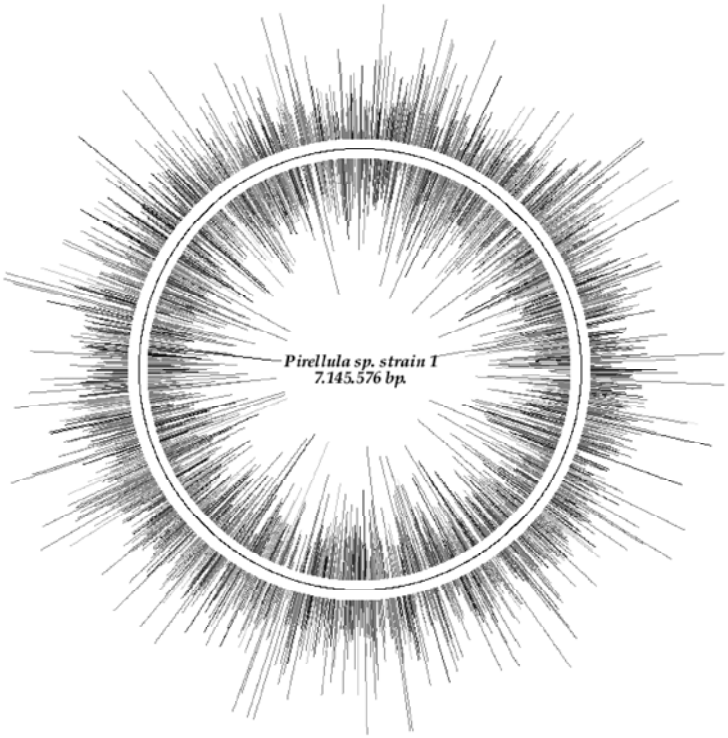
Sequenzierung und Strukturen von *Pirellula* sp. Stamm 1



Michael Kube

Die Grafik auf dem Titelblatt (Kube, Beck und Reinhardt, MPI für Molekulare Genetik Berlin) zeigt die Zusammenfassung der ersten automatischen Rohannotation des Genoms von *Pirellula* sp. Stamm1 mit HTGA (Rabus et al. 2002b) im Rahmen der Evaluierung des REGX Projektes.

Sequenzierung und Strukturen von *Pirellula* sp. Stamm 1

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
- Dr. rer. nat. -

dem Fachbereich Biologie/Chemie der
Universität Bremen
vorgelegt von

Michael Kube

Juli 2003

Die vorliegende Arbeit wurde in der Zeit von April 2000 bis Mai 2003 am Max Planck Institut für Molekulare Genetik Berlin Dahlem angefertigt.

1. Gutachter: Prof. Dr. Rudolf Amann
2. Gutachter: Prof. Dr. Dietmar Blohm

Tag des Promotionskolloquiums: 09.07.2003

Danksagung:

Herrn Prof. Dr. Amann danke ich für die Übernahme des Promotionsgutachtens und für sein persönliches Engagement. Ebenso danke ich Herrn Prof. Dr. Blohm für sein Interesse an der Arbeit und die Bereitschaft dieses Promotionsvorhaben als Vertreter der Universität Bremen zu unterstützen.
Mein besonderer Dank gilt Herrn Prof. Dr. Lehrach und Herrn Dr. Reinhardt für ihre fachliche Betreuung, ihre Anregungen und ihren persönlichen Einsatz für meinen beruflichen Werdegang. Meinen Freunden und Kollegen am Max Planck Institut für molekulare Genetik gilt mein Dank für die freundliche Arbeitsatmosphäre und Unterstützung bei dieser Arbeit.

Berlin, Juli 2003

Michael Kube

INHALTSVERZEICHNIS

INHALTSVERZEICHNIS

INHALTSVERZEICHNIS

1. Einleitung

1.1 Thematische Zielsetzungen

Die vorliegende Arbeit befaßt sich mit der Sequenzierung und den Strukturen des Genoms von *Pirellula* sp. Stamm 1. Die Sequenzierung ermöglicht zum ersten Mal Einblicke in das vollständige Genom eines Mitglieds der Gruppe der Planctomyceten. Die Bestimmung der genomischen Sequenz stellte alleine schon durch die Größe des Genoms von mehr als sieben Megabasen eine Herausforderung dar. Die hierfür gewählte Vorgehensweise sowie die hiermit verbundenen grundlegenden Überlegungen werden im ersten Teil der Arbeit dargelegt. An diesen Abschnitt schließen sich die bioinformatischen Analysen an, wobei nur einzelne Elemente der genomischen Sequenz fokussiert betrachtet werden konnten. Sie wurden ausgewählt, weil sie im Rahmen der Bestimmung der genomischen Sequenz als repetitive Elemente von besonderer Bedeutung sind. Andere Strukturen, wie das bei den Planctomyceten unterbrochen vorliegende rRNA Operon und die Verteilung der tRNAs im Genom, werden mit den wenigen bereits vor der Sequenzierung dieses Genoms bekannten genetischen Informationen aus den Planctomyceten und anderen mikrobiellen Genomen verglichen.

1.2 *Pirellula* sp. Stamm 1 als Teilprojekt von Real Environmental GenomiX (REGX)

Das *Pirellula* Genom Projekt stellt eines von drei Genomen mariner umweltrelevanter Organismen dar, die innerhalb der vom BMBF (Förderkennzeichen 03F0279C) geförderten *Real Environmental GenomiX* (REGX) Projekte analysiert werden sollen. Hierbei handelt es sich neben *Pirellula* sp. Stamm 1 um die marinen Sulfatreduzierer *Desulfobacterium autotrophicum* und *Desulfotalea psychrophila*.

Seit April 2000 wird im Rahmen des *Pirellula* sp. Stamm 1 Projektes versucht, möglichst umfangreich Daten zum Genom, Transkriptom und Proteom zu erheben. Langfristig sollen sich die gewonnenen Daten für ein auf DNA-Chiptechnik (Microarray) basierendes Biomonitoring für die Umweltmikrobiologie nutzen lassen. Hierbei kann im Idealfall die Expression von Genen und ihre Regulation in Abhängigkeit von Umwelteinflüssen gemessen werden. Mit einer derartigen Methode wird man in der Lage sein, die Anpassung des Organismus an seine Umweltbedingungen bzw. auch wechselnde Umweltbedingungen zu erfassen. *Pirellula* sp. Stamm 1 repräsentiert ein wichtiges Mitglied der mikrobiellen Ge-

sellschaft in marinen Systemen, die an der Degradation von Biopolymeren zu Kohlenstoffdioxid beteiligt sind.

Neben der Bedeutung für marine Kohlenstoffkreisläufe steht bei *Pirellula* die Zugehörigkeit zu den Planctomyceten im Vordergrund. Die eigenständige weit verbreitete Bakteriengruppe der Planctomyceten zeigt sich in der Forschung bisher weitgehend uncharakterisiert. Die vorliegende Arbeit als Teil des *Pirellula* Projektes versucht diese Lücken aufzufüllen. Als erstes Beispiel, für die sich aus der genomischen Sequenz ergebenen Möglichkeiten, ist die Verknüpfung von Proteom und Genomdaten bei der Analyse des Acetylglucosamin Metabolismus von *Pirellula* sp. Stamm 1 zu nennen (Rabus et al. 2002a).

1.3 Charakteristika der Planctomyceten

Die Gattung *Pirellula* gehört zur Ordnung der *Planctomycetales* (Schlesner & Stackebrandt 1986; Staley et al. 1992; Griepenburg et al. 1999), welche ein Teil des phylogenetisch tief abzweigenden Phylums *Planctomycetes* ist (Stackebrandt et al. 1984; Hugenholtz et al. 1998; Abb. 1).

Der ursprüngliche Name Planctomyces geht auf Gimesi (1924) zurück, der fälschlicherweise dieses Genus den Pilzen zuordnete. Die Bezeichnung etablierte sich dennoch (Staley et al. 1992). Früher wurden die Planctomyceten als rein aquatisch beschrieben (Staley et al. 1992), wo sie zum Teil in hohen Zahlen nachgewiesen werden konnten. Es gibt jedoch auch Nachweise für das Auftreten an terrestrischen Standorten z.B. im Zusammenhang mit einer 16S rDNA Studie mit Proben aus australischen Böden (Liesack & Stackebrandt, 1992). Auch andere Studien auf der Basis von 16S rDNA/rRNA unterstreichen die weite Verbreitung der Planctomyceten in unterschiedlichsten Habitaten (DeLong et al. 1993; Bond et al. 1995; Gray & Herwig 1996; Borneman et al. 1996; Lee et al. 1996; Fuerst et al. 1997; Glöckner et al. eingereicht). Reinkulturen konnten vorwiegend aus Süß-, Brack- und Salzwasser isoliert werden (Schmidt 1978; Schlesner 1986, 1994; Giovannoni et al. 1987).

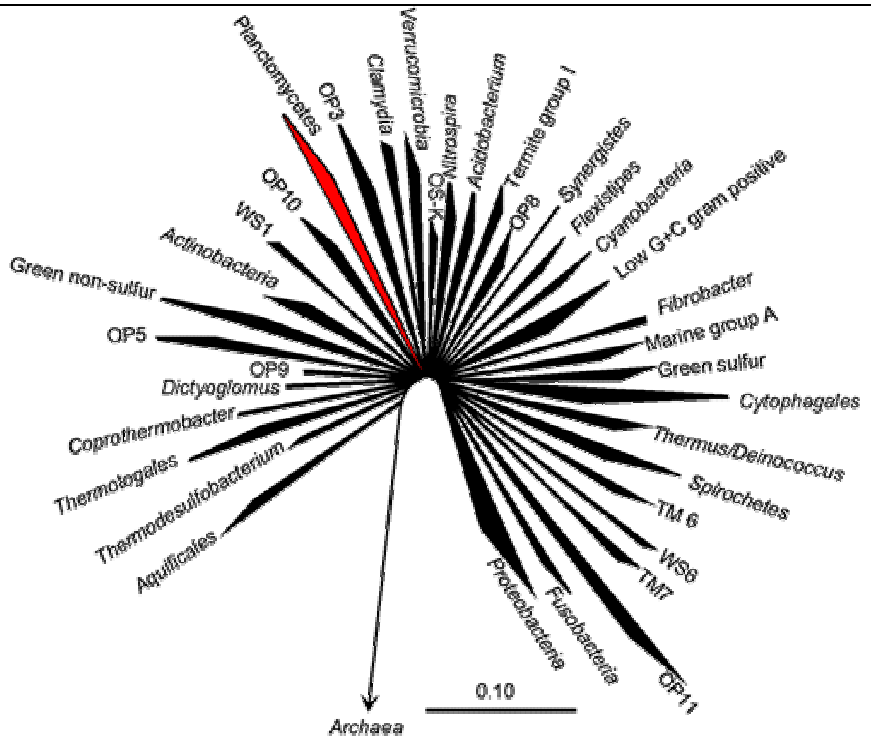


Abb. 1: Stammbaum der unterschiedlichen Bakterienphyla
Die Abbildung zeigt einen evolutionären Distanzbaum bekannter Phyla und potenzieller Kandidaten. Der Stammbaum basiert auf der ARB-Software und der ARB Sequenz Datenbank Stand März 1997. Die Gruppierung auf dem Phylumlevel von zwei oder mehr Sequenzen ist als Keile dargestellt. Die Tiefe des Keiles spiegelt die abzweigende Tiefe der Repräsentanten für eine Abteilung wider. Keile von Abteilungen mit bereits kultivierten Vertretern sind ausgefüllt dargestellt. Der Maßstab definiert sich durch 0,1 Austausche pro Nukleotid (Hugenholtz et al. 1998, modifiziert www.regx.de).

Bei den Planctomyceten handelt es sich um aerobe, Gram-negative, chemoheterotrophe Bakterien mit überwiegend kugelförmiger, ovoider oder birnenförmiger Zellform. Sie verbindet eine Reihe von einzigartigen Besonderheiten, die es ermöglichen, sie von anderen Bakterien zu unterscheiden. Hierzu gehört die Knospenbildung bei der Reproduktion (Schmidt 1978). Wichtig ist auch das Fehlen von Peptidoglykan in ihren Zellwänden, eine Besonderheit der Planctomyceten, die sie innerhalb der *Bacteria* nur mit den Chlamydiae und den zellwandlosen Mycoplasmen teilen (König et al. 1984; Liesack et al. 1986; Lindsay et al. 2001). An Stelle von Peptidoglykan besitzen sie eine proteinhaltige Zellmembran, die reich an Cystin und Prolin ist (Liesack et al. 1986; Stackebrandt et al. 1986); Aminozucker und neutrale Zucker liegen nur in geringen Mengen vor (Liesack et al. 1986).

Ein weiteres hervorhebenswertes Merkmal der Gruppe ist die von einer Membran umschlossene intracytoplasmatische Struktur, das so genannte Pirellulosome (Abb. 2).

Das Auftreten dieser intracytoplasmatischen Membran prägt die Zellorganisation der Planctomyceten. Für *Pirellula marina* und *Pirellula staleyi* konnte gezeigt werden, dass es sich hierbei um eine einschichtige Membran handelt (Lindsay et al. 1997; Lindsay et al. 2001). Diese Organismen stehen damit im Gegensatz zu *Gemmata obscuriglobus*, einem anderen Planctomyceten, bei dem die DNA von einer Doppelmembran umgeben ist. Diese Besonderheit der Zellorganisation ist besonders vor dem Hintergrund interessant, dass der von einer Doppelmembran umgebener Nucleus stets als eines der grundlegenden Unterscheidungsmerkmale zwischen Eukaryoten und Prokaryoten herangezogen wurde. Die elektronenmikroskopischen Aufnahmen und die daraus abgeleiteten differenten Zellorganisationen unterschiedlicher Vertreter der Planctomyceten werden von Lindsay et al. (2001) als evolutionäre Entwicklung einer Kompartimentierung, die sich an die Bedürfnisse des jeweiligen Organismus angepasst hat, gedeutet.

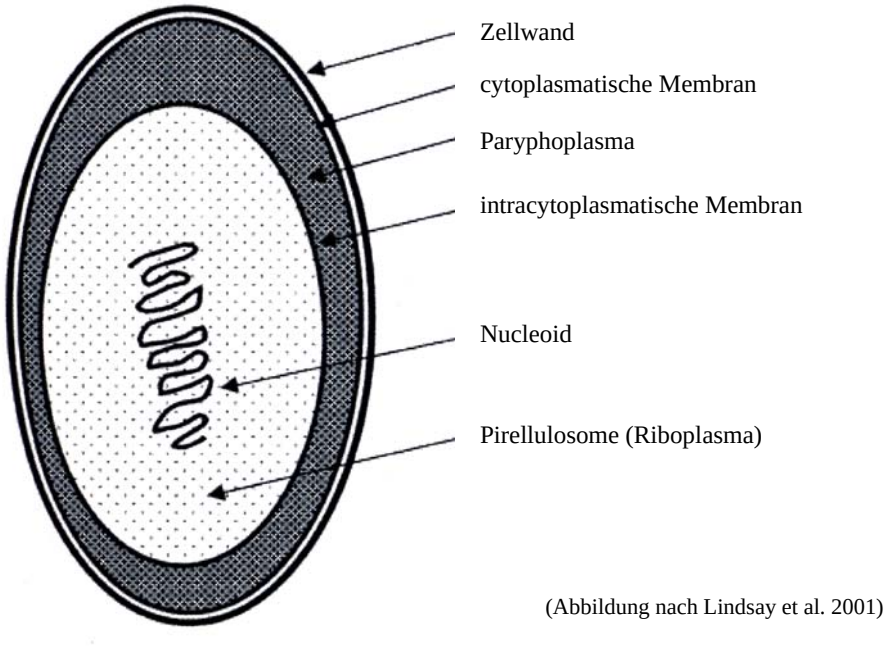


Abb. 2:
Beispiel der Zellorganisation und Aufteilung in einzelne Kompartimente in *Pirellula marina*
Die schematische Abbildung verdeutlicht die Kompartimierung in das ribosomenfreie Paryphoplasma und das Riboplasma, das die Nucleoidregion und ribosomenähnliche Partikel einschließt.

Hypothesen, die auf der Endosymbiontentheorie beruhen (Gupta & Golding 1996; Lake & Rivera 1994), werden von Lindsay et al. (2001) als unwahrscheinlicher angesehen, da er seine Ergebnisse an die Konzepte einer autonomen Kerngenese anlehnt (Lake & Rivera 1994) und diese bestätigt sieht. Fraglich bleibt, ob die Ergebnisse von Lindsay et al. (2001)

der Endosymbionten Theorie endgültig widersprechen können. Das Fehlen von DNA im Paryphoplasma aller untersuchten Planctomyceten stellt das wichtigste Argument gegen die Endosymbionten Hypothese dar. Das Auffinden eines ursprünglicheren Organismus mit DNA im Paryphoplasma und Riboplasma würde die Interpretationen zur Genese des Pirellulosoms infrage stellen. Aufgrund der nur eingeschränkt charakterisierten Gruppe der Planctomyceten ist eine derartige Entdeckung nicht auszuschließen.

Weitere Hinweise, die für eine Verifikation oder Falsifikation der Endosymbionten Hypothese dienen könnten, lassen sich vielleicht in den Genomen der Planctomyceten finden. Gene, die an der Produktion der paracytoplasmatischen und intracytoplasmatischen Membran beteiligt sind, könnten phylogenetisch unterschiedlichen Ursprungs sein. Diese Analysen setzen natürlich noch zu erwerbende Kenntnisse über die beteiligten Gene und deren Identifikation voraus. Auch doppelt vorliegende Gensets oder deren Relikte für bestimmte Funktionen müssten nachweisbar sein. Wenn auch die Entstehung und Entwicklung des Pirellulosoms letztendlich nicht geklärt ist, so ist seine Bedeutung für die Gruppe der Planctomyceten nicht zu übersehen. Die Bezeichnung Nucleoid (sinngemäß dem Nucleus ähnlich) erscheint im Falle von *Pirellula* noch einiges Potenzial für weitere Forschungsarbeiten aufzuweisen.

Über die Genome der Planctomyceten lagen bisher nur geringe Erkenntnisse vor. Sie beschränken sich auf einzelne Gene, wie z.B. die des rRNA-Operons. Letztere wurden für die phylogenetische Analysen auf der Basis von 16S und 5S rDNA im Vergleich zum Elongationsfaktor Tu (Jenkins & Fuerst 2001) sowie Analysen zur Organisation des Operons herangezogen (Liesack & Stackebrandt 1989). Diese Analysen zeigten ein Reihe von Besonderheiten auf, wie die kürzeren 5S rRNAs oder das aufgebrochene rRNA-Operon (Bomar et al. 1988; Liesack & Stackebrandt 1989; Menke et al. 1991). Die phylogenetische Position der Planctomyceten konnte nicht abschließend bestimmt werden, da in Abhängigkeit von der verwendeten Methode unterschiedliche Ergebnisse erzielt wurden (Jenkins & Fuerst 2001). In einigen 16S rRNA basierenden Studien wurden die *Chlamydiae* als nächste Verwandte bestimmt (Weisburg et al. 1986), in anderen konnte dies nicht bestätigt werden (Van de Peer et al. 1994; Embley et al. 1994). Beispiele für die inkonsistenten phylogenetischen Resultate liegen mit der 5S rRNA (Van den Eynde et al. 1990) und 23S rRNA Analysen (Schleifer & Ludwig 1989) vor. Hinzu kommt die vermutlich hohe Geschwindigkeit, mit der die Evolution bei den Planctomyceten abläuft, die zu Verzerrungen bei der Stammbaumanalyse führen kann (Liesack & Stackebrandt 1992).

Die nicht auf einzelne Gengruppen reduzierte Studie von Jenkins et al. (2002) zeigt ebenfalls die Problematik der Suche nach der nächsten Verwandtschaftsgruppe der Planctomyceten bzw. der Eigenständigkeit der Gruppe auf. Die hier mithilfe der Shotgun-Methode zufällig ausgewählten DNA Abschnitte zeigen in erster Linie Sequenzhomologien zu den *Bacteria*, aber es werden auch beste Ähnlichkeiten zu den *Eukaryota* und *Archaea* gefunden. Vergleichende Analysen zu den Genomen der Planctomyceten liegen für den GC-Gehalt und die Genomgröße vor. Für verschiedene *Pirellula* Stämme wurde der GC-Gehalt innerhalb einer Schmelzpunktanalyse bestimmt. Der GC-Gehalt variierte zwischen 54 und 58% in den untersuchten Genomen (Gebers et al. 1985).

Die Genomgrößenbestimmung bei knospenden *Bacteria* wurde 1985 unter Verwendung der Renaturierungs-Kinetik durchgeführt (Kölbel-Boelke et al. 1985). *E.coli* K12 wurde als Standard gewählt (1985 mit 4.5 x 10^6 bp angenommen). Für die Planctomyceten wurden unter Verwendung der alten taxonomischen Bezeichnungen *Planctomyces* und *Pirella* Genomgrößen mit 4.7 Mb bis 7.4 Mb bestimmt.

1.4 Sequenzierung des Genoms im Rahmen der Shotgun-Strategie

Das Genom von *Pirellula* sp. Stamm 1 wurde im Rahmen eines *whole genome shotguns* sequenziert. Das Sequenzieren nach der Shotgun Strategie unterscheidet sich grundlegend von den klassischen Sequenzierstrategien. Vermutlich der bedeutenste Vorteil des Shotguns ist die zufällige Generierung von DNA-Fragmenten. Wie der Name andeutet, sollte die Generierung der Fragmente so zufällig erfolgen wie das Auftreffen von Schrotkugeln auf ein Zielobjekt. Der ursprüngliche Ansatz des gerichteten Klonierens setzt die Kenntnis des Restriktionsmusters bzw. der Anordnung jedes Inserts voraus. Restriktionsfragmente werden hierbei in den Sequenzierungsvektor ligiert. Weitere benötigte Sequenzen wurde durch *Primer Walking* mit sequenzspezifischen Primern oder durch die Verkürzung der Sequenz, z.B. mit Exonuklease III und systematisches Klonieren erreicht (Martin-Gallardo et al. 1994; Bodenteich et al. 1994).

Die Shotgun-Strategie basiert auf dem Ansatz, dass mit hoher Redundanz zufällig erzeugte überlappende Fragmente eines Genoms sequenziert werden. Die erhaltenen Sequenzen (Reads) können aufgrund ihrer übereinstimmenden DNA-Sequenz zu zusammenhängenden Sequenzbereichen assembliert werden (Contigs) und ermöglichen schließlich so die Determination einer durchgängigen Sequenz. Die Bestimmung von doppelsträngigen Sequenzen oder eine Verlängerung der Sequenz wird durch die redundante Sequenzierung des Genoms und die Sequenzierung der Inserts von beiden Enden ermöglicht (Edwards et

al. 1990). Diese Sequenzierung der Inserts von beiden Enden ermöglicht die Bildung von zusammenhängenden Sequenzbereichen, die entweder eine durchgängige (unter Umständen zunächst einzelsträngige Sequenz) assemblierte Sequenz bilden oder eine zunächst von Lücken durchsetzte Sequenz (*Sequencing Gaps*) besitzen. *Sequencing Gaps* sind definierte Lücken. Die Verknüpfung der zusammengehörenden Einzel-Contigs, die durch die Lücke getrennt werden, ist durch das überspannende von beiden Seiten ansequenzierte Insert bestimmt. *Sequencing Gaps* lassen sich entsprechend durch gezieltes Nachsequenzieren der fehlenden Sequenz des Brückenklons schließen und werden zunächst wissentlich in Kauf genommen. Sie stehen damit im Gegensatz zu Lücken der genomischen Sequenz, die nicht durch Inserts abgedeckt werden (*Physical Gaps*).

Eine Shotgun-Bank erreicht mit zunehmender Insertgröße schneller eine physikalische Abdeckung (*Physical Coverage*) des Genoms. Mit zunehmender Insertgröße verringert sich jedoch erfahrungsgemäß die zufällige gleichmäßige Verteilung der Klone über das Genom und der Aufwand zur Bereitstellung sequenzierfähiger Templates steigt in manueller und materieller Hinsicht extrem an. Die Verwendung von auf *high copy* Plasmiden basierenden Shotgun-Banken mit unterschiedlichen Insertgrößen im Bereich von unter fünf Kilobasen stellt hier einen Kompromiss dar. In diesem Größenbereich stehen weiterhin Methoden aus dem HTS-Bereich (*high throughput system* Bereich) zur Verfügung und die statistische Abdeckung des Genoms bleibt gewährleistet. Die Verwendung von zwei Banken im kleineren (1,5 kb) und größeren Insertbereich (3,5 kb) ermöglicht die Bereitstellung der Templates für die Sequenzierreaktion im Hochdurchsatz mithilfe der Polymerasenkettenreaktion.

Überlappende Shotgun-Sequenzen führen zu einer redundanten Sequenzierung der genomischen Sequenz und damit zu hoher Sequenzqualität. Entwicklungen wie das *Oligonucleotide Fingerprinting*, die zu einer Absenkung der Redundanz in der Shotgun-Sequenzierung durch die Vorauswahl der Shotgun-Klone eingesetzt werden können (Radelof et al. 1998), konnten sich durch die technischen Fortschritte im *Shotgun Sequencing* nicht etablieren. Statt der Diskriminierung von Klonen stehen Überlegungen im Vordergrund gezielt Klone, die in unterrepräsentierten Bereichen liegen, durch die Bestimmung kurzer Endsequenzen der Shotgun-Klone aufzuspüren. Erste innovative Versuche nutzen die Massenspektroskopie zur Bestimmung der Sequenz (Nordhoff et al. 2000). Mit dieser Methode können mehr als 20000 Endsequenzen innerhalb eines Arbeitstages bestimmt werden.

Ausgehend von Plasmiden, Viren (Sanger et al. 1982), Cosmiden und BACs (*bacterial artificial chromosomes*) wurden schließlich die DNA-Sequenz der ersten mikrobiellen Genome mit der Shotgun-Strategie (*whole genome shotgun*) erschlossen. Ursprünglich eingesetzt zu Sequenzierung kleiner Genome wie *Haemophilus influenza* (1,8 Mb; Fleischmann et al. 1995) und *Methanococcus jannaschii* (1,7 Mb; Bult et al. 1996), folgte die Sequenzierungen größerer Genome wie das 6,3 Mb große Genom von *Pseudomonas aeruginosa* (Stover et al. 2000). Die Sequenz anderer großer Bakteriengenome (*Bacillus subtilis* mit 4,2 Mb Kunst et al. 1997; *Escherichia coli* K12 mit 4,6 Mb, Blattner et al. 1997; *Mycobacterium tuberculosis* mit 4,4 Mb, Cole et al. 1998, *Streptomyces coelicolor* A3(2) mit 8,7 Mb, Bentley et al. 2002) basierte zunächst auf einer angeordneten *Large Insert Library* (LIL), die in der Folge die Basis für die weiteren Subklonierungen und Sequenzierungen oder in Form einer Kartierung die Grundlage für die Anordnung der Sequenz bildete. Eine derartige Vorgehensweise wurde für das Genom von *Pirellula* sp. Stamm 1 auf Grund der bestehenden umfangreichen Erfahrungen mit der Shotgun-Strategie für nicht notwendig erachtet.

Neue Klonierungssysteme ermöglichen unabhängig von Restriktionsendonukleasen die Herstellung von Fosmid- und Cosmidbanken. Das große Genom von *Pirellula* sp. Stamm 1 legte in der Endphase der Sequenzierung ebenfalls die Erstellung einer weiteren Bank mit Inserts einer Größe von über 10 kb nahe. Diese *Library* sollte das Schließen der *Physical Gaps* und *Sequencing Gaps* unterstützen sowie der Überprüfung von Verknüpfungen dienen. Alle Aufgaben ließen sich jedoch mit den bereits vorhandenen Banken und PCR-Produkten auf der genomischen DNA lösen. Die Etablierung der Erstellung von Cosmidlibraries und die Präparation der Cosmide im HTS für die Sequenzierung wurde im Rahmen anderer Projekte benötigt, so dass parallel auch eine Cosmidlibrary für *Pirellula* sp. Stamm 1 erstellt wurde, die somit einer weiteren Verifikation der genomischen Sequenz diente.

Überzeugt von der hohen Effizienz der Shotgun-Strategie regten bereits 1997 Weber und Meyers die Sequenzierung des menschlichen Genoms mit der Shotgun-Methode an, wobei auf die BAC-Sequenzen und bekannten Marker zurückgegriffen werden sollte und später auch wurde. Autoren wie Green (1997) kritisierten diese Vorschläge wohl eher aus ideologischen Gründen, was zur bekannten Konkurrenzsituation im humanen Genom führte. Das Hauptargument gegen den Shotgun-Ansatz zeigt dessen Achillesferse auf. Repetitive Sequenzen aller Art, wie sie für eukaryotische Genome typisch sind, können zu falschen Assemblierungen der Daten führen, wenn sie in extremer Häufung auftreten (Green 1997)

und sie die Leselänge der Sequenzierreaktionen überschreiten. Die *whole genome shotgun* Strategie muss in diesen Fällen durch andere Methoden ergänzt werden, die sich im Wesentlichen auf die Bereitstellung zusätzlicher Informationen über die Verknüpfung der Sequenz zusammenfassen lässt. Hierzu lässt sich zur Kontrolle die Anordnung bzw. Orientierung der beiden Reads jedes Inserts im Contig heranziehen. Dieses Prinzip lässt sich solange anwenden, wie die Insertgröße nicht die Repeatlänge unterschreitet und die Reads nicht bereits wieder in einem neuen Repeat liegen. Voraussetzung für eine effektive Anwendung dieser Strategie sind neben der notwendigen Insertgröße auch die Identifikation der vorhandenen Repeats sowie das Fehlen von sehr komplexen als auch häufigen Repeats in der zu assemblierenden Datenmenge. Diese Situation tritt in Bakteriengenomen im Vergleich zu eukaryotischen Genomen nur in geringem Umfang auf, wodurch die Problematik begrenzt ist (Green 1997). Bei größeren Genomen mit einer Vielzahl an komplexen Repeatmustern wie z.B. dem humanen Genom erwies sich (wie gefordert; Weber & Meyers 1997) die Bereitstellung eines weiteren Datensatzes von überlappenden Sequenzen und Markern z.B. aus BAC libaries als hilfreich, die eine Überprüfung der Anordnung von großen zusammenhängenden DNA-Bereichen ermöglichen. Eine derartige Verknüpfung des *whole genome shotgun* wurde bei der Sequenzierung des Genoms der Fruchtfliege *D. melanogaster* genutzt (Adams et al. 2000). Bei vielen anderen eukaryotischen Projekten überwog die Strategie, die klonierten großen Fragmente einer Shotgun-Sequenzierung zu unterziehen. Dieser hierarchische Ansatz erwies sich bei der Sequenzierung von eukaryotischen Genomen als erfolgreich. Beispiele hierfür sind der Nematode *C. elegans* (C. elegans Sequencing Consortium 1998), die Brassicaceae *A. thaliana* (Arabidopsis Gen. Init. 2000) und das humane Genom (International Genome Sequencing Consortium 2001).

Genomprojekte wie *Drosophila melanogaster* (Adams et al. 2000) und das humane Genom spiegeln die Repeatproblematik des *whole genome shotgun* pointiert wider. Von beiden Genomen läge ohne den *whole genome shotgun* nur ein Bruchteil der Sequenz vor. Jedoch werden bei beiden Genomsequenzierungen noch Jahre bis zur lückenlosen Vervollständigung der Sequenz benötigt, was insbesondere auf die repeatreichen Centromerregionen zurückzuführen ist.

1.5 Fragmentierung der genomischen DNA als Basis der Shotgun-Sequenzierung

Eine Vielzahl von Methoden steht zur Fragmentierung der DNA zur Herstellung von Shotgun-Banken zur Verfügung. Diese Methoden verfolgen das Ziel der zufälligen Fragmentierung der DNA. Dominiert werden die Methoden zur Herstellung von Shotgun-Banken durch zwei unterschiedliche Verfahren: das hydrodynamische Scheren und das enzymatische Schneiden der DNA (Sambrook & Russel 2001).

Enzymatische Verfahren basierend auf Restriktionsendonukleasen (Sanger et al. 1980; Messing et al. 1981; Baer et al. 1984) aber auch DNase I, die z.B. in Gegenwart von hohen Mn^{2+} Konzentrationen Doppelstrangbrüche induziert (Campbell & Jackson 1980). DNase I soll eine weitgehend zufällige Verteilung der Fragmente ermöglichen, wobei hier jedoch nur ein kleiner Teil im gewünschten Bereich liegen soll (Sambrook & Russel 2001). Die Verwendung von Restriktionsendonukleasen zur Fragmentierung der DNA (Sanger et al. 1980; Messing et al. 1981) wurde ausgeschlossen, da eine zufällige Verteilung der zu generierenden Fragmente durch die vorgegebene spezifische Erkennungssequenz nicht zu realisieren ist. Das partielle Schneiden der DNA mit Restriktionsendonukleasen, um Fragmente in einem bestimmten Größenbereich zu erhalten, kann nie eine wirklich zufällige Verteilung erreichen, weil die möglichen Schnittstellen durch die Sequenz vorgegeben sind. Zeitkostende Vorversuche, die von möglichst hochmolekularer genomischer DNA ausgehen, begleiten hier das Bemühen um eine gute *Library*, die sich durch eine möglichst zufällige Verteilung der Inserts auszeichnet. Die Verwendung von Restriktionsendonukleasen wie CviJI, die ihre Stärke in der Generierung von Shotgun-Banken aus geringen Ausgangsmengen an DNA zeigen (0,2-0,5 µg DNA), verringert diesen Effekt nur (Fitzgerald et al. 1992; Davis et al. 1996).

Für die Herstellung von Banken durch mechanisches Scheren kann von hochmolekularer flüssiger DNA ausgegangen werden, die so gering wie möglich durch enzymatische Vorgänge degradiert ist. Eine eventuell zum Teil auftretende mechanische ungerichtete Degradierung der DNA bei der Präparation stellt jedoch kein Problem dar, da eine weitere Fragmentierung in den nachfolgenden Arbeitsschritten angestrebt wird.

Entscheidend für die Wahl des Systems zum Scheren war nicht die Möglichkeit einen möglichst engen Größenbereich zu generieren, sondern der Klonierungserfolg der Fragmente. Systeme wie die Nebulizer-Technik (Invitrogen, Karlsruhe) oder HydroShear (GeneMachines, San Carlos/USA) erreichten in durchgeführten Experimenten (nicht dargestellt) einen verhältnismäßig eng definierten Fragmentgrößenbereich mit einer Schwan-

kungsbreite von 4-6 kb. Die geringe Schwankungsbreite brachte eine Anreicherung der DNA in einem stärker definierten Größenbereich als z.B. bei der Beschallung mit Ultraschall mit sich. Eine später erfolgende Größenselektion bleibt aber unerlässlich, da kleinere Fragmente im Größenbereich von 200-900 bp, die zunächst nicht detektiert werden, später bevorzugt ligiert werden und so die Insertgrößenverteilung in den Banken ungünstig verschieben. Das Scheren der DNA mithilfe einer Spritze (Schriefer et al. 1990; Hengen 1997) führte nur zu einer schwachen Fragmentierung und zu keiner Bevorzugung eines Größenbereiches und wurde deshalb nicht weiter verfolgt (Ergebnisse nicht dargelegt). Wie beim Hydroshear-Verfahren wird bei dieser Methode die DNA mit Druck durch eine kleine Öffnung gepresst. Eine modifizierte HPLC-Pumpe, wie bei Oefner et al. (1996) beschrieben, stand nicht zur Verfügung. Die Transformationseffizienz lag bei der Hydroshear- und Nebulizer-Methode ca. um den Faktor 100 unter den Ergebnissen, die mit der Ultraschallmethode erreicht wurden. Die Gründe hierfür liegen vermutlich in dem notwendig hohen Eingangsvolumen für die Fragmentierungstechniken von Nebulizer und Hydroshearesystem, die eine Einengung des Volumens für die folgenden Schritte notwendig macht. Die hierbei entstehenden Verluste werden die Hauptursache für die späteren niedrigen Klonausbeuten darstellen.

Beim Scheren mit Ultraschall wird die DNA zum Schwingen angeregt und zerbricht dabei in Abhängigkeit von der Schallmenge und Zeitdauer zufällig in kleinere Stücke. Dieses eingesetzte Scheren mit Ultraschall (Deininger 1983) zeichnet sich durch mehrere Vorteile aus: (1) die hohe Reproduzierbarkeit, (2) ein geringes Ausgangsvolumen, wodurch eine folgende Präzipitation mit Verlusten überflüssig wird, (3) das Fehlen von Überführungsverlusten, da direkt im verwendeten Reaktionsgefäß weitergearbeitet werden kann, (4) die Möglichkeit in einem geringen Reaktionsvolumen zu arbeiten, was bei der Größenselektion von Vorteil ist, (5) die hohen Klonmengen, die den Ligationserfolg im Vergleich widerspiegeln und (6) die weitgehend sequenzunabhängige Fragmentierung der DNA.

Die Fragmentgrößen schwanken in der Regel zwischen 0,8 bis 5 kb und lassen sich nach Modifikationen in *high copy* Vektorsysteme ligieren. Hierdurch werden die typischen Fragmentgrößen bei der Shotgun-Sequenzierung von 2-4 kb erreicht (Martin-Gallardo et al. 1994). Plasmide in diesem Größenbereich lassen sich noch problemlos im *high through put* präparieren. Ebenso können ihre Inserts im Hochdurchsatz mit der Polymerasekettenreaktion (PCR) amplifiziert werden. Die simultane Verarbeitung von mindestens 96 oder 384 Proben auf einer Mikrotiterplatte (MTP) kann bis zur Bestimmung der Sequenz beibe-

halten werden. Die Möglichkeit der Verwendung von HTS sind zur Bereitstellung der großen Mengen an einzelnen DNA-Sequenzen notwendig.

1.6 Bestimmung der Sequenz

Methodische und technische Entwicklungen ermöglichten das Etablieren der Shotgun-Sequenzierung, die durch Sanger et al. (1977) eingeführt wurde. Die zu Grunde liegende, als Kettenabbruch- oder Didesoxynukleotidverfahren bezeichnete Methode, stellt in ihren aufbauenden Entwicklungen die Basis aller Genomsequenzierungen dar. Als Template für die Shotgun-Sequenzierung dienten in der vorliegenden Arbeit isolierte Plasmide und/oder deren amplifizierte Inserts. Beim verwendeten Kettenabbruchverfahren wird durch Hitzedenaturierung die doppelsträngige DNA in Einzelstränge (*single stranded DNA*; ssDNA) aufgeschmolzen, die dann als Matrize für die Sequenzierreaktion dient.

In der Sequenzierreaktion wird durch eine gezielte statistisch verteilte Unterbrechung der Komplementärstrang-Synthese eine Population von unterschiedlich langen Einzelsträngen erreicht, die analysiert wird. Die Unterbrechung der Komplementärstrang-Synthese erfolgt durch den Zusatz von Didesoxynukleotiden (ddNTPs) zu dem Gemisch von Desoxynukleotiden (dNTPs). Didesoxynukleotiden fehlt die entscheidende 3'-OH-Gruppe, die zur Strangverlängerung notwendig ist, so dass die Synthese abgebrochen wird, wenn ein ddNTP in die wachsende Kette eingebaut wird. Diese Methode wurde durch die zyklische Sequenzierung (C*ycle Sequencing*) weiterentwickelt, die eine Kombination der Didesoxymethode und PCR darstellt. Hierbei werden die Kernschritte der PCR, die Denaturierung der DNA, die Hybridisierung eines Primers und Polymerisation zur Gewinnung einzelsträngiger, unterschiedlich langer und endmarkierter linearer Produkte genutzt (lineare PCR mit Kettenabbruch).

Die Markierung der Sequenzierprodukte erfolgte durch die Verwendung des Big Dye Terminator Systems (Applied Biosystems; Madison/US). Als DNA Polymerase wurde die AmpliTaq FS (*Fluorescent Sequencing*) verwendet. Dieses Enzym stellt eine Variante der *Thermus aquaticus* DNA Polymerase dar, die in der aktiven Domäne eine Punktmutation (F667Y) besitzt. Diese Mutation führt zu einer geringeren Diskriminierung der fluoreszenzmarkierten ddNTPs und so zu einem verbesserten Einbau. Eine weitere Punktmutation (G46D) resultiert im Verlust der Exonukleaseaktivität. Die ddNTPs sind beim Dye-Terminator System mit unterschiedlichen Fluoreszensdonoren versehen. Zur Verbesserung der Lesequalität und –weite beinhaltet der dNTP Mix dITP (2'-Desoxy-Inosin-5'-

Triphosphat) statt dGTP zum Vermeiden von Kompressionen sowie dUTP statt dTTP zum verbesserten Lesen von T-Stretchen (P/N 4390037).

Zur Auftrennung der Sequenzierprodukte standen Kapillarsequenzierer (3700; ABI, Branchburg/USA) zur Verfügung, die in den letzten Jahre die gelbasierenden Systeme abgelöst haben. Die Etablierung von Kapillarsequenzern ist erst in den letzten Jahren erreicht worden, jedoch bereits seit längerer Zeit in der Erprobung (Huang & Mathies, 1994). Die Generierung der Sequenz wird durch das Anregen des jeweiligen Fluoreszenzmarker (Dyes) beim Verlassen der Kapillaren mit einem Laser (480 und 514,5 nm Wellenlänge) erreicht. Die Emission wird mittels Spektrograph sowie CCD-Kamera dokumentiert und im Elektropherogramm dargestellt.

Ein hoher Probendurchsatz basierend auf kürzeren Trennzeiten der Sequenzierprodukte, einen hohen Grad der Automation und das Entfallen einer manuellen Nachbearbeitung der Daten sind nur einige Gründe, die zur schnellen Etablierung des Systems in Sequenzier-Laboratorien führten.

Für die Assemblierung großer Mengen von Einzelreads ist am MPI für Molekulare Genetik Berlin die PHRAP-Software (Philip Green, Univ. of Washington, Seattle, USA) etabliert. Dieser Assembler ist zu den vorhandenen Datenformaten kompatibel und hat sich auch bei der Assemblierung großer bakterieller Genome bewährt (Kaneko et al. 2001).

1.7 Datenanalysen ausgewählter Bereiche des Genoms

Nur in geringem Maße standen zur Analyse experimentelle Daten zur Verfügung (Rabus et al. 2002a). Im Gegensatz zu anderen Genomen (z.B. *Schizosaccharomyces pombe;* Wood et al. 2002) stand weder ein umfangreicher Datensatz aus dem Transkriptom oder Proteom zur Verfügung noch ein nahe verwandter gut charakterisierter Organismus. Eingehender untersucht wird in der vorliegenden Arbeit das zum Teil aufgelöste rRNA-Operon der Planctomyceten (Kap. 3.2.3), die Identifikation und die Verteilung der tRNAs (Kap. 3.2.4) und insbesondere die repetitiven Elemente des Genoms, die untereinander hohe Sequenzidentität zeigen. Die repetitiven Elemente führten zu Problemen bei der Assemblierung der Shotgun-Daten, weshalb ihre Identifikation und die Analyse ihrer kodierenden Information von besonderen Interesse waren. Für die genannten Bereiche standen keine weiteren experimentellen Daten zur Verfügung, so dass die Analysen durch Vergleiche mit den zur Verfügung stehenden Datenbanken *in silico* durchgeführt wurden (Kap. 2.5).

2. Material und Methoden

2.1 Herstellung der Banken

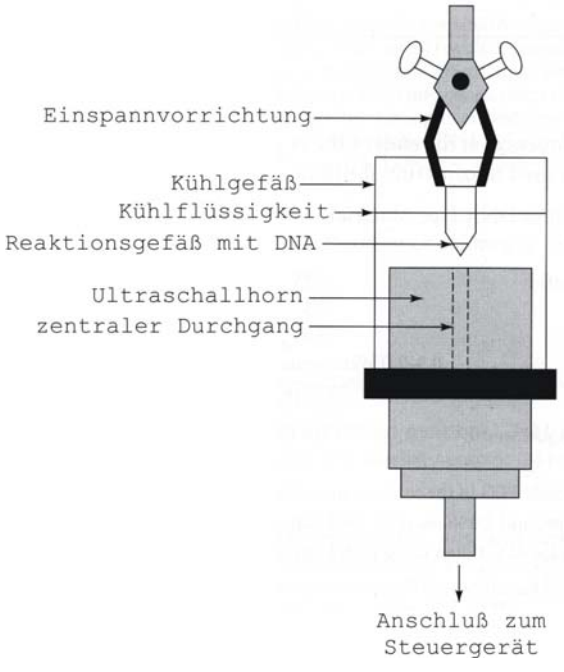
2.1.1 Kultivierung von *Pirellul*a sp. Stamm 1 und DNA-Isolierung

Pirellula sp. Stamm 1 wurde aus dem oberen aeroben Teil der Wassersäule in der Kieler Bucht isoliert (Schlesner 1994). Zellmaterial dieses Bakteriums wurde in den Arbeitsgruppen von Herrn Schlesner an der Universität Kiel und von Herrn Rabus am MPI für marine Mikrobiologie hergestellt. Die DNA Präparationen von Kulturen in der exponentiellen Wachstumsphase (Rabus et al. 2002) erfolgte am MPI für marine Mikrobiologie (Herr Rabus und Herr Wulf). *Pirellula* Kulturen wurden auch zur Anfertigung von selbst durchgeführten DNA-Isolationen (Genomic Tip System; Qiagen, Hilden/D) für Teilversuche genutzt (Herstellung der Cosmidlibrary, PCR-Reaktionen).

2.1.2 Scheren der genomischen DNA durch Ultraschall

Für die Herstellung der Shotgun-Banken wurden jeweils 10 µg genomische DNA in einem Volumen von 58 µl eingesetzt. Zum Scheren der DNA wurde ein Ultraschallhorn (Abb. 3) mit entsprechender Kontrolleinheit (Cell Disruptor B30, Branon Sonic, Heinemann Schwäbisch Gmünd) verwendet. Die DNA wurde für 16 sec. beschallt (Duty Cycle 40, Output 5), wobei das Reaktionsgefäß in Ethylenglycol bei 4°C gekühlt wurde.

Abb. 3: Ultraschallvorrichtung zur zufälligen Fragmentierung von DNA
Das Reaktionsgefäß mit der zu fragmentierenden DNA wurde direkt über dem Ausgang des Ultraschallhorns platziert. Während des Scherungsprozesses wurde die Probe mit Ethylenglycol (4°C) gekühlt. Der Versuchsaufbau entspricht dem von Birren et al. 1997 (Skizze aus Sambrook & Russel 2001).



2.1.3 Auffüllen der Fragmentenden

Zum Auffüllen der erhaltenen Fragmente bzw. zum Herstellen von *blunt-ends* (glatter Enden) wurden die Polymerasen *E. coli* DNA-Polymerase I/ Klenow (große Untereinheit; NEB, Frankfurt am Main/D) und die T4 Polymerase (Fermentas, St.Leon-Rot/D) verwendet. *E. coli* DNA-Polymerase I lässt sich in zwei Untereinheiten aufspalten, wobei die größere Untereinheit als Klenow Fragment bezeichnet wird. Auf der größeren Untereinheit sind folgenden Aktivitäten des Holoenzyms lokalisiert: die DNA-Polymerase sowie die 3'-5' Exonuklease Aktivität. Die 5'-3' Exonuklease Aktivität ist mit der kleineren Einheit assoziiert (Klenow & Henningsen 1970a; Klenow & Overgaard-Hansen 1970b).

Die Eigenschaften der T4-Polymerase ähneln der des Klenow-Fragmentes, jedoch ist die 3'-5' Exonuklease Aktivität deutlich höher (mehr als zweihundertfach). Die Reaktion muss dementsprechend bei hoher Konzentration von dNTPs ablaufen, um die Exonuklease zu kompensieren. Diese Exonuklease-Aktivität birgt jedoch auch die Gefahr, bei zu langen Inkubationszeiten die glatten Enden wieder zu verlieren. Die Kombination beider Enzyme resultiert in einer höheren Ausbeute an ligierbaren Fragmenten (Sambrook & Russel 2001). Die gescherte *Pirellula* DNA wurde mit den Polymerasen, dNTPs und Reaktionspuffer versetzt (Tab. 1) und für 30 min bei 23°C inkubiert.

Tab. 1: Auffüllreaktion der gescherten DNA Fragmente

Volumen	Komponente
58 µl	gescherte *Pirellula* sp. Stamm 1 DNA (10 µg)
7 µl	10 x Polymerase Puffer [670 mM Tris-HCl (pH 8,8), 66 mM $MgCl_2$, 10 mM DTT, 16 8mM $(NH_4)_2SO_4$
1,5 µl	0,5 mM dNTPs (10 0mM/µl; Amershan, New Jersey/US)
1,5 µl	T4-Polymerase (5 u/µl; Fermentas, St.Leon-Rot/D)
2 µl	Klenow large fragment (5 u/µl; NEB, San Carlos/US)
70 µl	Endvolumen

Die DNA wurde unmittelbar nach der Reaktion zur Größenselektionierung in einem 1%igem Agarosegel elektrophoretisch aufgetrennt.

2.1.4 Größenselektion und Isolierung von selektierten DNA-Fragmenten

Es wurden Gele mit 1% (w/v) Agarose (NEEO Agarose; Roth, Karlsruhe) in 1*TBE-Puffer (Sambrook & Russel 2001) hergestellt. Diese Gelzusammensetzung und der Laufpuffer wurden, wenn nicht anders ausgewiesen, auch in den anderen Experimenten verwendet. Die fragmentierte DNA wurde mit 15 µl Gel-Ladepuffer (Sambrook & Russel 2001) versetzt, der aufgrund der höheren Dichte das Absinken der DNA in die Geltaschen ermöglicht. Die Größenauftrennung wurde bei einer elektrischen Spannung von 8 V/cm für 20 min durchgeführt. Die Gelelektrophorese dient neben der Größenfraktionierung der DNA auch der Aufreinigung, da Enzyme und dNTPs abgetrennt werden.

Die Färbung der DNA erfolgte durch direkte Zugabe von Ethidiumbromid in das Gel (0,1 µg/ml). Ethidiumbromid interkaliert und färbt so die DNA an. Neben der Spur mit dem Größenstandard wurde jeweils eine Geltasche freigelassen, um Kontaminationen zu vermeiden. Unter UV-Licht (312 nm) wurde die angefärbte DNA sichtbar gemacht. Hierzu wurde das Gel in einer Haushaltsschale auf den UV-Transiluminator (Roth, Karlsruhe) gelegt. Unter Verwendung eines Skalpells wurde ein Gelbereich ausgeschnitten, der mittig zwischen den ein und zwei Kilobasen Banden des Markers und unterhalb der vier Kilobasen Bande des Größenstandards Markers lag. Die Verwendung der Haushaltschale erlaubte eine verminderte UV-Belastung der DNA. Das Ausschneiden der DNA beinhaltenden Agarose erfolgte zügig, um die UV-Exponierung der DNA so kurz wie möglich zu halten. Die Ergebnisse der Gelelektrophorese wurden mit einem Videosystem (Cybertech, Berlin) dokumentiert.

Zur Isolierung der DNA aus den Agarosestücken wurde der Easypure Kit (Biozym, Hessisch Oldendorf) verwendet. Die Aufreinigungsmethode basiert auf der Bindung von DNA in Gegenwart hoher Natriumjodidkonzentrationen an eine Silikatmatrix (Vogelstein & Gillespie 1979). Hierbei wird das Gel unter Hitzeeinwirkung zunächst in NaI aufgeschmolzen, die DNA an eine Silikatmatrix gebunden, gewaschen und von der Matrix mit 20 µl H_2O wieder eluiert. Abweichend zu den Angaben des Herstellers wurde die Menge der eingesetzten Silikatmatrix auf 9 µl erhöht und zur Elution nur 20 µl doppelt destilliertes Wasser verwendet.

Zur Überprüfung wurden 2 µl der eluierten DNA auf ein 1%iges Agarosegel gegeben und mit einer 1 kb Leiter auf ihre Größe und mit einem Massenmarker auf ihre ungefähre Konzentration überprüft. OD- Messungen zur Überprüfung der DNA-Konzentration lassen sich nicht zuverlässig durchführen, da bereits kleinste Mengen an Silikatmatrix zu einer nicht abschätzbaren Fehlerquelle werden.

2.1.5 Ligation der größenselektierten DNA

Die Vervielfältigung der Shotgun-DNA durch Klonierung erfolgt *in vivo*. Der Plasmidvektor pUC19/SmaI BAP (Fermentas, St.Leon-Rot) wurde für die Klonierungen eingesetzt. Der Vektor pUC19 (Yanish-Perron et al. 1985) ermöglichte die Verwendung der Ampicillinresistenz sowie die blau/weiß Färbung der Kolonien als Selektivmarker.

Das Auftreten von blau oder weiß gefärbten *E .coli* DH10B Kolonien in Gegenwart von IPTG und X-Gal ist auf die Promotersequenz des LacZ-Gens und die kodierende Sequenz der amino-terminalen 146 Aminosäuren der β-Galactosidase im pUC19 Vektor zurückzuführen. Das kleine amino-terminale Fragment der β-Galactosidase kodiert durch den pUC-Vektor wurde in elektrokompetente *E. coli* DH10B transformiert. Dieser Stamm verfügt über keine β-Galactosidase-Aktivität. Das amino-terminale Fragment (α-Fragment) kann verschiedene β-Galactosidase negative Mutanten komplementieren. α-Komplementation (Ullmann et al. 1967) tritt auf, wenn pUC-Plasmide in Bakterienstämme wie *E.coli* DH10B gebracht werden, die ein inaktiviertes carboxy-terminales Fragment der β-Galactosidase (ω-Fragment) besitzen. Die *multiple cloning site* (MCS) mit der verwendeten SmaI Schnittstelle liegt innerhalb des α-Fragmentes. Das Insertieren eines Shotgun-Fragmentes in den Bereich des α-Fragmentes führt zur Unterbrechung der kodierenden Sequenz, wodurch die α-Komplementation herabgesetzt wird oder unterbleibt; es kommt zu deutlich herabgesetzter oder ausbleibender ß-Galactosidase-Aktivität. Im Gegensatz hierzu stehen Klone, die kein Insert aber den rezirkulierten Vektor tragen, da sie in der Lage sind X-gal zu hydrolisieren, wodurch eine Blaufärbung der Kolonien auftritt (Horwitz et al. 1964; Davies & Jacob 1968).

Die MCS wurde durch die Verwendung der Restriktionsendonuklease SmaI (Position 271) geöffnet und ermöglichte so die Klonierung von *blunt end* Fragmenten. Durch die Dephosphorilisierung der Enden wird eine Religation des Vektors vermieden. Das Vektorsystem hatte sich in der Vergangenheit bei der Erstellung von Shotgun-Banken bereits bewährt (The chromosome 21 mapping and sequencing consortium, 2000; Seo et al. 2001; Rabus et al. 2002b). Zur Ligation wurden 100 ng *Pirellula* DNA und 20 ng Vector-DNA eingesetzt.

Tab. 2: Zusammensetzung der Ligationsansätze

Volumen	Komponente
7,2 µl	*Pirellula* Fragmente (100 ng)
1,2 µl	pUC19/SmaI/BAP (20 ng)
1,2 µl	10 x Ligationspuffer (500 mM Tris-HCl, pH 7.5; 100 mM $MgCl_2$; 100 mM DTT)
1,2 µl	5 mM ATP
1,2 µl	Ligase (1,2 u)
12 µl	Endvolumen

Es wurde jeweils ein Ligationsansatz für die 1,5 kb und die 3,5 kb großen Fragmente hergestellt (Tab. 2). Die Ligationsansätze wurden für 16 Stunden bei 16°C im Wasserbad inkubiert. Die erhaltenden rekombinanten DNA-Moleküle wurden in Wirtszellen eingeschleust.

