medgen 2010 · 22:242–247
DOI 10.1007/s11825-010-0220-5
Online publiziert: 10. Juni 2010

A.M. Stütz · J.O. Korbel
Genome Biology Research Unit, European Molecular Biology Laboratory (EMBL), Heidelberg

Potenzial und Herausforderungen der personalisierten Genomik und des 1000-Genom-Projekts

Fortschritte in der Genomik haben in den letzten Jahren zu einem verbesserten Verständnis des Zusammenhangs zwischen Genotyp und Phänotyp geführt. Die wichtigste genomische Ressource war bisher das HapMap-Projekt, ein Katalog „häufiger" (Allelfrequenz größer als 5%) Einzelnukleotidpolymorphismen (SNPs) in 4 verschiedenen Populationen. Die durch das HapMap-Projekt erfolgte Kartierung von SNPs als „Marker" bildete im Wesentlichen die Grundlage für genomweite Assoziationsstudien (GWAS), bei denen große Patientenkohorten mit phänotypisch unauffälligen Kontrollpersonen verglichen werden, um signifikant angereicherte SNP-Marker aufzufinden und so krankheitsrelevante Kandidatenregionen zu definieren. Für eine Reihe von Krankheiten konnten so neue genomische Kandidatenregionen identifiziert werden, so z. B. für Diabetes oder koronare Herzkrankheit [3]. Allerdings detektieren GWAS-Studien in den meisten Fällen nicht das kausale Allel, sondern stattdessen eine erweiterte Kandidatenregion. Das Auffinden der tatsächlich verantwortlichen, oft sehr seltenen, funktionellen Mutation, z. B. durch die traditionelle „Sanger-Sequenzierung", ist in Folge oft sehr aufwendig.

Hinzu kommt, dass in den meisten Fällen nur ein kleiner Bruchteil des genetisch vererbbaren Anteils einer Krankheit durch eine GWAS-Studie erklärt werden kann. Eine der möglichen Erklärungen ist, dass eine in GWAS-Studien nicht erfassbare Klasse von Mutationen – wie seltene Einzelnukleotidvarianten (Populationshäufigkeit unter 1%) oder genomische strukturelle Variationen [SVs, d. h. zum Beispiel Kopienzahlvariationen (CNVs), Inversionen oder Translokationen] – mitverantwortlich für den Phänotyp sind [11].

Aufgrund neuartiger Hochdurchsatz-Sequenziertechnologien (HDS; auch „next generation sequencing technologies" genannt) können menschliche Genome inzwischen sehr viel schneller und kostengünstiger als mit früheren Technologien sequenziert werden. Immer wieder ist in Fachkreisen wie auch in der Öffentlichkeit das „1000-Dollar-Genom" im Gespräch, mit dem die Zielsetzung verknüpft ist, eine menschliche Genomsequenzierung für diagnostische Zwecke zu ermöglichen. Obwohl die zzt. gängigen Technologien von dem 1000-Dollar-Genom preislich noch relativ weit entfernt liegen [6], ermöglicht die signifikante Verbilligung der Genomsequenzierung schon in der heutigen Zeit einen Durchbruch in der Molekularbiologie und Genetik. Das 1000-Genom-Projekt (1000GP, http://www.1000genomes.org) hat die Zielsetzung, in den kommenden Monaten etwa 2000 humane Genome zu sequenzieren und zu analysieren, um so eine systematische Analyse menschlicher Nukleotidvariationen mit bisher unerreichtem Detailgrad zu ermöglichen [10]. Es ist wahrscheinlich, dass die Ergebnisse des 1000GP die Aussagekraft von nachfolgenden GWAS-Studien durch die verbesserte Auflösung an Variationsmarkern weiter verstärken wird. Weiterhin sind schon jetzt viel versprechende Anschlussprojekte geplant, die beispielsweise die Ermittlung häufiger und seltener Mutationen in verschiedenen Tumoren zur Zielsetzung haben [9].

Die Möglichkeiten der „neuen Form" der genetischen Analyse durch HDS sind jedoch auch mit neuen Herausforderungen verbunden. Insbesondere erfordern die Analyse und die Speicherung der Daten eine professionelle IT-Infrastruktur und Bioinformatik. In diesem Übersichtsartikel werden wir zunächst auf Technologien und „genomische" Applikationen der HDS mit Fokus auf dem 1000GP eingehen und anschließend die Herausforderungen des Projekts an die IT-Infrastruktur und bioinformatische Analyse herausarbeiten.

