



(19)
 Bundesrepublik Deutschland
 Deutsches Patent- und Markenamt

(10) **DE 10 2004 016 437 A1** 2005.10.20

(12)

Offenlegungsschrift

(21) Aktenzeichen: **10 2004 016 437.1**

(22) Anmeldetag: **04.04.2004**

(43) Offenlegungstag: **20.10.2005**

(51) Int Cl.⁷: **C12Q 1/68**

(71) Anmelder:

oligene GmbH, 10117 Berlin, DE

(74) Vertreter:

BOEHMERT & BOEHMERT, 28209 Bremen

(72) Erfinder:

Häupl, Thomas, 15537 Erkner, DE; Grün, Joachim, 14057 Berlin, DE; Radbruch, Andreas, 10117 Berlin, DE; Burmester, Gerd-Rüdiger, 10117 Berlin, DE; Kaps, Christian, 10117 Berlin, DE; Grützkau, Andreas, 10117 Berlin, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:

WO 04/0 05 882 A2

WO 03/0 83 141 A1

WO 03/0 72 065 A2

VENET, D. u.a.: Separation of samples into their constituents using gene expression data.

Bioinformatics (2001) 17 (Suppl. 1) 279-87;

TÜRECI, Ö. u.a.: Computational dissection of tissue contamination for identification of colon cancer-specific expression profiles. FASEB J. (2003) 17 (3) 376-85;

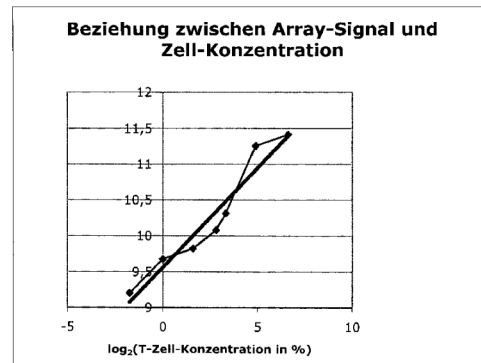
KIM, S. u.a.: Identification of combination gene sets for glioma classification. Mol. Cancer Ther. (2002) 1 (13) 1229-36;

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

Prüfungsantrag gemäß § 44 PatG ist gestellt.

(54) Bezeichnung: **Verfahren zur Erkennung von Signaturen in komplexen Genexpressionsprofilen**

(57) Zusammenfassung: Die vorliegende Erfindung betrifft ein Verfahren zur Erkennung von Signaturen in komplexen Genexpressionsprofilen, das die Schritte umfaßt von: a) zur Verfügung stellen einer zu untersuchenden biologischen Probe, b) zur Verfügung stellen mindestens eines geeigneten Expressionsprofils, wobei das mindestens eine Expressionsprofil einen oder mehrere Marker umfaßt, die ausschließlich für das Expressionsprofil typisch sind, c) Bestimmen des komplexen Expressionsprofils der biologischen Probe, d) Bestimmen der quantitativen zellulären Zusammensetzung der biologischen Probe mittels der in den Schritten b) und c) bestimmten Expressionsprofile. Weiterhin kann das erfindungsgemäße Verfahren die Schritte von e) Berechnung eines virtuellen Signals, das aufgrund der bestimmten Zusammensetzung der Expressionsprofile erwartet wird, f) Berechnung der Differenz aus dem tatsächlichen gemessenen komplexen Expressionsprofil und dem virtuellen Signal und g) Bestimmen der quantitativen Zusammensetzung des komplexen Expressionsprofils auf Basis der ermittelten Differenzen umfassen. Die vorliegende Erfindung betrifft weiterhin die Anwendung des erfindungsgemäßen Verfahrens in der Diagnose, Prognose und/oder Verfolgung einer Erkrankung. Schließlich werden entsprechende Computersysteme, Computerprogramme, computerlesbare Datenträgermedien und Laborroboter oder Auswertegeräte für molekulare Nachweismethoden offenbart.



Beschreibung

[0001] Die vorliegende Erfindung betrifft ein Verfahren zur Erkennung von Signaturen in komplexen Genexpressionsprofilen, das die Schritte umfaßt von: a) zur Verfügung stellen einer zu untersuchenden biologischen Probe, b) zur Verfügung stellen mindestens eines geeigneten Expressionsprofils, wobei das mindestens eine Expressionsprofil einen oder mehrere Marker umfaßt, die ausschließlich für das Expressionsprofil typisch sind, c) Bestimmen des komplexen Expressionsprofils der biologischen Probe, d) Bestimmen der quantitativen zellulären Zusammensetzung der biologischen Probe mittels der in den Schritten b) und c) bestimmten Expressionsprofile, e) Berechnung eines virtuellen Signals, das aufgrund der bestimmten Zusammensetzung der Expressionsprofile erwartet wird, f) Berechnung der Differenz aus dem tatsächlichem gemessenem komplexen Expressionsprofil und dem virtuellen Signal, und g) Bestimmen der quantitativen Zusammensetzung des komplexen Expressionsprofils auf Basis der ermittelten Differenzen. Die vorliegende Erfindung betrifft weiterhin die Anwendung des erfindungsgemäßen Verfahrens in der Diagnose, Prognose und/oder Verfolgung einer Erkrankung. Schließlich werden entsprechende Computersysteme, Computerprogramme, computerlesbare Datenträgermedien und Laborroboter oder Auswertegeräte für molekulare Nachweismethoden offenbart.

Stand der Technik

Einleitung

[0002] Die Expression von bestimmten Genen zu bestimmten Zeiten im Lebenszyklus der Zelle bestimmt letztendlich ihren Phänotyp. Die Analyse der Genexpression insbesondere bei der Diagnose und Behandlung ist von besonderer Bedeutung bei erkrankten und/oder entarteten Zellen und letztendlich Geweben, die besondere insbesondere komplexe, d.h. unbekannte Gemische an Expressionsprofilen verschiedener Zelltypen aufweisen können.

[0003] Die im Stand der Technik bekannten Hochdurchsatz-Verfahren wie die DNA- und Protein-Array Technologie, die Massenspektrometrie oder Verfahren zu epigenetischen Untersuchungen erlauben quantitative Bestimmung von komplexen molekularen Profilen. Mit DNA-Array Untersuchungen wird z.B. die Aktivität von Genen über die Expression der mRNA gemessen.

[0004] Auch die Proteinexpression wird zunehmend im Hochdurchsatzverfahren verfügbar über entsprechende Array-Technologien oder die Massenspektrometrie. Epigenetische Analysen erheben Profile zum DNA-Methylierungszustand von Genen und erlauben Rückschlüsse auf eine Inaktivierung bzw. die Aktivierbarkeit von Genen. Diese Methoden lassen weit reichende Entwicklungen für die molekulare Diagnostik erwarten. Es besteht die Hoffnung, daß verschiedene molekulare Profile mit besonderen klinischen Merkmalen assoziiert sein können, durch molekulare Merkmale Krankheiten in Untergruppen einteilbar werden und Interpretationsmöglichkeiten entwickelt werden können, die für Therapie und Krankheitsverlauf prognostische Informationen liefern. Ferner könnten aus den molekularen Profilen bzw. deren Interpretation auf Einzelfaktorebene Pathomechanismen ableitbar werden, die gezielte therapeutische Beeinflussung ermöglichen.

[0005] Die zu untersuchenden Proben tragen viele verschiedene molekulare Informationen. Zahlreiche Gene können bei einer veränderten Expression sowohl mit einer Verschiebung der zellulären Zusammensetzung der Probe (Einwanderung von Zellen) als auch einer Aktivierung einzelner oder mehrerer Stoffwechselwege verbunden sein.

[0006] Beide Informationen bilden sich im Expressionsmuster oder auch -Profil überlappend ab. Derzeitige bioinformatische Analysemethoden erlauben keine Unterscheidung zwischen diesen beiden Ursachen. Die Interpretation der Array Daten ist somit stark eingeschränkt. Um die Genregulationen in Zellpopulationen zu erkennen, ist heute vor der Array-Analyse eine Aufreinigung der Zellen erforderlich oder eine histologische Untersuchung von Geweben mit immunhistologischer Zuordnung zu Zelltypen. Zellaufreinigungen können aber zu artifiziellen Veränderungen der Genexpressionsmuster führen und histologische Möglichkeiten sind auf wenige Gene beschränkt.

[0007] Die nachteilige Bedeutung dieser Vermischung von Ursachen und Wirkungen wird um so eindrücklicher, als regulierte Gene normalerweise keiner Ein-/Aus-Aktivität unterliegen, sondern meist eine Grundaktivität (konstitutive Expression) aufweisen. Ferner können sie in verschiedenen Zelltypen und auch Stoffwechselwegen unterschiedlich aktiv sein.

[0008] Es fällt somit der Großteil der differentiell exprimierten Gene in diese ursächlich nicht eindeutig zuor-

denbare Gruppe. Es sind derzeit also für die meisten Gene weiterführende Untersuchungen erforderlich, um zu klären, ob eine Verschiebung in der Zellzusammensetzung oder eine Genregulation aufgetreten ist.

[0009] Haviv et al (Haviv I, Campbell IG. DNA microarrays for assessing ovarian cancer gene expression. *Mol Cell Endocrinol.* 2002 May 31;191(1):121-6.) beschreiben die gleichzeitige Expressionsanalyse von Genen innerhalb einer gegebenen Population mittels Array-Technologien. Danach kann die Expression von normalen und malignen Zellen verglichen werden und Gene identifiziert werden, die unterschiedlich reguliert werden. Vallat et al (Vallat L, Magdelenat H, Merle-Beral H, Masdehors P, Potocki de Montalk G, Davi F, Kruhoff M, Sabatier L, Orntoft TF, Delic J. The resistance of B-CLL cells to DNA damage-induced apoptosis defined by DNA microarrays. *Blood.* 2003 Jun 1;101(11):4598-606. Epub 2003 Feb 13.) beschreiben den Vergleich von getrennten B-Zell chronischer lymphoider Leukämie (BCLL)-Zellproben. Dabei werden 16 unterschiedlich exprimierte Gene identifiziert, so unter anderem Nuclear orphan receptor TR3, major histocompatibility complex (MHC) Class II glycoprotein HLA-DQA1, mtmr6, c-myc, c-rel, c-IAP1, mat2A und fmod, MIP1a/GOS19-1 homolog, stat1, blk, hsp27, und ech1.

[0010] Vasselli et al (Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, Torres-Cabala C, Tabios R, Mariotti A, Stearman R, Merino M, Walther MM, Simon R, Klausner RD, Linehan WM. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proc Natl Acad Sci U S A.* 2003 Jun 10;100(12):6958-63. Epub 2003 May 30.) beschreiben die Analyse von verschiedenen Geweben auf der Suche nach potentiellen molekularen Determinanten der Tumorbiologie und möglichem klinischen Ausgang in Nierenkrebs. Suzuki et al (Suzuki S, Asamoto M, Tsujimura K, Shirai T. Specific differences in gene expression profile revealed by cDNA microarray analysis of glutathione S-transferase placental form (GST-P) immunohistochemically positive rat liver foci and surrounding tissue. *Carcinogenesis.* 2004 Mar;25(3):439-43. Epub 2003 Dec 04.) beschreiben das Genexpressionsprofil in GST-P positiven Foci im Vergleich mit der Umgebung des Tumors. Die GST-P positiven Foci wurden durch Laser ausgeschnitten und mittels cDNA Microarrayassays getestet.

[0011] Favier et al (Favier J, Plouin PF, Corvol P, Gasc JM. Angiogenesis and vascular architecture in pheochromocytomas: distinctive traits in malignant tumors. *Am J Pathol.* 2002 Oct;161(4):1235-46.) beschreiben die Untersuchung von Genexpressionsprofilen im Rahmen der Angiogenese in Tumoren.

[0012] Pession et al (Pession A, Libri V, Sartini R, Conforti R, Magrini E, Bernardi L, Fronza R, Olivotto E, Prete A, Tonelli R, Paolucci G. Real-time RT-PCR of tyrosine hydroxylase to detect bone marrow involvement in advanced neuroblastoma. *Oncol Rep.* 2003 Mar-Apr;10(2):357-62.) beschreiben TH mRNA Expression als eine spezifischen Tumormarker und dessen Analyse in verschiedenen Geweben.

