3. Ergebnisse

3.1. Repräsentative Differenzanalyse

3.1.1. Subtraktive Hybridisierung

Zur Identifikation von Genen, die für die Blutbahnform und die prozyklische Form von *Trypanosoma brucei* spezifisch sind, wurde eine repräsentative Differenzanalyse (RDA) durchgeführt. Bei der RDA wird eine subtraktive Hybridisierung zwischen einer *Driver*- und *Tester*-cDNA Population vorgenommen. Nach der subtraktiven Hybridisierung erfolgt eine spezifische Amplifikation derjenigen *Tester*-cDNAs, die häufiger vorkommen als die homologen *Driver*-cDNAs. Auf diese Weise erfolgt eine sukzessive Anreicherung differentiell exprimierter *Tester*-Fragmente.

Für die Durchführung der RDA wurde aus der mRNA beider Lebensformen jeweils ein *Driver* und ein *Tester* hergestellt. Zur Anreicherung spezifischer Gene der Blutbahnform wurde für die RDA der *Tester* dieser Form mit dem *Driver* der prozyklischen Form eingesetzt. Zur Anreicherung prozyklisch-spezifischer Gene wurden *Tester*- und *Driver*-Populationen vertauscht. Für die cDNA-Synthese wurden jeweils 2 µg mRNA eingesetzt. Die Ausbeute betrug 1,6 µg cDNA für die Blutbahnform-cDNA-Synthese und 1,8 µg cDNA für die cDNA-Synthese von mRNA aus prozyklischen Zellen. Aus dieser cDNA wurden durch Restriktionsverdau und Ligation von PCR-Adaptoren *Tester* und *Driver* hergestellt. Anschließend wurden je 2 Subtraktions- und Amplifikationszyklen mit beiden *Driver/Tester* Kombinationen durchgeführt.

Zur Analyse der einzelnen Schritte wurden je 100 ng der Repräsentation des ersten Differenzproduktes und des zweiten Differenzproduktes in einem Agarosegel aufgetrennt (Abb. 3.1). Die Anreicherung differenzieller RDA-Fragmente im Verlauf der RDA läßt sich auf dem Gel durch das Auftreten diskreter Banden bei den Differenzprodukten erkennen. Es wird zwar im Verlauf der RDA ein Bandenmuster deutlich, allerdings bleibt ein Hintergrund erhalten, der darauf hindeutet, daß auch im zweiten Differenzprodukt noch ein gewisser Grad an Komplexität vorhanden ist. Die Verschiebung des Bandenmusters zwischen dem ersten und zweiten Differenzprodukt liegt daran, daß das zweite Differenzprodukt im Gegensatz zum ersten Differenzprodukt noch die Adaptoren enthält. Dadurch sind die Fragmente des zweiten Differenzprodukts 48 bp länger.

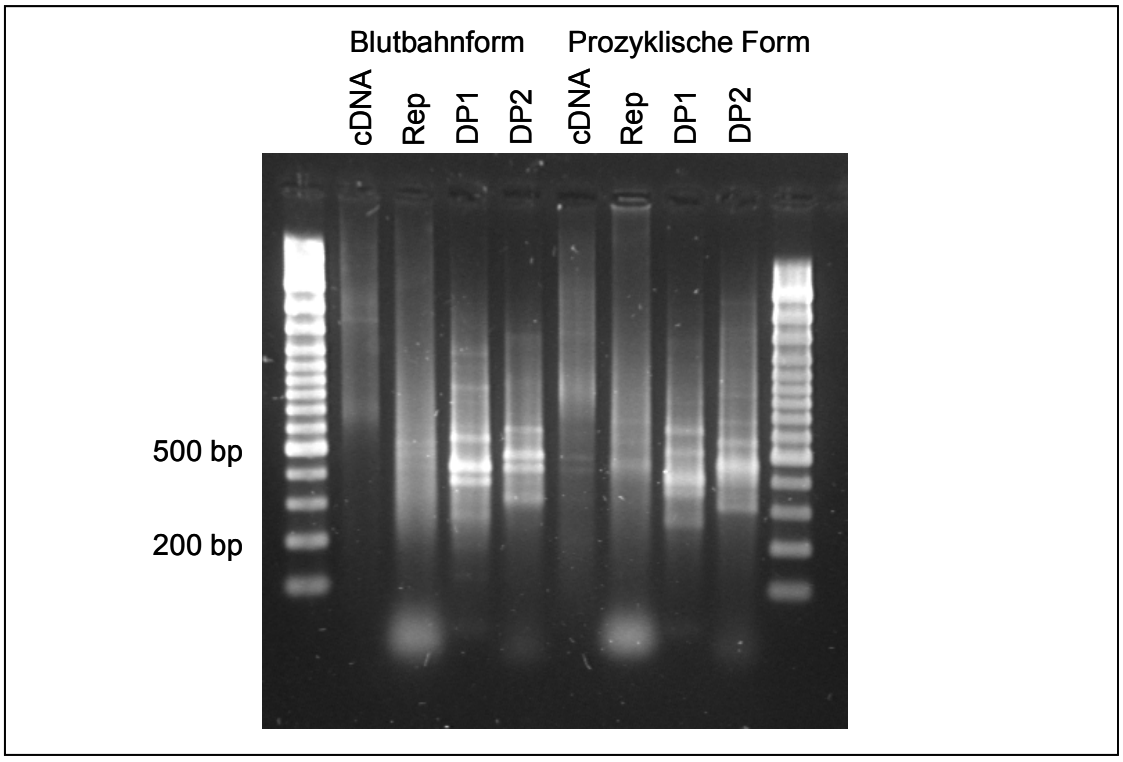


Abb. 3.1 Gelelektrophorese der RDA-Produkte. Jeweils 100 ng der cDNA, der Repräsentation, des Differenzproduktes 1 und des Differenzproduktes 2 wurden in einem 2,5% igem Agarosegel aufgetrennt.

3.1.2. Isolierung von RDA-Differenzprodukten durch Herstellung geordneter RDA-Bibliotheken

Nach zwei Runden von Subtraktion und Amplifikation wurden die Differenzprodukte in den TA-Klonierungsvektor (pCR2.1, Invitrogen) kloniert. In diesen Vektor können PCR-Produkte direkt ligiert werden. Die Ligationsansätze wurden zur Transformation kompetenter *E. coli*-Bakterien verwendet. Nach dem Ausstreichen der Bakteriensuspensionen auf Nährmediumplatten wurden 384 Klone jedes Differenzprodukts zur Herstellung geordneter Klonbibliotheken in eine Mikrotiterplatte mit 2YT-Nährmedium übertragen. Diese Bibliotheken wurden zur Verifizierung der differentiellen Genexpression der RDA-Produkte und für die Sequenzanalyse verwendet.

3.1.3. Verifizierung der differentiellen Expression durch Hybridisierung von cDNA auf RDA-Macroarrays

Um die differenzielle Expression der erhaltenen RDA-Produkte zu verifizieren, wurden mit den oben beschriebenen RDA-Bibliotheken Macroarrays hergestellt, auf die dann radioaktiv-markierte cDNA aus der Blutbahn- bzw. der prozyklischen Form hybridisiert wurde.

Zur Herstellung der Arrays wurden die Inserts aus den Klonen der RDA-Bibliotheken mit adaptorspezifischen Primern (N-Bgl-24) amplifiziert. Von 384 Klonen der Blutbahn-spezifischen Bibliothek ließen sich 309 Klone amplifizieren (80%). Von der prozyklisch-spezifischen Bibliothek ließen sich 226 Fragmente (59%) amplifizieren. Die PCR-Produkte beider RDA-Bibliotheken wurden mit einem Roboter (BioGrid, BioRobotics) direkt aus der PCR-Platte auf Nylonmembranen übertragen. Die PCR-Produkte beider Bibliotheken wurden jeweils zweimal in einem 2×2-Muster gespottet, in dem die Replika-Spots diagonal angeordnet waren. Diese Anordnung hatte den Vorteil, daß die resultierenden Hybridisierungssignale durch „Musterbildung“ einfach der jeweiligen Bibliothek zugeordnet werden konnten.

Für die Hybridisierung wurden jeweils 1 µg polyA$^+$RNA durch den direkten Einbau von P^{33}-dCTP während der cDNA-Synthese radioaktiv markiert und über Nacht auf die RDA-Arrays hybridisert. Nach der Hybridisierung wurden die Membranen mehrfach gewaschen. Die Membranen wurden anschließend mehrere Stunden auf Phosphoimaging-Platten aufgelegt. Die Platten wurden dann zur Erzeugung eines digitalen Bildes mit einem Laserscanner ausgelesen. Abbildung 3.2 zeigt die Hybridisierungssignale der RDA_Membranen nach der Hybridisierung mit radioaktiv-markierter prozyklischer- und Blutbahnform-cDNA.

Insgesamt 264 Klone (85%) der Blutbahnform-spezifischen Bibliothek zeigten nach der Hybridisierung mit cDNA der Blutbahnform ein starkes Hybridisierungssignal. Dagegen zeigten nur 26 Klone der prozyklisch-spezifischen Bibliothek ein Signal in der Hybridisierung mit prozyklischer cDNA.

Die meisten Klone in einer RDA-Bibliothek sind sehr redundant. Deshalb wurden durch Klonpoolhybridisierungen redundante Klone identifiziert, um den Sequenzieraufwand zu minimieren. Dazu wurden markierte PCR-Produkte einzelner Klone durch den Einbau Cy5-markierter Primer während der PCR hergestellt. Gemische von jeweils 8 PCR-Produkten wurden dann auf die Nylonmembranen hybridisiert. Insgesamt wurden 4 Pools auf die RDA-Arrays hybridisiert. Anhand der Hybridisierungsmuster wurden zehn nicht-redundante Klone aus der Blutbahnform identifiziert und anschließend sequenziert. Für die prozyklisch-spezifische Bibliothek wurden jeweils vier der verifizierten Klone ausgewählt und gemeinsam auf die Arrays hybridisiert. Nach zwei dieser Klonpoolhybrisierungen wurden fünf Klone für die Sequenzierung ausgewählt.

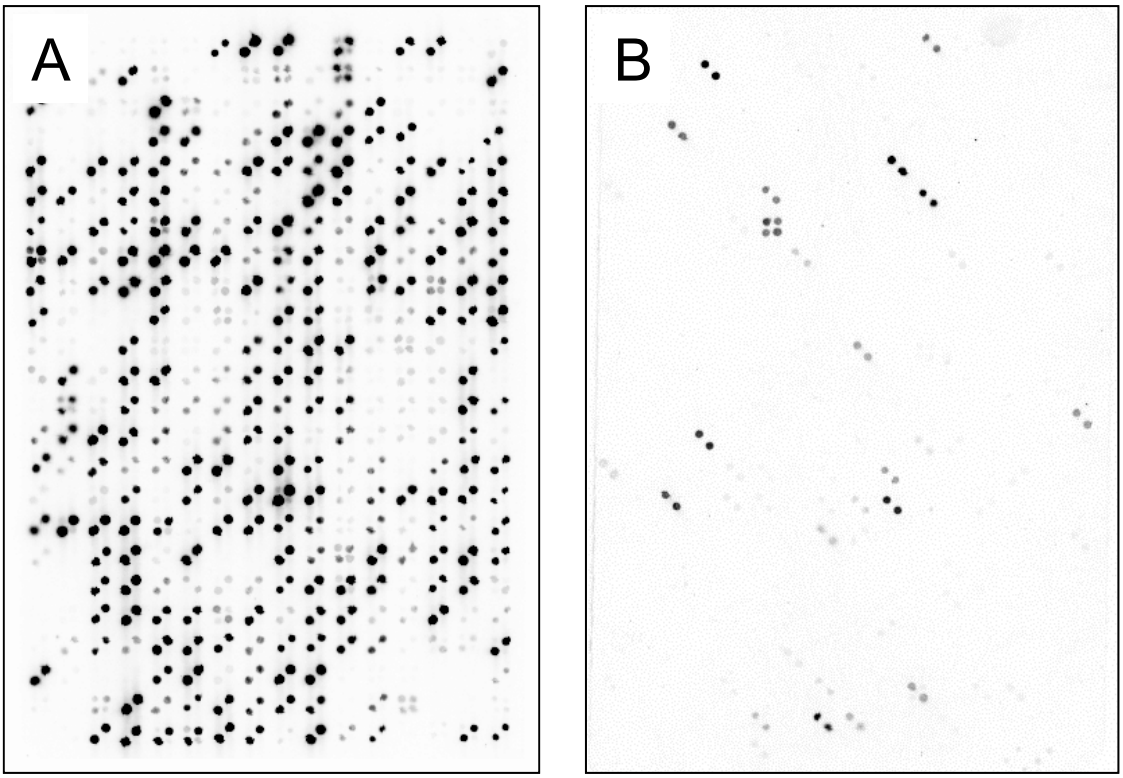


Abb. 3.2 Hybridisierung der RDA-Bibliotheken mit radioaktiv markierter cDNA zur Verifizierung der differentiellen Genexpression. Die RDA-Arrays enthalten die Klone aus beiden RDA-Bibliotheken. Die Arrays wurden mit radioaktiv markierter cDNA aus der Blutbahnform (A) und aus der prozyklischen Form (B) hybridisiert.

