Genomsequenzierung

Next Generation Sequencing in der mikrobiellen (Meta)Genomforschung

ANDREA THÜRMER
LABOR FÜR GENOMANALYSE, UNIVERSITÄT GÖTTINGEN

In the past two decades genome sequencing was getting more and more one of the most important research fields of life sciences. It was accelerated by the introduction of the next-generation sequencing technologies in 2005. With the further development of the technologies today, there is a bride field of features and applications that one can use for genome research.

10.1007/s12268-014-0424-3

Der initiale Meilenstein in der mikrobiellen Genomforschung wurde 1995 mit der ersten vollständigen Genomsequenz des Bakteriums *Haemophilus influenza* gelegt [1]. In nur 18 Monaten hatten damals Craig Venter und der Mikrobiologe Hamilton Smith zusammen mit drei Dutzend Mitarbeitern die 1,8 Millionen Basenpaare des Bakteriums mithilfe der Sanger-Sequenziertechnik [2] entschlüsselt.

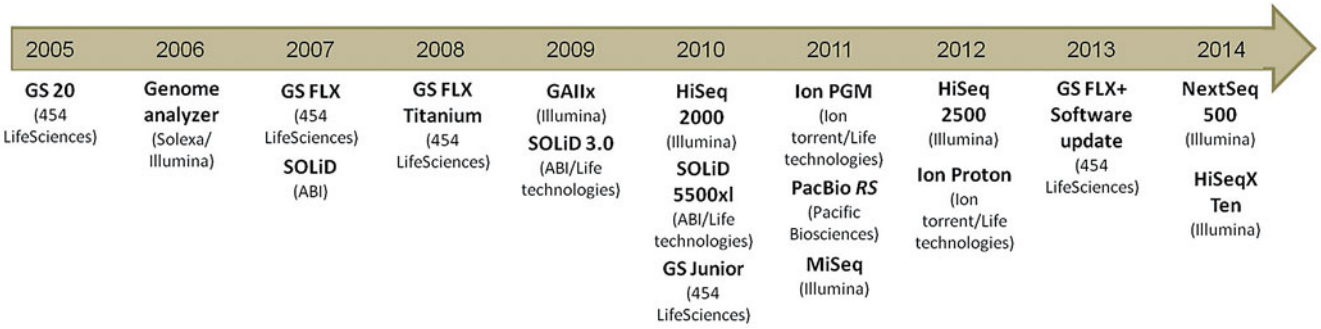
Die funktionelle Genomforschung der Mikroorganismen gehört heute zu den weltweit bestimmenden Gebieten der Lebenswissenschaften. Ergebnisse dieses Forschungszweigs haben bereits weitreichende Auswirkungen auf unterschiedliche Gesellschafts- und Wirtschaftsbereiche (Gesundheit, Klimawandel, Biotechnologie, Ernährung) und werden maßgeblich die Wissenschaftslandschaft der Zukunft prägen.

Mit dem Start des *Next Generation Sequencing* im Jahr 2005 wurde der Startschuss zum massiven parallelen Sequenzieren von Millionen DNA-Fragmenten in einem Sequenzierlauf gegeben. Von nun an hatte man die Möglichkeit, in kürzester Zeit die Erbinformation von mehreren Bakterien gleichzeitig oder auch von ganzen Metagenomen zu entschlüsseln. In den folgenden Jahren stieg nicht nur die Zahl der sequenzierten Genome, auch die Entwicklung neuer Sequenzierplattformen und die Anzahl der Genomsequenzen nahm rasant zu.

2001 kam es im Göttinger Labor für Genomanalyse zum Abschluss der ersten komplett geschlossenen Genomsequenz von *Methanosarcina mazei* Gö1 [3]. Dieses Bakterium hat eine sehr große Bedeutung in verschiedenen Fermentationsprozessen. Neben der Sequenzierung und Erforschung von diversen pathogenen Organismen oder auch Bakterien mit biotechnologischem Potenzial liegt ein großer wissenschaftlicher Fokus auf der Sequenzierung von mikrobiellen Lebensgemeinschaften in verschiedenen Ökosystemen. Zum einen wird die Diversität der verschiedenen Standorte untersucht und zum anderen deren funktionelles Zusammenspiel sowie die Wechselwirkungen auf sich ändernde Umweltbedingungen [4–6].

