

Bachelorarbeit

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

Antonie Vietor

Gutachter:

Dr. rer. nat. Johannes Köster

Prof. Dr. rer. nat. Sven Rahmann

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

<b>1</b>	<b>Einleitung</b>	<b>1</b>
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 . . . . .	9
1.3	RAD-Sequencing . . . . .	9
1.4	Sequenzierungsdaten und -formate . . . . .	9
1.4.1	Reads . . . . .	9
1.4.2	Sequenzalignments . . . . .	9
1.4.3	Varianten . . . . .	10
<b>2</b>	<b>Analyse von RAD-Seq-Daten</b>	<b>11</b>
2.1	Problemstellung . . . . .	11
2.2	Formale Definition der Problemstellung . . . . .	12
2.3	Lösungsansatz und Modell . . . . .	12
2.3.1	Graph und Zusammenhangskomponenten . . . . .	13
2.3.2	Approximation von PairHMM mittels Minimap2-Alignments . . . . .	15
2.3.3	Allele-Fractions mit maximaler Likelihood . . . . .	18
2.3.4	Locuszuordnung mit maximaler Likelihood . . . . .	20
2.3.5	Varianten und Genotyp . . . . .	21
<b>3</b>	<b>Implementierung</b>	<b>23</b>
3.1	Preprocessing . . . . .	23
3.2	Edit-Distanzen . . . . .	24

3.3	Konstruktion des Graphen . . . . .	24
3.3.1	Knoten des Graphen . . . . .	25
3.3.2	Kanten des Graphen . . . . .	25
3.3.3	Bestimmung der Zusammenhangskomponenten . . . . .	26
3.4	Bestimmung der Allele-Fractions mit maximaler Likelihood . . . . .	28
3.4.1	Bestimmung der Kandidatenallele und ihrer möglichen Häufigkeits- verteilung . . . . .	28
3.4.2	Berechnung der Likelihoods der Allele anhand der Allele-Fractions .	30
3.5	Bestimmung der maximalen Likelihood der Loci . . . . .	35
3.5.1	Bestimmung der möglichen Loci . . . . .	35
3.5.2	Bestimmung der wahrscheinlichsten Locuszuordnung der Allele-Fraction mit maximaler Likelihood . . . . .	41
3.6	Ausgabe der wahrscheinlichsten Loci als VCF-Datei . . . . .	45
<b>4</b>	<b>Laufzeitanalyse</b>	<b>47</b>
4.1	Graphkonstruktion und Bestimmung der Zusammenhangskomponenten . . .	47
4.2	Laufzeit zur Bestimmung der Allele-Fractions mit maximaler Likelihood . .	49
4.3	Laufzeit zur Bestimmung der Loci mit maximaler Likelihood . . . . .	55
4.4	Abschließende Laufzeitanalyse des gesamten Algorithmus . . . . .	57
<b>5</b>	<b>Evaluation an simulierten Datensätzen</b>	<b>59</b>
5.1	. . . . .	59
5.1.1	. . . . .	59
<b>6</b>	<b>Zusammenfassung und Ausblick</b>	<b>61</b>
6.1	. . . . .	61
6.1.1	. . . . .	61
	<b>Abbildungsverzeichnis</b>	<b>63</b>
	<b>Algorithmenverzeichnis</b>	<b>65</b>
	<b>Literaturverzeichnis</b>	<b>71</b>
	<b>Erklärung</b>	<b>71</b>

# 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 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

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. Punktmutationen in Form von Insertionen und Deletionen (Indels) 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. Auch bei Indels von mehreren Basen kann es zu einem Frame-Shift kommen, ändert sich dadurch allerdings die Länge des codierenden Abschnitts um drei Basen oder dem Vielfachen hiervon, so bleibt das Leseraster erhalten.

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 unterscheidet 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-Sequenzierung (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 SNPs und kleinen Indels. 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 jedem Probengemisch 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

Kosteneffizienz, gepoolte Probengemische gleichzeitige testung im hochdurchsatzverfahren

## 1.4 Sequenzierungsdaten und -formate

### 1.4.1 Reads

Identifikationsbezeichnungen, der Basensequenz, Basenqualität

### 1.4.2 Sequenzalignments

Häufig werden für die Analyse von Sequenzierungsdaten sogenannte Alignments erstellt, bei denen eine Readsequenz (Query) gegen eine andere Sequenz (Reference) verglichen wird. In Abhängigkeit von ihren Übereinstimmungen (Matches) und Unterschieden (Mismatches) werden die Sequenzen einander zugeordnet. Ein solches Alignment im SAM/BAM-Format wird auch in der vorliegenden Arbeit zur weiterführenden Analyse verwendet.

Das SAM/BAM-Format enthält unter anderem die ID's der Query- und Reference-Sequenzen, Informationen zur Basenqualität, Readlänge und -sequenz, den sogenannten CIGAR-String sowie verschiedene optionale Tags [20, 21], wie beispielsweise den NM-Tag, der die Edit-Distanz angibt (vgl. Kap. 2.3.1). Die Edit-Distanz ist die mindestens notwendige Anzahl von Ersetzungs-, Einfügungs- und Löschungsoperationen, um die Sequenz des Source-Knotens in die Sequenz des Target-Knotens zu transformieren. Auf DNA-Ebene entspricht dies den Punktmutationen im Sinne von Substitutionen, Insertionen und Deletionen (vgl. Kap. 1.1.5). Der CIGAR-String hingegen ist eine kondensierte Darstellung der Unterschiede zwischen Query- und Reference-Sequenz. In im SAM-Format werden darin Matches mit  $M$  oder  $=$ , Mismatches mit  $X$  oder  $S$ , Insertionen mit  $I$  und Deletionen mit  $D$  sowie jeweils mit der Anzahl der betroffenen Basen codiert (siehe auch Kap. 1.1.5). Aus ihrer Summe mit Ausnahme und abzüglich der Deletionen ergibt die Readlänge. So ergibt beispielsweise der CIGAR-String  $69 = 1X24 =$  ein Matching der ersten 69 Basen des Queryreads beim Abgleich mit dem Referenzread, anschließend ist bei einer Base ein Mismatch aufgetreten und schließlich folgen 24 Basen die auf den Referenzread matchen. Die Readlänge beträgt somit insgesamt 94 Basen und die Edit-Distanz im NM-Tag besitzt den Wert 1.

Daneben gibt es weitere Darstellungsmöglichkeiten von Veränderungen der Querysequenz gegenüber der Referenzsequenz. So können im cs-Tag des SAM-Formats verkürzt (short) oder vollständig mit Angabe der gesamten Readsequenz die genauen Veränderungen mit Bezeichnung der betreffenden Basen angegeben werden. Bezüglich des oben genannten Beispiels kann der dazugehörige short cs-Tag `69*tc : 24` lauten und zeigt damit an, dass nach Base 69 ein Basenaustausch von Thymin gegen Cytosin erfolgt ist.

Beim BAM-Format handelt es sich um einen komprimierten Binärfile [20] mit den gleichen Informationen wie sie in einem SAM-File enthalten sind. Mit gängigen Tools, wie beispielsweise Samtools-view [21], können beide Formate in einander umgewandelt werden.

### 1.4.3 Varianten

Variant Call Format (VCF, [22]) geschrieben. Wie für dieses Format üblich, gibt es acht notwendige tabulatorgetrennte Spalten: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO. Außerdem wird hier die optionale Spalte FORMAT sowie eine Spalte zur Probe verwendet.

## Kapitel 2

# Analyse von RAD-Seq-Daten

### 2.1 Problemstellung

Ausgangsmaterial für RAD-Seq-Analysen sind die Reads aus in der Regel gepoolten Probengemischen mehrerer Individuen. Durch entsprechende Barcodemarkierungen können die Reads den einzelnen Individuen zugeordnet werden. Durch die Verwendung gepoolter Probengemischen ist das Verfahren kostengünstiger und zeitsparender als die meisten anderen NGS-Verfahren. Die Reads stammen aus dem gesamten Genom, wobei sie nur kleinere Sequenzabschnitte zwischen den Schnittstellen der Restriktionsenzyme repräsentieren und damit nicht das gesamte Genom abbilden. Ein deutlicher Vorteil des Verfahrens besteht darin, die Reads ohne Kenntnis eines Referenzgenoms möglichen Loci zuordnen zu können. Auch die Loci und ihre Sequenz sind somit anfangs unbekannt und werden erst durch die Analyse ermittelt.

Das RAD-Sequencing-Verfahren bedingt, dass durch die Restriktionsenzyme die DNA nur in innerhalb spezifischer Sequenzen geschnitten wird und dass durch PCR und Sequenzierung mehrere Reads erzeugt werden, die jeweils vom gleichen Locus stammen. Das hier implementierte Tool NodeRAD stellt einen Prototypen dar, der im Gegensatz zu derzeit etablierten RADSeq-Tools, wie beispielsweise dem Tool Stacks [23], mit Hilfe von nur zwei Parametern diese Zuordnung realisiert. Das Clustering der Reads sowie die Identifizierung und Locizuordnung sollen dabei unter Berücksichtigung der Ploidie ohne weiteres Vorwissen allein auf Grundlage der Sequenzierfehlerrate und Heterozygotiewahrscheinlichkeit erfolgen. Die ermittelten Loci sollen für anschließende Diversitätsanalysen zur Verfügung stehen.

## 2.2 Formale Definition der Problemstellung

Gegeben sei eine Menge von Reads eines Individuums  $D = (s_1, \dots, s_m) \in \{A, C, G, T\}^{k^m}$  mit der Readlänge  $k$ . In den Reads sind zudem für jede Base Informationen zur Qualität der Sequenzierung  $Q = (q_1, \dots, q_m) \in [0, 1]^{k^m}$  enthalten. Weiterhin seien für Substitutionen, Insertionen und Deletionen die Heterozygotiewahrscheinlichkeit  $\eta = (\eta_{sub}, \eta_{ins}, \eta_{del})$  und für Indels die Sequenzierfehlerrate  $\epsilon = (\epsilon_{ins}, \epsilon_{del})$  gegeben. Ebenso ist die Ploidie  $\phi$  des untersuchten Organismus bekannt. In der vorliegenden Arbeit wird im Model und beim implementierten Prototypen ein diploider Chromosomensatz vorausgesetzt, für höherploide Organismen ist eine Anpassung hinsichtlich des Sequenzalignments notwendig (siehe Kap. 6).

Ziel ist die Zuordnung der Reads zu den Loci unter Berücksichtigung von  $\epsilon$  und  $\eta$  sowie die Ausgabe der Menge der ermittelten Loci mit den Sequenzen der beteiligten Allele. Dabei soll die Menge von Loci gefunden werden, welche die beobachteten Reads am besten erklären kann, das heißt welche die höchste Wahrscheinlichkeit in Zusammenschau mit  $\epsilon$  und  $\eta$  besitzt.

## 2.3 Lösungsansatz und Modell

Das Problem wird als gerichteter Graph betrachtet (siehe Kap. 2.3.1), dessen Knoten die einzelnen Reads beinhalten und dessen Kanten auf einem Sequenzalignment der Reads basieren. Für die Zuordnung der Reads soll das Problem in Teilprobleme aufgeteilt werden, für jedes Teilproblem wird dann eine Lösung ermittelt. Der Graph wird daher entsprechend seiner Zusammenhangskomponenten partitioniert. Die in einer Zusammenhangskomponente vorkommenden Sequenzen werden in Abhängigkeit von festgelegten Schwellwerten als Kandidatenallele betrachtet und jeweils mit sämtlichen Reads der Zusammenhangskomponente verglichen. Dabei sollen diejenigen Allele identifiziert werden, von denen die übrigen Reads der Komponente am wahrscheinlichsten durch Sequenzierfehler entstanden sind (Kap. 2.3.3). Die Likelihoodberechnung hierfür wird durch die Basenqualität der Reads und Sequenzierfehlerrate bestimmt und erfolgt über alle Kombination möglicher Häufigkeitsverteilungen der Allele (Allele-Fractions).

Die Allele-Fraction mit der höchsten Likelihood  $vaf_{max}$  soll anschließend für die Zuordnung der Loci verwendet werden (Kap. 2.3.4). Hierfür wird die Likelihood für alle Loci-Kombinationen bestimmt, welche zu der ermittelten  $vaf_{max}$  passen. Für die Berechnung der Likelihoods der Locizuordnung werden die Heterozygotiewahrscheinlichkeiten verwendet. Die Loci-Kombination mit der größten Likelihood wird schließlich als Lösung der



betreffenden Zusammenhangskomponente ausgegeben (Kap. 2.3.5).

Für jede Zusammenhangskomponente wird auf diese Weise also die wahrscheinlichste Lösung ermittelt, deren Komposition von Loci die beobachteten Reads innerhalb der Zusammenhangskomponente am besten erklären kann. Die Loci mit maximaler Likelihood aus allen Zusammenhangskomponenten werden schließlich durch NodeRAD im VCF-Format ausgegeben. Eine Übersicht der einzelnen Schritte des Workflows findet sich in Abbildung 2.1.

### 2.3.1 Graph und Zusammenhangskomponenten

Im Preprocessing wird aus den Reads zunächst ein paarweises Sequenzalignment mit Hilfe des Tools Minimap2 erzeugt [24]. Hierbei werden alle Reads mit einander verglichen und bei ausreichender Ähnlichkeit ihrer Sequenzen einander zugeordnet. Zusätzlich werden durch Minimap2 die Edit-Distanzen zwischen den jeweils verglichenen Readsequenzen bestimmt.

Wie bereits einleitend erwähnt, wird das Problem schließlich als gerichteter Graph  $G = (V, E)$  mit der Knotenmenge  $V$  und der Kantenmenge  $E$  betrachtet. Die Knoten repräsentieren dabei die einzelnen Reads. Entsprechend dem von Minimap2 erzeugtem Sequenzalignment werden die Knoten der Reads durch Kanten verbunden. Die Kanten des Graphen repräsentieren somit den Vergleich der Sequenzen der beiden angrenzenden Reads. Die Anzahl der Kanten kann zusätzlich durch die von Minimap2 bestimmten Edit-Distanzen reguliert werden. Diese soll über einen konfigurierbaren Schwellwert flexibel wählbar sein und ermöglicht ein zusätzliches Filtern der Kanten zugunsten von Laufzeit und Effizienz.

Da aufgrund des Sequenzalignments nur Reads mit einander verbunden werden, deren Sequenzen eine gewisse Ähnlichkeit zu einander aufweisen und da die RAD-Sequencing-Daten für gewöhnlich mehrere Loci beinhalten, ist der Graph in der Regel nicht zusammenhängend. Dies ermöglicht eine Unterteilung des Gesamtproblems in kleinere Teilprobleme durch Betrachtung der einzelnen Zusammenhangskomponenten des Graphen. Für diese Teilprobleme wird dann nach Lösungen gesucht. Da sich die Konstellation der Zusammenhangskomponenten direkt aus dem Sequenzalignment ergibt, sind die so entstandenen Cluster nicht notwendiger Weise nur einem einzigen Locus zuzuordnen. Vielmehr können die einzelnen Zusammenhangskomponenten auch jeweils mehrere Loci beinhalten.

