(19) **Bundesrepublik Deutschland**
Deutsches Patent- und Markenamt

(10) **DE 697 37 450 T2** 2007.11.29

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 0 923 650 B1**
(21) Deutsches Aktenzeichen: **697 37 450.5**
(86) PCT-Aktenzeichen: **PCT/US97/09472**
(96) Europäisches Aktenzeichen: **97 929 757.9**
(87) PCT-Veröffentlichungs-Nr.: **WO 1997/046704**
(86) PCT-Anmeldetag: **02.06.1997**
(87) Veröffentlichungstag der PCT-Anmeldung: **11.12.1997**
(97) Erstveröffentlichung durch das EPA: **23.06.1999**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **07.03.2007**
(47) Veröffentlichungstag im Patentblatt: **29.11.2007**

(51) Int Cl.[8]: ***C12Q 1/68*** (2006.01)

(30) Unionspriorität:

659453	**06.06.1996**	**US**
689587	**12.08.1996**	**US**

(84) Benannte Vertragsstaaten:
AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LI, LU, MC, NL, PT, SE

(73) Patentinhaber:
Solexa, Inc., Hayward, Calif., US

(74) Vertreter:
Vossius & Partner, 81675 München

(72) Erfinder:
ALBRECHT, Glenn, Redwood City, CA 94061, US;
BRENNER, Sydney, Cambridge CB2 3PJ, GB;
LLOYD, David H., Daly City, CA 94014, US;
DUBRIDGE, Robert B., Belmont, CA 94002, US;
PALLAS, Michael C., San Bruno, CA 94066, US

(54) Bezeichnung: **SEQUENZIERUNG DURCH LIGATION KODIERTER ADAPTER**

Beschreibung

Gebiet der Erfindung

[0001] Die Erfindung betrifft im Allgemeinen Verfahren zur Bestimmung der Nucleotidsequenz eines Polynucleotids und genauer ein Verfahren zur Identifizierung endständiger Nucleotide eines Polynucleotids durch spezifische Ligierung von codierten Adaptoren.

HINTERGRUND

[0002] Die DNA-Sequenzierungsverfahren der Wahl für nahezu alle wissenschaftlichen und kommerziellen Anwendungen basieren auf dem durch Sanger, zum Beispiel Sanger et al., Proc. Natl. Acad. Sci., 74 (1977): 5463–5467, in Pioneerarbeit etablierten Didesoxy-Ketten-Abbruch-Ansatz. Das Verfahren wurde in verschiedenen Arten verbessert und wird in einer Vielfalt von Formen in allen kommerziellen DNA-Sequenzier-Geräten verwandt, zum Beispiel Hunkapiller et al., Science, 254 (1991): 59–67.

[0003] Das Ketten-Abbruch-Verfahren erfordert die Erzeugung von einem oder mehreren Sätzen von markierten DNA-Fragmenten, wobei jedes einen gemeinsamen Ursprung hat und jedes mit einer bekannten Base endet. Der Satz oder die Sätze von Fragmenten müssen dann der Größe nach aufgetrennt werden, um die Sequenzinformation zu erhalten. Die Größenauftrennung wird für gewöhnlich durch hochauflösende Gelelektrophorese bewerkstelligt, welche die Leistungsfähigkeit besitzen muss, sehr große Fragmente, die sich in der Größe durch nicht mehr als ein einzelnes Nucleotid unterscheiden, zu unterscheiden. Trotz vieler wesentlicher Verbesserungen wie Auftrennungen in Anordnungen von Kapillaren und der Verwendung von elektrophoretischen Auftrennungsmedien, die nicht aus Gel bestehen, bietet sich das Verfahren nicht einfach für eine Miniaturisierung oder für massive parallele Anwendung an.

[0004] Als eine Alternative zu den auf Sanger basierenden Ansätzen zur DNA-Sequenzierung wurden einige so genannte "Base-um-Base" oder "Einzelbasen"-Sequenzieransätze untersucht, zum Beispiel Cheeseman, US-Patent 5,302,509; Tsien et al., internationale Anmeldung WO 91/06678; Rosenthal et al., internationale Anmeldung WO 93/21340; Canard et al., Gene, 148 (1994): 1–6; und Metzker et al., Nucleic Acids Research, 22 (1994): 4259–4267. Diese Ansätze sind durch die Bestimmung eines einzelnen Nucleotids pro Zyklus bei einer chemischen oder biochemischen Durchführung und dem Fehlen der Notwendigkeit eines Auftrennungsschrittes charakterisiert. Somit versprechen, wenn sie wie vorgesehen durchgeführt werden könnten, "Base-um-Base"-Ansätze die Möglichkeit, viele tausende Sequenzreaktionen parallel zum Beispiel auf an Mikropartikel oder an Festphasenanordnungen gebundenen Zielpolynucleotiden durchzuführen, zum Beispiel internationale Patentanmeldung PCT/US95/12678 (WO 96/12039).

[0005] WO 95/20053 offenbart ein weiteres Verfahren der Sequenzierung von Nucleinsäuren. Ein solches Verfahren verwendet die spezifische Hybridisierung von markierten Adaptoren, worin jede der vier Markierungen einem unterschiedlichen Nucleotid entspricht, gefolgt durch einen Abspaltungsschritt, welcher das Ziel um ein Nucleotid kürzt. In diesem Verfahren ist jede Runde der Ligierung und Abspaltung wirksam, um ein Nucleotid im Ziel zu identifizieren.

[0006] WO 96/12014 offenbart die Verwendung von Oligonucleotid-Tags, ausgewählt aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden, zur Verfolgung, Identifizierung und/oder Sortierung von Populationen von Molekülen. Das Verfahren kann verwendet werden, um eine Population von Polynucleotiden auf einem festen Träger zur simultanen Sequenzierung zu sortieren. Eine "Einzelbasen-Sequenzierverfahrensweise", welche wiederholte Schritte der Ligierung von markierten Sonden, der Identifizierung und Abspaltung mit einer Nuclease einschließt, wird für die Sequenzierung beschrieben.

[0007] EP-A2 0 630 972 offenbart ein weiteres Verfahren zur Sequenzierung von DNA. Ein Polynucleotid wird mit einer Restriktionsendonuclease gespalten, was Fragmente erzeugt, welche identische gespaltene Enden besitzen. Ein markierter Adaptor, der einen überhängenden Strang besitzt, welcher komplementär zu diesen gespaltenen Enden ist, wird dann an die Fragmente ligiert. Die Spaltung mit einer Exonuclease macht die End-Regionen des sich ergebenden Konstruktes einschließlich Abschnitte des Fragmentes einzelsträngig. Die Fragmente werden dann entsprechend dem Unterschied der Endsequenzen, die auf die ligierten bekannten Oligomersequenz folgen, aufgetrennt. Dies wird durch Hybridisierung der Konstrukte an Sonden auf einem festen Träger bewerkstelligt, von welchen jede einen Abschnitt der bekannten Adaptorsequenz und eine variable Sequenz einschließt.

[0008] Unglücklicherweise hatten "Base-um-Base"-Sequenzierungsschemata aufgrund vieler Probleme wie ineffizienter Chemien, welche die Bestimmung von mehr als ein paar Nucleotiden in einer kompletten Sequenzierungsoperation verhindern, keine weit verbreitete Anwendung. Mehr noch entstehen in Base-um-Base-Ansätzen, welche enzymatische Manipulationen erfordern, weitere Probleme mit Geräten, welche für die automatische Verarbeitung verwendet werden. Wenn eine Serie von enzymatischen Schritten in Reaktionskammern durchgeführt wird, welche hohe Oberfläche-zu-Volumen-Verhältnisse und enge Kanaldimensionen haben, können Enzyme an Oberflächenkomponenten haften, was Waschschritte und nachfolgende Verarbeitungsschritte sehr schwierig macht. Die Ansammlung von Protein beeinflusst auch molekulare Reportersysteme, insbesondere jene, welche fluoreszierende Markierungen verwenden, und macht die Interpretation von Messungen basierend auf solchen Systemen schwierig und ungünstig. Diese und ähnliche Schwierigkeiten haben die Anwendung von "Base-um-Base"-Sequenzierungsschemata in Versuchen der parallelen Sequenzierung wesentlich verlangsamt.

[0009] Es könnte ein wichtiger Fortschritt in der Base-um-Base-Sequenzierungstechnologie insbesondere bei automatisierten Systemen gemacht werden, wenn ein alternativer Ansatz zur Bestimmung der endständigen Nucleotide von Polynucleotiden verfügbar wäre, welcher sich wiederholende Prozessierungszyklen unter Verwendung vieler Enzyme minimiert oder eliminiert.

Zusammenfassung der Erfindung

[0010] Entsprechend ist ein Gegenstand unserer Erfindung ein DNA-Sequenzierungsschema bereitzustellen, welches nicht unter den Nachteilen der aktuellen Base-um-Base-Ansätze leidet.

[0011] Ein anderer Gegenstand unserer Erfindung ist es ein Verfahren der DNA-Sequenzierung bereitzustellen, welches zur parallelen oder gleichzeitigen Anwendung an tausenden von DNA-Fragmenten, die in einem gemeinsamen Reaktionsgefäß vorhanden sind, geeignet ist.

[0012] Ein weiteres erfindungsgemäßes Ziel ist es, ein Verfahren zur DNA-Sequenzierung bereitzustellen, welches die Identifizierung eines endständigen Abschnittes eines Ziel-Polynucleotids mit minimalen enzymatischen Schritten erlaubt.

[0013] Noch ein weiterer Gegenstand unserer Erfindung ist es einen Satz von codierten Adaptoren zur Identifizierung der Sequenz einer Vielzahl von endständigen Nucleotiden von einem oder mehreren Ziel-Polynucleotiden bereitzustellen.

[0014] Unsere Erfindung stellt diese und andere Gegenstände durch Bereitstellung eines Verfahrens der Sequenzanalyse von Nucleinsäuren basierend auf der Ligierung von einem oder mehreren Sätzen von codierten Adaptoren an ein Ende eines Ziel-Polynucleotids (oder an die Enden von vielen Ziel-Polynucleotiden, wenn in einer parallelen Sequenzierungsoperation angewendet) bereit. Jeder codierte Adaptor umfasst einen überhängenden Strang und einen Oligonucleotid-Tag, ausgewählt aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden.

[0015] Codierte Adaptoren, deren überhängende Stränge perfekt passende Duplexmoleküle mit den komplementären überhängenden Strängen des Ziel-Polynucleotids bilden, werden ligiert. Nach der Ligierung werden die Identität und die Anordnung der Nucleotide in den überhängenden Strängen durch spezifisches Hybridisieren eines markierten Tagkomplements an seinen korrespondierenden Tag auf dem ligierten Adaptor bestimmt oder "decodiert".

[0016] Wenn zum Beispiel ein codierter Adaptor mit einem überhängenden Strang von vier Nucleotiden, zum Beispiel 5'-AGGT, ein perfekt passendes Duplexmolekül mit einem komplementären überhängenden Strang eines Ziel-Polynucleotids bildet und ligiert wird, können die vier komplementären Nucleotide, 3'-TCCA, auf dem Polynucleotid durch einen einzigartigen Oligonucleotid-Tag, ausgewählt aus einem Satz von 256 solcher Tags, einen für jede mögliche vier-Nucleotidsequenz der überhängenden Stränge, identifiziert werden. Tagkomplemente werden unter Bedingungen an den ligierten Adaptoren angewendet, welche eine spezifische Hybridisierung nur für jene Tagkomplemente erlauben, welche perfekt passende Duplexmoleküle (oder Triplexmoleküle) mit den Oligonucleotid-Tags der ligierten Adaptoren bilden. Die Tagkomplemente können einzeln oder als ein Gemisch oder mehrere Gemische angewendet werden, um die Identität der Oligonucleotid-Tags und damit der Sequenzen der überhängenden Stränge zu bestimmen.

[0017] Wie nachstehend vollständiger erklärt wird, können die codierten Adaptoren in einer Sequenzanalyse

entweder (i) zur Identifizierung von einem oder mehreren Nucleotiden als ein Schritt eines Verfahrens, das wiederholte Zyklen von Ligierung, Identifizierung und Abspaltung, wie in Brenner, US-Patent 5,599,675 und PCT-Veröffentlichungsnr. WO 95/27080 beschrieben, oder (ii) als "eigenständiges" Identifikationsverfahren, worin Sätze codierter Adaptoren an Ziel-Polynucleotiden angewendet werden, so dass jeder Satz fähig ist, die Nucleotidsequenz eines unterschiedlichen Abschnitts eines Ziel-Polynucleotids zu identifizieren, verwendet werden; das heißt, dass in der letzteren Ausführungsform die Sequenzanalyse mit einer einzigen Ligierung für jeden Satz, gefolgt durch eine Identifizierung, durchgeführt wird.

[0018] Eine wichtige Eigenschaft der codierten Adaptoren ist die Verwendung von Oligonucleotid-Tags, die Mitglieder eines minimal kreuz-hybridisierenden Satzes von Oligonucleotiden, wie zum Beispiel beschrieben in den internationalen Patentanmeldungen PCT/US 95/12791 (WO 96/12014) und PCT/US 96/09513 (WO 96/141011), sind. Die Sequenzen von Oligonucleotiden eines solchen Satzes unterscheiden sich von den Sequenzen von jedem anderen Mitglied des gleichen Satzes durch wenigstens zwei Nucleotide. Somit kann jedes Mitglied eines solchen Satzes kein Duplexmolekül (oder Triplexmolekül) mit weniger als zwei Fehlpaarungen mit dem Komplement eines jeden anderen Mitglieds bilden. Bevorzugt unterscheidet sich jedes Mitglied eines minimal kreuz-hybridisierenden Satzes von jedem anderen Mitglied durch so viele Nucleotide wie möglich, soweit dies mit der Größe eines für eine bestimmte Anwendung benötigten Satzes vereinbar ist. Zum Beispiel ist, wo längere Oligonucleotid-Tags verwendet werden, wie 12- bis 20-mere zur Anbringung von Markierungen an codierte Adaptoren, der Unterschied zwischen Mitgliedern eines minimal kreuz-hybridisierenden Satzes dann bevorzugt signifikant größer als zwei. Bevorzugt unterscheidet sich jedes Mitglied solch eines Satzes von jedem anderen Mitglied durch wenigstens vier Nucleotide. Stärker bevorzugt unterscheidet sich jedes Mitglied eines solchen Satzes von jedem anderen Mitglied durch wenigstens sechs Nucleotide. Komplemente von erfindungsgemäßen Oligonucleotid-Tags werden hier als "Tagkomplemente" bezeichnet.

[0019] Oligonucleotid-Tags können einzelsträngig sein und können zur spezifischen Hybridisierung an einzelsträngige Tagkomplemente durch Duplex-Bildung ausgelegt sein. Oligonucleotid-Tags können auch doppelsträngig sein und zur spezifischen Hybridisierung an einzelsträngige Tagkomplemente durch die Triplex-Bildung ausgelegt sein. Bevorzugt sind die Oligonucleotid-Tags der codierten Adaptoren doppelsträngig und ihre Tagkomplemente sind einzelsträngig, so dass die spezifische Hybridisierung eines Tags mit seinem Komplement durch die Bildung einer Triplex-Struktur erfolgt.

[0020] Bevorzugt umfasst das erfindungsgemäße Verfahren die folgenden Schritte:
(a) Ligieren eines codierten Adaptors an das Ende eines Polynucleotids, wobei der Adaptor einen Oligonucleotid-Tag besitzt, der aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden ausgewählt ist, und einen überhängenden Strang besitzt, der komplementär zu einem überhängenden Strang des Polynucleotids ist; und (b) Identifizieren eines oder mehrerer Nucleotide in dem überhängenden Strang des Polynucleotids durch spezifisches Hybridisieren eines Tagkomplements an den Oligonucleotid-Tag des codierten Adaptors.

Kurze Beschreibung der Zeichnungen

[0021] Die Fig. 1A–Fig. 1E stellen diagrammartig die Verwendung von codierten Adaptoren zur Bestimmung der endständigen Nucleotidsequenzen einer Vielzahl von mit einem Tag versehenen Polynucleotiden dar.

[0022] Die Fig. 2 stellt das Phänomen der Selbst-Ligierung von identischen Polynucleotiden dar, welche an einem Festphasenträger verankert sind.

[0023] Die Fig. 3A stellt Schritte in einem bevorzugten erfindungsgemäßen Verfahren dar, in welchem ein doppelsträngiger Adaptor, der einen blockierten 3'-Kohlenstoff besitzt, an ein Ziel-Polynucleotid ligiert wird.

[0024] Die Fig. 3B stellt die Verwendung der bevorzugten Ausführungsform in einem Verfahren der DNA-Sequenzierung durch schrittweise Zyklen von Ligierung und Abspaltung dar.

[0025] Die Fig. 4 stellt Daten von der Bestimmung der endständigen Nucleotide eines Test-Polynucleotids unter Verwendung des erfindungsgemäßen Verfahrens dar.

[0026] Die Fig. 5 ist eine schematische Darstellung einer Flusskammer und eines Nachweisgerätes zur Beobachtung einer planaren Anordnung von Mikropartikeln, die mit cDNA-Molekülen zur Sequenzierung beladen sind.

Definitionen

[0027] Wie hier verwendet, wird der Begriff „codierter Adaptor" mit dem Begriff „codierte Sonde" des Prioritätsdokuments US-Patentanmeldung Seriennummer 08/689,587 synonym verwendet.

[0028] Wie hier verwendet, bedeutet der Begriff „Ligierung" die Bildung einer kovalenten Bindung zwischen den Enden von einem oder mehreren (für gewöhnlich zwei) Oligonucleotiden. Der Begriff bezieht sich für gewöhnlich auf die Bildung einer Phosphodiesterbindung welche sich aus der folgenden Reaktion ergibt, die für gewöhnlich durch eine Ligase katalysiert wird: $Oligo_1$(5')-OP(O-)(=O)O + HO-(3')$Oligo_2$-5' → $Oligo_1$(5')-OP(O-)(=O)O-(3')$Oligo_2$-5' worin $Oligo_1$ und $Oligo_2$ entweder zwei verschiedene Oligonucleotide oder verschiedene Enden des gleichen Oligonucleotids sind. Der Begriff umfasst die nichtenzymatische Bildung von Phosphodiester-Bindungen genauso wie die Bildung von kovalenten nicht-Phosphodiester-Bindungen zwischen den Enden von Oligonucleotiden wie Phosphorthioat-Bindungen, Disulfid-Bindungen und ähnliches. Eine Ligierungsreaktion ist für gewöhnlich durch eine Matrize angetrieben, indem die Enden von $Oligo_1$ und $Oligo_2$ durch spezifische Hybridisierung an einem Matrizen-Strang in Nebeneinanderstellung gebracht werden. Ein spezieller Fall der Matrizen-gesteuerten Ligierung ist die Ligierung von zwei doppelsträngigen Oligonucleotiden, die komplementäre, überhängende Stränge besitzen.

[0029] "Komplement" oder "Tagkomplement", wie hier in Bezugnahme auf Oligonucleotid-Tags verwendet, bezieht sich auf ein Oligonucleotid, an welches ein Oligonucleotid-Tag spezifisch hybridisiert, um ein perfekt passendes Duplexmolekül oder Triplexmolekül zu bilden. In Ausführungsformen, in denen die spezifische Hybridisierung ein Triplexmolekül ergibt, kann der Oligonucleotid-Tag so ausgewählt sein, dass er entweder doppelsträngig oder einzelsträngig ist. Somit soll der Begriff "Komplement", wo Triplexmoleküle gebildet werden, entweder ein doppelsträngiges Komplement eines einzelsträngigen Oligonucleotid-Tags oder ein einzelsträngiges Komplement eines doppelsträngigen Oligonucleotid-Tags umfassen.

[0030] Der Begriff „Oligonucleotid" schließt, wie hier verwendet, lineare Oligomere von natürlichen oder modifizierten Monomeren oder Bindungen einschließlich Desoxyribonucleoside, Ribonucleoside, anomere Formen davon, Peptid-Nucleinsäuren (PNAs) und ähnliches ein, welche fähig sind, spezifisch an ein Ziel-Polynucleotid durch ein regelmäßiges Muster von Monomer-zu-Monomer-Interaktionen wie dem Watson-Crick-Typ der Basenpaarung, der Basenstapelung, den Hoogsteen-oder Revers-Hoogsteen-Typen der Basenpaarung oder ähnliches zu binden. Für gewöhnlich werden Monomere durch Phosphodiester-Bindungen oder Analoga davon verbunden, um Oligonucleotide zu bilden, die in einem Größenbereich von einigen wenigen monomeren Einheiten, zum Beispiel 3–4, bis zu einigen 10 monomeren Einheiten, zum Beispiel 40–60 liegen. Wo immer ein Oligonucleotid durch eine Sequenz von Buchstaben wie "A T G C C T G" dargestellt ist, wird es verstanden werden, dass die Nucleotide 5' → 3' von links nach rechts angeordnet sind und dass "A" Desoxyadenosin bezeichnet, "C" Desoxycytidin bezeichnet, "G" Desoxyguanosin bezeichnet und "T" Thymidin bezeichnet, sofern dies nicht anders angegeben ist. Für gewöhnlich umfassen die erfindungsgemäßen Oligonucleotide die vier natürlichen Nucleotide; sie können jedoch auch nicht-natürliche Nucleotid-Analoga umfassen. Dem Durchschnittsfachmann ist klar, wann Oligonucleotide, die natürliche oder nicht-natürliche Nucleotide besitzen, verwendet werden können, zum Beispiel werden, wenn eine Prozessierung durch Enzyme benötigt wird, für gewöhnlich Oligonucleotide, die aus natürlichen Nucleotiden bestehen, benötigt.

[0031] "Perfekt passend" bedeutet im Bezug auf ein Duplexmolekül, dass die Poly- oder Oligonucleotidstränge, die das Duplexmolekül bilden, eine doppelsträngige Struktur miteinander bilden, so dass jedes Nucleotid in jedem Strang eine Watson-Crick-Basenpaarung mit einem Nucleotid in dem anderen Strang eingeht. Der Begriff umfasst auch die Paarung von Nucleosid-Analoga wie Desoxyinosin, Nucleoside mit 2-Aminopurin-Basen und ähnliches, die verwendet werden können. Im Bezug auf ein Triplexmolekül bedeutet der Begriff, dass das Triplexmolekül aus einem perfekt passenden Duplexmolekül und einem dritten Strang besteht, in welchem jedes Nucleotid eine Hoogsteen- oder Revers-Hoogsteen-Assoziation mit einem Basenpaar des perfekt passenden Duplexmoleküls eingeht. Umgekehrt bedeutet eine "Fehlpaarung" in einem Duplexmolekül zwischen einem Tag und einem Oligonucleotid, dass ein Paar oder Triplett von Nucleotiden in dem Duplexmolekül oder Triplexmolekül keine Watson-Crick- und/oder Hoogsteen- und/oder Revers-Hoogsteen-Bindung eingeht.

[0032] Wie hier verwendet, schließt "Nucleosid" die natürlichen Nucleoside, einschließlich 2'-Desoxy- und 2'-Hydroxyl-Formen, wie zum Beispiel in Kornberg und Baker, DNA Replication, 2. Aufl. (Freeman, San Francisco, 1992) beschrieben, ein. "Analoga" im Bezug auf Nucleoside schließt synthetische Nucleoside, die modifizierte Baseneinheiten und/oder modifizierte Zuckereinheiten, zum Beispiel beschrieben durch Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman und Peyman, Chemical Reviews, 90 (1990): 543–584, oder ähnliches, besitzen, mit der einzigen Voraussetzung, dass sie fähig sind spezifisch zu hybridisieren, ein.

Solche Analoga schließen synthetische Nucleoside ein, die dazu ausgelegt sind, die Bindungsfähigkeit zu verbessern, die Komplexität zu reduzieren, die Spezifität zu erhöhen und ähnliches.

[0033] Wie hier verwendet, schließt "Sequenzbestimmung" oder "Bestimmung einer Nucleotidsequenz" in Bezug auf Polynucleotide die Bestimmung einer teilweisen genauso wie einer vollen Sequenzinformation des Polynucleotids ein. Das heißt, dass der Begriff Sequenzvergleiche, Fingerprinting und ähnliche Ebenen der Information über ein Ziel-Polynucleotid genauso wie die ausdrückliche Identifizierung und Ordnung von Nucleosiden, für gewöhnlich eines jeden Nucleosids, in einem Ziel-Polynucleotid einschließt. Der Begriff schließt auch die Bestimmung oder die Identifizierung, Ordnung und Lokalisierung von einer, zwei oder drei der vier Arten von Nucleotiden in einem Ziel-Polynucleotid ein. Zum Beispiel kann in einigen Ausführungsformen die Sequenzbestimmung durch die Identifizierung der Ordnung und Lokalisierung einer einzigen Art von Nucleotid, zum Beispiel Cytosine, in dem Ziel-Polynucleotid "CATCGC ..." durchgeführt werden, so dass seine Sequenz als binärer Code dargestellt ist, zum Beispiel "100101 ..." für "C-(nicht C)-(nicht C)-C-(nicht C)-C ..." und ähnliches.

[0034] Wie hier verwendet bedeutet der Begriff "Komplexität" in Bezug auf eine Population von Polynucleotiden die Zahl der verschiedenen Spezies von Molekülen, die in der Population vorhanden ist.

DETAILLIERTE BESCHREIBUNG DER ERFINDUNG

[0035] Die Erfindung betrifft die Ligierung von codierten Adaptoren, die spezifisch an den Terminus oder die Termini von einem oder mehreren Ziel-Polynucleotiden hybridisiert sind. Die Sequenzinformation über die Region, wo die spezifische Hybridisierung auftritt, wird durch "Decodierung" der Oligonucleotid-Tags der so ligierten codierten Adaptoren gewonnen. In einem Aspekt der Erfindung werden multiple Sätze von codierten Adaptoren an gestaffelten Abspaltungspunkten an das Ziel-Polynucleotid ligiert, so dass die codierten Adaptoren Sequenzinformation von jedem aus einer Vielzahl von Abschnitten des Ziel-Polynucleotids bereitstellen. Solche Abschnitte können getrennt, überlappend oder zusammenhängend sein; die Abschnitte sind jedoch bevorzugt zusammenhängend und erlauben zusammen die Identifizierung einer Sequenz von Nucleotiden, die gleich der Summe der Länge der individuellen Abschnitte ist. In diesem Aspekt gibt es nur eine einzige Ligierung von codierten Adaptoren, gefolgt durch Identifizierung durch "Decodierung" der Tags der ligierten Adaptoren. In einem anderen Aspekt der Erfindung werden codierte Adaptoren als ein Identifizierungsschritt in einem Prozess verwendet, der, nachstehend vollständiger beschrieben, wiederholte Zyklen von Ligierung, Identifizierung und Abspaltung umfasst.

[0036] In der letzteren Ausführungsform verwendet die Erfindung Nucleasen, deren Erkennungsstellen von deren Schneidestellen verschieden sind. Bevorzugt sind solche Nucleasen Typ-II-Restriktionsendonucleasen. Die Nucleasen werden verwendet, um überhängende Stränge an Ziel-Polynucleotiden zu erzeugen, an welche codierte Adaptoren ligiert werden. Die Menge der Sequenzinformation, die in einer gegebenen erfindungsgemäßen Ausführungsform gewonnen wird, hängt teilweise davon, wie viele solcher Nuclease verwendet werden, und von der Länge des bei der Abspaltung erzeugten, überhängenden Stranges ab.

[0037] Ein wichtiger Aspekt der Erfindung ist die Fähigkeit, viele Ziel-Oligonucleotide parallel zu Sequenzieren.

[0038] Entsprechend stellt die vorliegende Erfindung ein Verfahren zur Ermittlung einer Nucleotidsequenz an einem Ende eines Polynucleotids bereit, wobei das Verfahren die Schritte umfasst:

(a) Anwenden einer Vielzahl verschiedener codierter Adaptoren auf das Polynucleotid, wobei jeder codierte Adaptor eine doppelsträngige Desoxyribonucleinsäure ist, umfassend (i) einen Oligonucleotid-Tag, ausgewählt aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden, und (ii) einen überhängenden Strang, welcher in einer bekannten Art und Weise dem Oligonucleotid-Tag entspricht;
wobei jedes Oligonucleotid des minimal kreuz-hybridisierenden Satzes von Oligonucleotiden sich von jedem anderen Oligonucleotid des Satzes in mindestens zwei Nucleotiden unterscheidet;
(b) Ligieren von codierten Adaptoren, deren überhängende Stränge perfekt passende Duplexe mit dem Ende bilden, an das Ende des Polynucleotids; und
(c) für jedes aus einer Vielzahl von Nucleotiden in dem Ende des Polynucleotids, spezifisches Hybridisieren eines markierten Tagkomplements an das Oligonucleotid-Tag jedes codierten Adaptors, der daran ligiert ist, wobei das hybridisierte Tagkomplement in einer bekannten Art und Weise der Identität des Nucleotids entspricht,

wodurch jedes Nucleotid aus der Vielzahl von Nucleotiden in dem Ende des Polynucleotids identifiziert wird.

[0039] In einer bevorzugten Ausführungsform ist das Verfahren nützlich zur Ermittlung von Nucleotidsequenzen einer Vielzahl von Polynucleotiden, wobei das Verfahren des weiteren, vor Schritt (a), die Schritte umfasst:

(i) Anbringen eines ersten Oligonucleotid-Tags aus einem Repertoire von Tags an jedes Polynucleotid in einer Population von Polynucleotiden,
wobei jeder erste Oligonucleotid-Tag aus dem Repertoire ausgewählt ist aus einem ersten minimal kreuz-hybridisierenden Satz von Oligonucleotiden und wobei jedes Oligonucleotid aus dem ersten minimal kreuz-hybridisierenden Satz sich von jedem anderen Oligonucleotid aus dem ersten Satz in mindestens zwei Nucleotiden unterscheidet;
(ii) Auswerten der Population von Polynucleotiden zur Erzeugung einer Probe von Polynucleotiden, so dass im Wesentlichen alle verschiedenen Polynucleotide in der Probe verschiedene erste Oligonucleotid-Tags gebunden haben; und
(iii) Sortieren der Polynucleotide der Probe durch spezifisches Hybridisieren der ersten Oligonucleotid-Tags mit deren entsprechenden Komplementen, wobei die entsprechenden Komplemente als einheitliche Populationen von im Wesentlichen identischen Oligonucleotiden in räumlich diskreten Regionen auf dem einen oder mehreren Festphasenträgern befestigt sind;

und wobei Schritte (a)–(c) auf jedes Polynucleotid der Vielzahl angewandt werden.

[0040] In einer weiteren bevorzugten Ausführungsform ist das Verfahren nützlich zur Identifizierung einer Population von mRNA-Molekülen, wobei die Polynucleotide cDNA-Moleküle sind und wobei Schritt (i) umfasst:
Erzeugen einer Population von cDNA-Molekülen aus der Population von mRNA-Molekülen, so dass jedes cDNA-Molekül einen ersten Oligonucleotid-Tag gebunden hat, wobei die ersten Oligonucleotid-Tags ausgewählt sind aus einem ersten minimal kreuz-hybridisierenden Satz von Oligonucleotiden, wobei jedes Oligonucleotid aus dem ersten minimal kreuz-hybridisierenden Satz sich von jedem anderen Oligonucleotid des ersten Satzes in mindestens zwei Nucleotiden unterscheidet;
und des weiteren umfassend den Schritt:
Identifizieren der Population von mRNA-Molekülen durch die Häufigkeitsverteilung der Abschnitte von Sequenzen der cDNA-Moleküle.