[0013] Sabek et al (Sabek O, Dorak MT, Kotb M, Gaber AO, Gaber L. Quantitative detection of T-cell activation markers by real-time PCR in renal transplant rejection and correlation with histopathologic evaluation. *Transplantation.* 2002 Sep 15;74(5):701-7.) beschreiben ein ein-Schritt RT-PCR Verfahren im Rahmen der Abstoßung von Transplantaten, die mit T-Zellmarkern einhergehen, z.B. Granzym B und Perforin

[0014] Schließlich beschreiben Hoffmann et al (Hoffmann R, Seidl T, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.* 2002 Jun 14;3(7):RESEARCH0033.) die Normalisierung von Arraysignalen mittels drei verschiedener statistischer Algorithmen zum Nachweis verschieden exprimierter Gene.

[0015] Ähnliche Analysen sind z. B. in Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl.* 2001;Suppl 37:120-5; 3: Dozmorov I, Centola M. An associative analysis of gene expression array data. *Bioinformatics.* 2003 Jan 22;19(2):204-11; Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild HH, Nielsen C, Brunak S, Knudsen S. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* 2002 Aug 30;3(9):research0048; Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics.* 2003 Feb 12;19(3):368-75; Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics.* 2002 Nov;18(11):1454-61 und Park PJ, Pagano M, Bonetti M. A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac Symp Biocomput.* 2001;;52-63 beschrieben.

[0016] Die molekularen Profile bilden verschiedene Veränderungen ab, die sich in den individuellen Messpunkten (das heißt einer spezifischen mRNA, einem Protein, einem Metaboliten, der Methylierung einer spe-

zifischen DNA-Sequenz) häufig überlagern und deshalb aus dem Gesamtwert eines Messpunktes nicht als Teilkomponenten erkennbar sind.

[0017] Am Beispiel der DNA-Array-Analyse soll dies veranschaulicht werden. Veränderungen im Genexpressionsprofil können durch Verschiebungen der zellulären Zusammensetzung der Probe (Einwanderung von Zellen) als auch Aktivierungen einzelner oder mehrerer Gene verursacht sein. Es treten z.B. Veränderungen in der zellulären Zusammensetzung bei jeder Entzündung auf und sind deshalb nicht spezifisch für eine bestimmte Erkrankung. Dagegen können Aktivierungen einzelner oder mehrerer Gene typisch oder sogar spezifisch für einen bestimmten Krankheitsprozeß sein. Beide Veränderungen, die der zellulären Zusammensetzung und die der Regulationen von Gene, bilden sich in der Hybridisierung jedoch miteinander ab, ohne daß derzeitige bioinformatische Analysemethoden eine Zuordnung zu den beiden möglichen Ursachen erlauben. Die Interpretation der Array Daten ist somit stark eingeschränkt.

[0018] Vergleichbare zur Genexpression treten diese Probleme auch bei der Abbildung von Proteinexpressionsmustern auf. Werden ganze Gewebe untersucht, überlappen sich Veränderungen in der zellulären Zusammensetzung mit Veränderungen in der der Proteinexpression einzelner Zelltypen. Vergleichbar kann die Bestimmung von DNA-Methylierungszuständen, die sich zwischen verschiedenen Zelltypen unterscheiden, bei variabler zelluläre Zusammensetzung unterschiedliche Ergebnisse liefern und eine krankheitsspezifische Veränderung in einem einzelnen Zelltyp verschleiern. Wird hingegen Serum oder eine andere Körperflüssigkeit untersucht, so können Veränderungen, die durch eine bestimmte Erkrankung ausgelöst werden, von anderen Einflüssen wie einer diabetischen Stoffwechsellage, einer Niereninsuffizienz oder einer bestimmten Therapie überlagert werden und eine Beurteilung erschweren oder gar unmöglich machen.

[0019] Um Genregulationen in Zellpopulationen zu erkennen, ist heute vor der Array-Analyse eine Aufreinigung der Zellen erforderlich oder eine histologische Untersuchung von Geweben mit immunhistologischer Zuordnung von Genen zu Zelltypen. Zellaufreinigungen können zu artifiziellen Veränderungen der Genexpressionsmuster führen und histologische Möglichkeiten sind auf wenige Gene beschränkt. Ferner sind Aufreinigungsschritte mit einem größeren technischen und damit auch Kostenaufwand verbunden. Hauptziel für eine Routineanwendung ist die Untersuchung von möglichst einfach zugängigen Proben und möglichst unkomplizierter Weiterverarbeitung. Die größte Attraktivität für eine Routineanwendung besitzt hierfür das Blut. Insbesondere das Blut unterliegt bei vielen Erkrankungen zum Teil erheblichen Schwankungen in der zellulären Zusammensetzung und erschwert deshalb die Interpretation von komplexen molekularen Profilen aus dieser Probenart.

[0020] Die Bedeutung dieser Vermischung von Ursachen und Wirkungen wird in **Fig. 5** dargestellt. Dies wird um so deutlicher, als die meisten regulierten Gene keiner Ein-/Aus-Aktivität unterliegen sondern meist eine Grundaktivität aufweisen. Ferner können sie nicht nur in einem sondern in verschiedenen Zelltypen und auch Stoffwechselwegen unterschiedlich aktiv sein. Es fällt somit der Großteil der differentiell exprimierten Gene in diese ursächlich nicht eindeutig zuordenbare Gruppe. Es sind derzeit also für die meisten Gene weiterführende Untersuchungen erforderlich, um zu klären, ob eine Verschiebung in der Zellzusammensetzung oder eine Genregulation aufgetreten ist.

[0021] Prinzipiell ist dieses Problem genereller Natur und trifft auch für Profile der Proteinexpression und Proteinmodifikation oder epigenetische Profile (d.h. unterschiedliche Methylierungsprofile der DNA aus verschiedenen Zelltypen oder komplexen Proben) zu.

Aufgabenstellung

[0022] Es ist somit eine Aufgabe der vorliegenden Erfindung, ein verbessertes Verfahren zur Verfügung zu stellen, das zur Aufschlüsselung der genannten komplexen Informationen z.B. aus Array-Analysen verwendet werden kann. Das Verfahren soll die schnelle und in der Hochdurchsatz-Technik anwendbare Analyse von komplexen Expressionsprofilen ermöglichen, ohne das besondere Aufreinigungsschritte erforderlich sind. Es ist eine weitere Aufgabe der vorliegenden Erfindung, ein für das erfindungsgemäß Verfahren geeignetes bioinformatisches Rechenprogramm zur Verfügung zu stellen. Zuletzt sollen geeignete verbesserte Vorrichtungen zur Verfügung gestellt werden.

[0023] Eine dieser Aufgaben wird erfindungsgemäß durch ein Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe gelöst, wobei das Verfahren die Schritte umfaßt von

- a) zur Verfügung stellen einer zu untersuchenden biologischen Probe,

- b) zur Verfügung stellen mindestens eines geeigneten Expressionsprofils, wobei das mindestens eine Expressionsprofil einen oder mehrere Marker umfaßt, die ausschließlich für das Expressionsprofil typisch sind,
- c) Bestimmen des komplexen Expressionsprofils der biologischen Probe, und
- d) Bestimmen der quantitativen zellulären Zusammensetzung der biologischen Probe mittels der in den Schritten b) und c) bestimmten Expressionsprofile.

[0024] In einer bevorzugten Ausführungsform umfaßt das erfindungsgemäße Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe die weiteren Schritte von e) Berechnung eines virtuellen Signals, das aufgrund der bestimmten Zusammensetzung der Expressionsprofile erwartet wird,
f) Berechnung der Differenz aus dem tatsächlichem gemessenem komplexen Expressionsprofil und dem virtuellen Signal, und
g) Bestimmen der quantitativen Zusammensetzung des komplexen Expressionsprofils auf Basis der ermittelten Differenzen.

[0025] Die vorliegende Erfindung zeigt hier ein Verfahren auf, das zur Aufschlüsselung komplexer Informationen aus Array-Analysen beiträgt. Dieses Verfahren gliedert sich erfindungsgemäß in mehrere Schritte.

[0026] Es werden zunächst folgende Profile für die Auftrennung der Einflüsse benötigt:
a) ein Expressionsprofil, das zum Beispiel den Normalzustand darstellt,
b) weitere definierte oder spezifische Expressionsprofile, die z.B. definierte Einflüsse oder Zustände einer Zelle oder Zellpopulation charakterisieren, und
c) das zu untersuchende komplexe Expressionsprofil der biologischen Probe, zum Beispiel des Erkrankungszustandes.

[0027] Die typischen „Expressionsprofile“ oder „Profile“ von definierten Einflüssen und/oder Zuständen werden nachfolgend auch „Signaturen“ oder „Fingerprints“ genannt. Für die Erkennung der Zellzusammensetzung sind Signaturen für die verschiedenen Zelltypen erforderlich, z.B. für Monozyten, für T-Zellen, für Granulozyten usw. Vergleichbar dazu kann auch eine sogenannte „funktionelle“ und/oder „charakterisierende“ Signatur, wie sie durch eine bestimmte Zytokineinwirkung entsteht, eine Signatur im Sinne der vorliegenden Erfindung darstellen.

[0028] Für jeden Einfluß, der erkannt und von anderen molekularen Informationen abgetrennt werden soll, müssen Markergene definiert werden. Diese lassen den Anteil einer Signatur am Gesamtprofil quantitativ abschätzen. Für die Erkennung unterschiedlicher zellulärer Zusammensetzungen werden also z.B. Markergene für Monozyten, T-Zellen oder Granulozyten identifiziert. Diese spiegeln den Anteil der jeweiligen Zellpopulation in einer gemischten Probe wieder. Für die zelluläre Zusammensetzung einer Probe könnten alternativ auch andere Meßverfahren, wie z. B. das Differentialblutbild oder eine FACS-Analyse verwendet werden.

[0029] Allerdings können zwischen molekular charakterisiertem Anteil und mit anderen Methoden gemessenem Anteil unterschiedliche Relationen auftreten, die nachfolgend zu einer fehlerhaften Berechnung führen können. Es ist daher anzustreben, daß die Grundlagen für die nachfolgenden Berechnung dem gleichen Meßverfahren entstammen.

[0030] Mit Hilfe der molekularen Signaturen von Zellpopulationen (bzw. Einflüsse) und deren quantitativen Beteiligung am Gesamtprofil kann ein virtuelles Signal berechnet werden, das aufgrund der Zusammensetzung erwartet wird. Die Differenz aus tatsächlichem gemessenem Signal und dem erwartenden Signal läßt erkennen, ob sich die Unterschiede lediglich durch die Mischung der verschiedenen Populationen (Einflüsse) erklären (keine Differenz) oder eine Aktivierung (positive Differenz) bzw. eine Unterdrückung (negative Differenz) der Genaktivität stattgefunden hat. Angewandt auf alle mit dem Array gemessenen Gene können die Profile virtuell in Teilkomponenten zerlegt werden.

[0031] Über Unterschiede in der Verteilung der verschiedenen Komponenten ist zu erwarten, daß Kriterien für eine Einteilung in unterschiedliche Gruppen definiert werden können. Gene, deren Expressionsverhalten keiner bekannten Teilkomponente zugeführt werden können, sind von besonderem Interesse für die weitere Abklärung und Suche nach noch unbekanntem Teilkomponenten.

[0032] Bevorzugt ist ein erfindungsgemäßes Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei das Bestimmen des ge-

eigneten Expressionsprofils das Bestimmen eines RNA-Expressionsprofils, Protein-Expressionsprofils, -Sekretionsprofils, DNA-Methylierungsprofils und/oder Metabolitenprofil umfaßt. Natürlich können auch Kombinationen davon bestimmt werden, was die Auswertung jedoch erschwert.

[0033] Weiter bevorzugt ist ein erfindungsgemäßes Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei das Bestimmen eines Expressionsprofils eine molekulare Nachweismethode, wie z. B. ein Genarray, Proteinarray, Peptidarray und/oder PCR-Array oder die Erstellung eines Differentialblutbilds oder eine FACS-Analyse umfaßt. Die vorliegende Erfindung ist somit nicht nur auf Nukleinsäure-Arrays beschränkt. Zudem können auch Expressionsprofile aus Gelanalysen (z. B. 2D), Massenspektrometrie und/oder enzymatischem Verdau (Nuklease- oder Protease-Pattern) verwendet werden.