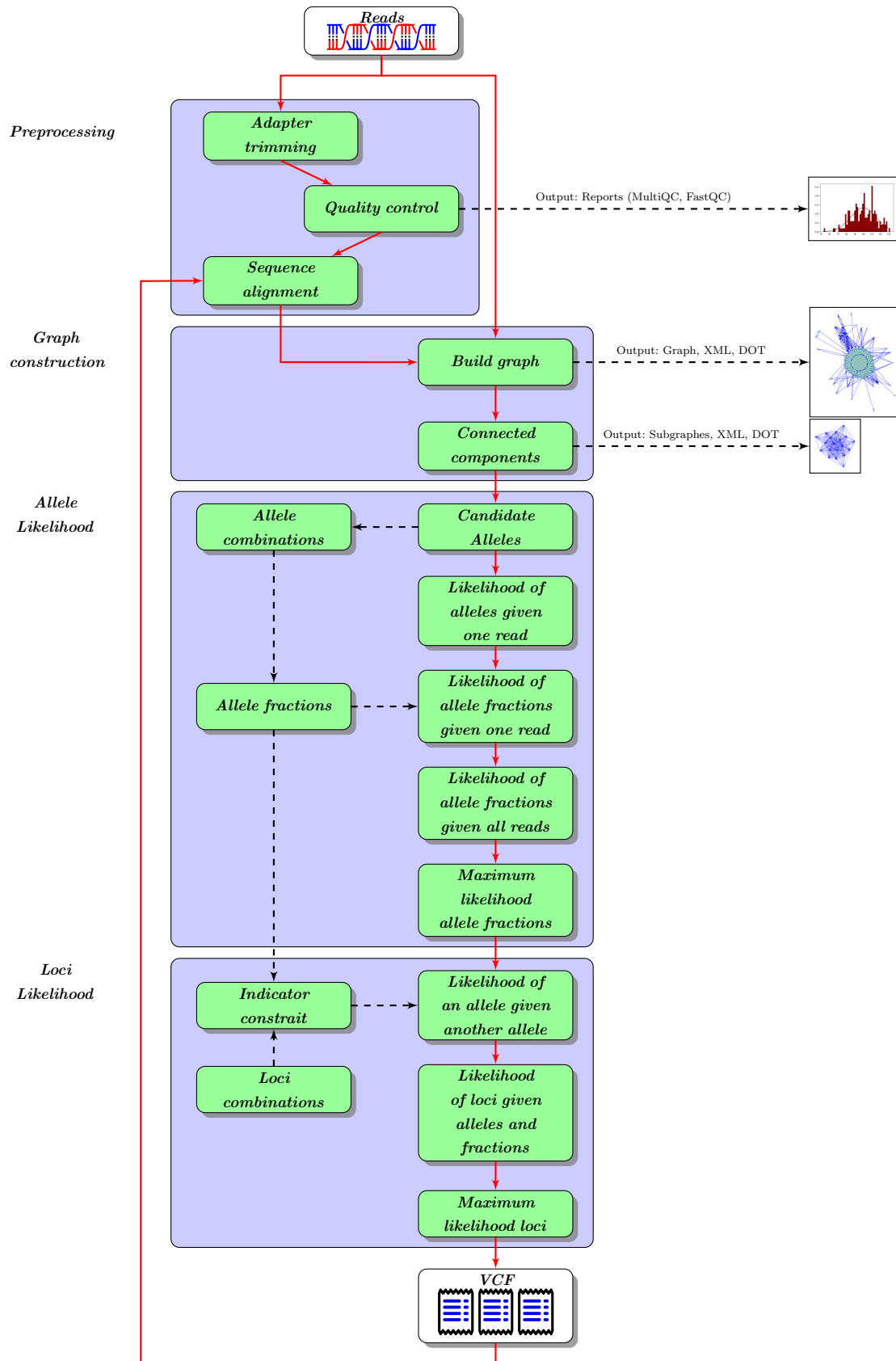


Abbildung 2.1: Übersicht der Prozesse des Workflows

### 2.3.2 Approximation von PairHMM mittels Minimap2-Alignments

Wie bereits erwähnt, werden die Kanten des Graphen durch das Sequenzalignment repräsentiert. Dabei führt Minimap2 einen paarweisen Readvergleich anhand der DNA-Sequenzen der Reads durch. Die Sequenz jedes Reads (Querysequenz) wird mit jeder Sequenz der übrigen Reads (Referenzsequenz) verglichen. Reads mit hoher Ähnlichkeit zueinander werden in das Alignment aufgenommen. Das Ergebnis ist ein Mapping derjenigen Reads, die eine hohe Wahrscheinlichkeit besitzen, durch Sequenzierfehler, Variationen oder Mutationen aus einem gemeinsamen Allel hervorgegangen zu sein. Somit basieren die Kanten des Graphen auf den Wahrscheinlichkeiten des Sequenzalignments.

Das zugrundeliegende Modell ist hierbei ein vollständiger gerichteter Graph dessen Kantengewichte durch den paarweisen Sequenzvergleich eines pair Hidden Markov Models (pairHMM) [25] errechnet werden. Das beobachtete Ergebnis der Referenzsequenz wird dabei durch Abfolgen von Mismatch-, Insertions und Deletions-Operationen der Querysequenz erklärt. Jede dieser Operationen stellt einen Zustand dar, der Übergang zwischen den Zuständen wird über Wahrscheinlichkeiten abgebildet. Die Zustände selbst sind allerdings nicht direkt beobachtbar. Die möglichen Kombinationen von Zustandsübergängen, die zum beobachteten Ergebnis der Referenzsequenz führen, können aber errechnet und als Matrix dargestellt werden. In der Matrix kann dann der wahrscheinlichste Pfad ermittelt werden. Hierfür muss gelten, dass auf diesem Pfad das Produkt über die Zustandsübergangswahrscheinlichkeiten der Matrix maximal ist. Als Ergebnis erhält man somit die wahrscheinlichste Zustandssequenz, um die Querysequenz in die Referenzsequenz zu transformieren.

===== draft =====

ergibt sich dabei aus dem Produkt der Zustandsübergangswahrscheinlichkeiten. kann aber die wahrscheinl Als Resultat entsteht eine Matrix mit den Daher soll im vorliegenden Modell das Sequenzalignment durch Minimap2 als Approximation des zugrundeliegenden PairHMM betrachtet werden, welches

stellt dabei eine Approximation des pair Hidden Markov Models zum paarweisen Readvergleich dar (siehe Kap. 2.3.2). Grundlage für die Berechnung der Likelihoods der Kandidatenallele und der Loci ist das durch Minimap2 ermittelte Sequenzalignment. Minimap2 berechnet hier die

Dieses repräsentiert einen Pfad im der pairHMM-Matrix, und zwar den Wahrscheinlichsten Pfad. Dieser dominiert ohnehin die Wahrscheinlichkeit die das pairHMM berechnen würde. Daher kannst du auch einfach nur die Wahrscheinlichkeit für diesen Pfad berechnen. Das ist ein Produkt über dem Alignment. Dieses liefert die wird als Approximation

als Approximation eine Pair hidden Markov Models [25] (PairHMM) betrachtet

[26] [27] [25] -> Berücksichtigung der Basenqualität  $q_r \Rightarrow$  Sequenzierfehler fließen mit geringerer likelihood in die weitere Berechnung ein -> werden weniger berücksichtigt und kommen seltener vor, z.T. als noise entfernt -> loci likelihood wird mit den sequenzen mit seq-fehlern geringer sein -> seq-fehler gefiltert, da es nicht möglich ist eine bessere lh für z.B. 3 Allele zu bekommen (von denen eines aber auf einem seq-fehler beruht), wenn es nur 2 tatsächliche Allele in einer komp gibt, d.h. lh für 2 Allele ist in diesem fall höher als für 3 -> loci likelihood enthält mutationen

paarweises Alignment aus Minimap2 als Teilproblem -> fast jede Beobachtung wird berücksichtigt

Bewertung/wslk ob ein read aus einem anderen read entstanden ist

Hierbei repräsentiert das durch Minimap2 bestimmte Sequenzalignment bereits den wahrscheinlichsten Pfad durch die pairHMM-Matrix. Da dieser Pfad ohnehin die Wahrscheinlichkeit des pairHMM dominieren würde, wird zugunsten der Laufzeit direkt auf das Alignment von Minimap2 zurückgegriffen, um die Likelihoods zwischen den Readsequenzen zu bestimmen. Dabei werden die Sequenzierfehlerrate  $\epsilon$  und Basenqualität  $q_{query}$  durch die bereits zuvor ermittelte geschätzte Fehlerrate  $p_{query}$  berücksichtigt. Die Likelihood  $pairHMM_{\epsilon, q_{query}}(s_{ref} | s_{query})$ , dass der Queryread aus dem Referenzread allein durch Sequenzierfehler und Mutationen entstanden ist, errechnet sich schließlich aus dem Produkt der Likelihoods  $L_i$  für jede Base  $b$  an jeder Position  $i$  innerhalb der Sequenz  $s$  des Queryreads  $s_{query}$  im Vergleich zum Referenzread  $s_{ref}$ .

$$pairHMM_{\epsilon, q_{query}}(s_{query} | s_{ref}) = \prod_{i=1}^k L_i \quad (2-5)$$

!!! [28] seq-error

Jede Base  $b$  an Position  $i$  einer Readsequenz  $s$  der Länge  $k$  lässt sich also definieren als  $b \in \{b_i \in \{A, C, G, T\}^k, b_i \in s \mid i = 1, \dots, k\}$ . Seien  $b_{i_{ref}}$  und  $b_{i_{query}}$  die Basen der Query- und der Referenzsequenzen an Position  $i$  einer Sequenz und  $p_{i_{query}}$  die geschätzte Fehlerrate von  $b_{i_{query}}$ , die sich aus dem Phred Quality Score  $Q$  nach (3-3) ergibt. Seien zudem  $m_{sub}$ ,  $m_{ins}$  und  $m_{del}$  die über die Konfigurationsdatei festgelegten Mutationsraten für Substitutionen, Insertionen und Deletionen. Dann errechnet sich die Likelihood  $L_i = Pr(b_{i_{ref}} | b_{i_{query}})$  an der Position  $i$  im Falle eine Matches unter Berücksichtigung der geschätzten Fehlerrate durch:

$$L_{i_{match}} = 1 - p_{i_{query}} \quad (2-6)$$

Bei einem Mismatch dagegen müssen die Wahrscheinlichkeiten von Mutationen und Sequenzierfehlern berücksichtigt werden. Im Falle einer Mutation muss in die Wahrschein-

lichkeit eines Matches auch die Mutationsrate des aufgetretenen Mismatches  $m_{rate} \in \{m_{sub}, m_{ins}, m_{del}\}$  einbezogen werden:

$$L_{i_{mut}} = m_{rate} \cdot (1 - p_{i_{query}}) \quad (2-7)$$

Die Wahrscheinlichkeit eines Sequenzierfehlers, also dass anstelle der sequenzierten Base tatsächlich eine der drei anderen Basen vorliegt, entspricht  $1/3$  der geschätzten Fehlerrate des Phred Quality Scores:

$$L_{i_{seqerr}} = \frac{1}{3} \cdot p_{i_{query}} \quad (2-8)$$

Aus (2-7) und (2-8) errechnet sich also die Likelihood bei einem Mismatch durch:

$$L_{i_{mismatch}} = (1 - m_{rate}) \cdot L_{seqerr} \cdot L_{mut} \quad (2-9)$$

Aus den Likelihoods von Matches (2-6) und Mismatches (2-7) kann somit schließlich nach (2-5) die Likelihood zwischen den Reads paarweise bestimmt werden.

———— nur bei diploidie -> sonst multiples Sequenzalignment notwendig.

(vgl. Kap. ??). Allerdings werden nun die Heterozygotiewahrscheinlichkeiten  $\eta_{sub}$ ,  $\eta_{ins}$  und  $\eta_{del}$  der Grapheigenschaften zur Ermittlung der Likelihood nach Formel (2-20) genutzt.

Das Sequenzalignment der Reads aus Minimapp2 repräsentiert bereits den wahrscheinlichsten Pfad in der pairHMM-Matrix, so dass über diesen Pfad die Likelihoods der Allele (Kap. 2.3.3) und der Loci (Kap. 2.3.4) bestimmt werden können.

Dabei entspricht im Falle eines Mismatches die Likelihood der in der Konfigurationsdatei angegebenen Heterozygotiewahrscheinlichkeit  $\eta$  für die betreffende Mutationsart, also  $\eta_{sub}$  für Substitutionen,  $\eta_{ins}$  für Insertionen bzw.  $\eta_{del}$  Deletionen. Sei  $i$  der Index der betreffenden Base innerhalb der betrachteten Allelsequenz und  $\eta_{rate} \in \{\eta_{sub}, \eta_{ins}, \eta_{del}\}$ , dann berechnet sich bei einem Mismatch die Likelihood  $L_i$  der Base nach Formel (2-10).

$$L_{i_{mismatch}} = \eta_{rate} \quad (2-10)$$

Im Falle eines Matches senkt die Möglichkeit eines Mismatches die Likelihood der betreffenden Base entsprechend um die Summe der Heterozygotiewahrscheinlichkeiten der genannten Mutationsarten (Formel (2-11)).

$$L_{i_{match}} = 1 - (\eta_{sub} + \eta_{ins} + \eta_{del}) \quad (2-11)$$

Die Likelihood  $Pr(T = a_{l_{j,2}} | S = a_{l_{j,1}}, \eta)$  des paarweisen Vergleichs der Allele  $a_{l_{j,1}}$  und  $a_{l_{j,2}}$  hinsichtlich der Zuordnung zu den Loci-Kombinationen errechnet sich schließlich aus dem Produkt der Wahrscheinlichkeiten der einzelnen Basen:

$$Pr(T = a_{l_{j,2}} | S = a_{l_{j,1}}, \eta) = pairHMM_{\eta}(a_{l_{j,1}}, a_{l_{j,2}}) = \prod_{i=1}^k L_i \quad (2-12)$$

### 2.3.3 Allele-Fractions mit maximaler Likelihood

Um die Allele zu identifizieren, von denen die beobachteten Reads einer Zusammenhangskomponente am wahrscheinlichsten stammen, muss zunächst die Menge der Kandidatenallele  $A = (a_1, \dots, a_n) \in \{A, C, G, T\}^{k^n}$  bestimmt werden. Bei größeren Clustern ist davon auszugehen, dass die Sequenzen solcher Kandidatenallele mehrfach in den Reads auftauchen. Um den Aufwand der Likelihoodberechnung bei größeren Cluster anpassen zu können, soll es möglich sein, selten auftretende Sequenzen herauszufiltern. Dies geschieht über zwei Grenzwerte, welche die Mindestgröße des Clusters und die Mindesthäufigkeit von Kandidatenallelsequenzen festlegen. Durch sie werden ab einer bestimmten Clustergröße nur diejenigen Sequenzen in die Auswahl der Kandidatenallele aufgenommen, die mit einer bestimmten Häufigkeit im Cluster vorkommen.

