Biochemie 2011

RNA: Deep Sequencing bedeutet die nächste Stufe der RNA-Revolution.

Epigenetik: 5-Hydroxymethylcytosin, die sechste Base im Genom, wird als neue epigenetische Modifikation diskutiert.

Deep Sequencing

◆ Im Jahr 2004 kündigte das National Institutes of Health (NIH) der USA das „1000-Dollar-Genom" als Ziel einer millionenschweren Förderinitiative an. In ihrem Verlauf sollten die Kosten für die Sequenzierung der $3 \cdot 10^9$ Basen des menschlichen Genoms von 10 Millionen auf 1000 Dollar sinken. Vor dem Hintergrund der Sequenzierung individueller Genome als Teil der medizinischen Routineversorgung wurde eine Kommerzialisierung binnen fünf Jahren anvisiert; das Erreichen der 1000-Dollar-Schwelle innerhalb von zehn Jahren galt als realistisch.

Nicht einmal acht Jahre später gibt es einen umkämpften Markt, dessen Innovationen den Umgang mit Nukleinsäuresequenzen in kurzen Abständen revolutionieren und sich als effiziente Impulsgeber der akademischen Forschung erweisen.

Während die erste Sequenzierung des menschlichen Genoms etwa 13 Jahre eines internationalen Konsortiums in Anspruch nahm, produziert ein einzelnes Gerät des Next Generation Sequencing (NGS) Daten, die in ihrem Umfang dem mehrfachen des Humangenoms entsprechen – nur noch nicht in ihrer Qualität. Hier wird die akademische Anwendung derzeit dadurch gebremst, dass es nötig ist, diese Datenfülle sinnvoll und schnell aufzubereiten; so schlägt (wieder einmal) die Stunde der Bioinformatiker.

Anwendungen des NGS haben auch die RNA-Forschung revolutioniert. Einige in diesem Rückblick aufgeführte Beispiele zeigen, wie die Technik längst die eindimensionale Erfassung simpler DNA-Sequenz hinter sich gelassen hat. Die Wissenschaft ist im Begriff, für alle traditionellen Formen der RNA-Sequenzerfassung NGS-basierte Methoden zu entwickeln. Diese könnten in der Zukunft prinzipiell nicht nur lang etablierte Standardtechniken der RNA-Biochemie, sondern auch vergleichsweise neue Chip-Technologien verdrängen. Gerade durch die ständig neu entdeckten Funktionen nicht kodierender RNAs in der Genregulation erhält die NGS immer mehr Aufwind.

In diesem Rückblick diskutieren wir zunächst die Funktionsweise der wichtigsten NGS-Techniken auf DNA-Ebene, um dann die Überführung von Sequenzinformation aus RNA in DNA zu beleuchten. Nach dieser technischen Einführung präsentieren wir einige Anwendung der jüngsten Zeit, die Meilensteine bei der Verbreitung des Next Generation Sequencing in der RNA-Welt darstellen.

High-Seq, Deep-Seq

◆ Für einfache Sequenzanalysen, wie sie etwa bei Klonierungen benötigt werden, ist das altbewährte Verfahren nach Sanger noch die Methode der Wahl, da es ohne großen technischen Aufwand mit geringen Kosten durchführbar ist. Für Genomanalysen oder für das Erstellen von Transkriptionsprofilen aller Art ist dieses Verfahren jedoch zu langsam und im Hinblick auf die zu analysierende Zahl an Sequenzen und die Sequenzlängen nicht geeignet. Hier finden Sequenzierungsmethoden der nächsten Generation ihren Platz. Dazu zählen gegenwärtig Pyrosequenzierung, die als 454-Sequencing bekannt ist, die Reversible-Terminator-Methode des Unternehmens Illumina (Illumina-Sequenzierung) sowie die Verfahren True Single Molecule Sequencing (tSMS) von Helicos und Sequencing by Oligonucleotide Ligation and Detection (Solid) von Life Technologies zur Verfügung (Abbildung 1).

Allen Methoden ist gemein, dass relativ kurze immobilisierte DNA-Fragmente in hochparallelen Ansätzen untersucht werden. Das 454-System weist – im Gegensatz zum Sanger-Verfahren – nicht den Nukleotideinbau direkt nach, sondern die bei jedem polymerasebasierten Desoxyribonukleosidtriphosphat(dNTP)-Einbau erfolgte Freisetzung von Pyrophosphat, das in einer Enzymkaskade zu einem protokollierbaren Lichtsignal umgewandelt wird. Da sich das gebildete Pyrophosphat nicht unmittelbar einer einzelnen Nukleoti-

dart zuordnen lässt, dürfen hier die Nukleotide nicht simultan als Gemisch angeboten werden, sondern einzeln zeitlich aufeinander folgend.

In der Illumina-Sequenzierung werden dNTPs ebenfalls einzeln zum Einbau angeboten. Jede Nukleotidart ist dabei mit einem spezifischen Fluoreszenzfarbstoff gekoppelt, so dass sich die beim Einbau registrierten Farbsignale den einzelnen Basen zuordnen lassen. Um jeweils nur den Einbau eines einzelnen Nukleotids pro DNA-Matrize verfolgen zu können, besitzen die eingesetzten Nukleotide keine freien 3'OH-Positionen, sondern sind dort als Terminatoren blockiert. Um einen Nukleotideinbau nach dem anderen beobachten zu können, werden nach dem jeweiligen Auslesen der Fluoreszenz sowohl Farbstoff wie auch Terminatorposition abgespalten. Ein ähnliches Vorgehen findet man beim True Single Molecule Sequencing, bei dem jedoch keine Amplifikation der zu analysierenden DNA vorangestellt wird, so dass hier tatsächlich einzelne DNA-Moleküle untersucht werden können.

