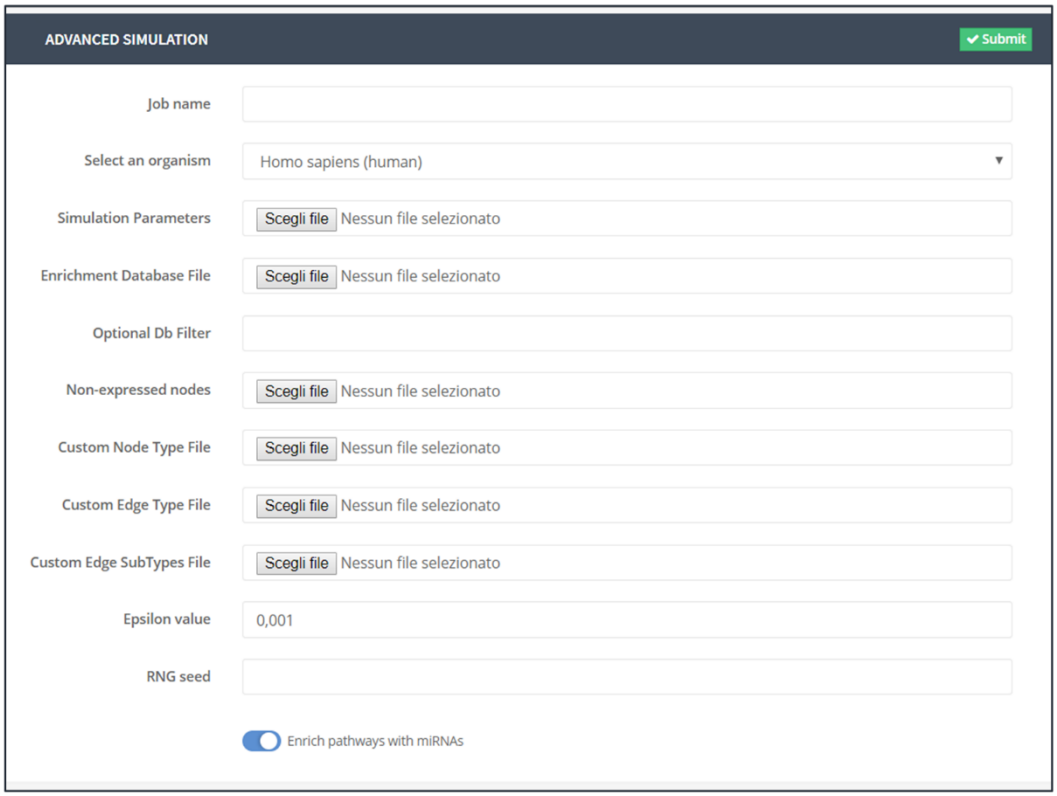


Figura 3.3 Screenshots che mostrano le schermate della Simple mode e della Advanced mode in PHENSIM.

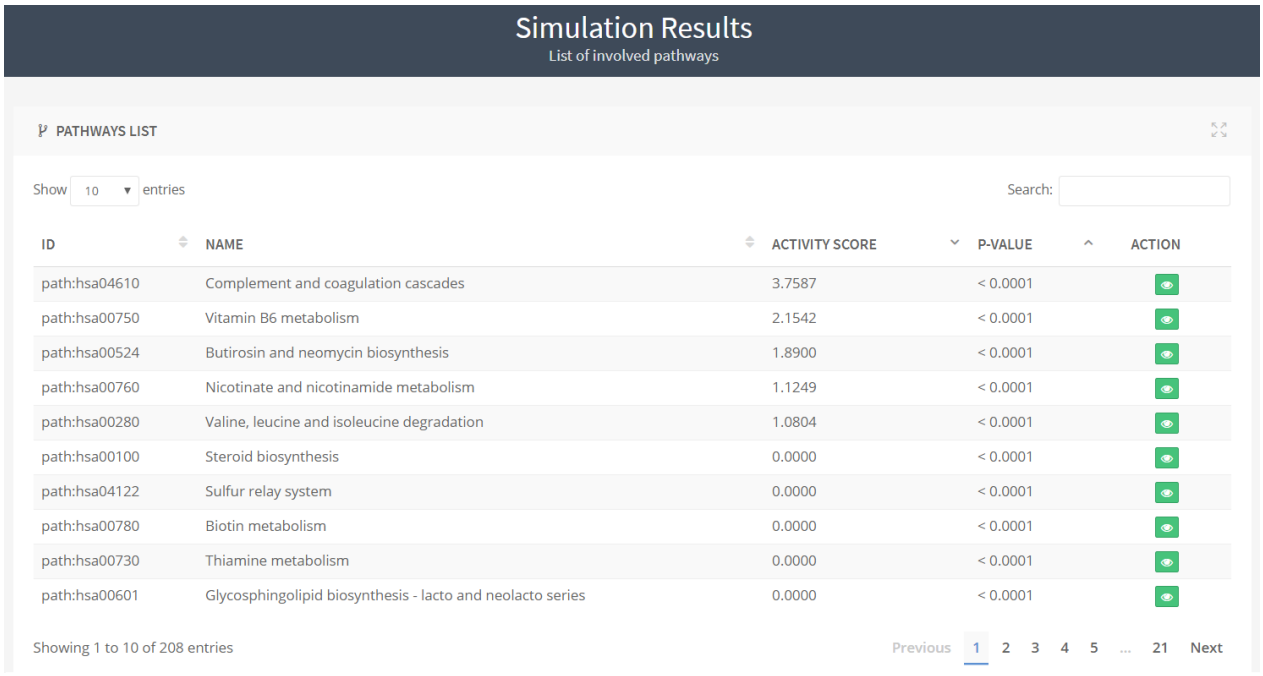


Figura 3.4 Tabella in cui vengono mostrati i risultati ottenuti dalla simulazione lanciata. La tabella riporta pathway ID, nome ufficiale del pathway in KEGG, activity score del pathway, e p-value calcolato per il pathway. Cliccando sui pulsanti verdi sulla destra si accede alla tabella che mostra le deregolazioni simulate per ogni singolo del pathway in questione.

4. Risultati

4.1 SPECifIC

Come riportato in sezione 2, al fine di valutarne le performance, SPECifIC è stato confrontato con tre altre metodiche di estrazione di subpathways, precisamente Subpathway-GM [112], Subpathway-GMir [113], e DEsubs [114]. Il confronto è stato operato sulla base dei risultati generati per mezzo di due analisi distinte. In particolare, le analisi sono state corse su dati di espressione forniti dal database TCGA, rispettivamente nel contesto del (i) carcinoma invasivo della mammella (BRCA) e (ii) dell'adenocarcinoma al colon (COAD). Per quanto riguarda la selezione dei nodi d'interesse (NoIs) in SPECifIC, questa è stata eseguita avvalendosi della procedura semi-automatizzata del sistema, allo scopo di avere una selezione unbiased (sezione 3.1.4). Nel caso dei tre tools "competitori", invece, il settaggio è stato eseguito secondo quanto suggerito dagli autori nei rispettivi manoscritti/manuali. Per evitare eventuali bias ed uniformare quanto più possibile le procedure analitiche, i disease pathways di KEGG sono stati rimossi manualmente, così che non se ne tenesse conto durante lo svolgimento delle pipelines. Alla fine delle procedure, i risultati ottenuti sono stati classificati sulla base dei p-values restituiti e, infine, confrontati (Figure 4.1-4.2, Tabelle 4.2-4.3 e Materiali supplementari). A tal proposito, l'unica eccezione ha riguardato la classificazione dei risultati ottenuti da DEsubs (Materiali supplementari). Infatti, il tool restituisce solamente l'insieme dei termini che arricchiscono significativamente i subpathway estratti (step (iv) della pipeline di DEsubs, si veda la sezione 3), ma non riporta i p-value delle sottostrutture stesse. Per questa ragione, si è deciso di procedere ordinando i termini restituiti dal tool sulla base del numero di subpathway in cui il termine in questione risultava statisticamente significativo.

4.1 Estrazione di sottostrutture e arricchimento funzionale operato da SPECifIC

A seguito dell'estrazione delle sottostrutture operata con il metodo SPECifIC, per ogni sottostruttura statisticamente significativa è stato eseguito un arricchimento funzionale, e tutti i termini dei pathway sono stati raccolti a prescindere dai p-values restituiti, per poter così analizzare anche i risultati non-significativi. Nei casi in cui un termine arricchiva più sottostrutture, è stato scelto il p-value con valore minore tra tutte le sottostrutture. Il risultante insieme di termini è stato quindi ordinato sulla base dei p-value restituiti e i primi 20 pathways (Tabella 4.1) sono stati utilizzati per l'analisi a valle, al fine di valutarne le performance.

Tabella 4.1 Lista dei primi 20 pathway ottenuti mediante l'analisi SPECifIC su dati di carcinoma mammario (BRCA) e adenocarcinoma al colon (COAD) a seguito dell'estrazione dei subpathway. I termini sono prima stati ordinati sulla base

dei loro p-value aggiustati per la molteplicità dei test, e i 20 termini statisticamente più significativi sono stati inclusi nella lista.

BRCA		COAD	
Pathway	***p***	**Pathway**	***p***
metabolism of xenobiotics by cytochrome p450	0	metabolism of xenobiotics by cytochrome p450	0
steroid hormone biosynthesis	0	drug metabolism cytochrome p450	0
drug metabolism cytochrome p450	0	chemical carcinogenesis	0
chemical carcinogenesis	0	steroid hormone biosynthesis	0
drug metabolism other enzymes	0	drug metabolism other enzymes	0
linoleic acid metabolism	0	linoleic acid metabolism	0
longevity regulating pathway	3.27×10^{-8}	ppar signaling pathway	0
egfr tyrosine kinase inhibitor resistance	2.17×10^{-7}	phenylalanine metabolism	0
endocrine resistance	2.35×10^{-7}	estrogen signaling pathway	1.09×10^{-30}
rap1 signaling pathway	5.32×10^{-7}	chemokine signaling pathway	1.28×10^{-30}
progesterone mediated oocyte maturation	5.39×10^{-7}	erbb signaling pathway	8.64×10^{-29}
hif 1 signaling pathway	5.89×10^{-7}	phospholipase d signaling pathway	1.11×10^{-27}
melanogenesis	6.15×10^{-7}	neurotrophin signaling pathway	4.63×10^{-27}
apoptosis	1.27×10^{-6}	insulin signaling pathway	7.95×10^{-26}
platinum drug resistance	1.27×10^{-6}	egfr tyrosine kinase inhibitor resistance	2.76×10^{-25}
phospholipase d signaling pathway	1.30×10^{-6}	prolactin signaling pathway	1.20×10^{-24}
mtor signaling pathway	1.55×10^{-6}	oxytocin signaling pathway	5.67×10^{-24}
ras signaling pathway	1.55×10^{-6}	platelet activation	5.67×10^{-24}
thyroid hormone signaling pathway	3.13×10^{-6}	endocrine resistance	5.87×10^{-24}
erbb signaling pathway	3.32×10^{-6}	focal adhesion	6.00×10^{-24}

4.1.1 Risultati ottenuti per BRCA

Attraverso il metodo SPECifIC sono stati individuati un totale di 163 termini di annotazione (pathways biologici di KEGG, ad esclusione dei disease pathways) potenzialmente associati a BRCA, di cui 74 statisticamente significativi (Tabella 4.2). L'ampio numero di termini identificati per questa patologia è stato interpretato come riflesso del fatto che, in realtà, BRCA è una patologia tumorale che include diversi sottotipi tumorali, ognuno associato a specifiche alterazioni oncogene o combinazioni di esse [149], come brevemente accennato nella sezione 2. Il coinvolgimento in questa patologia di diversi pathways biologici identificati dal tool, tra cui alcuni restituiti come statisticamente significativi (p-value < 0,01), trova conferma in letteratura (sezione 5).

Il metodo SubPathway-GM ha permesso l'identificazione di 124 termini di annotazione potenzialmente associati a BRCA, di cui 25 statisticamente significativi e 36 in sovrapposizione con quelli predetti tramite SPECifIC (senza considerare il p-value). Tra i termini con p-value < 0,01, 14 erano in comune con termini statisticamente significativi restituiti da SPECifIC (Figura 4.1).

Il metodo DEsubs ha identificato un totale di 22 termini di annotazione potenzialmente associati a BRCA, tutti statisticamente significativi (il metodo è stato progettato in modo da scartare automaticamente tutti i termini restituiti come non-significativi). Dei 22 termini totali, 13 erano condivisi con quelli statisticamente significativi predetti da SPECifIC (Figura 4.1).

SubPathway-GMir ha identificato un totale di 54 termini di annotazione potenzialmente associati a BRCA, di cui 53 restituiti come statisticamente significativi e solamente 7 condivisi con quelli risultanti dall'analisi predittiva di SPECifIC, tutti restituiti con p-value < 0,01 da entrambi i metodi (Figura 4.1). È da notare che il metodo SubPathway-GMir è in grado di annotare le sottostrutture analizzate esclusivamente con termini relativi a pathways metabolici.

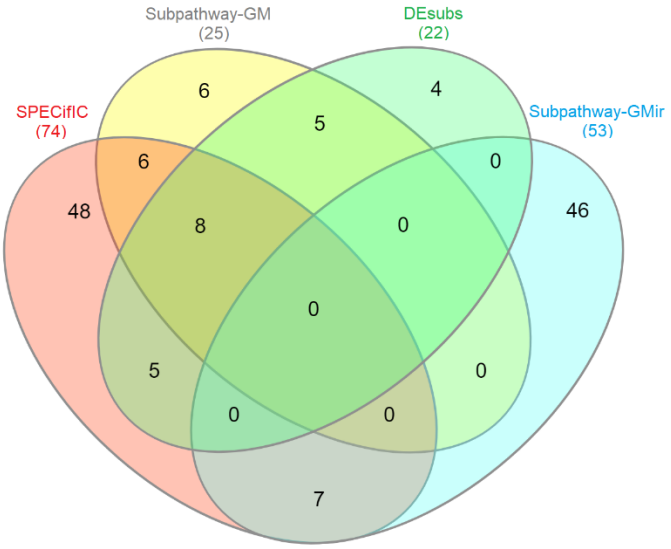


Figura 4.1 Confronti operati con diagrammi di Venn dei risultati ottenuti mediante SPECifIC, Subpathway- GM, Subpathway-GMir e DEsubs per il dataset BRCA estratto dal database TCGA. I dati riportati in figura fanno riferimento esclusivamente a pathway per cui il p-value calcolato era statisticamente significativo ($p < 0.01$).

Tabella 4.2 Risultati restituiti dal metodo SPECifIC per BRCA. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo, a prescindere dalla significatività statistica. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

Pathway	**SPECifIC**	**SubPathway-GM**	**SubPathway-GMir**	**DEsubs**
metabolism of xenobiotics by cytochrome p450	0		Yes	
drug metabolism cytochrome p450	0		Yes	
chemical carcinogenesis	0			
steroid hormone biosynthesis	0		Yes	
drug metabolism other enzymes	0		Yes	
linoleic acid metabolism	0		Yes	
ppar signaling pathway	0	Yes		Yes
phenylalanine metabolism	0		Yes	
estrogen signaling pathway	1,09E-30			
chemokine signaling pathway	1,28E-30	Yes		

erbb signaling pathway	8,64E-29		
phospholipase d signaling pathway	1,11E-27		
neurotrophin signaling pathway	4,63E-27	Yes	
insulin signaling pathway	7,95E-26	Yes	
egfr tyrosine kinase inhibitor resistance	2,76E-25		
prolactin signaling pathway	1,2E-24		
oxytocin signaling pathway	5,67E-24		
platelet activation	5,67E-24		
endocrine resistance	5,87E-24		
focal adhesion	6E-24	Yes	Yes
inflammatory mediator regulation of trp channels	1,98E-23		
cholinergic synapse	5,56E-21		
adrenergic signaling in cardiomyocytes	1,53E-20		
rap1 signaling pathway	2,43E-20		
vegf signaling pathway	2,53E-20		Yes
sphingolipid signaling pathway	2,68E-20		
natural killer cell mediated cytotoxicity	3,34E-20		Yes
thyroid hormone signaling pathway	9,45E-19		
ras signaling pathway	2,83E-18		
pathways in cancer	5,11E-18		
fc epsilon ri signaling pathway	1,36E-17		Yes
b cell receptor signaling pathway	7,78E-17	Yes	
cgmp pkg signaling pathway	2,27E-16		
phosphatidylinositol signaling system	7,35E-16		
choline metabolism in cancer	1,56E-15		
mtor signaling pathway	2,78E-15		
foxo signaling pathway	3,66E-15		
regulation of actin cytoskeleton	4,17E-15	Yes	Yes
t cell receptor signaling pathway	6,31E-15	Yes	
aldosterone regulated sodium reabsorption	1,2E-14		
central carbon metabolism in cancer	2,24E-14		
longevity regulating pathway multiple species	3,08E-14		
carbohydrate digestion and absorption	3,96E-14		
gnrh signaling pathway	4E-14		Yes
proteoglycans in cancer	1,87E-13		
serotonergic synapse	3,38E-13		
regulation of lipolysis in adipocytes	4,07E-13		
leukocyte transendothelial migration	4,29E-13		

fc gamma r mediated phagocytosis	8,62E-13	Yes		
signaling pathways regulating pluripotency of stem cells	8,89E-13			
microrнas in cancer	1,29E-12			
longevity regulating pathway	1,56E-12			
progesterone mediated oocyte maturation	1,86E-12			
long term depression	2,97E-12			Yes
camp signaling pathway	3,76E-12			
axon guidance	6,84E-12			
inositol phosphate metabolism	1,03E-11		Yes	
viral carcinogenesis	2,21E-11			
pi3k akt signaling pathway	2,7E-11			
jak stat signaling pathway	3,09E-11	Yes		Yes
dopaminergic synapse	3,84E-11			
toll like receptor signaling pathway	1,18E-10	Yes		Yes
tnf signaling pathway	1,5E-10			
glutamatergic synapse	1,67E-10			
ampk signaling pathway	4,93E-10			
osteoclast differentiation	6,68E-10			
apoptosis	1,13E-09	Yes		Yes
gap junction	1,32E-07	Yes		Yes
cell cycle	2,99E-07	Yes		Yes
oocyte meiosis	4,41E-07			
circadian entrainment	1,18E-05			
long term potentiation	1,63E-05			
vascular smooth muscle contraction	5,50E-05			
wnt signaling pathway	7,60E-05			
glycolysis gluconeogenesis	0,5000		Yes	
tyrosine metabolism	0,5000		Yes	
beta alanine metabolism	0,5000		Yes	
histidine metabolism	0,5000		Yes	
arginine and proline metabolism	0,5000		Yes	
glycine serine and threonine metabolism	0,5000		Yes	
platinum drug resistance	0,5000			
hif 1 signaling pathway	0,5000			
tryptophan metabolism	0,5000		Yes	
hippo signaling pathway	0,5000			
tight junction	0,5000	Yes		
retrograde endocannabinoid signaling	0,5000			
pyrimidine metabolism	0,5001		Yes	

purine metabolism	0,5002			Yes	
calcium signaling pathway	0,5004	Yes			Yes
mapk signaling pathway	0,5010	Yes			Yes
starch and sucrose metabolism	1			Yes	
insulin secretion	1				
melanogenesis	1				Yes
ovarian steroidogenesis	1				
thyroid hormone synthesis	1				
aldosterone synthesis and secretion	1				
glucagon signaling pathway	1				
arachidonic acid metabolism	1			Yes	
p53 signaling pathway	1	Yes			
mrna surveillance pathway	1				
protein processing in endoplasmic reticulum	1				
apoptosis multiple species	1				
nod like receptor signaling pathway	1				
transcriptional misregulation in cancer	1				
glutathione metabolism	1			Yes	
gabaergic synapse	1				
renin secretion	1				
salivary secretion	1				
gastric acid secretion	1				
pancreatic secretion	1				
dorso ventral axis formation	1				
th17 cell differentiation	1				
adipocytokine signaling pathway	1	Yes			Yes
endocrine and other factor regulated calcium reabsorption	1				
vasopressin regulated water reabsorption	1				
cytokine cytokine receptor interaction	1				Yes
endocytosis	1				
adherens junction	1	Yes			Yes
fatty acid degradation	1			Yes	
fatty acid metabolism	1				
citrate cycle tca cycle	1			Yes	
pyruvate metabolism	1			Yes	
proximal tubule bicarbonate reclamation	1				
carbon metabolism	1				
tgf beta signaling pathway	1				

cardiac muscle contraction	1			
lysine degradation	1		Yes	
ecm receptor interaction	1			Yes
pentose and glucuronate interconversions	1		Yes	
ascorbate and aldarate metabolism	1		Yes	
porphyrin and chlorophyll metabolism	1			
cytosolic dna sensing pathway	1			
neuroactive ligand receptor interaction	1			
taste transduction	1			
th1 and th2 cell differentiation	1			
olfactory transduction	1			
intestinal immune network for iga production	1			
antifolate resistance	1			
hematopoietic cell lineage	1			
cell adhesion molecules cams	1			
phototransduction	1			
glycerophospholipid metabolism	1		Yes	
ether lipid metabolism	1		Yes	
alpha linolenic acid metabolism	1			
glycerolipid metabolism	1		Yes	
primary bile acid biosynthesis	1			
alanine aspartate and glutamate metabolism	1		Yes	
butanoate metabolism	1		Yes	
fatty acid biosynthesis	1			
peroxisome	1			
fat digestion and absorption	1			
rig i like receptor signaling pathway	1	Yes		
synthesis and degradation of ketone bodies	1		Yes	
valine leucine and isoleucine degradation	1		Yes	
terpenoid backbone biosynthesis	1		Yes	
biosynthesis of unsaturated fatty acids	1			
phagosome	1	Yes		
propanoate metabolism	1		Yes	
vitamin b6 metabolism	1			
nicotinate and nicotinamide metabolism	1		Yes	
hippo signaling pathway multiple species	1			
thiamine metabolism	1			
one carbon pool by folate	1		Yes	

