medgen 2014 · 26:239–245
DOI 10.1007/s11825-014-0448-6
Online publiziert: 12. Juni 2014

R. Weißmann[1] · C. Gilissen[2]
[1] Institut für Humangenetik, Universitätsmedizin Greifswald, und Interfakultäres Institut für Genetik und Funktionelle Genomforschung, Universität Greifswald
[2] Department of Human Genetics, Radboud University Medical Center, Nijmegen

NGS Datenanalyse und Qualitätskontrolle

Hintergrund

Eine Vielzahl an Technologien zur DNA-Sequenzierung wird als Next Generation Sequencing (NGS) oder „massive parallel sequencing" bezeichnet. Alle diese Technologien produzieren sehr große Mengen digitaler Daten auf der Grundlage biologischen Materials. Dieser Artikel soll v. a. Lesern mit geringeren bioinformatischen Vorkenntnissen grundlegende Hintergrundinformationen zu Art und Struktur der anfallenden Daten vermitteln. Darüber hinaus werden der allgemeine Ablauf (**Abb. 1**) und wichtige Prinzipien der NGS-Datenanalyse vorgestellt. Es wird zudem ein erster Einblick in die Qualitätskontrolle beim NGS gegeben. Für detailliertere Ausführungen und weitere Programme wird der Artikel von Guo et al. empfohlen [5].

Primärdaten

Die Daten, die direkt im Sequenziergerät entstehen, werden auch als Primärdaten bezeichnet. Welcher Art diese Daten sind, ist vom jeweiligen Gerät abhängig. Die verschiedenen derzeit gängigen Geräteplattformen und die jeweiligen Besonderheiten der unterschiedlichen Technologien werden im Artikel „Einführung in die Grundlagen der Hochdurchsatzsequenzierung" von K. Neveling und A. Hoischen in diesem Themenheft vorgestellt.

Bei fast allen momentan auf dem Markt vorhandenen Geräten werden die Sequenzen mithilfe von Fluoreszenzmarkern erkannt. Die Signale dieser Marker werden mithilfe von Kameras aufgenommen und als Bilddateien abgespeichert. Die Primärdaten sind hier also digitale Bilder. Um aus den Leuchtsignalen auf die sequenzierten Nukleotide zu schließen, gibt es Computersoftware, die diesen als „base calling" bezeichneten Prozess durchführt. Eine Ausnahme stellt hier die Ion-Torrent™-Technologie dar. Diese basiert auf der Messung der pH-Änderung beim Nukleotideinbau während der Sequenzierreaktion. Die Primärdaten sind in diesem Fall keine Bilder, sondern Messkurven von pH-Werten, die dann aber ebenfalls in Nukleotidinformation übersetzt werden.

„Base calling"

Die Bilder, die beim „base calling" analysiert werden, haben Ähnlichkeit mit einem Sternenhimmel (**Abb. 2**). Vor einem dunklen Hintergrund befinden sich verstreute Punkte: die Leuchtsignale von Fluoreszenzmarkern. Jedes Signal stammt dabei nicht von einem einzelnen Molekül, sondern von einem Cluster aus Molekülen. Diese Cluster werden von einer Software, dem „base caller", auseinandergehalten und einem bestimmten Ort auf dem Reaktionsträger („flow cell" oder „flow chip") zugeordnet. Allerdings befinden sich manche Cluster zu dicht beieinander, um auseinandergehalten werden zu können, andere sind polyklonal und produzieren deshalb zweideutige Signale. In beiden Fällen können die

Begriffe und Abkürzungen

„Base caller"	Computerprogramm, das auf Grundlage der Primärdaten eine Nukleotidsequenz (Read) generiert.
BAM	*„B*inary sequence *a*lignment/*m*ap". Herstellerübergreifender Quasi-Standard für NGS-Reads.
dbSNP	Datenbank, in der bekannte „single nucleotide polymorphisms" (SNP) gesammelt werden.
„Flow cell/flow chip"	Glasträger, an den beim Sequenzieren DNA-Fragmente angeheftet sind.
GRC	Genome Reference Consortium.
Illumina®	Hersteller von NGS-Maschinen.
Ion Torrent™	NGS-Technik, bei der keine Bilder gemacht, sondern pH-Werte auf einem Siliziumchip gemessen werden.
Monoklonal	Cluster auf der „flow cell", die aus einem einzigen DNA-Fragment entstanden sind. Gegensatz: polyklonal
Polyklonal	Cluster auf der „flow cell", die aus einer Mischung von 2 oder mehreren DNA-Fragmenten entstanden sind und nicht weiter ausgewertet werden können.
Pyrosequenzierung	NGS-Methode der Fa. Roche (454™-Technologie).
Read	Kurze Nukleotidsequenz (26–1000 nt), die beim NGS produziert wird.
„Read mapper"	Computerprogramm, das Reads einer Position im Referenzgenom zuordnet.
SOLiD™	Sequencing by *o*ligonucleotide *l*igation and *d*etection. NGS-Technik der Fa. ABI (Life Technologies).

