

Genexpressionsanalyse mit SuperSAGE und hochparalleler Sequenzierung

Björn Rotter, Günter Kahl, Ralf Horres, Peter Winter,
GenXPro GmbH, Frankfurter Innovationszentrum Biotechnologie (FIZ), Frankfurt am Main

SuperSAGE, eine Variante des SAGE-Verfahrens, in Kombination mit modernen Hochdurchsatz-Sequenzierungsverfahren erfasst und quantifiziert selbst sehr seltene Transkripte mittels 26 Basenpaar-Tags. Die Genauigkeit der Tag-Zuordnung mithilfe dieses Verfahrens erlaubt die simultane Genexpressionsanalyse mehrerer Organismen, zum Beispiel zur Studie von Parasit-Wirt Beziehungen. Im Gegensatz zu vielen anderen Tag-basierten Verfahren, treten bei SuperSAGE keine PCR-bedingten systematischen Messabweichungen auf.

Genexpressionsanalysen sollen Daten darüber liefern, welche Gene in einer biologischen Probe wo abgelesen werden, und wie häufig die Transkripte dieser Gene vorkommen. Grundsätzlich lassen sich „Array-“ und „Tag“-basierte Verfahren unterscheiden.

Auf Microarrays werden spezifische Oligonukleotide komplementär zu bekannten cDNA-Sequenzen einzelsträngig synthetisiert und fixiert. Die Position für jedes Gen, das somit auf der Matrix repräsentiert wird, ist bekannt. Fluoreszenz-markierte RNA oder cDNA der zu untersuchenden Probe wird mit diesem Array hybridisiert. Auf Basis der Position der emittierten Fluoreszenzsignale wird dann ermittelt, welches Gen abgelesen wurde, wobei die Signalintensität als Anhaltspunkt für die Stärke der Expression dient.

Tag-basiertes Gene Expression Profiling

Tag-basierte Verfahren zur Analyse der Genexpression beruhen darauf, von jedem Transkript ein möglichst repräsentatives Stück – den „Tag“ – zu gewinnen und möglichst viele dieser Tags zu sequenzieren und zu zählen. Tag-basierte Verfahren zeigen gegenüber vielen Array-basierten Anwendungen mehrere Vorteile. Zum einen lässt sich mit Tag-basierten Verfahren sehr viel zuverlässiger feststellen, wie häufig ein bestimmtes Transkript vorliegt. Aufgrund der genauen und sensitiven Quantifizierung der Transkripte lassen sich auch seltene Transkripte erfassen und analysieren. Zum anderen können auch unbekannte Transkripte analysiert werden, was mit Microarrays nicht möglich ist. Mit dieser als „offene Architektur“ bezeichneten Eigenschaft können deshalb auch Organismen untersucht werden, deren Genom noch nicht oder nur unzureichend analysiert ist. Da

sich mit Microarrays nur Gene wiederfinden lassen, die zuvor auf den Microarray eingebracht wurden, handelt es sich hierbei um eine „geschlossene Architektur“. Interessanterweise werden mit Tag-basierten Verfahren selbst bei vermeintlich gut untersuchten Organismen wie Maus und Mensch immer noch viele bislang unbekannte Transkripte identifiziert^{1,2}.

Genau Quantifizierung und seltene Transkripte

Die hohe Genauigkeit Tag-basierter Verfahren bei der Quantifizierung der Genexpression resultiert daraus, dass das einfache Zählen von Tags wesentlich genauer ist als die Auswertung semi-quantitativer Lichtsignale der Microarrays. Zudem können bei Arrays falsch-positive Lichtsignale von kreuz-hybridisierten (zum Beispiel sehr ähnlichen) Transkripten emittiert werden. Hierdurch kommt es zu falsch-positiven Ergebnissen. Zudem entsteht ein Hintergrundsignal, durch das hauptsächlich die Information über selten abgelesene Gene verlorengeht. Dies ist besonders gravierend, da im Normalfall 90 % bis 95 % der verschiedenen Transkripte einer Zelle in nur ein bis fünf Kopien vorkommen, also seltene Transkripte sind (Abbildung 1). Unter ihnen befinden sich Transkripte von wichtigen Rezeptor- und Signaltransduktions-Proteinen sowie Transkriptionsfaktoren, die Reaktionskaskaden auslösen und von großer Bedeutung für den Stoffwechsel sind. Gelingt die Analyse der seltenen Transkripte, können bestimmte Stoffwechselwege – zum Beispiel in frühen Tumorstadien – erkannt werden, was sie möglicherweise für die Tumordiagnostik, unter Umständen auch Prognostik qualifiziert. Die Qualität Tag-basierter Verfahren hängt von zwei Faktoren ab:

