

UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO SCIENZE DELLA VITA E DELL'AMBIENTE

Corso di Laurea Magistrale in Biologia Molecolare e Applicata

Oltre l'analisi trascrittomica: identificazione e prioritizzazione funzionale di varianti somatiche da dati single-nucleus RNA-seq nel carcinoma duttale pancreatico

Beyond transcriptomic analysis: identification and functional prioritization of somatic variants from single-nucleus RNA sequencing data in pancreatic ductal adenocarcinoma

Tesi di Laurea Magistrale di:
Eleonora Carletti

Relatore Chiar.mo
Prof. Matteo Giulietti

Correlatore
Prof. Francesco Piva

SESSIONE AUTUNNALE
Anno Accademico 2025/2026

INDICE

1. INTRODUZIONE 5
1.1 L'adenocarcinoma duttale pancreatico 5
1.1.1 Epidemiologia e caratteristiche cliniche 6
1.1.2 Sviluppo e progressione del PDAC 8
1.1.3 Microambiente tumorale 11
1.1.4 Predisposizione genetica e forme sporadiche 13
1.1.5 Alterazioni genetiche e genomiche 15
1.1.6 Varianti a singolo nucleotide ed eterogeneità intratumorale 17
1.1.7 Eterogeneità trascrizionale 19
1.1.8 Dal profilo molecolare all'analisi a singola cellula 21
1.2 Single-cell RNA sequencing: metodologie e caratteristiche dei dati 10x Genomics 23
1.2.1 Single-cell RNA sequencing e single-nucleus RNA sequencing 23
1.2.2 Smart-seq: un approccio full-length alla trascrittomica single-cell 24
1.2.3 Tecnologia 10x Genomics Chromium 25
1.2.4 Vantaggi e limiti dei dati 10x per l'identificazione delle varianti 28
2. OBIETTIVO 30
3. MATERIALI E METODI 32
3.1 Selezione del dataset e caratteristiche dei campioni 32
3.2 Recupero e preparazione dei dati grezzi 33
3.3 Allineamento e processamento mediante STARsolo 35
3.3.1 Configurazione di STARsolo e riferimenti genomici 35
3.3.2 Configurazione della chimica 10x Genomics v2 e v3 36
3.3.3 Generazione e organizzazione degli output 37
3.4 Estrazione e organizzazione delle annotazioni cellulari 38
3.5 Identificazione delle varianti somatiche candidate mediante SComatic 39
3.5.1 Preparazione dei dati e suddivisione per tipo cellulare 40
3.5.2 Conteggio delle basi e identificazione delle varianti candidate 41
3.5.3 Filtraggio e selezione delle varianti 43
3.6 Selezione delle varianti associate alla componente epiteliale maligna 46
3.7 Annotazione genomica mediante UCSC Genome Browser 46
3.8 Predizione degli effetti funzionali mediante AlphaGenome 48
3.8.1 Configurazione dell'analisi 49
3.8.2 Predizioni multimodali e scoring delle varianti 49

3.8.3 Analisi e organizzazione degli output 51
3.9 Annotazioni funzionali mediante Ensembl Variant Effect Predictor 52
3.9.1 Frequenze alleliche e ulteriori annotazioni in VEP 56
3.10 Prioritizzazione funzionale delle varianti somatiche 57
4. RISULTATI 60
4.1 Identificazione delle varianti somatiche candidate 60
4.2 Distribuzione cromosomica 62
4.3 Distribuzione rispetto alle regioni geniche 63
4.4 Distribuzione delle varianti tra i geni 65
4.5 Identificazione delle varianti ricorrenti 65
4.6 Conseguenze molecolari predette mediante VEP 66
4.7 Score globali di impatto: AVI e CADD 68
4.8 Predizione dell'impatto delle varianti missenso 69
4.9 Analisi mediante popEVE 70
4.10 Analisi dello splicing 72
4.11 Predizioni regolatorie mediante AlphaGenome 74
4.12 Annotazioni in database di varianti note 75
4.13 Prioritizzazione integrata delle varianti somatiche candidate 76
5. DISCUSSIONE 81
5.1 Significato complessivo dei risultati 81
5.2 Prioritizzazione delle varianti in funzione del possibile meccanismo 83
5.3 Differenti profili funzionali tra le varianti prioritarie 84
5.4 Processi biologici rappresentati tra le candidate 86
5.5 Utilità e limiti delle predizioni AlphaGenome 86
5.6 Ricorrenza delle varianti tra pazienti 87
5.7 Limiti dello studio 88
5.8 Prospettive future 89
6. CONCLUSIONI 92
7. APPENDICI 93
Appendice A: Pipeline utilizzata per l'identificazione delle varianti somatiche candidate mediante SComatic. 93
Appendice B: Script Python sviluppato per l'analisi delle varianti mediante AlphaGenome. 94
Riferimenti bibliografici 100

1. INTRODUZIONE

1.1 L'adenocarcinoma duttale pancreatico

L'adenocarcinoma duttale pancreatico (PDAC) rappresenta la principale forma istologica di carcinoma del pancreas, costituendo oltre il 90% di tutte le neoplasie maligne di questo organo, ed è una delle neoplasie caratterizzate dalla maggiore aggressività biologica e dalla prognosi più sfavorevole. La sua elevata mortalità è il risultato della combinazione di diversi fattori, tra cui la difficoltà di una diagnosi precoce, perciò la malattia viene diagnosticata frequentemente già in fase localmente avanzata o metastatica, l'elevata capacità invasiva delle cellule neoplastiche e la limitata efficacia delle strategie terapeutiche disponibili. A questi aspetti si aggiunge una notevole complessità biologica, determinata dall'interazione fra alterazioni genetiche ed epigenetiche, programmi trascrizionali differenti, metabolismo tumorale e un microambiente caratterizzato da una componente stromale e immunitaria particolarmente sviluppata [1]. La sopravvivenza globale a cinque anni, tra il 5%-8% per decenni, è rimasta per lungo tempo tra le più basse in assoluto tra tumori solidi, al punto che il PDAC è stato proiettato come una delle principali cause di morte oncologica nei paesi industrializzati nei prossimi anni, in ragione dell'invecchiamento della popolazione e della crescente prevalenza di fattori di rischio metabolici [2]. Oggi la sopravvivenza a 5 anni è aumentata, ma purtroppo solo al 12%.

Dal punto di vista anatomico il pancreas è una ghiandola con funzioni sia esocrine sia endocrine, dove la componente esocrina costituisce la parte più grande del parenchima pancreatico ed è formata principalmente da cellule acinari e duttali. Le cellule acinari sono specializzate nella sintesi e nella secrezione degli enzimi digestivi, che vengono rilasciati negli acini e convogliati attraverso una rete duttale gerarchicamente organizzata fino al duodeno, mentre le cellule duttali contribuiscono alla produzione e al trasporto del fluido pancreatico, secreto da un epitelio cubico o colonnare semplice, ricco di bicarbonato. La componente endocrina è invece costituita principalmente dalle isole pancreatiche, responsabili della produzione di ormoni come l'insulina e il glucagone. Il PDAC appartiene alle neoplasie del compartimento esocrino e presenta caratteristiche morfologiche e molecolari di differenziazione duttale, quali l'espressione di citocheratine, di mucine (in particolare MUC1, MUC4, MUC5AC) e di proteine tipiche dell'epitelio duttale; tuttavia, l'origine cellulare della neoplasia è biologicamente più

complessa di quanto ci possa suggerire la sua denominazione. Infatti, studi sperimentali hanno dimostrato che differenti popolazioni cellulari del pancreas possono contribuire alla carcinogenesi in presenza di specifiche alterazioni genetiche e condizioni del microambiente. In particolare, le cellule acinari possiedono una significativa plasticità fenotipica e possono andare incontro a un processo denominato metaplasia acino-duttale (acinar-to-ductal metaplasia, ADM), attraverso il quale acquisiscono delle caratteristiche che sono simili a quelle delle cellule duttali. Questo processo può essere indotto da condizioni di danno o infiammazione e, in presenza di alterazioni oncogenetiche, può rappresentare un passaggio rilevante nella formazione di lesioni precursore del PDAC [1,2]. I modelli murini geneticamente ingegnerizzati, nei quali il gene KRAS mutato oncogenico condizionale viene attivato specificamente nei diversi compartimenti cellulari del pancreas, hanno dimostrato che le cellule acinari, le cellule centroacinari e le cellule duttali possono tutte dare origine, sebbene con efficienze e tempi differenti, a lesioni neoplastiche duttali, e che l'infiammazione cronica e le alterazioni della segnalazione TGF-β/SMAD4 influenzano fortemente la competenza dei diversi compartimenti a subire trasformazione [2,4,6]. Questo modello evidenzia come la carcinogenesi pancreatica non debba essere interpretata semplicemente come la trasformazione di una specifica popolazione cellulare, ma come il risultato dell'interazione tra identità cellulare, plasticità epiteliale, alterazioni genetiche e segnali provenienti dall'ambiente circostante [1].

1.1.1 Epidemiologia e caratteristiche cliniche

Il PDAC rappresenta una patologia di grande rilevanza clinica a causa dell'elevata mortalità associata alla malattia. Secondo i dati del Global Cancer Observatory relativi al 2024, a livello mondiale sono stati stimati 531.318 nuovi casi di tumore pancreatico e oltre 490.786 decessi. Il tumore pancreatico si colloca pertanto tra le principali cause di morte oncologica a livello globale, con un rapporto particolarmente elevato tra mortalità e incidenza [3]. Uno dei principali fattori che contribuiscono alla prognosi sfavorevole è la difficoltà di diagnosticare il tumore in una fase sufficientemente precoce. L'incidenza aumenta con l'età, con un picco nella settima e ottava decade di vita, ed è leggermente superiore nel sesso maschile; i fattori di rischio modificabili e non modificabili comprendono il tabagismo, l'obesità, il diabete di lunga durata, la pancreatite cronica,

alcune condizioni ereditarie e, in parte ancora da chiarire, alterazioni del microbiota intestinale e pancreatico [1,2]. La presentazione clinica del PDAC è quasi sempre tardiva e poco specifica. I sintomi iniziali comprendono dolore addominale o dorsale vago, perdita di peso non intenzionale, astenia, nausea e, in una proporzione rilevante di pazienti, ittero ostruttivo quando la lesione interessa la testa del pancreas e comprime il coledoco. Queste manifestazioni, tuttavia, compaiono generalmente quando il tumore ha già raggiunto dimensioni significative o si è esteso alle strutture circostanti, e la loro natura aspecifica ritarda la richiesta di accertamenti diagnostici. In molti casi il PDAC si manifesta con eventi acuti come pancreatite acuta di nuova insorgenza, trombosi venosa profonda o una rapida instabilizzazione metabolica, in particolare un peggioramento di un diabete preesistente o un'insorgenza di un diabete di tipo 2 in età adulta senza fattori predisponenti; quest'ultimo evento è riconosciuto come possibile manifestazione paraneoplastica precoce, ma raramente viene interpretato come segnale di allarme oncologico [1,2]. Alla diagnosi, solo una minoranza di pazienti presenta una malattia localizzata e potenzialmente resecabile: la maggioranza ha malattia localmente avanzata o metastatica. Questa prognosi sfavorevole deriva da tre elementi convergenti: la progressione locale rapida, con infiltrazione precoce delle strutture vascolari peripancreatiche, in particolare l'arteria mesenterica superiore, il tronco celiaco e la vena porta, che può compromettere la possibilità di un trattamento chirurgico radicale; la diffusione metastatica precoce, prevalentemente epatica e peritoneale; e la resistenza ai trattamenti, sostenuta da meccanismi farmacologici intrinseci alla biologia del tumore e dal ruolo barriera della matrice extracellulare [1,2,18]. Il ruolo del microambiente nel determinare questo quadro clinico è stato progressivamente chiarito dalla ricerca di base e dalla patologia clinica. Il PDAC presenta una reazione desmoplastica estremamente intensa, con un accumulo di matrice extracellulare densa e di cellule stromali attivate che circonda le cellule neoplastiche e le isola dal compartimento vascolare, riducendo la perfusione intratumorale e ostacolando la distribuzione dei farmaci chemioterapici [1,18]. A questo si aggiunge l'azione immunosoppressiva del microambiente, che limita l'efficacia delle strategie di immunoterapia e crea una barriera ulteriore alla risposta antitumorale [1,18]. L'approccio terapeutico attuale, basato sulla combinazione di resezione chirurgica quando possibile e chemioterapia sistemica con schemi come FOLFIRINOX o gemcitabine/nab-paclitaxel nella malattia avanzata, ha portato a un

progresso modesto ma reale della sopravvivenza [17], lasciando tuttavia un ampio spazio alla necessità di approcci molecolari in grado di identificare sottogruppi di pazienti con vulnerabilità specifiche.

1.1.2 Sviluppo e progressione del PDAC

La carcinogenesi pancreatica è generalmente considerata un processo multistep, nel quale le cellule pancreatiche acquisiscono progressivamente alterazioni genetiche ed epigenetiche che modificano i meccanismi responsabili del controllo della proliferazione, della sopravvivenza, della differenziazione e dell'integrità genomica. Il carcinoma invasivo può svilupparsi attraverso differenti percorsi di carcinogenesi che comprendono la progressione di lesioni precursori non invasive. Nel 2001 un congresso di patologi internazionali ha proposto una nomenclatura unificata per le lesioni proliferative dell'epitelio dei piccoli dotti pancreatici, definendole neoplasie intraepiteliali pancreatiche (pancreatic intraepithelial neoplasia, PanIN) e classificandole in tre gradi di gravità crescente (PanIN-1, PanIN-2 e PanIN-3) sulla base di criteri morfologici e architettonici standardizzati. La standardizzazione di questi criteri ha rappresentato un prerequisito essenziale per la correlazione tra gradi lesionali e le alterazioni molecolari, in quanto prima di allora esistevano più di settanta termini diagnostici differenti utilizzati per descrivere le stesse lesioni, con scarsa concordanza tra osservatori [5]. Dunque, le PanIN sono lesioni microscopiche caratterizzate da alterazioni progressive dell'epitelio duttale e possono evolvere da forme di basso grado verso lesioni di alto grado fino alla comparsa del carcinoma invasivo. Le PanIN-1 sono lesioni a basso grado, caratterizzate da un epitelio colonnare con mucina abbondante e atipia minima o assente; le PanIN-2 presentano atipia moderata, con pseudostratificazione nucleare, ipercromatismo e mitosi occasionali; le PanIN-3 sono lesioni di alto grado, con atipia severa, perdita di polarità, gemmazione papillare intraluminale e necrosi, e sono considerate lesioni precorritrici immediate del carcinoma invasivo. La sequenza temporale dell'acquisizione delle alterazioni genetiche lungo questa progressione è stata ricostruita mediante l'analisi delle lesioni precursori e ha stabilito che l'attivazione di KRAS mutato è l'evento genetico più precoce, presente già nelle PanIN-1 e in un'elevata percentuale di lesioni a basso grado [1,2,4]. La perdita di CDKN2A/p16 costituisce tipicamente l'evento successivo, osservato con frequenza crescente nelle lesioni intermedie e di alto grado, mentre l'inattivazione di

TP53, SMAD4 e di altri geni soppressori tumorali compare prevalentemente nelle PanIN-3 e nel carcinoma invasivo [1,2,4]. L'ordine di questi eventi, pur non essendo rigidamente deterministico, definisce un modello nel quale l'acquisizione precoce di alterazioni oncogeniche, in particolare l'attivazione costitutiva della via RAS, innesca inizialmente la proliferazione, mentre la perdita progressiva di meccanismi di controllo della crescita, dell'apoptosi e della sorveglianza genomica ne favorisce la progressione verso la malignità [2] (Figura 1).

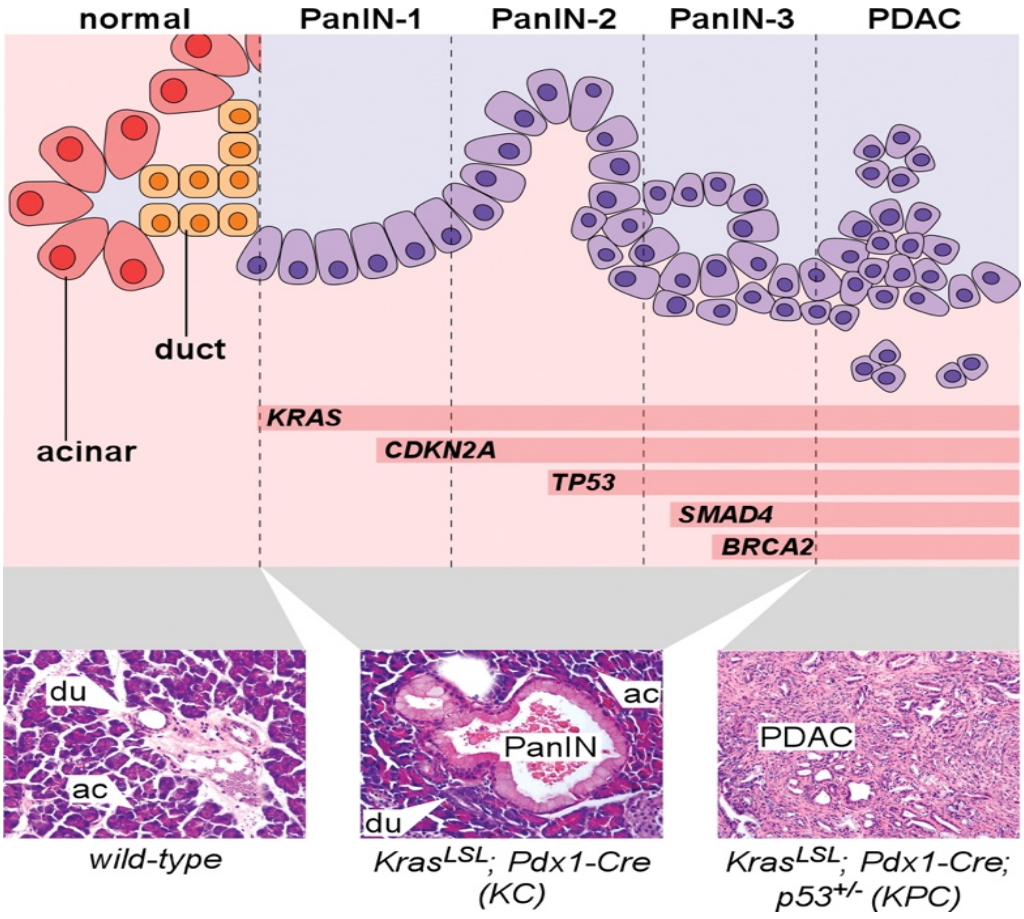


Figura 1. Progressione delle lesioni precursori verso l'adenocarcinoma duttale pancreatico. Rappresentazione schematica della progressione dal pancreas normale alle lesioni PanIN fino al PDAC, associata all'acquisizione progressiva di alterazioni genetiche, tra cui *KRAS, CDKN2A, TP53, SMAD4* e *BRCA2*. Nella parte inferiore sono riportati esempi istologici di pancreas normale, lesione PanIN e PDAC in modelli murini. Fonte: Murtaugh LC. *Pathogenesis of Pancreatic Cancer: Lessons from Animal Models*. Toxicologic Pathology. 2014;42(1):217–228. doi:10.1177/0192623313508250.

Un supporto sperimentale decisivo a questo modello è venuto dai modelli murini geneticamente ingegnerizzati (genetically engineered mouse models, GEMM). Nel

modello Pdx1-Cre;LSL-KrasG12D, un allele KRAS mutato condizionale viene attivato nel pancreas embrionale; questo modello riproduce alcuni aspetti della progressione osservata nella carcinogenesi pancreatica umana, in quanto l'attivazione di KRASG12D da sola genera PanIN a basso grado, la cui progressione verso un carcinoma metastatico richiede l'ulteriore inattivazione di un soppressore tumorale [4]. In particolare, la combinazione di KRASG12D con l'eliminazione di Ink4a/Arf determina l'insorgenza più precoce delle PanIN e una progressione rapida verso carcinoma duttale metastatico [7], mentre la combinazione con mutazioni attivanti di p53 genera carcinomi altamente metastatici con instabilità cromosomica, che riproducono la storia naturale della malattia umana [4]. Questi modelli hanno anche dimostrato che l'infiammazione cronica, come quella indotta da pancreatite sperimentale, accelera fortemente la progressione: in topi Kras-mutati, la pancreatite può favorire la progressione delle lesioni precancerose interferendo con i meccanismi che normalmente limitano la trasformazione neoplastica, tra cui la senescenza oncogena, che costituisce una barriera alla proliferazione delle cellule neoplastiche di basso grado, consentendo la loro evoluzione verso lesioni di alto grado [6]. Analogamente, l'eliminazione di Smad4 in un contesto di KRASG12D e infiammazione favorisce la progressione verso carcinoma metastatico, evidenziando un'interazione diretta tra segnali microambientali e destino neoplastico del clone epiteliale [2,4,7]. Oltre a queste neoplasie intraepiteliali pancreatiche, vi sono altri precursori del carcinoma pancreatico, tra cui le neoplasie papillari mucinose intraduttali (intraductal papillary neoplasms, IPMN) e le neoplasie cistiche mucinose (mucinous cystic neoplasms, MCN). Le IPMN sono neoplasie epiteliali mucinose che si sviluppano a livello del sistema duttale pancreatico e sono caratterizzate dalla proliferazione papillare dell'epitelio neoplastico e dalla produzione di quantità variabili di mucina. Esse sono classificate in sottotipi principali (gastrico, intestinale, pancreatico biliare, oncocitico) e possono progredire verso carcinoma invasivo con frequenza variabile in funzione del grado di displasia e della dilatazione duttale [1]. Le MCN sono invece neoplasie cistiche prevalentemente localizzate nel corpo e nella coda del pancreas, quasi esclusivamente in donne, caratterizzate dalla presenza di un epitelio mucinoso associato a uno stroma di tipo ovarico, elemento distintivo rispetto alle IPMN. Anche le MCN possono presentare diversi gradi di displasia e, nei casi più avanzati, evolvere verso una componente carcinomatosa invasiva. Dal punto di vista molecolare, KRAS e GNAS sono mutati con

elevata specificità negli IPMN, mentre RNF43 è frequentemente alterato nelle lesioni cistiche mucinose; l'acquisizione di mutazioni in TP53, SMAD4 e PIK3CA caratterizza invece la progressione verso l'alta displasia o il carcinoma invasivo, così come avviene nella via PanIN [1]. Questa variabilità nelle vie di sviluppo, insieme all'evoluzione clonale che caratterizza ciascun tumore, spiega perché tra pazienti differenti si osservino tumori con storie genomiche distinte e perché all'interno di un medesimo tumore possano coesistere subcloni geneticamente e fenotipicamente differenti. L'eterogeneità intratumorale, che verrà approfondita nelle sezioni successive, è quindi una diretta conseguenza della natura evolutiva e multistep della carcinogenesi pancreatica [1,2].

1.1.3 Microambiente tumorale

Il microambiente tumorale del PDAC costituisce uno degli aspetti più distintivi e clinicamente rilevanti di questa neoplasia (Figura 2). A differenza di altri tumori solidi, nei quali la componente neoplastica prevale nettamente su quella stromale, il PDAC è caratterizzato da un'abbondante reazione desmoplastica, cioè da un accumulo massiccio di matrice extracellulare e di cellule stromali che circonda e separa le cellule tumorali. La matrice extracellulare del PDAC è costituita da collagene di tipo I, fibronectina, laminina, acido ialuronico, in particolare dai fibroblasti associati al tumore e dalle cellule stellate pancreatiche attivate ed è organizzata da una rete tridimensionale che contribuisce a definire le proprietà meccaniche del tessuto, la perfusione vascolare e l'architettura del microambiente [1,18]. Le proprietà fisiche di questa matrice, inclusa la sua rigidità, la sua tensione e la sua porosità, non sono una conseguenza passiva della reazione stromale, ma esercitano un controllo attivo sul comportamento delle cellule neoplastiche, sulla segnalazione di meccanotrasduzione, sulla migrazione delle cellule e sulla capacità delle cellule immunitarie di raggiungere il tumore [1,18].

La componente cellulare del microambiente del PDAC è strettamente eterogenea. I fibroblasti associati al tumore (cancer-associated fibroblasts, CAF), che derivano in gran parte dall'attivazione delle cellule stellate pancreatiche, costituiscono la popolazione stromale più numerosa e presentano una notevole eterogeneità funzionale. Le analisi a livello di singola cellula ed imaging spaziale hanno identificato diverse sottopopolazioni di CAF, tra cui quelle con fenotipo miofibroblastico, caratterizzate dall'espressione di α-SMA e localizzazione iuxta-tumorale, quelle infiammatorie, che secernono citochine

proinfiammatorie come IL-6 e IL-11, quelle antigen-presenting, che esprimono MHC di classe II, e quelle mesenchimali, che risiedono nella periferia del tumore sono associate a segnalazione TGF-β e a un ruolo di rimodellamento della matrice extracellulare [1,18]. Queste sottopopolazioni svolgono funzioni contrastanti: alcune promuovono la crescita tumorale, l'invasione, la modulazione immunitaria e la resistenza ai farmaci, mentre altre esercitano un ruolo di contenimento della crescita neoplastica, come dimostrato dalla scomparsa della reazione desmoplastica nei modelli sperimentali di inattivazione delle cellule stellate, che paradossalmente si accompagna a tumori più aggressivi e meno differenziati. La funzione complessiva del compartimento stromale non può quindi essere ridotta a un'azione pro-tumorale o antitumorale semplice, ma deve essere interpretata come un sistema dinamico e contesto dipendente, in cui il bilancio netto dipende dal tipo cellulare, dallo stadio evolutivo e dalle condizioni microambientali locali [1,18].

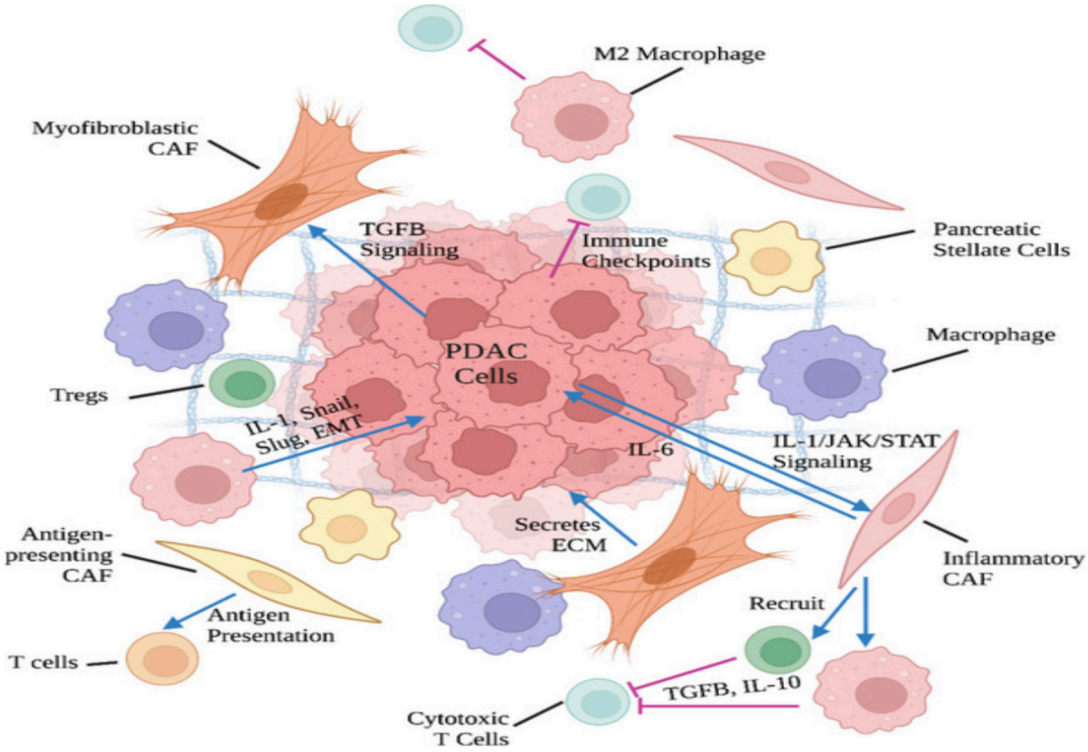


Figura 2. Rappresentazione schematica del microambiente tumorale dell'adenocarcinoma duttale pancreatico (PDAC). La figura mostra la complessa interazione tra le cellule neoplastiche e le principali componenti stromali e immunitarie del microambiente, tra cui differenti sottotipi di *cancer-associated fibroblasts* (CAF), macrofagi, linfociti T regolatori e citotossici, cellule B e altre popolazioni immunitarie coinvolte nella modulazione della risposta antitumorale. Fonte: Kung HC, Yu J. *Targeted therapy for pancreatic ductal adenocarcinoma: Mechanisms and clinical study*. MedComm. 2023;4(2):e216, Fig. 1. doi:10.1002/mco2.216.

Accanto ai CAF, il microambiente del PDAC contiene una componente immunitaria complessa, comprendente i linfociti T citotossici e T-helper, linfociti T regolatori, linfociti B, cellule NK, macrofagi associati al tumore, prevalentemente di fenotipo M2, cellule dendritiche e cellule mieloidi soppressorie derivate da precursori mieloidi. La composizione di questo infiltrato, la sua distribuzione spaziale e il suo stato funzionale determinano la capacità del sistema immunitario di riconoscere e distinguere le cellule neoplastiche, ma nel PDAC è sistematicamente orientato verso un'azione immunosoppressiva e protumorale. Le cellule endoteliali, infine, formano un letto vascolare anomalo, compresso dalla matrice extracellulare e caratterizzato da un flusso irregolare, che limita la perfusione e ostacola l'arrivo dei farmaci, creando un ulteriore elemento di eterogeneità spaziale nel microambiente. La fibrosi e la rigidità della matrice, oltre a favorire un ambiente ipossico e acido, agiscono come una barriera fisica alla distribuzione dei farmaci chemioterapici, spiegando in parte la resistenza relativa del PDAC anche agli agenti citotossici più potenti [1,18].

Questa complessità del microambiente ha implicazioni metodiche dirette. L'analisi trascrittomica di un campione tumorale nella sua totalità, chiamata “bulk”, in cui si produce una misura di espressione genica media di tutte le cellule presenti, mescola il segnale trascrizionale delle cellule neoplastiche con quello di un numero variabile e paziente specifico di cellule stromali, immunitarie ed endoteliali, rendendo difficile identificare le caratteristiche molecolari intrinseche del tumore, attribuire correttamente i programmi di espressione alle cellule che li esprimono e definire le interazioni tra i diversi compartimenti. Approcci computazionali di virtual microdissection hanno dimostrato che una parte rilevante della variabilità trascrizionale osservata nel PDAC è attribuibile alla composizione stromale del campione piuttosto che al tumore stesso, e che la separazione dei segnali tumorali e stromali permette di identificare sottotipi molecolari distinti con prognosi differenti [13]. Questo limite metodologico costituisce uno dei principali argomenti a favore dell'analisi a livello di singola cellula.

