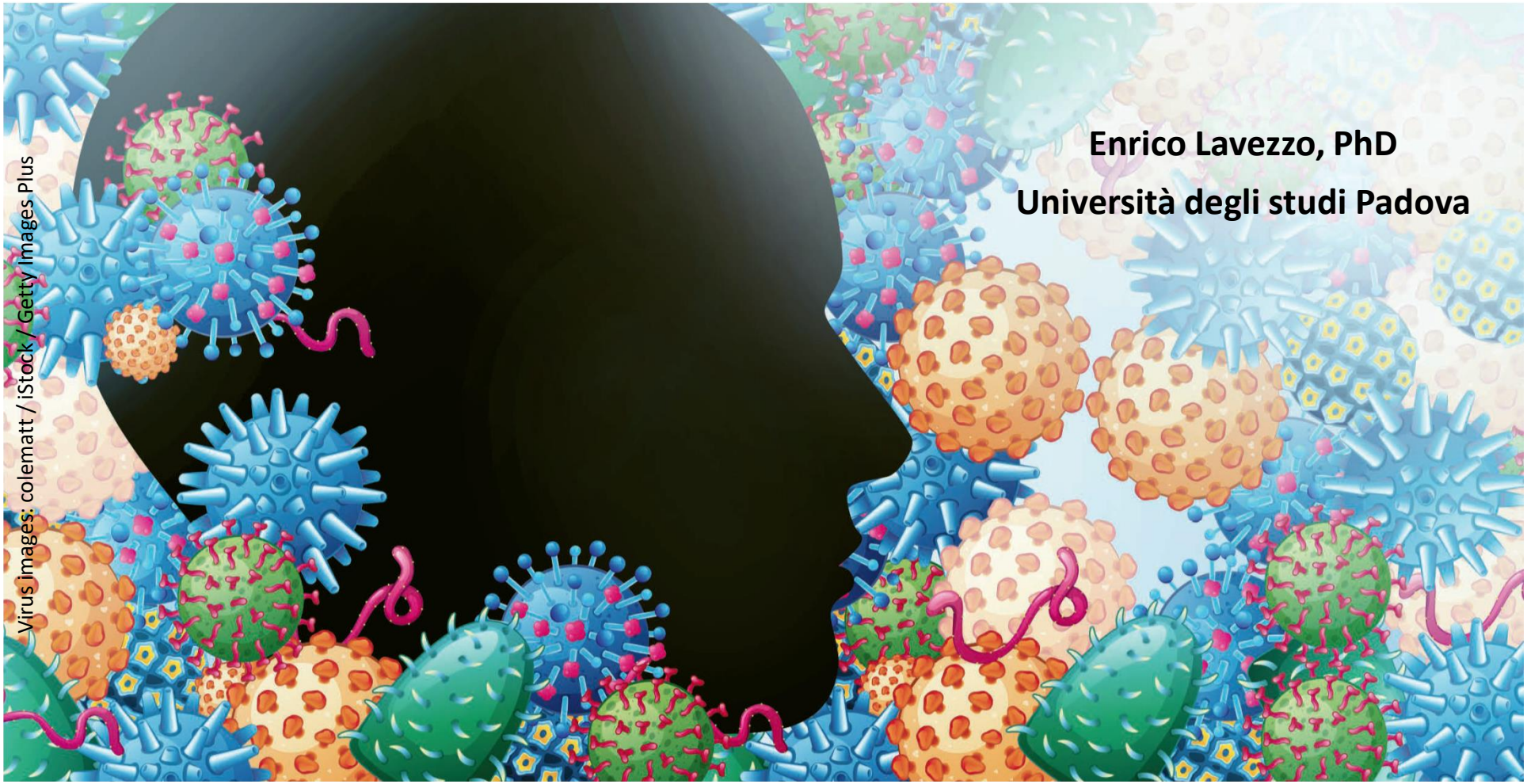
Aspetti bioinformatici associati all'NGS: dalla metagenomica alla ricerca di varianti minoritarie



Enrico Lavezzo, PhD

Università degli studi Padova

Virus images: colematt / iStock / Getty Images Plus

Sommario

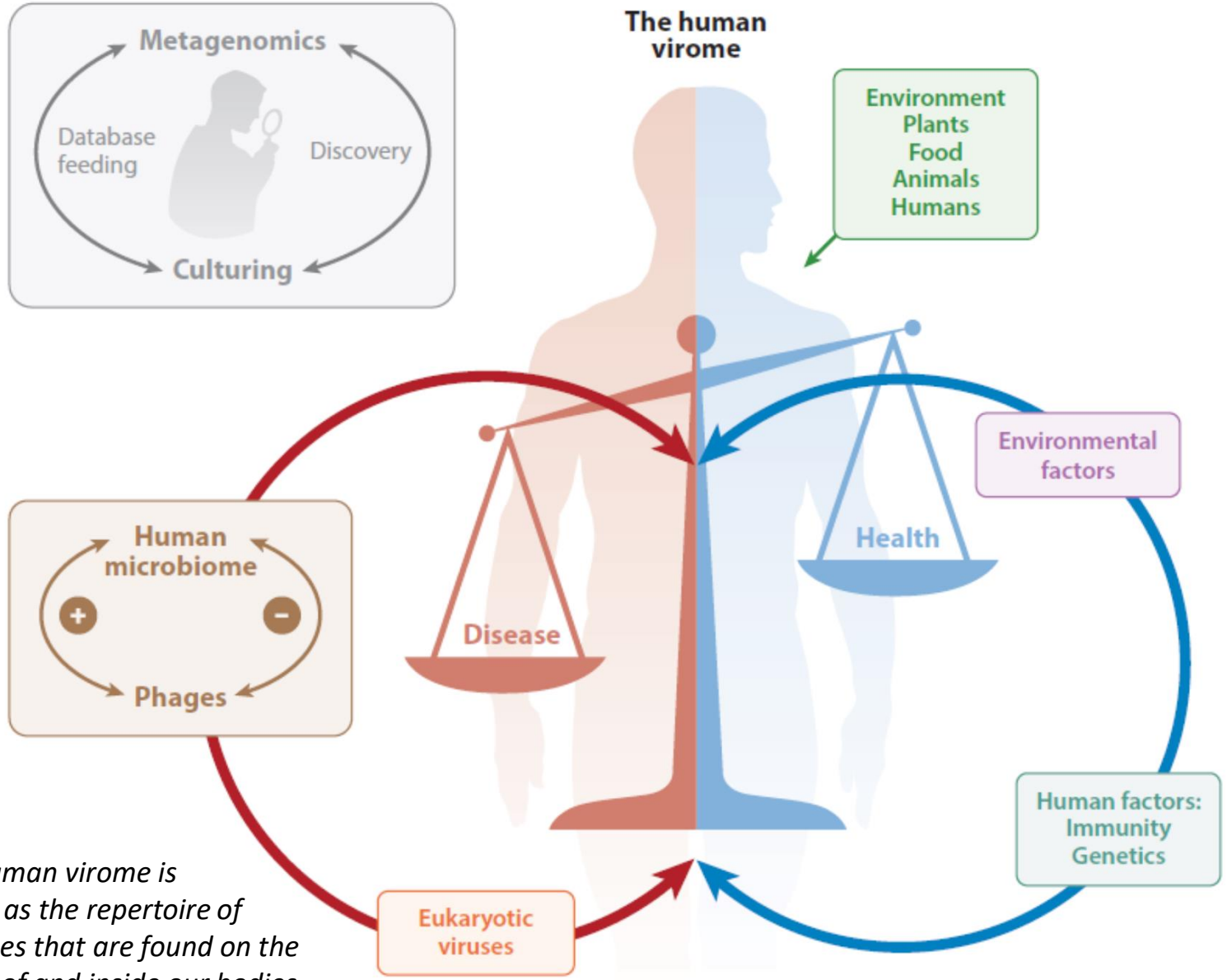
- La nuova accessibilità del viroma, grazie allo sviluppo delle piattaforme NGS

- La caratterizzazione del viroma attraverso NGS:

  - Le piattaforme di sequenziamento

  - Metagenomica e metatrascrittomica **shotgun:** who is there?

  - Deep sequencing mirato: genotipizzazione, farmaco-resistenze, identificazione di varianti minoritarie



"The human virome is defined as the repertoire of all viruses that are found on the surface of and inside our bodies in the absence of clinically significant symptoms of infection."

Edited from Rascovan et al., Annu. Rev. Microbiol. 2016.

Perché studiare il viroma?

"We know of only a minuscule fraction of the viruses out there, and our questions about the viral world are profound"

Edward Holmes, a virologist and professor at the University of Sydney in Australia.

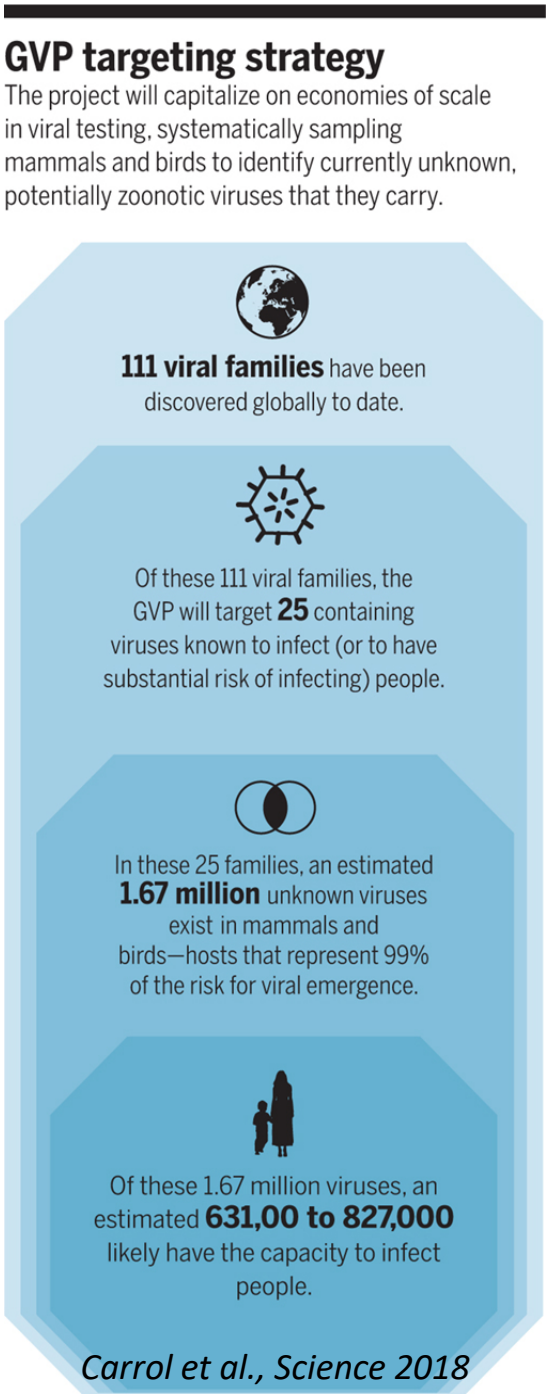
Global Virome Project → progetto internazionale che punta all'identificazione di nuovi virus potenzialmente pericolosi per l'uomo

Interfaccia uomo/animale

Obiettivi:

- facilitare la rapida identificazione dell'origine (ospite animale) di potenziali nuove epidemie virali
- identificare le modalità di trasmissione
- caratterizzare l'evoluzione dei virus nel tempo
- prevedere virus causeranno nuove infezioni ed epidemie nell'uomo

La disponibilità di tali informazioni nelle fasi iniziali di un'epidemia può essere determinante per mitigare o fermare l'epidemia stessa



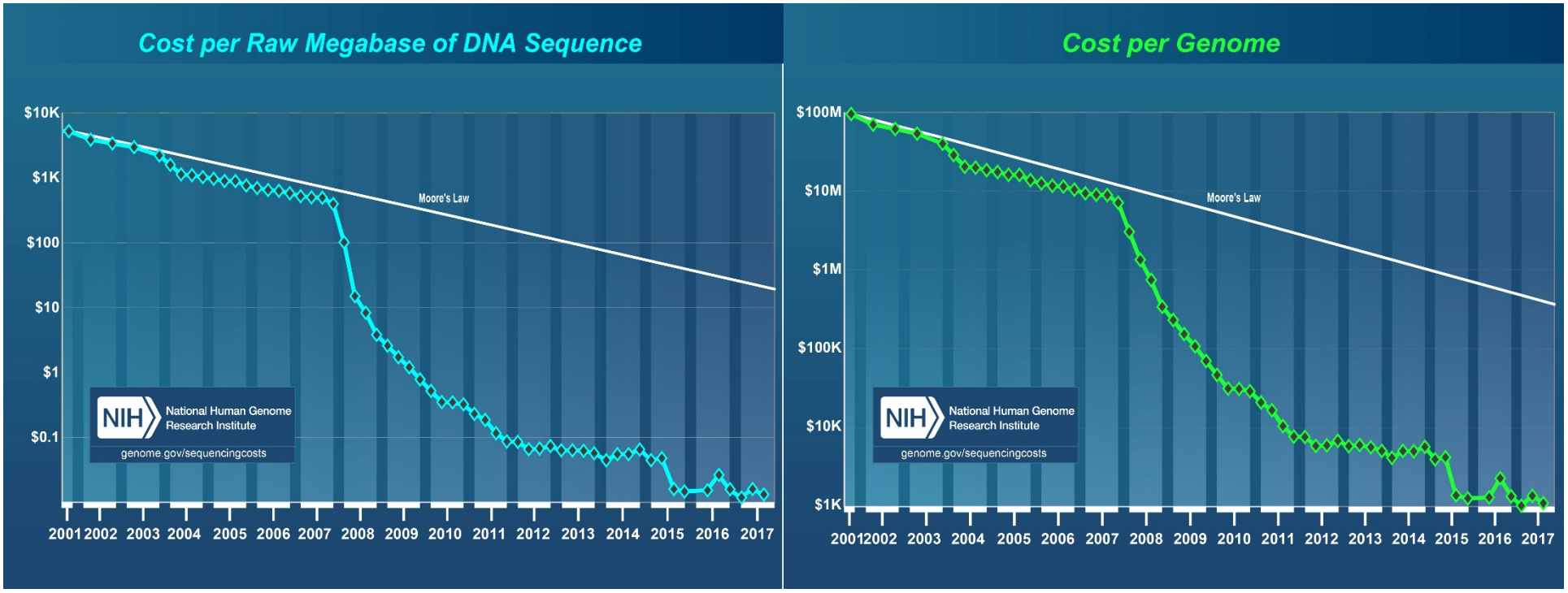
Carrol et al., Science 2018

Next generation sequencing e la rivoluzione genomica

La capacità di sequenziare genomi ha sorpassato la legge di Moore nel gennaio 2008

Legge di Moore: la performance raddoppia ogni due anni.