Anschließend wird die Wahrscheinlichkeit bestimmt, dass ein Read mit der Readsequenz  $s_r$  und mit der Fehlerrate  $\epsilon$  tatsächlich von einem bestimmten Allel stammt. Die Fehlerrate  $\epsilon$  ergibt sich dabei aus der Qualität der einzelnen Basen eines Reads. Jede Base der Sequenz des Source-Knotens wird mit der Sequenz des Target-Knotens verglichen. Bei einem sogenannten Mismatch unterscheiden sich die verglichenen Basen durch Substitutionen, Insertionen oder Deletionen. Im Falle eines Matches sind beide Basen identisch. Bei einem Match ergibt sich die Wahrscheinlichkeit  $L_{match}$ , dass es sich dabei im Source-Knoten tatsächlich um die korrekte Base handelt allein aus der Fehlerrate:

$$L_{match} = 1 - \epsilon \quad (2-13)$$

Bei einem Mismatch  $L_{mismatch}$  müssen die Wahrscheinlichkeiten für Sequenzierfehler  $L_{sequerr}$  (vgl. Formel (2-8)) und Mutationen  $L_{mut}$  (vgl. Formel (2-7)) sowie die Mutationsrate selbst  $m_{rate}$  mit einbezogen werden:

$$L_{mismatch} = (1 - m_{rate}) \cdot L_{sequerr} \cdot L_{mut} \quad (2-14)$$

Die genaue Berechnung der Likelihood der einzelnen Basen wird an späterer Stelle in Kap. 3.3.2 ausführlicher besprochen. Das Produkt aller Wahrscheinlichkeiten der einzelnen Basen ergibt schließlich die Likelihood für den paarweisen Vergleich zweier Reads im Sinne des pairHMM. Bei paarweisen Vergleich eines Allels  $a_i \in A$  mit einem Read mit der Sequenz  $s_r$  lässt sich die Likelihoodberechnung also wie folgt formulieren:

$$Pr(T = s_r | S = a_i, \epsilon) = pairHMM_{\epsilon, q_r}(a_i, s_r) \quad (2-15)$$

===== draft =====

0. noise nicht als Kandidaten-Allele -> liste lexicogra. sortierter Allele

1. lh zwischen 1 read und 1 allel; likelihood im pair: Wslk, dasss von 1 bestimmten Allel  $a_i$  mit Fehlerrate  $\epsilon$  tatsächlich die readsequenz  $s_r$  stammt

$s_r$  Sequenz von Read  $r$ ,  $q_r$  Basenqualität aller Basen von Read  $r$ ,  $a_i$  Allel  $i$  aus der Menge der Kandidatenallele  $A = (a_1, \dots, a_n)$  mit Anzahl  $n$  und  $\epsilon$  Sequenzierfehlerrate

$$Pr(T = s_r | S = a_i, \epsilon) = pairHMM_{\epsilon, q_r}(a_i, s_r) \quad (2-16)$$

2. Info über alle vafs aller Kandidatenallele für 1 read; Zusammenhangskomponenten kann auch >1 locus enthalten -> Fractions finden in der möglichen Menge von loci

Berechnung der Wslk, einen Read zu beobachten anhand der gegebenen Allele-Fraktion die Wslk der Fraktion  $\Theta_i$  muss sich aus den tatsächlichen Beobachtungen ergeben => Urnenmodell (entspricht binominal modell)

Urnenmodell: bsp. 2 farben in unterschiedl. Menge -> Wslk, wenn n Kugeln gezogen werden, dass dabei jeweils genau 50% von jeder Farbe gezogen wurden (Binomial Formel), Anzahl der roten Kugeln  $P$  (Erfolgswslk)  $\Theta_i$  und > 2 Farben => Multinomialverteilung  $\Theta_i$  + individuellen Verteilung  $Pr$ :  $\Theta_i * Pr$  es gilt  $\sum \Theta_i = 1$  => die mögliche Zuordnung entspricht der Häufigkeit die gezogen wurde  $\Theta_i$

Wslk.  $s_r$  zu beobachten unter der gegebenen Verteilung aus den Fraktionen; Wslk read zu sehen gegeben das Allel, nur ausgehend von Sequenzierfehlern

$\Theta = \theta_1, \dots, \theta_n) \in [0, 1]^n$  sind Allel-Fractionen,  $n$  ist Anzahl der Allele

$$Pr(s_r | \Theta = \theta_1, \dots, \theta_n) = \sum_{i=1}^n \theta_i \cdot Pr(T = s_r | S = a_i, \epsilon) \quad (2-17)$$

3. Für alle Reads:

1 read  $s_r$  mit  $n$  kandidaten allelen ->  $n$  mögliche Allele von denen  $s_r$  stammen kann -> es wird für jedes der allele wslk berechnet, dass die allele aus  $s_r$  entstanden sind

bsp. allel i hat frac 0.25, also 1/4 für i, allel i+1 mit 0.75 => zu 25% allel i und zu 75% allel i+1 -> über max lh wird bestimmt, wie die seq der allele zusammenpassen -> modellierung ähnlicher, unsicherheit bei der zuordnung der reads, wenn sie weder der dem eine noch dem anderen allel entstammen, z.b. bei 0.5/0.5 => Unsicherheit wird an schritt 3 weitergegeben

Produkt über alle reads für jede mögliche Kombination an fractions

max wählen innerhalb der conn. comp

insgesamt:

=>  $\Theta$  berechnen -> max lh aus allel-fractions

=> alles was an Unterschieden vorkommt, kann nur ein Sequenzierfehler sein

$m$  ist Anzahl der Reads,  $L \in \{l_i \in \mathbb{N}_{\leq n}^\phi | i = 1, \dots, g\}$  ist Menge der Loci,  $D = (s_1, \dots, s_m) \in \{A, C, G, T\}^{k^m}$  ist Menge der Readsequenzen/der Reads,

$$L(\Theta = \theta_1, \dots, \theta_n | D) = Pr(D | \Theta = \theta_1, \dots, \theta_n) = \prod_{r=1}^m Pr(s_r | \Theta = \theta_1, \dots, \theta_n) \quad (2-18)$$

für jede vaf wurde also likelihood über alle reads bestimmt -> Maximum

### 2.3.4 Locuszuordnung mit maximaler Likelihood

connected comp in loci aufsplitten

Zuordnung zu loci, also welche genomischen loci stecken dahinter, wie werden die allele den loci zugeordnet -> max parsimony, prinzip der einfachsten möglichen lösung

-> ploidy als eingabeparam -> zuordnung der kandidaten allele zu loci

lh ausrechnen aus anzahl der allele und loci -> gerade die loci, die die seq. der reads am besten unter der geg. ploidy und heterozygotie erklären haben max lh max\_lh\_vafs nur noch 1 Vektor

Zuordnung von cand-Allelen zu den loci => die beobachteten max\_lh\_allele\_fractions müssen sie durch die gewählten loci erklären, z.B. 2 loci => 1.locus: allel 0, 0 homozyg, 2. locus allel 0, 2 heterozygot

-> bei z.B 3 Allelen sind min 2 Loci notwendig => min # der loci, so dass alle ausgewählten allele passen -> diese menge muss gefunden werden und muss wieder max lh haben unter der annahme der loci-verteilung muss bestimmt werden ob sinnvoll, d.h allel-vafs müssen sich zu 1 summieren -> indikator-funktion: wenn 1, dann gilt  $\Pr(T|S)$ , d.h. Wslk, dass sich aus S die seq T gebildet hat

indicator-function  $z_l \in [0, 1]$ : wird 1, wenn anzahl des auftretens aller allele  $A = (a_1, \dots, a_n)$  in einer möglichen locus-verteilung  $l = (l_1, \dots, l_{g \cdot \phi})$  jeweils genau ihrer absoluten Häufigkeit aus der in Kap. 2.3.3 errechneten vaf mit maximaler Likelihood entspricht, also  $\theta_i \cdot g \cdot \phi$  entsprechen, andernfalls gilt  $z_l = 0$ . Seien dabei  $\Theta = \theta_1, \dots, \theta_n \in [0, 1]^n$  die vaf mit maximaler Likelihood,  $\phi$  die Ploidie und gilt für eine mögliche Locusverteilung  $l$  außerdem  $L \in \{l_j \in \mathbb{N}_{\leq n}^\phi \mid j = 1, \dots, g\}$ , so lässt sich die Indikatorfunktion wie folgt definieren:

$$z_l = \prod_{i=1}^n 1_{\sum_{j=1}^g \sum_{k=1}^{\phi} 1_{l_{j,k}=i} = \theta_i \cdot g \cdot \phi} \quad (2-19)$$

pHMM: Wslk, dass allel 2  $a_{l_{j,2}}$  aus allel 1  $a_{l_{j,1}}$  entstanden ist -> heterozygotie  $\eta$  (in conf konfigurierbar) =>  $match = 1 - (het_{sub} + het_{del} + het_{ins})$

$$\Pr(T = a_{l_{j,2}} \mid S = a_{l_{j,1}}, \eta) = pairHMM_\eta(a_{l_{j,1}}, a_{l_{j,2}}) \quad (2-20)$$

Kombi von loci -> lh berechnen

$$\Pr(\Theta, A \mid L = \{l_j \mid j = 1, \dots, g\}) = z_l \cdot \prod_{j=1}^g \Pr(T = a_{l_{j,2}} \mid S = a_{l_{j,1}}) \quad (2-21)$$

=> maximum der loci Likelihoods wählen



**2.3.5 Varianten und Genotyp**

mit den allelen  $A = (a_1, \dots, a_n)$  jedes locus  $L \in \{l_j \in \mathbb{N}_{\leq n}^\phi \mid j = 1, \dots, g\}$ , allelsequenz,  
genotyp

Menge von Loci, die die Beobachtungen erklären



## Kapitel 3

# Implementierung

Für das in Python implementierte RAD-Sequencing-Tool, NodeRAD, wurde zur Workflowintegration das Workflow Management System Snakemake verwendet [29, 30]. 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 (siehe Kap. 1.4.1), welche die single-end Reads der verschiedenen Individuen mit ihren IDs, der Basensequenz und Angaben zur Basenqualität enthält. Des Weiteren ist eine Tabelle im tsv-Format erforderlich, in der die Zuordnung der Probennamen zu den Individuen und ihren Barcode-Sequenzen definiert ist. Nach dem Preprocessing, der Qualitätskontrolle der Reads und dem Sequenzalignment erfolgt die RAD-Seq-Analyse durch NodeRAD. Hierbei werden die Wahrscheinlichkeiten der Allele-Fractions und der möglichen Loci bestimmt. Die Loci mit der höchsten Wahrscheinlichkeit werden schließlich mit den Sequenzen ihrer Allele in einer Datei im Variant Call Format (VCF) ausgegeben.

### 3.1 Preprocessing

Im Preprocessing werden durch das Tool Cutadapt [31] die Reads jedes Individuums anhand ihrer Barcodesequenzen identifiziert und extrahiert (Demultiplexing). Hiernach werden die Barcodesequenzen 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 [32]. Dabei werden einige allgemeine Statistiken zu den Rohdaten der Reads generiert,

wie beispielsweise zur Basenqualität, zum GC-Gehalt, dem Anteil an Duplikaten oder überrepräsentierten Sequenzen. Durch das Tool MultiQC [33] wird aus diesen Statistiken und den Log-Dateien von Cutadapt ein Report im HTML-Format 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 (vgl. Kap. 1.4.2) mit Hilfe des Tools Minimap2 [24] erstellt. Hierbei werden alle Readssequenzen paarweise mit einander verglichen. Das Ergebnis des Mappings wird anschließend durch Samtools-view [21] in das BAM-Format konvertiert und enthält neben den Angaben zur Query- und Referenzsequenz auch den CIGAR-String sowie den NM-Tag mit den Edit-Distanzen. Der CIGAR-String und der NM-Tag definieren wichtige Kanteneigenschaften des späteren Graphen.

## 3.3 Konstruktion des Graphen

NodeRAD benötigt als Input zu jedem Individuum die getrimmten single-end Reads sowie das Sequenzalignment. Zur Konstruktion des Graphen wird die Python-Library graph-tool [34] genutzt. Zunächst wird daraus für jedes Individuum ein eigener, gerichteter 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 Kap. 3.1) entnommen. Die Kanten,  $E$ , zwischen den Knoten ergeben sich aus dem Vergleich ihrer Sequenzen im Rahmen des Sequenzalignments mittels Minimap2 (siehe Kap. 3.2).

Zusätzlich entnimmt NodeRAD der Konfigurationsdatei des Workflows einige Konstanten und Grenzwerte für die späteren Berechnungen. Zu den Konstanten gehören die Sequenzierfehlerraten für Indels, die Heterozygotiewahrscheinlichkeiten für Substitutionen, Insertionen und Deletionen sowie die Ploidie des Chromosomensatzes der untersuchten Spezies. Diese werden im Script `noderad_main.py` im Graphen als Grapheigenschaften abgelegt. Als konfigurierbare Grenzwerte gibt es für NodeRAD einen Schwellenwert `edit-distance-graph` für die maximal zulässige Edit-Distanz, bei der zwei Knoten noch durch eine Kante verbunden werden sowie Schwellenwerte zum Filtern selten vorkommender Sequenzen (`threshold-seq-noise` für `small-clusters` und `large-clusters`) ober- und unterhalb einer bestimmten Clustergröße (`threshold-cluster-size`), die als Hinter-

grundrauschen nicht in der Berechnung Berücksichtigung finden sollen.

### 3.3.1 Knoten des Graphen

Die Knoten werden aus den FASTQ-Daten der getrimmten Reads mittels SeqIO aus der Library Biopython [35] ausgelesen und im Graphen mit den Knoteneigenschaften ihrer Basensequenz, einer internen ID sowie Angaben zur Basenqualität abgelegt. Die Codierung des Qualitystrings der Reads variiert je nach verwendeter Plattform. Daher wird er durch SeqIO ausgelesen und für jede Base in ein einheitliches Maß, den Phred Quality Score, decodiert [36]. Für jeden Knoten werden die Vektoren mit den Phred Quality Scores der Basen als Knoteneigenschaft gespeichert.

### 3.3.2 Kanten des Graphen

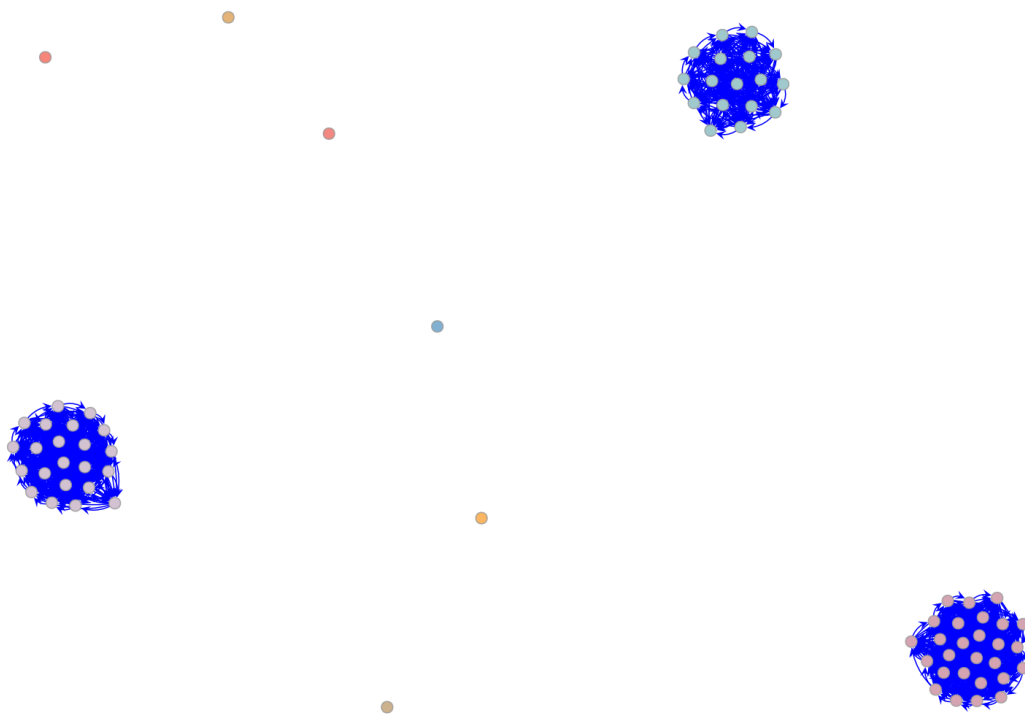
Die Kanten des Graphen definieren sich durch die mittels Minimap2 erzeugten Sequenzalignments (vgl. Kap. 3.2). Jedes Alignment zweier Reads entspricht im Graphen einer gerichteten Kante  $e = (source, target)$ , die den Vergleich der Query- zur Referenzsequenz repräsentiert. Sie verbindet somit zwei der zuvor aus der FASTQ-Daten erzeugten Knoten. Das Auslesen des Alignmentfiles im BAM-Format erfolgt mit Hilfe der Python-Library pysam [37]. Dabei werden die Edit-Distanzen aus dem NM-Tag genutzt, um nur Kanten in den Graphen aufzunehmen, die unterhalb des durch die Konfigurationsdatei festgelegten Grenzwertes `threshold_max_edit_distance` liegen. Neben den Edit-Distanzen wird als weitere Kanteneigenschaft der zu Tupeln decodierte CIGAR-String hinzugefügt. Zusätzlich kann zur Kontrolle oder für eine spätere Verwendung auch der cs-Tag (Kap. 1.4.2) als Kanteneigenschaft gespeichert werden, falls bei Minimap2 die Option zur Erzeugung des cs-Tags aktiviert wurde. Die CIGAR-Tupel werden durch pysam aus dem CIGAR-String (Kap. 1.4.2) geparsed. Hierbei handelt es sich um eine Liste von Tupeln, die jeweils aus Integer-Wertepaaren bestehen. Der erste Wert jedes Tupels gibt die spezifische Operation des Matches oder Mismatches an. So entspricht beispielsweise ein Wert von 7 oder 0 einem Match und ein Wert von 2 einer Deletion. Der zweite Werte jedes Tupels gibt die Anzahl der Basen an, die von der entsprechenden Operation betroffen sind. Die CIGAR-Tupel werden für die Berechnung der Likelihood zwischen zwei Knoten benötigt (Kap. 2.3.2). Dies erfolgt zum einen zur Bestimmung der Likelihood der Allele-Fractions (Kap. 3.4.2) und zum anderen, um die Likelihood der Loci-Kombinationen zu ermitteln.

