Analyse des Genoms von *Dictyostelium discoideum*

Habilitationsschrift

vorgelegt am
der Biologisch–Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von
Dr. Gernot Glöckner
aus Aschaffenburg

Gutachter

1.

2.

3.

Erteilung der Lehrbefähigung am

Inhaltsverzeichnis

Abbildungen und Tabellen 5

1. Einleitung 6

1.1. Geschichte der Genomanalyse 6

1.1.1. Techniken 7

1.1.1.1. Sequenzierung und Assemblierung 7
1.1.1.2. Analyse 8

1.1.2. Genome 10

1.1.2.1. Prokaryontengenome 10
1.1.2.2. Eukaryontengenome 11
1.1.2.3. Basenungleichgewichte 12

1.2. Der Organismus 13

1.2.1. Phylogenetische Position 13

1.2.2. Sexueller und vegetativer Lebenszyklus 15

1.2.3. Der Modellcharakter 16

1.2.4. Das Genom 18

2. Ergebnisse 19

2.1. Das Dictyostelium Genomprojekt - eine internationale Anstrengung 19

2.2. Sequenzierung 21

2.3. Globalanalyse 23

2.3.1. Basenungleichgewichte und Kodonnutzung 23

2.3.2. Komplexe Repetitive Elemente 29

2.3.3. Genkatalog 33

2.4. Assemblierungsstrategien 37

2.5. Kartierungen 41

2.6. Analyse von Chromosom 2 47

2.6.1 Das Chromosom im Überblick 47

2.6.2 Genmodelle und Proteinanalyse 52

2.7. Die Analyse des gesamten Genoms 59

2.8. Zugänglichkeit der Daten 60

3. Diskussion 61

3.1. Zweck der Arbeit 61

3.2. Verschiedene Sequenzierstrategien 62

3.3. Genkatalog bei niedriger Redundanz 64
3.4. Integration von Karte und Sequenz 65
3.5. Repetitive Elemente 69
3.6. Chromsomenenden 70
3.7. Junge Duplikationen 70
3.8. Basenungleichgewichte 71
3.9. Annotation von Genen 73
3.10. Gene und Domänen 77
3.11. *D. discoideum* als Modell 79
3.12. Das *D. discoideum* Genomprojekt im Vergleich zum Humangenomprojekt 81
3.13. Funktionelle Analyse 82
4. Zusammenfassung 84
5. Literatur 87

Abbildungen und Tabellen

1. Abbildungen

Abbildung 1: Phylogenie der Mycetozoa 14
Abbildung 2: Der vegetative Lebenszyklus von *D. discoideum* 16
Abbildung 3: Die Aufteilung der Chromosomen von *D. discoideum* an die Mitglieder des Sequenzierungskonsortiums 21
Abbildung 4: Asparagin Homopolymerregionen in verschiedenen Eukaryonten 28
Abbildung 5: Schematische Darstellung der möglichen Insertionen von TRE-Elementen vor oder hinter tRNA Genen 32
Abbildung 6: Das Assemblierungsschema 39
Abbildung 7: Bildung von Scaffolds anhand von read pair information 40
Abbildung 8: Eine durch HAPPY Map Marker definierte Kopplungsgruppe 45
Abbildung 9: Verteilung von Merkmalen auf der 8,5 MB Kopplungsgruppe von Chromosom 2 48
Abbildung 10: Schematische Darstellung der Entwicklung des Chromosom 2 aus AX2, dem Ursprungsstamm hin zur Situation in AX4, dem sequenzierten Stamm 51
Abbildung 11: Genvorhersageprogramme im Vergleich 52
Abbildung 12: Die Aufteilung von Proteinen zwischen Repräsentanten verschiedener Entwicklungslinien 55
Abbildung 13: Klassifizierung der Chromosom 2 kodierten Proteine anhand der Zuordnung von Interpro Domänen zur GO Terminologie 57

2. Tabellen

Tabelle 1: Die Anteile der verschiedenen Klonbibliotheken an der Gesamtsequenzproduktion in Jena 22
Tabelle 2: Verteilung von 6er Tupeln im Genom von *D. discoideum* 25
Tabelle 3: Kodonnutzung von *D. discoideum* (A) und *P. falciparum* (B) 27
Tabelle 4: Berechnete Häufigkeiten von ausgewählten Zytoskelettgenen im Genom 36
Tabelle 5: Anzahl der auf der größten Kopplungsgruppe von Chromosom 2 festgestellten Komplexen Element Familien 49
Tabelle 6: Generelle Eigenschaften der auf Chromosom 2 mit GeneID vorhergesagten Genmodelle 53
Tabelle 7: Ausgewählte Interpro Domänen von *D. discoideum* Chromosom 2 kodierten Proteinen, die in *S. cerevisiae* nicht vorhanden sind 59

1. Einleitung

1.1. Geschichte der Genomanalyse

Bis vor wenigen Jahren war an die Analyse eines gesamten Genoms wegen technischer Limitationen nicht zu denken. Zwar wurden schon früh Virengenome entschlüsselt (Cashdollar et al., 1984; Sanger et al., 1977a; Sharp et al., 1984), doch die Genomanalyse autarker Organismen schien wegen der schieren Größe der Genome lange außerhalb jeder Möglichkeit. Vielmehr konnte nur die Sequenz einzelner Gene, im Höchstfall etliche Kilobasen (kb), aufgeklärt werden. Ausgangspunkt für diese Genanalysen sind zumeist positionale Klonierungs- oder Komplementationsexperimente z.B. (Glöckner and Beck, 1997; Momeni et al., 2000; Stoyan et al., 2001). Für eine Genomanalyse fehlten lange Zeit die Voraussetzungen wie z.B. geeignete Klonierungssysteme oder die Möglichkeit, große Mengen an Daten zu verarbeiten. In den frühen 80er Jahren wurden erste technische Voraussetzungen für die Genomanalyse durch die Herstellung von Cosmidbanken geschaffen (Bates and Swift, 1983; Little and Cross, 1985). Diese ermöglichten es erstmals, längere DNA-Stücke zu amplifizieren und damit selektiv zugänglich zu machen. Mitte der 80er Jahre schließlich wurden erste Stimmen laut, die die Entzifferung des menschlichen Genoms forderten. Von einzelnen Genen zum drei Gigabasen großen Genom, das war jedoch ein gewaltiger Schritt. Deswegen wurden ab 1986 zunächst Pilotprojekte finanziert, mit denen die nötigen Technologien und Ressourcen entwickelt werden sollten. Schon zu diesem Zeitpunkt war klar, dass das Humangenomprojekt eingebunden sein müsse in einen größeren Rahmen, in dem auch so genannte Modellorganismen analysiert werden sollten. Merkmale, die bestimmte Spezies zu Modellorganismen machen, sind u.a. deren evolutionäre Stellung, die Zugänglichkeit ihres Genoms für Manipulationen und die geringe Größe ihres Genoms. Die dichte Abfolge von Genen in kompakten Genomen macht deren Auffindung weniger schwer als z.B. in komplexen Vertebratengenomen. Der Vergleich potentieller Genstrukturen mit ähnlichen Sequenzen aus dem komplexeren Genom des Menschen würde dann auch die Generkennung in *Homo sapiens* erleichtern. Zunächst wurden etliche Genome von Bakterien vollständig entschlüsselt. 1995 publizierte The Institute for Genomic Research (TIGR) die erste komplette DNA Sequenz eines frei lebenden Organismus, des Bakteriums *Haemophilus influenzae* (Fleischmann et al., 1995). Mittlerweile sind mehr als 100

komplette Genome von Bakterien (Eubacteria und Archaebacteria) in den öffentlichen Datenbanken verfügbar (Stand Ende 2005). Schon 1996 wurde dann die Sequenz des ersten Eukaryonten, der Bäckerhefe *Saccharomyces cerevisiae* von einem internationalen Konsortium veröffentlicht (Goffeau et al., 1996). Es dauerte dann nur noch knapp 5 Jahre, bis eine Skizze des Humangenoms verfügbar war (Lander et al., 2001; Venter et al., 2001). Diese Skizze wurde mit Hilfe zweier unterschiedlicher Strategien erstellt. Der öffentlich geförderte Ansatz verwendete BAC Bibliotheken zur Erstellung dieser Skizze (clone by clone), die private Initiative von Venter et al. versuchte, das gesamte Genom in Einzelstücken zu sequenzieren und zusammenzusetzen (whole genome shotgun). Inzwischen ist die gesamte euchromatische DNA des Menschen vollständig bekannt (The Human Genome Sequencing Consortium, 2004). Wie im Folgenden gezeigt werden wird, ist die Genomanalyse dennoch eine Herausforderung nicht nur logistischer Art geblieben.

1.1.1. Techniken

1.1.1.1. Sequenzierung und Assemblierung

Die Erfindung der biochemischen Sequenzierung (Sanger et al., 1977b) erlaubte relativ einfach die Herstellung kürzerer Sequenzen. Die Limitationen jedoch, die sich aus der Größenauftrennung der Sequenzierungsprodukte über eine Matrix ergeben, blieben bis heute erhalten. So kann im Höchstfall etwas mehr als eine Kilobase an Sequenz mit einer Sequenzierungsreaktion an einem Template erzeugt werden. Techniken, die unbegrenzte Sequenzlängen ergeben, sind bis heute im experimentellen Stadium. Zu nennen wären hier vor allem chipbasierte Methoden (Feldman and Pevzner, 1994; Milosavljevic, 1995). Diese werden aber momentan nur zur schnellen Re-Sequenzierung schon bekannter DNA-Abschnitte eingesetzt, ihrer breiteren Verwendbarkeit für die Genomanalyse stehen jedoch Schwierigkeiten entgegen. Darunter fallen u.a. die Kosten für einen Chip und das Auftreten von Kreuzhybridisierungen bei der Inkubation des Chips. Ein anderer Weg wurde mit der Entwicklung einer schnellen, billigen Technik, die nur relativ kurze Sequenzen von ca. 100 Basen erzeugt, begangen (Margulies et al., 2005). Hier wird der Nachteil einer kurzen, zusammenhängenden Sequenz durch eine wesentlich höhere Coverage des Zielgenoms abgefangen. Jedoch potenzieren sich damit auch die Fehlerquellen, die durch repetitive Elemente im Genom verursacht werden.

Generell macht es die geringe erzielbare Länge der Sequenzabschnitte nötig, die Gesamtsequenz aus diesen einzelnen Abschnitten später zusammenzusetzen, zu assemblieren. Während noch zu Zeiten der Einzelgenanalyse arbeitsaufwendige, halbmanuelle Methoden eingesetzt wurden, wurde für die Genomanalyse Software zur automatischen Assemblierung entwickelt (Allex et al., 1996; Huang and Madan, 1999). Diese Programme beziehen teilweise Qualitätsinformationen der Sequenzen mit ein. Im Zuge der Verbesserung der Rechnerleistungen war es möglich, immer größere DNA-Stücke zu assemblieren. Sequenzabschnitte, die länger als die zur Verfügung stehenden Einzelsequenzen sind und gleichzeitig mehrfach in einem Genom vorkommen, verursachen Probleme bei der Assemblierung von DNA Stücken, vor allem im Megabasen- Bereich. Um dieses Problem zu lösen, wurden in jüngerer Zeit Programme entwickelt, die auch Informationen über die Zuordnung von Sequenzen zueinander, die aus dem selben Klon stammten (read pair information), ermöglichen (Myers et al., 2000). Vollkommen automatische Assemblierung ist jedoch nur dann möglich, wenn mit Hilfe der Rohdaten eine Abdeckung (Coverage) der zu assemblierenden Sequenz von mindestens 5 mal erreicht wird. Aber auch dann stoßen automatische Methoden an Grenzen, wenn es gilt, nahezu identische, d.h. duplizierte oder repetierte Regionen zusammenzusetzen. Jede Assemblierung in solch kritischen Regionen muss mit Hilfe geeigneter Techniken, z.B. Polymerasekettenreaktion (PCR), überprüft werden. Des Weiteren bleiben nach einer automatischen Assemblierung immer Lücken übrig, die dadurch bedingt sind, dass:

a. bestimmte DNA Stücke nicht ausreichend häufig in der Klonbibliothek vertreten sind oder fehlen (cloning bias)
b. die verwendete Polymerase nicht in der Lage ist, über bestimmte Sequenzmotive hinaus zu gelangen (sequencing bias)

Diese Lücken können nur unter Einsatz spezifischer Methoden (Primer Walk, inverse PCR etc.) geschlossen werden.

1.1.1.2. Analyse

Ist die Sequenz mit hinreichender Genauigkeit erstellt, wird sie, ebenfalls mit Hilfe von Software, analysiert. Es gilt zunächst, die allgemeinen Eigenschaften der Sequenz zu berechnen. Dazu gehören z.B. die globale und lokale Basenzusammensetzung, der Anteil an einfachen Sequenzwiederholungen (simple repeats), potentielle stabile Haarnadeln

(hairpins). Die sogenannten komplexen repetitiven Elemente, zu denen alle Arten von Transposonen zählen, werden dann durch angepasste Software (z.B. RepeatMasker, http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) erkannt und maskiert.

Generkennung erfolgt durch Software, die in der Lage ist, organismenspezifisch Eigenschaften von Genen zu erkennen. Alle verfügbaren Programme arbeiten mit statistischen Modellen, die jedoch jeweils etwas unterschiedliche Algorithmen verwenden (Murakami and Takagi, 1998). Damit ein Programm in der Lage ist, Gene im Genom einer Spezies zu erkennen, muss ein Datensatz erstellt werden, der gut charakterisierte Genstrukturen dieses Organismus enthält. Idealerweise sind von diesen Genen sowohl die genomische als auch die mRNA Sequenz inklusive Splicevarianten bekannt. Dadurch sind dann allgemeine Eigenschaften eines Genes des Organismus wie 5´ und 3´ Umgebung des Genes, Splice-Orte und Introngrößen definiert. Dieser Datensatz wird dann dazu verwendet, die statistischen Modelle an die Wirklichkeit des zu untersuchenden Organismus anzugleichen. Die Sequenz wird dann nach Regionen mit solchen Eigenschaften abgesucht, die sich zu Genmodellen mit hoher statistischer Wahrscheinlichkeit zusammensetzen lassen.

Die Proteinsequenzen der Genmodelle werden dann auf Ähnlichkeiten zu bereits beschriebenen Genen aus anderen Organismen hin untersucht. Große Ähnlichkeit über die gesamte Proteinlänge hinweg ist ein Hinweis auf Orthologie, d.h. beide Proteine haben ein direktes gemeinsames Ursprungsgen in der Vergangenheit und könnten unter Umständen die gleiche Funktion in der Zelle ausüben. Paraloge Gene dagegen sind aus Duplikationen in einer oder mehreren Spezies hervorgegangen. Ortholgie/Paralogiebeziehungen in diesen Genen sind nur dann relativ eindeutig zu bestimmen, wenn eines der duplizierten Gene rasch Mutationen akkumuliert.

Funktionen von vielen Proteinen werden über Domänen vermittelt. Diese Domänen können dadurch gekennzeichnet sein, dass nur wenige Aminosäuren in bestimmten Abständen in der Proteinsequenz konserviert sind. Daher sind Domänen nicht unbedingt über Sequenzabgleiche zu entdecken. Vielmehr können über die Mustererkennung mit Hilfe von z.B. Hidden Markow Modellen wesentlich genauere Resultate erzielt werden. Selbst wenn mit den üblichen Alignmentalgorythmen (BLAST (Altschul et al., 1990); ClustalW (Thompson et al., 1994)) zu bereits funktionell untersuchten Proteinen keine offensichtlichen Ähnlichkeiten entdeckt werden können, können diese Domänen Aussagen über die Funktion des Proteins ermöglichen.

Um die vielfältigen Analyseprogramme zu bündeln und leichter zugänglich zu machen, wurde am IMB ein Programmpaket namens RUMMAGE entwickelt, das eine erste automatische Analyse fertiggestellter Sequenzen erlaubt (Glöckner et al., 1998; Taudien et al., 2000).

1.1.2. Genome

Aufgrund wesentlicher technischer Verbesserungen und damit einhergehender Automation stellt die Herstellung ausreichender Mengen an Sequenzrohdaten kein Problem mehr dar. Dementsprechend stieg auch die Anzahl der analysierten Genome, vor allem von Prokaryonten, in den letzten Jahren exponentiell an. Dies machte den Vergleich verschiedener Genome im Hinblick auf z.B. Kodierungskapazität, gemeinsame Gene und Spezies spezifische Gengruppen möglich.

1.1.2.1. Prokaryontengenome

Die Größe der Genome von Prokaryonten variiert von weniger als einer Megabase (MB) für Mycoplasmen (Fraser et al., 1995) bis ca. 10 MB wie z.B. bei Bradyrhizobium Arten (Kundig et al., 1993). Oft sind die Genome von Bakterien aufgeteilt in ein Chromosom und mehrere Megaplasmide (Freiberg et al., 1997). Das DNA-Molekül, das die essentiellen Gene kodiert, ist dabei unabhängig von der Größe des Chromosoms. Es können sich jedoch auch wenige essentielle Gene auf den Plasmiden befinden. Es gibt jedoch auch Prokaryonten mit mehr als einem Chromosom (Heidelberg et al., 2000). Meistens handelt es sich um ringförmige Moleküle, es wurden aber auch lineare Chromosomen und Plasmide gefunden (Fraser et al., 1997). Die Genome kodieren nur für wenige hundert bis wenige tausend Proteine. Aus einem Vergleich der ersten beiden vollständig sequenzierten Genome von Prokaryonten wurde abgeleitet, dass nur 256 Gene essentiell für zelluläres Leben sind (Mushegian and Koonin, 1996). Mit Hilfe von akkumulierten „Knockout“ Mutanten von *Mycoplasma genitalium* und *M. pneumoniae* durch Transposoninsertionen wurde versucht, das minimale Genom eines Prokaryonten experimentell zu bestimmen (Hutchison et al., 1999). Es zeigte sich auch hier, dass etwa 265 bis 350 Gene für zelluläres Leben unter Laborbedingungen notwendig sind. Eine Schwäche dieser Studie ist, dass nur Stämme mit wenigen Mutationen hergestellt wurden,

kein Stamm enthielt alle festgestellten Mutationen. Da Proteine oder Proteingruppen überlappende Funktionen haben können, kann eine in einem Minimalgenom letale Mutation unter diesen experimentellen Bedingungen ohne Auswirkungen bleiben. Auch muss beachtet werden, dass dieser Minimalsatz nur ein Leben unter Laborbedingungen ermöglichen würde. Da Mycoplasmen zudem obligat intrazellulär leben und viele Stoffe aus ihrer Umgebung aufnehmen, fehlen ihnen wichtige Synthesewege, die für ein vollständig autarkes Dasein nötig wären. Die Zahl an essentiellen Genen für ein minimales Genom eines autarken Organismus muss deshalb eher höher angesetzt werden.

1.1.2.2. Eukaryontengenome

Eukaryonten zeichnen sich gegenüber Prokaryonten durch eine stärkere Kompartimentierung der Zellen aus. Sie enthalten Organellen, die aus ehemaligen Endosymbionten entstanden sind. Diese Organellen besitzen Restgenome, die für Teile der Funktionen der Organellen wie z.B. Photosynthese kodieren. Je nach Spezies sind jedoch unterschiedliche Gene von der Zurückhaltung im Organellengenom betroffen (Glöckner et al., 2000). Offensichtlich konnten diese Restgenome nicht in den Kern transferiert und ins Kerngenom integriert werden. Die Gründe dafür sind unklar.

Die im Vergleich zu Prokaryonten höhere Komplexität der Eukaryontenzelle wird auch im nukleären Genom wiedergespiegelt. Das haploide Genom ist mindestens 5 MB groß, kann aber mehrere hundert Gigabasen (GB) erreichen. Die Zahl der Protein kodierenden Gene umspannt einen Bereich von unter 5000 z.B. bei Hefen bis knapp 25.000 Genen in Mammalia (The Human Genome Sequencing Consortium, 2004). Die Anzahl an Genen ist unabhängig von der Genomgröße, wobei die höchste bis jetzt gemessene Gendichte mit 1 Gen/2 kb bei *S. cerevisiae* auftritt (Mannhaupt et al., 1994). Die Aufblähung eines Genoms ist in erster Linie sogenannter „junk DNA“ geschuldet. Sie trägt zur Kodierungskapazität des Genoms nichts bei. Diese „nutzlose“ DNA hat teilweise eine Funktion als Regulator von Expressionshöhen und -orten. Sie könnte aber auch als ein Reservoir für evolutionäre Anpassungen dienen (Jurka, 1998; McDonald et al., 1997).

Der Einzeller *S. cerevisiae* wird als Modell für die einfachste eukaryontische Zelle angesehen, obwohl es eukaryontische Genome mit weniger Genen gibt (http://webace.sanger.ac.uk/cgi-bin/webace?db=pombase). Mit 5885 proteinkodierenden Genen können somit alle Funktionen einer frei lebenden Eukaryontenzelle erfüllt werden.

Theoretisch könnten nur 12 % dieser Gene nötig für ein Überleben der Zelle sein, wie aus Transposonmutagenese -Experimenten hervorgeht (Goebl and Petes, 1986). Mutationen in weiteren 18 % der vorhandenen Gene führten zu phänotypischen Veränderungen. Unbeeinträchtigtes Leben der Hefezelle unter Laborbedingungen müsste demnach von nur ca. 2000 Genen abhängen. Der Vergleich dieser Zahl mit dem essentiellen Genset der Prokaryonten zeigt, dass das Erreichen einer komplexeren Organisationsform nicht auf der Erfindung einiger weniger Gene beruht. Vielmehr müssen mehrere hundert Gene neue Funktionen übernehmen und deren Wirkung koordiniert werden. Mit den selben Einschränkungen wie in 1.1.2.1. kann der Komplexitätsunterschied zwischen dem minimalen Genset von Prokaryonten und Eukaryonten mit mindestens 4 - 5 fach angegeben werden.

1.1.2.3. Basenungleichgewichte

Abweichungen von der gleichverteilten Basenzusammensetzung von 50 % G+C treten häufig auf. Dies kann einzelne Gene betreffen oder auch größere Regionen bis hin zum ganzen Genom. Im Fall von einzelnen Genen ist die Ursache oft in einer anderen Expressionshöhe des Genes als im Durchschnitt des Organismus zu finden. Dieses Phänomen wird als translationale Selektion bezeichnet (Lloyd and Sharp, 1992; Pan et al., 1998). Asymetrische biochemische Prozesse wie DNA -Replikation oder -Reparatur können, bei Prokaryonten, sogar strangspezifisch, zu einer Gleichgewichtsverschiebung führen. Bei manchen Säugetieren wie dem Menschen wurden sogenannte Isochoren beobachtet (Bernardi, 1993; Sabeur et al., 1993). Sie definieren Regionen unterschiedlichen G+C Gehalts unabhängig von Genen und deren Transkriptionsniveaus. Woher diese Ungleichgewichte rühren ist unklar (Eyre-Walker and Hurst, 2001). Da sie z.B. bei Nagern nicht ausgeprägt auftreten, haben sie offensichtlich keinen speziellen funktionellen Aspekt für Säugetiere (Robinson et al., 1997).

Auch gesamte Genome können hin zu einer ungleichen Basenzusammensetzung verändert sein. Woher diese Verschiebungen rühren wird heftig diskutiert. Eine Fraktion spricht von „mutational bias“. Hervorgerufen werden kann er, wie vermutet wird, von veränderten Wahrscheinlichkeiten des Einbaus von Nukleotiden aufgrund von Mutationen in Schlüsselenzymen der Replikation und Translation (Lafay et al., 1999; McLean et al., 1998; Sueoka, 1993; Zsiros et al., 1999). Möglicherweise rührt diese Genomeigenschaft

aber auch von Selektionsvorteilen für ein ungleichgewichtiges Genom her. Extreme Basenungleichgewichte eines ganzen Genoms treten bei Bakterien (Pan et al., 1998) und bei niederen Eukaryonten (Glöckner, 2000) häufiger auf, jedoch nicht in gleichem Maße in allen Spezies einer Gattung (Gentles and Karlin, 2001; Lafay et al., 1999). Eine Verschiebung hin zu hohen A+T Werten kann häufiger beobachtet werden (Glöckner, 2000). Wenn, wie im Fall von *Dictyostelium discoideum* oder *Plasmodium falciparum*, dieses Basenungleichgewicht hin zu nur noch 20 % G+C verschoben ist, wird auch die Kodon- und Aminosäurenutzung stark davon beeinflusst (Musto et al., 1997; Singer and Hickey, 2000).

1.2. Der Organismus

Weltweit sind etliche Dictyostelienarten beschrieben worden. Die ersten Charakterisierungen stammen noch aus dem 19. Jahrhundert (Brefeld, 1869; Brefeld, 1884; van Tieghem, 1884). Die vegetative Organisationsform aller Arten dieser Gattung ist einzellig. Sie verfügen nicht über eine Zellwand oder Flagelle, vielmehr formen sie einen nackten Zellkörper, der sich amöboid über das Substrat fortbewegt. Der Lebensraum ist verrottendes Laub, Kot von Pflanzenfressern, etc.. Nahrung wird über Phago- und Pinocytose aufgenommen. Sie besteht in erster Linie aus Gram negativen Bakterien wie z.B. Klebsiellen (Depraitere and Darmon, 1978).

Auf den einzelnen Kontinenten sind jeweils verschiedene Arten vorherrschend. Allgemein scheint die häufigste Art in Europa und den USA *D. mucoroides* zu sein (William Loomis, pers. Mitteilung). *D. discoideum*, die hier untersuchte Art, wurde erstmals in den 30er Jahren beschrieben (Raper, 1935).

1.2.1. Phylogenetische Position

D. discoideum gehört zu den Mycetozoa. Diese Gruppe besteht aus 3 gut unterscheidbaren Gruppen: Echten Schleimpilzen (Myxogastria z.B. *Physarum polycephalum*), zellulären Schleimpilzen (Dictyostelia z.B. *Dictyostelium discoideum*) und den Protostelia (Baldauf and Doolittle, 1997). Lange Zeit war die Zusammengehörigkeit

und Stellung dieses evolutionären Astes innerhalb des Stammbaums der Arten ungeklärt. Die Berechnung von Stammbäumen aus SSU und LSU rRNA Sequenzen ergab, dass die Mycetozoa polyphyletisch entstanden sein könnten (De Rijk et al., 1995; Kumar and Rzhetsky, 1996). Da die Berechnungen mit Hilfe von Proteinsequenzen uneinheitliche Stammbäume lieferten, sahen die einen die Dictyostelidae an der Basis des Hauptastes zu Pflanzen und Tieren abzweigen. Andere wiederum ordneten sie zwischen echten Pilzen und Metazoa ein (Kessin, 1997). Erst die Verfügbarkeit größerer Mengen an Proteinsequenzen erlaubte die Berechnung einer relativ robusten Phylogenie (Abb. 1) (Baldauf and Doolittle, 1997; Baldauf et al., 2000). Dieser Stammbaum zeigt, dass der Ast der Mycetozoa erst kurz nach der Trennung von Pflanzen und Tieren aus den tierischen Vorläufern entstanden ist.

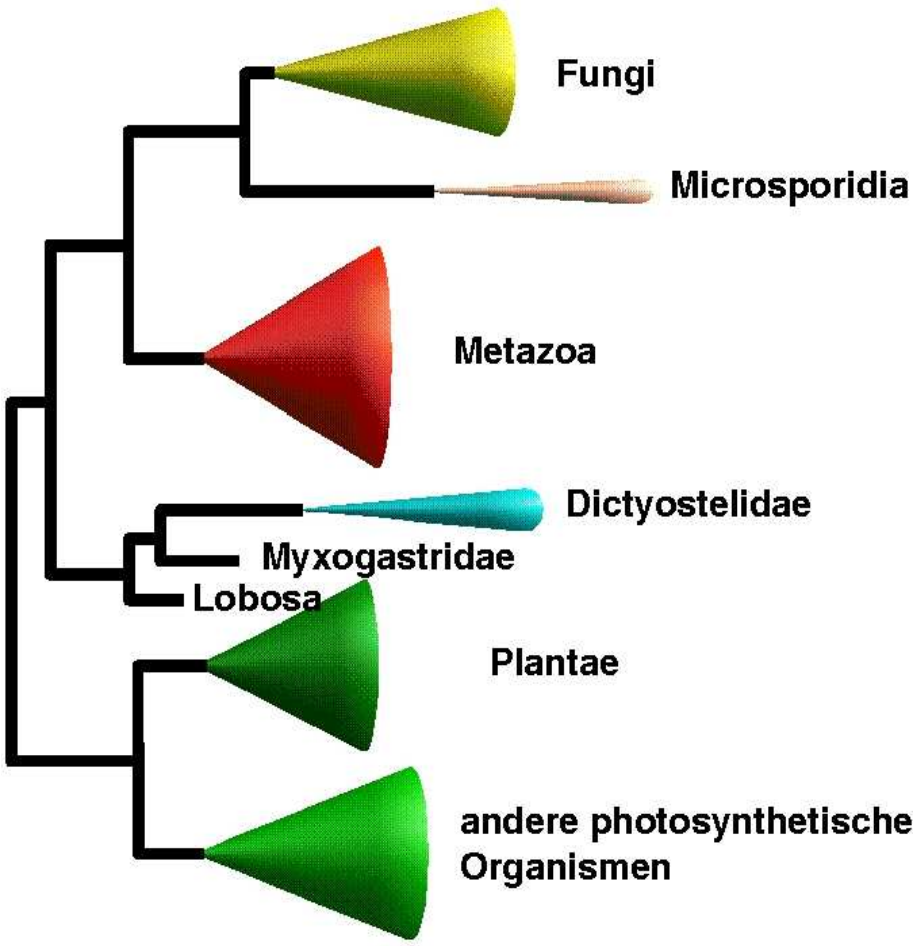


Abb. 1: Phylogenie der Mycetozoa (repräsentiert durch Dictyostelidae und Myxogastridae) im Kontext der Eukaryonten (nach (Baldauf et al., 2000) stark verändert). Der Berechnung liegen die Gene für α und ß-Tubulin, Aktin und EF1- α zugrunde. Die Größe der Dreiecke spiegelt die Anzahl der zur Konstruktion des Baumes verwendeten Taxa wieder.

Wegen des Alters dieser Verzweigung könnte man die Mycetozoa mit einigem Recht als weitere Hauptlinie neben der photosynthetischen und der zu den Tieren und Pilzen abzweigenden Linie bezeichnen. In allen drei Zweigen der Evolution wurde die Mehrzelligkeit unabhängig voneinander entwickelt. Die genetischen Grundlagen für dieses Phänomen müssen also schon in einem früheren, gemeinsamen Entwicklungsschritt vorhanden gewesen sein.

1.2.2. Sexueller und vegetativer Lebenszyklus

Damit *D. discoideum* in den sexuellen Lebenszyklus, bei dem sich zwei zu Isogameten umgebildete Zellen zu einer Zygote vereinigen und später meiotische Reifeteilungen durchmachen, eintritt, muss es dunkel und feucht sein (Habata et al., 1991). Die Rekombinationsrate bei der Meiose ist hoch und vergleichbar mit der von Hefe (Francis, 1998). Unter ungünstigen Witterungs- und Ernährungsbedingungen bevorzugt der Organismus jedoch einen vegetativen Zyklus (Raper, 1941). Der Eintritt in diesen Zyklus ist allerdings nur möglich, wenn sich genügend Zellen in ausreichender räumlicher Nähe zueinander befinden. Er wird eingeleitet durch das Aussenden von cAMP Pulsen durch eine bis wenige Amöben. Auf dieses Bildungszentrum zu fließen alle Amöben, die mit dem Chemotaktikum in Berührung gekommen sind. Die Zellen vereinigen sich schließlich zu einem Aggregationsplasmodium, wobei die Zellwände erhalten bleiben (Pseudoplasmodium). In diesem Aggregationszustand vermag das vielzellige Plasmodium wie ein einzelner, mehrzelliger Organismus zu agieren (Maree and Hogeweg, 2001). Das Aggregat selber macht eine streng geregelte Umformung durch. Zunächst wird ein frühes Hügelstadium gebildet (early mount). In dieser Phase bewegen sich die Zellen schon koordiniert, wie an wirbelförmig-spiralig verlaufenden Fronten unterschiedlicher Lichtdichte zu sehen ist (Gottmann and Weijer, 1986; Rietdorf et al., 1996). Daraufhin kriecht das Pseudoplasmodium mehrere Stunden umher. Dieses Stadium wird als slug bezeichnet. Dieses Kriechen kommt durch wellenförmige, koordinierte Zellbewegungen innerhalb des slug zustande. Nach Bildung des sogenannten tipped mount richtet sich der Zellhaufen auf, verschlankt sich, und bildet einen Sporenträger aus. Die nötigen Ausformungen zu verschiedenen Zelltypen (Fuß, Ständer, Sporenkapsel, Sporen) haben schon vor der sichtbaren Ausdifferenzierung stattgefunden. Durch Untersuchungen an aus verschiedenen Klonen gemischten Pseudoplasmodien konnte gezeigt werden, dass die

genetische Ausstattung in Abhängigkeit von der Zusammensetzung des Pseudoplasmodiums Einfluss auf die bevorzugte Bildung verschiedener Zelltypen nimmt (Queller et al., 2001). Abbildung 2 zeigt die typischen Phasen der vegetativen Entwicklung von *D. discoideum* als zusammengesetzte, elektronenmikroskopische Aufnahme. Der gesamte Zyklus zieht sich über mindestens 20 Stunden hin.

