4. Diskussion

4.1. Funktionelle Genomanalyse zur Identifizierung stadienspezifischer Gene von *Trypanosoma brucei*

4.1.1. Allgemeine Fragestellung

Die Entwicklung spezifischer Medikamente gegen parasitäre Erkrankungen erfordert genaue Kenntnisse der Biologie der Parasiten. Der molekulare Bauplan aller Lebewesen steht in der DNA-Sequenz des Genoms festgeschrieben. Durch die Sequenzierung ganzer Genome unterschiedlicher Organismen wird der Grundstein für das Verständnis der zellulären Vorgänge gelegt. Für eine Vielzahl von pathogenen Organismen liegt bereits die komplette genomische Sequenzinformation vor. Für *Trypanosoma brucei* wird die Sequenz in naher Zukunft bekannt sein. Einem Teil dieser Sequenz kann durch computergestützte Vergleiche mit der DNA-Sequenz bereits bekannter Gene anderer Organismen eine Funktion zugeordnet werden. Die Funktion des Großteils der identifizierten Gene erschließt sich allerdings nicht allein aus der Sequenz und muß deshalb durch experimentelle Arbeiten aufgeklärt werden. Einen Hinweis auf die Funktion eines Gens können die Umstände geben, unter denen ein Gen exprimiert wird.

Bei höheren Eukaryonten wird die Genaktivität hauptsächlich durch die Initiation der Transkription reguliert. Aus diesem Grund beginnt man bei der Suche nach differentiell exprimierten Genen zunächst mit der Untersuchung der Regulation auf der mRNA-Ebene. Im Gegensatz zu höheren Eukaryonten wird die Transkription bei Kinetoplastiden nicht auf der Ebene der Initiation reguliert, sondern erfolgt nachträglich durch die differentielle Degradation der mRNA (Clayton, 2002). Das Ausmaß der Steuerung der Genaktivität durch Regulation der Transkriptmenge in *T. brucei* ist bislang nicht bekannt. Bei essentiellen Gene wie z.B. den Oberflächen-Antigenen findet man jedoch eine starke Transkriptionskontrolle. Man kann deshalb vermuten, daß Gene, deren Expression große Unterschiede in verschieden Stadien zeigen, für das Überleben der Parasiten besonders wichtig sind.

Die Forschung an *T. brucei* wird zu einem großen Teil an den *in vitro* kultivierbaren, replikativen Formen des Parasiten vorgenommen, der *long slender* Blutbahnform und der prozyklischen Form. Die Identifizierung der für diese Stadien spezifischen Gene ermöglicht einen interessanten Einblick in das Ausmaß der Transkriptionsregulation in Kinetoplastiden

und schafft eine wichtige Grundlage für weitergehende experimentelle Arbeiten an diesen Stadien.

4.1.2. Repräsentative Differenzanalyse

4.1.2.1. Vor-und Nachteile der Repräsentativen Differenzanalyse

Die Repräsentative Differenzanalyse ist eine auf der subtraktiven Hybridisierung komplexer DNA-Proben beruhende Methode, die zur Isolierung differentiell exprimierter Gene angewendet werden kann. Die Methode der RDA wurde ursprünglich zum Auffinden spezifischer Unterschiede zwischen Genomen entwickelt und ist dazu auch erfolgreich angewendet worden (Lisitsyn und Wigler, 1993). Beim Einsatz der RDA zur Identifikation differentiell exprimierter Gene zeigen sich jedoch gewisse Limitierungen, die durch Unterschiede in der Reassoziationskinetik der cDNA gegenüber genomischer DNA bedingt sind. Die Unterschiede in der Häufigkeit einzelner Transkripte in den komplexen cDNA-Gemischen können 10-100.000-fach sein (Bishop *et al.*, 1974; Lockhart und Winzeler, 2000). Bei der subtraktiven Hybridisierung werden die doppelsträngigen RDA-Fragmente denaturiert. Die Geschwindigkeit der Reassoziation komplementärer Fragmente hängt von deren Konzentration in der Hybridisierungslösung ab. Je höher die Konzentration der *Tester*-Fragmente gegenüber den kompetitierenden *Driver*-Fragmenten ist, desto wahrscheinlicher ist die Bildung einer *Tester/Tester*-Duplex. Bei einer großen Anzahl differentieller Transkripte in der Repräsentation werden deshalb durch die Amplifikationsschritte bevorzugt Fragmente von stark exprimierten Genen durch die RDA angereichert, bei denen die Expressionsunterschiede besonders ausgeprägt sind (Sagerstrom *et al.*, 1997). Aus diesem Grund kann mit der RDA nur eine relativ geringe Anzahl differenziell exprimierter Gene identifiziert werden. Zudem können manche Gene grundsätzlich nicht durch die RDA identifiziert werden, weil sie kein amplifizierbares Fragment enthalten. Dies kann zum Beispiel der Fall sein, wenn der Abstand zwischen zwei Restriktionsschnittstellen zu groß oder klein ist. Ein weiterer Nachteil der RDA ist der geringe Durchsatz dieser Methode. So können immer nur zwei Proben miteinander verglichen werden. Aufgrund der vielen verschiedenen Arbeitsschritte ist es nicht möglich, den Durchsatz der RDA durch Automatisierung zu erhöhen. Ein Vorteil der RDA liegt jedoch darin, daß sie auch bei Organismen, für die keine Sequenzdaten vorliegen, erfolgreich angewendet werden kann. Außerdem erfordert die Durchführung einer RDA nicht die Anschaffung spezieller Geräte,

und stellt somit ein in jedem Labor kostengünstig durchzuführendes Verfahren zur Identifikation differenziell exprimierter Gene dar.

4.1.2.2. Die Repräsentative Differenzanalyse zur Identifizierung stadienspezifischer Gene

Bei der Anreicherung Blutbahn-spezifischer Fragmente sind mit der RDA vor allem Fragmente von VSG-Genen isoliert worden. Bei Wildtyp-Isolaten von *T. brucei* wird immer nur ein VSG-Gen exprimiert. Durch die RDA von Trypanosomen einer *in vitro* Kultur wurden dagegen vier verschiedene VSG-Gene identifiziert. Die Heterogenität der VSG-Expression in der *in vitro* Kultur von Trypanosomen wurde auch anderweitig beobachtet (van Deursen *et al.*, 2001). Sie ist vermutlich auf den in der Kultur fehlenden Selektionsdruck durch das Immunsystem des Wirts zurückzuführen.

