Metagenomik

Überblick, Hintergrund, Historisches

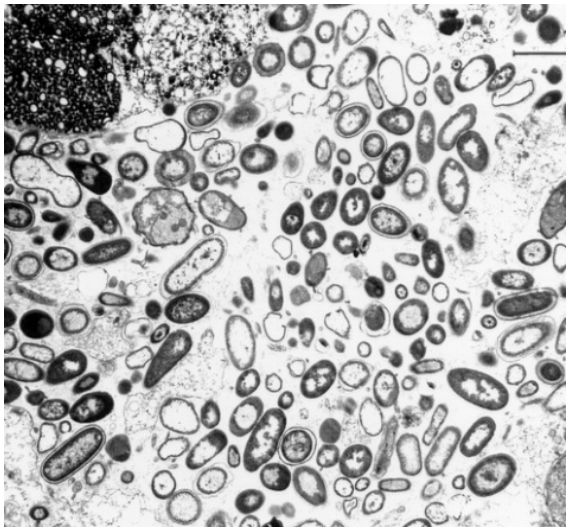


Abbildung 1 *Aufnahme einer mikrobiellen Gemeinschaft mit einem Transmissionselektronenmikroskop. (Abbildungsquelle: Fieseler et al., 2004, Appl. Environ. Microbiol. 70(6))*

Mikroorganismen sind universell und unabdingbar für das Leben auf der Erde. Sie sind die Primärproduzenten von Nährstoffen und recyceln totes Material in seine organischen Bestandteile. Mikroben kommen in nahezu sämtlichen Lebensräumen vor; von der Tiefsee über Wüsten bis hin zu den Verdauungstrakten fast aller multizellulären Organismen. Ein Gramm Waldboden beinhaltet schätzungsweise 4 x 10^7 prokaryotische Zellen und 2000-18'000 verschiedene Genome. Ähnlich vielfältig präsentiert sich die Situation in den Ozeanen. Ein Milliliter Meerwasser der Sargassosee beinhaltet schätzungsweise eine Million Bakterien und eine durchschnittliche Genomgrösse von zwei Millionen Basenpaaren. Trotz dieser riesigen Anzahl, Diversität, Abundanz und ihres grossen biotechnologischen Potentials wissen wir fast nichts über die grosse Mehrheit dieser Mikroorganismen.

Die Untersuchung mikrobieller Genome startete in den 1970er Jahren, als die Genome der Bakteriophagen MS2 und φ-X174 sequenziert wurden. Die Analyse von Mikroorganismen-Genomen ist jedoch ein schwieriges Unterfangen. Nur ein geringer Anteil (viele Habitate: 0.1-1%) der in der Natur vorkommender Mikroorganismen kann bisher im Labor kultiviert werden. Ausserdem leben Mikroben faktisch nie in Reinkultur bestehend aus einer einzelnen Spezies, sondern in Gemeinschaften, in denen sie miteinander und ihrem Lebensraum (z. B. Wirt-Organismen) interagieren (Abbildung 1). Eine klonale Kultur einer einzelnen Spezies repräsentiert somit den Zustand in der Natur nur eingeschränkt.

Die Gesamtheit der genomischen Information aller Mikroorganismen einer bestimmten Lebensgemeinschaft wird als Metagenom bezeichnet. Die Metagenomik befasst sich mit der Analyse genomischer DNA einer Lebensgemeinschaft, die durch direkte Entnahme einer Umweltprobe erhalten wurde (Abbildung 2). Die Metagenomik umgeht damit die Notwendigkeit, einzelne Mikroorganismen zu isolieren und zu kultivieren. Das Feld der Metagenomik begann mit dem Klonieren von DNA aus Umweltproben und darauffolgenden Expressions-Screenings. Die Entwicklung neuer Sequenzierungstechnologien und die damit verbundene Kostenreduktion haben die Erzeugung genetischer Daten von Mikroorganismen in ihrem natürlichen Lebensraum stark vereinfacht. So konnten Shotgun-Sequenzanalysen von metagenomischer DNA durchgeführt werden, um mikrobielle Gene zu identifizieren. Diese Studien bestätigten nicht nur die breiten Anwendungsgebiete der Metagenomik, sondern zeigten

