

Krebsdiagnostik mit Hilfe von *STATISTICA Data Miner*: Analyse von Patienten- und Genexpressionsdaten

Identifikation klinisch relevanter Gene für die Krebsdiagnostik und –therapie

Ansatz: Die Verknüpfung und Analyse von klinischen Patientendaten mit experimentell gewonnenen molekularen Signaturen von Tumoren eröffnen neue Möglichkeiten für die frühe Diagnose und Therapie von Krebspatienten. Mit Hilfe von *STATISTICA Data Miner* konnten homogene Patientengruppen für experimentelle Studien gebildet, komplexe Auswertungen mit tausenden von Datenpunkten durchgeführt und potenzielle Tumormarker identifiziert werden. Das Projekt wird in weiterführenden Studien fortgesetzt.

Hintergrund: Die Nutzung der Daten im Verbund sollte es ermöglichen, bessere Prognosen über den Krankheitsverlauf von Patienten zu stellen sowie den Therapieerfolg besser zu evaluieren. Alles Ziele, wie sie unter dem Stichwort „personalisierte Medizin“ verfolgt werden. In der Praxis zeigen sich jedoch mehrere Hindernisse, die eine Umsetzung dieser Ziele erschweren. Zum einen bedarf es als Ausgangsmaterial für molekulare Analysen qualitativ hochwertiger klinischer Proben, deren Gewinnungs-, Aufbereitungs- und Lagerungsschritte aus Gründen der Vergleichbarkeit dokumentiert sein sollten. Zum anderen sollten die Proben mit allen essentiellen klinischen Daten annotiert sein, wie etwa den durchgeführten medizinischen Behandlungen, den Operationsschritten und ihren Zeitpunkten etc., da all diese Parameter einen Einfluss auf die molekularen Expressionsmuster ausüben können und somit als präanalytische Variablen das spätere Analyseergebnis beeinflussen.

„Die einfache Bedienung der Projektoberfläche und die Funktionalität von *STATISTICA Data Miner* haben uns sofort begeistert. StatSofts zielgerichtete Dienstleistungen fügen sich gut in unser Szenario und unseren Anwendungsbereich. Wir haben diese deshalb in den vergangenen Jahren gerne in Anspruch genommen.“

Dr. Jörg Spangenberg
Director of Product Development, INDIVUMED GmbH

Wird zudem die sensitive DNA-Microarray-Technik als Werkzeug zur Detektion von molekularen Unterschieden in klinischen Proben eingesetzt, ist die Gefahr für artifizielle Messergebnisse groß. Um diese Gefahr zu minimieren und um vergleichbare klinische Proben für molekulare Analysen zu identifizieren, wurde eine internet-zugängliche oraclebasierte Datenbank (IndivUNET) konzipiert, die mehr als 300 verschiedene klinische und demographische Variablen pro Krebspatient erfasst sowie die Lagerbedingungen und Aufbereitungsschritte der aus diesen Patienten standardisiert gewonnenen Proben dokumentiert. Ziel der nachfolgend beschriebenen klinischen Studie war die Charakterisierung der Probenqualitäten und die integrative Nutzung klinischer Daten zur Identifikation von potenziellen Tumormarkern mittels Genexpressionsanalysen.

Erste Analysen: Für die Studie wurden sämtliche klinischen Daten der Datenbank via einer ODBC-Schnittstelle in *STATISTICA Data Miner* als Auswertungsplattform importiert, da diese Software umfangreiche Datenaufbereitungs- und Datenanalysemöglichkeiten bietet und

zusätzlich die Kapazität aufweist, die sehr umfangreichen Datenmengen resultierend aus Genexpressionsstudien mit tausenden von Datenpunkten pro Array zu integrieren und zu analysieren. Für die Studie wurden zwei geeignete Patientenkohorten, bezeichnet als Trainings- und Testdatensatz, bestehend aus jeweils 30 Patienten mit benignen Erkrankungen (Kontrollgruppe) und malignen Erkrankungen (Adenomcarcinoma-Patienten) zusammengestellt, insgesamt 120 Patienten. Die Zusammenstellung homogener Patientengruppen, die sich in ihren klinischen Variablen möglichst nicht unterscheiden sollten, um den Einfluss von präanalytischen Variablen zu minimieren, erfolgte durch Anwendung von parametrischen (T-tests) als auch nicht-parametrischen statistischen Verfahren (Mann-Whitney-U, Kruskal-Wallis-ANOVA; χ^2 -test), je nach Verteilung der analysierten klinischen Variable.