2.1.6 Elektroporation

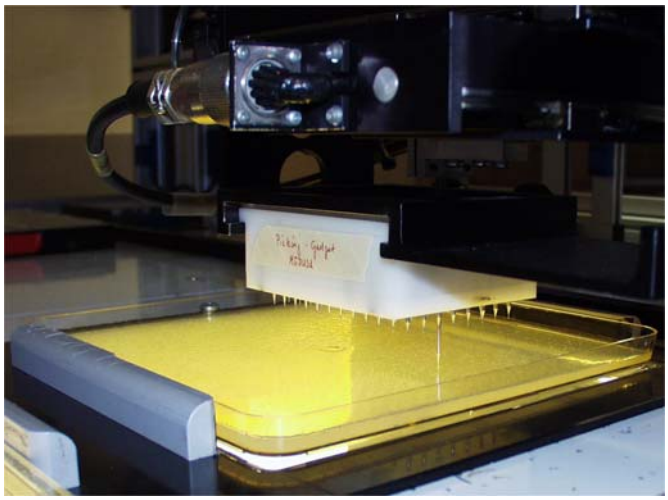
Die Plasmide wurden durch Elektroporation in kompetente Bakterienzellen überführt. Bei der Elektroporation werden die *E. coli* Zellen elektrischen Entladungen ausgesetzt, die zur reversiblen Destabilisierung ihrer Membranen führt und vorübergehend unter anderem die Formation von Membranporen induziert (Neumann & Rosenheck 1972; Neumann et al. 1982; Wong & Neumann 1982). Die Ligationsansätze wurden mit 2 µl Chloroform versehen und vorsichtig durchmischt. Anschließend wurden die Phasen durch Zentrifugation getrennt. Der Überstand wurde dann in ein frisches Reaktionsgefäß überführt. Die Ligationsansätze wurden anschließend auf Nitrocellulose Dialyse Filter (Millipore, Schwalbach) gegeben und 15 min zur Verringerung des Salzgehaltes gegen zweifach destilliertes Wasser dialysiert. Die Elektroporation erfolgte bei 200 Ohm, 25 µF und 2,5 kV (Gene Pulser; BioRad, München) in Küvetten (BioRad, München) mit einem Elektrodenabstand von 0,2 cm. Zur Elektroporation wurden 2 µl des Ligationsansatzes und 40 µl Electromax *E. coli* DH10B Zellen (Invitrogene, Karlsruhe) verwendet. Experimente zur Abschätzung der möglichen Ausbeute an Klonen wurden zunächst mit selbst hergestellten elektrokompetenten Zellen des genannten Bakterien Stammes (Sambrook & Russel 2001) durchgeführt, die nur eine Effizienz von 10^9 cfu/µg aufweisen. Die Küvetten wurden auf 0°C vorgekühlt und der Transformationsansatz nach der Elektroporation sofort mit vorgewärmten (Rabussay et al. 1987) 960 µl SOC Medium (Sambrook & Russel 2001) versetzt. Die Suspension wurde zum Regenerieren der Zellen und Aufbau der Antibiotikaresistenz bei 37°C und

225 rpm für 45 min inkubiert (Sambrook & Russel 2001). Das anschließende Ausplattieren erfolgte zur gleichmäßigeren Verteilung mit Glasperlen (2,85-3,3 mm; Roth, Karlsruhe). Überdimensionale Petrischalen (23 x 23cm; Genetix, Dorset/UK) wurden zum Vereinzeln und als Vorbereitung für den späteren Transfer der Klone verwendet. Der als Substrat verwendete LB-Agar (Sambrook & Russel 2001) wurde durch Zugabe von Ampicillin (100 mg/l; Sigma, Taufkirchen), X-GAL (70 mg/ml; Roth, Karlsruhe) und 1 mM IPTG (Roth, Karlsruhe) als Selektivmedium gestaltet (Sambrook & Russel 2001).

2.1.7 Überführung der Klone in Kulturmedien

Die Kolonien wurden mit Hilfe des Pickingroboters Medusa (MPI für Molekulare Genetik Berlin; Vorläufer des Q-BOT; Genetix, Dorset/UK; Abb. 4) in Flüssigkulturen überführt. Die Agarplatte mit den Klonen wird zunächst mit einer Videokamera im Pickingroboter aufgenommen. Ausgehend vom digitalisierten Bild werden die Koordinaten jedes Klones auf der Agarplatte errechnet. Eine Qualitätsbestimmung jedes Klones wird durch die Rundheit (Ausschluss von Kolonien, die miteinander verwachsen sind), den Mindestdurchmesser (Ausschluss von evt. vorhandenen Satelliten) und des Helligkeitswertes (Ausschluss von blauen Klonen) für jeden Klon erreicht.

Abb. 4: Pickingroboter Medusa
Die Abbildung zeigt den Pickingroboter beim Abnehmen einer Kolonie von der Agarplatte. Eine Nadel von insgesamt 96 ist ausgefahren. Im Anschluss überführt der Roboter die Klone in 384er MTPs mit Flüssigmedium.



Die abgenommenen Klone wurden in jeweils 70 µl Gefrier-Medium (Tab. 3) in 384er MTPs aufgenommen und für 16 Stunden bei 37°C inkubiert. Zur weiteren Verwendung der Kulturen wurde mit 384er-Replikatoren (Genetix, Dorset/UK) eine Kopie der 384er MTPs

gezogen. Für die Kopien wurde LB-Medium mit 100 mg/l Ampicillin verwendet, was zu besseren Ergebnissen bei der folgenden Amplifikation führt.

Tab. 3: Zusammensetzung des Gefrier-Mediums
Das verwendete Gefrier-Medium enthält Nährmedium (LB; Sambrook & Russel, 2001), Selektionsmarker und HMFM (Hagness modified freezing medium; Genetix). HMFM ermöglicht nach erfolgter Inkubation der Kulturen ein Aufbewahren bei -80°C, wodurch auf die Kulturen jederzeit zurückgegriffen werden kann. HMFM wird aus zwei Komponenten zusammengesetzt, die nach dem autoklavieren zu einer 10-fach konzentrierten Stammlösung gemischt werden (8 Teile der Komponente „A" und 2 Teile der Komponente „B"). Die HMFM Lösung wird dann im LB Nährmedium (Sambrook & Russel 2001) auf einfache Konzentration verdünnt und mit Ampicillin (100 mg/l „Freezing"-Medium) versetzt.

HMFM Komponente A		**HMFM Komponente B**	
Magnesiumsulfat-Heptahydrat:	0,9 g	Kaliumdihydrogenphosphat	18 g
Ammoniumsulfat:	9 g	Kaliumhydrogenphosphat-Trihydrat:	47 g
Glycerol (96%):	440 g		
Auf:	800 ml	Auf:	200 ml

LB (Luria-Bertani Medium)	
NaCl	9 g
Tryptone (Difco; Le Pont de Claix/F)	9 g
Yeast Extract (Difco ; Le Pont de Claix/F)	4,5 g
Auf:	900 ml

2.1.8 Weitere Absicherung der Sequenz durch die Cosmidlibrary

Die DNA von *Pirellula* sp. Stamm 1 wurde mit dem Genomic Kit (Genomic-tip 500-G, Qiagen, Hilden/D) mit Anionen-Austauscher-Säulen nach dem Herstellerprotokoll präpariert. Dieses Verfahren erlaubt die Isolation von DNA-Fragmenten mit einer durchschnittlichen Größe von 50-100 kb. Durch zusätzliches Vortexen wurde die DNA auf einen durchschnittlichen Größenbereich von 40-60 kb eingestellt, so dass die generierten Fragmente im gewünschten Größenbereich für die folgende Klonierung (34-45 kb) lagen. Die erhaltenen Fragmente wurden mit dem pWEB Cloning Kit (Biozym, Hessisch-Oldendorf/D) nach den Herstellerangaben in Cosmide überführt, wobei die Einengung der zu ligierenden Fragmente durch die Verwendung des Membransystemes membra-Spin Mini (membraPure, Bodenheim/D) erfolgte, die ein verlustfreieres Arbeiten als die vorgeschlagene Präzipitation mit Ethanol ermöglicht.

Das gewählte pWEB Klonierungssystem bietet als entscheidenden Vorzug gegenüber der traditionellen Herstellung von Cosmidbanken die Klonierung von zufällig erzeugten Fragmenten (Fiandt 1998). Die zeitaufwendige Präparation von hochmolekularer DNA und Vorversuche zum partiellen Schneiden der genomischen DNA mit Restriktionsendonukleasen entfällt hierbei, ebenso wird die ungleichmäßige Repräsentanz der Fragmente in der resultierenden Cosmidbank herabgesetzt.

Die erhaltene Cosmidbank wurde, wie für die Shotgun-Banken beschrieben, in Flüssigmedium überführt (Kap. 2.1.7). Die Präparation der Cosmide erfolgte, wie für die Plasmide beschrieben (Kap. 2.2.4); das Kulturvolumen wurde jedoch auf 8 ml erhöht. Die verhältnismäßig hohe Kopienzahl von 5-20 Kopien in *E. coli* erleichterte die Präparationen, wodurch durchschnittlich 2 µg DNA gewonnen wurden. Die Sequenzierung erfolgte nach den Angaben zur Sequenzierung von Cosmiden mit dem Big Dye Terminator System (Applied Biosystems; Madison/US).

2.2 Bereitstellung der Templates für die Sequenzierung

2.2.1 Anwendung der PCR zur standardisierten Insertamplifikation

Die Polymerasekettenreaktion (*polymerase chain reaction*; PCR) wurde als *in vitro* Verfahren zur gezielten Amplifikation der klonierten Fremd-DNA genutzt. Hierbei wurden sogenannte M13 Primer zur Amplifikation verwendet (Tab. 4), die in den Randbereichen der MCS des verwendeten pUC19 Vectors hybridisieren. Die Amplifikation von Insert-DNA wird seit der Einführung von thermostabilen Polymerasen durchgeführt (Saiki et al. 1988) und hat sich auch bei Projekten wie der Sequenzierung des humanen Chromosoms 21 (The chromosome 21 mapping and sequencing consortium 2000) bewährt.

Tab. 4: Für die standardisierte PCR verwendete Primer

Amplifikations-primer	Sequenz (5'-3')	Hybridisirungs-position auf pUC19	Schmelz-temperatur
pUC/M13 (forward Primer*)	CCC CAG GCT TTA CAC TTT ATG CTT CCG GCT CG	520-551	67,4°C
pUC/M13 (reverse Primer*)	GCT ATT ACG CCA GCT GGC GAA AGG GGG ATG TG	296-327	68,7°C

(*Radelof et al. 1998)

Die Primer wurden zu den anderen PCR-Komponenten gegeben und zu einem PCR-Mastermix vermischt (Tab. 5). Die thermostabilen DNA Polymerasen Taq (Ursprung

Thermus aquaticus) und Pfu (Ursprung *Pyrococcus litoralis*) werden als Enzymgemisch in der PCR eingesetzt. Die Vorteile dieses Enzymmixes spiegeln sich in einer erhöhten Ausbeute an PCR-Produkt wider, die vermutlich in der *proofreading* Eigenschaft der Pfu begründet ist. Die Fehlerrate von Pfu wird um den Faktor 7-10 niedriger als bei Taq angegeben (Cline et al. 1996). Die Amplifikation der kleinen Inserts lässt sich problemlos durchführen, während sich bei den größeren Inserts die Ausfälle bei alleiniger Verwendung von Taq häufen. Polymerasengemische haben sich in der Vergangenheit zur Amplifikation größerer DNA-Abschnitte (*long-range* PCR) bewährt. Die Zugabe eines geringen Anteils eines *proofreading* Enzyms zum Reaktionsmix erhöht die Leistungsfähigkeit der *non-proofreading* Polymerase durch die Korrektur von Fehlern bei der Neusynthese. Die Folge ist eine höhere Effizienz bei der Amplifikation der gesamten Länge des PCR-Produktes (Barnes 1994). Der pH 9 des PCR-Puffers gewährleistet optimale Bedingungen für die *proofreading* Aktivität der Pfu (Cline 1996). Das Unterschreiten der optimalen Arbeitstemperatur zwischen 70-80°C der thermostabilen Polymerasen hat sich in höheren Ausbeuten der PCR-Produkte positiv widergespiegelt und unterstützt die Hypothese, dass die optimale Temperatur für die erfolgreiche Polymerisation eher durch die Stabilität des Templates als durch die des Enzyms bestimmt wird (Dabrowski & Kur 1998). Der Zusatz von Betain (N,N,N-trimethylglycine; Sigma, Taufkirchen) als Enhancer wurde in die standardisierte PCR integriert. Betain dient der Reduzierung von Sekundärstrukturen, die auf GC-reiche Sequenzen zurückgehen (Henke et al. 1997).

Tab. 5: PCR-Mastermix der standardisierten PCR

Reagenz	Finale Konzentration im PCR-Reaktionsmix
M13 forward Primer	0.3 µM
M13 reverse Primer	0.3 µM
dNTP's	0,3 µM
10 x PCR-Puffer	20 mM $(NH_4)_2SO_4$ 75 mM Tris-HCl, pH 9,0 0.01% Tween 20 2,5 mM MgCl
$MgCl_2$	2,5 mM
Betain	0,5 M
Taq	2 u/rxn
Pfu	0,05 u/rxn

Die Verteilung des PCR-Mastermixes erfolgte durch die Verwendung eines automatisierten Pipettiersystems [siehe Kap. 2.2.2; Hydra-Twister Kombination mit Transferarm

(Zymark, Idstein) und einer modifizierten Hydra [(Robbins, Sunnyvale/USA); Ansteuerungssoftware und Verknüpfung der Systeme MPI für Molekulare Genetik Berlin].

Die PCR wurde in Thermocyclern (9700, Applied Biosystems, Norwalk/USA) durchgeführt, die 384er MTP (Thermo-Fast 384; Abgene House, Ashford/UK) verarbeiten können. Der Reaktionsansatz für die einzelne standardisierte PCR betrug 20 µl. Die DNA-Matrize für die PCR wurde in Form eines Teils der Bakterienkultur bereitgestellt. Hierbei wurde durch das Eintauchen von Replikatoren (384 PIN Replicators; Genetix, Dorset/UK) in die Bakterienkultur und anschließendes Eintauchen in die mit PCR-Reaktionsmix versehenden Mikrotiterplatten ein Teil der Kultur überführt. Die befüllten 384er MTPs wurden mit Folien abgedichtet (Micro Amp, Applied Biosystems; Norwalk/USA) und in den Thermocyclern inkubiert. Die Parameter für die PCR (Tab. 6) wurden an die jeweiligen Insertgrößen angepaßt, wobei sich die verhältnismäßig langen Polymerisationszeiten in erhöhten Ausbeuten positiv niederschlugen. Die PCR-Produkte wurden bis zur Weiterverarbeitung bei 4°C aufbewahrt.

Tab. 6: PCR-Parameter für die Amplifikation

	Initiale Denaturierung	*Denaturierung*	*Hybridisierung*	*Extension*	*Finale Extension*
∅1,5 kb Insertgröße	94 °C 5 min	94 °C 40 sek	63 °C 30 sek	68 °C 4 min	68 °C 6 min
∅3,5 kb Insertgröße	94 °C 5 min	94 °C 40 sek	63 °C 30 sek	68 °C 6 min	68 °C 8 min
		35 Zyklen			

2.2.2 Analyse der PCR-Produkte

Unter Verwendung eines automatisierten Pipettiersystems (modifizierte Hydra-Twister Kombination, Abb. 5) wurde 1 µl aus dem PCR-Ansatz mit 3 µl Loading-Dye vermischt und direkt in ein 1%iges Agarose-Gel pipettiert (Abb. 6). Als Laufpuffer wurde ein 1-fach TBE Puffer verwendet. Die Anfärbung erfolgte durch Zugabe von Ethidiumbromid in das Gel. Anschließend wurde das mit den 384 aufgetragenen PCR-Aliquots versehene Gel in eine mit Laufpuffer (1-fach TBE) vorbereitete Gelkammer gegeben. Als Größenstandard wurde ein aus PCR-Produkten hergestellter Größenstandard (0,5-4 kb) verwendet.

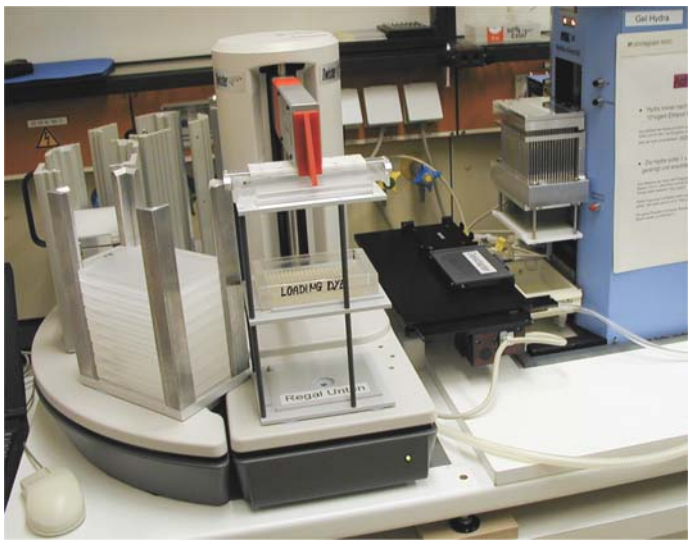


Abb. 5:
Twister-Hydra-Kombination zum Beladen der 384er Gele
Auf der rechten Abbildung ist der mit einem Gelträger beladene Twister-Arm (rote Achse) beim Transfer eines 384er Geles zum Pipettierroboter Hydra (blau) dargestellt. Im linken Bildausschnitt sind die Regale mit den PCR Platten und Gelträgern zu sehen. Im rechten Bildausschnitt sind die Hydra mit vorgelagerten beweglichen Arbeitstisch (schwarz) abgebildet. Der Loading Dye befindet sich auf dem Regal in der Bildmitte.
Der Pipettierroboter erhält von der Twisterkomponente die MTPs mit den PCR-Produkten, den Trog und die Gele im Gelträger zum Pipettieren. PCR-Produkte, ein Luftpolster und Ladepuffer werden nacheinander in den Spritzen der Hydra aufgezogen und dann gemeinsam in das Gel abgegeben. Acht Gele können so hintereinander ohne manuelle Eingriffe beladen werden.

Bei dem verwendeten Gelkammersystem handelt es sich um eine Diagonalelektrophorese (MPI für. Molekulare Genetik Berlin). Die Proben wurden bei 70 V für 50 min aufgetrennt. Proben, die kein oder ein in der Größe deutlich abweichendes PCR-Produkt zeigten, wurden verworfen. Alle verbleibenden PCR-Produkte wurden neu arrangiert (Xme, Software MPI Berlin; Pipettiergerät BioRobot 9600; Qiagen, Hilden), so dass 384er MTPs entstanden, die nur die gewünschten PCR-Produkte enthielten. Abschließend wurden die neu arrangierten PCR-Produkte (5 µl) mit 15 µl Wasser versetzt (Hydra; Robinson, Sunnyvale/USA).

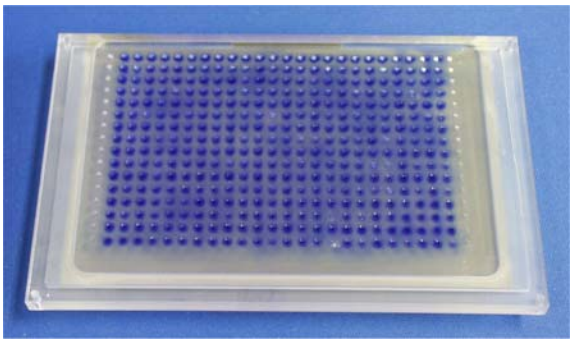


Abbildung 6: 384er-Gel
Agarosegel für die Diagonalelektrophorese mit 384 PCR-Proben beladen. Der Ladepuffer wurde mit Xylenblau angefärbt. An den Randbereichen befinden sich die noch nicht befüllten Taschen für die Größenstandards, die erst nach dem Eintauchen des Geles in den Laufpuffer aufgetragen wurden (vergleiche Kap. 3.1.1, Abb. 12).

2.2.3 Schließen von *Physical Gaps*

Einzelne Contigs ließen sich unter Verwendung der Readpair-Informationen zu großen, noch nicht lückenlosen, Super-Contigs zusammenstellen. Lücken, die in der Sequenz nicht durch Klone abgedeckt wurden (*Physical Gaps*), konnten mit Hilfe der PCR auf genomischer DNA geschlossen und die Anordnung der Super-Contigs geklärt werden. Die Primer wurden in der GAP4 Datenbank an den Enden der verbliebenden Super-Contigs zusammengestellt, wobei ein TM-Wert von 60-63 °C und eine Länge von 20-23 Basen für die Oligonukleotide angestrebt wurde. Die designten Primer sind in der genomischen Sequenz einmalig und lagen mindestens 500 bp vom Ende des jeweiligen Super-Contigs entfernt. Diese Primer wurden in allen möglichen Varianten miteinander kombiniert (Primersequenzen, Anhang Kap. 7.3, Tab. 73). Hierzu wurde der XL-Kit (Perkin Elmer, Branchburg/USA) verwendet, der sich zur Amplifikation direkt von genomischer DNA schon mehrfach bewährt hatte. Das Ansetzen der Reaktion erfolgte nach den Angaben des Herstellers mit ca. 800 ng genomischer DNA *Pirellula* sp. Stamm 1 pro Reaktion. Nach Optimierung der Hybridisierungszeit und einer verlängerten Polymerisierungszeit gelang die Amplifizierung reproduzierbar (Tab. 7). Die Enden der erhaltenen PCR-Produkte wurden im Anschluss sequenziert, die internen Sequenzen durch *Primer Walking* bestimmt.

Tab. 7: Verwendete PCR-Parameter zum Schließen der physikalischen Lücken

Primäre Denaturierung	94°C für 5 min
35 Zyklen	Denaturierung: 94°C für 90 sek Hybridisierung: 63°C für 30 sek Extension: 68°C für 12 min
Finale Extension	68°C für 12 min
Beendigung der Reaktion	4°C

(2400 Thermocycler, Applied Biosystems; Brachburg/USA)

2.2.4 Plasmid-Präparationen

Die Aufreinigung von Plasmiden erfolgte mit Hilfe des Multimac 96 Pipettiersystems (Beckman; Palo Alto/US) in Kombination mit dem Wizard SV 96 Plasmid DNA Purification System (Promega; Madison/US). Die Plasmidisolierung erfolgte nach den Herstellerangaben in drei Schritten, wobei die unter Tab. 8 angegebenen Puffer verwendet wurden. Zunächst wurden die Zellen lysiert (alkalische Lyse inklusive RNAase Inkubation), das Zellmaterial durch Filtration abgeschieden und die DNA in Waschschritten von Salzen gereinigt.

Tab. 8: Zur Plasmidaufreinigung verwendete Puffer:

P1 (Resuspensionspuffer)	50 mM 10 mM 100 g/ml	Tris-HCl, pH 8,0 EDTA RNase A
P2 (Lysispuffer)	200 mM 1 %	NaOH SDS
P3 (Neutralisationspuffer)	4090 mM 759 mM 2120 mM	Guanidin Hydrochlorid Kaliumacetat Essigsäure

Die Klone wurden durch Transferieren von 5 µl der jeweiligen 384er MTP Kultur in 1,5 ml Kulturmedium [2YT-Medium (Sambrook & Russel2001) mit Amp 100 $^{mg}/_{ml}$] im 96er Deepwellblock (Qiagen; Hilden/D) als Übernachtkultur (37°C, 260 rpm, 16 Stunden) angezogen. Nach der Inkubation wurden die Zellen bei 3200 x g für 10 min sedimentiert und der Überstand dekantiert. Die Bakterienpellets wurden jeweils in 80 µl P1-Puffer resuspendiert. Zu den Bakteriensupensionen wurden 80 µl P2-Puffer gegeben, kurz durchmischt und für 5 min inkubiert. Im Anschluss wurden jeweils 80 µl P3-Puffer zugegeben und das Lysat erneut durchmischt. 240 µl dieses Lysats wurden auf die Filterplatte transferiert und durch Zentrifugation (10 min bei 3200 x g) in eine 96er MTP filtriert. Das geklärte Lysat wurde aus der 96er MTP auf die Millipore Bindeplatte transferiert und erneut zentrifugiert (5 min bei 3200 x g). Die an die Filterplatte gebundene DNA wurde zweimal mit 200 µl 80% Ethanol gewaschen (3200 x g für 5min) und dann bei 70°C für 20 min getrocknet. 70 µl Elutionspuffer (10 mM Tris-HCl pH 8,5) wurden zum Lösen der DNA auf die Bindeplatte gegeben und nach 3 min Inkubationszeit bei 3200 x g für 10 min in eine 96er MTP eluiert. Die durchschnittliche Ausbeute lagen bei 8-10 µg DNA.

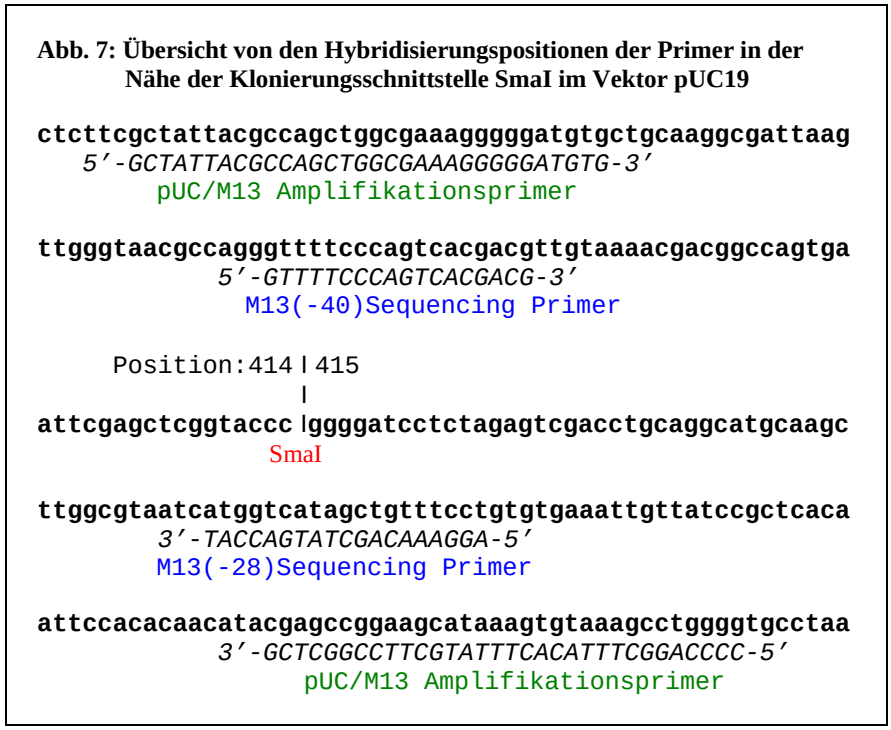
2.3 Sequenzierungen

2.3.1 Standardisierte Sequenzierung

Ausgehend von einer Genomgröße zwischen 4,7-8 Mb (Kölbel-Boelke et al. 1985) wurde mit einer Anzahl von mindesten 80000 Sequenzierreaktionen bei einer durchschnittlichen Leselänge von 400 b gerechnet. Hieraus resultiert bei einer Genomgröße von 5 Mb ein *Sequencing Coverage* von 6,4. Es wurde angestrebt, die Endsequenzen jedes Inserts von beiden Seiten zu bestimmen. Die Assemblierung aller Endsequenzen wurde kontinuierlich durchgeführt, um die Shotgun-Sequenzierung im Bereich nicht mehr absinkender Contiganzahl abzubrechen (Kap. 3.1.2).

Die Sequenzierung erfolgte mit Modifikationen nach der ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction Kit (P/N 4303151, Perkin Elmer); Anleitung für Sequenzierprodukte. Die zur Sequenzierung verwendeten Primer liegen im Gegensatz zu den Amplifikationsprimern in geringerem Abstand zur Klonierungsschnittstelle (Abb. 7). Dadurch wurde die sequenzierte Vektorsequenz minimiert, ein Erkennen von Vektor und Sequenzierrichtung jedoch ermöglicht.

Abb. 7: Übersicht von den Hybridisierungspositionen der Primer in der Nähe der Klonierungsschnittstelle SmaI im Vektor pUC19



Zur Durchführung der *Cycle Sequencing Reaction* wurden zunächst der jeweilige Sequenzierprimer, Wasser und BigDye (Amplittaq, dNTPs, ddNTPs und Reaktionspuffer) zum Sequenzier-Mastermix vermischt.

Tab. 9: Ansatz für die einzelne Sequenzierreaktion in der 384er MTP

Zusammensetzung:

1 µl Primer (5 pmol)
M13(-40) für die FDT (*forward*) oder M13(-28) für die RDT (*reverse*) Reaktion
2 µl zweifach destilliertes Wasser
1 µl Big Dye
1 µl verdünntes PCR-Produkt

Der Sequenzier-Mastermix wurde durch die Verwendung eines automatisierten Pipettiersystems [Hydra-Twister Kombination (Transferarm, Zymark; Hydra, Robbins; Ansteuerungssoftware und Verknüpfung der Systeme, MPI für Molekulare Genetik Berlin)] auf 384er MTPs verteilt. In diese vorpipettierten Platten wurde aus den neu arrangierten PCR-Platten von jeder PCR-Reaktion 1 µl übertragen (Tab. 9). Die Reaktionen wurden anschließend in der Zentrifuge (Eppendorf 5810R; Eppendorf, Hamburg/D) bei 3200 x g gesammelt und im Thermocycler inkubiert (Tab. 10).

Tab. 10: Parameter für die zyklische Sequenzierung (Cycle Sequencing)

	Prä-Denaturierung	Denaturierung	Hybridisierung	Extension	Aufbewahrung im Cycler nach der Reaktion
PCR-Produkte	-	96°C 10 sec	*°C 5 s	60°C 4 min	10°C
		25 Zyklen			
Plasmide	96°C 3 min	96°C 20 sec	*°C 10 sec	60°C 4 min	10°C
		35 Zyklen			

* Die Hybridisierungstemperatur für den M13(-40) (FDT) Primer beträgt 55°C und für den M13(-28) (RDT) Primer 50°C.

Die Sequenzierprodukte wurden durch eine Ethanolfällung zunächst präzipitiert und im Anschluss dreimal gewaschen (Tab. 11). Da ein direktes Dekantieren der jeweiligen Überstände in der 384er MTP nicht möglich war, wurden die Überstände durch kurzes abzentrifugieren auf Zellstoff entfernt.

Tab. 11: Konditionen für Präzipitation und Waschschritte

Ethanol Präzipitation	15 µl 100% Ethanol 1 h bei 20°C und 3200 x g
Waschschritte	20 µl 70% Ethanol 30 min bei 20°C 3200 x g

Durch die Umfällung wurde der Salzgehalt der Proben deutlich verringert. Die Proben wurden unter Vakuum getrocknet und im Anschluss in 10 µl Wasser gelöst.

2.3.2 Schließen von *Sequencing Gaps* und selektiertes Nachsequenzieren

Nach der Assemblierung der Daten auftretende Sequenzlücken wurden mit Hilfe des *Primer Walkings* auf den überspannenden Klonen geschlossen (*Sequencing Gaps*). Als Templates wurden entsprechende PCR-Produkte bzw. aufgereinigte Plasmide gewählt. Plasmiden wurde der Vorzug gegeben, da hier größere Leseweiten bei der Sequenzierung erreicht wurden. Diese Vorgehensweise wurde auch zum Erreichen der angestrebten Sequenzqualität von drei unabhängigen Reads, wobei beide Stränge der DNA sequenziert wurden, gewählt.

Die Auswahl der Primer geschah im Datenbank-Editor GAP4 (Bonfield et al. 1998), wobei ein TM-Wert von 60-63°C für die Oligonukleotide angestrebt wurde. Die Auswahl von Primern innerhalb dieses Temperaturbereiches war in der Regel möglich und die hohe Hybridisierungstemperatur ermöglichte eine hohe Spezifität der Primer, so dass ggf. auch eine Verwendung als PCR-Primer möglich war. Ein Verzeichnis aller im Rahmen des *Primer Walking* zur Absicherung der repetitiven Elemente verwendeten Primer befindet sich im Anhang (Anhang, Kap. 7.3, Tab. 73).

2.3.3 Auftrennung der Sequenzierprodukte

Die Sequenzierprodukte wurden auf einem Kapillarsequenzierer (3700 Capillary System; Applied Biosystems, USA) aufgetrennt und die Sequenz bestimmt. Die Verwendung dieses Systems ermöglichte einen hohen Grad der Automation, wodurch die Möglichkeit manueller Fehler herabgesetzt wurde. So wurden z.B. die zur Auftrennung verwendete Matrix und die Sequenzierproben vorgelegt. Entscheidend war die erreichte Leistungskapazität. Innerhalb von drei Stunden konnte die Sequenz von 96 Proben bestimmt werden. Ohne manuelle Eingriffe am Gerät konnte die Sequenz von insgesamt 1536 Proben bestimmt werden. Methodisch bestehen hier Unterschiede in der Art der Auftrennung gegenüber gelbasierenden Systemen. Zu den prägnanten Unterschieden gehörten: (1) das Kapillarsystem, welches mit dem flüssig bleibenden linearen Polymer POP6 (Applied Biosystems, Madison/USA) gefüllt wird, (2) das Anlegen einer weitaus höheren Spannung (begründet durch die verwendeten Kapillaren; 600-700 mA bei 6,8 kV konstant) und (3) das elektrokinetische Beladen der Kapillaren bei 1000 V für 20-40 sek in Abhängigkeit vom individuellen Gerät (P/N 4309125, Perkin Elmer).

2.4 Bioinformatische Methoden

2.4.1 Zusammenführen der genomischen Sequenz

2.4.1.1 Prozessierung der Rohdaten

Die im ABI-Format gewonnenen Daten wurden in das SCF- und Experiment-Format mit dem ASP-Package (http://www.sanger.ac.uk/Software/sequencing/docs/asp/processing.shtml) konvertiert. Hierbei wurden Qualitätswerte vergeben (PHRED, Sanger Centre; Hinxton/UK), schlechte Daten automatisch ausselektiert sowie die Positionen des Klonierungsvektors erfasst. Letzteres ermöglichte ein späteres Maskieren der Vektorsequenz. Das SCF-Format (Dear & Staden 1992) diente dem komprimierten Aufbewahren der Daten und der Elektropherogramme. Das EXP-Format (Bonfield & Staden 1996) ermöglichte das Verwenden der Sequenzen für unterschiedlichste Software.

2.4.1.2 Assemblierung der Shotgun-Sequenzen mit Phrap und Gap4

Die Einführung der Shotgun-Sequenzierung ist mit der Problematik der Datenzusammenführung der Shotgun-Reads verbunden, deren Lage auf dem Genom zunächst unbekannt ist. Zum Assemblieren der Daten hat sich PHRAP bewährt (Rieder et al. 1998; The chromosome 21 mapping and sequencing consortium, 2000). PHRAP benutzt die Sequenz- und Qualitätsinformation jeder einzelnen Base von PHRED zum „Alignen“ von überlappender Sequenzen bzw. Reads (Rieder et al. 1998). Als Parameter wurden PHRAP mit einem Minscore von 50 und einem Minmatch von 25 (mindestens 25 Basen der Reads müssen überlappen) eingestellt. Die Verknüpfung nicht durchgängiger Sequenzen wurde durch das Auslesen der Readpair-Informationen aus GAP4 erreicht. Readpärchen bzw. die *forward* und *reverse* Reads eines Inserts charakterisieren zusammengehörende Bereiche des Genoms. Klone, die *Sequencing Gaps* beinhalten, werden als Brückenklone bezeichnet (Abb. 10).

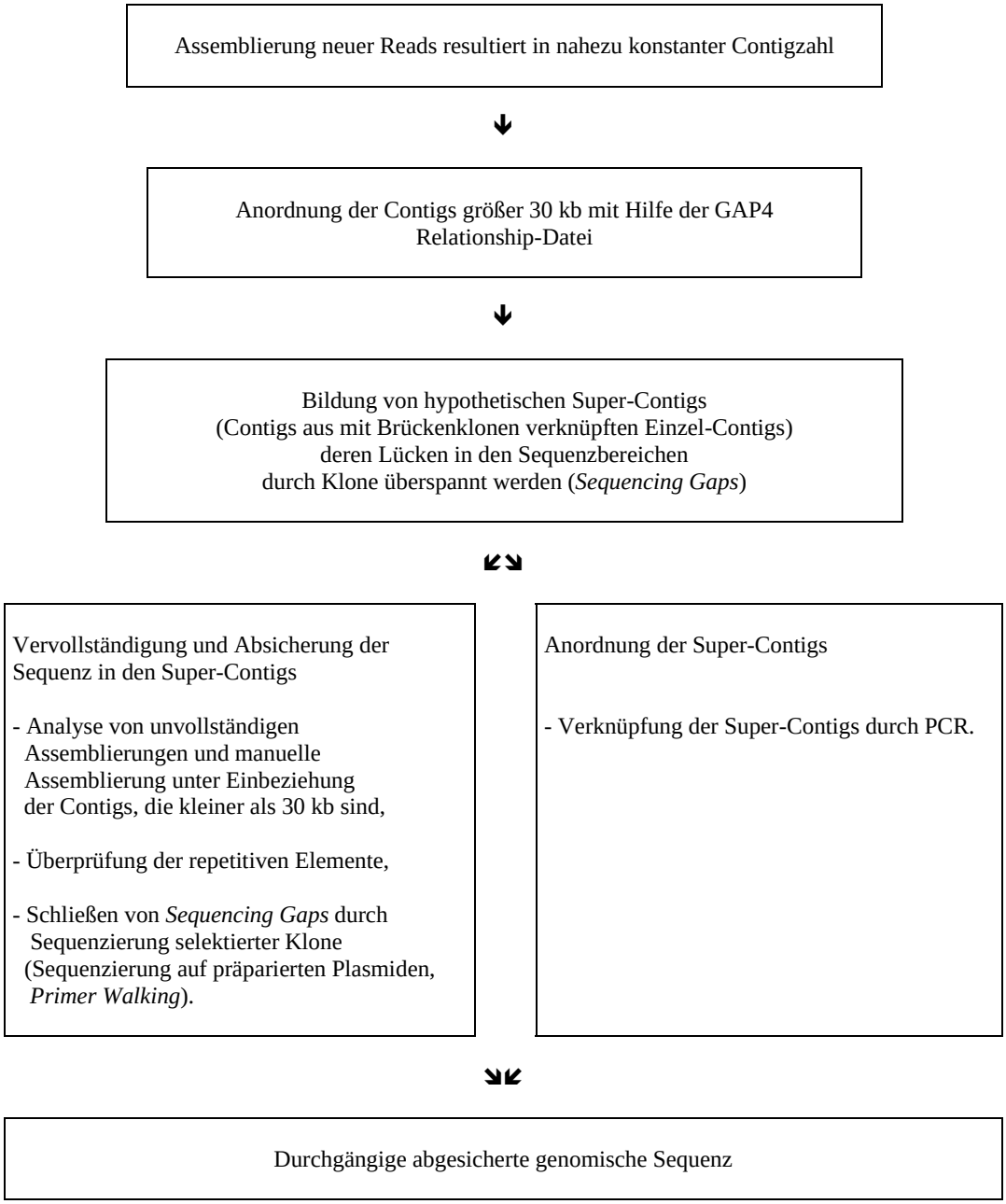
Durch die Verwendung der Software phrap2gap (http://www.sanger.ac.uk/Software/sequencing/docs/phrap2gap/) wurde die beschriebene Assemblierung und Überführung der Daten in eine GAP4 Datenbank (Staden et al. 1999) erreicht sowie die Clipping-Information (Maskierung der Vektorsequenz und Daten schlechter Sequenz, Informationen aus PHRED) mit einbezogen. Der methodische Ablauf von PHRAP wurde zur Begrenzung des temporär benötigten Arbeitsspeichers mithilfe eines Perl-Skriptes (Assembly-Split, A. Beck, MPI für Molekulare Genetik Berlin) modifiziert, so dass eine Assemblierung mit den zur Verfügung stehenden 2,5 Gigabyte Memory möglich war. GAP4 wurde im Anschluss als Editor zum manuellen Zusammenführen der Daten, Abrufen von Verknüpfun-

gen der Readpaare untereinander in der Datenbank, dem Entwickeln von Primern, manuellen Einfügen von zusätzlichen Reads zur Sequenzabsicherung, manuellen Zusammenführen von Contigs sowie dem Anbringen von Markierungen in der Sequenz, um nur einige Eigenschaften dieses Editors zu nennen, verwendet (Bonfield et al. 1995; Bonfield et al. 1998). Auf die eigentliche Assembly-Funktion dieser Softwareplattform wurde zugunsten des effektiveren und variableren PHRAP-Assemblers verzichtet. Die Datenprozessierung bis zur fertigen Sequenz wird im Überblick in der Abb. 8 dargestellt.

Die bestimmte Sequenz basiert nach Abschluss der Sequenzierung auf mindestens drei unabhängigen Reads von hoher Qualität, wobei Strang und Gegenstrang sequenziert werden. War die Bestimmung in dieser Form nicht für beide Stränge möglich, so wurde die Sequenz auf einem Strang durch drei unabhängige Reads bestimmt oder durch zwei unabhängige Reads und einen abhängigen, dessen Sequenzierung in der Methode abwich, z.B. Sequenzierung auf dem Plasmid. Durch die hohe Qualität der Sequenzierungen und Abdeckung liegt die Wahrscheinlichkeit eines Fehlers kleiner 1 zu 10000 Basen.

Identifizierte repetitive Elemente wurden zusätzlich durch PCR auf ihre Länge überprüft und das PCR-Produkt zusätzlich sequenziert, wenn die Readpärcheninformationen nicht ausreichend waren. Die gewählte Vorgehensweise stimmt in ihrer Qualität mit den für die Sequenzierung des humanen Genoms etablierten Bermuda-Qualität überein (Human Genome News 1998) und ist ähnlich der Qualität des Genoms von *Schizosaccharomyces pombe* (Wood et al. 2002). Die Qualität der Sequenz dieser beiden Genome weicht jedoch ab. So liegt die Sequenz bei *Schizosaccharomyces pombe* im Gegensatz zu *Pirellula* sp. Stamm 1 nicht lückenlos, sondern mit definierten Lücken vor. Qualitativ höherwertig ist jedoch die bei *Schizosaccharomyces pombe* durchgeführte Identifikation und experimentelle Überprüfung potentieller Frameshifts, die den Standard der Sequenzierung in anderen Genomprojekten bei weiten übertrifft und hier nicht geleistet werden konnte.

Abb. 8: Ablauf der Assemblierung bis zur fertigen Sequenz im Überblick

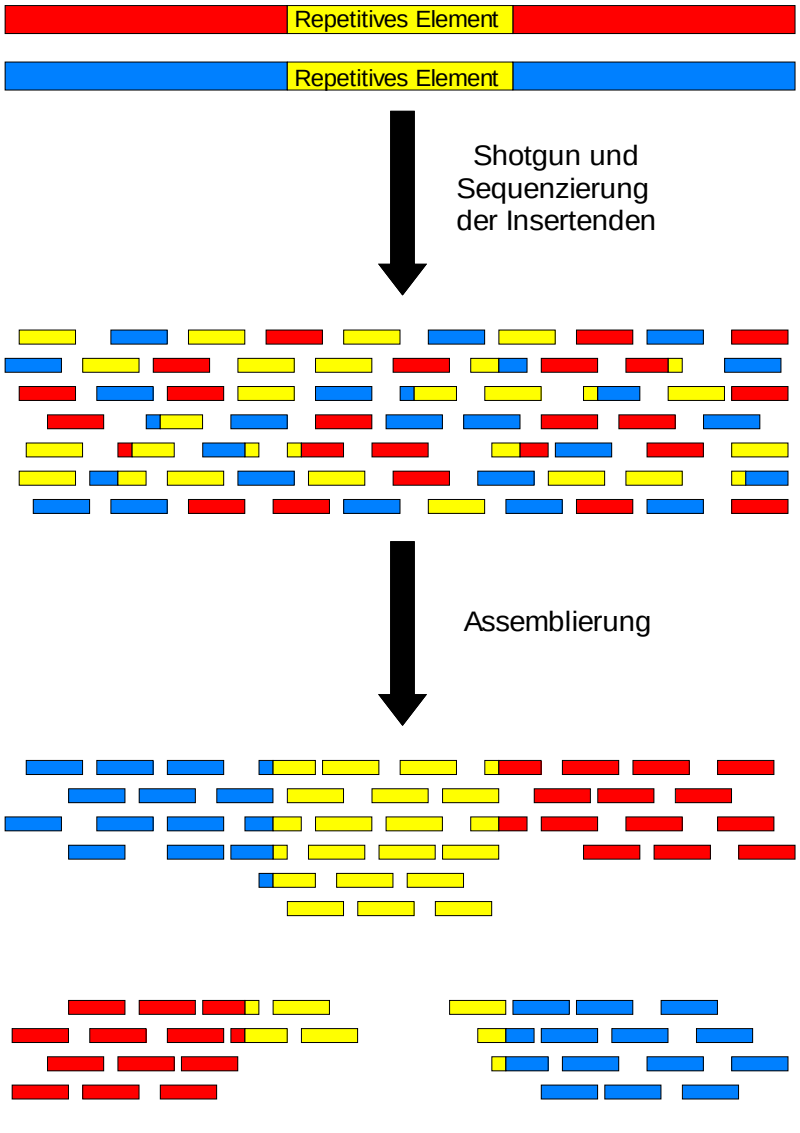


2.4.1.3 Identifizierung und Korrektur von fehlerhaften Assemblierungen

Nachdem die Assemblierung der Shotgun-Reads zu keiner weiteren Reduzierung der Contiganzahl führte, wurde der gesamte Datensatz gegen sich selbst mit BLASTN (Altschul et al. 1997) verglichen. Die Identifizierung aller in den genomischen Daten mehrfach auftretenden Sequenzbereiche war so möglich. Sequenzbereiche mit nahezu gleicher Nukleotidsequenz treten innerhalb aller bisher untersuchten Genome auf. Sie werden durch ihr wiederholtes Auftreten im Genom allgemein als Repeats bezeichnet. Der Begriff Repeats wird überwiegend für verhältnismäßig kurze Sequenzelemente, die nicht kodierend sind, verwendet. Der Begriff repetitive Elemente erscheint in diesem Zusammenhang als geeigneter, da er der komplexen Struktur und Sequenzlängen dieser Bereiche des Genoms gerechter wird und neutral in Hinsicht auf die unterschiedliche kodierte Information ist.
Alle Übereinstimmungen mit einer Identität von 90% und einer Größe von über 1000 bp wurden als Sequenzen herausgeschrieben und konnten anhand der Identitäten untereinander in Gruppen zusammengefasst werden. Jedes repetitive Element wurde als Suchpattern dem Mustererkennungsprogramm RepeatMasker (Smith & Green, http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl) übergeben, in der GAP4 Datenbank als repetitives Element markiert und automatisch mit einem Gruppenindex versehen. Hierdurch war es möglich auch nach bislang unentdeckten repetitiven Elementen im Genom zu suchen. Die Sequenz wurde anschließend abgesichert. In diesen identifizierten Bereichen können fehlerhafte Assemblierungen auftreten. Sie resultieren aus der Vorgehensweise des Phrap Algorithmuses, dessen Assemblierung unter Berücksichtigung der Phred Qualitätswerte auf Identitäten beruht. Diese Identitäten werden in den repetitiven Elementen häufig erreicht. In der Folge sammeln sich Reads in diesen Elementen an und führen im Alignment der Datenbank zu Stapelungen. Diese Stapelungen beinhalten neben den korrekt lokalisierten Reads auch Reads, die an einer anderen Stelle der genomischen Sequenz benötigt werden und so zur Lückenbildung führen. Problematisch ist darüber hinaus die Verknüpfung der flankierenden Sequenzen an einem repetitiven Element, da die Readlänge unter Umständen nicht ausreicht, um aus dem im Genom einmaligen Sequenzbereich über den repetitiven Bereich wieder einen einmaligen Sequenzbereich zu erreichen. Durch diese Problematik können zwei nicht zusammen gehörende Sequenzbereiche falsch über ein repetitives Element miteinander verknüpft werden (Abb. 9).

Abb. 9: Schematische Darstellung zur Entstehung von Assemblierungsproblemen

Dargestellt werden Ausschnitte an zwei unterschiedlichen Positionen der genomischen DNA. Beide Bereiche beinhalten einen nahezu identischen repetitiven Sequenzbereich (gelb).

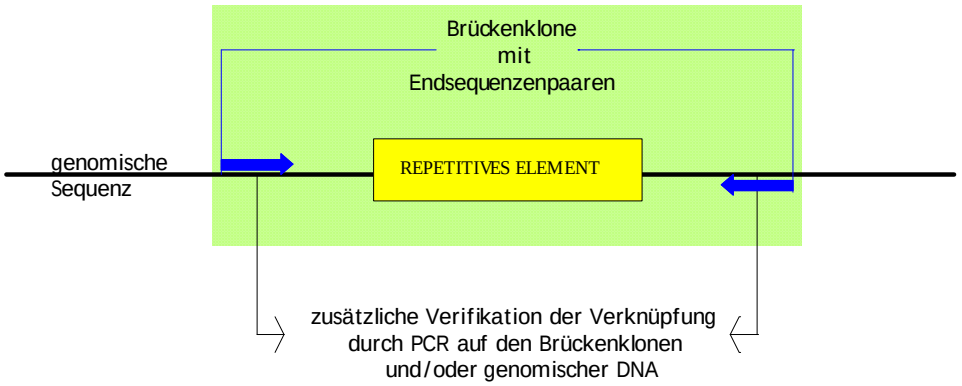


Die Shotgun-Reads können zum Teil oder vollständig repetitive Sequenzen beinhalten.

Die Assemblierung basiert auf Identitäten. Im Bereich der repetitiven Elemente können falsche Verknüpfungen auftreten. Reads, die mit ihrer Sequenz vollständig in repetitiven Elementen liegen, können fehlerhaft assembliert werden und fehlen an anderen Positionen der genomischen Sequenz.

Anschließend wurden Endsequenzen bzw. Inserts in der GAP4 Datenbank identifiziert, die den jeweiligen repetitiven Bereich umspannten. Diese Brückenklone haben ihre Endsequenzen in einmaligen Bereichen des Genoms außerhalb der repetitiven Elemente, beinhalten jedoch die repetitive Sequenz im Inneren ihres Inserts (Abb. 10).

Abb. 10: Absicherung der Verknüpfung an repetitiven Elementen



Diese Brückenklone stellen die erste Verifikation der Verknüpfung genomischer Sequenzen flankierend der repetitiven Elemente dar. So aufgespürte fehlerhafte oder fehlende Verknüpfungen wurden innerhalb der GAP4 Datenbank manuell korrigiert. Die Sequenz wurde innerhalb der repetitiven Elemente dann sukzessive ausgehend von den außerhalb der Elemente liegenden Sequenzbereiche manuell überprüft. Hierbei standen im Rahmen der Assemblierung die Read-Pärchen Informationen der Shotgun-Klone im Vordergrund. Zusammengehörende Reads wurden gegebenenfalls in der GAP4 Datenbank in das Alignment importiert bzw. widersprüchliche falsch lokalisierte Reads korrigiert. Die weitere Absicherung der Sequenzqualität wurde, wenn die vorliegenden Informationen nicht ausreichten, durch die gezielte Wiederholung der Sequenzierreaktion von Shotgun-Reads und *Primer Walking* Reaktionen auf den Brückenklonen sowie zusätzlich hergestellten PCR-Produkten erreicht. Im Anschluss wurden die Basenfolgen der repetitiven Elemente nochmals aus der genomischen Sequenz herausgeschrieben, wobei ausgehend vom Zentrum der Sequenz jedes repetitiven Elements in beiden Leserichtungen 10000 bp genomische Sequenz berücksichtigt wurden. Die resultierenden 20000 bp wurden dann mit dem Programm DOTPLOT$^{(+)}$ (Genetics Computer Group 1991) aligned. DOTPLOT$^{(+)}$ generiert eine grafische Darstellung eines Alignments bei der jede mögliche Übereinstimmung zwischen den beiden Sequenzen abgebildet wird. Dabei wird jede Base der 20 kb Sequenz (inklusive repetitives Element) mit einem Punkt (DOT) dargestellt. Dies ermöglicht mit variabler Sensitivität nach weiteren bisher nicht erkannten repetitiven Elementen im Umfeld des identifizierten Bereiches zu suchen. Die Positionen wurden aus der grafischen Darstellung entnommen, die verkürzten Sequenzbereiche herausgeschrieben und im Alignment gegenübergestellt. Die Alignments wurden auf die konservierten Bereiche der Sequenz verkürzt. Die erhaltenen Sequenzen wurden nochmals dem Programm Repeat-

Masker übergeben und die Positionen im Genom ermittelt. Die Ergebnisse dieser Analyse stimmten mit den zuvor erhaltenen BLASTN-Resultaten überein.

2.4.1.4 Zusätzliche Überprüfung der Assemblierung mit Hilfe der Cosmidbank

Die Endsequenzen von 907 Cosmidinserts wurden bestimmt. Diese Cosmide erreichen ein mehr als fünffaches *Physical Coverage* des Genoms. Die Positionen der erhaltenen Endsequenzen auf dem Genom wurden mit BLASTN (Altschul et al. 1997) bestimmt. Mehrfach auftretende Übereinstimmungen, die auf repetitive Elemente zurückzuführen sind, wurden unter Berücksichtigung der möglichen Insertgrößen der Cosmide interpretiert. Eine entsprechende Insertgröße von 28-45 kb, die dem Abstand der Endsequenzen auf dem Genom gleichgesetzt wurde, konnten als widerspruchfrei zur Shotgun-Assemblierung gewertet werden. Die grafische Umsetzung der Ergebnisse erfolgte mit dem Perl Skript ALIGN.pl (Georgi, MPI für Molekulare Genetik Berlin).

2.5 Sequenzanalysen ausgewählter Strukturen des Genoms

Ausgewählte Strukturen des Genoms wie die repetitiven Elemente, das rRNA-Operon und Teile des Replikationsapparates wurden mit bioinformatischen Methoden untersucht. Die Sequenzbereiche und im Detail analysierten ORFs wurden mit dem Bioedit Programmpaket (http://www.mbio.ncsu.edu/BioEdit/bioedit.html; T.Hall, North Carolina State University) bearbeitet. Nukleotid- sowie Peptidsequenzen wurden hier mit dem CAP-Assembler bzw. ClustalW zu multiplen Alignments zusammengeführt, Consensus-Sequenzen bestimmt und in grafische Darstellungen umgewandelt. Zur Unterstützung der Erstellung von Alignments wurde in Einzelfällen auf die GCG-Programme (Wisconsin Package Version 10.2, Genetics Computer Group (GCG), Madison, Wisc./USA) Bestfit und Gapshow für die graphische Umsetzung zurückgegriffen. Die ORF-Vorhersage wurde mit dem Programm ORPHEUS (Frishman et al. 1998) durchgeführt. Die Parameter wurden so gewählt, dass ORFs (*open reading frames* bzw. offene Leserahmen) mit einer Größe von ≥105 bp vorhergesagt wurden. Die ORFs wurden als Nukleotid- und entsprechend abgeleitete Aminosäuresequenz unter Angabe der Genomposition isoliert. ORPHEUS ermöglicht eine weitgehende Vorhersage aller Gene eines Genoms unter Berücksichtigung der statistischen Charakteristika von Protein kodierenden Regionen und der potenziellen ribosomalen Bindestellen. Die Analyse ausgewählter Sequenzbereiche wurde zusätzlich einer Überprüfung aller theoretisch möglichen ORFs mit dem Programm ORF-Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) unterzogen, um eine Diskriminierung poten-

ziell kodierender aber nicht vorhergesagter Bereiche zu vermeiden. Als weitere Absicherung wurden die selektierten Bereiche einer BLASTX-Analyse unterzogen (Altschul et al. 1997), die unabhängig von der Vorhersage der ORFs abläuft und auch trotz eventuell vorhandener *Frameshifts* Sequenzähnlichkeiten zu bekannten Sequenzen aufzeigt. Durch dieses Vorgehen besteht die Möglichkeit, auch eventuell im Laufe der Evolution degenerierte bzw. fragmentierte Gene zu identifizieren.

An diese Analysen schlossen sich die Untersuchungen der ORFs an. Diese ebenfalls auf Sequenzhomologien basierenden Analysen erstreckten sich von BLASTP-Suchen (Altschul et al. 1997) gegen NRPROT (http://www.ncbi.nlm.nih.gov/BLAST/), COGs (Clusters of Orthologous Groups of proteins; Tatusov et al. 2001) und InterPro (Apweiler et al. 2001) und nochmals einzeln gegen die Module Pfam (Bateman et al. 2002, http://www.sanger.ac.uk/Software/Pfam/, inklusive SMART sowie TIGRFAMs) und ProDom (Corpet et al. 2000). COGs, ProDom sowie die in Interpro beinhalteten Datenbanken ermöglichen neben dem Auffinden von sequenzhomologen Genen, Protein-Domänen oder Proteinfamilien, einen direkten Zugriff auf weitere Orthologe, deren Funktionenbeschreibungen und die Organisation der Gene bzw. die Charakteristika der jeweiligen Genfamilie.

ORFs, bei denen keine Funktionszuweisung oder Funktionszusammenhang sowie keine Ähnlichkeiten zu bereits beschriebenen ORFs bestehen (hypothetische ORFs), wurden in der Datenanalyse nur berücksichtig, wenn sie nicht im Widerspruch zu ORFs standen, die Sequenzhomologien zu anderen Organismen oder zu ORFs im eigenen Genom zeigten. Diese ORFs können nicht zur Analyse der Genese z.B. der repetitiven Elemente herangezogen werden. Ihre Vorhersage differiert in Abhängigkeit vom verwendeten Vorhersageprogramm. Derartige ORFs sind ohnehin zunehmend umstritten. Sie machten z.B. bei *Saccharomyces cerevisiae* ein Drittel aller annotierten kodierenden Regionen aus. Bezeichnungen wie „orphans“ oder „qORFs“, die für ORFs ohne zuordenbare Funktion oder bekanntes Proteinhomolog angewandt werden, werden in Bezug auf ihre tatsächliche kodierende Funktion zunehmend kritisch betrachtet bzw. als fragwürdig angesehen (Tatusov et al. 1997, Harrison et al. 2002).