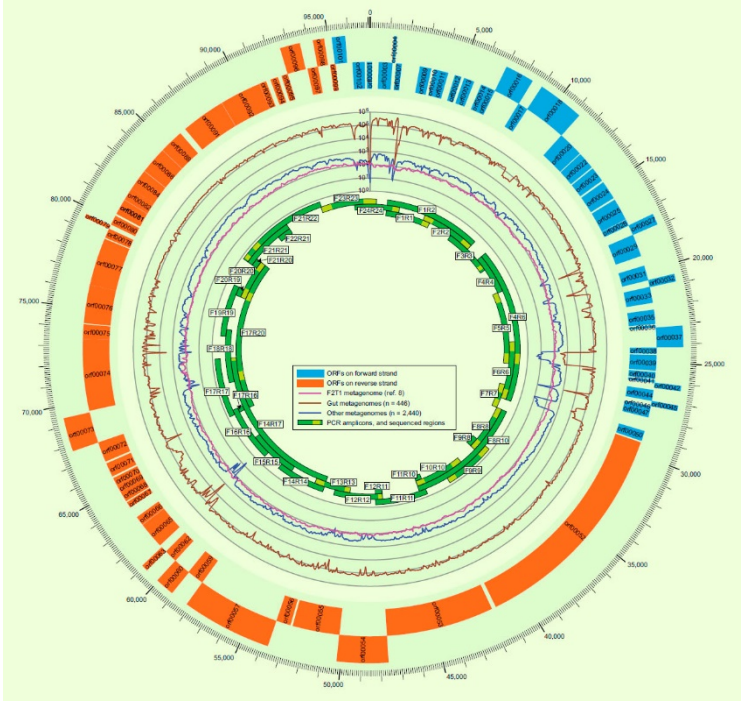
Le tecnologie che mantengono il passo della legge di Moore sono considerate molto positive



L'analisi dei dati prodotti rimane un collo di bottiglia!

Background: origini della metagenomica in ambito virologico

- Nei primi anni 2000, il sequenziamento (insieme ad alter tecniche come la microscopia) è stato applicato su campioni ambientali per la ricerca di acidi nucleici virali (no colture cellulari)
- 2002 → prima identificazione di un virus in un campione di acqua oceanica (Rowher F, San Diego State University)
- 65% di DNA ignoto trovato nello stesso campione ("dark matter")
- Da allora, approcci metagenomici sono stati applicati a molti altri tipi di campioni, sia ambientali che animali (prevalentemente intestine/feci)
- Identificazione di un virus (fago) che costituisce >90% delle sequenze virali: **crAssphage** (100kb di DNA che non allineavano a nulla di già presente nelle banche dati) (Robert Edwards and collaborators)



crAssphage genome

Background – strategie per lo studio del viroma

Microbioma vs Viroma

- **No 16S rRNA** nei virus
- Necessità di utilizzare **marker specifici per gruppi tassonomici diversi**, oppure allineare direttamente con sequenze genomiche di riferimento (es: NCBI Genome Database)
- Virus a RNA non identificabili tramite approcci metagenomica "classica" → necessità di **approcci metatrascrittomici** per identificare sia virus a DNA che a RNA
- Banche dati molto più ricche di sequenze di virus a DNA

Background - limitazioni

- Purificazione del contenuto virale di un campione
  (Perdita di materiale durante filtrazione e centrifugazione (grandi virus a DNA))

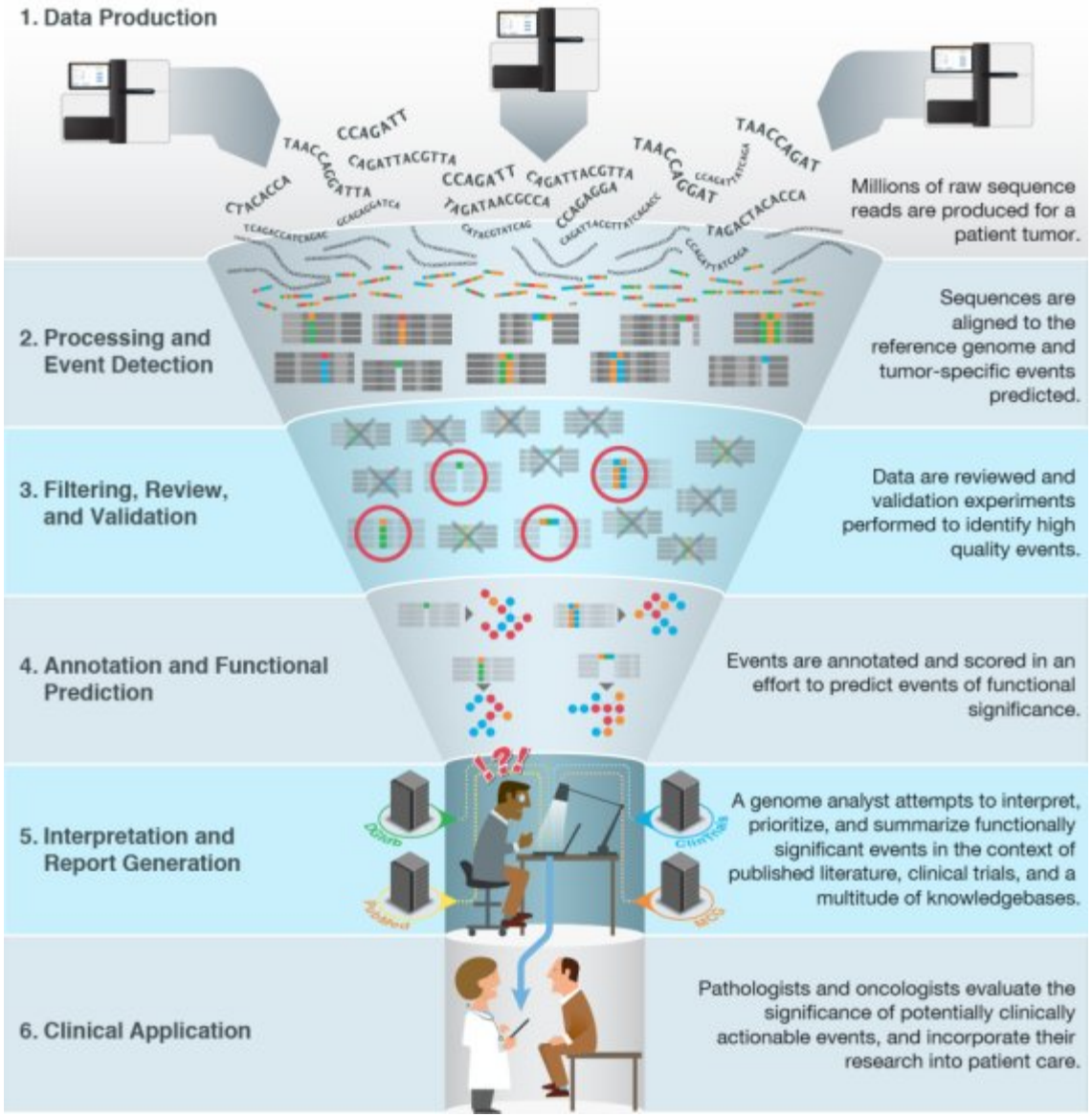
- Virus integrati/latenti (anche profagi, virus episomali)
  → spesso vengono persi

- Le sequenze (**corte**) ottenute spesso hanno origine incerta (virus? Batteri? Contaminanti?)
  Necessità di assemblare le **read** in sequenze più lunghe (**contig**)

- In alternativa si può evitare lo step di filtraggio → ma… poche read saranno virali! (< 0,01%)

- Contaminazioni sono frequenti (campioni, reagenti, linee cellulari)

Background – problemi bioinformatici



Esiste una **moltitudine di tool bioinformatici** per le applicazioni più svariate, ma:

- Necessitano elevato expertise bioinformatico e risorse computazionali
- Inaccessibili alla maggioranza dei potenziali utenti
- Molto spesso non sviluppati per applicazioni virologiche
- Difficile scegliere il tool da usare tra i molti disponibili
- Difficile comparare l'efficienza dei diversi tool

Sono necessari:

- **Tool robusti**, intuitivi, che non richiedano troppe risorse computazionali e di connessione (non disponibili in regioni di campionamento remote)

A che punto siamo nell'esplorazione del viroma?

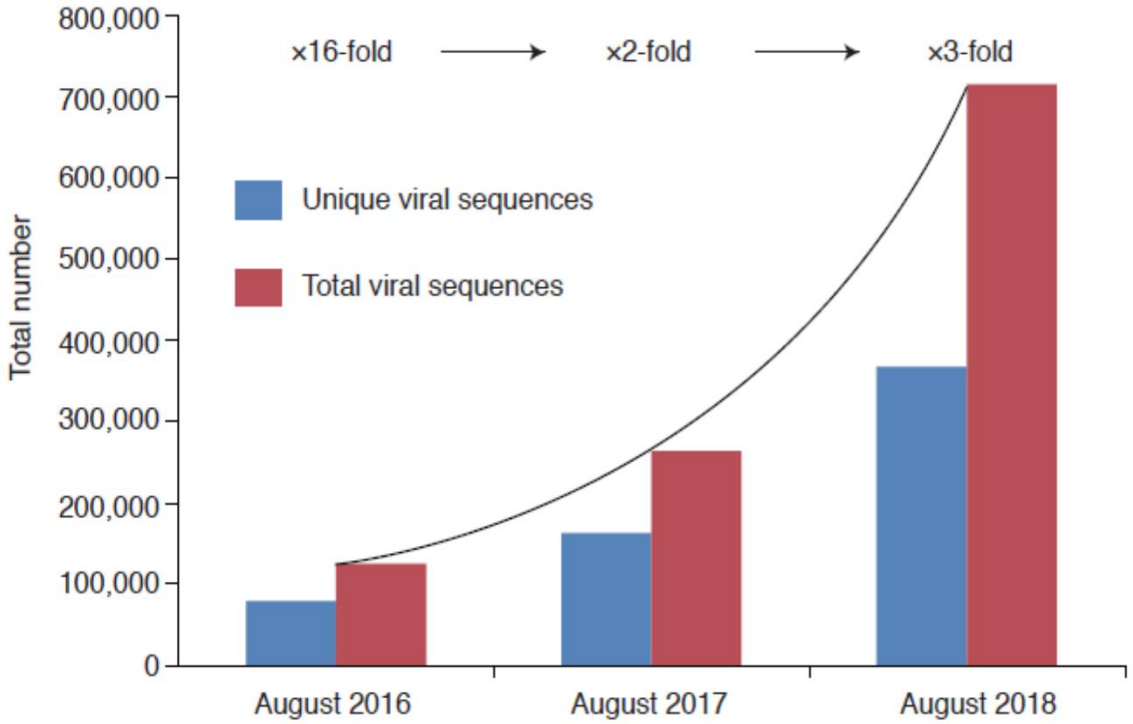


Figure 1 Growth rate of virus identification and microbial host prediction. Growth over time in the total and unique number of viral sequences in the January 2018 release of the IMG database. The first data point represents a 16-fold increase in the number of species in comparison to the number of previously identified viruses. Subsequent points are twofold and threefold higher, respectively (image by David Paez Espino, JGI).

IMG: Integrated Microbial Genomes (IMG) database (Dept. Of Energy, US)

Lo sviluppo software è in piena espansione

Sommario

- La nuova accessibilità del viroma, grazie allo sviluppo delle piattaforme NGS

- La caratterizzazione del viroma attraverso NGS:

  - Le piattaforme di sequenziamento

  - Metagenomica e metatrascrittomica **shotgun:** who is there?

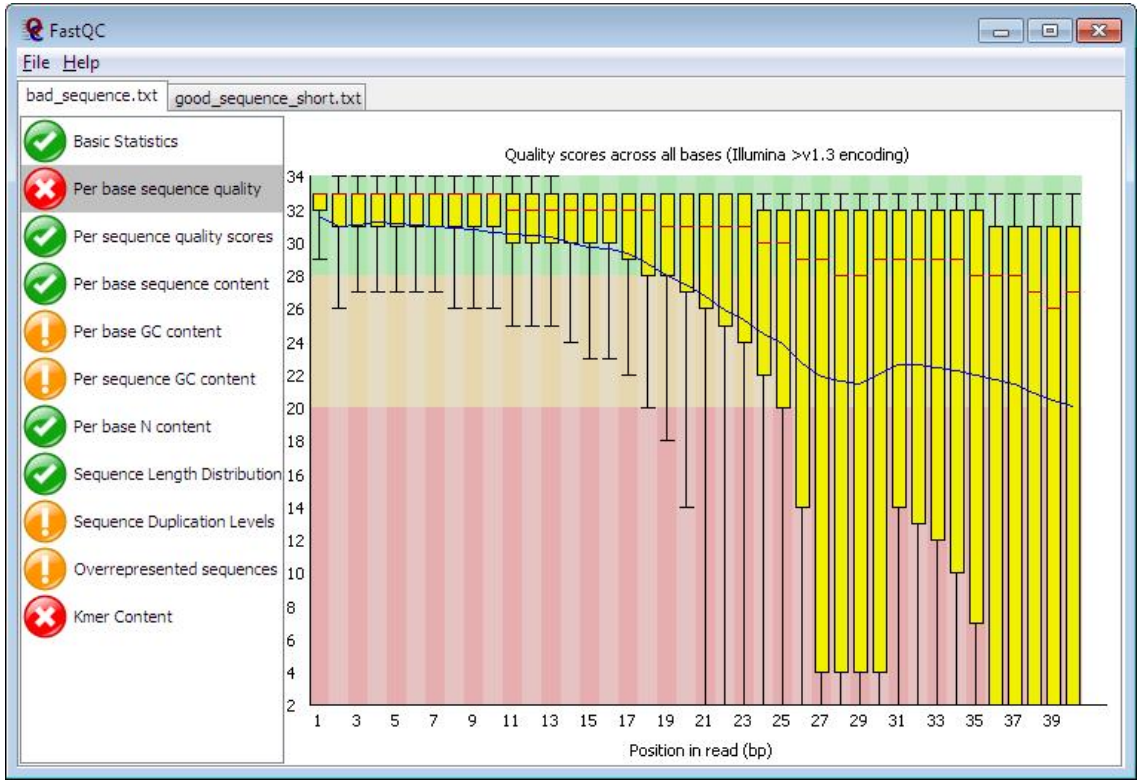
  - Deep sequencing mirato: genotipizzazione, farmaco-resistenze, identificazione di varianti minoritarie