Abkürzungen

1000 GP	1000-Genom-Projekt
SNPs	Einzelnukleotidpolymorphismen
GWAS	Genomweite Assoziationsstudien
SV	Strukturelle Variation
HDS	Hochdurchsatz-Sequenziertechnologien
CNV	Kopienzahlvariation
bp	Basenpaar
Gb	Gigabasen
ICGC	International Cancer Genome Consortium

Technologien zur Hochdurchsatzsequenzierung

Drei große Sequenzierplattformen haben sich bei den HDS in den letzten Jahren durchgesetzt, nämlich Roche/454 (derzeitige Plattform: Genome Sequencer FLX Titanium), Illumina/Solexa (derzeit Genome Analyzer IIx, bald HiSeq 2000) und Applied Biosystems (derzeit SOLiD v3+, bald SOLiD 4) (s. auch aktuelle Übersichtsarbeit [6]). Das Roche-System ermöglicht pro Sequenzierexperiement („sequencing run") die Generierung von über einer Million Sequenzierreads einer Länge von über 400 bp, während die anderen beiden Plattformen bei kürzeren Sequenzlängen (30–100 bp) eine höhere Anzahl an Reads erzeugen. Alle 3 Sequenzierplattformen ermöglichen bei der Probenbearbeitung eine Erzeugung von Einzelsequenzen (Single Ends) oder von Sequenzpaaren (Paired Ends). In der Genomsequenzierung, z. B. im 1000GP, werden fast ausschließlich Paired Ends eingesetzt, da sich diese bei der Alinierung von Reads gegen das Genom und beim Auffinden von SVs mittels „Paired-End-Kartierung" („paired-end mapping") von Vorteil erwiesen haben [4]. Zusätzlich zu dem hier vorgestellten Einsatz finden HDS-Technologien mittlerweile in der Molekularbiologie breite Anwendung, beispielsweise in der Transkriptionsanalyse (RNA-Seq) oder in der Analyse von DNA-Protein-Bindestellen (ChIP-Seq; [6]).

Bei der Neusequenzierung von Genomen und komplexen genomischen Regionen (wie z. B. SVs) ist die längere Leselänge des Roche/454-Systems von Vorteil, da sie eine genauere Erfassung repetitiver Stellen des Genoms erlaubt und daher qualitativ hochwertige Genom-Assemblies ermöglicht. Im Gegensatz dazu ist bei der Resequenzierung von Individuen zur Erfassung von SNPs eine hohe Gesamtanzahl an Reads ein wichtiger Faktor. Eine hohe Abdeckung von sequenzierten Polymorphismen erhöht die Chance, SNPs von Sequenzierfehlern zu unterscheiden und Genotypen (heterozygote und homozygote SNPs) korrekt zu detektieren. Häufig werden bei der Resequenzierung bis zu 30-fache Abdeckungen des Genoms erzeugt.

Zusammenfassung · Abstract

medgen 2010 · 22:242–247 DOI 10.1007/s11825-010-0220-5

A.M. Stütz · J.O. Korbel

Potenzial und Herausforderungen der personalisierten Genomik und des 1000-Genom-Projekts

Zusammenfassung

Vor Kurzem hat die Sequenzierung individueller menschlicher Genome mittels neuartiger Technologien ein neues Kapitel in der Humangenetik eingeläutet. So hat das 1000-Genom-Projekt (1000GP) die Genomanalyse in 2500 Individuen zur Aufgabe und wird unser Wissen über genetische Variationen durch die Erstellung einer hochauflösenden Karte im Menschen maßgeblich erweitern. So sollen im 1000GP sowohl Einzelnukleotidpolymorphismen als auch strukturelle Variationen mittels Sequenzierung in mehreren ethnischen Gruppen identifiziert werden. Außerdem werden die verwendeten Technologien auf ihre Eignung für Projekte dieses Maßstabs getestet. Letztlich sollen auch neue bioinformatische Lösungen erarbeitet werden, um die 1000GP-Daten effizient für die Forschung aufarbeiten zu können. Dieser neue Katalog an häufigen und seltenen genetischen Varianten wird in naher Zukunft die Entwicklung verbesserter Methoden zur Phänotypassoziation und zur Ermittlung der molekularen Ursache verschiedener Krankheiten ermöglichen.