Die resultierenden Patientengruppen wiesen eine hohe Inter- als auch Intrahomogenität in Bezug auf eine Vielzahl der klinischen Variablen (klinische Blutbilder, Tumorcharakteristika, demographische Daten, Risikofaktoren etc.) auf, so dass sich störende Begleitfaktoren durch unbalancierte Gruppengruppenzusammenstellungen minimieren ließen. Von allen Patienten wurden für die Studie unter standardisierten Bedingungen gewonnene Gewebeproben eingesetzt, wobei als weitere Auswahlparameter auf kurze und vergleichbare Ischämie- und Resektionszeiträume (< 10 Minuten) für die Proben geachtet wurde, um die durch die Prozessierung der Gewebe induzierten molekularen Expressionsveränderungen von Genen möglichst gering zu halten.

Als weiteres Auswahlkriterium wurde zusätzlich der Tumorgehalt von jeder Probe mit Hilfe eines Gewebeschnitts und einer nachfolgenden histochemischen Standardfärbung (H&E Färbung) bestimmt und nur Proben mit einem Gehalt > 40 - 50 Prozent in die Studie aufgenommen. Die Gewinnung des Gewebematerials für die RNA-Isolationen erfolgte durch konsekutive Kryoschnitte (10 μ m) und nachfolgender Lagerung in RNA-Later (Qiagen). In der Studie wurde RNA jeweils aus dem gesunden Gewebe von Patienten, die nicht an einem Krebsleiden erkrankt waren, jedoch chirurgisch behandelt wurden als Kontrolle, als auch als weitere Referenz aus dem Normalgewebe der Patienten mit einem malignen Befund verwendet. Die Isolation der RNAs und daraus abgeleiteter biotinmarkierter cRNAs erfolgte unter Nutzung molekularbiologischer Standardtechniken. Der Status und die Qualität jedes einzelnen Präparationsschrittes sowie die Bestimmung von Konzentrationen wurden mit Hilfe von Lab-on-a-chip-Techniken (Agilent Bioanalyzer) kontrolliert und bestimmt.

Die Messung aller Genaktivitäten in den Gewebeproben erfolgte mittels einer Affymetrix-Plattform unter Verwendung von U133 Plus 2.0 Arrays. Dies gestattet die simultane Detektion von über 47.000 humanen Transkripten. Die Umwandlung der gemessenen Spotintensitäten in Signalwerte zur weiteren Auswertung erfolgte durch Verwendung des MAS 5.0 Algorithmus (Affymetrix). Die auf diese Weise aufbereiteten Daten mit arrayspezifischen Parametern (.RPT-Dateien) und die Dateien mit den Gensignalwerten (.CHP-Dateien) von jedem der insgesamt 180 Arrays in der Studie wurden zur weiteren Auswertung in *STATISTICA Data Miner* importiert. Die hierbei generierten Datenblätter, teilweise mit mehr als 54.000 Zeilen und hunderten von Spalten, dienten als Basis für die mit der Software durchgeführten statistischen und multivariaten Analysen sowie den graphischen Darstellungen der Resultate. Zur Extraktion von Daten wurden zunächst, je nach interessierender Fragestellung, verschiedenste

Datenaufbereitungs- und Analyseschritte mit Hilfe der Software durchgeführt. Zur Vorbereitung für die einzelnen Analyseschritte wurden die Datensätze eines jeden Arrays entsprechend der Zugehörigkeit zur jeweiligen Gewebe- oder Patientengruppe sowie der Zuordnung zum Training oder Testsatz markiert und anhand verschiedenster Qualitätsparameter, wie Signalwerten von zugesetzten Hybridisierungskontrollen (spike in Kontrollen), Signalverhältniswerten von Kontrollgenen, Skalierfaktorwerten, Rauschfaktoren (RawQ), Hintergrundsignalwerte etc., mit Hilfe von nicht-parametrischen Tests (Mann-Whitney-U, Kruskal-Wallis-ANOVA) auf signifikante Unterschiede hin untersucht, um zum einen die Vergleichbarkeit der Datensätze zu gewährleisten und zum anderen Ausreißerarrays zu detektieren. Abbildung 1 zeigt beispielhaft anhand ausgewählter Parameter die Verteilung zwischen den Arrays der Trainings- und Testdatensätze (Skalierungsfaktorbandbreite, Rauschfaktor und Anzahl verlässlich gemessener, präsender Gene).