Scegliere la piattaforma di sequenziamento

	Piattaforma	Throughput	Lunghezza reads	Commenti
Sì PCR	Illumina	10^9 reads	150 nt paired-end (max 300 nt)	Grande throughput, ma sequenze corte
	Ion Torrent	10^6 reads	Fino a 600 nt	Ridotto throughput ma sequenze più lunghe
No PCR	PacBio	10^5 reads	>10k nt	Sequenze molto lunghe senza necessità di amplificazione Tasso di errore elevato (>10%) Costi
	Oxford Nanopore	10^5 reads	>10k nt	Sequenze molto lunghe senza necessità di amplificazione Tasso di errore ancora relativamente alto Portabilità

Lunghezza, **qualità** e numerosità (**profondità di sequenziamento**) delle sequenze prodotte sono fattori che influenzano enormemente le strategie di analisi e il risultato finale delle stesse

Controllo qualità delle read



Potenziali problemi:

- Artefatti di sequenziamento (adattatori, primer)
- Errori di sequenziamento (e.g. base miscalls)

Soluzione:

- Trimming delle reads o completa eliminazione

Metagenomica ‘Shotgun’

Caratteristiche specifiche della metagenomica virologica:

- Presenza preponderante di sequenze dell’ospite
- Necessità di identificare le sequenze di origine virale
- Assemblaggio o mappaggio delle reads virali per ricostruire i genomi full-length
- Identificazione tassonomica deli genomi virali (o frammenti) identificati

Arricchimento degli acidi nucleici virali

Prima del sequenziamento:
filtrazione e centrifugazione, DNAse e RNAse, amplificazione PCR

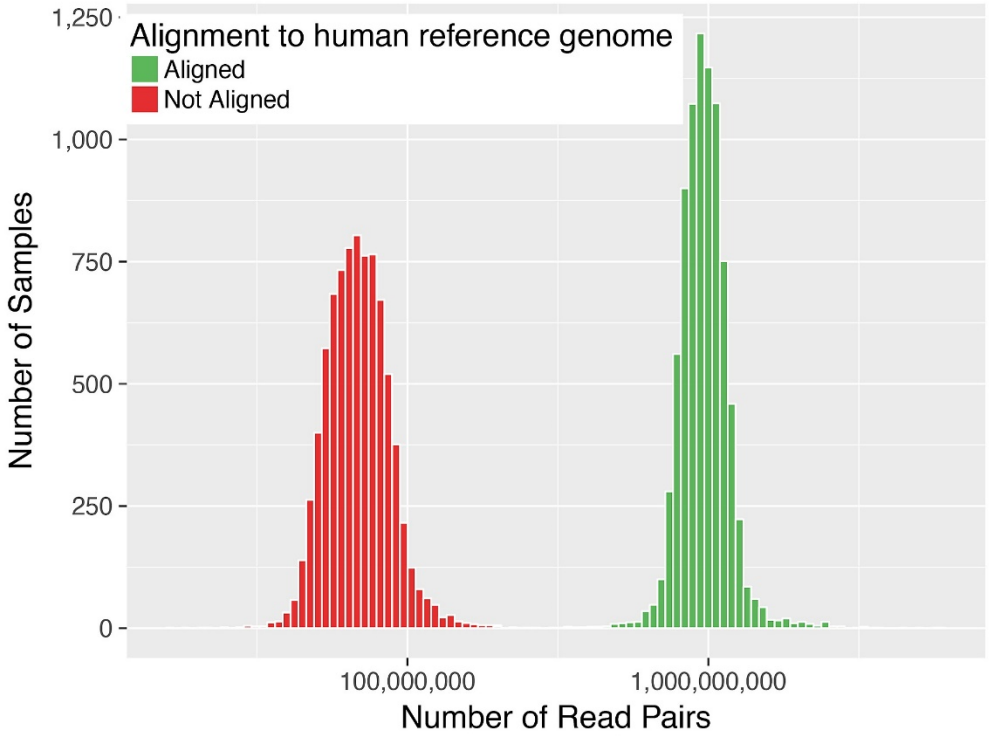
Problemi:

- perdita di virus grandi, di virus integrati e latenti
- Disegno dei primer complesso in caso di alta variabilità delle specie target

***In silico*:** filtraggio delle read dell'ospite attraverso allineamento vs un genoma reference

Problemi:

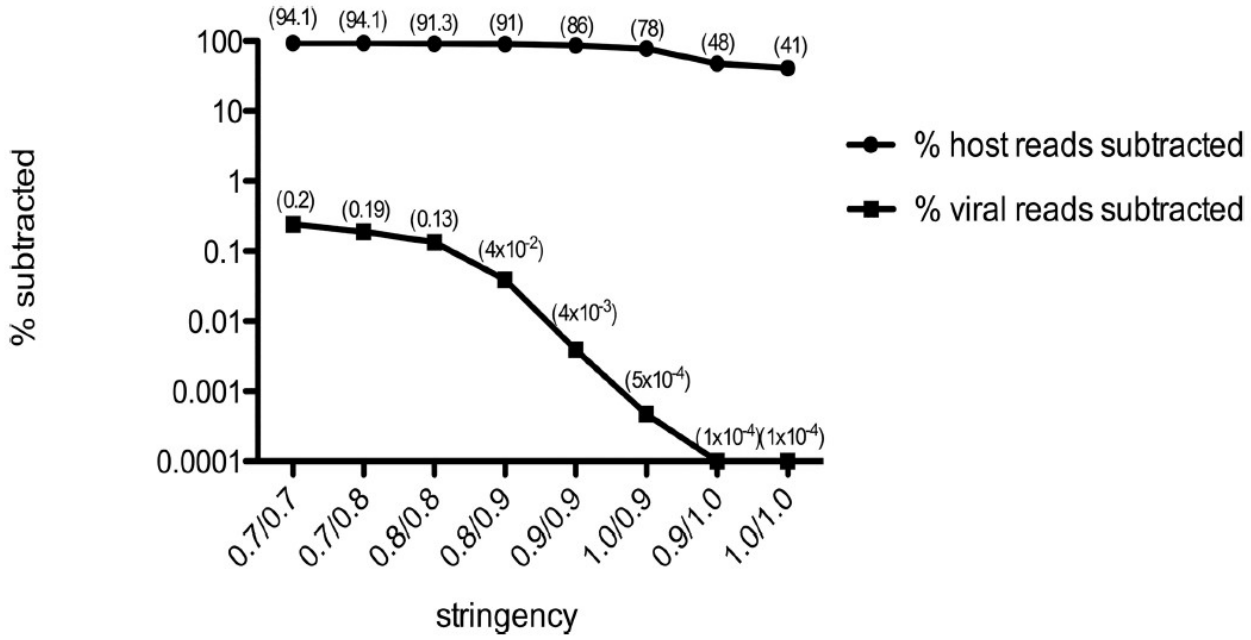
- Perdita di virus integrati se non annotati correttamente



Mediamente solo il 5% delle reads non mappa sul genoma ospite, e solo una piccolissima parte è di origine virale (<0.01%)
(Moustafa et al., Plos pathogens 2017)

Filtraggio delle read dell'ospite: quale stringenza?

Non solo sequenze del genoma ospite, ma anche rRNA, mtRNA, mRNA, sequenze batteriche o fungine o altri genomi non umani



La rimozione delle read dell'ospite richiede una scelta di parametri di stringenza dei tool di allineamento:

- Quale percentuale di similarità?
- Quale percentuale di lunghezza della read?

Necessità di un tradeoff → perdita di sequenze virali e mantenimento di sequenze dell'ospite

Altri potenziali problemi → presenza di **GTA** (gene-transfer agents), *viral-like particles* contenenti frammenti di genoma non virale (fagi)

Approccio 'Shotgun': mapping o *de novo* assembly?

Allineamento delle read su una (o molte) sequenza di riferimento

(Mapping)

Metodo computazionalmente efficiente

In genere ricostruzioni più omogenee

Assemblaggio delle read

(*De novo* assembly)

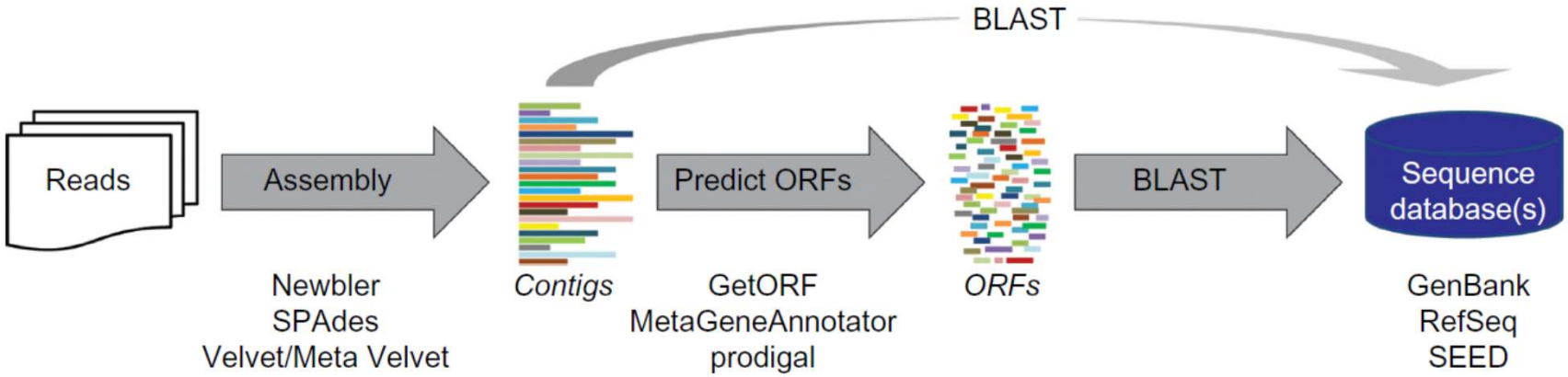
Necessario se il genoma target è ignoto o poco caratterizzato

Metodo computazionalmente esoso

Ricostruzioni più frammentate

Con entrambe le strategie si punta ad ottenere una sequenza consenso del (o dei) genoma target → mitigazione dell'errore di sequenziamento presente sulle singole read

Assemblaggio delle read de novo



Tool bioinformatici per l'assemblaggio *de novo* di metagenomi:

Generici

- Omega (Haider et al. 2014)
- Genovo (Laserson, Jojic, and Koller 2011)
- MEGAHIT (Li et al. 2015)
- MetaSPAdes (Nurk et al. 2016)
- Ray-Meta (Boisvert et al. 2012)
- MetAMOS (Treangen et al. 2013)
- MetaVelvet (Namiki et al. 2012; Afiahayati, Sato, and Sakakibara 2015)
- IDBA-UD (Peng et al. 2012)

Per ricostruzione aplotipi virali

- ShoRAH (Zagordi et al. 2011)
- PredictHaplo (Giallonardo et al. 2014)
- QuRe (Prosperi and Salemi 2012)

L'assemblaggio di genomi multipli è un processo **complesso** (possibilità di generare assemblaggi chimerici)

Gli assemblatori tendono a **ridurre la complessità** collassando la variabilità (strain diversi assemblati in una sequenza consenso)

Molto **problematica** la ricostruzione di **genomi virali a RNA** molto variabili

Profondità di sequenziamento molto variabile (intra- e inter-genomi, errori vs mutazioni)

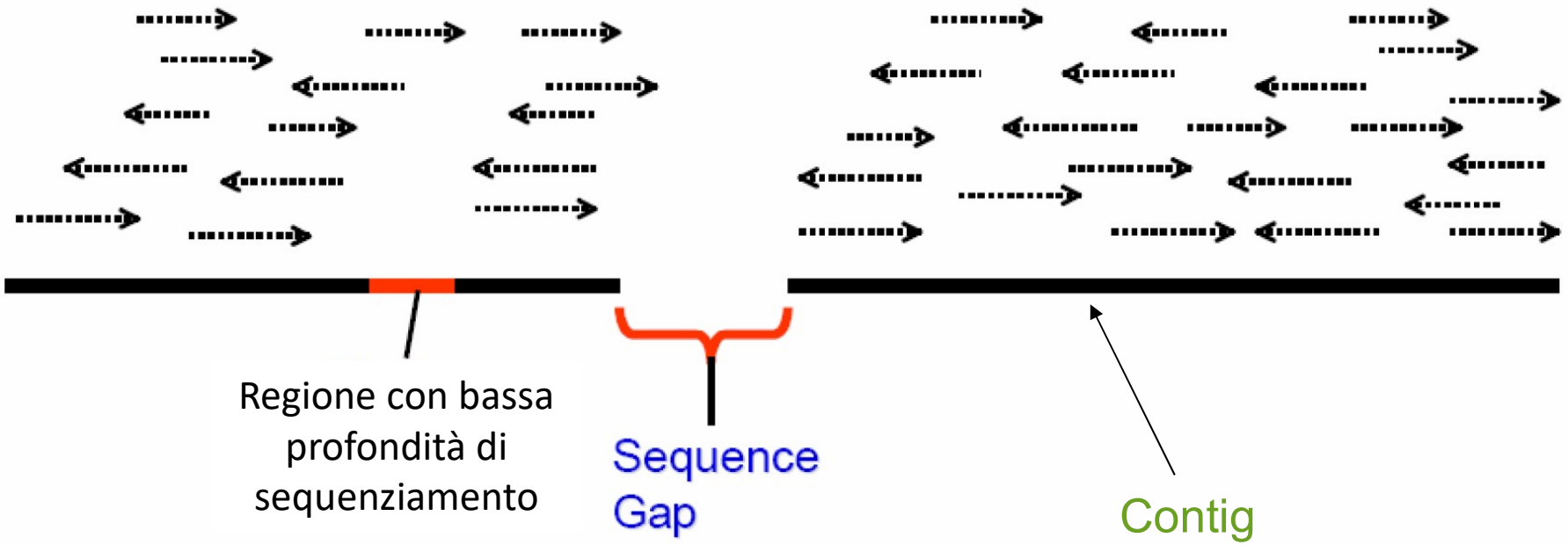
Molto sensibili alla qualità del campione in input (estrazione, arricchimento, preparazione delle librerie, …)