- Der Anzahl der Tags, die sequenziert werden. Je mehr Tags für eine Analyse zur Verfügung stehen, desto besser wird die Aussagekraft, und umso mehr seltene Transkripte können erfasst werden. Um etwa ein Transkript zu erfassen, das in hunderttausend Transkripten nur einmal vorkommt, müssen theoretisch hunderttausend Tags sequenziert werden. Sollen zudem noch statistisch abgesicherte Aussagen über die Menge des Transkriptes gemacht werden, so sollten etwa 5-mal so viele Tags analysiert werden.
- Der Genauigkeit der Zuordnung eines jeden Tags zum zugehörigen Transkript. Ist der Tag unspezifisch und passt zu mehreren Transkripten, ist er für eine Analyse unbrauchbar.

SAGE und Weiterentwicklungen

Das bekannteste Tag-basierte Verfahren ist die von Velculescu et al.³ vorgestellte SAGE (serial analysis of gene expression, Abb. 2). Diese und davon abgeleitete Techniken nutzen die Eigenschaft bestimmter Restriktionsenzyme, DNA entfernt von ihrer Erkennungssequenz zu schneiden. So wird von jedem zuvor an eine Matrix gebundenen cDNA-Molekül ein Tag mit diesem „Tagging-Enzym“, gewonnen. Die Stelle, an welcher der Tag entsteht, wird durch ein möglichst

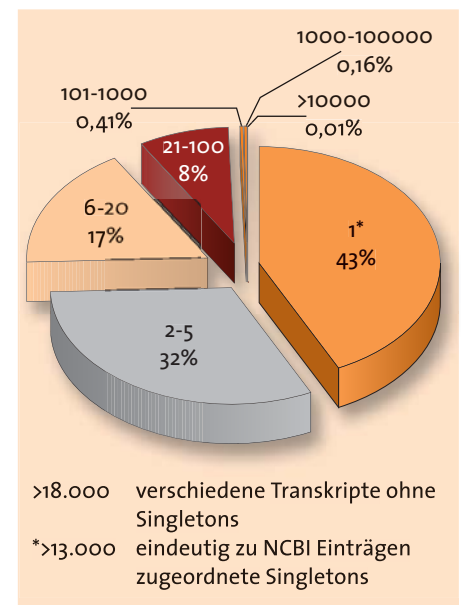
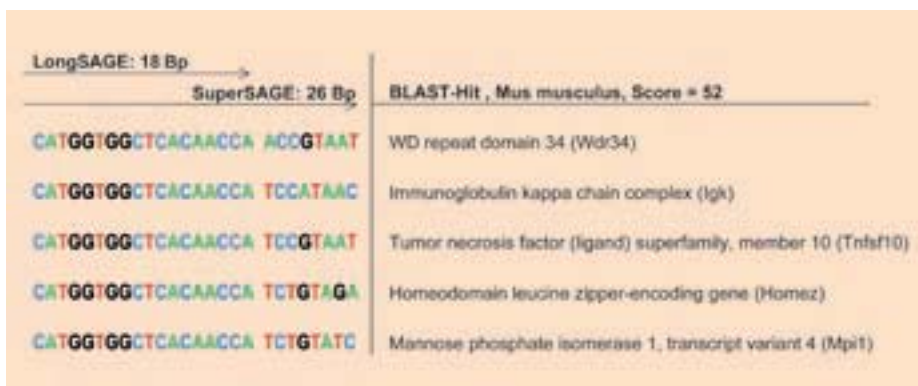


Abb. 1: SuperSAGE-Analyse der Milz von Mäusen (*Mus musculus*). Selbst bei mehr als 447.000 analysierten Tags stammen mehr als 75% von Transkripten, die in nur geringer Kopienzahl (1-5 Kopien) vorkommen. Lediglich 8% der Tags sind in 21-100 facher Kopienzahl vorhanden (Kooperationsprojekt mit Prof. Dr. Charles Dinarello, University of Colorado, Denver).

Tab. 1: Vorteil des 26 Bp langen SuperTags. Transkript-Tags, die sich in den ersten 18 Bp nicht unterscheiden und daher nicht mit LongSAGE (MmeI)-Tags identifiziert werden können, können erst durch die zusätzliche Information der SuperTags eindeutig zugeordnet werden.

