TECHNISCHER HINWEIS

# Verbesserung der Genauigkeit des Callings von Keimbahn-Varianten bei der DRAGEN™-Sekundäranalyse

## Optimierung der Performance beim Varianten-Calling mit maschinellem Lernen und Multigenom-Mapping von Illumina

illumina®

# Einleitung

Die Erschließung der im Genom enthaltenen Informationen durch Sequenzierung der nächsten Generation (NGS, Next-Generation Sequencing) ist entscheidend für die biomedizinische Forschung und die Präzisionsmedizin. NGS lässt sich bei Untersuchungen nur optimal einsetzen, wenn den Forscher Datenanalysetools zur Verfügung stehen, die durch Sequenzierung gewonnene Rohdaten in aussagekräftige Ergebnisse überführen. Die DRAGEN-Sekundäranalyse ermöglicht eine genaue, umfassende und effiziente Sekundäranalyse von NGS-Daten. Dank der umfassend konfigurierbaren FPGA-Technologie (Field-Programmable Gate Array, feldprogrammierbarer Gate-Array) verkürzt DRAGEN die Sekundäranalyse von NGS-Daten, z. B. Mapping, Alignment und Varianten-Calling, erheblich. Darüber hinaus beseitigt die DRAGEN-Sekundäranalyse häufige Herausforderungen bei der Genomanalyse, darunter solche im Zusammenhang mit langen Rechenzeiten, großen Datenmengen und dem Varianten-Calling in schwierigen Genomregionen.

Die DRAGEN-Sekundäranalyse liefert herausragend präzise Ergebnisse. 2020 setzte sich die DRAGEN v3.7-Sekundäranalyse bei der PrecisionFDA Truth Challenge V2 (PrecisionFDA V2) mit den präzisesten Daten bei allen Benchmark-Regionen und schwer zu mappenden Regionen gegen andere Lösungen wie Sentieon, Seven Bridges und BWA-GATK durch (Abbildung 1).[1, 2] Nach nur vier Jahren erzielt die neueste Version, die DRAGEN v4.3-Sekundäranalyse, eine deutliche Steigerung dieser bereits herausragenden Performance und bietet bislang unerreichte Genauigkeit beim Calling kleiner Varianten. Der F1-Score (ein kombiniertes Maß für Präzision und Recall) liegt dank der neuen und leistungsstarken Features in allen Benchmark-Regionen bei 99,89 %.

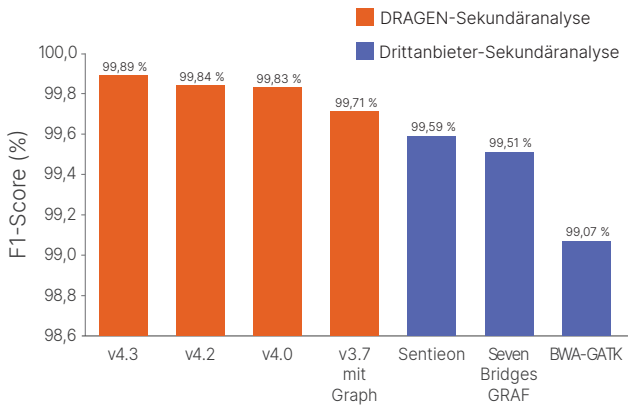


Abbildung 1: Genauigkeit der DRAGEN-Sekundäranalyse für die FDA-Analyse aller Benchmark-Regionen: Der F1-Score (%) wird anhand der Summe der richtig positiven und richtig negativen Ergebnisse als Anteil der Gesamtergebnisse berechnet.[5, 6] Höhere Scores weisen auf eine verbesserte, anhand von Referenzdaten ermittelte Genauigkeit hin.

Dieser technische Hinweis erläutert die jüngsten Verbesserungen, die zur hohen Genauigkeit der DRAGEN-Sekundäranalyse beitragen, darunter Multigenom-Mapper mit Pangenomreferenz, die Integration von maschinellem Lernen (ML), das Calling von Mosaikvarianten, Spezial-Caller und der Nachweis von strukturellen Varianten (SVs) und Kopienzahlvarianten (CNVs, Copy Number Variations).