Das Solid-Verfahren geht andere Wege. Diese Methode basiert nicht mehr auf der DNA-Synthese durch Polymerase, sondern verwendet Oligonukleotide, die auf Grund ihrer Komplementarität an die zu analysierende DNA angelagert und ligiert werden. Ein Fluoreszenzfarbcode erlaubt dabei die Identifi-

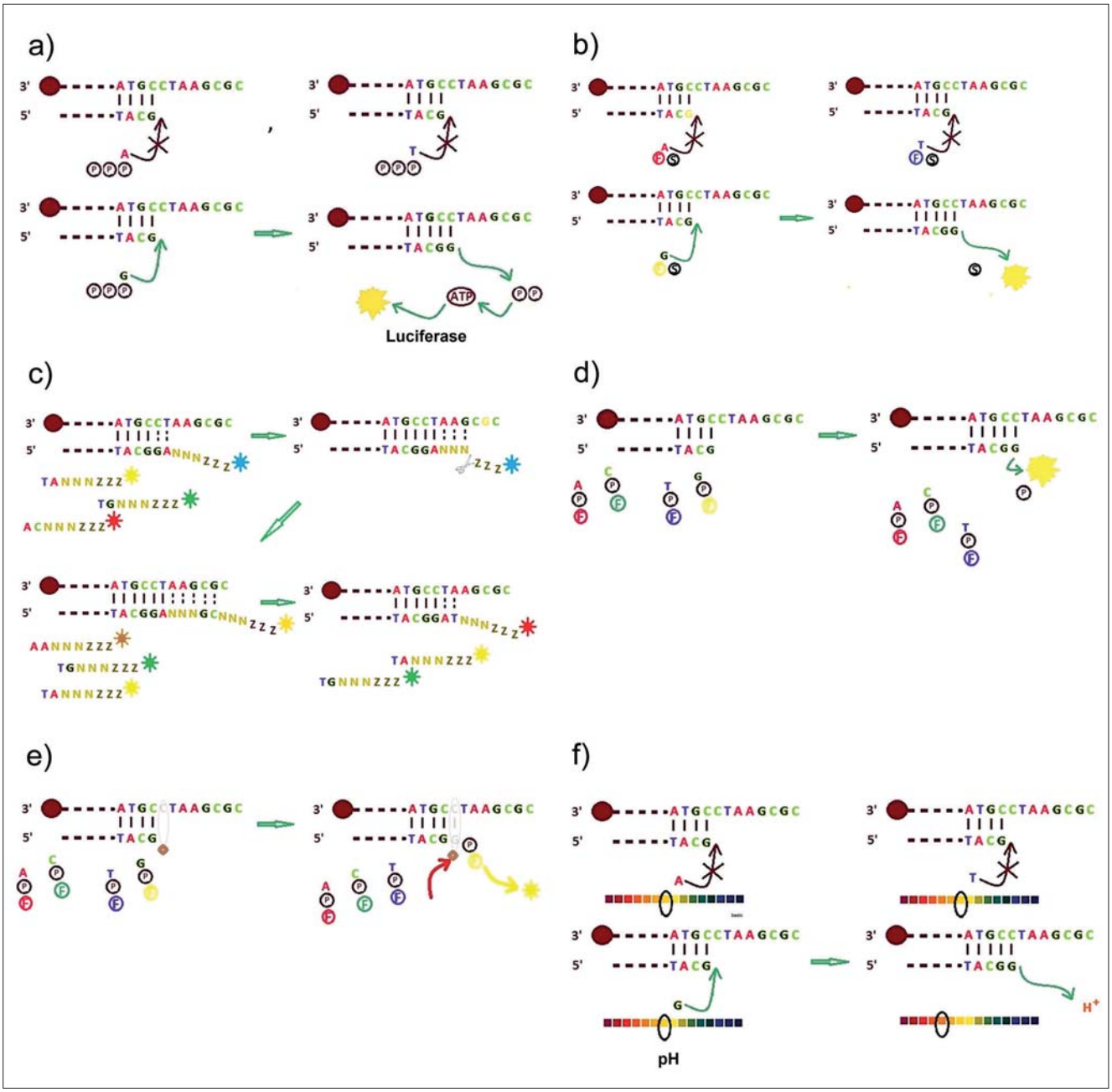

Abb. 1. Deep-Seq-Techniken. Gezeigt sind die Prinzipien der Generierung von Sequenzierdaten bei: a) 454-Pyrosequenzierung, b) Illumiina, c) Sequencing by Oligonucleotide Ligation and Detection (Solid), d) Single Molecule Realtime Sequencing, e) FRET-Sequencing und f) Semiconductor Sequencing. Die vier Nukleotide sind in jeweils charakteristischen Farben abgebildet. Gelbe Lichtblitze symbolisieren die Entstehung entsprechend detektierbarer Sequenzsignale.

zierung von jeweils zwei benachbarten Basenpositionen der einzelnen Oligonukleotide. Durch iterative Anlagerungen und Ligationen ist es in diesem relativ aufwendigen Verfahren dann möglich, die komplette Sequenzzusammensetzung der eingesetzten DNA aufzudecken.

Während diese Verfahren gegenwärtig den Stand der Technik repräsentieren und das angestrebte 1000-Dollar-Genom in Reichweite bringen, stehen die Methoden der dritten Generation bereits in den Startlöchern. Diese Verfahren arbeiten ebenfalls mit massiven Parallelansätzen, erlauben jedoch einen noch höheren Sequenzdurchsatz und damit nochmals wesentlich schnellere Genomanalysen.

Im Single-Molecule-Realtime-Sequencing-Verfahren (SMRT) von Pacific Biosciences werden dNTPs eingesetzt, die an der γ-Phosphatgruppe mit einem spezifischen Fluorophor gekoppelt sind. Eine immobilisierte DNA-Polymerase in einem Zero Mode Waveguide baut diese Nukleotide entsprechend der zu sequenzierenden DNA-Matrize in einen wachsenden DNA-Strang ein. Dabei wird die Verweildauer eines korrekt basenpaarenden dNTPs registriert, während falsche Nukleotide die dNTP-Bindungstasche der Polymerase wesentlich schneller wieder verlassen und somit nicht als Signal gewertet werden. Beim Einbau trennt die Polymerase die β- und γ-Phosphatgruppen und damit das Fluoreszenzsignal von der wachsenden DNA, dadurch lässt sich der Einbau des nächsten Nukleotids protokollieren.