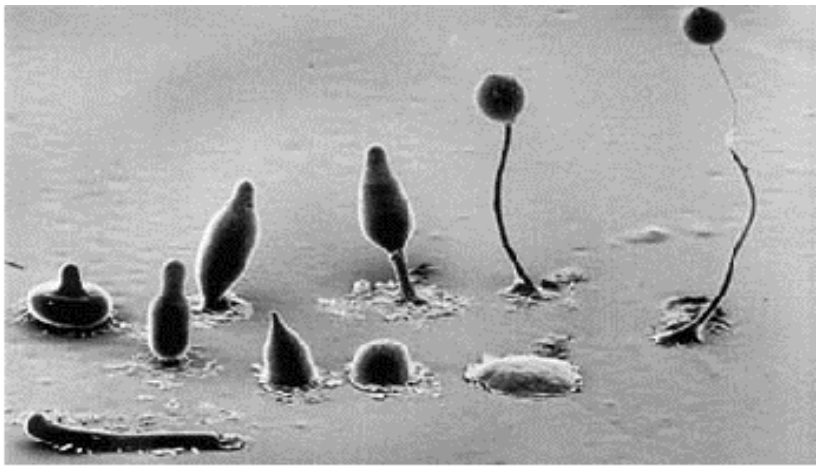


Abb. 2: Der vegetative Lebenszyklus von *D. discoideum* dargestellt als rasterelektronenmikroskopische Aufnahme, zusammengesetzt aus Einzelbildern (Bild freundlicherweise zur Verfügung gestellt von M. G. Grimson und R. L. Blanton, Texas Tech University, Lubbock, Tx, USA)

1.2.3. Der Modellcharakter

Um die vielfältigen Funktionen einer Eukaryontenzelle und auch das Zusammenspiel mehrerer differenzierter Zellen im mehrzelligen Organismus molekular verstehen zu können, müssen die genetischen Grundlagen dieser Funktionen aufgeklärt werden. In diesem Zusammenhang ist die Erkenntnis wichtig, dass komplexe Funktionen der Vertebratenzelle in den Grundzügen schon in einfacheren Eukaryonten angelegt sind. So besitzen viele Gene, die in mutierter Form Krankheiten beim Menschen auslösen können, signifikante Homologien zu Genen niederer Eukaryonten wie z.B. *Saccharomyces cerevisiae, D. discoideum, D. melanogaster* und *Caenorhabditis elegans.* In diesen Organismen können diese Gene wesentlich einfacher und ausführlicher funktionell charakterisiert werden als im Menschen (Aboobaker and Blaxter, 2000; Reiter et al., 2001; Resor et al., 2001; Sturley, 2000). Jedes Gen evolviert im Kontext des gesamten Genoms, in dem es kodiert ist. Deshalb ist es nicht verwunderlich, dass orthologe Gene in

verschiedenen Organismen für unterschiedliche Funktionen benötigt werden und dementsprechend angepasst wurden. Die Schnittmenge aller möglichen Funktionen kann nur in möglichst vielen verschiedenen Modellsystemen untersucht werden.

Als Vertebratenmodelle sind inzwischen mehrere Spezies (*Xenopus laevis, Danio rerio, Fugu rupripes, Mus musculus, Rattus norvegicus, Pan troglodides*) anerkannt, deren Genome entschlüsselt wurden, bzw. werden. Da zumindest die näher verwandten Arten sehr homologe Gene besitzen und Regionen gleicher Genanordnung konserviert haben, bietet ein Vergleich der Genome die Möglichkeit für das Auffinden von Genen sowie auch regulatorischer Einheiten. Niedere Eukaryonten wiederum eignen sich wegen der vergleichsweise einfachen Manipulierbarkeit dazu, funktionelle Analysen zu betreiben.

D. discoideum verfügt über mehrere Vorzüge, die diesen Organismus als Modell ausweisen. So ist diese Spezies einfach und billig in Massen zu kultivieren und kann über Jahre in gefrorenem Zustand gelagert werden. Das Genom ist haploid und kann relativ leicht manipuliert werden, das Methodenrepertoire ist hier ähnlich breit wie bei *S. cerevisiae*. Da mehrere Selektionsmarker vorhanden sind, ist Transformation möglich. Homologe Rekombination kann zur Generierung von knock-out Mutanten benutzt werden. Veränderte Gene können vor gleichbleibendem Genomhintergrund getestet werden (knock-in).

D. discoideum unterscheidet sich jedoch in wichtigen Aspekten wie der Beweglichkeit und der Entwicklungsfähigkeit zu einem mehrzelligen Organismus von *S. cerevisiae*. Diese Eigenschaften haben ihre genetische Grundlage in Genen z.B. des Zytoskeletts oder der Signaltransduktion, die in Hefe nicht oder nur rudimentär vorhanden sind. Wegen dieser zusätzlichen Funktionen wird *D. discoideum* vor allem bei der Analyse amöboider Zellbewegung, der Zytoskelettorganisation, und der Signaltransduktion bei Entwicklungsvorgängen eingesetzt (Eichinger et al., 1999; Escalante and Vicente, 2000; Firtel and Meili, 2000; Kay, 2000). Diese Arbeiten waren und sind entscheidend für ein besseres allgemeines Verständnis dieser Funktionen. Neuere Arbeiten, in denen gezeigt wurde, dass *D. discoideum* eine Wirtszelle für humanpathogene Bakterien (Legionella spec.) sein kann, haben ein weiteres Forschungsfeld eröffnet (Hagele et al., 2000). Nun können am Modell *D. discoideum* die Mechanismen der Aufnahme und Erhaltung der Bakterien in der Zelle untersucht werden. Des Weiteren ist *D. discoideum* inzwischen ein etabliertes Expressionssystem für Proteine (Linskens et al., 1999; Malnasi-Csizmadia et al., 2000).

1.2.4. Das Genom

Wie bei vielen niederen Eukaryonten ist das Genom von *D. discoideum* in der vegetativen Phase haploid. Mit 34 MB hat es nur ca. 1/3 bis 1/4 der Größe der kleinsten Genome der bis jetzt sequenzierten echten Vielzeller (*C. elegans*, *D. melanogaster*, *Arabidopsis thaliana* mit rund 100 MB bis 120 MB (Adams et al., 2000; The C. elegans Sequencing Consortium, 1998; The Arabidopsis Genome Initiative, 2000)). Das Genom ist auf sechs Chromosomen verteilt, die zwischen vier und acht MB groß sind (Loomis and Kuspa, 1997). Zusätzlich enthält der Kern das sogenannte rDNA-Palindrom, das ca. 90 kb groß ist und die Gene, die für die rRNA kodieren, enthält. Dieses Palindrom besteht aus einer perfekten inversen repetierten Region von 45 kb, getrennt nur durch asymetrische 42 Basen im Zentrum. Dieses Palindrom ist mit ca. 100 Kopien hoch amplifiziert und trägt 20 % zum Gesamt-DNA-Gehalt im Kern bei. Für die Chromosomen existieren integrierte physikalische und genetische Karten, die jedoch teilweise inkorrekt sind (Kuspa and Loomis, 1994; Kuspa and Loomis, 1996; Loomis et al., 1995).

Das Genom weist ein starkes Basenungleichgewicht auf, d.h. die Basen A und T sind stark überrepräsentiert. Mit im Durchschnitt über 78 % hat *D. discoideum* einen der höchsten A+T Gehalte, die bis jetzt festgestellt wurden (Firtel and Bonner, 1972). Übertroffen wird diese Zahl nur von *Plasmodium falciparum*, dem Verursacher einer Form der Malaria (Bowman et al., 1999; Gardner et al., 1998). Kodierende Regionen haben einen etwas niedrigeren A+T Gehalt, da hier Limitationen hinsichtlich der Kodonbenutzung bestehen. Dafür können in intergenischen Regionen G+C fast völlig fehlen, was zu bis zu 98 % A+T über längere Strecken (> 1 kb) führt. Der große Unterschied in der Basenzusammensetzung zwischen Genen und intergenischen Regionen bzw. Intronen erleichtert die manuelle Generkennung. Jedoch sind vorhandene Klonierungs- und Sequenzierungssysteme auf einen durchschnittlichen A+T Gehalt optimiert. Das führt dazu, dass es nicht möglich ist, bakterielle Klonbibliotheken aus A+T reicher DNA herzustellen, deren Plasmidinserts länger als ca. 5 kb sind und stabil bleiben (Triglia and Kemp, 1991). Vielmehr sind solche Plasmide in *Escherichia coli* vielfältigen Veränderungen wie Rearrangements und Deletionen unterworfen. Auch die Herstellung von Yeast Artificial Chromosomes (YACs) ist schwieriger, die erzielbare Länge beträgt nur einen Bruchteil dessen, was mit DNA ohne Basenungleichgewicht zu erreichen ist.

2. Ergebnisse

2.1. Das Dictyostelium Genomprojekt - eine internationale Anstrengung

Genomanalyse wird heute als Voraussetzung dafür gesehen, einen Organismus als experimentelles System weiter nutzen zu können (Kay and Williams, 1999). Aus diesem Grund bemühte sich die Dictyostelium Forschungsgemeinde in den USA und Europa um die Finanzierung eines Genomprojekts. Da die Analyse eines ganzen Genoms ein vergleichsweise großes Projekt ist, sollten mehrere Institute weltweit beteiligt werden. Im Frühjahr 1998 entschied die DFG, einen Teil der Genomanalyse zu finanzieren. Kurz darauf bewilligten auch das NIH und der MRC Mittel, so dass insgesamt 80 % der Genomanalyse von *D. discoideum* finanziert werden konnten. Da gleichzeitig an drei Stellen (IMB Jena, Abteilung Genomanalyse; Sanger Centre, Hinxton; Bayor College of Medicine, Huston) an der Genomanalyse gearbeitet werden sollte, mussten vorab die grundlegenden Vorgehensweisen abgesprochen werden.

Damit vergleichbare Daten von einem einzigen Genom entstehen, musste zunächst ein Konsens über den zu analysierenden Stamm erreicht werden. Im Falle von *D. discoideum* standen zwei etablierte Laborstämme (AX2 und AX4), die von einem seit über 50 Jahren verwendeten Stamm abstammten, zur Debatte. Obwohl beide Stämme einen gemeinsamen, bereits lange im Labor kultivierten Vorgänger hatten, unterscheiden sie sich doch physiologisch, und demnach auch genetisch. Da aber unklar war, wie groß diese Unterschiede sind, und außerdem für AX4 eine relativ hoch auflösende genetische und physikalische Karte existierte (Kuspa and Loomis, 1996), wurde dieser Stamm ausgewählt.

Da die Herstellung von Cosmiden oder Bacterial Artificial Chromosomes (BACs) wegen des hohen A+T Gehalts nicht möglich ist, musste eine alternative Strategie verfolgt werden. Die Idee, die YAC-Karte des Genoms von AX4 als Basis für eine YAC basierte Sequenzierung zu verwenden, wurde verworfen, da die Isolierung der YAC-DNA im größeren Maßstab aufwendig ist und Zweifel hinsichtlich des Grades der Degradierung und der Rearrangements dieser Klone bestanden. Obwohl heute die Durchführbarkeit einer Gesamtsequenzierung und Assemblierung größerer Genome oder Chromosomen bewiesen ist (Myers et al., 2000; Venter et al., 2001) stand sie zum Zeitpunkt der Diskussionen über das Projekt (1996/1997) noch im Mittelpunkt der Debatte. Deswegen wurde für dieses

Projekt eine gemischte Strategie angestrebt, wie sie auch für die wesentlich kleineren Chromosomen von *P. falciparum* angewandt wurde (Bowman et al., 1999; Gardner et al., 1998). Diese Strategie besteht aus dem so genannten YAC Skimming, bei dem wenige Sequenzen bis zu einer Abdeckung von bis zu 2 mal von einem YAC Klon hergestellt werden, und der chromosomenweisen Produktion von Sequenzen, die dann zur Vervollständigung der jeweiligen YAC Sequenz dienen. Zunächst werden kartierte YAC Klone ausgesucht, die einen so genannten „minimal tiling path" bilden. Die mit dem YAC skim produzierten Sequenzen dienen dazu, die aus den chromosomenspezifischen Klonbibliotheken stammenden Sequenzen zuzuordnen.

Bis auf die fast gleich großen Chromosomen 4 und 5 sind alle Chromosomen von *D. discoideum* auf einem PFGE Gel auftrennbar (Cox et al., 1990b). Die Aufreinigung der einzelnen Chromosomen wurde von Edward Cox von der Princeton University vorgenommen. Mit diesem Ausgangsmaterial war es Michael A. Quail vom Sanger Centre möglich, chromosomenspezifische Banken herzustellen, die allerdings einen hohen Anteil an Klonen von anderen Chromosomen enthielten (50 %). Da sich später die YAC Karten als nicht genau genug für die initiale Assemblierung herausstellten (Konfortov et al., 2000; Williams and Firtel, 2000), musste bei der Assemblierung auf die Unterstützung durch Sequenzen aus den YAC Klonen verzichtet werden. Die assemblierten Contigs konnten also nur mit Hilfe der Sequenzen aus den chromosomenspezifischen Banken zugeordnet werden. Um diese Assemblierung überprüfbar zu machen, begannen wir gleichzeitig, eine hochauflösende „Happy Mapping"Karte (Dear and Cook, 1993) zu erstellen.

Im Dictyostelium Genome Sequencing Consortium, das die nachfolgenden Arbeiten (Sequenzierung, Assemblierung und Annotation) übernehmen sollte, organisierten sich folgende Gruppen: Die Abteilung Genomanalyse des IMB Jena zusammen mit dem Institut für Biochemie I der Universität Köln, das Baylor College of Medicine, das Sanger Centre und der Medical Research Council (MRC). In regelmäßigen, ca. halbjährigen Abständen traf sich dieses Konsortium zur Abstimmung und Planung weiterer Arbeiten. Da mit chromosomenspezifischen Klonbibliotheken gearbeitet werden sollte, lag es nahe, die Arbeiten auch Chromosom für Chromosom zu verteilen (Abb. 3).

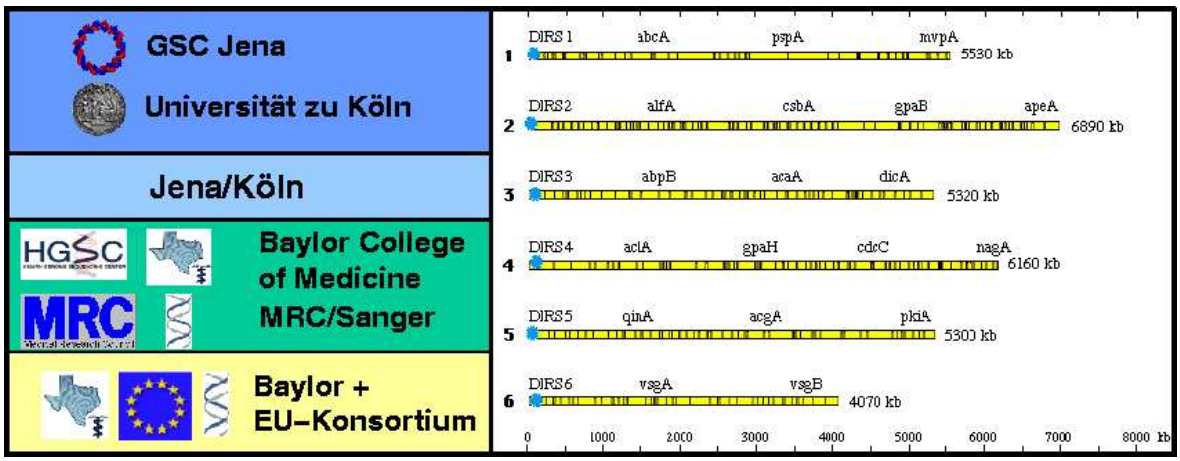


Abb. 3: Die Aufteilung der Chromosomen von *D. discoideum* an die Mitglieder des Sequenzierungskonsortiums. Für Chromosom 3 wurde erst im Jahr 2003 der Zuschlag an Jena gegeben. Das EU-Konsortium setzt sich zusammen aus: IMB Jena, Universität Köln, Institute Pasteur, Sanger Centre. Die verwendeten Chromosomenkarten stammen von: http://glamdring.ucsd.edu/others/dsmith/GenomicView.cgi.html.

Nicht nur die Sequenzierung sondern auch die nachfolgenden Arbeiten, nämlich Kartierung, Assemblierung und Annotation, sollten diesem Schema folgen. Damit aber vergleichbare Ergebnisse zustande kommen, sollten auch vergleichbare Analysemethoden in der Annotation angewandt werden. Jährliche Koordinierungstreffen sollten dies gewährleisten.

Weitere Gruppen außerhalb des Konsortiums wurden für unterstützende Arbeiten eingebunden. Die Gruppe von P. Dear am MRC in Cambridge erstellt Karten der Chromosomen, um die Assemblierung überprüfbar zu machen und die Contigs ordnen zu können. Für das Training eines Programmes zur Definition von Genmodellen in *D. discoideum* konnten wir die Gruppe um R. Guigo in Barcelona als Partner gewinnen.

2.2. Sequenzierung

Um einen Überblick über die zu erwartenden Schwierigkeiten bei der Sequenzierung zu bekommen, habe ich aus isolierten Kernen vegetativer Zellen eine Klonbibliothek der Gesamt DNA hergestellt. Damit die rDNA-Palindrom Kontamination möglichst niedrig wäre, wurden die Kerne in Agaroseblöcke eingegossen. Nach Abbau der Kernmembran mit Hilfe von Proteinase K wurde die DNA mittels PFGE aufgetrennt. Aus

der aufgereinigten DNA wurde eine Klonbibliothek hergestellt, aus deren Klonen knapp 50.000 Sequenzen generiert wurden. Da diese Klone nicht vorsortiert waren, sollten sie statistisch auf das gesamte Genom verteilt sein. Eine finale Analyse nach Fertigstellung des Genoms bestätigte diese Annahme. Im Laufe der Generierung der Sequenzen wurde die Sequenzqualität evaluiert und durch Anpassen der Sequenzreaktionsbedingungen an den hohen A+T Gehalt optimiert. Diese 50.000 Sequenzen waren auch die erste Ressource, die einen ersten Überblick über allgemeine Eigenschaften des Genoms ermöglichten (siehe 2.3.). Des Weiteren flossen sie natürlich als Ressource in die Assemblierung mit ein. Von weiteren Bibliotheken wurde dann der Hauptanteil der übrigen Sequenzen generiert (Tabelle 1).

Tabelle 1: Die Anteile der verschiedenen Klonbibliotheken an der Gesamtsequenzproduktion in Jena (Stand nach Beendigung des gesamten Projekts). Die Anzahl der Basen wurde nach Kürzung der Sequenzen auf Bereiche mit hoher Qualität bestimmt. Die eingeklammerten Kürzel geben die bibliotheksspezifische Benennung der einzelnen Sequenzen wieder.

Bibliothek	**Sequenzen**	**Basen**
Gesamt DNA (JAX*)	49044	18378672
Chromosom 1 (JC1*)	135882	52131018
Chromosom 2 (JC2*)	161260	45126849
Chromosom 3 (JC3*)	101642	38993459
Gesamt	**447828**	**154629998**
YAC shotgun	4964	1105794

Insgesamt wurde so eine rechnerische Abdeckung des gesamten Genoms von 4,55 mal erreicht. Dies entspricht in etwa der Hälfte der erforderlichen Anzahl an Sequenzen, die für die Assemblierung des gesamten Genoms nötig sind.

Im Rahmen des EU Konsortiums (Abb. 3) wurden zusätzlich fünf YAC Klone, die vom Sanger Centre ausgewählt worden waren, mit niedriger Abdeckung sequenziert (YAC shotgun, Tabelle 1). Diese sollten jeweils eine bestimmte Region von Chromosom 6 abdecken, dies traf jedoch nur für einen Klon zu (DY3850). Die übrigen YAC Klone wurden später auf anderen Chromosomen lokalisiert (Konfortov et al., 2000). Somit war

auch klar, dass die vorhandenen YAC-Karten der Chromosomen nur eingeschränkt als Rückgrat für die Assemblierung verwendet werden konnten.

Im internationalen Konsortium wurden weitere Sequenzen produziert, die aus für die Chromosomen 4, 5 und 6 spezifischen Klonbibliotheken stammten. Die Rohdaten aller Sequenzen wurden zwischen den drei Sequenzierungszentren ausgetauscht, um jedem Partner Zugriff auf alle Daten zu ermöglichen und die Assemblierung zu vereinfachen. Gegen Ende der Sequenzierphase im Jahre 2003 standen so insgesamt knapp eine Million Sequenzen vom gesamten Genom zur Verfügung.

2.3. Globalanalyse

Etwas weniger als 50.000 Sequenzen wurden aus einer Klonbibliothek des Gesamtgenoms hergestellt, um allgemeine Eigenschaften des Genoms untersuchen zu können. Weitere Sequenzen aus der Chromosom 2 spezifischen Klonbibliothek wurden unter Einbeziehung eines Korrekturfaktors zum Ausgleich der gewichteten Verteilung der Einzelsequenzen zu dieser Analyse herangezogen. Diese Ressource diente sowohl dazu, Basenzusammensetzung und Basenungleichgewichte zu untersuchen, als auch einen ersten Katalog der Gene und Komplexen Repetitiven Elemente (KRE) von *D. discoideum* zu erstellen. Es zeigte sich, dass mit einer relativ niedrigen Abdeckung aussagekräftige Schlussfolgerungen über das Gesamtgenom getroffen werden konnten, die sich später auch am vollständigen Genom bestätigen ließen.

2.3.1. Basenungleichgewichte und Kodonnutzung

Die Analyse der ersten 3000 Sequenzen ergab einen A+T Gehalt von 74 ± 7 %. Der Mittelwert lag damit deutlich unter dem für das Genom berichteten von etwa 78 %. Die Schwankungsbreite des A+T Gehalts bei dieser Berechnung indizierte aber auch, dass die Analyse ganzer Contigs besseren Aufschluss über die Basenzusammensetzung des Genoms geben würde. Die observierte Differenz bei Betrachtung einzelner Sequenzen ist der ungleichen Klonier- und Sequenzierbarkeit von unterschiedlich A+T reichen Sequenzen geschuldet. Erst die Analyse von assemblierten Contigs größer 3 kb ergab einen Wert von

77.8 %. Wie der Vergleich mit dem fertiggestellten Genom dann ergab, entsprach dieser Wert genau dem Gesamt A+T Gehalt. Obwohl die Zahl der Lücken, und damit die Länge der unbekannten Sequenz unbekannt war, konnten also anhand von Contigs ausreichender Länge Aussagen für das gesamte Genom getroffen werden.

Um über die Basenanteile des Genoms hinaus Aussagen treffen zu können, muss die Analyse von Sequenzmotiven mit einbezogen werden. Tupel ist ein Begriff aus der Mathematik. Er bezeichnet eine geordnete Zusammenstellung von Objekten, bei der im Gegensatz zu Mengen eine Reihenfolge festgelegt ist. Dieser Begriff kann also hervorragend zur Beschreibung von Sequenzabschnitten herangezogen werden. Eine Analyse der Sequenztupel ergab, dass, wenig überraschend, Tupel mit hohen A+T Anteilen überwiegen. Tabelle 2 zeigt die dreißig 6er Tupel mit den größten Häufigkeiten.

Tabelle 2: Verteilung von 6er Tupeln im Genom von *D. discoideum*. Die Berechnung erfolgte auf Basis von über 350.000 Einzelsequenzen, deren Bereiche niederer Qualität entfernt wurden.

Tupel	Anzahl	Häufigkeit
TTTTTT	393234	10.33%
AATAAT	113478	2.98%
TAATAA	102712	2.70%
ATAATA	96338	2.53%
AAATAA	87717	2.31%
AATAAA	85262	2.24%
ATTTTT	84637	2.22%
TTTTTA	84135	2.21%
TATTTT	73189	1.92%
ATAAAA	69984	1.84%
TTTTAA	62953	1.65%
AAAATT	56560	1.49%
AATTAA	50923	1.34%
TTTAAT	50772	1.33%
TAATTT	48908	1.29%
TAAAAT	46221	1.21%
ATAAAT	43576	1.15%
ATTTAA	43537	1.14%
TATTTA	40863	1.07%
TAAATT	40038	1.05%
ATAATT	38429	1.01%
TATTAA	35018	0.92%
AAATAT	34925	0.92%
ATATTT	34925	0.92%
TTGTTG	32385	0.85%
TTTAAA	30946	0.81%
TAATAT	28719	0.75%
TTTATA	28204	0.74%
TGTTGT	26792	0.70%
AAAGAA	26558	0.70%

In Abhängigkeit von der Gesamtbasenzusammensetzung des Genoms läge die zu erwartende Häufigkeit eines reinen AT-Tupels bei ca. 0,7 % während ein Tupel aus 4 A/T

und 2 G/C nur mit einer zu erwartenden Häufigkeit von 0,05 % auftritt. Bemerkenswert ist deshalb, dass ein Tupel (TTGTTG) mit einer stark erhöhten Häufigkeit auftritt. UUG kodiert für Leucin (L), ein weiteres für Leucin kodierendes Kodon (UUA) tritt mit 2.7 % ebenfalls stark gehäuft auf. Eine Analyse der Genmodelle und der Kodierungstabelle für *D. discoideum* im Verhältnis zu anderen Eukaryontentabellen zeigt jedoch, dass der Anteil an Leucinen in den Proteinen kaum erhöht ist. Eine genauere Analyse der Kodonnutzung aller Genmodelle von Chromosom 2 und weiterer, durch vorhandene mRNAs gesicherter Gene (Tabelle 3A) zeigte, dass vor allem Asparagin (N), Glutamin (Q) und Isoleucin (I) statistisch signifikant eine Sonderrolle in Proteinen spielen. Einige der zugehörigen Triplet-Kodonen (AAU, AAC (N); CAA (Q), AUU, AUA (I)) treten auch häufiger als bei einer Zufallsverteilung zu erwarten wäre auf. Die genauen Daten zu dieser Analyse können im Supplement von (Eichinger et al., 2005) nachgelesen werden.

Tabelle 3 (nächste Seite)**:** Kodonnutzung von *D. discoideum* (A) und *P. falciparum* (B). Die Genmodelle von *D. discoideum* wurden mit Hilfe von GeneID erstellt (siehe 2.6.2.), *P. falciparum* Daten wurden von http://www.kazusa.org.jp/codon entnommen. Die einzelnen Zellen enthalten: Kodon, Anteil des Kodons pro 1000 Kodonen, Gesamtzahl der Kodonen in der untersuchten Menge. Unterhalb der Teiltabellen ist der G+C Anteil in den jeweiligen Kodonpositionen angegeben.

A

UUU 31.0 (60604)	UCU 15.8 (30809)	UAU 28.9 (56399)	UGU 12.9 (25282)
UUC 14.4 (28118)	UCC 4.2 (8109)	UAC 5.3 (10387)	UGC 1.5 (2910)
UUA 55.0(107369)	UCA 50.4 (98474)	UAA 1.5 (3013)	UGA 0.1 (161)
UUG 11.5 (22523)	UCG 2.4 (4712)	UAG 0.1 (174)	UGG 7.6 (14851)
CUU 9.8 (19182)	CCU 5.8 (11338)	CAU 15.4 (30006)	CGU 7.6 (14788)
CUC 3.8 (7496)	CCC 1.0 (2027)	CAC 2.5 (4807)	CGC 0.1 (142)
CUA 4.7 (9129)	CCA 34.1 (66541)	CAA 49.6 (96820)	CGA 0.5 (945)
CUG 0.3 (576)	CCG 0.5 (984)	CAG 1.7 (3415)	CGG 0.1 (110)
AUU 50.4 (98463)	ACU 21.8 (42566)	AAU 94.5(184652)	AGU 22.5 (44043)
AUC 11.7 (22893)	ACC 8.5 (16521)	AAC 10.8 (21088)	AGC 2.4 (4757)
AUA 18.8 (36701)	ACA 29.3 (57288)	AAA 60.6(118474)	AGA 19.8 (38719)
AUG 16.6 (32476)	ACG 1.0 (1948)	AAG 12.0 (23377)	AGG 1.3 (2484)
GUU 25.7 (50176)	GCU 12.0 (23356)	GAU 47.0 (91803)	GGU 36.9 (72094)
GUC 3.9 (7659)	GCC 4.2 (8294)	GAC 4.3 (8401)	GGC 2.2 (4285)
GUA 13.7 (26675)	GCA 17.9 (34889)	GAA 48.3 (94436)	GGA 8.8 (17116)
GUG 2.3 (4583)	GCG 0.6 (1175)	GAG 9.3 (18165)	GGG 0.9 (1696)

Kodierend GC 28.63% 1. Position GC 37.53% 2. Position GC 33.45% 3. Position GC 14.90%

B

UUU 31.2 (20097)	UCU 17.0 (10951)	UAU 38.6 (24890)	UGU 16.4 (10575)
UUC 7.3 (4704)	UCC 6.0 (3879)	UAC 4.6 (2942)	UGC 2.4 (1516)
UUA 49.9 (32202)	UCA 19.7 (12720)	UAA 1.2 (785)	UGA 0.2 (104)
UUG 9.9 (6364)	UCG 2.4 (1551)	UAG 0.3 (184)	UGG 5.3 (3414)
CUU 8.8 (5659)	CCU 10.5 (6784)	CAU 17.5 (11291)	CGU 3.4 (2179)
CUC 1.5 (983)	CCC 3.0 (1930)	CAC 3.8 (2420)	CGC 0.3 (200)
CUA 5.3 (3422)	CCA 18.4 (11887)	CAA 25.8 (16640)	CGA 1.9 (1248)
CUG 1.0 (672)	CCG 0.9 (565)	CAG 3.1 (2020)	CGG 0.2 (106)
AUU 32.2 (20776)	ACU 14.2 (9169)	AAU 97.8 (63109)	AGU 22.0 (14219)
AUC 5.6 (3590)	ACC 5.4 (3500)	AAC 19.3 (12438)	AGC 3.7 (2379)
AUA 40.8 (26317)	ACA 21.8 (14085)	AAA 84.8 (54695)	AGA 16.9 (10923)
AUG 18.6 (11985)	ACG 3.4 (2191)	AAG 17.8 (11496)	AGG 3.3 (2111)
GUU 20.7 (13351)	GCU 14.8 (9572)	GAU 54.8 (35359)	GGU 17.9 (11514)
GUC 2.7 (1714)	GCC 4.1 (2650)	GAC 8.7 (5630)	GGC 1.9 (1233)
GUA 19.3 (12459)	GCA 17.5 (11268)	GAA 74.1 (47824)	GGA 20.1 (12947)
GUG 3.7 (2394)	GCG 1.3 (852)	GAG 9.9 (6408)	GGG 3.0 (1944)

Kodierend GC 27.46% 1. Position GC 38.00% 2. Position GC 27.93% 3. Position GC 16.43%

Die 3. Position ist stark hinsichtlich der Nutzung von A oder U selektioniert. Unter Ausnutzung der Wobble-Möglichkeiten wird damit der A+T Anteil innerhalb der kodierenden Genomanteile erhöht. Auffällig ist, dass *P. falciparum* mit etwas höherem A+T Anteil im Genom (Tabelle 3B) im Vergleich wesentlich stärker A+T Basen an Position 2 nutzt als *D. dicoideum*.