[0041] Bevorzugt schließt der Schritt der Ligierung das Ligieren einer Vielzahl von verschiedenen codierten Adaptoren an das Ende des Polynucleotids ein, so dass die überhängenden Stränge der Vielzahl der verschiedenen codierten Adaptoren komplementär zu einer Vielzahl von verschiedenen Abschnitten des Stranges des Polynucleotids sind und dass ein Eins-zu-Eins-Verhältnis zwischen den verschiedenen codierten Adaptoren und den verschiedenen Abschnitten des Stranges besteht. Bevorzugt sind die verschiedenen Abschnitte des Stranges des Polynucleotids zusammenhängend.

[0042] Eine stärker bevorzugte Ausführungsform des erfindungsgemäßen Verfahrens schließt die Schritte ein:

(d) Abspalten der codierten Adaptoren von den Polynucleotids mit einer Nuclease, die eine Nucleaseerkennungsstelle hat, die verschieden von ihrer Schneidestelle ist, so dass ein neuer überhängender Strang an dem Ende jedes Polynucleotids gebildet wird; und
(e) Wiederholen der Schritte (a) bis (d).

[0043] Weitere bevorzugte Ausführungsformen des erfindungsgemäßen Verfahrens sind durch die folgenden Merkmale charakterisiert:

(i) des weiteren einschließend die Schritte:
(d) Abspalten des codierten Adaptors vom Ende des Polynucleotids mit einer Nuclease, die eine Nucleaseerkennungsstelle hat, die verschieden von ihrer Schneidestelle ist, so dass ein neuer überhängender Strang an dem Ende des Polynucleotids gebildet wird; und (e) Wiederholen der Schritte (a) bis (d).
(ii) Der überhängende Strang des codierten Adaptors enthält zwei bis sechs Nucleotide und der Schritt des Identifizierens umfasst das spezifische Hybridisieren nacheinander der Tagkomplemente an das Oligonucleotid-Tag, so dass die Identität jedes Nucleotids in dem Abschnitt des Polynucleotids nacheinander bestimmt wird.
(iii) Der Schritt des Identifizierens schließt des weiteren die Bereitstellung einer Anzahl von Sätzen von Tagkomplementen ein, die Äquivalent zu der Anzahl von Nucleotiden sind, die in dem Abschnitt des Polynucleotids identifiziert werden sollen. Bevorzugt schließt der Schritt des Identifizierens des weiteren ein: Bereitstellen der Tagkomplemente in jedem der Sätze, die in der Lage sind, die Gegenwart eines zuvor bestimmten Nucleotids durch ein Signal anzuzeigen, das durch eine ein Fluoreszenzsignal erzeugende Einheit erzeugt wird, wobei für jede Art von Nucleotid eine unterschiedliche, ein Fluoreszenzsignal erzeugende Einheit vorhanden ist.

(iv) Die Oligonucleotid-Tags der in dem Verfahren verwendeten codierten Adaptoren sind einzelsträngig und die Tagkomplemente zu den Oligonucleotid-Tags sind einzelsträngig, so dass eine spezifische Hybridisierung zwischen einem Oligonucleotid-Tag und seinem entsprechenden Tagkomplement durch Watson-Crick-Basenpaarung erfolgt. Bevorzugt haben die codierten Adaptoren die folgende Form:

$$\begin{array}{l} 5'\text{-}p(N)_n(N)_r(N)_s(N)_q(N)_t\text{-}3' \\ \quad\quad z(N')_r(N')_s(N')_o\text{-}5' \end{array}$$

oder

$$\begin{array}{l} \quad\quad p(N)r(N)s(N)q(N)t\text{-}3' \\ 3'\text{-}z(N)n(N')r(N')s(N')q\text{-}5' \end{array} ,$$

wobei N ein Nucleotid ist und N' sein Komplement ist, p eine Phosphatgruppe ist, z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist, n eine ganze Zahl zwischen einschließlich 2 und 6 ist, r eine ganze Zahl zwischen einschließlich 0 und 18 ist, s eine ganze Zahl ist, welche entweder zwischen einschließlich 4 und 6 ist, wenn der codierte Adaptor eine Nucleaseerkennungsstelle hat, oder 0 ist, wenn keine Nucleaseerkennungsstelle vorhanden ist, q eine ganze Zahl größer oder gleich 0 ist und t eine ganze Zahl größer oder gleich 8 ist. Bevorzugt ist r zwischen einschließlich 0 und 12, t ist eine ganze Zahl zwischen einschließlich 8 und 20 und z ist eine Phosphatgruppe.

(v) Die Oligonucleotid-Tags der in dem Verfahren verwendeten codierten Adaptoren sind doppelsträngig und die Tagkomplemente zu den Oligonucleotid-Tags sind einzelsträngig, so dass eine spezifische Hybridisierung zwischen einem Oligonucleotid-Tag und seinem entsprechenden Tagkomplement durch die Bildung eines Hoogsteen- oder Revers-Hoogsteen-Triplexmoleküls erfolgt. Bevorzugt haben die codierten Adaptoren die Form:

$$\begin{array}{l} 5'\text{-}p(N)_n(N)_r(N)_s(N)_q(N)_t\text{-}3' \\ \quad\quad z(N')_r(N')_s(N')_q(N')_t\text{-}5' \end{array}$$

oder

$$\begin{array}{l} \quad\quad p(N)r(N)s(N)q(N)t\text{-}3' \\ 3'\text{-}z(N)n(N')r(N')s(N')q(N')t\text{-}5' \end{array} ,$$

wobei N ein Nucleotid ist und N' sein Komplement ist, p eine Phosphatgruppe ist, z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist, n eine ganze Zahl zwischen einschließlich 2 und 6 ist, r eine ganze Zahl zwischen einschließlich 0 und 18 ist, s eine ganze Zahl ist, welche entweder zwischen einschließlich 4 und 6 liegt, wenn der codierte Adaptor eine Nucleaseerkennungsstelle hat, oder 0 ist, wenn keine Nucleaseerkennungsstelle vorhanden ist, q eine ganze Zahl größer oder gleich 0 ist und t eine ganze Zahl größer oder gleich 8 ist. Bevorzugt ist r zwischen einschließlich 0 und 12, t eine ganze Zahl zwischen einschließlich 12 und 24 und z eine Phosphatgruppe.

(vi) Die Mitglieder des im erfindungsgemäßen Verfahren verwendeten minimal kreuz-hybridisierenden Satzes unterscheiden sich von jedem anderen Mitglied in mindestens sechs Nucleotiden.

Sequenzuntersuchung ohne Zyklen von Ligierung und Abspaltung

[0044] Eine Sequenzuntersuchung ohne Zyklen von Ligierung und Abspaltung in Übereinstimmung mit der Erfindung wird durch die Ausführungsform der Fig. 1A bis Fig. 1E dargestellt. In dieser Ausführungsform werden k Ziel-Polynucleotide, wie nachstehend und auch in Brenner, internationale Patentanmeldungen PCT/US 95/12791 (WO 96/12041) und PCT/US 96/09513 (WO 96/41011) beschrieben, hergestellt. Das heißt, aus einer Population von Polynucleotiden, die mit Oligonucleotid-Tags konjugiert ist, welche durch kleine "t's" bezeichnet sind, wird eine Probe genommen. Diese Tags werden manchmal als Oligonucleotid-Tags zur Sortierung oder als "erste" Oligonucleotid-Tags bezeichnet. Die Tag-Polynucleotid-Konjugate der Probe werden vervielfältigt,

zum Beispiel durch die Polymerasekettenreaktion (PCR) oder durch Clonierung, um 1 bis k Populationen von Konjugaten zu ergeben, angezeigt durch (14)–(18) in Fig. 1A. Bevorzugt werden die Enden der Konjugate, welche den (kleinen "t") Tags gegenüber liegen, zur Ligierung mit einem oder mehreren Adaptoren vorbereitet, von welchen jeder eine Erkennungsstelle für eine Nuclease enthält, deren Schneidestelle von ihrer Erkennungsstelle verschieden ist. In der dargestellten Ausführungsform werden drei solche Adaptoren, hier als "Abspaltungsadaptoren" bezeichnet, verwendet. Die Zahl solcher verwendeter Adaptoren hängt von verschiedenen Faktoren einschließlich der Menge der gewünschten Sequenzinformation, der Verfügbarkeit von Typ-II-Nucleasen mit geeigneten Reichweiten und Schneide-Charakteristiken und ähnlichem ab. Bevorzugt werden von eins bis drei Abspaltungsadaptoren verwendet und die Adaptoren werden darauf ausgelegt, verschiedene Arten von Typ-II-Nucleasen zu beherbergen, welche fähig sind, nach der Spaltung überhängende Stränge von wenigstens vier Nucleotiden zu erzeugen.

[0045] Wenn das erfindungsgemäße Verfahren für das Signatur-Sequenzieren einer cDNA-Population verwendet wird, dann können vor einer Ligierung der Abspaltungsadaptoren die Tag-Polynucleotid-Konjugate mit einer Restriktionsendonuclease mit einer hohen Frequenz von Erkennungsstellen wie TaqI, AluI, HinPII, DpnII, NlaIII oder ähnlichen gespalten werden. Bei Enzymen wie AluI, die glatte Enden hinterlassen, kann ein abgestuftes Ende mit der T4-DNA-Polymerase, zum Beispiel wie in Brenner, internationale Patentanmeldung PCT/US 95/12791, vorstehend zitiert, und Kuijper et al., Gene, 112 (1992): 147–155, beschrieben, erzeugt werden. Wenn die Ziel-Polynucleotide durch Abspaltung mit TaqI erzeugt werden, dann sind die folgenden Enden für die Ligierung verfügbar:
cgannn ... -3'
tnnnn ... -5'

[0046] Somit kann ein beispielhafter Satz von drei Abspaltungsadaptoren wie folgt konstruiert werden:

```
(1)   NN ... NGAAGA      cgannnnnnnnnnnnnnnnn ... -3'
      NN ... NCTTCTGCp     tnnnnnnnnnnnnnnnnnn ... -5'

(2)   NN ... NGCAGCA     cgannnnnnnnnnnnnnnnn ... -3'
      NN ... NCGTCGTGCp    tnnnnnnnnnnnnnnnnnn ... -5'

(3)   NN ... NGGGA       cgannnnnnnnnnnnnnnnn ... -3'
      NN ... NCCCTGCp      tnnnnnnnnnnnnnnnnnn ... -5'
```

wobei die Abspaltungsadaptoren (1), (2) und (3) in Großbuchstaben mit den jeweiligen Erkennungsstellen der Nucleasen BbsI, BbvI und BsmFI unterstrichen und einem 5'-Phosphat angezeigt als "p" gezeigt werden. Die doppelt unterstrichenen Abschnitte der Ziel-Polynucleotide zeigen die Positionen der überhängenden Stränge nach Ligierung und Abspaltung. In allen Fällen wird das Ziel-Polynucleotid mit einem 5'-überhängenden Strang von vier Nucleotiden belassen. Klarerweise können viele verschiedene Ausführungsformen unter Verwendung verschiedener Zahlen und Arten von Nucleasen konstruiert werden. Wie in Brenner, US-Patent 5,599,675 und WO 95/27080 diskutiert, werden interne BbsI, BbvI und BsmFI-Stellen bevorzugt vor der Spaltung zum Beispiel durch Methylierung blockiert, um unerwünschte Abspaltungen an internen Stellen des Ziel-Polynucleotids zu verhindern.

[0047] Zu der dargestellten Ausführungsform zurückkehrend werden die Abspaltungsadaptoren A_1, A_2 und A_3 in einem Konzentrationsverhältnis von 1 : 1 : 1 an die k Ziel-Polynucleotide ligiert (**20**), um die in Fig. 1B gezeigten Konjugate zu ergeben, so dass in jeder Population von Tag-Polynucleotid-Konjugaten annähernd gleiche Zahlen von Konjugaten vorliegen, die A_1, A_2 und A_3 gebunden haben. Nach der Ligierung (**20**) werden die Ziel-Polynucleotide nacheinander mit jeder der Nucleasen der Abspaltungsadaptoren gespalten und an einen Satz von codierten Adaptoren ligiert. Zuerst werden die Ziel-Polynucleotide mit der Nuclease des Abspaltungsadaptors A_1 abgespalten (**22**), wonach ein erster Satz von codierten Adaptoren an die sich ergebenden überhängenden Stränge ligiert wird. Die Abspaltung führt dazu, dass etwa ein Drittel der Ziel-Polynucleotide von jeder Art, das heißt t_1, t_2, ... t_k, für eine Ligierung zur Verfügung stehen. Bevorzugt werden die codierten Adaptoren als ein Gemisch oder mehrere Gemische von Adaptoren, welche zusammengenommen jede mögliche Sequenz eines überhängenden Stranges enthalten, angewendet. Die Reaktionsbedingungen werden so ausgewählt, dass nur codierte Adaptoren, deren überhängende Stränge perfekt passende Duplexmoleküle mit denen der Ziel-Polynucleotide bilden, ligiert werden, um codierte Konjugate (**28**), (**30**) und (**32**) (Fig. 1C) zu bilden. Die großgeschriebenen "T's" mit tiefer gestellten Indices zeigen an, dass einzigartige Oligonucleotid-Tags

von den codierten Adaptoren zur Markierung getragen werden. Die von den codierten Adaptoren getragenen Oligonucleotid-Tags werden manchmal als Tags zur Anbringung von Markierungen an die codierten Adaptoren oder als "zweite" Oligonucleotid-Tags bezeichnet. Wie nachstehend vollständiger beschrieben, bestehen einzelsträngige Oligonucleotid-Tags, welche zum Sortieren verwendet werden, bevorzugt aus nur drei der vier Nucleotide, so dass eine T4-DNA-Polymerase-"Stripping"-Reaktion, zum Beispiel Kuijper et al. (vorstehend zitiert), verwendet werden kann, um Ziel-Polynucleotide zur Beladung auf Festphasenträger vorzubereiten. Andererseits können Oligonucleotid-Tags, die zur Anbringung von Markierungen verwendet werden, aus allen vier Nucleotiden bestehen.

[0048] Wie vorstehend erwähnt, umfassen codierte Adaptoren einen überhängenden Strang (**24**) und einen Oligonucleotid-Tag (**26**). So dass, wenn die "A_1"-Abspaltung der t_1-Polynucleotid-Konjugate folgende Enden ergibt:
5'- ... nnnnnnnnn
3'- ... nnnnnnnnnacct
dann könnte der Oligonucleotid-Tag T_{24} folgende Struktur haben (SEQ ID NO: 1):

```
tggattctagagagagagagagagag -3'
    aagatctctctctctctctctc
                           ,
```

wobei der doppelsträngige Abschnitt einer aus einem Satz von 48 (= 12 Nucleotidpositionen × 4 Arten von Nucleotiden) doppelsträngigen 20-mer Oligonucleotid-Tags sein kann, welche ein perfekt passendes Triplexmolekül mit einem einzigartigen Tagkomplement bilden und ein Triplexmolekül mit wenigstens 6 Fehlpaarungen mit allen anderen Tagkomplementen bilden. Die codierten Adaptoren in diesem Beispiel können an die Ziel-Polynucleotide in einem oder mehreren Gemischen aus einer Gesamtheit von 768 (3 × 256) Mitgliedern ligiert werden. Optional kann ein codierter Adaptor auch eine Spacer-Region umfassen, wie dies in dem vorstehenden Beispiel gezeigt wird, worin die vier-Nucleotidsequenz "ttct" als Spacer zwischen dem überhängenden Strang und dem Oligonucleotid-Tag dient.

[0049] Nach Ligierung des ersten Satzes von codierten Adaptoren (**28**), (**30**) und (32) werden die Tag-Polynucleotid-Konjugate mit der Nuclease des Abspaltungsadaptors A_2 abgespalten (**34**), wonach ein zweiter Satz von codierten Adaptoren angewendet wird, um Konjugate zu bilden (**36**), (**38**) und (**40**) (Fig. 1D). Schließlich werden die Tag-Polynucleotid-Konjugate mit der Nuclease des Abspaltungsadaptors A_3 abgespalten (**42**), wonach ein dritter Satz von codierten Adaptoren angewendet wird, um Konjugate zu bilden (**44**), (**46**) und (**48**) (Fig. 1E). Nach Abschluss der Aufeinanderfolge von Abspaltungen und Ligierungen von codierten Adaptoren wird das Gemisch (**50**) über die Oligonucleotid-Tags t_1 bis t_k auf einen oder mehrere Festphasenträger, wie nachstehend vollständiger beschrieben und wie durch Brenner, zum Beispiel PCT/US 95/12791 oder PCT/US 96/09513 gelehrt, geladen. Wenn ein einzelnes Ziel-Polynucleotid untersucht wird, dann sind klarerweise vielfache Oligonucleotid-Tags t_1, t_2, ... t_k, nicht notwendig. In solch einer Ausführungsform kann Biotin oder eine ähnliche Einheit verwendet werden, um das Polynucleotid-codierte Adaptor-Konjugat zu verankern, da kein Sortieren benötigt wird. Auch hängt die Anordnung der Schritte der Abspaltung, Ligierungen und Beladung auf Festphasenträger von der speziellen durchgeführten Ausführungsform ab. Zum Beispiel können die Tag-Polynucleotid-Konjugate zuerst auf einen Festphasenträger geladen werden, gefolgt von Ligierung der Abspaltungsadaptoren, Abspaltungen derer und Ligierung von codierten Adaptoren; oder die Abspaltungsadaptoren können zuerst ligiert werden, gefolgt durch Laden, Abspalten und Ligierung von codierten Adaptoren; und so weiter.

[0050] Nachdem codierte Adaptoren in Übereinstimmung mit der Erfindung an die Enden eines Ziel-Polynucleotids ligiert worden sind wird die Sequenzinformation durch aufeinanderfolgende Anwendung markierter Tagkomplemente an die immobilisierten Ziel-Polynucleotide entweder einzeln oder als Gemisch unter Bedingungen, die die Bildung von perfekt passenden Duplexmolekülen und/oder Triplexmolekülen zwischen den Oligonucleotid-Tags der codierten Adaptoren und deren entsprechenden Tagkomplementen erlauben, gewonnen. Die Zahlen und Komplexität der Gemische hängt von verschiedenen Faktoren, einschließlich der Art des verwendeten Markierungssystems, der Länge der Abschnitte, deren Sequenzen zu identifizieren sind, ob die Komplexität reduzierende Analoga verwendet werden und ähnlichem ab. Bevorzugt wird für die in den Fig. 1a bis Fig. 1e dargestellte Ausführungsform ein einzelner Fluoreszenzfarbstoff verwendet, um jedes der 48 (= 3 × 16) Tagkomplemente zu markieren. Die Tagkomplemente werden einzeln angewendet, um die Nucleotide von jedem der vier-Nucleotid-Abschnitte des Ziel-Polynucleotids zu identifizieren (das heißt vier Tagkomplemente für jede von 12 Positionen für eine Gesamtheit von 48). Klarerweise würden Abschnitte von unterschiedlichen Längen verschiedene Zahlen von Tagkomplementen erfordern, zum Beispiel in Übereinstimmung mit dieser Ausführungsform würde ein 5-Nucleotid-Abschnitt 20 Tagkomplemente erfordern, ein 2-Nucleotid-Ab-

schnitt acht Tagkomplemente erfordern und so weiter. Die Tagkomplemente werden unter ausreichend stringenten Bedingungen angewendet, so dass nur perfekt passende Duplexmoleküle gebildet werden, Signale von den fluoreszierenden Markierungen auf den spezifisch hybridisierten Tagkomplementen werden gemessen und die Tagkomplemente werden von den codierten Tags abgewaschen, so dass das nächste Gemisch angewendet werden kann. Die 16 Tagkomplemente haben eine eins-zu-eins-Entsprechung mit den folgenden Sequenzen der vier-mer-Abschnitte der Ziel-Sequenz:

ANNN	NANN	NNAN	NNNA
CNNN	NCNN	NNCN	NNNC
GNNN	NGNN	NNGN	NNNG
TNNN	NTNN	NNTN	NNNT

,

wobei "N" ein jegliches der Nucleotide A, C, G oder T ist. Somit werden für jede Nucleotid-Position vier unabhängige Abfragen gemacht, eine für jede Art von Nucleotid. Diese Ausführungsform schließt einen signifikanten Grad der Redundanz (eine Gesamtheit von 16 Tagkomplementen wird verwendet, um vier Nucleotide zu identifizieren) im Austausch für eine gesteigerte Verlässlichkeit der Nucleotid-Bestimmung ein.

[0051] Alternativ können 12 Gemische von jeweils vier Tagkomplementen aufeinanderfolgend unter Verwendung von vier spektral unterscheidbaren Fluoreszenzfarbstoffen angewendet werden, so dass es eine eins-zu-eins-Entsprechung zwischen Farbstoffen und Arten von Nucleotiden gibt. Zum Beispiel kann ein Gemisch von vier Tagkomplementen das Nucleotid "x" in der überhängenden Strangsequenz "nnxn" identifizieren, so dass eine erste fluoreszierende Markierung beobachtet wird, wenn x = A, eine zweite fluoreszierende Markierung beobachtet wird, wenn x = C, eine dritte fluoreszierende Markierung beobachtet wird, wenn x = G und so weiter.

[0052] Weitere Sequenzinformation kann unter Verwendung der vorstehend beschriebenen Ausführungsform in einem Prozess analog zu dem "Multi-Stepping"-Verfahren, offenbart in Brenner, internationale Patentanmeldung PCT/US 95/03678 (WO 95/27080), erhalten werden. In dieser Ausführungsform wird ein vierter Adaptor, hier bezeichnet als "Stepping-Adaptor", an die Enden der Ziel-Polynucleotide zusammen mit den Abspaltungsadaptoren A_1, A_2 und A_3 zum Beispiel in einem Konzentrationsverhältnis von 3:1:1:1 ligiert. Somit werden annähernd die Hälfte der verfügbaren Enden an den "Stepping-Adaptor" ligiert. Der Stepping-Adaptor schließt eine Erkennungsstelle für eine Typ-II-Nuclease ein, die so positioniert ist, dass ihre Reichweite (nachstehend definiert) die Abspaltung des Ziel-Polynucleotids an dem Ende der über die Abspaltungsadaptoren A_1, A_2 und A_3 bestimmten Sequenz erlaubt. Ein Beispiel für einen Stepping-Adaptor, der mit dem vorstehenden Satz von Abspaltungsadaptoren verwendet werden kann, ist wie folgt:

```
NN ... NCTGGAGA      cgannnnnnnnnnnnnnnnnn ... -3'
NN ... NGACCTCTGCp     tnnnnnnnnnnnnnnnnnn ... -5'
                                                ,
```

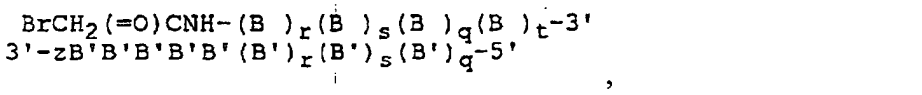
worin wie vorstehend die Erkennungsstelle der Nuclease, in diesem Fall BpMI, einfach unterstrichen ist und die Nucleotide an der Abspaltungsstelle doppelt unterstrichen sind. Die mit der Nuclease des Stepping-Adaptors gespaltenen Ziel-Polynucleotide können mit einem weiteren Satz von Abspaltungsadaptoren A_4, A_5 und A_6 ligiert werden, welche Nucleaseerkennungsstellen enthalten können, die die gleichen oder unterschiedliche sind wie jene, welche in den Abspaltungsadaptoren A_1, A_2 und A_3 enthalten sind. Ob ein vergrößerter Satz von codierten Adaptoren benötigt wird oder nicht, hängt davon ab, ob die Abspaltungs- und Ligierungsreaktionen in dem Signalmessgerät toleriert werden können. Wenn es wie vorstehend gewünscht ist, die Enzymreaktionen in Verbindung mit der Signalmessung zu minimieren, dann müssen zusätzliche Sätze von codierten Adaptoren verwendet werden. Das heißt, wo vorstehend 768 Oligonucleotid-Tags und Tagkomplemente erforderlich waren, wobei sechs Abspaltungsreaktionen jeweils überhängende Stränge von vier Nucleotiden erzeugen, wären 1536 Oligonucleotid-Tags und Tagkomplemente (jeweils 24 Gemische von 64 Tagkomplementen) erforderlich. Beispielhafte Abspaltungsadaptoren A_4, A_5 und A_6 mit den gleichen Nucleaseerkennungsstellen wie A_1, A_2 und A_3 und welche mit dem vorstehend gezeigten Stepping-Adaptor verwendet werden können, sind wie folgt:

```
(4)   NN ... NGAAGACNN        nnnnnnnnnnnnnnnnnn ... -3'
      NN ... NCTTCTGp         nnnnnnnnnnnnnnnnnn ... -5'

(5)   NN ... NGCAGCACNN       nnnnnnnnnnnnnnnnnn ... -3'
      NN ... NCGTCGTGp        nnnnnnnnnnnnnnnnnn ... -5'

(6)   NN ... NGGGACNN         nnnnnnnnnnnnnnnnnn ... -3'
      NN ... NCCCTGp          nnnnnnnnnnnnnnnnnn ... -5'
```

wobei die Abspaltungsstellen durch doppelte Unterstreichung gekennzeichnet sind. Die Abspaltungsadaptoren A_4, A_5 und A_6 werden bevorzugt als Gemische angewendet, so dass jeder mögliche überhängende zwei-Nucleotid-Strang repräsentiert ist.

[0053] Sind die codierten Adaptoren einmal ligiert, werden die Ziel-Polynucleotide für die Beladung auf Festphasenträger, bevorzugt Mikropartikel, wie offenbart in Brenner, internationale Patentanmeldung PCT/US 95/12791 (WO 96/12014), vorbereitet. Kurzgesagt werden die zu sortierenden Oligonucleotid-Tags unter Verwendung einer "Stripping"-Reaktion mit T4-DNA-Polymerase, zum Beispiel Kuijper et al. (vorstehend zitiert), einzelsträngig gemacht. Die einzelsträngigen Oligonucleotid-Tags werden spezifisch hybridisiert und an ihre Tagkomplemente auf Mikropartikeln ligiert. Die beladenen Mikropartikel werden dann in einem Gerät untersucht, wie beschrieben in Brenner (vorstehend zitiert), welches die aufeinander folgende Zuführung, spezifische Hybridisierung und Entfernung von markierten Tagkomplementen an codierte Adaptoren erlaubt.

[0054] In Ausführungsformen, worin codierte Adaptoren nur einmal an ein Ziel-Polynucleotid ligiert werden (oder an eine Population von Ziel-Polynucleotiden), existieren einige nicht-enzymatische, Matrizen-getriebene Verfahren der Ligierung, die in Übereinstimmung mit der Erfindung verwendet werden können. Solche Ligierungsverfahren schließen jene ein, welche in Shabarova, Biochimie 70 (1988): 1323–1334; Dolinnaya et al., Nucleic Acids Research, 16 (1988): 3721–3738; Letsinger et al., US-Patent 5,476,930; Gryaznov et al., Nucleic Acids Research, 22 (1994) 2366–2369; Kang et al., Nucleic Acids Research, 23 (1995): 2344–2345; Gryaznov et al., Nucleic Acids Research, 21 (1993): 1403–1408; Gryaznov, US-Patent 5,571,677; und ähnlichen Quellen offenbart sind, sind aber nicht auf diese beschränkt. Bevorzugt wird die nicht-enzymatische Ligierung durch das Verfahren von Letsinger et al. (vorstehend zitiert) durchgeführt. In diesem Verfahren wird ein codierter Adaptor, der ein 3'-bromacetyliertes Ende hat, mit einem Polynucleotid umgesetzt, das einen komplementären überhängenden Strang und eine Thiophosphoryl-Gruppe an seinem 5'-Ende besitzt. Ein beispielhafter codierter Adaptor, der eine solche Chemie verwendet, hat die folgende Struktur:



worin B und B' Nucleotide und deren Komplemente sind und z, r, s, q und t wie nachstehend beschrieben sind. Br, C, H und N haben ihre gewöhnlichen chemischen Bedeutungen. Wie in den vorstehenden Quellen erklärt, reagiert in einer Matrizen-getriebenen Reaktion das 3'-bromacetylierte Oligonucleotid unter wässrigen Bedingungen spontan mit einem Oligonucleotid, das eine 5'-Thiophosphoryl-Gruppe besitzt, um eine Thiophosphorylacetylamino-Bindung zu bilden. Eine Thiophosphoryl-Gruppe wird, wie in Kang et al. beschrieben (vorstehend zitiert), durch Behandlung mit T4-Kinase) in der Anwesenheit von Adenosin-5'-O-(1-thiotriphosphat), das heißt g-S-ATP, einfach an das 5'-Hydroxyl eines Ziel-Polynucleotids angehängt.

Sequenzuntersuchung mit Zyklen von Ligierung und Abspaltung

[0055] Codierte Adaptoren können in einem auf Adaptoren basierenden Verfahren der DNA-Sequenzierung verwendet werden, welches wiederholte Zyklen von Ligierung, Identifizierung und Abspaltung, wie das in Brenner, US-Patent 5,599,675 und PCT-Veröffentlichung Nr. WO 95/27080 beschriebene Verfahren, einschließt. Kurzgesagt umfasst so ein Verfahren die folgenden Schritte: (a) Ligieren eines codierten Adaptors an das Ende eines Polynucleotids, wobei der codierte Adaptor eine Nucleaseerkennungsstelle einer Nuclease, deren Schneidestelle von ihrer Erkennungsstelle verschieden ist, besitzt; (b) Identifizieren von einem oder mehreren Nucleotiden am Ende des Polynucleotids durch die Identität des daran ligierten codierten Adaptors, (c) Abspalten des Polynucleotids mit einer Nuclease, die die Nucleaseerkennungsstelle des codierten Adaptors erkennt, so dass das Polynucleotid um ein oder mehrerer Nucleotide gekürzt wird; und (d) Wiederholen der Schritte (a) bis (c), bis die Nucleotidsequenz des Polynucleotids bestimmt ist. Bei dem Identifizierungsschritt werden auf-

einander folgende Sätze von Tagkomplementen spezifisch an die jeweiligen Tags hybridisiert, welche von den codierten Adaptoren getragen werden, die an die Enden des Ziel-Polynucleotids, wie vorstehend beschrieben, ligiert sind. Die Art und Sequenz von Nucleotiden in den überhängenden Strängen des Polynucleotids werden durch die Markierung, welche durch das spezifisch hybridisierte Tagkomplement getragen wird, und den Satz, aus welchem das Tagkomplement stammt, wie vorstehend beschrieben, identifiziert.

Oligonucleotid-Tags und Tagkomplemente

[0056] Oligonucleotid-Tags werden in den bevorzugten erfindungsgemäßen Ausführungsformen für zwei verschiedene Zwecke verwendet: Oligonucleotid-Tags werden, wie in Brenner, internationale Patentanmeldungen PCT/US 95/12791 und PCT/US 96/09513 (WO 96/12014 und WO 96/41011), beschrieben, verwendet, um große Zahlen von Polynucleotiden, zum Beispiel einige tausend bis einige hunderttausend, aus einem Gemisch in uniforme Populationen von identischen Polynucleotiden zur Untersuchung zu sortieren und sie werden verwendet, um Markierungen an codierte Adaptoren, die sich zahlenmäßig im Bereich von einigen zehn bis einigen tausend bewegen, anzubringen. Für die erste Anwendung werden typischerweise große Zahlen oder Repertoires von Tags benötigt und deshalb ist die Synthese von individuellen Oligonucleotid-Tags problematisch. In diesen Ausführungsformen ist die kombinatorische Synthese der Tags bevorzugt. Andererseits können, wo extrem große Repertoires von Tags nicht benötigt werden – wie bei der Anbringung von Markierungen an codierte Adaptoren, Oligonucleotid-Tags eines minimal kreuz-hybridisierenden Satzes getrennt genauso wie kombinatorisch synthetisiert werden.