3.1.4. Identifizierung der RDA-Produkte durch Sequenzierung

Die ausgewählten RDA-Klone wurden von H. Delius, DKFZ, Heidelberg, sequenziert. Die erhaltenen Sequenzen wurden mit dem BLAST-Algorithmus (Altschul *et al.*, 1990) (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST) auf Homologien zu bekannten Sequenzen in öffentlichen Sequenzdatenbanken untersucht. Sechs Klone hatten starke Homologien zu VSG-Genen, wie z.B. VSG ILtat 1.3 und 1.2, VSG GUTat 10.2 und VSG WRATat A. Ein Klon zeigte eine starke Homologie zu ESAG 11, einem Gen aus den VSG-*expression sites* (Redpath *et al.*, 2000). Ein Klon der prozyklischen RDA-Bibliothek ist ein Fragment aus dem EP/GPEET-Lokus (Prozyklin). Für die Klone, für die keine Homolgie in den öffentlichen Datenbanken gefunden werden konnten, wurde in der TIGR-Datenbank nach identischen genomischen Sequenzen gesucht, um gegebenenfalls die komplette Sequenzinformation für ein Gen zu erhalten. Fast alle Sequenzen der RDA-Klone waren in der TIGR-Datenbank vertreten. Die verfügaren TIGR-Sequenzen wurden dann erneut zur BLAST-Suche und zur ORF-Suche eingesetzt wurden. Allerdings konnte auch durch diese Methode keinem dieser

RDA-Klone eine Identität zugeordnet werden. Es wurden auch keine benachbarten ORFs gefunden.

3.2. *Trypanosoma brucei* Microarray-Analyse

3.2.1. Aufbau des Microarrays

3.2.1.1. Allgemeine Einführung

Die *Trypanosoma brucei* Microarrays wurden durch das Immobilisieren von PCR-Produkten einer geordneten genomischen Klonbibliothek auf Glasobjektträger hergestellt. Zur Generierung der geordneten genomischen Bibliothek wurde genomische DNA vom Stamm TREU 927/4 verwendet. Dieser Stamm wird auch im Rahmen des *T. brucei* Genomprojektes sequenziert.

3.2.1.2. Herstellung einer geordneten genomischen Klonbibliothek von T. brucei *TREU 927/4*

Der Ligationsansatz für die Herstellung der Bibliotheken wurde von TIGR (Rockville, MD) zur Verfügung gestellt. Er enthielt 2-2,5 kb große Fragmente, die in den Sequenzierungsvektor pUC18 kloniert wurden. Dieser Ligationsansatz wurde zur Transformation von *E. coli* DH 10b verwendet. Positive Kolonien wurden mithilfe eines Roboters (*Max-Planck-Institut für molekulare Genetik, Berlin*) durch Blau/Weiß-Selektion identifiziert und in 384-well-Mikrotiterplatten überführt. Insgesamt wurden 24.960 Klone transferiert (Abb. 3.3).

3.2.1.3. Herstellung der Microarrays

PCR-Amplifikation

Zur Herstellung des genomischen Microarrays wurde die Bibliothek im 384-well Maßstab amplifiziert. Als Ausgangsmaterial für die Amplifikation wurden flüssige Bakterienkulturen eingesetzt. Die Verwendung von Bakterienkulturen als Ausgangsmaterial für die Amplifikation bietet den Vorteil, daß auf die Isolierung von Plasmiden verzichtet werden kann. Neben der Zeit- und Kostenersparnis gibt es bei der direkten Amplifikation aus Bakterienkulturen weniger Kreuzkontaminationen.

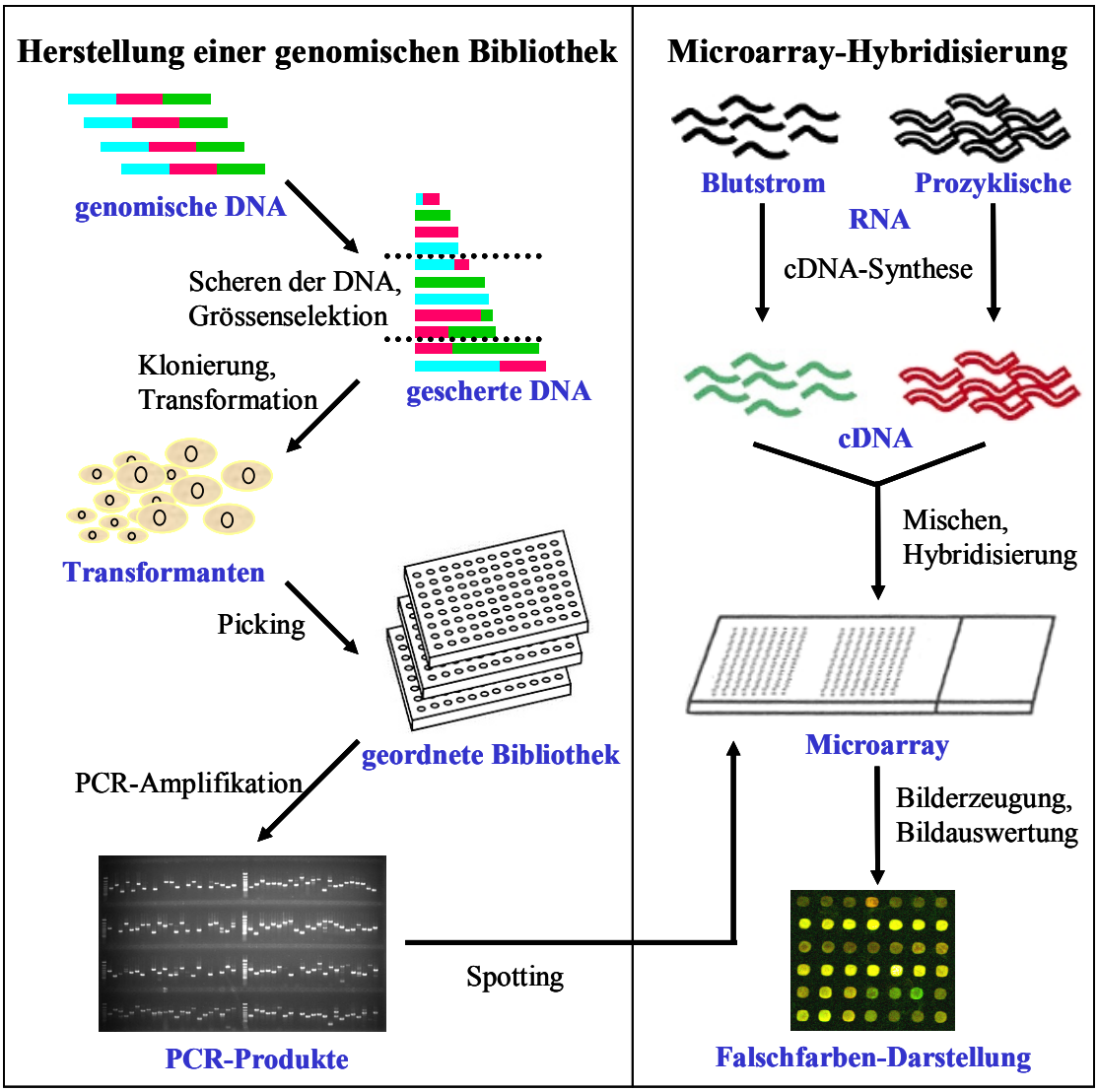


Abb. 3.3 Herstellung einer genomischen Bibliothek von *T. brucei* (links). Gescherte und grössenselektierte genomische DNA von *T. brucei* wurde kloniert und zur zur Transformation von *E. coli* verwendet. Die Transformanten wurden in 384-well Klonplatten übertragen. Diese geordnete Bibliothek wurde PCR-amplifiziert und zur Herstellung von Microarrays verwendet. Für die **Microarray-Hybridisierung (rechts)** wurde RNA aus der Blutbahnform und der prozyklischen Form von *T. brucei* isoliert und mit zwei verschiedenen Fluoreszenfarbstoffen markiert. Beide Proben wurden zusammen auf einen Microarray hybridisiert. Das Verhältnis der Fluoreszenzsignale diente zur Identifikation von Genen, deren Transkription in der Blutbahnform, bzw. der prozyklischen Form differentiell reguliert wird.

Alle Klone der genomischen *T. brucei* Bibliothek wurden mit Primern amplifiziert, die außerhalb der sogenannten *multiple cloning site (MCS)* und der M13-Primer-Stellen des Vektors liegen. Diese Primersequenzen wurden ausgewählt, um eine hohe Anlagerungstemperatur, und damit eine schnellere Amplifikation zu erreichen. Durch die hohe Anlagerungstemperatur der Primer von 65°C konnte die Dauer der Kühlphase bei der Amplifikation erheblich reduziert werden, so daß eine Amplifikation über 35 Zyklen in einer Stunde und 20 Minuten abgeschlossen werden konnte. Zur Überprüfung der Länge, Qualität und Ausbeute der PCR-Produkte wurde ein Aliquot aller Reaktionen auf Agarosegelen

analysiert (Abb. 3.4). Durch die Verwendung von Betain und Kresolrot im PCR-Reaktionsansatz konnte auf die Zugabe von Ladepuffer vor dem Auftragen auf das Agarosegel verzichtet werden. Die Erfolgsrate der PCR betrug ungefähr 95%. Lag die Ausfallrate der Amplifikation höher als 10%, wurde die PCR für die betreffende Platte wiederholt.

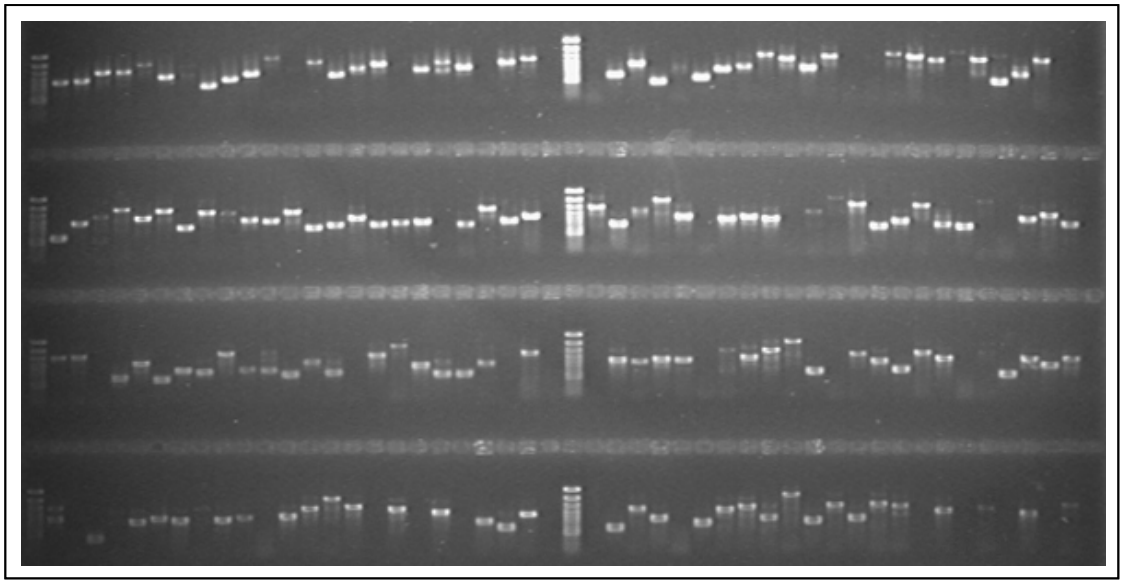


Abb. 3.4 Gelelektrophorese von PCR-Produkten der genomischen Klone von *T. brucei*. Die PCR-Produkte wurden in einem 1,5%igem Agarosegel aufgetrennt.

Spotten der Microarrays

Zum Spotten auf Objektträger wurden je 3 µl der PCR-Produkte in spezielle 384-well Platten überführt. Durch die Anwesenheit von Betain in der PCR-Lösung konnten die PCR-Produkte ohne vorherige Aufreinigung auf der Glasoberfläche immobilisiert werden (Diehl *et al.*, 2002a). Dies ist darauf zurückzuführen, das durch die Eigenschaften von Betain eine größere Menge an DNA auf die Oberfläche gebunden werden kann. Die Anwesenheit von Betain in der Spotting-Lösung verhindert ebenfalls die ungleichmäßige Anlagerung von DNA auf dem Microarray. Die erhöhte Homogenität der gebundenen DNA hat außerdem einen direkten Einfluß auf die Reproduzierbarkeit der Genexpressionsdaten (Diehl *et al.*, 2001).

Die Microarrays wurden mit einem Roboter der Firma Biorad hergestellt, der mit speziellen Spotnadeln (Telechem, Inc.) ausgerüstet war. Der verwendete Nadeltyp besitzt einen Spalt, der beim Eintauchen in die DNA-Lösung durch Kapillarkraft gefüllt wird. Durch den Kontakt der Nadel mit der Glasoberfläche wird ein kleines Volumen (ca. 1 nl) abgegeben. Der Durchmesser der resultierenden Spots beträgt ca. 90-110 µm. Die Proben wurden mit 16

Nadeln gleichzeitig aus 384-well Platten entnommen und auf die Objektträger aufgebracht. Die Herstellung von 75 Objektträgern mit insgesamt 22.177 Spots dauerte etwa 60 Stunden. Die Abstände zwischen den Spots in *x*- und *y*-Richtung betrugen 210 µm. Mit dieser Einstellung wäre es theoretisch möglich, 40.800 Spots auf einem Glasobjektträger (25×75 mm = 18,75 cm^2) unterzubringen, also eine Spotdichte von 2.177 Spots/cm^2 zu erreichen. Aus technischen Gründen konnte für die Herstellung der Microarrays allerdings nur 18×54 mm (9,7 cm^2) des Objektträgers für das Aufbringen von DNA genutzt werden. Der Abstand zwischen den Spotnadeln im Druckkopf beträgt 2,5 mm. Um eine möglichst hohe Spotdichte zu erreichen, wird der Druckkopf bei jedem neuen Spot um den geringstmöglichen Abstand in *x*-Richtung versetzt, bis der erste Spot der benachbarten Nadel in *x*-Richtung erreicht wird. Ist die erste Reihe vollständig bedruckt, wird der Druckkopf in y-Richtung versetzt, bis wiederum der erster Spot der benachbarten Spotnadel in *y*-Richtung erreicht wird. Auf diese Weise entstehen Blöcke, die jeweils von einer Nadel gespottet wurden. In der *x*-Richtung wurden so 21 Spots und in der *y*-Richtung 22 Spots pro Nadel gesetzt, also für jede Spotnadel 462 Spots in einem Block. Durch die Anordnung der Spotnadeln in einen 4×4-Raster (18×18 mm) ergab sich die Anordnung von insgesamt 7.392 Spots in einem Raster aus 16 Blöcken (*grid*). Insgesamt wurden drei Grids gespottet, daraus ergab sich eine Gesamtanzahl von 22.176 Spots auf einem Objektträger (Abb. 3.5).