Die Methode des FRET-Sequencings basiert dagegen auf dem Förster-Energie-Resonanz-Transfer (FRET) eines kovalent an die Polymerase gehefteten Quantenpunkts auf einen Fluoreszenzfarbstoff am einzubauenden Nukleotid. Auch hier ist dieser nukleotidspezifische Farbstoff an die γ-Phosphatgruppe gekoppelt, so dass er nach dem Einbau – und dem registrierten Energietransfer – wieder abgespalten wird.

Völlig ohne derartige Markierungen kommt die Halbleiter-Sequenzierung (Semiconductor Sequencing) von Ion Torrent und Life Technologies aus. Hier wird in einer Art pH-Meter die Freisetzung von Protonen während des Nukleotideinbaus durch eine immobilisierte Polymerase in einem Microwell-Array als Signal registriert. Wie bei der 454-Sequenzierung müssen hier die dNTP-Arten nacheinander appliziert werden, um die registrierten pH-Änderungen einzelnen Basen zuordnen zu können. Mit der rasanten Genomanalyse der im Sommer 2011 in Deutschland grassierenden Ehec-Keime stellte Ion Torrent die Leistungsfähigkeit dieser Methode unter Beweis.[1)]

Auch in der dritten Generation der Sequenzierungstechniken findet man eine Methode, die nicht auf Polymerisationsreaktionen beruht. Das Exonuclease-Sequencing-Verfahren von Oxford Nanopore beruht auf einer Nanopore aus Hämolysin, die in eine artifizielle Membran mit angelegtem elektrischem Potenzial eingelagert ist. Eine mit dieser Pore assoziierte Exonuklease „knabbert“ vom Ende der zu sequenzierenden DNA Nukleotid für Nukleotid ab und entlässt diese in die Pore. Auf Grund der unterschiedlichen Basen der freigesetzten Nukleotide ändert sich innerhalb der Pore die Spannung. Dies lässt sich spezifisch den einzelnen Basen zuordnen. Alternativ kann die Pore auch mit einer Polymerase gekoppelt werden, die den zu sequenzierenden DNA-Strang bei der Polymerisation durch die Pore zieht. Auch hierbei ergeben sich basenspezifische Spannungsschwankungen, die zur Bestimmung der Sequenzabfolge herangezogen werden.

In Kombination mit anderen biochemischen Techniken haben sich Anwendungen des Next Generation Sequencing in den Lebenswissenschaften unter diversen „Seq“-Kürzeln etabliert, darunter Chip-Seq zur Untersuchung von DNA-Protein-Wechselwirkungen, DNase-Seq zur Identifizierung aktiver regulatorischer Regionen, CNV-Seq, um DNA-Copy-Number-Variationen zu detektieren, oder Methyl-Seq zur Identifzierung von epigenetischen m^5C-Modifikationen in der DNA.

Von Deep-Seq zu RNA-Seq: zwischen quantitativ und selektiv

◆ Die Anwendung des Next Generation Sequencing auf die RNA-Sequenzierung (RNA-Seq) ist wohl eine der komplexesten, weil die Effizienz diverser biochemischer Transformationen auf RNA- und DNA-Ebene noch nicht quantitativ verstanden ist. Hier gab es im vergangenen Jahr Ergebnisse, die einerseits RNA-Seq zu einer zuverlässigen quantitativen Methode entwickeln könnten, anderseits den Weg weisen, um einzelne RNA-Populationen mit definierten biochemischen Eigenschaften selektiv zu erfassen.

Während einzelmolkülbasiertes Sequenzieren noch nicht die Marktreife erreicht hat, ist bei den derzeit gängigen Deep-Seq-Techniken noch ein PCR-basierter Amplifikationsschritt vor die eigentliche Datenerfassung geschaltet. Daher muss die zu sequenzierende RNA zunächst durch Reverse Transkriptase (RT) in cDNA umgeschrieben werden. Wie bei der DNA-Sequenzierung auch, wird die zu lesende Sequenz durch das Anfügen von Adapter-Oligonukleotiden an ihrem 5‘- und 3‘-Ende vorbereitet. Diese enthalten bekannte Sequenzen als Hybridisierungsanker für die Primer der anschließenden PCR. Verschiedene Varianten dieses Protokolls zeigt Abbildung 2. Am 3‘-Ende der RNA wird der Anker beispielsweise durch Ligation eines RNA-Adapters, durch Polyadenylierung oder durch enzymatisches Anfügen einer poly-C-Sequenz dargestellt. Dabei können Strukturen, die vom üblichen 3‘-OH-Ende abweichen, die Effizienz und damit die quantitative

Erfassung der jeweiligen RNA beeinträchtigen (Abbildung 2a). An dieser Stelle (Abbildung 2c) kann der 3‘-Adapter als Primersequenz zur Synthese eines ersten cDNA-Strangs durch Reverse Transkription genutzt werden. Diese Methode erfasst auch Abbruchprodukte der Reversen Transkriptase, wie sie unter anderem durch natürlich vorkommende oder chemisch generierte Nukleotidmodifikationen entstehen (Abbildung 2c).[2,3]

Weiterhin vermeidet diese – etwas missverständlich auch als 5‘-Ligations-unabhängig bezeichnete – Methode weitere biochemische Diskriminierung der zu erfassenden RNA auf Grund der chemischen Beschaffenheit des 5‘-Endes (Abbildung 2b), weil hier die anschließende Ligation eines 5‘-Adapters auf DNA-Ebene erfolgt. Anders verhält es sich, wenn beispielsweise zur Identifizierung kleiner RNAs die Ligation des 5‘-Adapters auf RNA-Ebene noch vor der cDNA-Synthese durchgeführt wird. Weil die anschließende PCR nur cDNA erfasst, welche die Erkennungssequenzen beider Adapter enthält, werden hier Abbruchprodukte der Reversen Transkription ignoriert. Ferner hat die Chemie des 5‘-Endes der RNA massiven Einfluss auf die Ligationseffizienz des 5’-Adapters: RNA-Ligasen benötigen ein 5‘-Monophosphat, um mit einem Molekül ATP

