TECHNISCHER HINWEIS

Design anwendungs-spezifischer Panels für Illumina Complete Long Read Prep with Enrichment, Human

Hochflexible, gezielte Long-Read-Anreicherung für Humangenome

illumina®

Einleitung

Bei der Sequenzierung des Humangenoms (WGS, Whole-Genome Sequencing) lassen sich bestimmte Regionen unter Umständen mit Short-Reads allein nur schwer mappen. Die Long-Read-Sequenzierung kann mit der herkömmlichen Short-Read-WGS gewonnene Daten ergänzen, sodass sich diese schwierigen Regionen erfolgreich analysieren lassen. Die Illumina Complete Long Reads-Technologie generiert im Rahmen eines herkömmlichen NGS-Workflows (Next-Generation Sequencing, Sequenzierung der nächsten Generation) auf Illumina-Sequenziersystemen anhand einer einzigen Analysepipeline zusammenhängende Long-Read-Sequenzen (Abbildung 1).[1–3] Zusätzlich macht Illumina Complete Long Read Prep with Enrichment, Human die Long-Read-Sequenzierung dank eines gezielten Verfahrens kostengünstiger.* Die Illumina Complete Long Read-Anreicherungschemie bietet hohe Flexibilität bei Targets und Sondendesign, sodass sich schwer zu mappende Regionen analysieren oder anhand einer phasierten Sequenzierung zusätzliche Erkenntnisse gewinnen lassen.

Design von Anreicherungs-sondenpanels für Long-Reads

Bei Illumina Complete Long Read Prep with Enrichment, Human kommt zur Erfassung längerer Fragmente (ca. 7–10 kb) eine andere Sondendesignstrategie zum Einsatz als beim herkömmlichen Verfahren für kurze Fragmente (ca. 200–500 bp). Die Illumina DesignStudio™-Software ist ein kostenloses, anwenderfreundliches Tool für das Design von Anreicherungssondenpanels.

Der DesignStudio-Algorithmus berücksichtigt GC-Gehalt, Zielspezifität und Sondenabstand, d. h., wie viele Sonden sich im Zielbereich befinden. Der Standardabstand für Short-Read-Anreicherungspanels (120 mer) ist ein Sondenfenster von 250–350 bp. Für das Design des Long-Read-Anreicherungspanels wurden unterschiedliche Sondenabstände getestet, wobei sich ein Fenster mit der Länge von einer Kilobase als optimal für die kostengünstige, hocheffiziente Erfassung erwiesen hat.

Die Wirksamkeit der Hybridisierungsanreicherung hängt stark von der Sondenspezifität ab. Der prozentuale Anteil der On-Target-Anreicherung wirkt sich direkt darauf aus, in welchem Umfang eine Sequenzierung erforderlich ist, um die angestrebte Coverage-Tiefe zu erreichen. Bei repetitiven Regionen ist es schwieriger, eine hohe Spezifität zu erreichen. Das größere Sondenfenster ermöglicht jedoch eine höhere Flexibilität zum Ausschluss von Sonden mit unzureichender Leistung, zur Vermeidung von Wiederholungsregionen (bis zur Fenstergröße von 1 kb) und zur Aufrechterhaltung der Anreicherungseffizienz mit weniger Sonden (Abbildung 2). Der DesignStudio-Algorithmus kann auf Basis dieser Faktoren die Sondenplatzierung empfehlen. Panels von Drittanbietern sollten zur Optimierung von Leistung und Kosteneffizienz ähnliche Richtlinien verwenden. Der Standardabstand für Anreicherungssonden ist ebenfalls vollständig kompatibel.

Flexibilität bei Sondendesign und Zielstrategie

Illumina Complete Long Read Prep with Enrichment, Human bietet eine hohe Flexibilität bei der Auswahl und dem Design anwendungsspezifischer Sondenpanels für bestimmte Studienziele. Die einzelnen Zielregionen können dabei von einzelnen Basen bis hin zu Hunderten von Kilobasen reichen. Die Gesamtgröße anwendungsspezifischer Panels kann von nur 2,5 Mb bis hin zu > 95 Mb reichen. Forscher können anhand von gezielten Long-Reads die Coverage in bestimmten Regionen erhöhen, die mit Short-Read-Daten nur eine geringe Mapping-Fähigkeit aufweisen.