1.1.4 Predisposizione genetica e forme sporadiche

Le neoplasie pancreatiche si dividono, dal punto di vista dell’origine delle alterazioni genetiche, in forme con componente ereditaria e forme sporadiche. Le alterazioni

genetiche che conducono al PDAC possono essere presenti nella linea germinale, cioè trasmesse ereditariamente e presenti in tutte le cellule dell'organismo, oppure acquisite in modo somatico nelle cellule del tessuto pancreatico nel corso della vita. Le varianti germinali patogeneticamente rilevanti costituiscono la base delle forme familiari o ereditarie, mentre le alterazioni somatiche sono responsabili della quasi totalità dei casi sporadici. Questa distinzione ha rilevanza clinica, perché la presenza di una mutazione germinale in un gene di predisposizione oncologica implica la necessità di counseling genetico, di screening selettivi nei familiari e, in alcuni casi, di scelte terapeutiche specifiche [1,14]. Le varianti germinali più frequentemente associate al PDAC interessano i geni coinvolti nella riparazione del danno al DNA, in particolare nella riparazione delle rotture a doppio filamento mediante ricombinazione omologa: BRCA1, BRCA2, PALB2 e ATM [1,14]. Le mutazioni germinali in BRCA2 rappresentano la causa genetica più comune di PDAC ereditario e sono anche associate a un aumentato rischio di carcinoma mammario, ovarico e prostatico. Anche le mutazioni germinali in BRCA1, PALB2 e ATM sono associate a una predisposizione al PDAC, sebbene conferiscano rischi differenti. Altri geni di predisposizione comprendono MLH1, MSH2, MSH6 e PMS2, implicati nella riparazione dei mismatch e responsabili della sindrome di Lynch cui si associa un rischio elevato di carcinoma del colon retto e, in misura minora, di carcinoma del pancreas, e CDKN2A, la cui inattivazione germinale determina la sindrome dei nei atipici multipli con melanoma e predisposizione al PDAC [1,14]. Inoltre, la pancreatite ereditaria causata da mutazioni del gene PRSS1, codificante un tripsinogeno cationico, determina un'infiammazione cronica ricorrente del pancreas che costituisce un substrato pro-tumorigenico e aumenta notevolmente il rischio di PDAC nel corso della vita [1,2]. La prevalenza complessiva di varianti germinali patogenetiche in geni di predisposizione oncologica nei pazienti con PDAC, anche in assenza di storia familiare, è rilevante e giustifica la raccomandazione, in molti contesti clinici, di offrire il test germinale a tutti i pazienti con PDAC indipendentemente dalla storia familiare [14]. Le forme sporadiche, cioè quelle prive di una causa genetica ereditaria riconoscibile, costituiscono la maggior parte dei casi e sono caratterizzate dall'accumulo di alterazioni somatiche acquisite nelle cellule pancreatiche, in assenza di un background di suscettibilità ereditata. Tuttavia, la distinzione tra forme ereditarie e sporadiche non è assoluta. In primo luogo, le stesse vie molecolari colpite dalle alterazioni germinali nei

casi ereditari vengono frequentemente inattivate in modo somatico nelle forme sporadiche: ad esempio, alterazioni somatiche di BRCA2 o di altri geni coinvolti nella riparazione del DNA sono state osservate in tumori privi di mutazione germinale, e la loro presenza ha implicazioni terapeutiche in quanto rende le cellule tumorali sensibili agli agenti che causano danni al DNA, come i sali di platino e gli inibitori di PARP [1,15]. In secondo luogo, i fattori ambientali, tra cui il tabagismo, l'obesità, l'abuso di alcol e la dieta, interagiscono con la suscettibilità genetica, sia ereditaria che somatica, nel modulare il rischio di sviluppare il tumore e la velocità della sua progressione [1,2]. La carcinogenesi pancreatica deve quindi essere intesa come risultato dell'interazione tra predisposizione genetica, alterazioni somatiche acquisite e fattori ambientali, in cui nessun singolo elemento è sufficiente a determinare l'esito, ma la combinazione di essi determina la probabilità, la tempestività e la modalità di insorgenza della neoplasia [1,2,14].

1.1.5 Alterazioni genetiche e genomiche

Il PDAC è caratterizzato da un panorama genetico complesso nel quale coesistono alterazioni ricorrenti e alterazioni presenti solamente in una frazione dei tumori. I maggiori studi di sequenziamento genomico hanno identificato quattro principali geni frequentemente alterati: KRAS, TP53, CDKN2A e SMAD4 [1,2]. L'attivazione oncogenica di KRAS, quasi sempre mediante mutazione missense del codone 12, ma anche dei codoni 13 e 61, è presente in oltre il 90% dei tumori e costituisce l'evento genetico più precoce e più universale della carcinogenesi pancreatica [1,2,4]. La proteina KRAS mutata, bloccata in uno stato legato al GTP, attiva costitutivamente le vie di segnalazione a valle, in particolare RAS/MEK/ERK/PI3K/AKT/mTOR e RalGDS, promuovendo proliferazione, sopravvivenza, metabolismo e invasione. TP53 è inattivato prevalentemente mediante mutazioni missense con perdita del secondo allele, in circa il 75% dei PDAC; l'inattivazione di p53 compromette l'arresto del ciclo cellulare in risposta al danno al DNA, l'apoptosi e la senescenza, e in molti casi le proteine p53 mutanti acquisiscono funzioni gain-of-function che promuovono l'invasione, il metabolismo e la resistenza alla terapia [1,2,4]. CDKN2A, che codifica l'inibitore delle chinasi ciclina-dipendenti p16INK4A e p14ARF, è inattivato in circa il 95% dei PDAC mediante delezione, mutazione o metilazione del promotore: la perdita di p16 favorisce la

progressione del ciclo cellulare dalla fase G1 alla fase S, e la perdita di p14ARF compromette la via di stabilizzazione di p53 [1,2]. SMAD4, effettore centrale della segnalazione TGF-β, è inattivato mediante delezione o mutazione in circa il 55% dei PDAC, con conseguente perdita dei segnali antiproliferativi e di differenziazione mediati da TGF-β e, in modo contesto dipendente, una promozione dell'infiammazione e della progressione verso carcinoma avanzato [1,2]. Oltre a questi quattro geni, gli studi genomici sistematici hanno identificato alterazioni ricorrenti in un numero di vie cellulari e processi funzionali limitato. Lo studio pionieristico di Jones et al (2008), mediante sequenziamento di tutti i trascritti codificanti e analisi del numero di copie, ha evidenziato che ogni PDAC contiene in media circa 63 alterazioni genetiche, prevalentemente in mutazioni puntiformi che si aggregano in 12 vie di segnalazione e processi funzionali, ognuna alterata in una percentuale compresa tra il 67 e il 100% dei tumori [8]. Queste vie comprendono la segnalazione pro-proliferativa, l'adesione cellula-cellula, l'invasione e il rimodellamento della matrice extracellulare, l'apoptosi, il controllo del ciclo cellulare, la segnalazione TGF-β, la segnalazione WNT/Notch, il rimodellamento della cromatina, la riparazione del DNA, il processamento dell'RNA, il metabolismo e la segnalazione RAS [8]. Studi successivi, basati sul sequenziamento dell'intero genoma, hanno confermato e ampliato questo quadro, identificando alterazioni ricorrenti in geni come RNF43, ARID1A, TGFβR2, BRAF, GNAS e KDM6A, e dimostrando che i PDAC KRAS wild-type rappresentano una piccola minoranza dei casi, contengono alterazioni attivanti in altre componenti della via RAS o in geni oncogenici come GNAS, BRAF o CTNNB1 [9,10]. Al di là delle mutazioni puntiformi, il genoma del PDAC presenta un carico rilevante di alterazioni strutturali e di numero di copia. Il sequenziamento del genoma completo ha rilevato che il PDAC è caratterizzato da un alto tasso di riarrangiamenti cromosomici, tra cui inserzioni, delezioni, duplicazioni in tandem, inversioni, traslocazioni e amplificazioni focali, che interessano geni coinvolti nella segnalazione, nel controllo del ciclo cellulare, nel rimodellamento della cromatina e nel metabolismo [9]. In particolare, l'amplificazione di regioni cromosomiche contenenti oncogeni come MYC, AKT2, ERBB2, MDM2 e GATA6, e la delezione di regioni contenenti soppressori tumorali come CDKN2A, TP53 e SMAD4, sono frequentemente osservate e contribuiscono al fenotipo tumorale indipendentemente dalla presenza di mutazioni puntiformi [1,9,10]. Un elemento distintivo del PDAC è il fenomeno del riarrangiamento

catastrofico, con eventi di chromothripsis, cioè la formazione e il riarrangiamento di una o più regioni cromosomiche in un singolo evento che può generare amplificazioni oncogeniche o inattivare soppressori tumorali in un solo passo [9]. Inoltre, una parte rilevante dei PDAC presenta instabilità strutturale e genomica di tipo specifico, caratterizzata da specifici pattern di riarrangiamento cromosomico. Sulla base di queste caratteristiche, i tumori sono stati classificati in quattro classi genomiche con prognosi e caratteristiche cliniche differenti: stabile, localmente riarrangiata, riarrangiata e instabile per riparazione dei mismatch. Quest'ultima è caratterizzata da alterazione nei geni coinvolti nel sistema di riparazione dei mismatch (MMR) ed è associata a una potenziale maggiore sensibilità all'immunoterapia [9]. Dal punto di vista evolutivo, la cronologia di questi eventi genomici è stata ricostruita mediante l'analisi delle lesioni precursori e mediante la datazione molecolare dei cloni. L'attivazione di KRAS precede di gran lunga le altre alterazioni, mentre le alterazioni di TP53 e SMAD4, pur essendo frequenti, compaiono in un momento più tardivo, probabilmente in associazione con la transizione da lesioni pre-invasive a carcinoma invasivo [1,2,10]. L'analisi dei riarrangiamenti genomici ha inoltre proposto un modello nel quale eventi di riarrangiamento catastrofici precoci, presenti già nelle lesioni precursori, precedono di molti anni la diffusione metastatica, suggerendo una finestra temporale estremamente lunga tra iniziazione genetica e manifestazione clinica della malattia [9]. Questo modello evolutivo, insieme alla descrizione delle vie genetiche alterate, costituisce la base molecolare su cui si fondano gli approcci diagnostici e terapeutici attuali, tra cui l'identificazione di sottotipi molecolari, la selezione di terapie mirate e la valutazione della suscettibilità a trattamenti che sfruttano specifiche vulnerabilità genetiche [9,10,15].

1.1.6 Varianti a singolo nucleotide ed eterogeneità intratumorale

Tra le alterazioni genetiche di maggior interesse rientrano le varianti a singolo nucleotide (single nucleotide variants, SNV), nelle quali un singolo nucleotide della sequenza viene sostituito da un nucleotide differente. A seconda della regione genomica interessata, una variante può avere conseguenze biologiche molto diverse. Nelle regioni codificanti, una SNV può determinare una sostituzione amminoacidica, dando origine a una mutazione missense, oppure introdurre un codone di stop prematuro, generando una mutazione nonsense. In altri casi, la sostituzione di un nucleotide non modifica l'amminoacido

codificato e viene definita mutazione sinonima. Le mutazioni frameshift, invece, sono generalmente causate da inserzioni o delezioni di nucleotidi (indel) che alterano la cornice di lettura, quindi in un numero non multiplo di 3. Nelle mutazioni missense, come le mutazioni attivanti di KRAS o le mutazioni missense di TP53, l'amminoacido sostituito può alterare la funzione della proteina, generando un'attività oncogenica gain-of-function, una perdita di funzione, o l'acquisizione di nuove funzioni; nelle mutazioni nonsense o frameshift, invece, la conseguenza più frequente è la perdita di funzione della proteina, mediante generazione di una proteina tronca e instabile o mediante degradazione mediata dal nonsense-mediated decay dell'mRNA. Un SNV in qualsiasi punto del gene può alterare lo splicing del pre-mRNA, ad esempio, una SNV può alterare i siti donatore o accettore degli introni, generando trascritti aberranti, con eson skipping, intron retention o utilizzo di siti di splicing criptici, che possono generare proteine tronche o alterate. Infine, nelle regioni regolatorie a monte del gene, tra cui promotore prossimale e distale, con enhancer e silencer, contenenti centinaia di siti di legame per i fattori di trascrizione, una SNV può alterare il livello o la specificità dell'espressione di un gene, senza modificare la sequenza codificante, e questo tipo di alterazione è progressivamente riconosciuto come un contributo rilevante alla variabilità trascrizionale osservata nei tumori [1,9,10]. Nel PDAC le SNV più frequenti e meglio caratterizzate interessano i geni driver già descritti: KRAS, TP53, CDKN2A, SMAD4, insieme ad un gruppo più variabile di mutazioni di geni come ARID1A, KDM6A, RNF43, TGFβR2 e nei geni coinvolti nella riparazione del DNA [8,9,10]. Tuttavia, la frequenza delle SNV nel PDAC è in assoluto relativamente moderata rispetto ad altri tumori, come il melanoma o il carcinoma polmonare a cellule non-piccole, in cui l'esposizione a mutageni ambientali genera un numero estremamente elevato di mutazioni puntiformi. Il carico mutazionale del PDAC è invece dominato da alterazioni strutturali, delezioni, amplificazioni e riarrangiamenti cromosomici. Nonostante ciò, le SNV conservano una rilevanza fondamentale, sia perché interessano i geni driver che determinano il comportamento biologico del tumore, sia perché generano neoepitopi potenzialmente riconoscibili dal sistema immunitario, che costituiscono la base delle strategie di immunoterapia personalizzata [1]. Una considerazione centrale nell'interpretazione delle SNV è l'eterogeneità intratumorale, cioè la coesistenza di subcloni cellulari con assetti genetici differenti all'interno di un medesimo tumore. L'evoluzione clonale del PDAC genera una struttura gerarchica, in cui

una parte delle mutazioni è presente in tutte le cellule tumorali (mutazioni troncali o clonali, acquisite nelle prime fasi della carcinogenesi), mentre altre mutazioni sono presenti solo in sottoinsiemi di cellule (mutazioni ramificate o subclonali, acquisite più tardivamente). Questo fenomeno ha conseguenze dirette sulla lettura dei dati genomici prodotti dall'analisi di un intero campione (bulk). In un esperimento di sequenziamento bulk, la frequenza allelica osservata di una SNV è il risultato della proporzione di cellule che portano quella variante, del numero di copie della regione genomica interessata e dalla purezza del campione; una variante presente in una frazione minoritaria di cellule può essere sottorappresentata o sfuggire alla rilevazione, e la stima dell'assetto clonale o subclonale di una mutazione richiede inferenze computazionali soggette a incertezza [1,9,17]. L'analisi di metastasi multiple dello stesso paziente e la ricostruzione filogenetica dei cloni hanno permesso di stabilire che, nel PDAC, le mutazioni di KRAS e di TP53 sono generalmente clonali e condivise tra le lesioni metastatiche di uno stesso paziente, mentre le alterazioni subclonali sono più frequentemente osservate a livello di singola metastasi e meno condivise tra siti differenti [16]. Questo quadro evidenzia la necessità di approcci capaci di preservare l'informazione relativa all'eterogeneità cellulare del tumore, consentendo di distinguere alterazioni associate a specifiche popolazioni cellulari e di approfondire la complessità genetica intratumorale.

1.1.7 Eterogeneità trascrizionale

Oltre alla variabilità genetica, il PDAC presenta una marcata eterogeneità trascrizionale, cioè una variabilità nei programmi di espressione genica osservata a parità di assetto genetico. Studi trascrittomici condotti su coorti di pazienti hanno identificato differenti sottotipi molecolari del PDAC, caratterizzati da programmi di espressione distinti e da prognosi differenti. Collisson e collaboratori, mediante analisi di dati di espressione di campioni chirurgici primari e linee cellulari, hanno definito tre sottotipi: classic, quasi-mesenchymal e exocrine-like. Questi sottotipi sono caratterizzati, rispettivamente, dall'espressione di geni coinvolti nell'adesione cellulare e associati al fenotipo epiteliale, dall'espressione di geni mesenchimali e dall'espressione di geni caratteristici del compartimento acinare [12]. Studi successivi, attraverso analisi integranti dati di espressione e dati genomici, hanno raffinato questa classificazione, identificando quattro sottotipi: squamous, pancreatic progenitor, immunogenic e aberrantly differentiated

endocrine exocrine (ADEX), caratterizzati rispettivamente da un programma squamoso con attivazione di TP63 e p63ΔN, da espressione di geni associati allo sviluppo pancreatico precoce come FOXA2/3, PDX1 e MNX1, da un'infiltrazione immunitaria accentuata e da programmi di differenziazione endocrina ed esocrina [11]. Approcci di virtual microdissection computazionale hanno infine separato i segnali epiteliali da quelli stromali, identificando due sottotipi intrinseci del tumore, classical e basal-like, quest'ultimo associato a prognosi più sfavorevole e a una maggiore sensibilità alla chemioterapia [13]. La rilevanza biologica di questa eterogeneità trascrizionale risiede nel fatto che la relazione tra fenotipo e genotipo trascrizionale non è lineare. Due tumori con lo stesso assetto di mutazioni possono esprimere programmi di geni distinti, in funzione di fattori epigenetici, microambientali, metabolici e dell'origine cellulare, e questo rende l'assetto mutazionale, considerato singolarmente, insufficiente a spiegare il comportamento biologico del tumore e della risposta terapeutica. Al contrario, programmi trascrizionali differenti in tumori geneticamente simili sono stati associati a una diversa sensibilità a specifici trattamenti e a differenti esiti prognostici [11,12,13]. Studi a livello di singola cellula e di organoidi hanno inoltre dimostrato che gli stati trascrizionali delle cellule neoplastiche non sono fissi ma plastici, e che il microambiente esercita un controllo diretto sugli stati di espressione delle cellule tumorali, mediante segnali solubili, interazione cellula-cellula e proprietà fisiche della matrice extracellulare [20]. L'integrazione di più livelli di informazione, includendo mutazioni, numero di copie, metilazioni, espressione e stato del microambiente, è quindi necessaria per una descrizione accurata della biologia del PDAC, e nessun singolo livello di informazioni è sufficiente da solo per predire la risposta terapeutica o il comportamento clinico [1,10,11,12]. Da un punto di vista metodologico, l'eterogeneità trascrizionale evidenzia i limiti dell'approccio bulk. Una misura trascrittomica bulk fornisce la media dell'espressione di tutti i geni in tutte le cellule del campione, e quindi può identificare correttamente un sottotipo molecolare solo quando il tumore è omogeneo trascrizionalmente, ma non è in grado di rilevare la coesistenza di stati cellulari differenti nello stesso campione, né di attribuire programmi di espressione a specifiche popolazioni cellulari. Inoltre, la composizione cellulare variabile dei campioni, in termini di proporzione di cellule neoplastiche, stromali, immunitarie ed endoteliali, influenza la misura bulk e può determinare l'attribuzione errata di un sottotipo molecolare, come

dimostrato dal fatto che alcuni sottotipi definiti bulk in particolare l'exocrine-like, sono stati in seguito reinterpretati come un artefatto di contaminazione da tessuto acinare normale piuttosto che come un sottotipo intrinseco del tumore [11,13]. L'approccio single-cell, che misura l'espressione in migliaia di singole cellule, supera questi limiti e fornisce una risoluzione sufficiente per identificare le popolazioni cellulari, i loro stati e le loro interazioni.

1.1.8 Dal profilo molecolare all'analisi a singola cellula

I limiti dell'analisi bulk, evidenziati nella sezione precedente, derivano dal fatto che una massa tumorale non è una popolazione cellulare omogenea, ma una combinazione di cellule neoplastiche, fibroblasti, cellule immunitarie, cellule endoteliali e matrice extracellulare, in proporzioni variabili da paziente a paziente e in differenti regioni del medesimo tumore. Nel PDAC, la componente non neoplastica è particolarmente abbondante, e in molti campioni chirurgici le cellule tumorali costituiscono una minoranza delle cellule presenti, al punto che la bassa percentuale di cellule neoplastiche nei campioni è stata esplicitamente menzionata come una delle sfide tecniche degli studi genomici sistematici del PDAC [11]. In un esperimento di sequenziamento genomico o trascrittomico bulk, il segnale prodotto da una popolazione minoritaria di cellule neoplastiche viene diluito dal segnale di una popolazione maggioritaria di cellule stromali e immunitarie, rendendo difficile identificare le alterazioni molecolari intrinseche del tumore, attribuire programmi di espressione alle cellule che li esprimono e ricostruire la struttura clonale del tessuto [1,11,14]. L'analisi a livello di singola cellula, in particolare il sequenziamento dell'RNA a livello di singola cellula (single-cell RNA sequencing, scRNA-seq), ha rivoluzionato la possibilità di caratterizzare la composizione cellulare e l'eterogeneità trascrizionale del PDAC. Con gli approcci a base di droplet, come la tecnologia 10x Genomics, è possibile profilare simultaneamente migliaia di cellule individuali, ottenendo per ciascuna di esse un profilo trascrizionale che permette di identificare il tipo cellulare, lo stato funzionale e la posizione in percorsi di differenziamento o di attivazione. Nel PDAC, studi di scRNA-seq hanno mappato l'eterogeneità cellulare dei tumori, identificando le popolazioni epiteliali neoplastiche, le differenti sottopopolazioni di CAF, le cellule immunitarie e le cellule endoteliali, e ricostruendo le interazioni di segnalazione tra questi compartimenti [19,20]. Questi studi

hanno rilevato un livello di eterogeneità non accessibile con gli approcci bulk: al di là della presenza di differenti tipi cellulari, hanno identificato differenti stati trascrizionali all'interno delle cellule neoplastiche, alcuni dei quali associati a un fenotipo proliferativo, altri a un fenotipo invasivo o a quello di una cellula staminale, e hanno dimostrato che questi stati possono coesistere nello stesso tumore e che la loro distribuzione è modulata dal microambiente tumorale [19,20]. Inoltre, l'analisi single-cell ha permesso di identificare popolazioni cellulari rare, che nell'analisi bulk erano invisibili perché diluite, ma che possono essere associate alla resistenza ai trattamenti, alla ricorrenza della malattia o alla metastatizzazione [19,20]. Un aspetto rilevante della scRNA-seq è che le reads di RNA prodotte da un esperimento di single-cell (cioè le sequenze di RNA lette dal sequenziatore) possono contenere varianti nucleotidiche. Poiché l'mRNA è trascritto dal DNA genomico, le SNV presenti in regioni genomiche espresse si riflettono nei trascritti. Questo significa che un esperimento di scRNA-seq, progettato principalmente per la caratterizzazione trascrizionale, può contenere anche informazioni utili per l'identificazione di varianti nucleotidiche espresse, che possono essere analizzate mediante algoritmi di variant calling specificamente sviluppati per dati single-cell. Questo approccio presenta tuttavia importanti sfide metodologiche, legate alla copertura bassa e non uniforme dei dati, alla natura trascrittomica dell'informazione, che limita l'osservazione alle regioni espresse e adeguatamente coperte, e alla necessità di distinguere le varianti somatiche candidate da varianti germinali, eventi di RNA editing e artefatti tecnici. Ma c'è il vantaggio che nei dati single-cell e single-nucleus, la presenza dei barcode consente inoltre di preservare l'informazione relativa all'origine delle reads e di utilizzare le annotazioni cellulari per valutare la distribuzione del segnale mutazionale tra differenti popolazioni cellulari. L'informazione mutazionale può quindi essere integrata con quella relativa all'identità cellulare, consentendo di attribuire le varianti candidate a specifiche popolazioni cellulari e di approfondire l'eterogeneità molecolare del campione. Questo principio costituisce il fondamento metodologico e concettuale del capitolo successivo.

1.2 Single-cell RNA sequencing: metodologie e caratteristiche dei dati 10x Genomics

Il sequenziamento dell'RNA mediante approcci bulk RNA-seq permette di caratterizzare il profilo trascrizionale complessivo di un campione, ma il segnale ottenuto rappresenta la media delle molecole provenienti dalle diverse cellule presenti nel tessuto. In campioni complessi, come quelli tumorali, questa caratteristica può rendere difficile distinguere le diverse popolazioni cellulari e può mascherare varianti di espressione specifiche di determinati sottotipi cellulari. Lo sviluppo delle tecnologie single-cell RNA sequencing ha permesso di superare questa limitazione, consentendo di associare il profilo trascrizionale alla singola cellula e di caratterizzare direttamente l'eterogeneità cellulare presente all'interno di un campione [23]. L'applicazione della scRNA-seq è diventata progressivamente più accessibile grazie allo sviluppo di metodologie ad alto rendimento (high throughput) capaci di analizzare migliaia di cellule in un singolo esperimento. L'aumento del numero di cellule analizzabili ha reso possibile identificare anche popolazioni cellulari rare e caratterizzare con maggior dettaglio la composizione dei tessuti complessi e delle neoplasie eterogenee [23].

Nel presente lavoro di tesi sono stati analizzati esclusivamente dati di single-cell RNA, nello specifico single-nucleus sequencing, generati mediante la tecnologia 10x Genomics Chromium, che è la più adottata al mondo. Le altre metodologie di trascrittomica a singola cellula, descritte nei paragrafi successivi, inclusa la tecnologia Smart-seq, sono molto meno diffuse e la disponibilità di dataset su campioni PDAC sequenziati è perciò molto limitata.

1.2.1 Single-cell RNA sequencing e single-nucleus RNA sequencing

La single-cell RNA sequencing ha introdotto una discontinuità concettuale rispetto al bulk RNA-seq: mentre quest'ultimo restituisce una misura aggregata, in cui il contributo delle singole cellule è irrimediabilmente confuso, la scRNA-seq stabilisce una corrispondenza esplicita tra le molecole di RNA sequenziate e la cellula di origine, producendo un profilo di espressione per singola cellula [23]. Questa associazione molecole-cellula è resa possibile da una fase di compartimentalizzazione e barcoding, in cui ogni cellula (o nucleo) viene isolata fisicamente o altrimenti etichettata con un

identificatore molecolare prima dell'amplificazione e del sequenziamento, cosicché le reads possano essere ricondotte alla loro origine al momento dell'analisi bioinformatica. La possibilità di ottenere profili per singola cellula ha immediate conseguenze biologiche: diventa possibile identificare tipi cellulari rari, ricostruire traiettorie di differenziamento, rilevare stati transitori e caratterizzare l'eterogeneità intratumorale a livello della singola cellula, tutti aspetti che sfuggono alle misure bulk [23]. Accanto alla scRNA-seq su cellule intere, un'alternativa è la single-nucleus RNA sequencing (snRNA-seq), in cui l'oggetto dell'analisi non è la cellula intera ma il nucleo isolato [22]. Il nucleo contiene l'RNA nascente che è ancora in fase di processing (pre-mRNA) e mRNA maturo pronto per essere esportato nel citoplasma, ed è quindi in grado di fornire un profilo di espressione informativo, pur con caratteristiche quantitative leggermente diverse rispetto alla cellula intera. Il vantaggio principale dell'approccio nucleare è di ordine pratico: i nuclei sono più facili da isolare rispetto alle cellule intere, in particolare da tessuti complessi come il sistema nervoso centrale o tessuti fibrotici, e sono meno suscettibili ai problemi di dissociazione che affliggono i protocolli basati su cellule intere. Questo è particolarmente rilevante per i tumori solidi, come il pancreas, dove la reazione stromale fibrotica rende la dissociazione enzimatica delle cellule intere difficile e selettiva, con il rischio di introdurre bias di composizione cellulare. L'isolamento dei nuclei, spesso effettuabile anche su campioni congelati, riduce in larga parte questi problemi e ha reso la snRNA-seq una scelta metodologica di riferimento per l'analisi di tessuti difficili da dissociare, come nel caso del PDAC analizzato in questo lavoro. È inoltre importante sottolineare che la snRNA-seq, pur avendo una copertura dell'RNA per cellula più bassa rispetto alla scRNA-seq su cellule intere, mantiene la capacità di discriminare tra popolazioni cellulari e, nel caso di questo studio, di fornire il substrato molecolare per l'identificazione di varianti genetiche e la loro attribuzione a specifiche popolazioni cellulari.

1.2.2 Smart-seq: un approccio full-length alla trascrittomica single-cell

Tra le metodologie sviluppate per la caratterizzazione del trascrittoma di singole cellule, Smart-seq rappresenta uno degli approcci più utilizzati per ottenere una copertura estesa ai trascritti. Il protocollo si articola in tre fasi principali: retrotrascrizione dell’RNA in cDNA, amplificazione tramite template switching (un meccanismo che sfrutta la capacità

della trascrittasi inversa di aggiungere nucleotidi indipendenti dallo stampo all'estremità 3' del cDNA per innescare una seconda trascrizione) e preamplificazione tramite PCR [21]. Questo approccio ha il pregio di generare librerie full-length, in cui il cDNA copre l'intera lunghezza dei trascritti anziché limitarsi a un'estremità. La conseguenza pratica è che la copertura dei trascritti è estesa: le reads si distribuiscono lungo tutta la molecola, anziché concentrarsi in prossimità di un'estremità come accade nei protocolli di tipo 3' end counting (es. 10x Genomics). Questa caratteristica rende Smart-seq particolarmente adatto alla caratterizzazione dei trascritti a bassa abbondanza e all'analisi dello splicing alternativo, dove la copertura full-length è essenziale per identificare isoforme specifiche [21]. La sensibilità elevata del metodo, che rileva in media un numero maggiore di geni per cellula rispetto alle piattaforme droplet-based, è un vantaggio sostanziale per l'identificazione di varianti: maggiore è la copertura, più alta è la probabilità di osservare una posizione variante del trascritto. Tuttavia, Smart-seq presenta anche svantaggi significativi. Il primo è il throughput limitato: essendo un protocollo su piastra, il numero di cellule analizzabili per esperimento è nell'ordine delle centinaia, anziché delle decine di migliaia. Il secondo è l'assenza di unique molecular identifiers (UMI), che rende impossibile distinguere tra molecole originali e copie PCR, introducendo il rumore di amplificazione che complica la quantificazione assoluta dell'espressione [21]. Per quanto riguarda l'identificazione di varianti, la copertura full-length di Smart-seq aumenta la probabilità di osservare una posizione variante, ma resta il fatto che anche questo approccio è basato su RNA e condivide quindi con 10x i limiti intrinseci legati alla dipendenza dell'espressione del gene, ai livelli di dropout e alla distinzione tra varianti somatiche, germinali e di RNA editing.

1.2.3 Tecnologia 10x Genomics Chromium

La tecnologia 10x Genomics Chromium appartiene alla categoria degli approcci droplet-based per la trascrittomica single-cell (Figura 3). Il principio generale di questi approcci consiste nella compartimentalizzazione di singole cellule (o nuclei) in goccioline d'emulsione, ciascuna contenente anche una micro perla gelatinosa (Gel Bead) recante un barcode cellulare specifico [23]. All'interno di ogni gocciolina, l'RNA della cellula viene retrotrascritto e il cDNA risultante viene marcato con il barcode della gel bead, in modo che tutte le molecole provenienti dalla stessa cellula condividano lo stesso barcode.

Accanto al barcode cellulare, ogni molecola riceve un unique molecular identifier (UMI), un'etichetta di 10-12 nucleotidi che contraddistingue la molecola originale di mRNA, permettendo al momento dell'analisi di distinguere le molecole originali dalle copie generate dall'amplificazione PCR e di ottenere una quantificazione digitale dell'espressione [23]. Questo sistema di doppio barcoding (barcode cellulare + UMI) è una delle principali innovazioni della piattaforma, poiché consente di superare il bias di amplificazione che affligge i metodi senza UMI. L'elevato throughput della piattaforma è il vantaggio principale rispetto ai metodi su piastra: le goccioline possono essere generate in pochi minuti, e il numero di cellule analizzabili per esperimento è nell'ordine di migliaia o decine di migliaia, con un'efficienza di cattura che permette di profilare un numero elevato di cellule anche a partire da campioni di dimensioni limitate [23].

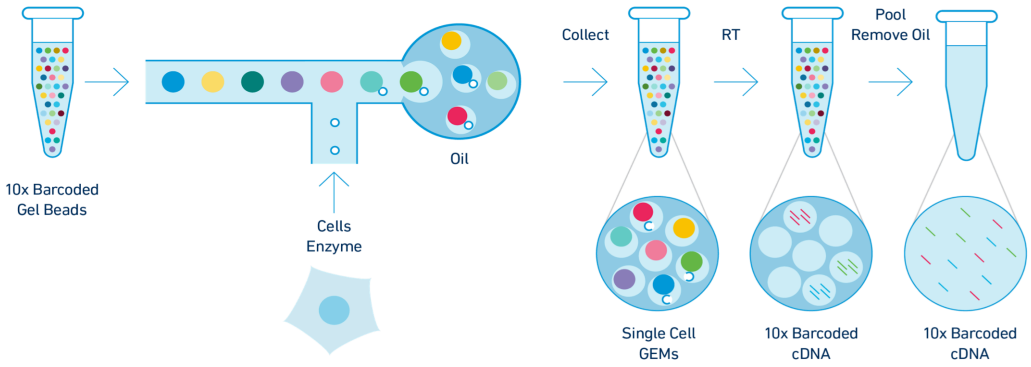


Figura 3. Schema del workflow di single-cell RNA sequencing (scRNA-seq) mediante tecnologia 10x Genomics. Le singole cellule vengono compartimentalizzate insieme a gel beads contenenti barcode specifici all'interno di GEM (Gel Beads-in-Emulsion), consentendo di associare i trascritti alla cellula di origine. Dopo la trascrizione inversa (RT), il cDNA barcodato viene recuperato e utilizzato per le successive fasi di preparazione della libreria e sequenziamento. Fonte: 10x Genomics, *Getting Started with Single Cell Gene Expression*, CG000360.

Nel presente lavoro di tesi sono stati analizzati dati ottenuti tramite il protocollo Chromium Single Cell 3' Gene Expression. Una caratteristica cruciale di questo protocollo è la cattura delle molecole di RNA poliadenilate mediante primer oligo-dT. Nei dati di snRNA-seq, tuttavia, il materiale di partenza è costituito dal contenuto nucleare e comprende, oltre a mRNA maturo non ancora esportato nel citoplasma, anche una quota rilevante di trascritti non completamente processati (pre-mRNA); di

conseguenza, una parte consistente delle reads può allinearsi anche a regioni introniche. Questo aspetto distingue parzialmente i profili ottenuti dai nuclei da quelli derivati da cellule intere e rende importante considerare sia il segnale esonico sia quello intronico durante la quantificazione dell'espressione.