Besonders häufig werden AAU und AAC Triplets in Proteinsequenz übersetzt. Das führt zu einem stark erhöhten Anteil an N Homopolymeren, die auch im Vergleich zu anderen Organismen, selbst mit ähnlichem A+T Gehalt wie *P. falciparum*, eine beträchtliche Länge erreichen (Abb. 4). Auch andere Aminosäuren formen Homopolymerregionen, z.B. Serin. Hier werden aber alle 6 möglichen Kodonen verwendet, während im Falle der Homopolymere, die für Asparagin bzw. Glutamin kodieren, meistens nur ein einziges Kodon auftritt. Dies legt nahe, dass diese Regionen durch Amplifizierung des jeweiligen Kodons entstanden sind und noch wenig Möglichkeiten der Akkumulation von Mutationen bestand. Das Basenungleichgewicht hin zu einem erhöhten A+T Gehalt des gesamten Genoms spiegelt sich also bei den kodierenden Regionen nicht nur in einer veränderten Kodon-Nutzung (die dritte Position enthält vorzugsweise A oder T), sondern auch in der spezifischen Nutzung bestimmter Aminosäuren als wahrscheinlich nicht-funktionale Sequenzabschnitte in den Proteinen wieder.

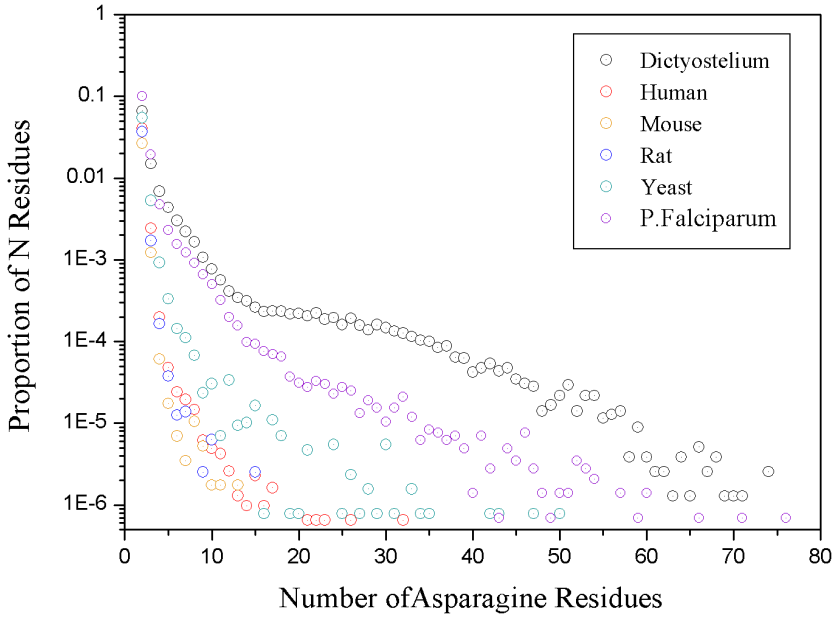


Abbildung 4: Asparagin Homopolymerregionen in verschiedenen Eukaryonten. Gezeigt ist die Länge der Homopolymerregionen im Verhältnis zur Anzahl der Aminosäuren, die sich in solchen Sequenzabschnitten befinden.

Bei der Analyse von *P. falciparum* Proteinen wurde festgestellt, dass diese vermehrt Homopolymerstücke aus gleichen Aminosäuren enthalten. Diese Abschnitte tragen zur Funktionalität des Proteins nichts bei. Vielmehr scheinen sie beim nativen Protein als Schleifen aus dem Protein heraus zu hängen und die Faltung zur Sekundär- und Tertiärstruktur nicht zu beeinflussen (Verra and Hughes, 1999). Möglicherweise ist die Einfügung von vielen gleichförmigen A+T reichen Kodonen in Gene ein Weg, den Gesamtanteil von A+T am Genom zu erhöhen, ohne die Gendichte zu stark herabzusetzen. Eine Erklärungsmöglichkeit dafür, warum eine hohe Gendichte anzustreben wäre, wäre z.B., weniger Raum für transposable Elemente zum Inserieren zu bieten. Generell ist jedoch nicht geklärt, welche Determinanten Genomgrößen, und damit das sogenannte C-Value Paradoxon, bestimmen.

Intergenische Regionen zeichnen sich bei *D. discoideum* durch einen sehr hohen A+T Anteil von bis über 90 % aus. Wie wir in einem Vergleich mit *P. falciparum* nachgewiesen haben (Szafranski et al., 2005), können offensichtlich diese Regionen nicht noch höhere A+T Anteile akquirieren. Damit würde mit der Einführung von sogenannten „homopolymer runs“ eine Steigerung des bereits hohen A+T Gehalts erreicht, ohne intergenische Regionen ausweiten oder im A+T Anteil erhöhen zu müssen. Wir sehen also bei zwei Organismen mit ähnlich hohem Anteil an A+T im Genom etwas divergierende Stratgien, dieses Basenungleichbewicht zu erreichen.

2.3.2. Komplexe Repetitive Elemente

Repetitive DNA kommt in allen Organismen vor. Sie kann unterteilt werden in einfache Sequenzwiederholungen (simple repeats), die auch die häufigen Tupel enthalten können (2.3.1.) und Komplexe Repetitive Elemente (KREs), die aus mehreren verschiedenen Einheiten aufgebaut sind, die teilweise kodierendes Potential besitzen. In der Regel haben KREs das Potential zu springen und sich zu vermehren, sind also Transposonen. Diese Eigenschaft macht sie zu einer die Evolution vorantreibenden Kraft, da jedes Sprungereignis eine Mutation darstellt, die potentiell schädlich oder nützlich ist. Man kann je nach Intermediärform während der Transposition zwischen RNA- und DNA-Transposonen unterscheiden. RNA Transposonen können an den Enden noch einen „Long Terminal Repeat“ (LTR) besitzen und werden demnach jeweils als LTR oder non-LTR

Tranposonen klassifiziert. Die kodierten Proteine (in der Regel 2) weisen über alle Organismengruppen hinweg ähnliche Domänen auf. Die Gruppe der DNA Tranposonen ist wesentlich heterogener, ihr Kodierungspotential muss keine Ähnlichkeit zu anderen DNA Transposonen zeigen. Dementsprechend sind sie über Ähnlichkeiten zu Transposonen aus anderen Organismen kaum zu identifizieren. Darüber hinaus gibt es noch kleinere Elemente wie z.B. MITEs oder „foldback elements“ (Rebatchouk and Narita, 1997; Surzycki and Belknap, 1999), die meist kein Kodierungspotential besitzen. Man vermutet, dass sie über Mechanismen, die nur zelluläre Komponenten benötigen oder mit Hilfe von Proteinen aus anderen Transposonen dennoch springen und sich vermehren können (Izsvak et al., 1999).

Während simple repeats in jedem Organismus einfach zu erkennen sind, muss der Gehalt an KRE durch eingehende Analyse geklärt werden. Zwar kann es auch zwischen KREs verschiedener Organismen Gemeinsamkeiten geben, diese können aber meist nur auf der Ebene der kodierten Proteine gefunden werden. Dies gilt zumal, wenn die Basenzusammensetzung des Organismus extrem ungleich gewichtet ist wie bei *D. discoideum*. Etliche KREs, auch ungewöhnlicher Natur wie z.B. DIRS-1, waren schon vor Beginn des Genomprojekts bekannt (Cappello et al., 1985; Leng et al., 1998; Marschalek et al., 1990). Weitere Elemente aus den selben Elementfamilien konnten wir mit Hilfe relativ weniger weiterer genomischer Sequenzen durch Ähnlichkeitsanalysen finden (Szafranski et al., 1999).

Da repetitive Elemente ein Hindernis für die Assemblierung darstellen können, sollten vor der Zusammensetzung aller Sequenzen zum Genom diese Elemente definiert sein. Eine Datenbank aller Elemente kann dann zur Maskierung entsprechender Sequenzen verwendet werden, damit sie im Assemblierungsprozess zunächst ausgeschlossen werden können. Mit Erreichung einer Genomabdeckung von ca. 0.5 mal war gewährleistet, dass aus den akkumulierten Sequenzdaten die Konsensussequenzen auch seltenerer Elemente abgeleitet werden konnten. Auch hier wurden neue KREs zunächst über eine Analyse der Ähnlichkeit zu schon vorhandenen Elementen definiert.

Zusätzlich wurden Sequenzen untersucht, die gehäuft im Datensatz auftraten. Nach Ausschluss von Genfamilien aufgrund von Ähnlichkeiten zu vorhandenen Genen in den öffentlich zugänglichen Datenbanken wurden die übrigen gehäuft auftretenden Sequenzabschnitte näher untersucht. Weitere Sequenzen konnten so noch unbekannten Genfamilien zugewiesen werden, da nur Homologien in offensichtlich kodierenden Abschnitten gefunden wurden. Die genaue Struktur aller Sequenzabschnitte, die auch Häufungen in nicht kodierenden Bereichen aufwiesen, wurde dann bestimmt. In Fällen, in

denen das Element nur selten im Datensatz vorhanden war, wurde zusätzliche Sequenzinformation durch „Primer walking“ auf vorhandenen bakteriellen Klonen gewonnen. Auf diese Weise konnten alle Elemente, die häufiger als 5 mal im Genom vorkommen, bestimmt werden. Die detaillierte Analyse des Gesamt- KRE Gehalts von *D. discoideum* kann in (Glöckner et al., 2001) nachgelesen werden. Insgesamt wurden neue Elemente sowohl der RNA Transposonen als auch der DNA Transposonen gefunden. Zusätzlich konnte ein kleineres repetitives Element (thug) beschrieben werden. Sicherlich handelt es sich bei diesem Element um ein Analogon zu kleineren Elementen in anderen Organismen. Solche Elemente sind sehr unterschiedlich in verschiedenen Organismen, weswegen deren Eingruppierung unklar bleibt. Die Analyse ergab des Weiteren, dass der Gesamtgehalt an KREs im *D. discoideum* Genom um 9,7 % liegt. Im Vergleich zu den sequenzierten Spezies aus dem Bereich der niederen Eukaryonten (*S. cerevisiae*: 1 %; *C. elegans*: 6 %) ist dies ein sehr hoher Wert. Er erreicht sogar den Anteil an der Gesamt-DNA, der im Genom von *A. thaliana* festgestellt wurde (10 %) (The Arabidopsis Sequencing Initiative, 2000).

Für die Abschätzung der möglichen Schwierigkeiten bei einer Assemblierung sind die Polymorphismen sowie der Anteil an individuell unvollständigen, trunkierten, Elementen innerhalb einer Elementfamilie entscheidend. Sehr heterogene Elementfamilien sowie kurze Elemente stellen in der Regel keine Schwierigkeit im Assemblierungsprozess dar, da die Sequenzen den Elementen aufgrund der Polymorphismen bzw. der eindeutigen Randsequenzen zweifelsfrei zugeordnet werden können. Die verschiedenen Elementfamilien im *D. discoideum* Genom weisen eine weite Streuung im Anteil an Polymorphismen auf. Vor allem die am häufigsten vertretenen Familien DIRS-1, TRE3-A und TRE5-A weisen niedrige Polymorphismenraten und niedrige Fragmentierungsindizes, d.h. wenige trunkierte Elemente auf (Glöckner et al., 2001). Da diese Elemente mit je über 5 kb auch länger sind als die längsten verfügbaren Inserts (bis 4,5 kb) einer sequenzierbaren Klonbibliothek, erscheint es aussichtslos, die einzelnen Mitglieder dieser Elementfamilien erfolgreich zusammenzusetzen.

Die Analyse der Trunkierungsstellen ergab außerdem, dass sehr viele Elemente an Orten inseriert sind, die schon von anderen Elementen besetzt waren. Oft handelt es sich dabei um Elemente der gleichen Art. Möglicherweise kommen solche aus gleichen oder ähnlichen Transposonen zusammengesetzten Loci durch das gleichzeitige Springen und Vermehren dieser Transposonen zustande. Die Häufung von Elementen gleicher Art gilt

nicht für die TRE Elementfamilien, die je nach Insertionspräferenz nur vor oder hinter tRNA- Genen inserieren (Abb. 5).

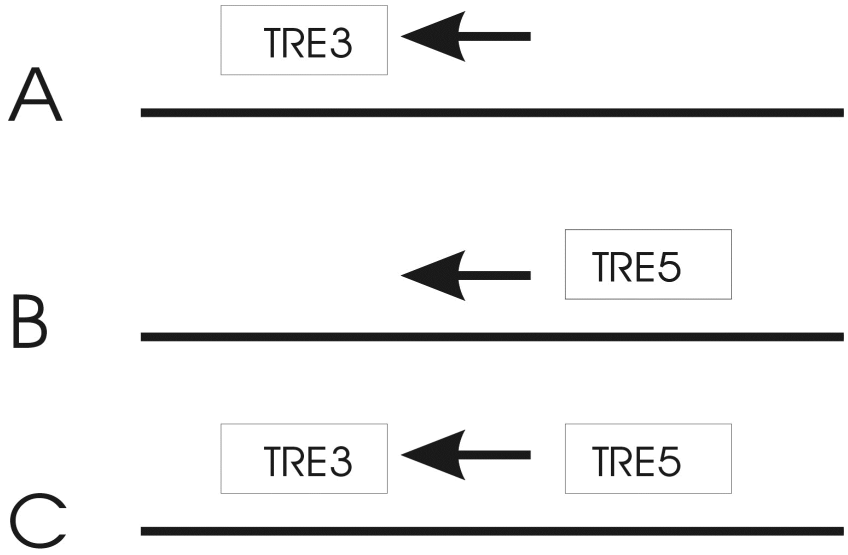


Abb. 5: Schematische Darstellung der möglichen Insertionen von TRE-Elementen vor oder hinter tRNA Genen. Die Pfeile repräsentieren die tRNA und deren Ableserichtung. Im Fall C kann die Lage der tRNA nur vermutet werden, die Region kann nicht assembliert werden.

Bei diesen Familien kann es zu Tandemanordnungen auf beiden Seiten einer tRNA kommen (Abb. 5 C). Es besteht auch die Möglichkeit, dass ein einziges tRNA Gen mehrmals für eine Insertion durch TRE Elemente benutzt wird. Das würde sich darin äußern, dass man in der Shotgundatenbank Sequenzstücke findet, die Bereiche von zwei aneinander stoßenden TRE Elementen enthalten. In der Tat lassen sich solche Sequenzbereiche assemblieren, diese mehrmalige Insertion scheint aber nur in etwa 5 % aller TRE Loci der Fall zu sein. Dagegen sind einige tRNA Gene nicht mit TRE Elementen assoziiert. Öfter findet man jedoch Mitglieder anderer Elementfamilien, die in TRE-Elemente integriert sind.

Die Analyse der KREs wurde dann dazu benutzt, eine Datenbank aller repetitiven Elemente zu bilden. Diese wurde eingesetzt, um alle produzierten Sequenzen auf repetitive Anteile zu testen und gegebenenfalls zu markieren. Die Markierung half dann, diese Sequenzanteile aus der Assemblierung auszuschließen.

Die Untersuchung der Verteilungsmuster der einzelnen KREs innerhalb der Chromsomen ergab, dass es eindeutig präferierte Regionen für die einzelnen Elemente gibt. So sind DIRS Elemente nur in einem riesigen Cluster jeweils an einem Chromosomenende anzutreffen, der Rest jeden Chromosoms, bis auf eine Ausnahme, ist frei davon. Nur auf Chromosom 2 findet sich mitten im Chromosom ein zweites Cluster. Da diese Region allen Chromosomen gemeinsam ist, könnte sie dem Centromer entsprechen. Ein Erklärungsversuch für das zweite Cluster auf Chromosom 2 wird in 2.6.1 gegeben. TRE Elemente sind nur in der Nähe von tRNA Genen anzutreffen, da sie diese als „Landeplatz" nutzen. Obwohl die weiteren Elemente nicht einer solchen Einschränkung unterliegen, sind sie doch bevorzugt in den Centromer-DIRS-Clustern anzutreffen.

2.3.3. Genkatalog

Ein Genkatalog stellt die Gesamtheit aller Gene eines Organismus dar. Wenn eine Genomsequenz vollständig aufgeklärt ist, kann unter Nutzung von Genvorhersageprogrammen ein solcher Genkatalog aufgebaut werden. Dennoch unterliegt ein solcher Katalog ständigen Veränderungen, da sich einerseits die Genvorhersagen durch die Verbesserung von Programmen ändern, andererseits die manuelle Annotation und damit verbundene weitere Experimente zu einzelnen Genen die Beschreibung von Genen und Genprodukten verbessert. Partielle Genkataloge können durch die Herstellung genügender Mengen von ESTs (Expressed Sequence Tags) oder genomischer Einzelsequenzen erstellt werden. Der partielle Charakter dieser Kataloge beruht auf unterschiedlichen Limitationen. Mit Hilfe von ESTs können nur exprimierte Gene detektiert werden, die Vollständigkeit eines solchen Katalogs hängt von der Qualität und der Vielfältigkeit der verwendeten cDNA Bibliotheken ab. Dennoch wird sich ein solcher Katalog nur der Vollständigkeit nähern können, da niemals alle möglichen Situationen, für die eine Zelle gerüstet sein muss, simuliert werden können. Ein aus Einzelsequenzen entwickelter Genkatalog kann nur Informationen über Gene liefern, die orthologe oder paraloge Gene in anderen Organismen besitzen oder gut konservierte funktionelle Domänen enthalten. Gene, die für den jeweiligen Organismus oder für die Gattung spezifisch sind, bleiben unentdeckt. Darüber hinaus ist die Kalkulation der Trefferwahrscheinlichkeit für ein bestimmtes Gen von Bedeutung. Sie gibt an, mit welcher Wahrscheinlichkeit ein Gen im Datenpool enthalten ist. Die dafür verwendete Formel folgt einer Binomialverteilung und lautet:

$$\sum_{h=1}^{\infty} P_{(h)} = \frac{N!}{h!(N-h)!} p^h (q)^{N-h}$$

wobei N die Gesamtzahl aller Sequenzen, h die Anzahl der Kopien ist. Es wird summiert über die Anzahlen 1 bis unendlich. p ist die Trefferwahrscheinlichkeit für die Ziehung einer bestimmten Sequenz, q die Wahrscheinlichkeit, bei einmaligen Ziehen das gesuchte Gen nicht zu erhalten.

Mit Hilfe dieser Formel wurde die Wahrscheinlichkeit, ein bestimmtes Gen (oder einen bestimmten Genabschnitt) in knapp 87.000 Sequenzen einer Gesamtgenom-Klonbibliothek zu finden, mit ca. 90 % bestimmt. Selbstverständlich wird dabei von einer Gleichverteilung der Sequenzen des Datenpools über das gesamte Genom ausgegangen, d.h. es besteht keine Limitierung bei der Klonierbarkeit bestimmter Sequenzen. Das Ungleichgewicht hin zu einer höheren Abdeckung von Chromosom 2 bedingt durch Sequenzen aus der Chromosom 2 spezifischen Bibliothek wurde entsprechend herausgerechnet. Alle Sequenzen wurden gegen die Swissprot Datenbank annotierter Proteine abgeglichen. Dabei wurden 4123 singuläre Treffer mit einem Wahrscheinlichkeitswert kleiner als $p=10^{-10}$ registriert. Wenn man davon ausgeht, dass *D. discoideum* wie alle anderen bis jetzt analysierten Organismen 40 - 50 % organismenspezifische Gene enthält, kann die Gesamtzahl aller *D. discoideum* Gene anhand dieser Zahlen auf 9000 bis 12000 geschätzt werden. *D. discoideum* enthält somit deutlich mehr Gene als der einzellige Eukaryont *S. cerevisiae* mit ca. 6000 Genen, liegt jedoch unterhalb der Zahlen echter Vielzeller wie *D. melanogaster* mit etwas über 13.000 und *C. elegans* mit 19.000 Genen (Adams et al., 2000; The C. elegans Sequencing Consortium, 1998). Die Anzahl der Gene in einem Genom spiegelt somit in gewisser Weise die Komplexität eines Organismus wieder. Eingeschränkt wird diese Schlussfolgerung allerdings durch die Tatsache, dass verschiedene Spezies verschiedene Gen- und Genomanteile amplifiziert haben. Solche Duplikationen, zumal wenn sie sich erst kürzlich in der Evolution ereignet haben, können mit Hilfe der angewandten Methode nicht entdeckt werden. Falls das Genom von *D. discoideum* größere oder mehr Duplikationen aufwiese als das anderer Organismen, müsste man von einer niedrigeren durchschnittlichen Genzahl am

Übergang zur Vielzelligkeit ausgehen. Jedoch kann die erwartete Kopienzahl eines bestimmten Genes aus der Anzahl der getroffenen Sequenzen errechnet werden. Sie ergibt sich aus:

$$\text{Kopienzahl} = \frac{\text{Trefferanzahl} \times \text{durchschnittliche Leselänge einer Sequenz}}{\text{Genlänge} \times \text{Genomabdeckung}}$$

Sowohl die Kopienzahlen bekannter Zytoskelettgene (Tabelle 4) als auch der Komplexen Repetitiven Elemente konnten mit hinreichender Genauigkeit aus der Anzahl an Sequenzen, die ähnliche Motive enthalten, bestimmt werden (Glöckner et al., 2001). Je größer die zu erwartende Anzahl an Genen ist, desto größer wird allerdings auch der Unsicherheitsfaktor, zufällige und gerichtete Häufungen (aufgrund der Klonierbarkeit spezifischer Sequenzen) haben einen großen Einfluss. Eine spätere Assemblierung und Vervollständigung aller Aktin-Gene ergab z.B. eine Anzahl von 27 Stück. Die berechneten Werte für die Kopienzahl dieser Genfamilie liegen weit höher. Dieses Beispiel zeigt, dass die Berechnung nur eine Aussage über den ungefähren Bereich der zu erwartenden Kopien im Genom erlaubt.

Tabelle 4: Berechnete Häufigkeiten von ausgewählten Zytoskelettgenen im Genom. Grundlage der Analyse waren 87.000 Sequenzen. Die Accession Nummer gibt den GenBank Eintrag an, der für die BLAST Analyse verwendet wurde.

Accession Nummer	Länge des Proteins	Trefferzahl	Gen	berechnete Kopienzahl (gerundet)
A25084	376	72	Actin	44
A23562	137	4	Cofilin	7
A23750	185	5	Comitin	6
A25084	118	2	Histactophilin I	4
A26655	124	2	Profilin II	4
A28517	610	8	Fimbrin	3
A31642	1111	12	Myosin heavy chain IB	2
A33284	959	10	Protovillin	2
A34400	117	1	Histactophilin II	2
A37098	1738	14	Interaptin	2
A38682	126	1	Profilin I	2
A43358	1113	8	Myosin heavy chain ID	2
A47106	418	3	Actin related protein (ARP3)	2
A48120	857	6	ABP120 gelation factor	2
A54818	143	1	Ponticulin	2
A57036	441	3	Cortexillin II	2
A61042	456	3	EF1 Alpha 50 KD Actin-binding protein	2
ACTZ_HUMAN,	2116	13	Myosin heavy chain	1
B61042	2491	14	Talin	1
DDP17	994	5	Myosin heavy chain IA	1
EF1A_DICDI,	444	2	Cortexillin I	1
FADO1	445	2	Coronin	1

Des Weiteren wurde mit Hilfe der akkumulierten Sequenzen getestet, ob wir alle Rho-ähnlichen Proteine in *D. discoideum* finden könnten. Bei dieser Analyse wurde zusätzlich zu den bereits bekannten ein weiteres Gen und ein Pseudogen entdeckt. Die Struktur aller dieser Gene wurde nach weiteren gezielten Sequenzierungen der entsprechenden Klone aufgeklärt (Rivero et al., 2001). Des Weiteren haben wir Kinesine und ABC-Transporter untersucht (Anjard et al., 2002; Kollmar and Glöckner, 2003). Diese Analyse zeigte, dass die Datenbasis, die mit der gesamtgenomischen Bibliothek geschaffen

wurde, ausreichend war, um gezielt Genfamilien beschreiben zu können. Da für derartige Analysen nur ungefähr 2700 Sequenzen/1MB nötig sind, könnte eine hinreichend genaue Beschreibung des Genoms anderer Organismen auf dieser Basis durchgeführt werden.

2.4. Assemblierungsstrategien

Das generelle Problem bei der Assemblierung eines Chromosoms ist die richtige Zuordnung der Sequenzen. Da die chromosomenspezifischen Klonbanken von *D. discoideum* nur zu 50 % „richtige" Sequenzen enthielten, war es nötig, nicht zum Chromosom gehörende Sequenzen und Contigs aus der Assemblierung und/oder der späteren Bearbeitung auszuschließen. Bisherige Methoden zur Assemblierung von Chromosomen oder auch ganzen Genomen bezogen Informationen von YAC- oder BAC-Klonen (Camargo et al., 1997; Tanaka et al., 1995; Thompson et al., 1999) mit ein. Dies ermöglichte eine chromosomen- oder genomweite Sequenzproduktion gepaart mit einer lokalen Assemblierung von Sequenzen, die durch die Klone einander zugeordnet waren. Im Fall von *D. melanogaster* und *Homo sapiens* wurden die Klonressourcen zur lokalen Überprüfung der globalen Assemblierung hinzugezogen (Myers et al., 2000; Venter et al., 2001). Beide Möglichkeiten entfielen für das hier diskutierte Projekt, da keine Klonbibliotheken mit großen Insertlängen hergestellt werden konnten und die vorhandene YAC Karte zu viele Fehler aufwies. Die einzigen Anhaltspunkte für die *ab initio* Zuordnung einzelner Sequenzen waren die auf die verschiedenen Chromosomen kartierten Gene. Für Chromosom 2 standen nur Sequenzen von 49 potentiell richtig kartierten Genen zur Verfügung. Mit Hilfe dieser Information konnten also nur wenige Contigs bzw. Sequenzen zugeordnet werden. Eine weitere Möglichkeit, Contigs bestimmten Chromosomen zuzuordnen, bestand in der Analyse ihrer Zusammensetzung aus Sequenzen von Klonen aus verschiedenen chromosomenspezifischen Klonbibliotheken. Bei geringer Größe oder Abdeckung innerhalb der Contigs ist eine eindeutige Bestimmung jedoch nicht möglich, da die geringe Zahl der zur Berechnung herangezogenen Sequenzen zu sehr großen statistischen Schwankungen führt.

Alle drei Chromosomen, die in Jena bearbeitet wurden, wurden mit Hilfe der im Folgenden skizzierten Technik assembliert. Mit der Assemblierung wurde schon während der Sequenzproduktion begonnen, im Gegensatz zu den üblichen Methoden wurde also eine

kontinuierliche Ausweitung der Contigs angestrebt. Die Vorteile liegen auf der Hand: 1. Frühe Assemblierung erlaubt frühe Datenanalyse. 2. Repetierte Regionen können analysiert und markiert, bzw. maskiert werden. 3. Fehlende Sequenzpaarpartner können schnell nachproduziert werden, wenn diese weitere Informationen über die Anordnung von Contigs und die Ausweitung der assemblierten Contigs versprechen.

Im einzelnen bestand das Procedere bei der Assemblierung aus folgenden Schritten (Abb. 6): Die initial mit Hilfe der kartierten Gene gebildeten Contigs waren die Keime der Assemblierung. Nur diese Contigs konnten relativ sicher dem Chromosom zugeordnet werden (1). Neu produzierte Sequenzen wurden gegen die vorhandenen Contigs durch Ähnlichkeitssuche (Blast (Altschul et al., 1990)) abgeglichen und automatisch zur Erhöhung der Abdeckung und Verlängerung der vorhandenen Contigs eingesetzt (2). Nicht passende Sequenzen und Fehlassemblierungen wurden aus der Datenbank wieder entfernt (3). Die Bildung von Gerüsten (Scaffolds), die über Sequenzpaare miteinander verbunden sind, wurde durch die Hinzufügung der Gegensequenzen von vorhandenen Klonen (read pairs) für alle Contigs mit mehr als 1800 Basen erreicht. Diese zunächst singulären Sequenzen dienten dem Aufbau weiterer Contigs. Hier begann der Zyklus neu. Später wurden größere Scaffolds aus der Gesamtdatenbank exportiert, manuell bearbeitet, und reimportiert (5). Vor allem die Sequenzen der Contigränder wurden in diesem Schritt bearbeitet. Diese waren, da meist A+T reich, von schlechterer Qualität als die Sequenzen der meist kodierende Regionen enthaltenden inneren Contigbereiche. Außerdem sind solche Sequenzen wegen ihrer geringen Komplexität für die Suche nach übereinstimmenden Sequenzen wenig geeignet. Vermeintliche Treffer erwiesen sich häufig als falsch, d.h. automatisch assemblierte Sequenzen mussten wieder entfernt werden. Auch Duplikationen von Genen wurden, wenn sie nur wenige Polymorphismen enthielten, falsch assembliert und mussten dann manuell bearbeitet werden.

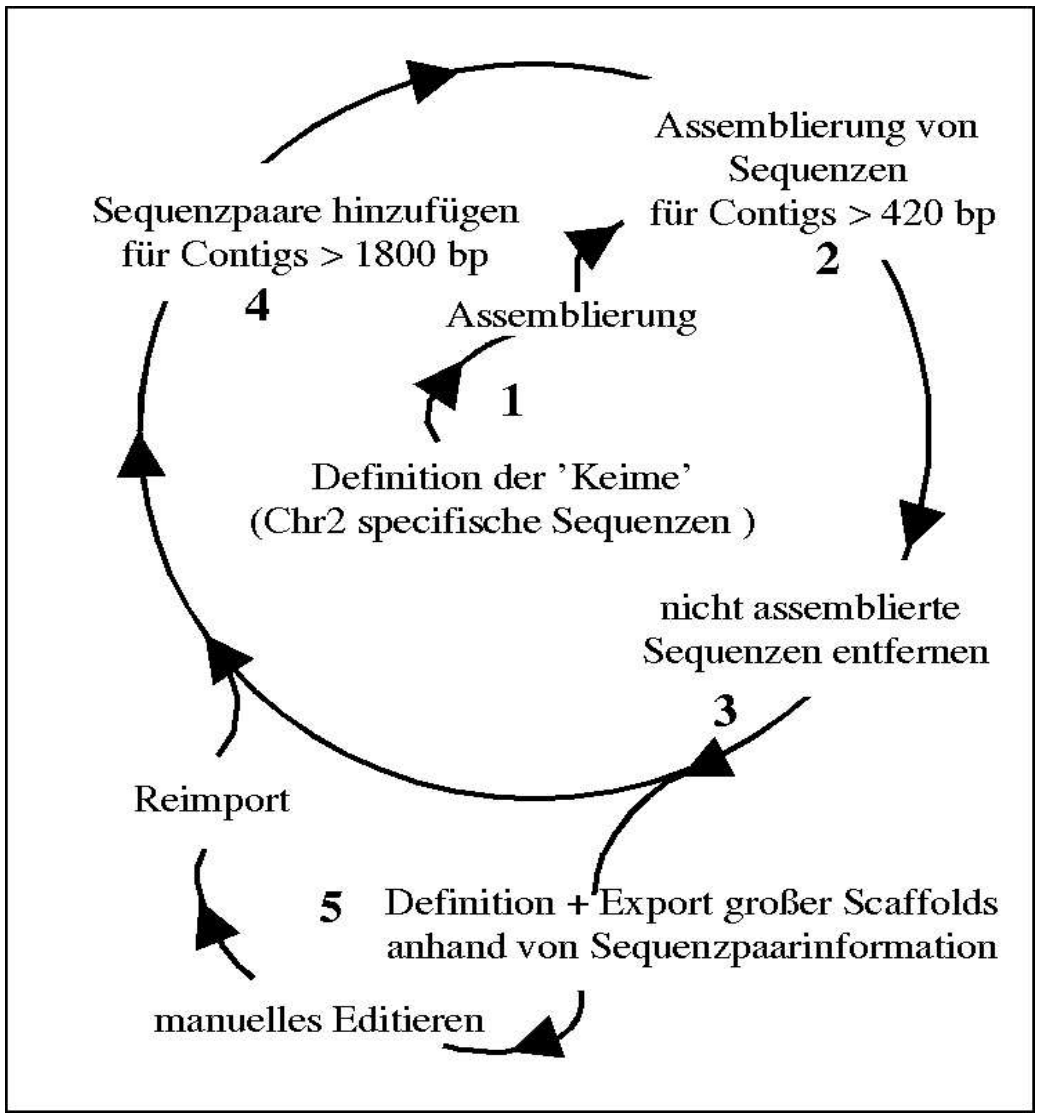


Abb. 6: Das Assemblierungsschema, hier mit Chromosom 2 Sequenzen als Startpunkt. Im Gegensatz zu üblichen Verfahren, die mit dem rechnerisch benötigten Satz an Sequenzen für eine vollständige Assemblierung arbeiten, wurde hier eine sequentielle Methode angewandt. Schritte 1 bis 4 wurden kontinuierlich automatisch durchgeführt, sobald mehr als 1000 neue Sequenzen produziert worden waren.

