



(19)  
Bundesrepublik Deutschland  
Deutsches Patent- und Markenamt

(10) **DE 600 36 645 T2 2008.07.17**

(12) **Übersetzung der europäischen Patentschrift**

(97) **EP 1 250 449 B1**

(51) Int Cl.<sup>8</sup>: **C12Q 1/68 (2006.01)**

(21) Deutsches Aktenzeichen: **600 36 645.6**

(86) PCT-Aktenzeichen: **PCT/US00/40517**

(96) Europäisches Aktenzeichen: **00 960 208.7**

(87) PCT-Veröffentlichungs-Nr.: **WO 2001/011088**

(86) PCT-Anmeldetag: **01.08.2000**

(87) Veröffentlichungstag

der PCT-Anmeldung: **15.02.2001**

(97) Erstveröffentlichung durch das EPA: **23.10.2002**

(97) Veröffentlichungstag

der Patenterteilung beim EPA: **03.10.2007**

(47) Veröffentlichungstag im Patentblatt: **17.07.2008**

(30) Unionspriorität:

**368810 05.08.1999 US**

**PCT/US99/20209 01.09.1999 WO**

**591435 09.06.2000 US**

(84) Benannte Vertragsstaaten:

**AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT,  
LI, LU, MC, NL, PT, SE**

(73) Patentinhaber:

**Evolutionary Genomics LLC, Lafayette, Col., US**

(72) Erfinder:

**MESSIER, Walter, Longmont, CO 80501, US;**

**SIKELA, James M., Englewood, CO 80111, US**

(74) Vertreter:

**Uexküll & Stolberg, 22607 Hamburg**

(54) Bezeichnung: **METHODEN ZUR IDENTIFIZIERUNG EVOLUTIONÄR SIGNIFIKANTER ÄNDERUNGEN IN POLY-  
NUKLEOTID- UND POLYPEPTIDSEQUENZEN IN DOMESTIZIERTEN PFLANZEN UND TIEREN**

Anmerkung: Innerhalb von neun Monaten nach der Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents kann jedermann beim Europäischen Patentamt gegen das erteilte europäische Patent Einspruch einlegen. Der Einspruch ist schriftlich einzureichen und zu begründen. Er gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist (Art. 99 (1) Europäisches Patentübereinkommen).

Die Übersetzung ist gemäß Artikel II § 3 Abs. 1 IntPatÜG 1991 vom Patentinhaber eingereicht worden. Sie wurde vom Deutschen Patent- und Markenamt inhaltlich nicht geprüft.

**Beschreibung**

Technisches Gebiet

**[0001]** Diese Erfindung betrifft die Verwendung von molekularen und Evolutionstechniken zum Identifizieren von Polynukleotid- und Polypeptidsequenzen, die kommerziell oder ästhetisch relevanten Eigenschaften bei domestizierten Pflanzen und Tieren entsprechen.

Technischer Hintergrund

**[0002]** Seit Tausenden von Jahren haben Menschen Pflanzen und Tiere gezüchtet, wobei auf bestimmte kommerziell wertvolle und/oder ästhetische Eigenschaften selektiert wurde. Domestizierte Pflanzen unterscheiden sich von ihren wilden Vorfahren in solchen Eigenschaften wie Ertrag, Kurztagblühen (short day length flowering), Protein- und/oder Ölgehalt, Leichtigkeit der Ernte, Geschmack, Krankheitsresistenz und Dürre-resistenz. Domestizierte Tiere unterscheiden sich von ihren wilden Vorfahren in solchen Eigenschaften wie Fett- oder Proteingehalt, Milchproduktion, Fügbarkeit, Fruchtbarkeit oder Zeit bis zur Reife. Momentan sind weder die meisten Gene, die den obigen Unterschieden zu Grunde liegen, noch, was ebenso wichtig ist, die spezifischen Veränderungen bekannt, die sich in diesen Genen entwickelt haben, um diese Fähigkeiten zu liefern. Wenn man die Grundlage dieser Unterschiede zwischen domestizierten Pflanzen und Tieren und ihren wilden Vorfahren versteht, liefert dies nützliche Informationen zum Aufrechterhalten und Verbessern dieser Eigenschaften. Im Falle der Feldfruchtpflanzen ermöglicht die Identifizierung der spezifischen Gene, die bestimmte Eigenschaften kontrollieren, direkte und rasche Verbesserung in einer zuvor unmöglichen Weise.

**[0003]** Obwohl der Vergleich homologer Gene oder Proteine zwischen domestizierten Spezies und ihren wilden Vorfahren brauchbare Informationen in Bezug auf konservierte molekulare Sequenzen und funktionale Merkmale liefert, ist dieser Ansatz bei der Identifizierung von Genen, deren Sequenzen sich während der von Menschen auferlegten Selektionsdrücke geändert haben, von begrenztem Nutzen. Mit dem Aufkommen hochentwickelter Algorithmen und Analyseverfahren können DNA-Sequenzveränderungen viel mehr Informationen entlockt werden, welche Gene positiv selektiert worden sind. Das Mächtigste dieser Verfahren, " $K_A/K_S$ ", beinhaltet paarweise Vergleiche zwischen Alignment unterzogenen Protein-kodierenden Nukleotidsequenzen mit den Verhältnissen von:

nicht-synonymen Nukleotidsubstitutionen pro nicht-synonymer Stelle ( $K_A$ )  
synonymen Substitutionen pro synonymer Stelle ( $K_S$ )

(wobei nicht-synonym Substitutionen bedeutet, die die kodierte Aminosäure ändern, und synonym Substitutionen bedeutet, die die kodierte Aminosäure nicht ändern). "Verfahren vom  $K_A/K_S$ -Typ" schließen diese und ähnliche Verfahren ein.

**[0004]** Diese Verfahren sind bereits verwendet werden, um das Vorkommen von Darwin'scher (d. h. natürlicher) positiver Selektion auf molekularer Ebene zu demonstrieren, was zu Aminosäuredifferenzen in homologen Proteinen führt. Mehrere Gruppen haben derartige Verfahren verwendet, um zu dokumentieren, dass sich ein spezielles Protein rascher entwickelt hat als die neutrale Substitutionsrate, und dies stützt die Existenz von Darwin'scher positiver Selektion auf molekularer Ebene. McDonald und Kreitman (1991) Nature 351:652-654, schlagen beispielsweise einen statistischen Test der neutralen Protein-Evolutionshypothese basierend auf dem Vergleich der Zahl der Aminosäureersatzsubstitutionen zu synonymen Substitutionen im Kodierbereich eines Locus vor. Wenn sie diesen Test auf den Adh-Locus dreier Drosophila-Spezies anwenden, folgern sie, dass er stattdessen zeigt, dass der Locus adaptive Fixierung selektiv vorteilhafter Mutationen eingegangen ist und dass selektive Fixierung adaptiver Mutationen eine lebensfähige Alternative zu der schematischen Akkumulation neutraler Mutationen sein kann, um meisten Proteinevolutionen zu erklären. Jenkins et al. (1995) Proc. R. Soc. Lond. B 261:203-207 verwenden den Test von McDonald & Kreitman, um zu untersuchen, ob adaptive Evolution in Sequenzen vorkommt, die Transkription kontrollieren (nicht-kodierende Sequenzen).

**[0005]** Nakashima et al. (1995) Proc. Natl. Acad. Sci USA 92:5606-5609, verwenden das Verfahren von Miya-ta und Yasunaga zur Durchführung paarweiser Vergleiche der Nukleotidsequenzen von zehn PLA2 Isozymgenen von zwei Schlangenspezies; bei diesem Verfahren wird die Zahl der Nukleotidsubstitutionen pro Stelle für die nicht-kodierenden Regionen einschließlich Introns ( $K_N$ ) und das  $K_A$  und  $K_S$  verglichen. Sie folgern, dass sich die Protein kodierenden Regionen mit viel höheren Raten entwickelt haben als die nicht-kodierenden Regionen einschließlich Introns. Die stark beschleunigte Substitutionsrate ist für die Darwin'sche Evolution der PLA2-Isozymgene auf molekularer Ebene verantwortlich, um neue physiologische Aktivitäten zu produzieren, die für ei-

nen starken Selektionsvorteil zum Beutefang oder zur Verteidigung gegen Raubtiere gesorgt haben müssen. Endo et al. (1996) *Mol. Biol. Evol.* 13(5):685-690 verwenden das Verfahren von Nei und Gojobori, wobei  $d_N$  die Zahl der nicht-synonymen Substitutionen ist und  $d_S$  die Zahl der synonymen Substitutionen ist, um in Frage kommende Gene zu identifizieren, mit denen die positive natürliche Selektion arbeitet. Metz und Palumbi (1996) *Mol. Biol. Evol.* 13(2):397-406 verwenden jenen Test von McDonald & Kreitman (supra) sowie ein Verfahren, das Nei und Gojobori, Nei und Jin, sowie Kumar, Tamura und Nei zugeschrieben wird, wobei die durchschnittlichen Proportionen von  $P_n$ , die Ersetzungssubstitutionen pro Ersetzungsstelle, und  $P_s$ , die stillen Substitutionen pro stiller Stelle, untersucht werden, um nach Anzeichen für positive Selektion an Bindungsgenen bei Seeigeln zu suchen, um zu untersuchen, ob sie sich als Vorspiel zu der Speziesbildung rasch entwickelt haben. Goodwin et al. (1996) *Mol. Biol. Evol.* 13(2):346-358 verwendet ähnliche Verfahren, um die Evolution einer speziellen murinen Genfamilie zu entwickeln, und sie folgern, dass die Verfahren wichtige fundamentale Einsichten liefern, wie Selektion genetische Divergenz in einem experimentell manipulierbaren System antreibt. Edwards et al. (1995) verwenden degenerierte Primers, um MHC-Loci aus verschiedenen Vogelspezies und einer Alligatorspezies herauszuziehen, welche dann nach den Verfahren von Nei und Gojobori ( $d_N:d_S$ -Verhältnisse) analysiert werden, um MHC-Studien auf Wirbeltiere zu erweitern, die keine Säugetiere sind. Whitfield et al. (1993) *Nature* 364:713-715 verwenden  $K_A/K_S$ -Analyse, um nach direktonaler Selektion in den Regionen zu suchen, die eine konservierte Region in dem SRY-Gen flankieren (welches das männliche Geschlecht bestimmt). Sie schlagen vor, dass die rasche Evolution von SRY eine signifikante Ursache für reproduktive Isolation sein könnte, welche zu neuen Spezies führt. Wettsetin et al. (1996) *Mol. Biol. Evol.* 13(1):56-66 wenden das MEGA-Programm von Kumar, Tamura und Nei und phylogenetische Analyse zur Erforschung der Diversifikation von MHC Klasse I Genes bei Eichhörnchen und verwandten Nagern an. Parham und Ohta (1996) *Science* 212:67-74 konstatieren, dass ein populationsbiologischer Ansatz, der Tests auf Selektion sowie auf Genumwandlung und neutralen Drift einschließt, zum Analysieren der Generierung und des Erhalts von menschlichem MHC Klasse I Polymorphismus erforderlich ist. Hughes (1997) *Mol. Biol. Evol.* 14(1): 1-5 vergleichen über hundert orthologe Immunoglobulin-C2-Domänen zwischen Mensch und Nager, wobei das Verfahren von Nei und Gojobori ( $d_N:d_S$ -Verhältnisse) zum Testen der Hypothese verwendet wird, dass Proteine, die in Zellen des Immunsystems des Wirbeltiers exprimiert werden, sich ungewöhnlich rasch entwickeln. Swanson und Vacquier (1998) *Science* 281:710-712 verwenden  $d_N:d_S$ -Verhältnisse, um die konzertierte Evolution zwischen dem Lysin und dem Ei-Rezeptor für Lysin zu zeigen, und diskutieren die Rolle einer derartigen konzertierten Evolution bei der Bildung neuer Spezies (Spezifikation). Messier und Stewart (1997) *Nature* 385:151-154, verwendeten  $K_A/K_S$  zur Demonstration der positiven Selektion bei Primatenlysozymen.

**[0006]** Die mit der Domestizierung zusammenhängenden genetischen Veränderungen sind am umfassendsten bei Mais (Korn) untersucht worden (Dorweiler (1993) *Science* 262:232-235). Bei Mais (*Zea spp. mays mays*) steht offensichtlich eine kleinere Anzahl von Einzelgenveränderungen für alle der Unterschiede zwischen unserer gegenwärtigen domestizierten Maispflanze und ihrem wilden Vorfahren, Teosinte (*Zea mays ssp. parviglumis*) (Dorweiler, 1993). QTL-(quantitative trait locus)-Analyse hat gezeigt (Doebley (1990) *PNAS USA* 87:9888-9892), dass nicht mehr als fünfzehn Gene interessierende Eigenschaften (traits) bei Mais kontrollieren, und erklärt den deutlichen morphologischen Unterschied zwischen Mais und Teosinte (Wang (1999) *Nature* 398:236-239).

**[0007]** Wichtig ist, dass eine ähnlich kleine Anzahl an Genen interessierende Eigenschaften in anderen von Gras abgeleiteten Feldfruchtplanzen kontrollieren kann, einschließlich Reis, Weizen, Hirse und Sorghum (Patterson (1995) *Science* 269:1714-1718). Bei den meisten dieser relevanten Gene in Mais kann das homologe Gen in der Tat ähnliche Eigenschaften in anderen von Gras abgeleiteten Feldfruchtplanzen kontrollieren (Patterson, 1995). Die Identifizierung dieser Gene in Mais würde somit die Identifizierung homologer Gene in Reis, Weizen, Hirse und Sorghum erleichtern.

**[0008]** Wie aus den oben zitierten Druckschriften ersichtlich ist, können Analysemethoden der molekularen Evolution zur Identifizierung sich rasch entwickelnder Gene ( $K_A/K_S$ -Typ-Methoden) verwendet werden, um vielen unterschiedlichen Zwecken zu dienen, allgemein zur Bestätigung der Existenz von Darwin'scher positiver Selektion auf molekularer Ebene, aber auch zur Bewertung der Frequenz der Darwin'schen positiven Selektion auf molekularer Ebene, zum Verständnis phylogenetischer Beziehungen, zur Bewertung von Mechanismen, nach denen neue Spezies gebildet werden, oder zum Nachweis von einzelner oder mehrfacher Herkunft bei spezifischen Genpolymorphismen. Aus den oben zitierten Druckschriften und anderen Literaturstellen geht hervor, dass keiner der Autoren  $K_A/K_S$ -Typ-Methoden zum Identifizieren von evolutionären Veränderungen bei domestizierten Pflanzen und Tieren verwendet hat, die durch künstliche Selektionsdrücke entstanden sind. Während Turcich et al. (1996) *Sexual Plant Reproduction* 9:65-74, die Verwendung der  $K_S$ -Analyse für Pflanzengene beschreiben, wird angenommen, dass niemand bislang Analyse vom  $K_A/K_S$ -Typ als systematisches Werkzeug zum Identifizieren jener Gene in domestizierten Pflanzen und Tieren verwendet hat, die evolutionär

signifikante Sequenzveränderungen enthalten, die zur Entwicklung, zum Erhalt oder zur Verbesserung erwünschter kommerzieller oder ästhetischer Merkmale verwendet werden können.

