

(12) **Österreichische Patentanmeldung**

(21) Anmeldenummer: A 1359/2007

(22) Anmeldetag: 30.08.2007

(43) Veröffentlicht am: 15.03.2009

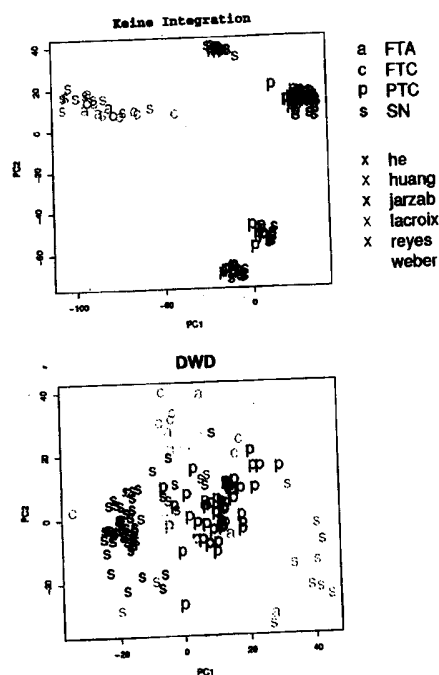
(51) Int. Cl.⁸: G01N 33/68 (2006.01),
G06F 19/00 (2006.01)

(73) Patentinhaber:

AUSTRIAN RESEARCH CENTERS GMBH
- ARC
A-1220 WIEN (AT)

(54) **SET VON TUMOR-MARKERN**

(57) Die vorliegende Erfindung sieht ein Set von Komponenten vor, welche für Tumormarker spezifisch sind, insbesondere von follikulärem Schilddrüsenkarzinom (FTC) und papillärem Schilddrüsenkarzinom (PTC), als auch ein Verfahren zum Identifizieren von Markern von einer genetischen Krankheit.



003116

- 35 -

Zusammenfassung

Die vorliegende Erfindung sieht ein Set von Komponenten vor, welche für Tumormarker spezifisch sind, insbesondere von follikulärem Schilddrüsenkarzinom (FTC) und papillärem Schilddrüsenkarzinom (PTC), als auch ein Verfahren zum Identifizieren von Markern von einer genetischen Krankheit.

NACHGEREICHT

Die vorliegende Erfindung betrifft den Bereich der Krebsdiagnose und diagnostische Mittel dafür.

Schilddrüsenknötchen sind in Gebieten mit Iodmangel, wie den europäischen Alpenregionen, endemisch, wo sie eine Prävalenz von 10-20% haben. Sie werden durch ihre Histologie in die 2 gutartigen Typen Struma nodosa (SN) und follikuläres Schilddrüsenadenom (FTA) und die bösartigen Gebilde follikuläres Schilddrüsenkarzinom (FTC), papilläres Schilddrüsenkarzinom (PTC), medulläres Schilddrüsenkarzinom (MTC) und anaplastisches Schilddrüsenkarzinom (ATC) klassifiziert. Herkömmlicherweise wird Unterscheidung zwischen gutartigen und bösartigen Schilddrüsenknötchen durch Szintigraphie und Feinnadelaspiration, gefolgt von Histologie durchgeführt. Trotz vieler Fortschritte bei der Diagnose und Therapie von Schilddrüsenknötchen und Schilddrüsenkrebs haben diese Verfahren einen gut bekannten Mangel an Spezifität, insbesondere für die Unterscheidung zwischen FTA und FTC, was dazu führt, dass eine Anzahl von Patienten unnötigerweise auf bösartige Krankheit behandelt wird.

In Anbetracht der diagnostischen Beschränkungen von vorhergehenden Verfahren, insbesondere Feinnadelaspiration gefolgt von Zytologie, haben mannigfaltige Forscher Expressionsprofilierungsstudien mit der Hoffnung durchgeführt, neue diagnostische Werkzeuge zu identifizieren. Solche Analysen versuchen unterschiedlich exprimierte Gene mit einer wichtigen Rolle bei Krankheitsentwicklung oder Fortschreiten unter Verwendung von Transkriptlevel Expressions-Profiling-Technologien in großem Maßstab wie cDNA Microarrays, Oligonukleotid-Arrays und Serialer Analyse von Genexpression (SAGE) zu identifizieren. Typischerweise werden Duzende oder Hunderte von Genen identifiziert, wobei von vielen erwartet wird, dass sie falsche Positive sind, und nur ein kleiner Bruchteil als diagnostische/prognostische Marker oder therapeutische Ziele brauchbar ist (Griffith et al., J. Clin. Oncol. 24 (31): 5043-5051 (2006)).

Bei anderen Krebstypen ist gezeigt worden, dass Genexpressions-Profiling wesentlichen Wert zur Unterscheidung der verschiedenen klinisch relevanten Tumorgebilde beitragen kann. US-2006/183141 A beschreibt z.B. Klassifizierung von Tumormarkern aus einer Kern-Serumantwort-Signatur. Verschiedene Studien haben versucht, die verschiedenen Gebilde von Schilddrüsenkarzi-

nom auf der Basis ihrer Genexpressionsprofile zu klassifizieren, von denen jedes zwischen 2 der 5 Gebilde unterscheidet. Jedoch haben die Studien keine oder nur sehr wenige Gene gemeinsam und Anwenden eines Klassifikators von einer Studie auf die Daten von einer anderen Studie ergeben im Allgemeinen schlechte Klassifizierungsergebnisse.

Es ist ein Ziel der vorliegenden Erfindung, verlässliche unterscheidende Marker für die Diagnose von Krebs vorzusehen, insbesondere um gutartige Schilddrüsenknötchen von bösartigem follikulärem Schilddrüsenkarzinom (FTC) und papillärem Schilddrüsenkarzinom (PTC) zu unterscheiden.

Daher sieht die vorliegende Erfindung ein Set von Komponenten vor, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern PI-1 bis PI-33, PII-1 bis PII-64, PIII-1 bis PIII-70, fi-1 bis fi-147, PIV-1 bis PIV-9, vorzugsweise PIV-4 oder PIV-5, und PV-1 bis PV-11, vorzugsweise PV-1, PV-2 und PV-4 bis PV-11. Diese Tumormarker betreffen verschiedene Gene, welche aberrierend in Tumoren exprimiert werden, und werden in Tabellen 1 bis 6 angegeben und können durch ihr Genidentifizierungszeichen, ihren deskriptiven Gennamen, aber am eindeutigsten durch ihre UniGeneID oder ihre Eingangsnummer identifiziert werden, welche sich auf spezifische Sequenzen in üblichen Sequenz-Datenbanken bezieht, wie NCBI GenBank, EMBL-EBI Database, EnSEMBL oder die DNA Data Bank von Japan. Diese Marker sind in Form von bevorzugten Sets (PI bis PV, FI) identifiziert worden, aber können für das Set der Erfindung in jeder Form als Ziele vereinigt werden.

Tabelle 1: PTC-Markerset PI-1 bis PI-33

Nummer PI-	Marker-gen	Beschreibung des Gens	Eingangsnummer	UniGeneID
1	BBS9	Bardet-Bied Syndrom 9	NM_198428 NM_001033605 NM_001033604 NM_014451	Hs.372360
2	C13orf1	Chromosom 13 offenes Leseraster 1	NM_020456	Hs.44235
3	CB-FA2T3	Kern-bindender Faktor, Runt Domäne, alpha Untereinheit 2	NM_005187 NM_175931	Hs.513811
4	CDT1	Chromatin Lizenzierung und DANN Replikationsfaktor 1	NM_030928	Hs.122908

NACHGEREICHT

00316

- 3 -

5	CRK	V-crk Sarkom Virus CT10 Onkogen Homolog (Avian)	NM_016823 NM_005206	Hs.638121
6	CTPS	CTP Synthase	NM_001905	Hs.473087
7	DAPK2	Tod-assoziierte Proteinkinase 2	NM_014326	Hs.237886
8	EIF5	Eukaryotischer Translation-Initiationsfaktor 5	NM_001969 NM_183004	Hs.433702
9	EREG	Epiregulin	NM_001432	Hs.115263
10	GK	Glycerinkinase	NM_203391 NM_000167	Hs.1466
11	GPAT-CH8	G Patch Domain enthaltend 8	NM_001002909	Hs.463129
12	HDGF	Hepatoma-abgeleiteter Wachstumsfaktor (High-Mobility-Group Protein 1-ähnlich)	NM_004494	Hs.506748
13	IRF2BP1	Interferon regulatorischer Faktor 2 Bindungsprotein 1	NM_015649	Hs.515477
14	KRT83	Keratin 83	NM_002282	Hs.661428
15	MYOD1	Myogene Differenzierung 1	NM_002478	Hs.181768
16	NME6	Nicht-metastatische Zellen 6, Protein exprimiert in (Nukleosid-Diphosphat Kinase)	NM_005793	Hs.465558
17	POLE3	Polymerase (DNA dirigiert), Epsilon 3 (p17 Untereinheit)	NM_017443	Hs.108112
18	PPP1R13B	Proteinphosphatase 1, regulatorische (Inhibitor) Untereinheit 13B	NM_015316	Hs.436113
19	PRPH2	Peripherin 2 (retinale Degeneration, langsam)	NM_000322	Hs.654489
20	RASSF7	RAS-Assoziations (RalGDS/AF-6) Domänen Familie 7	NM_003475	Hs.72925
21	ROCK2	Rho-assoziiert, Coiled-Coil enthaltende Proteinkinase 2	NM_004850	Hs.591600
22	RTN1	Reticulon 1	NM_021136 NM_206857 NM_206852	Hs.368626
23	S100B	S100 Calcium-bindendes Protein B	NM_006272	Hs.422181
24	SLIT2	Slit homolog 2 (Drosophila)	NM_004787	Hs.29802
25	SNRNPB2	Kleines nukleares Ribonucleoprotein Polypeptid B"	NM_003092 NM_198220	Hs.280378
26	SPAG7	Sperma-assoziiertes Antigen 7	NM_004890	Hs.90436
27	STAU1	Staufen, RNA Bindungsprotein, Homolog 1 (Drosophila)	NM_017453 NM_001037328 NM_004602 NM_017452 NM_017454	Hs.596704
28	SUPT5H	Suppressor von Ty 5 Homolog (S. cerevisiae)	NM_003169	Hs.631604
29	TBX10	T-Box 10	NM_005995	Hs.454480
30	TLK1	Tousled-ähnliche Kinase 1	NM_012290	Hs.655640
31	TM4SF4	TransMembran 4 L six Familie Mitglied 4	NM_004617	Hs.133527
32	TXN	Thioredoxin	NM_003329 NM_005659	Hs.435136
33	UFD1L	Ubiquitin fusion degradation 1 ähnlich (Hefe)	NM_001035247	Hs.474243

NACHGEREICHT

Tabelle 2: PTC Markerset PII-1 bis PII-64

Num- mer P II-	Marker- gen	Beschreibung des Gens	Eingangsnummer	UniGeneID
1	ADH1B	Alkohol Dehydrogenase IB (Klasse I), beta Polypeptid	NM_000668	Hs.4
2	AGR2	Anterior Gradient Homolog 2 (Xenopus laevis)	NM_006408	Hs.530009
3	AGTR1	Angiotensin II Rezeptor, Typ 1	NM_031850 NM_004835 NM_009585 NM_032049	Hs.477887
4	AGTR1	Angiotensin II Rezeptor, Typ 1	NM_000685	Hs.654382
5	ALDH1A1	Aldehyd Dehydrogenase 1 Familie, Mitglied A1	NM_000689	Hs.76392
6	ALDH1A3	Aldehyd Dehydrogenase 1 Familie, Mitglied A3	NM_000693	Hs.459538
7	AMIGO2	Adhäsionsmolekül mit Ig-ähnlicher Domäne 2	NM_181847	Hs.121520
8	ATP2C2	ATPase, Ca++ transportierend, Typ 2C, Mitglied 2	NM_014861	Hs.6168
9	BID	BH3 interagierende Domäne Tod-Agonist	NM_197966 NM_001196 NM_197967	Hs.591054
10	C7orf24	Chromosom 7 offenes Leseraster 24	NM_024051	Hs.530024
11	CA4	Carbonische Anhydrase IV	NM_000717	Hs.89485
12	CCL21	Chemokin (C-C Motiv) Ligand 21	NM_002989	Hs.57907
13	CD55	CD55 Molekül, Zerfall-Beschleunigungsfaktor für Komplement (Cromer blood group)	NM_000574	Hs.527653
14	CDH16	Cadherin 16, KSP-Cadherin	NM_004062	Hs.513660
15	CDH3	Cadherin 3, Typ 1, P-Cadherin (plazental)	NM_133458 NM_001793	Hs.461074
16	CFI	Komplement-Faktor I	NM_000204	Hs.312485
17	CHI3L1	Chitinase 3-ähnlich 1 (Cartilage Glycoprotein-39)	NM_001276	Hs.382202
18	CHST2	Kohlenhydrat (N-Acetylglucosamin-6-O) Sulfotransferase 2	NM_004267	Hs.8786
19	CITED2	Cbp/p300-interagierender Transaktivator, mit Glu/Asp-reicher carboxyterminaler Domäne, 2	NM_006079	Hs.82071
20	CLCNKB	Chlorid-Kanal Kb	NM_000085	Hs.352243
21	COMP	Cartilage oligomeres Matrixprotein	NM_000095	Hs.1584
22	CTSH	Cathepsin H	NM_004390 NM_148979	Hs.148641
23	DIO2	Deiodinase, Iodthyronin, Typ II	NM_013989 NM_000793 NM_001007023	Hs.202354

NACHGEFERTIGT

24	DIRAS3	DIRAS Familie, GTP-bindend RAS-ähnlich 3	NM_004675	Hs.194695
25	DUSP4	Dual spezifische Phosphatase 4	NM_057158 NM_001394	Hs.417962
26	DUSP5	Dual spezifische Phosphatase 5	NM_004419	Hs.2128
27	EDN3	Endothelin 3	NM_207032 NM_207034 NM_207033 NM_000114	Hs.1408
28	ENTPD1	Ectonucleosid Triphosphat Diphosphohydrolase 1	NM_001776 NM_001098175	Hs.576612
29	FHL1	Vier und eine Halbe LIM Domänen 1	NM_001449	Hs.435369
30	GDF15	Wachstumsdifferenzierungsfaktor 15	NM_004864	Hs.616962
31	GPM6A	Glycoprotein M6A	NM_201591 NM_005277 NM_201592	Hs.75819
32	HBA1	Hämoglobin, alpha 1	NM_000558	Hs.449630
33	IRS1	Insulin Rezeptor Substrat 1	NM_005544	Hs.471508
34	KCNJ2	Einwärts gleichrichtender Kaliumkanal, Unterfamilie J, Mitglied 2	NM_000891	Hs.1547
35	KCNN4	Calcium-aktivierter Kaliumkanal von mittlerer/niedriger Leit- fähigkeit, Unterfamilie N, Mitglied 4	NM_002250	Hs.10082
36	KLK10	Kallikrein-bezogene Peptidase 10	NM_002776 NM_001077500 NM_145888	Hs.275464
37	LAMB3	Laminin, beta 3	NM_001017402 NM_000228	Hs.497636
38	LCN2	Lipocalin 2 (Onkogen 24p3)	NM_005564	Hs.204238
39	LMOD1	Leiomodin 1 (Glattmuskel)	NM_012134	Hs.519075
40	MATN2	Matrilin 2	NM_002380 NM_030583	Hs.189445
41	MPPED2	Metallophosphoesterase Domäne enthaltend 2	NM_001584	Hs.289795
42	MVP	Major-Vault-Protein	NM_017458 NM_005115	Hs.632177
43	NELL2	NEL-ähnlich 2 (Huhn)	NM_006159	Hs.505326
44	NFE2L3	Nuklearer Faktor (Erythroid-abgeleitet 2)-ähnlich 3	NM_004289	Hs.404741
45	NPC2	Niemann-Pick Krankheit, Typ C2	NM_006432	Hs.433222
46	NRCAM	Neuronales Zelladhäsionsmolekül	NM_001037132 NM_005010 NM_001037133	Hs.21422
47	NRIP1	Nuklearrzeptor interagierendes Protein 1	NM_003489	Hs.155017
48	PAPSS2	3'-Phosphoadenosin 5'-Phosphosulfat Synthase 2	NM_001015880 NM_004670	Hs.524491
49	PDLIM4	PDZ und LIM Domäne 4	NM_003687	Hs.424312
50	PDZ- K1IP1	PDZK1 interagierendes Protein 1	NM_005764	Hs.431099
51	PIP3-E	Phosphoinositid-bindendes Protein PIP3-E	NM_015553	Hs.146100
52	PLAU	Plasminogen Aktivator, Urokinase	NM_002658	Hs.77274
53	PRSS2	Protease, Serin, 2 (Trypsin 2)	NM_002770	Hs.622865
54	PRSS23	Protease, Serin, 23	NM_007173	Hs.25338
55	RAP1- GAP	RAP1 GTPase aktivierendes Protein	NM_002885	Hs.148178
56	S100A11	S100 Calcium bindendes Protein A11	NM_005620	Hs.417004

NACHGEREICHT

57	SFTPB	Surfaktant, pulmonal-assoziiertes Protein B	NM_198843 NM_000542	Hs.512690
58	SLPI	Sekretorischer Leukozyten Peptidase Inhibitor	NM_003064	Hs.517070
59	SOD3	Superoxid Dismutase 3, extrazellulär	NM_003102	Hs.2420
60	SPINT1	Serin-Peptidase Inhibitor, Kunitz Typ 1	NM_181642 NM_003710 NM_001032367	Hs.233950
61	SYNE1	Spectrin repeat containing, nuclear envelope 1	NM_182961 NM_033071 NM_015293 NM_133650	Hs.12967
62	TACSTD2	Tumor-assoziiertes Calcium Signal-Transducer 2	NM_002353	Hs.23582
63	UPP1	Uridin-Phosphorylase 1	NM_181597 NM_003364	Hs.488240
64	WASF3	WAS Proteinfamilie, Mitglied 3	NM_006646	Hs.635221