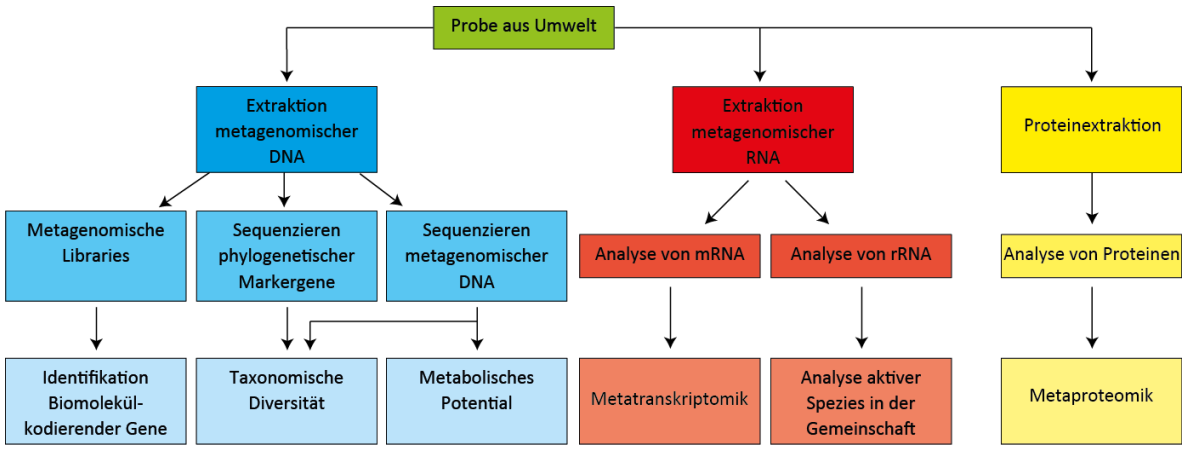


Abbildung 2 *Übersicht der Untersuchungsmethoden für mikrobieller Gemeinschaften durch metagenomische, metatranskriptomische und metaproteomische Analysen von Umweltproben. (angepasst von Simon C., and Daniel R., 2011, Appl. Environ. Microbiol. 77(4))*

auch die enorme funktionelle Gendiversität in Mikroorganismen auf. Während bei der Genomanalyse eines einzelnen Organismus meist das komplette Genom sequenziert werden kann, gibt es bei der Analyse von metagenomischer DNA viele kleine Fragmente, die oftmals nicht eindeutig einer Spezies zugeordnet werden können. Dennoch ermöglicht die Metagenomik die Generierung genetischer Information über potentiell neuartige Biomoleküle, genomische Beziehungen zwischen Funktion und Phylogenie unkultivierter Organismen und evolutionäre Profile von Struktur und Funktion einer mikrobiellen Lebensgemeinschaft.

Diversität und Rarefaction-Analyse

Um die mikrobielle Diversität einer Umweltprobe zu bestimmen, wird oftmals ribosomale RNA (rRNA) analysiert. Durch ihre essentielle und konservierte Funktion in jeder lebenden Zelle unterliegt sie kaum horizontalem Gentransfer (also Gentransfer von einer Spezies zu einer anderen) und eignet sich somit optimal als molekularer Marker, um verwandtschaftliche Beziehungen unter den Organismen zu bestimmen. Da die Analyse der kleinen rRNAs (5S und 5.8S) zu wenig phylogenetische Information liefert und die grossen rRNAs (23S und 28S) schwieriger zu untersuchen sind, hat sich die Analyse der 16S rRNA bei Prokaryonten und 18S rRNA bei Eukaryonten als Methode der Wahl durchgesetzt. Mit den generierten Daten können mit den Methoden die Sie im Bioinformatikabschnitt kennen gelernt haben, Sequenzalignments und phylogenetische Stammbäume erstellt werden. Ein kurzes Video zum 16S rRNA Sequencing und dem American Gut Project ist auf dem YouTube-Kanal unseres GGB-Kurses[1] zu finden.

Da die einzelnen Mikroorganismen einer Umweltprobe nicht direkt beobachtbar oder zählbar sind, ist es schwierig abzuschätzen, wie viele Proben für eine aussagekräftige Untersuchung notwendig sind. Die Rarefaction-Analyse ist ein Hilfsmittel, um den Anteil einer Probe am gesamten Artenreichtum abzuschätzen. Dazu werden Rarefaction-Graphen erstellt, wobei die Anzahl gefundener Spezies als Funktion der Anzahl Proben geplottet wird (Abbildung 3). Die Kurven weisen meistens eine starke Steigung zu Beginn auf und flachen danach ab, da mit steigender Probenzahl weniger neue Spezies pro Probe detektiert werden. Je kleiner die Steigung ist, desto weniger trägt das Sampling zur Identifikation von operational taxonomic units (OTUs) bei. Diese Methode eignet sich zudem gut, um die genetische Diversität zwischen Studien mit unterschiedlicher Probenanzahl zu vergleichen und durch Extrapolation auf die Gesamtzahl der OTUs zu schliessen.

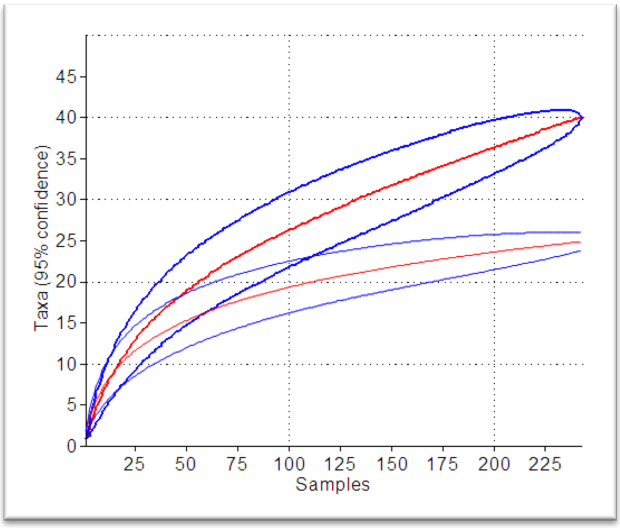


Abbildung 3 *Rarefaction-Graph. Die Anzahl gefundener Taxa ist als Funktion der Anzahl Proben geplottet (rot) mit den 95%-Konfidenzintervallen (blau). (Abbildungsquelle: Veterinary Population Medicine Departement of the University of Minnesota Stand März 2014)*

Metagenomische Libraries

Das Konstruieren metagenomischer Libraries (Abbildung 4) erlaubt die genetische Analyse mikrobieller Gemeinschaften, ohne die einzelnen Organismen zu kultivieren. Anfänglich wurden metagenomische Methoden hauptsächlich für die Suche nach neuen Biomolekülen angewandt. Dabei kommen entweder funktions- oder sequenzbasierte Screenings zum Einsatz. Beide Screening-Techniken beinhalten das Klonieren metagenomischer DNA aus der Umweltprobe und die anschliessende Konstruktion metagenomischer Libraries in einem passenden Host.