**[0009]** Die Identifizierung in domestizierten Spezies von Genen, die sich entwickelt haben, um verglichen mit homologen Vorfahrengenen einzigartige, verbesserte oder veränderte Funktionen zu verleihen, könnte zur Entwicklungen von Mitteln zur Modulierung dieser Funktionen verwendet werden. Die Identifizierung der zu Grunde liegenden Gene der domestizierten Spezies und der spezifischen Nukleotidveränderungen, die aufgetreten sind, und die weitere Charakterisierung der physikalischen und biochemischen Veränderungen der Proteine, die durch diese sich entwickelnden Gene kodiert sind, könnte wertvolle Informationen über den Mechanismus liefern, welcher der gewünschten Eigenschaft zu Grunde liegt. Diese wertvolle Information könnte für die Entwicklung von Mitteln verwendet werden, die die Funktion der Zielproteine weiter verbessern. Alternativ könnte weiteres Engineering der verantwortlichen Gene die gewünschte Eigenschaft modifizieren oder steigern. Es könnte zudem gefunden werden, dass die identifizierten Gene eine Rolle bei der Kontrolle interessierender Eigenschaften in anderen domestizierten Pflanzen spielen können. Ein ähnlicher Prozess kann Gene für interessierende Merkmale bei domestizierten Tieren identifizieren.

**[0010]** Doebley beschreibt in Trends in Genetics, Band 8, Nr. 9, September 1992, Seiten 302-307, die Anwendung von Genanalyse unter Verwendung molekularer Marker auf die Kartierung von Genen, die morphologische Eigenschaften bei Mais und seinem wilden Vorfahren Teosinte kontrollieren. Durch die Entwicklung gesättigter Verknüpfungskarten (Linkage-Karten) unter Verwendung von Restriktionsfragmentlängenpolymorphismen (RFLPs) und neue statistische Verfahren zur Kartierung und Charakterisierung von quantitative trait loci (QTLs) wurden fünf Regionen des Genoms identifiziert, von denen gesagt wird, dass sie die meisten der Unterschiede zwischen Mais und Teosinte kontrollieren.

**[0011]** EP 0 120 658 A2 offenbart ein Verfahren zum Charakterisieren einer unbekanntem Organismenspezies, bei dem man die Position evolutionär konservierter Sequenzen in genetischem Material des Organismus relativ zu einer bekannten Position von Restriktionsendonuklease-Spaltungsstellen in dem genetischen Material bestimmt, um eine identifizierende genetische Charakterisierung des unbekanntem Organismus zu erhalten, und man die Charakterisierung mit Informationen von mindestens zwei Sets identifizierender genetischer Charakterisierungen vergleicht, die von den gleichen konservierten Sequenzen stammen, wobei jeder der Sets eine bekannte Organismusspezies definiert.

#### OFFENBARUNG DER ERFINDUNG

**[0012]** Die vorliegende Erfindung liefert Verfahren zum Identifizieren von Polynukleotid- und Polypeptidsequenzen mit evolutionär signifikanten Veränderungen, die mit kommerziellen oder ästhetischen Eigenschaften bei domestizierten Organismen einschließlich Pflanzen und Tieren zusammenhängen. Die Erfindung verwendet vergleichende Genome, um spezifische Genveränderungen zu identifizieren, die mit strukturellen, biochemischen oder physiologischen Zuständen zusammenhängen und somit für diese verantwortlich sein können, wie kommerziell oder ästhetisch relevanten Eigenschaften, und verwendet die aus diesen Eigenschaften erhaltenen Informationen, um domestizierte Organismen mit verbesserten interessierenden Eigenschaften zu entwickeln.

**[0013]** In einer bevorzugten Ausführungsform hat ein Polynukleotid oder Polypeptid einer domestizierten Pflanze oder eines domestizierten Tieres künstliche Selektion erfahren, die zu einer evolutionär signifikanten Änderung führte, die in der domestizierten Spezies vorhanden und in dem wilden Vorfahren nicht vorhanden ist. Ein Beispiel für diese Ausführungsform ist, dass das Polynukleotid oder Polypeptid mit verbessertem Ertrag der Feldfrucht zusammenhängen kann, verglichen mit dem Vorfahren. Zu anderen Beispielen gehören Kurztagblüte (d. h. dass die Pflanze nur dann blüht, wenn die tägliche Lichtperiode unter irgendeiner kritischen Länge liegt), Proteingehalt, Ölgehalt, Leichtigkeit der Ernte, Geschmack, Dürresistenz und Krankheitsresistenz. Die vorliegende Erfindung kann somit brauchbar sein, um Einsicht in die molekularen Mechanismen zu erhalten, die Funktionen oder Eigenschaften domestizierter Organismen zu Grunde liegen. Diese Informationen können zum Design des Polynukleotids brauchbar sein, um so die Funktion oder die Eigenschaft weiter zu verbessern. Ein Polynukleotid, von dem bestimmt wurde, dass es für verbesserten Feldfruchtertrag verantwortlich ist, könnte beispielsweise statistischer oder gerichteter Mutagenese unterzogen werden, gefolgt von Testen der mutierten Gene, um jene zu identifizieren, welche die Eigenschaft weiter verbessern.

**[0014]** Daher werden gemäß einem Aspekt Verfahren zum Identifizieren einer Polynukleotidsequenz bereitgestellt, die ein Polypeptid eines domestizierten Organismus (z. B. einer Pflanze oder eines Tieres) kodiert, wobei das Polypeptid mit einer kommerziell oder ästhetisch relevanten Eigenschaft zusammenhängen kann,

die in dem domestizierten Organismus, verglichen mit einem wilden Vorfahren des domestizierten Organismus, einzigartig, verbessert oder verändert ist, umfassend die Stufen, in denen: a) man Protein kodierende Nukleotidsequenzen des domestizierten Organismus mit Protein kodierenden Nukleotidsequenzen des wilden Vorfahren vergleicht, und b) man eine Polynukleotidsequenz in dem domestizierten Organismus selektiert, die eine Nukleotidveränderung enthält, die verglichen mit einer entsprechenden Sequenz in dem wilden Vorfahren evolutionär signifikant ist.

**[0015]** Gemäß einem weiteren Aspekt der Erfindung werden Verfahren zum Identifizieren einer evolutionär signifikanten Veränderung in einer Protein kodierenden Nukleotidsequenz eines domestizierten Organismus (z. B. einer Pflanze oder einem Tier) bereitgestellt, welche die Schritte aufweisen, in denen a) man Protein kodierende Nukleotidsequenzen des domestizierten Organismus mit entsprechenden Sequenzen eines wilden Vorfahren des domestizierten Organismus vergleicht, und b) man eine Polynukleotidsequenz in dem domestizierten Organismus selektiert, die verglichen mit einer entsprechenden Sequenz in dem wilden Vorfahren eine Nukleotidveränderung enthält, wobei die Veränderung evolutionär signifikant ist.

**[0016]** Die nach irgendeiner der hier beschriebenen Verfahren identifizierte Nukleotidveränderung ist in einigen Ausführungsformen eine nicht-synonyme Substitution. In einigen Ausführungsformen wird die evolutionäre Signifikanz der Nukleotidveränderung gemäß der nicht-synonymen Substitutionsrate ( $K_A$ ) der Nukleotidsequenz bestimmt. In einigen Ausführungsformen werden die evolutionär signifikanten Veränderungen bewertet, indem das  $K_A/K_S$ -Verhältnis zwischen dem Polynukleotid des domestizierten Organismus und dem entsprechenden Polynukleotid des Vorfahren bestimmt wird. Das Verhältnis ist vorzugsweise mindestens etwa 0,75, oder das Verhältnis ist mit zunehmender Bevorzugung mindestens etwa 1,25, 1,50 und 2,00.

**[0017]** Die Erfindung liefert gemäß einem weiteren Aspekt ein Verfahren zum Identifizieren eines Mittels, welches die relevante Eigenschaft in dem domestizierten Organismus modulieren kann, bei dem man mindestens ein in Frage kommendes Mittel mit einer Zelle, einem Modellsystem oder einer transgenen Pflanze oder einem nicht menschlichen transgenen Organismus kontaktiert, die bzw. der die Polynukleotidsequenz mit der evolutionär signifikanten Änderung exprimiert, wobei das Mittel durch seine Fähigkeit identifiziert wird, Funktionen des Polypeptids zu modulieren.

**[0018]** Es wird auch ein Verfahren zum Sequenzvergleich im Großmaßstab zwischen Protein kodierenden Nukleotidsequenzen eines domestizierten Organismus und Protein kodierenden Sequenzen von einem wilden Vorfahren bereitgestellt, bei dem a) man die Sequenzen des domestizierten Organismus mit entsprechenden Sequenzen des wilden Vorfahren gemäß Sequenzhomologie Alignment unterzieht und b) man jegliche Nukleotidveränderungen innerhalb der Sequenzen des domestizierten Organismus identifiziert, verglichen mit den homologen Sequenzen aus dem wilden Vorfahrenprimaten.

**[0019]** Die vorliegende Erfindung liefert gemäß einem anderen Aspekt ein Verfahren zum Korrelieren einer evolutionär signifikanten Nukleotidveränderung mit einer kommerziell oder ästhetisch relevanten Eigenschaft, die in einem domestizierten Organismus einzigartig, verbessert oder verändert ist, bei dem a) man eine Nukleotidsequenz mit einer evolutionär signifikanten Veränderung gemäß den hier beschriebenen Verfahren identifiziert und b) man die funktionale Wirkung der Anwesenheit oder Abwesenheit der identifizierten Sequenz in dem domestizierten Organismus oder in einem Modellsystem analysiert.

**[0020]** Die in den vorliegenden Verfahren verwendeten domestizierten Pflanzen können Korn (Mais), Reis, Tomaten, Kartoffeln oder jede domestizierte Pflanze sein, deren wilder Vorfahre existent und bekannt ist. Der Vorfahre von Korn (Mais) ist beispielsweise Teosinte; Vorfahren von Weizen sind *Triticum monococcum*, *T. speltoides* und *Aegilops tauschii*; und Vorfahren von Reis sind *Oryza nivora* und *O. rufipogon*. Die relevante Eigenschaft kann jede kommerziell oder ästhetisch relevante Eigenschaft sein, wie Ertrag, Kurztagblüte, Proteingehalt, Ölgehalt, Dürresistenz, Geschmack, Leichtigkeit der Ernte oder Krankheitsresistenz.

**[0021]** Die in den vorliegenden Verfahren verwendeten domestizierten Tiere können jedes domestizierte Tier sein, für das ein Vorfahre zur Verfügung steht, einschließlich Schweinen, Rindern (Vieh), Pferden, Hunden oder Katzen. Ein Vorfahre des Pferdes ist beispielsweise das Pryzewalskii-Pferd, und Vorfahren von Rindern (Vieh) schließen einige indische Rassen ein. Die relevante Eigenschaft könnte beispielsweise Fettgehalt, Proteingehalt, Milchproduktion, Zeit bis zur Reife, Fruchtbarkeit, Fügsamkeit oder Krankheitsresistenz und Krankheitsanfälligkeit sein.

**[0022]** Die vorliegende Erfindung verwendet vergleichende Genome, um spezifische Genveränderungen zu identifizieren, die mit kommerziell oder ästhetisch relevanten Eigenschaften bei domestizierten Organismen (z. B. Pflanzen und Tieren) zusammenhängen und somit zu ihnen beitragen oder für sie verantwortlich sein können.

**[0023]** In einer bevorzugten Ausführungsform können die hier beschriebenen Verfahren zur Identifizierung der Gene verwendet werden, die interessierende Eigenschaften in landwirtschaftlich bedeutsamen domestizierten Pflanzen kontrollieren. Menschen haben seit Tausenden von Jahren domestizierte Pflanzen gezüchtet, ohne die Gene zu kennen, die diese Eigenschaften kontrollieren. Die Kenntnis des spezifischen beteiligten genetischen Mechanismus würde eine viel raschere und direkte Intervention auf molekularer Ebene ermöglichen, um Pflanzen mit erwünschten oder verbesserten Eigenschaften zu erzeugen.

**[0024]** Menschen haben durch künstliche Selektion enorme Selektionsdrücke auf Feldfruchtpflanzen ausgeübt. Dieser Druck spiegelt sich in evolutionär signifikanten Veränderungen zwischen homologen Genen domestizierter Organismen und ihrer wilden Vorfahren wider. Es ist gefunden worden, dass nur einige wenige Gene, z. B. 10 bis 15 pro Spezies, Eigenschaften von kommerziellem Interesse bei domestizierten Feldfruchtpflanzen kontrollieren. Diese wenigen Gene sind nach Standardverfahren der Molekularbiologie bei Pflanzen außerordentlich schwierig zu identifizieren. Die hier beschriebenen  $K_A/K_S$ - und verwandten Analysen können die Gene identifizieren, die interessierende Eigenschaften kontrollieren, wenn diese Gene in der Protein kodierenden Region Veränderungen eingegangen sind.

**[0025]** Es können für jede interessierende Feldfruchtpflanze cDNA-Bibliotheken von der domestizierten Spezies oder Subspezies und ihrem wilden Vorfahren aufgebaut werden. Wie in der USSN 09/240,915, eingereicht am 29. Januar 1999, beschrieben ist, werden die cDNA-Bibliotheken jeweils gegeneinander "geBLASTet", um homologe Polynukleotide zu identifizieren. Der versierte Fachmann kann alternativ auf kommerziell und/oder öffentlich verfügbare Genom- oder cDNA-Datenbanken zugreifen, wie die Folgenden:

[www.central.edu/homepages/liedlb/genetics/gene-site.html](http://www.central.edu/homepages/liedlb/genetics/gene-site.html)  
[www.ornl.gov/Techresources/Human-Genome/genetics.html](http://www.ornl.gov/Techresources/Human-Genome/genetics.html) and  
[www.mcb.harvard.edu/Biolinks/Sequences.html](http://www.mcb.harvard.edu/Biolinks/Sequences.html)  
[www.ncbi.nlm.gov/Web/Genbank/index.html](http://www.ncbi.nlm.gov/Web/Genbank/index.html)

anstatt cDNA-Bibliotheken aufzubauen. Als nächstes wird eine  $K_A/K_S$ - oder verwandte Analyse durchgeführt, um ausgewählte Gene zu identifizieren, die sich unter Selektionsdruck rasch entwickelt haben. Diese Gene werden dann unter Verwendung von Standardmolekularverfahren für transgene Pflanzen bewertet, um zu bestimmen, ob sie eine Rolle in den kommerziell oder ästhetisch interessierenden Eigenschaften spielen. Die interessierenden Gene werden dann z. B. durch statistische oder stellengerichtete Mutagenese manipuliert, um neue verbesserte Varietäten, Subspezies, Stämme oder Kulturvarianten zu entwickeln.