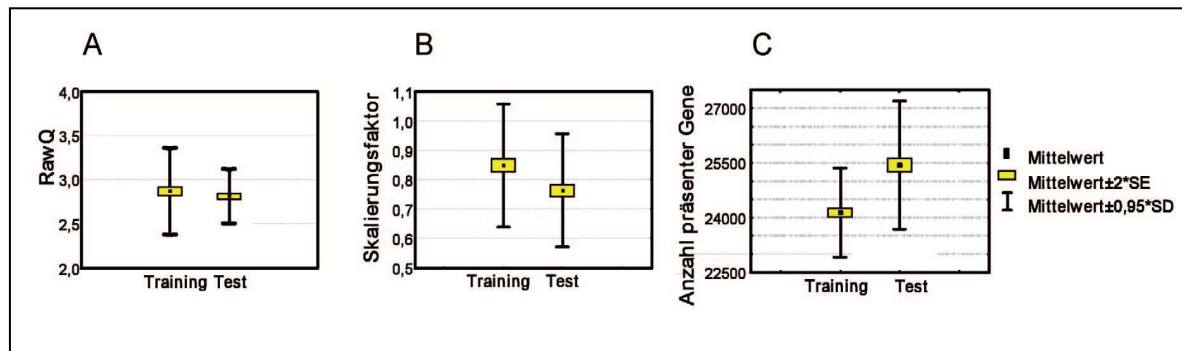


Abbildung 1 Vergleichende Analyse von ausgewählten Array-Qualitätsparametern zwischen den jeweiligen Arrays der beiden Datensätze. A: Rauschfaktor (RawQ); B: Skalierungsfaktor, C: Anzahl verlässlich gemessener Gene (präsende Gene). SE: Standardfehler; SD: Standardabweichung

Weiterführende Analysen: Zusätzliche Charakterisierungen der Daten erfolgten durch Ermittlung von Perzentil-Verteilungen der Signalintensitäten, Mittelwerten, Standardabweichungen und weiteren statistischen Parametern. Alle Analysen zeigten eine hohe Homogenität bezüglich der Qualitätsparameter und damit die Vergleichbarkeit aller Datensätze. Somit konnten alle 180 Arrays für weitergehende Analysen nach Durchführung der notwendigen Datentransformationen wie Varianzstabilisierung, Logarithmierung und Standardisierung herangezogen werden. Um einen ersten Überblick über die gesamten Datenstrukturen zu erlangen, wurde eine Hauptkomponentenanalyse, die eine Reduktion der Datendimensionalität

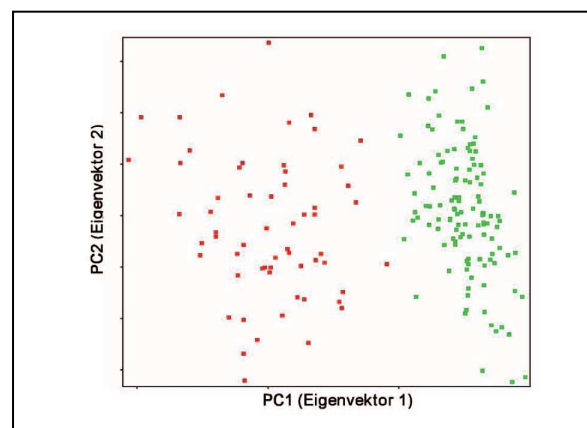


Abbildung 2 Hauptkomponentenanalyse von Arrays. Grün eingefärbt: Arrays hybridisiert mit RNA isoliert aus den Normalgeweben der Kontrollpatienten als auch von Tumorkontrollpatienten. Rot eingefärbt: Arrays hybridisiert mit RNA extrahiert aus Tumorgewebe.

ermöglicht, durchgeführt. An diese schlossen sich Korrelationsanalysen und unüberwachte Clusteranalysen an, die alle als Analysewerkzeuge von der Software bereitgestellt werden. Beispielhaft ist in Abbildung 2 eine Hauptkomponentenanalyse mit allen Arrays in der Studie gezeigt, aus der ersichtlich ist, dass sich anhand zweier extrahierbarer Faktoren (Hauptkomponenten) die mit Tumorgewebe-RNAs hybridisierten Arrays (rot markiert) deutlich von denjenigen Arrays unterscheiden lassen, die mit RNA isoliert aus Normalgewebe (grün markiert) versetzt wurden.