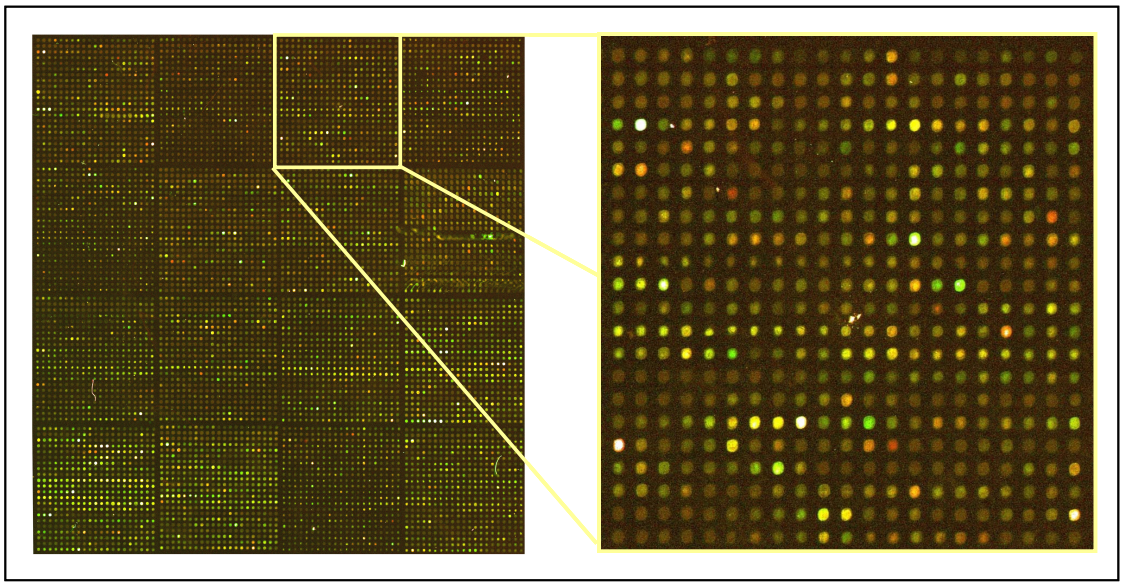


Abb. 3.5 *Trypanosoma brucei* Microarray (Ausschnitt). Der Microarray wurde mit einer Probe aus der Blutbahnform (grün) und der prozyklischen Form (rot) von *T. brucei* hybridisiert. Die Abbildung zeigt einen der insgesamt drei Blöcke des *T. brucei*-Microarrays. Jeder Block enthält 7.392 Spots. Innerhalb des Blocks sind die 16 Grids zu erkennen, die von jeweils einer Nadel gespottet wurden. Der vergrößerte Ausschnitt zeigt einen Grid, bestehend aus 462 Spots.

3.2.2. Datenbearbeitung

3.2.2.1. Allgemeine Einführung in die Datenbearbeitung

Die Microarray-Technik ermöglicht die Messung der differentiellen Transkription tausender Gene in einem einzigen Experiment. Die eigentlichen Messwerte einer Microarray-Analyse sind die Signalintensitätswerte für die Cy5- und Cy3-Hybridisierungen (R_i; G_i), aus denen man einen Quotienten (R_i/G_i) bildet, der den Grad der differenziellen Genexpression beschreibt. Dieser Wert wird im folgenden als Differenzwert bezeichnet. Da durch die Division der Signalintensitäten Dezimalbrüche entstehen, wenn die Signalintensität im Nenner größer ist als die des Zählers, empfiehlt sich die Logarithmierung der Differenzwerte (Nadon und Shoemaker, 2002). Durch die Logarithmierung erhält man positive und negative Differenzwerte.

Je geringer die Unterschiede der Transkripthäufigkeiten sind, die man detektieren möchte, desto genauer muß die Messung der Signalintensitäten sein, aus denen die Differenzwerte berechnet werden. Die Messung der Signalintensitäten ist jedoch mit einem Meßfehler behaftet. Dieser Fehler setzt sich aus einem Zufallsfehler und einem systematischen Fehler zusammen und führt zur Abweichung des gemessenen Werts vom tatsächlichen Wert. Durch die Bearbeitung der Daten kann dieser Fehler verringert werden. Wichtige Schritte bei der Datenbearbeitung sind die Normalisierung und die Filterung der Rohdaten. Durch die Normalisierung der erhaltenen Daten können systematische Meßfehler korrigiert werden und durch Datenfilterung unzuverlässige Datenpunkte (Ausreißer) eliminiert werden.

Der Meßfehler der Differenzwerte kann in einem Experiment, bei dem zwei verschiedene Proben auf einen Arrary hybridisiert werden, nicht bestimmt werden, da man den tatsächlichen Wert nicht kennt. In einem Kontrollexperiment wurden deshalb zwei identische Proben mit unterschiedlicher Markierung zusammen auf einen Microarray hybridisiert (Konkordanzhybridisierung). Bei der Hybridisierung identischer Proben sollte der zu erwartende Differenzwert theoretisch für alle Spots gleich 1 betragen.

Für die Konkordanzanalyse wurde je 240 µg prozyklische RNA in zwei getrennten Reaktionen mit Cy3- und Cy5-Farbstoffen markiert. Die markierte DNA wurde aufgeteilt und auf 4 Arrays hybridisiert. Anhand der Daten aus der Konkordanzanalyse wurde untersucht, wie sich verschiedene Verfahren zur Normalisierung und Datenfilterung auf die Variabilität der Differenzwerte auswirken.

3.2.2.2. Normalisierung

Vergleich zweier Normalisierungsverfahren

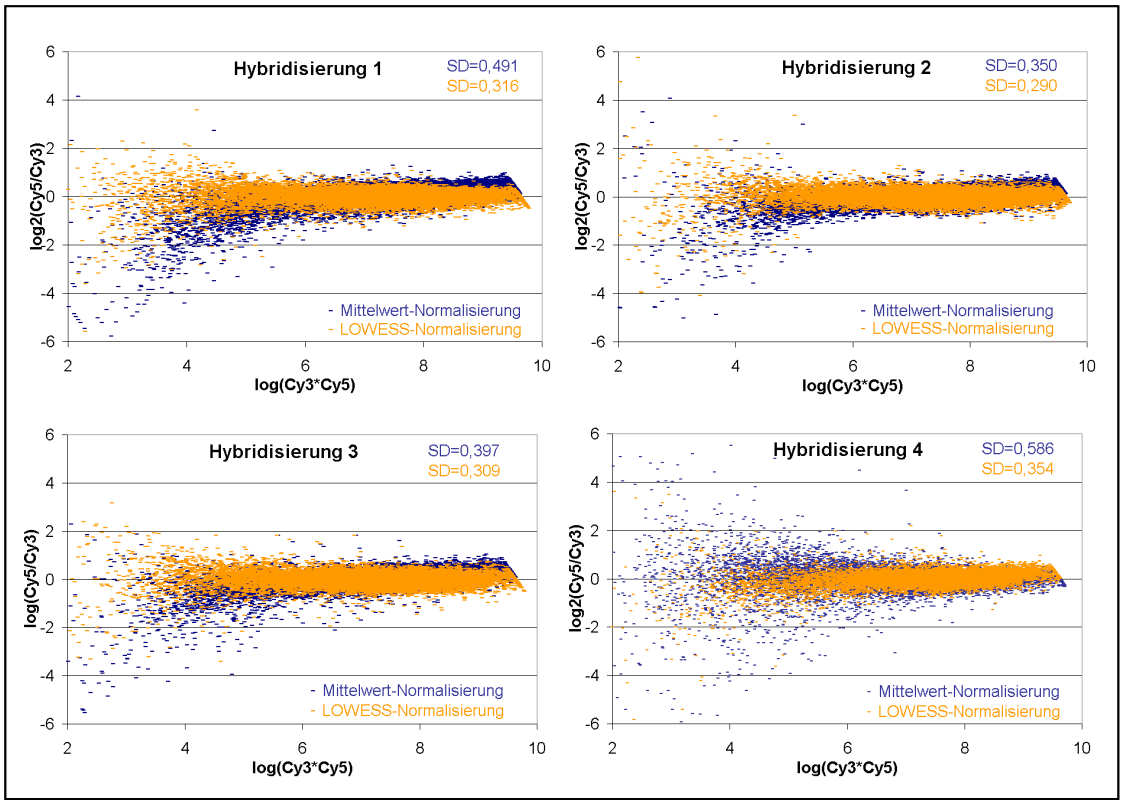
Unter Normalisierung versteht man die Anpassung der für die beiden Proben generierten Datensätze eines Arrays. Die Normalisierung ist notwendig, um die Signalintensitätswerte beider Farben für jeden Spot vergleichen zu können. Durch die Normalisierung können Meßfehler, die zu einer systematischen Verschiebung der Meßwerte führen, ausgeglichen werden. Zufallsfehler können durch Normalisierung nicht korrigiert werden.

Eine verbreitete Methode ist die Normalisierung über die Gesamtintensität. Dieses Normalisierungsverfahren beruht auf der Korrektur des farbspezifischen Meßfehlers durch die Subtraktion eines Korrekturfaktors von den logarithmierten Differenzwerten, so daß deren Mittelwert für jeden Array gleich Null beträgt. Diese Methode vollzieht also eine lineare Transformation der Daten und wird im folgenden als Mittelwert-Normalisierung bezeichnet.

Eine andere Normalisierungsmethode, die die Korrektur nichtlinearer Meßfehler erlaubt, ist die Anwendung der LOWESS-Regression (*locally weighted scatterplot smoother*) auf die Signalintensitätswerte (Dudoit *et al.*, 2002; Yang *et al.*, 2002b). Die LOWESS-Regression ist eine lokal-gewichtete Regression zur Glättung von Punktdiagrammen (Cleveland und Devlin, 1988).

Die meisten Normalisierungsverfahren können entweder auf den gesamten Datensatz eines Arrays (global) angewendet werden, oder nur auf einen Teil (lokal). Eine lokale Normalisierung wird vorgenommen, wenn einzelne Untergruppen der Daten eine systematische Verschiebung der Differenzwerte aufweisen. So können zum Beispiel Unterschiede in der Beschaffenheit der Spotnadeln dazu führen, daß die resultierenden Spots der jeweiligen Nadeln unterschiedliche DNA-Konzentrationen aufweisen. Diese Unterschiede können zu systematischen Verschiebungen zwischen den Differenzwerten der einzelnen Spotnadeln führen (Schuchhardt *et al.*, 2000). Aus diesem Grund wurde die Spotnadelzugehörigkeit der Signalintensitätswerte bei der LOWESS-Regression ebenfalls berücksichtigt. Abbildung 3.6 zeigt Punktdiagramme aller normalisierten Daten. Als Darstellungsform wurde der sogenannte MA-Plot (Dudoit *et al.*, 2000) gewählt, der die Differenzwerte $[\log_2 (R_i/G_i)]$ als Funktion des Produktes der Intensitäten $[\log_{10} (R_i*G_i)]$ darstellt. Die Standardabweichung der Differenzwerte ($\log_2 [R_i/G_i]$), die durch die jeweilige Normalisierungsmethode erhalten wurden, ist in jedem Diagramm angegeben. In dieser Abbildung ist auch die Abhängigkeit der Streuung der Differenzwerte von der Signalintensität

deutlich zu erkennen. Durch die LOWESS-Normalisierung konnte die Variabilität der Differenzwerte niedriger Signalintensitäten erheblich vermindert werden.



3.6 Konkordanzanalyse: Abhängigkeit der Differenzwerte ($log_2[R_i/G_i]$) von der Signalintensität ($log[R_i*G_i]$). Aufgetragen sind die logarithmierten Produkte der Signalintensitäten gegen die logarithmierten Quotienten der Signalintensitäten (Differenzwerte), für die LOWESS-Normalisierung (orange) bzw. Mittelwert-Normalisierung (blau).

Abbildung 3.7 zeigt die Verteilung der Differenzwerte der 16 Spotnadeln vor und nach der LOWESS-Normalisierung in einem Box-Plot. Die Unterschiede in der Verteilung der Differenzwerte der einzelnen Nadeln vor der Normalisierung ist darauf deutlich zu erkennen. Durch die spotnadelbezogene LOWESS-Normalisierung sind die Mittelwerte der einzelnen Gruppen ausgeglichen worden.