Infolge von Fehlbenennungen aufgrund technischer Limitationen (ungenaues line tracking etc.) stellte die Addierung der Gegenstücke bereits assemblierter Sequenzen eine nicht zu unterschätzende Fehlerquelle dar. Im Laufe des Assemblierungsprozesses akkumulierten deswegen immer mehr Contigs von anderen Chromosomen, die aber sehr kurz blieben. Dies lag an der hauptsächlichen Verwendung der chromosomenspezifischen Klonbibliotheken, Sequenzen aus den übrigen chromosomenspezifischen Klonbibliotheken wurden erst später zur Vervollständigung und Verbesserung der Sequenz verwendet.

Chromosom 2 war das erste Chromosom, an Hand dessen diese Strategie auf ihre Durchführbarkeit getestet wurde. Nach dem Abschluss dieser Assemblierungsarbeiten waren schließlich 2512 Contigs vorhanden, von denen aber nur 1113 für Chromosom 2 spezifisch waren. Durch die Verwendung der „read pair information“ konnten diese Contigs zu 480 größeren Einheiten zusammengefasst werden (Abb. 7)

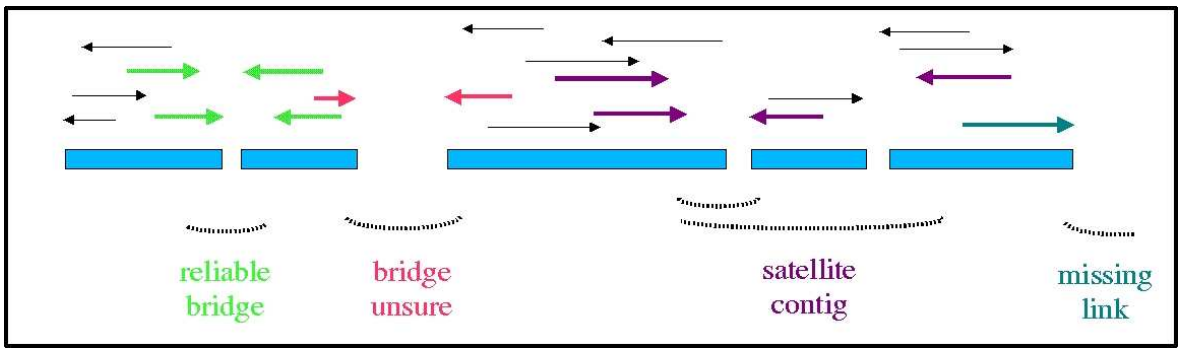


Abb. 7: Bildung von Scaffolds anhand von read pair information. Gegeneinander weisende Pfeile gleicher Farbe repräsentieren Sequenzen, die dem gleichen Klon zugeordnet sind. Mögliche Verbindungen zwischen Contigs über Klonbrücken sind gezeigt. Die für die Bildung von Scaffolds nutzbaren Informationen sind als unterbrochene Bogenlinien mit beschreibendem Text angezeigt.

Die Überbrückung von Lücken durch zwei oder mehr Klone wurde als sichere Verbindung zweier Contigs angesehen. Brücken aus Einzelklonen konnten jedoch auch durch Fehlbenennungen zustande kommen. Da aber die durchschnittliche Größe der Klone in den einzelnen Klonbibliotheken bekannt war, konnte die zu berücksichtigende Auswahl an Klonen reduziert werden. Nur Inserts, die einschließlich der Lücke eine rechnerische Größe von 4,5 kb nicht überstiegen, konnten überhaupt als Brückenklon in Frage kommen. Die Lücken innerhalb der Scaffolds konnten, da ja Klone zur Verfügung standen, relativ einfach durch die Ansequenzierung dieser Klone von spezifischen Primern aus geschlossen werden (Primer Walking). Auf diese Weise wurden 809 Lücken geschlossen. Im weiteren Verlauf der Arbeiten zeigte sich jedoch, dass die restlichen 109 Sequenzlücken nur noch schwer geschlossen werden können. An den Rändern dieser Lücken ließen sich nur schlecht Primer für Sequenzierungsraktionen definieren, da der A+T Gehalt hier sehr hoch war. Sequenzierung mit Hilfe von Transposonmutagenese (Devine et al., 1997) erbrachte ebenfalls nur schlechte Resultate. Es zeigte sich, dass das als Basis für einen Primer dienende Transposon präferentiell in bestimmte Positionen des Plasmides inserierte, so dass nicht die gesamte Lückensequenz bestimmt werden konnte.

Klonlücken sind, da die Information über benachbarte Sequenzen fehlt, nur schwer zu schließen. Sofern keine anderen Informationen bezüglich der Reihenfolge von Contigs vorliegen, kann nur ein beschränktes Arsenal an Methoden eingesetzt werden. Dazu zählt die inverse PCR (iPCR) und eine speziell für Genome A+T reicher Organismen entwickelte Methode, die für die unbekannte Sequenz in der Lücke gemischte Primer aus häufig vorkommenden kürzeren Sequenzmotiven als Ankerpunkt für eine konventionelle PCR einsetzt. Beide Methoden sind sehr arbeitsaufwendig und führen nicht immer zum Erfolg. Dementsprechend wurden nur drei Klonlücken mit Hilfe von iPCR geschlossen. Die Informationen über die Lage der Scaffolds innerhalb eines Chromosoms, die eine chromosomale Karte liefern, können jedoch effektiv genutzt werden, um mit einem PCR-Produkt zwei benachbarte Scaffolds zu verbinden. Auf diese Weise konnten 27 Klonlücken überbrückt werden. 25 Klonlücken verbleiben in der Sequenz von Chromosom 2, da kein Versuch der Lückenschließung zum Erfolg führte.

Nachdem exemplarisch am Chromosom 2 die Durchführbarkeit des ausgedachten Schemas bewiesen war, wurden auf die gleiche Weise auch Chromosom 1 und 3 assembliert.

2.5. Kartierungen

Karten sind essentiell für die Orientierung innerhalb eines Genoms. Drei Arten von genomischen Karten existieren: Genetische, physikalische und cytogenetische. Cytogenetische Karten können nur dann aufgestellt werden, wenn kondensierte Chromosomen beobachtet werden können. Dies ist für die meisten niederen Eukaryonten zur Zeit nicht möglich. Genetische Karten zeigen relative Abstände von Genen durch die Häufigkeit von Rekombinationsereignissen zwischen diesen während der Meiose (Gloeckner and Beck, 1995). Da Rekombinationshäufigkeiten nicht kontinuierlich, sondern diskontinuierlich über ein Chromosom verteilt sein können, können genetische Abstände (centiMorgan; cM) nicht direkt in physikalische Abstände umgerechnet werden. Physikalische Karten zeigen die echten Abstände zwischen Genen oder Markern. Sogar optische Methoden sind schon zur physikalischen Kartierung eingesetzt worden (Jing et al., 1999; Jing et al., 1998). Die physikalische Karte mit der höchsten Auflösung ist die DNA-Sequenz.

Wenn bakterielle oder YAC Klonkarten vorhanden sind, ermöglichen sie die zielgenaue Sequenzierung eines bestimmten Chromosomenabschnitts. Das wird vor allem bei der Auffindung von Genen, deren mutierte Varianten zu Krankheiten führen, genutzt (positionale Klonierung). Exakte Klonkarten bildeten auch das Gerüst für genomweite Sequenzierungen (McPherson et al., 2001). Darüber hinaus können solche Karten genutzt werden, um die zusammengesetzte Sequenz auf Assemblierungsfehler hin zu untersuchen. Umgekehrt muss eine Klon basierte Sequenzierung immer durch den Abgleich mit anderen Karten überprüft werden.

Als Ausgangsmaterial für weitere Kartierungen des *D. discoideum* Genoms standen integrierte genetische und YAC-Klon Karten zur Verfügung ("http://glamdring.ucsd.edu/others/dsmith/loci_maps.html") (Kuspa and Loomis, 1996; Loomis et al., 1995). Karten aus bakteriellen Klonen als Gerüst für eine klonbasierte Sequenzierung waren nicht vorhanden, da wegen des hohen AT Gehalts die Konstruktion von bakteriellen Klonen größer als wenige KB nicht möglich war. Die vorhandenen Karten erschienen jedoch als Basis für die Kontrolle der Sequenzierung zu ungenau, die Markerdichte war nicht hoch genug.

Für Chromosom 6 sollte zur Komplementierung dieser Karten eine sogenannte HAPPY Map erstellt werden. HAPPY mapping ist eine in vitro Technik, d.h. weder Meiose (klassische genetische Kartierung) noch lebende Zellen (radiation hybrid mapping (Cox et al., 1990a)) sind zur Lokalisierung der Marker nötig. Vielmehr wird genomische DNA an zufälligen Stellen in etwa gleich große Stücke gebrochen. Die Größe der Stücke bestimmt dann auch die maximale beobachtbare Kopplungsdistanz. Als Marker werden in diesem Fall Paare von Oligos verwendet, die ein singuläres PCR-Produkt ergeben, das also nur einmal im Genom existiert. Dann wird die Häufigkeit der Brüche zwischen den Markern gemessen, indem Stücke, die zusammen etwa ein halbes Genomäquivalent ergeben, auf die Anwesenheit verschiedener Marker getestet werden. Je häufiger Marker zusammen auftauchen („kosegregieren"), desto dichter liegen diese Marker auf dem Chromosom zusammen. Somit ist die Häufigkeit der Brüche ein Maß für den Abstand zwischen zwei Markern. Da diese Methode unabhängig von lebenden Zellen ist, können Phänomene wie Rekombinationshotspots oder Hybridinstabilitäten die Kartierung nicht beeinflussen.

Die mit dieser Methode erstellte Karte von Chromosom 6 zeigte große Diskrepanzen zu der vorher erstellten YAC Karte (Konfortov et al., 2000). Nur ein Bruchteil der kartierten YAC Klone befand sich laut HAPPY Map an der vorbestimmten Stelle, ein großer Teil der YAC Klone konnte überhaupt nicht mit der HAPPY Map von

Chromosom 6 in Einklang gebracht werden. Da YAC Klone aus A+T reicher DNA zudem instabiler sind als aus „normaler" DNA und damit mit vielen chimären Klonen zu rechnen war, wurden auf Grund dieser Resultate die YAC Karten als Assemblierungshilfe verworfen. Da die HAPPY Map Methode sehr schnell zu vollständigen Karten führt und sehr fehlerresistent ist, wurde im International Sequencing Consortium beschlossen, die HAPPY Map Kartierung des gesamten Genoms durchzuführen.

Für Chromosom 6 wurden noch Marker aus vorhandenen kartierten Genen sowie aus Einzelsequenzen der Chromosom 6 spezifischen Klonbibliothek definiert. Wegen der nur 50 % Anreicherung dieser Bibliothek mit chromosomenspezifischen Sequenzen konnten etliche dieser Marker nicht in die Karte für Chromosom 6 integriert werden, sie waren auf anderen Chromosomen lokalisiert. Bei der HAPPY Map Kartierung aller anderen Chromosomen konnte eine etwas andere Strategie zur Herstellung von Markern angewandt werden. Nach der Assemblierung und Lückenschließung (2.2.) waren jeweils Contigs vorhanden, die über ihre Zusammensetzung aus Sequenzen der verschiedenen Klonbibliotheken als zu den jeweiligen Chromosomen gehörig definiert werden konnten. Die aus der YAC Karte geschätzten Größen für alle Chromosomen waren etwas geringer bis gleich der aus den aneinander gehängten Contigs resultierenden Chromosomengrößen. Deshalb konnten wir davon ausgehen, dass der größte Teil der Chromosomen in diesem Datensatz vertreten war. Um nun eine Karte der Chromosomen zu erhalten, mussten auf den vorhandenen Contigs Markerregionen definiert werden, die möglichst nur einmal im Genom vorhanden waren. Darüber hinaus sollten die einzelnen Contigs bzw. Scaffolds möglichst wenige Marker enthalten, damit eine optimale Kombination von Sequenz- und Kopplungsgruppeninformation eine vollständige Karte ergeben würde. Nach dem Aussuchen der potentiellen Markerregionen trug jedes Contig bzw. jeder Scaffold mindestens zwei Markerregionen, die per BLAST-Analyse gegen den gesamten verfügbaren Sequenzdatensatz auf Einzigartigkeit hin überprüft waren. Einerseits wurde dadurch eine Markerdichte von 1 Marker/15 kb erreicht. Darüber hinaus erlaubte dies nicht nur die Lokalisierung, sondern auch die Orientierung der Contigs entlang des Chromosoms.

In P. Dear´ s Labor in Cambridge wurden die von uns definierten Markerregionen in Primerpaare für die PCR umgesetzt und nochmals auf ihre Einmaligkeit hin mittels einer PCR gegen genomische DNA untersucht. Für jedes Primerpaar wurden 96 PCR Reaktionen im Kartierungsprozess durchgeführt (P. Dear, pers. Mitteilung). Die Kartierung ergab jeweils eine Kopplungsgruppe für die Chromosomen 1 und 3, sowie insgesamt zwölf große Kopplungsgruppen für Chromosom 2, wobei durch den Abgleich mit den Chromosom 2

Sequenzen 3 kleinere Diskrepanzen in der Karte der HAPPY Map Marker ausgeräumt werden konnten. Solche Diskrepanzen können durch die Generierung von falsch positiven PCR Signalen, wie sie bei Hochdurchsatzmethoden anfallen können, entstehen. Aber auch die Hybridisierung der Primer an „falsche“ Stellen im Genom, die dann Produkte gleicher erwarteter Größe ergeben, kann ein Problem darstellen. Die falsche Zuordnung von Markergruppen kann bei der Kartierung nur erkannt werden, wenn eine vollständige Karte entsteht. Diese muss dann an allen Stellen konsistent sein. Im Falle von Chromosom 2 konnte eine vollständige Karte nach der HAPPY Map Methode nicht erstellt werden. Einerseits müssen weitere Marker, die die bereits vorhandenen Kopplungsgruppen verbinden können, definiert werden. Andererseits zeigte sich, dass eine kartierungstechnisch unüberbrückbare Lücke in diesem Chromosom vorhanden ist (siehe unten).

Ein Beispiel für eine der größeren Kopplungsgruppen ist in Abbildung 8 gezeigt. Sie überspannt knapp 500 kb, enthält mehrere Komplexe Repetitive Elemente und weißt noch mehrere Lücken auf. Nicht alle Contigs, die in diese Karte aufgenommen wurden, trugen HAPPY Map Marker. Vielmehr konnte über Klonbrücken, die Scaffolds definierten, ein großer Teil der Contigs in diese Kopplungsgruppe eingeordnet werden. Die anschließend erfolgreich durchgeführte Schließung von etlichen Lücken bestätigte die Reihenfolge der Contigs.

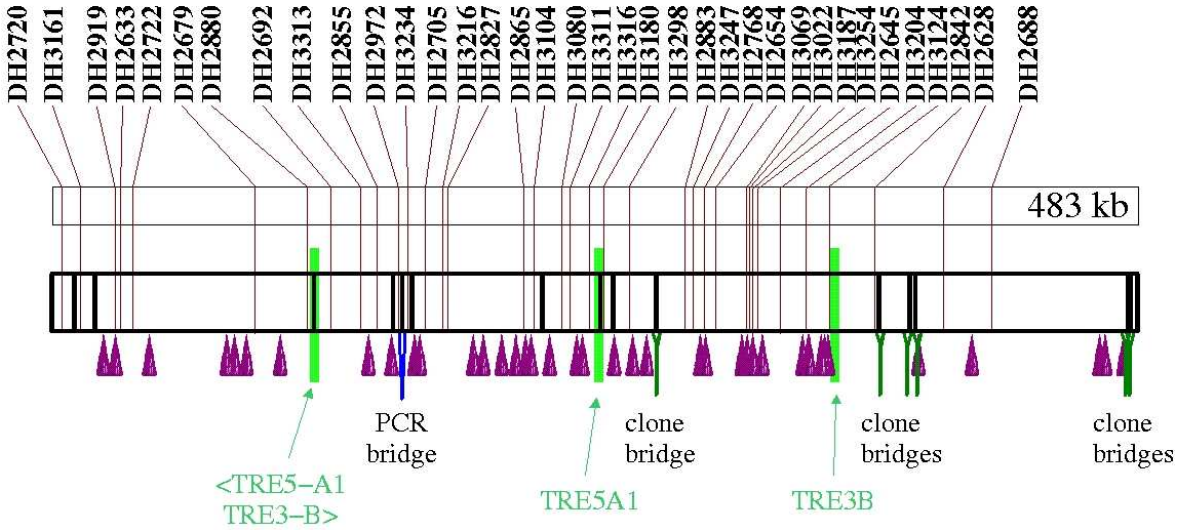


Abb. 8: Eine durch HAPPY Map Marker definierte Kopplungsgruppe. Gezeigt ist eine Region von Position ca. 2,5 MB bis 3 MB, von der Centromerregion aus gerechnet. Über der Kopplungsgruppe sind die HAPPY Map Marker, die zur Kartierung verwendet wurden, an den durch die Sequenz definierten Stellen angegeben. Schwarze Linien zeigen die noch vorhandenen Lücken. Die violetten Dreiecke zeigen die Positionen geschlossener Lücken. Auf Grund der Informationen, die durch die Kartierung entstanden, konnten auch neun Klonlücken geschlossen werden. Grün gefärbte Bereiche definieren die Lage von KREs. Sie stellen nicht schließbare Klonlücken dar.

Die repetitiven Element Loci definieren Lücken, die nicht durch Klone überspannt werden können. Wenn beide Enden das gleiche repetitive Element aufweisen, kann die Lückengröße über den Consensus dieses Elements abgeschätzt werden. Im Falle des ersten Locus auf der Karte ist die Lücke durch die Enden sowohl eines TRE3 als auch eines TRE5 Elementes definiert, das Aussehen des Innenbereichs des Locus kann nicht analysiert werden. Es ist jedoch zu vermuten, dass sich innerhalb dieses Locus eine tRNA verbirgt, da TRE Elemente in Abhängigkeit von ihrer Insertionspräferenz immer 5´ oder 3´ einer tRNA inserieren (Abb. 5).

Kopplungen von HAPPY Map Markern konnten unter den verwendeten Versuchsbedingungen nur bis zu einer Distanz von ca. 100 kb beobachtet werden. Die assemblierten Contigs und Scaffolds waren teilweise erheblich größer. Deshalb konnten in vier Fällen je zwei Kopplungsgruppen über eine Sequenz- bzw. Klonbrücke zusammengefasst werden. Zwei dieser Brücken wurden später durch Kopplung von HAPPY Map Markern bestätigt.

Um die Karte von Chromosom 2 zu vervollständigen, wurden weitere Ressourcen für die Kartierung herangezogen. Wie schon erwähnt, war die vorhandene YAC Karte wegen großer Ungenauigkeit nur eingeschränkt verwendbar. Deshalb wurde die Erstellung einer Teilkarte aus weiteren YAC Klonen anhand von Chromosom 2 spezifischen Markern angestrebt. Aus allen 3084 Klonen einer von P. deJong (Children´ s Hospital Oakland Research Institute in Oakland, Ca., USA) konstruierten zirkulären YAC Klonbibliothek (cYAC) wurde die DNA isoliert und so in Pools zusammengefasst, dass sie mit wenigen PCR Reaktionen (pro Primerpaar 49) auf passende Klone hin untersucht werden konnte. Die 100 größten Scaffolds wurden so verschiedenen cYAC Klonen zugeordnet. Mit Hilfe von cYAC Klonen, die mehrere Scaffolds überspannten, konnten weitere 4 Lücken in der HAPPY Map überbrückt werden. Eine weitere Informationsquelle stellten die genetischen Karten dar. Sie erlaubten, die großen Kopplungsgruppen relativ zueinander anzuordnen. Alle Ressourcen zusammen erlaubten die Konstruktion einer Chromosom 2 Karte, die mit 7,8 MB alle dem Chromosom zugewiesenen Contigs umfasste. Auf diesen 7.8 Mb waren 615 HAPPY Map Marker vertreten. Da 726 kb des Chromosoms verdoppelt sind, sind auch die entsprechenden Marker dieses Abschnitts zweimal vorhanden. Insgesamt bestand der assemblierte Teil des Chromosom 2 inklusive der Duplikation, jedoch ohne einige kleinere aus dem Centromer stammende Contigs, aus 8,5 MB. Ein Teil der KREs konnte nicht assembliert werden, der Anteil der KREs am Chromosom entspricht jedoch im Wesentlichen (ohne centromerlokalisierte DIRS Elemente) dem berechneten Anteil am Gesamtgenom von 9,6 %. Deshalb kann davon ausgegangen werden, dass nur ein unwesentlicher Teil dieser Elemente im Assembly fehlt. Die Gesamtgröße ist deutlich höher als nach der Messung anhand der, allerdings fehlerhaften, YAC-Karte zu erwarten war. Frühere Messungen der Chromosomengröße nach einer Auftrennung mit Hilfe der Pulsed Field Gel Elektrophorese ergaben um ca. 9 MB (Cox et al., 1990b). Die Messgenauigkeit ist jedoch gerade bei großen Chromosomen wegen der logarithmischen Natur der Auftrennungsabstände bezogen auf die Basenanzahl mit einem Fehler in der Größenordnung von bis zu 20 % behaftet. Weder die Chromosomenkarte noch die assemblierten Sequenzen gaben Hinweise auf das Fehlen größerer nicht repetitiver Stücke des Chromosoms. Die 8,5 Mb Sequenz repräsentieren somit höchstwahrscheinlich den Anteil des gesamten Chromosoms, der ohne Fehler assemblierbar ist. Sowohl in der Karte als auch in der Sequenz fehlen die Anteile, die das Centromer ausmachen, sowie die Teile, die direkt anschließend an die Duplikation in der kopplungslosen Region vorkommen, da diese hoch repetitiv sind.

Die Kopplungsgruppe von Chromosom 1 überspannt 4.8 Mb, die von Chromosom 3 6.8 Mb. In beiden Fällen ist das Centromer nicht in die Gesamtlänge einbezogen, da keine Marker für diese Regionen definiert werden konnten. Eine versuchte Assemblierung des Chromosom 1 Centromers ergab eine Länge von 200 kb für diesen chromosomalen Abschnitt, so dass die Gesamtlänge des Chromosoms etwas über 5 Mb beträgt.

2.6. Analyse von Chromosom 2

2.6.1 Das Chromosom im Überblick

Die Sequenzen der 7,8 MB Kopplungsgruppe wurden mit Hilfe von dafür entwickelter Software unter Integration aller vorhandenen Daten widerspruchsfrei in richtiger Reihenfolge und Orientierung zusammengesetzt. Die im Folgenden präsentierten Ergebnisse aus der Analyse des Chromosom 2 wurden im Jahr 2002 veröffentlicht (Glöckner et al., 2002). Die resultierende Sequenz wurde auf ihre Eigenschaften hin untersucht. Die Analyse zeigte, dass über die gesamte Region der A+T Gehalt nur schwach um den Durchschnittswert von 77,8 % schwankte (Abb. 9). Da diese Analyse unter Verwendung um 1 kb überlappender 20 kb Stücke durchgeführt wurde, konnten allerdings kleinere Bereiche mit G+C Minima nicht entdeckt werden. Aber auch in einer detailierteren Analyse von 5 kb Stücken wurden keine Regionen gefunden, die signifikante Abweichungen von diesem durchschnittlichen A+T Gehalt aufwiesen.

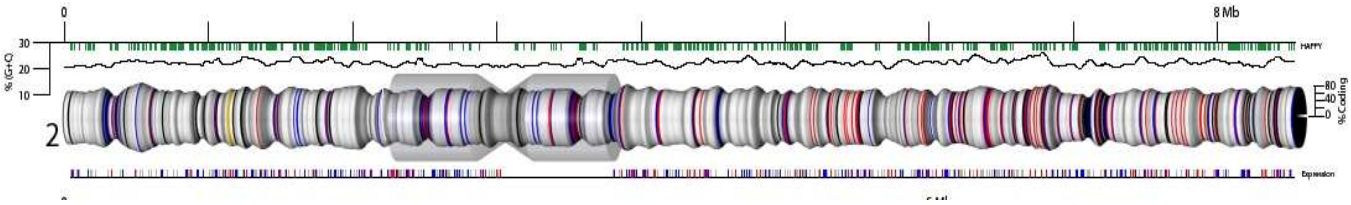


Abb. 9: Verteilung von Merkmalen auf der 8,5 MB Kopplungsgruppe von Chromosom 2. Die Sequenzen der Contigs wurden in der laut Karteninformation richtigen Reihenfolge und Orientierung aneinander gelegt, wobei für Klonierungslücken 100 und für Sequenzierungslücken 50 N eingeführt wurden. Die Kodierungsdichte ist als Durchmesser des Chromosoms dargestellt. Blaue Banden repräsentieren komplexe repetitive Elemente, rote Banden tRNAs und schwarze Lücken. Der GC Gehalt ist über dem Chromosom angegeben in Form einer Kurve, die mit einem „Sliding Window" von 100 kb und Schrittweite von 10 kb berechnet wurde. Die Position der HAPPY Marker für die physikalische Karte des Chromosoms sind als grüne Striche an der Messlatte angegeben. Unterhalb des Chromosoms befinden sich die Informationen über die hoch (rot) und heruntergeregelten (blau) Gene, wenn die Zellen in den vegetativen Entwicklungszyklus eintreten (im duplizierten Bereich nur auf einer Seite angegeben). Besonders hervorgehoben ist die Region der Duplikation durch die übergestülpte transparente Form.

Die Kodierungskapazität zeigt dagegen etwas größere Schwankungen. Offensichtlich hängen die beiden Parameter Kodierungsdichte und G+C Gehalt also nicht direkt zusammen. Die Untersuchung der strangspezifischen Kodierungskapazität ergab keine Präferenz für + oder – Strang.

KREs sind gleichmäßig über das gesamte Chromosom verteilt. Die meisten dieser KREs sind über HAPPY Map Marker miteinander verkoppelt, so dass die Wahrscheinlichkeit für einen ausgedehnten KRE Lokus (> 10 kb) innerhalb dieser Kopplungsgruppe gering ist. Der Anteil der verschiedenen KREs an den Loci ist in Tabelle 5 festgehalten. Die assemblierten Basen der KRE Loki in dieser Kopplungsgruppe machen zusammen 128,5 kb aus. Weitere 350 kb an KRE Sequenzen sind in einigen Contigs enthalten, die als Chromosom 2 spezifisch assembliert, aber nicht in die Karte integriert werden konnten. Ein Charakteristikum für diese Contigs ist der hohe Anteil an DIRS und Skipper Elementen, der signifikant vom Durchschnittswert für das Genom abweicht. Zugleich finden sich in diesen Contigs keinerlei TRE Elemente. Obwohl es keine Beweise dafür gibt, wird das Centromer jeden *D. discoideum* Chromosoms acro- bis telozentrisch eingebettet in einen großen, DIRS-Elemente enthaltenden, repetitiven Lokus vermutet. Für Chromosom 2 konnten wir zwei solcher Loci nachweisen, einer befindet sich an einem Ende des Chromosoms, einer am Ende der Duplikation. Die nicht in die Sequenz

integrierbaren Contigs sollten sich demnach an diesen beiden Stellen befinden. Wie sich allerdings die 350 kb auf diese Positionen verteilen, ist wegen der hoch repetitiven Natur dieser Contigs kaum aufzuklären.

Tabelle 5: Anzahl der auf der größten Kopplungsgruppe von Chromosom 2 festgestellten Fragmente von Komplexen Element Familien (KRE). Nicht assemblierbare oder integrierbare Teile der Loci fehlten bei der Analyse. So können weitere Elementfamilienmitglieder oder Bruchstücke derselben innerhalb der Loci und in Centromer bzw. Telomerregionen vorhanden sein

LTR Retrotransposonen			Non-LTR Retrotransposonen		DNATransposonen		Unklassifiziert
DIRS	skipper	DGLT	TRE3	TRE5	Tdd	DDT	thug
2	2	3	31	34	6	4	6

Die in Tabelle 5 aufgeführten DIRS Elemente befinden sich am Ende der Duplikation als sehr kurze (~150 b) Stücke, die wohl den Übergang zu einem Pseudocentromer anzeigen (Abb. 10). Insgesamt sehen wir also eine strenge Aufteilung der verschiedenen repetitiven Elemente auf bestimmte chromosomale Abschnitte: TREs befinden sich nur im Kern des Chromosoms, der auch den kodierenden Anteil des Genoms darstellt, während DIRS ausschließlich und Skipper fast ausschließlich in großen Clustern residieren, die womöglich die Centromerfunktion vermitteln.

Sowohl die Sequenz- als auch die Klonlücken sind annähernd gleich über die Kopplungsgruppe verteilt. Allenfalls könnte eine leichte Häufung gegen die Enden zu nicht zufällig sein. Auffällig ist hingegen die Häufung von tRNAs am dem Centromer gegenüber liegenden Telomer. Einhergehend damit häufen sich hier auch die non-LTR Transposonen, die TRE, da sie nur in der Nähe von tRNAs inserieren. Ob die Häufung der tRNA Gene an einem Ende des Chromosoms eine funktionelle Signifikanz hat, bleibt unklar.

Innerhalb von Chromosom 2 wurden Sequenzen aus dem rDNA Palindrom gefunden (Abb. 10), die fast das vollständige Palindrom umfassen könnten. Dieser Lokus konnte nicht assembliert werden, da nicht zu unterscheiden war, welche Sequenzen von der genomischen Kopie, und welche von den extrachromosomalen Palindromen stammten. An dieser Stelle konnte Kopplung von Markern weder über die HAPPY Map noch über cYAC

Klone festgestellt werden. Da Marker, die weiter als ca. 50 kb voneinander entfernt sind, bei der HAPPY Kopplungsanalyse als ungekoppelt erscheinen, muss diese Region einen größeren Bereich überspannen.

Hybridisierungen hatten schon früher ergeben, dass sich im Genom von *D. discoideum* eine Kopie des Palindroms befinden musste. Inzwischen haben wir diese Kopie auf Chromosom 4 lokalisiert. Damit ergibt sich für diesen Organismus eine ähnliche Situation wie bei *Tetrahymena thermophila*. Auch diese Spezies verfügt über ein extrachromosomales rDNA Element, das als Palindrom 21 kb umfasst. Dieses Palindrom wird aus der einfachen genomischen Kopie erzeugt und mit Telomeren versehen (Kapler, 1993). Ein ähnlicher Mechanismus könnte bei *D. discoideum* wirksam sein. Die von uns aufgefundenen Sequenzen konnten wegen des fehlenden Hybridisierungssignals gegen die rDNA Gene keine weitere „Master Copy“ darstellen. Also lag es nahe, die Umgebung dieser Sequenzen nach Hinweisen darauf zu durchsuchen, welchem Zweck diese Sequenzen dienen könnten.