**[0026]** Die hier beschriebenen Methoden können in ähnlicher Weise auf domestizierte Tiere angewendet werden, einschließlich Schweinen, Rindern, Pferden, Hunden, Katzen und anderen domestizierten Tieren, für die ein wilder Vorfahre zur Verfügung steht. Rinder und Pferde sind von besonders wichtigem kommerziellem Interesse. Wie bei den Pflanzen haben Menschen seit Tausenden von Jahren Tiere gezüchtet, und diese intensiven Selektionsdrücke spiegeln sich in erhöhten  $K_A/K_S$ -Raten für sich rasch entwickelnde interessierende Gene wider. Um homologe Polynukleotide zu identifizieren, können wiederum aufgebaute cDNA-Bibliotheken domestizierter Tiere und ihrer wilden Vorfahren gegeneinander geBLASTet werden, und/oder es kann auf Öffentliche oder private Genom- oder cDNA-Datenbanken zugegriffen werden. Für homologe Sequenzen können  $K_A/K_S$ - oder verwandte Analysen durchgeführt werden, welche die Polynukleotide identifizieren, die sich unter dem künstlichen Selektionsdruck rasch entwickelt haben. Diese Gene werden dann unter Verwendung von Standardmolekularverfahren für nicht-menschliche transgene Tiere bewertet, um zu bestimmen, ob sie eine Rolle in den kommerziell oder ästhetisch interessierenden Eigenschaften spielen. Diese Gene können dann manipuliert werden, um neue verbesserte Tiervarietäten oder Subspezies zu entwickeln.

**[0027]** Die Praxis der vorliegenden Erfindung verwendet, wenn nicht anders angegeben, konventionelle Techniken der Molekularbiologie, Genetik und Molekularevolution, die innerhalb des Wissens des Standes der Technik liegen. Derartige Techniken sind vollständig in der Literatur erläutert, wie: "Molecular Cloning: A Laboratory Manual", 2. Auflage (Sambrook et al., 1989); "Oligonucleotide Synthesis" (M. J. Gait, Herausgeber, 1984); "Current Protocols in Molecular Biology" (F. M. Ausubel et al., Herausgeber, 1987); "PCR: The Polymerase Chain Reaction", (Mullis et al., Herausgeber, 1994); "Molecular Evolution", (Li, 1997).

## Definitionen

**[0028]** Ein "Polynukleotid" bezieht sich hier auf eine polymere Form von Nukleotiden von beliebiger Länge, entweder Ribonukleotide oder Deoxyribonukleotide oder Analoga davon. Dieser Begriff bezieht sich auf die Primärstruktur des Moleküls und schließt somit doppel- und einsträngige DNA sowie doppel- und einsträngige RNA ein. Eingeschlossen sind auch modifizierte Polynukleotide, wie methylierte und/oder verkappte Polynukleotide. Die Begriffe "Polynukleotid" und "Nukleotidsequenz" werden austauschbar verwendet.

**[0029]** Ein "Gen" bezieht sich hier auf ein Polynukleotid oder einen Abschnitt eines Polynukleotids, das bzw. der eine Sequenz aufweist, die ein Protein kodiert. Es ist in der Technik wohl bekannt, dass ein Gen auch nicht-kodierende Sequenzen aufweist, wie 5'- und 3'-flankierende Sequenzen (wie Promotoren, Enhancer, Repressoren und andere Regulationssequenzen) sowie Introns.

**[0030]** Die Begriffe "Polypeptid", "Peptid" und "Protein" werden hier austauschbar zur Bezeichnung von Polymeren von Aminosäuren mit beliebiger Länge verwendet. Diese Begriffe schließen auch Proteine ein, die nach der Translation durch Reaktionen modifiziert worden sind, zu denen Glykolysierung, Acetylierung und Phosphorylierung gehören.

**[0031]** Der Begriff "domestizierter Organismus" bezieht sich auf einen individuellen lebenden Organismus oder Population desselben, eine Spezies, Subspezies, Variante, Kultur oder einen Stamm, die bzw. der künstlichem Selektionsdruck ausgesetzt gewesen ist und eine kommerziell oder ästhetisch relevante Eigenschaft entwickelt hat. In einigen bevorzugten Ausführungsformen ist der domestizierte Organismus eine Pflanze ausgewählt aus der Gruppe bestehend aus Mais, Weizen, Reis, Sorghum, Tomate oder Kartoffel oder eine beliebige andere domestizierte Pflanze von kommerziellem Interesse, deren Vorfahr bekannt ist. In anderen bevorzugten Ausführungsformen ist der domestizierte Organismus ein Tier ausgewählt aus der Gruppe bestehend aus Rind (Vieh), Pferden, Schweinen, Katzen und Hunden. Ein domestizierter Organismus und sein Vorfahre können als unterschiedliche Spezies, Subspezies, Varietäten, Kulturen oder Stämme oder beliebige Kombination davon verwandt sein.

**[0032]** Der Begriff "wilder Vorfahre" oder "Vorfahre" bedeutet eine(n) Vorläufer- oder Vorgängerorganismus, -spezies, -subspezies, -varietät, -kultur oder -stamm, aus dem bzw. der sich ein(e) domestizierte(r) Organismus, Spezies, Subspezies, Varietät, Kultur oder Stamm entwickelt hat. Ein domestizierter Organismus kann einen oder mehr als einen Vorfahren haben. Domestizierte Pflanzen können in der Regel einen oder mehrere Vorfahren haben, während domestizierte Tiere üblicherweise nur einen einzigen Vorfahren haben.

**[0033]** Der Begriff "kommerzielle oder ästhetisch relevante Eigenschaft" wird hier zur Bezeichnung von Eigenschaften verwendet, die in domestizierten Organismen, wie Pflanzen oder Tieren, vorliegen, deren Analyse Informationen (z. B. physikalische oder biochemische Daten) liefern kann, die für die Entwicklung von Mitteln relevant sind, die das für die Eigenschaft verantwortliche Polypeptid modulieren können. Die kommerziell oder ästhetisch relevante Eigenschaft kann in Bezug zu dem Vorfahren einzigartig, verbessert oder verändert sein. Mit "verändert" ist gemeint, dass sich die relevante Eigenschaft qualitativ oder quantitativ von Eigenschaften unterscheidet, die bei dem Vorfahren beobachtet wurden.

**[0034]** Der Begriff "Verfahren vom  $K_A/K_S$ -Typ" bedeutet Verfahren, die Unterschiede bewerten, die sich oft (aber nicht immer) als Verhältnis zwischen der Anzahl der nicht-synonymen Substitutionen und synonymen Substitutionen in homologen Genen zeigen (einschließlich der rigoroseren Verfahren, die nicht-synonyme und synonyme Stellen bestimmen). Diese Verfahren werden nach mehreren Nomenklatursystemen bezeichnet, einschließlich, aber nicht begrenzt auf  $K_A/K_S$ ,  $d_N/d_S$ ,  $D_N/D_S$ .

**[0035]** Die Begriffe "evolutionär signifikante Veränderung" und "adaptive evolutionäre Veränderung" beziehen sich auf eine oder mehrere Nukleotid- oder Peptidsequenzveränderung(en) zwischen zwei Organismen, Spezies, Subspezies, Varietäten, Kulturen und/oder Stämmen, die auf positiven Selektionsdruck zurückzuführen sein können. Ein Verfahren zur Bestimmung der Anwesenheit einer evolutionär signifikanten Veränderung ist die Anwendung eines Analyseverfahrens vom  $K_A/K_S$ -Typ, wie die Messung eines  $K_A/K_S$ -Verhältnisses. Ein  $K_A/K_S$ -Verhältnis von mindestens etwa 0,75, insbesondere mindestens etwa 1,0, insbesondere mindestens etwa 1,25, insbesondere mindestens etwa 1,5 und am meisten bevorzugt mindestens etwa 2,0 zeigt die Einwirkung einer positiven Selektion und wird als evolutionär signifikante Veränderung angesehen.

**[0036]** Der Begriff "positive evolutionär signifikante Veränderung" bedeutet eine evolutionär signifikante Veränderung in einem/einer speziellen Organismus, Spezies, Subspezies, Varietät, Kultur oder Stamm, die zu ei-

ner adaptiven Veränderung führt, die verglichen mit anderen verwandten Organismen positiv ist. Ein Beispiel für eine positive evolutionär signifikante Veränderung ist eine Veränderung, die zu einem erhöhten Ertrag bei Feldfruchtpflanzen führt.

**[0037]** Der Begriff "resistent" bedeutet, dass ein Organismus eine Fähigkeit hat, einen Erkrankungszustand und/oder eine Entwicklung der Erkrankung zu vermeiden oder den Erkrankungsgrad zu vermindern, vorzugsweise im Vergleich mit nicht-resistenten Organismen.

**[0038]** Der Begriff "Empfindlichkeit" bedeutet, dass ein Organismus nicht die Fähigkeit hat, einen Erkrankungszustand und/oder eine Entwicklung der Erkrankung zu vermeiden oder den Erkrankungsgrad zu vermindern, vorzugsweise im Vergleich mit Organismen, die bekanntermaßen resistent sind.

**[0039]** Es ist bekannt, dass Resistenz und Empfindlichkeit von Individuum zu Individuum variieren, und dass diese Begriffe für die erfindungsgemäßen Zwecke auch für eine Gruppe von Individuen innerhalb einer Spezies gelten, und dass Vergleiche von Resistenz und Empfindlichkeit sich im Allgemeinen auf gesamte durchschnittliche Unterschiede zwischen Spezies beziehen, obwohl intraspezifische Vergleiche verwendet werden können.

**[0040]** Der Begriff "homolog" oder "Homologes" oder "ortholog" ist bekannt und in der Technik allgemein üblich und bezieht sich auf verwandte Sequenzen, die einen gemeinsamen Vorfahren haben, und wird basierend auf dem Sequenzidentitätsgrad bestimmt. Diese Begriffe beschreiben die Beziehung zwischen einem Gen, das sich in einer/einem Spezies, Subspezies, Varietät, Kultur oder Stamm findet, und dem entsprechenden oder äquivalenten Gen in einer/einem anderen Spezies, Subspezies, Varietät, Kultur oder Stamm. Für die Zwecke dieser Erfindung werden homologe Sequenzen verglichen. Es wird angenommen, vermutet oder ist bekannt, dass "homologe Sequenzen" oder "homologe" oder "orthologe" funktionell verwandt sind. Eine funktionelle Verwandtschaft kann nach irgendeiner von zahlreichen Weisen gezeigt werden, einschließlich, aber nicht beschränkt auf (a) den Grad der Sequenzidentität und (b) die gleiche oder ähnliche biologische Funktion. Vorzugsweise werden sowohl (a) als auch (b) gezeigt. Der Sequenzidentitätsgrad kann variieren, ist vorzugsweise jedoch mindestens 50% (wenn im Stand der Technik bekannte Standard-Sequenz-Alignment-Programme verwendet werden), insbesondere mindestens 60% insbesondere mindestens etwa 75% insbesondere mindestens etwa 85%. Homologie kann unter Verwendung von Softwareprogrammen bestimmt werden, die in der Technik leicht erhältlich sind, wie jene, die in Current Protocols in Molecular Biology (F. M. Ausubel et al., Herausgeber, 1987) Ergänzung 30, Abschnitt 7.718, Tabelle 7.71, erörtert sind. Bevorzugte Alignment-Programme sind MacVector (Oxford Molecular Ltd, Oxford, Großbritannien) und ALIGN Plus (Scientific and Educational Software, Pennsylvania, USA). Ein weiteres bevorzugtes Alignment-Programm ist Sequencher (Gene Codes, Ann Arbor, Michigan, USA) unter Verwendung von Standard-Parametern.

**[0041]** Der Begriff "Nukleotidveränderung" bezieht sich auf Nukleotidsubstitution, -deletion oder -insertion, wie in der Technik wohl bekannt ist.

**[0042]** "Haushältergene" ist ein in der Technik wohl bekannter Begriff und bedeutet jene Gene, die mit der allgemeinen Zellfunktion zusammenhängen, einschließlich, aber nicht begrenzt auf Wachstum, Teilung, Stase, Metabolismus und/oder Tod. "Haushälter"-Gene führen im Allgemeinen Funktionen durch, die in mehr als einem Zelltyp gefunden werden. Im Unterschied dazu führen zellspezifische Gene allgemein Funktionen in einem speziellen Zelltyp und/oder einer speziellen Zellklasse durch.

**[0043]** Der Begriff "Mittel" bedeutet hier eine biologische oder chemische Verbindung, wie ein einfaches oder komplexes, organisches oder anorganisches Molekül, ein Peptid, ein Protein oder ein Oligonukleotid, das die Funktion eines Polynukleotids oder Polypeptids moduliert. Es kann eine umfangreiche Gruppierung (Array) von Verbindungen synthetisiert werden, beispielsweise Oligomere, wie Oligopeptide und Oligonukleotide; und synthetische organische und anorganische Verbindungen, die auf verschiedenen Kernstrukturen basieren, und auch diese sind in den Begriff "Mittel" eingeschlossen. Außerdem können verschiedene natürliche Quellen Verbindungen zum Screening liefern, wie pflanzliche oder tierische Extrakte und dergleichen. Verbindungen können einzeln oder in Kombination miteinander getestet werden.

**[0044]** Der Begriff "Modulieren der Funktion" eines Polynukleotids oder eines Polypeptids bedeutet, dass die Funktion des Polynukleotids oder Polypeptids verändert wird, verglichen mit der fehlenden Zugabe eines Mittels. Die Modulation kann auf jeder Ebene stattfinden, die die Funktion beeinflusst. Eine Polynukleotid- oder Polypeptidfunktion kann direkt oder indirekt sein und direkt oder indirekt gemessen werden.

**[0045]** Eine "Funktion eines Polynukleotids" schließt Replikation, Translation, Expressionsmuster, ein, ist je-

doch nicht darauf begrenzt. Eine Polynukleotidfunktion schließt auch Funktionen ein, die mit einem Polypeptid assoziiert sind, das in dem Polynukleotid kodiert ist. Ein Mittel, das auf ein Polynukleotid wirkt und die Proteinexpression, -konformation, -faltung (oder andere physikalische Charakteristika), Bindung an andere Einheiten (wie Liganden), Aktivität (oder andere funktionale Charakteristika), Regulierung und/oder andere Aspekte der Proteinstruktur oder -funktion bewirkt, wird als modulierte Polynukleotidfunktion aufweisend angesehen.