* Erfordert herkömmliche Short-Read-WGS-Daten mit ≥ 30-facher Coverage aus derselben Probe für die Analyse. Es können FASTQ-Dateien einer Probe aus einem vorherigen Lauf verwendet werden.

Abbildung 1: Teil eines integrierten Workflows: Zugriff auf kostengünstige, zielgerichtete Long-Read-WGS-Daten mithilfe eines skalierbaren, optimierten Bibliotheksvorbereitungsprotokolls mit Anreicherung, bewährter Illumina-Sequenzierungschemie und DRAGEN-Sekundäranalyse. Erfordert herkömmliche Short-Read-WGS-Daten mit ≥ 30-facher Coverage aus derselben Probe für die Analyse. Es können FASTQ-Dateien einer Probe aus einem vorherigen Lauf verwendet werden.

Alternativ können Long-Reads sich auch über gesamte Gene oder Multigenregionen erstrecken, was die Phasierung von Varianten und das Haplotypen-Calling ermöglicht.

Das DesignStudio-Tool enthält mehrere vordefinierte Panels (Tabelle 1). Diese Panels untersuchen CMRG (Challenging Medically Relevant Genes, schwierige medizinisch relevante Gene)[4], Gene, die üblicherweise mit pharmakogenetischen (PGx, Pharmacogenetics) Testassays untersucht werden[5–7], Gene auf der Liste der Sekundärergebnisse (ACMG SF v3.1)[8] des American College of Medical Genetics and Genomics (ACMG) oder die gesamte Region des Haupthistokompatibilitätskomplexes (MHC, Major Histocompatibility Complex)[9]. Das Illumina Human Comprehensive Panel, das primär diskrete Regionen mit geringer Coverage innerhalb proteincodierender Gene untersucht, ist auch als vordefiniertes oder versandfertiges vorgefertigtes Panel (Illumina, Katalog-Nr. 20113836) erhältlich.[10, 11] Die DesignStudio-Software ermöglicht das Design von anwendungsspezifischen Panels anhand von BED-Dateien† bzw. die Anpassung vordefinierter Panels.

Empfohlene Sequenzierungstiefe für anwendungsspezifische Sondenpanels

Illumina Complete Long Read Prep with Enrichment, Human zeichnet sich durch eine hochgradig konsistente und robuste Leistung aus. Für die getesteten vordefinierten Panels wurde eine optimale Leistung mit ca. 1,5 Gb Sequenzdaten (ca. 5 Mio. Paired-End-Reads) pro 1 Mb Zielpanelgröße erreicht (Abbildung 3). Für neu erstellte Panels mit unbekannter Leistung werden 3 Gb Sequenzdaten (ca. 10 Mio. Paired-End-Reads) pro 1 Mb Zielpanelgröße als Ausgangspunkt empfohlen. Im Rahmen einer weiteren Optimierung ist dabei eine Reduzierung möglich.

Hochgenaue Coverage und Phasierung schwieriger Regionen

Long-Read-Anreicherungssondenpanels zur Optimierung der Analyse bestimmter Regionen mit geringer Coverage, z. B. Illumina Human Comprehensive Panel und CMRG-Panel, verbessern die Genauigkeit des Varianten-Callings in schwierigen Zielregionen (Abbildung 4). Die Long-Read-Anreicherung mit dem CMRG-Panel optimiert zudem die Vollständigkeit der Coverage und den Nachweis von Varianten in proteincodierenden Regionen (Abbildung 5, Abbildung 6).