Direkt anschließend an die rDNA Sequenzen befindet sich eine inverse Duplikation. Sie ist ebenfalls nicht auflösbar, da sie keine Polymorphismen enthält. Diese Duplikation konnte nur indirekt durch die Anwesenheit des Genes carA, das für einen cAMP Rezeptor kodiert, und durch die Definition des Zentrums nachgewiesen werden. Der nur durch 30 Jahre separater Haltung unterschiedene Stamm AX2 enthält diese Duplikation noch nicht (Loomis and Kuspa, 1997), so dass die Abwesenheit von Polymorphismen nicht weiter verwunderlich ist. Die Etablierung dieser Duplikation muss jedoch mit weiteren genomischen Änderungen einhergegangen sein, damit die Zelle dieses Ereignis unbeschadet überstehen konnte. Eine dieser Änderungen könnte die Fusion mit rDNA Palindromsequenzen gewesen sein. Falls dies eine natürliche Reaktion der Zelle auf eine Verletzung der Integrität eines Chromosoms wäre, sollten sich weitere solcher Fälle aufspüren lassen.

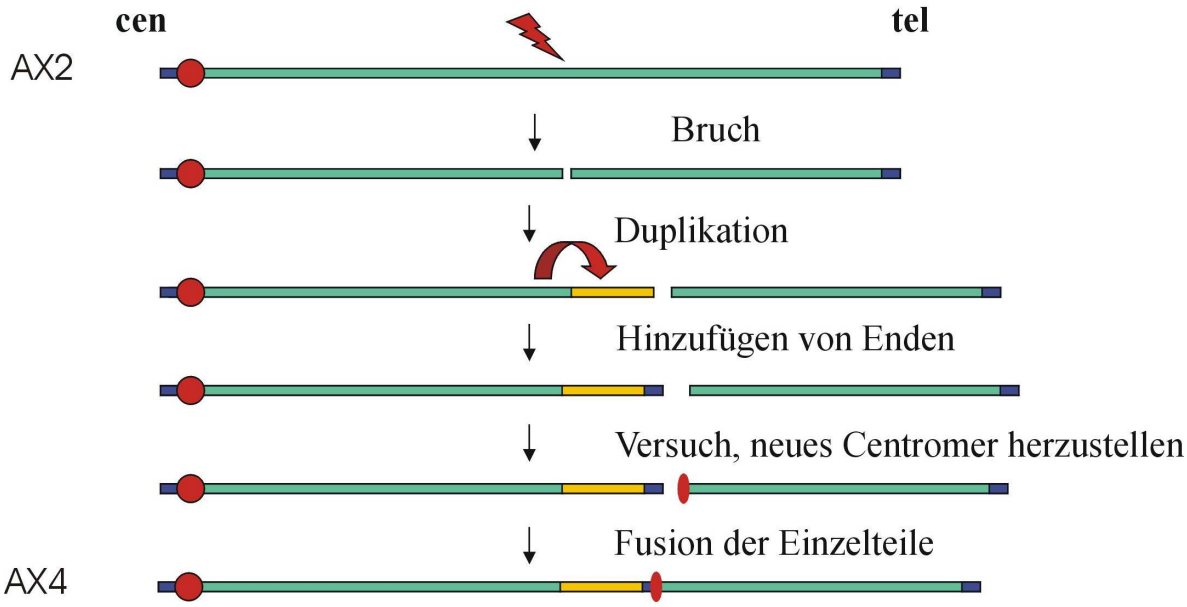


Abb. 10: Schematische Darstellung der Entwicklung des Chromosom 2 aus AX2, dem Ursprungsstamm hin zur Situation in AX4, dem sequenzierten Stamm. Die einzelnen Regionen sind nicht proportional gezeichnet. cen = Centromer; tel = Telomer.

Inspiriert vom gekoppelten Auftreten von rDNA Sequenzen mit chromosomalen Abschnitten, urchsuchten wir die Rohdaten nach weiteren solchen Regionen. Insgesamt konnten wir 16 solcher Sequenzen einschließlich derer von Chromosom 2 zweifelsfrei bestimmen. Auf Grund der Herkunft der Sequenzen aus einzelnen chromosomenspezifischen Bibliotheken konnten je zwei dieser Sequenzen einem Chromosom zugeordnet werden. Diese konnten nur jeweils an den Enden der Chromosomen lokalisiert sein. Somit laufen alle Chromosomen in Sequenzen aus, die vom rDNA Palindrom stammen. Diese Erkenntnis führte zu einem Modell, wie die aufgefundenen Strukturen in Chromosom 2 aus einem „normalen“ Chromosom entstanden sein könnten (Abb. 10). Nach einem Bruch des Chromosoms wurde in einer klassischen break- fusion-bridge Reaktion eine inverse Duplikation etabliert. Um die Integrität des so beschädigten Chromosoms über mehrere Zellzyklen zu erhalten, müssen Telomere sowie Centromer hinzugefügt werden. Offensichtlich gelang es der betroffenen Zelle nicht, diese Reparatur erfolgreich zum Abschluss zu bringen. Es blieb nur die Refusion der Einzelteile. Der Aufbau des heutigen AX4 Chromosom 2 ist also nun: Telomer-Centromer-Genregion-Duplikation-Telomer(pseudo)-Centromer(pseudo)-Genregion-Telomer.

2.6.2 Genmodelle und Proteinanalyse

Für die Auffindung von Genen stand zunächst kein für *D. discoideum* spezifisches Programm zur Verfügung. Da aber *P. falciparum* über eine ähnliche Basenzusammensetzung verfügt, wurde ein für diesen Organismus entworfenes Programm (GlimmerM, (Salzberg et al., 1999)) mit einer leichten Anpassung, was die Erkennung von Kodierungspotential betraf, verwendet. Dieses Programm postulierte 2995 Gene. Ein Vergleich von Genmodellen mit einer manuellen Annotation ergab jedoch, dass die vorhergesagten Genmodelle in der Regel kürzer als die tatsächlichen Gene waren. Da keine Möglichkeit bestand, dieses Programm weiter anzupassen (der Source Code wurde nicht zur Verfügung gestellt), wurde auf ein anderes Programm ausgewichen. Ein von R. Guigo in Barcelona entwickeltes Programm, GeneID (Parra et al., 2000), war unter anderem erfolgreich bei der Annotation von *D. melanogaster* eingesetzt worden. In Zusammenarbeit mit dieser Arbeitsgruppe wurde GeneID nun auf die speziellen Erfordernisse von *D. discoideum* umgestellt. Dazu wurde ein Trainingsset von 140 Genen erstellt, deren Exon-Intron Struktur und umgebende intergenische DNA bekannt war. Eine manuelle Feinabstimmung machte eine bessere Vorhersage von Genen möglich als mit anderen Programmen (Abb. 11).

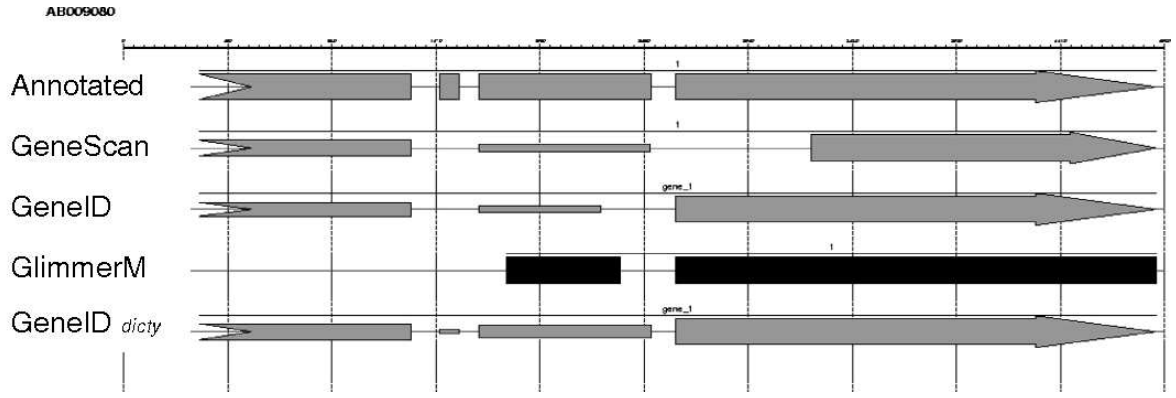


Abb. 11: Genvorhersageprogramme im Vergleich. Gezeigt ist die Vorhersage für TRFA (AB009080) aus *D. discoideum*. „Annotated" zeigt die Genstruktur so, wie sie in der Datenbank erfasst worden war. Für GeneScan und GeneID gibt die Höhe der Rechtecke die Qualität der Vorhersage an. GeneID*dicty* wurde mit einem Datensatz aus 140 Genen von *D. discoideum*, deren Genstruktur aufgrund einer vollständigen mRNA bekannt war, trainiert.

Auf dem gesamten Chromosom 2 wurden mit Hilfe von GeneID 2799 Genmodelle definiert. In dieser Zahl sind Genmodelle, die in Bereichen der KREs definiert wurden, nicht enthalten. Genmodelle, deren translatierte Proteine kleiner als 40 Aminosäuren waren, wurden ebenfalls nicht berücksichtigt. Alle Genmodelle an den Rändern der Contigs wurden auf Auffälligkeiten inspiziert und gegebenenfalls verändert, da Genvorhersageprogramme an Sequenzrändern generell mehr Fehler machen als inmitten eines Contigs. Dies hängt damit zusammen, dass Vorhersageprogramme den Regionenkontext bei der Definition von Genmodellen berücksichtigen, nicht nur das Kodierungspotenzial.

Tabelle 6: Generelle Eigenschaften der auf Chromosom 2 mit GeneID vorhergesagten Genmodelle.

Gene		2799
Gendichte (berechnet basierend auf den assemblierten Sequenzen)		1 Gen/2,7 kb
durchschnittliche Genmodelllänge		1626 Basen
Anzahl der Genmodelle mit ESTs (sexuelle und Hungerstadien)		1120 (40%)
kodierende Exonen		
	Anzahl	6398
	Gesamtlänge	4550112 Basen
	durchschnittliche Anzahl der Exons/Gen	2,29
	durchschnittliche Länge	711 Basen
Intronen		
	Anzahl	3587
	Gesamtlänge	523702 Basen
	durchschnittliche Länge	146 Basen

Wie aus Tabelle 6 ersichtlich, ist die Gendichte in *D. discoideum* mit 1 Gen/2,7 kb sehr hoch. Die spätere Wiederholung der Analyse mit den vollständigen Sequenzen ergab den gleichen Wert für das gesamte Genom. Als einziger freilebender eukaryontischer Organismus übertrifft nur *S. cerevisiae* diese Dichte mit 1 Gen/2 kb, die Gendichte z.B. von *P. falciparum* ist mit 1 Gen/4,5 kb wesentlich geringer. Zum Teil hängt der etwas höhere

A+T Gehalt im *P. falciparum* Genom mit dieser niedrigeren Gendichte zusammen, wie wir nachweisen konnten (Szafranski et al., 2005). Die durchschnittliche Intronlänge von *D. discoideum* ist etwas größer als bei *P. falciparum*, die Menge an Intronen pro Gen ist jedoch vergleichbar (Gardner et al., 1998).

Nur ein Teil der Genmodelle ist durch transkribierte Einheiten (ESTs) als exprimiert gekennzeichnet. Da die ESTs jedoch überwiegend aus speziellen Differenzierungsstadien stammen (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html; (Morio et al., 1998)) ist dies nicht weiter verwunderlich. Weitere cDNA Bibliotheken, die aus mRNA vegetativer Stadien stammten, werden inzwischen analysiert. Die Einbeziehung dieser erweiterten EST-Ressource ergab eine Expression von 58 % aller Gene in den untersuchten Zellzuständen.

Die translatierten Genmodelle wurden nun einer weiteren Analyse unterzogen. Diese beinhaltete den Vergleich mit allen bis jetzt bekannten vollständigen Proteinsätzen von Eukaryonten (*S. cerevisiae, S. pombe, C. elegans, D. melanogaster, A. thaliana*). Des Weiteren wurde zur Analyse des Proteinsets der vollständige Datensatz von *H. sapiens*, ergänzt um Einträge von Vertebratengenen in Swissprot, herangezogen. Auch die Gesamtdatenbanken von Swissprot und embl wurden in die Analyse mit einbezogen. Bei einem Schwellenwert von $p<e^{-15}$ hatten 47 % der translatierten Genmodelle einen Treffer in einer der Datenbanken. Ein Teil der übrigen Genmodelle mag trotzdem Orthologe in anderen Organismen besitzen, die aufgrund hoher Mutationsraten jedoch zu stark divergieren, um als solche erkannt zu werden. Insgesamt ist der Anteil an Genen ohne entsprechenden Gegenpart in anderen Genomen bei diesem Schwellenwert vergleichbar mit den Zahlen aus anderen Genomen. Anscheinend sind bis zu 50 % aller Gene in jedem Organismus individuelle Erfindungen, schnell mutierende Reservoirs für prospektive neue Funktionen oder Überbleibsel nicht weiter verfolgter Entwicklungslinien. Häufig sind diese Gene unter Laborbedingungen nicht exprimiert (Gad Shaulski, perönliche Mitteilung). Trotzdem ist für die möglichen Genprodukte eine Funktion in spezifischen Situationen denkbar, die in den jeweiligen ökologischen Nischen unter spezifischen Umständen auftreten.

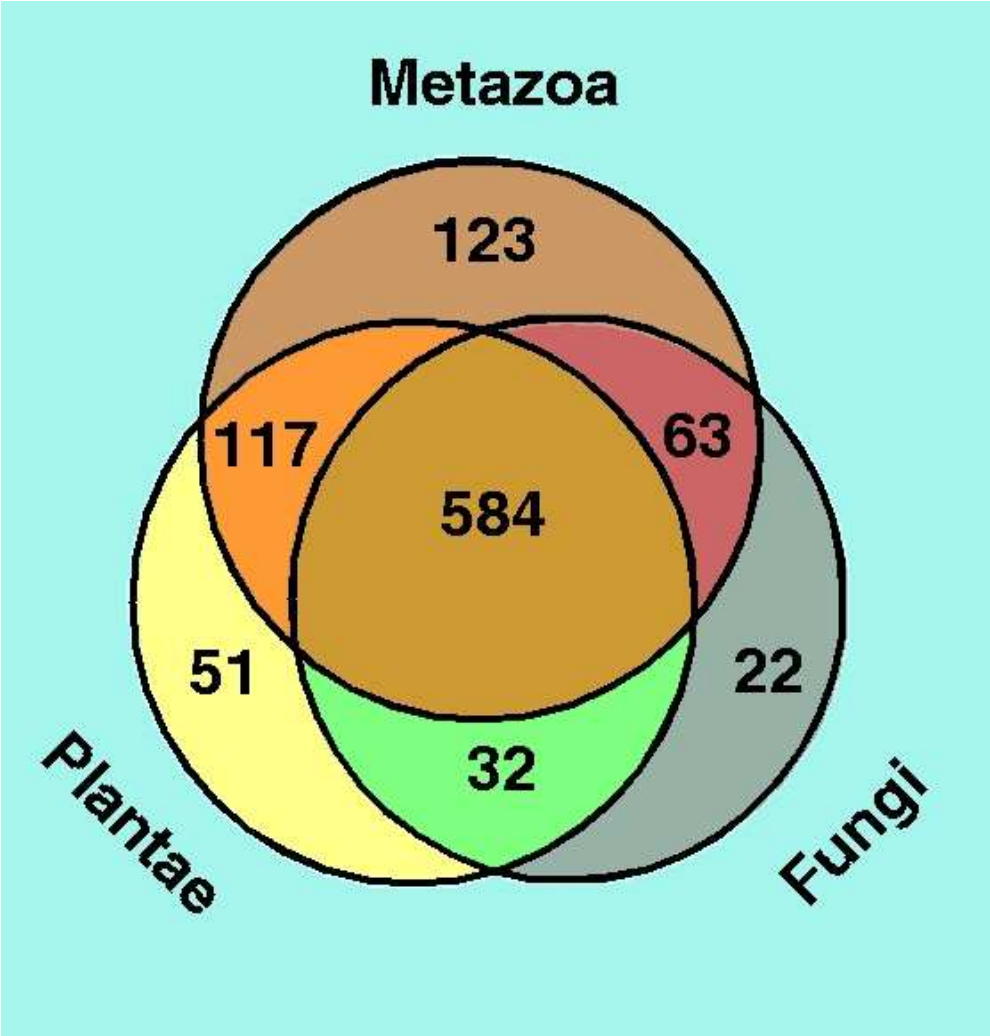


Abb. 12: Die Aufteilung von Proteinen zwischen Repräsentanten verschiedener Entwicklungslinien. Fungi = *S. cerevisiae* und *S. pombe*; Planta = *A. thaliana*; Metazoa = *C. elegans*, *D. melanogaster*, *H. sapiens* (inklusive anderer Vertebratenproteine, die nicht im bisher veröffentlichten Proteinsatz des Menschen vorhanden sind). Der Schwellenwert für die Ähnlichkeitsberechnungen war $p>e^{-15}$.

Da Chromosom 2 ca. 25 % des Genoms und damit auch des Gensatzes von *D. discoideum* beinhaltet, können mit Hilfe dieser Daten auf das gesamte Genom bezogene Aussagen getroffen werden. Spätere Analysen am vollständigen Genom bestätigten denn auch die Schlüsse, die aus den Daten zu Chromosom 2 gezogen wurden. Aus der Verteilung der Proteinähnlichkeiten zwischen den verschiedenen evolutionären Entwicklungslinien geht eindeutig hervor, dass *D. discoideum* eher den Metazoa als den Pflanzen zuzuordnen ist (Abb. 12). Diese Aussage stimmt mit dem aus wenigen Proteinsequenzen errechneten Stammbaum (Baldauf et al., 2000) gut überein. Bei einem Schwellenwert von $p>e^{-15}$ können jedoch auch nicht orthologe Proteine einander zugeordnet werden. Verschiedene Spezies der gleichen Gruppe können unterschiedliche

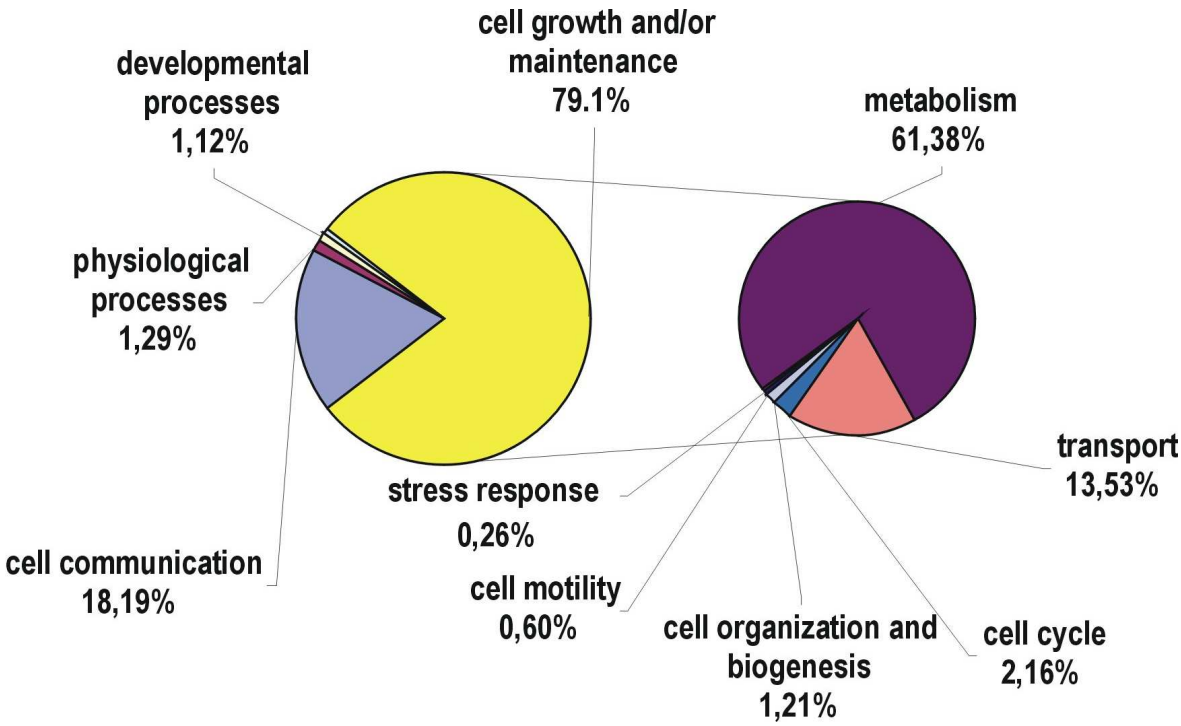
Proteinsets haben, je nachdem, welche Gene im Laufe der Evolution verloren gegangen sind. So besitzt *S. pombe* trotz der geringeren absoluten Genzahl einige Gene des „Eukaryontengensatzes", die *S. cerevisae* verloren hat. Möglicherweise wurde die Funktion der verloren gegangenen Genprodukte von anderen, nicht homologen Proteinen übernommen. Da das gezeigte Diagramm nur auf den Daten weniger Spezies beruht, werden sich mit der Addierung jeder weiteren Spezies die Zahlen etwas verändern, bis näherungsweise alle Gene, die es in einer Hauptgruppe gibt, erfasst sind. Der Kern an gemeinsamen Genen wird dadurch etwas größer und sollte die Obergrenze eines allen Eukaryonten gemeinsamen Gensatzes wiederspiegeln. Wenn man diese Unsicherheiten einrechnet, kann von 2000 - 2500 Genen ausgegangen werden, die von allen Eukaryontenhauptgruppen geteilt werden (homologer Gensatz). Das sind ungefähr gleich viele Gene, wie für das uneingeschränkte Funktionieren einer Eukaryontenzelle postuliert werden (essentieller Gensatz; siehe auch 1.1.2.2.). Der homologe und der essentielle Gensatz sind sicherlich nicht identisch, doch kann von einer weitgehenden Überlappung ausgegangen werden. Da innerhalb der Hauptgruppen verschiedene Organismen jeweils mit weniger Genen des „Eukaryontengensatzes" auskommen, spiegelt diese Diskrepanz sowohl eine gewisse Redundanz von Funktionen, als auch eine Reduktion von einer „omnipotenten" Eukaryontenzelle während der Speziation wieder.

Weitere Analysen beinhalteten Abgleiche der Proteine mit der COG Datenbank (http://www.ncbi.nlm.nih.gov/COG/cognitor.html (Tatusov et al., 2001)) und eine Suche nach Interpro Domänen. Alle gesammelten Daten wurden in eine Datenbank integriert. Über eine WWW basierte Schnittstelle können alle Daten inklusive Aminosäure- und genomische DNA-Sequenz abgerufen werden (http://genome.fli-leibniz.de/cgi-bin/dicty/map.pl) (Felder et al., 2005).

Das Gene Ontology Consortium (http://www.geneontology.org/) bemüht sich, einheitliche Klassifizierungsmethoden für eukaryontische Proteine hinsichtlich ihrer Funktion, Lokalisierung innerhalb der Zelle und Zuordnung zu einem biochemischen Prozess zu definieren. Am Ende dieser Bemühungen soll ein verbindliches, aber flexibles Vokabular (GO terms) zur Definition von Genprodukten stehen. Momentan werden Genprodukte der Organismen *C. elegans*, *S. cerevisiae*, *A. thaliana*, *Mus musculus* und *D. melanogaster* auf diese Weise klassifiziert. Wir haben zu den Organisatoren Kontakt aufgenommen, um auch *D. discoideum* in diesen Kreis aufnehmen zu können. Langfristig sollen auch GO terms für *D. discoideum* spezifische Funktionen wie z.B Sporenträgerbildung in das Vokabular eingeführt werden. Eine Liste, die Interpro Domänen

mit Hilfe der GO Terminologie zuordnet, wurde genutzt, um die Chromosom 2 Proteine zu klassifizieren (Abb. 13).

A



B

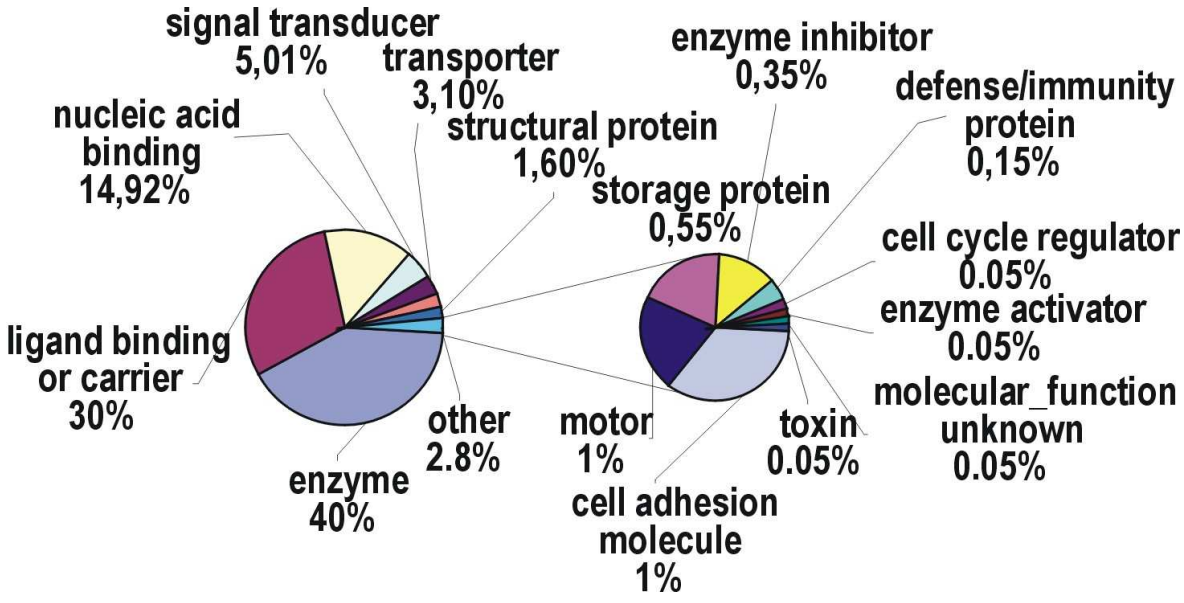


Abb. 13 (vorherige Seite)**:** Klassifizierung der Chromosom 2 kodierten Proteine anhand der Zuordnung von Interpro Domänen zur GO Terminologie. Benutzt wurde die unter http://www/geneontology.org/interpro2go abrufbare Klassifizierungsliste. Unklassifizierbare Proteine sind nicht gezeigt. Die prozentualen Werte der Ausschnittskreise sind stark gerundet. A: Zuordnung zu biochemischen Prozessen. B: Zuordnung zu molekularen Funktionen.

Die Verteilung der Proteine in die einzelnen Kategorien ist ähnlich der, die auch in anderen Organismen beobachtet werden kann. Im Vergleich zu Hefe finden sich aber funktionelle Kategorien, die wesentlich größeren Raum einnehmen. So benötigt *D. discoideum* zum Beispiel wesentlich mehr Proteine für Zellkommunikation und Entwicklungsprozesse. Die Analyse des gesamten Genoms ergab das gleiche Bild. Damit kann davon ausgegangen werden, dass Chromosom 2 dem Durchschnitt aller Chromsomen in Bezug auf das Kodierungspotential entspricht.

Insgesamt konnten so 48 % aller auf Chromosom 2 gefundenen translatierten Genmodelle zugeordnet werden. Allerdings ist eine rein automatische Zuordnung fehlerbehaftet und nicht komplett. So sind in dieser Klassifizierung Proteine, die gut dokumentierte Funktionen in *D. discoideum* haben, jedoch keinen Gegenpart in anderen Organismen, falsch eingeordnet. Eine Aufgabe für die Zukunft ist es, die Klassifizierung manuell zu überprüfen und zu verbessern.

Darüber hinaus sind einige Domänen in Hefe überhaupt nicht vertreten, die häufiger in *D. discoideum* vorkommen. In Tabelle 7 sind einige dieser Domänen zusammengefasst. Die Domäne Laminin-type EGF-like ist eine relativ schlecht definierte Domäne, da wenige Aminosäuren diesen Domänentyp bestimmen. Somit ist die falsch positive Rate relativ hoch. Involviert ist diese Domäne jedoch in wichtige Funktionen wie Differenzierung und Zelladhäsion. IPR001230 wiederum ermöglicht posttranslationale Modifikationen, die unter anderem bei G-Proteinen und Ras-Proteinen ausgeführt werden. Tropomyosin wiederum ist involviert in Zellbewegungen. Die relative Häufigkeit solcher Domänen korrespondiert direkt mit dem Lebensstil von *D. discoideum*. Einerseits müssen im vegetativen Lebenszyklus Zell-Zell-Kontakte hergestellt werden, die über Signaltransduktion initiiert werden müssen. Andererseits braucht natürlich ein amöboid beweglicher Organismus mehr Zytoskelettkomponenten als eine Zelle mit einer Wand wie *S. cerevisiae*.

Tabelle 7: Ausgewählte Interpro Domänen von *D. discoideum* Chromosom 2 kodierten Proteinen, die in *S. cerevisiae* nicht vorhanden sind.

Domäne	IPR Nummer	Anzahl der Proteine mit mindestens einer Domäne
Laminin-type EGF-like (LE) domain	IPR002049	25
Prenyl group binding site (CAAX box)	IPR001230	12
Neutrophil cytosol factor 2	IPR000108	10
HMW kininogen	IPR002395	9
Tropomyosin	IPR000533	8
TRAF-type zinc finger	IPR001293	8

2.7. Die Analyse des gesamten Genoms

In den vorhergehenden Kapiteln habe ich dargelegt, wie aus wenigen Sequenzen aussagefähige Daten über ein ganzes Genom gewonnen werden können. Dies einhaltet die Basenzusammensetzung, das Auftreten von Sequenzmotiven, die Abschätzung der Größe von Genfamilien und die Charakterisierung komplexer repetitiver Elemente. Durch die Einbeziehung eines ganzen Chromosoms in die Analyse konnten die statistischen Daten verfeinert, aber nicht wesentlich verändert werden. Dasselbe gilt für die Analyse des gesamten Genoms. Die Ausbeutung von Rohdaten stößt jedoch an Grenzen, wenn genaue Daten zu einzelnen Genen und Genfamilien erforderlich werden. In Einzelfällen und mit erheblichem Aufwand können Genfamilien aus Rohdaten definiert werden (Kollmar and Glöckner, 2003), identische Duplikate von Genen können jedoch nicht aufgelöst werden. Nur die genaue Kenntnis der 700 kb Duplikation auf Chromosom 2 erlaubt es zum Beispiel, die darin enthaltenen Gene für spätere Experimente als dupliziert zu erkennen und zu behandeln, eine statistische Analyse scheitert hier. Deshalb ist die Rekonstruktion des gesamten Genoms in möglichst genauer Weise nötig, um auf einer detailierteren Ebene einzelne Gene und Genfamilien zu beschreiben sowie die gesamte Genomstruktur

aufzuklären. An Chromosom 2 konnten mehrere wichtige Resultate gewonnen werden, die auch für alle anderen Chromosomen gültig sind. Vor allem die strukturellen Eigenschaften (Centromere, Telomere, Duplikationsmechanismen) konnten nur aus Chromosom 2 Daten wegen der vorhandenen Duplikation abgeleitet werden. Das Dictyostelium Genomprojekt mündete letztendlich in die Publikation aller Chromosomen im Jahre 2005 (Eichinger et al., 2005). Die Assemblierung und Analyse der Chromosomen 1 bis 3 (~20 Mb) beruhen im wesentlichen auf den in vorherigen Kapiteln skizzierten Herangehensweisen für Chromosom 2. Die Interpretation aller Daten, die das Konsortium generierte, wurde wesentlich durch diese Arbeiten erleichtert.