Il protocollo 10x Genomics Chromium Single Cell 3' Gene Expression rimane comunque basato su una strategia di 3' gene expression, nella quale l'informazione di sequenza è maggiormente concentrata verso l'estremità 3' dei trascritti rispetto agli approcci full-length come Smart-seq (Figura 4).

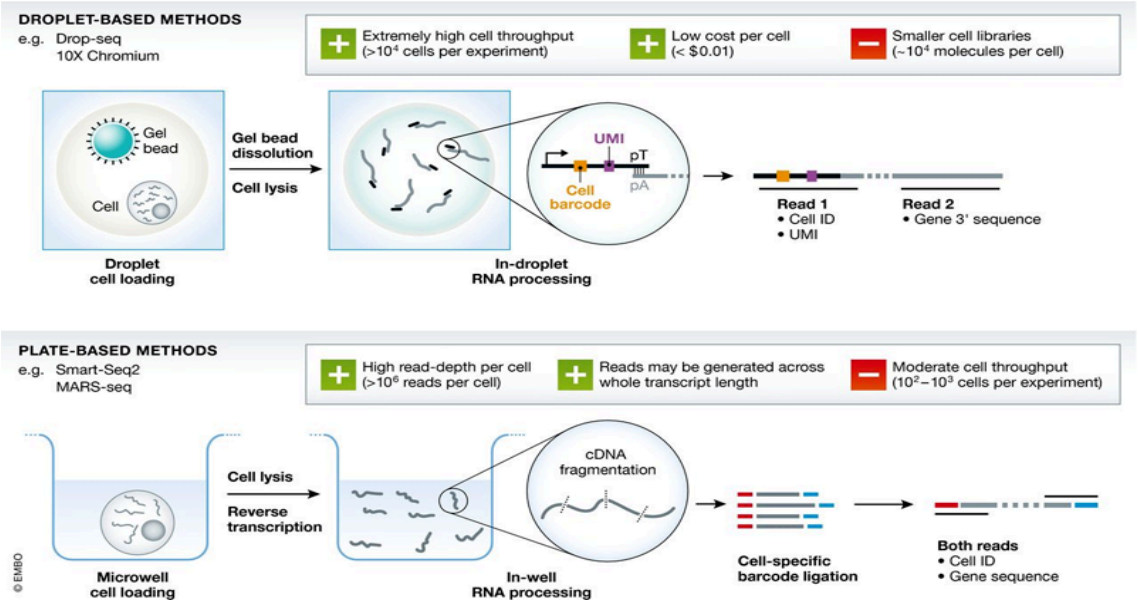


Figura 4. Confronto tra approcci droplet-based e plate-based per il sequenziamento dell'RNA a singola cellula. I metodi droplet-based, come 10x Genomics Chromium, consentono di analizzare un elevato numero di cellule mediante compartimentalizzazione in goccioline e utilizzo di barcode cellulari e UMI. Gli approcci plate-based, come Smart-seq2, presentano invece un throughput inferiore, ma permettono una maggiore profondità di sequenziamento per cellula e una copertura più estesa dei trascritti. Fonte: Griffiths et al., *Using single-cell genomics to understand developmental processes and cell fate decisions*, *Molecular Systems Biology* (2018).

Questa caratteristica determina una copertura non uniforme lungo il trascritto e costituisce un limite particolarmente rilevante quando i dati vengono riutilizzati per il variant calling. La probabilità di osservare una variante dipende infatti dalla posizione della variante rispetto alle regioni effettivamente sequenziate, dall'espressione del gene, dalla quantità di RNA disponibile nel nucleo e dalla profondità di sequenziamento. Una variante

genomica può pertanto essere presente nelle cellule analizzate ma non risultare osservabile nei dati trascrittomici qualora la posizione corrispondente non sia adeguatamente coperta [23,24].

1.2.4 Vantaggi e limiti dei dati 10x per l'identificazione delle varianti

L'utilizzo di dati 10x Genomics Chromium per l'identificazione di varianti genetiche presenta vantaggi e limiti. Un importante vantaggio è rappresentato dalla presenza dei barcode cellulari, che consentono di preservare l'informazione relativa all'origine cellulare delle reads. Tale informazione può essere sfruttata per valutare la distribuzione del segnale mutazionale tra differenti popolazioni cellulari e attribuire le varianti candidate a specifici tipi cellulari, un livello di informazione non direttamente disponibile nel normale sequenziamento bulk [24]. Questo aspetto è particolarmente rilevante in oncologia, dove l'eterogeneità intratumorale rende utile distinguere il segnale mutazionale associato alle differenti popolazioni cellulari presenti nel campione. Un ulteriore vantaggio è il throughput elevato, che aumenta il numero di osservazioni disponibili per ogni posizione variante. I limiti, tuttavia, sono sostanziali e derivano dalla natura stessa del dato trascrittomico. Il primo è la dipendenza dall'espressione: una variante può essere osservata solo se il gene che la contiene è espresso, se la popolazione interessata è effettivamente coperta e se la profondità di sequenziamento è sufficiente a consentirne la rilevazione [24]. Questo implica che il sequenziamento dell'RNA non fornisce un'immagine completa del genoma, ma solo dei trascritti attivi, con una copertura fortemente non uniforme. Il secondo limite è il dropout, il fenomeno per cui un trascritto effettivamente presente in una cellula non viene rilevato, particolarmente severo per i geni a bassa espressione sulla piattaforma 10x Genomics [24]. Il dropout implica che l'assenza di una variante in una cellula non corrisponde necessariamente alla sua assenza a livello genomico, ma può semplicemente riflettere il fatto che il trascritto non sia stato catturato o sequenziato. Questo è il limite cruciale nell'interpretazione dei dati: mentre in un esperimento di DNA sequencing l'assenza di una variante è un'informazione relativamente affidabile, in un esperimento di RNA sequencing non lo è. Infine, una sostituzione osservata nei dati di RNA può essere attribuibile a fattori diversi da una variante somatica: può essere una variante germinale già presente nel genoma dell'individuo, un evento di RNA editing (in particolare A-to-I, mediato dalla famiglia di

enzimi ADAR, che può essere confuso con una sostituzione A-to-G) o un semplice errore tecnico introdotto durante la preparazione delle librerie o il sequenziamento [24]. La distinzione tra queste alternative richiede filtri specifici e test statistici, come quelli implementati in tool di analisi come SComatic (Figura 5).

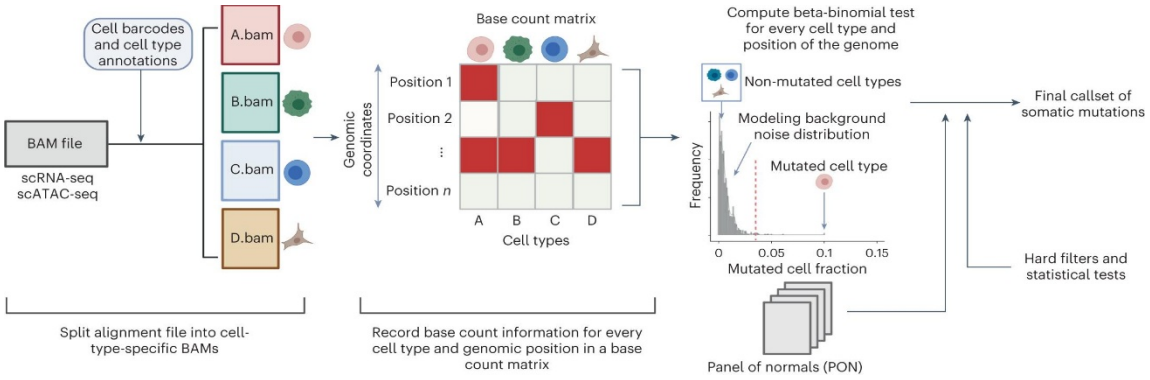


Figura 5. Workflow di SComatic per l'identificazione de novo di mutazioni somatiche da dati single-cell. Il file di allineamento BAM viene suddiviso in file specifici per tipo cellulare utilizzando i barcode e le annotazioni cellulari. I conteggi delle basi vengono quindi organizzati in una matrice per posizione genomica e tipo cellulare e analizzati mediante un modello beta-binomiale, insieme a filtri statistici e a un panel di campioni normali (Panel of Normals, PON), per distinguere il segnale mutazionale dal rumore di fondo e ottenere un insieme finale di mutazioni somatiche candidate. Fonte: adattata da Muyas et al., *De novo detection of somatic mutations in high-throughput single-cell profiling data sets*, *Nature Biotechnology* (2024).

2. OBIETTIVO

L'obiettivo generale del presente lavoro di tesi è valutare la fattibilità dell'identificazione de novo di varianti somatiche candidate a partire da dati di single-nucleus RNA sequencing (snRNA-seq), generati mediante tecnologia 10x Genomics Chromium, e della loro attribuzione a specifiche popolazioni cellulari in campioni di adenocarcinoma duttale pancreatico (PDAC). Come descritto nei capitoli precedenti, il PDAC è caratterizzato da una marcata eterogeneità genetica, trascrizionale e microambientale; pertanto, la possibilità di associare l'informazione mutazionale a specifiche popolazioni cellulari rappresenta un elemento di forte interesse per la comprensione della biologia di questa neoplasia. L'analisi di varianti a partire da dati trascrittomici costituisce, tuttavia, un approccio di difficile gestione, a causa della dipendenza della rilevazione dall'espressione genica, della copertura limitata e non uniforme dei dati single-cell e della necessità di distinguere le mutazioni somatiche da varianti germinali, eventi di RNA editing e artefatti tecnici. La presente tesi si colloca quindi in un ambito di studio di fattibilità metodologica, con l'obiettivo di verificare se un dataset single-cell originariamente generato per la caratterizzazione trascrittomica possa restituire un'informazione mutazionale biologicamente interpretabile.

Gli obiettivi specifici del lavoro possono essere articolati come segue:

1. Identificare varianti somatiche candidate a partire da dati snRNA-seq. Valutare la possibilità di rilevare de novo varianti nucleotidiche candidate nei dati 10x disponibili, mantenendo l'informazione relativa alle popolazioni cellulari nelle quali il segnale mutazionale viene osservato.
2. Integrare l'informazione mutazionale con quella trascrittomica. Valutare le varianti candidate nel contesto dell'espressione dei geni interessati e delle differenti popolazioni cellulari, al fine di interpretare correttamente la presenza o la mancata rilevazione del segnale mutazionale nei dati trascrittomici.
3. Caratterizzare e prioritizzare le varianti candidate. Contestualizzare le varianti identificate dal punto di vista genomico e funzionale, allo scopo di individuare gli eventi potenzialmente più rilevanti e distinguerli da quelli caratterizzati da una maggiore incertezza interpretativa.
4. Valutare criticamente la fattibilità dell'approccio. Definire potenzialità e limiti dell'utilizzo di dati snRNA-seq per il recupero di informazione mutazionale

associata a specifiche popolazioni cellulari, con particolare attenzione alla dipendenza dall'espressione genica e dalla copertura delle posizioni analizzate e alle conseguenze di tali fattori sull'interpretazione dei risultati.

Nel complesso, il lavoro mira quindi a stabilire se e in quale misura dati snRNA-seq già disponibili possano essere utilizzati non soltanto per descrivere l'eterogeneità trascrizionale e cellulare del PDAC, ma anche per ottenere un'informazione complementare relativa alle mutazioni somatiche osservabili nelle singole popolazioni cellulari.

3. MATERIALI E METODI

3.1 Selezione del dataset e caratteristiche dei campioni

La fase iniziale del presente lavoro di tesi è stata dedicata alla ricerca e alla selezione di un dataset di trascrittomica a singola cellula adeguato allo studio delle varianti somatiche nell'adenocarcinoma duttale pancreatico (PDAC). La scelta del dataset ha rappresentato un passaggio metodologico particolarmente rilevante, poiché l'identificazione di varianti a partire da dati trascrittomici presenta requisiti differenti rispetto a una tradizionale analisi di espressione genica. Oltre alla disponibilità di un numero adeguato di cellule o nuclei, risultavano infatti necessari l'accesso ai dati grezzi di sequenziamento, una quantità di reads sufficiente a garantire un'adeguata rappresentazione delle regioni trascritte e la presenza di informazioni che consentissero di associare i dati di sequenziamento alle singole cellule o nuclei e alle rispettive popolazioni cellulari.

È stata condotta una ricerca approfondita dei dataset di single-cell RNA sequencing disponibili per il PDAC, con particolare interesse iniziale per dati prodotti mediante protocolli di Smart-seq e Smart-seq2, perché permettono di sequenziare l'intero mRNA. La ricerca preliminare ha tuttavia evidenziato una disponibilità limitata di dataset che soddisfacessero contemporaneamente i requisiti necessari per il presente progetto. In alcuni studi, inoltre, l'accesso ai dati grezzi derivati da campioni umani era subordinato a procedure di accesso controllato.

Considerate tali limitazioni, la ricerca è stata successivamente estesa a dataset prodotti mediante tecnologie single-cell ad alto rendimento (10x Genomics). È stato selezionato il dataset pubblico GSE202051 [25], depositato nel Gene Expression Omnibus (GEO) del National Center for Biotechnology Information (NCBI). La serie GSE202051 contiene dati generati nell'ambito dello studio di Hwang et al., *"Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment" [25].* Lo studio è stato finalizzato alla caratterizzazione dell'eterogeneità cellulare e molecolare del PDAC e delle modificazioni associate al trattamento neoadiuvante attraverso l'impiego combinato di single-nucleus RNA sequencing (snRNA-seq) e tecniche di analisi spaziale.

La coorte complessiva descritta nei materiali supplementari dello studio comprende 48 pazienti affetti da PDAC, suddivisi in pazienti non sottoposti a trattamento neoadiuvante e pazienti trattati prima della resezione chirurgica. Non tutti i campioni inclusi nella

coorte disponevano tuttavia di dati snRNA-seq. Il dataset GSE202051 comprende 43 campioni di tumori primari di PDAC da 43 pazienti, tutti considerati nel presente lavoro. Di questi, 18 appartenevano al gruppo Untreated, identificato mediante le sigle PDAC_U_1 – PDAC_U_18, mentre 25 appartenevano al gruppo Treated, comprendente i campioni PDAC_T_1 – PDAC_T_25, da pazienti sottoposti a differenti regimi terapeutici neoadiuvanti, comprendenti differenti combinazioni di chemioterapia, radioterapia e ulteriori agenti terapeutici.

Nello studio originale, la caratterizzazione trascrittomica dei campioni selezionati è stata effettuata mediante single-nucleus RNA sequencing utilizzando la piattaforma 10x Genomics Chromium. Come riportato nei materiali supplementari dello studio, le librerie non sono state generate utilizzando una singola versione della chimica 10x Genomics. Per la maggior parte dei campioni è stata impiegata la chimica 10x Genomics v3, mentre 6 campioni sono stati processati mediante la versione v2 (U4, U13, U15, T5, T6 e T7).

In NCBI SRA (Sequence Read Archive), per ciascuno dei 43 campioni sono stati identificati i corrispondenti record di sequenziamento. I dati grezzi di sequenziamento associati a tali record sono identificati mediante specifici SRA Run accession number (SRR). La corrispondenza tra identificativo del paziente, record GEO e relativo SRR è stata verificata prima di procedere alla successiva elaborazione bioinformatica dei dati.

In aggiunta, il dataset GSE202051 contiene anche i dati snRNA-seq processati dagli Autori, contenenti le annotazioni dei tipi cellulari identificati. Nel presente lavoro, i dati grezzi e i dati processati sono stati utilizzati in maniera complementare: i primi sono stati sottoposti a nuova elaborazione bioinformatica, mentre i secondi sono stati utilizzati per il recupero delle annotazioni cellulari necessarie alle successive analisi. Le procedure adottate per il recupero e il processamento dei dati grezzi, nonché per l'estrazione delle annotazioni cellulari, sono descritte nelle sezioni successive.

3.2 Recupero e preparazione dei dati grezzi

Una volta individuati i run di sequenziamento da analizzare, è stato necessario convertirli in un formato utilizzabile nelle successive fasi. A tale scopo è stata utilizzata la piattaforma Galaxy (https://galaxy-main.usegalaxy.org/), attraverso la quale i dati depositati nel Sequence Read Archive (SRA) sono stati estratti in formato FASTQ.

L’estrazione delle reads è stata effettuata mediante il tool “Faster Download and Extract Reads in FASTQ format from NCBI SRA”, disponibile sulla piattaforma Galaxy e basato su fasterq-dump, componente dell’NCBI SRA Toolkit. Fasterq-dump è uno strumento sviluppato per estrarre i dati di sequenziamento contenuti nei record SRA e convertirli in formato FASTQ. Rispetto alla rappresentazione utilizzata per l’archiviazione dei dati nell’SRA, il FASTQ rende direttamente disponibili le singole reads insieme alle rispettive informazioni di qualità. Ogni record FASTQ contiene infatti un identificativo delle reads, la relativa sequenza nucleotidica e i valori di qualità associati alle basi sequenziate. La produzione dei FASTQ costituisce pertanto un passaggio necessario per rendere le reads accessibili agli strumenti utilizzati nelle successive fasi di processamento e allineamento. Per fasterq-dump, sono stati mantenuti i parametri di default, ad eccezione dell’opzione “Select how to split the spots”, per la quale è stata invece selezionata la modalità --split-files. Nell’organizzazione dei dati SRA, uno spot rappresenta l’insieme delle sequenze associate a un determinato evento di sequenziamento e può comprendere più reads. L’opzione --split-files consente di esportare separatamente le differenti reads associate agli spot, generando file FASTQ distinti anziché riunire tutte le componenti in un unico file. Questa scelta è particolarmente importante nel contesto dei dati single-cell e single-nucleus generati mediante tecnologia 10x Genomics, nei quali le diverse reads di una stessa molecola svolgono funzioni differenti. Nelle librerie 10x Genomics paired-end, in particolare, la Read 1 (R1) contiene generalmente le informazioni relative al cell barcode e all’UMI (Unique Molecular Identifier), mentre la Read 2 (R2) contiene la sequenza biologica trascritta, corrispondente al cDNA derivato dall’RNA sequenziato. Il mantenimento delle reads in file separati consente quindi di assegnare ciascun file allo specifico input previsto dai dati del software di analisi.

Infatti, le librerie 10x Genomics utilizzate per la generazione dei dati snRNA-seq associano alla sequenza trascrittomica informazioni tecniche necessarie per ricostruire l’origine cellulare e molecolare delle reads. Il Cell Barcode (CB) è una sequenza che consente di associare le molecole sequenziate al singolo droplet e, dopo le procedure di processamento e correzione del barcode, al corrispondente nucleo. L’Unique Molecular Identifier (UMI) è invece, come anche spiegato precedentemente, una breve sequenza utilizzata per distinguere le singole molecole originariamente catturate. Gli UMI permettono di riconoscere reads originate dall’amplificazione della stessa molecola,

limitando quindi la sovrastima dell'abbondanza trascrizionale dovuta alla duplicazione introdotta durante la PCR.

Al termine dell'esecuzione di fasterq-dump, per ciascun run i dati di interesse venivano restituiti da Galaxy sotto forma di tre dataset FASTQ, distinti dai suffissi _1 (contenente codice paziente), _2 (contenente cell barcode+UMI) e _3 (contenente la sequenza parziale del trascritto, lunghezza = 96-98 bp).

3.3 Allineamento e processamento mediante STARsolo

Solo i file FASTQ contraddistinti dai suffissi _2 e _3 ottenuti nella fase precedente sono stati processati mediante il tool STARsolo [26], adattamento del software di allineamento STAR per l'analisi di dati di scRNA-seq basati sull'impiego di Cell Barcode (CB) e Unique Molecular Identifier (UMI). Nel presente lavoro STARsolo è stato utilizzato attraverso la sua implementazione disponibile sulla piattaforma Galaxy.

I file con suffisso _2 forniscono al programma le informazioni necessarie per identificare Cell Barcode e UMI, mentre quelli con il suffisso _3 contengono la sequenza trascrittomica che viene sottoposta ad allineamento genomico. STARsolo può così associare la posizione genomica individuata per una read cDNA alle informazioni relative al nucleo e alla molecola originale. Questa associazione costituisce un elemento fondamentale del workflow, poiché le successive analisi non richiedono esclusivamente di conoscere dove una read si allinea sul genoma, ma anche di mantenere l'informazione relativa alla cellula o al nucleo dal quale essa deriva.

Nel contesto del presente lavoro di tesi, l'uso di STARsolo è finalizzato alla generazione di un file BAM necessario per l'identificazione delle varianti mediante SComatic. STARsolo produce anche matrici di conteggio e file relativi a geni e barcode, caratteristici del processamento dei dati single-cell, ma non utili allo scopo di questa tesi.

3.3.1 Configurazione di STARsolo e riferimenti genomici

STARsolo è stato configurato utilizzando la modalità default CB_UMI_Simple, appropriata per librerie nelle quali Cell Barcode e UMI occupano posizioni definite nella barcode read.

L'allineamento è stato effettuato utilizzando come riferimento il genoma umano hg38/GRCh38, disponibile in forma indicizzata all'interno di Galaxy. Un genoma di

riferimento rappresenta una sequenza genomica standard utilizzata come sistema di coordinate rispetto al quale vengono confrontate le reads ottenute dal sequenziamento. Durante l'allineamento, STAR individua per ciascuna read cDNA la posizione del genoma di riferimento con la quale essa presenta la migliore compatibilità secondo i criteri dell'algoritmo. In questo modo, alle sequenze presenti nei FASTQ vengono attribuite specifiche coordinate genomiche. Questo passaggio è particolarmente importante in un workflow finalizzato al variant calling, poiché le successive differenze nucleotidiche osservate nelle reads devono essere valutate rispetto alla sequenza presente nel medesimo riferimento genomico.

Al genoma è stata associata un'annotazione genica in formato GTF (Gene Transfer Format). Il GTF contiene informazioni relative alla struttura e alle coordinate degli elementi genici annotati e consente quindi di interpretare gli allineamenti rispetto a geni, trascritti, esoni e giunzioni di splicing.

Nel lavoro è stato utilizzato il file: Homo_sapiens.GRCh38.113.gtf, basato sull'annotazione di Homo sapiens GRCh38 release 113.

Per l'identificazione delle giunzioni di splicing è stato utilizzato come elemento del modello genico exon, mentre la lunghezza della sequenza genomica considerata intorno alle giunzioni annotate è stata impostata a 100 nucleotidi. Lo stesso genoma di riferimento e lo stesso file GTF sono stati utilizzati per tutti i campioni, indipendentemente dalla chimica 10x con cui erano stati generati.

3.3.2 Configurazione della chimica 10x Genomics v2 e v3

Poiché i campioni inclusi nel dataset erano stati generati utilizzando differenti versioni della chimica 10x Genomics Chromium, la configurazione di STARsolo è stata adattata alla chimica associata a ciascun campione. In particolare, sono state considerate separatamente le librerie v2 e v3. La distinzione è necessaria perché la chimica utilizzata determina la struttura della barcode read e, di conseguenza, il modo in cui devono essere interpretate le sequenze contenenti Cell Barcode e UMI. Nelle librerie 10x 3' v2 il Cell Barcode è costituito da 16 nucleotidi e l'UMI da 10 nucleotidi; nella v3 il Cell Barcode rimane di 16 nucleotidi, mentre l'UMI è costituito da 12 nucleotidi.

Tra i parametri di STARsolo, è richiesta una lista di barcode chiamata Cell Barcode whitelist.

La Cell Barcode whitelist, indicata nella documentazione 10x Genomics più recente anche come barcode inclusion list, contiene l'insieme delle sequenze barcode previste per una determinata chimica. I barcode osservati nei dati vengono confrontati con quelli presenti nella lista, permettendo al software di riconoscere le sequenze compatibili con i barcode previsti e di gestire, secondo l'algoritmo impostato, determinate discrepanze riconducibili a errori di sequenziamento.

L'utilizzo della whitelist corrispondente alla chimica v2 o v3 del campione è pertanto essenziale per una corretta interpretazione delle barcode reads. Le whitelist utilizzate sono state reperite attraverso la risorsa Single Cell Genomics Library Structure del Teichmann Lab (https://teichlab.github.io/scg_lib_structs/). Per la chimica v2 è stato utilizzato il file 737-august-2016.txt, mentre per la chimica v3 3M-february-2018.txt.gz.

Anche gli UMI vengono sottoposti a specifiche procedure di filtraggio e deduplicazione. Tale passaggio permette di ridurre l'effetto delle duplicazioni introdotte durante l'amplificazione e di riconoscere reads compatibili con la stessa molecola originaria a un singolo evento molecolare. Alcune impostazioni relative a queste operazioni differivano tra le configurazioni utilizzate per le chimiche v2 e v3. Nei campioni v3, ad esempio, STARsolo era configurato con l'algoritmo di deduplicazione 1MM_CR e con la rimozione degli UMI contenenti basi N e omopolimeri, mentre nei campioni v2 era configurato a 1MM_All. Per il matching dei Cell Barcode alla whitelist è stata invece utilizzata, per entrambe le versioni, la modalità 1MM_multi_nbase_pseudocounts, che consente la gestione di determinate corrispondenze multiple in presenza di basi non definite.

La quantificazione è stata impostata a livello di Gene, contando le reads/UMI attribuibili ai trascritti genici, mentre la strandedness è stata configurata indicando che il filamento della read corrisponde a quello della molecola di RNA originale. Il parametro relativo a cell_filtering è stato impostato su no_filter. Per gli ulteriori parametri avanzati di allineamento e processamento sono state mantenute, ove fosse necessaria una modifica specifica, le impostazioni previste di default.

3.3.3 Generazione e organizzazione degli output

Tra gli output prodotti da STARsolo, quello di maggiore rilevanza per la successiva identificazione delle varianti era costituito dal file degli allineamenti in formato BAM

(Binary Alignment/Map). Il BAM rappresenta la forma binaria e compressa del formato SAM e consente di memorizzare, per ciascuna read, la posizione di allineamento sul genoma insieme alle caratteristiche dell'allineamento e a informazioni aggiuntive codificate mediante specifici tag.

Nel presente workflow era particolarmente importante che il BAM conservasse, oltre all'informazione genomica, anche quella relativa all'origine cellulare delle reads. Tra i tag presenti sono stati inclusi CR e CY, relativi al Cell Barcode originale e alla sua qualità, UR e UY, relativi all'UMI originale e alla sua qualità, CB e UB, contenenti rispettivamente Cell Barcode e UMI processati, e GX e GN, associati all'identificazione genica. In particolare, la presenza del Cell Barcode nel tag CB permette di mantenere nel BAM il collegamento tra una read allineata e il nucleo al quale essa è stata attribuita. Questa informazione sarà successivamente necessaria per associare gli allineamenti alle annotazioni cellulari e consentire a SComatic di analizzare le reads in funzione delle differenti popolazioni cellulari.

Per gli allineamenti univoci il valore di MAPQ (Mapping Quality) è stato impostato a 60, che è il valore massimo. Il MAPQ rappresenta una misura della confidenza associata all'assegnazione di una read a una determinata posizione genomica. La qualità dell'allineamento assume particolare importanza in un workflow di Variant calling, poiché reads erroneamente assegnate a regioni genomiche non corrette possono introdurre differenze apparenti rispetto alla sequenza di riferimento.

L'esecuzione di STARsolo rappresentava una delle fasi computazionalmente più impegnative del preprocessing. Il tempo necessario dipendeva dalle dimensioni del campione e dalle risorse disponibili sulla piattaforma Galaxy; nelle analisi effettuate, il completamento di un singolo job e la generazione degli output richiedevano generalmente almeno 3-4 ore.

3.4 Estrazione e organizzazione delle annotazioni cellulari

Parallelamente al processamento dei dati grezzi di sequenziamento, è stato necessario recuperare le annotazioni cellulari associate ai nuclei analizzati, indispensabili per distinguere le differenti popolazioni cellulari presenti nei campioni tumorali. A tale scopo è stato utilizzato il file in formato h5ad reso disponibile dagli autori del dataset GSE202051, insieme ai dati processati dello studio.

Il formato h5ad è un formato gerarchico comunemente utilizzato per la memorizzazione di dataset di trascrittomica a singola cellula. Esso permette di raccogliere all'interno dello stesso oggetto differenti componenti del dataset, tra cui la matrice di espressione e le informazioni associate alle singole cellule. In questo caso, il file .h5ad fornito dagli Autori contiene anche le annotazioni cellulari da essi definite nell'ambito dello studio condotto sul dataset GSE202051. I risultati ottenuti mediante l'analisi di questo dataset sono stati pubblicati dagli stessi Autori nell'articolo associato al dataset [25]. La definizione del tipo cellulare per ogni nucleo analizzato è un'informazione necessaria per il tool SComatic, applicato nelle fasi successive. La lettura dei metadati di ogni paziente presenti nel file .h5ad è stata effettuata usando la libreria AnnData in ambiente Python.

Tra le annotazioni cellulari definite dagli Autori erano presenti diverse popolazioni rappresentative della componente tumorale e del microambiente, tra cui fibroblasti associati al tumore (Cancer-associated fibroblasts), cellule endoteliali, cellule endocrine, popolazioni linfoidi e mieloidi, periciti e cellule di Schwann. Di particolare rilevanza per il presente lavoro di tesi era la distinzione della componente epiteliale nelle categorie Epithelial (malignant) ed Epithelial (non-malignant). Nello studio originale, la distinzione tra cellule epiteliali maligne e non maligne è stata supportata dall'analisi delle alterazioni del numero di copie (copy number alterations, CNA) inferite dai profili di espressione mediante InferCNV. Tale distinzione risultava particolarmente importante, poiché SComatic richiede l'associazione dei barcode cellulari alle rispettive popolazioni cellulari.

3.5 Identificazione delle varianti somatiche candidate mediante SComatic

L'identificazione delle varianti somatiche candidate è stata eseguita mediante SComatic [24], utilizzando come input i file BAM ottenuti dall'allineamento con STARsolo e le annotazioni cellulari precedentemente estratte dal file h5ad e organizzate separatamente per ciascun paziente. L'analisi è stata condotta in ambiente Linux mediante un ambiente Conda/Python dedicato a SComatic. Per ogni campione è stata predisposta una directory di lavoro indipendente. SComatic richiede una sequenza genomica di riferimento in formato FASTA: è stato utilizzato il genoma umano hg38. Il file di riferimento era stato

preventivamente decompresso e indicizzato mediante “samtools faidx”. La pipeline SComatic è stata applicata separatamente ai diversi campioni ed è articolata in quattro fasi principali: i) suddivisione del file BAM in funzione del tipo cellulare, ii) conteggio delle basi nelle differenti popolazioni cellulari, iii) integrazione delle matrici di conteggio e iv) identificazione delle varianti somatiche candidate. Al termine del variant calling è stato inoltre applicato un ulteriore filtro basato sulla mappabilità delle regioni genomiche e sull’esito di ulteriori filtri implementati da SComatic (Appendice A).

3.5.1 Preparazione dei dati e suddivisione per tipo cellulare

La prima fase della pipeline è stata eseguita mediante lo script SplitBamCellTypes.py di SComatic. Per ciascun campione sono stati forniti allo script il file BAM precedentemente generato mediante STARsolo e il corrispondente file TSV contenente l’associazione tra barcode e annotazione cellulare, ottenuto come descritto nella sezione precedente. Attraverso la corrispondenza tra il barcode presente nel BAM e quello riportato nel file di annotazione, gli allineamenti sono stati separati in funzione del tipo cellulare di appartenenza, producendo vari file BAM distinti per le differenti popolazioni cellulari presenti nel campione. In questa fase sono stati impostati i parametri --n_trim 5, --max_nM 5, --max_NH 1 e --min_MQ 60. Il parametro --n_trim 5 è stato utilizzato per escludere dal successivo conteggio le cinque basi terminali delle reads, maggiormente suscettibili alla presenza di artefatti tecnici. Nell’implementazione di SComatic tale operazione non determina la rimozione fisica delle basi dalla sequenza, ma ne impedisce il contributo alle successive operazioni di conteggio attraverso la modifica della relativa qualità.

Il parametro --max_nM 5 è stato utilizzato per limitare l’analisi alle reads che non superavano cinque mismatch rispetto al genoma di riferimento, mentre --max_NH 1 ha consentito di mantenere esclusivamente reads con un singolo allineamento riportato, riducendo il contributo di allineamenti multipli o ambigui.