4.1.2 Risultati ottenuti per COAD

Nel caso di COAD, SPECifIC ha individuato un totale di 165 termini di annotazione potenzialmente associati alla patologia, di cui 31 statisticamente significativi (Tabella 4.3). Similmente a quanto accennato nel caso dei risultati restituiti per BRCA, diversi termini biologici identificati dal tool in relazione alla patologia del COAD trovano conferma in letteratura (sezione 5).

Attraverso il metodo SubPathway-GM sono state identificati di 83 termini di annotazione potenzialmente associati a COAD, di cui 27 statisticamente significativi e 11 in sovrapposizione con quelli predetti tramite SPECifIC (senza considerare il p-value). Di questi ultimi, 3 erano statisticamente significativi per entrambi imetodi (Figura 4.2).

Il metodo DEsubs ha identificato un totale di 19 termini di annotazione potenzialmente associati a COAD, tutti statisticamente significativi. Di questi, 5 erano condivisi con quelli predetti da SPECifIC e sono stati restituiti con p-value < 0,01 da entrambi i metodi (Figura 4.2).

SubPathway-GMir ha identificato un totale di 35 termini di annotazione potenzialmente associati a COAD, tutti con p-value < 0,01. Di questi, solamente 3 erano condivisi con quelli restituiti dall'analisi di SPECifIC ed erano statisticamente significativi per entrambi i metodi (Figura 4.2).

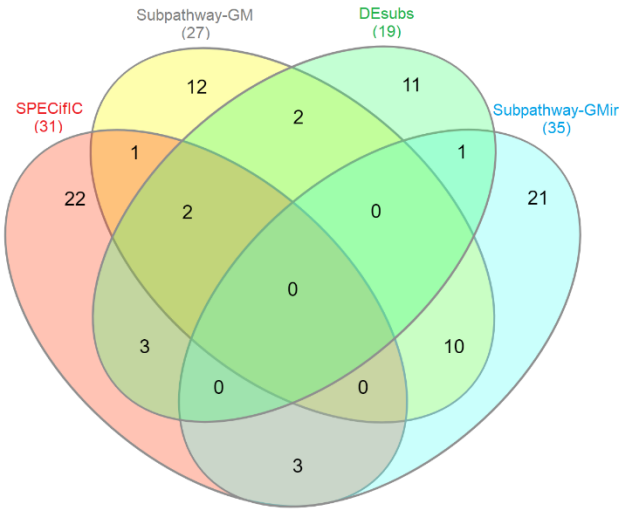


Figura 4.2 Confronti operati con diagrammi di Venn dei risultati ottenuti mediante SPECifIC, Subpathway- GM, Subpathway-GMir e DEsubs per il dataset COAD estratto dal database TCGA. I dati riportati in figura fanno riferimento esclusivamente a pathway per cui il p-value calcolato era statisticamente significativo (p < 0.01).

Tabella 4.3 Risultati restituiti dal metodo SPECifIC per COAD. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo, a prescindere dalla significatività statistica. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

KEGG Pathway	**SPECifIC**	**SubPathway-GM**	**SubPathway-GMir**	**DEsubs**
metabolism of xenobiotics by cytochrome p450	0		Yes	
steroid hormone biosynthesis	0		Yes	
drug metabolism cytochrome p450	0			
chemical carcinogenesis	0			
drug metabolism other enzymes	0			
linoleic acid metabolism	0			
longevity regulating pathway	3,27E-08			
egfr tyrosine kinase inhibitor resistance	2,17E-07			
endocrine resistance	2,35E-07			
rap1 signaling pathway	5,32E-07			
progesterone mediated oocyte maturation	5,39E-07			
hif 1 signaling pathway	5,89E-07			
melanogenesis	6,15E-07			Yes
apoptosis	1,27E-06	Yes		Yes
platinum drug resistance	1,27E-06			
phospholipase d signaling pathway	1,30E-06			
mtor signaling pathway	1,55E-06			
ras signaling pathway	1,55E-06			
thyroid hormone signaling pathway	3,13E-06			
erbb signaling pathway	3,32E-06			Yes
estrogen signaling pathway	3,97E-06			
inflammatory mediator regulation of trp channels	4,85E-06			
proteoglycans in cancer	5,00E-06			
pathways in cancer	7,08E-06			
platelet activation	9,18E-06			
chemokine signaling pathway	1,42E-05	Yes		
gnrh signaling pathway	1,50E-05			Yes
oxytocin signaling pathway	1,61E-05			
arachidonic acid metabolism	2,34E-05		Yes	
pi3k akt signaling pathway	0,0003			
mapk signaling pathway	0,0017	Yes		Yes
insulin secretion	0,5000			
ovarian steroidogenesis	0,5000			
beta alanine metabolism	0,5000			
long term depression	0,5000			Yes
thyroid hormone synthesis	0,5000			
aldosterone synthesis and secretion	0,5000			

gap junction	0,5000				
vegf signaling pathway	0,5000				Yes
b cell receptor signaling pathway	0,5000				Yes
camp signaling pathway	0,5000				
glucagon signaling pathway	0,5000				
micrornas in cancer	0,5000				
regulation of lipolysis in adipocytes	0,5000				
dopaminergic synapse	0,5000				
natural killer cell mediated cytotoxicity	0,5000	Yes			Yes
adrenergic signaling in cardiomyocytes	0,5001				
cgmp pkg signaling pathway	0,5001				
purine metabolism	0,5005			Yes	
fc epsilon ri signaling pathway	1,0000				Yes
cytokine cytokine receptor interaction	1,0000				Yes
inositol phosphate metabolism	1,0000			Yes	Yes
phosphatidylinositol signaling system	1,0000				Yes
leukocyte transendothelial migration	1,0000				Yes
focal adhesion	1,0000	Yes			Yes
toll like receptor signaling pathway	1,0000				Yes
jak stat signaling pathway	1,0000				Yes
regulation of actin cytoskeleton	1,0000				Yes
adipocytokine signaling pathway	1,0000				Yes
starch and sucrose metabolism	1,0000	Yes		Yes	
fatty acid degradation	1,0000	Yes		Yes	
pyrimidine metabolism	1,0000			Yes	
pentose and glucuronate interconversions	1,0000	Yes		Yes	
porphyrin and chlorophyll metabolism	1,0000			Yes	
ascorbate and aldarate metabolism	1,0000	Yes			
glutathione metabolism	1,0000	Yes		Yes	
galactose metabolism	1,0000	Yes			
glycerolipid metabolism	1,0000			Yes	
tyrosine metabolism	1,0000	Yes		Yes	
glycerophospholipid metabolism	1,0000			Yes	
amino sugar and nucleotide sugar metabolism	1,0000	Yes		Yes	
fructose and mannose metabolism	1,0000			Yes	
citrate cycle tca cycle	1,0000			Yes	
pentose phosphate pathway	1,0000			Yes	
cysteine and methionine metabolism	1,0000			Yes	
ether lipid metabolism	1,0000			Yes	

phenylalanine metabolism	1,0000		Yes
arginine and proline metabolism	1,0000	Yes	Yes
pyruvate metabolism	1,0000	Yes	Yes
glycolysis gluconeogenesis	1,0000	Yes	
ppar signaling pathway	1,0000	Yes	
tight junction	1,0000	Yes	
calcium signaling pathway	1,0000	Yes	
aldosterone regulated sodium reabsorption	1,0000	Yes	
wnt signaling pathway	1,0000	Yes	
alanine aspartate and glutamate metabolism	1,0000	Yes	
butanoate metabolism	1,0000	Yes	
oocyte meiosis	1,0000	Yes	
cell adhesion molecules cams	1,0000	Yes	
cell cycle	1,0000	Yes	
tgf beta signaling pathway	1,0000	Yes	
fc gamma r mediated phagocytosis	1,0000		
lysine degradation	1,0000		
olfactory transduction	1,0000		
long term potentiation	1,0000		
neurotrophin signaling pathway	1,0000		
nod like receptor signaling pathway	1,0000		
vascular smooth muscle contraction	1,0000		
phenylalanine tyrosine and tryptophan biosynthesis	1,0000		
axon guidance	1,0000		
salivary secretion	1,0000		
insulin signaling pathway	1,0000		
central carbon metabolism in cancer	1,0000		
ampk signaling pathway	1,0000		
sphingolipid signaling pathway	1,0000		
carbohydrate digestion and absorption	1,0000		
viral carcinogenesis	1,0000		
foxo signaling pathway	1,0000		
choline metabolism in cancer	1,0000		
p53 signaling pathway	1,0000		
mrna surveillance pathway	1,0000		
nf kappa b signaling pathway	1,0000		
protein processing in endoplasmic reticulum	1,0000		
apoptosis multiple species	1,0000		
hedgehog signaling pathway	1,0000		

transcriptional misregulation in cancer	1,0000
circadian entrainment	1,0000
retrograde endocannabinoid signaling	1,0000
glutamatergic synapse	1,0000
gabaergic synapse	1,0000
renin secretion	1,0000
gastric acid secretion	1,0000
pancreatic secretion	1,0000
abc transporters	1,0000
dorso ventral axis formation	1,0000
signaling pathways regulating pluripotency of stem cells	1,0000
t cell receptor signaling pathway	1,0000
prolactin signaling pathway	1,0000
osteoclast differentiation	1,0000
tnf signaling pathway	1,0000
th17 cell differentiation	1,0000
endocrine and other factor regulated calcium reabsorption	1,0000
vasopressin regulated water reabsorption	1,0000
autophagy	1,0000
endocytosis	1,0000
adherens junction	1,0000
fatty acid metabolism	1,0000
proximal tubule bicarbonate reclamation	1,0000
neomycin kanamycin and gentamicin biosynthesis	1,0000
carbon metabolism	1,0000
homologous recombination	1,0000
hippo signaling pathway	1,0000
cardiac muscle contraction	1,0000
other types of o glycan biosynthesis	1,0000
ecm receptor interaction	1,0000
protein digestion and absorption	1,0000
ubiquitin mediated proteolysis	1,0000
rna polymerase	1,0000
cytosolic dna sensing pathway	1,0000
neuroactive ligand receptor interaction	1,0000
taste transduction	1,0000
th1 and th2 cell differentiation	1,0000

selenocompound metabolism	1,0000
mineral absorption	1,0000
biosynthesis of amino acids	1,0000
folate biosynthesis	1,0000
intestinal immune network for iga production	1,0000
antifolate resistance	1,0000
hematopoietic cell lineage	1,0000
phototransduction	1,0000
alpha linolenic acid metabolism	1,0000
primary bile acid biosynthesis	1,0000
nucleotide excision repair	1,0000
taurine and hypotaurine metabolism	1,0000
histidine metabolism	1,0000

4.2 Metriche per la specificità dei metodi

Per esaminare la specificità dei risultati (in relazione all'associazione con le patologie), sono state calcolate due metriche sui disease genes (letteralmente "geni legati a malattie") inclusi all'interno delle sottostrutture estratte dai quattro metodi: (i) distanza media tra gene legato a malattia e sottostruttura; (ii) distanza media tra coppie di disease genes inclusi nelle sottostrutture (Tabella 4.4). I valori ottenuti sono stati confrontati con quelli ottenuti applicando le stesse metriche sui pathways di KEGG. I disease genes sono stati estrapolati dalle risorse online DisGeNET [150,151] e Genetic Association Database (GAD) [152], e sono stati successivamente filtrati tramite rimozione di tutti i geni non presenti in alcuno dei pathway presi in considerazione. La **metrica (i)** è un indice di specificità dei risultati, dal momento che le subpathway più vicine ai disease genes hanno maggiori probabilità di essere coinvolti nella malattia stessa. La **metrica (ii)** è un indice che mostra, in linea teorica, la tendenza che hanno i disease genes di manifestare funzioni sovrapposte e di causare manifestazioni fenotipiche sovrapposte. Tuttavia, questo parametro non sempre indica una reale rilevanza biologica dell'associazione.

Tabella 4.4 Metriche calcolate per determinare la specificità rispetto alle due patologie (BRCA e COAD) dei subpathway ricavati dai quattro metodi. Andando da sinistra verso destra, la tabella mostra: il numero di nodi inclusi nelle sottostrutture; il numero di nodi statisticamente significativi (p-value < 0.01); il numero di disease genes; il numero di disease genes statisticamente significativi (p-value < 0.01); numero di coppie di disease genes interni ai subpathway e raggiungibili mediante cammino diretto (parametro critico per il calcolo della metrica (ii)); distanza media tra disease gene e subpathway (†); distanza media tra disease genes contenuti all'interno di ogni sottostruttura (‡). I risultati sono stati confrontati con valori di riferimento calcolati direttamente in KEGG.

		# Nodes		**# Disease Genes**				
Dataset	**Algorithm**	$p < 0.01$	**All**	$p < 0.01$	**All**	**Reachable Pairs**	**†**	**‡**
BRCA	*KEGG Pathways*	1009	7121	30	104	283	-	7
	SPECifIC	466	466	15	15	6	**1.78**	3
	SubPathway-GM	101	214	9	14	6	1.89	3
	SubPathway-Gmir	142	722	4	8	1	2.09	2
	DESubs	34	34	0	0	0	2.48	-
COAD	*KEGG Pathways*	1009	7121	11	81	490	-	9
	SPECifIC	486	486	9	9	6	**1.67**	4
	SubPathway-GM	59	173	3	8	4	2.04	3
	SubPathway-Gmir	158	248	4	7	9	2.20	2
	DESubs	6	6	0	0	0	2.97	-

Tabella 4.5 Metriche calcolate per determinare la specificità rispetto alle due patologie (BRCA e COAD) dei subpathway ricavati dai quattro metodi dopo la rimozione delle interazioni miRNA-gene. Da sinistra verso destra: numero di nodi inclusi nelle sottostrutture; il numero di nodi statisticamente significativi (p-value < 0.01); numero di disease genes; il numero di disease genes statisticamente significativi (p-value < 0.01); numero di coppie di disease genes interni ai subpathway e raggiungibili mediante cammino diretto; distanza media tra disease gene e subpathway (†); distanza media tra disease genes contenuti all'interno di ogni sottostruttura (‡). I risultati sono stati confrontati con valori di riferimento calcolati direttamente in KEGG.

		# Nodes		**# Disease Genes**				
Dataset	**Algorithm**	$p < 0.01$	**All**	$p < 0.01$	**All**	**Reachable Pairs**	**†**	**‡**
BRCA	*KEGG Pathways*	983	6688	30	104	156	-	3
	SPECifIC	247	247	15	15	2	1,83	2
	SubPathway-GM	101	214	9	14	6	2,64	3
	SubPathway-Gmir	135	682	4	8	1	2,76	2
	DESubs	34	34	0	0	0	1,71	-
COAD	*KEGG Pathways*	995	6688	11	81	88	-	3
	SPECifIC	139	139	9	9	2	1,97	2
	SubPathway-GM	59	173	3	8	4	2,19	3
	SubPathway-Gmir	131	221	4	7	9	2,96	2
	DESubs	6	6	0	0	0	2,4	-

Poiché l'integrazione delle interazioni miRNA-gene nei pathway biologici con ne provoca cambiamenti topologici, le metriche (i) e (ii) sono state ricalcolate per SPECifIC a seguito della rimozione di tali interazioni dai grafi originali. Ciò ha reso possibile un migliore confronto tra i risultati restituiti dal tool SPECifIC e quelli restituiti da Subpathway-GM e DEsubs (Tabella 4.5), dal momento che questi ultimi non integrano alcuna informazione sulle interazioni miRNA-gene durante la procedura di estrazione delle sottostrutture. I risultati ottenuti hanno mostrato che SPECifIC individua sottostrutture significativamente più vicine a disease genes rispetto agli altri due metodi (Tabelle 4.6-4.7). Sebbene si sia osservato un incremento nella distanza media tra nodi legati a malattia interni alle sottostrutture (metrica (ii)), questi valori sono comunque rimasti inferiori rispetto a quelli osservati nei pathway di KEGG (Tabelle 4.8-4.9), indicando una maggiore specificità rispetto a quest'ultimo. A seguito della rimozione dei miRNAs dalle sottostrutture, sono stati ottenuti risultati simili a quelli ottenuti dagli altri metodi. Comunque, è bene ricordare che, nonostante questi risultati siano utilizzati come indice della specificità dei metodi, ciò potrebbe non implicare una reale valenza biologica.

Tabella 4.6 p-value utilizzati per il confronto delle le quattro metodologie in termini di distanze tra disease genes e subpathway. I p-value sono stati calcolati mediante Wilcoxon rank-sum test per i risultati ottenuti nel caso di BRCA.

