5 Diskussion

In dieser Arbeit ist das Erstellen sowie die Analyse zweier Bibliotheken beschrieben. Dabei handelt es sich um Zellen des Immunsystems, zum einen um eine T-Zell-Linie (Jurkat), zum anderen um eine Natural Killer-Zell-Linie (NKL). Beide Zell-Linien sind etablierte Zell-Linien leukämischer Patienten. Die Untersuchungen der Genexpression beider Linien sollen helfen, in großem Maßstab an der Erklärung der Krebsentstehung und -Progression beizutragen. Sie sollen einen Beitrag an den umfangreichen Studien über *drug targeting* sowie der Medikamentenresistenz leisten.

Die Analyse nur einzelner weniger Gene auf einmal weicht immer mehr der hoch parallelen Analyse mehrerer Tausend von Genen. In der heutigen Zeit stehen immer mehr globale Experimente im Vordergrund, die nur mit Hochdurchsatz-Methoden umzusetzen sind. Da viele Gene schon sequenziert und charakterisiert sind, ist es notwendig, den Arbeitsaufwand so gering wie möglich zu halten und Redundanzen zu vermeiden, die Versuche aber auch so detailliert wie möglich zu gestalten. Mit der Entwicklung neuer Techniken wurden so auch die Methoden der Molekulargenetik revolutioniert, die es überhaupt erst ermöglichen, immer komplexere und diffizilere Arbeitsschritte durchzuführen und aufwendige manuelle Abläufe zu automatisieren. Dafür ist das Oligonucleotid Fingerprinting, das sich die Array-Technologie zunutze macht, eine prädestinierte Methode. Das Oligonucleotid Fingerprinting ist ein Verfahren des Sequenzierens durch Hybridisieren in großem Maßstab und findet Einsatz beispielsweise bei der Analyse der Genexpression verschiedener Gewebe und Organe. Das ONF stellt eine Methode dar, mit der parallel Hunderttausende von Genen mit relativ geringem finanziellen und zeitlichen Aufwand analysiert werden können. Die konventionelle Art des Sequenzierens gestaltet sich wesentlich aufwendiger in der Durchführung und ist zeit- und kostenintensiver als das ONF.

In zunehmendem Maße findet eine weitere Miniaturisierung statt, die es in Zukunft ermöglichen wird, Arrays in immer dichteren Anordnungen zu erstellen. Das ONF ist eine Methode mit großen Möglichkeiten, die eine breite Anwendung in einer Anzahl von Applikationen findet. Die Hauptanwendung des ONF liegt in der Sequenzierung durch Hybridisierung in großem Maßstab mit relativ geringem Zeit- und Kostenaufwand und erleichtert den Arbeitsaufwand um ein Vielfaches (10-fach) (Maier, E. *et al.*, 1994). Sie erlaubt das Identifizieren von Genen sowie das Clustern von Bibliotheken, d.h., gleiche oder ähnliche Klone können in Gruppen zusammengefaßt und so Expressionsprofile der Bibliotheken erstellt werden. Mit Hilfe der in dieser Arbeit angewandten Methode des Oligonucleotid Fingerprintings können mittels der Co-Clustering-Analysen vergleichende Expressionsprofile verschiedener Zell-Linien erstellt werden. Sie erweist sich als sehr hilfreich und leistungsfähig beim Identifizieren neuer, noch unbekannter Gene, sowie

beim Erstellen verschiedenartiger Unigene-Sets und stellt damit ein wertvolles Tool für sich anschließende Applikationen dar.

An dieser Stelle sollen nun einige Vor- und Nachteile des Oligonucleotid Fingerprintings näher dargestellt werden. Das ONF mit all seinen zahlreichen Arbeitsschritten -von der Konstruktion der cDNA-Bibliotheken, über die ONF-Hybridisierungen bis hin zur Auswertung der Signalintensitäten und der Clustering-Analyse- birgt jedoch auch einige Schwachstellen in sich, die begründet liegen in der Methode selbst. Hier sollen auch einige Ansätze diskutiert werden, die dazu beitragen können, diese Methode der Genexpressionsanalyse zu verbessern und ihre Effektivität weiter zu erhöhen.

Die Identifizierung des Genoms der Zelle ist ein wesentlicher Bestandteil molekulargenetischer Forschung. Dabei ist die Identifizierung der Gene, die in der Zelle aktiv sind, die also exprimiert und in Proteine translatiert werden, und die Häufigkeit der Expression dieser Gene in spezifischen Geweben und Organen oder zu bestimmten Zeitpunkten der Entwicklung von besonderem Interesse.

Es ist schon lange Zeit möglich, den codierenden Teil der Gene, der Teil der DNA-Sequenz, der in Proteine translatiert wird, zu isolieren und in die stabilere cDNA revers zu transkribieren. Somit kann diese Information der Genexpression, die durch die mRNA widergespiegelt wird, in einer cDNA-Bibliothek gesammelt und für Genexpressionsanalysen bereitgestellt werden.

Da in der Zelle Tausende verschiedener Gene aktiv sind und die Mehrheit dieser Genen in mehrfacher Kopiezahl vorliegt, die zwischen 1 und etwa 1.000 variiert, also auch in der cDNA-Bibliothek redundant vertreten sind, ist die Analyse dieser Datenmenge nur unter Nutzung von Hochdurchsatz-Techniken der Datengeneration in Form von Robotertechniken und Datenanalyse-Methoden in großem Maßstab möglich, um diese aufwendige Aufgabe zu übernehmen.

