illumina®

Verbesserung der Genauigkeit des Callings kleiner Varianten bei Keimbahnen mit der DRAGEN™-Plattform

Verschiedene Algorithmen zur Verbesserung der Genauigkeit ermöglichen den Nachweis kleiner Varianten mit hoher Sensitivität und Spezifität unter Beibehaltung der standardmäßigen Berechnungsgeschwindigkeit des DRAGEN-Systems.

Einleitung

Durch die Fortschritte der NGS (Next-Generation Sequencing, Sequenzierung der nächsten Generation) nimmt die Menge der Sequenzierungsdaten exponentiell zu. Aufgrund dieser zunehmenden Datenmenge werden schnelle und effiziente Analysemethoden benötigt, die hohe Genauigkeitsstandards beim Varianten-Calling gewährleisten. Die Illumina DRAGEN (Dynamic Read Analysis for Genomics) Bio-IT-Plattform ermöglicht eine höchst genaue und extrem schnelle Sekundäranalyse von NGS-Daten. Die DRAGEN-Plattform beschleunigt dank der umfassend konfigurierbaren FPGA-Technologie (Field-programmable Gate Array Technology) die Sekundäranalyse von NGS-Daten, z. B. Mapping, Alignment und Varianten-Calling, erheblich.

Grundlegende Funktionen der DRAGEN-Plattform stellen eine Lösung für häufige Probleme bei der Genomanalyse bereit, beispielsweise die lange Berechnungsdauer und die riesigen Datenmengen. Die DRAGEN-Plattform bietet hohe Verarbeitungsgeschwindigkeiten, Flexibilität, Genauigkeit und Kosteneffektivität. Da sich die DRAGEN-Plattform neu programmieren lässt, können die Algorithmen hinsichtlich künftiger Anwendungen verbessert und angepasst werden. Dank der Geschwindigkeit der Plattform können Entwickler Algorithmen mithilfe rechenintensiver Methoden, für die reine Softwarelösungen ungeeignet wären, rasch überarbeiten. Daher konnte die Genauigkeit der DRAGEN-Plattform mit jeder neuen Version verbessert werden. DRAGEN ist inzwischen eine hervorragende Lösung für das Calling kleiner Varianten bei der Keimbahn-Gesamtgenom-Sequenzierung (WGS, Whole-Genome Sequencing).

In diesem Anwendungshinweis werden aktuelle Verbesserungen der Illumina DRAGEN Bio-IT-Plattform hinsichtlich der schnellen Sekundäranalyse beschrieben und Geschwindigkeit und Genauigkeit anhand von drei öffentlich zugänglichen WGS-Datensätzen demonstriert. Wir stellen Benchmarkvergleiche zwischen DRAGEN v3.2.8 und anderen Anwendungen bereit, z. B. BWA-MEM+GATK4 und DRAGEN v2 (Abbildung 1). Die Ergebnisse des Varianten-Callings der einzelnen Anwendungen wurden mit Referenz-Calls verglichen, um falsch positive (FP) und falsch negative (FN) Werte zu ermitteln. Als Metriken wurden für diese Vergleiche die Gesamtlaufzeiten sowie Recall, Präzision und FP+FN herangezogenen. Durch die Kombination von Geschwindigkeit, Genauigkeit und Anwendungsflexibilität revolutioniert die DRAGEN-Plattform die Genomanalyse.

DRAGEN v3-Algorithmen verbessern Genauigkeit

In DRAGEN v3 wurden die neuesten Algorithmusaktualisierungen für die Erkennung von Einzelnukleotid-Polymorphismen (SNPs, Single-Nucleotide Polymorphisms) und Insertionen/Deletionen (Indels) implementiert. Dies führt zu Verbesserungen bei der Geschwindigkeit und der analytischen Sensitivität. Das Varianten-Calling wurde in vier Bereichen optimiert: probenspezifisches Indel-Fehlermodell, präzise mathematische Modelle auf Grundlage korrelierter Pileup-Fehler, optimierte Darstellung einer exponentiellen Anzahl von Haplotyp-Kandidaten in variantenreichen oder Regionen mit starkem Rauschen und spaltenweise Ergänzung der Liste mit Ereignissen, die durch Assemblierung von De-Bruijn-Graphen generiert wurden. Im Vergleich zu den in dieser Untersuchung vorgestellten Anwendungen ermöglichen diese Upgrades leichte Verbesserungen der Geschwindigkeit sowie Steigerungen der Genauigkeitsstandards. Im Anhang werden die Verbesserungen der jeweiligen Algorithmen ausführlich erläutert.

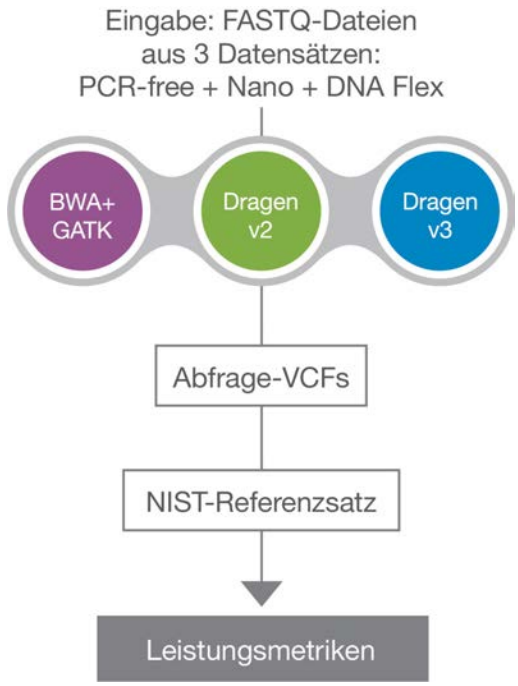

Abbildung 1: Entwurf einer Vergleichsstudie mit Benchmarks: FASTQ-Dateien aus drei Datensätzen wurden mit drei Analyseanwendungen verarbeitet, um VCF-Abfragedateien zu erstellen. Anschließend wurden auf Grundlage eines Vergleichs der Varianten-Calls mit Referenzvarianten im NIST-Vergleichssatz mithilfe des Variant Calling Assessment Tools (VCAT) TPs, FPs und FNs ermittelt.