Die essentielle Rolle der Oberflächenantigene für den Parasiten und deren differentielle Expression in der Blutbahnform und der prozyklischen Form steht zwar außer Zweifel, ist aber bereits hinlänglich bekannt. Durch die stark differentiell exprimierten Oberflächenproteine in der Blutbahnform und der prozyklischen Form wird die Anreicherung anderer differentieller Fragmente, die in geringerer Konzentration vorliegen, stark erschwert. Wie die Sequenzierungsergebnisse zeigten, entfiel ein Großteil der gefundenen differentiellen RDA-Fragmente auf die Oberflächenantigene, wogegen nur wenige andere Sequenzen gefunden werden konnten. Eine Möglichkeit, dieses Problem zu umgehen, wäre der Zusatz von VSG-Kompetitor-Fragmenten zur *Driver*-Population während der RDA. Wegen des Phänomens der Antigen-Variation müssten dazu allerdings für jeden Zellkulturansatz die exprimierten VSG-Gene vor der Durchführung der RDA bestimmt werden. Die Anwendung der RDA zur Isolation differentiell exprimerter Gene in *T. brucei* ist deshalb eher bei solchen Fragestellungen aussichtsreich, bei denen man nur sehr wenige differentiell exprimierte Gene erwartet.

Ein weitere interessantes Ergebniss der RDA war der relativ hohe Anteil differenzieller Sequenzen, die keinen ORF enthielten. Solche Sequenzen sind auch durch die Microarrayanalyse identifiziert worden. Die Funktion dieser Sequenzen in Trypanosomen ist unklar.

4.1.2.3. Eigenschaften des genomischen **T. brucei** *Microarrays*

Mit Microarrays kann die Transkripthäufigkeit von tausenden Genen zugleich untersucht werden. Die Herstellung genspezifischer Microarrays erfordert jedoch die Kenntnis der

Sequenz der zu untersuchenden Gene. Die Konstruktion eines genspezifischen Arrays für die Expressionsanalyse aller Gene von *T. brucei* ist zur Zeit nicht möglich, da das Genom noch nicht vollständig sequenziert ist. Die Herstellung genspezifischer Arrays wäre zudem kostspielig, da für jedes Gen ein Primerpaar synthetisiert werden müsste. Als Alternative zur direkten Amplifikation der DNA-Sonden aus genomischer DNA können Klone aus cDNA-Bibliotheken zur Herstellung genspezifischer DNA-Sonden verwendet werden. Diese Vorgehensweise erfordert allerdings das Vorliegen möglichst vollständiger, normalisierter cDNA-Bibliotheken. Eine vollständige cDNA-Bibliothek, die sämtliche Lebensstadien von *T. brucei* umfaßt, existiert ebenfalls nicht. Um dennoch eine genomweite Transkriptionsanalyse durchführen zu können, wurde ein Microarray aus genomischen Fragmenten hergestellt. Die Auswahl einer genomischen Bibliothek zur Herstellung eines Arrays bietet gegenüber der Verwendung von cDNA-Bibliotheken den Vorteil, daß selten transkribierte Sequenzen mit der gleichen Wahrscheinlichkeit in der Bibliothek enthalten sind wie häufige vorkommende Transkripte. Auch bei normalisierten cDNA-Bibliotheken kann eine völliger Ausgleich in der Repäsentation der cDNA-Fragmente nicht gewährleistet werden. Ein weiterer Vorteil genomischer Arrays gegenüber genspezifischen Arrays ist, daß auch nichttranslatierte Sequenzen, wie 3'UTRs oder kurze ORFs, die bei der Sequenzannotation häufig unberücksichtigt bleiben, auf einem genomischen *shotgun*-Array enthalten sind.

Zur Herstellung des genomischen Arrays wurden Klone einer sogenannten *Shotgun*-Bibliothek verwendet. Das *Shotgun*-Verfahren bezeichnet eine Methode, bei der das Genom in Fragmente zufälliger Sequenz und Länge zerlegt wird. Die verwendeten Klone enthalten gescherte DNA und sind durch Grössenselektion auf eine Insertlänge im Grössenbereich von 2 bis 2,5 kb beschränkt worden (El-Sayed *et al.*, 2003).

Ein Nachteil dieser Fragmentgröße ist, daß bei einer durchschnittlichen Grösse der Gene von 1-2 kb und der intergenischen Regionen von 0,5-1 kb (El-Sayed *et al.*, 2000) die Wahrscheinlichkeit relativ groß ist, daß ein Klon die Sequenz zweier verschiedener Gene enthält. Dadurch können schwache Fluoreszenzsignale einer Sequenz von starken Signalen einer komplementär markierten Probe überlagert werden. Bei einer kleineren Insertlänge müsste man allerdings wesentlich mehr Spots auf dem Array unterbringen, um die gleiche Abdeckung des Genoms zu erreichen. Die Anzahl der Spots, die pro Array gespottet werden können, ist jedoch aus technischen Gründen limitiert. Diese Limiterung hat ihre Ursache darin, daß die Spots ineinanderlaufen können, solange sie noch nicht eingetrocknet sind. Durch die technischen Limitierungen wurde die Anzahl der Spots pro Array auf 22.177 begrenzt.

Ein weiterer Nachteil der Verwendung genomischer Fragmente zur Konstruktion der Microarrays gegenüber der genspezifischer Arrays ist, daß die Unterscheidung der Expression von verschiedenen Genen einer Genfamilie wegen der undefinierten Fragmente kaum möglich ist.