Das Oligonucleotid Fingerprinting ist eine effiziente, da hoch parallele und in hohem Maße automatisierte, und schnelle Methode des Sequenzierens durch Hybridisieren und stellt damit eine prädestinierte Möglichkeit der Analyse der Genexpression einer Zellpopulation oder eines zu untersuchenden Gewebes in cDNA-Bibliotheken dar, die mit hohem Durchsatz realisiert werden kann. Das ONF und die sich daran anschließenden Normalisierungs- und Clustering-Verfahren vereinen viele Vorteile in sich. Diese Verfahren ermöglichen es, wesentlich schneller als das herkömmliche Sequenzieren genomischer DNA, große Datenmengen zu analysieren sowie wesentliche Informationen zu filtern und damit die Redundanz der Daten zu minimieren. So können repetitive Experimente und Analysen (wie Sequenzierungen) gleicher Sequenzen vermieden und schon bekannte Sequenzen aus der weiteren Analyse ausgeschlossen werden. Somit erweist sich die Methode des ONF als kostengünstiger und weniger zeitaufwendig im Erstellen von Genexpressionsprofilen und der Analyse exprimierter Sequenzen einer Zelle.

Das Fingerprinting beruht auf der Bindung von kurzen DNA-Sequenzen an viele Tausende von DNA-Targets. Es gibt eine Anzahl von Einflüssen verschiedenster Art, die in den komplexen Ablauf der Erstellung der Bibliotheken bis hin zur Analyse der Daten eingreifen können.

Die Bildanalyse der Hybridisierungssignale umfaßt die Detektion und Quantifizierung der Spots der Arrays, woraus die Rohdaten ermittelt werden. Die Signalintensitäten der Spots können aber durch eine Vielzahl von Faktoren beeinflußt werden. Diese Fehlerquellen sind niemals vollständig auszuschließen, wie auch die daraus resultierenden Abweichungen vom Ideal. Die Auswirkungen dieser Einflüsse zu minimieren, dafür wurden verschiedene Normalisierungsmethoden entwickelt.

Eine sehr robuste Methode, wenngleich auch verbunden mit einem höherem Informations- und Komplexitätsverlust, ist das *double ranking*, das unter Berücksichtigung des Einflusses experimenteller Faktoren Rohdaten in normalisierte Daten umwandelt und die Reproduzierbarkeit der Daten am ehesten bewerkstelligt (Herwig, R., 2000). Das Normalisieren der Rohdaten, dem ein spezieller sequentieller Algorithmus zugrunde liegt, ist ein Verfahren, das die Analyse der Daten erheblich verbessert und somit eine wesentliche Rolle in der Datenanalyse spielt.

Das ONF und das sich daran anschließende Clustering liefern keine absoluten Werte und Ergebnisse, sondern zeigen die Tendenz der Clustergröße und damit der relativen Expressionshäufigkeit eines Gens in einer Zellpopulation bzw. geben beim Vergleich der Expression zweier Populationen eine Wahrscheinlichkeit der Unterschiedlichkeit an. Dies trifft insbesondere auf kleinere Cluster zu. Die Grenze der Sensitivität dieser Methode liegt etwa bei Clustern mit einer Größe von kleiner als 20. Bei einer Clustergröße von kleiner als 20 Klonen je Cluster können keine verläßlichen Angaben mehr über die Clusterverteilung gemacht werden und es kann zu Ungenauigkeiten und damit zu Falschaussagen bezüglich der Expressionshöhe eines Gens oder der Signifikanz der differentiellen Expression kommen.

Das Erstellen der Fingerprints der Klone einer cDNA-Bibliothek beruht auf Signalintensitäten, die wieder beruhen auf einer Reihe von unabhängigen Hybridisierungsexperimenten. Das Clustering erfolgt anhand der Ähnlichkeiten der Fingerprints zueinander, die sich aus den vielen Signalintensitäten der ONF-Hybridisierungen ergeben. Demzufolge spielt die Beurteilung und Auswertung der Signalintensitäten sowie die Normalisierung der Rohdaten eine zentrale Rolle. Fehler in den Signalintensitäten können eine Veränderung des Fingerprints und somit deren Ähnlichkeit zu anderen Fingerprints zur Folge haben und verändern somit die Zugehörigkeit zu einem Cluster und können auch ein Grund für das Cluster-Splitting sein.

Diese relativen Aussagen über relevante oder interessierende Gene müssen also mit anderen Methoden des direkten Vergleichs, wie beispielsweise des Northern-Blots, validiert werden. Dabei ist nicht zu erwarten, daß sich alle Ergebnisse in den Northern-Hybridisierungen bestätigen lassen, die Tendenz der Verteilung der Transkripte in der jeweiligen Zellpopulation sollte jedoch erkennbar sein.

Zum Erstellen der Bibliotheken wurden zur reversen Transkription der mRNA in cDNA Oligo(dT)-Primer eingesetzt. Die Wahl dieser Primer beruht auf dem Umstand der Hybridbildung des Oligo(dT) mit dem poly(A)-Tail der eukaryotischen mRNA, was für die Konstruktion einer Bibliothek aus mRNA zur Expressionsanalyse von klarem Vorteil ist. Daher enthalten alle klonierten Fragmente die 3`-Enden der codierenden Sequenzen, das 5`-Ende wird bei dieser Art Priming seltener kloniert. Es kommt praktisch nur bei Fragmenten vor, bei denen die reverse Transkription die gesamte Länge der mRNA in cDNA transkribieren konnte (sogenannte *full length clones*). Bei der Nutzung von *random* Primern, die sich auch entlang der mRNA in Richtung 5`-Ende anlagern, könnten größere Fragmente synthetisiert werden, und es entstehen cDNAs, die mit gleicher Wahrscheinlichkeit 3`- und 5`-Enden enthalten. Für das ONF zur Clustering-Analyse und zum Erstellen eines Unigene-Sets ist das Priming mittels oligo(dT) jedoch ausreichend. Für andere Anwendungen solcher cDNA-Bibliotheken, wie beispielsweise zur Antikörper-Produktion oder der Expression von Proteinen, kann es aber von Bedeutung sein, *random* Primer zur Erstellung der Bibliothek zu bevorzugen, um im Durchschnitt größere Fragmente oder einen höheren Anteil an *full length* Klonen zu erhalten.