Methoden

Empfohlene Best Practices für das Festlegen von Benchmarks wurden genau befolgt.[1] Für die Demonstration der mit DRAGEN v3 erzielten Geschwindigkeit und Genauigkeit wurden drei Datensätze aus unterschiedlichen Bibliotheksvorbereitungen verwendet, die mit der NA12878-Probe erstellt wurden (Abbildung 1). Als Eingabe für die Sekundäranalyse durch unabhängige Anwendungen (DRAGEN v3.2.8, DRAGEN v2 und BWA+GATK[2]) wurde aus jedem Datensatz die FASTQ-Datei verwendet. Die VCF-Ergebnisdateien (QUERY VCF) von jeder Anwendung wurden in BaseSpace™ Sequence Hub in ein Projekt geladen. Mit dem Variant Calling Assessment Tool (VCAT v3.1.1 mit Hap.py v0.3.10) wurden die QUERY VCF-Dateien einzeln mit einem Referenzsatz verglichen, um wahre oder falsche Varianten-Calls zu identifizieren. Die Ergebnisse wurden für einen Vergleich der Anwendungen in Tabellen aufbereitet. Alle Eingabedaten, Analyseergebnisse und Auswertungstools stehen im BaseSpace-Projekt zur Verfügung.[3] Eine ausführlichere Beschreibung der Methoden finden Sie im Anhang.

Ergebnisse des Benchmarking-Vergleichs

Die Vergleiche der Laufzeiten und der erzielten Genauigkeit zeigen, dass DRAGEN eine leistungsstarke Lösung für die Sekundäranalyse von NGS-Daten darstellt.

DRAGEN – Genauigkeit: FP+FN, Recall und Präzision

Obwohl bereits DRAGEN v2 branchenführende Berechnungslösungen bereitstellte, bietet DRAGEN v3 mit verschiedenen Verbesserungen (beschrieben im Abschnitt zu den Algorithmusmethoden) eine erhebliche Steigerung der Genauigkeit. Die Ergebnisse dieses Benchmarking-Vergleichs zeigen außerdem, dass DRAGEN v3 dank der Verbesserungen hinsichtlich sämtlicher Genauigkeitsmetriken anderen Analyseanwendungen (einschließlich älterer DRAGEN-Versionen) überlegen ist.

In Bezug auf die FP+FN-Metrik bei der SNV-Erkennung war die Performance von DRAGEN v3 gegenüber der Performance von BWA+GATK4 und DRAGEN v2 bei allen drei Datensätzen deutlich überlegen (Abbildung 2). Hinsichtlich der FP+FN-Metrik bei der Indel-Erkennung zeigte DRAGEN v3 bei allen drei Datensätzen eine bessere Leistung als BWA+GATK4. Außerdem wurden die Verbesserungen gegenüber DRAGEN v2 deutlich (Abbildung 3).

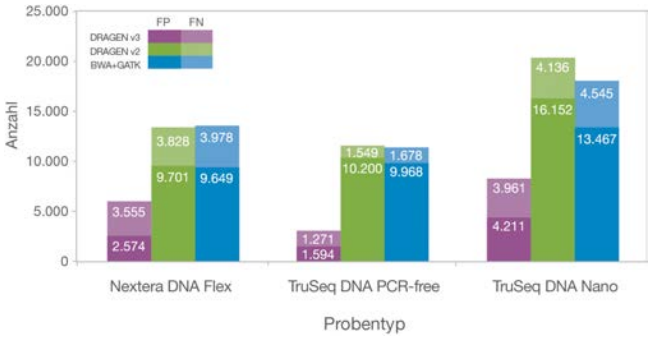

Abbildung 2: Falsch positive und falsch negative Werte bei der SNV-Erkennung: Es wurden Dateien mit Rohdaten (FASTQ) aus drei Datensätzen mit drei unabhängigen Anwendungen analysiert. Jeder Datensatz (TruSeq DNA PCR-free, Nextera DNA Flex und TruSeq DNA Nano) wurde aus NA12878-Proben-DNA erstellt. Varianten-Calls (VCF) aus jeder Analyseanwendung wurden mit einem (ebenfalls auf der NA12878-Probe basierenden) NIST-Referenzsatz verglichen, um FPs und FNs zu ermitteln.

