

Design personalizzato del pannello per Illumina Complete Long Read Prep with Enrichment, Human

Arricchimento a lettura lunga
altamente flessibile e mirato
per genomi umani

illumina[®]

Introduzione

Quando si esegue il sequenziamento dell'intero genoma (WGS, whole-genome sequencing) umano, molte regioni potrebbero essere difficili da mappare solo con letture brevi. Il sequenziamento a lettura lunga può integrare i dati WGS standard a lettura breve per aiutare a risolvere queste regioni difficili. La tecnologia Illumina Complete Long Reads utilizza un flusso di lavoro standard di sequenziamento di nuova generazione (NGS, next-generation sequencing) per generare sequenze a lettura lunga contigue sui sistemi di sequenziamento Illumina con una singola pipeline di analisi (Figura 1).¹⁻³ Illumina Complete Long Read Prep with Enrichment, Human aggiunge un approccio mirato per un sequenziamento a lettura lunga più conveniente.* La chimica di arricchimento dell'Illumina Complete Long Read fornisce un'elevata flessibilità con target e design sonda volti ad aiutare a risolvere regioni difficili da mappare o a fornire informazioni aggiuntive dal sequenziamento con determinazione delle fasi.

Progettazione di pannelli sonda di arricchimento per letture lunghe

Illumina Complete Long Read Prep with Enrichment, Human utilizza una strategia di progettazione della sonda diversa al fine di acquisire frammenti più lunghi (circa 7-10 kb) rispetto all'approccio tipicamente utilizzato per acquisire frammenti

* Per l'analisi richiede almeno 30x di dati WGS standard a lettura breve dallo stesso campione. È possibile utilizzare i file FASTQ di un campione precedentemente analizzato.

corti (circa 200-500 bp). Il software Illumina **DesignStudio™** è uno strumento gratuito e intuitivo per la progettazione di pannelli sonda di arricchimento. L'algoritmo DesignStudio considera il contenuto in GC, la specificità del target e la spaziatura delle sonde, ovvero quante sonde ci saranno nella regione target. La spaziatura standard per i pannelli di arricchimento a lettura breve da 120 mer è una finestra della sonda da 250-350 bp. Per la progettazione del pannello di arricchimento a lettura lunga, la spaziatura delle sonde è stata testata su più lunghezze ed è risultato che una finestra a un solo chilobase è ottimale per un'acquisizione conveniente e ad alta efficienza.

L'efficacia dell'arricchimento dell'ibridazione dipende fortemente dalla specificità della sonda. La percentuale di arricchimento sul target influisce direttamente sulla quantità di sequenziamento necessaria per raggiungere la profondità di copertura target. L'elevata specificità è più difficile da ottenere per le regioni ripetitive. Tuttavia, utilizzando una finestra della sonda più grande si ha una maggiore flessibilità per escludere le sonde con prestazioni scarse, evitare regioni ripetute (fino alla dimensione della finestra di 1 kb) e garantire l'efficienza di arricchimento con meno sonde (Figura 2). L'algoritmo DesignStudio può utilizzare queste considerazioni per raccomandare il posizionamento della sonda. I pannelli di terze parti devono utilizzare linee guida simili per assicurare le migliori prestazioni e la convenienza. Anche la spaziatura standard della sonda di arricchimento è completamente compatibile.

Flessibilità nella progettazione della sonda e nella strategia target

Illumina Complete Long Read Prep with Enrichment, Human offre un'elevata flessibilità per scegliere e progettare pannelli di sonde personalizzati in base agli obiettivi dello studio. Le singole regioni mirate possono coprire singole basi fino a centinaia di kilobasi.



Figura 1: parte di un flusso di lavoro integrato. Consente di accedere ai dati WGS a lettura lunga mirati e convenienti utilizzando un protocollo scalabile e ottimizzato per la preparazione delle librerie con arricchimento, la comprovata chimica di sequenziamento Illumina e l'analisi secondaria DRAGEN. Per l'analisi richiede almeno 30x di dati WGS standard a lettura breve dallo stesso campione. È possibile utilizzare i file FASTQ di un campione precedentemente analizzato.