# Multigenom-Mapper mit Pangeomreferenz

Das in der DRAGEN v3.7-Sekundäranalyse eingeführte Multigenom-Mapping verbessert die Genauigkeit des Varianten-Callings.[3] Die DRAGEN v4.3-Sekundäranalyse zeichnet sich durch eine signifikante Steigerung der Genauigkeit aus. Die Fehlerrate ist um 83 % niedriger als bei v3.6.3 und um 40 % niedriger als bei v4.2.7 (Abbildung 2).

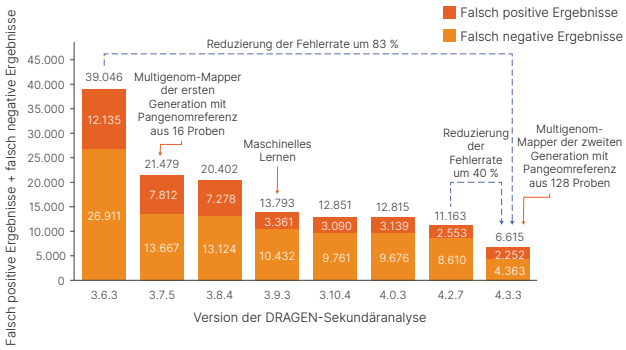


Abbildung 2: Kontinuierliche Innovationen zur Optimierung der DRAGEN-Sekundäranalyse: Verbesserungen bei Falsch-positiv- und Falsch-negativ-Raten für SNPs und Indels bei der „Genome in a Bottle"-Probe HG002, NIST v4.2.1[4] zeigen die signifikante Verringerung der Fehleranzahl, die in nur vier Jahren erreicht wurde.

Zur besseren Abbildung bestimmter Populationen bietet die DRAGEN v4.3-Sekundäranalyse Anwendern die Möglichkeit zur Erstellung einer anwendungsspezifischen Pangenomreferenz, die das Varianten-Calling bei Studien weiter verbessert. Anwender können eine anwendungsspezifische Pangenomreferenz mit eigenen Assemblys oder mit einer Auswahl der vom Human Pangenome Reference Consortium (HPRC) bereitgestellten Assemblys erstellen. Beispielsweise liefert eine anwendungsspezifische Pangenomreferenz auf Basis von 44 HPRC-Assemblys, die eine spezifische Forschungspopulation darstellen, im Vergleich zu früheren Versionen der DRAGEN-Sekundäranalyse wie DRAGEN v4.2 ein genaueres Varianten-Calling (Abbildung 3). Die in v4.3 enthaltene Standard-Pangenomreferenz (auf Basis von 128 Proben) sollte jedoch für allgemeine Anwendungen am besten geeignet sein.[4]

