Experimenta Schülerlabor

iGEM-Team Tübingen 2019
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

September 2019

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Herzlich Wilkommen im Skript zum Schülerlaborkurs des iGEM Team Tübingen 2019 an der Experimenta in Heilbronn. Dieses Skript soll dir bei der Nachbereitung einzelner Themen des Kurses zur Seite stehen.

Alle Themen die wir in den kommenden Tagen behandeln werden haben wir auch hier im Skript verschriftlicht. An der Seite gibt es auch extra noch einen Rand für Notizen den du gerne während des Kurses nutzen kannst.

Wir wünschen dir viel Spaß mit dem Kurs und hoffentlich viele neue Ideen und Erkenntnisse.

2. Sequenzierung

Als Sequenzierung bezeichnet man den Prozess der Bestimmung der Nukleotid-Abfolge in einem DNA-Molekül, also die Reihenfolge von Adenin, Guanin, Thymin und Cytosin. Sequenzierungen sind aus der Genetik nicht mehr wegzudenken, insbesondere im Rahmen von DNA-Klonierungen in einem molekularbiologischen Laborbetrieb. Auch wird die DNA-Sequenzierung zusammen mit DNA-analytischen Verfahren zur Untersuchung von genetisch bedingten Erkrankungen herangezogen.

In den vergangenen Jahren wurden mehrere verschiedene Sequenzierungsverfahren entwickelt. Diese werden im Allgemeinen in drei verschiedene Generationen unterteilt.

2.1. Erste Generation

Der wichtigste Vertreter der Sequenzierungsverfahren der 1. Generation ist die Sanger Sequenzierung. Sie wurde von Sanger und Coulson 1975 entwickelt, wofür Sanger zusammen mit Gilbert und Berg 1980 den Nobelpreis für Chemie erhielt.

Der generell Ablauf der Sanger Sequenzierung, welche auch als Kettenabbruch-Synthese oder Didesoxymethode nach Sanger genannt wird, ist in Abbildung 1 dargestellt. Er lässt sich wie folgt zusammenfassen:

Ausgehend von einem Primer wird durch eine DNA-Polymerase einer der beiden komplementären DNA-Stränge verlängert. Daraufhin wird zunächst die DNA-Doppelhelix durch Erwärmung denaturiert. In vier gleichen Ansätzen mit allen vier Nukleotiden, wird je eine der vier Basen als Didesoxynukleosidtriphosphat (ddNTP) zugegeben (also je ein Ansatz mit entweder ddATP, ddCTP, ddGTP oder ddTTP). Diese Kettenabbruch-ddNTPs besitzen keine 3'-Hydroxygruppe: Werden sie in den neusynthetisierten Strang eingebaut, ist eine Verlängerung der DNA durch die DNA-Polymerase nicht mehr möglich, da die OH-Gruppe am 3'-C-Atom für die Verknüpfung mit der Phosphatgruppe des nächsten Nukleotids fehlt. Des Weiteren sind die ddNTPs mit einem Fluorophor gelabeled. Sie beginnen also, wenn sie von einem Laser angestrahlt werden, zu leuchten. In der Folge entstehen DNA-Fragmente unterschiedlicher Länge, die in jedem einzelnen Ansatz stets mit dem gleichen ddNTP enden (also je Ansatz nur mit A oder C oder G oder T). Dies produziert alle möglichen Längen der Ketten. Die Produkte werden nun auf einem single lane kapillaren Gel aufgetrennt und von einem Laser bestrahlt. Software erlaubt es nun vom entstehenden Chromatogramm die DNA Sequenz abzulesen.

Die Sanger Sequenzierung wird auch heute noch für kleinere Genome verwendet und ist ein Standardverfahren der Molekularbiologie. Ein Vorteil der Sanger Sequenzierung ist ihre sehr hohe Präzision (>99,9%), allerdings ist sie für Genome, welche länger als 1000 Basen Paare sind, nicht geeignet.

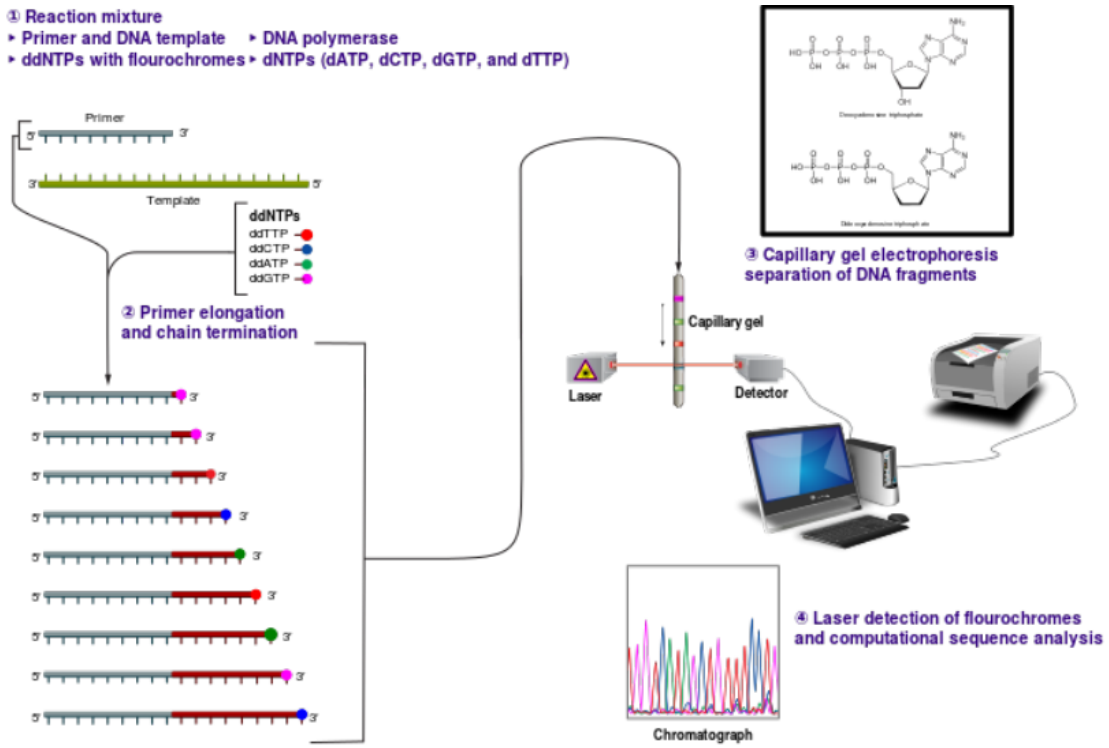

Abbildung 1: Prinzip der DNA-Sequenzierung nach der Didesoxy-Methode. dNTP ist die allgemeine Abkürzung für ein Nukleosidtriphosphat und kann für dATP, dCTP, dGTP oder dTTP stehen. ddNTPs sind die entsprechenden Didesoxy-Varianten der dNTPs. Der Einbau eines ddNTPs führt zum Abbruch der Polymerisationsreaktion. Die blauen Punkte am 5'-Ende des Primers stellen eine Markierung dar (z. B. eine fluoreszierende Gruppe), mittels der die Syntheseprodukte später im Gel sichtbar gemacht werden können. Da alle Fragmentlängen mindestens einmal gebildet werden, kann die vollständige DNA-Sequenz entschlüsselt werden.

2.2. Zweite Generation

Die Vertreter der 2. Generation sind neben der Sanger Sequenzierung die am Häufigsten verwendesten Sequenzierverfahren und bekannt unter dem Namen Next Generation Sequencing bekannt. Sie unterscheiden sich von der 1. Generation durch ihren extrem hohen Durchsatz und ihre Parallelisierbarkeit. Die Sequenzierverfahren sind für Genome aller Längen geeignet und somit auch für sehr große Genome, wie das des Menschen.

Die bekanntesten Vertreter der 2. Generation sind Roche 454 sequencing (basierend auf der *pyrosequencing* Technologie), ABIs SOLiD sequencing (basierend auf Hybridisierungen) und am Bekanntesten, Illumina sequencing. Im Folgenden wird nur das von Illumina entwickelte Sequenzierverfahren behandelt, da es am Weitesten verbreitet ist.

Das Kernprinzip der Illumina-Sequenzierung ist bekannt als Illumina sequencing. Die Illumina-Sequenzierung folgt dem Prinzip des Sequencing-by-Synthesis. Es

wird also die Synthese eines komplementären Stranges Nukleotid-für-Nukleotid verfolgt. Dies ist möglich, da spezielle Nukleotide verwendet werden. Diese dNTPs tragen ein fluoreszierendes Label (jedes mit einer anderen Wellenlänge). Wird ein Nukleotid eingebaut, bricht die Synthese ab, da das Label die Polymerase blockiert. Erst nachdem Enzyme das Label abspalten, kann die nächste Runde beginnen. Ein entscheidendes Charakteristikum der Illumina-Sequenzierung ist die Art der Signalverstärkung. Das Signal eines einzelnes DNA-Molekül wäre zu schwach um von einer Kamera detektiert zu werden. Deswegen wird vorher eine spezifische PCR durchgeführt. Die DNA wird dazu auf einer spezifischen Platte gebunden und vervielfältigt. Dadurch entstehen Cluster der identischen Sequenz, die ein messbares Signal erzeugen. Der gesamte Prozess ist in Abbildung 2 dargestellt.