Extraktion und Aufbereitung metagenomischer DNA

Die physikalische und chemische Struktur einer mikrobiellen Gemeinschaft beeinflusst die Grösse, Menge und Qualität an metagenomischer DNA, die extrahiert werden kann. Die Analyse von Lebensgemeinschaften im Wasser erfordert oftmals das Sammeln grosser Wassermengen, um genügend DNA zu erhalten. DNA Extraktion aus Bodenproben kann aufgrund von anorganischen Stoffen, Huminsäuren oder DNasen im Boden schwierig sein. Das Entfernen von Kontaminationen bestimmt sowohl die Klonierbarkeit als auch die Grösse der klonierten DNA Fragmente, denn viele Methoden zur effektiven Entfernung von Kontaminationen scheren metagenomische DNA bzw. hemmen den Klonierungsprozess. Nach der DNA

[1] http://youtu.be/vnF6y8xkusA

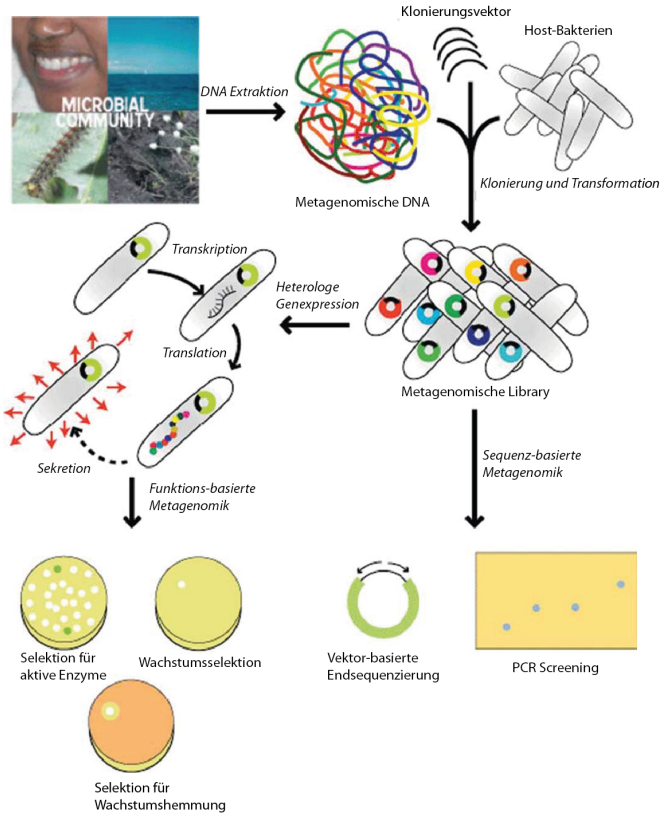


Abbildung 4 *Konstruktion einer metagenomischen Library. Metagenomische DNA wird direkt aus der Umweltprobe extrahiert, in ein Vektorsystem kloniert und in einen passenden Host-Organismus transformiert. Zwei verschiedene Ansätze können für die Analyse metagenomischer Libraries angewendet werden. Funktionelle Metagenomik erfordert die Expression der rekombinanten DNA im Host-Organismus. Dabei kann für aktive Enzyme, Antibiotikaproduktion oder Wachstum unter limitierenden Bedingungen selektioniert werden. Bei der Sequenz-basierten Metagenomik werden entweder zufällige DNA Fragmente mit Vektorprimern sequenziert oder metabolische Gene mittels PCR Screening detektiert. (Abbildungsquelle: Sabree et al., 2009, in Encyclopedia of Microbiology (Third Edition))*

Extraktion folgt meist eine Gelelktrophorese, um grosse DNA Fragmente (high molecular weight DNA) von kleineren Fragmenten, Huminsäuren und DNAsen zu trennen. Die Anwendung mehrerer aufeinanderfolgender Extraktionsmethoden führt zu minimal kontaminierter DNA und vereinfacht den Klonierungsprozess. Bei jedem Extraktionsschritt geht aber auch DNA verloren. Eine weitere Schwierigkeit besteht im Isolieren von DNA aus robusteren Organismen wie beispielsweise eingekapselten Bakterien oder Sporen. Sie erfordern heftigere Lysemethoden. So kann ein diverserer Anteil der mikrobiellen Gemeinschaft für die Klonierung erhalten werden. Die Gesamtmenge an extrahierter DNA kann jedoch durch diese heftigeren Bedingungen wieder reduziert werden. Der Prozess der DNA Extraktion und Aufreinigung ist somit ein Balanceakt, um die optimale Menge, Grösse und Qualität an DNA für die gewünschte Analyse zu generieren.