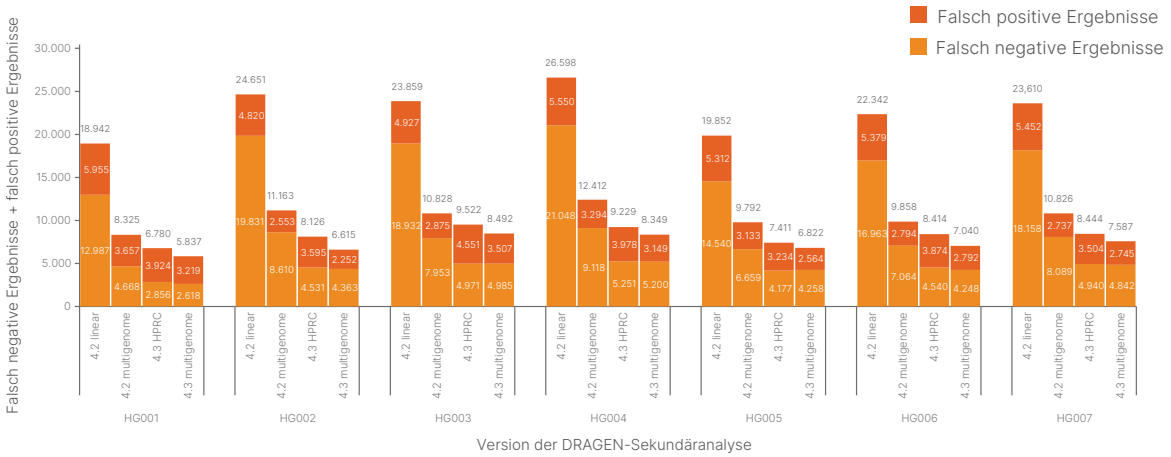


Abbildung 3: Genauigkeitsverbesserungen mit anwendungsspezifischen Referenzen beim Calling kleiner Varianten im Rahmen der DRAGEN-Sekundäranalyse: Mit der HPRC-basierten Multigenomreferenz werden bei der DRAGEN v4.3-Sekundäranalyse für die Analyse der „Genome in a Bottle"-Proben HG001–HG007 bessere Ergebnisse erzielt als mit v4.2.[4] Die Standard-Multigenomreferenz (4.3 multigenome) übertrifft nach Auswertung von 128 Proben die HPRC-basierte Referenz der Version 4.3 bei allgemeinen Anwendungen.

## Maschinelles Lernen

Das mit der DRAGEN v3.9-Sekundäranalyse eingeführte und in Version 3.10 verbesserte ML-Modul nutzt ein überwachtes Modell, das kontext- und Read-spezifische Merkmale auswertet, die anhand der Varianten-Caller der DRAGEN-Sekundäranalyse gewonnen wurden. Die Genauigkeit kleiner Varianten wird durch die Verringerung der Anzahl fehlerhafter Calls optimiert. Diese Verbesserung der Ergebnisse erfolgt mithilfe einer Kombination aus Multigenom-Mapping und ML (Abbildung 4). Bei sämtlichen Testsubjekten zeigten sich deutliche Zugewinne, auch bei Testdaten aus anderen Populationen, die beim Training nicht verwendet wurden.

## Nachweis von Mosaikvarianten

Die DRAGEN v4.3-Sekundäranalyse ermöglicht jetzt mithilfe eines neuen ML-Modells das Calling von Mosaikvarianten innerhalb des Callers für kleine Keimbahn-Varianten. Dank der Reduzierung des Schwellenwerts für die Allelfrequenz auf null erkennt die DRAGEN-Sekundäranalyse Varianten mit Allelfrequenzen < 20 %.

Die DRAGEN v4.3-Sekundäranalyse erkennt Mosaikvarianten mit höherer Genauigkeit und Präzision als frühere Versionen.