Next Generation Sequencing-Plattformen

Die Firma 454 Life Sciences brachte 2005 mit dem GS-20-Instrument das erste *Next Generation*-Sequenziergerät auf den Markt. Dieses System basiert auf der Pyrosequenzierung und hatte anfangs eine Leseweite von ca. 100 Basenpaaren [7]. Mit dem weiterentwickelten System GS FLX+ kommt die Plattform mittlerweile auf eine Leseweite von bis zu 1.000 Basenpaaren. Die kleinere Variante, GS Junior, ist mit etwa 400 Basenpaaren Leseweite das ideale System für schnelle Amplikon-Sequenzierungen. Die mittlerweile am weitesten verbreitete Technologie mit den meistverfügbaren Systemen ist die *Sequencing-by-Synthesis*(SBS)-Sequenzierplattform von Illumina. Mit ihren verschiedenen Geräten deckt Illumina den unterschiedlichen Bedarf an Durchsatzraten ab: das MiSeq-Benchtop-System, das in drei Tagen bis zu 30 Millionen *reads* generiert, der Genome Analyzer IIx mit einem mittleren



▲ **Abb. 1:** Zeitskala der Erscheinung von verschiedenen kommerziellen Sequenzierplattformen.

Tab. 1: Leistung und Merkmalen von Sequenzierplattformen.

System	GS FLX+ (454, LifeSciences)	GS Junior (454, LifeSciences)	GAIIx (Illumina)	HiSeq 2500 (Illumina)	MiSeq (Illumina)	NextSeq 500 (Illumina)	HiSeqX Ten (Illumina)	Solid 5500xl (ABI/Life Technologies)	Ion Proton (Ion Torrent/Life Technologies)	Ion PGM (Ion Torrent/Life Technologies)	PacBio *RS* (Pacific Biosciences)
Sequenzier-mechanismus	Pyro-sequen-zierung	Pyro-sequen-zierung	SBS	SBS	SBS	SBS	SBS	Ligation & zwei Basen-Codierung	Protonen-detektion	Protonen-detektion	Echtzeit-sequen-zierung
Readlänge (bp)	~ 800	~ 400	2x125 (H) 2x150 (R)	2x150	2x300	2x150	2x150	75	bis zu 200	200 oder 400	8.000 (avg)
Durchsatz/ Lauf (bis zu, GB)	0,7	0,035	1.000 (H) 180 (R)	95	15	120 (H) 39 (M)	1.800	15	10	1 oder 2	0,4
Laufzeit	24 h	8 h	6 d (H) 2 d (R)	10 d	3 d	30 h (H) 26 h (M)	< 3 d	8 d	4 h	4 h oder 7 h	3 h
Fehlerrate	>0,8 %	>0,5 %	>0,3 %	>0,8 %	>0,8 %	x	x	>0,5 %	>1 %	>1,5 %	>10 %

H: Hochdurchsatzmodus, R: Rapid-/Schnellmodus, M: mittlerer Modus, SBS: *sequencing by synthesis*, avg: Durchschnittswert, x: bisher keine Angaben.

Durchsatz von bis zu 640 Millionen *reads* und die Hochdurchsatzvariante, der HiSeq2500, der in sechs Tagen bis zu zwei Billionen *reads* generieren kann. Mit dem gerade im Januar 2014 erschienenen NextSeq500-System bietet Illumina ein weiteres Benchtop-System und schließt mit einem Durchsatz von 400 Millionen *reads* in nur 30 Stunden die Lücke vom kleinen Durchsatzgerät zum Hochdurchsatzsystem. Das ebenfalls im Januar erschienene System HiSeqX Ten erweitert die Hochdurchsatzsequenzierung (drei Billionen *reads* in weniger als drei Tagen). Allerdings wird bei dem großen Durchsatz auf die *read*-Länge verzichtet. Lediglich der MiSeq schafft es auf eine Länge von 2 × 300 Basenpaaren, während die anderen Systeme auf zurzeit maximal 2 × 150 Basenpaare kommen.