Klonierungsvektoren

Je nachdem wie gross die zu klonierenden DNA-Fragmente sind, werden Plasmide (bis zu 15 kb), Cosmide, Fosmide (beide bis zu 40 kb) oder künstliche Bakterienchromosomen (BACs, >40 kb) als Vektoren verwendet. Fosmide liegen normalerweise als einzelne Kopie in der Zelle vor, während Cosmide mit 20-70 Kopien pro Zelle vertreten sind. Fosmide sind dadurch etwas stabiler und in machen Fosmiden kann die Produktion von mehreren Kopien pro Zelle durch Zugabe von Arabinose induziert werden. Die Grösse und Art des Vektorsystems hängt dabei von der DNA-Qualität, den Zielgenen und der Screening-Strategie ab. Libraries mit kleinen DNA-Inserts können der Identifikation von Biomolekülen, welche durch ein einzelnes Gen kodiert sind, dienen. Mehrere Enzyme wie beispielsweise Amidasen, Hydrolasen oder Cellulasen wurden durch das Screening metagenomischer Libraries mit Inserts kleiner als 10 kb identifiziert. Libraries mit grossen DNA-Inserts sind für das Detektieren von komplexeren Biosynthese-Pathways, die durch mehrere Gene kodiert sind, von Nutzen. Dabei nutzt man die Tatsache, dass in bakteriellen Genomen funktional verknüpfte Gene (z.B. die Enzyme die zu einem bestimmten Stoffwechselweg gehören) oft in sogenannten Operons zusammen clustern.

Komplexität mikrobieller Gemeinschaften und Struktur der Library

Die Art der zu untersuchenden Gene und die Komplexität einer mikrobiellen Gemeinschaft bestimmen die Grösse des Zielfragments, das Vektorsystem und die minimal notwendige Anzahl Klone in der Library. Shotgun Sequencing wird meist zur Analyse von Libraries mit kleinen Inserts verwendet. In Aktivitäts-basierten Analysen sind hingegen grosse Inserts von Vorteil, da die Wahrscheinlichkeit, die gewünschte Aktivität in einem einzelnen Klon zu finden, in Libraries mit grossen Inserts erhöht ist.

Libraries mit kleinen Inserts in Plasmiden (bis zu 10 kb) erfordern 3-20 Mal mehr Klone als Libraries mit Inserts in Cosmiden, Fosmiden oder BACs, um einen ähnlichen Anteil der Diversität abzudecken. Die Erfassung einer kompletten Gemeinschaft und das Rekonstruieren von Genomen sind nur in simplen Systemen zu erreichen. Komplexe Metagenome können daher nur auf bestimmte Funktionen oder Sequenzen untersucht werden.

Transformation eines geeigneten Hosts

Der am häufigsten verwendete Organismus für das Erstellen metagenomischer Libraries ist *E. coli*. *E. coli* weist

viele nützliche Eigenschaften für metagenomische Anwendungen auf. Mutationen die Rekombination (*recA*) und DNA Degradierung (*endA*) hemmen sind ebenso wünschenswert wie solche, die ein rekombinantes blau/weiss Screening (*lacZ*) erlauben. Elektroporation ist die am häufigsten verwendete Methode, um metagenomische DNA in *E. coli* einzuführen. Das Packen der DNA durch Phagen und die anschliessende Transfektion von *E. coli* ist eine weitere Methode, um metagenomische DNA in den Host einzuschleusen. Das Screening metagenomischer Libraries in *E. coli* hat zu vielversprechenden Ergebnissen geführt. Die Expression von Genen aus Organismen, die nur entfernt mit *E. coli* verwandt sind, stellt jedoch ein Problem dar. Deshalb wurden auch andere Wirte verwendet, um funktionsbasierte Screenings durchzuführen. In vielen dieser Wirte ist es jedoch schwierig die grossen Klonzahlen zu erzeugen, die mit *E. coli* erreicht werden. sequenzbasierte Analysen werden momentan fast ausschliesslich in *E. coli* durchgeführt.

Sequenzbasierte Metagenomik

Im Gegensatz zu funktionellen Screenings beruht das sequenzbasierte Screening auf einer Sequenzanalyse und den darauffolgenden Schlussfolgerungen auf mögliche Funktionen oder Verwandtschaftsbeziehungen. Sequenzbasierte Screeningansätze beinhalten das Design von Primern, welche auf konservierten Regionen von bereits bekannten Gen- oder Proteinfamilien basieren. PCRs mit Primern für die konservierten Abschnitte von 16S rRNAs erlauben beispielsweise die Einschätzung der Diversität einer mikrobiellen Gemeinschaft. Mit Primern für konservierte Sequenzabschnitte der schon aus anderen Organismen bekannten metabolischen Gene lassen sich so in den untersuchten Proben neue Varianten dieser Gene identifizieren.

Die grossen Fortschritte der Sequenzierungstechnologien erlauben nun auch, grosse Libraries ohne Vorselektion von Klonen zu sequenzieren. Dabei verwendet man Primer gegen diejenigen Sequenzen der Klonierungsvektoren welche die klonierten Sequenzen flankieren. So kann man das Insert von beiden Seiten sequenzieren. Solche Analysen haben zur Akkumulation einer beträchtlichen Menge an Sequenzdaten von unkultivierten Mikroben geführt.