Die Abdeckung des Genoms von *T. brucei* durch den Array kann man bestimmen, indem man die Wahrscheinlichkeit berechnet, mit der eine gegebene Sequenz auf dem Array enthalten ist (Clarke und Carbon, 1976). Bei zufälliger Verteilung der DNA-Fragmente auf dem Genom ist die Wahrscheinlichkeit (*P*), daß eine bestimmte DNA Sequenz in *N* Klonen enthalten ist gleich:

$$P = 1 - e^{N \cdot \ln(1-f)}$$

P = Wahrscheinlichkeit, mit der eine bestimmte Sequenz in der Bibliothek enthalten ist

N = Anzahl der Klone in der Bibliothek

f = Verhältnis von Insertgröße zu Genomgröße

Daraus ergibt sich bei N = 20.640 Klonen einer durchschnittlichen Größe von 2,0 kb und einer Größe des chromosomalen Gesamtgenoms von 35 Mb eine Abdeckung des Microarrays von circa 82% des gesamten chromosomalen Genoms von *T. brucei*. Da die Erfolgsrate der PCR bei etwa 90% lag, beträgt die tatsächliche Abdeckung des Arrays ungefähr 74%.

4.2. Optimierung der Datenbearbeitung

4.2.1. Normalisierung

Im ersten Schritt der Datenbearbeitung werden die Intensitätswerte des roten und grünen Kanals aneinander angepasst (Normalisierung). Diese Anpassung ist notwendig, um die Signalintensitätswerte beider Farben für jeden Spot vergleichen zu können. Zur Normalisierung der Intensitätswerte stehen verschiedene Verfahren zur Verfügung, denen der Gedanke zugrunde liegt, daß die Expression der Mehrzahl der untersuchten Gene unverändert bleibt und dadurch der Mittelwert der Differenzwerte 1 beträgt (Chen *et al.*, 1997). Eine einfache Möglichkeit zum Angleich der Signalintensitäten besteht in der globalen Mittelwert-Normalisierung. Bei dieser Normalisierungsmethode wird ein Korrekturfaktor aus dem Verhältnis der Gesamtsumme der Intensitäten beider Farbkanäle berechnet. Dieser Wert wird dann von den logarithmierten Differenzwerten subtrahiert (Quackenbush, 2002). Neben dieser Art der Normalisierung sind eine Reihe anderer Methoden für die Normalisierung vorgeschlagen worden, wie zum Beispiel eine lineare Regressionsanalyse (Beissbarth *et al.*, 2000; Fellenberg *et al.*, 2001), Varianz-Regulierung (Durbin *et al.*, 2002; Huber *et al.*, 2002)

oder die Anwendung von ANOVA-Modellen (Kerr *et al.*, 2000). Der Nachteil dieser Methoden ist, daß lediglich lineare Abhängigkeiten ausgeglichen werden können. Durch grafische Darstellungen wurde jedoch gezeigt, daß zwischen den Differenzwerten und der Signalintensität der Spots ein nicht-linearer Zusammenhang besteht, der durch eine lineare Transformation wie der Mittelwert-Normalisierung oder einer linearen Regression nicht beseitigt werden kann (Abbildung 3.15 und 3.6). Eine nicht-lineare Abhängigkeit der Differenzwerte von der Signalintensität ist auch in anderen Veröffentlichungen diskutiert worden (Dudoit *et al.*, 2000; Yang *et al.*, 2002a; Yang *et al.*, 2002b). Aus diesem Grund ist eine lokal-gewichtete Regression zur Glättung von Punktdiagrammen (LOWESS-Regression) (Cleveland und Devlin, 1988) zur Beseitigung nicht-linearer Effekte in Microarraydaten angewendet worden.

Die meisten Normalisierungsalgorithmen, wie auch LOWESS, können sowohl global, als auch lokal, d.h. auf einen ausgewählten Teil der Daten angewendet werden. Bei globalen Normalisierungsverfahren werden weder die Abhängigkeiten von der räumlichen Orientierung auf dem Array noch die Zugehörigkeit der Spots zu Spotnadelgruppen berücksichtigt. Durch die lokale Anwendung der Normalisierungsalgorithmen können dagegen systematische räumliche Variationen des Hybridisierungssignals innerhalb eines Arrays korrigiert werden, die z.B. durch inhomogene Spotoberflächen oder lokale Schwankungen der Hybridisierungsbedingungen verursacht werden können. Ein weiterer Einflußfaktor ist die Zugehörigkeit der Spots zu einer bestimmten Spotnadelgruppe bei der Herstellung der Arrays. So können einzelne Spotnadeln besonders starke Verschiebungen der Differenzwerte der Spots aufweisen (Schuchhardt *et al.*, 2000).

Durch die Darstellung der Verteilung der Differenzwerte der einzelnen Spotnadelgruppen des *T. brucei*-Arrays in Boxplots konnte gezeigt werden, daß die Verteilung der Werte zwischen den einzelnen Spotnadeln stark schwankt. Diese Abhängigkeit der Differenzwerte von der Zugehörigkeit zu einer Spotnadelgruppe erklärt sich dadurch, daß jede Spotnadel eine etwas andere Menge an Flüssigkeit abgibt. Dadurch werden von den verschiedenen Nadeln systematisch etwas unterschiedliche Mengen von DNA auf den Array gespottet. Die Menge der immobilisierten DNA hat einen Einfluß auf die Signalintensität des Spots. Die systematischen Unterschiede zwischen den einzelnen Spotnadelgruppen bei der Verteilung der Differenzwerte hat also ihre Ursache in der Abhängigkeit der Differenzwerte von der Signalintensität.

Durch die lokale Anwendung der LOWESS-Regression konnten räumliche-Effekte vermindert und Printnadel-Effekte entfernt werden, die bei der globalen Mittelwert-

Normalisierung nicht entfernt werden können. Eine gewisse Schwankung der mittleren Differenzwerte der einzelnen Spotnadelgruppen könnte jedoch auch durch Zufallsprozesse zu erklären sein. Ob die Schwankungen der Differenzwerte zufälliger Natur sind, oder ob die festgestellten Abweichungen zwischen der Varianz der Differenzwerte der Spotnadelgruppen signifikant sind, könnte durch eine Varianzanalyse überprüft werden. Die Anzahl der Spots in den einzelnen Spotnadelgruppen war jedoch mit 1.386 Spots so groß, daß die festgestellten Unterschiede eher auf systematische Verschiebungen schliessen lassen.