Schlüsselwörter

1000-Genom-Projekt · Genomanalyse · Genetische Variationen · Einzelnukleotidpolymorphismen · Strukturelle Variationen

Potential and challenges of personalized genomics and the 1000 Genome Project

Abstract

The ability to sequence entire individual human genomes has heralded a new era in human genetics. Such advances in sequencing technologies make it possible to address new questions such as the generation of a comprehensive map of common and rare genetic variants in humans. The 1000 Genome Project will analyze 2500 genomes and is expected to greatly expand our knowledge about genomic variation, both on single nucleotide polymorphisms and genomic structural variants in a number of human ethnic populations. Furthermore, the possibility to use these new sequencing technologies for such large scale projects will be evaluated. Finally, new bioinformatics solutions will be developed to efficiently store and process such large volumes of data for the scientific community. This catalogue of common and rare variations will facilitate the development of better methods for phenotype-genotype associations and help uncover the molecular bases for a variety of diseases in the near future.

Keywords

1000 Genome Project · Genome analysis · Genetic variations · Single nucleotide polymorphisms · Structural variations

In den nächsten Monaten und Jahren werden eine Reihe weiterer Firmen ihre verbesserten oder preisgünstigeren Alternativen anbieten, welche beispielsweise die Sequenzierung einzelner DNA-Moleküle ermöglichen. So wurden erste Geräte von Helicos, PacBio oder Complete Genomics vor Kurzem bereits erfolgreich für individuelle Genomsequenzierungen verwendet, u. a. mit reinen Sequenzierkosten (laut Firmenangaben) von etwa 4000 $ [6].

Das persönliche Genom

In den Jahren 2007 und 2008 gelang erstmalig die Sequenzierung individueller Genome mit traditioneller Sanger-Sequenzierung (Craig Venters Genom) sowie mit der HDS-Technologie von 454/Roche (Jim Watsons Genom). Es folgten Publikationen anonymer Individuen mit Illumina und SOLiD/AB-Technologien Ende 2008 und Anfang 2009 (zusammengefasst in [6]). Dabei wurden bei Bentley et al. 4 Mio. SNPs (74% in dbSNP bekannt), bei Wheeler et al. 3,3 Mio. SNPs (82% bekannt), bei McKernan et al. 3,8 Mio. SNPs (81% bekannt) und bei Venter 3,5 Mio. (79% bekannt) SNPs detektiert. Außerdem hat jede personelle Genomstudie SVs identifiziert, wenngleich mit unterschiedlich hoher Auflösung.

Heutzutage deckt jede Studie eine Vielzahl neuer, potenziell seltener Varianten auf – der Fundus an bekannten Polymorphismen ist also bei Weitem noch nicht ausgeschöpft. Weiterhin konnte gezeigt werden, dass selbst normale, gesunde Menschen mehrere genetische Nukleotidvarianten in sich tragen, die nach unserem heutigen Wissen die Funktion eines Proteins stark beeinträchtigen sollten. Beispielsweise wurden in den Genomen von Venter und Watson 3882 und 3766 Kodon verändernde (d. h. nichtsynonyme) SNPs identifiziert. Außerdem fanden sich in bisher jedem publizierten „persönlichen Genom" eine Reihe von Stopkodonvarianten, Veränderungen des Leserahmens („frame shifts"), Deletionen funktionaler Exons oder Genfusionsvarianten („gene fusion variants").

Um aber einschätzen zu können, welcher Anteil der neu kartierten Variationen spezifisch für ein Individuum sind, sollte die Frequenz dieser Variation in einer größeren Anzahl an Leuten unterschiedlicher Herkunft bestimmt werden. Aus diesem Grund ist, unter Beteiligung von Forschungszentren in den USA, Großbritannien, Deutschland und China, das 1000GP ins Leben gerufen worden.