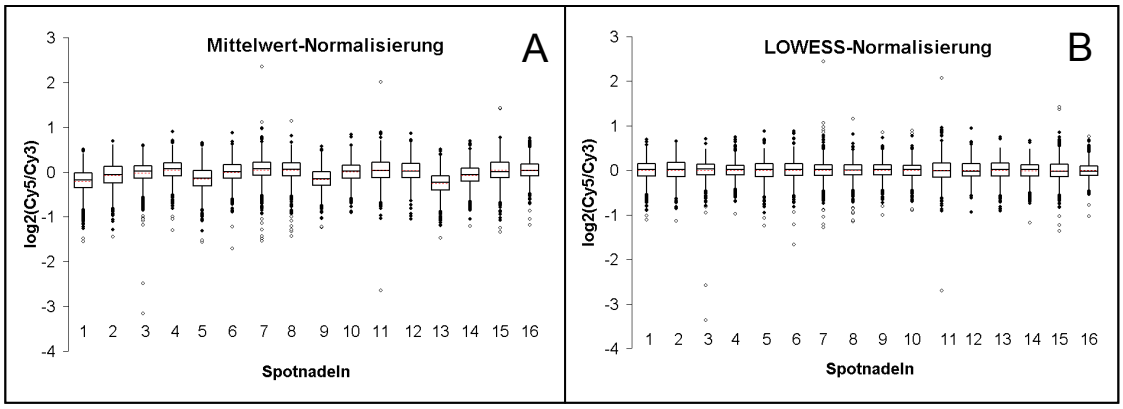


Abb. 3.7 Konkordanzanalyse: Verteilung der Differenzwerte nach Spotnadelzugehörigkeit. Der Interquartilsbereich (IQR) wird durch die *box* dargestellt, die durch den Median unterteilt wird. Die *whisker* stellen den höchsten und den niedrigsten Wert innerhalb des Bereichs vom 1. Quartil bis -1,5×IQR und vom 3. Quartil +1,5×IQR dar. Moderate Ausreißer, die außerhalb der *whisker* liegen, sind als (●) eingezeichnet, extreme Ausreißer, die mehr als 3×IQR vom 1., bzw. 3. Quartil entfernt liegen, sind als (○) eingezeichnet. (A) Differenzwerte nach der Mittelwert-Normalisierung. (B) LOWESS-normalisierte Differenzwerte.

Normalisierungseffizienz

Durch die Berechnung eines Normalisierungskoeffizienten J_{norm} (Yang *et al.*, 2003) läßt sich die Effizienz der globalen Mittelwert-Normalisierung und der Normalisierung durch LOWESS-Regression vergleichen. Für die Berechnung dieses Koeffizienten wurde folgende Formel verwendet:

$$J_{norm} = \frac{1}{p}\sum^{p}\left(\sum_{i=1}^{n}\left(\log\left[\frac{\overline{R}_i}{G_i}\right]\right)^2 \Big/ \sum_{i=1}^{n}\left(\log\left[\frac{R_i}{G_i}\right]\right)^2\right)$$

J_{norm}	Normalisierungskoeffizient
n	Anzahl der Spots
p	Anzahl der Arrays
R_i	Signalintensität Cy5
G_i	Signalintensität Cy3

Der Normalisierungskoeffizient beschreibt das Verhältnis zwischen den arithmetischen Mittelwerten der Differenzwerten aus normalisierten und nichtnormalisierten Signalintensitäten im Durchschnitt aller Hybridisierungen. Für eine perfekte Normalisierung, bei der das Verhältnis zwischen normalisierter Signalintensität R_i und Signalintensität G_i gleich 1 beträgt, berechnet sich ein Koeffizient von Null, ohne Normalisierung würde der

Koeffizient 1 betragen. Der Normalisierungskoeffizient für den Kontrolldatensatz betrug 0,96 bei globaler Normalisierung und 0,86 bei Normalisierung durch LOWESS-Regression. Die Effizienz der LOWESS-Normalisierung ist also deutlich höher als die Effizienz der globalen Mittelwert-Normalisierung.

3.2.2.3. Optimierung der Datenfilterung

Qualitätskriterien

Spots mit niedriger Qualität produzieren häufig extreme Differenzwerte und können deshalb zu einer erhöhten Falsch-Positivrate bei der Detektion differenziell exprimierter Gene führen. Extreme Differenzwerte sind Werte, die mehr als zwei Standardabweichungen vom Mittelwert entfernt liegen.

Eine wichtige Qualitätskontrolle ist die visuelle Inspektion bei der Bildanalyse, durch die Artefakte sicher identifizert werden können. Durch die Filterung der Daten nach zuvor definierten Qualitätskriterien kann die Variabilität der normalisierten Daten weiter verringert werden (Wang *et al.*, 2001). Da jede Datenfilterung auch mit dem Verlust von Daten einhergeht, ist es wichtig, effiziente Kriterien für die Filterung zu finden. Als mögliche Qualitätskriterien wurden Spotdurchmesser, Einheitlichkeit und Absolutwert des Hintergrundsignals sowie die Signalintensität des Spots und das Verhältnis zwischen Spot- und Hintergrundsignalintensität anhand ihrer Filtereffizienz verglichen. Zur Beurteilung der Effizienz der einzelnen Qualitätskriterien wurde das Verhältnis von herausgefilterten Daten zur dadurch erreichten Verringerung der Standardabweichung der Differenzwerte aus dem Konkordanzexperiment herangezogen. Bei dem Vergleich der Filtereffizienz haben sich die absolute Signalintensität des Spots und das Verhältnis zwischen Spot- und Hintergrundsignal als die effizientesten Qualitätskriterien herausgestellt. Für diese Qualitätskriterien wurden dann geeignete Schwellenwerte festgelegt.

Artefakte

Spots, die durch Hybridisierungsartefakte, wie zum Beispiel Fusseln oder Fluoreszenzflecken zu verfälschten Differenzwerten führen, wurden durch visuelle Inspektion während der Bildanalyse identifiziert, markiert und in der nachfolgenden Datenbearbeitung herausgefiltert. Alle Spots, die durch das Bildanalyseprogramm aufgrund fehlender Signalintensität nicht gefunden werden konnten, wurden ebenfalls aus der Datenanalyse herausgenommen. Betrug der Anteil der markierten Spots eines Arrays mehr als 25%, wurde die gesamte

Hybridisierung nicht für die Datenanalyse verwendet. Eine der insgesamt 4 Konkordanzhybridisierungen (Nr. 4) wies einen Artefakt-Anteil in Höhe von 28,7% auf und wurde deshalb nicht analysiert. Der Anteil der durch visuelle Inspektion als Artefakte eingeordneten Spots auf den anderen 3 Microarrays betrug im Durchschnitt 9%. Dies beinhaltet auch die Spots, die durch das Bildanalyseprogramm nicht gefunden werden konnten.

Signalintensität

Die hohe Variabilität von Spots mit geringer Signalintensität ist ein bekanntes Phänomen. Vielfach werden deshalb Intensitätsfilter eingesetzt, bei denen die Spots unterhalb eines Signalintensitätsgrenzwertes herausgefiltert werden (Beissbarth *et al.*, 2000; Quackenbush, 2002).

Um die Abhängigkeit der Variabilität der Differenzwerte von der Signalintensität der Spots zu überprüfen, wurden die Spots in Signalintensitätsklassen eingeteilt und die Standardabweichung der Differenzwerte der jeweiligen Klassen bestimmt.

Als Maß für die Signalintensität eines Spots wurde das Produkt aus beiden Signalintensitäten berechnet und logarithmiert ($\log_2[R_i*G_i]$). Da sich die Signalintensitätsprodukte einzelner Spots zwischen den Hybridisierungen nicht direkt vergleichen lassen, wurden die Signalintensitätswerte der einzelnen Arrays mittels Division durch den Maximalwert ($\log_2[R_i*G_i]/\log_2[R_i*G_i]_{max}$) standardisiert. Die standardisierten Werte werden im folgenden als qA-Werte bezeichnet. In Abbildung 3.6 wird deutlich, daß Spots mit niedrigem Signalintensitätsprodukt eine höhere Variabilität besitzen und häufig hohe Differenzwerte aufweisen. Der durchschnittliche qA-Wert von Spots, deren Differenzwerte mehr als 2 Standardabweichungen vom Mittelwert entfernt liegen, betrug 0,52, wogegen der qA-Wert im Gesamtdurchschnitt 0,75 betrug. Abbildung 3.8 zeigt den Zusammenhang zwischen dem qA-Wert der Spots und der Variabilität der Daten (Standardabweichung der logarithmierten Differenzwerte). Auf einer sekundären *y*-Achse ist der jeweilige prozentuale Anteil der Signalintensitätswerte an der gesamten Datenmenge aufgetragen. 95% der Spots weisen einen größeren qA-Wert als 0,5 auf. Durch Herausfiltern der Spots, deren Signalintensitätswert unter den 5% der niedrigsten Werte eines Arrays liegen, läßt sich eine deutliche Verringerung der Standardabweichung erzielen. Die Standardabweichung konnte durch die qA-Filterung im Durchschnitt von 0,305 auf 0,232, also um 24% verringert werden.

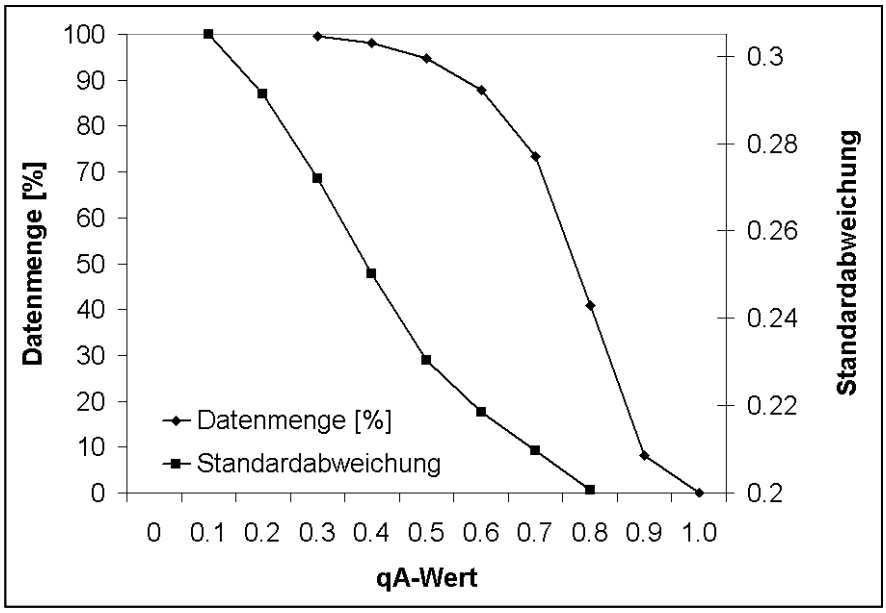


Abb. 3.8 Histogramm der prozentualen Verteilung der Signalintensität der Spots (qA-Wert) und Abhängigkeit der Standardabweichung der Differenzwerte von der Signalintensität. Die Spots sind anhand ihrer qA-Werte in Klassen eingeteilt worden. Für jede Klasse wurde die Standardabweichung der Differenzwerte und der Anteil an der Gesamtdatenmenge bestimmt. Die Prozentangaben sind kumuliert, d.h. in den jeweiligen Klassen sind alle Spots enthalten, deren qA-Wert mindestens so hoch ist wie die untere Klassengrenze. Die Standardabweichung der Differenzwerte wurde gleichermassen bestimmt, d.h. es wurde die Standardabweichung aller Spots berechnet, deren qA-Wert mindestens so hoch war wie die untere Klassengrenze. Die dargestellten Werte sind Durchschnittswerte aus der Konkordanzanalyse.

Verhältnis zwischen Spotintensität und Hintergrundintensität

Ein weiteres wichtiges Kriterium zur Beurteilung der Spotqualität ist das Verhältnis zwischen der Signalintenstät des Spots und der Signalintensität des Hintergrunds. Als Maß für die Intensitätsdifferenz zwischen Spot- und Hintergrundsignal wurde der Prozentsatz der Spotpixel herangezogen, die einen Signalintensitätswert von mindestens 2 Standardabweichungen über dem lokalen Hintergrundwert jedes Spots besitzen. Dieser Wert wird im folgenden als PP-Wert (Pixelprozent-Wert) bezeichnet. Für jeden Farbkanal eines Spots gibt es einen eigenen PP-Wert. Um aus den beiden PP-Werten eines Spots einen gemeinsamen Wert zu bilden, wurde die Quadratwurzel aus dem Produkt beider PP-Werte gebildet ($\sqrt{[R_i * G_i]}$). Abbildung 3.9 zeigt den Zusammenhang zwischen dem PP-Wert und der Variabilität der Daten (Standardabweichung der logarithmierten Differenzwerte). Aus dieser Grafik ist ersichtlich, daß Spots mit niedrigem PP-Wert häufig extreme Differenzwerte aufweisen. Der durchschnittliche PP-Wert von Spots, deren Differenzwert mehr als 2 Standardabweichungen vom Mittelwert entfernt lag, betrug 38%, wogegen der Gesamtdurchschnitt der PP-Werte aller Spots 81% betrug. Die Gesamtstandardabweichung

der Differenzwerte ($\log_2[R_i/G_i]$) wurde durch Herausfiltern von Spots mit einem PP-Wert von weniger als 40 von 0,305 auf 0,229 verringert. Der Anteil der Spots, die einem geringeren PP-Wert als 40 aufwiesen, betrug im Durchschnitt 9,6%.