	SPECifIC	**Subpathay-GM**	**Subpathway-Gmir**	**DESubs**
SPECifIC	-	0,0062	1,68E-09	3,98E-05
Subpathay-GM		-	2,98E-06	1,01E-05
Subpathway-Gmir			-	0,1794
DESubs				-

Tabella 4.7 p-value utilizzati per il confronto delle le quattro metodologie in termini di distanze tra disease genes e subpathway. I p-value sono stati calcolati mediante Wilcoxon rank-sum test per i risultati ottenuti nel caso di COAD.

	SPECifIC	**Subpathay-GM**	**Subpathway-Gmir**	**DESubs**
SPECifIC	-	2,26E-07	8,32E-12	8,41E-19
Subpathay-GM		-	1,03E-06	2,44E-13
Subpathway-Gmir			-	4,08E-07
DESubs				-

Tabella 4.8 p-value utilizzati per il confronto delle le quattro metodologie in termini di distanze tra disease genes e subpathway a seguito della rimozione delle interazioni miRNA-gene dai pathway. I p-value sono stati calcolati mediante Wilcoxon rank-sum test per i risultati ottenuti nel caso di BRCA. Non è stato possibile calcolare alcun p-value per i confronti con DESubs.

	KEGG P.	**SPECifIC**	**Subpathay-GM**	**Subpathway-Gmir**	**DESubs**
KEGG Pathways	**-**	7,23E-24	0,0118	5,49E-24	**-**
SPECifIC		-	0,7974	4,56E-11	**-**
Subpathay-GM			-	6,50E-06	**-**
Subpathway-Gmir				**-**	**-**
DESubs					**-**

Tabella 4.9 p-value utilizzati per il confronto delle le quattro metodologie in termini di distanze tra disease genes e subpathway a seguito della rimozione delle interazioni miRNA-gene dai pathway. I p-value sono stati calcolati mediante Wilcoxon rank-sum test per i risultati ottenuti nel caso di COAD. Non è stato possibile calcolare alcun p-value per i confronti con DESubs.

	KEGG P.	**SPECifIC**	**Subpathay-GM**	**Subpathway-Gmir**	**DESubs**
KEGG Pathways	**-**	0,0155	1,74E-05	6,77E-23	**-**
SPECifIC		-	1,67E-07	6,51E-33	**-**
Subpathay-GM			-	4,05E-05	**-**
Subpathway-Gmir				**-**	**-**
DESubs					**-**

Infine, data l'elevata ridondanza dei pathways biologici – infatti pathways distinti spesso condividono un certo numero di sottostrutture che svolgono ruoli simili o uguali nei due pathway [153] – è stato eseguito un enrichment delle sottostrutture estratte per COAD e BRCA con il metodo SPECifIC utilizzando i termini "disease" ricavati da DisGeNET per valutare quali patologie condividono le stesse sottostrutture (Tabella 4.10).

Tabella 4.10 Arricchimento (enrichment) dei subpathway estratti dall'algoritmo di SPECifIC per le due tipologie tumorali BRCA e COAD, eseguito utilizzando i termini disponibili sul database DisGeNET. In tabella sono stati elencati solamente i termini statisticamente significativi (p-value < 0.01).

COAD		BRCA	
Disease	**Adjusted p-value**	**Disease**	**Adjusted p-value**
Prostatic Neoplasms	0	*Prostatic Neoplasms*	0
Mammary Neoplasms	0	*Mammary Neoplasms*	0
Osteosarcoma	0	*Osteosarcoma*	0
Hepatitis C	0	*Hepatitis C*	0
Torsades de Pointes	0	*Torsades de Pointes*	0
Esophageal Neoplasms	0.0001	*MICROPHTHALMIA. ISOLATED 8*	0
Adenocarcinoma	0.0002	*Prostatic Neoplasms. Castration-Resistant*	0
Colorectal Neoplasms	0.0005	*Neoplasms. Hormone-Dependent*	1.41×10^{-08}
Hypertensive disease	0.0005	*Disorders of Sex Development*	7.05×10^{-08}
Lung Neoplasms	0.0005	*Obesity*	3.38×10^{-06}
Liver carcinoma	0.0011	*Substance-Related Disorders*	0.0002
Reperfusion Injury	0.0012	*Polycystic Ovary Syndrome*	0.0002
		Hypertensive disease	0.0004
		Myocardial Infarction	0.0006
		Stomach Neoplasms	0.0010
		Diabetes Mellitus. Experimental	0.0011
		Colorectal Neoplasms	0.0022

4.2 PHENSIM

Di seguito verranno mostrati i risultati restituiti dal metodo predittivo PHENSIM nei quattro casi simulati, elencati nel capitolo 2.

4.2.1 Trattamento con metformina

Per simulare lo scenario (i) è stato creato un nodo fittizio che connettesse il farmaco metformina ai tre target riportati in letteratura: STK11 (sovraregolato), INS (insulina, sottoregolato) e IGF-1 (sottoregolato) [122]. Nessuna lista di geni non espressi è stata tenuta in considerazione durante la simulazione, al fine di ottenere dei risultati quanto più possibile generalizzabili. Al termine dell'analisi, PHENSIM ha restituito una sottoregolazione significativa delle vie di segnalazione dell'insulina e di mTOR (pathway activity score = -2.2981 and -2.3491, rispettivamente; p-value < 0.0001). Tutte le isoforme delle chinasi PI3K e AKT, come anche il metabolita PIP3, sono state predette come sottoregolate (es. activity score = -3.3524 and -2.3567 per le isoforme PIK3CA e AKT1, rispettivamente; p-value < 0.0001). mTOR risultava regolato negativamente (activity score = -3.2626; p-value < 0.0001), causando a sua volta la sovraregolazione del repressore dei fattori iniziali della traduzione 4EBP (activity score = 3.1847; p-value < 0.0001) e la sottoregolazione dei nodi a valle coinvolti nella sintesi proteica (Figura 4.3). Il ciclo degli acidi tri-carbossilici è stato restituito anch'esso come sottoregolato (pathway activity score = -3.0033; p-value < 0.0001), insieme alle vie di segnalazione MAPK e NF-kB (pathway activity score = -2.0641 and -1.7574; p-value < 0.0001) (Figura 4.4). È da notare che il tipo di perturbazione predetto per diversi degli enzimi e dei metaboliti inclusi in questi pathway è in linea con quanto precedentemente riportato in letteratura [122]. Tuttavia, la sottoregolazione della via di segnalazione di p53 e la non-deregolazione di alcune interleuchine restituite da PHENSIM erano in contrasto con i dati di letteratura [122]. Nel contesto della presente simulazione, PHENSIM ha mostrato una sensibilità = 0.84 ed una specificità = 0 (Tabella 4.11).

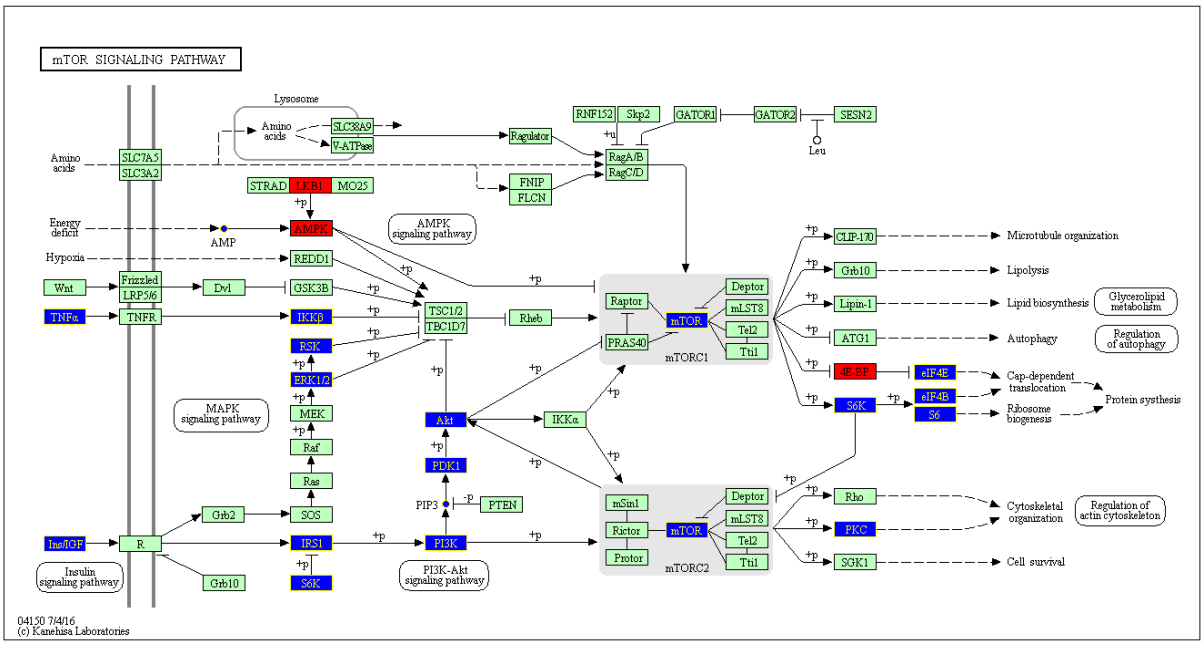


Figura 4.3 Effetti anti-tumorali della metformina simulati dal tool PHENSIM: predizioni per la via di segnalazione di mTOR.

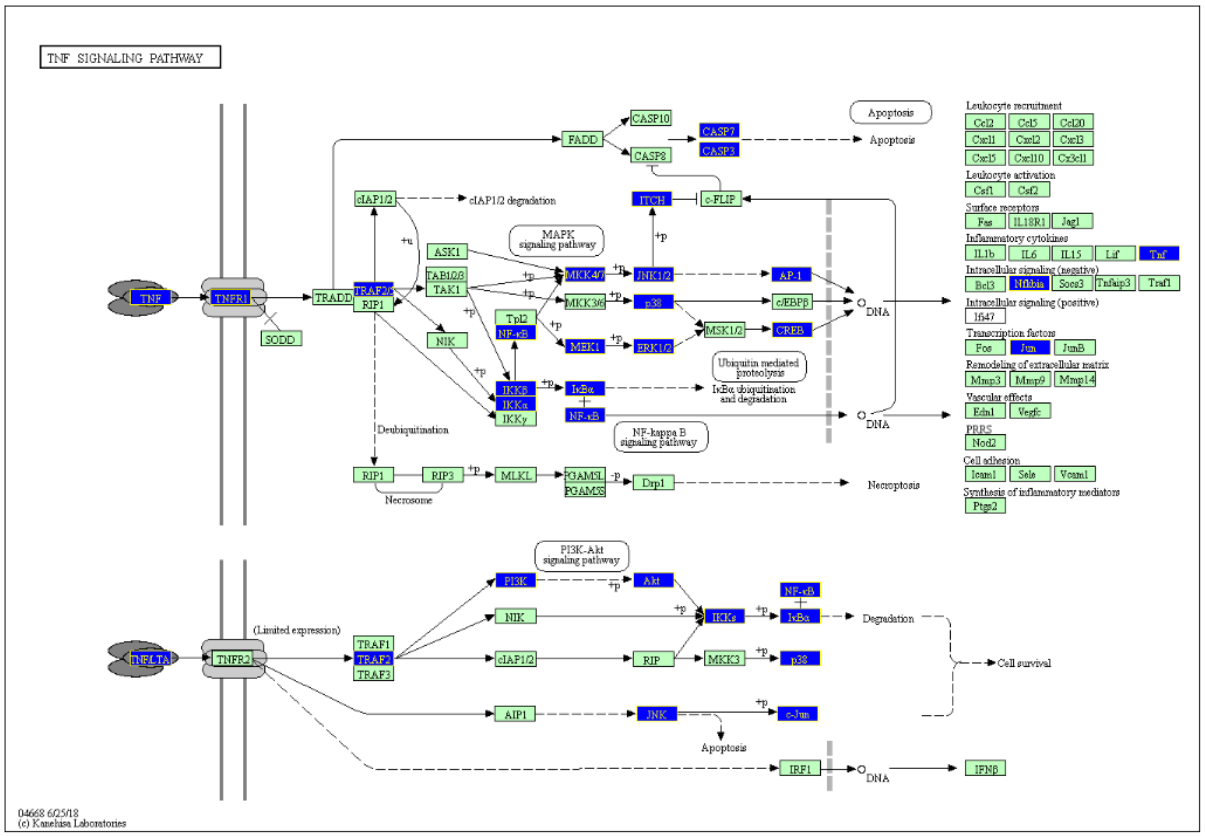


Figura 4.4 Effetti anti-tumorali della metformina simulati dal tool PHENSIM: predizioni per la via di segnalazione di TNF. In questo pathway vengono mostrate le perturbazioni di diversi nodi facenti parte delle vie di segnalazione MAPK e NF-kB.

Tabella 4.11 Numero assoluto di true positives, true negatives, false positives e false negatives ottenuti per la simulazione (i).

			CONDIZIONE REALE: Positivo	CONDIZIONE REALE: Negativo
CONDIZIONE PREDETTA	Predetto	positivo	37	20
	Predetto	negativo	7	0

4.2.2 Trattamento di cellule del tessuto mammario con everolimus

Nel caso studio (ii) non è stato possibile simulare gli effetti di RAD001 settando una semplice sottoregolazione di mTOR, in quanto KEGG non fa distinzione tra mTOR di mTORC1 ed mTOR di mTORC2. Perciò, al fine di superare tale limitazione, è stato creato un nodo fittizio che connettesse mTORC1 con i nodi adiacenti a valle, la cui attività è notoriamente modificata a seguito dell'inibizione del suddetto macchinario molecolare. In particolare, la simulazione è stata condotta ipotizzando una sottoregolazione di S6K, e una sovraregolazione di 4E-BP e ULK. Inoltre, durante l'analisi si è tenuto conto della lista di geni non espressi nel tessuto mammario, ricavata dal database GTEx, avendo l'obiettivo di simulare uno scenario in cui il farmaco viene somministrato direttamente ai tipi cellulari che compongono tale tessuto. Alla fine dell'analisi, sia la via di trasporto dell'RNA che la via di segnalazione di mTOR sono state restitutite come significativamente sottoregolate, laddove gli activity score di questi pathway risultavano essere tra i più bassi in assoluto (-2.9570 and -2.3177, respectively; p-value < 0.0001), secondo le aspettative (Figura 4.5). Infatti, alcuni fattori coinvolti nella sintesi dell'RNA e nella sintesi proteica, come ad esempio EIF4E, EIF4B, EIF4A e S6, sono stati restituiti come nodi fortemente sottoregolati (activity score = -3.3524, -4.6565, -4.6052 e -4.6565, rispettivamente; p-value < 0.0001). Inoltre, PHENSIM prediceva la sovraregolazione della via dell'autofagia (pathway activity score = 1.4966, p-value < 0.0001), conseguentemente al decremento dei livelli di fosforilazione di ULK1/2 da parte di mTOR [123,124]. Tuttavia, PHENSIM non ha restituito alcun risultato circa la deregolazione di p21, ciclina D ed NF-

kB a seguito del trattamento con RAD001 [124]. Nel contesto di questa simulazione, PHENSIM ha mostrato una sensibilità = 0.25 e una specificità = 1 (Tabella 4.12).

PATHWAYS LIST

Show 10 entries Search:

ID	NAME	ACTIVITY SCORE	P-VALUE	ACTION
path:hsa03013	RNA transport	-2.9570	< 0.0001	
path:hsa04910	Insulin signaling pathway	-2.3919	< 0.0001	
path:hsa04350	TGF-beta signaling pathway	-2.3544	< 0.0001	
path:hsa04213	Longevity regulating pathway - multiple species	-2.3335	< 0.0001	
path:hsa04150	mTOR signaling pathway	-2.3177	< 0.0001	
path:hsa04151	PI3K-Akt signaling pathway	-2.2784	< 0.0001	
path:hsa04012	ErbB signaling pathway	-2.2779	< 0.0001	
path:hsa04666	Fc gamma R-mediated phagocytosis	-2.1125	< 0.0001	
path:hsa04066	HIF-1 signaling pathway	-2.0759	< 0.0001	
path:hsa04144	Endocytosis	0.0000	< 0.0001	

Figura 4.5 Predizione della perturbazione dei pathway a seguito della somministrazione di everolimus: tessuto mammario. In figura viene riportata la lista dei top-10 pathway sottoregolati, tra cui figurano le vie di trasporto dell'RNA e la segnalazione mediata da mTOR.

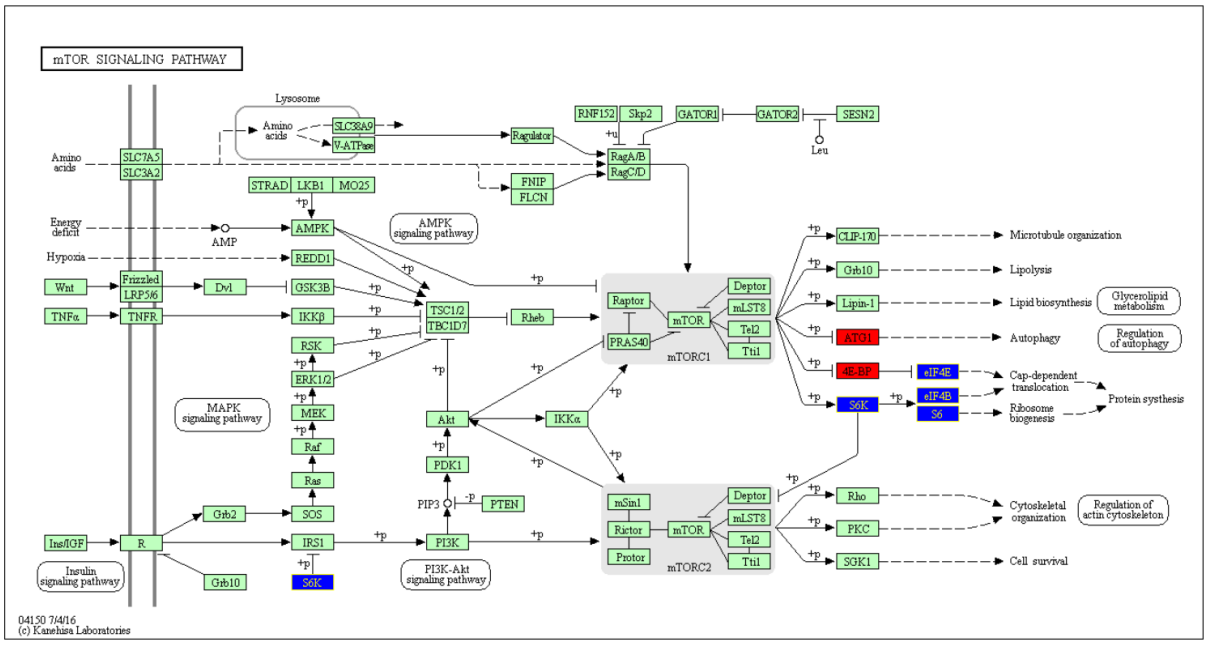


Figura 4.6 Predizione della perturbazione dei pathway a seguito della somministrazione di everolimus: tessuto mammario. Qui vengono mostrate nel dettagli le perturbazioni sui nodi del pathway di segnalazione di mTOR.