Zum Erstellen der cDNA-Bibliotheken wurde sich für die gerichtete (direktionale) Klonierung der Fragmente entschieden, was durch die Nutzung von *Not*I-Primer- bzw. *Sal*I-Adaptern sowie zweier verschiedener Restriktionsendonucleasen (*Not*I, *Sal*I) zum Verdau der cDNAs (*sticky end*) realisiert wurde. Prinzipiell wäre auch eine ungerichtete Klonierung (*blunt end*) für die Fingerprinting- und Clustering-Analysen möglich. Der klare Vorteil der gerichteten Klonierung gegenüber der ungerichteten liegt darin begründet, daß bei einer Klonierung in einen Expressionsvektor, wie sie in dieser Arbeit durchgeführt wurde, beispielsweise eine Proteinexpression induziert werden kann, wofür die direktionale Klonierung eine Voraussetzung ist.

Unabhängig von ihrer Komplexität wurden bisher alle Oligos mit der gleichen Prozedur hybridisiert und gewaschen. Jedoch weisen sie einen z.T. sehr stark unterschiedlichen GC-Gehalt und damit voneinander abweichende optimale Hybridisierungsbedingungen auf, so daß ideale Bedingungen für ein Oligo nur suboptimal für ein anderes sind. Dementsprechend können auch die Hybridisierungen von unterschiedlicher Qualität sein. So müßten also jeweils auch verschiedene Bedingungen in den Hybridisierungen und den Waschschritten gewählt werden. Das ist jedoch nicht für jedes einzelne Oligo möglich, jedoch würde die Einteilung der einzusetzenden Oligonucleotide in verschiedene Gruppen mit jeweils gleichen optimalen Bedingungen zu einer verbesserten Qualität der Hybridisierungen führen.

Die Wahl der Oligos ist entscheidend für das Identifizieren der Sequenzen der cDNA-Klone. Dabei werden Oligos ausgewählt, die aufgrund ihrer Sequenz die Fähigkeit besitzen, eine Population von cDNA-Klonen jeweils zu etwa gleichen Teilen in hybridisierende und nicht hybridisierende Klone zu teilen. Dies kann in verschiedenen Organismen recht unterschiedlich sein und ist u.a. abhängig vom GC-Gehalt der DNA des zu untersuchenden Organismus, so daß es sich als sinnvoll erweisen könnte, verschiedene Oligosets zu erstellen, die spezifischer für bestimmte Organismen sind, wie pflanzenspezifische oder tierspezifische Sets, oder aber auch klassen-spezifische Sets.

Der GC-Gehalt eines Oligos ist für die Auswahl von Wichtigkeit, da die Oligos eine hohe Hybridisierungsstabilität aufweisen sollten. GC-Basenpaarungen formen aufgrund ihrer drei Wasserstoffbrücken stabilere Hybride als AT-Basenpaarungen. Ein zu hoher GC-Gehalt der Oktamere jedoch beeinflußt die Teilung der Klon-Pools negativ, d.h., es müßten mehr Oligos zur Hybridisierung eingesetzt werden, um den cDNA-Pool zu teilen (Herwig, R., 2000; Herwig, R. *et al.*, 2000).

Für die ONF-Hybridisierungen wurden Dekamere mit einem oktameren Kernbereich verwendet. Ein optimales Oligo sollte idealerweise eine Hybridisierungsfrequenz von etwa 50% aufweisen, also den Klon-Pool teilen in eine Hälfte der Klone, die zum Oligo komplementäre Bereiche aufweist und es binden kann und in eine andere Hälfte, die es nicht binden kann. Bei der Verwendung von Dekameren liegt diese Hybridisierungsfrequenz jedoch nur bei etwa 30%. Kürzere Oligos, wie beispielsweise radioaktiv markierte Heptamere, würden höhere Hybridisierungsfrequenzen aufweisen und somit das ONF noch effektiver gestalten und den Arbeitsaufwand weiter minimieren. Jedoch zeigten sie sich in Experimenten als weniger stabil, d.h., es traten häufiger unspezifische Hybridisierungssignale auf.

Durch Studien auf diesem Gebiet konnten DNA-Analoga entwickelt werden, die Peptid-Nucleinsäuren (PNA), deren Rückgrat strukturell ähnlich der Deoxyribose ist. Die Zucker-Phosphate jedoch wurden durch N-(2-aminoethyl)glycin-Einheiten ersetzt, einer Peptid-ähnlichen Struktur, an die die Nucleobasen gebunden sind. PNA-Oligomere hybridisieren an komplementäre Oligonucleotide und formen dabei wahrscheinlich Watson-Crick-Hoogsteen $(PNA)_2$-DNA Triplexe, die wesentlich stabiler sind als ihre korrespondierenden DNA-DNA-Duplexe und binden an dsDNA durch Strangverdrängung. PNA, die alle vier natürlichen Basen enthält, hybridisiert an komplementäre Oligonucleotide nach den Regeln der Watson-Crick-Basenpaarung, verhalten sich also hinsichtlich der Erkennung komplementärer Basenpaare wie DNA. Aufgrund dieser Eigenschaften sind sie besonders gut geeignet für Hybridisierungs-Experimente, wie dem ONF (Egholm, M. et al., 1993; Guerasimova, A. et al., 2001).