Funktionsbasierte Metagenomik

Funktionelle Metagenomik beruht auf der Identifikation von Klonen, welche eine bestimmte Aktivität aufweisen, die durch metagenomische DNA hervorgerufen wird. Die funktionsbasierte Metagenomik erlaubt die Analyse verschiedener Funktionen einer Gemeinschaft. Im Gegensatz zu sequenzbasierten Methoden erfordern funktionelle Studien keine Homologie zu Genen mit bekannter Funktion. Sie ermöglichen somit die Erforschung bisher unbekannter Gene mit neuartigen Funktionen. Das funktionsbasierte Screening ist für die Identifikation der meisten Gene, welche für neuartige Biomoleküle kodieren, verantwortlich. Dabei wird die metabolische Aktivität von einzelnen Klonen in der metagenomischen Library untersucht. Falls die Analyse einer spezifischen Biomolekülklasse gefragt ist, können mikrobielle Gemeinschaften vor der Konstruktion der metagenomischen Library manipuliert werden, um Mikroben mit der gewünschten Aktivität in der Probe anzureichern.

Gene, die Bakterien Antibiotikaresistenzen verleihen, sind von grossem wissenschaftlichem und gesundheitlichem Interesse. Horizontaler Gentransfer und die vielfältige Nutzung von Antibiotika haben zu einer weiten Verbreitung von Antibiotikaresistenzgenen in mikrobiellen Gemeinschaften geführt. Die Charakterisierung von Resistenzgenen aus unkultivierten Gemeinschaften ist wichtig für die zukünftige Anwendung von Antibiotika. Das Verständnis von Resistenzmechanismen in unkultivierten Organismen könnte neue Aufschlüsse in der Bekämpfung von Resistenzen geben und damit die Entwicklung neuer, effektiverer Medikamente vorantreiben. Durch die Verwendung von Selektionsstrategien, die Sie bereits aus der Lerneinheit zur Bakteriellen Genetik kennen, lassen sich antibiotische Resistenzgene auch aus grossen metagenomischen Libraries sehr effizient anreichern. In solchen Studien wurden sowohl Homologe zu bereits bekannten Resistenzgenen aus kultivierbaren Organismen, als auch gänzliche neue Resistenzgene ohne Ähnlichkeit zu bereits bekannten Genen gefunden.

Sequenzierung und Datenanalyse

Sequenzierungstechnologien

Für das Sequenzieren prokaryotischer Genome wird typischerweise Shotgun Sequencing verwendet. Diese Methode schert die DNA zufällig in kleinere Fragmente, sequenziert viele dieser kurzen Fragmente und rekonstruiert diese zu einer zusammenhängenden Sequenz. Shotgun Metagenomik gibt Auskunft über die vorhandenen Organismen in der Probe und das metabolische Potential einer Gemeinschaft. Da die Entnahme von DNA aus einer Umweltprobe nicht kontrolliert werden kann, sind die am meisten verbreiteten Organismen in den Sequenzdaten am besten repräsentiert. Dennoch erlaubt die zufällige Natur des Shotgun Sequencing eine teilweise Abdeckung von unterrepräsentierten Organismen in der Probe.

Durch die Entwicklung von high-throughput Sequenzierungsmethoden konnte das Sequenzieren metagenomischer DNA stark vereinfacht und parallelisiert werden. Die ersten metagenomischen Studien, die mit high-throughput Sequencing durchgeführt wurden, verwendeten paralleles 454 Pyrosequencing. Andere Technologien, die

Sequenzlänge (bp)	***Genomelement***
25-75	SNPs, short frameshift mutations
100-400	kurze funktionale Elemente
500-1000	Domänen, single domain genes
1000-5000	kurze Operone, multidomain genes
5000-10‘000	längere Operone, *cis*-control elements
> 100’000	Prophagen, pathogenicity islands, mobile insertion elements
> 1’000’000	Chromosomorganisation Prokaryoten

Abbildung 5 *(Wooley et al., 2010, PLoS Comput. Biol. 6(2))*

häufig in Metagenomikstudien Anwendung finden, sind das Ion Torrent Personal Genome Machine System, Illumina MiSeq oder HiSeq und das SOLiD System von Applied Biosystems. Diese Methoden generieren zwar kürzere Fragmente als Sanger Sequencing, dafür ist die Anzahl Reads viel höher. Ein weiterer Vorteil dieser Technologien ist, dass nicht notwendige Klonieren der DNA, bevor sequenziert werden kann.

Assemblierung

Bei der Sequenzierung eines Genoms werden die einzelnen Fragmente zu immer länger werdenden, zusammenhängenden Sequenzen zusammengefügt bis schliesslich das gesamte Genom vollständig ist. Durch die Analyse langer, zusammenhängender DNA-Fragmente können open reading frames (ORFs), Operone, operational transcriptional units und Bindungsstellen von Transkriptionsfaktoren identifiziert werden. Der Gewinn an Information korreliert dabei mit der Länge der genomischen Fragmente.

Im Gegensatz zur relativ einfachen Assemblierung einzelner Genome von kultivierten Mikroorganismen ist eine vollständige Assemblierung eines Metagenoms gegenwärtig meist nicht realisierbar. Erstens ist die Probenahme unvollständig, wodurch die meisten Genome nur teilweise oder unzureichend abgedeckt werden. Zudem ist die Information über die verschiedenen Spezies meist unvollständig und es ist schwierig, einzelne Fragmente der richtigen Spezies zuzuordnen. Aus diesen Gründen ist die Identifikation genomischer Elemente aus metagenomischen Daten auf die ersten drei oder vier Zeilen in der obigen Tabelle limitiert.