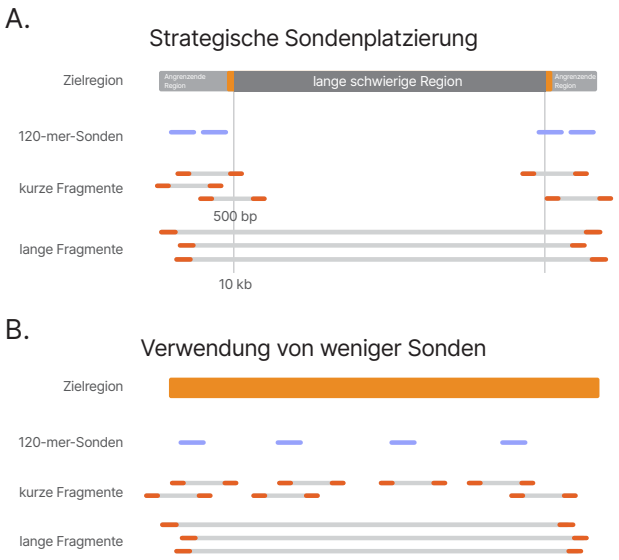

Abbildung 2: Die Hybridisierung mit langen Fragmenten erhöht die Effizienz der Anreicherung: Die Hybridisierung mit langen Fragmenten bietet gegenüber der Erfassung kurzer Fragmente Vorteile, darunter die (A) strategische Platzierung von Sonden außerhalb schwieriger Regionen für das Sondendesign wie solchen mit extremem GC-Gehalt, geringer Komplexität oder Wiederholungen und die (B) Erfassung der Zielregionen mit weniger Sonden. Der DesignStudio-Algorithmus sucht zur Platzierung von Sonden in 1-kb-Abschnitten der Zielregionen nach Regionen mit optimalem GC-Gehalt und der höchsten Spezifität.

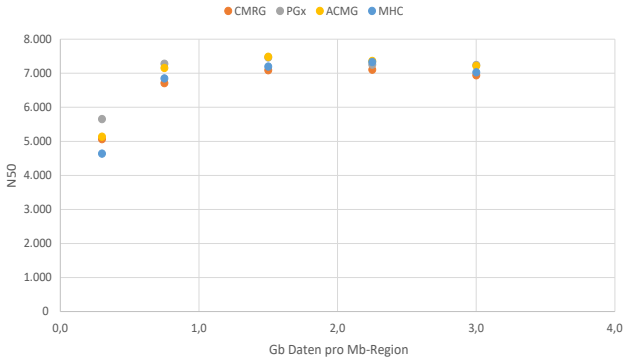

Abbildung 3: Sequenzierungsanforderungen für anwendungsspezifische Sondenpanels: Die aus der Titration resultierenden Sequenzierungsdaten, die für einen maximalen N50-Wert erforderlich sind, zeigen, dass 1,5 Gb (ca. 5 Mio. Paired-End-Reads) pro Mb-Zielregion zur Analyse von Zielregionen mit Illumina Complete Long Read geeignet sind.

† BED, Browser Extensible Data (browsererweiterbare Daten), ein Datenformat.

Tabelle 1: Vordefinierte Sondenpanels für Illumina Complete Long Read Prep with Enrichment, Human

Panels[a]	CMRG-Panel	PGx-Panel	ACMG-Panel	MHC-Panel
Zielgene	391 medizinisch relevante Gene, von denen bekannt ist, dass sie mit Short-Reads schwer zu analysieren sind[5]	98 Gene, die häufig mit pharmakogenetischen Test-Assays untersucht werden[6–8]	78 eindeutige Gene aus der ACMG-Liste der Sekundärergebnisse (ACMG SF v3.1)[9]	> 140 Gene in der gesamten MHC-Region in der GRCh38. p14-Assemblierung[10]
Größe der Zielregion[b]	22,5 Mb	8,1 Mb	7 Mb	4,9 Mb
Sequenzierungsausgabe je Probe[c]	ca. 67,5 Gb	ca, 24,3 Gb	ca. 21 Gb	ca. 14,7 Gb
Anzahl der Sonden	ca. 22.500	ca. 8.200	ca. 6.900	ca. 5.000
N50[d]	6,1 kb	7,3 kb	7,3 kb	7,3 kb
Phasenblock N50[d, e]	82,8 kb	94,4 kb	94,4 kb	357 kb
Mittlere Größe der Zielregion[e]	58 kb	83 kb	88 kb	5.000 kb
Einheitlichtkeit[d, f]	97,9 %	99,0 %	99,5 %	97,8 %
Padded-Read-Anreicherung (PRE, Padded Read Enrichment)[d, f]	80,1 %	79,3 %	66,3 %	67,5 %
Anteil phasierter heterozygoter SNV in %[d]	98,9 %	98,9 %	99,6 %	98,6 %