Aus der Möglichkeit der Anwendung der PNAs ergeben sich eine Reihe weiterer Vorteile. Beispielsweise können aufgrund ihrer stabileren Hybridbildung mit der DNA auch kürzere

Oligomere, wie Hexamere verwendet werden. Damit erhöht sich die Hybridisierungsfrequenz und der cDNA-Pool kann so mit dem Einsatz einer geringeren Anzahl von Oligos untersucht werden. PNAs lassen sich mit Fluoreszenzfarbstoffen markieren und mit dem Laserscanner detektieren. Durch den Einsatz fluoreszenzmarkierter PNAs können Signalinterferenzen, die durch Radioaktivität entstehen, verringert werden. Kombiniert man jeweils mehrere PNAs mit unterschiedlichen Fluoreszenzfarbstoffen, können mehrere Oligos gleichzeitig in den Oligonucleotid-Hybridisierungen eingesetzt werden, was den Arbeits- und Zeitaufwand um ein Vielfaches verringert. Fingerprinting-Hybridisierungen mit den PNAs sind jedoch bisher nur experimentell und noch in der Entwicklung und sind in näherer Zukunft fern von einem Einsatz in der Praxis und Routine.

Um die Verläßlichkeit der Aussagen des Fingerprintings und des sich daran anschließenden Clusterings zu gewährleisten, müssen bei der Auswertung der Daten der Hybridisierungen, der Analyse der Signalintensitäten, einige Punkte berücksichtigt und einige Fehlerquellen ausgeglichen werden.

Eine erste Möglichkeit der Überprüfung sind die verschiedenen Kontroll-Hybridisierungen, wie zum einen die Background-Hybridisierungen, die dazu dienen, die Menge an PCR-Produkt abschätzen zu können und den gleichmäßigen Transfer der Produkte auf die Membran zu überprüfen, oder zum anderen die Longprobe- oder Backhybridisierungen, die als Kontrolle für das Clustering genutzt werden.

Das Spotten von Duplikaten dient als weitere Möglichkeit zur Kontrolle, beispielsweise, um falsch positive Signale ausschließen zu können. Bei der Hybridisierung mit einem zu ihnen komplementären Oligo sollten beide Duplikate ein Signal geben, im Idealfall mit gleicher Signalintensität. Die Duplikatintensität ist ein Wert der Übereinstimmung der Signalintensitäten der jeweils zusammengehörigen Duplikate x und y zum Ausschluß falsch positiver Signale. Sie setzt sich zusammen aus dem maximalen und dem minimalen Wert des x- und des y-Duplikates eines Duplikatpaares einer Hybridisierung verglichen mit einem bestimmten Schwellenwert. Sie spiegelt den Grad der Übereinstimmung der Duplikate eines Duplikatpaares wider. Für die Analyse der Signalintensität wird zunächst das Verhältnis der Duplikatintensitäten berechnet (max : min). Bei optimaler Übereinstimmung liegt der Wert bei 1, der praktisch jedoch kaum erreichbar ist. Ist dieser Quotient zu groß, im Allgemeinen >1.5, so wird das Minimum der Duplikatsignale verwendet, ansonsten das arithmetische Mittel aus beiden Signalen. Die Duplikatintensität eines jeden Klons einer Hybridisierung wird in die sogenannte uhd-Datei (*unformatted hybridization data*) geschrieben, die später zur Normalisierung verwendet wird. Schlechte Werte der Duplikatintensität -zu große Abweichungen der Signalintensitäten der einzelnen Duplikate voneinander- können z.B. auch zustande kommen, wenn starke Guide Dots oder andere starke

Signale (lokale Erscheinungen) benachbarte schwächere Klone überstrahlen, somit ihre schwachen Signale verstärken und damit die x:y-Duplikatintensitäten verfälschen.

Betrachtet man alle 27.648 Duplikatsignale auf dem Filter, können die Duplikate auch zur numerischen Bewertung eines Hybridisierungsexperiments herangezogen werden, und es läßt sich der Korrelationskoeffizient berechnen (Herwig, R. *et al.*, 1999). Dieser Korrelationskoeffizient ist ein Wert der Übereinstimmung der Signalintensitäten aller Duplikate x und y einer Hybridisierung und dient wiederum dem Ausschluß falsch positiver Signale. Bei optimaler Übereinstimmung beider Duplikatsignale liegt der Wert bei 1.00, die Duplikatsignale sind perfekt korreliert, was praktisch jedoch kaum erreichbar ist. Nimmt der Korrelationskoeffizient kleine Werte an, so besteht eine schlechte Korrelation der Duplikatsignale. Liegt der Korrelationskoeffizient unter 0.6 (empirisch ermittelter Grenzwert), so wird das gesamte Experiment verworfen. Eine ausreichende Übereinstimmung beider Duplikatintensitäten ist noch bei einem Wert von etwa 0.8 gewährleistet, Werte über 0.9 stellen eine für die Praxis sehr gute Übereinstimmung dar.