Tabelle 3: PTC Markerset PIII-1 bis PIII-70

Num-mer P III-	Markergen	Beschreibung des Gens	Eingangsnummer	UniGeneID
1	APOE	Apolipoprotein E	NM_000041	Hs.654439
2	ATIC	5-Aminoimidazol-4-carboxamid Ribonukleotid Formyltransferase/IMP Cyclohydrolase	NM_004044	Hs.90280
3	BASP1	Brain abundant, Membran-gebundenes Signalprotein 1	NM_006317	Hs.201641
4	C9orf61	Chromosom 9 offenes Leseraster 61	NM_004816	Hs.118003
5	CCL13	Chemokin (C-C Motiv) Ligand 13	NM_005408	Hs.414629
6	CD36	CD36 Molekül (Thrombospondin-Rezeptor)	NM_001001548 NM_001001547 NM_000072	Hs.120949
7	CDH6	Cadherin 6, Typ 2, K-Cadherin (fötal Niere)	NM_004932	Hs.171054
8	CFB	Komplement-Faktor B	NM_001710	Hs.69771
9	CFD	Komplement-Faktor D (Adipsin)	NM_001928	Hs.155597
10	CLDN10	Claudin 10	NM_182848 NM_006984	Hs.534377
11	COL11A1	Collagen, Typ XI, alpha 1	NM_080629 NM_001854 NM_080630 NM_005203 NM_080804 NM_080798 NM_080803 NM_080802 NM_080799 NM_080800 NM_080801 NM_080808 NM_080809 NM_080805 NM_080807 NM_080806 NM_080811 NM_080810 NM_080812 NM_080813 NM_080814 NM_080815	Hs.523446
12	COL13A1	Collagen, Type XIII, alpha 1	NM_080815	Hs.211933
13	CORO2B	Coronin, Actin-bindendes Protein, 2B	NM_006091	Hs.551213
14	CRLF1	Cytokin-Rezeptor-ähnlicher Faktor 1	NM_004750	Hs.114948
15	CXorf6	Chromosom X offenes Leseraster 6	NM_005491	Hs.20136
16	DDB2	Defekt-spezifisches DNA-Bindungsprotein 2, 48kDa	NM_000107	Hs.655280
17	DPP6	Dipeptidyl-Peptidase 6	NM_001039350 NM_130797 NM_001936	Hs.490684
18	ECM1	Extrazelluläres Matrixprotein 1	NM_004425 NM_022664	Hs.81071

NACHGEREICHT

19	EFEMP1	EGF-enhaltendes Fibulin-ähnliches extrazelluläres Matrixprotein 1	NM_004105 NM_001039348 NM_001039349	Hs.76224
20	ESRRG	Estrogen-bezogener Rezeptor gamma	NM_206594 NM_001438 NM_206595	Hs.444225
21	ETHE1	Ethylmalonsäure-Enzephalopathie 1	NM_014297	Hs.7486
22	FAS	Fas (TNF-Rrezeptor Überfamilie, Mitglied 6)	NM_000043 NM_152872 NM_152871 NM_152873 NM_152875 NM_152874 NM_152877 NM_152876	Hs.244139
23	FMOD	Fibromodulin	NM_002023	Hs.519168
24	GABBR2	Gamma-Aminobuttersäure (GABA) B Rezeptor, 2	NM_005458	Hs.198612
25	GALE	UDP-Galactose-4-Epimerase	NM_000403 NM_001008216	Hs.632380
26	GATM	Glycine-Amidinotransferase (L-Arginine:Glycin-Amidinotransferase)	NM_001482	Hs.75335
27	GDF10	Wachstumsdifferenzierungsfaktor 10	NM_004962	Hs.2171
28	GHR	Wachstumshormonrezeptor	NM_000163	Hs.125180
29	GPC3	Glypican 3	NM_004484	Hs.644108
30	ICAM1	Interzelluläres Adhäsionsmolekül 1 (CD54), humaner Rhinovirus-Rezeptor	NM_000201	Hs.643447
31	ID3	Inhibitor von DNA bindendem 3, dominant negativem Helix-Schleife-Helix Protein	NM_002167	Hs.76884
32	IER2	Immediate-Early-Respons 2	NM_004907	Hs.501629
33	IGFBP6	Insulin-ähnliches Wachstumsfaktor bindendes Protein 6	NM_002178	Hs.274313
34	IQGAP2	IQ Motiv enthaltendes GTPase aktivierendes Protein 2	NM_006633	Hs.291030
35	ITGA2	Integrin, alpha 2 (CD49B, alpha 2 Untereinheit von VLA-2 Rezeptor)	NM_002203	Hs.482077
36	ITGA3	Integrin, alpha 3 (Antigen CD49C, alpha 3 Untereinheit von VLA-3 Rezeptor)	NM_002204 NM_005501	Hs.265829
37	ITM2A	Integrales Membranprotein 2A	NM_004867	Hs.17109
38	KIAA0746	KIAA0746 Protein	NM_015187	Hs.479384
39	LRIG1	Leucin-reiche Wiederholungen und Immunoglobulin-ähnliche Domänen 1	NM_015541	Hs.518055
40	LRP2	niedrigdichtes Lipoprotein-bezogenes Protein 2	NM_004525	Hs.470538
41	LY6E	Lymphozyten Antigen 6 Komplex, Lokus E	NM_002346	Hs.521903
42	MAPK13	Mitogen-aktivierte Proteinkinase 13	NM_002754	Hs.178695
43	MDK	Midkin (Neurit wachstumsfördernder Faktor 2)	NM_001012334 NM_001012333 NM_002391	Hs.82045
44	MLLT11	Myeloid/Lymphoid oder Leukämie von gemischter Abstammung (trithorax homolog, Drosophila)	NM_006818	Hs.75823
45	MMRN1	Multimerin 1	NM_007351	Hs.268107
46	MTMR11	Myotubularin-bezogenes Protein 11	NM_181873	Hs.425144
47	MXRA8	Matrix-Remodelling assoziiertes 8	NM_032348	Hs.558570
48	NAB2	NGFI-A Bindungsprotein 2 (EGR1 Bindungsprotein 2)	NM_005967	Hs.159223
49	NMU	Neuromedin U	NM_006681	Hs.418367
50	OCA2	Okulokutaner Albinismus II (pink-eye dilution homolog, Maus)	NM_000275	Hs.654411
51	PDE5A	Phosphodiesterase 5A, cGMP-spezifisch	NM_001083 NM_033430 NM_033437	Hs.647971
52	PLAG1	Pleiomorphes Adenom-Gen 1	NM_002655	Hs.14968

53	PLP2	Proteolipidprotein 2 (Kolonepithelium-angereichert)	NM_002668	Hs.77422
54	PLXNC1	Plexin C1	NM_005761	Hs.584845
55	PRKCQ	Proteinkinase C, theta	NM_006257	Hs.498570
56	PRUNE	Prune homolog (Drosophila)	NM_021222	Hs.78524
57	RAB27A	RAB27A, Mitglied RAS Onkogen Familie	NM_004580 NM_183234 NM_183235 NM_183236	Hs.654978
58	RYR2	Ryanodin-Rezeptor 2 (Herz)	NM_001035	Hs.109514
59	SCEL	Sciellin	NM_144777 NM_003843	Hs.534699
60	SE-LENBP1	Selenium bindendes Protein 1	NM_003944	Hs.632460
61	SORBS2	Sorbin und SH3 Domäne enthaltend 2	NM_021069 NM_003603	Hs.655143
62	STMN2	Stathmin-ähnlich 2	NM_007029	Hs.521651
63	TBC1D4	TBC1 Domänen-Familie, Mitglied 4	NM_014832	Hs.210891
64	TM4SF4	Transmembran 4 L six Familie Mitglied 4	NM_004617	Hs.133527
65	TNC	Tenascin C (Hexabrachion)	NM_002160	Hs.143250
66	TPD52L1	Tumorprotein D52-ähnlich 1	NM_001003395 NM_003287 NM_001003396 NM_001003397	Hs.591347
67	TSC22D1	TSC22 Domänen-Familie, Mitglied 1	NM_183422 NM_006022	Hs.507916
68	TTC30A	Tetratricopeptid Wiederholungsdomäne 30A	NM_152275 NM_003383	Hs.128384
69	VLDLR	Sehr niedriglichter Lipoproteinrezeptor	NM_001018056	Hs.370422
70	WFS1	Wolfram-Syndrom 1 (Wolframin)	NM_006005	Hs.518602

Tabelle 4: FTC Markerset FI-1 bis FI-147

Nummer FI-	Marker gen	Beschreibung des Gens	Eingangsnummer	UniGeneID
1	AATF	Apoptose antagonisierender Transkriptionsfaktor	NM_012138	Hs.195740
2	ACOX3	Acyl-Coenzym A Oxidase 3, Pristanoyl	NM_003501	Hs.479122
3	AHDC1	AT-Haken, DNA bindendes Motiv, enthaltend 1	NM_001029882	Hs.469280
4	ALAS2	Aminolevulinat, delta-, Synthase 2 (sideroblastische/hypochrome Anämie)	NM_000032 NM_001037968 NM_001037967 NM_001037969	Hs.522666
5	ALKBH1	AlkB, Alkylierung Repair Homolog 1 (E. coli)	NM_006020	Hs.94542
6	ANGPTL2	Angiopoietin-ähnlich 2	NM_012098	Hs.653262
7	AP2A2	Adaptor-bezogener Proteinkomplex 2, alpha 2 Untereinheit	NM_012305	Hs.19121
8	APOBEC3G	Apolipoprotein B mRNA editierendes Enzym, katalytisch Polypeptid-ähnlich 3G	NM_021822	Hs.660143
9	APRIN	Androgen-induzierter Proliferationshemmer	NM_015032	Hs.693663
10	ARNT	Aryl Kohlenwasserstoff Rezeptor nuklearer Translokator	NM_001668 NM_178427 NM_178426	Hs.632446
11	AZGP1	Alpha-2-Glycoprotein 1, Zink-bindend	NM_001185	Hs.546239
12	BAT2D1	BAT2 Domäne enthaltend 1	NM_015172	Hs.494614
13	BATF	Basischer Leukin-Zipper Transkriptionsfaktor, ATF-ähnlich	NM_006399	Hs.509964

NACHGEFÜRCHT

14	BPHL	Biphenyl hydrolase-ähnlich (Serin-Hhydro-lase)	NM_004332	Hs.10136
15	C13orf1	Chromosom 13 offenes Leseraster 1	NM_020456	Hs.44235
16	C14orf1	Chromosom 14 offenes Leseraster 1	NM_007176	Hs.15106
17	C2orf3	Chromosom 2 offenes Leseraster 3	NM_003203	Hs.303808
18	CBFB	Kern-bindender Faktor, beta Untereinheit	NM_001755 NM_022845	Hs.460988
19	CBR3	Carbonyl-Rreduktase 3	NM_001236	Hs.154510
20	CBX5	Chromobox Homolog 5 (HP1 alpha Homo-log, Drosophila)	NM_012117 NM_057749	Hs.632724
21	CCNE2	Cyclin E2	NM_057735	Hs.567387
22	CD46	CD46 Molekül, Komplement regulator- isches Protein	NM_002389 NM_172354 NM_172351 NM_172355 NM_172352 NM_172359 NM_172357 NM_172360 NM_153826 NM_172358 NM_172356 NM_172353 NM_172361 NM_172350	Hs.510402
23	CHPF	Chondroitin polymerisierender Faktor	NM_024536	Hs.516711
24	CHST3	Kohlenhydrat (Chondroitin 6) Sulfotrans- ferase 3	NM_004273	Hs.158304
25	CLCN2	Chlorid-Kanal 2	NM_004366	Hs.436847
26	CLCN4	Chlorid-Kanal 4	NM_001830	Hs.495674
27	CLIC5	Chlorid intrazellulärer Kanal 5	NM_016929	Hs.485489
28	CNOT2	CCR4-NOT Transkriptionskomplex, Unter- einheit 2	NM_014515	Hs.133350
29	COPS6	COP9 konstitutive photomorphogene ho- mologe Untereinheit 6 (Arabidopsis)	NM_006833	Hs.15591
30	CPZ	Carboxypeptidase Z	NM_001014448 NM_001014447 NM_003652	Hs.78068
31	CSK	C-src Tyrosinkinase	NM_004383	Hs.77793
32	CTDP1	CTD (Carboxy-terminale Domäne, RNA Po- lymerase II, Polypeptide A) Phosphatase, Untereinheit 1	NM_004715 NM_048368	Hs.465490
33	DDEF2	Entwicklung and Differenzierung ver- stärkender Faktor 2	NM_003887	Hs.555902
34	DKFZ- P586H2123	Regenerations-assoziierte Muskelprotease	NM_015430 NM_001001991	Hs.55044
35	DLG2	Scheiben, großes Homolog 2, Chapsyn-110 (Drosophila)	NM_001364	Hs.654862
36	DPAGT1	Dolichyl-Phosphat (UDP-N-Acetylglucosa- min) N-Acetylglucosaminphosphotransfera- se 1 (GlcNAc-1-P Transferase)	NM_001382 NM_203316	Hs.524081
37	DSCR1	Down Syndrom kritische Region Gen 1	NM_004414 NM_203418 NM_203417	Hs.282326
38	DUSP8	Dual-spezifische Phosphatase 8	NM_004420	Hs.41688

39	EI24	Etoposid-induzierte 2.4 mRNA	NM_004879 NM_001007277	Hs.643514
40	ENOSF1	Enolase Überfamilie Mitglied 1	NM_017512	Hs.369762
41	ERCC1	Excision-Repair kreuz-komplementierende Nager Repair-Deficiency Komplementierungsgruppe 1 (schließt überlappende Antisense-Sequenz ein)	NM_202001 NM_001983	Hs.435981
42	ERCC3	Excision Repair kreuz-komplementierende Nager Repair-Deficiency, Komplementierungsgruppe 3 (Xeroderma pigmentosum Gruppe B komplementierend)	NM_000122	Hs.469872
43	ERH	Verstärker von rudimentärem Homolog (Drosophila)	NM_004450	Hs.509791
44	F13A1	Koagulationsfaktor XIII, A1 Polypeptid	NM_000129	Hs.335513
45	FAM20B	Familie mit Sequenz-Ähnlichkeit 20, Mitglied B	NM_014864	Hs.5737
46	FBP1	Fructose-1,6-bisphosphatase 1	NM_000507	Hs.494496
47	FCGR2A	Fc Fragment von IgG, niedrige Affinität IIa, Rezeptor (CD32)	NM_021642	Hs.352642
48	FGF13	Fibroblast-Wachstumsfaktor 13	NM_004114 NM_033642	Hs.6540
49	FGFR1OP	FGFR1 Onkogen-Partner	NM_007045 NM_194429	Hs.487175
50	FLNC	Filamin C, gamma (Actin Bindungsprotein 280)	NM_001458	Hs.58414
51	FMO5	Flavin enthaltende Monooxygenase 5	NM_001461	Hs.642706
52	FRY	Furry Homolog (Drosophila)	NM_023037	Hs.591225
53	GADD45G	Wachstumsarrest und DNA-schädigungsinduzierbar, gamma	NM_006705	Hs.9701
54	GCH1	GTP Cyclohydrolase 1 (Dopa-responsive Dystonie)	NM_000161 NM_001024024 NM_001024070 NM_001024071	Hs.86724
55	GFRA1	GDNF Familie Rezeptor alpha 1	NM_005264 NM_145793	Hs.591913
56	GLB1	Galactosidase, beta 1	NM_001039770 NM_000404 NM_001079811	Hs.443031
57	GOLGA8A	Golgi Autoantigen, Golgin Unterfamilie a, 8A	NM_181077 NM_001023567	Hs.182982
58	HCLS1	Hämatopoietisches Zell-spezifisches Lyn-Substrat 1	NM_005335	Hs.14601
59	HDGF	Hepatom-abgeleiteter Wachstumsfaktor (High-Mobility-Group Protein 1-ähnlich)	NM_004494	Hs.506748
60	HRC	Histidin-reiches Calcium-Bindungsprotein	NM_002152	Hs.436885
61	ICMT	Isoprenylcystein Carboxyl-Methyltransferase	NM_012405	Hs.562083
62	IFNA5	Interferon, alpha 5	NM_002169	Hs.37113
63	IGF2BP3	Insulin-ähnlicher Wachstumsfaktor 2 mRNA Bindungsprotein 3	NM_006547	Hs.648088
64	IL12A	Interleukin 12A (natürlicher Killerzellen stimulierender Faktor 1, zytotoxischer Lymphozyten Reifungsfaktor 1, p35)	NM_000882	Hs.673
65	ITI2H	Inter-alpha (Globulin) Inhibitor H2	NM_002216	Hs.75285
66	ITPKC	Inositol 1,4,5-Trisphosphat 3-Kinase C	NM_025194	Hs.515415
67	JMJD2A	Jumonji Domäne enthaltend 2A	NM_014663	Hs.155983