Die Abdeckung eines Genoms ist definiert als die durchschnittliche Häufigkeit, mit der ein einzelnes Nukleotid sequenziert werden muss, um das komplette Genom zu identifizieren. Wenn ein Genom in einem Read sequenziert werden könnte, würde demnach eine einfache Deckung für die Sequenzierung ausreichen. Je nach Sequenzierungstechnologie werden Read Lengths von 25-700 bp erreicht und so müssen einzelne Nukleotide mehrmals sequenziert werden, um zu gewährleisten, dass alle Reads überlappen und diese Überlappungen einzigartig genug sind, um das Genom durch Assemblierung der Fragmente zu rekonstruieren.

Unter der Annahme, dass die Fähigkeit, eine Überlappung zwischen zwei tatsächlich überlappenden Reads zu detektieren, bei allen Klonen gleichbleibt, kann die Anzahl notwendiger Reads, um ein komplettes Genom zu sequenzieren, mit einem Poisson-Modell beschrieben werden. Das Modell ist durch die Lander-Waterman Gleichung gegeben, in der *L* die Read Length, *N* die Anzahl Reads und *G* die Genomgrösse darstellen. *C* stellt die oben erwähnte Abdeckung (Coverage) dar.

$$C = \frac{L \times N}{G}$$

Der prozentuale Anteil des Genoms, der durch die Sequenzen abgedeckt wird, ist demnach gegeben durch P_0.

$$P_0 = 1 - e^{-C} = 1 - e^{-(\frac{LN}{G})}$$

Diese Formel kann umgestellt werden, um die Anzahl benötigter Reads *N* für die gewünschte Abdeckung P_0 zu erhalten.

$$N = -\frac{\ln(1 - P_0)}{L} \times G$$

Die Metagenom-Grösse G_m einer Probe mit *l* Spezies und n_i Kopien eines Genoms ist:

$$G_m = \sum_{i=1}^{l} n_i G_i$$

G_i ist dabei die Grösse eines beliebigen Genoms in der Probe. Die verschiedenen Spezies in einer Probe erscheinen jedoch mit unterschiedlicher Häufigkeit im Metagenom. Deswegen kann die Metagenom-Grösse G_m als Summe der einzelnen Genomgrössen G_i mit ihren relativen Häufigkeiten p_i geschrieben werden.

$$\hat{G}_m = p_1 G_1 + p_2 G_2 + \ldots + p_l G_l \qquad \sum_{i=1}^{l} p_i = 1$$

Durch die Anwendung von Spezies-spezifischen Genmarkern (z. B. rDNA) können die Diversität in der Probe und somit die p_i-Werte abgeschätzt werden. Dennoch

sind komplette oder adäquate Abdeckungen für Spezies-reiche Gemeinschaften schwierig zu erreichen.

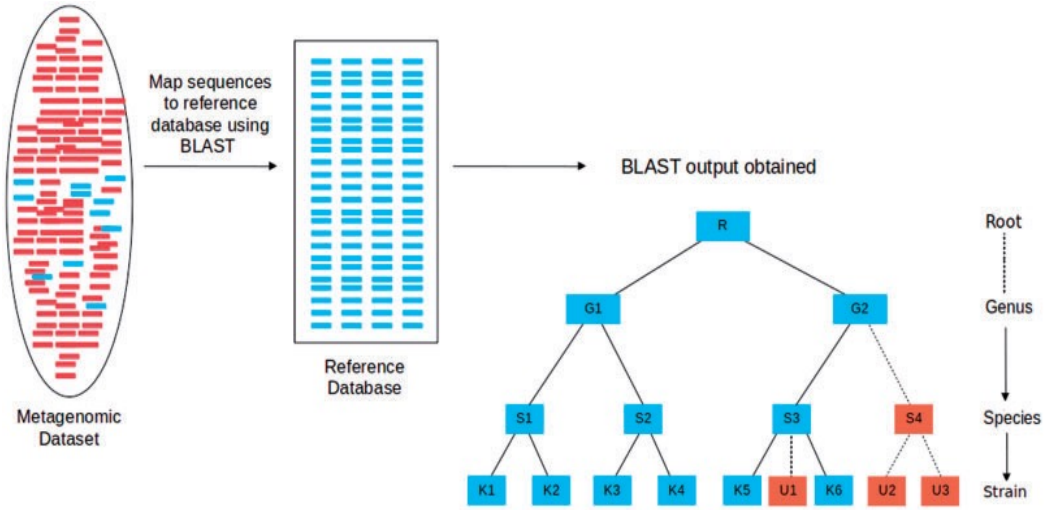
Operational taxonomic units (OTUs) sind mikrobielle Diversitätseinheiten, die Mikroben anhand von Ähnlichkeiten zwischen DNA Sequenzen (typischerweise rDNA) klassifizieren. OTUs dienen somit der Spezies-Differenzierung von Mikroorganismen. Da die Erfassung eines Metagenoms normalerweise unvollständig ist, besteht die Gefahr, dass bei der Assemblierung Sequenzen verschiedener operational taxonomic units (OTUs) zusammengefügt werden und Inter-Spezies Chimären entstehen. Phrap, Forge, Arachne, JAZZ und der Celera Assembler sind Assemblierungsprogramme, die für die Assemblierung einzelner, Sanger-sequenzierter Genome entwickelt wurden. Dennoch liefern diese Programme relativ gute Resultate, wenn metagenomische, Sanger-sequenzierte Fragmente assembliert werden. Die meisten dieser Algorithmen verwenden mate-pair Informationen, um die zusammengefügten Fragmente zwischen rohen Reads und ganzen Chromosomen zu vergleichen.