**[0046]** Eine "Funktion eines Polypeptids" schließt Konformation, Faltung (oder andere physikalische Charakteristik), Bindung an andere Einheiten (wie Liganden), Aktivität (oder andere funktionale Charakteristika) und/oder andere Aspekte der Proteinstruktur oder -funktionen ein, ohne darauf begrenzt zu sein. Ein Mittel, das auf ein Polypeptid wirkt und seine Konformation, Faltung (oder andere physikalische Charakteristika), Bindung an andere Einheiten (wie Liganden), Aktivität (oder andere funktionale Charakteristika) und/oder andere Aspekte der Proteinstruktur oder -funktion bewirkt, wird beispielsweise als modulierte Polypeptidfunktion aufweisend angesehen. Die Weisen, nach denen ein wirksames Mittel wirken kann, um die Funktion eines Polypeptids zu modulieren, schließen 1) Veränderung der Koformations-, Faltungs- oder anderer physikalischer Charakteristika, 2) Änderung der Bindungsstärke an seinen natürlichen Liganden oder Änderung der Spezifität der Bindung an Liganden und 3) Änderung der Aktivität des Polypeptids ein, sind jedoch nicht darauf begrenzt.

**[0047]** Der Begriff "Zielstelle" bedeutet eine Position in einem Polypeptid, die eine einzelne Aminosäure sein kann und/oder Teil eines strukturellen und/oder funktionalen Motivs sein kann, z. B. eine Bindungsstelle, eine Dimerisierungsdomäne oder eine katalytisch aktive Stelle. Zielstellen können für die direkte oder indirekte Wechselwirkung mit einem Mittel, wie einem therapeutischen Mittel, brauchbar sein.

**[0048]** Der Begriff "molekularer Unterschied" schließt jeglichen strukturellen und/oder funktionalen Unterschied ein. Nachfolgend werden Verfahren zur Detektierung dieser Unterschiede sowie Beispiele für derartige Unterschiede beschrieben.

**[0049]** Eine "funktionale Wirkung" ist ein in der Technik gut bekannter Begriff und bedeutet jegliche Wirkung, die direkt oder indirekt auf irgendein Aktivitätsniveau gezeigt wird.

**[0050]** Der Begriff "Leichtigkeit der Ernte" bezieht sich auf Pflanzencharakteristika oder -merkmale, die manuelles oder automatisiertes Sammeln von Strukturen oder Teilen (z. B. Früchten, Blättern, Wurzeln) zum Verbrauch oder für weitere kommerzielle Verarbeitung erleichtern.

**[0051]** Allgemeine Verfahren, die in der Technik bekannt sind Die Quelle des Polynukleotids von der domestizierten Pflanze oder dem domestizierten Tier kann für die Zwecke dieser Erfindung jegliche geeignete Quelle sein, z. B. Genomsequenzen oder cDNA-Sequenzen. Vorzugsweise werden cDNA-Sequenzen verglichen. Protein kodierende Sequenzen können aus verfügbaren privaten, öffentlichen und/oder kommerziellen Datenbanken, wie den hier beschriebenen, erhalten werden. Diese Datenbanken dienen als Fundgruben für molekulare Sequenzdaten, die durch fortlaufende Forschungsprojekte herausgefunden werden. Protein kodierende Sequenzen können alternativ aus beispielsweise Sequenzierung der cDNA aus der reversen Transkription von in Zellen exprimierter m-RNA oder nach der PCR-Amplifizierung gemäß im Stand der Technik wohl bekannten Verfahren erhalten werden Alternativ können zum Sequenzvergleich Genomsequenzen verwendet werden. Genomsequenzen können aus verfügbaren öffentlichen, privaten und/oder kommerziellen Datenbanken oder aus einer Sequenzierung kommerziell erhältlicher Genom-DNA-Bibliotheken oder aus Genom-DNA nach PCR erhalten werden.

**[0052]** In einigen Ausführungsformen wird die cDNA aus mRNA hergestellt, die aus einem Gewebe in einem festgelegten Endwicklungsstadium erhalten wurde, oder aus einem Gewebe, das erhalten wurde, nachdem der Organismus bestimmten Umweltbedingungen ausgesetzt war. cDNA-Bibliotheken, die zum Sequenzvergleich der vorliegenden Erfindung verwendet wurden, können mit konventionellen cDNA-Bibliothekkonstruktionstechniken aufgebaut werden, die in der Literatur des Standes der Technik vollständig erläutert sind. Als Template (Schablonen) für Reverses Transkribieren von cDNAs werden vollständige mRNAs verwendet. Transkribierte cDNAs werden in geeignete Vektoren subkloniert, um eine cDNA-Bibliothek zu erzeugen. Die erzeugte cDNA-Bibliothek kann auf Gehalt an cDNA mit vollständiger Länge maximiert werden, obwohl cDNAs mit weniger als vollständiger Länge verwendet werden können. Die Sequenzfrequenz kann beispielsweise gemäß Bonaldo et al. (1996) Genome Research 6:791-806, normalisiert werden. cDNA-Klone, die statistisch aus der aufgebauten cDNA-Bibliothek ausgewählt worden sind, können mit automatisierten Standard-Sequenzieretechniken sequenziert werden. Für das Sequenzieren werden vorzugsweise cDNA-Klone mit vollständiger Länge verwendet. Es können entweder die vollständigen cDNA-Klone oder ein großer Abschnitt davon aus ei-

ner cDNA-Bibliothek sequenziert werden, obwohl einige Ausführungsformen der Erfindung auch durchgeführt werden können, indem so wenig wie ein einzelner cDNA-Klon oder mehrere cDNA-Klone sequenziert werden.

**[0053]** In einer bevorzugten Ausführungsform der vorliegenden Erfindung können zu sequenzierende cDNA-Klone gemäß ihrer Expressionsspezifität vorselektiert werden. Um cDNAs gemäß aktiven Genen zu selektieren, die spezifisch exprimiert werden, können die cDNAs mit mRNAs, die aus andere Organen, Geweben oder Zellen desselben Tieres erhalten wurden, Subtraktionshybridisierung unterzogen werden. Unter bestimmten Hybridisierungsbedingungen mit geeigneter Stringenz und Konzentration werden jene cDNAs, die mit nicht-gewebespezifischen mRNAs hybridisieren und somit wahrscheinlich "Haushälter"-Gene repräsentieren, aus dem cDNA-Pool ausgeschlossen. Die restlichen zu sequenzierenden cDNAs hängen demnach mit größerer Wahrscheinlichkeit mit gewebespezifischen Funktionen zusammen. Zum Zwecke der Subtraktionshybridisierung können nicht-gewebespezifische mRNAs aus einem Organ oder vorzugsweise aus einer Kombination unterschiedlicher Organe und Zellen erhalten werden. Die Menge der nicht-gewebespezifischen mRNAs wird maximiert, um die gewebespezifischen cDNAs zu sättigen.

**[0054]** Alternativ können Informationen aus Online-Datenbanken verwendet werden, um cDNAs zu selektieren oder ihnen Priorität zu geben, die wahrscheinlich zu spezifischen Funktionen gehören. Die Vorfahren-cDNA-Kandidaten zur Sequenzierung können beispielsweise durch PCR mittels Primern ausgewählt werden, die aus Kandidaten-cDNA-Sequenzen des domestizierten Organismus entworfen wurden. Kandidaten-cDNA-Sequenzen aus domestizierten Organismen sind beispielsweise jene, die sich nur in einem spezifischen Gewebe finden, wie Skelettmuskeln, oder die Genen entsprechen, die wahrscheinlich in der speziellen Funktion von Bedeutung sind. Derartige gewebespezifische cDNA-Sequenzen können durch Durchsuchen von Online-Sequenzdatenbanken erhalten werden, in denen Informationen in Bezug auf das Expressionsprofil und/oder die biologische Aktivität für cDNA-Sequenzen angegeben sein können.

**[0055]** Sequenzen für Vorfahren-Homolog(e) eines Gens eines domestizierten Organismus können nach Standardverfahren des Standes der Technik erhalten werden, wie PCR-Verfahren (wobei beispielsweise GeneAmp PCR System 9700 Temperaturwechselgeräte (Applied Biosystems, Inc.)) verwendet werden. Die Vorfahren-cDNA-Kandidaten zur Sequenzierung können beispielsweise durch PCR mittels Primern ausgewählt werden, die aus in Frage kommenden cDNA-Sequenzen des domestizierten Organismus entworfen wurden. Bei PCR können die Primer aus den Sequenzen des domestizierten Organismus unter Verwendung von Standardverfahren des Standes der Technik hergestellt werden, einschließlich öffentlich zugänglichen Primer-Designprogrammen, wie PRIMER<sup>®</sup> (Whitehead Institute). Die amplifizierte Vorfahrensequenz kann dann unter Verwendung von Standardverfahren und -geräten des Standes der Technik, wie automatisierten Sequenzern (Applied Biosystems, Inc.), sequenziert werden.

#### Allgemeine Verfahren der Erfindung

**[0056]** Das allgemeine erfindungsgemäße Verfahren ist wie folgt. Kurz gesagt werden Nukleotidsequenzen von einem domestizierten Organismus und einem wilden Vorfahren erhalten. Die Nukleotidsequenzen des domestizierten Organismus und des Vorfahren werden miteinander verglichen, um Sequenzen zu identifizieren, die homolog sind. Die homologen Sequenzen werden analysiert, um jene zu identifizieren, die Nukleinsäuresequenzdifferenzen zwischen dem domestizierten Organismus und dem Vorfahren aufweisen. Dann wird eine molekulare Evolutionsanalyse durchgeführt, um die evolutionäre Signifikanz der Unterschiede quantitativ und qualitativ zu bewerten. Bei Genen, die positiv selektiert worden sind, kann eine Fremdgruppenanalyse durchgeführt werden, um jene Gene zu identifizieren, die in dem domestizierten Organismus (im Unterschied zu dem Vorfahren) positiv selektiert worden sind. Als nächstes wird die Sequenz hinsichtlich der molekularen/genetischen Identität und der biologischen Funktion charakterisiert. Die Information kann schließlich verwendet werden, um Mittel zu identifizieren, die die biologische Funktion des von dem Gen kodierten Polypeptids zu modifizieren.

**[0057]** Die allgemeinen Verfahren der Erfindungen beinhalten das Vergleichen der Protein kodierenden Nukleotidsequenzen von Vorfahren- und domestizierten Organismen. Für den Vergleich wird Bioinformatik verwendet, und es werden Sequenzen selektiert, die eine Nukleotidänderung oder mehrere Nukleotidänderungen enthalten, die eine oder mehrere evolutionär signifikante Änderung(en) ist/sind. Die Erfindung ermöglicht die Identifizierung von Genen, die sich herausgebildet haben, um irgendeinen evolutionären Vorteil zu verleihen, und die Identifizierung der speziellen herausgebildeten Veränderungen.

**[0058]** Protein kodierende Sequenzen eines domestizierten Organismus und seines Vorfahren werden verglichen, um homologe Sequenzen zu identifizieren. Diese Erfindung umfasst jeglichen geeigneten Mechanis-

mus zur Durchführung dieses Vergleichs. Manuell oder mittels Software kann der Abgleich durchgeführt werden (Beispiele für geeignete Alignmentprogramme sind in der Technik bekannt). Vorzugsweise werden Protein kodierende Sequenzen eines Vorfahren über Datenbankrecherchen, z. B. BLAST-Recherchen, mit der Sequenz der domestizierten Spezies verglichen. Die hochrangigen "Hits", d. h. Sequenzen, die nach BLAST-Analysen eine signifikante Ähnlichkeit zeigen, werden abgerufen und analysiert. Sequenzen, die eine signifikante Ähnlichkeit zeigen, können jene mit mindestens etwa 60% mindestens etwa 75%, mindestens etwa 80%, mindestens etwa 85% oder mindestens etwa 90% Sequenzidentität sein. Sequenzen, die mehr als etwa 80% Identität zeigen, werden vorzugsweise weiter analysiert. Die über Datenbankrecherchen identifizierten homologen Sequenzen können vollständig mittels Sequenzalignmentverfahren und -programmen, die bekannt und in der Technik verfügbar sind, wie dem üblicherweise verwendeten einfachen Alignmentprogramm CLUSTAL V von Higgins et al. (1992) CABIOS 8:189-191, Alignment unterzogen werden.

**[0059]** Alternativ kann der Sequenzier- und Homologievergleich der Protein kodierenden Sequenzen zwischen dem domestizierten Organismus und seinem Vorfahren simultan unter Verwendung der neu entwickelten Sequenzierchiptechnologie durchgeführt werden. Siehe beispielsweise Rava et al., US-5,545,531.

**[0060]** Die abgeglichenen Protein kodierenden Sequenzen von domestiziertem Organismus und Vorfahren werden analysiert, um Nukleotidsequenzunterschiede an speziellen Stellen zu identifizieren. Wiederum ist erfindungsgemäß jedes geeignete Verfahren eingeschlossen, um dies zu erreichen. Wenn es keine Nukleotidsequenzunterschiede gibt, wird die Vorfahren-Proteinkodiersequenz üblicherweise nicht weiter analysiert. Die detektierten Sequenzänderungen werden allgemein und vorzugsweise anfangs auf Genauigkeit geprüft. Die Anfangsprüfung beinhaltet vorzugsweise die Durchführung von einer oder mehreren der folgenden Stufen, die alle jeweils in der Technik bekannt sind: (a) Ermitteln der Punkte, an denen es Veränderungen zwischen den Sequenzen des Vorfahren und des domestizierten Organismus gibt; (b) Prüfen des Sequenzfluorogramms (Chromatogramms), um zu bestimmen, ob die Basen, die für den Vorfahren oder den domestizierten Spezies einzigartig zu sein scheinen, klaren starken Signalen entsprechen, die für die angesprochene Base spezifisch sind; (c) Prüfen der "Hits" des domestizierten Organismus, um zu sehen, ob es mehr als eine Sequenz des domestizierten Organismus gibt, die einer Sequenzveränderung entspricht. Mehrere Einträge der Sequenz des domestizierten Organismus für dasselbe Gen, die dasselbe Nukleotid an einer Position haben, an der es in einer Vorfahrensequenz ein anderes Nukleotid gibt, sprechen unabhängig voneinander dafür, dass die domestizierte Sequenz akkurat ist und die Änderung signifikant ist. Derartige Veränderungen werden unter Verwendung von Datenbankinformationen und dem genetischen Kode untersucht, um zu bestimmen, ob diese Nukleotidsequenzveränderungen zu einer Veränderung der Aminosäuresequenz des kodierten Proteins führen. Wie die Definition von "Nukleotidänderung" verdeutlicht, umfasst die vorliegende Erfindung mindestens eine Nukleotidänderung, entweder eine Substitution, eine Deletion oder eine Insertion, in einer Protein kodierenden Polynukleotidsequenz eines domestizierten Organismus, verglichen mit einer entsprechenden Sequenz des Vorfahren. Die Änderung ist vorzugsweise eine Nukleotidsubstitution. Insbesondere ist in der identifizierten Sequenz mehr als eine Substitution vorhanden und wird molekularer Evolutionsanalyse unterzogen.