[0057] Wie in Brenner (vorstehend zitiert) beschrieben, werden die Nucleotidsequenzen von Oligonucleotiden eines minimal kreuz-hybridisierenden Satzes durch einfache Computerprogramme wie jene, die durch die Programme, deren Quellcodes in den Anhängen I und II aufgeführt sind, veranschaulicht werden, spezifiziert. Ähnliche Computerprogramme zur Auflistung von Oligonucleotiden eines minimal kreuz-hybridisierenden Satzes für jegliche erfindungsgemäße Ausführungsform sind einfach zu schreiben. Die nachstehende Tabelle 1 stellt eine Anleitung für die Größe von Sätzen von minimal kreuz-hybridisierenden Oligonucleotiden für die angegebenen Längen und Zahlen von Nucleotid-Differenzen zur Verfügung. Die vorstehenden Computerprogramme wurden verwendet, um die Zahlen zu erzeugen.

Tabelle I

Minimal kreuz-hybridisierende Sätze von aus vier Nucleotiden bestehenden Worten

Oligonucleotid-Wortlänge	Nucleotidunterschied zwischen Oligonucleotiden eines minimal kreuz-hybridisierenden Satzes	Maximale Größe eines minimal kreuz-hybridiserenden Satzes	Größe eines Repertoires mit drei Worten	Größe eines Repertoires mit vier Worten
4	3	11	1331	14,641
6	4	25	15,625	3.9×10^5
6	5	4	64	256
8	4	225	1.14×10^7	
8	5	56	1.75×10^5	
8	6	17	4913	
12	8	62		

[0058] Sätze, die einige hundert bis einige tausend oder sogar einige zehntausend Nucleotide enthalten, können direkt durch eine Vielzahl von parallelen Syntheseansätzen synthetisiert werden, wie zum Beispiel offenbart in Frank et al., US-Patent 4,689,405; Frank et al., Nucleic Acids Research, 11 (1983): 4365–4377; Matson et al., Anal. Biochem., 224 (1995) 110–116; Fodor et al., internationale Anmeldung PCT/US 93/04145 (WO 93/22684); Pease et al., Proc. Natl. Acad. Sci., 91 (1994): 5022–5026; Southern et al., J. Biotechnology, 35 (1994): 217–227; Brennan, internationale Anmeldung PCT/US 94/05896 (WO 94/27719); Lashkari et al., Proc. Natl. Acad. Sci., 92 (1995): 7912–7915; oder ähnliches.

[0059] Bevorzugt werden Tagkomplemente in Gemischen, ob kombinatorisch oder individuell synthetisiert, so ausgewählt, dass sie im Vergleich miteinander ähnliche Duplex- oder Triplex-Stabilitäten besitzen, so dass perfekt passende Hybride ähnliche oder im Wesentlichen identische Schmelztemperaturen besitzen. Dies erlaubt es, dass fehlgepaarte Tagkomplemente einfacher, zum Beispiel durch Waschen unter stringenten Bedingen, von perfekt passenden Tagkomplementen unterschieden werden können, wenn sie an codierten Adaptoren angewendet werden. Für kombinatorisch synthetisierte Tagkomplemente können minimal kreuz-hybridisierende Sätze aus Untereinheiten konstruiert werden, die annähernd gleichwertige Beiträge zur Duplex-Stabilität wie jede andere Untereinheit in dem Satz leisten. Eine Anleitung zur Durchführung solcher Auswahlen wird durch veröffentlichte Verfahren zur Auswahl optimaler PCR-Primer und der Berechnung von Duplex-Stabilitäten bereitgestellt, zum Beispiel Rychlik et al., Nucleic Acids Research, 17 (1989): 8543–8551 und 18 (1990): 6409–6412, Breslauer et al., Proc. Natl. Acad. Sci., 83 (1986): 3746–3750; Wetmur, Crit. Rev. Biochem. Mol. Biol., 26 (1991): 227–259; und ähnliches. Wenn kleinere Zahlen von Oligonucleotid-Tags benötigt werden, wie zur Anbringung von Markierungen an codierte Adaptoren, können die Computerprogramme der Anhänge I und II verwendet werden, um die Sequenzen von minimal kreuz-hybridisierenden Sätzen von Oligonucleotiden zu erzeugen und aufzulisten, welche direkt verwendet werden (das heißt ohne Verkettung in "Sätze"). Solche Listen können weiter nach zusätzlichen Kriterien wie GC-Gehalt, Verteilung von Fehlpaarungen, theoretischen Schmelztemperaturen und ähnlichem durchmustert werden, um zusätzliche minimal kreuz-hybridisierende Sätze zu erzeugen.

[0060] Für kürzere Tags, zum Beispiel etwa 30 Nucleotide oder weniger, wird der durch Rychlik und Wetmur beschriebene Algorithmus zur Berechnung der Duplex-Stabilität bevorzugt und für längere Tags, zum Beispiel etwa 30–35 Nucleotide oder größer, kann ein durch Suggs et al., Seiten 683–693 in Brown, Herausgeber, ICN-UCLA Symp. Dev. Biol., Bd. 23 (Academic Press, New York, 1981) offenbarter Algorithmus günstigerweise verwendet werden. Klarerweise gibt es viele Ansätze, die dem Durchschnittsfachmann zum Entwurf von Sätzen von minimal kreuz-hybridisierenden Untereinheiten im Umfang der Erfindung zur Verfügung stehen. Zum Beispiel können zur Minimierung der Auswirkungen von verschiedenen Energien der Basen-Stapelung von endständigen Nucleotiden beim Zusammenbau von Untereinheiten Untereinheiten bereitgestellt werden, welche die gleichen endständigen Nucleotide besitzen. In dieser Weise werden, wenn Untereinheiten verbunden werden, die Summen der Energien der Basen-Stapelung von allen aneinander angrenzenden endständigen Nucleotiden gleich sein, wodurch die Variabilität in Tag-Schmelztemperaturen reduziert oder eliminiert wird.

[0061] Bei Multi-Untereinheiten-Tags kann ein "Wort" aus endständigen Nucleotiden, nachstehend kursiv gezeigt, auch zu jedem Ende eines Tags hinzugefügt werden, so dass jedes Mal ein perfektes Zusammenpassen zwischen ihm und einem ähnlichen endständigen "Wort" auf jedem anderen Tagkomplement gebildet wird. Solch ein vergrößerter Tag würde die Form haben:

W	W_1	W_2	... W_{k-1}	W_k	*W*
W'	W_1'	W_2'	... W_{k-1}'	W_k'	*W'*

worin die mit einem Apostroph versehenen Ws Komplemente anzeigen. Da die Enden der Tags immer perfekt passende Duplexe bilden, werden alle fehlgepaarten Wörter interne Fehlpaarungen sein, wodurch die Stabilität von Tagkomplement-Duplexmolekülen reduziert wird, die ansonsten fehlgepaarte Wörter an ihren Enden besitzen würden. Es ist wohl bekannt, dass Duplexmoleküle mit internen Fehlpaarungen signifikant weniger stabil sind als Duplexmoleküle mit der gleichen Fehlpaarung an einem Ende.

[0062] Bei für das Sortieren verwendeten Oligonucleotid-Tags ist eine bevorzugte Ausführungsform von minimal kreuz-hybridisierenden Sätzen jene, deren Untereinheiten aus drei der vier natürlichen Nucleotide aufgebaut sind. Wie nachstehend vollständiger diskutiert werden wird, erlaubt die Abwesenheit von einer Art von Nucleotid in den Oligonucleotid-Tags, dass Ziel-Nucleotide auf Festphasenträger unter Verwendung der 5'→3'-Exonucleaseaktivität einer DNA-Polymerase geladen werden. Das folgende ist ein beispielhafter minimal kreuz-hybridisierender Satz von Untereinheiten, von denen jede vier Nucleotide, ausgewählt aus der Gruppe bestehend aus A, G und T, umfasst:

Wort:	w_1	w_2	w_3	w_4
Sequenz:	GATT	TGAT	TAGA	TTTG
Wort:	w_5	w_6	w_7	w_8
Sequenz:	GTAA	AGTA	ATGT	AAAG

[0063] In diesem Satz würde jedes Mitglied ein Duplexmolekül bilden, das drei fehlgepaarte Basen mit dem Komplement von jedem anderen Mitglied besitzt.

[0064] Bei Oligonucleotid-Tags, die zur Anbringung von Markierungen an codierte Adaptoren verwendet werden, werden alle vier Nucleotide verwendet.

[0065] Die erfindungsgemäßen Oligonucleotid-Tags und ihre Komplemente werden günstigerweise auf einem automatisierten DNA-Syntheseautomaten, zum Beispiel einem Applied Biosystems, Inc. (Foster City, Kalifornien)-Modell 392 oder 394-DNA/RNA Synthesizer, unter Verwendung von Standard-Chemien wie Phosphoramidit-Chemien synthetisiert, zum Beispiel offenbart in den folgenden Quellen: Beaucage und Iyer, Tetrahedron, 48 (1992): 2223-2311; Molko et al., US-Patent 4,980,460; Koster et al. US-Patent 4,725,677; Caruthers et al., US-Patente 4,415,732; 4,458,066; und 4,973,679; und ähnliches. Alternative Chemien, zum Beispiel nicht-natürliche Rückgrat-Gruppen ergebend, wie Peptid-Nucleinsäuren (PNAs), N3'→P5'-Phosphoramidate, und ähnliches können auch verwendet werden. In einigen Ausführungsformen können Tags natürlich auftretende Nucleotide umfassen, die die Prozessierung oder Manipulation durch Enzyme erlauben, während das korrespondierende Tagkomplement nicht-natürliche Nucleotid-Analoga wie Peptid-Nucleinsäuren oder ähnliche Verbindungen, welche die Bildung von stabileren Duplexmolekülen während des Sortierens fördern, umfassen kann. Im Fall von Tags, die zur Anbringung von Markierungen an codierte Adaptoren verwendet werden, können sowohl die Oligonucleotid-Tags als auch die Tagkomplemente aus nicht-natürlichen Nucleotiden oder Analoga konstruiert werden, vorausgesetzt eine Ligierung kann entweder chemisch oder enzymatisch stattfinden.

[0066] Doppelsträngige Formen von Tags können durch getrenntes Synthetisieren der komplementären Stränge, gefolgt durch Mischen, unter Bedingungen, welche eine Bildung von Duplexmolekülen erlauben, hergestellt werden. Alternativ können doppelsträngige Tags zuerst durch Synthetisieren eines einzelsträngigen Repertoires, gebunden an eine bekannte Oligonucleotidsequenz, welche als Primerbindungsstelle dient, erzeugt werden. Der zweite Strang wird dann durch Kombinieren des einzelsträngigen Repertoires mit einem Primer und Verlängerung mit einer Polymerase synthetisiert. Dieser letztere Ansatz wird in Oliphant et al., Gene, 44 (1986): 177–183 beschrieben. Solche Duplex-Tags können dann zusammen mit Ziel-Polynucleotiden zur Sortierung und Manipulation der Ziel-Polynucleotide in Übereinstimmung mit der Erfindung in Clonierungsvektoren eingeführt werden.

[0067] Wenn Tagkomplemente verwendet werden, die aus Nucleotiden aufgebaut sind, welche verbesserte Bindungscharakteristiken besitzen, wie PNAs oder Oligonucleotid-N3'→P5'-Phosphoramidate, kann eine Sortierung durch die Bildung von D-Schleifen zwischen Tags, die natürliche Nucleotide umfassen, und deren PNA- oder Phosphoramidat-Komplementen als eine Alternative zu der "Stripping"-Reaktion, welche die 3'→5'-Exonucleaseaktivität einer DNA-Polymerase nutzt, um einen Tag einzelsträngig zu machen, durchgeführt werden.

[0068] Oligonucleotid-Tags zur Sortierung können in einem Längenbereich von 12 bis 60 Nucleotiden oder Basenpaaren liegen. Bevorzugt liegen die Oligonucleotid-Tags in einem Längenbereich von 18 bis 40 Nucleotiden oder Basenpaaren. Stärker bevorzugt liegen die Oligonucleotid-Tags in einem Längenbereich von 25 bis 40 Nucleotiden oder Basenpaaren. In Begriffen von bevorzugten und stärker bevorzugten Zahlen von Untereinheiten können diese Bereiche wie folgt ausgedrückt werden:

Tabelle III

Zahlen von Untereinheiten in Tags in bevorzugten Ausführungsformen

Monomere in Untereinheiten	Nucleotide im Oligonucleotid-Tag		
	(12 - 60)	(18 - 40)	(25 - 40)
3	4 - 20 Untereinheiten	6 - 13 Untereinheiten	8 - 13 Untereinheiten
4	3 - 15 Untereinheiten	4 - 10 Untereinheiten	6 - 10 Untereinheiten
5	2 - 12 Untereinheiten	3 - 8 Untereinheiten	5 - 8 Untereinheiten
6	2 - 10 Untereinheiten	3 - 6 Untereinheiten	4 - 6 Untereinheiten

[0069] Am stärksten bevorzugt sind Oligonucleotid-Tags zum Sortieren einzelsträngig und die spezifische Hybridisierung tritt über Watson-Crick-Paarung mit einem Tagkomplement auf.

[0070] Bevorzugt enthalten Repertoires von einzelsträngigen Oligonucleotid-Tags zum Sortieren wenigstens 100 Mitglieder; stärker bevorzugt enthalten Repertoires solcher Tags wenigstens 1000 Mitglieder; und am stärksten bevorzugt enthalten Repertoires solcher Tags wenigstens 10000 Mitglieder.

[0071] Bevorzugt enthalten Repertoires von Tagkomplementen zur Anbringung von Markierungen wenigstens 16 Mitglieder; stärker bevorzugt enthalten Repertoires solcher Tags wenigstens 64 Mitglieder. Noch stärker bevorzugt enthalten solche Repertoires von Tagkomplementen von 16 bis 1024 Mitglieder, zum Beispiel eine Menge zur Identifizierung von Nucleotiden in überhängenden Strängen von einer Länge von zwei bis fünf Nucleotiden. Am stärksten bevorzugt enthalten solche Repertoires von Tagkomplementen von 64 bis 256 Mitglieder. Repertoires der gewünschten Größen werden durch die direkte Erzeugung von Sätzen von Wörtern oder Untereinheiten der gewünschten Größe ausgewählt, zum Beispiel mit Hilfe der Computerprogramme der Anhänge I und II, oder die Repertoires werden durch die Erzeugung eines Satzes von Wörtern gebildet, welche dann in einem kombinatorischen Syntheseschema verwendet werden, um ein Repertoire der gewünschten Größe zu ergeben. Bevorzugt ist die Länge von einzelsträngigen Tagkomplementen zur Anbringung von Markierungen zwischen 8 und 20. Stärker bevorzugt ist die Länge zwischen 9 und 15.

Triplex-Tags

[0072] In Ausführungsformen, in denen die spezifische Hybridisierung über die Bildung von Triplexmolekülen auftritt, folgt die Codierung der Tag-Sequenzen den gleichen Prinzipien wie für Duplex-bildende Tags; es gibt jedoch weitere Einschränkungen für die Auswahl der Sequenzen von Untereinheiten. Im Allgemeinen ist die Assoziation des dritten Stranges über eine Bindung vom Hoogsteen-Typ am stabilsten entlang Homopyrimidin-Homopurin-Abschnitten in einem doppelsträngigen Ziel. Für gewöhnlich bilden sich Basen-Tripletts in T-A*T- oder C-G*C-Motiven (worin "–" eine Watson-Crick-Paarung anzeigt und "*" einen Hoogsteen-Typ der Bindung anzeigt); es sind jedoch auch andere Motive möglich. Zum Beispiel erlaubt die Hoogsteen-Basenpaarung abhängig von den Bedingungen und der Zusammensetzung der Stränge parallele und antiparallele Orientierungen zwischen dem dritten Strang (dem Hoogsteen-Strang) und dem Purin-reichen Strang des Duplexmoleküls, an welches der dritte Strang bindet. Es gibt ausführliche Anleitung in der Literatur zur Auswahl geeigneter Sequenzen, Orientierungen, Bedingungen, Nucleosidart (zum Beispiel ob Ribose- oder Desoxyribose-Nucleoside verwendet werden), Basenmodifizierungen (zum Beispiel methylierte Cytosine und ähnliches), um die Stabilität des Triplexmoleküls, wie in bestimmten Ausführungsformen gewünscht, zu maximieren oder anderweitig zu regulieren, zum Beispiel Roberts et al., Proc. Natl. Acad. Sci., 88 (1991): 9397–9401; Roberts et al., Science, 258 (1992): 1463–1466; Roberts et al., Proc. Natl. Acad. Sci., 93 (1996): 4320–4325; Distefano et al., Proc. Natl. Acad. Sci., 90 (1993): 1179–1183; Mergny et al., Biochemistry, 30 (1991): 9791–9798; Cheng et al., J. Am. Chem. Soc., 114 (1992): 4465–4474; Beal und Dervan, Nucleic Acids Research, 20 (1992): 2773–2776; Beal und Dervan, J. Am. Chem. Soc., 114 (1992): 4976–4982; Giovannangelli et al., Proc. Natl. Acad. Sci., 89 (1992): 8631–8635; Moser und Dervan, Science, 238 (1987): 645–650; McShan et al., J. Biol. Chem. 267 (1992): 5712-5721; Yoon et al., Proc. Natl. Acad. Sci., 89 (1992): 3840-3844; Blume et al., Nucleic Acids Research, 20 (1992): 1777–784; Thuong und Helene, Angew. Chem. Int. Ed. Engl. 32 (1993): 666–690; Escude et al., Proc. Natl. Acad. Sci., 93 (1996): 4365–4369; und ähnliches. Bedingungen zum Aneinanderla-

gern einzelsträngiger oder Duplex-Tags an ihre einzelsträngigen oder Duplex-Komplemente sind wohl bekannt, zum Beispiel Ji et al., Anal. Chem. 65 (1993): 1323–1328; Cantor et al., US-Patent 5,482,836; und ähnliches. Die Verwendung von Triplex-Tags beim Sortieren hat den Vorteil, eine "Stripping"-Reaktion mit einer Polymerase zur Freisetzung des Tags zur Aneinanderlagerung an sein Komplement nicht zu benötigen.

[0073] Bevorzugt sind erfindungsgemäße Oligonucleotid-Tags, welche eine Triplex-Hybridisierung verwenden, doppelsträngige DNA und die entsprechenden Tagkomplemente sind einzelsträngig. Stärker bevorzugt wird 5-Methylcytosin anstatt von Cytosin in den Tagkomplementen verwendet, um den Bereich der pH-Stabilität des zwischen dem Tag und seinem Komplement gebildeten Triplexmoleküls zu erweitern. Bevorzugte Bedingungen zur Bildung von Triplexmolekülen sind in den vorstehenden Quellen vollständig offenbart. Kurzgesagt findet die Hybridisierung in einer konzentrierten Salzlösung, zum Beispiel 1,0 M NaCl, 1,0 M Kaliumacetat oder ähnlichem bei einem pH-Wert unter 5,5 (oder 6,5, wenn 5-Methylcytosin verwendet wird) statt. Die Hybridisierungstemperatur hängt von der Länge und der Zusammensetzung des Tags ab; für einen 18-20-mer-Tag oder länger ist die Hybridisierung bei Raumtemperatur adäquat. Waschschritte können mit weniger konzentrierten Salzlösungen, zum Beispiel 10 mM Natriumacetat, 100 mM $MgCl_2$, pH 5,8, bei Raumtemperatur durchgeführt werden. Tags können von ihren Tagkomplementen durch Inkubation in einer ähnlichen Salzlösung bei pH 9,0 eluiert werden.

[0074] Minimal kreuz-hybridisierende Sätze von Oligonucleotid-Tags, die Triplexmoleküle bilden, können durch das Computerprogramm von Anhang II oder ähnliche Programme erzeugt werden. Ein beispielhafter Satz von doppelsträngigen 8-mer-Wörtern ist nachstehend in Großbuchstaben mit den entsprechenden Komplementen in Kleinbuchstaben aufgeführt. Ein jedes solches Wort unterscheidet sich von jedem anderen Wort in dem Satz durch drei Basenpaare.

Tabelle IV

Beispielhafter minimal kreuz-hybridisierender Satz von doppelsträngigen 8-mer-Tags

5'-AAGGAGAG 3'-TTCCTCTC 3'-ttcctctc	5'-AAAGGGGA 3'-TTTCCCCT 3'-tttccccT	5'-AGAGAAGA 3'-TCTCTTCT 3'-tctcttct	5'-AGGGGGGG 3'-TCCCCCCC 3'-tccccccc
5'-AAAAAAAA 3'-TTTTTTTT 3'-tttttttt	5'-AAGAGAGA 3'-TTCTCTCT 3'-ttctctct	5'-AGGAAAAG 3'-TCCTTTTC 3'-tccttttc	5'-GAAAGGAG 3'-CTTTCCTC 3'-ctttcctc
5'-AAAAAGGG 3'-TTTTTCCC 3'-ttttcccc	5'-AGAAGAGG 3'-TCTTCTCC 3'-tcttctcc	5'-AGGAAGGA 3'-TCCTTCCT 3'-tccttcct	5'-GAAGAAGG 3'-CTTCTTCC 3'-cttcttcc
5'-AAAGGAAG 3'-TTTCCTTC 3'-tttccttc	5'-AGAAGGAA 3'-TCTTCCTT 3'-tcttcctt	5'-AGGGGAAA 3'-TCCCCTTT 3'-tccccttt	5'-GAAGAGAA 3'-CTTCTCTT 3'-cttctctt

Tabelle V

Repertoiregröße von verschiedenen doppelsträngigen Tags, die Triplexmoleküle mit ihren Tagkomplementen bilden

Oligonucleotid-Wortlänge	Nucleotidunterschied zwischen Oligonucleotiden eines minimal kreuzhybridisierenden Satzes	Maximale Größe eines minimal kreuzhybridiserenden Satzes	Größe eines Repertoires mit vier Wörtern	Größe eines Repertoires mit fünf Wörtern
4	2	8	4096	3.2×10^4
6	3	8	4096	3.2×10^4
8	3	16	6.5×10^4	1.05×10^6
10	5	8	4096	
15	5	92		
20	6	768		
20	7	484		
20	8	189		
20	9	30		

Synthese und Struktur von Adaptoren

[0075] Die codierten Adaptoren und Abspaltungsadaptoren werden günstigerweise auf automatisierten DNA-Syntheseautomaten unter Verwendung von Standard-Chemien wie Phosphoramidit-Chemie, zum Beispiel offenbart in den folgenden Quellen: Beaucage und Iyer, Tetrahedron, 48 (1992): 2223–2311; Molko et al., US-Patent 4,980,460; Koster et al., US-Patent 4,725,677; Caruthers et al., US-Patente 4,415,732; 4,458,066; und 4,973,679; und ähnliches synthetisiert. Alternative Chemien, zum Beispiel nicht-natürliche Rückgrat-Gruppen ergebend, wie Phosphorthioate, Phosphoramidate und ähnliches können auch verwendet werden, vorausgesetzt, dass die sich ergebenden Oligonucleotide mit den in einer speziellen Ausführungsform verwendeten Ligierungs- und/oder Abspaltungsreagenzien kompatibel sind. Typischerweise werden nach der Synthese der komplementären Stränge die Stränge kombiniert, um einen doppelsträngigen Adaptor zu bilden. Der überhängende Strang eines codierten Adaptors kann als ein Gemisch synthetisiert werden, so dass jede mögliche Sequenz in dem überhängenden Abschnitt repräsentiert ist. Solche Gemische werden einfach unter Verwendung wohl bekannter Techniken synthetisiert, wie zum Beispiel in Telenius et al., Genomics, 13 (1992): 718–725; Welsh et al., Nucleic Acids Research, 19 (1991): 5275–5279; Grothues et al., Nucleic Acids Research, 21 (1993): 1321–1322; Hartley, europäische Patentanmeldung 90304496.4 (EP-Veröffentlichungsnummer 395398); und ähnlichem offenbart. Im Allgemeinen erfordern diese Techniken, wo man multiple Nucleotide einführen möchte, einfach während der Kopplungsschritte die Anwendung von Gemischen der aktivierten Monomere an den wachsenden Oligonucleotiden. Wie vorstehend diskutiert, kann es in einigen Ausführungsformen wünschenswert sein, die Komplexität der Adaptoren zu reduzieren. Dies kann unter Verwendung die Komplexität reduzierender Analoga wie Desoxyinosin, 2-Aminopurin oder ähnlichem bewerkstelligt werden, wie zum Beispiel in Kong Thoo Lin et al., Nucleic Acids Research, 20: 5149–5152, oder durch US-Patent 5,002,867, Nichols et al., Nature, 369 (1994): 492–493; und ähnlichem gelehrt.

[0076] In einigen Ausführungsformen kann es wünschenswert sein, die codierten Adaptoren oder Abspaltungsadaptoren als einzelnes Polynucleotid zu synthetisieren, welches selbst-komplementäre Regionen enthält. Nach der Synthese wird es den selbst-komplementären Regionen erlaubt sich aneinanderzulagern, um einen Adaptor mit einem überhängenden Strang an einem Ende und einer einzelsträngigen Schleife am anderen Ende zu bilden. Bevorzugt kann in solchen Ausführungsformen die Schleifen-Region von etwa drei bis 10 Nucleotide oder andere vergleichbare Verknüpfungseinheiten, zum Beispiel Alkylether-Gruppen, so wie im US-Patent 4,914,210 offenbart, umfassen. Viele Techniken zur Anfügung reaktiver Gruppen an die Basen oder Internucleosid-Bindungen zur Markierung sind, wie in den nachstehend zitierten Quellen diskutiert, verfügbar.

[0077] Wenn in der Erfindung übliche Ligasen verwendet werden, wie dies vollständiger nachstehend be-

schrieben ist, kann in einigen Ausführungsformen das 5'-Ende des Adaptors phosphoryliert sein. Ein 5'-Monophosphat kann an ein zweites Oligonucleotid entweder chemisch oder enzymatisch mit einer Kinase angehängt werden, zum Beispiel Sambrook et al., Molecular Cloning: A Laboratory Manual, zweite Auflage (Cold Spring Harbor Laboratory, New York, 1989). Die chemische Phosphorylierung wird durch Horn und Urdes, Therahedron Lett., 27 (1986): 4705, beschrieben und Reagenzien zur Durchführung der offenbarten Protokolle sind kommerziell erhältlich, zum Beispiel 5'-Phosphate-ON™ von Clontech Laboratories (Palo Alto, Kalifornien).

[0078] Codierte Adaptoren, welche im erfindungsgemäßen Verfahren verwendet werden können, können einige Ausführungsformen haben, zum Beispiel abhängig davon, ob einzel- oder doppelsträngige Tags verwendet werden, ob verschiedene Tags verwendet werden, ob ein 5'-überhängender Strang oder ein 3'-überhängender Strang verwendet wird, ob eine 3'-Blockierungsgruppe verwendet wird und ähnliches. Formeln für einige Ausführungsformen von codierten Adaptoren werden nachstehend gezeigt. Bevorzugte Strukturen für codierte Adaptoren, die einen einzelsträngigen Tag verwenden, sind wie folgt:

$$\begin{array}{l} 5'\text{-}p(N)_n(N)_r(N)_s(N)_q(N)_t\text{-}3' \\ \qquad\quad z(N')_r(N')_s(N')_q\text{-}5' \end{array}$$

oder

$$\begin{array}{l} \qquad\quad p(N)_r(N)_s(N)_q(N)_t\text{-}3' \\ 3'\text{-}z(N)_n(N')_r(N')_s(N')_q\text{-}5' \end{array},$$

worin N ein Nucleotid ist und N' sein Komplement ist, p eine Phosphatgruppe ist, z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist, n eine ganze Zahl zwischen einschließlich 2 und 6 ist, r eine ganze Zahl größer als oder gleich 0 ist, s eine ganze Zahl ist, welche entweder zwischen 4 und 6 ist, wenn der codierte Adaptor eine Nucleaseerkennungsstelle hat, oder 0 ist, wenn keine Nucleaseerkennungsstelle vorhanden ist, q eine ganze Zahl größer oder gleich 0 ist und t eine ganze Zahl zwischen einschließlich 8 und 20 ist. Stärker bevorzugt ist n 4 oder 5 und t ist zwischen einschließlich 9 und 15. Wenn ein codierter Adaptor eine Nucleaseerkennungsstelle enthält, wird der Bereich von "r" Nucleotidpaaren ausgewählt, so dass eine vorbestimmte Zahl von Nucleotiden von dem Ziel-Polynucleotid abgespalten wird, wenn die Nuclease, welche die Stelle erkennt, angewendet wird. Die Größe von "r" in einer bestimmten Ausführungsform hängt von der Reichweite der Nuclease (wie der Begriff in US-Patent 5,599,675 und WO 95/27080 definiert ist) und der Zahl von Nucleotiden, die vom Ziel-Polynucleotid abgespalten werden sollen, ab. Bevorzugt ist r zwischen 0 und 20; stärker bevorzugt ist r zwischen 0 und 12. Der Bereich von "q" Nucleotidpaaren ist ein Spacerabschnitt zwischen der Nucleaseerkennungsstelle und der Tag-Region der codierten Sonde. Die Region von "q" Nucleotiden kann weiterhin Nucleaseerkennungsstellen, Markierungs- oder Signal-erzeugende Einheiten oder ähnliches enthalten. Das einzelsträngige Oligonucleotid von "t" Nucleotiden ist ein "t-mer"-Oligonucleotid-Tag ausgewählt aus einem minimal kreuz-hybridisierenden Satz.

[0079] Die 3'-Blockierungsgruppe "z" kann eine Vielfalt von Formen haben und kann fast jede chemische Einheit einschließen, die eine Ligierung verhindert und die keine anderen Schritten des Verfahrens, zum Beispiel die Entfernung des 3'-blockierten Stranges, Ligierung oder ähnliches, stört. Beispielhafte 3'-Blockierungsgruppen schließen Wasserstoff (das heißt 3'-Desoxy), Phosphat, Phosphorothioat, Acetyl und ähnliches ein, sind aber nicht auf diese begrenzt. Wegen der Einfachheit der Addition der Gruppe während der Synthese des 3'-blockierten Stranges und der Einfachheit bei der Entfernung der Gruppe mit einer Phosphatase, um den Strang fähig zur Ligierung mit einer Ligase zu machen, ist die 3'-Blockierungsgruppe bevorzugt ein Phosphat. Ein Oligonucleotid, welches ein 3'-Phosphat besitzt, kann unter Verwendung des in Kapitel 12 von Eckstein, Herausgeber, Oligonucleotides and Analogues: A Practical Approach (IRL Press, Oxford, 1991) beschriebenen Protokolls synthetisiert werden.

[0080] Weitere 3'-Blockierungsgruppen sind aus den Chemien, welche für umkehrbare Kettenabbruch-Nucleotide in Base-um-Base-Sequenzierungsschemata entwickelt wurden, zum Beispiel offenbart in den folgenden Quellen: Cheeseman, US-Patent 5,302,509; Tsien et al., internationale Anmeldung WO 91/06678; Canard et al., Gene, 148 (1994): 1–6; und Metzker et al., Nucleic Acids Research, 22 (1994): 4259–4267; erhältlich. Allgemein gesagt erlauben diese Chemien die chemische oder enzymatische Entfernung von spezifischen Blockierungsgruppen (die für gewöhnlich eine angehängte Markierung haben), um ein freies Hydroxyl am 3'-Ende

eines Starter-Stranges zu erzeugen.

[0081] Bevorzugt ist z, wenn es eine 3'-Blockierungsgruppe ist, eine Phosphatgruppe und der doppelsträngige Abschnitt der Adaptoren enthält eine Nucleaseerkennungsstelle einer Nuclease, deren Erkennungsstelle von ihrer Schneidestelle verschieden ist.