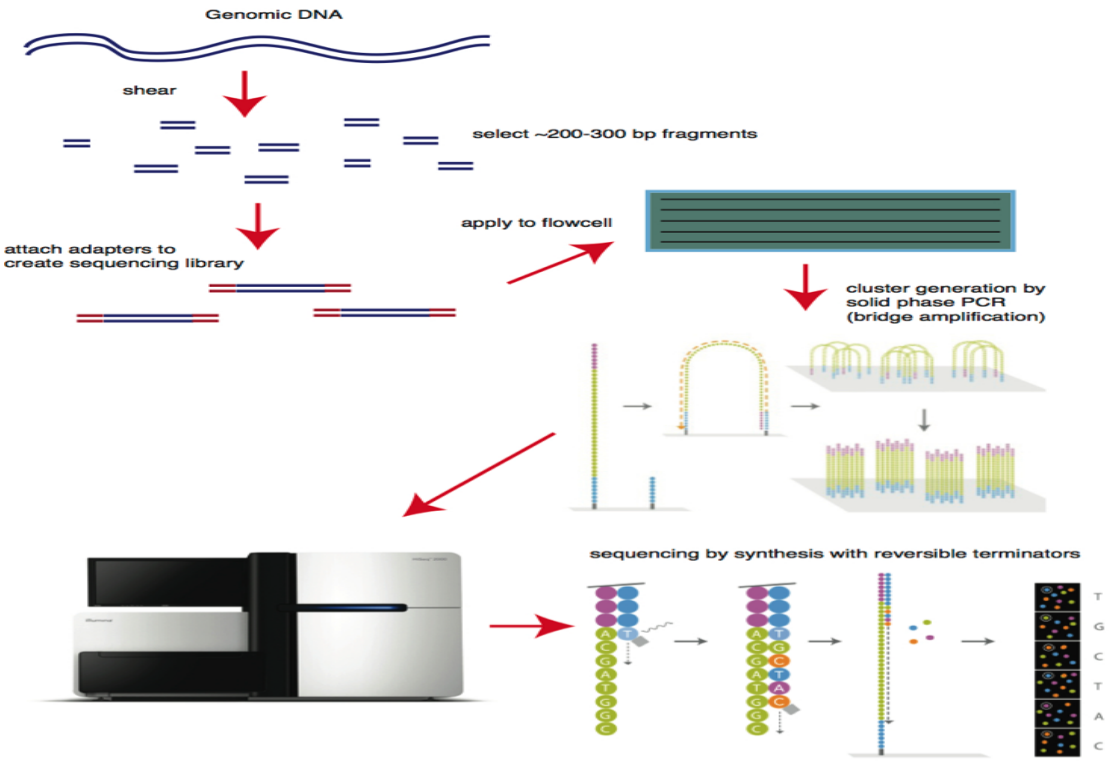

Abbildung 2: Die DNA wird durch mechanische Verfahren fragmentiert und es entstehen DNA-Fragmente mit einer Länge von ca. 300 bp. An beide Enden des Fragments werden nun Adapter ligiert, deren Sequenz komplementär zu den später verwendeten Primern ist. Die Fragmente werden nun über die Adapter auf einer Platte gebunden. Auf dieser Platte befindet sich außerdem ein dichter Rasen von ebenfalls gebundenen Primern, die komplementär zu den Adaptern sind. Die DNA-Fragmente "beugenßich mit ihrem freien Adapter zu einem Primer hin. Wegen dieser Beugung wird dies als Brücken-PCR bezeichnet. Eine Polymerase beginnt nun ab diesem Primer den komplementären Strang zu synthetisieren. Durch Denaturierung werden diese wieder getrennt. Es entstehen zwei komplementäre Einzelstränge, ein Forward-Strang und ein Reverse-Strang. Fällt bei der Sequenzierung eine eigenartige Sequenz auf, kann im Gegenstrang geschaut werden, ob diese dort ebenfalls präsent ist. Dieser Prozess der Amplifizierung wird mehrfach wiederholt. Dadurch entstehen überall auf der Platte unterscheidbare Cluster, die jeweils die identische Sequenz besitzen. Es werden nun ein Primer, die Polymerase und die gelabelten dNTPs zugegeben. Nach dem Einbau eines Nukleotids stoppt die Synthese. Ein Laser regt die fluoreszenten Labels an. Überall auf der Platte leuchtet jeweils das im Cluster eingebaute Nukleotid auf. Die Kamera speichert das Bild. Jeder der einzelnen Cluster wird von der Software registriert. Ein Enzym spaltet nun das Label ab und das zweite Nukleotid wird eingebaut. Dieser Vorgang wird wiederholt, bis das Fragment komplett synthetisiert ist.

Ein Vergleich der Sequenzierverfahren der 2. Generation ist in Tabelle 1 aufgeführt.

Tabelle 1: Vergleich der Sequenzierverfahren der 2. Generation.

	Roche 454	Illumina	ABI SOLiD
Template preparation	emPCR	solid-phase	emPCR
NGS Chemie	Pyrosequenzierung	Rervsible Terminatoren	Sequenzierung durch Ligation
Read Länge	330	75-300	35
Laufzeit in Tagen	0.35	9	10
Gigabyte pro Sequenziervorgang	0.45	35	50
Kosten pro Maschine	500.000$	540.000$	595.000$
Vorteile	Lange reads, schnelle Laufzeiten	Am meisten verwendente Plattform	2-Basen Kodierung kann Fehler korrigieren
Nachteile	Hohe Kosten der Reagenzien, hohe Fehlerraten in homopolymer repeats	Schwierigkeiten viele Proben auf einmal zu Sequenzieren	Hohe Laufzeiten

Das Ergebnis einer Sequenzierung der 2. oder 3. Generation ist eine Sammlung von sequenzierten DNA-Fragmenten, sogenannte reads. Diese werden im Allgemeinen in sogenannten FASTQ Dateien gespeichert. Diese werden im Kapitel 2.4 eingeführt. Die Länge der reads unterscheidet sich je nach Sequenziertechnologie. Im Allgemeinen ist eine möglichst hohe read Länge wünschenswert. Jedoch sind mit einer höheren read Länge bisher noch höhere Fehlerraten verbunden. Beide Probleme, sowie die Notwendigkeit von PCR Schritten (Mögliche Fehlerquelle) sind die Motivation für die 3. Generation von Sequenziertechnologien.

2.3. Dritte Generation

Eine dritte Generation von Sequenzierungstechniken steht seit kurzem zur Verfügung: Sequenzer, die von Pacific Biosystems (PacBio) vertrieben werden, verwenden Single Molecule Real-Time Sequenzierung, während Oxford Nanopore Technologies (ONT) ein Gerät zur Nanopore Sequenzierung entwickelt hat. Im Gegensatz zu Techniken der zweiten Generation (454, Illumina, Ion Torrent) beinhalten diese Verfahren keine Amplifikation während der library preparation und ermöglichen somit die Sequenzierung einzelner Moleküle.

Für SMRT befestigt man ein einzelnes DNA Molekül am Boden eines Zero Mode Waveguide (ZMW), dieses wird dann von unten durch einen Laser beleuchtet. Die Nanostrukturierung des ZMW sorgt für eine Einschränkung der Anregung innerhalb eines Zeptoliters. Dies ermöglicht die Detektion einzelner mit phospholinked Nukleotide während diese durch die Polymerase in den DNA Strang eingebaut werden. Dieser Prozess ist in Abbildung 3 dargestellt. Nach dem Einbau des phos-

pholinked Nucleotides wird das Farbmolekül durch die DNA Polymerase entfernt wenn auch die Phosphatkette geschnitten wird. Dies kann danach wegdiffundieren und hinterlässt somit ein DNA Stück komplett ohne Label. Die Sequenzierung läuft also ohne Stops.

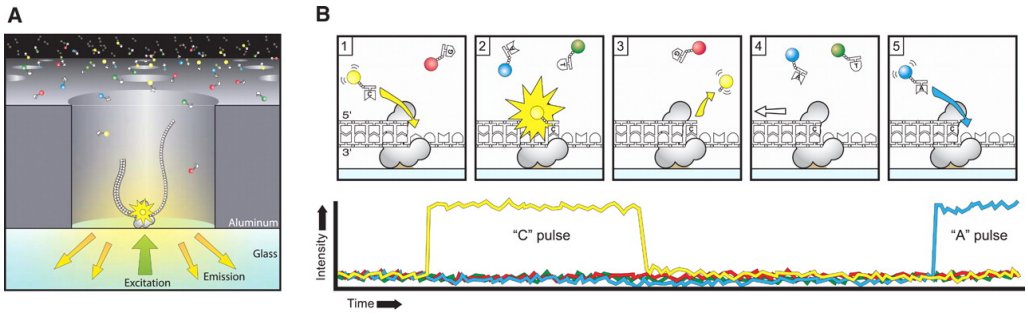

Abbildung 3: Prinzip von Single Molecule Real-Time Sequenzierung. (A) Experimenteller Aufbau. (B) Funktionsschema des Einbaus der phospholinked dNTPs mit korrespondierendem time trace der detektierten Fluoreszenzintensität. Quelle: Eid, John, et al. "Real-time DNA sequencing from single polymerase molecules."*Science* 323.5910 (2009): 133-138.