Die Analysen von direkten und invertierten Repeats wurden mit den Programmen REPEAT (Wisconsin Package Version 10.2, Genetics Computer Group, Madison, Wisc./USA) und Palindrome (EMBOSS, http://www.hgmp.mrc.ac.uk/Software/EMBOSS/) durchgeführt. Palindrome ermöglicht durch die Vielzahl an Variationsmöglichkeiten auch eine Suche von imperfekten invertierten Repeats.

3. Ergebnisse und Diskussion

3.1 Sequenzierung

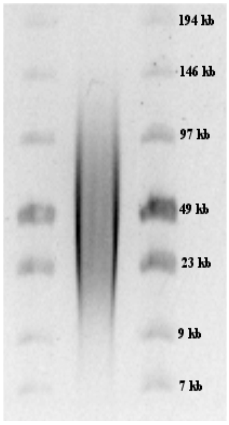
3.1.1 Genomische Shotgun-Banken

Als Basis für die Sequenzierung des Genoms von *Pirellula* sp. Stamm 1 wurden zwei Plasmid Banken mit unterschiedlicher Insertgröße erfolgreich hergestellt und eine Cosmid-Bank zur weiteren Absicherung der Assemblierung herangezogen.

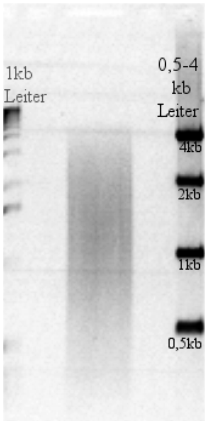
Die genomische *Pirellula* DNA wies vor der Fragmentierung im Maximum eine Größe von 9 - 146 kb auf (Abb. 11a). Nach der Fragmentierung wurde ein maximaler Größenbereich von 0,3 - 4 kb erreicht (Abb. 11b). Im Gegensatz zu der Aussage von Oefner et al. (1996) stellte sich die verwendete Fragmentierung durch Ultraschall unter den gewählten Versuchsbedingungen reproduzierbar dar.

Die Klone der Plasmidbanken zeigten nur einen geringen Anteil von blauen Klonen, der mit 0,5% abgeschätzt wurde. Diese blauen Klone werden zum Großteil auf nicht geschnittene Vektoren zurückgeführt, die entsprechend nicht ligiert werden konnten. Blaugefärbte Klone wurden im Picking generell nicht überführt.

Abb. 11 a und b: Verwendete *Pirellula* DNA vor und nach der Ultraschallfragmentierung



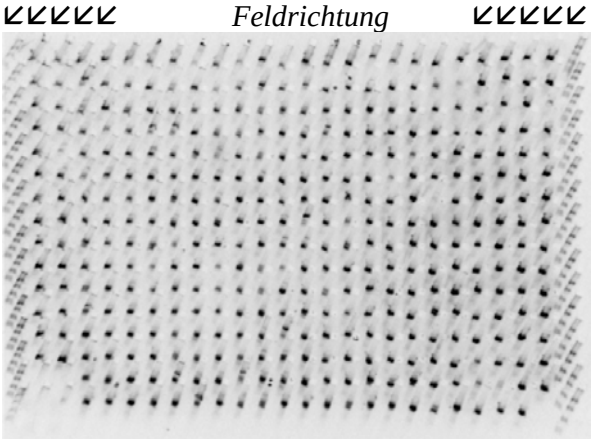
1% Agarose, 0,5 x TBE, 6 V/cm, 15°C, 14 h, switch time 1-12 sek; PFGE Low Range Marker (NEB, San Carlos/US) in den Randspuren, mittlere Spur 5 µg *Pirellula* sp. Stamm 1 DNA



1% Agarose,1 x TBE, 6 V/cm, 30 min; linke Spur 1 kb Leiter (Invitrogen, Karlsruhe), rechte Spur eine aus PCR-Produkten hergestellte Leiter, mittlere Spur 10 µg gescherte *Pirellula* sp. Stamm 1 DNA

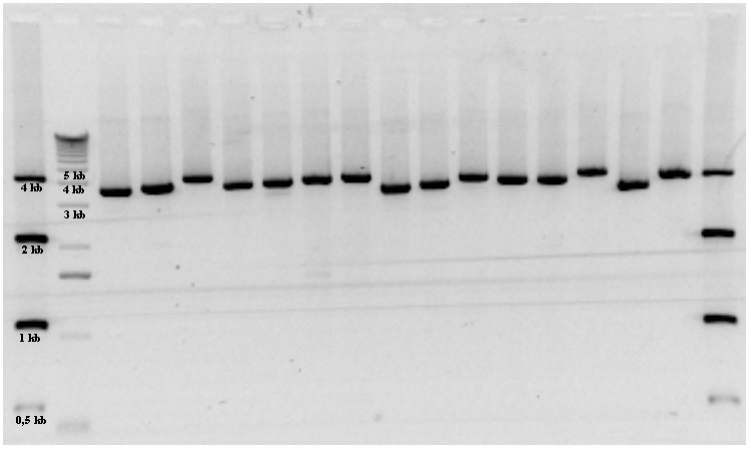
Die in pUC19 ligierten größenselektierten Fragmente zeigten in der *small insert library* eine Größe von durchschnittlich 1,5 kb. Durchschnittlich beinhalteten 85% der Klone ein amplifizierbares Insert im antizipierten Größenbereich und zeigten somit ein eindeutiges PCR-Produkt (Abb. 12).

Abbildung 12:
Beispiel eines 384er Agarosegels nach der Diagonalgelektrophorese
Als Größenstandard wurden PCR-Produkte mit einer Größe von 0,5, 1, 2 und 4 kb gewählt. Dieser Marker findet sich beidseitig in jeder Probenreihe (A1-A12 und P1-P12). Im obersten linken Probenviereck der PCR-Produkte (Position A1, A2, B1, B2) befinden sich die Positivkontrollen und unteren linken Viereck (O1, O2, P1, P2) die Negativkontrollen.



Die für die Shotgun-Sequenzierung verwendeten Plasmid Klone mit größeren Inserts zeigten eine Insertgröße von durchschnittlich 3,5 kb (Abb. 13). Bei durchschnittlich 95% der isolierten Plasmide konnte das Insert erfolgreich ansequenziert werden. Die Amplifizierung der 3,5 kb großen Inserts gelang nur bei durchschnittlich 70% im Hochdurchsatz. Diese geringe Ausbeute resultiert unter anderem aus der Verwendung eines nicht aufgereinigten Templates für die Amplifizierung, letztere gestaltete sich mit zunehmender Größe des Amplifikats als schwieriger.

Abb. 13: Größenabschätzung der 3,5 kb Insert Plasmid-Bank

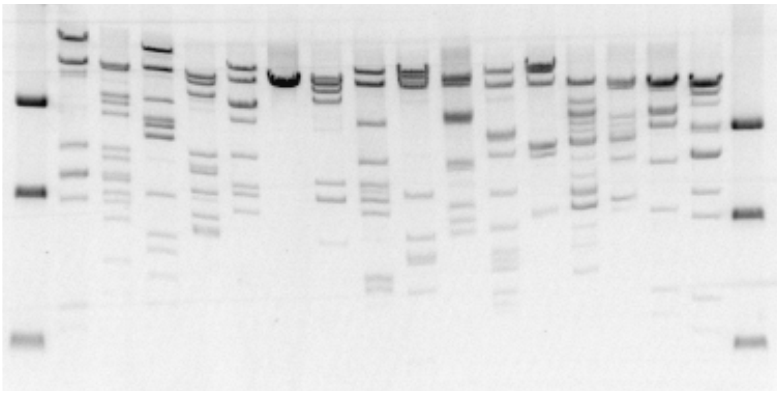


In den Außenspuren wurden jeweils PCR-Produkte als Marker aufgetragen. In der zweiten Spur eine 1 kb Leiter (Invitrogen, Karlsruhe). Größenunterschiede der Marker ergeben sich durch Rundungen und Konzentrationsunterschiede. In den restlichen Spuren befinden sich jeweils 3 µl von 25 µl PCR-Produkt der Inserts der 3,5 kb Insert-Bank. Von den dargestellten Insertgrößen sind jeweils 256 bp durch die Entfernung der Primer zur Klonierungsstelle abzuziehen.

Nach anfänglichen Schwierigkeiten wurde durch Anpassung der Parameter (Kap. 2.1.5) der Ligationserfolg deutlich gesteigert, wodurch die Transformationsrate auf 1,5 x 10^7 cfu/µg anstieg. Auch die Cosmid-Bank mit 1 x 10^7 cfu/µg lag in diesem Bereich. Die hergestellte Cosmid-Bank zeigte eine durchschnittliche Insertgröße von 30-45 kb. Die beschriebenen Insertgrößen wurden in der GAP4 Datenbank anhand der beiden Endsequenzen der Inserts verifiziert.

Lediglich bei 5% der Cosmide konnten keine Inserts nachgewiesen werden. Die erhaltenen Inserts zeigten eine weitgehend zufällige Verteilung. Mit Hilfe von Restriktionsendonukleasen wurden Stichproben der Cosmid-Bank untersucht (Abb. 14).

Abb. 14: Testserie der Cosmid-Bank nach dem Restriktionsverdau mit Eco RI
Die untersuchten Klone beinhalten wie antizipiert unterschiedliche Inserts.
Ein Klon von 20 (Spur 7) beinhaltet vermutlich kein Insert.



Auf einem 1%igem Agaosegel wurden 1,3 µg mit EcoRI verdaute Cosmid-DNA aufgetrennt.
In den Randspuren befindet sich ein selbst hergestellter Marker (4, 2 und 1 kb).

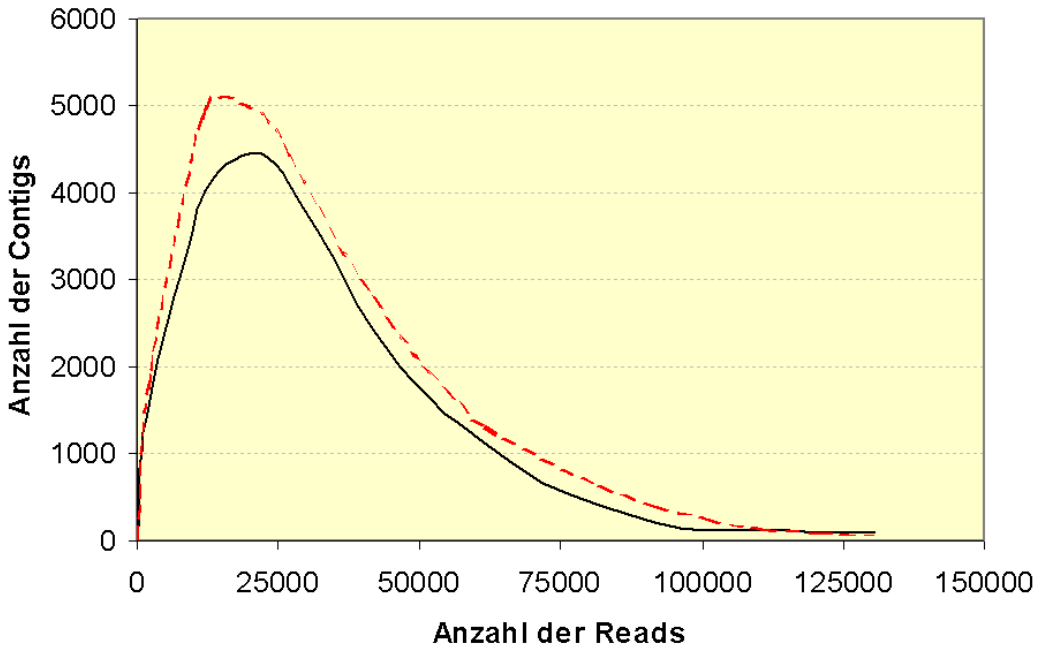
3.1.2 Bestimmung der genomischen Sequenz

Insgesamt 132055 Shotgun-Sequenzen mit einer durchschnittlichen Leselänge von 420 b wurden bestimmt (Tab. 13). Diese Reads ergeben eine Gesamtmenge an 56 Mb generierter Sequenz bzw. ein *Sequencing Coverage* von 7,76. Die Shotgun-Sequenzierung erreichte bei 90000 Reads (Abb. 15) und 205 Contigs (Abb. 16) die Plateauphase. Die Sequenzierung weiterer Shotgun-Klone wurde auch nach dem Erreichen der Plateauphase weitergeführt, um die Anzahl der *Physical Gaps* sowie der zusätzlichen Sequenzierreaktionen zur weiteren Absicherung zu reduzieren.

Abb. 15: Länge der bestimmten genomischen Sequenz in Abhängigkeit von der Anzahl der Reads

Die Assemblierungen der Shotgun-Sequenzen zeigt sich weitgehend in Übereinstimmung (Abb. 16) mit den theoretischen Überlegungen von Fraser und Fleischmann (1997).

Abb. 16: Entwicklung der Contigzahl in Abhängigkeit von der Readzahl
Der schwarz abgebildete Graph stellt die bei der Assemblierung der Shotgun–Reads aufgetretenden Contigzahlen (schwarze Linie) dar. Diesem Graphen wurden mithilfe der Formel $ne^{(-nw/l)}$ (Frazer & Fleischmann, 1997) die statistisch zu erwartenden Contigzahlen (unterbrochene rote Linie) gegenübergestellt. Die theoretisch zu erwartende Anzahl der Contigs (n) wird zu den jeweilig ereichten Readzahlen (*n*), den durchschnittlichen Leselängen der Reads (*w*) und der determierten Genomgröße von 7'145'576 b (*L*) in Beziehung gesetzt. Die Werte für *w* und *n* stammen aus den einzelnen Assemblierung der Shotgun-Reads, die alle 10'000 Reads durchgeführt wurden.



Die schließlich erreichte *Sequencing Coverage* von acht ist als angemessen einzuschätzen und ist auf eine ausgewogene Abdeckung des Genoms durch die verwendeten Banken zurückzuführen. Es stehen nur wenige Vergleichsdaten zur Verfügung, da Angaben zum *Sequencing Coverage* natürlich auch Angaben zur Leseweite der Reads mit sich bringen und häufig vermieden werden (Tab. 12).

Tab. 12:
Beispiele der Shotgun *Sequencing Coverage* bei mikrobiellen Genomprojekten

Organismus	Genomgröße in Megabasen	*Sequencing Coverage*	Referenz
Pseudomonas aeruginosa	6,3	6,9	Stover et al. 2000
Streptomyces avermitilis	8,7	10	Omura et al. 2001
Pirellula sp. Stamm 1	7,2	7,8	Glöckner et al. eingereicht
Salmonella enterica serovar Typhi CT18	5,1*	7.9	Parkhill et al. 2001
Bacillus halodurans	4,2	7,1	Takami et al. 2000
Methanosarcina mazei	4,1	8,6	Deppenmeier et al. 2002
Vibrio cholerae	4,0	7	Heidelberg et al. 2000
Campylobacter jejuni	1,6	10	Parkhill et al. 2000
Chlamydia pneumoniae J138	1,2	10	Shirai et al. 2000
Buchnera sp. APS	0,6	7	Shigenobu et al. 2000

(*inklusive Plasmide)

Die Assemblierung der Shotgun-Klone resultierte zunächst in 87 Contigs. Diese Contigs ließen sich mit den aus der Relationship-Datei der GAP4 Datenbank gewonnenen Informationen mithilfe von Brückenklonen in fünf „Super-Contigs" zusammenfassen. Jedes Super-Contig stellte einen zusammenhängenden Sequenzbereich dar, der von *Sequencing Gaps* unterbrochen wurde. Die Anordnung der Einzel-Contigs, die jedes Super-Contig formten, und die Länge der *Sequencing Gaps* waren durch die Insertgrößen der Brückenklone bekannt. Die fünf Super-Contigs erreichten eine Größe von 2,54 Mb, 2,45 Mb, 1,10 Mb, 0,68 Mb sowie 0,37 Mb und deckten somit das Genom fast vollständig ab. Lediglich fünf *Physical Gaps* blieben bestehen, die durch PCR-Produkte geschlossen werden konnten. Von allen zum Schließen der *Physical Gaps* entwickelten Primern an den Contigenden erbrachten nur fünf Kombinationen Amplifikationsprodukte mit einer Größe von 0,5 bis 4,3 kb. Die Gesamtlänge der *Physical Gaps* betrug 8951 bp.Es konnten keine weiteren PCR-Produkte nachgewiesen werden. Vier der ehemaligen *Physical Gaps* liegen in Cosmiden (Abb. 17). Die internen Sequenzen wurden durch *Primer Walking* bestimmt. Die geringe Anzahl an *Physical Gaps* wird neben der statistischen Entwicklung der Assemblierung (Abb. 16) auf den mit 47% hohen Anteil an Klonen mit einer Insertgröße bei 3,5 kb in der Sequenzierung zurückgeführt, der ein hohes *Physical Coverage* (30-fach) mit sich brachte.

Der Ausfall der Sequenzierreaktionen in einem Rahmen von 13% wird als durchschnittlich angesehen. Der Ausfall von Sequenzierreaktionen hat diverse Ursachen. Reaktionen können durch Pipettierfehler, die auch bei der Verwendung von Robotersystemen auftreten, ausfallen. Andere Ursachen liegen z.B. in der Verarbeitung der Proben im Hochdurch-

satz Maßstab, so können individuelle Eigenschaften der Proben wie z.B. Konzentrationsschwankungen nicht berücksichtigt werden.

Kritisch ist die Leselänge von nur 420 b zu sehen. Die Ursache hierfür ist in der Verwendung von verdünnten PCR-Produkten und einfach aufgereinigten Sequenzierreaktionen (Umfällung) zu sehen. Für die Sequenzierung stand entsprechend in der Mehrzahl der Fälle nur mit dNTPs, Primern und Salzen versetzte DNA zur Verfügung, welches für die folgende Sequenzierreaktion als suboptimal anzusehen ist. Auch die Aufreinigung der Sequenzierprodukte durch eine Ethanolpräzipitation ist als Kompromiss anzusehen. Kommerziell erhältliche Aufreinigungsprodukte für PCR-Produkte sowie Sequenzierprodukte konnten durch ihre hohen Kosten nicht im Hochdurchsatz eingesetzt werden. Hinzu kommt die geringere mögliche Leseweite der verwendeten Kapillarsequenzierer. Die erreichbaren Leseweiten von über 1000 bp beim gelbasierenden Sequenzieren (z.B. Li-Cor, Frankfurt/D) werden durch vorausgehende Aufreinigungsschritte und einen hohen manuellen Aufwand erreicht. Dieser Ansatz stand im Rahmen der vorliegenden Arbeit nicht zur Verfügung. Die gewählte Vorgehensweise ermöglichte bei höherer Anzahl von Sequenzierreaktionen und kürzeren Leseweiten die lückenlose Sequenzbestimmung in einem finanziell tragbaren Rahmen. Die Qualität der für die Sequenzierung eingesetzten DNA bestimmt hier eindeutig die Länge der Leseweite. Mit isolierten Plasmiden wurden auf den Kapillarsequenzierern Leseweiten von über 600 b erreicht. Die Bereitstellung derartiger Vorlagen für die Sequenzierung liegt jedoch alleine bei den Materialkosten um den Faktor 50 über denen der PCR-Produkte. Der Zeitfaktor für die Präparation ist in einem ähnlichen Verhältnis anzusetzen, so dass eine derartige Sequenzierungsstrategie nicht als Option erster Wahl zur Verfügung stand. Deshalb beruhen nur rund 10% der gesamten zufälligen Shotgun-Sequenzierungen auf Plasmidpräparationen. Dieser Anteil an Klonen mit 3,5 kb Inserts sollte auch die Repräsentierung des gesamten Genoms mit einem 6,6-fachen *Physical Coverage* gewährleisten. Diese Maßnahme basierte auf der Überlegung, dass durch die Amplifizierung der Inserts und die Sequenzierung von PCR-Produkten Sekundärstrukturen der DNA wie z.B. Hairpins in der PCR oder kurze Tandemrepeats häufiger Sequenzierreaktionen ausfallen und so Lücken in der Genomsequenz entstehen. Lücken, die nicht auf die Klonierbarkeit der DNA zurückzuführen sind, sollten zumindest eingegrenzt werden.
Die gewählte Vorgehensweise führte zu einer Erhöhung der Anzahl der benötigten Reads bzw. PCR-Produkte und Klone. Durch die Verwendung von weitgehend automatisierten Systemen ließen sich diese Aufgaben lösen.

Tab. 13: Überblick über die durchgeführten Sequenzierungen:

Shotgun-Sequenzierungen:

Analysierte Klone:	75789			
Anzahl der Reaktionen insgesamt:	151578	(100%)		
Beinhaltet:				
Anzahl der Reaktionen guter Qualität:	132055	(87%)	⇒ 7,8 faches *Sequencing Coverage*	
			Anzahl der Reaktionen der Plasmide (PCR-Produkte) mit 1,5 kb Inserts:	70614
			Anzahl der Reaktionen der Plasmide (PCR-Produkte) mit 3,5 kb Inserts:	48031
			Anzahl der Reaktionen der Plasmide (isolierte Plasmide) mit 3,5 kb Inserts:	13410
Anzahl der Reaktionen unbrauchbarer Qualität:	19523	(13%, davon 3% ohne identifizierbares Insert)		

Ausgewählte Wiederholungen, *Primer Walking* und Sequenzierungen auf speziellen PCR-Produkten:

Anzahl der Reaktionen insgesamt:	8584			
Beinhaltet:				
Anzahl der Reaktionen guter Qualität:	7261	(85%)	Anzahl der Reaktionen der Plasmide (isolierte Plasmide):	5575
			Anzahl der Reaktionen mit spezifischen Oligonukleotidprimern:	1686
Anzahl der Reaktionen unbrauchbarer Qualität:	1323	(15%)		

Anzahl aller zur Erstellung der genomischen Sequenz durchgeführten Reaktionen: 160162

Zur Absicherung und dem Erreichen der zirkulären genomischen Sequenz nach dem Abschluss der Shotgun-Sequenzierung wurden noch 7261 Reads benötigt. 76,8% der Reads gehen auf das Wiederholen von ausgewählten Sequenzierreaktionen zurück. In diesen Fällen wurden die Plasmide selektierter Klone präpariert, um größere Leseweiten und höhere Sequenzqualitäten zu erreichen. In fast 1700 Fällen mussten Sequenzreaktionen mit

spezifischen Primern durchgeführt werden. Insgesamt wurden 383 Primer zur Absicherung der Sequenz verwendet (Anhang Kap. 7.3, Tab. 73). Die hohe Anzahl der Reaktionen erklärt sich aus der Notwendigkeit, die Sequenzen auf unterschiedlichen Vorlagen zur Absicherung zu generieren. So erfolgte z.B. die Absicherung der repetitiven Elemente neben der Sequenzierung auf Brückenklonen in vielen Fällen notwendigerweise auf spezifisch generierten PCR-Produkten, von denen zur Verifikation der Sequenzen zunächst die Endsequenzen und dann weitere interne Sequenzen bestimmt wurden. Auch für diverse weitere Reaktionen im Rahmen des *Primer Walkings* zum Schließen von *Sequencing Gaps* musste mit spezifischen Primern auf den Brückenklonen sequenziert werden. Zusätzlich erforderten GC- und AT-reiche Regionen des Genoms die Verwendung von spezifisch für die Problemregion generierten Primern. Rückblickend hätte eine Erhöhung des *Sequencing Coverages* positiv betrachtet vermutlich nicht mehr als maximal eine Halbierung der 8584 Sequenzierreaktionen erreicht, da Reduzierungen der Wiederholungen lediglich in Bereichen mit geringer Abdeckung oder *Sequencing Gaps* auftreten können. Die Erhöhung des *Sequencing Coverages* um den Faktor eins hätte mehr als 17000 weitere Reads erfordert.
Die genomische Sequenz wurde mit einer Länge von 7'145'576 bp determiniert (noch verdeckt hinterlegt unter BX119912). Dominiert wurde die lückenlose Bestimmung der Sequenz durch die Verwendung des *whole genome shotgun* Ansatzes, der sich auch bei einem deratig großen Genom als erfolgreich erwies. Die genomische DNA liegt in einem ringförmigen geschlossenen Chromosom organisiert vor.

3.1.3 Ursachen für die Assemblierungsproblematik der repetitiven Elemente

Repetitive Elemente führten bei der Assemblierung zu Problemen, die eine gesonderte experimentelle Überprüfung jedes einzelnen Bereiches im Umfeld und verstärkte manuelle Eingriffe erforderten. Die gewählte Vorgehensweise des Generierens von Reads unterliegt einer Reihe von Überlegungen, die sicherlich mit zur Problematik der repetitiven Elemente beigetragen haben. Maßgebliche Ursache für das Auftreten der Assemblierungsproblematik dürften die erreichten Readlängen sein. Längere Leseweiten von durchschnittlich mehr als 600 b hätten durch das Auftreten von Überlappungen der Reads bei vielen Inserts zu einer Reduzierung von manuellen Eingriffen bei den repetitiven Elementen führen können.
Auf der bioinformatischen Seite hätte die Verwendung einer Assemblierungssoftware, die im Gegensatz zu PHRAP die Readpairinformation nutzt, vermutlich nur geringfügig zur Lösung der Problematik mit den repetitiven Elementen beigetragen. Bei der Verwendung

derartiger Assembler wie z.B. Arachne (Batzoglou et al. 2002) würden lediglich fehlerhafte Verknüpfungen nicht durchgeführt werden. Die Identifikation und Überprüfung müsste dennoch erfolgen. Eine Reduzierung der durch die Assemblierung bedingten *Gaps* würde nicht erfolgen. Zusätzlich verbleibt die Problematik, dass die zur Verfügung stehenden Assembler zurzeit in vielen Punkten nicht mit der GAP4 Datenbank kompatibel sind und dadurch die manuelle Nachbearbeitung nicht unerheblich erschweren.

3.1.4 Absicherung der Sequenz mit Hilfe der Cosmidbank

Eine Cosmidbank wurde erfolgreich hergestellt. Insgesamt 1152 Cosmidklone wurden in Kulturen überführt. Die beiden Endsequenzen von 907 Cosmidinserts wurden bestimmt, wodurch die Cosmidbank ein *Physical Coverage* des Genoms von 4,8 erreicht.

Insgesamt zehn Bereiche des Genoms mit einer Gesamtlänge von 124008 bp (1,7% des Genoms) konnten nicht durch die Cosmidbank abgedeckt werden. Dieses Phänomen von Bereichen, die von der Klonierung ausgenommen sind, zeigte sich auch schon bei Cosmidbanken anderer Organismen mit höherem *Physical Coverage* wie z.B. bei *Pseudomonas aeruginosa* PA01 (Stover et al. 2000).

Trotz einer zufälligen Fragmentierung der DNA gelang die Klonierung einiger Bereiche des Genoms deutlich besser als bei anderen (Abb. 17). Über die Ursachen für die unterrepräsentierten Bereiche kann nur spekuliert werden. Die Möglichkeiten reichen von toxischen exprimierten Produkten für den verwendeten *Escherichia coli* Stamm bis zu stabilieren Bereichen des Chromosoms, die die Präparation und damit die mechanische Beanspruchung besser überstanden haben und deshalb als längere Fragmente vorliegen.

Die Cosmidinserts konnten auf dem Genom angeordnet werden. Die Sequenzen bestätigten die determinierte genomische Sequenz und zeigten keine Widersprüche zur bereits bestimmten Anordnung der bestimmten Sequenz auf.

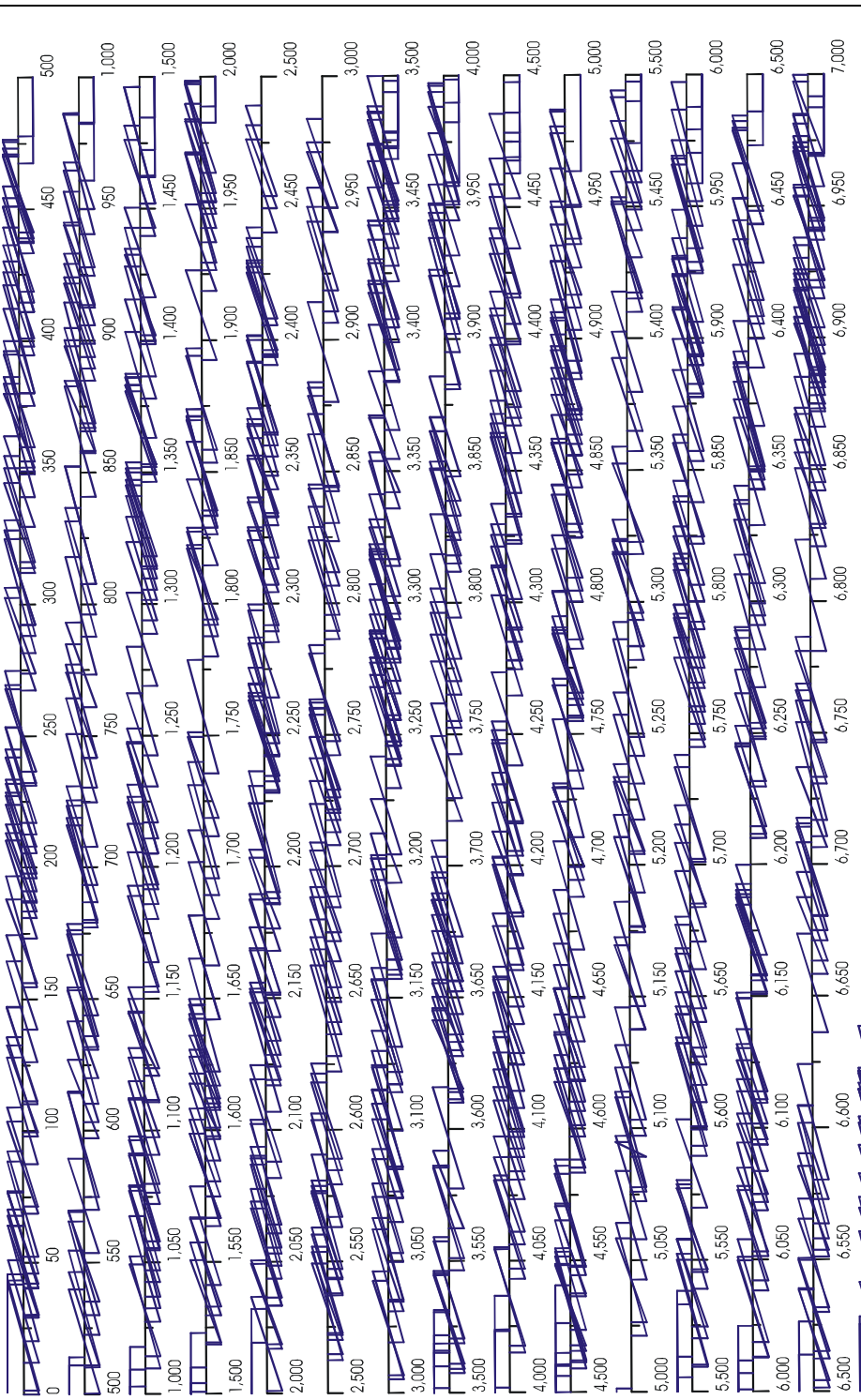


Abb. 17: Verteilung der Cosmide über das Genom von *Pirellula* sp. Stamm 1

Die Skalierung gibt die Positionen im Genom an. Die senkrechten Striche kennzeichnen die Start – und Endpositionen der Cosmidinserts. Zur Verdeutlichung der Zusammengehörigkeit sind die senkrechten Striche miteinander verbunden.

3.2 Strukturen des Genoms

3.2.1 Generelle Charakteristika des Genoms

Mit 7,15 Mb handelt es um eines der größten zirkulären Genome von mehr als 100 bereits sequenzierten mikrobiellen Genomen. *Pirellula* sp. Stamm 1 stellt nur einen der Vertreter der großen Bakteriengenome dar (Tab. 14), denen in der nächsten Zeit noch eine Vielzahl folgen werden. Besonders die Gruppe der Planctomyceten wird hier noch mit einigen Überraschungen aufwarten, wobei selbst das mit 9 Mb abgeschätzte Genom von *Gemmata obscuriglobus* (http://www.tigr.org/tdb/mdb/mdb/mdbinprogress.html) vermutlich nicht den größten Vertreter repräsentiert (Tab. 14).

Tab. 14: Beispiele großer mikrobieller Genome

Organismus	Genomgröße in Megabasen
Bradyrhizobium japonicum	9,11
Gemmata obscuriglobus	~9
Streptomyces coelicolor	8,67
Anabaena sp. strain PCC7120	7,21*
Pirellula sp. Stamm 1	7,15
Wigglesworthia brevipalpis	6,98
Nostoc sp. PCC 7120	6,41
Pseudomonas aeruginosa PA01	6,26
Pseudomonas putida KT2440	6,18

*inklusive Plasmide

Datengrundlage:
31.12.2002, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html;
26.01.2003, http://www.tigr.org/tdb/mdb/mdb/mdbinprogress.html;
Glöckner et al. eingereicht.

Der GC-Gehalt liegt bei 55,4% und damit in dem typischen Bereich der Planctomyceten-Genome (Gebers et al. 1985). Entsprechend dem großen Genom von *Pirellula* sp. Stamm 1 weist das Genom eine hohe Anzahl von Genen auf, die in Abhängigkeit der verwendeten Vorhersagemethode schwankt. Für das Genom von *Pirellula* sp. Stamm 1 wird zur Zeit eine Anzahl von 7394 Genen angenommen. Sie stellt das Resultat einer Genvorhersage der Bioinformatikgruppe des MPI für marine Mikrobiologie Bremen dar, die auf der Verwendung der Vorhersageprogramme Critica, Glimmer und ORPHEUS basiert. Durch die Verwendung unterschiedlicher ORF-Vorhersageprogramme resultierte hierbei zunächst ein redundanter Datensatz, der im Rahmen der manuellen Annotation bereinigt wurde (Glöckner et al. eingereicht 2003). Die unabhängig am MPI Berlin durchgeführte ORF-Vorhersage mit den Standardparametern von ORPHEUS führte zu einer Anzahl von 7359 ORFs. Eine derartig hohe Genanzahl legt den Schluss nahe, dass sich auch die Kom-

plexität des Organismuses in ihr widerspiegelt (Tab. 15), da prokaryotische Genome mit kodierender Sequenz eng gepackt vorliegen (Rogozin et al. 2002). Zum gegenwärtigen Zeitpunkt steht die funktionale Analyse des Genoms erst am Anfang.

Tab. 15: Komplexität des Lebens und Anzahl der Gene in *Bacteria, Archaea* und *Eukaryota*

Komplexitätsstufen des Lebens	Zahl der Gene	Modellorganismus
Zelle eines intrazellulären Parasiten	517	*Mycoplasma genitalium* (1)
Freilebende Bakterienzelle	1512	*Aquifex aeolicus* (2)
Freilebende Bakterienzelle mit Pirellulosome	7394	*Pirellula* sp. Stamm 1 (3)
Archaebakterium	3371	*Methanosarcina mazei* Stamm Goe1 (4)
Freilebende Zelle mit Zellkern	4824	*Schizosaccharomyces pombe* (5)
Insekt	13600	*Drosophila melanogaster* (6),
Nematode	19000	*Caenorhabditis elegans* (7)
Säugetier	30000 – 40000	*Human Genome* (8)
Pflanze	46022 – 55615	*Oryza sativa* (9)

(1) Fraser et al. 1995, Hutchison et al. 1999; (2) Deckert et al. 1998; (3) Glöckner et al. eingereicht; (4) Deppenmeier et al. 2002; (5) Wood et al. 2002; (6) Adams et al. 2000;
(7) The *C. elegans* Sequencing Consortium, 1998; (8) Yeh et al. 2001; (9) Yu et al. 2002.

3.2.2 Feinanalysen der repetitiven Elemente

3.2.2.1 Repetitive Elemente

In den folgenden Kapiteln werden die Analysen der im Genom von Pirellula bestimmten repetitiven Elemente mit einer Größe von mehr als 1000 bp vorgestellt (Kap. 2.4.1.3). Die Sequenzhomologien innerhalb der repetitiven Elemente betragen mindestens 90% und ermöglichten die Einteilung in 13 Gruppen (A-M; Kap. 7.3, Tab. 71).

Repetitive Elemente mit unterschiedlicher kodierender Information treten in allen Genomen auf. Sie entstehen zum Beispiel aus Duplikationen ganzer Genomabschnitte, Genamplifikationen und oder der Transposition mobiler Elementen, die sich wiederholt in das Genom einbauen. In mikrobiellen Genomen machen insbesondere sogenannte IS-Elemente (*insertion sequence elements*) einen Großteil der repetitiven Elemente aus. Sie gehören zur Gruppe der im Genom "beweglichen DNA", wovon mehr als 500 bereits beschrieben worden sind. In der Regel liegt ihre Größe bei maximal 2,5 kb (Mahillon & Chandler 1998), was auch mit den im *Pirellula* Genom identifizierten potenziellen IS-Elementen übereinstimmt (Anhang 7.3, Tab. 71). Bewegliche DNA-Elemente lassen sich als molekulare Parasiten des Genoms beschreiben, die für die Lebensvorgänge des Wirtes keine bestimmte Funktion besitzen, sondern nur für sich selbst existieren, sogenannte egoistische DNA (Lodish et al. 2001). Die Transposition eines IS-Elements ist jedoch ein sehr seltenes Ereignis, das in Abhängigkeit von der Art des Elements pro Generation nur in einer von 10^5 bis 10^7 Zellen auftritt. Höhere Transpositionsraten sind mit einem höheren Risiko für den Wirt verbunden. Zahlreich auftretende Transpositionen könnten zur Deaktivierung essenzieller Gene im Wirtgenom führen und somit einen Selektionsnachteil darstellen. Werden nicht essenzielle Bereiche getroffen, so können die Insertionssequenzen sich im Genom ansammeln (Mahillon & Chandler 1998).

Die Transposase des IS-Elements bindet beim *cut and paste* Mechanismus an die Ziel-DNA (Integrationsstelle) und erzeugt in kurzen Abständen versetzte Schnitte, so dass einzelsträngige Schnitte entstehen. Im Anschluss verknüpft die Transposase die 3'-Enden der Insertionssequenz mit den 5'-Enden der geschnittenen Ziel-DNA. Die durch Schnitte entstandenen einzelsträngigen Lücken in der Ziel-DNA werden mithilfe der DNA-Polymerase des Wirtes aufgefüllt. Die zwei einzelsträngigen Zielsequenzen liegen im Anschluss flankierend zu der Insertionssequenz als direkte Repeats vor. Bei der komplexen replikativen Transposition wird eine Kopie der Insertionssequenz erzeugt und das Original bleibt im Genom erhalten (Mahillon & Chandler 1998).

Bei bakteriellen IS-Elementen und Transposons kann der Mechanismus der Transposition nichtreplikativ oder replikativ erfolgen (Craig 1996). Während des nichtreplikativen Transpositionsmechanismuses wird die Insertionssequenz herausgeschnitten und an einer anderen Stelle des Bakteriengenoms wieder eingefügt. Hierbei bindet die Transposase an den invertierten Repeat der Insertionssequenz der Donor-DNA und spaltet die DNA, so dass die Insertionssequenz direkt glatt herausgeschnitten wird. Das Prinzip der replikativen Transposition beruht darauf, dass während der Replikation das Transposon dupliziert wird. Es entstehen Kopien an der Donor- und Zielstelle. Das Produkt ist ein Cointegrat. Bei der replikativen Transposition werden hierzu Einzelstrangbrüche am Transposon und der Zielstelle generiert. Die offenen Enden des Transposons und der Zielstelle werden über eine Crossing-over Struktur miteinander verbunden. Die Crossing–over Struktur beinhaltet jeweils eine einzelsträngige Region. Diese Regionen stellen Pseudoreplikationsgabeln dar. Wenn die Replikationen von den Pseudoreplikationsgabeln über die Transposons fortschreitet, kommt es zur Separierung der Stränge, gefolgt von der Termination an den Enden des jeweiligen Transposons. Als Resultat liegt eine weitere Kopie der Insertionssequenz vor. Dieser Zustand wird als Cointegrate bezeichnet und kann durch die Resolvase wieder aufgehoben werden.

Neben den IS-Elementen liegen als weitere Gruppe der beweglichen Elemente bakterielle Transposons in vielen Bakterien vor. Sie werden häufig auch als komplexe oder zusammengesetzte Transposons bezeichnet, da sie unter anderen aus mehr als einem IS-Element bestehen können (Klasse I Transposons). Transposons beinhalten neben der Information, die zur Transposition notwendig ist, weitere Gene (Lengeler et al. 1999). Derartige Strukturen können auch im Genom von *Pirellula* vorliegen. Ihre Identifikation ist jedoch bedingt durch fehlende Vergleichsdaten zurzeit nicht möglich. Es kann lediglich ausgeschlossen werden, dass bereits bekannte komplexe Transposons als repetitive Elemente im Genom auftreten. Auch das Vorliegen von replikativen Transposons, im Reaktionsmechanismus ähnlich dem des Phagen Mu (Klasse 2 Transposons; Lengeler et al. 1999), kann nicht abschließend bestimmt werden.

3.2.2.2 Die repetitiven Elemente der Gruppe A

Das repetitive Element A tritt mit zehn Kopien im Genom auf. Weitere Fragmente des repetitiven Elementes konnten mit BLASTN nicht aufgefunden werden. Die Varianten des repetitiven Elements liegen mit einer Länge von 1560 bp vor. Die Kopien zeigen sich in der Nukleotidsequenz weitgehend konserviert (Tab. 16). Die hochkonservierten Bereiche des repetitiven Elements (ab Base 8 bis einschließlich 1557) zeigen Abweichungen zur Consensussequenz von 0-21 Nukleotiden; teilweise sehr variable Außenbereiche wurden ausgeschlossen. Das repetitive Element A2 zeigt sich am variabelsten.

Tab. 16: Abweichungen der Varianten zum gemeinsamen Consensus

Als Basis wurde der hochkonservierte 1550b lange Sequenzbereich gewählt.

Variante	Anzahl der Abweichungen	Prozentuale Abweichung
A1	2	0,13%
A2	21	1,35%
A3	2	0,13%
A4	4	0,26%
A5	5	0,32%
A6	1	0,07%
A7	7	0,45%
A8	1	0,13%
A9	0	0,00%
A10	3	0,19%

Die repetitiven Elemente der Gruppe A zeigen im BLASTX die größten Ähnlichkeiten zu Transposasen (Tab. 17).

Tab. 17: BLASTX Resultate am Beispiel vom repetitiven Element A9

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identitäten	Positive	Gaps	Frame
Ralstonia solanacearum	NP_522696	ISRSO17-Transposase-Protein	453	158 bits (399)	2e-37	119/398 (29%)	187/398 (46%)	10/398 (2%)	+2
Bradyrhizobium japonicum	NP_71827	blr8270	457	158 bits (399)	2e-37	127/411 (30%)	186/411 (45%)	15/411 (3%)	+2
Azotobacter vinelandii	ZP_00092676	Hypothetisches Protein	453	155 bits (393)	1e-36	120/406 (29%)	183/406 (45%)	10/406 (2%)	+2
Streptomyces netropsis	AAD45539	InsA	390	155 bits (392)	2e-36	116/409 (28%)	183/409 (44%)	6/409 (1%)	+2
Xanthomonas oryzae pv. oryzae	AAO20850	mutmaßlich ISXo8 Transposase	439	143 bits (360)	8e-33	108/406 (26%)	185/406 (45%)	8/406 (1%)	+2

Die Vorhersagen mit ORPHEUS sagen für jedes der Elemente einen ORF (*orfA*; Tab. 18) mit Ähnlichkeiten zu den Transposasen voraus. Alle Kopien von *orfA*, außer der *orf4731* der Variante A2, weisen identische Start- und Stoppcodons auf. Der *orf4731* weist ein alternatives Startcodon auf, was zu einer Verlängerung des ORFs führt, sowie die Deletion zweier Aminosäuren, zwei Aminosäurenaustausche und zwei konservierte Substitutionen. Die im Gegensatz zu den anderen ORFs fehlenden zwei Aminosäuren gehen auf die Deletion eines durchgehenden sechs Basen langen Sequenzabschnittes zurück. Auf Nukleotidebene zeigt der *orf4731* mit insgesamt elf Abweichungen die höchste Abweichung vom Consensus.

Durch die sonst eindeutigen Vorhersagen wurde das Startcodon von *orf4731* als zu früh vorhergesagt eingestuft und das zweite einheitlich vorliegende Startcodon gewählt. Diese Interpretation wird durch die typische Struktur eines IS-Elements mit flankierenden Repeats unterstützt, in die der *orf4731* hineinragen würde.

Tab. 18: *orfA* im repetitiven Element A

Repetitives Element	*orfA*	Position im Genom	Länge (aa)	Abweichungen zum Consensus vom einheitlichen Startcodon ausgehend (na/aa)
A1	*orf1766*	1'153’972-1'152’695	426	0/0
A2	*orf4731*	3'076’293-3'077'564 (3'076’242-3'077'564)*	424 (441*)	11/6
A3	*orf4840*	3'151’526-3'150’249	426	1/0
A4	*orf5369*	3'527’026-3'528’303	426	4/0
A5	*orf5450*	3'578’831-3'577’554	426	1/0
A6	*orf6785*	4'478’644-4'479’921	426	1/0
A7	*orf8057*	5'338’206-5'339’483	426	2/0
A8	*orf9925*	6'593’355-6'594’632	426	0/0
A9	*orf10044*	6'593’355-6'594’632	426	0/0
A10	*orf10704*	7'081’786-7'083’063	426	0/0

*In der Klammer werden die nicht verkürzten Daten nach der ORPHEUS Vorhersage angegeben.

Der *orfA* zeigte in der BLASTP Analyse gegen die Proteindatenbank von NCBI lediglich signifikante Ähnlichkeiten zum ISRSO17-Transposase Protein von *Ralstonia solanacearum* auf (Tab. 19). Suchen der ORF Sequenzen gegen Interpro und COGs zeigten keine Ähnlichkeiten auf.

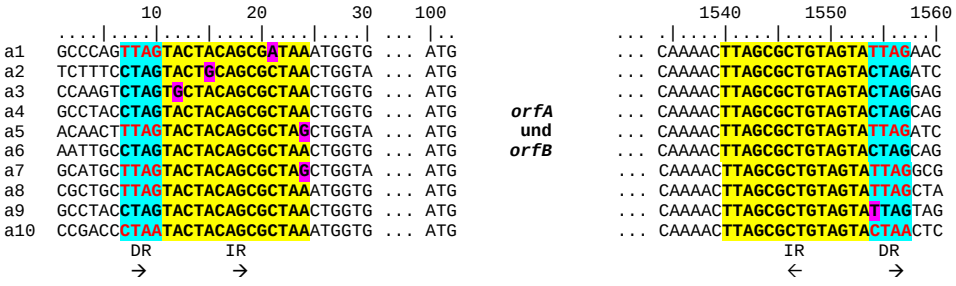
Tab. 19: Sequenzhomologie von *orfA* gegen das repetitive Element A8 gegen die NCBI Datenbank mit BLASTP

Ähnlichstes Protein:	ISRSO17-Transposase Protein von *Ralstonia solanacearum*
Zugriffsnummer:	NP_522696
Länge:	453 aa
Ähnlichkeiten im alignbaren Bereich:	
e-value:	7e-37
Identitäten :	119/398 (29%)
Positive :	187/398 (46%)
Gaps :	15/398 (3%)

Das propagierte IS-Element weist die charakteristischen invertierten Repeats und direkten Repeats auf (Abb. 18).

Abb. 18: Direkte Repeats der Zielstellenverdopplung und invertierte Repeats flankieren die vermutliche Transposase und den benachbarten ORF im repetitiven Element A

Dargestellt werden die Randbereiche des repetitiven Elements A mit einer Gesamtlänge von 1560 bp. Die repetitiven Elemente sind mit „r_“ abgekürzt. Farbliche Unterlegungen kennzeichnen Gelb die invertierten Repeats, lila Abweichungen der Sequenz im Alignment, Blau die direkten Repeats und in Rot die im direkten Repeat überwiegende Sequenz.



Ein weiterer potenzieller ORF, *orfB*, mit 50 aa liegt im Randbereich des repetitiven Elements vor, wird aber durch ORPHEUS nicht vorausgesagt. Die fehlende Vorhersage durch ORPHEUS resultiert aus dem bevorzugten Vorhersagen langer ORFs, die bei diesem Element in anderen Frames erfolgen. Diesen vermutlich falsch vorhergesagten ORFs ließen sich keine Funktionen nach der BLASTP Analyse zuweisen. Auch zu dem bestimmten *orfB* (Tab. 20) lassen sich keine Aussagen treffen, weshalb der *orfB* als hypothetischer ORF angesprochen werden soll. Er liegt jedoch konserviert im postulierten IS-Element vor, so dass es sich hierbei um einen mit der Transposase assoziierten ORF handeln kann. *IS Rso17* von *Ralstonia solanacearum* besitzt ebenfalls einen ähnlichen konservierten Sequenzbereich bis zu den invertierten Repeats des IS Elementes. Bedingt durch die kurze Länge der zur Verfügung stehenden Sequenz für einen weiteren ORF wurde auf die Aus-

weisung verzichtet. Auch im IS-Element Rso7 liegt dieser Sequenzbereich konserviert vor (http://sequence.toulouse.inra.fr/ralsto/Complete/InsertionSequences/ISRso17.20.gif).

Tab. 20: *orfB* repetitiven Element A

Variante	Position im Genom	Länge (aa)	Abweichungen zum Consensus (na/aa)
A1	1'152'681-1'152'532	50	0/0
A2	3'077'578-3'077'727	50	5/1
A3	3'150'235-3'150'086	50	0/0
A4	3'528'317-3'528'466	50	0/0
A5	3'577'540'-3'577'391	50	2/1
A6	4'479'935-4'480'084	50	0/0
A7	5'339'497-5'339'646	50	2/1
A8	6'594'646-6'594'795	50	0/0
A9	6'665'894-6'665'745	50	0/0
A10	7'083'077-7'083'226	50	0/0

Die identifizierten Strukturen (Abb. 19) weisen auf ein neues bisher uncharakterisiertes IS-Element hin, welches sich nicht den bisher beschriebenen IS-Elemente zuordnen lässt, ohne dass die Abweichungen von den Gruppen überwiegen. Die charakteristischen Grundstrukturen ließen sich jedoch identifizieren.

Abb. 19: Schematische Darstellung des IS-Elements des repetitiven Elements A
Die im repetitiven Element A identifizierten Bereiche wie die direkten Repeats (DR), die flankierenden invertierten Repeats (IR), die vermutliche Transposase und ein konservierter ORF ohne Hinweise auf eine Funktion durch Sequenzvergleiche füllen das repetitive Element aus. Die Pfeile geben die Transkriptionsrichtung an.

3.2.2.3 Die repetitiven Elemente der Gruppe B

Das repetitive Element B wurde zunächst mit einer Länge von insgesamt 1313 bp bestimmt. Es stellt sich mit seinen neun Kopien im Genom als weitgehend konserviert dar (Tab. 21). Weitere Fragmente konnten im Genom nicht aufgefunden werden.

Tab. 21: Abweichungen der repetitiven Elemente zum gemeinsamen Consensus

Als Basis wurde der hochkonservierte 1313 b lange Sequenzbereich gewählt. Die variablen Randbereiche (ersten zwei bzw. letzten zwei Basen) der zunächst mit 1309 b bestimmten Variante wurden nicht berücksichtigt.

Repetitives Element	Anzahl der Abweichungen	Prozentuale Abweichung
B1	2	0,15
B2	0	0,00
B3	0	0,00
B4	0	0,00
B5	0	0,00
B6	0	0,00
B7	0	0,00
B8	1	0,08
B9	1	0,08

Die BLASTX Analyse der Sequenz des repetitiven Elements B gegen die Datenbank von NCBI zeigt Sequenzhomologien zu Genen mit Transposasefunktion (Tab. 22). Auffällig sind hierbei die Sequenzhomologien in unterschiedlichen Leserastern (z.B. mögliche Transposase von *Brucella melitensis* biovar Abortus) und die Ähnlichkeiten zu einem zweiten Gen (z.B. *ISRSO8*-Transposase *orfB* Protein von *Ralstonia solanacearum*).

Tab. 22: BLASTX Resultate am Beispiel vom repetitiven Element B7

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Brucella melitensis biovar Abortus	AF454951	mutmaßliche Transposase	401	219 bits (559)	2e-63	118/296 (39%)	179/296 (59%)	2/296 (0%)	+3
				46.2 bits (108)	2e-63	24/85 (28%)	44/85 (51%)	-	+1
Ralstonia solanacearum	NP_518699	ISRSO8-Transposase *orfB* Protein	296	210 bits (534)	3e-53	120/289 (41%)	165/289 (56%)	2/289 (0%)	+3
Mesorhizobium loti	NP_106702	Transposase	286	205 bits (522)	8e-52	111/284 (39%)	166/284 (58%)	1/284 (0%)	+3
Lactococcus lactis subsp. lactis	S14450	wahrscheinlich Transposase, insertion sequence *IS1076, orf1*	384	183 bits (465)	1e-49	106/304 (34%)	161/304 (52%)	2/304 (0%)	+3
Leptospira interrogans	AAA88919	*orfB*, mutmaßliche Transposase	282	198 bits (503)	1e-49	108/278 (38%)	158/278 (55%)	-	+3

Durch die teilweise außerhalb des repetitiven Elements liegenden potenziellen ribosomalen Bindungsstellen sagt das Vorhersageprogramm ORPHEUS unterschiedliche ORFs für das Element voraus. Die Struktur des hier vorliegenden besonderen IS-Elements wird vom Programm in seiner Komplexität nicht berücksichtigt. Ausgehend von der Variante B2 wurden deshalb alle theoretisch möglichen ORFs bestimmt (Tab. 23) und auf konservierte Sequenzbereiche untersucht. Hierbei konnten zwei überlappende ORFs identifiziert werden, die Homologien zu IS-Elementen zeigen. Die anderen möglichen ORFs wurden verworfen, da keine Hinweise auf mögliche Funktionszuordnungen gefunden wurden.

Tab. 23: ORFs im repetitiven Element B

Repetitives Element	Position *orfA*	Länge (aa)	Abweichungen zum Consensus (na/aa)	Position *orfB*	Länge (aa)	Abweichungen zum Consensus (na/aa)
B1	2214533 - 2214820	96	1/1	2214799 - 2215701	301	0/0
B2	378137 - 378424	96	0/0	378403 - 379305	301	0/0
B3	606838 - 607125	96	0/0	607104 - 608006	301	0/0
B4	6328064 - 6328351	96	0/0	6328330 - 6329232	301	0/0
B5	6361950 - 6362237	96	0/0	6362216 - 6363118	301	0/0
B6	6966246 - 6966533	96	0/0	6966512 - 6967414	301	0/0
B7	89333 - 89620	96	0/0	89599 - 90501	301	0/0
B8	3087349 - 3087636	96	0/0	3086468 - 3087370	301	1/1
B9	3563859 - 3564146	96	0/0	3562978 - 3563880	301	1/1

Überlappende ORFs treten in der *IS1* (Sekine et al. 1992) und *IS3*-Familie (Gerischer et al. 1996) auf. Intensiv analysierte Vertreter der *IS3*-Familie sind *IS3* (Sekine et al. 1994), *IS150* (Vögele et al. 1991), *IS911* (Polard et al. 1991) und *IS1236* (Gerischer et al. 1996). In diesen IS-Elementen wird davon ausgegangen, dass zwei überlappende ORFs, *orfA* und *orfB* durch einen –1 translationalen Frameshift ein Fusionsprotein produzieren können (Sekine & Ohtsubo 1989), wobei das Fusionsprotein die aktive Transposase darstellt. Programmiertes *Frameshifting* ist in bakteriellen Insertions-Elementen am weitesten verbreitet (Rettberg et al. 1999). Für *IS1236* und *IS1* konnte experimentell aufgezeigt werden, dass die im Überlappungsbereich enthaltene Sequenz A AAA AAA (A_7) ein *Frameshift Window* darstellt. Auch im repetitiven Element B findet sich das *Frameshift Window*, das sich mit seinen Randbereichen erstaunlich konserviert zu einer Transposase (*Tn1953*) in *Brucella melitensis* biovar Abortus (AF454951) zeigt (Abb. 20). In *IS911* von *E. coli* und der Mehrheit der *IS3* Elemente lässt sich die konservierte Sequenz des *Frameshift Windows* noch auf A_7G erweitern (Rettberg et al. 1999). Dieses *Frameshifting Window* wird auch im programmierten *Frameshifting* von *dnaX* verwendet (Flower & McHenry 1990; Tsuchihashi & Kornberg 1990; Blinkowa & Walter 1990).