Profondità di sequenziamento disomogenea

Nella tecnica shotgun vengono sequenziati frammenti casuali della libreria

Essendo il campione molto eterogeneo, alcune regioni di alcuni genomi sono sequenziati con bassa profondità

Altri non sono sequenziati affatto



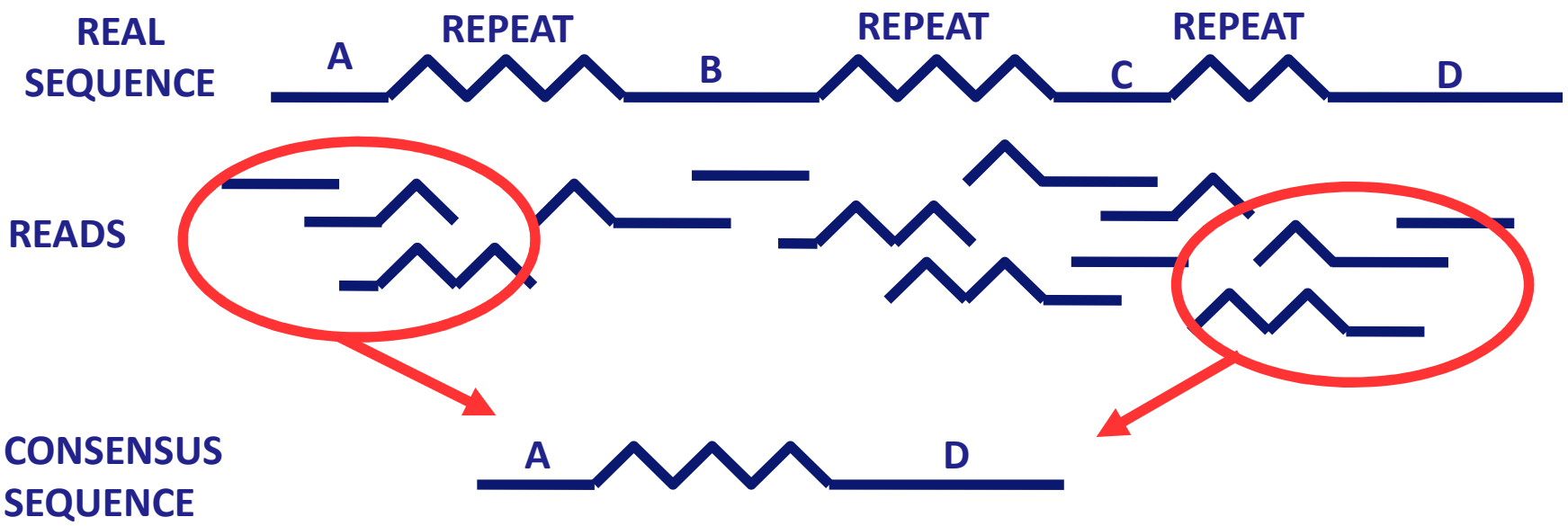
Generalmente per un buon assemblaggio de novo è necessario un sequenziamento 10x (10 volte la lunghezza del genoma)

Nelle regioni a bassa profondità di sequenziamento è difficile distinguere mutazioni da errori introdotti in fase di amplificazione

Assemblaggio dei dati: il problema dei repeats

Reads overlappanti non necessariamente provengono da regioni contigue di un genoma

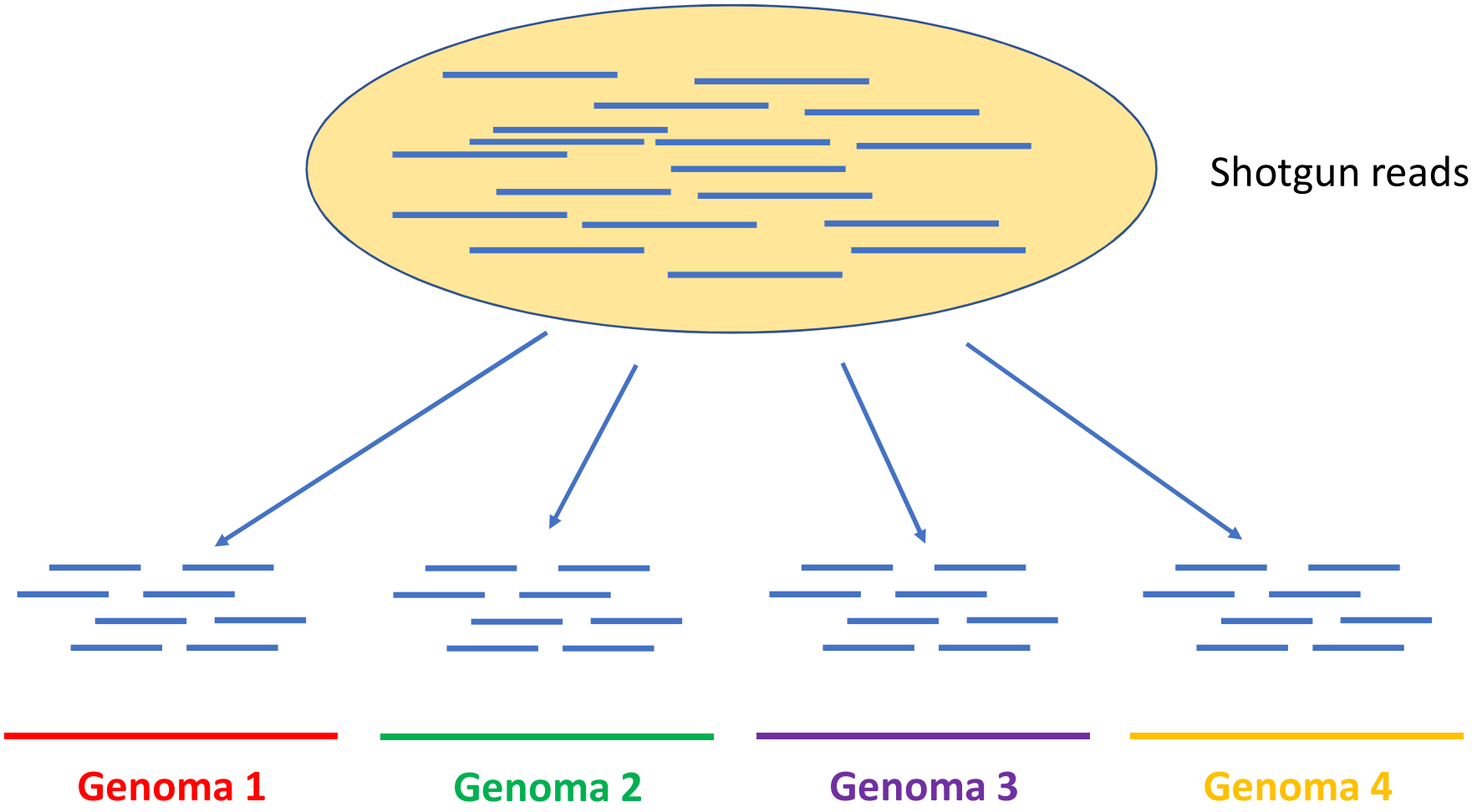
Regioni molto simili in genomi diversi (virus diversi) possono produrre **assemblaggi chimerici**



Gli assemblatori genomici in genere non fondono le reads “multi-mappers”

Gli assemblatori metagenomici spesso lo fanno (profondità di sequenziamento bassa o disomogenea)

Mapping delle read

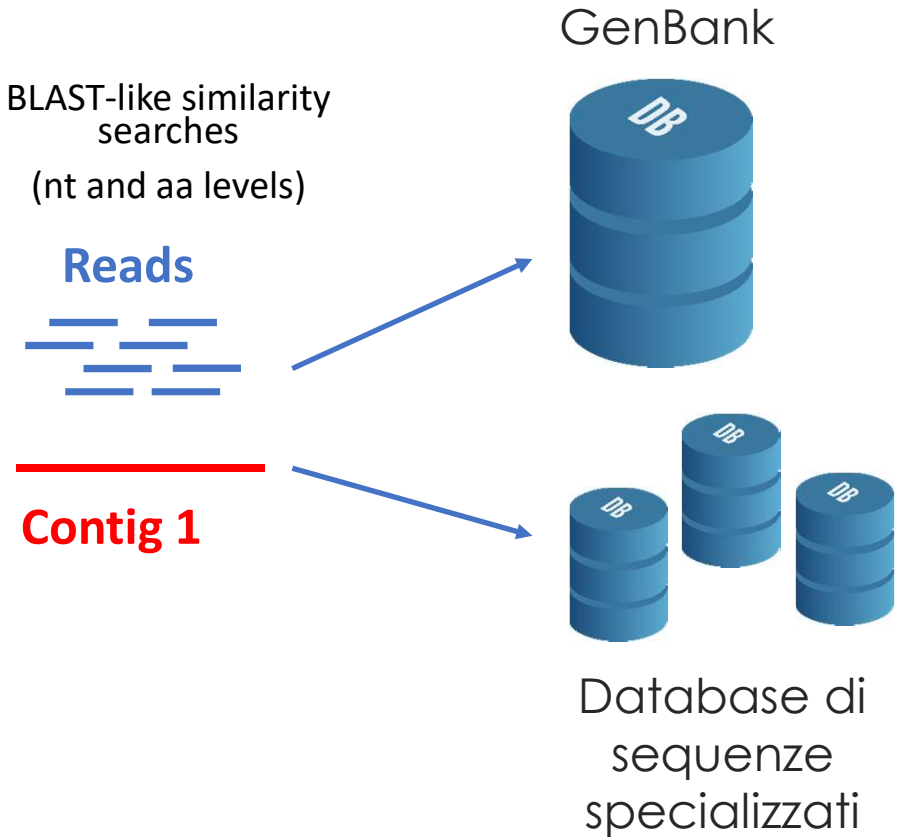


Scarsa disponibilità di reference e di varianti sub-specie delle stesse possono impedire l'ottenimento di mappaggi soddisfacenti

Mapping e *de novo* assembly possono essere usati insieme

Classificazione delle sequenze: allineamento

Classificazione → ricerca di similarità tra una read (o un contig) e un database di sequenze di riferimento annotate



- Velocità VS Sensibilità
- Elevate risorse computazionali
- Output complesso, necessità di ulteriori analisi (molti dati)
- Richiesta integrazione di tool diversi (application-specific pipelines)
- Complessità di installazione e configurazione, necessità di competenze bioinformatiche

Tool ottimizzati per la classificazione di sequenze virali:

Kraken (Wood and Salzberg 2014)(k-mer based)
Clark (Ounitet al. 2015)
One Codex (Minot, Krumm, and Greenfield)
Lambda (Hauswedell, Singer, and Reinert 2014)
Diamond (Buchfink, Xie, and Huson 2015)

Classificazione delle sequenze alignment-free

- Poche sequenze virali annotate nei database
- Alta diversità all'interno delle famiglie virali
- Mancanza di marker genes (es: 16S nei batteri)

→ Difficoltà nel trovare parametri di similarità validi per tutti i virus presenti in un campione

Metodi che non si basano sulla ricerca di similarità di sequenza

PhyloPythia (McHardy et al. 2007) → usa la frequenza dei nucleotidi per classificare le reads

PHYMM (Brady and Salzberg 2009) → usa Markov models per cercare oligonucleotidi caratteristici di alcune specie (NCBI RefSeq database)

PHYMMBL (Brady and Salzberg 2011) → combina PHYMM and BLAST

Altri potenziali approcci → frequenza di di-nucleotidi, codon usage, domini proteici (anche piccoli) conservati all'interno di famiglie di virus

Detection di nuovi genomi virali full-length (potenziali patogeni)

Assemblaggio delle read in contig

↓

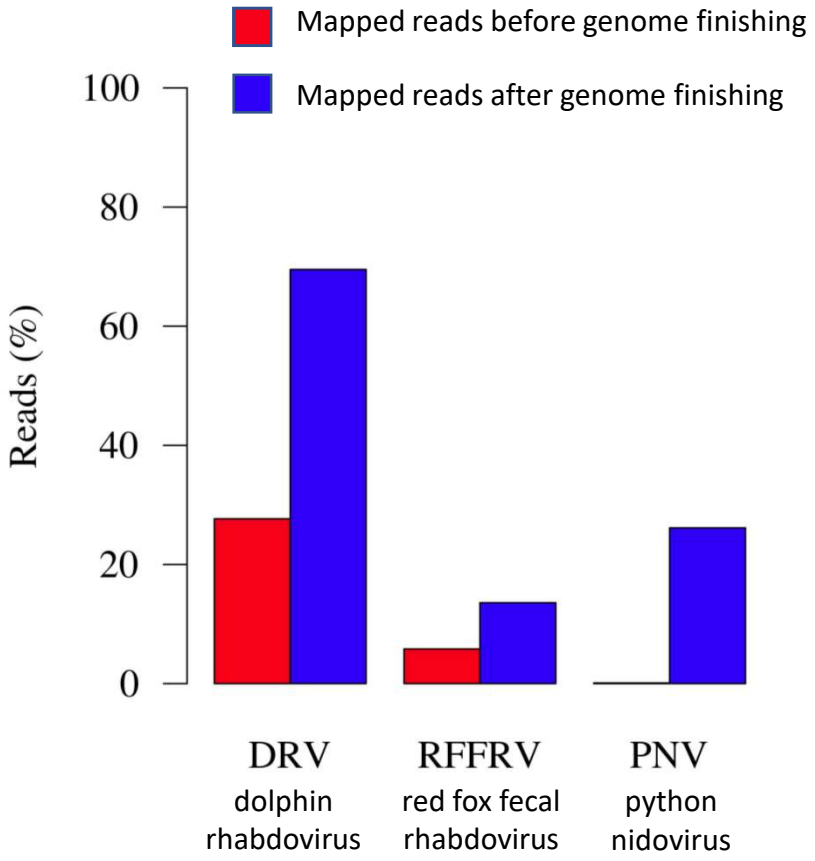
Determinazione dell'origine virale di almeno un contig

↓

Linkage di altri contig a questo contig 'seed'