2.8. Zugänglichkeit der Daten

Die Fertigstellung eines Genoms nimmt mehrere Jahre in Anspruch, zumal wenn sich die Arbeiten nicht nur im Sequenzieren erschöpfen, sondern eine möglichst vollständige Repräsentation des Genoms zum Ziel haben. Daten, also Einzelsequenzen, des Genoms stehen jedoch schon von Beginn des Projekts an zur Verfügung. Diese früh und ständig aktualisiert öffentlich zu machen, ist ein wichtiger Teilaspekt jedes Genomprojekts, damit die wissenschaftliche Gemeinschaft davon profitieren kann. Im Falle des deutschen Teils des Dictyostelium Genomprojekts wurde dem Rechnung getragen, indem ein WWW basiertes Informationssystem am Anfang des Projekts installiert und ständig ausgebaut wurde und wird. So ist eine täglich aktualisierte Datenbank aller Rohdaten via Sequenzvergleich (blast) abfragbar. Des Weiteren können Contigs, Genmodelle und weitere Analysen abgerufen werden (http://genome.imb-jena.de/dictyostelium/). Eine genaue Kopie dieser Seiten wurde in Köln eingerichtet. Inzwischen sind unsere Daten auch über die Website der „Dictyostelium Community“ erhältlich (http://dictybase.org/).

Zu Beginn des Genomprojekts verpflichteten wir uns außerdem, alle von uns produzierten Ressourcen und Analyseresultate allen Interessenten zur Verfügung zu stellen. Dementsprechend wurden mehrere hundert Klone weltweit verschickt, die für die jeweiligen Forschergruppen interessante Gene enthielten. Auch die schnelle Isolierung von Genen mittels PCR Methoden war durch die Abfragbarkeit der Daten gewährleistet.

3. Diskussion

3.1. Zweck der Arbeit

Genomanalyse unterscheidet sich wesentlich von der Analyse einzelner Gene nicht nur in der Methodik, sondern auch in ihrem Blickwinkel auf biologische Phänomene. Bei der Betrachtung der Wirkungsweise eines Proteins muss notgedrungen in Kauf genommen werden, dass der sogenannte genomische Hintergrund im Versuchssystem Zelle als vernachlässigbar eingestuft wird. Da jedoch Proteine nicht alleine agieren, sondern in vielfältige Wechselbeziehungen zu weiteren Bestandteilen der Zelle treten, können in Abhängigkeit von diesem genomischen Hintergrund unterschiedliche Ergebnisse resultieren. Besonders auffällig ist dies, wenn Proteine untersucht werden, die die Antwort der Zelle auf die Umgebung optimieren, aber nicht essentiell sind (Glöckner and Beck, 1997). Die Aufgabe der strukturellen Genomanalyse ist es nun, das Potenzial einer Zelle, das in ihrem Genom festgeschrieben ist, zu beschreiben. Sie gibt die Mittel an die Hand, funktionelle Analyse im Kontext des gesamten Genoms zu betreiben.

In einer kürzlichen Untersuchung wurde gezeigt, dass die Ergebnisse genomweiter Analysen „objektiver" sein könnten als die Analyse einzelner Ge ne. Im konkreten Fall untersuchten die Autoren die Daten, die aus Two-Hybrid-Analysen an *S. cerevisiae* gewonnen worden waren. Einzelanalysen legten nahe, dass interagierende Proteine auch öfter in gleicher Weise transkriptionell reguliert werden. Die Ergebnisse, die aus Analysen aller Interaktionen in der Zelle gewonnen wurden, zeigten eine solche Kopplung nicht (Mrowka et al., 2001). Zwei mögliche Begründungen werden für dieses Phänomen angegeben. Einerseits könnten die genomweiten Scans eine sehr hohe Fehlerrate aufweisen, was diesen Ansatz in Frage stellen würde. Andererseits könnte die Aufstellung einer Arbeitshypothese, wie bei an einzelnen Genen durchgeführten Experimenten nötig, statistisch signifikante Zusammenhänge nur simulieren. Wenn letzteres zutreffen sollte, muss die genomweite Analyse Einzelgenanalysen zur Seite gestellt werden, um eine objektive Beurteilung der Ergebnisse zu ermöglichen.

Voraussetzung für die funktionelle Analyse des Genoms ist die Aufklärung der Struktur des Genoms, d. h. die Bestimmung der Basenabfolge und die Erkennung von Genen und sonstiger Eigenschaften sowie deren Verteilung. Da jedes Genom andere Eigenschaften aufweist, muss auch das zur Verfügung stehende Methodenrepertoire den jeweiligen Gegebenheiten angepasst werden. Mit dieser Arbeit soll exemplarisch gezeigt

werden, welcher Art die Erkenntnisse sind, die auf verschiedenen Ebenen der strukturellen Genomanalyse erzielt werden können.

Mit *D. discoideum* wurde für diese Analysen ein Organismus gewählt, der schon lange als Modell für verschiedenste Fragestellungen genutzt wird. Welcher Leistungen schon frühe Eukaryontenzellen fähig waren, läßt sich aus dem Vergleich von lange sich unabhängig entwickelnden Linien extrapolieren. Aus der Berechnung von phylogenetischen Stammbäumen wurde schon lange gefolgert, dass *D. discoideum* auf einem evolutionären Ast sitzt, der tief am Stamm der Eukaryonten ansetzt. Damit ist diese Art ideal geeignet um die Evolution auf dem Weg zu den Metazoa, den Tieren, besser zu beschreiben. Früh verwurzelte Entwicklungstrends konnten also anhand der Genomanalyse von *D. discoideum* aufgespürt werden. Es sollte jedoch nicht Aufgabe dieser Schrift sein, alle Gene und Genfamilien aufzuzählen, die mögliche interessante Forschungswege eröffnen. Vielmehr sollte sie sich auf den allgemeinen Rahmen und einige Highlights beschränken, da viele Einzelaspekte sehr detailiert in anderweitigen Veröffentlichungen dargelegt wurden.

3.2. Verschiedene Sequenzierstrategien

Um die Basenabfolge eines bestimmten genomischen Abschnitts zu bestimmen, müssen viele Einzelsequenzen aus diesem Bereich zusammengesetzt werden. Um die Komplexität dieser Aufgabe zu reduzieren, wurden üblicherweise diese Abschnitte vorher in geeignete Vektoren kloniert oder, wenn die Sequenzabschnitte sehr klein sind, per PCR amplifiziert. So konnte mit relativ wenigen Sequenzstücken umgegangen werden, der Aufwand für die Berechnung der Anordnung der Einzelsequenzen ist relativ gering. Ein großer Teil des öffentlich geförderten Humangenomprojektes war deshalb der exakten Zuordnung von YAC, BAC und Cosmid -Klonen zu Chromosomenpositionen gewidmet. Die kartierten Klone können dann sequentiell sequenziert und assembliert werden („clone by clone“ Strategie). Dem Vorteil der Handhabbarkeit der Sequenzen steht jedoch gegenüber, dass genomische Regionen, die nicht in einer dieser Klonbibliotheken repräsentiert sind, auch nicht analysiert werden können. Wenn zudem die Klonbibliotheken eine zu geringe Repräsentation des Genoms enthalten, können chimäre Klone die Kartierung stören. Die zügige Verbesserung von Rechnerleistungen und Programmen ließ eine weitere Möglichkeit ist Blickfeld rücken: Die reine Shotgun Sequenzierung von

ganzen Genomen (whole genome shotgun, WGS) oder Chromosomen (whole chromosome shotgun, WCS). Hier stellt aber die Assemblierung eine Herausforderung dar. Stören können dabei vor allem repetitive Elemente, aber auch große Genfamilien stellen ein Problem bei der richtigen Zuordnung der Sequenzen zueinander dar. Die teilweise Lösung für dieses Problem besteht in der Verwendung eines Sequenzierungsvektors, der im Gegensatz zum einzelsträngigen Standardvektor M13 die Sequenzgeneration von beiden Klonenden her erlaubt wie pUC18. Die Verwendung von Shotgun Klonbibliotheken mit verschieden langen Kloninserts verbessert hier noch die Ausbeute an Informationen, die zur Anordnung von Sequenzen verwendet werden können. Für die endgültige Zusammensetzung zu ganzen Chromosomen ist allerdings immer noch eine genaue Karte nötig (siehe Integration von Karte und Sequenz).

Die „clone by clone" Strategie konnte mangels erhältlicher BAC oder PAC Klone nicht in Erwägung gezogen werden. So blieb nur die WCS Strategie als durchführbare Alternative. Zum Zeitpunkt dieser Entscheidung gab es nur Vorschläge, dass eine solche Strategie auch bei größeren Genomen als denen von Bakterien angewendet werden könnte (Weber and Myers, 1997). Für das Dictyostelium Sequenzierungsprojekt bedeutete das, dass mit dem Aufnehmen dieser Strategie absolutes Neuland betreten wurde. Um zu einem Erfolg zu kommen, mussten vor allem neuartige Assemblierungsroutinen entwickelt werden, die einerseits mit den großen Datenmengen umgehen, andererseits aber auch Nutzen aus den „read pair" Informationen ziehen konnten. Wegen der hohen „Verunreinigung" der Ausgangsdaten mit Sequenzen aus anderen Chromosomen konnten nicht einfach alle Daten assembliert werden. Eine Vorauswahl an Sequenzen anhand vorhandener genetischer Informationen wurde getroffen, die dann über Ähnlichkeitsanalysen (BLAST) erweitert wurde. Bei dieser Vorauswahl können Fehler auftreten, wenn Sequenzen, die eigentlich zum Chromosom gehören, nicht berücksichtigt werden. Ein Abgleich mit allen anderen Chromosomensequenzen am Ende der Analyse ergab jedoch, dass kein Contig größer als 5 kb einem falschen Chromosom zugeordnet worden war. Mit einiger Sicherheit lässt sich für das gesamte fertiggestellte Genom behaupten, dass neben den repetitiven Sequenzen und den wenigen Sequenzen, die in den Lücken fehlen, alle signifikanten Teile des Genoms repräsentiert sind. Ein Hinweis für die Richtigkeit dieser Aussage ist, dass sowohl die Chromosomengrößen als auch die Gesamtgenomgröße mit vorher abgeschätzten Daten übereinstimmen.

3.3. Genkatalog bei niedriger Redundanz

Die Festlegung auf die WCS Strategie für die Sequenzierung des Genoms erwies sich als vorteilhaft zur Aufstellung eines vorläufigen Genkataloges. Nicht zuletzt durch die „Verunreinigung“ der chromosomenspezifischen Klonbibliotheken mit Klonen, deren Inserts aus anderen Chromosomen stammten, war gewährleistet, dass am gesamten Genom in einer frühen Phase der Sequenzproduktion Analysen durchgeführt werden konnten. Ursprünglich war geplant, einen Genkatalog zu erstellen, der dann auch später als Basis für weitere Analysen genutzt werden könnte. Zunächst wurde untersucht, wie die Sequenzen repetitiver Elemente aus den Rohdaten extrahiert werden könnten. Da sie recht häufig im Genom vorkommen, waren genug Daten schon in einer kleinen Menge an Sequenzen vorhanden. Zwar konnten alle Elemente in einem relativ kleinen Datenpool, der eine genomische Abdeckung von nur 0,4 repräsentierte, gefunden werden. Doch zeigte sich, dass erst mit einer ca. 0,9 fachen Abdeckung des Genoms die Sequenzen der repetitiven Elemente zweifelsfrei und korrekt aufgeklärt werden konnten.

Die Analysen, die zu einer vollständigen Beschreibung der repetitiven Elemente führten, sollten prinzipiell auch auf einzelne Gene übertragbar sein. Wegen der angenommenen statistischen Verteilung der Einzelsequenzen statt einer optimalen Überdeckung des Genoms können aber gar nicht alle Gene getroffen werden. Des Weiteren kommt beim *D. discoideum* Genom hinzu, dass mit starken Klonierungsungleichgewichten zu rechnen war, die die Trefferwahrscheinlichkeit beeinflussen würden. Es stellte sich später heraus, dass diese Ungleichgewichte eher die Wahrscheinlichkeit für Gentreffer erhöhten, da intergenische Regionen etwa 2 fach unterrepräsentiert im Pool der Sequenzierungsrohdaten sind.

Die Kopienzahl der untersuchten Elemente oder Gene spielt eine große Rolle bei deren korrekten Beschreibbarkeit. Die Berechnung der Kopienzahlen unterliegt, wie schon gezeigt, Limitationen, die nicht durchbrochen werden können. So ist es zwar möglich, Genfamilien ab einer Größenordnung von 4 Mitgliedern von Einzelkopiegenen zu unterscheiden. Ob aber ein Gen in ein-, zwei- oder dreifacher identischer oder nahezu identischer Ausführung vorhanden ist, läßt sich nicht exakt genug berechnen. So bietet die Anzahl an Treffern nur einen Hinweis auf die Redundanz des untersuchten Genes. Fatal kann sich diese Abschätzungsungenauigkeit z.B. dann auswirken, wenn versucht wird, „knock out“ Experimente mit Genen zu machen, die trotz einer berechneten Menge von einer Kopie mehrfach im Genom vorliegen.

Eine weitere Limitation eines solchen Genkatalogs neben der nicht hundertprozentigen Trefferwahrscheinlichkeit und der Unsicherheit bei der Kopienzahl liegt in der Tatsache begründet, dass nur Gene mit Ähnlichkeiten zu bereits bekannten Genen gefunden werden können. Zwar könnte man versuchen, Genabschnitte mit Hilfe des großen Unterschieds im A+T Gehalt zwischen kodierender und nicht kodierender Sequenz im *D. discoideum* Genom zu definieren. Doch dieser Ansatz würde zwangsläufig den Wert des Genkatalogs reduzieren. Denn da der Leserahmen nicht bekannt wäre, müsste korrekterweise für jeden potenziell kodierenden Abschnitt sechs Translationsmöglichkeiten als möglicher Genabschnitt in den Katalog aufgenommen werden. Darüber hinaus muss nicht jeder kurze Genomabschnitt mit erhöhtem G/C Gehalt ein Gen repräsentieren.

Ein Vergleich des vorläufigen Genkatalogs mit dem nun vollendeten Genom ergab im Wesentlichen Übereinstimmung, was die generellen Trends und Strukturen in diesem Genom angeht. So wurde die Gendichte korrekt vorhergesagt (Glöckner et al., 2002), die Mitglieder einzelner Genfamilien korrekt definiert (Kollmar and Glöckner, 2003; Rivero et al., 2001; Wilkins et al., 2005), und weitere Analysen bezüglich der Unterscheidung zwischen genischen und intergenischen Regionen damit durchgeführt (Szafranski et al., 2005). Trotz allfälliger Limitationen gibt ein Genkatalog, der bei niedriger Redundanz erstellt wurde, also ein wichtiges Hilfsmittel zu statistischen und anderen Analysen an die Hand. Gleichzeitig erlaubt er einen groben Überblick über das Genom, der begleitend zur Vollendung des Genoms genutzt und verfeinert werden kann.

3.4. Integration von Karte und Sequenz

Zu Beginn des Projekts konnte noch davon ausgegangen werden, dass eine verlässliche Beschreibung des Genoms von *D. discoideum* in Form von integrierten genetischen und physikalischen Karten vorhanden war. Im Laufe der Arbeiten stellte sich jedoch heraus, dass diese Karten so fehlerbehaftet waren, dass sie nur eingeschränkt verwendet werden konnten. Ohne Karten kann jedoch kein Sequenzierprojekt erfolgreich abgeschlossen werden, wie gerade die Entwicklung im Humangenomprojekt zeigte. Zwar wurde von Celera angekündigt, einen „whole genome shotgun“ (WGS) durchzuführen und wesentlich schneller als die öffentlich geförderten Institute, die eine Karten- und BAC-Klonbasierte Sequenzierung durchführten, eine vollständige Version vorzulegen. Doch da die durch die öffentlichen Institute erzeugten Daten frei zugänglich waren, konnte auch der

WGS auf die Karteninformationen bauen. Tatsächlich unterschieden sich dann die beiden vorgelegten Skizzen des Humangenoms nur unwesentlich (Aach et al., 2001). Dort, wo keine BAC Klone vorhanden waren, konnten auch die aus dem gesamten Genom stammenden Sequenzen diese Lücken nicht füllen. Die Annahme, dass mit einer WGS Strategie nur wenige Lücken bleiben würden, die die Rekonstruktion der Chromosomen nicht beeinflussen würden, war also falsch. Das zeigte, dass eine vollständige genomische Karte unerlässlich auch für die Durchführung von Sequenzierungen von ganzen Genomen oder Chromosomen ist.

Genomische DNA von *D. discoideum* ist nur in gewissen Grenzen klonierbar. Nicht nur, dass Fragmente über ca. 5 kb in bakteriellen Systemen instabil sind, war bekannt. Auch mit einer Selektion gegen besonders A/T reiche Sequenzen musste gerechnet werden. Zwar waren schon 2 Chromosomen mit ähnlich hohem A+T Gehalt aus *Plasmodium falciparum* sequenziert, doch dabei handelte es sich um Genomabschnitte im Bereich von einer Mb. Darüber hinaus waren die Kosten für diese Projekte enorm (M-A Rajandream, pers. Mitteilung), da eine sehr hohe Abdeckung des Chromosoms von 15 bis 20 fach angestrebt wurde. Trotzdem blieben nach der Shotgun Phase in diesen Projekten sehr viele Lücken übrig, die mit Hilfe von PCR Methoden und Transposoninsertionssequenzierungen geschlossen wurden. Da die Chromosomen von *D. discoideum* wesentlich größer sind als die bis dahin assemblierten Chromosomen von *P. falciparum*, musste damit gerechnet werden, dass auch nach Abschluss der Arbeiten etliche Lücken ungeschlossen bleiben würden. Jedoch sollte kaum Information über Gene verloren gehen, da die Lücken wegen des hohen A+T Gehalts in intergenischen Regionen zumindest vorwiegend dort angesiedelt sein würden. In der Tat erwies sich später, dass die Lücken, die geschlossen werden konnten, nie Kodierungspotenzial enthielten. Gleichzeitig bestätigte sich bei den erfolgreich geschlossenen Lücken, dass sie im Schnitt längere Abschnitte, die nur aus A und T Nukleotiden bestehen, enthalten als das übrige Genom.

Nach der initialen Assemblierung der Chromosomen 1 bis 3 waren knapp 2000 Contigs vorhanden, die größer als 2 kb waren und den jeweiligen Chromosomen zugeordnet werden konnten. Diese Zuordnung erfolgte über die Kalkulation ihrer Zusammensetzung aus Sequenzen aus den verschiedenen Klonbanken. Die Ausschlussgrenze für eine eindeutige Zuordnung lag dabei bei einer Mindestlänge von 5 kb. Bei kleineren Contigs reichte oftmals die Anzahl der Sequenzen im Contig nicht aus, um eine statistisch signifikante Zuweisung zu treffen. Viele dieser Konfliktfälle konnten aufgelöst werden, indem diese kleinen Contigs mit anderen Contigs zu größeren Einheiten

über die „read pair" Info rmation zusammengefasst wurden. Doch in einigen Fällen blieben Zweifel über die Zuordnung bestehen, die erst mit fortschreitender Lückenschließung ausgeräumt werden konnten. Insgesamt betraf dies etwas weniger als 50 kb an Sequenz.

Über die Sequenzinformation allein würde es schon wegen der Klonlücken nicht möglich sein, ein Chromosom zu rekonstruieren. Die Zuordnung von Chromosomenregionen zu YAC Klonen mittels YAC Skims von ausgewählten YAC Klonen versprach Abhilfe, obwohl diese Methode, wie ebenfalls am *P. falciparum* Projekt erwiesen, sehr aufwendig ist. Ein Pilotprojekt zur HAPPY Map Kartierung des Chromosoms 6 ergab aber erhebliche Diskrepanzen zwischen dieser Karte und der YAC Karte. Die Möglichkeit der Zuordnung von Regionen über die YACs musste deshalb wieder in Frage gestellt werden. Die sichere Zuordnung von Sequenzen zu distinkten genomischen Regionen ist jedoch unabdingbar, um einen Überblick über die Eigenschaftsverteilungen auf dem Chromosom zu erhalten. Eine genauere YAC Karte zu erstellen verbot sich aus Zeitgründen von selbst. Statt dessen wurde die Lage einzelner Contigs auf YAC Klonen mittels PCR bestimmt. Der Vorteil: Diese Methode ist schnell zu realisieren. Außerdem ist sie, wenn die YAC Klonbibliothek eine hinreichende Abdeckung des Genoms von mindestens 5 fach aufweist, tolerant gegen chimäre Klone. Von Nachteil ist, dass mit vertretbarem Aufwand nur die Zuordnung von größeren Einheiten zueinander möglich ist. Diese größeren Kopplungsgruppen mittels der HAPPY Map Methode zu schaffen, übernahmen J. Pachebat und P. Dear in Cambridge. Ein Problem, das sich dabei ergab, war die Definition von Markerregionen. Zwar konnten wir die schon assemblierten Sequenzen dafür nutzen, diese Regionen auszusuchen. Diese ausgewählten Regionen wurden dann auch auf Einmaligkeit im Genom hin untersucht. Doch zeigte sich, dass nicht alle in diesen Markerregionen ausgewählten Primer einmalig im Genom vertreten waren oder dass es Kreuzreaktionen mit ähnlichen Sequenzen gab. Dies mag mit bedingt sein durch den hohen A+T Gehalt des Genoms, der relativ niedrige Annealing-Temperaturen bei der PCR erfordert. Es gelang uns nicht, durch geeignete Auswahl aus den assemblierten Contigs eine Gleichverteilung der Markerregionen zu erreichen. Vielmehr konnten Abstände zwischen Markerregionen von mehr als 100 kb statt der angestrebten 10 kb – 15 kb auftreten.

Ein Vorteil der Markerdefinition anhand vorhandener chromosomenspezifischer Sequenzen ist, dass die Kopplungsanalyse vereinfacht wird. Zur Konstruktion einer Karte müssen nicht Marker aus anderen Chromosomen erst ausgeschlossen werden. Man kann davon ausgehen, dass alle Marker im Pool in die selbe Kopplungsgruppe gehören. Ein

Nachteil der zielgerichteten Markerwahl ist, dass möglicherweise Regionen nicht berücksichtigt werden, die eigentlich zum Chromosom gehören. Dies gilt vor allem für kleinere Contigs, deren Zuordnung zum Chromosom unsicher ist. Um diese Fehlerquelle möglichst auszuschließen, wurden auch für Contigs zwischen 5 und 10 kb, deren Zuordnung unsicher war, Marker definiert.

Die genetische Karte wurde für die Zusammenfassung der HAPPY Kopplungsgruppen verwendet. Die genetischen Karten beruhten auf Analysen, die mit Hilfe eines sogenannten „parasexuellen" Kreislaufs gewonnen wurden. In klassischen Kreuzungsanalysen konnten zu wenig lebende Nachkommen erzeugt werden, um Rekombinationshäufigkeiten festzustellen. In einem parasexuellen Zyklus werden zunächst Diploide aus verschiedenen, selektierbare Marker tragenden Stämmen hergestellt. In geeigneten Medien überleben nur Zellen, die beide Marker tragen, also diploid sind. Wird der Selektionsdruck entfernt, können die Diploiden durch Zellteilungen in Gegenwart von Mikrotubuli destabilisierenden Agentien wieder haploide Nachkommen erzeugen. Die untersuchten Marker segregieren dabei in die haploiden Zellen. Gene können so leicht den einzelnen Kopplungsgruppen zugewiesen werden. Ein großer Nachteil dieser Methode ist, dass nur mitotische Rekombinationsereignisse oder Rearrangements zu Aufschlüssen über die Abfolge der Gene auf den Chromosomen beitragen können. Diese Ereignisse sind irregulär und führen in der haploiden Generation nicht unbedingt zu Chromosomenstrukturen, wie sie die Ursprungsstämme besaßen. Deshalb können sie nicht vollkommmen mit einer meiotischen Rekombination gleichgesetzt werden. Daraus folgt, dass die resultierenden genetischen Karten recht grob bleiben müssen und manche Gene falsch positioniert werden. In der Tat zeigte sich, dass einige wenige Gene, die auf der genetischen Karte weit voneinander entfernt lagen, in den Sequenzen der Kopplungsgruppen benachbart waren. Die Konsequenz daraus war, dass nur dann genetische Information benutzt werden konnte, wenn mindestens 2 Gene widerspruchsfrei sowohl in der genetischen Karte als auch in den Kopplungsgruppen lokalisiert werden konnten.

Für die Chromosomen 1 und 3 konnte je eine einzige Kopplungsgruppe definiert werden. Für Chromosom 2 konnten jedoch nur 2 Kopplungsgruppen etabliert werden, die nicht verbunden werden konnten. Auch frühere genetische Kartierungen wiesen hier eine Lücke auf. Diese Region muss größer als ~100 kb sein, da dies die obere Grenze für die Detektion einer Kopplung, bedingt durch die Versuchsbedingungen, war. Wie die Analysen zeigen enthält diese Region eine lange Duplikation und Strukturen, wie wir sie als

Telomere und Centromere definiert haben (siehe 3.6). Trotz dieser einen unüberbrückbaren Lücke in Chromosom 2 können alle Chromosomen als vollständig angesehen werden. Nur mit diesen genau geordneten Sequenzen war eine weitere Analyse der Chromosomenenden und der Verteilung von Repetitiven Elementen sowie eine repräsentative Darstellung der Chromosomen möglich.

3.5. Repetitive Elemente

Alle repetiven Elemente zeigen einen ausgeprägten Verteilungsdualismus innerhalb des Genoms. Während die einen Elemente (DIRS, skipper, Tdd) ausschließlich oder fast ausschließlich auf die Centromerregion beschränkt sind, sind andere Elemente (TRE) an das Vorhandensein von tRNA Genen in ihrer Nachbarschaft gebunden. Wenn andere Elemente als TREs jenseits der Cenromerregion aufgefunden werden, sind sie fast immer mit TRE Genen assoziiert und bilden Cluster von repetitiven Elementen (Winckler et al., 2002; Winckler et al., 2005). Die Anhäufung von KREs in distinkten Loci kann auch in anderen Spezies beobachtet werden. Man vermutet, dass Sprünge in andere, genhaltige Positionen den Organismus negativ beeinflussen können. Die dadurch einsetzende Selektion gegen „ungünstige“ Insertionspunkte führt dann dazu, dass über mehrere Generationen die Träger solcher „ungünstiger“ KRE Orte wieder aus der Population verschwinden. So können nur noch an bestimmten Stellen Loci von repetitiven Elementen beobachtet werden.

Dieses Erklärungsmuster müsste jedoch erlauben, dass DIRS Elemente in schon vorhandene TREs oder andere Cluster hineinspringen können. Da dies nicht geschieht sorgt möglicherweise ein aktiver Clearingmechanismus für das Entfernen von DIRS und anderen Elementen aus dem kodierenden Teil der Chromosomen. Alternativ denkbar wäre, dass alle DIRS Elemente inaktiv sind. In der Tat gibt es nur wenige Hinweise für eine Transkription dieses Elements (Cohen et al., 1984). Initiale Analysen von anderen sozialen Amöben zeigen jedoch, dass DIRS Elemente erst in *D. discoideum* zur vollen Verbreitung gekommen sind (unveröffentlichte Daten). Da diese Expansion eher nach der Verbreitung der TREs erfolgte, ist negative Selektion auf Grund schädlicher Einflüsse auf die Fitness eher unwahrscheinlich.

3.6. Chromsomenenden

Eine Suche nach möglichen Telomerrepeatsequenzen in den Rohdaten verlief erfolglos. Das kann daran liegen, dass *D. discoideum*, ähnlich wie *D. melanogaster* keine solchen Repeats besitzt. Andererseits könnten sie so kurz sein, dass sie nicht ohne weiteres klonierbar sind. Jedoch fiel beim Suchen nach ungewöhnlichen Sequenzmotiven auf, dass es eine begrenzte Anzahl an Sequenzen gab, die sich aus rDNA Palindrom Anteilen und chromosomaler DNA zusammensetzten. Eine weitere Untersuchung ergab, dass sie zwanglos paarweise den einzelnen Chromosomen zugeordnet werden konnten. Da manche dieser Contigs centromerspezifische KREs enthalten, postuliere ich, dass jedes Chromosom von einer rDNA Palindromsequenz an beiden Enden abgeschlossen wird. Da es keinerlei Polymorphismen zwischen den Palindromsequenzen unterschiedlicher Herkunft gibt, muss ein reger Austausch zwischen Chromosomenenden und Palindrom zum Zwecke des Sequenzabgleichs stattfinden. Damit würde das amplifizierte rDNA Palindrom als Reservoir für die Generierung neuer Enden der Chromosomen dienen. Potentiell könnte die Non-Homologous End Joining Maschinerie hier involviert sein, aber auch völlig andere Mechanismen, die durch neue Proteinfunktionen ausgeführt werden, sind vorstellbar. Diese Genomorganisation ist völlig neu, jedoch nicht völlig überraschend, da z.B. *D. melanogaster* KREs zur Erhaltung der Chromosomenenden verwendet (Abad et al., 2004).

Kürzlich wurden die subtelomeren Regionen im Humangenom näher untersucht (Linardopoulou et al., 2005). Die Autoren finden, dass hier häufige Austausche zwischen den Chromosomen stattfinden. Der für *D. discoideum* postulierte Mechanismus könnte also bis zum Menschen hin konserviert sein.

3.7. Junge Duplikationen

Die Unfähigkeit der vorhandenen Laborstämme, nach einer Kreuzung mit dem entsprechenden Paarungspartner lebende Nachkommen in genügender Zahl hervorzubringen, könnte ein Hinweis auf eine hohe Instabilität des Genoms sein. Untersuchungen an verschiedenen Karyotypen haben gezeigt, dass unterschiedliche Genom- und Chromosomengrößen auftreten können (Cox et al., 1990b; Loomis and Kuspa, 1997). Da diese Unterschiede bei Laborstämmen auftraten, die noch nicht sehr lange selektiert worden waren, ist dies ebenfalls ein Hinweis für ein flexibles Genom. Für das

Chromosom 2 von AX4, den zu sequenzierenden Stamm, haben wir eine inverse Duplikation von über 700 kb beschrieben, die erst kürzlich stattgefunden haben musste.

Einen ersten Hinweis auf die Art des Zustandekommens der Duplikation lieferte die Erkenntnis, dass rDNA-palindromische Sequenzen an der Erhaltung der Chromosomenenden beteiligt sein könnten. Eine Suche nach solchen Sequenzen, die chromosomale sowie Palindromabschnitte enthielt, lieferte für jedes Chromosom ein Paar, für Chromosom 4, das eine vollständige rDNA Kopie enthält, vier Enden und für Chromosom 2 ebenfalls vier Enden. Da aus anderen Untersuchungen klar war (fluorescent in situ hybridisation; R. Sucgang, persönliche Mitteilung), dass keine rRNA Gene auf Chromosom 2 lokalisiert sind, mussten diese Stücke ehemalige Chromosomenenden repräsentieren. Da sie in der Nachbarschaft der Duplikation auftauchen, kann diese Duplikation zwanglos mit dem in Abbildung 9 beschriebenen Szenario erklärt werden. Die Entstehung der Duplikation lässt sich im Stammbaum der Laborstämme auf nur 30 Jahre Lebensdauer (jedoch mit unbestimmter Generationszahl) einschränken.

Es konnten weitere, relativ kleine Duplikationen auf Chromosom 2 observiert werden, die keine Gene umfassten. Sie stammen von allen Chromosomen und wirken wie Einsprengsel in die Chromosomenstruktur. Auch hier sind bemerkenswert wenige Polymorphismen zu finden, was ein Hinweis auf jüngere Insertionen ist. Insgesamt wurden im Verhältnis zur Zahl der jüngeren Insertionen kaum Insertionen mit hohen Polymorphismenraten gefunden. Das mag daran liegen, dass solche Insertionen sehr schnell mutieren und damit nicht mehr als dupliziert auffallen. Andererseits könnte es bei der Etablierung des Stammes in jüngerer Vergangenheit zu mehreren solcher Insertionsereignisse gleichzeitig gekommen sein. Die Anwesenheit der großen Duplikation ist ein starkes Indiz dafür, das es bei der Stammetablierung von AX4 auch weitere, das Genom formende Ereignisse gegeben hat. Insgesamt deuten alle genannten Aspekte darauf hin, dass *D. discoideum*, zumindest in seiner Laborform, ein sehr flexibles Genom besitzt.