Außerdem zeigt sich, dass keine weitere Unterscheidung der Normalgewebe-Arrays der Patienten, etwa anhand ihres Ursprunges aus Patienten mit benignen oder malignen Erkrankungen, basierend auf dieser Methode, detektiert werden kann.

Zusätzliche Analysen mit Hilfe von Korrelationsanalysen oder Ähnlichkeitsuntersuchungen mittels hierarchischer Clusteranalyse (complete linkage-Verfahren) bestätigten das Ergebnis und zeigten die hohe Homogenität und Ähnlichkeit innerhalb aller vermessenen Normalgewebe, ersichtlich an den hohen Korrelationskoeffizienten und kleinen Distanzwerten.

Dabei hoben sich die untersuchten Tumorgewebe deutlich vom Normalgewebe ab und wiesen eine gewisse Heterogenität ihrerseits auf (Abbildung 3). Die geringe Varianz innerhalb der Normalgewebe im

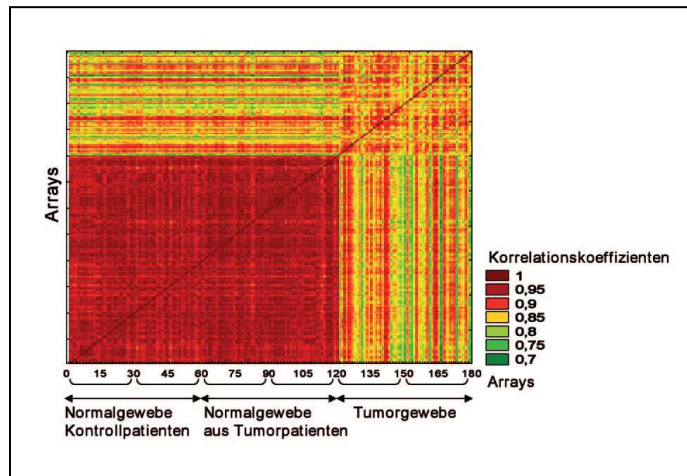


Abbildung 3 Korrelationsanalyse der Arrays in der Studie (N = 180). Eine hohe Korrelation weisen alle Arrays hybridisiert mit RNA isoliert aus den Normalgeweben auf. Tumorgewebe-Arrays zeigen eine geringere Korrelation zu den Normalgeweben und weisen in sich eine höhere Heterogenität auf.

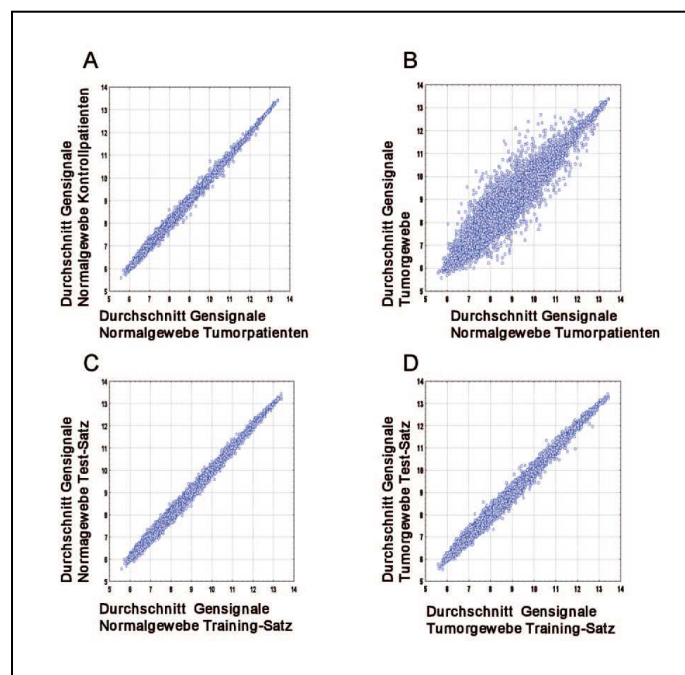


Abbildung 4 Streudiagramm-Analyse von Arrays innerhalb des Trainingsatzes (A,B; jeweils 30 Arrays pro Gewebe) und zwischen Trainings- und Testsatz (C,D; jeweils 30 Arrays pro Gewebe) mit den gemittelten Signalwerten von 10512 Genen.