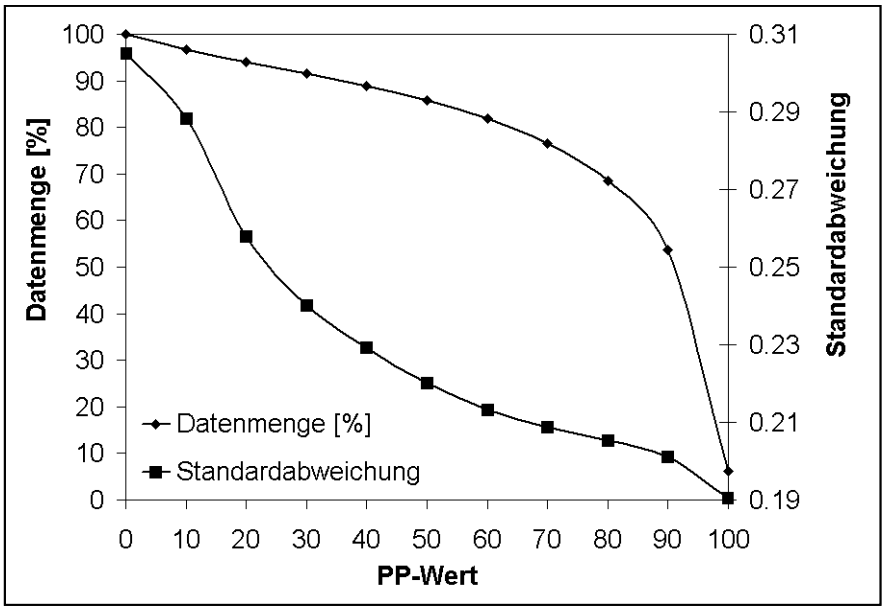


Abb 3.9 Histogramm der prozentualen Verteilung der Spots anhand ihres Signal/Hintergrundverhältnis (PP-Wert) und Abhängigkeit der Standardabweichung der Differenzwerte von der Signalintensität. Die Spots sind anhand ihrer PP-Werte in Klassen eingeteilt worden. Für jede Klasse wurde die Standardabweichung der Differenzwerte und der Anteil an der Gesamtdatenmenge bestimmt. Die Prozentangaben sind kumuliert, d.h. in den jeweiligen Klassen sind alle Spots enthalten, deren PP-Wert mindestens so hoch ist wie die untere Klassengrenze. Die Standardabweichung der Differenzwerte wurde gleichermassen bestimmt, d.h. es wurde die Standardabweichung aller Spots berechnet, deren PP-Wert mindestens so hoch war wie die untere Klassengrenze. Die dargestellten Werte sind Durchschnittswerte aus der Konkordanzanalyse.

Vergleich der Effizienz von qA-Wert und PP-Wert

Sowohl der qA-Wert als auch der PP-Wert sind zur effizienten Filterung der Daten geeignet, d.h., bei relativ geringem Datenverlust kann die Variabilität der Daten deutlich verringert werden. Die Effizienz der Filterung wurde für die Qualitätskriterien qA-Wert und PP-Wert verglichen. Dazu wurden in einer Grafik die nach der Filterung verschiedener Prozentsätze erreichten Standardabweichungen der Differenzwerte für jedes Kriterium aufgetragen (Abbildung 3.10). Durch diese Darstellung wurde deutlich, daß beide Kriterien ähnlich effizient sind. Die Kurven beider Kriterien schneiden sich bei einer Datenmenge von ca. 84%, d.h. bei einer Filterung von 16% wird die gleiche Standardabweichung für beide Kriterien erzielt. Bei der Filterung von weniger als 16% ist der qA-Wert etwas effizienter als der PP-

Wert. Bei der Filterung einer größeren Datenmenge ist der PP-Wert etwas effizienter. Bei beiden Kriterien wird jedoch unterhalb eines Werts von 20% der herausgefilterten Daten nur noch eine geringe Reduktion der Standardabweichung erzielt.

Der qA-Wert und der PP-Wert korrelieren miteinander. Der Korrelationskoeffizient dieser beiden Werte betrug 0,73. Durch die kombinierte Filterung der Daten nach den Qualitätskriterien „qA-Wert >0,5“ und „PP-Wert >40“ wurden insgesamt 10,27% der Daten herausgefiltert. Durch die Qualitätsfilterung der Daten nach LOWESS-Normalisierung konnte die Standardabweichung der logarithmierten Differenzwerte von 0,305 auf 0,225 gesenkt werden, das entspricht einer Verringerung der Variabilität der Differenzwerte um 25,6%.

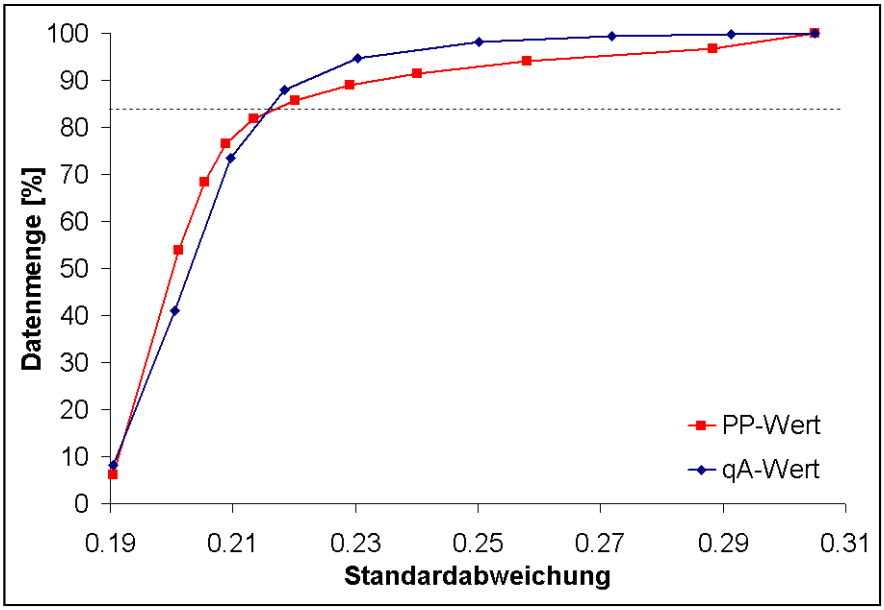


Abb. 3.10 Vergleich der Effizienz der Qualitätskriterien qA-Wert und PP-Wert. Die Effizienz der Qualitätskriterien kann durch den Vergleich der Datenmenge, die zur Verringerung der Standardabweichung auf einen bestimmten Wert herausgefiltert werden muß, beurteilt werden. Der Schnittpunkt beider Kurven liegt bei einer Datenmenge von 84% (Linie).

3.2.3. Validierung der Microarray-Technologie

3.2.3.1. Allgemeine Einführung

Die Validierung einer analytischen Methode befaßt sich vorrangig mit der Identifizierung potentieller Fehlerquellen und der Quantifizierung der Meßfehler. Dabei wird die Leistungsfähigkeit einer Methode durch quantifizierbare, statistische Methoden charakterisiert. Zur Validierung einer analytischen Methode betrachtet man die Parameter Akkuratheit, Präzision, Detektionsgrenze, Spezifizität, dynamischer Bereich und Robustheit.

Die eigentlichen Meßwerte aus Microarray-Experimenten sind Signalintensitätswerte, aus denen durch diverse Transformationen Differenzwerte berechnet werden, die ein Maß für den Grad der differentiellen Expression der untersuchten Gene darstellen. Für eine rein technische Validierung betrachtet man die primären Meßdaten. Um jedoch Aussagen über die Stichhaltigkeit der verwendeten Methode für die Untersuchung biologischer Fragestellungen beurteilen zu können, müssen die Endergebnisse der Microarray-Analyse validiert werden. Aus diesem Grund wurden für die Validierung die Differenzwerte betrachtet. Als Parameter für die Validierung der Microarray-Technologie wurden die Präzision und die Akkuratheit der Messung herangezogen. Ein für die Microarray-Technologie entscheidender Parameter ist zusätzlich das Detektionslimit der Differenzwerte. Da das Detektionslimit für die differentielle Genexpression besonders stark von der verwendeten Datenanalysemethode abhängt, wird dieser Parameter im Kapitel Datenanalyse besprochen. Die Spezifizität, der dynamische Bereich und die Robustheit der verwendeten Methodik wurde bei der Validierung außer Acht gelassen, da bei der Detektion differentieller Gene eher qualitative Aussagen getroffen werden.

3.2.3.2. Präzision

Die Präzision einer Methode bezeichnet den Grad der Übereinstimmung wiederholter Messungen der gleichen Probe. Die Präzision der Differenzwerte kann durch die Varianz, die Standardabweichung oder den Variationskoeffizienten (relative Standardabweichung [%]) der Differenzwerte eines Klons aus verschiedenen Hybridisierungen ausgedrückt werden. Zur Bestimmung der Präzision wurde aus den Konkordanzdaten der durchschnittliche Variationskoeffizient der Messung der Differenzwerte eines Spots aus den wiederholten Hybridisierungen berechnet. Der Variationskoeffizient der wiederholten Differenzwerte betrug durchschnittlich 11,75%.

Als Maß für die Reproduzierbarkeit einer Arrayhybridisierung wurden ebenfalls die Korrelationskoeffizienten der Datensätze aus dem Konkordanzdatensatz verglichen. Der durchschnittliche Korrelationskoeffizient der Signalintensitätswerte betrug 0,834 für den Cy5-Kanal (R) und 0,837 für den Cy3-Kanal (G), wogegen der durchschnittliche Korrelationskoeffizient der aus den Signalintensitätswerten ermittelten Differenzwerte nur 0,422 betrug.

Der Vergleich der Verteilung der aus den Konkordanzhybridisierungen erhaltenen Differenzwerte gibt ebenfalls Aufschluß über die Variabilität der einzelnen Hybridisierungen. Abbildung 3.11 zeigt die Verteilung der Differenzwerte der einzelnen Arrays, dargestellt in

einem Box-Whisker-Plot. Durch die Darstellung der Verteilung der Differenzwerte im Box-Whisker-Plot wird deutlich, daß die Interquartilsbereiche aller Hybridisierungen relativ schmal sind und sich zwischen den Hybridisierungen nicht sehr unterscheiden. Der Anteil der Extremwerte schwankt dagegen stark, obwohl es sich um bereits gefilterte Datensätze handelt. Diese Ausreißer haben einen starken Einfluß auf die Gesamtvariabilität der Differenzwerte.

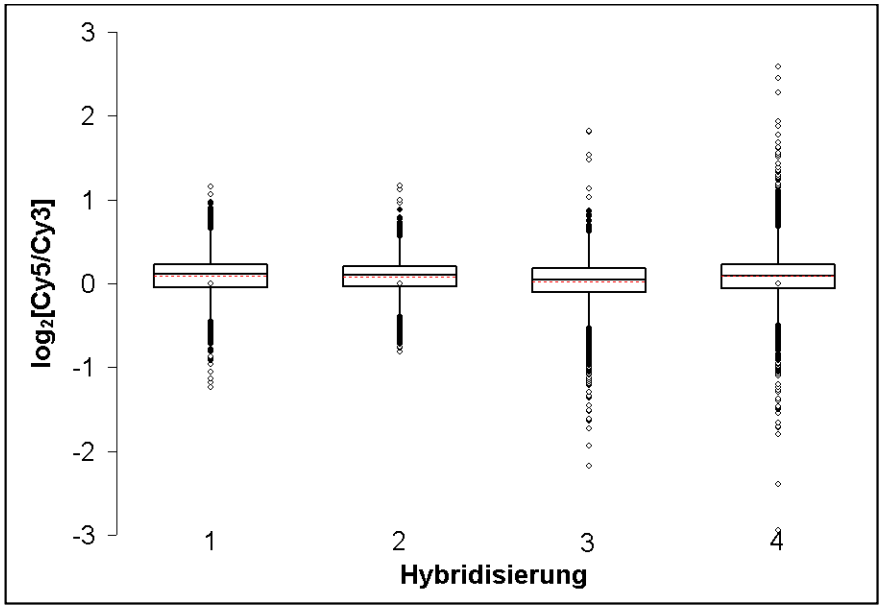


Abb. 3.11 Verteilung der Differenzwerte aus den wiederholten Hybridisierungen des Konkordanzexperiments. Der Interquartilsbereich (IQR) wird durch die *box* dargestellt, die durch den Median unterteilt wird. Die *whisker* stellen den höchsten und den niedrigsten Wert innerhalb des Bereichs vom 1. Quartil bis -1,5×IQR und vom 3. Quartil +1,5×IQR dar. Moderate Ausreißer, die außerhalb der *whisker* liegen, sind als (●) eingezeichnet, extreme Ausreißer, die mehr als 3×IQR vom 1., bzw. 3. Quartil entfernt liegen, sind als (○) eingezeichnet.

3.2.3.3. Akkuratheit der Messung der Signalintensitäten

Die Akkuratheit ist ein Maß für die Exaktheit einer Methode, d.h. für die Abweichung zwischen dem gemessenen und dem tatsächlichen Wert oder einem Referenzwert. Da man den tatsächlichen Wert der Transkriptionsunterschiede nicht kennt, dient das Ausmaß der Übereinstimmung (Konkordanz) der Signalintensitätswerte (R_i; G_i) zur Bestimmung der Akkuratheit. Die Verteilung der Differenzwerte der Konkordanzanalyse ermöglichte die Abschätzung des technisch bedingten Meßfehlers innerhalb eines Arrays. Die Konkordanz der roten und grünen Intensitätwerte auf einem Array läßt sich durch den Pearsonschen Korrelationskoeffizienten beschreiben. Der Korrelationskoeffizient betrug durchschnittlich

0,98. Betrachtet man jedoch die aus den Intensitäten berechneten Differenzwerte, wird deutlich, daß die Differenzwerte trotz hoher Korrelation der zugehörigen Intensitätswerte stark streuen. Der Variationskoeffizient der Differenzwerte eines Arrays betrug nach LOWESS-Normalisierung und Qualitätsfilterung im Durchschnitt 11,34%. Die Häufigkeit der aus den Konkordanzhybridisierungen erhaltenen Differenzwerte nach Normalisierung und Qualitätsfilterung ist in Abbildung 3.12 dargestellt. Das Histogramm liefert eine grafische Darstellung der Variabilität der Meßwerte der Mittelwert-normalisierten Daten und der LOWESS-normalisierten, gefilterten Daten. Die angegebenen Häufigkeiten beziehen sich auf die normalisierten Differenzwerte [$\log_2(R_i/G_i)$].