68	KCNJ15	Einwärts gleichrichtender Kaliumkanal Unterfamilie J, Mitglied 15	NM_170736 NM_002243 NM_170737	Hs.411299
69	KCTD12	Kaliumkanal Tetramerisationsdomäne enthaltend 12	NM_138444	Hs.693617
70	KIAA0652	KIAA0652	NM_014741	Hs.410092
71	KIAA0913	KIAA0913	NM_015037	Hs.65135
72	KLKB1	Kallikrein B, Plasma (Fletcher-Faktor) 1	NM_000892	Hs.646885
73	KRT37	Keratin 37	NM_003770	Hs.673852
74	LAMB3	Laminin, beta 3	NM_001017402 NM_000228	Hs.497636 Hs.694758 Hs.649524
75	LPHN3	Latrophilin 3	NM_015236	
76	LRIG1	Leuzin-reiche Wiederholungen und Immunoglobulin-ähnliche Domänen 1	NM_015541	Hs.518055
77	LSR	Lipolyse-stimulierender Lipoprotein-Rezeptor	NM_205834 NM_015925 NM_205835	Hs.466507
78	MANBA	Mannosidase, beta A, lysosomal	NM_005908	Hs.480415
79	MAP7	Mikrotubulus-assoziiertes Protein 7	NM_003980	Hs.486548
80	MAPKAPK5	Mitogen-aktivierte Proteinkinase-aktivierte Proteinkinase 5	NM_139078 NM_003668	Hs.413901
81	MET	Met Proto-Onkogen (Hepatozyten Wachstumsfaktor Rezeptor)	NM_000245	Hs.132966
82	MMP14	Matrix-Metallopeptidase 14 (Membran-insertiert)	NM_004995	Hs.2399
83	MX1	Myxovirus (Influenza-Virus) Resistenz 1, Interferon-induzierbares Protein p78 (Maus)	NM_002462	Hs.517307
84	MYL9	Myosin, leichte Kette 9, regulatorisch	NM_006097 NM_181526	Hs.504687
85	MYO9B	Myosin IXB	NM_004145	Hs.123198
86	NCOR1	Nuklearer Rezeptor Co-Repressor 1	NM_006311	Hs.462323
87	NDRG4	NDRG Familie Mitglied 4	NM_020465 NM_022910	Hs.322430
88	NDUFA5	NADH Dehydrogenase (Ubiquinon) 1 alpha Subkomplex, 5, 13kDa	NM_005000	Hs.651219
89	NEUROD2	Neurogene Differenzierung 2	NM_006160	Hs.322431
90	NFKB2	Nuklearer Faktor von kappa leichtem Polypeptid Gene-Verstärker in B-Zellen 2 (p49/p100)	NM_001077494 NM_001077493 NM_002502	Hs.73090
91	NME6	Nicht-metastatische Zellen 6, Protein exprimiert in (Nucleoside-Diphosphat-Kinase)	NM_005793	Hs.465558
92	NPY1R	Neuropeptid Y Rezeptor Y1	NM_000909	Hs.519057
93	NUP50	Nucleoporin 50kDa	NM_007172 NM_153645	Hs.475103
94	PDGFRA	Plättchen-abgeleiteter Wachstumsfaktor Rezeptor, alpha Polypeptid	NM_006206	Hs.74615
95	PDHX	Pyruvat-Dehydrogenase-Komplex, Bestandteil X	NM_003477	Hs.502315
96	PDLIM1	PDZ und LIM Domäne 1 (elfin)	NM_020992	Hs.368525
97	PEX1	Peroxisom-Biogenese-Faktor 1	NM_000466	Hs.164682
98	PEX13	Peroxisom-Biogenese-Faktor 13	NM_002618	Hs.567316
99	PIB5PA	Phosphatidylinositol (4,5) Bisphosphat 5-Phosphatase, A	NM_014422 NM_001002837	Hs.517549

100	PICK1	Protein interagierend mit PRKCA 1	NM_012407 NM_001039583 NM_001039584	Hs.180871
101	PLEC1	Plectin 1, Zwischenfilament Bindungsprotein 500kDa	NM_201380 NM_201384 NM_000445 NM_201379 NM_201383 NM_201382 NM_201381 NM_201378	Hs.434248
102	POLE2	Polymerase (DNA-dirigiert), Epsilon 2 (p59 Untereinheit)	NM_002692	Hs.162777
103	POLE3	Polymerase (DNA-dirigiert), Epsilon 3 (p17 Untereinheit)	NM_017443	Hs.108112
104	PPIF	Peptidylprolyl-Isomerase F (Cyclophilin F)	NM_005729	Hs.381072
105	PPP2R5A	Protein-Phosphatase 2, regulatorische Untereinheit B', alpha Isoform	NM_006243	Hs.497684
106	PSCD2	Pleckstrin Homologie, Sec7 und Coiled-Coil Domänen 2 (Cytohesin-2)	NM_017457 NM_004228	Hs.144011
107	PSMA5	Proteasom (Prosom, Makropain) Untereinheit, alpha Typ, 5	NM_002790	Hs.485246
108	PTPN12	Protein Tyrosin Phosphatase, Nicht-Rezeptor Typ 12	NM_002835	Hs.61812
109	PTPN3	Protein Tyrosin Phosphatase, Nicht-Rezeptor Typ 3	NM_002829	Hs.436429
110	PTPRCAP	Protein Tyrosin Phosphatase, Rezeptor Typ, C-assoziiertes Protein	NM_005608	Hs.155975
111	QKI	Quaking Homolog, KH Domäne RNA-bindend (Maus)	NM_206855 NM_206854 NM_206853 NM_006775	Hs.510324
112	RASAL2	RAS Protein Aktivator ähnlich 2	NM_170692 NM_004841	Hs.656823
113	RASSF7	Ras Assoziierungs (RalGDS/AF-6) Domänen-Familie 7	NM_003475	Hs.72925
114	RBM10	RNA-bindendes Motiv Protein 10	NM_005676 NM_152856	Hs.401509
115	RBM38	RNA-bindendes Motiv Protein 38	NM_017495 NM_183425	Hs.236361
116	RER1	RER1 Retention in endoplasmischem Reticulum 1 Homolog (S. cerevisiae)	NM_007033	Hs.525527
117	RGL2	Ral Guanin Nukleotid Dissoziierungsstimulator-ähnlich 2	NM_004761	Hs.509622
118	RHOG	Ras Homolog Gen Familie, Mitglied G (rho G)	NM_001665	Hs.501728
119	RNASE1	Ribonuklease, RNase A Familie, 1 (pankreatisch)	NM_198235 NM_198234 NM_198232 NM_002933	Hs.78224
120	RTN4	Retikulum 4	NM_020532 NM_207521 NM_207520 NM_153828 NM_007008	Hs.645283
121	RYR2	Ryanodin Rezeptor 2 (Herz)	NM_001035	Hs.109514
122	SCC-112	SCC-112 Protein	NM_015200	Hs.331431
123	SDS	Serin-Dehydratase	NM_006843	Hs.654416
124	SF3B2	Spleißfaktor 3b, Untereinheit 2, 145kDa	NM_006842	Hs.406423

125	SH3PXD2A	SH3 und PX Domänen 2A	NM_014631	Hs.594708
126	SIX6	Sine oculis Homeobox Homolog 6 (Drosophila)	NM_007374	Hs.194756
127	SLC10A1	Solute Carrier Familie 10 (Natrium/Gal-lensäure Co-Transporter Familie), Mitglied 1	NM_003049	Hs.952
128	SLC6A8	Solute Carrier Familie 6 (Neurotransmitter Transporter, Creatin), Mitglied 8	NM_005629	Hs.540696
129	SMG6	Smg-6 Homolog, nonsense vermittelter mRNA Zerfallsfaktor (C. elegans)	NM_017575	Hs.448342
130	SNRPB2	Kleines nukleares Ribonukleoprotein Poly-peptid B"	NM_003092 NM_198220	Hs.280378
131	SOX11	SRY (Geschlecht bestimmende Region Y)-Box 11	NM_003108	Hs.432638
132	SPI1	Milz Fokus bildendes Virus (SFFV) provirale Integration Onkogen spi1	NM_001080547 NM_003120	Hs.502511
133	SRGAP3	SLIT-ROBO Rho GTPase aktivierendes Protein 3	NM_014850 NM_001033117	Hs.654743
134	STX12	Syntaxin 12	NM_177424	Hs.523855
135	SYK	Milz Tyrosinkinase	NM_003177	Hs.371720
136	TAF4	TAF4 RNA Polymerase II, TATA Box Bindungsprotein (TBP)-assoziiertes Faktor, 135kDa	NM_003185	Hs.18857
137	TCN2	Transcobalamin II	NM_000355	Hs.417948
138	TGOLN2	Trans-golgi Netzwerk-Protein 2	NM_006464	Hs.593382
139	TIA1	TIA1 cytotoxisches Granalien-assoziiertes RNA Bindungsprotein	NM_022173 NM_022037	Hs.516075
140	TOMM40	Translokase von äußerer mitochondrialer Membran 40 Homolog (Hefe)	NM_006114	Hs.655909
141	TXN2	Thioredoxin 2	NM_012473	Hs.211929
142	UGCG	UDP-Glucose Ceramid Glucosyltransferase	NM_003358	Hs.304249
143	USP11	Ubiquitin spezifische Peptidase 11	NM_004651	Hs.171501
144	VDR	Vitamin D (1,25- Dihydroxyvitamin D3) Rezeptor	NM_001017535 NM_000376	Hs.524368
145	VEGFC	Vaskulärer endothelialer Wachstumsfaktor C	NM_005429	Hs.435215
146	YWHAQ	Tyrosin 3-Monooxygenase/Tryptophan 5-Monooxygenase Aktivierungsprotein, theta Polypeptid	NM_006826	Hs.74405
147	ZNF140	Zinkfinger-Protein 140	NM_003440	Hs.181552

Tabelle 5: PTC Markerset PIV-1 bis PIV-9

Num- mer PIV-	Markergen	Beschreibung des Gens	Eingangsnummer	UniGeneID
1	WAS	Wiskott-Aldrich Syndrom (Ekzema-Thrombozytopenie)	BC012738	Hs.2157
2	LRP4	niedrigdichtes Lipoprotein Rezeptor-bezogenes Protein 4	BM802977	Hs.4930
3	TFF3	Trefoil Faktor 3 (intestinal)	BC017859	Hs.82961
4	ST3GAL6	ST3 beta-Galactosid alpha-2,3-Sialyl-transferase 6	BC023312	Hs.148716
5	STK39	Serin-Threoninkinase 39 (STE20/SPS1 Homolog, Hefe)	BM455533	Hs.276271

NACHGEREICHT

6	DPP4	Dipeptidyl-Peptidase 4 (CD26, Adenosin Deaminase komplexierendes Protein 2)	BC065265	Hs.368912
7	CHI3L1	Chitinase 3-ähnlich 1 (Cartilage Glycoprotein-39)	BC038354	Hs.382202
8	FABP4	Fettsäure Bindungsprotein 4, Adipozyt	BC003672	Hs.391561
9	LAMB3	Laminin, beta 3	BC075838	Hs.497636

Tabelle 6: PTC Markerset PV-1 bis PV-11

Numer PV-	Marker gen	Beschreibung des Gens	Eingangsnummer	UniGeneID
1	GPR4	G Protein-gekoppelter Rezeptor 4	BC067535	Hs.17170
2	STAM2	Signal transduzierendes Adaptor-Molekül (SH3 Domäne und ITAM Motiv) 2	BC028740	Hs.17200
3	QPCT	Glutaminyl-Peptid Cyclotransferase (Glutaminyl-Cyclase)	BC047756	Hs.79033
4	CDK7	Cyclin-abhängige Kinase 7 (MO15 Homolog, Xenopus laevis, cdk-aktivierende Kinase)	BC000834	Hs.184298
5	SFTPD	Surfactant, pulmonal-assoziiertes Protein D	BC022318	Hs.253495
6	CYB5R1	Cytochrome b5 Reduktase 1	BC018732	Hs.334832
7	VWF	Von Willebrand Faktor	BI490763	Hs.440848
8	VWF	Von Willebrand Faktor	BQ888783	Hs.440848
9	PDHX	Pyruvat Dehydrogenase Komplex, Bestandteil X	BC010389	Hs.502315
10	HOXA4	Homeobox A4	BM996071	Hs.654466
11	HOXA4	Homeobox A4	BI521357	Hs.654466

Das Set der Erfindung kann verwendet werden, um Krebs oder Tumorzellen nachzuweisen, insbesondere Schilddrüsenkrebs, und sogar, um gutartige Schilddrüsenknötchen von bösartigem follikulärem Schilddrüsenkarzinom (FTC) und papillärem Schilddrüsenkarzinom (PTC) zu unterscheiden. In bevorzugten Ausführungsformen umfasst das Set Komponenten, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern PI-1 bis PI-33, PII-1 bis PII-64, PIII-1 bis PIII-70 und PIV-1 bis PIV-9, vorzugsweise PIV-4 oder PIV-5, und PV-1 bis PV-11, vorzugsweise PV-1, PV-2 und PV-4 bis PV-11, insbesondere von den Tumormarkern PI-1 bis PI-33. Diese Marker sind für papilläres Schilddrüsenkarzinom (PCT) spezifisch und der diagnostizierte Schilddrüsenkrebs kann als PTC gekennzeichnet werden.

In einer ähnlichen bevorzugten Ausführungsform umfasst das Set Komponenten, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern FI-1 bis FI-147. Diese Marker sind für follikuläres Schilddrüsenkarzinom (FTC) spezifisch und der diagnostizierte Schilddrüsenkrebs kann als FTC ge-

kennzeichnet werden.

Besonders bevorzugt umfasst das Set eine Komponente, welche für den Tumormarker SERPINA1 (Serin (oder Cystein) Protease-Inhibitor, Stamm A (alpha-1 Antiproteinase, Antitrypsin), Mitglied 1; NM_000295, NM_001002236, NM_001002235) spezifisch ist, welcher ein sehr wirkkraftiger Marker für PTC ist. Dieser Marker kann als einzelnes Mitglied des Sets PTC von gutartigen Zuständen unterscheiden.

Vorzugsweise umfasst das Set mindestens 5 oder mindestens 10, vorzugsweise mindestens 15, bevorzugter mindestens 20, besonders bevorzugt mindestens 25, insbesondere bevorzugt mindestens 30 Komponenten, welche für die Tumormarker von Tabelle 1 bis 6 oben spezifisch sind. Das Set kann ausgewählt werden aus Komponenten, welche für beliebige mindestens 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 33, 35, 40, 45, 50, 55, 60, 64, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 145, 147, 150, 160, 170, 180, 190 oder 200 der obigen Tumormarker spezifisch sind, z.B. ausgewählt aus PI-1 bis PI-33, PII-1 bis PII-64, PIII-1 bis PIII-70, FI-1 bis FI-147, PIV-1 bis PIV-9, vorzugsweise PIV-4 oder PIV-5, und PV-1 bis PV-11, vorzugsweise PV-1, PV-2 und PV-4 bis PV-11, insbesondere von beliebigen von PI-1, PI-2, PI-3, PI-4, PI-5, PI-6, PI-7, PI-8, PI-9, PI-10, PI-11, PI-12, PI-13, PI-14, PI-15, PI-16, PI-17, PI-18, PI-19, PI-20, PI-21, PI-22, PI-23, PI-24, PI-25, PI-26, PI-27, PI-28, PI-29, PI-30, PI-31, PI-32, PI-33, PII-1, PII-2, PII-3, PII-4, PII-5, PII-6, PII-7, PII-8, PII-9, PII-10, PII-11, PII-12, PII-13, PII-14, PII-15, PII-16, PII-17, PII-18, PII-19, PII-20, PII-21, PII-22, PII-23, PII-24, PII-25, PII-26, PII-27, PII-28, PII-29, PII-30, PII-31, PII-32, PII-33, PII-34, PII-35, PII-36, PII-37, PII-38, PII-39, PII-40, PII-41, PII-42, PII-43, PII-44, PII-45, PII-46, PII-47, PII-48, PII-49, PII-50, PII-51, PII-52, PII-53, PII-54, PII-55, PII-56, PII-57, PII-58, PII-59, PII-60, PII-61, PII-62, PII-63, PII-64, PIII-1, PIII-2, PIII-3, PIII-4, PIII-5, PIII-6, PIII-7, PIII-8, PIII-9, PIII-10, PIII-11, PIII-12, PIII-13, PIII-14, PIII-15, PIII-16, PIII-17, PIII-18, PIII-19, PIII-20, PIII-21, PIII-22, PIII-23, PIII-24, PIII-25, PIII-26, PIII-27, PIII-28, PIII-29, PIII-30, PIII-31, PIII-32, PIII-33, PIII-34, PIII-35, PIII-36, PIII-37, PIII-38, PIII-39, PIII-40, PIII-41, PIII-42, PIII-43, PIII-44, PIII-45,

PIII-46, PIII-47, PIII-48, PIII-49, PIII-50, PIII-51, PIII-52, PIII-53, PIII-54, PIII-55, PIII-56, PIII-57, PIII-58, PIII-59, PIII-60, PIII-61, PIII-62, PIII-63, PIII-64, PIII-65, PIII-66, PIII-67, PIII-68, PIII-69, PIII-70, FI-1, FI-2, FI-3, FI-4, FI-5, FI-6, FI-7, FI-8, FI-9, FI-10, FI-11, FI-12, FI-13, FI-14, FI-15, FI-16, FI-17, FI-18, FI-19, FI-20, FI-21, FI-22, FI-23, FI-24, FI-25, FI-26, FI-27, FI-28, FI-29, FI-30, FI-31, FI-32, FI-33, FI-34, FI-35, FI-36, FI-37, FI-38, FI-39, FI-40, FI-41, FI-42, FI-43, FI-44, FI-45, FI-46, FI-47, FI-48, FI-49, FI-50, FI-51, FI-52, FI-53, FI-54, FI-55, FI-56, FI-57, FI-58, FI-59, FI-60, FI-61, FI-62, FI-63, FI-64, FI-65, FI-66, FI-67, FI-68, FI-69, FI-70, FI-71, FI-72, FI-73, FI-74, FI-75, FI-76, FI-77, FI-78, FI-79, FI-80, FI-81, FI-82, FI-83, FI-84, FI-85, FI-86, FI-87, FI-88, FI-89, FI-90, FI-91, FI-92, FI-93, FI-94, FI-95, FI-96, FI-97, FI-98, FI-99, FI-100, FI-101, FI-102, FI-103, FI-104, FI-105, FI-106, FI-107, FI-108, FI-109, FI-110, FI-111, 112, FI-113, FI-114, FI-115, FI-116, FI-117, FI-118, FI-119, FI-120, FI-121, FI-122, FI-123, FI-124, FI-125, FI-126, FI-127, FI-128, FI-129, FI-130, FI-131, FI-132, FI-133, FI-134, FI-135, FI-136, FI-137, FI-138, FI-139, FI-140, FI-141, FI-142, FI-143, FI-144, FI-145, FI-146, FI-147, PIV-1, PIV-2, PIV-3, PIV-4, PIV-5, PIV-6, PIV-7, PIV-8, PIV-9, PV-1, PV-2, PV-3, PV-4, PV-5, PV-6, PV-7, PV-8, PV-9, PV-10, PV-11. Vorzugsweise ist das Set spezifisch für ein vollständiges Subset, ausgewählt aus PI, PII, PIII, PIV, PV oder FI. Jedoch ist es ebenfalls möglich, eine kleine Anzahl aus diesen Subsets oder vereinigttem Set auszuwählen, da eine Unterscheidung zwischen gutartigen und bösartigen Zuständen oder die Diagnose von Krebs auch mit annehmbarer Sicherheit durchgeführt werden kann. Zum Beispiel umfasst in einer bevorzugten Ausführungsform das Set der Erfindung mindestens 5 (oder eine der oben erwähnten Anzahlen) Komponenten, welche für die Tumormarker spezifisch sind, ausgewählt aus FI-1 bis FI-147. Figuren 4 und 5 zeigen solche diagnostischen Klassifizierungswahrscheinlichkeiten für PTC und FTC. Zum Beispiel hat ein Set, spezifisch für eine Anzahl von Markern aus Tabelle 2 (Subset PII), spezifisch für 5 Marker, nur eine Fehlerspanne von 4%, also 96% aller Fälle würden korrekt klassifiziert. Ein Fehlerwert von 1% (99% Sicherheit) wird mit mindestens 20 Mitgliedern erreicht. Im Fall der FTC-spezifischen Marker wird ein stabiler

Wert von 8% Fehlern mit mindestens 11 unterschiedlichen Markern erreicht, ausgewählt aus dem FI Subset.