La dimensione totale del pannello personalizzato può variare da 2,5 Mb a più di 95 Mb. I ricercatori possono utilizzare letture lunghe mirate per migliorare la copertura in regioni specifiche notoriamente a bassa mappabilità con dati di lettura breve. In alternativa, le letture lunghe possono essere mirate per coprire interi geni fino a lunghe regioni multigeniche per consentire la determinazione delle fasi delle varianti e l'identificazione degli aplotipi.

Diversi pannelli pre-progettati sono disponibili nello strumento DesignStudio (Tabella 1). Questi pannelli si occupano di geni rilevanti dal punto di vista medico (CMRG, challenging medically relevant gene),⁴ di geni comunemente coperti dai saggi di analisi farmacogenetica (PGx, pharmacogenetic),⁵⁻⁷ di geni presenti nell'elenco dei risultati secondari dell'American College of Medical Genetics and Genomics (ACMG) (ACMG SF v3.1)⁸ o dell'intera regione del complesso maggiore di istocompatibilità (MHC, major histocompatibility complex).⁹ L'Illumina Human Comprehensive Panel, che si occupa principalmente di regioni separate a bassa copertura all'interno dei geni codificanti le proteine, è disponibile anche come pannello pre-progettato o pre-fabbricato pronto per la spedizione (Illumina, N. di catalogo 20113836).^{10,11} Il software DesignStudio supporta la progettazione di pannelli personalizzati da file BED[†] o la modifica di progetti predefiniti esistenti.

Profondità di sequenziamento consigliata per i pannelli sonda personalizzati

Illumina Complete Long Read Prep with Enrichment, Human fornisce prestazioni solide e altamente coerenti. Per i pannelli pre-progettati testati, le prestazioni ottimali sono state ottenute con circa 1,5 Gb di dati di sequenza (circa 5 milioni di letture paired-end) per 1 Mb di dimensione del pannello target (Figura 3). Per i pannelli di nuova progettazione con prestazioni sconosciute, 3 Gb di dati di sequenza (circa 10 milioni di letture paired-end) per 1 Mb di dimensione del pannello target rappresentano il punto di inizio raccomandato, soggetto a riduzione con ulteriore ottimizzazione.

Copertura e determinazione delle fasi ad alta precisione delle regioni difficili

I pannelli delle sonde di arricchimento a lettura lunga dedicati al miglioramento di specifiche regioni a bassa copertura, come l'Illumina Human Comprehensive Panel e il pannello CMRG, migliorano l'accuratezza di identificazione di varianti in regioni difficili mirate (Figura 4).

† BED, formato dati estensibile del browser.



Figura 2: l'ibridazione a frammenti lunghi aumenta l'efficienza di arricchimento. L'ibridazione a frammenti lunghi offre vantaggi rispetto all'acquisizione di frammenti corti, tra cui (A) il posizionamento strategico delle sonde al di fuori delle regioni difficili per la progettazione delle sonde, come contenuto in GC estremo, bassa complessità o ripetizioni, e (B) la necessità di un numero inferiore di sonde per acquisire ciascuna regione target. L'algoritmo DesignStudio cerca tra le regioni target in sezioni da 1 kb per ottenere un contenuto ottimale in GC e regioni con la massima specificità per posizionare le sonde.

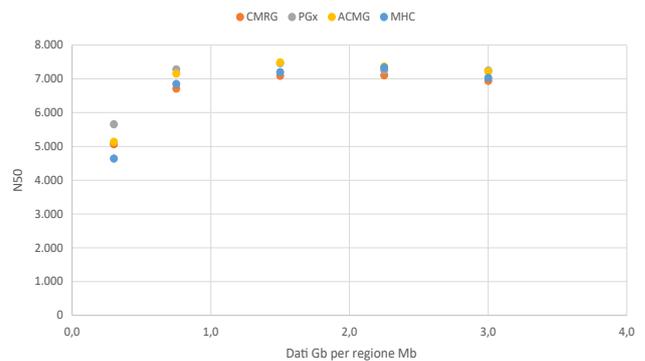


Figura 3: requisiti di sequenziamento per pannelli di sonde personalizzati. La titolazione dei dati di sequenziamento richiesti per massimo N50 mostra che 1,5 Gb (circa 5 milioni di letture paired-end) per regione target Mb risolvono in modo efficiente le regioni target per la generazione di dati Illumina Complete Long Read.