Die Berechnung des Normalisierungskoeffizienten für beide Methoden zeigte deutlich die größere Effizienz der LOWESS-Normalisierung gegenüber der globalen Mittelwert-Normalisierung. Eine Gefahr bei der lokalen Normalisierung der Signalintensitäten liegt in der Überanpassung. Das bedeutet, daß die Daten so stark transformiert werden, daß nicht nur Messartefakte ausgeglichen werden, sondern auch tatsächliche Expressionsunterschiede. Die Stärke der Datentransformation wird bei der LOWESS-Regression über den *Smoothing*-Parameter festgelegt. Da die Festlegung der Größe dieses Parameters empirisch erfolgt, ist die Gefahr einer zu starken Anpassung der Werte durchaus gegeben.

4.2.2. Datenfilterung

Trotz Normalisierung findet man auch bei den Differenzwerten aus einer Konkordanzanalyse extreme Werte. Ein Teil dieser Ausreißer sind durch Hybridisierungsartefakte verursacht, die durch visuelle Inspektion identifiziert werden können. Die Intensitätsquotienten von Spots mit niedriger Signalintensität weisen ebenfalls häufig eine besonders hohe Variabilität auf (Newton *et al.*, 2001). Durch Herausfiltern solcher Ausreißer kann die Qualität von Microarraydaten nach zuverlässigen Qualitätskriterien erheblich verbessert werden. In vielen Veröffentlichungen werden die Daten durch empirisch festgelegte Schwellenwerte für die Signalintensität herausgefiltert (Quackenbush, 2002). Da die Filterung der Daten mit einem mehr oder weniger großen Datenverlust einher geht, ist es wichtig, die Effizienz der eingesetzten Filtermethode zu überprüfen.

Es konnte gezeigt werden, daß zwischen dem absoluten Hybridisierungssignal der Spots (qA-Wert) und der Standardabweichung der Differenzwerte ein deutlicher Zusammenhang besteht. Als weiteres effizientes Filterkriterium wurde das Verhältnis zwischen Spot- und Hintergrundsignal (PP-Wert) ermittelt. Beide Qualitätskriterien zeigen eine deutliche Korrelation und sind in ihrer Effizienz der Datenfilterung etwa vergleichbar. Die starke Korrelation der beiden Werte ist dadurch zu erklären, daß der PP-Wert ein Mass für den Abstand des Spotsignals vom Hintergrundsignal darstellt und damit direkt vom Spotsignal

abhängt. Trotz der hohen Korrelation ist es sinnvoll, beide Kriterien zur Filterung anzuwenden, da so ein größerer Teil der Spots mit niedrigen Signalintensitäten beibehalten werden kann.

Bei der Datenfilterung in einen Experiment zur Identifikation differentiell exprimierter Gene hängt die Festlegung der jeweiligen Schwellenwerte davon ab, wieviele Werte man beibehalten möchte und welche Falsch-Positivrate bei den gefundenen differentiellen Genen tolerabel ist.

Durch die verwendete Filtermethodik konnte die Variabilität der Daten erheblich reduziert werden. Dazu muß man jedoch insgesamt auf etwa 10% der Daten verzichten. Dieser Prozentsatz liegt deutlich niedriger als in anderen Veröffentlichungen, bei denen etwa 25% der Daten herausgefiltert werden (Yang *et al.*, 2002a; Yue *et al.*, 2001). Welcher Prozentsatz der Daten herausgefiltert werden muß, hängt vor allem von der Qualität der Hybridisierungsdaten ab. Durch die Auswahl effizienter Filterungsverfahren kann die Variabilität der Daten bei geringem Datenverlust erheblich verringert werden.

Über die Auswahl und Effizienz verschiedener Filtermethoden gibt es bislang nur wenige Veröffentlichungen. Wang *et al*. haben zur Beurteilung der Spotqualität einen aus verschiedenen Einzelwerten ermittelten Qualitätskoeffizienten vorgeschlagen (Wang *et al.*, 2001; Wang *et al.*, 2003). Zur Ermittlung des Qualitätskoeffizienten sind die Größe des Spots, das Verhältnis zwischen Hybridisierungssignal und Hintergrundsignal, die Homogenität und der Absolutwert der Hintergrundsignale sowie der Sättigungswert des Hybridisierungssignals herangezogenen worden. Ein Test der oben genannten Qualitätsparameter mit den Daten aus der Konkordanzanalyse ergab jedoch, daß es zwischen der Standardabweichung der Differenzwerte von der Größe des Spots, dem Sättigungswert, der Uniformität und des Absolutwertes des Hintergrundsignals keinen Zusammenhang gibt. Ein Zusammenhang z.B. zwischen der Variabilität der Differenzwerte und der Größe der Spots ist zudem auch in der Grafik in der oben genannten Veröffentlichung nicht zu erkennen.

Im Gegensatz zur kompletten Filterung der Gene ermöglicht die Berechnung eines Qualitätskoeffizienten die Korrektur der Differenzwerte, so daß diese in der Analyse beibehalten werden können. In der praktischen Anwendung dürfte dieses Verfahren jedoch analog zur Filterung sein, da durch die willkürliche Festlegung der “Strafpunkte” bei Qualitätsmängeln fast alle Differenzwerte niedriger Qualität gegen Null tendieren. Ein Vorteil dieser Methode besteht darin, daß bei der Analyse der Daten nicht mit leeren Werten gearbeitet werden muß, was zum Beispiel bei *Clustering*-Verfahren problematisch ist (Troyanskaya *et al.*, 2001). Andererseits kann die Miteinbeziehung eines auf Null gesetzten

Wertes den Mittelwert der Replika-Differenzwerte ungerechtfertigterweise beeinflussen und so zu einer erhöhten Falsch-Negativrate führen.
Außer den hier angewendeten Verfahren zur Datenbearbeitung gibt es zahlreiche andere Verfahren, die zur Datentransformation angewendet werden können. Keine noch so ausgefeilte Datenbearbeitung oder -analyse kann jedoch schlechte technische Qualität kompensieren. Der wichtigste Schritt bei der Microarray-Analyse bleibt darum die Generierung qualitativ hochwertiger Daten sowie eine Versuchsplanung mit einer ausreichenden Anzahl an wiederholten Hybridisierungen. Bei jedem Schritt der Microarray-Analyse, von der RNA-Isolierung und der Herstellung der Arrays bis zur Datenbearbeitung sollte deshalb versucht werden, die Variabilität der Messung zu verringern.

