Die Transposons im Genom von *Dictyostelium discoideum*

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von Dipl.-Biol. Karol Szafranski
geboren am 25.05.1970 in Wuppertal

Danksagungen

Den Herren Prof. Dr. André Rosenthal und Dr. Matthias Platzer gilt mein Dank für das „schützende Dach“ in der Abteilung für Genomanalyse und ihre Unterstützung während der gut dreijährigen Doktorandenzeit. Herrn Dr. Gernot Glöckner danke ich für sein Vertrauen und seine Unterstützung während der Zusammenarbeit im Genomprojekt. Diese Arbeit wurde finanziell unterstützt durch die Deutsche Forschungsgemeinschaft.

Herrn Dr. Thomas Winckler danke ich für die fruchtbare Zusammenarbeit an Fragestellungen um die TRE-Elemente und die damit verbundenen aufschlussreichen Diskussionen. Mein Dank gilt auch den Kollegen der Abteilung Genomanalyse, die mir vielfach mit Rat und Verbesserungsvorschlägen zur Seite gestanden haben: Cornelia Baumgart, Petra Galgoczy, Dr. Ulrike Gaussmann, Dr. Klaus Huse, Dr. Gernot Glöckner, Niels Jahn, Rüdiger Lehmann, Dr. Matthias Platzer, Prof. Dr. André Rosenthal, Dr. Ruben Schattevoy, Bernd Senf, Dr. Roman Siddiqui.

Dr. Ludwig Eichinger stellte der Arbeitsgruppe freundlicherweise eine cDNA-Bibliothek und genomische DNA von *D. discoideum* zur Verfügung. Dank auch an Dr. Edward Cox und Dr. Mark Quail, die die chromosomenspezifischen Schrotschussbibliotheken herstellten, und an die Kollegen vom Sanger-Centre (Hinxton) sowie vom Baylor College of Medicine (Houston) für die gute Kooperation und den regen Informationsaustausch. Den technischen Mitarbeiterinnen Silke Förste, Sabine Landmann, Sandra Rothe, Regina Schulz und Nadine Zeise danke ich für Ihren tatkräftigen Eifer bei der Laborarbeit und ein angenehmes Klima an der Bench.

Last but not least danke ich meiner Lebensgefährtin Beate Fischer für die entgegengebrachte Motivation und Geduld, mit der sie den Fortgang der wissenschaftlichen Arbeit und das Entstehen dieses Textes begleitet hat.

Jena, 2. Feb. 2002

Kurzfassung

Der Schleimpilz *Dictyostelium discoideum* dient als Modell für zelluläre Bewegung, Signaltransduktion und zelluläre Differenzierungs- und Entwicklungsprozesse. Im Jahr 1998 wurde in internationaler Zusammenarbeit ein Genomprojekt für *D. discoideum* AX4 in Angriff genommen. In der Anfangsphase des Genomprojekts habe ich eine detaillierte Charakterisierung der komplexen repetitiven Elemente, d.h. Transposons, dieses Genoms vorgenommen. Alle zuvor in der Literatur für *D. discoideum* beschriebenen Transposons konnten im sequenzierten Stamm AX4 wieder aufgefunden werden (DIRS-1, H3R, skipper, Tdd-4, TRE3-A, TRE3-B, TRE5-A). Einige neuentdeckte Elemente zeigen eine strukturelle Ähnlichkeit zu den bekannten Transposons und gut charakterisierten Transposongruppen (DCLT-A, DGLT-A, Tdd-5, TRE3-C, TRE3-D, TRE5-B, TRE5-C). Darüber hinaus stelle ich neu aufgefundene transposable Elemente vor, die keine verwandtschaftliche Beziehung zu bekannten Transposons zeigen (DDT-A, DDT-B, DDT-S; thug-S, thug-T). Bei der Elementfamilie DDT handelt es sich offenbar um einen bisher unbekannten Typ von autonomen DNA-Transposons.

Basierend auf Trefferstatistiken von genomischen Schrotschusssequenzen habe ich den Sequenzanteil und Fragmentierungsgrad von Transposonkopien im Genom von *D. discoideum* geschätzt. Die Zahl der Kopienfragmente je Transposonspezies reicht danach von etwa 10 bis 300, der Gesamtgehalt der transposablen Elemente des Genoms beträgt 9,6 %. Ein Teil der Transposons zeigt eine strikte Spezifität für die Insertion in die Nachbarschaft von tRNA-Genen (TRE-Familie und DGLT-A). Für die übrigen Transposons kann ich aus der Untersuchung von Transposonflanken folgern, dass sie in Form von dichten Clustern über das Genom verteilt sind. Insbesondere aggregieren Transposons der gleichen Familie zu genomischen Clustern. Dieses Phänomen konnte ich zur Identifizierung und Analyse neuer Transposonspezies ausnutzen. Daneben werden die Implikationen für die Genomorganisation diskutiert.

Eine Analyse der Polymorphismendichte für die verschiedenen Kopien jedes Transposons in Verbindung mit der Kopiendichte des Transposons im Genom liefert ein quantitatives Maß für die Schwierigkeiten, die bei der Assemblierung repetitiver Loci des Genoms von *D. discoideum* zu erwarten sind. Anhand dieser Berechnungen und ausgewählter Assemblierungsbeispiele zeige ich Möglichkeiten und Grenzen für die fehlerfreie Assemblierung von genomischen Transposonkopien auf. Die Analyseergebnisse an Transposons erweisen sich schließlich als eine wertvolle Ressource für die laufende Assemblierung des Genoms von *D. discoideum* mit einer chromosomenorientierten Schrotschussstrategie.

Abkürzungen

aa	Längenangabe für Polypeptide in Aminosäureresten
Acc.No.	engl. „Accession Number“, eindeutiger, alphanumerischer Identifikationscode einer Sequenz des internationalen Datenbankkonsortiums
bp	Längenangabe für Nukleinsäuren in Basenpaaren, kbp = 10^3 bp, Mbp = 10^6 bp
cDNA	„codierende DNA“. Gemeint ist DNA, die durch reverse Transkription aus einer mRNA entsteht.
CDS	engl. „Coding Sequence“, Sequenzabschnitte von genomischer DNA oder mRNA, die in Polypeptid translatiert wird
g	Erdgravitation, als Maßeinheit für die Kraftwirkung bei der Zentrifugation
gDNA	genomische DNA
h	Stunde(n)
HSP	engl. „High-Scoring Segment Pair“, ein lokales Subalignment als Bestandteil des Ergebnisses einer BLAST-Suche
InDel	Kategorie von Mutationen, die auf Insertion oder Deletion beruhen
ITR	engl. „Inverted Terminal Repeat“.
LTR	engl. „Long Terminal Repeat“.
min	Minute(n)
msDNA	engl. „multi-copy single-stranded DNA“, RT-codierender Transposontyp, der ausschließlich bei Prokaryonten verbreitet ist
ORF	engl. „Open Reading Frame“, offener Leserahmen
PBS	engl. „Primer Binding Site“, Startpunkt für die reverse Transkription bei LTR-Retrotransposons
PCR	engl. „Polymerase Chain Reaction“
PPT	Polypurintrakt
RT	reverse Transkriptase
SNP	engl. „Single-Nucleotide-Polymorphism“, Sequenzdivergenz durch Basenaustausch
TSD	engl. „Target Site Duplication“.
UTR	engl. „Untranslated Region“, Teil eines Transkripts, das nicht proteincodierend ist

Inhaltsverzeichnis

1 EINLEITUNG

1.1 Transposons – ubiquitäre Bestandteile eukaryotischer Genome

Als Barbara McClintock in den 40er Jahren DNA-Elemente entdeckte, die innerhalb des Genoms einer Zelle „springen“ können (MCCLINTOCK 1951), war nicht absehbar, dass diese mobilen Elemente, genannt Transposons, universelle Bestandteile von Genomen sind. Bisher ist unter den entschlüsselten eukaryotischen Genomen keines bekannt, das nicht zu einem gewissen bis beachtlichen Anteil transposable Elemente beherbergt. Die Anteile reichen von 3 % bei der Hefe (KIM ET AL. 1998) über 45 % beim Menschen (INT. HUM. GENOME SEQ. CONS. 2001) und mehr als 70 % beim Mais (KUMAR & BENNETZEN 1999).

Von Genomforschern wurde dieser Genomanteil vielfach als sogenannte „junk DNA“ oder sogar „selfish DNA“ interpretiert (FLAVELL 1995). Aber die ubiquitäre Verbreitung von Transposons spricht gegen eine Anschauung von einem nutzlosem Ballast oder gar molekularem Parasitismus und deutet vielmehr auf einen positiven Beitrag hin, den Transposons zur evolutiven Kompetenz des Wirtsgenoms leisten. Konkret provozieren die zahlreichen genomischen Kopien solcher repetitiven Elemente durch Rekombination häufig Inversionen sowie umfangreiche Duplikationen oder Deletionen und tragen so zur Flexibilität und zum Reorganisationspotenzial eines Genoms bei.

1.2 Gruppen und Charakteristika von Transposons

Aus Sicht der Genomforschung sind Transposons zunächst als repetitive Elemente zu betrachten und sind somit in einem Atemzug mit **Satelliten-DNA** zu nennen. Sie alle haben gemeinsam, dass sie aufgrund ihrer hohen Kopienzahl im Zuge niedrigredundanter Schrotschusssequenzierung eines Genoms bald auffällig werden. Satelliten-DNA hat durch eine monotone Abfolge von Sequenzeinheiten die Eigenschaft, in einem Dichtegradienten scharfe Sedimentationsbanden (sogenannte „Satellitenbanden“) zu bilden. Nach der Länge der repetitiven Einheit werden Mikrosatelliten (wenige Basenpaare) und Minisatelliten (wenige hundert Basenpaare) unterschieden. Minisatelliten sind typische Sequenzstrukturen der Centromere bei den Metazoa (CROLLIUS ET AL. 2000), Mikrosatelliten kommen gleichmäßiger verteilt in allen bisher untersuchten eukaryotischen Genomen vor. Die Entstehung von Satelliten-DNA beruht auf einem „Stolper“-Verhalten von DNA-Polymerasen, und ihre Anreicherung ist vermutlich auf Selektion begründet (NEEF & GROSS 2001, SIBLY ET AL. 2001). Insofern spricht das Verbreitung von Satelliten-DNA für eine funktionelle Rolle in der Genomorganisation, und diese funktionelle Bedeutung kann durch Analogieschluss auch für transposable Elemente postuliert werden.

Im Gegensatz zu Satelliten-DNA sind Transposons wesentlich komplexer aufgebaut, denn sie

codieren Proteine, die ihre Mobilisierung ermöglichen. Transposons können nach ihrer Struktur und ihrem Replikationsmechanismus in zwei Hauptgruppen eingeteilt werden: Retrotransposons und DNA-Transposons; es sind auch die Bezeichnungen Klasse-I- und Klasse-II-Transposons in Gebrauch. **Retrotransposons** replizieren über ein mRNA-Intermediat, das die vollständige Sequenzinformation des Elements trägt. Eine reverse Transkriptase (RT), die der Transposonklasse ihren Namen gibt, erzeugt von dem RNA-Intermediat neue DNA-Kopien, die in das Genom eingefügt werden. Im Gegensatz zu Retrotransposons, die durch einen „Copy-and-Paste"-Mechanismus replizieren, vermehren sich **DNA-Transposons** genaugenommen nicht, sondern verschieben durch Wirkung einer Transposase Transposonkopien von einem genomischen Locus zum anderen. Ihre Transposition erfolgt also nach einem „Cut-and-Paste"-Mechanismus.

Das mit großer öffentlicher Aufmerksamkeit verfolgte Humangenomprojekt hat leider wenig zu einem breiteren Verständnis von Transposontypen beigetragen. Die eingeführten kategorialen Begriffe **LINEs und SINEs** sind im organismenübergreifenden Kontext unbrauchbar. So ist „Long Interspersed Nuclear Element" (LINE) eine Bezeichnung für den Zweig von non-LTR-Retrotransposons, der bei Säugern anzutreffen ist. Vom Gebrauch des wenig deskriptiven Ausdrucks „LINE" in einem weiter gefassten Sinne, etwa als synonyme Bezeichnung für non-LTR-Retrotransposons, wird abgeraten. Als „Short Interspersed Nuclear Element" (SINE) wird ein Typ von passiv replizierten, kurzen, PolIII-transkribierten Retroelementen bezeichnet (SMIT 1996). Der wohl bekannteste Vertreter dieser Gruppe ist das Alu-Element, das sich unter Primaten sehr schnell ausgebreitet hat (JURKA & MILOSAVLJEVIC 1991). Gleichermaßen erfolgreich ist das B1-Element bei Nagern (DEININGER & DANIELS 1986). Bei den SINEs handelt es sich um einen speziellen Transposontyp, in dem sich das Phänomen von passiver Mobilität widerspiegelt. Auf die Abhängigkeit der SINEs von aktiven non-LTR-Retrotransposons wird noch im folgenden Abschnitt eingegangen.

1.2.1 Retrotransposons

Charakteristisches und namensgebendes Merkmal der Retrotransposons ist die RT. Die RTs von Retrotransposons, verschiedenen Viren, einer Gruppe von prokaryotischen Insertionssequenzen (engl. „multi-copy single-stranded DNA", msDNA) und Gruppe-II-Introns zeigen über ihre gesamte Länge Sequenzähnlichkeit, so dass der Schluss auf einen gemeinsamen Ursprung der RT aller sogenannten „Retroelemente" naheliegt (XIONG & EICKBUSH 1990).

Einen Einblick in die verwandtschaftlichen Beziehungen zwischen den Retroelementen ermöglicht eine Baumberechnung am Alignment der RT (EICKBUSH 1992, Abb. 1). Aus der Überlegung, dass das organische Leben ursprünglich nur RNA als Informationsträger kannte, ergibt sich die Annahme, dass eine RNA-abhängige RNA-Polymerase das Vorläuferenzym der RT ist (XIONG & EICKBUSH 1990). Eine solche wurde als Wurzel für den dargestellten Verwandtschaftsbaum gewählt. Alternative Hypothesen platzieren die Urform der RT unter msDNAs oder Gruppe-II-Introns (MALIK ET AL. 1999).

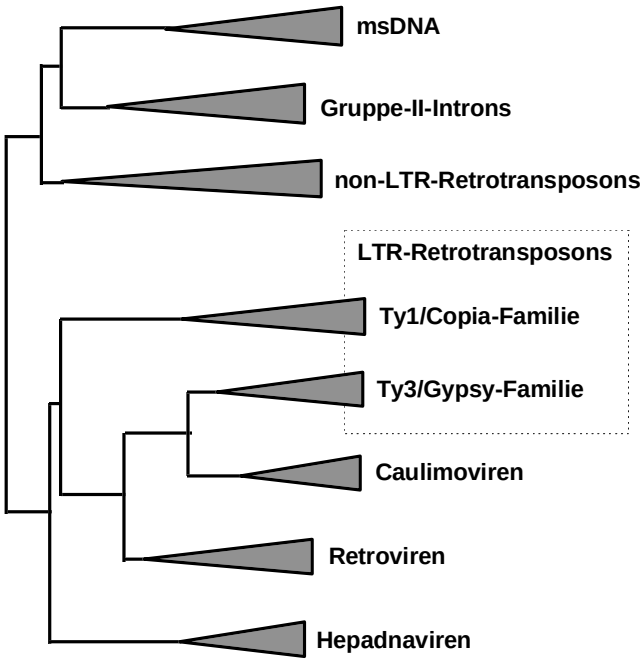


Abb. 1. Verwandtschaftliche Beziehungen zwischen Retroelementen, basierend auf einer Baumberechnung durch die Methode des Neighbor Joining anhand eines Alignments der reversen Transkriptase (RT), modifiziert nach EICKBUSH 1992. Die Wurzel bilden RNA-abhängige RNA-Polymerasen (vgl. dazu Text). Die Dreiecke an den Zweigenden geben in horizontaler Ausdehnung die Verzweigungstiefe in den jeweiligen Gruppen wieder.

Unabhängig von der Frage nach der ursprünglichsten Form der RT unterstützen alle geläufigen Vorstellungen die These, dass **non-LTR-Retrotransposons** eine verwandtschaftliche Basisposition gegenüber den LTR-Retrotransposons einnehmen. Dafür sprechen die einfachere strukturelle Organisation und der schlichtere Replikationsmechanismus, der sich ganz deutlich von dem der LTR-Retrotransposons unterscheidet. Bei den non-LTR-Retrotransposons wird die Transposition durch einen durch die Endonuklease generierten Einzelstrangbruch am genomischen Ziel eingeleitet. Die reverse Transkription beginnt anschließend direkt am freien 3'-OH dieses Einzelstrangbruchs (LUAN & EICKBUSH 1995). Die neu synthetisierte DNA-Kopie des Transposons ist also vom Zeitpunkt ihrer Entstehung an kovalent an den neuen genomischen Locus gebunden. Einzelheiten zur Bildung des DNA-Doppelstrangs und zum Einfügen des Gegenstrangs in die genomische DNA sind noch unklar.

Retrotransposons besitzen – auch bei eukaryotischen Wirten – niemals Introns. Sie würden während der Replikation zwangsläufig verlorengehen. Da die Transposon-mRNA als Replikationszwischenstadium die gesamte Sequenz des Elements beherbergen muss, ist sie kompakt organisiert und polycistronisch. Endonuklease, RT und Integrase werden dementsprechend in Form eines einzigen, komplexen Multidomänenenzyms exprimiert, des sogenannten Polyproteins (Pol). In einem separaten ORF wird das Gag-Protein codiert, das auch bei LTR-Retroelementen anzutreffen ist. Gag hat die Funktion eines Komplexierungsproteins, das die Bildung virusähnlicher Partikel herbeiführt, welche ein RNA-Transkript und Polyprotein enthalten. Neben der Nukleinsäurebindung, die typischerweise durch das CCHC-Motiv, $CX_2CX_4HX_4C$, vermittelt wird, kommt dem Gag-Protein eine maßgebliche Rolle bei der Vermittlung des Kernimports zu (DANG & LEVIN 2000). Eine zweckmäßige Über-

produktion von Gag im Verhältnis zum Polyprotein wird durch getrennte, im Leseraster versetzte ORFs ermöglicht. Das Polyprotein im zweiten ORF wird nur dann translatiert, wenn nach Passieren des Stopcodons in ORF1 eine ribosomale Reinitiation am nahegelegenen Startcodon für ORF2 erfolgt.

Um die Replikation ohne Sequenzverlust für das 5'-Ende zu gewährleisten, muss die Transkription eines non-LTR-Retrotransposons über einen internen Promotor initiiert werden (MCLEAN ET AL. 1993). Die Erkennung der eigenen mRNA durch das Polyprotein wird durch Protein-DNA-Wechselwirkung allein mit dem 3'-UTR der mRNA bewerkstelligt (LUAN & EICKBUSH 1995). Da die Vollständigkeit des 5'-Endes der mRNA auf diese Weise nicht sichergestellt werden kann, kommt es häufig zur Insertion 5' unvollständiger Elementkopien (WINCKLER 1998). Die relativ geringe Sequenzspezifität bei der Vermittlung der reversen Transkription ist auch verantwortlich für das Auftreten prozessierter Pseudogene, d.h. intronloser Genduplikate. Sie entstehen, wenn eine wirtseigene mRNA durch Aktion der RT eines non-LTR-Retrotransposons in das Genom revers transkribiert wird. Darüber hinaus sind SINEs obligate Trittbrettfahrer der transpositionellen Maschinerie von non-LTR-Retrotransposons. Sie sind äußerst kurz – wenige hundert Basenpaare –, besitzen kein codierendes Potenzial und sind daher nur passiv mobil (SMIT 1996).

Bei den **LTR-Retrotransposons** ist die reverse Transkription von der Insertion räumlich und zeitlich entkoppelt. Das RNA-Intermediat wird erst in ein DNA-Intermediat transkribiert und dieses dann in einen genomischen Locus eingefügt. Die reverse Transkription des RNA-Intermediats wird durch Priming mit einer häufigen zellulären RNA-Spezies, meist einer tRNA, initialisiert (Abb. 2A). Da die Flanken der LTR-Retroelemente über eine Länge von 200-400 bp identisch sind (die namensgebenden „Long Terminal Repeats", LTRs), kann über Dissoziation und Reassoziation der Nukleotidstränge die DNA-Synthese zwischen Sinn- und Gegensinnstrang des Elements hin- und herspringen (Abb. 2D). Nach diesem Prinzip wird der im primären Transkript unvollständige linke LTR auf Basis der Sequenz des rechten LTRs rekonstruiert. Dissoziation und Reassoziation von RNA-Template und Zwischenprodukten der DNA-Polymerisation werden erleichtert durch simultanes Einwirken einer transposonbürtigen Ribonuklease (RNAse H, Abb. 2E,F,H). So führt der Prozess schließlich zu einem doppelsträngigen cDNA-Molekül, das die komplette Sequenz des LTR-Retrotransposons repräsentiert. Eine Integrase-Aktivität inseriert die cDNA an einen genomischen Locus, so dass eine neue genomische Kopie des Transposons entstanden ist. Als Bindemotiv für die Integrase besitzen die äußeren Enden der LTR-Retroelemente die stark konservierte Nukleotidsequenz TG...CA.

Insgesamt ist die Genorganisation von LTR-Retrotransposons der von non-LTR-Retrotransposons sehr ähnlich. In zwei bis drei ORFs wird neben RT, RNAse und Integrase ein Gag-Protein codiert, das als Hüll- oder Komplexierungsuntereinheit die Bildung von virusähnlichen Partikeln unterstützt. Hinzu kommt meist eine Protease, welche die multifunktionellen Translationsprodukte in ihre Einheiten zerlegt.

Solo-LTRs sind häufig beobachtete einzelnstehende LTR-Sequenzen, die vermutlich aus vollständigen LTR-Retrotransposons durch Rekombination hervorgehen. Sie machen in vielen Fällen den

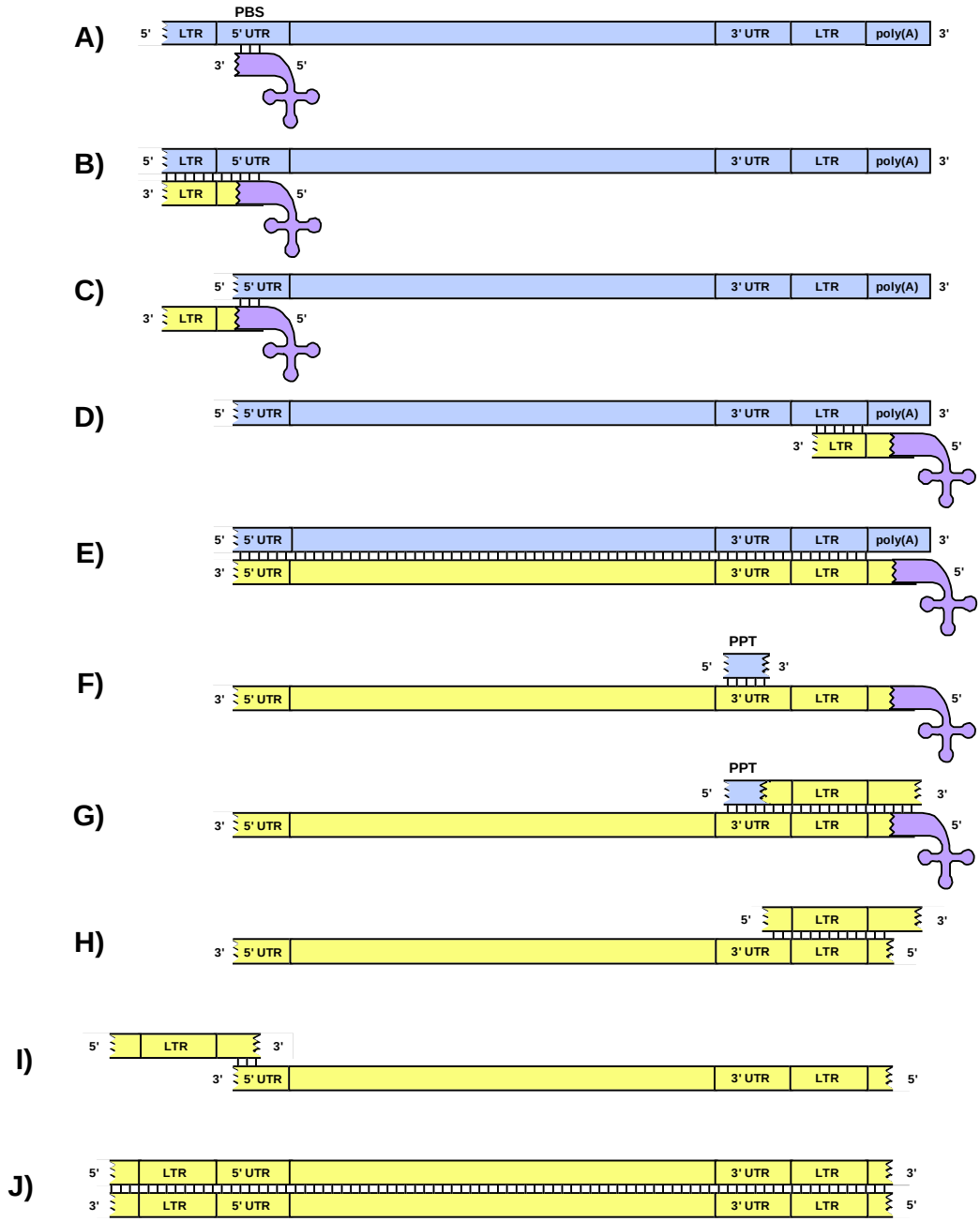


Abb. 2. Replikationsmechanismus eines LTR-Retroelements, modifiziert nach GÖTTE ET AL. 1999. Anhand ihrer Farbe sind RNA-Transkript (blau), Primermolekül (violett) und synthetisierte DNA (gelb) zu unterscheiden. Die Größenverhältnisse zwischen LTRs, UTRs und Zentralregion sind nicht maßstabsgetreu. Erklärung der Reaktionsschritte: **A)** Dem Transposon-Transkript fehlt ein Teil des linken LTR, da der Transkriptionsstart stromabwärts eines Promotors im LTR liegt. 3' hingegen besitzt das Transkript einen Sequenzüberschuss von flankierender Sequenz und Poly-A-Schwanz. An die sogenannte „Primer Binding Site" (PBS) bindet eine zelluläre RNA-Spezies **B)** Es beginnt die reverse Transkription des Gegensinnstrangs **C)** RNAse H verdaut hybridisierte Teile der mRNA **D)** Dissoziation und Reassoziation von mRNA- und cDNA-Strang, bevorzugt intramolekular **E)** Weitgehende DNA-Polymerisation des Gegensinnstrangs **F)** Die mRNA wird größtenteils durch RNAse H verdaut, die Region des Polypurintrakts (PPT) bleibt jedoch bevorzugt erhalten **G)** Die verbleibenden Reste der mRNA dienen als Primer für die cDNA-Doppelstrang-Synthese **H)** RNAse H beseitigt alle verbliebenen RNA-Reste **I)** Erneute Dissoziation und Reassoziation der Stränge, bevorzugt intramolekular **J)** Komplettierung des cDNA-Doppelstrangs.

numerischen Hauptteil der Transposonkopien aus (KIM ET AL. 1998, GOODWIN & POULTER 2000). Offenbar findet die Rekombination zwischen den LTRs relativ häufig statt. Dieser Deletionsmechanismus geht einher mit einer Verminderung der Zahl aktiver Transposonkopien und erlaubt so, die Zahl von Transpositionsereignissen im Wirtsgenom über lange Zeiträume weitgehend konstant zu halten.

1.2.2 DNA-Transposons

Der Mobilisierungsmechanismus von DNA-Transposons beruht auf einer Folge von Exzision und Insertion. Dieser Prozess wird durch ein einziges Enzym bewerkstelligt, die **Transposase**. Der Prozess ist analog zu einer Abfolge von Rück- und Hinreaktion des Prozesses, der bei den Retroelementen durch Integrase katalysiert wird. Daher überrascht es nicht, dass zumindest für einen Teil der DNA-Transposons eine Sequenzähnlichkeit zwischen Transposasen und den Integrasen verschiedener Retroelemente besteht. Die sequenzkonservierte Domäne wird D35E-Motiv genannt und stellt eine Verallgemeinerung des DDE-Motivs dar, das für die Integrase der LTR-Retroelemente definiert wurde (DOAK ET AL. 1994). Als Bindemotiv für die Transposase besitzen die Enden der DNA-Transposons einen **invertierten terminalen Repeat (ITR)**, der meist viel länger ist als die analoge Sequenzwiederholung bei den LTR-Retroelementen, wenige bis knapp 100 bp. Die Sequenz der ITRs variiert zwischen den verschiedenen Gruppen von DNA-Transposons.

Zwar ist die Klasse der DNA-Transposons heterogen, doch zumindest lassen sich zwei Gruppen mit jeweils vielen Mitgliedern in eukaryotischen Genomen voneinander abgrenzen. Die erste ist die **Tc1/mariner-Subklasse**: Die Transposase ihrer Mitglieder weist das beschriebene D35E-Motiv auf, die Target-Site-Duplikationen (vgl. Abschnitt 1.2.3) haben überwiegend eine Länge von 2 bp und typischerweise die Sequenz TA. Gut untersuchte Vertreter sind das Element Tc1 aus *Caenorhabditis elegans* und die mariner-Elemente bei verschiedenen Insekten und beim Menschen (Überblick in DOAK ET AL. 1994). Die andere Gruppe ist die **hAT-Familie** von DNA-Transposons: Die Transposase dieser Familie weist drei charakteristische konservierte Sequenzblöcke auf (CALVI ET AL. 1991). Der Consensus der ITR-Sequenzen lautet CAGRGA...TCYCTG, und die Target-Site-Duplikationen (vgl. Abschnitt 1.2.3) haben typischerweise eine Länge von 8 bp. Die namensgebenden Vertreter dieser Familie sind die Elemente hobo aus *Drosophila*, Activator aus Mais und Tam3 aus *Antirrhinum majus* (vgl. CALVI ET AL. 1991).

Vergleiche der Abstammungsverhältnisse zwischen DNA-Transposons und der sie beherbergenden Wirtsorganismen haben vielfach zur Postulierung einer horizontalen Ausbreitung der DNA-Transposons geführt (KOGA ET AL. 2000, HARING ET AL. 2000 u.v.m.). Die in der Literatur dokumentierten Befunde beziehen sich meist auf kürzliche Ereignisse horizontalen Transfers, d.h. wenige Jahre bis wenige Millionen Jahre, so dass anzunehmen ist, dass diese Ereignisse in der Natur relativ häufig auftreten. Da der Transpositionsmechanismus von DNA-Transposons – anders als bei den Retrotransposons – nicht sicherstellt, dass bevorzugt aktive Transposonkopien mobilisiert werden, ist der Bestand eines DNA-Transposons innerhalb seines Wirts stark gefährdet, wenn inaktive Kopien sich anreichern und die aktiven Kopien zunehmend verdrängen. Aus dieser Überlegung ergibt sich ein Erklärungsansatz für einen hohen evolutiven Druck zum häufigen Wirtswechsel (SMIT 1996).

1.2.3 Charakteristika der Transposoninsertion

Die genomische Insertion von LTR-Retrotransposons und die Transposition bei den DNA-Transposons sind nicht nur analoge biochemische Prozesse, die Enzyme – Integrasen der LTR-Retrotransposons und Transposasen der DNA-Transposons – weisen auch strukturell eine deutliche Homologiebeziehung auf (CRAIG 1995, DOAK ET AL. 1994). Abb. 3 veranschaulicht die Uniformität der Insertionsreaktion.

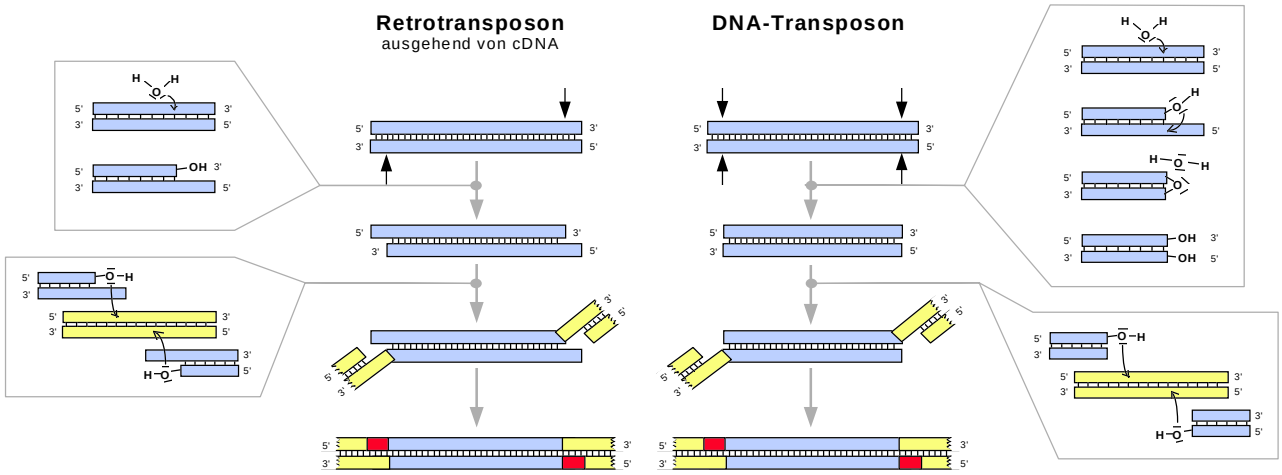


Abb. 3. Uniformität der Transposoninsertionsreaktion im Vergleich zwischen Retrotransposons und DNA-Transposons, modifiziert nach CRAIG (1995) und DAVIES ET AL. (2000). Nach ihrer Farbe sind Transposon-DNA (blau), genomische DNA (gelb) und durch Reparatur eingefügte DNA (rot) zu unterscheiden. Oben markieren schwarze Pfeile die Orte von nachfolgenden DNA-Einzelstrangspaltungen. **linke Seite:** Reaktionsschema für die Insertion von Retrotransposons. **rechte Seite:** Reaktionsschema für die Insertion von DNA-Transposons. **Kästen außen:** Erklärungen zu den Einzelschritten der Reaktionen.

Da die Insertion fast immer an versetzten Positionen des DNA-Doppelstrangs erfolgt, wird die Sequenz zwischen den Einzelstrangbrüchen dupliziert – man spricht von einer **„Target-Site-Duplikation“** (TSD). Die Länge der TSD ergibt sich aus der räumlichen Konstellation zwischen aktivem Zentrum, DNA-bindender Domäne(n) und Dimerisierungsdomäne(n) der Integrase bzw. Transposase. Sie ist also ein spezifisches Merkmal jedes Transposons.

1.3 Wissenschaftliche und ökonomische Bedeutung von Transposons

1.3.1 Auf Transposons basierende molekularbiologische Arbeitsmethoden

Nicht allein die Faszination von der Minimalausgabe einer weitgehend autonom replizierenden DNA hat die Forschung an Transposons in den letzten Jahrzehnten vorangetrieben. Transposons bilden den funktionellen Kern verschiedener molekularbiologischer Methoden:

Da ist zunächst an die ***in-vitro*-Insertionsmutagenese** an Vektor-DNA zu denken, die der randomisierten Erzeugung neuer Primerbindungsstellen zum Zwecke der DNA-Sequenzierung dient (PHADNIS ET AL. 1989, HAAPA ET AL. 1999). Die Systeme basieren im Wesentlichen auf einem mobilen DNA-

Fragment eines artifiziellen Transposons oder Phagen, das ein Resistenzgen als Selektionsmarker für Insertionsmutanten zwischen den ITRs beherbergt. Durch Zugabe von Transposase- bzw. Integrase-Protein werden aktive Integrationskomplexe gebildet und die Insertionsreaktion an der Ziel-DNA ausgelöst. Die Technik ist eine hervorragende Alternative zum „Primer-Walking“, wenn die Sequenz nicht die Definition einer spezifischen Primerbindungsstelle zulässt, beispielsweise durch Repetitivität oder extreme Werte im G/C-Gehalt. Heute sind Systeme als sogenannte „Transposon-Kits“ kommerziell bei verschiedenen Anbietern verfügbar.

Früh wurde der Wert von Transposons erkannt, den sie durch ***in-vivo*-Insertionsmutagenese** zur Erzeugung von Knock-Out-Mutanten besitzen. Beruht doch diese Technik auf der einfachen Beobachtung, dass die natürliche Mobilität von Transposons regelmäßig zu Gendisruptionen führt. Der naheliegendste und technisch einfache Ansatz besteht in einer Ausnutzung des Transposon-Repertoires des zu untersuchenden Organismus (KEMPKEN & KÜCK 1996). Ein Nachteil dieser Methode besteht darin, dass die mutierten Genloci nicht spezifisch vor dem Hintergrund der natürlich etablierten Transposonkopien identifiziert werden können. Ein methodischer Fortschritt wird daher erreicht, wenn man Transposonsysteme aus anderen Organismen überträgt (BESSEREAU ET AL. 2001) oder indem man artifizielle Transposonvektoren konstruiert (TSUGEKI ET AL. 1998). Artifizielle Systeme können auf hohen Durchsatz optimiert werden, Mutanten lassen sich durch Marker selektieren oder per PCR screenen, und schließlich können Loci von Insertionen zügig durch inverse PCR identifiziert und charakterisiert werden.

1.3.2 Transposons im Kontext der Genomforschung

Aus Sicht eines Projekts zur Genomsequenzierung gilt das Augenmerk besonders dem repetitiven Charakter der Transposons. Je nach Kopiendichte, Länge und Divergenz stellen sie ein mäßiges bis unüberwindbares Problem bei der Assemblierung der genomischen Sequenz aus Schrotschussfragmenten dar. Die zu erwartenden Probleme hängen auch ganz entscheidend von der Klonierungs- und Sequenzierungsstrategie ab. Nähert man sich der Genomsequenz über kartierte Klone mit großen Inserts – wie beim öffentlichen Humangenomprojekt (INT. HUM. GENOME SEQ. CONS. 2001) – ist die Wahrscheinlichkeit eher gering, dass man zwei sehr ähnlichen Kopien des gleichen Transposons auf einem Klon begegnet. Verfolgt man jedoch eine genomweite Schrotschussstrategie, ist die Wahrscheinlichkeit für derartige Problemfälle groß, und es erfordert sorgfältige Planung, ihnen zu begegnen (ANSON & MYERS 1999). Es waren genau diese Überlegungen, die zu Beginn eines Genomsequenzierungsprojekts für *Dictyostelium discoideum* (nächster Abschnitt) die vorliegende Arbeit zur Aufklärung des Transposongehalts motiviert haben.

1.4 Ein Genomprojekt für den Modellorganismus *Dictyostelium discoideum*

1.4.1 Allgemeines

Dictyostelium discoideum, eine bodenbewohnende Amöbe, wird vielfach herangezogen für Modellstudien an zellulärer Bewegung, Signaltransduktion und zellulären Differenzierungs- und Entwick-

lungsprozessen (KAY & WILLIAMS 1999, NOEGEL & SCHLEICHER 2000). Das wissenschaftliche Interesse an diesem Organismus motivierte zunächst ein cDNA-Sequenzierungsprojekt in Japan (MORIO ET AL. 1998) und bald auch ein Genomprojekt, das im Jahre 1998 durch ein internationales Konsortium (Tab. 4) aufgenommen wurde. Dieses Genomprojekt ist Anlass und Rahmen für die hier vorgestellte Arbeit.

Tab. 4. Das internationale Konsortium zur Sequenzierung des Genoms von *D. discoideum* und weitere, kooperativ beteiligte Gruppen.

Institut	Beitrag	Ansprechpartner
Baylor College of Medicine (Houston)	Sequenzierung Chr6, Chr4/Chr5; YAC-Ressourcen	A. Kuspa R. Sucgang R. Gibbs
Institut für Molekulare Biotechnologie (Jena) / Institut für Biochemie (Köln)	Sequenzierung Chr2, Chr6-YACs, Chr1, Chr3	A. Rosenthal, Jena M. Platzer, Jena G. Glöckner, Jena L. Eichinger, Köln A.A. Noegel, Köln
Institut de Pasteur (Paris)	Sequenzierung Chr6-YACs	M. Veron C. Buchrieser
MRC Laboratory of Molecular Biology (Cambridge)	HAPPY-Mapping	P.H. Dear
Princeton University	Chromosomen-Auftrennung	E. Cox
Sanger Centre (Hinxton)	Sequenzierung Chr6, Chr4/Chr5	M.-A. Rajandream M.A. Quail – DNA-Bibliotheken

Das Genom von *D. discoideum* AX4 hat eine geschätzte Größe von 34 Mbp und ist zum überwiegenden Teil in 6 akro- bis telozentrischen Chromosomen organisiert (LOOMIS ET AL. 1995). Daneben existiert ein extrachromosomales Multi-Copy-Palindrom von ca. 90 kbp Größe, das die RNA-Gene der ribosomalen Untereinheiten beherbergt. Schätzungen für die Chromosomen- und schließlich die Genomgröße basieren neben einer Beurteilung elektrophoretisch aufgetrennter Chromosomen (COX ET AL. 1990: ca. 40 Mbp beim Stamm AX3) im Wesentlichen auf einer genomweiten YAC-Karte mit einer berichteten Abdeckung von > 98 % (KUSPA & LOOMIS 1996: 34 Mbp). Da Ergebnisse des HAPPY-Mappings von Chromosom 6 im deutlichen Widerspruch mit den Ergebnissen der YAC-Karte stehen, wonach die geschätzte Fehlkartierung der YACs 60 % oder mehr beträgt (KONFORTOV ET AL. 2000), sind die Literaturdaten mit großer Vorsicht zu behandeln.

Der mittlere G/C-Gehalt des Genoms wurde auf 23 % geschätzt (FIRTEL & BONNER 1972). Dabei ist der G/C-Gehalt deutlich ungleichmäßig auf codierende Sequenz (31 %) und nicht codierende Sequenz (12 %) verteilt (PARRA ET AL. 2001). Der hohe A/T-Gehalt bringt mit sich, dass DNA-Fragmente von mehr als etwa 5 kbp Länge in *E. coli* nicht stabil kloniert werden können (GLÖCKNER 2000). Die fehlende Verfügbarkeit von Klonen mit großen Inserts wiederum lässt nur eine genom- oder chromosomenweite Schrotschussstrategie als begehbaren Weg zur Sequenzierung des Genoms erscheinen. Im Abschnitt 1.3.2 wurde bereits darauf hingewiesen, dass eine genaue Kenntnis des

Transposongehalts für die Durchführung einer Schrotschussassemblierung derart großer Sequenzierungsziele unabdinglich ist.

1.4.2 Transposons im Genom von *D. discoideum*

Das erste im Genom von *D. discoideum* identifizierte und offenbar anteilmäßig häufigste transposable Element ist das LTR-Retrotransposon **DIRS-1** (ROSEN ET AL. 1983, CAPPELLO ET AL. 1985). Es zeichnet sich durch einen Hitzeschock-Promotor aus und wird in der späten Phase des Entwicklungszyklus stark transkribiert. DIRS-1 besitzt ungewöhnlicherweise palindromische LTRs, die zudem nicht vollkommen sequenzidentisch sind. Der spezialisierte Mechanismus, der im Gegensatz zum Standardschema der LTR-Retroelemente für eine sequenzkonservierende Replikation auch der LTRs sorgt, konnte beim bisher einzigen strukturellen Verwandten des DIRS-1, dem Element PAT aus dem Nematoden *Panagrellus redivivus* aufgeklärt werden: proximal benachbart zum rechten LTR befindet sich eine dritte, unvollständige LTR-Kopie, die als Matrize zum Vervollständigen des linken LTRs dient (CHASTONAY ET AL. 1992).

Ein weiteres LTR-Retrotransposon im Genom von *D. discoideum* ist das Element **skipper**. Seiner RT-Sequenz zufolge handelt es sich um einen typischen Vertreter aus der Gruppe der Ty3/Gypsy-ähnlichen Retrotransposons (LENG ET AL. 1998).