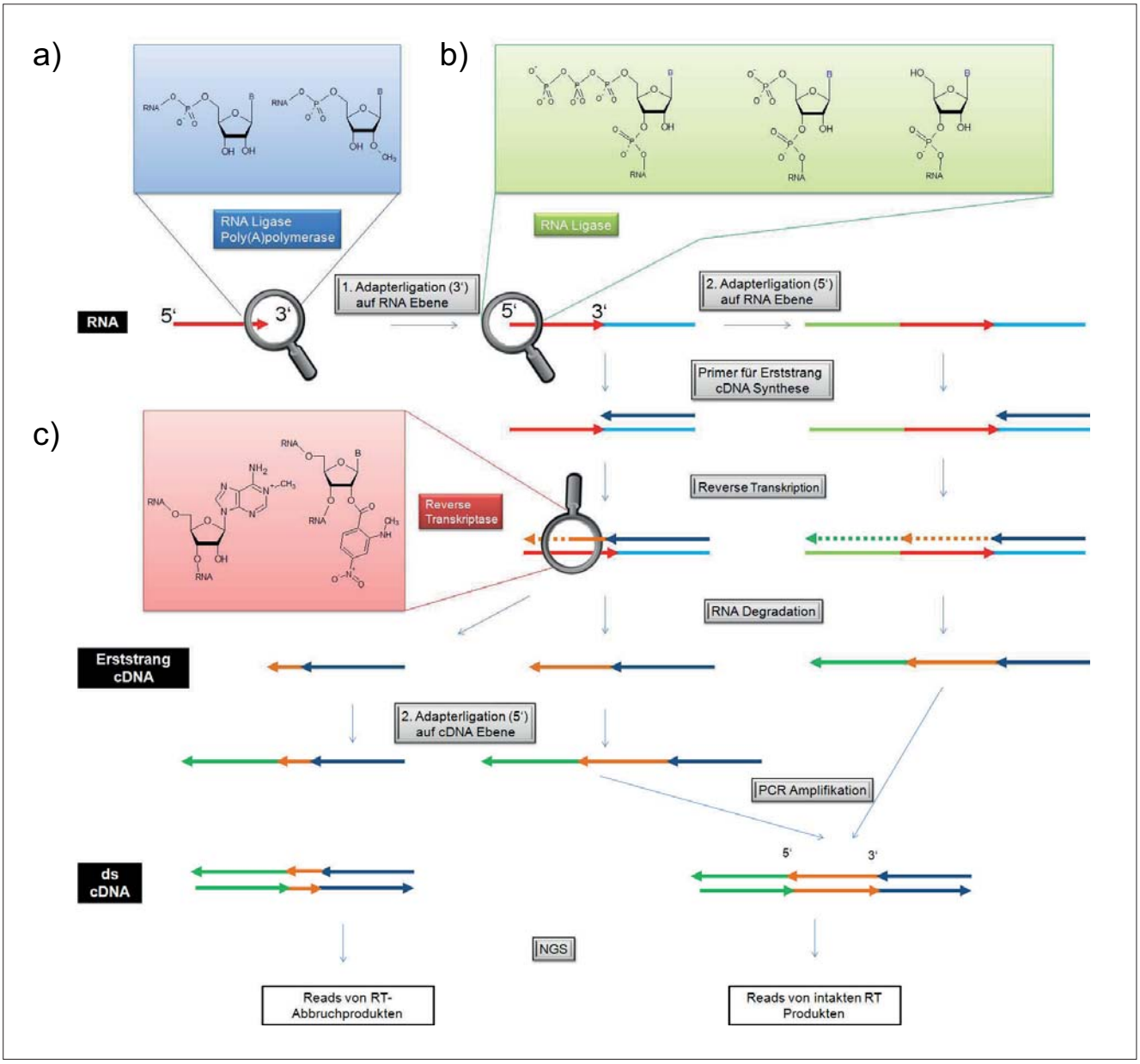

Abb. 2. Biochemische Einflüsse auf die Probenvorbereitung von RNA für RNA-Seq. Gezeigt sind zwei Varianten der Adapterligation, auf der linken Seite die 5‘-ligationsunabhängige Variante, bei welcher der entsprechende 5‘-Adapter auf der cDNA-Ebene angefügt wird. Die Ligation bleibt so von den biochemischen Details der RNA unbeeinflusst. Die Nukleotidstränge sind farbkodiert, wobei rot und orange die zu lesende Sequenz auf RNA- bzw. cDNA-Ebene darstellen. Der 3’-Adapter ist in Blautönen für RNA (hell) und Primer (dunkelblau) dargestellt. Der 5‘-Adapter ist in entsprechenden Grüntönen für RNA (hell) und cDNA/Primer (dunkelgrün) dargestellt. a) bis c): Chemische Modifikationen der zu lesenden RNA, welche die einzelnen Schritte des Protokolls beeinflussen können, sind in den entsprechenden Farben kodiert.

ein Ligationsintermediat zu generieren, das eine Cap-ähnliche Struktur aufweist. Dementsprechend werden prozessierte RNAs mit natürlichem 5'-Monophosphat bevorzugt ligiert, während Triphosphate oder 5'-Hydroxylenden nicht oder stark unterrepräsentiert erfasst werden (Abbildung 2b, S. 303).

Diese biochemischen Unterschiede und die sich daraus ergebende Ungenauigkeiten bei der quantitativen Erfassung von RNAs sind im Prinzip literaturbekannt, aber in vielen Publikationen nicht berücksichtigt. Raabe et al.[4] verglichen am Beispiel des Choleraerregers *Vibrio cholerae* Daten konventioneller Sequenzierung per Klonierung mit vorher von Liu et al.[5] publizierten RNA-Seq-Daten. Bemerkenswerterweise gab es bei zirka 200 bzw. 600 entdeckten kleinen nichtkodierenden RNAs beider Studien nur eine Schnittmenge von etwa 40 RNA-Sequenzen. Dies demonstriert die Ungenauigkeiten, unter denen beide Methoden leiden.