Bei der Nanopore Sequenzierung wird ein DNA Strang durch eine Pore mit einem Durchmesser von < 5 nm durchgezogen. Die Basen werden einzeln abgelesen, basierend auf der Amplitude eines gemessenen Stroms. Dadurch kann die Sequenz der DNA einfach gescannt werden, ohne dass eine Amplifikation durch PCR oder chemische Markierung erforderlich ist.

Für die Ionenstromdetektion wird eine Spannung an die Membran angelegt, die in eine Elektrolytlösung eingebracht wird, die zu einer Ionenbewegung durch die Poren führt (siehe Abbildung 4). Die DNA-Stränge werden mit einem prozessualen Enzym vermischt, das einen DNA-Protein-Komplex bildet. Wenn sich dieser Komplex der Nanopore nähert, öffnet das Enzym die DNA, die dann durch die Öffnung in der Membran gezogen wird. Ein Nukleotid nach dem anderen wird in die Pore geleitet und erzeugt charakteristische Störungen im elektrischen Strom, da es die Anzahl der durch die Pore wandernden Ionen beeinflusst.

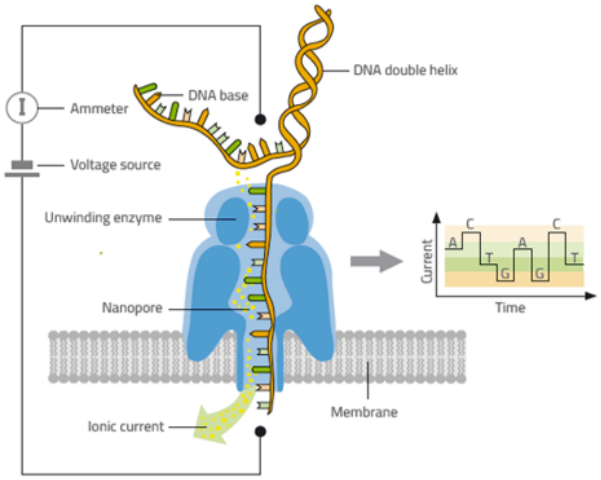

Abbildung 4: Ionenstrom Detektion. Quelle: `https://www.scienceinschool.org/content/decoding-dna-pocket-sized-sequencer` - 08.02.19

Tabelle 2: Vergleich der zwei verschiedenen third generation Sequenzierungsverfahren.

	SMRT	**MinION**
+	• Longest average read lengths (∼ 30 kbp, up to 500 kbp)	• Ultra-long reads (up to 200 kbp)
	• 2-4 nucleotides per s per cell	• 450 bases per s per nanopore per MinION Flow Cell
	• Highest consensus accuracy	• Low cost & Label-free
	• Uniform coverage	• Low sample volume & High yields
	• Simultaneous epigenetic characterization	• High throughput reads (REAL Real-time)
−	• Short reads make assembly & related analyses difficult	• Fast passage of DNA
	• Utilizes hairpin adaptors for sequencing double-stranded DNA, which can result in chimeric reads if the sequencing reaction processes both strands of the DNA	• Error rate of 2-13%
>	• Assembly results expected to improve as the read lengths of the technology improve	• Slow down passage of DNA while maintaining high signal-to-noise ratio

2.4. FASTQ Dateien

In der Bioinformatik (siehe Kapitel 3) sind standartisierte Dateiformate eine Notwendigkeit. Dies ermöglicht es allen Wissenschaftlern ihre generierten Daten auszutauschen und alle Software Tools auf diesen anzuwenden.
Das FASTQ Format ist ein text basiertes Format für Nukleotidsequenzen und ihre dazugehörigen Qualitätswerte. Ein Beispiel für eine FASTQ Datei ist in Abbildung 5 zu sehen.

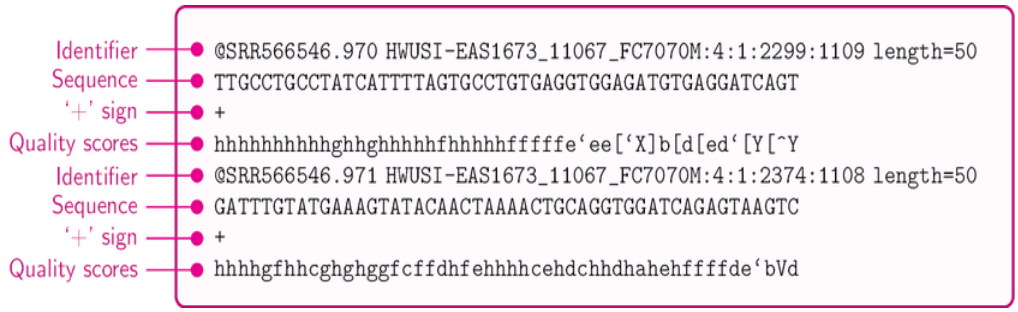

Abbildung 5: Beispiel einer FASTQ Datei, die von einem Sequenziergerät der 2. oder 3. Generation generiert wird.

Im Detail:

1. Zeile 1 startet mit einem '@' Buchstaben gefolgt von einem Sequenzidentifier und einer optionalen Beschreibung
2. Zeile 2 beinhaltet die rohe DNA Sequenz
3. Zeile 3 startet mit einem '+' Buchstaben und ist optional gefolgt von demselben Sequenzidentifier und einer optionalen Beschreibung
4. Zeile 4 codiert die Qualitätswerte für die Sequenz in Zeile 2 im ASCII Format

3. Bioinformatik

In der Bioinformatik verschmilzen Biologie, Informatik und Informationstechnologie zu einer einzigen Disziplin welche Methoden und Werkzeuge entwickelt, die zum Verständnis insbesondere der molekularen Mechanismen des Lebens auf der Erde beitragen. Der Begriff "Bioinformatik"wurde bereits 1970 von Ben Hesper und Pauline Hogeweg definiert als "das Studium informatischer Prozesse in biotischen Systemen", um die Bioinformatik als Disziplin von der Biophysik und Biochemie zu unterscheiden.

Das ultimative Ziel des Feldes ist es, die Entdeckung neuer biologischer Erkenntnisse zu ermöglichen und eine globale Perspektive zu schaffen, aus der heraus vereinheitlichende Prinzipien in der Biologie erarbeitet werden können. Zu Beginn der "genomischen Revolution"war ein großes bioinformatisches Anliegen die Erstellung und Pflege einer Datenbank zur Speicherung biologischer Informationen wie Nukleotid- und Aminosäuresequenzen.

Heute hat sich der Bereich der Bioinformatik so entwickelt, dass der Schwerpunkt auf der Entwicklung und Anwendung von Methoden zur Analyse und Interpretation verschiedener Arten von Daten liegt, darunter Nukleotid- und Aminosäuresequenzen, Proteindomänen und Proteinstrukturen. Der eigentliche

Prozess der Analyse und Interpretation von Daten wird als Computational Biology bezeichnet.

3.1. BLAST

Basic Local Alignment Search Tool (BLAST) ist das vermutlich am weitesten verbreitete bioinformatische Werkzeug welches jemals geschrieben wurde. Es handelt sich um eine sogenannte alignment heuristic die 'lokale Übereinstimmungen' zwischen einer Abfragesequenz und Sequenzen in einer Datenbank bestimmt und wird hauptsächlich zur Suche nach paralogen und orthologen Genen und Proteinen innerhalb eines oder mehrerer Organismen verwendet.

BLAST kann sowohl als Commandline Tool als auch als webbasierte Anwendung verwendet werden. Unter `https://blast.ncbi.nlm.nih.gov/Blast.cgi` ist der BLAST-Algorithmus öffentlich zugänglich. Abbildung 6 zeigt die Weboberfläche von BLAST.