Das 1000-Genom-Projekt (1000GP)

Wissenschaftliche Ziele

Eines der Hauptziele des 1000GP ist die Kartierung aller Variationen (sowohl SNPs als auch SVs) mit einer Allelhäufigkeit von mehr als 1% in verschiedenen Populationen [10]. Ferner sollen die Bereiche im Genom mit wahrscheinlicher funktioneller Relevanz (proteinkodierende Exons, regulatorische Bereiche) durch den Einsatz von Anreicherungsmethoden noch eingehender analysiert werden, um in diesen Regionen die Detektion von Variationen bis zu einer Allelhäufigkeit von 0,1–0,5% zu ermöglichen. Neben Allelhäufigkeiten sollen hochauflösende Haplotypen bestimmt werden. Weiterhin soll auch das Referenzgenom mittels der neuen Daten verbessert und vervollständigt werden. In der Tat wird man in vielen Fällen erst nach Ablauf des 1000GP eindeutig bestimmen können, ob eine Variante das Hauptallel oder das seltene Allel (häufig die Neumutation) darstellt.

Darüber hinaus werden die immer noch relativ neuen HDS-Technologien auf ihre Tauglichkeit überprüft, solche Fragestellungen verlässlich zu beantworten. Die genomische DNA der bereits im HapMap-Projekt verwendeten Personen ist in dieser Hinsicht als Qualitätskriterium gut geeignet, da es einen direkten Vergleich zu bereits bekannten SNP-Daten ermöglicht. Zuletzt befasst sich eine eigene Gruppe innerhalb des 1000-Genom-Projekts mit den ethischen Fragestellungen, die mit der Veröffentlichung kompletter menschlicher Genome und dem Schutz von Individuen zusammenhängen. In Summe werden die Ergebnisse des 1000GP der Wissenschaft einen verbesserten Katalog seltener und häufiger genetischer Variationen liefern. Dieser Katalog kann in zukünftigen Forschungsprojekten als Nachschlagewerk verwendet werden, z. B. um relevante biomedizinische Fragen wie „Kommt eine in Patienten identifizierte Mutation auch in gesunden Menschen vor und wie oft?" zu beantworten.

Pilotphase

In der Planungsphase des 1000GP wurde die optimale Vorgehensweise in Bezug auf Sequenziertiefe und Anzahl der zu untersuchenden Individuen – bei einem gegebenen Forschungsbudget – diskutiert. So ermöglicht ein tiefes Sequenzieren (z. B. eine 30-fache Abdeckung jedes einzelnen Nukleotids) die Genotypisierung eines Großteils der SNPs. Alternativ führt eine breite Suche mit vielen Samples bei geringer Abdeckung des Genoms (2- bis 4-fach) zum Auffinden einer höheren Zahl unterschiedlicher Variationen, da zum gleichen Preis mehr individuelle Genome sequenziert werden können. So wurde eine Pilotphase mit 3 Teilzielen gestartet, welche einen Kompromiss beider Vorteile vereint, um den besten Ansatz für die Hauptphase des 1000GP zu bestimmen.

- Der 1. Ansatz hat zum Ziel, in jeweils 60 nicht verwandten Personen europäischen, nigerianischen und ostasiatischen Ursprungs 2- bis 4-fache genomische Abdeckung zu erreichen (Datenmenge etwa 1080 Gb; d. h. $1*10^{12}$ Basen).
- Der 2. Ansatz zielt darauf ab, je ein Familientrio (Eltern und Kind) europäischen und nigerianischen Ursprungs mit 20- bis 60-facher Abdeckung zu sequenzieren.
- Der 3. Ansatz soll mittels Sequenzanreicherung („sequence capture") in 1000 Genen von fast 700 Personen eine 50-fache Abdeckung von DNA-Sequenz erzielen.

Erste Lektionen

Die Pilotphase des 1000GP steht kurz vor dem Abschluss [10]. Beispielsweise wurden im Rahmen des ersten Pilotansatzes des 1000GP zum Zeitpunkt der Einreichung dieses Papers 180 Individuen mit der angestrebten 2- bis 4-fachen haploiden Abdeckung durchsequenziert, was einer Gesamtmenge von mehr als einer Terabase entspricht. Im März 2010 wurden vom 1000GP die neuesten SNP-Genotypisie-