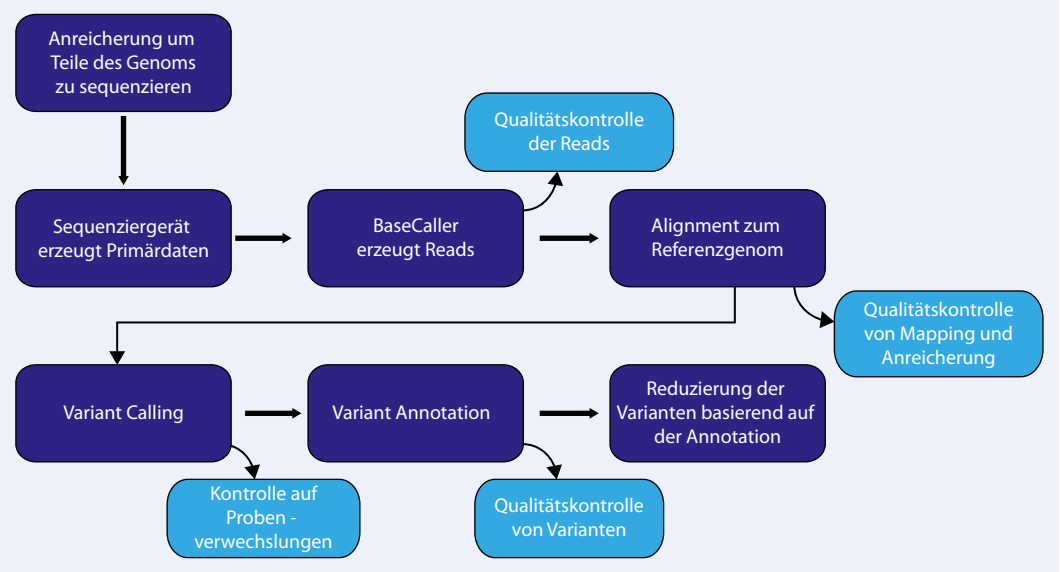

Abb. 1 ◀ Arbeitsablauf beim Next Generation Sequencing. Der Anreicherungsschritt entfällt, wenn das ganze Genom sequenziert wird

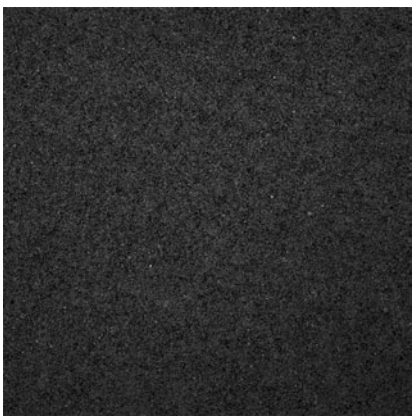

Abb. 2 ▲ Ausschnitt einer „flow cell". Jeder weiße Punkt ist das Fluoreszenzsignal eines Clusters von DNA-Fragmenten. Jedes Cluster führt beim Sequenzieren zu einem Read (Single-end-Sequenzierung) oder zu 2 Reads (Paired-end-Sequenzierung)

Signale nicht verwendet werden. Darüber hinaus gibt es Unterschiede in der Signalstärke. Alle diese Informationen fasst der „base caller" zu 2 Informationen zusammen: dem detektierten Nukleotid und dem zugehörigen Qualitätswert.

Qualitätswerte der Reads

Die Nukleotidinformationen, die aus den sequenzierten DNA-Fragmenten gewonnen werden, bezeichnet man als Einzelsequenzen bzw. Reads. Die Sequenzen der Reads bestehen neben Nukleotiden (A, C, G, T) auch aus „no calls" (N), falls an einer Stelle in der Sequenz ein Nukleotid nicht eindeutig erkannt werden konnte. Zu jedem Nukleotid gibt es außerdem einen Qualitätswert, der die Wahrscheinlichkeit angibt mit der das Nukleotid falsch erkannt wurde. Diese Werte werden in der sog. Phred-Skala angegeben. Gute Werte liegen bei ≥30, wobei 30 eine Wahrscheinlichkeit von 0,001 für ein falsch identifiziertes Nukleotid bedeutet [3]. Es gibt allerdings bisher in der Literatur keinen Konsens über einen Schwellenwert für akzeptable Qualität.

Read-Mengen

Wenn Hersteller für ihre Geräte werben, wird oft der Durchsatz in Mb angegeben. Dieser Wert ergibt sich aus der Anzahl und der Länge der Reads. Eine Mio. Reads einer Länge von 36 Nukleotiden entsprechen z. B. 36 Mb. Bei den Geräten der Fa. Illumina® und bei der SOLiD™-Plattform (Fa. Life Technologies) sind alle Reads eines Sequenzierlaufs jeweils gleich lang. Bei Geräten mit Ion-Torrent™- (Ion PGM™ und Ion Proton™, Fa. Life Technologies) oder Pyrosequenzierungstechnologie (Genome Analyzer, Fa. Roche) variieren die Längen der Reads eines Sequenzierlaufs. Die Anzahl der Reads befindet sich bei allen Anwendungen üblicherweise im Bereich mehrerer Millionen. Beim Single-end-Sequenzieren entsteht aus jedem Fragment ein Read. Beim Paired-end-Sequenzieren bekommt man pro Fragment 2 Reads. Das Fragment wird hier von beiden Enden her ansequenziert. Zur weiteren Auswertung der Reads wird i. d. R. ein Referenzgenom benutzt.