Technik	Tag-Länge	e-value (NCBI BLAST)
Sage	14 bp	105
LongSage	18 bp	0,34
SuperSage	26 bp	0,00002



häufig schneidendes Enzym bestimmt, dem „Anchoring-Enzym“. Nach Verdau mit dem Anchoring-Enzym entsteht eine Schnittstelle, an welche ein Linker ligiert wird, der die Erkennungstelle für das Tagging-Enzym enthält. Bei SAGE schneidet das Tagging-Enzym Bsmfl 14 Basenpaare (Bp) entfernt von seiner künstlich erzeugten Erkennungsstelle den Tag ab, der Tag ist deshalb bei SAGE 14 Bp lang. Je zwei Tags werden daraufhin zu einem „Ditag“ ligiert und dieser mittels PCR amplifiziert. Vor der Markteinführung hochparallelierter Hochdurchsatz-Sequenzierverfahren wurden die Ditags anschließend konkatamisiert, also seriell hintereinander ligiert und kloniert. In einem Klon konnten so 30 bis 50 Tags sequenziert werden, was die Sequenzierkosten seinerzeit erheblich reduzierte.

Ditag-Sicherheit

Der Ditag erfüllt eine wichtige Funktion: Da Ditags aufgrund ihrer geringen Menge per PCR vervielfältigt werden müssen, besteht die Gefahr, dass bestimmte Ditags präferentiell amplifiziert werden, es entsteht ein systematischer Messfehler. Der Ditag bietet eine Sicherheit, um dies zu verhindern: Da statistisch jede Tag-Kombination im Ditag nur einmal vorkommt, lassen sich alle durch die PCR entstandenen Kopien von Ditags erkennen, die später im Datensatz auftauchen. Werden also bestimmte Ditags durch die PCR präferentiell kopiert, können diese „Ausreißer“ bioinformatisch eliminiert werden. Zudem haben alle Ditags die gleiche, durch die Tags definierte Größe, so dass der bekannte, Größen-basierte PCR-Messfehler verhindert wird. Tag-basierte Verfahren, die auf diese „Ditag-Sicherheit“ verzichten, wei-

sen mitunter einen starken systematischen Messfehler auf (z. B. MPSS⁴; cDNA-Enden⁵).

Da mit SAGE und davon abgeleiteten Methoden auch polyadenylierte Antisense-Transkripte erfasst werden, ist die Technik nicht allein auf Protein-kodierende mRNA beschränkt. Mit steigender Kenntnis über „non-coding“ RNAs (ncRNAs) gewinnt diese Information zunehmend an Bedeutung.

Die Nachteile der ursprünglichen SAGE-Technik, die letztlich den Microarrays den Weg bereiteten, waren einerseits die geringe Genauigkeit bei der Zuordnung der nur 14 Bp kurzen Tags (zu kurz als PCR-Sonde), andererseits die aufwendige und teure Her-

stellung, Klonierung und Sequenzierung der Tag-Konkatamere. Bei „LongSAGE“, einer von Saha et al. 2002⁶ vorgestellten, verbesserten Version, konnte der Tag mit Hilfe des Tagging-Enzyms MmeI zwar auf 18 Bp verlängert werden. Die Genauigkeit der Zuordnung des Tags zum Transkript war jedoch immer noch unzureichend. In einer deutsch-japanischen Kooperation wurde deshalb das „SuperSAGE“-Verfahren entwickelt⁷. Bei dieser neuen SAGE-Version werden die Tags mit dem Tagging-Enzym EcoP15I um weitere acht Basenpaare auf 26 Bp verlängert, es entstehen sogenannte SuperTags. Ein Vergleich der Zuverlässigkeit der Zuordnung, ausgedrückt durch den „e-value“ des NCBI-BLAST-Programms (Tabelle 1) zeigt die enorme qualitative Verbesserung durch den Einsatz des 26 Bp langen Tags. Wie aus Tabelle 2 ersichtlich, lassen sich viele Transkripte erst durch die zusätzliche Information der 26 Bp langen SuperTAGs unterscheiden. Mit SuperSAGE ist damit der erste Nachteil der ursprünglichen SAGE-Technik – die schlechte Zuordnung von Tag zum Transkript – überwunden. Der zweite große Nachteil von SAGE – die hohen Kosten und der enorme Aufwand zur Herstellung der SAGE-Bibliotheken – wird durch eine weitere technologische Neuerung aufgehoben: neue Hochdurchsatz-Sequenzierverfahren, insbesondere das PicoLiter-Pyrosequenzieren von Roche („454-Sequencing“). Mit diesen Verfahren können heute Hunderttausende bis Millionen von 25 bis 400 Bp kurzen Sequenzabschnitten parallel sequenziert werden. Werden die Ditags mit diesen Verfahren direkt sequenziert, entfällt zum einen das sehr aufwendige Konkatamerisieren und Klonieren, zum anderen reduzieren sich die Kosten der Sequenzierung um ein Vielfaches.