[0082] Wenn doppelsträngige Oligonucleotid-Tags verwendet werden, die spezifisch mit einzelsträngigen Tagkomplementen hybridisieren, um Triplex-Strukturen zu bilden, haben erfindungsgemäße codierte Tags bevorzugt die folgende Form:

$$\begin{array}{r} 5'\text{-}p(N)_n(N)_r(N)_s(N)_q(N)_t\text{-}3' \\ z(N')_r(N')_s(N')_q(N')_t\text{-}5' \end{array}$$

oder

$$\begin{array}{r} p(N)_r(N)_s(N)_q(N)_t\text{-}3' \\ 3'\text{-}z(N)_n(N')_r(N')_s(N')_q(N')_t\text{-}5' \end{array} ,$$

wobei N, N', p, q, r, s, z und n wie vorstehend definiert sind. Bevorzugt ist in dieser Ausführungsform t eine ganze Zahl im Bereich von 12 bis 24.

[0083] Klarerweise gibt es zusätzliche Strukturen, welche Elemente des grundlegenden Entwurfsatzes, welcher vorstehend dargestellt ist, enthalten, die für den Durchschnittsfachmann offensichtlich sein werden. Zum Beispiel schließen erfindungsgemäße codierte Adaptoren Ausführungsformen mit multiplen Tags ein, so wie die folgenden:

$$\begin{array}{r} 5'\text{-}p(N)_n(N)_r(N)_s(N)_q(N)_{t1} \ldots (N)_{tk}\text{-}3' \\ z(N')_r(N')_s(N')_q(N')_{t1} \ldots (N')_{tk}\text{-}5' \end{array}$$

oder

$$\begin{array}{r} p(N)_r(N)_s(N)_q(N)_{t1} \ldots (N)_{tk}\text{-}3' \\ 3'\text{-}z(N)_n(N')_r(N')_s(N')_q(N')_{t1} \ldots (N')_{tk}\text{-}5' \end{array} ,$$

wobei der codierte Adaptor k doppelsträngige Tags einschließt. Bevorzugt ist $t_1 = t_2 = \ldots t_k$ und k entweder 1, 2 oder 3.

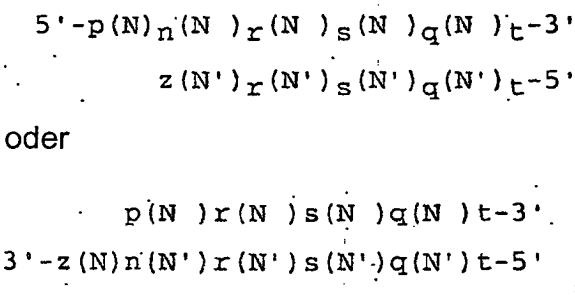
[0084] Die vorliegende Erfindung betrifft auch eine Zusammensetzung von Material, umfassend eine Vielzahl von doppelsträngigen Oligonucleotidadaptoren, wobei die Adaptoren die Form haben:

$$\begin{array}{r} 5'\text{-}p(N)_n(N)_r(N)_s(N)_q(N)_t\text{-}3' \\ z(N')_r(N')_s(N')_q\text{-}5' \end{array}$$

oder

$$\begin{array}{r} p(N)r(N)s(N)q(N)t\text{-}3' \\ 3'\text{-}z(N)n(N')r(N')s(N')q\text{-}5' \end{array} ,$$

wobei jedes $(N)_t$ ein einzigartiger, einzelsträngiger Oligonucleotid-Tag ist und ausgewählt ist aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden, so dass sich jedes Oligonucleotid des Satzes von jedem anderen Oligonucleotid des Satzes in mindestens zwei Nucleotiden unterscheidet;
oder die Form:



oder

wobei jedes $\binom{(N)_t}{(N')_t}$ ein einzigartiger, doppelsträngiger Oligonucleotid-Tag ist, ausgewählt aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden, so dass jedes Oligonucleotid des Satzes sich von jedem anderen Oligonucleotid des Satzes in mindestens zwei Basenpaaren unterscheidet;
wobei: N ein Nucleotid und N' sein Komplement ist,
p eine Phosphatgruppe ist,
z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist,
n eine ganze Zahl zwischen einschließlich 2 und 6 ist,
r eine ganze Zahl zwischen einschließlich 0 und 18 ist,
s eine ganze Zahl zwischen einschließlich 4 und 6 ist,
der Adaptor schließt in einem doppelsträngigen Abschnitt, getrennt von dem Oligonucleotid-Tag, eine Nucleaseerkennungsstelle einer Nuclease ein, deren Erkennungsstelle separat von ihrer Schneidestelle ist,
q eine ganze Zahl größer oder gleich 0 ist und
t eine ganze Zahl größer oder gleich 8 ist.

Markierung von Tagkomplementen

[0085] Die erfindungsgemäßen Tagkomplemente können in einer Vielfalt von Arten zur Decodierung von Oligonucleotid-Tags markiert werden, einschließlich des direkten oder indirekten Anbringens von radioaktiven Einheiten, fluoreszierenden Einheiten, colorimetrischen Einheiten, chemilumineszenten Einheiten und ähnlichem. Viele umfassende Übersichten über Verfahrensweisen zur Markierung von DNA und der Konstruktion von DNA-Adaptoren stellen Anleitungen bereit, die zur Konstruktion von erfindungsgemäßen Adaptoren anwendbar sind. Solche Übersichten schließen Matthews et al., Anal. Biochem., Bd. 169, Seiten 1–25 (1988); Haugland, Handbook of Flourescent Probes and Research Chemicals (Molecular Probes, Inc., Eugene, 1992); Keller und Manak, DNA Probes, zweite Auflage (Stockton Press, New York, 1993); und Eckstein, Herausgeber, Oligonucleotides and Analogues: A Practical Approach (IRL Press, Oxford, 1991); Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26 (1991): 227–259; und ähnliche ein. Viel mehr spezielle Verfahrensweisen, die auf die Erfindung anwendbar sind, werden in der folgenden Auswahl von Quellen offenbart: Fung et al., US-Patent 4,757,141; Hobbs Jr., et al. US-Patent 5,151,507; Cruickshank, US-Patent 5,091,519; (Synthese von funktionalisierten Oligonucleotiden für das Anbringen von Reporter-Gruppen); Jablonski et al., Nucleic Acids Research, 14 (1986): 6115–6128 (Enzym-Oligonucleotid-Konjugate); Ju et al., Nature Medicine, 2 (1996): 246–249; und Urdea et al., US-Patent 5,124,246 (verzweigte DNA). Stellen für das Anbringen von Markierungseinheiten sind nicht kritisch, vorausgesetzt, dass solche Markierungen die Ligierungs- und/oder Abspaltungsschritte nicht stören.

[0086] Bevorzugt werden ein oder mehrere Fluoreszenzfarbstoffe als Markierungen für die Tagkomplemente verwendet, zum Beispiel wie offenbart durch Menchen et al., US-Patent 5,188,934; Bergot et al., PCT-Anmeldung PCT/US 90/05565 (WO 91/05060). Wie hier verwendet, bedeutet der Begriff "Fluoreszenzsignal erzeugende Einheit" ein Signalmittel, welches Informationen durch die Fluoreszenzabsorptions- und/oder Emissionseigenschaften von einem oder mehreren Molekülen überträgt. Solche Fluoreszenzeigenschaften schließen die Fluoreszenzintensität, Fluoreszenzüberlebenszeit, Charakteristiken der Emissionsspektren, Energietransfer und ähnliches ein.

Ligierung von Adaptoren und Verhinderung der Selbst-Ligierung

[0087] In Übereinstimmung mit der bevorzugten erfindungsgemäßen Ausführungsform werden Abspaltungsadaptoren an die Enden von Ziel-Polynucleotiden ligiert, um solche Enden für die schließliche Ligierung von codierten Adaptoren vorzubereiten. Bevorzugt wird die Ligierung enzymatisch unter Verwendung einer Ligase in einem Standardprotokoll durchgeführt. Viele Ligasen sind bekannt und sind für die Verwendung in der Erfindung geeignet, zum Beispiel Lehmann, Science, 186 (1974): 790–797; Engler et al., DNA Ligases, Seiten 3–30 in Boyer, Herausgeber, The Enzymes, Bd. 15B (Academic Press, New York, 1982); und ähnliches. Bevorzugte

Ligasen schließen T4-DNA-Ligase, T7-DNA-Ligase, E. coli-DNA-Ligase, Taq-Ligase, Pfu-Ligase und Tth-Ligase ein. Protokolle für deren Verwendung sind wohl bekannt, zum Beispiel Sambrook et al. (vorstehend zitiert); Barany, PCR Methods and Applications, 1 (1991): 5–16; Marsh et al., Strategies, 5 (1992): 73–76, und ähnliches. Im Allgemeinen erfordern Ligasen, dass eine 5'-Phosphatgruppe für die Ligierung an das 3'-Hydroxyl eines anstoßenden Stranges anwesend ist. Dies wird günstigerweise wenigstens für einen Strang des Ziel-Polynucleotids durch Auswahl einer Nuclease, welche ein 5'-Phosphat hinterlässt, wie zum Beispiel Fok I, bereitgestellt.

[0088] Ein besonderes Problem kann beim Arbeiten mit entweder Polynucleotid-Enden oder Adaptoren, welche zur Selbst-Ligierung fähig sind, so wie in Fig. 2 dargestellt, wo die überhängenden vier-Nucleotid-Stränge der verankerten Polynucleotide miteinander komplementär (**114**) sind, auftreten. Dieses Problem ist in Ausführungsformen besonders schwerwiegend, wo die zu untersuchenden Polynucleotide (**112**) den Adaptoren als gleichförmige Populationen von identischen Polynucleotiden, die an einen Festphasenträger gebunden sind (**110**), angeboten werden. In diesen Situationen können sich die freien Enden der verankerten Polynucleotide gewunden sein, um perfekt passende Duplexmoleküle miteinander zu bilden (**116**). Wenn die 5'-Stränge der Enden phosphoryliert sind, werden die Polynucleotide in der Anwesenheit einer Ligase leicht ligiert. Ein analoges Problem besteht auch für doppelsträngige Adaptoren. Wenn nämlich deren 5'-Stränge phosphoryliert sind, kann der 5'-Strang des einen Adaptors an das freie 3'-Hydroxyl eines anderen Adaptors ligiert werden, wann immer die Nucleotidsequenzen von deren überhängenden Strängen komplementär sind. Wenn Selbst-Ligierung auftritt, sind weder die überhängenden Stränge der Adaptoren noch der Ziel-Polynucleotide für Untersuchung oder Verarbeitung verfügbar. Dies führt seinerseits zu einem Verlust oder einem Verschwinden von in Reaktion auf korrekte Ligierung von Adaptoren an Ziel-Polynucleotide erzeugten Signalen. Da die Wahrscheinlichkeit eines palindromischen Vier-mers in einer zufälligen Sequenz die gleiche ist wie die Wahrscheinlichkeit eines wiederholten Paares von Nucleotiden (6,25%), haben auf Adaptoren basierende Verfahren der de novo-Sequenzierung einen hohen Erwartungswert des Versagens wegen Selbst-Ligierung nach einigen Zyklen. Wenn dies auftritt, wird eine weitere Untersuchung des Polynucleotids unmöglich.

[0089] Die vorstehenden Probleme können durch Umsetzung der Erfindung mit den folgenden Schritten, welche für eine bevorzugte Ausführungsform in Fig. 3A dargestellt sind, angegangen werden. (a) Ligieren (**120**) eines codierten Adaptors an das Ende des Polynucleotids (**122**), wobei das Ende des Polynucleotids ein dephosphoryliertes 5'-Hydroxyl besitzt und das Ende des zu ligierenden codierten Adaptors (**124**) einen ersten Strang (**126**) und einen zweiten Strang (**128**) besitzt und der zweite Strang des codierten Adaptors eine 3'-Blockierungsgruppe (**130**) besitzt; (b) Entfernen der 3'-Blockierungsgruppe des zweiten Stranges nach Ligierung, zum Beispiel durch Waschen (**132**) oder durch enzymatisches oder chemisches Entfernen der Gruppe in situ, zum Beispiel durch Behandlung mit einer Phosphatase, wenn die Blockierungsgruppe ein Phosphat ist; (c) Phosphorylieren (**134**) des 5'-Hydroxyls des Polynucleotids; (d) Ligieren (**136**) eines zweiten Stranges (**142**), welcher eine entblockierte 3'-Einheit besitzt, um den codierten Adaptor zu regenerieren (**138**); und (e) Identifizieren (**144**) eines oder mehrerer Nucleotide am Ende des Polynucleotids, durch die Identität des daran ligierten codierten Adaptors, zum Beispiel über ein fluoreszierend markiertes (**140**) Tagkomplement. Die codierten Adaptoren und Ziel-Polynucleotide können für die Ligierung entweder einzeln oder als Gemische kombiniert werden. Zum Beispiel kann eine einzelne Art von Adaptor, welche eine definierte Sequenz besitzt, mit einer einzelnen Art von Polynucleotid kombiniert werden, welche eine gemeinsame (und vielleicht unbekannte) Nucleotidsequenz besitzt; oder eine einzelne Art von Adaptor, welche eine definierte Sequenz besitzt, kann mit einem Gemisch von Polynucleotiden kombiniert werden, wie eine Vielzahl von gleichförmigen Populationen von identischen Polynucleotiden, welche an verschiedene Festphasenträger im gleichen Reaktionsgefäß gebunden ist, zum Beispiel beschrieben durch Brenner et al., internationale Anmeldung PCT/US 96/09513 (WO 96/41011); oder ein Gemisch von codierten Adaptoren, insbesondere Gemische, welche verschiedene Nucleotidsequenzen in ihren überhängenden Strängen besitzen, kann mit einer einzelnen Art von Polynucleotid kombiniert werden; oder ein Gemisch von codierten Adaptoren kann mit einem Gemisch von Polynucleotiden kombiniert werden. Wenn der Begriff "Adaptor" oder "codierter Adaptor" in der Einzahl verwendet wird, so soll seine Bedeutung ein Gemisch von Adaptoren, welche verschiedene Sequenzen von überhängenden Strängen besitzen, genauso wie eine einzelne Art von Adaptor, welche die gleiche Sequenz des überhängenden Stranges besitzt, in einer Weise umfassen, welche der Verwendung des Begriffs "Sonde" analog ist.

[0090] Neben der Entfernung durch Schmelzen kann eine 3'-Desoxygruppe von einem zweiten Strang durch eine Polymerase-"Austausch"-Reaktion, offenbart in Kuijper et al., Gene, 112 (1992): 147–155; Aslanidis et al., Nucleic Acids Research, 18 (1990): 6069–6074; und ähnlichen Quellen, entfernt werden. Kurzgesagt kann die 5'→3'-Exonucleaseaktivität der T4-DNA-Polymerase und ähnlicher Enzyme verwendet werden, um Nucleotide in einem Starterstrang durch deren Triphosphat-Gegenpart in Lösung, zum Beispiel Kuijper et al. (vorstehend zitiert), auszutauschen. Somit kann mit solch einer Reaktion ein 3'-Didesoxynucleotid durch ein 2'-Deso-

xy-3'-Hydroxynucleotid aus einem Reaktionsgemisch ausgetauscht werden, was nach einer Behandlung mit Polynucleotidkinase den zweiten Strang mit einem Ziel-Polynucleotid ligierbar macht.

[0091] Eine bevorzugte Ausführungsform, welche Zyklen von Ligierung und Abspaltung verwendet, umfasst die folgenden Schritte: (a) Ligieren (**220**) eines codierten Adaptors an das Ende des Polynucleotids (**222**), wobei das Ende des Polynucleotids ein dephosphoryliertes 5'-Hydroxyl besitzt, das Ende des zu ligierenden doppelsträngigen Adaptors (**224**) einen ersten Strang (**226**) und einen zweiten Strang (**228**) besitzt, der zweite Strang des doppelsträngigen Adaptors eine 3'-Blockierungsgruppe besitzt (**230**) und der doppelsträngiger Adaptor eine Nucleaseerkennungsstelle (**250**) einer Nuclease, deren Erkennungsstelle von ihrer Schneidestelle verschieden ist, besitzt (**250**); (b) Entfernen der 3'-Blockierungsgruppe nach Ligierung, zum Beispiel durch Wegwaschen des zweiten Stranges (**232**); (c) Phosphorylieren (**234**) des 5'-Hydroxyls des Polynucleotids; (d) Ligieren (**236**) eines zweiten Stranges (**242**), welcher eine entblockierte 3'-Einheit besitzt, um den doppelsträngigen Adaptor (**238**) und die Nucleaseerkennungsstelle (**250**) zu regenerieren; (e) Identifizieren (**244**) von einem oder mehreren Nucleotiden am Ende des Polynucleotids durch die Identität des daran ligierten Adaptors; (f) Abspalten (**252**) des Polynucleotids mit einer Nuclease, welche die Erkennungsstelle erkennt, so dass das Polynucleotid um eines oder mehrere Nucleotide gekürzt wird, wobei die Erkennungsstelle in dem dargestellten Adaptor positioniert ist (**224**), so dass die Abspaltung (**254**) zwei Nucleotide von dem Polynucleotid entfernt (**222**); (g) Dephosphorylieren (**256**) des 5'-Endes des Polynucleotids; und (h) Wiederholen der Schritte (**258**) (a) bis (g).

[0092] Typischerweise werden die Enden von zu untersuchenden Polynucleotiden vor der Ligierung durch Spaltung mit einer oder mehreren Restriktionsendonucleasen, welche vorbestimmte Spaltungen erzeugen, welche für gewöhnlich 3'- oder 5'-überhängende Stränge besitzen, das heißt "klebrige" Enden, vorbereitet. Solche Spaltungen belassen für gewöhnlich die 5'-Stränge phosphoryliert. Bevorzugt werden diese 5'-phosphorylierten Enden durch Behandlung mit einer Phosphatase wie alkalische Kälberdarmphosphatase oder ähnlichen Enzymen unter Verwendung von Standardprotokollen, zum Beispiel wie beschrieben in Sambrook et al., Molecular Cloning, zweite Auflage (Cold Spring Harbor Labroratory, New York, 1989) dephosphoryliert. Durch Entfernung der 5'-Phosphate werden die Ziel-Polynucleotide unfähig gemacht, in der Anwesenheit einer Ligase ligiert zu werden. Der Schritt der Dephosphorylierung hinterlässt bevorzugt ein freies 5'-Hydroxyl.

Bevorzugte Nucleasen

[0093] "Nuclease" bedeutet, so wie der Begriff in Übereinstimmung mit der Erfindung verwendet wird, jedes Enzym, jede Kombination von Enzymen oder anderen chemischen Reagenzien oder Kombinationen chemischer Reagenzien und Enzyme, die, wenn sie, wie nachstehend vollständiger diskutiert, an einem ligierten Komplex angewendet werden, den ligierten Komplex spalten, um einen vergrößerten Adaptor und ein gekürztes Ziel-Polynucleotid zu erzeugen. Eine erfindungsgemäße Nuclease muss nicht ein einzelnes Protein sein oder ausschließlich aus einer Kombination von Proteinen bestehen. Ein Schlüsselmerkmal der Nuclease oder der Kombination von Reagenzien, die als Nuclease verwendet werden, ist, dass ihre Schneidestelle verschieden von ihrer Erkennungsstelle ist. Der Abstand zwischen der Erkennungsstelle einer Nuclease und ihrer Schneidestelle wird hier als ihre "Reichweite" bezeichnet werden. Durch Konvention wird die "Reichweite" durch zwei ganze Zahlen definiert, welche die Zahl von Nucleotiden zwischen der Erkennungsstelle und der hydrolysierten Phosphodiesterbindung von jedem Strang angeben. Zum Beispiel werden die Erkennungs- und Spaltungseigenschaften von FokI typischerweise als "GGATG (9/13)" dargestellt, weil es eine doppelsträngige DNA wie folgt erkennt und schneidet (SEQ ID NO: 2):

```
5'-... NNGGATGNNNNNNNNN          NNNNNNNNNN ...
3'-... NNCCTACNNNNNNNNNNNNN          NNNNNN ...
                                              ,
```

wobei die Nucleotide in Fett-Schrift die Erkennungsstelle von FokI sind und die N's beliebige Nucleotide und ihre Komplemente sind.

[0094] Es ist wichtig, dass die Nuclease das Ziel-Polynucleotid nur spaltet, nachdem nie einen Komplex mit seiner Erkennungsstelle gebildet hat; und die Nuclease hinterlässt nach der Spaltung bevorzugt einen überhängenden Strang auf dem Ziel-Polynucleotid.

[0095] Bevorzugt sind in dieser Erfindung verwendete Nucleasen natürliche Protein-Endonucleasen, (i) deren Erkennungsstelle von ihrer Schneidestelle verschieden ist und (ii) deren Spaltung einen überhängenden Strang auf dem Ziel-Polynucleotid ergibt. Am stärksten bevorzugt werden Klasse-II-Restriktionsendonucleasen als erfindungsgemäße Nucleasen verwendet, zum Beispiel wie beschrieben in Szybalski et al., Gene, 100

(1991): 13–26; Roberts et al., Nucleic Acids Research, 21 (1993): 3125–3137; und Livak und Brenner, US-Patent 5,093,245. Beispielhafte Klasse-II-Nucleasen für die erfindungsgemäße Verwendung schließen AlwXI, BsmAI, BbvI, BsmFI, StsI, HgaI, BscAI, BbvII, BcefI, Bce85I, BccI, BcgI, BsaI, BsgI, BspMI, Bst71I, EarI, Eco57I, Esp3I, FauI, FokI, GsuI, HphI, MboII, MmeI, RleAI, SapI, SfaNI, TaqII, Tth111II, Bco5I, BpuAI, FinI, BsrDI und Isoschizomere davon ein. Bevorzugte Nucleasen schließen BbvI, FokI, HgaI, EarI und SfaNI ein. BbvI ist die am stärksten bevorzugte Nuclease.

[0096] Bevorzugt wird vor den Nuclease-Spaltungsschritten, für gewöhnlich beim Beginn einer Sequenzierungsoperation, das Ziel-Polynucleotid behandelt, um die Erkennungsstellen und/oder Schneidestellen der verwendeten Nuclease zu blockieren. Dies verhindert die unerwünschte Spaltung des Ziel-Polynucleotids wegen des zufälligen Vorkommens von Nucleaseerkennungsstellen an internen Stellen in dem Ziel-Polynucleotid. Die Blockierung kann in einer Vielfalt von Weisen einschließlich der Methylierung und der Behandlung durch Sequenz-spezifische Aptamere, DNA-Bindungsproteine oder Oligonucleotide, welche Triplexmoleküle bilden, erreicht werden. Wann immer natürliche Protein-Endonucleasen verwendet werden, können Erkennungsstellen bequem durch Methylierung des Ziel-Polynucleotids mit der zugehörigen Methylase der verwendeten Nuclease blockiert werden. Das heißt, für die meisten, wenn nicht alle bakteriellen Typ-II-Restriktionsendonucleasen existiert eine so genannte "zugehörige" Methylase, die deren Erkennungsstelle methyliert. Viele solche Methylase werden in Roberts et al. (vorstehend zitiert) und Nelson et al., Nucleic Acids Research, 21 (1993): 3139— 3154, offenbart und sind auch von einer Vielfalt von Quellen kommerziell erhältlich, insbesondere New England Biolabs (Beverly, MA). Alternativ können, wenn ein PCR-Schritt bei der Vorbereitung des Ziel-Polynucleotids zum Sequenzieren verwendet wird, 5-Methylcytosintriphosphate während der Vervielfältigung verwendet werden, so dass das natürliche Cytosin im Amplifikat durch methyliertes Cytosin ersetzt wird. Dieser letztere Ansatz hat den zusätzlichen Vorteil, die Notwendigkeit zu eliminieren, ein an einen Festphasenträger gebundenes Ziel-Polynucleotid mit einem anderen Enzym zu behandeln.

[0097] Klarerweise kann der Durchschnittsfachmann Merkmale der vorstehend ausgeführten Ausführungsformen kombinieren, um noch weitere Ausführungsformen zu entwerfen, die mit der Erfindung übereinstimmen, aber nicht ausdrücklich vorstehend dargestellt sind.

[0098] Es wird eine Vielzahl von Kits bereitgestellt, um verschiedene Ausführungsformen der Erfindung durchzuführen. Im Allgemeinen schließen erfindungsgemäße Kits codierte Adaptoren, Abspaltungsadaptoren und markierte Tagkomplemente ein. Kits schließen weiterhin die Nucleasereagenzien, die Ligierungsreagenzien und Anweisungen zur Durchführung der spezifischen erfindungsgemäßen Ausführungsform ein. In Ausführungsformen, welche natürliche Protein-Endonucleasen und Ligasen verwenden, können Ligasepuffer und Nucleasepuffer eingeschlossen sein. In einigen Fällen können diese Puffer identisch sein. Solche Kits können auch eine Methylase und ihren Reaktionspuffer einschließen. Bevorzugt schließen Kits auch einen oder mehrere Festphasenträger, zum Beispiel Mikropartikel, die Tagkomplemente zum Sortieren und Verankern von Ziel-Polynucleotiden tragen, ein.

Anbringen von Tags an Polynucleotide zur Sortierung auf Festphasenträger

[0099] Ein wichtiger Aspekt der Erfindung ist das Sortieren und Anbringen von Populationen von Polynucleotiden, zum Beispiel aus einer cDNA-Genbank, an Mikropartikel oder getrennte Regionen auf einem Festphasenträger, so dass an jedem Mikropartikel oder an jeder Region im Wesentlichen nur eine Art von Polynucleotid angebracht ist. Dieses Ziel wird dadurch bewerkstelligt, dass sichergestellt wird, das im Wesentlichen alle verschiedenen Polynucleotide verschiedene Tags angehängt haben. Diese Bedingung wird ihrerseits dadurch bewerkstelligt, dass eine Probe aus der vollständigen Gesamtheit von Tag-Polynucleotid-Konjugaten zur Untersuchung gezogen wird. (Es ist akzeptabel, dass identische Polynucleotide verschiedene Tags haben, da dies nur dazu führt, dass das gleiche Polynucleotid zweimal an zwei verschiedenen Stellen bearbeitet oder untersucht wird.) Die Probennahme kann, nachdem die Tags an die Polynucleotide angehängt worden sind, entweder offen durchgeführt werden – zum Beispiel durch das Abnehmen eines kleinen Volumens von einem größeren Gemisch – sie kann inhärent als ein Nebeneffekt des zum Verarbeiten der Polynucleotide und Tags verwendeten Verfahrens durchgeführt werden oder die Probennahme kann offen und inhärent als Teil der Verarbeitungsschritte durchgeführt werden.

[0100] Bevorzugt wird bei der Konstruktion einer cDNA-Genbank, wo im Wesentlichen alle verschiedenen cDNAs verschiedene Tags haben, ein Tag-Repertoire verwendet, dessen Komplexität oder dessen Zahl von unterschiedlichen Tags die gesamte Zahl von mRNAs, welche aus einer Zell- oder Gewebeprobe extrahiert worden sind, in großem Maße überschreitet. Bevorzugt ist die Komplexität des Tag-Repertoires wenigstens zehnmal so groß wie jene der Polynucleotidpopulation; und stärker bevorzugt ist die Komplexität des Tag-Reper-

toires wenigstens 100-mal so groß wie jene der Polynucleotidpopulation. Nachstehend wird ein Protokoll zur Konstruktion einer cDNA-Genbank unter Verwendung eines Primer-Gemisches, welches ein vollständiges Repertoire von beispielhaften neun-Wort-Tags enthält, offenbart. Solch ein Gemisch von Tag-enthaltenden Primern hat eine Komplexität von 8^9 oder etwa $1{,}34 \times 10^8$. Wie durch Winslow et al., Nucleic Acids Research, 19 (1991): 3251–3253, angedeutet, kann mRNA für die Konstruktion einer Genbank aus so wenig wie 10–100 Säugerzellen extrahiert werden. Da eine einzelne Säugerzellen etwa 5×10^5 Kopien von mRNA-Molekülen von etwa $3{,}4 \times 10^4$ verschiedenen Arten enthält, kann man durch Standardverfahren die mRNA von etwa 100 Zellen oder (theoretisch) etwa 5×10^7 mRNA-Moleküle isolieren. Der Vergleich dieser Zahl mit der Komplexität des Primer-Gemisches zeigt, dass ohne zusätzliche Schritte und selbst wenn man annimmt, dass mRNAs mit perfekter Effizienz in cDNAs umgewandelt werden (1% Effizienz oder weniger ist genauer), das Protokoll der cDNA-Genbank-Konstruktion eine Population, welche nicht mehr als 37% der gesamten Zahl der verschiedenen Tags enthält, ergibt. Das heißt, dass das Protokoll mit überhaupt keinem offensichtlichen Schritt der Probennahme inhärent eine Probe erzeugt, die 37% oder weniger des Tag-Repertoires umfasst. Die Wahrscheinlichkeit, unter diesen Bedingungen einen Doubles zu erhalten, ist etwa 5%, was innerhalb des bevorzugten Bereichs liegt. Mit mRNA von 10 Zellen wird die Fraktion des ausgewerteten Tag-Repertoires auf nur 3,7% reduziert, selbst wenn angenommen wird, dass alle Verfahrensschritte mit einer Effizienz von 100% stattfinden. Tatsächlich sind die Effizienzen der Verfahrensschritte zur Konstruktion von cDNA-Genbanken sehr niedrig, wobei eine „Daumenregel" ist, dass eine gute Genbank etwa 10^8 cDNA-Clone aus mRNA, die aus 10^6 Säugerzellen extrahiert worden ist, enthalten sollte.

[0101] Bei der Verwendung von größeren Mengen von mRNA in dem vorstehenden Protokoll oder bei größeren Mengen von Polynucleotiden im Allgemeinen, wo die Zahl solcher Moleküle die Komplexität des Tag-Repertoires übersteigt, enthält ein Tag-Polynucleotid-Konjugat-Gemisch potentiell jede mögliche Paarung von Tags und Arten von mRNA oder Polynucleotid. In solchen Fällen kann eine offensichtliche Probennahme durch Entnahme eines Probenvolumens nach einer seriellen Verdünnung des Start-Gemisches des Tag-Polynucleotid-Konjugates durchgeführt werden. Die Stärke der benötigten Verdünnung hängt von der Menge des Startmaterials und den Effizienzen der Verfahrensschritte ab, welche leicht geschätzt werden können.

[0102] Wenn mRNA von 10^6 Zellen extrahiert würde (was etwa 0,5 µg Poly(A)$^+$-RNA entspricht) und wenn Primer in einem etwa 10–100-fachen Konzentrationsüberschuss vorhanden wären – wie dies in einem typischen Protokoll gefordert wird, zum Beispiel Sambrook et al., Molecular Cloning, zweite Auflage, Seite 8.61 (10 µl 1,8 kB mRNA bei 1 mg/ml entspricht etwa $1{,}68 \times 10^{-11}$ Mol und 10 µl 18-mer Primer bei 1 mg/ml entspricht etwa $1{,}68 \times 10^{-9}$ Mol), dann wäre die gesamte Zahl von Tag-Polynucleotid-Konjugaten in einer cDNA-Genbank einfach gleich oder geringer als die Ausgangszahl von mRNAs oder etwa 5×10^{11} Tag-Polynucleotid-Konjugate enthaltende Vektoren (wobei man erneut annimmt, dass jeder Schritt in der Konstruktion der cDNA – Synthese des ersten Stranges, Synthese des zweiten Stranges und Ligierung in einen Vektor – mit perfekter Effizienz auftritt), was eine sehr konservative Schätzung ist. Die tatsächliche Zahl ist wesentlich geringer.