Das Cluster-Splitting, also das Aufspalten eigentlich zusammengehöriger Cluster oder Co-Cluster in zwei oder mehrere Cluster, ist ein bekanntes Phänomen der Clustering-Analysen. Im Allgemeinen ist es auf die unterschiedlichen Klonlängen der ein und dasselbe Gen repräsentierenden cDNA-Fragmente zurückzuführen, die, daraus resultierend, einen signifikant unterschiedlichen Fingerprint aufweisen. Das ist insbesondere dann der Fall, wenn der Fingerprint nicht sehr viele positive Hybridisierungssignale enthält (Meier-Ewert, S. *et al.*, 1998).

Als Beispiele für das Aufsplitten eigentlich zusammengehöriger Cluster können u.a. GAPDH und α-Tubulin genannt werden. GAPDH splittet in mindestens 3 größere Cluster (Co-Cluster 22, 29, 199). Dabei fällt auf, daß die einzelnen Cluster jeweils nur aus entweder Jurkat- bzw. NKL-Klonen zusammengesetzt sind (NKL:Jurkat jeweils 132:0, 0:111, 0:25). Ähnlich verhält es sich auch mit dem α-Tubulin, das in mindestens 5 Cluster (Co-Cluster 37 (85:1), 63 (0:64), 125 (0:39), 610 (8:1), 978 (0:6)) splittet.

Da die reverse Transkriptase-Reaktion oligo(dT) geprimt ist, beginnt die reverse Transkription immer am 3`-Ende der mRNA und setzt sich in Richtung des 5`-Endes fort. An unterschiedlichen Positionen der Erststrang-Synthese der mRNA bricht die Polymerisation irgendwann ab, so daß unterschiedlich große cDNA-Fragmente auch derselben mRNA entstehen können. Relativ kurze Bereiche, um die ein Fragment länger oder kürzer ist, reichen aus, um den Fingerprint so unterschiedlich werden zu lassen, daß die cDNAs in verschiedene Cluster geordnet werden.

Gibt es interne poly(A)-Bereiche in der mRNA, so kann sich der poly(dT)-Primer anstatt am 3`-Ende der mRNA auch weiter in 5`-Richtung anlagern und so cDNA-Fragmente entstehen lassen, die zwar eine gleiche Klonlänge aufweisen können, aber die zum einen nicht den gleichen 3`-

Startpunkt wie seine Homologe aufweisen und zum anderen viel weiter in 5‘-Richtung synthetisiert werden können und somit einen stark unterschiedlichen Fingerprint aufweisen werden.

In den speziellen, hier genannten Fällen konnten jedoch keine unterschiedlichen Fragmentlängen der Cluster nachgewiesen werden, die somit wahrscheinlich auch nicht Ursache für das Aufsplitten der Cluster sein können. Speziell in dem geschilderten Fall der Co-Cluster 22 und 29 ist lediglich ein Größenunterschied von etwa 20 - 50 bp festzustellen, wodurch sich ein Splitting nicht erklären läßt. Im direkten Vergleich der jeweiligen Visualisierungen der Fingerprints jedoch fällt auf, daß das Co-Cluster 22 (NKL) zwar einen im Prinzip ähnlichen Fingerprint aufweist wie das Co-Cluster 29 (Jurkat), jedoch z.T. auch weniger oder schwächere Signale enthält.

Gerade beim Co-Clustering zweier unterschiedlicher Bibliotheken ist das Phänomen des Splittings ein bekanntes Problem. Es gibt viele verschiedene Faktoren innerhalb der vielfältigen Arbeitsschritte des ONF, die die Qualität zweier unterschiedlicher Bibliotheken zum Teil stärker beeinflussen können. Durch verschiedene Normalisierungsmöglichkeiten wird ein Großteil dieser Fehler innerhalb einer Bibliothek ausgeglichen, so daß die Daten der Hybridisierungen reproduzierbar werden. Jedoch gibt es einige zusätzliche Fehlerquellen, die auftreten können, wenn zwei oder mehrere Bibliotheken miteinander verglichen werden sollen (unterschiedliche Qualitäten der verwendeten Membranen, unterschiedlich starke Amplifikation der Bibliotheken während der PCR, verschiedenartige Bedingungen beim Spotting der Bibliotheken (damit verbunden ist eine verschiedene Transfermenge an amplifizierten Produkten), die Bedingungen bei den Hybridisierungen an sich, das Versagen einzelner Oligo-Hybridisierungen u.a.). Alles dies sind Möglichkeiten, die in veränderten, eventuell schwächeren Signalen resultieren und somit die Ergebnisse, und damit den Fingerprint der Klone verändern können, so daß ein homologes Cluster weniger Fingerprints aufweisen kann. Das Co-Clustering zweier Bibliotheken erfolgt anhand der Ähnlichkeiten der jeweiligen Consensus-Fingerprints. Der Consensus-Fingerprint eines Cluster spiegelt die Fingerprints der zugehörigen Klone wider. Veränderte Fingerprints können also zu veränderten Consensus-Fingerprints eines Clusters und damit zum Aufspalten der Co-Cluster führen. Somit läßt sich auch das Cluster-Splitting mehrerer ein und dasselbe Gen repräsentierender Co-Cluster erklären. Northern-Hybridisierungen interessierender Cluster/Gene sollten als Nachweis und Bestätigung der Ergebnisse des Oligonucleotid Fingerprintings und Clusterings dienen.

Die Gene, die bisher mit der konventionellen Weise des Sequenzierens identifiziert werden konnten, zählen eher zu den in der Zelle häufiger vorkommenden Genen. Die meisten Gene, die bisher noch unbekannt sind, sind wahrscheinlich seltener transkribierte Gene (sogenannte *low copy genes*). Die Cluster, die im ONF gefunden werden, sind fast alle relativ klein, d.h., sie stehen für weniger in der Zelle vorkommende Gene. So ist es sehr wahrscheinlich, daß mit dem ONF solche neuen, noch unbekannten Gene gefunden werden können.