a. CMRG, Challenging Medically Relevant Genes (schwierige medizinisch relevante Gene); PGx, Pharmacogenomics (Pharmakogenomik); ACMG, American College of Medical Genetics and Genomics; MHC, Major Histocompatibility Complex (Haupthistokompatibilitätskomplex).
b. Die Größe der Zielregion ist die Summe der Sondenpositionslängen mit Auffüllung, die an der Überlappung zusammengeführt werden.
c. Erfordert einen Sequenzierungslauf mit 2 × 150 bp und 5 Mio.–10 Mio. Paired-End-Reads (ca. 1,5–3 Gb-Daten) pro Mb-Zielregion, wodurch die finale ca. 30-fache Coverage von Illumina Complete Long Reads generiert wird. Bei den Anforderungen hinsichtlich der Daten je Probe für anwendungsspezifische Panels handelt es sich lediglich um einen empfohlenen Ausgangspunkt. Anwender können zugewiesene Daten abhängig von der Panelleistung optimieren.
d. Daten, die mit 50 ng genomischer HG002-DNA (Corielle, Katalog-Nr. NA24385) generiert wurden. Die Leistung kann je nach DNA-Zugabe und Probenqualität variieren.
e. Die Phasenblockgrößen sind auf die Größen einzelner zusammenhängender Zielregionen begrenzt.
f. Coverage-Einheitlichkeit berechnet als % > 0,2 * Mittelwert. PRE berechnet als 100 * (alignierte Padded-Target-Reads / insgesamt alignierte Reads).

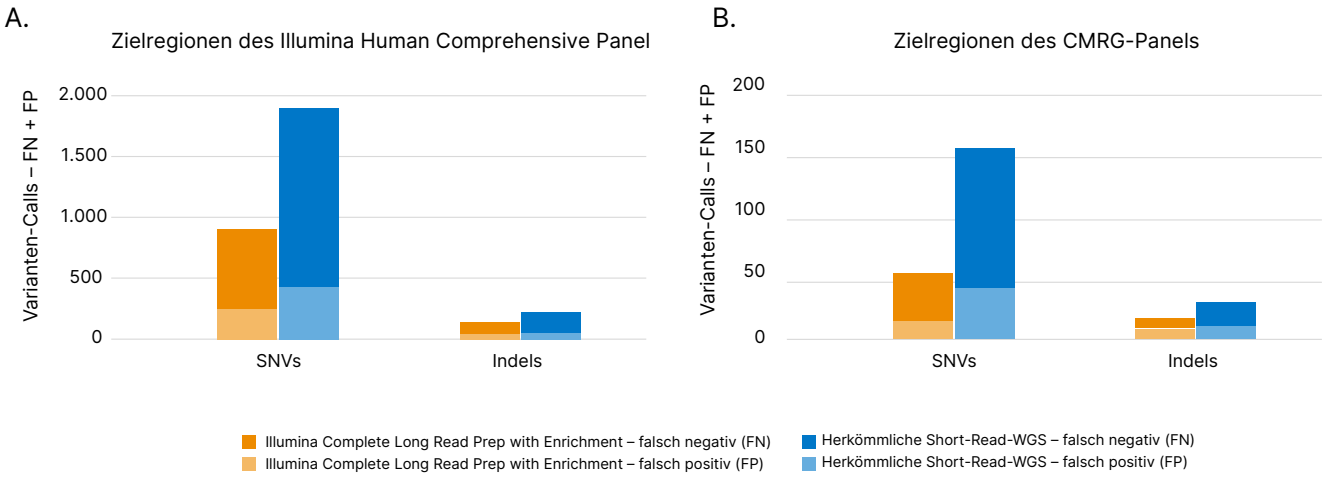

Abbildung 4: Zielgerichtete Long-Reads zur Verbesserung der Genauigkeit des Varianten-Callings in schwierigen Regionen: Falsch negative (FN) und falsch positive (FP) Varianten-Calls für Einzelnukleotid-Varianten (SNVs, Single-Nucleotide Variants) und Insertionen/Deletionen (Indels) in genetischen HG002-Regionen, die mit dem (A) Human Comprehensive Panel oder dem (B) CMRG-Panel untersucht wurden, wobei Illumina Complete Long Read Prep with Enrichment (orange) im Vergleich zur herkömmlichen Short-Read-WGS (blau) verwendet wurde.