Die Auswertung der Metriken für Präzision und Recall veranschaulicht die Vorteile der verbesserten Algorithmen in DRAGEN v3 für die SNP- und Indel-Erkennung. Die Werte für Präzision und Recall übertreffen für jede Anwendung und bei jedem Datensatz für die SNV-Erkennung regelmäßig 99 % (Tabelle 1). Was die SNP-Erkennung betrifft, war die Leistung von DRAGEN v2 mit der von BWA+GATK4 vergleichbar. DRAGEN v3 weist jedoch gegenüber diesen beiden Anwendungen deutliche Verbesserungen sowohl im Bereich Präzision als auch im Bereich Recall auf. Bei der Indel-Erkennung wies DRAGEN v2 eine höhere Genauigkeit als BWA+GATK4 auf. DRAGEN v3 zeigte eine weitere Verbesserung gegenüber DRAGEN v2 bei Recall und Präzision (Tabelle 2).

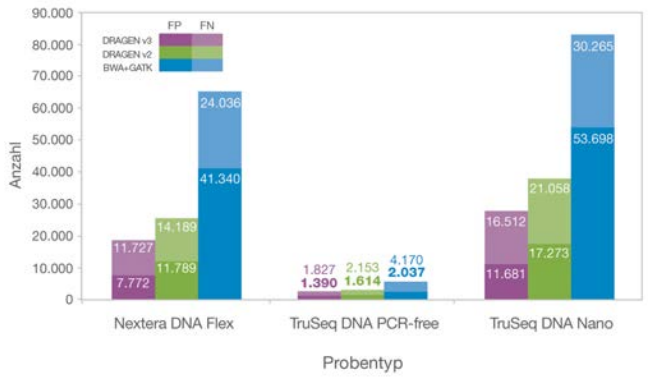

Abbildung 3: Falsch positive und falsch negative Werte bei der Indel-Erkennung: Es wurden Dateien mit Rohdaten (FASTQ) aus drei Datensätzen mit drei unabhängigen Anwendungen analysiert. Jeder Datensatz (TruSeq DNA PCR-free, Nextera DNA Flex und TruSeq DNA Nano) wurde aus NA12878-Proben-DNA erstellt. Varianten-Calls (VCF) aus jeder Analyseanwendung wurden mit einem (ebenfalls auf der NA12878-Proben-DNA basierenden) NIST-Referenzsatz verglichen, um FPs und FNs zu ermitteln.

Tabelle 1: Sensitivität und Spezifität der SNV-Erkennung

	Präzision			Recall		
Datensätze	DRAGEN v3	DRAGEN v2	BWA+GATK	DRAGEN v3	DRAGEN v2	BWA+GATK
TruSeq DNA PCR-free	99,95 %	99,68 %	99,69 %	99,96 %	99,95 %	99,95 %
Nextera DNA Flex	99,92 %	99,70 %	99,70 %	99,89 %	99,88 %	99,88 %
TruSeq DNA Nano	99,87 %	99,50 %	99,58 %	99,88 %	99,87 %	99,86 %

Tabelle 2: Sensitivität und Spezifität der Indel-Erkennung

	Präzision			Recall		
Datensätze	DRAGEN v3	DRAGEN v2	BWA+GATK	DRAGEN v3	DRAGEN v2	BWA+GATK
TruSeq DNA PCR-free	99,71 %	99,66 %	99,58 %	99,62 %	99,55 %	99,13 %
Nextera DNA Flex	98,37 %	97,54 %	91,53 %	97,56 %	97,05 %	95,01 %
TruSeq DNA Nano	97,56 %	96,39 %	89,37 %	96,57 %	95,63 %	93,71 %

DRAGEN – Geschwindigkeit

Im Rahmen der Untersuchungen zur Laufzeit wurden die entsprechenden Werte sowohl für die cloudbasierte als auch für die lokal implementierte DRAGEN-Lösung gesammelt. Die lokale Installation von DRAGEN v3 wurde mit einer BWA+GATK-Installation verglichen, die auf dem gleichen Server ausgeführt wurde. Die cloudbasierte Version von DRAGEN v3 auf BaseSpace Sequence Hub wurde mit BWA+GATK auf Terra verglichen.[4]

DRAGEN beschleunigt sowohl das Mapping als auch das Varianten-Calling. Beide Vorgänge können unabhängig voneinander ausgeführt werden. Obwohl nicht in dieser Untersuchung berücksichtigt, sei darauf hingewiesen, dass im Vorfeld der Sekundäranalyse DRAGEN auch eine schnellere BCL2FASTQ-Konvertierung unterstützt. Identische FASTQs werden noch schneller und effizienter erstellt. Außerdem gibt DRAGEN automatisch eine ausführliche Liste mit Qualitätssicherungsmetriken für Mapping und Varianten-Calling aus. Dieser Vorgang hat kaum Auswirkungen auf die Laufzeit. Dies stellt gegenüber Anwendungen, die für die Bereitstellung von Qualitätssicherungsmetriken auf langsame Drittanbietertools angewiesen sind (z. B. Samtools, Picard), eine erhebliche Verbesserung dar.

Bei der Messung der Ausführungsgeschwindigkeiten der Anwendungen auf dem gleichen lokalen Server zeigte sich DRAGEN v3 16- bis 18-mal schneller als BWA+GATK (Abbildung 4).