↓

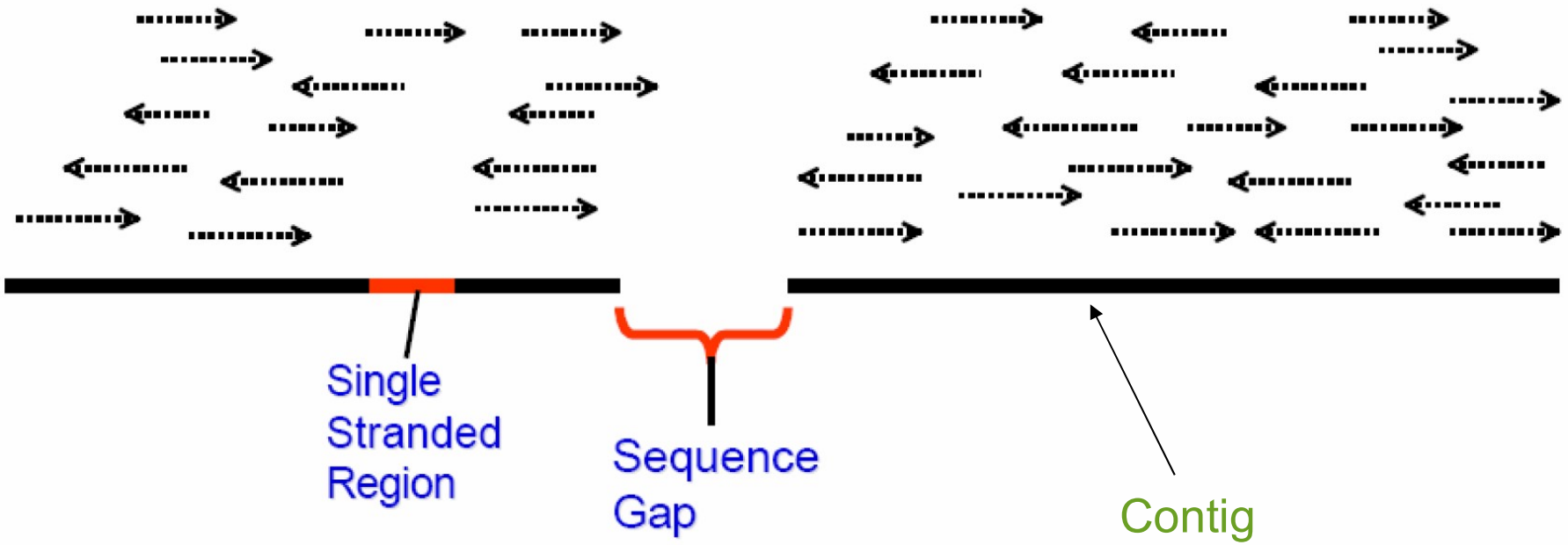
Finishing del genoma virale



L'efficienza dei metodi di classificazione basati su allineamento di assegnare le sequenze virali alla specie di appartenenza variano molto in presenza o assenza di un genoma di riferimento completo

Assemblaggio di un nuovo virus (full length)

A partire dalle read corte, creazione di contig più lunghi attraverso fusione di read overlappanti

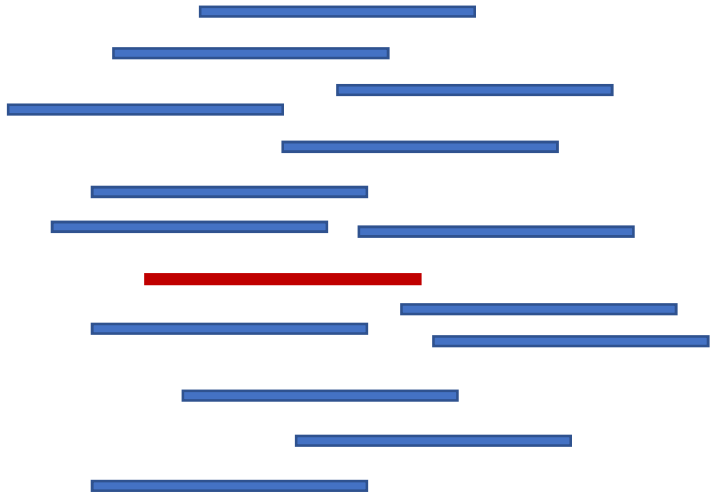


- Pochi tool bioinformatici sviluppati per genomi virali (maggior tasso di mutazioni)
- **Assemblaggio iterativo** produce discreti risultati (es: PRICE, Grard et al., 2009; IVA, Hunt et al., 2012)

Identificazione di un frammento "seed"

Obiettivo:

- identificare almeno un contig appartenente al virus ignoto che si sta cercando di identificare.



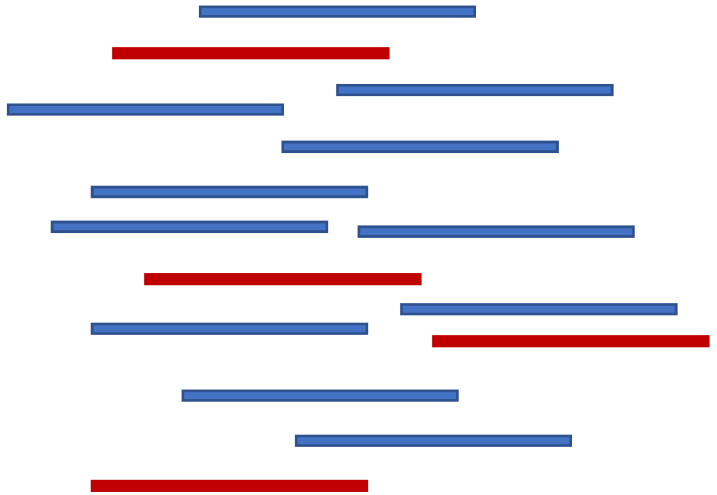
Metodi:

- ricerca di similarità, anche debole, con altre sequenze di virus noti (nt e aa)

- in alternativa (nessuna similarità): si parte dai contig ignoti più lunghi o che contengono più reads (indicazione di presenza del microorganismo in grande quantità)

Linkage dei frammenti

Obiettivo: collegare frammenti di sequenze appartenenti allo stesso virus ma che non sono stati assemblati insieme (es: profondità di sequenziamento non omogenea, virus segmentato)



Metodi:

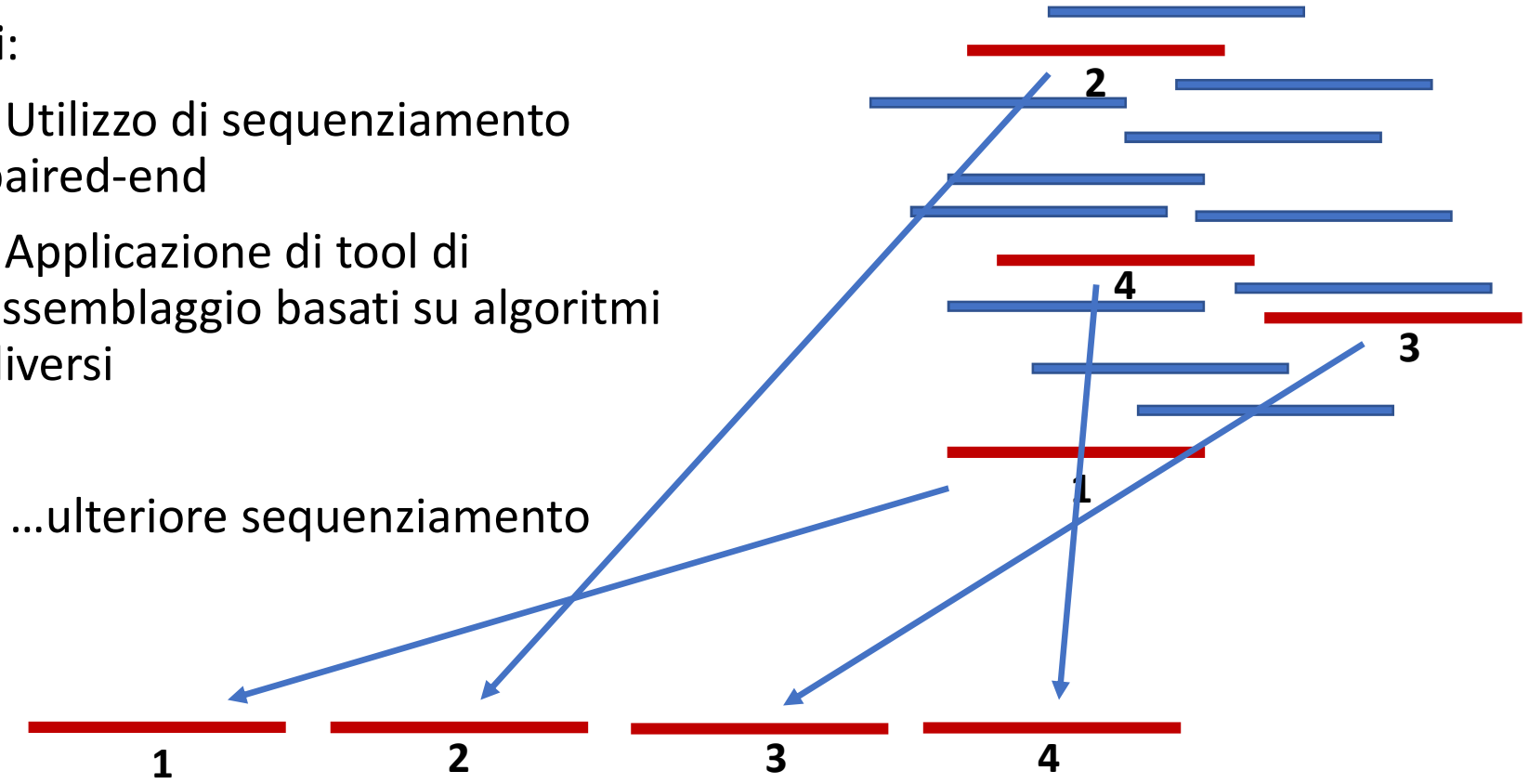
- ricerca di **motivi conservati** su frammenti diversi (es: motivi di inizio/fine trascrizione)
- ricerca di frammenti con coverage simile (**coverage profile binning**)
- valutazione della frequenza de tutti le possibili sequenze di lunghezza k all'interno di una sequenza di DNA (**k-mer profiling**)
- ricerca di similarità di sequenza remote (**remote homology detection**)

Finishing del genoma

Obiettivo: ridurre il numero di contig (idealmente) a 1

Metodi:

- Utilizzo di sequenziamento paired-end
- Applicazione di tool di assemblaggio basati su algoritmi diversi
- …ulteriore sequenziamento



(Ultra) Deep sequencing mirato

Approccio che prevede amplificazione di una specifica regione genomica e sequenziamento ad altissima profondità

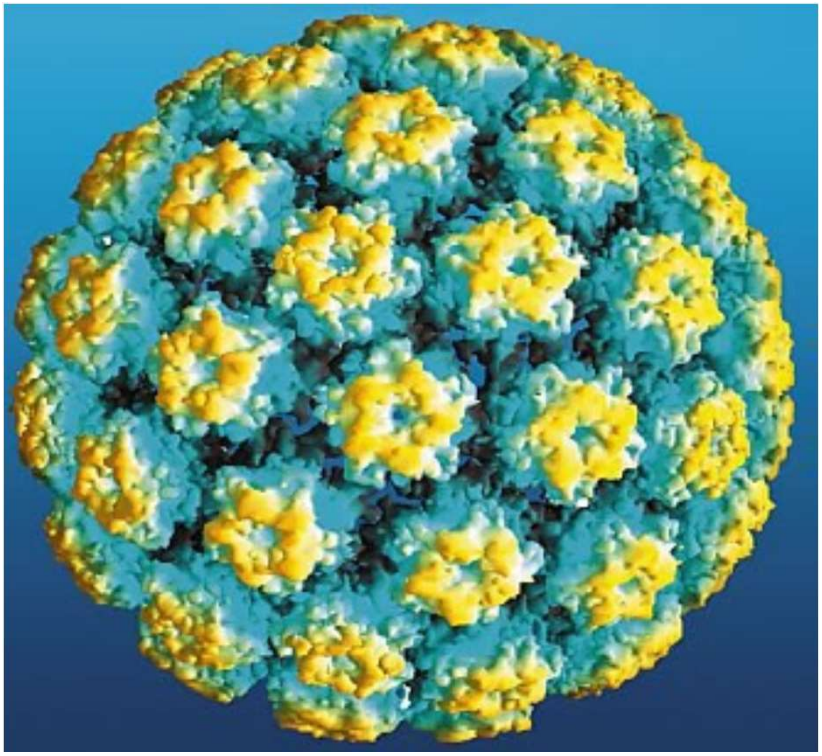
Consente di rilevare mutazione presenti a livelli estremamente bassi che possono raggiungere 1%

Metodo utile per identificare mutazioni somatiche a bassa frequenza in campioni tumorali o, ad esempio, scoprire variant virali rare (es: HIV)

Deep sequencing mirato: applicazioni

- HIV:
  - genotipizzazione del tropismo di HIV relativamente ai co-recettori CCR5 and CXCR4
- HCV:
  - Deep sequencing di NS3, NS4A, NS5 per la caratterizzazione delle farmaco-resistenze ai DAA (direct-acting antivirals), anche in aplotipi minoritari del virus
  - Genotipizzazione basata su NS5B, 5'-UTR, Core
- HBV:
  - Genotipizzazione e farmacoresistenze (Core e pre-core)