Tabella 1: pannelli di sonde di arricchimento pre-progettati per Illumina Complete Long Read Prep with Enrichment, Human

Pannelli ^a	Pannello CMRG	Pannello PGx	Pannello ACMG	Pannello MHC
Geni target	391 geni clinicamente rilevanti notoriamente difficili da risolvere con letture brevi ⁵	98 geni comunemente coperti dai saggi di analisi farmacogenetica ⁶⁻⁸	78 geni unici dall'elenco dei risultati secondari dell'ACMG (ACMG SF v3.1) ⁹	Più di 140 geni nella regione MHC completa nel gruppo GRCh38.p14 ¹⁰
Dimensione regione target ^b	22,5 Mb	8,1 Mb	7 Mb	4,9 Mb
Output di sequenziamento per campione ^c	Circa 67,5 Gb	Circa 24,3 Gb	Circa 21 Gb	Circa 14,7 Gb
N. di sonde	Circa 22.500	Circa 8.200	Circa 6.900	Circa 5.000
N50 ^d	6,1 kb	7,3 kb	7,3 kb	7,3 kb
Blocco di fase N50 ^{d,e}	82,8 kb	94,4 kb	94,4 kb	357 kb
Dimensione media della regione target ^e	58 kb	83 kb	88 kb	5.000 kb
Uniformità ^{d,f}	97,9%	99,0%	99,5%	97,8%
Arricchimento della lettura "padded" (PRE) ^{d,f}	80,1%	79,3%	66,3%	67,5%
% di SNV eterozigote a fasi ^d	98,9%	98,9%	99,6%	98,6%

- a. CMRG, geni rilevanti dal punto di vista medico; PGx, farmacogenomica; ACMG, American College of Medical Genetics and Genomics; MHC, complesso maggiore di istocompatibilità.
- b. La dimensione della regione target corrisponde alla somma delle lunghezze della posizione della sonda "padded", unite dove si sovrappongono.
- c. Richiede una corsa di sequenziamento di 2 x 150 bp e 5-10 milioni di letture paired-end (circa 1,5-3 Gb di dati) per regione target Mb, generando circa 30x di copertura finale di Illumina Complete Long Read. I requisiti relativi ai dati del pannello personalizzato per campione rappresentano solo il punto di partenza raccomandato. È possibile ottimizzare i dati assegnati in base alle prestazioni del pannello.
- d. Dati generati utilizzando 50 ng di DNA genomico HG002 (Coriell, n. di catalogo NA24385). Le prestazioni possono variare in base all'input di DNA e alla qualità del campione.
- e. Le dimensioni dei blocchi di fase sono limitate alle dimensioni delle singole regioni target contigue.
- f. Uniformità della copertura calcolata come % >0,2 * media. PRE calcolato come 100 * (letture allineate target "padded"/letture allineate totali).