Der **HindIII-Repeat (H3R)**, ist eine repetitive Sequenz von 268 bp Länge, die mehrfach stromaufwärts von tRNA-Genen beobachtet wurde (WINCKLER 1998). Es besteht der Verdacht, dass es sich um den solo-LTR eines bisher nicht identifizierten LTR-Retrotransposons handelt. Dafür sprechen besonders die palindromischen Enden des Repeats mit der Sequenz TGT...ACA, eine typische Erkennungssequenz der Integrase dieser Elementklasse.

Mindestens drei verschiedene, aber nah verwandte non-LTR-Retrotransposons wurden im Genom von *D. discoideum* regelmäßig in der Nachbarschaft von tRNA-Genen beobachtet (WINCKLER 1998). Namentlich sind dies die Elemente Tdd-3 (WINCKLER ET AL. 1998), DRE (MARSCHALEK ET AL. 1992a), RED (WINCKLER ET AL. 1998, konkrete Charakterisierung nur unter Acc.No. AF067198) und das fragmentarische Element Tdd-2 (POOLE & FIRTEL 1984). Um Namensverwirrungen zu vermeiden, sei bereits an dieser Stelle angemerkt, dass diese Arbeit ein systematisches Benennungsschema für die Mitglieder dieser Familie einführt, wobei alle Namen mit dem Stamm „**TRE**“, einer Abkürzung für die englische Bezeichnung „tRNA gene-targeted retroelement“, beginnen (vgl. SZAFRANSKI ET AL. 1999, Abschnitt 3.2.1).

Kürzlich ist das erste DNA-Transposon, **Tdd-4**, in *D. discoideum* nachgewiesen worden (WELLS 1999). Es sticht aus den beschriebenen DNA-Transposons heraus durch seine ausgeprägte Ähnlichkeit der Transposase mit Integrasen von Retroelementen. Insofern handelt es sich beim Tdd-4 um so etwas wie ein „lebendes Fossil“, das den gemeinsamen Ursprung von Integrase der LTR-Retroelemente und Transposase der Tc1/mariner-Subklasse von DNA-Transposons verdeutlicht.

Es sei angemerkt, dass die Namensformel Tdd-X ursprünglich als einheitliches Schema zur Benennung der Transposons von *D. discoideum* gedacht war. Analog werden die Benennungsschemata TaX in *Arabidopsis*, TcX in *Caenorhabditis*, TxX in *Xenopus*, TyX in Hefe verwendet.

Tdd-1 ist eine alternative Bezeichnung für DIRS-1 (ROSEN ET AL. 1983), es hat sich aber der Name der zweiten, unabhängigen Erstbeschreibung durchgesetzt (ZUKER ET AL. 1983). Tdd-2 und Tdd-3 sind die erstbeschreibenden Namen für die non-LTR-Retrotransposons TRE3-C und TRE3-A (vgl. Abschnitt 3.2.1 ab S. 32 für das hier propagierte systematische Benennungsschema).

2 DEFINITIONEN, MATERIAL, METHODEN

2.1 Begriffe, Codierungen

2.1.1 Degenerierte Basencodierung

Die folgenden Definitionen werden entsprechend den Vorschlägen von NC-IUB (1985) verwendet.

Tab. 5. Degenerierte Basencodierung laut NC-IUB (1985).

Code	Basen	Erklärung
B	G/T/C	nicht A
D	A/G/T	nicht C
H	A/C/T	nicht G
K	G/T	Ketobasen
M	A/C	Aminobasen
N	A/C/G/T	engl. „any“, jede Base
R	A/G	Purine
S	C/G	engl. „strong“, starke Paarung
V	A/C/G	nicht T
W	A/T	engl. „weak“, schwache Paarung
Y	C/T	Pyrimidine

2.1.2 Begriffe im Kontext der Genomsequenzierung

Die verfolgte Methode der Genomsequenzierung ist „**schrotschussbasiert**“, was bedeutet, dass genomische DNA durch ein weitgehend zufällig wirkendes Verfahren (Ultraschall) in kleine Bruchstücke zerlegt wird. Das Ergebnis der Sequenzierung der klonierten Bruchstücke sind dann „Schrotschusssequenzen“ (vgl. technische Definition unter Abschnitt 2.2.3).

Im Prozess der **Assemblierung** wird der Vorgang des Zerstückelns wieder umgekehrt, um möglichst die durchgehende Sequenz ganzer Chromosomen zu rekonstruieren. Dazu werden die Schrotschuss-sequenzen nach sequenzähnlicher Überlappung in Alignments gebracht, sogenannte „**Contigs**“. Eine wichtige Maßzahl bei der Assemblierung ist die mittlere Abdeckung der genomischen Sequenz durch Schrotschusssequenzen, als anglo-amerikanischer Fachbegriff „**Coverage**“. In dieser Arbeit wird der Begriff „Coverage“ auch angewandt auf die Abdeckung einer repetitiven Sequenz durch ihre Einzelkopien oder durch Schrotschusssequenzen.

2.1.3 Taxonomische Begriffsdefinitionen für die Klassifizierung von Transposons

Die taxonomische Bearbeitung von Transposons ist ähnlich schwierig wie bei den Bakterien, bei denen ein überzeugendes **Artkonzept** fehlt. Es hat sich im Verlauf dieser Arbeit als praktikabel

erwiesen, Transposonkopien – also „**Transposonindividuen**“ – nach ihrer Sequenzähnlichkeit in Gruppen – also Taxa – zusammenzufassen. Ordnet man Transposons nach ihrer Sequenzähnlichkeit, so bilden sich scharf unterscheidbare Elementartaxa bei einem Ähnlichkeitsschwellenwert im Bereich von 90 bis 95 %, in Ausnahmefällen bis zu 85 %. Diese Elementartaxa bezeichne ich als „Art“ oder „**Spezies**“. Nach dieser Auffassung handelt es sich bei den Veröffentlichungen anderer Autoren über die Transposons DIRS-1, skipper, Tdd-3, Tdd-4 usw. um Artbeschreibungen. Eine Einteilung auf höherer taxonomischer Ebene ist schwieriger. Hier wird die Bezeichnung „**Familie**“ für eine Gruppe von Transposons verwendet, die auf der Ebene ihrer Proteinsequenz oder anderer struktureller Eigenschaften deutlich eine Verwandtschaft erkennen lassen. Als „Klasse“ und „Subklasse“ werden – wie in der Einleitung – diejenigen Gruppen bezeichnet, die aus der funktionellen Grobeinteilung aller Transposonformen resultieren.

2.1.4 Begriffsdefinitionen: Polymorphismen und polymorphe Ausprägung

Die einzelnen Kopien von Multi-Copy-Sequenzen, in unserem Fall Transposons, zeigen untereinander Unterschiede, d.h. sie besitzen eine lokale Diversität. Ich nenne die beobachteten Unterschiede „**Polymorphismen**“, die Positionen der Multi-Copy-Sequenz, an denen Unterschiede zwischen den Kopien beobachtet werden, „**polymorphe Positionen**“. Die individuellen Sequenzformen an einer polymorphen Position nenne ich „Sequenzausprägungen“ oder „**Allele**“. Die relative Häufigkeit, mit der eine einzelne Sequenzausprägung an einer bestimmten polymorphen Sequenzposition auftritt, heisst „Allelfrequenz“. Allgemein betrachte ich die Sequenzvariation in ihrem kleinstmöglichen regionalen Ausmaß, d.h. an einer einzelnen Basenposition bei Basenaustauschen (sog. „Single Nucleotide Polymorphisms“, SNPs) oder über wenige Basenpositionen bei InDels. Die Begriffe „Allel“ und „Polymorphismus“ werden in Anlehnung an den populationsgenetischen Sprachgebrauch verwendet, der sich üblicherweise auf die Sequenzdiversität zwischen Kopien gleicher Loci von verschiedenen Individuen bezieht. Sie lassen sich analog auf die hier behandelte Sequenzdiversität unter verschiedenen Kopien eines Transposons übertragen, wenn man die einzelnen Transposonkopien jeweils als „Minigenome“ betrachtet.

2.2 Molekulargenetische Methoden

2.2.1 Schrotschuss-DNA-Bibliotheken von *D. discoideum*

Eine gesamtgenomische Schrotschuss-DNA-Bibliothek von *D. discoideum* wurde durch G. GLÖCKNER (IMB Jena) nach der folgenden Vorschrift hergestellt: Zellen von *D. discoideum* AX4 werden in Suspensionskultur kultiviert, während der logarithmischen Phase geerntet und anschließend gewaschen. Auf Phosphat-gepufferten Agarplatten lässt man die Zellen zu Aggregaten in einer Dichte von $1 \cdot 10^8$ Zellen je Platte auswachsen. Aus den geernteten Zellen werden Zellkerne nach der Methode von ROGGE & RISSE (1974) hergestellt und in LMP-Agarose (FMC) eingebettet. Durch Einwirken von SDS und Proteinase K (Roth) wird die DNA freigesetzt. Zur Abtrennung des extrachromosomalen rDNA-Palindroms wird eine Pulsfeldgelelektrophorese (18 h 160 V, Wechselzeit 150 s) durchgeführt und Agaroseblöcke mit den langsam wandernden Banden isoliert.

Nach Behandlung mit β-Agarase (New England Biolabs) wird die DNA durch zwei Impulse von je 5 s im Sonikator (Heat Systems) fragmentiert, die Fragmente werden bei 6 V/cm für 4 h auf einem 0,8-%-igen Agarosegel aufgetrennt und im Größenbereich von 1 kbp bis 2 kbp isoliert. Die Agarose wird durch Anwendung des Jetsorb-Kits (Genomed) abgetrennt, dann werden die DNA-Fragmente in die SmaI-Schnittstelle des Plasmidvektors pUC18 (Acc.No. L08752) ligiert.

Durch M. QUAIL (Sanger Centre, Hinxton) wurden der Arbeitsgruppe weitere chromosomenspezifische Bibliotheken zur Verfügung gestellt. Die Erstellung erfolgte analog der Methode für die gesamtgenomische Bibliothek, allerdings wurde eine verfeinerte Pulsfeldgelelektrophorese durchgeführt, um die Chromosomen gegeneinander aufzutrennen (Trennung durch E. COX, Princeton University). Aus der gleichen Quelle stammen auch DNA-Bibliotheken verschiedener YACs, die gemäß Arbeiten von A. KUSPA und W. LOOMIS auf Chromosom 6 kartiert wurden.

2.2.2 cDNA aus dem Einzelzellstadium von *D. discoideum*

In Flüssigkultur wird *Dictyostelium discoideum* AX4 bis zu einer Dichte von $4 \cdot 10^6$ Zellen kultiviert (logarithmische Phase). Die Zellen werden durch Zentrifugieren für 5 min 180×g pelletiert und anschließend mit Sörensen-Puffer (17 mM Na-K-Phosphat, pH 6,0) gewaschen. Aus $2 \cdot 10^8$ Zellen wird mit dem RNeasy Midi-Kit (Qiagen) nach dem Protokoll „Isolation cytoplasmatischer RNA aus tierischen Zellen“ Gesamt-RNA isoliert. Unter Anwendung des Oligotex Midi-Kit (Qiagen) wird mRNA gewonnen, die mit Superscript RT (GibcoBRL) und poly(dT)-Primer in cDNA umgeschrieben wird.

2.2.3 Plasmidklonierung und Sequenzierung

Die Transformation erfolgt durch Elektroporation von elektroporationskompetenten *E. coli* XL1 Blue (Stratagene) im MicroPulser (BIORAD) durch 2,5 kV für ca. 5 ms. Die Kolonienanzucht und Selektion von Transformanten wird auf LB-Agar mit 100 µg/ml Ampicillin durchgeführt. Zur präparativen Klonkultivierung wird in 10 ml 2×-TY-Medium mit 10 µg/ml Ampicillin überimpft und 16 h bei 37 °C inkubiert. Die Plasmid-DNA, die als Sequenzierungsmatrize dient, wird mit dem QiaPrep Turbo-Kit (Qiagen) mit Hilfe eines Roboters BioRobot 8000 (Qiagen) isoliert und gereinigt. Zur Sequenzierung werden ca. 0,3 pmol Plasmid-DNA, 3 µl Big Dye Terminator Mix (PE Biosystems), 3 µl Half Term (GenPak) und 0,04 µl 100 mM Primer (typischerweise M13-fwd oder M13-rev) in 20 µl Reaktionsvolumen folgender Cycle-Sequencing-Prozedur unterworfen: 5 min 95 °C, 30 Zyklen von je 30 s 95 °C, 10 s 55 °C, 4 min 60 °C. Nach einer alkoholischen Fällung wird das Produkt der Sequenzierreaktion auf Sequenzierautomaten der Typen ABI377 oder ABI3700 aufgetrennt.

Die erzeugten Sequenzierrohdaten werden zu Dateien im Staden-Experiment-Format und SCF-Format prozessiert (BONFIELD ET AL. 1995). Diese Umwandlung wird durch die Software REAP bewerkstelligt (IMB Jena, unveröffentlicht). Im Verlauf dieser Prozessierung werden auch die Bereiche von Sequenziervektor und brauchbarer Signalqualität bestimmt und markiert. Wenn nicht

anders vermerkt, bezieht sich die Bezeichnung „Schrotschusssequenz“ in dieser Arbeit auf den qualitativ positiv bewerteten und vektorfreien Sequenzbereich des Ergebnisses einer Sequenzierreaktion von einem genomischen Fragment gemäß voran beschriebener Methoden.

2.2.4 Oligonukleotide

Alle Oligonukleotide werden über den Syntheseservice von MWG Biotech bezogen.

Tab. 6. Verwendete Oligonukleotide.

Name	Sequenz 5'→3'
DDT-A.812	GTTCCCCGTCATTTTTTGAAAATGG
DDT-A.981	TTCCAGCCAGCAAACAAAATG
DDT-A.1671R	CCATATATTGAGGTGTGTATCTTATTTG
DDT-A.2220	GAGCATCSTGTAAACTTAATGGAC
DDT-A.2415R	CACCCTCTGYCCACACC
DDT-A.3591	GCAGACTGAAAAAGCGAAAATG
DDT-A.3660R	CAGGTACATTACCATTYGGTGC
DDT-A.3790	CAAACCCAATGGAAAACCAGAG
DDT-A.4584R	GGGAAGAATTTAAGTAAAGTGCAG
DDT-A.4672R	TTAWGTGAACAATAAAATACACAGA
DDT-A.4876R	GCCCTTTAAAATTAACTTTATAATGCTC
DDT-S.21R	GGTGGTTTTTTCTTCGCTGTG
DDT-S.738	GGTGCATTTTTCTTYGCTGTG
JP0013	CTATACACTATGGCATTTTGAGAAG
JP0137	CTCCTTTTGTGTGGTTGATTTGG
JP0138	CGAAGGAATTTTTATGGTTGGCG
JP0179	GGGATCATCCAACACAATCAG
JP0180	CTTCTTTTGATGGACAATATTCAGTAG
JP0186	GAATTTTCAAGTAAAGGTGGTGAAG
JP3222	GAGAAAGGATAGATTTATCTATACAATATG
JP3223	CTCCACCAATAATTGTACCAC
JP3240	CCCATTCATTTGCTTTTGGTG
JP3265	CCATTGGACTCCACCATTC
JP3266	GATCAAATGTCTATCAATAAGGAGAAAC
JP3267	GAGAAACAGAGGGTAATAATGAAGATTG
M13-fwd	CGACGTTGTAAAACGACGGCCAGT
M13-rev	AGCGGATAACAATTTCACACAGGA
pahA.A	CAACCCTTCAAGAAAGTTCAAACC
pahA.B	GAGAGTAAACCTTGTACTGGAC
pahA.C	CCTTGTCCTTGAGAATTTGTGTTAG
thug-T.2R	CCGACCACCAGTTGTTAC
thug-T.74R	CAACATTACAACACCAATAACTTCAG
thug-T.693	TATCTGGTATCTAATATCTGTTATGAAAAC
thug-T.820	GGAGAAAACCCCTTTAAACAAAAC

2.2.5 PCR

Längencharakterisierung und Amplifizierung von pUC18-Inserts. Ca. 30 fmol Template-DNA, entsprechend 0,1 µl Vektorpräparat (Abschnitt 2.2.3), wird mit 3,75 µl $4 \times 2{,}5$ mM dNTP (Amersham Pharmacia), zweimal je 0,25 µl 100 µM Primern (MWG Biotech, $T_m \geq 55$ °C) und 2 U rekombinanter Taq-Polymerase (Qiagen) in 50 µl Reaktionsvolumen folgender thermischer Prozedur unterzogen: 1 min 94 °C, 22 Zyklen von je 40 s 94 °C, 40 s 53 °C, 90 s 70 °C, dann 5 min 72 °C. Bei einem großen zu erwartenden Produkt wird die Verlängerungsphase der Zyklen auf bis zu 8 min (bei ≥ 5 kbp Produktlänge) gesteigert und die Enzymmenge auf bis zu 4 U angehoben. Der Reaktionsansatz verbraucht bei einer Produktlänge von 750 bp Primer und Nukleotide in gleichem Maße. Bei abweichender Produktlänge wird eine Anpassung des Verhältnisses von Primer- zu dNTP-Konzentration vorgenommen, wenn das PCR-Produkt kloniert werden soll, wobei ein Überschuss von Primern störend wirkt.

Amplifizierung von genomischen Loci und Transkripten. 20 ng genomische DNA (gDNA) oder cDNA werden einer PCR-Reaktion unterzogen, wie sie im vorigen Absatz „Amplifizierung von pUC18-Inserts“ beschrieben ist. Es werden jedoch mindestens 28 Reaktionszyklen durchgeführt. Als Positivkontrolle wird das Intron-überspannende Primerpaar pahA.A / pahA.B für das Phenylalanin-Hydroxylasegen von *D. discoideum* verwendet. Das zu erwartende Produkt hat eine Länge von 780 bp (gDNA) oder 455 bp (cDNA).

Screening von bakteriellen Transformanten. Von Übernacht-Plattenkulturen werden Einzelkolonien auf Wells einer 96-Well-Platte mit jeweils 50 µl LB-Medium oder 2×TY-Medium mit 10 µg/ml Ampicillin überimpft und 3 h bei 37 °C inkubiert. 1 µl der Bakteriensuspension wird mit 1,5 µl $4 \times 2{,}5$ mM dNTP (Amersham Pharmacia), zweimal je 0,1 µl 100 µM Primern (MWG Biotech, $T_m \geq 55$ °C) und 1 U rekombinanter Taq-Polymerase (Qiagen) in 20 µl Reaktionsvolumen folgender thermischer Prozedur unterzogen: 3 min 94 °C, 28 Zyklen von je 40 s 94 °C, 40 s 53 °C, 90 s 70 °C, dann 5 min 72 °C. Die Verlängerungsphase der Zyklen wird je nach Größe des zu erwartenden Produkts auf bis zu 8 min (bei ≥ 5 kbp Produktlänge) gesteigert. Die PCR-Produkte werden auf einem Agarosegel elektrophoretisch aufgetrennt und durch Färben mit Ethidiumbromid visualisiert.

Durch **Sequenzierung** (Abschnitt 2.2.3) werden homogene PCR-Produkte nach Reinigung an Microcon PCR-Filterröhrchen (Millipore) ohne weitere Auftrennung sofort charakterisiert. Bei einer Längeninhomogenität der Produkte wird der PCR-Ansatz durch Gelelektrophorese unter Vermeidung von Ethidiumbromid und UV-Licht aufgetrennt, Agaroseblöcke ausgeschnitten und DNA mit dem QiaQuick-Kit (Qiagen) isoliert. Falls erforderlich, erfolgt eine Klonierung von PCR-Produkten mit dem pGEM-T Easy-System (Promega) oder pDrive-System (Qiagen) – man beachte die Ausführungen zum Verhältnis von Primer- und dNTP-Konzentration weiter oben in diesem Abschnitt. Die Aufarbeitung von bakteriellen Klonen und Vektor-DNA erfolgt dann wie unter Abschnitt 2.2.3 beschrieben.

2.2.6 Inverse PCR

Die zugrunde liegende Situation und die Versuchsplanung sind in Abb. 7 skizziert: Die bekannte Sequenz für ein Polynukleotid – hier genomische DNA – soll verlängert werden. Proximal der Abbruchstelle befindet sich ein Sequenzmotiv, das von einem Restriktionsenzym erkannt wird. Jenseits der Abbruchstelle wird sich ebenfalls das gleiche Motiv – allerdings an einer unbekannten Position – befinden (Abb. 7A). Spaltet man die DNA mit dem Restriktionsenzym und ligiert anschließend wieder, entstehen ringförmige Fragmente (Abb. 7B,C). In diesen Fragmenten ist der unbekannte DNA-Abschnitt beiderseits von der bekannten Sequenz flankiert. Er lässt sich also mit geeigneten Primern per PCR amplifizieren (Abb. 7D). Der Einsatz von verschachtelten Primern erlaubt eine sensitive und gleichzeitig spezifische Amplifizierung.

In einer Variante der Versuchsdurchführung wird das Gelingen der PCR-Amplifikation erleichtert, indem die ringförmigen Template-Moleküle (vgl. Abb. 7C) durch eine zweite Endonuklease linearisiert werden. Das zur Linearisierung verwendete Restriktionsenzym soll möglichst selten schneiden, um die Wahrscheinlichkeit einer Restriktion auch in der zu amplifizierenden DNA-Sequenz niedrig zu halten. Daher erschwert diese erweiterte Technik die Definition von geeigneten Restriktionsenzymen und Primern.

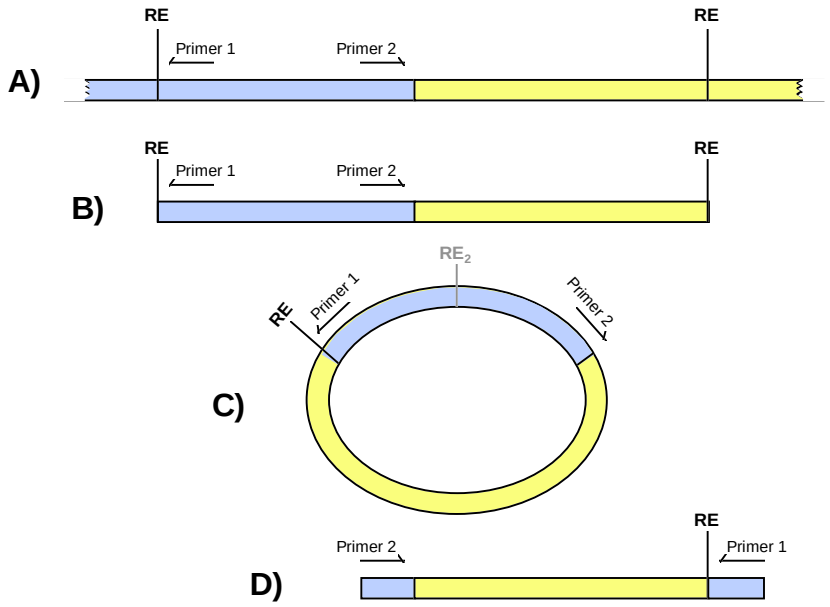


Abb. 7. Experimentverlauf der inversen PCR. Anhand ihrer Farbe sind bekannte Sequenz (blau) und unbekannte Sequenz (gelb) zu unterscheiden. Abkürzung: RE – Restriktionsschnittstelle, die die Inversion des PCR-Templates vermittelt, RE_2 – Restriktionsschnittstelle zum optionalen Öffnen des PCR-Templates.

Template. Ca. 400 ng genomische DNA werden mit 10 U Restriktionsenzym (New England Biolabs) – zur Wahl des Enzyms vgl. Abschnitt 3.3.1, besonders Tab. 12 auf S. 37 – in 100 µl Reaktionsvolumen nach Herstellerangaben inkubiert. Die DNA-Lösung wird durch Chloroform/Phenol-Extraktion und alkoholische Fällung gereinigt. Die in 20 µl gelöste DNA wird mit 2 U T4-Ligase (Boehringer) in 50 µl Reaktionsvolumen ligiert. Es schließen sich wiederum Chloroform/Phenol-

Extraktion und alkoholische Fällung an. Die gefällte DNA wird in 20 µl 1 mM TRIS/HCl pH 8.0 gelöst.

PCR-Template öffnen. 2,5 µl ligierte DNA, entsprechend 50 ng genomischer DNA, wird mit 2 U Restriktionsenzym (New England Biolabs) in 10 µl Reaktionsvolumen gespalten. Es werden nur solche Restriktionsenzyme verwendet, die sich durch Hitze inaktivieren lassen.

PCR-Reaktion. 2 µl des Restriktionsansatzes oder 0,5 µl des ungeöffneten Templates werden mit 3,75 µl $4 \times 2{,}5$ mM dNTP (Amersham Pharmacia), zweimal je 0,25 µl 100 µM Primern (MWG Biotech, $T_m \geq 55$ °C) und 4 U rekombinanter Taq-Polymerase (Qiagen) in 50 µl Reaktionsvolumen folgender thermischer Prozedur unterzogen: 3 min 94 °C, 35 Zyklen von je 40 s 94 °C, 40 s 53 °C, 5 min 70 °C, dann 5 min 72 °C. Nested-Primer-PCR wird zweimal mit je 25 Zyklen durchgeführt. Die Analyse der PCR-Produkte erfolgt durch Auftrennung an Agarose und Sequenzierung (Abschnitt 2.2.5).

2.3 Verwendete Programme und Datenbanken

Tab. 8. Verwendete Programme und Datenbanken.

Programm	Version	Referenz
BLAST	WU 2.0	ALTSCHUL ET AL. 1990
Clustal W	1.74 bis 1.81	THOMPSON ET AL. 1994
GenBank	bis 127.0	BENSON ET AL. 2000
GeneID	1.0	PARRA ET AL. 2000
GAP4	4.2 bis 4.5	BONFIELD ET AL. 1995
gff2ps	0.97	ABRIL & GUIGÓ 2000
Perl	5.004 und 5.6.0	WALL ET AL. 1996
Pfam	7.0	BATEMAN ET AL. 2002
PHYLIP	3.5	FELSENSTEIN
ProDom	2001.8	CORPET ET AL. 2000
Prosite	17	HOFMANN ET AL. 1999
REBASE	104	ROBERTS & MACELIS 2001
tRNAscan-SE	1.0	LOWE & EDDY 1997
UNIX	SunOS 5.6	

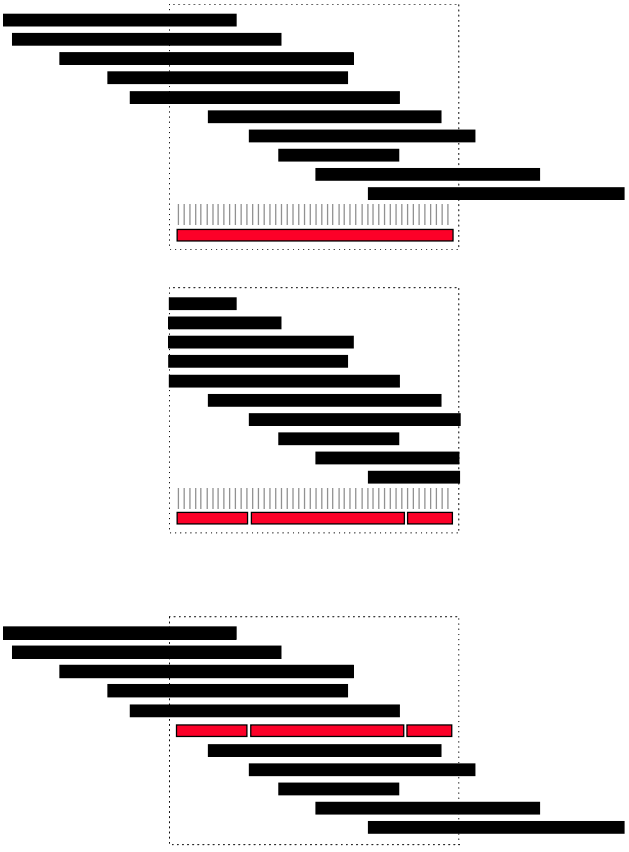
2.4 Sequenz-Clustering

2.4.1 Aufbau und Management von Sequenz-Clustern

Eine gegebene Keimsequenz – das Fragment einer (putativen) Transposonsequenz oder der vorläufige Consensus eines Transposons – wird als Probe verwendet, um per BLASTN Treffer unter den verfügbaren genomischen Schrotschusssequenzen zu finden. Homopolymere Regionen der Sequenzprobe mit einer Länge ≥ 12 bp werden maskiert. Für die BLAST-Parameter N, M und W werden die

Anfangswerte M=6, N=-18, W=12 verwendet. Wenn es später im Verlauf der zyklischen Alignmenterweiterung zu „Lawineneffekten“ durch unspezifische Treffer kommt, werden die Parameter N und W angepasst, um möglichst ein Optimum zwischen Spezifität und Sensitivität zu erreichen. Alle BLAST-Treffer werden gegen Schwellenwerte von relativer Nukleotididentität (90 % und höher) und BLAST-Score (250 und höher) gefiltert. Transposonfragmente mit größeren Insertionen oder Deletionen können zerstückelte BLAST-Treffer verursachen. In solchen Fällen wird nur das lokale BLAST-Alignment (sog. „HSP“) mit dem höchsten Trefferwert in das wachsende Alignment aufgenommen (Abb. 9A). Anderenfalls wäre die Gefahr von Fehlern beim nachfolgenden Alignmentschritt gegeben.

Ein Profil-Alignment der Trefferregion zwischen Cluster und Treffersequenz wird mit CLUSTAL W durchgeführt (Abb. 9B). Abweichungen von den Standardparametern sind dabei die Parameter -gapopen=0.385 und -gapext=2.20. Unähnliche Teile der Treffersequenz bleiben zwar vom sichtbaren Alignment ausgeschlossen, werden aber im Datenformat mitgeführt und stehen der späteren Analyse zur Verfügung. Schließlich wird der alignierte Block in das Sequenz-Cluster eingefügt, das damit um eine Schrotschusssequenz gewachsen ist (Abb. 9C).



A) Der überstrichene Alignmentabschnitt eines künftigen Cluster-Mitglieds wird per BLAST ermittelt. Die alignierbare Sequenz des neuen Cluster-Mitglieds ist rot gekennzeichnet.

B) Der betroffene Alignmentabschnitt wird aus dem Cluster ausgegliedert und gemeinsam mit der hinzuzufügenden Sequenz an CLUSTAL W übergeben. Mit CLUSTAL W wird aus der Eingabe ein optimales Subalignment errechnet. Eingefügte Gaps sind schematisch dargestellt.

C) Durch Wiedereinfügen des Subalignments in das gesamte Cluster wird das neue Mitglied schließlich sauber in das Cluster eingefügt.

Abb. 9. Sukzessiver Aufbau eines Sequenz-Clusters. Aufrufe der Programme BLAST und CLUSTAL W werden durch Skriptcode eingekapselt, um einen neuen Sequenzabschnitt in ein Cluster überlappender Sequenzfragmente einzufügen.

Die Abfolge der Einzelschritte und eine zyklische Wiederholung der Cluster-Erweiterung werden durch ein Perl-Programm bewerkstelligt. Nach Sättigung des Clusters mit passenden Sequenzen,

kann in einem Alignmenteditor (vgl. Abschnitt 2.4.2) eine Verlängerung der Flanken des Clusters durch Offenlegen der überhängenden Regionen der Schrotschusssequenzen erfolgen. Für den Fall, dass daraufhin weiterer Sequenzzuwachs für das Cluster zu erwarten ist, wird die gesamte Prozedur mit dem neuen Consensus als BLAST-Probe wiederholt.

2.4.2 Repräsentations- und Bearbeitungsformen für Sequenz-Cluster

In der Praxis wird zwischen zwei Repräsentations- und Bearbeitungsformen der Sequenz-Cluster unterschieden. Ein sogenanntes „**Cluster-Projekt**“ wird automatisiert aus den unbearbeiteten Schrotschusssequenzen aufgebaut. Es werden keine Editierungen vorgenommen und keine Sequenzen außerhalb der automatisch definierten Qualitätsfenster (vgl. BONFIELD ET AL. 1995) aufgedeckt. Diese Bearbeitungsform erfüllt die Anforderungen der statistischen Repräsentativität.

In den Fällen, in denen die vollständige Transposonsequenz aus den Sequenzdaten erst ermittelt werden muss, und speziell, wenn dabei Engpässe in der Datenlage auftreten (niedrige Coverage im Alignment, hohe Sequenzvariabilität besonders in Form von InDels, aussageschwache Sequenzdaten z.B. verursacht durch lange poly(A)- oder poly(T)-Segmente) wird auf die Möglichkeiten einer Assemblierung mit dem Programm GAP4 zurückgegriffen, das dem Bearbeitenden umfangreichen Einblick in die Sequenzdaten erlaubt und Freiraum zur Korrektur von Baseninterpretationen ermöglicht. In einem solchen „**GAP4-Projekt**“ wird durch menschliche Interpretation, die der automatischen auch zum derzeitigen Stand der Technik weit überlegen ist, eine Datenlage geschaffen, die höchsten Qualitätsanforderungen gerecht wird. Aus dieser Bearbeitungsform wird eine zuverlässige, fehlerrobuste Consensussequenz gewonnen. In ihr sind auch die Sequenzpolymorphismen, die zwischen den verschiedenen genomischen Kopien der Transposons bestehen, artefaktfrei herausgearbeitet.

2.5 Statistische Analyse

2.5.1 Wahrscheinlichkeitsmodell für Schrotschusstreffer

Es soll die Wahrscheinlichkeit berechnet werden, mit der eine Schrotschusssequenz einen Treffer auf ein beliebiges genomisches Feature liefert. Ein genomisches Feature kann dabei ein Gen, eine Transposonkopie, ein regulatives Element oder ähnliches sein. Wenn man annimmt, dass die Längen aller Schrotschusssequenzen gleich sind, hat jede Sequenz die gleiche Chance P(H), ein genomisches Feature in solcher Weise zu treffen, dass die Identifizierung des Features in der Schrotschusssequenz möglich ist.

$P(H) = N_H / N_\Omega$ mit:

N_H := Zahl der Positionen von Schrotschusssequenzen im Genom, welche die Identifizierung des Features zulassen

N_Ω := Zahl aller möglichen Positionen der Schrotschusssequenz im Genom, mit:

$N_\Omega \approx L_G$ (Größe des Genoms ≈ 34 Mbp, vgl. Abschnitt 3.1.2 ab S. 32)

N_H setzt sich zusammen aus:

$$N_H = n_F \times [\, L_{Sh} - L_O + \max (L_F - L_O, 0) \,] \quad \text{mit:}$$

n_F := Anzahl von gleichwertigen Loci, an denen das Feature im Genom enthalten ist.

L_{Sh} := Länge der Schrotschusssequenzen

L_F := redundante Länge des genomischen Features, wobei eine Überlappung an einer beliebigen Position der Region L_F gleichermaßen die Identifizierung des Features ermöglicht

L_O := minimale Länge der Überlappung zwischen Schrotschusssequenz und dem genomischen Feature. Bei Unterschreitung dieser Länge kann das Feature in der Schrotschusssequenz nicht mehr erkannt werden. Es muss gefordert werden: $L_O < L_{Sh}$.

Weitere Zusammenhänge ergeben sich aus der binomialen Verteilung der Elementarereignisse „Treffer“ mit der Wahrscheinlichkeit P(H) zu „kein Treffer“ mit der Wahrscheinlichkeit 1 - P(H). So ist die zu erwartende Trefferzahl n_H für eine Zahl von Schrotschusssequenzen n_{Sh}:

$$E(n_H) = n_{Sh} \times P(H)$$

In Umkehrung errechnet sich der Maximum-Likelihood-Schätzer für die Zahl der Features basierend auf der Zahl der Schrotschusstreffer nach:

$$ML(n_F) = n_H / [\, n_{Sh} \times P(H \mid n_F=1) \,]$$

wobei $P(H \mid n_F=1)$ die elementare Trefferwahrscheinlichkeit unter der Annahme $n_F=1$ bezeichnet.

Die Wahrscheinlichkeit, ein Feature mindestens einmal oder kein einziges Mal unter n_{Sh} Schrotschusssequenzen anzutreffen, ist:

$$p(n_H \geq 1, n_{Sh}) = 1 - [\, 1 - P(H) \,]^{n_{Sh}}$$

$$p(n_H = 0, n_{Sh}) = [\, 1 - P(H) \,]^{n_{Sh}}$$

Die Beträge anderer Vertrauensintervalle können in Binomialtabellen nachgeschlagen werden.

Zur Veranschaulichung des vorgestellten Modells ein Berechnungsbeispiel zu der Frage: Wie hoch ist die Wahrscheinlichkeit, ein bestimmtes Gen im Genom anhand von Schrotschusssequenzen zu identifizieren? Als Datengrundlage sei der tatsächliche Umfang repräsentativer Schrotschusssequenzen vom *D.-discoideum*-Genom gewählt (s. Abschnitt 3.1.1 ab S. 31). Danach existieren 45150 chromosomale Schrotschusssequenzen (n_{Sh} = 45150) mit einer mittleren Länge von 379 bp (L_{Sh} = 379). Das zugrunde liegende Genom habe eine angenommene Größe von 34 Mbp (L_G = $3{,}4 \cdot 10^7$, vgl. Abschnitt 3.1.2 ab S. 32). Es sei angenommen, dass ein gesuchtes Gen genau einmal im chromosomalen Teil des Genoms vorkommt (n_F = 1). Ein beliebige, 50 bp lange Sequenzregion (L_O = 50) aus dem Bereich der Genlänge von 1000 bp (L_F = 1000) reiche aus, um das Gen per BLAST-Suche eindeutig zu identifizieren. Nach dem beschriebenen Modell ist die Elementarwahrscheinlichkeit, das Gen mit einer einzigen Schrotschusssequenz zu treffen:

$$P(H) = n_F \times \frac{L_{Sh} - L_O + \max(L_F - L_O, 0)}{L_G}$$

$$= 1 \times \frac{379 - 2 \times 50 + 1000}{3{,}4 \cdot 10^{-7}} = 3{,}76 \cdot 10^{-5}$$

Die Wahrscheinlichkeit, das Gen in mindestens einer der 45150 Schrotschusssequenzen anzutreffen ist:

$$p(n_H \geq 1, n_{Sh}) = 1 - [\, 1 - P(H) \,]^{n_{Sh}}$$

$$= 1 - [\, 1 - 3{,}76 \cdot 10^{-5} \,]^{45150} = 0{,}82$$

Das bedeutet, dass bei einer 0,5-fachen Abdeckung des Genoms durch Schrotschusssequenzen unter den beschriebenen Bedingungen eine 82-%-ige Chance besteht, ein existierendes Gen in den Schrotschusssequenzen anzutreffen.

2.5.2 Schätzung von Nukleotidanteil und Kopienzahlen

Der Nukleotidanteil einer Transposonspezies am gesamten Genom wird nicht über das voran beschriebene Treffermodell geschätzt, da neben vollständigen Transposonkopien auch mehr oder weniger große Kopienfragmente über das Genom verteilt sind. Die Länge des Features (L_F in Abschnitt 2.5.1), dessen Häufigkeit geschätzt werden soll, ist variabel. Es wird daher folgende Dreisatzgleichung zur Schätzung angewandt:

$$\frac{ntG_{Tn}}{ntG_{\Sigma}} = \frac{ntSh_{Tn}}{ntSh_{\Sigma}} \Leftrightarrow ntG_{Tn} = ntSh_{Tn} \times \frac{ntG_{\Sigma}}{ntSh_{\Sigma}}$$

ntG_{Tn} := Zahl der genomischen Nukleotide des Features (Transposons)

ntG_{Σ} := Zahl aller Nukleotide des Genom $\approx 3{,}4 \cdot 10^7$

$ntSh_{Tn}$:= Zahl der Nukleotide des Features in Schrotschusssequenzen

$ntSh_{\Sigma}$:= Zahl der Nukleotide aller Schrotschusssequenzen

Setzt man ntG_{Σ} gleich 1, so erhält man aus der Gleichung für ntG_{Tn} den relativen Nukleotidanteil des in Frage stehenden Features am Genom.

Setzt man den absoluten Nukleotidanteil ntG_{Tn} in Bezug zur maximalen Kopiengröße in Nukleotiden, erhält man einen wichtigen Grenzwert für die Schätzung der Kopienzahl des Features: Unter der Annahme, dass die Kopien nicht fragmentiert sind, gilt dieser Grenzwert als Schätzwert für die Zahl der Kopien im Genom.

2.5.3 Berechnung der Sequenzdiversität π

Das Maß π für die Sequenzdiversität geht auf NEI & LI (1979) zurück. Seine Berechnung erfolgt nach:

$$\pi = \sum_{i=2}^{n_A} \sum_{j=1}^{i-1} 2 \times f_i \times f_j$$

n_A := Anzahl aller verschiedenen Allele an der polymorphen Sequenzposition

f_i, f_j := relative Häufigkeit des Allels i oder j

i,j := Iteratoren für die Allele einer polymorphen Sequenzposition

In π drückt sich die Dichte von Polymorphismen und die lokale Verschiedenartigkeit von Sequenzen quantitativ aus. In anschaulicher Formulierung: Es gebe ein Pool von Sequenzen, die sich alle untereinander sehr ähnlich sind, sich aber teilweise durch Polymorphismen voneinander unterscheiden. Für den Fall, dass man zwei beliebige Sequenzen aus dem Pool herausgreift, sie aligniert und miteinander vergleicht, gibt π an, mit welcher Wahrscheinlichkeit der paarweise Sequenzvergleich an einer einzelnen Position des Alignments einen Unterschied zu Tage fördert.

2.5.4 Maß für die Assemblierbarkeit einzelner Transposonspezies

Die Sequenzdiversität π (Abschnitt 2.5.3) gibt keinen direkten Aufschluss über die Schwierigkeit, eine gewisse Anzahl von Repeatloci anhand von Polymorphismen zu unterscheiden, denn die Schwierigkeiten nehmen auch linear mit der Anzahl der Repeat-Kopien zu, welche ihrerseits aber nicht in die Berechnung von π einfließt. Aus diesem Grund sei eine weitere Größe, R_A, eingeführt, die quantitativ beschreibt, welchen „Widerstand“ ein Pool von Transposonkopien gegen eine korrekte Assemblierung setzt.

$$R_A = \frac{N_{Tn}}{n_P}$$

n_P := Anzahl aller Polymorphismen für die Kopien der Transposonspezies

N_{Tn} := Zahl der genomischen Basenpaare, die der Transposonspezies zuzuschreiben sind

Anschaulich formuliert, ist das Maß R_A eine Näherung für die Sequenzstrecke, die eine Schrotschusssequenz überspannen muss, um so viele polymorphe Ausprägungen zu sammeln, dass sie trotz ihres repetitiven Charakters einzigartig wird und dadurch einer spezifischen Kopie des betreffenden Transposons zugeordnet werden kann.

2.6 Gensuche und Proteinanalyse

2.6.1 Das Genanalyseprogramm GeneID

Das Genanalyseprogramm GENEID (PARRA ET AL. 2000) hat sich als flexibles Werkzeug erwiesen, um die Gensuche in genomischer Sequenz von *D. discoideum* zu implementieren. Das Programm bietet über eine textbasierte Parameter-Datei ein offenes Interface, um eine artspezifische Optimierung von relevanten Parametern der Genanalyse zu ermöglichen.

Konkret basiert die Anpassung von GENEID an *D. discoideum*, die in Kooperation mit G. PARRA und R. GUIGÓ (IMIM, Barcelona) erfolgte, auf einem Datensatz von komplett annotierten Genstrukturen, die über das Datenbankkonsortium GenBank/EMBL/DDBJ (BENSON ET AL. 2000) bis Oktober 2000 veröffentlicht wurden. Nachdem Redundanzen eliminiert wurden, die durch Anwendung von BLASTP

(E=1e-20) nachweisbar sind, verblieben 140 Datenbankeinträge. Aus der Analyse der Trainingsdaten ergaben sich folgende Ressourcen für die Gensuche: Eine Tabelle der Tupel-Übergangswahrscheinlichkeiten für CDS und nicht-CDS als Grundlage für einen Entscheidungsalgorithmus auf Grundlage eines Markov-Modells, Positional-Weight-Matrizen für Translationsstart und -stop sowie die Splice-Signale (PARRA ET AL. 2001).