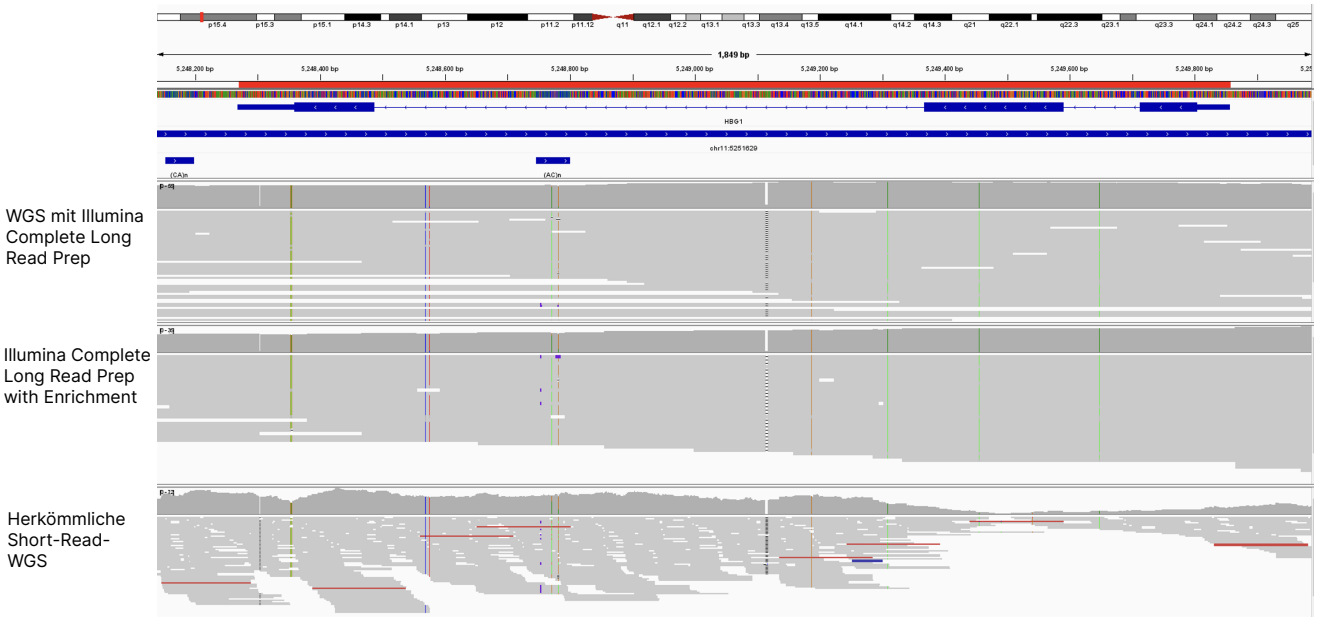

Abbildung 5: Gezielte Long-Reads optimieren die Analyse von Regionen mit geringer Coverage: Integrative Genomics Viewer(IGV)-Plots aus der Long-Read-Sequenzierung von *HBG1* unter Verwendung von WGS mit Illumina Complete Long Read Prep, Human (oben), Illumina Complete Long Read Prep with Enrichment, Human und CMRG-Panel (Mitte) im Vergleich zur herkömmlichen Short-Read-WGS (unten).