[0034] Noch weiter bevorzugt ist ein erfindungsgemäßes Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei die oben in Schritt b) des Verfahrens bestimmten Expressionsprofile ausgewählt sind aus der Gruppe von Expressionsprofilen, die funktionelle Einflüsse oder Zustände charakterisieren, wie z.B. Expressionsprofile, die die Aktivität von bestimmten Botenstoffen, der Signaltransduktion oder der Genregulation charakterisieren. Weiterhin können diese die Ausprägung bestimmter molekularer Vorgänge charakterisieren, wie z.B. der Apoptose, Zellteilung, Zelldifferenzierung, Gewebeentwicklung, Entzündung, Infektion, Tumorgenese, Metastasierung, Gefäßneubildung, Invasion, Zerstörung, Regeneration, Autoimmunreaktion, Immunkompatibilität, Wundheilung, Allergie, Vergiftung und/oder der Sepsis. Auch können diese die Ausprägung bestimmter klinischer Zustände charakterisieren, wie z.B. des Erkrankungsstatus oder der Wirkung von Medikamenten. Die Wahl der Expressionsprofile hängt vom Ursprung der zu untersuchenden biologischen Probe ab, sowie deren Zusammensetzung und/oder erwarteten Zusammensetzung. Gegebenenfalls müssen die Profile im Vorgang zu der Messung definiert und geeignet bestimmt werden, oder lassen sich aus öffentlichen Expressions-Datenbanken entnehmen.

[0035] Noch weiter bevorzugt ist ein erfindungsgemäßes Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei die Berechnung der Gesamtkonzentration aus den Anteilen A_i der verschiedenen Zelltypen bzw. Einflüssen (z.B. eingewanderte Zelltypen) i mit ihren unterschiedlichen Konzentrationen K_i mittels der Beziehung

$$K_{Probe} = K_1 \cdot A_1 + K_2 \cdot A_2 + \dots = \sum_{i=1}^n (K_i \cdot A_i) \text{ mit } i \in N \quad (\text{Gleichung 3})$$

erfolgt.

[0036] Immer noch weiter bevorzugt ist ein erfindungsgemäßes Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei der SLR-Wert eines Markergens mittels der Formel

$$A_{Zelltyp} = 2^{\frac{1}{k} (SLR_{Probe|Kontrolle} - SLR_{Zelltyp|Kontrolle})} \quad (\text{Gleichung (14)})$$

bestimmt wird. Für jeden Einfluß, der erkannt und von anderen molekularen Informationen abgetrennt werden soll, müssen Markergene definiert werden. Diese lassen den Anteil einer Signatur am Gesamtprofil quantitativ abschätzen. Für die Erkennung unterschiedlicher zellulärer Zusammensetzungen werden also z.B. Markergene für Monozyten, T-Zellen oder Granulozyten identifiziert. Diese spiegeln den Anteil der jeweiligen Zellpopulation in einer gemischten Probe wieder.

[0037] Bevorzugt ist ein erfindungsgemäßes Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei der Marker ausgewählt ist aus den in Tabelle 2 unten angegebenen Markern. Diese Marker sind jedoch nur beispielhaft für die dort angegebenen Zelltypen und können mittel der hier offenbarten Lehre leicht auch für andere Gewebe entsprechend ermittelt werden.

[0038] Weiter bevorzugt ist ein erfindungsgemäßes Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, umfassend die beispielhafte qualitative und/oder quantitative Erkennung von Expressionsprofilen eines T-Zell-, Monozyten- und/oder Granulozyten-Expressionsprofils.

[0039] Ein weiterer Aspekt der vorliegenden Erfindung betrifft ein Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei das Bestimmen der quantitativen Zusammensetzung des komplexen Expressionsprofils auf Basis der ermittelten Differenzen weiterhin die Identifizierung eines bisher unbekanntes Expressionsprofils umfaßt.

[0040] Der Vergleich zwischen zwei komplexen Proben liefert zunächst eine differentielle Genexpression, die sowohl durch Unterschiede der zellulären Zusammensetzung als auch durch Genregulation hervorgerufen sein kann. Im ersten Schritt ist deshalb die zelluläre Zusammensetzung aufzuschlüsseln. Dies erfolgt unter Verwendung von Signaturen, die verschiedene Zelltypen charakterisieren. Durch die Verwendung von Normalsignaturen für Gewebe und einzelne Zelltypen wird ein zu erwartendes Profil errechnet, das nur die normale Genexpression berücksichtigt. Der Unterschied aus diesem virtuellen Profil und dem tatsächlich gemessenen Profil ergibt die Gene, die entweder durch weitere, noch nicht berücksichtigte Zelltypen oder durch Regulation verändert sind. Funktionelle Veränderungen in der Genexpression sind deshalb in dieser Differenz zu erwarten. Eine Zuordnung zu einem bestimmten Zelltyp ist zunächst nicht möglich. Diese Gene gehen aber aus der funktionellen Veränderung der beteiligten Zellen hervor. Werden Markergene für die vom Zelltyp bereinigte funktionelle Signatur definiert, kann der Anteil dieser Signatur im Unterschied zwischen virtuellem Profil und tatsächlich gemessenem Profil quantitativ abgeschätzt werden. Diese funktionellen Profile lassen sich nun schrittweise aus dem Unterschied zwischen virtuellem Profil und tatsächlich gemessenem Profil erschließen.

[0041] Insgesamt werden so Parameter für die zelluläre Zusammensetzung und molekulare Funktionen geschaffen, die untereinander sowie mit klinischen Merkmalen korreliert werden können. Dadurch ergeben sich neue Bewertungsmaßstäbe für die Interpretation von Array Daten, die sowohl für die Diagnostik, als auch für die Identifizierung von therapeutisch bedeutsamen Zielstrukturen (insbesondere Proteine (z. B. Enzyme, Rezeptoren) und/oder deren Komplexe) bzw. Regulationsmechanismen entscheidende Verbesserung liefern.

[0042] Ein weiterer Aspekt der vorliegenden Erfindung betrifft somit ein Verfahren zur quantitativen Bestimmung und qualitativen Charakterisierung eines komplexen Expressionsprofils in einer biologischen Probe, wobei das Bestimmen der quantitativen Zusammensetzung des komplexen Expressionsprofils auf Basis der ermittelten Differenzen weiterhin die Identifizierung von molekularen Kandidaten für die diagnostische, prognostische und/oder therapeutische Anwendung umfaßt.

[0043] Ein noch weiterer Aspekt der vorliegenden Erfindung betrifft dann einen molekularen Kandidaten oder auch Zielstruktur für die diagnostische, prognostische und/oder therapeutische Anwendung, identifiziert mittels des erfindungsgemäßen Verfahren. Bevorzugt ist ein molekularer Kandidat für die diagnostische, prognostische und/oder therapeutische Anwendung, der eine in einer der Tabellen 5 bis 8 aufgeführte Sequenz aufweist.

[0044] Erfindungsgemäß können die molekularen Kandidaten der Erfindung zum Beispiel a) zur Charakterisierung der entzündlichen Zellinfiltration in ein entzündetes Gewebe mit Genen der Tabelle 5 abgrenzend von der Genaktivierung durch Entzündung, b) zur Charakterisierung der Genaktivierung in einem entzündeten Gewebe mit Genen der Tabelle 6 abgrenzend von der Zellinfiltration, c) zur Charakterisierung der Genaktivierung bzw. der entzündlichen Zellinfiltration in ein entzündetes Gewebe über den berechneten Anteil an Aktivierung bzw. Infiltration der Gene in Tabelle 7 und/oder d) zur Charakterisierung von Untergruppen entzündlicher Genaktivierung mit Genen der Tabellen 6, 7 und/oder 8.

[0045] Ein weiterer Aspekt der vorliegenden Erfindung betrifft dann diese Kandidaten und/oder Zielstrukturen als „Werkzeuge“ zur Diagnostik, molekularen Definition und Therapieentwicklung von Erkrankungen, insbesondere chronischer entzündlicher Gelenkerkrankungen, und anderer entzündlicher, infektiöser oder tumoröser Erkrankungen beim Menschen. Dabei können die Sequenzen einzelner Gene, eine Auswahl von Genen oder alle Gene, die in Tabellen 5 bis 8 genannt sind sowie deren kodierte Proteine verwendet werden. Diese erfindungsgemäßen Werkzeuge können weiterhin Gensequenzen einbeziehen, die in ihrer Sequenz identisch zu den in Tabellen 5 bis 8 genannten Genen bzw. zu deren kodierten Proteine sind oder mindestens 80% Sequenzidentität in den Protein-kodierenden Abschnitten besitzen. Weiterhin sind entsprechende (DNA oder RNA oder Aminosäure-) Sequenzabschnitte oder Teilsequenzen einbezogen, die in ihrer Sequenz eine Sequenzidentität von mindestens 80% zu den entsprechenden Abschnitten der genannten Gene besitzen.

[0046] Die erfindungsgemäßen Werkzeuge können in vielen Aspekten der Prognose, Therapie und/oder Diagnose von Erkrankungen Verwendung finden. Bevorzugte Verwendungen sind high-throughput Verfahren in der Protein-Expressionsanalytik (hochauflösende, zweidimensionale Protein-Gelelektrophorese, MALDI-Techniken), high-throughput Verfahren in der Protein-Spotting Technik (Protein Arrays) zum Screening von Autoantikörpern als diagnostisches Werkzeug für entzündliche Gelenkerkrankungen und andere entzündliche, infek-

tiöse oder tumoröse Erkrankungen beim Menschen, high-throughput Verfahren in der Protein-Spotting Technik (Protein Arrays) zum Screening von autoreaktiven T-Zellen als diagnostisches Werkzeug für entzündliche Gelenkerkrankungen und andere entzündliche, infektiöse oder tumoröse Erkrankungen beim Menschen, nicht-high-throughput Verfahren in der Protein-Spotting Technik zum Screening von autoreaktiven T-Zellen als diagnostisches Werkzeug für entzündliche Gelenkerkrankungen und andere entzündliche, infektiöse oder tumoröse Erkrankungen beim Menschen oder zur Erzeugung von Antikörpern (auch humanisierten oder humanen), die spezifisch für die genannten Proteine oder Teilsequenzen der Werkzeuge sind, die in Tabellen 5 bis 8 aufgeführt sind oder für die Analytik in Tierexperimenten oder zur Diagnostik bei Tieren mit entzündlichen Gelenkerkrankungen und anderen entzündlichen, infektiösen oder tumorösen Erkrankungen mittels entsprechenden homologen Sequenzen einer entsprechenden anderen Spezies.

[0047] Weiter Verwendungen betreffen die Werkzeuge als diagnostische Werkzeuge zum Nachweis genetischer Veränderungen (Mutationen), in den genannten Genen oder deren Regulationssequenzen (Promotor, Enhancer, Silencer, spezifische Sequenzen für die Bindung weiterer regulatorischer Faktoren).

[0048] Weiterhin können die erfindungsgemäßen Werkzeuge zum Therapie-Entscheid und/oder zur Verlaufskontrolle/Therapiekontrolle entzündlicher Gelenkerkrankungen und/oder anderen entzündlichen, infektiösen oder tumorösen Erkrankungen beim Menschen unter Verwendung der genannten Gene, DNA-Sequenzen oder davon abgeleitete Proteine oder Peptide und/oder zur Entwicklung von Therapiekonzepten, die direkte oder indirekte Beeinflussung der Expression der genannten Gene oder Gensequenzen, die Expression der genannten Proteine oder Protein-Teilsequenzen oder die direkte oder indirekte Beeinflussung autoreaktiver T-Zellen, gerichtet gegen die genannten Proteine oder Protein-Teilsequenzen, beinhalten verwendet werden, oder um unter Design und Verwendung von Interpretationsalgorithmen die genannten Gene und Sequenzen und deren Regulationsmechanismen verwenden, um Therapiekonzepte, -wirkungen, -optimierungen oder Krankheitsprognosen erkennen zu lassen oder vorauszusagen.

[0049] Weiterhin können die erfindungsgemäßen Werkzeuge zur Beeinflussung der biologischen Wirkung der aus den genannten Gensequenzen abgeleiteten Proteine, der unmittelbaren molekularen Regelkreise, in die die genannten Gene und davon abgeleiteten Proteine eingebunden sind, und zur Entwicklung von biologisch wirksamen Medikamenten (Biologicals) unter Verwendung von Genen, Gensequenzen, Regulation von Genen oder Gensequenzen, oder unter Verwendung von Proteinen, Proteinsequenzen, Fusionsproteinen oder unter Verwendung von Antikörpern oder autoreaktiven T-Zellen wie oben genannt, verwendet werden.

[0050] Ein weiterer Aspekt der vorliegenden Erfindung betrifft einen Array als molekulares Werkzeug, bestehend aus verschiedenen Antikörpern oder Molekülen mit vergleichbarem proteinspezifischem Bindungsverhalten, die zum Nachweis aller oder einer Auswahl der von den Genen der Tabellen 5 bis 8 abgeleiteten Proteine oder aller bzw. einer Auswahl dieser Proteine dienen. Dieser Array kann auch als Kit vorliegen, z.B. zusammen mit herkömmlichen Inhaltsstoffen und Gebrauchsanweisungen.