Tabella 4.12 Numero assoluto di true positives, true negatives, false positives e false negatives ottenuti per la simulazione (ii).

		CONDIZIONE REALE	
		Positivo	Negativo
CONDIZIONE PREDETTA	Predetto positivo	4	0
	Predetto negativo	12	38

4.2.3 Impatto dei miRNA esosomiali derivati da cellule di LMA nelle cellule riceventi del midollo osseo

La simulazione (iii) è stata corsa prendendo in considerazione gli 8 miRNA più rappresentativi isolate dagli esosomi della linea cellulare Molm-14 (linea cellulare continua umana di LMA), ovvero miR-155-5p, -150-5p, -146a-5p, -191-5p, -221-3p, -99b-5p, -1246 e let-7a-5p, secondo quanto riportato in [125]. Tali miRNA sono stati inclusi in un unico nodi fittizio appositamente creato. Ai fini della simulazione, nessuna lista di geni non espressi è stata caricata in memoria, dal momento che il target erano cellule immature del midollo osseo (BM). Alla fine della simulazione, sia la via di differenziazione degli osteoclasti che le interazioni tra citochine erecettori delle citochine sono state restituite come significativamente sottoregolate (pathway activity score = -2.4196 e -2.5975, rispettivamente; p-value < 0.0001) (Figura 4.7-4.8). In particolare, alcuni geni coinvolti nella modulazione dei processi ematopoietici, come ad esempio CXCL12 e IGF1, sono stati restituiti come nodi sottoregulated (activity score = -5.2005 e -2.8102, rispettivamente; p-value < 0.0001) [127]. Anche il fattore di trascrizione c-MYB, coinvolto nella differenziazione e proliferazione delle HSPCs, è stato restituito con un valore di attività negativo (activity score = -4.1044; p-value < 0.0001) [126] (Figura 4.9). Tuttavia, PHENSIM non è stato in grado di predire la sovraregolazione di DKK1 [127]. Poiché la letteratura scientifica manca di molte informazioni riguardo agli effetti esercitati dal carico

esosomiale di LMA sulle cellue riceventi, per questa simulazione non è stato possibile calcolare i valori di sensibilità e specificità.

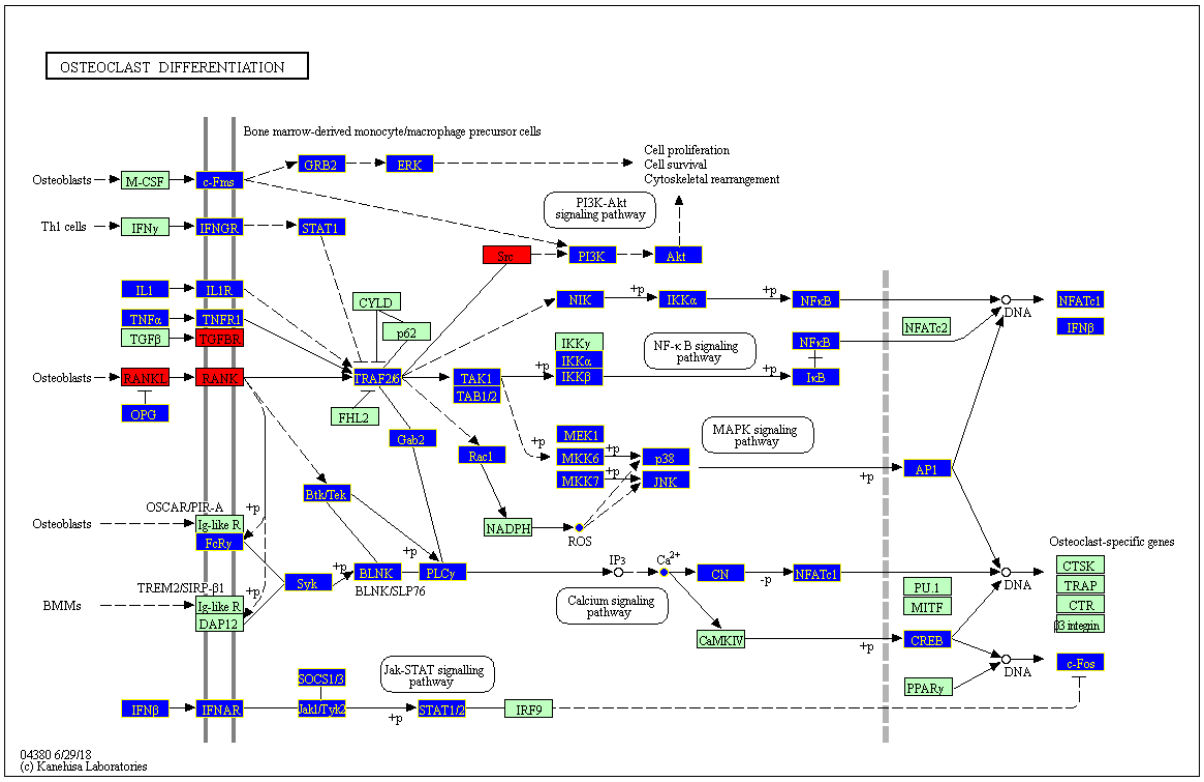


Figura 4.7 Risultati ottenuti per la via di differenziazione degli osteoclasti a seguito della simulazione (iii).

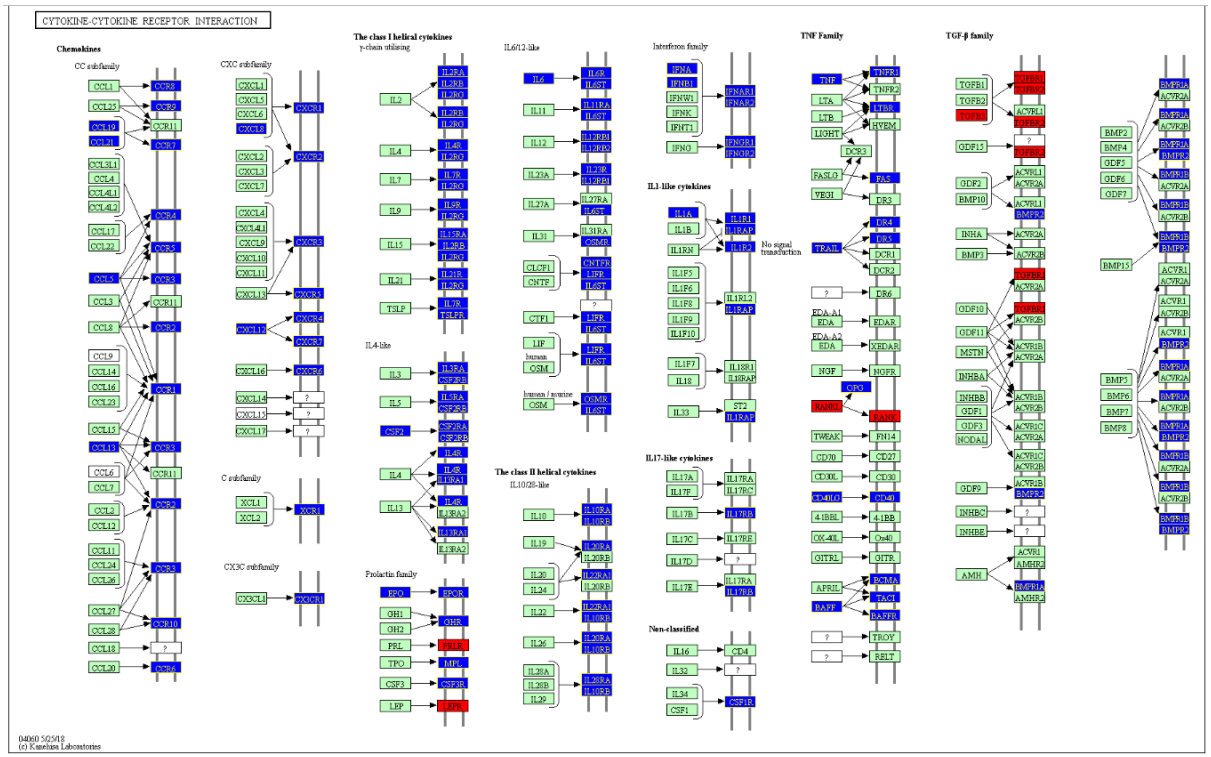


Figura 4.8 Risultati ottenuti per interleuchine e loro recettori a seguito della simulazione (iii).

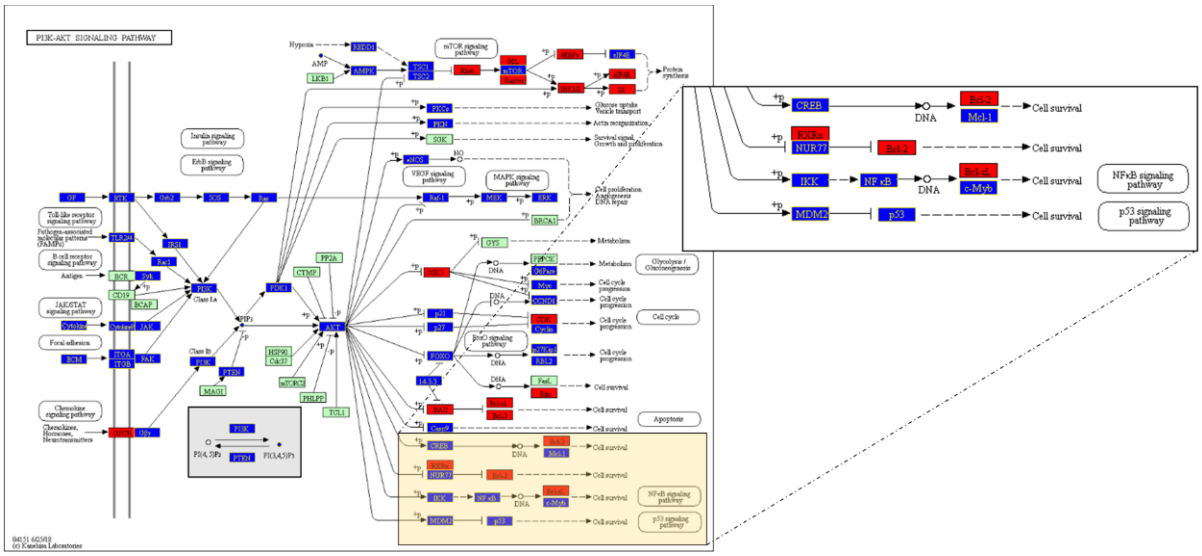


Figura 4.9 Risultati ottenuti per la via di segnalazione PI3K-Akt a seguito della simulazione (iii). Lo zoom mette in evidenza la sottoregolazione di c-Myb.

4.2.4 Efficacia di un modello di letalità sintetica in sei linee cellulari tumorali

Per generare un quadro completo dello scenario (iv), le stesse simulazioni (sei) sono state eseguite due volte. In particolare, le simulazioni avevano lo scopo di predire il fenotipo di 6 diverse linee cellulari continue umane (HeLa, HCT116, U2-OS, RKO, SW480 e CaCo-2) a seguito di due particolari

trattamenti capaci di indurre, o almeno predisporre, la cellula all'apoptosi (i.e. ogni linea cellulare avrebbe dovuto ricevere due diversi trattamenti). Per correre ognuna di queste simulazione, è stata caricata nel sistema la lista di geni non espressi relativa ad ognuna di queste cellule. Il primo dei due trattamenti consisteva nel semplice KD di TPL2 (1), per cui le simulazioni sono state lanciate in simple mode settando TPL2 come nodo sottoregolato. Il secondo trattamento, invece, consisteva in una combinazione tra KD di TPL2 e aggiunta di TNFα alla coltura cellulare, per cui le simulazioni sono state lanciate in advanced mode impostando la sottoregolazione di TPL2 e la sovraregolazione di TNFα e TNF-R1.

In (1) l'analisi di PHENSIM non è stata in grado di individuare, nelle line cellular sensibili al trattamento (HeLa, HCT116, U2-OS), la sovraregolazione di CASP8, che invece è stato identificato come il meccanismo responsabile dell'induzione apoptotica nel lavoro recentemente condotto dal gruppo di ricerca del Professore Tsichlis (articolo in revisione su PNAS). Inoltre, la sottoregolazione di BAK e la sovraregolazione di PARP erano in contrasto con i dati ottenuti dal gruppo di ricerca. D'altro canto, PHENSIM ha restituito la sovraregolazione degli inibitori dell'apoptosi BCL2 e BCL-XL, e la sottoregolazione dell'induttore dell'apoptosi mitocondriale BAX. Inoltre, questo risultato è stato restituito solamente per le linee cellulari sensibili al trattamento, ma non per le resistenti, ad eccezione di CaCo-2 (Figura 4.10-4.11). Nonostante l'apparente contrasto tra i risultati attesi e quelli restituiti in questo particolare contesto, tale predizione ha trovato riscontro nei risultati di riferimento, mostrando anche una discreta capacità di distinguere tra linee cellulari sensibili e resistenti a specifici trattamenti. Inoltre, PHENSIM è stato in grado di predire un forte decremento dei livelli di fosforilazione delle chinasi ERK, MEK, JNK e p38, come anche dell'attività di cIAP2 per tutte e sei le linee cellulari. Anche questi risultati hanno trovato positivo nei dati di riferimento.

In (2), I risultati restituiti da PHENSIM hanno riportato solo pochi cambiamenti rispetto al caso (1) (Figura 4.12-4.13). In particolare, gli activity score dei nodi evidenziano un increment significativo degli effetti causati dal KD di TPL2, confermando, per sommi capi, quanto ottenuto dal gruppo Tsichlis. Nemmeno in questo caso è stata predetta l'attivazione di CASP8 per le linee cellulari sensibili al trattamento.

Nel suo insieme, nella la simulazione (iv) PHENSIM ha mostrato sensibilità = 0.83 e specificità = 0.14 (Tabella 4.13).

Tabella 4.13 Numero assoluto di true positives, true negatives, false positives e false negatives ottenuti per la simulazione (iv).

		CONDIZIONE REALE	
		Positivo	Negativo
CONDIZIONE PREDETTA	Predetto positivo	33	71
	Predetto negativo	7	12

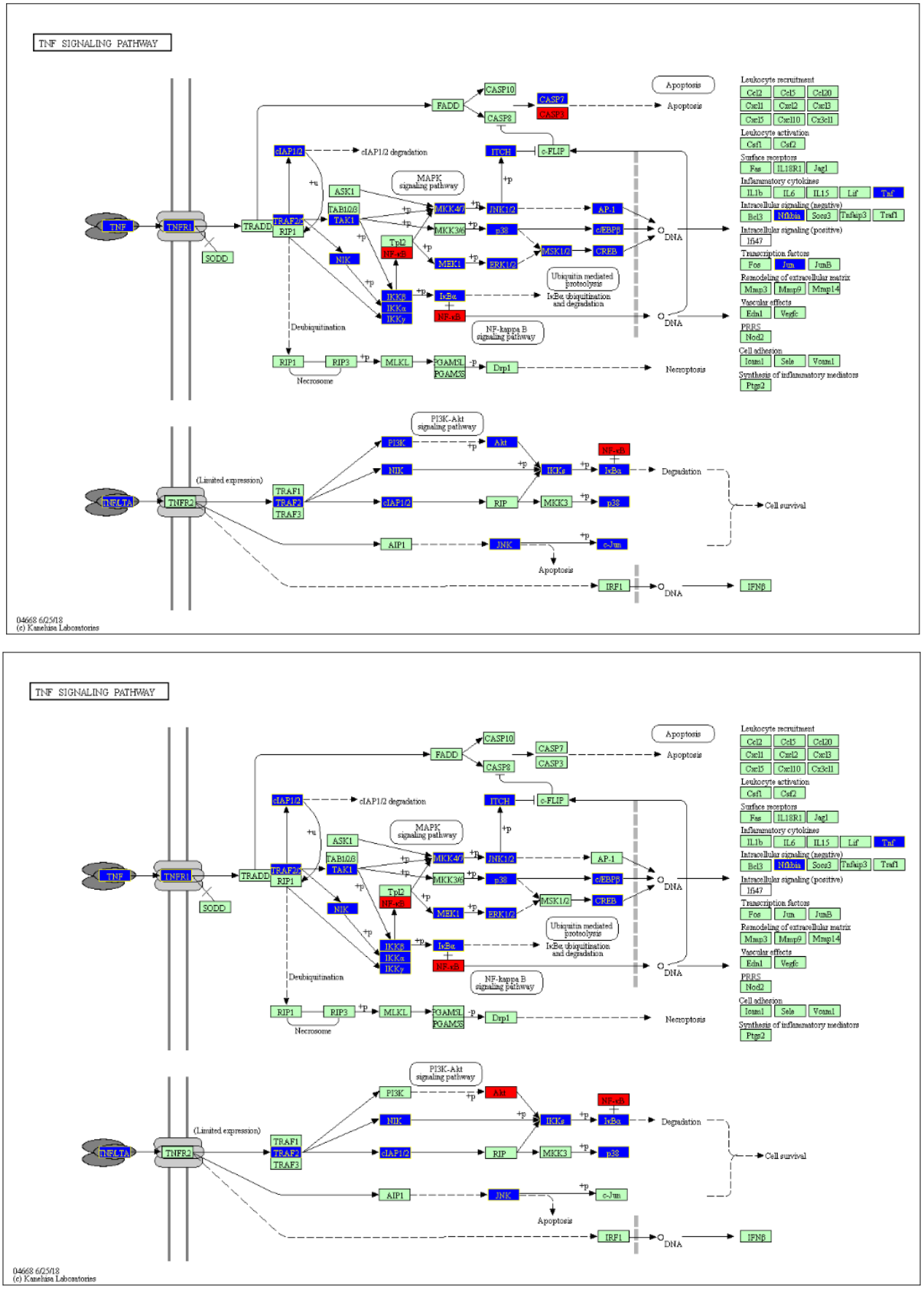
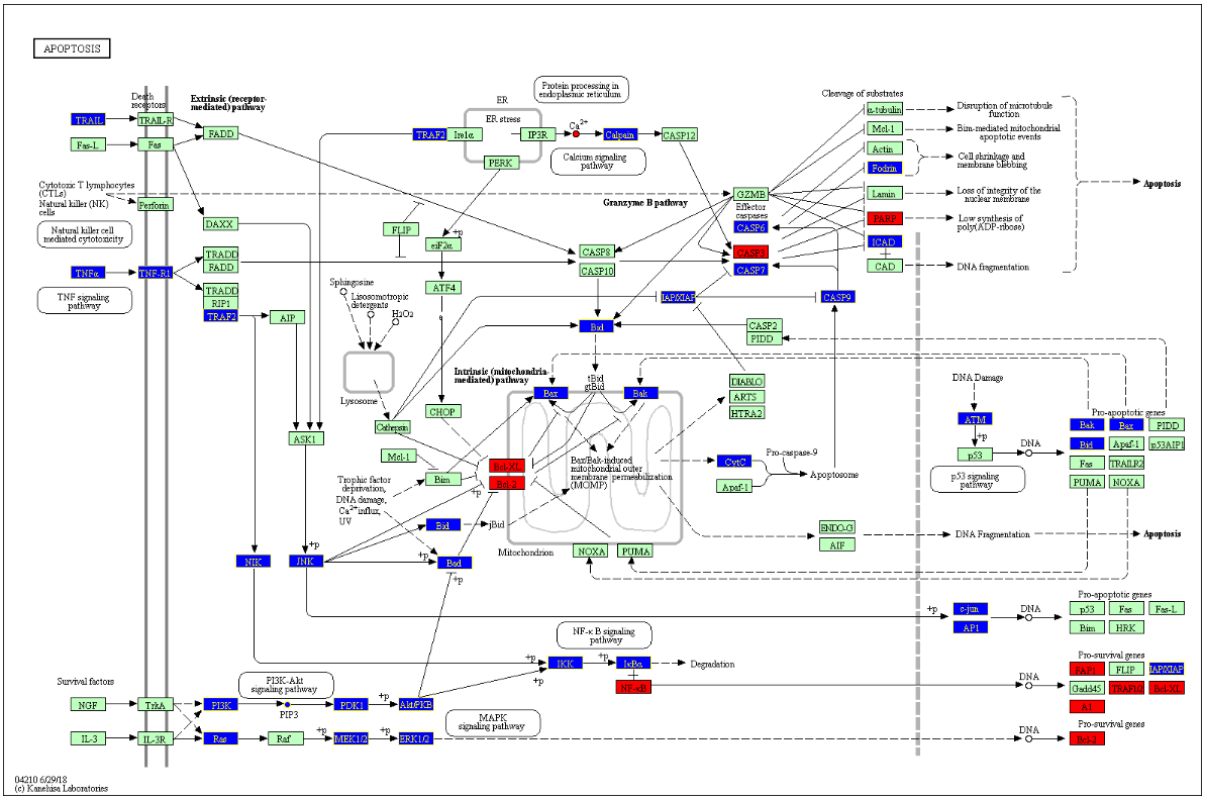


Figura 4.10 Confronto dei risultati restituiti da PHENSIM a seguito del KD di TPL2 in HeLa cells (linea cellulare sensibile al trattamento, riquadro superiore) e RKO cells (linea cellulare resistente al trattamento, riquadro inferiore): TNF pathway.