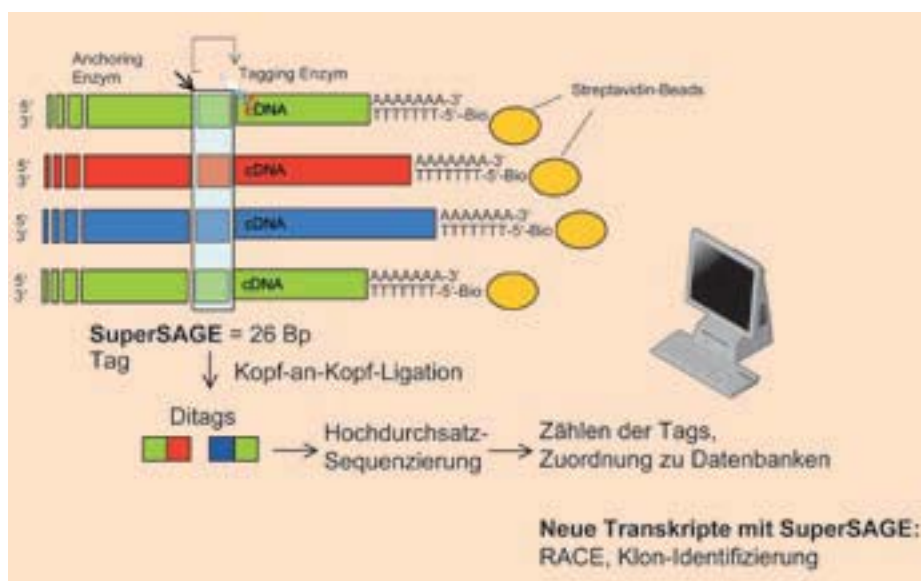


Abb.2: Schema von SAGE und Varianten. Ein spezifischer Tag jedes Transkripts wird enzymatisch gewonnen, je zwei der Tags zu einem Ditag ligiert. Diese Ditags können heute mit Hochdurchsatzsequenzierern kostengünstig und in großer Anzahl sequenziert werden. Bei SuperSAGE werden längere Tags von 26 Bp Länge analysiert.

Tab. 2: Gene Expression Profiling mittels SuperSAGE im Vergleich zu Microarrays

Microarrays (geschlossene Architektur)	SuperSAGE (offene Architektur)
Analyse beschränkt auf bekannte Transkripte; neue Transkripte können nicht entdeckt werden	Bekannte und unbekannte Transkripte können analysiert werden
seltene Transkripte werden schlecht oder nicht erfasst	seltene Transkripte werden genau quantifiziert
Falsch-positive Signale durch Kreuz-Hybridisierung	verlässliche Daten auf Basis von 26 Bp-Tags
semiquantitative Lichtsignale	genaue Quantifizierung durch Zählen der Tags
starke Umwelteinflüsse (Ozon etc.) beeinträchtigen Datenanalyse	kein Umwelteinfluss auf Datenanalyse

Genexpressionsanalyse mit SuperSAGE/454Sequencing

Keine der bisherigen Hochdurchsatz-Sequenzierungsmethoden zur Tag-Analyse kommt ohne PCR aus, und auch die Emulsion-PCR der 454-Sequenziermethode ist sehr anfällig für einen PCR-basierten Fehler. Deshalb ist die genaue Quantifizierung der Transkripte von Hochdurchsatz-sequenzierten DNA-Fragmenten unterschiedlicher Länge und Eigenschaften („Monotags“, c-DNA-Enden) ohne die zuvor beschriebene Ditag-Sicherheit nicht möglich^{4,5}. SuperSAGE vereint deshalb als einziges Verfahren eine Messabweichungs-freie Amplifikation von Ditags gleicher Größe mit genauer Annotation der Tags und erreicht mit moderner Hochdurchsatz-Sequenzierung eine höhere Genauigkeit der Transkriptions-Analyse. So ermöglicht SuperSAGE erstmals zu vertretbaren Kosten hochauflösende Genexpressionsanalysen, wobei seltene Transkripte erfasst und neue Transkripte identifiziert werden können.