4.3. Validierung der Microarray-Technologie

4.3.1. Allgemeine Einführung

Die zentralen Parameter, die über die Qualität eines Microarray-Experiments entscheiden, sind die Akkuratheit und die Präzision der Differenzwerte, die aus den Meßwerten der Signalintensität des Spots für jede Wellenlänge errechnet werden. Unter der Präzision einer Messung versteht man den Grad der Übereinstimmung wiederholter Messungen der gleichen Probe, während die Akkuratheit die Abweichung zwischen gemessenem und tatsächlichen Wert oder einem Referenzwert bezeichnet. Je geringer die Unterschiede der Transkripthäufigkeit sind, die man detektieren möchte, desto größer müssen die Akkuratheit und die Präzision der Differenzwerte sein.

4.3.2. Präzision

Die Präzision der Differenzwerte wurde durch wiederholte Messungen des gleichen Spots auf verschiedenen Arrays bestimmt. Faktoren, die zur Abweichung der wiederholten Messwerte führen, sind die Spot-Variation, Unterschiede bei der Scannereinstellung, Unterschiede in der Beschaffenheit der Arrayoberfläche und Unterschiede in der Hybridisierungseffizienz sowie Hintergrundschwankungen und Hybridisierungsartefakte.
Die Präzision einer Methode wird durch die Streuung der Messwerte bestimmt. Als Masszahl für die Streuung der Messdaten wurde der durchschnittliche Variationskoeffizient der normalisierten, gefilterten Differenzwerte aus allen Hybridisierungen pro Spot betrachtet. Der durchschnittliche Variationskoeffizient der Differenzwerte betrug 11,75%. Übereinstimmend mit diesem Wert wurde in einer Microarray-Validierungsstudie der Firma *Incyte Genomics*

ein Variationskoeffizient der wiederholten Differenzwerte von 12% festgestellt (Yue *et al.*, 2001).

Ist die gleiche Sequenz auf verschiedenen Spots auf einem Objekträger repräsentiert, kann man auch das durchschnittliche Verhältnis dieser Replika-Spots heranziehen, um die Präzision der Messung zu beurteilen (Quackenbush, 2002). In einer Studie von Bartosiewicz *et al.* wurde so die Spot-Variabilität innerhalb eines Arrays in Abhängigkeit von der Sequenz bestimmt. Dabei wurde ein Variationskoeffizient der Differenzwerte von 8-18% festgestellt (Bartosiewicz *et al.*, 2000). Bei der Herstellung des *T. brucei*-Arrays wurde jedoch auf Replika-Spots verzichtet, um eine möglichst große Abdeckung des Genoms zu erreichen.

Die in Abbildung 3.11 dargestellte Verteilung der Differenzwerte der einzelnen Konkordanzhybridisierungen machen deutlich, daß für die Gesamtvariabiltät der Differenzwerte vorallem der Anteil der mittleren und extremen Ausreißer entscheidend ist. Die Interquartilsbereiche zeigen in allen Hybridisierungen nur geringfügige Abweichungen. Die Ausreißer in einer Hybridisierung werden vor allem durch starke Hintergrundsignale verursacht, die durch Hybridisierungsartefakte wie Fluoreszenzflecken entstehen. Die Qualität der Hybridisierungsdaten könnte durch ein Verfahren, das die Entstehung solcher Hybridisierungsartefakte verhindert, erheblich verbessert werden.

4.3.3. Akkuratheit der Signalintensitätswerte (Konkordanzanalyse)

Unter der Akkuratheit einer Messung versteht man die Übereinstimmung der Messwerte einer Probe mit dem tatsächlichen Wert oder einem Referenzwert. Bei einer Konkordanzanalyse kann man die Signalintensitätswerte (R_i; G_i) als Mess- und Referenzwertepaar betrachten, da theoretisch beide Werte gleich sein sollten. Auch in anderen Veröffentlichungen wurde die Abweichung der Signalintensitätswerte aus einer Konkordanzhybridisierung als Mass für die Akkuratheit der Messung herangezogen (Yue *et al.*, 2001). Zur Bestimmung der Akkuratheit der Microarray-Analyse wurde die Korrelation der Intensitätswerte und die Streuung der aus den Intensitätswerten resultierenden Differenzwerte eines Arrays berechnet. Dazu wurden der durchschnittliche Pearsonsche Korrelationskoeffizient und der durchschnittliche Variationskoeffizient der Differenzwerte aus allen Hybridisierungen berechnet.

Die Streuung der Differenzwerte innerhalb eines Arrays entspricht mit einem Variationskoeffizienten von durchschnittlich 11,34% ungefähr dem durchschnittlichen Variationskoeffizient der wiederholten Messung der Differenzwerte auf verschiedenen Arrays, der 11,75% betrug.

4.4. Datenanalyse

4.4.1. Prüfung auf Normalverteilung

Viele statistische Tests, die zur Identifizierung differentiell exprimierter Gene angewendet werden, setzen die Normalverteilung der Differenzwerte voraus. Aus diesem Grund sind die Daten aus der Konkordanzanalyse sowohl auf der Ebene der Differenzwerte als auch der Signalintensitätswerte durch ein graphisches Verfahrens auf das Vorliegen einer Normalverteilung überprüft worden. Die Überprüfung der Verteilung der Differenzwerte auf Normalverteilung mithilfe eines Normalverteilungsplots hat ergeben, daß die Messdaten aus der Konkordanzanalyse deutlich von einer Normalverteilung abweichen. In einer Veröffentlichung von Brody *et al.* ist dieses Phänomen ebenfalls beschrieben worden (Brody *et al.*, 2002). Als Ursache für die von der Normalverteilung abweichende Verteilung der Differenzwerte wurde das zur Berechnung der Differenzwerte angewendete Verfahren vermutet. Bei den meisten Microarray-Analysen werden die Differenzwerte durch die Bildung des Quotienten aus den medianen Signalintensitäten aller Pixel des Spots berechnet. Die Verteilung des Quotienten [x/y] zweier normalverteilter Zufallsvariablen x und y folgt einer Lorentzverteilung, wenn die Standardabweichung der Messwerten relativ hoch ist. Die Lorentzverteilung zeichnet sich durch einen hohen Anteil an Extremwerten aus, was für die Microarray-Analyse unvorteilhaft ist. Als alternatives Verfahren der Differenzwertberechnung wird von Brody *et al.* die Berechnung des Medians der Quotienten aller einzelnen roten und grünen Pixel-Intensitäten eines Spots vorgeschlagen. Die grafische Überprüfung der Verteilung der anhand dieses Verfahrens berechneten Differenzwerte ergab, daß die so ermittelten Differenzwerte eine zwar weniger starke, aber dennoch eindeutige Abweichung von einer Normalverteilung zeigen (Abb. 4.1).