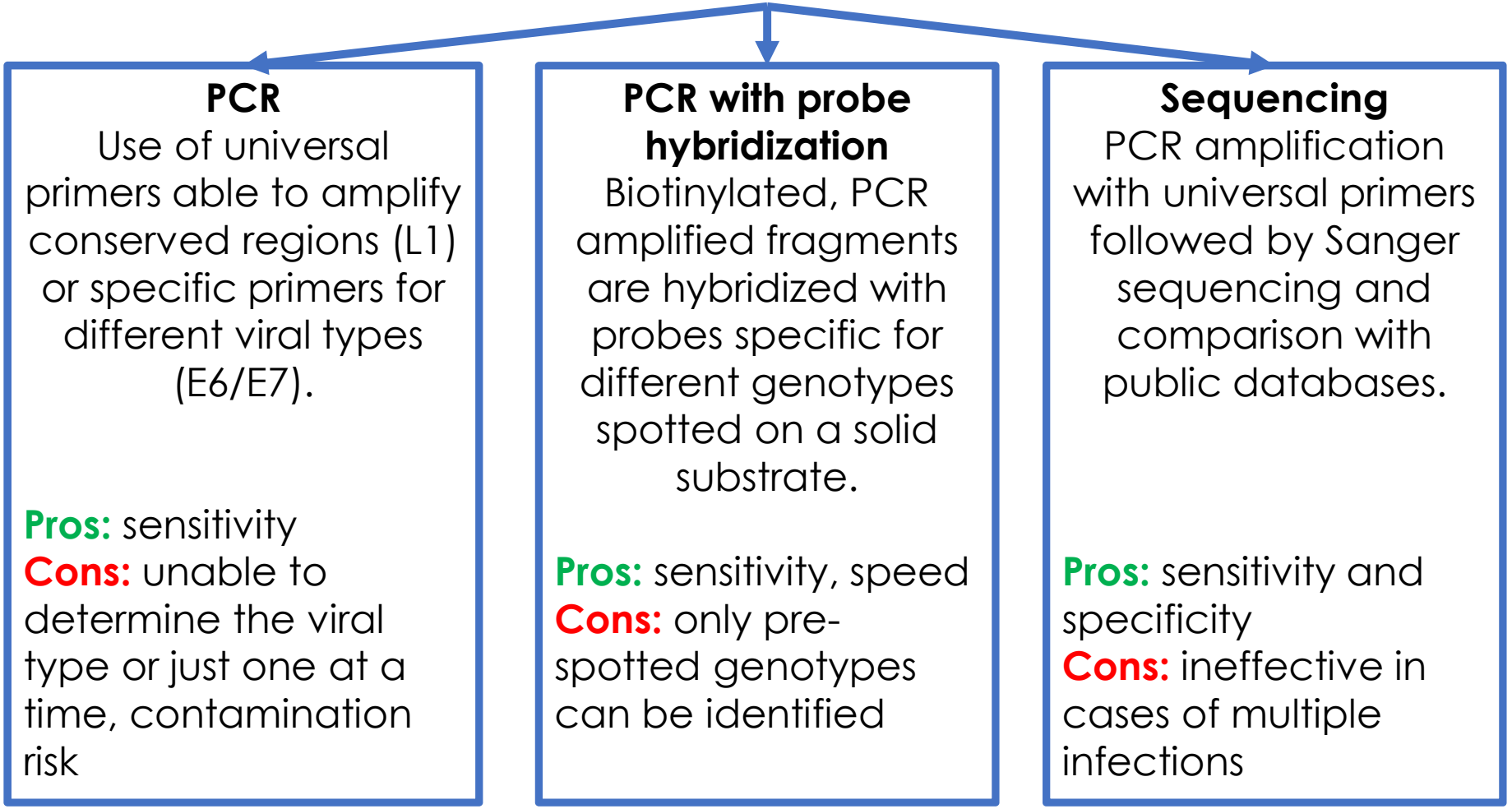
Human papillomavirus (HPV)



- Piccolo virus (55nm), genoma a dsDNA (~8kb)
- Trasmissione attraverso rapporti sessuali, tropismo preferenziale per cellule dell'epitelio basale (pelle, mucose)
- **Più di 200 genotipi umani** già noti (>10% divergenza nucleotidica)
- Infezione persistente da parte di **genotipi ad alto rischio** può causare **cancro** alla **cervice** e altri tumori **testa-collo**
- Altri genotipi causano lesioni più comuni (verruche a mani, piedi, genitali, condilomi)

HPV laboratory diagnosis

Based on viral nucleic acids detection in clinical samples (swabs, fresh or paraffin embedded biopsies)



HPV laboratory diagnosis

Based on viral nucleic acids detection in clinical samples (swabs, fresh or paraffin embedded biopsies)

(E6/E7).

Pros: sensitivity
Cons: unable to determine the viral type or just one at a time, contamination risk

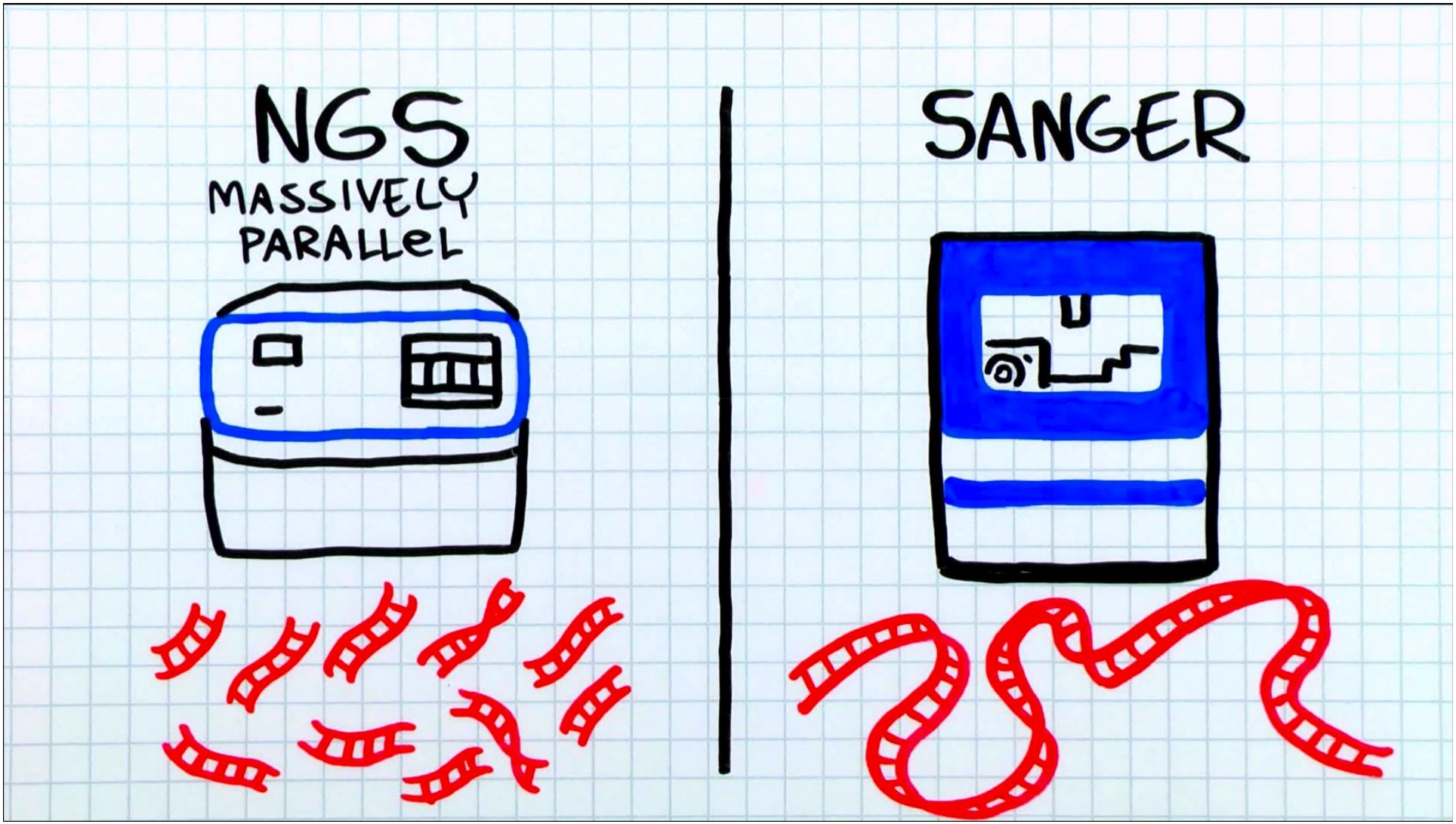
different genotypes spotted on a solid substrate.

Pros: sensitivity, speed
Cons: only pre-spotted genotypes can be identified

public databases.

Pros: sensitivity and specificity
Cons: ineffective in cases of multiple infections

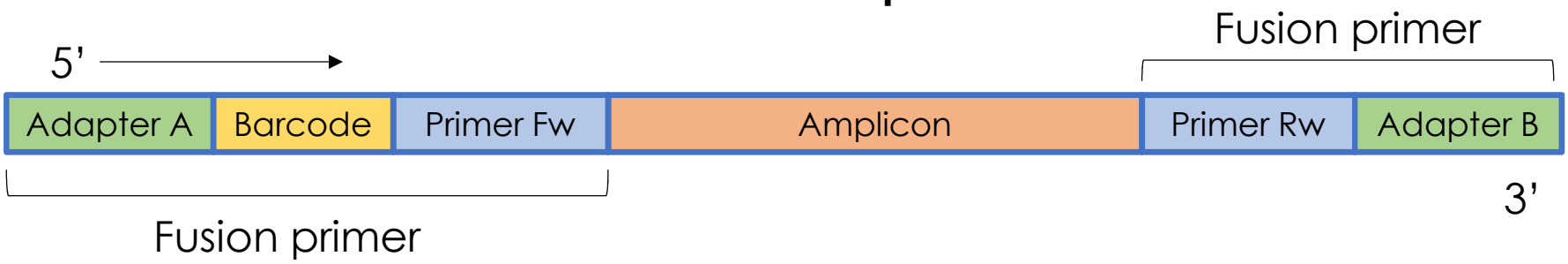
Excellent opportunity to exploit NGS!



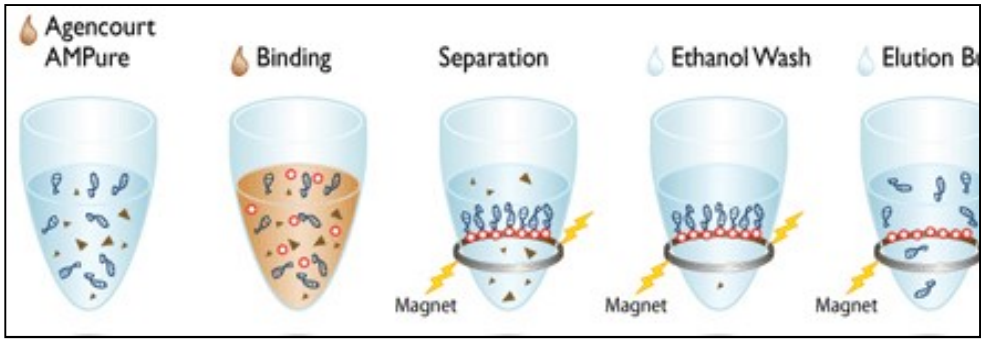
Le piattaforme NGS sequenziano i singoli frammenti, permettendo di identificare genotipi/varianti diverse nello stesso campione

Il metodo Sanger legge un pool di frammenti per ogni campione, restituendo una sequenza consenso

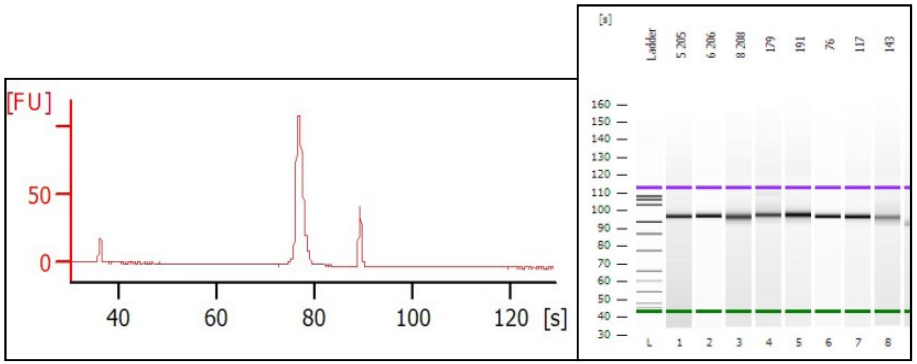
Generation of PCR amplicons



Amplicon purification



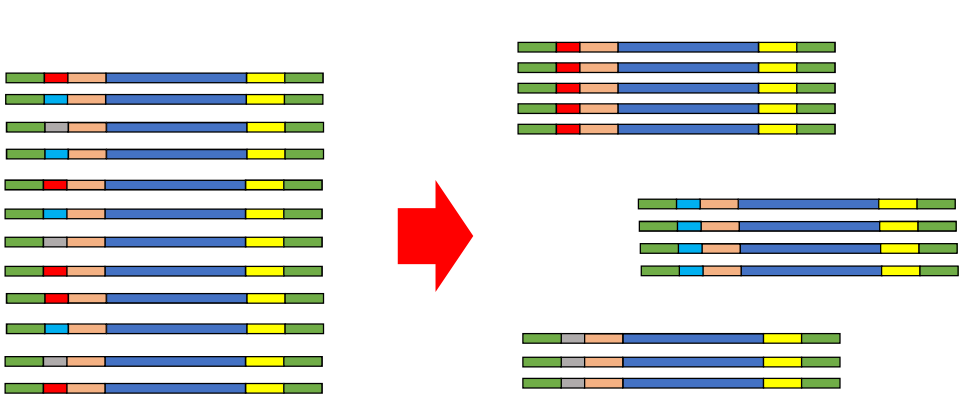
Quality control and quantification

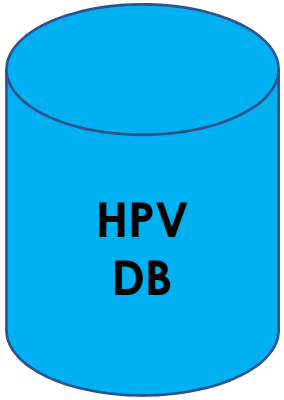


Sequencing

164 samples (10 negative controls)

Data analysis





DEMULTIPLEXING AND PRIMER TRIMMING

GENERATION OF A CUSTOM DATABASE WITH REFERENCE SEQUENCES

BLAST ALIGNMENT AND AUTOMATIC EVALUATION OF RESULTS

REQUIREMENTS FOR A READ TO BE ASSIGNED TO A SPECIFIC GENOTYPE:

- Alignment length >= 90% of read length (to avoid spurious short alignments)
- At least 90% of sequence identity, in agreement with HPV classification guidelines