Vorzugsweise sind die Komponenten Nukleinsäuren, insbesondere Oligonukleotide oder Primer, welche für Tumormarker-Nukleinsäuren spezifisch sind. In einer weiteren Ausführungsform sind die Komponenten Antikörper (monoklonal oder polyklonal) oder Antikörperfragmente, vorzugsweise ausgewählt aus Fab, Fab', Fab₂, F(ab')₂ oder scFv (einkettige variable Fragmente), welche für Tumormarker-Proteine spezifisch sind.

In einer bevorzugten Ausführungsform sind die Komponenten des Sets auf einem festen Träger immobilisiert, vorzugsweise in Form eines Microarrays oder Nanoarrays. Der Begriff „Microarray“, gleichermaßen „Nanoarray“ wird verwendet, um eine Reihe einer mikroskopischen Anordnung (Nanoarray für eine Reihe im Nanometer-Maßstab) zu beschreiben oder betrifft einen Träger, der solch eine Reihe umfasst. Beide Definitionen widersprechen einander nicht und sind im Sinn der vorliegenden Erfindung anwendbar.

In einem weiteren Aspekt betrifft die vorliegende Erfindung ein Verfahren für den Nachweis von einem oder mehreren Schilddrüsenkrebs-Markern in einer Probe, umfassend Verwenden des Sets der Erfindung und Nachweisen der Anwesenheit oder Messen der Menge des Vorkommens von Tumormarkern in der Probe. Das Auftreten oder Muster der nachgewiesenen Marker kann spezifisch die Anwesenheit dieser Marker identifizieren, was für Krebsdiagnose oder als eine Referenz von gesunden Proben relevant, oder einfach eine genetische Erforschung von Objekten sein kann.

Vorzugsweise umfasst die Probe Zellen, vorzugsweise Säugerkzellen, besonders bevorzugt menschliche Zellen, welche von einer Biopsie oder Körperflüssigkeit beschafft werden können. Insbesondere wird die Anwesenheit oder Menge der Tumormarker in diesen Zellen nach z.B. Zellzerfall nachgewiesen oder gemessen.

Das Verfahren kann einen Nachweis oder eine Messung durch RNA-Expressionsanalyse, vorzugsweise durch Microarray oder quantitative PCR oder Proteinanalyse, vorzugsweise durch Gewebe-Microarray-Nachweis, Protein-Microarray-Nachweis, mRNA-Microarray-Nachweis, ELISA, Multiplex-Tests, Immunhistochemie oder DNA-Analyse, komparative Genomhybridisierung (CGH)-Arrays oder Einzel-Nukleotid-Polymorphismus (SNP)-Analyse umfassen. Diese Verfahren

sind nach Stand der Technik bekannt und können als Beispiele für das weite Feld der genetischen Markeranalyse leicht für das Verfahren der vorliegenden Erfindung verwendet werden.

In einem weiteren Aspekt sieht die vorliegende Erfindung ein Verfahren für die Diagnose von Krebs in einem Patienten vor, umfassend Beschaffen einer Probe, vorzugsweise einer Probe von Zellen, von dem Patienten, Nachweisen von einem oder mehreren Tumormarkern durch Messen von Tumormarker-Signalen mit dem Set gemäß der vorliegenden Erfindung, Vergleichen der gemessenen Signalwerte der Tumormarker mit Werten der Tumormarker in gesunden Proben und Diagnostizieren von Krebs falls mehr als 50%, vorzugsweise mehr als 60%, bevorzugter mehr als 70%, insbesondere bevorzugt mehr als 80% der Werte sich verglichen mit den Werten der gesunden Proben durch mindestens die Standardabweichung, vorzugsweise zweimal die Standardabweichung, sogar bevorzugter dreimal die Standardabweichung des Messverfahrens unterscheiden. Der Unterschied in genetischer Expression zwischen Proben von erkrankten Objekten und gesunden Objekten kann von jeder Art sein und schließt Hochregulation (z.B. von Onkogenen) oder Herunterregulation (z.B. von Tumorsuppressorgenen) ein. Es ist möglich, dass in gesunden Proben ein Gen nicht exprimiert ist, während Expression in erkrankten Proben auftritt. Andererseits ist es ebenfalls möglich, dass ein Gen in erkrankten Proben nicht exprimiert ist, während Expression in gesunden Proben auftritt.

Krebs kann auch diagnostiziert werden, falls sich mehr als 50%, bevorzugt mehr als 60%, bevorzugter mehr als 70%, insbesondere bevorzugt mehr als 80% der Werte der Probe verglichen mit den Werten der gesunden Proben um mindestens einen Faktor 1,5, mindestens einen Faktor 2, mindestens einen Faktor 3 oder mindestens einen Faktor 4 unterscheiden. Üblicherweise werden die Tumormarker-Expressionsprodukte durch einen Faktor von 2 bis 6 hoch- oder herunterreguliert, aber Unterschiede um einen Faktor 60 sind ebenfalls möglich.

In noch einem weiteren Aspekt betrifft die Erfindung ein Verfahren für die Identifizierung von krankheitsspezifischen Markern, wie z.B. in Tabellen 1 bis 6 angegeben, vorzugsweise Genen oder Genexpressionsmustern, umfassend:

- Vorsehen von Genexpressionsdaten zu mehrfachen potentiellen

krankheitsspezifischen Genen von mindestens zwei unterschiedlichen Expressionsdatensätzen,

- Bestimmen gemeinsamer Gene der Datensätze,
- Normalisieren von jedem Genexpressionsdatensatz, vorzugsweise durch Lowess- oder Quantil-Normalisierung,
- Vereinigen der Genexpressionsdatensätze zu einem vereinigten Datensatz und vorzugsweise Normalisieren des vereinigten Datensatzes und Integrieren des vereinigten Datensatzes,
- Bestimmen von Genen des vereinigten Datensatzes durch Bestimmen seines Nearest-Shrunken-Centroids, was die Bestimmung eines kreuzvalidierten Fehlerwerts von Zuweisen der Gene zur Krankheit und Minimieren des Fehlerwerts durch Verringern der Anzahl an Mitgliedern des vereinigten, vorzugsweise normalisierten Datensatzes einschließt,

wobei die Gene des verringerten Datensatzes die für die Krankheit spezifischen Marker sind. Die Kreuzvalidierung kann z.B. das Leave-One-Out Verfahren sein. Vorzugsweise umfasst der Bestimmungsschritt (der Klassifizierungsschritt) die Bestimmung eines maximierten Schwellenwerts des Unterschieds des normalisierten Expressionswerts für jedes Gen zum Zentroid-Wert durch die Kreuzvalidierung. Dann werden die Gene mit niedrigeren normalisierten Expressionswerten als der Schwellenwert aus dem verringerten (oder geschrumpften) Satz entfernt und Gene mit größeren Werten als der Schwellenwert zum Zentroid sind für die Krankheit spezifisch.

Der Bestimmungsschritt kann mehrere Male durch Auslassen der sich ergebenden Marker aus jedem vorhergehenden Schritt wiederholt werden. Das Nearest-Shrunken-Centroid Verfahren wird einen neuen Ergebnissatz von weiteren Markern ergeben, welche für die Krankheit spezifisch sind. Vorzugsweise wird der Bestimmungsschritt 2, 3, 4, 5, 6, 7, 8, 9, 10 oder mehr Male wiederholt. Je nach Größe des vereinigten Datensatzes wird er weitere spezifische Marker ergeben. Vorzugsweise wird für jedes Ergebnis eine Kreuzvalidierung durchgeführt. Die Bestimmung kann wiederholt werden, bis die Kreuzvalidierung einen Fehlerwert von z.B. unter 50%, 60%, 70% oder 80% anzeigt. Bei niedrigeren Werten kann erwartet werden, dass alle Marker identifiziert worden sind.

Die anfänglichen Genexpressionsdatensätze sind rohe Expressionsprofile, z.B. jeweils aus einer multi-genetischen Microar-

ray-Analyse erhalten. Von den meisten gemessenen Genen wird erwartet, dass sie nicht an der Krankheit beteiligt sind und das Verfahren der Erfindung kann kennzeichnende Markergene von mindestens zwei, vorzugsweise drei, vier, fünf, sechs, sieben oder acht Expressionsdatensätzen identifizieren. Daher umfassen die Expressionsdaten der anfänglichen Datensätze vorzugsweise Daten von mindestens zwei unterschiedlichen Microarray-Datensätzen, insbesondere mit studien- oder plattformspezifischen Verzerrungen. Solche Verzerrungen können durch Verwenden von nur einem spezifischen Aufbau während der Messung der Expressionsdaten, z.B. eines Microarrays, auftreten, welcher sich von Aufbauten von anderen Datensätzen signifikant unterscheiden kann. Die vorliegende Erfindung hat den Vorteil, dass während der Vereinigung solcher Sätze die Probleme solcher Messungsverzerrungen überwunden werden. Ferner sind die erhaltenen (anfänglichen) Genexpressionsdaten rohe, unverarbeitete Genexpressionsdaten, also es wurde keine Verfeinerung oder Datenumwandlung vor dem Verfahren der Erfindung durchgeführt.

Vorzugsweise ist die Erkrankung eine genetische Störung, vorzugsweise eine Störung mit veränderter Genexpression, insbesondere bevorzugt Krebs. Andere Typen von Störungen mit veränderter Genexpression können z.B. pathogene Infektionen, insbesondere Virus-, einschließlich Retrovirusinfektionen, Strahlungsschaden und altersbezogene Störungen sein.

Der Schritt des Vereinigens und Integrierens des vereinigten Datensatzes entfernte studienspezifische Verzerrungen. In bevorzugten Ausführungsformen wird dieser Schritt durch schrittweise Vereinigung von zwei Genexpressionsdatensätzen pro Schritt und Integration des vereinigten Datensatzes, vorzugsweise durch DWD (Distance-Weighted-Discrimination) durchgeführt. Im Fall von 3 Datensätzen wird z.B. zuerst Satz 1 mit Satz 2 vereinigt und der verschmolzene Satz 1+2 wird mit Satz 3 vereinigt. Integration kann z.B. Berechnen des Normalenvektors des vereinigten Datensatzes und nachfolgend einer Hyperebene, welche Cluster (z.B. der anfänglichen Datensätze) von Datenwerten des Datensatzes trennt, und Subtrahieren der Datensatz-Mittelwerte wie im DWD-Verfahren einschließen. Im Prinzip kann jedes Datenintegrationsverfahren, welches Verzerrungen entfernt, für das Verfahren der Erfindung verwendet werden.

Vorzugsweise umfassen die mindestens ein, vorzugsweise zwei, drei, vier, fünf, sechs, sieben oder acht erhaltenen Expressionsdatensätze Daten von mindestens 10, vorzugsweise mindestens 20, bevorzugter mindestens 30, noch bevorzugter mindestens 40, mindestens 50, mindestens 70, mindestens 100, mindestens 120, mindestens 140, mindestens 160 oder mindestens 200 unterschiedlichen Genen. Das Verfahren der Erfindung ist besonders geeignet, um durch große Datensätze zu filtrieren und die kennzeichnenden Marker darin zu identifizieren. Das erhaltene Set dieser Marker wird auch als „Klassifikator“ bezeichnet.

Dieses Verfahren, Krebs-spezifische Marker und somit Komponenten, also Oligonukleotide oder Antikörper, welche für Krebs spezifisch sind, zu identifizieren, kann ebenfalls im obigen Verfahren des Diagnostizierens von Krebs verwendet werden. Also werden die Marker entsprechend dem für das diagnostische Verfahren verwendeten Set von Komponenten gemäß dem obigen Verfahren identifiziert (auch „klassifiziert“ genannt), was die Verfeinerung und Festsetzen von Zentroid-Werten der gemessenen Werte der anfänglichen Datensätze einschließt. Dieses Muster kann dann verwendet werden, um Krebs zu diagnostizieren, falls die Werte der Probe des Patienten näher am geclusterten Zentroid-Wert der Tumormarker sind. Demgemäß wird ein Verfahren für die Diagnose von Krebs in einem Patienten vorgesehen, umfassend Beschaffen einer Probe, vorzugsweise einer Probe von Zellen, von dem Patienten, Nachweisen von einem oder mehreren Tumormarkern durch Messen von Tumormarker-Signalen mit dem Set gemäß der vorliegenden Erfindung, Vergleichen der gemessenen Signalwerte der Tumormarker mit Werten der Tumormarker in Krebsproben durch das oben erwähnte Identifizierungsverfahren und Diagnostizieren von Krebs, falls der Nearest-Shrunken-Centroid von Werten der Probe des Patienten für mindestens 50%, vorzugsweise mindestens 60%, bevorzugter mindestens 70% oder sogar mindestens 80%, insbesondere bevorzugte 90% Marker des Sets innerhalb der Standardabweichung, vorzugsweise zweimal der Standardabweichung, noch bevorzugter dreimal der Standardabweichung des Messverfahrens zum Nearest-Shrunken-Centroid der mit den Krebsproben identifizierten Tumormarker ist.

Die vorliegende Erfindung wird ferner durch die folgenden Figuren und Beispiele veranschaulicht, ohne darauf spezifisch

NACHGEREICHT

beschränkt zu werden.

Figuren:

Figur 1: Die ersten beiden Hauptbestandteile vor und nach DWD-Integration. Datensätze werden durch Farbe kodiert und Tumorgebilde werden durch Buchstaben gemäß der Legende kodiert.

Figur 2: Dendrogramm der DWD-integrierten Daten für alle Gene. Die Farben der Zweige des Dendrogramms zeigen den Datensatz der entsprechenden Probe an, die Farbe der Blattmarkierung zeigt das Tumorgebilde an.

Figur 3: Unterscheidung zwischen papillärem Karzinom und gutartigen Knötchen über vier unterschiedliche Datensätze durch nur ein Gen (SERPINA1).

Figur 4 zeigt einen Graph der durchschnittlichen Fehlerwahrscheinlichkeit während PTC-Klassifizierung von reduzierten Sets (Klassifikator) von Markern von Tabelle 2.

Figur 5 zeigt einen Graph der durchschnittlichen Fehlerwahrscheinlichkeit während FTC-Klassifizierung von reduzierten Sets (Klassifikator) von Markern von Tabelle 4.

Beispiele

Beispiel 1: Datensätze

Datensätze wurden von entweder Webseiten oder von öffentlichen Quellen (GEO, ArrayExpress) herunter geladen. Tabelle 7 zeigt eine Zusammenfassung der in dieser Studie verwendeten Datensätze (He et al., PNAS USA 102 (52): 19075-80 (2005); Huang et al., PNAS USA 98 (26): 15044-49 (2001); Jarzab Cancer Res. 65 (4): 1587-97 (2005); Lacroix Am. J. Pathol. 167 (1): 223-231 (2005); J. Clin. Endocrinol. Metab. 90 (5): 2512-21 (2005)). Hier werden drei unterschiedliche Kategorien von Nicht-Krebsgeweben verwendet: kontralateral (c.lat) für gesundes umgebendes Gewebe gepaart mit einer Tumorseite, andere Krankheit (o.d.) für Schilddrüsengewebe, operiert für andere Krankheit, und SN (Struma nodosa) für gutartige Schilddrüsenknötchen. Für alle nachfolgenden Analysen wurden diese als gesund vereinigt.

Tabelle 7: Microarray-Daten, verwendet für Meta-Analyse

veröffentlicht	FTA	FTC	PTC	SN	o.d.	c.lat	Platform
He PNAS 2005	0	0	9	0	0	9	Affy U133plus
Huang PNAS 2001	0	0	8	8	0	0	Affy U133A
Jarzab Cancer Res 2005	0	0	23	0	11	17	Affy U133A
Lacroix Am J Path 2005	4	8	0	11	0	0	Agilent Custom
Reyes nicht veröffentlicht?	0	0	7	0	0	7	Affy U133A
Weber J Clin Endocr Metabol 2005	12	12	0	0	0	0	Affy U95A

Beispiel 2: Finden der Gen-Überschneidung

Der erste Schritt in jeder MetaAnalyse von Microarray-Daten ist es, das Set von Genen zu finden, welches von allen in der Analyse verwendeten Microarray-Plattformen geteilt wird. Traditionell wird Überschneidung durch das Finden gemeinsamer UniGen-Identifikatoren bestimmt. Dies lässt jedoch alle möglichen Spleißvariationen in den untersuchten Genen außer Acht. Falls ein Gen zum Beispiel 2 Spleißvarianten hatte, von welchen eine in dem Versuch differentiell exprimiert war und die andere nicht und falls eine Plattform ein Oligo enthalten würde, welches nur für die differentiell exprimierte Variante spezifisch ist, und die andere Plattform nur ein Oligo zur anderen Variante, dann würde ein auf UniGene basiertes Abgleichen Sonden verschmelzen, welche unterschiedliche Dinge messen.

