

Bachelorarbeit

**Analyse von RAD-Seq-Daten als Optimierungsproblem unter Berücksichtigung von
Sequenzierfehler- und Mutationsraten**

Antonie Vietor

Gutachter:

Name des Erstgutachters

Name des Zweitgutachters

Technische Universität Dortmund

Fakultät für Informatik

Lehrstuhl 11

Bioinformatics for High-Throughput Technologies

<http://ls11-www.cs.tu-dortmund.de/>

In Kooperation mit:

Universität Duisburg-Essen

Genome Informatics

<http://genomeinformatics.uni-due.de/>

Inhaltsverzeichnis

| | | |
|----------|--|-----------|
| 1 | Einleitung | 1 |
| 1.1 | Biologischer Hintergrund | 1 |
| 1.1.1 | Aufbau und Struktur der DNA | 1 |
| 1.1.2 | Bindungen innerhalb und zwischen DNA-Molekülen | 2 |
| 1.1.3 | RNA und die Proteinbiosynthese | 3 |
| 1.1.4 | DNA-Replikation | 4 |
| 1.1.5 | Mutationen und SNPs | 5 |
| 1.2 | Molekulargenetische Verfahren und Techniken | 7 |
| 1.2.1 | Sanger-Sequenzierung | 7 |
| 1.2.2 | PCR | 8 |
| 1.2.3 | Next Generation Sequencing | 8 |
| 1.3 | RAD-Sequencing | 8 |
| 1.3.1 | Verfahren | 8 |
| 2 | Analyse von RAD-Seq-Daten | 9 |
| 2.1 | Problemstellung | 9 |
| 2.2 | Formale Definition | 9 |
| 2.3 | Lösungsansatz | 9 |
| 3 | Algorithmus | 11 |
| 3.1 | Preprocessing | 11 |
| 3.2 | Edit-Distanzen | 12 |
| 3.3 | Konstruktion des Graphen und Bestimmung der Zusammenhangskomponenten | 12 |
| 3.4 | Bestimmung der maximalen Likelihood der Allele | 15 |
| 3.4.1 | | 15 |
| 3.5 | Bestimmung der maximalen Likelihood der Loci | 15 |
| 3.5.1 | | 15 |
| 3.6 | Ausgabe der Loci im VCF-Format | 15 |
| 3.6.1 | | 15 |

| | | |
|----------|--|-----------|
| 4 | Evaluation an simulierten Datensätzen | 17 |
| 4.1 | | 17 |
| 4.1.1 | | 17 |
| 5 | Zusammenfassung und Ausblick | 19 |
| 5.1 | | 19 |
| 5.1.1 | | 19 |
| A | Weitere Informationen | 21 |
| | Abbildungsverzeichnis | 23 |
| | Algorithmenverzeichnis | 25 |
| | Literaturverzeichnis | 29 |
| | Erklärung | 29 |

Kapitel 1

Einleitung

1.1 Biologischer Hintergrund

1.1.1 Aufbau und Struktur der DNA

In den vergangenen Jahren wurden durch die molekulargenetische Methoden in Medizin und Biologie enorme Fortschritte erzielt. Heute sind sie nicht nur ein wesentliches Instrument bei der Erforschung, Diagnostik und Therapie verschiedenster Erkrankungen sondern sind auch bei der Entdeckung und Klassifikation von Organismen oder ganzen Ökosystemen von entscheidender Bedeutung.

Einer der ersten und wichtigsten Meilensteine auf dem noch eher jungen Gebiet der Molekulargenetik wurde 1953 durch die Entdeckung der Doppelhelixstruktur der DNA und die Beschreibung ihres Aufbaus erreicht [1]. Die DNA (desoxyribonucleic acid) ist ein langkettiges, aus zwei gegenläufigen, komplementären Strängen bestehendes und zu einer Helix gewundenes Molekül, welches die Erbinformation der meisten Zellen codiert. Die Komplementarität und Gegenläufigkeit werden in Kap. 1.1.2 gesondert beschrieben. Jeder Strang besteht aus vielen aneinandergereihten Nukleotiden, die sich jeweils aus einem Zuckermolekül (Desoxyribose), einem Phosphatrest und einer von vier möglichen Basen zusammensetzen. Als Basen kommen in der DNA Adenin (A), Thymin (T), Guanin (G) und Cytosin (C) vor. Die Kombinationen dieser Basen codieren über ihre Sequenz die genetische Information.

Hierbei codieren im Rahmen der Proteinbiosynthese (Kap. 1.1.3) Kombinationen aus jeweils drei Basen, sog. Basentriplets bzw. Codons, entweder für eine von i.d.R. 20 Aminosäure [2, 3] oder für Start- bzw. Stop-Sequenzen, welche den Anfang bzw. das Ende von Genen signalisieren (genetischer Code). Hiervon codieren 61 Triplets für die bereits erwähnten 20 Aminosäuren. Somit codieren für die meisten Aminosäuren mehrere verschiedene Triplets, diese Eigenschaft des genetischen Codes wird auch als Degeneration

bezeichnet.

Die Basensequenz der DNA entspricht somit einer Aminosäuresequenz, welche die Primärstruktur von Proteinen (Eiweißmolekülen) darstellt. Proteine erfüllen in lebenden Organismen umfangreiche Funktionen, sie können als Hormone, Enzyme, Strukturproteine fungieren und sind an den meisten Stoffwechselprozessen und Signalwegen von Zellen beteiligt. Die verschiedenen Proteine werden jeweils von für sie spezifischen Abschnitten auf der DNA codiert. Solche DNA-Abschnitte werden als Gene bezeichnet. Ein DNA-Strang besteht also aus vielen verschiedenen Genen. In komplexeren Zellen befinden sich im Zellkern in der Regel mehrere DNA-Stränge, die jeweils ein Chromosom repräsentieren, auf dem sich jeweils mehrere Gene befinden. Die Anzahl der Chromosomen innerhalb der Zellen ist speziesabhängig.