Nach Abschluss der Graphkonstruktion werden für jedes Individuum Anzahl der Knoten und Kanten des Graphen in den Log-Dateien festgehalten. Als optionaler Output können über die Konfigurationsdatei und die Snakemake-Regel `noderad` auch die detaillierten

Graphinformationen sowie eine Visualisierung des Graphen ausgegeben werden. Die Graphinformationen wie Knoten, Kanten und ihre Eigenschaften können dabei im GraphML-, DOT-, GML- oder im binären gt-Format gespeichert werden. Die graphische Darstellung wird als pdf-Datei ausgegeben.

### 3.3.3 Bestimmung der Zusammenhangskomponenten

Die Bestimmung der Zusammenhangskomponenten erfolgt ebenfalls durch graph-tool [38]. Die Indexnummer jeder Zusammenhangskomponente wird den in ihr enthaltenen Knoten als Knoteneigenschaft hinzugefügt. Zusammenhangskomponenten mit mehr als einem Knoten werden als neuer eigenständiger Graph initialisiert und in einer Liste abgelegt. Hierfür wird aus dem Graphen für jede Komponente eine gefilterte Sicht erzeugt, die als neues Graph-Object gespeichert wird.



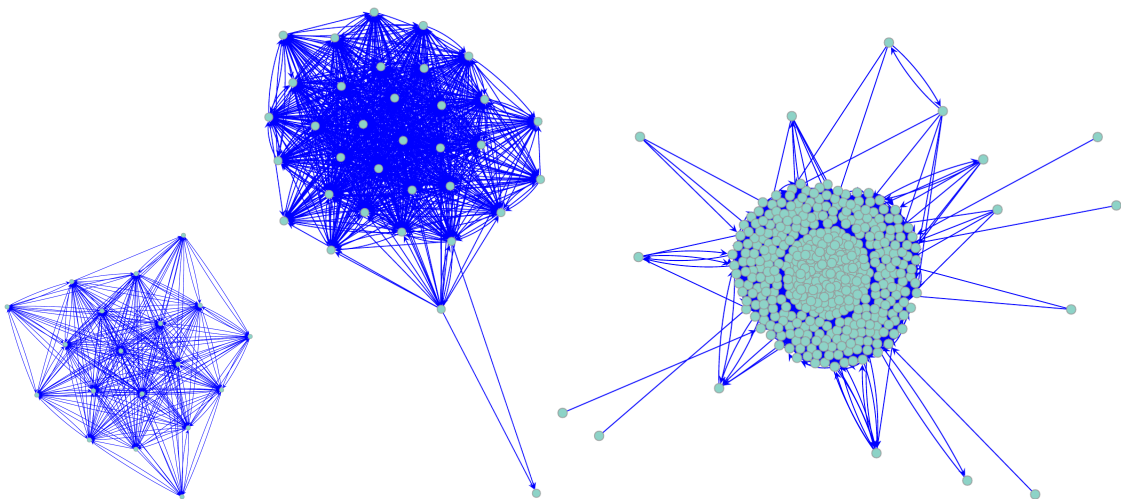
**Abbildung 3.1:** Ausschnitt einiger Zusammenhangskomponenten des Gesamtgraphen

Da alle weiteren Schritte des Algorithmus jeweils auf den einzelnen Komponenten durchgeführt werden, kann durch die Verwendung einer Liste von Graphen im Folgenden eine einfache Iteration über die Komponenten ausgeführt werden, ohne dass der Filtervorgang über alle Knoten jeder Komponente wiederholt werden muss. Zudem ermöglicht diese Datenstruktur eine effizientere Traversierung und Suche innerhalb der Zusammen-

hangskomponente, ohne dass für jede Komponente der gesamte Graph betrachtet werden muss. Der ursprüngliche Graph wird anschließend entfernt, um Arbeitsspeicher freizugeben.

In der Log-Datei wird die Anzahl der Knoten aller Zusammenhangskomponenten als Histogramm festgehalten. Ebenso wird dort für alle Komponenten mit mehr als einem Element die Anzahl ihrer Knoten, Kanten und Eigenschaften aufgelistet.

Über die Konfigurationsdatei und die Snakemake-Regel `noderad` können optional auch für die Zusammenhangskomponenten jeweils Visualisierungen (siehe Abbildung 3.2) und detaillierte Graphinformationen in den oben genannten Formaten des Graphen (Kap. 3.3.2) ausgegeben werden.



**Abbildung 3.2:** Beispiele von einzelnen Zusammenhangskomponenten unterschiedlicher Größe

Zudem kann optional auch der gesamte Graph mit den Komponentenindizes als Knoteigenschaften gespeichert werden. In der visuellen Darstellung werden dort seine Knoten entsprechend der zugehörigen Zusammenhangskomponente eingefärbt (siehe Abbildung 3.1).

## 3.4 Bestimmung der Allele-Fractions mit maximaler Likelihood

### 3.4.1 Bestimmung der Kandidatenallele und ihrer möglichen Häufigkeitsverteilung

Die einzelnen Zusammenhangskomponenten repräsentieren einen oder mehrere Loci. Alle weiteren Berechnungen aus diesem und den folgenden Kapiteln 3.5 und 3.6 werden für jede Komponente einzeln durchgeführt, dabei sollen die wahrscheinlichsten Allelsequenzen und die dazugehörigen Loci identifiziert werden. Die hierfür implementierten Funktionen finden sich im Modul `likelihood_operations.py`.

Zunächst werden mit der Funktion `get_candidate_alleles()` aus dem Modul `likelihood_operations.py` die Allelsequenzen ermittelt, die in der Zusammenhangskomponenten vorkommen und deren Häufigkeiten bestimmt. Übersteigt die Größe eines Clusters, d.h. die Knotenanzahl einer Zusammenhangskomponenten, den in der Konfigurationsdatei unter `threshold-seq-noise` festgelegten Grenzwert `large-clusters`, so werden nur diejenigen Sequenzen als Kandidatenallele ausgewählt, deren Häufigkeit über dem `large-clusters` Schwellenwert liegt. Dies dient dazu, bei großen Clustern Komplexität und Rechenaufwand zu reduzieren, da davon auszugehen ist, dass innerhalb eines großen Clusters echte Varianten einer Sequenz mehrfach vorkommen, wohingegen Artefakte und Sequenzierfehler eher vereinzelt auftreten. Dieses sog. Rauschen kann durch Anpassung der Schwellwerte in der Konfigurationsdatei herausgefiltert werden. Ebenso kann es jedoch vorkommen, dass auch bei kleinen Clustern viele verschiedene Sequenzen auftreten, so dass es durch eine hohe Anzahl von Kandidatenallele zu einem erheblichen Rechenaufwand bei der späteren Analyse kommen kann. Um dem entgegenwirken, kann durch einen weiteren unter `threshold-seq-noise` festgelegten Grenzwert `small-clusters` die Mindesthäufigkeit der Sequenzen bei kleineren Clustern unterhalb von `threshold-cluster-size` angegeben werden. Da bei kleinen Clustern auch einmalig registrierte Varianten einer Sequenz von Bedeutung sein können, sollte der Filtervorgang hier möglichst gering gehalten werden. Gilt `small-clusters = 1`, so wird die Option für kleine Cluster vollständig deaktiviert.

Die Kandidatenallele werden anschließend für deterministischere Ergebnisse lexikographisch sortiert und als Liste  $A_{cand}$  der Länge  $n_{cand}$  von `get_candidate_alleles()` zurückgegeben. Aus dieser Liste sollen nun die möglichen Häufigkeitsverteilungen in Abhängigkeit von der Ploidie bestimmt werden. Hierzu muss zunächst die aufgrund der Ploidie tatsächlich zu erwartende Anzahl von Allelen  $n_{alleles}$  bestimmt werden. Dies geschieht durch die Funktion `get_max_parsimony_n_alleles()` in `likelihood_operations.py`. Dadurch werden unnötige bzw. nicht mögliche Allelkombinationen eingespart und in der weiteren Berech-



nung nicht berücksichtigt. Ist die Ploidie  $\phi$  höher als die Anzahl der Kandidatenallele, so muss es mindestens genauso viele und bei Homozygotie auch mehrfach vorkommende Allele geben, damit die Ploidie erfüllt werden kann. Es muss also gelten  $n_{alleles} = \phi$ . Wurden dagegen mehr Allele beobachtet als aufgrund der Ploidie möglich sind und die Ploidie ist Teiler von  $n_{cand}$ , so können alle beobachteten Allele auch tatsächlich vorkommen, da die Zusammenhangskomponente auch mehrere Loci enthalten kann. Es gilt dann also  $n_{alleles} = n_{cand}$ . Ist dagegen die Anzahl der beobachteten Allele höher als die Ploidie, aber nicht ganzzahlig durch die Ploidie teilbar, so muss die Anzahl der Allele entsprechend erhöht werden. Das heißt, es müssen tatsächlich so viele Allele vorkommen, dass eine korrekte Ploidie erreicht wird, dass also die Ploidie eine Teiler von  $n_{alleles}$  wird. Die Anzahl der Allele muss also um die Ploidie abzüglich des Restes aus der Restdivision erhöht werden und es gilt  $n_{alleles} = n_{cand} + \phi - (n_{cand} \bmod \phi)$ .

Mit Hilfe der tatsächlich zu erwarteten Anzahl von Allelen  $n_{alleles}$  können nun alle Kombinationen möglicher Häufigkeitsverteilungen der Allele bestimmt werden (Funktion `get_candidate_vafs()` im Modul `likelihood_operations.py`). Diese Häufigkeitsverteilungen werden auch als Allele-Fractions bezeichnet. Hierfür werden zunächst alle möglichen Allelkombinationen ermittelt. Dies erfolgt nach dem Urnenmodell unter Auswahl von  $n_{alleles}$  Elementen mit Zurücklegen aus insgesamt  $n_{cand}$  verschiedenen Elementen und ohne Berücksichtigung der Reihenfolge. Dadurch berechnet sich die Anzahl möglicher Kombination nach (3-1) aus dem Binomialkoeffizienten der  $k$ -ten Ordnung aus  $n$  Elementen mit Zurücklegen [39, 40].

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)! \cdot k!} = \frac{(n_{alleles} + n_{cand} - 1)!}{(n_{alleles} - 1)! \cdot n_{cand}!} \quad (3-1)$$

Die Allelkombinationen werden mit Hilfe der Funktion `combinations_with_replacement` aus der Python-Library `itertools` erzeugt [41].

#### Beispiele möglicher Allelkombinationen:

$ploidy = 2, n_{cand} = 2, n_{alleles} = 2$ :

$[(0, 0), (0, 1), (1, 1)]$

$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$ :

$[(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 0, 2), (0, 0, 1, 1), (0, 0, 1, 2), (0, 0, 2, 2), (0, 1, 1, 1), (0, 1, 1, 2), (0, 1, 2, 2), (0, 2, 2, 2), (1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 2, 2), (1, 2, 2, 2), (2, 2, 2, 2)]$

Für jedes Allel  $a_i$  können im Anschluss aus seinen absoluten Häufigkeiten  $H_{a_i}$  innerhalb jeder Allelkombination die relativen Häufigkeiten  $h_{a_i}$  nach (3-2) bestimmt werden.

$$h_{a_i} = \frac{H_{a_i}}{n_{alleles}} \quad (3-2)$$

#### Beispiele möglicher Allele-Fractions:

$ploidy = 2, n_{cand} = 2, n_{alleles} = 2$ :

[1.0, 0.0], [0.5, 0.5], [0.0, 1.0]

$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$ :

[1.0, 0.0, 0.0], [0.75, 0.25, 0.0], [0.75, 0.0, 0.25], [0.5, 0.5, 0.0], [0.5, 0.25, 0.25], [0.5, 0.0, 0.5], [0.25, 0.75, 0.0], [0.25, 0.5, 0.25], [0.25, 0.25, 0.5], [0.25, 0.0, 0.75], [0.0, 1.0, 0.0], [0.0, 0.75, 0.25], [0.0, 0.5, 0.5], [0.0, 0.25, 0.75], [0.0, 0.0, 1.0]

### 3.4.2 Berechnung der Likelihoods der Allele anhand der Allele-Fractions

In diesem Kapitel soll diejenige Allele-Fraction bestimmt werden, welche die beobachteten Reads am besten erklärt (vgl. Kap. 2.3.3). Für jeden Read wird dafür zunächst die Wahrscheinlichkeit bestimmt, dass der Read durch Sequenzierfehler aus einem der Kandidatenallele entstanden ist. Hiernach wird für jeden Read die Wahrscheinlichkeit errechnet, diesen anhand einer gegebenen Allele-Fraction zu beobachten. Diese Berechnungen werden für alle Allele-Fractions durchgeführt. Aus der so gewonnenen Lösungsmenge wird die Allele-Fraction mit der höchsten Likelihood für die spätere Zuordnung der Loci (Kap. 3.5) ausgewählt.

Zunächst wird in der Funktion `get_allele_likelihood_read()` eine Kante zwischen dem betrachteten Read und dem Kandidatenallel gesucht. Während der Ausführung werden die Sequenzen und Likelihoods der bereits gefundenen Read-Kandidatenallel-Paare in einem Dictionary abgelegt. Als Key wird dabei das Tupel aus der Read-ID und der Allelsequenz verwendet, die berechnete Likelihood wird als Value eingetragen. Da später für die Likelihoodberechnungen der verschiedenen Allele-Fractions mehrfach die Werte der Read-Allel-Likelihoods verwendet werden, ermöglicht die Verwendung eines Dictionary eine schnellere und effizientere Suche der passenden Kante ohne wiederholte Likelihoodberechnung.

Existiert für ein Read-Kandidatenallel-Paar noch kein Eintrag im Dictionary, so werden zuerst alle ausgehenden Nachbarn des Reads hinsichtlich ihrer in den Knoteneigenschaften

abgelegten Sequenz geprüft. Besitzt einer der ausgehenden Nachbarknoten die Sequenz des Kandidatenallels, so wird die Likelihood durch `get_alignment_likelihood()` berechnet (Algorithmus 3.2), ein entsprechender Eintrag im Dictionary angelegt und die Likelihood zurückgegeben.

Gibt es keine solche Kante, die vom Read zu einem Knoten mit Kandidatenallelsequenz führt, so wird geprüft, ob es eine Kante in umgekehrter Richtung gibt. Hierfür werden nun alle eingehenden Nachbarn des Reads betrachtet. Findet sich unter ihnen ein Knoten mit der gesuchten Kandidatenallelsequenz, so wird mit `get_alignment_likelihood()` die Likelihood berechnet, wobei die Verwendung der Gegenrichtung durch das Argument `reverse=True` markiert wird. Existiert auch keine entgegengesetzt gerichteten Kante, so wird der Wert 0 zurückgegeben.

Zur Veranschaulichung ist der beschriebene Algorithmus zudem in Pseudocode dargestellt (Algorithmus 3.1). Dabei sind  $C_k \in \{C_1, \dots, C_p\}$  der Subgraph einer Zusammenhangskomponente,  $r_i \in C_k$  der Knoten des Reads mit der ID  $idx_{r_i}$  und  $a_j \in A_{cand}$  ein Kandidatenallel mit der Sequenz  $s_{a_j}$ , wobei  $k, i, j \in \mathbb{N}$ . Das oben beschriebene Dictionary *dict* besitzt als Keys die Tupel der Sequenzen  $(idx_{r_i}, s_{a_i})$  und als Values die Likelihoods  $L_{r_i, a_i}$ , so dass gilt  $((idx_{r_i}, s_{a_i}), L_{r_i, a_i}) \in dict$ . Die Verwendung fester Knoten- oder Kanteneigenschaften ist durch eckige Klammern gekennzeichnet, so ruft beispielsweise  $r_i[q\_values]$  die  $q$ -Werte des Phred Quality Scores von  $r_i$  auf.