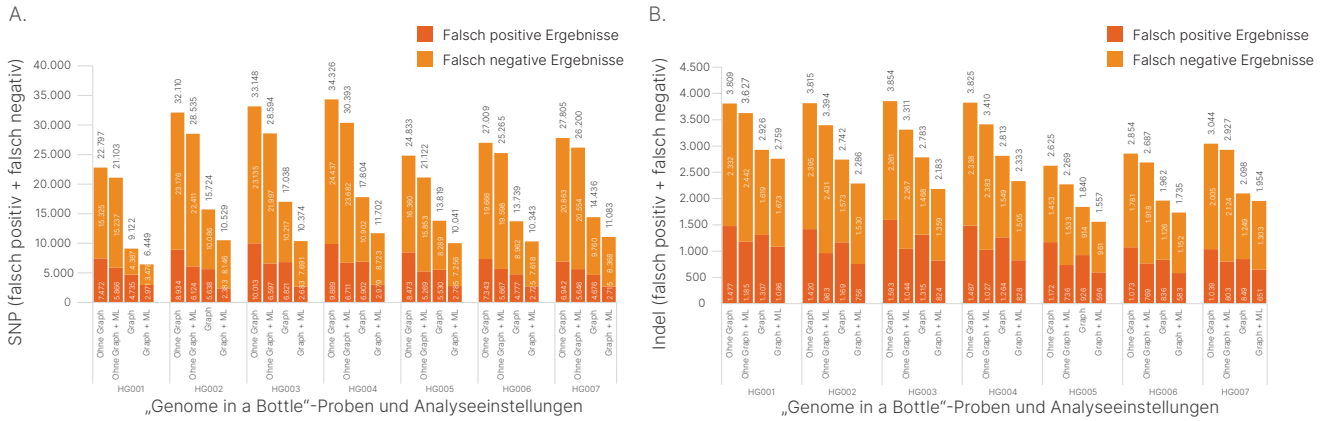


Abbildung 4: ML und Multigenom-Mapping reduzieren die Anzahl falsch positiver und falsch negativer Ergebnisse: Bei einer Analyse der „Genome in a Bottle"-Proben HG001–HG007[4] verringert ML die Anzahl der Fehler um 10 % bei deaktivierter Multigenome (graph)-Referenz und um ca. 30 % bei aktivierter Multigenome (graph)-Referenz. Wenn sowohl die Multigenomreferenz als auch ML aktiviert sind, verringert sich die Anzahl falscher Calls für (A) SNVs und (B) Indels um 62 %.

Zur Demonstration wurden vier DRAGEN-Sekundäranalyse-Pipelines mit den Mosaic-Truth-Set-Daten des National Institute of Standards and Technology (NIST) getestet: DRAGEN v4.2-Sekundäranalyse, DRAGEN v4.2-Sekundäranalyse im Hochsensitivitätsmodus (HSM), DRAGEN v4.3-Sekundäranalyse und DRAGEN v4.3-Sekundäranalyse mit aktiviertem Mosaikmodus. Das Mosaic-Truth-Set des NIST enthält 73 bekannte Mosaikvarianten in 300×-Daten, die von v4.2 und v4.3 der DRAGEN-Sekundäranalyse nicht erkannt wurden, jedoch von der DRAGEN v4.2-Sekundäranalyse im HSM und von der DRAGEN v4.3-Sekundäranalyse im Mosaikmodus. Die DRAGEN v4.3-Sekundäranalyse im Mosaikmodus erzielte jedoch eine höhere Genauigkeit beim Calling der Mosaikvarianten. Die Anzahl falsch positiver Ergebnisse war 73 % geringer als bei der DRAGEN v4.2-Sekundäranalyse im HSM (Abbildung 5).

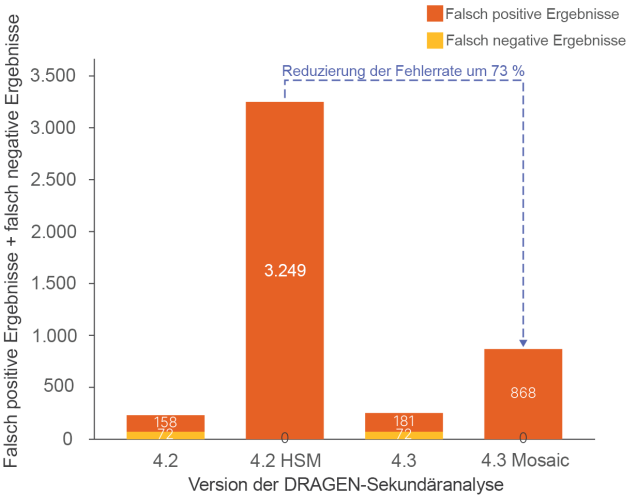


**Abbildung 5: Höhere Genauigkeit und Präzision durch den Mosaiknachweismodus:** Gegenüber der DRAGEN v4.2-Sekundäranalyse im Hochsensitivitätsmodus (HSM) zeigt sich bei DRAGEN v4.3 im Mosaiknachweismodus eine um 73 % verringerte Fehleranzahl. Aus den Daten geht zudem die hohe Anzahl falsch negativer Ergebnisse ohne aktivierten HSM-Modus oder Mosaiknachweis hervor.