[0103] Wenn eine Probe von n Tag-Polynucleotid-Konjugaten zufällig aus einem Reaktionsgemisch gezogen wird – wie durch das Ziehen eines Probenvolumens durchgeführt werden könnte, wird die Wahrscheinlichkeit des Ziehens von Konjugaten, welche den gleichen Tag besitzen, durch die Poisson-Verteilung beschrieben, $P(r) = e^{-\lambda} (\lambda)^r/_r$, wobei r die Zahl von Konjugaten, welche den gleichen Tag besitzen, ist und $\lambda = np$, wobei p die Wahrscheinlichkeit ist, dass ein gegebener Tag ausgewählt wird. Wenn $n = 10^6$ und $p = 1/(1{,}34 \times 10^8)$, dann $\lambda = 0{,}0746$ und $P(2) = 2{,}76 \times 10^{-5}$. Somit ergibt eine Probe von einer Million Molekülen eine erwartete Zahl von Doubles, die gut innerhalb des bevorzugten Bereichs liegt. Solch eine Probe wird leicht wie folgt gewonnen: Nimm an, dass die 5×10^{11} mRNAs perfekt in 5×10^{11} Vektoren mit Tag-cDNA-Konjugaten als Insertionen konvertiert werden und dass sich die 5×10^{11} Vektoren in einer Reaktionslösung befinden, die ein Volumen von 100 µl hat. 4 zehnfache serielle Verdünnungen können durch Übertragung von 10 µl aus der Ursprungslösung in ein Gefäß, welches 90 µl eines geeigneten Puffers wie TE enthält, durchgeführt werden. Dieser Vorgang kann für drei zusätzliche Verdünnungen wiederholt werden, um eine Lösung von 100 µl zu erhalten, welche 5×10^5 Vektormoleküle pro Mikroliter enthält. Eine Probe von 2 µl aus dieser Lösung ergibt 10^6 Vektoren, welche Tag-cDNA-Konjugate als Insertionen enthalten. Diese Probe wird dann durch die einfache Transformation einer kompetenten Wirtszelle gefolgt durch Züchtung vervielfältigt.

[0104] Natürlich erfolgt, wie vorstehend erwähnt, kein Schritt in dem vorstehenden Prozess mit perfekter Effizienz. Insbesondere wenn Vektoren verwendet werden, um eine Probe von Tag-Polynucleotid-Konjugaten zu vervielfältigen, ist der Schritt der Transformation eines Wirts sehr ineffizient. Für gewöhnlich werden nicht mehr als 1% der Vektoren durch den Wirt aufgenommen und vervielfältigt. Somit wären für solch ein Verfahren der Vervielfältigung sogar weniger Verdünnungen erforderlich, um eine Probe von 10^6 Konjugaten zu erhalten.

[0105] Ein Repertoire von Oligonucleotid-Tags kann an eine Population von Polynucleotiden in einer Vielfalt von Arten konjugiert werden, einschließend der direkten enzymatischen Ligierung, Vervielfältigung, zum Beispiel über PCR, unter Verwendung von Primern, die die Tag-Sequenzen enthalten, und ähnliches. Der initiale Ligierungsschritt erzeugt eine sehr große Population von Tag-Polynucleotid-Konjugaten, so dass ein einzelner Tag im Allgemeinen an viele verschiedene Polynucleotide angebracht ist. Wie jedoch vorstehend erwähnt ist, kann durch Ziehen einer ausreichend kleinen Probe der Konjugate die Wahrscheinlichkeit "Doubles" zu erhalten, das heißt der gleiche Tag an zwei verschiedenen Polynucleotiden, vernachlässigbar gemacht werden. Im Allgemeinen ist die Wahrscheinlichkeit, eine Doubles zu erhalten, umso größer je größer die Probe ist. Es besteht also ein Entwurfskonflikt zwischen der Auswahl einer großen Probe von Tag-Polynucleotid-Konjugaten – was zum Beispiel die adäquate Abdeckung eines Ziel-Polynucleotids in einer Shotgun-Sequenzieroperation oder die adäquate Repräsentation eines sich schnell ändernden mRNA-Pools sicherstellt, und der Auswahl einer kleinen Probe, was sicherstellt, dass eine minimale Zahl von Doubles anwesend sein wird. In den meisten Ausführungsformen fügt die Anwesenheit von Doubles nur eine zusätzliche Quelle von Rauschen oder im Falle des Sequenzierens eine geringfügige Erschwerung beim Abtasten und der Signalverarbeitung hinzu, da Mikropartikel, welche multiple Fluoreszenzsignale liefern, einfach ignoriert werden können.

[0106] Wie hier verwendet soll der Begriff „im Wesentlichen alle" in Bezug auf das Anbringen von Tags an Moleküle, insbesondere Polynucleotide, die statistische Natur des Verfahrens der Probennahme reflektieren, welches verwendet wird, um eine Population von Tag-Molekül-Konjugaten, die im Wesentlichen frei von Doubles ist, zu erhalten. Die Bedeutung von im Wesentlichen alle in Begriffen von tatsächlichen Prozentsätzen von Tag-Molekül-Konjugaten hängt davon ab, wie die Tags verwendet werden. Bevorzugt bedeutet im Wesentlichen alle für das Sequenzieren von Nucleinsäuren, dass wenigstens 80% der Polynucleotide einzigartige Tags angebracht haben. Stärker bevorzugt bedeutet es, dass wenigstens 90% der Polynucleotide einzigartige Tags angebracht haben. Noch stärker bevorzugt bedeutet es, dass wenigstens 95% der Polynucleotide einzigartige Tags angebracht haben. Und am stärksten bevorzugt bedeutet es, dass wenigstens 99% der Polynucleotide einzigartige Tags angebracht haben.

[0107] Bevorzugt können, wenn die Population von Polynucleotiden aus Messenger-RNA (mRNA) besteht, Oligonucleotid-Tags durch reverse Transkription der mRNA mit einen Satz von Primern angebracht werden, welche bevorzugt Komplemente der Tag-Sequenzen enthalten. Ein beispielhafter Satz von solchen Primern könnte die folgende Sequenz haben:

```
5'-mRNA-  [A]n -3'
          [T]19GG[W,W,W,C]9ACCAGCTGATC-5'-biotin
```

,

wobei "$[W,W,W,C]_9$" die Sequenz von einem Oligonucleotid-Tag aus neun Untereinheiten von jeweils vier Nucleotiden darstellt und „[W,W,W,C]" die vorstehend aufgeführten Sequenzen der Untereinheiten darstellt, das heißt „W" stellt T oder A dar. Die unterstrichenen Sequenzen identifizieren eine optionale Erkennungsstelle einer Restriktionsendonuclease, die verwendet werden kann, um das Polynucleotid aus der Verankerung an einem Festphasenträger über das Biotin, wenn eines verwendet wird, freizusetzen. Für den vorstehenden Primer könnte das an das Mikropartikel angebrachte Komplement die Form haben: 5'-$[G,W,W,W]_9$ TGG-Linker-Mikropartikel

[0108] Nach der reversen Transkription wird die mRNA entfernt, zum Beispiel durch eine RNase H-Spaltung, und der zweite Strang der cDNA wird unter Verwendung von zum Beispiel einem Primer der folgenden Form synthetisiert (SEQ ID NO: 3):
5'-NRRGATCYNNN-3'
wobei N jegliches von A, T, G oder C ist; R ein Purin enthaltendes Nucleotid ist und Y ein Pyrimidin enthaltendes Nucleotid ist. Dieser spezielle Primer erzeugt eine BstYI-Restriktionsstelle in der sich ergebenden doppelsträngigen DNA, welche zusammen mit der SalI-Stelle die Clonierung in einen Vektor mit zum Beispiel BamHI- und XhoI-Stellen ermöglicht. Nach Spaltung mit BstYI und SalI hätten die beispielhaften Konjugate die Form:

```
5'-RCGACCA[C,W,W,W]9GG[T]19-   cDNA  -NNNR
       GGT[G,W,W,W]9CC[A]19-   rDNA  -NNNYCTAG-5'
```

[0109] Die Polynucleotid-Tag-Konjugate können dann unter Verwendung von molekularbiologischen Standardverfahren manipuliert werden. Zum Beispiel kann das vorstehende Konjugat – welches tatsächlich ein Gemisch ist – in kommerziell erhältliche Clonierungsvektoren, zum Beispiel Stratagene Cloning System (La Jolla, CA) eingefügt werden; in einen Wirt wie kommerziell erhältliche Wirtsbakterien transfiziert werden; welcher

dann gezüchtet wird, um die Zahl von Konjugaten zu erhöhen. Die Clonierungsvektoren können dann unter Verwendung von Standardverfahren, zum Beispiel Sambrook et al., Molecular Cloning, zweite Auflage (Cold Spring Harbor Laboratory, New York, 1989), isoliert werden. Alternativ können geeignete Adaptoren und Primer verwendet werden, so dass die Konjugat-Population durch PCR vermehrt werden kann.

[0110] Bevorzugt werden, wenn das Ligase-basierte Verfahren der Sequenzierung verwendet wird, die mit BstYI und SalI gespaltenen Fragmente in einen mit BamHI/XhoI gespaltenen Vektor cloniert, welcher die folgenden Restriktionsstelle in einfacher Kopie besitzt:

```
5'-GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA-3'
     FokI          BamHI  XhoI
```

[0111] Dies fügt die FokI-Stelle hinzu, was die Einleitung des nachstehend vollständiger diskutierten Sequenzierprozesses erlaubt.

[0112] Tags können an cDNAs von existierenden Genbanken durch Standard-Clonierverfahren konjugiert werden. cDNAs werden aus ihrem existierenden Vektor ausgeschnitten, isoliert und dann in einen Vektor ligiert, welcher ein Repertoire von Tags enthält. Bevorzugt wird der Tag enthaltende Vektor durch Spaltung mit zwei Restriktionsenzymen linearisiert, so dass die ausgeschnittenen cDNAs in einer vorbestimmten Orientierung ligiert werden können. Die Konzentration des linearisierten, Tag enthaltenden Vektors ist im wesentlichen Überschuss gegenüber jener der cDNA-Insertionen, so dass die Ligierung eine inhärente Auswertung von Tags bereitstellt.

[0113] Ein allgemeines Verfahren zur Freisetzung des einzelsträngigen Tags nach Vervielfältigung beinhaltet das Spalten eines Ziel-Polynucleotid enthaltenden Konjugates mit der 5'→3'-Exonucleaseaktivität der T4-DNA-Polymerase oder eines ähnlichen Enzyms. Wenn in der Anwesenheit eines einzelnen Desoxynucleosidtriphosphats verwendet, wird solch eine Polymerase Nucleotide von 3' zurückgesetzten Enden, welche auf dem nicht-Matrizenstrang eines doppelsträngigen Fragments vorhanden sind, abspalten, bis ein Komplement des einzelnen Desoxynucleosidtriphosphats auf dem Matrizenstrang erreicht wird. Wenn solch ein Nucleotid erreicht wird, endet die 5'→3'-Spaltung effektiv, da die Extensionsaktivität der Polymerase Nucleotide mit einer höheren Rate hinzufügt, als die Exzisionsaktivität Nucleotide entfernt. Folglich sind einzelsträngige Tags, welche aus drei Nucleotiden konstruiert sind, einfach für eine Beladung auf Festphasenträger vorzubereiten.

[0114] Das Verfahren kann auch verwendet werden, um interne FokI-Stellen eines Ziel-Polynucleotids bevorzugt zu methylieren, während eine einzelne FokI-Stelle am Ende des Polynucleotids unmethyliert belassen wird. Zuerst wird die endständige FokI-Stelle unter Verwendung einer Polymerase mit Desoxycytidintriphosphat einzelsträngig gemacht. Der doppelsträngige Anteil des Fragmentes wird dann methyliert, wonach das einzelsträngige Ende mit einer DNA-Polymerase in der Anwesenheit von allen vier Nucleosidtriphosphaten aufgefüllt wird, wodurch die FokI-Stelle regeneriert wird. Klarerweise kann dieses Verfahren auf andere Endonucleasen als FokI verallgemeinert werden.

[0115] Nach dem die Oligonucleotid-Tags für eine spezifische Hybridisierung, zum Beispiel indem sie, wie vorstehend beschrieben, einzelsträngig gemacht wurden, vorbereitet wurden, werden die Polynucleotide mit Mikropartikeln, welche die komplementären Sequenzen der Tags enthalten, unter Bedingungen gemischt, welche die Bildung von perfekt passenden Duplexmolekülen zwischen den Tags und ihren Komplementen begünstigen. Es gibt ausführliche Anleitung in der Literatur zur Herstellung dieser Bedingungen. Beispielhafte Quellen, welche solche Anleitung bereitstellen, schließen Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26 (1991): 227–259; Sambrook et al., Molecular Cloning: A Laboratory Manual, zweite Auflage (Cold Spring Harbor Laboratory, New York, 1989); und ähnliches ein. Bevorzugt sind die Hybridisierungsbedingungen ausreichend stringent, so dass nur perfekt passende Sequenzen stabile Duplexmoleküle bilden. Unter solchen Bedingungen können die durch ihre Tags spezifisch hybridisierten Polynucleotide an die an die Mikropartikel angebrachten komplementären Sequenzen ligiert werden. Schließlich werden die Mikropartikel gewaschen, um Polynucleotide mit nicht ligierten und/oder fehlgepaarten Tags zu entfernen.

[0116] Wenn üblicherweise als Syntheseträger verwendete CPG-Mikropartikel verwendet werden, ist die Dichte von Tagkomplementen auf der Oberfläche der Mikropartikel typischerweise größer als jene, die für einige Sequenzierungsoperationen notwendig ist. Das heißt, dass in Sequenzierungsansätzen, die die aufeinanderfolgende Behandlung der angebrachten Polynucleotide mit einer Vielfalt von Enzymen erfordern, dicht gesetzte Polynucleotide dazu neigen können, den Zugang der relativ sperrigen Enzyme zu den Polynucleoti-

den zu hemmen. In solchen Fällen werden die Polynucleotide bevorzugt so mit den Mikropartikeln gemischt, dass die Tagkomplemente gegenüber den Polynucleotiden in einem wesentlichen Überschuss vorhanden sind, zum Beispiel von 10:1 bis 100:1 oder mehr. Dies stellt sicher, dass die Dichte der Polynucleotide auf der Oberfläche der Mikropartikel nicht so hoch sein wird, dass sie den Zugang der Enzyme hemmen wird. Bevorzugt ist der durchschnittliche Abstand zwischen den Polynucleotiden auf der Oberfläche der Mikropartikel in der Größenordnung von 30–100 nm. Anleitung für die Auswahl der Verhältnisse für Standard-CPG-Träger und Ballotini-Kügelchen (eine Art von festen Glas-Trägern) ist in Maskos und Southern, Nucleic Acids Research, 20 (1992) 1679–1684, zu finden. Bevorzugt werden für Sequenzierungsanwendungen Standard-CPG-Kügelchen mit einem Durchmesser im Bereich von 20–50 µm mit etwa 10^5 Polynucleotiden beladen und Glycidalmethacrylat (GMA)-Kügelchen, erhältlich von Gangs Laboratories (Carmel, IN), mit einem Durchmesser im Bereich von 5–10 µm werden mit einigen zehntausend Polynucleotiden beladen, zum Beispiel 4×10^4 bis 6×10^4.

[0117] In der bevorzugten Ausführungsform werden Tagkomplemente zur Sortierung kombinatorisch auf Mikropartikeln synthetisiert; so erhält man am Ende der Synthese ein komplexes Gemisch von Mikropartikeln, von welchem eine Probe zur Beladung mit mit einem Tag versehenen Polynucleotiden genommen wird. Die Größe der Probe von Mikropartikeln wird von einigen Faktoren einschließlich der Größe des Repertoires von Tagkomplementen, der Art des Gerätes, welches zur Beobachtung der beladenen Mikropartikel verwendet wird – zum Beispiel seiner Kapazität, der Toleranz gegenüber multiplen Kopien von Mikropartikeln mit dem gleichen Tagkomplement (das heißt „Kügelchen-Doubles") und ähnlichem abhängen. Die folgende Tabelle stellt eine Anleitung bezüglich der Probengröße der Mikropartikel, des Durchmessers der Mikropartikel und den annähernden physikalischen Dimensionen einer gepackten Anordnung von Mikropartikeln von verschiedenen Durchmessern bereit.

Mikropartikel-durchmesser	5 µm	10 µm	20 µm	40 µm
Max. Zahl von geladenen Polynucleotiden bei 1 pro 10^5 Quadratangström		3×10^5	1.26×10^6	5×10^6
Annähernde Oberfläche einer einlagigen Schicht von 10^6 Mikropartikeln	.45 x .45 cm	1 x 1 cm	2 x 2 cm	4 x 4 cm

[0118] Die Wahrscheinlichkeit, dass die Probe der Mikropartikel ein gegebenes Tagkomplement enthält oder dass dieses in multiplen Kopien vorliegt, wird, wie in der folgenden Tabelle angezeigt, durch die Poisson-Verteilung beschrieben.

Tabelle VI

Zahl von Mikropartikeln in der Probe (als Bruchteil der Repertoiregröße), m	In der Probe vorhandener Bruchteil des Repertoires der Tagkomplemente, $1-e^{-m}$	Bruchteil der Mikropartikel in der Probe mit einem angebrachten einzigartigen Tagkomplement, $m(e^{-m})/2$	Bruchteil der Mikropartikel in der Probe, die das gleiche Tagkomplement wie ein anderes Mikropartikel in der Probe tragen („Kügelchen-Doubles"), $m^2(e^{-m})/2$
1.000	0.63	0.37	0.18
.693	0.50	0.35	0.12
.405	0.33	0.27	0.05
.285	0.25	0.21	0.03
.223	0.20	0.18	0.02
.105	0.10	0.09	0.005
.010	0.01	0.01	

Sortieren und Panning mit hoher Spezifität

[0119] Die Kinetik des Sortierens hängt von der Rate der Hybridisierung von Oligonucleotid-Tags an ihre Tagkomplemente ab, welche ihrerseits von der Komplexität der Tags in der Hybridisierungsreaktion abhängt. Es besteht also ein Konflikt zwischen der Rate der Sortierung und der Komplexität der Tags, so dass eine Zunahme in der Rate der Sortierung auf Kosten der Verminderung der Komplexität von an der Hybridisierungsreaktion beteiligten Tags erreicht werden kann. Wie nachstehend erklärt, können die Auswirkungen dieses Konflikts durch "Panning" gemildert werden.

[0120] Die Spezifität der Hybridisierungen kann durch Nehmen einer ausreichend kleinen Probe gesteigert werden, so dass sowohl ein hoher Prozentsatz der Tags in der Probe einzigartig ist und auch die nächsten Nachbarn von im Wesentlichen allen Tags in einer Probe sich um wenigstens zwei Worte unterscheiden. Diese letztere Bedingung kann durch Nehmen einer Probe erfüllt werden, die eine Zahl von Tag-Polynucleotid-Konjugaten enthält, die etwa 0,1% oder weniger der Größe des verwendeten Repertoires hat. Wenn zum Beispiel die Tags mit acht Worten, ausgewählt aus Tabelle 2, konstruiert werden, wird ein Repertoire von 8^8 oder etwa $1{,}67 \times 10^7$ Tags und Tagkomplementen erzeugt. In einer wie vorstehend beschriebenen Genbank von Tag-cDNA-Konjugaten bedeutet eine 0,1%-Probe, dass etwa 16700 verschiedene Tags vorhanden sind. Wenn diese direkt auf ein Repertoire-Äquivalent von Mikropartikeln oder in diesem Beispiel eine Probe von $1{,}67 \times 10^7$ Mikropartikeln geladen würde, dann wäre nur eine kleine Untergruppe der getesteten Mikropartikel beladen. Die Dichte der beladenen Mikropartikel kann erhöht werden – zum Beispiel für effizienteres Sequenzieren – indem ein „Panning"-Schritt, in welchem die getesteten Tag-cDNA-Konjugate verwendet werden, um beladene Mikropartikel von unbeladenen Mikropartikeln zu trennen, durchgeführt wird. Somit können im vorstehenden Beispiel, obwohl eine „0,1%"-Probe nur 16700 cDNAs enthält, die Probennahme- und Panning-Schritte wiederholt werden, bis so viele beladene Mikropartikel wie gewünscht angesammelt sind. Alternativ können beladene Mikropartikel von unbeladenen Mikropartikeln durch ein Fluoreszenzaktiviertes Zell-Sortier (FACS)-Gerät unter Verwendung üblicher Protokolle getrennt werden, zum Beispiel können Tag-cDNA-Konjugate in dem nachstehend beschriebenen Verfahren durch Bereitstellung eines fluoreszierend markierten rechten Primers fluoreszierend markiert werden. Nach Beladung und FACS-Sortierung kann die Markierung vor der Ligierung codierter Adaptoren zum Beispiel durch DpnI oder ähnliche Enzyme, welche methylierte Stellen erkennen, abgespalten werden.

[0121] Ein Panning-Schritt kann durch Bereitstellung einer Probe von Tag-cDNA-Konjugaten durchgeführt werden, von welchen jedes eine Fänger-Einheit an einem Ende gegenüberliegend oder entfernt von dem Oligonucleotid-Tag enthält. Bevorzugt ist die Fänger-Einheit von einer Art, welche von den Tag-cDNA-Konjugaten freigesetzt werden kann, so dass die Tag-cDNA-Konjugate mit einem Einzelbasen-Sequenzierverfahren sequenziert werden können. Solche Einheiten können Biotin, Digoxigenin oder ähnliche Liganden, eine Triplex-bindende Region oder ähnliches umfassen. Bevorzugt umfasst solch eine Fänger-Einheit eine Biotin-Komponente. Biotin kann an Tag-cDNA-Konjugate durch eine Vielzahl von Standardverfahren angebracht werden. Wenn geeignete, Adapter enthaltende PCR-Primerbindungsstellen an Tag-cDNA-Konjugaten angebracht werden, kann Biotin durch Verwendung eines biotinylierten Primers in einer Vervielfältigung nach der Probennahme angebracht werden. Alternativ kann, wenn die Tag-cDNA-Konjugate Insertionen von Clonierungsvektoren sind, nach Ausschneiden der Tag-cDNA-Konjugate durch Spaltung mit einem geeigneten Restriktionsenzym, gefolgt durch Isolierung und Auffüllen eines überhängenden Stranges entfernt von den Tags mit einer DNA-Polymerase in der Anwesenheit von biotinyliertem Uridintriphosphat, Biotin angebracht werden.

[0122] Nachdem ein Tag-cDNA-Konjugat gefangen ist, kann es von der Biotin-Einheit in einer Vielfalt von Arten wie durch eine chemische Bindung, die durch Reduktion gespalten wird, zum Beispiel Herman et al., Anal. Biochem., 156 (1986): 48–55, oder die photochemisch gespalten wird, zum Beispiel Olejnik et al., Nucleic erste Zusatz, 24 (1996): 361–366, oder die enzymatisch durch Einführung einer Restriktionsstelle in den PCR-Primer gespalten wird, freigesetzt werden. Die letztere Ausführungsform kann durch Betrachtung der vorstehend beschriebenen Genbank von Tag-Polynucleotid-Konjugaten veranschaulicht werden:

```
5'-RCGACCA[C,W,W,W]9GG[T]19- cDNA -NNNR
       GGT[G,W,W,W]9CC[A]19- rDNA -NNNYCTAG-5'
```

[0123] Die folgenden Adapter können an die Enden dieser Fragmente ligiert werden, um eine Vervielfältigung durch PCR zu erlauben:

```
5'- XXXXXXXXXXXXXXXXXXXX
    XXXXXXXXXXXXXXXXXXXXYGAT
```

Rechter Adapter

```
GATCZZACTAGTZZZZZZZZZZZZ-3'
    ZZTGATCAZZZZZZZZZZZZ
```

Linker Adapter

```
ZZTGATCAZZZZZZZZZZZZ-5'-biotin
```

Linker Primer

wobei "ACTAGT" eine SpeI-Erkennungsstelle ist (welche eine zurückgesetzte Spaltung hinterlässt, die zum Einzelbasen-Sequenzieren bereit ist) und die X's und Z's ausgewählte Nucleotide sind, so dass die Aneinanderlagerungs- und Dissoziationstemperaturen der jeweiligen Primer annähernd die gleichen sind. Nach Ligierung der Adapter und Vervielfältigung durch PCR unter Verwendung der biotinylierten Primer werden die Tags der Konjugate durch die Exonucleaseaktivität der T4-DNA-Polymerase einzelsträngig gemacht und die Konjugate werden mit einer Probe der Mikropartikel, zum Beispiel ein Repertoire-Äquivalent, mit den angebrachten Tagkomplementen kombiniert. Nach dem Aneinanderlagern unter stringenten Bedingungen (um das fehlerhafte Anbringen von Tags zu minimieren), werden die Konjugate bevorzugt an ihre Tagkomplemente ligiert und die beladenen Mikropartikel werden von den unbeladenen Mikropartikeln durch Fangen mit mit Avidin versehenen magnetischen Kügelchen oder einer ähnlichen Fänger-Technik getrennt.

[0124] Zu dem Beispiel zurückkehrend, ergibt dieser Prozess die Akkumulation von etwa 10500 (= 16700 × 63) beladenen Mikropartikeln mit verschiedenen Tags, welche von den magnetischen Kügelchen durch Spaltung mit SpeI freigesetzt werden können. Bei 40–50-maliger Wiederholung dieses Prozesses mit neuen Proben von Mikropartikeln und Tag-cDNA-Konjugaten können durch Zusammenführen der freigesetzten Mikropartikel 4–5 × 10^5 cDNAs angesammelt werden. Die zusammengeführten Mikropartikel können dann simultan durch eine Einzelbasen-Sequenzierungstechnik sequenziert werden.

[0125] Das Bestimmen, wie oft die Probennahme- und Panning-Schritte zu wiederholen sind – oder allgemeiner, das Bestimmen wie, viele cDNAs zu untersuchen sind, hängt von der Zielsetzung einer Person ab. Wenn die Zielsetzung ist, Änderungen der Häufigkeit von relativ häufigen, zum Beispiel bis zu 5% oder mehr einer Population ausmachend, Sequenzen zu überwachen, dann können relativ kleine Proben, das heißt ein kleiner Bruchteil der gesamten Populationsgröße, statistisch signifikante Schätzungen der relativen Häufigkeiten erlauben. Wenn man andererseits die Häufigkeit von seltenen Sequenzen, zum Beispiel bis zu 0,1% oder weniger einer Population ausmachend, zu überwachen sucht, dann sind große Proben erforderlich. Im Allgemeinen gibt es eine direkte Beziehung zwischen der Probengröße und der Verlässlichkeit von Schätzungen der relativen Häufigkeiten, basierend auf der Probe. Es gibt ausführliche Anleitung in der Literatur zur Bestimmung der geeigneten Probengröße zur Durchführung verlässlicher statistischer Schätzungen, zum Beispiel Koller et al., Nucleic Acids Research, 23 (1994): 185–191; Good, Biometrika, 40 (1953): 16–264; Bunge et al., J. Am. Stat. Assoc., 88 (1993): 364–373; und ähnliches. Bevorzugt wird zur Überwachung der Änderung der Genexpression basierend auf der Untersuchung einer Serie von cDNA-Genbanken, welche 10^5 bis 10^8 unabhängige Clone von 3,0–3,5 × 10^4 verschiedenen Sequenzen enthalten, eine Probe von wenigstens 10^4 Sequenzen zur Untersuchung jeder Genbank akkumuliert. Stärker bevorzugt wird eine Probe von wenigstens 10^5 Sequenzen für die Untersuchung jeder Genbank akkumuliert; und am stärksten bevorzugt wird eine Probe von wenigstens 5 × 10^5 Sequenzen für die Untersuchung jeder Genbank akkumuliert. Alternativ ist die Zahl der untersuchten Sequenzen bevorzugt ausreichend, um die relative Häufigkeit einer Sequenz, die mit einer Frequenz im Bereich von 0,1% bis 5% vorhanden ist, mit einer 95%-Konfidenzgrenze von nicht größer als 0,1% der Populationsgröße zu schätzen.

Konstruktion einer Tag-Genbank

[0126] Eine beispielhafte Tag-Genbank wird wie folgt konstruiert, um die chemisch synthetisierten 9-Wort-Tags aus den Nucleotiden A, G und T zu bilden, definiert durch die Formel:
3'-TGGC-[^{4}(A,G,T)$_{9}$]-CCCCp,
wobei „[^{4}(A,G,T,)$_{9}$]" ein Tag-Gemisch anzeigt, worin jeder Tag aus neun 4-mer-Wörtern aus A, G und T besteht; und "p" ein 5'-Phosphat anzeigt. Dieses Gemisch wird an die folgenden rechten und linken primerbindenden Regionen ligiert (SEQ ID NO: 4 & 5):

```
5'- AGTGGCTGGGCATCGGACCG        5'- GGGGCCCAGTCAGCGTCGAT
    TCACCGACCCGTAGCCp                   GGGTCAGTCGCAGCTA
```

Links Rechts

[0127] Die rechten und linken primerbindenden Regionen werden an das vorstehende Tag-Gemisch ligiert, wonach der einzelsträngige Abschnitt der ligierten Struktur mit einer DNA-Polymerase aufgefüllt, dann mit den nachstehend gezeigten rechten und linken Primern gemischt und vervielfältigt wird, um eine Tag-Genbank zu ergeben.

Linker Primer

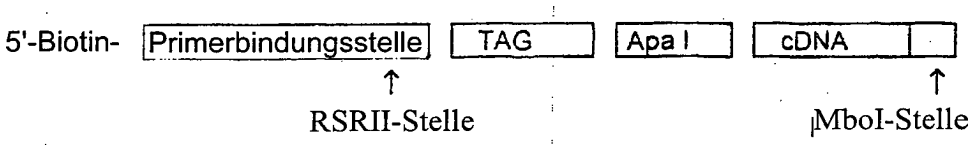
```
5'- AGTGGCTGGGCATCGGACCG

5'- AGTGGCTGGGCATCGGACCG-  [4(A,G,T)9]-GGGGCCCAGTCAGCGTCGAT
    TCACCGACCCGTAGCCTGGC-  [4(A,G,T)9]-CCCCGGGTCAGTCGCAGCTA

                                        CCCCGGGTCAGTCGCAGCTA-5'
```

Rechter Primer

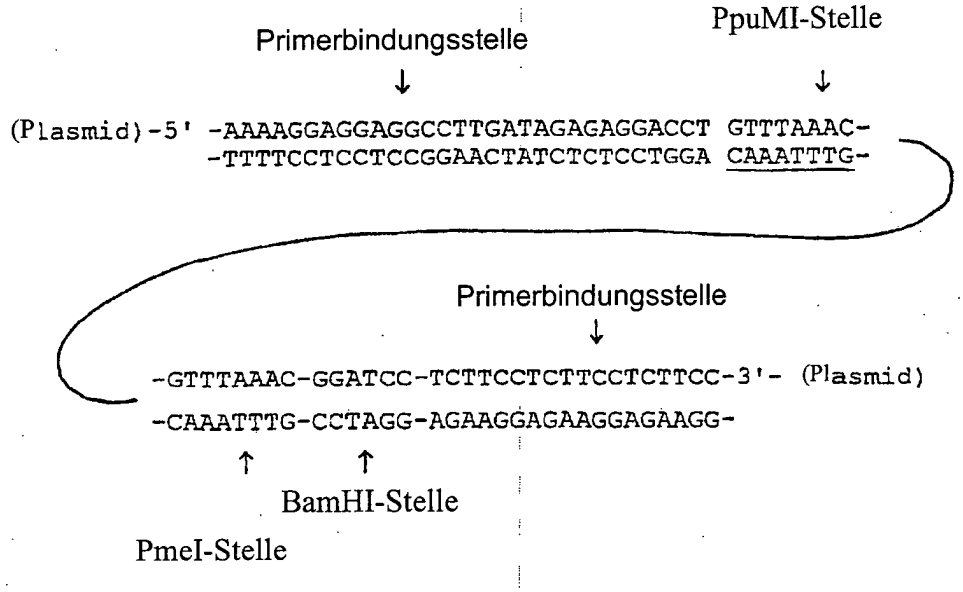
[0128] Der unterstrichene Abschnitt der linken primerbindenden Region zeigt eine RsrII-Erkennungsstelle an. Die unterstrichene Region der rechten primerbindenden Region am weitesten links zeigt die Erkennungsstellen von Bsp120I, ApaI, EcoO109I und eine Schneidestelle für HgaI an. Die unterstrichene Region der rechten primerbindenden Region am weitesten rechts zeigt die Erkennungsstelle für HgaI an. Optional kann der rechte oder der linke Primer mit einem angebrachten Biotin synthetisiert werden (unter Verwendung üblicher Reagenzien, zum Beispiel erhältlich von Clontech Laboratories, Palo Alto, CA), um die Reinigung nach Vervielfältigung und/oder Spaltung zu erleichtern.