Um die oft implizit als quantitativ eingestuften RNA-Seq-Daten tatsächlich als solche zu nutzen, begannen im letzten Jahr manche Gruppen, einige der unbekannten Größen zu kalibrieren. In einer breit angelegten Studie zur Quantifizierung von microRNAs verglichen Hafner et al.[6] das Verhalten von einigen Hundert microRNAs und Kalibrieroligonukleotiden in einem Protokoll mit zwei RNA-basierten Adapterligationen (Abbildung 2b, S. 303). Dabei erstellten sie unter anderem Korrekturfaktoren für die prinzipiell bekannten Verzerrungen, die durch die Präferenz von RNA-Ligasen für bestimmte Nukleotidkombinationen am 3'- und 5'-Ende der zu ligierenden Fragmente entstehen. Ermutigend ist hier die Beobachtung, dass Reverse Transkription und PCR nicht mit nennenswerten Verzerrungen behaftet waren. Ebenfalls kalibriert wurden Barcodes, also Informationen in Form zusätzlicher Nukleotide in den RT-Primern, die es erlauben, die cDNA mehrerer unabhängiger Experimente in einem einzigen Sequenzierlauf zu lesen.

Während sich das Feld so einerseits in Richtung auf eine quantitative Erfassung des Transkriptoms bewegt, lassen sich biochemische Diskriminierungen durch Enzyme (Abbildung 2b, S. 303) auch nutzen, um gezielt RNA-Populationen mit unterschiedlichen Termini zu identifizieren. So haben Sharma et al. und Mitschke et al. [7,8] ein differenzielles RNA-Seq(dRNA-Seq)-Protokoll entwickelt, das auf der Tatsache beruht, dass primäre, also unprozessierte Transkripte, ein 5'-Triphosphat aufweisen. Dabei werden Sequenzdaten zweier RNA-Populationen verglichen, die sich durch die Behandlung mit einer Nuklease unterschieden, die prozessierte Transkripte abbaut, die ein 5'-Monophosphat enthalten. Mit diesem differenziellen RNA-Seq-Verfahren wurden transkriptomweit Transkriptionsstartstellen (TSS) im Krankheitserreger *Helicobacter pylori* und in einem Cyanobakterium kartiert.[7,8]

Innovative Anwendungen von RNA- Seq

◆ Abbruchprodukte der Reversen Transkription entstehen unter anderem durch starke Sekundärstruktur, aber auch durch in der Matrizen-RNA vorhandenen Nukleotidmodifikationen. Eine derartige Blockade der RT (Abbildung 2c, S. 303) beschrieben Findeiss et al. bei der Untersuchung von tRNA-Sequenzen, die in konventionellen RNA-Seq-Daten vielfach als Nebenprodukt auftauchen.[2]

Die gezielte Acylierung von 2'-OH Gruppen in nativ gefalteter RNA durch *N*-Methlyisatoinsäureanhydrid und dessen Derivate ist auch als Shape-Chemie bekannt (Shape = Selective 2'-Hydroxyl Acylation analyzed by Primer Extension). Shape ist eine besonders effiziente Variante des Chemical Probing: nur 2'-OH Gruppen von relativ unstrukturierten Nukleotiden, wie sie typischer Weise in Schlaufen und ungepaarten RNA-Bereichen vorkommen, reagieren und arretieren die anschließende Reverse Transkription (Abbildung 2c, S. 303). Kürzlich beschrieben Lucks, Aviran und Kollegen Shape-Seq, bei der sie die Shape-Chemie[3,9] mit einem RNA-Seq-Verfahren unter Einsatz der Illumina-Technik kombinierten. Dabei demonstrierten sie, wie sich eine große Datenmenge in einer Qualität erheben lässt, die es erlaubt, kleinste Mengen an RNA in einem Gemisch komplexer Sequenzen zu analysieren und eine semiautomatische Strukturmodellierung anzuschließen.

Eine weitere originelle Anwendung spielt gezielt die Stärke von RNA-Seq aus, nämlich die ansonsten unerreichte Datentiefe, mit der auch selten vorkommende Sequenzen einer RNA-Population erfasst werden. Pitt et al.[10] wandten dies auf kombinatorische RNA-Methoden, d.h. auf die Systematische Evolution von Liganden durch exponentielle Anreicherung (Selex) an. Bei Selex-Experimenten werden durch ein iteratives Verfahren von physischer Isolierung und Amplifikation die fittesten RNA-Sequenzen aus einem komplexen Pool zufälliger Sequenzen angereichert und in einem finalen Schritt typischerweise kloniert und sequenziert. Dabei war bisher die Anreicherung, die auf Bindungseigenschaften von Aptamer- oder auf katalytischer Aktivität von Ribozymsequenzen beruht, praktisch immer ein Blindflug. Die Standortbestimmung bezüglich der enthaltenen Sequenzen erfolgte erst mit der finalen Sequenzierung der selektierten RNA. Die neue Methode kann nun auch Zwischenstufen des Selex-Verfahrens mit RNA-Seq charakterisieren, um Aufschluss über den Fortgang der In-vitro-Selektion zu geben. Da dies einer Standortbestimmung im Sequenzraum (im Gegensatz zum vorherigen Blindflug) gleichkommt, wurde der Effekt auch mit dem eines

Global-Positioning-Systems (GPS) verglichen.[11)]

Bioinformatik in der NGS

◆ Die kontinuierliche Erhöhung der Durchsatzraten der Sequenzierverfahren hat eine Datenflut bisher – nicht nur in biologischen Labors – unbekannten Ausmaßes zur Folge.[12)] Ein Hochdurchsatzsequenzieransatz produziert in wenigen Stunden Rohdaten im zweistelligen Terabytebereich. Selbst bei der Löschung der Primärdaten – etwa Bilder einer CCD-Kamera oder Spannungssignale – verbleiben bei einigen Systemen mehrere Terabyte an Sequenzdaten. Schlimmer noch: Die Datenmenge vervielfacht sich im Laufe der Auswertung. Dies erfordert ein durchdachtes Datenmanagement. In Abhängigkeit von der Art der Sequenzanalyse werden Computersysteme angewandt, die nicht nur erhebliche Rechenkraft, sondern auch ausreichend große Arbeitsspeicher von bis zu mehreren Terabyte zur Verfügung stellen. Neben stationären Supercomputern und Compute-Clustern kommen immer häufiger auch Clouds zum Einsatz.[13,14)] Dazu werden die Daten, meist via Internet, auf die Systeme eines in der Regel kommerziellen Anbieters überspielt. Der Cloud-Betreiber stellt die Rechenkraft als Dienstleistung zur Verfügung. Endnutzer haben den Vorteil, keine eigene Hardwareinfrastruktur betreiben zu müssen und können den Rahmen der Dienstleistung flexibler an die Daten- und Rechenlast anpassen. ➜