Abb. 20: ***Frameshift Windows*** **und flankierende Bereiche des repetitiven Elements B und** ***Brucella melitensis*** **biovar Abortus (AF454951, Position 48780-48804) im Alignment**

Das potenzielle *Frameshift Window* ist im Fettdruck hervorgehoben.

```
Element B : taaaaaaagcgacggcattcttcgc
            |||||||||||||||| ||||||||
Brucella:   taaaaaaagcgacggccttcttcgc
```

In IS-Elementen wie *IS911* konnte nachgewiesen werden, dass *orfA* (in *IS1* als *InsA* bezeichnet) auch exprimiert wird, wenn der *Frameshift* ausbleibt (Polard et al. 1991). Einige Studien in *IS1* weisen darauf hin, dass das Verhältnis zwischen *InsA* zur Transposase die Effizienz des *Frameshift*-Ereignisses beeinflusst, wodurch eine Kontrolle der Transposition des Insertionsereignisses bestehen würde (Machida & Machida 1989; Sekine & Ohtsubo 1989). Für *IS1* konnte gezeigt werden, dass *Frameshiftings* nur in sehr geringen Umfang auftreten (Sekine et al. 1992). Eine dem *Frameshifting Window* vorgelagerte ribosomale Bindungsstelle, wie in *IS911*, die als Enhancer fungiert (Rettberg et al. 1999), konnte nicht identifiziert werden. Weitere Anhaltspunkte für ein hypothetisch gering angenommenes Auftreten der *Frameshiftings* sind das Fehlen von Sekundärstrukturen, wie Hairpin Strukturen am Ende von *orfA*, die das Ereignis räumlich durch ihre Faltung unterstützen (Sekine et al. 1992; Rettberg et al. 1999). Derartige Strukturen, die z.B. in *IS3*, *IS911* und *IS1236* potenziell vorliegen (Gerischer et al. 1996), fehlen benachbart zum *Frameshift Window* im repetitiven Element B und führen somit zu einer niedrigen Expression der Transposase (Sekine et al. 1992). Hierdurch wird das Risiko eines letalen chromosomalen Re-Arrangements für das Bakterium durch den Einbau des IS-Elementes gering gehalten (Sekine et al. 1992).

Der im repetitiven Element B lokalisiert *orfA* besitzt mit einer Länge von 96 aa eine typische Länge für den ersten ORF im IS-Element und eindeutige Homologien zu Transposasen (Tab. 24).

Tab. 24: BLASTP Resultate am Beispiel von *orfA* des repetitiven Element B2

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Agrobacterium tumefaciens	AAL46059	Transposase	96	55.1 bits (131)	2e-07	36/94 (38%)	54/94 (57%)	1/94 (1%)
Brucella melitensis	NP_541161	Transposase	93	52.0 bits (123)	2e-06	31/93 (33%)	52/93 (55%)	3/93 (3%)
Mesorhizobium loti	NP_106546	Transposase	98	50.4 bits (119)	4e-06	32/93 (34%)	50/93 (53%)	1/93 (1%)
Escherichia coli O157:H7 EDL933	NP_289541	unbekanntes Protein kodiert durch *ISEc8*	224	45.1 bits (105)	2e-04	19/50 (38%)	34/50 (68%)	1/50 (2%)
Pantoea agglomerans	AF327445	mutmaßliche Transposase A	94	40.8 bits (94)	0.003	29/91 (31%)	45/91 (48%)	3/91 (3%)

orfA besitzt das für den ersten ORF typische Helix-Turn-Helix Motiv (HTH Scan: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hth.html; Dodd & Egan 1990), das z.B. auch in IS1236 (Gerischer et al. 1996) identifiziert wurde. Die konserviert auftretenden Strukturen zeigen Ähnlichkeiten zum COG2963, der eine Transposasengruppe repräsentiert, und zum Pfam01527 Eintrag (Abb. 21), in der die Transposase 8 Familie zusammengefasst wird.

Abb. 21: Vollständiges Alignment von *orfA* zum Consensus von Pfam01527, Transposase 8

```
orfA:       5   RTFSREYKLAAVKKVIEQGLSYTAVAKDLGIGDSLIRKWKKSFDE-DGTFQAEVVGSQSI  63
Pfam01527:  1   RRYSEEFKAIAVKLY-EAGRSVSEVAREHGVSPATLYKWRKKYGEKAGMEVSDAKRLKAL  59

orfA:       64  EAELRRLREENRQLKMERDILKKATAFFA  92
Pfam01527:  60  EKENRELRKELARLKLENEILKKAAAKKS  88
```

Ein invertierter Repeat, der die potenzielle ribosomale Bindestelle mit einschließt und sie so evtl. blockiert, wie bei *IS1236* (Gerischer et al. 1996; Timmerman & Tu 1985), konnte nicht identifiziert werden.

Tab. 25 : BLASTP Resultate am Beispiel von *orfB* des repetitiven Elements B2

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Brucella melitensis biovar Abortus	AF454951	mutmaßliche Transposase	401	207 bits (527)	1e-52	111/285 (38%)	171/285 (59%)	3/285 (1%)
Ralstonia solanacearum	NP_518699	*ISRSO8*-Transposase orfB Protein	296	202 bits (513)	5e-51	120/289 (41%)	165/289 (56%)	5/289 (1%)
Mesorhizobium loti	NP_106702	Transposase	286	199 bits (507)	3e-50	111/284 (39%)	166/284 (58%)	4/284 (1%)
Leptospira interrogans	AAA88919	*orfB*	282	194 bits (492)	1e-48	108/278 (38%)	158/278 (55%)	1/278 (0%)
Agrobacterium tumefaciens str. C58	AAL46058	Transposase	286	193 bits (491)	2e-48	107/286 (37%)	163/286 (56%)	4/286 (1%)

Die BLASTP Ergebnisse (Tab. 25) spiegeln zum Teil bereits in der Kurzbeschreibung der Genfunktion die erwartete *orfB* Zuordnung (NP_518699; AAA88919) des IS-Elements wider. Diese Sequenzhomologien zeigen sich auch in der Zuordnung zu COG2801 mit der angenommenen Funktion einer Transposase. Charakteristischer ist die eindeutige Zuordnung des *orfB* im repetitiven Element B zur Integrasen Kerndomäne (Pfam00665; Abb. 22), was in retroviralen Integrasen und der *IS3*-Familie konserviert vorliegt und vermutlich katalytisch aktiv ist (Fayet et al. 1990; Polard & Chandler, 1995). *IS1* von *E. coli* K12 zeigt keine Ähnlichkeiten zu Elementen der retroviralen Integrasen, sondern zur Transposase 27 (Pfam03400). Diese Zuordnung ist charakteristisch für die Mitglieder der *IS1*-Familie.

Abb. 22: Alignment von *orfB* mit dem Consensus von Pfam00665

Die Domäne Pfam00665 beinhaltet die Integrase Core Domäne. Integrase vermitteln die Integration von DNA Kopien eines viralen Genoms in das Wirtchromosom. Sie treten jedoch auch als katalytisch aktiver Bereich in IS-Elementen auf. Die Domäne wird vollständig gegenübergestellt.

```
orfB:      131  TTEAINRVWLTDITYIP--TQEGSTYLCAFVDLHSRKIVSWKTSRNMDSELVVGAFDQAL  188
Pfam00665:   1  RASRPNELWQMDFTPLPVLGKGGKKYLLVIVDDFSRFVVAYPLKSKTSAETVFDLLEAAL  60

orfB:      189  TFRKPNAGLIVHSDRGSQFASDHFRRRLAASGLVQSMSRRGNCYDNAPMESFFKSYKTEE  248
Pfam00665: 61  ERRGG-KPKTIHSDNGSEFTSKAFQELLKELGIKHSFSRPYSPQDNGVVERFNRTLKREL  119

orfB:      249  AQQIYDTHEHATRGVSDYIERFYNPHRLHSSLGYLSPIDFEQ  290
Pfam00665:120  RKLLRFLSLEEWEEALETALYLYNRRRHSLLGG-TPAERLA  160
```

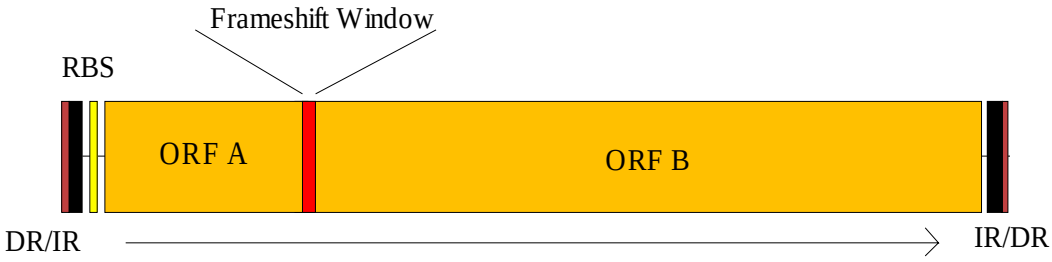
Die Initiation der Translation durch das seltene AUU Codon im *orfB* wurde auch in *IS911* nachgewiesen, wobei die Lokalisation des Startcodons für den zweiten ORF strangabwärts des *Frameshifting-Windows* in repetitiven Element B von der in *IS911* strangaufwärts liegenden Position abweicht. Ein derartiges *orfB* Protein konnte im Gegensatz zu *IS3* in *IS1* nicht nachgewiesen werden (Polard et al. 1991). Die Expression von *orfB* unabhängig von *Frameshifting* Ereignis wurde unter anderem für *IS911* (Rettberg et al. 1999) nachgewiesen.

Die beiden im repetitiven Element B identifizierten ORFs werden von zwei invertierten terminalen Repeats mit geringen Abweichungen und einer Länge von 34 bp eingeschlossen. Flankierend tritt eine Zielstellenverdopplung in Form von direkten Repeats mit einer Länge von drei Basen auf (Abb. 23), die mit der Zielstellenverdopplung von Vertretern der *IS3*-Familie wie *IS1236* übereinstimmt (Gerischer et al. 1996). Eine Zielstellenpräferenz konnte nicht identifiziert werden.

Abb. 23: Alignment der Randbereiche des repetitiven Elements B
Farbunterlegungen kennzeichnen in Blau die außerhalb des repetitiven Elements liegenden direkten Repeats, in Gelb die invertierten Repeats und die Fehlpaarungen in Lila. In B1 konnte der linke direkte Repeat nur versetzt aufgefunden werden.

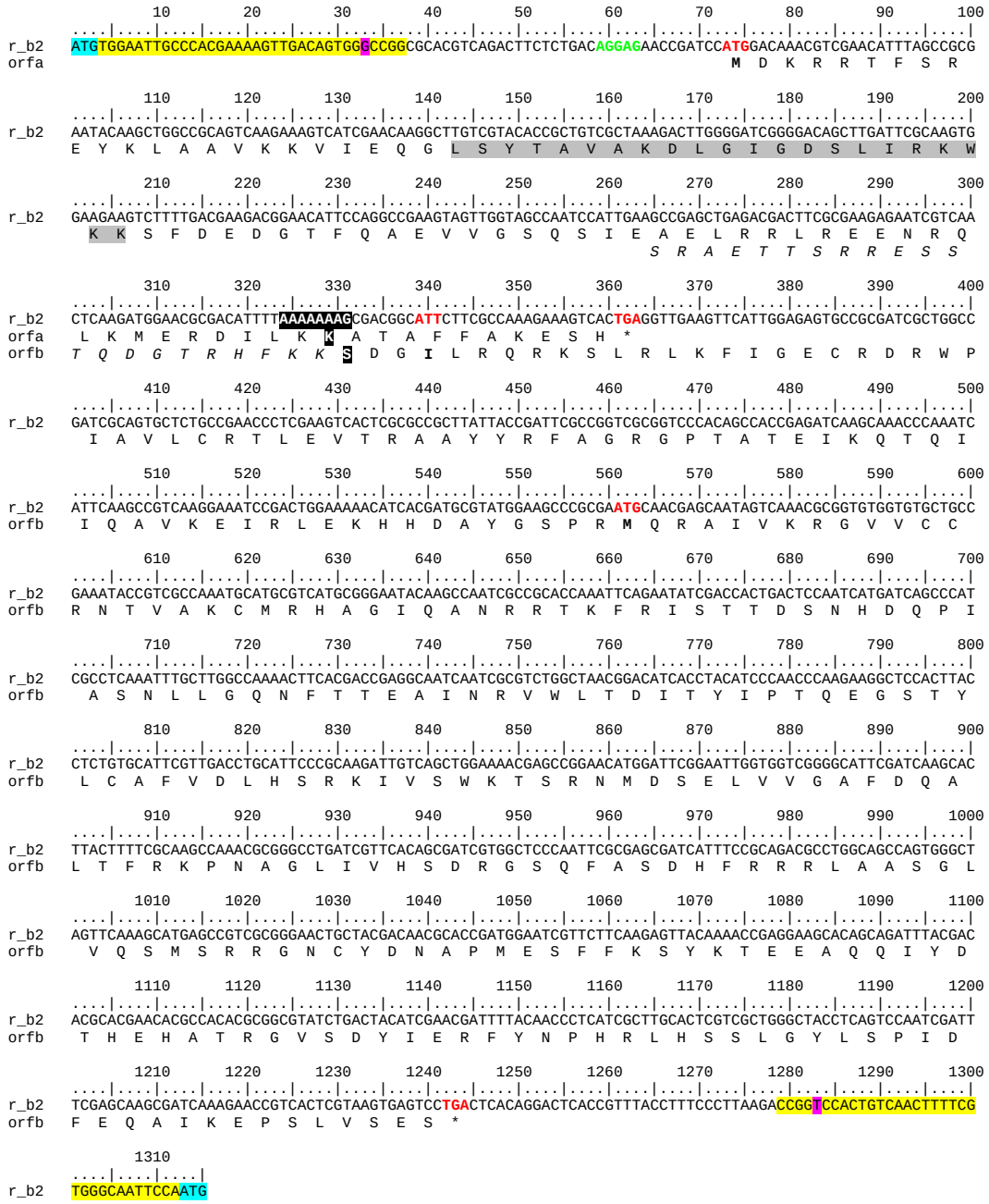
Die identifizierten Strukturen (vgl. Abb. 22, 23, 24 und 25) weisen auf ein IS-Element hin, bei dem vermutlich die Transposase als Fusionsprotein synthetisiert wird.

Abb. 24: Schematische Darstellung des IS-Elements im repetitiven Element B
Die identifizierten Elemente wie die direkten Repeats (DR), die flankierenden invertierten Repeats (IR), eine potenzielle ribosomale Bindestelle (RBS) sowie die im *Frameshift Window* überlappenden ORFs A und B spiegeln ein komplex aufgebautes IS-Element wider. Der Pfeil gibt die Transkriptionsrichtung unter Berücksichtigung der *Frameshift* Unterdrückung an.



IS1 und *IS911* (*IS3*-Familie) unterscheiden sich grundsätzlich in einigen Punkten. Zunächst besteht ein quantitativer Unterschied, da das *Frameshifting* Ereignis in *IS1* wesentlich seltener auftritt als in *IS3*. Ein weiterer Punkt ist die Expression eines zweiten Proteins als Resultat der Translation von *orfB* in *IS911*. Ein derartiges Produkt konnte in den *IS1* Elementen nie nachgewiesen werden. Ein weiterer Unterschied besteht vermutlich im Mechanismus der Transposition (Polard et al. 1991). *IS1* generiert kointegrative Strukturen, die mit dem replikativen Modus der Transposition übereinstimmen, zuzüglich zu simplen Insertionen (Galas & Chandler 1982). *IS911* wird im Gegensatz hierzu dem nicht replikativen Typ zugesprochen (Prère et al. 1990). Auf der Basis der zur Verfügung stehenden Informationen können zu diesen Unterschieden keine Aussagen getroffen werden.

Abb. 25: Überblick über das IS-Element am Beispiel des repetitiven Elements B2
Farbunterlegungen kennzeichnen in Blau die direkten Repeats, in Gelb die terminalen invertierten Repeats, in Grün die potentielle ribosomale Bindestelle, in Grau das Helix-Turn-Helix-Motiv, in Schwarz das *Frameshifting Window* und Substitutionen in Lila. Potenzielle Start- und Stoppcodon sowie die ribosomale Bindestelle werden farblich im Fettdruck dargestellt.



Frameshifting Windows konnten bisher bei *IS3* und *IS1* nachgewiesen werden. Das IS-Element vom repetitiven Element B zeigt die größten Übereinstimmungen zu Parametern der *IS3*-Familie (Mahillon & Chandler 1998) wie der Länge, den 3 bp langen direkten Re-

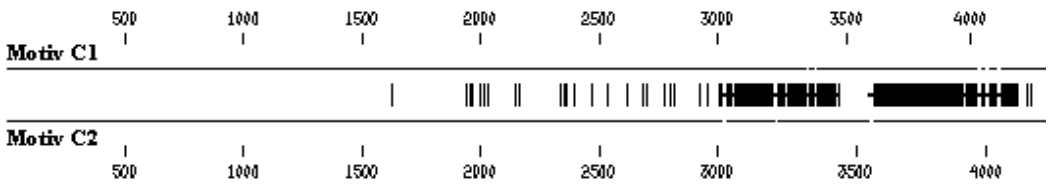
peats und den invertierten Repeatlängen von 31/43 bp (z.B. *Streptococcus mutans* L23843). Wie bereits beschrieben, konnten bei *orfB* keine Sequenzhomologien zu Sekundärstrukturen, die unterstützend auf das *Frameshifting* Ereignis einwirken könnten, identifiziert werden. Hieraus resultiert, dass die Regulation des identifizierten Elementes dem von *IS1* ähneln wird, verbunden mit einer niedrigen Expression eines potenziellen Fusionproteins.

Abschließend wird das IS-Element im repetitiven Element B trotz einer Reihe von Abweichungen der *IS3*-Familie zugeordnet, wobei hauptsächlich der Zusammenhang mit der retroviralen Integrase Domäne von *orfB* zu sehen ist.

3.2.2.4 Die repetitiven Elemente der Gruppe C

Das repetitive Element C liegt mit zwei Varianten im Genom vor, weitere Kopien oder Fragmente konnten nicht identifiziert werden. Die beiden Kopien differieren in der Länge um 15 Basen mit 4458 bp für das repetitive Element C1 und 4443 bp für C2. Sie zeigen erhebliche Abweichungen untereinander. So wird eine durchschnittliche Identität (BESTFIT) von 90,5% bei 42 Lücken erreicht. Die geringe Anzahl an Kopien im Genom verbunden mit dem hohen Anteil an Abweichungen in der Sequenz der Kopien (Abb. 26) lässt nicht vermuten, dass es trotzdem zu Problemen im Assembly der Shotgun-Daten kam. Die Ursachen liegen in den komplett kopiert vorliegenden Sequenzbereichen.

Abb. 26: Alignment der Varianten C1 und C2 in der GAPSHOW Übersicht
Die Grafik verdeutlicht den Wechsel zwischen höchst homologen und differenten Sequenzbereichen im Alignment der beiden Varianten des repetitiven Elements (in der Abbildung als Motiv C1 und C2 bezeichnet).
C1 und C2 wurden hierzu mit BESTFIT gegenübergestellt, mit GAPSHOW grafisch umgesetzt und überarbeitet abstrahiert dargestellt. Die Skalierung gibt die Positionen der jeweiligen Variante wider. Die beiden Sequenzen sind als waagerechte Linien dargestellt; die Unterbrechungen kennzeichnen Lücken des jeweiligen Stranges im Alignment. Die im letzten Drittel des Alignments zunehmenden Fehlpaarungen erscheinen durch die limitierte grafische Auflösung deutlich als schwarze Balken.
Lücken, die zur Aufrechterhaltung des Alignments eingefügt wurden, werden als kurze senkrechte Striche dargestellt.



Die BLASTX Analysen (Tab. 26) des repetitiven Elements C gegen die Datenbank von NCBI zeigen hohe Sequenzhomologien zu den Untereinheiten des Typ I Restriktions- und Modifikationssystems, die sich auch in den folgenden Analysen der ORFs widerspiegeln.

Tab. 26: Beispiel für das beste BLASTX Ergebnis für C1

Höchste Sequenzhomologie:
site-specific DNA-methyltransferase (adenine-specific), subunit M; *Methanosarcina acetivorans* Stamm C2A (NP_617320)

Length = 420	Identities = 243/356 (68%)
Score = 498 bits (1283)	Positives = 282/356 (78%)
e-value = e-144	Frame = +3

Mithilfe des Programmes ORPHEUS wurden für das repetitive Element C1 insgesamt neun ORFs und für C2 sechs ORFs vorhergesagt. Mit den repetitiven Elementen überlappende ORFs wurden in die Analyse einbezogen. Die vorhergesagten ORFs, die sich auf dem Gegenstrang befinden und zu denen keine Funktionsaussagen getroffen werden konnten, wurden in der folgenden Analyse nicht weiter berücksichtigt. Zusätzlich wurde *orf9453* in C2 (1560 bp, Position 6282816-6281257; Pendant zum in C1 nicht verwendeten *orf9430*, 1140 bp, Position 6267625-6266486), der sich im Widerspruch zu einer ehemals kodierenden Region befindet, nicht berücksichtigt.

Die bereits dargelegten Sequenzunterschiede (Abb. 26) zwischen den beiden Varianten spiegeln sich in den ausgewählten ORFs wider (Tab. 27). Insbesondere die ORFs *9428* (C1) und *9452* (C2) sowie *orf9431* (C1) und der nachträglich bestimmte *orfC2_verk* weichen bereits in ihrer Länge erheblich voneinander ab.

Tab. 27: Sequenzhomologien der einzelnen ORFs in den beiden Varianten des repetitiven Elements C zueinander

C1- ORFs	Positionen (Nukleotid/ Aminosäurenlänge)	Identitäten zueinander in der Nukleotidsequenz	Identitäten zueinander in der Aminosäure-sequenz	C2- ORFs	Positionen (Nukleotid-/ Aminosäurenlänge)
9423	6262455-6262622 (168 na/56 aa)	100%	100%	*9448*	6277280-6277447 (168 na/56 aa)
9424	6262799- 6263689 (891 na/297 aa)	100%	100%	*9449*	6277624-6278514 (891 na/297 aa)
9425	6263708- 6265363 (1656 na/552 aa)	98,5%	99,3%	*9450*	6278533- 6280188 (1656 na/552 aa)
9428	6265363-6266541 (1179 na/393 aa)	94,2% bei einer Alignmentlänge von 162 na!	46,3% bei einer Alignment-länge von 334 aa!	*9452*	6280320-6281354 (1035 na/345 aa)
9431	6266684-6269977 (3294 na/1098 aa)	93% bei einer Align-mentlänge von 137 na!	97,7% bei einer Alignment-länge von 44 aa!	*C2_verk*, entspricht in der Lokalisation *orf9431* der Variante C1	6281494- 6281721 (225 na/75 aa)

Unabhängig von den auftretenden Unterschieden zwischen den ORFs der beiden repetitiven Elemente sind die Ergebnisse der BLASTP Analyse eindeutig (Tab. 28).

Tab. 28: Ergebnisse der Datenbanksuchen zu den ORFs im repetitiven Element C

ORF[*1]	BLASTP	PFAM	COG	Zuordnung
C1_9423	-	-	-	Konservierter hypothetischer ORF
C2_9448	-	-	-	siehe oben
C1_9424	BAB83495: ORF24 [*Staphylococcus hominis*]; e-value 6e-22	-	-	Konservierter hypothetischer ORF, ähnlich ORF24 *(Staphylococcus hominis)*
C2_9449	siehe oben	-	-	siehe oben
C1_9425	AAM05800: *site-specific* DNA-Methyltransferase (adeninspezifisch), Untereinheit M [*Methanosarcina acetivorans* Stamm C2A]; e-value e-142.	1.) pfam02506, Methylase_M., Type I Restriktions-Modifikationssystem, M Protein 2.) pfam02384, N-6 DNA Methylase. Familie beinhaltet N-6 adeninspezifische DNA Methylase EC:2.1.1.72 vom Typ I und Typ IC Restriktionssystem.	COG0286: Type I Restriktions-Modifikations-system Methyl-transferase-Untereinheit	Restriktions-Modifikationssystem, Untereinheit M (Methylase)
C2_9450	siehe oben	siehe oben	siehe oben	siehe oben
C1_9428	AAM05801: Typ I Restrik-tions/ Modifikationsenzyme Protein S [*Methanosarcina acetivorans* Stamm C2A]; e-value 5e-27.	Methylase_S, Type I Restriktions-Modifikationssystem DNA spezifitäts Domä-ne (pfam01420), liegt zweimal vor.	COG0732: Re-striktionendo-nuclease S Unter-einheiten	Restriktions/Modifikationssystem, Untereinheit S (Spezifitäts-Domäne). Das doppelte Auftreten dieser Domäne erklärt sich aus den zwei spezifischen DNA-Bindestellen.
C2_9452	siehe oben; e-value 4e-25.	siehe oben	siehe oben	Restriktions/Modifikationssystem, Untereinheit S (Spezifitäts-Domäne); zum Teil homolog zu ORF9428 (C1)
C1_9431	AAM05804: Typ I site-specific Deoxyribonuclease Protein R [*Methanosarcina acetivorans* Stamm C2A]; e-value 0.0	DEAD, DEAD/DEAH Box Helicase (pfam00270)	COG0610: Restriktions-Modifikations-system Type I, Helicase Unter-einheiten und verwandte Helica-sen	Restriktions/Modifikationssystem, Untereinheit R (Restriktionsendo-nuclease)
C2_verk	siehe oben; e-value 9e-12	-	-	Restriktions/Modifikationssystem, Untereinheit R (Restriktionsendo-nuclease); zum Teil homolog zu ORF9431 (C1); deutlich verkürzt gegenüber vergleichbaren Unter-einheiten, vermutlich funktionslos (Pseudogen)

[*1]Index vor der ORF-Bezeichnung gibt die Variante des repetitiven Elements an.

Die ORFs *9423* und *9424* (C1), bzw. *9448* und *9449* (C2), stellen konservierte hypothetische ORFs dar, denen zurzeit keine Funktion zugeordnet werden kann. Ihre entgegengesetzte Orientierung zu den folgenden ORFs wird als Anhaltspunkt für einen fehlenden Funktionszusammenhang zum Restriktions-Modifikationssystem interpretiert. Die 100% Identität dieser ORFs im Vergleich der Kopien zueinander legt die Vermutung nahe, dass beide ORFs für aktive Proteine kodieren.

Die ORFs lassen sich eindeutig einem Restriktions-Modifikationssystem des Typ I zuordnen. Restriktions-Modifikationssysteme treten exklusiv in einzelligen Organismen, überwiegend Bakterien sowie einigen Viren, auf und schützen die Bakterienzelle vor der Invasion durch Fremd-DNA. Die Quellen dieser Invasionsversuche stellen überwiegend Bakteriophagen Genome und konjugative Plasmide dar (Redaschi & Bickle 1996). Das Prinzip des klassischen Restriktions-Modifikationssystems beruht auf der Annahme, dass Fremd-DNA durch die Endonuclease des Wirtes geschnitten wird, während die Schnittstellen im Wirtgenom durch sequenzspezifische Methylierung der Andenosinreste geschützt sind (Thorpe et al. 1997). Wie effektiv diese Strategie in der Natur wirklich ist, kann abschließend nicht beurteilt werden, da es vorkommt, dass Phagenpartikel dem Restriktionsenzym entkommen und methyliert werden (Schouler et al. 1998b).

In Typ I Restriktions-Modifikationssystemen sind Restriktions- und Modifikationsaktivitäten in einem heteromeren Enzymkomplex zusammengefasst. Dieser Enzymkomplex besteht aus der Spezifitäts-Untereinheit (Untereinheit S; beinhaltet eine Reihe von Einzel-Domänen, zwei DNA-Bindestellen), der Modifikationsuntereinheit (Untereinheit M; zwei Domänen) und der Restriktionsuntereinheit (Untereinheit R; zwei Restriktionsdomänen). M, S und R Untereinheit bilden das komplette Restriktions-Modifikationssystem (Fuller-Pace & Murray 1986; Thorpe et al. 1997). Diese Enzyme binden an die charakteristische zweiteilige DNA-Zielstelle, meist eine 3 bp lange Sequenz sowie eine 4-5 bp lange Sequenz separiert durch einen 6-8 bp lange unspezifische Spacer-Sequenz (Thorpe et al. 1997). M und S Untereinheit sind zur erfolgreichen Methylierung notwendig; gemeinsam bilden sie die aktive Methyltransferase (Thorpe et al. 1997). Die Methyltransferase transferiert eine Methylgruppe von S-Adenosyl-Methionin zur N6-Position eines spezifischen Adeninrestes auf jeden Strang der DNA-Zielstelle. Die Methylasen besitzen die gleiche Sequenzspezifität wie die zugehörige Restriktionsendonuclease (Thorpe et al. 1997).
M und S Untereinheit liegen in einem Operon vor, während die Untereinheit R durch einen eigenen Promoter exprimiert wird. Die Untereinheit R ist nur für die Restriktion verantwortlich (Redaschi & Bickle 1996). Die Untereinheit R der Type I Restriktions-Modifikationssysteme besitzt vermutlich eine Helicase-Aktivität, die eine Rolle beim Entwinden der DNA an der Schnittstelle und bei der DNA-Translokation spielen könnte (Schouler et al. 1998a).

Falls die Erkennungsstelle nicht modifiziert ist, wird die DNA durch den DNA-Enzym-Komplex translokalisiert. Der DNA-Enzym-Komplex erhält durch die Mg^{2+} abhängige ATP-Hydrolyse freie Energie. Bei Enzymen wie EcoB1 und EcoK1 konnten entspannte

Loop- und supercoiled Loop-DNA beobachtet werden. Diese Beobachtungen führten zu der Überlegung, dass die DNA durch das Zusammenspiel zweier Enzym-Komplexe translokalisiert werden kann. Die DNA wird dann an einer zufälligen Stelle, die entfernt von der Erkennungsstelle liegt, geschnitten (Janscak & Bickle 1998).

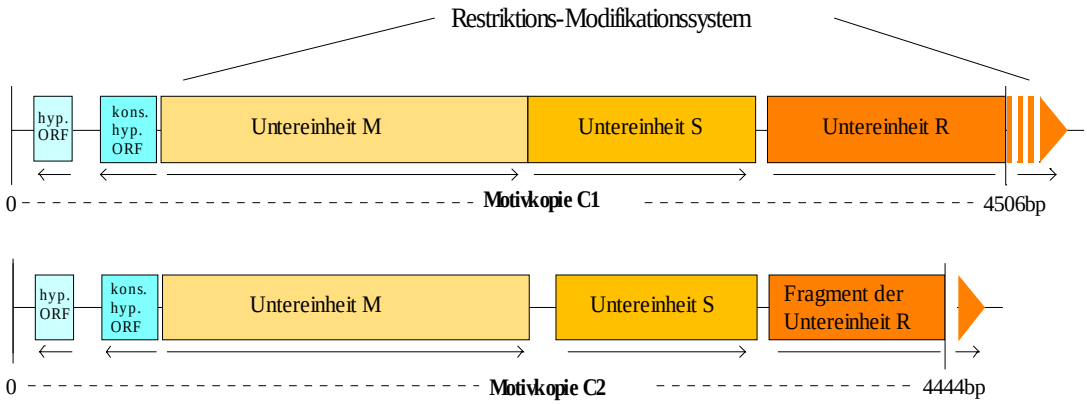
Im Fall einer hemimethylierten Erkennungsstelle oder eines ATP-Mangels bei nichtmodifizierter DNA fehlt dem Enzym-Komplex N6-Adenin DNA-Methyltransferaseaktivität und es kommt zu einer Methylierung eines spezifischen Adeninrestes in jedem Strang der Erkennungsstelle unter Nutzung von S-Adenosylmethionine als Methyldonor. Die Untereinheit R ist für die Methylierung nicht notwendig, so kann eine monofunktionale DNA-Methyltransferase mit einer Untereinheitenzusammensetzung von M_2S_1 geformt werden (Janscak & Bickle 1998).

Die im repetitiven Element C1 identifizierten ORFs *9425, 9428* und *9431* zeigen weitgehend die beschriebenen Strukturen und Sequenzhomologien. *orf9425* und *orf9428* (C1) bilden hierbei die charakteristische Operonstruktur der Methyltransferase. Die Untereinheit M (*orf9425*) zeigt die zwei Methylase Domänen, darunter die charakteristische N-6 adeninspezifische DNA-Methylase. Die Untereinheit S (*orf9428*) zeigt die entsprechende Methylase-S Domäne. Wie erwartet, zeigt die Untereinheit R (*orf9431*) Ähnlichkeiten zum Helicase Motiv; wirtspezifische R Domänen können zurzeit mit bioinformatischen Methoden nicht bestimmt werden (Tab. 28)

In der Variante C2 liegt die Operonstruktur von Untereinheit M (*orf9450*) und S (*orf9452*) nicht mehr vor. In der Variante C2 liegt *orf9452*, der für die Untereinheit M kodiert, weitgehend konserviert vor. Die Untereinheit S zeigt sich durch Mutationen verändert, erreicht jedoch eine ähnliche Länge in der abgeleiteten Peptidsequenz. Obwohl die funktionale Domäne identifiziert werden konnte, kann zurzeit keine Aussage getroffen werden, ob die Bildung einer funktionstüchtigen Methyltransferase beeinträchtigt wird. Mit hoher Wahrscheinlichkeit ist die Untereinheit R (*C2_verk*) durch ihre extreme Verkürzung (Abb. 27), die unter anderem den Verlust der Helicase Domäne zur Folge hat, funktionslos. Das verstärkte Auftreten von Deletionen und Substitutionen unterstützt diese Interpretation.

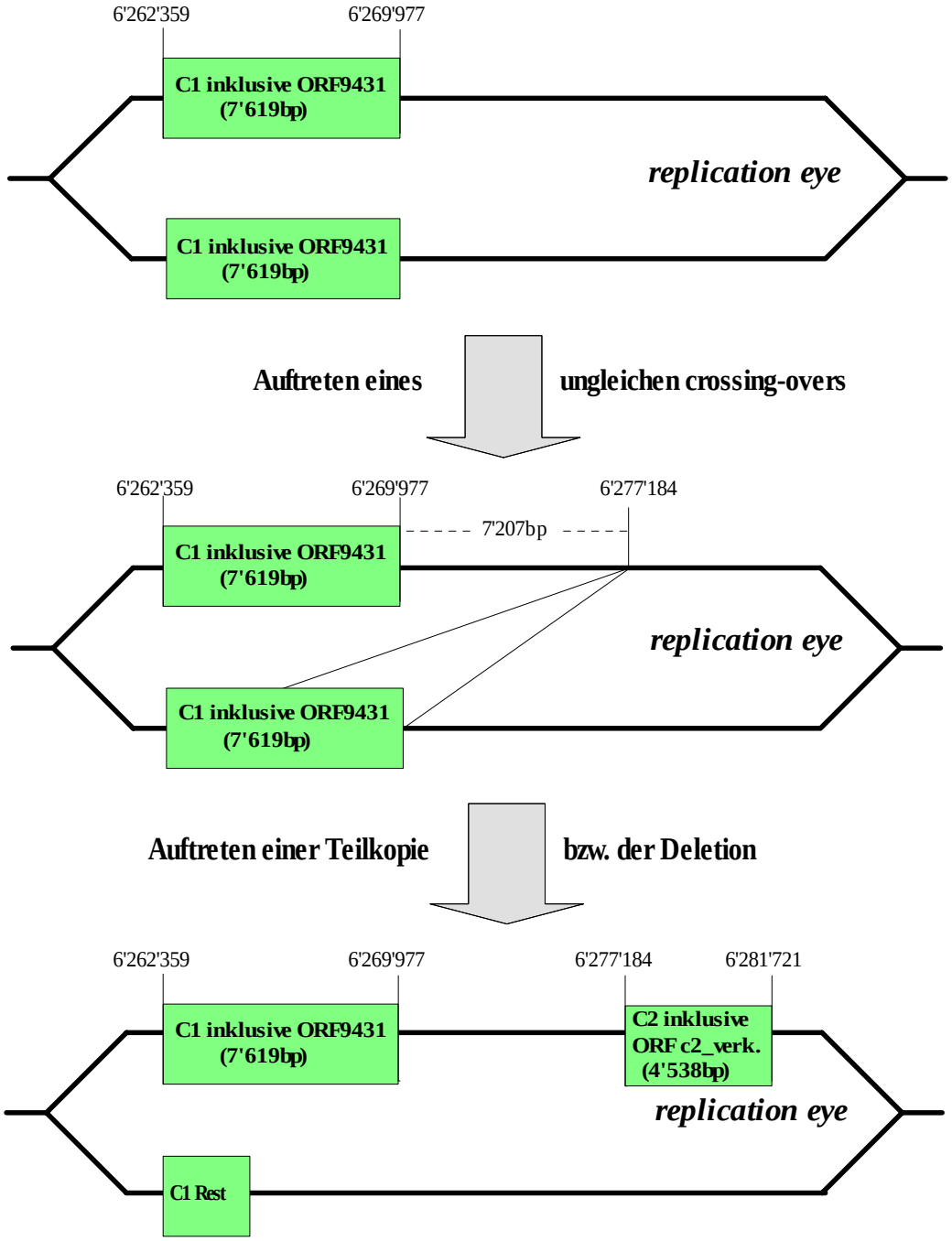
Abb. 27: Organisation, Vergleich und Funktionszuweisung der im repetitiven Element C enthaltenen ORFs

Dargestellt werden die beiden hypothetischen ORFs und die drei Untereinheiten des identifizierten Restriktions-Modifikationssystems. In beiden Kopien reicht die Untereinheit R über die identifizierte Länge des repetitiven Elements hinaus. Die Kopie der Untereinheit R liegt im repetitiven Element C2 jedoch innerhalb des Elements als auch außerhalb deutlich verkürzt vor. Durch umfangreiche Deletionen liegt das gesamte Element C2 deutlich verkürzt vor.



Über die Ursachen, die zum Auftreten der Kopie (C2) im Genom führten, kann nur spekuliert werden. Für eine Transposition, die im Zusammenhang mit der Genese einer zweiten Kopie im Genom steht, konnten keine Anhaltspunkte gefunden werden. Eine mögliche Erklärung bietet das Auftreten eines ungleichen Crossing-overs während der Replikation, die zu einer Genamplifikation in diesem prokaryotischen System führte (Romero & Palacios 1997). Während der Replikation wäre es dann im *replication eye* zum ungleichen Crossing-over gekommen (Abb. 28, die zu einer homologen Kopie des Restriktions-Modifikationssystems im Genom führte (Redaschi & Bickle 1996). In Übereinstimmung mit dieser Hypothese (Abb. 28) stehen die benachbarte Lokalisation der Kopien im Genom und die fehlende Möglichkeit *orf9431* und *C2_verk* (inklusive der benachbarten Sequenz) sinnvoll im Alignment gegenüberzustellen. Homologien des *orf9431* zu *orfc2_verk* konnten außerhalb des Leserasters nicht mehr nachgewiesen werden. Es kann deshalb angenommen werden, dass diese Bereiche des *orf9431* vermutlich niemals dupliziert vorlagen. Unter dieser Annahme wäre der *orfc2_verk* bereits nach der Duplikation funktionslos gewesen. Auftretende Mutationen hätten vermutlich zu keinem Selektionsnachteil für den Organismus geführt und treten somit im *orfc2_verk* verstärkt auf.

Abb. 28: Erklärungsversuch zur Genese der repetitiven Elemente C1 und C2 im Genom
Die schematische Darstellung orientiert sich an dem Erklärungsmodell zum Auftreten von ungleichen Crossing-overs (*unequal crossing-over*) während der Replikation in Bakteriengenomen (Romero & Palacios 1997). Die schwarzen Linien symbolisieren hierbei den DNA-Doppelstrang. Auf eine differenzierte Ausweisung der einzelsträngigen Regionen oder der Replikationsgabel wurde aus Gründen der Übersichtlichkeit verzichtet. Der obere und der untere Balken stehen für die semikonserativ replizierten Stränge des Genoms. Das repetitive Element wird als grün unterlegte Box mit den Genompositionen kenntlich gemacht. Der grün dargestellte Bereich umfasst neben den Varianten, den zum Restriktions-Modifikationssystem gehörenden *orf9431* (C1) sowie den neu bestimmten homologen *ORFc2_verk*.



3.2.2.5 Die repetitiven Elemente der Gruppe D

Das repetitive Element D wurde während der Assemblierung mit einer maximalen Länge von insgesamt 1774 bp bestimmt. Es liegt insgesamt fünfmal im Genom vor, davon einmal, Variante D1, als Fragment. Die erhaltenen Bereiche liegen konserviert vor (Tab. 29), weitere Fragmente konnten nicht aufgefunden werden.

Tab. 29: Abweichungen der Varianten zum gemeinsamen Consensus
Als Basis wurde der hochkonservierte 1760 bp lange Sequenzbereich gewählt.

Variante	Anzahl der Abweichungen	Prozentuale Abweichung
D1*	35 Substitutionen und 2 Deletionen	-*
D2	41 Substitutionen	2,3
D3	-	0,0
D4	1 Substitution und 1 Insertion	0,1
D5	9 Substitutionen	0,5

*Variante D1 liegt lediglich mit einer Gesamtlänges von 572 bp verkürzt vor.

Die BLASTX Analyse der repetitiven Elemente der Gruppe D gegen die Datenbank von NCBI zeigt Verwandtschaften zu Genen, denen eine Transposasefunktion zugesprochen wird (Tab. 30).

Tab. 30: BLASTX Resultate am Beispiel von Variante D3

Organismus	Acc. Number	Gen-funktion	Länge (aa)	Score	e-value	Identitäten	Positives	Gaps	Frame
Porphyromonas gingivalis	AAD38020	Transposase	385	82.8 bits (203)	1e-14	68/240 (28%)	109/240 (45%)	-	-1
Bacillus subtilis	BAA92234	Transposase	374	76.3 bits (186)	1e-12	83/343 (24%)	140/343 (40%)	1/343 (0%)	-1
Bacillus halodurans	NP_241557	Transposase	371	67.8 bits (164)	4e-10	59/203 (29%)	90/203 (44%)	16/203 (7%)	-1
Bacillus stearothermophilus	CAA48047	Transposase	377	65.5 bits (158)	2e-09	60/224 (26%)	95/224 (41%)	1/224 (0%)	-1
Rhizobium sp. NGR234	NP_444173	mutmaßliche Transposase *Y4ZB*	356	62.0 bits (149)	2e-08	49/226 (21%)	103/226 (44%)	3/226 (1%)	-1

Das Programm ORPHEUS sagt unterschiedliche ORFs für die jeweiligen Varianten voraus (Tab. 31). Die Ursachen liegen (1) in den Sequenzunterschieden, die unterschiedlich lange Leserahmen ermöglichen, (2) in außerhalb der Elemente liegenden Sequenzrandbereichen, in denen mögliche ORFs beginnen und (3) in der fehlenden Berücksichtigung der vorliegenden Struktur.

Tab. 31: Mit ORPHEUS vorhergesagte ORFs im Element D
In den Elementen einander entsprechende ORFs wurden nur jeweils einmal pro Element gefunden (Fettdruck). Die Verwendung unterschiedlicher Startcodons wurde hierbei berücksichtigt. Abweichungen in der ORF-Vorhersage resultieren in Differenzen der Sequenzen der Varianten zueinander und der Einbeziehung der ORFs, die aus den sich dem Element anschließenden Sequenzbereiche in das Element hinein- oder herausreichen (in Grün).

Kopie	ORF	Position im Genom	Länge in Nukleotide/Aminosäuren
D1	*orf9206*	6133854 - 6134243	390/130
	orf9207*	6134589 - 6134347	243/81
	orf9208	6134563 - 6134673	111/37
	orf9209	6136270 - 6134660	1611/537
D2	*orf3116*	2001756 - 2001289	468/156
	orf3117	2001755 - 2002951	1197/399
	orf3118	2002923 - 2003249	327/109
D3	*orf3853*	2465774 - 2465421	354/118
	orf3854	2465920 - 2465735	186/62
	orf3855	2465885 - 2467081	1197/399
	orf3856	2467011 - 2467286	276/92
D4	*orf7630*	5048582 - 5048169	414/138
	orf7631	5048517 - 5049890	1374/458
	orf7632	5050041 - 5049913	129/43
D5	*orf7975*	5283128 - 5282688	441/147
	orf7976	5283274 - 5283089	186/62
	orf7977	5283239 - 5284435	1197/399
	orf7978	5284365 - 5284670	306/102

*liegt als Fragment vor

Nur jeweils einem ORF pro Elementvariante kann eine potenzielle Funktion zugeordnet werden. Diese ORFs (D2_*orf3117*, D3_*orf3855*, D4_*orf7631*, D5_*orf7977*) spiegeln zudem die Transposasenzuweisung aus den BLASTX-Suchen im BLASTP (Tab. 32) wider.

Tab. 32: BLASTP Resultate der innerhalb von Element D identifizierten ORFs am Beispiel des *orf3855* (D3)

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Porphyromonas gingivalis	AAD38020	Transposase	385	85.5 bits (210)	1e-15	68/240 (28%)	109/240 (45%)	24/240 (10%)
Bacillus subtilis	BAA92234	Transposase	374	76.6 bits (187)	5e-13	89/356 (25%)	140/356 (39%)	55/356 (15%)
Bacillus halodurans	NP_241557	Transposase	371	72.0 bits (175)	1e-11	59/203 (29%)	91/203 (44%)	40/203 (19%)
Bacillus stearothermophilus	Q45620	vermutliche Transposase für das IS-Element *IS5377*	377	67.0 bits (162)	4e-10	60/225 (26%)	99/225 (43%)	29/225 (12%)
Rhizobium sp. NGR234	NP_444173	vermutliche Transposase *Y4ZB*	356	60.5 bits (145)	4e-08	49/226 (21%)	103/226 (44%)	28/226 (12%)

Die Elementvariante D1 kann durch die vorliegende Verkürzung nur einen entsprechend fragmentierten ORF (*orf9207*) aufweisen. Die maximale bei D4 erreichte ORF-Länge lässt sich nicht auf die anderen Elementvarianten übertragen, weil eine in D4 vorliegende Insertion die Verlängerung des Leserahmens erzwingt. Da diese Insertion im Element D einmalig ist, wird sie als zufällig interpretiert, zudem die im *orf7631* erreichte ORF-Länge über der, der vergleichbaren Transposasen liegt.

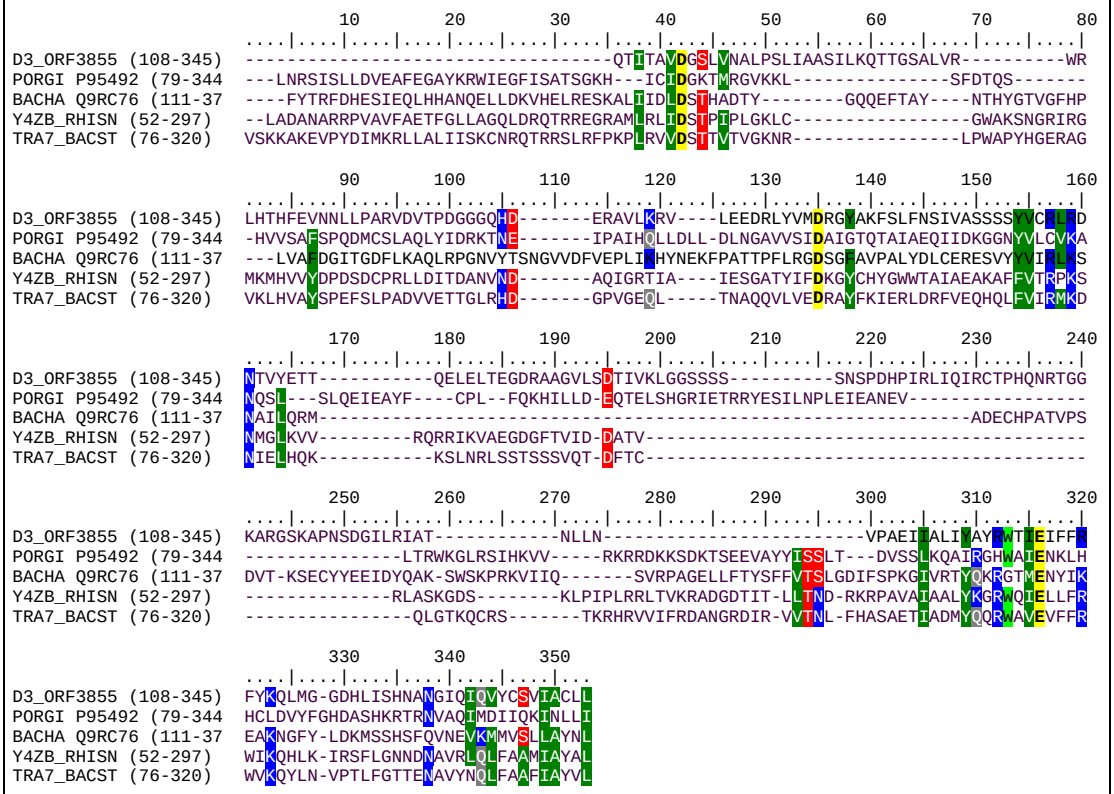
Ausgehend von dem *orf3855* der Elementvariante D3, die keine Abweichungen zum gemeinsamen Consensus der Elementvarianten zeigt, soll die Charakterisierung des Elements D aufgebaut werden. Diese Interpretation steht mit den anderen Elementvarianten in Übereinstimmung. Auf der Basis von *orf3855* wurde für die Elementvariante D4 *orf7631* verkürzt, der hierdurch eine Länge von 1197 bp erreicht (Positionen 5048694-5049890). Abweichungen zur gemeinsamen in Aminosäuren übersetzten Sequenz treten im *orf3117* (D2) mit 9 Austauschen (vier konservierte Austausche) und *orf9207+* (D1) mit 3 Austauschen (ein konservierter Austausch) im vorliegenden Leserahmen auf.

Die Zuordnung zu einer Transposasefunktion wird durch die Sequenzhomologien zur Transposase 11 (Pfam01609) der *IS4*-Familie (IPR002559) und COGs (COG3385: predicted Transposase) bekräftigt. Die potenziellen Transposasen des Elementes D beinhalten das charakteristische DDE-Motiv (Abb. 29). Der charakteristische Abstand von D2 zu E von über 100 aa (Mahillon & Chandler 1998) im *IS4* wird in der Transposase des Elements D erreicht.

Die identifizierten Homologien weisen darauf hin, dass es sich bei dem repetitiven Element D um ein IS-Element handelt. In den Randbereichen des Elements D konnten die charakteristischen invertierten und direkten Repeats eines IS-Elements identifiziert werden (Abb. 30). In Übereinstimmung mit den einzelnen Elementvarianten kann nur ein gemeinsamer ORF identifiziert werden, der als kodierend angesehen wurde, weshalb die anderen vorhergesagten ORFs vernachlässigt wurden. Das Auftreten lediglich eines langen ORFs im IS-Element, der den größten Raumanteil einnimmt, tritt innerhalb der *IS4*-Familie häufiger auf (Mahillon & Chandler 1998; Wang et al. 1997).

Abb. 29: Alignment zwischen dem *orf3855* und Transposasen der Pfam 01609 Transposase 11 zur Verdeutlichung des konservierten DDE Motives

Dargestellt werden vier Transposasen, die an der Bildung der Transposase 11 Domäne in Pfam beteiligt sind, versus dem *orf3855* aus der Elementvariante D3 stellvertretend für die anderen D Varianten. Identische und ähnliche Aminosäuren der konservierten Bereiche sind farblich markiert. Das charakteristische DDE-Motiv wird im Fettdruck und Gelb unterlegt dargestellt.



Abkürzungen der Organismennamen:
Porphyromonas gingivalis: PORGI
Bacillus halodurans: BACHA
Rhizobium sp.: RHISN
Bacillus stearothermophilus: BACST

Die Elemente D3 und D4 besitzen jeweils 44 bp lange invertierte Repeats mit zwei Fehlpaarungen. Die Zielstellenverdopplung resultierte in 3 bp langen in unterschiedlichen direkten Repeats, die an einem Ende durch 2 bp vom invertierten Repeat getrennt sind (Abb. 30). Derartig lange invertierte Repeats wurden nur selten aufgefunden, z.B. in *IS408* mit bis zu 48 bp (Mahillon & Chandler 1998). Längere invertierte Repeats konnten im *IS5376* mit 50b identifiziert werden, deren Transposase Sequenzhomologien zu den potenziellen im repetitiven Element D zeigt (Xu et al. 1993). Beide werden in die *IS21* Familie gruppiert, weichen jedoch in Sequenz, Organisation und direkten Repeats von dem vorliegenden IS-Element ab (Mahillon & Chandler 1998).

Element D2 und D5 weichen in ihren Repeatmustern von D3 und D4 ab. Ihre invertierten Repeats zeigen bei zwei Fehlpaarungen lediglich eine Länge von 30 b, an die sich 9 bp

lange direkte Repeats flankierend anschließen. 9 bp lange direkte Repeats erinnern an das *IS6120* (Guilhot et al. 1992), bei dem die invertierten Repeats mit einer Länge von 24 bp sich im gleichen Bereich bewegen.

Abb. 30: Alignment der Element D Randbereiche
Farbunterlegungen kennzeichnen in Blau die außerhalb des repetitiven Element liegenden direkten Repeats, in Gelb die invertierten Repeats. Die graue Farbunterlegung markiert Bereiche der invertierten Repeats, die Fehlpaarungen darstellen.

Hiermit weichen die IS-Elemente des Elements D gleich in zwei Charakteristika voneinander ab. Element D1, welches nur noch als Fragment vorliegt, ist aufgrund der invertierten Repeats auch dem Typus von D2 und D5 zuzuordnen. Über die Ursachen der unterschiedlichen Strukturen lässt sich nur spekulieren. Die ungewöhnliche Länge der invertierten Repeats ist vermutlich nicht zwingend zur Basenpaarung notwendig, der Verlust eines Teiles der invertierten Repeats scheint die Transposition nicht zu beeinträchtigen. Die Veränderung des Ablaufes der Insertion, aus der die deutlich verlängerten direkten Repeats hervorgehen, können durch unterschiedliche Hypothesen erklärt werden. Möglicherweise kommt es bei der Insertion zu einer 3 bp langen oder 9 bp langen Zielstellenverdopplung. Derartige Schwankungen, z.B. 8-14 bp bei *IS1b* (Mahillon & Chandler 1998), sind bekannt. Sie können nicht in der Transposase des Elements D kodiert sein, da D2 und D5 keine gemeinsamen Abweichungen vom Consensus bzw. D3 und D4 besitzen. Eine weitere Hypothese könnte auf die gemeinsamen Abweichungen zurückgehen. Danach könnten die invertierten Repeats auch Einfluss auf die Zielstellenverdopplung nehmen.

Abschließend werden die im repetitiven Element D identifizierten Strukturen (Abb. 31) denen eines IS-Elements zugeordnet. Die festgestellten Varianten im Element D beeinträchtigen diese Zuordnung nicht, lassen jedoch auf unterschiedliche Mechanismen der Transposition während der Genese schließen. Sequenzhomologien der Transposase lassen auf eine weitläufige Verwandtschaft zur *IS4*–Familie schließen. Die festgestellte Verkürzung des Elements D1 kann die Folge eines ungleichen Crossing-overs sein (vgl. Kap. 3.2.2.4).

Abb. 31: Schematische Darstellung des IS-Elements im Element D
Die schematische Darstellung basiert auf den IS-Elementen der Elemente D3 und D4, die deutlich längere invertierte Repeats beinhalten. Die im repetitiven Element D identifizierten Elemente wie die direkten Repeats (DR), die flankierenden invertierten Repeats (IR) und die potenzielle Transposase spiegeln ein einfach aufgebautes IS-Element mit Ähnlichkeiten zu *IS4* wider. Der Pfeil gibt die Transkriptionsrichtung an.

3.2.2.6 Die repetitiven Elemente der Gruppe E

Die repetitiven Elemente der Gruppe E wurden mit einer Länge von insgesamt 1474 bp bestimmt. Die Elemente stellen sich mit ihren drei Kopien im Genom weitgehend konserviert dar (Tab. 33). Weitere Fragmente konnten im Genom nicht aufgefunden werden.

Tab. 33: Abweichungen der repetitiven Elemente zum gemeinsamen Consensus
Als Basis wurde der hochkonservierte 1474 bp lange Sequenzbereich gewählt.

Element	Anzahl der Abweichungen	Prozentuale Abweichung
E1	0	0,00%
E2	2	0,14%
E3	12	0,81%

Die BLASTX Analyse der repetitiven Elemente der Gruppe E gegen die Datenbank von NCBI zeigt Verwandtschaften zu Genen, denen eine Transposasefunktion zugesprochen wird (Tab. 34).

Tab. 34: BLASTX Resultate am Beispiel von E1

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Halobacterium sp. NRC-1	AAG21037	Vng6442h (Transposase)*	454	123 bits (308)	6e-27	117/448 (26%)	188/448	25/448 (5%)	+1
Methanosarcina acetivorans str. C2A	NP_616360	Transposase	477	95.9 bits (237)	1e-18	109/453 (24%)	188/453 (41%)	27/453 (5%)	+1
Vibrio cholerae	BAA33622	vermutliche Transposase	372	95.1 bits (235)	2e-18	88/338 (26%)	151/338 (44%)	11/338 (3%)	+1
Deinococcus radiodurans	NP_051698.1	vermutliche Transposase	333	86.7 bits (213)	7e-16	86/316 (27%)	130/316 (40%)	10/316 (3%)	+1
Bacillus stearothermophilus	T44628	vermutliche Transposase *ISBst12*	482	80.1 bits (196)	6e-14	101/432 (23%)	173/432 (39%)	13/432 (3%)	+1

Das Programm ORPHEUS sagt jeweils einen kodierenden Bereich innerhalb des Elements E voraus (Tab. 35). Überlappende ORFs mit Hinweisen auf ihre Funktion bestehen nicht.