Unterschied zu einem Vergleich mit den Tumorgeweben lässt sich auch aus Streudiagrammen ersehen, die zudem die hohe Übereinstimmung der gemessenen Signale für die jeweiligen Gene zwischen den Trainings- und den Testdatensätzen verdeutlichen (Abbildung 4).

Zur Identifikation potenzieller Tumormarker, das heißt zur Auffindung von Gensignaturen, die eine eindeutige und zuverlässige Unterscheidung zwischen dem Genexpressionsstatus von normalem Gewebe und Tumorgewebe gewährleisten bzw. einen Übergang vom normalen Gewebe hin zu einem malignen Zustand indizieren können, wurden zusätzliche Filterschritte und Schwellenwertsetzungen notwendig. Beschränkt man sich, wie in dieser Studie geschehen, nur auf die Gene, die zuverlässig in allen 180 Arrays der Studie detektiert wurden (10512 Gene), so lassen sich durch statistische Gruppenvergleichstests (parametrische T-tests/ANOVA und nicht-parametrische Tests Mann-U-Whitney/Kruskal-Wallis-ANOVA)

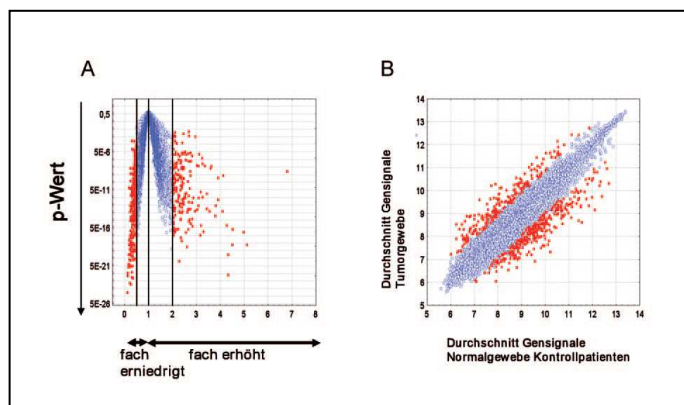


Abbildung 5 Vulkan (A)- und Streudiagrammanalyse (B) von 10.512 Genen der Trainingsdatensätze. Gene, die sich signifikant ($p < 0,05$) in ihrer Expression um mindestens das Zweifache zwischen dem Normalgewebe der Kontrollpatienten und Tumorgewebe unterscheiden, sind rot markiert.

diejenigen Gene identifizieren, die sich mit einer bestimmten Signifikanz (hier $p < 0,05$) um mindestens das Zweifache ihres Expressionsniveaus voneinander in den jeweiligen analysierten Gruppen unterscheiden. Abbildung 5 zeigt beispielhaft die Identifikation dieser Gene für einen Vergleich zwischen den Arrays hybridisiert mit Normalgewebe-RNAs aus Patienten mit benignen Erkrankungen gegenüber den Arrays mit Tumorgewebe-RNAs mit Hilfe eines Vulkan- und eines Streudiagramms. Rot markiert sind diejenigen Gene mit einem mindestens zweifachen Expressionsunterschied und einem p -Wert $< 0,05$.

Ergebnisse: Die gesamte statistische Auswertung aller möglichen Gruppenvergleiche und die Gegenüberstellung von Test- und Trainingsätzen ergab etwa 400 Gene mit einer mindestens zweifach verschiedenen Expression zwischen benignem Normalgewebe und dem Tumorgewebe und eine ebenso große Anzahl, die zwischen dem Normalgewebe und dem Tumorgewebe der Krebspatienten als unterschiedlich expremiert klassifiziert wurde. Die erhaltenen Anzahlen beruhen darauf, dass diese Gene in beiden Datensätzen übereinstimmend nachgewiesen wurden. Für Validierungsstudien dieser identifizierten Kandidatengene war eine noch weitergehende Einengung dieser Anzahl gewünscht, um einen effektiven Ressourceneinsatz zu ermöglichen. Eine Möglichkeit hierfür bot die multiple Korrelationsanalyse unter Verwendung der klinischen Parameter der Patienten in der Studie. Mit ihrer Hilfe wurden diejenigen Gene herausgefiltert, die signifikante Korrelationen zu tumorspezifischen klinischen Variablen wie etwa Tumormarkerkonzentrationen im Blut, Tumorgroße, Tumorstadium usw. aufwiesen, um sich auf sie zu fokussieren, sowie diejenigen identifiziert, die eine Vielzahl von Korrelationen zu klinischen Variablen aufwiesen, um sie zu

priorisieren (Abbildung 6). Außerdem wurden die Signalmuster der Gene auf Assoziationen zur Verabreichung von Pharmaka untersucht. Dadurch ließ sich die Anzahl der nun in weiteren experimentellen Studien bearbeiteten potenziellen Tumormarker beträchtlich einschränken.