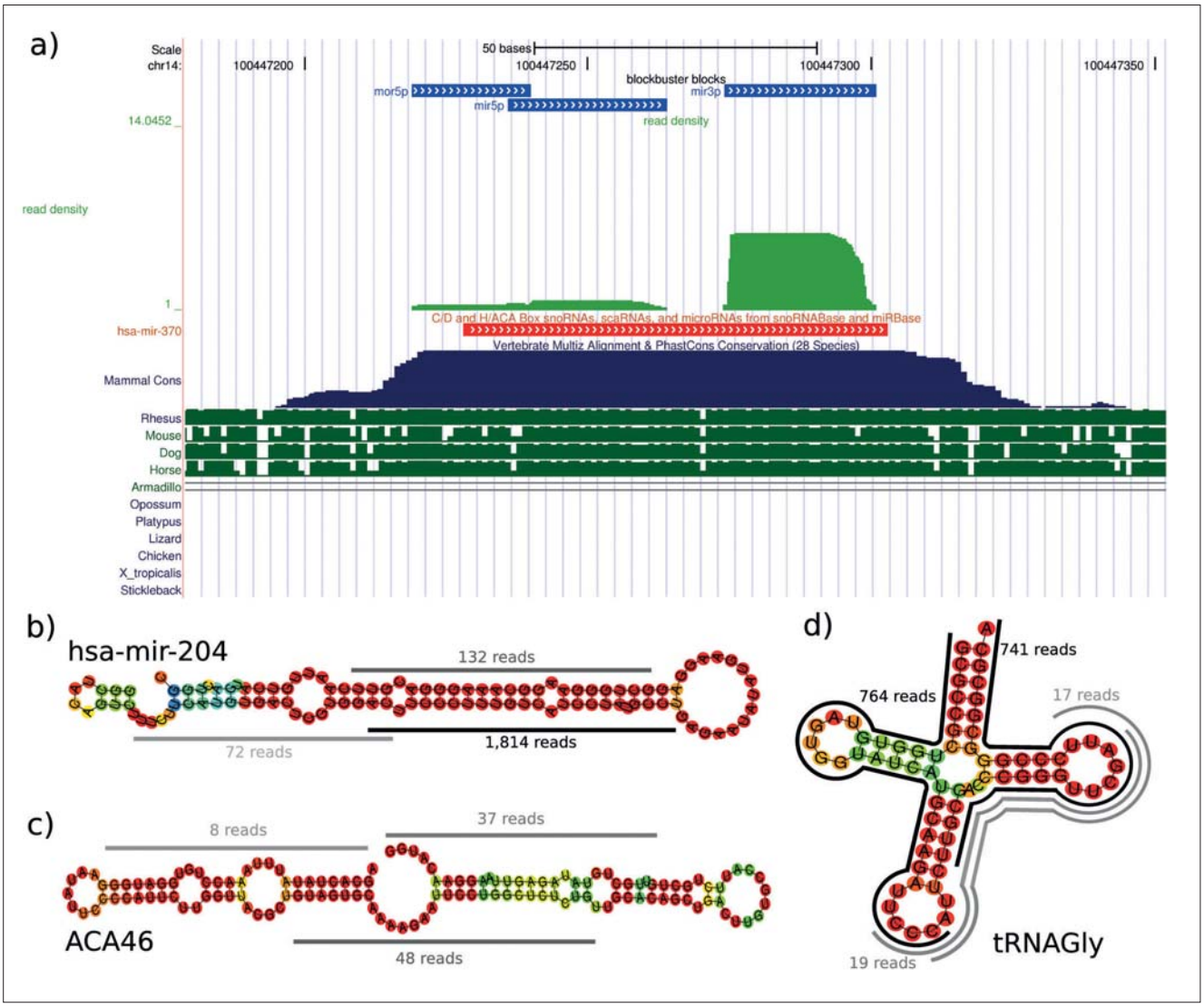

Abb. 3. Stapel von NGS-Reads sind bei vielen nicht kodierenden RNAs zu beobachten. a) Am Lokus der menschlichen microRNA 370 lassen sich drei klar getrennte Stapel erkennen. Sie korrespondieren mit drei getrennten RNA-Spezies: miR (mir3p), miR (mir5p) und mors (mor5p). Sie werden aus den microRNA Transkripten prozessiert. Unterhalb der microRNA-Annotation (roter Balken) ist die hohe evolutionäre Sequenzkonservierung des gesamten Bereiches angezeigt.*
b) Ein ähnliches Muster lässt sich, wie hier bei der hsa-mir-204, bei allen microRNAs erkennen. In dieser Darstellung sind die Read-Stapel zusammen mit der Sekundärstruktur dargestellt. Dieses Muster entsteht maßgeblich durch das Enzym Dicer, dass die doppelsträngigen microRNAs zerschneidet. In den Sekundärstrukturen zeigt die Farbe Rot eine hohe, Blau dagegen eine geringe Basenpaarungwahrscheinlichkeit an. Eine Stapelung von Reads lässt sich ebenfalls in snoRNAs (c) und tRNAs wie hier am Beispiel einer humanen Glycin tRNA (d) beobachten. Diese eindeutig nichtgleichförmige Verteilung der Reads reicht in vielen Fällen aus, um ncRNAs korrekt als tRNA, snoRNA oder miRNA zu klassifizieren.[15)]

Das Schlagwort Cloud beschreibt lediglich ein Konzept; in Abhängigkeit von der technischen Realisation kann es auch Nachteile bei der Laufzeit und der Verfügbarkeit mit sich bringen. Dies gilt insbesondere bei großen Datenmengen, zum Beispiel beim Datentransfer zwischen Prozessor, Festspeicher und Netzwerk (Eingabe-/Ausgabe-Flaschenhals).