3.8. Basenungleichgewichte

Genomische Basenungleichgewichte treten in den verschiedensten Organismen auf (Glöckner, 2000). Wie in der Einleitung angeführt, wird über die Ursachen und Wirkungen dieser Ungleichgewichte noch gestritten. Da wir nun mit *D. discoideum* (DD) und *P. falciparum* (PF) zwei Genome mit sehr ausgeprägten Basenpräferenzen vergleichen

können, könnten sich Hinweise zur Lösung dieses Problems ergeben. Trotz vergleichbarer A+T Gehalte unterscheiden sich diese beiden Organismen erheblich in ihren Genomen (Szafranski et al., 2005). Zum Einen ist die Gendichte in PF wesentlich geringer als in DD. Zum zweiten sind die Gene in PF wesentlich länger als in anderen Organismen. Wie der Versuch einer Genvorhersage mit verschiedenen Programmen zeigt, unterscheiden sich beide Organismen auch hinsichtlich der Eigenschaften, mit denen Gene identifiziert werden können. Somit scheint der hohe A+T Gehalt die Geneigenschaften weniger zu prägen als ihnen als sekundäre Eigenschaft übergestülpt zu sein (Szafranski et al., 2005). Ein weiteres interessantes Merkmal der Gene in DD sind die sehr langen, nur aus einem einzigen A+T reichen Kodon bestehenden Sequenzabschnitte, die in Protein umgesetzt werden. Diese langen „Homopolymer Runs" sind die längsten, die bis jetzt in einem Genom festgestellt wurden (Abb. 4). Offensichtlich schaden sie dem Organismus nicht, während manche Krankheiten des Menschen, wie z.B. Chorea Huntington (Kremer et al., 1994) gerade durch solche „triplet repeat expansions" hervorgerufen werden. Diese A+T reichen Regionen in den Genen addieren sich, je nach benutzten Schwellenwert für die Berechnung, auf bis zu 1 % des Genoms und tragen deshalb signifikant zum Gesamt A+T Gehalt bei. Da diese Regionen bevorzugt in bestimmte Aminosäuren umgesetzt werden, findet eine Selektion sicherlich auf Proteinebene statt (Eichinger et al., 2005). Interessanterweise sind diese Regionen sogar länger als in PF, möglicherweise auch deswegen leistet sich DD eine etwas weniger auf A+T reiche Kodonen fixierte Kodierung als PF. Eine detailierte Analyse zeigt nämlich, dass der zusätzlich gegenüber DD erhöhte A+T Anteil (1,3 %), über den PF verfügt, nicht auf längeren intergenischen Regionen oder Introns oder erhöhtem A+T Gehalt in denselben beruht, sondern durch konsequentere Nutzung von A+U reichen Kodonen in den kodierenden Regionen erzielt wird. Daraus ergeben sich folgende Schlussfolgerungen: i) Es gibt ein oberes Limit von ca. 87 % für den durchschnittlichen A+T Gehalt, der in nichtkodierenden Regionen erreicht werden kann. Je länger ein solcher Bereich, desto weniger A+T darf er enthalten. ii) teilweise ist die Akkumulation von A+T ein interner Prozess, der von präferentiell bestimmte Nukleotide verwendenden Proteinen getrieben wird. Sonst wäre der teilweise erhebliche Unterschied in der Basenzusammensetzung sehr nahe verwandter Arten mit wahrscheinlich ähnlichem Selektionsdruck (unveröffentlichte Daten) nicht zu erklären. iii) zumindest in der Phase nach der Etablierung des A+T Reichtums kommen Selektionskräfte zum Zuge. Anders wäre ein Eingriff in das Kodierungspotential wie bei PF kaum vorstellbar.

3.9. Annotation von Genen

In prokaryontischen Genomen ist die Genvorhersage relativ einfach. Alle Sequenzabschnitte, die mit einem Startkodon (ATG oder GTG) beginnen und mit einem der drei Stopkodonen enden, werden als offene Leserahmen (ORF) definiert. Je länger diese ORFs sind, desto wahrscheinlicher ist es, dass sie Gene darstellen. Zufällige ORFs sollten wegen der statistischen Wahrscheinlichkeit des Auftretens eines Stopkodons relativ kurz sein. Jedoch lässt sich keine untere Ausschlussgrenze für ORF -Größen angeben, weil auch sehr kurze Gene existieren (z.B. psaM *Cyanidium caldarium* mit 30 Aminosäuren (Glöckner et al., 2000)). Prokaryontische Gene können auch überlappen, aber dabei handelt es sich meist um nur wenige Basen. Auch kann manchmal nicht das erste mögliche Startkodon, sondern erst das zweite oder dritte genutzt werden, was den ORF etwas verkürzt. Generell können bei Prokaryonten Gene gleichgesetzt werden mit allen nicht oder nur sehr wenig überlappenden ORFs, die eine gewisse Länge übersteigen. Deshalb ist deren Auffinden mit einfachen Mitteln zu bewältigen.

Ganz anders bei Eukaryonten. Hier können die kodierenden Regionen eines Genes zersplittert über mehrere Exonen verteilt sein. Im Extremfall kann die Region, die ein einziges Gen enthält, mehrere Mb lang sein. Exonen selber können sehr kurz, aber auch sehr lang sein. Durch alternatives Spleissen können aus einer als Gen definierten Region tausende Proteinvarianten entstehen (Graveley, 2001). Die Auffindung von Genen ist hier also nicht über die Definition von ORFs möglich. Vielmehr wird in den meisten Fällen von Generkennungssoftware eine Matrize eines idealen Genes aus den Strukturen bereits bekannter Gene erstellt. Mit Hilfe dieser Matrize wird dann jede Region auf eine Übereinstimmung oder Ähnlichkeit mit dieser Matrize hin überprüft.

Die Erstellung dieser Matrize zerfällt in mehrere Schritte: Zunächst muss die Wahrscheinlichkeit dafür bestimmt werden, ob ein bestimmter Genomabschnitt ein Kodierungspotenzial besitzt oder nicht. Da kodierende Regionen einem gewissen selektiven Druck zur Nutzung bestimmter Triplettkodonen unterliegen, um die Herstellung funktionierender Proteine zu ermöglichen, können sie am gehäuften Auftreten ansonsten weniger genutzter Kodonen erkannt werden. Im Falle von *D. discoideum* ist dies relativ einfach. Die Verschiebung der Nukleotidverteilung des gesamten Genoms hin zu einem erhöhten A/T Gehalt lässt gleichzeitig die Unterschiede zwischen kodierenden und nicht kodierenden Regionen schärfer hervortreten. So weisen intergenische Regionen sowie Introns einen durchschnittlichen A/T Gehalt von 86 bzw. 87 % aus, während die

kodierenden Regionen mit 72 % A/T einen deutlich erhöhten Anteil an G/C Nukleotiden zeigen. Dieser Unterschied ist so groß, dass er selbst beim oberflächlichen Betrachten einer *D. discoideum* Sequenz ins Auge fällt. Damit wird die Einteilung der Sequenz in kodierende und nichtkodierende Abschnitte ähnlich einfach wie bei Prokaryonten.

Im Gegensatz dazu stehen Genome mit einer annähernd gleichen Verteilung an Nukleotiden wie z.B. das menschliche Genom. Hier sind die Unterschiede zwischen kodierenden und nicht kodierenden Sequenzen nicht sehr stark ausgeprägt, obwohl auch hier gehäufte Nutzungen bestimmter Kodonen in kodierenden Regionen nachgewiesen werden können. Zusätzliche Anhaltspunkte für Kodierungspotenzial können in diesen Genomen Ähnlichkeiten zu bereits bekannten Proteinen bieten. In der Tat wird dies von manchen Generkennungsprogrammen genutzt (Solovyev and Salamov, 1997).

Kodierungspotenzial alleine macht einen Sequenzabschnitt noch nicht zu einem Teil eines Genes. Vielmehr müssen weitere Eigenschaften eines Genes hinzutreten. Diese Eigenschaften werden in einem weiteren Schritt untersucht. Intron/Exon Grenzen haben eine stark konservierte Basenabfolge in allen Organismen. Der Donor Bereich enthält ein GT, der Akzeptor ein AG. Darüber hinaus finden sich weitere relativ häufige Basen in der Umgebung von Donor und Akzeptor wie z.B. ein C vor dem Akzeptor AG. Vor dem Akzeptor befindet sich normalerweise außerdem ein Bereich variabler Länge, der stark mit Pyrimidinen angereichert ist. Da die Exon/Intron Grenzen jedoch im Kontext des Genoms relativ kurz sind, können diese Basenabfolgen natürlich öfter auftreten, jedoch funktionslos sein. Erst die Kombination von Kodierungspotenzial und das Vorhandensein eines durchgehendes Leserasters nach dem Ausspleißen eines potenziellen Intronen ist ein starker Hinweis auf die Anwesenheit eines Genes. Man sollte annehmen, dass wegen des relativ kleinen Genoms bei gleichzeitig hoher Gendichte und kurzen Intronen die korrekte Erkennung von Spleissorten in *D. discoideum* kein Problem darstellen sollte. Wahrscheinlich aufgrund des sehr hohen A/T Gehalts des Genoms sind nur rudimentäre Spleisssignale vorhanden. Sie beschränken sich auf die oben genannten Donoren (GT) und Akzeptoren (AG). Darüber hinaus sind bei den meisten Genmodellen keinerlei konservierte Basen an Intron/Exon Grenzen auszumachen. Allenfalls eine leichte Häufung des Donormotivs GTAAGT und von Thymidinen an der Akzeptorstelle war bei einer späteren Untersuchung aller gefundenen Genmodelle festzustellen. Zwei konservierte Basen sind jedoch zu wenig, um in jedem Fall die richtige Spleißstelle zu finden. Dies schränkt den Wert dieser Stellen für die korrekte Definition von Genmodellen ein.

Der Bereich vor Startkodonen ist meist etwas A+T reicher als der Durchschnitt. Wenn Kodierungspotenzial kurz hinter einer solchen A+T reicheren Stelle in Verbindung mit einem Startkodon auftritt, kann von einem Genstart ausgegangen werden. Diese Geneigenschaft ist jedoch bei Betrachtung des *D. discoideum* Genoms wertlos, da nichtkodierende Regionen hier immer sehr A+T reich sind. Zusammenfassend läßt sich sagen, dass zwar die Erkennung von kodierenden Abschnitten im *D. discoideum* Genom ähnlich einfach ist wie in prokaryontischen Genomen. Durch die eingeschränkte Spezifität von Genstart-Regionen und Intron/Exon Grenzen jedoch wird die Vorhersage der korrekten Struktur von Genen erschwert.

Genvorhersageprogramme nutzen die oben beschriebenen Eigenschaften von Genen, um möglichst korrekt Genmodelle vorzuschlagen. Die Qualität der Vorhersagen wird mit zwei Parametern, der Spezifität und der Sensitivität, beschrieben. In anderen Worten: Ein Programm sollte möglichst alle Gene erkennen (Sensitivität), aber dabei möglichst wenig falsche Vorhersagen machen (Spezifität). Die vorhandenen Programme unterscheiden sich hinsichtlich dieser beiden Eigenschaften, was einerseits an den jeweils zugrunde liegenden Algorithmen liegt. Eine Rolle spielen aber auch die verwendeten Schwellenwerte. Wenn ein Programm möglichst alle Gene erfassen soll, werden die Vorhersagen unspezifischer, die falsch positiven Resultate steigen. Umgekehrt geht eine erhöhte Spezifität zu Lasten der Sensitivität, echte Gene werden nicht erkannt. Dieses Dilemma bei der automatischen Generkennung ist bei der Nutzung eines einzigen Programmes nicht zu umgehen. Bessere Ergebnisse sind zu erzielen, wenn man mehrere Programme in die Analyse einbezieht (Glöckner et al., 1998; Taudien et al., 2000).

Alle Programme müssen allerdings jeweils an die Geneigenschaften des zu untersuchenden Organismus angepasst werden, um eine möglichst hohe Sensitivität und Spezifität bei der Vorhersage zu erreichen. Das wird dadurch erreicht, dass die Strukturen bereits bekannter Gene dieses Organismus als Trainingsset für das jeweilige Programm verwendet werden. Je größer das Trainingsset, desto besser sollten die zu erwartenden Vorhersagen sein, da mit einer großen Menge an Daten auch etwas abweichende Genstrukturen erfasst werden.

Zwar waren ca. 300 Gene von *D. discoideum* bekannt, jedoch von nur 140 Genen war die vollständige Genstruktur aufgeklärt worden. So war auch die Möglichkeit, Programme zu trainieren, auf diesen Datensatz eingeschränkt. Ein weiterer Nachteil dieses Trainingssets war es, dass es hauptsächlich aus Genen bestand, die in Signaltransduktion und Aufbau und Funktion des Zytoskeletts involviert sind. Diese beiden Forschungsgebiete

machen einen Hauptteil der Arbeiten mit *D. discoideum* aus. Es ist nicht auszuschließen, dass die Strukturen dieser Gengruppen einem Selektionsdruck unterliegen, der mit ihrer Funktion gekoppelt ist. Somit könnte die hier vorgestellte Genvorhersage auf einer Matrize beruhen, die nicht allgemeine Genstrukturen des Organismus wiederspiegelt. Somit würde die Genstrukturanalyse nicht in allen Fällen exakte Voraussagen treffen können. Da aber das Kodierungspotenzial der Gene mehrheitlich erfasst wurde, wie ein Abgleich mit vorhandenen EST Sequenzen ergab, bleiben die Aussagen zur Funktionsanalyse der Gene jedoch von dieser potentiellen Fehlerquelle unberührt.

Die zunächst mit GlimmerM durchgeführten Genvorhersagen mussten notgedrungen ungenau bleiben, weil dieses Programm für uns nur hinsichtlich des Kodierungspotenzials trainierbar war. Im Vergleich zu dem später verwendeten Programm waren die vorhergesagten Genmodelle durchgehend kürzer. Da wir mit der Gruppe um R. Guigo in Barcelona zur Erstellung einer für *D. discoideum* spezifischen GeneID Version kooperieren konnten, wurden dann alle weiteren Vorhersagen mit diesem Programm getroffen. Zusätzlich zu einem Training wurde es noch einer Feinabstimmung unterzogen. Ausgewählte Genmodelle wurden auf Stimmigkeit hin überprüft. Einerseits können dafür EST Daten herangezogen werden, andererseits zeigen orthologe Gene die vermutlich richtige Genstruktur an. Wegen der schieren Zahl der Genmodelle musste mit Stichproben vorlieb genommen werden. Manche später untersuchten Genmodelle konnten mit vorhergehenden oder nachfolgenden Genmodellen, die auf dem gleichen Strang kodiert sind, verschmolzen werden, wenn als Vergleich ein orthologes Gen aus einem anderen Organismus zur Verfügung stand. Da jedoch in den meisten untersuchten Fällen die Genmodelle der mRNA Struktur oder ESTs bzw. Orthologen entsprach, gehe ich davon aus, dass die Vorhersagen insgesamt vertrauenswürdig sind.

Ein späterer Vergleich mit einem weiteren trainierten Programm, das von A. Krogh entwickelt worden war, zeigte weitgehend übereinstimmende Resultate. Auffällig war jedoch, dass gerade an Spleißstellen und im 5´ Bereich der Genmodelle Diskrepanzen zwischen den Vorhersagen der beiden Programme bestanden. Dies ist ein Hinweis dafür, dass diese Signale wegen ihrer relativ schwachen Ausprägung von den beiden Programmen unterschiedlich bewertet und interpretiert werden. Welches der Programme eine höhere Vorhersagegenauigkeit an diesen Stellen hat, lässt sich aus den vorliegenden Daten nicht bestimmen. Ein vorhergesagtes Genmodell kann zwar als Orientierungshilfe für die weitere Analyse dienen, man kann dieses auch nutzen, um ein genomisches Microarray für die Transkriptanalyse herzustellen. Doch wenn Gene ausgeschaltet, überexprimiert oder sonstig

verändert werden sollen, ist eine genaue Kenntnis der Genstruktur unerlässlich. Auch die Analyse von Promotoren kommt ohne diese Kenntnis nicht aus. Eine Überprüfung aller Genmodelle durch mRNA Sequenzierungen erscheint deshalb unausweichlich.

3.10. Gene und Domänen

D. discoideum ist der erste freilebende Protist, von dem nun das gesamte Genom bekannt ist. Überraschend ist der Befund, dass dieser meist einzellig lebende Organismus eine Menge von Genen benötigt, die vergleichbar ist mit der eines wesentlich komplexeren Tieres, von *Drosophila melanogaster*. Diese Menge könnte natürlich aus vielen Genduplikationen oder einer Genomduplikation hervorgegangen sein. Die Clusterung von allen Genen von Chromosom 2 ergab jedoch, dass nur ca. 12 % dieser Gene eine so weitgehende Ähnlichkeit zu anderen Genen aufweisen, dass sie leicht erkennbar aus der Duplikation eines Vorgängers herrühren. Interessanterweise sind das meistens Gene, die im Tandem auf dem Genom angeordnet sind. Eine Analyse eines Genes mit mehreren Kopien auf Chromosom 2 (Eichinger et al., 2005) zeigte auch, dass mit zunehmender Distanz vom Ursprungsgen die Ähnlichkeit rasch abnimmt. Diese Diversifizierung kann natürlich auch mit einer Änderung der Funktion einhergehen. Im Falle von identischen Genfamilien wie z.B. den Aktinen ist auch die Beibehaltung der Funktion anzunehmen. Selbst bei Annahme identischer Funktionen von weit diversifizierten Gentandems wäre die hohe Anzahl an Genen nicht mit einer hohen Funktionsredundanz erklärbar.

Der Anteil an funktional redundanten Genen und Gengruppen könnte einen erheblichen Einfluss auf die Genzahl haben. Streng genommen müsste also der Satz an einzigartigen Funktionsträgern für einen Vergleich der Genmengen herangezogen werden. Solange aber die Funktionen aller Genprodukte nicht aufgeklärt sind, läuft man Gefahr, Genprodukten innerhalb einer Familie die gleiche Funktion zuzuordnen und damit die Zahl der für den Organismus unabdingbaren Gene zu unterschätzen. Deshalb mag für diese Betrachtung die absolute Genzahl als Grundlage ausreichen. Natürlich kann die Menge an Genen nur ein grober Orientierungspunkt sein, doch mit der Anzahl an Genen erhöht sich die Anzahl der möglichen Verknüpfungen der Genprodukte untereinander. Damit würde bis zu einem gewissen Grade die Anzahl an Genen auch die Komplexität bzw. Organisationsstufe eines Organismus beschreiben, auch wenn sie sich möglicherweise unter Laborbedingungen nicht ausprägt.

Die schiere Zahl an Genen weist also *D. discoideum* als einen Organismus aus, der wesentlich komplexer ist als Hefe. Diese Komplexität manifestiert sich z.B. in der höheren Zellbeweglichkeit und der Möglichkeit, Phasen des Lebens als koordinierter Mehrzeller zu verbringen. Die berechnete Anzahl von über 12.000 Genen im *D. discoideum* Genom ist nahe an den Werten von echten Vielzellern wie *D. melanogaster*. Vielzelligkeit, oder die Fähigkeit dazu, könnte also einen großen Teil des Mehr an Genen erfordern, das wir in *D. discoideum* im Verhältnis zu reinen Einzellern beobachten können. Mehrzelligkeit wurde in der Evolution öfter unabhängig voneinander erfunden. Es ist schwer vorstellbar, dass diese Fähigkeit sich jeweils aus dem Nichts entwickelt hat, vielmehr ist anzunehmen, dass auch der letzte gemeinsame Vorfahre aller Eukaryonten schon das Potenzial dafür gehabt hat. *D. discoideum* steht nun ausweislich des Stammbaumes zwar nicht an der Basis der Entwicklung von Pflanzen und Tieren, aber diese Spezies hat sich sehr früh von den Hauptästen abgetrennt. Durch einen Vergleich des Genoms von *D. discoideum* mit den Genomen von tierischen Vielzellern konnten ~1000 evolutionär konservierte Gene definiert werden, die nur in der Linie der Metazoa auftreten. Zu diesen gehören sicherlich auch die Gene, die ursprüngliche Voraussetzungen für tierische Vielzelligkeit geliefert haben (unveröffentlichte Daten).

Die Analyse von Domänen, also besonders konservierten Teilen von Proteinen, bietet einen vertiefenden Blick jenseits der Erkenntnisse, die durch die Analyse von Orthologen erzielt werden können. Metazoa verfügen generell über mehr verschiedene Domänen als Hefe, wenn auch jeweils verschiedene Genfamilien in den einzelnen Organismen amplifiziert sein können (http://www.ebi.ac.uk/proteome/). Die im Vergleich zu Metazoa ähnliche Vielfalt an Domänen bei *D. discoideum* könnte wiederum durch die freilebend räuberische Lebensart und die komplexe Entwicklung eines Dauerstadiums aus einer vielzelligen Assoziationsvorstufe bedingt sein. Lebensbedingungen scheinen also den bewahrten Gensatz zu beeinflussen. Ein Blick auf die Frequenz von Interpro Domänen zeigt außerdem, dass manche Domänen wesentlich häufiger in *D. discoideum* vorkommen als in anderen, vollständig analysierten Organismen. Diese Häufungen stellen sicherlich spezielle Anpassungen des Genoms an die Erfordernisse dieses Organismus dar. Gerade Hefen scheinen im Laufe der Evolution etliche Gengruppen, die für ihre Lebensweise nicht benötigt wurden, als Ballast über Bord geworfen zu haben. Anhand des Vergleichs von vorhandenen Interpro Domänen kann dies leicht festgestellt werden (Tabelle 7). Auffällig ist, dass sehr viele der in Hefe fehlenden Domänen mit der Etablierung von Vielzelligkeit und Differenzierungen zu tun haben. Es finden sich des Weiteren im Genom von *D.*

discoideum auch Domänen, von denen bislang mangels vorhandener Daten gedacht wurde, dass sie Erfindungen der Metazoa sind (Eichinger et al., 2005). *D. discoideum* bietet also die Möglichkeit, evolutionäre Entwicklungslinien weiter zurück zu verfolgen als bisher. Offensichtlich benutzt *D. discoideum* den selben Werkzeugkasten wie mehrzellige Tiere zur Etablierung seiner Lebensweise. In jedem Modellorganismus fehlen auch Domänen und Gene, die sonst allgemein verbreitet sind. So besitzt *S. cerevisiae* z.B. nicht das komplexe Zytoskelett von *D. discoideum*, das für Zellbewegungen nötig ist (Eichinger et al., 1999; Noegel and Schleicher, 2000). Abwesenheit von Domänen vor allem in niederen Eukaryonten würde eher auf einen Verlust hinweisen als auf die mehrmalige unabhängige Erfindung der selben Domänen. Da auch in nahe verwandten Arten diesbezüglich Unterschiede bestehen können, empfiehlt sich die Analyse eines möglichst breiten Spectrums an Arten, um einen detaillierten Überblick über Entwicklungslinien zu bekommen.

Die Verteilung von Proteinähnlichkeiten zwischen den bis jetzt analysierten Spezies erhärtet die These, dass *D. disoideum* den Metazoa näher steht als den Pflanzen. Damit steigt auch die Relevanz, die Untersuchungen an Genen haben, die in veränderter Form beim Menschen zu Krankheiten führen können. Besonders interessant sind hierbei Gene, die nicht in Hefe vorkommen. Da *D. discoideum* fast gleichwertige molekulare Untersuchungsmöglichkeiten wie Hefe bietet, können auch in diesem System sehr schnell Genwirkungen untersucht werden. Eine parallele Untersuchung von Funktionen von Genprodukten in Hefe und in *D. discoideum* bietet sich an, um systemunabhängige Erkenntnisse über Funktionen gewinnen zu können.

3.11. *D. discoideum* als Modell

D. discoideum ist der erste freilebende Vertreter der Protozoa (dieser Begriff ist keine taxonomische Einheit, da sich dahinter sehr diverse Organismen verbergen), dessen Genom vollständig entschlüsselt wurde. Dass es gerade dieser Organismus war, der an erster Stelle gewählt wurde, ist kein Zufall. Studien an diesem Organismus hatten schon früher zu allgemeineren Erkenntnissen geführt. Während die vegetative Entwicklung dieses Organismus zunächst als Kuriosum angesehen wurde, konnte später gezeigt werden, dass

hier viele Gene eine Rolle spielen, die Orthologe in anderen Organismen haben. Spezialisierungen innerhalb einer Organismengruppe können also trotzdem eine genetische Grundlage haben, die von vielen Spezies auf anderen evolutionären Ästen geteilt wird. Die Erkenntnis, dass auch niedere Organismen Gene haben, deren Orthologe im Menschen in mutierter Form Krankheiten auslösen können, ist nicht neu (Friedberg, 1985). Um solche Gene zu studieren kann deshalb auf leichter zugängliche Organismen als den Menschen ausgewichen werden. Da aber Proteinfunktionen je nach Organismus variieren können, sollten solche Studien an mehreren Modellen durchgeführt werden. Die Aufschlüsselung des Genoms liefert die Basis für weitergehende funktionale Analysen. Die Kenntnis aller Gene des Organismus ermöglicht erst die Konzeption von Experimenten, die die Wirkungsweise ganzer Gengruppen untersuchen sollen. Es ist heute kaum vorstellbar, dass Forschung zu grundlegenden Fragen in der Biologie in Zukunft an Organismen durchgeführt werden kann, deren Genom nicht vollständig entschlüsselt wurde.

Die Genomanalyse von *D. discoideum* erweitert den Blick auf Gemeinsamkeiten und Unterschiede zwischen den Gensätzen von Modellorganismen auf verschiedenen Entwicklungsstufen. Bis jetzt stehen jedoch nur wenige vollständig analysierte Eukaryontengenome für eine vergleichende Analyse zur Verfügung. Deshalb sind Untersuchungen von Genomen weiterer niederer Eukaryonten mit unterschiedlichen Lebensbedingungen geboten. Einerseits könnte dadurch der allen Eukaryonten gemeinsame Gensatz besser definiert werden, andererseits könnte das Verständnis der Funktionen speziesspezifischer Genprodukte wachsen. Vergleiche von mehr als jetzt zur Verfügung stehenden Genomen würden auch klären helfen, wie die zellulären Funktionen auf Genomebene in verschiedenen Organismen abgestimmt werden. Momentan wird vergleichende Genomanalyse vor allem im Bereich der Vertebraten durchgeführt, aber auch weitere Nematoden und Insekten werden untersucht. Da die Genomanalyse von *D. discoideum* nun abgeschlossen ist, wäre es an der Zeit, weitere soziale Amöben wie z.B. Polysphondyllium oder eine weiter entfernte Art aus der Gruppe der Mycetozoa für Vergleichszwecke heranzuziehen.

3.12. Das *D. discoideum* Genomprojekt im Vergleich zum Humangenomprojekt

Das Humangenomprojekt wurde vor knapp 20 Jahren initiiert. In diesem Projekt sollten auch gleichzeitig Techniken und Methoden entwickelt werden, die für jedes Projekt anwendbar wären. Es brachte, gerade in letzter Zeit, einen gewaltigen Fortschritt bei der Produktion und Assemblierung von Daten zu Wege. Die Assemblierung und automatische Annotation von BAC Klonen in der Größenordnung von 100 bis 200 kb ist inzwischen Standard (Glöckner et al., 1998; Momeni et al., 2000; Riboldi Tunnicliffe et al., 2000; Taudien et al., 2000). Es erscheint also als natürlich, die Standardtechniken, die so erfolgreich am Humangenom angewandt wurden, auch auf das viel kleinere *D. discoideum* Genom anzuwenden. Doch wie sich herausstellte, galten für das *D. discoideum* Projekt völlig andere Bedingungen als für das Humangenomprojekt. Teilweise waren diese Bedingungen dem Umstand geschuldet, dass *D. discoideum* ein A+T reiches Genom besitzt. So konnte keine stabile bakterielle Klonbibliothek hergestellt werden, wie sie für das klonbasierte Sequenzieren verwendet wird. YAC Klone stellen keine leicht zu verwendende Alternative dar, weil deren DNA nicht in hinreichenden Mengen sauber von der genomischen Hefe-DNA des Wirtes zu trennen ist. Damit war es nicht möglich, dem Weg der akademischen Zentren zu folgen, den sie erfolgreich für die Humangenomsequenzierung einschlugen.
Die WCS Strategie dagegen kann nur mit einer genauen Chromosomenkarte erfolgreich sein. Daran ermangelte es zu Beginn des Projekts. Diese Karte musste erst erstellt werden, nachdem der größte Teil der Sequenz schon vorhanden war.

Im Humangenomprojekt ist die Annotation von Genen jeweils an die einzelnen BAC oder PAC Kloncontigs gebunden. Dementsprechend fallen für eine Arbeitsgruppe relativ wenig Annotationsarbeiten an. Das gesamte Chromosom 21 des Menschen z.B. enthält etwas mehr als 250 Gene. Die Annotation teilten sich die 3 beteiligten Sequenzierzentren (Tsukuba, Jena, Berlin) auf, wobei später die Resultate abgeglichen und vereinheitlicht wurden. Die Annotation der Gene der *D. discoideum* Chromosomen stellte eine ungleich größere Herausforderung dar. Zum Beispiel enthält das Chromosom 2 von *D. discoideum* mit 2799 Genen mehr als 10 mal so viele Gene wie Chromosom 21 des Menschen. Annotationen einiger hundert Gene können noch in Tabellenform aufbewahrt und gesichtet werden. Bei mehreren tausend Genen wird es nötig, eine Datenbankstruktur aufzubauen. Sie erlaubt gezielte Abfragen, so dass der Nutzer besser von den Informationen

profitieren kann. Je größer die Menge an Genen ist, die analysiert werden sollen, desto mehr muss auf computerisierte Analyse zurückgegriffen werden. Im Humangenomprojekt übernehmen spezialisierte Institute wie das EBI in Hinxton die Aufbereitung und Aktualisierung der Daten. Im Gegensatz dazu existierte für das *D. discoideum* Projekt kein Institut, das die Erstellung, Pflege und Verfügbarmachung der Daten übernommen hätte. So war meine Arbeitsgruppe in der Lage, von der initialen Sequenzierung über die Assemblierung und Annotation bis zur Datenanalyse und Aufbereitung alle für ein Genomprojekt nötigen Schritte zu übernehmen. Somit kann an diesem Modell gezeigt werden, wie die Planung und Koordinierung aller Arbeiten in einem Genomanalyseprojekt durchgeführt werden.

3.13. Funktionelle Analyse

Die Herstellung und automatische Analyse des Genkatalogs eines Organismus ist die Voraussetzung für die genomweite funktionelle Analyse. Automatische Generkennung unterliegt aber Limitationen, die nicht aufgehoben werden können. Ein automatisch gewonnener Gensatz kann deshalb nur bis zu einem gewissen Grade richtig sein. Nicht nur Genanfänge und -Enden sind schwierig zu bestimmen, auch fehlen manchmal Gene in den Datensätzen, was durch falsche Vorhersagen oder „frame shifts“ bedingt ist, die wiederum von Sequenzfehlern herrühren. Deswegen schließt sich an die hier gezeigte Analyse die manuelle Bearbeitung des Datensatzes an. Mit Hilfe von passenden ESTs und vollständigen mRNA Sequenzen müssen alle Genmodelle verifiziert und die Sequenzen in der Datenbank aktualisiert werden. Darüber hinaus müssen Experten für die Annotation einzelner Genfamilien gefunden werden, die z.B. Unsicherheiten betreffs einer Genstruktur besser beurteilen können als jedes Computerprogramm (Galperin and Koonin, 1998). Einbindung von Literaturhinweisen und weiteren wichtigen Informationen in die Proteindatenbank erleichtert dann deren Nutzung. Zur Zeit ist die manuelle Annotation der *D. discoideum* Gene noch in den Anfängen, erste Vereinbarungen über die Vorgehensweise hierbei wurden getroffen. Erst wenn der vollständige Genkatalog auf diese Weise annotiert worden ist, besteht die Gewähr dafür, dass für die Herstellung von Expressionschips für jedes Genmodell passende PCR -Produkte bzw. Oligos generiert werden. Trotzdem können schon vorher Expressionschips für einen Teil der Gene hergestellt und mit unter verschiedenen Bedingungen isolierten mRNAs hybridisiert werden. Die für die

Sequenzierungsarbeiten verwandten Klone stellen eine billige Ressource für solche Chips dar. Deswegen wurden schon erste Versuche mit diesen Klonen zur Herstellung von Expressionschips unternommen.