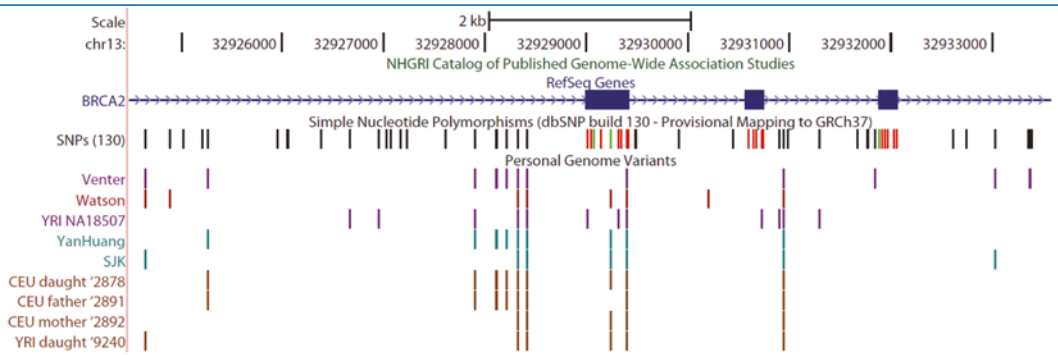

Abb. 1 ▶ Vergleich der SNP-Sequenzunterschiede einiger persönlicher Genome inklusive der Pilot-2-Individuen des 1000 GP in einem Teilbereich des *BRCA2*-Gens im UCSC Genome Browser

rungsergebnisse der Pilotprojekte veröffentlicht (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/). Insgesamt wurden schon weit mehr als 9 Mio. neue SNPs im Rahmen des 1000GP identifiziert, die vor dem Projekt noch nicht in der öffentlichen SNP Datenbank dbSNP enthalten waren. Da die Validierungsrate hoch ist, eignet sich die HDS-Sequenzierung demzufolge generell auch im großen Maßstab für das Auffinden von SNPs. Aufgrund der relativ kurzen Sequenzierungsreads von HDS-Technologien ist das Auffinden von SVs im Vergleich zu SNPs wesentlich schwieriger. Deswegen entwickeln Mitglieder der Analysegruppe des 1000GP derzeit optimierte SV-Kartierungsansätze.

Erste Sequenzdaten können bereits im UCSC Browser (http://genome.ucsc.edu/) angezeigt werden (**Abb. 1**).

Anhand der bisher erhaltenen Daten wird geplant, in der Hauptphase des Projekts die Genome von 2500 Personen aus 25 Populationen mit 4-facher Abdeckung zu sequenzieren, was zu einem rasanten Anstieg durchsequenzierter Genome führen wird (**Abb. 2**). Die Datengenerierung des Pilotprojekts steht Anfang bis Mitte 2010 unmittelbar vor der Vollendung.

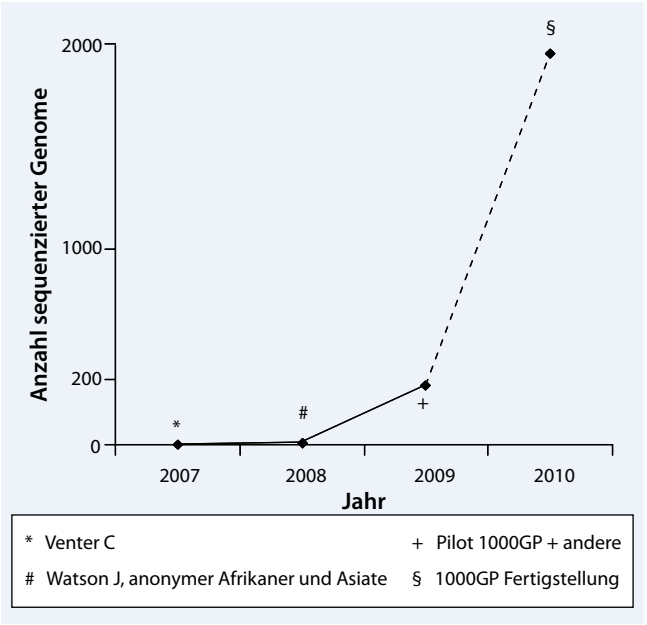

Abb. 2 ▶ Einfluss des 1000GP auf die Anzahl der publizierten menschlichen Genome. *Gestrichelte Linie* Vorhergesagte Zahl der sequenzierten Genome bis Ende 2010/Anfang 2011