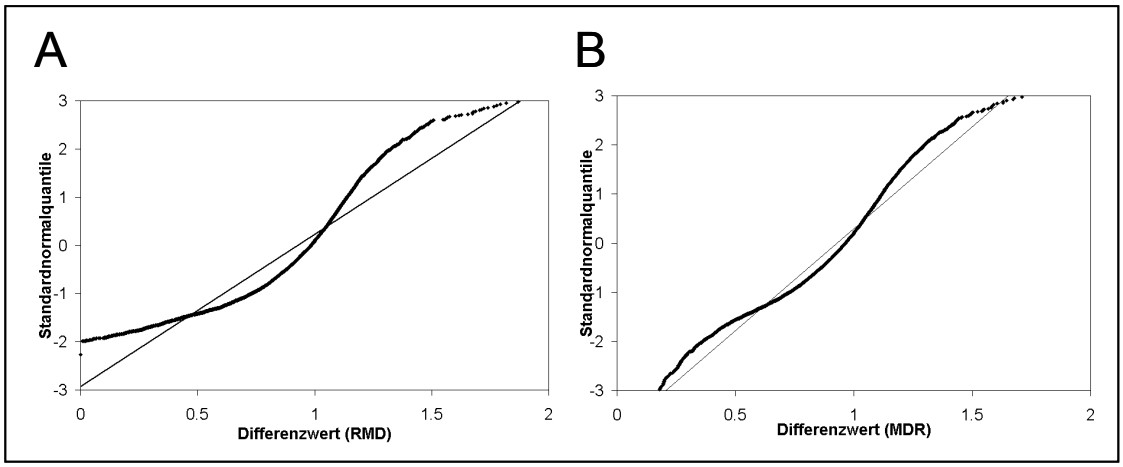

Abb. 4.1 Normal-Quantilplot der Differenzwerte aus der Konkordanzanalyse. (A) Berechnung der Differenzwerte aus dem Quotienten der Medianwerte aller Spotpixel. (B) Berechnung der Differenzwerte aus dem Medianwert der Intensitätsquotienten der einzelnen Spotpixel.

4.4.2. Signifikanzanalyse für den Konkordanzdatensatz

Da die Verteilung der Daten der Forderung nach Normalverteilung für parametrische Prüfverfahren nicht genügten, wurde die *Signifikanz-Analyse für Microarrays* (SAM) als statistisches Verfahren zur Detektion differenziell exprimierter Gene ausgewählt. Dieses Verfahren beruht auf einem nicht-parametrischen Test, und setzt deshalb nicht die Normalverteilung der Differenzwerte voraus (Tusher *et al.*, 2001).

Bislang sind drei verschiedene statistische Verfahren zur Detektion differentiell exprimierter Gene beschrieben worden, die auf nicht-parametrischen Tests beruhen. Neben der verwendeten Signifikanz-Analyse sind *empirical Bayes-Statistik* (Efron und Tibshirani, 2002) und das *Mixed-Model*-Verfahren (Pan, 2002) als nicht-parametrische Prüfverfahren vorgeschlagen worden. Allen Verfahren ist gemeinsam, daß die Nullverteilung (Verteilung ohne Signifikanzen) durch Datenpermutation geschätzt wird.

Ein Vergleich der Effizienz und Prüfstärke verschiedener Verfahren zur Identifikation differentiell exprimierter Gene ist bislang nur für den *Mixed-Model*-Ansatz mit verschiedenen parametrischen Verfahren wie dem *t*-Test und dem *Regression Modeling*-Verfahren (Thomas *et al.*, 2001) vorgenommmen worden (Pan, 2002).

Die Bestimmung des Detektionslimits für differenzielle Expression für SAM lag bei einem *delta*-Wert von 0,26. Bei einer vergleichenden Hybridisierung kommt neben der technischen Variabilität allerdings noch die biologische Variabilität hinzu, die ein beträchtlich höheres Ausmaß als die technische Variabilität haben kann (Callow *et al.*, 2000; Churchill, 2002), sowie weitere Variabilität durch die unterschiedliche Handhabung zweier verschiedener

Proben bei der RNA-Extraktion. Aus diesem Grund kann das bestimmte Detektionslimit nur als Anhaltspunkt dienen. Tatsächlich sollte ein höherer *delta*-Wert zur Detektion verwendet werden, um zusätzliche Variabilität durch die Verwendung verschiedener Proben und biologische Variabilität zu berücksichtigen.

Durch das Konkordanzexperiment läßt sich die Falsch-Positivrate bei gegebenem *delta*-Wert gut abschätzen, wogegen die Ermittlung der Falsch-Negativrate nicht möglich ist. Die Falsch-Negativrate wäre z.B. durch ein Konkordanzexperiment möglich, dem Kontroll-RNAs in unterschiedlichen Konzentrationen zugesetzt wird.

Ein Vorzug der verwendeten Methode ist die Möglichkeit des Experimentators, die tolerierbare Falsch-Positivrate im vorhinein festzulegen. Dadurch kann das Verfahren an verschiedene Situationen angepaßt werden. Sucht man beispielsweise nach einem einzigen Gen mit relativ niedriger differenzieller Expression, so ist es durchaus vertretbar 10-20 Kandidaten zu sequenzieren und einer genaueren Analyse zu unterziehen, wogegen bei einer größeren Zahl differenzieller Gene eine wesentlich geringere Falsch-Positivrate notwendig ist.

