Nel gennaio del 2020 un nuovo betacoronavirus, designato come SARS-CoV-2, viene identificato come agente di alcuni casi di polmonite nella città di Wuhan in Cina. A questa nuova malattia verrà dato il nome di COVID-19, e l'11 marzo 2020 la World Healt Organization dichiara lo stato di pandemia. Il COVID-19 ha colpito più di 200 stati in tutto il mondo, causando 162 milioni di contagi individuali e oltre 3 milioni di morti. Questa nuova malattia viene principalmente trasmessa attraverso le gocce di saliva e il contatto fisico. I periodi di incubazione possono variare tra i 2 e i 14 giorni. I sintomi più comuni sono febbre, tosse secca e fatica; mentre sintomi meno comuni sono dolore ai muscoli, congestiona nasale, gola infiammata e diarrea. Una minoranza di pazienti ha inoltre avuto come sintomo anche la polmonite, una malattia respiratoria grave.
La prima sequenza genomica completa del nuovo betacoronavirus su trovata alla fine di dicembre 2019 attraverso approcci meta-trascrittomici, supportati da PCR e dal sequenziamento Sanger. La scoperta di questo genoma di riferimento ha poi facilitato la scoperta dei test diagnostici basati sulla real-time PCR.
Il genoma di SARS-CoV-2 è di RNA ed è lungo all'incirca 30 000 nucleotidi. È formato da due reading frame, ORF1a e ORF1b, che occupano i due terzi del genoma alla fine 5'. Da ORF1a si traduce la poliproteina 1, mentre da ORF1a e ORf1b si traduce la poliproteina 11b; queste due poliproteine vengono poi processate in 16 proteine che servono per la replicazione virale del genoma e la trascrizione. Dalla parte della fine terminale 3' si hanno 4 proteine strutturali e altre 6 proteine accessorie che sono meno caratterizzate e non sono universalmente presenti in tutti i genomi di coronavirus.

Recenti esperienze con nuove malattie infettive, come SARS, MERS, Zika o Ebola, hanno dimostrato che le tecnologie di Next Generation Sequencing rappresentano uno strumento importante per identificare le origini e per tenere traccia della propagazione del virus e delle catene di trasmissione dei focolai.

Sample collection
Le sequenze di dati disponibili di SARS-CoV-2 derivano principalmente da campioni clinici diagnosticati, e quindi presentano solitamente alte cariche virali che permettono l'estrazione di una quantità di RNA tale per cui si riesce a sequenziarlo e a ricostruire un genoma virale.
La WHO ha fatto una lista dei vari tipi di campioni clinici che possono essere raccolti e dai quali si può estrarre l'RNA di SARS-CoV-2.
La maggior parte dei campioni che si hanno deriva dal tratto respiratorio superiore o inferiore. Ci sono stati anche casi in cui il genoma è stato però ricostruito da campioni clinici non di origine respiratoria, provenienti per esempio da urine o feci. Altra origine di campioni potrebbe essere quella di linee cellulari infette, ma questo viene poco usato perché in questo modo spesso si accumulano nuove varianti genetiche durante il passaggio in laboratorio, e di conseguenza ciò implica profonde modifiche nel risultato dello studio.
Un numero molto limitato di campioni proviene poi da elementi ambienti come acque reflue o campioni di aria, questo per la poca caria infettiva che presentano.
È chiaro che la maggior parte dei campioni disponibili provengano da tratti respiratori, ma in molti casi non è comunque specificata la reale provenienza del dato all'interno dei repository di genomi di SARS-CoV-2 attualmente in circolazione, questo sottolinea già l'incompletezza della diffusione dei metadati associati al genoma virale.

RNA extraction
L'RNA può essere estratto da campioni clinici, da colture isolate o da campioni ambientali, utilizzando una larga varietà di kit commercialmente disponibili. Le metodologie standard includono l'uso di sale di Guanidina, per inibire i nucleasi [è un enzima capace di idrolizzare i legami fosfodiestere fra le subunità nucleotidiche degli acidi nucleici.] e assicurare che l'RNA non si degradi, e l'uso di fenolo, per denaturare e dissolvere le proteine che effettivamente attivano il virus.
Prima del sequenziamento dell'RNA bisogna però valutare la presenza e la quantità di SARS-CoV-2 utilizzando una qRT-PCR ((Real-Time Quantitative Reverse Transcription PCR) è un importante sviluppo

della tecnologia PCR che consente il rilevamento e la misurazione affidabili dei prodotti generati durante ogni ciclo del processo PCR).

Strategie di sequenziamento

Le strategie di sequenziamento NGS sono il metodo maggiormente scelto per l'identificazione di nuovi virus, per la ricostruzione delle sequenze genomiche virali e per l'analisi dell'evoluzione del virus. Uno dei vantaggi di queste tecnologie è che è possibile ricostruire l'intera sequenza genomica a partire sia da colture virali amplificate sia direttamente da campioni clinici.
Ovviamente in base agli obiettivi sperimentali di ogni progetto si deve scegliere l'opportuna tecnica di sequenziamento, bisogna perciò tenere in considerazione il tipo di campione che si ha, il suo carico virale, la procedura con la quale è estratto l'RNA e la qualità dell'RNA.
Al giorno d'oggi vengono principalmente applicati quattro approcci per il sequenziamento.