2.6.2 Genstrukturanalyse durch Sequenzvergleich

Eine Möglichkeit zur Analyse der Genstruktur liegt im Vergleich zweier homologer DNA-Sequenzen, die jeweils in allen denkbaren Leserastern in ihre Peptidsequenz translatiert wurden. Das Programm BLAST (engl. „Basic Logical Alignment Search Tool“) eignet sich hervorragend, um in den Peptidsequenzen lokale Bereiche von kreuzweiser Sequenzähnlichkeit zu identifizieren. Es wird also ein BLAST-Vergleich zwischen dem Paar von Sequenzen durchgeführt – mit der einen Sequenz als Sonde und der zweiten Sequenz in der BLAST-Datenbank. Die BLAST-Parameter werden in der Weise angepasst, dass ein Optimum zwischen Sensitivität und Spezifität der Treffer erzielt wird:

- Hohe Sensitivität wird mit einer primären Suchwortlänge W=2 und geringem Schwellenscore S=30 erreicht. Mit der Definition von hspmax=10000 wird das Verhalten von BLAST unterdrückt, nur eine begrenzte Zahl der besten HSPs zu berichten.
- Gaps in den lokalen Alignments werden durch die Einstellung Q=24 möglichst unterdrückt.
- Eine PAM100-Matrix wird in der Weise angepasst, dass ein Paar von Sequenzsymbolen eine hohe Strafe enthält (Score -24), wenn ein Stopcodon im lokalen Alignment enthalten ist.

Die Analyse und Darstellung der BLAST-Treffer in automatisierter Weise wird möglich durch eine Umwandlung der hierarchisch gegliederten originalen Ausgabeform in eine tabellarische Datenstruktur. Die Treffer werden dann nach den gewählten Kriterien gefiltert und in einem geeigneten Format – hier GFF (DURBIN & HAUSSLER) – ausgegeben. Eine graphische Darstellung ist daraufhin mit Hilfe des Programms GFF2PS möglich (ABRIL & GUIGÓ 2000).

2.6.3 Maße lokaler Proteinähnlichkeit

Funktionsrelevante Strukturinformation ist niemals homogen über eine Proteinsequenz verteilt, sie häuft sich in Domänen und konzentriert sich dort an einzelnen invarianten Aminosäureresten, wie etwa am aktiven Zentrum eines Enzyms oder Schlüsselpositionen, die maßgeblich zur Bildung der Tertiärstruktur beitragen. Der Sequenzvergleich von homologen Peptiden liefert Aufschluss über die Verteilung funktioneller Strukturinformation über die Proteinsequenz.

Für ein Proteinsequenzalignment, in dem spaltenweise die homologen Sequenzsymbole jeweils aller Sequenzen übereinander stehen, lässt sich die Analyse des lokalen Konservierungsgrads sukzessiv Spalte für Spalte abarbeiten. Ich habe drei verschiedene Methoden angewandt, um den Konservierungsgrad einer Alignmentspalte zu quantifizieren:

- Es wird der relative Anteil der Sequenzen in der Alignmentspalte bestimmt, deren Sequenzsymbol

mit der Consensussequenz übereinstimmt. Diese Methode soll im Folgenden **„Methode der relativen Identität“** genannt werden.

- Durch die **„Entropie-Methode“** findet eine Bewertung der Sequenzdiversität aus der Sicht der Informationstheorie statt. Kommen in einer Alignmentspalte alle möglichen Sequenzsymbole mit gleicher Häufigkeit vor, so befindet sich die Spalte im Zustand größtmöglicher Unentschiedenheit, die Entropie erreicht ihren Maximalwert. Jeder andere Verteilungszustand mit einer Anhäufung einzelner oder weniger Symbole lässt sich quantitativ durch eine Abnahme der Entropie H beschreiben. Aus Gründen der Proportionalität zu den anderen hier vorgestellten Bewertungskriterien wird hier das Maß -H nach folgender Gleichung berechnet:

$$-H = -\sum_{i=1}^{n} h_i \times \ln (h_i)$$

n := Anzahl aller Arten von Sequenzsymbolen in der Alignmentspalte

h_i := relative Häufigkeit des Symbols i

- Die **„Matrix-Methode“** erlaubt gegenüber der „Entropie-Methode“ zusätzlich, die auftretenden Sequenzsymbole (Aminosäuren) nach ihrer biologischen bzw. chemischen Ähnlichkeit zu bewerten. Eine solche Ähnlichkeitsbewertung ist mit Hilfe von Austauschmatrizen, beispielsweise den klassischen PAM-Matrizen (DAYHOFF ET AL. 1978), möglich. Die praktische Herangehensweise liegt in einem kreuzweisen Vergleich aller in einer Alignmentspalte beobachteten Sequenzsymbole. Mittelung der Matrixwerte führt zu einem Wert für die durchschnittliche Distanz aller Aminosäuren in der Alignmentspalte.

$$\bar{d} = \frac{2}{n\,(n-1)} \times \sum_{i=2}^{n} \sum_{j=1}^{i-1} M\,(s_i, s_j)$$

n := Anzahl aller Sequenzen und deren Symbole in der Alignmentspalte

s_i,s_j := Symbol der Sequenz i oder j

M(x,y) := Distanz zwischen den Symbolen x und y laut Austauschmatrix

Auftragungen der beschriebenen lokalen Diversitätsmaße gegen die Alignmentpositionen werden geglättet, um die Übersichtlichkeit der graphischen Darstellung zu verbessern. Dazu wird in einem kontinuierlich fortbewegten Fenster von 6 Sequenzpositionen der Mittelwert des berechneten Maßes gegen den Mittelpunkt des Sequenzfensters aufgetragen.

2.6.4 Berechnung phylogenetischer Bäume

Ein Alignment von Polypeptidsequenzen wird mit CLUSTAL W erstellt. Dabei werden die voreingestellten Parameter verwendet außer einem Wert für Gap-Bestrafung von 5,0 anstatt 10,0. Vor der Baumberechnung werden alle Bereiche mit Gaps und solche ohne nachvollziehbare Sequenzähnlichkeit aus dem Alignment eliminiert. Zur Berechnung des phylogenetischen Baums wird die Parsimony-Methode aus dem PHYLIP-Paket (FELSENSTEIN) herangezogen.

3 ERGEBNISSE

3.1 Genomsequenzierung und Genomgröße

3.1.1 Schrotschusssequenzierung und -assemblierung

Die Sequenzierung von Klonen der Schrotschussbibliotheken lieferte reichlich Sequenzmaterial für die Analysen (Tab. 10). Eine wichtige Grundlage für die statistische Interpretation der Sequenzierungsergebnisse bildet die repräsentative, genomische Schrotschussbibliothek mit dem Namen „JAX4“. Von ihr wurden 49260 Sequenzen erzeugt (Abschnitt 2.2.3 ab S. 19), die insgesamt 18,67 Mbp informative Sequenz umfassen. Die durchschnittliche Länge der Schrotschusssequenzen beträgt 379 bp. Der Gehalt der Bibliothek an DNA von extrachromosomalen Elementen (rDNA, mtDNA) wurde mittels BLAST auf mindestens 8,33 % bestimmt, es entfallen also 45150 Sequenzen auf den chromosomalen Teil des Genoms.

Tab. 10. Sequenzressourcen aus der Schrotschusssequenzierung.

Bibliothek	Sequenzen	extrachromosomal [%]	Spezifität	
JAX4	49260	8,33	–	
JC1	129980	1,39	Chr1	80 %
JC2	158468	10,11	Chr2	47 %

Die Sequenzen wurden durch beidseitiges Sequenzieren von Schrotschussklonen erzeugt. Die meisten Schrotschusssequenzen sind demnach topologisch gekoppelt mit einer zweiten Sequenz, die vom gleichen Klon stammt. Dieser Umstand hat einen wesentlichen Einfluss auf die Varianzbetrachtung von Schrotschusstreffern. Aus den 20695 beidseitig und 7870 einseitig sequenzierten Klonen ergibt sich für die genomische Position der erzeugten Sequenzen ein mittlerer statistischer Freiheitsgrad von 0,58.

Der Umgang mit Sequenzdaten und das anschließende Sequenz-Clustering einschließlich Analyse bringt die Auseinandersetzung mit einer Fülle von verschiedenen Formaten und Dienstprogrammen mit sich. Zur Automatisierung von Datenflüssen habe ich verschiedene Programme entwickelt. In aller Kürze dargestellt, handelt es sich dabei um:

- Hantieren mit Sequenzdaten in den Formaten fastA (Pearson), GenBank, Staden Experiment, GFF. Kommandozeilenorientierter Hochdurchsatz von Umformatierung, Schneiden/Verknüpfen, Selektieren, Tupelanalyse, Motivsuche.
- Verarbeiten von Alignmentdaten im CLUSTAL-W-Format und GAP4-Datenbankformat. Kommandozeilenorientierter Durchsatz von Umformatierung, Teilen/Verknüpfen, Topologiereport und

Scaffold-Berechnung, Consensusbildung, Polymorphismenanalyse.

- Statistische Berechnungen: Stichprobenanalyse, Gaußsche Grenzwerte, Binomialtabellen.
- Routinen zur Assemblierung.

Die Assemblierung der Schrotschusssequenzen ist inzwischen besonders für Chromosom 2, das etwa 25 % des Genoms ausmacht, weit vorangeschritten. Die Chromosomensequenz wird in Kürze veröffentlicht (GLÖCKNER ET AL. 2002). Daneben wurden ausgewählte Loci von verschiedenen Genomabschnitten assembliert: Genloci von V4a / V4b (Abschnitt 3.5.4 ab S. 53), Genloci der Aktingene (Abschnitt 3.4.1 ab S. 44) und ausgedehnte repetitive Loci (Abschnitt 3.7.3 ab S. 65).

3.1.2 Schätzung der Genomgröße

Aus der voran beschriebenen Sequenzierung steht ein repräsentativer, chromosomaler Satz von Schrotschusssequenzen zur Verfügung, der 45150 Sequenzen umfasst. Daneben habe ich ein auf Chromosom 6 kartiertes YAC mit einer Insert-Größe von 145194 bp durchgehend sequenziert und assembliert. Sein mittlerer G/C-Gehalt beträgt 23,2 %, steht also in gutem Einklang mit dem genomischen Mittel von 23 % (FIRTEL & BONNER 1972), so dass angenommen werden kann, dass das YAC-Insert eine repräsentative Stichprobe des Genoms darstellt. Ein Teil des YAC-Inserts entfällt auf längere Abschnitte repetitiver Sequenzen (2 Kopien von TRE5-A.2 mit je 2 kbp Länge). Sie können keine zweifelsfreien BLAST-Hits liefern und werden deshalb von der Analyse ausgeschlossen. Auf die verbleibenden 141247 bp YAC-Sequenz passen 187 genomische Schrotschusssequenzen. Nach Dreisatz ergibt sich daraus eine geschätzte Genomgröße von 45150 / 187 × 141247 bp = 34,1 Mbp (vgl. Abschnitt 2.5.2 ab S. 27).

Zur Ermittlung der Schätzungsungenauigkeit werden diejenigen abweichenden YAC-zu-Genom-Größenverhältnisse gesucht ($p_l < p = 0{,}00414 < p_u$), die die beobachteten Trefferdaten nicht mehr in einem einseitig begrenzten Vertrauensintervall liefern können. Der Umstand, dass die Treffer der Schrotschusssequenzen auf das YAC nicht gänzlich unabhängig voneinander zu werten sind, wird durch einen Freiheitsgrad von 0,58 berücksichtigt. Er wird direkt korrigierend auf die Werte von n und x angewandt wird – damit ergibt sich: n = 26080, x = 108 (auf ganzzahliges x gerundet). Für diese nunmehr normalisierte Situation können die gesuchten Werte für p_l und p_u aus Binomialtabellen abgelesen werden. Die entsprechenden Binomialtabellen sind im Anhang 6.1.1 ab S. 93 aufgeführt, und sie liefern für das Ereignis x = 108 aus n = 26080 die Intervallgrenzen $p_l = 0{,}00394$ und $p_u = 0{,}00436$ für ein jeweils einseitiges Verlassen des 68-%-Vertrauensintervalls (Standardfehlerintervall der Normalverteilung). Damit ergeben sich die Grenzen für die Schätzgröße: 32,3 und 35,9 Mbp. Die Schätzung einschließlich Vertrauensintervall lautet **34,1 ± 1,8 Mbp**.

3.2 Transposonsequenzen

3.2.1 Benennungsschema für tRNA-Gen-assoziierte non-LTR-Retrotransposons

In der Literatur wurden drei verschiedene, nah verwandte tRNA-Gen-assoziierte non-LTR-Retro-

transposons in ihrer vollständigen Sequenz beschrieben: Tdd-3, DRE, RED (Rückblick in WINCKLER 1998). Da die Zahl der Mitglieder der Familie durch die Neubeschreibungen im Rahmen dieser Arbeit auf sieben ansteigt, erscheint es sinnvoll, eine systematische Nomenklatur einzuführen (s. auch SZAFRANSKI ET AL. 1999). Es wird dazu vorgeschlagen, das Namenskürzel TRE (für „tRNA Gene-Targeted Retroelement“) als Namensstamm einzuführen, gefolgt von einer Ziffer, die angibt, ob das Element stromaufwärts („5“) oder stromabwärts („3“) des tRNA-Gens inseriert. Es wird an mehreren Stellen des Textes noch die Rede von der Konserviertheit dieses Merkmals sein. Schließlich wird an den Namen, durch Bindestrich getrennt, ein laufender Buchstabe angehängt, um zu einem eindeutigen Namen zu gelangen. Es ergeben sich die folgenden Umbenennungen von bereits in der Literatur beschriebenen Transposons: Tdd-3 wird zu TRE3-A, RED zu TRE3-B, und DRE heißt nun TRE5-A. Es wird auch eine Regel zur Namensbildung für Subtypen einzelner Transposonspezies vorgeschlagen: Durch einen Punkt getrennt, wird eine laufende Zahl an den Namen angehängt. Eine „1“ für das Stammelement, weitere Zahlen für abgeleitete Formen. Beim TRE5-A ist eine Unterscheidung von Subtypen angebracht. Es gibt eine Stammform mit voll ausgebildeten ORFs, TRE5-A.1 (alter Name DREa), und eine häufig anzutreffende Subform, TRE5-A.2 (alter Name DREb), die durch eine große interne Deletion stark verkürzt ist und sich durch verschiedene charakteristische Basensubstitutionen von der Stammform unterscheidet (MARSCHALEK ET AL. 1992b).

3.2.2 Korrektur und Ergänzung von Sequenzen beschriebener Transposons

Von allen bereits in der Literatur beschriebenen Transposons aus *D. discoideum* wurden Sequenz-Cluster aus den Schrotschusssequenzen des Genomprojekts angefertigt. Diese Cluster wurden vorrangig in Hinblick auf die statistische Auswertung von Sequenzanteil und Diversität angefertigt. Darüber hinaus waren sie jedoch auch geeignet, zuvor nur partiell beschriebene Transposonsequenzen zu vervollständigen und nicht-repräsentative Sequenzveröffentlichungen durch repräsentative Daten zu ergänzen.

Im Falle der zuvor veröffentlichten Sequenz des **TRE3-B** (Acc.No. AF067198) wurde festgestellt, dass sie im Vergleich zum Consensus der Transposonkopien (Acc.No. AF134170) zwei größere Deletionen und eine umfangreiche Inversion enthält (Abb. 11). Die Analyse der jetzt vorliegenden Schrotschussdaten zeigt, dass die veröffentlichte Sequenz tatsächlich als einmalige, degenerierte Kopie im Genom von *D. discoideum* existiert, also nicht auf Klonierungs- oder Assemblierungsartefakten beruht.

Vom **TRE3-C** war vor dieser Untersuchung nur ein Sequenzstück von etwa 300 bp unter dem Namen Tdd-2 veröffentlicht worden (POOLE & FIRTEL 1984, Acc.No. K02642). Die Identifizierung der vollständigen Sequenz dieses Elements im Verlauf dieser Arbeit erfolgte unabhängig von der Kenntnis dieses Fragments. Wie im Abschnitt 3.2.3 beschrieben, wurde nach genomischen Loci gesucht, die zwar auf Proteinebene Sequenzähnlichkeit mit dem 3’-Ende eines bekannten TRE-Elements aufweisen, aber auf Nukleotidebene signifikante Sequenzähnlichkeit vermissen lassen (vgl. Abschnitt 2.4 ab S. 23). Unter den Kandidaten neuer TRE-Elemente ist eines, das nach Vervollständigung in 5’-Richtung am 5’-Ende sequenzidentisch mit dem veröffentlichten Tdd-2-Fragment ist. Die Consensussequenz von TRE3-C wurde unter Acc.No. AF134171 abgelegt.

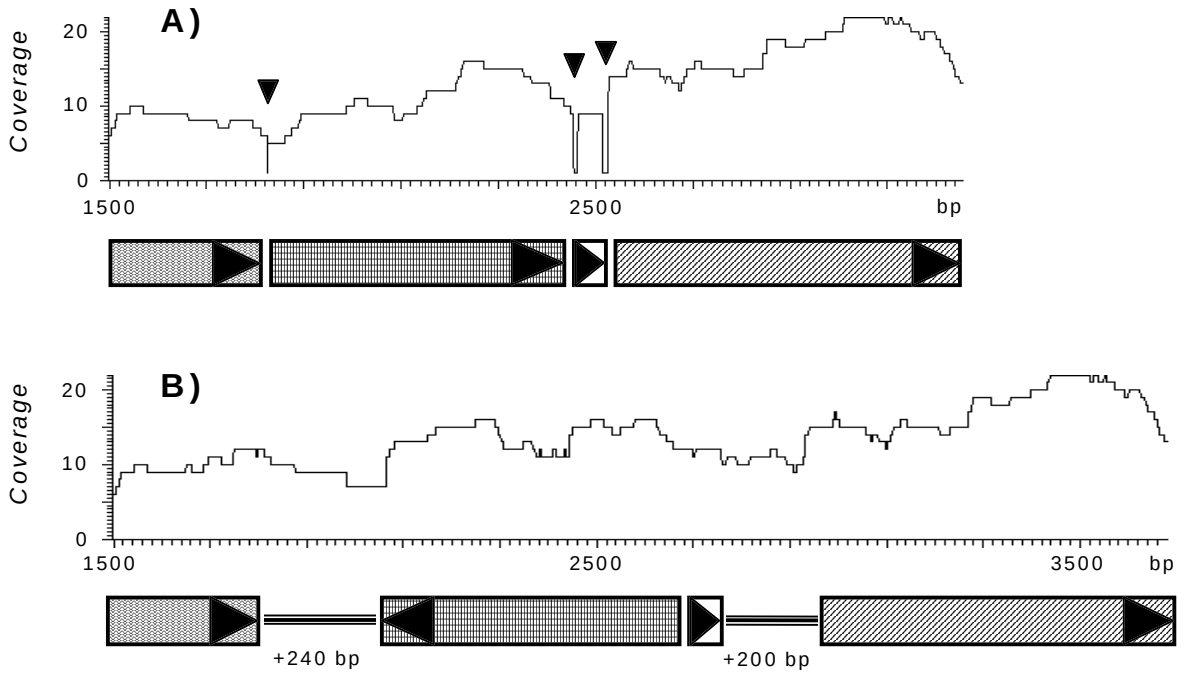


Abb. 11. Diskrepanz zwischen veröffentlichter Einzelklon-Sequenz (Acc.No. AF067198) und der auf Schrotschussdaten basierenden Consensussequenz für TRE3-B. **A)** Konstruktion eines Sequenz-Clusters auf Basis der Sequenz AF067198 führt zu auffälligen Lücken in der Coverage des Clusters, gekennzeichnet durch Dreiecke. **B)** Durch Ausdehnen des Sequenz-Clusters nach Maßgabe der Schrotschusssequenzen ergibt sich ein neues, konsistentes Alignment. Die schematischen Blöcke unter dem Graph der Coverage zeigen die Veränderungen gegenüber der Sequenz AF067198 an. Es wurden zweimal je ca. 200 bp Sequenz hinzugefügt, ein zentraler Block von ca. 700 bp wurde invertiert.

Der **HindIII-Repeat (H3R)** war bereits vor dieser Untersuchung beschrieben und stand im Verdacht, den solo-LTR eines bislang nicht identifizierten autonomen LTR-Retrotransposons darzustellen (WINCKLER 1998). Im Verlauf der fortschreitenden Assemblierung von Chromosom 2 von *D. discoideum* durch unsere Arbeitsgruppe wurden Multi-Copy-Sequenzen auffällig, die ausnahmslos mit Loci von tRNA-Genen und einer Kopie des H3Rs assoziiert waren. Bei der anfänglichen Sequenzanalyse der repetitiven Contig-Abschnitte fiel in direkter Nachbarschaft zur H3R-Sequenz ein übereinstimmender guaninreicher Abschnitt auf – eine hochsignifikante Erscheinung bei dem gegebenen hohen A/T-Gehalt des *D.-discoideum*-Genoms –, der an den typischen Polypurintrakt von LTR-Retrotransposons denken lässt. Somit waren die ersten Indizien gesammelt, dass es sich bei den beobachteten Multi-Copy-Sequenzen tatsächlich um die Fragmente eines autonomen LTR-Retrotransposons handelt, dessen LTR-Sequenz mit der des H3Rs identisch ist. Das neu entdeckte autonome Element wurde wegen seiner Zugehörigkeit zu den Gypsy-ähnlichen LTR-Retrotransposons „*Dictyostelium* Gypsy-Like Transposon“ (DGLT-A) genannt (Consensussequenz unter Acc.No. AF134171).

3.2.3 Auffinden neuer Transposons

Die Identifizierung neuer Transposonspezies habe ich durch drei unabhängige methodische Ansätze

verfolgt. Ein **kreuzweiser Vergleich von Schrotschusssequenzen** wurde mit dem BLASTN-Programm durchgeführt. Alle Schrotschusssequenzen aus der gesamtgenomischen Bibliothek (vgl. Abschnitt 3.1.1 ab S. 31) wurden per BLASTN-Vergleich gegen die bekannten Sequenzen der extrachromosomalen Genomanteile gefiltert. In den verbleibenden Sequenzen wurden Anteile von homopolymeren A- und T-Abschnitten sowie einfach-repetitive Anteile (Mikrosatelliten) maskiert. Die so behandelten Schrotschusssequenzen wurden per BLASTN-Suche jeweils mit allen anderen Schrotschusssequenzen verglichen. Eine hohe Trefferrate einer Schrotschusssequenz auf andere Schrotschusssequenzen wurde als Kriterium für die repetitive Natur der enthaltenen Sequenz benutzt. Teilregionen der Elemente Tdd-5, thug-S und thug-T wurden auf diese Weise aufgefunden. Daneben wurden aber auch ca. zehn repetitive Sequenzabschnitte identifiziert, für die bisher nicht geklärt werden konnte, ob es sich um Teile von Transposons oder Multigenfamilien handelt. Die Länge dieser unklassifizierten Sequenzen beträgt um 1 kbp, die geschätzte Kopienzahl beträgt maximal zehn, so dass ihre Bedeutung für die Genomassemblierung und die Betrachtung des Transposonanteils am Genom eine untergeordnete Rolle spielen.

Für alle bekannten Transposonfamilien wurde eine **Suche nach Transposonhomologen** unternommen. Dazu wurde die Proteinsequenz mindestens eines Mitglieds jeder bekannten Transposonfamilie verwendet, um per TBLASTN nach Sequenzähnlichkeiten unter den Schrotschusssequenzen zu suchen. Unter den erhaltenen Treffern wurden mit BLASTN diejenigen herausgefiltert, die Fragmente von bekannten Transposonspezies repräsentierten. Übrig blieben TBLASTN-Treffer unter den Schrotschusssequenzen, die mutmaßlich neuen Transposonspezies zuzuschreiben waren. Auf diese Weise wurden mehrere neue Elemente der TRE-Familie der non-LTR-Retrotransposons identifiziert: TRE3-C (s. auch Abschnitt 3.2.2), TRE3-D, TRE5-B und TRE5-C. Darüber hinaus wurden verschiedene Fragmente von TRE-ähnlichen Sequenzen gefunden, die durchweg kurz sind – etwa 500 bp – und jeweils nur in einer bis wenigen einander ähnlichen Kopien existieren. Es handelt sich vermutlich um Reste von ausgestorbenen Spezies dieser Familie. Da sie weder für die Genomassemblierung noch für die Gesamtbeurteilung des Transposongehalts in *D. discoideum* von wesentlicher Bedeutung sind, wurden diese fragmentarischen Transposonspezies nicht weiter bearbeitet.

Die **Analyse von Transposonenden und internen Transposonbruchpunkten** ist eine Methode, die mit sehr hoher Effizienz zur Identifizierung neuer Transposonspezies geführt hat. Der Methode liegen zwei Ansatzpunkte zugrunde. Der eine lautet: Findet man in einer Transposonkopie einer bekannten Spezies eine scharf begrenzte umfangreiche Insertion, handelt es sich mit hoher Wahrscheinlichkeit um die Insertion eines transposablen Elements. Der Weg zur Identifizierung einer neuen Transposonspezies aus Schrotschusssequenzen führt hier über die Analyse von internen Transposonbruchpunkten. Dabei muss in jedem Einzelfall ausgeschlossen werden, dass es sich um die Verkürzung einer Transposonkopie durch Rekombination handelt. Das Erscheinungsbild der gesuchten verschachtelten Transposontopologien in Schrotschusssequenzen wird im Zusammenhang mit der Ermittlung von TSD-Sequenzen näher beschrieben (Abschnitt 3.3.3 ab S. 42). Der zweite Ansatzpunkt ergibt sich aus der Beobachtung, dass alle Transposons, die nicht strikt sequenzspezifisch an bestimmten Orten des Genoms inserieren, signifikant häufig in vorhandene Transposonkopien inserieren (vgl. Abschnitt 3.6.2 ab S. 62). Demnach ist jede Transposonkopie

wahrscheinlich von weiterer Transposonsequenz flankiert. Die Frage, ob dieser wahrscheinliche Fall für eine gegebene Transposonflanke tatsächlich zutrifft, kann durch sorgfältige Sequenzanalyse beantwortet werden. Das Aufspüren von Transposonverschachtelungen hat zur Entdeckung verschiedener neuer Transposonspezies beigetragen: DCLT-A, DDT-A, DDT-B, DDT-S sowie wesentliche Teile der Elemente thug-S und thug-T.

Die Consensussequenzen aller neu aufgefundenen transposablen Elemente wurden bei GenBank deponiert. Die Acc.Nos. können der Tab. 43 auf S. 74 oder Anhang 6.3 ab S. 98 entnommen werden. Die Klassifizierung zu Familien und Klassen ist ebenfalls aus Tab. 43 zu ersehen. Die eingehende Sequenzanalyse wird im Abschnitt 3.5 ab S. 47 beschrieben.

3.3 Analyse von Target-Site-Duplikationen

Die Länge der „Target-Site-Duplikation“ (TSD) ist ein spezifisches Merkmal jedes Transposons. Daher war es ein Ziel der vorliegenden Arbeit, alle in *D. discoideum* angetroffenen Transposonspezies bezüglich der Länge ihrer TSD zu charakterisieren.

3.3.1 Transposonflankenpaare durch Anwendung der inversen PCR

Der Experimentverlauf der inversen PCR (Abschnitt 2.2.6 ab S. 22) hängt von einer unbekannten Größe ab, nämlich der Position einer Restriktionsstelle. Ist die unbekannte Restriktionsstelle sehr nah zum Ende der bekannten Sequenz, bringt das Experiment geringen Sequenzzuwachs. Ist die unbekannte Restriktionsstelle sehr weit jenseits des Sequenzabbruchs, ist der intramolekulare Verlauf der Ligation eher unwahrscheinlich, und die Amplifizierung der großen Region zwischen den beiden Primerstellen wird wahrscheinlich scheitern. Der **Wahl des Restriktionsenzyms** kommt also eine wesentliche Bedeutung für die Erfolgsaussicht der inversen PCR zu. Eine wichtige Orientierungshilfe ist das Einbeziehen von Information über die Häufigkeitsverteilung von Erkennungsmotiven der verfügbaren Restriktionsenzyme. Wenn möglich, sollte dabei die organismenspezifische Basenverteilung, oder besser noch die organismenspezifische Motivverteilung zugrunde gelegt werden. Die entsprechenden Werte für *D. discoideum* wurden aus den größeren Contigs (> 5000 bp) der akkumulierten Assemblierungsergebnisse (Abschnitt 3.1.1 ab S. 31) ermittelt. Als Datengrundlage für die verfügbaren Restriktionsenzyme diente die Datenbank REBASE (ROBERTS & MACELIS 2001). In einer Übersicht sind die kommerziell verfügbaren Enzyme zusammengestellt, welche die methodischen Voraussetzungen erfüllen und deren Schnittstellenabstand in der Genomsequenz im Mittel 500 bis 2600 bp beträgt (Tab. 12). Diese Auswahl von Restriktionsenzymen verspricht nach den statistischen Überlegungen die höchsten Erfolgsquoten.

Der Verlauf der DNA-Präparation wird durch **PCR-Kontrollen** verfolgt (Tab. 13). Aufgrund der relativ geringen DNA-Konzentrationen geht die DNA besonders leicht im Verlauf der Reinigungsprozeduren verloren.

Tab. 12. Auswahl kommerziell verfügbarer Typ-II-Restriktionsendonukleasen, die versetzte Einzelstrangbrüche setzen (sog. „klebrige Enden") und deren Erkennungsmotive im Bereich der später überhängenden Enden eindeutig definiert sind (kein sog. „Wobbling"). Fett hervorgehoben sind Restriktionsenzyme, die im Verlauf dieser Arbeit zur inversen PCR verwendet wurden. Drei verschiedene Methoden liefern Werte für den mittleren Motivabstand von Erkennungsmotiven: **Δ_{RE} allgemein** – berechneter mittlerer Motivabstand unter der Annahme von Gleichverteilung der vier Basen. **Δ_{RE} nach Basenverteilung** – berechneter mittlerer Motivabstand auf Basis der tatsächlichen Basenverteilung in *D. discoideum*. **Δ_{RE} empirisch** – beobachteter mittlerer Motivabstand in Contig-Sequenzen von *D. discoideum*. Er ist das Sortierkriterium der Tabelle.

Restriktionsenzym	Erkennungsmotiv	Δ_{RE} allgemein	Δ_{RE} nach Basenverteilung	Δ_{RE} empirisch
...				
NlaIII	,CATG'	256	503	572
TaqI	T'CG,A	256	503	727
TatI	W'GTAC,W	1024	854	967
MfeI	C'AATT,G	4096	3417	1004
TscI	'ACGT,	256	503	1108
BstYI	R'GATC,Y	1024	2015	1478
BfaI	C'TA,G	256	503	1486
NspI	R,CATG'Y	1024	2015	2028
BclI	T'GATC,A	4096	3417	2326
PacI	TTA,AT'TAA	65536	2115	2541
AciI	C'CG,C	256	2761	2580
...				

Tab. 13. Primerpaare für PCR-Kontrollen, die die Existenz von genomischer DNA aus *D. discoideum* während der Durchführung präparativer Methoden nachzuverfolgen erlauben. Die Auswahl befriedigt insbesondere das Bedürfnis nach restriktionsresistenten PCR-Markern (Spalte 1).

Restriktionsenzym	Primerpaar	Produkt [bp]	Locus
BclI	DDT-A.2220 / DDT-A.3660R	1450	DDT-A
BfaI	DDT-A.2220 / DDT-A.3660R	1450	DDT-A
BstYI	DDT-A.2220 / DDT-A.3660R	1450	DDT-A
MfeI	pahA.A / pahA.B	780	pahA
NspI	DDT-A.3790 / DDT-A.4584R	800	DDT-A
TatI	DDT-A.981 / DDT-A.2415R	1400	DDT-A
TscI	pahA.A / pahA.B	780	pahA

Die inverse PCR wurde außer zur Identifizierung von Transposonflanken auch im Rahmen des Genomprojekts zur Vervollständigung genomischer Loci eingesetzt, für die eine Abdeckung mit Klonen der Schrotschussbibliotheken fehlt (nicht dargestellt). Die dabei gesammelten Ergebnisse wurden bei nachfolgenden Experimenten als Positivkontrolle eingesetzt, um die Verwendbarkeit der

Template-Präparate zu bestätigen (Tab. 14).

Tab. 14. Positivkontrollen für Template-Präparate zur inversen PCR. Die Primersequenzen sind im Abschnitt 2.2.4 ab S. 20 definiert. Abkürzung: RE = Restriktionsenzym.

RE: Template generieren	RE: Template linearisieren	Primerpaar(e)	Produkt [bp]	Locus
BfaI	BglII	pahA.C / pahA.A	950	pahA
BstYI	-	JP3222/JP3223 + JP0013/JP0186	930	Contig JAX4b14d09.r1
MfeI	AclI	JP0179 / JP0180	440	Contig JC2a113f09.s1
NspI	NsiI	JP0137 / JP0138	1100	Flanke von gpbA
TatI	-	JP3265/JP3266 + JP3240/JP3267	457	Contig JC2a177b02.r1

Die Anwendung der inversen PCR zur Identifizierung von Flankenpaaren von Transposons beinhaltet eine geringfügige Abwandlung des Grundschemas (Abschnitt 2.2.6 ab S. 22). Die Transposonkopie stellt hier das bekannte Sequenzstück dar, und beiderseits der bekannten Sequenz schließt sich unbekannte Sequenz an, die es zu ermitteln gilt. Deshalb wird ein Restriktionsenzym gewählt, das in der Transposonsequenz überhaupt nicht schneidet. Der experimentelle Prozedur bleibt davon unbeeinflusst.

3.3.1.1 Identifizierung von Flankenpaaren des Transposons thug-T

Für die Consensussequenz des Elements thug-T gibt es zwei nicht-schneidende Typ-II-Endonukleasen unter der methodisch erschlossenen Auswahl von Enzymen (vgl. Tab. 12 ab S. 37): BclI und NspI. Es wurden ungeöffnete invers-PCR-Templates verwendet, und in Nested-Primer-Ansätzen wurde mit den Primerpaaren thug-T.74R / thug-T.693 sowie thug-T.2R / thug-T.820 amplifiziert. Aus der Primerbenennung geht die Position der Primer in der 1132 bp langen Consensussequenz des thug-T hervor.

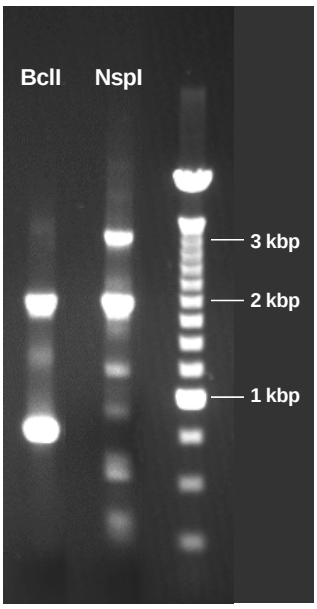


Abb. 15. Amplikons der inversen PCR zur Identifizierung von Flankenpaaren des Transposons thug-T.

Abb. 15 zeigt das Ergebnis der PCR-Reaktionen. Auf beiden Templates werden jeweils mindestens zwei Produkte erhalten (BclI-Template: Produkte von ca. 750 und 2000 bp Länge; NspI-Template: Produkte von ca. 2000 und 3000 bp Länge neben mehreren schwachen Produkten zwischen 0 und 1500 bp Länge). Nach Trennung im Agarosegel wurden jeweils DNA-Banden der Größenbereiche 600-1000 bp, 1500-2500 bp (BclI-Template) und 0-1500 bp, 1500-2500 bp, 2500-4000 bp (NspI-Template) ausgeschnitten, die DNA extrahiert, im Vektor pDrive (Qiagen) kloniert und mit den Primern M13-fwd und M13-rev sequenziert. Die Sequenzen der Einzelklone wurden nach hoher Sequenzüberein-

stimmung zu einer Consensussequenz assembliert. Die resultierenden fünf Consensussequenzen sind in Anhang 6.2.1 ab S. 95 vollständig wiedergegeben.

Fünf verschiedene Amplikons konnten unter den Klonen identifiziert werden (Tab. 16). Zwei davon sind hinsichtlich ihrer Sequenz inkonsistent mit der angenommenen Entstehung aus einem inversen Template (vgl. Abschnitt 2.2.6 ab S. 22): Eines mit einer Länge von 557 bp enthält nicht die zu

Tab. 16. Flankenpaare des Transposons thug-T als Ergebnis inverser PCR. Abkürzung: RITR – rechter ITR.

inverser Template	Produkt-länge [bp]	TSD	Sequenzanalyse
BclI	786	-	konsistent, aber verkürzter RITR der Transposonkopie
BclI	2097	AACT	konsistent; Insertion in TRE3-B; Locus identisch mit 3350-bp-Produkt auf NspI-Template
NspI	557	-	inkonsistent, keine Schnittstelle für NspI,
NspI	924	GTAAT	Ligationsstelle mit Einschub von mtDNA, nur links davon NspI-Schnittstelle; Flankenpaarung konsistent mit Contig JC2d04c02.r1
NspI	3349	AACT	konsistent; Insertion in TRE3-B; Locus identisch mit 2100-bp-Produkt auf BclI-Template

erwartende Schnittstelle für BclI, ein anderes mit einer Länge von 924 bp enthält einen artifiziellen Sequenzeinschub von mitochondrialer DNA – der chimäre Status wird durch Sequenzvergleich mit Schrotschusssequenzen bestätigt. Beide verfälschten Produkte sind jeweils nur in geringer Menge im Produktgemisch aufzufinden (Abb. 15). Obwohl das Amplikon mit 924 bp Länge einen artifiziellen Charakter hat, bilden die enthaltenen Elementflanken von thug-T ein topologisches Paar im Genom, wie die Sequenz eines Contigs zeigt, in dem beide Flanken konsistent zu einem gemeinsamen Transposonlocus assembliert wurden. Es lässt sich eine TSD-Länge von 5 bp ableiten.

Ein Amplikon mit 786 bp Länge repräsentiert einen Locus des thug-T, dessen rechtes Ende abgeschnitten ist. Es kann demnach keine TSD abgeleitet werden. Zwei weitere Amplikons stammen vom gleichen thug-T-Locus (2097 bp von BclI-Template, 3349 bp von NspI-Template). Die abzuleitende TSD ist 4 bp lang.

3.3.1.2 Suche nach Flankenpaaren des Transposons DDT-S

Mehrere Restriktionsendonukleasen der Auswahl in Tab. 12 auf S. 37 haben keine Schnittstelle in der Consensussequenz des Transposons DDT-S: BclI, BfaI, MfeI, NspI, TatI. Es wurde zunächst nur auf dem inversen Template des Enzyms BclI mit dem Primerpaar DDT-S.21R / DDT-S.738 eine PCR durchgeführt. Aus der Primerbenennung geht die Position der Primer in der 758 bp langen Consensussequenz des DDT-S hervor. Es ist vorab zu bemerken, dass der Primer DDT-S.21R unvorteilhafterweise auch an die Sequenz des selteneren Transposons DDT-A binden kann. Wie im Abschnitt 3.5.5 ab S. 56 näher ausgeführt wird, hat der rechte ITR des DDT-S sehr hohe Sequenzähnlichkeit mit dem linken ITR des Elements DDT-A. Andere Regionen des DDT-S kommen

als Primerbindestellen weniger in Frage, weil die Sequenz von DDT-S überwiegend sehr A/T-reich ist.

Aus der PCR-Reaktion wurden dominante Produkte von je 450 bp, 750 bp und 1000 bp erhalten (Ergebnisse nicht dargestellt). Nach Trennung im Agarosegel wurden Banden entsprechend dieser Produkte ausgeschnitten, die DNA isoliert, im Vektor pGEM-T Easy (Promega) kloniert und mit Primern M13-fwd und M13-rev sequenziert. Die Sequenzen der Einzelklone wurden nach hoher Sequenzübereinstimmung zu einer Consensussequenz assembliert. Die resultierenden vier Consensus-sequenzen sind in Anhang 6.2.2 ab S. 97 vollständig wiedergegeben.

Tab. 17. Flanken der Transposons DDT-A und DDT-S als Ergebnis inverser PCR. Aus der Sequenzanalyse geht für alle identifizierten PCR-Produkte hervor, dass sie nicht von einem invertierten Template stammen.

inverser Template	Produkt [bp]	Identität der Flanke	Sequenzanalyse
BclI	486	TRE5-A.1 4748-5186	Das Produkt enthält beiderseits die Primersequenz DDT-S.21R; keine BclI-Schnittstelle
BclI	741	Tdd-5 418-1116	Das Produkt enthält beiderseits die Primersequenz DDT-S.21R; keine BclI-Schnittstelle
BclI	1017	skipper 3839-4696	Das Produkt enthält beiderseits die Primersequenz DDT-S.21R; keine BclI-Schnittstelle
BclI	1092	Tdd-4 1862-2910	Das Produkt enthält beiderseits die Primersequenz DDT-S.21R; keine BclI-Schnittstelle

Die Amplikons enthalten nicht das eingesetzte Primerpaar, sondern jeweils nur paarweise denselben Primer DDT-S.21R. Die identifizierten Flanken enthalten auch keine Schnittstellen für BclI. Dementsprechend können die erhaltenen Produkte nicht aus DNA-Abschnitten hervorgehen, die durch Template-Inversion entstanden sind (vgl. Beschreibung zu Theorie und Methodik, Abschnitt 2.2.6 ab S. 22). Die Sequenzen der PCR-Produkte weisen auch keine Anzeichen von chimärem Charakter auf (Tab. 17). Vielmehr handelt es sich bei den amplifizierten DNA-Abschnitten um Sequenzen anderer Transposons, die im vollen Einklang mit deren Consensussequenz stehen. Dieser Befund deutet darauf hin, dass die amplifizierten Fragmente in genau dieser Topologie im Genom von *D. discoideum* existieren. Es sind offenbar Flanken von jeweils zwei sehr eng benachbarten Elementen der Spezies DDT-A oder DDT-S (s. Bemerkungen zur Primerspezifität). Demnach hätten die Produkte ebensogut durch ein konventionelles PCR-Experiment von genomischer DNA amplifiziert werden können.

3.3.2 Sequenzanalyse kompletter Transposonloci

Die komplette Sequenz einer Transposonkopie liegt vielfach vor, wenn die Sequenz *per se* sehr kurz ist (DDT-S, thug-Familie) oder wenn verkürzte Kopien transponiert worden sind, was regelmäßig bei den non-LTR-Retrotransposons vorkommt (TRE-Familie). In solchen Fällen besteht eine gute Chance, dass einzelne Schrotschusssequenzen oder Schrotschussklone die eine vollständige

Transposonkopie mitsamt der TSD abdecken (Tab. 18). Beispielsweise konnten auch für das LTR-Retrotransposon DGLT-A mehrere TSDs aus der Sequenzanalyse von solo-LTR-Kopien ermittelt werden. In einigen Fällen waren längere, vollständige Transposonkopien auf assemblierten Contigs des Genomprojekts vorhanden (Abschnitt 3.1.1 ab S. 31). Durch Analyse der Paarung von Schrotschusssequenzen, die vom gleichen Klon stammen, wurde die korrekte Assemblierung dieser Transposonkopien verifiziert.

Tab. 18. Identifizierte Target-Site-Duplikationen (TSDs) durch Auswertung von vollständig sequenzierten Transposonloci.