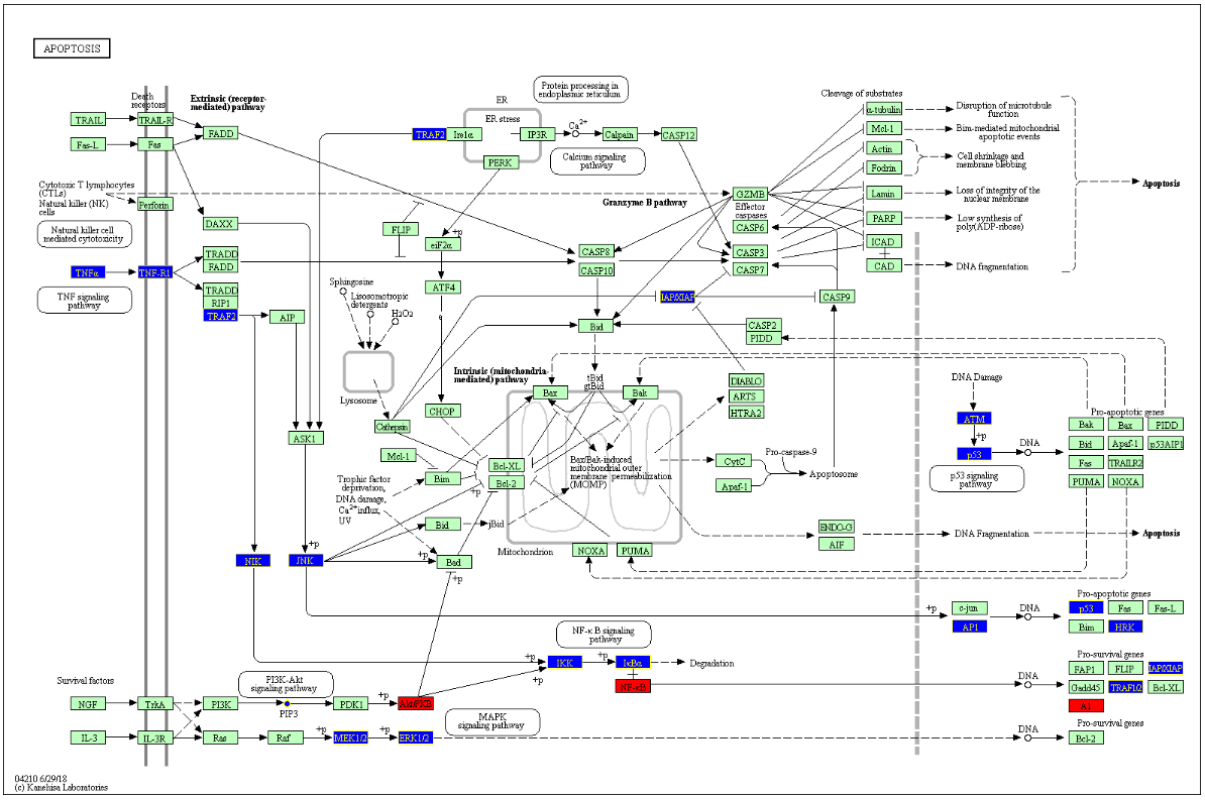


Figura 4.11 Confronto dei risultati restituiti da PHENSIM a seguito del KD di TPL2 in HeLa cells (linea cellulare sensibile al trattamento, riquadro superiore) e RKO cells (linea cellulare resistente al trattamento, riquadro inferiore): Apoptosi.

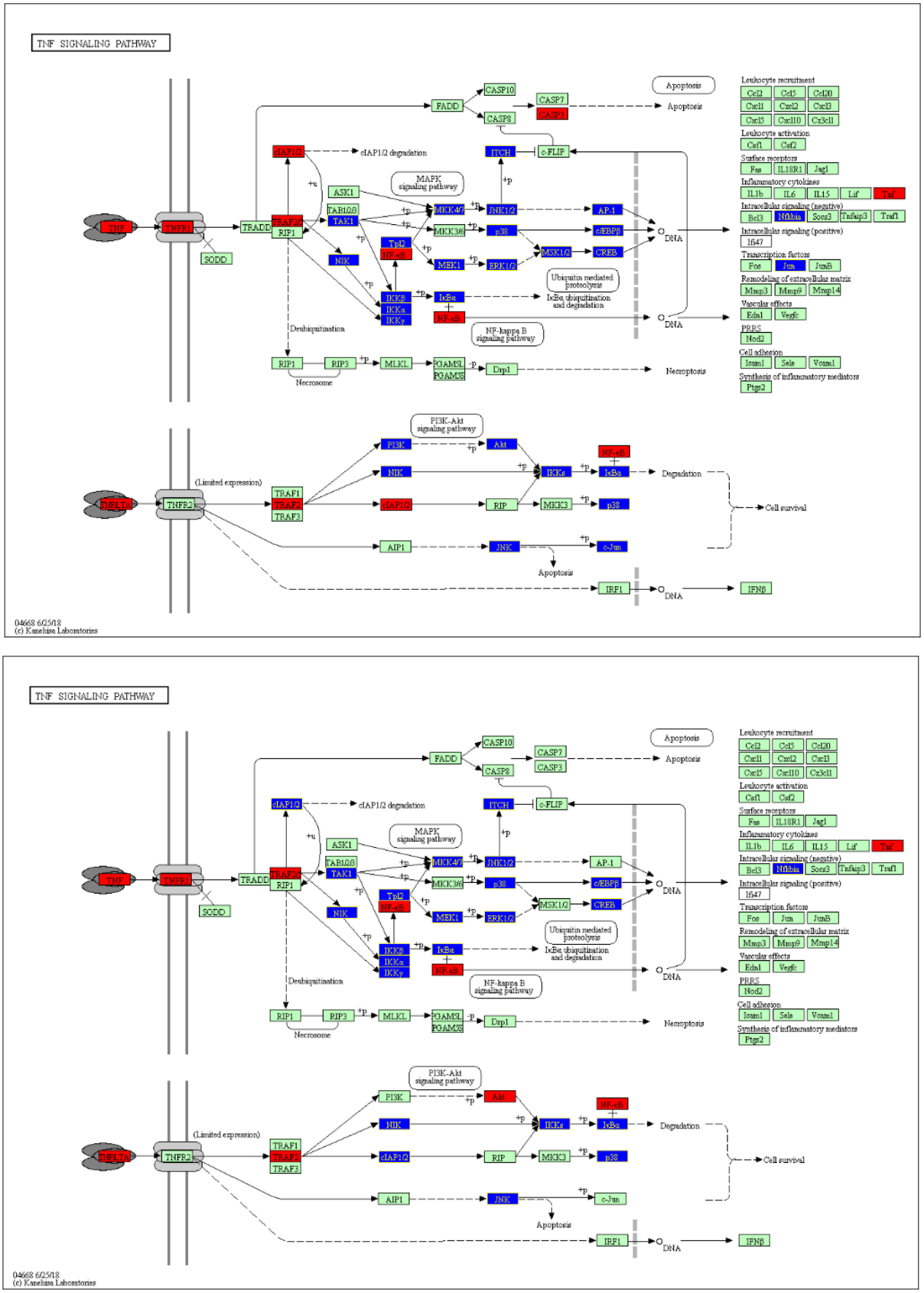


Figura 4.12 Confronto dei risultati restituiti da PHENSIM a seguito del KD di TPL2 più somministrazione di TNFα in HeLa cells (linea cellulare sensibile al trattamento, riquadro superiore) e RKO cells (linea cellulare resistente al trattamento, riquadro inferiore): TNF pathway.

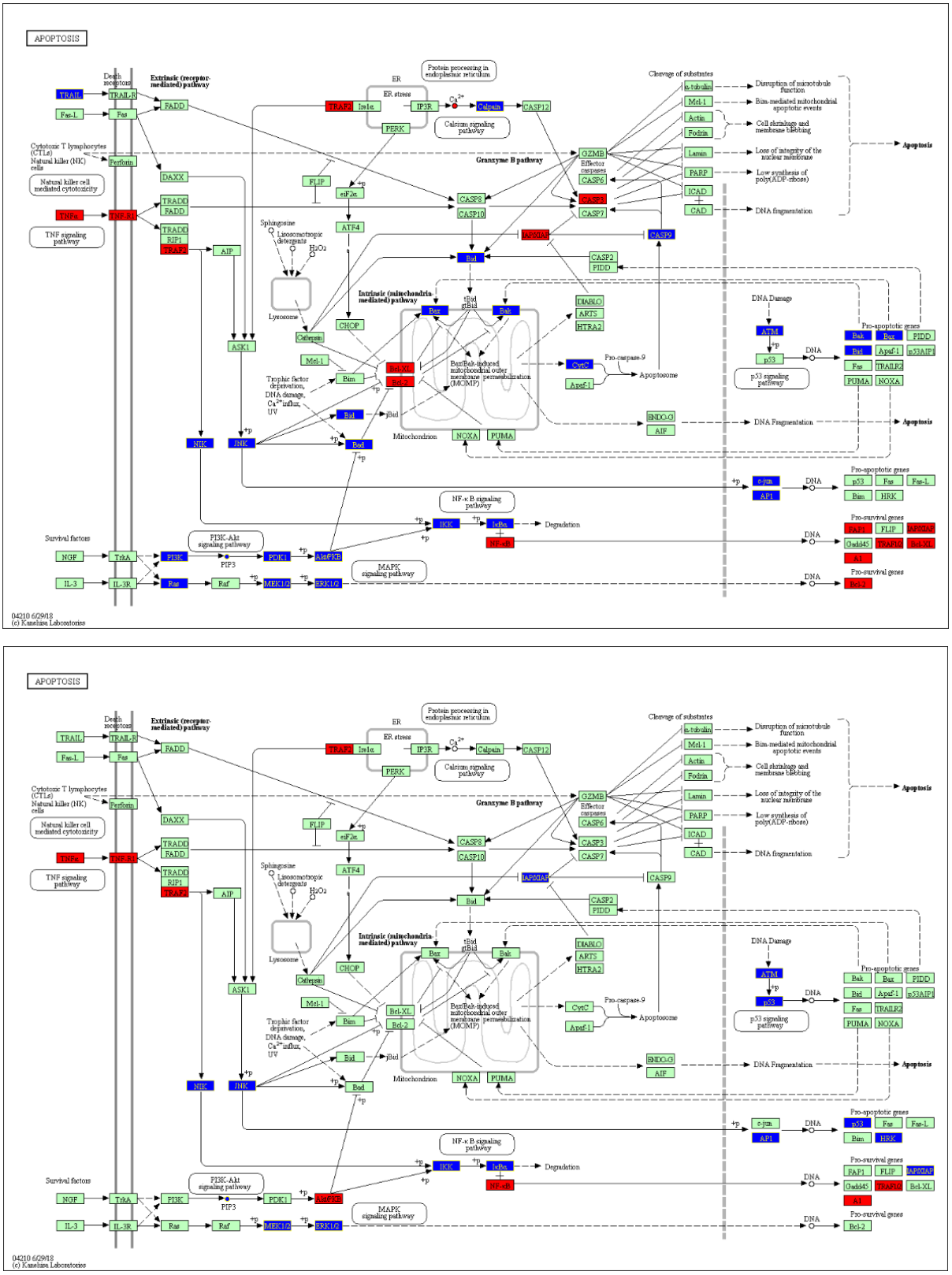


Figura 4.13 Confronto dei risultati restituiti da PHENSIM a seguito del KD di TPL2 più somministrazione di TNFα in HeLa cells (linea cellulare sensibile al trattamento, riquadro superiore) e RKO cells (linea cellulare resistente al trattamento, riquadro inferiore): Apoptosi.

5. Discussioni

L'applicazione delle odierne tecnologie high-throughput sequencing (HTS) e high-throughput detection (HTD) al profiling genomico, trascrittomico e proteomico ha dato il via a nuove, potenti metodologie di indagine nel campo biomedico, fornendo i mezzi necessari per il monitoraggio di sistemi biologici sperimentali e di pazienti. A prescindere dalla particolare tecnologia adoperata per il profiling, il risultato finale consiste nella produzione di una notevole mole di dati grezzi, inizialmente filtrati e trattati mediante tecniche bioinformatiche, al fine di ottenere datasets qualitativamente e quantitativamente analizzabili. Ciò permette la potenziale identificazione di sottoinsiemi di geni, isoforme proteiche e processi molecolari coinvolti in specifiche condizioni fisiopatologiche, principalmente grazie all'applicazione di metodi analitici appropriati sviluppati nel campo della biologia dei sistemi [3]. In tal senso, le tecniche di pathway analysis hanno visto un impiego sempre maggiore da parte della comunità scientifica, grazie anche alla progettazione e all'implementazione di applicazioni web o standalone tools che permettono di trattare i dati in input calcolando opportune statistiche e correndo algoritmi che seguono schemi logici ben definiti. In particolare, i metodi di pathway analysis di terza generazione (PTB) hanno contribuito grandemente all'incremento dell'accuratezza delle analisi grazie all'inclusione della topologia delle vie molecolari biologiche e al calcolo di fattori d'impatto a livello di singolo gene e di pathway [100]. Essenziale allo sviluppo di tali analisi è stata, a sua volta, l'implementazione di un'ampia gamma di knowledgebase, dove le varie componenti del sistema cellulare vengono suddivise per tipologia molecolare, categoria funzionale e compartimenti cellulari, o dove vengono descritte e definite reti di interazioni e pathway biologici a cui i singoli geni, RNA o proteine prendono parte, includendo processi fisiologici e patologici.

Nel presente elaborato sono stati presentati e descritti due metodi di pathway analysis recentemente sviluppati dal gruppo di ricerca dell'Università degli Stui di Catania guidato dal professore A. Ferro: (1) SPECIFIC, metodo SPTB, e (2) PHENSIM, metodo PTB. Entrambi i metodi sono stati implementati come applicazioni web e integrano nei loro calcoli le interazioni miRNA::mRNA e TF → miRNA, alterando la topologia originale dei pathway di KEGG. In quanto metodo SPTB, SPECifIC è in grado di identificare ed estrarre sottostrutture (subpathway) potenzialmente perturbate in specifiche tipologie tumorali. Al termine del processo computazionale, SPECifIC assegna un p-value ad ogni sottostruttura, determinandone la significatività statistica. Allo stato attuale, l'applicazione web di SPECIFIC consta di dieci differenti datasets estrapolati da TCGA, in cui vengono resi disponibili dati di espressione genica in diversi tipi di tumore e sui quali è possibile correre le analisi (Tabella 3.1). Per valutarne le potenzialità, il metodo SPECIFIC è stato confrontato

con tre altri metodi SPTB precedentemente pubblicati. Secondo quanto suggerito in [100] il confronto è stato operato sulla base dei valori di specificità ottenuti per i quattro metodi, ricavati calcolando opportune metriche sui risultati restituiti dalle analisi corse sui dati TCGA disponibili per BRCA e COAD. Le metriche hanno mostrato che in confronto agli altri metodi adoperati, SPECifIC riesce ad individuare sottostrutture significativamente più vicine ai disease genes d'interesse (Tabelle 4.4-4.7); tuttavia, i risultati ottenuti per la metrica (ii) mostravano un incremento nella distanza media tra disease genes interni alle sottostrutture rispetto a quanto ottenuto con gli altri tre metodi. Comunque, i valori ottenuti per la metrica (ii) mostravano un maggiore livello di specificità se confrontati con i risultati ottenuti sui pathway di KEGG.

Tra i top- 20 pathway statisticamente significativi riportati da SPECIFIC, diversi trovano conferma in dati di letteratura precedentemente pubblicati. Ad esempio, nel caso di BRCA, è ormai ampiamente risaputo che l'alterazione dei livelli di espressione di geni coinvolti nelle vie di segnalazione degli estrogeni e di ERBB giocano un ruolo chiave nello sviluppo e nella progressione dei vari sottotipi tumorali di BRCA, nonché nella resistenza ai trattamenti ormonali somministrati a pazienti con determinate sottotipologie di tumore al seno [154]. Similmente, alcuni studi hanno dimostrato il coinvolgimento di geni inclusi nella via di segnalazione della prolattina in certe sottoclassi di pazienti BRCA, che inoltre sembra presentare un cross-talk con la via di segnalazione degli estrogeni [155]. La resistenza agli inibitori dell'attività tirosinchinasica dei recettori EGFR caratterizza anch'essa una buona parte dei pazienti con tumore al seno in stadio avanzato [156]. Geni inclusi nelle vie di segnalazione delle chemochine [157], della fosfolipasi D [158], di PPAR [159], e nei processi di adesione focale [160,161] e di attivazione delle piastrine [162] sono stati trovati anch'essi sovrarappresentati in questo tipo tumorale come anche in altri.

Per quando riguarda il caso di COAD, alterazioni nei livelli di espressione di geni appartenenti alle vie di segnalazione di mTOR [163], RAS [164], fosfolipasi D [165,166], HIF-1 [167,168], nonché di geni coinvolti nei processi apoptotici [169], sono stati precedentemente confermati da vari studi sperimentali, sebbene sia da notare che in questo caso si tratta di pathway generalmente alterate in diversi tipi tumorali, e che perciò non sono caratterizzanti per COAD. Similmente, alcuni lavori scienitifici hanno dimostrato alterazioni nei livelli di espressione di geni legati al metabolismo degli xenobiotici del pathway del citocromo p-450 [170,171].