Tab. 35: Transposase-verwandte ORFs im Element E

Element	ORF	Position im Genom	Länge in Aminosäuren	Abweichungen zum Consensus vom einheitlichen Startcodon ausgehend (na/aa)
E1	*3864*	2'471'205 - 2'472'596	464	0/0
E2	*9194*[*]	6'128'565 - 6'127'208 (6'128'599 - 6'127'208)	464 (462)	0/0
E3	*3450*	2'216'575 - 2'217'966	464	11/5

[*] *orf9194* wurde durch ORPHEUS zu kurz vorhergesagt (Originalvorhersage in Klammern).

Die innerhalb der Gruppe E lokalisierte ORFs spiegeln die Resultate der BLASTX Suchen wider (Tab. 36).

Tab. 36: BLASTP Resultate der innerhalb von Gruppe E identifizierten ORFs am Beispiel des *orf3864*

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Halobacterium sp. NRC-1	NP_395902	Vng6442h	454	112 bits (280)	1e-23	117/448 (26%)	188/448 (41%)	45/448 (10%)
Vibrio cholerae	T44318	vermutliche Transposase	372	93.2 bits (230)	6e-18	85/335 (25%)	148/335 (43%)	17/335 (5%)
Methanosarcina acetivorans str. C2A	AAM04840	Transposase	477	89.7 bits (221)	7e-17	109/453 (24%)	190/453 (41%)	38/453 (8%)
Deinococcus radiodurans	H75637	vermutliche Transposase	333	82.4 bits (202)	1e-14	86/316 (27%)	130/316 (40%)	19/316 (6%)
Methanosarcina mazei Goe1	NP_633760	Transposase	377	79.0 bits (193)	1e-13	90/367 (24%)	154/367 (41%)	35/367 (9%)

Eine Zuordnung zu Interpro-Einträgen konnte nicht vorgenommen werden. Im Gegensatz hierzu ist eine Zuordnung zu COGs möglich, da die Transposasen aus den BLAST-Suchen zusammengefasst vorliegen und einer Transposasenfunktion zugeordnet werden.

Für das charakteristische DDE-Motiv liegen in den ORFs und in COGs eine ganze Reihe von Möglichkeiten vor, so dass es leider nicht möglich war, es präzise zu identifizieren. Ursache für diese Problematik ist die Heterogenität des Alignments von COG3436, in dem vermutlich Fehler vorliegen.

Die vermutliche Transposase wird durch 13 bp lange perfekte terminale IRs flankiert, die sich unter Berücksichtigung von zwei Fehlpaarungen verlängern lassen (Abb. 32). Die Zielstellenverdopplung resultiert in acht Basen langen direkten Repeats flankierend der

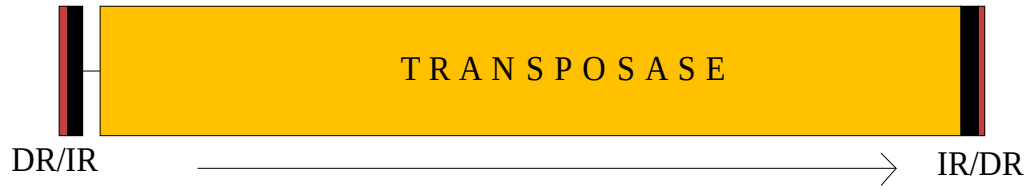
Elemente E1 und E2. Im Element E3 liegt die Zielstellenverdopplung vermutlich überformt vor und als weiterer Repeat im Element. Zielstellenlängen und die damit verbundenen direkten Repeats werden als charakteristisch für die jeweilige Transposase angesehen (Haren et al. 1999).

Abb. 32: Alignment der Randbereiche der repetitiven Elemente der Gruppe E
Farbunterlegungen kennzeichnen in Blau die außerhalb des Element liegenden direkten Repeats, in Gelb die invertierten Repeats. Die graue Farbunterlegung markiert den Bereich um die invertierten Repeats, die nur unter Berücksichtigung von zwei Fehlpaarungen verlängerbar sind. Start- und Stoppcodons der vermutlichen Transposase sind in Rot markiert. Unterstrichen wird ein direkter Repeat im Element E3 dargestellt, der im Bereich der Zielstellenverdopplung liegt.

```
                10        20              70        80              1470       1480
     ....|....|....|....|....|. ... ....|..    ..| ...  ...  .|....|....|....|....|...
e1   AAACGAAGTGTACTTGTTTACCCAGC     AATTCGC     ATG             GCACGGTAAACAAGTACAACGAAGT
e2   ATCTTTGTTGTACTTGTTTACCCAGC     AATTCGC     ATG  Transp.   GCACGGTAAACAAGTACTCTTTGTT
e3   AATTCGCGCGTACTTGTTTACCCAGC     AATTCGC     ATG             GCACGGTAAACAAGTACAATTGTAT
       DR        IR                                                IR        DR
        →         →                                                 ←         →
```

Die Einzelelemente der Gruppe E zeigen die charakteristischen Elemente eines IS-Elements auf (Abb. 33). Eine Zuordnung zu einer bekannten IS-Familie konnte jedoch nicht vorgenommen werden.

Abb. 33: Schematische Darstellung der repetitiven Elemente der Gruppe E
Die in der Gruppe E identifizierten Elemente, wie die direkten Repeats (DR), die flankierenden invertierten Repeats (IR), die vermutliche Transposase, spiegeln ein einfach aufgebautes IS-Element wider. Der Pfeil gibt die Transkriptionsrichtung an.



3.2.2.7 Die repetitiven Elemente der Gruppe F

Die Gruppe F wurde während der Assemblierung mit einer Länge von insgesamt 1511-1513 bp bestimmt. Sie stellt sich unter Vernachlässigung der Randbereiche mit seinen vier Kopien im Genom als hochkonserviert dar, Sequenzabweichungen treten nicht auf. Weitere Fragmente konnten im Genom nicht aufgefunden werden.

Die BLASTX Analyse des repetitiven Elements F gegen die Datenbank von NCBI zeigt Verwandtschaften zu Transposasen, die Teile von bekannten *IS4*-Elementen darstellen (Tab. 37).

Tab. 37: BLASTX Resultate am Beispiel von F1

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	Expect	Identities	Positives	Gaps	Frame
Pantoea agglomerans	CAA57483	ORF	440	52.0 bits (123)	2e-05	99/458 (21%)	173/458 (37%)	21/458 (4%)	-3
Shigella flexneri	NP_085177	IS4 orf	448	49.3 bits (116)	1e-04	87/400 (21%)	152/400 (37%)	9/400 (2%)	-3
Escherichia coli K12	NP_418698	Transposase insG für IS-Element	442	48.5 bits (114)	2e-04	85/400 (21%)	152/400 (37%)	9/400 (2%)	-3

Das Programm ORPHEUS sagt jeweils zwei kodierende Bereiche im repetitiven Element F voraus (Tab. 38). Weitere ORFs ohne Ähnlichkeiten zu Datenbankeinträgen ragen in die Elemente herein. Sie werden als fraglich angesehen und in den folgenden Analysen nicht berücksichtigt. Die im Element F liegenden ORFs weichen lediglich im Startcodon um ein Codon voneinander ab. Eine generelle Verlängerung dieser ORFs bis zum gemeinsamen Start (ATG) wurde vorgenommen, da eine zugehörige potenzielle ribosomale Bindestelle identifiziert werden konnte. Die betroffenen ORFs wurden entsprechend verlängert und mit einem „+“ gekennzeichnet. Abweichungen in der Nukleotid- und Aminosäuresequenz zueinander bestehen nicht.

Tab. 38: Transposase-verwandte ORFs in den repetitiven Elementen der Gruppe F

Element	*orfA* (na/aa)	Position im Genom	*orfB* (na/aa)	Position im Genom
F1	*orf8644+* (1371/457)	2765787-2764417	*orf8645* (480/160)	2764996-2765475
F2	*orf3453* (1371/457)	2221207-2219837	*orf3453* (480/160)	2220416-2220895
F3	*orf3465+* (1371/457)	2226866-2228236	*orf3466* (480/160)	2227657-2227178
F4	*orf4301* (1371/457)	5759903-5761273	*orf4302* (480/160)	5760694-5760215

Die innerhalb des repetitiven Elements F lokalisierten ORFs der Gruppe A spiegeln die Resultate der BLASTX Suchen im BLASTP (Tab. 39) wider. Der abweichende beste Treffer zu einem hypothetischen Protein von *Nostoc punctiforme* wird als Folge einer automatischen ORF-Vorhersage eingestuft und vernachlässigt. Die vorhergesagte Länge von 438 aa liegt hierbei im gleichen Bereich wie die sequenzhomologen Transposasen. COG Zuordnungen konnten nicht vorgenommen werden.

Tab. 39: BLASTP Resultate der innerhalb der repetitiven Elemente der Gruppe F identifizierten ORFs der Gruppe A

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Nostoc punctiforme	ZP_00105941	hypothetisches Protein	238	53.1 bits (126)	8e-06	25/77 (32%)	42/77 (54%)	-
Shigella flexneri	NP_085177	*IS4* ORF	448	45.1 bits (105)	0.002	89/401 (22%)	151/401 (37%)	35/401 (8%)
Shigella flexneri 2	AAL72389	hypothetisches Protein	447	43.5 bits (101)	0.005	88/401 (21%)	150/401 (36%)	35/401 (8%)
Escherichia coli K12	AAC77234	Transposase insG für IS-Element *IS4*	442	43.1 bits (100)	0.007	87/401 (21%)	151/401 (36%)	35/401 (8%)
Pantoea agglomerans	CAA57483	ORF	440	42.4 bits (98)	0.012	73/324 (22%)	126/324 (38%)	44/324 (13%)

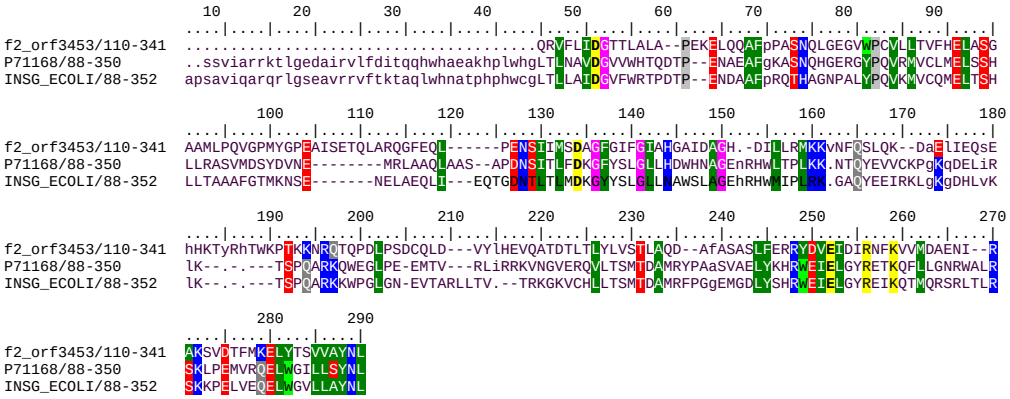
Die ORF-Gruppe B zeigt bei sehr schwachem e-value (4.0) und abweichender Länge (329 aa) Ähnlichkeiten zu einer Cytokinin Synthase von *Arabidopsis thaliana* (BAB59033), die als zufällig eingestuft wird. Weitere Anhaltspunkte für eine derartige Zuordnung finden sich nicht.

Die ORF-Gruppe *A* zeigt Sequenzhomologien zu Pfam- und Interpro-Einträgen, wobei Interpro000719 Eintrag (*eukaryotic protein kinase*), der auf einen Prodom-Eintrag zurückzuführen ist (PD000001), sich unter ProDom nicht nachvollziehen läßt. *OrfA* zeigt unter Prodom eindeutig die beste Zuordnung zu einem plasmidassoziierten *IS4* ORF (PD082068), weshalb die InterPro-Zuordnung als zweifelhaft anzusehen ist.

Die Zuordnung zu einem *IS4* assoziierten ORF wird durch die Identifizierung der Transposase 11 Homologie (Pfam01609) innerhalb der ORFs der Gruppe *A* des Elements F und die charakteristische hochkonservierten DDE Super-Familie (Abb. 34) gestützt.

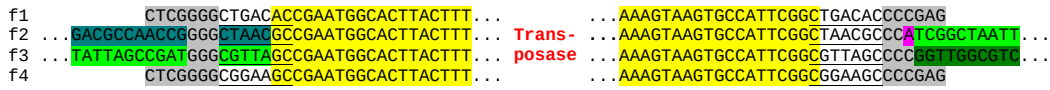
Abb. 34: Auszug aus dem Alignment zwischen dem orf3453 des repetitiven Elements F2 und Transposasen der Pfam 01609 Transposase 11 zur Verdeutlichung des konservierten DDE Motives

Dargestellt werden zwei Transposasen (P71168, *Porphyromonas gingivalis*; *INSG, E. coli* K12), die an der Bildung der Transposase 11 Domäne in Pfam beteiligt sind, versus dem *orf3453* aus dem repetitiven Element F2 stellvertretend für die anderen Element F Varianten. Identische und ähnliche Aminosäuren der konservierten Bereiche sind farblich markiert. Das charakteristische DDE-Motiv wird im Fettdruck und Gelb unterlegt dargestellt. Nur in Gelb hervorgehoben wurde, dass sich innerhalb von sieben Aminosäuren häufig anschließende R-K Motiv der *IS4*-Familie.



Das DDE-Motiv liegt bei allen Transposase 11 Mitgliedern durchgängig vor, während die restliche Sequenz deutlich variiert.Eine Reihe von IS Elementen wie *IS4* und *IS421* werden der Transposase 11 Familie zugeordnet. Eine weitere Zuordnung zu einem beschriebenen IS-Element aufgrund charakteristischer Aminosäuren im Umfeld des DDE Consensus (Mahillon & Chandler 1998) lässt sich durchführen. Die Verlängerung des DDE Consensus (Jenkins et al. 1997) innerhalb der folgenden 7 Aminosäuren um Lysin (K) und Arginin (R) liegt in den ORFs der Gruppe *A* vor und zeigt die Zuordnung zur *IS4*-Familie auf. Abweichungen zu den beschriebenen Strukturen in *IS4* liegen in den Abständen der ORFs der Gruppe *A* innerhalb des DDE Motivs vor. So wird für *IS4* eine Anordnung von DD(94-154)E angegeben (Mahillon & Chandler, 1998) während in der ORF-Gruppe *A* ein Abstand von DD(90)E vorliegt. Diese Abweichung wird jedoch als zulässig im Rahmen der ohnehin großen Schwankungsbreite angesehen. Auch die Struktur eines langen, das gesamte IS-Element auf einem Strang einnehmenden, ORFs (*A*) stimmt mit der *IS4* Zuordnung überein.

Abb. 35: Alignment der Randbereiche des repetitiven Elements F
Farbunterlegungen kennzeichnen in Gelb die invertierten Repeats und die enthaltenen Substitutionen in Lila. Die mögliche Erweiterung der terminalen Repeats liegt Grau unterlegt vor. Die in den terminalen invertierten Repeats enthaltenen direkten Repeats sind unterstrichen hervorgehoben. In Grüntönen werden die sich im Anschluss an die unterschiedlichen repetitiven Elemente F2 und F3 fortsetzenden invertierten Repeats gekennzeichnet. Die hellgrünen invertierten Repeats erreichen eine Länge von 59 bp (Genompositionen: F2 links 2219821-2219763, F3 rechts 2228321-2228252). Die dunkelgrünen invertierten Repeats erreichen eine Länge von 70 bp (Genompositionen: F3 links 2226764-2226706, F2 rechts 2221378-2221319).



Die IS-Elemente des Elements F liegen in den Strukturen ihrer terminalen invertierten Repeats differenziert vor. Element F2 und F3 zeigen untereinander eine invertierte Repeatstruktur mit einer Länge von 59 bzw. 70 bp (Abb. 35, grün unterlegte Bereiche). Die Sequenz in diesem Bereich, wie auch allen anderen Elementen im Genom von *Pirellula*, wurde zusätzlichen Absicherungen unterzogen und ist somit als sicher anzusehen. Über die Funktion der in F2 und F3 auftretenden langen invertierten Repeats kann nur spekuliert werden, möglicherweise agieren F2 und F3 gemeinsam als Transposon bzw. als ein zusammengesetztes IS-Modul (Lewin 2000). Der geringe Abstand von 5431 bp zwischen F2 und F3 würde für eine derartige Hypothese sprechen. Der kodierende Bereich zwischen diesen beiden IS-Elementen zeigt jedoch keine charakteristischen Marker Gene (z.B. Antibiotikaresistenzen wie bei Tn903, Tn10 oder Tn5; Lewin 2000) auf. Auch die deutlich verlängerten invertierten Repeats im Gegensatz zu F1 und F4 müssten als Sondererscheinung gedeutet werden.

Ausgehend von der Hypothese, dass es sich bei F2 und F3 um Sonderfälle handelt, würden die Elemente F1 und F4 die einzelnen IS-Elemente des Transposons repräsentieren. Sie zeigen eine hohe Übereinstimmung in den terminalen invertierten Repeats. Auch in diesen Fällen lässt die Datenlage verschiedene Interpretationen der Bewertung für den Typ der IS-Elemente zu.

1. IS-Element mit Zielstellenverdopplung und Zielstellenpräferenz.

In diesem Fall würden die terminalen invertierten Repeats eine Länge von 19 bp erreichen (Abb. 35, gelb unterlegt), gefolgt von einem 4 bp langen Repeat aus der Zielstellenverdopplung (Abb. 35, unterstrichen). Der folgende Bereich (Abb. 35, grau unterlegt) würde die präferierte Zielstelle kennzeichnen.

2. IS-Element ohne Zielstellenverdopplung und langen terminalen Repeats.

Diese Hypothese würde davon ausgehen, dass die invertierten terminalen Repeats einen

konservierten Fehlpaarungsbereich besitzen und damit eine Gesamtlänge von 27-31 bp erreichen (in Abb. 35 gelb und grau unterlegter Bereich).
Eine Verifikation oder Falsifikation der vorgestellten Hypothesen lässt sich ohne weitere Experimente nicht durchführen.

Abschließend lassen sich die repetitiven Elemente der Gruppe F als *IS4* verwandte IS-Elemente einstufen, die von einem für eine potenzielle Transposase kodierenden ORF dominiert werden und einen weiteren hypothetischen ORF aufweisen.

3.2.2.8 Die repetitiven Elemente der Gruppe G

Die repetitiven Elemente der Gruppe G wurden während der Assemblierung mit einer maximalen Länge von insgesamt 2526 bp bestimmt. Sie liegen insgesamt fünfmal im Genom vor, davon viermal mit geringen Längenunterschieden (Tab. 40). In einem Fall konnte lediglich ein Fragment nachgewiesen werden. Weitere Fragmente konnten im Genom nicht aufgefunden werden.

Tab. 40: Abweichungen der repetitiven Elemente der Gruppe G zum gemeinsamen Consensus

Bedingt durch die größeren Längenunterschiede werden die Anzahl der Gesamtabweichungen und die Anzahl der Abweichungen im gemeinsamen Alignment zum jeweiligen Consensus angegeben. Bei den Gesamtabweichungen werden Abweichungen nur im gegenübergestellten Bereich gezählt; völlig fehlende Bereiche wurden nicht gezählt.

Element	Länge der Varianten in Nukleotiden	Anzahl der Abweichungen zum gesamten Alignment	Abweichungen zum gemeinsamen Alignment (prozentuale Abweichung)
G1	2526	1[*1]	1 (0,1%)
G2	2525	6[*1]	0 (0,0%)
G3	2525	30	11 (0,7%)
G4	2507	2[*2]	0 (0,0%)
G5	1645	11	11 (0,7%)

[*1] davon jeweils eine Deletion [*2] davon eine Insertion

Die BLASTX Analyse am Beispiel des repetitiven Elementes G1 gegen die Datenbank von NCBI zeigt Verwandtschaften zu Genen, denen eine Transposasefunktion zugesprochen wird. Hinzu kamen Gene, die den Integrasen/Rekombinasen zugeordnet werden (Tab. 41).

Tab. 41: BLASTX Resultate am Beispiel von G1

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Mesorhizobium loti	NP_106943	Transposase	397	175 bits (443)	3e-42	125/379 (32%)	184/379 (47%)	26/379 (6%)	+3
Rhizobium sp. NGR234	NP_444037	vermutliche Transposase Y4QJ	398	171 bits (432)	6e-41	121/384 (31%)	186/384 (47%)	29/384 (7%)	+3
Burkholderia fungorum	ZP_00031419	hypothetisches Protein	389	170 bits (430)	1e-40	123/379 (32%)	184/379 (48%)	24/379 (6%)	+3
Bergeyella zoohelcum	AAA50501	Transposase	388	157 bits (397)	7e-37	95/296 (32%)	151/296 (50%)	5/296 (1%)	+3
Rhizobium sp. NGR234	NP_444038	Vermutliche Integrase/ Rekombinase Y4QK	308	125 bits (315),	2e-27	77/267 (28%)	124/267 (45%)	3/267 (1%)	+1
Mesorhizobium loti	NP_085856	Integrase/ Recombinase	299	125 bits (315)	2e-27	82/267 (30%)	122/267 (44%)	3/267 (1%)	+1

Da die Sequenzhomologien in zwei unterschiedlichen Leserahmen und in den Positionen versetzt identifiziert wurden, wurden zwei potenziell kodierende ORFs zumindest antizipiert (ORF-Gruppe *B* und *C*; Tab. 42). Die zugeordnete Funktion einer Integrase, gefolgt von einer Transposase, lassen aufgrund ihrer Struktur ein replikatives Transposon vermuten (Lodish et al. 1996).

Die durch ORPHEUS vorhergesagten ORFs für die Elemente der Gruppe G variieren in ihren Start- und Terminationspositionen erheblich (Tab. 42). Diese Unterschiede sind das Resultat der variierenden Gesamtlängen der Elemente sowie einer Reihe von Sequenzunterschieden.

Tab. 42: ORPHEUS-vorhergesagte ORFs im repetitiven Element G
Die sequenzhomologen ORFs der einzelnen Elemente sind in Gruppen zusammengefasst. ORF-Gruppen, die mit einem „-“ versehen sind, zeigen einen verkürzten ORF auf. Die Verkürzungen entstehen durch Sequenzabweichungen, die zu vorzeitigen Terminationen oder *Frameshifts* führen. ORFs, die nur aufgrund einer Abweichung in der Vorhersage kürzer vorhergesagt wurden, werden in Rot dargestellt und wurden durch „+“ ORFs ersetzt.
Die ORF-Gruppe *A* wurde in G3 vorhergesagt und auf die anderen Elemente übertragen.
Die ORF-Gruppen *A*, *B* und *D* liegen in Element G5 durch die verkürzte Sequenz des Elements nicht oder als Fragmente vor und wurden deshalb nicht berücksichtigt.

Element	ORF-Gruppe	ORF	Position im Genom	Länge in Nukleotide/Aminosäuren
G1	A	*G1_755*	3234941 – 3235063	123/41
	B	*4969*	3235152 – 3235937	786/262
	C	*4970*	3235937 – 3237061	1125/375
	D	*4968*	3235856 – 3235014	843/281
	E	*4971*	3236833 – 3236243	591/197
	F	*4973*	3237400 – 3236912	489/163
G2	A	*G2_755*	3952594 – 3952716	123/41
	B	*5986*	3952805 – 3953590	786/262
	C	*5987*	3953590 – 3954714	1125/375
	D	*5985*	3953509 – 3952667	843/281
	E	*5988*	3954486 – 3953896	591/197
	F	*5990*	3955053 – 3954565	489/163
G3	A	*755*	492148 – 492026	123/41
	B-	*754*	491937 – 491257	681/227
	C	*751+*	491152 - 490028	1125/375
	C	*751*	*491137 – 490028*	*1110/370*
	D-	*753*	491233 – 491757	525/175
	E	*752*	490256 – 490846	591/197
	F	*G3_4973*	489689- 490177	489/163
G4	A	*G4_755*	505889 – 505767	123/41
	B-	*779*	505476 – 504892	585/195
	C	*777+*	504892 - 503768	1125/375
	C	*777*	*504877 – 503768*	*1110/370*
	D-	*780*	504973 – 505728	756/252
	D-	*781*	505304 – 505816	513/171
	E	*778*	503996 – 504586	591/197
G5*	C	*760+*	494203 – 493079	1125/375
	C	*760*	*494188 – 493079*	*1110/370*
	E	*761*	493307 – 493897	591/197
	F	*758*	492740 – 493228	489/163

*liegt als Fragment vor

Alle vorhergesagten ORFs wurden einer BLASTP-Analyse und der Suche nach konservierten Elementen bei NCBI unterzogen. Lediglich zwei ORF-Gruppen zeigen Sequenzhomologien (ORF-Gruppen *B* und *C*, Tab. 42; BLASTP nach ORF-Gruppen, Tab. 43). Es handelt sich hierbei um Ergebnisse, die mit denen der BLASTX-Analyse korrespondieren.

Mit diesen ORFs überlappend vorhergesagte ORFs werden in den weiteren Analysen, ebenso wie ORFs, die aus den Randbereichen in die Elemente ragen, nicht berücksichtigt. Die gemeinsamen Sequenzen der ORFs, denen eine potenzielle Funktion zugeordnet werden konnte, zeigten sich deutlich differenziert. ORF-Gruppe *B* (potenzielle Integrase) zeigt lediglich im Element G4 einen Unterschied zur gemeinsamen Sequenz und im Element G3 zehn Unterschiede. In der abgeleiteten Aminosäuresequenz führen die Unterschiede im Element G4 zu *Frameshift* Ereignissen. In ORF-Gruppe *C* resultieren die Sequenzunterschiede auf Nukleotidebene in acht Abweichungen in Element G3 und neun in Element G5. In der abgeleiteten Aminosäuresequenz ergeben diese Abweichungen lediglich zwei Substitutionen in dem jeweiligen Element jedoch keine *Frameshifts*.

Die überarbeiteten ORFs zeigen nach den Ergänzungen für jeden Strang der vollständigen Elemente drei ORFs auf, von denen vier als hypothetische ORFs charakterisiert wurden (Abb. 39).

Tab. 43: BLASTP Resultate für die ORF-Gruppen *B* (Tab. 43a) und *C* (Tab. 43b) der innerhalb der repetitiven Elemente der Gruppe G identifizierten ORFs

Tab. 43a: BLASTP Resultate der ORF-Gruppe *B* am Beispiel von *orf4969*

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Rhizobium sp. NGR234	NP_444038	vermutliche Integrase/ Rekombinase Y4QK	308	115 bits (288)	5e-25	71/238 (29%)	113/238 (46%)	11/238 (4%)
Mesorhizobium loti	NP_085856	Integrase/ Rekombinase	299	112 bits (281)	3e-24	76/238 (31%)	112/238 (46%)	11/238 (4%)
Bergeyella zoohelcum	AAA50502	Integrase	259	99.8 bits (247)	3e-20	50/182 (27%)	98/182 (53%)	1/182 (0%)
Burkholderia fungorum	ZP_00031418	hypothetisches Protein	347	94.4 bits (233)	1e-18	72/213 (33%)	105/213 (48%)	18/213 (8%)
Pyrococcus abyssi Stamm Orsay	NP_126073	Integrase/ Rekombinase xerd PAB0255	286	89.0 bits (219)	6e-17	61/204 (29%)	109/204 (52%)	12/204 (5%)

Tab. 43b: BLASTP Resultate der ORF-Gruppe *C* am Beispiel von *orf4970*

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Mesorhizobium loti	NP_106943	Transposase	397	172 bits (437)	5e-42	126/379 (33%)	184/379 (48%)	40/379 (10%)
Rhizobium sp. NGR234	NP_444037	vermutliche Transposase Y4QJ	398	167 bits (423)	2e-40	123/395 (31%)	189/395 (47%)	38/395 (9%)
Burkholderia fungorum	ZP_00031419	hypothetisches Protein	389	167 bits (422)	3e-40	129/399 (32%)	192/399 (47%)	40/399 (10%)
Bergeyella zoohelcum	AAA50501	Transposase	388	160 bits (404)	3e-38	98/299 (32%)	156/299 (51%)	20/299 (6%)
Mesorhizobium loti	NP_085855	Transposase	394	154 bits (390)	1e-36	106/342 (30%)	163/342 (46%)	31/342 (9%)

Die Zuordnung der ORF-Gruppe *B* als potenzielle Integrase wird zusätzlich durch die Sequenzähnlichkeiten zum Consensus der Phagen Integrasen Familie (Pfam00589) bestätigt (Abb. 36). Die potenzielle Transposasefunktion der ORF-Gruppe *C* konnte durch Pfam nicht weiter erhärtet werden. Eine weitere Bestätigung liegt jedoch in Form einer ganzen Reihe von Sequenzhomologien zu Transposasen-Domänen in ProDom vor (PD130812, PD130811, PD467276, PD130817 und PD014119).

Abb. 36: Alignment von *orf4969* (ORF-Gruppe *B*) zur Pfam00589
Die angegebenen Positionen beziehen sich auf die ORF-Sequenz und die Consensus-Sequenz des Pfam-Eintrages. 94% der Pfam-Sequenz wird im Alignment gegenübergestellt. Der Pfam-Consensus hat eine Länge von 175 aa.

```
orf4969:     76 WQLIDATVASHLQVIFRAMYSCGLRGVDVRHLRPQDVD--ADRMMLRV-CTTKGHRQREV 132
Pfam00589:   12 ASELARPIGARDRAAVELLLLTGLRISELLSLRWSDIDFDKGTIFIPVRTSGKGRKSRTV  71

orf4969:    133 PLPQATLDAFRAHWATHRNPNWLFPATQRNTPASKADQPISARTIQRGFTKVTESLGWQD 192
Pfam00589:   72 PLSDKAVEALKQYLEIYGRRDLLGGERSDALFPSAVGKPLSRR-LLRRAGKDAGE----- 125

orf4969:    193 SGLTPHTLRHSYATAMLDAGVNLKVLQGYLGHKNLQATEVYLHLTRLGDER  243
Pfam00589:  126 -ELTPHDLRHTFATHLLEAGVDLRVIQKLLGHSSISMTQRYTHVAAEELAE  175
```

Anhand der dargelegten Strukturen wird das Element G von zwei ORFs dominiert, die Sequenzhomologien zu einer Integrase und einer Transposase zeigen. Eine derartige Struktur weist auf ein potenziell replikatives IS-Element hin (Lodish et al. 1996). Ein Beispiel hierfür ist eine in *Bergeyella zoohelcum* identifizierte Transposon ähnlicher Struktur. Wie im repetitiven Element G zeigt ein ORF signifikante Ähnlichkeiten zur Phagen Integrasen Familie, an den sich mit einer Base überlappend ein weiterer für eine potenzielle Transposase kodierender ORF anschließt. Bei *Bergeyella zoohelcum* wurde die Funktion des ersten ORFs (ORF-Gruppe *B* im Element G) als Transposon Resolvase wie in *Tn4430* gedeutet. Terminale invertierte oder direkte Repeats wurden für das Transposon ähnliche Element von *Bergeyella zoohelcum* nicht beschrieben (Brassard et al. 1995). Die Bestimmung von terminalen direkten und indirekten Repeats zeigt keine eindeutigen Strukturen. So zeigen die Varianten des repetitiven Elements G in der Nähe der Enden imperfekte invertierte Repeats (Abb. 37). Da diese Strukturen nicht terminal lokalisiert sind und zudem nicht unmittelbar von direkten Repeats flankiert werden, kann eine eindeutige Zuordnung nicht vorgenommen werden. Der potenzielle invertierte Repeat würde die angenommene Integrase sowie die Transposase in der klassischen Art eines IS-Elements umrahmen, was für eine derartige Hypothese spricht. Unklar bleibt bei dieser Hypothese jedoch, wie es zur Genese des restlichen Bereiches des repetitiven Elements kam.

Abb. 37: Möglicher invertierter Repeat in den Randbereichen des repetitiven Elements G1
Die Sequenz zeigt eine Variante eines invertierten Repeats im Randbereich des Elements G auf. In Gelb werden die im Alignment ohne Fehlpaarungen gegenüberstellbaren Sequenzbereiche unterlegt, auftretende Fehlpaarungen sind in Lila hervorgehoben. Im Endbereich der invertierten Repeats befindet sich ein direkter Repeat, bei dem es sich um eine potenzielle Zielstelle handeln kann.

GCCTCAGCGTAAAAACCATGTTCAAGTGAGCCGCTTTGCGGCGGGG...CCCCACCGCAAAGCGACTCGCCTTGAACAAGGATTTGGGTTCAGGC

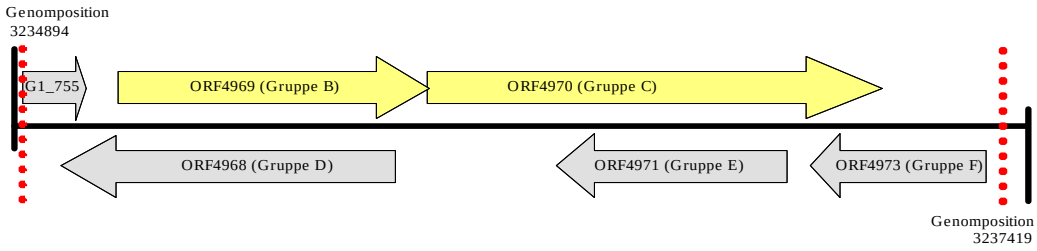
An einem Ende wird der potenzielle imperfekte invertierte Repeat vom restlichen Bereich des Elements über eine potenzielle Loop-Struktur (Abb. 38) getrennt.

Abb. 38:
Potenzielle Loop-Struktur am 3'-Ende des repetitiven Elements G

Die potenzielle Loop-Struktur unterstützt die Termination der ORFs der Gruppe C und verhindert eine Transkription in das Element G in der Gegenrichtung. Eine ähnlich positionierte Hairpin Struktur findet sich im bereits angeführten einem Transposon ähnlichen Element von *Bergeyella zoohelcum*. Die *Hairpin* Struktur von *Bergeyella zoohelcum* befindet sich jedoch innerhalb der terminalen invertierten Repeats.

Abschließend zeigt das Element G eine Reihe von Strukturen, die replikativen Transposons ähneln. Inwiefern diese Strukturen bereits umfangreich überformt vorliegen, kann nicht abschließend beurteilt werden. Hierzu fehlten weitere experimentelle Daten, die über den Mechanismus der Transposition Aufschluss geben könnten. Abweichende Längen der Varianten des repetitiven Elements wurden als Indizien für derartige Überformungen angesehen. Bedingt durch die enge räumliche Lage der Varianten der repetitiven Elemente G2, G3 und G4 können Rekombinationsereignisse zu einer weiteren Überformung beigetragen haben.

Abb. 39: Beispiel für die Anordnung der ORF-Gruppen im repetitiven Element G
Die Abbildung zeigt die Anordnung der identifizierten ORF-Gruppen am Beispiel des Elements G1. ORFs mit Hinweisen auf die Funktion werden in Gelb hervorgehoben, während die hypothetischen ORFs in Grau dargestellt werden. Die ORF-Gruppe *A* wird durch den zusätzlich bestimmten *ORFG1_755* repräsentiert. Die rot gestrichelten Linien geben die Positionen der potenziellen invertierten Repeats an.



3.2.2.9 Die repetitiven Elemente der Gruppe H

Die repetitiven Elemente der Gruppe H treten dreimal im Genom auf. Sie liegen mit einer Länge von 1429 bp (H1 und H2) sowie in einer verkürzten Sequenz mit 1143 bp (H3) vor. Die repetitiven Elemente der Gruppe H zeigen sich in der Nukleotidsequenz (0,3-4,5% Abweichung zur gemeinsamen Consensussequenz im Element) weitgehend konserviert. Element H2 zeigt sich am variabelsten, wobei die größte Variabilität in den nicht kodierenden Sequenzabschnitten gefunden wurde.

Die Gruppe H zeigt im BLASTX (Tab. 45) die größten Ähnlichkeiten zu Genen mit Integrase bzw. Rekombinase Funktion. Die Genfamilie der Integrasen weisen Funktionen wie z.B. der Dekatenation oder Segregation von gerade replizierten Chromosomen, konjugative Transposition, Regulation der Plasmidkopienzahl oder die Expression von Proteinen an der Zelloberfläche aus. Zu den bekanntesten Aufgaben zählt die Integration und Exzession von viralen Genomen in das Wirtgenom wie z.B. bei der λ Integrase. Charakteristisch für die Integrasen ist die Fähigkeit, eine sequenzspezifische Rekombinationsreaktion zwischen zwei DNA-Abschnitten ohne zusätzlich Energiezufuhr durch Cofaktoren durchzuführen. Integrasen schneiden ihr DNA-Substrat mit einer Serie von gestaffelten Schnitten und rekombinieren im Anschluss einen DNA-Abschnitt in die Ziel-DNA. Es wird deshalb von einer Integrase/Rekombinase Funktion gesprochen. Schneiden der DNA und Wiederverbinden des rekombinanten Produktes geschehen in zwei Schritten. Im ersten Schritt attackiert ein Tyrosin Hydroxyl eine bestimmte Phosphatgruppe, wodurch es zum Nicken der DNA kommt. Gleichzeitig wird ein 3’ Phosphotyrosine gebundener DNA Komplex gebildet. Dieses Protein-DNA Zwischenprodukt wird aufgelöst, wenn die 5’-terminale Hydroxyl-Gruppe des „angreifenden“ DNA-Stranges an die phosphotyrosine Bindung an-

gelegt wird, das Protein wird verdrängt und die Holiday Verbindung erzeugt. Durch Wiederholung des Vorganges mit dem zweiten DNA Strang wird die Rekombination abgeschlossen, ein neuer DNA Abschnitt wurde integriert (Ellenberger et al. 1997; Gopaul et al. 1997).

Vorhersagen mit ORPHEUS sagen für jedes der Elemente einen ORF im Leserahmen der BLASTX Resultate voraus. Weitere ORFs, die konserviert in den ORFs vorliegen oder zu denen sich Aussagen zu deren Funktion treffen lassen, liegen nicht vor. Die ORFs liegen in unterschiedlicher Länge vor. Ein einheitliches Startcodon lässt sich nicht eindeutig bestimmen.

Tab. 44: ORFs im repetitiven Element H nach ORPHEUS-Vorhersage mit Ähnlichkeiten zu Integrasen

Element	orf	Position im Genom	Länge in Nukleotiden/ Aminosäuren
H1	*orf4976*	3'239'090-3'238'050	1161/387
H2	*orf5998*	3'958'684-3'957'302	1383/461
H3	*orf745*	487'594-488'469	876/292

Die Sequenzen der repetitiven Elemente stellen sich in der Nukleotidsequenz konserviert dar, was sich in höherem Maße in der abgeleiteten Aminosäuresequenz widerspiegelt (Abb.40).

Abb. 40: Ausgewählte ORFs des repetitiven Elements H im Alignment
Abweichungen im Alignment werden in Blau dargestellt. Es handelt sich hierbei um konservierte Austausche. Relevante Start- und Stoppcodons sind im Fettdruck dargestellt.

```
orfH1 LRNVSSPYLLEQLMSRLVAPDAPRLYTNGQPWRDARGRGAPRHAVETEFKAEHAVVDLSTQPPKPLVVKCLVRELRIRFYAVSTIKNYRSAWVCFFRWYR
orfH2 ----------------------------------------------------------------------------LRIRFYAVSTIKNYRSAWVCFLRWYR
orfH3 ----------------------------------------------------------------------------------------------------

orfH1 GPLDQIDQEDIREYLELLVNGGASASEVSVTLSALRTGLDKFCLLRCTVGLVSPRKSKQLPVVMSKKEVQRMMEAARTLRDKLLLTVLYATGLRVAEVAR
orfH2 GPLDQIDQEDIREYLELLVAGGASASEVSVTLSALRTGLDKFCLLRCTVGLVSPRKSKQLPVVMSKKEVQRMMEAARTLRDKLLLTVLYATGLRVAEVAR
orfH3 ------------------------------------------------------------------LRMMEAARTLRDKLLLTVLYATGLRVAEVAR

orfH1 LQWSDFDFDRQQIRVQLGKGKKDRYVMLADDLLPLMRQLWRHTKGVGYLFPSEGRRVDRHLSPRTIQRAVKQARILSGIGKAVTPHSFRHSFATHLIESG
orfH2 LQWSDFDFDRQQIRVQLGKGKKDRYVMLADDLLPLMRQLWRHTKGVGYLFPSEGRRVDRHLSPRTIQRAVKQARILSGIGKAVTPHSFRHSFATHLIESG
orfH3 LQWSDFDFDRQQIRVQLGKGKKDRYVMLADDLLPLMRQLWRHTKGVGYLFPSEGRRVDRHLSPRTIQRAVKQARILSGIGKAVTPHSFRHSFATHLIESG

orfH1 TDIRFIQKLLGHTNLETTSLYTKVARMKATAVASPLDQLRDEPGSSSESSGRQPKPRPSVGRMRLEVDPNPDSNGAYAVTLGVWKDGQLLPLPGMRA
orfH2 TDIRFIQKLLGHTNLETTSLYTKVARMKATAVASPLDRLRDEPGSSSESSGRQPKPRPSVGRMRLEVDPNPDSNGAYAVTLGVWKDGQLLPLPGMRA
orfH3 TDIRFIQKLLGHTNLETTSLYTKVARMKATAVASPLDQLRDEPGSSSESSGRQPKPRPSVGRMRLEVDPNPDSNGAYAVTLGVWKDGQLLPLPGMRA

orfH1 TMPRQDWVSLQIPLQDSWEPTLRCLPTAQRERLESPEFFSQVQREVAKQILRIRDAEPSQAIKT
orfH2 TMPRQDWVSLQIPLQDIWEPTLRCLPTAQRERLESPEFFSQLQREVAKRILRIRDAEPFQALKT
orfH3 TMPRQDWVSLQIPLQDSWEPTLRCLPTAQRERLESPEFFSQVQREVAKQILRIRDAEPSQAIKT
```

Die als kodierend vorhergesagten Sequenzen weisen im BLASTP (Tab. 45) die zu erwartenden Homologien zu Genen aus der Integrase Familie aus dem BLASTX auf, wobei

die Längen der zum Vergleich herangezogenen ORFs im Element H1 überschritten und bei H2 sowie H3 unterschritten werden. Im Anhang sind die Sequenzen sowie die analysierten ORFs abgelegt.

Die ORFs zeigen eine klare Zuordnung zum COGs Eintrag COG0582, der den Integrasen zugeordnet wird. Der Vergleich mit der Interpro-Datenbank ermöglicht eine Zuordnung zur Familie der Phagen Integrasen. Diese Zuordnung erklärt das Auftreten mehrerer Kopien in der Genomsequenz von *Pirellula* sp. Stamm 1 und lässt Rückschlüsse auf den externen Urprung der Sequenz zu. Die Consensussequenz des Pfam-Eintrags ist das Resultat der Inhomogenität der Integrasen in ihrer Sequenz, die vermutlich die Diversität in der Form und Funktion widerspiegelt, in der sie ihre genetischen Umordnungen durchführen. Die katalytische Domäne der Integrasen stellt sich im Gegensatz zum Rest der Sequenz typisch hochkonserviert (Arg-His-Arg Motiv) dar.

Tab. 45: Zusammenfassung der Funktionshinweise der ORFs im Element

BLASTX bester Treffer gegen die NCBI Protein Datenbank der Consensus-Sequenz des repetitiven Elements H	NC_002678 Integrase/ Rekombinase *Mesorhizobium loti* Length = 302 Expect = 3e-40 Identities = 105/241 (43%) Positives = 143/241 (58%) Gaps = 10/241 (4%)

ORF	**BLASTP gegen die NCBI Protein Datenbank**	**Interpro/Pfam**	**COGs**
H1 *orf4976* 387 aa	NCBI: NP_444038 vermutliche Integrase/ Rekombinase Y4QK *Rhizobium* sp. NGR234 Length = 308 e-value = 2e-38 Identities = 100/269 (37%) Positives = 154/269 (57%) Gaps = 7/269 (2%)	Phagen Integrase pfam00589.4 e-value: 4.4e-42 **(**Pfam 7.3**)** pfam02899, Phage_integr_N, Phagen-Integrase, N-terminal SAM-ähnliche Domäne e-value = 0.001	COG0582 Integrase
H2 orf5998 461 aa	NCBI: AAA50502 Integrase *Bergeyella zoohelcum* Length = 259 e-value = 5e-38 Identities = 98/247(39%) Positives = 142/247 (56%) Gaps = 7/247 (2%)	Phagen Integrase pfam00589.4 e-value: 4.4e-42 **(**Pfam 7.3**)** pfam02899, Phage_integr_N, Phagen-Integrase, N-terminal SAM-ähnliche Domäne e-value = 0.001	COG0582 Integrase
H3 orf745 292 aa	NCBI: AAA50502 integrase *Bergeyella zoohelcum* Length = 259 e-value = 5e-31 Identities = 75/179 (41%), Positives = 108/179 (59%), Gaps = 5/179 (2%)	Phagen Integrase pfam00589.4 e-value: 4.4e-42 **(**Pfam 7.3**)**	COG0582 Integrase

Die konservierten Bereiche stellen sich in der Pfam Consensussequenz (pfam00589.4) als Arg-173, His-289, Arg-292 sowie für das attackieren Tyr-324 dar (Abb. 41).

Abb. 41: Beispiel des erweiterten ORFs im repetitiven Element H im Alignment mit der Consensus Sequenz der Phagen Integrasen Familie (pfam00589.4)
Das Alignment zeigt die Übereinstimmungen der konservierten Aminosäuren in Rot, die hochkonservierten Aminosäuren im Reaktionszentrum im Rotfettdruck sowie die konservierten Austausche in Blau.

```
H1 erw. ORF: MSKKEVQRMMEAARTL-----RDKLLLTVLYATGLRVAEVARLQWSDFDFDRQQIRVQ--
H2 erw. ORF: MSKKEVQRMMEAARTL-----RDKLLLTVLYATGLRVAEVARLQWSDFDFDRQQIRVQ--
H3 erw. ORF:       LRMMEAARTL-----RDKLLLTVLYATGLRVAEVARLQWSDFDFDRQQIRVQ—
pfam00589.4: LTEDQIEKLLAASELARPIGARDRAAVELLLLTGLRISELLSLRWSDIDLEKGTITIPVR

H1 erw. ORF: -LGKGKKDRYVMLADDLLPLMRQLWRHTKGVGYLFPSEGRRVDRHLSPRTIQRAVKQARI
H2 erw. ORF: --GKGKKDRYVMLADDLLPLMRQLWRHTKGVGYLFPSEGRRVDRHLSPRTIQRAVKQARI
H3 erw. ORF: -LGKGKKDRYVMLADDLLPLMRQLWRHTKGVGYLFPSEGRRVDRHLSPRTIQRAVKQARI
pfam00589.4: TSGKGRKERTVPLSDKAVEALKQYLEIYGRRDLLGGERSHDALFPSAVGKPLSRRLLRRA

H1 erw. ORF: LSGIGKAVTPHSFRHSFATHLIESGTDIRFIQKLLGHTNLETTSLYTKVAR
H2 erw. ORF: LSGIGKAVTPHSFRHSFATHLIESGTDIRFIQKLLGHTNLETTSLYTKVAR
H2 erw. ORF: LSGIGKAVTPHSFRHSFATHLIESGTDIRFIQKLLGHTNLETTSLYTKVAR
pfam00589.4: GKDAGEELTPHDLRHTFATHLLEAGVDLRVIQKLLGHSNISMTQRYTHVAA
```

Die Zuordnung der potenziell kodierenden Sequenz im Element zu den Phagen Integrasen wird durch das vollständige Treffen der Interpro Familie unterstützt. Phagen Integrasen besitzen neben der katalytischen Domäne fast immer noch einen weiteren schwächer konservierten Bereich im N-Terminus der Sequenz. Es ist nicht auszuschließen, dass hier eine neue Variante der Integrasenfamilie vorliegt, der dieser Bereich oder ein Äquivalent fehlt, was aber von allen über 60 bisher beschriebenen Integrasen grundsätzlich abweichen würde. Dieser Bereich wird als Phagen Integrase N-terminal SAM-like Domäne in der Pfam02899 geführt und ist dem Phagen Integrasen Familienmotiv in der Sequenz fast immer vorgelagert. Ähnlichkeiten zu dieser Domäne lassen sich nur in den längeren Leserahmen der Elemente H1 und H2 identifizieren (Beschreibung Interpro IPR004107). Auch die im BLASTP identifizierte ähnlichste Sequenz von Y4qK (*Rhizobium* sp. NGR234; Acc. NP_444038) beinhaltet erwartungsgemäß beide Domänen (Abb. 42).

Abb. 42: Repetitive Elemente im Alignment mit der Consensus Sequenz der Pfam02899, Phage_integr_N, Phagen Integrase, N-terminale SAM-ähnliche Domäne
Konservierte Aminosäuren werden in Rot dargestellt und die konservierten Austausche in Blau.

```
H1 erw. ORF: LRI-RFYAVSTIKNYRSAWVCFFRWYR----GPLDQIDQEDIREYLELLVNGGASASEV
H2 erw. ORF: LRI-RFYAVSTIKNYRSAWVCFLRWYR----GPLDQIDQEDIREYLELLVAGGASASEV
pfam02899:   LRVERGLSPHTVRAYRRDLKAFLRFLAERGGLSWDQLTAEDVRAFLAELLAKGLSAASL

H1 erw.ORF:  SVTLSALR
H2 erw.ORF:  SVTLSALR
pfam02899:   ARRLSALR
```

ORFs der repetitiven Elemente H1 und H2 zeigen Ähnlichkeiten zur N-terminalen SAM-like Domäne. Hypothetisch könnte es sich hiernach bei H1 und H2 um vollständig erhaltene Integrasengene handeln. Zumindest der ORF der Variante H3 wäre demnach evolutiv überformt und lediglich ein Relikt eines Insertionsereignisses oder möglicherweise ein Duplikat, welches als Pseudogen (Lodisch et al. 1996) vorliegt. Weitere Sequenzähnlichkeiten in den unterschiedlichen Leserahmen in 5'-Richtung aufwärts vom H3 ORFs, die diese Hypothese unterstützen und womöglich fragmentiert vorliegen, konnten weder durch Blast2Seq, ClustalW oder BLASTP aufgefunden werden. Als Basis für diese Analyse wurden die übersetzten ORFs von H1 und H2 gewählt. Es kann letztendlich nicht ausgeschlossen werden, dass die schwachen Ähnlichkeiten zur zweiten Domäne zufälliger Natur sind und es sich um einen neuen Integrasentyp handelt. In diesem Fall könnten die ORFs deutlich kürzer ausfallen und noch immer das typische katalytische Zentrum beinhalten. Diese Möglichkeit wird im Nukleotidalignment ausgewiesen jedoch nicht favorisiert, da sie die Hinweise auf eine weitere Domäne und die BLASTP-Alignments vernachlässigen würde.

Alle Varianten des repetitiven Elements weichen in der Länge von den im BLASTP ausgewiesenen ORFs in 3'-Richtung nach dem katalytischen Zentrum ab. Diese Abweichung von zusätzlichen 80 aa kann in den speziellen Eigenschaften dieser Integrase begründet sein. Da das Stoppcodon in allen Kopien einheitlich ist, erscheint eine Sequenzveränderung, die zu einem Verlust eines früher auftretenden Stoppcodons führte, als unwahrscheinlich. Der unterschiedliche Start in den ORFs der repetitiven Elemente H1 und H2 lässt sich nicht eindeutig auflösen. Zwei Hypothesen bieten sich hier an. Durch Verkürzung des ORFs im Element H2 kann ein gemeinsames Startcodon gefunden werden. Das erste gemeinsame Startcodon würde in der Nukleotidsequenz uneinheitlich sein und die Längen der anderen Integrasen im Bereich der SAM Domäne unterschreiten. Es besteht auch die Möglichkeit, dass lediglich der ORF im Element H2 vollständig ist. Dieser ORF würde ca. 20 aa über die ORF-Längen bisher beschriebener Integrasen im Bereich der SAM-like Domäne hinausgehen. Der ORF in H1 wäre demnach ebenfalls nur ein Fragment des ursprünglichen ORFs. Gegen diese Auffassung spricht die konservierte Nukleotidsequenz.

Weitere Kopien oder Fragmente der analysierten ORFs konnten in der gesamten Genomsequenz mit TBLASTN nicht aufgefunden werden. Hinweise auf einen Transfer von Fremd-DNA in der Umgebung der Integrase konnten nicht gefunden werden.

3.2.2.10 Die repetitiven Elemente der Gruppe I

Das repetitive Element I wurde während der Assemblierung mit einer Länge von insgesamt 1375 bp bestimmt. Das repetitive Element I liegt insgesamt fünfmal im Genom vor. Weitere Fragmente konnten im Genom nicht aufgefunden werden. Die Varianten des repetitiven Elements I weichen in nur geringem Maße voneinander ab. Die Variante I4 zeigt sich mit sechs Substitutionen zum gemeinsamen Consensus mit 1372 bp am variabelsten (Tab. 46).

Tab. 46: Abweichungen der Varianten des repetitiven Elements I zum gemeinsamen Consensus
Als Basis wurde der hochkonservierte 1372 bp lange Sequenzbereich gewählt.

Variante	Anzahl der Abweichungen	Prozentuale Abweichung
I1	2	0,2
I2	3	0,2
I3	0	0,0
I4	6	0,4
I5	2	0,2

Die BLASTX Analyse des repetitiven Elements I gegen die Datenbank von NCBI weist auf Verwandtschaften zu Genen, denen eine Transposasefunktion zugesprochen wird, hin (Tab. 47).

Tab. 47: BLASTX Resultat am Beispiel von Element I3
I3 zeigt keine Abweichungen zum gemeinsamen Consensus. Die Zuordnung zeigt lediglich einen signifikanten Hit zu potenziellen Transposasen in *Nostoc* sp. PCC 7120, von denen der beste Treffer dargestellt wird. Diverse Zuordnungen zu dieser nicht charakterisierten Transposasen Gruppe existieren bei diesem Organismus. Leider sind die ORF-Vorhersagen von *Nostoc* sp. noch nicht weitergehend experimentell überprüft worden, so dass die ORF-Längen zwischen 410 und 316 aa schwanken.

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Nostoc sp. PCC 7120	BAB77349	Transposase	410	177 bits (448)	3e-50	99/272 (36%)	150/272 (54%)	2/272 (0%)	-2

Das Programm ORPHEUS sagt jeweils zwei ORFs für die jeweiligen repetitiven Elemente der Gruppe I voraus (Tab. 48). Weitere ORFs ohne eine mögliche Funktionszuweisung ragen in diversen Varianten in die repetitiven Elemente der Gruppe I hinein, die als fragwürdig eingestuft werden. Die Ursachen für ihre Vorhersage liegen in den außerhalb der Elemente liegenden Sequenzrandbereichen, in denen mögliche ORFs beginnen, und in der fehlenden Berücksichtigung der vorliegenden Struktur.

Die zwei in jedem der repetitiven Elemente der Gruppe I vorhandenen ORFs zeigen zum Teil unterschiedlich vorhergesagte Startcodons. Zu berücksichtigen ist in diesem Zusammenhang, dass durch die Verwendung eines anderen Startcodons sich eine Verlängerung einer Gruppe von ORFs ergibt. Diese ORFs wurden um die potenziell kodierende Sequenz auf die maximale Länge bis zum gemeinsamen Start (ATG) verlängert (Kennzeichnung mit +). Für jede Variante des Elements I wurden so zwei ORFs bestimmt, die vermutlich in Zusammenhang mit der Transposasefunktion stehen (Tab. 50, Abb. 45).

Tab. 48: Mit ORPHEUS vorhergesagte ORFs im repetitiven Element I
In den Elementen einander entsprechende ORFs wurden nur jeweils einmal pro Element gefunden. Farblich markiert werden die ORFs dargestellt, die potenziell im Zusammenhang mit der Transposasefunktion stehen. Alle Elemente der Gruppe I beinhalten einen kurzen ORF (blaues Schriftbild; 109 aa) sowie einen längeren (grünes Schriftbild; 309 aa). Die manuell bis zum gemeinsamen Startcodon verlängerte ORFs tragen das Schriftzeichen „+“ im Namen. Abweichungen in der ORF-Vorhersage resultieren aus Differenzen der Sequenzen der Elemente zueinander und der Einbeziehung der ORFs, die in das Element hinein- oder herausreichen (braunes Schriftbild).