Um dieses Problem zu überwinden, verschmilzt die hier angenommene Herangehensweise nur Sonden, welche den gleichen Satz von RefSeq Identifikatoren annotieren. Zu diesem Zweck wurden alle passenden RefSeqs für jede(s) Sonde(nset) herunter geladen, entweder durch Bioconductor Annotationspakete (hgu133a, hgu95a und hgu133plus2; erhältlich im Web www.bioconductor.org) oder durch eine BLAST-Suche der Sequenzen bei NCBI-Database. Dann wurden für jede Sonde die RefSeqs sortiert und konzentriert. Dies ist die genaueste Darstellung der an dem Array gemessenen Einheit. Der Medianwert wurde verwendet, falls ein Set von RefSeqs durch Mehrfachsonden an dem Array repräsentiert war. Es waren 5707 unterschiedliche Sets von RefSeqs an allen Arrays vorhanden.

Beispiel 3: Vorverarbeitung und Datenintegration

Zuerst wurde jeder Datensatz untergrundkorrigiert und getrennt normalisiert, wie für jede Plattform empfohlen (Lowess für duale Farbe und Quantil-Normalisierung für einfarbige Experimente) (Bolstad et al. *Bioinformatics* 19 (2): 185-193 (2003); Smyth et al. *Methods* 31 (4): 265-273 (2003)), dann wurden sie verschmolzen und gemeinsam Quantil-normalisiert. Trotz aller Vorverarbeitung ist gezeigt worden, dass auf unterschiedlichen Microarray-Plattformen oder auf unterschiedlichen Generationen der selben Plattform erzeugte Daten aufgrund der Plattform-spezifischen Verzerrungen nicht vergleichbar sein können (Eszlinger et al., *Clin. Endocrinol. Metab.* 91 (5): 1934-1942 (2006)). Dies ist von Hauptbestandteilsanalyse der verschmolzenen Daten wie in Fig. 1 gezeigt ebenfalls offensichtlich. Um diese Verzerrungen zu bereinigen, sind Verfahren für Integration von Microarray-Daten entwickelt worden. Eines dieser Verfahren ist Distance-Weighted-Discrimination (DWD), welches woanders detailliert beschrieben wird (Benito et al., *Bioinformatics* 20 (1): 105-114 (2004)). Kurz: DWD projiziert Datenpunkte auf den Normalenvektor einer Klasse (Datensatz) - welcher wie durch eine modifizierte Support-Vector-Machine (SVM) berechnet Hyperebene trennt, und subtrahiert die Klassen (Datensatz) Mittelwerte. Daher müssen für ein Mehrklassenproblem (mehr als 2 Datensätze zu verschmelzen) die Datensätzen sequentiell verschmolzen werden. Für 6 Datensätze führt dies zu 720 unterschiedlichen Möglichkeiten für Verschmelzung, was Baum-strukturierte Herangehensweisen nicht einschließt, z.B. statt $((1 + 2) + 3) + 4$ erwäge $((1 + 2) + (3 + 4))$. Die hier angewandten Verschmelzungsreihenfolgen wurden nach der allgemeinen Idee ausgewählt, dass ähnliche und größere Datensätze zuerst verschmolzen werden sollten und disparatere später. Es lohnt sich ebenfalls anzumerken, dass Zugeben einer Probe zu einem DWD-verschmolzenen Datensatz den ganzen Datensatz verändern wird, genau wie Zugeben einen neuen Zahl zu einem Vektor von Zahlen seinen Mittelwert verändern wird.

Datenintegration durch DWD wird in Figur 1 veranschaulicht, welche die Wirkung des Datenintegrationsverfahrens auf die ersten beiden Hauptbestandteile zeigt. In dieser Analyse war DWD in der Lage, die Trennung zwischen den Datensätzen zu entfernen, wie durch die PC-Plots und durch das Mischen der Zweige in dem Dendrogramm (siehe Fig. 2) angezeigt. Jedoch trennen sich selbst

in dem DWD-integrierten Datensatz die Lacroix-Daten noch teilweise von den anderen Daten. Sehr wahrscheinlich liegt dies an der Plattform; die Lacroix-Daten sind die einzigen Daten von einer Nicht-Affymetrix Plattform. Figur 2 zeigt Dendrogramme der jeweiligen integrierten Datensätze. Auch scheint DWD-Integration nicht die Unterscheidung zwischen den Tumorgebilden zu behindern (siehe Tabelle 8 unten).

Beispiel 4: Klassifizierung

Für Sondenselektion wurde Klassifizierung und Kreuzvalidierung eines Nearest-Shrunken-Centroid Verfahrens gewählt (Tibshirani et al. PNAS USA 99 (10): 105-114 (2004)) (implementiert im Bioconductor-Paket pamr). Es wurde aus mehreren Gründen gewählt: es ermöglicht Mehrklassen-Klassifizierung und es betreibt Merkmalsselektion, Klassifizierung und Kreuzvalidierung in einem Durchgang. Kurz: es errechnet mehrere unterschiedliche mögliche Klassifikatoren unter Verwendung unterschiedlicher Schrumpfungsschwellenwerte (also unterschiedlich Anzahl von Genen) und findet den besten Schwellenwert aus Kreuzvalidierung. Der Klassifikator wurde mit der kleinsten Anzahl an Genen (größter Schwellenwert) ausgewählt, falls mehr als ein Schwellenwert die gleichen Kreuzvalidierungsergebnisse ergaben.

Beispiel 5: Papilläres Schilddrüsenkarzinom (PTC)

Zuerst und als Qualitätsmaßstab für jede Studie wurde jeder Datensatz getrennt genommen (vor DWD-Integration) und eine pamr-Klassifizierung und Leave-One-Out Kreuzvalidierung (loocv) wurden durchgeführt. Die Ergebnisse der Kreuzvalidierung sind nahezu perfekt bei einzelnen falsch klassifizierenden Proben. Jedoch kann mit Ausnahme des Klassifikators vom He-Datensatz keiner dieser Klassifikatoren auf einen der anderen Datensätze angewendet werden. Klassifizierungsergebnisse sind selten auch nur höher als durch Zufall erwartet. Falls man jedoch die DWD-integrierten Daten (unten) verwendet, passen die Klassifikatoren schon viel besser (siehe Tabelle 8).

Tabelle 8: Klassifizierungsergebnisse bei Anwenden von Klassifikatoren von einer Studie auf eine andere Studie. Vor Datenintegration (links) und nach DWD-Integration (rechts)

NACHGEREICHT

Test-reihe	he	huang	jarzab	reyes	Test-reihe	he	huang	jarzab	reyes
he	1,00	1,00	0,98	1,00	he	1,00	1,00	0,96	1,00
huang	0,50	1,00	0,55	0,50	huang	0,50	1,00	0,90	0,71
jarzab	0,50	0,81	1,00	0,57	jarzab	0,89	1,00	1,00	1,00
reyes	0,78	0,50	0,92	1,00	reyes	0,89	0,88	0,90	1,00

Dann wurde ein pamr-Klassifikator für den vollständigen DWD-integrierten Datensatz konstruiert und in einer Leave-One-Out Kreuzvalidierung validiert. Dies identifizierte einen (!) Gen-Klassifikator, welcher 99% von Proben korrekt in loocv klassifiziert. Dieses unterscheidende Gen ist SERPINA1. Fig. 3 zeigt die Unterscheidung von PTC gegenüber SN vor und nach DWD. Man könnte bis zu 422 Gene zum Klassifikator dazugeben und noch 99% Genauigkeit (von loocv) erreichen. Falls man die SERPINA1-Sonde aus der Analyse entfernt, kann man wieder einen Klassifikator (nachfolgend Klassifikator 2 genannt) mit 99% Genauigkeit in loocv konstruieren, dieses Mal unter Verwendung einer 9-Gen Signatur (siehe Tabelle 3). Entfernen dieser 9 Gene ergibt einen weiteren 9-Gen Klassifikator mit einer ähnlichen Leistung (99% Genauigkeit) und ferner einen 11-Gen Klassifikator mit 99% Genauigkeit. Solche weiteren Klassifikator werden z.B. in Tabellen 1 bis 3, 5 und 6 (oben) für PTC angegeben.

Jedoch werden ähnliche Ergebnisse unter Durchführung der gleichen Analyse von nicht-integrierten Daten erhalten. Unter Berücksichtigung der Ergebnisse von PCA (Fig. 1), wo es offensichtlich war, dass die Varianz, erklärt durch die unterschiedlichen Datensätze, viel größer ist, als die durch Tumorgebilde erklärte Varianz, könnte man sich vorstellen, dass die durch die Datensätze eingeführte Verzerrung Klassifizierung unterstützen (oder behindern) kann. Daher wurde eine Studien-Kreuzvalidierung durchgeführt, wobei sequenziell eine Studie aus dem Datensatz genommen wurde, ein Klassifikator aus den verbleibenden Proben konstruiert wurde und am eliminierten Datensatz getestet wurde. Für die DWD-integrierten Daten war die Genauigkeit der Vorhersage 100, 100, 98 und 100% beim Auslassen von jeweils He, Huang, Jarzab und Reyes aus dem Klassifikator. Für nicht-integrierte Daten waren die Ergebnisse ähnlich (100, 100, 94 und 100%).

Tabelle 9: Gene in Klassifikator 2 (nach Auslassen von SERPINA1)

Symbol	Titel	Cluster	Eingang
WAS	Wiskott-Aldrich-Syndrom (Ekzem-Thrombozytopenie)	Hs.2157	BC012738
LRP4	Niedrigdichtes Lipoprotein Rezep- tor-bezogenes Protein 4	Hs.4930	BM802977
TFF3	Trefoil-Faktor 3 (intestinal)	Hs.82961	BC017859
ST3GAL6	ST3 beta-Galactosid alpha-2,3-Sia- lyltransferase 6	Hs.148716	BC023312
STK39	Serin-Threonin-Kinase 39 (STE20/SPS1 Homolog, Hefe)	Hs.276271	BM455533
DPP4	Dipeptidyl-Peptidase 4 (CD26, Ade- nosin-Deaminase-komplexierendes Protein 2)	Hs.368912	BC065265
CHI3L1	Chitinase 3-ähnlich 1 (Cartilage- Glycoprotein-39)	Hs.382202	BC038354
FABP4	Fettsäure-bindendes Protein 3, Adipozyt	Hs.391561	BC003672
LAMB3	Laminin, beta 3	Hs.497636	BC075838

Beispiel 6: Follikuläres Karzinom

Eine ähnliche Analyse wurde ebenfalls für die FTC-Daten durchgeführt, aber Kreuzvalidierung wurde wegen der sehr begrenzten Verfügbarkeit von Daten behindert. Wieder wurde ein Klassifikator für jeden Datensatz konstruiert (Lacroix und Weber). Sie erreichten eine loocv-Genauigkeit von 96% (Weber) und 100% (Lacroix) an 25 und 3997 Genen. Die Anzahl an Genen in den Lacroix-Daten legt bereits Überanpassung nahe, welche durch Kreuzklassifizierung mit dem anderen Datensatz (jeweils 25 und 35% Genauigkeit) bestätigt wurde. Auch ist die Gen-Überschneidung zwischen den beiden Klassifikatoren niedrig (zwischen 0 und 10% je nach Schwellenwert). Wenn jedoch die 2 Datensätze unter Verwendung von DWD vereinigt werden, konnte ein 147-Gen Klassifikator (Tabelle 4 oben) konstruiert werden, welcher in der Lage war, Proben richtig zu identifizieren (mit einer 92% Genauigkeit).

Beispiel 7: Diskussion

Die vorliegende Erfindung stellt die größte Kohorte von Schilddrüsenkarzinom-Microarray-Daten dar, welche bis jetzt analysiert wurde. Sie nutzt das neuartige kombinatorische Verfahren unter Verwendung der neuesten Algorithmen für Microarray-Daten-Integration und Klassifizierung. Dennoch stellt Meta-Analyse von Microarray-Daten noch eine Herausforderung dar, hauptsächlich weil einzelne Microarray-Untersuchungen auf zumindest teilweise unterschiedliche Fragen zielen und somit unterschiedliche experimentelle Entwürfe verwenden. Darüber hinaus ist die Anzahl an bis jetzt verfügbaren Schilddrüsentumor-Microarray-Daten noch vergleichsweise niedrig (verglichen mit z.B. Brustkrebs). Daher ist man beim Durchführen von Meta-Analyse gezwungen, alle verfügbaren Daten zu verwenden, selbst wenn die Patientenkohorten eine eher heterogene und potentiell verzerrte Population repräsentieren. Insbesondere ist es schwierig, eine homogene Sammlung von Kontrollmaterial (von gesunden Patienten) zu erhalten. Diese werden üblicherweise von Patienten genommen, welche wegen anderer Schilddrüsenerkrankung operiert wurden, was seinerseits sehr wahrscheinlich eine Veränderung in Genexpression wie an Microarrays gemessen verursacht. Die Erzeugung von homogenen Patientenkohorten wird weiter durch die begrenzte Verfügbarkeit von Patientendaten wie Alter, Geschlecht, genetischer Hintergrund usw. behindert.

Beim Durchführen von Meta-Analyse von Microarray-Daten haben viele Forscher ihre Herangehensweise auf Vergleichen von Genlisten von veröffentlichten Studien (Griffith et al., oben zitiert) basiert. Dies ist sehr nützlich, da man alle Studien in der Analyse einschließen kann und nicht auf die Studien beschränkt ist, wo Rohdaten verfügbar sind. Jedoch folgen die Studien im Allgemeinen sehr unterschiedlichen Analysestrategien, einige strenger als andere. Es ist nicht unter der Kontrolle des Meta-Analysierers, wie die Autoren zu den Genlisten kamen. Daher können diese Analysen verzerrt sein.

Hinsichtlich Datenintegration leistet DWD gemäß dem ursprünglichen DWD-Dokument am besten, wenn mindestens 25-30 Proben pro Datensatz vorhanden sind. In der vorliegenden Studie enthielten 4 von 6 Datensätzen weniger als 20 Proben. Dennoch

arbeitete DWD für das Entfernen von Plattform-Verzerrungen vergleichsweise gut (siehe Tabelle 8).

DWD verbesserte in hohem Maße die Ergebnisse von PCA (Figur 1), hierarchischem Clustern (Figur 2) und die Klassifizierungsgenauigkeit beim Anwenden eines Klassifikators von einer Studie auf eine andere Studie (Tabelle 8). In diesem Licht war es überraschend zu sehen, dass die nicht-integrierten Daten in der Studien-Kreuzvalidierung gleich gut leisteten, verglichen mit den DWD-integrierten Daten. Eine Erklärung hierfür ist, dass jede studienspezifische Verzerrung weniger wichtig wird, je mehr Studien bewertet werden. Vorausgesetzt, dass die Studienverzerrung einige Gene mehr als andere beeinträchtigt, werden die beeinträchtigteren Gene wegen der durch die Studienverzerrung eingeführte Varianz weniger wahrscheinlich das pamr-Schwellenwertverfahren überleben. Jedoch gibt es, wie oben gezeigt, eine große Fülle von Genen, welche PTC und gutartige Knötchen unterscheiden. So lange eines (oder ein paar) jener Gene nicht durch die Studienverzerrung beeinträchtigt wird (werden), wird (werden) es (sie) Schwellenwertverfahren überleben und Unterscheidung zwischen Tumorgebilden wird noch möglich sein.

Es gibt eine offensichtliche Diskrepanz, wenn man Fig. 3 anschaut: Vor DWD haben die PTC-Proben eine höhere SERPINA1-Expression, während es nach DWD anders herum ist. Jedoch wie im Abschnitt Materialien und Methoden angemerkt, subtrahiert DWD die Klassenmittelwerte für jede Probe. Dies bedeutet einfach, dass vor DWD die Studienverzerrung für SERPINA1 höher ist als die Differenz in der Expression zwischen den Tumorklassen. Dies erklärt ebenfalls, warum SERPINA1 bei den nicht-integrierten Daten kein gut arbeitender Klassifikator ist.

Eine kürzliche Meta-Analyse und Meta-Übersicht von Griffith et al. (oben zitiert) hat Gene mit einem diagnostischen Potential im Zusammenhang von Schilddrüsenerkrankung zusammengefasst. Sie veröffentlichten Listen von Genen, welche in mehr als einer Schilddrüsenerkrankung analysierenden Studie mit hohem Durchsatz (Microarray, SAGE) erschienen, und wendeten ein Rangsystem an. In ihrer Analyse erreichte SERPINA1 den höchsten Rang, und TFF3, welches Teil von Klassifikator 2 ist (wenn SERPINA1 ausgelassen wird) erreichte den zweiten Rang. Vier von neun Genen von Klas-

sifikator 2 erschienen in der Liste von Griffith et al. (LRP4, TFF3, DPP4 und FABP4).

Die meisten dieser Listen wurden von Microarray-Analyse erzeugt. Jedoch selbst beim Vergleichen der Gene in den Klassifikatoren mit Genlisten, welche mit unabhängigen Technologien erzeugt wurden, wie cDNA-Bibliothek Erzeugung, gibt es wesentliche Überschneidung. SERPINA1 erscheint in ihren Listen, wie auch vier der neun Gene von Klassifikator 2 (TFF3, DPP4, CHI3L1 und LAMB3).