Zwischen den einzelnen Genen eines DNA-Stranges befinden sich nicht-codierende und oft repetitive Sequenzen. Ebenso gibt es auch innerhalb der Gene codierende Abschnitte (Exons) und nicht-codierende Sequenzen (Introns). Insgesamt machen die für Proteine codierenden Bereiche der DNA nur einen geringen Anteil des Genoms, also der gesamten Erbinformation einer Zelle aus. Beim Menschen wird dieser Anteil auf etwa 2 % geschätzt, d.h. ca. 98 % des menschlichen Genoms besteht aus DNA, die nicht für Proteine codiert. Über diese nicht-codierenden Abschnitte ist bislang nur wenig bekannt, teilweise werden ihnen regulatorische Funktionen zugeschrieben [4, 5].

1.1.2 Bindungen innerhalb und zwischen DNA-Molekülen

DNA liegt meist in Form eines Doppelstrangs vor. Die Basen beider Stränge sind dabei intermolekular über schwache chemische Bindungen, sogenannte Wasserstoffbrücken, mit einander verbunden. Dabei kann Adenin nur an Thymin unter Ausbildung von zwei Wasserstoffbrückenbindungen binden. Ebenso kann Cytosin nur mit Guanin über insgesamt drei Wasserstoffbrücken eine Bindung eingehen. Diese selektive Basenpaarung wird auch als Komplementarität bezeichnet. Es sind also Adenin und Thymin ebenso wie Cytosin und Guanin jeweils komplementär zu einander. Bezogen auf einen DNA-Doppelstrang sind auch seine beiden Einzelstränge komplementär, so dass sich für jede Base des einen Stranges auf dem anderen Strang jeweils die komplementäre Base an der entsprechenden Position befindet. Es genügt also so die Sequenz von einem der beiden Stränge zu kennen, um die Sequenz des jeweils anderen rekonstruieren zu können. Dies wird sowohl zum Auslesen der genetischen Informationen bei der Transkription (Kap. ??) als auch zum Kopieren von DNA im Rahmen der DNA-Replikation (Kap. 1.1.4) genutzt.

Wie bereits in Kap. 1.1.1 erwähnt, ist doppelsträngige DNA gegenläufig orientiert. Die Einzelstränge besitzen eine Polarität, welche durch die intramolekulare Bindung zwischen den einzelnen Nukleotiden über sogenannte Phosphodiesterbindungen zustande kommt. Dabei bindet der Phosphatrest, der sich jeweils am 5. Kohlenstoffatom (C5) des Zuckermoleküls der Nukleotide befindet, an das 3. Kohlenstoffatom (C3) des Zuckermoleküls des nachfolgenden Nukleotids. An den Enden eines DNA-Strangs fehlt jedoch diese Phosphodiesterbindung, so dass an einem Ende das C3 ungebunden bleibt (3'-Ende) während am anderen Ende der Phosphatrest nur an C5 gebunden ist (5'-Ende) und somit die zweite Esterbindung am Phosphatrestes fehlt. Aufgrund der Gegenläufigkeit befindet sich also an beiden Enden eines Doppelstrangs jeweils ein 3'-Ende des einen und ein 5'-Ende des anderen Einzelstrangs. Diese Polarität spielt eine wichtige Rolle bei der Lese- und Syntheserichtung im Rahmen der Transkription und der DNA-Replikation.

1.1.3 RNA und die Proteinbiosynthese

Ebenso wie DNA gehört auch RNA (ribonucleic acid) zu den Nukleinsäuren. Sie unterscheidet sich in ihrem Aufbau von der DNA durch die Base Uracil (U) statt Thymin und den Zucker Ribose statt Desoxyribose. Meist liegt RNA einzelsträngig oder nur über kürzere Abschnitte doppelsträngig vor. Während DNA insbesondere der Speicherung der Erbinformation dient, hat RNA eher die Funktion der Informationsübertragung. RNA nimmt daher zahlreiche regulatorische Funktionen wahr.

Eine ihrer wichtigsten Aufgaben ist die Übertragung der genetischen Information von der DNA in die Aminosäuresequenz der Proteine bei der Proteinbiosynthese. Dabei werden von dem für das herzustellende Protein codierenden DNA-Abschnitt zunächst Arbeitskopien in Form von mRNA (messenger RNA) hergestellt. Ein solches Umschreiben von DNA in mRNA wird auch als Transkription bezeichnet. Nach der Transkription erhalten die mRNA-Fragmente noch einige Modifikationen und werden aus dem Zellkern hinaus in das Cytoplasma transportiert. Im Cytoplasma erfolgt schließlich mit Hilfe sogenannter tRNAs (transfer RNA) die Übersetzung der Basentriplets in eine Aminosäuresequenz (Translation, siehe auch Kap. 1.1.1). tRNAs besitzen eine Bindungsstelle bestehend aus jeweils drei Nukleotiden, mit der sie komplementär an ein passendes Basentriplett der mRNA binden. In Abhängigkeit vom Basentriplett an ihrer mRNA-Bindungsstelle trägt jede tRNA entsprechend dem genetischen Code eine spezifische Aminosäure. Entlang der mRNA wird nun ab der Startsequenz für jedes Basentriplett die passende tRNA nacheinander angelagert. Sobald eine tRNA bindet, wird die an sie gebundene Aminosäure gelöst und an die Aminosäure der nachfolgenden tRNA gebunden. Dadurch entsteht eine Kette von aneinander gebundene Aminosäuren, die sich jeweils entsprechend der mRNA Sequenz an die Aminosäure der nächsten passenden tRNA anlagert und um deren Aminosäure verlängert.