[0051] Ein weiterer Aspekt der vorliegenden Erfindung betrifft schließlich die Verwendung eines erfindungsgemäßen molekularen Kandidaten zum Screenen auf pharmakologisch aktive Substanzen, insbesondere Bindpartner. Entsprechende Verfahren sind im Stand der Technik gut bekannt einschließlich unter anderem der folgenden Publikationen: Abagyan R, Totrov M. High-throughput docking for lead generation. *Curr Opin Chem Biol.* 2001 Aug;S(4):375-82. Review. Bertrand M, Jackson P, Walther B. Rapid assessment of drug metabolism in the drug discovery process. *Eur J Pharm Sci.* 2000 Oct;11 Suppl 2:S61-72. Review. Panchagnula R, Thomas NS. Biopharmaceutics and pharmacokinetics in drug research. *Int J Pharm.* 2000 May 25;201(2):131-50. Review. White RE. High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery. *Annu Rev Pharmacol Toxicol.* 2000;40:133-57. Review. Zuhlsdorf MT. Relevance of pheno- and genotyping in clinical drug development. *Int J Clin Pharmacol Ther.* 1998 Nov;36(11):607-12. Review. Chu YH, Cheng CC. Affinity capillary electrophoresis in biomolecular recognition. *Cell Mol Life Sci.* 1998 Ju1;54(7):663-83. Review. Kuhlmann J. Drug research: from the idea to the product. *Int J Clin Pharmacol Ther.* 1997 Dec;35(12):541-52. Review. *J Hepatol.* 1997;26 Suppl 2:26-36. Review. Shaw I. Receptor-based assays in screening for biologically active substances. *Curr Opin Biotechnol.* 1992 Feb;3(1):55-8. Review. Matula TI. Validity of in vitro testing. *Drug Metab Rev.* 1990;22(6-8):777-87. Review. Bush K. Screening and characterization of enzyme inhibitors as drug candidates. *Drug Metab Rev.* 1983;14(4):689-708. Review.

[0052] Ein weiterer Aspekt der vorliegenden Erfindung betrifft ein Verfahren zur Diagnose, Prognose und/oder Verfolgung einer Erkrankung, umfassend ein wie oben genanntes Verfahren. Die entsprechende Verknüpfung der Expressionsprofildaten mit der Diagnose, Prognose und/oder Verfolgung einer Erkrankung ist dem Fachmann aus dem Stand der Technik bekannt und kann entsprechend an die jeweiligen Verhältnisse angepaßt

werden (siehe z. B. Simon R. Using DNA microarrays for diagnostic and prognostic prediction. Expert Rev Mol Diagn. 2003 Sep;3(5):587-95. Review.; Franklin WA, Carbone DP. Molecular staging and pharmacogenomics. Clinical implications: from lab to patients and back. Lung Cancer. 2003 Aug;41 Suppl 1:S147-54. Review. Kallow W. Pharmacogenetics and personalised medicine. Fundam Clin Pharmacol. 2002 Oct;16(5):337-42. Review; Jain KK. Personalized medicine. Curr Opin Mol Ther. 2002 Dec;4(6):548-58. Review.).

[0053] Ein weiterer Aspekt der vorliegenden Erfindung betrifft dann ein Computersystem, das mit Mitteln zur Durchführung des erfindungsgemäßen Verfahrens versehen ist. Ein Computersystem im Sinne der vorliegenden Erfindung kann aus einem oder mehreren einzelnen Rechnern bestehen, die zentral oder dezentral miteinander vernetzt sein können. Ein noch weiterer Aspekt der vorliegenden Erfindung betrifft ein Computerprogramm, umfassend einen Programmiercode, um die Schritte des erfindungsgemäßen Verfahrens durchzuführen, wenn auf einem Computer ausgeführt. Ein noch weiterer Aspekt der vorliegenden Erfindung betrifft schließlich ein computerlesbares Datenträgermedium, umfassend ein erfindungsgemäßes Computerprogramm in Form eines computerlesbaren Programmcodes.

[0054] Ein noch weiterer Aspekt der vorliegenden Erfindung betrifft einen Laborroboter oder Auswertegerät für molekulare Nachweismethoden (z. B. ein computerisiertes CCD-Kamera-Auswertesystem), umfassend ein erfindungsgemäßes Computersystem und/oder ein erfindungsgemäßes Computerprogramm. Entsprechende Geräte sind dem Fachmann gut bekannt und können an die vorliegende Erfindung ohne weiteres angepaßt werden.

[0055] Die Erfindung soll nun im folgenden anhand der beigefügten Beispiele weiter verdeutlicht werden, ohne darauf beschränkt zu sein. In den beigefügten Figuren zeigt:

[0056] Fig. 1: ein Verdünnungsexperiment zur Abschätzung der Konzentration nichtregulierter Markergene

[0057] Fig. 2: den Kurvenverlauf in Grenzbereichen bei niedriger und hoher Konzentration des Markers

[0058] Fig. 3: die verschiedenen Beziehungsgrößen, die für Berechnungen angenommen werden

[0059] Fig. 4: Beziehung zwischen Signal und Konzentration bei den Extremzuständen M_1 und M_2

[0060] Fig. 5: die hierarchische Clusteranalyse unter Verwendung der Gene aus Tabelle 5

[0061] Fig. 6: die hierarchische Clusteranalyse unter Verwendung der Informationen aus der Berechnung von Infiltrationsanteilen der verschiedenen Zelltypen (Tabelle 4)

[0062] Fig. 7: A) Hierarchische Clusteranalyse unter Verwendung der Gene aus Tabelle 6. Die Vertreter RA3, RA6, RA7 und RA9 stellen in der hierarchischen Clusteranalyse mit euklidischer Distanzrechnung eine separate Gruppe dar, die zwischen der OA Gruppe und der übrigen RA Gruppe steht. B) Veranschaulichung mittels Principle Component Analysis (PCA); Gene der Tabelle 6

[0063] Fig. 8: die hierarchische Clusteranalyse mit den Genen der Tabelle 7

[0064] Fig. 9: A) die hierarchische Clusteranalyse mit den Genen der Tabelle 8. B) die Veranschaulichung der Unterschiede mittels PCA der Experimente, die sich durch Verwendung der Gene aus Tabelle 8 ergeben

Beispiele

Ausgangssituation

[0065] Es können die beiden folgenden unterschiedlichen Ausgangssituationen vorliegen:

- 1.) Ein Zelltyp (zu messender Einfluß) kann in der Kontroll-Probe vollständig fehlen. Erst im veränderten (erkrankten) Zustand werden in der Probe unterschiedliche und für die Erkrankung bedeutsame Zellen (oder Einflüsse) gefunden. Beispiel: Synovialgewebe weist im Normalzustand keine Infiltrate aus T-Zellen, Monozyten, etc. auf. Erst durch Entzündungsvorgänge gelangen diese Zellen in das Gewebe und erfahren dort weitere Aktivierung.
- 2.) Andererseits kann bereits in der Normalsituation eine Mischung aus verschiedenen Zelltypen (oder Einflüssen) vorliegen. So setzt sich z.B. das Blut aus verschiedenen Zellen zusammen, die auch im Normalzustand Variationen unterliegen. Bei Erkrankungen können diese Variationen sehr stark ausgeprägt sein.

Sie sind nicht krankheitsspezifisch, können aber möglicherweise die für eine Krankheit typischen Genregulationen verschleiern.

Einstellungen der verwendeten Software

Identifizierung von Markergenen

[0066] Unterschiedliche Zelltypen lassen sich durch Zelloberflächenmarker unterscheiden. Vergleichbar sind auch aus Genexpressionsanalysen unterschiedliche Merkmale zu erwarten, die für einzelne Zelltypen charakteristisch sind und eine quantitative Abschätzung erlauben.

[0067] Genexpressionsprofile von Geweben und aufgereinigten Zellen wurden miteinander verglichen. Es werden Gene ausgewählt, die nur in einer Zellpopulation oder einem Gewebe vorhanden sind, nicht aber in den anderen. Diese sind Kandidaten für die Abschätzung, mit welchem Anteil diese Population in einer Probe mit gemischten Zelltypen vorhanden ist.

[0068] Die in Tabelle 1 angegebenen Zellpopulationen und Gewebe wurden miteinander verglichen. Die Auswahlkriterien für die erste Stufe der Genselektion waren, daß

- alle Messungen in der Markenpopulation eine signifikant höhere Expression ergeben als alle Messungen in anderen Populationen und Geweben und
- der mittlere Unterschied zwischen den Signalen ein Ausmaß übersteigt, das auch in einem geringen Anteil am Gesamtprofil noch meßbare Unterschiede erwarten läßt.

[0069] Mit dieser Auswahl wurden die in Tabelle 2 angegeben Gene identifiziert. Diese Gene sind nicht für alle Proben geeignet. Zum Beispiel können einige dieser Gene bei niedrigen Zellkonzentrationen nicht mehr nachweisbar sein und dann zu einer quantitativen Unterschätzung des Einflusses führen. Deshalb sind weitere Einschränkungskriterien erforderlich, die auf die zu untersuchenden komplexen Proben angepaßt werden müssen.

- Die Markergene müssen ausreichende Signale und Unterschiede in der zu untersuchenden komplexen Probe liefern, wenn eine Infiltration/Anteil am Gesamtprofil erwiesen ist (z.B. über Bestimmung des Differentialblutbildes).
- Im Vergleich zur Kontrolle darf keine Regulation dieser Gene in der zu untersuchenden Probe stattfinden.
- Die Gene dürfen im Signaturprofil im Vergleich zur untersuchten Probe nicht artifiziell induziert oder unterdrückt sein.

[0070] Für die Untersuchung von Synovialgeweben bzw. Vollblut-Proben wurden die in Tabelle 2 gesondert bezeichneten Gene verwendet. Zur Berechnung der Anteile wurden die im nachfolgenden Abschnitt etablierten Bedingungen und aufgestellten Gleichungen verwendet. Zur Auswahl dienten die in Tabelle 3 genannten Einschränkungskriterien.

Beziehung zwischen Signal und RNA- bzw. Zell-Konzentration

[0071] Es wird von der Grundbeziehung ausgegangen, daß sich die logarithmierten Werte von gemessenem Signal und Konzentration der RNA zueinander linear verhalten (Gleichung 1).

$$\log_b(y) = k \cdot \log_b(x) + a \quad (\text{Gleichung 1})$$

mit $y :=$ Signal; $x :=$ Konzentration der RNA und $b \in \mathbb{R}$.

[0072] Die praktische Anwendbarkeit wurde in einem Verdünnungsexperiment mit unterschiedlichen Konzentrationen von CD4-positiven T-Zellen in CD4-depletierten peripheren mononuklearen Blutzellen überprüft. Für nicht regulierte Gene, die ausschließlich in einer Population vorkommen, stellt die Konzentration dieser Population eine „Konzentrationseinheit“ für das Gen dar. Damit verhält sich auch der Logarithmus der Konzentration der CD4-positiven Zellen zum Logarithmus des Signals linear. Diese Annäherung wird in [Fig. 1](#) an dem Verdünnungsexperiment veranschaulicht.

[0073] Aus dieser Modellannahme ergibt sich folgende theoretischen Beziehungen:

- Mit Annäherung an eine Konzentration von 0 geht der Logarithmus gegen $-\infty$.
- Mit einer Annäherung der Signale an 0 geht auch der Logarithmus der Signale gegen $-\infty$.

[0074] In der Realität ergeben sich jedoch andere Grenzbedingungen. Bei niedrigen Konzentrationen eines Gens wird die Nachweisgrenze erreicht. Geringe Signale der spezifisch bindenden Probe werden durch Signale aus fehlerhaften Hybridisierungen und Hintergrundintensitäten überlagert. Dadurch kommt es zu einer Abflachung, wie sie in [Fig. 2](#) dargestellt ist. Dieser Übergang erweist sich in der Praxis sehr vielfältig. Wird für diesen Grenzbereich eine lineare Beziehung angenommen, ergeben sich fälschlicherweise zu hohe Werte für die Konzentration des Gens in einer Probe.

[0075] Darüber hinaus wird die Hybridisierungsstärke und damit der Anstieg des Signals mit der Zunahme der Konzentration für jede Sequenz einer individuellen Dynamik folgen. Diese ist bestimmt von der Sequenz der Probe, aber auch von den Hybridisierungsbedingungen, der Hybridisierungsdauer und den Stringenzbedingungen der anschließenden Waschschrte.