„Alignment" zum Referenzgenom

Im Gegensatz zur Sanger-Sequenzierung, bei der man Primer entwirft, die die zu sequenzierende Region bestimmen, weiß man beim NGS zunächst nicht, zu welcher Sequenz die einzelnen Reads im Genom korrespondieren. Es werden daher Computerprogramme benötigt, die die Reads einer Position im Referenzgenom zuordnen. Diese Programme werden auch als „read mapper" oder „short read aligner" bezeichnet. Sie finden die Positionen im Referenzgenom, die eine identische oder sehr ähnliche Reihenfolge an Nukleotiden aufweisen, wie die Reads. Es ist sehr wichtig, dass „read mapper" gewisse Abweichungen der Reads vom Referenzgenom zulassen, denn ansonsten könnte man keine Varianten finden. Je mehr Abweichungen zugelassen werden, desto wahrscheinlicher kann ein Read nicht mehr eindeutig einer genomischen Position zugeordnet werden. Die erlaubte Anzahl an Abweichungen kann i. d. R. dem Programm über entsprechende Parameter vorgegeben werden. Die Bestimmung der Read-Position im Referenzgenom wird auch als Read-Mapping bezeichnet. Die dabei erzeugten Ergebnisdateien werden konventionell inzwischen im BAM-Dateiformat ausgegeben. Die Auswahl eines geeigneten „read

mappers“ richtet sich nach verschiedenen Kriterien, wie Schnelligkeit, Genauigkeit, Hardwareanforderungen, Kompatibilität mit Sequenzformaten und anderen Programmen. Ein Vergleich findet sich z. B. bei Hatem et al. [6].

Referenzsequenz

Das Referenzgenom des Menschen entstand ursprünglich aus dem Human Genome Project. Die dort erstmals ermittelte Sequenz wurde aber im Anschluss daran immer genauer, so dass es inzwischen unterschiedliche Versionen der menschlichen Referenzsequenz gibt. Die momentan aktuelle humane Referenzsequenz ist GRCh38 (http://genomeref.blogspot.de, Zugegriffen 19. Januar 2014). Da Analysen schlechter vergleichbar sind, wenn unterschiedliche Versionen des Referenzgenoms benutzt wurden, sollte man sich vor der Analyse bewusst für eine bestimmte Version entscheiden. Weiterhin ist das Referenzgenom ohne Annotationen kaum brauchbar. Annotationen beschreiben u. a. die Lage von Chromosomenbanden, Genen, Varianten, konservierten Bereichen, repetitiven Elementen und Zentromeren. Annotationen sind immer nur für eine konkrete Version der Referenzsequenz verwendbar. Wenn eine neue Version des Referenzgenoms verfügbar wird, dann dauert es einige Zeit, bis Annotationen angepasst werden. Für manche Annotationen geschieht das auch gar nicht. Daher ist es nicht immer die beste Wahl, die aktuellste Version des Referenzgenoms zu verwenden.

„Mappability“

Nicht jede Teilsequenz kommt nur ein einziges Mal im Genom vor. Der Grad der Einzigartigkeit wird auch als „mappability“ bezeichnet. Er kann Werte zwischen 0 und 1 annehmen, wobei ein Wert von 1 bedeutet, dass die Teilsequenz nur einmal vorkommt. Ein Read, der einer genomischen Position mit hoher „mappability“ zugeordnet wird, liegt dort höchstwahrscheinlich richtig. Eine niedrige „mappability“ lässt vermuten, dass der Read möglicherweise auch anderen genomischen Positionen zugeordnet werden kann.

Anreicherung

Oft wird bei der Suche nach Mutationen nicht das gesamte Genom sequenziert. Günstiger und schneller ist es, sich auf den Bereich zu beschränken, der proteincodierende Gene enthält. Man spart einerseits Geld, weil weniger sequenziert werden muss, und bewegt sich andererseits auf viel sichererem Terrain bei der Beurteilung der gefundenen Mutationen in Bezug auf mögliche Krankheitsursachen: Eine Mutation, die in einem proteincodierenden Gen ein Stop-Codon produziert, ist einfacher zu erklären als eine Mutation in wenig bis gar nicht charakterisierten intergenischen Bereichen. Außerdem wird die Zahl an möglichen Zusatzbefunden drastisch reduziert. Es gibt verschiedene Anreicherungsverfahren, die von kommerziellen Anbietern angeboten werden. Die fast allen Verfahren zugrundeliegenden Prinzipien werden in diesem Themenheft im Beitrag von K. Neveling und A. Hoischen kurz vorgestellt.