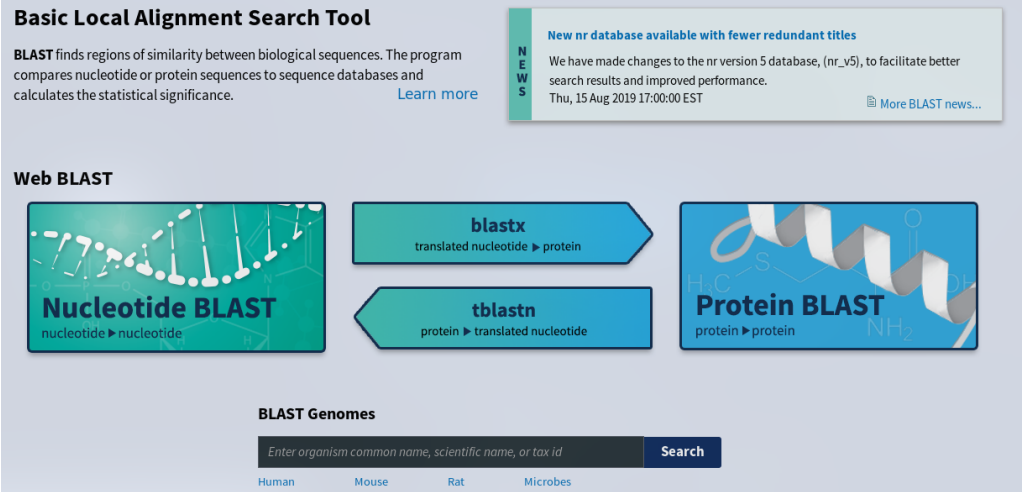

Abbildung 6: Screenshot der landing page von BLAST. Hier können die verschiedenen Typen von BLAST je nach gewünschter Anwendung ausgewählt werden.

Über Score und E-Wert wird die Homologie der bearbeiteten Abfragesequenz mit der Datenbank bestimmt: Der Score ist eine quantitative Bewertung der Ähnlichkeit der Suchsequenz mit einer bekannten Sequenz (je höher, desto höher ist auch die Identität der Sequenzen) und der E-Wert gibt die erwartete Anzahl der Hits an, deren Score mindestens so groß ist wie der beobachtete (je kleiner, desto besser).

Es gibt außerdem verschiedene BLAST Typen, diese sind in Abbildung 7 dargestellt.

Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses
BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein
TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

Abbildung 7: Typische BLAST Programme. Quelle: `http://etutorials.org/Misc/blast/Part+III+Practice/Chapter+5.+BLAST/5.1+The+Five+BLAST+Programs/`

3.2. Galaxy

Bioinformatiker arbeiten im Alltag in der Regel mit sogenannten Commandline Tools. Vorteile dieser Tools sind ihre Möglichkeiten sie auf Hochleistungsrechnern zu nutzen, die einfache Integration in Datenanalysepipelines und viele mehr. Diese verwenden keine graphische Oberfläche und sind somit für Laien wenig geeignet.

Galaxy ist eine webbasierte Platform mit dem Ziel Software Tools der Bioinformatik für Wissenschaftler bereitzustellen, die keine Programmiererfahrungen besitzen.

Galaxy erlaubt es einzelne Tools auf Datensätzen anzuwenden und die Ergebnisse online zwischenzuspeichern, um sie für weitere Tools zu verwenden.

Die europäischen Server sind unter `https://usegalaxy.eu/` erreichbar. Abbildung 8 zeigt die Weboberfläche von Galaxy.

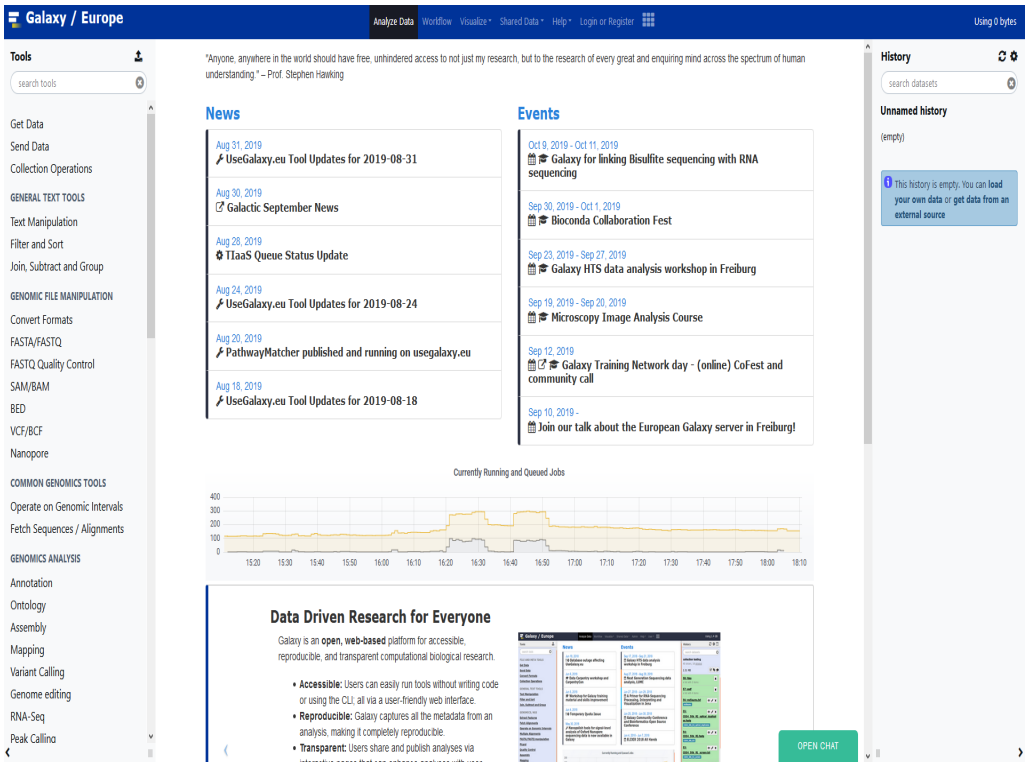

Abbildung 8: Screenshot der landing page von Galaxy. Links oben kann nach den spezifischen Tools gesucht werden. Alle ausgeführten Tools und Ergebnisse finden sich rechts in der history.

3.3. DNA-Seq Datenanalyse

Die im Labor generierten Sequenzdaten müssen von Bioinformatikern analysiert werden.
Häufige Fragestellungen sind hierbei:

1. Wie hoch ist die Qualität der generierten Daten? War die DNA verunreinigt?
2. Können Gene detektiert werden?
3. Gibt es auffällige SNPs, also einzelne ausgetauschte Basen in der DNA, die mit einer Krankheit in Verbindung stehen können (Siehe Sichelzellenanämie)?
4. Gibt es ein erhöhtes Risiko für eine spezifische genetische Krankheit?
5. Welche Organismen sind in einer Wasser oder Bodenprobe vorhanden?
6. Wie hat sich die DNA eines Organismuses im Laufe der Evolution verändert?
7. Viele Weitere...

Die allerwichtigsten und groben Schritte sollen im folgenden Abschnitt kurz erwähnt und eingeführt werden.

Abbildung 9 zeigt einen beispielhaften Workflow einer Datenanalysepipeline in der Bioinformatik für DNA-Sequenzen. Zuallererst, muss die Qualität der Reads überprüft werden. Durch technische Fehler, sowie biologische Variationen kann die Qualität der reads leiden. Häufig wird hierfür das Tool FastQC [1] verwendet. Daraufhin muss aus den kurzen reads die gesamte DNA wieder rekonstruiert werden. Dies ist auch als das assembly Problem bekannt (siehe Abbildung 10). Hierfür gibt es zwei Ansätze:

1. Reference-free oder auch bekannt als de-novo: Ohne ein bereits bekanntes Referenzgenom
2. Reference-based oder auch bekannt als mapping: Mit einem bekannten Referenzgenom

Im Allgemeinen ist die Variante mit einem Referenzgenom zu bevorzugen, da hierfür die Fehlerrate deutlich geringer ist. Auch dieser Prozess, der als mapping bekannt ist, sollte einer Qualitätskontrolle unterliegen. Die weiteren Schritte sind variant calling, sowie somatic mutation Detektion. Diese beiden Prozesse beschreiben die Identifikation von auffälligen Unterschieden von den sequenzierten Daten und einer Referenz dar. Hierbei wird unterschieden zwischen erblich bedingten Mutationen und durch Umwelteinflüssen Mutationen (somatische Mutationen). Schlussendlich werden die identifizierten Varianten mit ihren bekannten Funtionen assoziiert.