Transposon-spezies	TSD	Belege
DDT-S	CC	Klon JAX4a29b01
DDT-S	GT	Klon JAX4a168h12
DDT-S	TA	Klon JAX4b16a11
DGLT-A	TTTA	JC1a140f10.s1 JC1b82g12.s1 JAX4b63d12.r1 JC1a139h01.s1
DGLT-A	TTTGT	JC2a23h07.r1 JC2a37h12.r2
DGLT-A	TTTC	JC2d01d02.r1 JC2y01h09.s1 JC2a99g02.r1 JC2c38f04.r1 ...
DGLT-A	ACTT	Contig JC2a234e08.r1 / JC2e38h05.s1 (Chr2)
Tdd-5	GTTAA	Klon JC1b45h04, Contig V4a
TRE3-B	TCGTCTCG(GAAAAC)	Klon JAX4a16f12
TRE3-C	TCTTGGCATTTTGAT(A)	Klon JAX4b09c08
TRE3-D	GGTTCAATTCTTT(AAT)	Klon JAX4a48d09
TRE5-A	(AAAA)GTACGTTATTATCTA	YAC DY3850
TRE5-A	GTACGTTATTATCT	YAC DY3850
TRE5-C	GTCCATTTTTAAA	Klon JC2b186b11
TRE5-C	TATACCCTTTTATTTA	Klon JC2c38e08
thug-T	GTAAT	Contig JC2d04c02.r1

Für längere Individuen von TRE-Elementen konnten Paare von flankierenden Sequenzen auch unabhängig von einer gemeinsamen Klonierung identifiziert werden. Die TSDs dieser Elementgruppe sind so lang – im Bereich von 10 bis 16 bp –, dass jede TSD genügend Sequenzinformation enthält, um die korrespondierenden Transposonflanken aufgrund ihrer Sequenz eindeutig zuordnen zu können. So wurden mehrere Flankenpaare zusammengetragen, die jeweils die Analyse der TSD erlauben (Tab. 19). Bei den non-LTR-Retrotransposons der TRE-Familie erschwert der Umstand die TSD-Interpretation, dass die Elementgrenzen weder 3’ noch 5’ exakt definiert sind. Die Polyadenylierung kann verschiedene Längen haben, und das 5’-Ende kann an einer beliebigen Position der Consensussequenz abgebrochen sein. Dadurch bestehen für die Beurteilung der TSD-Länge Freiräume.

Tab. 19. Identifizierte Target-Site-Duplikationen (TSDs) durch Paarung langer, individueller TSD-Sequenzen im Falle der Transposons der TRE-Familie.

Transposon-spezies	TSD	Belege durch Schrotschusssequenzen
TRE3-A	(A)GTATCCGTTC	JAX4a83f02.r1 JAX4b21c08.s1
TRE3-A	(A)CTTGCTGCT	JC2a10c08.s1 JAX4b05d06.s1 JAX4a98g08.s1
TRE3-A	(AAA)GGTACC(C)	JAX4a78b05.s1 JAX4b18e01.r1
TRE3-B	CTCTCAAATATC	JAX4a86b06.s1 JAX4a69g04.r1
TRE3-B	TGAGAAAGGAA(A)	JAX4b21a02.s1 JAX4d01d09.s1
TRE3-C	GTGTACCGATCTCCC	JAX4a28h06.r1 JAX4b11a07.r1 JAX4b17f07.r1
TRE3-C	(A)TTTTAAAAGTTTTTT	JC2a103e07.s1 JAX4a170e04.s1 JAX4b09c12.s1 ...
TRE3-C	(A)TTGATGTTGATGATGT	JAX4a30d10.r1 JAX4a68c09.s1 JC2a20f11.r1
TRE3-C	(AA)TTCCTTTTTTCAAT	JAX4a18f07.s1 JAX4a14h01.s1
TRE3-D	(AA)TCAAGATATTGTT(ACA)	JAX4a222b10.r1 JAX4a212c10.s1
TRE5-B	(AAA)TTTATATTTT(GAT)	JC2a115a09.s1 JC2a75c12.r1 JC1c198e04.s1 JC2b142d04.r1
TRE5-B	(A)TACTTCCATAA	JC2a25f06.r1 JAX4a44b09.r1

3.3.3 Analyse von verschachtelten Transposonloci

Eine Methode, die für die meisten Transposons aus *D. discoideum* eine Analyse der TSD erlaubt, basiert auf der **Auswertung verschachtelter Transposonloci**, wie sie in den Sequenz-Clustern für einzelne Transposonspezies (s. Abschnitt 2.4 ab S. 23) zu beobachten sind. Verschachtelte Transposonloci verursachen in den Sequenz-Clustern Erscheinungen, wie sie in Abb. 20 gezeigt sind. Insertionen einer Transposonspezies B in eine Transposonspezies A offenbaren sich in den Sequenz-Clustern durch eine charakteristische Topologie von divergenten Abschnitten der Schrotschusssequenzen. Es finden sich zwei Subpopulationen von divergierenden Schrotschusssequenzen, deren divergente Sequenzabschnitte in entgegengesetzte Richtungen zeigen.

Es müssen zwei Voraussetzungen in der Transposonverteilung gegeben sein, damit die beschriebenen Phänomene in regelmäßiger Weise beobachtet werden können:

- Verschachtelte Transposonloci müssen überhaupt im Genom vorhanden sein. Daraus folgt, dass Transposons mit einer gewissen Häufigkeit im Genom vorkommen müssen, damit in rein statistischer Konsequenz auch eine gewisse Zahl von verschachtelten Transposonloci zu beobachten ist. Oder es muss alternativ einen Trend zum Verschachteln von Transposonkopien geben, der über das Maß der allgemeinen Wahrscheinlichkeit für das Zusammentreffen von zwei Transposoninsertionen am gleichen genomischen Locus hinausgeht. Letzteres ist im Genom von *D. discoideum* der Fall: Es besteht ein hoch signifikanter Trend zum Verschachteln von Transposonkopien (Abschnitt 3.6.2 ab S. 62).
- Die Abdeckung des Genoms mit Schrotschusssequenzen muss ausreichend groß sein, damit verschachtelte Transposonkopien unter den Sequenzen erkannt und analysiert werden können. In Anknüpfung an das entworfene Treffermodell (Abschnitt 2.5.1 ab S. 25) geht es um die Identifizierung der beiden Sequenzübergänge zwischen Transposonkopie A und Transposonkopie B –

also ein genomisches Feature der Länge 0. Die beiden zu identifizierenden Sequenzfeature sind über die Länge der Transposonkopie B voneinander getrennt, welche in der Regel die Länge der Schrotschusssequenzen und -klone übersteigt. Die Treffer auf die Sequenzfeature sind demnach statistisch unabhängig.

Abb. 20. Identifikation von Target-Site-Duplikationen (TSDs) durch Auswertung von verschachtelten Transposonloci wie sie in Sequenz-Clustern sichtbar werden. **A)** Mehrere Transposonkopien der Spezies A (gelb) befinden sich im Genom. Eine Kopie enthält eine inserierte Kopie einer Transposonspezies B (blau), mit den flankierenden Sequenzen der TSD (rot). **B.1)** Aus genomischen Schrotschusssequenzen, die Fragmente der Transposonspezies A enthalten, wird ein Sequenz-Cluster aufgebaut. Das Bild zeigt den schematisierten Blick auf einen Ausschnitt des Sequenz-Cluster für diejenige Region, in der die Insertion von Transposonspezies B stattgefunden hat. **B.2)** Wie B.1 ein Blick auf das Sequenz-Cluster, hier in der Repräsentationsform eines GAP4-Projekts (vgl. Abschnitt 2.4.2 ab S. 25). Dieses explizite Beispiel zeigt die Insertion einer DDT-S-Kopie in eine TRE5-B-Kopie. Die TSD-Sequenz lautet auf AT.

Da die beiden Voraussetzungen gegeben sind, besteht ein Ansatzpunkt, in systematischer Weise TSDs aus der Beobachtung von verschachtelten Transposonkopien zu analysieren. Schließlich belegt die Ausbeute der mit dieser Methode gesammelten TSDs, dass die aufgezählten Voraussetzungen erfüllt sind. Tab. 21 listet die aufgefundenen verschachtelten Transposonloci und die davon abgeleiteten TSD-Sequenzen der nachfolgend inserierten Transposonspezies auf.

Tab. 21. Identifizierte Target-Site-Duplikationen (TSDs) durch Auswertung von verschachtelten Transposonloci. Die Tabelle dokumentiert Fälle der Insertion einer Transposonkopie der Spezies A in eine Transposonkopie der Spezies B. Eine Vielzahl von Insertionen der Spezies DDT-S und DIRS-1 wurde in dieser Zusammenstellung vernachlässigt.

Transposon-spezies A	Transposon-spezies B	TSD	Belege durch Schrotschusssequenzen
DCLT-A	DDT-S	AATT	JC2b287d12.r1 JAX4a99f08.s1 JC2b230b11.r1 JC3a45c08.r1 ...
DDT-A	TRE3-A	AA	JAX4a34d11.r1 JAX4c01d03.r1
DDT-B	DDT-B	AT	JC2c176c11.r1 JAX4d03b12.s1 JC2a162g04.s1 ...
DDT-S	Tdd-5	TA	JC2d21e02.r1 JC2b325c09.r1 JC2b395c11.s2 JC2d104c05.r1 ...
DDT-S	Tdd-5	TA	JAX4a88h03.r1 JC2b364h10.s1 JC2a87c04.r1 JC2b134g11.s1 ...
DDT-S	Tdd-5	TA	JAX4a87b05.s1 JAX4a56c10.r1 JAX4a97d10.s1
DDT-S	TRE5-B	AT	JC2d23b07.s1 JC2a20c07.s1 JAX4b40f06.r1
DIRS-1	DDT-B	-	JAX4a103a10.s1 JC2a130b11.r1 JC2a118b09.s1 JC2a112a02.s1 ...
DIRS-1	DDT-B	-	JAX4d06a05.r1 JC2e63f11.r1 JC2a16h12.s1 JAX4a154c01.r1 ...
DIRS-1	DDT-B	-	JC2d72a02.s1 JAX4a15f08.r1 JAX4a144h09.r1
skipper	TRE5-B	TTGTT	JAX4a72c10.r1 JAX4b60b05.s1
Tdd-5	TRE3-A	ATTAG	JAX4a34d11.r1
thug-T	TRE3-B	AACT	JAX4a239e06.r1 JAX4a21d03.s2
TRE3-A	TRE3-C	(A)GAAAAA CATA	gb|K02643 gb|AH001350
TRE3-C	TRE3-D	GTATTCAA TGAAG(A)	JC1a212c08.r1 JAX4a229e09.s1 JC1c87e06.r1
TRE5-C	DGLT-A	TTCCTATTA AAT(A)	JC2b386h05.r1 JC2a23h07.s1 JC1b57f10.r1

Aus verschachtelten Transposonloci konnten demnach systematisch TSDs für die meisten Transposonspezies von *D. discoideum* ermittelt werden: DIRS-1, skipper, Tdd-4, -5, DDT-A, -B, -S, thug-S, -T. Unermittelt bleiben meist die TSDs der streng insertionsortspezifischen Transposonspezies, die äußerst selten in andere Transposonkopien inserieren – wenn doch, dann nur in Kopien von Transposons mit derselben Insertionsortspezifität. Für diese Transposongruppe konnten nur wenige verschachtelte Transposontopologien aufgefunden werden: Ein Fall eines TRE3-C in TRE3-D, einer von TRE5-C in DGLT-A. Glücklicherweise besitzen die streng insertionsortspezifischen Transposons der TRE-Familie sehr lange TSDs, so dass die paarweise Zuordnung von Flanken durch deren Sequenz möglich ist (Abschnitt 3.3.2).

3.4 Häufigkeit der Transposons im Genom von *D. discoideum*

3.4.1 Genomischer Anteil von Transposons

In Hinblick auf die angestrebte Assemblierung des *Dictyostelium*-Genoms aus Schrotschusssequenzen ist es wichtig, den Anteil einzelner Transposonspezies an der Gesamtsequenz des Genoms zu kennen. Die Zahl von Transposonindividuen einzelner Spezies ist eine wesentliche Größe bei der

Beurteilung der Komplikationen, die bei der Assemblierung zu erwarten sind (vgl. Abschnitt 3.7 ab S. 63). Im Abschnitt 2.5.2 ab S. 27 sind die Methoden zur Schätzung von Nukleotidanteil und Mindestkopienzahl eines Transposons im Genom beschrieben. Die Schätzergebnisse für alle identifizierten Transposonspezies sind in Tab. 43 auf S. 74 zusammengefasst.

Wichtig für die Interpretation der Schätzergebnisse ist eine Kenntnis der Unschärfe der Schätzung. Als **Kontrolle der Schätzmethode** habe ich die Aktingenfamilie mit ihren zahlreichen Genkopien herangezogen. Die Kopienzahl dieser Genfamilie wurde in zurückliegenden Untersuchungen aufgrund von Blot-Analysen auf 17 bis 20 geschätzt, und es wurden 15 dieser Kopien durch Sequenzierung charakterisiert (Rückblick in ROMANS & FIRTEL 1985). Die Ähnlichkeit der Aktingenkopien auf Nukleotidebene beträgt mindestens 90 %. Die Verhältnisse lassen sich also gut mit denen von Transposonkopien einer Spezies vergleichen.

Aufgrund der Trefferstatistik von Schrotschusssequenzen wird für die Aktingene eine Kopienzahl von 39 geschätzt – also die doppelte Zahl im Vergleich zum Literaturwert. Um die Ursache für den gravierenden Unterschied in den Kopienzahlschätzungen zu untersuchen, wurden genomische Contigs assembliert, die jeweils individuelle Aktingenkopien repräsentieren. Für diese Assemblierung wurde auf alle Ressourcen von Schrotschusssequenzen zurückgegriffen, d.h. auch die Sequenzen der chromosomenspezifischen Banken, um eine maximale Ausbeute der vorhandenen Aktingenkopien zu erreichen. Schließlich konnte auf diese Weise die Existenz von mindestens 23 vollständigen Genkopien nachgewiesen werden. Daneben fanden sich zwei kurze genomische Fragmente mit einer Sequenzähnlichkeit zum Aktin. Es handelt sich dabei sicherlich um Pseudogene. Es sind also mindestens drei Aktingenkopien mehr im Genom von *D. discoideum* vorhanden als ursprünglich geschätzt. Es bleibt dennoch eine gravierende Diskrepanz zwischen der Kopienzahlschätzung von 39 und der anzunehmenden tatsächlichen Kopienzahl von 23, maximal 24 (unter Berücksichtigung der fragmentarischen Kopien als 2 × ½). Zur Klärung dieses Missverhältnisses wurde eine alternative Schätzmethode angewandt, die auf der Auswertung von treffenden Sequenzen beruht, nicht auf der Zählung der getroffenen Nukleotide (Abschnitt 2.5.1 ab S. 25). Die Länge der Aktingenkopien ist invariabel – von den zwei kürzeren Fragmenten abgesehen –, daher greift diese Methode in dieser Situation. Man erhält auf diesem Wege einen deutlich geringeren Schätzwert:

$$P(H \mid n_F=1) = [\, L_{Sh} - L_O + \max(L_F - L_O, 0) \,] / L_G$$

mit $L_{Sh} = 379$, $L_O = 50$, $L_F = 1131$, $L_G = 3{,}4 \cdot 10^7$ (vgl. Abschnitt 3.1.2 ab S. 32)

$$= (379 - 2 \times 50 + 1131) / 3{,}4 \cdot 10^7 = 4{,}1 \cdot 10^{-5}$$

$$ML(n_F) = n_H / [\, n_{Sh} \times P(H \mid n_F=1) \,]$$

mit $n_{Sh} = 45150$ und gefundenen Schrotschusstreffern $n_H = 61$

$$= 61 / (\, 45150 \times 4{,}1 \cdot 10^{-5} \,) = 33$$

Dieser Schätzwert von 33 Genkopien steht bereits in einem besseren Einklang mit der anzunehmenden Zahl. Ist die Abweichung signifikant? Angenommen, im Genom befinden sich tatsächlich nicht mehr als die nachgewiesenen 24 Kopien – und eine sehr viel höhere Zahl ist aufgrund der Datenlage der Contigsequenzen und aufgrund der Ergebnisse vieler veröffentlichter wissen-

schaftlicher Arbeiten auszuschließen –, dann berechnet sich die Wahrscheinlichkeit aufgrund einer binomial verteilten Kombination der Ereignisse „Treffer“ / „kein Treffer“ gemäß Abschnitt 2.5.2 ab S. 27. Ich korrigiere für die paarweise Kopplung von Schrotschusssequenzen mit einem Freiheitsgrad von 0,58 und arbeite demzufolge mit der Ziehung 35 aus 26480 bei $p = 24 \times 4{,}1 \cdot 10^{-5}$. Das Integral der Binomialverteilung für $x \geq 35$ beträgt 0,054 (Tab. in Anhang 6.1.2). Für die Fragestellung nach beidseitigem Verlassen des zentralen Wahrscheinlichkeitsintervalls wird ein Integral von 0,108 erhalten. Die Abweichung vom Erwartungswert ist demnach nur schwach signifikant.

Die allerdings hochsignifikante Diskrepanz zwischen der Schätzung aufgrund von Nukleotidumfang von treffenden Schrotschusssequenzen und zu der anzunehmenden Kopienzahl ist ein starkes Indiz für einen systematischen Fehler bei der Schätzung von Kopienzahlen. Die Erfahrungen bezüglich der Klonierbarkeit und Sequenzierbarkeit deuten auf die mögliche Ursache hin, dass G/C-reichere Sequenzen, beispielsweise CDS mit 31 % G/C gegenüber nicht codierender Sequenz mit 12 % G/C, in den Schrotschusssequenzen deutlich überrepräsentiert sind. Aufgrund der Zweifel an der Haltbarkeit des statistischen Modells gegenüber der Ziehung von Schrotschusssequenzen aus einem Genom wurde davon abgesehen, auf Basis des statistischen Modells Vertrauensintervalle für die Schätzungen von Sequenzanteil und Kopienzahlen zu berechnen. Es muss deutlich hervorgehoben werden: Die Unschärfe der vorgestellten Schätzungen geht weniger auf die Begrenztheit der herangezogenen Stichproben als vielmehr auf die eingeschränkte Anwendbarkeit des statistischen Modells zurück. Die notwendige Einschränkung geht auf den spürbaren Einfluss von Klonierungs- und Sequenzierungs-Bias zurück. Leider ist der Einfluss dieser Größen kaum quantifizierbar.

Alle Schätzergebnisse für die Transposonhäufigkeiten sind gemeinsam mit anderen Charakteristika der Transposonspezies in Tab. 43 auf S. 74 zusammengefasst.

3.4.2 Fragmentierungsindex und Fragmentanzahl der Transposonsspezies

Unter Fragmentierung verstehe ich hier das Phänomen, dass Transposonanteile im Genom nicht durchweg in vollständiger Länge ihrer Consensussequenz auftauchen. Vielmehr findet man mehr oder weniger häufig Fragmente, die nur Teile der Consensussequenz repräsentieren, z.B. nur die linke Hälfte oder ein Mittelstück der Gesamtsequenz. Falls für eine Transposonspezies nur vollständige Insertionen im Genom existieren, ermittelt sich die Anzahl der Transposonloci dieser Spezies im Genom aus dem genomischen Anteil der Transposonspezies (s. vorigen Abschnitt) und der Länge der Consensussequenz. Man erhält als Wert die „minimale Fragmentanzahl“ – kleinere Werte können nur im Rahmen der Schätzungenauigkeit vorkommen. Falls die Transposonindividuen aber fragmentiert sind, ist die Zahl der Transposonloci größer. Als Maß führe ich hier das Verhältnis der Anzahl von tatsächlichen Transposonfragmenten im Genom zur Anzahl der Transposonindividuen, die sich ergeben würden, wenn alle Transposonloci unfragmentierte und vollständige Sequenzen enthalten. Ich nenne das Maß „Fragmentierungsindex“. Es hat einen Wertebereich von ≥ 1.

Die Fragmentierung der Loci einer Transposonspezies lässt sich belegen und ermitteln durch Schrotschusssequenzen, welche die Sequenz der Transposonspezies tragen, die Sequenz aber an einer Consensusposition, die nicht dem Ende der intakten Transposonspezies entspricht, in nicht-

Transposonsequenz übergeht. Die statistische Auswertung solcher Funde führt zu einem Maximum-Likelihood-Schätzwert für Bruchereignisse unter den Transposonkopien (Abschnitt 2.5.1 ab S. 25).

Ein Berechnungsansatz bezieht auch die aus der Consensussequenz zu erwartenden Transposonenden ein, und aus dem Verhältnis von extrapolierter Endenanzahl und minimaler Fragmentanzahl ergibt sich der Fragmentierungsindex. Dieser Berechnungsweg berücksichtigt auch den Einfluss interner Deletionen quantitativ. Unter den Ergebnissen des Fragmentierungsindex für die Transposonspezies befinden sich viele Werte von kleiner 1 (nicht dargestellt), die aus theoretischer Sicht keinen Sinn ergeben. Ich habe festgestellt, dass diese irritierenden Ergebnisse darauf zurückgehen, dass die natürlichen Enden der betroffenen Transposonspezies in den Schrotschusssequenzen deutlich unterrepräsentiert sind. Die Sequenz-Cluster sind in diesen Fällen an den Enden der Consensussequenz schwächer durch Schrotschusssequenzen abgedeckt als im Mittelbereich, wobei das Absinken der lokalen Coverage im Alignment nicht mit internen Bruchpunkten der Transposonsequenz einhergeht – es gibt keine Anzeichen von Fragmentierung. Als Beispiel für das beschriebene Phänomen ist in Abb. 22 die lokale Verteilung der Coverage über das Sequenz-Cluster von Tdd-4 dargestellt.

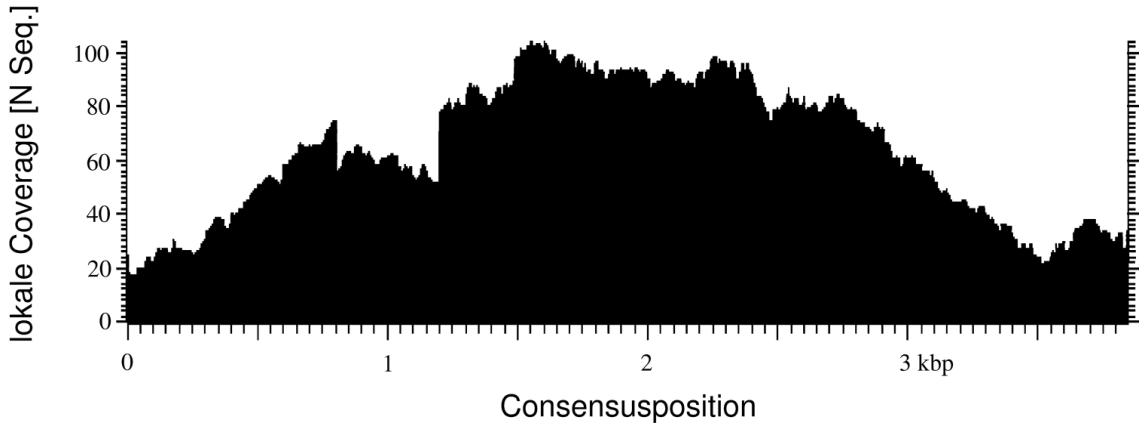


Abb. 22. Lokale Abdeckung der Consensussequenz des Transposons Tdd-4 durch Schrotschusssequenzen.

In den beschriebenen Fällen, die Hinweise auf einen Klonierungs- und/oder Sequenzierungs-Bias für die Endabschnitte der Transposonsequenzen zeigen, wurden schließlich nur die beobachteten internen Sequenzbruchpunkte für die Schätzberechnungen des Fragmentierungsindex herangezogen. Die minimale Fragmentanzahl wird zunächst als Schätzung für die tatsächliche Fragmentanzahl zugrunde gelegt. Anschließend werden die beobachteten Hinweise auf Fälle interner Fragmentierung hinzugerechnet. Alle Berechnungsergebnisse sind gemeinsam mit anderen Charakteristika der Transposonspezies in Tab. 43 auf S. 74 zusammengefasst.

3.5 Sequenzanalyse der neu aufgefundenen Transposons aus *D. discoideum*

In den vorigen Abschnitten wurden die Transposonspezies immer unter allgemeingültigen Gesichtspunkten betrachtet: Sequenzermittlung, Kopienzahl, Fragmentierung und TSD-Analyse. In den folgenden Abschnitten soll hingegen auf die charakteristischen biologischen Eigenschaften der neu entdeckten Transposonspezies näher eingegangen werden.

3.5.1 LTR-Retrotransposons

Es wurde bereits früher vermutet, dass der **HindIII-Repeat (H3R)**, ein 268 bp langes Element, das häufig in der Nachbarschaft von tRNA-Genen beobachtet wird, den solo-LTR eines LTR-Retrotransposons darstellt (WINCKLER 1998). Ich habe ein autonomes LTR-Retrotransposons identifiziert, dem der H3R als LTR zuzuschreiben ist (Abschnitt 3.2.2). Die Benennung dieses neu entdeckten Elements als ***„Dictyostelium* Ty3/Gypsy-Like Transposon“ (DGLT-A)** ergibt sich aus der Sequenzähnlichkeit der codierenden Sequenz zu den Ty3/Gypsy-ähnlichen LTR-Retrotransposons (Einzelheiten später). DGLT-A kommt in zwei Subformen vor, DGLT-A.1 und DGLT-A.2, die durch eine Reihe von charakteristischen Basensubstitutionen, Insertionen und Deletionen voneinander verschieden sind. Die Kopienzahl ist für beide Formen sehr niedrig, und nur für die Subform DGLT-A.1 konnte die vollständige Consensussequenz ermittelt werden.

Die 5 kbp lange Consensussequenz des DGLT-A.1 besitzt einem einzigen, langen ORF (Abb. 23). Ein Sequenzvergleich der korrespondierenden Peptidsequenz mit den nächstähnlichen annotierten Transposons, Gypsy-ähnlichen LTR-Retrotransposons aus *Oryza sativa* (Acc.No. AB014741) und aus *Pisum sativum* (Acc.No. AF083074), zeigt, dass alle wesentlichen funktionellen Elemente in

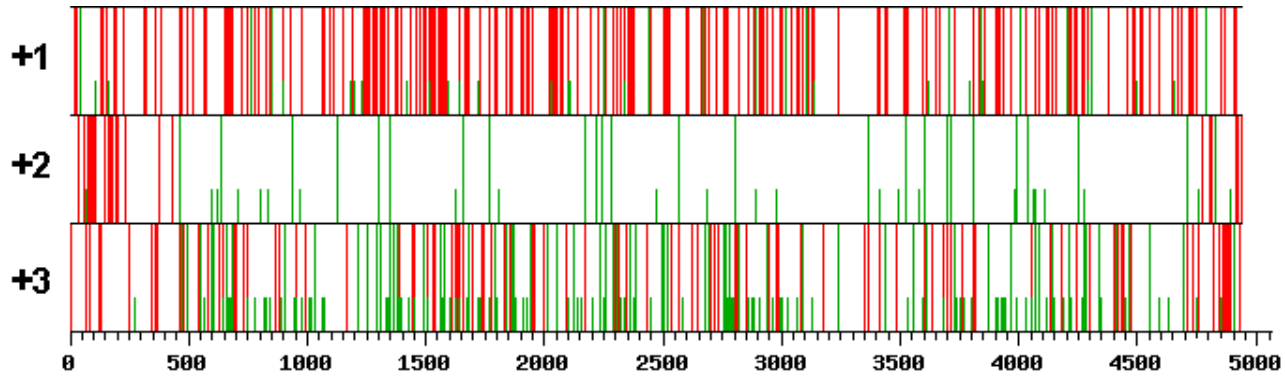


Abb. 23. Auftragung von Start- (grüne Balken) und Stopcodons (rote Balken) über die drei Leseraster im Sinnstrang der Consensussequenz von DGLT-A.1. Verkürzte grüne Balken markieren nicht-kanonische Startcodons. Auf die Darstellung der negativ orientierten Leseraster wird verzichtet, da sie keine längeren Leserahmen enthalten.

dieser Peptidsequenz enthalten sind: Ein Zinkfinger-ähnliches Motiv der Formel $CX_2CX_4HX_4C$ im Sequenzabschnitt 243-256, ein Aspartyl-Protease-Motiv der Sequenz LVDTGS im Abschnitt 310-315, eine Reverse-Transkriptase-Domäne etwa im Abschnitt 500-680 mit außerordentlicher hoher Aminosäureähnlichkeit zwischen den verglichenen Proteinen, eine RNAse-H-Domäne im Abschnitt

773-786 und schließlich eine Integrase-Domäne im Bereich 1076-1434 einschließlich einem aminoterminal gelegenen HHCC-Motiv im Bereich 1089-1130. Dem DGLT-A-Element fehlt ein Gag-ähnlicher Proteinabschnitt. Ein solcher ist üblicherweise 5' in einem separaten ORF codiert (LENG ET AL. 1998). Dadurch ist die Transposonsequenz ungewöhnlich kurz. Von etwa 50 näher überprüften Ty3/Gypsy-ähnlichen LTR-Retrotransposons war keines kürzer als 6000 bp. Da, wie dargelegt, alle funktionell relevanten Peptidabschnitte identifiziert werden konnten, handelt es sich offenbar um ein autonomes Transposon voller Länge.

Ein für LTR-Retrotransposons typischer Polypurintrakt befindet sich beim DGLT-A 2 bp proximal vom rechten LTR mit der Sequenz AAA AGG GGG GGA. Insbesondere die Aufeinanderfolge von sieben Guanidinnukleotiden ist für das A/T-reiche Genom von *D. discoideum* eine hochsignifikante Erscheinung. Proximal benachbart zum linken LTR ist typischerweise die „Primer Binding Site" (PBS, Abb. 2 auf S. 9) lokalisiert. Die Sequenz TTT GGC GAC ATC GTC TTT CAA AAA hat jedoch keine Ähnlichkeit mit dem 3'-Ende einer tRNA, dem häufig beobachteten Primermolekül für die Initialisierung der reversen Transkription des Minusstrangs. Die Teilsequenz TGG CGA CAT CGT C konnte zwar in einem assemblierten Contig unseres fortschreitenden Genomsequenzierungsprojekts aufgefunden werden, der entsprechende Sequenzabschnitt ist aber im zentralen Teil eines PolII-transkribierten Gens, einem Homologen des „nuclear receptor binding factor-1" der Ratte (Acc.No. AB015724), lokalisiert. Schließlich muss die Frage nach dem Primermolekül offen gelassen werden. Auch beim LTR-Retrotransposon skipper konnte der natürliche Primer für die reverse Transkription des Minusstrangs bisher nicht identifiziert werden (LENG ET AL. 1998).

DGLT-A hat eine ausgeprägte Zielpräferenz für das 5'-Ende von tRNAs – wie auch die non-LTR-Retrotransposons der TRE5-Subfamilie. Diese Eigenschaft hat zur Entdeckung des Elements geführt. Allerdings inseriert das DGLT-A mit umgekehrter Orientierung. Die acht inspizierten Kopien sind alle in paralleler Leserichtung zur tRNA ausgerichtet.

3.5.2 non-LTR-Retrotransposons: Die TRE-Familie

In *D. discoideum* wird die Klasse der non-LTR-Retrotransposons durch eine einzige, mitgliederreiche, aber sehr homogene Transposonfamilie vertreten. In Abschnitt 3.2 wurde ein einheitliches Benennungsschema für diese Transposons eingeführt. Der Grundbauplan der Elemente ist recht übersichtlich und typisch für die gesamte Subklasse der non-LTR-Retrotransposons (Abb. 27A). Wiederholte Sequenzen an den beiden Enden kommen nicht vor. Eine ausgedehnte zentrale intronlose codierende Region enthält zwei überlappende – oder im Falle des TRE5-B zumindest eng benachbarte – ORFs.

Obwohl die Überlappungsregion zwischen ORF1 und ORF2 wesentliche Bedeutung für das Gleichgewicht der Translationsprodukte hat (Abschnitt 1.2.1 ab S. 6), zeigt sie unter den verschiedenen TRE-Elementen eine hohe Formenvielfalt (Abb. 24):

- Es werden in keinem Fall Stem-Loop-Strukturen nahe dem ORF-Überlapp gefunden (JACKS ET AL. 1988, LENG ET AL. 1998). Derartige Strukturen verursachen eine gelegentliche Unterdrückung des Stopsignals. Alle ORF-Übergänge sind in Form von Leserasterverschiebungen oder – nur im Falle

von TRE5-B – durch räumliche Trennung der ORFs realisiert.

- TRE3-A besitzt einen „Single-Nucleotide Polymorphism“ (SNP) für das Startcodon von ORF2. Alternativ zur DNA-Sequenz TAATG (Stop, gefolgt von Start im -1-benachbarten Leseraster) findet sich TAATA mit einer Allelfrequenz von 30 %. Entweder wird hier das ungewöhnliche Startcodon AUA benutzt (s. auch nächster Punkt) oder dieser Polymorphismus führt zu Transkripten, von denen ORF2 nicht translatiert wird. Im letzteren Fall stellt sich allerdings die Frage, wie ein solches Transkript den Replikationsmechanismus durchlaufen kann. Die reverse Transkription könnte nur durch das Translationsprodukt eines anderen Transkripts erfolgen.
- TRE5-A.1 besitzt das Codon AUA im Leseraster -2 relativ zum Stopcodon, das als einzig möglicher Reinitialisierungspunkt in Frage kommt. Das nächste mögliche kanonische Startcodon AUG ist 300 bp stromaufwärts entfernt.
- Eine deutliche Sonderstellung nimmt das Transposon TRE5-B ein. Dessen ORFs sind zwischen Stop- und Startcodon durch einen über 100 bp langen Polyadenintrakt getrennt.

TRE3-A

```
          CTATACGACTCCCAAGAGGAGATCGATATCATCTCATAATGGTAGTAATCAAAATCAACCAATGGAATGT
          CTATACGACTCCCAAGAGGAGATCGATATCATCTCATAATAGTAGTAATCAAAATCAACCAATGGAATGT
Frame +1    I  R  L  P  R  G  D  R  Y  H  L  I  M  V  V  I  K  I  N  Q  W  N
Frame +2  L  Y  D  S  Q  E  E  I  D  I  I  S  *  W  *  *  S  K  S  T  N  G  M
Frame +3   Y  T  T  P  K  R  R  S  I  S  S  H  N  G  S  N  Q  N  Q  P  M  E  C
```

TRE3-B

```
          ACCAGTCAATACTGGTGCTGCCAAACCTAGTAAGGTATGAGCTCAAAAATCAACTTTATAACATGGAATT
Frame +1   P  V  N  T  G  A  A  K  P  S  K  V  *  A  Q  K  S  T  L  *  H  G  I
Frame +2    Q  S  I  L  V  L  P  N  L  V  R  Y  E  L  K  N  Q  L  Y  N  M  E
Frame +3  T  S  Q  Y  W  C  C  Q  T  *  *  G  M  S  S  K  I  N  F  I  T  W  N
```

TRE3-C

```
          TGGTACCAAGTTTCCAATTGCCAGTAGTATTAGAAAATAATGGAACAATTAAAATTATTACTATGGAACT
Frame +1   G  T  K  F  P  I  A  S  S  I  R  K  *  W  N  N  *  N  Y  Y  Y  G  T
Frame +2    V  P  S  F  Q  L  P  V  V  L  E  N  N  G  T  I  K  I  I  T  M  E
Frame +3  W  Y  Q  V  S  N  C  Q  *  Y  *  K  I  M  E  Q  L  K  L  L  L  W  N
```

TRE5-A.1

```
          TCAAAAGGTAGATAAAAATCTACCCCAAATAAAACTATAAAGAAAAACACAATAAGAATAGGTGTTTGGA
Frame +1   Q  K  V  D  K  N  L  P  Q  I  K  L  *  R  K  T  Q  *  E  *  V  F  G
Frame +2    K  R  *  I  K  I  Y  P  K  *  N  Y  K  E  K  H  N  K  N  R  C  L
Frame +3  S  K  G  R  *  K  S  T  P  N  K  T  I  K  K  N  T  I  R  I  G  V  W
```

TRE5-B

```
          ATACTTCCATCTGTCTCATTTGTTTAAAAAAAAAA| 90 bp |AAAAAAAAATGATAAAAATAATGACC
Frame +1  I  L  P  S  V  S  F  V  *  K  K  K | 30 aa | K  K  N  D  K  N  N  D  L
Frame +2   Y  F  H  L  S  H  L  F  K  K  K  K| 30 aa |  K  K  M  I  K  I  M  T
Frame +3    T  S  I  C  L  I  C  L  K  K  K  | 30 aa |K  K  K  *  *  K  *  *  P
```

Abb. 24. ORF-Übergänge bei den TRE-Elementen. In Rot hervorgehoben sind jeweils die Peptidsequenzen von ORF1 und ORF2. In Blau hervorgehoben beim Element TRE3-A ein Polymorphismus, der das Startcodon von ORF2 betrifft.

Von den zwei abzuleitenden Proteinprodukten ist letzteres deutlich dem für Retroelemente – Retrotransposons und Retroviren – typischen Polyprotein homolog. In der Auftragung des Konservierungsgrads der alignierten Polyproteine aller vollständig bekannten TREs zeichnen sich drei funktionelle Domänen ab (Abb. 25): Eine Endonuklease-Domäne vom Typ der apurinisch/apyrimidinischen Endonukleasen, die Domäne der reversen Transkriptase (RT) und ein CCHC-Motiv, das vermutlich an der DNA-Bindung beteiligt ist. Diese Zuweisung wird durch Sequenzähnlichkeit zu verschiedenen Retroelementen unterstützt (BLASTP).

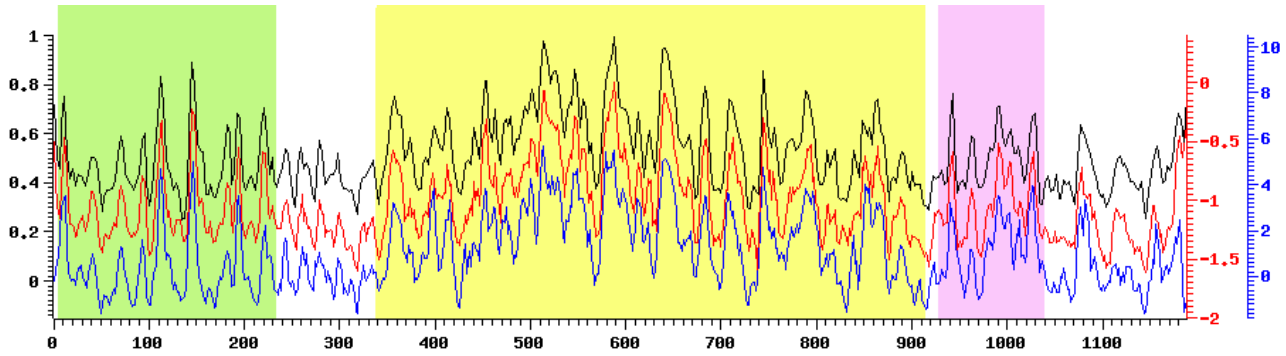


Abb. 25. Lokaler Konservierungsgrad im Alignment der Polyproteine (ORF2) aller vollständig bekannten TRE-Elemente. Methoden zur Bemessung des lokalen Konservierungsgrads (Abschnitt 2.6.3 ab S. 29): **schwarz** – Methode der relativen Identität, **rot** – Entropie-Methode, **blau** – Matrix-Methode. Die Lage der funktionellen Domänen ist durch farbliche Unterlegung gekennzeichnet: **grün** – AP-Endonuklease, **gelb –** reverse Transkriptase, **violett –** CCHC-Motiv mit aminoterminaler Nachbarschaft.

Das Polyprotein und speziell die RT-Domäne eignen sich als phylogenetische Indikatoren (EICKBUSH 1992). Da sich die Polyproteine der TREs in ihrer Gesamtlänge gut alignieren lassen, wurde dieses Alignment unter Ausschluss zweifelhaft alignierter und Gap-reicher Bereiche (ca. 15 % im carboxyterminalen Abschnitt) zum Berechnen eines phylogenetischen Baums herangezogen (Abb. 26). Da zwei der TRE-Elemente und deren Polyproteine unvollständig bekannt sind (TRE3-D und TRE5-C), wurde in einem ersten Schritt die Gesamttopologie unter Ausschluss der unvollständigen Elemente berechnet. In einem zweiten Schritt wurden Alignmentsegmente aller Elemente entsprechend der Länge der unvollständigen Elemente herangezogen, um die Verzweigungspunkte der verbleibenden Elemente zu ermitteln und in den Baum einzufügen.

Bemerkenswert ist, dass sich in der Reihe von TRE3-Elementen über TRE5-B und -C hin zum Element TRE5-A eine ausgeprägte Modulstruktur entwickelt hat (Abb. 27B, vgl. auch phylogenetische Analyse in Abb. 26). Es wurde bereits durch zurückliegende Untersuchungen gezeigt, dass das Transposon TRE5-A am 5'-Ende eine charakteristische Tandemwiederholung einer 202 bp langen Sequenz, das sogenannte A-Modul, besitzt (MARSCHALEK ET AL. 1992a). Auch das neu aufgefundene Element TRE5-B zeigt diese modulare Struktur des 5'-UTR mit einer 84-%-igen Konservierung seiner 198 bp langen Modulsequenz im Vergleich zum TRE5-A. Einzelne Transposonkopien von TRE5-B weisen Sequenzabweichungen in der Modulstruktur auf, und die Abweichungen sind dabei in jeweils allen Wiederholungseinheiten strikt gleich. Zwei unabhängige Fälle mit einer Consensus-Übereinstimmung von 96,5 % und 99,5 % und jeweils zwei bis drei exakten

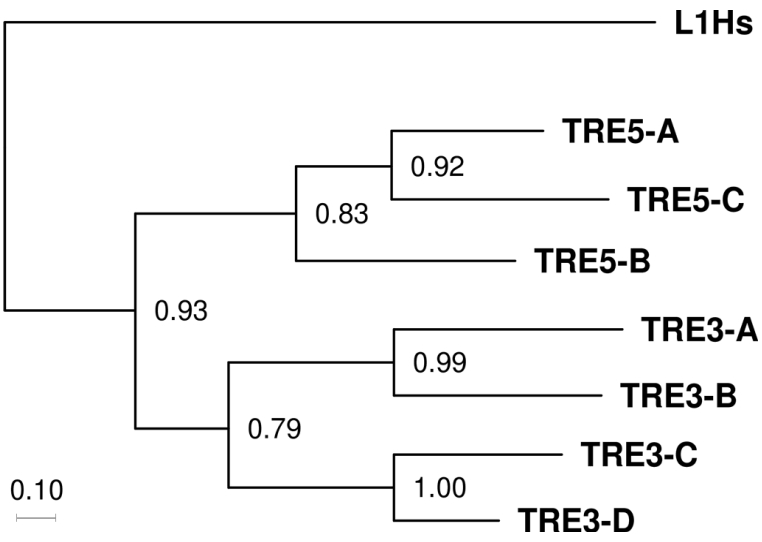


Abb. 26. Phylogenetischer Baum auf Basis des Polyproteins (ORF2) aller TRE-Elemente. Als Wurzel wurde ORF2 vom LINE1-Element des Menschen herangezogen („L1Hs", Acc.No. AAB59368). Die Astlängen im Baum korrelieren mit der Jukes-Cantor-Distanz (Maßstab unten links). Die Werte an den Knoten geben die relative Bootstrap-Unterstützung aus 1000 Ziehungen wieder.