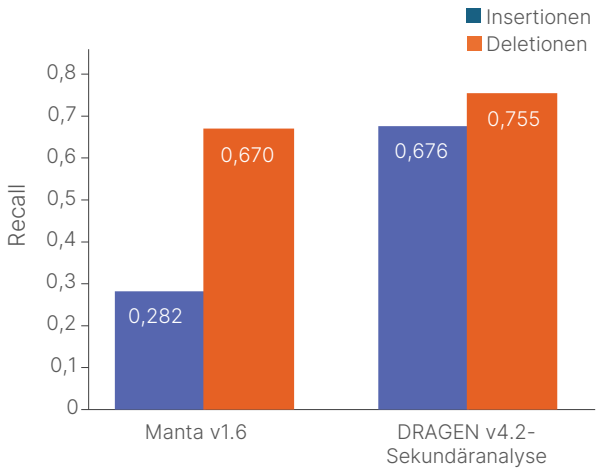


**Abbildung 6: Hochgenaue SV-Calling-Daten mit der DRAGEN-Sekundäranalyse:** Recall-Vergleich für SV-Indels zwischen der DRAGEN v4.2-Sekundäranalyse und Manta v1.6, der anhand von Benchmark-Daten für Genome in a Bottle (GIAB SV v0.6) vorgenommen wurde.[7]

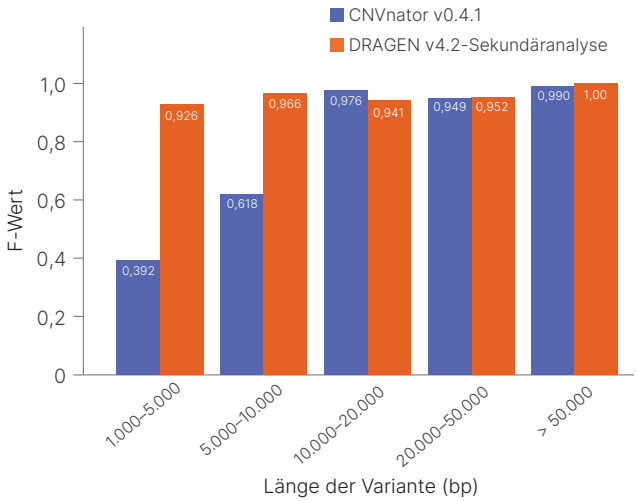


**Abbildung 7: Hochgenaues CNV-Calling mit der DRAGEN-Sekundäranalyse:** CNV-Calling mithilfe der DRAGEN v4.2-Sekundäranalyse im Vergleich zu CNVnator v1.6 über verschiedene Größen von Deletionen anhand von Benchmark-Daten für Genome in a Bottle (GIAB SV v0.6).[7]

## Nachweis von SV und CNV

Strukturelle Varianten (SVs) sind Veränderungen im Genom mit einer Länge von mindestens 50 bp. Bei Kopienzahlvarianten (CNVs) handelt es sich um einen bestimmten Typ von SVs mit reduzierter (Deletionen) oder erhöhter (Insertionen) Anzahl der Kopien einer genomischen Sequenz. Die DRAGEN-Sekundäranalyse zeichnet sich im Vergleich zu anderen Lösungen durch größere Genauigkeit beim SV-Calling (Abbildung 6) und CNV-Calling (Abbildung 7) aus.[7] Die fortschrittlichen Algorithmen und neuartigen Verfahren speziell für komplexe Regionen des Genoms heben die DRAGEN-Sekundäranalyse von anderen Lösungen ab.

Der SV-Caller von DRAGEN stellt im Vergleich zu den Manta-Calling-Verfahren für strukturelle Varianten eine Verbesserung dar und enthält Informationen aus der Pangenomreferenz, was beim Nachweis von SVs die Filterung präzisiert und die Genauigkeit erhöht. Diese Verbesserungen sind u. a. dem neuen Mobilelement-Insertionsdetektor zur Bestimmung großer Insertionen, optimierten Paarparametern zur Optimierung des Callings großer Deletionen und dem verfeinerten Contig-Alignment zur besseren Erkennung von Insertionen zu verdanken.

Des Weiteren kommen bei der DRAGEN-Software Verbesserungen an den Assemblierungsschritten, der Read-Likelihood-Berechnung sowie bei der Handhabung von überlappenden Mates sowie von geclippten Basen hinzu.