Die Ion-Torrent-Systeme, Ion Torrent Personal Genome Machine (PGM) und Ion Proton, sind als Benchtop-Varianten eine schnelle und kostengünstige Alternative, gerade wenn es um Genomsequenzierungen geht, die einen mittleren Durchsatz benötigen. Diese Systeme

Tab. 2: Sequenzierabdeckung für eine optimale *de novo*-Sequenzierung. Grundlage der Daten sind *assemblies* von mehr als 100 sequenzierten bakteriellen Genomen im Göttinger Labor für Genomanalyse.

Abdeckung / Geräte/kombination	GS FLX+	Illumina	PacBio RS II
GS FLX+	12–15 x	–	–
Illumina	–	100 – 120 x	–
Illumina/GS FLX+	5–10 x	50 – 80 x	–
Illumina/GS FLX+/PacBio RSII	4–8 x	50 – 80 x	5 x

Tab. 3: Hybridassemblierungen mit dem MIRA-Assembler (Bastien Chevreux, http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html).

assembly Ergebnisse / Sequenzierplattform	Illumina	Illumina/ GS FLX+	Illumina/ GS FLX+/ PacBio RS II
contig-Zahl	392	273	120
consensus (bp)	5.995.535	6.073.540	6.143.325
größter *contig* (bp)	304.824	572.524	937.459
N50 *contig* (bp)	90.032	141.292	351.682

beruhen auf der Protonendetektion bei der Sequenzierung. Das SOLiD-5500-XL-System von Life Technologies hat mit seiner Ligation und Zwei-Basencodierung eine sehr geringe Fehlerrate innerhalb aller Plattformen. Alle diese Systeme basieren auf der Amplifikation von DNA-Fragmenten und führen zu einem mehr oder weniger starken *bias* in der relativen Häufigkeit bestimmter DNA-Fragmente. Um dieses Problem zu umgehen und auch längere Fragmente zu generieren, wurde mit der Einzel-DNA-Molekül-Sequenzierung die *Third Generation Technology* eingeläutet. Dabei wird in Echtzeit die DNA-Sequenz direkt von einem einzigen Molekül bestimmt, ohne einen Amplifikationsschritt. Der erste Einzelmolekül-Sequenzierer wurde 2010 von Heliscope auf den Markt gebracht, gefolgt vom SMRT™(*single molecule real-time*)-Sequenzierer, PacBio RS II, der 2011 von Pacific Biosciences eingeführt wurde.

Die verfügbaren Technologien bieten ein großes Spektrum in der Anwendung, wie z. B. *de novo*-Sequenzierung von prokaryotischer und eukaryotischer gDNA/cDNA, Resequenzierung, Transkriptomanalysen, Methylom-/Epigenomstudien, ChIP-Seq (*chromatin immunoprecipitation DNA sequencing*) sowie Amplikon-basierende Markergenanalysen und Metagenomsequenzierung. Je nach spezifischer Anwendung kann man die unterschiedlichen Leistungen, Leseweiten oder den gewünschten Durchsatz nutzen (**Abb. 1**, **Tab. 1**).

Auch wenn Pacific Biosciences im vergangenen Jahr erhebliche Fortschritte in der Leseweite und der Reduktion der Fehlerrate gemacht hat, ist es derzeit immer noch ein Engpass, direkt Leseweiten über zehn Kilobasen mit einer Fehlerrate unter einem Prozent zu generieren. Die 2012 auf der Konferenz *Advances in Genome Biology and Technology* (AGBT) angekündigte Nanopore-Technologie von Oxford Nanopores, welche lange Leseweiten bis zehn Kilobasen verspricht, ist bis zum jetzigen Zeitpunkt lediglich ein Prototyp. Die einzige Alternative mit qualitativ hochwertigen *reads* von ca. zehn Kilobasen bietet die von Illumina 2012 gekaufte *long-read*-Sequenziertechnologie Moleculo (www.illumina.com/technology/moleculo-technology.ilmn) [8, 9]. Dabei wurde eine neue Probenvorbereitung generiert, bei der lange DNA-Fragmente hergestellt werden, die mittels Illumina-Standardsequenzierer sequenziert und hinterher bioinformatisch wieder zusammengesetzt werden. Diese Methode ist bis jetzt allerdings noch nicht kommerziell verfügbar. Es wird sich zeigen, ob eine markttaugliche Technologie angeboten werden kann, die es ermöglicht, kostengünstige und labortaugliche lange Leseweiten in Echtzeit zu generieren und somit ein direktes Genom-*finishing* zu realisieren.