Sequenziertiefe

Die Anzahl der Reads, die einer gemeinsame genomische Position zugeordnet wurden, wird als Sequenziertiefe („coverage“) bezeichnet. Je größer die Sequenziertiefe ist, desto sicherer kann man heterozygote von homozygoten Varianten unterscheiden und umso besser kann man Sequenzierfehler und echte Varianten auseinanderhalten. Wenn man sich bei der Sequenzierung auf einen kleineren Bereich des Genoms beschränkt, kann man bei gleicher Menge an erzeugten Reads eine größere Sequenziertiefe erreichen.

„Variant calling“

Herauszufinden, wo und wie sich die sequenzierte DNA vom Referenzgenom unterscheidet, ist das Ziel des „variant calling“. Hierbei muss zwischen Veränderungen, die in der sequenzierten DNA vorliegen und Veränderungen, die nur aufgrund von Sequenzierfehlern oder falsch zugeordneten Reads auftauchen, unterschieden werden. Je größer die Sequenziertiefe ist, desto zuverlässiger gelingt das „variant calling“ und desto höher ist der Qualitätswert, der für die erkannte DNA-Variante ermittelt wird.

Für das „ variant calling“ stehen mehrere Programme zur Auswahl, die über individuelle Vor- und Nachteile verfügen [14]. Genome Analysis Toolkit (GATK, [12]) wurde beispielsweise im Rahmen des 1000-Genome-Projekts entwickelt und wird oft eingesetzt. Weitere Beispiele sind SAMTools [8], glfTools (http://www.sph.umich.edu/csg/abecasis/glfTools/. Zugegriffen: 24. Mai 2014) und Atlas2 [1]. Ein Vergleich dieser Werkzeuge findet sich bei [11]. Um vererbte Varianten zuverlässiger erkennen zu können, gibt es die Möglichkeit, die Varianten von mehreren Proben gleichzeitig zu bestimmen.

medgen 2014 · 26:239–245
DOI 10.1007/s11825-014-0448-6

R. Weißmann · C. Gilissen

NGS Datenanalyse und Qualitätskontrolle

Zusammenfassung
Next Generation Sequencing (NGS) wird immer häufiger in der Humangenetik eingesetzt. Die Analyse der anfallenden Datenmengen birgt allerdings andere und größere Herausforderungen als bisher eingesetzte Verfahren. In diesem Artikel werden einige Grundlagen, die dem Verständnis der anfallenden Daten und Analyseschritte beim NGS dienen sollen, beschrieben. Ein besonderer Schwerpunkt ist dabei die Qualitätskontrolle.

Schlüsselwörter
Datenanalyse · Qualitätskontrolle · Datenformate · Bioinformatik · DNA

NGS data analysis and quality assessment

Abstract
Next generation DNA sequencing (NGS) is rapidly becoming a pervasive technique within the human genetics community. The analysis of NGS data is however much more challenging than with previous genetic and genomics techniques. In this article, the basic data formats and analysis steps that are involved in any NGS DNA resequencing experiment are described. Special emphasis is placed on methods for quality control.