Basierend auf der bekannten Sequenz der hybridisierten Oligonucleotide ergibt sich ein experimenteller Fingerprint der Klone. Der Consensus-Fingerprint eines Clusters kann mit den theoretischen Fingerprints, die von Genen aus den Datenbanken erstellt wurden, verglichen werden. So können in der Datenbank bekannte Gene schnell identifiziert und solche, die keinen signifikanten Treffer aufweisen, als eventuelle neue Kandidatengene ermittelt werden.

In den hier durchgeführten Untersuchungen konnten 135 Cluster ermittelt werden, die keiner bekannten Sequenz zugeordnet werden konnten. Diese könnten für mögliche neue Kandidatengene, Gene immunologischer Relevanz codieren, bei denen es von großem Interesse wäre, die Sequenz mittels konventioneller Sequenzierung zu ermitteln. Von diesen 135 unbekannten Genen werden 43 differentiell exprimiert. 19 Gene davon kommen differentiell in der Zell-Linie Jurkat zur Expression, 24 in der NKL-Linie. Bei all diesen Angaben, besonders bei den nicht näher identifizierten Clustern, muß allerdings auch das Vorkommen des Cluster-Splittings berücksichtigt werden, welches die Angaben etwas relativiert. So ist es auch möglich, daß nicht alle der als unbekannt identifizierten Cluster wirklich neue Gene darstellen. Es ist denkbar, daß diese bekannt sind, aber aufgrund eines veränderten Fingerprints im Vergleich zu den theoretischen Fingerprints der Datenbanken nicht identifiziert werden konnten.

Die Bestrebungen der Identifizierung bisher noch unbekannter Gene gehen dahin, mehr Licht in das Wirkungsgefüge der Proteine und Gene bei der Entstehung und Proliferation der Krebserkrankungen und anderer Erkrankungen des Immunsystems zu bringen. Die in dieser Arbeit gefundenen, bisher noch unbekannten Gene, können einen Beitrag an der Aufklärung dieser Vorgänge leisten. Sie können helfen, neue Targets für Leukämietherapien zu finden und neue Therapeutika zu entwickeln. Von wesentlicher Bedeutung für die onkologische Forschung ist dabei auch die Identifizierung prognostischer Marker für Leukämien sowie die Aufklärung der Medikamentenresistenz. Damit eröffnet sich die Perspektive, in naher Zukunft Leukämiepatienten die Möglichkeit einer gezielteren Therapie zuteil werden zu lassen, diese malignen Erkrankungen besser bekämpfen oder gar vorbeugen zu können.

Im Vergleich zweier Bibliotheken gibt es gemeinsam exprimierte Cluster, also solche, die in beiden Zellarten in etwa gleich stark exprimiert werden, und solche Cluster, die differentiell exprimiert in einer Zellart vorkommen, somit häufiger in einer der zu untersuchenden Bibliotheken bzw. Zellarten zu finden sind, bis hin zu einer ausschließlichen Expression bestimmter Gene in nur einer der zu vergleichenden Bibliotheken.

Einige von diesen ausschließlich in einer Zellart vorkommenden Genen sind beispielsweise das HLA-A2 MHC Klasse I Antigen, das in einem Verhältnis von 47:2 in der NKL-Zell-Linie vorkommt, oder das CD3E-Gen, das mit einem Verhältnis von 0:12 ausschließlich in der T-Zell-Linie Jurkat zur Expression kommt. Bei der Interpretation der Clustering-Ergebnisse ist jedoch zu beachten, daß diese lediglich eine Tendenz anzeigen und keine exakten Werte der Expression des

Gens in einer Zelle wiedergeben. Die Werte stehen für die in der Bibliothek repräsentierten Gene (ermittelt anhand der Fingerprints der Klone), die somit zwar Repräsentanten für die Transkripte in der Zelle sind, jedoch nur einen relativen Expressionslevel widerspiegeln. Bei einigen, als differentiell exprimiert identifizierten Clustern kann es sich auch um gesplittete Cluster handeln, besonders wenn es sich dabei um solche Cluster handelt, in denen nur Klone in einer größeren Anzahl der einen Bibliothek vorkommen und keine Klone der jeweils anderen Bibliothek vertreten sind. So ist immer eine Validierung mit direkten Nachweismethoden notwendig.

Die Co-Cluster 1 bis 4 mit jeweils mehr als 1.000 Mitgliedern in der einen sowie keinen oder nur sehr wenigen Mitgliedern in der anderen Bibliothek (Co-Cluster 2) stellen sehr große Cluster mit differentieller Expression dar. Bei diesen Clustern könnte es sich um solche Fälle des Cluster-Splittings handeln. Die Co-Cluster 1, 3 sowie 4 beinhalten ausschließlich Klone der NKL-Bibliothek, wohingegen dem Co-Cluster 2 wenigstens 4 Jurkat-Klone zugeordnet werden konnten. Möglich wäre auch, daß es sich bei diesen Clustern um Sequenzen handelt, die nur in bestimmten Bereichen hohe Homologie aufweisen, jedoch kein Gen, sondern eher eine Genfamilie repräsentieren. Solche Sequenzen würden auch einen ähnlichen Fingerprint ergeben und so in ein Cluster geordnet werden.