Binning

Die Zuordnung von Sequenzierungsdaten zu den entsprechenden OTUs wird als Binning bezeichnet. Häufig sind jedoch phylogenetische Markergene (z.B. 16S rRNA-Gene) nicht vorhanden, da sie nicht auf den assemblierten Fragmenten liegen oder bei unterrepräsentierten Organismen sogar im gesamten Datensatz fehlen. Es wird dabei zwischen Taxonomie-abhängigen und -unabhängigen Methoden unterschieden.

Taxonomie-abhängige Methoden

In Taxonomie-abhängigen Methoden werden Reads mit Sequenzen aus Referenzdatenbanken verglichen und aligniert. Reads welche eine zu geringe Ähnlichkeit mit Referenzsequenzen aufweisen, werden als unassigned kategorisiert. Taxonomie-abhängige Methoden können weiter in Alignment-basierte, Kompositions-basierte und Hybridmethoden unterteilt werden. Alignment-basierte Methoden benutzen BLAST, um individuelle Reads der Umweltproben mit Referenzsequenzen aus bereits charakterisierten Genomen zu alignen. Solche Referenzsequenzen sind in vielen öffentlichen Sammlungen wie NCBI, EMBL oder UniProt vorhanden. Reads werden schliesslich anhand ihrer Alignments mit verschiedenen Hit-Sequenzen aus der Datenbank in taxonomische Gruppen eingeordnet. Da bei metagenomischen Proben viele Reads zu unbekannten Taxa gehören, wäre es falsch, diese Reads dem passendsten BLAST-Hit zuzuordnen. Deshalb verwenden viele Programme zusätzlich



Reads originate from	Significant BLAST Hits	Assingment Strategies	
		Best BLAST Hit Approach	LCA
K1	K1, K2, K3	K1 (√)	G1 (√)
U1	K5, K6	K5 (X)	S3 (√)
U2 and U3	K5, K6	K5 (X)	S3 (X)

Abbildung 6 *Genauigkeit taxonomischer Zuordnungen der „best BLAST hit" Methode und der „lowest common ancestor (LCA)" Methode. Reads, die von bekannten Genomen stammen (K1) können mit beiden Methoden richtig zugeordnet werden. Reads aus Genomen, die nicht in Referenzdatenbanken vorhanden sind (U1), können mit der LCA Methode besser zugeordnet werden, da die „best BLAST hit" Methode in diesem Fall falsche Zuweisungen macht. Wenn es sich jedoch um eine komplett neue Spezies (U2 und U3) oder Gattung handelt, sind beide Methoden fehleranfällig. (Abbildungsquelle: Mande et al., 2012, Brief. Bioinform. 13(6))*

die LCA-Methode (lowest common ancestor), welche ein Read nicht einem einzelnen Stamm, sondern dem Vorfahren der Organismen mit signifikanten BLAST-Hits zuordnet (Abbildung 6).

Kompositions-basierte Methoden verwenden Eigenschaften wie GC-Gehalt, Codon-Verwendung und Oligonukleotidmuster, um Reads mit Sequenzen in Referenzdatenbanken zu vergleichen (Abbildung 7). Die endgültigen taxonomischen Zuordnungen werden anhand von absoluten oder relativen Ähnlichkeiten gemacht. Diese Methoden sind im Gegensatz zu Alignment-basierten Techniken schneller und brauchen weniger Rechnerleistung. Die einzelnen Reads müssen jedoch genügend lang sein, damit eine taxonomische Differenzierung möglich ist.

Hybridmethoden verwenden eine Kombination aus Alignment- und Kompositions-basierten Strategien für eine taxonomische Klassifikation. Der SPHINX Algorithmus besteht aus einem Zweiphasen-Binning. In der ersten Phase wird die Zusammensetzung eines Reads mit geclusterten Referenzsequenzen verglichen. Das Ziel dieser Phase ist die Identifizierung eines Subsets an Referenzsequenzen mit kompositioneller Ähnlichkeit zum Read. In der zweiten Phase wird der Read mit dem Subset an Referenzsequenzen aligniert und auf Ähnlichkeiten untersucht. Während die erste Phase die Menge an Referenzsequenzen reduziert, gewährleistet die zweite Phase die Genauigkeit und Spezifität der Zuordnung.

Taxonomie-unabhängige Methoden

Taxonomie-unabhängige Techniken verwenden ausschliesslich intrinsische Informationen und sind nicht auf Referenzsequenzen aus Datenbanken angewiesen. Die einfachste Strategie unter diesen Methoden weist der TETRA Algorithmus auf. Dieser errechnet paarweise Korrelationsmuster verschiedener Tetranukleotide für die einzelnen Reads. Die Verwendung von 4-mer Häufigkeiten ist gestützt auf die Tatsache, dass 4-Mere ein optimales Diskriminierungspotential aufweisen. Die so gewonnene Information wird anschliessend verwendet, um die einzelnen Reads in unterschiedliche Gruppen zu ordnen. Es existieren noch zahlreiche weitere und komplexere Taxonomie-unabhängige Binning-Algorithmen (SOMs, CompostBin, MetaCluster), die hier nicht weiter beschrieben werden.