Una modifica specifica rispetto alla configurazione comunemente indicata da SComatic per BAM di scRNA-seq prodotti mediante Cell Ranger ha riguardato la soglia minima di mapping quality (MAPQ). Poiché nel presente lavoro i file BAM erano stati generati mediante STARsolo, il parametro --min_MQ è stato impostato a 60, che è anche il valore massimo possibile. Questi parametri sopra indicati sono tutti volti a selezionare reads di altissima qualità per le successive fasi.

3.5.2 Conteggio delle basi e identificazione delle varianti candidate

I BAM specifici per tipo cellulare ottenuti nella fase precedente sono stati elaborati mediante lo script BaseCellCounter.py. L'esecuzione è stata automatizzata in modo da processare consecutivamente tutti i file BAM presenti nella directory relativa al singolo campione.

Il conteggio è stato effettuato utilizzando il genoma hg38 come riferimento ed estendendo l'analisi a tutti i cromosomi mediante l'operazione --chrom all. Sono state impostate una soglia minima di base quality pari a 30 (--min_bq 30) e una soglia minima di mapping quality pari a 60 (--min_mq 60). La base quality esprime l'affidabilità della chiamata del singolo nucleotide all'interno della read, mentre la mapping quality rappresenta l'affidabilità dell'assegnazione della read alla corrispondente posizione genomica. L'impostazione di --min_bq 30 ha quindi determinato l'utilizzo, per il conteggio, delle sole basi che soddisfacevano tale requisito di qualità. Tale soglia è più stringente rispetto al valore predefinito pari a 20. La soglia --min_mq 60 è stata invece mantenuta coerente con quella applicata nello step precedente.

L'elaborazione è stata parallelizzata utilizzando otto processori (--nprocs 8), al fine di ridurre i tempi di calcolo che erano indicativamente compresi tra le 8 e le 10 ore per paziente.

Al termine di questa fase sono stati ottenuti, per ciascun tipo cellulare, i conteggi delle basi osservate nelle posizioni genomiche analizzate, utilizzati come input per la successiva integrazione dei dati.

Le informazioni di conteggio prodotte separatamente per le diverse popolazioni cellulari sono state successivamente integrate mediante lo script MergeBaseCellCounts.py. Lo script ha prodotto, per ciascun campione, un'unica matrice contenente congiuntamente le informazioni relative ai differenti tipi cellulari. Il file risultante è stato denominato secondo la struttura: <sample>.BaseCellCounts.AllCellTypes.tsv e ha costituito l'input per la successiva fase di identificazione delle varianti.

L'identificazione delle varianti somatiche candidate è stata effettuata attraverso i due passaggi consecutivi previsti da SComatic, utilizzando gli script BaseCellCalling.step1.py e BaseCellCalling.step2.py. In termini generali, SComatic utilizza una combinazione di filtri e test statistici per distinguere le varianti candidate dagli eventi non somatici e dagli artefatti. In particolare, il metodo utilizza un modello beta-binomiale per

caratterizzare il tasso di errore di fondo della tecnologia e valutare se il numero di reads contenenti l'allele alternativo sia superiore a quello atteso in condizioni di rumore. Il modello viene parametrizzato utilizzando dati provenienti da campioni non neoplastici, permettendo di tenere conto delle caratteristiche specifiche del rumore associato ai dati single-cell [24].

In particolare, nel primo passaggio, BaseCellCalling.step1.py è stato applicato alla matrice <sample>.BaseCellCounts.AllCellTypes.tsv ottenuta nella fase precedente, utilizzando il genoma hg38 come riferimento. In questa fase non sono stati modificati i parametri di default di SComatic per la selezione iniziale delle varianti candidate e per l'applicazione del modello statistico implementato dalla pipeline. Tali parametri consentono di stabilire requisiti minimi di copertura e di supporto alle varianti, riducendo il rischio che errori di sequenziamento, errori di allineamento o osservazioni sporadiche vengano erroneamente identificate come mutazioni somatiche.

In particolare, il parametro --min_cov è stato impostato a 5 e definisce la profondità minima di copertura richiesta affinché una posizione genomica possa essere considerata nell'analisi. L'utilizzo di una copertura minima permette di escludere le posizioni per le quali il numero di reads disponibili è troppo basso per consentire una valutazione affidabile della presenza di una variante.

Il parametro --min_cells è stato impostato a 5 e definisce il numero minimo di cellule che devono presentare una copertura sufficiente della posizione genomica affinché il sito venga incluso nelle analisi successive. Questo filtro è particolarmente importante nei dati single-cell, caratterizzati da una copertura genomica generalmente più limitata e variabile rispetto ai dati di sequenziamento bulk. Richiedere che il sito sia adeguatamente coperto in più cellule consente di evitare che le chiamate di variante siano basate su informazioni provenienti da un numero eccessivamente ridotto di cellule.

Il parametro --min_ac_cells è stato impostato a 2 e definisce il numero minimo di cellule che devono supportare l'allele alternativo affinché una variante possa essere considerata una potenziale mutazione. La richiesta di supporto da parte di almeno due cellule riduce la probabilità che una variante osservata in una singola cellula rappresenti un errore di sequenziamento, un errore di allineamento o un altro artefatto tecnico.

Infine, il parametro --min_ac_reads è stato impostato a 3 e definisce il numero minimo di reads che devono supportare l'allele alternativo affinché la variante venga considerata

nella chiamata. Questo requisito fornisce un'ulteriore evidenza a supporto della variante e contribuisce a ridurre i falsi positivi dovuti a singole reads contenenti errori di sequenziamento.

Il primo passaggio ha prodotto, per ciascun campione, il file: <sample>.calling.step1.tsv contenente le posizioni candidate sottoposte al successivo livello di filtraggio.

Nel secondo passaggio, il file prodotto da BaseCellCalling.step1.py è stato elaborato mediante BaseCellCalling.step2.py. In questa fase sono state utilizzate due risorse specifiche messe a disposizione da SComatic: il file AllEditingSites.hg38.txt, contenente i siti di RNA editing considerati dalla pipeline, e il Panel of Normals PoN.scRNAseq.hg38.tsv, specifico per dati di single-cell RNA sequencing. In particolare, SComatic considera i siti di RNA editing conosciuti e gli SNP comuni presenti nel database gnomAD. SComatic utilizza inoltre un Panel of Normals (PoN), ottenuto da un insieme di campioni non neoplastici, per identificare varianti ricorrenti associate a errori tecnici o di allineamento. L'utilizzo di questo tipo di riferimento permette di caratterizzare gli errori sistematici della tecnologia e di ridurre la probabilità che essi vengano interpretati come mutazioni somatiche [24]. Tali risorse consentono a SComatic di sottoporre le varianti candidate a ulteriori criteri di filtraggio finalizzati a ridurre la presenza di eventi compatibili con fonti di variazione non riconducibili a mutazioni somatiche.

L'esecuzione del secondo passaggio ha prodotto il file: <sample>.calling.step2.tsv contenente le varianti candidate e le relative informazioni sull'esito dei criteri di filtraggio applicati dalla pipeline.

3.5.3 Filtraggio e selezione delle varianti

Al termine del secondo step di variant calling, le varianti candidate sono state sottoposte a un ulteriore filtraggio sulla base delle caratteristiche delle regioni genomiche nelle quali erano localizzate. A tale scopo è stato utilizzato BEDtools, mediante il comando bedtools intersect, confrontando le coordinate delle varianti contenute nel file <sample>.calling.step2.tsv con quelle riportate nel file messo a disposizione da SComatic denominato UCSC.k100_umap.without.repeatmasker.bed. Questo file BED contiene regioni genomiche selezionate sulla base delle informazioni di mappabilità Umap. La componente k100_umap fa riferimento alla valutazione della mappabilità basata su k-mer

di lunghezza pari a 100 nucleotidi, utilizzata per individuare regioni nelle quali sequenze di tale lunghezza possono essere ricondotte in maniera univoca alla corrispondente posizione del genoma. La restrizione dell'analisi a tali regioni permette di ridurre il contributo di posizioni nelle quali l'elevata similarità di sequenza tra loci differenti potrebbe determinare allineamenti ambigui e, conseguentemente, aumentare il rischio di segnali artefatti nel variant calling.

Il file utilizzato permette di escludere inoltre le regioni annotate come ripetitive mediante RepeatMasker, come indicato dalla componente without.repeatmasker del nome del file. Le sequenze ripetitive possono infatti presentare elevata similarità tra differenti regioni genomiche e rappresentare pertanto una fonte di ambiguità nell'assegnazione delle reads alla corretta posizione genomica. L'intersezione mediante BEDTools delle coordinate presenti nel file BED di riferimento con quelle presenti nel nostro file .tsv, ha consentito quindi di selezionare solo le varianti ricadenti nelle regioni genomiche comprese nel file BED. Sul risultato ottenuto è stata successivamente applicata un'ulteriore selezione mediante awk, mantenendo esclusivamente le righe nelle quali la colonna di filtraggio riportava il valore PASS.

Per ciascun campione è stato così generato il file finale: <sample>.calling.step2.pass.tsv

Questo file prodotto da SComatic conserva, per ciascuna variante, sia le informazioni genomiche necessarie alla sua identificazione sia le evidenze quantitative utilizzate durante il processo di chiamata. Le colonne #CHROM, Start ed End riportano rispettivamente il cromosoma e le coordinate genomiche dell'evento, mentre REF e ALT identificano l'allele di riferimento e l'allele alternativo osservato. La colonna FILTER riporta l'esito complessivo dei criteri di filtraggio applicati dalla pipeline; poiché in questa fase sono stati utilizzati i file calling.step2.pass.tsv, le varianti considerate erano già state selezionate sulla base del valore PASS. Un'informazione centrale ai fini della successiva selezione è contenuta nella colonna Cell_types, che identifica il tipo cellulare al quale SComatic attribuisce la variante candidata. Tale informazione deriva dal confronto effettuato dalla pipeline tra le differenti popolazioni cellulari presenti nello stesso campione e consente quindi di mantenere, insieme all'identità genomica della variante, anche l'informazione relativa alla popolazione cellulare alla quale essa è stata attribuita.

Le colonne successive forniscono informazioni quantitative relative al supporto della variante nel tipo cellulare considerato. In particolare, Dp esprime la profondità di copertura del sito in termini di reads e Nc il numero di cellule che contribuiscono alla copertura della posizione. Bc indica il numero di reads che supportano l'allele alternativo, mentre Cc riporta il numero di cellule nelle quali tale allele è stato osservato.

A partire da tali informazioni, VAF (Variant Allele Frequency) descrive la frazione di reads che supportano l'allele alternativo rispetto alla copertura disponibile, mentre CCF esprime la corrispondente frazione a livello cellulare, fornendo quindi una misura del supporto alla variante basata sulle cellule informative. Questi parametri permettono di descrivere l'evidenza a sostegno della variante secondo due livelli complementari, rispettivamente riferiti alle reads e alle singole cellule.

L'output comprende inoltre informazioni relative ai criteri statistici e di filtraggio applicati durante il processo di chiamata. Le colonne BCp e CCp riportano i valori associati alla valutazione statistica effettuata rispettivamente a livello delle reads e delle cellule, mentre Cell_types_min_BC e Cell_types_min_CC indicano il numero di tipi cellulari che soddisfano i requisiti minimi previsti dalla pipeline per la valutazione del sito. Rest_BC e Rest_CC descrivono l'eventuale supporto ad ulteriori alleli alternativi osservati nella medesima posizione.

Nella parte finale del file sono inoltre presenti informazioni specifiche per le differenti popolazioni cellulari considerate nel campione. La struttura di tali informazioni è indicata nella colonna INFO e comprende parametri relativi alla copertura, al numero di cellule informative, ai conteggi delle basi e alle relative caratteristiche qualitative. La presenza di queste informazioni consente di conservare nell'output finale le evidenze utilizzate da SComatic per valutare la posizione genomica nelle differenti popolazioni cellulari.

L'espressione "varianti somatiche candidate" è stata mantenuta nelle successive fasi dell'analisi, poiché l'identificazione computazionale mediante SComatic costituisce una procedura di chiamata e filtraggio delle varianti e non una validazione sperimentale indipendente della loro origine somatica.

L'intera procedura è stata eseguita separatamente per i campioni inclusi nell'analisi, mantenendo per ciascuno una struttura di directory organizzata secondo le principali fasi della pipeline: BAM suddivisi per tipo cellulare, conteggi delle basi, matrice integrata e risultati del variant calling. Questa organizzazione ha consentito di preservare gli output

intermedi e, quando necessario, di riprendere l'elaborazione da uno specifico passaggio senza dover rieseguire le fasi precedenti.

3.6 Selezione delle varianti associate alla componente epiteliale maligna

Al termine del workflow SComatic, per ciascun campione è stato ottenuto il file <sample>.calling.step2.pass.tsv, contenente le varianti somatiche candidate che avevano superato i criteri di filtraggio descritti nelle sezioni precedenti. Prima di procedere alle successive fasi di annotazione genomica e funzionale, i file sono stati ulteriormente esaminati al fine di restringere l'analisi alle varianti attribuite alle cellule tumorali. In particolare, sono state considerate separatamente le varianti rilevate sia nelle cellule tumorali sia nelle popolazioni cellulari non tumorali dello stesso paziente. La presenza di una variante in entrambi i compartimenti è infatti compatibile con un'origine germinale, poiché indica che la variante è condivisa da differenti tipi cellulari dell'individuo e non è quindi specificamente associata alla trasformazione tumorale. Al contrario, le varianti rilevate esclusivamente nelle cellule tumorali sono state considerate candidate varianti somatiche, in quanto compatibili con alterazioni acquisite dalle cellule tumorali e non rilevate nelle popolazioni cellulari non tumorali. Tale distinzione ha permesso di concentrare le successive analisi di annotazione genomica e funzionale sulle varianti maggiormente compatibili con un'origine somatica e quindi potenzialmente associate al fenotipo tumorale. Per ciascun campione, ogni file <sample>.calling.step2.pass.tsv è stato filtrato in base al contenuto della colonna Cell_types selezionando esclusivamente le righe corrispondenti alla categoria Epithelial_(malignant). Le varianti attribuite alle altre popolazioni cellulari, comprese quelle identificate nelle cellule epiteliali annotate come non maligne, non sono state incluse nelle successive fasi di annotazione e caratterizzazione.

3.7 Annotazione genomica mediante UCSC Genome Browser

Le varianti somatiche candidate attribuite alla popolazione delle cellule epiteliali maligne, selezionate secondo la procedura descritta nelle sezioni precedenti, sono state sottoposte a una fase sistematica di annotazione finalizzata a integrarne le coordinate genomiche con informazioni relative al contesto genico, alla possibile rilevanza funzionale e alle

evidenze disponibili nei principali database di varianti. Come primo livello di caratterizzazione è stato utilizzato UCSC Genome Browser (Figura 6), facendo riferimento all'assembly umano GRCh38/hg38. Le coordinate cromosomiche di ogni variante sono state cercate in UCSC Genome Browser [27], consentendo di visualizzare il locus genomico corrispondente nel contesto delle annotazioni genomiche disponibili. L'identificazione del gene e la valutazione del contesto trascrizionale sono state effettuate utilizzando principalmente le track GENCODE V50 e NCBI RefSeq. In questo modo, alle informazioni originariamente prodotte dal variant calling, ovvero coordinate genomiche e sostituzione REF/ALT, è stato possibile associare il gene interessato e la relativa localizzazione nel contesto genico (esone, introne, UTR).
L'identificazione del gene rappresentava inoltre un passaggio fondamentale per le analisi successive.

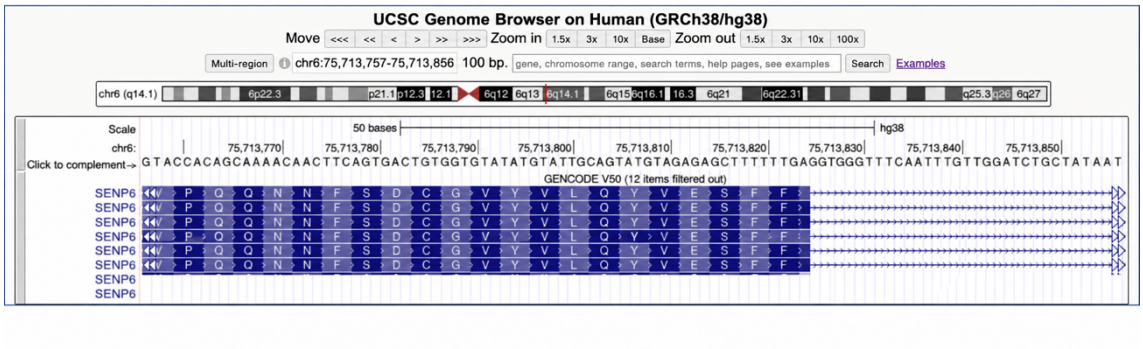


Figura 6. Esempio di contestualizzazione genomica di una variante somatica candidata mediante UCSC Genome Browser. Visualizzazione della regione genomica comprendente la variante candidata localizzata sul cromosoma 6 in posizione 75.713.807 (GRCh38/hg38), in corrispondenza del gene *SENP6*. La track GENCODE V50 consente di contestualizzare la posizione della variante rispetto alla struttura trascrizionale e alla sequenza codificante del gene. Fonte: elaborazione originale mediante UCSC Genome Browser.

L'esame del locus non è stato limitato alla sola annotazione genica. Sono state annotate informazioni anche usando le track dedicate alla caratterizzazione funzionale e clinica delle varianti, tra cui AlphaMissense, CADD 1.6 e 1.7, CIViC, ClinGen, ClinVar Variants, COSMIC e COSMIC Regions, Deleteriousness Predictions, G2P Project, GenCC, GWAS Catalog, HGMD Public 2025, Orphanet, PanelApp, REVEL Scores, SNPedia, Splicing Impact, Variants in Papers e dbSNP 155. La consultazione congiunta di queste risorse ha consentito di contestualizzare le varianti rispetto a differenti livelli

informativi, comprendenti la presenza in database di varianti note, eventuali associazioni con fenotipi patologici o neoplastici, predizioni computazionali di impatto e informazioni riportate in letteratura.

3.8 Predizione degli effetti funzionali mediante AlphaGenome

Per approfondire le possibili conseguenze funzionali delle varianti candidate è stato utilizzato AlphaGenome, un modello di deep learning sequence-to-function sviluppato da Google DeepMind. Il modello rappresenta un approccio innovativo per l'interpretazione funzionale delle varianti genomiche, consentendo di effettuare predizioni direttamente a partire dalla sequenza di DNA. AlphaGenome è stato descritto per la prima volta nel 2026 nella pubblicazione di Avsec et al. su *Nature* [28].

A differenza degli strumenti orientati principalmente alla classificazione delle conseguenze di varianti codificanti, AlphaGenome permette di esplorare simultaneamente differenti livelli della regolazione genomica. Il modello riceve in input fino a 1 Mb di sequenza di DNA e predice migliaia di tracce funzionali riferite a differenti tipi cellulari e tessuti, comprendendo informazioni relative all'espressione genica, all'inizio della trascrizione, all'accessibilità della cromatina, alle modificazioni istoniche, al legame di fattori di trascrizione e allo splicing. Il modello comprende inoltre predizioni relative all'organizzazione tridimensionale della cromatina [28].

Una caratteristica particolarmente rilevante per l'interpretazione delle varianti consiste nella possibilità di confrontare le predizioni ottenute utilizzando la sequenza contenente l'allele di riferimento (REF) con quelle generate dopo l'introduzione dell'allele alternativo (ALT). La differenza tra le due predizioni permette quindi di stimare, in silico, il possibile effetto della variante su diversi output molecolari modellizzati. Questo approccio è risultato particolarmente interessante nel presente studio poiché le varianti identificate da SComatic non erano tutte limitate a regioni codificanti.

AlphaGenome ha pertanto consentito di affiancare alle annotazioni genomiche tradizionali una valutazione predittiva di possibili effetti regolatori e trascrizionali, inclusi quelli associati a varianti localizzate al di fuori degli esoni codificanti.

Considerato il numero di varianti analizzate e la necessità di applicare una procedura uniforme ai diversi pazienti, l'interrogazione di AlphaGenome è stata automatizzata mediante uno script sviluppato in Python, utilizzando l'API e l'SDK del modello. Lo

script completo è riportato in Appendice B; nella presente sezione vengono pertanto descritti esclusivamente gli aspetti metodologici necessari a comprendere e riprodurre l'analisi.

3.8.1 Configurazione dell'analisi

Per ciascun paziente è stato predisposto un file tabulato contenente le varianti da sottoporre all'analisi. L'input comprendeva obbligatoriamente un identificativo univoco della variante (variant_id), il cromosoma (CHROM), la posizione genomica (POS), l'allele di riferimento (REF) e l'allele alternativo (ALT). A tali informazioni è stato associato il nome del gene precedentemente identificato durante l'annotazione mediante UCSC Genome Browser.

Prima dell'interrogazione del modello, lo script uniformava la nomenclatura cromosomica nel formato chr1, chr2, ..., chrX. Le coordinate sono riferite a hg38, coordinate 1-based.

Le predizioni di AlphaGenome sono strettamente dipendenti dal contesto genomico: all'aumentare dell'ampiezza della finestra considerata, cresce proporzionalmente l'accuratezza del modello. Di conseguenza, si è scelto di adottare la massima finestra consentita, pari a 1 Mb, centrandola attorno alla coordinata cromosomica di interesse. È stato inoltre specificato l'identificativo anatomico UBERON:0001264 (pancreas), utilizzato per contestualizzare le predizioni e selezionare, quando disponibili, gli score associati al contesto biologico considerato.

3.8.2 Predizioni multimodali e scoring delle varianti

La pipeline è stata configurata per interrogare differenti categorie di output biologico disponibili in AlphaGenome. In particolare, sono state richieste predizioni relative a RNA-seq, CAGE, PRO-cap, DNase, ATAC, ChIP-seq per modificazioni istoniche, ChIP-seq per fattori di trascrizione, splice sites, splice junctions e splice-site usage.

L'impiego di output differenti ha consentito di estendere la caratterizzazione oltre le sole conseguenze sulla sequenza codificante, considerando potenziali modificazioni dell'attività trascrizionale, dell'accessibilità della cromatina, dell'attività regolatoria e dei processi di splicing.

Le predizioni RNA-seq forniscono informazioni relative all'output trascrizionale e permettono di valutare se la sostituzione REF → ALT sia associata a una variazione del livello di espressione del gene di interesse.

Le predizioni CAGE (*Cap Analysis of Gene Expression*) e PRO-cap (*Precision Run-On and sequencing*) forniscono informazioni sull'attività trascrizionale in prossimità dei siti di inizio della trascrizione (TSS). Gli output DNase e ATAC riguardano invece l'accessibilità della cromatina, ossia la maggiore o minore disponibilità della regione genomica all'interazione con proteine regolatorie. Le predizioni ChIP-histone riguardano segnali associati a specifiche modificazioni degli istoni, che possono essere correlate a differenti stati funzionali e regolatori della cromatina. Le predizioni ChIP-TF sono relative al legame dei fattori di trascrizione al DNA.

Infine, splice sites, splice usage e splice junctions descrivono aspetti complementari dello splicing. Le relative predizioni sono state utilizzate per valutare possibili modificazioni dei siti di splicing e dell'utilizzo delle giunzioni conseguenti alla presenza dell'allele alternativo.

L'utilizzo simultaneo di differenti modalità ha permesso di ottenere, per ciascuna variante, una caratterizzazione multidimensionale del possibile effetto regolatorio. AlphaGenome, infatti, è stato sviluppato come modello multimodale, nel quale i diversi output vengono predetti all'interno di un unico framework sequence-to-function [28].

Per ciascuna variante sono stati richiesti gli score raccomandati disponibili nell'SDK di AlphaGenome. Gli score prodotti non rappresentavano una misura diretta della patogenicità, ma quantificavano il possibile effetto della sostituzione REF → ALT rispetto a differenti modalità biologiche.

Di conseguenza, un valore elevato per un determinato output AlphaGenome non è stato interpretato automaticamente come indicazione di maggiore patogenicità. Gli score sono stati considerati nel contesto della specifica modalità analizzata, prestando attenzione alla presenza, all'entità e, quando è interpretabile, alla direzione della variazione predetta tra allele di riferimento e allele alternativo.

Tale distinzione è particolarmente importante poiché AlphaGenome costituisce uno strumento di predizione funzionale, non un sistema di classificazione clinica delle varianti. I risultati sono stati pertanto utilizzati come evidenze computazionali da integrare con le annotazioni ottenute mediante altri strumenti. Per la selezione delle

predizioni AlphaGenome riportate nella matrice finale sono stati considerati i risultati caratterizzati da quantile_score < −0,99 oppure quantile_score > 0,99, corrispondenti quindi a |quantile_score| > 0,99. Tale criterio è stato utilizzato per restringere le informazioni inserite nella tabella ai risultati collocati agli estremi della distribuzione dello score. La soglia è stata pertanto utilizzata come criterio di selezione delle predizioni e non come misura della probabilità di correttezza del risultato. Le evidenze AlphaGenome che soddisfacevano tale criterio sono state riportate mediante una codifica sintetica, concepita per preservare sia la presenza di una predizione selezionata sia, quando disponibile, l'informazione relativa al gene o alla specifica caratteristica molecolare a cui essa era associata. In questo modo, i risultati delle diverse modalità sono stati ricondotti a una struttura comune, mantenendo tuttavia il livello di dettaglio necessario per distinguere la natura dell'effetto predetto.

Per alcune componenti, oltre alla presenza di un risultato che soddisfacesse il criterio di selezione, è stata mantenuta anche l'informazione relativa alla direzione o alla natura specifica della predizione. Nel caso dell'RNA-seq, ad esempio, è stata conservata l'indicazione relativa all'aumento o alla diminuzione del segnale predetto per l'allele alternativo rispetto a quello di riferimento. Per le modalità caratterizzate da informazioni più specifiche, come quelle relative alle modificazioni istoniche, sono state invece riportate le denominazioni presenti negli output di AlphaGenome, preservando così l'indicazione della specifica caratteristica molecolare associata alla predizione. Un criterio analogo è stato adottato per le altre modalità quando l'output forniva informazioni aggiuntive utili alla caratterizzazione dell'effetto.

La codifica adottata ha inoltre permesso di distinguere le evidenze riferite al gene associato alla variante da quelle eventualmente attribuite dal modello a geni differenti.

3.8.3 Analisi e organizzazione degli output

Per supportare la rappresentazione grafica delle predizioni, la pipeline utilizzava le annotazioni GENCODE v46 riferite a hg38. Lo script selezionava i trascritti protein-coding con transcript support level 1 e, ai fini della visualizzazione, manteneva il trascritto più lungo, riducendo la complessità derivante dalla presenza di molteplici isoforme. Grazie al nome del gene, le annotazioni GENCODE venivano inoltre interrogate per determinarne lo strand, utilizzato per orientare le tracce ai fini della visualizzazione.

Un elemento centrale dell'analisi mediante AlphaGenome è stato il confronto tra le predizioni ottenute per la sequenza contenente l'allele di riferimento (REF) e quelle generate in presenza dell'allele alternativo (ALT).

Per gli output compatibili, la pipeline produceva una rappresentazione grafica sovrapposta dei profili REF e ALT, accompagnata da una seconda visualizzazione della differenza ALT – REF. Tale rappresentazione consentiva di evidenziare le regioni nelle quali la sostituzione nucleotidica determinava una variazione della predizione del modello. La posizione della variante veniva indicata all'interno dei grafici e, quando disponibili, venivano sovrapposte le informazioni relative alla struttura trascrizionale.

Gli output della pipeline sono stati organizzati separatamente per paziente e variante in score numerici, rappresentazioni grafiche e file log. Gli score filtrati sono stati esportati in formato XLSX e successivamente integrati nel dataset complessivo.

3.9 Annotazioni funzionali mediante Ensembl Variant Effect Predictor

A completamento della caratterizzazione, le varianti sono state sottoposte ad annotazione mediante Ensembl Variant Effect Predictor (VEP) [29]. VEP consente di determinare le conseguenze di una variante rispetto alle annotazioni genomiche e trascrizionali e di integrare numerose risorse esterne dedicate alla predizione dell'impatto funzionale, allo splicing, alla frequenza di popolazione e alle caratteristiche proteiche e regolatorie. L'analisi è stata effettuata mediante l'interfaccia web di VEP e applicata alle varianti precedentemente selezionate per ciascun paziente. Come transcript set è stato utilizzato GENCODE 50.

Poiché i tool esterni usati da VEP utilizzano modelli, scale e criteri differenti, i relativi valori non sono stati considerati direttamente sovrapponibili né interpretati come dimostrazione autonoma di patogenicità.

Per le varianti missenso sono stati selezionati i tool AlphaMissense, REVEL, ClinPred e EVE, che erano già stati considerati precedentemente mediante UCSC Genome Browser. AlphaMissense è un modello di deep learning sviluppato per stimare il possibile effetto patogenetico delle sostituzioni missenso integrando informazioni relative al contesto strutturale ed evolutivo della proteina. Lo score am_pathogenicity è compreso tra 0 e 1, con valori maggiori associati a una maggiore probabilità predetta di patogenicità. Secondo le

categorie definite dagli autori e implementate anche nel plugin AlphaMissense di Ensembl VEP, valori inferiori a 0,34 sono classificati come likely benign, valori superiori a 0,564 come likely pathogenic, mentre l'intervallo compreso tra tali soglie è considerato ambiguous. Tali cut-off sono stati definiti per ottenere un'elevata precisione nelle rispettive categorie e sono stati utilizzati come riferimento interpretativo, senza considerarli equivalenti a una classificazione clinica della variante [30,31].

REVEL (Rare Exome Variant Ensemble Learner) è un metapredittore sviluppato specificamente per le varianti missenso rare e combina le informazioni provenienti da molteplici strumenti di predizione. Lo score varia da 0 a 1 e valori maggiori indicano una maggiore evidenza computazionale a favore di un possibile effetto patogenetico. Nella pubblicazione originale e in successive applicazioni è stato frequentemente utilizzato il valore 0,5 come riferimento per distinguere predizioni maggiormente orientate verso un effetto benigno o patogenetico; valori più elevati, ad esempio maggiori di 0,75, rappresentano un'evidenza predittiva più stringente. Tali soglie sono state considerate come riferimenti interpretativi e non come criteri diagnostici autonomi [32].

ClinPred è un classificatore sviluppato per distinguere varianti missenso potenzialmente rilevanti per malattia da varianti verosimilmente neutrali, integrando differenti score di patogenicità con informazioni relative alla frequenza allelica. Anche ClinPred restituisce uno score compreso tra 0 e 1; valori > 0,5 sono classificati come damaging, mentre valori inferiori a tale soglia sono classificati come tolerated [33].

EVE (Evolutionary model of Variant Effect) utilizza un approccio basato sulla variazione evolutiva delle sequenze proteiche per stimare il possibile effetto delle varianti missenso. Poiché l'output e le relative categorie non sono direttamente sovrapponibili a quelli degli altri predittori, l'annotazione EVE è stata mantenuta separatamente e utilizzata come ulteriore evidenza a supporto della valutazione complessiva [34].

È stato inoltre considerato CADD (Combined Annotation Dependent Depletion), uno score che integra numerose annotazioni genomiche e funzionali con lo scopo di stimare la deleteriousness relativa delle varianti. Anche questo valore è stato annotato sia da VEP che da UCSC. Per l'interpretazione è stato considerato il valore CADD PHRED, per il quale i valori maggiori identificano varianti progressivamente più elevate nel ranking di potenziale impatto. La scala consente un'interpretazione percentile: un valore pari a 10 corrisponde approssimativamente al 10% delle varianti con score più elevato, mentre un

valore pari a 20 corrisponde approssimativamente all'1% delle varianti maggiormente prioritarie secondo CADD. Valori elevati non dimostrano tuttavia la patogenicità delle varianti, ma costituiscono un'indicazione computazionale di possibile rilevanza funzionale [35].

La possibile influenza delle varianti sui processi di splicing è stata valutata mediante SpliceAI, un algoritmo di deep learning sviluppato per predire, direttamente a partire dalla sequenza nucleotidica, possibili alterazioni dei siti di splicing conseguenti alla presenza di una variante [37]. Nel presente lavoro le informazioni prodotte da SpliceAI sono state consultate sia attraverso UCSC che attraverso VEP, permettendo di integrare la valutazione dello splicing nelle differenti fasi del processo di annotazione.