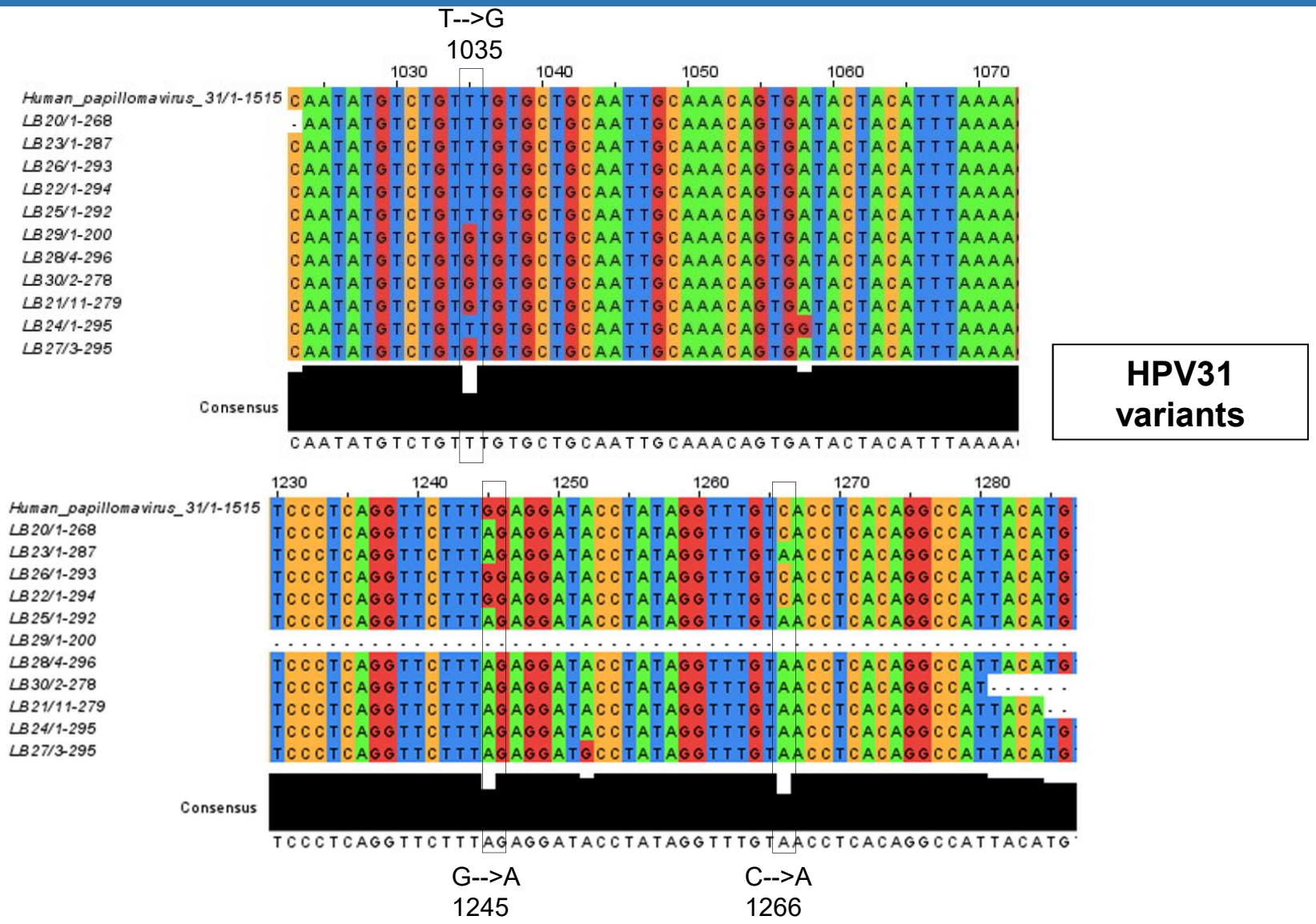
Risultati

- Il DNA di HPV è stato rilevato in tutti i 154 campioni, e tutti sono stati genotipizzati

- Confronto vs. Sanger
  - 100% conconrdanza dei risultati nelle infezioni singole (66% compatibilità nelle infezioni multiple)

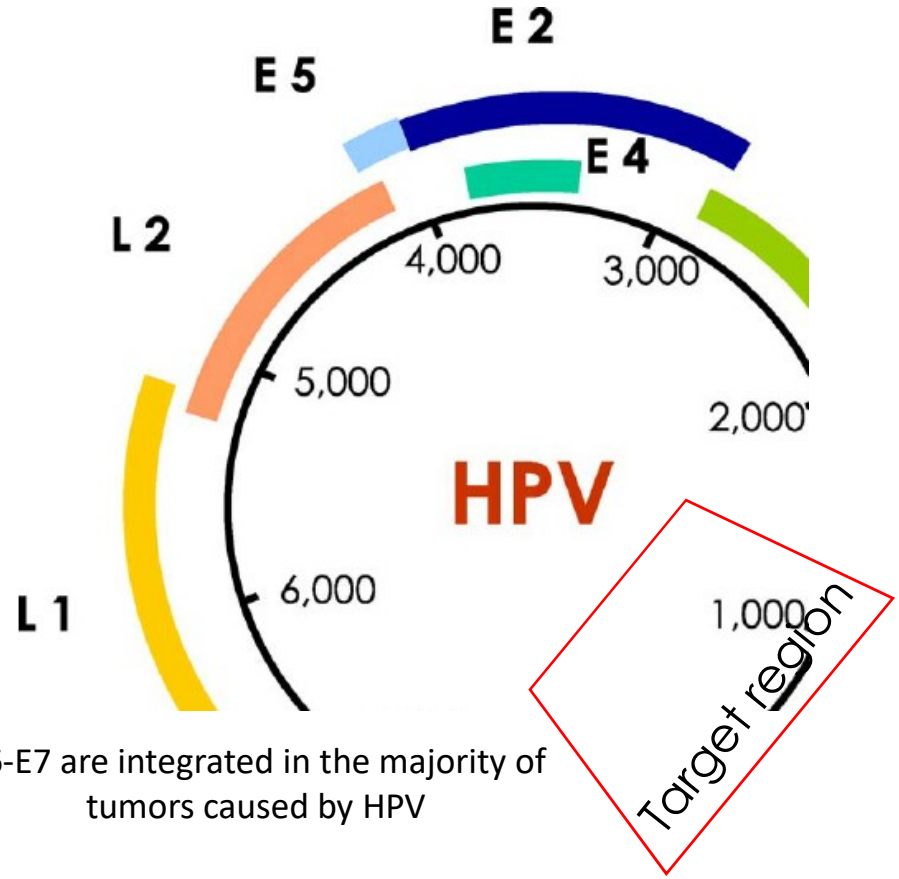
- Confronto vs. saggi di ibridazione
  - Buona concordanza sui genotipi target (85-99%)
  - Maggiore sensibilità delle NGS nelle infezioni multiple
  - Possibilità, per il metodo NGS, di identificare varianti intra-genotipo (spesso associate a diversi fenotipi clinici)

Scansione delle sequenze per rilevare varianti nucleotidiche



Barzon et al., J Clin Virol 2011
Militello & Lavezzo, Clin Microbiol Infect 2013

Primer design

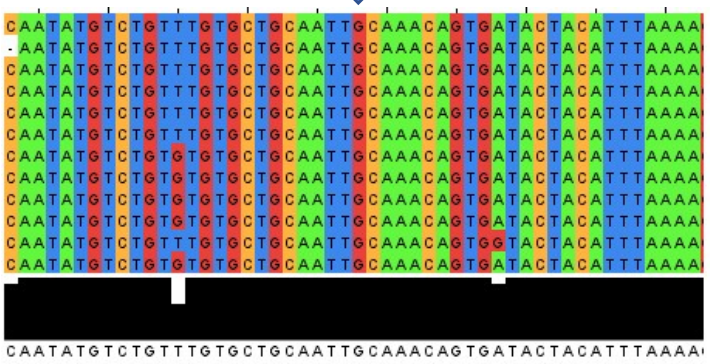


E6-E7 are integrated in the majority of tumors caused by HPV

Generation of all possible haplotypes (2^n) and primer design with Hyden

Download of all the available HPV sequences

HPV type	No. of available sequences
51	6
56	126
18	130
39	25
45	104
59	13
68	92
16	1897
31	276
33	82
35	98
52	363
58	795



Intra-type multiple alignments and SNPs detection

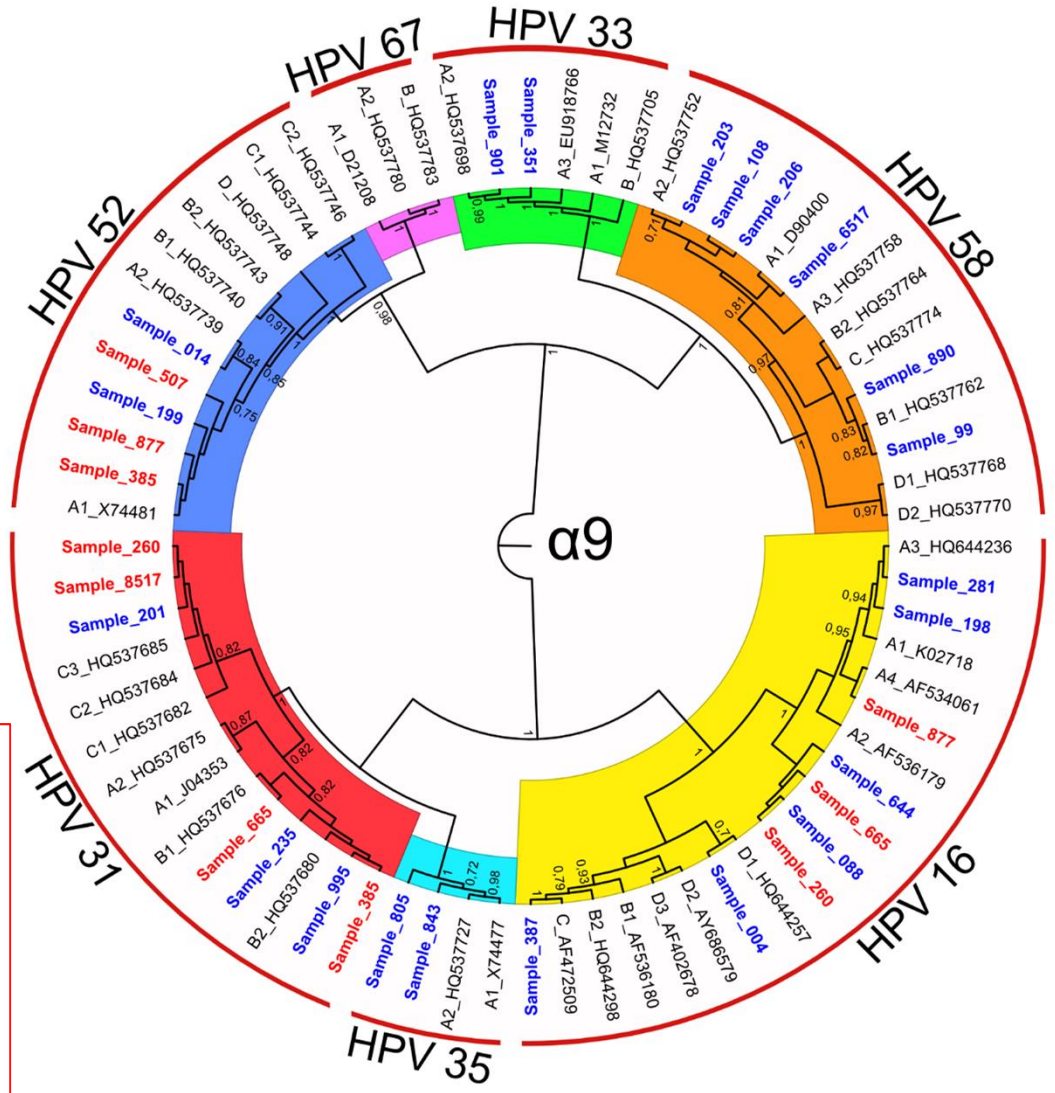
HPV Species	Primer Type	Primer Sequence *	HPV Type and Variant §	Amplicon Length (HPV Type)
α 5	F	AACCGAAAAGGGTTATGACCGA	51 (all known variants)	815
	R	TCTGCTGTACAACGCGAAGG		
α 6	F	AGGCAGCTTATTCTGTGTGGA	56 (all known variants)	873
	R	CAGAGTGGGCACGTTACTGT		
α 7	F	AGGGAGTRACCRAAAACGGT	18, 39 (all known variants)	814 (HPV18) 807 (HPV39)
	R	GGAATGCTCGAAGGTCGTCT	18 (all known variants)	
	R	CCCGTGAGGCTTCTACTACC	39 (all known variants)	
	F	TGCAACCAAAAACGGTGCAT	45 (all known variants)	849
	R	TTAGTTGCACACCACGGACA		
	F	GCATGGCACGCTTTGAGG	59 (all known variants)	806
	R	GTTTGCTGCACACAAAGGACA		
	F	GGTCACGACCGAAAACGG	68 (A2, B, D1, D2, E, F1, F2)	815
	R	AGCAGYTSYAGCTTCCGCA	68 (C1, C2, D1, D2, E, F1, F2)	
	F	GKGACCGRAARCGGTCAT	68 (C1, C2)	830
	R	AACAGCTGYTSTAGTGTCCG	68 (A2, B)	
α 9	F	AGGGYGTAACCGAAAACGGT	52, 58 (all known variants)	801 (HPV52) 828 (HPV58)
	R	CCGGGGCACACAACTTGTAA	52 (all known variants)	
	R	ACAGCTAGGGCACACAATGG	58 (A1, A2, A3, B1, B2, D1)	
	R	GCTGTAGGGTTCGTSCTTCA	58 (C, D2)	785
	F	AGGGCGTAACCGAAATCGGT	16 (A1, A2, A3, A4, D1, D2, D3)	830
	R	TGAGAACAGATGGGGCACAC	16 (A1, A2, A3, B1, B2, C, D1, D2, D3)	
	F	TTGMACCGAAACCGGTTAGT	16 (B1, B2, C)	807
	R	RCAGATGGGGCACACAATTC	16 (A4)	
	F	GGTGAACCGAAAACGGTTGG	31 (all known variants)	793
	R	GGGGCACACGATTCCAAATG		
	F	AAGTAGGGTGTAACCGAAAGCG	33 (all known variants)	787
	R	TGCTGTATGGTTCGTAGGTCAC		
	F	ACGGTTGCCATAAAAGCAGAA	35 (all known variants)	827
	R	TCTCTGTGAACAGCCGGGG		

F: forward; R: reverse; * The 454-specific adapters were added to the 5′ end of both F and R primers, together with a 10-base multiplex identifier; § For each primer, the target genotype is reported; the variants that are covered by the corresponding primer are reported within round brackets.

