Neue Sequenzierungstechnologien

Genomanalyse von Modellorganismen: *Saccharomyces cerevisiae*

MARKUS RALSER¹, BERND TIMMERMANN²

¹DEPT. OF BIOCHEMISTRY AND CAMBRIDGE SYSTEMS BIOLOGY CENTRE, UNIVERSITY OF CAMBRIDGE, UK

²SEQUENCING CORE FACILITY, MAX PLANCK INSTITUTE FOR MOLECULAR GENETICS (MPIMG), BERLIN

New sequencing technologies are currently revolutionizing the field of genomic research. They enable researchers to investigate genomes and transcriptomes at previously unachieved depth and level of detail. The combination of functional genomics with next generation sequencing will facilitate new experimental strategies for addressing complex biological phenomena and will give us fundamental insights in the underlying genome organization and biology of model organism.

DOI: 10.1007/s12268-012-0222-8

© Springer-Verlag 2012

■ Durch die neuen Sequenzierungsverfahren (*Next Generation Sequencing*, NGS) hat es nicht nur einen Umbruch in der humanen Genomanalyse gegeben [1], sondern auch im

Bereich der Systembiologie. Es ist nun möglich, die gesamte Genomsequenz von Modellorganismen innerhalb kürzester Zeit zu bestimmen. Hierbei ist mit den neuen Technologien sowohl die Erstentschlüsselung (*de novo*-Sequenzierung) als auch die vergleichende Analyse ganzer Genome zu einer Referenzsequenz (Re-Sequenzierung) durchführbar, womit sich völlig neue Ansätze für die funktionelle Genomforschung ergeben [2, 3].

Technologische Voraussetzungen

Am Max Planck Institute for molecular Genetics (MPIMG) werden gegenwärtig für die Gesamtgenomanalyse zwei unterschiedliche NGS-Technologien komplementär eingesetzt: einerseits die Illumina-Plattformen, die auf einer Sequenzierung durch Synthese beruhen. Hier werden in der Sequencing Core Facility des MPIMG die HiSeq 2000- und Genome Analyzer IIX-Geräte verwendet (**Abb. 1A, B**). Der Hauptvorteil dieser Technologie ist der enorme Sequenzdurchsatz. So kann mit dem HiSeq 2000-System eine Datenmenge von bis zu 600 Gigabasen pro Geräte-lauf produziert werden. Allerdings ist bei diesem System gegenwärtig die Leselänge auf 100 Basen limitiert, wobei mittels der sogenannten *paired end*-Sequenzierung auch Fragmente von beiden Seiten mit dieser Leselänge entschlüsselt werden können. Ein Geräte-lauf benötigt hier jedoch einen Zeitraum von ungefähr zehn Tagen. Die zweite verwendete technologische Plattform ist der Genome Sequencer FLX+ von Roche/454 (**Abb. 1C**). Bei diesem Gerät ist der Durchsatz zwar wesentlich geringer, jedoch können hier Leselängen von bis zu 1.000 Basen erreicht werden. Die Länge der Sequenzen unterliegt hierbei einer Verteilung, die bei den meisten Projekten einen Hauptpeak (*modal read length*) von ungefähr 800 Basen erreicht (**Abb. 2**). Der einzelne Sequenzierungs-lauf dauert hier nur 20 Stunden. Für kleinere Projekte, wie z. B. die Sequenzierung bakterieller Genome, kommt auch das Benchtop-Gerät von Roche, der GS Junior (**Abb. 1D**), zum Einsatz. Bei diesem Gerät werden zwar nur ca. 40 Megabasen

▲ **Abb. 1:** Next Generation Sequencing-Systeme in der Sequencing Core Facility am Max Planck Institute for molecular Genetics. **A**, Illumina HiSeq 2000. **B**, Illumina Genome Analyzer IIX. **C**, Roche GS FLX+. **D**, Roche GS Junior.

Hier steht eine Anzeige.

▲ **Abb. 2:** Längenverteilung der Sequenzen eines typischen Roche GS FLX+-Gerätauflaufs. **A**, Sequenzlängen unterliegen einer Verteilungskurve. Die Hauptverteilung liegt bei diesem Lauf bei 823 Basen. **B**, Die Sequenzmenge der *high quality reads* liegt bei 1,4 Millionen Sequenzen und einer Gigabase.