Für den Fall von follikulärer Schilddrüsenkrankheit ist Aufbauen eines robusten Klassifikators viel schwieriger. Dies ist hauptsächlich wegen der begrenzten Verfügbarkeit von Daten. Auch waren die beiden Datensätze bezüglich der verwendeten Plattformen sehr unterschiedlich; während alle anderen Datensätze an Affymetrix GeneChips Arrays von unterschiedlichen Generationen erzeugt wurden, wurden die Lacroix-Daten auf einer maßgeschneiderten Agilent Plattform erzeugt. Dennoch war der Klassifikator (Set) von Tabelle 4 in der Lage, die meisten Proben in loocv richtig zu identifizieren.

Das Leistungsvermögen der hier angenommenen Herangehensweise der Meta-Analyse wird durch eine 99% loocv-Genauigkeit (97,9% gewichtete durchschnittliche Genauigkeit in der Studien-Kreuzvalidierung) für die Unterscheidung zwischen papillärem Schilddrüsenkarzinom und gutartigen Knötchen gezeigt. Dies ist mit dem bislang größten und mannigfaltigsten Datensatz erreicht worden (99 Proben von 4 unterschiedlichen Studien).

Eine Probe wurde falsch klassifiziert und obwohl es nicht möglich ist, die Proben von dieser Analyse zur ursprünglichen Analyse abzubilden, ist die fehlklassifizierte Probe von der selben Gruppe (PTC, Validierungsgruppe) wie die Probe, welche in der ursprünglichen Analyse falsch klassifiziert war. Gemäß Jarzab et al. war die Probe ein Ausreißer, weil sie nur ~20% Tumorzellen enthielt.

P a t e n t a n s p r ü c h e:

1. Set von Komponenten, spezifisch für mindestens 3 Tumormarker, ausgewählt aus den Tumormarkern PI-1 bis PI-33, PII-1 bis PII-64, PIII-1 bis PIII-70, FI-1 bis FI-147, PIV-1 bis PIV-9, vorzugsweise PIV-4 oder PIV-5, und PV-1 bis PV-11, vorzugsweise PV-1, PV-2 und PV-4 bis PV-11.
2. Set gemäß Anspruch 1, dadurch gekennzeichnet, dass das Set Komponenten umfasst, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern PI-1 bis PI-33, PII-1 bis PII-64, PIII-1 bis PIII-70 und PIV-1 bis PIV-9, vorzugsweise PIV-4 oder PIV-5, und PV-1 bis PV-11, vorzugsweise PV-1, PV-2 und PV-4 bis PV-11.
3. Set gemäß Anspruch 1, dadurch gekennzeichnet, dass das Set Komponenten umfasst, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern PI-1 bis PI-33.
4. Set gemäß Anspruch 1, dadurch gekennzeichnet, dass das Set Komponenten umfasst, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern FI-1 bis FI-147.
5. Set gemäß einem von Ansprüchen 1 bis 4, dadurch gekennzeichnet, dass das Set eine Komponente umfasst, welche für den Tumormarker SERPINA1 spezifisch ist.
6. Set gemäß einem von Ansprüchen 1 bis 5, dadurch gekennzeichnet, dass das Set mindestens 5, vorzugsweise mindestens 10, ebenfalls bevorzugt mindestens 15, bevorzugter mindestens 20, besonders bevorzugt mindestens 25, insbesondere bevorzugt 30 Komponenten, spezifisch für die Tumormarker und ausgewählt aus Tabellen 1 bis 6, umfasst.
7. Set gemäß einem von Ansprüchen 1 bis 6, dadurch gekennzeichnet, dass die Komponenten Oligonukleotide sind, welche für Tumormarker-Nukleinsäuren spezifisch sind.
8. Set gemäß einem von Ansprüchen 1 bis 7, dadurch gekennzeichnet, dass die Komponenten Antikörper oder Antikörperfragmente sind, vorzugsweise ausgewählt aus Fab, Fab' Fab₂, F(ab')₂ oder scFv, welche für Tumormarker-Proteine spezifisch sind.
9. Set gemäß einem von Ansprüchen 1 bis 8, dadurch gekennzeichnet, dass die Komponenten auf einem festen Träger immobilisiert sind, vorzugsweise in Form eines Microarrays.

NACHGEREICHT

10. Verfahren für den Nachweis von einem oder mehreren Schilddrüsenkrebs-Markern in einer Probe, welches Verwenden des Sets nach einem von Ansprüchen 1 bis 9 und Nachweisen der Anwesenheit oder Messen der Menge des Vorkommens von Tumormarkern in der Probe umfasst.

11. Verfahren nach Anspruch 10, dadurch gekennzeichnet, dass die Probe Zellen, vorzugsweise Säugerzellen, besonders bevorzugt menschliche Zellen umfasst.

12. Verfahren gemäß Anspruch 10 oder 11, dadurch gekennzeichnet, dass der Nachweis oder die Messung durch RNA-Expressionsanalyse, vorzugsweise durch Microarray oder quantitative PCR, oder Proteinanalyse, vorzugsweise durch Gewebe-Microarray-Nachweis, Protein-Microarray-Nachweis, mRNA-Microarray-Nachweis, ELISA, Multiplex-Tests, Immunhistochemie oder DNA-Analyse, komparative Genomhybridisierung (CGH)-Arrays oder Einzel-Nukleotid-Polymorphismus (SNP)-Analyse durchgeführt wird.

13. Verfahren für die Diagnose von Krebs in einem Patienten, umfassend Beschaffen einer Probe, vorzugsweise einer Probe von Zellen, von dem Patienten, Nachweisen von einem oder mehreren Tumormarkern gemäß einem Verfahren nach einem der Ansprüche 10 bis 12 oder durch Messen von Tumormarker-Signalen mit dem Set gemäß einem von Ansprüchen 1 bis 9, Vergleichen der gemessenen Signalwerte der Tumormarker mit Werten der Tumormarker in gesunden Proben und Diagnostizieren von Krebs falls (a) mehr als 50%, vorzugsweise mehr als 60%, bevorzugter mehr als 70%, insbesondere bevorzugt mehr als 80% der Werte sich verglichen mit den Werten der gesunden Proben durch mindestens die Standardabweichung, vorzugsweise zweimal die Standardabweichung, sogar bevorzugter dreimal die Standardabweichung des Messverfahrens unterscheiden und/oder (b) sich mehr als 50%, bevorzugt mehr als 60%, bevorzugter mehr als 70%, insbesondere bevorzugt mehr als 80% der Werte der Probe verglichen mit den Werten der gesunden Proben um mindestens einen Faktor 1,5 unterscheiden.

14. Verfahren für die Identifizierung von krankheitsspezifischen Markern, vorzugsweise Genen oder Genexpressionsmustern, umfassend:

- Vorsehen von Genexpressionsdaten zu mehrfachen potentiellen krankheitsspezifischen Genen von mindestens zwei unterschiedlichen Expressionsdatensätzen,

- Bestimmen gemeinsamer Gene der Datensätze,
- Normalisieren von jedem Genexpressionsdatensatz, vorzugsweise durch Lowess- oder Quantil-Normalisierung,
- Vereinigen der Genexpressionsdatensätze zu einem vereinigten Datensatz und vorzugsweise Normalisieren des vereinigten Datensatzes und Integrieren des vereinigten Datensatzes,
- Bestimmen von Genen des vereinigten Datensatzes durch Bestimmen seines Nearest-Shrunken-Centroids, was die Bestimmung eines kreuzvalidierten Fehlerwerts von Zuweisen der Gene zur Krankheit und Minimieren des Fehlerwerts durch Verringern der Anzahl an Mitgliedern des vereinigten, vorzugsweise normalisierten Datensatzes einschließt,

wobei die Gene des verringerten Datensatzes die für die Krankheit spezifischen Marker sind.

15. Verfahren nach Anspruch 14, dadurch gekennzeichnet, dass die Expressionsdaten Daten von mindestens zwei unterschiedlichen Microarray-Datensätzen umfassen, insbesondere mit studienspezifischen Verzerrungen.

16. Verfahren nach Ansprüchen 14 oder 15, dadurch gekennzeichnet, dass die Krankheit eine genetische Störung, vorzugsweise eine Störung mit veränderter Genexpression, insbesondere bevorzugt Krebs ist.

17. Verfahren nach einem von Ansprüchen 14 bis 16, dadurch gekennzeichnet, dass die erhaltenen Genexpressionsdaten rohe, unverarbeitete Genexpressiondaten sind.

18. Verfahren nach einem von Ansprüchen 14 bis 17, dadurch gekennzeichnet, dass der Vereinigungsschritt durch schrittweise Vereinigung von zwei Genexpressionsdatensätzen und Integration der vereinigten Daten, vorzugsweise durch DWD, durchgeführt wird.

19. Verfahren nach einem von Ansprüchen 14 bis 18, dadurch gekennzeichnet, dass jeder Expressionsdatensatz Daten von mindestens 10, vorzugsweise mindestens 20, bevorzugter mindestens 30, noch bevorzugter mindestens 40, insbesondere bevorzugt mindestens 50 unterschiedlichen Genen umfasst.

20. Verfahren nach einem von Ansprüchen 14 bis 19, dadurch gekennzeichnet, dass der Bestimmungsschritt für einen vereinigten Satz wiederholt wird, ohne dass die Gene in einem vorhergehenden Bestimmungsschritt bestimmt wurden.

21. Verfahren nach einem von Ansprüchen 14 bis 20, dadurch gekennzeichnet, dass der Bestimmungsschritt die Bestimmung eines maximierten Schwellenwerts des Unterschieds des normalisierten Expressionswerts für jedes Gen zum Zentroid durch die Kreuzvalidierung umfasst, und worin die Gene mit niedrigeren normalisierten Expressionswerten als der Schwellenwert aus dem verringerten Satz entfernt werden.

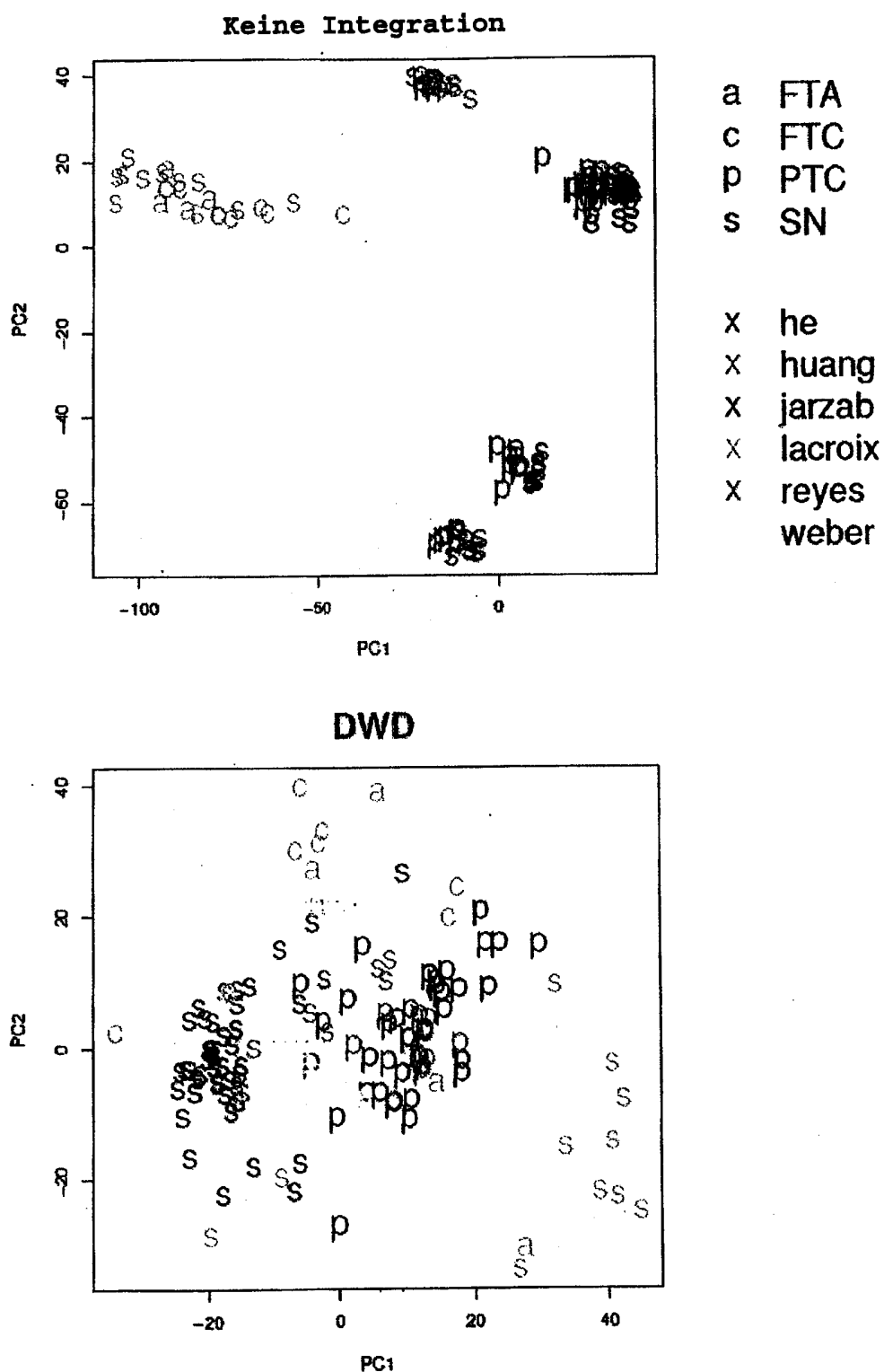
22. Verfahren für die Diagnose von Krebs in einem Patienten, umfassend Beschaffen einer Probe, vorzugsweise einer Probe von Zellen, von dem Patienten, Nachweisen von einem oder mehreren Tumormarkern gemäß einem Verfahren nach einem von Ansprüchen 10 bis 12 oder durch Messen von Tumormarker-Signalen mit dem Set gemäß einem von Ansprüchen 1 bis 9, Vergleichen der gemessenen Signalwerte der Tumormarker mit Werten der Tumormarker in Krebsproben durch das Identifizierungsverfahren nach einem von Ansprüchen 14 bis 21 und Diagnostizieren von Krebs, falls der Nearest-Shrunken-Centroid von Werten der Probe des Patienten für mindestens 50% Marker des Sets innerhalb der Standardabweichung, vorzugsweise zweimal der Standardabweichung, noch bevorzugter dreimal der Standardabweichung des Messverfahrens zum Nearest-Shrunken-Centroid der mit den Krebsproben identifizierten Tumormarker ist.