**[0061]** Es kann jede von mehreren verschiedenen molekularen Evolutionsanalysen oder Verfahren vom  $K_A/K_S$ -Typ verwendet werden, um die evolutionäre Signifikanz der identifizierten Nukleotidänderungen zwischen Gensequenzen domestizierter Spezies und denjenigen der entsprechenden Vorfahren quantitativ und qualitativ zu bewerten. Kreitman and Akashi (1995) Annu. Rev. Ecol. Syst. 26:403-422; Li, Molecular Evolution, Sinauer Associates, Sunderland, MA, 1997. Positive Selektion von Proteinen (d. h. adaptive Evolution auf molekularer Ebene) kann beispielsweise in Protein kodierenden Genen detektiert werden, indem die Verhältnisse nicht synonyme Nukleotidsubstitutionen pro nicht synonyme Stelle ( $K_A$ ) zu synonymen Substitutionen pro synonyme Stelle ( $K_S$ ) (Li et al., 1985; Li, 1993) paarweise verglichen werden. Es kann jeder beliebige Vergleich von  $K_A$  und  $K_S$  verwendet werden, obwohl es besonders zweckmäßig und am effektivsten ist, diese beiden Variablen als Verhältnis zu vergleichen. Sequenzen werden unter Verwendung von statistischen Standardverfahren dadurch identifiziert, dass sie einen statistisch signifikanten Unterschied zwischen  $K_A$  und  $K_S$  zeigen.

**[0062]** Vorzugsweise wird zur Durchführung der vorliegenden Erfindung die  $K_A/K_S$ -Analyse von Li et al. verwendet, obwohl auch andere Analyseprogramme verwendet werden können, die positiv selektierte Gene zwischen Spezies detektieren können. Li et al. (1985) Mol. Biol. Evol. 2:150-174; Li (1993); siehe auch J. Mol. Evol. 36:96-99; Messier und Stewart (1997) Nature 385:151-154; Nei (1987) Molecular Evolutionary Genetics (New York, Columbia University Press). Das  $K_A/K_S$ -Verfahren, das einen Vergleich der Rate der nicht-synonymen Substitutionen pro nicht-synonymen Stelle mit der Rate der synonymen Substitutionen pro synonymen Stelle zwischen homologen Protein kodierenden Genregionen in Form eines Verhältnisses beinhaltet, wird zur Identifizierung von Sequenzsubstitutionen verwendet, die im Unterschied zu neutralen Selektionen während der Evolution durch adaptive Selektionen angetrieben sein können. Eine synonyme ("stumme") Substitution ist

eine, die wegen der Degeneriertheit des genetischen Kodes keine Veränderung an der kodierten Aminosäuresequenz vornimmt, eine nicht-synonyme Substitution führt zum Ersetzen einer Aminosäure. Das Ausmaß jedes Änderungstyps kann als  $K_A$  beziehungsweise  $K_S$ , die Anzahl der synonymen Substitutionen pro synonyme Stelle und der nicht-synonymen Substitutionen pro nicht-synonymer Stelle geschätzt werden. Die Berechnungen von  $K_A/K_S$  können manuell oder mit Software durchgeführt werden. Ein Beispiel für ein geeignetes Programm ist MEGA (Molecular Genetics Institute, Pennsylvania State University).

**[0063]** Zur Schätzung von  $K_A$  und  $K_S$  werden entweder vollständige oder partielle Protein kodierende Sequenzen verwendet, um die Gesamtanzahl der synonymen und nicht-synonymen Substitutionen sowie der nicht-synonymen und synonymen Stellen zu berechnen. Die Länge der analysierten Polynukleotidsequenz kann jede geeignete Länge sein. Vorzugsweise wird die gesamte Kodiersequenz verglichen, um jegliche und alle signifikanten Veränderungen zu ermitteln. Allgemein zugängliche Computerprogramme, wie Li93 (Li (1993) J. Mol. Evol. 36:96-99) oder INA, können verwendet werden, um die  $K_A$ - und  $K_S$ -Werte für alle paarweisen Vergleiche zu berechnen. Diese Analyse kann ferner angepasst werden, um Sequenzen in wie in einem "Schiebefenster" zu untersuchen, so dass geringe Zahlen von wichtigen Änderungen nicht von der Gesamtsequenz zugedeckt werden. "Schiebefenster" bezieht sich auf die Untersuchung fortlaufender Unterabschnitte des Gens (die Unterabschnitte können eine beliebige Länge haben).

**[0064]** Der Vergleich der nicht-synonymen und synonymen Substitutionsraten wird durch das  $K_A/K_S$ -Verhältnis wiedergegeben. Es ist gezeigt worden, dass  $K_A/K_S$  den Grad widerspiegelt, bis zu dem die adaptive Evolution in der untersuchten Sequenz gewirkt hat. Für die  $K_A/K_S$ -Analyse können die vollständige Länge oder partielle Segmente einer Kodiersequenz verwendet werden. Je höher das  $K_A/K_S$ -Verhältnis ist, um so größer ist die Wahrscheinlichkeit, dass eine Sequenz adaptive Evolution eingegangen ist und die nicht-synonymen Substitutionen evolutionär signifikant sind. Siehe beispielsweise Messier and Stewart (1997). Das  $K_A/K_S$ -Verhältnis ist vorzugsweise mindestens etwa 0,75, insbesondere mindestens etwa 1,0, insbesondere mindestens etwa 1,25, insbesondere mindestens etwa 1,50 oder bevorzugter mindestens etwa 2,00. Die statistische Analyse wird vorzugsweise mit allen bewerteten  $K_A/K_S$ -Verhältnissen durchgeführt, einschließlich, aber nicht begrenzt auf Standardtestverfahren, wie der Student-T-Test und Wahrscheinlichkeitsverhältnistests, die von Yang (1998) Mol. Biol Evol. 37:441-456 beschrieben worden sind.

**[0065]** Bei einem paarweisen Vergleich homologer Sequenzen sind  $K_A/K_S$ -Verhältnisse deutlich größer als 1 ein deutlicher Hinweis darauf, dass positive Selektion eine größere Zahl von Aminosäureersetzungen fixiert hat, als lediglich als Ergebnis des Zufalls zu erwarten wäre, und steht im Gegensatz zu dem üblicherweise beobachteten Muster, bei dem das Verhältnis kleiner als oder gleich 1 ist. Nei (1987); Hughes and Nei (1988) Nature 335:167-170; Messier und Stewart (1994) Current Biol. 4:911-913; Kreitman und Akashi (1995) Ann. Rev. Ecol. Syst. 26:403-422; Messier und Stewart (1997). Verhältnisse kleiner als 1 betonen allgemein die Rolle der negativen oder reinigenden Selektion: es gibt einen starken Druck auf die Primärstruktur funktionaler, wirksamer Proteine, unverändert zu bleiben.

**[0066]** Alle Verfahren zur Berechnung von  $K_A/K_S$ -Verhältnissen basieren auf einem paarweisen Vergleich der Zahl der nichtsynonymen Substitutionen pro nicht-synonymer Stelle zu der Zahl der synonymen Substitutionen pro synonyme Stelle für die Protein kodierenden Regionen homologer Gene von den Vorfahren- und domestizierten Organismen. Jedes Verfahren implementiert unterschiedliche Korrekturen, um "mehrere Hits" zu erreichen (d. h. mehr als eine Nukleotidsubstitution an der gleichen Stelle). Jedes Verfahren verwendet auch unterschiedliche Modelle, wie sich DNA-Sequenzen im Verlauf der Evolution verändern. Somit wird vorzugsweise eine Kombination von Ergebnissen unterschiedlicher Algorithmen verwendet, um das Empfindlichkeitsniveau zur Detektion positiv selektierter Gene und das Vertrauen in das Ergebnis zu erhöhen.

**[0067]** Es sollten vorzugsweise  $K_A/K_S$ -Verhältnisse für orthologe Genpaare berechnet werden, im Unterschied zu paralogen Genpaaren (d. h. einem Gen, das aus Speziesbildung resultiert, im Unterschied zu einem Gen, welches das Ergebnis von Genduplizierung ist), Messier und Stewart (1997). Diese Unterscheidung kann vorgenommen werden, indem zusätzliche Vergleiche mit weiteren Vorfahren vorgenommen werden, die das Aufbauen eines phylogenetischen Baums ermöglichen. Orthologe Gene ergeben, wenn sie zum Aufbau des Baumes verwendet werden, den bekannten "Speziesbaum", d. h. sie erzeugen einen Baum, der wieder den bekannten biologischen Baum ergibt. Im Unterschied dazu ergeben paraloge Gene Bäume, die mit dem bekannten biologischen Baum nicht in Einklang zu bringen sind.

**[0068]** Es sei darauf hingewiesen, dass die hier beschriebenen Verfahren zur Identifizierung von Polynukleotidsequenzen von Vorfahren- oder domestizierten Organismen führen können, die mit den Protein kodierenden Sequenzen funktional verwandt sind. Zu derartigen Sequenzen können nicht-kodierende Sequenzen oder

Kodiersequenzen gehören, die keine Proteine kodieren. Diese verwandten Sequenzen können beispielsweise physikalisch neben den Protein kodierenden Sequenzen in dem Genom liegen, wie Introns oder 5'- und 3'-flankierende Sequenzen (einschließlich Kontrollelementen, wie Promotern und Enhancern). Diese verwandten Sequenzen können durch Recherchen in verfügbaren öffentlichen, privaten und/oder kommerziellen Genomdatenbanken oder alternativ durch Screening und Sequenzieren der Genombibliothek des Organismus mit einer Proteinkodiersequenz als Sonde erhalten werden. Dem Fachmann sind Verfahren und Techniken zum Erhalten von Nicht-Kodiersequenzen unter Verwendung von verwandten Kodiersequenzen gut bekannt.

**[0069]** Die evolutionär signifikanten Nukleotidänderungen, die durch molekulare Evolutionsanalysen, wie  $K_A/K_S$ -Analyse detektiert werden, können ferner nach ihrem einzigartigen Auftreten in dem domestizierten Organismus oder dem Grad, bis zu dem die Änderungen in dem domestizierten Organismus einzigartig sind, bewertet werden. Die identifizierten Änderungen in dem domestizierten Gen können beispielsweise auf Anwesenheit/Abwesenheit in anderen Sequenzen verwandter Spezies, Subspezies oder anderer Organismen mit einem gemeinsamen Vorfahren mit dem domestizierten Organismus getestet werden. Dieser Vergleich ("Fremdgruppenanalyse") ermöglicht die Bestimmung, ob das positiv selektierte Gen für den fraglichen domestizierte Organismus (im Unterschied zu dem Vorfahren) positiv selektiert worden ist.

**[0070]** Die Sequenzen mit mindestens einer evolutionär signifikanten Änderung zwischen einem domestizierten Organismus und seinem Vorfahren können als Primer für PCR-Analyse anderer Vorfahren-Proteinkodiersequenzen verwendet werden, und die resultierenden Polynukleotide werden sequenziert, um zu sehen, ob die gleiche Änderung in anderen Vorfahren vorhanden ist. Diese Vergleiche ermöglichen die weitere Diskriminierung, ob die adaptiven evolutionären Veränderungen für die domestizierte Linie einzigartig sind, verglichen mit anderen Vorfahren, oder ob die adaptive Änderung für die Vorfahren einzigartig ist, verglichen mit der domestizierten Spezies und anderen Vorfahren. Eine Nukleotidänderung, die in einem domestizierten Organismus, jedoch nicht in anderen Vorfahren detektiert wird, steht mit größerer Wahrscheinlichkeit für eine adaptive evolutionäre Veränderung in dem domestizierten Organismus. Alternativ steht eine Nukleotidänderung, die in einem Vorfahren detektiert wird, jedoch nicht in dem domestizierten Organismus oder anderen Vorfahren detektiert ist, wahrscheinlich für eine adaptive evolutionäre Veränderung des Vorfahren. Andere zum Vergleich herangezogene Vorfahren können bezogen auf ihre phylogenetische Verwandtschaft mit dem domestizierten Organismus gewählt werden. Die statistische Signifikanz dieser Vergleiche kann mit etablierten verfügbaren Programmen bestimmt werden, z. B. T-Test, verwendet von Messier und Stewart (1997) Nature 385:151-154. Jene Gene, die statistisch hohe  $K_A/K_S$ -Verhältnisse zeigen, sind sehr wahrscheinlich adaptive Evolution eingegangen.

**[0071]** Sequenzen mit signifikanten Änderungen könne als Sonden in Genomen aus unterschiedlich domestizierten Populationen eingesetzt werden, um zu sehen, ob die Sequenzänderungen von mehr als einer domestizierten Population geteilt werden. Gensequenzen von unterschiedlichen domestizierten Populationen können aus Datenbanken oder alternativ aus direkter Sequenzierung PCR-amplifizierter DNA aus mehrere nicht verwandten, diversen, domestizierten Populationen erhalten werden. Die Anwesenheit der identifizierten Änderungen in verschiedenen domestizierten Populationen würde ferner die evolutionäre Bedeutung der Änderungen zeigen.

**[0072]** Sequenzen mit signifikanten Änderungen zwischen Spezies können ferner durch ihre molekularen/genetischen Identitäten und biologischen Funktionen unter Verwendung von Verfahren und Techniken charakterisiert werden, die Fachleuten bekannt sind. Die Sequenzen können beispielsweise genetisch und physikalisch mit öffentlich zugänglichen Bioinformatikprogrammen innerhalb des Genoms des Organismus lokalisiert werden. Die neu identifizierten signifikanten Änderungen in der Nukleotidsequenz können eine potentielle Rolle des Gens in der Evolution des Organismus und eine potentielle Assoziation mit einzigartigen, verbesserten oder veränderten funktionalen Möglichkeiten nahe legen. Das putative Gen mit den identifizierten Sequenzen kann ferner durch beispielsweise Homologiesuche gekennzeichnet werden. Gemeinsame Homologie des putativen Gens mit einem bekannten Gen kann eine ähnliche biologische Rolle oder Funktion anzeigen. Ein weiteres beispielhaftes Verfahren zur Charakterisierung einer putativen Gensequenz ist auf Basis bekannter Sequenzmotive. Bestimmte Sequenzmuster kodieren bekanntermaßen Regionen von Proteinen mit speziellen biologischen Charakteristika, wie Signalsequenzen, DNA-Bindungsdomänen oder Transmembrandomänen.

**[0073]** Die identifizierten Sequenzen mit signifikanten Veränderungen können ferner auch bewertet werden, indem danach gesucht wird, wo das Gen in Form von Gewebe- oder Zelltypspezifität exprimiert wird. Die identifizierte Kodiersequenz kann beispielsweise als Sonden zur Durchführung von in situ-mRNA-Hybridisierung verwendet werden, die das Expressionsmuster der Sequenzen offenbaren wird. Gene, die in bestimmten Geweben exprimiert werden, können bessere Kandidaten sein, da sie mit wichtigen Funktionen assoziiert sind,

die mit jenem Gewebe assoziiert sind, beispielsweise Skelettmuskelgewebe. Es kann auch der Zeitpunkt der Genexpression während jedes Entwicklungsstadiums eines Spezieselements bestimmt werden.