4.5. Identifizierung und Validierung von stadienspezifischen Genen in der Blutbahnform und der prozyklischen Form von *T. brucei* mittels Microarray-Analyse

4.5.1. Experimenteller Aufbau

Für die Detektion stadienspezifisch exprimierter Gene in der Blutbahnform und der prozyklischen Form wurden 4 Hybridisierungen mit Proben vorgenommen, die aus getrennt präparierter RNA und separaten Zellkulturen hergestellt wurden. Eine der Hybridisierungen wurde mit invers markierten Proben durchgeführt. Durch diesen Versuchsaufbau sollte verhindert werden, daß in der Microarray-Analyse Klone fälschlicherweise als stadienspezifisch exprimiert identifiziert werden, die aufgrund variierender Präparation oder aufgrund von Markierungsartefakten einen stark abweichenden Differenzwert aufweisen.

4.5.2. Identifikation differentiell exprimierter Gene

Die Verteilung der erhaltenen Daten aus den Hybridisierungen mit den Blutbahnformproben und den prozyklischen Proben zeigte einen deutlichen Unterschied zur Verteilung der Konkordanzdaten. Neben dem Vorhandensein stadienspezifisch-exprimierter Transkripte kann ein Teil der höheren Variabilität der Daten auch durch biologische Variabilität zwischen verschiedenen Zellkulturansätzen und geringfügige Abweichungen bei der experimentellen Handhabung bedingt sein.

Auffällig war der deutlich höhere Anteil positiv-signifikanter Klone, d.h., der Blutbahnform-spezifischen Klone. Durch eine inverse Markierung konnte gezeigt werden, daß es sich hierbei nicht um einen Farbstoffartefakt handelte. Dieses Ungleichgewicht ist vermutlich auf die große Anzahl von VSG-Genen im Genom von *T. brucei* zurückzuführen, die starke homologe Bereiche besitzen, so daß die Kreuzhybridisierung von VSG-mRNA an verschiedenen Gene wahrscheinlich ist.

Durch die Sequenzierung eines Teils der differentiellen Klone konnten zahlreiche bislang nicht beschriebene Gene identifiziert werden. Darunter waren Gene mit Homologien zu putativen Proteinen anderer Mitglieder der Kinetoplastiden, die durch Genomprojekte identifiziert wurden, wie z.B. eine putative CAAX-Prenyl-Protease aus *T. cruzi,* ein mutmasslicher Aminosäureransporter (*L. major*) und zwei hypothetische Membranproteine, von denen eines spezifisch für die infektiöse Insektenform von *L. donovani* ist (META1). Unter den sequenzierten Klonen befanden sich ebenfalls zahlreiche der bereits bekannten Gene mit differentieller Expression.

Ein Großteil der identifizierten Blustromform-spezifischen Gene zeigte starke Homologien zu VSG-Genen, oder anderen Genen aus den VSG-Expressionsstellen. Es ist durchaus möglich, daß einige dieser Sequenzen nicht unbedingt in den Blutbahnformen exprimiert werden, sondern das diese Sequenzen mit VSG mRNAs kreuzhybridisieren, da sie in einigen Bereichen mit den VSG-Sequenzen identisch sind. Unter den sequenzierten Klonen fanden sich einige bekannte Gene mit differentieller Expression in der Blutbahnform oder der prozyklischen Form, wie z.B. Cytochrom-Oxidase Untereinheit IV (Matthews und Gull, 1998), *Major Surface Protease* (GP63) (LaCount *et al.*, 2003), RNA-Helikase, Kinetoplastid Membran Protein 11, ESAG11, Calpain und Pyruvat-Kinase (El-Sayed *et al.*, 2000).

Einige der regulierten Gene in *T. brucei* werden durch eine Microarray-Analyse nicht zu finden sein. Darunter sind z.B. solche Gene, die sich lediglich durch ihre 3'-untranslatierten Regionen unterscheiden, wie es beispielsweise bei drei verschiedenen stadienspezifisch exprimierten Phosphoglycerat-Kinase mRNAs der Fall ist (Blattner und Clayton, 1995). Ein anderes Beispiel sind differentiell exprimierte Hexose-Transporter, bei denen die Regulation der Transkriptmenge über einen kurzen nicht homologen Abschnitt innerhalb der 3'-untranslatierten Regionen erfolgt (Hotz *et al.*, 1995).

Ein interessanter Aspekt der Ergebnisse aus der Sequenzierung der differentiellen Klone sind die Klone, die weder ORFs aufweisen, noch eine andere bereits bekannte nicht-codierende RNA-Sequenz. Einige dieser Klone weisen repetitive Segmente auf. Diese Klone könnten Matritzen für nicht-codierende Transkripte sein (Eddy, 2001). Es besteht auch die

Möglichkeit, daß diese Fragmente mit anderen, kodierenden RNAs kreuzhybridisieren. Diese könnten lange 3'-untranslatierte Regionen oder intergenische Regionen sein, die als Teil von polycistronischen Transkripten transkribiert werden und prozessiert werden, weil sie trans-spleiss-Sequenzen tragen. Bislang gibt es keinen Hinweis darauf, das das sogenannte *nonsense-mediated decay*-System in Trypanosomen aktiv ist, das den Abbau von mRNA mit *nonsense*-Codons bewerkstelligt (Liniger *et al.*, 2001; Pays *et al.*, 1989; Vassella *et al.*, 1994). Es sind jedoch prozessierte, aber scheinbar nicht kodierende Transkripte beschrieben worden (Scholler *et al.*, 1988). Durch die Sequenzierung von etwa 20% der gefundenen Klone mit differentiellen Sequenzen sind 20 entweder neue Gene oder Sequenzen, deren stadienspezifische Transkriptregulation bislang nicht bekannt war. Mit dem Sequenzieren der verbleibenden Klone liessen sich vermutlich noch etwa die fünffache Anzahl an neuen differentiellen Sequenzen finden. Diese recht übersichtliche Anzahl ermöglicht eine detailiertere funktionelle Analyse der gefundenen Gene.

Die Regulation der Transkriptmenge erfolgt bei Trypanosomen zum großen Teil durch Sequenzen in der 3'-untranslatierten Region. Deshalb könnten durch die vorliegenden Ergebnisse spezifische Sequenzen innerhalb der 3'-untranslatierten Regionen gefunden werden, durch die spezifische Differenzierungsmuster festgelegt werden.