Herausforderungen in Bezug auf Bioinformatik und IT-Infrastruktur

Eine Reihe bioinformatischer Arbeitsschritte ist im Zusammenhang mit dem 1000GP notwendig, um aus den riesigen Datenmengen im Petabyte-Bereich (10^{15} Bytes), die im Rahmen des 1000GP entstehen werden, sinnvolle Informationen zu gewinnen. Für diesen Zweck wurde eine Vielzahl von bioinformatischen Algorithmen und Analyseprogrammen entwickelt. Ein wichtiger erster Schritt ist das korrekte Zuweisen einer Positionsangabe des Sequenzstücks (Alinierung), ein Schritt der nur in nichtrepetitiven Stellen des Genoms eindeutig durchführbar ist. Hierfür gibt es inzwischen eine ganze Reihe von bioinformatischen Methoden (z. B. BWA, Eland, MAQ, Bowtie, SOAP2, NovoAlign, SHRiMP und RazerS). Diese Methoden können direkt zur SNP-Detektion eingesetzt werden.

Zur wesentlich schwierigeren SV-Detektion werden separate Ansätze wie die Paired-End-Kartierung („pair-end mapping"; [4]) angewendet, die eine abweichende Distanz oder Orientierung von Readpaaren mittels nachgeschalteter statistischer Analyse zur SV-Detektion verwendet. Ein alternativer Ansatz ist die Sequenztiefenanalyse („depth-of-coverage analysis"; [1]), wo anhand der erhaltenen Anzahl der sequenzierten Reads Rückschlüsse auf die Kopienzahl einer genomischen Region gemacht werden. Weiterhin können Readsequenzen in einigen Fällen direkt zur SV-Identifizierung mittels Alinierung verwendet werden, wenn der Bruchpunkt innerhalb der Sequenz enthalten ist („Split-Read-Analyse"). Schließlich entstanden in den letzten Monaten erste Algorithmen zur Assemblierung kurzer DNA-Sequenzen, die in vielen Fällen SVs identifizieren können.

Ein wichtiger Aspekt der neuen Sequenzierprojekte ist die IT-Infrastruktur. Die in Zukunft benötigten Speicherkapazitäten für die Archivierung und Bearbeitung der Daten übersteigen die aktuellen Kapazitäten vieler biologischer und medizinischer Institute. Große Sequenzierprojekte wie das 1000GP halten mittlerweile Rohdaten in Form von Bildern, die mehrere Terabyte groß sind und zur Erstellung der Sequenz erzeugt werden, nur noch für kurze Zeit. Stattdessen veröffentlichen sie die vom „Basecaller" identifizierten Sequenzierreads sowie Alinierungen dieser Reads (inklusive Qualitätsscores) gegen das Referenzgenom.

Während des Pilotprojekts wurden bereits mehr als 3,8 Terabasen sequenziert und vorab ins Netz gestellt (s. ftp://ftp.1000genomes.ebi.ac.uk). Diese Daten enthalten sowohl gefilterte Sequenzreads mit Qualitätsscores (FASTQ-Dateiformat) als auch komprimierte Ergebnisse von Readalinierungen (BAM-Dateiformat). Die Datenmengen im Tera- und Petabytebereich sind derzeit außerhalb spezialisierter Zentren nur schwer handhabbar. Genomdaten können jedoch prinzipiell bis auf einzeln kartierbare Unterschiede im Vergleich zum Referenzgenom komprimiert werden. Im letzteren Fall reduzieren sich die Datenmengen auf wenige Megabytes. Im Beispiel des Genoms von James Watson erfolgte eine Komprimierung auf 4 MB [2], eine Datenmenge, die sich an normale E-Mails anhängen lässt und deshalb relativ schnell zwischen verschiedenen Forschungsgruppen ausgetauscht werden kann.

Insgesamt wird im Zusammenhang mit der HDS die Bedeutung der Bioinformatik in der Genomik, Genetik und der medizinischen Forschung vermutlich weiter zunehmen. Neben Datenspeicherungskapazitäten werden zukünftig im Zusammenhang mit dem 1000GP und verwandten Projekten in der biomedizinischen Forschung schnelle Computercluster mit ausreichendem Arbeitspeicher benötigt. Dadurch verschiebt sich der Hauptanteil des benötigten Zeitaufwands und der personellen und finanziellen Ressourcen in der Genetik und Genomik weg von der Generierung der Daten (welche nur wenige Tage dauert) hin zur Datenanalyse und -speicherung. In der Tat veranschlagen Strom- bzw. Kühlkosten der benötigten Datenserver einen nennenswerten Betrag der Gesamtkosten solcher Projekte. Die Kosten zur Haltung von Daten liegen derzeit – wenn Bilddaten vorgehalten werden sollen – sogar in derselben Größenordnung wie die Resequenzierung.