Durante l'annotazione mediante UCSC è stata utilizzata la track Splicing Impact – SpliceAI Variants, disponibile per l'assembly GRCh38/hg38. Tale track contiene gli score di SpliceAI e consente di visualizzare, in corrispondenza della posizione genomica analizzata, il possibile effetto della sostituzione nucleotidica sullo splicing. Questo valore, infatti, confronta le predizioni ottenute sulla sequenza di riferimento con quelle generate dopo l'introduzione della variante, valutando possibili modificazioni dei siti donatori e accettori. La track UCSC distingue inoltre le predizioni relative a SNV e indel e rende disponibili informazioni sulla tipologia e sulla posizione dell'evento di splicing potenzialmente interessato.

La stessa tipologia di informazione è stata successivamente acquisita in maniera sistematica durante l'annotazione mediante Ensembl VEP. Per ciascuna variante possono essere restituiti quattro principali delta score: acceptor gain (DS_AG), acceptor loss (DS_AL), donor gain (DS_DG) e donor loss (DS_DL), corrispondenti rispettivamente alla possibile acquisizione o perdita di un sito accettore o donatore. Il massimo dei quattro delta score assume valori compresi tra 0 e 1, con valori progressivamente maggiori associati a una maggiore probabilità predetta che la variante determini un'alterazione dello splicing.

Per l'interpretazione sono stati considerati i cut-off proposti per SpliceAI: uno score pari a 0,2 presenta una soglia caratterizzata da una maggiore sensibilità, 0,5 costituisce il cutoff raccomandato e 0,8 rappresenta una soglia caratterizzata da maggiore precisione. UCSC classifica inoltre valori ≥ 0,2 come indicativi di un impatto predetto sullo splicing,

mentre valori ≤ 0,1 sono riportati come non indicativi di impatto e l'intervallo intermedio come non informativo [38].

La disponibilità di questo valore sia durante l'esplorazione genomica mediante UCSC sia nella successiva annotazione sistematica mediante VEP ha consentito di mantenere l'informazione relativa allo splicing durante i differenti livelli di analisi. Le predizioni di SpliceAI sono state inoltre considerate congiuntamente agli output specifici di AlphaGenome relativi a splice sites, splice-site usage e splice junctions, fornendo così evidenze computazionali complementari sul possibile coinvolgimento delle varianti nei processi di maturazione dell'RNA.

Tra gli indicatori utilizzati per la caratterizzazione e la successiva prioritizzazione delle varianti è stato considerato AVI (AlphaGenome Variant Impact), uno score di impatto sviluppato nell'ambito di AlphaGenome [39,40]. AVI sintetizza le informazioni derivate dalle predizioni sequence-to-function di AlphaGenome allo scopo di fornire una misura quantitativa dell'impatto potenzialmente associato a una singola sostituzione nucleotidica. Nel presente lavoro AVI rappresenta pertanto un punto di collegamento tra le due strategie di analisi basate su AlphaGenome. Da una parte, l'interrogazione diretta del modello mediante la pipeline Python descritta nella sezione 3.8 ha consentito di analizzare separatamente differenti modalità molecolari e di confrontare le predizioni ottenute per gli alleli REF e ALT; dall'altra, lo score AVI ha fornito una misura sintetica dell'impatto complessivamente predetto della variante. L'annotazione AVI è stata inoltre recuperata mediante Ensembl VEP, che integra uno specifico plugin dedicato agli AlphaGenome Variant Impact scores [41].

AVI è stato reso disponibile da Ensembl in una fase avanzata dello svolgimento del presente lavoro e, considerata la sua stretta relazione con AlphaGenome e la potenziale utilità ai fini della prioritizzazione, è stato incorporato nella fase finale dell'analisi. La sua introduzione ha permesso di affiancare alle predizioni multimodali precedentemente generate mediante lo script Python un indicatore sintetico e standardizzato utilizzabile per il confronto tra le diverse varianti candidate.

Lo score è disponibile in due principali rappresentazioni. AVI_RAW costituisce lo score grezzo e rappresenta approssimativamente valori compresi tra -1,3 e +4,0, con valori maggiori associati a una maggiore probabilità di impatto deleterio. AVI_PHRED, restituito di default dal plugin VEP, rappresenta una trasformazione PHRED-scaled

basata sulla posizione della variante nella distribuzione delle possibili sostituzioni genomiche e presenta valori approssimativamente compresi tra 0 e 80.

L'interpretazione della scala PHRED permette di contestualizzare il valore rispetto alla distribuzione genome-wide: ad esempio, un valore AVI_PHRED pari a 20 indica che la variante si colloca approssimativamente nell'1% delle sostituzioni con maggiore impatto deleterio predetto a livello genomico [40]. Valori progressivamente maggiori corrispondono quindi a una posizione più estrema nella distribuzione degli score. AVI non deve tuttavia essere interpretato come una probabilità clinica di patogenicità né come dimostrazione sperimentale dell'effetto delle varianti. La presenza di valori elevati e concordanti di AVI e CADD PHRED è stata considerata di particolare interesse ai fini della prioritizzazione delle varianti, poiché i due score forniscono stime di impatto derivate da approcci differenti. Tra le annotazioni selezionate è stato inoltre incluso GERP, che fornisce una misura del grado di conservazione e di constraint evolutivo della posizione genomica [36]. A differenza di AVI e CADD, GERP non rappresenta direttamente uno score di patogenicità: valori più elevati indicano una maggiore conservazione della posizione, compatibile con una più intensa pressione di selezione purificante durante l'evoluzione. Tale informazione è stata pertanto utilizzata come ulteriore elemento di supporto alla prioritizzazione. Gli score AVI, CADD e GERP sono stati quindi considerati congiuntamente alle ulteriori evidenze funzionali, regolatorie e di splicing ottenute mediante AlphaGenome, VEP e UCSC Genome Browser, al fine di identificare le varianti candidate di maggiore interesse biologico.

3.9.1 Frequenze alleliche e ulteriori annotazioni in VEP

Sono state richieste le informazioni relative alle frequenze alleliche disponibili in gnomAD Exomes v4.1.1, gnomAD Genomes v4.1.1 e NIH All of Us. Tali informazioni sono state utilizzate per contestualizzare la presenza delle varianti nelle popolazioni di riferimento [42]. Una frequenza relativamente elevata nella popolazione generale rappresenta un elemento da considerare con cautela nelle interpretazioni di una variante candidata come evento somatico tumorale raro. Al contrario, l'assenza o la rarità della variante nella popolazione di riferimento può aumentare l'interesse, ma non costituisce una dimostrazione della sua origine somatica.

Sono state inoltre selezionate annotazioni relative a differenti livelli funzionali. UTRAnnotator è stato utilizzato per caratterizzare possibili conseguenze delle varianti localizzate nelle regioni UTR; le informazioni relative al nonsense-mediated decay (NMD) hanno permesso di considerare il possibile coinvolgimento dei meccanismi di degradazione dei trascritti contenenti codoni di stop prematuri; RiboSeqORFs ha fornito informazioni aggiuntive relative a regioni con evidenza di traduzione derivata da dati di ribosome profiling [43].

Per la componente proteica sono stati richiesti Protein ID e ProtVar, comprendendo le informazioni relative alla stabilità della struttura proteica e alle interfacce di interazione proteina-proteina [44]. È stato inoltre selezionato mutfunc, includendo le annotazioni disponibili relative a motivi lineari, interazioni e struttura proteica [45]. Sono stati infine incluse le annotazioni di Gene Ontology, GENCODE promoters e Phenotypes, in modo da integrare ulteriori informazioni funzionali, regolatorie e fenotipiche.

L'assenza di un'annotazione per uno specifico strumento non è stata considerata equivalente all'assenza di un possibile effetto, poiché la disponibilità dei diversi output dipende dalla posizione, dalla conseguenza e dalla tipologia della variante.

3.10 Prioritizzazione funzionale delle varianti somatiche

Al termine delle procedure di annotazione descritte nelle sezioni precedenti, le informazioni ottenute mediante UCSC Genome Browser, AlphaGenome e Ensembl VEP sono state raccolte e organizzate in un'unica tabella di lavoro. Questa fase è stata effettuata con l'obiettivo di ricondurre a una struttura comune informazioni provenienti da strumenti differenti e caratterizzate da natura e significato biologico eterogenei, consentendo di confrontare, per ciascuna variante somatica candidata, le diverse evidenze disponibili. La tabella integrata ha quindi rappresentato il punto di raccordo tra la fase di identificazione delle varianti mediante SComatic e le successive analisi comparative e di prioritizzazione.

Le varianti somatiche precedentemente identificate sono state sottoposte a una procedura di prioritizzazione finalizzata a selezionare gli eventi con maggiore probabilità di determinare conseguenze funzionali. Poiché il dataset comprendeva sia varianti codificanti sia varianti localizzate in regioni introniche e regolatorie, è stato adottato un approccio differenziato in funzione della conseguenza molecolare prevista, evitando l'applicazione di un singolo criterio di patogenicità a tutte le categorie di varianti. Come

parametro primario di prioritizzazione genome-wide è stato utilizzato l'AlphaGenome Variant Impact score (AVI). Lo score AVI è espresso su scala PHRED e può essere applicato anche a varianti non codificanti; un valore pari a 20 identifica approssimativamente l'1% delle sostituzioni nucleotidiche con maggiore impatto predetto a livello genomico. Poiché per AVI non è stata definita una soglia clinica di patogenicità, il valore AVI ≥ 20 è stato utilizzato nel presente studio esclusivamente come soglia operativa di prioritizzazione e non come criterio di classificazione patogenetica. Parallelamente, le conseguenze delle varianti sono state valutate mediante Ensembl Variant Effect Predictor (VEP). Le varianti classificate da VEP come ad alto impatto, quali stop_gained, splice_acceptor_variant, splice_donor_variant e altre eventuali varianti loss-of-function, sono state mantenute indipendentemente dagli altri predittori. Pertanto, il filtro iniziale di selezione è stato definito come AVI ≥ 20 oppure conseguenza VEP ad alto impatto. Il punteggio CADD è stato utilizzato come informazione complementare, ma non come requisito obbligatorio per l'inclusione. Questa scelta è stata adottata per non penalizzare varianti non codificanti o di splicing caratterizzate da un elevato AVI e da evidenze funzionali specifiche, ma da un punteggio CADD inferiore a 20. Per le varianti missenso sono state considerate le predizioni di REVEL, ClinPred e PrimateAI-3D. Ai fini della prioritizzazione sono stati utilizzati i seguenti criteri: REVEL ≥ 0,70, ClinPred ≥ 0,50 e classificazione PrimateAI-3D come *pathogenic*. La soglia di ClinPred pari a 0,50 è coerente con quella utilizzata dagli autori per distinguere predizioni damaging da tolerated, mentre per REVEL è stata adottata una soglia più conservativa pari a 0,70. La predizione di AlphaMissense è stata esaminata come informazione aggiuntiva, ma non è stata utilizzata come criterio indipendente di assegnazione ai Tier, poiché AlphaMissense contribuisce già alla costruzione dello score AVI per le varianti protein-altering.

Il possibile effetto sullo splicing è stato valutato mediante SpliceAI, considerando il massimo delta score tra acceptor gain, acceptor loss, donor gain e donor loss. È stato utilizzato un valore SpliceAI ≥ 0,20 come soglia ad alta sensibilità per identificare varianti potenzialmente in grado di modificare lo splicing. Tale valore corrisponde alla soglia *high-recall* proposta per SpliceAI. Per le varianti non codificanti e sinonime sono state inoltre considerate le predizioni funzionali ottenute mediante AlphaGenome. I diversi output sono stati raggruppati in quattro classi funzionali: alterazioni dell'espressione

dell'RNA; alterazioni della cromatina o dell'accessibilità, comprendenti ChIP, DNase e ATAC; alterazioni dello splicing, comprendenti splice-site usage e splice junctions; alterazioni dell'attività trascrizionale valutata mediante CAGE. La presenza di effetti predetti in almeno due classi molecolari differenti è stata considerata evidenza di supporto funzionale.

Sulla base di tali criteri, le varianti sono state suddivise in quattro categorie di priorità. Il Tier 1A comprendeva le varianti con evidenza funzionale più forte, definite dalla presenza di almeno una delle seguenti condizioni: conseguenza VEP ad alto impatto; SpliceAI ≥ 0,20; oppure, per le varianti missenso, concordanza tra REVEL ≥ 0,70, ClinPred ≥ 0,50 e predizione *pathogenic* di PrimateAI-3D. Il Tier 1B comprendeva le varianti con AVI ≥ 20 che non soddisfacevano i criteri del Tier 1A ma mostravano evidenza convergente: per le missenso erano richiesti almeno due criteri positivi tra REVEL, ClinPred e PrimateAI-3D, mentre per le varianti non codificanti e sinonime erano richiesti effetti AlphaGenome in almeno due differenti classi funzionali. Il Tier 2 comprendeva le restanti varianti con AVI ≥ 20 prive di sufficiente corroborazione da parte degli altri predittori. Infine, le varianti con AVI < 20 e senza conseguenze VEP ad alto impatto sono state assegnate al Tier 3.

4. RISULTATI

4.1 Identificazione delle varianti somatiche candidate

L'applicazione della pipeline di variant calling ai dati di single-nucleus RNA sequencing ha portato all'identificazione di 427 osservazioni paziente-specifiche di varianti somatiche candidate nella popolazione cellulare classificata come *Epithelial (malignant)*. Poiché alcune coordinate genomiche erano condivise da pazienti differenti, tali osservazioni corrispondevano complessivamente a 414 varianti genomiche distinte.

Nelle analisi successive è stata pertanto mantenuta una distinzione esplicita tra variante genomica e osservazione paziente-specifica. Questa scelta ha permesso, da un lato, di riconoscere le alterazioni presenti in più individui e, dall'altro, di conservare l'informazione relativa al campione nel quale ciascuna variante era stata rilevata.

L'annotazione genica ha inoltre individuato sette casi in cui una medesima variante risultava associata a due geni differenti. Di conseguenza, la tabella utilizzata per le successive analisi di annotazione comprendeva 434 associazioni variante-gene, pur derivando dalle 427 osservazioni paziente-specifiche. Le righe aggiuntive non corrispondevano quindi a ulteriori eventi mutazionali, ma riflettevano l'associazione della stessa coordinata genomica con più di una feature genica.

Le varianti candidate erano distribuite tra 30 pazienti, di cui 18 appartenenti al gruppo untreated e 12 al gruppo treated. Complessivamente, nei campioni untreated sono state rilevate 316 varianti candidate, mentre nei campioni treated ne sono state osservate 111.

Il numero di varianti identificate nei singoli pazienti mostrava una marcata eterogeneità (Figura 7). Nel gruppo untreated i valori erano compresi tra 1 e 42 varianti, con una mediana di 13,5. I conteggi più elevati riguardavano U9, con 42 varianti, U11, con 39, U5 e U6, entrambi con 36, e U18.2, con 33. Nel gruppo treated, invece, il numero di candidate variava da 1 a 35 e la mediana era pari a 2,5. I valori maggiori erano quelli dei campioni T5 e T1.2, nei quali sono state identificate rispettivamente 35 e 31 varianti.

Considerando tutti i 43 campioni sottoposti al workflow di variant calling, compresi quelli nei quali nessuna variante aveva superato i criteri di selezione, è emersa una marcata differenza tra i gruppi untreated e treated. Nei 18 campioni untreated la mediana del numero di varianti candidate era pari a 13,5, mentre nei 25 campioni treated era pari a 0 (Mann–Whitney, $p = 2{,}8 \times 10^{-5}$). Tutti i campioni untreated presentavano almeno una

variante candidata, mentre nel gruppo treated ciò avveniva in 12 dei 25 campioni; nei restanti 13 non era stata identificata alcuna variante che avesse superato i filtri applicati. Tale differenza si accompagnava a una marcata diversa rappresentazione della componente epiteliale maligna. Il numero mediano di nuclei classificati come *Epithelial (malignant)* era pari a 3175,5 nei campioni untreated e a 58 nei campioni treated (Mann–Whitney, $p = 1{,}7 \times 10^{-6}$). Anche la proporzione di nuclei maligni rispetto al totale della componente epiteliale risultava inferiore nei campioni treated, con valori mediani rispettivamente pari al 26,5% e al 98,0% nei gruppi treated e untreated ($p = 6{,}5 \times 10^{-4}$).

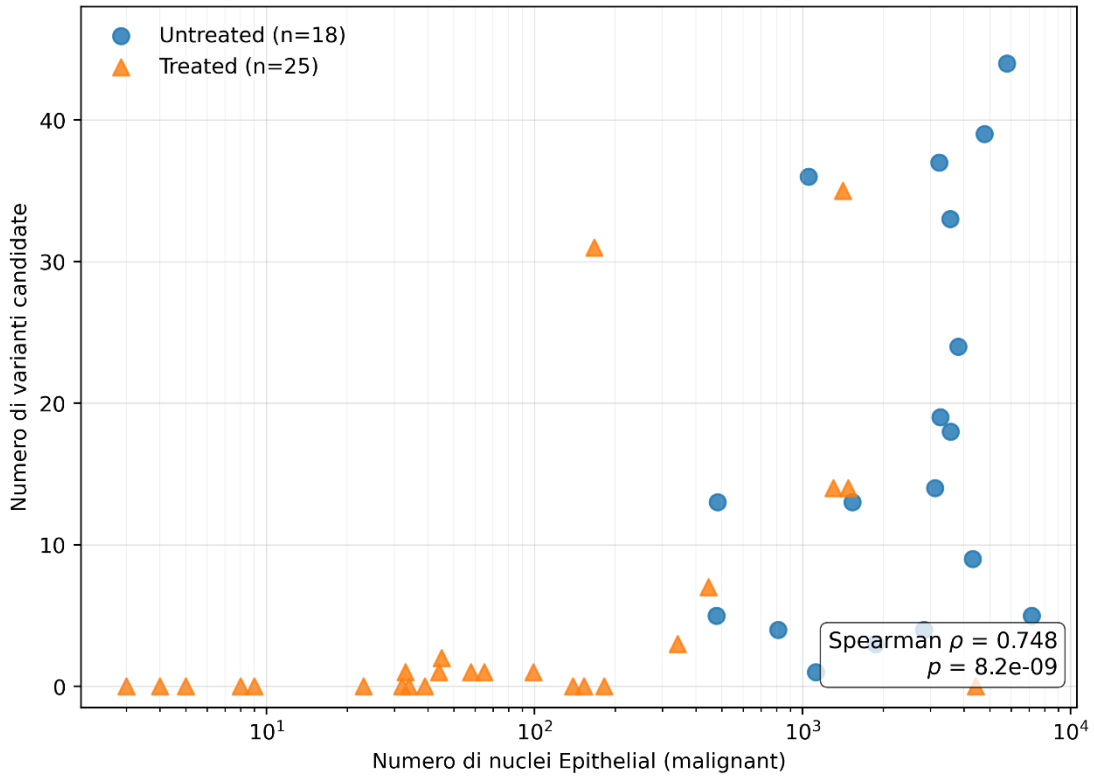


Figura 7. Relazione tra numero di nuclei epiteliali maligni e varianti candidate identificate. Ogni punto rappresenta un paziente; i campioni untreated e treated sono indicati con simboli differenti. Il numero di varianti candidate mostrava una forte correlazione positiva con il numero di nuclei classificati come *Epithelial (malignant)* disponibili per il variant calling (Spearman ρ = 0,748; p = 8,2 × 10^{-9}; n = 43). L'asse X è rappresentato in scala logaritmica.

Considerando l'intera coorte, il numero di varianti candidate identificate mostrava una forte correlazione positiva con il numero di nuclei epiteliali maligni disponibili per l'analisi (Spearman ρ = 0,748; $p = 8{,}2 \times 10^{-9}$) (Figura 7). L'associazione rimaneva significativa anche considerando separatamente i campioni treated (ρ = 0,608; p =

0,0013), indicando che la relazione non dipendeva esclusivamente dalla separazione tra i due gruppi.

In un'analisi esplorativa mediante regressione binomiale negativa, nella quale il numero di varianti era valutato in funzione del numero di nuclei epiteliali maligni e dello stato di trattamento, il numero di nuclei maligni rimaneva significativamente associato al numero di varianti identificate ($p < 0,001$), mentre l'associazione indipendente con lo stato di trattamento non risultava significativa ($p = 0,43$). L'inclusione della chimica 10x v2/v3 nel modello non modificava sostanzialmente questo risultato.

Questi dati indicano quindi una stretta relazione tra la quantità di componente tumorale disponibile nel campione e il numero di varianti rilevabili mediante snRNA-seq.

4.2 Distribuzione cromosomica

Le 427 osservazioni paziente-specifiche erano distribuite su tutti i cromosomi considerati (Figura 8). Il cromosoma 1 presentava il numero più elevato di varianti, con 39 osservazioni, seguito dal cromosoma 11 con 35, dal cromosoma 2 con 29, dal cromosoma 5 con 28 e dal cromosoma 7 con 27. Conteggi relativamente elevati erano presenti anche nei cromosomi 14, 16 e 17, nei quali sono state rilevate rispettivamente 26, 24 e 22 varianti. I cromosomi 19 e 3 comprendevano 20 e 19 osservazioni, mentre sia il cromosoma 8 sia il cromosoma 10 ne presentavano 18. Sul cromosoma X sono state identificate 13 varianti. All'estremo opposto della distribuzione, il cromosoma 22 mostrava il numero più basso di candidate, pari a quattro. Seguivano il cromosoma 18, con sette osservazioni, e i cromosomi 13 e 21, entrambi con otto.

Per tenere conto delle differenti dimensioni cromosomiche, i conteggi sono stati inoltre normalizzati rispetto alla lunghezza dei cromosomi del genoma di riferimento GRCh38 e riportati come numero di osservazioni per 100 Mb. Dopo normalizzazione, i valori più elevati erano osservati nei cromosomi 19, 16, 17, 11, 20 e 14. Questa analisi corregge il semplice effetto della dimensione cromosomica, ma non tiene conto della diversa densità genica né della frazione effettivamente trascritta e coperta nei dati snRNA-seq (Figura 9).

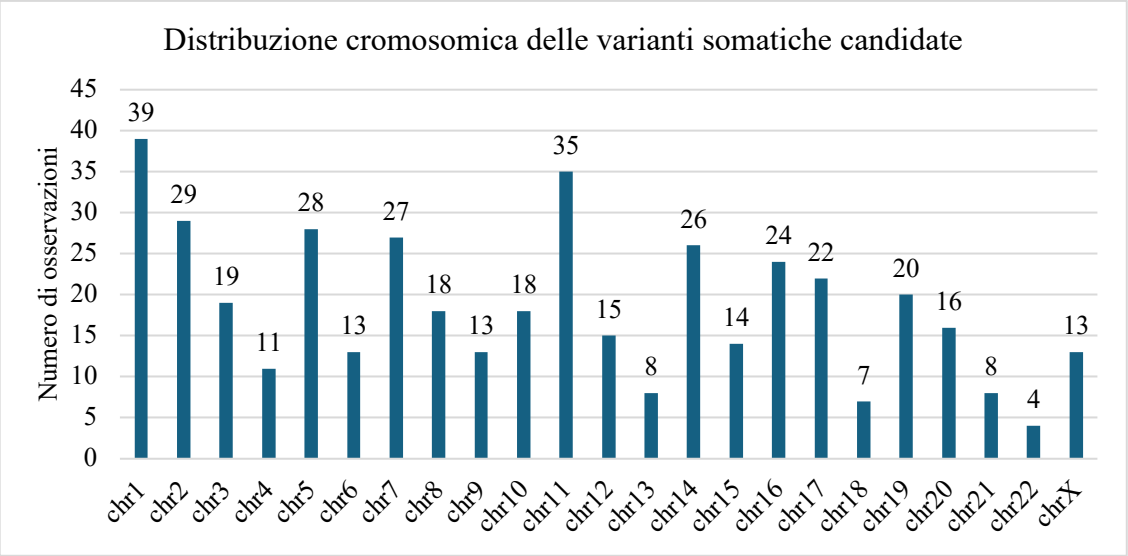


Figura 8. Distribuzione cromosomica delle varianti somatiche candidate. Numero di osservazioni paziente-specifiche di varianti identificate nei diversi cromosomi. I conteggi si riferiscono alle 427 osservazioni complessive e non sono normalizzati rispetto alla lunghezza cromosomica.

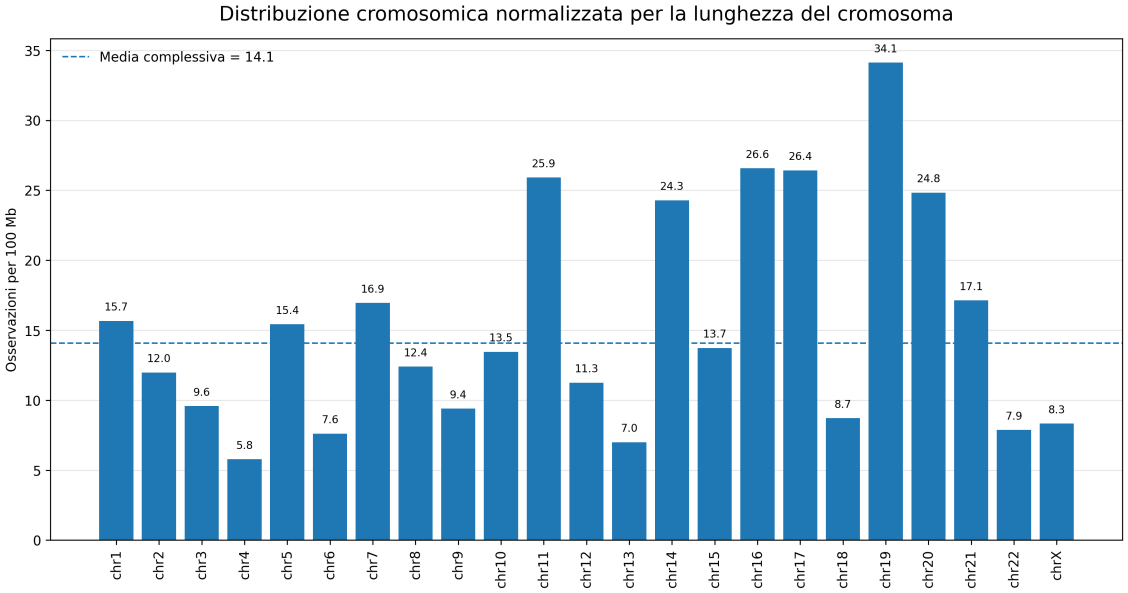


Figura 9. Distribuzione cromosomica normalizzata delle varianti candidate. I conteggi sono stati normalizzati rispetto alla lunghezza dei cromosomi di GRCh38.

4.3 Distribuzione rispetto alle regioni geniche

La classificazione delle varianti sulla base della loro posizione rispetto alla struttura genica ha evidenziato una prevalenza di alterazioni localizzate nelle regioni introniche e non codificanti (Figura 10).

Poiché una stessa variante poteva essere associata a geni differenti e, di conseguenza, assumere una diversa classificazione rispetto alla struttura genica, questa analisi è stata effettuata considerando le 434 associazioni variante-gene presenti nella tabella annotata. Di queste, 177 varianti (40,8%) ricadevano in regioni introniche, che rappresentavano quindi la categoria più numerosa. Le regioni 3' UTR comprendevano 139 osservazioni, pari al 32,0% del totale, mentre 108 varianti (24,9%) erano localizzate in regioni esoniche. Una quota più ridotta interessava la 5' UTR, con sei varianti (1,4%). Infine, quattro osservazioni non erano state assegnate a nessuna delle principali categorie considerate in quanto presentavano una regione genica non definita.

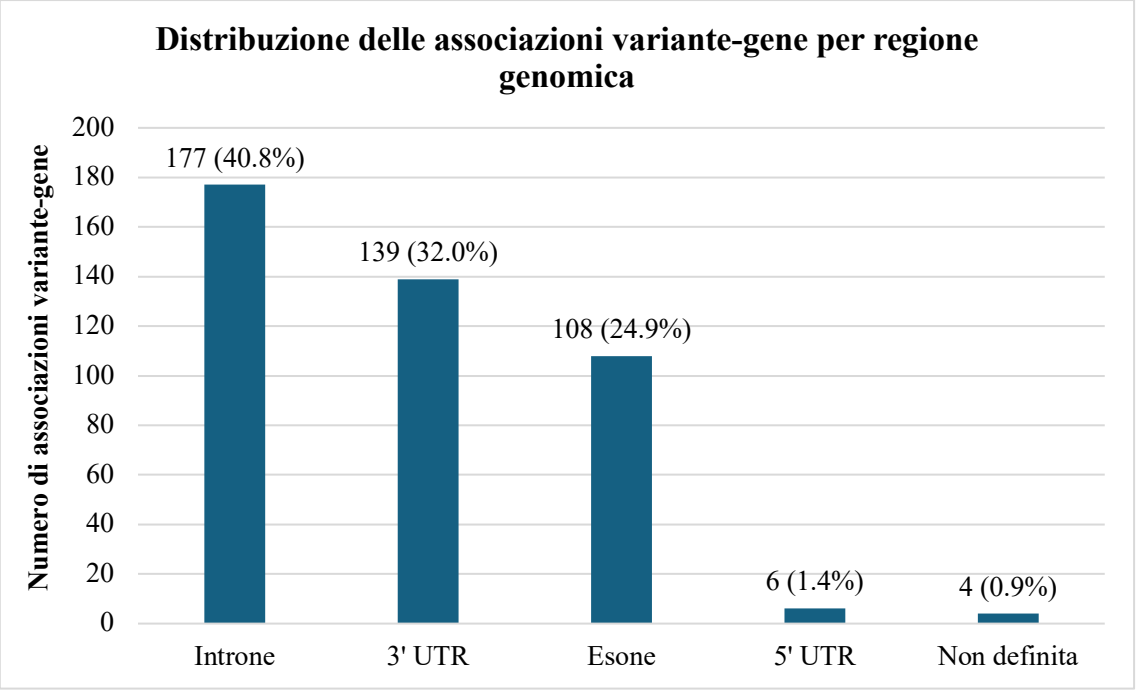


Figura 10. Distribuzione delle associazioni variante-gene rispetto alle regioni geniche. Distribuzione delle 434 associazioni variante-gene in funzione della regione genica annotata. Una stessa variante poteva contribuire a più associazioni quando risultava attribuita a geni differenti.

La composizione del dataset indicava dunque che una parte consistente delle candidate non coinvolgeva direttamente sequenze codificanti. Questo elemento ha assunto particolare rilievo nelle successive fasi di caratterizzazione funzionale, rendendo necessario affiancare ai predittori focalizzati sull'impatto proteico strumenti capaci di valutare anche possibili conseguenze sullo splicing, sull'espressione genica e sui meccanismi di regolazione trascrizionale.

4.4 Distribuzione delle varianti tra i geni

La distribuzione delle varianti candidate a livello genomico ha mostrato un quadro fortemente disperso, poiché la maggior parte dei geni era interessata da una sola variante o da un numero limitato di osservazioni.

Tra quelli maggiormente rappresentati, RBMX presentava quattro osservazioni, tutte provenienti dallo stesso paziente. TMC5, NEAT1, ANXA4, PSMC3 e MUC5AC comparivano invece in tre osservazioni ciascuno. Il significato di questi conteggi non era tuttavia equivalente, poiché variavano sia il numero di coordinate genomiche distinte sia quello dei pazienti coinvolti.

Nel dettaglio, le tre osservazioni relative a TMC5 corrispondevano a due varianti genomiche distinte rilevate complessivamente in tre pazienti. Per NEAT1 erano presenti tre varianti genomiche distribuite tra due pazienti. ANXA4 e PSMC3 comprendevano entrambi tre osservazioni, riconducibili in ciascun caso a due varianti genomiche identificate in due pazienti. Per MUC5AC, al contrario, le tre osservazioni rappresentavano tre varianti genomiche distinte riscontrate nello stesso paziente.

Questi risultati sottolineano l'importanza di mantenere separati tre livelli informativi: il numero di associazioni variante-gene, quello delle varianti genomiche distinte e il numero di pazienti nei quali esse vengono osservate. Di conseguenza, la frequenza con cui un gene compariva nella tabella non è stata considerata, da sola, un criterio sufficiente per stabilirne la priorità.

4.5 Identificazione delle varianti ricorrenti

L'analisi ha permesso di riconoscere 13 varianti genomiche ricorrenti, ognuna delle quali era presente in due pazienti al massimo (Tabella 1).

Le alterazioni ricorrenti interessavano i geni NPM1, ANXA4, ARID4B, EFNA5, NFKB2, ESYT2, PSMC3, PTPRG, CTSH, ZSWIM6, TMC5, MGRN1 e QTRT1. Tra queste, la variante chr10:102402366 A>C di NFKB2 era condivisa dai pazienti U6 e U11; chr11:47418898 G>C di PSMC3 era presente in U3 e U18.2; chr3:62281608 A>T di PTPRG compariva in U3 e U11. La variante chr16:4668266 A>C di MGRN1 era stata rilevata in U6 e U9, mentre chr19:10713023 C>A di QTRT1 risultava condivisa da U11 e U18.2.