Beim CNV-Caller von DRAGEN handelt es sich in erster Linie um einen auf der Read-Tiefe basierenden Caller, der unterschiedliche Segmentierungs- und Scoringmodelle für zahlreiche Anwendungen unterstützt. Anhand zusätzlicher Signale von diskordanten und Split-Reads verbessert der CNV-Caller (wie dies auch beim SV-Calling erfolgt) die Sensitivität bei der Erfassung von Ereignissen von nur 1 kbp.

Zusätzlich umfasst der CNV-Caller von DRAGEN auch das Segmental Duplication Extension-Modul. Dieses Feature ermöglicht den CNV-Nachweis in Regionen des Genoms mit segmentalen Duplikationen. Bei Regionen mit segmentalen Duplikationen handelt es sich um Regionen des Genoms mit einer Übereinstimmung der Sequenz von > 90 %, die ca. 5 % des Genoms ausmachen. Diese Regionen zeichnen sich durch unzureichende Mappability aus, was den Variantennachweis in diesen Regionen schwierig macht. Segmental Duplication Extension gewinnt ca. eine Million Basen in CNV-Regionen zurück, die zuvor von der Analyse ausgeschlossen wurden. Dies ermöglicht den CNV-Nachweis bei 43 medizinisch relevanten Genen und erhöht die allgemeine Genauigkeit des Varianten-Callings.

## Spezialisierte und gezielte Caller

Gezielte Caller ermöglichen die genaue Genotypisierung spezifischer Gene, die aufgrund von Faktoren wie hoher Ähnlichkeit der Sequenz mit Pseudogenen, repetitiven Regionen und starkem Polymorphismus schwer zu analysieren sind. Die DRAGEN-Sekundäranalyse begegnet diesen Herausforderungen mithilfe unterschiedlicher gezielter Caller (Tabelle 1), die eine präzise Genotypisierung medizinisch relevanter Gene ermöglichen. Zur Gewinnung von Erkenntnissen für die Pharmakogenomik (PGx) bestimmt der PGx Star Allele Caller Stern-Allele und den Metabolisierungstyp für 22 Gene (Tabelle 2).

Der DRAGEN-Caller für humane Leukozytenantigene (HLA) ermöglicht eine hochpräzise Genotypisierung von HLA-Allelen der Klassen I und II. Der Caller aligniert die Reads anhand einer umfassenden Datenbank, die über 9.000 Allele enthält, und kann in Bereichen wie der Übereinstimmungsprüfung für die Organtransplantation, der Immunogenetik und bei Krankheitsassoziationsstudien eingesetzt werden.

Tabelle 1: Zusammenfassung der Zielgene der gezielten und spezialisierten Caller.

| Gezielter Caller | Forschungsbereich | Assoziierte Erkrankung |
|---|---|---|
| *CYP21A2* | Carrier-Screening | Kongenitale adrenale Hyperplasie (CAH, Congenital Adrenal Hyperplasia) |
| *HBA* | Carrier-Screening | α-Thalassämie |
| *GBA* | Carrier-Screening | Morbus Gaucher, Morbus Parkinson |
| *SMN* | Carrier-Screening | Spinale Muskelatrophie |
| *LPA* | Kardiovaskuläre Erkrankungen | Koronare Herzkrankheit |
| *RH* | Blutgruppen-bestimmung | – |
| *CYP2B6* | PGx | – |
| *CYP2D6* | PGx | – |
| *HLA* | Übereinstimmungs-prüfung für die Transplantation, Immunogenetik | – |

Tabelle 2: Zielgene des PGx Star Allele Caller mit Relevanz für die PGx

| Gensymbol | | |
|---|---|---|
| *ABCG2* | *CYP4F2* | *RYR1* |
| *BCHE* | *DPYD* | *SLCO1B1* |
| *CACNA1S* | *F5* | *TPMT* |
| *CFTR* | *G6PD* | *UGT1A1* |
| *CYP2C19* | *IFNL3* | *UGTB17* |
| *CYP2C9* | *MT-RNR1* | *VKORC1* |
| *CYP3A4* | *NAT2* | |
| *CYP3A5* | *NUDT15* | |

Mit der DRAGEN v4.3-Sekundäranalyse wird eine neue Caller-Klasse für den Nachweis von *De-novo*-Varianten in Regionen mit segmentalen Duplikationen eingeführt. Der MRJD-Caller (Multiregion Joint Detection) implementiert einen haplotypbasierten *De-novo*-Caller für kleine Varianten für sechs medizinisch relevante Gene in Regionen mit segmentalen Duplikationen (Tabelle 3).