Variante	ORF	Position im Genom	Länge in Nukleotiden/Aminosäuren
I1	*orf3478*	2234837-2233920	918/306
	orf3478+	2234846-2233920	927/309
	orf3479	2234849-2235175	327/109
	orf3480	2235597-2235166	432/144
I2	*orf10438*	6925424-6925642	219/73
	orf10439	6926544-6925627	918/306
	orf10439+	6926553-6925627	927/309
	orf10440	6926556-6926882	327/109
	orf10441	6926995-6926873	123/41
I3	*orf4314*	2773896-2773639	258/86
	orf4315	2774817-2773900	918/306
	orf4315+	2774826-2773900	927/309
	orf4316	2774829-2775155	327/109
I4	***orf9031***	6022597-6022271	327/109
	orf9032	6022618-6023526	909/303
	orf9032+	6022600-6023526	927/309
I5	***orf7486***	4937008-4936682	327/109
	orf7487	4937029-4937937	909/303
	orf7487+	4937011-4937937	927/309

Da die im Zusammenhang mit der Transposase kodierenden ORFs externen Ursprungs sind, ist die Vorhersage mit ORPHEUS unter Verwendung eines auf *Pirellula* trainierten Sets stets problematisch. Erschwerend kommt hinzu, dass eine konservierte Shine-Dalgarno-Sequenz nicht identifiziert werden konnte. Dieses Phänomen ist nicht ungewöhn-

lich und ist im Zusammenhang mit den Auswirkungen der Transposition zu sehen. Eine Translation findet trotzdem statt, wenn auch mit geringerer Effizienz (Sato et al. 1989).
Die beiden identifizierten ORF-Gruppen weichen nur in geringem Maße voneinander ab. So zeigen die kurzen gemeinsamen ORFs (*orfA*; blaues Schriftbild, Tab. 48) maximal eine Abweichung zueinander und die längeren gemeinsamen ORFs maximal fünf Abweichungen (Tab. 49). Durch die vorgenommene Verlängerung von *orfB* entstehen keine zusätzlichen Abweichungen.

Tab. 49: Abweichungen der ORFs im Element I zueinander

Variante	*orfA* (na/aa)	Abweichungen zum gemeinsamen Consensus (na/aa)	*orfB* (na/aa)	Abweichungen zum gemeinsamen Consensus (na/aa)
I1	327/109	1/1*	927/309	1/1
I2	327/109	1/0	927/309	2/1
I3	327/109	0/0	927/309	0/0
I4	327/109	0/0	927/309	5/2*
I5	327/109	1/0	927/309	0/0

* konservierte Austausche

Nur jeweils einem ORF pro Variante kann eine potenzielle Funktion zugeordnet werden. Diese ORFs der ORF-Gruppe *B* spiegeln zudem die Transposasenzuweisung sowie die geringen Vergleichsmöglichkeiten aus den BLASTX-Suchen im BLASTP wider (Tab. 50).

Tab. 50: BLASTP Resultate der innerhalb des repetitiven Elements I identifizierten ORFs am Beispiel des *orf4315+*

Organismus	**Acc. Number**	**Genfunktion**	**Länge (aa)**	**Score**	**e-value**	**Identities**	**Positives**	**Gaps**
Nostoc sp. PCC 7120	NP_478444	Transposase, unbekanntes Protein	410	174 bits (440)	1e-42	98/271 (36%)	149/271 (54%)	3/271 (1%)
Nostoc sp. PCC 7120	NP_486185	Transposase	316	129 bits (324)	4e-29	72/204 (35%)	113/204 (55%)	3/204 (1%)
Nostoc sp. PCC 7120	NP_478209	Transposase, unbekanntes Protein	405	79.7 bits (195)	4e-14	69/301 (22%)	142/301 (46%)	31/301 (10%)
Streptomyces coelicolor A3(2)	NP_628515	vermutliche Transposase	281	38.9 bits (89)	0.073	23/80 (28%)	42/80 (51%)	4/80 (5%)
Streptomyces coelicolor A3(2)	NP_639736	vermutliche Transposase	374	38.1 bits (87)	0.12	23/80 (28%)	41/80 (50%)	4/80 (5%)

Weitere Anhaltspunkte, die Rückschlüsse auf die Funktion der Gene zulassen, finden sich ebenfalls nur für *orfB* in ProDom (Abb. 43), wobei die hier vorgenommene Transposasen-Zuordnung auch auf dem BLASTP Ergebnis basiert.

Abb. 43: ProDom Alignment zwischen dem *orf4315+* und ProDom Domäne PD399532
Dargestellt werden zwei Transposasen aus *Streptomyces coelicolor* (STRCO) der ProDom Domäne im Alignment zu *orf4315+* aus der Variante I3 stellvertretend für die anderen I Varianten.

Die schwachen Sequenzhomologien weisen im Zusammenhang mit den Randbereichen des Elements I auf ein IS-Element hin. In diesen Bereichen konnten charakteristische invertierte Repeats eines IS-Elements identifiziert werden, die eine Länge von 33-37 bp besitzen (Abb. 44).

Abb. 44: Flankierende invertierte Repeats im repetitiven Element I
Die einzelnen Varianten werden durch imperfekte terminale invertierte Repeats flankiert. Fehlerhafte Paarungen in den invertierten Repeats wurden grau unterlegt. Hierbei fällt eine zwei Basen lange konservierte Fehlpaarungsstelle auf. Der in allen Varianten konservierte invertierte Repeatbereich wird im Fettdruck hervorgehoben. Flankierende direkte Repeats, die direkt an die gelb unterlegten invertierten Repeats anschließen, konnten nicht identifiziert werden.

Die flankierenden direkten Repeats konnten nicht identifiziert werden. Über die Ursachen kann folgendermaßen spekuliert werden. Ein völliges Fehlen der direkten Repeats ist durchaus möglich (Mahillon & Chandler 1998) und würde mit den Beobachtungen in *IS605* sowie *IS606* übereinstimmen, bei denen jedoch auch die invertierten Repeats fehlen (Kersulyte et al. 1998). Auch ein leicht veränderter Mechanismus der Transposition könnte zu einer leicht veränderten Positionierung der direkten Repeats führen, die z.B. bei einer Länge von 2 bp nicht mehr eindeutig alleine anhand der Sequenz zu identifizieren ist. Das weitreichende Auftreten von Rekombinationsereignissen sowie die Überformung durch weitere IS-Elemente (Mahillon & Chandler 1998) kann nicht als Erklärungsmodell für alle Varianten herangezogen werden.

In Übereinstimmung mit den einzelnen Varianten liegen innerhalb dieses IS-Elements zwei ORFs vor, die potenziell für die Transposition kodieren. Das Auftreten von zwei ORFs im IS-Element ist bei einer ganzen Reihe von IS-Elementen der Fall (Mahillon & Chandler 1998). Eine seltene Besonderheit liegt in der Orientierung der ORFs zueinander vor. *OrfA* und *orfB* liegen in einander entgegengerichteter Orientierung auf dem gleichen Strang (Abb. 45). Eine derartige Struktur findet sich zum Beispiel bei *IS605* und *IS606*

(Kersulyte et al. 1998; Mahillon & Chandler 1998). Ergänzend ist anzumerken, dass bei dem teilweise sequenzhomologen *IS607*, bei dem im Gegensatz zu *IS605* und *IS606* beide ORFs in gleicher Orientierung vorliegen, hier nur ein ORF (der kürzere ORF) zur Transposition benötigt wird. Hiermit ist das ursprüngliche Modell für *IS605* und *IS606*, das von einem heterodimeren Protein ausging, in Frage gestellt (Kersulyte et al. 2000).

Abb. 45: Schematische Darstellung des IS-Elements im repetitiven Element I
Die schematische Darstellung basiert auf den Elementen in der repetitiven Gruppe I. Die identifizierten Elemente wie die flankierenden invertierten Repeats (IR) und die potenziell für die Tranposation kodierenden ORFs spiegeln die Strukturen eines IS-Elements wider. *OrfB* zeigte in der BLASTP-Analyse Ähnlichkeiten zu Transposasen. Die Pfeile geben die Transkriptionsrichtung an.

Die beiden innerhalb der repetitiven Elemente der Gruppe I bestimmten ORFs füllen den größten Bereich des potenziellen IS-Elements aus. Die repetitiven Elemente der Gruppe I stellen eine zurzeit noch nicht charakterisierte Familie von IS-Elementen dar. Invertierte Repeats und ORF-Organisation im Inneren spiegeln bekannte Strukturen wider. Das Fehlen von direkten Repeats lässt sich anhand der zur Verfügung stehenden Daten nicht erklären.

3.2.2.11 Die repetitiven Elemente der Gruppe J

Die repetitiven Elemente der Gruppe J wurde mit einer Länge von insgesamt 2658 bp bestimmt. Es stellt sich mit seinen sechs Varianten im Genom unterschiedlich konserviert dar (Tab. 51). Abweichungen in den Sequenzlängen entstehen durch die Insertion eines weiteren IS-Elements des repetitiven Elements B3 (vgl. Kap. 3.2.2.3), woraus die Varianten J1 und J2 entstehen. Am 3'-Ende von J1 bzw. am 5'-Ende von J2 findet sich die Zielstellenverdopplung (CAT) des IS-Elements des Elements B3. Für die folgenden Analysen wurden die Elemente J1 und J2 wieder zusammengeführt, um sie als vollständiges Element analysieren zu können.

Element J6 liegt mit lediglich 668 bp vor. Aufgrund der geringen Identität zu den anderen repetitiven Elementen der Gruppe J kann J6 bestenfalls als Relikt angesehen werden.

Möglicherweise hängt die Genese von Element J6 nicht mit denen der anderen Elemente dieser Gruppe zusammen. Das Element J6 wird deshalb in den folgenden Analysen nicht weiter berücksichtigt.

Weitere Fragmente konnten im Genom nicht aufgefunden werden.

Tab. 51: Abweichungen der repetitiven Elemente der Gruppe J zum gemeinsamen Consensus
Als Basis wurde der 2658 bp lange Sequenzbereich gewählt. Die Abweichungen im Element J6 beziehen sich auf die 668 bp lange Sequenz.

Variante	Anzahl der Abweichungen	Prozentuale Abweichung
J1/J2	26	0,98%
J3	120	4,52%
J4	6	0,23%
J5	2	0,08%
J6*	193	7,26%

Die BLASTX Analyse der Sequenz des Elements J5 gegen die Datenbank von NCBI zeigt Sequenzhomologien zu Genen mit Integrase/Rekombinase sowie Transposasefunktion (Tab. 52). Somit finden sich Hinweise auf zwei potenzielle Gene im Element J.

Tab. 52: BLASTX Resultate am Beispiel der Variante J5

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Rhizobium sp. NGR234	NP_444038	Integrase/ Rekombinase Y4QK	308	172 bits (435)	3e-41	94/271 (34%)	150/271 (54%)	1/271 (0%)	-2
Mesorhizobium loti	NP_085856	Integrase/ Rekombinase	299	156 bits (394)	2e-36	95/299 (31%)	153/299 (50%)	1/299 (0%)	-2
Burkholderia fungorum	ZP_00031418	hypothetisches Protein	347	155 bits (392)	3e-36	101/311 (32%)	164/311 (52%)	10/311 (3%)	-2
Bergeyella zoohelcum	AAA50501	Transposase	388	154 bits (389)	6e-36	90/284 (31%)	152/284 (52%)	5/284 (1%)	-1
Mesorhizobium loti	NP_106943	Transposase	397	150 bits (378)	1e-34	124/390 (31%)	174/390 (43%)	18/390 (4%)	-1

Vier bis fünf ORFs werden durch das ORF-Vorhersageprogamm ORPHEUS vorhergesagt. Abweichungen in der Vorhersage resultieren aus unterschiedlich vorhergesagten Startcodons, Abweichungen in der Sequenz und der Länge der repetitiven Elemente. Hinzu kommt die durch die Insertion durchbrochene Variante der Elemente J1 und J2. Die vorhergesagten ORFs lassen sich in fünf Gruppen zusammenstellen (Tab. 53).

Tab. 53: ORPHEUS-Vorhersagen für die repetitiven Elemente der Gruppe J
Dargestellt werden die mit ORPHEUS vorhergesagten ORFs des Elements J. In den vollständigen Varianten J3, J4 und J5, werden fünf ORFs vorhergesagt, die einander entsprechen und gruppiert wurden (ORF-Gruppen *A*-*E*). Die ORF-Gruppe *C* wurde im wieder zusammengesetzten repetitiven Element J1/J2 nicht vorhergesagt. Sie fallen weiterhin durch unterschiedliche Positionen der Stoppcodons auf. Zur Berücksichtigung der auftretenden Verkürzung des ORFs im Element J3 (-12b) wurden Abweichungen nur für den Teilbereich des gemeinsamen Alignments bestimmt. In ORF-Gruppe *D* weichen drei von vier ORFs in der Position ihres Startcodons voneinander ab, die auftretenden Abweichungen wurden nur für den gemeinsamen Teil des Alignments 582 Nukleotiden bzw. 195 Aminosäuren bestimmt. ORF-Gruppe *E* wurde im repetitiven Element J3 nicht vorhergesagt. Die Ursachen liegen in Sequenzabweichungen zum gemeinsamen Alignment, die unter anderem in einem deutlich reduzierten offenen Leserahmen (-106 b) resultieren.

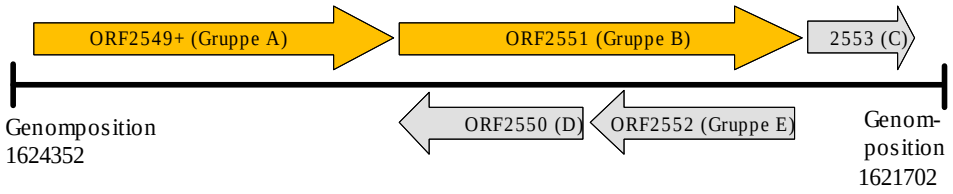
Variante	ORF-Gruppe	ORF	Position im Genom	ORF-Länge (na/aa)	Abweichungen zum gemeinsamen Consensus (na/aa)
J2	*A*	*953*	609496-608453	1044/348	26/11
J1	*B*	*948**	606742-605978	765/255	0/0
	C	-	-	-	-
J2	*D*	*952*	608109-608534	426/142	0/0
J1	*E*	*949*	606002-606586	585/195	0/0
J3	*A*	*3221-*	2064477-2063434	1044/348	7/2
	A	*3221*	*2064741-2063434*	*1308/436*	-
	B	*3219*	2063416-2062271	1146/382	76/15
	C	*3218*	2062264- 2061983	282/94	37/39
	D	*3220*	2062934-2063515	582/194	7/7
	E	-	-	-	-
J4	*A*	*2079+*	1335881-1336924	1044/348	3/2
	A	*2079*	*1335884-1336924*	*1041/347*	-
	B	*2081*	1336942-1338087	1146/382	0/0
	C	*2083*	1338094- 1338387	294/98	0/0
	D	*2080*	1337469-1336843	627/209	0/0
	E	*2082*	1338063-1337479	585/195	
J5	*A*	*2549+*	1621739-1622782	1044/348	0/0
	A	*2549*	*1621742-1622782*	*1041/347*	-
	B	*2551*	1622800-1623945	1146/382	0/0
	C	*2553*	1623952- 1624245	294/98	0/0
	D	*2550*	1623327-1622701	627/209	0/0
	E	*2552*	1623921-1623337	585/195	1/0

* unvollständig, da Element J1 nur als Fragment vorliegt.

Die Integrase der ORF-Gruppe *A* ist vermutlich zu lang vorhergesagt. Eine Berücksichtigung von ATG als erstes gemeinsames Startcodon führt jedoch ebenfalls zu einer signifikanten Abweichung zu den beschriebenen Integrasen, die Sequenzhomologien zeigen. Eine eindeutige Bestimmung des Startcodons erscheint deshalb zurzeit nicht möglich. Unklar bleibt, wie präzise die Bestimmung des Startcodons bei einem Großteil der Sequenzhomologien-zeigenden Integrasen in der Datenbank durchgeführt wurde. Die ORFs der Gruppe *A* werden deshalb mit dem ersten gemeinsamen alternativen Startcodon genutzt. Diesem Startcodon ist eine potenzielle ribosomale Bindestelle (CGGCGGU) in einem Abstand von sieben Nukleotiden vorgelagert, die an den Consensus von *E. coli* erinnert (AGGAGGU; Lewin 2000). Der *orf3221* des Elementes J3, dessen vorhergesagtes Startcodon außerhalb des Elements liegt, wurde ebenfalls bis zum gemeinsamen potenziellen Startcodon zurückgesetzt.

Die ORFs der Gruppen *A*, *B* und *C* liegen gemeinsam auf einen Strang und die ORFs der Gruppe *D* und *E* auf dem Gegenstrang (Abb. 46).

Abb. 46: Darstellung der überlappenden ORF-Vorhersagen durch ORPHEUS am Beispiel des repetitiven Elements J5
Die potenziell kodierenden ORFs sind in Orange dargestellt.



Die ORF-Gruppe *A* zeigt eine hohe Sequenzhomologie zu Genen, denen eine Integrase-funktion zugeordnet wird (Tab. 54). Die Zuordnung wird durch die Sequenzhomologien zu Pfam00589 (Abb. 47) und COG0582 unterstützt.

Tab. 54: BLASTP Resultate am Beispiel von *orf2549+* (ORF-Gruppe *A*) des Elements J5

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Rhizobium sp. NGR234	NP_444038	vermutliche Integrase/ Re-kombinase Y4QK	308	170 bits (430)	3e-41	94/271 (34%)	150/271 (54%)	13/271 (4%)
Mesorhizobium loti	NP_085856	Integrase/ Re-combinase	299	154 bits (390)	1e-36	95/299 (31%)	153/299 (50%)	13/299 (4%)
Burkholderia fungorum	ZP_00031418	hypothetisches Protein	347	147 bits (370)	3e-34	101/311 (32%)	164/311 (52%)	23/311 (7%)
Shewanella oneidensis MR-1	NP_717639	site-specific Re-kombinase, Phagen-Integrase Familie	287	141 bits (356)	1e-32	86/270 (31%)	147/270 (53%)	11/270 (4%)
Bacteroides thetaiotaomicron	CAC47935	TpnF -Protein	279	117 bits (293)	2e-25	64/202 (31%)	112/202 (54%)	21/202 (10%)

Abb. 47: Vollständiges Alignment von *orf2549+* (ORF-Gruppe *A*) des Elementes J5 zum Consensus von Pfam 00589

orf2549+ verdeutlicht beispielhaft die Sequenzhomologien zum pfam00589, einer Phagen-Integrasen Familie. Hierbei ist es möglich 94,3 % der Consensussequenz mit dem *orf2549+* im Alignment gegenüberzustellen.

```
orf2549+ : 150  LPEVLTIEQVHELIGSATTQRMFVYFWTVYSLGLRLNEALHLQVSDIDAERGWVHVH---  206
pfam00589:   6  IEKLLAASELARPIGA----RDRAAVELLLLTGLRISELLSLRWSDIDFDKGTIFIPVRT  61

orf2549+ : 207  RGKGAKDRYVPLPTTTVRLLRNYWASHRHPSFLFPADGRKHDLAKDGVSEATTPMSETAV  266
pfam00589:  62  SGKGRKSRTVPLSDKAVEALKQYLEIYGRRDLLG---GERSDALFP--SAVGKPLSRRLL  116

orf2549+ : 267  QGAMKQITKNLRFGKKVSIHTLRHSYATHLLEAGVGLKVIQKYLGHSSLQTTMVYLHLTD  326
pfam00589: 117  RRAGKDA------GEELTPHDLRHTFATHLLEAGVDLRVIQKLLGHSSISMTQRYTHVAA  170
```

Die ORF-Gruppe *B* zeigt hohe Sequenzhomologien zu Genen, denen eine Transposasefunktion zugesprochen wird (Tab. 55). Diese Zuordnung wird durch diverse Sequenzhomologien zu transposonassoziierte Domänen in ProDom unterstützt (PD130812, PD328794, PD130811 und PD130817).

Tab. 55: BLASTP Resultate am Beispiel von *orf2551* (ORF-Gruppe *B*) des Elementes J5

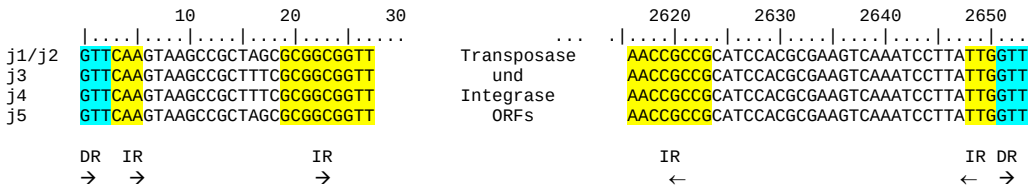
Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Bergeyella zoohelcum	AAA50501	Transposase	388	155 bits (393)	7e-37	90/284 (31%)	153/284 (53%)	13/284 (4%)
Mesorhizobium loti	NP_085855	Transposase	394	150 bits (378)	3e-35	99/327 (30%)	147/327 (44%)	37/327 (11%)
Rhizobium sp. NGR234	NP_444037	vermutliche Transposase Y4QJ	398	149 bits (376)	6e-35	99/326 (30%)	157/326 (47%)	28/326 (8%)
Shewanella oneidensis MR-1	NP_717128	Transposase, vermutlich *IS91* Familie	372	142 bits (358)	8e-33	92/307 (29%)	153/307 (48%)	12/307 (3%)
Burkholderia fungorum	ZP_00031419	hypothetisches Protein	389	129 bits (323)	9e-29	96/330 (29%)	156/330 (47%)	31/330 (9%)

Die ORF-Gruppen *C*, *D* und *E* zeigen keine Sequenzhomologien im BLASTP oder gegen COGs.

Die Gesamtstruktur der repetitiven Elemente der Gruppe J lässt sich nur schwer interpretieren. In den Randbereichen treten terminale imperfekte invertierte Repeats mit flankierenden direkten Repeats auf. Eine derartige Struktur würde auf ein IS-Element ähnlich dem der repetitiven Elemente der Gruppe G schließen lassen. Ungewöhnlich ist die Struktur der invertierten Repeats (Abb. 48). Hier scheinen zwei perfekt invertierte Bereiche die Paarung der Enden zu ermöglichen. Durch die hierbei auftretenden differenten Entfernungen (13 bp bzw. 24 bp) zwischen diesen Bereichen müsste es zusätzlich zu der Ausbildung

einer Schleifenstruktur kommen. Die Ausbildung einer derartigen Struktur erscheint möglich, kann aber abschließend nicht belegt werden. Die im Alignment auftretenden Übereinstimmungen außerhalb des potenziellen IS-Elementes können zufälliger Natur sein oder eine Präferenz der Zielstelle widerspiegeln.

Abb. 48: Alignment der Randbereiche der repetitiven Elemente der Gruppe J
Farbunterlegungen kennzeichnen in Blau die außerhalb des repetitiven Elements liegenden direkten Repeats und in Gelb die invertierten Repeats.



Die repetitiven Elemente der Gruppe J werden von zwei ORFs dominiert, die Sequenzhomologien zu einer Integrase und einer Transposase zeigen. Derartige Strukturen weisen wie beim repetitiven Element G auf ein potenziell replikatives Transposon hin. Ein Beispiel hierfür ist eine in *Bergeyella zoohelcum* identifizierte Transposon-ähnliche Struktur (Brassard et al. 1995).

3.2.2.12 Die repetitiven Elemente der Gruppe K

Das repetitive Element K wurde während der Assemblierung mit einer Länge von insgesamt 1726-1826 bp bestimmt. Es stellt sich unter Betrachtung des 1744 bp langen gemeinsamen Alignments als konserviert dar.

Tab. 56: Abweichungen zum gemeinsamen Consensus innerhalb der repetitiven Elemente der Gruppe K

Variante	Abweichungen in Nukleotiden	Prozentuale Abweichungen
K1	0	0,0
K2	7	0,4
K3	7*	0,4

*davon zwei Deletionen

Die BLASTX Analyse des Elements K gegen die Datenbank von NCBI zeigt Verwandtschaften zu Transposasen, die Teile von bekannten IS-Elementen darstellen (Tab. 57).

Tab. 57: BLASTX Resultate am Beispiel von Variante K1

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Nostoc punctiforme	ZP_00108891	hypothetisches Protein	320	183 bits (465)	6e-45	112/308 (36%)	175/308 (56%)	1/308 (0%)	-2
Nostoc sp. PCC 7120	NP_484293	Transposase	320	176 bits (446)	9e-43	97/255 (38%)	154/255 (60%)	-	-2
Pseudomonas atlantica	A32816	hypothetisches Protein, 33K Insertionsequenz *IS492*	318	154 bits (389)	4e-36	95/307 (30%)	164/307 (52%)	-	-2
Caulobacter crescentus CB15	NP_419459	ISCc2, Transposase	311	150 bits (380)	4e-35	92/259 (35%)	141/259 (53%)	4/259 (1%)	-2
Mesorhizobium loti	NP_102910	vermutliche Transposase	311	149 bits (376)	1e-34	101/304 (33%)	157/304 (51%)	6/304 (1%)	-2

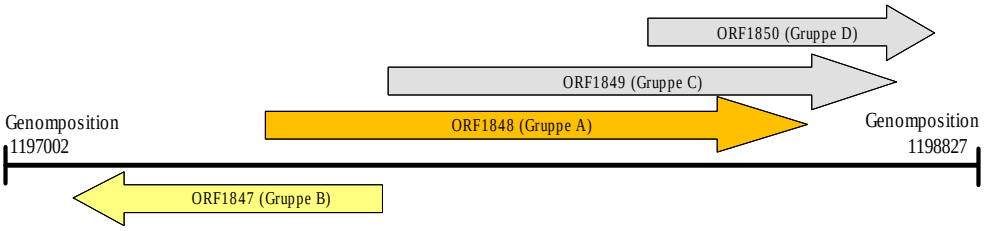
Das Programm ORPHEUS sagt jeweils vier kodierende Bereiche im repetitiven Element K voraus (Tab. 58). Weitere ORFs ohne Ähnlichkeiten zu Datenbankeinträgen ragen in die Elemente herein. Sie werden als fraglich angesehen (*orf1846*, Position 1196961-1197125; *orf1851*, Position 1199020-1198832) und in den folgenden Analysen nicht berücksichtigt, da sie im Widerspruch zur Struktur in dem vorliegenden IS-Element stehen.

Tab. 58: ORPHEUS-Vorhersagen für die repetitiven Elemente der Gruppe K
Dargestellt werden die mit ORPHEUS vorhergesagten ORFs des repetitiven Elements K. In jeder Variante werden vier ORFs vorhergesagt, die einander entsprechen und gruppiert wurden (ORF-Gruppen *A*-*D*). Abweichungen innerhalb der ORF-Gruppe *B* sind unter anderem die Folge von zwei Deletionen (*orf8216*), die zu einer Verschiebung des Stoppcodons (neun zusätzliche Nukleotide bzw. 3 Aminosäuren) führten. In ORF-Gruppe *D* wurde *orf8213* bis zum gemeinsamen Startcodon der anderen Gruppe 4 ORFs verlängert (*orf8213+*). Dem jetzt verwendeten Startcodon ist im Gegensatz zum Ursprünglichen eine ribosomale Bindestelle vorgelagert. Insgesamt ist die ORF-Vorhersage durch überlappende ORFs auf einem Strang geprägt (ORF-Gruppen *A*, *C* und *D*).

Variante	ORF-Gruppe	ORF	Position im Genom	ORF-Länge (na/aa)	Abweichungen zum gemeinsamen Consensus (na/aa)
K1	*A*	*1848*	1197486-1198508	1023/341	0/0
	C	*1849*	1198676-1197723	954/318	0/0
	D	*1850*	1198213-1198755	543/181	0/0
	B	*1847*	1197719-1197105	615/205	0/0
K2	*A*	*8215*	5446891-5445869	1023/341	0/0
	C	*8214*	5445701-5446654	954/318	0/0
	D	*8213*	*5446134-5445622*	*513/171*	-
	D	*8213+*	5446164-5445622	543/181	0/0
	B	*8216*	5446658-5447281	624/208	27/11
K3	*A*	*2804*	1792382-1793404	1023/341	0/0
	C	*2805*	1793572-1792619	954/318	7/5
	D	*2806*	1793109-1793651	543/181	7/4
	B	*2803*	1792615-1792001	615/205	0/0

Die ORFs der Gruppen *A*, *C* und *D* liegen gemeinsam auf einem Strang und die ORFs der Gruppe B auf dem Gegenstrang (Abb. 49).

Abb. 49: Darstellung der überlappenden ORF-Vorhersagen durch ORPHEUS am Beispiel des Elements K1



Die ORFs der Gruppe *A* zeigen keine Abweichungen zueinander. In den anderen ORF-Gruppen weicht immer ein ORF vom gemeinsamen Consensus ab.

Funktionszuweisungen anhand von Sequenzhomologien lassen sich nur zu den ORFs der Gruppe *A* treffen. Die mit den Sequenzen der ORF-Gruppe *A* überlappenden ORF-Gruppen *C* und *D* werden als falsch vorhergesagt beurteilt und in den folgenden Analysen nicht mehr berücksichtigt. Die hier auftretenden Sequenzhomologien zu bekannten Transposasen lassen auf Grund ihrer Zuordnung auf ein IS-Element schließen (Tab. 59).

Tab. 59: BLASTP Resultate am Beispiel von *orf1848* (ORF-Gruppe *A*)

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Nostoc punctiforme	ZP_00108891	hypothetisches Protein	320	178 bits (452)	7e-44	112/308 (36%)	175/308 (56%)	4/308 (1%)
Nostoc sp. PCC 7120	NP_484293	Transposase	320	173 bits (439)	3e-42	97/255 (38%)	154/255 (60%)	1/255 (0%)
Pseudoalteromonas atlantica	A32816	hypothetisches Protein, 33K - *Pseudomonas atlantica* Insertionsequenz IS492	318	153 bits (387)	3e-36	95/307 (30%)	165/307 (52%)	2/307 (0%)
Mesorhizobium loti	NP_102910	Transposase	311	144 bits (362)	2e-33	101/304 (33%)	157/304 (51%)	14/304 (4%)
Caulobacter crescentus CB15	NP_419459	ISCc2, transposase	311	143 bits (360)	4e-33	92/259 (35%)	141/259 (53%)	8/259 (3%)

Die Zuordnung der ORFs der Gruppe *A* zu einer Transposasefunktion wird durch eine ganze Reihe von weiteren Hits in den Datenbankeinträgen bestätigt. Die ORFs der Gruppe *A* können weiterhin dem COG-Eintrag 3547 (Transposase) zugeordnet werden sowie einer Reihe von ProDom Einträgen, denen eine Transposasefunktion zugesprochen wird (PD317542, PD358241, PD001698 und PD455746), die den größten Teil des ORFs abdekken. In Übereinstimmung hiermit erfolgt eine weitere Zuordnung zu den Pfam-Einträgen

der Transposase 20 (PF02371) und Transposase 9 (PF01548), beide assoziiert mit einer ganzen Reihe von IS-Elementen (*IS116, IS110, IS902, IS111A, IS1328* und *IS1533*). Die erreichten Übereinstimmungen der Pfam-Einträge stehen nicht im Widerspruch zueinander, da der Transposase 20 Eintrag im Bereich von 180-291 aa der ORFs Übereinstimmungen zeigt (Abb. 50b) und die Transposase 9 Übereinstimmungen im Bereich 87-166 aa auftreten (Abb. 50a). Derartige Übereinstimmungen zwischen den beiden Familien treten häufiger auf (vgl. Beschreibung Pfam PF02371).

Abb. 50:
Alignments der ORFs der Gruppe *A* zu sequenzhomologen Transposasen aus Pfam
Stellvertretend für die ORF-Gruppe *A* wurde *orf1848* der Variante K1 den entsprechenden Pfams gegenübergestellt.

Abb. 50a: Alignment der Transposase_9 Consensussequenz versus *k1_orf1848*/85-181

```
                  *->laaaglkVvyvnplavarfakayggsraKtDakDAqviAryartdlh
                     + ++  +V++vn ++v+ fak  g+ + KtD++DA v+ ++++
k1_orf1848    85    AHDNSVDVAVVNARQVRDFAKGQGR-LEKTDQIDAGVLCQFGQDV-- 128

                  rlrpllpdddivaeLreLtrrredLvadrtrlaNRlrrllrevfpalera
                  +  + +p +  +   ++L+ rre L ++r ++  Rl++  ++ +++ ++
k1_orf1848   129 KVHLTAPRTAQQKHHTALVNRREALLKMRGQERMRLEHTHDAEAIKFLEE 178

                  fds<-*
                    +
k1_orf1848   179 MLE    181
```

Abb. 50b: Alignment der Transposase_20 Consensussequenz versus *k1_orf1848*/184-291

```
                  *->lreldeqikdldaeieellrlhadaqiLlsiPGiGpitAatllaeiG
                     + +l+ + k l + ++el++++++++iLls  G+G++tA++ll+ ++
k1_orf1848   184    QKQLKSVEKRLHEILKELAKEDPKVDILLSHTGVGKVTASVLLTRLP 230

                  dDpsrFksarqlAayaGLaPrqrsSGrktgrggiskrGnrrLRraLymgA
                    + +  +++q+A ++G++P+  +SGrk+g++ i ++ ++ +R a+ym+A
k1_orf1848   231 --ELGTLNRKQVAKLVGVSPIANQSGRKDGKRPIRGG-RQDVRNAMYMAA 277

                  lvalrhpedpgsrayy<-*
                   +a+rh  dp+ +a+y
k1_orf1848   278 NSARRH--DPATKAFY    291
```

Typische invertierte terminale sowie direkte Repeats konnten im repetitiven Element K nicht identifiziert werden. Repeats treten innerhalb des Elements auf, sind jedoch nicht in den charakteristischen Regionen lokalisiert. Diese Beobachtung steht in diesem Fall jedoch nicht im Widerspruch zur Zuordnung der K Elemente zu einem IS-Element. In der BLASTP Analyse (Abb. 57) zeigt die potenzielle Transposase Sequenzhomologien zum ausführlich charakterisierten *IS492* von *Pseudomonas atlantica* (Bartlett & Silverman 1989), das wiederum ein Mitglied der *IS110*-Familie ist (Mahillon & Chandler 1998).

Hinweise auf eine Verwandtschaft zur *IS110*-Familie finden sich bereits in den Homologien zu Pfam02371 (Transposase 9). *IS492* besitzt ebenfalls keine terminalen invertierten Repeats. Die experimentell evaluierte Zielstellenverdopplung mit 5 bp Länge *IS492* findet sich unmittelbar im Anschluss an das IS-Element. Das Auftreten einer Zielstellenverdopplung in Form direkten Repeats stellt innerhalb der *IS100*-Familie jedoch eine Ausnahme dar. Abweichend ist jedoch die Größe von bis zu 1826 bp der Varianten zu den Größen in der *IS100*-Familie, die zwischen 1136 - 1558 bp liegen (Mahillon & Chandler 1998). Übereinstimmend ist wiederum die Dominanz eines langen ORFs. Das Auftreten eines weiteren ORFs im Element K muss im Zusammenhang mit der abweichenden Größe des gesamten IS-Elements gesehen werden. Die auftretenden Längenvariationen der Varianten treten auch bei anderen IS-Elementen der *IS100*-Familie, wie z.B. bei *IS900* und *IS901* der Mycobacterien (Mahillon & Chandler 1998), auf. Unbeantwortet muss in diesem Zusammenhang jedoch die Frage nach den Ursachen bleiben, da unklar ist, ob die Varianten Fragmente, Variationen oder Rudimente eines IS-Elements darstellen. Abschließend kann das repetitive Element K als ein IS-Element oder als Teil eines solchen beschrieben werden, das Übereinstimmungen mit der *IS100*-Familie aufweist.

3.2.2.13 Die repetitiven Elemente der Gruppe L

Das repetitive Element L wurde zunächst mit einer Länge von insgesamt 1306 bp bestimmt. Es stellt sich mit seinen fünf Kopien im Genom als weitgehend konserviert dar (Tab. 60). Weitere Fragmente konnten im Genom nicht aufgefunden werden.

Tab. 60: Abweichungen der Varianten des repetitiven Elements zum gemeinsamen Consensus
Als Basis wurde der hochkonservierte 1306 bp lange Sequenzbereich gewählt.

Variante	Anzahl der Abweichungen	Prozentuale Abweichung
L1	0	0,00%
L2	29	2,22%
L3	0	0,00%
L4	0	0,00%
L5	2	0,15%

Die BLASTX Analyse des Elements L (Tab. 61) gegen die Datenbank von NCBI zeigt Verwandtschaften zu Genen, die möglicherweise Proteine kodieren, die an der Rezeptor-

bildungen beteiligt sind oder mit geringerer Wahrscheinlichkeit für mögliche Transposasen kodieren.

Tab. 61: BLASTX Resultate am Beispiel von Variante L1

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Escherichia coli O157:H7 EDL933	NP_286420	vermutlicher Rezeptor	357	251 bits (640)	2e-65	140/366 (38%)	207/366 (56%)	6/366 (1%)	+1
Vibrio cholerae	AAA76604	Transposase	375	222 bits (566)	7e-57	134/383 (34%)	201/383 (51%)	7/383 (1%)	+1
Listonella anguillarum	AAD00759	vermutliche Transposase	377	218 bits (554)	2e-55	130/383 (33%)	201/383 (51%)	7/383 (1%)	+1
Aeromonas salmonicida	AAA72049	hypothetisches Protein	371	196 bits (498)	5e-49	119/376 (31%)	192/376 (50%)	4/376 (1%)	+1
Sinorhizobium meliloti	NP_386255	vermutliches Transposase Protein	358	196 bits (497)	7e-49	132/383 (34%)	185/383 (47%)	5/383 (1%)	+1

Das Programm ORPHEUS sagt jeweils einen kodierenden Bereich im repetitiven Element L voraus (Tab. 62). Überlappende ORFs, die vorausgesagt werden, sind als fraglich anzusehen. Ihnen kann im Vergleich zu anderen Genen oder Elementen, im Gegensatz zu den innerhalb des Elementes liegendem ORF, keine Funktion zugeordnet werden. Zusätzlich stehen diese ORFs im Widerspruch zu den in der Folge dargestellten eindeutigen Strukturen im Element. Der im repetitiven Element L liegende ORF zeigt sich hochkonserviert in der abgeleiteten Peptidsequenz. Abweichungen treten gehäuft nur in der Nukleotidsequenz der Variante L2 auf.

Tab. 62: Transposase verwandte ORFs im repetitiven Element L

Variante	ORF	Position im Genom	Länge in Aminosäuren	Abweichungen zum Consensus vom einheitlichen Startcodon ausgehend (na/aa)
L1	*661*	427861 - 429036	392	0/0
L2	*1035*	654572 - 655747	392	27/0
L3	*2007*	1294260 - 1295435	392	0/0
L4	*3283*	2106265 - 2107440	392	0/0
L5	*3299*	2116506 - 2117681	392	1/0

Die innerhalb des repetitiven Elements L lokalisierten ORFs spiegeln im BLASTP (Tab. 63) die Resultate der BLASTX Suchen wider.

Tab. 63: BLASTP Resultate am Beispiel von *orf661*

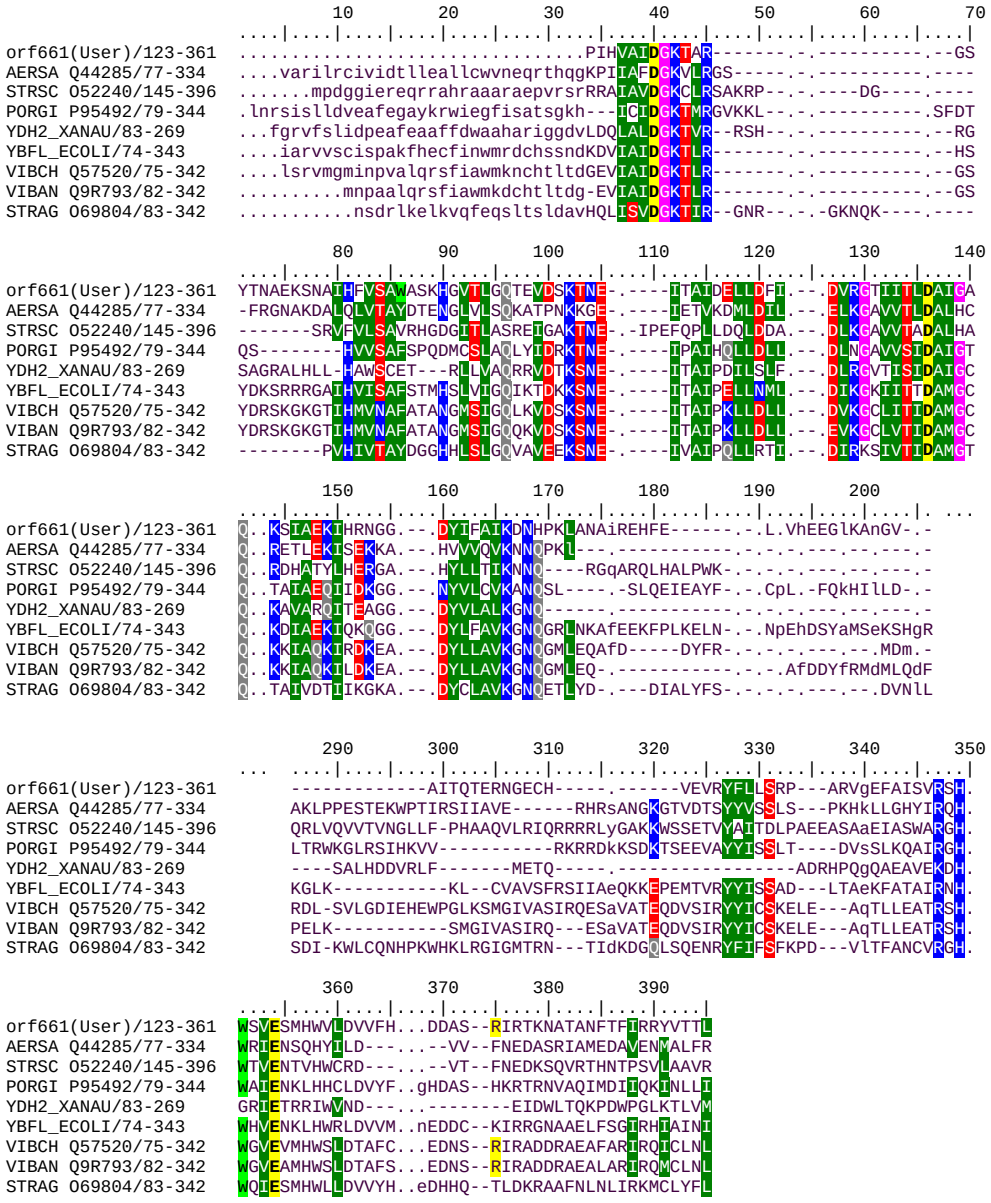
Organismus	Acc. Number	Genfunktion	Länge in aa	Score	Expect	Identities	Positives	Gaps
Escherichia coli O157:H7 EDL933	NP_286420	vermutlicher Rezeptor	357	249 bits (637)	3e-65	140/366 (38%)	207/366 (56%)	23/366 (6%)
Vibrio cholerae	S70960	Transposase	375	223 bits (568)	3e-57	134/383 (34%)	201/383 (51%)	27/383 (7%)
Listonella anguillarum	AAD00759	vermutliche Transposase	377	218 bits (555)	1e-55	130/383 (33%)	201/383 (51%)	27/383 (7%)
Klebsiella pneumoniae	CAB82577	hypothetisches Protein	375	195 bits (496)	7e-49	116/381 (30%)	191/381 (49%)	23/381 (6%)
Sinorhizobium meliloti	NP_386255	vermutliche Transposase für die Insertionsequenz ISRm21 Protein	358	190 bits (482)	3e-47	132/383 (34%)	185/383 (47%)	37/383 (9%)

Durch die Identifikation des Transposase 11 Musters (Pfam01609) innerhalb der ORFs des Elements L und die charakteristische konservierte DDE Super-Familie (Abb. 51) können die ORFs eindeutig einer Transposase Funktion zugeordnet werden. Eine entsprechende Ähnlichkeit zu COGs konnte nicht aufgefunden werden.

Eine ganze Reihe von IS Elementen wie *IS4* und *IS421* werden der Transposase 11 Familie zugeordnet. Eine weitere Zuordnung zu einem beschriebenen IS-Element aufgrund charakteristischer Aminosäuren im Umfeld des DDE Consensus (Mahillon & Chandler 1998) lässt sich nicht durchführen. Eine Verlängerung des DDE Consensus um Lysin (K) liegt nicht vor. Ein DDER(/K) Motiv ist möglich aber nicht eindeutig (Abb. 51). Ein DDER Motiv würde mit *IS982* übereinstimmen, wobei die Abstände und Consensus vom *IS982* sich deutlich unterscheiden (Mahillon & Chandler 1998), so dass eine Zuordnung zu diesem IS-Element nicht erfolgen kann.

Abb. 51: Alignment zwischen dem *orf661* der Variante L1 und Transposasen der Pfam01609 Transposase 11 zur Verdeutlichung des konservierten DDE Motives

Dargestellt werden neun Transposasen, die an der Bildung der Transposase 11 Domäne in Pfam beteiligt sind, versus dem *orf661* aus der Variante L1 stellvertretend für die anderen L Kopien. Identische und ähnliche Aminosäuren der konservierten Bereiche sind farblich markiert. Das charakteristische DDE Motiv wird im Fettdruck und Gelb unterlegt dargestellt. Die mögliche Erweiterung zum DDER Motiv ist nur Gelb unterlegt dargestellt.



Abkürzungen der Organismennamen:

Aeromonas salmonicida:	AERSA
Streptomyces scabies:	STRSC
Porphyromonas gingivalis:	PORGI
Xanthobacter autotrophicus:	XANAU
Escherichia coli:	ECOLI
Vibrio cholerae:	VIBCH
Vibrio anguillarum:	VIBAN
Streptococcus agalactiae:	STRAG

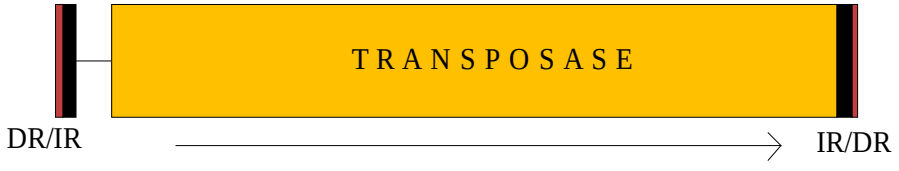
Das repetitive Element L weist die typischen Randstrukturen eines IS Elementes auf (Abb. 52). Die entsprechenden IR mit 18 bp Länge ohne Abweichungen liegen an den Enden des Elementes. Die invertierten Repeats werden von direkten Repeats aus der Zielstellenverdopplung flankiert, die mit 10 bp eine ungewöhnlich große Länge aufweisen. Die Zielsequenzen lassen keine Präferenzen für die Orte der Transposition erkennen (Abb. 51).

Abb. 52: Alignment der Randbereiche der Varianten des repetitiven Elements L
Farbunterlegungen kennzeichnen in Blau die außerhalb des Element liegenden direkten Repeats, in Gelb die invertierten Repeats und die Substitutionen in Lila. In Rot werden die Start- und Stoppcodons gekennzeichnet. In der Variante L3 liegt im zweiten direkten Repeat eine Deletion vor.

```
            10        20        30   115          1291     1300      1310      1320
   ....|....|....|....|....|....| ... |.. ...   ... |...|....|....|....|....|....|..
l1 CCGGCGTATCCACGGCGATCGCATCAAATT     ATG           TGATTTTGATGCGATCGCCGTGCCGGCGTATC
l2 GATACAAGTGCACGGCGATCGCATCAAAAT     ATG           TGATTTTGATGCGATCGCCGTGGATACAAGTG
l3 GCGTCAAGGGCACGGCGATCGCATCAAATT     ATG   TRANS-  TGATTTTGATGCGATCGCCGTG-CGTCAAGGG
l4 CGCCCGCACCCACGGCGATCGCATCAAATT     ATG   POSASE  TGATTTTGATGCGATCGCCGTGCGCCCGCACC
l5 CGCTCGACCTCACGGCGATCGCATCAAATT     ATG           TGATTTTGATGCGATCGCCGTGCGCTCGACCT
       DR          IR                                      IR          DR
       →           →                                       ←           →
```

Auffällig ist die Lücke von 85 bp zwischen dem invertierten Repeat und dem Methionin der Transposase. In diesem Bereich lassen sich keine ORFs identifizieren, die nicht weitgehend mit den invertierten Repeat überlappen. Eine Verlängerung der ORFs zu einem alternativen Startcodon (GTG) ist möglich. Dagegen spricht, dass keines der vollständigen ORFs der mehr als 60 Pfam01609 Transposasen ein alternatives Startcodon nutzt. Aus diesem Grund wird das Startcodon Methionin favorisiert. ORFs, die nicht im Widerspruch zu den charakterisierten Strukturen stehen und innerhalb der invertierten Repeats liegen, konnten nicht aufgefunden werden.

Abb. 53: Schematische Darstellung des IS-Elements im repetitiven Element L
Die im repetitiven Element L identifizierten Elemente wie die direkten Repeats (DR), die flankierenden invertierten Repeats (IR) und die vermutliche Transposase spiegeln ein einfach aufgebautes IS-Element wider. Die Pfeile geben die Transkriptionsrichtung an.



3.2.2.14 Die repetitiven Elemente der Gruppe M

Das repetitive Element M wurde während der Assemblierung mit einer Länge von insgesamt 1281 bp bestimmt. Das repetitive Element M liegt zweimal im Genom vor. Weitere Fragmente konnten im Genom nicht aufgefunden werden. Die Varianten des repetitiven Elements M weichen in nur geringem Maße voneinander ab (Tab. 64). Im repetitiven Element M2 tritt eine Deletion auf.

Tab. 64: Abweichungen der Varianten des repetitiven Elements M zum gemeinsamen Consensus
Als Basis wurde der 1281 bp lange Sequenzbereich von M2 gewählt.

Variante	Anzahl der Abweichungen	Prozentuale Abweichung
M1	8	0,6
M2	9	0,7

Die BLASTX Analyse des repetitiven Elements I gegen die Datenbank von NCBI weist auf Verwandtschaften zu Genen, denen eine Transposasefunktion zugesprochen wird, hin (Tab. 65). Auffällig sind bei den homologen Transposasen die unterschiedlichen Längen.

Tab. 65: BLASTX Resultat am Beispiel von Element M1

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps	Frame
Brucella melitensis biovar Abortus	AAL59362	Vermutliche Transposase	392	174 bits (441)	3e-42	127/380 (33%)	188/380 (49%)	6/380 (1%)	-1
Brucella suis 1330	AAN34276	IS3 Familie, Transposase orfB	293	157 bits (398)	3e-37	98/262 (37%)	139/262 (53%)	2/262 (0%)	-1
Shigella dysenteriae	AAF28144	InsB	307	155 bits (392)	1e-36	100/280 (35%)	149/280 (53%)	6/280 (2%)	-1
Escherichia coli O157:H7 EDL933	D85653	Vermutliche Transposase	272	154 bits (390)	2e-36	96/264 (36%)	144/264 (54%)	3/264 (1%)	-1
Escherichia coli CFT073	AAN82023	Transposase insF für Insertions-sequenz IS3A/B/C/D/E/fA	272	153 bits (386)	7e-36	95/264 (35%)	143/264 (54%)	3/264 (1%)	-1

Das Programm ORPHEUS sagt einen ORF für das repetitive Element M1 und zwei ORFs für das repetitive Element M2 voraus (Tab. folgende). Die Ursachen für diese unterschiedliche Vorhersage liegen in der auftretenden Deletion und den Substitutionen (Tab. 66).

Tab. 66: Mit ORPHEUS vorhergesagte ORFs im repetitiven Element M
Das repetitive Element M1 wird durch *orf150* dominiert, dieser orf spiegelt sich entsprechend in *orf666* und *orf667* von M2 wieder. *Orf667* läßt sich über das Startcodon hinaus im Leserahmen verlängern (*orf667+*).

Variante	ORF	Position im Genom	Länge in Nukleotiden/Aminosäuren
M1	*orf150*	88726 – 87560	1167/389
M2	*orf666*	429942 – 430238	297/99
	orf667	430238 - 431107	870/290
	orf667+	430199 - 431107	909/303

orf150 aus dem repetitiven Element M1 zeigt hohe Homologien zu *orf666* und *orf667*. Die Sequenzen von *orf666* und *orf667* finden sich mit Abweichungen in der gesamten Länge in *orf150* aus M1 wieder.

Vor *orf667* befindet sich ein mögliches *Frameshift Window* (vgl. Kap. 3.2.2.3), welches durch Verlängerung der Sequenz von *orf667* mit einbezogen werden kann (*orf667+*). Die ORFs in den beiden repetitiven Elementen weisen Homologien zu Genen auf, die in Zusammenhang mit Transposasefunktionen stehen (Tab. 67).

orf666 und *orf667* führen vermutlich über ein programmiertes *Frameshifting* Ereignis (Sato et al. 1989) zum Fusionsprotein, der Transposase. Dieses *Frameshifting Window* (AAAAAAAG) hat im repetitiven Element M1 keine Bedeutung, da ein durchgängiges ORF vorliegt (Abb. 55). Lediglich fünf abweichende Aminosäuren treten zwischen *orf150* des Elements M1 und den ORFs des repetitiven Elements M2 (*orf666* und *orf667+*) auf, entsprechend übereinstimmend fallen die Zuweisungen zu den übergeordneten Gruppen aus. Hier spiegelt sich die Zuordnung der ORFs zur Integrasen-/Transposasenfunktion wider (vgl. auch Kap. 3.2.2.3): Integrasenkerndomäne (Pfam00665), Transposase 8 Familie (Pfam01527) und orthologe Transposasen (COG2801, COG2963). Gemeinsam zeigen alle diese Zuordnungen auch eine Verwandschaft zur *IS3*-Familie auf.

Tab. 67: BLASTP Resultate der innerhalb des repetitiven Elements M identifizierten ORFs

M1: *orf150*

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Brucella melitensis biovar Abortus	AAL59362	Transposase	392	171 bits (433)	2^{e}-41	126/380 (33%)	187/380 (49%)	12/380 (3%)
Brucella suis 1330	AAN34276	IS3 Familie, Transposase orfB	293	158 bits (400)	1^{e}-37	98/262 (37%)	139/262 (53%)	4/262 (1%)
Shewanella oneidensis MR-1	AE015890	ISSod1, Transposase OrfB	269	148 bits (374)	1^{e}-34	85/263 (32%)	137/263 (52%)	3/263 (1%)
Escherichia coli (strain O157:H7, substrain RIMD 0509952)	G90792	Transposase	272	147 bits (372)	2^{e}-34	98/265 (36%)	146/265 (55%)	Gaps = 8/265 (3%)
Erwinia amylovora	S21562	ORF B	285	147 bits (370)	4^{e}-34	83/273 (30%)	147/273 (53%)	4/273 (1%)

M2: *orf666*

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Ralstonia solanacearum	CAD15250	Transposase-verwandtes Protein	95	42.0 bits (97)	0.002	25/55 (45%)	38/55 (69%)	3/55 (5%)
Brucella suis 1330	AAN34240	IS3 Familie, Transposase orfA	93	35.8 bits (81)	0.14	29/90 (32%)	46/90 (51%)	6/90 (6%)
Pseudomonas aeruginosa	AAO47354	Vermutliche Transposase Untereinheit	102	34.7 bits (78)	0.30	23/52 (44%)	33/52 (63%)	3/52 (5%)
Pseudomonas putida KT2440	AAN69996	Transposase, OrfA	102	33.1 bits (74)	0.83	20/53 (37%)	30/53 (56%)	2/53 (3%)
Corynebacterium glutamicum ATCC 13032	BAB99127	Transposase	98	32.7 bits (73)	1.2	19/38 (50%)	24/38 (63%)	2/38 (5%)

M2: *orf667+*

Organismus	Acc. Number	Genfunktion	Länge (aa)	Score	e-value	Identities	Positives	Gaps
Brucella suis 1330	AAN34276	IS3 Familie, Transposase orfB	293	154 bits (388)	2e-36	97/262 (37%)	137/262 (52%)	4/262 (1%)
Brucella melitensis biovar Abortus	AAL59362	Vermutliche Transposase	392	150 bits (379)	2e-35	96/262 (36%)	137/262 (52%)	4/262 (1%)
Leptospira interrogans	AAC05649	Vermutliches transposase-verwandtes Protein	278	148 bits (374)	1e-34	87/271 (32%)	141/271 (52%)	8/271 (2%)
Escherichia coli CFT073	AAN82023	Transposase insF für Insertions-sequenz IS3A/B/C/D/E/Fa	272	147 bits (370)	2e-34	95/264 (35%)	142/264 (53%)	6/264 (2%)
Shigella dysenteriae	AAF28120	InsB	292	147 bits (370)	3e-34	95/264 (35%)	142/264 (53%)	6/264 (2%)

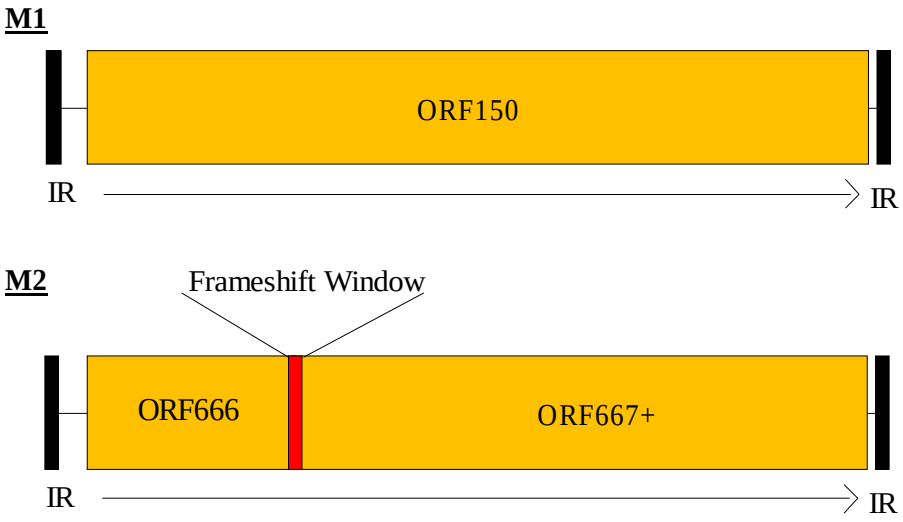
Das Vorliegen eines IS-Elements wird durch die in den Randbereichen identifizierten invertierten Repeats mit einer Länge von 31 bp gestützt (Abb. 54). Sich unmittelbar anschließende direkte Repeats konnten nicht bestimmt werden.