[1] https://www.bioinformatics.babraham.ac.uk/projects/FastQC/

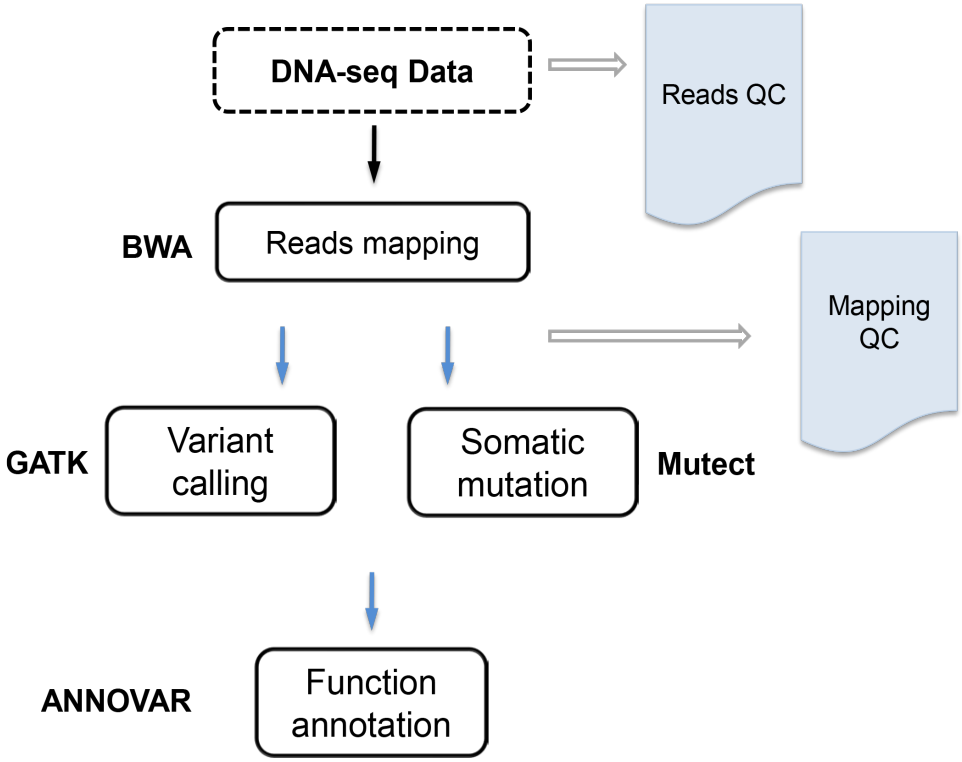

Abbildung 9: Beispielhafter Workflow für DNA-Sequenz Daten.
1. Kontrolle der Qualität der reads (FastQC)
2. Mapping der reads auf ein Referenzgenom, falls vorhanden
3. Kontrolle der Qualität des Mappings (Qualimap)
4. Variant Calling - Identifizierung von Varianten in Genomen
5. Somatic Mutation - Identifizierung von somatischen Mutationen (Mutationen, die nicht geerbt wurden)
6. Function Annotation - Annotation der gefundenen Mutationen mit ihren Funktionen und Verantwortlichkeiten
Die Tools, die für die spezifischen Schritte verwendet werden können sind fett neben den Kästen vermerkt.

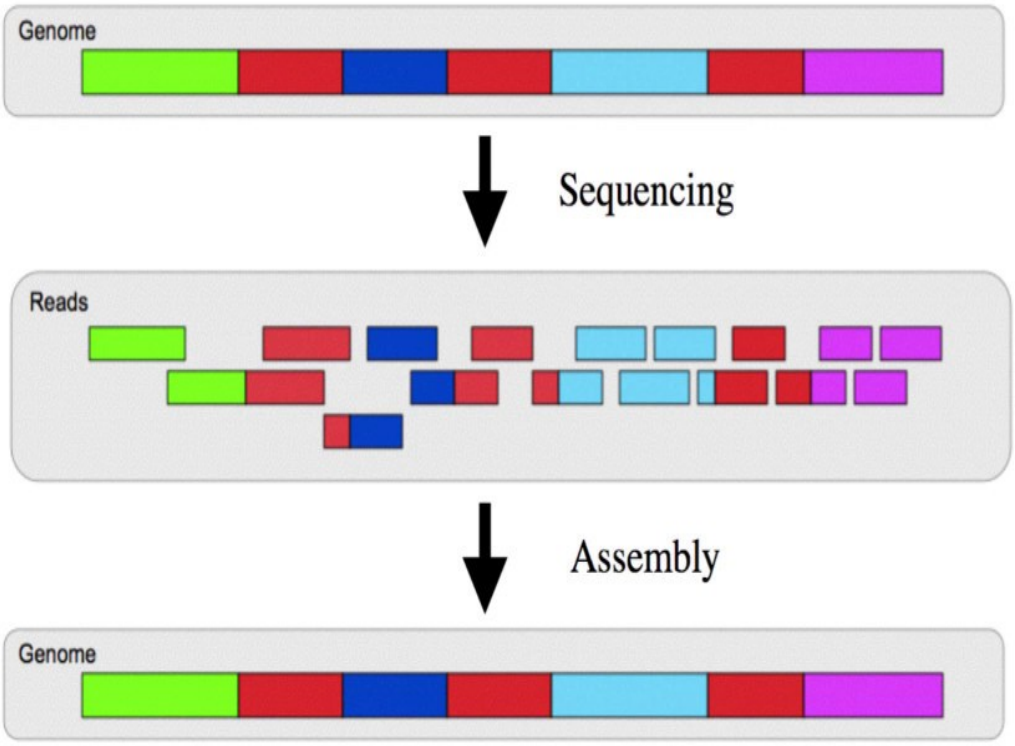

Abbildung 10: Das Assembly Problem. Aus den generierten Sequencing reads muss die vollständige DNA-Sequenz wieder rekonstruiert werden.

3.4. FastQC

FastQC [2]ist ein Software Tool für die Kontrolle der Qualität von reads. Ein Beispiel Report ist in Abbildung 11 zu sehen. Der Report wird als html Datei für den User bereitgestellt. Die Qualität wird anhand von 3 Symbolen (grüner Check, gelbes Ausrufezeichen, rotes X) einfach bewertet.
Für die Interpretation der Ergebnisse können folgende Links hilfreich sein:

1. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/`
2. https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

[2]`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

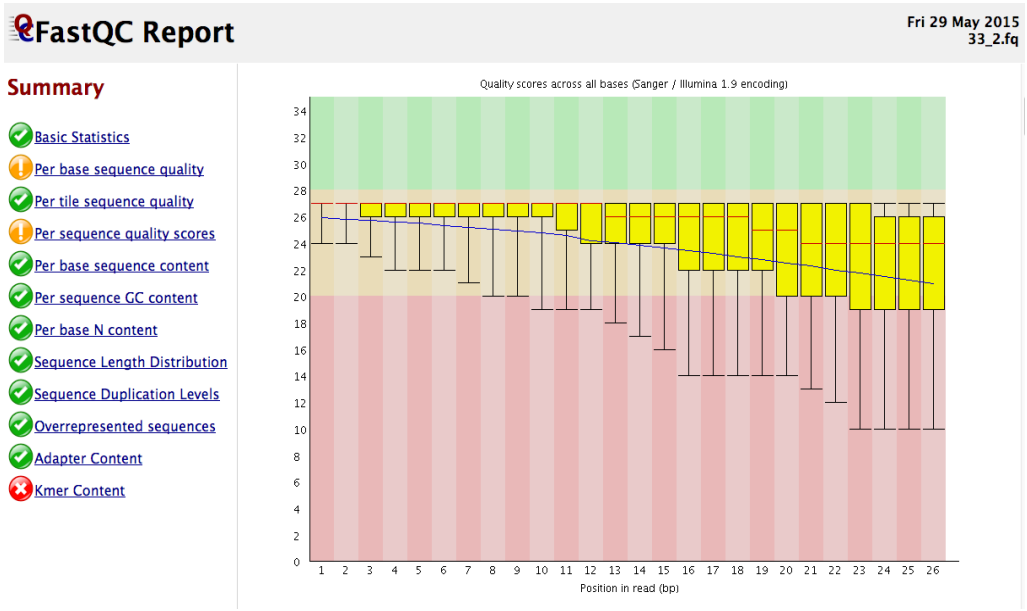

Abbildung 11: Beispiel Report von FastQC für generierte reads.

3.5. Mapping

Mapping beschreibt den Prozess der Rekonstruktionen der Positionen von generierten reads zu einem Referenzgenom. Das Ergebnis eines Mapping Prozesses ist in Abbildung 12 zu erkennen. Dies ist extrem wichtig um in darauffolgenden Schritten Unterschiede zum Referenzgenom zu entdecken. Gerade die Unterschiede sind zumeist die für Wissenschaftler und Ärzte besonders interessante Stellen im Genom.

Das Ergebnis eines Mapping ist ein sogennantes SAM-File. Der CIGAR String beinhaltet die nötigen Operationen, wie Insertionen oder Deletionen, um das Mapping zu generieren. Abbildung 14

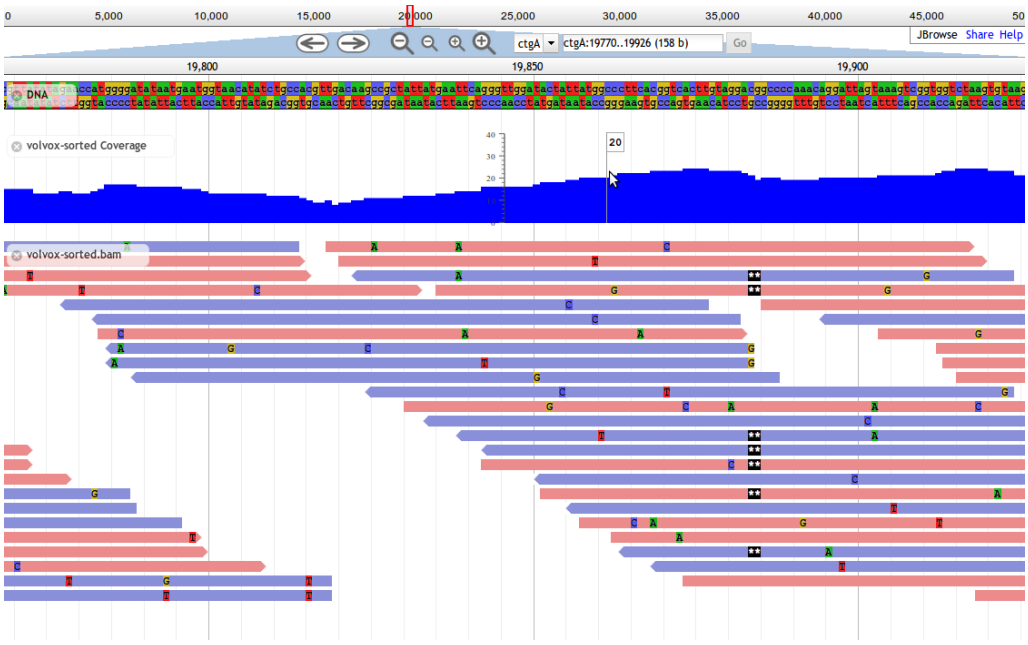

Abbildung 12: Visualisierung eines Mappings. Insertionen oder Deletionen im Vergleich zur Referenz sind mit Buchstaben annotiert.