Die Klone der Co-Cluster 3 und 4 ließen sich jedoch auch in Wiederholungen der PCR nicht amplifizieren. In einer sich anschließenden Plasmidpräparation und Restriktion mit den beiden Enzymen *Not*I und *Sal*I ließ sich nur eine Vektorbande bei etwa 3.5 kb nachweisen, jedoch kein enthaltenes Insert. Die Co-Cluster 3 und 4 enthalten somit keine amplifizierbaren Inserts, weisen aber spezifische Fingerprints auf. Beide Co-Cluster weisen zudem einen sehr ähnlichen Fingerprint auf, bilden möglicherweise ein gemeinsames Co-Cluster. Es wurde zuerst vermutet, daß es sich bei diesen Klonen möglicherweise um Verunreinigungen bei der PCR handeln könnte, was allerdings bei einer Doppelselektion mit Ampicillin und Kanamycin sowie der Nutzung spezifischer Primer nahezu auszuschließen ist, auch waren keine Regelmäßigkeiten hinsichtlich der Plattenpositionen festzustellen. Auch wurde diskutiert, daß eventuell Oligos mit Vektormatch zum Einsatz kamen. Dies jedoch sollte von vornherein ausgeschlossen werden und konnte auch nach nochmaliger Überprüfung nicht bestätigt werden.

Im Falle dieser beiden Co-Cluster 3 und 4 ist es auch möglich, daß es sich dabei um sogenannte „leere“ Klone, also Klone ohne Inserts handelt. Auch solche Klone geben in Hybridisierungen Signale, wenngleich auch nur sehr schwache, die als das sogenannte Hintergrundrauschen (*background noise*) bezeichnet werden. Bei der Normalisierung wird jedem Signal ein numerischer Wert zugeordnet, dem stärksten Signal eines Filters mit einer bestimmten Anzahl von gespotteten Klonen der Wert=1, dem zweitstärksten der Wert 1-1/N, dem dritten 1-2/N usw., bzw. für alle Hybridisierungsexperimente wiederum dem stärksten Signal der Wert=1, dem zweitstärksten 1-1/p usw., wobei N die Anzahl der Klone auf der Membran und p die Anzahl der Experimente darstellt.

So wird also auch diesen sehr schwachen Signalen während der Normalisierung ein numerischer Wert zugeordnet, so daß sich letztendlich ein Fingerprint daraus ergeben kann.

Zur Analyse kommen nur 85% der Hybridisierungssignale, die 15% schwächsten Klone, also alle Klone mit einer Signalintensität unterhalb dieses Schwellenwertes, werden aus der Analyse ausgeschlossen. Gibt es jedoch mehr Klone mit sehr schwachen Signalen, beispielsweise 25%, so bleiben immer noch 10% der Klone mit eben diesen schwachen Signalen in die Analyse einbezogen und ergeben Fingerprints. So ließe sich erklären, warum im Falle der Co-Cluster 3 und 4 keine Amplifikation von Fragmenten möglich war, jedoch Fingerprints errechnet werden konnten. Die Anzahl von 2.477 Klonen in diesen Clustern entspräche auch in etwa dem 8%-igen Anteil an in der PCR festgestellten Klonen ohne Insert.

In einer cDNA-Bibliothek ist das gesamte Spektrum aller in der Zelle oder einem Gewebe vorkommenden mRNAs etwa im Verhältnis ihrer Expression enthalten. Stark exprimierte Gene sind demnach häufiger in der Bibliothek vertreten als nur schwach exprimierte. cDNA-Bibliotheken, die aus mRNA erstellt wurden, sind demnach mehrfach redundant, da viele Gene, die in der Zelle stärker exprimiert werden als andere, in der Bibliothek überrepräsentiert sind. Um nun aber für zukünftige Experimente, wie der Komplex-Hybridisierung oder der Sequenzierung, den Analyse-Aufwand so gering wie möglich zu halten, ist es notwendig, diese Redundanz zu reduzieren. Es ist von Vorteil und auch ausreichend, daß jeweils nur ein Vertreter dieser Klone -jedes ein Gen repräsentierendes cDNA-Fragment- im Idealfall nur einmal vorkommt. Solch eine Bibliothek nicht redundanter cDNA bezeichnet man als ein Unigene-Set.

Um diese Reduktion der Redundanz zu realisieren, können mit Hilfe des ONF und mittels Clustering-Analysen gleiche Sequenzen in Cluster zusammengefaßt werden. Jeweils nur ein repräsentativer Klon eines jeden Clusters wird nun selektiert und in Microtiterplatten neu angeordnet (*rearraying*) und somit ein Set nicht (oder nur gering) redundanter Klone erstellt. Ein solches Unigene-Set hat aufgrund der geringen Redundanz der Klone einige Vorteile. So kann der Arbeitsaufwand beim konventionellen Sequenzieren um ein Vielfaches reduziert werden.

Das Jurkat-Unigene-Set wurde aus einer T-Lymphocyten-Linie erstellt und stellt damit eine umfassende nicht redundante Bibliothek T-Zell-spezifischer Gene dieser Zell-Linie dar, das auch besonders Gene immunologischer Relevanz exprimiert, die sonst in anderen Geweben und Organen nicht oder kaum zur Expression gelangen. Da diese Zell-Linie ursprünglich aus Leukämiepatienten isoliert wurde, sind im Unigene-Set desweiteren auch solche Gene exprimiert -oder stärker exprimiert-, die in die Entstehung und Progression von malignen Erkrankungen involviert sein können. Zum Erstellen des Jurkat Unigene-Sets wurden die jeweiligen Consensus-Klone der Cluster selektiert. Der Consensus-Klon eines Clusters ist der das Cluster am besten repräsentierende Klon, der, zu dem die Ähnlichkeit aller anderen Klone am höchsten ist (der durchschnittliche Fingerprint aller dem Cluster zugeordneten Fingerprints). Das Jurkat-Unigene-

Set besteht aus 10.506 Klonen in 28 MTP, was einer 2.25-fachen Normalisierung der Bibliothek entspricht. Aufgrund des Cluster-Splittings ist es jedoch nicht auszuschließen, daß einige Gene mehrfach repräsentiert sind. Klone nicht eindeutig identifizierter Cluster und Singletons sollten für die folgenden Komplex-Hybridisierungen sequenziert werden, um ihre Identität zu verifizieren, was sich nun aufgrund der reduzierten Redundanz als wesentlich effektiver gestaltet.