Konstruktion einer Plasmid-Genbank von Tag-Polynucleotid-Konjugaten für die cDNA-"Signatur"-Sequenzierung unter Verwendung codierter Adaptoren

[0129] Aus einer mRNA-Probe wird durch konventionelle Protokolle unter Verwendung von pGGCCCT$_{15}$ (A oder G oder C) als Primer zur Synthese des ersten Stranges, verankert an der Grenze der poly A-Region der mRNAs, und N$_{8}$ (A oder T)GATC als Primer für die Synthese des zweiten Stranges cDNA hergestellt. Das heißt, beide sind degenerierte Primer, so dass der Primer für den zweiten Strang in zwei Formen vorliegt und der Primer für den ersten Strang in drei Formen vorliegt. Die GATC-Sequenz im Primer für den zweiten Strang entspricht der Erkennungsstelle von MboI; andere Erkennungsstellen mit vier Basen könnten genauso gut verwendet werden, so wie jene für BamHI, SphI, EcoRI oder ähnliche. Die Anwesenheit des A und T angrenzend an die Restriktionsstelle des Primers für den zweiten Strang stellt sicher, dass eine Stripping- und Austauschreaktion im nächsten Schritt verwendet werden kann, um einen 5'-Überhang von fünf Basen von "GGCCC" zu erzeugen. Der Primer für den ersten Strang wird an die mRNA-Probe angelagert und mit reverser Transkriptase verlängert, wonach der RNA-Strang durch die RNase H-Aktivität der reversen Transkriptase, eine einzelsträngige cDNA hinterlassend, abgebaut wird. Der Primer für den zweiten Strang wird angelagert und mit einer DNA-Polymerase unter Verwendung üblicher Protokolle verlängert. Nach der Synthese des zweiten Stranges werden die sich ergebenden cDNAs mit CpG-Methylase unter Verwendung des Protokolls des Herstellers methyliert (New England Biolabs, Beverly, MA). Die 3'-Stränge der cDNAs werden dann mit der vorstehend erwähnten Stripping- und Austauschreaktion unter Verwendung von T4-DNA-Polymerase in der Anwesenheit von dATP und dTTP zurückgeschnitten, wonach die cDNAs an die vorstehend beschriebene Tag-Genbank ligiert werden, welche vorher mit HgaI gespalten wurde, um das folgende Konstrukt zu ergeben:



[0130] Getrennt wird der folgende Vektor konstruiert, zum Beispiel ausgehend von einem kommerziell erhältlichen Plasmid wie einem Bluescript-Phagemid (Stratagene, La Jolla, CA) (SEQ ID NO: 6)



[0131] Das Plasmid wird mit PpuMI und PmeI gespalten (um ein RsrII-kompatibles Ende und ein glattes Ende zu ergeben, so dass die Insertion orientiert ist) und dann mit DAM-Methylase methyliert. Das den Tag enthaltende Konstrukt wird mit RsrII gespalten und dann mit dem offenen Plasmid ligiert, wonach das Konjugat mit Mbo I und BamHI gespalten wird, um eine Ligierung und das Schließen des Plasmids zu erlauben. Die Plasmide werden dann zur Verwendung in Übereinstimmung mit der Erfindung vervielfältigt und isoliert.

Beispiel 1

Sequenzierung eines aus pGEM7Z vervielfältigten Ziel-Polynucleotids: Identifizierung von Nucleotiden durch Zyklen von Ligierung und Abspaltung

[0132] In diesem Beispiel wird ein Abschnitt des Plasmids pGEM7Z (Promega, Madison, WI) vervielfältigt und über einen doppelsträngigen DNA-Linker, von welchem ein Strang direkt auf (und deshalb kovalent an diese gebunden) die Kügelchen synthetisiert wird, an Glaskügelchen angebracht. Nachdem die Enden der Ziel-Polynucleotide zur Ligierung an die codierten Adaptoren vorbereitet worden sind, wird in jedem Zyklus von Ligierung und Abspaltung ein Gemisch von codierten Adaptoren (insgesamt 1024 verschiedene Adaptoren) an den Ziel-Polynucleotiden angewendet, so dass nur jene Adaptoren, deren überhängende Stränge perfekt passende Duplexmoleküle mit den Ziel-Polynucleotiden bilden, ligiert werden. Jedes von 16 fluoreszierend markierten Tagkomplemente wird dann an den Polynucleotid-Adaptor-Konjugaten unter Bedingungen angewendet, die nur die Hybridisierung von korrekten Tagkomplementen erlauben. Die Anwesenheit oder Abwesenheit eines Fluoreszenzsignals nach dem Waschen zeigt die Anwesenheit oder Abwesenheit eines bestimmten Nucleotids an einem bestimmten Ort. Das Sequenzierungsprotokoll dieses Beispiel ist es auf multiple Ziel-Polynucleotide anwendbar, die auf einen oder mehrere Festphasenträger, wie in Brenner, internationale Patentanmeldungen PCT/US 95/12791 und PCT/US 96/09513 (WO 96/12014 und WO 96/41011) beschrieben, sortiert sind.

[0133] Ein 47-mer-Oligonucleotid wird direkt auf Ballotini-Kügelchen (0,040–0,075 mm, Jencons Scientific, Bridgeville, PA) unter Verwendung eines Standardprotokolls für automatisierte DNA-Synthese synthetisiert. Der komplementäre Strang zu dem 47-mer wird separat synthetisiert und durch HPLC gereinigt. Wenn er hybridisiert ist, hat das sich ergebende Duplexmolekül eine BstXI-Restriktionsstelle am vom Kügelchen entfernten Ende. Der komplementäre Strang wird in dem folgenden Gemisch an das angebrachte 47-mer hybridisiert: 25 µl komplementärer Strang bei 200 pMol/µl; 20 mg Ballotini-Kügelchen mit dem 47-mer; 6 µl New England Biolabs #3-Restriktionspuffer (aus einer 10 × Stammlösung); und 25 µl destilliertes Wasser. Das Gemisch wird

auf 93°C erhitzt und dann langsam auf 55°C gekühlt, wonach 40 Einheiten von BstXI (zu 10 Einheiten/µl) hinzugefügt werden, um das Reaktionsvolumen auf 60 µl zu bringen. Das Gemisch wird bei 55°C über zwei Stunden inkubiert, wonach die Kügelchen dreimal in TE (pH 8,0) gewaschen werden.

[0134] Der Abschnitt von pGEM7Z, welcher an den Kügelchen angebracht werden soll, wird wie folgt hergestellt: zwei PCR-Primer wurden unter Verwendung von Standardprotokollen hergestellt (SEQ ID NO: 7 und SEQ ID NO: 8):

```
Primer 1:  5'-CTAAACCATTGGTATGGGCCAGTGAATTGTAATA

Primer 2:  5'-CGCGCAGCCCGCATCGTTTATGCTACAGACTGTC-
                                   AGTGCAGCTCTCCGATCCAAA
```

[0135] Das PCR-Reaktionsgemisch besteht aus dem folgenden: 1 µl pGEM7Z mit 1 ng/µl; 10 µl Primer 1 mit 10 pMol/µl; 10 µl Primer 2 mit 10 pMol/µl; 10 µl Desoxyribonucleosidtriphosphate mit 2,5 mM; 10 µl 10 × PCR-Puffer (Perkin-Elmer); 0,5 µl Taq-DNA-Polymerase mit 5 Einheiten/µl; und 58 µl destilliertes Wasser, um ein Endvolumen von 100 µl zu ergeben. Das Reaktionsgemisch wurde 25 Zyklen von 93°C über 30 Sekunden; 60°C über 15 Sekunden; und 72°C über 60 Sekunden unterzogen, um ein Produkt von 172 Basenpaaren zu ergeben, welches aufeinanderfolgend mit BbvI (100 µl PCR-Reaktionsgemisch, 12 µl 10 × #1 New England Biolabs-Puffer, 8 µl BbvI mit 1 Einheiten/µl inkubiert bei 37°C über sechs Stunden) und mit Bst XI (zu dem BbvI-Reaktionsgemisch wird hinzugefügt: 5 µl 1 M NaCl, 67 µl destilliertes Wasser und 8 µl BstXI mit 10 Einheiten/µl und das sich ergebende Gemisch wird bei 55°C über zwei Stunden inkubiert) gespalten wird.

[0136] Nach Hindurchschicken des vorstehenden Reaktionsgemisches durch eine Centricon 30-Zentrifugationssäule (Amicon, Inc.) gemäß dem Herstellerprotokoll wird das mit BbvI/BstXII gespaltene Fragment in dem folgenden Gemisch an den an den Ballotini-Kügelchen angebrachten doppelsträngigen Linker ligiert: 17 µl mit BbvI/BstXI gespaltenes Fragment (10 µg), 10 µl Kügelchen (20 mg), 6 ml 10 × Ligierungspuffer (New England Biolabs, nachstehend als NEB bezeichnet) 5 µl T4-DNA-Ligase mit 2000 Einheiten/µl und 22 µl destilliertes Wasser; dieses Gemisch wird bei 25°C über vier Stunden inkubiert, wonach die Kügelchen dreimal mit TE (pH 8,0) gewaschen werden, was das folgende Ziel-Polynucleotid (SEQ ID NO: 9) zum Sequenzieren hinterlässt, welches ein 5'-Phosphat besitzt:

```
                 . . .    AGCTACCCGATC
[Kügelchen] --   . . .    TCGATGGGCTAGATTTp-5'
```

[0137] Das 5'-Phosphat wird durch Behandlung des Kügelchen-Gemisches mit einer alkalischen Phosphatase, zum Beispiel aus Kälberdarm, erhältlich von New England Biolabs (Beverly, MA), unter Verwendung des Protokolls des Herstellers entfernt.

[0138] Die oberen Stränge der folgenden 16 Sätze von 64 codierten Adaptoren (SEQ ID NO: 10 bis SEQ ID NO: 25) werden jeder getrennt auf einem automatisierten DNA-Syntheseautomaten (Modell 392 Applied Biosystems, Foster City) unter Verwendung von Standardverfahren synthetisiert. Der untere Strang, welcher für alle Adaptoren der gleiche ist, wird getrennt synthetisiert und dann an die jeweiligen oberen Stränge hybridisiert:

SEQ ID NO.	Codierter Adoptor
10	5'-pANNNTACAGCTGCATCCCttggcgctgagg pATGCACGCGTAGGG-5'
11	5'-pNANNTACAGCTGCATCCCtgggcctgtaag pATGCACGCGTAGGG-5'
12	5'-pCNNNTACAGCTGCATCCCttgacgggtctc pATGCACGCGTAGGG-5'
13	5'-pNCNNTACAGCTGCATCCCtgcccgcacagt pATGCACGCGTAGGG-5'
14	5'-pGNNNTACAGCTGCATCCCttcgcctcggac pATGCACGCGTAGGG-5'
15	5'-pNGNNTACAGCTGCATCCCtgatccgctagc pATGCACGCGTAGGG-5'
16	5'-pTNNNTACAGCTGCATCCCttccgaacccgc pATGCACGCGTAGGG-5'
17	5'-pNTNNTACAGCTGCATCCCtgaggggatag pATGCACGCGTAGGG-5'
18	5'-pNNANTACAGCTGCATCCCttcccgctacac pATGCACGCGTAGGG-5'
19	5'-pNNNATACAGCTGCATCCCtgactccccgag pATGCACGCGTAGGG-5'
20	5'-pNNCNTACAGCTGCATCCCtgtgttgcgcgg pATGCACGCGTAGGG-5'
21	5'-pNNNCTACAGCTGCATCCCtctacagcagcg pATGCACGCGTAGGG-5'
22	5'-pNNGNTACAGCTGCATCCCtgtcgcgtcgtt pATGCACGCGTAGGG-5'
23	5'-pNNNGTACAGCTGCATCCCtcggagcaacct pATGCACGCGTAGGG-5'
24	5'-pNNTNTACAGCTGCATCCCtggtgaccgtag pATGCACGCGTAGGG-5'
25	5'-pNNNTTACAGCTGCATCCCtcccctgtcgga pATGCACGCGTAGGG-5'

wobei N und p wie vorstehend definiert sind und die in Kleinbuchstaben angezeigten Nucleotide 12-mer-Oligonucleotid-Tags sind. Jeder Tag unterscheidet sich von jedem anderen durch sechs Nucleotide. Gleiche molare Mengen von jedem Adaptor werden in NEB-Puffer Nr. 2 (New England Biolabs, Beverly, MA) kombiniert, um ein Gemisch mit einer Konzentration von 1000 pMol/µl zu bilden.

[0139] Jedes der 16 Tagkomplemente wird getrennt als Amino-derivatisiertes Oligonucleotid synthetisiert und wird jeweils mit einem Fluoresceinmolekül markiert (zum Beispiel FAM, ein NHS-Ester von Fluorescein, erhältlich von Molecular Probes, Eugene, OR), welches durch einen Polyethylenglycol-Linker an das 5'-Ende des Tagkomplements angebracht wird (Clonetech Laboratories, Palo Alto, CA). Die Sequenzen der Tagkomplemente sind einfach die 12-mer-Komplemente der vorstehend aufgeführten Tags.

[0140] Die Ligierung der Adaptoren an das Ziel-Polynucleotid wird in einem Gemisch durchgeführt, das aus

5 µl Kügelchen (20 mg), 3 µl NEB 10 × Ligasepuffer, 5 µl Adaptor-Gemisch (25 nM), 2,5 µl NEB-T4-DNA-Ligase (2000 Einheiten/µl) und 14,5 µl destilliertes Wasser besteht. Das Gemisch wird bei 16°C über 30 Minuten inkubiert, wonach die Kügelchen dreimal in TE (pH 8,0) gewaschen werden.

[0141] Nach Zentrifugation und Entfernung von TE, werden die 3'-Phosphate der ligierten Adaptoren durch Behandlung des Polynucleotid-Kügelchen-Gemisches mit alkalischer Phosphatase aus Kälberdarm (CIP) (New England Biolabs, Beverly, MA) unter Verwendung des Protokolls des Herstellers entfernt. Nach Entfernung der 3'-Phosphate kann die CIP durch proteolytische Spaltung zum Beispiel unter Verwendung von Pronase™ (erhältlich von Boehringer Mannheim, Indianapolis, IN) oder einer äquivalenten Protease mit dem Protokoll des Herstellers inaktiviert werden. Das Polynucleotid-Kügelchen-Gemisch wird dann gewaschen und mit einem Gemisch von T4-Polynucleotid-Kinase und T4-DNA-Ligase (New England Biolabs, Beverly, MA) behandelt, um ein 5'-Phosphat an der Lücke zwischen dem Ziel-Polynucleotid und dem Adaptor hinzuzufügen, um die Ligierung des Adaptors an das Ziel-Polynucleotid zu vervollständigen. Das Kügelchen-Polynucleotid-Gemisch wird dann in TE gewaschen.

[0142] Getrennt wird jedes der markierten Tagkomplemente an das Polynucleotid-Kügelchen-Gemisch unter Bedingungen angewandt, welche die Bildung von perfekt passenden Duplexmolekülen nur zwischen den Oligonucleotid-Tags und ihren jeweiligen Komplementen erlauben, wonach das Gemisch unter stringenten Bedingungen gewaschen wird und die Anwesenheit oder Abwesenheit eines Fluoreszenzsignals gemessen wird. Tagkomplemente werden in einer Lösung, die aus 25 nM Tagkomplement, 50 mM NaCl, 3 mM Mg, 10 mM Tris-HCl (pH 8,5) besteht, bei 20 °C angewendet und über 10 Minuten inkubiert, dann in der gleichen Lösung über 10 Minuten bei 55°C gewaschen (ohne Tagkomplement).

[0143] Nachdem, wie vorstehend beschrieben, die vier Nucleotide identifiziert sind, werden die codierten Adaptoren von den Polynucleotiden mit BbvI unter Verwendung des Protokolls des Herstellers abgespalten. Nach einer ersten Ligierung und Identifizierung wird der Zyklus von Ligierung, Identifizierung und Abspaltung dreimal wiederholt, um die Sequenz der 16 endständigen Nucleotide des Ziel-Polynucleotids zu ergeben. Die **Fig. 4** stellt die relative Fluoreszenz von jedem von vier Tagkomplementen dar, welche angewendet werden, um Nucleotide an den Positionen 5 bis 16 zu identifizieren (von der am weitesten vom Kügelchen entfernten zu der, die dem Kügelchen am nächsten ist).

Beispiel 2

Konstruktion und Sortieren einer cDNA-Genbank zum Signatur-Sequenzieren mit codierten Adaptoren

[0144] In diesem Beispiel wird eine cDNA-Genbank konstruiert, in welcher ein Oligonucleotid-Tag, bestehend aus 8 vier-Nucleotid-"Worten", an jeder cDNA angebracht ist. Wie vorstehend beschrieben, ist das Repertoire von Oligonucleotid-Tags dieser Größe ausreichend groß (etwa 10^8), so dass, wenn die cDNAs aus einer Population von etwa 10^6 synthetisiert werden, eine hohe Wahrscheinlichkeit besteht, dass dann jede cDNA einen einzigartigen Tag für das Sortieren haben wird. Nach der Extraktion der mRNA wird die Synthese des ersten Stranges in der Anwesenheit von 5-Me-dCTP (um bestimmte cDNA-Restriktionsstellen zu blockieren) und eines biotinylierten Primer-Gemisches, welches die Oligonucleotid-Tags enthält, durchgeführt. Nach üblicher Synthese des zweiten Stranges werden die Tag-cDNA-Konjugate mit DpnII (welches durch die 5-Me-Desoxycytosine unbeeinflusst ist) gespalten, die biotinylierten Anteile werden aus dem Reaktionsgemisch unter Verwendung von mit Streptavidin beschichteten magnetischen Kügelchen getrennt und die Tag-cDNA-Konjugate werden durch ihre Abspaltung von den magnetischen Kügelchen über eine BsmBI-Stelle, welche durch den biotinylierten Primer getragen wird, gewonnen. Das BsmBI-DpnII-Fragment, welches die Tag-cDNA-Konjugate enthält, wird dann in ein Plasmid eingeführt und vervielfältigt. Nach Isolierung des Plasmids werden die Tag-cDNA-Konjugate aus diesen Plasmiden durch PCR in der Anwesenheit von 5-Me-dCTP unter Verwendung von biotinylierten und fluoreszierend markierten Primern, welche im Vorhinein definierte Restriktionsendonucleasestellen enthalten, vervielfältigt. Nach einer Affinitätsreinigung mit mit Streptavidin beschichteten magnetischen Kügelchen werden die Tag-cDNA-Konjugate von den Kügelchen abgespalten, mit T4-DNA-Polymerase in der Anwesenheit von dGTP behandelt, um die Tags einzelsträngig zu machen, und dann mit einem Repertoire von GMA-Kügelchen kombiniert, welche Tagkomplemente angebracht haben. Nach einer stringenten Hybridisierung und Ligierung werden die GMA-Kügelchen über FACS sortiert, um eine angereicherte Population von GMA-Kügelchen zu erzeugen, die mit cDNAs beladen sind. Die angereicherte Population von beladenen GMA-Kügelchen wird in einer planaren Anordnung in einer Flusszelle immobilisiert, worin eine Base-um-Base-Sequenzierung unter Verwendung codierter Adaptoren stattfindet.

[0145] Annähernd 5 µg von poly (A^+)-mRNA wird aus DBY746-Hefezellen unter Verwendung üblicher Proto-

kolle extrahiert. Die Synthese des ersten und zweiten Stranges der cDNA wird durch Kombinieren von 100–150 pMol des folgenden Primers (SEQ ID NO: 26):
5'-biotin-ACTAATCGTCTCACTATTTAATTAA[W,W,W,G]$_8$CC(T)$_{18}$V-3'
mit der poly(A$^+$)-mRNA unter Verwendung eines Stratagene (La Jolla, CA) cDNA Synthesis Kit in Übereinstimmung mit dem Protokoll des Herstellers durchgeführt. Dies ergibt cDNAs, deren Desoxycytosine des ersten Stranges an der 5-Kohlenstoff-Position methyliert sind. In der vorstehenden Formel ist „V" G, C oder A, ist „[W,W,W,G]" ein vier-Nucleotid-Wort, wie vorstehend beschrieben, ausgewählt aus Tabelle II, der einfach unterstrichene Abschnitt ist eine BsmBI-Erkennungsstelle und der doppelt unterstrichene Abschnitt ist eine PacI-Erkennungsstelle. Nach Größenfraktionierung (GIBCO-BRL cDNA Size Fractionation Kit) unter Verwendung üblicher Protokolle werden die cDNAs mit DpnII (New England Biolabs, Beverly, MA) unter Verwendung des Protokolls des Herstellers gespalten und mit mit Streptavidin beschichteten magnetischen Kügelchen affinitätsgereinigt (M-280-Kügelchen, Dynal A.S., Oslo, Norwegen). Die durch die Kügelchen gefangene DNA wird mit BsmBI gespalten, um die Tag-cDNA-Konjugate für die Clonierung in einen modifizierten pBCSK$^-$-Vektor (Stratagene, La Jolla, CA) unter Verwendung von Standardprotokollen freizusetzen. Der pBCSK$^-$-Vektor ist durch die Hinzufügung einer BbsI-Stelle durch Einführen des folgenden Fragments (SEQ ID NO: 27) in den mit KpnI/EcoRV gespaltenen Vektor modifiziert.

```
       CGAAGACCC
3'-CATGGCTTCTGGGGATA-5'
```

[0146] Mit BsmBI/DpnII-gespaltenes Tag-cDNA-Konjugat wird in den pBCSK$^-$-Vektor eingefügt, welcher vorher mit BbsI und BamHI gespalten wird. Nach der Ligierung wird der Vektor zur Vervielfältigung in den durch den Hersteller empfohlenen Wirt transfiziert.

[0147] Nach der Isolierung des vorstehenden pBCSK$^-$-Vektors aus einer Standard-Plasmid-Mini-Präparation werden die Tag-cDNA-Konjugate durch PCR in der Anwesenheit von 5-Me-dCTP unter Verwendung von 20-mer-Primern, welche komplementär zu Vektorsequenzen sind, die Tag-cDNA-Insertionen flankieren, vervielfältigt. Der "stromaufwärtige" Primer, das heißt angrenzend an den Tag, ist biotinyliert und der "stromabwärtige" Primer, das heißt angrenzend an die cDNA, ist mit Fluorescein markiert. Nach der Vervielfältigung wird das PCR-Produkt affinitätsgereinigt, dann mit PacI gespalten, um fluoreszierend markierte Tag-cDNA-Konjugate freizusetzen. Die Tags der Konjugate werden durch Behandlung mit T4-DNA-Polymerase in der Anwesenheit von dGTP einzelsträngig gemacht. Nachdem die Reaktion gedämpft ist, wird das Tag-cDNA-Konjugat durch Phenol-Chloroform-Extraktion gereinigt und mit 5,5 mm-GMA-Kügelchen kombiniert, welche Tagkomplemente tragen, wobei jedes Tagkomplement ein 5'-Phosphat besitzt. Die Hybridisierung wird unter stringenten Bedingungen in der Anwesenheit einer thermisch stabilen Ligase durchgeführt, so dass nur Tags, welche perfekt passende Duplexmoleküle mit ihren Komplementen bilden, ligiert werden. Die GMA-Kügelchen werden gewaschen und die beladenen Kügelchen werden durch FACS-Sortieren unter Verwendung der fluoreszierend markierten cDNAs zur Identifizierung beladener GMA-Kügelchen konzentriert. Die Tag-cDNA-Konjugate, welche an den GMA-Kügelchen angebracht sind, werden mit Dpn II gespalten, um die fluoreszierende Markierung zu entfernen, und mit alkalischer Phosphatase behandelt, um die cDNAs für das Sequenzieren vorzubereiten.

[0148] Der folgende Abspaltungsadaptor (SEQ ID NO: 28) wird an die mit DpnII-gespaltenen und mit Phosphatase behandelten cDNAs ligiert:

```
5'-pGATCAGCTGCTGCAAATTT
       pTCGACGACGTTTAAA
```

,

wonach das 3'-Phosphat durch alkalische Phosphatase entfernt wird, der 5'-Strang der cDNA mit T4-DNA-Kinase behandelt wird und der Bruch zwischen dem Abspaltungsadaptor und der cDNA ligiert wird. Nach Spaltung mit BbvI werden die codierten Adaptoren von Beispiel 1, wie vorstehend beschrieben, an die Enden der cDNAs ligiert.

[0149] Eine Fluss-Kammer (**500**), welche diagrammartig in Fig. 5 dargestellt ist, wird durch Ätzen einer Kavität, welche einen Flüssigkeitseinlass (**502**) und -Auslass (**504**) besitzt, unter Verwendung von Standard-Mikroverarbeitungsverfahren, zum Beispiel Ekstrom et al., internationale Patentanmeldung PCT/SE 91/00327 (WO 91/16966); Brown, US-Patent 4,911,782; Harrison et al., Anal. Chem., 64 (1992): 1926–1932; und ähnliches, in eine Glasplatte (**506**) hergestellt. Die Abmessungen der Fluss-Kammer (**500**) sind solche, dass beladene Mikropartikel (**508**), zum Beispiel GMA-Kügelchen, in einer dicht gepackten planaren einlagigen Schicht von 100.000–200.000 Kügelchen in die Kavität (**510**) gesetzt werden können. Die Kavität (**510**) wird durch anodisches Binden einer Glas-Abdeckplatte (**512**) auf die geätzte Glasplatte (**506**), zum Beispiel Pomerantz, US-Pa-

tent 3,397,279, zu einer geschlossenen Kammer mit einem Einlass und einem Auslas gemacht. Reagenzien werden in die Fluss-Kammer aus Spritzen-Pumpen (**514** bis **520**) durch den Ventilblock (**522**), gesteuert durch einen Mikroprozessor, wie er für gewöhnlich in automatisierten DNA- und Peptid-Syntheseautomaten, zum Beispiel Bridgham et al., US-Patent 4,668,479; Hood et al., US-Patent 4,252,769; Barstow et al., US-Patent 5,203,368; Hunkapiller, US-Patent 4,703,913 oder ähnliches, verwendet wird, dosiert.

[0150] Drei Zyklen von Ligierung, Identifizierung und Abspaltung werden in der Fluss-Kammer (**500**) durchgeführt, um die Sequenzen von 12 Nucleotiden an den Enden von jeder von annähernd 100.000 cDNAs zu ergeben. Nucleotide der cDNAs werden durch die wie in Beispiel 1 beschriebene Hybridisierung von Tagkomplementen an die codierten Adaptoren identifiziert. Spezifisch hybridisierte Tagkomplemente werden durch Anregung ihrer fluoreszierenden Markierungen mit dem Beleuchtungsstrahl (**524**) aus der Lichtquelle (**526**), welche ein Laser, eine Quecksilberdampflampe oder ähnliches sein kann, nachgewiesen. Der Beleuchtungsstrahl (**524**) tritt durch den Filter (**528**) und regt die fluoreszierenden Markierungen auf den Tagkomplementen, welche spezifisch an die codierten Adaptoren hybridisiert sind, in der Fluss-Kammer (**500**) an. Die sich ergebende Fluoreszenz (**530**) wird durch ein konfokales Mikroskop (**532**) gesammelt, durch den Filter geleitet (**534**) und auf eine CCD-Kamera (**536**) gerichtet, welche ein elektronisches Bild der Kügelchen-Anordnung zur Bearbeitung und Untersuchung durch den Arbeitsplatzrechner (**538**) erzeugt. Bevorzugt werden die cDNAs nach jedem Ligierungs- und Abspaltungsschritt mit Pronase™ oder einem ähnlichen Enzym behandelt. Codierte Adaptoren und T4-DNA-Ligase (Promega, Madison, WI) mit etwa 0,75 Einheiten pro µl werden durch die Fluss-Kammer mit einer Flussrate von etwa 1–2 µl pro Minute über etwa 20–30 Minuten bei 16°C geleitet, wonach 3'-Phosphate von den Adaptoren entfernt werden und die cDNAs durch Leiten eines Gemisches von alkalischer Phosphatase (New England Bioscience, Beverly, MA) mit 0,02 Einheiten pro µl und T4-DNA-Kinase (New England Bioscience, Beverly, MA) mit 7 Einheiten pro µl durch die Fluss-Kammer bei 37°C mit einer Flussrate von 1-2 µl pro Minute über 15–20 Minuten auf die Ligierung des zweiten Stranges vorbereitet werden. Die Ligierung wird durch T4-DNA-Ligase (0,75 Einheiten pro ml, Promega) durch die Fluss-Kammer über 20–30 Minuten bewerkstelligt. Tagkomplemente mit einer Konzentration von 25 nM werden durch die Fluss-Kammer mit einer Flussrate von 1–2 µl pro Minute über 10 Minuten bei 20°C geleitet, wonach die fluoreszierenden Markierungen, die durch die Tagkomplemente getragen werden, beleuchtet werden und die Fluoreszenz gesammelt wird. Die Tagkomplemente werden von den codierten Adaptoren mittels Durchleiten von Hybridisierungspuffer durch die Fluss-Kammer mit einer Flussrate von 1–2 µl pro Minute bei 55°C über 10 Minuten abgeschmolzen. Codierte Adaptoren werden von den cDNAs mittels Durchleiten von BbvI (New England Biosciences, Beverly, MA) mit 1 Einheit/µl mit einer Flussrate von 1–2 µl pro Minute über 20 Minuten bei 37°C abgespalten.