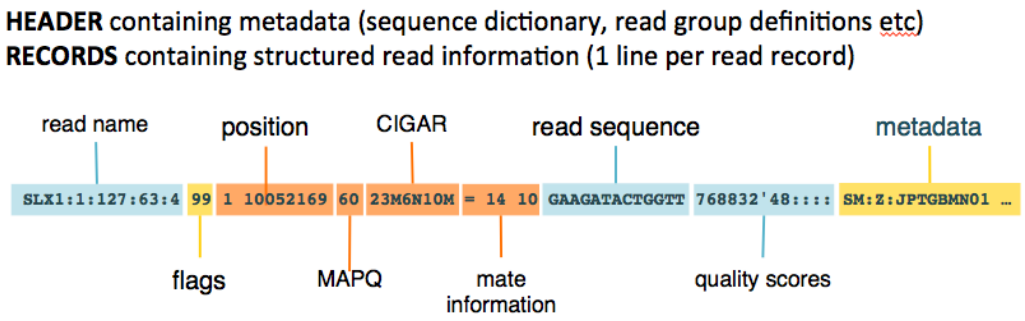

Abbildung 13: Struktur einer SAM Datei. Eine SAM Datei enthält mehrere Einträge mit dem read Namen, der Position bei welcher der read auf das Referenzgenome mapped, die read Sequenz und den CIGAR String (siehe Abbildung 14)..

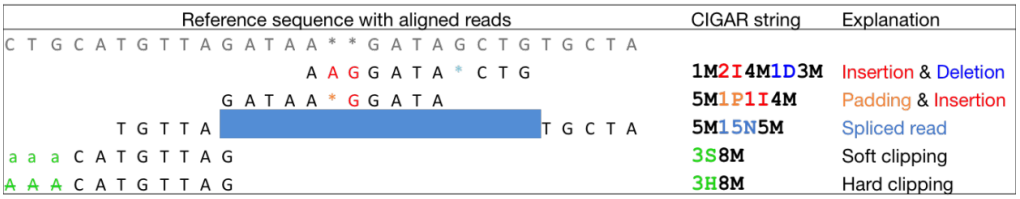

Abbildung 14: CIGAR String. Der CIGAR String enthält die nötigen Operationen um einen read auf eine Position im Genom zu mappen. Operationen können sein:
1. Matches
2. Missmatches
3. Insertionen
4. Deletionen
5. Viele weitere...

4. Primer Design

4.1. Was sind Primer?

Primer sind kurze DNA-Sequenzen, so gennante Oligonukleotide, die in der Regel aus 15-30 Nukleotiden aufgebaut sind. In einer PCR (polymerase chain reaction) dienen sie als Startpunkt für die Polymerase, da sie die von der Polymerase zum Starten benötigte freie Hydroxygruppe bereitstellen. Primer sind ein essentieller Bestandteil jeder PCR und erlauben es, durch ihre spezifische Bindung an bestimmte DNA-Regionen, ausgewählte Bereiche einer DNA-Sequenz zu amplifizieren.

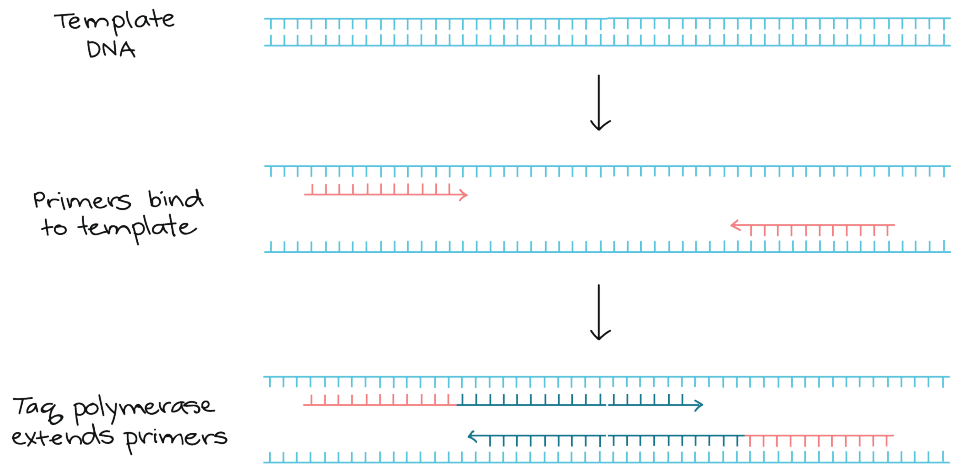

Abbildung 15: Die Abbildung zeigt das Prinzip der Primer in einer PCR. Die template DNA wird aufgeschmolzen und die Primer binden. Sie dienen dann als Startpunkt für die Polymerase. Quelle: `https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/polymerase-chain-reaction-pcr`

4.2. Worauf ist beim Primer Design zu achten?

Primer Design bezeichnet den gezielten Entwurf von Primern für eine Zielsequenz. Hierbei sind einige Dinge zu beachten, damit die anschließende PCR erfolgreich abläuft. Zunächst müssen die erstellten Primer die gewünschte Zielsequenz flankieren, d.h. sie müssen zu beiden Seiten des Zielsequenz binden und zwar an beiden Strängen. Die Primer nennt man dann meistens "forward” und "reverse” Primer. Die Länge der Primer ist ebenfalls entscheidend, da zu kurze Primer die Gefahr bergen noch an anderen Stellen in der Sequenz zu binden, was zu unspezifischen Amplifikationen führt. Primer haben meistens eine Länge von 15-30 Basen können je nach Fragestellung und Zielsetzung aber auch länger sein. PCR-Primer sollten in ihrer Sequenz einen Anteil von Guanin und Cytosin

von 45-60% haben und keine intramolekularen Sekundärstrukturen ausbilden. Zur Überprüfung von eventuellen Sekundärstrukturen stehen online-tools und Datenbanken zur Verfügung. Neben den Sekundärstrukturen sollten die erstellten Primer auch nicht aneinander binden, also nicht komplementär sein. Viel Guanin und Cytosin (3x G oder 3x C) sollten am 3'-Ende der Primer vermieden werden. Ein letzter wichtiger Aspekt ist darauf zu achten, dass die Schmelztemperatur (Tm) der beiden Primer ungefähr gleich ist (Abweichung von bis zu 5°C) und sich im Bereich von 45-65°C befindet.

Beachte:

1. Länge 15-30 Basen
2. Orientierung / Flankieren
3. GC-Gehalt (40-60%)
4. Schmelztemperatur (45-65°C, 5°C Abweichung)
5. Nicht komplementär
6. Keine Sekundärstrukturen

4.3. Was geht mit Primern noch?

Neben den üblichen PCR-Primern kann man mit dem richtigen Primer-Design auch gezielt Punktmutationen in eine Sequenz einbringen. Dazu wird im Primer eine Base verändert, sodass der Primer nur an dieser Stelle nicht komplementär zur Zielsequenz ist. Im Laufe der PCR entstehen dann immer mehr Amplifikate, die die veränderte Sequenz beinhalten. Eine weitere Möglichkeit ist die Nutzung von verlängerten Primern, die noch die Sequenz für eine Restriktionsstelle beinhalten. Somit kann man die Zielsequenz mit Restriktionsstellen flankieren, die zum Beispiel für eine Klonierung (Einbringung dieser Sequenz in ein Plasmid oder Ähnliches) genutzt werden können. Hierbei ist darauf zu achten nach der Schnittstelle noch 4-6 Basen extra anzuhängen, damit das Restriktionsenzym auch binden kann. Dieser Primerüberhang sollte nicht unmittelbar mit der Zielsequnz komplementär sein.