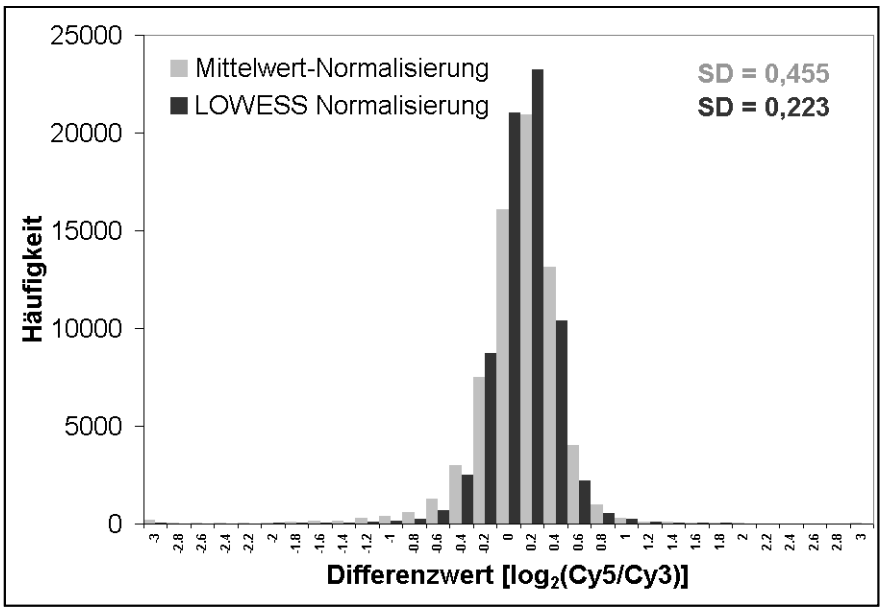


Abb. 3.12 Histogramm Konkordanzanalyse. Ein Histogramm der Differenzwerte aller Konkordanzhybridisierungen mit prozyklischer RNA nach Mittelwert-Normalisierung (grau) und nach LOWESS-Normalisierung (schwarz).

3.2.4. Datenanalyse

3.2.4.1. Überprüfung auf Normalverteilung

Statistische Verfahren zur Detektion differentiell exprimierter Gene beruhen zumeist auf der Berechnung einer Prüfstatistik zur Beurteilung der Signifikanz der gemessenen Differenzwerte. So kann für jedes Gen oder jeden auf dem Array repräsentierten Klon ein sogenannter *p*-Wert berechnet werden. Die Berechnung der *p*-Werte setzt Kenntnisse über die Verteilung der Differenzwerte unter einer Nullhypothese (keine differenzielle Expression)

voraus. Aus pragmatischen Gründen wird häufig die Normalverteilung der Differenzwerte zur Berechnung der *p*-Werte angenommmen. Um zu prüfen, ob die Normalverteilungsannahme bei den erhaltenen Daten gerechtfertigt ist, wurde die erhaltene Verteilung der Signalintensitätswerte und der Differenzwerte mithilfe eines Normal-Quantilplots (Rice, 1995) auf Vorliegen einer Normalverteilung überprüft. Dabei werden die beobachteten Datenquantile gegen die Quantile einer Standardnormalverteilung aufgetragen. Bei Vorliegen einer Normalverteilung würde man eine Gerade erhalten. Abbildung 3.13 zeigt einen repräsentativen Normal-Quantilplot der Differenzwerte und der Signalintensitätswerte aus der Konkordanzhybridisierung Nr. 3. Auf der Abbildung ist zu erkennen, daß der Graph der Verteilung der Differenzwerte keine Gerade bildet, sondern eine starke Krümmung aufweist. Die Krümmung deutet darauf hin, daß die Extremwerte der Verteilung häufiger sind, als es bei Normalverteilung der Fall wäre. Die starke Wölbung des Graphen bedeutet, daß die Verteilung der Differenzwerte eine starke Kurtosis (Exzess) gegenüber einer Normalverteilung besitzt. Das bedeutet, daß Werte nahe dem Mittelwert häufiger vorkommen. Aus dem Normal-Quantilplot geht also eindeutig hervor, daß sowohl die Verteilung der Signalintensitäten als auch die der Differenzwerte nicht einer Normalverteilung entspricht. Im Gegensatz zur Verteilung der Differenzwerte nimmt die Verteilung der Signalintensitäten einen konkaven Verlauf gegenüber der Normalverteilung an, die auf eine rechtschiefe Verteilung der Signalintensitäten schließen läßt. In dieser Verteilung sind die hohen Signalintensitäten häufiger als die niedrigen Werte.

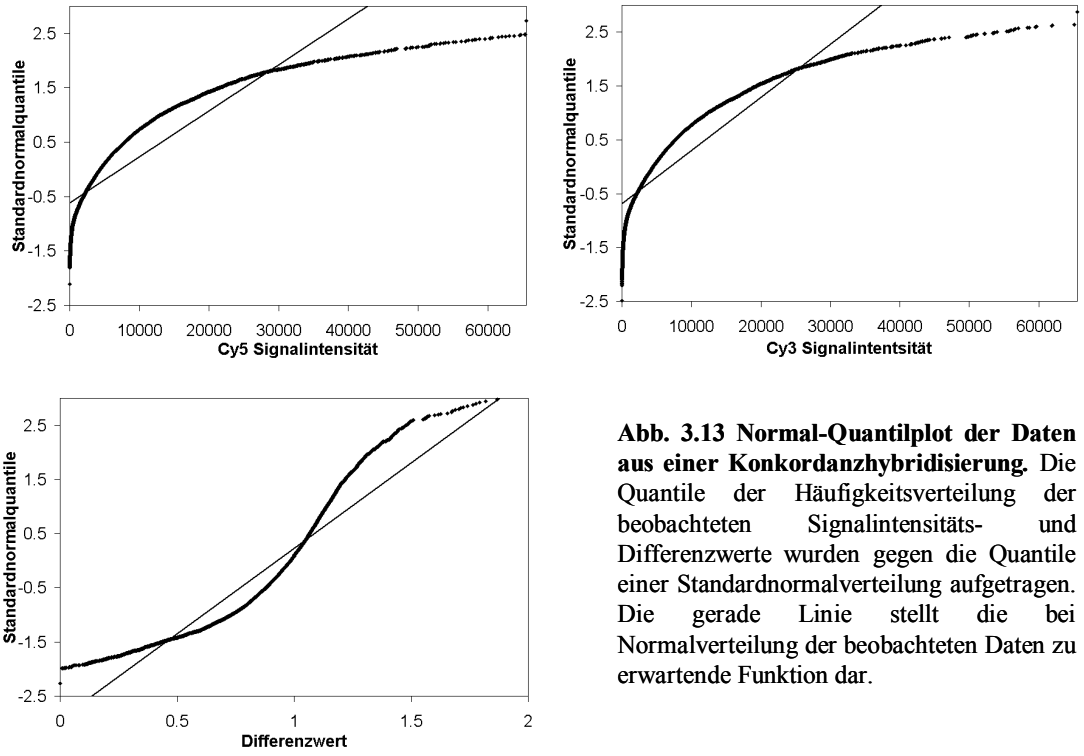


Abb. 3.13 Normal-Quantilplot der Daten aus einer Konkordanzhybridisierung. Die Quantile der Häufigkeitsverteilung der beobachteten Signalintensitäts- und Differenzwerte wurden gegen die Quantile einer Standardnormalverteilung aufgetragen. Die gerade Linie stellt die bei Normalverteilung der beobachteten Daten zu erwartende Funktion dar.

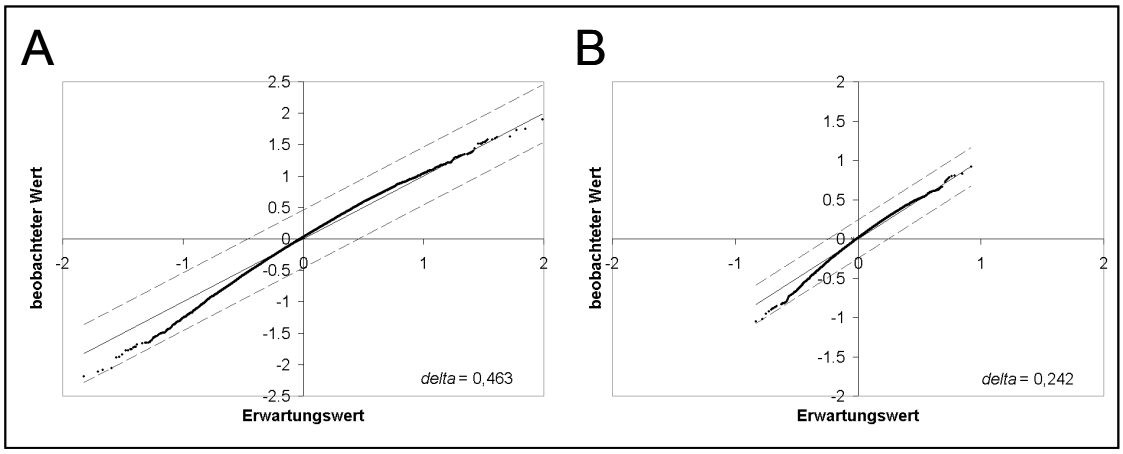
3.2.4.2. Signifikanzanalyse für den Kontrolldatensatz

Die am häufigsten angewendeten statistischen Testverfahren zur Detektion differenzieller Gene sind parametrische Verfahren, die eine Normalverteilung als Nullverteilung (Verteilung ohne Signifikanzen) zur Vorraussetzung haben. Die Konkordanzanalyse kann als experimentelles Verfahren zur Ermittlung der Nullverteilung der Differenzwerte betrachtet werden, da in dem Konkordanz-Datensatz keine Signifikanzen (differentiell exprimierte Gene) enthalten sind. Die Verteilung der Differenzwerte aus der Konkordanzanalyse ist durch einen Normal-Quantilplot auf Normalverteilung überprüft worden. Da die aus der Konkordanzanalyse erhaltenen Daten der Normalverteilungsannahme parametrischer Testverfahren nicht genügten, wurde das auf einem nicht-parametrischen Testverfahren beruhende SAM-Verfahren (*significance analysis for microarray;* Signifikanzanalyse) zur Detektion differenziell exprimierter Gene ausgewählt.

Um das Detektionslimit dieser Methode zu bestimmen, wurde eine Signifikanzanalyse für den Konkordanzdatensatz durchgeführt. Das Detektionslimit für die Signifikanzanlyse ist die Signifikanzgrenze der Methode, d.h. der niedrigste Wert, bei der keine Signifikanzen gefunden werden. Die Signifikanzgrenze wird bei der Signifikanzanalyse über den Parameter *delta* festgelegt. Zur Detektion differentiell exprimierter Gene wird bei der Signifikanzanalyse

für jedes Gen ein sogenannter *score* (d_i) aus dem arithmetischen Mittel und der Standardabweichung der Differenzwerte berechnet. Aus den *scores* aller Gene wird eine Rangfolge berechnet. Durch wiederholte Permutationen der Daten wird ein Schätzwert für eine zufällige Rangfolge ohne Signifikanzen (Nullverteilung) ermittelt. Der *delta*-Wert ist ein Maß für die Abweichung zwischen beobachteter und erwarteter Rangfolge.

Um den Einfluß der Datenfilterung auf die Signifikanzanalyse zu überprüfen, wurde eine Signifikanzanalyse mit den ungefilterten und den nach den kombinierten Qualitätskriterien gefilterten Daten durchgeführt. Die Ergebnisse der Signifikanzanalyse mit den ungefilterten Daten zeigt Abbildung 3.14 A, die Ergebnisse des gefilterten Datensatzes sind in Abbildung 3.14 B dargestellt. Wie der Vergleich der Ergebnisse der Signifikanzanalyse deutlich zeigt, konnte durch die Qualitätsfilterung der Daten der minimale *delta*-Wert deutlich vermindert werden. Ohne Datenfilterung lag der *delta*-Wert, bei dem keine signifikanten Klone mehr gefunden wurden, bei 0,49, wogegen der *delta*-Wert durch die Datenfilterung auf 0,26 gesenkt werden konnte. Die Filterung der Daten ermöglicht also eine sensitivere Signifikanzanalyse, da diese dann mit einem niedrigeren *delta*-Wert durchgeführt werden kann.



3.14 Signifikanz-Analyse des Konkordanzexperimentes. Die bei zufälliger Verteilung zu erwartenden Differenzwerte [$d_E(i)$] sind in einem Punktediagramm gegen die beobachteten Werte [$d(i)$] aufgetragen. Für die durchgezogenen Linie gilt $d(i) = d_E(i)$. Die gestrichelten Linien geben die Distanz des jeweiligen *delta*-Werts von der durchgezogenen Linie an. (A) Ergebnisse der Signifikanz-Analyse mit den ungefilterten Daten, (B) Ergebnisse der Signifikanz-Analyse für die gefilterten Datensätze.

3.2.5. Identifizierung von stadienspezifischen Genen in der Blutbahnform und der prozyklischen Form von *T. brucei*

3.2.5.1. Experimenteller Aufbau

Das Ziel dieser Arbeit war es, die Genexpression der prozyklischen Form und der Blutbahnform von *T. brucei* zu charakterisieren. Dazu wurden 4 Microarray-Hybridisierungen durchgeführt. Für die Genexpressionanalyse wurden die Blutbahnform und die prozyklische Form *in vitro* kultiviert. Um zufällige Unterschiede der Handhabung beider Kulturen auszugleichen, wurden für jede Hybridisierung separate Zellkulturen angesetzt, und die RNA aus diesen Kulturen wurde getrennt präpariert.

Die Blutbahnformen wurden bei einer Zelldichte von $1*10^6$/ml und die prozyklischen Formen bei einer Zelldichte von $2*10^6$/ml geerntet. Anschließend wurde die Gesamt-RNA beider Formen isoliert. Die Gesamt-RNA wurde durch reverse Transkription in cDNA umgeschrieben. Während der reversen Transkription wurde Cy5- bzw. Cy3-markiertes dCTP in die cDNA eingebaut. Für jede Hybridisierung wurden je 60 µg prozyklische RNA und 60 µg der Blutbahnform-RNA markiert. Es wurden drei Hybridisierungen mit Cy5-markierter Blutbahnform RNA und Cy3-markierter prozyklischer RNA vorgenommen. Um Farbstoffeffekte auszugleichen wurde außerdem eine inverse Hybridisierung vorgenommen. Bei der Markierung der Proben für diese Hybridisierung wurden die Farbstoffe vertauscht, d.h. die Blutbahnform wurde Cy3-markiert und die prozyklische RNA wurde Cy5-markiert.