[0076] Auch in hohen Signalbereichen werden sich die Hybridisierungs- und Detektionsbedingungen nicht mehr linear verhalten sondern an ein Maximum des Meßsystems annähern. In diesem Bereich wird die echte Konzentration eines Gens unterschätzt ([Fig. 2](#)).

[0077] Die tatsächlichen Konzentrationen für ein Gen in einer gegebenen Probe sind unbekannt. Sie ließen sich theoretisch aus der Array-Hybridisierung nur abschätzen, wenn eine entsprechende Eichkurve für jedes Gen vorliegen würde. Diese Eichkurven sind aber nicht vorhanden und auch zu aufwendig, um sie für alle Gene zu erstellen. Für den Vergleich zweier Arrays ist zunächst die Kenntnis der Konzentrationen auch unerheblich. Bedeutend ist lediglich die Abstimmung der Arrays aufeinander durch Normalisierung der Signale.

[0078] [Fig. 3](#) veranschaulicht die verschiedenen Beziehungsgrößen, die für Berechnungen angenommen werden.

[0079] Es ergibt sich aus Gleichung 1 für die Bestimmung von Unterschieden zwischen zwei Arrays A und B folgende Beziehung:

$\log_b(S_A) - \log_b(S_B) = [k \cdot \log_b(K_A) + a] - [k \cdot \log_b(K_B) + a]$ oder zusammengefaßt

$$\log_b\left(\frac{S_A}{S_B}\right) = k \cdot \log_b\left(\frac{K_A}{K_B}\right) \quad (\text{Gleichung 2})$$

[0080] Damit ist die Bestimmung des Unterschieds zwischen den logarithmierten Werten der Signale S_A und S_B , die auch Signal Log Ratio genannt wird, ein Maß für die Unterschiede zwischen den Konzentrationen K_A und K_B in den beiden Proben A und B.

[0081] Für die Berechnung der Gesamtkonzentration aus den Anteilen A_i der verschiedenen Zelltypen bzw. Einflüssen i mit ihren unterschiedlichen Konzentrationen K_i ergibt sich die folgende Beziehung:

$$K_{probe} = K_1 \cdot A_1 + K_2 \cdot A_2 + \dots = \sum_{i=1}^n (K_i \cdot A_i) \text{ mit } i \in N \quad (\text{Gleichung 3})$$

[0082] Damit wird ersichtlich, daß für die Aufschlüsselung der Gesamtprofile in Einzelkomponenten die Festlegung von absoluten Bezugsgrößen für die RNA- bzw. Zell-Konzentration erforderlich ist.

[0083] Abschätzung der Detektionsgrenzen und des Dynamikbereichs des Arrays Aus Gleichungen 1 bis 3 und den Überlegungen zu [Fig. 2](#) ergeben sich folgende unbekannte Größen, die für die Berechnung erforderlich sind:

- der Anstieg k als Ausdruck der Dynamik des Meßbereichs für ein Gen und
- die Zuordnung eines definierten Signalwerts zu einer definierten Konzentration für die Festlegung der Geraden im Koordinatensystem

[0084] Als Fixpunkt für die Festlegung der Geraden im Koordinatensystem wird die untere Nachweisgrenze S_{min} gewählt. Die Nachweisgrenze kann theoretisch durch Verdünnungsexperimente für jedes Gen ermittelt werden. Alternativ kann zur Abschätzung eine fehlerhafte Hybridisierung mit nicht vollständig übereinstimmenden Sequenzen (mismatch-Oligonukleotiden) gemessen werden. Die Affymetrix Technologie verwendet diese perfect match/mismatch Technologie und errechnet daraus eine Wahrscheinlichkeit, ob das gemessene Signal eines Gens vorhanden oder abwesend ist („present“ oder „absent“).

[0085] Um S_{\min} für jedes Gen individuell zu bestimmen, wurden 123 Messungen mit Affymetrix HG-U133A Arrays von verschiedenen Zelltypen, Zellmischungen und Gewebeproben analysiert. Es wurden die maximalen und minimalen Werte für jedes gemessene Gen bestimmt.

[0086] Gleichzeitig wurde die Präsenz dieser Gene geprüft. Es ergaben sich von insgesamt 22283 Affymetrix "Probe-Sets" dieses Arrays drei Gruppen:

- 1.) 4231 Probe-Sets, die in allen 123 Messungen als "absent" eingestuft wurden,
- 2.) 2197 Probe-Sets, die ausschließlich den Status "present" lieferten und
- 3.) 15855 Probe-Sets, die zum Teil mit "present" und zum Teil mit "absent" eingestuft wurden.

[0087] Die Gene, die nur absent gefunden wurden, spielen in den gemessenen Proben offensichtlich keine Rolle und müssen in der Berechnung nicht näher berücksichtigt werden. Sollten diese Gene in anderen Probenarten nachweisbar werden, kann analog zur 3. Gruppe die Berechnung erfolgen. Für Gene, die ausschließlich als "present" eingestuft werden, kann eine Nachweisgrenze nur geschätzt werden. Als Maß kann der Median oder Mittelwert aller Nachweisgrenzen dienen, die für die 3. Gruppe definiert wurden.

[0088] Die Signalthöhe S_{\min} als Grenze des Übergangs von „absent“ zu „present“ wurde ebenfalls aus den 123 Messungen für jedes Gen individuell bestimmt. Zunächst wurden die niedrigsten „present“ Signale und höchsten „absent“ Signale festgestellt. Von allen zwischen diesen Grenzen liegenden Werten wurde der Median als die Grenze S_{\min} definiert. Bei fehlender Überlappung wurde der höchste „absent“ Wert als S_{\min} festgelegt. Für alle Gene, die keine „absent“ Bestimmungen aufwiesen, wurde der Median aller S_{\min} Grenzwerte als einheitlicher S_{\min} festgelegt (68,6). Alternativ könnte auch eine andere Form der Abschätzung wie der Mittelwert oder ein gewichteter Mittelwert dienen.

[0089] Die Abschätzung des Dynamikbereichs kann aus den gemessenen Signalwerten einer Vielzahl verschiedener Experimente mit unterschiedlichen Proben wie folgt abgeschätzt werden:

S_1 kann als der größte gemessene Wert in einer Reihe von Experimenten unabhängig vom Gen als oberer Grenze des Meßspektrums definiert werden.

S_0 kann als der niedrigste verlässlich gemessene Wert dieser Reihe von Experimenten unabhängig vom Gene definiert werden.

[0090] Die Signal Log Ratio ergibt sich dann als

$$\log_b \frac{S_1}{S_0} \quad \text{(Gleichung 4)}$$

[0091] In dem hier verwendeten Beispiel wurden aus den 123 Messungen das größte Signal mit $S_1 = 31581,5$ arbitrary units; AU) und das kleinste Signal mit $S_0 = 1,2$ AU unabhängig von einem individuellen Gen über alle Gene bestimmt.

[0092] Die Signal Log Ratio errechnet sich damit unter Verwendung von $b = 2$ für die Basis des Logarithmus wie folgt:

$$\log_2 \frac{S_1}{S_0} = \log_2 \left(\frac{31581,5}{1,2} \right) \approx 14,7$$

[0093] Zum Vergleich ergab der Unterschied zwischen größtem Signal und kleinstem Signal unter Betrachtung jedes Gens für sich eine Signal Log Ratio von 15,4. Wurden nur „present“ Signale einbezogen und jedes Gen für sich betrachtet, lag die größte Signal Log Ratio bei 10,5. Alle absoluten Zahlenwerte für Signalgrößen sind abhängig von der Einstellung der Normierungsgrößen im jeweiligen Software Paket für das Auslesen und Vergleichen von DNA Arrays. Es ist nicht die Einstellung auf bestimmte Normierungsgrößen und damit die hier genannten Zahlenwerte entscheidend sondern die einheitliche Verwendung der gleichen Einstellung für alle Array-Analysen, die für die Berechnung benötigt werden. Mit der Einstellung auf andere Normierungsgrößen ergeben sich somit andere Zahlenwerte, die entsprechend den genannten Auswahlbedingungen festzulegen. Entscheidend ist dann die einheitliche Anwendung.

[0094] Der Wert aus Gleichung 4 wurde in dem hier dargestellten Beispiel als theoretisches Maß für den maximalen Dynamikbereich der Signale festgelegt. Für die angestrebten relativen Berechnungen sind die exakten Größen für beide Skalen nicht entscheidend. Die Signaleinheiten werden bei jeder Array Plattform arbiträr festgelegt. Ebenso können die Konzentrationseinheiten arbiträr festgelegt werden. Entscheidend sind die relativen

Beziehungen zwischen den Signalen und Konzentrationen sowie die Festlegung der Nachweisgrenze. Ferner muß bei einem Gen für alle verschiedenen Zelltypen und Proben die gleiche Beziehung gelten, um Berechnungen zwischen den verschiedenen Proben und Signaturen durchzuführen. Die Anwendung ähnlicher Größenverhältnisse für die Beziehung zwischen Konzentration und Signal bei allen verschiedenen Genen ermöglicht den Anteil einer Signatur von einem Gen auf andere Gene angenähert zu übertragen. Es wird hier die Konvention getroffen, daß für den Konzentrationsbereich eine Größenordnung vergleichbar zum Signalbereich zugeordnet wird.

[0095] Für die Beziehung zwischen Signal und Konzentration ergeben sich die in [Fig. 4](#) dargestellten Extremzustände M_1 und M_2 . Sie zeigen die beiden Grenzbereiche auf, wie die Beziehung zwischen Konzentration und Signal in Abhängigkeit von der Nachweisgrenze in das Modell einfließen kann.

[0096] M_0 zeigt dabei den Verlauf unter optimalen Bedingungen. Im diesem idealen Fall wird bereits bei sehr niedrigen Signalen S_{\min} eine lineare Beziehung zur minimalen Konzentration K_{\min} bestehen. Für viele Gene liefert die Analyse der Hybridisierung jedoch ein relativ hohes Einstiegssignal $S_{\min G}$, über welchem erst verlässlich die Präsenz eines Gens angezeigt wird und von dem aus eine lineare Beziehung angenommen werden darf.

[0097] Im Modell M_1 wird davon ausgegangen, daß eine Hintergrundaktivität die Nachweisgrenze $K_{\min 1}$ eines Gens nicht wesentlich beeinträchtigt. Es wird lediglich der Detektionsbereich des Signals verkleinert und damit die Dynamik des Signalanstiegs vermindert. Im Modell M_2 wird davon ausgegangen, daß niedrige Konzentrationen durch den hohen Hintergrund verborgen bleiben und erst ab einer höheren Konzentration $K_{\min M_2}$ ein Gen detektiert werden kann. Die [Fig. 4](#) veranschaulicht die Auswirkungen auf die Konzentrationsbestimmungen $K_{\text{Probe}M_1}$ bzw. $K_{\text{Probe}M_2}$ in Abhängigkeit von der Wahl des Modells M_1 oder M_2 .

[0098] Im Modell M_1 wird für jedes Gen individuell der Signalwert S_{\min} berechnet und diesem eine minimale Konzentration K_{\min} zugeordnet. Dabei muß gelten $K_{\min} < K_1$. Aus praktischen Gründen wurde hier $K_{\min} = 1$ zugeordnet. Dem höchsten gemessenen Signalwert S_1 wird K_1 zugeordnet. Aus praktischen Gründen wurde eine vergleichbar zum Signalmessbereich liegende Konzentration von $K_1 = 2^{14,7}$ zugeordnet. Die Steigung der Geraden ergibt sich über Gleichung 1 für jedes Gen individuell wie folgt:

$$k = \frac{\log_b(S_1) - \log_b(S_{\min})}{\log_b(K_1) - \log_b(K_{\min})} \quad (\text{Gleichung 5})$$

[0099] Im Modell M_2 ist $K_{\min 1} = 1$ und damit $K_{\min M_2}$ deutlich größer als $K_{\min 1}$. Die Steigung der Geraden ergibt sich aus den besten gemessenen und hier als ideal anzusehenden Detektionsgrenzen $K_{\min 1} = 1$ und $S_{\min 1} = 1,2$ sowie den zugehörigen maximalen Werten $S_1 = 31581,5$ und $K_1 = 2^{14,7}$ wie folgt:

$$k = \frac{\log_2(S_1) - \log_2(S_{\min 1})}{\log_2(K_1) - \log_2(K_{\min 1})} = \frac{14,7}{14,7} = 1 \quad (\text{Gleichung 6})$$

[0100] Signalwerten unter der Detektionsgrenze können in beiden Modellen keine eindeutigen Konzentrationswerte zugeordnet werden. Im grau unterlegten Bereich der [Fig. 4](#) liegt die mögliche Schwankungsbreite der Beziehung zwischen Signal und Konzentration. Theoretisch könnte über aufwendige Verdünnungsreihen für jedes Gen individuell eine spezifische Beziehungsgleichung aufgestellt werden. Diese müßte dann auch für jede Probenart überprüft werden und bei Weiterentwicklungen des Arrays wieder neu erhoben werden. Derzeit stehen solche Daten nicht zur Verfügung. Es werden deshalb anhand beider Modelle M_1 und M_2 berechnet und die Ergebnisse miteinander verglichen.