Shotgun metatrascriptomics

Il sequenziamento shotgun è una tecnica indipendente dalla coltura e permette di sequenziare l'intero DNA in un campione, permettendo quindi la caratterizzazione di complessi microrganismi senza avere nessuna conoscenza a priori riguardo la loro sequenza genomica. Per questo motivo questo metodo è uno strumento potente per l'identificazione di patogeni non ancora caratterizzati, poiché offre informazioni dettagliate e quantitative sulla composizione delle comunità microbiotiche e inoltre possono anche fornire informazioni riguardo infezioni secondarie.
Le tecnologie commercialmente disponibili per questo approccio provengono dalla Oxford Nanopore Technology (ONT) e dalla Pacifici Bioscense (PacBio).
La maggiore limitazione di questo metodo è il requisito di avere campioni con un alto carico virale di partenza per poter ricostruire la sequenza genomica completa del virus; inoltre è richiesta una profondità di sequenziamento alta (> di 2 giga basi). In aggiunta a tutto ciò è un approccio che richiede tempo, molto lavoro e l'accesso a laboratori con requisiti specifici.

Amplicon-base sequencing

Con il metodo basato su ampliconi si restringe la portata dell'analisi solo ad un numero limitato di sequenze scelte, per cui questo è un approccio altamente specifico, però richiede una significativa conoscenza a priori della sequenza che viene scelta per il test.
Il workflow che si usa in questi casi è quello di sintetizzare il primo filamento di cDNA a partire dal campione di RNA, e poi amplificarlo con PCR multiple fino ad avere degli ampliconi che riescano a coprire l'intera lunghezza o la porzione di genoma desiderata.
È un approccio robusto in quanto può essere utilizzato con poche quantità di RNA, che quindi può anche provenire da campioni degradati; è conveniente e poco costoso. Presenta come tutte le cose delle limitazioni: prima di tutto le differenze nell'efficienza dei primer nella reazione di PCR possono portare a amplificazioni che non coprono tutta la regione genomica richiesta, e questo si traduce in una ricostruzione incompleta; poi i primer vengono generati in riferimento alla sequenza di riferimento del genoma di SARS-CoV-2, e avendo piccole sequenze potrebbero non riuscire ad identificare grandi varianti strutturali.
Questo è un metodo usato prevalentemente per tenere sotto controllo la catena di trasmissione e per lo studio di campioni ambientali, che come già detto presentano cariche virali molto basse.
In commercio sono disponibili protocolli di sequenziamento che provengono dalla piattaforma PacBio.

Hibryd capture enrichment sequencing

È un tipo di sequenziamento che fa in modo che vengano sequenziate solo regioni di un genoma scelte che siano effettivamente rilventi per l'obiettivo dello studio. È un metodo simile al precedente.
La cattura ibrida amplifica il materiale genetico che è stato scelto attraverso l'ibridazione di specifiche probes biotilinate, questo comporta una riduzione della profondità di sequenziamento rispetto per esempio a quella richiesta per il sequenziamento shotgun.

Questi approcci sono generalmente basati su un grande numero di probes e fornisce un profilo generale più completo, poiché la cattura per ibridazione è più robusta della variabilità genomica.
Viene utilizzato sia per campioni con alta carica virale sia per campioni con bassa carica virale.
Commercialmente sequenziatori di questo tipo provengono da Illumina.

Direct RNA sequencing
Le precedenti strategie di sequenziamento richiedono tutte la retrotrascrizione dell'RNA, e un certo grado di manipolazioni per la costruzione della libreria, queste operazioni possono portare alla perdita di informazioni.
Tra le tecnologie di sequenziamento recentemente scoperte si ha la SMS, che permette di studiare direttamente la sequenza di una singola molecola di acido nucleico, senza amplificazioni e senza retrotrascrizioni. Queste tecnologie forniscono reads più lunghe ma con tasso di errore maggiore.
Con questa tecnica è possibile la ricostruzione accurata di trascritti singoli e di pattern trascrizionali complessi, come quelli che si hanno durante l'infezione da coronavirus.
Commercialmente questa tecnologie si ha da ONT.

Analisi dei dati, deposizione e accesso

Assemblaggio genomico
Poiché il genoma di SARS-CoV-2 è relativamente compatto e non contiene nessuna grande sequenza ripetuta, la ricostruzione del genoma dovrebbe essere un processo semplice, ma questo solo in teoria perché tutto dipende dai dati che vengono forniti dai vari approcci di sequenziamento.
Gli approcci metatrascrittomici e di cattura ibrida sembrano fornire in media una più completa rappresentazione del genoma di SARS-CoV-2. Per il sequenziamento basato su ampliconi, ONT tende ad essere più completo di Illumina.
La quantità di dati generati da ogni approccio è in line acon le aspettative.
Le librerie metatrascrittomiche presentano una copertura genomica molto più uniforme, anche se si ha una riduzione considerevole alle due estremità. Inoltre questi metodi possono presentare però porzioni variabili in base al carico virale del campione di partenza. I metodi basati sulla cattura ibrida forniscono una copertura abbastanza uniforme. Invece i metodi basati su ampliconi forniscono una copertura genomica più distorta, la presenza di picchi è dovuta alle regioni in cui si sovrappongono ampliconi diversi.
Quindi alla fine in generale il processo di assemblaggio del genoma di SARS-CoV-2 è un processo molto complesso e computazionalmente inteso e può essere influenzato da moltissimi fattori.