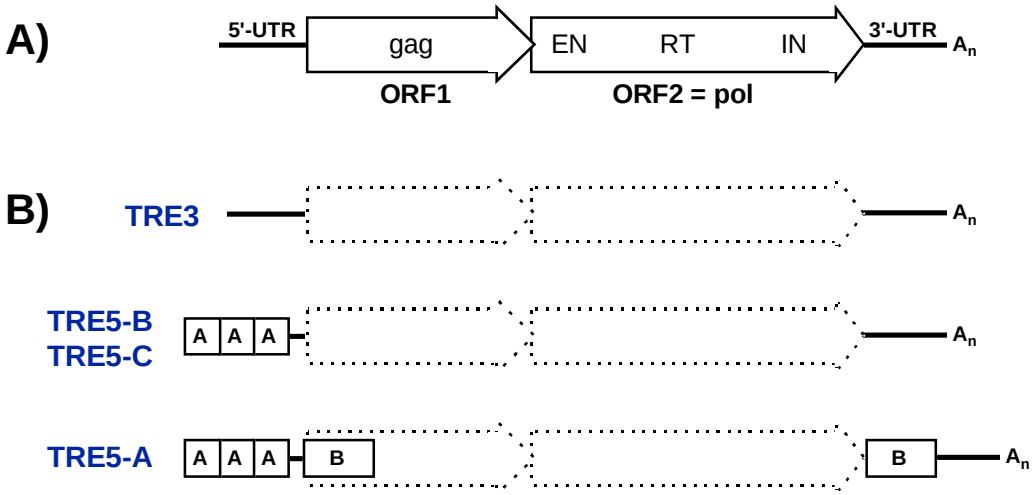


Abb. 27. Struktur der TRE-Elemente. **A)** Grundbauplan eines TRE-Elements, der ganz dem üblichen Aufbau eines non-LTR-Transposons folgt **B)** Zugewinn modularer Strukturelemente in der Reihe der Elemente TRE3 → TRE5-B/-C → TRE5-A.

Wiederholungen konnten identifiziert werden.

3.5.3 Unklassifizierte, unautonome Transposons: Die thug-Familie

Die neu aufgefundenen Transposonspezies thug-S und thug-T lassen sich aufgrund ihrer ähnlichen ITR-Sequenz zu einer gemeinsamen Familie gruppieren. Da viele Transposonkopien der Familie mit Rearrangements der Zielsequenz assoziiert sind, habe ich die Familie „thug" (engl. für „Schläger") genannt. Die spezifizierenden Buchstaben „S" (für engl. „shortened") und „T" sollen deutlich machen,

dass es sich nicht um autonome Transposonspezies handelt.

Die Elemente enthalten kaum codierendes Potenzial; beim Element thug-S ist im Minusstrang der 2192 bp langen Consensussequenz ein ORF vorhanden, der einer Proteinsequenz von 178 aa Länge entspricht. Es könnte sich um ein Genfragment handeln, jedoch lässt sich an der translatierten Sequenz mit BLASTP oder TBLASTN keine Sequenzähnlichkeit zu irgendeinem beschriebenen Polypeptid erkennen. Die ITRs sind das wesentliche Kriterium, die thug-Elemente einer gemeinsamen Familie zuzuordner. Sie haben eine Länge von 18 bp (thug-S) und 8 bp (thug-T) lauten distal auf die konservierte Sequenz TGT...ACA. Die Sequenz stellt eine Verbindung zu den LTR-Retrotransposons her. Gleiches gilt für die TSD-Länge von 4 bis 5 bp, die beim thug-T ermittelt wurde (Abschnitt 3.3.1.1 ab S. 38).

3.5.4 Die DNA-Transposons Tdd-4 und Tdd-5

Das Element Tdd-4 wurde bereits in der Einleitung als das erste beschriebene DNA-Transposon aus *D. discoideum* genannt. Ein neu aufgefundenes Element mit starker Ähnlichkeit auf den Ebenen von Organisation und abgeleiteter Proteinsequenz (s. unten) wurde Tdd-5 genannt, um es namentlich dem Tdd-4 zur Seite zu stellen. Wie in der Einleitung (Abschnitt 1.4.2 ab S. 14) ausgeführt, ist das Benennungsschema Tdd-X nur noch für die DNA-Transposons Tdd-4 und Tdd-5 im aktiven Gebrauch. Und ich schlage vor, das Schema nur noch für die Elemente dieser Familie zu verwenden.

Tdd-4 hat eine relativ komplizierte Genstruktur mit sechs Introns (WELLS 1999). Ich postuliere die Genstruktur des neu identifizierten Tdd-5 auf Grundlage zweier unabhängiger *in-silico*-Methoden (Abb. 28).

A. Das **Genanalyseprogramm GENEID** sagt ein einziges Gen mit 4 Introns vorher, aus dem sich ein Polypeptid von ca. 600 aa Länge erwarten lässt. Eine Vorhersage für das Startcodon fehlt (nicht dargestellt).

B. Ein **paarweiser Sequenzvergleich zwischen Tdd-5 und Tdd-4** unter Verwendung von TBLASTX bestätigt das durch GENEID vorhergesagte Gen in weiten Teilen. Die ersten vier Exons werden in weitgehend übereinstimmenden Regionen vorgeschlagen. Nur für die linke Grenze des ersten Exons, die rechte Grenze des dritten Exons und den carboxyterminalen Genbereich bestehen Unterschiede. Am 3'-Ende ist die Sensitivität der Genanalyse durch TBLASTX-Vergleich schwach und möglicherweise lückenhaft, weil die Proteinsequenzähnlichkeit zwischen Tdd-5 und Tdd-4 dort sehr niedrig ist.

Ich habe versucht, aus den teilweise voneinander abweichenden Vorhersagen unter Zuhilfenahme der ausgeprägten Proteinsequenzähnlichkeit zwischen Tdd-5 und Tdd-4 eine **bestmögliche Gesamtvorhersage** abzuleiten (Abb. 28 unten). Zu den Diskrepanzen zwischen den beiden Vorhersageergebnissen muss bemerkt werden, dass das Genanalyseprogramm GENEID zwar eigens für die Gensuche in *D. discoideum* trainiert wurde, dass jedoch andererseits bekannt ist, dass Transposons im Vergleich zu „normalen" Wirtsgenen eine verzerrte Codon-Nutzung zeigen (MRÁZEK & KARLIN 1999), so dass Genvorhersagen für Transposons stärker fehlerbehaftet sind als bei anderen Genen. Im Zweifelsfall sollte daher der Methode des TBLASTX-Vergleichs der Vorzug gegeben werden, sofern

konsistente Splice-Signale an den Rändern der Inseln von Sequenzähnlichkeit identifiziert werden können. Ich schlage eine Genstruktur mit vier Introns vor, die sich exakt an homologen Positionen zu den ersten vier Introns des Tdd-4 befinden (WELLS 1999), während die Intronflanken kanonische Splice-Signale besitzen. Für das fünfte und sechste Intron von Tdd-4 konnte keine Entsprechung beim Tdd-5 gefunden werden – insgesamt sinkt die Proteinsequenzähnlichkeit zwischen Tdd-5 und Tdd-4 im carboxyterminalen Bereich auf niedrige Werte ab (mittlere Zeile in Abb. 28, Abb. 29). Die postulierte Transposasesequenz von Tdd-5 hat eine Länge von 702 aa.

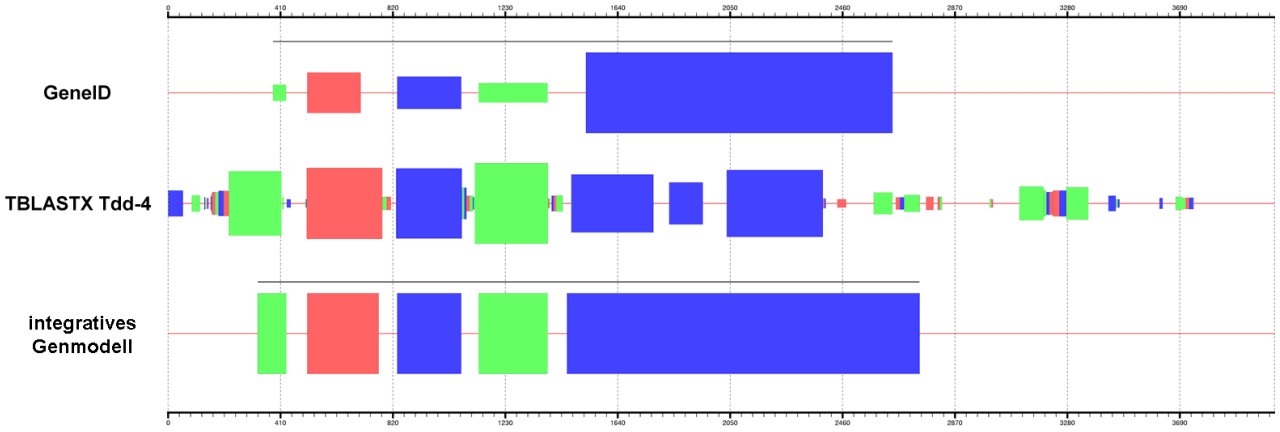


Abb. 28. Genstrukturanalyse für die Nukleotidsequenz des Transposons Tdd-5. Das skalierte Grafikfeld repräsentiert in horizontaler Ausrichtung die Sequenz von Tdd-5 in gesamter Länge. Die Vorhersagen und ein integratives Genmodell sind darin zeilenweise angeordnet, Kästen markieren postulierte CDS. Die Farbe jedes Kastens gibt die absolute Leserasterposition der vorhergesagten CDS in Bezug auf die Basenposition 1 wieder. Die Höhe der Kästen in den Zeilen „GENEID" und „TBLASTX_Tdd-4" korreliert mit log(Score) des zugrunde liegenden Vorhersageergebnisses.

Aus dem Alignment der Peptidsequenzen der Transposasen von Tdd-4 und Tdd-5 geht durch Analyse des lokalen Konservierungsgrads deutlich die Domänenstruktur des Proteins hervor (Abb. 29). Zwei konservierte Regionen können aufgrund von Sequenzähnlichkeit charakteristischen funktionellen Domänen zugeordnet werden. Zunächst liegt aminoterminal ein **HHCC-Motiv**, das dem der Integrase-Domäne von diversen Retroelement-Pol-Proteinen ähnlich ist, besonders dem des Woot-Retrotransposons aus *Tribolium castaneum* und dem des LTR-Retrotransposons skipper aus *D. discoideum*. Interessanterweise schließt sich carboxyterminal ein konservierter Sequenzbereich an, der keine Sequenzähnlichkeit zu anderen verwandten Proteinen zeigt. Es ist möglich, dass dieser konservierten Region eine Rolle bei der familienspezifischen Vermittlung der Substraterkennung, d.h. DNA-Bindung zukommt.

Weiter proximal befindet sich eine ausgedehnte konservierte Proteinregion, die deutliche Sequenzähnlichkeit zum **DD35E-Motiv** verschiedener Retroelement-Integrasen zeigt. Durch diese Sequenzähnlichkeiten wird die phylogenetische Brückenstellung der Elemente Tdd-4 und Tdd-5 zwischen den Integrasen der Retrotransposons einerseits und den Transposasen der DNA-Transposons andererseits eindringlich belegt (vgl. DOAK ET AL. 1994). Wie beim HHCC-Motiv setzt sich auch beim DD35E-Motiv carboxyterminal die Sequenzkonservierung fort, ohne dass die zusätzliche Region eine

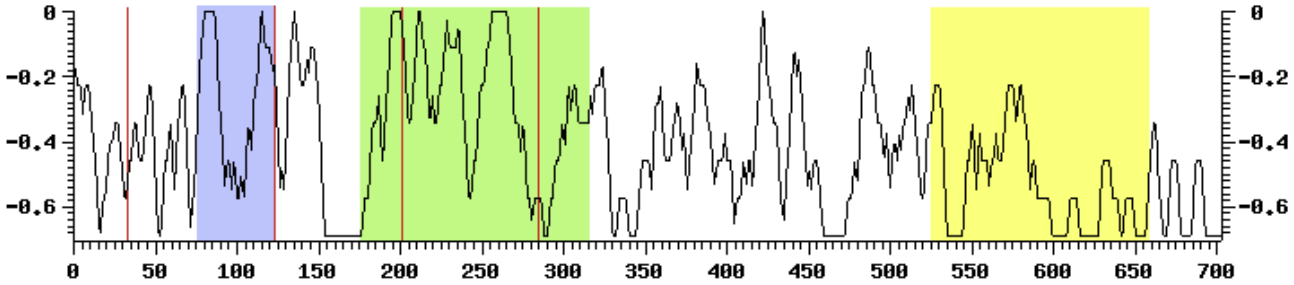


Abb. 29. Lokaler Konservierungsgrad im kompletten Alignment der Transposasen von Tdd-4 und Tdd-5. Zur Bemessung des lokalen Konservierungsgrads wurde die Entropie-Methode herangezogen (Abschnitt 2.6.3 ab S. 29). Hervorgehobene Regionen: **blau** – HHCC-Motiv, **grün** – DD35E-Motiv, **gelb** – prolinreiche Domäne. **Rote Balken** geben die Positionen der Introns im Transposasegen des Elements Tdd-5 wieder.

Sequenzähnlichkeit zu verwandten Proteinen aufweist. Auch hier kann ohne experimentelle Befunde vorerst nur vermutet werden, dass die Region eine familienspezifische Funktion vermittelt.

Carboxyterminal ist die Sequenzähnlichkeit zwischen den Transposasen von Tdd-4 und Tdd-5 sehr gering. Trotzdem ist diese Region beim Tdd-5 wie beim Tdd-4 sehr **prolinreich**, ähnlich wie die Integrase-Nachbarschaft bei einzelnen Viren (WELLS 1999). Eine differenzierte Funktion kann dieser „prolinreichen Domäne“ gegenwärtig nicht zugeschrieben werden. Ein **SPXX-Motiv**, das in der Transposase von Tdd-4 carboxyterminal in Tandemwiederholungen auftritt (WELLS 1999), wird beim Tdd-5 nicht gefunden. Es existiert auch nicht in den verschiedenen Sinnstrang-Translationen der Nukleotidsequenz stromabwärts des postulierten Stopcodons – für den Fall, dass die Vorhersage des Stopcodons beim Tdd-5 nicht zuträfe. Demnach lässt sich dem SPXX-Motiv der Transposase von Tdd-4 keine wesentliche funktionelle Bedeutung zuschreiben. Die Längen der Consensussequenzen von Tdd-4 und Tdd-5 unterscheiden sich nur um ca. 50 bp (1,3 %), und daher scheint es unwahrscheinlich, dass die von mir abgeleitete Sequenz des Tdd-5 unvollständig ist.

Das Element Tdd-4 ist sicherlich das Transposon mit der höchsten aktuellen Transpositionsaktivität in *D. discoideum*. Dieser Befund wird bereits gestützt durch die Tatsache, dass die Entdeckung (WELLS 1999) auf ein zufälliges Transpositionsereignis bei einem Transfektionsexperiment zurückgeht. Weiterhin habe ich von den 20 berichteten Flanken von Tdd-4 in *D. discoideum* AX4 (WELLS 1999) nur eine einzige (5 %) in den 20 unabhängigen Flanken aus unseren Schrotschusssequenzen wiedergefunden, während die Wiederauffindungsrate bei einer geschätzten Zahl von 80 Enden (vgl. Tab. 43 auf S. 74) etwa 25 % betragen sollte. Und drittens habe ich im Verlauf der Assemblierung genomischer Sequenz von *D. discoideum* immer wieder Abschnitte beobachtet, die in jeweils zwei Varianten auftauchen, die sich nur durch Vorhandensein oder Fehlen einer Tdd-4-Insertion voneinander unterscheiden. Eine Locus-Duplikation mit Disruption der einen Locus-Kopie durch Transposoninsertion erscheint in den meisten Fällen unwahrscheinlich, weil die Tdd-4-Variante immer nur durch einen einzigen Schrotschussklon repräsentiert wird. Ein gutes Beispiel für dieses Phänomen der „flüchtigen“ Tdd-4-Kopien ist der genomische Locus des Gens V4b. Es wurden zwei unabhängige Tdd-4-Insertionen gefunden, die jeweils nur durch einen einzigen Schrotschussklon repräsentiert sind (Abb. 30 oben). Der Locus des eng sequenzverwandten V4a-Gens trägt interessanterweise eine permanente Insertion des Transposons Tdd-5 (Abb. 30 unten). So scheinen

beide V4-Genorte ein attraktives Sprungziel gerade für diese Transposonfamilie zu sein. Die Existenz eines dritten V4-Locus wird weder durch die Sequenzen des Genomprojekts noch durch die molekulargenetische Analyse anderer Autoren unterstützt (MCPHERSON & SINGLETON 1993).

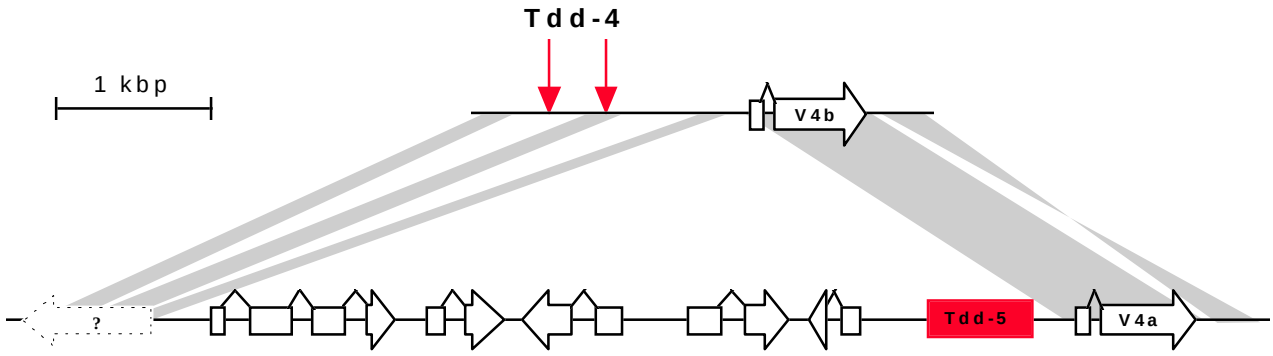


Abb. 30. Genomische Loci der V4-Gene. oben: V4b (Acc.No. X15380 und flankierende Sequenzen aus dem Assembly des laufenden Genomprojekts), unten: V4a (Acc.No. X15381 und flankierende Sequenzen aus dem Assembly des laufenden Genomprojekts). Graue Strahlen verbinden Regionen mit Sequenzähnlichkeit zwischen den beiden Loci. Transposoninsertionen sind rot hervorgehoben: Ein Kasten im V4a-Locus symbolisiert eine stabile Kopie, Pfeilmarkierungen am V4b-Locus markieren unabhängige, jeweils einmalig durch Schrotschusssequenzen repräsentierte Insertionen.

3.5.5 Neuartige DNA-Transposons: Die DDT-Familie

Drei neu aufgefundene repetitive Elemente aus dem Genom von *D. discoideum* bilden eine Verwandtschaftsgruppe. Sie haben einen invertierten terminalen Repeat (ITR) gemeinsam, dessen äußerster Bereich die streng konservierte Sequenz CACAGC...GCTGTG aufweist und der noch über weitere ca. 50 bp eine signifikante kreuzweise Sequenzähnlichkeit zeigt (Abb. 31). Diese gemeinsame strukturelle Eigenschaft deutet bereits darauf hin, dass es sich um eine Familie von DNA-Transposons handelt. Zwei vermutlich autonome Elemente von 5,2 und 5,5 kbp Länge haben jeweils eine ausgeprägte codierende Kapazität, die durch Vergleich der translatierten Sequenz hervortritt (s. unten). Ein drittes Element ist gegenüber den anderen stark verkürzt, mit einer Gesamtlänge von nur 758 bp. Es enthält offenbar keine codierenden Regionen, und demnach ist es kaum ein autonomes Element. Ich gebe der Familie den Namen „*Dictyostelium* DNA-Transposons“ (DDT). Die einzelnen Elemente werden DDT-A, DDT-B und DDT-S („S“ für shortened) genannt.

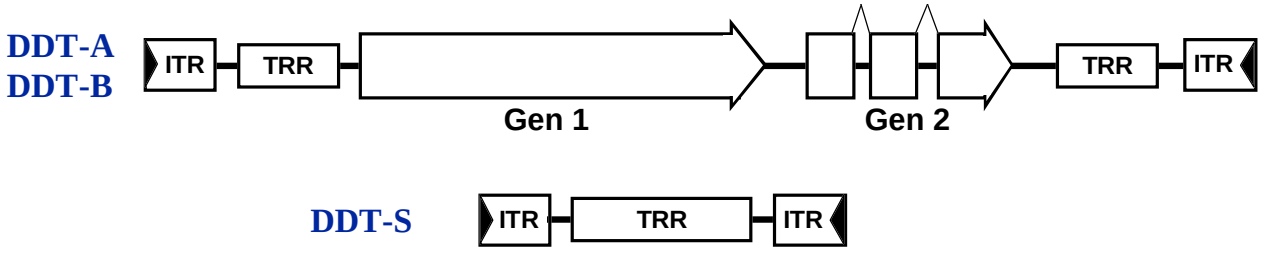


Abb. 31. Struktur der Transposons aus der DDT-Familie. Zur Entwicklung der Genmodelle s. Text und Abb. 34. Abkürzungen: ITR - invertierter terminaler Repeat, TRR - Tandem-Repeat-Region.

Die Herkunft des **DDT-S**-Elements und seine Beziehung zu den vermeintlich autonomen Elementen DDT-A und DDT-B ist wegen der kaum verfügbaren Vergleichskriterien schwierig. Die Orientierung des DDT-S ist aufgrund der fehlenden codierenden Kapazität frei wählbar bzw. augenscheinlich unbestimmbar. Um eine Klärung in dieser Frage zu finden, wurde ein phylogenetischer Baum für die ITRs aller drei DDT-Elemente berechnet (Abb. 32). Es fällt auf, dass der linke ITR des DDT-A sehr eng mit dem rechten ITR des DDT-S gruppiert. Gemessen an der Kürze der untersuchten Sequenzregionen ist die Bootstrap-Unterstützung gut. Ein weiteres diagnostisches Merkmal liegt in Tandem-Repeat-Regionen, die bei allen drei DDT-Elementen proximal benachbart zu den ITRs zu finden sind. Das Verteilungsmuster der Repeats und die Länge der repetitiven Einheiten sind jeweils für die Elemente charakteristisch (Abb. 33). Trotzdem lässt sich eine sehr starke Ähnlichkeit des Musters von DDT-S mit dem von DDT-A nicht verleugnen. Insgesamt sprechen die Befunde *uni sono* dafür, dass sich DDT-S von DDT-A oder einem DDT-A-nahen Element durch Deletion ableitet. Die Strukturen von ITRs und Tandem-Repeat-Regionen werden kongruent, wenn die vorab veröffentlichte Sequenz von DDT-S (Acc.No. AF298203) invertiert wird. Wahrscheinlich ist auch, dass DDT-S durch Translationsprodukte von DDT-A passiv mobilisiert wird.

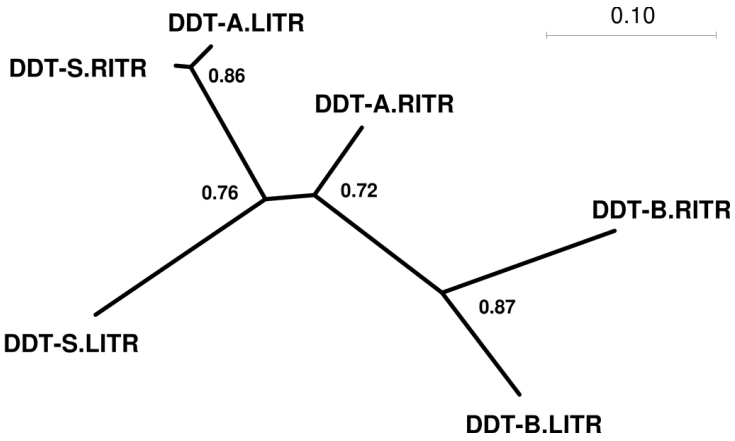


Abb. 32. Phylogenetischer Baum für die ITRs aller drei DDT-Elemente, berechnet mit den Programmen DNADIST / NEIGHBOR aus dem PHYLIP-Paket. Spezielle Abkürzungen: LITR = linker ITR, RITR = rechter ITR. Beim Element DDT-S beziehen sich die Bezeichnungen „links" und „rechts" auf die unter Acc.No. AF298203 veröffentlichte Sequenz. Die Astlängen im Baum korrelieren mit der Jukes-Cantor-Distanz (Maßstab oben rechts). Die Werte an den Knoten geben die relative Bootstrap-Unterstützung aus 100 Ziehungen wieder.

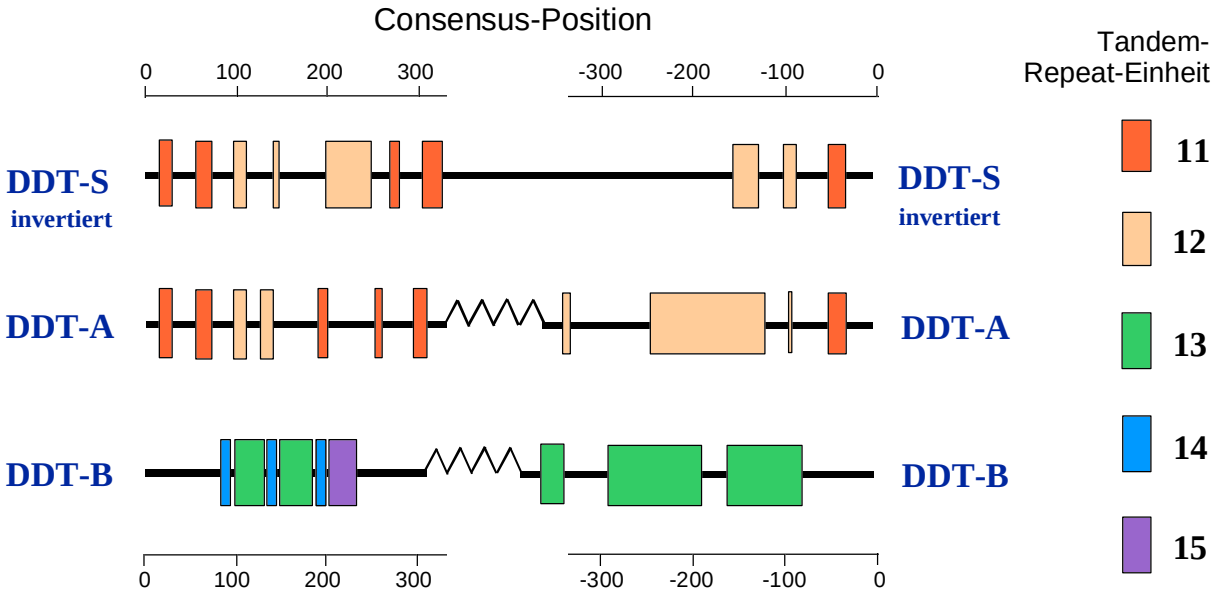


Abb. 33. Karte der Tandem-Repeats bei den DDT-Elementen, welche die Nachbarschaft der ITRs bilden. Eine Region wird als repetitiv gewertet, wenn in einem Fenster von ≥ 25 bp eine Basenkorrelation von mindestens 85 % besteht. Negative Positionsangaben beziehen sich auf das Sequenzende. Die Sequenz von DDT-S wurde gegenüber der veröffentlichten Referenz (Acc.No. AF298203) invertiert, um die Ähnlichkeiten zwischen DDT-S und DDT-A aufzuzeigen (vgl. Text).

Auf der Suche nach dem **Gengehalt der DDT-Elemente** wurde eine sorgfältige Analyse von Genstrukturen der Transposonspezies DDT-A und DDT-B vorgenommen. Dabei wurden drei unabhängige Methoden verfolgt (Abb. 34):

A. Das Genanalyseprogramm GENEID sagt zwei separate Gene vorher: Ein sehr langes, intronloses, das ein Polypeptid von ca. 800 aa Länge erwarten lässt. Ein zweites kürzeres mit drei Introns, das zu einem Polypeptid von ca. 300 aa Länge korreliert.

B. Ein paarweiser Sequenzvergleich zwischen DDT-A und DDT-B unter Verwendung von TBLASTX bestätigt das erste durch GENEID vorhergesagte Gen vollauf. Im Falle des zweiten Gens liefert TBLASTX jedoch nur für die ersten drei putativen Exons Treffer gering oberhalb des Rauschens, und die Exongrenzen variieren zwischen den beiden Vorhersagen erheblich.

C. Durch RT-PCR konnte ein Transkript für das zweite vorhergesagte Gen amplifiziert werden (Abb. 35 und Abb. 34 unten). Der linke Primer liegt mit einem Abstand von 23 bp stromaufwärts des Startcodons, der rechte Primer liegt proximal des putativen Stopcodons. Es konnten also die linke Gengrenze und die Intronpositionen und -grenzen unzweifelhaft bestimmt werden, die rechte Gengrenze bleibt jedoch durch den PCR-Befund unbestimmt. Es konnte kein Produkt für die Verbindung zwischen Gen 1 und Gen 2 erhalten werden. Überhaupt war es nicht möglich, ein Transkript für Gen 1 zu amplifizieren.

Anhand der Sequenzähnlichkeit lassen sich die Ergebnisse der Genanalyse für DDT-A weitgehend auf DDT-B übertragen. Danach codiert Gen 1 jeweils für ein Polypeptid von 839 und 815 aa Länge (in der Reihenfolge DDT-A, DDT-B). Die Startcodons sind dabei untereinander konserviert, das

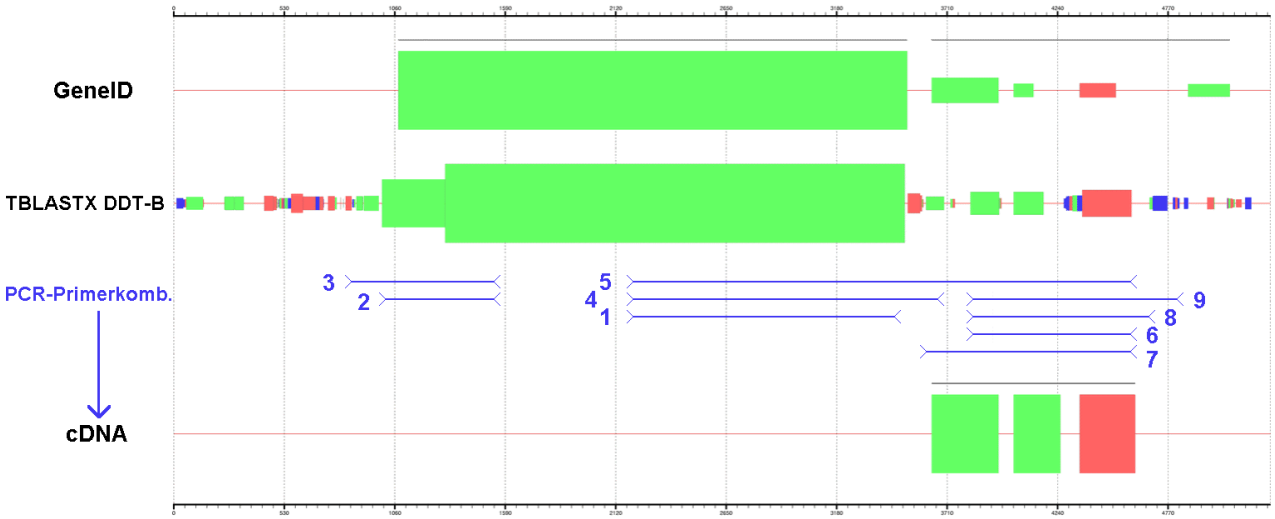


Abb. 34. Genanalyse für die Nukleotidsequenz des Transposons DDT-A. Das skalierte Grafikfeld repräsentiert in horizontaler Ausrichtung die Sequenz von DDT-A in gesamter Länge. Die Vorhersagen sind darin zeilenweise angeordnet, Kästen markieren postulierte CDS. Die Farbe jedes Kastens gibt die absolute Leserasterposition der vorhergesagten CDS in Bezug auf die Basenposition 1 wieder. Die Höhe der Kästen in den Zeilen „GeneID" und „TBLASTX_DDT-B" korreliert mit log(Score) des zugrunde liegenden Vorhersageergebnisses. Linien über den Kästen zeigen eine Gruppierung von Exons zu einem Gen an.

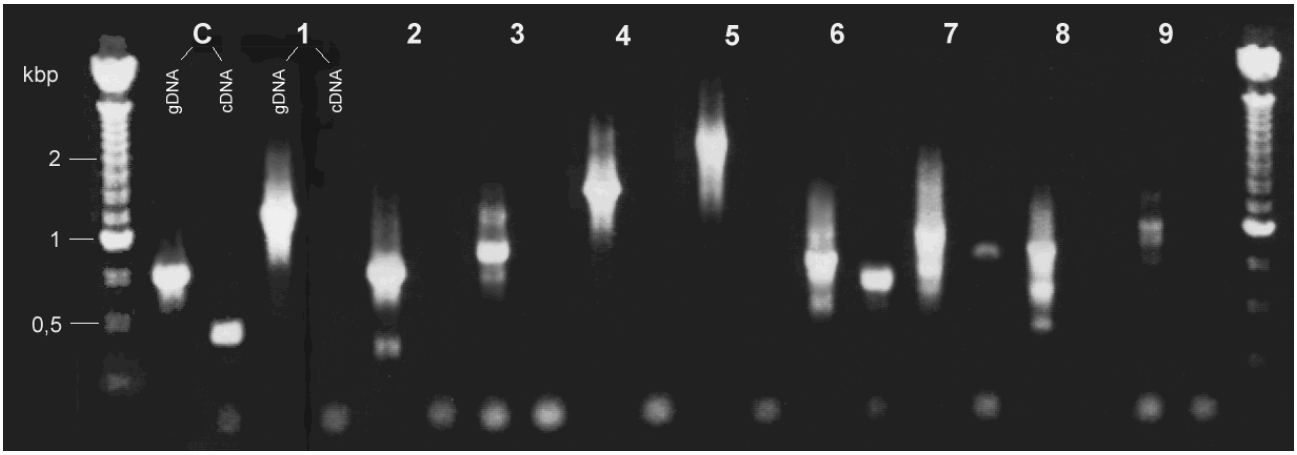


Abb. 35. PCR-Experimente zur Etablierung eines Genmodells für das Transposon DDT-A. Mit insgesamt 9 Primerpaaren wurden PCR-Reaktionen von gDNA und cDNA durchgeführt. Zahlen bezeichnen die Doppelbahnen für Produkte von gDNA (links) und cDNA (rechts) mit den Primerkombinationen 1 bis 9. Die Sequenzintervalle der verwendeten Primerpaare auf der Consensussequenz von DDT-A sind in Abb. 34 eingezeichnet: 1 – DDT-A.2200/DDT-A.3429R, 2 – DDT-A.981/DDT-A.1671, 3 – DDT-A.812/DDT-A.1671R, 4 – DDT-A.2220/DDT-A.3660R, 5 – DDT-A.2220 /DDT-A.4584R, 6 – DDT-A.3790/DDT-A.4584, 7 – DDT-A.3591/DDT-A.4584R, 8 – DDT-A.3790/ DDT-A.4672R, 9 – DDT-A.3790/DDT-A.4876R. „C" bezeichnet die Doppelbahn für Produkte der PCR-Positivkontrolle.

Stopcodon liegt beim DDT-A vier Codons hinter dem alignierten Genende von DDT-B. Gen 2 codiert jeweils für ein Polypeptid von 269 und 256 aa Länge (in der Reihenfolge DDT-A, DDT-B). Startcodon und beide Intronpositionen stimmen zwischen den Transposons überein, die Carboxy-termini lassen sich nicht zweifelsfrei alignieren, haben aber annähernd gleiche Längen. Die

Sequenzähnlichkeit der Polypeptide ist deutlich erkennbar: 52 % Identität für Polypeptid 1, 33 % für Polypeptid 2.

Der **lokale Konservierungsgrad**, ermittelt aus dem Alignment von Protein 1 der Transposons DDT-A und DDT-B (Abb. 36), ist sehr viel unschärfer verteilt als in den vorangegangenen Anwendungsfällen (Abb. 25 auf S. 51, Abb. 29 auf S. 55). Es findet sich ein scharfer Doppelpeak um Alignmentposition 410 und ausgedehntere Bereiche mit hoher Konservierung um die Positionen 500, 560, 640 und 740.

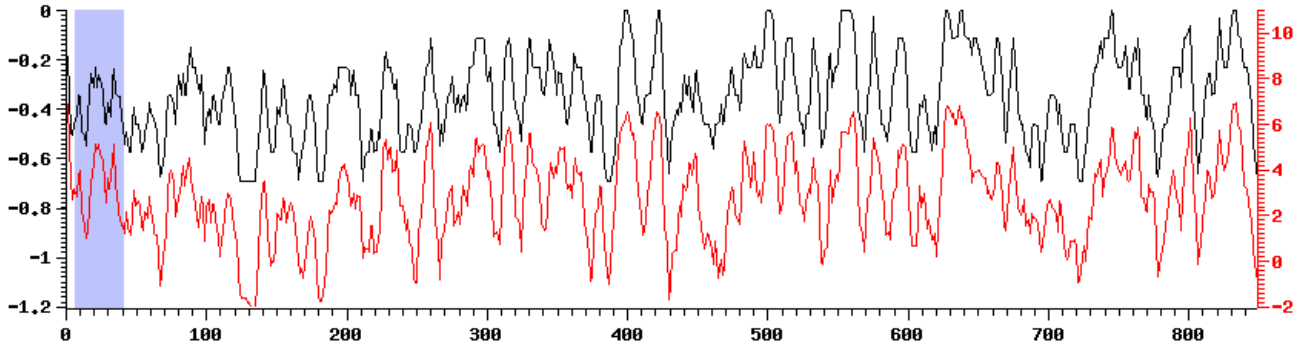


Abb. 36. Lokaler Konservierungsgrad im Alignment von Protein 1 der Transposons DDT-A und DDT-B. Methoden zur Bemessung des lokalen Konservierungsgrads (Abschnitt 2.6.3 ab S. 29): **schwarz** – Entropie-Methode, **rot** – Matrix-Methode. Hervorhebung der Domänenstruktur: **blau** – Zinkfinger-ähnliche Domäne.

Ungewöhnlicherweise ist durch BLASTP-Suchen mit den Proteinsequenzen dieser Transposongruppe **keine Sequenzähnlichkeit zu irgendeinem zuvor beschriebenen transposablen Element** – auch nicht zu irgendeinem zuvor beschriebenen Protein – nachweisbar. Eine BLASTP-Suche gegen PRODOM liefert Treffer im Bereich der Signifikanzgrenze, die nicht mit dem Profil der Konservierung zwischen den Elementen DDT-A und DDT-B zusammenfallen. Abfragen gegen Proteinmotivdefinitionen aus den Datenbanken PROSITE und PFAM, basierend auf Hidden-Markov-Modellen, weisen auf eine Zinkfinger-ähnliche Domäne im aminoterminalen Bereich von Protein 1 hin. Als Abwandlung des Grundmotivs $CX_{2\text{-}4}CX_{19}HX_{3\text{-}5}H$ (PROSITE PS00028 und PS50157) findet sich hier eine konservierte Sequenz der Formel $CX_2CX_{12}HX_3HX_3CX_2C$. Eine Ähnlichkeit besteht auch zum sogenannten „RING finger"-Motiv (PFAM PF00097).

3.6 Zielpräferenzen bei der Transposoninsertion

3.6.1 Spezifische Insertion in tRNA-Genflanken

Alle Transposons der TRE-Familie zeigen eine strikte Zielortspezifität bei ihrer Insertion (WINCKLER 1998). Um diese Spezifität näher zu untersuchen, habe ich die Abstände zwischen dem 5'-Ende von TRE-Insertionen und Promotormotiven der benachbarten tRNA-Gene ausgewertet. tRNA-Gene enthalten sogenannte A-Box- und B-Box-Motive als essentielle Bestandteile ihres internen PolIII-Promotors. Zusätzlich kann auch 3' ein zweites B-Box-Motiv auftauchen („exB-Box", MARSCHALEK & DINGERMANN 1991). 5'-verkürzte Elementkopien, die bei non-LTR-Retrotransposons häufig durch vorzeitigen Abbruch der reversen Transkription zustande kommen, wurden gleichermaßen ausge-

wertet, da der Insertionsort schon zu Beginn der reversen Transkription endgültig determiniert ist. Die Suche nach den Promotormotiven wurde mit Positional-Weight-Matrizen durchgeführt, die an veröffentlichten tRNA-Gensequenzen trainiert wurden. Die Ergebnisse sind in Tab. 37 zusammengefasst.

Tab. 37. Stichprobenanalysen der Abstände zwischen TRE3-Elementen und den Promotormotiven der benachbarten tRNA-Gene. Als maßgebliche Positionen gelten jeweils die ersten Positionen von Übereinstimmung mit der TRE-Consensussequenz und der Promotormotive (Consensus: A-Box TRNYNNRRN$_{1-2}$GG, B-Box und exB-Box RGGTTANNCCY). Abkürzungen: - = das Motiv ist nicht vorhanden/nachweisbar, ? = Die verfügbare Sequenz reicht nicht aus, um das Motiv zu identifizieren. Fett hervorgehoben ist der Abstand zu dem Motiv, das mutmaßlich die Insertion dirigiert hat.

Transposon	tRNA-Gen	Distanz [bp] gegenüber:			Beleg, Referenz
		A-Box	B-Box	exB-Box	
TRE3-A	Arg (UCU)	?	**72**	-	JAX4a77f03
TRE3-A	Ser (UGA)	130	**76**	26	JAX4b24d09
TRE3-A	Ala (AGC)	134	**91**	41	X53444
TRE3-A	Thr (AGU)	136	**94**	49	JAX4a35h11
TRE3-A	Val (AAC)	142	**98**	48	M24053
TRE3-A	Asp (GUC)	144	**102**	50	JAX4a56c10
TRE3-A	Ser (UGA)	159	**107**	-	JAX4a92d10
TRE3-A	Cys (ACA)	-	**108**	-	AF133115
TRE3-A	Arg (ACG)	155	**111**	57	X59561
TRE3-A	Val (UAC)	174	**114**	-	AF133116
TRE3-A	Val (UAC)	160	**116**	-	X03499
TRE3-A	Arg (ACG)	161	**117**	61	JAX4a134e01
TRE3-A	Glu (UUC)	162	**120**	68	M24566
TRE3-A	Tyr (GUA)	176	**123**	75	X53447
TRE3-A	Met (CAU)	?	**124**	74	K02645
TRE3-A	Ile (AAU)	168	**125**	70	U46204
TRE3-A	Thr (AGU)	176	**134**	84	JAX4a69b10
TRE3-A	Lys (UUU)	181	**138**	85	X59577
TRE3-A	Thr (UGU)	203	**141**	-	JAX4a20e08
TRE3-A	Glu (UUC)	187	145	**96**	M24567
TRE3-A	Leu (CAA)	201	148	**93**	JAX4a135d05
TRE3-A	Asn (GUU)	201	157	**105**	X53443
TRE3-A	Leu (UAA)	?	174	**120**	JAX4a186a09
TRE3-B	Thr (AGU)	94	**52**	-	JAX4a88h08
TRE3-B	Val (AAC)	99	**53**	-	AF067199, JAX4a135b01
TRE3-B	Gly (GCC)	113	**72**	20	JAX4a63a02
TRE3-B	Ile (AAU)	117	**74**	20	JAX4a143g01
TRE3-B	Phe (GAA)	?	**79**	29	U46205
TRE3-B	Val (AAC)	131	**87**	33	AF067200
TRE3-B	Met (CAU)	135	**92**	37	JC1a236g06
TRE3-B	Gln (UUG)	136	**93**	42	JAX4d01d09
TRE3-B	Lys (UUU)	150	**107**	36	JAX4a42b03

Transposon	tRNA-Gen	Distanz [bp] gegenüber:			Beleg, Referenz
		A-Box	B-Box	exB-Box	
TRE3-B	Gly (GCC)	153	112	**60**	JAX4a117a10
TRE3-B	Arg (UCU)	159	116	**68**	X59563
TRE3-C	Asp (GUC)	111	**69**	-	JAX4b19c10
TRE3-C	Asp (GUC)	118	**76**	-	JAX4b23h11
TRE3-C	Asp (GUC)	127	**85**	-	JAX4a137h06
TRE3-C	Lys (UUU)	147	**104**	49	X59576
TRE3-C	Pro (UGG)	146	**105**	48	JAX4b13e08
TRE3-C	Ile(AAU)	149	**106**	52	JAX4d10a02
TRE3-C	Met(CAU)	151	**109**	58	JAX4a16a06
TRE3-C	Gln (UUG)	153	**110**	-	JC2a103e07
TRE3-C	Pro (UGG)	169	**128**	-	JAX4b11a07
TRE3-C	Cys (GCA)	178	**136**	-	JAX4b07h04
TRE3-C	Thr (AGU)	188	146	**98**	JAX4a121b08

3.6.2 Allgemeine Verschachtelung von Transposons

Neben der funktionell begründeten Insertionsortspezififät der vorangenannten Transposons wurde in dieser Arbeit auch für unspezifisch inserierende Transposons eine Konzentrierung in Clustern beobachtet. Das geht hervor aus der Sequenzanalyse der Transposonflanken. Für Transposonindividuen kann in durchschnittlich 67 % der Fälle in der flankierenden Sequenz ein beliebiges anderes Transposon identifiziert werden (Tab. 38). Einerseits zeigt die Transposonassoziation innerhalb einer Transposonfamilie eine gleichbleibende Rate, zueinander sind die Assoziationsraten der Transposonfamilien jedoch ungleich: Tdd-4 und Tdd-5 sind jeweils zu ca. 50 % in andere Transposons inseriert, für die Mitglieder der DDT-Familie liegt diese Rate bei ca. 90 %.