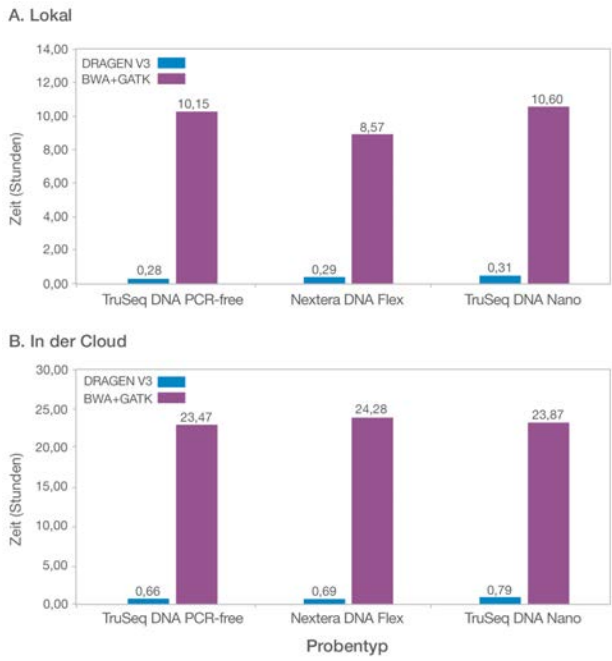

Abbildung 4: Laufzeitvergleich von lokal und in der Cloud durchgeführten Analysen: (A) Die Analyse mit DRAGEN v3 und BWA+GATK wurde auf dem gleichen lokalen Server ausgeführt. (B) Die Analyse mit DRAGEN v3 auf BaseSpace Sequence Hub wurde mit der Analyse mit BWA+GATK auf Terra verglichen.

Ein Vergleich der cloudbasierten Anwendungen ergab, dass DRAGEN v3 auf BaseSpace Sequence Hub 13- bis 16-mal schneller als BWA+GATK auf Terra war.

Zusammenfassung

Genomanwendungen ermöglichen zunehmend präzise Bestimmungen komplexer Genomregionen sowie das Messen von Calls mit niedriger Variantenallelhäufigkeit in Proben mit starkem Rauschen. DRAGEN erweist sich als optimale Plattform für die effiziente und genaue Verarbeitung von NGS-Daten.

Dank der Geschwindigkeit von DRAGEN sind Forscher auf den steigenden Durchsatz von NGS-Geräten vorbereitet. Genauso wichtig ist die Möglichkeit rascher Überarbeitung und kontinuierlicher Verbesserung der Algorithmen, um eine hohe Genauigkeit zu gewährleisten.

Anhang

Ausführliche Beschreibung neuer Algorithmen

Probenspezifisches PCR-Fehlermodell

Eine der Schwierigkeiten beim Varianten-Calling ist die Unterscheidung zwischen Indel-Fehlern und echten Varianten. Zu diesem Zweck wird für Varianten-Calls häufig ein Hidden Markov Model (HMM) verwendet, um im Rahmen der Wahrscheinlichkeits- berechnung das statistische Verhalten von Indel-Fehlern zu modellieren. Die Eingabeparameter des HMM, Gap Open Penalty (GOP) und Gap Continuation Penalty (GCP), stehen in direkter Beziehung zur Indel-Fehlerrate (Indel-Fehlerrate = f(GOP,GCP)). Indel-Fehler sind bei Vorkommen von kurzen Tandemwiederholungen (STR, Short Tandem Repeat) wahrscheinlicher. Die Fehlerwahrscheinlichkeit (und dementsprechend GOP und GCP) kann von der Dauer und der Länge der STR abhängen. Der Fehlerprozess kann in Abhängigkeit verschiedener Faktoren wie PCR-Amplifikation zwischen den einzelnen Datensätzen erheblich abweichen. Für eine genaue Erkennung sind HMM-Parameter erforderlich, die den Fehlerprozess auf Einzelprobenbasis präzise modellieren. Typische Varianten-Caller arbeiten jedoch mit starren Parametern oder nicht probenspezifischen, im Voraus festgelegten Funktionen, mit denen der Fehlerprozess nicht genau modelliert werden kann. Dies beeinträchtigt die Erkennungsleistung.

Die in DRAGEN v3 implementierte automatische HMM-Kalibrierung behebt das oben genannte Problem, indem die PCR-Parameter direkt anhand des zu verarbeitenden Datensatzes geschätzt werden. Dieser Vorgang erfolgt im Anschluss an Mapping und Alignment vor dem Varianten-Calling, ohne Einbeziehung von Referenzdaten und ohne Verwendung von externen Datenquellen mit bekannten Mutationen. Die Parameter hängen von der STR-Sequenz und der Wiederholungslänge ab.

Für eine gegebene STR-Sequenz und -Länge wird ein Satz von N Loci mit der gewünschten Sequenz und Länge gewählt. Der Algorithmus ermittelt die diesen Loci zugeordneten Read-Pileups und zählt die an den einzelnen Loci ermittelten Indels. Das Konzept beruht darauf, dass es durch die Einbeziehung einer ausreichenden Anzahl an Loci möglich ist, die Parameter von Interesse genau zu schätzen. Hierfür ermitteln wir die Parameter, die die Wahrscheinlichkeit einer Generierung des Satzes von N ermittelten Pileups maximieren. Wenn die Anzahl der Parameter zur Maximierung der Wahrscheinlichkeit klein genug ist (z. B. 2), kann eine erschöpfende Suche erfolgen. In der aktuellen Implementierung von DRAGEN v3 erfolgt die Optimierung anhand von zwei Parametern: GOP und Alpha. Letzterer gibt die Wahrscheinlichkeit von Indel-Varianten von einer beliebigen Länge außer null an. Für jede berücksichtigte STR-Sequenz und -Länge gibt die Suche die Werte GOP und Alpha aus, welche die Wahrscheinlichkeit einer Generierung des Satzes von N ermittelten Pileups maximieren. Diese Werte werden als Eingabe für das HMM verwendet. Eine Erweiterung der Suche auf mehr als 2 Parameter ist ebenfalls möglich und würde weitere Verbesserungen bringen.