**[0074]** Als weiteres beispielhaftes Verfahren zur Sequenzcharakterisierung können die funktionalen Rollen der identifizierten Nukleotidsequenzen mit signifikanten Änderungen bewertet werden, indem funktionale Assays für unterschiedliche Allele eines identifizierten Gens in dem transfektierten domestizierten Organismus durchgeführt werden, z. B. in der transgenen Pflanze oder dem transgenen Tier.

**[0075]** Als weiteres beispielhaftes Verfahren zur Sequenzcharakterisierung ermöglicht die Verwendung von Computerprogrammen Modellierung und Visualisierung der dreidimensionalen Struktur der homologen Proteine aus domestizierten Organismen und Vorfahren. Die genaue Kenntnis, welche Aminosäuren in dem Vorfahrenprotein/den Vorfahrenproteinen ersetzt worden ist, ermöglicht spezifisch die Detektierung struktureller Veränderungen, die mit funktionalen Unterschieden zusammenhängen können. Die Verwendung von Modellierungstechniken steht somit in engem Zusammenhang mit der Identifizierung von funktionalen Rollen, die in dem vorhergehenden Absatz erörtert wurden. Die Verwendung von individuellen oder Kombinationen dieser Techniken bildet somit einen Teil der vorliegenden Erfindung.

**[0076]** Ein Gen eines domestizierten Organismus, das nach dem vorliegende Verfahren identifiziert worden ist, kann somit zur Identifizierung homologer Gene in anderen Spezies eingesetzt werden, die einen gemeinsamen Vorfahren haben. Mais, Reis, Weizen, Millet und Sorghum haben beispielsweise einen gemeinsamen Vorfahren, und in Mais identifizierte Gene können direkt zu homologen Genen in diesen anderen Gräsern führen. In ähnlicher Weise haben Tomaten und Kartoffeln einen gemeinsamen Vorfahren, und nach dem vorliegenden Verfahren in Tomaten identifizierte Gene haben vermutlich Homologe in Kartoffeln.

**[0077]** Die nach den hier beschriebenen Verfahren identifizierten Sequenzen können zur Identifizierung von Mitteln verwendet werden, die zur Modulierung von einzigartigen, verbesserten oder geänderten funktionalen Möglichkeiten des domestizierten Organismus brauchbar sind und/oder Defekte dieser Möglichkeiten unter Verwendung dieser Sequenzen korrigieren. Diese Verfahren verwenden beispielsweise in der Technik bekannte Screening-Techniken, wie in vitro-Systeme, Expressionssysteme auf Zellbasis sowie transgene nicht-menschliche Tiere und Pflanzen. Der von der vorliegenden Erfindung bereitgestellte Ansatz identifiziert nicht nur rasch entwickelte Gene, sondern zeigt auch Modulationen, die an dem Protein vorgenommen werden können und möglicherweise nicht zu toxisch sind, da sie in einer andern Spezies existieren.

#### Screening-Verfahren

**[0078]** Die vorliegende Erfindung liefert auch Screeningverfahren unter Verwendung der Polynukleotide und Polypeptide, die nach den oben beschriebenen Verfahren identifiziert und charakterisiert wurden. Diese Screening-Verfahren sind zur Identifizierung von Mitteln brauchbar, die die Funktion(en) der Polynukleotide oder Polypeptide in einer Weise modulieren können, die zur Verbesserung oder Verminderung eines Charakteristikums in einem domestizierten Organismus nützlich ist. Bei diesen Verfahren bringt man allgemein mindestens ein zu testendes Mittel mit entweder einem transgenen Organismus oder einer transgenen Zelle in Kontakt, der/die mit einer nach den oben beschriebenen Verfahren identifizierten Polynukleotidsequenz transfektiert worden ist, oder einer Präparation des Polypeptids, das durch diese Polynukleotidsequenz kodiert worden ist, wobei ein Mittel durch seine Fähigkeit zur Modulierung der Funktion entweder der Polynukleotidsequenz oder des Polypeptids identifiziert wird. Ein Mittel kann beispielsweise eine Verbindung sein, die mit einer domestizierten Pflanze oder einem domestizierten Tier eingesetzt oder in Kontakt gebracht wird, um die Expression des identifizierten Gens zu einer gewünschten Zeit zu induzieren. In Hinsicht auf Pflanzen könnte speziell ein Mittel verwendet werden, um das Blühen zu einer geeigneten Zeit zu induzieren.

**[0079]** Der Begriff "Mittel" bedeutet hier eine biologische oder chemische Verbindung, wie ein einfaches oder komplexes, organisches oder anorganisches Molekül, ein Peptid, ein Protein oder ein Oligonukleotid. Es kann eine umfangreiche Gruppierung (Array) von Verbindungen synthetisiert werden, beispielsweise Oligomere, wie Oligopeptide und Oligonukleotide, und synthetische organische und anorganische Verbindungen, die auf verschiedenen Kernstrukturen basieren, und auch diese sind in den Begriff "Mittel" eingeschlossen. Außerdem können verschiedene natürliche Quellen Verbindungen zum Screening liefern, wie pflanzliche oder tierische Extrakte und dergleichen. Verbindungen können einzeln oder in Kombination miteinander getestet werden.

**[0080]** Das "Modulieren der Funktion" eines Polynukleotids oder eines Polypeptids bedeutet, dass die Funktion des Polynukleotids oder Polypeptids verändert wird, verglichen mit der fehlenden Zugabe eines Mittels. Die Modulation kann auf jeder Ebene stattfinden, die die Funktion beeinflusst. Eine Polynukleotid- oder Poly-

peptidfunktion kann direkt oder indirekt sein und direkt oder indirekt gemessen werden. Eine "Funktion" eines Polynukleotids schließt Replikation, Translation, Expressionsmuster, ein, ist jedoch nicht darauf begrenzt. Eine Polynukleotidfunktion schließt auch Funktionen ein, die mit einem Polypeptid assoziiert sind, das in dem Polynukleotid kodiert ist. Ein Mittel, das auf ein Polynukleotid wirkt und die Proteinexpression, -konformation, -faltung (oder andere physikalische Charakteristika), Bindung an andere Einheiten (wie Liganden), Aktivität (oder andere funktionale Charakteristik), Regulierung und/oder andere Aspekte der Proteinstruktur oder -funktion bewirkt, wird als modulierte Polynukleotidfunktion aufweisend angesehen. Die Wirkweisen eines effektiven Mittels zur Modulation der Expression eines Polynukleotids schließen 1) Modifizieren der Bindung eines Transkriptionsfaktors an das auf einen Transkriptionsfaktor reagierende Element in dem Polynukleotid, 2) Modifizieren der Wechselwirkung zwischen zwei Transkriptionsfaktoren, die für die Expression des Polynukleotids erforderlich sind; 3) Ändern der Fähigkeit eines Transkriptionsfaktors, der für die Expression des Polynukleotids erforderlich ist, zum Eintreten in den Kern; 4) Inhibieren der Aktivierung eines an der Transkription des Polynukleotids beteiligten Transkriptionsfaktors; 5) Modifizieren eines Zelloberflächenrezeptors, der normalerweise mit einem Liganden in Wechselwirkung tritt und dessen Bindung des Liganden zur Expression des Polynukleotids führt; 6) Inhibieren der Inaktivierung einer Komponente der Signaltransduktionskaskade, die zu Expression des Polynukleotids führt, und 7) Steigerung der Aktivierung eines an der Transkription des Polynukleotids beteiligten Transkriptionsfaktors ein, ohne darauf begrenzt zu sein.

**[0081]** Eine "Funktion" eines Polypeptids schließt Konformation, Faltung (oder andere physikalische Charakteristik), Bindung an andere Einheiten (wie Liganden), Aktivität (oder andere funktionale Charakteristika) und/oder andere Aspekte der Proteinstruktur oder -funktionen ein, ohne darauf begrenzt zu sein. Ein Mittel, das auf ein Polypeptid wirkt und seine Konformation, Faltung (oder andere physikalische Charakteristika), Bindung an andere Einheiten (wie Liganden), Aktivität (oder andere funktionale Charakteristik) und/oder andere Aspekte der Proteinstruktur oder -funktion bewirkt, wird beispielsweise als modulierte Polypeptidfunktion aufweisend angesehen. Die Weisen, nach denen ein wirksames Mittel wirken kann, um die Funktion eines Polypeptids zu modulieren, schließen 1) Veränderung der Konformations-, Faltungs- oder anderer physikalischer Charakteristika, 2) Änderung der Bindungsstärke an seinen natürlichen Liganden oder Änderung der Spezifität der Bindung an Liganden und 3) Änderung der Aktivität des Polypeptids ein, sind jedoch nicht darauf begrenzt.

**[0082]** Die Wahl der Mittel für das Screening wird allgemein durch mehrere Parameter bestimmt, wie das spezielle angestrebte Polynukleotid oder Polypeptid, seine ausgeübte Funktion, seine dreidimensionale Struktur (falls bekannt oder vermutet) sowie andere Aspekte des rationalen Drug Designs. Es können auch Techniken der kombinatorischen Chemie verwendet werden, um zahlreiche Permutationen von Kandidaten zu erzeugen. Fachleute können geeignete Mittel zum Test ersinnen und/oder erhalten.

**[0083]** Die hier beschriebenen in vivo-Screening-Assays können gegenüber konventionellen Drug-Design-Assays etliche Vorteile haben: 1) wenn ein Mittel in eine Zelle eintreten muss, um eine gewünschte therapeutische Wirkung zu erreichen, kann ein in vivo-Assay eine Angabe darüber machen, ob das Mittel in eine Zelle eintreten kann; 2) ein in vivo-Screening-Assay kann Mittel identifizieren, die in dem Zustand, in dem sie zu dem Assaysystem gegeben werden, unwirksam zum Zeigen mindestens eines Charakteristikums sind, das mit der Modulation der Polynukleotid- oder Polypeptidfunktion zusammenhängt, jedoch, nachdem sie sich in einer Zelle befinden, durch zelluläre Komponenten in einer solchen Weise modifiziert werden, dass sie wirksame Mittel werden, 3) ein in vivo-Assaysystem erlaubt, was am wichtigsten ist, die Identifizierung von Mitteln, die irgendeine Komponente eines Stoffwechselwegs beeinflussen, die letztendlich zu Charakteristika führt, die mit Polynukleotid- oder Polypeptidfunktion zusammenhängen.

**[0084]** Das Screening kann allgemein durchgeführt werden, indem einer Probe geeigneter Zellen, die mit einem unter Verwendung der erfindungsgemäßen Verfahren identifizierten Polynukleotid transfektiert worden sind, ein Mittel zugefügt wird und die Wirkung, d. h. Modulation einer Funktion des Polynukleotids oder des in dem Polynukleotid kodierten Polypeptids, überwacht wird. Das Experiment schließt vorzugsweise eine Kontrollprobe ein, die das in Frage kommende Mittel nicht erhält. Die behandelten und unbehandelten Zellen werden danach durch jedes geeignete phänotypische Kriterium verglichen, einschließlich, aber nicht begrenzt auf mikroskopische Analyse, Vitalitätstest, Fähigkeit zur Replikation, histologische Untersuchung, die Konzentration einer speziellen RNA oder eines speziellen Polypeptids, die/das mit den Zellen zusammenhängt, das Niveau der enzymatischen Aktivität, welches von den Zellen oder Zelllysaten exprimiert wird, die Wechselwirkungen der Zellen, wenn sie infektiösen Agentien ausgesetzt werden, und die Fähigkeit der Zellen zur Wechselwirkung mit anderen Zellen oder Verbindungen. Unterschiede zwischen behandelten und unbehandelten Zellen zeigen Wirkungen, die auf das in Frage kommende Mittel zurückzuführen sind. Das Mittel hat optimalerweise eine größere Wirkung auf die experimentellen Zellen als auf Kontrollzellen. Zu geeigneten Wirtszellen

gehören, ohne darauf begrenzt zu sein, eukariotische Zellen, vorzugsweise Säugerzellen. Die Wahl der Zelle hängt mindestens teilweise von der Art des vorgesehenen Assays ab.

**[0085]** Um Mittel zu testen, die die Expression eines Polynukleotids nach oben regulieren, wird eine geeignete Wirtszelle, die mit einem interessierenden Polynukleotid transfektiert ist, so dass das Polynukleotid exprimiert ist (Expression schließt hier Transkription und/oder Translation ein) mit einem zu testenden Mittel in Kontakt gebracht. Ein Mittel würde auf seine Fähigkeit getestet, zu erhöhter Expression von mRNA und/oder Polypeptid zu führen. Verfahren zur Herstellung von Vektoren und Transfektion sind in der Technik wohl bekannt. "Transfektion" umfasst jedes Verfahren zur Einführung der exogenen Sequenz einschließlich beispielsweise Lipofektion, Transduktion, Infektion oder Elektroporation. Das exogene Polynukleotid kann als nicht-integrierter Vektor (wie ein Plasmid) aufrechterhalten bleiben, oder in das Wirtsgenom integriert werden.

**[0086]** Um Mittel zu identifizieren, die spezifisch Transkription aktivieren, können Transkriptionsregulationsregionen an ein Reporter-gen gelinkt werden und das Konstrukt an eine passende Wirtszelle addiert werden. Der Begriff "Reporter-gen" bedeutet hier ein Gen, welches ein Genprodukt kodiert, das identifiziert werden kann (d. h. ein Reporterprotein). Zu Reporter-genen gehören, ohne darauf begrenzt zu sein, alkalische Phosphatase, Chloramphenicol, Acetyltransferase,  $\beta$ -Galactosidase, Luciferase und grünes Fluoreszenzprotein (GFP). Identifizierungsverfahren für die Produkte von Reporter-genen schließen enzymatische Assays und fluorimetrische Assays ein, ohne darauf begrenzt zu sein. Reporter-gene und Assays zur Detektierung ihrer Produkte sind in der Technik wohl bekannt und beispielsweise in Ausubel et al. (1987) und periodischen Aktualisierungen beschrieben. Reporter-gene, Reporter-genassays und Reagenzkits sind auch aus kommerziellen Quellen leicht erhältlich. Zu Beispielen für geeignete Zellen gehören, ohne darauf begrenzt zu sein, Pilz-, Hefe-, Säuger- und andere eukariotische Zellen. Ein durchschnittlich versierter Praktiker ist mit Techniken zur Transfektierung eukariotischer Zellen einschließlich der Herstellung eines geeigneten Vektors, wie eines viralen Vektors, dem Einschleusen des Vektors in die Zelle, wie durch Elektroporation, und dem Selektieren von Zellen, die transformiert worden sind, wie durch Verwendung eines Reporter- oder Arzneimittelempfindlichkeitselements, durchaus vertraut. Die Wirkung eines Mittels auf die Transkription von der Regulationsregion in diesen Konstrukten wird durch die Aktivität des Reporter-genprodukts bewertet.