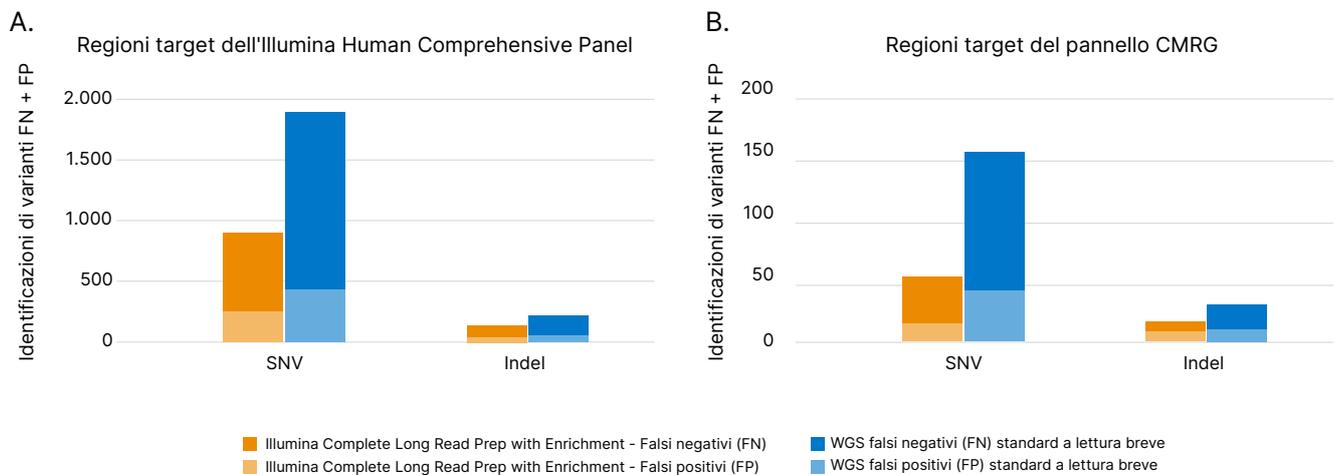


Figura 4: letture lunghe mirate per migliorare l'accuratezza dell'identificazione di varianti nelle regioni difficili. Identificazioni di varianti false negative (FN) e false positive (FP) per le varianti di singolo nucleotide (SNV, single nucleotide variant) e delezione/ inserzione (indel) nelle regioni geniche HG002 coperte (A) dallo Human Comprehensive Panel o (B) dal pannello CMRG, utilizzando Illumina Complete Long Read Prep with Enrichment (arancione) rispetto al WGS standard a lettura breve (blu).



Figura 5: le letture lunghe mirate migliorano le regioni a bassa copertura. Grafici Integrative Genomics Viewer (IGV) relativi al sequenziamento a lettura lunga di *HBG1* utilizzando Illumina Complete Long Read Prep, Human WGS (in alto) e Illumina Complete Long Read Prep with Enrichment, Human e il pannello CMRG (al centro) rispetto al WGS standard a lettura breve (in basso).

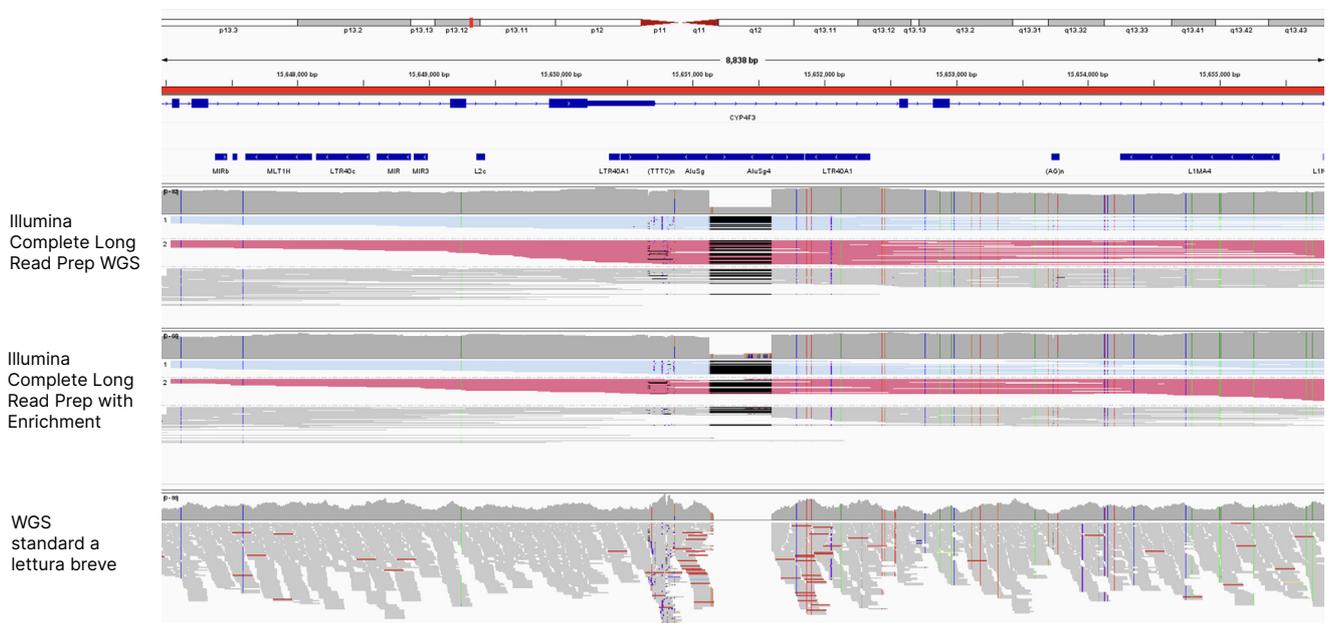


Figura 6: risoluzione chiara dei bordi di delezione con letture lunghe mirate. Grafici IGV dal sequenziamento a lettura lunga e dalla determinazione delle fasi di *CYP4F3* utilizzando Illumina Complete Long Read Prep, Human WGS (in alto) e Illumina Complete Long Read Prep with Enrichment, Human e il pannello CMRG (al centro) rispetto al WGS standard a lettura breve (in basso). Allele 1 in blu, allele 2 in rosa.