Le 13 varianti ricorrenti rappresentavano il 3,1% delle 414 varianti genomiche distinte. Poiché ciascuna era presente in due pazienti, esse corrispondevano complessivamente a 26 delle 427 osservazioni paziente-specifiche, pari al 6,1%.

Nel complesso, gli eventi ricorrenti rappresentavano soltanto una frazione limitata del dataset. Poiché le varianti analizzate erano già state sottoposte al processo di identificazione e filtraggio degli eventi somatici, la ricorrenza inter-paziente è stata mantenuta come un ulteriore elemento di interesse biologico, da integrare con le evidenze relative al possibile impatto funzionale.

Gene	Variante genomica	Pazienti
NPM1	chr5:171392798 T>G	U17.2; U9
ANXA4	chr2:69781662 C>G	U11; U3
ARID4B	chr1:235166987 A>C	U16; U7
EFNA5	chr5:107669847 G>C	U2; U9
NFKB2	chr10:102402366 A>C	U11; U6
ESYT2	chr7:158753845 A>C	T5; U18.2
PSMC3	chr11:47418898 G>C	U18.2; U3
PTPRG	chr3:62281608 A>T	U11; U3
CTSH	chr15:78921878 T>C	U16; U3
ZSWIM6	chr5:61375547 T>C	U11; U3
TMC5	chr16:19449613 T>G	T4; U3
MGRN1	chr16:4668266 A>C	U6; U9
QTRT1	chr19:10713023 C>A	U11; U18.2

Tabella 1. Varianti genomiche ricorrenti identificate in due pazienti differenti.

4.6 Conseguenze molecolari predette mediante VEP

L'annotazione effettuata mediante Variant Effect Predictor (VEP) ha consentito di definire in modo più dettagliato le possibili conseguenze molecolari associate alle varianti candidate (Figura 11). Considerando le consequence VEP associate alle varianti, le categorie maggiormente rappresentate erano intron_variant, seguita da 3_prime_UTR_variant e missense_variant. In particolare, 77 varianti erano annotate

come missense_variant; a queste si aggiungevano due osservazioni classificate come missense_variant&splice_region_variant.

Sono state inoltre rilevate 17 synonymous_variant, due osservazioni splice_region_variant&synonymous_variant, una splice_acceptor_variant e una stop_gained. Il dataset comprendeva anche consequence relative alla 5' UTR, a regioni upstream e downstream e a trascritti non codificanti.

Per alcune delle varianti associate a più geni, la consequence dipendeva dal gene preso in considerazione. Un esempio è rappresentato dalla variante U6 chr5:112891872 A>G, classificata come intronica rispetto a REEP5 ma associata a una consequence comprendente la 3' UTR e una regione downstream nel caso di SRP19. Analogamente, la variante U15 chr19:45474782 A>C interessava la 3' UTR di FOSB, mentre risultava intronica rispetto a ERCC1. Le consequence gene-specifiche non sono state pertanto considerate come categorie necessariamente mutuamente esclusive quando riferite alla medesima coordinata genomica.

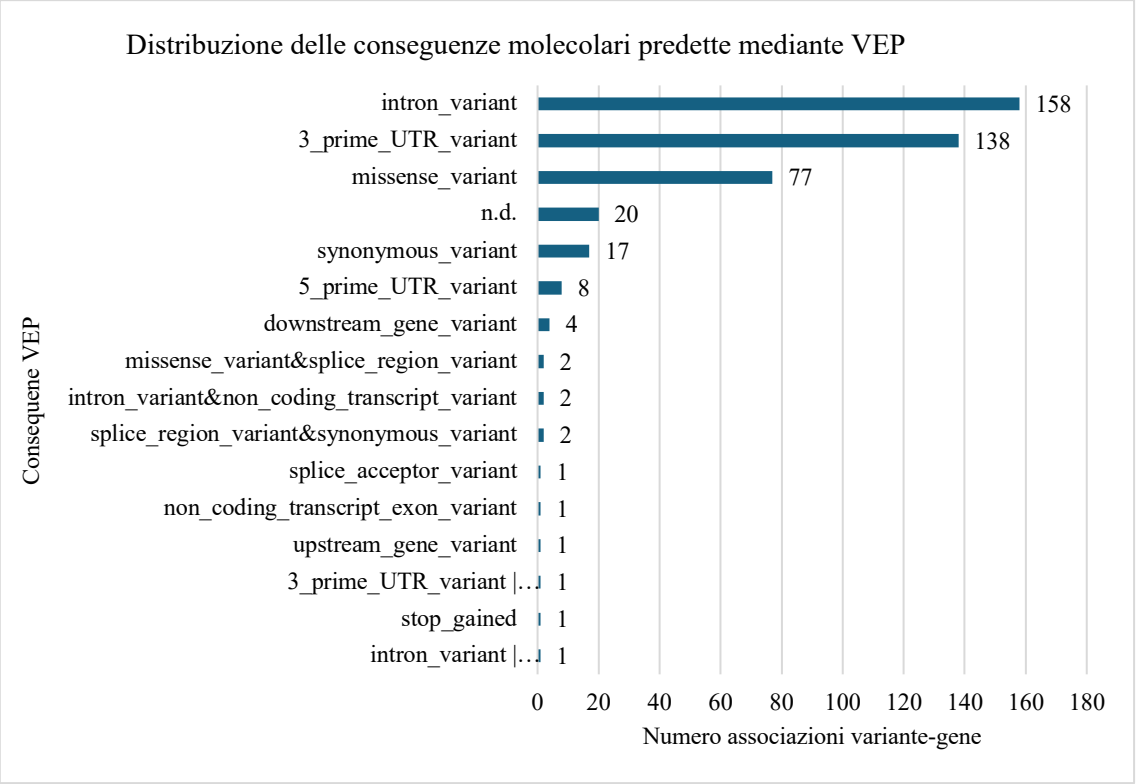


Figura 11. Distribuzione delle conseguenze molecolari delle varianti identificate mediante Ensembl VEP.

Due alterazioni mostravano consequence di particolare interesse. La variante chr6:75713807 T>G di SENP6, rilevata nel paziente U1, era classificata come stop_gained. Nel paziente T1.2, invece, la variante chr9:128153310 G>T di LCN2 risultava annotata come splice_acceptor_variant.

Le consequence ottenute mediante VEP sono state mantenute come un livello informativo distinto rispetto agli score generali di deleteriousness e ai predittori sviluppati specificamente per le sostituzioni missenso. Questi strumenti, infatti, non presentano la stessa applicabilità per tutte le categorie di variante.

4.7 Score globali di impatto: AVI e CADD

Gli score AVI e CADD sono stati impiegati per ottenere una valutazione quantitativa iniziale del potenziale impatto delle varianti (Figura 12).

Considerando le 434 associazioni variante-gene, 69 varianti mostravano un valore AVI ≥20; tra queste, 26 raggiungevano almeno il valore 25 e 6 presentavano valori pari o superiori a 30. Per CADD, 62 varianti avevano uno score ≥20, 28 raggiungevano almeno 25 e 7 almeno 30.

Poiché la successiva prioritizzazione è stata effettuata a livello di variante genomica, i conteggi sono stati successivamente ricondotti alle 414 coordinate genomiche distinte. In questo caso, 62 varianti presentavano AVI ≥20, 23 AVI ≥25 e cinque AVI ≥30. Per CADD, 56 varianti genomiche distinte presentavano valori ≥20, 24 valori ≥25 e sette valori ≥30.

L'intersezione tra i due score comprendeva 49 varianti genomiche distinte caratterizzate contemporaneamente da AVI ≥20 e CADD ≥20. Tredici varianti con AVI ≥20 presentavano invece CADD <20, mentre sette varianti con CADD ≥20 mostravano AVI <20. Questi risultati hanno indicato che l'utilizzo congiunto dei due score come filtro obbligatorio avrebbe prodotto una selezione più restrittiva, ma avrebbe escluso una parte delle candidate identificate mediante AVI. Per questo motivo, CADD è stato mantenuto come informazione complementare e non utilizzato come requisito necessario per la successiva prioritizzazione.

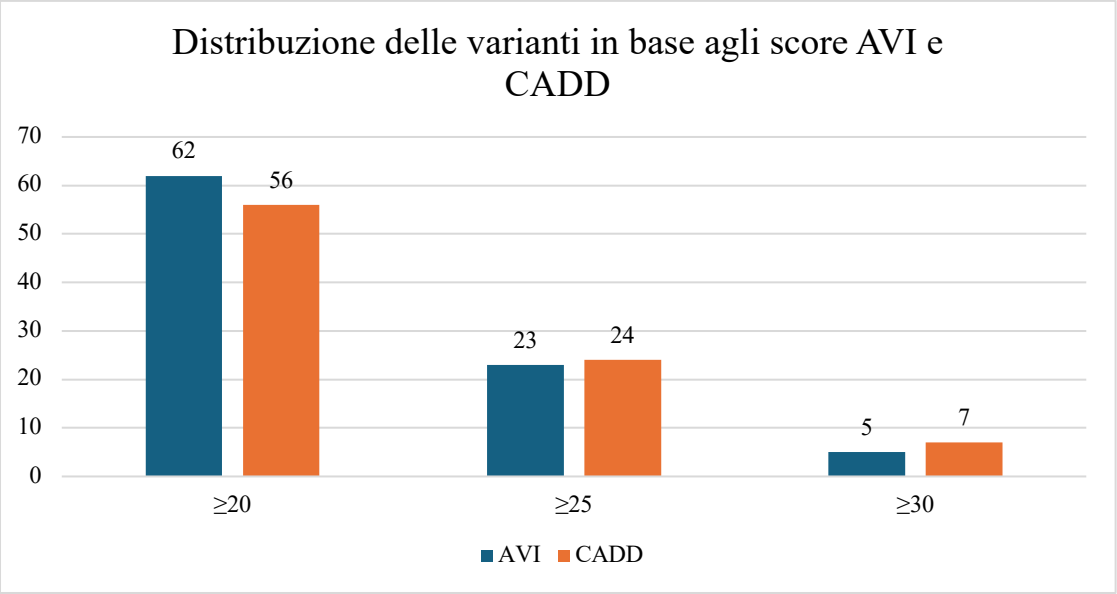


Figura 12. Distribuzione delle varianti in base agli score AVI e CADD.

Un caso particolarmente evidente era quello di SENP6, caratterizzato da AVI pari a 35,4 e CADD pari a 38, oltre alla consequence stop_gained.

Tra le varianti missenso con valori elevati per entrambi gli score figuravano invece QTRT1, PNPLA8, CHD2, SRSF1, KIF13B, CNOT2, LMAN1, SUPT20H, PSMC6, PSMC3 e BPTF.

La scelta di non applicare AVI e CADD come criteri rigidi di esclusione si è rivelata particolarmente importante per candidate come LCN2, caratterizzata da AVI pari a 26 ma CADD pari a 15,25, la cui evidenza funzionale principale derivava dall'analisi dello splicing piuttosto che da score globali particolarmente elevati.

4.8 Predizione dell'impatto delle varianti missenso

Le varianti missenso sono state ulteriormente caratterizzate mediante AlphaMissense [30], REVEL [32], ClinPred [33] e PrimateAI-3D [46], così da confrontare predizioni ottenute attraverso modelli differenti (Tabella 2).

AlphaMissense forniva un'annotazione per 80 varianti. In base alle categorie restituite dal modello, 30 erano classificate come *likely pathogenic*, 13 come ambigue e 37 come *likely benign*. REVEL era invece disponibile per 82 varianti: 24 presentavano uno score ≥0,50 e, tra queste, 17 raggiungevano o superavano la soglia di 0,75.

Per ClinPred erano disponibili 76 annotazioni. Venticinque varianti mostravano valori inferiori a 0,50, cinque erano comprese nell'intervallo 0,50–0,75 e sette tra 0,75 e 0,90. Le restanti 39 varianti presentavano uno score ≥0,90. PrimateAI-3D forniva infine una classificazione per 70 varianti, di cui 22 indicate dal modello come *pathogenic* e 48 come *benign*. Tali categorie sono state considerate esclusivamente come output predittivi e non come classificazioni cliniche di patogenicità.

Il confronto tra i diversi strumenti ha evidenziato un sottogruppo di varianti caratterizzato da una marcata concordanza delle predizioni. La variante associata a QTRT1, per esempio, presentava AlphaMissense pari a 0,980, REVEL 0,833 e ClinPred 0,9954. Valori analogamente elevati erano osservati per CNOT2 (AlphaMissense 0,9966; REVEL 0,875; ClinPred 0,9945), KIF13B (0,9981; 0,915; 0,9991), PSMC6 (0,9685; 0,921; 0,9993) e BPTF (1,000; 0,933; 0,9998). Una concordanza elevata tra più predittori era presente anche per PNPLA8, CHD2, SRSF1, LMAN1, SUPT20H e PSMC3.

Non tutte le candidate, tuttavia, mostravano predizioni concordanti. Nei casi di discordanza, le differenze tra gli strumenti sono state conservate come parte dell'informazione disponibile, evitando di sintetizzarle attraverso un punteggio medio che avrebbe potuto mascherare la diversa natura dei modelli utilizzati.

Predittore	N. annotate	Distribuzione
AlphaMissense	80	30 likely pathogenic; 13 ambiguous; 37 likely benign
REVEL	82	24 ≥0,50; di cui 17 ≥0,75
ClinPred	76	25 <0,50; 5 0,50–<0,75; 7 0,75–<0,90; 39 ≥0,90
PrimateAI-3D	70	22 pathogenic; 48 benign

Tabella 2. Risultati dei predittori di impatto applicati alle varianti missenso.

4.9 Analisi mediante popEVE

L'analisi mediante popEVE ha fornito un ulteriore livello di caratterizzazione delle varianti missenso (Figura 13). Considerando le 69 osservazioni per le quali era disponibile uno score, 61 varianti presentavano valori superiori alla soglia definita per la fascia moderatamente deleteria, 4 rientravano nella categoria moderatamente deleteria e 4 nella fascia severamente deleteria [47].

Le quattro osservazioni appartenenti alla classe severamente deleteria corrispondevano tuttavia a tre varianti genomiche distinte, poiché la variante di NFKB2 era stata identificata in due pazienti. Le tre varianti distinte appartenenti a questa categoria interessavano quindi SRSF1 con un valore pari a −5,736, NFKB2 con −5,505 e CNOT2 con −5,340. Invece, PSMC6, con un valore di −4,948, apparteneva alla fascia moderatamente deleteria.

L'informazione fornita da popEVE si è rivelata particolarmente utile nei casi caratterizzati da predizioni non uniformi. NFKB2 ne costituisce un esempio: pur non mostrando un forte supporto da parte di AlphaMissense e PrimateAI-3D, la variante presentava un valore popEVE appartenente alla fascia severamente deleteria, insieme ad altre evidenze indipendenti.

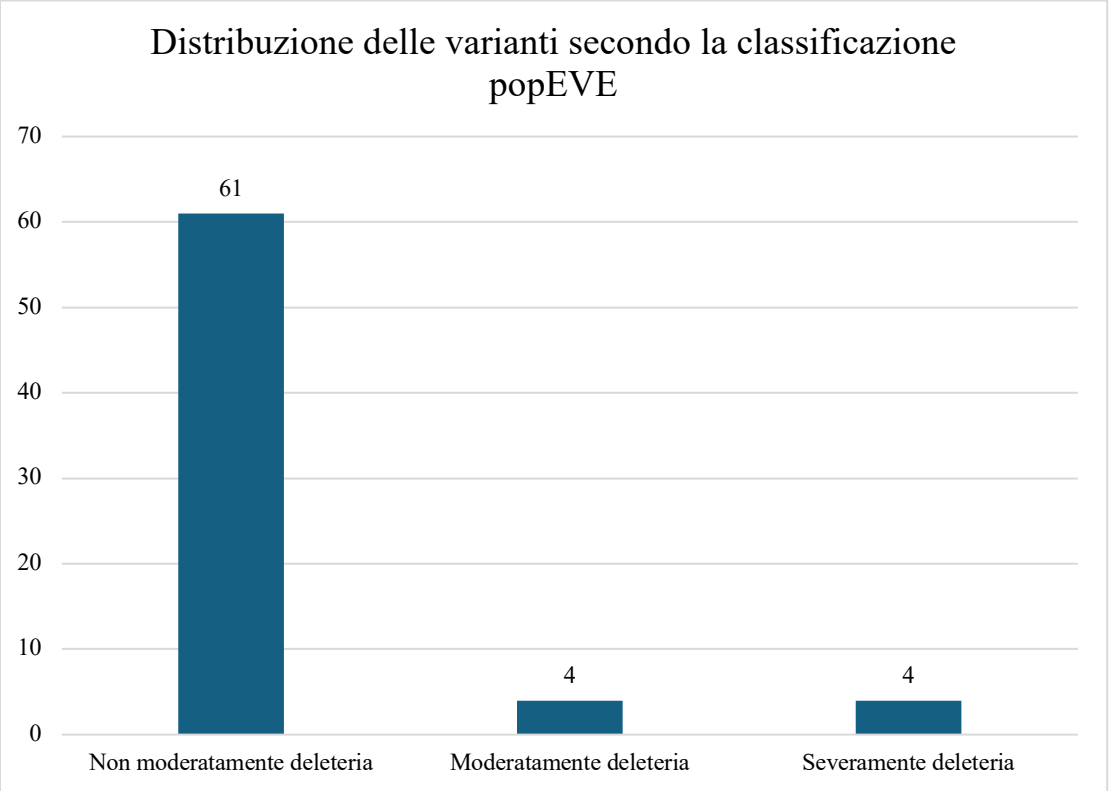


Figura 13. Distribuzione delle varianti secondo la classificazione popEVE.

4.10 Analisi dello splicing

L'integrazione delle annotazioni dedicate allo splicing ha consentito di individuare un sottogruppo di varianti supportate, con diverso grado di concordanza, da strumenti indipendenti (Tabella 3).

Tra i casi più informativi, la variante associata a LCN2 presentava una predizione SpliceAI di *acceptor gain* pari a 0,42. Per RPF2 era invece riportato un *acceptor gain* di 0,40, mentre SENP6 mostrava una predizione di *donor loss* pari a 0,22.

Questi risultati sono stati confrontati con le predizioni di AlphaGenome relative a due modalità distinte dello splicing, *splice-site usage* e *splice junctions*. LCN2 mostrava evidenza in entrambe, mentre RPF2 presentava un segnale nelle *splice junctions*. Per SENP6, al contrario, non era stata rilevata una corrispondente evidenza AlphaGenome nelle due modalità considerate.

Gene	SpliceAI	Score	AlphaGenome splice-site usage	AlphaGenome splice junctions
LCN2	acceptor gain	**0,42**	Sì	Sì
RPF2	acceptor gain	**0,40**	No*	Sì
SENP6	donor loss	**0,22**	No	No

Tabella 3. Confronto delle principali predizioni relative allo splicing.

Il profilo di LCN2 mostrava quindi una convergenza particolarmente articolata (Figura 14). VEP classificava la variante come splice_acceptor_variant, SpliceAI indicava un *acceptor gain* di 0,42 e AlphaGenome forniva evidenza sia per *splice-site usage* sia per *splice junctions*. All'annotazione era inoltre associato un segnale di predizione di splicing aberrante.

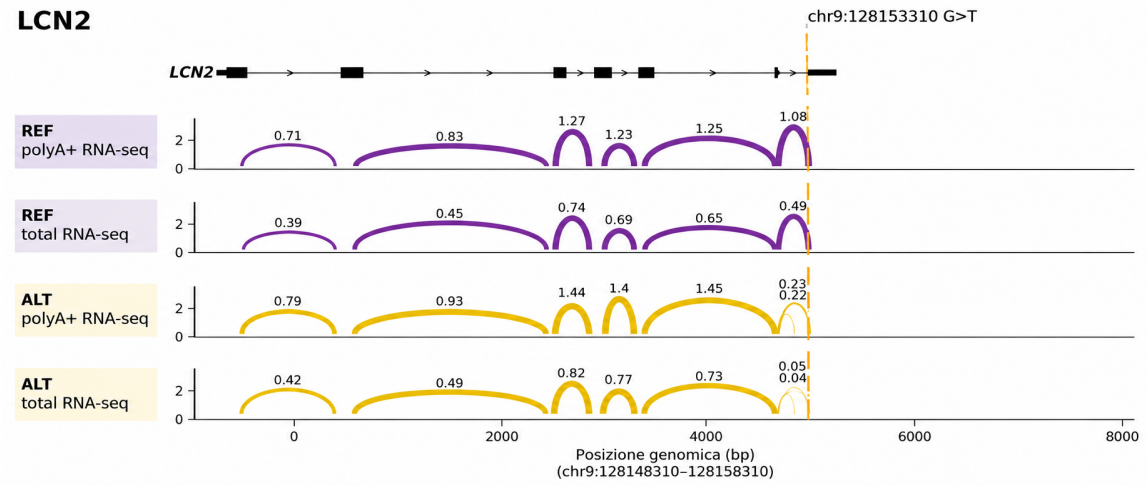


Figura 14. Predizioni AlphaGenome relative all'effetto della variante chr9:128153310 G>T di LCN2 sulle splice junctions. Sono riportate le predizioni ottenute per l'allele di riferimento (REF) e per l'allele alternativo (ALT), considerando gli output *polyA plus RNA-seq* e *total RNA-seq*. Gli archi rappresentano le diverse giunzioni di splicing predette, mentre i valori riportati in corrispondenza di ciascun arco indicano il valore predetto da AlphaGenome per la relativa giunzione nei diversi output considerati. Il confronto tra REF e ALT evidenzia variazioni delle predizioni in corrispondenza di più giunzioni, incluse quelle prossime alla posizione della variante, indicata dalla linea verticale. La visualizzazione è stata generata mediante lo script sviluppato per l'elaborazione e la rappresentazione grafica degli output di AlphaGenome.

Complessivamente, 17 varianti genomiche distinte presentavano evidenza concomitante nelle predizioni AlphaGenome di *splice-site usage* e *splice junctions*. Le varianti interessavano PLSCR1, CDK13, VMP1, NFKB2, MEIS1, STK35, ANKRD36C, ITGA10, TAF1D, MKNK2, VILL, SRRM1, COQ8B, LCN2, PARP2, COX6C e YTHDC2.

Ulteriori profili di interesse riguardavano MGRN1, la cui consequence VEP era missense_variant&splice_region_variant, e PARP2, classificato come splice_region_variant&synonymous_variant e contemporaneamente supportato dalle predizioni AlphaGenome sia per *splice-site usage* sia per *splice junctions*.

Nel complesso, l'analisi ha mostrato come la valutazione di un possibile effetto sullo splicing non possa essere ricondotta a un singolo algoritmo. Il confronto tra consequence molecolare, SpliceAI e AlphaGenome ha infatti consentito di evidenziare candidate che avrebbero potuto ricevere minore attenzione utilizzando esclusivamente predittori orientati all'impatto proteico.

4.11 Predizioni regolatorie mediante AlphaGenome

L'impiego di AlphaGenome ha esteso la caratterizzazione funzionale oltre le sole conseguenze sulla sequenza codificante, permettendo di esaminare possibili effetti delle varianti su differenti output molecolari.

Considerando le predizioni relative all'RNA del gene associato alla variante, 137 delle 427 osservazioni paziente-specifiche, pari al 32,1%, presentavano una predizione AlphaGenome relativa all'espressione del gene associato. Tra queste, 85 erano associate a una predizione UP (sovra-espressione) e 52 a una predizione DOWN (sotto-espressione), equivalenti rispettivamente al 62,0% e al 38,0% delle osservazioni con evidenza RNA.

La direzione prevista è stata conservata come informazione funzionale senza attribuirle automaticamente un significato clinico.

Tra le candidate successivamente selezionate per l'analisi integrata, QTRT1 e NFKB2 presentavano una predizione RNA UP, mentre per PSMC3, PTPRG e RPF2 era previsto un effetto RNA DOWN (Figure 15, 16).

AlphaGenome ha quindi aggiunto un ulteriore livello interpretativo, particolarmente utile per le varianti localizzate al di fuori delle regioni codificanti e per quelle candidate nelle quali i predittori dell'impatto proteico fornivano risultati discordanti.

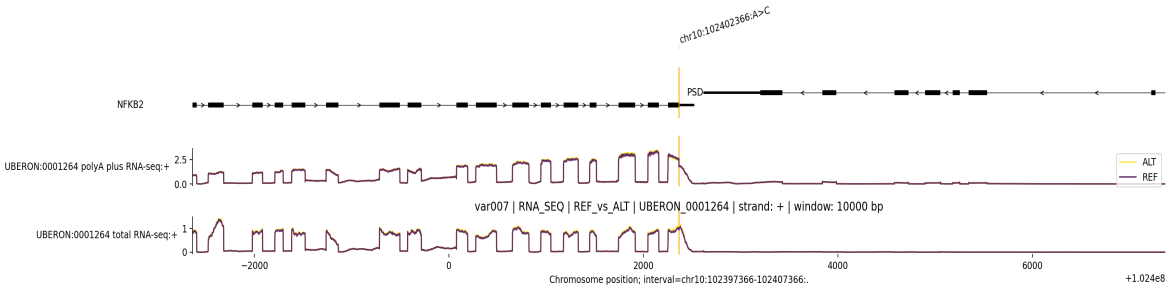


Figura 15. Predizione AlphaGenome dell'effetto della variante chr10:102402366 A>C associata a NFKB2 sui segnali RNA-seq. Sono mostrati i profili predetti da AlphaGenome per l'allele di riferimento (REF) e per l'allele alternativo (ALT) in una finestra genomica di 10 kb, per gli output *polyA plus RNA-seq* e *total RNA-seq*. La linea verticale indica la posizione della variante. Il confronto REF–ALT consente di visualizzare le modificazioni locali del segnale RNA predetto associate alla sostituzione, classificata nell'analisi come RNA UP.

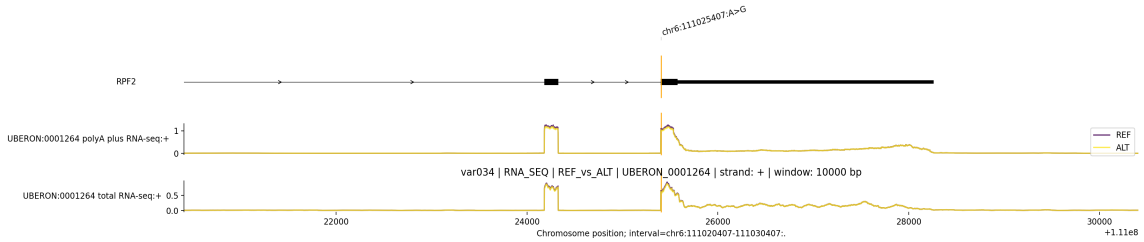


Figura 16. Predizione AlphaGenome dell'effetto della variante chr6:111025407 A>G associata a RPF2 sui segnali RNA-seq. Sono mostrati i profili predetti da AlphaGenome per l'allele di riferimento (REF) e per l'allele alternativo (ALT) in una finestra genomica di 10 kb, per gli output *polyA plus RNA-seq* e *total RNA-seq*. La linea verticale indica la posizione della variante. Il confronto REF–ALT evidenzia le modificazioni locali del segnale RNA predetto associate alla sostituzione, classificata nell'analisi come RNA DOWN.

4.12 Annotazioni in database di varianti note

La consultazione dei database di varianti già catalogate ha mostrato che 89 varianti paziente-specifiche possedevano un identificativo dbSNP. La presenza di un rsID è stata interpretata esclusivamente come indicazione della precedente catalogazione della variante. Nel dataset erano inoltre riportati otto identificativi ClinVar. Tuttavia, un'interpretazione clinica esplicita era disponibile soltanto per tre varianti, associate ai geni PPFIBP1, SYNE2 e NADSYN1. In tutti e tre i casi, ClinVar riportava una classificazione come variant of uncertain significance (VUS). Le informazioni cliniche disponibili risultavano quindi limitate rispetto all'ampiezza del dataset e non erano sufficienti, da sole, a definire la priorità della maggior parte delle candidate. Ciò ha ulteriormente motivato l'integrazione con predizioni di natura funzionale, proteica, regolatoria e di splicing (Tabella 4).

Tipo di annotazione/evidenza	N. osservazioni	Risultato
AlphaGenome – RNA	137/427	85 UP (62,0%); 52 DOWN (38,0%)
dbSNP	89/427	Presenza di identificativo rsID
ClinVar – identificativo	8	Identificativo ClinVar presente
ClinVar – interpretazione clinica esplicita	3	PPFIBP1, SYNE2 e NADSYN1: VUS

Tabella 4. Sintesi delle principali annotazioni ed evidenze disponibili per le varianti analizzate.

4.13 Prioritizzazione integrata delle varianti somatiche candidate

Nell'ultima fase dell'analisi, le informazioni ottenute mediante i diversi strumenti di annotazione e predizione sono state integrate al fine di selezionare un insieme ristretto di varianti somatiche candidate da sottoporre a successivo approfondimento biologico.

Non è stato costruito un punteggio complessivo derivato dalla semplice somma dei diversi predittori. Una strategia di questo tipo avrebbe infatti assegnato un peso maggiore alle varianti missenso, per le quali erano disponibili numerosi algoritmi parzialmente correlati, penalizzando al tempo stesso alterazioni in zone non codificanti o associate a splicing e trascrizione. Le diverse tipologie di evidenza sono state pertanto mantenute separate e valutate in funzione della consequence e del possibile meccanismo molecolare associato alla variante.

Come primo livello di selezione è stato utilizzato uno score AVI ≥20, mantenendo inoltre indipendentemente eventuali varianti classificate da VEP come ad alto impatto. Nel dataset analizzato, le due varianti caratterizzate da consequence VEP ad alto impatto, stop_gained e splice_acceptor_variant, presentavano entrambe AVI ≥20 e risultavano quindi già comprese nella selezione basata su AVI. Complessivamente, 62 delle 414 varianti genomiche distinte superavano il criterio iniziale, corrispondenti al 15,0% delle varianti identificate.

Le varianti selezionate sono state successivamente suddivise in differenti livelli di priorità sulla base della convergenza delle evidenze disponibili. Il Tier 1A comprendeva le varianti caratterizzate dall'evidenza funzionale più forte, definita dalla presenza di una consequence VEP ad alto impatto, da uno score SpliceAI ≥0,20 oppure, nel caso delle varianti missenso, dalla concordanza tra REVEL, ClinPred e PrimateAI-3D secondo le soglie prestabilite. Il Tier 1B comprendeva varianti con AVI ≥20 supportate da evidenze convergenti ma meno stringenti: per le varianti missenso era richiesta la positività di almeno due dei tre predittori considerati, mentre per le varianti sinonime o non codificanti era richiesta la presenza di effetti AlphaGenome riconducibili ad almeno due differenti classi funzionali. Le rimanenti varianti con AVI ≥20 sono state assegnate al Tier 2, mentre quelle che non soddisfacevano il criterio iniziale sono state classificate come Tier 3.

L'applicazione di questi criteri ha identificato 15 varianti Tier 1A, 22 Tier 1B e 25 Tier 2. Le restanti 352 varianti sono state assegnate al Tier 3. Complessivamente, quindi, 37

varianti appartenenti ai Tier 1A e 1B sono state considerate ad alta priorità funzionale. Esse rappresentavano l'8,9% delle 414 varianti genomiche distinte iniziali e il 59,7% delle 62 varianti che avevano superato il primo livello di selezione (Figura 17).

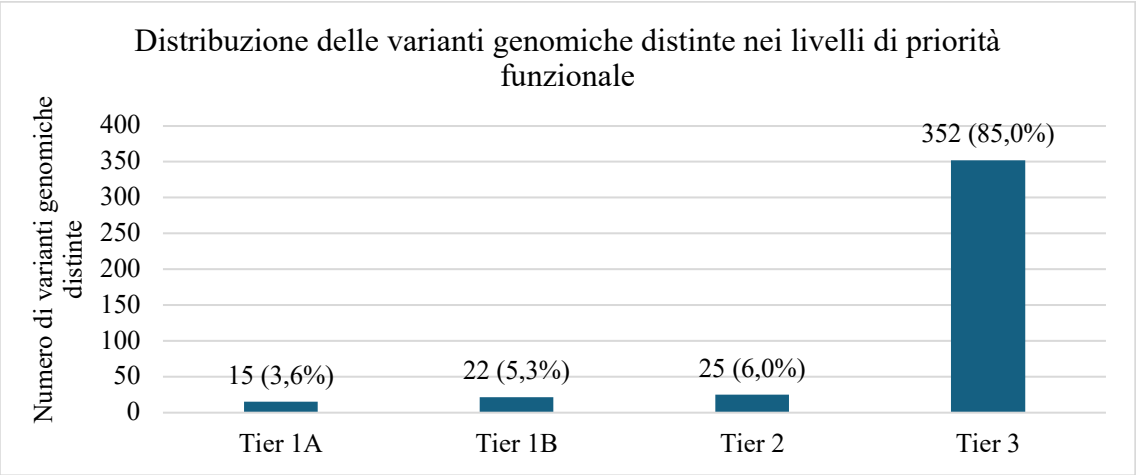


Figura 17. Distribuzione delle varianti genomiche distinte secondo i livelli di priorità funzionale.