In einer anderen Genexpressionsanalyse zur Identifikation differentiell exprimierter Gene in der Blutbahnform und der prozyklischen Form ist ein Microarray mit 400 cDNA und GST-Klonen hergetellt worden (El-Sayed *et al.*, 2000). 57 dieser Klone wurden als differenziell eingeordnet. Dieser, mit 14% relativ hohe Prozentsatz ist wahrscheinlich damit zu erklären, daß bei der Auswahl der Klone des verwendeten Arrays ein großer Anteil von stark exprimierten und oft regulierten Sequenzen verwendet worden ist.

Vergleicht man die sich aus der Microarray-Analyse ergebenden Differenzwerte mit den in der Literatur angegebenen Werten die durch *Northern*-Analyse ermittelt werden, so liegen diese Werte im Arrayexperiment deutlich niedriger. So weisen zum Beispiel die EP/GPEET *Loci* in der Microarray-Analyse nur einen Differenzwert von 2-3 auf, wogegen der Unterschied bei der *Northern*-Analyse etwa um ein hundertfaches größer ist. So ist auch die in der Blutbahnform stark exprimierte alternative Oxidase in prozyklischen Zellen kaum nachzuweisen, wogegen in der Microarray-Analyse nur ein zweifacher Transkriptionsunterschied zwischen Blutbahn- und prozyklischem Stadium nachgewiesen werden konnte. Diese niedrigeren Differenzwerte in der Microarray-Analyse sind ein bekanntes Phänomen (Taniguchi *et al.*, 2001). Für den deutlich geringeren dynamischen Bereich bei der Messung von Transkriptionsunterschieden mit Microarrays sind

wahrscheinlich mehrere Einflußfaktoren verantwortlich. Die Verwendung unaufgereinigter DNA für die Herstellung der Microarrays kann sicher eine Ursache sein, da die Messwerte für differentielle Expression etwas geringer sind als bei Verwendung aufgereingter Proben (Diehl *et al.*, 2002b). Allerdings ist bekannt, daß die durch Microarrays ermittelten Expressionsunterschiede oft wesentlich geringer sind, als die durch andere Verfahren wie RT-PCR und *Northern*-Analyse ermittelten Werte. Die unterschiedlichen Ergebnisse sind wahrscheinlich darauf zurückzuführen, daß sich die Kinetik von Microarray- und *Northern*-Hybridisierungen stark unterscheiden (Yuen *et al.*, 2002). Am zuverlässigsten sind wahrscheinlich die mittels quantitativer Echtzeit-RT-PCR ermittelten Expressionswerte (Chuaqui *et al.*, 2002).

4.5.3. Schlußfolgerung und Ausblick

Die Ergebnisse der Analyse der Genexpression in der Blutbahnform und der prozyklischen Form haben gezeigt, daß Microarrays aus genomischen Fragmenten erfolgreich zur Identifikation differentiell exprimierter Transkripte eingesetzt werden können. Durch eine Genexpressionsanalyse mit genomischen Microarrays konnten zahlreiche neue stadienspezifisch exprimierte Gene der Blutbahnform und der prozyklischen Form von *T. brucei* identifiziert werden, die zur molekularbiologischen Charakterisierung dieser experimentell sehr wichtigen Formen beitragen.

Für die Entwicklung von Vakzinen sind die für den Menschen infektiöse metazyklische Form besonders interessant. Diese ist allerdings experimentell nur schwer zugänglich, da sie aus den Speicheldrüsen infizierter Fliegen präpariert werden muß. Die geringe Menge verfügbarer RNA stellt besondere methodische Anforderungen. Die PCR-Amplifikation der Gesamt-RNA über die leader-Sequenz und den polyA-Schwanz bietet eine einfache Möglichkeit, die nötige Menge an Probe für ein Array-Experiment herzustellen. Der Nachteil einer PCR-Amplifikation ist, daß es leicht zu artifiziellen Verschiebungen in der Zusammensetzung der Transkriptmenge kommt. Zudem wären diese Transkriptionsprofile wegen der unterschiedlichen Kinetik der Hybridisierung von doppelsträngigen Proben, bei denen der Gegenstrang als effektiver Kompetitor in der Hybridisierungslösung vorliegt, nur bedingt mit den Profilen aus cDNA-Hybridisierungen vergleichbar. Eine andere Möglichkeit zur Amplifikation des Probenmaterials ist die Amplifikation von RNA über die Ligation eines T7-Promoters. Diese Methode ist bereits erfolgreich für die Microarray-Analyse angewendet worden (Luo *et al.*, 1999b), und stellt eine erfolgversprechende Ansatz für die Analyse derjenigen Lebensformen dar, die nicht *in vitro* kultiviert werden können.

Eine Möglichkeit zur Microarray-Analyse geringer RNA-Mengen ohne vorherige Amplifikation ist die radiaktive Markierung, mit der Expressionsprofile aus weniger als 1 µg gesamt-RNA erstellt werden können (Salin *et al.*, 2002).

Neben der Transkriptionsanalyse mit Microarrays ermöglicht die Proteomanalyse aufschlußreiche Einblicke in die differentielle Genexpression verschiedener Lebensstadien. Ein Vergleich zwischen Blutbahnform und prozyklischer Form durch Protein-2D-Gelelektrophorese hat ebenfalls nur wenige Unterschiede zwischen der Blutbahnform und der prozyklischen Form festgestellt (van Deursen *et al.*, 2001). Dies legt die Vermutung nahe, daß die physiologischen und morphologische Unterschiede der Lebensformen mit nur wenigen Genen erreicht werden können. Für ein genaueres Verständnis der Differenzierung der verschiedenen Lebensstadien von *T. brucei* sind jedoch weitere, detailliertere Untersuchungen der intermediären Formen und gezielte genetische Manipulationen nötig, zum Beispiel durch RNAi oder Knockout-Zelllinien.

Für ein tiefergehendes Verständnis der Anpassungen der verschiedenen Lebensstadien von *T. brucei* müsste die Genexpression aller Stadien verglichen werden. So ist beispielsweise für den Malaria-Parasiten *Plasmodium falciparum* eine Microarray-Analyse von neun Stadien durchgeführt worden (Le Roch *et al.*, 2003). Durch diese Studie konnten 43% aller Gene anhand ihres Expressionsmusters den verschiedenen Stadien zugeordnet werden.