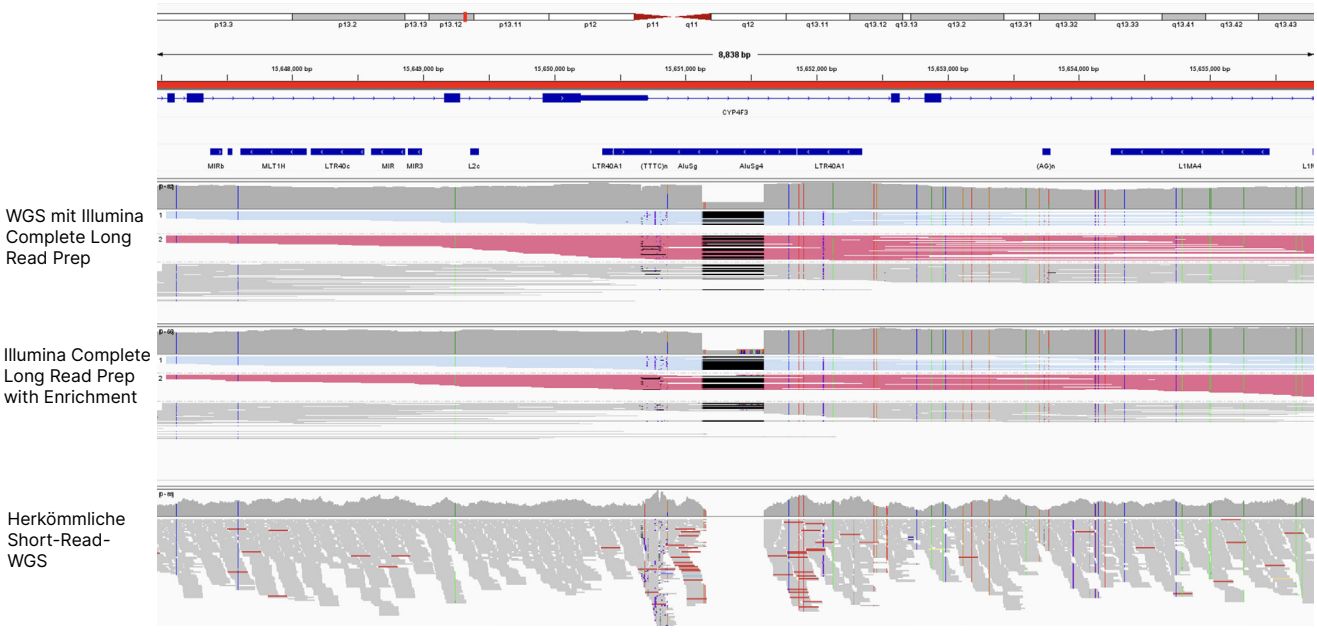

Abbildung 6: Klare Analyse von Deletionsgrenzen mit gezielten Long-Reads: IGV-Plots aus der Long-Read-Sequenzierung und Phasierung von *CYP4F3* unter Verwendung von WGS mit Illumina Complete Long Read Prep, Human (oben), Illumina Complete Long Read Prep with Enrichment, Human und CMRG-Panel (Mitte) im Vergleich zur herkömmlichen Short-Read-WGS (unten). Allel 1 in Blau, Allel 2 in Rosa.

Lange Phasenblöcke zur Analyse von Haplotypen

Der Phasenblock N50‡ der einzelnen Panels steht in Bezug zur zusammenhängenden Länge der Zielregionen (Abbildung 7, Tabelle 1). Die CMRG-, PGx- und ACMG-Panels untersuchen Gene von Interesse in voller Länge und ergaben einen mittleren Phasenblock N50 von ca. 80–95 kb für die vollständige Phasierung heterozygoter Allele (Abbildung 8). Das MHC-Panel untersucht eine einzelne zusammenhängende Region von ca. 4,9 Mb und ergab einen mittleren Phasenblock N50 von über 350 kb, was die Analyse der Genregion in voller Länge ermöglicht (Abbildung 9).

‡ Phasenblock N50 bezeichnet die Länge des kürzesten Blocks einer zusammenhängenden Sequenz bei 50 % der Gesamt-Assemblierungslänge der Zielregionen.

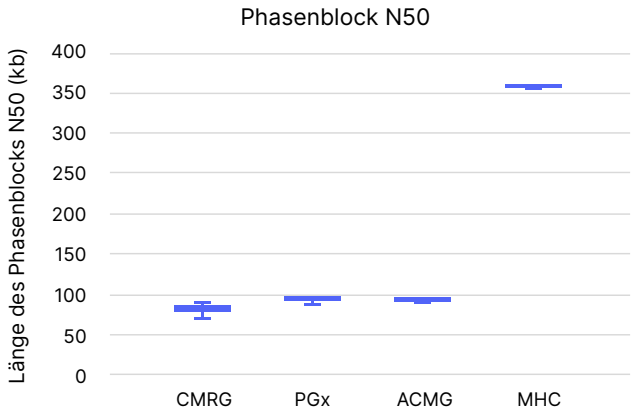

Abbildung 7: Phasenblock N50 ist abhängig von der Länge zusammenhängender Zielregionen: Die CMRG-, PGx- und ACMG-Panels untersuchen Gene von Interesse in voller Länge und ergaben einen mittleren Phasenblock N50 von ca. 80–95 kb. Das MHC-Panel untersucht die gesamte Genregion des Haupthistokompatibilitätskomplexes und ergab einen mittleren Phasenblock N50 von über 350 kb. Die mittlere Größe der Zielregionen beträgt beim CMRG-Panel 58 kb, beim PGx-Panel 83 kb, beim ACMG-Panel 88 kb und beim MHC-Panel 5.000 kb.