Beim Erreichen einer Stop-Sequenz kann diese keine tRNA angelagern, die Synthese wird abgebrochen und die Aminosäurekette löst sich von der zuletzt gebundenen tRNA und wird weiteren Modifikationen unterzogen.

1.1.4 DNA-Replikation

Die DNA-Replikation dient der Verdopplung der DNA im Rahmen der Zellteilung, so dass jede der beiden resultierenden Tochterzellen das gleiche genetische Material erhält. Die DNA-Replikation ist also ein natürlicher Vorgang zur Erzeugung von DNA-Kopien. Ihr grundlegendes Prinzip findet bei der PCR (Polymerase-Ketten-Reaktion, siehe Kap. 1.2.2) Anwendung und wird für verschiedene molekulargenetische Verfahren genutzt, um DNA-Kopien synthetisch herzustellen. In diesem Zusammenhang sei hier auf die verschiedenen Sequenzierungstechniken insbesondere im Hinblick auf das RAD-Sequencing verwiesen (Kap. 1.2.1, Kap. 1.2.3 und Kap. 1.3). Daher soll der Vorgang der DNA-Replikation im Folgenden kurz umrissen werden [6, 7, 8].

Bei eukaryotischen Zellen, die im Gegensatz zu Bakterien (Prokaryoten) einen Zellkern besitzen, liegt die DNA im Zellkern häufig gewunden und stark kondensiert vor. Um ihre Sequenz Base für Base kopieren zu können, muss sie zunächst entwunden werden. Dies geschieht durch Enzyme aus der Gruppe der Topoisomerasen. Diese erzeugen gezielt am Replikationsursprung temporäre Strangbrüche in der DNA und entspannen so den Doppelstrang. Anschließend setzt am Replikationsursprung ein weiteres Enzym, die Helikase, an und trennt die beiden komplementären Stränge auf. Es entsteht die sogenannte Replikationsgabel. Während der Replikation schiebt sich die Helikase unter fortlaufender Auftrennung der beiden Stränge auf der DNA entlang. Auch bei diesem Prozess sorgen Topoisomerasen immer wieder für eine Entspannung des DNA-Fadens hinsichtlich seiner Windung.

Nun kann der eigentliche Kopiervorgang an den beiden von einander getrennten Einzelsträngen mit Hilfe von DNA-Polymerasen erfolgen. Dabei fahren die DNA-Polymerasen an den Strängen entlang und fügen an jeder Position des Elternstranges Nukleotide mit der jeweils komplementären Base an. Im Ergebnis entsteht also an jedem der beiden Elternstränge ein neuer komplementärer DNA-Strang, der also die gleiche Basensequenz wie der jeweils andere Elternstrang besitzt. Bei den DNA-Polymerasen handelt es sich um Enzyme, die für die Initiation des Kopiervorgangs eine Startsequenz benötigen. Diese Startsequenzen sind kleine RNA-Fragmente mit spezifischer Basensequenz, die sich an die jeweils passende, komplementäre Stelle auf dem Elternstrang anlagern. An diese, auch als RNA-Primer bezeichneten Fragmente kann nun die DNA-Polymerase binden und mit der Replikation

beginnen.

Die DNA-Polymerase kann den Elternstrang nur in eine Richtung lesen, nämlich vom 3' Ende zum 5' Ende (3'-5'-Richtung). Aufgrund der gegenläufigen, antiparallelen Ausrichtung zweier komplementärer Stränge zu einander, kann folglich die Synthese des Tochterstranges nur in 5'-3'-Richtung erfolgen. Für die beiden antiparallel ausgerichteten Elternstränge bedeutet dies, dass nur bei einem Strang die Richtung der sich öffnenden Replikationsgabel der 5'-3'-Richtung des Stranges entspricht. Dieser Strang kann kontinuierlich repliziert werden, da sich die DNA-Polymerase auf ihm in Richtung der voranschreitenden Aufspaltung des Doppelstranges bewegt. Der auf diese Weise kontinuierlich synthetisierte Strang wird als Leitstrang bezeichnet.

Der andere Strang ist jedoch in Gegenrichtung orientiert, so dass seine Syntheserichtung, also die 5'-3'-Richtung, entgegengesetzt zur Bewegungsrichtung der Replikationsgabel orientiert ist. Dadurch können jeweils nur kleinere Fragmente synthetisiert werden die von der Replikationsgabel bis zum bereits replizierten Teil des Strangs reichen. Schreitet die Öffnung der Replikationsgabel weiter fort, muss der nun frei gewordene Strangabschnitt ebenfalls synthetisiert werden. Es muss also erneut ein RNA-Primer angelagert werden und dann mit Hilfe der DNA-Polymerase der Bereich zwischen Replikationsgabel und bereits replizierten Strang synthetisiert werden. Die Replikation erfolgt somit diskontinuierlich. Der so synthetisierte Strang wird als Folgestrang bezeichnet und besteht zunächst aus multiplen Fragmenten (Okazaki-Fragmente). Nach dem Replikationsvorgang werden mit Hilfe der DNA-Polymerase die RNA-Primer durch DNA ersetzt. Im Anschluss werden die multiplen Fragmente des Folgestrangs durch Ligasen zu einem kontinuierlichen Strang verbunden. Im Ergebnis sind also nach Abschluss der DNA-Replikation zwei identische Kopien der beiden Elternstränge entstanden.