Abfallen der Basenqualität (Base Quality Dropoff, BQD)

Beim Aufbau herkömmlicher Varianten-Caller wird davon ausgegangen, dass Sequenzierungsfehler von Read zu Read unabhängig auftreten. Folgt man dieser Annahme, ist es sehr unwahrscheinlich, dass mehrere identische Fehler an einem bestimmten Locus auftreten. Nach der Analyse von NGS-Datensätzen fiel jedoch auf, dass Häufungen von Fehlern wesentlich häufiger auftreten, als bei der Unabhängigkeitsvermutung anzunehmen war. Diese Häufungen können eine hohe Zahl falsch positiver Ergebnisse zur Folge haben.

Glücklicherweise unterscheiden sich diese Fehler in deutlichen Merkmalen von echten Varianten. Bei dem in DRAGEN v3 implementierten BQD (Base Quality Drop-off)-Algorithmus handelt es sich um einen Erkennungsmechanismus, der sich bestimmte Eigenschaften solcher Fehler zunutze macht (Strangverschiebung, Lokalisierung des Fehlers im Read, geringe mittlere Basenqualität am Locus von Interesse) und sie auf einfache und robuste Weise in die Wahrscheinlichkeits-berechnung zur Genotypisierung einbezieht. Neue Hypothesen für mögliche Genotypen werden zur vorhandenen Liste diploider Genotypen (die auf der Annahme unabhängiger Pileup-Fehler basieren) hinzugefügt. So fügen wir beispielsweise im Falle eines Locus mit einem ALT-Allel zusätzlich zur Berücksichtigung von P(G00|R), P(G01|R) und P(G11|R) als zwei weitere Hypothesen P(G00,E1|R) und P(G11,E0|R) hinzu, wobei die Allele E0 und E1 das Referenzallel und das aus einem Sequenzierungsfehler stammende ALT-Allel repräsentieren. Die Eigenschaften dieser Fehler wie Strangverschiebung, Lokalisierung des Fehlers im Read und die mittlere Basenqualität werden in die Berechnung von P(G00,E1|R) und P(G11,E0|R) mit einbezogen. Anschließend wird der resultierende Genotyp max (max(P(G00|R), P(G00,E1|R)), P(G01|R), max(P(G11|R), P(G11,E0|R))) übernommen.

Durch die Möglichkeit, korrelierende Sequenzierungsfehler aus dem Kern des Varianten-Caller-Ergebnisses zu charakterisieren, ergibt sich ein deutlicher Anstieg der Spezifität, da viele falsch

positive Calls beseitigt werden. Durch die Korrektur von Genotypfehlern wird zudem die Sensitivität verbessert.

Foreign Read Detection (FRD)

Herkömmliche Varianten-Caller behandeln Mapping-Fehler für jeden Read als unabhängige Fehler und ignorieren dabei die Tatsache, dass diese Art von Fehlern üblicherweise als Häufung auftritt. Das kann zu Varianten-Calls führen, die trotz geringer MAPQ und/oder verzerrter AF mit Scores mit sehr hoher Konfidenz ausgegeben werden. Um dieses Problem zu minimieren, filtern herkömmliche Varianten-Caller im Vorfeld des Varianten-Callings Reads anhand eines MAPQ-Grenzwertes heraus (d. h. Reads mit einer MAPQ < Grenzwert werden aus der Berechnung ausgeschlossen). Dadurch wird jedoch wertvolle Evidenz aus dem Varianten-Caller ausgesondert und das Verfahren ist zur Unterdrückung falsch positiver Varianten weniger geeignet.

In DRAGEN v3 ist Foreign Read Detection (FRD) implementiert, eine Erweiterung des bestehenden Genotypisierungs-Algorithmus durch Einbindung der zusätzlichen Hypothese, dass es sich bei einigen Reads im Pileup um fremde Reads handelt (deren eigentlicher Locus eine andere Stelle im Referenzgenom ist und/oder deren Herkunft außerhalb des Referenzgenoms liegt (Probenkontamination)). Der Algorithmus nutzt mehrere Eigenschaften aus (verzerrte Allelfrequenz und geringe MAPQ) und bindet diese Evidenz mathematisch präzise in die Wahrscheinlichkeitsberechnung ein.

Neue Hypothesen für mögliche Genotypen werden zur vorhandenen Liste diploider Genotypen (die auf der Annahme unabhängiger Pileup-Fehler basieren) hinzugefügt. So fügen wir beispielsweise im Falle eines Locus mit einem ALT-Allel zusätzlich zur Berücksichtigung von P(G00|R), P(G01|R) und P(G11|R) als zwei weitere Hypothesen P(G00,F1|R) und P(G11,F0|R) hinzu, wobei die Allele F0 und F1 das Referenzallel und das aus einem Mapping-Fehler stammende ALT-Allel repräsentieren. Die Eigenschaften dieser Fehler wie Alleltiefe und MAPQ werden in die Berechnung von P(G00,F1|R) und P(G11,F0|R) mit einbezogen. Anschließend wird der resultierende Genotyp max (max(P(G00|R), P(G00,F1|R)), P(G01|R), max(P(G11|R), P(G11,F0|R))) übernommen.