Il Tier 1A comprendeva varianti associate ai geni SENP6, QTRT1, PNPLA8, CHD2, SRSF1, KIF13B, CNOT2, LMAN1, SUPT20H, PSMC6, LCN2, PTP4A2, PSMC3, BPTF e RPF2. Sebbene appartenessero alla stessa classe di priorità, tali varianti emergevano attraverso modalità differenti di convergenza delle evidenze.

La variante chr6:75713807 T>G di SENP6 mostrava il profilo complessivamente più marcato. Presentava infatti il valore AVI più elevato dell'intero dataset, pari a 35,4, associato a CADD 38 e alla consequence VEP stop_gained. A queste evidenze si aggiungeva una predizione SpliceAI di *donor loss* pari a 0,22. In questo caso, la prioritizzazione era quindi sostenuta principalmente dalla previsione di una perdita di funzione a livello della sequenza codificante, ulteriormente supportata dagli score globali di impatto.

Un diverso profilo caratterizzava le varianti missenso supportate dalla concordanza dei predittori dell'impatto proteico. Tra queste, QTRT1 chr19:10713023 C>A presentava AVI 32,7, CADD 26,5, REVEL 0,833 e ClinPred 0,9954, oltre alla classificazione *pathogenic* di PrimateAI-3D. La stessa variante era stata inoltre identificata in due pazienti differenti. Una concordanza analoga tra i predittori caratterizzava, con differenti combinazioni di

score, anche le varianti associate a PNPLA8, CHD2, SRSF1, KIF13B, CNOT2, LMAN1, SUPT20H, PSMC6, PTP4A2, PSMC3 e BPTF.

Per alcune candidate, invece, la principale evidenza di impatto derivava dalla possibile alterazione dello splicing. La variante chr9:128153310 G>T di LCN2 presentava AVI pari a 26 ma CADD pari a 15,25. VEP la classificava tuttavia come splice_acceptor_variant, mentre SpliceAI indicava un *acceptor gain* pari a 0,42. A ciò si aggiungeva l'evidenza AlphaGenome sia per *splice-site usage* sia per *splice junctions*. LCN2 rappresentava pertanto un esempio particolarmente evidente di variante che sarebbe stata esclusa adottando come criterio obbligatorio la contemporanea presenza di AVI ≥20 e CADD ≥20, nonostante la convergenza di differenti predizioni specificamente orientate verso un possibile effetto sullo splicing.

Un profilo analogo, ma meno direttamente evidente dalla consequence VEP, era osservato per RPF2 chr6:111025407 A>G. Sebbene la variante fosse annotata come missenso e i predittori focalizzati sull'alterazione aminoacidica fornissero un supporto relativamente contenuto, SpliceAI mostrava un *acceptor gain* pari a 0,40. La sua assegnazione al Tier 1A derivava pertanto dal possibile impatto sullo splicing piuttosto che dalla sola conseguenza sulla sequenza proteica.

Il Tier 1B comprendeva 22 varianti caratterizzate da un grado di supporto inferiore rispetto al Tier 1A, ma comunque sostenute dalla convergenza di più evidenze. Undici erano varianti missenso per le quali almeno due tra REVEL, ClinPred e PrimateAI-3D fornivano una predizione compatibile con un potenziale effetto funzionale. Queste varianti interessavano ZMYND8, ACAP2, RPL36AL, RNF31, PSMC3, IFT172, ITCH, PTPRG, EIF5, CFLAR e PLSCR1.

Le altre undici varianti Tier 1B erano sinonime o non codificanti e risultavano supportate da predizioni AlphaGenome distribuite su almeno due differenti classi funzionali. Le candidate interessavano SMARCE1, BRF2/ADGRA2, PPP3CB, STK35, YTHDC2, TACC1, INO80C, SAT1, USP25, SPINT2 e MEIS1.

Questo gruppo risultava particolarmente rilevante per valutare l'utilità di un approccio non limitato alle regioni codificanti. La variante chr8:37843857 G>A, associata a BRF2 e ADGRA2 e localizzata nella 3′UTR, presentava ad esempio AVI 23,2 ma CADD 15,56. AlphaGenome indicava contemporaneamente possibili effetti sull'RNA, su parametri associati alla cromatina e sulle *splice junctions*. Analogamente, diverse varianti

localizzate nelle regioni UTR, tra cui quelle associate a STK35, TACC1, SAT1, USP25, SPINT2 e MEIS1, mostravano predizioni distribuite su più output molecolari pur non essendo valutabili attraverso i principali algoritmi sviluppati per le sostituzioni aminoacidiche.

Le 25 varianti assegnate al Tier 2 presentavano AVI ≥20 ma non raggiungevano il livello di convergenza richiesto per l'inclusione nei Tier 1A o 1B. La classificazione nel Tier 2 non implicava pertanto l'assenza di un possibile effetto funzionale, ma indicava un livello di supporto computazionale inferiore secondo i criteri prestabiliti.

Alcune di queste varianti presentavano comunque profili meritevoli di attenzione. Un esempio era rappresentato da NFKB2, variante ricorrente in due pazienti e caratterizzata da una predizione AlphaGenome RNA UP, da evidenza sia per *splice-site usage* sia per *splice junctions* e da un valore popEVE appartenente alla fascia severamente deleteria. Al contrario, AlphaMissense e PrimateAI-3D non suggerivano un marcato effetto proteico. NFKB2 evidenziava quindi come l'applicazione di criteri formali di prioritizzazione possa mantenere separate situazioni caratterizzate da forte concordanza all'interno di uno specifico meccanismo da profili più complessi e multidimensionali, nei quali le diverse fonti di evidenza non risultano completamente concordanti.

Anche la ricorrenza inter-paziente è stata considerata come informazione aggiuntiva e non come criterio sufficiente per l'assegnazione ai livelli di priorità. Tra le varianti comprese nelle categorie di maggiore interesse, QTRT1, PSMC3, PTPRG, oltre a varianti quali MGRN1 e NFKB2, erano state rilevate in due pazienti differenti. La loro ricorrenza è stata pertanto interpretata nel contesto delle ulteriori evidenze funzionali disponibili.

Nel complesso, la procedura di prioritizzazione ha ridotto le 414 varianti genomiche distinte iniziali a 37 candidate ad alta priorità funzionale, appartenenti ai Tier 1A e 1B. La selezione finale comprendeva profili eterogenei: varianti associate a potenziale perdita di funzione, alterazioni dello splicing, sostituzioni missenso supportate da predittori concordanti e varianti non codificanti caratterizzate da effetti regolatori multimodali.

La classificazione in Tier non è stata interpretata come una classificazione clinica di patogenicità, ma come uno strumento operativo di prioritizzazione. L'approccio adottato ha consentito di integrare differenti meccanismi molecolari evitando che il maggior numero di strumenti disponibili per le varianti missenso determinasse una selezione sistematicamente sbilanciata verso le regioni codificanti. Le varianti appartenenti ai livelli

di priorità più elevati sono state quindi considerate il gruppo principale di candidate da approfondire successivamente sulla base della funzione dei geni coinvolti e delle evidenze disponibili in letteratura.

Le 15 varianti appartenenti al Tier 1A, considerate quelle con maggiore evidenza funzionale, sono riportate in Tabella 5.

Gene	Variante	Conseguenza VEP	AVI	CADD	Principale criterio di prioritizzazione
SENP6	chr6:75713807 T>G	stop_gained	35,4	38,0	VEP high-impact; SpliceAI donor loss 0,22
QTRT1	chr19:10713023 C>A	missense	32,7	26,5	REVEL/ClinPred/PrimateAI-3D concordanti
PNPLA8	chr7:108479297 C>T	missense	30,6	28,4	REVEL/ClinPred/PrimateAI-3D concordanti
CHD2	chr15:92979208 T>G	missense	29,4	27,0	REVEL/ClinPred/PrimateAI-3D concordanti
SRSF1	chr17:58005970 A>G	missense	28,7	32,0	REVEL/ClinPred/PrimateAI-3D concordanti
KIF13B	chr8:29188527 A>G	missense	28,3	28,9	REVEL/ClinPred/PrimateAI-3D concordanti
CNOT2	chr12:70346307 T>G	missense	28,3	33,0	REVEL/ClinPred/PrimateAI-3D concordanti
LMAN1	chr18:59353221 T>C	missense	27,5	32,0	REVEL/ClinPred/PrimateAI-3D concordanti
SUPT20H	chr13:37044097 T>G	missense	27,5	31,0	REVEL/ClinPred/PrimateAI-3D concordanti
PSMC6	chr14:52724019 A>T	missense	27,2	32,0	REVEL/ClinPred/PrimateAI-3D concordanti
LCN2	chr9:128153310 G>T	splice_acceptor	26,0	15,25	VEP high-impact; SpliceAI 0,42
PTP4A2	chr1:31910057 C>A	missense	24,9	26,8	REVEL/ClinPred/PrimateAI-3D concordanti
PSMC3	chr11:47418890 T>C	missense	24,9	24,9	REVEL/ClinPred/PrimateAI-3D concordanti
BPTF	chr17:67854505 T>G	missense	24,6	23,8	REVEL/ClinPred/PrimateAI-3D concordanti
RPF2	chr6:111025407 A>G	missense	23,5	24,2	SpliceAI acceptor gain 0,40

Tabella 5. Varianti appartenenti al Tier 1A e principali evidenze utilizzate per la prioritizzazione funzionale.

5. DISCUSSIONE

5.1 Significato complessivo dei risultati

Il presente lavoro si proponeva di valutare la fattibilità dell'identificazione di varianti somatiche a partire da dati di single-nucleus RNA sequencing generati mediante tecnologia 10x Genomics Chromium in PDAC. Il principale vantaggio dell'approccio del snRNA-seq riguarda l'identificazione dei tipi cellulari. Il barcode cellulare permette di conservare l'associazione tra il segnale della variante e la popolazione cellulare nella quale esso viene rilevato. In un tumore fortemente eterogeneo come il PDAC, questa informazione aggiunge un livello che il normale sequenziamento bulk non fornisce direttamente. Il valore del metodo consiste quindi soprattutto nell'utilizzare il dato trascrittomico per formulare ipotesi mutazionali riferite a specifici compartimenti cellulari, pur mantenendo chiari i limiti imposti dalla natura dell'RNA sequencing.

La pipeline adottata ha portato all'identificazione di 427 osservazioni paziente-specifiche, corrispondenti a 414 varianti genomiche distinte, attribuite alla popolazione delle cellule epiteliali tumorali. Questo risultato dimostra innanzitutto che dati prodotti originariamente per finalità trascrittomiche possono contenere un ulteriore livello informativo, cioè gli stessi dati possono essere sfruttati anche per recuperare informazioni sul carico mutazionale. Tale informazione, tuttavia, non è equivalente a quella ottenibile mediante sequenziamento del DNA genomico, ma rappresenta un livello complementare che conserva il collegamento tra variante osservata e identità cellulare. Infatti, la possibilità di rilevare una variante mediante snRNA-seq dipende dalla trascrizione del locus genico, dalla presenza del trascritto nel nucleo, dalla posizione della variante rispetto alle regioni effettivamente sequenziate e dalla sufficiente profondità di copertura. Nel caso dei dati utilizzati in questa tesi si aggiungono due caratteristiche tecniche. Da un lato, il snRNA-seq include una quota rilevante di trascritti non ancora completamente processati, con conseguente presenza di reads introniche; dall'altro, il protocollo 10x Genomics 3′ Gene Expression produce una copertura non uniforme lungo il trascritto, maggiormente concentrata verso l'estremità 3′. La distribuzione delle varianti candidate riflette probabilmente anche queste proprietà del dato. Infatti, una quota rilevante delle varianti identificate ricadeva in regioni introniche e nelle UTR, mentre solo una parte interessava direttamente la sequenza codificante. Non sarebbe quindi corretto concludere che tale distribuzione rappresenti la frequenza relativa delle diverse classi di mutazioni

nel genoma del PDAC. Essa descrive piuttosto la frazione delle alterazioni che risulta osservabile nelle condizioni imposte dal dato snRNA-seq.

Lo stesso principio è utile per interpretare la mancata identificazione, tra le principali candidate, dei geni driver più frequentemente alterati nel PDAC quali KRAS, TP53, CDKN2A e SMAD4 [1,2,4,8–10]. In generale, la mancata rilevazione di una specifica mutazione in un dataset snRNA-seq non costituisce evidenza della sua assenza dal tumore. Infatti, una variante può non essere osservata perché la specifica porzione del trascritto non è adeguatamente coperta dalle reads disponibili, soprattutto in una libreria con forte bias 3′, oppure perché appartenente ad un gene poco espresso. Inoltre, alcuni dei principali geni alterati nel PDAC possono essere inattivati mediante delezioni o variazioni del numero di copie, eventi che non rientrano nel tipo di varianti considerate in questo studio.

Un aspetto interessante emerso dall'analisi riguarda il minor numero di varianti candidate osservato nei campioni sottoposti a trattamento neoadiuvante. Questa differenza si accompagnava a una marcata riduzione della componente epiteliale maligna nei campioni treated e il numero di varianti identificate risultava fortemente associato al numero di nuclei tumorali disponibili per il variant calling. Nei dati snRNA-seq, infatti, la probabilità di rilevare una variante dipende anche dal numero di cellule che possono contribuire all'osservazione del locus e dalla quantità di reads disponibili. Quando la componente tumorale è scarsamente rappresentata, diminuisce quindi la probabilità di rilevare una variante, soprattutto se presente soltanto in una sottopopolazione delle cellule maligne. L'analisi multivariata è risultata coerente con questa interpretazione: una volta considerato il numero di nuclei epiteliali maligni, lo stato treated/untreated non mostrava più un'associazione indipendente con il numero di varianti identificate. In altre parole, le cellule tumorali residue dopo trattamento non possiedono un numero inferiore di mutazioni, ma la differenza osservata deriva dalla minore quantità di componente neoplastica disponibile nei campioni treated. Questa osservazione evidenzia anche un limite generale del variant calling da snRNA-seq: il numero di varianti recuperate non dipende esclusivamente dal genoma delle cellule tumorali, ma anche dalla loro rappresentazione nel campione. Di conseguenza, il conteggio delle varianti non può essere utilizzato direttamente come misura del carico mutazionale né confrontato tra campioni senza considerare la quantità di materiale tumorale effettivamente disponibile.

5.2 Prioritizzazione delle varianti in funzione del possibile meccanismo

Dopo il variant calling, la difficoltà principale era rappresentata dall'eterogeneità delle alterazioni ottenute. I risultati comprendevano varianti missenso, sinonime, introniche, localizzate nelle UTR, varianti associate allo splicing e una variante *stop_gained*. L'applicazione dello stesso criterio a tutte queste categorie avrebbe inevitabilmente favorito le varianti per le quali esiste un maggior numero di algoritmi predittivi, in particolare le missenso.

Per questo motivo non è stato costruito un punteggio complessivo ottenuto sommando i risultati dei diversi strumenti. La prioritizzazione è stata invece adattata al tipo di conseguenza prevista. Le varianti con possibile effetto sulla proteina sono state valutate soprattutto mediante predittori dedicati alle sostituzioni missenso; per quelle associate allo splicing sono stati considerati VEP, SpliceAI e gli output AlphaGenome specifici; per le varianti non codificanti sono state utilizzate soprattutto le predizioni regolatorie. Questa impostazione ha permesso di mantenere nella selezione anche alterazioni che sarebbero state penalizzate da un filtro costruito prevalentemente sulle caratteristiche delle varianti codificanti.

L'utilizzo di AVI ≥20 come primo criterio operativo ha ridotto considerevolmente il numero di candidate, mentre le consequence VEP ad alto impatto sono state mantenute indipendentemente dagli altri score. I successivi criteri hanno portato alla classificazione delle 414 varianti genomiche distinte in 15 Tier 1A, 22 Tier 1B, 25 Tier 2 e 352 Tier 3. Le 37 varianti comprese nei primi due livelli rappresentano quindi un sottoinsieme relativamente ristretto rispetto alle candidate iniziali.

Questi Tier non devono essere letti come categorie cliniche. Una variante Tier 1A non può essere definita patogena sulla sola base della classificazione ottenuta, così come una variante Tier 2 non può essere considerata priva di effetto biologico. I Tier indicano esclusivamente il grado di priorità assegnato secondo i criteri utilizzati in questa analisi.

Anche la concordanza tra strumenti differenti richiede cautela. Molti algoritmi di predizione non sono completamente indipendenti, perché utilizzano informazioni parzialmente sovrapposte. Lo stesso vale per il rapporto tra AVI e alcune componenti AlphaGenome. Per questo motivo, nelle varianti non codificanti la presenza di più output

AlphaGenome concordanti è utile soprattutto perché suggerisce quale meccanismo potrebbe essere interessato.

Questa distinzione è importante per comprendere il senso della prioritizzazione. L'obiettivo non era individuare il maggior numero possibile di strumenti "positivi", ma selezionare le varianti per le quali fosse possibile formulare un'ipotesi funzionale sufficientemente definita.

5.3 Differenti profili funzionali tra le varianti prioritarie

Le varianti a maggiore priorità sono state individuate per ragioni diverse. Alcune emergono principalmente per la consequence molecolare prevista, altre per la concordanza dei predittori missenso, altre ancora per un possibile effetto sullo splicing.

La variante di SENP6 rappresenta il caso più evidente di una possibile alterazione della sequenza codificante. La sostituzione chr6:75713807 T>G presenta AVI pari a 35,4, il valore più elevato dell'intero dataset, CADD pari a 38 ed è classificata da VEP come *stop_gained*. È presente, inoltre, una predizione SpliceAI di *donor loss* pari a 0,22. In questo caso la consequence VEP è più informativa dei predittori sviluppati specificamente per le varianti missenso. La presenza di un codone di stop prematuro suggerisce un possibile effetto importante sulla proteina, ma non permette di concludere automaticamente che SENP6 venga completamente inattivato. Sarà necessario stabilire sperimentalmente quale trascritto sia interessato, se venga attivato il *nonsense-mediated decay* e se possa essere prodotta una proteina tronca stabile. La variante di SENP6 rappresenta quindi una candidata ad alta priorità soprattutto perché l'ipotesi funzionale è chiara e direttamente verificabile.

Un secondo tipo di profilo è rappresentato dalle varianti per le quali l'elemento principale è un possibile effetto sullo splicing. Il caso più convincente è quello di LCN2. La variante chr9:128153310 G>T è classificata come *splice_acceptor_variant* da VEP, mentre SpliceAI riporta un *acceptor gain* pari a 0,42. AlphaGenome indica inoltre effetti sia sullo *splice-site usage* sia sulle *splice junctions*. L'interesse di questa candidata è ulteriormente rafforzato dal fatto che LCN2 è già stato studiato nel carcinoma pancreatico e associato a fenomeni quali invasività, crescita tumorale e resistenza alla gemcitabina [48]. Questo collegamento non costituisce però una validazione della variante, perché riguardano due livelli di evidenza distinti. I risultati della tesi suggeriscono che la specifica sostituzione

possa modificare lo splicing di LCN2; gli studi precedenti dimostrano invece che la funzione o l'espressione del gene possono avere conseguenze biologiche nel PDAC.

Anche RPF2 emerge soprattutto per un possibile effetto sul processamento dell'RNA. La variante è annotata come missenso, ma AlphaMissense e REVEL non forniscono un supporto particolarmente elevato. SpliceAI indica invece un *acceptor gain* pari a 0,40; AlphaGenome segnala un effetto sulle *splice junctions* e una predizione RNA DOWN. In questo caso è quindi plausibile che la consequence più immediatamente visibile, cioè la sostituzione aminoacidica, non rappresenti necessariamente l'effetto funzionale più importante della variante.

Altre candidate emergono principalmente per la concordanza dei predittori relativi all'impatto proteico. SRSF1, CNOT2 e BPTF sono tra gli esempi più evidenti. Per tutte e tre le varianti, i principali predittori missenso forniscono valori elevati; inoltre, per questi geni sono disponibili evidenze sperimentali, con diverso grado di specificità, nel contesto del carcinoma pancreatico.

SRSF1 è particolarmente interessante perché il gene è direttamente coinvolto nello splicing e il suo aumento di espressione è stato associato alla carcinogenesi pancreatica KRAS^G12D-dipendente [49]. CNOT2 appartiene al complesso CCR4-NOT, coinvolto nel controllo della stabilità degli RNA, mentre BPTF partecipa al rimodellamento della cromatina. In tutti questi casi, tuttavia, la letteratura riguarda prevalentemente la funzione del gene o la sua modulazione sperimentale. Non permette quindi di attribuire automaticamente alla variante identificata lo stesso effetto descritto per la sovraespressione o il silenziamento del gene.

Un caso diverso è NFKB2. La variante chr10:102402366 A>C è formalmente classificata nel Tier 2. Il suo profilo rimane tuttavia interessante perché mostra chiaramente i limiti di una classificazione basata su un'unica classe di predittori. AlphaMissense e PrimateAI-3D suggeriscono un impatto proteico modesto, mentre popEVE ricade nella fascia severamente deleteria. La variante è inoltre presente in due pazienti e AlphaGenome predice un effetto RNA UP insieme a segnali relativi a *splice-site usage* e *splice junctions*. Il caso di NFKB2 non dimostra che la variante sia funzionalmente rilevante, ma mostra perché può essere utile conservare l'informazione completa anche quando gli algoritmi non concordano. In situazioni di questo tipo la discordanza non deve essere eliminata

mediante una media degli score, ma interpretata alla luce delle differenze tra ciò che i singoli modelli cercano di predire.

5.4 Processi biologici rappresentati tra le candidate

Pur non essendo stata effettuata un'analisi formale di enrichment (considerato il numero relativamente contenuto di geni associati alle varianti Tier 1), i geni interessati dalle varianti prioritarie mostrano alcune convergenze che meritano di essere considerate.

Una prima riguarda il processamento e l'omeostasi dell'RNA. SRSF1 è un regolatore dello splicing; CNOT2 appartiene al complesso CCR4-NOT e partecipa al controllo della stabilità dei trascritti; RPF2 interviene nella biogenesi ribosomiale; QTRT1 è coinvolto nella modificazione dei tRNA. Nella lista più ampia compaiono inoltre altri geni associati al metabolismo o al processamento dell'RNA. Anche se non è possibile, sulla base di questi dati, concludere che nel campione sia alterata una specifica "via dell'RNA", perché il numero degli eventi è limitato, rimane interessante che più varianti indipendenti coinvolgano processi che convergono sulla produzione, maturazione e utilizzo dell'RNA.

Un secondo gruppo riguarda la proteostasi. PSMC3 e PSMC6 codificano subunità ATPasiche del proteasoma 26S, mentre MGRN1 è una E3 ubiquitin ligasi. La presenza di più geni associati al sistema ubiquitina-proteasoma non è sufficiente per dimostrare un'alterazione coordinata di questa via, ma suggerisce un possibile tema da verificare in dataset indipendenti.

Un terzo gruppo interessa la regolazione della cromatina. BPTF e CHD2 partecipano direttamente al rimodellamento cromatinico; nella shortlist estesa sono presenti anche geni quali SMARCE1 e ZMYND8.

Queste convergenze possono essere considerate soprattutto come indicazioni per studi successivi. In una coorte più ampia potrebbe essere più informativo osservare mutazioni diverse in geni appartenenti allo stesso processo biologico piuttosto che attendersi necessariamente la ricorrenza della stessa identica variante in numerosi pazienti.

5.5 Utilità e limiti delle predizioni AlphaGenome

AlphaGenome ha ampliato la caratterizzazione delle varianti permettendo di valutare possibili effetti non limitati alla sequenza proteica. Nel dataset, 137 osservazioni paziente-specifiche presentavano una predizione relativa all'espressione del gene associato; per

altre varianti erano disponibili informazioni relative a cromatina, attività trascrizionale o splicing.

L'aspetto più utile di AlphaGenome è probabilmente la possibilità di rendere più specifica l'ipotesi funzionale. Per una variante non codificante, sapere che AVI è elevato suggerisce che possa esistere un impatto funzionale; osservare che le differenze REF/ALT riguardano prevalentemente splicing, RNA o accessibilità cromatinica permette di formulare un'ipotesi più precisa da sottoporre successivamente a verifica.

Va inoltre ricordato che AVI e gli output AlphaGenome non rappresentano livelli completamente indipendenti di evidenza. È più corretto considerare gli output specifici come informazioni che aiutano a definire il possibile meccanismo sottostante allo score globale.

Nel presente lavoro le predizioni di AlphaGenome sono state contestualizzate, quando possibile, utilizzando il termine ontologico relativo al pancreas. Questo rende l'analisi più pertinente rispetto all'utilizzo di tracce provenienti da tessuti diversi o di un valore globale medio, ma permane un limite importante: non è possibile selezionare in AlphaGenome il tessuto tumorale, ma solo pancreas sano. Lo stato trascrizionale ed epigenetico della cellula tumorale può differire notevolmente da quello dei tessuti di riferimento utilizzati dal modello. Le predizioni AlphaGenome devono quindi essere considerate ipotesi funzionali contestualizzate a livello tissutale, non una ricostruzione dello stato molecolare specifico delle cellule tumorali del singolo paziente.

5.6 Ricorrenza delle varianti tra pazienti

Nel dataset sono state identificate 13 varianti genomiche presenti in due pazienti differenti al massimo. La ricorrenza interessa quindi una quota limitata delle varianti e, da sola, non può essere utilizzata come criterio per definire una variante come driver.

Nel caso di dati derivati da RNA, anche la probabilità di osservare più volte la stessa variante dipende dalla copertura del locus. Una posizione appartenente a un trascritto abbondante e ben rappresentato ha maggiori possibilità di essere interrogata in più pazienti rispetto a una regione scarsamente coperta. Per questo motivo la ricorrenza è stata considerata come informazione aggiuntiva, e non come criterio sufficiente di prioritizzazione.

QTRT1 rappresenta il caso nel quale questo elemento si associa al profilo computazionale più convincente. La variante chr19:10713023 C>A è presente in U11 e U18.2 e mostra

contemporaneamente AVI e CADD elevati e una forte concordanza dei predittori missenso. In questo caso la ricorrenza aggiunge quindi un elemento a una variante già prioritaria sulla base di altre caratteristiche.

Anche PSMC3 è interessante, ma richiede una distinzione importante. Nel dataset sono presenti due differenti varianti associate allo stesso gene: chr11:47418890 T>C, compresa nel Tier 1A, e chr11:47418898 G>C, osservata in U3 e U18.2. È quest'ultima a essere ricorrente e a presentare il profilo descritto nella discussione, comprendente una predizione RNA DOWN e la discordanza di PrimateAI-3D.

PTPRG, MGRN1 e NFKB2 costituiscono ulteriori esempi nei quali la ricorrenza si associa ad altri elementi di interesse. Anche in questi casi, tuttavia, la presenza in due pazienti deve essere considerata un motivo per approfondire la variante e non una dimostrazione della sua selezione durante la carcinogenesi.

5.7 Limiti dello studio

Il limite principale deriva dalla natura del materiale analizzato. Il variant calling è stato effettuato a partire da RNA e non da DNA, per cui il dataset contiene esclusivamente le alterazioni che possono essere osservate nei trascritti sufficientemente rappresentati. La mancata rilevazione di una variante non equivale alla sua assenza dal genoma tumorale. Questo limita anche la possibilità di confrontare direttamente il numero di varianti tra pazienti o di utilizzare tale numero come misura del carico mutazionale.

Un secondo limite riguarda la definizione della natura somatica delle varianti. SComatic è stato sviluppato specificamente per ridurre il contributo di varianti germinali, RNA editing e artefatti sfruttando anche il confronto tra differenti popolazioni cellulari. Ma ciò non sostituisce il confronto con il sequenziamento del DNA. In particolare, una variante non osservata nelle popolazioni non maligne non può essere considerata certamente assente da tali cellule: il locus potrebbe semplicemente non essere sufficientemente coperto. Per questo motivo è opportuno mantenere la definizione di varianti somatiche candidate.

Anche i parametri VAF e CCF devono essere interpretati nel contesto del dato RNA. Questi parametri descrivono il supporto della variante rispettivamente a livello delle reads e delle cellule informative, ma non rappresentano direttamente la frazione clonale genomica. Espressione allele-specifica, dropout, diversa quantità di trascritto e

degradazione selettiva di un allele possono modificare la relazione tra genotipo e segnale osservato nell'RNA. Nel caso di una variante nonsense, per esempio, l'attivazione del *nonsense-mediated decay* potrebbe ridurre proprio il trascritto che contiene l'allele mutato.

Un ulteriore limite riguarda la caratterizzazione funzionale. AVI, CADD, AlphaMissense, REVEL, ClinPred, PrimateAI-3D, popEVE, SpliceAI e AlphaGenome producono predizioni computazionali. Anche quando più strumenti concordano, il risultato non dimostra che la variante abbia realmente l'effetto previsto. Inoltre, alcuni modelli condividono fonti informative o caratteristiche utilizzate per la predizione e non possono quindi essere considerati repliche indipendenti.

Un altro aspetto riguarda il rapporto tra variante e gene. La presenza di letteratura che associa un determinato gene al PDAC non dimostra che la specifica variante identificata abbia un ruolo nella malattia. Questo è particolarmente evidente per geni quali LCN2, SRSF1 e BPTF, per i quali sono disponibili dati sperimentali sulla funzione del gene ma non sulla sostituzione osservata in questa tesi.

La numerosità della coorte e la dipendenza dalla copertura trascrittomica impediscono infine di utilizzare le frequenze osservate per stimare la prevalenza delle singole varianti nel PDAC. Anche le alterazioni ricorrenti dovranno essere verificate in dataset indipendenti prima di attribuire loro un significato generale.

5.8 Prospettive future

Il dataset potrebbe essere ulteriormente sfruttato analizzando le varianti a livello di singola cellula. Per le posizioni sufficientemente coperte, sarebbe possibile distinguere cellule con supporto per l'allele ALT da cellule con supporto per l'allele REF, mantenendo separate come non informative quelle nelle quali il locus non è coperto. Il confronto ALT vs REF potrebbe permettere di verificare se la presenza della variante si associa a differenze nell'espressione del gene interessato o a specifici programmi trascrizionali.

Questa analisi potrebbe essere adattata al possibile meccanismo della variante. Per le candidate con predizione AlphaGenome RNA UP/DOWN si potrebbe valutare se le cellule ALT mostrino variazioni coerenti dell'espressione; per SENP6 si potrebbe esplorare una possibile riduzione del trascritto compatibile con *nonsense-mediated decay*; per LCN2 e RPF2 si potrebbero riesaminare le reads locali alla ricerca di differenze nello splicing. Per geni come NFKB2, invece, potrebbe essere più informativo verificare se le

cellule ALT mostrino una diversa attività di specifici programmi molecolari, ad esempio della via NF-κB.

Un ulteriore sviluppo potrebbe consistere nella ricostruzione di possibili sottocloni tumorali integrando più varianti contemporaneamente. La co-occorrenza di differenti alleli ALT nelle stesse cellule, eventualmente insieme a informazioni di copy number inferite dal trascrittoma, potrebbe permettere di identificare gruppi cellulari geneticamente distinti e di confrontarne i profili trascrizionali. La stessa strategia potrebbe essere particolarmente interessante per le varianti ricorrenti osservate in due pazienti, verificando se le cellule ALT mostrino caratteristiche trascrizionali simili nei due soggetti.

Parallelamente, il workflow potrebbe essere applicato a coorti indipendenti di PDAC per verificare la ricorrenza non solo delle stesse varianti, ma anche di alterazioni negli stessi geni o negli stessi processi biologici. Le candidate potrebbero inoltre essere confrontate con dataset genomici pubblici WES/WGS per ricercare eventi analoghi o localizzati negli stessi domini funzionali.