1.1.5 Mutationen und SNPs

ToDo: ggf. Literatur zu den verschiedenen Mutationsarten ergänzen (DD: als Grundlagenwissen voraussetzen?)

Veränderungen in der DNA-Sequenz werden als Mutationen bezeichnet. Sie können durch zellinterne Faktoren verursacht werden, wie beispielsweise Fehler beim Kopiervorgang der DNA (DNA-Replikation) während der Zellteilung. Ebenso können sie durch zahlreiche Umwelteinflüsse entstehen.

Mutationen können unterschiedlich große Abschnitte der DNA betreffen, von ganzen Chromosomen oder großen Chromosomenabschnitten über Veränderungen von mehreren Basen bis hin zu sogenannten Punktmutationen, bei denen nur eine einzige Base verändert ist. Auf DNA-Ebene können Punktmutationen in Form Substitutionen, Insertionen und

Deletionen auftreten. Bei der Substitution wird eine Base durch eine andere ausgetauscht, bei der Insertion wird eine zusätzliche Base in den DNA-Strang eingefügt und bei der Deletion kommt es zum Verlust einer Base.

Liegt eine solche Punktmutation in den codierenden DNA-Abschnitten, so können sich auf Proteinebene verschiedene Konsequenzen daraus ergeben. Insertionen und Deletionen bewirken durch die zusätzliche bzw. fehlende Base eine Verschiebung des Leserasters, so dass sich die Tripletstruktur für alle nachfolgenden Basen verschiebt. Dies wird als Frame-Shift bezeichnet und verursacht meist eine gravierende Veränderung des resultierenden Proteins, da viele der nachfolgenden Triplets nun für andere Aminosäuren codieren. Dies führt häufig zu einem deutlich veränderten Protein, welches seine reguläre Funktion nicht mehr oder nur noch unvollständig wahrnehmen kann.

Bei Substitutionen bleibt dagegen das Leseraster erhalten. Aufgrund der Degeneration des genetischen Codes können verschiedene Basentriplets für die gleiche Aminosäure codieren. Dadurch kann ein Basentriplett mit einer Punktmutation trotz des Basenaustauschs noch für die ursprüngliche Aminosäure codieren, so dass das resultierende Protein unverändert bleibt. In diesem Fall spricht man von einer silent-Mutation. Codiert das Basentriplett aber aufgrund der Mutation für eine andere Aminosäure, so handelt es sich um eine missense-Mutation. Die Proteinsequenz wird dadurch in einer Aminosäure geändert, so dass es je nach Position der betreffenden Aminosäure zu verschieden starken Effekten hinsichtlich der Proteinfunktion kommen kann.

Zudem können Substitutionen und Frame-Shifts dazu führen dass ein für eine Aminosäure codierendes Basentriplett zu einem Stop-Codon umgewandelt wird (nonsense-Mutation) oder ein Stop-Codon durch die Mutation für eine Aminosäure codiert (readthrough-Mutation).

Hinsichtlich der Eigenschaften des resultierenden Proteins unterscheiden man im Zusammenhang mit Mutationen zudem zwischen sogenannten loss-of-function- und gain-of-function-Mutationen. Loss-of-function-Mutationen führen zu einer verringerten Funktionalität oder dem vollständigen Funktionsverlust des Proteins. Gain-of-function-Mutationen bewirken dagegen eine verstärkte oder veränderte Aktivität bzw. Funktionalität des Proteins.

Wie bereits erwähnt führen aber nicht alle Veränderungen der DNA-Sequenz zu Störungen der Genfunktion. Veränderungen ohne unmittelbaren Krankheitswert werden als genetische Varianten bezeichnet, wenn sie innerhalb einer Spezies vermehrt auftreten [9, 10]. Am häufigsten finden sich dabei Varianten einzelner Basenpaare, sogenannte SNP's (single nucleotide polymorphisms). SNP's kommen sowohl in codierenden als auch nicht-

codierenden DNA-Abschnitten vor und treten regionsabhängig in unterschiedliche Häufigkeit auf. SNP's können als genetische Marker benutzt werden [11, 12], ihr Auftreten und ihre Verteilung spielen vor allem in der Populationsgenetik eine wichtige Rolle. Sie können Aufschluss hinsichtlich der Diversität, Selektion und Demographie einer Population geben [13, 14, 15].

Während große strukturelle Chromosomenaberrationen unter Umständen bereits lichtmikroskopisch erkennbar sind, ist bei Punktmutationen oder SNP's lediglich eine einzige Base verändert. Solche Veränderungen lassen durch verschiedene molekulargenetische Verfahren detektieren [12, 16]. Insbesondere die direkte Analyse der DNA-Sequenz mittels Sequenzierung (siehe Kap. 1.2.1) ist durch die Entwicklung der sogenannten Next-Generation-Sequencing-Verfahren (NGS) im Hochdurchsatzverfahren und mit hoher Parallelisierung durchführbar (siehe Kap. 1.2.3). Diese Techniken ermöglichen inzwischen umfangreiche, genomweite Analysen hinsichtlich einer großen Vielfalt molekulargenetischer Fragestellungen. Die vorliegende Arbeit befasst sich mit der Analyse und Auswertung von RAD-Sequencing-Daten. Auch die RAD-Sequenzierung gehört zu den NGS-Verfahren und dient insbesondere der Detektion von SNP's. Daher wird dieses Verfahren in Kap. 1.3 detaillierter vorgestellt.

1.2 Molekulargenetische Verfahren und Techniken

1.2.1 Sanger-Sequenzierung

Nach der Erforschung der DNA-Struktur war schließlich die Entwicklung der Sanger-Sequenzierung im Jahr 1975 ein entscheidender Meilenstein der molekulargenetischen Forschung [17]. Durch sie war es erstmals möglich die genaue Basensequenz eines DNA-Strangs zu bestimmen.