**Algorithmus 3.1** Suche eines Alignments zwischen Read und Kandidatenallel

---

```

1: function GET_ALLELE_LIKELIHOOD_READ( $C_k, r_i, s_{a_j}, dict$ )
2:    $idx_{r_i} \leftarrow r_i[name]$ 
3:    $qual \leftarrow r_i[q\_values]$ 
4:    $\epsilon \leftarrow C_k[\epsilon_{ins}] \cup C_k[\epsilon_{del}]$ 
5:   if  $\exists (idx_{r_i}, s_{a_i}) : ((idx_{r_i}, s_{a_i}), L_{r_i, a_i}) \in dict$  then
6:     return  $L_{r_i, a_i}$ 
7:   end if
8:    $out\_neighbors \leftarrow get\_out\_neighbors(r_i)$ 
9:   for  $out\_neighbor \in out\_neighbors$  do
10:    if  $s_{a_j} = out\_neighbor[sequence]$  then
11:       $cig \leftarrow edge(r_i, out\_neighbor)[cigar\_tuples]$ 
12:       $rev \leftarrow False$ 
13:       $L_{r_i, a_i} \leftarrow get\_alignment\_likelihood(\epsilon, cig, qual, rev)$   $\triangleright Alg. 3.2$ 
14:       $dict \leftarrow dict \cup ((r_i, out\_neighbor[sequence]), L_{r_i, a_i})$ 
15:      return  $L_{r_i, a_i}$ 
16:    end if
17:  end for
18:   $in\_neighbors \leftarrow get\_in\_neighbors(r_i)$ 
19:  for  $in\_neighbor \in in\_neighbors$  do
20:    if  $s_{a_j} = in\_neighbor[sequence]$  then
21:       $cig \leftarrow edge(in\_neighbor, r_i)[cigar\_tuples]$ 
22:       $rev \leftarrow True$ 
23:       $L_{r_i, a_i} \leftarrow get\_alignment\_likelihood(\epsilon, cig, qual, rev)$   $\triangleright Alg. 3.2$ 
24:       $dict \leftarrow dict \cup ((r_i, in\_neighbor[sequence]), L_{r_i, a_i})$ 
25:      return  $L_{r_i, a_i}$ 
26:    end if
27:  end for
28:  return 0
29: end function

```

---

Die eigentliche Likelihoodberechnung erfolgt in der Methode `get_alignment_likelihood()` (Algorithmus 3.2) nach Formel (2-15) des Modells. Wie in Kap. 2.3.2 beschrieben wird hierfür die Approximation des PairHMM aus den Sequenzalignments von Minimap2 verwendet.

Von der in `get_allele_likelihood_read()` ermittelten Kante zwischen Read und Kandidatenallel werden der Funktion die CIGAR-Tupel aus den Kanteneigenschaften, der Vektor mit dem Phred Quality Scores  $Q = (q_1, \dots, q_m)$  aus den Knoteneigenschaften des

Reads (siehe Kap. 3.3.1) sowie die Kantenrichtung als boolescher Wert übergeben. Aus den Phred Quality Scores des Reads wird zunächst die geschätzte Fehlerwahrscheinlichkeit  $P = (p_1, \dots, p_m)$  für jede Base nach Formel (3-3) errechnet [42].

$$p_i = 10^{\frac{-q_i}{10}} \quad (3-3)$$

Im Anschluss wird aus den p-Werten der Basenqualität für jede Base des Reads die Wahrscheinlichkeit errechnet, dass es sich im Falle eines Matches um die korrekte Base handelt (2-6) bzw. im Falle eines Mismatches, dass sich der Read durch Sequenzierfehler aus dem Allel erklären lässt (2-8). Für Indels werden die empirisch ermittelten Sequenzierfehlerraten von Illumina für Sequenzierplattformen [43] aus der Konfigurationsdatei unter dem Parameter **errors-per-base** entnommen. Entsprechend den in den CIGAR-Tupeln angegebenen Operationen wird für jede Base die Likelihood berechnet. Das Produkt der Likelihoods der einzelnen Basen ergibt schließlich die Likelihood zwischen Read und Allel (2-5).

Bei Bedarf kann die Methode `get_alignment_likelihood()` auch die Likelihood der entgegengesetzt gerichteten Kante bestimmen. Dabei wird die Querysequenz als Referenzsequenz betrachtet und umgekehrt. Dies ist über das boolesche Argument **reverse** steuerbar. Gilt **reverse = True**, so werden für die übergebenen CIGAR-Tupel Insertionen zu Deletionen und Deletionen zu Insertionen umgewandelt.

Zur Veranschaulichung ist die Methode in Algorithmus 3.2 zusätzlich als Pseudocode verfasst. Die Wahrscheinlichkeiten für Sequenzierfehler bei Insertionen und Deletionen aus den Grapheigenschaften werden dort mit  $\epsilon_{ins}$  und  $\epsilon_{del}$  bezeichnet, der Vektor der Phred Quality Scores der Querysequenz mit  $q_{query}$

**Algorithmus 3.2** Berechnung der Likelihood zwischen Read und Kandidatenallel

---

```

1: function GET_ALIGNMENT_LIKELIHOOD( $\epsilon_{ins}$ ,  $\epsilon_{del}$ , cigar_tuples, q_query, reverse)
2:   likelihood  $\leftarrow$  1.0, index  $\leftarrow$  0, p_query  $\leftarrow$  [ ]
3:   for  $q_i \in q_{query}$  do
4:     p_query  $\leftarrow$  p_query  $\cup$  ( $10^{\frac{-q_i}{10}}$ )
5:   end for
6:   if reverse then
7:     swap values of  $\epsilon_{ins}$  and  $\epsilon_{del}$ 
8:   end if
9:   for all (operation, length)  $\in$  cigar_tuples do
10:    while index  $<$  (index + length) do
11:      if operation  $\in$  match then
12:        likelihood  $\leftarrow$  likelihood  $\cdot$  ( $1 - p_{query}[index]$ )
13:      end if
14:      if operation  $\in$  substitution then
15:        likelihood  $\leftarrow$  likelihood  $\cdot \frac{1}{3} \cdot p_{query}[index]$ 
16:      end if
17:      if operation  $\in$  insertion then
18:        likelihood  $\leftarrow$  likelihood  $\cdot \epsilon_{ins}$ 
19:      end if
20:      if operation  $\in$  deletion then
21:        likelihood  $\leftarrow$  likelihood  $\cdot \epsilon_{del}$ 
22:      end if
23:      index  $\leftarrow$  index + 1
24:    end while
25:  end for
26:  return likelihood
27: end function

```

---

Durch `get_allele_likelihood_read()` und `get_alignment_likelihood()` wurden also die Wahrscheinlichkeiten jedes Reads bestimmt, dass er von den verschiedenen Allelen der Zusammenhangskomponente stammt (siehe oben). Die Wahrscheinlichkeiten im Bezug auf jedes Allel sollen nun in die zuvor bestimmten Allele-Fractions (siehe Kap. 3.4.1) nach Formel (2-17) einbezogen werden. In der Funktion `calc_vafs_likelihood_read()` aus dem Modul `likelihood_operations.py` wird nun für jeden Read die Wahrscheinlichkeit berechnet, diesen unter der gegebenen Allel-Fraction zu beobachten. Dazu wird über die relativen Häufigkeiten jedes Allels der Allele-Fraction iteriert und ihr Produkt mit der Readlikelihood aufsummiert.

Aus dem Produkt über alle Reads wird anschließend für die Allele-Fraction nach Formel (2-18) in `calc_vafs_likelihood()` die Gesamtwahrscheinlichkeit errechnet.

Auf diese Weise wird in `noderad_main.py` für jede Allele-Fraction die Gesamtwahrscheinlichkeit über alle Reads und alle relativen Häufigkeiten der Allele berechnet und in einer Liste abgelegt.

Aus dieser Liste wird die Allele-Fraction mit maximaler Likelihood als wahrscheinlichste Lösung gewählt und in Kap. 3.5 verwendet, um die wahrscheinlichsten Loci zu ermitteln, die diese Häufigkeitsverteilung der Allele erklären können.

Als zusätzliche Statistiken werden in den Log-Dateien für jede Komponente die Anzahl der Allele, die Ploidie sowie Allele-Fraction mit der maximaler Likelihood und den dazugehörigen Allelsequenzen festgehalten.

## 3.5 Bestimmung der maximalen Likelihood der Loci

In Abhängigkeit von der Anzahl der Kandidatenallele und der Ploidie können die Zusammenhangskomponenten auch mehr als nur einen Locus beinhalten. Daher soll die in Kap. 3.4 identifizierte Allelfraktion mit maximaler Likelihood nun passenden Loci-Kombinationen zugeordnet werden. Unter Einbeziehung der Heterozygotiewahrscheinlichkeiten soll daraus die wahrscheinlichste Loci-Zuordnung ermittelt werden.

### 3.5.1 Bestimmung der möglichen Loci

Für die Zuordnung der Loci müssen zunächst aus der Allele-Fraction mit maximaler Likelihood die verschiedenen Kombinationen möglicher Loci durch `get_candidate_loci()` erzeugt werden. Die Funktion ist bereits in einer effizienteren Form implementiert und die zugrundeliegenden Verbindungen zum Modell sind dadurch nicht direkt erkennbar. Aus diesem Grund sollen an dieser Stelle zunächst die Überlegungen aufgezeigt werden, auf denen die effizientere Implementierung basiert, im Anschluss werden die Änderungen zur Effizienzverbesserung aufgezeigt.

Analog zu `get_candidate_vafs()` (Kap. 3.4.1) sollen aus der beobachteten  $n_{cand}$  und der tatsächlich zu erwartenden Anzahl von Allelen  $n_{alleles}$  alle Allelkombinationen mit Wiederholung und ohne Berücksichtigung der Reihenfolge (vgl. Formel (3-1)) durch die Funktion `combinations_with_replacement()` aus der Python-Library `itertools` gebildet. Allerdings sollen nun die Allelkombinationen zu Tupeln entsprechend ihrer Ploidie zusammengefasst werden. Jedes Tupel repräsentiert einen möglichen Locus in der Kom-

ination. Die Allelkombinationen durch `combinations_with_replacement()` sind lexikographisch sortiert. Für jede der Kombinationen muss nun aber die Reihenfolge so variiert werden, dass die Allele, falls mehrere Loci möglich sind, in verschiedenen Kombinationen den Loci zugeordnet werden. Hierfür wird aus `itertools` die Funktion `permutations()` verwendet. Diese bildet aus  $n$  Elementen alle Kombinationen der Länge  $k$  mit Beachtung der Reihenfolge und ohne Wiederholungen.

Auch wenn sich die Allele in den Allelkombinationen durchaus wiederholen können, so werden sie doch in `permutations()` als einzigartiges Element behandelt unabhängig von jeweiligen Wert. Folglich entstehen durch die Bildung der Permutationen auch einige Duplikate. Die Anzahl der entstehenden Elemente wird für `permutations()` mit  $\frac{n!}{(n-k)!}$  angegeben [41]. Die Permutationen werden über jede Allelkombination in ihrer vollständigen Länge gebildet, so dass gilt  $n = k$ , folglich werden insgesamt  $n!$  Elemente erzeugt (3-4).

$$\frac{n!}{(n-k)!} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n! \quad (3-4)$$

Mit Hilfe der `itertools`-Funktion `grouper()` werden die Permutationen entsprechend der Ploidie in Tupel der Länge  $\phi$  unterteilt. Es resultieren Tupel von  $\phi$ -Tupeln, wobei die Anzahl der  $\phi$ -Tupel der Anzahl möglicher Loci entspricht. Mehrere Loci in einer Zusammenhangskomponenten sind nach `get_max_parsimony_n_alleles()` genau dann möglich, wenn gilt  $n_{cand} > \phi$ .

Nach der Gruppierung werden die übergeordneten Tupel sowie die darin enthaltenen  $\phi$ -Tupel jeweils sortiert, um anschließend durch die Python-Funktion `set()` die oben erwähnten Duplikate zu entfernen.



**Beispiele möglicher Loci-Kombinationen:**

$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$ :

**Kombinationen der Allele:**

[(0, 0, 0, 0), (0, 0, 0, 1), (0, 0, 0, 2), (0, 0, 1, 1), (0, 0, 1, 2), (0, 0, 2, 2), (0, 1, 1, 1), (0, 1, 1, 2), (0, 1, 2, 2), (0, 2, 2, 2), (1, 1, 1, 1), (1, 1, 1, 2), (1, 1, 2, 2), (1, 2, 2, 2), (2, 2, 2, 2)]

**Permutationen der Allele:**

(Zur anschaulicheren Darstellung wurden Duplikate entfernt, das geschieht eigentlich erst nach der Erzeugung und Sortierung der Tupel)

(1, 2, 1, 1), (2, 1, 0, 0), (2, 1, 1, 1), (0, 1, 2, 1), (0, 1, 1, 2), (0, 1, 0, 0), (2, 2, 1, 0), (0, 2, 2, 1), (2, 2, 0, 1), (1, 0, 2, 2), (0, 2, 0, 1), (2, 0, 0, 1), (1, 0, 1, 0), (0, 2, 1, 2), (0, 0, 2, 0), (2, 2, 2, 1), (1, 1, 0, 1), (2, 0, 1, 1), (2, 0, 2, 0), (0, 0, 2, 2), (1, 1, 2, 0), (1, 2, 1, 0), (2, 0, 2, 2), (2, 1, 1, 0), (2, 1, 0, 2), (1, 2, 0, 1), (0, 1, 2, 0), (1, 2, 1, 2), (1, 2, 2, 1), (0, 1, 1, 1), (1, 1, 1, 0), (0, 0, 0, 0), (2, 1, 1, 2), (2, 1, 2, 1), (1, 0, 0, 1), (0, 1, 0, 2), (2, 2, 1, 2), (0, 2, 2, 0), (1, 0, 2, 1), (2, 0, 0, 0), (0, 2, 1, 1), (1, 1, 1, 2), (0, 0, 0, 2), (0, 0, 1, 1), (1, 0, 1, 2), (2, 0, 0, 2), (0, 0, 2, 1), (1, 1, 2, 2), (2, 1, 0, 1), (1, 2, 0, 0), (0, 1, 2, 2), (1, 2, 2, 0), (0, 1, 1, 0), (2, 2, 0, 0), (0, 2, 0, 0), (2, 1, 2, 0), (1, 0, 0, 0), (1, 2, 0, 2), (0, 1, 0, 1), (2, 2, 1, 1), (2, 2, 2, 0), (1, 1, 0, 0), (1, 0, 2, 0), (0, 2, 2, 2), (1, 2, 2, 2), (2, 2, 0, 2), (0, 2, 1, 0), (0, 2, 0, 2), (2, 0, 1, 0), (0, 0, 0, 1), (1, 1, 1, 1), (0, 0, 1, 0), (2, 1, 2, 2), (1, 0, 0, 2), (1, 0, 1, 1), (2, 2, 2, 2), (1, 1, 0, 2), (2, 0, 1, 2), (2, 0, 2, 1), (0, 0, 1, 2), (1, 1, 2, 1)

**Mögliche Loci-Kombinationen:**

((0, 0), (0, 2)), ((1, 1), (1, 1)), ((0, 2), (0, 2)), ((1, 1), (2, 2)), ((0, 1), (0, 2)), ((1, 1), (1, 2)), ((1, 2), (2, 2)), ((1, 2), (1, 2)), ((0, 0), (1, 1)), ((0, 0), (2, 2)), ((0, 2), (1, 1)), ((0, 1), (1, 1)), ((0, 0), (0, 0)), ((0, 2), (2, 2)), ((0, 0), (1, 2)), ((0, 1), (2, 2)), ((2, 2), (2, 2)), ((0, 0), (0, 1)), ((0, 2), (1, 2)), ((0, 1), (1, 2)), ((0, 1), (0, 1))

Für jede Loci-Kombination wird später in Kap. 3.5.2 durch die Indikatorfunktion nach Formel (2-19) geprüft, ob sie die Allele-Fraction mit maximaler Likelihood  $\Theta_{max\_lh}$  überhaupt repräsentieren kann. Das heißt die Häufigkeit jedes Allels über alle Loci einer gegebenen Loci-Kombination muss unter Berücksichtigung der Ploidie und der zu erwartenden Anzahl der Allele  $n_{alleles}$  die relativen Häufigkeiten der ermittelten Allele-Fraction mit maximaler Likelihood erklären können.

Die hier zunächst vorgestellte Bildung von Permutationen über alle Allelkombinationen führt bei großen Clustern zu einer enormen Anzahl von Kombinationen. Da diese später ohnehin durch die Indikatorfunktion wieder reduziert werden, ermöglicht die direkte Erzeugung passender Loci-Kombinationen eine weitaus effizientere Implementierung. Im Hinblick auf die relativen Häufigkeiten in der Allel-Fraction mit maximaler Likelihood gibt es also eine feste Anzahl für jedes Allel, wie oft es über die Loci hinweg in einer Loci-Kombination auftreten darf. Diese Anzahl wird für jedes Allel durch die Indikatorfunktion getestet.