**[0087]** Neben dem Anstieg der Expression unter Bedingungen, unter denen sie wie bereits erwähnt normalerweise unterdrückt wird, könnte die Expression herabgesetzt werden, wenn normalerweise eine Expression erfolgen würde. Ein Mittel sollte dies durch eine Abnahme der Transkriptionsrate bewirken, und das oben beschriebene Reporter-gen-system wäre ein Mittel, um dies zu untersuchen. Die Wirtszellen zur Bewertung dieser Mittel müssten die Expression zulassen.

**[0088]** Zellen, die mRNA (von dem interessierenden Polynukleotid) transkribieren, könnten zur Identifizierung von Mitteln verwendet werden, die die Halbwertszeit von mRNA und/oder die Translation von mRNA spezifisch modulieren. Derartige Zellen würden auch zur Bewertung der Wirkung eines Mittels auf die Verarbeitung und/oder post-translationalen Modifizierung des Polypeptids verwendet. Ein Mittel könnte die Menge des Polypeptids in einer Zelle modulieren, indem der Umsatz des Polypeptids modifiziert wird (d. h. die Halbwertszeit erhöht oder verringert wird). Die Spezifität des Mittels in Bezug auf die mRNA und das Polypeptid würde durch Untersuchen der Produkte in Abwesenheit des Mittels und durch Untersuchen der Produkte von nicht verwandten mRNAs und Polypeptiden bestimmt. Fachleute kennen Verfahren zur Untersuchung der Halbwertszeit von mRNA, der Proteinverarbeitung und des Proteinumsatzes.

**[0089]** In vivo-Screeningverfahren können zur Identifizierung von Mitteln, die die Polypeptidfunktion durch die Wechselwirkung mit dem Polypeptid direkt modulieren, auch brauchbar sein. Diese Mittel können normale Polypeptid-Ligand-Wechselwirkungen blockieren, falls vorhanden, oder können diese Wechselwirkungen erhöhen oder stabilisieren. Diese Mittel können auch eine Konformation des Polypeptids verändern. Die Wirkung des Mittels kann unter Verwendung von Immunopräzipitationsreaktionen bestimmt werden. Es werden geeignete Antikörper verwendet, um das Polypeptid und jedes eng damit assoziierte Protein auszufällen. Durch Vergleichen der aus behandelten Zellen und aus unbehandelten Zellen immunopräzipitierten Polypeptide kann ein Mittel identifiziert werden, das die Polypeptid-Ligand-Wechselwirkungen, soweit vorhanden, erhöht oder hemmt. Polypeptid-Ligand-Wechselwirkungen können auch unter Verwendung von Vernetzungsreagentien bewertet werden, die eine enge, jedoch nicht-kovalente Wechselwirkung zwischen Polypeptiden in eine kovalente Wechselwirkung umwandeln. Techniken zur Untersuchung von Protein-Protein-Wechselwirkungen sind Fachleuten wohl bekannt. Techniken zur Untersuchung von Proteinkonformation sind Fachleuten ebenfalls wohl bekannt.

**[0090]** Es sei auch darauf hingewiesen, dass Screening-Verfahren in vitro-Verfahren, wie zellfreie Transkrip-

tions- oder Translationsysteme, beinhalten können. In diesen Systemen wird Transkription oder Translation stattfinden gelassen, und ein Mittel wird auf seine Fähigkeit zur Modulierung der Funktion getestet. Für einen Assay, der bestimmt, ob ein Mittel die Translation von mRNA oder einem Polynukleotid moduliert, kann ein in vitro-Transkriptions-/Translationssystem verwendet werden. Diese Systeme sind im Handel erhältlich und liefern ein in vitro-Mittel zur Produktion von mRNA, die einer interessierenden Polynukleotidsequenz entspricht. Nachdem die mRNA hergestellt worden ist, kann sie in vitro Translation unterzogen werden, und die Translationsprodukte können verglichen werden. Der Vergleich der Translationsprodukte zwischen einem in vitro-Expressionssystem, das kein Mittel enthält (negative Kontrolle), mit einem in vitro-Expressionssystem, das ein Mittel enthält, zeigt, ob das Mittel die Translation beeinflusst. Der Vergleich von Translationsprodukten zwischen Kontroll- und Testpolynukleotiden zeigt, ob das Mittel, wenn es auf dieser Ebene wirkt, selektiv Translation bewirkt (im Unterschied zum Bewirken der Translation in einer allgemeinen, unselektiven oder unspezifischen Weise). Die Modulierung der Polypeptidfunktion kann auf vielerlei Weise bewirkt werden, einschließlich, aber nicht begrenzt auf die oben aufgeführten in vivo- und in vitro-Assays sowie die in vitro-Assay unter Verwendung von Proteinpräparationen. Polypeptide können aus natürlichen oder rekombinanten Quellen extrahiert und/oder gereinigt werden, um Proteinpräparationen zu erzeugen. Ein Mittel kann einer Probe einer Proteinpräparation zugefügt und die Wirkung überwacht werden, das bedeutet, ob und wie das Mittel auf ein Polypeptid und seine Konformation, Faltung (oder andere physikalische Charakteristika), Bindung an andere Einheiten (wie Liganden), Aktivität (oder andere funktionale Charakteristika) und/oder andere Aspekte der Proteinstruktur oder -funktionen wirkt, und dies wird als modulierte Polypeptidfunktion aufweisend angesehen.

**[0091]** In einem Beispiel für einen Assay für ein Mittel, das an ein durch ein Polynukleotid kodiertes Polypeptid bindet, welches durch die hier beschriebenen Verfahren identifiziert wird, wird zuerst ein Polypeptid rekombinant in einem prokariotischen oder eukariotischen Expressionssystem als natives oder Fusionsprotein exprimiert, wobei ein Polypeptid (kodiert durch ein wie oben beschrieben identifiziertes Polynukleotid) mit einem gut charakterisierten Epitop oder Protein konjugiert wird. Dann wird das rekombinante Polypeptid durch beispielsweise Immunopräzipitation unter Verwendung geeigneter Antikörper oder Anti-Epitop-Antikörper gereinigt, oder durch Bindung an den immobilisierten Liganden des Konjugats. Eine aus Polypeptid oder Fusionsprotein hergestellte Affinitätssäule wird dann zum Screening einer Mischung von Verbindungen verwendet, die geeignet markiert worden sind. Zu geeigneten Markierungen gehören, ohne darauf begrenzt zu sein, Fluorochrome, Radioisotope, Enzyme und Chemilumineszenzverbindungen. Die ungebundenen und gebundenen Verbindungen können durch Wäschen unter Verwendung von verschiedenen Bedingungen (z. B. hoher Salzgehalt, Detergens) getrennt werden, die von Fachleuten routinemäßig verwendet werden. Die unspezifische Bindung an die Affinitätssäule kann durch Vorreinigung der Verbindungsmischung unter Verwendung einer Affinitätssäule minimiert werden, die lediglich das Konjugat oder das Epitop enthält. Ähnliche Verfahren können zum Screening auf ein oder mehrere Mittel verwendet werden, die um die Bindung an Polypeptide konkurrieren. Zusätzlich zu der Affinitätschromatographie gibt es andere Techniken, wie das Messen der Veränderung der Schmelztemperatur oder der Fluoreszenzanisotropie eines Proteins, die sich nach Binden eines anderen Moleküls verändern wird. Zur Bestimmung der Bindungsaktivität unterschiedlicher Mittel kann beispielsweise ein BIAcore-Assay unter Verwendung eines Sensorchips (angeboten von Pharmacia Biosensor, Stitt et al. (1995) Cell 80: 661-670), der kovalent an Polypeptid gebunden ist, durchgeführt werden.

**[0092]** Es sei auch darauf hingewiesen, dass die in vitro-Screening-Verfahren dieser Erfindung strukturelles oder rationales Drug Design einschließen, wobei die Aminosäuresequenz, dreidimensionale Atomstruktur oder andere Eigenschaft (oder Eigenschaften) eines Polypeptids eine Basis zum Design eines Mittels liefern, von dem erwartet wird, dass es an ein Polypeptid bindet. Allgemein werden das Design und/oder die Wahl der Mittel in diesem Kontext durch mehrere Parameter beherrscht, wie Seite-an-Seite-Vergleich der Strukturen von homologen Polypeptiden eines domestizierten Organismus und des Vorfahren, der erkennbaren Funktion des angestrebten Polypeptid, seiner dreidimensionalen Struktur (falls bekannt oder vermutet) und andere Aspekte des rationalen Drug Design. Es können auch Techniken der kombinatorischen Chemie verwendet werden, um zahlreiche Permutationen von in Frage kommenden Mitteln zu erzeugen.

**[0093]** In den erfindungsgemäßen Screening-Verfahren kommen auch nicht-menschliche, transgene, tierische und pflanzliche Systeme in Frage, die in der Technik bekannt sind.

**[0094]** Die oben beschriebenen Screening-Verfahren sind Primärscreens, die zum Detektieren von jeglichem Mittel vorgesehen sind, das Aktivität aufweisen kann, welche die Funktion eines Polynukleotids oder Polypeptids moduliert. Der Fachmann wird erkennen, dass wahrscheinlich Sekundärscreens erforderlich sind, um ein Mittel weiter zu bewerten. Ein Sekundärscreen kann beispielsweise das Testen des Mittels/der Mittel in einem Infektiositäts-Assay mit Mäusen und anderen Tiermodellen (wie der Ratte) beinhalten, die in der Technik bekannt sind, oder an der domestizierten Pflanze oder dem domestizierten Tier selbst. Es wird zudem ein Zytotoxizi-

täts-Assay als weitere Bestätigung durchgeführt, dass ein Mittel, das in einem Primärscreen positiv getestet wurde, zur Verwendung in lebenden Organismen geeignet wäre. Zu diesem Zweck ist jeder beliebige Zytotoxizitäts-Assay geeignet, einschließlich beispielsweise dem MTT-Assay (Promega).

**[0095]** Die Erfindung schließt auch Mittel ein, die durch die hier beschriebenen Screening-Verfahren identifiziert werden.

**[0096]** Die folgenden Beispiele werden Fachleuten zur näheren Erläuterung gegeben. Derartige Beispiele werden als veranschaulichend angesehen und sollten daher nicht als die Erfindung einschränkend angesehen werden. In dieser Anmeldung sind zahlreiche beispielhafte Modifikationen und Varianten beschrieben, und Fachleuten ergeben sich weitere. Derartige Varianten werden als unter den Schutzzumfang der hier beschriebenen und beanspruchten Erfindung fallend angesehen.

## BEISPIELE

### Beispiel 1: Aufbau der cDNA-Bibliothek

**[0097]** Eine cDNA-Bibliothek einer domestizierten Pflanze oder eines domestizierten Tieres wird mit einem geeigneten Gewebe der Pflanze oder des Tieres aufgebaut. Ein durchschnittlich versierter Fachmann kennt die geeigneten Gewebe, die gemäß der interessierenden Aufgabenstellung zu analysieren sind. Alternativ kann der gesamte Organismus verwendet werden. 1 Tage alte Pflanzensämlinge exprimieren beispielsweise bekanntermaßen die meisten Gene der Pflanze.

**[0098]** Die Gesamt-RNA wird aus dem Gewebe extrahiert (RNeasy kit, Quiagen; RNase-freies Rapid Total RNA Kit, 5 Prime-3 Prime, Inc.) und die Integrität und Reinheit der RNA werden gemäß konventionellen molekularen Klonierungsverfahren bestimmt. Polo A+ RNA wird isoliert (Mini-Oligo(dT) Cellulose Spin-Säulen, 5 Prime-3 Prime, Inc.) und als Templat für die reverse Transkription der cDNA mit Oligo (dT) als Primer verwendet. Die synthetisierte cDNA wird mit im Handel erhältlichen Kits zum Klonieren behandelt und modifiziert. Die Rekombinanten werden danach gepackt und in einer Wirtszelllinie vermehrt. Teile der verpackten Gemische werden amplifiziert, und der Rest vor der Amplifizierung behalten. Die Bibliothek kann normalisiert werden, und die Anzahl der unabhängigen Rekombinanten in der Bibliothek wird bestimmt.

### Beispiel 2: Sequenzvergleich

**[0099]** Basierend auf einem in Frage kommenden Gen des domestizierten Organismus werden geeignete Primer hergestellt und zur PCR-Amplifizierung der Vorfahren-cDNA entweder aus einer cDNA-Bibliothek oder aus einer aus mRNA hergestellten cDNA verwendet. Die selektierten Vorfahren-cDNA-Klone aus der cDNA-Bibliothek werden mit einem automatischen Sequenziergerät, wie einer ABI 377, sequenziert. Zur Durchführung der Sequenzierung werden üblicherweise an dem Kloniervektor verwendete Primer, wie die M13 Universal- und Reverse-Primer, verwendet. Bei Inserts, die durch Endsequenzierung nicht vollständig sequenziert werden, können farbstoffmarkierte Terminatoren oder individuell gefertigte Primer verwendet werden, um die restlichen Lücken zu füllen.

**[0100]** Die detektierten Sequenzdifferenzen werden anfangs auf Genauigkeit geprüft, beispielsweise durch Aufsuchen der Punkte, an denen es Unterschiede zwischen den domestizierten und Vorfahren-Sequenzen gibt; durch Prüfen des Sequenzfluorogramms (Chromatogramm), um zu ermitteln, ob die Basen, die in dem domestizierten Organismus einzigartig zu sein scheinen, starken, klaren Signalen entsprechen, die für die genannte Base spezifisch sind; durch Prüfen der Hits des domestizierten Organismus, um zu sehen, ob es mehr als eine Sequenz gibt, die einer Sequenzänderung entspricht, sowie nach Bedarf nach anderen Verfahren, die in der Technik bekannt sind. Mehrere Einträge der Sequenz des domestizierten Organismus für dasselbe Gen, die dasselbe Nukleotid an einer Position haben, an der es in einer Vorfahrensequenz ein anderes Nukleotid gibt, sprechen unabhängig voneinander dafür, dass die domestizierte Sequenz akkurat ist und die Änderung von domestiziert zu Vorfahren signifikant ist. Derartige Veränderungen werden unter Verwendung von öffentlichen oder kommerziellen Datenbankinformationen und dem genetischen Kode verwendet, um zu bestimmen, ob diese DNA-Sequenzänderungen zu einer Veränderung der Aminosäuresequenz des kodierten Proteins führen. Die Sequenzen können auch durch direktes Sequenzieren des kodierten Proteins untersucht werden.

### Beispiel 3: Molekulare Evolutionsanalyse

**[0101]** Die zu vergleichenden Sequenzen der domestizierten Pflanze oder des domestizierten Tieres und des

wilden Vorfahren werden  $K_A/K_S$ -Analyse unterzogen. In dieser Analyse werden öffentlich zugängliche oder kommerziell erhältliche Computerprogramme, wie Li 93 und INA, zur Bestimmung der Zahl der nicht-synonymen Änderungen pro Stelle ( $K_A$ ), geteilt durch die Zahl der synonymen Änderungen pro Stelle ( $K_S$ ), für jede untersuchte Sequenz wie oben beschrieben bestimmt. Es können Kodierregionen mit vollständiger Länge oder Teilsegmente einer Kodierregion verwendet werden. Je höher das  $K_A/K_S$ -Verhältnis ist, um so größer ist die Wahrscheinlichkeit, dass eine Sequenz adaptive Evolution eingegangen ist. Die statistische Signifikanz von  $K_A/K_S$ -Werten wird mit etablierten statistischen Verfahren und verfügbaren Programmen, wie dem T-Test, bestimmt.