Hierbei wird die zu sequenzierende DNA-Probe in vier Teile aufgeteilt, denen jeweils eine der vier DNA-Basen in Form von radioaktiv markierten synthetischen Nukleotiden, sowie anteilig einige modifizierte Nukleotiden dieser Base hinzugefügt werden. Die jeweils anderen drei Basen werden als unmarkierte und unmodifizierte Nukleotide hinzugegeben. In jeder der Probengemische ist also eine andere Base markiert und zum Teil auch modifiziert.

Wie bei der natürlichen DNA-Replikation (siehe Kap. 1.1.4) während der Zellteilung kann die Proben-DNA durch Hinzugabe der DNA-Polymerase I kopiert werden. Dabei werden auch die radioaktiv markierten Nukleotide in den kopierten Strang eingebaut. Die Kopiervorgänge starten jeweils an einem kleinen Fragment mit bekannter DNA-Sequenz, dem sog. Primer. Auch die Primer werden vorab den Probengemischen beigelegt. Der Primer

bindet komplementär an die Ausgangs-DNA der Probe und ermöglicht dadurch schließlich die Bindung der DNA-Polymerase. Diese fährt vom Primer aus am Ausgangsstrang entlang und fügt dabei zu jeder Base des Ausgangsstrangs ein Nukleotid mit komplementärer Base an die Kopie an. Wird dabei eines der modifizierten Nukleotide eingefügt, so kann im nächsten Schritt kein weiteres Nukleotid mehr an den kopierten Strang angefügt werden und der Synthesevorgang wird abgebrochen. Dadurch entstehen multiple, radioaktiv markierte DNA-Fragmente unterschiedlicher Länge. In jedem der Probenansätze enden diese Fragmente mit der selben Base, da nur eine der vier Basen in modifizierter Form hinzugegeben wurde.

Die vier Proben werden nun nebeneinander auf ein Gel aufgetragen. Da DNA negativ geladen ist bewegt sie sich im elektrischen Feld zur Anode. Wird also an das Gel ein elektrisches Feld angelegt, so werden die DNA-Fragmente durch das Gel bewegt. Kleinere Fragmente werden dabei schneller bewegt als größeren. Dadurch ist es möglich die DNA-Fragmente der Proben entsprechend ihrer Länge aufzutrennen. Es entstehen im Gel Anhäufungen von Fragmenten gleicher Länge, die auch als Banden bezeichnet werden. Die radioaktive Markierung der Banden kann auf Röntgenfolie sichtbar gemacht werden. Bei moderneren Verfahren ist die Markierung mit radioaktiven Isotopen durch Fluoreszenzfarbstoffe abgelöst worden. Da bekannt ist in welchen Proben welche der Basen markiert ist, kann die DNA-Sequenz direkt aus der aufsteigenden Länge der DNA-Fragmente an den Banden abgelesen werden.

1.2.2 PCR

Zunächst waren für die Sanger-Sequenzierung große Mengen an Zellmaterial notwendig, um daraus ausreichend DNA für sichtbare Banden auf dem Gel extrahieren zu können. Die Sequenzierung mit nur geringen DNA-Mengen war nicht möglich. Durch die Entwicklung des Verfahrens der Polymerasekettenreaktion (PCR, polymerase chain reaction) [18] wurden die Möglichkeiten der Sequenzierung revolutioniert. Die PCR ermöglicht es aus kleinsten DNA-Proben multiple Kopien herzustellen. Und so ist es inzwischen möglich, selbst aus der DNA einer einzigen Zelle umfangreiche Analysen durchzuführen [19].

1.2.3 Next Generation Sequencing

1.3 RAD-Sequencing

1.3.1 Verfahren

Kapitel 2

Analyse von RAD-Seq-Daten

2.1 Problemstellung

2.2 Formale Definition

2.3 Lösungsansatz

Kapitel 3

Algorithmus

Für das hier implementierte RAD-Sequencing-Tool, NodeRAD, wurde zur Workflowintegration das Workflow Management System Snakemake verwendet [20, 21]. Die einzelnen Analyseschritte werden dabei über Regeln abgebildet. Für jede Regel können neben dem zu verwendenden Script oder Shell-Kommando sowie den Pfadangaben für In- und Output auch zusätzliche Optionen festgelegt werden. Dazu gehören beispielsweise Angaben zu Parametern bzw. Argumenten für die verwendeten Tools, Pfadangaben für Log-Dateien oder die Anzahl der zu verwendenden Threads.

Als Input benötigt der Workflow eine Datei im FASTQ-Format, welche die single-end Reads der verschiedenen Individuen mit ihren Identifikationsbezeichnungen, der Basensequenz und Angaben zur Basenqualität enthält. Des Weiteren wird eine Tabelle im tsv-Format benötigt, in der die Zuordnung der Probenamen zu den Individuen und ihren Barcode-Sequenzen definiert ist. Nach dem Preprocessing, der Qualitätskontrolle der Reads und dem Sequence-Alignment erfolgt die RAD-Seq-Analyse durch NodeRAD. Hierbei werden die Wahrscheinlichkeiten der Allelsequenzen und der möglichen Loci bestimmt. Die Loci mit der höchsten Wahrscheinlichkeit werden schließlich mit den Sequenzen ihrer Allele und den möglichen Varianten entsprechend dem ermittelten Genotyp in einer Datei im Variant Call Format (VCF) ausgegeben.