Genomsequenzierung

Betrachtet man nur die entstehenden Kosten pro Megabase, würde man sich bei der Sequenzierung eines bakteriellen Genoms für ein Hochdurchsatz-Sequenziergerät von Illumina entscheiden. Doch um eine komplette hochqualitative Genomsequenz zu generieren, sind mehrere Parameter entscheidend. Jede Technologie hat ihre Vor- und Nachteile hinsichtlich der Leseweite, Qualität oder auch dem Umgang mit z. B. GC-reichen bzw. AT-reichen DNA-Regionen. Hinzu kommt, dass die Programme für die anschließenden Auswertungen der Daten ebenfalls unterschiedliche Ansprüche haben. Daher ist es notwendig, verschiedene Technologien miteinander zu kombinieren.

Die längeren Leseweiten der Systeme GS FLX+ (454) und PacBio RS II bieten die Möglichkeit, repetitive Bereiche in mikrobiellen Genomen, wie z. B. genomische Inseln, Phagenbereiche oder auch Genduplikationen, aufzulösen, während man mit den Illumina-Systemen durch die hohe Sequenzabdeckung die Qualität deutlich verbessern kann. Zusätzlich kann mit dem Illumina-System die GS-FLX+-Schwäche in den Homopolymerbereichen ausgeglichen werden.

Ein weiterer Punkt, den man bei der Genomsequenzierung berücksichtigen muss, ist die Menge und vor allem die Qualität der DNA, die man für die Sequenzierung einsetzt. Möchte man auf die langen *reads* der 454-Technologie zurückgreifen, braucht man ca. ein Mikrogramm hochmolekularer und -qualitativer DNA. Bei Nutzung der vorteilhaften *paired-end*-Technologie, mit der die resultierenden *contigs* geordnet werden können, benötigt man sogar bis zu 30 Mikrogramm hochqualitativer DNA. Für die Herstellung einer Genbank für eine PacBio-Sequenzierung braucht man mindestens zehn Mikrogramm hochqualitativer und vor allem hochmolekularer DNA. Nicht jedes Bakterium wächst in einer so hohen Zelldichte, die es ermöglicht, diese Mengen an DNA herzustellen. Eine Alternative bieten die Transposon-basierten Genombanken von Nextera (Nextera DNA Sample Preparation Kit, Illumina), die bereits bei einem Nanogramm hochmolekularer DNA zum Einsatz kommen. Allerdings muss man hier in Kauf nehmen, dass diese Bibliotheken einen gewissen *bias* in der Fragmentierung aufweisen und nur mit dem Illumina-System sequenziert werden können.

Wenn die DNA nicht der limitierende Faktor ist und man auf alle drei Systeme (454, Illumina, PacBio) zugreifen kann, so wäre es der Goldstandard, die Systeme für eine bakterielle *de novo*-Sequenzierung zu nutzen.

Bei der Genomsequenzierung muss man unterscheiden, ob man von einer komplett geschlossenen Genomsequenz spricht oder von einer *scaffold*-Sequenz. Ein Genom mit einem *scaffold* (Gerüst) bedeutet, dass alle resultierenden überlappenden Fragmente (*contigs*), die bei der Assemblierung einer Rohsequenzierung entstehen, über Paarinformationen zu einem *scaffold* geordnet werden können. Es fehlen aber noch Sequenzteile zwischen den *contigs*, die teilweise wichtige Informationen beinhalten können. Fehlende Sequenzinformationen können z. B. über die Vollständigkeit eines potenziellen *open reading frame* (ORF) entscheiden und