Gegen die Hypothese, dass die Transposoninsertionen unterliegen der Freiheit, an jeder beliebigen Stelle des Genoms gleichermaßen stattfinden zu können, spricht die beobachtete Transposonballung statistisch hoch signifikant: $p << 1 \cdot 10^{-20}$ unter Anwendung des zentralen Grenzwertsatzes. Fortführend lässt sich der Anteil des Genoms, der frei von Transposonsequenzen ist, der aber die Insertion von Transposons unter neutraler Auswirkung auf die Vitalität des Wirts erlaubt, berechnen: Die vorhandenen Transposonsequenzen machen ca. 10 % der chromosomalen Sequenz aus, und etwa halb so groß ist die zusätzlich verfügbare potenzielle „Landefläche“ für Transposoninsertionen (33 % gegenüber 66 %, s. Summenzeile in Tab. 38), also 5 % der chromosomalen Sequenz. Aus den Ergebnissen der Assemblierung von Chromosom 2 (Abschnitt 3.1.1 ab S. 31) kann als Vergleich die Gendichte im Genom ermittelt werden: Gene einschließlich Introns machen 72 % der nichtrepetitiven Sequenz von Chromosom 2 aus. Introns sind bei *D. discoideum* relativ kurz (durchschnittlich 177 bp), daher kommen sie als neutrale Insertionsorte für Transposons nicht in Frage. Da in intergenischer Sequenz, die 25 % des Genoms ausmacht, nur halb so viele Transposoninsertionen stattgefunden haben wie im repetitiven Anteil des Genoms (10 %), kann

Tab. 38. Verteilung der Transposoninsertionsorte auf bestehende Transposonloci und anonyme genomische Sequenz. Die Zusammenstellung enthält alle identifizierten Transposonspezies, für die keine ausgesprochene Insertionsortspezifität nachgewiesen werden kann.

		Insertionsziel			
		Transposon		anonym	
Transposon	**analysierte Enden**	**N**	**%**	**N**	**%**
DIRS-1	15	13	87	2	13
skipper	10	9	90	1	10
DCLT-A	4	4	100	0	0
Tdd-4	20	10	50	10	50
Tdd-5	16	7	44	9	56
DDT-A	24	20	83	4	17
DDT-B	29	26	90	3	10
DDT-S	21	18	86	3	14
thug-S	14	3	21	11	79
thug-T	17	4	24	13	76
Summe	**170**	**114**	**67**	**56**	**33**

geschlossen werden, dass die intergenische Sequenz überwiegend (rechnerisch 80 %) funktionell relevant ist.

3.7 Assemblierung von Transposonkopien aus Schrotschusssequenzen

3.7.1 Relevante Parameter

Zwei wesentliche Parameter, die in Hinblick auf die Assemblierbarkeit von Transposonindividuen eine Rolle spielen, sind die Dichte von Polymorphismen und die korrespondierenden Allelfrequenzen für die beobachteten Polymorphismen. Diese Parameter gehen in die Berechnung der Sequenzdiversität π ein, die als ein Maß für die Abschätzung der Assemblierbarkeit herangezogen wird (vgl. Abschnitt 2.5.3 ab S. 27). Um die Beziehung zwischen Sequenzdiversität und Assemblierbarkeit zu untersuchen, habe ich Simulationen mit künstlich erzeugten Sequenzpopulationen durchgeführt (Abb. 39A). Der Zusammenhang zwischen sinkendem π aufgrund veränderter Allelfrequenz und der Sequenzlänge, die nötig ist, um mehrere Transposonindividuen spezifisch voneinander zu unterscheiden, hat einen schwach hyperbolischen Verlauf (Abb. 39B). Er kann ohne gravierenden Fehler linear angenähert werden.

Eine dritte Einflussgröße, nämlich die Anzahl der Transposonindividuen einer Spezies, bestimmt darüber hinaus die Schwierigkeit der Assemblierbarkeit. Sie geht wie die Allelfrequenz linear in die „Hartnäckigkeit“ gegen eine spezifische Assemblierung ein (nicht dargestellt). Die Anzahl der Transposonindividuen wird in dem neu eingeführten Maß R_A berücksichtigt (vgl. Abschnitt 2.5.3 ab S. 27).

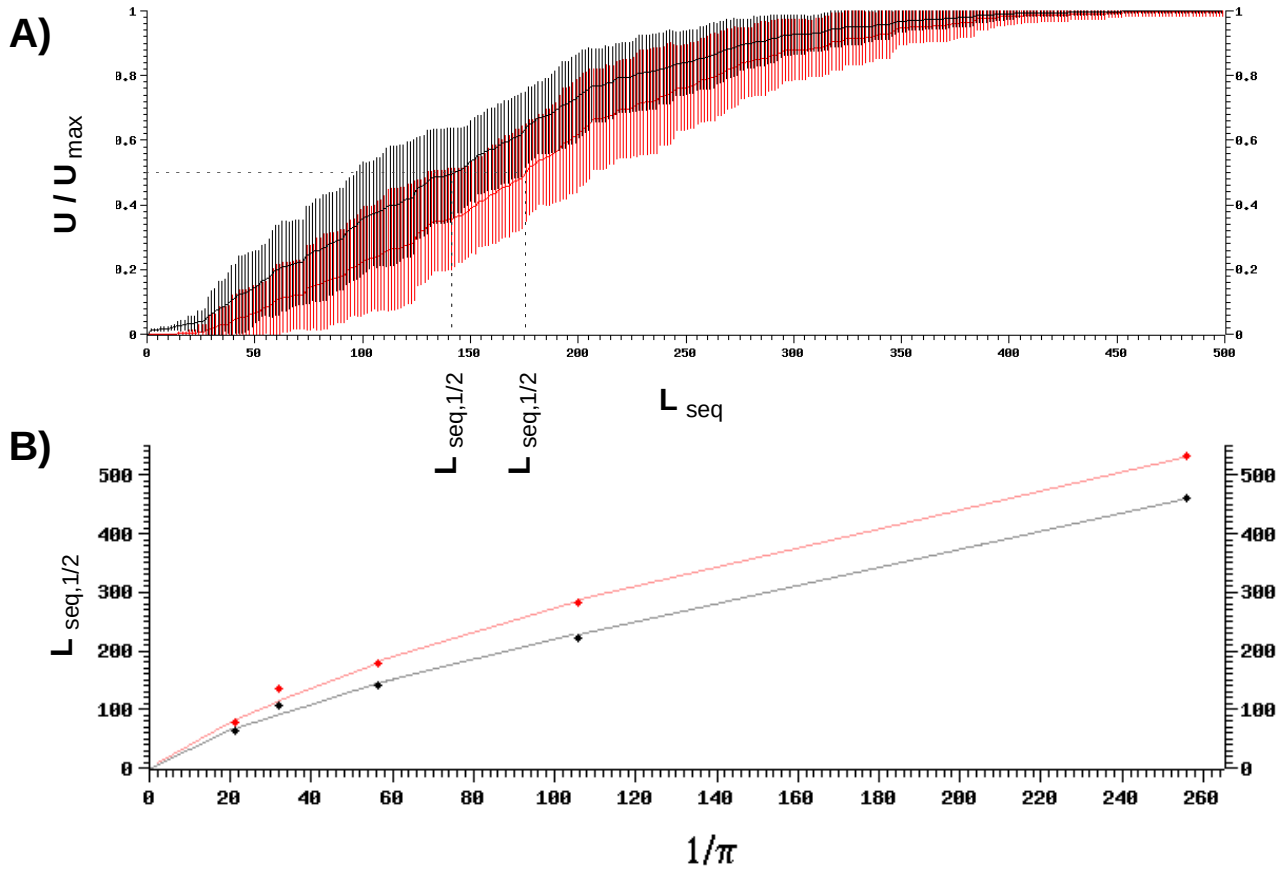


Abb. 39. Beziehung zwischen Sequenzdiversität und Unterscheidbarkeit (U / U_{max}) dargestellt an Simulationen mit künstlich erzeugten Alignments. **A)** Unterscheidbarkeit der Sequenzen bei π = 0,018, abhängig von der analysierten Sequenzstrecke – anhand der Zahl der unterscheidbaren Sequenzgruppen (schwarz) oder des Anteils der Sequenzen im Alignment, die eindeutig identifizierbar sind (rot). Senkrechte Balken markieren Intervalle der Standardabweichung. **B)** Halbwertspunkt $L_{seq,1/2}$ aus A) in Abhängigkeit von 1/π. Farbsymbolik wie unter A).

3.7.2 Maße π und R_A zur Abschätzung der Assemblierbarkeit von Transposonindividuen

Es wurden für alle identifizierten Transposonspezies in *D. discoideum* zwei Maße berechnet, um die Assemblierbarkeit für die Transposonspezies quantitativ zu fassen: π und R_A (vgl. Abschnitte 2.5.3 und 2.5.4 ab S. 27). Dazu wurden verifizierte Polymorphismen und deren Allelfrequenzen aus den Sequenz-Clustern ausgewertet. Die Ergebnisse für alle identifizierten Transposonspezies sind in Tab. 40 zusammengefasst. Eine Analyse der gefundenen Werte erfolgt gemeinsam mit anderen Parametern unter Abschnitt 4.4 ab S. 83.

Tab. 40. Sequenzdiversität π und das Assemblierbarkeitsmaß R_A für alle identifizierten Transposonspezies in *D. discoideum*.

Transposon	Sequenzdiversität π	Assemblierungs-widerstand R_A
DIRS-1	0,0234	1537
skipper	0,0215	512
DGLT-A	0,0042	417
DCLT-A	0,0072	255
TRE3-A	0,0121	967
TRE3-B	0,0143	486
TRE3-C	0,0071	771
TRE3-D	0,0032	450
TRE5-A	0,0104	1982
TRE5-B	0,0079	479
TRE5-C	0,0180	51
Tdd-4	0,0072	667
Tdd-5	0,0253	76
DDT-A	0,0211	159
DDT-B	0,0246	158
DDT-S	0,0294	842
thug-S	0,0154	161
thug-T	0,0417	51

3.7.3 Assemblierung eines genomischen Transposon-Clusters

In den vorigen Abschnitten wurden allgemein die Parameter berechnet, die eine objektive Einschätzung der Assemblierbarkeit von Transposonkopien ermöglichen. Da die verfügbaren Maße keine absolute Antwort geben, ob die Kopien einer bestimmten Transposonspezies erfolgreich assemblierbar sind oder nicht, wurde die Frage beispielhaft von der praktischen Seite angegangen.

Ein assembliertes genomisches Contig von ca. 33 kbp Länge trägt ein Cluster von Transposonfragmenten (Abb. 41). Die Assemblierung wurde durch Verfolgen einzigartiger polymorpher Ausprägungen (diagnostischer Positionen) geleitet, und sie wurde validiert in Hinblick auf die Paarung von Schrotschusssequenzen, die vom gleichen Schrotschussklon stammen. Es gibt keine Hinweise auf mögliche Chimärismen in dem Contig. Die Abdeckung stimmt mit der rechnerischen Abdeckung für eine unikale genomische Sequenz überein. Die Assemblierung wurde an einer beliebig gewählten Kopie des Transposons DDT-A begonnen; die Wahl der Region ist also weitgehend zufällig. Daher ist auch der bewältigte Schwierigkeitsgrad bei der Assemblierung der Transposonfragmente repräsentativ - jedenfalls für die in diesem Contig dominierenden Transposonspezies, die hauptsächlich der DDT-Familie angehören. Das Contig - und die Assemblierbarkeit - endet an Fragmenten des DIRS-1, einer Transposonspezies mit einem besonders hohen Schwierigkeitsgrad.

Das Contig ist ausschließlich aus Fragmenten derjenigen Transposonspezies zusammengesetzt, die kein ausgesprochen zielortspezifisches Insertionsverhalten zeigen: Elemente der DDT-Familie, DIRS-1 und – vertreten durch einige sehr kurze Fragmente – das Retrotransposon skipper und Elemente der thug-Familie. Die an dem Contig beobachtete Ballung von Transposonkopien entspricht den Verhältnissen, die sich aus Schrotschusssequenzen extrapolieren lassen (Abschnitt 3.6.2). In das

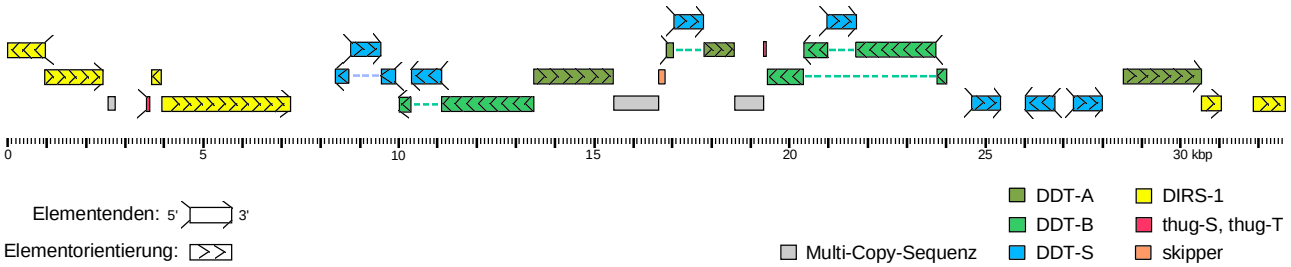


Abb. 41. Assemblierte Sequenz eines genomischen Clusters aus Transposonkopien (GENBANK Acc.No. AF298624). Die Zuordnung von Sequenzregionen zu einzelnen Transposonspezies geht aus der farbigen Kennzeichnung hervor (Legende rechts). Eine Pfeilsymbolik kennzeichnet Transposonenden und die Orientierung der Transposonfragmente (Legende links).

Cluster eingestreut sind längere Abschnitte von nicht-Transposonsequenz (fünf mit ca. 1 kbp Länge), die etwa zur Hälfte an mindestens einer anderen Stelle des Genoms sequenzähnlich wiederholt werden. Einer dieser niedrig-repetitiven Abschnitte trägt tRNA-Pseudogene. Der Anteil von Transposonsequenzenden auf dem Contig, die unmittelbar mit einer anderen Transposonkopie assoziiert sind, beträgt 60 %. Im genomweiten Mittel beträgt dieser Wert für die Transposonspezies der DDT-Familie 83 bis 90 % (Abschnitt 3.6.2).

3.7.4 Gesamtabschätzung der Assemblierbarkeit von Transposonindividuen

Die berechneten Maße π und R_A erlauben, die Transposonspezies in eine Reihenfolge bezüglich ihrer Assemblierbarkeit zu bringen (Abschnitt 3.7.2). Des Weiteren wurden praktische Versuche zur korrekten Assemblierung von Transposonloci unternommen (Abschnitt 3.7.3). Aus den Ergebnissen habe ich abgeschätzt, welcher Anteil der Transposonsequenzen assemblierbar ist und welcher nicht. Kopien der Transposonspezies mit $R_A < 160$ – das sind TRE5-C, Tdd-5, DDT-A, DDT-B, thug-S, thug-T – lassen sich zweifelsfrei assemblieren (Abschnitt 3.7.3). Fehlerfrei assemblierbar sind auch die Kopien der Transposonspezies und -individuen mit einer Länge < 1,5 kbp – das sind DDT-S, alle bekannten Kopien von TRE5-C, die meisten Kopien von TRE3-D –, da die Transposonsequenz mit großer Wahrscheinlichkeit durch Schrotschussklone überbrückt wird.

Nach ihrem Verteilungsmuster im Genom unterscheiden sich zwei Transposongruppen:

- Gruppe 1 bildet umfangreiche Kopien-Cluster aus zielortunspezifisch inserierenden Transposonspezies (Abschnitt 3.6.2). Diese Cluster sind ausschließlich um die Centromere angesiedelt. Sie haben einen Gesamtumfang von 5,9 % der chromosomalen Sequenz. Eingestreut in diese Cluster sind auch Sequenzen, die nicht von Transposons gebildet werden (Abschnitt 3.7.3). Deren Anteil lässt sich momentan nur grob auf 10 % der repetitiven Regionen schätzen.

- Gruppe 2 enthält die insertionsortspezifischen Transposonspezies, TRE-Elemente und DGLT-A, die Kopien-Cluster um tRNA-Genloci bilden. Die Kopienzahl dieser Spezies beläuft sich auf etwa 350. Die Kopienzahl von tRNA-Genen wird ausgehend von der beobachteten Zahl auf Contigs der Assemblierung von Chromosom 2 auf 400 hochgerechnet (GLÖCKNER ET AL. 2002). Gemäß Untersuchungen von MARSCHALEK & DINGERMANN (1991) betragen die Anteile von nicht, einseitig und beidseitig mit Transposons assoziierten tRNA-Genloci jeweils 62, 29 und 8 %. Schätzungsweise lassen sich durchschnittlich 2 kbp Randbereich dieser Transposon-Cluster aus Sequenzen randständiger Schrotschussklone assemblieren.

Abb. 42 stellt Anteile und Größenumfang der Transposonsequenzen dar, die nach obigen Überlegungen und Schätzungen fehlerfrei assembliert werden können. Die Rechnung unterscheidet zwischen

- durchweg repetitiven Regionen, die nicht assembliert werden können,
- assemblierbaren Transposonkopien, die von nicht assemblierbaren Regionen flankiert werden. Das in Abschnitt 3.7.3 beschriebene Contig fällt in diese Kategorie.
- nicht repetitiven Abschnitten größerer repetitiver Regionen und tRNAs, die beiderseits von tRNA-assoziierten Transposons flankiert werden,
- Rändern repetitiver Regionen, die an unikale genomische Sequenz stoßen.

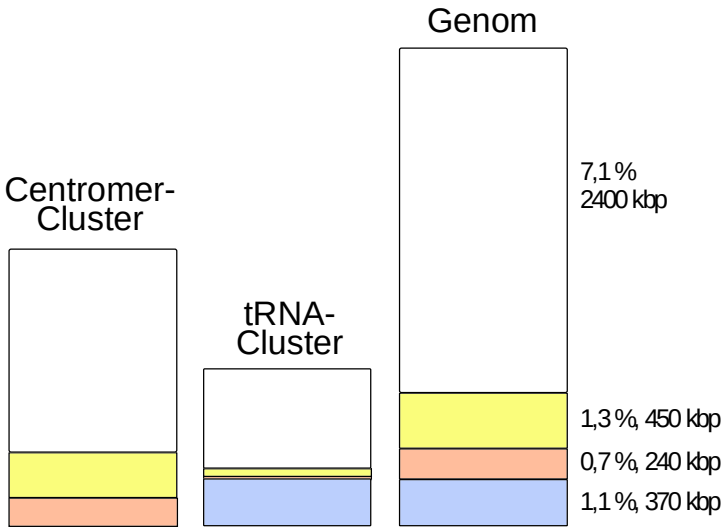


Abb. 42. Assemblierbare und nicht assemblierbare Anteile von Transposonsequenz im *Dictyostelium*-Genom. Durch farbliche Kennzeichnung werden unterschieden: **weiß –** nicht assemblierbare Transposonsequenz, **gelb –** Transposonsequenz, die assemblierbar ist, aber voraussichtlich zu kleinen, nicht kartierbaren Contigs führt, **orange –** nicht-Transposonsequenz, die Bestandteil von Transposon-Clustern ist und voraussichtlich zu kleinen, möglicherweise kartierbaren Contigs führt, **blau –** assemblierbare Transposonsequenz, die voraussichtlich Teil großer, kartierbarer Contigs sein wird.

4 DISKUSSION

4.1 Allgemeine Charakteristika der Transposons

4.1.1 Aufspüren neuer Transposonsequenzen

Verschiedene methodische Ansätze erlaubten mir die Identifizierung neuer transposabler Elemente. Die Strategien haben unterschiedliche Vor- und Nachteile hinsichtlich Arbeitsaufwand, Sensitivität, Spezifität und Ausbeute. Ein **kreuzweiser Vergleich aller Schrotschusssequenzen** wurde bereits in zurückliegenden Studien durchgeführt (GOODWIN & POULTER 2000), um repetitive Sequenzen zu identifizieren. So vielversprechend dieser methodische Ansatz aus theoretischer Sicht sein mag, in der Praxis erweist sich gerade die hohe Sensitivität der Methode als Nachteil, da sie eine große Zahl von falsch positiven Kandidaten liefert. Dies sind entweder unikale chromosomale Regionen mit ungewöhnlich hoher Coverage oder tatsächliche Multi-Copy-Sequenzen, die keine Transposons darstellen. Es existieren im Genom von *D. discoideum* sehr viele Multigenfamilien, die sich *a priori* nicht von multiplen Kopien einer Transposonspezies unterscheiden lassen (GLÖCKNER ET AL. 2002). Eine Unterscheidung ist erst durch Sequenzanalyse möglich, wobei sich Transposons durch terminale Sequenzwiederholungen (LTRs oder ITRs) und/oder codierende Kapazität mit Sequenzähnlichkeit zu bekannten Transposonformen diagnostizieren lassen. Die gezielte **Suche nach Mitgliedern gut charakterisierter Transposongruppen** mittels BLAST lässt Transposons unter den Schrotschusssequenzen mit hoher Spezifität erkennen. Sensitiv ist die Methode freilich nur für den Verwandtschaftsbereich der Transposonspezies, die zur Suche eingesetzt wurden. Eine Reihe von Transposonspezies in *D. discoideum* wäre zwangsläufig unerkannt geblieben, wenn ausschließlich diese Methode angewandt worden wäre: Die Elemente der DDT-Familie und der thug-Familie. Trotzdem stützen sich andere Studien allein auf diese Strategie (CROLLIUS ET AL. 2000). Eine neuartige Methode zur Suche nach Transposonsequenzen ergibt sich aus der **Analyse von Transposonverschachtelungen**, wie sie im Genom von *D. discoideum* häufig vorkommen (vgl. Abschnitt 3.6.2 ab S. 62). Diese Methode ist speziell geeignet, Transposons neuartiger Typen und solche mit geringen Kopienzahlen aufzuspüren. In dieser Hinsicht bildet sie eine sinnvolle Ergänzung zu den anderen beiden Analysemethoden.

Im Anschluss an diese Studie stellt sich die Frage: Wurden alle Transposonspezies im Genom identifiziert? Die Methode der Analyse von verschachtelten Transposonloci wurde erschöpfend angewandt. Es können dadurch aber nur diejenigen Spezies identifiziert werden, deren Loci in Verschachtelungen vorkommen. Ein Beispiel für einen widrigen Fall ist DGLT-A, ein streng insertionsortspezifisches Transposon. Es wurde während der fortschreitenden Assemblierung des Genoms von *D. discoideum* entdeckt, wäre aber auch durch die später erfolgte Suche nach Transposonhomologen aufgefunden worden. Aber auch nach einer erschöpfenden Homologiesuche

können Transposonformen unentdeckt bleiben, wenn es sich um neuartige Formen handelt, die keine deutliche strukturelle Beziehung zu den bekannten Transposons besitzen und in recht niedriger Kopienzahl – weniger als zehn – im Genom vertreten sind.

4.1.2 Aufbau von Alignments für Transposonsequenzen

Aus mehreren Gründen war es nützlich, Sequenz-Cluster für alle zu untersuchenden Transposonspezies aufzubauen:

- Zunächst erweist sich eine erschöpfende Sammlung von Schrotschusssequenzen für jede Transposonspezies in Form einer Liste als vorteilhaft. Sie enthält die Sequenzen, die während der Assemblierung der genomischen Sequenzen von *D. discoideum* gesondert zu berücksichtigen sind. Die Sammlung von Schrotschusstreffern ist als repräsentative Stichprobe gleichzeitig eine solide Grundlage für die Abschätzung der Kopienzahl jeder Transposonspezies.

- Die vermutlich naheliegendste Analyse eines solchen Transposonsequenz-Clusters ist die Ableitung einer Consensussequenz. Eine Consensussequenz, die sich aus Schrotschusssequenzen ableitet, hat gegenüber Klonierung und Analyse einzelner Transposonkopien den Vorteil, dass Einzelphänomene von Deletionen, Inversionen und nicht-repräsentativen Basenaustauschen nicht einfließen.

- In einem Sequenz-Cluster können Polymorphismen unter den verschiedenen Transposonkopien identifiziert und eingehend analysiert werden. Subformen der Transposons mit wiederkehrenden Deletionen oder diagnostischen polymorphen Ausprägungen lassen sich erkennen.

- Aus den alignierten Sequenzen des Clusters lassen sich auch die Übergänge von Transposon- zu nicht-Transposon-Sequenzen ablesen. Durch Analyse der nicht-Transposon-Sequenzen lassen sich Schlüsse über die Nachbarschaft der Transposons ziehen. Aufgrund der Repräsentativität der beobachteten Übergänge lässt sich aus beobachteten transposoninternen Übergangspunkten der Fragmentierungsgrad der Transposonspezies schließen.

GOODWIN & POULTER (2000) haben zur Untersuchung des Transposongehalts im Genom von *Candida albicans* einen ähnlichen Weg zur Organisation und Analyse der Schrotschusssequenzen eingeschlagen. Die Autoren beschränken sich jedoch im Wesentlichen auf die Identifizierung und Abgrenzung der Transposonspezies. Die hier vorgestellte Prozedur zum Erstellen der Sequenz-Cluster ist in ihrem Grundprinzip dem von PSI-BLAST verwandt (ALTSCHUL & KOONIN 1998). Diese Anwendung zielt jedoch allein auf ein Domänen-Clustering von Proteinsequenzen ab. Durch diese Spezialisierung auf Proteindomänen, die im Vergleich zu Nukleotidsequenzen immer relativ kurz ausfallen, erübrigt sich das Problem des Hantierens mit versetzt überlappenden, kurzen Sequenzfragmenten, die unter Umständen sehr lange Alignments bilden. Darüber hinaus ist PSI-BLAST als abgeschlossene Anwendung in der Programmiersprache C entwickelt, was die Weiterentwicklung oder Anpassung komplizierter gestaltet als das hier entwickelte integrative Konzept mit den Komponenten BLASTN und CLUSTAL W.

4.1.3 Genomischer Anteil und Fragmentierungsindex

Ganz wesentlich für die Schätzung des genomischen Anteils von Transposonsequenzen ist die Größe des Genoms, und zwar des chromosomalen Anteils des Genoms. Das extrachromosomale rDNA-Element spielt für die Betrachtung der transposablen Elemente keine Rolle, da es keine konstitutiven Transposoninsertionen enthält (nicht dargestellte Ergebnisse). In Abschnitt 3.1.2 ab S. 32 habe ich eine **Schätzung der Genomgröße** vorgestellt. Sie basiert auf der statistischen Interpretation von Sequenzen aus einer repräsentativen Sequenzbibliothek, die mit einer beobachteten Dichte auf einen unikalen Ausschnitt des Genoms fallen. Es ist damit erstmals eine Schätzung der Genomgröße verfügbar, deren Ergebnis und Unsicherheitsanteil rational validierbar ist. COX ET AL. (1990) haben aus der elektrophoretischen Auftrennung ganzer Chromosomen des *D.-discoideum*-Stamms AX3k den chromosomalen Anteils des Genoms auf etwa 40 Mbp geschätzt. Es liegt auf der Hand, dass die Größenbeurteilung von DNA-Molekülen mit mehr als 4 Mbp Länge auf einem Agarosegel sehr unsicher ist. Mein Schätzergebnis von 34 Mbp unterscheidet sich überhaupt nicht von der Genomgröße, die sich aus der YAC-basierten physikalischen Karte des Genoms ergibt (KUSPA & LOOMIS 1996). Wie in der Einleitung dargelegt, ist das Vertrauen in die YAC-Karte durch Ergebnisse eines unabhängigen Mappings getrübt (KONFORTOV ET AL. 2000). Eine Erklärung dafür, dass die Größeneinschätzung durch die Arbeiten von KUSPA & LOOMIS dennoch zutrifft, liegt möglicherweise darin, dass der physikalischen Karte ein Grundstock von identifizierten Restriktionsfragmenten aus dem Verdau mit drei verschiedenen selten schneidenden Endonukleasen zugrunde liegen (KUSPA & LOOMIS 1994). Die Größe der Restriktionsfragmente, die maximal 2 Mbp erreicht, lässt sich aus der Auftrennung durch Pulsfeldgelelektrophorese recht zuverlässig ermitteln.

Ich habe **Schätzungen von Nukleotidanteil** und Mindestkopienzahl aller Transposonspezies des Genoms beschrieben (Abschnitt 3.4.1 ab S. 44). Wird die Schätzmethode anhand der zahlreichen Genkopien des Aktins kontrolliert, so ist eine signifikante Abweichung von der anzunehmenden Kopienanzahl zu erkennen. Ein Abweichungstrend vom erwarteten Wert zeichnet sich ebenfalls ab, wenn der Schätzung die Zahl der Schrotschusstreffer anstelle des Nukleotidumfangs zugrunde gelegt wird. Diese Ergebnisse sind ein Indiz für einen Bias hinsichtlich Klonierbarkeit und Sequenzierbarkeit lokaler DNA-Abschnitte des Genoms von *D. discoideum*, auf den bereits mehrfach hingewiesen wurde. Als Konsequenz beruht die Unschärfe der vorgestellten Schätzungen nicht nur auf dem Umfang der untersuchten Schrotschusssequenzen sondern auch merklich auf der eingeschränkten Anwendbarkeit des statistischen Modells. Ich halte es daher nicht für zweckmäßig, Vertrauensintervalle der Schätzungen zu berechnen. Für das verfolgte Leitziel dieser Arbeit, dem Einschätzen von Schwierigkeiten und Möglichkeiten bei der Assemblierung von repetitiver genomischer Sequenz, reicht es aus, Kopien- und Fragmentzahlen in ihren Größenordnungen zu bestimmen. Immerhin gilt für die berechneten Schätzwerte (Tab. 43 auf S. 74), dass sie die tatsächliche Situation auf Grundlage der verfügbaren Daten bestmöglich beschreiben.

In vorangehenden Studien haben andere Autoren die Kopien- und Fragmentzahlen durch Southern Blot bestimmt. Für DIRS-1 wurden 40 intakte und etwa 200 degenerierte Kopien in einem nicht näher benannten Stamm von *D. discoideum* ermittelt (CAPPELLO ET AL. 1984). In der vorliegenden Arbeit wurde etwa der gleiche Gesamtwert gefunden, jedoch eine weitaus geringere Fragmentierung.

Die Diskrepanz erklärt sich möglicherweise in einer hohen Sequenzdiversität des DIRS-1, die auch die Restriktionsstellen betreffen kann und so zusätzliche Restriktionslängenpolymorphismen hervorruft, die von den Autoren als Anzeichen für Fragmentierung gewertet wurden. Das Element skipper kommt laut LENG ET AL. (1998) in 15 bis 20 wenig fragmentierten Kopien in den Stämmen AX-2 und AX-3 vor. Meine Schätzung ergibt einen mehr als doppelt so großen Wert. Außer der Schätzungsungenauigkeit könnte ein Unterschied in den untersuchten Schleimpilzstämmen die Diskrepanz der Werte erklären. MARSCHALEK & DINGERMANN (1991) schätzten Kopienzahlen von 200 TRE5-A-Elementen und 800 tRNA-Gene für die Genome der Stämme AX-2 und AX-3. Diese Schätzung der Zahl der tRNA-Gene ist verglichen mit der beobachteten Dichte auf 6,5 Mbp Sequenz von Chromosom 2 sehr hoch (vgl. Abschnitt 3.1.1 ab S. 31). Unsere Daten lassen 340 bis 400 tRNA-Gene im Genom des Stamms AX-4 erwarten. Entsprechend beträgt auch meine Schätzung der Zahl von TRE5-A-Elementen nur das 0,45-fache des Literaturwerts. Die Kopienzahl von TRE3-C wurde von POOLE & FIRTEL (1984) für verschiedene Stämme auf 10 bis 15 geschätzt. Wenn man berücksichtigt, dass in der zitierten Arbeit mit einer Sequenz des 5'-Endes hybridisiert wurde und dass 5'-Verkürzungen bei den non-LTR-Retrotransposons häufig sind, korrespondiert der Literaturwert gut mit der hier ermittelten Zahl von über 30.

4.1.4 Target-Site-Duplikationen

Es wurden mehrere Strategien verfolgt, um die Länge der TSD für möglichst alle Transposonspezies zu bestimmen. Die Analyse der **Sequenz von kompletten Transposonloci** ist der meistbeschrittene Weg (JURKA 1997, WELLS 1999). Im Rahmen dieser Arbeit war die Methode nur bedingt einsetzbar, weil die vollständige Sequenz von Transposonkopien nur in Einzelfällen zur Verfügung stand. Über die Schrotschusssequenzen stehen komplette Transposonloci von kurzen oder verkürzten Transposons zur Verfügung. Einen Sonderfall stellen die TRE-Elemente mit ihren sehr langen TSDs dar. Hier konnten – unabhängig von einer vollständigen Sequenzierung und evtl. Assemblierung von vollständigen Transposonkopien – zusammengehörige Flanken allein aufgrund der unzweifelhaften Identität der TSD-Sequenz identifiziert werden. Diese Möglichkeit eröffnet auch einen vielversprechenden Ansatzpunkt für die Paarung von Contig-Enden während der Assemblierung, wenn durch eine nicht assemblierbare Transposonkopie eine Sequenzlücke verursacht wird.

Ein alternativer methodischer Ansatz ist die **inverse PCR**. Mit Hilfe dieser Technik ist es möglich, gezielt Paare von flankierenden Sequenzen einer einzelnen Transposonkopie zu erhalten (WELLS 1999). Im Fall des Transposons DDT-S bestanden zum einen Probleme, in der A/T-reichen Sequenz von *D. discoideum* Primerregionen zu definieren. Des weiteren ergab sich, dass durch die nachbarschaftliche Topologie der zahlreichen Kopien von DDT-S ein dominanter Hintergrund von unerwünschten PCR-Produkten erhalten wurde (Abschnitt 3.3.1.2 ab S. 39). Aus meiner Sicht eignet sich daher der methodische Ansatz der inversen PCR nicht zur systematischen Analyse von Transposonflanken, und er wurde daher nicht weiterverfolgt.

Die beiden zuvor beschriebenen Methoden zielen darauf ab, Flankenpaare von Transposonloci in weitgehend konservierter Topologie in eine interpretierbare Sequenz zu überführen. Eine demgegenüber andersartige Herangehensweise ist die **Analyse von verschachtelten Transposonloci**, die durch

Auswertung der Transposonsequenz-Cluster gefunden werden können (Abschnitt 3.3.3 ab S. 42). Wenn eine Transposonkopie A in eine Transposonkopie B inseriert ist, dann müssen in den Schrotschusssequenzen zwei Zeugnisse von dieser Topologie zu finden sein: Der Übergang von A nach B und der Übergang von B nach A. Die Übergangspositionen, bezogen auf die Consensussequenz von Transposon A, geben exakt die Länge der TSD für Transposonspezies B an. Der Erfolg dieser Methode wird nicht durch Klonierbarkeit und Länge der Transposonsequenz oder deren Assemblierbarkeit beeinträchtigt. Lediglich muss die Abdeckung des Genoms durch Schrotschusssequenzen ein gewisses Mindestmaß erreichen, damit die benötigten Sequenzen der Übergangspunkte gefunden werden können. Die grundlegende Voraussetzung, dass verschachtelte Transposonloci überhaupt existieren, ist im Genom von *D. discoideum* klar erfüllt (Abschnitt 3.6.2 ab S. 62).

Für die meisten der identifizierten Transposonspezies konnte ich TSD-Sequenzen ermitteln (Abschnitt 3.3 ab S. 36, Zusammenfassung in Tab. 43). Die vorhandenen Sequenzdaten lieferten keine Informationen für die TSD des Transposons Tdd-4. In einer zurückliegenden Arbeit konnte deren Länge auf (4-)5 bp bestimmt werden (WELLS 1999), und im Einklang damit steht die hier bestimmte TSD-Länge von 5 bp für die neu aufgefundene, verwandte Transposonspezies Tdd-5. Für die Transposonspezies thug-S liegt ebenfalls kein Ergebnis vor. Grund dafür mag die abzuschätzende niedrige Kopienzahl und der hohe Fragmentierungsindex dieser Transposonspezies sein. Experimente mit inverser PCR schlugen fehl (Ergebnisse nicht dargestellt), so dass vermutlich keine intakten Kopien dieses Transposons im Genom von *D. discoideum* existieren und daher auch keine TSDs ermittelt werden können. Für das verwandte Transposon thug-T konnten zwei Flankenpaare und die korrespondierenden TSD-Sequenzen ermittelt werden, allerdings mit uneinheitlicher Länge von jeweils 4 und 5 bp. Auch wenn daher die Ergebnislage für die Transposons der thug-Familie dünn und unscharf ist, wird doch ein bemerkenswerter Bezug zu den LTR-Retrotransposons hergestellt, deren TSD-Länge ebenfalls 4 bis 5, selten 3 bp beträgt (KUMAR & BENNETZEN 1999). Auch die TSD-Länge des Transposons DGLT-A, die mehrmals 4 bp und in einem Fall 5 bp beträgt, ist mit der Zugehörigkeit des DGLT-A zu den LTR-Retrotransposons konform.

Hinsichtlich alternativer Längen der TSD stellen die Transposons der TRE-Familie einen Sonderfall dar. Für die non-LTR-Retrotransposons gelten spezielle Unwägbarkeiten aufgrund der variablen Länge der Transposonkopien und der dadurch häufig undefinierbaren Kopiengrenzen. Es sind lediglich die minimalen Wertebereiche angegeben, die aufgrund der Datenlage gesichert sind (Tab. 43). Demnach erzeugen mindestens drei der sieben TRE-Elemente eine TSD mit variierender Länge. Damit im Einklang stehen Berichte anderer Autoren: 9-10 bp für TRE3-A (POOLE & FIRTEL 1984) und 13-15 bp für TRE5-A (MARSCHALEK ET AL. 1993). Offenbar ist die TSD-Länge bei den TRE-Elementen ein wenig konserviertes Merkmal; das zeigt auch der Vergleich zwischen den einzelnen Vertretern der TRE-Familie. Nur in der Krone des Verwandtschaftsbaums kann eine Korrelation zum Verwandtschaftsgrad festgestellt werden (vgl. Abb. 26 auf S. 52).

Tab. 43. Überblick über die transposablen Elemente in *Dictyostelium discoideum* und deren Charakteristika. Eine Angabe der Consensussequenzlänge in Klammern zeigt an, daß das Element nicht vollständig aufgelöst werden konnte; dementsprechend sind alle anderen Angaben vorläufig. Symbole und Abkürzungen: * = Die Ergebnisse dieser Arbeit führen zu gravierenden Korrekturen oder zur Vervollständigung des Elements; ** = Das Transposon wurde im Verlauf dieser Arbeit entdeckt; # = Die Transposonspezies kommt in zwei deutlich unterscheidbaren Subformen vor; **DR** = direkter Repeat; **FI** = Fragmentierungsindex (= ML(N) / N_{min}); **IR** = invertierter Repeat; **ML(N)** = Maximum-Likelihood-Schätzer für die tatsächliche Fragmentanzahl im Genom; **N_{min}** = Kopienzahl auf Basis des Nukleotidgehalts im Genom; **n.u.** = nicht untersucht; **TSD** = Länge der Target-Site-Duplikation; **z1** = Wert zitiert aus WELLS 1999.

Transposon					LTR				Fragmentanzahl		
Klasse, Subklasse	Familie	Spezies	Acc.No.	Consensus–Länge [bp]	Typ	Länge [bp]	TSD [bp]	Genom–anteil [% nt]	N_{min}	ML(N)	FI
LTR–Retrotransposons	DIRS–1	DIRS–1	M11339	4826	IR	320	0	3,260	235	302	1,3
	Gypsy–ähnl.	skipper	AF049230	6994	DR	390	5	0,997	48	82	1,7
		skipper LTR		390				0,011	9	n.u.	n.u.
		DGLT–A **#	AF298204	5054	DR	268	4	0,067	5	7	1,5
		H3R		268				0,013	16	n.u.	n.u.
	Copia–ähnl.	DCLT–A **	AF4740004	6178	DR	208	4	0,109	6	7	1,1
non–LTR–Retrotransposons	TRE	TRE3–A	AF134169	5243	–	–	10	0,960	62	82	1,3
		TRE3–B *	AF134170	5292	–	–	12	0,770	52	68	1,3
		TRE3–C *	AF134171	4751	–	–	15–16	0,450	33	36	1,1
		TRE3–D **	AF135841	(1571)	–	–	15	0,041	9	16	1,8
		TRE5–A #	X57034	~6200	–	–	14–15	1,220	71	92	1,3
		TRE5–B **	AF298209	~5700	–	–	11	0,200	17	36	2,1
		TRE5–C **	AF298210	(890)	–	–	13–16	0,012	4	12	3,0
DNA–Transposons	Tdd–4	Tdd–4	U57081	3839	IR	146	[z1] 5	0,425	38	55	1,4
		Tdd–5 **	AF298206	3783	IR	183	5	0,076	6	20	3,3
	DDT	DDT–A **	AF298201	5169	IR	48	2	0,309	20	65	3,3
		DDT–B **	AF298202	5471	IR	38	2	0,314	19	68	3,6
		DDT–S **	AF298203	758	IR	27	2	0,295	132	175	1,3
unklassifiziert	thug	thug–S **	AF298207	2192	IR	18	?	0,058	9	19	2,1
		thug–T **	AF298208	1132	IR	8	4–5	0,038	11	18	1,6
	Summe							**9,625**	**802**	**1185**	

4.2 Sequenzanalyse der aufgefundenen Transposonspezies

Im Verlauf dieser Studie habe ich 18 Transposonspezies im Genom von *D. discoideum* identifiziert und analysiert, deren Consensussequenzen eine Gesamtlänge von 76 kbp ergeben. 12 der beobachteten Transposonspezies werden hier erstmals beschrieben (Gesamtconsensuslänge 44 kbp). Diese Sequenzinformation stellt nicht nur eine wertvolle Ressource für die Assemblierung des Genoms dar (Abschnitt 4.4), die Analyse der Transposonsequenzen erlaubt auch einen Einblick in die molekulare Biologie der mobilen Elemente.

4.2.1 Analysemethoden

Für die Elemente Tdd-5, DDT-A und DDT-B wurde ein vergleichender Ansatz zur **Analyse von Genstrukturen** verfolgt. Es steht zwar mit GENEID (PARRA ET AL. 2000, Abschnitt 2.6.1) ein eigens für *D. discoideum* trainiertes Genanalyseprogramm zur Verfügung. Allerdings haben MRÁZEK & KARLIN (1999) gezeigt, dass Transposons im Vergleich zu „normalen" Wirtsgenen eine atypische Codon-Nutzung an den Tag legen, so dass Genvorhersagemethoden, die auf Annahmen über die Sequenztupelhäufigkeit beruhen, für Transposons stärker fehlerbehaftet sind als bei anderen Genen. Der TBLASTX-Vergleich besitzt eine Vorhersagekraft, die unabhängig vom Typ des Proteins ist. Es muss lediglich ein homologes Gegenstück existieren. Es sei angemerkt, dass auch das Programm PROCRUSTES Sequenzähnlichkeiten gegen Einträge einer Proteindatenbank zur Vorhersage von Genstrukturen heranzieht (GELFAND ET AL. 1996).