Annotation

Nach der Assemblierung des Metagenoms und der Identifikation mutmasslicher ORFs möchte man das funktionelle Potential der mikrobiellen Gemeinschaft untersuchen. Die erste Phase des Annotierens besteht in der Zuweisung biologischer Funktionen zu einzelnen ORFs. Unvollständige ORFs und grosse Fraktionen ohne annotierte Homologe machen dies für metagenomische Proben zu einer anspruchsvollen Aufgabe. Der zweite Schritt beinhaltet die Identifikation von Genen, die ein biologisches Netzwerk darstellen und beispielsweise einen metabolischen Pathway ausmachen. Diese Aufgabe ist für metagenomische Proben ebenfalls schwierig, da nicht jedes annotierte ORF einer einzelnen Spezies zugeordnet werden kann. Eine Strategie für die funktionelle Annotation

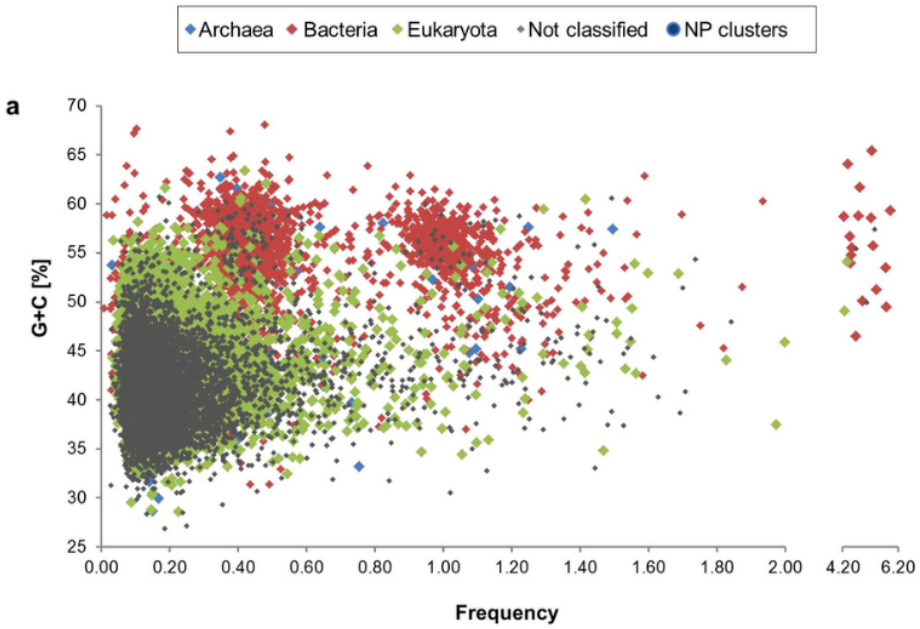


Abbildung 7 *Binning-Graph. Der GC-Gehalt ist als Funktion der relativen Häufigkeiten der sequenzierten, zusammengehängten Sequenzen (Contigs) geplottet. Eine BLASTX-Analyse wurde gegen eine Referenzdatenbank durchgeführt und jedem Contig wurde die taxonomische Domäne des besten Hits zugeordnet (rot = Bakterien; blau = Archaea; grün = Eukaryoten; schwarz = nicht klassifiziert). (Abbildungsquelle: Wilson et al., 2014, Nature 506(7486))*

metagenomischer Information ist die Anwendung von six-frame Translationen auf die verschiedenen Reads. Bei ausreichend langen Reads können so ORFs detektiert werden. Für kürzere Reads werden möglicherweise partielle ORFs identifiziert. Diese können dann nach Motiven Sequenzsignaturen durchsucht werden, die auf Funktionalität hinweisen. Einzelne, nicht-assemblierte Reads lassen ebenfalls Rückschlüsse auf Funktionalität zu, falls sie lang genug sind, um kurze Motive oder signifikante BLAST Hits zu finden.

Zwei vielseitige und für metagenomische Daten nützliche Annotations-Pipelines, welche die oben aufgeführten Prinzipien anwenden sind MG-RAST (Metagenomic Rapid Annotations using Subsystems Technology) und RAMMCAP (Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline). MG-RAST ist eine open source Webapplikation, die eine phylogenetische und funktionelle Analyse von Metagenomen durchführt. Funktionelle Zuweisungen werden durch Sequenzvergleiche mit Datenbanken auf dem Nukleotid- und Proteinlevel gemacht. RAMMCAP verwendet den Clustering-Algorithmus CD-HIT, um translatierte ORFs nach Sequenzähnlichkeit zu kategorisieren. Das Auftreten vieler ähnlicher ORFs bestärkt dabei die Hypothese, dass es sich um wahre ORFs handelt. CD-HIT reduziert so die Menge an Annotationen auf die repräsentativen Sequenzen im Metagenom. So ergeben sich Ähnlichkeits-basierte Sequenzcluster, in denen nur der Teil der Fragmente mit der grössten Ähnlichkeit mit Datenbanksequenzen verglichen wird.