Infine, degno di nota è il fatto che otto dei top-20 pathway ricavati per BRCA e per COAD risultavano essere sovrapposti tra loro (Tabella 4.1), sebbene le sottostrutture estratte per i relativi pathway si sovrapponessero solo parzialmente o non si sovrapponessero affatto (dati non riportati). Questo dimostra come l'estrazione di sottostrutture significative per la patologia in questione sia caratterizzata da maggiore accuratezza, e dunque da maggiore potenziale applicativo in ambito

biomedico, rispetto all'estrazione di pathway, che invece è caratterizzata da una più elevata ridondanza dei termini associati alle analisi. L'assegnazione di termini quali farmaci o disease genes alle sottostrutture estratte rendono il tool applicabile al design di esperimenti di riposizionamento del farmaco.

Sebbene si tratti di un metodo analitico appartenente alla classe degli approcci PTB, che durante il primo step computazionale sono tipicamente caratterizzati da analisi statistiche a livello di gene basate sui LFC (sezione 1.6.3), PHENSIM si discosta dallo schema operativo tipico di questi ultimi, non tenendo conto dei valori di LFC reali dei geni deregolati presi in considerazione. Da questo punto di vista, PHENSIM utilizza un approccio più simile a quello adoperato in prima fase dai metodi ORA, che tengono conto semplicemente di DEGs statisticamente significativi, scartando tutti gli altri (sezione 1.6.1). Ciononostante, PHENSIM è in grado di generare inferenze circa l'impatto delle variazioni a valle dei geni deregolati inseriti in input dall'utente, fornendo una stima dell'impatto sulle attività dei corrispettivi prodotti genici, nonché dei pathway che li contengono. Dunque, PHENSIM ha principalmente lo scopo di aiutare il ricercatore nel disegno di esperimenti di laboratorio piuttosto che di analizzare dati HTS o HTD precedentemente prodotti. Tra le possibili funzioni attuali, PHENSIM include simulazioni degli effetti di KD genico, sovra- o sottoespressione genica ed uptake del carico esosomiale o di singoli farmaci sul fenotipo cellulare, a condizione che siano note le deregolazioni iniziali. Degno di nota è il fatto che PHENSIM permette di generare inferenze su cellule e tessuti di *Mus musculus* e *Rattus norvegicus* oltre che su quelli di *Homo sapiens*. Oltre alle funzioni appena elencate, il tool verrà prossimamente implementato con una funzione capace di simulare gli effetti del KO genetico sul fenotipo cellulare.
Nel caso di PHENSIM, le potenzialità del tool non sono state dimostrate mediante confronti con altri strumenti analitici della stessa categoria, ma sono state valutate calcolando sensibilità e specificità sulla base dei dati di riferimento forniti dalla letteratura. I valori di sensibilità e specificità sono stati calcolati per i risultati restituiti dalle simulazioni (i), (ii) e (iv) (sezione 2), mostrando in generale un discreto livello di accuratezza. Non è invece stato possibile stimare l'accuratezza del tool nel caso della simulazione (iii) a causa dello scarso numero di informazioni presenti in letteratura.

Dal punto di vista metodologico, è importante sottolineare che procedure computazionali discusse nel presente elaborato mirano a superare tre limiti comunamente presentati da altri approcci PTB per pathway analysis. Anzitutto, l'annotazione dei pathway con dati di espressione dei miRNA (interazioni TF → miRNA) e delle interazioni miRNA-mRNA permette di incrementare la risoluzione della rete di interazioni interna ai pathway, contribuendo così all'incremento

dell'accuratezza dei risultati in output. In realtà, procedure d'annotazione simili erano state già seguite da metodi precedentemente pubblicati, come ad esempio Micrographite [108] e Subpathway-GMir [113], sebbene tali approcci non tenevano in considerazioni le relazioni esistenti tra TFs e trascrizione dei miRNA, a differenza del metodo di pathway analysis MITHrIL [109], sviluppato più recentemente. Tuttavia, entrambi i metodi qui descritti prendono in considerazione solamente interazioni TF → miRNA e miRNA-mRNA validate con metodi sperimentali forti, aggiornate rispettivamente al 2009 e 2013. Ciò implica la presenza di diversi gaps all'interno dei rispettivi datasets, che introducono inevitabilmente un certo grado di inesattezza nei risultati restituiti a valle del processo di pathway analysis. Inoltre, bisogna anche considerare il basso livello di risoluzione e l'inesattezza presentati dagli odierni knowledgebase a cui i PTB tools fanno riferimento per l'analisi dei pathway, incluso KEGG. Ad esempio, studi basati su RNA-seq hanno dimostrato che più del 90% del trascrittoma umano è sottoposto a splicing alternativo, laddove trascritti codificanti maturi (mRNA) derivanti dallo stesso trascritto primario (pre-mRNA) talvolta svolgono funzioni completamente diverse o addirittura opposte [172]. In maniera del tutto simile, studi basati su "whole-exome sequencing" (WES) hanno identificato un gran numero di varianti geniche potenzialmente o effettivamente coinvolte in condizioni fisiopatologiche ben determinate [173]. Tuttavia, gli attuali knowledgebase specificano solamente quale gene è attivo in un determinato pathway, senza distinguere tra isoforme o varianti. Similmente a quanto appena affermato, una seconda limitazione imposta dagli attuali knowledgebase consiste nella mancata o frammentaria contestualizzazione dell'espressione genica e della topologia dei pathway in dipendenza del tipo cellulare considerato. Inoltre, una comune limitazione presentata da queste banche dati consiste nell'erronea inclusione di più geni, indipendenti tra loro, in un unico nodo, a partecipare alla stessa reazione. Dunque, la possibilità di tener conto di liste di geni non espressi in un certo contesto cellulare permette anch'essa di accrescere l'accuratezza dei risultati in output, al fine di simulare la rete di interazioni interna ad uno specifico tipo cellulare e tissutale in maniera quanto più accurata possibile. Tale obiettivo viene raggiunto cambiando la topologia del meta-pathway, in particolare eliminando dal grafo direzionato i nodi corrispondenti ai geni indicati sulla lista ed eliminando o ridistribuendo gli archi in maniera opportuna. Infine, l'esecuzione di statistiche e analisi topologiche a livello di gene applicate a un unico meta-pathway a monte della procedura analitica permette di superare, almeno in parte, il problema della inter-dipendenza dei pathway nel contesto cellulare a motivo dei cross-talk che intercorrono tra essi, sebbene le analisi eseguite a valle durante il calcolo delle statistiche a livello di pathway tornino a considerare i singoli pathway come entità tra loro indipendenti.

Alla pari di diversi altri metodi PTB di pathway analysis, i due metodi qui presentati non tengono conto dei cambiamenti dinamici che avvengono internamente ad un generico sistema cellulare a seguito di una perturbazione. Tale limitazione è stata recentemente trattata e parzialmente superata da Vrahatis et al. [174] grazie alla possibilità di integrare dati di espressione relativi a mRNA e miRNA con i pathway di KEGG e con informazioni disponibili sulle interazioni miRNA-mRNA. Nel loro metodo, denominato CHRONOS e appartenente alla categoria dei metodi SPTB, le sottostrutture vengono estratte e valutate statisticamente sulla base delle transizioni temporali e delle variazioni dei relativi livelli di espressione, fornendo così una misura delle proprietà dinamiche dei pathway perturbati. Inoltre, il suddetto metodo restituisce delle metriche che riassumono caratteristiche strutturali e funzionali dei subpathway in relazione all'insieme delle mappe dei pathway di KEGG.

6. Conclusioni e prospettive future

Nel presente elaborato, sono stati descritti due metodologie top-down di pathway analysis, i.e. SPECifIC e PHENSIM.

SPECifIC è un metodo SPTB in grado di estrarre, visualizzare e arricchire sottostrutture incluse nel meta-pathway di KEGG, specifiche di una data tipologia tumorale, a partire dai dati di espressione forniti in input dall'utente. Questo metodo rappresenta un'estensione dell'algoritmo MITHrIL, precedentemente sviluppato dallo stesso gruppo di ricerca, e permette un'efficiente valutazione statistica della perturbazione e della significatività delle sottostrutture estratte, incrementando l'accuratezza dei risultati grazie all'annotazione dei pathway di KEGG con informazioni sulle interazioni miRNA-mRNA e sull'induzione dell'espressione dei miRNA da parte di specifici fattori della trascrizione.

PHENSIM è un metodo PTB, principalmente progettato per pronosticare l'esito della propagazione della perturbazione di uno o più nodi inclusi nei pathway di KEGG su una data linea cellulare o su una specifica tipologia tissutale. La procedura computazionale di PHENSIM si discosta notevolmente da quella di MITHrIL, dal momento che non tiene in considerazione i fold-change dei nodi in input e che non permette un'appropriata gestione di ampi set di geni, come ad esempio quelli che si otterrebbero da un esperimento di microarray. Il tool assegna ad ogni pathway di KEGG un valore di significatività statistico, predicendone il tipo di deregolazione e il potenziale incremento o decremento dei pathway e dei sengoli nodi in termini numerici. È da sottolineare che in PHENSIM è possibile ottenere simulazioni sulla perturbazione dei pathway per tre diverse specie: *Homo sapiens*, *Mus musculus* e *Rattus norvegicus*. Lo scopo principale di questo tool è quello di formire un valido supporto a gruppi di ricerca per il design di esperimenti di laboratorio.

Gli sviluppi futuri includono l'integrazione nei processi computazionali di informazioni derivanti dall'espressione di altre classi di trascritti non codificanti, primi tra tutti lncRNA, nonché le informazioni riguardanti le reciproche interazioni dei vari elementi tenuti in cosiderazione in seno al sistema cellulare. Alterazioni epigenetiche quali metilazione dei promotori saranno anch'esse da tenere in considerazione, data la loro comporvata importanza in ambito clinico. Infine, nel caso particolare di PHENSIM, si prevede di introdurre una funzione in grado di simulare il KO genico in seno ad una data linea cellulare.

In termini di prospettive future, ciò permetterebbe di tendere verso un ulteriore arricchimento delle conoscenze attuali sui processi fisiopatologici e farmacologici, nonché verso la capacità di predire in maniera sempre più accurata i possibili esiti di una malattia o di una terapia.

Materiali supplementari: SPECifIC

Risultati restituiti da Subpathway-GM, Subpathway-GMir e DESubs nel caso di BRCA

Tabella S1 Risultati restituiti dal metodo Subpathway-GM per BRCA. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo, a prescindere dalla significatività statistica. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

Pathway	SPECifIC	SubPathway-GM	SubPathway-GMir	DEsubs
cell cycle	Yes	7,08E-11		Yes
focal adhesion	Yes	2,25E-10		Yes
complement and coagulation cascades		2,00E-08		
ppar signaling pathway	Yes	2,56E-06		Yes
tight junction		8,74E-08		
mapk signaling pathway		1,07E-06		Yes
p53 signaling pathway		1,07E-06		
neurotrophin signaling pathway	Yes	2,71E-06		
jak stat signaling pathway	Yes	1,69E-05		Yes
chemokine signaling pathway	Yes	6,46E-05		
regulation of actin cytoskeleton	Yes	7,65E-05		Yes
insulin signaling pathway	Yes	0,0002		
t cell receptor signaling pathway	Yes	0,0004		
fc gamma r mediated phagocytosis	Yes	0,0007		
calcium signaling pathway		0,0008		Yes
notch signaling pathway		0,0016		Yes
phagosome		0,0041		
glycosaminoglycan biosynthesis chondroitin sulfate dermatan sulfate		0,0072		
b cell receptor signaling pathway	Yes	0,0078		
gap junction	Yes	0,0078		Yes
adherens junction		0,0078		Yes
apoptosis	Yes	0,0096		Yes
toll like receptor signaling pathway	Yes	0,0099		Yes
adipocytokine signaling pathway		0,0099		Yes
rig i like receptor signaling pathway		0,0099		
vegf signaling pathway	Yes	0,0104		Yes

dorso ventral axis formation		0,0133		
gnrh signaling pathway	Yes	0,0185		Yes
phenylalanine metabolism	Yes	0,0207	Yes	
hedgehog signaling pathway		0,0407		Yes
glycerolipid metabolism		0,0408	Yes	
leukocyte transendothelial migration	Yes	0,0443		
ecm receptor interaction		0,0450	Yes	
mtor signaling pathway	Yes	0,0458		
progesterone mediated oocyte maturation	Yes	0,0458		
tgf beta signaling pathway		0,0458		
glycosylphosphatidylinositol gpi anchor biosynthesis		0,0458		
vascular smooth muscle contraction	Yes	0,0506		
nod like receptor signaling pathway		0,0511		
natural killer cell mediated cytotoxicity	Yes	0,0540		Yes
axon guidance	Yes	0,0560		
n glycan biosynthesis		0,0560	Yes	
drug metabolism cytochrome p450	Yes	0,0619	Yes	
long term potentiation	Yes	0,0780		
tyrosine metabolism		0,0780	Yes	
circadian rhythm		0,0812		
pathways in cancer	Yes	0,0868		
pyruvate metabolism		0,0887	Yes	
wnt signaling pathway	Yes	0,1024		
aldosterone regulated sodium reabsorption	Yes	0,1028		
phosphatidylinositol signaling system	Yes	0,1162		
cell adhesion molecules cams		0,1162		
erbb signaling pathway	Yes	0,1454		
glycolysis gluconeogenesis		0,1572	Yes	
cysteine and methionine metabolism		0,1593	Yes	
butanoate metabolism		0,1615	Yes	
pancreatic secretion		0,1630		
pyrimidine metabolism		0,1671	Yes	
arginine and proline metabolism		0,1747	Yes	
mucin type o glycan biosynthesis		0,1870	Yes	
inositol phosphate metabolism	Yes	0,2146	Yes	
histidine metabolism		0,2146	Yes	
vasopressin regulated water reabsorption		0,2146		
long term depression	Yes	0,2267		Yes
beta alanine metabolism		0,2440	Yes	

glycerophospholipid metabolism		0,2440	Yes	
ether lipid metabolism		0,2440	Yes	
sphingolipid metabolism		0,2593	Yes	
galactose metabolism		0,2717	Yes	
amino sugar and nucleotide sugar metabolism		0,2721	Yes	
gastric acid secretion		0,2761		
glycine serine and threonine metabolism		0,2797	Yes	
glycosphingolipid biosynthesis globo and isoglobo series		0,2864		
salivary secretion		0,2947		
valine leucine and isoleucine degradation		0,2947	Yes	
fructose and mannose metabolism		0,3022	Yes	
taste transduction		0,3127		
purine metabolism		0,3660	Yes	
endocytosis		0,3660		
pantothenate and coa biosynthesis		0,4057		
drug metabolism other enzymes	Yes	0,4292	Yes	
caffeine metabolism		0,4307		
glycosphingolipid biosynthesis ganglio series		0,4307		
nitrogen metabolism		0,4307	Yes	
oocyte meiosis	Yes	0,4833		
melanogenesis		0,4870		Yes
fc epsilon ri signaling pathway	Yes	0,4894		Yes
propanoate metabolism		0,5207	Yes	
tryptophan metabolism		0,5248	Yes	
synthesis and degradation of ketone bodies		0,5248	Yes	
phenylalanine tyrosine and tryptophan biosynthesis		0,5248		
riboflavin metabolism		0,5248	Yes	
one carbon pool by folate		0,5316	Yes	
terpenoid backbone biosynthesis		0,5957	Yes	
cytosolic dna sensing pathway		0,6002		
alanine aspartate and glutamate metabolism		0,6023	Yes	
arachidonic acid metabolism		0,6041	Yes	
glycosaminoglycan degradation		0,6041	Yes	
citrate cycle tca cycle		0,6053	Yes	
glycosaminoglycan biosynthesis heparan sulfate heparin		0,6053		
glycosphingolipid biosynthesis lacto and neolacto series		0,6231	Yes	

selenocompound metabolism		0,6296	Yes	
protein processing in endoplasmic reticulum		0,6463		
pentose phosphate pathway		0,6463	Yes	
fatty acid biosynthesis		0,6569		
lysine degradation		0,6803	Yes	
glyoxylate and dicarboxylate metabolism		0,6936		
lysine biosynthesis		0,6936		
steroid biosynthesis		0,7141	Yes	
folate biosynthesis		0,7141	Yes	
primary bile acid biosynthesis		0,7374		
cytokine cytokine receptor interaction		0,7499		Yes
sulfur metabolism		0,8436		
fatty acid degradation		0,8545	Yes	
glutathione metabolism		0,8911	Yes	
nicotinate and nicotinamide metabolism		0,9075	Yes	
steroid hormone biosynthesis	Yes	0,9227	Yes	
vitamin b6 metabolism		0,9464		
starch and sucrose metabolism		0,9551	Yes	
antigen processing and presentation		0,9874		
pentose and glucuronate interconversions		1,0000	Yes	
porphyrin and chlorophyll metabolism		1,0000		
olfactory transduction		1,0000		
fatty acid elongation		1,0000		Yes

Tabella S2 Risultati restituiti dal metodo Subpathway-GMir per BRCA. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo, a prescindere dalla significatività statistica. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

Pathway	**SPECifIC**	**SubPathway-GM**	**SubPathway-GMir**	**DEsubs**
glycolysis gluconeogenesis			0	
pyrimidine metabolism			0	
purine metabolism			0	
glycerophospholipid metabolism			0	
fatty acid degradation			1,49E-14	
inositol phosphate metabolism	Yes		1,13E-13	
lysine degradation			3,73E-13	
pyruvate metabolism			1,93E-12	
valine leucine and isoleucine degradation			5,51E-12	
glycerolipid metabolism			6,46E-12	
tyrosine metabolism			2,42E-11	

n glycan biosynthesis		3,51E-11
arachidonic acid metabolism		5,60E-11
ether lipid metabolism		6,38E-11
arginine and proline metabolism		9,10E-10
amino sugar and nucleotide sugar metabolism		2,21E-09
glutathione metabolism		2,41E-09
sphingolipid metabolism		2,61E-09
propanoate metabolism		1,15E-08
glycosphingolipid biosynthesis lacto and neolacto series		1,15E-08
galactose metabolism		1,40E-08
fructose and mannose metabolism		1,40E-08
phenylalanine metabolism	Yes	4,24E-08
cysteine and methionine metabolism		5,24E-08
drug metabolism other enzymes	Yes	1,01E-07
alanine aspartate and glutamate metabolism		1,01E-07
fatty acid elongation		1,21E-07
starch and sucrose metabolism		1,37E-07
butanoate metabolism		2,19E-07
tryptophan metabolism		2,84E-07
aminoacyl trna biosynthesis		5,87E-07
metabolism of xenobiotics by cytochrome p450	Yes	1,02E-06
steroid hormone biosynthesis	Yes	1,03E-06
drug metabolism cytochrome p450	Yes	2,04E-06
citrate cycle tca cycle		2,04E-06
mucin type o glycan biosynthesis		2,04E-06
pentose phosphate pathway		2,82E-06
linoleic acid metabolism	Yes	7,36E-06
glycine serine and threonine metabolism		7,49E-06
histidine metabolism		2,10E-05
pentose and glucuronate interconversions		0,0002
steroid biosynthesis		0,0002
terpenoid backbone biosynthesis		0,0002
one carbon pool by folate		0,0002
folate biosynthesis		0,0014
ascorbate and aldarate metabolism		0,0019
synthesis and degradation of ketone bodies		0,0019
riboflavin metabolism		0,0019
nicotinate and nicotinamide metabolism		0,0024
glycosaminoglycan degradation		0,0057

beta alanine metabolism	0,0058
nitrogen metabolism	0,0058
selenocompound metabolism	0,0058
glycosphingolipid biosynthesis ganglio series	0,0188