L'arricchimento a lettura lunga con il pannello CMRG contribuisce anche a migliorare la completezza della copertura e il rilevamento delle varianti nelle regioni codificanti le proteine (Figura 5, Figura 6).

Blocchi di fase lunghi per risolvere gli aplotipi

Il blocco di fase N50[†] di ciascun pannello è correlato alla lunghezza contigua delle regioni target (Figura 7, Tabella 1). I pannelli CMRG, PGx e ACMG sono stati progettati per occuparsi dei geni di interesse a lunghezza intera e hanno prodotto un blocco di fase medio N50 di circa 80-95 kb per consentire la determinazione completa delle fasi degli alleli eterozigoti (Figura 8). Il pannello MHC si occupa di una singola regione contigua di circa 4,9 Mb e ha prodotto un blocco di fase medio N50 di oltre 350 kb per consentire la risoluzione della regione genica a lunghezza intera (Figura 9).

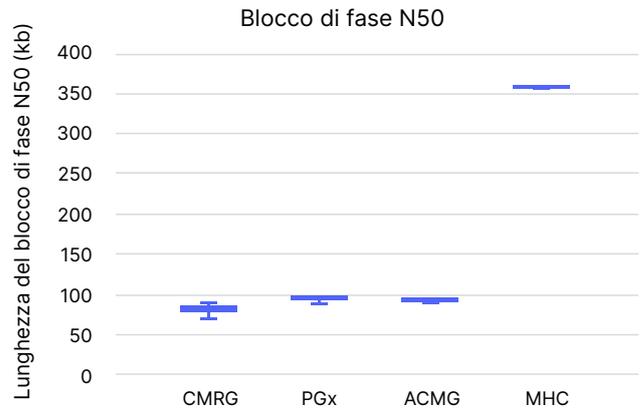


Figura 7: il blocco di fase N50 dipende dalla lunghezza delle regioni target contigue. I pannelli CMRG, PGx e ACMG si occupano dei geni di interesse a lunghezza intera e hanno prodotto un blocco di fase medio N50 di circa 80-95 kb. Il pannello MHC si occupa dell'intera regione genica del complesso maggiore di istocompatibilità e ha prodotto un blocco di fase medio N50 di oltre 350 kb. La dimensione media della regione target è di 58 kb per il pannello CMRG, 83 kb per il pannello PGx, 88 kb per il pannello ACMG e 5.000 kb per il pannello MHC.

† Il blocco di fase N50 riflette la lunghezza del blocco più breve della sequenza contigua al 50% della lunghezza totale del gruppo delle regioni target.