Mit drei verschiedenen Methoden wurde der **Grad der lokalen Konservierung** in alignierten Proteinsequenzen gemessen: Die Methode der relativen Identität, die Entropie-Methode und die Matrix-Methode. Abgesehen davon, dass bei nur zwei alignierten Sequenzen die Methode der relativen Identität und die Entropie-Methode proportionale Zahlenwerte liefern, sind die mit den drei Methoden erhaltenen Werte insgesamt überraschend einheitlich. „Überraschend" deswegen, weil nachvollziehbare Bewertungsunterschiede zwischen den Methoden bestehen: Die Methode der relativen Identität unterscheidet sich von der Entropie- und Matrix-Methode darin, dass eine Beschränkung der beobachteten Symbole an einer Alignmentposition auf eine kleine Teilmenge von Symbolen, z.B. nur hydrophobe Reste oder nur saure, nicht angemessen als Konservierung gewertet wird.

4.2.2 Retrotransposons

Retrotransposons dominieren das Genom von *D. discoideum*. 8,1 % des Genoms fallen Sequenzen dieser Transposonklasse zu, während der Gesamttransposongehalt auf 9,6 % geschätzt wird (Übersicht in Tab. 43). Das gut untersuchte **LTR-Retrotransposon** DIRS-1 trägt mit einem Genomanteil von 3,3 % wesentlich zum strukturellen Aufbau der centromernahen Regionen bei (vgl. Abschnitt 4.3.1). Die beiden homogenen und mitgliederreichen Transposonfamilien der Ty3/Gypsy- und Ty1/Copia-ähnlichen LTR-Retrotransposons besitzen Vertreter in fast allen bisher untersuchten eukaryotischen Genomen (KUMAR & BENNETZEN 1999). Erstere wird in *D. discoideum* durch die Elemente skipper (LENG ET AL. 1998) und DGLT-A (Abschnitt 3.5.1 ab S. 48) vertreten, letztere durch

das Element DCLT-A (Abschnitt 3.5.1). Die neu aufgefundenen Transposonspezies DGLT-A und DCLT-A kommen jeweils mit weniger als 10 Kopien im Genom vor. Da bei vielen Kopien Leserasterverschiebungen in den ORFs beobachtet werden, sind – wenn überhaupt – wenige intakt. Um die Interpretation von Artefakten zu vermeiden, wurde von einer tiefschürfenden Sequenzanalyse abgesehen, zumal die Elemente Transposonfamilien angehören, die bereits durch hunderte wenig diverse Mitglieder repräsentiert werden (KUMAR & BENNETZEN 1999).

Ein charakteristischer und vermutlich wirtstypischer Bestandteil des Transposon-Repertoires von *D. discoideum* sind die **non-LTR-Retrotransposons der TRE-Familie**. Die vorliegende Arbeit hat die Zahl der Mitglieder von drei auf sieben erhöht, und an der vorliegenden Sequenzinformation lässt sich durch Auftragung des lokalen Konservierungsgrads entlang eines Protein-Alignments ein sehr detailliertes Bild über die Verteilung der funktionellen Information im Protein gewinnen. Die grundsätzliche Domänenaufteilung in AP-Endonuklease, RT und Zinkfinger-ähnliches CCHC-Motiv tritt deutlich hervor (vgl. MALIK ET AL. 1999). Zukünftige experimentelle Studien erhalten durch diese Information eine nützliche Datengrundlage über die Topologie funktionell relevanter Eigenschaften des Polyproteins. Wichtige offene Fragen sind der Mechanismus der Integration in den zweiten DNA-Einzelstrang während der späten Integrationsphase und besonders die molekularen Grundlagen zur Vermittlung der Zielortspezifität, die bei den TRE-Elementen und anderen Retrotransposons ausgebildet ist.

Da Retrotransposons virusähnliche Partikel bilden, ist oft diskutiert worden, inwiefern sie sich horizontal, d.h. über Zell- und Artgrenzen hinweg ausbreiten können. Im Falle der non-LTR-Retrotransposons kamen Untersuchungen zum Ergebnis, dass die phylogenetischen Beziehungen sehr eng mit den Entwicklungslinien ihrer Wirte korreliert sind, so dass bei ihnen offenbar keine horizontale Ausbreitung stattgefunden hat (MALIK ET AL. 1999). Im Einklang damit konnte ich zeigen, dass die TRE-Elemente, die ausschließlichen Vertreter der non-LTR-Retrotransposons im Genom von *D. discoideum*, monophyletisch gegenüber allen übrigen non-LTR-Retrotransposons sind (Abb. 26 auf S. 52). Ihr entwicklungsgeschichtlicher Ursprung liegt an der Wurzel der Verzweigung von L1-ähnlichen Elementen (Metazoa & Chlorophyta, MALIK ET AL. 1999) und in Nachbarschaft zur Gruppe von R2/R4 (Arthropoden, BURKE ET AL. 1999). Die Korrelation von Wirts- und Transposonevolution dieser Klasse legt nahe, sie als **phylogenetischen Marker** heranzuziehen. Die Vorteile dieses Markers sind: A) Weite und durchdringende Verbreitung zumindest bei Metazoa und Fungi. B) Der sich aus der Transposition ergebende Kopienüberschuss, der eine Identifizierung bereits durch niedrigredundante Sequenzierung erlaubt und C) Der Konservierungsgrad der reversen Transkriptase, der eine Detektion mittels universeller Hybridisierungssonden und möglicherweise auch PCR-Techniken erlaubt. Gleichermaßen finden sich variable Genabschnitte, die auch kurze Verwandtschaftsdistanzen aufzulösen erlauben.

Die TRE5-Subfamilie weist eine Besonderheit in der strukturellen Organisation auf: **Das repetitive A-Modul** als distalen Teil des 5'-UTR (Abb. 27B auf S. 52). Die Nukleotidsequenz dieses Moduls ist zu 84 % zwischen TRE5-A und TRE5-B konserviert. Da die Entwicklungslinien von TRE5-A und -B sich gemäß der Peptidsequenz ihrer Pol-Proteine bereits vor einiger Zeit voneinander getrennt haben (Abb. 26), ist anzunehmen, dass das A-Modul eine wesentliche funktionelle Bedeutung hat.

Alternativ ist auch ein horizontaler Sequenzaustausch zwischen TRE5-A und -B eine denkbare Erklärung für die bemerkenswerte strukturelle Ähnlichkeit der 5'-Enden. Gleichermaßen muss aber eine funktionelle Bedeutung für die Tandemwiederholungen als selektives Kriterium hinter einem solchen Sequenzaustausch vermutet werden. Eine interessante Frage in Hinblick auf die Funktion des A-Moduls lautet: Wie erhält sich eigentlich die repetitive Struktur der A-Module? Das A-Modul enthält laut vorangehender Studien am TRE5-A den internen Promotor für die Bildung des Transkripts (SCHUMANN ET AL. 1994). Durch die repetitive Natur des Moduls hat jede Elementkopie mehrere potenzielle Transkriptionsstartpunkte, aber nur der erstmögliche Transkriptionsstartpunkt liefert ein Transkript voller Länge. Wenn jeder dieser Transkriptionsstartpunkte gleichwertig ist, müsste das Element nach mehreren Runden der Replikation am 5'-Ende auf eine einzige Kopie des A-Moduls reduziert sein. Wie erhält sich also die repetitive Struktur der A-Module? Möglicherweise sind die TRE5-Elemente in der Lage, ihre 5'-Enden durch neue Einheiten des A-Moduls zu ergänzen. Diese Theorie wird durch die folgenden Argumente gestützt:

- Am Beispiel des Elements R2 wurde gezeigt, dass die reverse Transkriptase von non-LTR-Retrotransposons ungewöhnliche Seitenaktivitäten besitzen kann: Eine Template-unabhängige Polymerase-Aktivität, die eine 3'-Verkürzung des mRNA-Templates durch Synthese einer Zufallssequenz am Zielmolekül ausgleicht (LUAN & EICKBUSH 1995) und die Fähigkeit zum Template-Wechsel während der reversen Transkription (BILBILLO & EICKBUSH 2002). Letzterer Mechanismus könnte durch zyklische Eigenverdrängung des Templates zu repetitiven Sequenzen führen.

- Es wurden beim TRE5-B zwei Varianten der Modulsequenz identifiziert, die zwar charakteristische Abweichungen vom Consensus aufweisen, die Abweichungen sind jedoch in jeder wiederholten Sequenzeinheit exakt gleich.

Ich postuliere daher eine Telomerase-ähnliche Aktivität der Transposons aus der TRE5-Familie. In diesem Zusammenhang ist hervorhebenswert, dass die non-LTR-Retrotransposons TART und TRAS aus *Drosophila* und GilM aus *Giardia lamblia* durch ihr zielortspezifisches Insertionsverhalten mit den Telomeren assoziiert sind (Tab. 45 mit Literaturverweisen). Das TART-Element hat bei *Drosophila* die Funktion der Telomerase, die diesem Organismus fehlt, vollständig übernommen. Zugegebenermaßen sind die repetitiven Einheiten der Telomere bei Eukaryonten nur wenige Basen lang – also wesentlich kürzer als die zur Diskussion stehende repetierte Modulstruktur. Jedoch gibt es im Bakterienreich, z.B. bei Streptomyceten, Telomere mit wiederholten Einheiten von mehr als 100 bp Länge (QUIN & COHEN 1998).

4.2.3 DNA-Transposons

Das neu aufgefundene Transposon **Tdd-5** besitzt gemäß seiner Struktur, speziell der Genorganisation und der abzuleitenden Peptidsequenz seiner Transposase eine deutliche verwandtschaftliche Beziehung zu Tdd-4. Tdd-5 hat einen sehr geringen Genomanteil, der einer Zahl von 6 vollständigen Kopien entspricht. Aktive Kopien dieser Transposonspezies sind wegen des hohen Fragmentierungsindex nicht zu erwarten. Wertvoll erweist sich der Fund dieses Elements, weil es durch den

paarweisen Sequenzvergleich mit Tdd-4 tiefere Einblicke in die strukturelle Organisation dieser Transposonfamilie erlaubt.

Die beiden neu aufgefundenen **thug-Elemente** sind der Sequenz ihrer ITRs zufolge miteinander verwandt. Das Vorhandensein eines ITR ist vorläufig das einzige Argument, sie der Klasse der DNA-Transposons zuzurechnen. Sie sind nicht autonom, da sie keinerlei codierendes Potenzial besitzen. Die distale Sequenz des ITRs lautet auf TGT...ACA, ebenso wie bei den Elementen Tdd-4 und Tdd-5. So kann vermutet werden, dass die Transposase autonomer Vorfahren der thug-Elemente eine ähnliche Zwischenstellung zwischen der Tc1/Mariner-Familie und den Integrasen der LTR-Retrotransposons einnimmt. Dafür spricht auch die nachgewiesene TSD-Länge von 4 oder 5 bp. In der Literatur sind viele DNA-Transposonspezies beschrieben worden, die sich durch Sequenzdeletionen von autonomen Formen ableiten lassen – klassisches Beispiel sind die Ds-Elemente, verkürzte Formen des Activator-Transposons beim Mais (FEDOROFF 1989). Da die Analyse der unautonomen thug-Elemente keine endgültigen Feststellungen über verwandtschaftliche Beziehungen zu anderen Transposongruppen erlaubt, würde es sich lohnen, in verwandten Organismen nach autonomen Formen dieser Familie zu suchen.

Unter den aufgefundenen Transposons befinden sich Elemente eines neuartigen Typs, die **DDT-Elemente** (Abschnitt 3.5.5 ab S. 56). Das Vorhandensein von ITRs und einem intronhaltigen Gen legen nahe, die drei Elemente dieser Familie als DNA-Transposons zu klassifizieren. Hinweise auf die verwandtschaftliche Stellung oder den Transpositionsmechanismus konnten jedoch wegen der fehlenden Sequenzähnlichkeit zu irgendeiner publizierten Proteinsequenz nicht gesammelt werden. Diese Frage nach der phylogenetischen Stellung der DDT-Elemente muss offen bleiben. Durch eine *in-silico*-Sequenzanalyse wurde ein Transkriptionsmodell mit zwei Genen aufgestellt. RT-PCR lieferte einen Nachweis für die linke Gengrenze und die Intronpositionen von Gen 2. Hingegen wurde für Gen 1 kein Transkript gefunden. An der Existenz von Gen 1 besteht jedoch wenig Zweifel, da ein ca. 2400 bp langer ORF vorliegt. So werden die Gene sehr wahrscheinlich separat transkribiert, Gen 1 – wie für eine Transposase zu erwarten – mit einer sehr geringen Rate. Möglicherweise wird es nur in bestimmten Entwicklungsstadien des Wirtes transkribiert, nicht während der vegetativen Entwicklungsphase, aus der die verwendete cDNA stammt.

Die Möglichkeit zur Interpretation von Transposon-Phänomenen allein anhand von genomischer Sequenz ist limitiert, da die Genomsequenz ein weitestgehend statisches Bild vom dynamischen Prozess der Transposition liefert. Es ist zu bedenken, dass Transpositionsereignisse über mehrere Wirtsgenerationen hinweg nur selten vorkommen. In der Natur sind die beobachteten transpositionellen Systeme immer von deutlich ausgeprägten Moderationsmechanismen begleitet (vgl. Abschnitt 4.3.2). Aus diesen Überlegungen heraus ist geplant, die funktionelle Charakterisierung der DDT-Elemente in einem System fortzuführen, das von der natürlichen Situation im Wirt *D. discoideum* unabhängig ist (Abb. 44). Die funktionelle Analyse der Transposons aus der DDT-Familie wird einen Schwerpunkt weiterführender experimenteller Arbeiten bilden. Geplant ist im Einzelnen:

- Eine Absicherung der Genmodelle.
- Die Untersuchung der natürlichen transposablen Aktivität im Genom von *D. discoideum*.

Ansatzpunkt ist hier die Charakterisierung von Transkripten und potenziellen Transkript-Vorlagen, d.h. wenig fragmentierten Transposonkopien.

- Versuche zum *in-vitro*-Nachweis für die Transpositionsaktivität (Abb. 44).

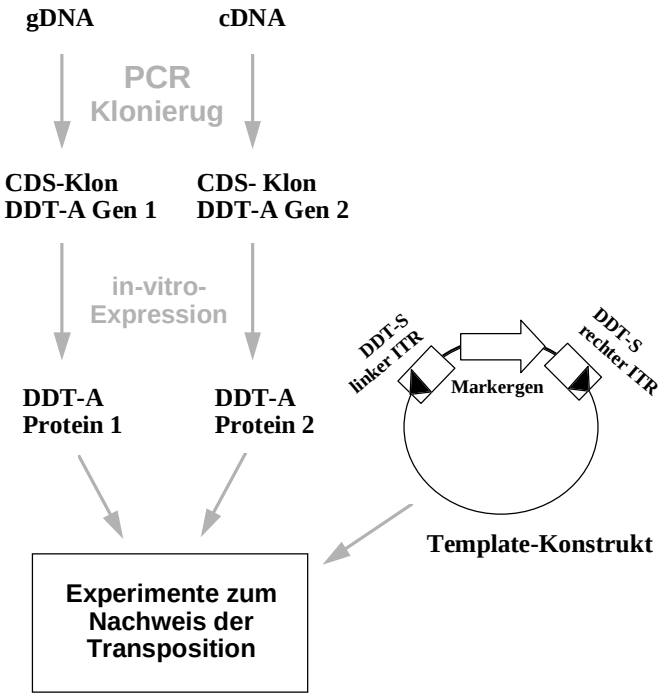


Abb. 44. Experimentplanung zur funktionellen Charakterisierung der Transposons aus der DDT-Familie.

Die DNA-Transposons bilden eine außerordentlich heterogene Transposongruppe, und ihre weite Verbreitung unter Prokaryonten lässt vermuten, dass sie sehr alt sind (MAHILLON & CHANDLER 1998). Die nähere Charakterisierung der DDT-Elemente könnte einen Beitrag leisten, das Verständnis von dieser Transposonklasse weiter voranzubringen.

4.3 Einfluss von Transposons auf die Genomorganisation

4.3.1 Verteilung von Transposonkopien

Viele Retrotransposons inserieren **strikt zielortspezifisch** (Tab. 45). Dabei wurden fast immer Präferenzen für die Flanken von PolIII-transkribierten Genen beobachtet, d.h. tRNA-Gene und Gene für ribosomale Untereinheiten. Auch im Genom von *D. discoideum* tauchen solche insertionsort-spezifischen Transposons auf, und zwar alle Mitglieder der TRE-Familie und das neu beschriebene Gypsy-like LTR-Retrotransposon DGLT-A. Die Elemente der TRE-Familie inserieren zielsicher in die stromaufwärts (Subfamilie TRE5) oder stromabwärts liegende Flanke (Subfamilie TRE3) von tRNAs. Die Untersuchung der Phylogenie der TRE-Elemente zeigt (Abb. 26 auf S. 52), dass die zwei Subfamilien jeweils monophyletische Linien darstellen. Demnach hat im Verlauf der Evolution ein Wechsel oder eine Differenzierung des Insertionsverhaltens der TRE-Elemente stattgefunden.

Mit dem LTR-Retrotransposon Ty3 aus Hefe, das gleichermaßen 4-10 bp stromaufwärts von tRNA-

Genen inseriert, liegt ein wertvolles Modellsystem für die zielspezifische Transposonintegration in tRNA-Flanken vor. *In-vitro*-Studien (KIRCHNER ET AL. 1995, CONNOLLY & SANDMEYER 1997) zeigen eine Beteiligung der Transkriptionsfaktoren TFIIIB und TFIIIC auf (YIEH ET AL. 2000). Diese essentiellen

Tab. 45. Streng spezifische Insertionsorte für Transposons aus *D. discoideum* und anderen Organismen.

Transposon	Transposontyp	Wirt	Insertionsort	Referenz
beta	LTR-Retrotransposon	*C. albicans*	5' von tRNA	PERREAU ET AL 1997
DGLT-A	LTR-Retrotransposon	*D. discoideum*	5' von tRNA	
GilM	non-LTR-Retrotransposon	*Giardia lamblia*	telomere Übergänge	ARKHIPOVA & MORRISON 2001
R1	non-LTR-Retrotransposon	Insekten	28S rRNA	JACUBCZAK ET AL. 1991
R2	non-LTR-Retrotransposon	Arthropoden	28S rRNA	BURKE ET AL. 1999
TART	non-LTR-Retrotransposon	*Drosophila*	telomere Substanz	SHEEN & LEVIS 1994
TRAS	non-LTR-Retrotransposon	Lepidoptera	telomere Repeats	KUBO ET AL. 2001
TRE3	non-LTR-Retrotransposon	*D. discoideum*	3' von tRNA	SZAFRANSKI ET AL. 1999
TRE5	non-LTR-Retrotransposon	*D. discoideum*	5' von tRNA	WINCKLER 1998
Ty1	LTR-Retrotransposon	*S. cerevisiae*	nahe tRNA	VOYTAS & BOEKE 1993
Ty3	LTR-Retrotransposon	*S. cerevisiae*	5' von tRNA	KIRCHNER ET AL. 1995

Bestandteile des aktiven RNA-Polymerase-III-Transkriptionskomplexes stehen durch Protein-DNA-Wechselwirkung in Verbindung mit den Promotormotiven A-Box und B-Box. Neben diesen konstitutiven Bestandteilen des internen tRNA-Gens ist bei *D. discoideum* häufig ein zusätzliches B-Box-Motiv stromabwärts des tRNA-Gens zu finden, genannt exB-Box-Motiv (für „extra B-Box“, vgl. MARSCHALEK & DINGERMANN 1991). Zur Untersuchung der Insertionsortspezifität von TRE-Elementen habe ich für 5'-Enden von verschiedenen TRE3-Transposonkopien die Abstände zu tRNA-Promotormotiven zusammengetragen. Die gefundenen Abstände deuten darauf hin, dass die Insertion in gleicher Weise durch das B-Box- oder exB-Box-Motiv vermittelt werden. Als Folgeschluss wäre der Transkriptionsfaktor TFIIIC ein wichtiger Kandidat für eine Protein-Protein-Wechselwirkung mit dem Insertionskomplex des Transposons als dirigierende Wirkung zu Beginn der Insertion. Angesichts der hohen Schwankungsbreite der B-Box-zu-Transposon-Abstände sind jedoch gewisse Zweifel gerechtfertigt. Ein endgültiges Urteil zu diese Frage sollte sich auf experimentelle Arbeiten stützen. Ein Schritt in diese Richtung ist die erfolgreiche Isolation von TFIIIC aus *D. discoideum* (BUKENBERGER ET AL. 1994).

Über die lange Zeit, in der LTR-Retrotransposon DIRS-1 in *D. discoideum* untersucht worden ist, wurde vielfach festgestellt, dass Kopien von DIRS-1 in Kopien der eigenen Spezies hinein inseriert sind (z.B. CAPPELLO ET AL. 1984). Kartierungsergebnisse weisen darauf hin, dass Cluster von DIRS-1-Kopien fast ausschließlich an jeweils einem Ende aller Chromosomen von *D. discoideum* lokalisiert sind (KUSPA & LOOMIS 1996). Die Anhäufung von DIRS-1 an jeweils einer einzigen eng begrenzten Region der Chromosomen, wahrscheinlich der Centromeren, konnte auch durch ein FISH-Assay an Metaphase-Chromosomen bestätigt werden (R. SUCGANG, persönliche Mitteilung). Andere haben aus

den Beobachtungen geschlossen, dass eine spezifische Gerichtetheit für das Verschachteln der DIRS-1-Kopien sorgt (CAPPELLO ET AL. 1984). Ich konnte in den Schrotschusssequenzen umfangreiches Belegmaterial für einen **allgemeinen Ballungstrend der Kopien aller Transposonspezies** finden – die streng insertionsortspezifischen ausgenommen (Abschnitt 3.6.2 ab S. 62). DIRS-1 besitzt also insofern keine Sonderstellung. Die Centromere sind vermutlich allein deshalb ein stark frequentiertes Insertionsziel, weil diese Region frei von wirtseigenen Genen und transkriptionell wenig aktiv ist. Insofern ist es nicht überraschend, dass im Genom von *Candida albicans* eine vergleichbare Ballung von Transposonkopien in den Centromerregionen gefunden wurde (GOODWIN & POULTER 2000). In einem kompakten Genom wie dem von *D. discoideum* oder *C. albicans* sind jenseits des Centromers offenbar kaum „inerte" Bereiche vorhanden. Nur spezialisierte Transposonformen wie DGLT-A und TRE-Elemente gelangen mittels ihrer zielortspezifischen Insertion in eng begrenzte Areale der genreichen Regionen des Genoms.

Kopien von DIRS-1 habe ich nicht nur als Insertion in Kopien der eigenen Art sondern auch in Kopien anderer Transposonspezies gefunden. Die Schlussfolgerung anderer Autoren auf ein spezifisch begründetes Verschachteln ist demnach auf begrenztes Sequenzmaterial zurückzuführen. Tatsächlich stelle ich aber einen Trend zur Ballung von Transposonkopien der gleichen Art oder Familie fest (Abschnitt 3.7.3 ab S. 65). Dieser Befund betrifft wiederum alle Transposonspezies. Für mehrere Transposonspezies haben KAPITONOV & JURKA (1999) eine ebensolche ungleichmäßige Verteilung auf den Chromosomen von *Arabidopsis thaliana* beschrieben. Eine mögliche Erklärung ist, dass die Transposition einer Spezies in zeitlichen Schüben erfolgt, zu einem gegebenen Zeitpunkt aber nur eng begrenzte genomische Areale als Landefläche zur Verfügung stehen.

4.3.2 Dynamik des Transposonbestands

Die verschiedenen Transposonspezies im Genom von *D. discoideum* zeigen deutliche Unterschiede hinsichtlich des Kopienumfangs und des Fragmentierungsindex. Retrotransposons sind mit einem Genomanteil von 8,1 % viel zahlreicher und weniger fragmentiert (mittlerer Fragmentierungsindex von 1,4 – gewichtet nach dem Genomanteil) als DNA-Transposons mit einem Genomanteil von 1,5 % und einem Fragmentierungsindex von 2,7. Ähnliche Verhältnisse für die Genomanteile nach Transposonklassen liegen bei Wirbeltieren vor (INT. HUM. GENOME SEQ. CONS. 2001). Die Unterschiede in der Ausbreitung lassen darauf schließen, dass die Populationsdynamik der Transposonspezies je nach ihrer Klassenzugehörigkeit sehr verschieden ist.

Ich werte wiederkehrende interne Deletionen bei Transposonkopien, die mit einem Verlust essentieller codierender Regionen und damit auch der Autonomie einher gehen, als Zeichen für einen Moderationseffekt. Ein weiterer Hinweis auf einen solchen Moderationseffekt ist beispielsweise die Transkription des LTR-Retrotransposons DIRS-1, die über einen Hitzeschock-Promotor gesteuert wird (COHEN ET AL. 1984). Sie unterliegt damit Umwelteinflüssen, die auf die Wirtszelle einwirken und durch die ihr eigenen Signaltransduktionssysteme mitgeteilt werden – ohne unmittelbaren Vorteil für das Transposon. Anbetracht dieser Tendenzen zur Moderation ist eine Modellvorstellung, die Transposons als „selfish DNA" beschreibt, nicht haltbar (HICKEY 1982, FLAVELL 1995).

Der Begriff der Moderation spielt eine entscheidende Rolle für ein Modell, das ich für die Dynamik

von Transposonpopulationen vorschlage. Dazu ist scharf zwischen Retrotransposons und DNA-Transposons zu unterscheiden. Bei den Retrotransposons geht die Transposition direkt mit der Vermehrung der Kopien einher. Da bei der Transposition – zumindest tendenziell – das Transkript und die Proteine, die von demselben Transkript abstammen, funktionell aneinander gekoppelt sind, selektiert der Vermehrungsmechanismus auf die Fitness zur Transposition. Durch eine steigende Transpositionsrate, die sich proportional zur Fitness und Kopienzahl verhält, kommt es leicht zu einem „Lawinenereignis“ von Transpositionen. Als entgegen wirkende Kraft werden unmoderierte (sehr fitte) Transposons dadurch eliminiert, dass infolge unkontrollierter Transposition der Wirt und damit auch die Transposons zugrunde gehen. Anders sind die Verhältnisse bei den DNA-Transposons, deren Transposition nicht direkt mit einer Vermehrung einhergeht. Die Vermehrung kann durch anschließende Rekombination in einer diploiden (oder polyploiden) Zelle, während der DNA-Replikation oder durch Reparaturmechnismen erfolgen. Ein Teil der Mechanismen kann aber auch zu einer Verringerung der Kopienzahl führen. Somit besteht kein ausgeprägter Trend zur Veränderung der Kopienzahl in Abhängigkeit von der Transpositionsrate. Das Überleben der DNA-Transposons hängt damit wesentlich von ihrem Beitrag ab, den sie zur Fitness des Wirts beitragen.

Das Modell erklärt verschiedene Beobachtungen an den Transposonpopulationen von *D. discoideum*:

- Die Kopienzahl für Retrotransposons ist höher als die von DNA-Transposons, weil nur beim replikativen Transpositionsmechanismus der Retrotransposons die Vermehrungsfähigkeit als positives Selektionskriterium wirken kann.
- Eine Vielzahl von Retrotransposons besitzt eine strikte Zielortspezifität bei der Insertionsreaktion (Beispiele in Tab. 45). Für DNA-Transposons ist eine solche Spezifität nicht bekannt. Retrotransposons bergen die Gefahr, sich „lawinenartig“ im Genom auszubreiten – das Risiko eines Schadens für den Wirt ist hoch, die Überlebenschance der Transposons hängt aber gleichsam vom Überleben des Wirts ab. Infolge dieser Abhängigkeit ist der Selektionsdruck für Mechanismen, die zur Vermeidung von wirtsschädigenden Mutationen dienen, groß. Es ist daher kein Zufall, dass sich ein solcher Mechanismus in Form einer Zielortspezifität, die Transposoninsertionen zu inerten Regionen des Genoms dirigiert, bei verschiedenen Retrotransposons entwickelt hat. Die Tatsache, dass sich eine Zielortspezifität mehrfach unabhängig in verschiedenartiger Form in einzelnen Transposonspezies entwickelt hat, ist ein Nachweis für den Selektionsdruck.
- Für DNA-Transposons besteht ein ausgesprochener Trend hin zum Verlust der Autonomie eines Teils der Transposonkopien. Der Beginn dieses Prozesses lässt sich am Tdd-4 beobachten, das zu ca. 25 % in der unautonomen Form Tdd-4d mit einer Deletion des aktiven Zentrums der Transposasedomäne vorkommt (WELLS 1999, Abb. 22 auf S. 47). Endpunkt dieser Entwicklung ist eine Trennung in unmobile Transposase und passiv mobile Transposons, beobachtet in hoch-fragmentierten, aber codierenden Elementen DDT-A und DDT-B gegenüber unautonomem DDT-S. Der selektionswirksame Vorteil einer solchen Populationsentwicklung ist, dass die Kopienzahl der Transposons zunehmen kann – mit dem positiven Aspekt des Gewinns potenzieller Rekombinationspunkte im Genom –, ohne dass die Transpositionsaktivität proportional zunimmt – das Risiko von potenziell nachteiligen Transpositionsereignissen bleibt

konstant.

4.4 Transposons im Kontext der Genomassemblierung

Transposonkopien stellen die größte Herausforderung bei der Assemblierung von Genomen dar. Auf dieser Einsicht fußt die vorliegende Arbeit. Um dem Problem eine Kontur zu geben, wurde der genomische Anteil der verschiedenen Transposonspezies von *D. discoideum* bestimmt und darüber hinaus die Sequenzdiversität jeder einzelnen Transposonspezies ermittelt (Abschnitt 3.7.2 ab S. 64). Herangezogen habe ich die Sequenzdiversität π, die durch NEI & LI (1979) als Maß zur Beschreibung der Variabilität unter sehr ähnlichen Sequenzen eingeführt wurde, sei es im Falle von Populationsgenetik an variablen Genloci oder wie im hier betrachteten Fall der Sequenzvariation unter Instanzen von Multi-Copy-DNA. Zusätzlich habe ich das Maß R_A eingeführt, um auch die Kopienzahl in einer Abschätzung der Assemblierbarkeit quantitativ zu berücksichtigen (Abschnitte 2.5.3 und 2.5.4 ab S. 27). Mit Hilfe dieser Maße wurde die Betrachtung der Assemblierbarkeit von Transposonkopien unabhängig von konkreten Verfahren zur Assemblierung durchgeführt.

Wie sind nun die Werte in Hinblick auf die Frage nach der tatsächlichen Assemblierbarkeit zu interpretieren? Unter günstigsten Voraussetzungen gilt, dass sich alle Transposonspezies assemblieren lassen, deren Wert R_A kleiner ist als die Länge der Schrotschusssequenzen – also ca. 350 bp oder sogar 2×350 bp bei Heranziehen der Sequenzpaare, die von einem Klon stammen. Auf dieser Sequenzstrecke sollten sich im statistischen Mittel genügend polymorphe Ausprägungen ansammeln, um eine eindeutige Zuordnung der Schrotschusssequenz zu einer Transposonkopie zu erlauben. Mehrere günstige Voraussetzungen gelten jedoch in der vorliegenden realen Situation nicht.

- Die Gleichverteilung polymorpher Positionen über die Transposonsequenz ist nur annähernd gegeben (Ergebnisse nicht dargestellt). Die vollständige, fehlerfreie Assemblierung von Transposonkopien wird nach dem Bottle-Neck-Prinzip gerade an solchen Stellen schwierig sein, an denen polymorphe Sequenzausprägungen selten sind. Eine Abschätzung der Erfolgsaussichten sollte also eher vom ungünstigeren Fall als vom günstigsten Fall ausgehen. Erschwerend ist, dass adäquate Verteilungsfunktionen für die Anhäufung oder Ausdünnung von polymorphen Merkmalen aus den Schrotschusssequenzen nicht abgeleitet werden können.

- Die Entscheidung, ob zwei Schrotschusssequenzen in ihren polymorphen Sequenzausprägungen voneinander verschieden sind, beruht auf einem paarweisen Vergleich. Die Sequenzstrecke, über die ein paarweiser Vergleich möglich ist, ist maximal so lang wie die kürzere der beiden Schrotschusssequenzen, die miteinander verglichen werden, eventuell betrifft die Überlappung aber jeweils nur Teile beider Sequenzen. Versucht man, die Länge der Überlappung statistisch zu fassen, so geht die Coverage des Genoms mit Schrotschusssequenzen wesentlich in die Berechnung ein. Und auch hier – wie auch beim vorigen Punkt – ist bei der Betrachtung von vollständigen Transposonkopien, deren Länge ein Vielfaches der Schrotschusssequenzen beträgt, von einem Worst-Case-Szenario an bestimmten Stellen der Transposonloci auszugehen (vgl. Abb. 22 auf S. 47).

- Die Kopplung polymorpher Sequenzausprägungen wirkt sich negativ auf die Prognose zur Unterscheidbarkeit von Transposonkopien aus. Mit steigendem Maß an Kopplung nimmt der diagnostische Wert polymorpher Sequenzausprägungen ab. Die Extrapolation der Maße π und R_A auf vollständige Transposonloci legt ein Diversitätsmodell zugrunde, bei dem jede polymorphe Ausprägung ungekoppelt mit benachbarten polymorphen Ausprägungen ist. Wahrscheinlich liegt aber eine gewisse Kopplung von polymorphen Merkmalen vor. Diese Annahme ergibt sich aus der anzunehmenden dichotomen Entwicklungsgeschichte der verschiedenen Transposonkopien. Eine neue Transposonkopie ist zum Zeitpunkt der Entstehung zwangsläufig sequenzidentisch zu einer anderen Transposonkopie, von der sie sich ableitet. Nur die nachfolgend in die Transposonkopien eingeführten Mutationen sind von den übrigen polymorphen Merkmalen entkoppelt. Entkopplung kann außerdem erfolgen durch homologe Rekombination zwischen zwei verschiedenen Transposonkopien.

Es existieren auch Effekte, die eine Unterschätzung der Unterscheidbarkeit von Transposonkopien bedingen:

- Transposonspezies mit einer hohen Dichte polymorpher Sequenzmerkmale zeigen auch einen erhöhten Grad von Kopienfragmentierung (vgl. Tab. 40 mit Tab. 43). Die Fragmentierungspunkte können als zusätzliches sehr spezifisches diagnostisches Kriterium gewertet werden, das in die Berechnung der Maße π und R_A nicht eingeht.

- Es besteht für drei der Transposonspezies aus *D. discoideum* allein aufgrund der Kürze ihrer Sequenz eine gute Prognose für ihre Assemblierbarkeit, denn viele der beidseitig sequenzierten Schrotschussklone sollten vollständige Kopien dieser Transposons überspannen können. Dies gilt für die Elemente DDT-S (Consensuslänge 758 bp), thug-S (2192 bp), thug-T (1132 bp) und TRE3-D und TRE5-C, die jeweils fast ausschließlich als verkürzte Kopien im Genom vorkommen.

Die Schrotschusstechnik zur Genomsequenzierung gewinnt immer mehr an Bedeutung. Daher ist der Entwicklungsdruck in diesem Zweig der Bioinformatik groß, und es gibt eine Reihe von algorithmischen Lösungsvorschlägen für die Assemblierung von repetitiven Sequenzen (ANSON & MYERS 1999, PEVZNER ET AL. 2001). Sie sind Ausdruck eines kontinuierlichen Fortschritts in der Erkennung und Behandlung von repetitiven Anteilen unter Schrotschusssequenzen. Das bedeutet aber nicht, dass eine eingehende Analyse des Gehalts von repetitiven Elementen eines Genoms in Zukunft überflüssig wird, vielmehr verschiebt sich die Grenze zwischen lösbaren und unlösbaren Assemblierungsproblemen. Die tatsächliche Lage der Grenze muss in jedem Anwendungsfall neu ermittelt werden. Untersuchungen zur Leistungsfähigkeit von Assemblierungsprogrammen wurden stets an beispielhaften Genomen durchgeführt und die Assemblierungslücken und -fehler ausgezählt (PEVZNER ET AL. 2001). Ziel sollte jedoch sein, praktikable Maße zu entwickeln, mit denen Assemblierungsschwierigkeiten bewertet werden können. Mittels dieser Maße könnte eine objektive und quantitative Bewertung von Assemblierungsproblemen erfolgen. Ich habe im Verlauf dieser Arbeit versucht, solche Maße zu entwerfen. Anleihen von Methoden aus dem Forschungsfeld der Populationsgenetik waren dabei

hilfreich aber nicht hinreichend. Die Anstrengungen sollten weitergeführt werden.

In Hinblick auf das Sequenzierungsprojekt für das Genom von *D. discoideum* konnte eine wichtige Grundlage zur Assemblierung des Genoms aus Schrotschusssequenzen geschaffen werden. Was aber geschieht mit den ungelösten Assemblierungsproblemen? Im Falle der Genomsequenzierung von *D. discoideum* liegt eine besondere Schwierigkeit darin, dass bakterielle Klone mit großen Inserts nicht verfügbar sind. Dadurch fehlen beispielsweise BAC-Endsequenzen, die bei der Überbrückung von Transposonkopien eine wichtige Rolle spielen könnten (ANSON & MYERS 1999). YAC- und cYAC-Klone sind verfügbar, haben aber eine hohe Quote an Chimären (KONFORTOV ET AL. 2000). Solange diese Klonressourcen zur sequenzspezifischen Auflösung oder zur Überbrückung von Transposonloci fehlen, können etliche Transposonkopien nicht aufgelöst werden (Abschnitt 3.7.4 ab S. 66). In einigen Einzelfällen können PCR-Experimente helfen, um Abschnitte kartierter Lücken zu amplifizieren und zu sequenzieren. Allerdings wird gerade während der Endphase des Genomprojekts abgewogen werden, ob der ökonomische Aufwand eines kompromisslosen Lückenschließens in einem akzeptablen Verhältnis zum Informationsgewinn steht.

5 LITERATUR

Abril, J.F., R. Guigó. gff2ps: visualizing genomic annotations. Bioinformatics 16, 743-744 (2000).

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

Altschul, S.F., E.V. Koonin. Iterated profile searches with psi-BLAST – a tool for discovery in protein databases. Trends Biochem. Sci. 23, 444-447 (1998).

Anson, E., E.W. Myers. Algorithms for whole genome shotgun sequencing. RECOMB 1-7 (1999).

Arkhipova, I.R., H.G. Morrison. Three retrotransposon families in the genome of *Giardia lamblia*: Two telomeric, one dead. Proc. Natl. Acad. Sci. USA 98, 14497-14502 (2001).

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, E.L.L. Sonnhammer. The Pfam protein families database. Nucl. Acids Res. 30, 276-280 (2002).

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, D.L. Wheeler. GenBank. Nucl. Acids Res. 28, 15-18 (2000). Online unter http:// www.ncbi.nih.gov.

Bessereau, J.L., A. Wright, D.C. Williams, K. Schuske, M.W. Davis, E.M. Jorgensen. Mobilization of a *Drosophila* transposon in the *Caenorhabditis elegans* germ line. Nature 413, 70-74 (2001).

Bilbillo, A., T.H. Eickbush. The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. J. Mol. Biol. 316, 459-473 (2002).

Bonfield, J.K., K.F. Smith, R. Staden. A new DNA sequence assembly program. Nucl. Acids Res. 23, 4992-4999 (1995). Online unter www.mrc-lab.cam.ac.uk/pubseq.

Bukenberger, M., T. Dingermann, W. Meissner, K.H. Seifart, T. Winckler. Isolation of transcription factor IIIC from *Dictyostelium discoideum*. Eur. J. Biochem. 220, 839-846 (1994).

Burke, W.D., H.S. Malik, J.P. Jones, T.H. Eickbush. Conserved structure and mechanism of integration of the R2 retrotransposable element in all arthropods. Mol. Biol. Evol. 16, 502-511 (1999).

Calvi, B.R., T.J. Hong, S.D. Findley, W.M. Gelbert. Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants: hobo, Activator, and Tam3. Cell 66, 465-471 (1991).

Cappello, J., S.M. Cohen, H.F. Lodish. *Dictyostelium* transposable element DIRS-1 preferentially inserts into DIRS-1 sequences. Mol. Cell. Biol. 4, 2207-2213 (1984).

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

Chastonay, Y. de, Chastonay, Y. de, H. Felder, C. Link, P. Aeby, H. Tobler, F. Müller. Unusual

features of the retroid element PAT from the nematode *Panagrellus redivivus*. Nucl. Acids Res. 20, 1623-1628 (1992).

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

Connolly, C., S. Sandmeyer. RNA polymerase III interferes with Ty3 integration. FEBS Lett. 405, 305-311 (1997).

Corpet, F., F. Servant, J. Gouzy, D. Kahn. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucl. Acids Res. 28, 267-269 (2000).

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

Craig, N.L. Unity in transposition reactions. Science 270, 253-254 (1995).

Crollius, H.R., O. Jaillon, C. Dasilva, C. Ozoul-Costaz, C. Fizames, C. Fischer, L. Bouneau, A. Billaut, F. Quetier, W. Saurin, A. Bernot, J.Weissenbach. Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. Genome Res. 10, 939-949 (2000).

Dang, V.-D., H.L. Levin. Nuclear import of the retrotransposon Tf1 is governed by a unique nuclear localization signal that possesses a unique requirement for the FXFG nuclear pore factor Nup124p. Mol. Cell. Biol. 20, 7798-7812 (2000).

Davies, D.R., I.Y. Goryshin, W.S. Reznikoff, I. Rayment. Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. Science 289, 77-84 (2000).

Dayhoff, M.O., R.M. Schwartz, B.C. Orcutt. A model of evolutionary change in proteins. In: Dayhoff, M.O. (ed.). Atlas of protein sequence and structure. Natl. Biomed. Res. Found. Washington, DC. 1978, pp. 345-352.

Deininger, P.L., G.R. Daniels. The recent evolution of mammalian repetitive DNA elements. Trends Genet. 2, 76-80 (1986).

Doak, T.G., F.P. Doerder, C.L. Jahn, G. Herrick. A proposed superfamily of transposase genes: Transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA 91, 942-946 (1994).

Durbin, R., D. Haussler. GFF (General Feature Format) Specifications Document. Online unter http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml.

Eickbush, T.H. Transposing without ends: the non-LTR retrotransposable elements. New Biologist 4, 430-440 (1992).

Fedoroff, N.V. About maize transposable elements and development. Cell 56, 181-191 (1989).

Felsenstein, J. The PHYLIP package. http://evolution.genetics.washington.edu/phylip.html

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

Flavell, A.J. Retroelements, reverse transcriptase and evolution. Comp. Biochem. Physiol. B. Biochem. Mol. Biol. 110, 3-15 (1995).

Gelfand, M.S., A.A. Mironov, P.A. Pevzner. Gene recognition via spliced sequence alignment. Proc. Natl. Sci. USA 93, 9061-9066 (1996).

Glöckner, G. Large scale sequencing and analysis of AT rich eukaryote genomes. Curr. Genomics. 1, 289-299 (2000).

Glöckner, G., L. Eichinger, K. Szafranski, J.A. Pachebat, A.T. Bankier, P.H. Dear, R. Lehmann, C. Baumgart, G. Parra, J.F. Abril, R. Guigó, K. Kumpf, B. Tunggal, the Dictyostelium Genome Sequencing Consortium, E. Cox, Q.M. Quail, M. Platzer, A. Rosenthal, A.A. Noegel. Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. Nature (2002), im Druck.

Götte, M., X. Li, M.A. Wainberg. HIV-1 reverse transcription: a brief overview focussed on structure-function relationships among molecules involved in initiation of the reaction. Arch. Biochem. Biophys. 365, 199-210 (1999).