Die rasante technische Entwicklung führt auch zu Problemen bei der Standardisierung der Datenformate. Während vor allem die Hersteller der Sequenziermaschinen viele proprietäre Datenformate entwickelt haben, konnte sich das vom Sanger Institute in Hinxton entwickelte FASTQ-Format als De-facto-Standard für Sequenzdaten durchsetzen. Das FASTQ-Format speichert neben der Nukleotidsequenz auch den Qualitätswert jeder einzelnen Base, der während der Auswertung des Sequenziersignals (basecalling) ermittelt wurde. Der Qualitätswert oder Phred-Score ist der negativ dekadische Logarithmus der Wahrscheinlichkeit für einen falschen Basecall. Im FASTQ-Format werden die Werte auf das ASCII-Alphabet abgebildet.[16] Dass Veränderungen der Wertebereiche dieser Abbildung für erhebliche Probleme bei Datenauswertung gesorgt haben, ist nur ein Beispiel für die Standardisierungsschwierigkeiten bei der Hochdurchsatzsequenzierung.

In den meisten Anwendungen schließt sich das Alignment (Mapping) der Sequenzen (Reads) auf ein oder mehrere Referenzgenome an. Neben den proprietären Mappingprogrammen gibt es eine Reihe freier Programme um Deep-seq-Daten zu alignieren.[17–19] Die Programme implementieren unterschiedliche Algorithmen, die es im Einzelfall erfordern, das richtige Programm für die gewünschte Anwendung auszuwählen. So ist eine große Zahl der Programme nicht in der Lage, mit multiplen Hits oder allzu fehlerhaften Reads umzugehen. Die meisten dieser Programme geben die Alignments im SAM-Format oder einem nach SAM konvertierbaren Datenformat aus. SAM ist ein speziell für die Hochdurchsatzsequenzierung entwickeltes Format und gibt neben der Position des Alignments im Referenzgenom, der Anzahl der Mismatches, Insertionen und Deletionen auch Informationen über die Positionen von Mate-pairs oder Paired Ends.[20] Die komprimierte Version des SAM-Formats ist das vollständig konvertierbare BAM-Format. Die meisten Programme für die auf dem Mapping aufsetzenden Analysen (SNP-calling; Expressionprofiling) akzeptieren eines dieser Formate.

Nicht zuletzt auf Grund des außerordentlichen Publikationsdrucks und dem Mangel an Bioinformatikern[21] häuft die große Publikationsgeschwindigkeit von Datensätzen, Sequenzierprotokollen und deren zielgerichteten Analysen bisher ungehobene Schätze in öffentlichen Datenbanken an. Beispielsweise gelang es im Bereich der Sequenzierung kleiner RNAs (small RNAseq), eine Reihe posttranskriptioneller Modifikationen nachzuweisen.[2] Es wurde gezeigt, dass neben den tRNAs eine Reihe von anderen RNAs nicht genomisch kodierte CCA-Enden aufweisen. Diese – vermutlich posttranskriptionellen – Additionen fanden sich unter anderem in miRNAs, aber auch in kodierenden Sequenzen. Bei der Analyse der Mapping-Ergebnisse kurzer RNAs fielen hohe, nicht uniformverteilte Fehlerraten auf. Diese ließen sich zum Teil direkt auf chemische Modifikationen zurückführen. Ein Beispiel ist das 1-Methyladenosin an der kanonischen Position 58 der tRNAs.[22,23] An dieser Position springt die Fehlerrate schlagartig von unter 5% auf bis zu 60% hoch, um danach wieder auf unter 5% zu sinken. Eine mögliche Erklärung für dieses Phänomen ist eine durch die Methylgruppe verursachte Störung der Reversen Transkriptase während der cDNA-Synthese. Nur auf der Grundlage von Mapping-Algorithmen lassen sich diese Effekte detektieren, die eine relative große Zahl von Substitutionen und vor allem auch Insertionen und Deletionen relativ zum Referenzgenom zulassen. Einfachere Ansätze würden Reads, die von chemisch modifizierten RNAs stammen, schlicht als Messfehler aussondern.

Vor allem bei der Sequenzierung kurzer RNA-Moleküle ist eine hochgradig nichtgleichförmige Verteilung der Reads festzustellen. Häufig ist dies auf posttranskriptionelle Regulationen zurückzuführen. So lassen sich in den Alignments von microRNA-Loci oft zwei klar voneinander getrennte Stapel von Reads beobachten: jeweils einer für beide Seiten des Hairpins und damit vermutlich ein direktes Resultat der Aktivität von Dicer oder anderen Nukleasen. Auch im Fall der chemischen Modifikationen von tRNAs sind Schwankungen der Read-Coverage zu beobachten: die Varianz der Read-Coverage steigt etwa in der Nähe von Methylguanosin und Dihydrouridin an. Spezialisierte Algorithmen können derartige Sprünge in den Read-Mustern detektieren und als Hinweis auf chemische Modifikationen oder RNA-Prozessierung interpretieren. Ein relativ simples Beispiel ist die Erkennung von Transkriptionsstarts aus den bereits erwähnten dRNA-seq-Datensätzen.[24]

Mark Helm, Jahrgang 1969, ist seit dem Jahr 2009 Professor für pharmazeutische/medizinische Chemie an der Universität Mainz. Er promovierte im Jahr 1999 in Molekularbiologie an der Universität Straßburg. Es folgten Postdoktorate am California Institute of Technology und an der FU Berlin. Danach war er Forschungsgruppenleiter in der Abteilung Chemie am Institut für Pharmazie und molekulare Biotechnologie der Universität Heidelberg, wo er sich im Jahr 2008 in den Fächern pharmazeutische Chemie und Biochemie habilitierte. Seine Forschungsgebiete: Ribonukleotidmodifikationen, Strukturdynamik von RNAs, Aufnahme und intrazelluläre Verteilung von siRNA, RNA-Biokonjugate.