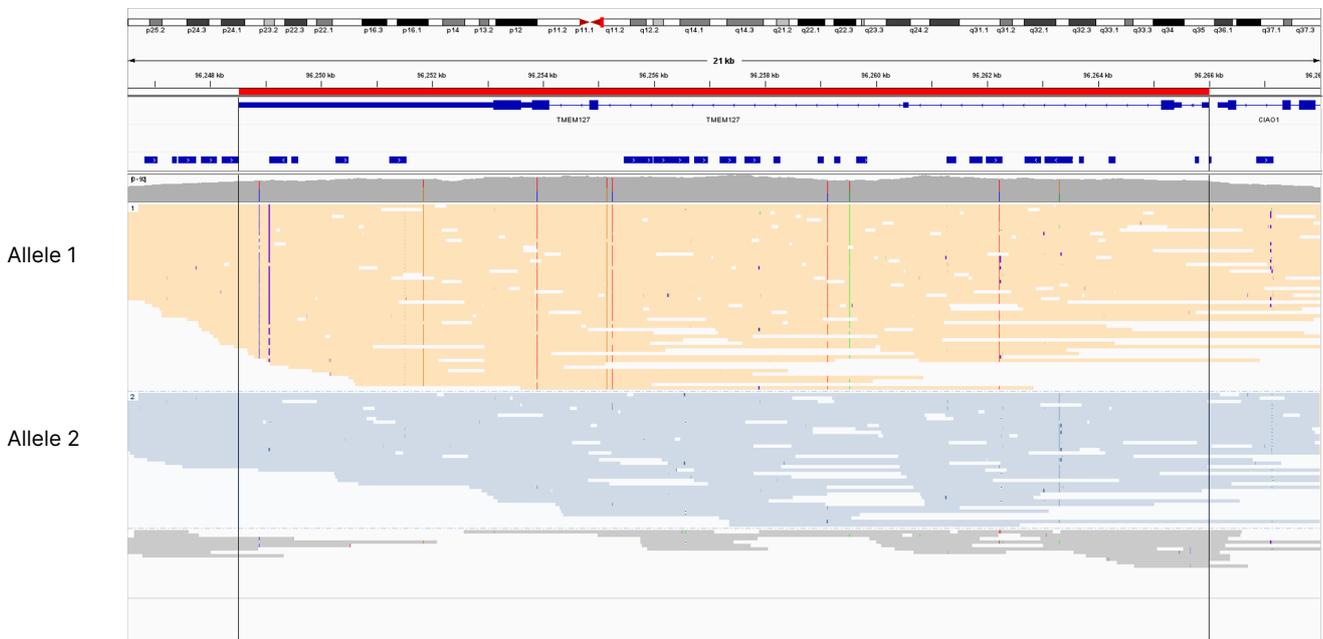


Figura 8: le letture lunghe mirate consentono la determinazione delle fasi delle regioni con SNV eterozigoti. I grafici IGV dal sequenziamento a lettura lunga mostrano la determinazione delle fasi completa in un blocco di fase per *TMEM127*, un gene da 21 kb, utilizzando Illumina Complete Long Read Prep with Enrichment, Human e il pannello ACMG. Allele 1 in giallo. Allele 2 in blu.

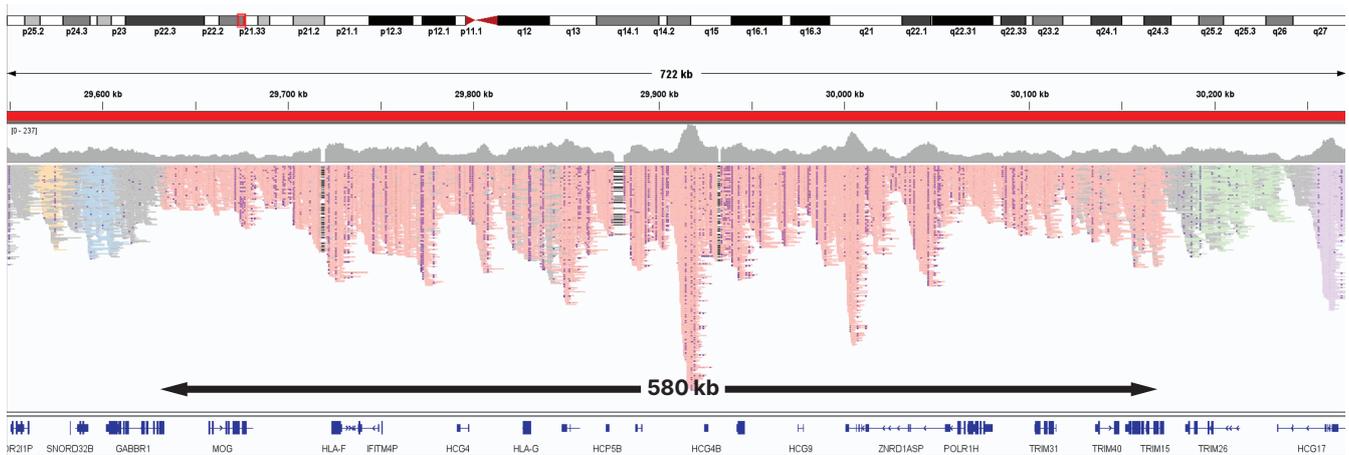


Figura 9: le letture lunghe mirate aiutano a risolvere gli aplotipi nei geni polimorfici. Grafici IGV del sequenziamento a lettura lunga con Illumina Complete Long Read Prep with Enrichment, Human. Determinazione delle fasi su una regione di 722 kb nel locus MHC. Una regione da 580 kb (rosa) è incapsulata in un blocco di fase.