3.2.5.2. Analyse der Hybridisierungsdaten

Datenbearbeitung

Mit den aus den Hybridisierungen erhaltenen Daten wurde eine LOWESS-Normalisierung durchgeführt. Anschließend wurde die Datenfilterung anhand der in Kapitel 3.2.2.3. aufgeführten Qualitätskriterien durchgeführt. Insgesamt wurden etwa 20% der Daten herausgefiltert. Abbildung 3.15 zeigt die MA-Plots der erhaltenen Daten nach Mittelwert-Normalisierung und nach LOWESS-Normalisierung.

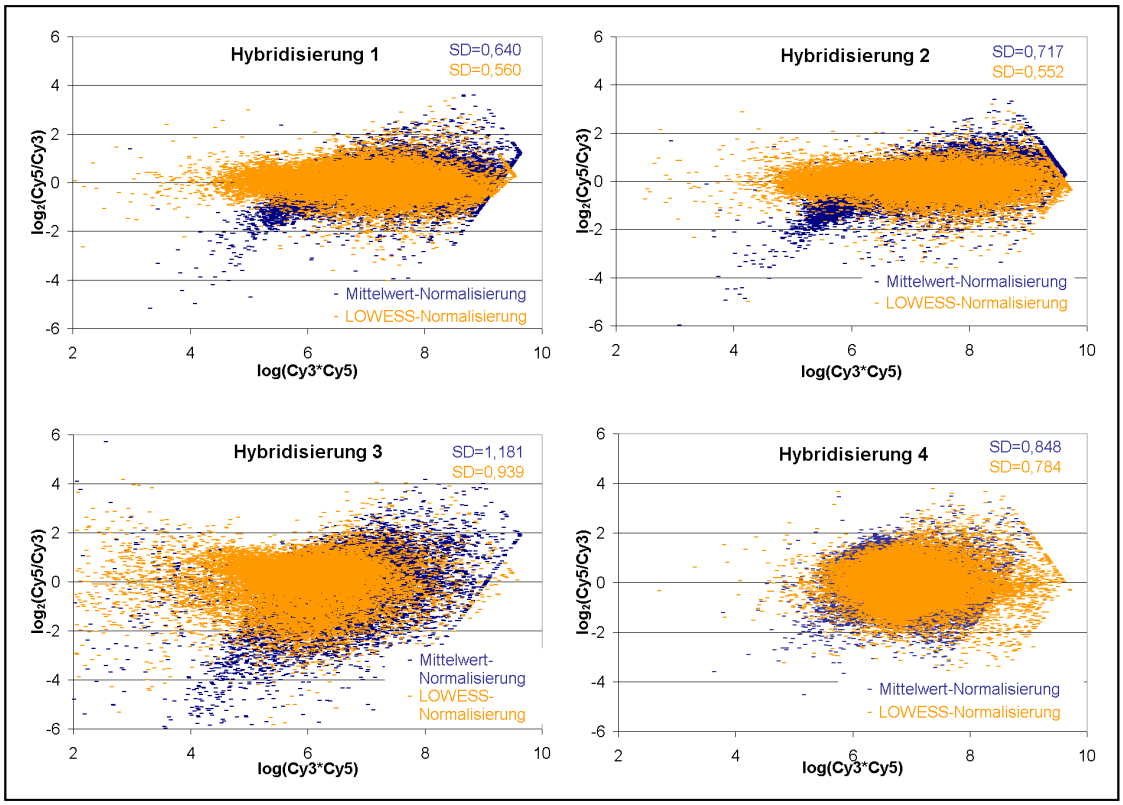


Abb. 3.15 Punktdiagramme der Daten aller Hybridisierungen mit RNA aus der Blutbahnform und der prozyklischen Form. Im Diagramm sind die Signalintensitätsprodukte [log(Cy5*Cy3)] gegen die Differenzwerte [$\log_2$(Cy5/Cy3)] eines Spots aufgetragen. Die dargestellten Werte sind LOWESS- (orange) bzw. Mittelwert-normalisiert (blau). Die Standardabweichungen der jeweiligen Datensätze sind in der Grafik angegeben.

Verteilung der Differenzwerte

Im Vergleich zur Konkordanzanalyse war die Standardabweichung der Differenzwerte bei den Hybridisierungen mit RNA aus der Blutbahnform und der prozyklischen Form wesentlich höher. Die Standardabweichung der Differenzwerte betrug im Durchschnitt 0,71 [$\log_2$(Cy3/Cy5)] im Vergleich zu 0,22 bei der Konkordanzanalyse. Abbildung 3.16 zeigt ein Histogramm der Differenzwerte aller Hybridisierungen mit den Proben aus Blutbahnform und prozyklischer cDNA.

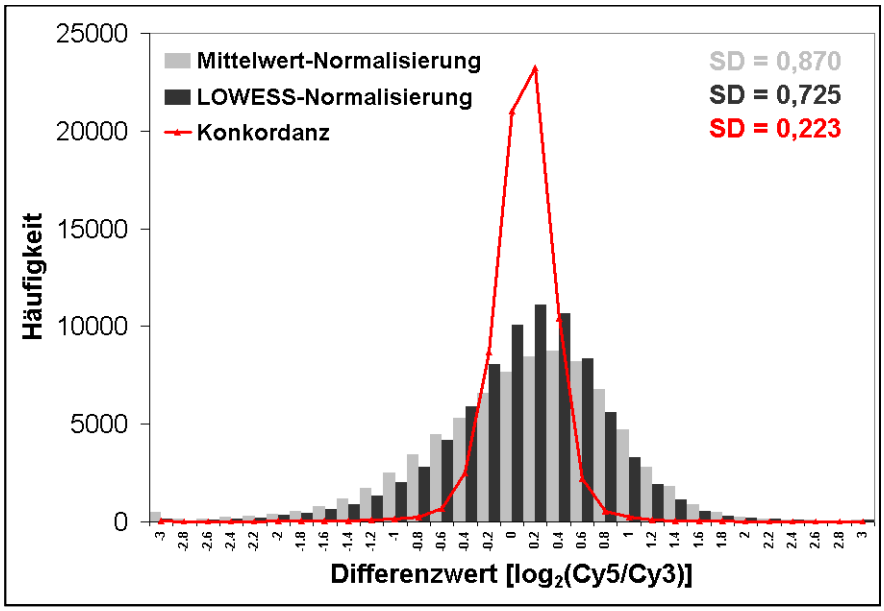


Abb 3.16 Histogramm der Differenzwerte aller Hybridisierungen mit RNA aus der Blutbahnform und der prozyklischen Form nach der Mittelwert-Normalisierung (grau) und nach LOWESS-Normalisierung (schwarz). Die rote Kurve gibt die Verteilung der LOWESS-normalisierten und gefilterten Differenzwerte der Konkordanzhybridisierung wieder.

Prüfung auf Normalverteilung

Um die erhaltene Verteilung der Differenzwerte aus den Hybridisierungen mit Blutbahn- und prozyklischer RNA zu charakterisieren, wurde die Verteilung der Differenzwerte mithilfe eines Normal-Quantilplots auf Normalverteilung überprüft. Der Normal-Quantilplot in Abbildung 3.17 läßt eine starke Abweichung der Verteilung der Differenzwerte von der Normalverteilung erkennen. Der Normal-Quantilplot zeigt eine konkave Form, die auf eine rechtssteile Verteilung schliessen lässt. Bei einer rechtssteilen Verteilung sind die negativen Extremwerte (prozyklisch-spezifisch) der Verteilung wesentlich weniger häufig, als bei einer Normalverteilung zu erwarten wäre, wogegen die positiven Extremwerte (Blutbahnform-spezifisch) viel häufiger vorkommen.

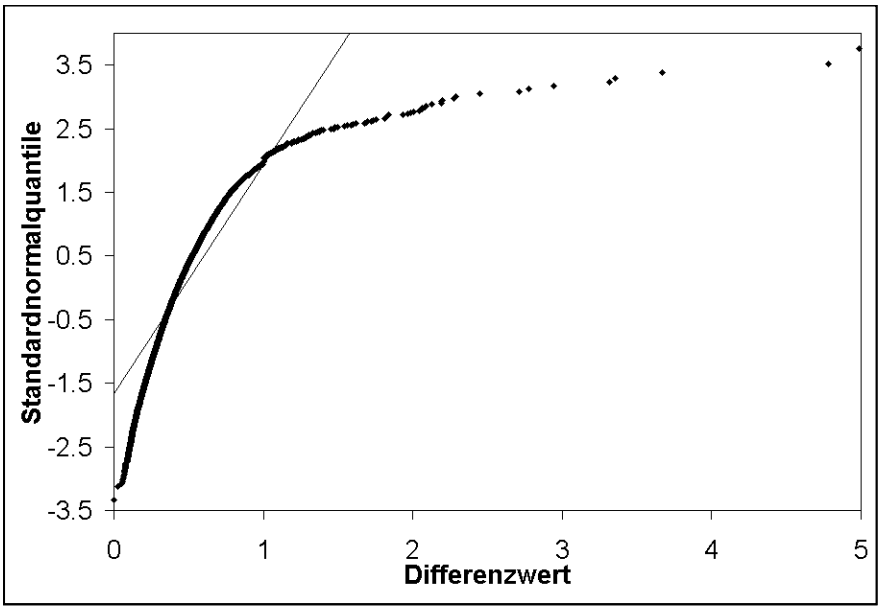


Abb. 3.17 Normal-Quantilplot der Daten aus dem Experiment zur Identifikation stadienspezifisch exprimierter Gene in der Blutbahnform und der prozyklischen Form. Die Quantile der Häufigkeitsverteilung der LOWESS-normalisierten Differenzwerte wurden gegen die Quantile einer Standardnormalverteilung aufgetragen. Die gerade Linie stellt die bei Normalverteilung der beobachteten Daten zu erwartende Funktion dar.

Signifikanzanalyse

Mit den LOWESS-normalisierten und gefilterten Datensätzen wurde eine Signifikanzanalyse durchgeführt. Die Differenzwerte wurden dabei so formuliert, daß die in der Blutbahnform stärker exprimierten Klone positive Differenzwerte und die in der prozyklischen Form exprimierten Sequenzen negative Differenzwerte bilden. Für die Signifikanzanalyse wurden die Differenzwerte der inversen Hybridisierung mit -1 multipliziert, um die umgekehrte Markierungsrichtung der Proben auszugleichen.

Zunächst wurde eine *delta*-Tabelle für diesen Datensatz berechnet, die die Anzahl der falsch-signifikanten Gene bei verschiedenen *delta*-Werten angab. Anhand dieser Tabelle wurde ein *delta*-Wert von 0,63 gewählt. Bei diesem *delta*-Wert betrug der Medianwert für die Anzahl der falsch-signifikanten Ergebnisse 3,04. Anhand dieses *delta*-Wertes wurden Klone mit signifikanten Differenzwerten ausgewählt. Das Ergebnis der Signifikanzanalyse zeigt Abbildung 3.18. Bei einem *delta*-Wert von 0,63 wurden 400 signifikante Klone detektiert, davon 309 positiv signifikante und 91 negativ signifikante Klone. Der niedrigste positiv signifikante Differenzwert betrug 1,56, der höchste negativ signifikante Differenzwert betrug –1,81. Der Medianwert für die FDR (*false discovery rate*) betrug 0,75. Der höchste *q*-Wert für positiv und negativ signifikante Klone betrug 0,67%.

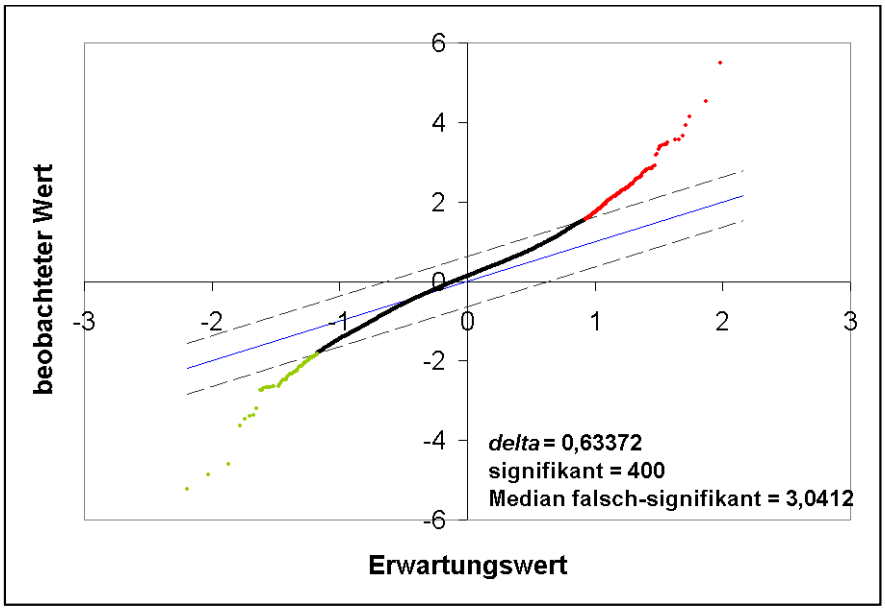


Abb. 3.18 Signifikanz-Analyse der Daten aus dem Experiment zur Identifikation differentiell exprimierter Gene der Blutbahnform und der prozyklischen Form. Die bei zufälliger Verteilung zu erwartenden Differenzwerte [$d_E(i)$] sind in einem Punktediagramm gegen die beobachteten Werte [$d(i)$] aufgetragen. Für die durchgezogenen Linie gilt $d(i) = d_E(i)$. Die gestrichelten Linien geben die Distanz des jeweiligen *delta*-Werts von der durchgezogenen Linie an.