Da `combinations_with_replacement()` eine sortierte Ausgabe von Kombinationen ohne Beachtung der Reihenfolge erzeugt, erfüllt nur eine einzige der resultierenden Allelkombinationen tatsächlich die Bedingung der Indikatorfunktion (Satz 1) und es genügt, nur über dieser Kombination die Permutationen zu bilden. Es genügt also, aus der Allel-Fraction mit maximaler Likelihood, die absoluten Häufigkeiten der beteiligten Allele aus ihren relativen Häufigkeiten durch Umformung von Formel (3-2) zu bestimmen. Ein sortiertes Tupel der betreffenden Allele in der jeweils ermittelten Häufigkeit entspricht dann genau der einzigen gültigen Lösung für die Indikatorfunktion, die sich aus den Kombinationen bei Ausführung von `combinations_with_replacement()` ergeben hätte.

Durch die Umformung der Allel-Fraction mit maximaler Likelihood zu einer Allelkombination, müssen nun nur noch die Permutationen für diese Allelkombination bestimmt werden. Das unten stehenden Beispiel illustriert die Änderungen bei der effizienteren Implementierung. Nach anschließender Sortierung und Gruppierung zu Loci können die Duplikate entfernt werden.

**Beispiele möglicher Loci-Kombinationen bei effizienter Implementierung:**

$ploidy = 2, n_{cand} = 3, n_{alleles} = 4$ :

**Errechnete Allel-Fraction mit maximaler Likelihood:**

[0.25, 0.5, 0.25]

**Daraus resultierende sortierte Allelkombination:**

(0, 1, 1, 2)

**Permutationen der resultierenden Allelkombination:**

(Duplikate wurden auch hier bereits entfernt)

(1, 0, 1, 2), (2, 1, 1, 0), (1, 1, 2, 0), (0, 1, 2, 1), (1, 0, 2, 1), (1, 2, 0, 1), (2, 1, 0, 1), (2, 0, 1, 1), (1, 1, 0, 2), (1, 2, 1, 0), (0, 2, 1, 1), (0, 1, 1, 2)

**Mögliche Loci-Kombinationen:**

((0, 1), (1, 2)), ((0, 2), (1, 1))

**Indikatorfunktion:**

$$z_l = \prod_{i=1}^3 1_{\sum_{j=1}^g \sum_{k=1}^{\phi} 1_{l_{j,k}=i} = \theta_i \cdot g \cdot \phi}$$

$$z_{((0,1),(1,2))} = \prod_{i=1}^n 1_{\sum_{j=1}^2 \sum_{k=1}^2 1_{l_{j,k}=i} = \theta_i \cdot 2 \cdot 2} = 1_{(1=0.25 \cdot 4)} \cdot 1_{(2=0.5 \cdot 4)} \cdot 1_{(1=0.25 \cdot 4)} = 1$$

$$z_{((0,2),(1,1))} = \prod_{i=1}^n 1_{\sum_{j=1}^2 \sum_{k=1}^2 1_{l_{j,k}=i} = \theta_i \cdot 2 \cdot 2} = 1_{(1=0.25 \cdot 4)} \cdot 1_{(2=0.5 \cdot 4)} \cdot 1_{(1=0.25 \cdot 4)} = 1$$

Das Ergebnis der Indikatorfunktion hängt allein von der absoluten Häufigkeiten der Allele ab und wird durch die Reihenfolge ihres Auftretens in einer Allelkombination nicht beeinflusst. Alle resultierenden Loci-Kombinationen der Funktion `get_candidate_loci()` erfüllen daher auch die Indikatorfunktion, da sie Permutationen einer Allelkombination sind, welche die Indikatorfunktion erfüllt. Daher müsste mit diesen Änderungen die Indikatorfunktion eigentlich nicht mehr geprüft werden. Sie wurde dennoch in der Implementierung belassen, um das Modell besser zeigen zu können und wird an entsprechender Stelle noch einmal erwähnt.

**Satz (1).** *Unter allen Allelkombinationen mit Wiederholung und ohne Beachtung der Reihenfolge gibt es nur eine Kombination, welche die Indikatorfunktion (2-19) erfüllt.*

*Beweis.* Für Kombinationen der Länge  $p$  mit Wiederholung aber ohne Beachtung der Reihenfolge gilt, dass sich jedes Element bis zu  $p$  Mal wiederholen kann. Da die Reihenfolge nicht beachtet werden soll, gilt außerdem dass jede Häufigkeitsverteilung der Elemente innerhalb einer Kombination in der Menge aller erzeugten Kombinationen nur genau einmal vorkommen darf. Andernfalls wäre dies eine Beachtung der Reihenfolge. Daraus kann abgeleitet werden:

1. Die absoluten Häufigkeiten einer Kombination sind in der Menge der erzeugten Kombinationen zu einer gegebenen Anzahl von Allelen und einer gegebenen Ploidie einzigartig.
2. Nach Formel 3-2 sind somit auch die relativen Häufigkeiten einer Allel-Fraction in der Menge der erzeugten Allel-Fractions zu einer gegebenen Anzahl von Allelen und einer gegebenen Ploidie einzigartig.

Es bleibt also zu beweisen, dass die Bedingung der Indikatorfunktion für genau eine Allel-Fraction gilt:

$$\begin{aligned}
z_l &= \prod_{i=1}^n 1_{\sum_{j=1}^g \sum_{k=1}^{\phi} 1_{l_{j,k}=i} = \theta_i \cdot g \cdot \phi} \wedge \theta_i \in \Theta_{max\_lh} \\
&\Rightarrow z_l = 1 \Leftrightarrow \forall a_i \in A_{cand} : H_{a_i} = \theta_{a_i} \cdot g \cdot \phi \\
&\Rightarrow z_l = 1 \Leftrightarrow \forall a_i \in A_{cand} : H_{a_i} \stackrel{3-2}{=} \left( \frac{H_{a_i}}{n_{alleles}} \right)_{max\_lh} \cdot g \cdot \phi \\
&\Rightarrow z_l = 1 \Leftrightarrow \forall a_i \in A_{cand} : H_{a_i} = \left( \frac{H_{a_i}}{n_{alleles}} \right)_{max\_lh} \cdot n_{alleles} \\
&\Rightarrow z_l = 1 \Leftrightarrow \forall a_i \in A_{cand} : H_{a_i} = (H_{a_i})_{max\_lh}
\end{aligned}$$

Für die Indikatorfunktion gilt also, dass ihre Bedingung nur dann erfüllt wird, wenn die absoluten Häufigkeiten der Allele einer Allelkombination den relativen Häufigkeiten der Allel-Fraction mit maximaler Likelihood  $\Theta_{max\_lh}$  entsprechen. Da absolute und relative Häufigkeitsverteilungen jeweils nur genau einmal bei den Kombinationen mit Wiederholung und ohne Beachtung der Reihenfolge auftreten, gibt es auch nur eine Allelkombination, welche über die Indikatorfunktion die Häufigkeitsverteilung der Allel-Fraction mit maximaler Likelihood erfüllen kann.  $\square$

### 3.5.2 Bestimmung der wahrscheinlichsten Locuszuordnung der Allele-Fraction mit maximaler Likelihood

In diesem Schritt sollen schließlich für die Allel-Fraction mit maximaler Likelihood (Kap. 3.4.2) die ermittelten Kandidatenloci (Kap. 3.5.1) unter Berücksichtigung der Heterozygotie analysiert werden und die wahrscheinlichste Kombination der Loci ermittelt werden.

Für jede Loci-Kombination aus `get_candidate_loci()` wird aus der Funktion `calc_loci_likelihoods()` heraus zunächst geprüft, ob sich die Loci dieser Kombination der zuvor bestimmten Allelfraction mit maximaler Likelihood zuordnen lassen. Dies geschieht mit Hilfe der Indikatorfunktion `indicator_constraint()` nach Formel (2-19). Ist die Bedingung der Indikatorfunktion erfüllt, so wird die Likelihood der Loci-Kombination aus den Heterozygotiewahrscheinlichkeiten der darin enthaltenen Allele berechnet und zurückgegeben. Ist die Bedingung der Indikatorfunktion nicht erfüllt, so wird eine Likelihood von 0 zurückgegeben. Wie bereits besprochen, ist durch die Änderungen an der Implementierung von `get_candidate_loci()` dieser Schritt nicht mehr notwendig und wurde nur belassen, um das dahinterstehende Modell strukturierter zeigen zu können.

Für die Likelihoodberechnung müssen nun einer Loci-Kombination zunächst die passenden Allelsequenzen zugeordnet werden. Die einzelnen Allele innerhalb einer Loci-Kombination werden durch Integerwerte repräsentiert, die dem Index des dazugehörigen Kandidatenalleles entsprechen (siehe Kap. 3.4.1). Zu jeder Loci-Kombination wird also eine Liste  $S_{loc, alleles}$  generiert, welche die Kandidatenallelsequenzen der einzelnen Loci als Sublisten beinhaltet.

Die Berechnung der Likelihood ist im vorliegenden Modell zunächst nur für diploide Organismen möglich. Analog zur Likelihoodberechnung beim paarweisen Vergleich der Reads unter Berücksichtigung der Sequenzierfehlerwahrscheinlichkeit (Kap. 3.3.2) erfolgt auch die Likelihoodberechnung für einer Loci-Kombination durch den paarweisen Vergleich der Allele im Sinne eines pair Hidden Markov Models  $pairHMM_{\eta}(a_{l_{j,1}}, a_{l_{j,2}})$  (vgl. Kap. 2.3.2). Allerdings werden nun die Heterozygotiewahrscheinlichkeiten  $\eta_{sub}$ ,  $\eta_{ins}$  und  $\eta_{del}$  der Grapheigenschaften zur Ermittlung der Likelihood nach Formel (2-20) genutzt.

Hierzu werden zunächst in `get_allele_likelihood_alleles()` (Algorithmus 3.3) auf jeweils einer der Listen  $S_{loc, alleles}$  für jeden Locus  $loc$  die darin enthaltenen Allelsequenzen  $s_{a_i}, s_{a_j}$  der Allele  $a_i, a_j$  paarweise mit einander kombiniert. Dies wird durch die Funktion `combinations()` aus `itertools` realisiert, die Kombinationen ohne Wiederholung und ohne Beachtung der Reihenfolge erzeugt. Die paarweise Kombination der Allele ist allerdings nur für Daten diploider Organismen möglich, bei höherer Ploidie wird ein multiples Se-

quenzalignment benötigt (vgl. Kap 6). Für jedes Allelpaar werden dann die CIGAR-Tupel durch die Funktion `get_cigar_tuples()` (Algorithmus 3.5) ermittelt und für die Likelihoodberechnung in `get_heterozygosity()` verwendet (Algorithmus 3.4). Die Gesamtwahrscheinlichkeit errechnet sich aus dem Produkt der Likelihood aller Allelpaare nach 2-21. Analog zur Likelihoodberechnung der Allele (vgl. Algorithmus 3.1) werden auch die ermittelten Locilikelihoods für die verschiedenen Loci-Kombinationen mehrfach benötigt. Daher wird die Effizienz auch hier durch Integration eines Dictionary mit den bereits ermittelten Likelihoods deutlich erhöht. Die Einträge des Dictionary sollen dabei die Tupel aus den beiden Allelsequenzen als Keys und die bereits errechneten Likelihoods als Values verwenden. Für alle Allelsequenzpaare wird noch vor der Suche nach geeigneten CIGAR-Tupeln das Dictionary  $dict_{loc}$  auf bereits vorhandene Einträge überprüft. Falls es einen Eintrag gibt, so wird auf eine erneute Likelihoodberechnung verzichtet und stattdessen der Wert des Dictionary verwendet. Falls kein Eintrag existiert, so wird im Anschluss an die Likelihoodberechnung ein entsprechender Eintrag dem Dictionary hinzugefügt.

---

**Algorithmus 3.3** Bestimmung der Likelihood der Allele innerhalb einer Loci-Kombination

---

```

1: function GET_ALLELE_LIKELIHOOD_ALLELES( $C_k, S_{loc,alleles}, dict_{loc}$ )
2:    $likelihood \leftarrow 1.0$ 
3:   for  $loc \in S_{loc,alleles}$  do
4:      $pairs \leftarrow combinations(loc)$   $\triangleright itertools.combinations()$ 
5:     for  $(s_{a_i}, s_{a_j}) \in pairs$  do
6:       if  $\exists (s_{a_i}, s_{a_j}) : ((s_{a_i}, s_{a_j}), L_{a_i,a_j}) \in dict_{loc}$  then
7:          $likelihood \leftarrow likelihood + L_{a_i,a_j}$ 
8:       else
9:          $cigar \leftarrow get\_cigar\_tuples(C_k, s_{a_i}, s_{a_j})$   $\triangleright Alg. 3.5$ 
10:        if  $cigar$  exists then
11:           $cig \leftarrow cigar[0]$ 
12:           $rev \leftarrow cigar[1]$ 
13:           $\eta_{rates} \leftarrow C_k[\eta_{sub}] \cup C_k[\eta_{ins}] \cup C_k[\eta_{del}]$ 
14:           $L_{a_i,a_j} \leftarrow \log(get\_heterozygosity(\eta_{rates}, cig, rev))$   $\triangleright Alg. 3.4$ 
15:           $dict_{loc} \leftarrow dict_{loc} \cup ((s_{a_i}, s_{a_j}), L_{a_i,a_j})$ 
16:           $likelihood \leftarrow likelihood + L_{a_i,a_j}$ 
17:        end if
18:      end if
19:    end for
20:  end for
21:  return  $e^{likelihood}$ 
22: end function

```

---

In `get_heterozygosity()` erfolgt die Likelihoodberechnung der paarweisen Kombinationen der Allelsequenzen eines Locus aus den CIGAR-Tupeln und den Heterozygotiewahrscheinlichkeiten. Dies beschränkt den Prototypen derzeit auf diploide Organismen (siehe auch Kap. 6). Wie in Algorithmus 3.2 wird auch hier die Likelihood über aller Matches und Mismatches aus den Informationen der CIGAR-Tupel bestimmt. Die Likelihoodberechnung der Matches erfolgt dabei nach Formel 2-11, die der Mismatches nach Formel 2-10. Das Argument `reverse` markiert zudem die Richtung der Kante, von der die CIGAR-Tupel stammen. Für Kanten in Rückrichtung, die von der Referen- zur Query-Sequenz verlaufen gilt `reverse=True`. In diesem Fall werden Deletionen als Insertionen behandelt und umgekehrt.

---

**Algorithmus 3.4** Bestimmung der Likelihood zwischen zwei Kandidatenallelen hinsichtlich der Heterozygotiewahrscheinlichkeiten

---

```

1: function GET_HETEROZYGOSITY( $\eta_{sub}$ ,  $\eta_{ins}$ ,  $\eta_{del}$ , cigar_tuples, reverse)
2:   likelihood  $\leftarrow$  1.0
3:   if reverse then
4:     swap values of  $\eta_{ins}$  and  $\eta_{del}$ 
5:   end if
6:   for all (operation, length)  $\in$  cigar_tuples do
7:     if operation  $\in$  match then
8:       likelihood  $\leftarrow$  likelihood  $\cdot$   $(1 - (\eta_{sub} + \eta_{ins} + \eta_{del}))^{length}$ 
9:     end if
10:    if operation  $\in$  substitution then
11:      likelihood  $\leftarrow$  likelihood  $\cdot$   $(\eta_{sub})^{length}$ 
12:    end if
13:    if operation  $\in$  insertion then
14:      likelihood  $\leftarrow$  likelihood  $\cdot$   $(\eta_{ins})^{length}$ 
15:    end if
16:    if operation  $\in$  deletion then
17:      likelihood  $\leftarrow$  likelihood  $\cdot$   $(\eta_{del})^{length}$ 
18:    end if
19:  end for
20:  return likelihood
21: end function

```

---

Wie bereits erwähnt, werden für den paarweisen Vergleich zweier Kandidatenallele im Rahmen der Likelihoodberechnung die dazugehörigen CIGAR-Tupel benötigt. Diese werden durch die Funktion `get_cigar_tuples()` ermittelt. Als Eingabeparameter erhält die Funktion den Subgraphen der Zusammenhangskomponente sowie die beiden Kandidaten-

allelsequenzen: die Query-Sequenz  $s_{source}$  und die Referenz-Sequenz  $s_{target}$ . Der Subgraph wird nach einer Kante durchsucht, die Knoten miteinander verbindet, welche die beiden gegebenen Sequenzen besitzen. Dabei werden nicht nur Kanten berücksichtigt, die von der Query- zur Referenzsequenz verlaufen, sondern auch Kanten in Gegenrichtung. Zur Markierung der Kantenrichtung gibt die Funktion neben den CIGAR-Tupeln auch einen booleschen Wert zurück. Dieser wird in der Funktion `get_heterozygosity()` (Kap. 3.5.2) für das Argument `reverse` verwendet.