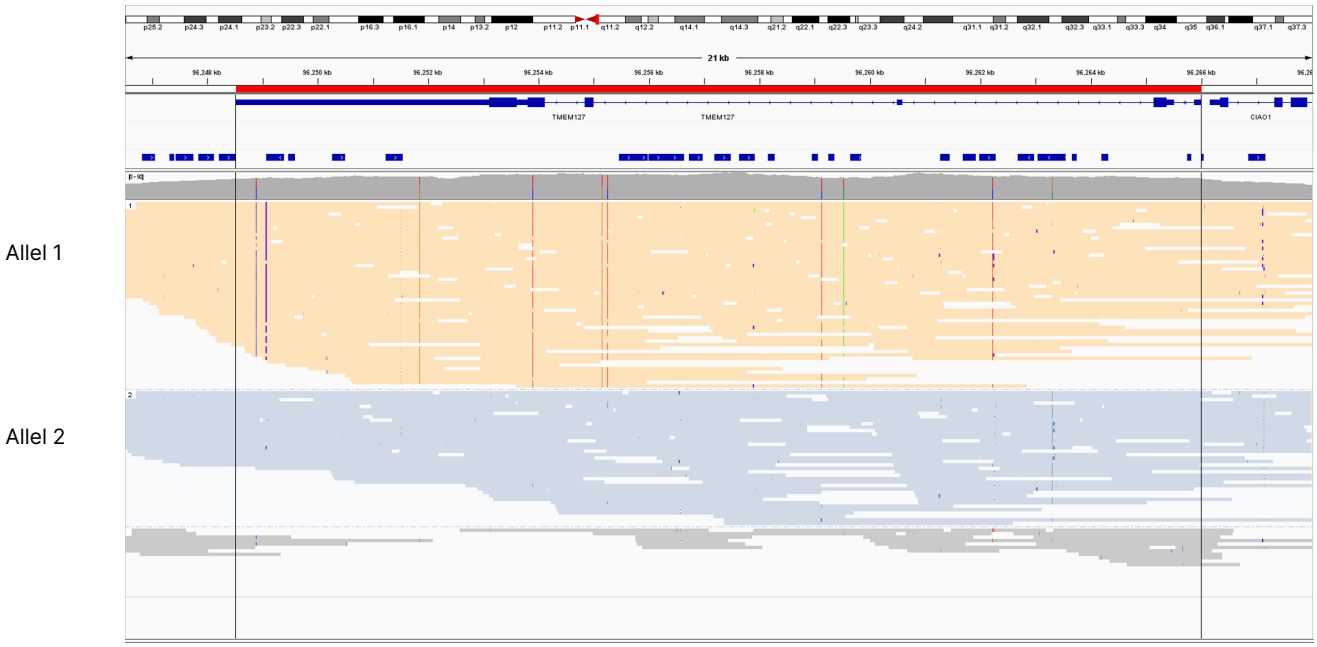

Abbildung 8: Gezielte Long-Reads ermöglichen die Phasierung von Regionen mit heterozygoten SNVs: IGV-Plots aus der Long-Read-Sequenzierung zeigen bei Illumina Complete Long Read Prep with Enrichment, Human und dem ACMG-Panel für *TMEM17* (Gen mit 21 kb) die vollständige Phasierung in einem Phasenblock. Allel 1 in Gelb. Allel 2 in Blau.