La deposizione in brevi tempi delle nuove scoperte e quindi dei nuovi dati riguardanti ogni aspetto di SARS-CoV-2 è essenziale, soprattutto per l'implementazione di nuove strategie di mitigazioni efficace, e quindi per la ricerca farmaceutica, per la scoperta di vaccini, e per capire in generale con sempre più chiarezza la malattia e i suoi effetti.
Al momento attuale il portale GISAD EpiCov rappresenta il repository di dati genomici su SARS-CoV-2 più largamente usato. Contiene oltre 100 000 genomi completi di SARS-CoV-2, il suo limite è però quello di avere associati ad ogni genoma pochissimi metadati che potrebbero avere informazioni potenzialmente importanti, come per esempio l'origine del campione, la tecnica di sequenziamento o semplici annotazioni cliniche di base.
Un altro repository disponibile è Research Data Alliance, in più rispetto a GISAD, contiene dati sul sequenziamento e sul genoma di SARS-CoV-2 più conformi ai principi FAIR
(**F**indability, **A**ccessibility, **I**nteroperability, and **R**euse).
[l'espressione genica dei dati può essere depositia in ArrayExpress o Gene Express Omnibus. I cataloghi EGA e GWAS contengono invece i dati associati ai genomi]

Oltre al materiale grezzo e ai metadati, anche la riproducibilità dell'analisi bioinformatica e dei workflow che vengono usati costituisce un elemento chiave della biologia moderna. Anche in questo casi queste conoscenze dovrebbero essere subito rese disponibili attraverso repository specifici. Un esempio di catalogo molto curato di software bioinformatici e di applicazione è *bio.tools* . Altri esempi sono la piattaforma Galaxy o il portale Microreat.

Esistono poi ulteriori portali online come Nextstrain e Hyphy COVID-19 che permettono il monitoraggio in tempo reale dell'evoluzione dei vari ceppi di SARS-CoV-2. In particolare il primo fornisce informazioni sulla distribuzione dei diversi tipi di SARS-CoV-2 nel mondo, mentre il secondo contiene analisi dettagliate riguardanti i geni che codificano le proteine di SARS-CoV-2.

Infine i portali EBI COVID-19 Data Portal e NCB forniscono un catalogo abbastanza completo di risorse per poter accedere e recuperare i dati riguardanti il SARS-CoV-2 e i relativi strumenti bioinformatici.

Sebbene tutti questi strumenti forniscono una grande ricchezza di dati di sequenziamento di SARS-CoV-2 e dei metadati relativi, la loro integrazione non è semplice.
Analisi esplorative ottenute dai vari dati genomici di disponibili sul SARS-CoV-2 sono raccolte in tre diversi portali: COG-UK, GISAID EpiCoV e NCBI Virus Portal. Questi tre database presentano però delle inconsistenze. Nel caso di GISAID, che contiene più di 100 000 genomi, si sa che circa 22 599 di questi derivano da COG-UK. Tuttavia questi genomi non rappresentano la totalità di COG-UK, che contiene 48 000 genomi totali. In modo analogo il 10% dei genomi in NCBI Virus Portal provengono da GISAID.
Al momento attuale il repository che collettivamente fornisce un accesso il più completo possibile è INSDC.

Conclusioni
Negli ultimi decenni si è avuta una significativa attenzione politica focalizzata sul bisogno di identificare e limitare l'emergere di focolai che avrebbero potuto o hanno portato a pandemie. In questo contesto dei metodi rapidi e convenienti per la ricostruzione di sequenze genomiche di patogeni rappresentano un importante strumento per il monitoraggio e il contrasto della diffusione di nuove malattie infettive nell'uomo. Nel caso di SARS-CoV-2 i metodi NGS sono stati molto utili e si è visto come possono essere potenzialmente applicati anche in altri ambiti biologici associati. In soli pochi mesi infatti si è riusciti ad avere grazie a NGS i dati sulla sequenza genomica di SARS-CoV-2, e da questi poi si sono potuti fare vari studi per l'identificazione dell'evoluzione del virus e per la scoperta di possibili metodi di prevenzione e attacco. Una ricchezza di database fornisce accesso a tutti questi dati su SARS-CoV-2, è importante però massimizzare il suo utilizzo e curare anche l'analisi secondare e l'integrazione dei metadati, in modo da facilitare analisi e studi successivi. La disponibilità e l'integrazione di questi dati in maniera pubblica è dunque fondamentale per aprire nuove porte alla scienza e per permettere il progresso.