Tabelle 3: Zielgene des MJRD-Callers

| Gezielter Caller | Forschungsbereich | Assoziierte Erkrankung |
|---|---|---|
| *PMS2* | Screening auf erblich bedingte Krebserkrankungen | Lynch-Syndrom-assoziierte Kolorektal-/Endometriumkarzinome |
| *SMN1* (kleine Varianten) | Carrier-Screening | Spinale Muskelatrophie |
| *STRC* | Carrier-Screening | Nicht syndromaler Hörverlust |
| *NEB* | Carrier-Screening | Nemalin-Myopathie |
| *TTN* | Neugeborenen-Screening, seltene Erkrankungen | Kardiomyopathie |
| *IKBKG* | Neugeborenen-Screening | Incontinentia pigmenti, hypohidrotische ektodermale Dysplasie |

# Zusammenfassung

Die DRAGEN-Sekundäranalyse ermöglicht bei NGS-Anwendungen eine hochgenaue, umfassende und effiziente Sekundäranalyse. Kontinuierliche Verbesserungen erhöhen die Genauigkeit und erweitern die Coverage schwieriger Regionen des Genoms, was den Nachweis herausfordernder und medizinisch relevanter Varianten ermöglicht.

# Anhang

## Multigenom-Mapping mit Pangenomreferenz

Die Verwendung von Populationshaplotypen phasierter Varianten und die Erweiterung des Referenzindex durch auf Basis der Population ermittelte Alt-Contigs ermöglicht der DRAGEN-Sekundäranalyse das effektive Mapping anhand einer Pangenomreferenz und die Optimierung des Mappings von Illumina-Reads in schwierigen Regionen. Dieses neue Feature erweitert wirksam die Reichweite von Illumina-Reads und ermöglicht das genaue Mapping und ein entsprechendes Varianten-Calling in Regionen, die zuvor nicht ausgewertet werden konnten.

Bei einem Multigenom-Mapper handelt es sich um ein Verfahren zur Unterstützung des Mappings anhand von Populationsdaten, bei dem in der Population ermittelte alternierende Sequenzinhalte als unterschiedliche divergierende und konvergierende Pfade dargestellt werden (Abbildung 8). Proben-Reads lassen sich mit dem Multigenom-Mapper auf den Pfad mit der besten Übereinstimmung alignieren.

Weitere Informationen finden Sie unter The quest for accuracy gains in the dark regions of the genomes: Presenting the DRAGEN multigenome mapper and pangenome reference updates in version 4.3 (Genauigkeitssteigerung in den dunklen Regionen des Genoms: Neuerungen in der DRAGEN v4.3: Multigenom-Mapper und Pangenomreferenz).

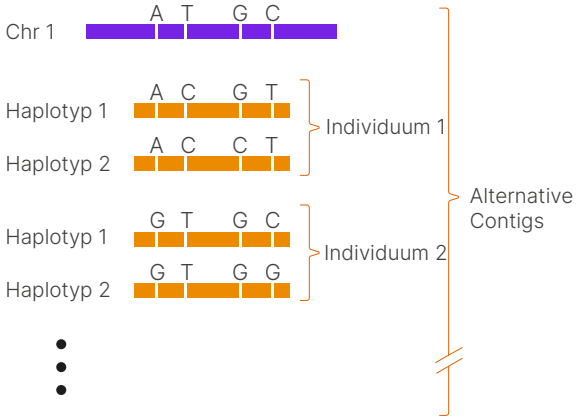


Abbildung 8: Multigenom-Mapper mit Pangenomreferenz: In einer Referenz wird der in einer Population erfasste alternierende Sequenzinhalt in Form unterschiedlicher divergierender und konvergierender Pfade dargestellt.