[0101] Zusammengefaßt ergibt sich nun unter Anwendung der Gleichung 1 für das Modell M_1 die Beziehung

$$\log_b(S_{\text{Probe}}) = \frac{\log_b(S_1) - \log_b(S_{\min})}{\log_b(K_1) - \log_b(K_{\min})} \cdot \log_b(K_{\text{Probe}}) + \log_b(S_{\min}) \quad (\text{Gleichung 7})$$

und für das Modell M_2 entsteht unter Verwendung der in Gleichung 6 angenommenen Bezugsgrößen zwischen Signal und Konzentration die Beziehung

$$\log_b(S_{\text{Probe}}) = \log_b(K_{\text{Probe}}) + \log_b(S_{\min}) \quad (\text{Gleichung 8})$$

Quantitative Abschätzung der Anteile einer Zellpopulation in einer Probe mit verschiedenen Zelltypen

[0102] Die dargestellten Berechnungsgrundlagen lassen sich zunächst auf die Markergene für einzelne Zelltypen anwenden. Für die in Tabelle 2A bis C genannten Gene ergibt dies die in Tabelle 2A bis C genannten S_{\min} Werte.

[0103] Aus Gleichung 7 und 8 kann die RNA Konzentration für ein Markergen in einer gemessenen Probe wie folgt abgeleitet werden:

Modell M_1

$$K_{Probe} = b^{\left[\log_b(S_{Probe}) - \log_b(S_{\min})\right] \frac{\log_b(K_I) - \log_b(K_{\min})}{\log_b(S_I) - \log_b(S_{\min})}} \quad \text{bzw.} \quad K_{Zelltyp} = b^{\left[\log_b(S_{Zelltyp}) - \log_b(S_{\min})\right] \frac{\log_b(K_I) - \log_b(K_{\min})}{\log_b(S_I) - \log_b(S_{\min})}} \quad (\text{Gleichung 9})$$

[0104] Modell M_2 unter Verwendung der in Gleichung 6 angenommenen Bezugsgrößen zwischen Signal und Konzentration:

$$K_{Probe} = b^{\left[\log_b(S_{Probe}) - \log_b(S_{\min})\right]} \quad \text{bzw.} \quad K_{Zelltyp} = b^{\left[\log_b(S_{Zelltyp}) - \log_b(S_{\min})\right]} \quad (\text{Gleichung 10})$$

[0105] Ein Markergen für einen bestimmten Zelltyp wurde so definiert, daß es in anderen Zell- oder Gewebetypen nicht oder vernachlässigbar gering zu finden ist. Damit ergibt sich folgende Berechnung:

$$A_{Zelltyp} \cdot K_{Zelltyp} + A_{Kontrolle} \cdot K_{Kontrolle} = K_{Probe}$$

[0106] Da Anteil der Zellpopulation und Konzentration des Markergens in der Kontrolle gegen null geht ($A_{Kontrolle} < 0,01$, $S_{Kontrolle} < S_{\min}$ und damit $K_{Kontrolle} < 1$), ergibt sich für den Anteil des Zelltyps in einer gemischten Probe:

$$A_{Zelltyp} = \frac{K_{Probe}}{K_{Zelltyp}} \quad (\text{Gleichung 11})$$

[0107] Für die Berechnung der Konzentrationen stehen verschiedene Ausgangsdaten zur Verfügung. So liefern zahlreiche Plattformen und Software Pakete normalisierte Signalwerte, mit denen weitere Berechnungen durchgeführt werden können. Hierfür sind die genannten Gleichungen unmittelbar anwendbar.

[0108] Eine besondere Stellung nimmt die Affymetrix Technologie ein. Bei dieser Plattform werden mehrere verschiedene Oligonukleotide pro Gen und zugehörige „mis-match“ Oligonukleotide verwendet. Auch hier können Signale zur unmittelbaren weiteren Berechnung generiert werden (z.B. über die robust multiarray analysis; RMA). Sowohl Signalbestimmung als auch Vergleiche können aber auch über spezielle Algorithmen vorgenommen werden, die sowohl perfect match als auch mis-match Informationen einbeziehen. Die Ergebnisse aus der Vergleichsberechnung werden ebenfalls als Signal Log Ratio (SLR) angegeben und können in die hier durchgeführten Berechnungen integriert werden. Ferner kann auf diese Weise eine Bezugspopulation als Norm verwendet werden. Dies ist in [Fig. 3](#) veranschaulicht. Diese Bezugsgröße wird Kontrolle genannt. Für das Beispiel der Synovialgewebsanalyse ist dies das Normalgewebe (siehe auch Tabelle 1). Es ergeben sich hierbei für die Berechnung der Infiltration folgende Beziehungen:

$$SLR_{Zelltyp/Kontrolle} = \log_b \left(\frac{S_{Zelltyp}}{S_{Kontrolle}} \right) \quad \text{und} \quad SLR_{Probe/Kontrolle} = \log_b \left(\frac{S_{Probe}}{S_{Kontrolle}} \right).$$

[0109] Zusammen mit Gleichung 1 folgt daraus:

$$\log_b(K_{Zelltyp}) = SLR_{Zelltyp/Kontrolle} \cdot \frac{1}{k} + \log_b(K_{Kontrolle}) \quad \text{bzw.}$$

$$K_{Zelltyp} = K_{Kontrolle} \cdot 2^{\frac{1}{k} \cdot SLR_{Zelltyp/Kontrolle}} \quad (\text{Gleichung 12})$$

und analog

$$K_{Probe} = K_{Kontrolle} \cdot 2^{\frac{1}{k} \cdot SLR_{Probe/Kontrolle}} \quad (\text{Gleichung 13})$$

[0110] Unter Verwendung der Gleichungen 11, 12 und 13 folgt für den Anteil eines Zelltyps gemessen an den SLR Werten von Markergenen:

$$A_{Zelltyp} = 2^{\frac{1}{k} (SLR_{Probe/Kontrolle} - SLR_{Zelltyp/Kontrolle})} \quad (\text{Gleichung 14})$$

[0111] Für die beiden Modelle M_1 und M_2 ergibt sich der Wert für die Steigung k aus den Gleichungen 5 und 6.

[0112] Die Gleichung 14 kann auf mehrere Gene angewendet werden, die für die Abschätzung der Anteile eines Zelltyps in einer Zellmischung geeignet sind (siehe Tabelle 2 und 3). Der Mittelwert aus den je Gen berechneten Anteilen liefert ein Maß für den Anteil des Zelltyps in der zu untersuchenden Probe.

Identifizierung der regulierten Gene durch Berechnung der virtuellen Profile aus der zellulären Zusammensetzung

[0113] Sind die verschiedenen zellulären Komponenten einer Probe und ihre anteilige Verteilung bekannt, kann aus den Profilen für jeden Zelltyp ein zu erwartendes Mischprofil errechnet werden.

1. Ausgangssituation: Der Zelltyp fehlt in der Normalsituation

[0114] Für das Synovialgewebe ergibt sich die Ausgangssituation, daß das Normalgewebe keine Immunzellen enthält. Dies entspricht dem oben genannten Kontrollgewebe. Die Infiltration bei Erkrankung kann über die Markergene verschiedener Zellpopulationen wie oben dargestellt berechnet werden (Gleichung 11 bzw. 14). Die Anteile der jeweiligen Zelltypen und des Normalgewebes summieren sich zu 100%.

[0115] Weiterhin kann über Gleichung 12 für jedes in einem Zelltyp exprimierte Gen die Konzentration $K_{Zelltyp}$ bestimmt werden. Die Konzentration $K_{Kontrolle}$ im Kontrollgewebe, dem normalen Synovialgewebe, wird über das Signal $S_{Kontrolle}$ des betreffenden Gens gemäß Gleichung 8 ermittelt.

[0116] Die zu erwartende Konzentration K'_{Probe} eines Gens, die aufgrund der zellulären Zusammensetzung zu erwarten ist, errechnet sich dann gemäß Gleichung 3 wie folgt:

$$K'_{Probe} = A_{Kontrolle} \cdot K_{Kontrolle} + \sum_{i=1}^n (A_i \cdot K_i) \quad (\text{Gleichung 15})$$

[0117] Der zugehörige logarithmierte Wert des Signals ergibt sich über Gleichung 1 mit

$$\log_b(S'_{Probe}) = k \cdot \log_b(K'_{Probe}) + \log_b(S_{min}) \quad (\text{Gleichung 16})$$

mit k gemäß Modell M_1 bzw. M_2 aus den Gleichungen 5 und 6

[0118] Der gemessene Unterschied zwischen erkranktem Synovialgewebe und normalem Synovialgewebe ergibt sich als

$$SLR_{Probe/Kontrolle}$$

[0119] Der Anteil der Regulation $SLR_{reguliert}$ ergibt sich durch Abzug der Infiltration:

$$SLR_{reguliert} = \log_b \frac{S_{Probe}}{S'_{Probe}} = SLR_{Probe/Kontrolle} - \log_b \frac{S'_{Probe}}{S_{Kontrolle}} \quad (\text{Gleichung 17})$$

[0120] Alternativ kann in gleicher Weise der Konzentrationsunterschied (Concentration Log Ratio; CLR) berechnet werden unter Verwendung von Gleichung 13 und 15:

$$CLR_{reguliert} = \log_b \frac{K_{Probe}}{K'_{Probe}} = \frac{K_{Kontrolle} \cdot 2^{\frac{1}{k} \cdot SLR_{Probe/Kontrolle}}}{A_{Kontrolle} \cdot K_{Kontrolle} + \sum_{i=1}^n (A_i \cdot K_i)} \quad (\text{Gleichung 18})$$

mit k gemäß Modell M_1 bzw. M_2 aus den Gleichungen 5 und 6.

2. Ausgangssituation: Der Zelltyp ist in der Normalsituation vorhanden

[0121] Im Vollblut sind bereits in der Normalsituation die verschiedenen Immunzellen vorhanden. Daher wird zunächst die „Normalsituation“ analysiert.

Bestimmung der Normalsituation

[0122] Die Berechnungen erfolgen direkt mit den ermittelten und aufeinander abgestimmten Signalen. Alternativ kann mit Hilfe des von Affymetrix entwickelten Vergleichs-Algorithmus unter Berücksichtigung der perfect match und mis-match Informationen der Bezug zu einem Kontrollgewebe dienen, das die verschiedenen Zelltypen nicht enthält wie z.B. das normale Synovialgewebe. Die Konzentration $K_{\text{Kontrolle}}$ errechnet sich somit aus Gleichung 10 bzw. 13. Die Anteile der einzelnen Zelltypen werden gemäß Gleichung 11 aus den Konzentrationen der Markergene bzw. den SLRs gemäß Gleichung 14 abgeschätzt.

[0123] Zur Berechnung der Gesamtkonzentration fehlt der Anteil der Restpopulationen, die nicht als Einzelprofile vorliegen. Dieser kann zu einer eigenen virtuellen „Restpopulation“ zusammengefaßt werden. Ihr Anteil ergibt sich wie folgt:

$$A_{K, \text{Rest}} = 1 - \sum_{i=1}^n A_{K,i} \quad (\text{Gleichung 19})$$

[0124] Der Anteil der Restpopulation kann verschwindend gering sein und die errechnete erwartete Konzentration aus den Signaturen und ihren Anteilen den tatsächlich gemessenen Wert übersteigen, also

$$K_{\text{Kontrolle}} - \sum_{i=1}^n (A_{K,i} \cdot K_i) < 0$$

[0125] Für diesen Fall ist eine einheitliche Anpassung der Konzentrationen K_i für jeden Zelltyp i erforderlich. Unter der Annahme, daß kein Beitrag durch das Restprofil erfolgt, d.h. die Expression des Gens im Restprofil unterhalb der Nachweisgrenze liegt, ergibt sich der Korrekturfaktor wie folgt:

$$KF = \frac{K_{\text{Kontrolle}}}{A_{K, \text{Rest}} \cdot K_{\text{Rest}} + \sum_{i=1}^n (A_{K,i} \cdot K_i)} \quad (\text{Gleichung 20})$$

mit $K_{\text{Rest}} < K_{\text{min}}$. Hier kann z.B. ein Wert von $K_{\text{Rest}} = 0,5$ eingesetzt werden.