Die Sensitivität wird durch die Erhaltung von FN, die Korrektur von Genotypen und die Möglichkeit, den MAPQ-Grenzwert für in den Varianten-Caller eingehende Reads zu senken, verbessert. Die Spezifität wird durch das Entfernen von FP und die Korrektur von Genotypen verbessert.

Bei FRD handelt es sich verglichen mit der Filterung nach den Varianten-Calls um ein leistungsstärkeres Tool. Anstatt nur auffällige Ergebnisse (z. B. aufgrund von Alleltiefe oder Read-Fehlern) nach dem Varianten-Calling zu erkennen, berücksichtigt der Erkennungsalgorithmus durch eine strenge Maximum-Likelihood-Erkennung unmittelbar das Vorhandensein fremder Reads.

PDHMM und spaltenweise Erkennung

Varianten-Caller wie GATK Haplotype Caller und DRAGEN reassemblieren Reads mithilfe des De-Bruijn-Graphen, um Haplotyp-Kandidaten zu bestimmen und potenzielle Varianten-Stellen zu ermitteln. Für Genomregionen mit Tandem-wiederholungen, strukturellen Varianten oder Clustern von Sequenzierungsfehlern kann eine geringere Sensitivität verursacht werden, wenn mithilfe der Graph-Assemblierung keine vollständige Liste der Haplotyp-Kandidaten und Variantenstellen ausgegeben werden kann.

Die spaltenweise Ereigniserkennung ergänzt den De-Bruijn-Graphen durch das Scannen jeder Spalte einer aktiven Region nach potenziellen Variantenstellen (SNPs und Indels) und durch das Vervollständigen der Liste mit den Haplotyp-Kandidaten. Durch dieses Verfahren lässt sich die Sensitivität in Regionen gewährleisten, in denen der Graph nicht angewendet werden kann.

Auswirkungen von FRD/BQD auf QUAL/GQ/QD und manuelle Filterung nach Erstellung der VCF-Datei

In den DRAGEN v3-Varianten-Caller wurden zwei Algorithmen implementiert, die korrelierte Fehler in den einzelnen Reads eines bestimmten Pileups modellieren, FRD (Foreign Read Detection, Erkennung fremder Reads) zur Erkennung falsch zugeordneter Reads sowie ein BQD-Algorithmus (Base Quality Drop off, Verringerung der Base-Qualität) zur Erkennung korrelierter Base-Call-Fehler. Neben der Verbesserung der Spezifität und Sensitivität bieten diese beiden Algorithmen weitere Vorteile:

Die Konfidenzwerte (QUAL, GQ, QD) liegen in einem realistischen Bereich der Phred-Skala.

Herkömmliche Varianten-Caller geben in der Regel überhöhte QUAL-Werte in den Tausenderbereichen der Phred-Skala aus, die ohne jede statistische Bedeutung sind. Mit modellierten korrelierten Fehlern aus dem Varianten-Caller können diese Werte in einen statistisch realistischen und aussagekräftigen Bereich zurückgeführt werden.

Die Abhängigkeit von Filterregeln nach Erstellen der VCF-Datei wird erheblich verringert.

Da herkömmliche Varianten-Caller nicht zwischen korrelierten Fehlern und tatsächlichen Abweichungen unterscheiden können, war es erforderlich, nach der Erstellung der VCF-Dateien mit Regeln für die manuelle Filterung die zahlreichen FP-Calls herauszufiltern. Verschiedene VCF-Annotationen (z. B. QD, MQ, FS, MQRankSum) wurden mit Ad-hoc-Schwellenwerten verglichen, um Calls als FP zu markieren. Außerdem kann mit diesen Annotationen und einem Referenzsatz auch ein KI-Algorithmus trainiert werden, um falsch positive Werte herauszufiltern (z. B. VQSR).

In DRAGEN v3 wurden die Algorithmen des Varianten-Callers grundlegend verbessert. So konnte die Abhängigkeit von Filtern nach Erstellung der VCF-Datei erheblich verringert werden. Die Standardregel für die manuelle Filterung in DRAGEN v3 verwendet QUAL mit einem Schwellenwert, der mit dem besten Fmeas-Wert (optimales Verhältnis zwischen Sensitivität und Spezifität) korrespondiert.

Umfangreiche Methoden

Eingabedatensätze

Für die Darstellung verschiedener Methoden der Bibliotheksvorbereitung wurden drei Datensätze sowohl mit als auch ohne ohne PCR (TruSeq DNA Nano, TruSeq DNA PCR-Free und Nextera DNA Flex) ausgewählt. Jeder Datensatz wurde mit DNA der NA12878-Probe erstellt. Nach der Vorbereitung der DNA-Bibliotheken entsprechend den jeweiligen Referenzhandbüchern[5-7] wurden die Ergebnisbibliotheken mit dem NovaSeq™ 6000-System in Paired-End-Läufen mit 2 x 150 bp sequenziert. Für die Normalisierung der Read-Anzahl wurde jeder Datensatz mit dem FASTQ Toolkit in BaseSpace Sequence Hub auf 30-fache Abdeckung heruntergerechnet. Alle drei Datensätze stehen in BaseSpace Sequence Hub zur Verfügung, sodass unabhängige Ergebnisprüfungen ausgeführt werden können.