## Alt-Masking

Seit der DRAGEN v3.9-Sekundäranalyse umfasst die DRAGEN Software Alt-Masking, ein Verfahren zur Handhabung nativer Referenz-ALT-Contigs, bei dem strategische Positionen der ALT-Contigs maskiert werden, um die Genauigkeit zu erhöhen. Dieses Verfahren lässt sich im Laufe der Zeit einfach definieren, pflegen und verfeinern.

Weitere Informationen finden Sie unter DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph (DRAGEN setzt neue Standards für die Datengenauigkeit bei PrecisionFDA-Benchmarkdaten. Optimierung der Performance beim Varianten-Calling mit maschinellem Lernen und DRAGEN graph von Illumina).

## Maschinelles Lernen

Seit der DRAGEN v3.9-Sekundäranalyse umfasst die Software eine leistungsstarke und effiziente ML-Rekalibrierungspipeline, die im Rahmen des Workflows für kleine Keimbahn-Varianten eingesetzt werden kann. In der DRAGEN v4.0-Sekundäranalyse ist diese Option standardmäßig aktiviert. Wenn aktiviert, führt die Pipeline das ML-Modell nach dem standardmäßigen Varianten-Calling aus. In diesem Schritt werden die QUAL- und GQ-Felder neu kalibriert, die in der finalen VCF-Datei ausgegeben werden. In bestimmten Fällen kann ML den GT ändern. Die Werte dieser Felder vor dem Einsatz von maschinellem Lernen werden in den Feldern DQUAL, DGT und DGQ beibehalten, sodass keine Informationen verloren gehen. Dieser Schritt verlängert den Standardworkflow für einen WGS-Keimbahnlauf mit 30-facher Coverage um ca. 5 Minuten. Die höhere Genauigkeit beeinträchtigt die Gesamtlaufzeit also nur geringfügig.

Das ML-Modell wurde mithilfe von überwachtem Offline-Training generiert. Das Modell verarbeitet eine Reihe von Read- und kontextabhängigen Merkmalen, was die Genauigkeit der Qualitäts-Scores des Callers für kleine Varianten erhöht. Die Merkmale, die zum Training des Modells verwendet werden, umfassen die Mappability, AF, VC-Qual., DP, GC-Inhalt, Abweichungen und andere interne Mapping-, Alignment- und VC-Metriken.

## Berechnung des F1-Scores

$$F1 = 2 \times (Recall \times Präzision)/(Recall + Präzision)$$

$$F1_{übergeordnet} = \sqrt{F1_{HG003} \times F1_{HG004}}$$

## DRAGEN-Befehlszeile

Starterrezepte finden Sie auf der Seite DRAGEN Recipe Germline WGS (DRAGEN-Rezept – Keimbahn-WGS).

# Quellen


1. Food and Drug Administration. Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. precision.fda.gov/challenges/10/results. Aufgerufen am 19. September 2024.
2. Illumina. DRAGEN sets new standard for data accuracy in PrecisionFDA benchmark data. Optimizing variant calling performance with Illumina machine learning and DRAGEN graph. illumina.com/science/genomics-research/articles/dragen-shines-again-precisionfda-truth-challenge-v2.html. Veröffentlicht am 12. Januar 2022. Aufgerufen am 19. September 2024.
3. Illumina. The quest for accuracy gains in the dark regions of the genomes: Presenting the DRAGEN multigenome mapper and pangenome reference updates in version 4.3. illumina.com/science/genomics-research/articles/second-gen-multigenome-mapping.html. Veröffentlicht am 12. August 2024. Aufgerufen am 30. September 2024.
4. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025. Veröffentlicht am 7. Juni 2016. doi:10.1038/sdata.2016.25
5. Illumina. DRAGEN wins at PrecisionFDA Truth Challenge V2 showcase accuracy gains from alt-aware mapping and graph reference genomes. illumina.com/science/genomics-research/articles/dragen-wins-precisionfda-challenge-accuracy-gains.html. Aufgerufen am 19. September 2024.
6. Interne archivierte Daten. Illumina, Inc., 2022.
7. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. Online veröffentlicht am 25. Oktober 2024. doi:10.1038/s41587-024-02382-1


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


M-GL-01016 DEU v3.0