Abb. 54: Flankierende invertierte Repeats im repetitiven Element M
Die repetitiven Elemente der Gruppe M werden durch konservierte perfekte terminale invertierte Repeats flankiert.

		Trans-posase ORF(s)	
M1	GTGAGGTGGCCTGGGTTTTGGGGGCCACGCG		CGCGTGGCCCCCAAAACCCAGGCCACCTCAC
M2	GTGAGGTGGCCTGGGTTTTGGGGGCCACGCG		CGCGTGGCCCCCAAAACCCAGGCCACCTCAC

Abschließend werden die repetitiven Elemente der Gruppe M der *IS3*-Familie zugeordnet. M1 und M2 werden als Varianten angesehen, deren Genprodukt die Transposase ist. Das vermutliche Auftreten eines Fusionsproteins erinnert an das repetitive Element B. Gemeinsamkeiten bestehen hier bei den Zuordnungen der Domänen und Familien.

Abb. 55: Schematische Darstellung des IS-Elements im repetitiven Element M
Die im Element identifizierten Elemente wie die flankierenden invertierten Repeats (IR) und die potenziell für die Transposition kodierende(n) ORF(s) spiegeln die Strukturen eines IS-Elements wider. Die Pfeile geben die Transkriptionsrichtung an.



3.2.2.15 Zusammenfassung der Analysen der repetitiven Elemente

Repetitive Elemente verhinderten bei der Assemblierung in 52 Fällen das automatische Erstellen einer durchgängigen Consensussequenz. Fehler und für den Assembler Phrap nicht auflösbare Widersprüche führten zu diesem Resultat. Unklar blieb zunächst, welche Information diese Sequenzen beinhalten.

Insgesamt 62 dieser repetitiven Elemente, die sich in 13 Gruppen (A-M) einteilen lassen, konnten im Genom von *Pirellula* identifiziert werden (Anhang 7.3, Tab. 71). Die repetitiven Elemente liegen in unterschiedlicher Anzahl von zwei bis zu zehn Kopien im Genom

vor. Auch die Länge von 668 bp bis zu 4458 bp variiert deutlich und spiegelt die Heterogenität der Gruppen wider. Alle repetitiven Elemente beinhalten potenzielle kodierende Funktionen, die in den meisten Fällen Rückschlüsse auf ihre Genese zulassen. Der Großteil dieser repetitiven Sequenzabschnitte lässt sich mobilen Elementen zuordnen, hierzu lässt sich im erweiterten Sinne eine phagenverwandte Integrasestruktur zählen. Lediglich eine Gruppe der repetitiven Elemente fällt heraus. Hierbei handelt es sich um ein mit zwei Kopien im Genom vorliegendes Restriktions/Modifikationssystem (Gruppe C), dessen Duplikation vielleicht mit einem ungleichen Crossing-over zu erklären ist.

Zehn der zwölf identifizierten repetitiven Elemente weisen Strukturen auf, die auf IS-Elemente oder Transposons schließen lassen.

Bei den 46 repetitiv auftretenden potenziellen IS-Elementen konnten keine bereits beschriebenen Insertionssequenzen aufgefunden werden, die mit mehr als einer Kopie im Genom vorliegen. Lediglich Sequenzhomologien zu bekannten IS-Elementen oder Familien (repetitive Elemente B, E, K, L, M) konnten identifiziert werden (Tab.69).

Tab. 68: IS-Elemente und potenzielle Transposons mit mehr als einer Kopie im *Pirellula* Genom

Nicht bei allen potenziellen IS-Elementen und Transposons konnten direkte und/oder invertierte terminale Repeats identifiziert werden. Die nicht erfolgte Identifikation muss nicht auf ein generelles Fehlen dieser Strukturen hindeuten, sondern kann das Resultat einer Überformung in der Evolution des Genoms darstellen. Die in Klammern angegebenen Zahlen weisen auf eine weitere Variante hin.

Element	Länge der terminalen direkten Repeats in bp	Länge der terminalen invertierten Repeats in bp	Anzahl im Genom	IS-Familie
A	4	14	10	-
B	3	34	9	*IS3*
D	9 (3)	44 (29)	5	*IS4*
E	8 (4)	13	3	-
F	nicht identifiziert	19	4	*IS4*
G	nicht identifiziert	max. 36*	5	-
I	nicht identifiziert	24*	5	-
J	3	24	6	-
K	nicht identifiziert	nicht identifiziert	3	*IS100*
L	10	18	5	-
M	31	nicht identifiziert	2	*IS3*

*Die genaue Länge des invertierten Repeats ließ sich nicht eindeutig bestimmen.

Größere Mengen an IS-Elementen sind nicht ungewöhnlich und kommen auch in anderen bakteriellen Genomen vor (Tab. 69). Ihre Anzahl im Genom korreliert nicht zwingend mit der Genomgröße.

Tab. 69: IS-Elemente in bakteriellen Genomen

Organismus	Genomgröße (Mb)	Anzahl an IS-Elementen
Archaeoglobus fulgidus	2,2	13[1]
Deinococcus radiodurans	3,3	52[1]
Bacillus subtilis	4,2	0[1]
Escherichia coli	4,6	37[1]
Mycobacterium tuberculosis	4,4	32[1]
Pirellula sp. Stamm 1	7,2	61[2]

[1] nach Makarova et al. 2001

[2] in repetitiven Elementen identifizierte IS-Elemente, ergänzt durch die Anzahl weiterer potenzieller IS-Elemente (15; Glöckner et al. eingereicht) im Genom

Neben den scheinbar negativen Folgen der mit der Insertion verbundenen Mutationen verleihen IS-Elemente dem Genom auch Plastizität bzw. die Möglichkeit chromosomaler Neuordnung und fördern intramolekulare Rearrangements. Insertionssequenzen kommen in vielen, jedoch nicht in allen Bakteriengenomen vor und fehlen unter anderem in den Chlamydiengenomen *C. trachomatis* und *C. pneumoniae* (Read et al. 2000). IS-Elemente sind auch in Phagen, Plasmiden oder lysogenen Viren nachgewiesen und können somit auch auf andere Zellen übertragen werden (Deonier 1996).

An beiden Enden der Insertionssequenz befinden sich häufig perfekte oder imperfekte umgekehrte Sequenzwiederholungen (invertierte Repeats, IRs) mit einer Länge von 10 bis 40 bp (Mahillon & Chandler 1998). Die im *Pirellula* Genom im Rahmen dieser Analyse identifizierten invertierten Repeats haben eine Länge von 13- 44 bp (Tab. 68). Die invertierten Repeats beinhalten zwei funktionale Domänen, eine ist involviert in der Bindung der Transposase, die andere in der Restriktion und Strangtransferreaktion (Mahillon & Chandler 1998).

Zwischen dem invertierten Repeat liegt der für ein oder zwei Gene kodierende Sequenzbereich (Transposase und ggf. Transposonassoziierter ORF), der für die Transposition eines IS-Elements an eine neue Stelle des Genoms erforderlich ist, die Transposase. Das Protein der Transposase kann auch aus zwei ORFs entstehen, die über translationales *Frameshifting* ein einzelnes Protein bilden. Dieser Fall tritt im *Pirellula* Genom in den repetitiven Element B und M auf (vgl. Kapitel 3.2.2.3 und 3.2.2.14).

Meist unmittelbar flankierend zum *inverted Repeat* befinden sich direkte Repeats mit einer Länge von 2-14 bp. Die identifizierten terminalen Repeats im Genom von *Pirellula* erreichen eine Länge von 2-10 bp. Die Länge der direkten Repeats ist für das jeweilige Element charakteristisch, nicht jedoch die Sequenz. Fehlende Repeats sind mit Vorsicht zu interpretieren, da sie durch Rekombinationsereignisse nicht mehr identifizierbar sein können (Mahillon & Chandler 1998). Analysen zahlreicher Genome lassen aber die Frage offen, ob direkte und invertierte terminale Repeats entweder schnell im Genom überformt werden oder ob auch andere Transpositionsmechanismen wirken, die zu Varianten oder dem Fehlen derartiger Strukturen in IS-Elementen führen (vgl. Datenzusammenstellung Mahillon & Chandler 1998). Beispiele hierfür sind die Vertreter der *IS1*-Familie *IS1A*, *IS1B*, *IS1C* und *IS1F*, die nicht von direkten Repeats flankiert werden (Deonier 1996).
Sequenzhomologien zu potenziell replikativen Transposons finden sich in den repetitiven Elementen der Gruppe G und J. Die potenziellen replikativen Transposons dieser Gruppen fallen auch durch ihre größere Länge von über 2,5 kb auf.

Die überwiegende Anzahl an IS-Elementen und Transposasen präferieren keine Zielsequenzen für die Transposition, so dass die direkten Repeats sich in ihrer Sequenz unterscheiden können. Auch Präferenzen von Zielsequenzen oder Bereichen des Genoms können auftreten und sich in *Hot Spots* widerspiegeln (Mahillon & Chandler 1998, Lewin 2000). Präferenzen von Zielsequenzen treten in den repetiven Elementen der Gruppe A und J auf. Hot Spots stellen sicherlich die Lokalisationen 605’571-609’533 (repetitive Elemente J1, B3, J3) sowie 2’106’160-2'117’786 (repetitive Elemente L4 und L5) im Genom dar. Aber auch 487’599-505'917 (repetitive Elemente H3, G3, G5 und G4) und 2'214’462-2'228'263 (repetitive Elemente B1, E3, F2 und F3) stellen dynamische Bereiche des Genoms dar. Die restlichen repetitiven Elemente verteilen sich überwiegend gleichmäßig über das Genom, wobei es auch Bereiche über mehrere 100 kb ohne das Vorkommen eines repetitiven Elements gibt (Abb. 56).

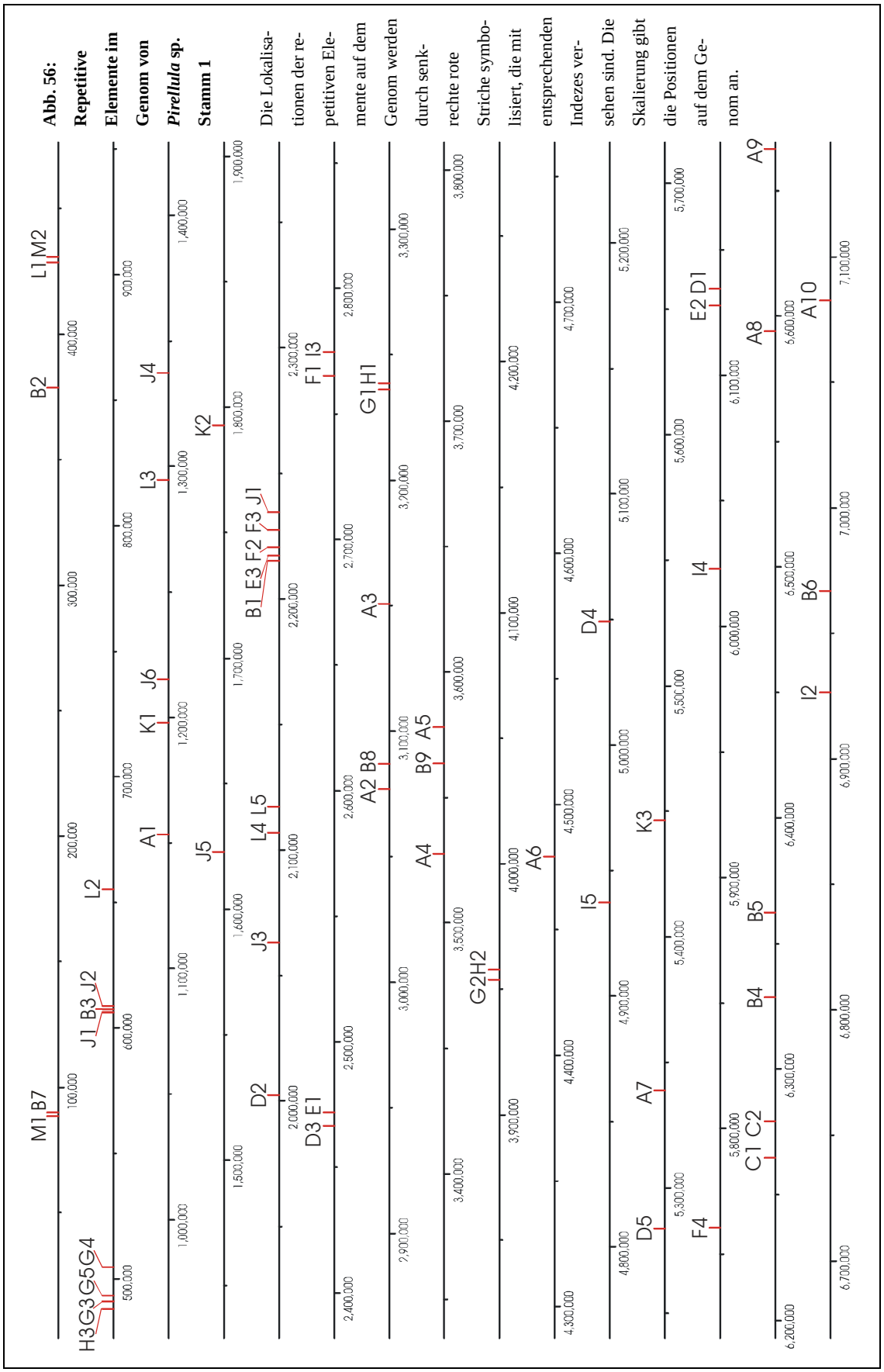


Abb. 56: Repetitive Elemente im Genom von *Pirellula* sp. Stamm 1

Die Lokalisationen der repetitiven Elemente auf dem Genom werden durch senkrechte rote Striche symbolisiert, die mit entsprechenden Indezes versehen sind. Die Skalierung gibt die Positionen auf dem Genom an.

3.2.3 tRNAs

Mithilfe des Programms tRNAscan-SE (Lowe & Eddy 1997) konnten alle 20 Aminosäure kodierenden tRNAs in *Pirellula* sp. Stamm 1 identifiziert werden. Insgesamt wurden 70 tRNAs und sechs Pseudogene identifiziert (Anhang, Tab. 72). Auffällig erscheint lediglich der hohe Wert von elf für die Aminosäure Leucin kodierenden tRNAs.

Die tRNAs weisen im Genom bei weitgehend gleichförmiger Verteilung zwei Bereiche auffälliger Konzentrierung auf. Neun tRNAs sind von der Position 425’655 bis 427'552 lokalisiert und 24 (davon vier Pseudo-tRNAs) befinden sich zwischen Position 6'198’097 und 6'201'871 (Abb. 57). Dieser *Hot Spot* (maximaler Abstand zwischen zwei tRNAs 500b) von 24 tRNAs innerhalb von 3,8 kb erscheint bemerkenswert. Lokale Häufungen von tRNAs treten in bakteriellen Genomen häufig in kleinen inselartigen Gruppen auf. *Hot Spots* von mehr 16 oder mehr tRNAs finden sich nur in einer Minderheit der bisher sequenzierten mikrobiellen Genome (Tab. 70). Ähnlichkeiten zu der tRNA-Abfolge im *Pirellula Hot Spot* konnten in anderen *Hot Spots* nicht gefunden werden.

Tab. 70: Mikobielle Genome mit tRNA Hot Spots von mehr als 16 tRNAs

Die Bestimmung der tRNAs erfolgte mit tRNAscan-SE (Lowe & Eddy 1997). Die Hot Spots bestehen aus mindestens 16 tRNAs, deren Abstand zueinander maximal 500 bp beträgt.

Organismus (Accession number)	**Anzahl der tRNAs im *Hot Spot* (davon Pseudo-tRNAs)**	**Länge des *Hot Spots (kb)***	**Position im Genom**	**Gesamtzahl der tRNAs im Genom**	**Prozent lokalisiert im Hotspot**
Staphylococcus aureus subsp. aureus N315 (NC_002745)	26 (1)	2,4	1916267 – 1918711	62	41,9
Pirellula sp. Stamm 1	24 (4)	3,8	6198097 - 6201871	76	31,6
Lactobacillus plantarum WCFS1 (NC_004567)	22	2,1	2013685 – 2015774	72	30,6
Listeria monocytogenes Stamm EGD (NC_003210)	21 16	1,9 4,7	1740916 – 1742814 2441257 – 2436576	67	55,2
Streptococcus pneumoniae (NC_003028)	17	1,5	1808658 – 1810161	58	29,3
Bacillus subtilis (NC_000964)	16	1,7	951088 – 952800	86	18,6

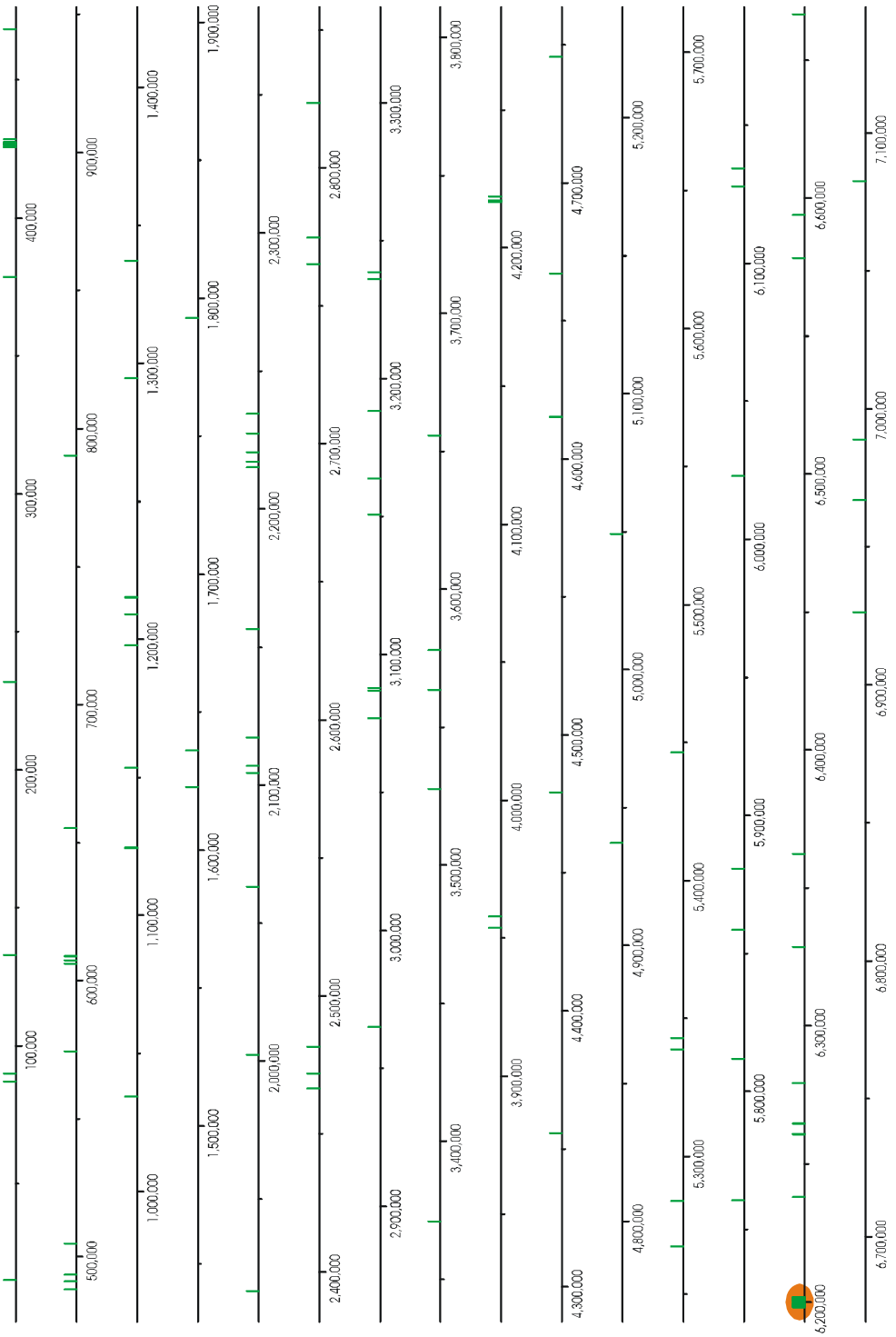


Abb. 57: Verteilung der tRNAs im Genom von *Pirellula* sp. Stamm 1
Die Lokalisationen der tRNAs auf dem Genom werden durch senkrechte grüne Striche symbolisiert. Der tRNA-Hotspot ist in Orange unterlegt. Die Skalierung gibt die Positionen auf dem Genom an.

3.2.4 rRNA-Operon

Eine Besonderheit der Gruppe der Planctomyceten stellt das nicht vollständig geschlossen vorliegende rRNA-Operon dar, das sich auch im Genom von *Pirellula* sp. Stamm 1 findet. Für die beschriebenen *Bacteria* stellt der enge Verbund der rRNA Gene als eine Transkriptionseinheit in der Anordnung 5'-16S-23S-5S-3' den Normalfall dar. Diese Aussage trifft für alle *Proteobacteria*, gram-positiven Bakterien und Cyanobakterien zu. Ausnahmen stellen z.B. *Mycoplasma hypopneumoniae* mit der separiert vorliegenden 5S rRNA und *Vibrio (Beneckea) harveyi* mit der Anordnung 5'-23S-16S-5S-3' dar. In allen *Bacteria* liegen 16S und 23S rDNAs als intakte Transkriptonseinheit vor. Im Gegensatz dazu liegen die zwei 23S-5S rRNA-Operons in *Thermus thermophilus* separiert von der 16S rRNA vor. In der Gruppe der Planctomyceten konnte für den bisher nähesten untersuchten Verwandten von *Pirellula* sp. Stamm 1, *Pirellula marina*, ebenfalls ein unterbrochenes rRNA Operon nachgewiesen werden. *P. marina* weist zwei 23S-5S rRNA Operons auf, die 8,5 und 4,4 kb von ihren 16S rRNA Genen entfernt lokalisiert sind (Liesack & Stackebrandt 1989). *Pirellula* sp. Stamm 1 weist ebenfalls dieses aufgelöste 16S-23S-5S rRNA Operon auf. Hierbei tritt die Auflösung der Transkriptionseinheit jedoch wesentlich deutlicher auf. 16S und die 23S-5S rRNA Einheit liegen ca. 467 kb voneinander entfernt im Genom vor. Da eine experimentelle Bestimmung des rRNA-Operons für *Pirellula* sp. Stamm 1 noch aussteht, konnte das Operon nicht präzise bestimmt werden. Sequenzhomologien zu den Genen des Operons finden sich jedoch nur einmal im Genom. Es wurden hierzu über 100000 Einträge der Genembl Datenbank herangezogen. Im Vergleich zum experimentiell verifizierten (Liesack et al. 1992b; Liesack et al. 1988; Bomar et al. 1988) rRNA Operon von *P. marina* (16S rRNA: X62912; 23S rRNA: X07408; 5S rRNA: M35165) lassen sich Sequenzhomologien zu *Pirellula* sp. Stamm 1 an den Positionen 5'078'494 - 5'076'959 (16S rRNA), 4'614'312 – 4'611'419 (23S rRNA) und 4'611'312 – 4'611'203 (5S rRNA) auffinden. Weitere potenzielle Sequenzhomologien lassen sich nicht im Genom auffinden, so dass nicht von weiteren degenerierten rRNA-Genen des Operons ausgegangen werden kann.

Ein gravierender Unterschied ist die Anzahl der Kopien im Genom. *Pirellula* sp. Stamm 1 weist lediglich jeweils ein Gen der 5S, 16S und 23S rRNA im Genom auf, während im Genom von *P. marina* zwei Kopien auftreten (Liesack & Stackebrandt 1989). Das Auftreten lediglich eines Sets an rRNA Genen ist auch für große Genome durchaus nicht ungewöhnlich. So besitzt das über 9 Mb große Genom von *Bradyrhizobium japo-*

nicum ebenfalls nur ein rRNA Operon, andere Genome wie *Streptomyces coelicolor* A3(2) mit 8,67 Mb besitzen sechs (Bentley et al. 2002).

4. Ausblick

Die Wahl des Organismus für mikrobielle Genomprojekte befindet sich im Wandel. Standen in der Vergangenheit mehrheitlich pathogene Erreger oder biotechnologisch bedeutsamen Organismen bei Sequenzierprojekten im Vordergrund, so gelangen umweltrelevante Organismen, wie im Rahmen des REGX-Projekts, zunehmend in den Mittelpunkt des Interesses. Im Fokus liegen somit Organismen, die auch aus weitestgehend uncharakterisierten Gruppen kommen, so dass es möglich sein wird, einen diverseren Einblick in die Struktur von mikrobiellen Genome zu erlangen.

Die Methoden zur Datenerhebung im Rahmen von mikrobiellen Genomprojekten stehen bereits zur Verfügung. Sie werden sich in naher Zukunft jedoch wieder wandeln. Die Klonierung der genomischen DNA in großen BAC-Banken wird trotz aller Schwierigkeiten in den Vordergrund treten, da die zeitaufwendigen DNA-Präparationen für die Subklonierungsschritte durch die Verwendung neuer Polymerasen wie Phi29 in (Amersham, New Jersey/US) in den Hintergrund treten werden. Die Verwendung der hiermit verbundenen *Rolling Circle* Methode (Dean et al. 2002) wird sich auch bei der Bereitstellung der Ausgangs-DNA für die Sequenzierung als Standard etablieren, wodurch der Hochdurchsatz eine neue Geschwindigkeit erreichen wird. Nach der Etablierung dieser Methoden wird die Bedeutung von BAC-Banken für Genomprojekte ein neues Gewicht bekommen, da die aufwendige Isolierung der BAC-DNA für die folgenden Subklonierungen entfällt.

Die exponentiell ansteigende Datenflut wirkt sich bereits heute auf die Datenanalyse der Genomsequenz aus. Zunehmende Vergleichsdaten ermöglichen detailliertere hypothetische Funktionsbeschreibungen. Das Durchführen der Analyse *in silico* ist für einen Großteil der Genomanalyse Standard. Diese Datenanalysen sind auf leistungsfähige Software-Plattformen angewiesen, die den Annotatoren mit Vorschlägen ein schnelles fundiertes Urteil ermöglichen. Software Plattformen wie Ergo (Integrated Genomics, Chicago/US) leisten dies bereits im kommerziellen Sektor, andere Entwicklungen wie HTGA (Rabus et al. 2002b) gehen bereits weiter und geben den Annotatoren eine automatisch generierte Annotation zur Überprüfung vor.

Die Analyse der Strukturen des Genoms wird mit der Publikation der Sequenz und der ersten Annotation (Glöckner et al. eingereicht) keinesfalls abgeschlossen sein. Die folgende Sequenzierung und Analyse der Genome von *Gemmata obscuriglobus* UQM2246 (http://www.tigr.org/tdb/mdb/mdbinprogress.html) und *Gemmata* sp. Wa1-1 (http://wit.integratedgenomics.com/GOLD/) werden weitere Möglichkeiten der Analyse geben, wie zum Beispiel zur Klärung der offenen Fragen nach den gemeinsamen Ursprün-

gen der Chlamydien und Planctomyceten. So zeigt das Genom von *Pirellula* sp. Stamm 1 zum Beispiel, wie auch die Chlamydien Genome von *C. trachomatis* (AE001273) und *C. muridarum* (AE002160), jeweils zwei potenzielle Gene die Sequenzhomologien für dnaA und für die Gyrase-Untereinheiten gyrA und gyrB aufweisen. Derartige Beobachtungen würden für sich genommen bei der Genomgröße von *Pirellula* sp. Stamm 1 nicht verwundern, während die reduzierten Genome der Chlamydien mit unter 1,5 Megabasen Fragen offen lassen. Das doppelte Auftreten dieser Gene in einem Genom ist bisher nur bei den *Chlamydien* und dem *Planctomyceten Pirellula* sp. Stamm 1 bekannt. Weitere Genomsequenzen können hier und in vielen anderen Fällen Ansatzpunkte zu Klärung bereitstellen.

Die weiteren Analysen des Genoms von *Pirellula* sp. Stamm 1 werden sich jedoch keinesfalls auf die *in silico* gewonnenen Daten beschränken, so laufen bereits die ersten Studien zum Transkriptom und Proteom von *Pirellula* an. Die Effizienz dieser Analysen wird durch verbesserte Geräte und Materialien kontinuierlich steigen, so dass die Herstellung von DNA-Chips und der massenspektroskopische Fingerabdruck zur Charakterisierung in der Zukunft zu jedem Genom gehören wird.

Die Verknüpfung der Daten des Genoms, Transkriptoms und Proteoms wird weitere Genomanalysen auf einem Qualitätsniveau ermöglichen, das wir zurzeit leider nur für wenige Organismen kennen.

5. Zusammenfassung

Sequenzierung und Strukturen des Genoms von *Pirellula* sp. Stamm 1 stehen im Mittelpunkt der vorliegenden Arbeit. Bei dem Genom von *Pirellula* sp. Stamm 1 handelt es sich um das erste sequenzierte Genom aus dem tief abzweigenden Phylum Planctomycetales. Durch die Verwendung der *whole genome shotgun* Strategie gelang es auf hohem Qualitätsniveau die Sequenz des zirkulären 7'145'576 bp großen Genoms lückenlos zu bestimmen. Das notwendige fast achtfache *Sequencing Coverage* ließ sich durch einen hohen Grad der Automatisierung erreichen. Bei dem Genom von *Pirellula* sp. Stamm 1 handelt es sich um eines der größten bisher sequenzierten Bakteriengenome. Die im Genom auftretenden repetitive Elemente führten zunächst zu Problemen bei der Assemblierung der Einzelsequenzen. 62 identifizierte repetitive Elemente wurden in 13 Gruppen eingeteilt. Diese Gruppen beinhalten bisher unbekannte bakterielle Insertionssequenzen.

Als weitere Besonderheiten dieses Genoms, die in dieser Arbeit hervorgehoben werden, sind die Verteilung der tRNAs und das nicht als klassische Einheit vorliegende rRNA-Operon zu nennen. Die tRNA-Verteilung zeigt außergewöhnliche lokale Konzentrationen, die in einem Fall ein Drittel der gesamten identifizierten tRNAs auf sich vereint.
Der Abstand der 16S rRNA Untereinheit zu der 23S und der 5S rRNA Untereinheit von 467 kb bestätigt die bisherigen Erkenntnisse über das nicht als Einheit vorliegende rRNA-Operon im Phylum Planctomycetales auf deutliche Weise.

Die vorliegende vollständige Sequenz von *Pirellula* sp. Stamm 1 stellt die Basis für weitere Analysen des Genoms dar.

6. Literatur

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams, SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC (2000). The genome sequence of *Drosophila melanogaster*. Science 287: 2185-95.

Alm EW, Oerther DB, Larsen N, Stahl DA, Raskin L (1996). The oligonucleotide probe database. Appl Environ Microbiol 62: 3357-3559.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

Amann RI, Binder BJ, Olson RJ, Chrisholm SW, Devereux R, Stahl DA (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. Appl Environ Microbiol 56: 1919-1925.

Amann R, Ludwig W, Schleifer KH (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Appl Environ Microbiol 59: 143-169.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29: 37–40.

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

Aznar, R., Amaro, C., Garay, E. & Alcaide, E. (1991). Physicochemical and bacteriological parameters in a hypereutrophic lagoon (Albufera lake, Valencia, Spain). Zentbl Mikrobiol. 146: 311-321.

Baer R, Bankier AT, Biggin MD, Deininger PL, Farrel PJ, Gibson TL, Hatfull G, Hudson GS, Satchwell SC, Séguin C, Tuffnell PS, Barrel BG (1984). DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310: 207-211.

Barnes WM (1994). PCR Amplification of up to 35-kb DNA with High Fidelity and High Yield from λ Bacteriophage Templates. PNAS 91: 2216-2220.

Bartlett DH, Silverman M (1989). Nucleotide sequence of IS492, a novel insertion sequence causing variation in extracellular polysaccharide production in the marine bacterium *Pseudomonas atlantica*. J Bacterio.171:1763-6.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002). The Pfam Protein Families Database. Nucleic Acids Research 30:276-280

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002). ARACHNE: a whole-genome shotgun assembler. Genome Res. 12:177-89

Bentley SD, Chater KF, Cerdeno-Tárraga et al. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417: 141-147.

Blinkowa AL, Walker JR. (1990). Programmed ribosomal frameshifting generates the *Escherichia coli* DNA polymerase III gamma subunit from within the tau subunit reading frame. Nucleic Acids Res. 18:1725-9.

Blattner FR (1999). Hot papers - Genetix - The complete genome sequence of *Escherichia coli* K-12 by FR Blattner, G Plunkett, CA Bloch, NT Perna, V Burland, M Riley, J Collado-Vides, JD Glasner, CK Rode, GF Mayhew, J Gregor, NW Davis, HA Kirkpatrick, MA Goeden, DJ Rose, B Mau, Y Shao - Comments. Scientist13:17.

Birren BW, Lai E, Clark SM, Hood L, Simon MI (1988). Optimized conditions for pulsed field gel electrophoretic separations of DNA. Nucleic Acids Res 16:7563-82.

Bodenteich A, Chissoe S, Wang YF, Roe BA (1994). Shotgun Cloning as the Strategy of Choice to Generate Templates for High-throughput Dideoxynucleotide Sequencing. In: H.D. Adams, C. Fields, J.C.Venter (ed.). Automated DNA Sequencing and Analysis. Acad. Press.

Bomar D, Giovannoni S, Stackebrandt E (1988). A unique type of eubacterial 5SrRNA in members of the order Planctomycetales. J Mol Evol 27: 121- 125.

Bond PL, Hugenholz P, Keller J, Blackall LL (1995). Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactors. Appl Environ Microbiol 61: 1910-1916.

Bonfield JK, Smith KF, Staden R (1995). A new DNA sequence assembly programm. Nucl. Acid Res. 24: 4992-4999.

Bonfield JK, Staden R (1996). Experiment files and their application during large-scale sequencing projects. DNA-Sequence 6: 109-117.

Bonfield JK, Rada C, Staden R (1998). Automated detection of point mutations using fluorescent sequence trace substraction. Nucl. Acid Res. 14: 3404-3409.

Borneman J, Skroch PW, O'Sullivan KM, Pallus JA, Rumjanek NG, Jansen JL, Nienhuis J, Triplett EW (1996). Molecular microbial diversity of an agricultural soil in Wisconsin. Appl Environ Microbiol 62: 1935-1943.

Brosius J, Dull TL, Sleeter DD, Noller HF (1981). Gene organization and primary structure of a ribosomal RNA operon from Escherichia coli. J Mol Biol 148: 107-127.

Brassard S, Paquet H, Roy PH. (1995).A transposon-like sequence adjacent to the AccI restriction-modification operon. Gene. 157:69-72.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, Fitzgerald LM, Clayton RA, Gocayne JD et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273: 1058-1073.

Campell VW, Jackson DA (1980). The Effect of Divalent Cations on the Mode of Action of DNase I. Journ. Of Biological Chemistry 8: 3726-3735.

Craig NC (1996). Transposition. In: in *Escherichia coli* and *Salmonella*, Cellular and Molecular Biology (Neidhardt FC Editor in Chief) pp. 2339-2362, American Society for Microbiology, Washington, DC.

The chromosome 21 mapping and sequencing consortium (2000). The DNA sequence of human chromosome 21. Nature 405: 311-319.

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

Cline J, Braman J, Hogrefe HH (1996). PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerase. Nucleic Acids Res. 22: 3259-3260.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornby T, Jagels K, Krogh A, Barrell BG et al. (1998). Deciphering the Biology of Mycobacterium tuberculosis from the complete Genome Sequence. Nature 393:537 ff.

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

Dabrowski S, Kur J (1998). Cloning and Expression in Escherichia coli of the Recombinant His-Tagged DNA Polymerases from *Pyrococcus furiosus* and *Pyrococcus woesei*. Protein Expression and Purification 14: 131-138.

Davies J, Jacob F (1968). Genetic mapping of the regulator and operator genes of the lac operon. J. Mol. Biol. 36: 413-417.

Davis HL, Schleef M, Moritz P, Mancini M, Schorr J, Whalen RG (1996). Partial CviJI Digestion as an Alternative Approach to Generate Cosmid Sublibraries for Large-Scale Sequencing Projects. BioRechniques 21: 99-104.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, Lasken RS (2002). Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci 99:5261-6.

Dear S, Staden R (1991). A sequence assembly and editing program for efficient management of large projects. Nucl. Acid Res. 19: 3907-3911.

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 392:353-8.

DeLong EF, Franks DG, Alldredge AL (1993). Phylogentic diversity of aggregate-attached vs. free living marine bacterial assemblages. Limnol Oceanogr 38: 924-934.

Deininger PL (1983). Random Subcloning of sonicated DNA: Application to shotgun DNA sequence analysis. Anal. Biochem. 129:216-223.

DeLong EF, Franks DG, Alldredge AL (1993). Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemlages. Limnol. Oceanogr. 38:924-934.

Deonier RC (1996). Native Insertion Sequence Elements: Locations, Distributions, and Sequence Relationships. In: in Escherichia coli and Salmonella, Cellular and Molecular Biology (Neidhardt FC Editor in Chief) pp. 2339-2362, American Society for Microbiology, Washington, DC.

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A,Wiezer A, Baumer S, Jacobi C, Bruggemann H, Lienard T, Christmann A, Bomeke M, Steckel S, Bhattacharyya A, Lykidis A, Overbeek R, Klenk HP, Gunsalus RP, Fritz HJ, Gottschalk G (2002). The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol 4:453-61.

Dodd IB, Egan JB (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res. 18:5019-5026.

Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, Erfle H, Caskey CT, Ansorge W (1990). Automated DNA sequencing of the human HPRT locus. Genomics 6:593-608.

Ellenberger T, Landy A, Kwon HJ, Tirumalai R (1997). Flexibility in DNA recombination: structure of the lambda integrase catalytic core. Science 276: 126-131.

Embley TM, Hirt RPO, Williams DM (1994). Biodiversity at the molecular level: the domains, kingdoms and phyla of life. Philos Trans R Soc Lond B Biol Sci. 345: 21-33.

Fiandt M (1998). High Efficiency Packaging of Methylated DNA for Genomic Library Construction using MaxPlax Lambda Packaging Extracts. Epicentre: 14

Fitzgerald MC, Skowron P, Van Etten JL, Smith LM, Mead DA (1992). Rapid shotgun cloning utilizing the two base recognition endonuclease CviJI. Nucleic Acid Res. 14: 3753-3762.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J, Dougherty BA, Merrick JM et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. Science 269:496-512.

Flower AM, McHenry CS (1990). The gamma subunit of DNA polymerase III holoenzyme of *Escherichia coli* is produced by ribosomal frameshifting. Proc Natl Acad Sci USA 87:3713-7.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA, Venter JC (1995). The minimal gene complement of *Mycoplasma genitalium*. Science 270:397-403.

Fraser CM, Fleischmann RD (1997). Strategies for whole microbial genome sequencing and analysis. Electrophoresis 18, 1207-1216.

Frishman D, Mironov A, Mewes HW, Gelfand M (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res. 26:2941-7.

Fuller-Pace FV, Murray NE (1986). Two DNA recognition domains of the specificity polypeptides of a family of type I restriction enzymes. Proc. Natl. Acad. Sci. USA 83:9368-9372.

Fuerst JA (1995). The planctomycetes: emerging models for microbial ecology, evolution and cell biology. Microbiology 141: 1493-1506.

Fuerst JA, Gwilliam HG, Lindsay M, Lichanska A, Beicher C, Vickers JE, Hugenholtz P (1997). Isolation and molecular identification of planctomycete bacteria from post-larvae of the giant tiger prawn, Penaeus monodon. Appl Environ Microbiol 63: 254-262.

Galas DJ, Chandler M (1982). Structure and stability of Tn9-mediated cointegrates. Evidence for two pathways of transposition. J Mol Biol. 154:245-72.

Genetics Computer Group (1991). Program Manual for the GCG Package, Version 7, April, 575 Science Drive, Madison, Wisconsin, USA 53711.

Gimesi N (1924). Hydrobiologiai Tanulmányok [Hydrobiologische Studien]. I. Planctomyces Békefii Gim. Nov. gen. et sp. [Ein neues Glied des Phytoplanktons.] Budapest: Kiadja a Magyar Ciszterci Rend, pp. 1-8. [Hungarian, with German translation.]

Giovannoni SJ, Schabtach E, Castenholz RW (1987). *Isophaera pallida* gen. nov., a gliding budding eubacterium from hot springs. Arch Microbiol 147: 276-284.

Glöckner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzam K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. submitted.

Gopaul DN, Guo F, van Duyne GD (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature 389: 40-46.

Gray JP, Herwig RP (1996). Phylogentic analysis of the bacterial communities in marine sediments. Appl Environ Microbiol 62, 4049-4059.

Green P (1997). Against a Whole-Genome Shotgun. Cold Spring Harbour Lab. Pr. 7: 410-417

Griepenburg U, Ward-Rainey N, Mohamed S, Schlesner H, Marxsen H, Rainey FA, Stackebrandt E, Auling G (1999). Phylogenetic diversity, polyamine pattern and DNA base composition of members of the Order Planctomycetales. International Journal of Systematic Bacteriol. 49: 689-696.

Guilhot C, Gicquel B, Davies J, Martin C. (1992). Isolation and analysis of IS6120, a new insertion sequence from *Mycobacterium smegmatis*. Mol Microbiol. 6:107-13.

Gupta RS, Golding GB (1996). The origin of the eukaryotic cell. Trends Biochem Sci. 21:166-71.

Haren L, Ton-Hoang B, Chandler M (1999). Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol 53:245-81.

Harrison PM, Kumar A, Lang N, Snyder M, Gerstein M (2002). A question of size: the eukaryotic proteome and the problems in defining it. Nucleic Acids Res. 30:1083-90.

Hengen PN (1997). Shearing DNA for genomic library construction. Trends in Biochem. Sci. 22:273-274.

Henke W, Herdel K, Jung K, Schnorr D, Loening SA (1997). Betaine improves the PCR amplification of GC-rich DNA sequences . Nucleic Acid Research 19: 3957-3958.

Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Umayam L, Gill SR, Nelson KE, Read TD, Tettelin H, Richardson D, Ermolaeva MD, Vamathevan J, Bass S, Qin H, Dragoi I, Sellers P, McDonald L, Utterback T, Fleishmann RD, Nierman WC, White O (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477-83.

Holt JG, Krieg NR, Sneath PH, Staley JT, Williams ST (1994). Bergey's Manual of Determinative Bacteriology, 9th edn. Baltimore: Williams & Wilkins.

Hugenholtz P, Goebel BM, Pace NR (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. J. Bacteriol. 180: 4765-4774.

Gebers R, Wehmeyer U, Roggentin T, Schlesner H, Kölbel-Boelke J, Hirsch P (1985). Deoxyribonucleic Acid Base Compositions and Nucleotide Distributions of 65 Strains of Budding Bacteria. International Journal of Systematic Bacteriology 35:260-269.

Henrici AT, Johnson DE (1935). Studies of freshwater bacteria. II. Stalked bacteria, a new order of Schizomycetes. Journal of Bacteriology 30: 61-93.

Horwitz JP, Chua J, Curby RJ, Tomson RJ, DaRooge MA, Fisher BE, Mauricio J, Klundt I, (1964). Substrates for cytochemical demonstration of enzyme activity. Some substituted 3-indolyl-β-D-glycopypyranosides. J. Med. Chem. 7: 574-575.

Huang XC, Maties RA (1994). Application of Capillary Array Electrophoresis to DNA Sequencing. In: H.D. Adams, C. Fields, J.C.Venter (ed.). Automated DNA Sequencing and Analysis. Acad. Press.

Human Genome News (1998). JGI and "Bermuda-Quality" Sequence. HGN 9(3).

International Human genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409:860-941

Janscak P, Bickle TA (1998). The DNA recognition subunit of the type IB restriction-modification enzyme EcoAI tolerates circular permutions of its polypeptide chain. J Mol Biol. 284:937-48.

Jenkins TM, Esposito D, Engelman A, Craigie R (1997). Critical contacts between HIV-1 integrase and viral DNA identified by structure-based analysis and photo-crosslinking. EMBO J. 16:6849-59.

Jenkins C, Fuerst JA (2001). Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain bacteria based on amino acid sequences of elongation factor Tu. J Mol Evol. 52:405-18.

Jenkins C, Kedar V, Fuerst JA. (2002). Gene discovery within the planctomycete division of the domain Bacteria using sequence tags from genomic DNA libraries. Genome Biol. 3:RESEARCH0031.

Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 8:205-13.

Kerger BD, Mancuso CA, Nichols PD, White DC, Langworthy T, Sittig M, Schlesner H, Hirsch P (1988). The budding bacteria, Pirellula and Planctomyces, with atypical 16S rRNA and absence of peptidoglycan, show eubacterial phospholipids and uniquely high proportions of long chain beta-hydroxy fatty acids in the lipopolysaccharide lipid A. Arch Microbiol 149: 255-260.

Kersulyte D, Akopyants NS, Clifton SW, Roe BA, Berg DE (1998). Novel sequence organization and insertion specificity of IS605 and IS606: chimaeric transposable elements of *Helicobacter pylori*. Gene 223:175-86.

Kersulyte D, Mukhopadhyay AK, Shirai M, Nakazawa T, Berg DE (2000). Functional organization and insertion specificity of IS607, a chimeric element of *Helicobacter pylori*. J Bacteriol 182:5300-8.

Klenow H, Henningsen I (1970a). Selective Elimination of Exonuclease Activity of Deoxynucleic Acid Polymerase from *Escherichia coli* B by Limited Proteolysis. Proc. Nat. Acad. Sci. 2: 168-175.

Klenow H, Overgaard-Hansen K (1970b). Proteolytic Cleavage of DNA Polymerase from *Escherichia coli* B into an exonuclease unit an a polymerase unit. FEBS Letters 6: 25-27.

Kunst F (1999) Hot papers - Genetix - The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis* by Kunst, Ogasawara, Moszer, Albertini, Alloni, Azevedo, Bertero, Bessieres, Bolotin, Borchert, Borriss, Boursier, Brans, Braun, Brignell, Bron, Brouillet, Bruschi, Caldwell, Capuano, Carter, Choi, Codani, Connerton, Cummings, Daniel, Denizot, Devine, Dusterhoft, Ehrlich, Emmerson, Entian, Errington, Fabret, Ferrari, Foulger, Fritz, Fujita, Fujita, Fuma, Galizzi, Galleron, Ghim, Glaser, Goffeau, Golightly, Grandi, Guiseppi, Guy, Haga, Haiech, Harwood, Henaut, Hilbert, Holsappel, Hosono, Hullo, Itaya, Jones, Joris, Karamata, Kasahara, Klaerr-Blanchard, Klein, Kobayashi, Koetter, Koningstein, Krogh, Kumano, Kurita, Lapidus, Lardinois, Lauber, Lazarevic, Lee, Levine, Liu, Masuda, Mauel, Medigue, Medina, Mellado, Mizuno, Moestl, Nakai, Noback, Noone, O'Reilly, Ogawa, Ogiwara, Oudega, Park, Parro, Pohl, Portetelle, Porwollik, Prescott, Presecan, Pujic, Purnelle, Rapoport, Rey, Reynolds, Rieger, Rivolta, Rocha, Roche, Rose, Saie, Sato, Scanlan, Schleich, Schroeter, Scoffone, Sekiguchi, Sekowska, Seror, Seror, Shin, Soldo, Sorokin, Tacconi, Takagi, Takahashi, Takemura, Takeuchi, Tamakoshi, Tanaka, Terpstra, Tognoni, Tosato, Uchiyama, Vandenbol, Vannier, Vassarotti, Viari, Wambutt, Wedler, Wedler, Weitzenegger, Winters, Wipat, Yamamoto, Yamane, Yasumoto, Yata, Yoshida, Yoshikawa, Zumstein, Yoshikawa, Danchin. Scientist. 13:14.

Kölbel-Boelke J, Gebers R, Hirsch P (1985). Genome Size Determinations for 33 Strains of Budding Bacteria. International Journal of Systematic Bacteriology 35: 270-273.

König H, Schlesner H, Hirsch P (1984). Cell wall studies on budding bacteria of the Planctomyces/Pasteuria group and on a Prosthecomicrobium sp. Arch Microbiol 138: 200-205.

Lake JA, Rivera MC (1994). Was the nucleus the first endosymbiont? Proc Natl Acad Sci USA 91:2880-1.

Lee SY, Bollinger J, Bezdicek D, Ogram A (1996). Estimation of the abundance of an uncultured soil bacterial strain by a competitive quantitative PCR method. Appl Environ Microbiol 62: 3787-3793.

Liesack W, König H, Schlesner H, Hirsch P (1986). Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the Pirella/Planctomyces group. Arch. Microbiol 145: 361-366.

Liesack W, Hopfl P, Stackebrandt E (1988). Complete nucleotide sequence of a 23S ribosomal RNA gene from *Pirellula marina*. Nucleic Acids Res. 16: 5194.

Liesack W, Stackebrandt E (1989). Evidence of unlinked rrn operons in the planctomycete *Pirellula marina*. J. Bacteriol 171: 5025-5030.

Liesack W, Stackebrandt E (1992). Occurence of novel groups of the domain Bacteria as revealed by analysis of genetic material isolated from Australian terrestrial environment. J Bacteriol 174: 5072-5078.

Liesack W, Soeller R, Stewart T, Haas H, Giovannoni S, Stackebrandt E (1992b). The influence of tachytelically (rapidly) evolving sequences on the topology of phylogenetic trees- intrafamily relationships and phylogenetic position of Planctomycetaceae as revealed by comparative analysis of 16S ribosomal RNA sequences. Syst. Appl. Microbiol. 15: 357-362

Lindsay MR, Webb RI, Fuerst JA (1997). Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus Pirellula. Microbiology 143: 739-748.

Lindsay MR, Webb RI, Strous M, Jetten MSM, Butler MK, Forde RJ, Fuerst JA (2001). Cell compartmenttalisation in planctomycetes: novel types of structural organisation for the bacterial cell. Arch. Microbiol. 175: 413-429.

Llobet-Brossa E, Rossellò-Mora R, Amann R (1998). Microbial community composition of wadden sea sediments as revealed by fluorescence in situ hybridization. Appl. & Environ. Microbiol. 64: 2691-2696.

Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J (1996). Molekulare Zellbiologie. 2. Aufl. Walter de Gruyter, Berlin.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J (2001). Molekulare Zellbiologie. 4. Aufl. Walter de Gruyter, Berlin.

Lowe TM, Eddy SR (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. Mar 1;25(5):955-64.

Machida C, Machida Y (1989). Regulation of IS1 transposition by the insA gene product. J Mol Biol. 208:567-74.

Mahillon J, Chandler M (1998). Insertion sequences. Microbiol Mol Biol Rev. 62:725-74.

Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, Daly MJ (2001).Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. Microbiol Mol Biol Rev. 65:44-79.

Martin-Gallardo A, Lamerdin J, Carrano A (1994). Shotgun Sequencing. In: H.D. Adams, C. Fields, J.C.Venter (ed.). Automated DNA Sequencing and Analysis. Acad. Press.

Menke MAOH, Liesack W, Stackebrandt E (1991). Ribotyping of 16S and 23S rRNA genes and organisation of rrn operons in members of the bacterial genera Gemmata , Planctomyces, Thermotoga, Thermus and Verrucomicrobium. Arch Microbiol 155:263-271

Messer W (1999). DNA, Chromosomes and Plasmids. In: Biology of the Prokaryotes. Lengeler J.W., G.Drews, H.G.Schlegel (eds.), Thieme Stuttgart.

Messing JR, Crea P, Seeburg H (1981). A system for shotgun DNA sequencing. Nucleic Acid Res. 2: 309-321

Neef A, Amann R, Schlesner H, Schleifer KH (1998). Monitoring a widespread bacterial group: in situ detection of planctomycetes with 16s rRNA- targeted probes. Microbiology 144: 3257-3266.

Neumann E, Rosenheck K (1972). Permeability changes indused by electric impulses in vesicular membranes. J. Membr. Biol. 10: 279-290

Neumann E, Schäfer-Ridder M, Wang Y, Hofschneider PH (1982). Gene transfer into mouse lyoma cells by electroporation in high electric fields. EMBO J. 1: 841-845

Nordhoff E, Lübbert C, Thiele G, Heiser V, Lehrach H (2000). Rapid determination of short DNA sequences by the use of MALDI-MS. Nucleic Acid Research 28: e86

Oefner PJ, Hunicke-Smith SP, Chiang L, Dietrich F, Mulligan J, Davis RW (1996). Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. Nucleic Acid Research 20: 3879-3886.

Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H, Nakazawa H, Osonoe T, Kikuchi H, Shiba T, Sakaki Y, Hattori M (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci USA 98:12215-20.

P/N 4390037. ABI PRISM BigDye Terminator v3.0 Ready Reaction Cycle Sequencing Kit. http://docs.appliedbiosystems.com/pebiodocs/04390037.pdf

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, Jagels K, Karlyshev AV, Moule S, Pallen MJ, Penn CW, Quail MA, Rajandream MA, Rutherford KM, van Vliet AH, Whitehead S, Barrell BG (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665-8.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T,Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature 413:848-52.

Polard P, Prere MF, Chandler M, Fayet O. (1991). Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911.J Mol Biol. 222: 465-77.

Prère MF, Chandler M, Fayet O (1990). Transposition in Shigella dysenteriae: isolation and analysis of IS911, a new member of the IS3 group of insertion sequences. J Bacteriol 172: 4090-9.

Rabus R, Gade D, Helbig R, Bauer M, Glöckner FO, Kube M, Schlesner H, Reinhardt R, Amann R (2002a). Analysis of N-acetylglucosamine metabolism in the marine bacterium *Pirellula* sp. strain 1 by a proteomic approach. Proteomics 2, No. 6, 649-655

Rabus R, Kube M, Beck A, Widdel F, Reinhardt R (2002b). Genes involved in the anaerobic degradation of ethylbenzene in a denitrifying bacterium, strain EbN1. Arch Microbiol. 178:506-16.

Rabussay D, Uher L, Bates G, Piastuch W (1987). Electroporation of mammalian and plant cells. Focus (Life Technologies) 9:1-3.

Radelof U, Hennig S, Seranski Z, Steinfath M, Ramser J, Reinhardt R, Poustka A, Francis F, Lehrach H (1998). Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large scale sequencing projects. Nucleic Acis Research 26: 5358-5364.

Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, Bass S, Linher K, Weidman J, Khouri H, Craven B, Bowman C, Dodson R, Gwinn M, Nelson W, DeBoy R, Kolonay J, McClarty G, Salzberg SL, Eisen J, Fraser CM (2000). Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. 28:1397-406.

Redaschi N, Bickle TA (1996). DNA Restriction and Modification Systems. IN: Neidhardt, Frederick, ed. *Escherichia coli* & *Salmonella* ed. 2, p. 773-781, American Society Microbiology.

Rettberg CC, Prere MF, Gesteland RF, Atkins JF, Fayet O (1999). A three-way junction and constituent stem-loops as the stimulator for programmed -1 frameshifting in bacterial insertion sequence IS911. J Mol Biol. 286:1365-78.

Rieder MJ, Taylor SL, Tobe VO, Nickerson DA (1998). Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. Nucl. Acid Res. 4: 967-973.

Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV (2002).Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. 30:4264-71.

Romero D, Palacios R (1997). Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet. 31:91-111. Review.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988). Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. Science, Vol. 239: 487-491.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. 74:5463-5467.

Sanger F, Coulson AR, Barrell BG, Smith AJH, Roe BA (1980). Cloning in Single-stranded Bacteriophage as an Aid to Rapid DNA Sequencing. J. Mol. Biol. 143: 161-178

Sanger F, Coulson AR, Hong GF, Hill DF, Peterson GB (1982). Nucleotide sequence of bacteriophage λ DNA. J. Mol. Biol. 162:729-773.