000116

002

1/5

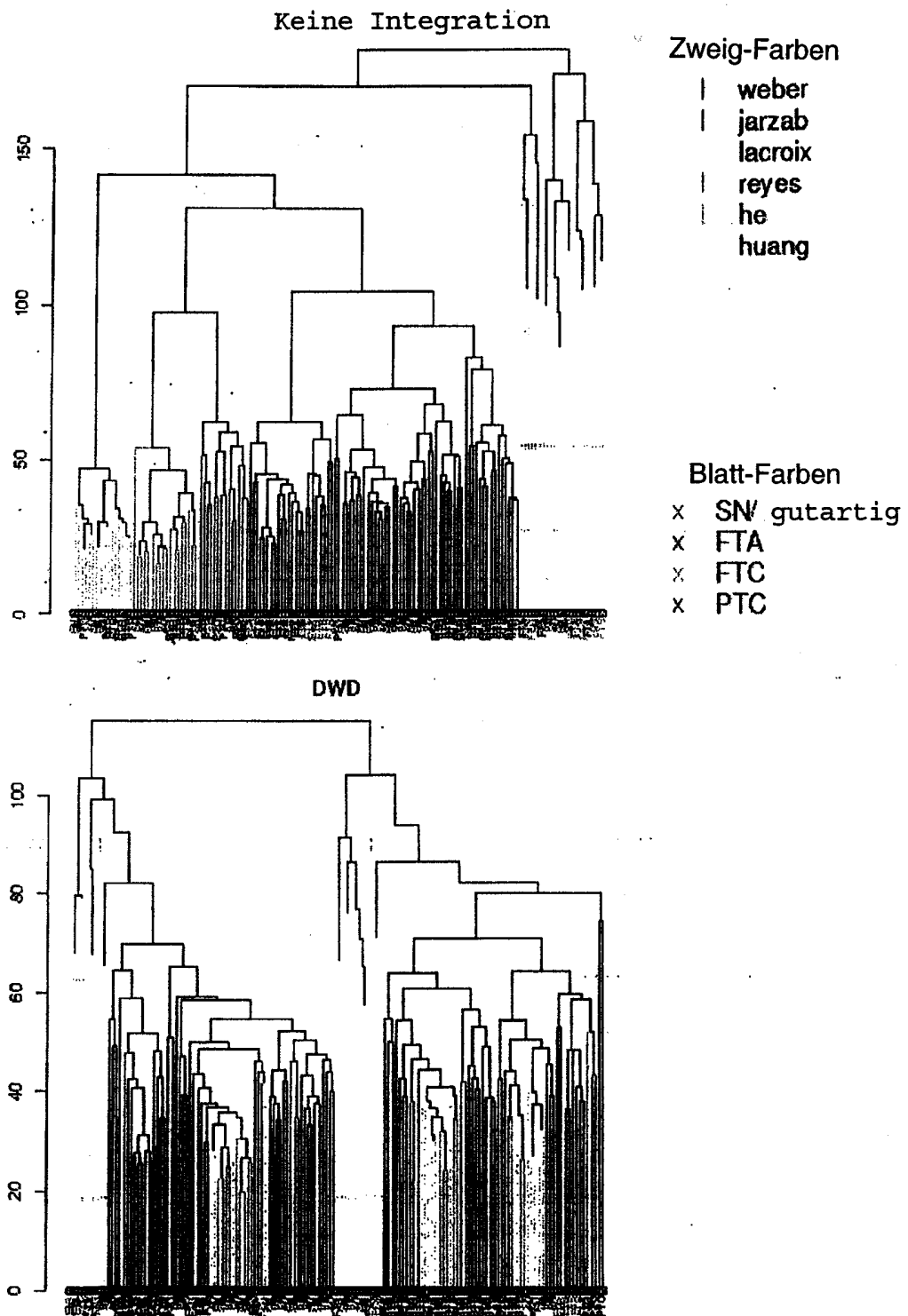
Fig. 1



NACHGEREICHT

Fig. 2

Dendrogramm von integrierten Datensätzen für alle Gene

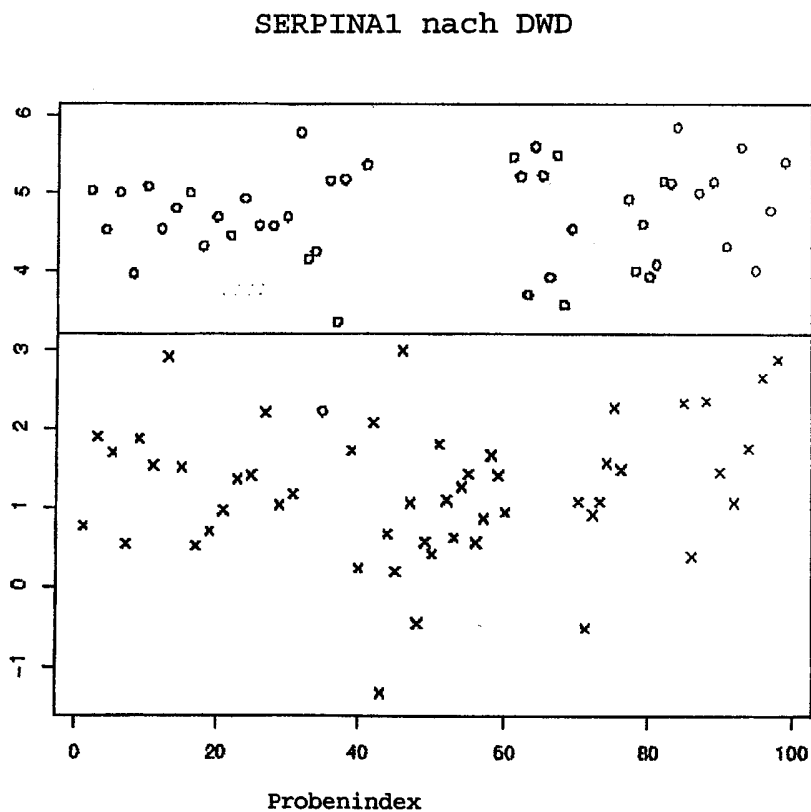
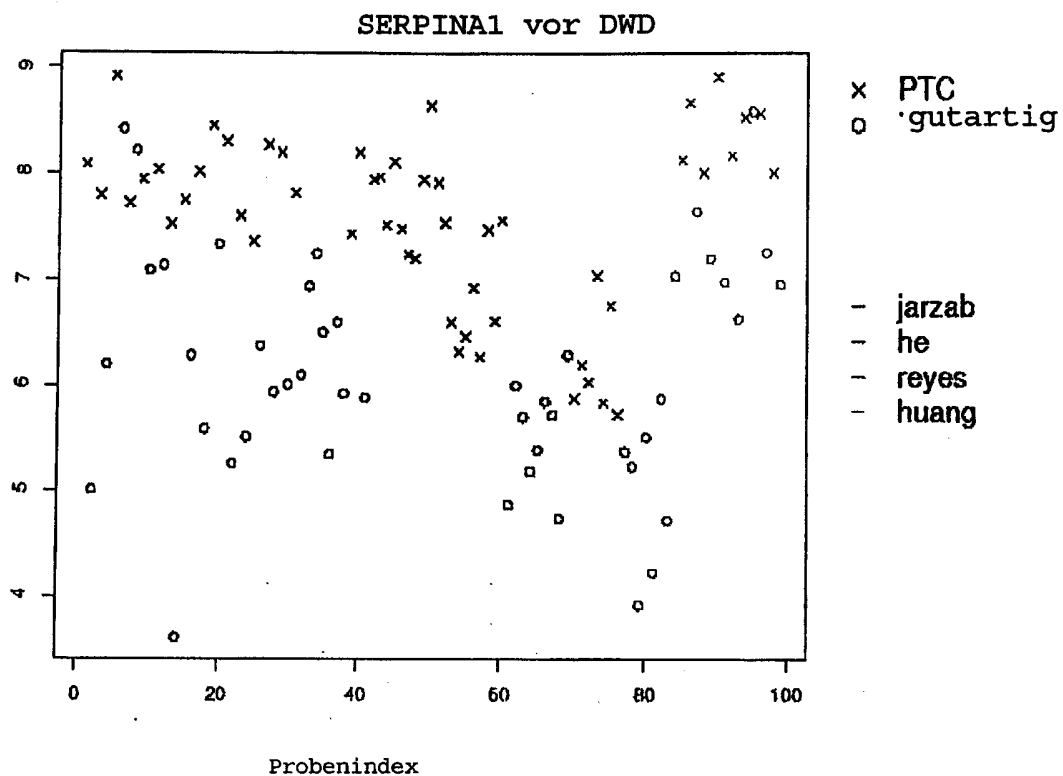


NACHGEREICHT

003116

3/5

Fig. 3



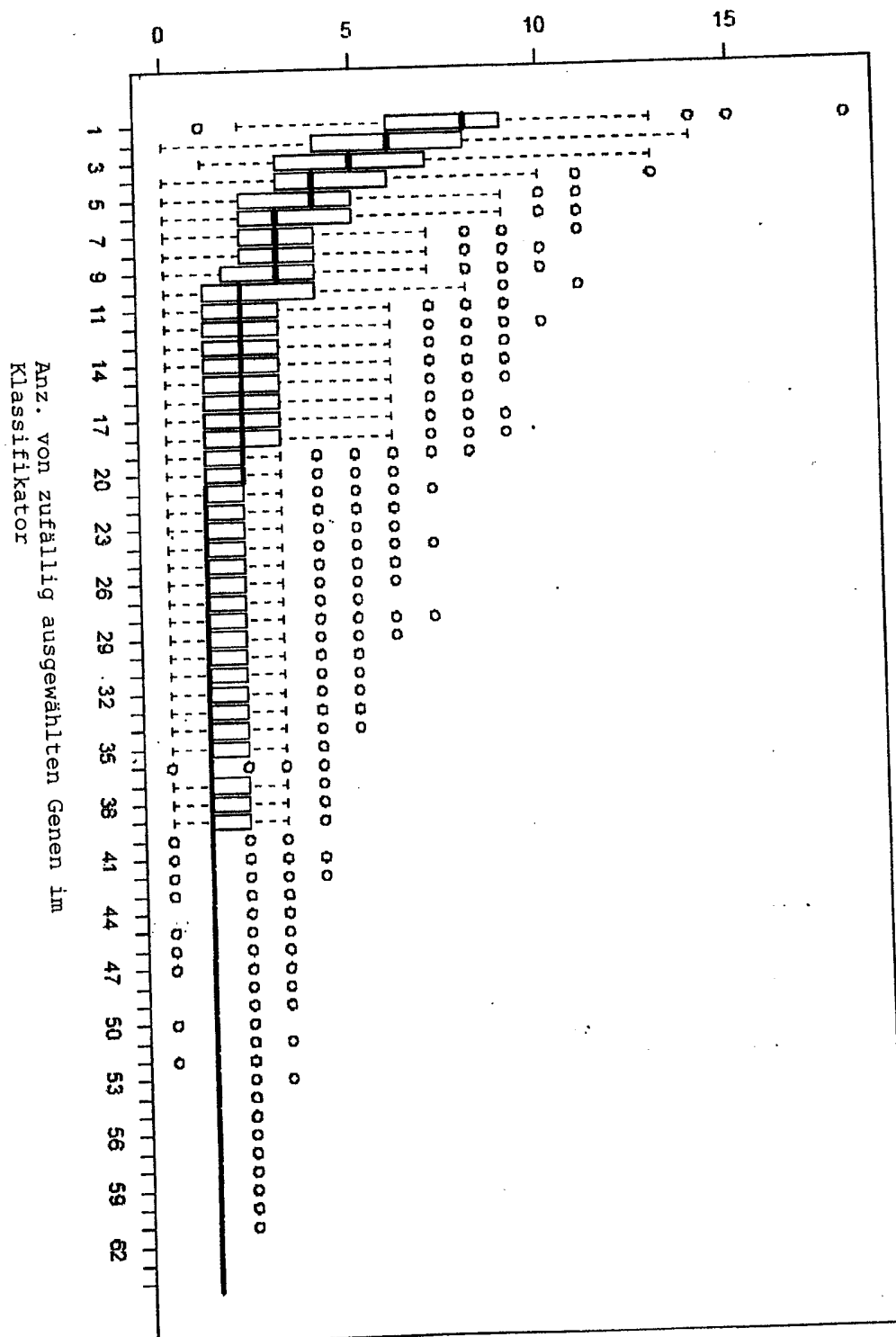
NACHGEREICHT

003116

4/5

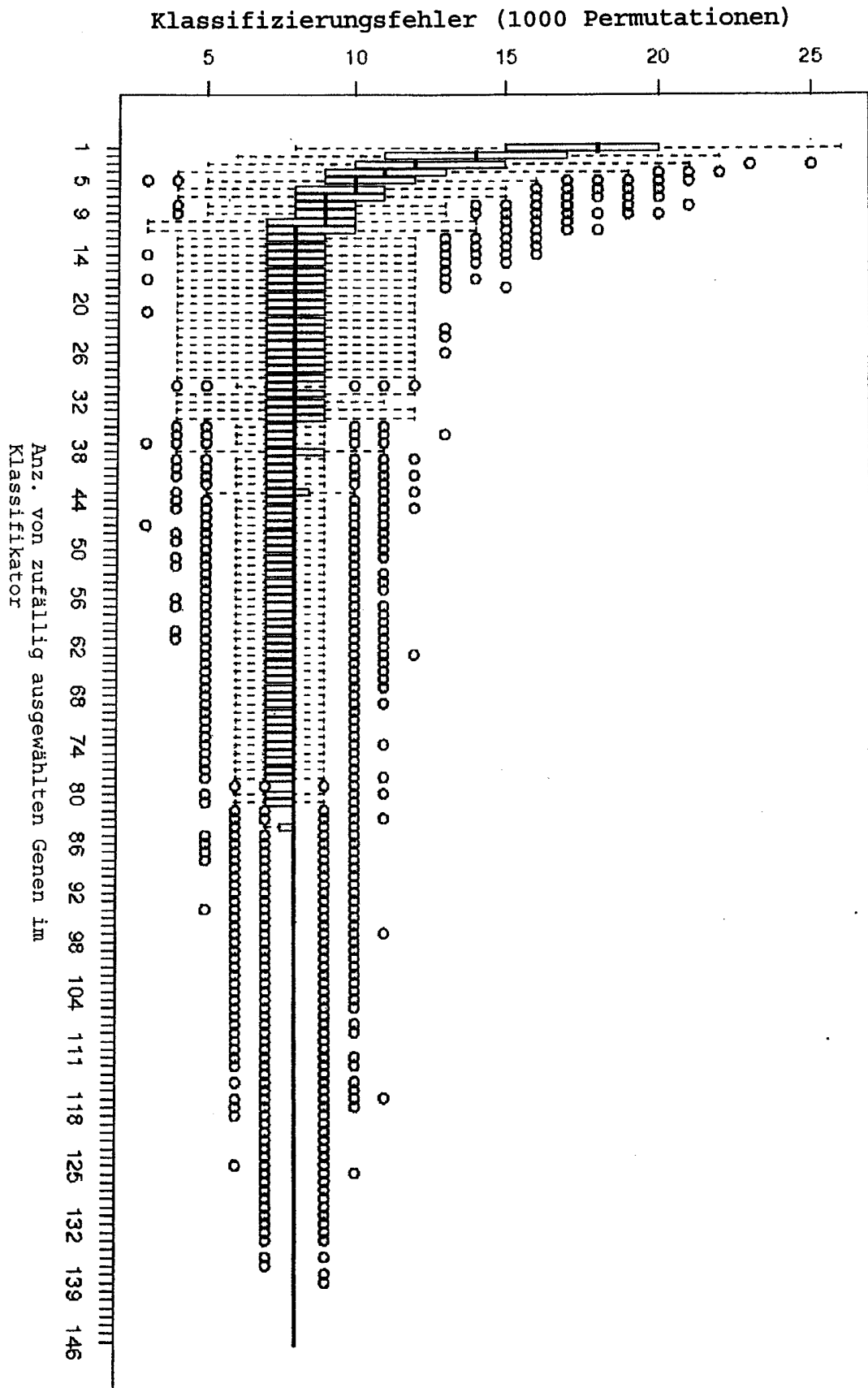
Fig. 4

Klassifizierungsfehler (1000 Permutationen)



NACHGEREICHT

Fig. 5



FTC

NACHGEREICHT

P a t e n t a n s p r ü c h e:

1. Set von Einheiten, spezifisch für mindestens 3 Tumormarker, ausgewählt aus den Tumormarkern BBS9, C13orf1, CBFA2T3, CDT1, CRK, CTPS, DAPK2, EIF5, EREG, GK, GPATCH8, HDGF, IRF2BP1, KRT83, MYOD1, NME6, POLE3, PPP1R13B, PRPH2, RASSF7, ROCK2, RTN1, S100B, SLIT2, SNRPB2, SPAG7, STAU1, SUPT5H, TBX10, TLK1, TM4SF4, TXN, UFD1L, ADH1B, AGR2, AGTR1 (Hs.477887), AGTR1 (Hs.654382), ALDH1A1, ALDH1A3, AMIGO2, ATP2C2, BID, C7orf24, CA4, CCL21, CD55, CDH16, CDH3, CFI, CHI3L1, CHST2, CITED2, CLCNKB, COMP, CTSH, DIO2, DIRAS3, DUSP4, DUSP5, EDN3, ENTPD1, FHL1, GDF15, GPM6A, HBA1, IRS1, KCNJ2, KCNN4, KLK10, LAMB3, LCN2, LMOD1, MATN2, MPPED2, MVP, NELL2, NFE2L3, NPC2, NRCAM, NRIP1, PAPSS2, PDLIM4, PDZK1IP1, PIP3-E, PLAU, PRSS2, PRSS23, RAP1GAP, S100A11, SFTPB, SLPI, SOD3, SPINT1, SYNE1, TACSTD2, UPP1, WASF3, APOE, ATIC, BASP1, C9orf61, CCL13, CD36, CDH6, CFB, CFD, CLDN10, COL11A1, COL13A1, CORO2B, CRLF1, CXorf6, DDB2, DPP6, ECM1, EFEMP1, ESRRG, ETHE1, FAS, FMOD, GABBR2, GALE, GATM, GDF10, GHR, GPC3, ICAM1, ID3, IER2, IGFBP6, IQGAP2, ITGA2, ITGA3, ITM2A, KIAA0746, LRIG1, LRP2, LY6E, MAPK13, MDK, MLLT11, MMRN1, MTMR11, MXRA8, NAB2, NMU, OCA2, PDE5A, PLAG1, PLP2, PLXNC1, PRKCQ, PRUNE, RAB27A, RYR2, SCEL, SELENBP1, SORBS2, STMN2, TBC1D4, TM4SF4, TNC, TPD52L1, TSC22D1, TTC30A, VLDLR, WFS1, AATF, ACOX3, AHDC1, ALAS2, ALKBH1, ANGPTL2, AP2A2, APOBEC3G, APRIN, ARNT, AZGP1, BAT2D1, BATF, BPHL, C13orf1, C14orf1, C2orf3, CBFB, CBR3, CBX5, CCNE2, CD46, CHPF, CHST3, CLCN2, CLCN4, CLIC5, CNOT2, COPS6, CPZ, CSK, CTDP1, DDEF2, DKFZP586H2123, DLG2, DPAGT1, DSCR1, DUSP8, EI24, ENOSF1, ERCC1, ERCC3, ERH, F13A1, FAM20B, FBP1, FCGR2A, FGF13, FGFR1OP, FLNC, FMO5, FRY, GADD45G, GCH1, GFRA1, GLB1, GOLGA8A, HCLS1, HDGF, HRC, ICMT, IFNA5, IGF2BP3, IL12A, ITIH2, ITPKC, JMJD2A, KCNJ15, KCTD12, KIAA0652, KIAA0913, KLKB1, KRT37, LAMB3, LPHN3, LRIG1, LSR, MANBA, MAP7, MAPKAPK5, MET, MMP14, MX1, MYL9, MYO9B, NCOR1, NDRG4, NDUFA5, NEUROD2, NFKB2, NME6, NPY1R, NUP50, PDGFRA, PDHX, PDLIM1, PEX1, PEX13, PIB5PA, PICK1, PLEC1, POLE2, POLE3, PPIF, PPP2R5A, PSCD2, PSMA5, PTPN12, PTPN3, PTPRCAP, QKI, RASAL2, RASSF7, RBM10, RBM38, RER1, RGL2, RHOG, RNASE1, RTN4, RYR2, SCC-112, SDS, SF3B2, SH3PXD2A, SIX6, SLC10A1, SLC6A8, SMG6, SNRPB2, SOX11, SPI1, SRGAP3, STX12,

NACHGEREICHT

SYK, TAF4, TCN2, TGOLN2, TIA1, TOMM40, TXN2, UGCG, USP11, VDR, VEGFC, YWHAQ, ZNF140, WAS, LRP4, TFF3, ST3GAL6, STK39, DPP4, CHI3L1, FABP4, LAMB3, vorzugsweise ST3GAL6 oder STK39, und GPR4, STAM2, QPCT, CDK7, SFTPD, CYB5R1, VWF, PDHX, HOXA4, vorzugsweise GPR4, STAM2, und CDK7, SFTPD, CYB5R1, VWF, PDHX, HOXA4.