Mario Mörl, Jahrgang 1960, studierte Biologie an der LMU München. Im Anschluss promovierte er über Ribozyme und wechselte danach als Postdoc und wissenschaftlicher Assistent zu Svante Pääbo. Im Jahr 1999 wurde er Gruppenleiter am MPI für Evolutionäre Anthropologie in Leipzig. Er habilitierte im Jahr 2002 in Genetik und ist seit 2004 Professor für Biochemie und Molekularbiologie an der Universität Leipzig. Sein Forschungsgebiet umfasst die Evolution von RNA-spezifischen Nukleotidyltransferasen, tRNA-Editing Ereignissen sowie die in vitro Evolution von RNA-Aptameren.

Peter F. Stadler, Jahrgang 1965, ist seit dem Jahr 2002 Professor für Bioinformatik an der Universität Leipzig. Er promovierte 1990 in Chemie an der Universität Wien. Nach einem Postdoktorat am Max-Planck-Institute für Biophysikalische Chemie in Göttingen ging er 1991 zurück nach Wien und habilitierte 1994 in Theoretischer Chemie. Seit 1994 ist er auch externer Professor am Santa Fe Institute in Neu Mexico. Seit 2009 wirkt er als Auswärtiges Wissenschaftliches Mitglied auch am Max-Planck-Institute für Mathematik in den Naturwissenschaften. Seine Forschungsgebiete: Bioinformatik mit Schwerpunkt auf Strukturbiologie und vergleichende Genomik, nichtkodierende RNAs, mathematische Grundlagen der Evolutionsbiologie, und diskrete Mathematik.

Steve Hoffmann, Jahrgang 1978, ist Leiter der Nachwuchsgruppe für Transcriptome Bioinformatics im LIFE-Zentrum für Zivilisationskrankheiten der Universität Leipzig. Er ist promovierter Mediziner und Bioinformatiker. Er beschäftigt sich mit dem Design von Algorithmen und Methoden zur Auswertung von Hochdurchsatzsequenzierungs-Daten.

Literatur

1) *A. Mellmann, D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S. F. McLaughlin, J. K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P. M. Brzoska, R. L. Moore, S. Guenther, J. M. Rothberg, H. Karch,* PLoS One *2011,* 6, *e22751.*
2) *S. Findeiss, D. Langenberger, P. F. Stadler, S. Hoffmann,* Biol. Chem. *2011,* 392, *305.*
3) *J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, A. P. Arkin,* Proc. Natl. Acad. Sci. USA *2011,* 108, *11063.*
4) *C. A. Raabe, C. H. Hoe, G. Randau, J. Brosius, T. H. Tang, T. S. Rozhdestvensky,* RNA *2011, 17, 1357.*
5) *J. M. Liu, J. Livny, M. S. Lawrence, M. D. Kimball, M. K. Waldor, A. Camilli,* Nucleic Acids Res. *2009,* 37, *e46.*
6) *M. Hafner, N. Renwick, M. Brown, A. Mihailovic, D. Holoch, C. Lin, J. T. Pena, J. D. Nusbaum, P. Morozov, J. Ludwig, T. Ojo, S. Luo, G. Schroth, T. Tuschl,* RNA *2011,* 17, *1697.*
7) *C. M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiss, A. Sittka, S. Chabas, K. Reiche, J. Hackermuller, R. Reinhardt, P. F. Stadler, J. Vogel,* Nature *2010,* 464, *250.*
8) *J. Mitschke, J. Georg, I. Scholz, C. M. Sharma, D. Dienst, J. Bantscheff, B. Voss, C. Steglich, A. Wilde, J. Vogel, W. R. Hess,* Proc. Natl. Acad. Sci. USA *2011,* 108, *2124.*
9) *S. Aviran, C. Trapnell, J. B. Lucks, S. A. Mortimer, S. Luo, G. P. Schroth, J. A. Doudna, A. P. Arkin, L. Pachter,* Proc. Natl. Acad. Sci. USA *2011,* 108, *11069.*
10) *J. N. Pitt, A. R. Ferre-D'Amare,* Science *2010,* 330, *376.*
11) *C. Kluwe, A. D. Ellington,* Science *2010,* 330, *330.*
12) Nat. Biotechnol. *2008,* 26, *1099.*
13) *S. V. Angiuoli, J. R. White, M. Matalka, O. White, W. F. Fricke,* PLoS One *2011,* 6, *e26624.*
14) *V. A. Fusaro, P. Patil, E. Gafni, D. P. Wall, P. J. Tonellato,* PLoS Comput. Biol. *2011,* 7, *e1002147.*
15) *D. Langenberger, C. I. Bermudez-Santana, P. F. Stadler, S. Hoffmann,* Pac. Symp. Biocomput. *2010, 80.*
16) *P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, P. M. Rice,* Nucleic Acids Res. *2010,* 38, *1767.*
17) *S. Hoffmann, C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, J. Hackermuller,* PLoS Comput. Biol. *2009,* 5, *e1000502.*
18) *H. Li, R. Durbin,* Bioinformatics *2010,* 26, *589.*
19) *B.Langmead, C. Trapnell, M. Pop, S. L. Salzberg,* Genome Biol. *2009,* 10, *R25.*
20) *H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin,* Bioinformatics *2009,* 25, *2078.*
21) *E. R. Mardis,* Genome Med. *2010,* 2, *84.*
22) *K. Iida, H. Jin, J. K. Zhu,* BMC Genomics *2009,* 10, *155.*
23) *H. A. Ebhardt, H. H. Tsang, D. C. Dai, Y. Liu, B. Bostan, R. P. Fahlman,* Nucleic Acids Res. *2009, 37, 2461.*
24) *C. Schmidtke, S. Findeiss, C. M. Sharma, J. Kuhfuss, S. Hoffmann, J. Vogel, P. F. Stadler, U. Bonas,* Nucleic Acids Res. *2011, doi: 10.1093/nar/gkr904.*