Keywords
Data analysis · Quality control · Data formats · Bioinformatics · DNA

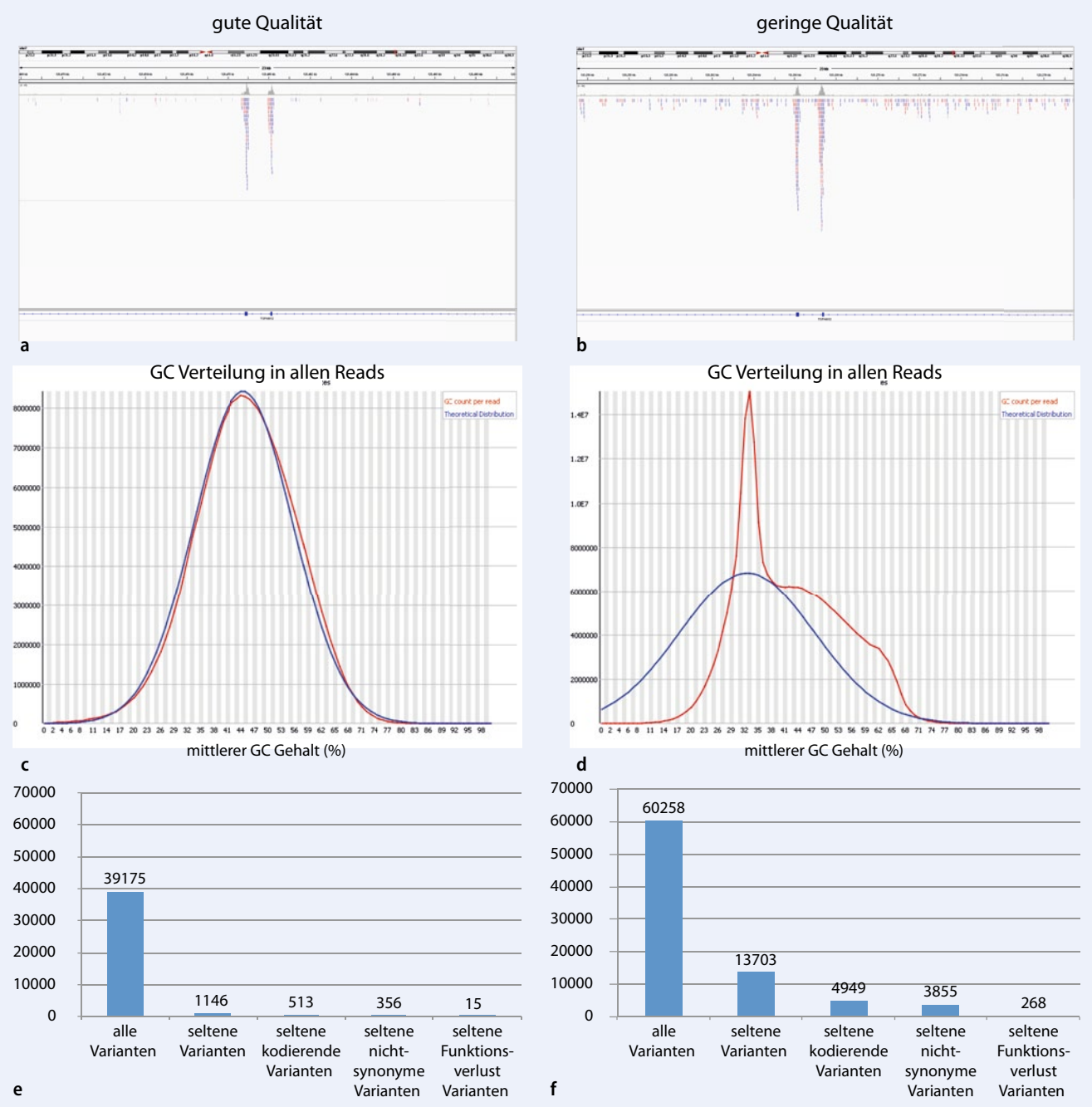

Abb. 3 ▲ Beispiele für Qualitätsmerkmale die auf geringe Probenqualität hinweisen. **a**, **b** Mittels Integrative Genome Viewer (IGV) sind 2 angereicherte Exons dargestellt. **a** Mehr als 90% der Reads liegen in den Exons. Dies deutet auf eine gute Anreicherung hin. **b** Viele Reads liegen außerhalb der angereicherten Exons. Dies deutet auf eine Anreicherung mit geringer Qualität hin. **c**, **d** Mit FastQC erstellte Darstellungen der Verteilung des GC-Gehalts aller Reads. **c** Die beobachteten Werte (*rote Kurve*) stimmen gut mit den erwarteten Werten (*blaue Kurve*) überein. **d** Die beobachteten Werte (*rote Kurve*) weichen stark von den erwarteten Werten (*blaue Kurve*) ab – ein Hinweis auf ein mögliches Problem mit dieser Probe. **e**, **f** Säulendiagramme zur Anzahl der Varianten. Jede Säule steht für die Anzahl der Varianten (Substitutionen und Indels) einer Kategorie. **e** Säulendiagramm zu den Sequenzierergebnissen eines zufällig ausgewählten Exoms mit guter Qualität. **f** Säulendiagramm für eine Probe, bei der es Probleme mit der Sequenzierchemie gab. Die Anteile der seltenen codierenden Varianten und der seltenen Funktionsverlustvarianten sind hier im Vergleich zu den in **e** abgebildeten Werte viel höher

Qualitätskontrolle

Das gezielte Sequenzieren ausgewählter genomischer Bereiche (z. B. „gene panels“ oder „whole exome sequencing“) ist ein kompliziertes Verfahren, das viele verschiedene Schritte beinhaltet, von denen jeder eine Fehlerquelle sein kann. Hier werden nun einige Methoden für die Qualitätskontrolle bei den verschiedenen Schritten der bioinformatischen Analyse von der Rohsequenz bis zur Annotation von DNA-Varianten diskutiert.

Rohsequenz

Obwohl die meisten Sequenziergeräte bereits Qualitätskontrollen durchführen, bedeutet dies nicht, dass die produzierten Reads automatisch von guter Qualität sind. Daher verwendet man zur weitergehenden Überprüfung der Read-Qualität geeignete Softwarelösungen, wie z. B. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Zugegriffen: 24. Mai 2014). Dieses Programm bewertet die Reads in ihrer Gesamtheit. Es bildet dabei die Qualitätswerte der Nukleotide und deren Verteilung über die Länge der Reads ab, wobei die Qualitätswerte i. d. R. mit zunehmender Länge der Reads abnehmen. Außerdem werden überrepräsentierte (Teil-)Sequenzen („k-mer“) erkannt und die Verteilung der Read-Längen erfasst. Die Länge der erzeugten Reads bei den Sequenziergeräten mit Pyrosequenzierung oder Ion-Torrent™-Technologie ist nicht konstant, sondern variiert von Read zu Read. Bei diesen Geräten können die Sequenzlängen und deren Verteilung als Qualitätsindikator für den Sequenzierlauf verwendet werden. Im Allgemeinen werden längere Reads bevorzugt, da dies verschiedene Analysen vereinfacht, wie z. B. das Erkennen von Haplotypen. Aber auch Read-Mapping und „variant calling“ werden durch größere Read-Längen erleichtert. Eine solche Fehleranalyse kann mit FastQC sowohl vor als auch nach dem Read-Mapping durchgeführt werden.