Die neuen Hochdurchsatz-Sequenzierungsmöglichkeiten erlauben außerdem die kostengünstige Analyse normalisierter cDNA, etwa von nicht-Modell-Organismen. Die quantitative Analyse der SuperTAGs kann somit um die qualitative Analyse der Vollängen-Transkripte ergänzt werden.

Entdeckung neuer Gene

Darüber hinaus können die mit SuperSAGE neu entdeckten Transkripte sehr einfach genauer analysiert werden, da sich von den 26 Bp langen Tags hochspezifische PCR-Primer ableiten lassen. So kann beispielsweise mittels 3'- und 5'-RACE ein zuvor unbekanntes Vollängen-Transkript schnell und zuverlässig ermittelt werden. Zur Identifikation von Klonen in einer Gen- oder cDNA-Bank können SuperSAGE-Tags zudem als spezifische Sonden eingesetzt werden.

Mit SuperSAGE lassen sich zudem die Transkriptome mehrerer Organismen in ihrem Wechselspiel untersuchen⁷. Aufgrund der Transkript-Spezifität der 26 Bp-Tags können die Transkripte eindeutig den verschiedenen Organismen zugeordnet werden. So können etwa Wechselwirkungen zwischen einem

Parasit oder Pathogen und dessen Wirt auf Transkriptom-Ebene sehr genau erfasst werden. Damit gehören die Isolation und separate Untersuchung der Transkriptionsmuster von Pathogenen und ihren Wirten der Vergangenheit an.

Um sich den einzigen Vorteil der Microarrays, die geringeren Analysenkosten, zunutze zu machen, können die 26 Bp-SuperTAGs für Hochdurchsatz-Analysen auch direkt auf einen Microarray gespottet werden und so in kürzester Zeit für jeden Organismus ein Expressions-Microarray hergestellt werden⁸.

Der Vergleich von Genexpressionsanalysen mittels SuperSAGE und Microarrays (Tabelle 2) zeigt zahlreiche Vorteile von SuperSAGE gegenüber Microarrays. Für Studien, die eine zuverlässige Quantifizierung aller Transkripte – inklusive neuer, seltener und non-coding-Transkripte – erfordern, ist SuperSAGE deshalb die Technologie der Wahl.

Literatur

- [1] Richards M, Tan S-P, Chan W-K, Bongso A. *Stem Cells*. 24 (2006), 1162-73.
- [2] Zhao Y, Raouf A, Kent D, Khattra J, Delaney A, Schnerch A, Asano J, McDonald H, Chan C, Jones S, Marra MA, Eaves CJ. *Stem Cells*. 25 (2007), 1681-1689.
- [3] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. *Science*. 270(5235) (1995), 484-7.
- [4] Hene L, Sreenu VB, Vuong MT, Abidi SH, Sutton JK, Rowland-Jones SL, Davis SJ, Evans EJ. *BMC Genomics*; 8 (2007), 333.
- [5] Tatiana Teixeira Torres, Muralidhar Metta, Birgit Ottenwälder and Christian Schlötterer *Genome Res.* 18 (2008), 172-177.
- [6] Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. *U Nat Biotechnol*. 20(5) (2002), 508-12.
- [7] Matsumura H, Reich S, Ito A, Saitoh H, Kamoun S, Winter P, Kahl G, Reuter M, Kruger DH, Terauchi R. *Proc Natl Acad Sci* 100(26) (2003), 15718-23
- [8] Matsumura H, Bin Nasir KH, Yoshida K, Ito A, Kahl G, Krüger DH, Terauchi R. *Nature Methods*. 3(6) (2006), 469-474.

Korrespondenzadresse

Dr. Björn Rotter
GenXPro GmbH
Frankfurter Innovationszentrum
Biotechnologie (FIZ)
Altenhöferallee 3
60438 Frankfurt am Main
www.genxpro.de

Introducing something totally new to DNA/RNA and Protein purification:


Ease.

PrepEase™

Kits for:

- Plasmid DNA Purification
- BAC Purification
- RNA Purification
- Gel Extraction
- Sequencing Dye Clean-Up
- His-Tagged Protein Extraction

Easy Familiar Methods
Easy to Understand Protocols



Fueling Innovation in Life Science™

Now in Europe

USB Europe GmbH
Hauptstrasse 1
79219 Staufen
Germany

T: +49 (0)76 33-933 40 0
F: +49 (0)76 33-933 40 20
www.usbweb.com
custserv@usbweb.de