3.1 Preprocessing

Im Preprocessing werden durch das Tool Cutadapt [22] die Reads jedes Individuums anhand ihrer Barcode-Sequenzen identifiziert und extrahiert. Hiernach werden die Barcode-Sequenzen entfernt (trimming) und die Reads jedes Individuums in separaten Dateien im FASTQ Format abgelegt.

Im Anschluss an das Trimming erfolgt eine Qualitätskontrolle durch das Tool FastQC [23]. Dabei werden einige allgemeine Statistiken zu den Rohdaten der Reads generiert, wie beispielsweise zur Basenqualität, zum GC-Gehalt, dem Anteil an Duplikaten oder überre-

präsentierten Sequenzen. Durch das Tool MultiQC [24] wird aus diesen Statistiken und den Log-Dateien von Cutadapt ein html-Report mit diversen Plots zur Veranschaulichung erstellt.

3.2 Edit-Distanzen

Für die spätere Konstruktion eines Graphen basierend auf den Edit-Distanzen zwischen den Readsequenzen wird für jedes Individuum zunächst ein Sequenzalignment mit Hilfe des Tools Minimap2 [25] erstellt. Hierbei werden alle Readssequenzen paarweise verglichen und in Abhängigkeit von ihren Übereinstimmungen (Matches) und Unterschieden (Mismatches) einander zugeordnet. Das Ergebnis des Mappings wird im sam-Format ausgegeben und enthält Angaben zur betrachteten Sequenz (Query), die gegen einen anderen Read (Reference) verglichen wurde. Neben den ID's der Query- und Reference-Sequenzen, werden dort unter anderem auch der CIGAR-String, Informationen zur Basenqualität der Query-Sequenz, sowie optional verschiedene Tags angegeben. Ein für die späteren Berechnungen wichtiges Maß sind die Edit-Distanzen, die durch den NM-Tag repräsentiert werden. Die Edit-Distanz gibt hierbei die minimale Anzahl von Editieroperationen an, um die Query-Sequenz in die Referenzsequenz zu transformieren. Als Editieroperationen sind hierbei ersetzen, einfügen und löschen von Basen möglich. Auf DNA-Ebene entspricht dies den Punktmutationen im Sinne von Substitutionen, Insertionen und Deletionen (vgl. Kap. 1.1.5). Der CIGAR-String ist eine kondensierte Darstellung der Unterschiede zwischen Query- und Reference-Sequenz. In ihm werden Matches und Mismatches wie Insertionen, Substitutionen und Deletion jeweils mit der Anzahl der betroffenen Basen angegeben. Sowohl der CIGAR-String als auch der NM-Tag definieren wichtige Kanteneigenschaften des späteren Graphen.

3.3 Konstruktion des Graphen und Bestimmung der Zusammenhangskomponenten

Das hier implementierte Tool, NodeRAD, benötigt als Input zu jedem Individuum die getrimmten single-end Read-Daten sowie das Sequenzalignment. Zunächst wird daraus für jedes Individuum ein separater Graph G mit $G = (V, E)$ erstellt. Seine Knoten, V , werden durch die einzelnen Reads repräsentiert. Entsprechend ergeben sich die Knoteneigenschaften aus den Daten der Reads, diese werden den FASTQ-Dateien nach Ausführung von Cutadapt (siehe 3.1) entnommen. Die Kanten, E , zwischen den Knoten ergeben sich aus dem Vergleich ihrer Sequenzen im Rahmen des Sequenzalignments mittels Minimap2 (siehe 3.2).

Zusätzlich entnimmt NodeRAD der Konfigurationsdatei des Workflows einige Konstanten und Grenzwerte für die späteren Berechnungen. Dazu gehören die Mutationsraten

3.3. KONSTRUKTION DES GRAPHEN UND BESTIMMUNG DER ZUSAMMENHANGSKOMPONENTEN

und Heterozygotiewahrscheinlichkeiten für Substitutionen, Insertionen und Deletionen, die Ploidie des Chromosomensatzes der untersuchten Spezies und Grenzwerte. Als konfigurierbare Grenzwerte gibt es für NodeRAD einen Schwellenwert für die maximal zulässige Editierdistanz, bei dem zwei Knoten noch durch eine Kante verbunden werden sowie Schwellenwerte zum Filtern selten vorkommender Sequenzen ab einer bestimmten Clustergröße, die als Hintergrundrauschen nicht in der Berechnung Berücksichtigung finden sollen.

=====

Zur Konstruktion des Graphen wird graph-tool verwendet,
Extraktion der Fastq-Daten je Individuum (später auch results für den Vergleich verschiedener Individuen)
gerichteter Graph
Knoten des Graphen entsprechen jeweils den Reads jedes Individuums aus den getrimmten Fastq-Dateien

Knoteneigenschaften: $O(n)$, wird für jeden Read durchgeführt

https://graph-tool.skewed.de/static/doc/graph_tool.html?highlight=add_edge#graph_tool.Graph.add_edge

-> aus fastq der einzelnen Individuen

ID: wird zusätzlich vergeben, Kombination aus Index als laufende Nummer und Probenname, String

Name: Sequenzidentifizier, entspricht Zeile 1 des Reads im Fastq-Format, wird als String gespeichert

Sequenz: wird als String gespeichert

Quality: aus der Quality-Sequenz des Reads aus den Fastq-Daten wird für jede Base der p-Wert berechnet. Dazu erfolgt zunächst das platformabhängige Encoding des Q-Wertes mit Hilfe von SeqIO aus dem Bio Python Package. Aus den Q Werten kann nun für jede Base des Reads der p-Wert errechnet werden: $10^{-Q/10}$, wird als Vektor von Float-Werten gespeichert

https://www.drive5.com/usearch/manual/quality_score.html

https://en.wikipedia.org/wiki/Phred_quality_score

Kanteneigenschaften:

-> aus Minimap2-Output im sam-Format, werden mit Hilfe von pysam ausgelesen

Kante Hinzufügen graph-tool in $O(1)$, gerichtete Kante von source/query zu target/ref -> ermöglicht die Betrachtung der Nachbarn bzw. aller ausgehenden Kanten eines Knotens

Edit-Distanz: aus dem nm-tag wird zunächst genutzt, um nur Kanten in den Graphen aufzunehmen, die bereits einen optimierten Minimap2-Path darstellen. Hierfür besteht optional die Möglichkeit den Schwellenwert für die maximale Edit-Distanz in der Snakemake-rule festzulegen.

cs-tag und cigar-string: neben dem klassischen cigar-string des sam-files kann der cs-tag im short- oder long-Format in den Graphen aufgenommen werden und im NodeRAD-Output mit ausgegeben werden, um für weitere Analysen zur Verfügung zu stehen. Der cigar-string selbst wird nicht in seiner üblichen Formatierung gespeichert, sondern liegt intern in Tupeln codiert vor und dient der Berechnung der Likelihood jeder Kante.

likelihood: Für jede Kante wird für jede Base der Read-Sequenz die Likelihood aus dem Cigar-String in Kombination mit den für die Knoten bestimmten p-Werten und der Mutationsrate berechnet:

bei Match: $(1 - \text{mutrate}) * \text{float}(1/3) * i + \text{mutrate} * (1 - i)$

bei Mismatch: $(1 - \text{mutrate}) * \text{float}(1/3) * i + \text{mutrate} * (1 - i)$

Die Likelihood für eine Kante ergibt sich dann aus dem Produkt der Wahrscheinlichkeiten aller Basen eines Reads. Bei einem Mismatch können die Mutationsraten auch explizit für Substitutionen, Deletionen und Insertionen spezifiziert werden.

zusätzliche konfigurierbare Optionen:

Schwellwert für die maximal im Graphen zu berücksichtigende edit-Distanz kann festgelegt werden, default 23

Mutationsrate kann für Substitutionen, Insertionen und Deletionen individuell festgelegt werden

Zwischenergebnisse als Output:

Graph als xml-Datei und als visuelle Darstellung im pdf-Format

zusätzliche Statistiken/Logs: In den Log-Files werden für jedes Individuum einige Statistiken zur Konstruktion des Graphen festgehalten. Neben der Anzahl der Knoten und Kanten des Graphen, wird auch die maximale edit distance aus den NM-tags der Reads angegeben. Diese kann entweder dem Schwellenwert oder seinem Default-Wert entsprechen oder auch je nach gewählten Minimap2-Option kleiner sein, als der Schwellenwert. Zudem wird die Anzahl der Substitutionen/SNP's, Insertionen und Deletionen angegeben.

Zusammenhangskomponenten extrahieren:

Die Bestimmung der Zusammenhangskomponenten inkl. Indexierung erfolgt durch graph-tool selbst und kann in $O(V+E)$ Zeit durchgeführt werden. Die Indexnummern (integer) werden dabei für jeden Knoten der Zusammenhangskomponente seinen Knoteneigenschaften hinzugefügt. Zusammenhangskomponenten die mehr als nur 1 Element besitzen, werden anschließend für die Berechnung des ILP einer Liste hinzugefügt. Im Log file wird die Anzahl der Knoten aller Komponenten als Histogramm festgehalten. Ebenso wird dort für alle Komponenten mit mehr als einem Element die Anzahl ihrer Knoten und Kanten und deren Eigenschaften aufgelistet.

3.4 Bestimmung der maximalen Likelihood der Allele

3.4.1

3.5 Bestimmung der maximalen Likelihood der Loci

3.5.1

3.6 Ausgabe der Loci im VCF-Format

3.6.1

Kapitel 4

Evaluation an simulierten Datensätzen

4.1

4.1.1

Kapitel 5

Zusammenfassung und Ausblick

5.1

5.1.1

Anhang A

Weitere Informationen

Abbildungsverzeichnis

Algorithmenverzeichnis

Literaturverzeichnis

- [1] WATSON, J. D. and F. H. CRICK: *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid*. Nature, 171:737–8, Apr 1953.
- [2] MARTIN, R. G., J. H. MATTHAEI, O. W. JONES, and M. W. NIRENBERG: *Ribonucleotide composition of the genetic code*. Biochemical and biophysical research communications, 6:410–4, Jan 1962.
- [3] MATTHAEI, H. and M. W. NIRENBERG: *The dependence of cell-free protein synthesis in e. coli upon rna prepared from ribosomes*. Biochemical and biophysical research communications, 4:404–8, Apr 1961.
- [4] *An integrated encyclopedia of dna elements in the human genome*. Nature, 489:57–74, Sep 2012.
- [5] TSAGAKIS, IOANNIS, KATERINA DOUKA, ISABEL BIRDS, and JULIE L. ASPDEN: *Long non-coding rnas in development and disease: conservation to mechanisms*. The Journal of pathology, 250:480–495, Apr 2020.
- [6] O'DONNELL, MICHAEL, LANCE LANGSTON, and BRUCE STILLMAN: *Principles and concepts of dna replication in bacteria, archaea, and eukarya*. Cold Spring Harbor perspectives in biology, 5, Jul 2013.
- [7] CHAGIN, VADIM O., JEFFREY H. STEAR, and M. CRISTINA CARDOSO: *Organization of dna replication*. Cold Spring Harbor perspectives in biology, 2:a000737, Apr 2010.
- [8] PRIOLEAU, MARIE-NOËLLE and DAVID M. MACALPINE: *Dna replication origins-where do we begin?* Genes & development, 30:1683–97, Aug 2016.
- [9] VIGNAL, ALAIN, DENIS MILAN, MAGALI SANCRISTOBAL, and ANDRÉ EGGEN: *A review on snp and other types of molecular markers and their use in animal genetics*. Genetics, selection, evolution : GSE, 34:275–305, May-Jun 2002.
- [10] SACHIDANANDAM, R., D. WEISSMAN, S. C. SCHMIDT, J. M. KAKOL, L. D. STEIN, G. MARTH, S. SHERRY, J. C. MULLIKIN, B. J. MORTIMORE, D. L. WILLEY, S. E. HUNT, C. G. COLE, P. C. COGGILL, C. M. RICE, Z. NING, J. ROGERS, D. R.