Neben der Qualitätskontrolle für einzelne Proben kann FastQC auch zur Durchführung von Trendanalysen verwendet werden. Damit kann erkannt werden, ob bzw. inwiefern Änderungen bei den Arbeitsanweisungen oder den verwendeten Reagenzien einen Einfluss auf die Sequenzierungsqualität haben.

Mapping und Anreicherung

Das Read-Mapping ist bei der Analyse von NGS-Daten ein grundlegender Schritt, denn die derzeit von den verschiedenen NGS-Plattformen erzeugten Reads sind zu kurz und das menschliche Genom ist zu groß, um ein „de novo assembly“ durchzuführen. Schon das Read-Mapping selbst bietet dabei die Möglichkeit einer Überprüfung der Güte von Sequenzierdaten und kann daher zur Qualitätssicherung mit eingesetzt werden. Hierbei ist zunächst der Anteil derjenigen Reads, denen tatsächlich eine Position auf dem Referenzgenom zugeordnet werden konnte, ein wichtiges Kriterium. Abhängig von der Menge der Reads, die wegen schlechter Qualität herausgefiltert wurden, und ihrer Länge sollten etwa 80–90% aller Reads mindestens einer Position im Referenzgenom zugeordnet werden. Kurze Reads (≤50 Nukleotide) werden häufiger mehreren Positionen im Referenzgenom zugeordnet, während andererseits Paired-end-Reads häufiger nur einer einzigen Position zugeordnet werden, da man zwischen zusammengehörigen Reads, die vom selben sequenzierten Fragment der Probe stammen, nur einen kurzen Abstand im Referenzgenom erwartet. Weil der Anteil der im Referenzgenom zugeordneten Reads stark von der Sequenziertechnik und von der Probenvorbereitung abhängt, ist es am hilfreichsten, wenn beides gleich bleibt und nur die sequenzierte Probe variiert.

Sobald man die zugeordneten Reads erhalten hat, ist es möglich, den Anteil der Basenpaare, die nicht der Referenzsequenz entsprechen, als Schätzgröße für die Sequenzierfehler zu benutzen, da die Anzahl der Sequenzierfehler viel größer ist als die Anzahl der echten Sequenzunterschiede. So bieten z. B. die Softwarepakete SAMtools [8] und Picard (http://picard.sourceforge.net. Zugegriffen: 24. Mai 2014) eine Reihe von Dienstprogrammen zur Berechnung dieser und anderer Statistiken für zugeordnete Sequenzdaten im BAM-Format. Allerdings sind diese Werkzeuge nur als Kommandozeilenver-

sion verfügbar, da die Dateien, mit denen sie operieren, i. d. R. zu groß sind, um sie in einer graphischen Benutzeroberfläche sinnvoll darzustellen.

Nach dem Read-Mapping ist es auch möglich, das verwendete Anreicherungsverfahren zu beurteilen, indem man die Anzahl der Reads/Basen, die innerhalb des angereicherten genomischen Bereichs zugeordnet wurden, mit denen vergleicht, die außerhalb liegen. Bei der Verwendung eines Anreicherungsverfahrens ist es dabei weniger wichtig, welche Read-Mengen für eine Probe produziert wurden. Wichtiger ist, wie viele Reads innerhalb des angereicherten Bereichs gefunden wurden („reads on target"). Bedingt durch das Prinzip der Anreicherung liegen manche Reads auch etwas außerhalb des angereicherten Bereichs. Man kann Reads daher in 3 Klassen einteilen: Reads *im* Bereich, Reads *am* Bereich und Reads *außerhalb* des Bereichs, wobei „am Bereich" typischerweise „nicht weiter als eine Read-Länge vom Bereich entfernt" bedeutet. Für eine gute Anreicherung ist zu erwarten, dass 80–90% der Reads im oder am angereicherten Bereich liegen. Für eine Anreicherung auf Grundlage von Amplikons erwartet man nahezu 100%. Reads, die außerhalb des Bereichs liegen, weisen nicht unbedingt auf ein Problem bei der Anreicherung hin, sondern können auch auf Duplikationen im Genom zurückgehen, die dazu führen, dass Reads von mehreren genomischen Regionen stammen oder einfach zum falschen Ort im Genom zugeordnet wurden. Daher ist es ratsam, das gesamte Genom und nicht nur den Anreicherungsbereich als Referenzsequenz zu verwenden, da letzterer den „read mapper" dazu zwingt, die Reads dort zu verorten, obwohl die Reads möglicherweise von ganz anderen genomischen Lozi stammen.