Von großem Interesse wird sein, Mutanten verschiedener Gene herzustellen, und die Änderungen ihrer globalen Expression gegenüber dem Wildtyp zu untersuchen. Vor allem die Untersuchung von Genen, die in Vertebraten, aber nicht in Hefe vorkommen, könnte tiefere Einblicke in die Funktion von komplexeren Zellen und Zellsystemen liefern. Mit der Analyse des Chromosoms 2 von *D. discoideum* wurden erste Voraussetzungen für die eingehende funktionelle Analyse des gesamten Organismus geschaffen. Zwei weitere Chromosomen werden gerade analysiert. Die aus diesen Arbeiten resultierenden Daten zu aktualisieren und zu verbessern, aber auch funktionell zu analysieren, werden die Aufgaben in der Zukunft sein.

4. Zusammenfassung

1. **Sequenzierung und Assemblierung.** Das gesamte Genom von *D. disocideum* konnte durch Integration von Sequenzdaten mit Kartierungsdaten assembliert werden. Die Chromosomen 1 bis 3 sind insgesamt durch weniger als 50 Lücken unterbrochen und die Sequenzen selber sind größtenteils mehrfach überprüft worden. Die hochgenaue Sequenz ist nun der Ausgangspunkt für weitere funktionelle Analysen.
2. **Basenzusammensetzung.** Das Genom von *D. discoideum* besitzt eine ungewöhnliche Basenzusammensetzung. Aus diesem Grund konnten Standardmethoden für die Genomanalyse nicht angewandt werden. Vielmehr mussten etliche Methoden und Programme entwickelt werden, die eine Rekonstruktion von Chromosomen dieses Organismus erlaubten. Zu nennen ist hier vor allem die Assemblierungsroutine. Ein Vergleich mit *P. falciparum*, dessen A+T Gehalt noch etwas höher liegt als der von *D. discoideum*, zeigte, dass unterschiedliche Strategien zur Erreichung dieses Basenungleichgewichtes in den beiden Organismen verfolgt werden. Zudem scheint eine Obergrenze für den A+T Reichtum abhängig von der Länge der betroffenen Region zu bestehen. Sowohl ein ursprünglicher „mutational bias" als auch spätere Selektionsvorteile könnten für das beobachtete Basenungleichgewicht verantwortlich sein.
3. **Genomstruktur.** Chromosom 2 enthält eine inverse Duplikation von über 700 kb, die nach einem Chromosomenbruch entstanden ist. Die Zusammensetzung der Bruchstelle (DIRS Elemente und rDNA Palindromsequenzen) gibt Hinweise auf die Struktur von Centromerregionen und subtelomere Regionen aller Chromosomen. Vermutlich enden alle *D. discoideum* Chromosomen in Sequenzen, die aus dem rDNA Palindrom stammen. Dies legt einen Mechanismus nahe, der das amplifizierte Palindrom als Reservoir zur Herstellung von Chromosomenenden benutzt.
4. **Repetitive Elemente und Transposonanalyse.** Mit dieser Arbeit wurde gezeigt, dass eine Überblicksanalyse eines Genoms erfolgen kann, ohne dass eine hohe, saturierende Abdeckung des Genoms erreicht werden müsste oder eine Assemblierung nötig wäre. Für die Auffindung von repetitiven Elementen,

Genfamilien oder auch Genen, die Ähnlichkeiten zu Genen in anderen Organismen aufweisen, haben wir Einzelsequenzen genutzt. Wir konnten damit unter anderem eine detailierte Analyse aller Transposonen in diesem Genom durchführen. Die spätere Assemblierung zeigte dann, dass die unterschiedlichen Transposonenklassen jeweils scharf auf bestimmte genomische Regionen beschränkt sind. Diese scharfe Abgrenzung ist möglicherweise durch einen akiven Clearingmechanismus zustande gekommen.

5. **Gene.** 2/3 des Genoms eines niederen Eukaryonten wurden assembliert und mit Hilfe selbst entworfener Software annotiert. Diese Entwicklungen wurden auch für die Analyse des gesamten Genoms eingesetzt. Mit 2,7 kb je Gen hat *D. discoideum* eine sehr hohe Gendichte, die nur von Hefe übertroffen wird. Es war überraschend zu sehen, dass ein so einfacher Organismus eine so große Zahl von Genen besitzt. Frühere Schätzungen gingen von 30 % weniger Genen aus. Obwohl die reine Anzahl wenig über die Menge an wirklich benötigten Gene aussagt (es könnte eine gewisse Funktionsredundanz bestehen) ist doch zu vermuten, dass die Steuerung und Ausformung der verschiedenen Zell-Zustände und –Differenzierungen erheblich komplexer ist als vorher vermutet.
6. **Domänen.** Viele Domänen wurden in vorhergesagten Genmodellen gefunden, die in Hefe nicht existieren. Diese Domänen umfassen unter anderem Funktionseinheiten, die für Signaltransduktion, Beweglichkeit und Zell-Zell Kontakte zuständig sind. Für viele Fragestellungen, die über die Analyse von Stoffwechselwegen und anderen Basisfunktionen aller eukaryontischen Zellen hinausgehen, scheint deshalb *D. discoideum* besser als einfachstes Modell geeignet zu sein als *S. cerevisiae*.
7. **Triplet Expansionen.** Viele kodierende Regionen sind charakterisiert durch sequentiell wiederholte Triplets, bevorzugt aus AAT oder AAC bestehend. Dies spiegelt sich in den Proteinen in der Anwesenheit von Abschnitten, die nur aus einer Aminosäure bestehen. Offensichtlich stellen diese Expansionen eine Strategie dar, wie innerhalb kodierender Regionen der A+T Gehalt angehoben werden kann. *P. falciparum* ist hier einen anderen Weg gegangen, dieser Organismus nutzt häufiger A+T reiche Kodonen.
8. **Beschreibungen von Genfamilien.** Einige ausgewählte Genfamilien konnten wir erschöpfend unter Ausnutzung der Rohdaten beschreiben. Unter anderem charakterisierten wir die Familien der Rho-ähnlichen Proteine, Kinesine und Aktine.

Einige Mitglieder der Familien weisen Besonderheiten wie z.B Domänenzusammenstellungen auf, die in keinem anderen Organismus anzutreffen sind. So ist zu vermuten, dass diese sich als spezielle Adaptationen des Organismus gebildet haben.

9. **Die Stellung von *D. discoideum* in der Evolution.** Ein Vergleich der Proteine, die Ähnlichkeiten mit den anderen bis jetzt analysierten Modellorganismen aufweisen, bestätigt, dass *D. dicoideum* an der Basis der Entwicklung von höheren Tieren (Metazoa) steht, obwohl dieser Organismus auch manche für Pflanzen spezifische Gene enthält. Vom letzten gemeinsamen Vorfahren aller Eukaryonten auf dem Weg zu heutigen Pflanzen und Tieren traten also differentielle Verluste auf. *D. discoideum* als nahe an der Verzweigung der beiden Hauptgruppen ebenfalls eine eigene Gruppe bildend zeigt dies deutlich.
10. **Gemeinsame Gene und Entwicklungslinien.**Vergleiche der Proteinsätze von Genomen mehrerer Modellorganismen ergaben, dass alle eukaryontischen Lebewesen nur ungefähr 2500 Gene gemeinsam zu haben scheinen. Da Verluste von Funktionen, die für spezialisierte Lebensweisen nicht benötigt werden, ein wesentlicher Aspekt von Evolution ist, kann davon ausgegangen werden, dass der erste Eukaryont eine Mischung von Genen aus den heutigen Entwicklungslinien besessen hat. Selbst innerhalb einer Gruppe von Organismen können differentielle Genverluste auftreten. Illustriert wird dies z.B. dadurch, dass Spalthefen einige Gene, die auch in Metazoa anzutreffen sind, noch enthalten, während *S. cerevisiae* diese verloren hat. Zur Abschätzung der Vielfalt innerhalb einer Entwicklungslinie müssen deshalb mehrere Organismengenome zum Vergleich zur Verfügung stehen. Separate Entwicklungslinien können dagegen aufdecken, welche Gene wann in der Evolution erstmals aufgetreten sind. Das Genom von *D. disocideum* zeigt, dass etliche Gene, die als Erfindung der Metazoa angesehen wurden, schon in diesem einfachen Organismus vorhanden sind.

5. Literatur

Aach, J., Bulyk, M. L., Church, G. M., Comander, J., Derti, A., and Shendure, J. (2001). Computational comparison of two draft sequences of the human genome. Nature *409*, 856-859.

Abad, J. P., De Pablos, B., Osoegawa, K., De Jong, P. J., Martin-Gallardo, A., and Villasante, A. (2004). Genomic analysis of Drosophila melanogaster telomeres: full-length copies of HeT-A and TART elements at telomeres. Mol Biol Evol *21*, 1613-1619. Epub 2004 May 1626.

Aboobaker, A. A., and Blaxter, M. L. (2000). Medical significance of Caenorhabditis elegans. Ann Med *32*, 23-30.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Allex, C. F., Baldwin, S. F., Shavlik, J. W., and Blattner, F. R. (1996). Improving the quality of automatic DNA sequence assembly using fluorescent trace-data classifications. Proc Int Conf Intell Syst Mol Biol *4*, 3-14.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410

Anjard, C., and Loomis, W. F. (2002). Evolutionary analyses of ABC transporters of Dictyostelium discoideum. Eukaryot Cell *1*, 643-652.

Baldauf, S. L., and Doolittle, W. F. (1997). Origin and evolution of the slime molds (Mycetozoa) [see comments]. Proc Natl Acad Sci U S A *94*, 12007-12012

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*, 972-977.

Bates, P. F., and Swift, R. A. (1983). Double cos site vectors: simplified cosmid cloning. Gene *26*, 137-146.

Bernardi, G. (1993). The vertebrate genome: isochores and evolution. Mol Biol Evol *10*, 186-204

Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. M., Devlin, K., Feltwell, T., *et al.* (1999). The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum [see comments]. Nature *400*, 532-538

Brefeld, O. (1869). Dictyostelium mucoroides. Ein neuer Organismus aus der Verwandtschaft der Myxomyceten. Abhandlungen der Senckenbergischen Naturforschenden Gesellschaft Frankfurt *7*, 85-107

Brefeld, O. (1884). Polysphondylium violaceum und Dictyostelium mucoroides nebst Bemerkungen zur Systematik der Schleimpilze. Untersuchungen aus dem Gesammtgebiete der Mykologie *6*, 1-34

Camargo, A. A., Fischer, K., and Lanzer, M. (1997). Construction and rapid screening of a representative yeast artificial chromosome library from the Plasmodium falciparum strain Dd2. Parasitol Res *83*, 87-89

Cappello, J., Handelsman, K., and Lodish, H. F. (1985). Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell *43*, 105-115

Cashdollar, L. W., Chmelo, R., Esparza, J., Hudson, G. R., and Joklik, W. K. (1984). Molecular cloning of the complete genome of reovirus serotype 3. Virology *133*, 191-196.

Cohen, S. M., Cappello, J., and Lodish, H. F. (1984). Transcription of Dictyostelium discoideum transposable element DIRS-1. Mol Cell Biol *4*, 2332-2340

Cox, D. R., Burmeister, M., Price, E. R., Kim, S., and Myers, R. M. (1990a). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science *250*, 245-250.

Cox, E. C., Vocke, C. D., Walter, S., Gregg, K. Y., and Bain, E. S. (1990b). Electrophoretic karyotype for Dictyostelium discoideum. Proc Natl Acad Sci U S A *87*, 8247-8251

De Rijk, P., Van de Peer, Y., Van den Broeck, I., and De Wachter, R. (1995). Evolution according to large ribosomal subunit RNA. J Mol Evol *41*, 366-375.

Dear, P. H., and Cook, P. R. (1993). Happy mapping: linkage mapping using a physical analogue of meiosis. Nucleic Acids Res *21*, 13-20

Depraitere, C., and Darmon, M. (1978). Growth of "Dictyostelium discoideum" on different species of bacteria. Ann Microbiol (Paris) *129 B*, 451-461.

Devine, S. E., Chissoe, S. L., Eby, Y., Wilson, R. K., and Boeke, J. D. (1997). A transposon-based strategy for sequencing repetitive DNA in eukaryotic genomes. Genome Res *7*, 551-563

Eichinger, L., Lee, S. S., and Schleicher, M. (1999). Dictyostelium as model system for studies of the actin cytoskeleton by molecular genetics. Microsc Res Tech *47*, 124-134.

Eichinger, L., Pachebat, J. A., Glockner, G., Rajandream, M. A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., *et al.* (2005). The genome of the social amoeba Dictyostelium discoideum. Nature *435*, 43-57.

Escalante, R., and Vicente, J. J. (2000). Dictyostelium discoideum: a model system for differentiation and patterning. Int J Dev Biol *44*, 819-835.

Eyre-Walker, A., and Hurst, L. D. (2001). The evolution of isochores. Nat Rev Genet *2*, 549-555.

Felder, M., Szafranski, K., Lehmann, R., Eichinger, L., Noegel, A. A., Platzer, M., and Glockner, G. (2005). DictyMOLD-a Dictyostelium discoideum genome browser database. Bioinformatics *21*, 696-697. Epub 2005 Jan 2028.

Feldman, W., and Pevzner, P. (1994). Gray code masks for sequencing by hybridization. Genomics *23*, 233-235.

Firtel, R. A., and Bonner, J. (1972). Characterization of the genome of the cellular slime mold Dictyostelium discoideum. J Mol Biol *66*, 339-361.

Firtel, R. A., and Meili, R. (2000). Dictyostelium: a model for regulated cell movement during morphogenesis. Curr Opin Genet Dev *10*, 421-427.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., and et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd [see comments]. Science *269*, 496-512

Francis, D. (1998). High frequency recombination during the sexual cycle of Dictyostelium discoideum. Genetics *148*, 1829-1832

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., *et al.* (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi [see comments]. Nature *390*, 580-586

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., and et al. (1995). The minimal gene complement of Mycoplasma genitalium [see comments]. Science *270*, 397-403

Freiberg, C., Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal, A., and Perret, X. (1997). Molecular basis of symbiosis between Rhizobium and legumes. Nature *387*, 394-401.

Friedberg, E. C. (1985). Nucleotide excision repair of DNA in eukaryotes: comparisons between human cells and yeast. Cancer Surv *4*, 529-555.

Galperin, M. Y., and Koonin, E. V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol *1*, 55-67.

Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., *et al.* (1998). Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum [published erratum appears in Science 1998 Dec 4;282(5395):1827]. Science *282*, 1126-1132

Gentles, A. J., and Karlin, S. (2001). Genome-scale compositional comparisons in eukaryotes. Genome Res *11*, 540-546.

Glöckner, G. (2000). Large Scale Sequencing and Analysis of AT Rich Eukaryote Genomes. Current Genomics *1*, 289-299

Glöckner, G., and Beck, C. F. (1997). Cloning and characterization of LRG5, a gene involved in blue light signaling in Chlamydomonas gametogenesis. Plant J *12*, 677-683

Glöckner, G., Eichinger, L., Szafranski, K., Pachebat, J. A., Bankier, A. T., Dear, P. H., Lehmann, R., Baumgart, C., Parra, G., Abril, J. F., *et al.* (2002). Sequence and analysis of chromosome 2 of Dictyostelium discoideum. Nature *418*, 79-85.

Glöckner, G., Rosenthal, A., and Valentin, K. (2000). The structure and gene repertoire of an ancient red algal plastid genome. J Mol Evol *51*, 382-390.

Glöckner, G., Scherer, S., Schattevoy, R., Boright, A., Weber, J., Tsui, L. C., and Rosenthal, A. (1998). Large-scale sequencing of two regions in human chromosome 7q22: analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci reveals 17 genes. Genome Res *8*, 1060-1073

Glöckner, G., Szafranski, K., Winckler, T., Dingermann, T., Quail, M. A., Cox, E., Eichinger, L., Noegel, A. A., and Rosenthal, A. (2001). The complex repeats of Dictyostelium discoideum. Genome Res *11*, 585-594.

Gloeckner, G., and Beck, C. F. (1995). Genes involved in light control of sexual differentiation in Chlamydomonas reinhardtii. Genetics *141*, 937-943

Goebl, M. G., and Petes, T. D. (1986). Most of the yeast genomic sequences are not essential for cell growth and division. Cell *46*, 983-992.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Gottmann, K., and Weijer, C. J. (1986). In situ measurements of external pH and optical density oscillations in Dictyostelium discoideum aggregates. J Cell Biol *102*, 1623-1629.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics *17*, 100-107

Habata, Y., Urushihara, H., Fang, H., and Yanagisawa, K. (1991). Possible existence of a light-inducible protein that inhibits sexual cell fusion in Dictyostelium discoideum. Cell Struct Funct *16*, 185-187.

Hagele, S., Kohler, R., Merkert, H., Schleicher, M., Hacker, J., and Steinert, M. (2000). Dictyostelium discoideum: a new host model system for intracellular pathogens of the genus Legionella. Cell Microbiol *2*, 165-171.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., *et al.* (2000). DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature *406*, 477-483.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res *9*, 868-877.

Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., and Venter, J. C. (1999). Global transposon mutagenesis and a minimal Mycoplasma genome. Science *286*, 2165-2169.

International Human Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H., and Hackett, P. B. (1999). Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. J Mol Evol *48*, 13-21.

Jing, J., Lai, Z., Aston, C., Lin, J., Carucci, D. J., Gardner, M. J., Mishra, B., Anantharaman, T. S., Tettelin, H., Cummings, L. M., *et al.* (1999). Optical mapping of Plasmodium falciparum chromosome 2. Genome Res *9*, 175-181

Jing, J., Reed, J., Huang, J., Hu, X., Clarke, V., Edington, J., Housman, D., Anantharaman, T. S., Huff, E. J., Mishra, B., *et al.* (1998). Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. Proc Natl Acad Sci U S A *95*, 8046-8051

Jurka, J. (1998). Repeats in genomic DNA: mining and meaning. Curr Opin Struct Biol *8*, 333-337

Kapler, G. M. (1993). Developmentally regulated processing and replication of the Tetrahymena rDNA minichromosome. Curr Opin Genet Dev *3*, 730-735.

Kay, R. R. (2000). Development at the edge of multi-cellularity: Dictyostelium discoideum. Int J Dev Biol *44*, 35-38.

Kay, R. R., and Williams, J. G. (1999). The Dictyostelium genome project: an invitation to species hopping. Trends Genet *15*, 294-297

Kessin, R. H. (1997). The evolution of the cellular slime molds. In Dictyostelium - A model system for cell and developmental biology., Y. Maeda, K. Inouye, and I. Takeuchi, eds. (Tokyo, Japan, Universal Academy Press), pp. 3-13

Kollmar, M., and Glöckner, G. (2003). Identification and phylogenetic analysis of Dictyostelium discoideum kinesin proteins. BMC Genomics *4*, 47.

Konfortov, B. A., Cohen, H. M., Bankier, A. T., and Dear, P. H. (2000). A high-resolution HAPPY map of Dictyostelium discoideum chromosome 6. Genome Res *10*, 1737-1742.

Kremer, B., Goldberg, P., Andrew, S. E., Theilmann, J., Telenius, H., Zeisler, J., Squitieri, F., Lin, B., Bassett, A., Almqvist, E., and et al. (1994). A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. N Engl J Med *330*, 1401-1406.

Kumar, S., and Rzhetsky, A. (1996). Evolutionary relationships of eukaryotic kingdoms. J Mol Evol *42*, 183-193.

Kundig, C., Hennecke, H., and Gottfert, M. (1993). Correlated physical and genetic map of the Bradyrhizobium japonicum 110 genome. J Bacteriol *175*, 613-622.

Kuspa, A., and Loomis, W. F. (1994). REMI-RFLP mapping in the Dictyostelium genome. Genetics *138*, 665-674

Kuspa, A., and Loomis, W. F. (1996). Ordered yeast artificial chromosome clones representing the Dictyostelium discoideum genome. Proc Natl Acad Sci U S A *93*, 5562-5566

Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M., and Wolfe, K. H. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res *27*, 1642-1649

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Leng, P., Klatte, D. H., Schumann, G., Boeke, J. D., and Steck, T. L. (1998). Skipper, an LTR retrotransposon of Dictyostelium. Nucleic Acids Res *26*, 2008-2015

Linardopoulou, E. V., Williams, E. M., Fan, Y., Friedman, C., Young, J. M., and Trask, B. J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature *437*, 94-100.

Linskens, M. H., Grootenhuis, P. D., Blaauw, M., Huisman-de Winkel, B., Van Ravestein, A., Van Haastert, P. J., and Heikoop, J. C. (1999). Random mutagenesis and screening of complex glycoproteins: expression of human gonadotropins in Dictyostelium discoideum. Faseb J *13*, 639-645.

Little, P. F., and Cross, S. H. (1985). A cosmid vector that facilitates restriction enzyme mapping. Proc Natl Acad Sci U S A *82*, 3159-3163.

Lloyd, A. T., and Sharp, P. M. (1992). Evolution of codon usage patterns: the extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae. Nucleic Acids Res *20*, 5289-5295

Loomis, W. F., and Kuspa, A. (1997). The Genome of Dictyostelium discoideum. In Dictyostelium- A Model System for Cell and Developmental Biology (Universal Academic Press Inc.), pp. 15-30

Loomis, W. F., Welker, D., Hughes, J., Maghakian, D., and Kuspa, A. (1995). Integrated maps of the chromosomes in Dictyostelium discoideum. Genetics *141*, 147-157

Malnasi-Csizmadia, A., Woolley, R. J., and Bagshaw, C. R. (2000). Resolution of conformational states of Dictyostelium myosin II motor domain using tryptophan (W501) mutants: implications for the open-closed transition identified by crystallography. Biochemistry *39*, 16135-16146.

Mannhaupt, G., Stucka, R., Ehnle, S., Vetter, I., and Feldmann, H. (1994). Analysis of a 70 kb region on the right arm of yeast chromosome II. Yeast *10*, 1363-1381

Maree, A. F. M., and Hogeweg, P. (2001). How amoeboids self-organize into a fruiting body: multicellular coordination in Dictyostelium discoideum. Proc Natl Acad Sci USA *98*, 3879-3883

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380. Epub 2005 Jul 2031.

Marschalek, R., Borschet, G., and Dingermann, T. (1990). Genomic organization of the transposable element Tdd-3 from Dictyostelium discoideum. Nucleic Acids Res *18*, 5751-5757

McDonald, J. F., Matyunina, L. V., Wilson, S., Jordan, I. K., Bowen, N. J., and Miller, W. J. (1997). LTR retrotransposons and the evolution of eukaryotic enhancers. Genetica *100*, 3-13

McLean, M. J., Wolfe, K. H., and Devine, K. M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol *47*, 691-696

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., *et al.* (2001). A physical map of the human genome. Nature *409*, 934-941.

Milosavljevic, A. (1995). DNA sequence recognition by hybridization to short oligomers. J Comput Biol *2*, 355-370.

Momeni, P., Glöckner, G., Schmidt, O., von Holtum, D., Albrecht, B., Gillessen-Kaesbach, G., Hennekam, R., Meinecke, P., Zabel, B., Rosenthal, A., *et al.* (2000). Mutations in a new gene, encoding a zinc-finger protein, cause tricho-rhino-phalangeal syndrome type I. Nat Genet *24*, 71-74

Morio, T., Urushihara, H., Saito, T., Ugawa, Y., Mizuno, H., Yoshida, M., Yoshino, R., Mitra, B. N., Pi, M., Sato, T., *et al.* (1998). The Dictyostelium developmental cDNA project: generation and analysis of expressed sequence tags from the first-finger stage of development. DNA Res *5*, 335-340.

Mrowka, R., Patzak, A., and Herzel, H. (2001). Is There a Bias in Proteome Research? Genome Research *11*, 1971-1973

Murakami, K., and Takagi, T. (1998). Gene recognition by combination of several gene-finding programs. Bioinformatics *14*, 665-675

Mushegian, A. R., and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A *93*, 10268-10273.

Musto, H., Caccio, S., Rodriguez-Maseda, H., and Bernardi, G. (1997). Compositional constraints in the extremely GC-poor genome of Plasmodium falciparum. Mem Inst Oswaldo Cruz *92*, 835-841

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., *et al.* (2000). A whole-genome assembly of Drosophila. Science *287*, 2196-2204.

Noegel, A. A., and Schleicher, M. (2000). The actin cytoskeleton of dictyostelium: a story told by mutants [In Process Citation]. J Cell Sci *113*, 759-766

Pan, A., Dutta, C., and Das, J. (1998). Codon usage in highly expressed genes of Haemophillus influenzae and Mycobacterium tuberculosis: translational selection versus mutational bias. Gene *215*, 405-413

Parra, G., Blanco, E., and Guigo, R. (2000). GeneID in Drosophila. Genome Res *10*, 511-515.

Queller, D. C., Fortunato, A., and Strassmann, J. E. (2001). Competition between clones in chimeras. Paper presented at: Dicty 2001 (San Diego, Ca, USA)

Raper, K. B. (1935). Dictyostelium discoideum, a new species of slime mold from decaying forest leaves. J Agr Res *50*, 135-147

Raper, K. B. (1941). Developmental patterns in simple slime molds. Growth *5 (Suppl.)*, 41-76

Rebatchouk, D., and Narita, J. O. (1997). Foldback transposable elements in plants. Plant Mol Biol *34*, 831-835

Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., and Bier, E. (2001). A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. Genome Res *11*, 1114-1125.

Resor, L., Bowen, T. J., and Wynshaw-Boris, A. (2001). Unraveling human cancer in the mouse: recent refinements to modeling and analysis. Hum Mol Genet *10*, 669-675.

Riboldi Tunnicliffe, G., Gloeckner, G., Elgar, G. S., Brenner, S., and Rosenthal, A. (2000). Comparative analysis of the PCOLCE region in Fugu rubripes using a new automated annotation tool. Mamm Genome *11*, 213-219

Rietdorf, J., Siegert, F., and Weijer, C. J. (1996). Analysis of optical density wave propagation and cell movement during mound formation in Dictyostelium discoideum. Dev Biol *177*, 427-438.

Rivero, F., Dislich, H., Glöckner, G., and Noegel, A. A. (2001). The Dictyostelium discoideum family of Rho-related proteins. Nucleic Acids Res *29*, 1068-1079.

Robinson, M., Gautier, C., and Mouchiroud, D. (1997). Evolution of isochores in rodents. Mol Biol Evol *14*, 823-828

Sabeur, G., Macaya, G., Kadi, F., and Bernardi, G. (1993). The isochore patterns of mammalian genomes and their phylogenetic implications [published erratum appears in J Mol Evol 1994 May;38(5):547]. J Mol Evol *37*, 93-108

Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J., and Tettelin, H. (1999). Interpolated Markov models for eukaryotic gene finding. Genomics *59*, 24-31

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., and Smith, M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. Nature *265*, 687-695.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463-5467.

Sharp, P. M., Rogers, M. S., and McConnell, D. J. (1984). Selection pressures on codon usage in the complete genome of bacteriophage T7. J Mol Evol *21*, 150-160.

Singer, G. A., and Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol *17*, 1581-1588.

Solovyev, V., and Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Proc Int Conf Intell Syst Mol Biol *5*, 294-302.

Stoyan, T., Gloeckner, G., Diekmann, S., and Carbon, J. (2001). Multifunctional centromere binding factor 1 is essential for chromosome segregation in the human pathogenic yeast Candida glabrata. Mol Cell Biol *21*, 4875-4888.

Sturley, S. L. (2000). Conservation of eukaryotic sterol homeostasis: new insights from studies in budding yeast. Biochim Biophys Acta *1529*, 155-163.

Sueoka, N. (1993). Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. J Mol Evol *37*, 137-153

Surzycki, S. A., and Belknap, W. R. (1999). Characterization of repetitive DNA elements in Arabidopsis. J Mol Evol *48*, 684-691

Szafranski, K., Glöckner, G., Dingermann, T., Dannat, K., Noegel, A. A., Eichinger, L., Rosenthal, A., and Winckler, T. (1999). Non-LTR retrotransposons with unique

integration preferences downstream of Dictyostelium discoideum tRNA genes. Mol Gen Genet *262*, 772-780

Szafranski, K., Lehmann, R., Parra, G., Guigo, R., and Glöckner, G. (2005). Gene organization features in A/T-rich organisms. J Mol Evol *60*, 90-98.

Tanaka, M., Hirai, H., LoVerde, P. T., Nagafuchi, S., Franco, G. R., Simpson, A. J., and Pena, S. D. (1995). Yeast artificial chromosome (YAC)-based genome mapping of Schistosoma mansoni. Mol Biochem Parasitol *69*, 41-51

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res *29*, 22-28.

Taudien, S., Rump, A., Platzer, M., Drescher, B., Schattevoy, R., Gloeckner, G., Dette, M., Baumgart, C., Weber, J., Menzel, U., and Rosenthal, A. (2000). RUMMAGE--a high-throughput sequence annotation system. Trends Genet *16*, 519-520.

The Arabidopsis Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. Science *282*, 2012-2018.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680

Thompson, J. K., Caruana, S. R., and Cowman, A. F. (1999). YAC contigs and restriction maps of chromosomes 4 and 5 from the cloned line 3D7 of Plasmodium falciparum [In Process Citation]. Mol Biochem Parasitol *102*, 197-204

Triglia, T., and Kemp, D. J. (1991). Large fragments of Plasmodium falciparum DNA can be stable when cloned in yeast artificial chromosomes. Mol Biochem Parasitol *44*, 207-211

van Tieghem, M. P. h. (1884). Coenonia, genre nouveau de Myxomycetes a plasmode agrege. Bull Soc Bot Fr *31*, 303-306

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Verra, F., and Hughes, A. L. (1999). Biased amino acid composition in repeat regions of Plasmodium antigens. Mol Biol Evol *16*, 627-633

Weber, J. L., and Myers, E. W. (1997). Human whole-genome shotgun sequencing. Genome Res *7*, 401-409

Wilkins, A., Szafranski, K., Fraser, D. J., Bakthavatsalam, D., Muller, R., Fisher, P. R., Glöckner, G., Eichinger, L., Noegel, A. A., and Insall, R. H. (2005). The Dictyostelium genome encodes numerous RasGEFs with multiple biological roles. Genome Biol *6*, R68. Epub 2005 Jul 2028.

Williams, J. G., and Firtel, R. A. (2000). HAPPY days for the Dictyostelium genome project. Genome Res *10*, 1658-1659.

Winckler, T., Dingermann, T., and Glöckner, G. (2002). Dictyostelium mobile elements: strategies to amplify in a compact genome. Cell Mol Life Sci *59*, 2097-2111.

Winckler, T., Szafranski, K., and Glöckner, G. (2005). Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact Dictyostelium discoideum genome. Cytogenet Genome Res *110*, 288-298.

Zsiros, J., Jebbink, M. F., Lukashov, V. V., Voute, P. A., and Berkhout, B. (1999). Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications. J Mol Evol *48*, 102-111

Danksagung

Diese Arbeit wäre nicht zustande gekommen ohne die Mithilfe vieler Kollegen und Freunde. Danken möchte ich Karol Szafranski und Rüdiger Lehmann für die Programmierung etlicher nützlicher Skripte, für die Realisierung vieler Ideen und fruchtbare Diskussionen. Des Weiteren danke ich Silke Förste für den Einsatz ihres organisatorischen Talents. Sie und auch Nadine Zeisse, Sandra Rothe und Regine Schultz sorgten für den reibungslosen Ablauf der Laborarbeiten. Allen Mitarbeitern der Abteilung Genomanalyse sei für die produktive Arbeitsatmosphäre gedankt.

Bedanken möchte ich mich auch bei den aufeinander folgenden Leitern der Genomanalyse, Herrn Andre Rosenthal und Herrn Matthias Platzer, bedanken, die mir beide die Freiheit für eigenständige Arbeit gewährt haben.

Besonderer Dank gilt Petra für ihre oft strapazierte Geduld und das kritische Lesen des Manuskripts.