Um eine passende Kante im Subgraphen zu finden, erstellt `get_cigar_tuples()` zunächst eine Liste aller Knoten  $R_{source} = (v_1, \dots, v_i)$ , welche die Query-Sequenz  $s_{query}$  besitzen sowie eine Liste aller Knoten  $R_{target} = (w_1, \dots, w_j)$ , die die Referenz-Sequenz  $s_{ref}$  tragen. Hierfür werden für beide Sequenzen die Knoten des Subgraphen mit `find_vertex()` aus graph-tool durchsucht. Anschließend wird nach einer Kante gesucht, die einen der Knoten aus  $R_{source}$  mit einem der Knoten aus  $R_{target}$  verbindet. Es erfolgt also ein Kantenaufsuch für jeden Knoten aus  $R_{source}$  in Kombination mit jedem Knoten aus  $R_{target}$ . Der Kantenaufsuch wird sowohl für die angegebene Kantenrichtung und falls erfolglos auch für die entgegengesetzte Richtung durchgeführt. Existiert eine solche Kante, dann werden die CIGAR-Tupel sowie der boolesche Wert entsprechend der Kantenrichtung zurückgegeben, andernfalls wird *None* zurückgegeben.

Zur Veranschaulichung ist der Algorithmus in 3.5 zudem als Pseudocode dargestellt. Dabei sei  $C_k \in \{C_1, \dots, C_p\}$  der Subgraph einer Zusammenhangskomponente mit seiner Kantenmenge  $E_k$  und  $s_{query}$  bzw.  $s_{ref}$  seien die Query- bzw. die Referenz-Sequenz. Die Verwendung fester Knoten- oder Kanteneigenschaften ist durch eckige Klammern gekennzeichnet, so ruft beispielsweise  $v_i[sequence]$  die Sequenz des Knotens  $v_i$  aus den Knoteneigenschaften ab.



**Algorithmus 3.5** CIGAR-Tupel bestimmen

---

```

1: function GET_CIGAR_TUPLES( $C_k, s_{query}, s_{ref}$ )
2:    $R_{source} \leftarrow \{v_i \in C_k \wedge i, k \in \mathbb{N} \mid v_i[sequence] = s_{query}\}$ 
3:    $R_{target} \leftarrow \{w_j \in C_k \wedge j, k \in \mathbb{N} \mid w_j[sequence] = s_{ref}\}$ 
4:   for  $v_i \in R_{source}$  do
5:     for  $w_j \in R_{target}$  do
6:       if  $(v_i, w_j) \in E_k$  then
7:         return (  $E_k[(v_i, w_j)][cigar\_tuples], False$  )
8:       end if
9:       if  $(w_j, v_i) \in E_k$  then
10:        return (  $E_k[(w_j, v_i)][cigar\_tuples], True$  )
11:       end if
12:     end for
13:   end for
14:   return None
15: end function

```

---

In `noderad_main.py` werden die beschriebenen Schritte zur Likelihoodberechnung für alle Permutationen ausgeführt und anschließend die Loci-Kombination mit maximaler Likelihood als wahrscheinlichste Loci-Verteilung ausgewählt.

In den Log-Dateien werden zusätzlich einige Statistiken zur Berechnung der wahrscheinlichsten Loci-Verteilung eingetragen. Dazu gehören die ermittelten Likelihoods der Loci-Kombination sowie die Loci mit maximaler Likelihood.

## 3.6 Ausgabe der wahrscheinlichsten Loci als VCF-Datei

Die Loci-Kombinationen mit maximaler Likelihood  $Loc_{max}$  aller Zusammenhangskomponenten werden für jedes Individuum jeweils in eine Datei im Variant Call (siehe Kap. 1.4.3) Format geschrieben. Dabei werden zusätzlich die optionalen FORMAT- und Proben-Spalten zur Kennzeichnung des Genotyps verwendet.

In die Spalten REF und ALT werden die Sequenzen der Allele aus  $Loc_{max}$  lexikographisch sortiert eingetragen. Hierfür wird in der Funktion `get_sorted_loci_alleles()` eine sortierte Liste der Allelsequenzen aus  $Loc_{max}$  gebildet. Durch die Pythonfunktion `set()` wird über dieser Liste die Menge der Allele extrahiert, so dass Duplikate herausgefiltert werden. Aus der so entstandenen Liste der Allelsequenzen  $A_{res}$  wird der erste Eintrag in

die Spalte REF eingetragen, die übrigen Einträge werden in die Spalte ALT geschrieben.

Da jeweils die vollständige Sequenz der Allele für den Eintrag in die VCF-Datei verwendet wird, wird für jeden Locus in der Spalte POS der Wert 1 eingetragen.

In der FORMAT-Spalte wird durch den Eintrag "GT" festgelegt, dass in der Probenspalte der Genotyp spezifiziert wird. Da die Indizes der Loci der Liste der Kandidatenallele zugeordnet sind, aber nicht alle Allele dieser Liste auch in  $Loc_{max}$  vorkommen, muss ihre Indizierung nun auf die bereits erstellte Liste mit den zu  $Loc_{max}$  gehörigen Allelsequenzen  $A_{res}$  angepasst werden.

Hierfür werden in der Funktion `get_alleles_matched_to_loci()` zunächst jedem Locus aus  $Loc_{max}$  die entsprechenden Sequenzen aus der Liste der Kandidatenallele zugeordnet. Anschließend werden diese Sequenzen in `get_gt_indices()` den Indizes der passenden Sequenz aus  $A_{res}$  zugeordnet. Dadurch wird das lexikographisch erste Allel, also REF mit 0 im Genotyp indiziert, die Allele aus ALT erhalten höhere Indizes entsprechend ihrer Sortierung in  $A_{res}$ . Nun lässt sich der Genotyp im Bezug auf die Sequenzen in REF und ALT aus diesen Indizes und getrennt durch Slashes direkt angeben. Dies geschieht in der Funktion `get_genotype()`, deren Ergebnis dann in die Probenspalte eingetragen wird.

# Kapitel 4

## Laufzeitanalyse

### 4.1 Graphkonstruktion und Bestimmung der Zusammenhangskomponenten

#### Laufzeit für das Hinzufügen der Knoten des Graphen

Die Laufzeit für das Hinzufügen eines Knotens beträgt nach der Dokumentation von graph-tool  $O(|V|)$ , da es sich hierbei um eine Einfügeoperation in die bereits bestehende Knotenmenge handelt und ein neuer Iterator über alle Knoten erzeugt und zurückgegeben wird [38]. Die Zuweisung der Knoteneigenschaften erfolgt in  $O(1)$ . Über alle Reads, also über die resultierende Anzahl der Knoten  $|V|$  ergibt sich daraus eine Gesamtlaufzeit von  $O(|V|^2)$ .

#### Laufzeit für das Hinzufügen der Kanten des Graphen

Hinsichtlich der Laufzeit benötigt das Hinzufügen einer Kante nach Angaben der graph-tool Dokumentation [38] eine Laufzeit von  $O(1)$ . Da aber die Query- und die Referenzreads den bereits zuvor angelegten Knoten zugeordnet werden müssen, erfordert dies eine Suche der betreffenden Knoten. Dabei durchsucht graph-tool mit seiner Funktion `find_vertex()` allein die Knoten und prüft auf die gesuchte Read-ID aus den FASTQ-Daten. Die ein- und ausgehenden Kanten der Knoten werden nicht beachtet, so dass eine Tiefen- oder Breitensuche des Graphen nicht notwendig ist und die Suche in  $O(|V|)$  durchgeführt werden kann [44]. Die Zuweisung der Kanteneigenschaften erfolgt jeweils in  $O(1)$ , da diese direkt bei der Erzeugung der Kante hinzugefügt werden und keine vorherige Suche der Kante erforderlich ist. Für die Berechnung der Likelihood wird die geschätzte Fehlerrate  $p_{query}$  jeder Base verwendet, so dass die Anzahl der Berechnungen für jede Kante der Länge der Readsequenz  $k$  entspricht. Die Laufzeit für das Hinzufügen einer Kante beträgt somit  $O(k)$ . Für alle Kanten ergibt sich daraus eine Gesamtlaufzeit von  $O(|E| \cdot (k + |V|))$ . Bei realen Datensätzen gilt in der Regel  $k \ll |V|$  und die Länge der Reads variiert nur in einem

engen Bereich, so dass  $k$  als vernachlässigbar klein und als nahezu konstant betrachtet werden kann. Dann ergibt sich aus Formel (4-1) eine Laufzeit von  $O(|E| \cdot |V|)$ .

$$O(|E| \cdot (k + |V|)) = O(|E| \cdot |V|) \quad (4-1)$$

Sind bei kleinen Datensätzen nur wenige Reads vorhanden, so dass  $k \leq |V|$ , dann ergibt sich unter zusätzlicher Berücksichtigung von  $k$  im Worst Case mit  $k = |V|$  nach Formel (4-2) ebenfalls eine Laufzeit von  $O(|E| \cdot |V|)$ .

$$O(|E| \cdot (k + |V|)) = O(|E| \cdot (|V| + |V|)) = O(|E| \cdot 2 \cdot |V|) = O(|E| \cdot |V|) \quad (4-2)$$

Unter der Annahme, dass in seltenen Fällen die Readlänge, die meist nur wenige hundert Basenpaare zählt, tatsächlich die Anzahl der Reads übersteigt und somit  $k > |V|$  gilt, dann dominiert  $k$  die Laufzeit. In diesem Fall wäre im Worst Case  $k$  eine obere Schranke für  $|V|$ , so dass gilt:

$$O(|E| \cdot (k + |V|)) = O(|E| \cdot (k + k)) = O(|E| \cdot 2 \cdot k) = O(|E| \cdot k) \quad (4-3)$$

Da die Länge der Reads durch die gewählten Restriktionsenzyme sowie durch das Sequenzierverfahren selbst beschränkt ist, müsste ein solcher Datensatz relativ klein sein, so dass die damit verbundene Laufzeiterhöhung nur geringfügige Auswirkungen hätte.

Zusammenfassend soll daher für die folgenden Berechnungen vereinfachend der Mittelwert der Readlänge als konstant betrachtet werden, also  $\bar{k} = \text{const}$ , so dass für die Berechnung der Likelihoods zwischen den Reads eine Laufzeit von  $O(\bar{k}) = O(1)$  veranschlagt wird. Für das Hinzufügen der Kanten des Graphen wird somit die Gesamtlaufzeit auf  $O(|E| \cdot |V|)$  geschätzt.

### Laufzeit zur Bestimmung der Zusammenhangskomponenten

Die Bestimmung und Indexierung der Zusammenhangskomponenten durch graph-tool kann in  $O(|V| + |E|)$  durchgeführt werden [38]. Zusammenhangskomponenten mit mehr als einem Knoten werden zunächst als gefilterte View des Gesamtgraphen erzeugt und anschließend als neuer eigenständiger Graph initialisiert. Da der Filtervorgang für alle Knoten jeder Zusammenhangskomponente durchgeführt werden muss, beträgt die Laufzeit  $O(|C| \cdot |V|)$ .

Die Subgraphen der Zusammenhangskomponenten werden als Elemente einer Liste zusammengefasst. Durch Erstellen einer Liste von Subgraphen kann später eine einfache Iteration über die Komponenten in  $O(|C|)$  ausgeführt werden, ohne dass der Filtervorgang über alle Knoten jeder Komponente wiederholt werden muss. Das anschließende Löschen des ursprünglichen Graphen erfolgt in konstanter Zeit.

Die Laufzeit für die Extraktion der Zusammenhangskomponenten wird also bestimmt durch Identifikation, Indexierung und Filterung der Komponenten mit  $O(|C| \cdot |V|) + O(|V| +$

$|E|) = O(|V| \cdot (|C| + 1) + |E|)$ . Bei realen Daten gibt es in der Regel deutlich mehr Knoten als Cluster bzw. Zusammenhangskomponenten, so dass gilt  $|C| < |V|$ . Würde im Worst Case aber jede Zusammenhangskomponente aus nur einem Knoten bestehen, also  $|C| = |V|$ , so kann die maximale Laufzeit auf  $O(|V| \cdot (|C| + 1) + |E|) = O(|V| \cdot (|V| + 1) + |E|) = O(|V|^2 + |E|)$  geschätzt werden kann.

### Gesamtlaufzeit der Konstruktion des Graphen und der Zusammenhangskomponenten

Wie an entsprechender Stelle bereits beschrieben, ist für die Erzeugung der Knoten eine Laufzeit von  $O(|V|^2)$  (Kap. 3.3.1) erforderlich, das Hinzufügen der Kanten kann in  $O(|E| \cdot |V|)$  erfolgen. Somit wird für den vollständigen Aufbau des Graphen nach (4-4) eine Laufzeit von  $O(|V| \cdot (|V| + |E|))$  benötigt.

$$O(|V|^2) + O(|E| \cdot |V|) = O(|V| \cdot (|V| + |E|)) \quad (4-4)$$

In Zusammenschau mit der für die Extraktion der Zusammenhangskomponenten erforderliche Laufzeit von  $O(|V|^2 + |E|)$  (Kap. 3.3.3) ergibt sich daraus nach (4-5) eine Laufzeit von  $O(|V| \cdot (|V| + |E|))$ .

$$\begin{aligned} & O(|V| \cdot (|V| + |E|)) + O(|V|^2 + |E|) \\ &= O(|V|^2 + |E| \cdot |V| + |V|^2 + |E|) \\ &= O(2 \cdot |V|^2 + |E| \cdot (|V| + 1)) \\ &= O(|V|^2 + |E| \cdot |V|) \\ &= O(|V| \cdot (|V| + |E|)) \end{aligned} \quad (4-5)$$

## 4.2 Laufzeit zur Bestimmung der Allele-Fractions mit maximaler Likelihood

### Laufzeit zur Ermittlung der Allele-Fractions

Funktion `get_candidate_alleles()` Für das Erstellen der Liste muss jeder Knoten in jeder Komponente betrachtet werden, so dass der Vorgang für alle Knoten in  $O(|V|)$  durchführbar ist.

Funktion `get_max_parsimony_n_alleles()` Die Anpassung der Anzahl der Allele kann für jede Komponente aus der Menge der Zusammenhangskomponenten  $C$  in  $O(1)$  erfolgen, so dass die Laufzeit für alle Komponenten  $O(|C|)$  beträgt. Im Worst Case, bei dem jede Komponente nur einen Knoten enthält und somit  $|C| = |V|$  gilt, würde die Laufzeit maximal  $O(|V|)$  betragen.

Funktion `get_candidate_vafs()` Kombinationen mit Wiederholung werden sowohl für die Bestimmung der Allelkombinationen aus diesem Kapitel als auch später in Kap. 3.5.1 für die Ermittlung möglicher Kombinationen von Loci genutzt. Hinsichtlich der Laufzeit und des Speicherplatzbedarfs kann diese Berechnung bei sehr großen Clustern zum dominierenden Faktor werden. Daher soll die Laufzeit an dieser Stelle genauer abgeschätzt werden. Seien dabei für eine bessere Übersichtlichkeit  $n_{alleles}$  als  $n$  und  $n_{cand}$  als  $k$  bezeichnet und es gilt  $n, k \in \mathbb{N}$ . Nach Formel (3-1) werden also  $\binom{n+k-1}{k}$  Kombinationen jeweils in  $O(1)$  erzeugt, d.h. in  $O(\binom{n+k-1}{k})$  für jede Zusammenhangskomponente.