Ein weiterer Qualitätsparameter ist die durchschnittliche Sequenziertiefe im angereicherten Bereich und deren Schwankung. Regionen im Genom müssen eine ausreichende Sequenziertiefe aufweisen, damit jedes der beiden Allele mindestens einmal sequenziert wird, aber auch, damit genomische Varianten von Sequenzierfehlern unterschieden werden können. In der Literatur gibt es unterschiedliche Auffassungen über eine ausreichende durchschnittliche Sequenziertiefe. Dort liegen die Angaben zwischen 15-fach [9] und 40-fach [13]. Die durchschnittliche Sequenziertiefe ist jedoch kaum aussagekräftig, wenn die Varianz der zugrunde liegenden Read-Zahlen pro genomischer Position zu groß ist. Vorzugsweise sollten daher alle angereicherten Bereiche eine möglichst gleichförmige Sequenziertiefe haben. Es ist jedoch bekannt, dass besonders GC-reiche Bereiche schwer anzureichern sind. Daher weisen Bereiche mit extrem hohem (oder auch extrem niedrigem) GC-Gehalt weniger Sequenziertiefe auf oder können z. T. gar nicht sequenziert werden. Dies betrifft v. a. die ersten Exons von Genen, da diese oft einen hohen GC-Gehalt haben. Bei einigen Anreicherungsverfahren wird daher versucht, dieses Problem durch die Gestaltung der verwendeten Sonden zu umgehen. Sie werden so entworfen, dass sie nicht direkt auf dem Zielbereich liegen, sondern etwas daneben. Man hofft, dass dadurch die zugehörigen Reads um die Sonde herum gestreut auftreten und so das eigentliche Ziel mit abdecken. Aufgrund der geschilderten Variabilität der Sequenziertiefe ist es oft notwendig, eine wesentlich höhere als die minimale erforderliche Menge zu sequenzieren, um auch in den Problembereichen eine ausreichende Sequenziertiefe zu erreichen. Es ist zu beachten, dass die Sequenziertiefe meist als Durchschnittswert über den zu sequenzierenden Bereich ermittelt wird und für individuelle Positionen stark schwanken kann. Man muss also dafür sorgen, dass eine ausreichende Sequenziertiefe erreicht wird, damit jedes der beiden Allele häufig genug sequenziert wird und genomische Varianten von Sequenzierfehlern unterschieden werden können.

Im Allgemeinen ist es daher nicht nur von Interesse zu wissen, für welche Regionen eine für das „variant calling" ausreichende Sequenziertiefe erreicht wurde, sondern auch, die Regionen zu bestimmen, für die das nicht zutrifft. Es gibt bereits mehrere Werkzeuge zur Berechnung der Anzahl von zugeordneten Reads in einem genomischen Bereich, z. B. das bereits erwähnten Programm SAMtools [8], aber auch BEDTools [16] und das GATK-Softwarepaket [12]. Eine sehr intuitive Art und Weise der Bewertung von Sequenz- und Anreicherungsqualität stellt die visuelle Inspektion der zugeordneten Reads dar, wofür verschiedene Werkzeuge existieren. Eines der hierzu am häufigsten verwendeten Programme ist der Integrative Genome Viewer (IGV; [17]).

Varianten

Nach dem Read-Mapping findet das „variant calling" statt. Die Anzahl der Varianten, die gefunden werden, hängt von der Größe des angereicherten Bereichs und von den gewählten Einstellungen für das „variant calling" ab. Für ein typisches Exom kann man am Ende mit zwischen 10.000 und 400.000 Varianten rechnen – abhängig von den Einstellungen. Eine Menge von rund 40.000 Varianten wird jedoch normalerweise erwartet. Ergebnisse, die davon zu sehr abweichen, sind ein Hinweis auf zu restriktive oder zu lockere Einstellungen. Jedoch ist die Anzahl der Varianten nur schlecht als Qualitätskontrolle geeignet, weil sie zu sehr von der Anzahl der Sequenzierfehler und der Sequenziertiefe abhängt. Ein besserer Indikator ist der Anteil der bekannten Varianten, der erkannt wird. Typischerweise wird überprüft, wie viele der bei der Sequenzierung beobachteten Varianten auch in der Datenbank dbSNP [18] vorkommen. Dieser Wert sollte in etwa 95% betragen [4]. Eine andere Möglichkeit der Charakterisierung der Probenqualität ist die Berechnung des Verhältnisses von Transitionen zu Transversionen bei „single nucleotide polymorphisms" (SNP) im Exom. Dieses sollte zwischen 3,0 und 3,5 liegen [10]. Eine fortschrittlichere Methode, um abweichende Proben zu identifizieren, besteht darin, die Genotypen der Proben mit einer Kohorte derselben Ethnie zu vergleichen [7].