Fazit: Es zeigt sich somit, dass die Kombination von multifunktionaler Software, die in der Lage ist eine Vielzahl von klinischen und molekularen Daten zu integrieren und zu analysieren, mit qualitativ hochwertigen detailliert annotierten Gewebeprobe

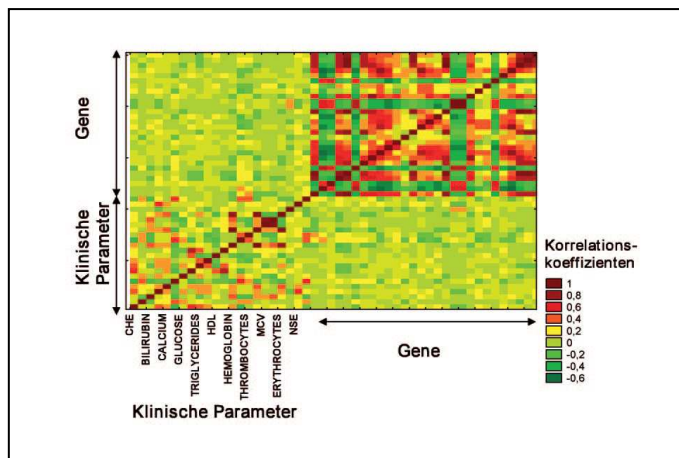
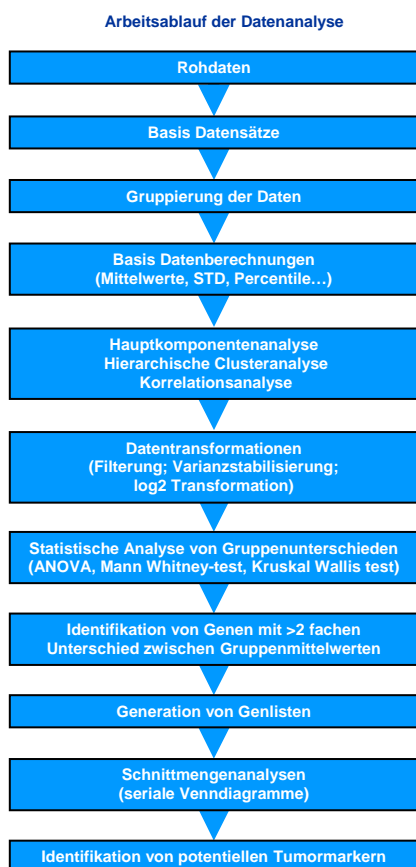


Abbildung 6 Multiple Korrelationsanalyse von klinischen Parametern und Expressionssignalen ausgewählter Gene.

das Auffinden von tumorrelevanten Biomarkern in molekularen Analysen entscheidend unterstützen kann.



Entscheidung für STATISTICA Data Miner. Die Entscheidung für *STATISTICA Data Miner* fiel nach einer durch StatSoft begleiteten, ausführlichen Testphase. Durch die strukturierte Projektoberfläche und intuitive Benutzerführung der Software können sich die Fachanwender auf ihre Aufgaben konzentrieren und gleichzeitig die analytische Mächtigkeit von *STATISTICA Data Miner* nutzen (z.B. die Dimensionalität der Genexpression zu reduzieren). Dabei stehen Statistik und Data Mining in einem Werkzeug zur Verfügung.

Über Indivumed: Die INDIVUMED GmbH in Hamburg ist ein innovatives Biotechnologieunternehmen im Bereich der Krebsforschung. Das Hauptziel besteht darin, molekulare Diagnostika und individualisierte medikamentöse Behandlungen von Krebspatienten zu entwickeln. Durch den Einsatz eigener Mitarbeiter zur Proben- und Informationsgewinnung in einem Verbund führender Krankenhäuser und der Entwicklung einer integrierten analytischen Plattform hat Indivumed einen neuen Qualitätsstandard in der Krebsforschung gesetzt. Das Unternehmen wurde mehrfach mit Preisen ausgezeichnet.