Menschliches Referenzgenom

Als Referenzgenom wurde hs37d5 in der DRAGEN BaseSpace-App verwendet. In der lokalen Analyse der einzelnen untersuchten Anwendungen wurde das entsprechende Referenzgenom verwendet. In dieser Referenz sind Scheinwerte enthalten.[8]

Anwendungen für die Sekundäranalyse

Es werden drei Anwendungen für die Sekundäranalyse verglichen. Die erste Anwendung ist DRAGEN v2 für den gesamten Prozess (Mapping, Alignment und Varianten-Calling). Bei der zweiten Anwendung handelt es sich um DRAGEN v3 für den gesamten Prozess. In der dritten Anwendung werden Mapping und Alignment mit BWA-MEM und das Varianten-Calling mit GATK4-HC durchgeführt.

Um den Vergleich aussagekräftig zu gestalten, wurden für alle drei Anwendungen die gleichen Regeln für die manuelle Filterung angewendet: Ein GQ-Schwellenwert wurde auf die VCF-Dateien vor deren Filterung angewendet. Der Schwellenwert wurde so festgelegt, dass er für jede Anwendung in der Nähe des optimalen Fmeas-Punkts liegt (Tabelle 3).

Tabelle 3: QC-Schwellenwerte für optimale Fmeas

GQ für optimale Fmeas	SNP	Indel
DRAGEN v3.2.8	9	9
DRAGEN v2.5	2	8
BWA+GATK	1	2

DRAGEN wurde auf einem lokalen Server und in der Cloud in BaseSpace Sequence Hub ausgeführt. Die Berechnungsdauer in der Cloud war unwesentlich länger, doch die Ergebnisse des Varianten-Callings wiesen keine Abweichungen auf. Die BWA+GATK-Anwendung wurde auf dem gleichen lokalen Server wie DRAGEN mit installiertem BCBIO-Framework ausgeführt.[9] BCBIO führt BWA+GATK entsprechend den Best-Practice-Richtlinien von GATK aus. Außerdem wird mit zusätzlichen Optimierungen der Parallelität die Laufzeit beschleunigt. Für die Analyse der Cloudversion wurde die BWA+GATK-Anwendung auf Terra ausgeführt.

DRAGEN 3.3.0

Version der DRAGEN-App:
DRAGEN Germline Pipeline 3.2.8
DRAGEN-Host-Softwareversion 05.011.281.3.2.8

BWA-Mem (0.7.17) + GATK4 (4.0.2)

Tabelle 4: Parameter aus der Konfigurationsdatei der BCBIO-Algorithmen

Parameter	Wert
align_split_size	5000000
aligner	BWA
coverage_depth	Hoch
coverage_interval	Regional
mark_duplicates	True
merge_bamprep	False
platform	Illumina
quality_format	Standard
realign	False
recalibrate	False
tools_off	Vqsr
variantcaller	GATK-haplotype

Analyse: variant2
Ressourcen: gatk-haplotype

BWA+GATK auf Terra

Als Eingaben für die Ausführung von GATK auf Terra wurden analysebereite BAM-Dateien von BWA-Mem (aus BCBIO-Läufen) verwendet. Im Allgemeinen wurde der Workflow GATK4-germline-snps-indels (https://github.com/gatk-workflows/gatk4-germline-snps-indels) befolgt. Dabei wurden spezifische Parameter auf die Parameter der BCBIO-Läufe abgestimmt. Alle Läufe wurden mit einem kostenlosen Testkonto auf Terra ausgeführt.

Die genaue WDL-Methode steht in BaseSpace Sequence Hub zur Verfügung.

Konfigurationen der WDL-Methode:
GATK-Docker-Bild: broadinstitute/gatk:4.0.2.0
GITC-Docker: broadinstitute/genomes-in-the-cloud:2.3.1-1500064817

Referenz-FASTA: hs37d5 (wie für die anderen Anwendungen)

In dieser Anwendung wurden nur VCF-Dateien mit Rohdaten erstellt. Die Filterung nach der Dateierstellung wurde lokal ausgeführt. VCF-Dateien mit Rohdaten stehen in BaseSpace Sequence Hub zur Verfügung.

BaseSpace (Januar 2019) – Legen Sie die Spezifikation der verwendeten AWS F1-Instanz fest (AWS F1 4-fach vergrößert). Version der BaseSpace Sequence Hub-App: 3.2.8

Tabelle 5: Lokaler Server (CentOS 7 x86_64, Supermicro 1029)

Teil	Vollständige Modellbezeichnung	Hinweise
Chassis	SYS-1029GQ-TNRT	1 Rack-Einheit
CPU	2 x Intel(R) Xeon(R) Gold 6126 CPU mit 2,60 GHz	24 Kerne, 48 Threads
RAM	384 GB	DDR4, 2666 MHz
Staging	Intel SSDPE2KE020T7	2 TB NVME

Benchmark-Referenzdatensatz (NIST)

Für das Benchmarking von Varianten-Calls sind ein spezifisches Referenzgenom und ein zugeordneter Satz an Calls, der die richtigen Ergebnisse für dieses Genom darstellt, erforderlich. Diese Call-Sätze können als Referenzsätze verwendet werden, um falsch positive und falsch negative Werte zuverlässig zu ermitteln. Der in dieser Untersuchung verwendete Referenzsatz basierte auf Referenz-Calls, die mit der gleichen DNA (NA12878) und entsprechend den Standards des National Institute of Standards and Technology (NIST) erstellt wurden. Das Genome in a Bottle Consortium (GIAB) ist eine Vereinigung öffentlicher und privater akademischer Organisationen, die von NIST ins Leben gerufen wurde. Von GIAB wurde ein Benchmarksatz mit Calls kleiner Varianten und Referenzcalls für das Pilotgenom NA12878 veröffentlicht, der einen Genotyp mit hoher Zuverlässigkeit für ca. 90 % von GRCh37 und GRCh38 angibt.