Bemerkt man eine geringe Qualität bei einer Probe (**Abb. 3**), dann ist es nicht immer machbar oder erstrebenswert, sofort die Probe erneut zu sequenzieren. Manchmal ist es in solchen Fällen möglich, die Reads geringer Qualität einfach herauszufiltern. Auch das Kürzen der Reads kann sich lohnen, da die Enden der Reads i. d. R. die niedrigsten Qualitätswerte aufweisen. Eine pragmatische

Herangehensweise ist oft, die Ergebnisse trotz ihrer niedrigen Qualität zu nutzen und die Proben nur dann erneut zu sequenzieren, wenn keine interessanten Varianten erkannt werden konnten.

Annotation der Varianten

Alle Sequenzierungen resultieren schließlich in einer Liste der erkannten Varianten. Diese enthalten zu jeder Variante mindestens die Angaben zur genomischen Position, zur Abweichung von der Referenzsequenz und einen zugehörigen Qualitätswert. Je nach benutztem Programm sind zusätzliche Informationen vorhanden. Diese sind z. B. die Zygotie oder die Sequenzen der Reads, die zur Bestimmung der Variante geführt haben. Die Interpretation der Varianten erfordert eine bioinformatische Verknüpfung mit verschiedenen Datenquellen. Hierfür stehen viele verschiedene Programme zur Verfügung, die alle mindestens diejenigen Informationen bereitstellen, die für die Interpretation von Varianten als wesentlich angesehen werden. Besonders wichtig sind hierbei Angaben zu eventuell betroffenen Genen (z. B. aus RefSeq oder von Ensembl). Zusätzlich wird berechnet, ob die betrachtete Variante die Proteincodierung verändert und ob dies wiederum einen möglichen Einfluss auf die Funktion des Proteins hat. Es gibt darüber hinaus auch viele Onlinewerkzeuge, die genutzt werden können, um Varianten von NGS-Daten mit Annotationen zu versehen[15]. Das am häufigsten verwendete Werkzeug ist wahrscheinlich Annovar ([2, 19]; http://wannovar.usc.edu/. Zugegriffen: 24. Mai 2014). Es beschreibt für jede Variante, welchen Einfluss diese auf das betroffene Gen hat (z. B. Aminosäureveränderungen) und wie die Variante von verschiedenen Analysewerkzeugen (z. B. SIFT und PolyPhen) beurteilt wird. Ferner werden die Allelfrequenzen aus öffentlichen Datenbanken abgegriffen (z. B. 1000-Genome-Projekt und NHLBI-ESP-6500-Exomes-Projekt). Ein Protokoll zur Reduzierung von Varianten ist implementiert, um eine Teilmenge der potenziell schädlichen Varianten zu identifizieren.

Fazit für die Praxis

- **Generell ist bei der Planung und Durchführung von NGS-Projekten die Verfügbarkeit von bioinformatischer Expertise im Umgang mit NGS-Daten empfehlenswert.**
- **Von besonderer Bedeutung ist außerdem die Qualitätskontrolle. Sie sichert nicht nur die Qualität der Ergebnisse für einzelne Proben, sondern ist auch eine wichtige Voraussetzung für die Vergleichbarkeit von Auswertungen, die über einen längeren Zeitraum hinweg durchgeführt werden. Andernfalls wird z. B. die Metaanalyse solcher Daten erschwert bis unmöglich gemacht.**

Korrespondenzadresse

C. Gilissen
Department of Human Genetics, Radboud University Medical Center
Nijmegen
Die Niederlande
Christian.Gilissen@radboudumc.nl

Danksagung. Die Autoren danken Prof. Dr. Ute Felbor und Prof. Dr. Andreas W. Kuss für ihre Unterstützung bei der Erstellung und Korrektur des Manuskripts.

Einhaltung ethischer Richtlinien

Interessenkonflikt. R. Weißmann und C. Gilissen geben an, dass kein Interessenkonflikt besteht.

Dieser Beitrag beinhaltet keine Studien an Menschen oder Tieren.

Literatur

1. Challis D, Yu J, Evani US et al (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13:8
2. Chang X, Wang K (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet 49:433–436
3. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194
4. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20:490–497
5. Guo Y, Ye F, Sheng Q et al (2013) Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform (Epub ahead of print)
6. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics 14:184
7. Heinrich V, Kamphans T, Stange J et al (2013) Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. Genome Med 5:69
8. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079
9. Li Y, Vinckenbosch N, Tian G et al (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet 42:969–972
10. Liu Q, Guo Y, Li J et al (2012) Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics 13(Suppl 8):S8
11. Liu X, Han S, Wang Z et al (2013) Variant callers for next-generation sequencing data: a comparison study. PLoS One 8:e75619
12. McKenna A, Hanna M, Banks E et al (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303
13. Ng SB, Buckingham KJ, Lee C et al (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35
14. O'Rawe J, Jiang T, Sun G et al (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5:28
15. Pabinger S, Dander A, Fischer M et al (2013) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. doi:10.1093/bib/bbs086 (Advance access published January 21, 2013)
16. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
17. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26
18. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311
19. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164