Schleifer KH, Ludwig W (1989). Phylogenetic relationships among bacteria. In: Fernholm B, Bremer K, Jörnvall H (eds). The hierarchy of life. Elsevier Science, Amsterdam, pp. 103-116.

Schlesner H, Hirsch P (1984). Assignment of ATCC 25377 to Pirella gen. nov. as *Pirella staleyi* comb. Int J system Bact 34: 492- 495.

Schlesner H (1986). *Pirellula marina* sp. nov., a budding, peptidoglycan-less bacterium from brackish water. Syst Appl Microbiol 8: 177-180.

Schlesner, H. & Stackebrandt, E. (1986). Assignment of the genera Planctomyces and Pirella to a new family Planctomycetaceae fam. nov. and description of the order Planctomycetales ord. nov. Syst Appl. Microbiol 8: 174-176.

Schlesner, H. (1994). The development of media suitable for the microorganisms morphologically resembling Planctomyces spp., Pirellula spp., and other Planctomycetales from various aquatic habitats using dilute media. Syst Appl Microbiol 17: 135-145.

Schmidt JM (1978). Isolation and ultrastructure of freshwater strains of Planctomyces. Curr Microbiol 1: 65-70.

Schmidt JM, Starr MP (1978). Morphological Diversity of Freshwater Bacteria Belonging to the Blastocaulis-Planctomyces Group as Observed in Natural Populations and Enrichments. Current Microbiology, 1: 325-330.

Schouler C, Clier F, Lerayer AL, Ehrlich SD, Chopin MC (1998a). A type IC restriction-modification system in *Lactococcus lactis*. J Bacteriol. 180:407-11.

Schouler C, Gautier M, Ehrlich SD, Chopin MC (1998b). Combinational variation of restriction modification specificities in *Lactococcus lactis*. Mol Microbiol. 28:169-78.

Schriefer LA, Gebauer BK, Qiu LQQ, Waterston RH, Wilson RK (1990). Low pressure DNA shearing: amethod for random DNA sequence analysis. Nucleic Acid Res. 24: 7455-7456.

Sekine Y, Ohtsubo E (1989). Frameshifting is required for production of the transposase encoded by insertion sequence 1. Proc Natl Acad Sci USA. 86:4609-13.

Sekine Y, Nagasawa H, Ohtsubo E (1992). Identification of the site of translational frameshifting required for production of the transposase encoded by insertion sequence IS 1.Mol Gen Genet. 235:317-24.

Sekine Y, Eisaki N, Ohtsubo E (1994). Translational control in production of transposase and in transposition of insertion sequence IS3. J Mol Biol. 235:1406-20.

Seo HC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, Chourrout D (2001). Miniature Genome in the Marine Chordate *Oikopleura dioica.* Science 294: 2506

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407:81-6.

Shirai M, Hirakawa H, Kimoto M, Tabuchi M, Kishi F, Ouchi K, Shiba T, Ishii K, Hattori M, Kuhara S, Nakazawa T (2000). Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucleic Acids Res. 28:2311-4.

Stackebrandt E, Ludwig W, Schubert W, Klink F, Schlesner H, Roggentin T, Hirsch P. (1984). Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less eubacteria. Nature 307:735-7.

Stackebrandt E, Wehmeyer U, Liesack W (1986). 16S ribosomal RNA- and cell wall analysis of *Gemmata obscuriglobus*, a new member of the order Planctomycetales. FEMS Microbiology Letters 37: 289-292.

Stackebrandt E, Ludwig W, Schubert W, Klink F, Schlesner H, Roggentin T, Hirsch P (1984). Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less eubacteria. Nature 307: 735-737.

Staden R, Beal KF, Bonfield JK (1999). The Staden Package, 1998. Computer Methods in Molecular Biology. In: Misener S, Krawetz SA (eds). Bioinformatics Methods and Protocols. The Humana Press Inc., pp 115–130

Staley JT, Fuerst JA, Giovannoni S, Schlesner H (1992). The order Planctomycetales and the genera Planctomyces, Pirellula, Gemmata and Isophaera. In: The Prokaryotes, 2nd edn, vol, IV, pp. 3710-3731. Edited by A. Balows, H.G. Truper, M. Dworkin, W. Harder & K.H. Schleifer. New York:Springer.

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV, Davis RW (1998). Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754-9.

Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FSL, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GKS, Wu Z, Olson MV et al. (2000). Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature. 406:959-964

Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K (2000). Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with Bacillus subtilis. Nucleic Acids Res. 28:4317-31.

Tatusov RL, Koonin EV, Lipman DJ (1997). A genomic perspective on protein families. Science 278:631-7. Review.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22-8

Timmerman KP, Tu CP. Complete sequence of IS3 (1995). Nucleic Acids Res. 13:2127-39.

Thorpe PH, Ternent D, Murray NE (1997). The specificity of sty SKI, a type I restriction enzyme, implies a structure with rotational symmetry. Nucleic Acids Res. 25:1694-700.

Tsuchihashi Z, Kornberg A. Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme (1990). Proc Natl Acad Sci USA 87:2516-20.

Ullmann A., Perrin D. (1970). Complementation in β-galactosidase. In The lactose operon (ed. J.R. Beckwith and D. Zipser), pp. 143-172. Cold Spring Harbour Laboratory, Cold Spring Harbour, New York.

Van Den Eynde, Van de Peer Y, Perry J, De Wachter R (1990). 5S rRNA sequences of representatives of the genera Chlorobium, Prosthecochloris, Thermomicrobium, Cytophaga, Flavobacterium, Flexibacter and Saprospira and a discussion of the evolution of eubacteria in general. J Gen Microbiol 136:11-18

Van de Peer Y, Neefs J, De Rijk P, De Vos P, De Wachter R (1994). About the order of divergence of the major bacterial taxa during evolution. Syst Apll Microbiol 17:32-38.

Vögele K, Schwartz E, Welz C, Schiltz E, Rak B (1994). High-level ribosomal frameshifting directs the synthesis of IS150 gene products. Nucleic Acids Res. 19:4377-85.

Vogelstein B, Gillespie D (1979). Preparative and analytical purification of DNA from agarose. Proc. Natl. Acad. Sci. 2: 615-619.

Wang CY, Bond VC, Genco CA. (1997). Identification of a second endogenous Porphyromonas gingivalis insertion element. J Bacteriol. 179:3808-12.

Ward N, Rainey FA, Stackebrandt E, Schlesner H (1995). Unraveling the extend of diversity within the order Planctomycetales. Appl Environ Microbiol 61: 2270-2275.

Weber JL, Myers EW (1997). Human Whole-Genome Shotgun Sequencing. Cold Spring Harbour Press, Vol.7 : 401-409

Weisburg WG, Hatch TP, Woese CR (1986). Eubacterial origin of Chlamydiae. J Bacteriol. 167:570-4.

Woese CR (1987). Bacterial evolution. Microbiol Rev 51:221-271.

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, Collins M, Connor R, Cronin A, Davis P, Feltwell T, Fraser A, Gentles S, Goble A, Hamlin N, Harris D, Hidalgo J, Hodgson G, Holroyd S, Hornsby T, Howarth S, Huckle EJ, Hunt S, Jagels K, James K, Jones L, Jones M, Leather S, McDonald S, McLean J, Mooney P, Moule S, Mungall K, Murphy L, Niblett D, Odell C, Oliver K, O'Neil S, Pearson D, Quail MA, Rabbinowitsch E, Rutherford K, Rutter S, Saunders D, Seeger K, Sharp S, Skelton J, Simmonds M, Squares R, Squares S, Stevens K, Taylor K, Taylor RG, Tivey A, Walsh S, Warren T, Whitehead S, Woodward J, Volckaert G, Aert R, Robben J, Grymonprez B, Weltjens I, Vanstreels E, Rieger M, Schafer M, Muller-Auer S, Gabel C, Fuchs M, Fritzc C, Holzer E, Moestl D, Hilbert H, Borzym K, Langer I, Beck A, Lehrach H, Reinhardt R, Pohl TM, Eger P, Zimmermann W, Wedler H, Wambutt R, Purnelle B, Goffeau A, Cadieu E, Dreano S, Gloux S, Lelaure V, Mottier S, Galibert F, Aves SJ, Xiang Z, Hunt C, Moore K, Hurst SM, Lucas M, Rochet M, Gaillardin C, Tallada VA, Garzon A, Thode G, Daga RR, Cruzado L, Jimenez J, Sanchez M, del Rey F, Benito J, Dominguez A, Revuelta JL, Moreno S, Armstrong J, Forsburg SL, Cerrutti L, Lowe T, McCombie WR, Paulsen I, Potashkin J, Shpakovski GV, Ussery D, Barrell BG, Nurse P (2002). The genome sequence of *Schizosaccharomyces pombe*. Nature 415:871-80.

Wong TK, Neumann E (1982). Electric field mediated gene transfer. Biochem. Biophys. Res. Commun. 107: 584-587

Xu K, He ZQ, Mao YM, Sheng RQ, Sheng ZJ. (1993). On two transposable elements from Bacillus stearothermophilus. Plasmid 29:1-9.

Yanisch-Perron, C., Vieira, J., Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene 33: 103-119.

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296:79-92.

Yeh RF, Lim LP, Burge CB (2001). Computational inference of homologous gene structures in the human genome. Genome Res. 11:803-16.

7. Anhang

7.1 Abkürzungen

aa:	*amino acids*
Abb.:	Abbildung
b:	*bases*
bp:	*base pairs*
HTS:	*high throughput system*
IS-Element:	*insertion sequence element*
Kap.:	Kapitel
kb:	Kilobasen
mb:	Megabasen
MCS:	*multiple cloning site*
MPI:	Max Planck Institut
MTP(s):	*microtiterplate(s)*
na:	*nucleotide acids*
ORF(s):	*open readings frame(s)*
PCR:	*polymerase chain reaction*
rpm:	*revolutions per minute*
Tab.:	Tabelle
TM:	*melting temperatur*
vgl.:	vergleiche

Abkürzungen für Länder:

D:	Deutschland
F:	Frankreich
USA:	United States of America
UK:	United Kingdom

7.2 Veröffentlichungen unter Hervorhebung der eigenen Beiträge

Seo HC, Kube M, Edvardsen RB, Jensen MF, Beck A, Spriet E, Gorsky G, Thompson EM, Lehrach H, Reinhardt R, Chourrout D (2001).
Miniature genome in the marine chordate *Oikopleura dioica*. Science 294:2506.
Verantwortlich für die DNA-Isolierung, Erstellung von Banken im Rahmen eines whole genome shotguns, Sequenzierung und Assemblierung. Durchführung der statistischen Abschätzung der Genomgröße des Urochordatengenoms als Teil der Publikation auf der Basis der assemblierten Sequenz, die die Hälfte des Genoms (~ 30 Mb) repräsentiert.

Rabus R, Gade D, Helbig R, Bauer M, Glockner FO, Kube M, Schlesner H, Reinhardt R, Amann R (2002). Analysis of *N*-acetylglucosamine metabolism in the marine bacterium *Pirellula* sp. strain 1 by a proteomic approach. Proteomics 2:649-55.
Verantwortlich für die Erstellung von Shotgun Banken, Sequenzierung, Assemblierung im Rahmen der vorliegenden Arbeit zur Dissertation und Mitarbeit an der Genidentifikation.

Rabus R, Kube M, Beck A, Widdel F, Reinhardt R (2002). Genes involved in the anaerobic degradation of ethylbenzene in a denitrifying bacterium, strain EbN1. Arch Microbiol. 178:506-16.
Verantwortlich für die DNA-Isolierung, Erstellung von Banken im Rahmen eines whole genome shotguns, Sequenzierung, Assemblierung, Identifikation und Interpretation von Schlüsselgenen.

Glöckner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzym K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R.
Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1.
Verantwortlich für die Erstellung von Shotgun Banken, der Sequenzierung, der Assemblierung, der Rohannotation und Detailanalysen zur Absicherung der Sequenz im Rahmen der vorliegenden Arbeit zur Dissertation. Das Manuskript wurde bei Proceedings of the National Academy of Sciences eingereicht.

Kube M, Heider J, Hufnagel P, Kühner S, Beck A, Widdel F, Reinhardt R, Rabus R.
Genes involved in the anaerobic degradation of toluene in a denitrifying bacterium, strain EbN1.
Verantwortlich für die Erstellung von Shotgun Banken, der Sequenzierung, Assemblierung, Identifikation und Interpretation von Schlüsselgenen im ausgewählten Datensatz. Das Manuskript wurde bei Archives of Microbiology eingereicht.

Schübbe S, Kube M, Scheffel A, Wawer C, Heyen U, Meyerdierks A, Madkour MH, Mayer F, Reinhardt R, Schüler D (2003). Characterization of a spontaneous nonmagnetic mutant of *Magnetospirillum gryphiswaldense* reveals a large deletion comprising a putative magnetosome island.
Verantwortlich für die Erstellung von Shotgun Banken, Subklonierungen, der Sequenzierung, der Assemblierung, der Detailanalysen zur Absicherung der Sequenz und Einreichung der Daten (EBI). Das Manuskript wurde bei Journal of Bacteriology eingereicht.

7.3 Zusätzliche Materialien

Tab. 71: Repetitive Elemente im Genom von Pirellula

Repetitives Element	Linke Position	Rechte Position	Strang	Länge	Anzahl	Potenzielle Funktion
A1	1152512	1154071	C	1560	10	IS-Element
A2	3076194	3077747	+	1554		
A3	3150066	3151625	C	1560		
A4	3526927	3528486	+	1560		
A5	3577371	3578930	C	1560		
A6	4478545	4480104	+	1560		
A7	5338107	5339666	+	1560		
A8	6593256	6594815	+	1560		
A9	6665725	6667284	C	1560		
A10	7081687	7083246	+	1560		
B1	2214462	2215774	+	1313	9	IS-Element
B2	378066	379378	+	1313		
B3	606767	608079	+	1313		
B4	6327993	6329305	+	1313		
B5	6361879	6363191	+	1313		
B6	6966175	6967487	+	1313		
B7	89262	90574	+	1313		
B8	3086395	3087707	C	1313		
B9	3562905	3564217	C	1313		
C1	6262359	6266816	+	4458	2	Restriktions-Modifikations-System
C2	6277184	6281626	+	4443		
D1	6134145	6134714	+	570	5	IS-Element
D2	2001397	2003155	C	1759		
D3	2465513	2467285	C	1773		
D4	5048321	5050094	C	1774		
D5	5282867	5284639	C	1773		
E1	2471136	2472609	+	1474	3	IS-Element
E2	6127195	6128668	C	1474		
E3	2216506	2217979	+	1474		
F1	2764390	2765900	+	1511	4	IS-Element
F2	2219810	2221320	+	1511		
F3	2226753	2228263	C	1513		
F4	5759790	5761300	C	1511		
G1	3234894	3237419	+	2526	5	Replikatives Transposon
G2	3952548	3955072	+	2525		
G3	489670	492194	C	2525		
G4	503410	505917	C	2508		
G5	492721	494365	C	1645		
H1	3237782	3239210	+	1430	3	Integrase/Rekombinase
H2	3957035	3958463	+	1429		
H3	487599	488738	C	1142		
I1	2233884	2235258	+	1375	5	IS-Element
I2	6925591	6926965	+	1375		
I3	2773864	2775238	+	1375		
I4	6022188	6023562	C	1375		
I5	4936599	4937973	C	1375		
J1	605571	606768	+	1198	6	Replikatives Transposon
J2	608078	609533	+	1456		
J3	2061863	2064514	+	2652		
J4	1335844	1338494	C	2651		
J5	1621702	1624352	C	2651		
J6	1214863	1215530	C	668		
K1	1197002	1198827	C	1826	3	IS-Element
K2	1791898	1793698	C	1801		
K3	5445550	5447275	+	1726		
L1	427836	429141	C	1306	5	IS-Element
L2	654547	655852	C	1306		
L3	1294155	1295460	+	1306		
L4	2106160	2107465	+	1306		
L5	2116481	2117786	C	1306		
M1	87515	88475	+	1281	2	IS-Element
M2	429873	431152	C	1280		
				∑ der Kopien: 62		

Tab. 72: Verteilung der tRNAs im Genom von *Pirellula* sp. Stamm 1

	tRNA		Anti	Intron Bounds	
Nummer	Start	Ende	Type	Codon	Score
1	15141	15217	Arg	TCT	85.59
2	86868	86941	Val	CAC	78.76
3	133023	133099	Val	TAC	86.79
4	231979	232052	Asp	GTC	84.19
5	468301	468382	Ser	CGA	38.88
6	574208	574291	Leu	CAA	66.67
7	790270	790354	Leu	CAG	67.89
8	1034374	1034447	Thr	CGT	70.78
9	1124293	1124381	Ser	GCT	70.78
10	1124509	1124581	Lys	TTT	82.99
11	1209090	1209161	Gln	TTG	52.03
12	2104395	2104468	Arg	ACG	73.21
13	2156459	2156532	His	GTG	79.60
14	2823426	2823502	Arg	CCT	75.25
15	3163724	3163810	Ser	TGA	55.02
16	3188257	3188330	Pro	TGG	76.55
17	3370937	3371021	Leu	TAA	63.83
18	4216684	4216756	Thr	TGT	81.25
19	4216923	4217004	Tyr	GTA	65.65
20	4217122	4217192	Gly	TCC	74.03
21	4217195	4217269	Thr	GGT	75.24
22	4218751	4218823	Trp	CCA	73.51
23	4355716	4355788	Glu	TTC	56.70
24	4667213	4667295	Ser	GGA	45.48
25	5880607	5880681	Pseudo	CAT	24.23
26	6260778	6260850	Cys	GCA	74.01
27	6260960	6261032	Lys	CTT	82.95
28	6578179	6578252	Val	GAC	83.33
29	6988936	6989010	Pro	GGG	74.82
30	6238082	6238009	Met	CAT	73.09
31	6201871	6201799	Pro	TGG	47.08
32	6201631	6201559	Lys	TTT	60.16
33	6201547	6201475	Ile	TAT	72.90
34	6201397	6201325	Lys	CTT	50.52
35	6201304	6201233	Ala	CGC	30.18
36	6201180	6201110	Gly	GCC	47.00
37	6201101	6201030	Asn	GTT	56.40
38	6200944	6200873	Gly	TCC	45.89
39	6200766	6200692	Gln	CTG	53.85
40	6200675	6200602	Pseudo	GTT	37.89
41	6200329	6200255	Glu	TTC	43.34
42	6200026	6199956	Trp	CCA	42.37
43	6199936	6199865	Val	TAC	49.17
44	6199480	6199407	Leu	TAA	40.38
45	6199388	6199318	Ala	TGC	40.25
46	6199312	6199241	Leu	TAG	51.01
47	6199228	6199155	Leu	CAA	39.47
48	6199141	6199058	Leu	CAG	54.67
49	6198884	6198809	Leu	GAG	39.53
50	6198798	6198726	Val	GAC	47.52
51	6198714	6198644	Pseudo	GGT	25.97
52	6198444	6198373	Arg	TCT	48.31
53	6198249	6198178	Pseudo	TCG	28.23
54	6198169	6198097	Pseudo	GCG	24.93
55	5858464	5858391	Arg	CCG	70.81
56	5811682	5811610	Phe	GAA	82.98
57	5343101	5343020	Leu	TAG	63.77
58	5267412	5267340	Asn	GTT	79.29
59	4745616	4745532	Leu	GAG	62.60
60	4615352	4615279	Ile	GAT	84.64
61	4615201	4615129	Ala	TGC	82.89
62	3655905	3655829	Met	CAT	83.32
63	3088138	3088066	Ala	GGC	74.80
64	2965068	2964996	Gly	GCC	83.01
65	2481660	2481587	Arg	TCG	75.27
66	1916814	1916738	Pro	CGG	61.04
67	1636151	1636075	Met	CAT	77.91
68	427552	427480	Pro	TGG	46.00
69	427215	427143	Ile	TAT	57.92
70	427117	427045	Lys	CTT	50.52
71	426924	426854	Gly	GCC	47.00
72	426845	426774	Asn	GTT	56.40
73	426504	426431	Pseudo	GTG	22.87
74	426236	426162	Glu	TTC	47.24
75	425742	425672	Ala	TGC	39.71
76	425655	425584	Leu	TAG	51.01

Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide
In der Tabelle werden für jeden Primer, der zur Sequenzierung verwendet wurde, die Arbeitsbezeichnung, die Lokalisation auf dem Genom und die Strangorientierung auf der genomischen Sequenz angegeben. Im Kommentarfeld ist im Falle der Verwendung der Primer zur Absicherung der repetitiven Elemente das jeweilige Element vermerkt. Die Primer, die zum Schließen der *Physical Gaps* verwendet wurden, sind ebenfalls gekennzeichnet. Alle Primer sind nach den Positionen auf dem Genom sortiert.

Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir2_42	71178	71202	+	gagttggattggagaaggctcatgg	B7
pir2_07	89191	89215	+	cgagaagaacagtcgctttacagtg	B7
pir2_08	90673	90649	-	gacagtgtccgtctttcgagctttg	B7
pir2_41	98628	98604	-	tcgacattgggaaacaactcggttc	
pir2_40	130131	130155	+	gtacttgcagcactttgccgtgcag	
pir2_39	138388	138412	+	tcgtgctggcacatttggaataccc	
pir2_38	316957	316933	-	tgagacgtgcattcaacaactcggg	
pir2_03	324081	324105	+	atcacgcccgacctgcgattgttcc	*Physical Gap*
pir2_05	324592	324615	+	ggcgacaatttgaaggcacgctgg	
pir2_52	325316	325340	+	tacaacggaacgacggtcgatcaag	
pir2_48	325466	325442	-	cccaggacaaactgttctggtacac	
pir2_51	326000	325976	-	tcaggaagtccgcaaacagcgaaac	
pir2_02	326047	326024	-	atgatcgagcaatgtcccgtcgcc	*Physical Gap*
pir1_37	329984	329960	-	aaagatgtcggtgcagtaaccttcc	
pir2_36	362562	362586	+	cttctcgatattgccagctcgattc	
pir2_09	377944	377968	+	gagcgttgattctcagtccttgttg	B2
pir2_10	379561	379537	-	aagacgaacggagtgtcggtgttac	B2
pir2_35	392810	392786	-	gagttcaaagacgttcagcttcgag	
pir2_47	406411	406387	-	ttgagcgagcattgaaacgagtcgg	
pir2_34	419156	419180	+	tcggtctcgatgtttggtcgtgaac	
pir2_11	427749	427773	+	aaaaggctcgctgtccgtatgacac	L1
pir2_12	429224	429200	-	gactcgaacagaaacacgaaggcag	L1
pir2_49	438198	438174	-	actaccagatcgacgtggaaatccc	
pir2_33	475392	475416	+	attcacatccaacgaactgatcgtc	
pir2_13	487463	487487	+	aaaacgcctgtgttcccaacgcaag	H3
pir2_14	488851	488827	-	gacatcgtcggagctttagtgtcag	H3
pir2_15	489544	489568	+	tcatcaacggcaagggtggaaactg	G3
pir2_16	492265	492241	-	ctgaacgaagcgaacgatgtctgag	G3
pir2_17	492623	492647	+	cgatgagacttacgaggcagatttc	G5
pir2_18	494521	494497	-	actcctgcgtatctgcttaatcccg	G5
pir2_19	503296	503320	+	actcggacatgcctgttgacattag	G4
pir2_46	504051	504075	+	cgcatctgcatcggcttgtattgac	G4
pir2_50	505039	505015	-	caacctcaaggttctgcaaggctac	G4
pir2_45	505757	505733	-	caagcggtttcgagaagacttgctg	G4
pir2_20	506118	506094	-	aatggtcccttcgattcgttggatg	G4
pir2_32	516183	516159	-	caaatgcagcgtcgcattccgaatc	
pir2_31	530027	530051	+	gagattggcatgaggttgctgcatc	
pir2_30	530603	530579	-	cccaaggtcgaagtcatcgaagaag	
pir2_29	540356	540380	+	acgatgcttgcgatgtgttcgttgg	

Fortsetzung Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide

Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir2_28	552934	552910	-	cgattgctgggatcaacggaaagcc	
pir2_27	565324	565300	-	ttttcgcatcgtcacgtagccatcg	
pir2_21	605424	605448	+	tctcgttgacctgatcgtgagatcg	J1/J2
pir4_71	607254	607277	+	caaacccaaatcattcaagccgtc	B3
pir2_22	609640	609613	-	gcattaattatctcaattcgatctcatc	J1/J2
pir2_25	614460	614484	+	ttcaagcgttcgatgcagtagtcgg	
pir2_26	615244	615220	-	gtactcatctacaatctgagcctcc	
pir2_23	654473	654497	+	tcacggcgatgagaagacttctaag	l2
pir2_24	655972	655948	-	tgcagacattatcggagtttgccag	l2
pir2_43	680542	680566	+	ctttcgtccaccaaaatcgacttcc	
pir2_01	680772	680795	+	gatcgtagcgagacaagacttcac	
pir4_104	681118	681139	+	cgtgccgtttccgttgcgatcc	
pir4_106	681149	681128	-	tgcaacgattggatcgcaacgg	
pir4_105	681411	681390	-	ctgctcatccccgctcttgccc	
pir4_links	681664	681643	-	catgtcctgtttccttgacgag	
pir4_81	854658	854679	+	atcgtgaaggattgcgatgccg	
pir4_82	881138	881159	+	cgtttgcgttggtccgttcatc	
pir4_83	881739	881718	-	agtcggatagcacgtagtagcc	
pir4_01	1036759	1036782	+	cgatttgttcgattgtgtgatgcc	
pir4_84	1037052	1037073	+	atgcaaagctggcgggctgggg	
pir4_02	1037351	1037330	-	cgaagccaagaacgaagaacgc	
pir4_85	1045177	1045156	-	cctcttccaaaccgttgaagcc	
pir4_86	1118174	1118195	+	gcagcaacttggtcgggatggg	
pir4_21	1119067	1119089	+	cattcttggtcatcggattgctg	
pir4_22	1119402	1119381	-	ccaatgtaaccgactcgacgtg	
pir4_15	1152413	1152434	+	tcgttttcgtccaccgtcgttc	a1
pir4_17	1152803	1152824	+	cagatagctcaacgacgtgatg	a1
pir4_18	1153607	1153586	-	caagcaggagaactgtatcgtc	a1
pir4_16	1154184	1154163	-	gtagccaagcgtaccgataatc	a1
pir4_87	1182261	1182283	+	ggtatcttgaatttcgcaccgac	
pir4_88	1182641	1182619	-	ctacgatcatcaaaacgatcccg	
pir4_62	1196914	1196935	+	gctcccagcgtgatggaacggc	k1
pir4_63	1199060	1199038	-	aggcgttgccccaggcgaggttg	k1
pir4_19	1294097	1294118	+	atggaatcccgcacatagtctg	l3
pir4_20	1295566	1295545	-	tggtgctcgattgggaggatgc	l3
pir4_89	1304239	1304260	+	gagcttatcgcacttctactcg	
pir4_90	1304564	1304541	-	gcgaaattcgagtttcaagttgcc	
pir4_23	1335644	1335668	+	caaaatcaatggatcaacggagacc	j4
pir4_111	1336160	1336181	+	gtgaacgagcagcatgtgcgac	j4
pir4_122	1336774	1336795	+	agcgggcgtgggactgaaggtg	j4
pir4_123	1337501	1337480	-	aagttcttcggtgtgctctgcc	j4
pir4_110	1338158	1338137	-	atcgcccccattcgcccttctg	j4
pir4_24	1338683	1338661	-	ggtgatgctccttgctggttctcg	j4
pir4_91	1341109	1341088	-	cgacatcggtcgccgcatctac	
pir4_92	1341871	1341848	-	ggcaaacgattcattcaactgcac	
pir4_03	1412728	1412749	+	ggggtatgtcggcaagcacgag	
pir4_04	1413071	1413050	-	ttgcgtccgctccgctggcaac	
pir4_120	1413568	1413589	+	gcagtgccaaatcccaacgacc	
pir4_121	1414013	1413992	-	cgatgtatgaatcggcgatggc	
pir4_64	1622106	1622128	+	gcatcaagttcttcttcacgcac	J5

Fortsetzung Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide

Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir4_65	1623880	1623859	-	tgacctcgatcactcgcatcac	J5
pir4_05	1746412	1746433	+	gccagttccattgttagtccag	
pir4_06	1746675	1746654	-	cgtctgtaccgcaacaatggtg	
pir4_28	1791811	1791832	+	ttcatcatggaacgagccgaag	K2
pir4_66	1792539	1792561	+	tccacaagcatttgctcgatcac	K2
pir4_29	1793738	1793717	-	cgccgaatggcacttactttac	K2
pir4_114	1909525	1909546	+	cgccttctctctttgttactcc	
pir4_115	1909754	1909728	-	cgagttatcacttcgttttcacaagac	
pir4_116	1958322	1958345	+	catgaagaatttgaaaagcggggc	
pir4_117	1958634	1958613	-	gtacccgcttagacgcacaatg	
pir4_07	1985795	1985816	+	aatgtttcacgcccttggttcg	
pirkb_08	1986254	1986233	-	tttccgaaatcgagtcgcaacc	
pir4_30	2001297	2001318	+	gtgatagcccaccatcctgttc	D2
pir4_67	2001847	2001868	+	gctcgtgttgctgtatgtcctg	D2
pir4_68	2002678	2002657	-	aattgtccatcggtaggcgtag	D2
pir4_31	2003181	2003160	-	gcaaaacacaaaaccgacagcc	D2
pir4_32	2061750	2061771	+	atgccagattcccacgcacgag	J3
pir4_64	2064110	2064088	-	gcatcaagttcttcttcacgcac	J3
pir4_33	2064573	2064552	-	cggcggcgacgagaaaatgtcc	J3
pir4_34	2106110	2106131	+	gagcggcgaccaactgacttcc	L4
pir4_35	2107530	2107509	-	attgagcgattagctgcggtag	L4
pir4_36	2116226	2116247	+	tgctattgactgcgttccagcc	L5
pir4_69	2116800	2116821	+	tccatttcgctccgtttgcgtg	L5
pir4_70	2117423	2117402	-	atcaagccagagcagtttcagc	
pir4_37	2118026	2118005	-	cgttggcggcggttaagcagac	L5
pir4_93	2141046	2141067	+	aagcaacggtcgatgatgacgg	
pir4_09	2142782	2142803	+	cagcggggactccaagcaaagc	
pir4_10	2143090	2143069	-	gcgtcacttgtcataacgtggc	B1
pir4_38	2214418	2214439	+	tcggcttccgattgctgtgtgc	B1
pir4_71	2214949	2214972	+	caaacccaaatcattcaagccgtc	B1
pir4_39	2215821	2215800	-	cggtgtgggtgtaatgtcggtc	B1
pir4_40	2216446	2216467	+	cgttcactcccgatacttaccc	E3
pir4_72	2216942	2216963	+	ggcaattcaacatcgacgctgg	E3
pir4_73	2217550	2217529	-	cgagcagcaagtctttcactcc	E3
pir4_40/1	2218058	2218034	-	gactaccaacatggtttttacgctg	E3
pir4_42	2219734	2219755	+	tcatccaaattcggtagacccc	F2
pir4_74	2220345	2220366	+	tagacatccaactggcagtcgg	F2
pir4_75	2220895	2220874	-	gaagtcgttcgttggtttgccg	F2
pir4_43	2221372	2221351	-	ggctgttgatttaatgagccgc	F2
pir4_44	2226612	2226633	+	agcgttcaccccaactaaagag	F3
pir4_76	2227012	2227033	+	gatgctcgtttaccagcgactc	F3
pir4_77	2227939	2227918	-	gtaagccaccaccgacgtatag	F3
pir4_45	2228353	2228332	-	cttccgaacgaaacccaggcg	F3
pir4_46	2233807	2233828	+	caagattcaccctcctgaacgc	I1
pir4_78	2234757	2234734	-	gattgtcagagaattgttccagcg	I1
pir4_47	2235418	2235397	-	tgctgtgttcgtgacatcttgc	
pir4_94	2284489	2284468	-	gcagtcagttcagtgccatcgg	
pir3_101	2367664	2367689	+	ttgggggcggcactggcatagacggc	

Fortsetzung Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide

Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir4_100	2368112	2368091	-	ccgctggtgacaacggcaacgc	
pir4_48	2465456	2465477	+	agccacgcaaacatttggggac	D3
pir4_67	2465977	2465998	+	gctcgtgttgctgtatgtcctg	D3
pir4_68	2466808	2466787	-	aattgtccatcggtaggcgtag	D3
pir4_49	2467320	2467298	-	gcctagagggtgaagatggagag	D3
pir4_50	2471034	2471055	+	cttaactccattgtcaccctcc	E1
pir4_72	2471572	2471591	+	ggcaattcaacatcgacgct	E1
pir4_73	2472180	2472159	-	cgagcagcaagtctttcactcc	E1
pir4_51	2472797	2472772	-	ctttaaactctacaagaacggtcagc	E1
pir4_11	2532921	2532943	+	cattgatcgaatgcgtcatcagg	
pir4_12	2533444	2533423	-	gatggcattcccgttctatgtg	
pir4_95	2565623	2565644	+	ggatcaccacaacgccacgctg	
pir4_96	2565989	2565968	-	cccgttgcgttgatcgttcgcc	
pir4_97	2567477	2567498	+	gggctccaattcggtctcgatg	
pir4_52	2764293	2764314	+	aaccgacatgcctcacccatcc	F1
pir4_77	2764714	2764735	+	gtaagccaccaccgacgtatag	F1
pir4_74	2764925	2764946	+	tagacatccaactggcagtcgg	F1
pir4_75	2765475	2765454	-	gaagtcgttcgttggtttgccg	F1
pir54_76	2765641	2765620	-	gatgctcgtttaccagcgactc	F1
pir4_53	2766018	2765997	-	acctgccgcacagacgcattcg	F1
pir4_54	2773709	2773730	+	acaacatcgtccgctacgtcag	I3
pirX_22	2773774	2773796	+	tcccacgccgccactctccatcc	I3
pir4_78	2774737	2774714	-	gattgtcagagaattgttccagcg	I3
pir4_55	2775330	2775309	-	ttggcgttctgatagcgaatgc	I3
pir4_98	2947617	2947638	+	tggtcccaagaaacccatcgcc	
pir4_56	3076108	3076129	+	gccatgcgttcaacccccgtgc	A2
pir4_18	3076658	3076679	+	caagcaggagaactgtatcgtc	A2
pir4_17	3077455	3077435	-	agatagctcaacgacgtgatg	A2
pir4_57	3077804	3077783	-	ttaccctgatttcttcggctgg	A2
pir4_113	3083625	3083649	+	gacagacatcattcaagaatctccc	
pir4_13	3083935	3083956	+	tgtcgcctcgcccattcttcgg	
pir4_14	3084329	3084308	-	tgcatccatgcgaagagcaggg	
pir4_103	3084395	3084416	+	gacagacatcattcaagaatctccc	
pir4_58	3086301	3086323	+	cgaaaagttcttgaatccgtgcg	B8
pir4_71	3087220	3087197	-	caaacccaaatcattcaagccgtc	B8
pir4_59	3087794	3087773	-	gacgagaggaatgccaatgccc	B8
pir4_60	3149939	3149960	+	caccccaatcgtttcagccgcc	A3
pir4_17	3150357	3150378	+	cagatagctcaacgacgtgatg	A3
pir4_18	3151161	3151140	-	caagcaggagaactgtatcgtc	A3
pir4_61	3151716	3151691	-	gagttcaaactcgattttgctatggg	A3
pir4_118	3160683	3160706	+	cgatttcccgaagatatgcaactg	A3
pir4_119	3160912	3160891	-	agatttggaaggagcactcggg	A3
pir4_99	3161911	3161934	+	caacatctcgtaaagtgtgactcc	
pir4_rechts	3220061	3220082	+	actcttgatgagttgcacgggc	*Physical Gap*
pir4_107	3220280	3220301	+	ccatgtcttccaatcgagcagc	
pir1_84	3220281	3220305	+	catgtcttccaatcgagcagcaacg	
pir4_108	3220612	3220633	+	aaatgctcggaccaaacctgat	
pir1_85	3220612	3220639	+	aaatgctcggaccaaacctgatccaacg	

Fortsetzung Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide

Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir4_109	3220763	3220742	-	gatgacgcctcgctccgatcca	
pir1_83	3220967	3220943	-	aacttggtcacgaaggtttatcccg	
pir1_16	3221059	3221036	-	cattcctgcgattgatgctccgtg	*Physical Gap*
pir1_27	3234847	3234870	+	agtacaagagatcataccaatgcc	G1/H1
pir1_28	3237552	3237529	-	acatcaacacgattgccaactacg	G1/H1
pir1_29	3237657	3237680	+	atgtagccaactttgagcccatcg	G1/H1
pir1_30	3239270	3239247	-	gggagggtgaagtaagtgccattc	G1/H1
pir1_18	3247372	3247348	-	cagcacttcaaagctcatggttatc	
pir1_52	3360704	3360728	+	acgatggacacatcaatggcttcgg	
pir1_20	3383350	3383374	+	gcctgctacgtgctgaaaaagttcg	
pir1_76	3383678	3383654	-	cttctgatcttcggtcatggtcagg	
pir1_19	3384766	3384742	-	gtttatgcgttggcggtacagatgg	
pir1_77	3477515	3477539	+	actgacaagccacttcacatccacg	
pir1_22	3478070	3478046	-	gcaactcgaacgaggtgaagtaagg	
pir1_24	3522688	3522664	-	cgtattgatcgagccatttctgtag	A4
pir1_31	3526791	3526814	+	tgcgtgccatcactcaagagatcg	A4
pir1_32	3528588	3528565	-	gtggctggcatccatcttcaaatc	A4
pir1_23	3530465	3530489	+	ctccagccaatcaatgaatagcgac	A4
pir1_33	3562735	3562758	+	tgctgtcgctgagattcggacttg	B9
pir1_34	3564278	3564255	-	gcaacgccgtgaacgattttgaag	B9
pir1_25	3577291	3577315	+	aacaaacgaatcgtcagtcggctgg	A5
pir1_26	3579019	3578995	-	gtctggataccagttacgggttgtg	A5
pir1_21	3649985	3650009	+	ggcaatgaaacccaactcgcaaatc	
pir1_54	3766429	3766405	-	cgacaccagtgattcgcaaattcag	
pir1_17	3840894	3840918	+	gcgtatcagcagcgaattgctcacc	
pir1_53	3905148	3905124	-	cttggcgaacgaagtgtcgaaatag	
pir1_35	3952433	3952457	+	aaacatcgctcatacgaaccaagcc	G2
pir1_36	3955209	3955185	-	aacatctgttgtcggcatcgagcac	G2
pir1_37	3956889	3956913	+	gtcatgccaaagtgctgcgagtacc	H2
pir1_38	3958625	3958601	-	gatgcaccccgtctctataccaatg	H2
pir1_63	3962535	3962559	+	ttcttgttggatcgtgtcctgaaag	
pir1_78	3962988	3962964	-	ctccgtctcgtcctgtgcaaattgg	
pir1_55	3992591	3992615	+	tcgcatcttccagatcaaagacttg	
pir1_56	4008020	4008044	+	gatttggaccctgtgctggaacttg	
pir1_15	4116699	4116722	+	ttcaaatcaacttggtcgagtgtg	
pir1_14	4117130	4117107	-	tgatcgccacgaatcatctcttac	
pir1_13	4147370	4147393	+	aaacggtggcacttttggtggtgg	
pir1_12	4147822	4147799	-	ccaatatcgtcggcaagggtcaac	
pir1_58	4206857	4206881	+	gagatgattgtcggcggttggaaac	
pir1_59	4231884	4231860	-	tggtgagttcgatgagtcgattgag	
pir1_60	4335704	4335680	-	acaaagatggcgagcttttggcttc	
pir1_57	4341212	4341236	+	caactccgattgcgaaacgtgttcc	
pir1_61	4466046	4466070	+	aagcatttgagcaagggtgtccatc	
pir1_39	4478516	4478540	+	accgagccctttttgtttgttgag	A6
pir1_40	4480213	4480189	-	ctgttgatttattgagccgcaccgc	A6
pir1_P9	4611219	4611243	+	tttcacgctggtgggcactatcatc	
pir1_69	4611279	4611255	-	tcccattccgaacacagcagtcaag	
pir1_P11	4614131	4614158	+	ggttacttagatgtttcagttcaccagg	
pir1kh _P10	4614537	4614513	-	ttcttcggaagcgtcgctgcaatag	

Fortsetzung Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide

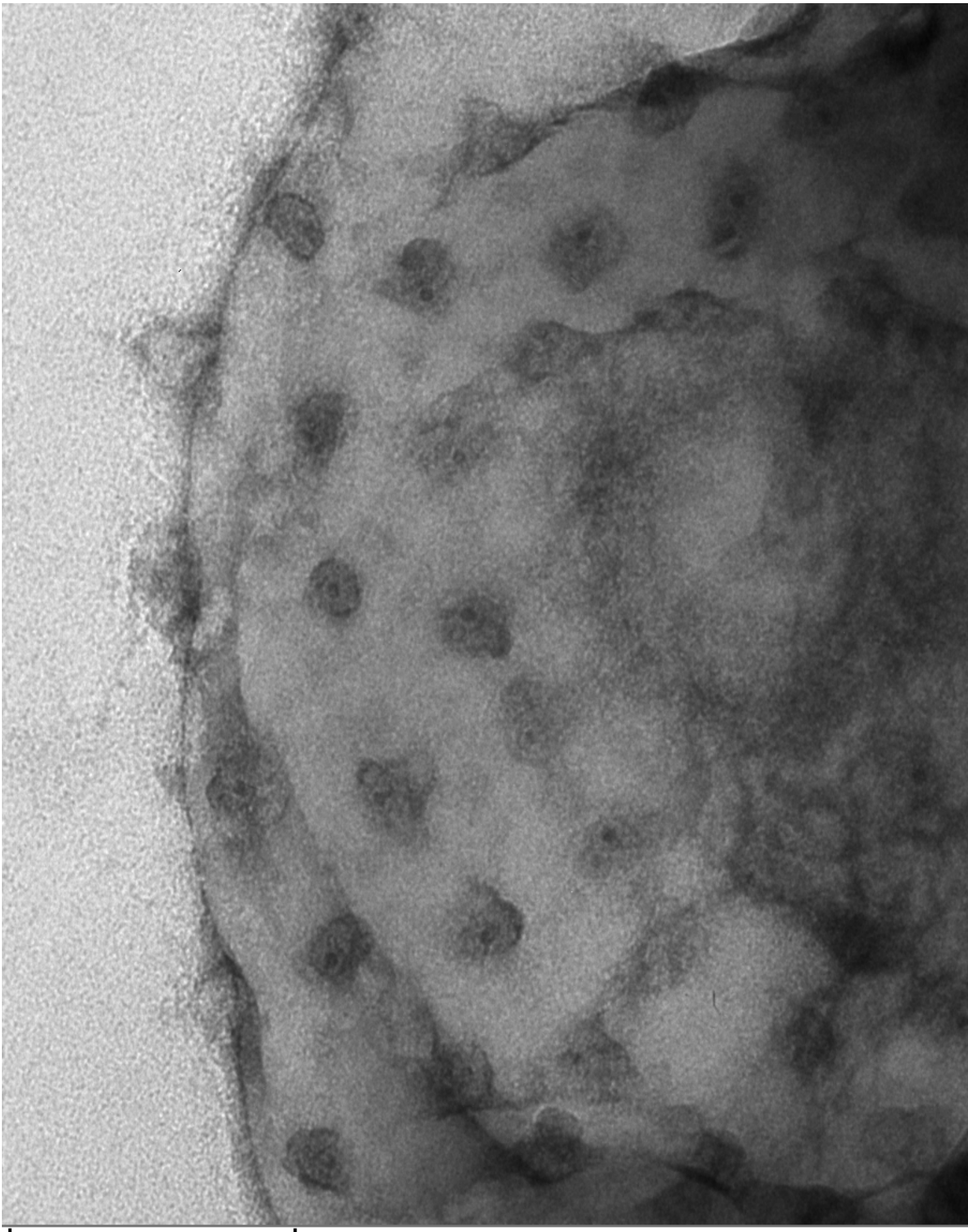
Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir1kh _P12	4617301	4617277	-	gccacgcagcaatttcatgtcgttc	
pir1_72	4625456	4625432	-	ccctccagtgcctggttacagattc	
pir1_62	4630725	4630701	-	accaagctccagcgagatttgcgtc	
pir1_73	4669235	4669211	-	gatcagcgatgggactgttgatgtg	
pir1_79	4696502	4696526	+	gtgtactgcgtactcatcatcacgg	
pir1_64	4696930	4696906	-	cagtgtgtgaagccctaagccaaac	
pir1_65	4787778	4787754	-	tgtggatctgaagggagaagcagtc	
pir1_11	4836257	4836234	-	tacagcaatgagtatccaacggtc	
pir1_10	4836831	4836808	-	atcgttcggtttgagctgtcagcg	
pir1_09	4880478	4880501	+	ctggttcgtgagatcctgcgtttc	
pir1_08	4880834	4880811	-	gcgagttgattcgtagtccactgc	
pir1_66	4892588	4892612	+	ggcagctcaaaaccagatcaagaac	
pir1_41	4936536	4936560	+	atcactgatttcaagtcaggcgagc	I5
pir1_42	4938071	4938047	-	tgttgtggcgtcgttgatatgacgg	I5
pir1_67	4994716	4994740	+	ttgcagcaaactgagcgatctctgg	
pir1_74	4996647	4996671	+	tgaatcgaaacgcaatggcggacag	
pir1_43	5048225	5048250	+	caaagttatgccgaccatccccaaag	D4
pir1_44	5050174	5050150	-	ggacattcgtgaggtcttgcttagg	D4
pir1kh _P8	5067960	5067984	+	gatcttattcagcgtgtagtacgcc	
pir1kh _P6	5070893	5070917	+	attcaatggcggatgtctgcgaagc	
pir1_P7	5071188	5071164	-	tgcccgtgatgacatcgcttattgg	
pir1kh _P4	5073949	5073973	+	catagctgctgcaaatcgcaaacgg	
pir1kh _P5	5074204	5074180	-	gatgggattgtgtgttcgattggag	
pir1kh _P2	5076707	5076731	+	cgattgaacgaagcggaagcaactc	
pir1kh _P3	5077168	5077144	-	ggaatcgctagtaatcgtaggtcag	
pir1kh _P1	5078764	5078740	-	acccgccttcttcacttcgcaatac	
pir1_68	5137594	5137618	+	aatcgccgctagacgatacaaatag	
pir1_45	5282725	5282749	+	cctcgacgctgatctttgggtgtag	D5
pir1_46	5284772	5284748	-	ggttgccttgacctagcgattgatg	D5
pir1_07	5308980	5309003	+	gcgagccgcatacgattcgatttg	
pir1_06	5309523	5309500	-	attccattgcttgcgtcagcagtg	
pir1_47	5338012	5338036	+	attcagttggtttggacgttggcgg	A7
pir4_18	5338571	5338592	+	caagcaggagaactgtatcgtc	A7
pir4_17	5339375	5339354	-	cagatagctcaacgacgtgatg	A7
pir1_48	5339766	5339742	-	tcgaacgtagatgtgtcaactttac	A7
pir1_05	5377297	5377320	+	gggagagcgatatactaggtcgag	
pir1_04	5377620	5377597	-	tcgtttccagatgctatcgagtcc	
pir1_80	5406721	5406745	+	ttgagcgatgtagttgcgaggattg	
pir1_70	5407171	5407147	-	cgattgcgtttacatgcgaacgacc	
pir1_49	5445402	5445426	+	gtgtggtcgaaattgatggcatcgg	K3
pir1_76	5446357	5446381	+	tctcttcaaggaacttgattgcctc	K3
pir1_50	5447495	5447471	-	tgctgcgtcgtcgatgctgatattg	K3
pir1_51	5451259	5451235	-	atcaacgtcggagtgatcgcagatg	
pir1_71	5584257	5584281	+	cgatgtaattcgccaaagactcgcc	
pir1_81	5584923	5584899	-	atcgctgacgacaatccagtttggc	
pir1_03	5631415	5631438	+	aacttggaaactattggaagtcac	
pir1_02	5632362	5632339	-	aacactcgtgtggatcaggattgg	
pir1_75	5632699	5632675	-	cgaaggcgacggaatcttcaatttc	
pir1_82	5670175	5670199	+	gggatgagcttgaccgaaatcttgg	

Fortsetzung Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide

Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir1_01	5670369	5670391	+	cggactggatgactgatgcaaag	*Physical Gap*
pir5_20	5670727	5670703	-	gtagcgtgacagtcaatgaagctcg	
pir5_01	5670874	5670851	-	tcgaacagtatggcgaagtgtcag	*Physical Gap*
pir5_07	5705725	5705749	+	tttgatgtcgtcaggcgtgacatag	
pir5_08	5714506	5714482	-	tgtccatctggaacaagatcacacg	
pir5_09	5730058	5730082	+	tcaggcttcgtttgctggtcagttc	
pir5_15	5740645	5740620	-	gcagcgttgtgactcacgatcatttg	
pir5_03	5759656	5759680	+	cgcttcgggttaggattagatcaac	F4
pir5_04	5761492	5761468	-	ttgctgagacgaaaacggcagttgg	F4
pir5_16	5792289	5792264	-	acggactttggaccgtcggtaatttg	
pir5_17	5793209	5793184	-	gtacgaaatcgtcgtccgcaatgaag	
pir5_10	5848353	5848377	+	aacgtagtcttggtaggtgatgccg	
pir5_11	5876986	5877010	+	atcgcaaatcaatcccgcagaatcc	
pir5_18	5900621	5900646	+	aaggggacggaggtatattgaccaag	
pir5_12	5901150	5901126	-	cccgaggctgttgatgaaccaaatg	
pir5_13	5965567	5965591	+	cgttcatatcatgtgggatgaagtg	
pir5_19	5988705	5988730	+	ctttgattacccgagcggatcaagac	
pir5_14	5989633	5989609	-	cagcgatgagtacaaacgcaatgag	
pir5_05	6022096	6022120	+	gtcctagttgttttgccgccaaacc	I4
pir5_06	6023695	6023671	-	acattcctgcgtacctgctcagaac	I4
pir5_02	6043233	6043256	+	gtgctgtgccaaccacgccgtatc	*Physical Gap*
pir5_21	6043764	6043740	-	tgaggatgacgcaaaggtgttggtg	
pir5_22	6044068	6044092	+	gtcatcacaaacggttccaatcgag	
pir3_rechts	6044380	6044359	-	ccttccgagaaaccaagacacc	*Physical Gap*
pir3_44	6044544	6044523	-	cttgttcgtagcgaccacttcc	
pir3_19	6127115	6127136	+	gagaggttgccgattcgctcag	E2
pir3_20	6128727	6128706	-	ggctttttgctccgcttgctgc	E2
pir3_21	6134070	6134091	+	ctccagcgtcccgagttgaacc	D1
pir3_22	6134778	6134757	-	cgggcgaacttaatgctcttgg	D1
pir3_15	6208391	6208417	+	gcctcaatgattcgatttaattcagcc	
pir3_14	6208438	6208459	+	cgtatcgacggtttgccacagc	
pir3_13	6208713	6208692	-	ctgcatgatcgactgcgtagcg	
pir3_16	6208852	6208830	-	gcgttcaatgaacaacgactacc	
pir3_12	6232969	6232990	+	ctggttgatgatgcactctcgg	
pir3_11	6233321	6233300	-	tcatcgaacagcgaatccagtg	
pir3_57	6253709	6253730	+	cgactgcttcgatgtgtttgtg	
pir3_58	6253959	6253938	-	tgagccatctcttcgagttgcg	
pir3_23	6262228	6262249	+	cgctcaggttgttctagggcac	C1
pir3_42	6262749	6262770	+	gtgtgggggcttgctaggtagg	C1
pir3_43	6263799	6263778	-	gttgtctcttcgacggtcccgc	C1
pir3_24	6264379	6264358	-	cggtacggtcagttccaccatc	C1
pir3_25	6277058	6277079	+	gtcgattcccgaaaccgttgcg	C1/C2
pir3_42	6277574	6277595	+	gtgtgggggcttgctaggtagg	C2
pir3_26	6279198	6279174	-	ggtcagttccaccattaattcgatg	C2
pir3_47	6294946	6294967	+	aaactactgcgtgtgttgcagg	
pir3_48	6297041	6297062	+	gcctgatagctctgggttgctg	
pir3_27	6327899	6327920	+	caatgctcttccacctgctctc	B4
pir3_28	6329339	6329318	-	ctgaatgacaagactccagtgc	B4
pir3_49	6329730	6329751	+	gagttgactccattcgccgccg	B4

Fortsetzung Tab. 73: Zusammenstellung der zur Absicherung verwendeten Oligonukleotide

Primer	Start	Ende	Strang	Sequenz 5' - 3'	Kommentar
pir3_50	6330049	6330028	-	ctcgacagcgaagcgattgcgg	
pir3_40G	6341389	6341413	+	cacctacgtgttggctttgggtggg	
pir3_39	6343365	6343386	+	cagggaccgacggaatcttcag	
pir3_17	6343804	6343825	+	taccgctttaggcaccgcagtc	
pir3_52	6344630	6344651	+	caaagtggctcacccgtgaccg	
pir3_51	6344756	6344778	+	gacgaattgacgactgattacgg	
pir3_53	6345069	6345045	-	aacagacgcacgtcggcacccgtcg	
pir3_54	6345200	6345179	-	atgtctgcgtatcgccttcatc	
pir3_09	6345293	6345272	-	gtggatcaagctggacctcgac	
pir3_18	6345454	6345433	-	gttgtcgccagcggtataggtc	
pir3_08	6345706	6345727	+	ggtggattacacgttctcgctg	
pir3_07	6346163	6346142	-	cggtattgaccgtgatgtcggc	
pir3_10	6346918	6346939	+	cgtgacagcttggaacactgcc	
pir3_06	6346944	6346965	+	caacctcgggtttcctcacagc	
pir3_05	6347666	6347644	-	caggattgtcaatcgtgatctcg	
pir3_38	6348754	6348733	-	actcacgttcttgtcggcgtcc	
pir3_41G	6349862	6349838	-	cggcggctgtcacattcagtgatcc	
pir3_29	6361703	6361724	+	gttgtgaaacgcacggttgggc	B5
pir3_30	6363302	6363281	-	tgcccaacaagcgagagtcatc	B5
pir3_69	6415319	6415340	+	cgttcaaacgtgaacaagcctg	
pir3_68	6418145	6418124	-	gactcagcgaagacgtttactc	
pir3_45	6503814	6503838	+	caaacatgactcgctgaaattctgg	
pir3_31	6593195	6593216	+	gccaaagtcaaaacggctgatg	A8
pir3_31a	6594972	6594951	-	ttcccaagcctcgaccaacgac	A8
pir3_32	6665650	6665671	+	gtcattagcgtctggcgttgag	A9
pir3_33	6667328	6667305	-	ccaacgaaatgaacgcattaccac	A9
pir3_04	6727735	6727756	+	cgtgctgtcatcgctcgcttcg	
pir3_03	6728095	6728073	-	gtgattgaggtcaaggtattccc	
pir3_66	6734604	6734625	+	gtgctttccgccacgcagaatg	
pir3_67	6734952	6734928	-	attgcgaaccgctcggcgtttgctc	
pir3_01	6736270	6736249	-	tgccatcgccaatggtcttggg	
pir3_55	6875827	6875848	+	ttgtccgctcatggaactgacg	
pir3_56	6876556	6876535	-	ttggttgcagtagatggctcgg	
pir3_34	6925565	6925586	+	gcatgtcgtttggacgcctcgc	I2
pir3_34a	6927130	6927109	-	gccatcggatgtggtgacggac	I2
pir3_35	6966126	6966147	+	ctgaatcctggtcttgcacatc	B6
pir3_36	6967525	6967504	-	agagcacaacccatcctgattc	B6
pir3_59	7040891	7040912	+	gcgacccttccctgttcgctc	
pir3_60	7041353	7041332	-	acaacgactcgacctttccctg	
pir3_37	7081549	7081570	+	gccataactcccaccaatactc	A10
pir3_37a	7083282	7083261	-	gggttctttatcaagcgatgcg	A10
pir3_links	7140024	7140045	+	cgcaatccgcttcttgagactg	*Physical Gap*
pir3_46	7141243	7141264	+	gacaggaaccaaccagctcaag	
pir3_65	7141345	7141366	+	gaaggcactgtggtagttgtgg	
pir3_62	7143224	7143245	+	gacttccgtgcttgggttcgag	
pir3_61	7143697	7143718	+	aagggcgtatagacttccgttg	
pir2_06	7143852	7143828	-	aggcaatcaagcaatatcagcgggc	
pir3_64	7143923	7143944	+	aatcagggcgttttggaagcac	
pir3_63	7144308	7144287	-	tggtgctgttcttcatcctggg	
pir2_04	7144354	7144331	-	aggctgtggctgcactacctgttg	*Physical Gap*
pir2_44	7144477	7144453	-	gttgagagcatgtacgtctttcggg	



Elektonenmikroskopische Aufnahme der Kraterstrukturen von *Pirellula* sp. Stamm 1, Negativfärbung (Kube & Lurz, MPI für Molekulare Genetik Berlin)