Goodwin, T.J.D., R.T.M. Poulter. Multiple LTR-retrotransposon families in the asexual yeast *Candida albicans*. Genome Res. 10, 174-191 (2000).

Haapa, S., S. Suomalainen, S. Eerikainen, M. Airaksinen, L. Paulin, H. Savilahti. An efficient DNA sequencing strategy based on the bacteriophage mu *in vitro* DNA transposition reaction. Genome Res. 9, 308-315 (1999).

Haring, E., S. Hagemann, W. Pinsker. Ancient and recent horizontal invasions of drosophilids by P elements. J. Mol. Evol. 51, 577-586 (2000).

Hickey, D.A. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101, 519-531 (1982).

Hofmann, K., P. Bucher, L. Falquet, A. Bairoch. The PROSITE database, its status in 1999. Nucl. Acids Res. 27, 215-219 (1999).

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).

Jacks, T., H.D. Madhani, F.R. Masiarz, H.E. Varmus. Signals for ribosomal frameshifting in the Rous Sarcoma Virus gag-pol region. Cell 55, 447-458 (1988).

Jacubczak, J.L., W.D. Burke, T.H. Eickbush. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. Proc. Natl. Acad. Sci. USA 88, 3295-3299 (1991).

Jurka, J., A. Milosavljevic. Reconstruction and analysis of human Alu genes. J. Mol. Evol. 32, 105-121 (1991).

Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. 94, 1872-1877 (1997).

Kapitonov, V.V., J. Jurka. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica 107, 27-37 (1999).

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

Kempken, F., U. Kück. restless, an active Ac-like transposon from the fungus *Tolypocladium inflatum*: structure, expression and alternative RNA splicing. Mol. Cell. Biol. 16, 6563-6572 (1996).

Kim, J.M., S. Vanguri, J.D. Boeke, A. Gabriel, D.F. Voytas. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. 8, 464-478 (1998).

Kirchner, J., C. Conolly, S. Sandmeyer. Requirement of RNA polymerase III transcription factors for *in vitro* position-specific integration of a retrovirus-like element. Science 267, 1488-1491 (1995).

Koga, A., A. Shimada, M. Sakaiyumi, H. Tachida, H. Hori. Evidence for recent invasion of the Medaka fish genome by the Tol2 transposable element. Genetics 155, 273-281 (2000).

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

Kubo, Y., S. Okazaki, T. Anzai, H. Fujiwara. Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. Mol. Biol. Evol. 18, 848-857 (2001).

Kumar, A., J.L. Bennetzen. Plant retrotransposons. Ann. Rev. Genet. 33, 479-532 (1999).

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

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

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

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

Lowe, T.M., S.R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 25, 955-964 (1997).

Luan, D.D., T.H. Eickbush. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. Mol. Cell. Biol. 15, 3882-3891 (1995).

Mahillon, J., M. Chandler. Insertion sequences. Microbiol. Mol. Biol. Rev. 62, 725-774 (1998).

Malik, H.S., W.D. Burke, T.H. Eickbush. The age and evolution of non-LTR retrotransposable elements. Mol. Biol. Evol. 16, 793-805 (1999).

Marschalek, R., T. Dingermann. Structure, organization and function of transfer RNA genes from the cellular slime mold *Dictyostelium discoideum*. Adv. Gene Tech. 2, 103-143 (1991).

Marschalek, R., J. Hofmann, G. Schumann, R. Gösseringer, T. Dingermann. Structure of DRE, a retrotransposable element which integrates with position specificity upstream of *Dictyostelium discoideum* tRNA genes. Mol. Cell. Biol. 12, 229-239 (1992a).

Marschalek, R., J. Hofmann, G. Schumann, T. Dingermann. Two distinct subforms of the retro-transposable DRE element in NC4 strains of *Dictyostelium discoideum*. Nucl. Acids Res. 20, 6247-6252 (1992b).

Marschalek, R., J. Hofmann, G. Schumann, M. Bach, T. Dingermann. Different organization of the transfer RNA-gene associated repetitive element, DRE, in NC4-derived strains and in other wild-type *Dictyostelium discoideum* strains. Eur. J. Biochem. 217, 627-631 (1993).

McClintock, B. Chromosome organization and genic expression. Cold Spring Harbor Symp. Quant. Biol. 16, 13-47 (1951).

McLean, C., A. Bucheton, D. Finnegan. The 5' untranslated region of the I factor, a long interspersed nuclear element-like retrotransposon of *Drosophila melanogaster*, contains an internal promoter and sequences that regulate expression. Mol. Cell. Biol. 13, 1042-1050 (1993).

McPherson, C.E., C.K. Singleton. Nutrient-responsive promoter elements of the V4 gene of

Dictyostelium discoideum. J. Mol. Biol. 232, 386-396 (1993).

Morio, T., H. Urushihara, T. Saito, Y. Ugawa, H. Mizuno et al. The *Dictyostelium* developmental cDNA project: Generation and analysis of expressed sequence tags from the first-finger stage of development. DNA Res. 5, 335-340 (1998).

Mrázek, J., S. Karlin. Detecting alien genes in bacterial genomes. Ann. N.Y. Acad. Sci. 870, 314-329 (1999).

Neef, B.D., M.R. Gross. Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evol. Int. J. Org. Evol. 55, 1717-1733 (2001).

Nei, M., W.-H. Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76, 5269-5273 (1979).

Noegel, A.A., M. Schleicher. The actin cytoskeleton of *Dictyostelium*: A story told by mutants. J. Cell. Sci. 113, 759-766 (2000).

Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Biochem. J. 15, 281-286 (1985).

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

Parra, G., K. Szafranski, R. Guigó. Gene structure, splice signalling and gene prediction in the *Dictyostelium discoideum* genome. Unveröffentlicht, 2001. Ressourcen unter http://www1.imim.es/~gparra/dicty/, http://genome.imb-jena.de/dictyostelium/strategy.html.

Perreau, V.M., M.A.S. Santos, M.F. Tuite. beta, a novel repetitive DNA element associated with tRNA genes in the pathogenic yeast *Candida albicans*. Mol. Microbiol. 25, 229-236 (1997).

Pevzner, P.A., H. Tang, M.S. Waterman. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA 98, 9748-9753 (2001).

Phadnis, S.H., H.V. Huang, D.E. Berg. Tn5supF, a 264-base-pair transposon derived from Tn5 for insertional mutagenesis and sequencing DNAs cloned in phage lambda. Proc. Natl. Acad. Sci. USA 86, 5908-5912 (1989).

Poole, S.J., R.A. Firtel. Genomic instability and mobile genetic elements in regions surrounding two discoidin I genes of *Dictyostelium discoideum*. Mol. Cell. Biol. 4, 671-680 (1984).

Quin, Z., S.N. Cohen. Replication at the telomeres of the *Streptomyces* linear plasmid pSLA2. Mol. Microbiol. 28, 893-903 (1998).

Roberts, R.J, D. Macelis. REBASE – restriction enzymes and methylases. Nucl. Acids Res. 29, 268-269 (2001). Online unter http://rebase.neb.com.

Rogge, H., H.J. Risse. A procedure for the isolation of *Dictyostelium* nuclei. H.-S. Z. Physiol. Chem. 355, 1467-1470 (1974).

Romans, P., R.A. Firtel. Organization of the actin multigene family of *Dictyostelium discoideum* and analysis of variability in the protein coding regions. J. Mol. Biol. 186, 321-335 (1985).

Rosen, E., A. Sivertsen, R.A. Firtel. An unusual transposon encoding heat shock inducible and developmentally regulated transcripts in *Dictyostelium*. Cell 35, 243-251 (1983).

Schumann, G., I. Zündorf, R. Marschalek, J. Hofmann, T. Dingermann. Internally located and oppositely oriented polymerase II promoters direct convergent transcription of a LINE-like

retroelement, the *Dictyostelium* repetitive element, from *Dictyostelium discoideum*. Mol. Cell. Biol. 14, 3074-3084 (1994).

Sheen, F., R. Levis. Transposition of the LINE-like retrotransposon TART to *Drosophila* chromosome termini. Proc. Natl. Acad. Sci. USA 91, 12510-12514 (1994).

Sibly, R.M., J.C. Whittaker, M. Talbot. A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Mol. Biol. Evol. 18, 413-417 (2001).

Smit, A.F.A. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6, 743-748 (1996).

Szafranski, K., G. Glöckner, T. Dingermann, K. Dannat, A.A. Noegel, L. Eichinger, A. Rosenthal, T. Winckler. Non-LTR retrotransposons with unique integration preferences downstream of *Dictyostelium discoideum* tRNA genes. Mol. Gen. Genet. 262, 772-780 (1999).

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

Tsugeki, R., M.L. Olson, N.V. Fedoroff. Transposon tagging and the study of root development in *Arabidopsis*. Gravit Space Biol. Bull. 11, 79-87 (1998).

Voytas, D., J. Boeke. Yeast retrotransposons and tRNAs. Trends in Genet. 9, 421-427 (1993).

Wall, L., T. Christiansen, R.L. Schwartz. Programming Perl. O'Reilly Sebastopol 1996(2). Online unter www.perl.org oder www.perl.com.

Wells, D.J. Tdd-4, a DNA transposon of *Dictyostelium* that encodes proteins similar to LTR retroelement integrases. Nucl. Acids Res. 27, 2408-2415 (1999).

Winckler, T, C. Tschepke, E. de Hostos, A. Jendretzke, T. Dingermann. Tdd-3, a tRNA gene-associated poly(A) retrotransposon from *Dictyostelium discoideum*. Mol. Gen. Genet. 257, 655-661 (1998).

Winckler, T. Retrotransposable elements in the *Dictyostelium discoideum* genome. Cell. Mol. Life Sci. 54, 383-393 (1998).

Xiong, Y., T.H. Eickbush. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 9, 3353-3362 (1990).

Yieh, L., G. Kassavetis, E.P. Geiduschek, S.B. Sandmeyer. The Brf and TATA-binding protein subunits of the RNA polymerase III transcription factor IIIB mediate position-specific integration of the gypsy-like element, Ty3. J. Biol. Chem. 275, 29800-29807 (2000).

Zuker, C., J. Cappello, R.L. Chisholm, H.F. Lodish. A repetitive *Dictyostelium* gene family that is induced during differentiation and by heat shock. Cell 34, 997-1005 (1983).

6 ANHANG

6.1 Binomialtabellen

Die Werte der folgenden Tabellen spiegeln die Werte der kumulativen Funktion $F(x \mid n,p)$ zu einer Binomialverteilung wider.

6.1.1 Vertrauensintervall bei der Schätzung der Genomgröße

Binomialtabellen für n=26080, $96 \leq x \leq 118$ und verschiedene Werte von p im Bereich $0.00390 \leq p \leq 0.00438$.

	p								
x	**0,00390**	**0,00396**	**0,00402**	**0,00408**	**0,00414**	**0,00420**	**0,00426**	**0,00432**	**0,00438**
...	...	...	...	...	...	...	...	...	...
96	0,30653	0,25490	0,20867	0,16816	0,13341	0,10420	0,08014	0,06069	0,04528
97	0,34281	0,28825	0,23865	0,19453	0,15612	0,12336	0,09599	0,07356	0,05552
98	0,38046	0,32340	0,27073	0,22317	0,18115	0,14480	0,11397	0,08835	0,06746
99	0,41916	0,36008	0,30471	0,25396	0,20846	0,16852	0,13416	0,10520	0,08126
100	0,45852	0,39796	0,34034	0,28674	0,23796	0,19451	0,15660	0,12419	0,09702
101	0,49816	0,43671	0,37734	0,32128	0,26950	0,22271	0,18130	0,14539	0,11486
102	0,53769	0,47594	0,41537	0,35732	0,30290	0,25301	0,20821	0,16881	0,13485
103	0,57672	0,51528	0,45409	0,39455	0,33792	0,28523	0,23725	0,19444	0,15703
104	0,61490	0,55435	0,49312	0,43265	0,37429	0,31919	0,26828	0,22222	0,18140
105	0,65188	0,59277	0,53210	0,47127	0,41169	0,35461	0,30112	0,25204	0,20792
106	0,68735	0,63021	0,57065	0,51004	0,44978	0,39122	0,33555	0,28374	0,23651
107	**0,72107**	**0,66634**	0,60842	0,54859	0,48823	0,42871	0,37130	0,31712	0,26705
108	0,75282	0,70088	0,64508	0,58657	0,52667	0,46673	0,40809	**0,35196**	**0,29935**
109	0,78243	0,73361	0,68034	0,62365	0,56474	0,50494	0,44560	0,38797	0,33322
110	0,80981	0,76432	0,71394	0,65951	0,60212	0,54300	0,48348	0,42487	0,36839
111	0,83489	0,79290	0,74567	0,69388	0,63846	0,58055	0,52139	0,46232	0,40459
112	0,85766	0,81923	0,77537	0,72654	0,67350	0,61727	0,55901	0,49999	0,44152
113	0,87814	0,84330	0,80291	0,75727	0,70698	0,65286	0,59598	0,53756	0,47885
114	0,89641	0,86509	0,82823	0,78596	0,73867	0,68705	0,63202	0,57468	0,51626
115	0,91256	0,88466	0,85131	0,81249	0,76842	0,71962	0,66683	0,61105	0,55342
116	0,92672	0,90207	0,87216	0,83682	0,79611	0,75036	0,70016	0,64637	0,59001
117	0,93901	0,91743	0,89084	0,85894	0,82165	0,77913	0,73181	0,68038	0,62574
118	0,94961	0,93086	0,90742	0,87888	0,84501	0,80583	0,76160	0,71284	0,66032
...	...	...	...	...	...	...	...	...	...

	p								
x	**0,00390**	**0,00392**	**0,00394**	**0,00398**	**...**	**0,00432**	**0,00434**	**0,00436**	**0,00438**
...	...	...	...	...	...	...	...	...	...
96	0.30653	0.28875	0.27154	0.25490	...	0.06069	0.05514	0.05001	0.04528
97	0.34281	0.32412	0.30592	0.28825	...	0.07356	0.06708	0.06108	0.05552
98	0.38046	0.36101	0.34198	0.32340	...	0.08835	0.08089	0.07394	0.06746
99	0.41916	0.39912	0.37941	0.36008	...	0.10520	0.09669	0.08871	0.08126
100	0.45852	0.43808	0.41788	0.39796	...	0.12419	0.11457	0.10552	0.09702
101	0.49816	0.47753	0.45703	0.43671	...	0.14539	0.13463	0.12445	0.11486
102	0.53769	0.51706	0.49646	0.47594	...	0.16881	0.15689	0.14557	0.13485
103	0.57672	0.55631	0.53581	0.51528	...	0.19444	0.18137	0.16890	0.15703
104	0.61490	0.59488	0.57468	0.55435	...	0.22222	0.20802	0.19441	0.18140
105	0.65188	0.63244	0.61272	0.59277	...	0.25204	0.23676	0.22205	0.20792
106	0.68735	0.66865	0.64959	0.63021	...	0.28374	0.26746	0.25171	0.23651
107	0.72107	0.70325	**0.68499**	0.66634	...	0.31712	0.29994	0.28324	0.26705
108	0.75282	0.73600	0.71867	0.70088	...	0.35196	0.33399	**0.31644**	0.29935
109	0.78243	0.76670	0.75042	0.73361	...	0.38797	0.36935	0.35109	0.33322
110	0.80981	0.79523	0.78006	0.76432	...	0.42487	0.40575	0.38691	0.36839
111	0.83489	0.82150	0.80750	0.79290	...	0.46232	0.44287	0.42361	0.40459
112	0.85766	0.84547	0.83266	0.81923	...	0.49999	0.48038	0.46087	0.44152
113	0.87814	0.86715	0.85554	0.84330	...	0.53756	0.51796	0.49837	0.47885
114	0.89641	0.88658	0.87615	0.86509	...	0.57468	0.55527	0.53578	0.51626
115	0.91256	0.90385	0.89455	0.88466	...	0.61105	0.59199	0.57276	0.55342
116	0.92672	0.91906	0.91085	0.90207	...	0.64637	0.62781	0.60901	0.59001
117	0.93901	0.93235	0.92516	0.91743	...	0.68038	0.66247	0.64424	0.62574
118	0.94961	0.94385	0.93761	0.93086	...	0.71284	0.69571	0.67819	0.66032
...	...	...	...	...	...	...	...	...	...

6.1.2 Wahrscheinlichkeitsintervalle für Schrotschusstreffer auf die Aktingenfamilie

Binomialtabelle für n = 26480, p = 0.000984 und verschiedene Werte für x im Bereich $5 \leq x \leq 50$ wiedergeben.

x	F(x|n,p)
...	...
5	0,00000
6	0,00000
7	0,00001
8	0,00004
9	0,00011
10	0,00030
11	0,00075
12	0,00174
13	0,00371
14	0,00738
15	0,01377
16	0,02417
17	0,04013
18	0,06322
19	0,09491

x	F(x|n,p)
20	0,13621
21	0,18745
22	0,24816
23	0,31695
24	0,39164
25	0,46949
26	0,54751
27	0,62280
28	0,69287
29	0,75582
30	0,81049
31	0,85643
32	0,89383
33	0,92336
34	**0,94598**
35	0,96282

x	F(x|n,p)
36	0,97500
37	0,98358
38	0,98945
39	0,99338
40	0,99593
41	0,99756
42	0,99856
43	0,99917
44	0,99953
45	0,99974
46	0,99986
47	0,99993
48	0,99996
49	0,99998
50	0,99999
...	...

6.2 Sequenzen von Transposonflanken aus inverser PCR

Die folgende Sequenzaufstellung gibt eine Zusammenfassung der Ergebnisse aus den Experimenten mit inverser PCR, beschrieben im Abschnitt 3.3.1 ab S. 36. In Kleinbuchstaben sind die Sequenzen der Primer hervorgehoben, mit denen das PCR-Produkt amplifiziert wurde.

6.2.1 Flankenpaare des Transposons thug-T

```
>JPCRa05d01  785 letters
ccgaccaccagttgttacAATTAAAATGAATATTATTTATATTTTTAAGATTTTTTTTAT
TTTTTTTTTTTCAAAAGATATGAAAAAATTTTGATCAAATTAAGATAAGTTGTAATAAAA
TTCAATTAAATACAATTGACACATCAATTTTAAAACATCCATCAGTTTTAGAATGTTGTT
CTATTGGTATTTTAAGTCCAGATTGTCGTACTGCACCGATTGGTATATTGGTATTAAAAG
AAAATCCATCAATTGATTTAAATAAATTACAAAATGAAATTAATAATATCATCACGCAAG
ATATTGAATCACTGGCAGTCTTAAAAAAAAATAATAGTTATTAACCAATTACCAAAAACAA
AAGTTGGTAAAATTCCAAGACAAATTTTATCAAATTTATTAAATGATCCAAACTATCAAT
TACCAGACGATGTCAATGACTCTGAACTATTTTATAAAATCAAAGAGTTATATATGAAAA
ATTAATTTTAGCAAAGTTTTTGTAAATTGTTTTTGTTCTAAAAAAAAAAAAAAAAAAAAA
GAGTAATAATAATTAAAAAAGAAAATAATTTGGATTATTTTTTTTTAATTTAATAAAAGT
AGGAGATATAAAAAATAAATAAAAAATATAAAAAAAAAATAAAAAAATTTAAACTTTTAA
ATAATATGTTTGAAATTATTTGATTTTTTTTTTTTTTTGACTAAAAATAGTGTTTTACTA
TGGGACAAAAAGAGTTTTTTTAATCCACCTCCATCCTTATTgttttgtttaaaggggttt
tctcc
>JPCRa05e01  2097 letters
ccgaccaccagttgttacAAACTAATCACAATAACACATCAATCAATCAATTTTCAAATT
GGTCAAAAATTATTAAAACTCTAAAAACTATCAATATAACAAAGACAAACAAACTCAACG
```

TGCAAAACAAATACTATTAGAAAAAGCAGCAAAATTCAAACATCTTGAAGAACAATGCAT
CGCTGAATACAATATTATGTGTAACAAAAACAAATTACATCAAAACCACACAAACATATT
AAAAAAACATATAAATGGAGAAATACCTAGCAAGTATCTATCAGCAATACTAAACAAAAG
ATCAAAAGATGCAAAAATGAATCAATCCAATATGGTAACATCATTACATCCGACCCCGAC
CAAATTGAAAATGCGTTTGTTGAATTCTACACAAATTTATATAGCTTACAAATATGTTGT
CCAATTACTCACCAGCTCATGCTAAACACATGGCCCATCATAAAAAATGAATATTGGAAT
GGTTTAGACTCACCATTTATACAAGATGAAGTTGAAGCTGCAATTAAAATCTGTAACCCC
AACAAATCACCCGGTCCAGATGGCGTTACAAATGCATTCTACATTAATCATTTAAACCAA
GTTAAACCAATTCTCACTACATTATTCAACGATATACTAGAAAATCCCCACCACATCACA
ACAGAATTCACCCAAGGTCTTATACACACAATATACAAGAAAGGCAACCCCTTACTCAAA
TCAAATCGTCGTCCCATTACACTTCTAAACACTGATTACAAGATCCTCTCAAAAGTAATC
AATACACGTCTTCTGAGAATACTTCCTTTCATCATCAACAACTACCAAACTGGTTTCGTA
CCCCACAGATTCATCAAAGACAACATCATCAACATCAATGAATTAATCAACTACCTCAAA
TCTAAAAACCTCCCTGGTATCATCACACTATTTGACTTCTTTAAAGCATTTGATTCTATC
TCTCACGATAGTATTAAAAGAACATTTATTCATATTGGTATACCAATAAAATTAATAAAT
TTAATCCACAAGCTATTATCTGACTCACAAGCAAAAATTTCAATTAATGGTAAAACCAGA
AAATTCGATATTAAAAGAGGTGTAAAACAAGGTGACCCAATCTCAGCCACTTTATTTGTA
ATTGTAATTGAAATATTAGCAAGAACAATAAATGCAGACAATTCAATAATTGGATTACCA
ATTTCACCCAATCCACAAATTAAAATCAAATTTACCCAATTTGCAGACGACCTTACAACA
TACAACATCAACTACGAACAACAACAACAGTCAATCAAACATTTCGATAATTTCTGCGCT
TCAACATCATCATCTTTAAATTTTGACAAAAGTGCAATAATAGAAATCAACCCCCACAAA
ATCACTGATAAACACATAATAAACAACATTCCACAATCAAAAAGAATTCCAATAACCAAA
AAAGATCAATCAGAAAGAGTTCTTGGCTATTTCTTTAATCATAATGGTTTACATAGGAAA
TTACCAGAAACAATGAAAACACTGATCAAGGTGAATTTAAAAACACTCCAACTCATACAC
ATGGCAACAGATTAGATAGAATATATACTCAATACAATAGAATAGACCCTGAAAATATAT
ATACACAAGTCCACCAAATTCCATCTTCATTATTAGATCATAATCCAGTATCAACAACAA
TTTCTTCACCTTTTCAAATAAATAAACATAAGAAAAGATGGATATTATCACCTGGTACTG
CAAATGACCATGAAGCAATCAATATTATCAATCATTTAATAACTTGTAACAAGATCTCAA
AAAAATATCTTGTTAATACTAATATTTATTTAATTTTATGAAATGTTCCAAACAGAACAT
TTAAAAATAACTATTAATATTGAATTGATAATTCAATATTAATTATTACTAACTTAATTC
ATTTTAATAAATTGTTTCATTTTATATAATAAACTCACTAAACACAACAAAACAATAATT
CAAAAATAAATAATAATAATTTAGAATATAGTTTTGTTTCAACATTTTGTTATGACACAT
ACAATTCAGAAAATAATAATAATTTATTTTATTgttttgtttaaaggggttttctcc
>JPCRa05a01 557 letters
ccgaccaccagttgttacAAACTAATCACAATAACACATCAATCAATCAATTTTCAAATT
GGTCAAAAATTATTAAAACTCTAAAAACTATCAATATAACAAAGACAAACAAACTCAACG
TGCAAAACAAATACTATTAGAAAAAGCAGCAAAATTCAAACATCTTGAAGAACAATGCAT
CGCTGAATACAATATTATGTGTAACAAAAACAAATTACATCAAAACCACACAAACATATT
AAAAAAACATATAAATGGAGAAATACCTAGCAAGTATCTATCAGCAATACTAAACAAAAG
ATCAAAAGATGCAAAAATGAATCAATCCAATATGGTAACATCATTACATCCGACCCCGAC
CAAATTGAAAATGCGTTTGTTGAATTCTACACAAATTTATATAGCTTACAAATATGTTGT
CCAATTACTCACCAGCTCATGCTAAACACATGGCCCATCATAAAAAATGAATATTGGAAT
GGTTTAGACTCACCATTCATACAATTCAGAAAATAATAATAATTTATTTTATTgttttgt
ttaaaggggttttctcc
>JPCRa05b03 924 letters
ccgaccaccagttgttacAGTAATGATTACAATTAAACAATAATAATTAAAATAATAATT
TATTTAGTTTTATATAATTTAGAGTTTTCATTCGAAACTTCTCGGTCGAAAATTGTAAAA
ACTAAAAAAAAATAAAGTTAGTTCTGATTGACCTTTACAGCATTCTTATAATATTAATTG
TTTTCCTTTGTTTACTAAGACAAAAACATTAAAGCATAACCCAAACACACTCAAACTCAA
ACACATATCAACAATTGATGGTAGAGCTATTTTAGCAGACGTTATATTTCCAAATCAATT
ACCAATTGGTATTCTTGGAATTTATGCACCAGCATCAAATGATACCATCAAAAAGAATTC
ATAACAACTTCACTACAAACACTTCTCCACAACCACTCAACAAACCCAATACCAAACCAA
TCCCAACATGTAATAATGCAGCTACGGGGGTTGGTCCTTCCATTGCATCTGATAACCACA
TATGTAAACCTAATTGTGCTGATTTTCCCACAACTCCTATTACTATGAAACTACCTATTA
AGGTAAACAAAATTTAATTATTAAATTACATTGTTAATATTAATAAGTTTTTAATAAAAT
TAATAGTAATTGTAACAAGATCTCAAAAAAATATCTTGTTAATACTAATATTTATTTAAT
TTTATGAAATGTTCTAAACAGAACATCTAAAAAATAACTATTAATATTGAATTGATAATT
CAATATTAATTATTACTAACTTAATTCATTTTAATAAATTGTTTCATTTTATATAATAAA
CTCACTAAACACAACTAAACAATAATTCAAAAATAAATAATAATAATTTAGAATATAGTT
TTGTTTCAACATTTTGTTATGACACATACAATTCAGAAAATAATAATAATTTATTTTATT
gttttgtttaaaggggttttctcc
>JPCRa05c03 3349 letters
ccgaccaccagttgttacAAACTAATCACAATAACACATCAATCAATCAATTTTCAAATT
GGTCAAAAATTATTAAAACTCTAAAAACTATCAATATAACAAAGACAAACAAACTCAACG
TGCAAAACAAATACTATTAGAAAAAGCAGCAAAATTCAAACATCTTGAAGAACAATGCAT

CGCTGAATACAATATTATGTGTAACAAAAACAAATTACATCAAAACCACACAAACATATT
AAAAAAACATATAAATGGAGAAATACCTAGCAAGTATCTATCAGCAATACTAAACAAAAG
ATCAAAAGATGCAAAAATGAATCAATCCAATATGGTAACATCATTACATCCGACCCCGAC
CAAATTGAAAATGCGTTTGTTGAATTCTACACAAATTTATATAGCTTACAAATATGTTGT
CCAATTACTCACCAGCTCATGCTAAACACATGGCCCATCATAAAAAATGAATATTGGAAT
GGTTTAGACTCACCATTTATACAAGATGAAGTTGAAGCTGCAATTAAAATCTGTAACCCC
AACAAATCACCCGGTCCAGATGGCGTTACAAATGCATTCTACATTAATCATTTAAACCAA
GTTAAACCAATTCTCACTACATTATTCAACGATATACTAGAAAATCCCCACCACATCACA
ACAGAATTCACCCAAGGTCTTATACACACAATATACAAGAAAGGCAACCCCTTACTCAAA
TCAAATCGTCGTCCCATTACACTTCTAAACACTGATTACAAGATCCTCTCAAAAGTAATC
AATACACGTCTTCTGAGAATACTTCCTTTCATCATCAACAACTACCAAACTGGTTTCGTA
CCCCACAGATTCATCAAAGACAACATCATCAACATCAATGAATTAATCAACTACCTCAAA
TCTAAAAACCTCCCTGGTATCATCACACTATTTGACTTCTTTAAAGCATTTGATTCTATC
TCTCACGATAGTATTAAAAGAACATTTATTCATATTGGTATACCAATAAAATTAATAAAT
TTAATCCACAAGCTATTATCTGACTCACAAGCAAAAATTTCAATTAATGGTAAAACCAGA
AAATTCGATATTAAAAGAGGTGTAAAACAAGGTGACCCAATCTCAGCCACTTTATTTGTA
ATTGTAATTGAAATATTAGCAAGAACAATAAATGCAGACAATTCAATAATTGGATTACCA
ATTTCACCCAATCCACAAATTAAAATCAAATTTACCCAATTTGCAGACGACCTTACAACA
TACAACATCAACTACGAACAACAACAACAATCAATCAAACATTTCGATAATTTCTGCGCT
TCAACATCATCATCTTTAAATTTTGACAAAAGTGCAATAATAGAAATCAACCCCCACAAA
ATCACTGATAAACACATAATAAACAACATTCCACAATCAAAAAGAATTCCAATAACCAAA
AAAGATCAATCAGAAAGAGTTCTTGGCTATTTCTTTAATCATAATGGTTTACATAGGAAA
TTACCAGAAACAATGAAAACACTGATCAAATCATTAGTACTATGGAAAACTAGTGGCACA
ACATTAAAAACAAAAACAACCATCATAAACACCTACTCACTATCACCAATAACATATTTA
TCATACCTTGAAGAATTTACAAAAGATGAAGAAATTCAAATAAATAAATTAATCTCAAGG
TTTATGAATTCTCCCGCAAATTCTGAATCACCAACTCTCGATACCAATTCCATTGAAAAC
AACACATCAACCATCCATTCACGTGCAAAAATACCTTTGATGTGTTATGATAGATCATTA
AAACCATTAAAAGAAGGTGGTTGGGGTATGTGGAATATACAATTACGACAAGTTGCTCAA
AAGATTTGGATATACAACAGATTTTTACAAATGCATAAATCTGCAAACAATTCAATATAC
ATGATAAGTTGGATGGATCAAATCATCAACAAATCAATCTCTTCTCCATACCTTATAAAA
ATCAAAAAAGAATGGGAAAACTATGCAACTCAAATTGGACATCTAAAAGATAAAGTTCAA
ATCCTACAACCAATATTAACAAAAAACCAACCCTCTCAAACATTAAACACAAATAATATT
TCACTACCACCAACACTCAAAGAAATCTACTCGACAATACTAAACAATCACCAATGCAAA
TCAAAAGACTATATTGGCAAGAAATATTCAGATCTACTTCTTACTTCGCACCAACAATCA
ATACAACTATTATGGAAATACACATACGACCAACTTTTTGTCAAAATTCAAAAATTAAAA
GATCCAAAAGGCCGTGATACAATGCAAAGATTCCATGCTAGATGTCTTCCGATTAATCAT
CTACACAACAAAGTTTGCCCCATTTGCAACAATGAAATGAATAATGATCCCTATGGTCAT
TTGTTTTTCAATTGTCAGCACACAATCAACTTCATAAACCACGACAAATTAAAATATTTT
ATATATAAAAATTGCAATGGAAACAAAAACTGGTCACTAACAAAAAACCAAACAACAAAA
TTATACACACTCATTTATAAACCACAAACAAACAACAAAGCATTTAAACCAGATCCAACT
ATCAATATTAATTTTCATTTTGCAGAAAACGCACAATATAAAAAATACAGGTTCAACTGG
AATTATATAAACACAAATCTTGATCTAGTCAGAACACATGTATCTATTATCGCAGGAGAT
TTCAACTGTATCCACAACGACAATTCACTATTAGATGATATTATCCAATCCTANANATAC
CAACACCATTTAATTGATCAAGGTGAATTTAAAAACACTCCAACTCATACACATGGCAAC
AGATTAGATAGAATATATACTCAATACAATAGAATAGACCCTGAAAATATATATACACAA
GTCCACCAAATTCCATCTTCATTATTAGATCATAATCCAGTATCAACAACAATTTCTTCA
CCTTTTCAAATAAATAAACATAAGAAAAGATGGATATTATCACCTGGTACTGCAAATGAC
CATGAAGCAATCAATATTATCAATCATTTAATAACTTGTAACAAGATCTCAAAAAAATAT
CTTGTTAATACTAATATTTATTTAATTTTATGAAATGTTCCAAACAGAACATTTAAAAAT
AACTATTAATATTGAATTGATAATTCAATATTAATTATTACTAACTTAATTCATTTTAAT
AAATTGTTTCATTTTATATAATAAACTCACTAAACACAACAAAACAATAATTCAAAAATA
AATAATAATAATTTAGAATATAGTTTTGTTTCAACATTTTGTTATGACACATACAATTCA
GAAAATAATAATAATTTATTTTATTgttttgtttaaaggggttttctcc

6.2.2 Flanken der Transposons DDT-A und DDT-S

>JPCRa38a06 486 letters
ggtgcatttttctttgctgtgCAACTACACCCTAAATAAGTGTGGACACACCACTCACAC
CTGGAATGTGAAGATTTTAAACAATCTACAAATTGCACTAGTAGCCAATTTAATAGCTAT
TATATTCGAAAAAATTTGGCACAAAAGAAACAAACTCATTCACGATAAAAAAGAAATAAT
AATTCATAGACAACAAGTCATACGTGAACTAATTAAAACACAAAGAGCTGCATGGGACAG
GATACAAGCGGTTATAAACAAAACATTAAGAATCAAATCAAAGCAACGGCCAGAAGAACA
AAATAAATTAGACTCACTAATCTCGCTAAGCTATTACAATTTAGAAATGGAACTCACCTC
TTCACTTAATAGCACTTCCGAAACATCTCAAAAAATACAATAATTCACTCAGTACTTTCT

ATAAATAAAAAAAAAAAAAAAAAAAAAAAAAAACTTGGGAACCCcacagcgaagaaaaa
tgcacc
>JPCRa39a07 741 letters
ggtgcatttttctttgctgtgTAGTCTCTAAAGAGTATGTTTATAAGTTTCATAATGTTG
AATCAAATTCATTATTAAAATATCTAATAATATAATCTTAATTTTTTAAAGAAATGGTCA
CGAAGAGTTACTTAAATCCGAAAAGAAAGTTTTTACAACCAATGATGGCCAAACCATTTC
ATCTAGTGTCCTCCGTAAGGTCATTAATGATGTTGAGTTTTCAAACCATTGGGACGAGAT
GCATAAAGGTCATATTGGAAGAGATGCCACTTACCAAAAATTCAAGTCAATGTATTTTTG
TACTGGTATGTGGGTAATGGTTGATAATGCAGTCAAGCAATGTGATATATGCCAAAGAAA
CAAAATTAAGGGTAAAATATTAATTTAAATTAATAAAAATTATTATTTATATCAAATGAT
ATTAATAATTTTAATATAGGTATCAATAAAGAATATGTTGCAATTGAAGATACTGAGGAG
TATTCAAGAATGGTTTTTGATTTAACATCTTTAAAAGGAGAGCATAGAAATAATGATATC
ATTATAGAATCTACCGATAGATTCTCTTCTAATTGGGAAACTATCAAAAATGATAATGTA
AAAGAAGCAAACGATTCAAACATTGTCTATATTCTTATTTGTGTCAATTCTTTCACAAAA
TTTGCAACTGGAAGGTAATTACCTTAATACTCTCCTAAAAGTTCAATATTATTTATATTA
cacagcaaagaaaaatgcacc
>JPCRa40a04 1017 letters
ggtgcatttttctttgctgTGACACCGCTTTCATTACACATATTGGTAAATTTGAATATA
CTAGAATGCCACAAGGTTTAGTCAACTCTCCATCCACATTCGCCAGATTGATGGTCGAAA
TATTTGGAAAAATCAAAAGTTTATTACAATACTTTGATGATCTTTTGGTTCATTCAAAAC
TCGACTACATGGTACATTTCATTGAAATCATTAGAATGCTTCTATATTGTAGAAAGTACC
TATTATTCATTTCAAGAGAAAAGAGTGAAATGTTAAAGACTGAAGTTGATTTCCTTGGTT
TTCACATTCATAAAGATGGTATATCTCCAAGAGCTGCAAAGGTTAGAGCTATCTCTGAGT
TACCTGAACCAAGAAACGCCAAAGAAGCTGAAGCTGCATTAGGTCTTTTTGGATTCTTCA
GGAGACACATTGAAAATTACGCTGAAAAAACCTATCACCTTTCCAAAGAATCAAAAGGGA
AAAACAAGAAAACTCTTTCTGATGAATCACTCAAAGAATTCAATAATCTAAAGAAAGAGT
TTGAAGGTGAAAATATTGTCGCTATTCCAATTGAACAAGATAACTCCATATCAATAGATA
TCGAAAAGGTTAAAGCATCAACAGACATGCCAATCCATTCAGATAATAATAATATAAACA
ATGGTAGTTTTCACTTGTATTGTGATGTTAGTGATAAAGCATTATCAGGTGTGTTATATC
AAATCCAAGGTGATAAATTCAAAGTCATTTGGTTCCACAGTAGAAAACTTACTGATACTC
AAAAGAGGTACAGCATAGGTGATAGAGAGTTCCTTTCAATCATTGATTCTCTAAAGAAGT
TTCAACATTTATTAATTGGTAAAAAGGTTTCAATCTACACTGATCACCAAAATCTTACAT
ATATTATCAATAAGTCAAACGATAAACCGTTCACAAAGAGACAAGATAATTATATGAAAT
ATATTAAAGAATTTGATTATGAATTAAGACATATAAcacagcaaagaaaaatgcacc
>JPCRa40a08 1092 letters
ggtgcatttttcttcgctgtgTTGGTAATACAGGAAGACACTAAACAACTTTATAAACTT
GAATTTTTGGAAGATGGAATCAATAGCCTTCAAAAAAAAGGTTTATACTCTGGTTTTGTA
GGAGGTAACAAGTTGGTTCTTTATAAACAATCAACAGTTGATATTTTTAGAACATCGCCA
AACATACAAAAAGATATTGATTCTTTTACTTCTGGTCTTTATTTAAATAATGATGGGAAA
GTTATTGAATTTCTGGTTCAGTTTATTAAAAATCTTGGTAAGGATTTATCCAATTTCAAC
TTGCATCAATTGGTAACACAAAATGATCCATTTACGGTTTTGGAAGATGCTCTTAACAAT
CCATCAAATATTCCACCTATCATTAACGATAATCCATTTCAACACAAACTCAATAATGAA
ATGGAACAAGAAGAAATACTCCCATTTTTAAATAATAATCCAACAGCGCCAAATTTAATG
AATTGTTTGAGAAAGGATATGAATTTAAAAGAAATAGTACCACCAATACCTCAAGTGCAA
ATAATACCATCGTATACTATACCACAATCTGGTAGTGGTATAGTAACAACTGTTAAACGT
CTCAGAGGCCGTCCTCATAAGATACCAATTGTTAATAAACCTCCTTTAAAATCACATTCT
AAACCATCCAAATCACCTCTCAAATCACCATCCATTTCACCTTCCAATTCACCATTTAAA
TCACCATTTAAATCACCTTCCAAATCACCATTTAAATCACCTTCCAAATCACCATCCATT
TCACCTTCCAATTCACCATTTAAATCACCTTCCAATTCACCTTTTAAATCACCATTTAAA
TCACCTTCCAAATCACCATCAATAGGAAAAATCGTTAATGTGATACCTTCAAAAAAAGTT
GCAAAAACAAAGTTAACAAGGACACAAAAATTAGAACTTGATTTGGTTGCAATTAATAGA
AGGGGGTATGGGGAAATTAGAAAATAAAAACAAGTGGCCAAATCAAAAAAACAAAAAAAA
CACAAATCAACCCAAAAAATTAAAAAAAAAAAAAAAATTAAAAATTAAAAAcacagcaaa
gaaaaatgcacc

6.3 Einträge in Sequenzdatenbanken

Tab. 46. Veröffentlichungen der erarbeiteten Sequenzdaten in Form von GENBANK-Einträgen.

Acc.No.	Titel
AF134169	Dictyostelium discoideum retrotransposon TRE3-A
AF134170	Dictyostelium discoideum retrotransposon TRE3-B
AF134171	Dictyostelium discoideum retrotransposon TRE3-C
AF135841	Dictyostelium discoideum retrotransposon TRE3-D
AF298201	Dictyostelium discoideum putative transposon DDT-A
AF298202	Dictyostelium discoideum putative transposon DDT-B
AF298203	Dictyostelium discoideum putative transposon DDT-S
AF298204	Dictyostelium discoideum gypsy-like LTR retrotransposon DGLT-A.1
AF298206	Dictyostelium discoideum transposon Tdd-5
AF298207	Dictyostelium discoideum transposon thug-S
AF298208	Dictyostelium discoideum transposon thug-T
AF298209	Dictyostelium discoideum non-LTR retrotransposon TRE5-B
AF298210	Dictyostelium discoideum non-LTR retrotransposon TRE5-C
AF298624	Dictyostelium discoideum chromosome 2 repeat region
AF4740004	Dictyostelium discoideum copia-like LTR retrotransposon DCLT-A

Selbständigkeitserklärung

Hiermit erkläre ich, daß

- mir die geltende Promotionsordnung der Fakultät bekannt ist,
- ich die vorliegende Dissertation selbst angefertigt habe und alle benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen angegeben habe,
- ich alle Personen, die mich bei Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts unterstützt haben, benannt habe,
- ich nicht die Hilfe eines Promotionsberaters in Anspruch genommen habe,
- Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- ich die vorliegende Arbeit nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe, auch nicht als Dissertation bei einer anderen Hochschule.

Jena, 12. Mai. 2002

Lebenslauf und wissenschaftlicher Werdegang

25.05.1970	geboren in Wuppertal
07.1976-06.1980	Grundschule in Wuppertal und Hanerau-Hademarschen
07.1980-06.1989	Besuch des Werner-Heisenberg-Gymnasiums in Heide (Holst.) und Kreisgymnasiums Itzehoe, Abitur am 02.06.1989
10.1991-09.1998	Studium der Biologie (Diplom) an der Heinrich-Heine-Universität Düsseldorf
01.1998-09.1998	Diplomarbeit am Institut für Physiologische Chemie I (Prof. H. Sies) zum Thema „Charakterisierung von Promotorelementen als Bindungsstelle für Vitamin-D-Rezeptoren“
10.1998-01.2002	Doktorarbeit am Institut für Molekulare Biotechnologie, Abt. Genomanalyse (Prof. A. Rosenthal, Dr. M. Platzer)

Wissenschaftliche Veröffentlichungen und Vorträge

Quack, M., K. Szafranski, J. Rouvinen, C. Carlberg. The role of the T-box for the function of the vitamin D receptor on different types of response elements. Nucl. Acids Res. 26, 5372-5378 (1998).

Szafranski, K. Charakterisierung von Promotorelementen als Bindungsstelle für Vitamin-D-Rezeptoren. Diplomarbeit Biologie, Düsseldorf 1998.

Szafranski, K., G. Glöckner, T. Dingermann, K. Dannat, A.A. Noegel, L. Eichinger, A. Rosenthal, T. Winckler. Non-LTR retrotransposons with unique integration preferences downstream of *Dictyostelium discoideum* tRNA genes. Mol. Gen. Genet. 262, 772-780 (1999).

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

Glöckner, G., L. Eichinger, K. Szafranski, J.A. Pachebat, A.T. Bankier, P.H. Dear, R. Lehmann, C. Baumgart, G. Parra, J.F. Abril, R. Guigó, K. Kumpf, B. Tunggal, the Dictyostelium Genome Sequencing Consortium, E. Cox, Q.M. Quail, M. Platzer, A. Rosenthal, A.A. Noegel. Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. Nature (2002), im Druck.