ANHANG I

Beispielhaftes Computerprogramm zur Erzeugung minimal kreuz-hybridisierender Sätze

(einzelsträngiger Tag/einzelsträngiges Tagkomplement)

Programm tagN

```
c
c           Programm tagN erzeugt minimal kreuzhybridisierende Sätze von
c           Untereinheiten, wobei gegeben ist i) N--Untereinheitenlänge und ii) einer
c           anfänglichen Untereinheitensequenz. tagN gibt vor, dass nur drei der vier
c           natürlichen Nucleotide in den Tags verwendet werden.
c
c
      character*1 subl(20)
      integer*2 mset(10000,20), nbase(20)
c
c
      write(*,*)'Gib Untereinheitenlänge ein'
      read(*,100)nsub
100   format(i2)
c
c
      write(*,*)'Gib Untereinheitensequenz ein'
      read(*,110)(subl(k),k=1,nsub)
110   format(20a1)
c
c
      ndiff=10
c
c
c                 Let a=1 c=2 g=3 & t=4
c
c
        do 800 kk=1,nsub
        if(subl(kk).eq.'a') then
           mset(1,kk)=1
           endif
                if(subl(kk).eq.'c') then
                   mset(1,kk)=2
                   endif
                        if(subl(kk).eq.'g') then
                           mset(1,kk)=3
                           endif
                                if(subl(kk).eq.'t') then
                                   mset(1,kk)=4
                                   endif
800   continue
c
c                 Erzeuge Satz von Untereinheiten, die sich von Unter 1
c                 durch mindestens n untersch. Nucleotide unterscheiden
c
c
c
c
        jj=1
c
c
```

```
          do 1000 k1=1,3
            do 1000 k2=1,3
              do 1000 k3=1,3
                do 1000 k4=1,3
                  do 1000 k5=1,3
                    do 1000 k6=1,3
                      do 1000 k7=1,3
                        do 1000 k8=1,3
                          do 1000 k9=1,3
                            do 1000 k10=1,3
          do 1000 k11=1,3
            do 1000 k12=1,3
              do 1000 k13=1,3
                do 1000 k14=1,3
                  do 1000 k15=1,3
                    do 1000 k16=1,3
                      do 1000 k17=1,3
                        do 1000 k18=1,3
                          do 1000 k19=1,3
                            do 1000 k20=1,3
c
c
                      nbase(1)=k1
                      nbase(2)=k2
                      nbase(3)=k3
                      nbase(4)=k4
                      nbase(5)=k5
                      nbase(6)=k6
                      nbase(7)=k7
                      nbase(8)=k8
                      nbase(9)=k9
                      nbase(10)=k10
                      nbase(11)=k11
                      nbase(12)=k12
                      nbase(13)=k13
                      nbase(14)=k14
                      nbase(15)=k15
                      nbase(16)=k16
                      nbase(17)=k17
                      nbase(18)=k18
                      nbase(19)=k19
                      nbase(20)=k20
c
c
                   do 1250 nn=1,jj
c
                    n=0
                    do 1200 j=1,nsub
                       if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1                    mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
     2                    mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
     3                    mset(nn,j).eq.4 .and. nbase(j).ne.4) then
                          n=n+1
                          endif
1200                      continue
c
c
            if(n.lt.ndiff) then
               goto 1000
               endif
1250        continue
c
c
```

```
                jj=jj+1
                write(*,130)(nbase(i),i=1,nsub),jj
                 do 1100 i=1,nsub
                    mset(jj,i)=nbase(i)
1100                   continue
c
c
1000        continue
c
c
             write(*,*)
130          format(10x,20(1x,i1),5x,i5)
             write(*,*)
             write(*,120) ii
120          format(1x,'Anzahl der Worte  =',i5)
c
c
             end
c
c                  ********************************************
c                  ********************************************
```

ANHANG II

Beispielhaftes Computerprogramm zur Erzeugung minimal kreuz-hybridisierender Sätze

(doppelsträngiger Tag/einzelsträngiges Tagkomplement)

```
Programm 3tagN
c
c
c       Programm 3tagN erzeugt minimal kreuzhybridisierende Sätze von
c       Triplexwörtern, wobei gegeben ist i) N--Untereinheitenlänge, ii) eine
c       anfängliche Untereinheitensequenz und iii) die Identität der Nucleotide, die die
c       Untereinheiten ausmacht, d.h. ob die Untereinheiten aus allen vier Nucleotiden
c       bestehen, oder ein Untersatz von Nucleotiden
c
c
c
c
      character*1 subl(20)
      integer*2 mset(10000,20), nbase(20)
c
c
      nsub=20
      ndiff=6
c
c
      write(*,*)'Gib Untereinheitensequenz ein: nur a & g'
      read(*,110)(subl(k),k=1,nsub)
110   format(20a1)
c
c             Erzeuge Satz von Wörtern, die sich von Unter1 durch
c             mindestens drei n untersch. Nucleotide unterscheiden.
c
c
c
c
c               Übersetze  a's & g's  in Zahlen mit        a=1 & g=2
c
c
        do 800 kk=1,nsub
        if(subl(kk).eq.'a') then
           mset(1,kk)=1
           endif
                if(subl(kk).eq.'g') then
                   mset(1,kk)=2
                   endif
800   continue
c
c
        jj=1
c
c
        do 1000 k1=1,2
          do 1000 k2=1,2
            do 1000 k3=1,2
```

```
                do 1000 k4=1,2
                  do 1000 k5=1,2
                    do 1000 k6=1,2
                      do 1000 k7=1,2
                        do 1000 k8=1,2
                          do 1000 k9=1,2
                            do 1000 k10=1,2
            do 1000 k11=1,2
               do 1000 k12=1,2
                  do 1000 k13=1,2
                     do 1000 k14=1,2
                        do 1000 k15=1,2
                           do 1000 k16=1,2
                              do 1000 k17=1,2
                                 do 1000 k18=1,2
                                    do 1000 k19=1,2
                                       do 1000 k20=1,2
c
c
                  nbase(1)=k1
                  nbase(2)=k2
                  nbase(3)=k3
                  nbase(4)=k4
                  nbase(5)=k5
                  nbase(6)=k6
                  nbase(7)=k7
                  nbase(8)=k8
                  nbase(9)=k9
                  nbase(10)=k10
                  nbase(11)=k11
                  nbase(12)=k12
                  nbase(13)=k13
                  nbase(14)=k14
                  nbase(15)=k15
                  nbase(16)=k16
                  nbase(17)=k17
                  nbase(18)=k18
                  nbase(19)=k19
                  nbase(20)=k20
c
c
               do 1250 nn=1,jj
c
                n=0
                do 1200 j=1,nsub
                   if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1                mset(nn,j).eq.2 .and. nbase(j).ne.2) then
                      n=n+1
                      endif
1200                  continue
c
c
         if(n.lt.ndiff) then
            goto 1000
            endif
1250     continue
c
```

```
c
              jj=jj+1
              write(*,130)(nbase(i),i=1,nsub),jj
               do 1100 i=1,nsub
                  mset(jj,i)=nbase(i)
1100                 continue
c
c
1000      continue
c
c
           write(*,*)
130        format(5x,20(1x,i1),5x,i5)
           write(*,*)
           write(*,120) jj
120        format(1x, 'Anzahl der Worte   =',i5)
c
c
           end
```

SEQUENZPROTOKOLL

(1) ALLGEMEINE INFORMATION

(i) ANMELDER: Lynx Therapeutics, Inc.

(ii) TITEL DER ERFINDUNG: Massiv paralleles Signatur-Sequenzieren durch Ligierung von codierten Adaptoren

(iii) ZAHL DER SEQUENZEN: 28

(iv) KORRESPONDENZADRESSE:
- (A) ADRESSAT: Dehlinger & Associates
- (B) STRASSE: 350 Cambrdige Avenue, Suite 250
- (C) STADT: Palo Alto
- (D) STAAT: CA
- (E) LAND: USA
- (F) PLZ: 94306

(v) IM COMPUTER LESBARE FORM:
- (A) MEDIUMTYP: Diskette
- (B) COMPUTER: IMB-PC-kompatibel
- (C) BETRIEBSSYSTEM: PC-DOS/MS-DOS
- (D) SOFTWARE: PatentIn Ausgabe #1.0, Version #1.25

(vi) DATEN DER AKTUELLEN ANMELDUNG:
- (A) ANMELDUNGSNUMMER:
- (B) EINREICHUNGSDATUM:

(vii) DATEN FRÜHERER ANMELDUNG:
- (A) ANMELDUNGSNUMMER: US 08/689,587
- (B) EINREICHUNGSDATUM: 12-AUG-96

(vii) DATEN FRÜHERER ANMELDUNG:
- (A) ANMELDUNGSNUMMER: US 08/659,453
- (B) EINREICHUNGSDATUM: 06-JUN-96

(viii) ANWALT / VERTRETERINFORMATION:
- (A) NAME: Powers, Vincent M.
- (B) REGISTRIERUNGSNUMMER: 36,246
- (C) REFERENZ/AKTENNUMMER: 5525-0029.41/808-1wo

(ix) TELEKOMMUNIKATIONSINFORMATION:
- (A) TELEFON: (415) 324-0880
- (B) TELEFAX: (415) 324-0960

(i) SEQUENZCHARAKTERISTIK:
- (A) LÄNGE: 28 Nucleotide
- (B) ART: Nucleinsäure
- (C) STRANGFORM: doppelt
- (D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 1:

TGGATTCTAG AGAGAGAGAG AGAGAGAG 28

(2) INFORMATION FÜR SEQ ID NO: 2:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 16 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 2:

NNGGATGNNN NNNNNN 16

(2) INFORMATION FÜR SEQ ID NO: 3:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 11 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: einzeln
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 3:

NRRGATCYNN N 11

(2) INFORMATION FÜR SEQ ID NO: 4:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 20 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 4:

AGTGGCTGGG CATCGGACCG 20

(2) INFORMATION FÜR SEQ ID NO: 5:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 20 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 5:

GGGGCCCAGT CAGCGTCGAT 20

(2) INFORMATION FÜR SEQ ID NO: 6:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 70 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 6:

AAAAGGAGGA GGCCTTGATA GAGAGGACCT GTTTAAACGT TTAAACGGAT 50
CCTCTTCCTC TTCCTCTTCC 70

(2) INFORMATION FÜR SEQ ID NO: 7:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 34 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: einzeln
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 7:

CTAAACCATT GGTATGGGCC AGTGAATTGT AATA 34

(2) INFORMATION FÜR SEQ ID NO: 8:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 55 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: einzelne
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 8:

CGCGCAGCCC GCATCGTTTA TGCTACAGAC TGTCAGTGCA 40

GCTCTCCGAT CCAAA 55

(2) INFORMATION FÜR SEQ ID NO: 9:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 16 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 9:

TCGATGGGCT AGATTT 16

(2) INFORMATION FÜR SEQ ID NO: 10:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 10:

ANNNTACAGC TGCATCCCTT GGCGCTGAGG 30

(2) INFORMATION FÜR SEQ ID NO: 11:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 11:

NANNTACAGC TGCATCCCTG GGCCTGTAAG 30

(2) INFORMATION FÜR SEQ ID NO: 12:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 12:

CNNNTACAGC TGCATCCCTT GACGGGTCTC 30

(2) INFORMATION FÜR SEQ ID NO: 13:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 13:

NCNNTACAGC TGCATCCCTG CCCGCACAGT 30

(2) INFORMATION FÜR SEQ ID NO: 14:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 14:

GNNNTACAGC TGCATCCCTT CGCCTCGGAC 30

(2) INFORMATION FÜR SEQ ID NO: 15:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 15:

NGNNTACAGC TGCATCCCTG ATCCGCTAGC 30

(2) INFORMATION FÜR SEQ ID NO: 16:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 16:

TNNNTACAGC TGCATCCCTT CCGAACCCGC 30

(2) INFORMATION FÜR SEQ ID NO: 17:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 17:

NTNNTACAGC TGCATCCCTG AGGGGGATAG 30

(2) INFORMATION FÜR SEQ ID NO: 18:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 18:

NNANTACAGC TGCATCCCTT CCCGCTACAC 30

(2) INFORMATION FÜR SEQ ID NO: 19:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 19:

NNNATACAGC TGCATCCCTG ACTCCCCGAG 30

(2) INFORMATION FÜR SEQ ID NO: 20:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 20:

NNCNTACAGC TGCATCCCTG TGTTGCGCGG 30

(2) INFORMATION FÜR SEQ ID NO: 21:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 21:

NNNCTACAGC TGCATCCCTC TACAGCAGCG 30

(2) INFORMATION FÜR SEQ ID NO: 22:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 22:

NNGNTACAGC TGCATCCCTG TCGCGTCGTT 30

(2) INFORMATION FÜR SEQ ID NO: 23:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide

(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 23:

NNNGTACAGC TGCATCCCTC GGAGCAACCT 30

(2) INFORMATION FÜR SEQ ID NO: 24:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 24:

NNTNTACAGC TGCATCCCTG GTGACCGTAG 30

(2) INFORMATION FÜR SEQ ID NO: 25:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 30 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 25:

NNNTTACAGC TGCATCCCTC CCCTGTCGGA 30

(2) INFORMATION FÜR SEQ ID NO: 26:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 78 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: einzeln
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 26:

ACTAATCGTC TCACTATTTA ATTAANNNNN NNNNNNNNNN 40
NNNNNNNNNN NNNNNNNGGT TTTTTTTTTT TTTTTTTV 78

(2) INFORMATION FÜR SEQ ID NO: 27:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 17 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 27:

ATAGGGGTCT TCGGTAC 17

(2) INFORMATION FÜR SEQ ID NO: 28:

(i) SEQUENZCHARAKTERISTIK:
(A) LÄNGE: 19 Nucleotide
(B) ART: Nucleinsäure
(C) STRANGFORM: doppelt
(D) TOPOLOGIE: linear

(xi) SEQUENZBESCHREIBUNG: SEQ ID NO: 28:

GATCAGCTGC TGCAAATTT 19

Patentansprüche

1. Verfahren zur Ermittlung einer Nucleotidsequenz an einem Ende eines Polynucleotids, wobei das Verfahren die Schritte umfasst:
(a) Anwenden einer Vielzahl verschiedener codierter Adaptoren auf das Polynucleotid, wobei jeder codierte Adaptor eine doppelsträngige Desoxyribonucleinsäure ist, umfassend (i) einen Oligonucleotid-Tag ausgewählt aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden und (ii) einen überhängenden Strang, welcher in einer bekannten Art und Weise dem Oligonucleotid-Tag entspricht;
wobei jedes Oligonucleotid des minimal kreuz-hybridisierenden Satzes von Oligonucleotiden sich von jedem anderen Oligonucleotid des Satzes in mindestens zwei Nucleotiden unterscheidet;
(b) Ligieren von codierten Adaptoren, deren überhängende Stränge perfekt passende Duplexe mit dem Ende bilden, an das Ende des Polynucleotids; und
(c) für jedes aus einer Vielzahl von Nucleotiden in dem Ende des Polynucleotids, spezifisches Hybridisieren eines markierten Tagkomplements an das Oligonucleotid-Tag jedes codierten Adaptors, der daran ligiert ist, wobei das hybridisierte Tagkomplement in einer bekannten Art und Weise der Identität des Nucleotids entspricht, wodurch jedes Nucleotid aus der Vielzahl von Nucleotiden in dem Ende des Polynucleotids identifiziert wird.

2. Verfahren nach Anspruch 1 zur Ermittlung von Nucleotidsequenzen einer Vielzahl von Polynucleotiden, wobei das Verfahren des weiteren, vor Schritt (a), die Schritte umfasst:
(i) Anbringen eines ersten Oligonucleotid-Tags aus einem Repertoire von Tags an jedes Polynucleotid in einer Population von Polynucleotiden, wobei jeder erste Oligonucleotid-Tag aus dem Repertoire ausgewählt ist aus einem ersten minimal kreuz-hybridisierenden Satz von Oligonucleotiden, und wobei jedes Oligonucleotid aus dem ersten minimal kreuz-hybridisierenden Satz sich von jedem anderen Oligonucleotid aus dem ersten Satz in mindestens zwei Nucleotiden unterscheidet;
(ii) Auswerten der Population von Polynucleotiden zur Erzeugung einer Probe von Polynucleotiden, so dass im Wesentlichen alle verschiedenen Polynucleotide in der Probe verschiedene erste Oligonucleotid-Tags gebunden haben; und
(iii) Sortieren der Polynucleotide der Probe durch spezifisches Hybridisieren der ersten Oligonucleotid-Tags mit deren entsprechenden Komplementen, wobei die entsprechenden Komplemente als einheitliche Populationen von im Wesentlichen identischen Oligonucleotiden in räumlich diskreten Regionen auf dem einen oder mehre-

ren Festphasenträgern befestigt sind;
und wobei Schritte (a)–(c) auf jedes Polynucleotid der Vielzahl angewandt werden.

3. Verfahren nach Anspruch 2 zur Identifizierung einer Population von mRNA-Molekülen, wobei die Polynucleotide cDNA-Moleküle sind und wobei Schritt (i) umfasst:
Erzeugen einer Population von cDNA-Molekülen aus der Population von mRNA-Molekülen, so dass jedes cDNA-Molekül einen ersten Oligonucleotid-Tag gebunden hat, wobei die ersten Oligonucleotid-Tags ausgewählt sind aus einem ersten minimal kreuz-hybridierenden Satz von Oligonucleotiden, wobei jedes Oligonucleotid aus dem ersten minimal kreuz-hybridisierenden Satz sich von jedem anderen Oligonucleotid des ersten Satzes in mindestens zwei Nucleotiden unterscheidet;
und des weiteren umfassend den Schritt:
Identifizieren der Population von mRNA-Molekülen durch die Häufigkeitsverteilung der Abschnitte von Sequenzen der cDNA-Moleküle.

4. Verfahren nach Anspruch 1, wobei der Schritt der Ligierung das Ligieren einer Vielzahl von verschiedenen codierten Adaptoren an das Ende des Polynucleotids einschließt, so dass die überhängenden Stränge der Vielzahl der verschiedenen codierten Adaptoren komplementär zu einer Vielzahl von verschiedenen Abschnitten des Stranges des Polynucleotids sind, und dass ein Eins-zu-Eins-Verhältnis zwischen den verschiedenen codierten Adaptoren und den verschiedenen Abschnitten des Stranges besteht.

5. Verfahren nach Anspruch 4, wobei die verschiedenen Abschnitte des Stranges des Polynucleotids zusammenhängend sind.

6. Verfahren nach Anspruch 2, des weiteren einschließend die Schritte (d) des Abspaltens der codierten Adaptoren von den Polynucleotiden mit einer Nuclease, die eine Nucleaseerkennungsstelle hat, die verschieden von ihrer Schneidestelle ist, so dass ein neuer überhängender Strang an dem Ende von jedem der Polynucleotide gebildet wird, und (e) des Wiederholens der Schritte (a) bis (d).

7. Verfahren nach Anspruch 1, des weiteren einschließend die Schritte:
(d) Abspalten des codierten Adaptors vom Ende des Polynucleotids mit einer Nuclease, die eine Nucleaseerkennungsstelle hat, die verschieden von ihrer Schneidestelle ist, so dass ein neuer überhängender Strang an dem Ende des Polynucleotids gebildet wird; und
(e) Wiederholen der Schritte (a) bis (d).

8. Verfahren nach Anspruch 1 oder 7, wobei der überhängende Strang des codierten Adaptors zwei bis sechs Nucleotide enthält, und wobei der Schritt des Identifizierens das spezifische Hybridisieren nacheinander der Tag-Komplemente an das Oligonucleotid-Tag umfasst, so dass die Identität jedes Nucleotids in dem Abschnitt des Polynucleotids nacheinander bestimmt wird.

9. Verfahren nach einem der Ansprüche 1, 7 oder 8, wobei der Schritt des Identifizierens des weiteren die Bereitstellung einer Anzahl von Sätzen von Tag-Komplementen einschließt, die äquivalent zu der Anzahl von Nucleotiden sind, die in dem Abschnitt des Polynucleotids identifiziert werden sollen.

10. Verfahren nach Anspruch 9, wobei der Schritt des Identifizierens des weiteren einschließt: Bereitstellen der Tag-Komplemente in jedem der Sätze, die in der Lage sind, die Gegenwart eines zuvor bestimmten Nucleotids durch ein Signal anzuzeigen, das durch eine ein Fluoreszenzsignal erzeugende Einheit erzeugt wird, wobei für jede Art von Nucleotid eine unterschiedliche, ein Fluoreszenzsignal erzeugende Einheit vorhanden ist.

11. Verfahren nach einem der Ansprüche 1 oder 7 bis 10, wobei die Oligonucleotid-Tags der codierten Adaptoren einzelsträngig sind und wobei die Tag-Komplemente zu den Oligonucleotid-Tags einzelsträngig sind, so dass eine spezifische Hybridisierung zwischen einem Oligonucleotid-Tag und seinem entsprechenden Tag-Komplement durch Watson-Crick-Basenpaarung erfolgt.

12. Verfahren nach Anspruch 11, wobei die codierten Adaptoren die folgende Form haben:

5'-p(N)n(N)r(N)s(N)q(N)t-3'

z(N')r(N')s(N')q-5'

oder

p(N)r(N)s(N)q(N)t-3'

3'-z(N)n(N')r(N')s(N')q-5'

wobei N ein Nucleotid ist und N' sein Komplement ist, p eine Phosphatgruppe ist, z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist, n eine ganze Zahl zwischen 2 und einschließlich 6 ist, r eine ganze Zahl zwischen 0 und einschließlich 18 ist, s eine ganze Zahl ist, welche entweder zwischen 4 und einschließlich 6 ist, wenn der codierte Adaptor eine Nucleaseerkennungsstelle hat, oder 0 ist, wenn keine Nucleaseerkennungsstelle vorhanden ist, q eine ganze Zahl größer oder gleich 0 ist und t eine ganze Zahl größer oder gleich 8 ist.

13. Verfahren nach Anspruch 12, wobei r zwischen 0 und einschließlich 12 liegt, t eine ganze Zahl zwischen 8 und einschließlich 20 ist und z eine Phosphatgruppe ist.

14. Verfahren nach Anspruch 1, wobei die Oligonucleotid-Tags der codierten Adaptoren doppelsträngig sind und wobei die Tag-Komplemente zu den Oligonucleotid-Tags einzelsträngig sind, so dass spezifische Hybridisierung zwischen einem Oligonucleotid-Tag und seinem entsprechenden Tag-Komplement durch die Bildung eines Hoogsteen- oder Revers-Hoogsteen-Triplex erfolgt.

15. Verfahren nach Anspruch 14, wobei die codierten Adaptoren die Form haben:

5'-p(N)n(N)r(N)s(N)q(N)t-3'

z(N')r(N')s(N')q(N')t-5'

oder

p(N)r(N)s(N)q(N)t-3'

3'-z(N)n(N')r(N')s(N')q(N')t-5'

wobei N ein Nucleotid und N' sein Komplement ist, p eine Phosphatgruppe ist, z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist, n eine ganze Zahl zwischen 2 und einschließlich 6 ist, r eine ganz Zahl zwischen 0 und einschließlich 18 ist, s eine ganze Zahl ist, welche entweder zwischen 4 und einschließlich 6 liegt, wenn der codierte Adaptor eine Nucleaseerkennungsstelle hat, oder 0 ist, wenn keine Nucleaseerkennungsstelle vorhanden ist, q eine ganze Zahl größer oder gleich 0 ist und t eine ganze Zahl größer oder gleich 8 ist.

16. Verfahren nach Anspruch 15, wobei r zwischen 0 und einschließlich 12 liegt, t eine ganze Zahl zwischen 12 und einschließlich 24 ist und z eine Phosphatgruppe ist.

17. Verfahren nach einem der Ansprüche 14 bis 16, wobei die Mitglieder des minimal kreuz-hybridisierenden Satzes sich von jedem anderen Mitglied in mindestens sechs Nucleotiden unterscheiden.

18. Zusammensetzung, umfassend eine Vielzahl von doppelsträngigen Oligonucleotidadaptoren, wobei die Adaptoren die Form haben:

5'-p(N)n(N)r(N)s(N)q(N)t-3'

z(N')r(N')s(N')q-5'

oder

p(N)r(N)s(N)q(N)t-3'

3'-z(N)n(N')r(N')s(N')q-5'

wobei jedes (N)t ein einzigartiger, einzelsträngiger Oligonucleotid-Tag ist und ausgewählt ist aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden, so dass jedes Oligonucleotid des Satzes sich von jedem anderen Oligonucleotid des Satzes in mindestens zwei Nucleotiden unterscheidet;

wobei: N ein Nucleotid ist und N' sein Komplement ist,
p eine Phosphatgruppe ist,
z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist,
n eine ganze Zahl zwischen 2 und einschließlich 6 ist,
r eine ganze Zahl zwischen 0 und einschließlich 18 ist,
s eine ganze Zahl zwischen 4 und einschließlich 6 ist,
der Adaptor in einem doppelsträngigen Abschnitt, getrennt von dem Oligonucleotid-Tag, eine Nucleaseerkennungsstelle einer Nuclease aufweist, deren Erkennungsstelle separat von ihrer Schneidestelle ist,
q eine ganze Zahl größer oder gleich 0 ist und
t eine ganze Zahl größer oder gleich 8 ist.

19. Zusammensetzung, umfassend eine Vielzahl doppelsträngiger Oligonucleotidadaptoren, wobei die Adaptoren die Form haben:

5’-p(N)n(N)r(N)s(N)q(N)t-3’

z(N’)r(N’)s(N’)q(N’)t-5’

oder

p(N)r(N)s(N)q(N)t-3’

3’-z(N)n(N’)r(N’)s(N’)q(N’)t-5’

wobei jedes (N)t ein einzigartiger, doppelsträngiger Oligonucleotid-Tag (N')t ist, ausgewählt aus einem minimal kreuz-hybridisierenden Satz von Oligonucleotiden, so dass jedes Oligonucleotid des Satzes sich von jedem anderen Oligonucleotid des Satzes in mindestens zwei Basenpaaren unterscheidet;
wobei: N ein Nucleotid und N' sein Komplement ist,
p eine Phosphatgruppe ist,
z eine 3'-Hydroxyl- oder eine 3'-Blockierungsgruppe ist,
n eine ganze Zahl zwischen 2 und einschließlich 6 ist,
r eine ganze Zahl zwischen 0 und einschließlich 18 ist,
s eine ganze Zahl zwischen 4 und einschließlich 6 ist
der Adaptor in einem doppelsträngigen Abschnitt, getrennt von dem Oligonucleotid-Tag, eine Nucleaseerkennungsstelle einer Nuclease aufweist, deren Erkennungsstelle separat von ihrer Schneidestelle ist,
q eine ganze Zahl größer oder gleich 0 ist und
t eine ganze Zahl größer oder gleich 8 ist.

Es folgen 10 Blatt Zeichnungen

Anhängende Zeichnungen

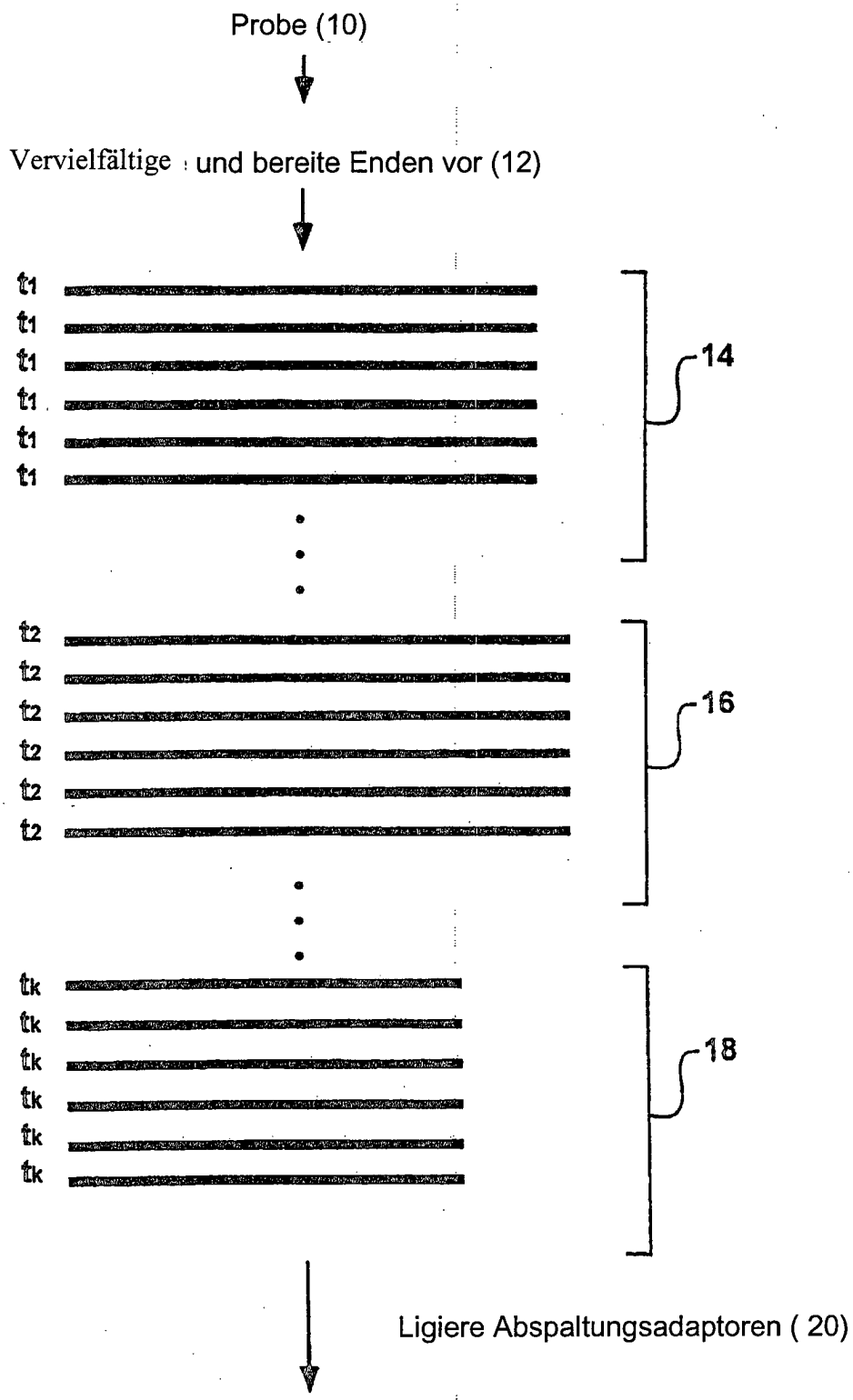


Fig. 1A

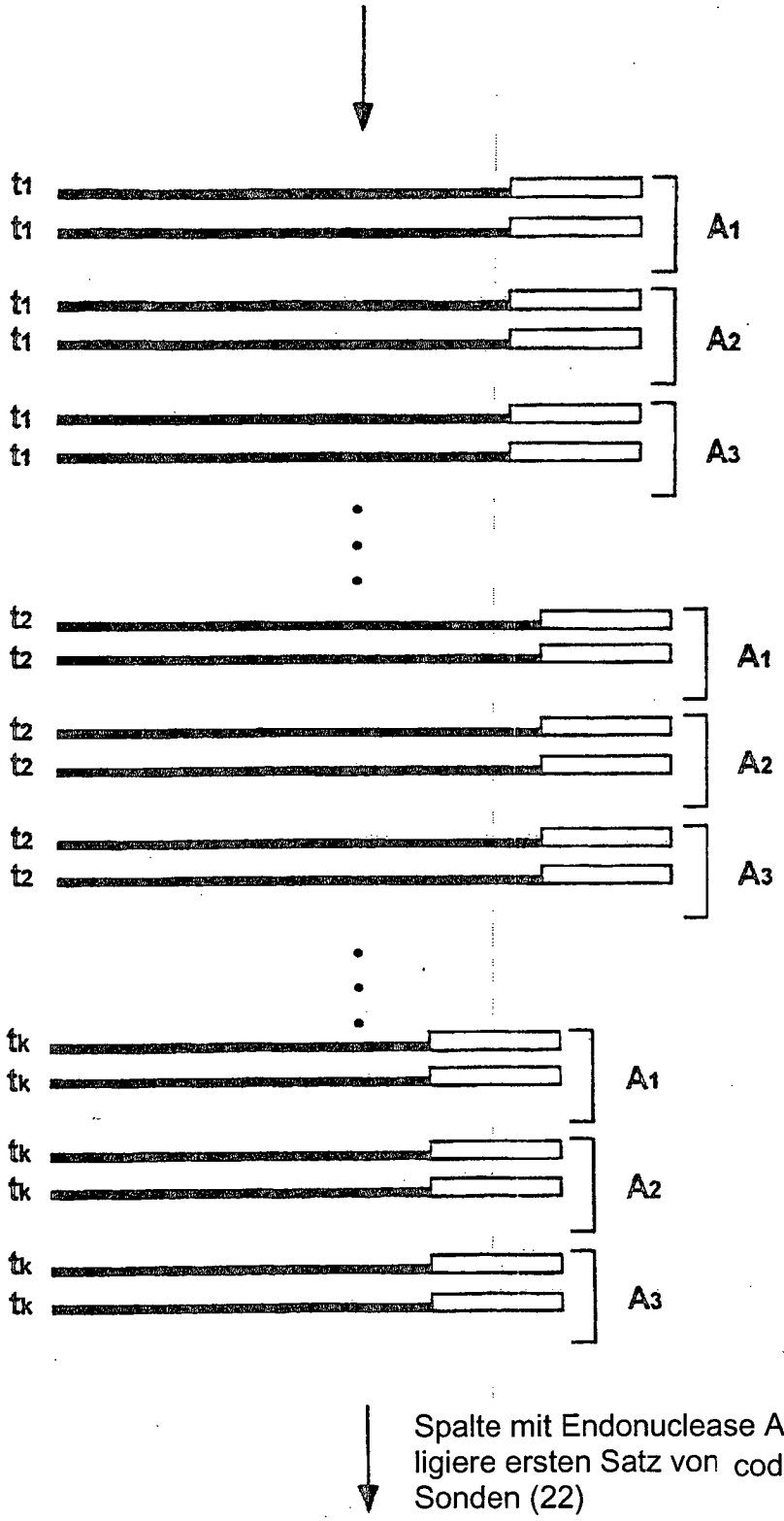

t1
t1
A1
t1
t1
A2
t1
t1
A3
t2
t2
A1
t2
t2
A2
t2
t2
A3
tk
tk
A1
tk
tk
A2
tk
tk
A3
Spalte mit Endonuclease A1 und ligiere ersten Satz von codierten Sonden (22)