▲ **Abb. 3:** Filterung der detektierten Sequenzvarianten zur Identifikation der funktionell relevanten Variante. **A**, Durch Entfernen von nicht-unique Sequenzen und Abgleich zum Referenzstamm konnten 13 reale Unterschiede identifiziert werden. **B**, Vier dieser Varianten führen zu einem Aminosäureaustausch (rot).

Sequenzmenge mit einer Leselänge von 400 bis 500 Basen erzeugt, jedoch benötigt das Gerät hierfür auch nur zehn Stunden, sodass es insbesondere für diagnostische Anwendungen eine starke Verbreitung gefunden hat.

De novo-Genomsequenzierung

Für die *de novo*-Sequenzierung von Genomen werden in der Sequencing Core Facility am MPIMG diese beiden Technologien in der Regel kombiniert. Hierbei wird mithilfe der langen 454-Sequenzen eine sogenannte *backbone*-Sequenzierung durchgeführt. Es wird hier zunächst eine fünf- bis sechsfache Abdeckung der erwarteten Genomgröße angestrebt. Diese Sequenzen werden dann mit einer bedeutend größeren Menge an Illumina-

Daten erweitert. Hierbei sind die Fragmentgröße und damit der Abstand der beiden Sequenzen der *paired end*-Reaktionen entscheidend, um eine möglichst gute Abdeckung des Genoms zu erreichen. In Abhängigkeit von der Genomgröße werden zunächst Sequenzierungen mit 500 und 1.000 Basen großen Fragmenten durchgeführt. Mit diesen wird eine bis zu 50-fache Abdeckung angestrebt. Ergänzt werden hierzu noch sogenannte *mate pair*-Sequenzierungen durchgeführt. Bei diesem Verfahren kann mittels einer Zirkularisierung von großen DNA-Fragmenten ein Abstand von bis zu 20 Kilobasen zwischen den beiden komplementären Sequenzen erreicht werden. Diese großen definierten Abstände sind wichtig, um die aus den vor-

herigen Sequenzierungen mittels Assemblierung (Zusammenbau der einzelnen *reads* zu größeren Sequenzabschnitten) erzeugten Genomfragmente in der richtigen Reihenfolge auf dem Genom anzuordnen (auch als *Scaffolding* bezeichnet). Der Aufwand bei einem solchen Projekt hängt stark von der Genomgröße, jedoch auch von der Beschaffenheit der DNA-Sequenz ab. So verfügen insbesondere Pflanzengenome über einen hohen Anteil an repetitiven Elementen (sich wiederholende Sequenzmotive). Diese erschweren durch ihre schlechte Zuordenbarkeit die Assemblierung der entsprechenden Sequenzen. Generell ist neben der Sequenzierung die Datenauswertung, also im Falle der *de novo*-Sequenzierung die Assemblierung, ein entscheidender Schritt. Es können hier unterschiedlichste Programme verwendet werden, wie z. B. Celera, SOAPdenovo oder auch Programme der Gerätehersteller, wie Newbler von Roche. Zusammenfassend lässt sich hier sagen, dass die NGS-Technologien die klassische Sanger-Sequenzierung inzwischen vollständig aus dem Bereich der *de novo*-Projekte verdrängt haben [4].

Re-Sequenzierung von Modellorganismen

Gerade die Kombination der funktionellen Genomik mit den *Next Generation Sequencing*-Technologien ermöglicht es heute, neue experimentelle Strategien zu entwickeln, um so komplexe biologische Phänomene zu adressieren. So konnte kürzlich in einer Studie im Hefestamm *Saccharomyces cerevisiae* W303-K6001, der insbesondere in der Altersforschung als Screeningsystem verwendet wird, mittels NGS eine neue funktionelle Variante des Enzyms Peroxiredoxin (Tsa1p) identifiziert werden. Bei dieser Studie wurde der *S. cerevisiae*-K6001-Stamm zunächst mittels chemischer Mutagenese behandelt, und es konnte eine Mutante (K6001-B7) isoliert werden, die sich durch eine hohe Resistenz gegenüber oxidativem Stress auszeichnet. Nach der gängigen Lehrmeinung sind oxidativer Stress und die Entstehung freier Radikale ursächlich am Alterungsprozess beteiligt [5]. Im Folgenden wurde sowohl das Genom des Ausgangsstamms K6001 als auch das der K6001-B7-Variante vergleichend sequenziert. Insgesamt konnten mehr als 12.400 Abweichungen von der Referenzsequenz des Ausgangsstamms (S288c) identifiziert werden. Durch einen Abgleich der sequenzierten Variante mit dem ebenfalls analysierten Ausgangsstamm konnte diese Anzahl auf 13 Varianten

eingegrenzt werden, von denen vier zu einer *missense*-Mutation führen (**Abb. 3**). Durch funktionelle Analysen konnte aus diesen vier die ursächliche Variante für eine dominante Resistenz gegenüber oxidativem Stress identifiziert werden [6].