Abbildung 16: Zwei weitere Anwendungsmöglichkeiten von Primern sind das gezielte Einbringen von Punktmutationen (oben), sowie das Versehen der zu amplifizierenden Sequenz mit Restriktionsschnittstellen (unten). Beim Einbringen von Punktmutationen stimmt die Primersequenz an einer Stelle nicht mit der Zielsequenz überein wodurch eine falsche Base eingebaut wird. Durch die mehreren PCR-Zyklen entsteht so eine an einer Stelle veränderte Sequenz. Verlängert man Primer und versieht sie mit Restriktionsschnittstellen, kann die neu entstehende vermehrte DNA-Sequenz gezielt mit Restriktionsenzymen geschnitten werden, eine Methode, die besonders in der Klonierung Anwendung findet.

5. CRISPR/Cas

5.1. Was ist CRISPR/Cas?

Die CRISPR/Cas Methode ist heutzutage eine weit verbreitete molekularbiologische Methode zur gezielten Veränderung von DNA und hat verschiedene Anwendungen. Das System beruht auf dem gleichnamigen System, das Bakterien zum Schutz vor Viren, Phagen und anderer eindringender Nukleinsäuren benutzen. CRISPR steht für Clustered Regularly Interspaced Short Palindromic Repeats und bezeichnet in Bakterien einen Abschnitt in der DNA der von vielen Repeats mit dazwischen liegenden variablen Regionen gekennzeichent ist. In unmittelbarer Nähe zu dieser CRISPR-Region befinden sich zusätzlich die Cas-Gene (CRISPR associated), die für Proteine, wie Cas9, kodieren.

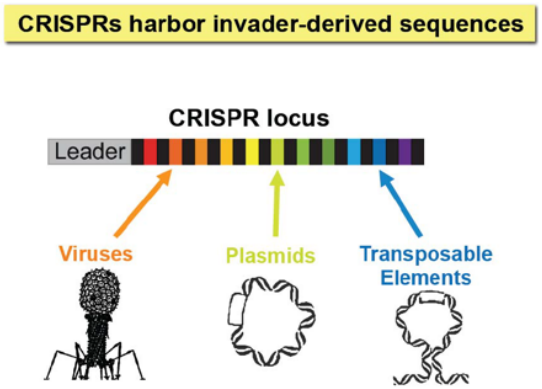

Abbildung 17: Das aus Bakterien stammende CRISPR/Cas System fungiert in Bakterien als Schutz vor eindringender Nukleinsäure von zum Beispiel Viren oder Phagen. Dringen diese in ein Bakterium ein werden kleine Fremd-DNA Stücke in das bakterielle Genom in den CRISPR-Locus eingebaut und sorgen somit später für eine Art Immunität gegen die Fremd-DNA. Quelle: http://www.bmb.uga.edu/mterns/bcmb8020/lectures/ppts/006%20%208020%20Transc.%20cis%20trans.pdf

5.2. Mechanismus

Wenn Fremd-DNA in ein Bakterium eingeschleust wird, z.B. durch Viren oder Phagen, besteht der erste Schritt des CRISPR/Cas System darin, Teile der eindringenden DNA zu ”akquirieren". Dazu schneiden einige der Cas-Proteine einen Teil aus der Fremd-DNA heraus. Anschließend wird dieses DNA-Stück in das bakterielle Genom in den CRISPR-Bereich zwischen zwei repeat-Sequenzen eingefügt. Das herausgeschnitte DNA-Fragment nennt man "Protospacer” und sobald es im CRISPR-Lokus zwischen den repeats eingebaut ist ”Spacer".

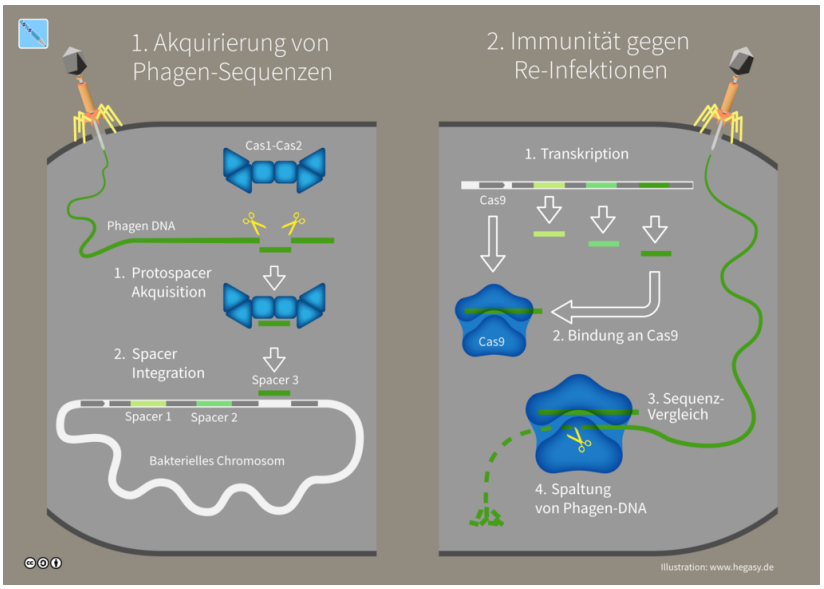

Abbildung 18: Darstellung des Mechanismus Des CRISPR/Cas Systems. Eindringende Fremd-DNA wird in einem ersten Schritt von Cas-Proteinen geschnitten. Die dabei entstehenden kurzen DNA-Fragmente werden als "Protospacer" in den CRISPR-Locus zwischen zwei repeat Sequenzen eingebaut. Diese "Spacer" werden in Folge abgelesen und bilden zusammen mit den Cas-Proteinen einen RNA-Protein-Komplex. Trifft dieser Komplex erneut auf die Fremd-DNA, aus der die Spacer-Region herausgeschnitten wurde, kann die Spacer-DNA an die Fremd-DNA komplementär binden und somit die Cas-Nuklease aktivieren. Als Folge wird die Fremd-DNA zerschnitten und unschädlich gemacht. Quelle: Von Guido4 - Eigenes Werk, CC BY-SA 4.0, `https://commons.wikimedia.org/w/index.php?curid=63789039`

Die CRISPR-Sequenzen und die Cas-Gene werden nun von der RNA-Polymerase dauerhaft abgelesen und die ”Spacer-repeat”-Regionen bilden zusammen mit einer so genannten tracrRNA und einem Cas-Protein, z.B. Cas9, einen RNA-Protein-Komplex. Die repeat-Sequenz und die tracrRNA dienen zur Bindung des Cas9-Proteins, wohingegen die ”Spacer”-Sequenz als spezifische Sequenz den Komplex zur Zielsequenz leitet. Trifft der Komplex erneut auf die Fremd-DNA, aus der die ”Spacer”-Sequenz ursprünglich herausgeschnitten wurde, bindet sie an die komplementäre Sequenz und das Cas-Protein, eine Nuklease, wird aktiv und schneidet die DNA. Als Einschränkung muss jedoch unmittelbar neben der Zielsequenz, die von der ”Spacer”-Sequenz gebunden wird eine so genannte PAM-Sequenz (Protospacer adjacent motif) liegen. Diese besteht für Cas9 aus drei Nukleotiden, einem belibigen und zwei Guaninen (NGG). Befindet sich so eine Sequenz unmittelbar neben der Zielsequenz schneidet die Nuklease drei Basen vor der PAM-Sequenz. Die Fremd-DNA wird so unschädlich gemacht und das Bakterium schützt sich durch diese Immunsystem-artige Funktion.

5.3. gRNA, Anwendungen & Typen

In der Forschung wird das CRISPR/Cas System vielseitig genutzt. Dazu werden so genannte guide RNA (gRNA) eingesetzt. Eine gRNA ist eine Fusions-RNA aus der crRNA (”Spacer”-Sequenz) und der tracrRNA. Die crRNA besteht aus 20 Nukleotiden und muss komplementär zu einer Region in der Zielsequenz sein, die unmittelbar neben einer PAM-Sequenz liegt. Prinzipiell ist somit jede DNA-Sequenz modifizierbar solange eine PAM-Sequenz vorhanden ist.