13 forward and 16 reverse primers

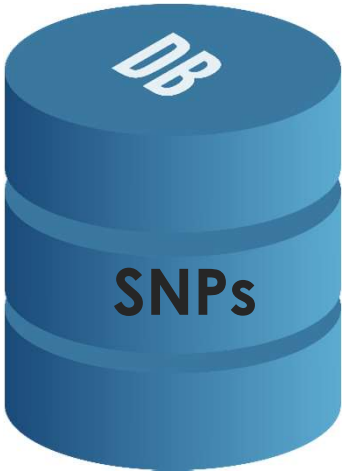
Characterization of HR-HPV intratype variants: the phylogenetic method

- HPV types are highlighted with different colors
- **Reference sequences for lineages and sublineages**
- **Sanger samples**
- **454 samples**

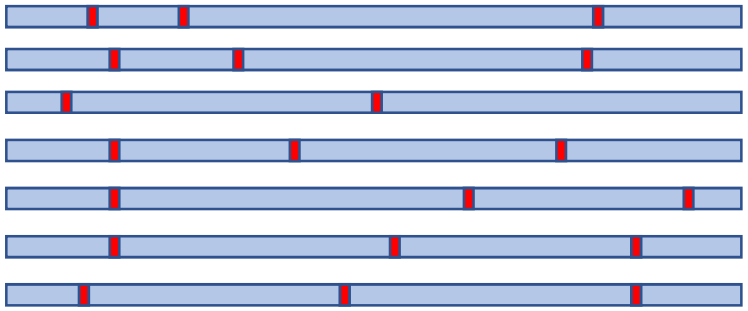


In most cases (just one exception) sample sequences clustered with a single variant reference and a reliable bootstrap value (>= 0.7)

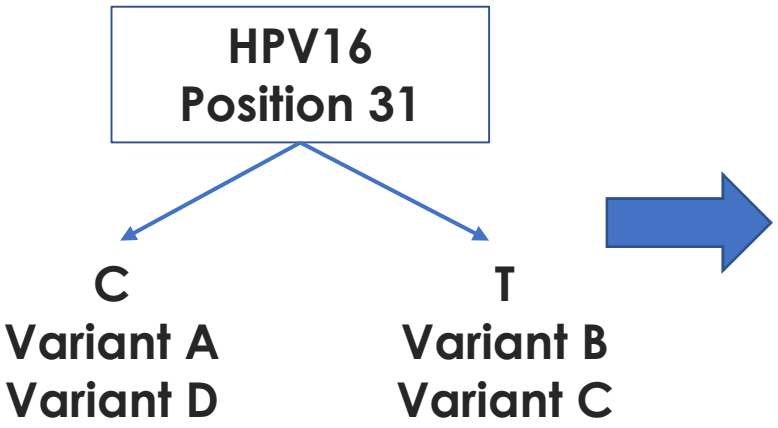
Characterization of HR-HPV intratype variants: SNPs scoring

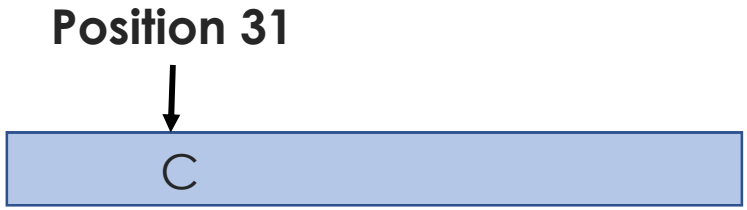


Variants-associated SNPs



Detection of SNPs on sample sequences, rewarding associated variants

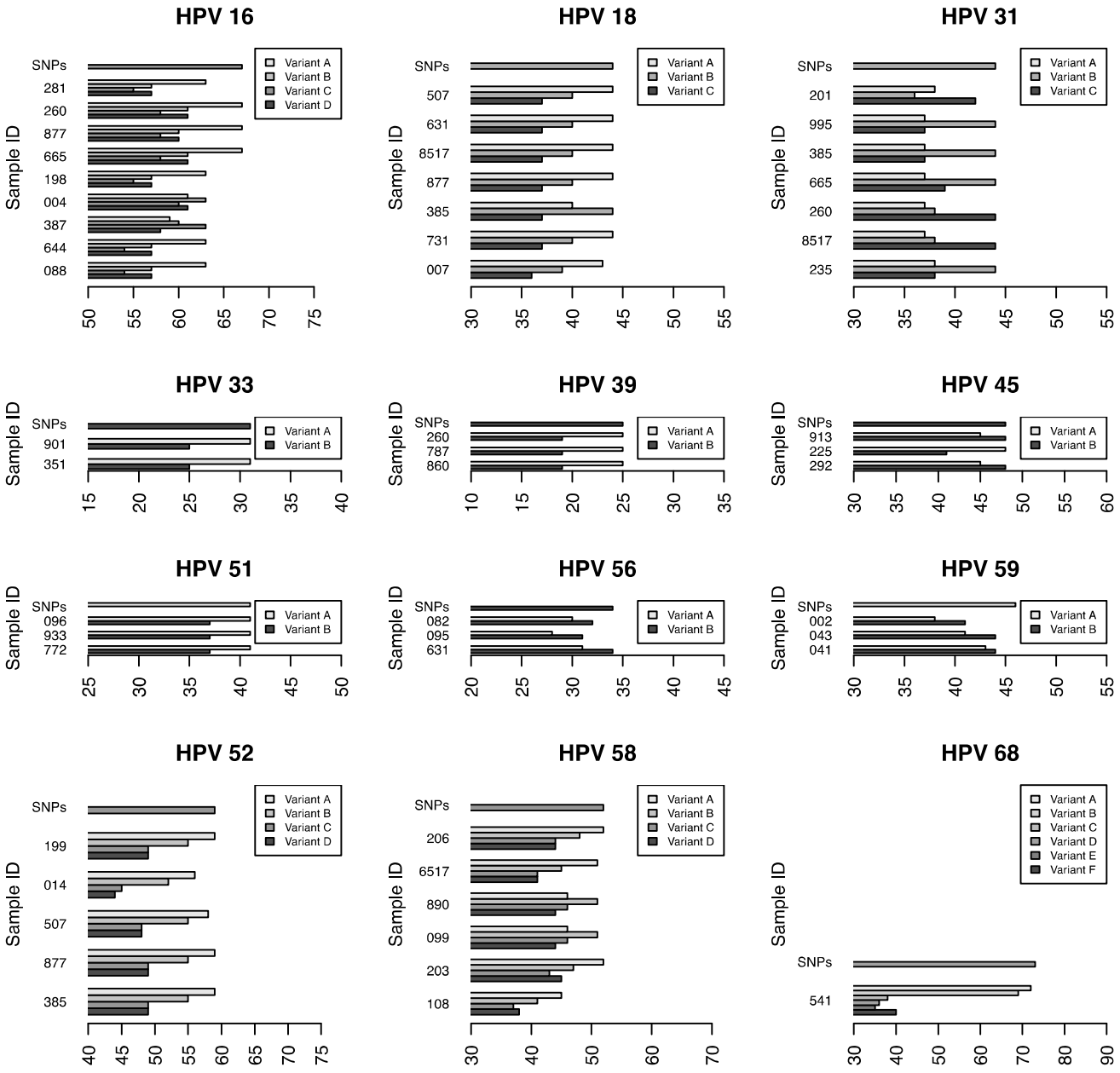




Sample sequence

Variant A = 1
Variant B = 0
Variant C = 0
Variant D = 1

SNP score to infer viral variants



Conclusioni

- Lo sviluppo di tool bioinformatici per l'analisi di campioni di metagenomica virale è un fenomeno relativamente recente

- **Non c'è uno standard**, il target è un ristretto gruppo di utenti esperti, il confronto tra tool è difficile

- I tool sono **difficili da mantenere** perché sono spesso progetti non finanziati, o sviluppati per applicazioni troppo specifiche

- Per una 'democratizzazione' dell'accessibilità all'analisi del viroma c'è la necessità di **software affidabile e di facile utilizzo** distribuito attraverso canali solidi e standardizzati (**necessità di risorse dedicate alla bioinformatica**)

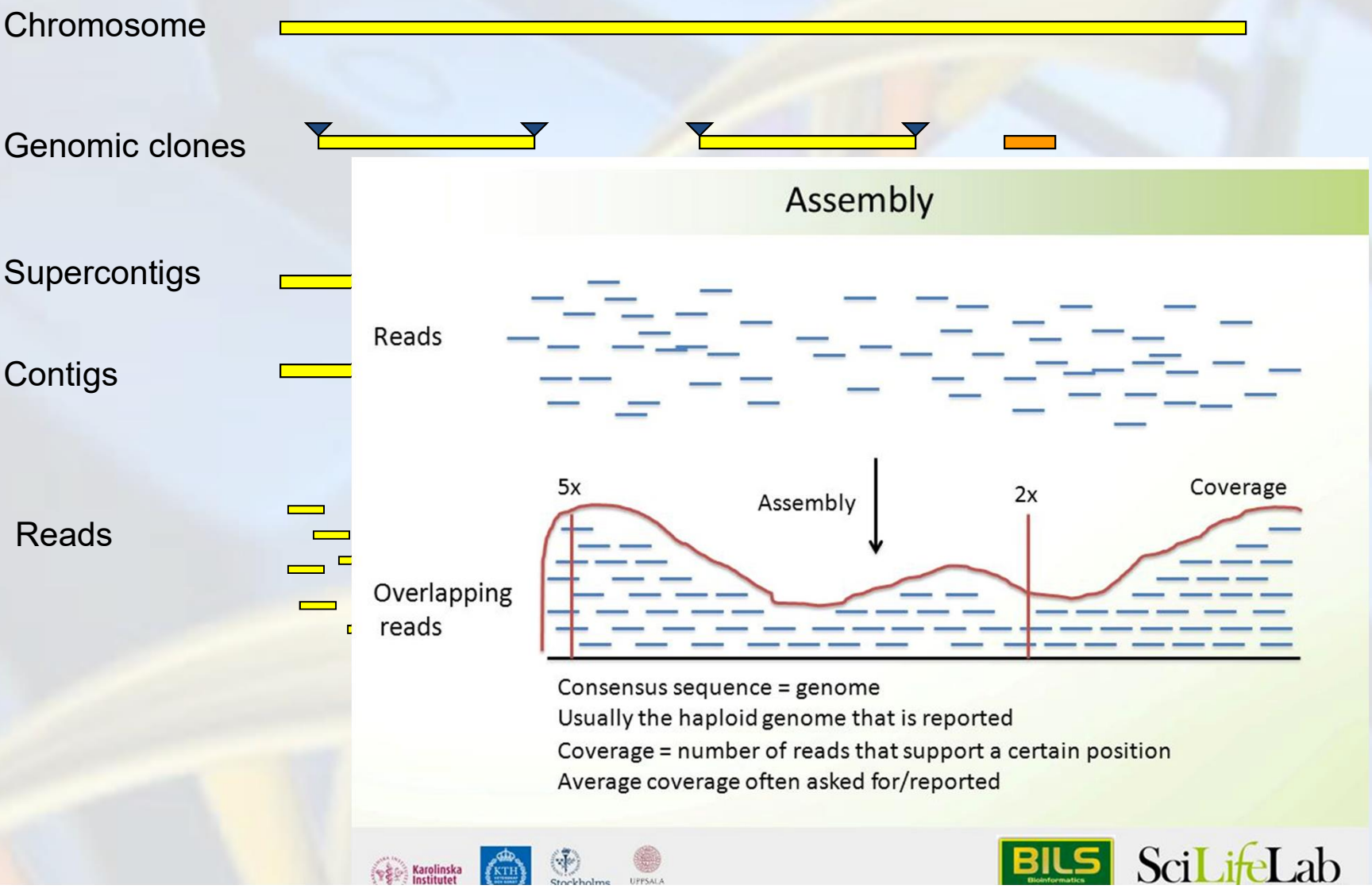
- Necessità di investimenti anche nei database di sequenze e condivisione dei dati, anche pre-analizzati

- La disponibilità di dati consentirà benchmarking più accurati dei tool

Usability and validation

- Disponibilità:
    - Alcuni tool forniscono servizi web con interfacce grafiche, e funzionano su molti sistemi operativi
    - Altri sono disponibili solo da linea di comando, in genere solo per sistemi Linux (in genere con poca documentazione)
    - Alcuni richiedono grandi risorse computazionali (memoria RAM e storage)
    - Altri mettono a disposizione computazione e storage in sistemi cloud (necessità di connessioni ad alta velocità)
- Validazione:
    - molto variabile a seconda dei tool (difficoltà nella definizione di benchmark datasets)
    - Performance diverse in applicazioni diverse

Sequencing substrates: from raw reads to finishing



Glossario

- Contigs: Contiguous nucleotide sequences assembled from multiple overlapping reads.
- Coverage: The number of times a genome (or part thereof) has been sequenced.
- de Bruijn graph: A network of nodes and edges, where each edge represents a k-mer found in the collection of reads, and each node represents either the prefix or suffix of the k-mer.
- De novo assembly: Reconstruction of short sequences into longer sequences (or contigs), without use of a reference sequence
- Digital signal processing data transformation: Analytical techniques for transforming sequential data into a domain representative of data features.
- Discrete Fourier transform: A spectral analysis technique for identifying sine and cosine frequency components in numerical signal data.
- Discrete wavelet transform: A spectral analysis technique for decomposing data to its frequency and spatial components.
- k-mer: A subsequence of length k. Many genomic analyses involve decomposition of sequences into all possible subsequences of a specified length k.
- Numerical sequence representation: Numerical mapping of nucleotide sequences, permitting the application of signal processing transformation approaches.
- Paired-end reads: Reads generated from both 5 and 3 ends of the same DNA molecule. Depending on the length of the molecule and that of the reads, these pairs may or may not overlap in the middle.
- Read overlap graphs: A network of nodes and edges, where each edge represents a read and each vertex represents an overlap between two nodes.
- Reference-based alignment: Orientation/alignment of reads with respect to a specified reference sequence.
- Scaffolds: DNA sequences comprising contigs with gaps between them, often generated using read pairing information.
- Suffix array: A sorted array of all suffixes of a string, such as a DNA sequence, enabling efficient sequence comparison.