- BENTLEY, P. Y. KWOK, E. R. MARDIS, R. T. YEH, B. SCHULTZ, L. COOK, R. DAVENPORT, M. DANTE, L. FULTON, L. HILLIER, R. H. WATERSTON, J. D. MCPHERSON, B. GILMAN, S. SCHAFFNER, W. J. VAN ET TEN, D. REICH, J. HIGGINS, M. J. DALY, B. BLUMENSTIEL, J. BALDWIN, N. STANGE-THOMANN, M. C. ZODY, L. LINTON, E. S. LANDER, and D. ALTSHULER: *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms*. Nature, 409:928–33, Feb 2001.
- [11] KRUGLYAK, L.: *The use of a genetic map of biallelic markers in linkage studies*. Nature genetics, 17:21–4, Sep 1997.
- [12] KWOK, PUI-YAN and XIANGNING CHEN: *Detection of single nucleotide polymorphisms*. Current issues in molecular biology, 5:43–60, Apr 2003.
- [13] NIELSEN, RASMUS: *Population genetic analysis of ascertained snp data*. Human genomics, 1:218–24, Mar 2004.
- [14] SHRIVER, MARK D., GIULIA C. KENNEDY, ESTEBAN J. PARRA, HEATHER A. LAWSON, VIBHOR SONPAR, JING HUANG, JOSHUA M. AKEY, and KEITH W. JONES: *The genomic distribution of population substructure in four populations using 8,525 autosomal snps*. Human genomics, 1:274–86, May 2004.
- [15] AKEY, JOSHUA M., GE ZHANG, KUN ZHANG, LI JIN, and MARK D. SHRIVER: *Interrogating a high-density snp map for signatures of natural selection*. Genome research, 12:1805–14, Dec 2002.
- [16] WANG, D. G., J. B. FAN, C. J. SIAO, A. BERNO, P. YOUNG, R. SAPOLSKY, G. GHANDOUR, N. PERKINS, E. WINCHESTER, J. SPENCER, L. KRUGLYAK, L. STEIN, L. HSIE, T. TOPALOGLOU, E. HUBBELL, E. ROBINSON, M. MITTMANN, M. S. MORRIS, N. SHEN, D. KILBURN, J. RIOUX, C. NUSBAUM, S. ROZEN, T. J. HUDSON, R. LIPSHUTZ, M. CHEE, and E. S. LANDER: *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome*. Science (New York, N.Y.), 280:1077–82, May 1998.
- [17] SANGER, F. and A. R. COULSON: *A rapid method for determining sequences in dna by primed synthesis with dna polymerase*. Journal of molecular biology, 94:441–8, May 1975.
- [18] MULLIS, K., F. FALOONA, S. SCHARF, R. SAIKI, G. HORN, and H. ERLICH: *Specific enzymatic amplification of dna in vitro: the polymerase chain reaction*. Cold Spring Harbor symposia on quantitative biology, 51 Pt 1:263–73, 1986.

- [19] GAWAD, CHARLES, WINSTON KOH, and STEPHEN R. QUAKE: *Single-cell genome sequencing: current state of the science*. Nature reviews. Genetics, 17:175–88, Mar 2016.
- [20] KÖSTER, JOHANNES and SVEN RAHMANN: *Snakemake—a scalable bioinformatics workflow engine*. Bioinformatics, 28(19):2520–2522, 08 2012.
- [21] KÖSTER, JOHANNES and SVEN RAHMANN: *Building and Documenting Workflows with Python-Based Snakemake*. In BÖCKER, SEBASTIAN, FRANZISKA HUFISKY, KERSTIN SCHEUBERT, JANA SCHLEICHER, and STEFAN SCHUSTER (editors): *German Conference on Bioinformatics 2012*, volume 26 of *OpenAccess Series in Informatics (OASIs)*, pages 49–56, Dagstuhl, Germany, 2012. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- [22] MARTIN, MARCEL: *Cutadapt removes adapter sequences from high-throughput sequencing reads*. EMBnet.journal, 17(1):10–12, 2011.
- [23] ANDREWS, SIMON, FELIX KRUEGER, ANNE SEGONDS-PICHON, LAURA BIGGINS, CHRISTEL KRUEGER, and STEVEN WINGETT: *FastQC*. Babraham Institute, January 2012. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [24] EWELS, PHILIP, MÅNS MAGNUSSON, SVERKER LUNDIN, and MAX KÄLLER: *Multic: summarize analysis results for multiple tools and samples in a single report*. Bioinformatics (Oxford, England), 32:3047–8, Oct 2016.
- [25] LI, HENG: *Minimap2: pairwise alignment for nucleotide sequences*. Bioinformatics, 34(18):3094–3100, 05 2018.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie Zitate kenntlich gemacht habe.

Dortmund, den 10. Dezember 2020

Muster Mustermann