2. Set gemäß Anspruch 1, dadurch gekennzeichnet, dass das Set Einheiten umfasst, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern BBS9, C13orf1, CBFA2T3, CDT1, CRK, CTPS, DAPK2, EIF5, EREG, GK, GPATCH8, HDGF, IRF2BP1, KRT83, MYOD1, NME6, POLE3, PPP1R13B, PRPH2, RASSF7, ROCK2, RTN1, S100B, SLIT2, SNRPB2, SPAG7, STAU1, SUPT5H, TBX10, TLK1, TM4SF4, TXN, UFD1L, ADH1B, AGR2, AGTR1 (Hs.477887), AGTR1 (Hs.654382), ALDH1A1, ALDH1A3, AMIGO2, ATP2C2, BID, C7orf24, CA4, CCL21, CD55, CDH16, CDH3, CFI, CHI3L1, CHST2, CITED2, CLCNKB, COMP, CTSH, DIO2, DIRAS3, DUSP4, DUSP5, EDN3, ENTPD1, FHL1, GDF15, GPM6A, HBA1, IRS1, KCNJ2, KCNN4, KLK10, LAMB3, LCN2, LMOD1, MATN2, MPPED2, MVP, NELL2, NFE2L3, NPC2, NRCAM, NRIP1, PAPSS2, PDLIM4, PDZK1IP1, PIP3-E, PLAUI, PRSS2, PRSS23, RAP1GAP, S100A11, SFTPB, SLPI, SOD3, SPINT1, SYNE1, TACSTD2, UPP1, WASF3, APOE, ATIC, BASP1, C9orf61, CCL13, CD36, CDH6, CFB, CFD, CLDN10, COL11A1, COL13A1, CORO2B, CRLF1, CXorf6, DDB2, DPP6, ECM1, EFEMP1, ESRRG, ETHE1, FAS, FMOD, GABBR2, GALE, GATM, GDF10, GHR, GPC3, ICAM1, ID3, IER2, IGFBP6, IQGAP2, ITGA2, ITGA3, ITM2A, KIAA0746, LRIG1, LRP2, LY6E, MAPK13, MDK, MLLT11, MMRN1, MTMR11, MXRA8, NAB2, NMU, OCA2, PDE5A, PLAG1, PLP2, PLXNC1, PRKCQ, PRUNE, RAB27A, RYR2, SCEL, SELENBP1, SORBS2, STMN2, TBC1D4, TM4SF4, TNC, TPD52L1, TSC22D1, TTC30A, VLDLR, WFS1, und WAS, LRP4, TFF3, ST3GAL6, STK39, DPP4, CHI3L1, FABP4, LAMB3, vorzugsweise ST3GAL6 oder STK39, und GPR4, STAM2, QPCT, CDK7, SFTPD, CYB5R1, VWF, PDHX, HOXA4, vorzugsweise GPR4, STAM2, und CDK7, SFTPD, CYB5R1, VWF, PDHX, HOXA4.

3. Set gemäß Anspruch 1, dadurch gekennzeichnet, dass das Set Einheiten umfasst, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern BBS9, C13orf1, CBFA2T3, CDT1, CRK, CTPS, DAPK2, EIF5, EREG, GK, GPATCH8, HDGF, IRF2BP1, KRT83, MYOD1, NME6, POLE3, PPP1R13B, PRPH2, RASSF7, ROCK2, RTN1, S100B, SLIT2, SNRPB2, SPAG7, STAU1, SUPT5H, TBX10, TLK1, TM4SF4, TXN, UFD1L.

4. Set gemäß Anspruch 1, dadurch gekennzeichnet, dass das Set

Einheiten umfasst, welche für mindestens 3 Tumormarker spezifisch sind, ausgewählt aus den Tumormarkern AATF, ACOX3, AHDC1, ALAS2, ALKBH1, ANGPTL2, AP2A2, APOBEC3G, APRIN, ARNT, AZGP1, BAT2D1, BATF, BPHL, C13orf1, C14orf1, C2orf3, CBFB, CBR3, CBX5, CCNE2, CD46, CHPF, CHST3, CLCN2, CLCN4, CLIC5, CNOT2, COPS6, CPZ, CSK, CTDP1, DDEF2, DKFZP586H2123, DLG2, DPAGT1, DSCR1, DUSP8, EI24, ENOSF1, ERCC1, ERCC3, ERH, F13A1, FAM20B, FBP1, FCGR2A, FGF13, FGFR1OP, FLNC, FMO5, FRY, GADD45G, GCH1, GFRA1, GLB1, GOLGA8A, HCLS1, HDGF, HRC, ICMT, IFNA5, IGF2BP3, IL12A, ITIH2, ITPKC, JMJD2A, KCNJ15, KCTD12, KIAA0652, KIAA0913, KLKB1, KRT37, LAMB3, LPHN3, LRIG1, LSR, MANBA, MAP7, MAPKAPK5, MET, MMP14, MX1, MYL9, MYO9B, NCOR1, NDRG4, NDUFA5, NEUROD2, NFKB2, NME6, NPY1R, NUP50, PDGFRA, PDHX, PDLIM1, PEX1, PEX13, PIB5PA, PICK1, PLEC1, POLE2, POLE3, PPIF, PPP2R5A, PSCD2, PSMA5, PTPN12, PTPN3, PTPRCAP, QKI, RASAL2, RASSF7, RBM10, RBM38, RER1, RGL2, RHOG, RNASE1, RTN4, RYR2, SCC-112, SDS, SF3B2, SH3PXD2A, SIX6, SLC10A1, SLC6A8, SMG6, SNRPB2, SOX11, SPI1, SRGAP3, STX12, SYK, TAF4, TCN2, TGOLN2, TIA1, TOMM40, TXN2, UGCG, USP11, VDR, VEGFC, YWHAQ, ZNF140.

5. Set gemäß einem von Ansprüchen 1 bis 4, dadurch gekennzeichnet, dass das Set eine Einheit umfasst, welche für den Tumormarker SERPINA1 spezifisch ist.

6. Set gemäß einem von Ansprüchen 1 bis 5, dadurch gekennzeichnet, dass das Set mindestens 5, vorzugsweise mindestens 10, ebenfalls bevorzugt mindestens 15, bevorzugter mindestens 20, besonders bevorzugt mindestens 25, insbesondere bevorzugt 30 Einheiten, spezifisch für die Tumormarker und ausgewählt aus BBS9, C13orf1, CBFA2T3, CDT1, CRK, CTPS, DAPK2, EIF5, EREG, GK, GPATCH8, HDGF, IRF2BP1, KRT83, MYOD1, NME6, POLE3, PPP1R13B, PRPH2, RASSF7, ROCK2, RTN1, S100B, SLIT2, SNRPB2, SPAG7, STAU1, SUPT5H, TBX10, TLK1, TM4SF4, TXN, UFD1L, ADH1B, AGR2, AGTR1 (Hs.477887), AGTR1 (Hs.654382), ALDH1A1, ALDH1A3, AMIGO2, ATP2C2, BID, C7orf24, CA4, CCL21, CD55, CDH16, CDH3, CFI, CHI3L1, CHST2, CITED2, CLCNKB, COMP, CTSH, DIO2, DIRAS3, DUSP4, DUSP5, EDN3, ENTPD1, FHL1, GDF15, GPM6A, HBA1, IRS1, KCNJ2, KCNN4, KLK10, LAMB3, LCN2, LMOD1, MATN2, MPPED2, MVP, NELL2, NFE2L3, NPC2, NRCAM, NRIP1, PAPSS2, PDLIM4, PDZK1IP1, PIP3-E, PLAU, PRSS2, PRSS23, RAP1GAP, S100A11, SFTPB, SLPI, SOD3, SPINT1, SYNE1, TACSTD2, UPP1, WASF3, APOE, ATIC, BASP1, C9orf61,

CCL13, CD36, CDH6, CFB, CFD, CLDN10, COL11A1, COL13A1, CORO2B, CRLF1, CXorf6, DDB2, DPP6, ECM1, EFEMP1, ESRRG, ETHE1, FAS, FMOD, GABBR2, GALE, GATM, GDF10, GHR, GPC3, ICAM1, ID3, IER2, IGFBP6, IQGAP2, ITGA2, ITGA3, ITM2A, KIAA0746, LRIG1, LRP2, LY6E, MAPK13, MDK, MLLT11, MMRN1, MTMR11, MXRA8, NAB2, NMU, OCA2, PDE5A, PLAG1, PLP2, PLXNC1, PRKCQ, PRUNE, RAB27A, RYR2, SCEL, SELENBP1, SORBS2, STMN2, TBC1D4, TM4SF4, TNC, TPD52L1, TSC22D1, TTC30A, VLDLR, WFS1, AATF, ACOX3, AHDC1, ALAS2, ALKBH1, ANGPTL2, AP2A2, APOBEC3G, APRIN, ARNT, AZGP1, BAT2D1, BATF, BPHL, C13orf1, C14orf1, C2orf3, CBF3, CBR3, CBX5, CCNE2, CD46, CHPF, CHST3, CLCN2, CLCN4, CLIC5, CNOT2, COPS6, CPZ, CSK, CTDP1, DDEF2, DKFZP586H2123, DLG2, DPAGT1, DSCR1, DUSP8, EI24, ENOSF1, ERCC1, ERCC3, ERH, F13A1, FAM20B, FBP1, FCGR2A, FGF13, FGFR1OP, FLNC, FMO5, FRY, GADD45G, GCH1, GFRA1, GLB1, GOLGA8A, HCLS1, HDGF, HRC, ICMT, IFNA5, IGF2BP3, IL12A, ITIH2, ITPKC, JMJD2A, KCNJ15, KCTD12, KIAA0652, KIAA0913, KLKB1, KRT37, LAMB3, LPHN3, LRIG1, LSR, MANBA, MAP7, MAPKAPK5, MET, MMP14, MX1, MYL9, MYO9B, NCOR1, NDRG4, NDUFA5, NEUROD2, NFKB2, NME6, NPY1R, NUP50, PDGFRA, PDHX, PDLIM1, PEX1, PEX13, PIB5PA, PICK1, PLEC1, POLE2, POLE3, PPIF, PPP2R5A, PSCD2, PSMA5, PTPN12, PTPN3, PTPRCAP, QKI, RASAL2, RASSF7, RBM10, RBM38, RER1, RGL2, RHOG, RNASE1, RTN4, RYR2, SCC-112, SDS, SF3B2, SH3PXD2A, SIX6, SLC10A1, SLC6A8, SMG6, SNRPB2, SOX11, SPI1, SRGAP3, STX12, SYK, TAF4, TCN2, TGOLN2, TIA1, TOMM40, TXN2, UGCG, USP11, VDR, VEGFC, YWHAQ, ZNF140, WAS, LRP4, TFF3, ST3GAL6, STK39, DPP4, CHI3L1, FABP4, LAMB3, GPR4, STAM2, QPCT, CDK7, SFTPD, CYB5R1, VWF, und PDHX, HOXA4, umfasst.

7. Set gemäß einem von Ansprüchen 1 bis 6, dadurch gekennzeichnet, dass die Einheiten Oligonukleotide sind, welche für Tumormarker-Nukleinsäuren spezifisch sind.

8. Set gemäß einem von Ansprüchen 1 bis 7, dadurch gekennzeichnet, dass die Einheiten Antikörper oder Antikörperfragmente sind, vorzugsweise ausgewählt aus Fab, Fab', Fab₂, F(ab')₂ oder scFv, welche für Tumormarker-Proteine spezifisch sind.

9. Set gemäß einem von Ansprüchen 1 bis 8, dadurch gekennzeichnet, dass die Einheiten auf einem festen Träger immobilisiert sind, vorzugsweise in Form eines Microarrays.

10. Verfahren für den Nachweis von einem oder mehreren Schilddrüsenkrebs-Markern in einer Probe, welches Verwenden des Sets

nach einem von Ansprüchen 1 bis 9 und Nachweisen der Anwesenheit oder Messen der Menge des Vorkommens von Tumormarkern in der Probe umfasst.

11. Verfahren nach Anspruch 10, dadurch gekennzeichnet, dass die Probe Zellen, vorzugsweise Säugerzellen, besonders bevorzugt menschliche Zellen umfasst.

12. Verfahren gemäß Anspruch 10 oder 11, dadurch gekennzeichnet, dass der Nachweis oder die Messung durch RNA-Expressionsanalyse, vorzugsweise durch Microarray oder quantitative PCR, oder Proteinanalyse, vorzugsweise durch Gewebe-Microarray-Nachweis, Protein-Microarray-Nachweis, mRNA-Microarray-Nachweis, ELISA, Multiplex-Tests, Immunhistochemie oder DNA-Analyse, komparative Genomhybridisierung (CGH)-Arrays oder Einzel-Nukleotid-Polymorphismus (SNP)-Analyse durchgeführt wird.

13. Verfahren für die Diagnose von Schilddrüsenkrebs in einem Patienten, Nachweisen von einem oder mehreren Tumormarkern gemäß einem Verfahren nach einem der Ansprüche 10 bis 12 oder durch Messen von Tumormarker-Signalen mit dem Set gemäß einem von Ansprüchen 1 bis 9 in einer Probe des Patienten, vorzugsweise einer Probe von Zellen, Vergleichen der gemessenen Signalwerte der Tumormarker mit Werten der Tumormarker in gesunden Proben und Diagnostizieren von Krebs falls (a) mehr als 50%, vorzugsweise mehr als 60%, bevorzugter mehr als 70%, insbesondere bevorzugt mehr als 80% der Werte sich verglichen mit den Werten der gesunden Proben durch mindestens die Standardabweichung, vorzugsweise zweimal die Standardabweichung, sogar bevorzugter dreimal die Standardabweichung des Messverfahrens unterscheiden und/oder (b) sich mehr als 50%, bevorzugt mehr als 60%, bevorzugter mehr als 70%, insbesondere bevorzugt mehr als 80% der Werte der Probe verglichen mit den Werten der gesunden Proben um mindestens einen Faktor 1,5 unterscheiden.

14. Verfahren für die Identifizierung von krankheitsspezifischen Markern, vorzugsweise Genen oder Genexpressionsmustern, umfassend:

- Vorsehen von Genexpressionsdaten zu mehrfachen potentiellen krankheitsspezifischen Genen von mindestens zwei unterschiedlichen Expressionsdatensätzen,
- Bestimmen gemeinsamer Gene der Datensätze,
- Normalisieren von jedem Genexpressionsdatensatz, vorzugsweise

durch Lowess- oder Quantil-Normalisierung,

- Vereinigen der Genexpressionsdatensätze zu einem vereinigten Datensatz und vorzugsweise Normalisieren des vereinigten Datensatzes und Integrieren des vereinigten Datensatzes,
- Bestimmen von Genen des vereinigten Datensatzes durch Bestimmen seines Nearest-Shrunken-Centroids, was die Bestimmung eines kreuzvalidierten Fehlerwerts von Zuweisen der Gene zur Krankheit und Minimieren des Fehlerwerts durch Verringern der Anzahl an Mitgliedern des vereinigten, vorzugsweise normalisierten Datensatzes einschließt,

wobei die Gene des verringerten Datensatzes die für die Krankheit spezifischen Marker sind.

15. Verfahren nach Anspruch 14, dadurch gekennzeichnet, dass die Expressionsdaten Daten von mindestens zwei unterschiedlichen Microarray-Datensätzen umfassen, insbesondere mit studienspezifischen Verzerrungen.

16. Verfahren nach Anspruch 14 oder 15, dadurch gekennzeichnet, dass die Krankheit eine genetische Störung, vorzugsweise eine Störung mit veränderter Genexpression, insbesondere bevorzugt Krebs ist.

17. Verfahren nach einem der Ansprüche 14 bis 16, dadurch gekennzeichnet, dass die erhaltenen Genexpressionsdaten rohe, unverarbeitete Genexpressiondaten sind.

18. Verfahren nach einem der Ansprüche 14 bis 17, dadurch gekennzeichnet, dass der Vereinigungsschritt durch schrittweise Vereinigung von zwei Genexpressionsdatensätzen und Integration der vereinigten Daten, vorzugsweise durch DWD, durchgeführt wird.

19. Verfahren nach einem der Ansprüche 14 bis 18, dadurch gekennzeichnet, dass jeder Expressionsdatensatz Daten von mindestens 10, vorzugsweise mindestens 20, bevorzugter mindestens 30, noch bevorzugter mindestens 40, insbesondere bevorzugt mindestens 50 unterschiedlichen Genen umfasst.

20. Verfahren nach einem der Ansprüche 14 bis 19, dadurch gekennzeichnet, dass der Bestimmungsschritt für einen vereinigten Satz wiederholt wird, ohne dass die Gene in einem vorhergehenden Bestimmungsschritt bestimmt wurden.

21. Verfahren nach einem der Ansprüche 14 bis 20, dadurch gekennzeichnet, dass der Bestimmungsschritt die Bestimmung eines

NACHGEREICHT

maximierten Schwellenwerts des Unterschieds des normalisierten Expressionswerts für jedes Gen zum Zentroid durch die Kreuzvalidierung umfasst, und worin die Gene mit niedrigeren normalisierten Expressionswerten als der Schwellenwert aus dem verringerten Satz entfernt werden.

22. Verfahren für die Diagnose von Krebs in einem Patienten, Nachweisen von einem oder mehreren Tumormarkern gemäß einem Verfahren nach einem der Ansprüche 10 bis 12 oder durch Messen von Tumormarker-Signalen mit dem Set gemäß einem der Ansprüche 1 bis 9, in einer Probe des Patienten, vorzugsweise einer Probe von Zellen, Vergleichen der gemessenen Signalwerte der Tumormarker mit Werten der Tumormarker in Krebsproben durch das Identifizierungsverfahren nach einem der Ansprüche 14 bis 21 und Diagnostizieren von Krebs, falls der Nearest-Shrunken-Centroid von Werten der Probe des Patienten für mindestens 50% Marker des Sets innerhalb der Standardabweichung, vorzugsweise zweimal der Standardabweichung, noch bevorzugter dreimal der Standardabweichung des Messverfahrens zum Nearest-Shrunken-Centroid der mit den Krebsproben identifizierten Tumormarker ist.

NACHGEFICHT