Richtig positive Ergebnisse (TPs) sind Varianten-Calls, die mit Referenz-Calls aus dem NIST-Referenzsatz übereinstimmen. Falsch positive Ergebnisse (FPs) sind Varianten-Calls, die nicht im Referenzsatz vorhanden sind. Bei falsch negativen Ergebnissen (FNs) handelt es sich um Varianten im Referenzsatz, die nicht in der QUERY VCF-Datei enthalten sind.

Mithilfe des Variant Calling Assessment Tools (VCAT) wurde jede QUERY VCF-Datei mit dem NIST-Referenzsatz v3.3.2 verglichen. Dieses Tool führt hap.py mithilfe des Evaluierungsmoduls RTG vcfeval aus. TPs, FPs und FNs wurden mit folgenden Ausgabedateien von hap.py bestimmt: *roc.Locations.INDEL.csv und *roc.Locations.SNP.csv von TRUTH TP, QUERY.TP, QUERY.FP und TRUTH.FN.

Die entsprechende Stringenz für die Berechnung von TP, FP und FN lautet „genotype match“ (Übereinstimmung des Genotyps) (cf. [1]) – nur Stellen mit übereinstimmenden Allelen und Genotypen werden als TP gewertet. Genotypfehler und nicht übereinstimmende Allele werden als FPs und FNs gewertet.

Tabelle 6: Definitionen und Berechnungen für Metriken mit Bezug zu Präzision und Recall

Metrik	Gebräuchliche Bezeichnung	Definition	Formel
TRUTH.TP	Richtig positive Ergebnisse (Referenz)	Anzahl der Referenz-Calls mit Abfrage-Call, der mit dem Referenz-Call und dessen Genotyp übereinstimmt	
QUERY.TP	Richtig positive Ergebnisse (Abfrage)	Anzahl der Abfrage-Calls mit Referenz-Call, der mit dem Abfrage-Call und dessen Genotyp übereinstimmt	
TRUTH.FN	Falsch negative Ergebnisse	Anzahl der Referenz-Calls ohne Abfrage-Call, der mit dem Referenz-Call und dessen Genotyp übereinstimmt	
QUERY.FP	Falsch positive Ergebnisse	Anzahl der Abfrage-Calls ohne Referenz-Call, der mit dem Abfrage-Call und dessen Genotyp übereinstimmt	
METRIC.Recall	Recall, Sensitivität	Anteil der Referenz-Calls, die mit einem Abfrage-Allel und Genotypaufruf innerhalb der zuverlässigen Regionen übereinstimmen	TRUTH.TP / (TRUTH.TP + TRUTH.FN)
METRIC.Precision	Präzision, positiver prädiktiver Wert	Anteil der Abfrage-Calls, die mit einem Referenz-Allel und Genotypaufruf innerhalb der zuverlässigen Regionen übereinstimmen	QUERY.TP / (QUERY.TP + QUERY.FP)

Benchmarking – Evaluierungsmetriken

Für den Vergleich der Geschwindigkeit wurde die Gesamtlaufzeit in Sekunden von FASTQ bis VCF aus den Analyseprotokolldateien und/oder den Analysezeiten aus den Berichten verwendet.

Für den Vergleich der Genauigkeit der verschiedenen Anwendungen wurden empfohlene Standards für Leistungsmetriken (Tabelle 6) verwendet.[1] Mit Präzision wird die analytische Spezifität oder die Fähigkeit, das Nichtvorhandensein von Varianten korrekt zu ermitteln, bzw. das Nichtvorhandensein von falsch positiven Ergebnissen angegeben. Mit Recall wird die analytische Sensitivität oder die Möglichkeit, das Vorhandensein bekannter Varianten zu ermitteln, bzw. das Nichtvorhandensein von falsch negativen Ergebnissen angegeben.

Die Definitionen und Berechnungen für die Metriken hinsichtlich der Präzision und der Anzahl der Recalls entsprechen der Referenz.

Quellen

1. Krusche P, Trigg L, Boutros PC, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol*. 2019;37(5):555-560.
2. GATK Best Practices. software.broadinstitute.org/gatk/best-practices/. Aufgerufen am 9. Mai 2019.
3. The BaseSpace project. basespace.illumina.com/s/3ExEZMlH8Lkq. Aufgerufen am 15. Mai 2019.
4. FireCloud Powered by Terra. firecloud.terra.bio/. Aufgerufen am 15. Mai 2019.
5. Illumina (2017) TruSeq DNA PCR-Free Reference Guide. Aufgerufen am 6. März 2019.
6. Illumina (2017) TruSeq DNA Nano Reference Guide. Aufgerufen am 6. März 2019.
7. Illumina (2018) Nextera DNA Flex Library Prep Reference Guide. Aufgerufen am 6. März 2019.
8. hs37d5 Reference Genome. ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/. Aufgerufen am 9. Mai 2019.
9. Bcbio-nextgen. Docs. bcbio-nextgen.readthedocs.io/en/latest/. Aufgerufen am 9. Mai 2019.

Illumina, Inc. • Tel. USA (gebührenfrei) 1.800.809.4566 • Tel. außerhalb Nordamerikas +1.858.202.4566 • techsupport@illumina.com • www.illumina.com