[0126] Die Konzentration für jedes Gen im Profil der virtuellen Restpopulation ergibt sich unter Anwendung der Gleichung 3 als

$$K_{\text{Rest}} = \frac{1}{A_{K, \text{Rest}}} \cdot \left(K_{\text{Kontrolle}} - \sum_{i=1}^n (A_{K,i} \cdot K_i) \right) \quad (\text{Gleichung 21})$$

[0127] Damit wird die Summe aus den errechneten Einzelkomponenten der Konzentrationen identisch mit der aus der tatsächlichen Messung errechneten Konzentration, also

$$K_{\text{Kontrolle}} = A_{K, \text{Rest}} \cdot K_{\text{Rest}} + \sum_{i=1}^n (A_{K,i} \cdot K_i) \quad (\text{Gleichung 22})$$

[0128] Für jedes Gen werden die berechneten Konzentrationen K_{Rest} der Restpopulation aus allen Normal Spendern gemittelt. So entsteht eine virtuelle Signatur für die Restpopulation des Normal spenders vergleichbar zu den gemessenen Signaturen der verschiedenen Zelltypen. Hiermit sind alle Voraussetzungen für die Berechnung der Normalsituation auf der Basis der vorhandenen Zellsignaturen und eines virtuellen normalen Restprofils gegeben.

Bestimmung in der Krankheitssituation

[0129] Die Berechnungen erfolgen analog zur Normalsituation direkt mit den ermittelten und aufeinander abgestimmten Signalen. Alternativ kann mit Hilfe des von Affymetrix entwickelten Vergleichs-Algorithmus der Be-

zug zum gleichen Kontrollgewebe wie für Normalspender verwendet werden. Die Konzentration K_{Probe} errechnet sich somit aus Gleichung 10 bzw. 13. Die Anteile der einzelnen Zelltypen werden gemäß Gleichung 11 aus den Konzentrationen der Markergene bzw. den SLRs gemäß Gleichung 14 abgeschätzt. Der Anteil der Restpopulation ergibt sich aus Gleichung 19.

[0130] Die erwartete Konzentration gemäß der zelluläre Zusammensetzung errechnet sich aus den Einzelkomponenten gemäß Gleichung 22:

$$K'_{\text{Probe}} = A_{P,\text{Rest}} \cdot K_{\text{Rest}} + \sum_{i=1}^n (A_{P,i} \cdot K_i)$$

[0131] Die erwarteten Signale errechnen sich aus Gleichung 16. Die regulierten Gene, die sich nicht auf die bekannten Signaturen zurückführen lassen, ergeben sich entweder über die SLRs gemäß Gleichung 17 oder die CLRs gemäß Gleichung 18.

Anwendung des Berechnungsverfahrens zur Charakterisierung von Genexpressionsprofilen

[0132] Die Zerlegung in einzelne Komponenten erfolgt schrittweise.

1. Aufteilung in Teilkomponenten von Zelltyp-Signaturen.
2. Erkennen funktioneller Signaturen
3. Prüfung von gegenseitigen Abhängigkeiten zwischen 1. und 2.
4. Korrelation mit klinischen Merkmalen

[0133] Der Vergleich zwischen zwei komplexen Proben liefert zunächst eine differentielle Genexpression, die sowohl durch Unterschiede der zellulären Zusammensetzung als auch durch Genregulation hervorgerufen sein kann. Im Ersten Schritt ist deshalb die zelluläre Zusammensetzung aufzuschlüsseln. Dies erfolgt unter Verwendung von Signaturen, die verschiedene Zelltypen charakterisieren. Durch die Verwendung von Normalsignaturen für Gewebe und einzelne Zelltypen wird ein zu erwartendes Profil errechnet, das nur die normale Genexpression berücksichtigt. Der Unterschied aus diesem virtuellen Profil und dem tatsächlich gemessenen Profil ergibt die Gene, die entweder durch weitere, noch nicht berücksichtigte Zelltypen oder durch Regulation verändert sind. Funktionelle Veränderungen in der Genexpression sind deshalb in dieser Differenz zu erwarten. Eine Zuordnung zu einem bestimmten Zelltyp ist zunächst nicht möglich. Diese Gene gehen aber aus der funktionellen Veränderung der beteiligten Zellen hervor.

$$K_{\text{Probe}} = \sum_{i=1}^n A_i \cdot K_i + \sum_{i=1}^n A_i \cdot K_{i,\text{reg}}$$

mit der Konzentration K_i im Normalzustand und der Konzentrationsänderung $K_{i,\text{reg}}$, die sich zusätzlich durch die funktionelle Regulation ergibt mit i als der Anzahl der verschiedenen beteiligten Zelltypen.

[0134] Die Untersuchung von einzelnen Zelltypen unter funktionellen Einflüssen kann eine funktionelle Signatur für einen Zelltyp liefern. Diese funktionelle Veränderung läßt sich wie folgt darstellen:

$$K_{i,\text{f}} = K_i + K_{i,\text{reg}}$$

[0135] Daraus ergibt sich eine von der Signatur des Zelltyps bereinigte funktionelle Konzentrationsänderung

$$K_{i,\text{reg}} = K_{i,\text{f}} - K_i$$

[0136] Werden Markergene für die vom Zelltyp bereinigte funktionelle Signatur definiert, kann der Anteil dieser Signatur im Unterschied zwischen virtuellem Profil und tatsächlich gemessenem Profil quantitativ abgeschätzt werden. Diese funktionellen Profile lassen sich nun schrittweise aus dem Unterschied zwischen virtuellem Profil und tatsächlich gemessenem Profil erschließen.

[0137] Insgesamt werden so Parameter für die zelluläre Zusammensetzung und molekulare Funktionen geschaffen, die untereinander sowie mit klinischen Merkmalen korreliert werden können. Dadurch ergeben sich neue Bewertungsmaßstäbe für die Interpretation von Array Daten, die sowohl für die Diagnostik, als auch für die Identifizierung von therapeutisch bedeutsamen Zielstrukturen bzw. Regulationsmechanismen entscheidende Verbesserung liefern.

Anwendung am Beispiel des Synovialgewebes.

[0138] Das genannte Verfahren wurde auf die Analyse von insgesamt 10 verschiedenen Proben von Patienten mit rheumatoider Arthritis (RA), 10 Patienten mit Osteoarthritis (OA) und 10 normale Synovialgewebe angewendet. Die in Tabelle 2 unter Auswahl mit 1 markierten Gene wurden für die Abschätzung der Anteile der CD4+ T-Zellen, der Monozyten und der Granulozyten im Synovialgewebe von den RA und OA Patienten verwendet. Es ergab sich die in Tabelle 4 genannte anteilige Verteilung für RA bzw. OA.

[0139] Anhand der dargestellten Berechnungsgrundlagen und der Anwendung von Modell M_1 wurden die Anteile bestimmt, die je Gen durch Infiltration von T-Zellen, Monozyten bzw. Granulozyten zu erwarten sind. Aus dem Unterschied zwischen zu erwartendem Expressionsniveau über die Berechnungsgrundlage nach Modell M_1 und dem tatsächlich gemessenem Expressionsniveau ergab sich der Anteil der durch Aktivierung bedingten Expressionsunterschiede. Es wurden zunächst die Gene bestimmt, die mittels der von Affymetrix entwickelten Software MAS 5.0 einen Unterschied in mehr als 50% aller paarweisen Vergleiche zwischen RA und Normalgewebe ergaben mit einer mittleren SLR größer 1,5. Die so erhaltenen Geneinträge wurden weiter unterteilt in Gruppen, die folgende Bedingungen erfüllen:

1. infiltrierte Gene, wenn das Verhältnis von $SLR_{\text{Probe/Probe}}$ zu $SLR_{\text{Probe/Kontrolle}}$ unter 0,25 lag
2. regulierte Gene bzw. Gene von weiteren eingewanderten Zelltypen, die noch nicht berücksichtigt wurden, wenn das Verhältnis von $SLR_{\text{Probe/Probe}}$ zu $SLR_{\text{Probe/Kontrolle}}$ über 0,75 lag
3. Gene, die sowohl infiltriert als auch reguliert bzw. von weiteren nicht berücksichtigten Zelltypen stammen können, wenn das Verhältnis von $SLR_{\text{Probe/Probe}}$ zu $SLR_{\text{Probe/Kontrolle}}$ zwischen 0,25 und 0,75 lag.

[0140] Die unter der 1. Bedingung gefundenen Geneinträge sind unten in Tabelle 5 angegeben. Sie stellen einen Genpool dar, der bei einer chronisch entzündlichen Gelenkerkrankung wie der rheumatoiden Arthritis als Diagnostikum für das Ausmaß der Infiltration, insbesondere von T-Zellen, Monozyten bzw. Granulozyten, verwendet werden kann. Diese Gene alleine können bereits Kriterium für die Diagnostik entzündlicher Gelenkerkrankungen darstellen. Für die Osteoarthritis ergab sich eine vergleichsweise deutlich geringere Infiltration (**Fig. 5**, Hierarchische Clusteranalyse mit den Genen der Tabelle 5 zwischen RA, OA und Normalgewebe) Auch für eine Einteilung in Untergruppen von verschiedenen RA Patienten ergeben sich Infiltrationsunterschiede, die sowohl an dieser Auswahl von Genen als auch über den Vergleich der Infiltrationsanteile anhand der Markergene identifiziert werden können (**Fig. 6**). Die Signale dieser Gene können ohne vorherige Berechnung für die diagnostischen Untersuchungen eingesetzt werden, da sie vorwiegend durch Infiltration entstanden sind.

[0141] Die unter der 2. Bedingung gefundenen Geneinträge sind unten in Tabelle 6 angegeben. Sie stellen einen Genpool dar, der als Diagnostikum für die charakteristische An der Genregulation verwendet werden kann. Hier können Unterschiede zwischen einzelnen RA Patienten identifiziert werden und Untereinteilungen möglich werden. Hierzu gehören Einteilungen nach An der Arthritis, Stadium der Erkrankung, Krankheitsprognose, Zuweisung zu einer optimalen Therapieform, Abschätzung bzw. Verlaufskontrolle der Ansprechrate auf eine spezifische Therapie. Es ergeben sich somit neue Marker bzw. Markergruppen, die als molekulare Merkmale mit verschiedenen klinischen Merkmalen bzw. zu erwartenden Merkmalsentwicklungen korreliert sein können und deshalb diagnostische Bedeutung erlangen. Auch diese Signale könnten unmittelbar ohne vorherige Berechnung der Infiltration bzw. Aktivierung diagnostisch verwendet werden, da sie vorwiegend durch Aktivierung entstanden sind. Dennoch kann auch hier die Berechnung des Genaktivierung entstandenen Signalanteils eine Verbesserung in der Interpretation bewirken. Eine Unterteilung in Untergruppen ist in **Fig. 7** dargestellt.

[0142] Die unter der 3. Bedingung identifizierten Geneinträge sind in Tabelle 7 angegeben. Sie stellen ebenfalls eine diagnostisch bedeutenden Genpool dar, der aber zur Differenzierung von Infiltration und Aktivierung erst in Signale umgerechnet werden muß, die den Regulations- bzw. Infiltrationsanteil widerspiegeln (Auflösung der Gleichung 16 nach S'_{Probe}).

[0143] Es wurde der durch Regulation bedingte Signalanteil für die Gene bestimmt, die sich zusammengefaßt durch die 2. bzw. 3. Bedingung ergeben. Auch der durch Infiltration bedingte Anteil könnte in analoger Weise weiter untersucht werden. Nach Umrechnung in den regulierten Signalanteil wurde eine hierarchische Clusteranalyse durchgeführt. Das Ergebnis ist in **Fig. 8** dargestellt. Es ergeben sich offensichtlich Unterscheidungsmerkmale für die beiden Untergruppen RA 1, 2, 4, 5, 8, 10 und RA 3, 6, 7, 9. Zur Identifizierung der für die Unterscheidung relevanten Gene wurde eine t-Test Analyse auf die berechneten Signale von allen Genen aus

den Bedingungen 2 und 3 angewandt. Diese führte zu den in Tabelle 8 angegebene Geneinträgen, die eine Unterscheidung ermöglichen. **Fig. 9** stellt die Clusteranalyse und zugehörige Principle Component Analyse dar.