3.2.5.3. Identifikation der differenziell exprimierten Gene

Sequenzierung

Zur Identifikation der Identität eines Teils der differenziellen Sequenzen wurden 70 Klone sequenziert. Diese Klone wurden anhand ihres Differenzwerts und der Standardabweichung ausgewählt. Alle Klone wurden beidseitig mit einer Lauflänge von jeweils ca. 1.000 Basenpaaren sequenziert. Die Sequenzierungen wurden von der Firma Genotype (Hirschhorn) durchgeführt. Mithilfe der Datenbank des *T. brucei* Genom-Projekts (http:www.ebi.ac.uk) wurden aus den erhaltenen Sequenzen nach Möglichkeit Contigs gebildet, um die gesamte Sequenzinformation für die Inserts zu erhalten. Die rekonstruierten Sequenzen wurden mit dem BLAST-Algorithmus (Altschul *et al.*, 1990) auf Homologien zu bekannten Sequenzen in den Nukleotid- und Protein-Sequenzdatenbanken des *National Center for Biotechnology Information* (NCBI, NIH, Bethesda, USA) und des *European Molecular Biology Laboratory* (EMBL) überprüft. Die Ergebnisse der Sequenzierung und der Sequenzanalyse sind in Tabelle 3.1 aufgeführt.

Spezifische Gene der Blutbahnform

Von den Klonen, die vorwiegend in der Blutbahnform exprimiert wurden, zeigte etwa die Hälfte eine starke Homologie zu mRNAs aus den VSG-Expressionsstellen. Unter diesen waren Klone mit Homologien zu ESAGs (*expression site associated gene*) und GRESAGs (*gene related to expression site associated gene*) und solche mit starken Homologien zu bereits bekannten VSG-Genen.

Weitere Klone enthielten die Sequenz bereits bekannter Blutbahnspezifischer-Gene, darunter GPIPLC (*glycosylphosphatidylinositol phospholipase C*), (Carrington *et al.*, 1989), ISG 65 (*invariant surface glykoprotein*), (Ziegelbauer und Overath, 1992), EP-Locus Transkriptionsterminator (Berberof *et al.*, 1996), und alternative Oxidase (Chaudhuri und Hill, 1996).

Andere Klone enthielten Sequenzen mit Homologien zu Genen, deren differentielle Expression bislang nicht bekannt war. Unter diesen Genen waren eine putative NADH-Dehydrogenase-Untereinheit, ein Aminosäure-Transporter, CAAX-prenyl Protease, axonemales Dynein und eine ATPase. Unter den Klonen mit differentiell exprimierten Sequenzen befanden sich auch 6 Klone, die kein offenes Leseraster enthielten (Tab. 3.1).

Prozyklisch-spezifische Gene

Klone mit prozyklisch-spezifischen Sequenzen beinhalteten neben Genen mit bereits bekannter differentieller Expression wie dem EP/GPEET Locus, Pyruvat-Phosphat-Dikinase, glykosomale Malat-Dehydrogenase und Hitzeschock-Protein 83 (HSP 83) auch einige interessante neue Gene. Es wurden drei ORFs mit Homologien zu drei verschiedenen Membranproteinen von *Leishmania* gefunden, von denen eines (META1) bereits als spezifisch für die Insektenstadien von Leishmania beschrieben wurde (Nourbakhsh *et al.*, 1996). Ein weiterer interessanter Fund war eine Sequenz mit Homologien zu *activator of G-Protein signaling* in der Maus. Wie bereits bei den Klonen mit Blutbahn-spezifischen Sequenzen beobachtet wurde, fanden sich auch acht prozyklisch-spezifische Klone, deren Inserts kein offenes Leseraster enthielten.

Tab. 3.1 Liste sequenzierter Klone mit differenzieller Expression in der prozyklischen Form und der Blutbahnform von *T. brucei.*

Nr.	Acc. Nr.	Mögliche Identität	Organismus	Homolog	BF/Pro
1	AJ438524, AJ438525	ESAG7	*T. brucei*		9,9 ± 4,0
3	AJ438529, AJ438530	65kDa invariant surface protein oder VSG	*T. brucei*		3,9 ± 1,3
5	AJ438508	VSG	*T. brucei*		3,8 ± 1,6
6	AJ438547, AJ438548	ESAG8	*T. brucei*		3,1 ± 0,9
8	AJ438526	Homolog zu VSG	*T. brucei*		2,7 ± 1,7
9	AJ438495, AJ438496	VSG Promoter	*T. brucei*		2,6 ± 0,9
11	AJ438560	VSG	*T. brucei*		2,5 ± 0,8
12	AJ438564	VSG	*T. brucei*		2,4 ± 0,9
13	AJ43856, AJ438562	ESAG2	*T. brucei*		2,3 ± 0,7
15	AJ438520, AJ438521	VSG	*T. brucei*		2,3 ± 0,8
17	AJ438545	VSG	*T. brucei*		2,1 ± 0,7
18	AJ438543	ESAG4	*T. brucei*		1,9 ± 0,6
19	AJ438544	VSG	*T. brucei*		1,9 ± 0,5
20	AJ438538	GRESAGESAG4	*T. brucei*		1,7 ± 0,2
21	AJ438556	GRESAG2ESAG2	*T. brucei*		1,6 ± 0,5
22	AJ438518	Hypothetisches Protein , gehört zur Pfam-B_9175	mehrere Arten		3,3 ± 1,3
23	AJ438498	Protein (Fragment)	*L. major*	Q9N7K2	2,9 ± 0,9
24		GPI-PLC	*T. brucei*	CAB60085	2,7 ± 1,0
25	AJ438532	75kD *invariant surface glycoprotein*	*T. brucei*		2,5 ± 0,7
26	AJ438539	putative CAAX Prenyl-Protease	*T. cruzi*	AF252543	2,3 ± 1,0
27	AJ438565	Ribosomales P0 Protein	*T. congolense*	AB056702	2,3 ± 1,0
28		*Alternative Oxidase*	*T. brucei*	U52964	2,2 ± 0,7
29		DNA fuer Prozyklin PARP A Transkriptionsterminator	*T. brucei*	Z96932	2,1 ± 0,6
30		axonemales Dynein, schwere Kette	*T. brucei*		2,0 ± 0,8
31	AJ438567, AJ438568	kein ORF			4,6 ± 1,3
33	AJ438559	ORF, keine signifikanten Homologien			3,2 ± 1,3
34	AJ438552	kein ORF			3,1 ± 1,2
35	AJ438533	repetitive Sequenz, kein ORF			2,6 ± 1,1
36	AJ438540	ORF, keine signifikanten Homologien			2,6 ± 0,9
37	AJ438499	kein ORF			2,4 ± 0,9
38	AJ438492, AJ438542	kein ORF			2,0 ± 0,5
40	AJ438502	ORF, keine signifikanten Homologien			1,9 ± 0,6
41	AJ438515	kurze ORFs, keine signifikanten Homologien			1,9 ± 0,6
42		PARP-A Locus	*T. brucei*	X52584	-2,3 ± 0,8
43		PARP-A Locus	*T. brucei*	X52585	-3,9 ± 1,1
44		Pyruvat-Phosphatdikinase	*T. brucei*		-8,9 ± 2,9
45		glykosomale Malat-Dehydrogenase	*T. brucei*		-6,7 ± 2,4
46	AJ438566	putativer Aminosaeuretransporter	*L. major*	Q9BHF5	-4,8 ± 1,8
47		Hitzeschockprotein 83	*T. brucei*	X14176	-4,7 ± 0,4
48	AJ438513	*activator of G-protein signaling*	Maus		-4,6 ± 1,4
49		Hitzeschockprotein 83	*T. brucei*	X14176	-3,5 ± 0,4
50	AJ438510	Hypothetisches Membranprotein	*L. major*	AL133435	-3,3 ± 1,2
51	AJ438528	putativer Aminosaeuretransporter	*L. major*	Q90BHG8	-3,1 ± 0,8
52	AJ438501	Retrotransposon Hotspot Protein	*T. brucei*	CAD21755	-3,1 ± 0,8
53	AJ438549, AJ438550	F1 ATPase alpha Untereinheit	*T. brucei*	AY007705	-3,1 ± 0,8
55	AJ438519	infective insect stage-specific protein META1	*L. donovani*	O43990	-3,1 ± 0,9
56	AJ43851, AJ438512	Tropomysin	mehrere Arten		-2,9 ± 0,9
58	AJ438509	Hypothetisches Protein	*L. donovani*	AL139794	-2,8 ± 0,9
59	AJ438503, AJ438504	NADH-Cytochrom B5 Reduktase	mehrere Arten		-2,6 ± 0,9
61	AJ438554, AJ438555	Importin *beta*	mehrere Arten		-2,3 ± 0,8
63	AJ438551	kein ORF			-4,1 ± 1,1
64	AJ438516, AJ438517	kein ORF			-4,0 ± 1,4
66	AJ438493, AJ438494	kein ORF			-3,8 ± 1,0
68	AJ438527	kein ORF			-3,7 ± 1,9
69	AJ438557, AJ438558	kein ORF			-3,7 ± 1,4
71	AJ438563	ORF, keine signifikanten Homologien			-3,4 ± 1,2
72	AJ438506, AJ438507	ORF, keine signifikanten Homologien			-3,1 ± 1,1
74	AJ438541	ORF, keine signifikanten Homologien			-3,0 ± 1,0
75	AJ438553	Match *T. brucei* Contig 1.0.4383			-2,7 ± 1,1
76	AJ438514	ORF, keine signifikanten Homologien			-2,5 ± 1,1
77	AJ438536, AJ438537	ORF, keine signifikanten Homologien			-2,5 ± 0,8
79	AJ438505	kein ORF			-2,4 ± 0,6

3.2.5.4. Verifikation der Genexpression mit RT-PCR

Die differentielle Expression der gefundenen Gene wurde durch semi-quantitative RT-PCR überprüft (St Croix *et al.*, 2000). Zur Quantifikation der Transkriptmenge wurde eine Verdünnungsreihe von der prozyklischen und der Blutbahnform Gesamt-RNA hergestellt. Für die Amplifikation wurden Primer verwendet, die einen Bereich von 200-300 Basenpaaren des entsprechenden Transkripts amplifizieren. Die PCR-Amplifikation wurde für alle ausgewählten Klone erfolgreich durchgeführt. Diese Ergebnisse bestätigten die differentiellen Transkriptmengen der putativen CAAX-prenyl-Protease, der alternativen Oxidase, und von GRESAG Protein 2.1-1 in der Blutbahnform. Ebenfalls wurde die prozyklisch-spezifische Expression des putativen Aminosäure-Transporters, des Hitzeschock-Proteins 83 (HSP 83), *activator of G-Protein-Signaling*, des hypothetischen Proteins L71710.1, der ATPase α-Untereinheit, der putativen Importin β-Untereinheit und *EP* (Abb. 3.19) bestätigt.

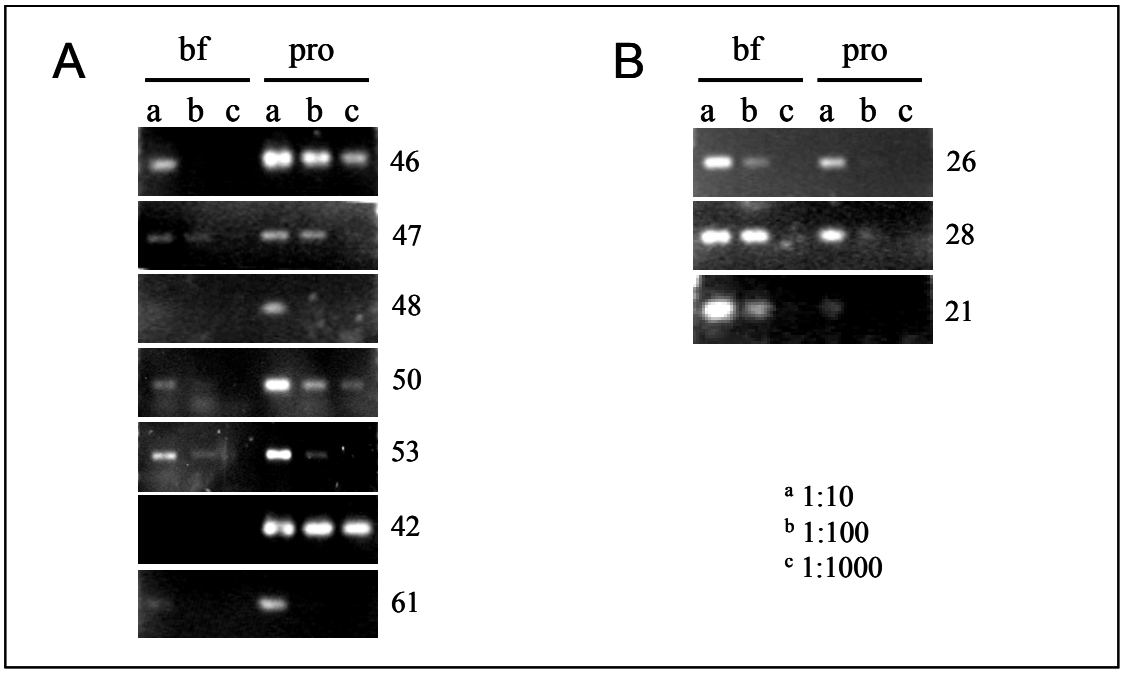


Abb. 3.19 Validierung der differentiellen Expression mittels semi-quantitativer RT-PCR. Als Ausgangsmaterial für die PCR-Reaktion wurde cDNA von prozyklischen Trypanosomen (pro) und von Trypanosomen der Blutbahnform (bf) eingesetzt. Die cDNA wurde 1:10 (a), 1:100 (b), und 1:1000 (c) verdünnt. Die Gene sind mit den Nummern aus Tabelle 3.1 bezeichnet.