Bestehende Limitierungen

Ein Schwerpunkt der Forschung an Daten des 1000GP ist die Behebung derzeitiger Limitierungen. Beispielsweise ist derzeit unklar, ob eine 2- bis 4-fache Abdeckung des Genoms für eine akkurate SV-Analyse ausreicht, und eine Reihe von Arbeitsgruppen, inklusive unserer EMBL-AG, beschäftigen sich deshalb mit der Verbesserung bioinformatischer Methoden zur SV-Detektion.

Bei der Generierung von Datensätzen der Größenordnung des 1000GP muss eine Lösung für die Speicherung und Distribution der Daten gefunden werden. In diesem Zusammenhang ist beispielsweise die Bandbreite des Internets ein limitierender Faktor. Selbst relativ schnelle Verbindungen, wie die mittlerweile an vielen Institutionen üblichen 1Gbit/s-Verbindungen, werden in Zukunft nicht ausreichen, um die Datenmengen zu bewältigen.

Eine weitere Limitierung entsteht durch die Verwendung der erzeugten kurzen Sequenzreads, welches sowohl Einfluss auf die Qualität der Daten als auch auf die Verwendung von Analysemethoden hat. So kann das volle Potenzial der Split-Read-Analyse erst dann entfaltet werden, wenn die beiden Seiten, die den Bruchpunkt einer strukturellen Variation flankieren, lang genug bleiben, um eindeutigen Positionen im Genom zugeordnet werden zu können. Durch das stetige Ansteigen der möglichen Leselänge wird diese Methode bald an Einfluss gewinnen. So wird wahrscheinlich in Zukunft eine simultane Verwendung mehrerer Technologien – u. U. auch die zukünftig verfügbare Einzelmolekülsequenzierung („single molecule sequencing") – das bestmögliche Analyseergebnis liefern können.

Resequenzierungsstudien sind wesentlich von der Qualität des Referenzgenoms abhängig, da alle Sequenzierdaten zunächst gegen dieses Genom verglichen werden. Die aktuelle, lineare Version des humanen Referenzgenoms, welche auf der Sequenz weniger Individuen basiert, hat noch mehrere Nachteile. Erstens fehlen einige Bereiche, die durch die technische Limitierung der tradionellen Sanger-Sequenzierung damals nicht erforscht werden konnten und somit heutzutage nicht zugeordnet werden können. Weiterhin kommt der Frage, auf wessen Genom die Referenz eigentlich basiert, eine neue Bedeutung zu, da diese als Referenzmaßstab verwendet werden soll. So hat sich gezeigt, dass in manchen Fällen das Referenzgenom nicht das häufige, sondern das seltene SNP- oder SV-Allel beinhaltet. Vermutlich werden Sequenzalinierungen in naher Zukunft gegen umfassende Kataloge von Sequenzvarianten – und nicht nur gegen ein (prinzipiell arbiträres) Referenzgenom – durchgeführt.

Ausblick

Die weitere Verbilligung der Genomsequenzierung wird in den nächsten Jahren dazu führen, dass immer mehr genetisch bzw. medizinisch orientierte Forschungsprojekte individuelle Genomsequenzen hervorbringen werden. Dabei werden sowohl die Genome phänotypisch unauffälliger Personen als auch Genome definierter Patientengruppen in großer Zahl durchsequenziert werden, z. B. um mit neuer Methodik den genomischen Ursachen von Krankheiten auf den Grund zu gehen. Der mögliche Nutzen der HDS-Sequenzierung in der genetischen Medizin ist enorm, wie bereits erste Studien belegen. Beispielsweise haben Ng et al. vor Kurzem durch eine gezielte Anreicherung der proteinkodierenden Sequenzen des Genoms in mehreren Patienten krankheitserzeugende Variationen für das seltene, dominant vererbte Freeman-Sheldon-Syndrom mittels HDS-Sequenzierung identifiziert [7]. Diese Anreicherung ermöglicht eine relativ kostengünstige Fokussierung des Sequenzierungsansatzes auf genomische Kandidatenregionen – das ist ein Ansatz, der in den kommenden Monaten sicherlich vielseitige Anwendung finden wird.