[0144] Anhand des dargestellten Beispiels wurde gezeigt, wie die Methode beiträgt, neue Bedeutungen für Gene und Gengruppen zu definieren, die sowohl für die Diagnostik als auch für die Entwicklung neuer Therapiestrategien bedeutsam sind. Es wurden damit Gene bzw. ihre Bedeutung in der Beurteilung entzündlicher Gelenkerkrankungen neu definiert hinsichtlich Infiltration und insbesondere hinsichtlich Aktivierung als Maß der aktiven Beteiligung und somit pathophysiologischen Bedeutung am Krankheitsprozeß.

Tabelle 1:

Verwendete Proben und Signaturen für die Etablierung der Berechnung

Probe bzw. Zelltyp	Information	Verwendung als
Normalspender Synovialgewebe	gesundes Gewebe ohne Infiltration	Kontrolle, Signatur eines fibroblastoiden Gewebes
Rheumatoide Arthritis Synovialgewebe	erkranktes Gewebe	zu untersuchende Probe
Normalspender Vollblut	gesundes „Gewebe“ mit variabler Zusammensetzung	Kontrolle

Rheumatoide Arthritis Vollblut	erkranktes „Gewebe“ mit variabler Zusammensetzung	zu untersuchende Probe
Arthrose Synovialgewebe	erkranktes Gewebe	Vergleich zwischen verschiedenen Erkrankungen
Normalspender CD4+ T-Zellen	Expressionsprofil im Normalzustand	CD4+ T-Zell Signatur
Rheumatoide Arthritis CD4+ T-Zellen	Expressionsprofil in der Krankheitssituation	Identifikation von regulierten T-Zell-Genen
Normalspender CD8+ T-Zellen	Expressionsprofil im Normalzustand	CD8+ T-Zell Signatur
Normalspender CD14+ Monozyten	Expressionsprofil im Normalzustand	Monozyten Signatur
Rheumatoide Arthritis CD14+ Monozyten	Expressionsprofil in der Krankheitssituation	Identifikation von regulierten Monozyten-Genen
Normalspender CD15+ Granulozyten	Expressionsprofil im Normalzustand	Granulozyten Signatur
Rheumatoide Arthritis CD15+ neutrophile Granulozyten	Expressionsprofil in der Krankheitssituation	Identifikation von regulierten Granulozyten-Genen
Knorpelzellen, Knorpelgewebe und kultivierte Synovialfibroblasten	unabhängige Gewebe	erweiterte Datengrundlagen für die Bestimmung des Dynamikbereichs

Tabelle 2: Verwendete Markergene

Tabelle 2A:

Auswahlliste für Monozyten-Markergene:

[0145] Die Gene wurden in allen untersuchten Monozyten-Populationen im Vergleich zu anderen Zelltypen oder nicht infiltrierten Geweben mindestens 4-fach erhöht exprimiert.

Affymetrix ID	Gen Symbol	Unigen	Name	Auswahl	S_min
201850_at	CAPG	Hs.82422	capping protein (actin filament), gelsolin-like	0	126,8
202295_s_at	CTSH	Hs.114931	cathepsin H	0	76,3

20/62

at	202944_at	NAGA	Hs.75372	N-acetylgalactosaminidase, alpha-adaptor-related protein complex 1, sigma subunit	0	77,8
at	203300_x_at	APIS2	Hs.40368		2	68,6
at	203922_s_at	CYBB	Hs.88974	cytochrome b-245, beta polypeptide (chronic granulomatous disease)	0	54,55
at	203923_s_at	CYBB	Hs.88974	cytochrome b-245, beta polypeptide (chronic granulomatous disease)	0	58,6
at	203932_at	HLA-DMB	Hs.11162	major histocompatibility complex, class II, DM beta	0	74,4
at	204057_at	ICSBP1	Hs.14453	interferon consensus sequence binding protein 1	0	78,95
at	204081_at	NRGN	Hs.232004	neurogranin (protein kinase C substrate, RC3)	0	110,4
at	204588_s_at	SLC7A7	Hs.194693	solute carrier family 7 (cationic amino acid transporter, y+ system), member 7	0	193,1
at	204619_s_at	CSPG2	Hs.434488	chondroitin sulfate proteoglycan 2 (versican)	0	34,7
at	205076_s_at	CRA	Hs.425144	cisplatin resistance associated	0	122,8
at	205552_s_at	OAS1	Hs.442936	2',5'-oligoadenylate synthetase 1, 40/46kDa	0	86,4
at	205685_at	CD86	Hs.27954	CD86 antigen (CD28 antigen ligand 2, B7-2 antigen)	2	46,9
at	205686_s_at	CD86	Hs.27954	CD86 antigen (CD28 antigen ligand 2, B7-2 antigen)	2	112,6
at	205789_at	CD1D	Hs.1799	CD1D antigen, d polypeptide	0	28,1
at	205859_at	LY86	Hs.184018	lymphocyte antigen 86	1	219,5
at	206120_at	CD33	Hs.83731	CD33 antigen (gp67)	1	124,8
at	206130_s_at	ASGR2	Hs.1259	asialoglycoprotein receptor 2	0	186,1
at	206214_at	PLA2G7	Hs.93304	phospholipase A2, group VII (platelet-activating factor acetylhydrolase, plasma)	1	16,8
at	206715_at	TFEC	Hs.125962	transcription factor EC	0	45,6
at	206743_s_at	ASGR1	Hs.12056	asialoglycoprotein receptor 1	0	55,5
at	206978_at	CCR2	Hs.511794	chemokine (C-C motif) receptor 2	1	69
at	208146_s_at	CPVL	Hs.95594	carboxypeptidase, vitellogenic-like	0	68,2
at	208450_at	LGALS2	Hs.113987	lectin, galactoside-binding, soluble, (galectin 2)	2	54,05
at	208771_s_at	LTA4H	Hs.81118	leukotriene A4 hydrolase	0	68,6
at	208890_s_at	PLXNB2	Hs.3989	plexin B2	0	188,5
at	209555_s_at	CD36	Hs.443120	CD36 antigen (collagen type I receptor, thrombospondin receptor)	1	116,85
at	210222_s_at	RTN1	Hs.99947	reticulin 1	1	37,2
at	210314_x_at	TNFSF13	Hs.54673	tumor necrosis factor (ligand) superfamily, member 13	0	54,9
at	210895_s_at	CD86	Hs.27954	CD86 antigen (CD28 antigen ligand 2, B7-2)	0	170,35

21/62

at	213385_at	CHN2	Hs.407520	antigen (chimerin (chimaerin) 2)	0	52,85
at	214058_at	MYCL1	Hs.437922	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived (avian)	1	61,25
at	217478_s_at	HLA-DMA	Hs.351279	major histocompatibility complex, class II, DM alpha	0	109,1
at	219574_at	FLJ20668	Hs.136900	hypothetical protein FLJ20668	0	32,55
at	219714_s_at	CACNA2D3	Hs.435112	calcium channel, voltage-dependent, alpha 2/delta 3 subunit	0	95,6
at	219806_s_at	FN5	Hs.416456	FN5 protein	0	121,8
at	220091_at	SLC2A6	Hs.244378	solute carrier family 2 (facilitated glucose transporter), member 6	0	103,95
at	220307_at	CD244	Hs.157872	natural killer cell receptor 2B4	0	252,45

Tabelle 2B:

Auswahlliste für 7-Zell-Markergene:

[0146] Die Gene wurden in allen untersuchten T-Zell-Populationen im Vergleich zu anderen Zelltypen oder nicht infiltrierten Geweben mindestens 8-fach erhöht exprimiert.

Affymetrix ID	Gen Symbol	Unigen	Name	Auswahl	S_min
202478_at	TRB2	Hs.155418	tribbles homolog 2	0	14,8
202524_s_at	SPOCK2	Hs.436193	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2	0	83,6
203385_at	DGKA	Hs.172690	diacylglycerol kinase, alpha 80kDa	0	86,95
203413_at	NELL2	Hs.79389	NEL-like 2 (chicken)	0	75
203685_at	BCL2	Hs.79241	B-cell CLL/lymphoma 2	0	49,5
203828_s_at	NK4	Hs.943	natural killer cell transcript 4	0	255,35
204777_s_at	MAL	Hs.80395	mal, T-cell differentiation protein	0	53,2
204890_s_at	LCK	Hs.1765	lymphocyte-specific protein tyrosine kinase	0	43,2
204891_s_at	LCK	Hs.1765	lymphocyte-specific protein tyrosine kinase	0	61,85
204960_at	PTPRCA	Hs.155975	protein tyrosine phosphatase, receptor type, C-associated protein	0	224,7
205255_x_at	TCF7	Hs.169294	transcription factor 7 (T-cell specific, HMG-box)	0	229,8
205456_at	CD3E	Hs.3003	CD3E antigen, epsilon polypeptide (TIT3 complex)	0	85,4
205488_at	GZMA	Hs.90708	granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)	0	53,3
205590_at	RASGRP1	Hs.189527	RAS guanyl releasing protein 1 (calcium and DAG-regulated)	0	2,6
205790_at	SCAP1	Hs.411942	src family associated phosphoprotein 1	0	91,65

22/62

at	205798_at	IL7R	Hs.362807	interleukin 7 receptor	0	82,5
at	205831_at	CD2	Hs.89476	CD2 antigen (p50), sheep red blood cell receptor	0	66,5
at	206150_at	TNFRSF7	Hs.355307	tumor necrosis factor receptor superfamily, member 7	0	65,6
at	206337_at	CCR7	Hs.1652	chemokine (C-C motif) receptor 7	0	66,65
at	206545_at	CD28	Hs.1987	CD28 antigen (Tp44)	0	25
at	206761_at	CD96	Hs.142023	CD96 antigen	0	54,4
at	206804_at	CD3G	Hs.2259	CD3G antigen, gamma polypeptide (TIT3 complex)	0	34,5
at	206828_at	TXK	Hs.29877	TXK tyrosine kinase	0	32,4
at	206980_s_at	FLT3LG	Hs.428	fms-related tyrosine kinase 3 ligand	0	109
at	206983_at	CCR6	Hs.46468	chemokine (C-C motif) receptor 6	0	14
at	207651_at	H963	Hs.159545	platelet activating receptor homolog	0	38,8
at	209504_s_at	PLEKHB1	Hs.445489	pleckstrin homology domain containing, family B (evectins) member 1	0	16,8
at	209602_s_at	GATA3	Hs.169946	GATA binding protein 3	0	23,9
at	209604_s_at	GATA3	Hs.169946	GATA binding protein 3	0	72,1
at	209670_at	TRA@	Hs.74647	T cell receptor alpha locus	1	93,7
at	209671_x_at	TRA@	Hs.74647	T cell receptor alpha locus	1	77,1
at	209871_s_at	APBA2	Hs.26468	amyloid beta (A4) precursor protein-binding family A, member 2 (X11-like)	0	26
at	209881_s_at	LAT	Hs.498997	linker for activation of T cells	0	237,8
at	210031_at	CD3Z	Hs.97087	CD3Z antigen, zeta polypeptide (TIT3 complex)	0	137,75
at	210038_at	PRKCQ	Hs.408049	protein kinase C, theta	0	159,95
at	210116_at	SH2D1A	Hs.151544	SH2 domain protein 1A, Duncan's disease, (lymphoproliferative syndrome)	0	45,9
at	210370_s_at	LY9	Hs.403857	lymphocyte antigen 9	0	322,7
at	210439_at	ICOS	Hs.56247	inducible T-cell co-stimulator	0	46,3
at	210607_at	FLT3LG	Hs.428	fms-related tyrosine kinase 3 ligand	0	19,75
at	210847_x_at	TNFRSF2	Hs.299558	tumor necrosis factor receptor superfamily, member 25	0	19,15
at	210915_x_at	---	Hs.419777	Homo sapiens T cell receptor beta chain, BV20S1 BJ1-5 BC1 mRNA, complete cds	1	79,2
at	210948_s_at	LEF1	Hs.44865	lymphoid enhancer-binding factor 1	0	57,55
at	210972_x_at	TRA@	Hs.74647	T cell receptor alpha locus	1	124,8
at	211005_at	LAT	Hs.498997	linker for activation of T cells	0	74,7
at	211272_s_at	DGKA	Hs.172690	diacylglycerol kinase, alpha 80kDa	0	54,15
at	211282_x_at	TNFRSF2	Hs.299558	tumor necrosis factor receptor superfamily, member 25	0	223,8
at	211339_s_at	ITK	Hs.211576	IL2-inducible T-cell kinase	0	22,3

23/62