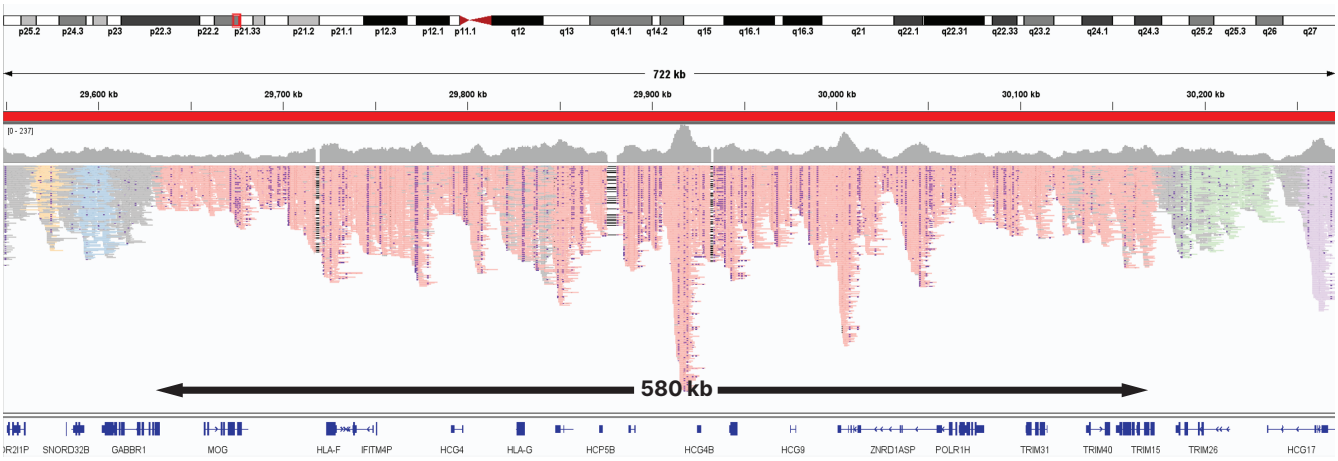

Abbildung 9: Zielgerichtete Long-Reads unterstützen die Analyse von Haplotypen mit polymorphen Genen: IGV-Plots aus der Long-Read-Sequenzierung mit Illumina Complete Long Read Prep with Enrichment, Human. Phasierung über eine Region von 722 kb im MHC-Locus. Eine Region von 580 kb (rosa) ist in einem Phasenblock eingekapselt.

Zusammenfassung

Illumina Complete Long Read Prep with Enrichment, Human ergänzt die bewährte Short-Read-WGS von Illumina und nutzt die Long-Read-Sequenzierung dort, wo sie die größten Vorteile bietet. Forscher profitieren von hoher Flexibilität, da sie vordefinierte Panels auswählen oder den DesignStudio-Algorithmus zum Design anwendungsspezifischer Panels für die gezielte Long-Read-Anreicherung verwenden können. Anreicherungssondenpanels lassen sich gezielt einsetzen, um die Coverage zu erhöhen oder durch die Phasierung vollständiger Gene im Rahmen einer kostengünstigen, hochpräzisen WGS mit umfassender Workflowlösung neue Erkenntnisse zu gewinnen.

Weitere Informationen

Illumina Complete Long Read Prep with Enrichment, Human

DesignStudio-Assaydesigntool

Technologie für die Long-Read-Sequenzierung

Quellen

1. Illumina. Illumina Complete Long Read Prep, Human – Datenblatt. 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. Veröffentlicht 2022. Aufgerufen am 22. September 2023.
2. Illumina. Comprehensive whole-genome sequencing with Illumina Complete Long Read Prep, Human – Technischer Hinweis. 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. Veröffentlicht 2022. Aufgerufen am 22. September 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. Veröffentlicht 2023. Aufgerufen am 22. September 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. Aufgerufen am 22. September 2023.
6. National Library of Medicine. GTR: Genetic Testing Registry. Precision HealthPGx Panel (25 Genes). ncbi.nlm.nih.gov/gtr/tests/593428/. Aktualisiert am 29. November 2022. Aufgerufen am 22. September 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. Veröffentlicht 2021. Aufgerufen am 30. August 2023.
11. Illumina. Illumina Human Comprehensive Panel – Datenblatt. 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. Veröffentlicht 2024. Aufgerufen am 26. Januar 2024.

illumina®

1 800 8094566 (USA, gebührenfrei) | +1 858 2024566 (Tel. außerhalb der USA)
techsupport@illumina.com | www.illumina.com

M-GL-02189 DEU v1.0