**[0102]** Um die Signifikanz eines hohen  $K_A/K_S$ -Verhältnisses weiter zu stützen, kann die untersuchte domestizierte Sequenz mit anderen evolutionär verwandten Spezies verglichen werden. Diese Vergleiche ermöglichen die weitere Diskriminierung, ob die adaptiven evolutionären Änderungen für die domestizierte Pflanzen- oder Tierlinie einzigartig sind, verglichen mit anderen eng verwandten Spezies. Die Sequenzen können auch durch direkte Sequenzierung des interessierenden Gens aus Vertretern mehrerer diverser domestizierter Populationen untersucht werden, um zu bewerten, bis zu welchem Grad die Sequenz in der domestizierten Pflanze oder dem domestizierten Tier konserviert wird.

#### Beispiel 4: Aufbau der cDNA-Bibliothek

**[0103]** Eine cDNA-Bibliothek der Teosinte wird mit 1 Tag alten vollständigen Teosinte-Sämlingen oder anderen geeigneten Pflanzengewebe aufgebaut. Aus dem Sämlingsgewebe wird die gesamte RNA extrahiert, und die Integrität und Reinheit der RNA werden gemäß konventionellen molekularen Klonierungsverfahren bestimmt. Poly A+ RNA wird gewählt und als Templat für die Reverse-Transkription von cDNA mit Oligo (dT) als Primer verwendet. Die synthetisierte cDNA wird mit im Handel erhältlichen Kits zum Klonieren behandelt und modifiziert. Die Rekombinanten werden danach gepackt und in einer Wirtszelllinie vermehrt. Teile der verpackten Gemische werden amplifiziert, und der Rest vor der Amplifizierung behalten. Rekombinante DNA wird nach etablierten Verfahren zum Transfektizieren von *E. coli*-Wirtszellen verwendet. Die Bibliothek kann normalisiert werden, und die Anzahl der unabhängigen Rekombinanten in der Bibliothek wird bestimmt.

#### Beispiel 5: Sequenzvergleich

**[0104]** Statistisch ausgewählte Teosinte-cDNA-Klone aus der cDNA-Bibliothek wurden mit einem automatischen Sequenziergerät, wie einer ABI 377, sequenziert. Zur Durchführung der Sequenzierung werden üblicherweise an dem Kloniervektor verwendete Primer, wie die M13 Universal- und Reverse-Primer, verwendet. Bei Inserts, die durch Endsequenzierung nicht vollständig sequenziert werden, wurden farbstoffmarkierte Terminatoren verwendet werden, um die restlichen Lücken zu füllen.

**[0105]** Die resultierenden Teosintesequenzen wurden mittels Datenbank-Recherchen mit den Sequenzen von domestiziertem Mais verglichen. Genomdatenbanken für zahlreiche Spezies einschließlich Mais sind öffentlich oder kommerziell zugänglich. Ein Beispiel für eine Mais-Datenbank findet sich auf [www.central.edu/homepages/liedlb/genetics/gene-site.html](http://www.central.edu/homepages/liedlb/genetics/gene-site.html). Andere geeignete Mais-EST (exprimiertes Sequenz-Tag)-Datenbanken sind in Privatbesitz und werden privat gepflegt. Die hochrangigen "Hits", d. h. Sequenzen, die nach Homologieanalysen eine signifikante (z. B. > 80%) Ähnlichkeit zeigen, werden abgerufen und analysiert. Die beiden homologen Sequenzen werden danach unter Verwendung des von Higgins et al. entwickelten Alignment-Programms CLUSTAL V Alignment unterzogen. Durch das Alignment kann jegliche Sequenzdivergenz einschließlich Nukleotidsubstitution, Insertierung und Deletion detektiert und aufgezeichnet werden.

**[0106]** Die detektierten Sequenzdifferenzen werden anfangs auf Genauigkeit geprüft, beispielsweise durch Aufsuchen der Punkte, an denen es Unterschiede zwischen den Teosinte- und Maissequenzen gibt; durch Prüfen des Sequenzfluorogramms (Chromatogramm), um zu ermitteln, ob die Basen, die in Mais einzigartig zu sein scheinen, starken, klaren Signalen entsprechen, die für die genannte Base spezifisch sind; durch Prüfen der Hits des Mais, um zu sehen, ob es mehr als eine Maissequenz gibt, die einer Sequenzänderung entspricht, sowie nach Bedarf nach anderen Verfahren, die in der Technik bekannt sind. Mehrere Einträge der Maissequenz für dasselbe Gen, die dasselbe Nukleotid an einer Position haben, an der es in einer Vorfahrensequenz ein anderes Nukleotid gibt, sprechen unabhängig voneinander dafür, dass die Maissequenz akkurat ist und die Änderung von Teosinte zu Mais real ist. Derartige Veränderungen werden unter Verwendung von öffentlichen oder kommerziellen Datenbankinformationen und dem genetischen Kode untersucht, um zu bestimmen, ob diese DNA-Sequenzänderungen zu einer Veränderung der Aminosäuresequenz des kodierten Proteins führen. Die Sequenzen können auch durch direktes Sequenzieren des kodierten Proteins untersucht werden.

## Beispiel 6: Molekulare Evolutionsanalyse

**[0107]** Die zu vergleichenden Teosinte- und Maissequenzen wurden  $K_A/K_S$ -Analyse unterzogen. In dieser Analyse werden öffentlich zugängliche oder kommerziell erhältliche Computerprogramme, wie Li 93 und INA, zur Bestimmung der Zahl der nicht-synonymen Änderungen pro Stelle ( $K_A$ ), geteilt durch die Zahl der synonymen Änderungen pro Stelle ( $K_S$ ), für jede untersuchte Sequenz wie oben beschrieben bestimmt. Es ist gezeigt worden, dass dieses Verhältnis,  $K_A/K_S$ , den Grad widerspiegelt, bis zu dem die adaptive Evolution, d. h. positive Selektion, in der untersuchten Sequenz gewirkt hat. Typischerweise sind Kodierregionen mit vollständiger Länge in diesen Vergleichsanalysen verwendet worden. Es können jedoch auch Teilsegmente einer Kodierregion effektiv verwendet werden. Je höher das  $K_A/K_S$ -Verhältnis ist, um so größer ist die Wahrscheinlichkeit, dass eine Sequenz adaptive Evolution eingegangen ist. Die statistische Signifikanz von  $K_A/K_S$ -Werten wird mit etablierten statistischen Verfahren und verfügbaren Programmen, wie dem T-Test, bestimmt. Jene Gene, die statistisch hohe  $K_A/K_S$ -Verhältnisse zwischen Teosinte- und Maisgenen zeigen, sind sehr wahrscheinlich adaptive Evolution eingegangen.

**[0108]** Um die Signifikanz eines hohen  $K_A/K_S$ -Verhältnisses weiter zu stützen, kann die untersuchte Sequenz mit anderen Vorfahren-Maisspezies verglichen werden. Diese Vergleiche ermöglichen die weitere Diskriminierung, ob die adaptiven evolutionären Änderungen für die domestizierte Maislinie einzigartig sind, verglichen mit anderen Vorfahren. Die Sequenzen können auch durch direkte Sequenzierung des interessierenden Gens aus Vertretern mehrerer diverser Maispopulationen untersucht werden, um zu bewerten, bis zu welchem Grad die Sequenz in den Maisspezies konserviert wird.

Beispiel 7: Anwendung des  $K_A/K_S$ -Verfahrens auf homologe Sequenzen von Mais und Teosinte, die aus einer Datenbank erhalten wurden

**[0109]** Der Vergleich der Sequenzen von domestiziertem Mais und Teosinte, die auf Genbank erhältlich waren ([www.ncbi.nlm.gov/Web/Genbank/index.html](http://www.ncbi.nlm.gov/Web/Genbank/index.html)), zeigte mindestens vier homologe Gene: waxy, A1\*, A1 und globulin. Alle verfügbaren Sequenzen für diese Gene für sowohl Mais als auch Teosinte wurden verglichen. Die  $K_A/K_S$ -Verhältnisse wurden mit Li93 und/oder INA verglichen:

| Gen            | durchschnittliche Zahl der synonymen Substitutionen | durchschnittliche Zahl der nicht-synonymen Substitutionen | $K_A/K_S$ |
|----------------|---|---|-----------|
| Waxy           | 4   | 1   | 0,068     |
| A1*            | 10  | 3   | 0,011     |
| A <sub>1</sub> | 3   | 2   | 0,44–0,89 |
| Globulin       | 10  | 7   | 0,42      |

**[0110]** Obwohl man erwartet hatte, dass der Polymorphismus (mehrere allele Kopien) und/oder die Polyploidität (mehr als zwei Chromosomensätze pro Zelle), die bei Mais beobachtet wurden, die  $K_A/K_S$ -Analyse komplex oder schwierig machen würden, wurde gefunden, dass dies nicht der Fall war.

**[0111]** Während die obigen  $K_A/K_S$ -Werte zeigten, dass diese Gene nicht positiv selektiert wurden, illustriert dieses Beispiel, dass das  $K_A/K_S$ -Verfahren für Mais und seine Teosinte-Sequenzen verwendet werden kann, die aus einer Datenbank erhalten wurden.

Beispiel 8: Untersuchung der Proteinfunktion unter Verwendung einer transgenen Pflanze

**[0112]** Die funktionalen Rollen eines positiv selektierten Maisgens, das nach den Verfahren von Beispielen 4 bis 7 erhalten wurde, können bewertet werden, indem Bewertungen jedes Allelen des Gens in einer transgenen Maispflanze durchgeführt werden. Eine transgene Pflanze kann durch eine Anpassung des in Peng et al. (1999) Nature 400:256-261 beschriebenen Verfahrens erzeugt werden. Physiologische, morphologische und/oder biochemische Untersuchung der transgenen Pflanze oder ihrer Proteinextrakte ermöglichen die Assoziation jedes Allelen mit einem speziellen Phänotyp.

Beispiel 9: Kartieren von positiv selektierten Genen auf QTLs

**[0113]** QTL (Quantitative trait locus (Eigenschafts-Ort))-Analyse hat Chromosomenregionen definiert, welche die Gene enthalten, die mehrere interessierende phänotypische Eigenschaften in Mais kontrollieren, einschließlich Pflanzenhöhe und Ölgehalt. Durch Kartieren jedes nach diesem Verfahren identifizierten positiv se-

lektierten Gens auf einen der bekannten QTLs kann die spezielle Eigenschaft, die durch jedes positiv selektierte Gen kontrolliert wird, rasch und eindeutig identifiziert werden.

**[0114]** Obwohl die vorhergehende Erfindung in einigen Details zur Deutlichkeit und zum Verständnis veranschaulicht und beispielhaft beschrieben wurde, wird Fachleuten klar werden, dass bestimmte Änderungen und Modifikationen durchgeführt werden können. Daher sollen die Beschreibung und die Beispiele nicht als den Schutzzumfang der Erfindung einschränkend angesehen werden, welcher durch die angefügten Ansprüche gegeben ist.

### Patentansprüche

1. Verfahren zum Identifizieren einer Polynukleotidsequenz, die ein Polypeptid eines domestizierten Organismus kodiert, wobei das Polypeptid mit einer kommerziell oder ästhetisch relevanten Eigenschaft assoziiert sein kann, die in dem domestizierten Organismus, verglichen mit einem wilden Vorfahren des domestizierten Organismus, einzigartig, verbessert oder verändert ist, bei dem

(a) man Protein kodierende Nukleotidsequenzen des domestizierten Organismus mit Protein kodierenden Nukleotidsequenzen des wilden Vorfahren vergleicht, und

(b) man eine Protein kodierende Polynukleotidsequenz in dem domestizierten Organismus auswählt, die eine Nukleotidveränderung enthält, die verglichen mit einer entsprechenden Sequenz in dem wilden Vorfahren evolutionär signifikant ist, definiert durch ein Verhältnis ( $Ka/Ks$ ) der nicht-synonymen Substitutionsrate ( $Ka$ ) zu der synonymen Substitutionsrate ( $Ks$ ) in der partiellen oder vollständigen Protein kodierenden Polynukleotidsequenz von  $\geq 0,75$ .

2. Verfahren nach Anspruch 1, bei dem man weiter die gewählte Sequenz hinsichtlich von molekularen/genetischen Identitäten und biologischer Funktion charakterisiert.

3. Verfahren nach Anspruch 1 oder 2, bei dem der domestizierte Organismus eine domestizierte Pflanze ist, deren Vorfahre bekannt ist.

4. Verfahren nach Anspruch 3, bei dem die Eigenschaft Ertrag, Kurztagblühen (short day length flowering), Protein- oder Ölgehalt, Geschmack, Leichtigkeit der Ernte oder Krankheits- oder Dürresistenz ist.

5. Verfahren nach Anspruch 3 oder 4, bei dem die domestizierte Pflanze Mais, Reis, Tomate oder Kartoffel ist.

6. Verfahren nach Anspruch 6, bei dem die domestizierte Pflanze Mais ist und der wilde Vorfahre Teosinte ist.

7. Verfahren nach Anspruch 1 oder 2, bei dem der domestizierte Organismus ein domestiziertes Tier ist, dessen Vorfahre bekannt ist.

8. Verfahren nach Anspruch 7, bei dem die Eigenschaft Fett- oder Proteingehalt, Milchproduktion, Zeit bis zur Reife, Fügbarkeit, Fruchtbarkeit oder Krankheitsresistenz ist.

9. Verfahren nach Anspruch 7 oder 8, bei dem das domestizierte Tier ein Schwein, Rind (Vieh), Pferd, Hund oder eine Katze ist.

10. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die verglichenen Sequenzen cDNA sind.

11. Verfahren zur Identifizierung eines Mittels, welches die Eigenschaft von Anspruch 1 modulieren kann, bei dem man mindestens ein in Frage kommendes Mittel mit einer Zelle oder einem nicht menschlichen transgenen Organismus kontaktiert, die bzw. der die in Anspruch 1 identifizierte Polynukleotidsequenz exprimiert, wobei das Mittel durch seine Fähigkeit identifiziert wird, Funktionen des Polypeptids zu modulieren, das durch das in Anspruch 1 identifizierte Polynukleotid kodiert wird.

12. Verfahren zum Korrelieren einer evolutionär signifikanten Nukleotidveränderung mit der Eigenschaft nach Anspruch 1, bei dem

(a) man eine Nukleotidsequenz nach Anspruch 1 identifiziert und

(b) man die funktionale Wirkung der Anwesenheit oder Abwesenheit der identifizierten Sequenz in dem domes-

tizierten Organismus analysiert.

Es folgt kein Blatt Zeichnungen