Ansätze, die im Rahmen des 1000GP entwickelt worden sind, werden seit Neuestem schon in vielversprechenden Krebsforschungsprojekten angewandt, wie dem International Cancer Genome Consortium (ICGC; http://www.icgc.org; [9]). Im Rahmen des ICGC hat sich das deutsche Konsortium „ICGC PedBrain Tumor" zur Aufgabe gemacht, das Genom von etwa 600 kindlichen Gehirntumoren (Medulloblastom und pilozytisches Astrozytom) sowie von Blutgewebe derselben Patienten komplett durchzusequenzieren und sowohl das Genom als auch das Transkriptom auf tumorspezifische Aberrationen zu untersuchen. Deren Auftreten wird dann mit den klinischen Phänotypen korre-

liert, um die pathogenen Mutationen herauszufiltern. Unser Labor ist gemeinsam mit anderen Forschungszentren in Heidelberg, Berlin und Düsseldorf an dieser Studie beteiligt und wird dabei v. a. größere genomische Aberrationen, wie Deletionen, Amplifikationen und Translokationen, mittels HDS-Sequenzierung und bioinformatischer Analyse untersuchen.

Die im Rahmen des 1000GP sequenzierten und analysierten Daten stellen einen Meilenstein im Verständnis des humanen Genoms dar. Wir erwarten, dass die Auswirkungen auf zukünftige genomische Studien mit medizinischer Relevanz beträchtlich sein werden – sowohl durch eine Verbesserung von GWAS-Studien (mehr SNPs) als auch durch die Etablierung der personalisierten Genomik als neuer Ansatz in der medizinischen Genetik. Die personalisierte Genomik hat das Potenzial, funktionale Varianten für Syndrome mit derzeit unbekannter Ursache (Mendel- und komplexe Krankheiten) durch eine Sequenzierung von Kandidatengenen in vielen Individuen zu ermöglichen. Aktuelle erste Beispiele für Morbus Charcot-Marie-Tooth [5] und Miller-Syndrom bzw. Kartagener-Syndrom [8] zeigen, dass diagnostische Anwendungen mit den neuen Sequenzierverfahren bereits möglich sind. Ultimativ erwarten wir, dass die Existenz von vielen kompletten individuellen Genomen in naher Zukunft die Voraussetzung schaffen wird, die viel zitierte „personalisierte Medizin" Wirklichkeit werden zu lassen.

Korrespondenzadresse

J.O. Korbel
Genome Biology Research Unit
European Molecular Biology Laboratory (EMBL)
Meyerhofstraße 1, 69117 Heidelberg
korbel@embl.de

Danksagung. Wir danken den Mitgliedern des 1000-Genom-Projekts und des ICGC PedBrain Tumor Konsortiums für ausführliche Diskussionen. Des Weiteren möchten wir uns bei allen Kollegen entschuldigen, auf deren Originalarbeiten wir aufgrund der Limitierung der Literaturzitate nicht verweisen konnten.

Interessenkonflikt. Der korrespondierende Autor gibt an, dass kein Interessenkonflikt besteht.

Literatur

1. Campbell PJ, Stephens PJ, Pleasance ED et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40:722–729
2. Christley S, Lu Y, Li C, Xie X (2009) Human genomes as email attachments. Bioinformatics 25:274–275
3. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251
4. Korbel JO, Urban AE, Affourtit JP et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426
5. Lupski JR, Reid JG, Gonzaga-Jauregui C et al (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362:1181–1191
6. Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11:31–46
7. Ng SB, Turner EH, Robertson PD et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276
8. Roach JC, Glusman G, Smit AF et al (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science Mar 10. [Epub ahead of print]
9. The International Cancer Genome Consortium (2010) International network of cancer genome projects. Nature 464:993–998
10. Via M, Gignoux C, Burchard EG (2010) The 1000 Genomes Project: new opportunities for research and social challenges. Genome Med 2:3
11. Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451–481