Fig. 1B

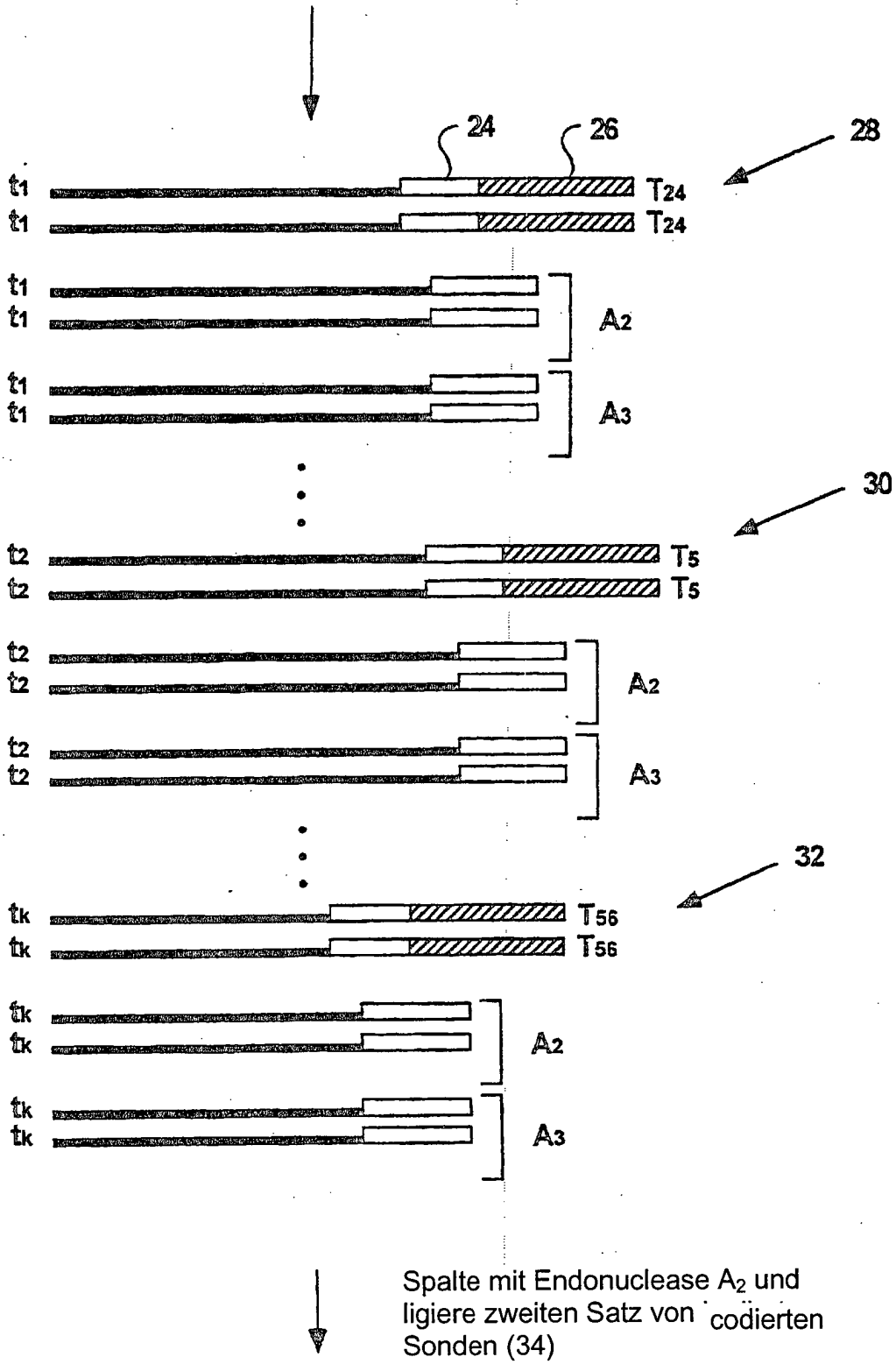

24
26
28
t1
t1
T24
T24
t1
t1
A2
t1
t1
A3
30
t2
t2
T5
T5
t2
t2
A2
t2
t2
A3
32
tk
tk
T56
T56
tk
tk
A2
tk
tk
A3
Spalte mit Endonuclease A2 und ligiere zweiten Satz von codierten Sonden (34)

Fig. 1C

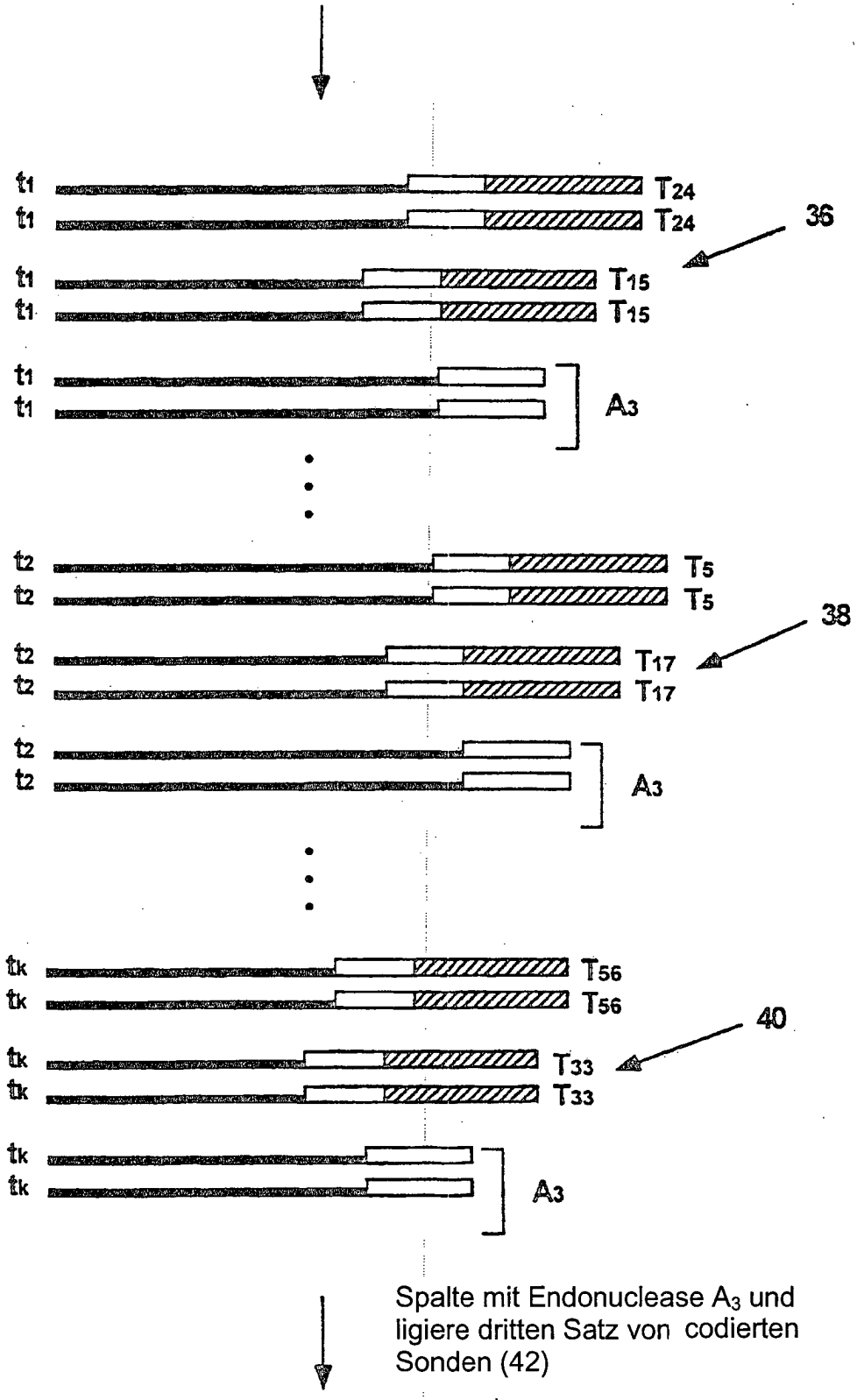

t1
T24
36
T15
A3
t2
T5
38
T17
tk
T56
40
T33
Spalte mit Endonuclease A3 und ligiere dritten Satz von codierten Sonden (42)

Fig. 1D

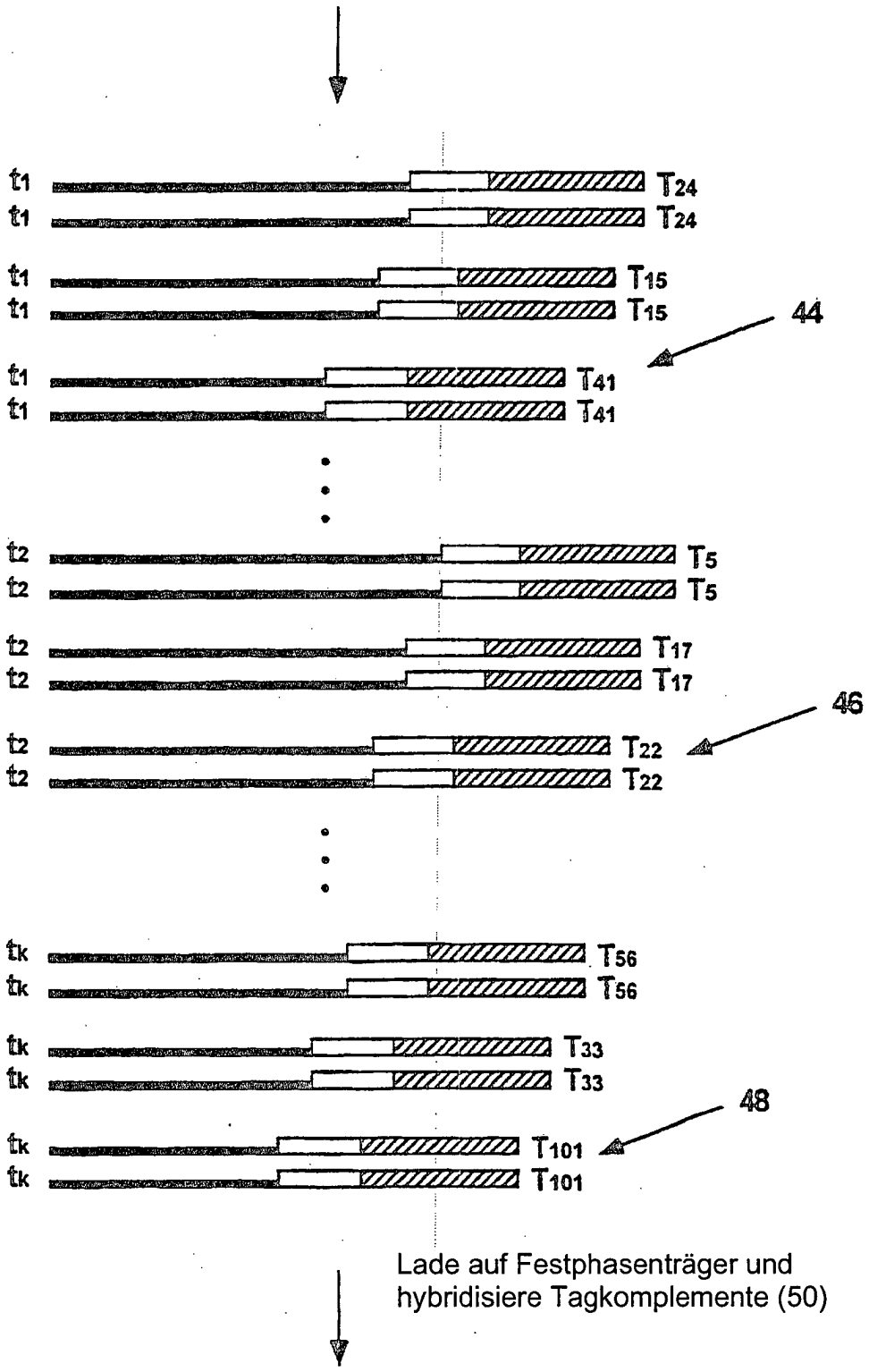

t1
T24
t1
T24
t1
T15
t1
T15
44
t1
T41
t1
T41
t2
T5
t2
T5
t2
T17
t2
T17
46
t2
T22
t2
T22
tk
T56
tk
T56
tk
T33
tk
T33
48
tk
T101
tk
T101
Lade auf Festphasenträger und
hybridisiere Tagkomplemente (50)

Fig. 1E

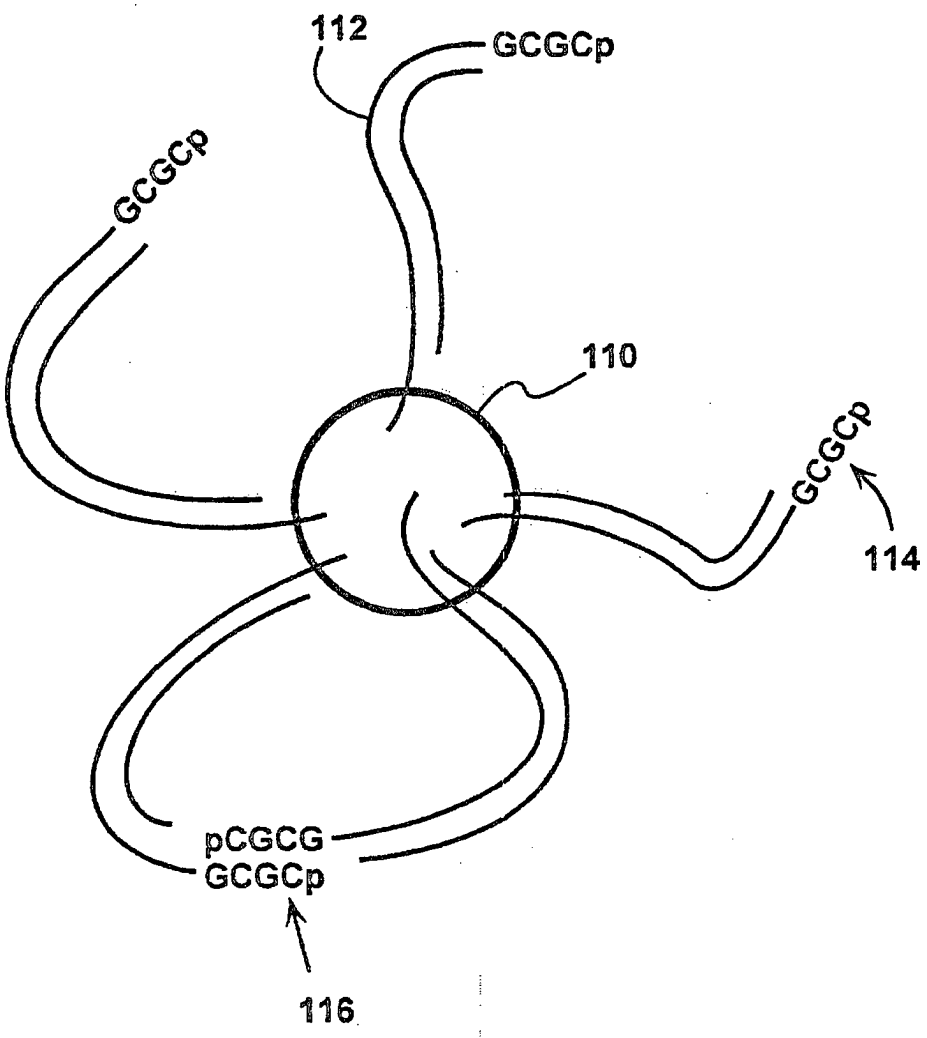


Fig. 2

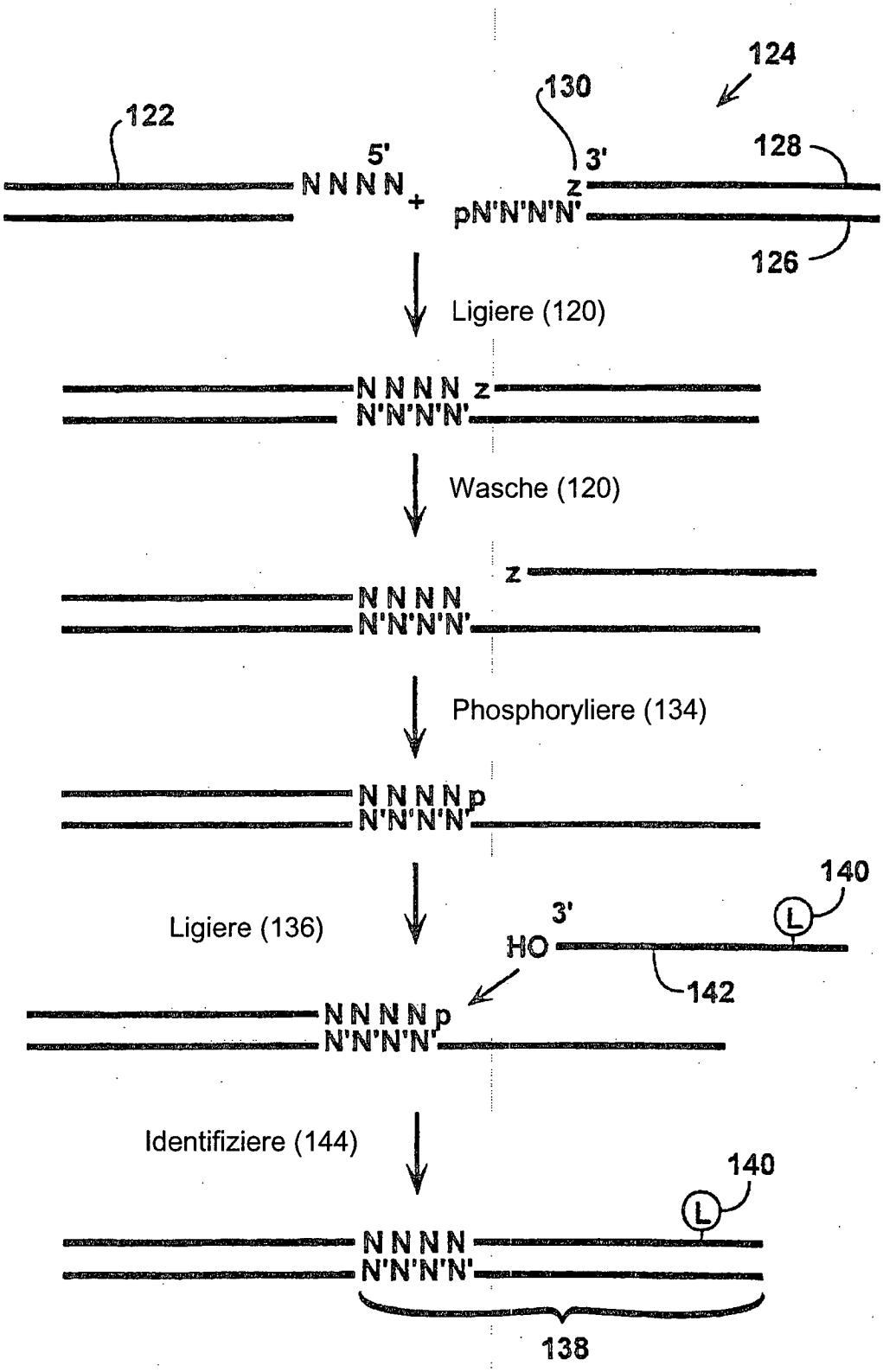

124
130
122
5'
3'
128
NNNN
+
z
pN'N'N'N'
126
Ligiere (120)
NNNN z
N'N'N'N'
Wasche (120)
z
NNNN
N'N'N'N'
Phosphoryliere (134)
NNNNp
N'N'N'N'
140
Ligiere (136)
3'
L
HO
142
NNNNp
N'N'N'N'
Identifiziere (144)
140
L
NNNN
N'N'N'N'
138

Fig. 3A

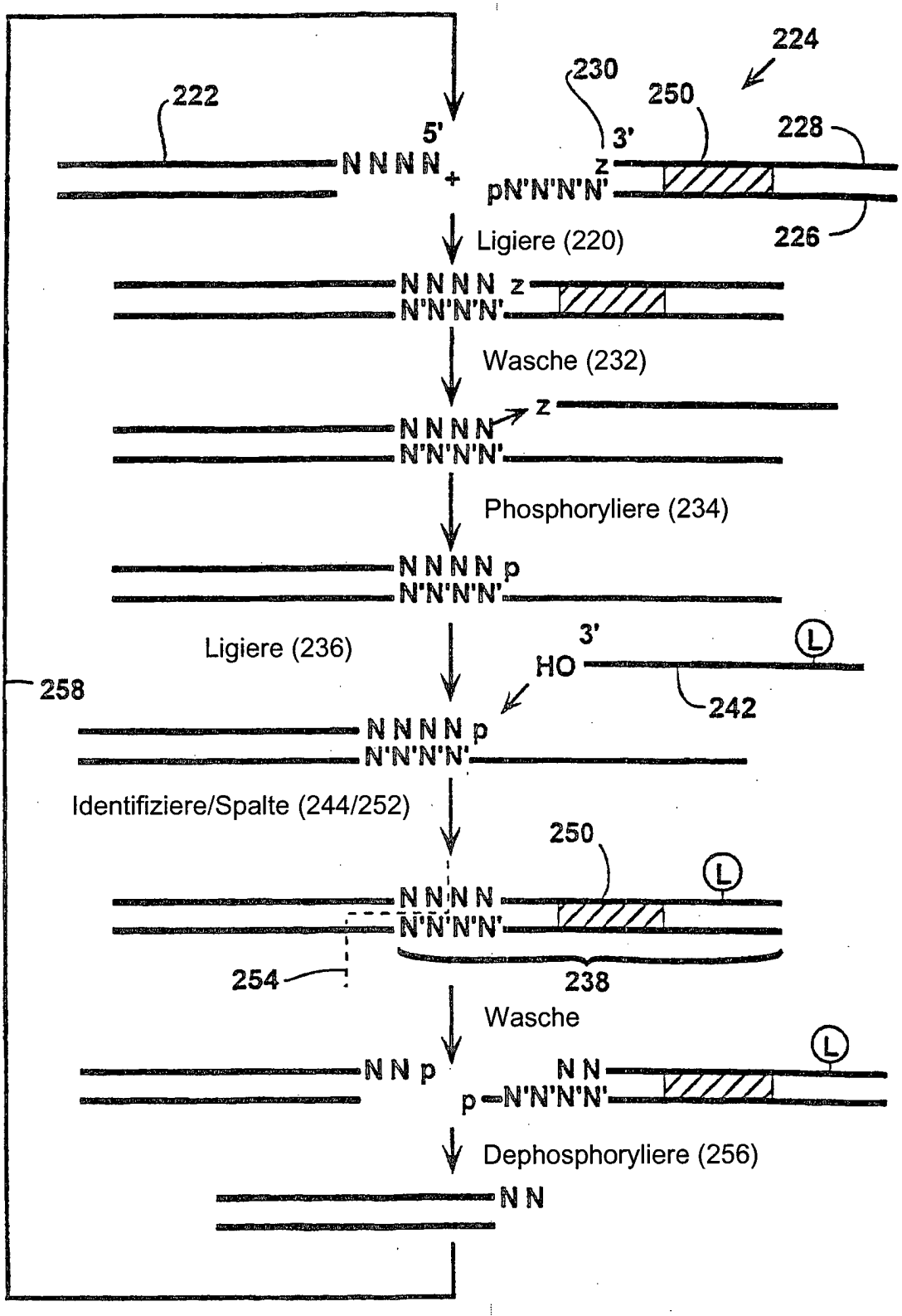

224
230
250
222
5'
3'
228
NNNN
z
pN'N'N'N'
226
Ligiere (220)
NNNN z
N'N'N'N'
Wasche (232)
z
NNNN
N'N'N'N'
Phosphoryliere (234)
NNNN p
N'N'N'N'
Ligiere (236)
3'
HO
L
258
242
NNNN p
N'N'N'N'
Identifiziere/Spalte (244/252)
250
L
NNNN
N'N'N'N'
254
238
Wasche
L
NN p
NN
p N'N'N'N'
Dephosphoryliere (256)
NN

Fig. 3B

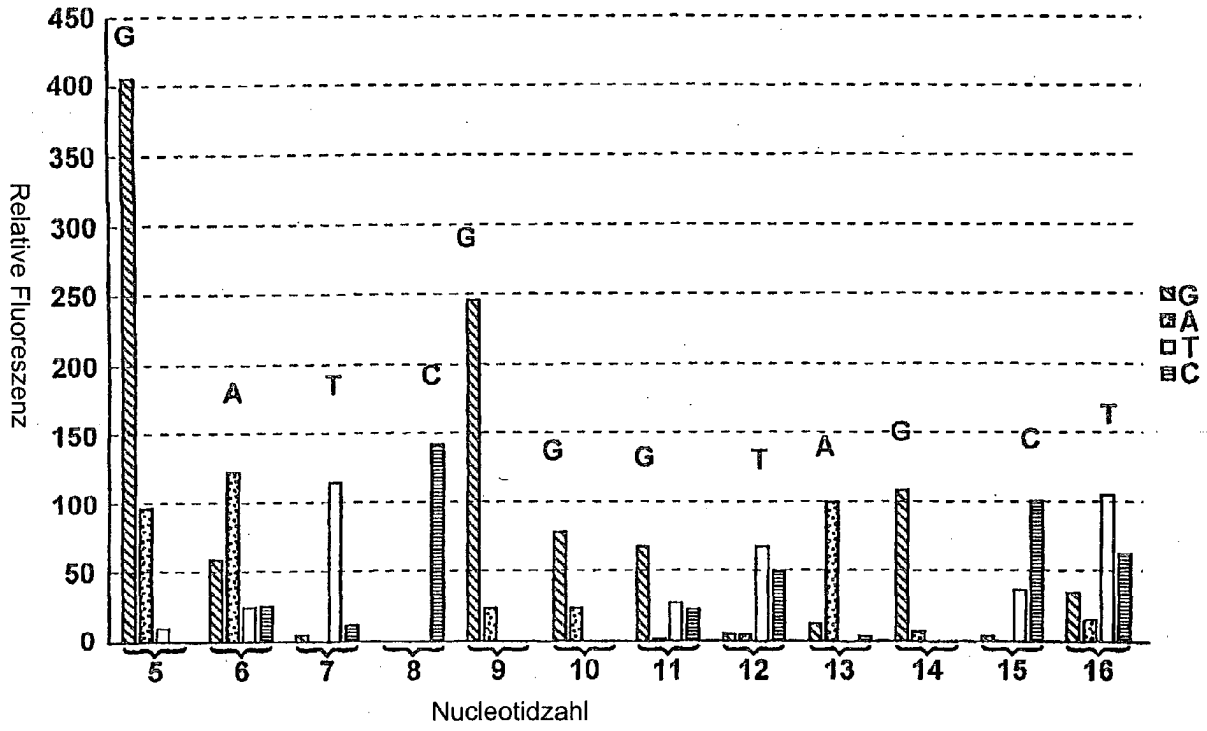

450
400
350
300
250
200
150
100
50
0
Relative Fluoreszenz
G
A
T
C
5
6
7
8
9
10
11
12
13
14
15
16
Nucleotidzahl

Fig. 4

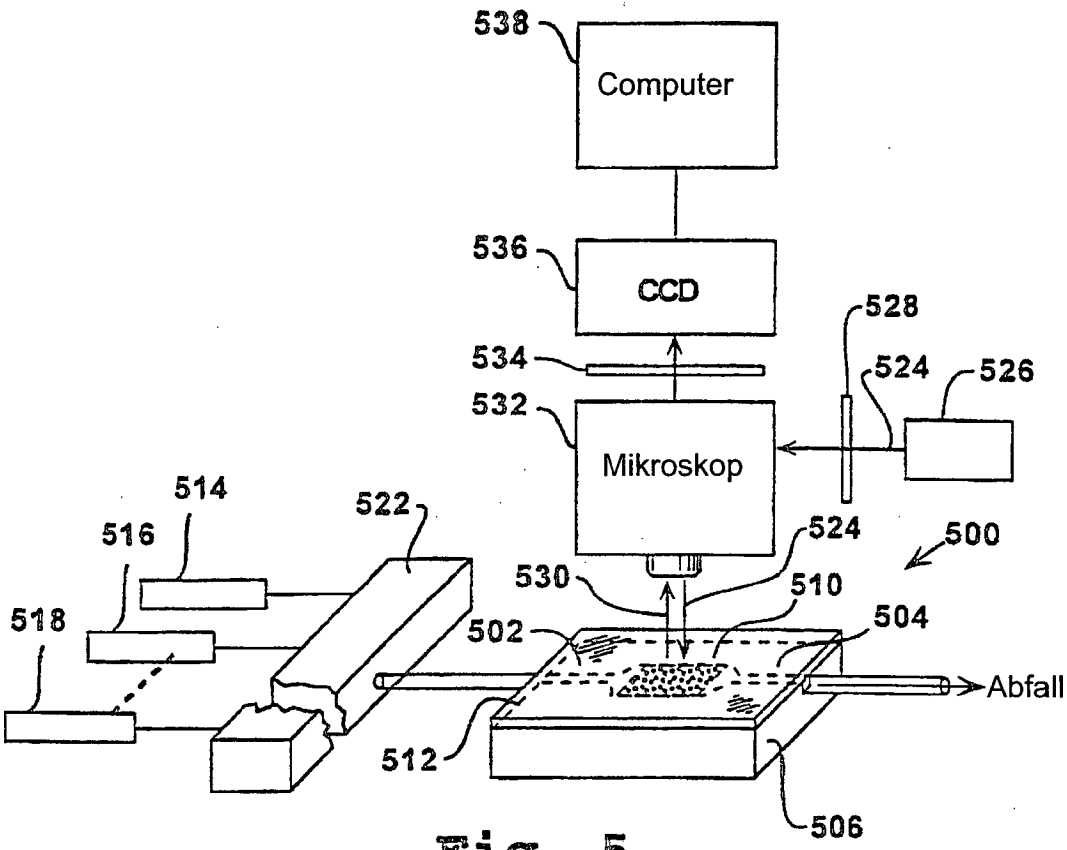

538
Computer
536
CCD
528
534
524
526
532
Mikroskop
514
522
516
524
500
530
510
518
502
504
Abfall
512
506

Fig. 5