Riepilogo

Illumina Complete Long Read Prep with Enrichment, Human integra il comprovato WGS a lettura breve Illumina e utilizza il sequenziamento a lettura lunga laddove questo fornisce il massimo valore. I ricercatori hanno un'elevata flessibilità poiché possono scegliere pannelli pre-progettati o utilizzare l'algoritmo DesignStudio per progettare pannelli personalizzati per un arricchimento mirato a lettura lunga. I pannelli sonda di arricchimento possono essere utilizzati per migliorare la copertura o aggiungere informazioni tramite la determinazione delle fasi di geni interi per un WGS conveniente e altamente accurato con una soluzione completa per il flusso di lavoro.

Maggiori informazioni

[Illumina Complete Long Read Prep with Enrichment, Human](#)

[Strumento di progettazione del saggio DesignStudio](#)

[Tecnologia di sequenziamento delle letture lunghe](#)

Bibliografia

1. Illumina. Scheda tecnica di Illumina Complete Long Read Prep, Human. [illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-data-sheet-m-gl-01420/illumina-long-read-prep-data-sheet-m-gl-01420.pdf](https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-data-sheet-m-gl-01420/illumina-long-read-prep-data-sheet-m-gl-01420.pdf). Pubblicato nel 2022. Consultato il 22 settembre 2023.
2. Illumina. Comprehensive whole-genome sequencing with Illumina Complete Long Read Prep, Human technical note. [illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-tech-note-m-gl-01421/ilmn-long-read-hu-tech-note-m-gl-01421.pdf](https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-tech-note-m-gl-01421/ilmn-long-read-hu-tech-note-m-gl-01421.pdf). Pubblicato nel 2022. Consultato il 22 settembre 2023.
3. Roessler K. Illumina Complete Long Reads software analysis workflow for human WGS. <https://www.illumina.com/science/genomics-research/articles/complete-long-read-software-analysis.html>. Pubblicato nel 2023. Consultato il 22 settembre 2023.
4. Wagner J, Olson ND, Harris L, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol.* 2022;40(5):672-680. doi:10.1038/s41587-021-01158-1.
5. PharmGKB. VIPs: Very Important Pharmacogenes. [pharmgkb.org/vips](https://www.pharmgkb.org/vips). Consultato il 22 settembre 2023.
6. National Library of Medicine. GTR: Genetic Testing Registry. Precision HealthPGx Panel (25 Genes). [ncbi.nlm.nih.gov/gtr/tests/593428/](https://www.ncbi.nlm.nih.gov/gtr/tests/593428/). Aggiornato il 29 novembre 2022. Consultato il 22 settembre 2023.

7. Pratt VM, Everts RE, Aggarwal P, et al. Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. *J Mol Diagn.* 2016;18(1):109-123. doi:10.1016/j.jmoldx.2015.08.005
8. Miller DT, Lee K, Abul-Husn NS, et al. ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med.* 2022;24(7):1407-1414. doi:10.1016/j.gim.2022.04.006.
9. Kulski JK, Suzuki S, Shiina T. Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum Genome Var.* 2022;9(1):49. doi:10.1038/s41439-022-00226-5.
10. Bekritsky MA, Colombo C, Eberle MA. Identifying genomic regions with high quality single nucleotide variant calling. Pubblicato nel 2021. Consultato il 30 agosto 2023.
11. Illumina. Scheda tecnica di Illumina Human Comprehensive Panel. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-enrich-hu-comp-panel-data-sheet-m-gl-02191/long-read-hu-comp-panel-data-sheet-m-gl-02191.pdf. Pubblicato nel 2024. Consultato il 26 gennaio 2024.



Numero verde 1.800.809.4566 (U.S.A.) | Tel. +1.858.202.4566
techsupport@illumina.com | www.illumina.com

© 2024 Illumina, Inc. Tutti i diritti riservati. Tutti i marchi di fabbrica sono di proprietà di Illumina, Inc. o dei rispettivi proprietari. Per informazioni specifiche sui marchi di fabbrica, visitare la pagina web www.illumina.com/company/legal.html.
M-GL-02189 ITA v1.0