Tabella S3 Risultati restituiti dal metodo DESubs per BRCA. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo. DESubs scarta automaticamente tutti i risultati restituiti con p-value < 0,001, per cui in tabella sono mostrati solamente termini statisticamente significativi. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

Pathway	SPECifIC	SubPathway-GM	SubPathway-GMir	DEsubs
focal adhesion	Yes	Yes		76
cytokine cytokine receptor interaction				59
cell cycle	Yes	Yes		56
ecm receptor interaction				30
jak stat signaling pathway	Yes	Yes		24
ppar signaling pathway	Yes	Yes		21
mapk signaling pathway		Yes		18
toll like receptor signaling pathway	Yes	Yes		14
adherens junction		Yes		13
gnrh signaling pathway	Yes			10
long term depression	Yes			10
hedgehog signaling pathway				10
natural killer cell mediated cytotoxicity	Yes			8
regulation of actin cytoskeleton	Yes	Yes		8
notch signaling pathway		Yes		8
apoptosis	Yes	Yes		6
adipocytokine signaling pathway		Yes		6
vegf signaling pathway	Yes			5
fc epsilon ri signaling pathway	Yes			5
gap junction	Yes	Yes		4
calcium signaling pathway		Yes		4
melanogenesis				4

Risultati restituiti da Subpathway-GM, Subpathway-GMir e DESubs nel caso di COAD

Tabella S4 Risultati restituiti dal metodo Subpathway-GM per COAD. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo, a prescindere dalla significatività statistica. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

KEGG Pathway	SPECifIC	SubPathway-GM	SubPathway-GMir	DEsubs
ppar signaling pathway		1,45E-08		
starch and sucrose metabolism		4,93E-07	Yes	
natural killer cell mediated cytotoxicity		3,86E-06		Yes
pentose and glucuronate interconversions		3,86E-06	Yes	
tight junction		4,53E-06		
complement and coagulation cascades		0,0003		
arginine and proline metabolism		0,0004	Yes	
apoptosis	Yes	0,0005		Yes
calcium signaling pathway		0,0006		
aldosterone regulated sodium reabsorption		0,0006		
mapk signaling pathway	Yes	0,0012		Yes
fatty acid degradation		0,0012	Yes	
focal adhesion		0,0013		Yes
wnt signaling pathway		0,0013		
alanine aspartate and glutamate metabolism		0,0014		
butanoate metabolism		0,0014		
pyruvate metabolism		0,0020	Yes	
oocyte meiosis		0,0026		
glycosphingolipid biosynthesis lacto and neolacto series		0,0026	Yes	
tyrosine metabolism		0,0029		Yes
chemokine signaling pathway	Yes	0,0030		
glutathione metabolism		0,0044	Yes	
cell adhesion molecules cams		0,0049		
cell cycle		0,0050		
amino sugar and nucleotide sugar metabolism		0,0065	Yes	
tgf beta signaling pathway		0,0065		
one carbon pool by folate		0,0089	Yes	
phosphatidylinositol signaling system		0,0101	Yes	
galactose metabolism		0,0131	Yes	
toll like receptor signaling pathway		0,0147		Yes
phenylalanine metabolism		0,0147	Yes	
inositol phosphate metabolism		0,0148	Yes	Yes
fc gamma r mediated phagocytosis		0,0176		
lysine degradation		0,0176		

glycosphingolipid biosynthesis globo and isoglobo series		0,0195		
vegf signaling pathway		0,0218		Yes
fructose and mannose metabolism		0,0230	Yes	
valine leucine and isoleucine degradation		0,0230	Yes	
adipocytokine signaling pathway		0,0253		Yes
glycerophospholipid metabolism		0,0253	Yes	
olfactory transduction		0,0253		
jak stat signaling pathway		0,0282		Yes
citrate cycle tca cycle		0,0282	Yes	
pentose phosphate pathway		0,0282	Yes	
sulfur metabolism		0,0282	Yes	
long term potentiation		0,0286		
glycosphingolipid biosynthesis ganglio series		0,0294	Yes	
progesterone mediated oocyte maturation	Yes	0,0338		
pathways in cancer	Yes	0,0358		
porphyrin and chlorophyll metabolism		0,0437	Yes	
neurotrophin signaling pathway		0,0437		
arachidonic acid metabolism	Yes	0,0467	Yes	
glycerolipid metabolism		0,0467	Yes	
cysteine and methionine metabolism		0,0471	Yes	
drug metabolism cytochrome p450	Yes	0,0482		
long term depression		0,0546		Yes
propanoate metabolism		0,0581		
mtor signaling pathway	Yes	0,0678		
glycolysis gluconeogenesis		0,0700	Yes	
nod like receptor signaling pathway		0,0700		
vascular smooth muscle contraction		0,0700		
phenylalanine tyrosine and tryptophan biosynthesis		0,0700		
glycine serine and threonine metabolism		0,0700	Yes	
tryptophan metabolism		0,0700		
regulation of actin cytoskeleton		0,0812		Yes
purine metabolism		0,0831	Yes	
sphingolipid metabolism		0,0831	Yes	
fatty acid elongation		0,0831		
synthesis and degradation of ketone bodies		0,0831		
circadian rhythm		0,0831		
melanogenesis	Yes	0,0982		Yes

axon guidance		0,0982		
gnrh signaling pathway	Yes	0,1498		Yes
gap junction		0,1558		
pyrimidine metabolism		0,1713	Yes	
steroid hormone biosynthesis	Yes	0,1858	Yes	
salivary secretion		0,1923		
phagosome		0,2123		
ether lipid metabolism		0,2859	Yes	
fc epsilon ri signaling pathway		0,2921		Yes
insulin signaling pathway		0,3107		
mucin type o glycan biosynthesis		0,3288	Yes	
nicotinate and nicotinamide metabolism		0,4014		

Tabella S5 Risultati restituiti dal metodo Subpathway-GMir per COAD. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo, a prescindere dalla significatività statistica. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

Pathway	**SPECifIC**	**SubPathway-GM**	**SubPathway-GMir**	**DEsubs**
metabolism of xenobiotics by cytochrome p450	Yes		0	
purine metabolism			0	
starch and sucrose metabolism		Yes	0	
fatty acid degradation		Yes	0	
steroid hormone biosynthesis	Yes		1,09E-14	
pyrimidine metabolism			2,4E-14	
pentose and glucuronate interconversions		Yes	2,68E-14	
inositol phosphate metabolism			2,9E-14	Yes
porphyrin and chlorophyll metabolism			2,48E-13	
ascorbate and aldarate metabolism			3,93E-13	
glutathione metabolism		Yes	8,26E-11	
galactose metabolism			8,26E-11	
arachidonic acid metabolism	Yes		1,18E-09	
aminoacyl trna biosynthesis			2,28E-09	
glycerolipid metabolism			2,98E-09	
valine leucine and isoleucine degradation			3,01E-09	
tyrosine metabolism		Yes	1,37E-08	
glycerophospholipid metabolism			3,75E-08	
amino sugar and nucleotide sugar metabolism		Yes	1,71E-07	
fructose and mannose metabolism			7,94E-07	
citrate cycle tca cycle			9,41E-07	

one carbon pool by folate		Yes	2,09E-06
pentose phosphate pathway			2,89E-06
cysteine and methionine metabolism			2,89E-06
glycosphingolipid biosynthesis lacto and neolacto series		Yes	1,71E-03
ether lipid metabolism			4,29E-06
phenylalanine metabolism			8,71E-06
mucin type o glycan biosynthesis			2,15E-05
sulfur metabolism			4,04E-05
arginine and proline metabolism		Yes	0,0001
pyruvate metabolism		Yes	0,0001
glycolysis gluconeogenesis			0,0001
sphingolipid metabolism			0,0002
glycosphingolipid biosynthesis ganglio series			0,0013
glycine serine and threonine metabolism			0,0013

Tabella S6 Risultati restituiti dal metodo DESubs per COAD. In tabella sono mostrati tutti i termini KEGG (pathway) ottenuti mediante questo metodo. DESubs scarta automaticamente tutti i risultati restituiti con p-value < 0,001, per cui in tabella sono mostrati solamente termini statisticamente significativi. Per ogni pathway riportato, è stato operato un confronto con i risultati restituiti dagli altri metodi.

Pathway	**SPECifIC**	**SubPathway-GM**	**SubPathway-GMir**	**DEsubs**
fc epsilon ri signaling pathway				67
cytokine cytokine receptor interaction				58
erbb signaling pathway	Yes			46
b cell receptor signaling pathway				29
natural killer cell mediated cytotoxicity		Yes		29
inositol phosphate metabolism			Yes	29
phosphatidylinositol signaling system				29
leukocyte transendothelial migration				29
gnrh signaling pathway	Yes			28
mapk signaling pathway	Yes	Yes		16
focal adhesion		Yes		16
long term depression				12
toll like receptor signaling pathway				10
jak stat signaling pathway				7
apoptosis	Yes	Yes		6
vegf signaling pathway				6
regulation of actin cytoskeleton				6
adipocytokine signaling pathway				2

melanogenesis	Yes	1

Riferimenti bibliografici

1. Castiglioni A. A History of Medicine [Internet]. 2019. doi: 10.4324/9780429019883.

2. National Research Council, Division on Earth and Life Studies, Board on Life Sciences, Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. National Academies Press; 2012. 142 p.

3. Wang R-S, Maron BA, Loscalzo J. Systems medicine: evolution of systems biology from bench to bedside. Wiley Interdiscip Rev Syst Biol Med. 2015; 7: 141–61.

4. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015; 372: 793–5.

5. Jain RK, Duda DG, Willett CG, Sahani DV, Zhu AX, Loeffler JS, Batchelor TT, Gregory Sorensen A. Biomarkers of response and resistance to antiangiogenic therapy [Internet]. Nature Reviews Clinical Oncology. 2009. p. 327–38. doi: 10.1038/nrclinonc.2009.63.

6. Chen P-L, Roh W, Reuben A, Cooper ZA, Spencer CN, Prieto PA, Miller JP, Bassett RL, Gopalakrishnan V, Wani K, De Macedo MP, Austin-Breneman JL, Jiang H, et al. Analysis of Immune Signatures in Longitudinal Tumor Samples Yields Insight into Biomarkers of Response and Mechanisms of Resistance to Immune Checkpoint Blockade. Cancer Discov. 2016; 6: 827–37.

7. Dietel M, Jöhrens K, Laffert MV, Hummel M, Bläker H, Pfitzner BM, Lehmann A, Denkert C, Darb-Esfahani S, Lenze D, Heppner FL, Koch A, Sers C, et al. A 2015 update on predictive molecular pathology and its role in targeted cancer therapy: a review focussing on clinical relevance. Cancer Gene Ther. 2015; 22: 417–30.

8. Barrett LW, Fletcher S, Wilton SD. Untranslated Gene Regions and Other Non-coding Elements [Internet]. Untranslated Gene Regions and Other Non-coding Elements. 2013. p. 1–56. doi: 10.1007/978-3-0348-0679-4_1.

9. Wilczynska A, Bushell M. The complexity of miRNA-mediated repression. Cell Death Differ. 2015; 22: 22–33.

10. St Laurent G, Wahlestedt C, Kapranov P. The Landscape of long noncoding RNA classification. Trends Genet. 2015; 31: 239–51.

11. Weidhaas JB, Babar I, Nallur SM, Trang P, Roush S, Boehm M, Gillespie E, Slack FJ. MicroRNAs as potential agents to alter resistance to cytotoxic anticancer therapy. Cancer Res. 2007; 67: 11111–6.

12. Bader AG, Brown D, Winkler M. The promise of microRNA replacement therapy. Cancer Res. 2010; 70: 7027–30.

13. Ma J, Dong C, Ji C. MicroRNA and drug resistance [Internet]. Cancer Gene Therapy. 2010. p. 523–31. doi: 10.1038/cgt.2010.18.

14. Iorio MV, Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. EMBO Mol Med. 2017; 9: 852.

15. Guil S, Esteller M. RNA–RNA interactions in gene regulation: the coding and noncoding players [Internet]. Trends in Biochemical Sciences. 2015. p. 248–56. doi: 10.1016/j.tibs.2015.03.001.

16. Karreth FA, Pandolfi PP. ceRNA cross-talk in cancer: when ce-bling rivalries go awry. Cancer Discov. 2013; 3: 1113–21.

17. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014; 505: 344–52.

18. Fisher J, Henzinger TA. Executable cell biology [Internet]. Nature Biotechnology. 2007. p. 1239–49. doi: 10.1038/nbt1356.

19. Waterman MS. Introduction to Computational Biology: Maps, Sequences and Genomes [Internet]. Biometrics. 1998. p. 398. doi: 10.2307/2534039.

20. Attwood TK, Gisel A, Eriksson N-E, Bongcam-Rudloff E. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective [Internet]. Bioinformatics - Trends and Methodologies. 2011. doi: 10.5772/23535.

21. Baker D. Protein Structure Prediction and Structural Genomics [Internet]. Science. 2001. p. 93–6. doi: 10.1126/science.1065659.

22. Kelley LA, Sternberg MJE. Protein structure prediction on the Web: a case study using the Phyre server [Internet]. Nature Protocols. 2009. p. 363–71. doi: 10.1038/nprot.2009.2.

23. Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. Nat Rev Genet. 2017; 18: 275–91.

24. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization [Internet]. Nature Reviews Genetics. 2004. p. 101–13. doi: 10.1038/nrg1272.

25. Noble D. The Music of Life: Biology beyond genes. OUP Oxford; 2008. 176 p.

26. Sauer U, Heinemann M, Zamboni N. Genetics. Getting closer to the whole picture. Science. 2007; 316: 550–1.

27. Holmans P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. Adv Genet. 2010; 72: 141–79.

28. Newman M. Networks: An Introduction. OUP Oxford; 2010. 784 p.

29. Jarvik J, Botstein D. A Genetic Method for Determining the Order of Events in a Biological Pathway [Internet]. Proceedings of the National Academy of Sciences. 1973. p. 2046–50. doi: 10.1073/pnas.70.7.2046.

30. Novick P, Ferro S, Schekman R. Order of events in the yeast secretory pathway. Cell. 1981; 25: 461–9.

31. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010; 11: 31–46.

32. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016; 17: 333–51.

33. Consortium IHGS, International Human Genome Sequencing Consortium. Erratum: correction: Initial sequencing and analysis of the human genome [Internet]. Nature. 2001. p. 565–6. doi: 10.1038/35087627.

34. Consortium †the International Hapmap, †The International HapMap Consortium. The International HapMap Project [Internet]. Nature. 2003. p. 789–96. doi: 10.1038/nature02168.

35. Consortium TEP, The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project [Internet]. Science. 2004. p. 636–40. doi: 10.1126/science.1105136.

36. Chen JC, Alvarez MJ, Talos F, Dhruv H, Rieckhof GE, Iyer A, Diefes KL, Aldape K, Berens M, Shen MM, Califano A. Identification of Causal Genetic Drivers of Human Disease through Systems-Level Analysis of Regulatory Networks. Cell. 2016; 166: 1055.

37. Koh W, Pan W, Gawad C, Fan HC, Kerchner GA, Wyss-Coray T, Blumenfeld YJ, El-Sayed YY, Quake SR. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. Proc Natl Acad Sci U S A. 2014; 111: 7361–6.

38. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016; 17: 257–71.

39. Torkamani A, Andersen KG, Steinhubl SR, Topol EJ. High-Definition Medicine [Internet]. Cell. 2017. p. 828–43. doi: 10.1016/j.cell.2017.08.007.

40. Trevino V, Falciani F, Barrera-Saldaña HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol Med. 2007; 13: 527–41.

41. Nelson PT, Baldwin DA, Scearce LM, Oberholtzer JC, Tobias JW, Mourelatos Z. Microarray-based, high-throughput gene expression profiling of microRNAs. Nat Methods. 2004; 1: 155–61.

42. Calin GA, Croce CM. MicroRNA signatures in human cancers. Nat Rev Cancer. 2006; 6: 857–66.

43. Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 2005; 11: 241–7.

44. Templin MF, Stoll D, Schrenk M, Traub PC, Vöhringer CF, Joos TO. Protein microarray technology. Drug Discov Today. 2002; 7: 815–22.

45. Hall DA, Ptacek J, Snyder M. Protein microarray technology [Internet]. Mechanisms of Ageing and Development. 2007. p. 161–7. doi: 10.1016/j.mad.2006.11.021.

46. Lee TI. Transcriptional Regulatory Networks in Saccharomyces cerevisiae [Internet]. Science. 2002. p. 799–804. doi: 10.1126/science.1075090.

47. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437: 376–80.

48. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005; 309: 1728–32.

49. Reuter JA, Spacek DV, Snyder MP. High-Throughput Sequencing Technologies [Internet]. Molecular Cell. 2015. p. 586–97. doi: 10.1016/j.molcel.2015.05.004.

50. Thudi M, Li Y, Jackson SA, May GD, Varshney RK. Current state-of-art of sequencing technologies for plant genomics research. Brief Funct Genomics. 2012; 11: 3–11.

51. Voelkerding KV, Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics [Internet]. Clinical Chemistry. 2009. p. 641–58. doi: 10.1373/clinchem.2008.112789.

52. Martin JA, Wang Z. Next-generation transcriptome assembly [Internet]. Nature Reviews Genetics. 2011. p. 671–82. doi: 10.1038/nrg3068.

53. Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 2010; 38: e159.

54. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011; 475: 348–52.

55. Nishi K, Nishi A, Nagasawa T, Ui-Tei K. Human TNRC6A is an Argonaute-navigator protein for microRNA-mediated gene silencing in the nucleus. RNA. 2013; 19: 17–35.

56. Cernilogar FM, Onorati MC, Kothe GO, Burroughs AM, Parsi KM, Breiling A, Lo Sardo F, Saxena A, Miyoshi K, Siomi H, Siomi MC, Carninci P, Gilmour DS, et al. Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. Nature. 2011; 480: 391–5.

57. Pitchiaya S, Heinicke LA, Park JI, Cameron EL, Walter NG. Resolving Subcellular miRNA Trafficking and Turnover at Single-Molecule Resolution. Cell Rep. 2017; 19: 630–42.

58. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014; 15: 509–24.

59. Kim Y-K, Narry Kim V. Processing of intronic microRNAs [Internet]. The EMBO Journal. 2007. p. 775–83. doi: 10.1038/sj.emboj.7601512.

60. de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, Åström G, Babina M, Bertin N, Burroughs AM, Carlisle AJ, Daub CO, Detmar M, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol. 2017; 35: 872–8.