Für einige Applikationen, wie der Sequenzanalyse oder der Proteinexpression, wäre es von größerer Bedeutung, nicht den Consensus-Klon, sondern den längsten Klon des Clusters zu selektieren. Der längste Klon wäre der mit den meisten Hybridisierungssignalen verglichen mit allen anderen Klonen desselben Clusters. Dann ist die möglichst längste DNA-Sequenz erwünscht, um beispielsweise eine möglichst vollständige Aminosäuresequenz zu translatieren. Um jedoch weitere Hybridisierungen, wie Komplex-Hybridisierungen, durchzuführen, ist es ausreichend, den jeweiligen Consensus-Klon der Cluster für das Unigene-Set zu selektieren.

Es liegen auch die entsprechenden Daten zur Erstellung eines Unigene-Sets der NKL-Zell-Linie, und damit einer Natural Killer-Zell spezifischen, nicht redundanten Bibliothek vor. Dieses Set besteht aus insgesamt 13.094 Klonen, davon 2.169 repräsentative Klone aus Clustern mit einer Mindestgröße von 2 sowie 10.925 Singletons, was einer 2.7-fachen Normalisierung entspricht. Ein solches Unigene-Set könnte weiterhelfen, in näherer Zukunft genauere Einblicke in die Natur der NK-Zellen zu erhalten, über die bisher weit weniger bekannt ist, als über ihre nah verwandten B- oder T-Lymphocyten.

Mit dem Jurkat- bzw. dem NKL-Unigene-Set stehen bedeutende Tools zweier Leukämie-Zell-Linien (Zellen des Immunsystems) für weitere, speziellere Untersuchungen zur Verfügung, die detailliertere Einblicke in die Carcinogenese und Proliferation geben können. Komplex-Hybridisierungen ermöglichen es, die Genexpression verschiedener Zellpopulationen oder Gewebe, beispielsweise verschiedener Lymphocytenisolate von Patienten, direkt miteinander zu vergleichen und können zum Screening von Leukämiepatienten und zum Nachweis minimal residualer Zellen genutzt werden, sowie auch in Studien zur molekularen Charakterisierung der Medikamentenresistenz und um Targets für neue Medikamente zu finden.

Als eine sehr erfolgversprechende Applikation des T-Zell- und NK-Zell-spezifischen Unigene-Sets gelten die Komplex-Hybridisierungen mit unterschiedlichen und sehr interessanten Zielsetzungen. Dabei kann mRNA verschiedener Patienten bestimmter Leukämietypen oder unterschiedlicher Zeitpunkte der Erkrankung hybridisiert werden. So können umfangreiche Studien zur Klassifizierung der Leukämien durchgeführt werden und mit der Identifizierung neuer Subtypen eine detailliertere Diagnose ermöglichen. Eine möglichst genaue Diagnose der Leukämieform ist Voraussetzung für die Wahl der entsprechenden Therapie und stellt somit die Grundlage dafür dar, die Therapierbarkeit dieser verschiedenen Subtypen zu verbessern und damit die Überlebenschancen der Patienten zu erhöhen. Ebenso ist die Identifizierung von

Risikofaktoren, an einer bestimmten Krebserkrankung zu erkranken, sowie das Aufklären prognostischer Faktoren für die Therapierbarkeit von wesentlichem Interesse für die Onkologie der heutigen Zeit. Mit dem Nachweis minimal residualer Zellen zu verschiedenen Zeitpunkten während der Dauer einer Behandlung bzw. vor und nach einer spezifischen Therapie kann das Fortschreiten der Krankheit (Progression) oder das Fortbestehen einer Remission -und damit der Therapieerfolg- überwacht werden.

Ein noch recht neues Forschungsgebiet befaßt sich mit den molekularen Grundlagen der Medikamentenresistenz. Hybridisierungen mit RNA von Patienten, bei denen es zum Versagen der Therapie kam, könnten dabei helfen, diese Resistenz -die häufig mit der Ausprägung einer Mehrfach-Medikamentenresistenz verbunden ist- aufzuklären. In engem Zusammenhang damit steht die Identifizierung neuer Targets für neue Medikamente und Therapien, wie beispielsweise auch der Krebsvorbeugung und -bekämpfung durch Vaccination.

Das Oligonucleotid Fingerprinting kann andere Methoden der Genexpressionsanalyse mittels Chiptechnologie -wie beispielsweise der Fa. Affymetrix- nicht ersetzen, jede hat ihre Vorzüge und Nachteile, die gemeinsam einander ergänzen können. Aber es vereint den Umfang an Daten mit einem relativ geringen Aufwand an Zeit und Kosten. Es kann als vorbereitendes Experiment für eine Reihe weiterführender Experimente eingesetzt werden und dient dazu, einen schnellen und globalen Überblick über die Genexpression eines Gewebes zu erlangen.