demnach auch über vorhandene Funktionen eines Organismus. Für den kompletten Genomschluss benötigt man zusätzlich die Sanger-Sequenziertechnologie. Denn mit den verfügbaren *Next Generation*-Technologien ist es nur in wenigen Fällen möglich, eine „Ein-*contig*-Genomsequenz" zu erstellen, da es neben schwer sequenzierbaren DNA-Bereichen zusätzlich innerhalb der Genomstruktur *repeats* gibt, die außerhalb der Leseweiten der Sequenziersysteme liegen. Die verbleibenden Lücken werden durch PCR-basierte Methoden und deren anschließende Sequenzierung nach Sanger geschlossen. Um ein „Genom-*finishing*" voranzutreiben, ist die Kombination verschiedener Technologien mit bestimmten Abdeckungen, vor allem in der Kombination mit Systemen längerer *reads* sehr vorteilhaft (**Tab. 2**).

Die Nutzung der unterschiedlichen Technologien hat einen erheblichen Einfluss auf die komplette Fertigstellung eines bakteriellen Genoms. Mit Hybridassemblierungen aus verschiedenen Datentypen aus der Sequenzierung eines *Bacillus*-Stamms konnten wir zeigen, dass die *contig*-Zahl mit der Nutzung längerer Leseweiten deutlich reduziert wird (**Tab. 3**).

Mit den zusätzlichen langen *reads* aus der PacBio-Sequenzierung konnten in diesem Organismus über 70.000 Basenpaare *repeats* aufgelöst werden.

Einen ebenso vielversprechenden Ansatz haben A. Voskoboynik *et al.* im Jahr 2013 mit der Genomsequenzierung von *Botryllus schlosseri* gezeigt [9]. Dabei haben sie das 725 Megabasen große und 16 Chromosomen umfassende Genom in zwei Ansätzen sequenziert. Zum einem wurde die DNA mit dem *shot-gun*- und *paired-end*-Protokoll jeweils auf dem GS FLX (454) und auf dem Genome Analyzer II (Illumina) sequenziert und kombiniert assembliert, und zum anderen wurde die DNA mit der neuen *long-read*-Genomsequenzierungs- und Assemblierungsmethode analysiert. Die Kombination aus beiden Sequenzierungen konnte die Daten der neuen *long-read*-Methode bestätigen. Diese Erfolge zeigen, wie wichtig es für die Genomforschung ist, die vorhandenen Systeme mit langer Leseweite zu verbessern oder aber neue Systeme für diese Anwendungen zu etablieren. ■

Literatur

[1] Fleischmann RD, Adams MD, White O et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

[2] Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

[3] Deppenmeier U, Johann A, Hartsch T et al. (2002) The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol 4:453–461

[4] Wemheuer B, Güllert S, Billerbeck S et al. (2014) Impact of a phytoplankton bloom on the diversity of the active bacterial community in the southern North Sea as revealed by metatranscriptomic approaches. FEMS Microbiol Ecol 87:378–389, doi: 10.1111/1574-6941.12230

[5] Pfeiffer B, Fender A-C, Lasota S et al. (2013) Leaf litter is the main driver for changes in bacterial community structures in the rhizosphere of ash and beech. Appl Soil Ecol 72:150–160

[6] Nacke H, Thürmer A, Wollherr A et al. (2011) Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. PLoS One 6:e17000

[7] Margulies M, Egholm M, Altman WE et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

[8] McCoy RC, Taylor RW, Blauwkamp TA et al. (2014) Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly repetitive transposable elements. bioRxiv, doi: 10.1101/001834

[9] Voskoboynik A, Neff NF, Sahoo D et al. (2013) The genome sequence of the colonial chordate, *Botryllus schlosseri*. Elife 2:e00569

Korrespondenzadresse:
Dr. Andrea Thürmer
Labor für Genomanalyse
Institut für Mikrobiologie und Genetik
Universität Göttingen
Grisebachstraße 8
D-37077 Göttingen
Tel.: 0551-39-33841
athuerm@gwdg.de
http://appmibio.uni-goettingen.de

AUTORIN

Andrea Thürmer
Jahrgang **1980**. **1999–2004** Biologiestudium an der Universität Kassel. **2008** Promotion an der Universität Göttingen, dort seit **2008** wissenschaftliche Mitarbeiterin im Labor für Genomanalyse am Institut für Mikrobiologie und Genetik.