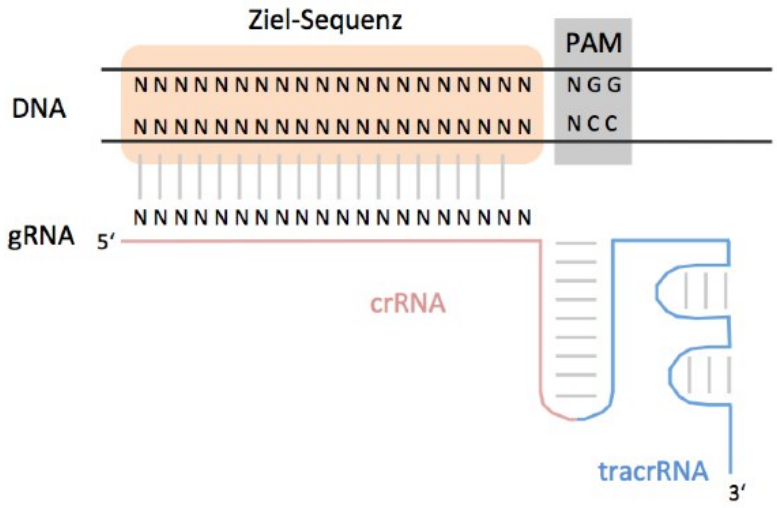

Abbildung 19: Aufbau einer gRNA. Die gRNA (guide RNA) besteht aus 20 Nkleotiden, die spezifisch für ihre Zielsequenz sind und die crRNA bzw. Spacer DNA repräsentieren. Im Gegenstz zu den normalen Umständen in BAkterien ist die gRNA eine Fusions-RNA aus crRNA und tracrRNA. Sie bindet an die Zielsequenz unmittelbar in der Nähe von PAM-Sequenzen (Protospacer adjacent motif). Prinzipiell ist mit einer gRNA jede DNA-Sequenz als target möglich solange auf die Präsenz der PAM-Sequenz geachtet wird. Quelle: `https://flexikon.doccheck.com/de/CRISPR/Cas-System`

Mit Hilfe der Nuklease Cas9, die einen Doppelstrangbruch in der Zielsequenz verursacht, lassen sich Mutationen, Insertionen und Deletionen in der Zielsequenz verursachen. Dies geschieht vor allem dadurch, dass der DNA-Reparaturmechsnismus bei einem Doppelstrangbruch fehleranfällig ist und teilweise Basen wegschenidet oder neue einsetzt um den Bruch zu reparieren. Eine weitere Möglichkeit ist durch zusätzliches Einbringen einer ”Template”-DNA gezielt ein größeres DNA-Fragment, was auch aus einem anderen Organismus stammen kann, in die Schnittstelle einzubringen. Durch Modifikation des Cas-Proteins kann auch ein Cas-Protein hergestellt werden, welches nicht mehr in der Lage ist DNA zu schneiden. In diesem Fall wird der Komplex aus gRNA und Cas-Protein zur Zielsequenz geleitet und bindet dort die DNA ohne sie zu schneiden. Kombiniert man so ein Cas-Protein mit einem Aktivator-Protein können die in der Nähe gelegenen Gene aktiviert werden oder durch Kombination mit einem Fluoreszenzprotein die

Bindestelle sichtbar gemacht werden.

Es gibt verschiedene CRISPR/Cas Typen. Das meistgenutzte Cas9-System gehört zu Typ II und fügt einen Schnitt in doppelsträngige DNA ein. Das Cas3-System gehört hingegen zu Typ I und bewirkt bei Bindung und Aktivierung an DNA den kompletten Abbau der DNA. Aus diesem Grund wollen wir in unserem diesjährigen Projekt das Cas3-System nutzen um es als Biosicherheitssystem zu verwenden. Sobald unser Bakterium die von uns vorgesehene Umgebung verlässt, soll das Cas3-System aktiviert werden und die bakterielle DNA abbauen.

6. iGEM Wettbewerb

6.1. Was ist iGEM?

International Genetically Engineered Machine (iGEM) ist ein internationaler Wettbewerb im Bereich der synthetischen Biologie, der auf Initiative des Massachusetts Institute of Technology in Boston ins Leben gerufen wurde. Ziel des Wettbewerbs ist es, innovative und intelligente Lösungen für relevante und aktuelle Probleme zu entwickeln.

Die Teilnahme bei iGEM ermöglicht es unserem Team das in der Universität vermittelte Wissen interdisziplinär und eigenständig anzuwenden. Nicht selten kommt es vor, dass aus den Projekten wissenschaftliche Publikationen entstehen, beispielsweise bei unserem Projekt 2017, oder Start-Ups gegründet werden. Im Rahmen von zahlreichen Symposien auf internationaler Ebene - innerhalb Europas und darüber hinaus - und unserem Engagement für Öffentlichkeitsarbeit, Lehre und interfakultäre Vernetzung, vertreten wir die Universität Tübingen, sowie die Mathematisch-Naturwissenschaftliche Fakultät. Final werden alle Projekte der teilnehmenden Teams bei einem großen Kongress in Boston, dem Giant Jamboree mit mehreren Tausend Teilnehmern, vorgestellt.

Das iGEM Team Tübingen stellt seit 2012 jährlich ein iGEM Team, das durch seine herausragende Leistung in der Lage ist die Eberhard-Karls-Universität Tübingen, Baden-Württemberg und Deutschland erfolgreich zu repräsentieren.

6.2. Projekt iGEM Team Tübingen 2019

Ziel des Projekt 2019 ist es, ein Probiotikum für die Therapie von Typ II Diabetes mellitus zu entwickeln, basierend auf einem Plasmid-Chassis mit einem Cas3-Selbsttötungsmechanismus, um die Ausbreitung von GVOs zu verhindern, was sowohl die vielen Bedenken gegenüber GVOs anspricht als auch die notwendigen

Biosicherheitsstandards erfüllt. Darüber hinaus ermöglicht der Plasmid-Chassis einen kontrollierten Einsatz des Probiotikum und beschränkt seinen Lebensraum auf den des menschlichen Darms. Kontrolle erhält man durch das Verwenden von drei verschiedenen Regulationsmechanismen, die das Cas3-Selbsttötungssystem hemmen, wenn sich das Bakterium in der vorgesehenen Umgebung befindet.

Im Darm soll unser Probiotikum Inkretin-Mimetika als Reaktion auf Glucose bilden und absondern. Die Inkretin-Mimetika werden anschließend von den Enterozyten aufgenommen und über das Blut in die Bauchspeicheldrüse transportiert, wo sie die Insulinsekretion verstärken. Da Diabetes eine der am weitesten verbreiteten Krankheiten ist, sprechen wir ein Problem an, das nicht nur für Kliniken, sondern auch für die Wirtschaft und die weltweite Gesundheitssituation relevant ist. Neben der Entwicklung eines solchen GVO planen wir auch die Modellierung der Effizienz, Stabilität und Kinetik des Selbsttötungsmechanismus. Unser probiotischer Cas3-Biosicherheitschassis kann auch für andere Anwendungen eingesetzt werden, indem man das Gen von Interesse im Plasmid austauscht. Daher kann unser Chassis von Forschern verwendet werden, die auf der Suche nach einem funktionalen Chassis sind, der die Biosicherheitsstandards erfüllt.

Noch mehr Informationen zu unserem Team, iGEM und dem Projekt gibt es auf unserer Website: `igem-tuebingen.com`

Akronyme

BLAST Basic Local Alignment Search Tool. 12

Cas CRISPR associated. 22

CRISPR Clustered Regularly Interspaced Palindromic Repeats. 22

DNA Desoxyribonukleinsäure. 3

gRNA guide RNA. 25

GVO Genetisch Veränderter Organismus. 27

iGEM International Genetically Engineered Machine. 3, 26, 27

NGS Next Generation Sequencing. 5

PAM Protospacer adjacent motif. 24

PCR polymerase chain reaction. 20–22

SAM Sequence Alignment Map. 18

SMRT Single-molecule real time Sequenzierung. 8

SNP Single Nucleotide Polymorphism. 14

ZMV Zero Mode Waveguide. 29

Glossar

assembly Rekonstruktion der gesamten DNA-Sequenz aus kurzen reads. 15

Commandline Tool Software, die ausschließlich von der Kommandozeile gestartet werden kann und keine graphische Oberfläche hat. 12, 13

Illumina sequencing Sequenzierverfahren der 2. Generation von Illumina - verwendet sequencing-by-synthesis. 5

Illumina sequencing Sequencing by synthesis ist ein Sequenzierverfahren, welches die wiederholte Synthese eines komplementären Stranges Nukleotid-für-Nukleotid verfolgt.. 5

library preparation Jedes DNA-Fragment wird in einen Klonierungsvektor eingefügt, der Pool rekombinanter DNA-Moleküle wird dann in eine Population von Bakterien oder Hefen überführt, so dass jeder Organismus ein Konstrukt (Vektor + Insert) enthält. 8

mapping Assoziation von reads zu ihren Positionen im Genom. 15

Nanopore Sequenzierung Sequenzierverfahren bei dem die DNA durch eine Pore in Nanometergröße transportiert wird. Jede Nukleobase wird anhand ihrer spezifischen Spannungsänderung identifiziert. 8, 9

reads Kurzes DNA-Fragment, welches von im Rahmen einer Sequenzierung gemessen wurde.. 8

Sanger Sequenzierung Enzymatische Sequenziermethode der 1. Generation durch Kettenabbruch-Synthese. 4

Sequenzierung Prozess, bei welchem die Abfolge von Nukleotiden einer DNA bestimmt wird. 3

Single Molecule Real-Time Sequenzierung Sequenzierverfahren bei dem die Basenreihenfolge der DNA mithilfe von Fluoreszenzmarkern an den verschiedenen Nukleobasen innerhalb eines kleinen Beobachtungsvolumens (ZMV) identifiziert wird. 8, 9

somatic mutation Identifikation von auffälligen Varianten in reads im Vergleich zu einem Referenzgenom, die ausschließlich auf Umwelteinflüsse zurückzuführen sind. 15

variant calling Identifikation von auffälligen Varianten in reads im Verleich zu einem Referenzgenom. 15