61. Creugny A, Fender A, Pfeffer S. Regulation of primary microRNA processing. FEBS Lett. 2018; 592: 1980–96.

62. Alarcón CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF. N6-methyladenosine marks primary microRNAs for processing. Nature. 2015; 519: 482–5.

63. Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs. Nat Rev Mol Cell Biol. 2016; 17: 83–96.

64. Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. Nat Rev Genet. 2015; 16: 421–33.

65. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136: 215–33.

66. Roberts JT, Borchert GM. Computational Prediction of MicroRNA Target Genes, Target Prediction Databases, and Web Resources. Methods Mol Biol. 2017; 1617: 109–22.

67. Moretti F, Thermann R, Hentze MW. Mechanism of translational regulation by miR-2 from sites in the 5′ untranslated region or the open reading frame [Internet]. RNA. 2010. p. 2493–502. doi: 10.1261/rna.2384610.

68. Forman JJ, Legesse-Miller A, Coller HA. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci U S A. 2008; 105: 14879–84.

69. Ipsaro JJ, Joshua-Tor L. From guide to target: molecular insights into eukaryotic RNA-interference machinery. Nat Struct Mol Biol. 2015; 22: 20–8.

70. Morozova N, Zinovyev A, Nonne N, Pritchard L-L, Gorban AN, Harel-Bellan A. Kinetic signatures of microRNA modes of action. RNA. 2012; 18: 1635–55.

71. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, et al. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A. 2002; 99: 15524–9.

72. Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A. 2004; 101: 2999–3004.

73. Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, Shimizu M, Wojcik SE, Aqeilan RI, Zupo S, Dono M, Rassenti L, Alder H, Volinia S, et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci U S A. 2005; 102: 13944–9.

74. Small EM, Olson EN. Pervasive roles of microRNAs in cardiovascular biology. Nature. 2011; 469: 336–42.

75. Wu YE, Parikshak NN, Belgard TG, Geschwind DH. Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. Nat Neurosci. 2016; 19: 1463–76.

76. Deiuliis JA. MicroRNAs as regulators of metabolic disease: pathophysiologic significance and emerging role as biomarkers and therapeutics. Int J Obes . 2016; 40: 88–101.

77. Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol. 2007; 9: 654–9.

78. Weber JA, Baxter DH, Zhang S, Huang DY, Huang KH, Lee MJ, Galas DJ, Wang K. The microRNA spectrum in 12 body fluids. Clin Chem. 2010; 56: 1733–41.

79. Turchinovich A, Weiz L, Langheinz A, Burwinkel B. Characterization of extracellular circulating microRNA. Nucleic Acids Res. 2011; 39: 7223–33.

80. Montecalvo A, Larregina AT, Shufesky WJ, Stolz DB, Sullivan MLG, Karlsson JM, Baty CJ, Gibson GA, Erdos G, Wang Z, Milosevic J, Tkacheva OA, Divito SJ, et al. Mechanism of transfer of functional microRNAs between mouse dendritic cells via exosomes. Blood. 2012; 119: 756–66.

81. Fabbri M. TLRs as miRNA receptors. Cancer Res. 2012; 72: 6333–7.

82. Wang K, Zhang S, Weber J, Baxter D, Galas DJ. Export of microRNAs and microRNA-protective protein by mammalian cells. Nucleic Acids Res. 2010; 38: 7248–59.

83. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF, Pogosova-Agadjanyan EL, Stirewalt DL, Tait JF, Tewari M. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. Proc Natl Acad Sci U S A. 2011; 108: 5003–8.

84. Anfossi S, Babayan A, Pantel K, Calin GA. Clinical utility of circulating non-coding RNAs - an update. Nat Rev Clin Oncol. 2018; 15: 541–63.

85. Alaimo S, Micale G, La Ferlita A, Ferro A, Pulvirenti A. Computational Methods to Investigate the Impact of miRNAs on Pathways. Methods Mol Biol. 2019; 1970: 183–209.

86. Sobie EA, Lee Y-S, Jenkins SL, Iyengar R. Systems biology--biomedical modeling. Sci Signal. 2011; 4: tr2.

87. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5: 621–8.

88. Bhalla US, Iyengar R. Emergent properties of networks of biological signaling pathways. Science. 1999; 283: 381–7.

89. Sobie EA, Dilly KW, dos Santos Cruz J, Jonathan Lederer W, Saleet Jafri M. Termination of Cardiac Ca2 Sparks: An Investigative Mathematical Model of Calcium-Induced Calcium Release [Internet]. Biophysical Journal. 2002. p. 59–78. doi: 10.1016/s0006-3495(02)75149-7.

90. Rzhetsky A, Wajngurt D, Park N, Zheng T. Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci U S A. 2007; 104: 11694–9.

91. Zhou X, Menche J, Barabási A-L, Sharma A. Human symptoms-disease network. Nat Commun. 2014; 5: 4212.

92. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015; 347: 1257601.

93. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes [Internet]. Bioinformatics. 2004. p. 3710–5. doi: 10.1093/bioinformatics/bth456.

94. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27: 1739–40.

95. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009; 37: D619–22.

96. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014; 42: D472–7.

97. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003; 13: 2129–41.

98. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. Nucleic Acids Research. 2000. p. 27–30. doi: 10.1093/nar/28.1.27.

99. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014; 42: D199–205.

100. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012; 8: e1002375.

101. Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, Voichiţa C, Drăghici S. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol. 2013; 4: 278.

102. Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genet Mol Biol. 2004; 3: Article16.

103. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis [Internet]. Bioinformatics. 2009. p. 75–82. doi: 10.1093/bioinformatics/btn577.

104. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. A systems biology approach for pathway level analysis [Internet]. Genome Research. 2007. p. 1537–45. doi: 10.1101/gr.6202607.

105. Khatri P, Draghici S, Tarca AL, Hassan SS, Romero R. A System Biology Approach for the Steady-State Analysis of Gene Signaling Networks [Internet]. Lecture Notes in Computer Science. 2007. p. 32–41. doi: 10.1007/978-3-540-76725-1_4.

106. Shojaie A, Michailidis G. Analysis of Gene Sets Based on the Underlying Regulatory Network [Internet]. Journal of Computational Biology. 2009. p. 407–26. doi: 10.1089/cmb.2008.0081.

107. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010; 26: i237–45.

108. Calura E, Martini P, Sales G, Beltrame L, Chiorino G, D'Incalci M, Marchini S, Romualdi C. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. Nucleic Acids Res. 2014; 42: e96.

109. Alaimo S, Giugno R, Acunzo M, Veneziano D, Ferro A, Pulvirenti A. Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. Oncotarget. 2016; 7: 54572–82.

110. Bauer-Mehren A, Furlong LI, Sanz F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol Syst Biol. 2009; 5: 290.

111. Alaimo S, Marceca GP, Ferro A, Pulvirenti A. Detecting Disease Specific Pathway Substructures through an Integrated Systems Biology Approach. Noncoding RNA [Internet]. 2017; 3. doi: 10.3390/ncrna3020020.

112. Li C, Han J, Yao Q, Zou C, Xu Y, Zhang C, Shang D, Zhou L, Zou C, Sun Z, Li J, Zhang Y, Yang H, et al. Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. Nucleic Acids Res. 2013; 41: e101.

113. Feng L, Xu Y, Zhang Y, Sun Z, Han J, Zhang C, Yang H, Shang D, Su F, Shi X, Li S, Li C, Li X. Subpathway-GMir: identifying miRNA-mediated metabolic subpathways by integrating condition-specific genes, microRNAs, and pathway topologies. Oncotarget. 2015; 6: 39151–64.

114. Vrahatis AG, Balomenos P, Tsakalidis AK, Bezerianos A. DEsubs: an R package for flexible identification of differentially expressed subpathways using RNA-seq experiments [Internet]. Bioinformatics. 2016. p. 3844–6. doi: 10.1093/bioinformatics/btw544.

115. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015; 19: A68–77.

116. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012 [Internet]. CA: A Cancer Journal for Clinicians. 2015. p. 87–108. doi: 10.3322/caac.21262.

117. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, Forman D, Bray F. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 2013; 49: 1374–403.

118. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L, Akslen LA, Ragaz J, Gown AM, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res. 2004; 10: 5367–74.

119. Hollestelle A, Nagel JHA, Smid M, Lam S, Elstrodt F, Wasielewski M, Ng SS, French PJ, Peeters JK, Rozendaal MJ, Riaz M, Koopman DG, Ten Hagen TLM, et al. Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. Breast Cancer Res Treat. 2010; 121: 53–64.

120. Minsky BD. Unique considerations in the patient with rectal cancer. Semin Oncol. 2011; 38: 542–51.

121. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487: 330–7.

122. Gong J, Kelekar G, Shen J, Shen J, Kaur S, Mita M. The expanding role of metformin in cancer: an update on antitumor mechanisms and clinical development. Target Oncol. 2016; 11: 447–67.

123. Hurvitz SA, Kalous O, Conklin D, Desai AJ, Dering J, Anderson L, O'Brien NA, Kolarova T, Finn RS, Linnartz R, Chen D, Slamon DJ. In vitro activity of the mTOR inhibitor everolimus, in a large panel of breast cancer cell lines and analysis for predictors of response [Internet]. Breast Cancer Research and Treatment. 2015. p. 669–80. doi: 10.1007/s10549-015-3282-x.

124. Lui A, New J, Ogony J, Thomas S, Lewis-Wambi J. Everolimus downregulates estrogen receptor and induces autophagy in aromatase inhibitor-resistant breast cancer cells. BMC Cancer. 2016; 16: 487.

125. Hornick NI, Huan J, Doron B, Goloviznina NA, Lapidus J, Chang BH, Kurre P. Serum Exosome MicroRNA as a Minimally-Invasive Early Biomarker of AML. Sci Rep. 2015; 5: 11295.

126. Hornick NI, Doron B, Abdelhamed S, Huan J, Harrington CA, Shen R, Cambronne XA, Chakkaramakkil Verghese S, Kurre P. AML suppresses hematopoiesis by releasing exosomes that contain microRNAs targeting c-MYB. Sci Signal. 2016; 9: ra88.

127. Kumar B, Garcia M, Weng L, Jung X, Murakami JL, Hu X, McDonald T, Lin A, Kumar AR, DiGiusto DL, Stein AS, Pullarkat VA, Hui SK, et al. Acute myeloid leukemia transforms the bone marrow niche into a leukemia-permissive microenvironment through exosome secretion. Leukemia. 2018; 32: 575–87.

128. Serebrennikova OB, Paraskevopoulou MD, Aguado-Fraile E, Taraslia V, Ren W, Thapa G, Roper J, Du K, Croce CM, Tsichlis PN. The combination of knockdown and TNFα causes synthetic lethality via caspase-8 activation in human carcinoma cell lines. Proc Natl Acad Sci U S A. 2019; 116: 14039–48.

129. Ben-Sahra I, Manning BD. mTORC1 signaling and the metabolic control of cell growth. Curr Opin Cell Biol. 2017; 45: 72–82.

130. Eisenberg-Lerner A, Bialik S, Simon H-U, Kimchi A. Life and death partners: apoptosis, autophagy and the cross-talk between them. Cell Death Differ. 2009; 16: 966–75.

131. Michels AA. MAF1: a new target of mTORC1. Biochem Soc Trans. 2011; 39: 487–91.

132. Lapid K, Itkin T, D'Uva G, Ovadya Y, Ludin A, Caglio G, Kalinkovich A, Golan K, Porat Z, Zollo M, Lapidot T. GSK3β regulates physiological migration of stem/progenitor cells via cytoskeletal rearrangement. J Clin Invest. 2013; 123: 1705–17.

133. Huan J, Hornick NI, Shurtleff MJ, Skinner AM, Goloviznina NA, Roberts CT Jr, Kurre P. RNA trafficking by acute myelogenous leukemia exosomes. Cancer Res. 2013; 73: 918–29.

134. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, Chan W-L, Tsai W-T, Chen G-Z, Lee C-J, Chiu C-M, Chien C-H, Wu M-C, Huang C-Y, et al. miRTarBase: a database curates experimentally validated microRNA–target interactions [Internet]. Nucleic Acids Research. 2011. p. D163–9. doi: 10.1093/nar/gkq1107.

135. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009; 37: D105–10.

136. Wang J, Lu M, Qiu C, Cui Q. TransmiR: a transcription factor–microRNA regulation database [Internet]. Nucleic Acids Research. 2010. p. D119–22. doi: 10.1093/nar/gkp803.

137. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011; 39: D152–7.

138. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; 42: D68–73.

139. Cormen TH, Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction To Algorithms. MIT Press; 2001. 1180 p.

140. Benjamini Y, Hochberg Y. On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics [Internet]. Journal of Educational and Behavioral Statistics. 2000. p. 60. doi: 10.2307/1165312.

141. Cormen TH. Introduction to Algorithms. MIT Press; 2009. 1292 p.

142. Poole W, Gibbs DL, Shmulevich I, Bernard B, Knijnenburg T. Combining Dependent P-values with an Empirical Adaptation of Brown's Method [Internet]. 2016. doi: 10.1101/029637.

143. Brown MB. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance [Internet]. Biometrics. 1975. p. 987. doi: 10.2307/2529826.

144. Fisher RA. Statistical Methods for Research Workers [Internet]. Springer Series in Statistics. 1992. p. 66–70. doi: 10.1007/978-1-4612-4380-9_6.

145. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics. 1979; 6: 65–70.

146. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing [Internet]. Journal of the Royal Statistical Society: Series B (Methodological). 1995. p. 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

147. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2013.

148. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 2016; 32: 309–11.

149. Hu X, Stern HM, Ge L, O'Brien C, Haydu L, Honchell CD, Haverty PM, Peters BA, Wu TD, Amler LC, Chant J, Stokoe D, Lackner MR, et al. Genetic alterations and oncogenic pathways associated with breast cancer subtypes. Mol Cancer Res. 2009; 7: 511–22.

150. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database . 2015; 2015: bav028.

151. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017; 45: D833–9.

152. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004; 36: 431–2.

153. Chen X, Xu J, Huang B, Li J, Wu X, Ma L, Jia X, Bian X, Tan F, Liu L, Chen S, Li X. A sub-pathway-based approach for identifying drug response principal network. Bioinformatics. 2011; 27: 649–54.

154. Osborne CK, Schiff R. Mechanisms of endocrine resistance in breast cancer. Annu Rev Med. 2011; 62: 233–47.

155. LaPensee EW, Ben-Jonathan N. Novel roles of prolactin and estrogens in breast cancer: resistance to chemotherapy. Endocr Relat Cancer. 2010; 17: R91–107.

156. Normanno N, Campiglio M, Maiello MR, De Luca A, Mancino M, Gallo M, D'Alessio A, Menard S. Breast cancer cells with acquired resistance to the EGFR tyrosine kinase inhibitor gefitinib show persistent activation of MAPK signaling. Breast Cancer Res Treat. 2008; 112: 25–33.

157. Fang WB, Jokar I, Zou A, Lambert D, Dendukuri P, Cheng N. CCL2/CCR2 chemokine signaling coordinates survival and motility of breast cancer cells through Smad3 protein- and p42/44 mitogen-activated protein kinase (MAPK)-dependent mechanisms. J Biol Chem. 2012; 287: 36593–608.

158. Chen Y, Rodrik V, Foster DA. Alternative phospholipase D/mTOR survival signal in human breast cancer cells [Internet]. Oncogene. 2005. p. 672–9. doi: 10.1038/sj.onc.1208099.

159. Krishnan A, Nair S, Pillai M. Biology of PPARγ in Cancer: A Critical Review on Existing Lacunae [Internet]. Current Molecular Medicine. 2007. p. 532–40. doi: 10.2174/156652407781695765.

160. Chen L-C, Tu S-H, Huang C-S, Chen C-S, Ho C-T, Lin H-W, Lee C-H, Chang H-W, Chang C-H, Wu C-H, Lee W-S, Ho Y-S. Human breast cancer cell metastasis is attenuated by lysyl oxidase

inhibitors through down-regulation of focal adhesion kinase and the paxillin-signaling pathway. Breast Cancer Res Treat. 2012; 134: 989–1004.

161. Emery LA, Tripathi A, King C, Kavanah M, Mendez J, Stone MD, de las Morenas A, Sebastiani P, Rosenberg CL. Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. Am J Pathol. 2009; 175: 1292–302.

162. Lal I, Dittus K, Holmes CE. Platelets, coagulation and fibrinolysis in breast cancer progression. Breast Cancer Res. 2013; 15: 207.

163. Johnson SM, Gulhati P, Rampy BA, Han Y, Rychahou PG, Doan HQ, Weiss HL, Evers BM. Novel expression patterns of PI3K/Akt/mTOR signaling pathway components in colorectal cancer. J Am Coll Surg. 2010; 210: 767–76, 776–8.

164. Benvenuti S, Sartore-Bianchi A, Di Nicolantonio F, Zanon C, Moroni M, Veronese S, Siena S, Bardelli A. Oncogenic activation of the RAS/RAF signaling pathway impairs the response of metastatic colorectal cancers to anti-epidermal growth factor receptor antibody therapies. Cancer Res. 2007; 67: 2643–8.

165. Saito M, Iwadate M, Higashimoto M, Ono K, Takebayashi Y, Takenoshita S. Expression of phospholipase D2 in human colorectal carcinoma. Oncol Rep. 2007; 18: 1329–34.

166. Kang DW, Min DS. Positive feedback regulation between phospholipase D and Wnt signaling promotes Wnt-driven anchorage-independent growth of colorectal cancer cells. PLoS One. 2010; 5: e12109.

167. Zhong H, De Marzo AM, Laughner E, Lim M, Hilton DA, Zagzag D, Buechler P, Isaacs WB, Semenza GL, Simons JW. Overexpression of hypoxia-inducible factor 1alpha in common human cancers and their metastases. Cancer Res. 1999; 59: 5830–5.

168. Kaidi A, Qualtrough D, Williams AC, Paraskeva C. Direct transcriptional up-regulation of cyclooxygenase-2 by hypoxia-inducible factor (HIF)-1 promotes colorectal tumor cell survival and enhances HIF-1 transcriptional activity during hypoxia. Cancer Res. 2006; 66: 6683–91.

169. Zeestraten ECM, Benard A, Reimers MS, Schouten PC, Liefers GJ, van de Velde CJH, Kuppen PJK. The prognostic value of the apoptosis pathway in colorectal cancer: a review of the literature on biomarkers identified by immunohistochemistry. Biomark Cancer. 2013; 5: 13–29.

170. Kumarakulasingham M, Rooney PH, Dundas SR, Telfer C, Melvin WT, Curran S, Murray GI. Cytochrome p450 profile of colorectal cancer: identification of markers of prognosis. Clin Cancer Res. 2005; 11: 3758–65.

171. Tamási V, Monostory K, Prough RA, Falus A. Role of xenobiotic metabolism in cancer: involvement of transcriptional and miRNA regulation of P450s. Cell Mol Life Sci. 2011; 68: 1131–46.

172. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456: 470–6.

173. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014; 59: 5–15.

174. Vrahatis AG, Dimitrakopoulou K, Balomenos P, Tsakalidis AK, Bezerianos A. CHRONOS: a time-varying method for microRNA-mediated subpathway enrichment analysis [Internet]. Bioinformatics. 2016. p. 884–92. doi: 10.1093/bioinformatics/btv673.