Diese Studie zeigt, dass die Re-Sequenzierung ganzer Genome mithilfe der neuen Technologien eine vielversprechende Alternative für die Identifizierung neuer funktioneller Allele im Bereich der klassischen molekularen Genetik ermöglicht.

Ein weiterer Punkt, der bei der vollständigen Re-Sequenzierung der Genome von Modellorganismen auffällt, ist, dass wir bisher relativ wenig über die wirkliche genetische Komposition dieser Organismen wissen. Im Falle des *Saccharomyces cerevisiae*-Stamms W303-K6001 wurde in einer Folgestudie das Genom erneut mit einer zweiten NGS-Technologie analysiert. Auch ein Abgleich dieser beiden Sequenzierungen führte zu mehr als 8.100 von der bisher bekannten Referenzsequenz abweichenden Positionen. Hierbei waren allein 799 Gene betroffen, die für Faktoren mit einer hohen Relevanz für die Physiologie und den Alterungsprozess codieren [7]. Wenn man berücksichtigt, dass gerade dieser Hefestamm ein Standardwerkzeug in der Altersforschung darstellt, ist dies ein erstaunliches Ergebnis, das deutlich macht, dass es zwingend erforderlich ist, die neuen Sequenzierungstechnologien auch zunächst für eine umfassende genetische Charakterisierung aller gegenwärtig verwendeten Modellorganismen zu verwenden. ■

Literatur

- [1] The 1000 Genomes Project Consortium (2010) A map of human genome variation from population scale sequencing. *Nature* 467:1061–1073
- [2] Burhans WC, Weinberger M (2010) Revisiting the free radical theory using next-generation sequencing technology. *Aging* 2:459–460
- [3] Gut IG (2010) 2nd generation DNA sequencing meets functional genomics. *Aging* 2:541
- [4] Kuhl H, Tine M, Beck A et al. (2011) Sequencing and annotation of several *Dicentrarchus labrax* L. chromosomes by applying Sanger- and pyro-sequencing technologies on pooled DNA of comparatively mapped BAC clones. *Genomics* 98:202–212
- [5] Finkel T, Holbrook NJ (2000) Oxidants, oxidative stress and the biology of ageing. *Nature* 408:239–247
- [6] Timmermann B, Jarolim S, Rußmayer H et al. (2010) A new dominant peroxiredoxin allele identified by whole-genome resequencing of random mutagenized yeast causes oxidant-resistance and premature aging. *Aging* 2:475–486
- [7] Ralser M, Kuhl H, Ralser M et al. (2012) The *Saccharomyces cerevisiae* W303-K6001 cross-platform genome sequence: insights into ancestry and physiology of a laboratory mutt. *Open Biol* 2:120093

Korrespondenzadresse:

Dr. Bernd Timmermann
Max Planck Institute for molecular Genetics
Sequencing Core Facility
Innstraße 63
D-14195 Berlin
Tel.: 030-8413-1542
timmermann@molgen.mpg.de

AUTOREN

Markus Ralser

1999–2003 Biologiestudium an der Universität Salzburg, Österreich. 2003–2006 Promotion am Max Planck Institute for molecular Genetics (MPIMG). 2006–2007 Postdoc in der Abteilung von Prof. Dr. H. Lehrach am MPIMG. 2007–2011 Gruppenleiter der AG Molecular Biology of Metabolism und der quantitativen Massenspektroskopie am MPIMG. Seit 2010 Gruppenleiter der AG Systembiologie an der Universität Cambridge, UK.

Bernd Timmermann

1989–1994 Biologiestudium an der Universität Osnabrück. 1995–1997 wissenschaftlicher Mitarbeiter in der AG Genomanalyse am Max-Delbrück-Centrum, Berlin. 1998–2002 Gruppenleiter DNA-Sequenzierung bei der GenProfile AG, Berlin. 2002–2007 Senior Scientist am Max-Planck-Institut für molekulare Genetik (MPIMG), Berlin. Seit 2007 Leiter der Servicegruppe Next Generation Sequencing am MPIMG. Seit 2010 Leiter der Sequencing Core Facility der Max-Planck-Gesellschaft am MPIMG.