Infine, le varianti più prioritarie potrebbero essere validate sperimentalmente anche senza disporre dei campioni originali, introducendo artificialmente gli alleli wild type e variante in sistemi appropriati. Per LCN2 e RPF2 potrebbero essere utilizzati sistemi *minigene* per studiare lo splicing; per SENP6 potrebbero essere valutate la stabilità del trascritto e la produzione della proteina; per le principali varianti missenso potrebbero essere confrontati direttamente gli effetti funzionali delle forme wild type e delle varianti più interessanti (Tabella 6).

Gene	Variante	Tier	Evidenza distintiva	Possibile interpretazione funzionale
SENP6	chr6:75713807 T>G	1A	*stop_gained*; SpliceAI donor loss 0,22	Possibile introduzione di uno stop prematuro, con produzione di una proteina tronca e/o attivazione del *nonsense-mediated decay*
LCN2	chr9:128153310 G>T	1A	*splice_acceptor*; SpliceAI 0,42; AlphaGenome: splicing	Possibile alterazione dello splicing del trascritto

RPF2	chr6:111025407 A>G	1A	SpliceAI 0,40; AlphaGenome: *splice junctions* e RNA DOWN	Possibile effetto sul processamento dell'RNA, oltre alla sostituzione aminoacidica
SRSF1	chr17:58005970 A>G	1A	Concordanza dei predittori missenso	Possibile alterazione della funzione proteica in un regolatore dello splicing
CNOT2	chr12:70346307 T>G	1A	Concordanza dei predittori missenso	Possibile impatto sulla funzione proteica e sul controllo dell'omeostasi dell'RNA
BPTF	chr17:67854505 T>G	1A	Concordanza dei predittori missenso	Possibile impatto sulla funzione proteica e sulla regolazione della cromatina
QTRT1	chr19:10713023 C>A	1A	Predittori missenso concordanti; variante ricorrente	Possibile alterazione proteica; la presenza in due pazienti costituisce un'ulteriore evidenza di interesse
PSMC3	chr11:47418898 G>C	1B	Variante ricorrente; AlphaGenome RNA DOWN	Possibile effetto proteico/trascrizionale; distinta dalla variante PSMC3 Tier 1A
NFKB2	chr10:102402366 A>C	2	popEVE severamente deleteria; AlphaGenome RNA UP e splicing; variante ricorrente	Profilo discordante tra predittori, compatibile con più possibili livelli di effetto funzionale

Tabella 6. Sintesi delle principali varianti discusse e dei possibili meccanismi funzionali.

6. CONCLUSIONI

Il presente lavoro ha mostrato che dati di single-nucleus RNA sequencing possono essere utilizzati, oltre che per l'analisi trascrittomica, anche come fonte complementare per l'identificazione e la prioritizzazione di varianti somatiche candidate nella componente epiteliale maligna del PDAC.

A partire da 414 varianti genomiche distinte, l'integrazione di annotazioni genomiche, predittori di impatto proteico, informazioni sullo splicing e predizioni regolatorie ha consentito di identificare un insieme ristretto di varianti ad alta priorità funzionale. I risultati evidenziano tuttavia che il variant calling da snRNA-seq fornisce una rappresentazione necessariamente parziale del panorama mutazionale, dipendente dall'espressione e dalla copertura dei loci, e non può sostituire il sequenziamento genomico.

Il principale valore dell'approccio risiede quindi nella possibilità di associare il segnale della variante alla popolazione cellulare nella quale viene rilevato e di generare ipotesi funzionali da sottoporre a successiva validazione genomica e sperimentale.

7. APPENDICI

Appendice A: Pipeline utilizzata per l'identificazione delle varianti somatiche candidate mediante SComatic.

```
cd /media/sf_CONDIVISA/SComatic-main
SCOMATIC=$PWD
output_dir=$SCOMATIC/PAZIENTI/U1/results
mkdir -p $output_dir
sample=U1
output_dir1=$output_dir/Step1_BamCellTypes
mkdir -p $output_dir1
python $SCOMATIC/scripts/SplitBam/SplitBamCellTypes.py --bam
$SCOMATIC/PAZIENTI/U1/U1.bam \
        --meta $SCOMATIC/PAZIENTI/ANNOTAZIONI/annotazioniU1.tsv \
        --id ${sample} \
        --n_trim 5 \
        --max_nM 5 \
        --max_NH 1 \
        --min_MQ 60 \
        --outdir $output_dir1
REF=$SCOMATIC/GENOMA/hg38.fa
        output_dir=$SCOMATIC/PAZIENTI/U1/results
output_dir2=$output_dir/Step2_BaseCellCounts
mkdir -p $output_dir2
for bam in $(ls -d $output_dir1/*bam);do
    cell_type=$(basename $bam | awk -F'.' '{print $(NF-1)}')
temp=$output_dir2/temp_${cell_type}
  mkdir -p $temp
  python $SCOMATIC/scripts/BaseCellCounter/BaseCellCounter.py --bam $bam \
    --ref $REF \
    --chrom all \
    --out_folder $output_dir2 \
    --min_bq 30 \
    --tmp_dir $temp \
    --min_mq 60 \
    --nprocs 8
  rm -rf $temp
done
sample=U1
output_dir3=$output_dir/Step3_BaseCellCountsMerged
mkdir -p $output_dir3
python $SCOMATIC/scripts/MergeCounts/MergeBaseCellCounts.py --tsv_folder ${output_dir2}
\
  --outfile ${output_dir3}/${sample}.BaseCellCounts.AllCellTypes.tsv
output_dir4=$output_dir/Step4_VariantCalling
mkdir -p $output_dir4
python $SCOMATIC/scripts/BaseCellCalling/BaseCellCalling.step1.py \
          --infile ${output_dir3}/${sample}.BaseCellCounts.AllCellTypes.tsv \
```

```
          --outfile ${output_dir4}/${sample} \
          --ref $REF
editing=$SCOMATIC/RNAediting/AllEditingSites.hg38.txt
PON=$SCOMATIC/PoNs/PoN.scRNAseq.hg38.tsv
python $SCOMATIC/scripts/BaseCellCalling/BaseCellCalling.step2.py \
          --infile ${output_dir4}/${sample}.calling.step1.tsv \
          --outfile ${output_dir4}/${sample} \
          --editing $editing \
          --pon $PON
bedtools intersect -header -a ${output_dir4}/${sample}.calling.step2.tsv -b
$SCOMATIC/bed_files_of_interest/UCSC.k100_umap.without.repeatmasker.bed | awk '$1 ~ /^#/
|| $6 == "PASS"' > ${output_dir4}/${sample}.calling.step2.pass.tsv
```

Appendice B: Script Python sviluppato per l'analisi delle varianti mediante AlphaGenome.

```
#!/usr/bin/env python3
import argparse
import os
import re
import traceback
from pathlib import Path
import matplotlib.pyplot as plt
import pandas as pd
from tqdm import tqdm
from alphagenome.data import gene_annotation, genome, junction_data, ontology, transcript
from alphagenome.models import dna_client, variant_scorers
from alphagenome.visualization import plot_components

OUTPUT_NAMES = ("RNA_SEQ", "CAGE", "PROCAP", "DNASE", "ATAC", "CHIP_HISTONE",
    "CHIP_TF", "SPLICE_SITES", "SPLICE_JUNCTIONS", "SPLICE_SITE_USAGE", "CONTACT_MAPS",)
LENGTH_NAMES = {"16KB": "SEQUENCE_LENGTH_16KB", "100KB": "SEQUENCE_LENGTH_100KB",
    "500KB": "SEQUENCE_LENGTH_500KB", "1MB": "SEQUENCE_LENGTH_1MB",}
GENCODE_URL = ("https://storage.googleapis.com/alphagenome/reference/gencode/hg38/"
    "gencode.v46.annotation.gtf.gz.feather")

def load_variants(path):
    df = pd.read_csv(path, sep="\t", dtype={
        "variant_id": str, "CHROM": str, "REF": str, "ALT": str, "gene": str,
    })
    missing = set(("variant_id", "CHROM", "POS", "REF", "ALT")) - set(df.columns)
    if missing:
        raise ValueError(f"Colonne mancanti: {sorted(missing)}")
    for column in ("variant_id", "CHROM", "REF", "ALT"):
        df[column] = df[column].astype(str)
    df["POS"] = df["POS"].astype(int)
    df["CHROM"] = df["CHROM"].map(lambda x: x if x.startswith("chr") else f"chr{x}")
    if df["variant_id"].duplicated().any():
        raise ValueError("variant_id duplicati: " + str(df.loc[
            df["variant_id"].duplicated(), "variant_id"].head(10).tolist()))
    return df

def load_annotations():
    gtf = pd.read_feather(GENCODE_URL)
    filtered = gene_annotation.filter_protein_coding(gtf)
    filtered = gene_annotation.filter_transcript_support_level(filtered, ["1"])
    longest = gene_annotation.filter_to_longest_transcript(filtered)
    return gtf, transcript.TranscriptExtractor(longest)
```

```
def gene_strand(gtf, gene):
    if gtf is None or not gene or gene.lower() == "nan":
        return None
    if not {"gene_name", "Strand"}.issubset(gtf.columns):
        print("[STRAND WARNING] GENCODE privo di gene_name/Strand")
        return None
    rows = gtf.loc[gtf["gene_name"].astype(str).eq(gene), "Strand"].dropna()
    rows = rows[rows.isin(["+", "-"])]
    if rows.empty:
        print(f"[STRAND WARNING] gene/strand non trovato: {gene}")
        return None
    return rows.value_counts().idxmax()

def serializable(df):
    df = df.copy()
    for col in df.select_dtypes(include="object").columns:
        df[col] = df[col].map(
            lambda x: "" if pd.api.types.is_scalar(x) and pd.isna(x) else str(x))
    return df

def filtered_scores(raw_scores, term, input_row, input_columns, strand):
    scores = serializable(variant_scorers.tidy_scores(raw_scores))
    cols = [c for c in scores if any(s in c.lower() for s in ("ontology", "curie", "term"))]
    mask = pd.Series(False, index=scores.index)
    for col in cols:
        mask |= scores[col].fillna("").astype(str).str.contains(term, regex=False)
    scores = scores.loc[mask].copy()  # Nessuna colonna ontologica: tabella vuota.
    for col in input_columns:
        if col not in scores:
            scores[col] = input_row[col]
    scores["input_gene"] = input_row.get("gene", "")
    scores["input_gene_strand"] = "" if strand is None else strand
    return serializable(scores)

def has_track(data):
    if data is None:
        return False
    try:
        if hasattr(data, "values") and (data.values is None or data.values.size == 0):
            return False
    except Exception:
        pass
    try:
        if hasattr(data, "metadata") and len(data.metadata) == 0:
            return False
    except Exception:
        pass
    return True

def track_on_strand(data, strand):
    if strand not in {"+", "-"}:
        return data
    names = (["filter_to_positive_strand", "filter_positive_strand", "positive_strand"]
             if strand == "+" else ["filter_to_negative_strand", "filter_negative_strand",
                                    "negative_strand",
"filter_to_nonpositive_strand"])
    for name in names:
        try:
            method = getattr(data, name, None)
            if method is not None:
                filtered = method()
                if has_track(filtered):
                    return filtered
        except Exception:
            pass
    try:
        keep = data.metadata["Strand"].astype(str).values == strand
        if keep.sum():
            filtered = data[:, keep]
            if has_track(filtered):
                return filtered
    except Exception:
        pass
    return data
```

```
def junctions_on_strand(data, strand):
    if strand not in {"+", "-"}:
        return data
    methods = (["filter_to_positive_strand", "filter_to_strand"] if strand == "+"
               else ["filter_to_negative_strand", "filter_to_strand"])
    for name in methods:
        try:
            method = getattr(data, name, None)
            if method is not None:
                return method(strand) if name == "filter_to_strand" else method()
        except Exception:
            pass
    return data

def plot_interval(data, variant, zoom_bp):
    try:
        return data.interval.resize(zoom_bp)
    except Exception:
        return variant.reference_interval.resize(zoom_bp)

def transcripts_for(extractor, interval):
    if extractor is None:
        return []
    try:
        return extractor.extract(interval)
    except Exception:
        return []

def draw(components, interval, variant, title, filename):
    try:
        plot_components.plot(components, interval=interval,
                             annotations=[plot_components.VariantAnnotation([variant],
alpha=0.8)])
        plt.title(title)
        plt.tight_layout()
        plt.savefig(filename, dpi=180)
    finally:
        plt.close("all")

def plot_tracks(prediction, variant, name, filename, mode, extractor, strand, zoom_bp,
source):
    ref = getattr(prediction.reference, name.lower(), None)
    alt = getattr(prediction.alternate, name.lower(), None)
    if not has_track(ref) or not has_track(alt):
        raise ValueError(f"Nessun TrackData valido per {name}")
    ref, alt = track_on_strand(ref, strand), track_on_strand(alt, strand)
    interval = plot_interval(ref, variant, zoom_bp)
    tx = transcripts_for(extractor, interval)
    components = [plot_components.TranscriptAnnotation(tx)] if len(tx) else []
    if name == "CONTACT_MAPS":
        # Le contact maps con più tracce vengono ridotte alla prima, come nello script
originale.
        ref = ref[:, :, 0] if ref.values.ndim == 3 else ref
        alt = alt[:, :, 0] if alt.values.ndim == 3 else alt
        if mode == "REF_vs_ALT":
            components              +=              [plot_components.ContactMaps(ref),
plot_components.ContactMaps(alt)]
        else:
            components.append(plot_components.ContactMapsDiff(alt - ref))
    elif mode == "REF_vs_ALT":
        components.append(plot_components.OverlaidTracks(tdata={"REF": ref, "ALT": alt}))
    elif mode == "ALT_minus_REF":
        components.append(plot_components.Tracks(alt - ref))
    else:
        raise ValueError(f"Tipo di grafico sconosciuto: {mode}")

    label = strand if strand in {"+", "-"} else "both_or_unknown"
    title = f"{variant.name} | {name} | {mode} | {source} | strand: {label} | window:
{zoom_bp} bp"
    draw(components, interval, variant, title, filename)
```

```
def plot_sashimi(prediction, variant, filename, delta, extractor, strand, zoom_bp,
source):
    ref,            alt            =            prediction.reference.splice_junctions,
prediction.alternate.splice_junctions
    if ref is None or alt is None:
        raise ValueError("splice_junctions assenti nella predizione")
    interval = plot_interval(ref, variant, zoom_bp)
    try:
        ref,           alt           =           ref.intersect_with_interval(interval),
alt.intersect_with_interval(interval)
    except Exception:
        pass
    ref, alt = junctions_on_strand(ref, strand), junctions_on_strand(alt, strand)
    tx = transcripts_for(extractor, interval)
    components = [plot_components.TranscriptAnnotation(tx)] if len(tx) else []
    if delta:
        # Come nell'originale, la sottrazione presume lo stesso ordine delle giunzioni
REF/ALT.
        if ref.values.shape != alt.values.shape:
            raise  ValueError(f"Forme  diverse  delle  giunzioni:  {ref.values.shape}  vs
{alt.values.shape}")
        diff = junction_data.JunctionData(
            junctions=alt.junctions, values=alt.values - ref.values,
            metadata=alt.metadata.copy(), interval=alt.interval, uns=getattr(alt, "uns",
None))
        components.append(plot_components.Sashimi(diff,  ylabel_template="ALT  -  REF  |
{name}",
                                                  filter_threshold=None,
annotate_counts=True,
                                                  normalize_values=False))
        mode = "Sashimi ALT - REF"
    else:
        for label, data in (("REF", ref), ("ALT", alt)):
            components.append(plot_components.Sashimi(data,  ylabel_template=label  +  "  |
{name}",
                                                      filter_threshold=None,
annotate_counts=True,
                                                      normalize_values=True))
        mode = "Sashimi REF/ALT"
    label = strand if strand in {"+", "-"} else "both_or_unknown"
    title = (f"{variant.name} | SPLICE_JUNCTIONS | {mode} | {source} | "
             f"strand: {label} | window: {zoom_bp} bp")
    draw(components, interval, variant, title, filename)

def plot_output(model,  interval,  variant,  output,  term,  label,  directory,  extractor,
strand, zoom_bp):
    name = output.name if hasattr(output, "name") else str(output).split(".")[-1]
    result = []
    for source, selected_term in ((label, term), ("GENERAL", None)):
        try:
            prediction = model.predict_variant(
                interval=interval,                                          variant=variant,
organism=dna_client.Organism.HOMO_SAPIENS,
                requested_outputs={output},
                ontology_terms=[selected_term] if selected_term is not None else [])
            stem      =      directory      /      f"{re.sub(r'[^A-Za-z0-9_.-]+',      '_',
variant.name)[:180]}.{name}.{source}"
            if name == "SPLICE_JUNCTIONS":
                plot_sashimi(prediction, variant, Path(f"{stem}.SASHIMI_REF_vs_ALT.png"),
                             False, extractor, strand, zoom_bp, source)
                try:
                    plot_sashimi(prediction,                                        variant,
Path(f"{stem}.SASHIMI_ALT_minus_REF.png"),
                                 True, extractor, strand, zoom_bp, source)
                except Exception as exc:
                    result.append(dict(output=name, source=source, status="delta_failed",
error=repr(exc)))
            else:
                for mode in ("REF_vs_ALT", "ALT_minus_REF"):
                    plot_tracks(prediction, variant, name, Path(f"{stem}.{mode}.png"),
                                mode, extractor, strand, zoom_bp, source)
        except Exception as exc:
            result.append(dict(output=name,        source=source,        status="failed",
error=repr(exc)))
```

```
        else:
            result.append(dict(output=name, source=source, status="plotted", error=""))
            break  # Se i grafici UBERON riescono, il fallback GENERAL non serve.
    return result

def arguments():
    p = argparse.ArgumentParser(description="Score per variante e grafici AlphaGenome su
hg38")
    p.add_argument("--variants", required=True, help="TSV: variant_id, CHROM, POS, REF,
ALT; gene opzionale")
    p.add_argument("--outdir", required=True, help="Cartella dei risultati")
    p.add_argument("--api-key",     help="Chiave     API;     alternativa:     variabile
ALPHAGENOME_API_KEY")
    p.add_argument("--sequence-length", choices=LENGTH_NAMES, default="1MB")
    p.add_argument("--ontology-term", default="UBERON:0001264")
    p.add_argument("--ontology-label", default="UBERON_0001264")
    p.add_argument("--zoom-bp", type=int, default=10000)
    p.add_argument("--keep-both-strands", action="store_true")
    p.add_argument("--skip-plots", action="store_true")
    p.add_argument("--overwrite", action="store_true")
    return p.parse_args()

def main():
    args = arguments()
    api_key = args.api_key or os.environ.get("ALPHAGENOME_API_KEY")
    if not api_key:
        raise ValueError("Manca la chiave: usare --api-key o ALPHAGENOME_API_KEY")
    outdir = Path(args.outdir)
    for folder in ("scores_per_variant", "plots", "logs"):
        (outdir / folder).mkdir(parents=True, exist_ok=True)
    variants = load_variants(args.variants)
    model = dna_client.create(api_key)
    length = dna_client.SUPPORTED_SEQUENCE_LENGTHS[LENGTH_NAMES[args.sequence_length]]
    outputs = [getattr(dna_client.OutputType, n) for n in OUTPUT_NAMES
               if hasattr(dna_client.OutputType, n)]
    scorers = list(variant_scorers.RECOMMENDED_VARIANT_SCORERS.values())
    term = ontology.from_curie(args.ontology_term)
    gtf, extractor = None, None
    if not args.skip_plots:
        try:
            gtf, extractor = load_annotations()
        except Exception as exc:
            print(f"GENCODE non disponibile; grafici senza annotazioni dei trascritti:
{exc!r}")
    failures, plot_log = [], []
    for _, row in tqdm(variants.iterrows(), total=len(variants)):
        variant_id = str(row["variant_id"])
        safe_id = re.sub(r"[^A-Za-z0-9_.-]+", "_", variant_id)[:180]
        score_dir = outdir / "scores_per_variant" / safe_id
        plot_dir = outdir / "plots" / safe_id
        score_dir.mkdir(parents=True, exist_ok=True)
        plot_dir.mkdir(parents=True, exist_ok=True)
        try:
            variant = genome.Variant(chromosome=row["CHROM"], position=int(row["POS"]),
                                     reference_bases=row["REF"],
alternate_bases=row["ALT"],
                                     name=variant_id)
            interval = variant.reference_interval.resize(length)
            gene = str(row.get("gene", "")).strip() if "gene" in row.index else ""
            strand = None if args.keep_both_strands else gene_strand(gtf, gene)
            stem = score_dir / f"{safe_id}.scores.{args.ontology_label}"
            if    args.overwrite    or    not    (Path(f"{stem}.tsv").exists()    and
Path(f"{stem}.xlsx").exists()):
                raw = model.score_variant(interval=interval, variant=variant,
                                          variant_scorers=scorers,
                                          organism=dna_client.Organism.HOMO_SAPIENS)
                scores = filtered_scores(raw, args.ontology_term, row, variants.columns,
strand)
                scores.to_csv(f"{stem}.tsv", sep="\t", index=False)
                scores.to_excel(f"{stem}.xlsx", index=False)
            if not args.skip_plots:
                for output in outputs:
```

```
                    rows = plot_output(model, interval, variant, output, term,
args.ontology_label,
                                       plot_dir, extractor, strand, args.zoom_bp)
                    for record in rows:
                        record.update(variant_id=variant_id,                    gene=gene,
gene_strand=strand)
                        plot_log.append(record)
                        if record["status"] in {"failed", "delta_failed"}:
                            print(f"[PLOT        {record['status']}]        {variant_id}
{record['output']} "
                                  f"{record['source']}: {record['error']}")
        except Exception as exc:
            failures.append(dict(variant_id=variant_id,           CHROM=row.get("CHROM"),
POS=row.get("POS"),
                                 REF=row.get("REF"),                  ALT=row.get("ALT"),
gene=row.get("gene", ""),
                                 error=repr(exc), traceback=traceback.format_exc()))

    pd.DataFrame(failures).to_csv(outdir / "logs" / "failed_variants.tsv", sep="\t",
index=False)
    pd.DataFrame(plot_log).to_csv(outdir / "logs" / "plot_log.tsv", sep="\t", index=False)
    print(f"Finito: {len(variants)} varianti, {len(failures)} fallite. Risultati in
{outdir}")

if __name__ == "__main__":
    main()
```

Riferimenti bibliografici

1. Roth, S., Apte, M., Balachandran, V. P., Biffi, G., Conroy, T., Cukierman, E., ... & Michalski, C. W. (2026). Pancreatic cancer. *Nature Reviews Disease Primers*, *12*(1), 23.
2. Hezel, A. F., Kimmelman, A. C., Stanger, B. Z., Bardeesy, N., & DePinho, R. A. (2006). Genetics and biology of pancreatic ductal adenocarcinoma. *Genes & development*, *20*(10), 1218-1249.
3. International Agency for Research on Cancer. (2024). *Global Cancer Observatory: Cancer today. Pancreas*. https://gco.iarc.who.int/media/globocan/factsheets/cancers/13-pancreas-fact-sheet.pdf
4. Hingorani, S. R., Petricoin, E. F., Maitra, A., Rajapakse, V., King, C., Jacobetz, M. A., ... & Tuveson, D. A. (2003). Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer cell*, *4*(6), 437-450.
5. Hruban, R. H., Adsay, N. V., Albores–Saavedra, J., Compton, C., Garrett, E. S., Goodman, S. N., ... & Offerhaus, G. J. A. (2001). Pancreatic intraepithelial neoplasia: a new nomenclature and classification system for pancreatic duct lesions. *American Journal of Surgical Pathology*, *25*(5), 579-586.
6. Guerra, C., Collado, M., Navas, C., Schuhmacher, A. J., Hernández-Porras, I., Cañamero, M., ... & Barbacid, M. (2011). Pancreatitis-induced inflammation contributes to pancreatic cancer by inhibiting oncogene-induced senescence. *Cancer cell*, *19*(6), 728-739.
7. Aguirre, A. J., Bardeesy, N., Sinha, M., Lopez, L., Tuveson, D. A., Horner, J., ... & DePinho, R. A. (2003). Activated Kras and Ink4a/Arf deficiency cooperate to produce metastatic pancreatic ductal adenocarcinoma. *Genes & development*, *17*(24), 3112-3126.
8. Jones, S., Zhang, X., Parsons, D. W., Lin, J. C. H., Leary, R. J., Angenendt, P., ... & Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *science*, *321*(5897), 1801-1806.
9. Waddell, N., Pajic, M., Patch, A. M., Chang, D. K., Kassahn, K. S., Bailey, P., ... & Greater Glasgow & Clyde National Health Service Biankin Andrew V. 36 37 57 Chang David K. 36 37 57 Jamieson Nigel B. 37 57 Graham Janet S. 37 57 Oien

Karen 37 57 Hair Jane 57. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, *518*(7540), 495-501.

10. Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., ... & Lolla, L. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, *32*(2), 185-203.
11. Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A. M., Gingras, M. C., ... & Grimmond, S. M. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, *531*(7592), 47-52.
12. Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., Truitt, M., Gu, S., ... & Gray, J. W. (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine*, *17*(4), 500-503.
13. Moffitt, R. A., Marayati, R., Flate, E. L., Volmar, K. E., Loeza, S. G. H., Hoadley, K. A., ... & Yeh, J. J. (2015). Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics*, *47*(10), 1168-1178.
14. Grant, R. C., Selander, I., Connor, A. A., Selvarajah, S., Borgida, A., Briollais, L., ... & Gallinger, S. (2015). Prevalence of germline mutations in cancer predisposition genes in patients with pancreatic cancer. *Gastroenterology*, *148*(3), 556-564.
15. Golan, T., Hammel, P., Reni, M., Van Cutsem, E., Macarulla, T., Hall, M. J., ... & Kindler, H. L. (2019). Maintenance olaparib for germline BRCA-mutated metastatic pancreatic cancer. *New England Journal of Medicine*, *381*(4), 317-327.
16. Makohon-Moore, A. P., Zhang, M., Reiter, J. G., Bozic, I., Allen, B., Kundu, D., ... & Iacobuzio-Donahue, C. A. (2017). Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nature genetics*, *49*(3), 358-366.
17. Conroy, T., Desseigne, F., Ychou, M., Bouché, O., Guimbaud, R., Bécouarn, Y., ... & Ducreux, M. (2011). FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *New England journal of medicine*, *364*(19), 1817-1825.
18. Sherman, M. H., & Beatty, G. L. (2023). Tumor microenvironment in pancreatic cancer pathogenesis and therapeutic resistance. *Annual Review of Pathology: Mechanisms of Disease*, *18*(1), 123-148.

19. Peng, J., Sun, B. F., Chen, C. Y., Zhou, J. Y., Chen, Y. S., Chen, H., ... & Wu, W. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research*, *29*(9), 725-738.
20. Raghavan, S., Winter, P. S., Navia, A. W., Williams, H. L., DenAdel, A., Lowder, K. E., ... & Shalek, A. K. (2021). Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell*, *184*(25), 6119-6137.
21. Ramsköld, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., ... & Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*, *30*(8), 777-782.
22. Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., ... & Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, *352*(6293), 1586-1590.
23. Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., ... & Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, *8*(1), 14049.
24. Muyas, F., Sauer, C.M., Valle-Inclán, J.E. *et al.* De novo detection of somatic mutations in high-throughput single-cell profiling data sets. *Nat Biotechnol* **42**, 758–767 (2024).
25. Hwang, W. L., Jagadeesh, K. A., Guo, J. A., Hoffman, H. I., Yadollahpour, P., Reeves, J. W., ... & Regev, A. (2022). Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment. *Nature genetics*, *54*(8), 1178-1191.
26. Kaminow, B., Yunusov, D., & Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *Biorxiv*, 2021-05.
27. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, A. D. (2002). The human genome browser at UCSC. *Genome research*, *12*(6), 996-1006.
28. Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., Taylor, K. R., Ward, T., ... & Kohli, P. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, *649*(8099), 1206-1218.

29. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... & Cunningham, F. (2016). The ensembl variant effect predictor. *Genome biology*, *17*(1), 122.
30. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., ... & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, *381*(6664), eadg7492.
31. Ensembl. AlphaMissense – Ensembl Variant Effect Predictor (VEP) plugin documentation. Ensembl. Consultato nel 2026.
32. Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., ... & Sieh, W. (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, *99*(4), 877-885.
33. Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J., & Hocking, T. D. (2018). ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. *The American Journal of Human Genetics*, *103*(4), 474-483.
34. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., ... & Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, *599*(7883), 91-95.
35. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, *47*(D1), D886-D894.
36. Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, *6*(12), e1001025.
37. Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... & Farh, K. K. H. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, *176*(3), 535-548.
38. UCSC Genome Browser. *SpliceAI Variants – Splicing Impact*. University of California, Santa Cruz. Disponibile presso: UCSC Genome Browser, SpliceAI track. Consultato nel 2026.
39. AlphaGenome Atlas Team. *AlphaGenome Atlas: A predictive map of every possible DNA letter change in the human genome*. Google DeepMind; 2026.

40. Cheng, J., Taylor, K. R., Nicolaisen, L., Pan, J., Bycroft, C., Perino, M., ... & Avsec, Z. (2026). AlphaGenome Atlas: in silico mutagenesis of the entire human genome improves prioritization and interpretation of non-coding variants. *medRxiv*, 2026-09.
41. Ensembl. *Variant Effect Predictor (VEP): AVI plugin – AlphaGenome Variant Impact scores*. Ensembl Documentation. Consultato nel 2026.
42. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434-443.
43. Zhang, X., Wakeling, M., Ware, J., & Whiffin, N. (2021). Annotating high-impact 5′ untranslated region variants with the UTRannotator. *Bioinformatics*, *37*(8), 1171-1173.
44. Stephenson, J. D., Totoo, P., Burke, D. F., Jänes, J., Beltrao, P., & Martin, M. J. (2024). ProtVar: mapping and contextualizing human missense variation. *Nucleic Acids Research*, *52*(W1), W140-W147.
45. Wagih, O., Galardini, M., Busby, B. P., Memon, D., Typas, A., & Beltrao, P. (2018). A resource of variant effect predictions of single nucleotide variants in model organisms. *Molecular systems biology*, *14*(12), MSB188430.
46. Gao, H., Hamp, T., Ede, J., Schraiber, J. G., McRae, J., Singer-Berk, M., ... & Farh, K. K. H. (2023). The landscape of tolerated genetic variation in humans and primates. *Science*, *380*(6648), eabn8153.
47. Orenbuch, R., Shearer, C. A., Kollasch, A. W., Spinner, A. D., Hopf, T., van Niekerk, L., ... & Marks, D. S. (2025). Proteome-wide model for human disease genetics. *Nature Genetics*, *57*(12), 3165-3174.
48. Leung, L., Radulovich, N., Zhu, C. Q., Organ, S., Bandarchi, B., Pintilie, M., ... & Tsao, M. S. (2012). Lipocalin2 promotes invasion, tumorigenicity and gemcitabine resistance in pancreatic ductal adenocarcinoma.
49. Wan, L., Lin, K. T., Rahman, M. A., Ishigami, Y., Wang, Z., Jensen, M. A., ... & Krainer, A. R. (2023). Splicing factor SRSF1 promotes pancreatitis and KRASG12D-mediated pancreatic cancer. *Cancer discovery*, *13*(7), 1678-1695.
50. Jee, W., Ko, H. M., Park, D. I., Park, Y. R., Park, S. M., Kim, H., ... & Jang, H. J. (2023). Momordicae Semen inhibits migration and induces apoptotic cell death

by regulating c-Myc and CNOT2 in human pancreatic cancer cells. *Scientific Reports*, *13*(1), 12800.

51. Munoz Velasco, R., Jimenez Sanchez, P., Garcia Garcia, A., Blanco Martinez-Illescas, R., Pastor Senovilla, A., Lozano Yagüe, M., ... & Sanchez-Arevalo Lobo, V. J. (2022). Targeting BPTF sensitizes pancreatic ductal adenocarcinoma to chemotherapy by repressing ABC-transporters and impairing multidrug resistance (MDR). *Cancers*, *14*(6), 1518.
52. Dumartin, L., Whiteman, H. J., Weeks, M. E., Hariharan, D., Dmitrovic, B., Iacobuzio-Donahue, C. A., ... & Crnogorac-Jurcevic, T. (2011). AGR2 is a novel surface antigen that promotes the dissemination of pancreatic cancer cells through regulation of cathepsins B and D. *Cancer research*, *71*(22), 7091-7102.
53. Hassan, Z., Schneeweis, C., Wirth, M., Müller, S., Geismann, C., Neuß, T., ... & Schneider, G. (2021). Important role of Nfkb2 in the KrasG12D-driven carcinogenesis in the pancreas. *Pancreatology*, *21*(5), 912-919.