Die Formel (3-1) ergibt sich aus der allgemeinen Formel des Binomialkoeffizienten (4-6), durch den die Anzahl aller Kombinationen ohne Zurücklegen berechnet wird [39].

$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!} \quad (4-6)$$

Mit anderen Worten, die Formel (3-1) kann auf zwei Arten interpretiert werden: es werden aus  $n$  Elementen  $k$  Elemente mit Zurücklegen gezogen oder es werden aus  $n + k - 1$  Elementen  $k$  Elemente ohne Zurücklegen gezogen. Sei also  $m = n + k - 1$ , dann gilt:

$$\binom{m}{k} = \frac{m!}{(m-k)! \cdot k!} \leq m! \quad (4-7)$$

Daraus ergibt sich eine Laufzeit zunächst eine obere Schranke der Laufzeit von  $O(m!) = O((n+k)!)$ . Hierfür soll nun eine kleinere obere Schranke gefunden werden. Um die Fakultät näherungsweise zu berechnen kann die Stirlingsche Formel [40] verwendet werden:

$$n! \approx \left(\frac{n}{e}\right)^n \cdot \sqrt{2 \cdot \pi \cdot n} \quad (4-8)$$

Somit gilt auch (4-9) und ist eine untere Schranke der Fakultät [45].

$$n! \geq \left(\frac{n}{e}\right)^n \quad (4-9)$$

Durch (3-1) und (4-9) lässt sich durch einige Umformungen die obere Schranke der Laufzeit von ursprünglich  $O(n!)$  auf  $O\left(\left(\frac{e \cdot (n+k-1)}{k}\right)^k\right)$  eingrenzen:

$$\begin{aligned}
\binom{n+k-1}{k} &\stackrel{(3-1)}{=} \frac{(n+k-1)!}{(n-1)! \cdot k!} = \frac{\prod_{i=1}^{n+k-1} i}{\prod_{i=1}^{n-1} i \cdot \prod_{i=1}^k i} \\
&= \frac{\prod_{i=n}^{n+k-1} i \cdot \prod_{i=1}^{n-1} i}{\prod_{i=1}^{n-1} i \cdot \prod_{i=1}^k i} = \frac{\prod_{i=n}^{n+k-1} i}{\prod_{i=1}^k i} \\
&= \frac{(n+k-1) \cdot (n+k-2) \cdot \dots \cdot n}{k!} \\
&\leq \frac{(n+k-1)^k}{k!} \\
&\stackrel{(4-9)}{\leq} \frac{(n+k-1)^k}{\left(\frac{k}{e}\right)^k} \\
&= \left(\frac{e \cdot (n+k-1)}{k}\right)^k
\end{aligned} \tag{4-10}$$

Der Binomialkoeffizient verfügt über bestimmte Eigenschaften, insbesondere gelten der Symmetrie- (4-11) und der Additionssatz (4-12), die sich auch in der Struktur des Pascalschen Dreiecks widerspiegeln [40].

$$\binom{n}{k} = \binom{n}{n-k} \tag{4-11}$$

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1} \tag{4-12}$$

Aufgrund der Symmetrie sind die Binomialkoeffizienten jeder Zeile des Dreiecks  $\binom{n}{k_i}$  mit  $k_i \in 1, \dots, n$  und  $i \in \mathbb{N}$  symmetrisch im Bezug auf die Zeilenmitte. Und da nach (4-12) jeder Koeffizient aus der Summe der beiden darüber liegenden Koeffizienten entsteht, sind die Werte der Koeffizienten zu den Zeilenrändern hin abnehmend und weisen in der Zeilenmitte ihr Maximum auf. Ist  $n$  gerade, so befindet sich die Zeilenmitte bei  $k = \frac{n}{2}$ . Ist  $n$  ungerade so befinden sich die Maxima bei  $k = \lceil \frac{n}{2} \rceil$  und  $k = \lfloor \frac{n}{2} \rfloor$  und besitzen aufgrund der Symmetrie den gleichen Wert. Vereinfacht gilt also für ein gegebenes  $n$ , dass der Binomialkoeffizient  $\binom{n}{k_i}$  bei  $k = \frac{n}{2}$  maximal ist.

Daher soll  $k = \frac{n}{2}$  als Worst Case angenommen werden, dann ergibt sich aus (4-10) eine Laufzeit von  $O(3e^{\frac{n}{2}}) \in O(e^n)$ .

$$\begin{aligned}
\left(\frac{e \cdot (n + k - 1)}{k}\right)^k &= \left(\frac{e \cdot (n + \frac{n}{2} - 1)}{\frac{n}{2}}\right)^{\frac{n}{2}} \\
&= \left(\frac{3en - 2e}{\frac{n}{2}}\right)^{\frac{n}{2}} \\
&= \left(\frac{3en - 2e}{n}\right)^{\frac{n}{2}} \\
&\leq \left(\frac{3en}{n}\right)^{\frac{n}{2}} \\
&= 3e^{\frac{n}{2}}
\end{aligned} \tag{4-13}$$

Hinsichtlich der Allelkombinationen entspricht dies einer Zeit von  $O(e^{n_{alleles}})$  für jede Zusammenhangskomponente. Für die Laufzeit über alle Komponenten  $C_i$  gilt dann  $O\left(\sum_{i=1}^{|C|} e^{n_{alleles}}\right)$ , es dominiert also die Komponente, welche die meisten Knoten enthält.

Für sehr große Cluster ist daher die Laufzeit problematisch. Ebenso kommt es zu einem hohen Bedarf an Arbeitsspeicher, aufgrund der großen Anzahl von Kombinationen die dabei erzeugt werden. Um dies bedingt kompensieren zu können, wurden die oben beschriebenen Schwellenwerte eingeführt, die ab einer festgelegten Clustergröße **threshold-cluster-size** nur Allelkandidaten ab einer bestimmten Häufigkeit ihre Sequenz in den Reads **threshold-seq-noise** berücksichtigen.

Die Berechnung der Häufigkeitsverteilungen erfordert weitere Iterationen über die ermittelten Allelkombinationen, dabei werden sämtliche Allele für jede Kombination betrachtet. Dadurch erhöht sich die Laufzeit auf  $O(e^{n_{alleles}}) \cdot O(n_{alleles}) \cdot O(n_{cand})$ . Den höchsten Wert nimmt  $n_{alleles}$  in dem Fall an, dass die Restdivision  $n_{cand} \bmod \phi = 1$ , denn dann gilt  $n_{alleles} = n_{cand} + \phi - 1$ . Da  $\phi$  für alle Zusammenhangskomponenten den gleichen Wert besitzt, gilt  $\phi = const$ . Daraus ergibt sich nach (4-14) eine Laufzeit von  $O(e^{n_{cand}} \cdot n_{cand}^2)$  für die Funktion **get\_candidate\_vafs()**.

$$\begin{aligned}
O(e^{n_{alleles}} \cdot n_{alleles} \cdot n_{cand}) &= O(e^{n_{cand} + \phi - 1} \cdot (n_{cand} + \phi - 1) \cdot n_{cand}) \\
&= O(e^{n_{cand} + \phi} \cdot n_{cand}^2) \cdot \phi \\
&= O(e^{n_{cand}} \cdot n_{cand}^2)
\end{aligned} \tag{4-14}$$

### Laufzeit zur Likelihoodberechnung der Allele anhand der Allele-Fractions

Funktion **get\_allele\_likelihood\_read()**

Im folgenden soll nun die Laufzeit der verschiedenen Suchvorgänge in **get\_allele\_likelihood\_read()** genauer betrachtet werden. Dabei ist zu berücksichtigen,



dass im Allgemeinen der Aufruf einer Kante bei bekannten Source- und Target-Knoten von ihren Knotengraden abhängt. Im Bezug auf die Knotengrade des Readknotens  $d(r_i)$  und des Kandidatenallels  $d(a_j)$  ist eine solche Suche durch graph-tool in  $O(\min(d(r_i), d(a_j)))$  durchführbar [38]. Im Worst Case sind alle Knoten einer Zusammenhangskomponente mit den beiden Knoten  $r_i$  und  $a_j$  verbunden, so dass diese jeweils einen Knotengrad besitzen, welcher der Anzahl der Knoten der Komponente  $V_c$  mit Ausnahme des Knotens selbst entspricht. In diesem Fall beträgt die Laufzeit für den Aufruf einer Kante im Graphen  $O(V_c)$  (4-15).

$$O(\min(d(r_i), d(a_j))) = O(\min(d(V_c + 1), d(V_c - 1))) = O(V_c) \quad (4-15)$$

Im Hinblick auf `get_allele_likelihood_read()` müssen im Worst Case alle Schritte (??) bis (??) ausgeführt werden, so dass sich ihre Laufzeiten summieren. In der unter (??) beschriebenen Suche, werden alle ausgehenden Nachbarn des Reads betrachtet, ihre Anzahl entspricht also dem ausgehenden Knotengrad des Readknotens  $d(r_{i_{out}})$ . War die Suche erfolgreich, so muss die Likelihood der betreffenden Kante aus den Kanteneigenschaften ausgegeben werden. Hierfür ist ein Kantenaufruf wie oben beschrieben in  $O(V_c)$  notwendig. Im ungünstigsten Fall hat jeder Knoten nur ausgehende Kanten und zwar zu allen anderen Knoten der Zusammenhangskomponente, dann beträgt die Laufzeit nach (4-16) für alle Reads der Komponente  $O(V_c^2)$ .

$$O(V_c) \cdot (O(d(r_{i_{out}})) + O(V_c)) = O(V_c \cdot ((V_c - 1) + V_c)) = O(V_c^2) \quad (4-16)$$

Die Suche bei (??) erfolgt analog über alle eingehenden Nachbarn in  $O(V_c^2)$ , allerdings mit zusätzlicher Likelihoodberechnung der rückläufigen Kante in  $O(k)$ . Daraus ergibt sich insgesamt eine Laufzeit von  $O(V_c^2 + k) = O(V_c^2)$ , da  $\bar{k} = \text{const}$  (siehe Kap. 4.1).

Die Laufzeit für das Auffinden eines geeigneten CIGAR-Tupels (??) kann in  $O(V_c^3)$  erfolgen (siehe Formel (4-18)). Auch hier führt die Likelihoodberechnung einer einzelnen Kante zu keiner relevanten Änderung der Laufzeit.

Die Rückgabe eines konstanten Wertes bei (??) benötigt lediglich eine Laufzeit von  $O(1)$ .

Aus den insgesamt vier Schritten, die im ungünstigsten Fall alle ausgeführt werden ergibt sich für `get_allele_likelihood_read()` eine Gesamtlaufzeit von  $O(V_c^3)$ .

$$\begin{aligned} & O(V_c^2) + O(V_c^2) + O(V_c^3) + O(1) \\ &= O(2 \cdot V_c^2 + V_c^3) \\ &= V_c^3 \end{aligned} \quad (4-17)$$

Funktion `get_cigar_tuples()`

Die hierfür erforderliche Laufzeit setzt sich zusammen aus den Suchvorgängen für die beiden Knotenmengen  $R_{source}$  und  $R_{target}$ , diese ist in graph-tool in  $O(V_c + V_c) = O(V_c)$  durchführbar (Kap. 3.3.2). Seien im Worst Case die Hälfte aller Knoten der Zusammenhangskomponente in  $R_{source}$  und die andere Hälfte in  $R_{target}$ . Um die Eigenschaft der Zusammenhangskomponente zu gewährleisten muss daher eine Kante zwischen beiden Knotenmenge existieren. Sei dies zudem eine entgegen gerichtete Kante. Sind die beiden Knoten  $(v, w)$  dieser Kante allerdings jeweils an letzter Position in  $R_{source}$  und  $R_{target}$ , so erfolgen für alle übrigen Knoten der Komponente Kantenaufrufe für die Hin- und Rückrichtung. Wie bereits oben beschrieben, benötigt ein Kantenaufruf eine Laufzeit von  $O(\min(d(v), d(w)))$ , mit den Knotengraden  $d(v)$  von Knoten  $v \in R_{source}$  und  $d(w)$  von Knoten  $w \in R_{target}$  [38]. Seien nun außerdem  $v$  mit allen Knoten in  $R_{source}$  und  $w$  mit allen Knoten in  $R_{target}$  verbunden, so besitzt  $v$  einen Knotengrad von  $d(v) = \frac{V_c-1}{2}$ . Der Knoten  $w$  besitzt einen Knotengrad von  $d(w) = \frac{V_c}{2}$ , da eine entgegen gerichtete Kante von  $w$  zu  $v$  verläuft. Dann ergibt sich nach (4-18) eine Gesamtlaufzeit für `get_cigar_tuples()` von  $O(V_c^3)$ .

$$\begin{aligned}
& O(V_c) + O(V_c^2 \cdot \min(d(v), d(w))) \\
&= O\left(V_c + V_c^2 \cdot \min\left(\frac{V_c-1}{2}, \frac{V_c}{2}\right)\right) \\
&= O(V_c + V_c^2 \cdot V_c) \\
&= O(V_c^3)
\end{aligned} \tag{4-18}$$

Funktion `calc_vafs_likelihood_read()` Formel (2-17)

Es über die relativen Häufigkeiten jedes Allels einer Allele-Fraction iteriert und ihr Produkt mit der Readlikelihood aufsummiert. Da die Länge der Allelfraction genau der Anzahl der Allele einer Zusammenhangskomponenten  $n_{cand}$  entspricht, erhöht sich hierdurch die Laufzeit um den Faktor  $O(n_{cand})$ .

Funktion `calc_vafs_likelihood()` Formel (2-18) Dadurch wird die Laufzeit zusätzlich um den Faktor der Anzahl der Reads der Zusammenhangskomponente  $O(V_c)$  erhöht.

in `noderad_main.py` für alle Allele-Fractions Dazu muss über alle VAFs iteriert werden. Die Anzahl der VAFs entspricht dabei genau der Anzahl der Allelkombinationen. Wie bereits in Kap. 4.2 besprochen, kann diese hinsichtlich der Laufzeit mit  $O(e^{n_{alleles}})$  abgeschätzt werden (Formel (4-13)). Da im Worst Case  $n_{alleles} = n_{cand} + \phi - 1$ , ergibt sich eine Laufzeit von  $O(e^{n_{cand} + \phi - 1}) \in O(e^{n_{cand}})$ . Um diesen Faktor erhöht sich also die Gesamtlaufzeit bei der Berechnung der Liste der Gesamtwahrscheinlichkeiten.

Maximumsbestimmung: Hinsichtlich der Laufzeit muss jede Lösung der Liste betrachtet werden. Da die Länge der Liste der Anzahl der VAFs entspricht, kann die Laufzeit hierfür nach (4-13) auf maximal  $O(e^{n_{cand}})$  geschätzt werden.

### Gesamtlaufzeit zur Bestimmung der Allele-Fractions mit maximaler Likelihood

Die geschätzten Laufzeiten der einzelnen Schritte wurden bereits an entsprechender Stelle diskutiert und sollen hier noch einmal zusammenfassend aufgeführt werden. Zur übersichtlicheren Darstellung soll die Anzahl der Allele hier nur noch mit  $n$  bezeichnet werden, dies entspricht der bisher verwendeten Variable  $n_{cand}$ . Für jede Zusammenhangskomponente werden vorab die Kandidatenallele bestimmt  $O(V_c)$ , die Anzahl der tatsächlich zu erwartenden Allele berechnet  $O(V_c)$  und die Kombinationen möglicher Häufigkeitsverteilungen ermittelt  $O(e^n \cdot n^2)$ . Anschließend erfolgt die Likelihoodberechnung in  $O(V_c^3)$  über alle Allelfraktionen  $O(e^n)$ , für alle Reads  $O(V_c)$  und im Vergleich jedes Reads zu allen Kandidatenallelen  $O(n)$ . Aus den Lösungen wird im Anschluss das Maximum bestimmt  $O(e^n)$ . Daraus ergibt sich nach 4-19 eine Gesamtlaufzeit für die Bestimmung der Allelfraktion mit maximaler Likelihood von  $O(e^n \cdot (n^2 + n \cdot V_c^4))$ .

$$\begin{aligned}
 O(V_c) + O(V_c) + O(e^n \cdot n^2) + O(e^n) \cdot O(V_c) \cdot O(n) \cdot O(V_c^3) + O(e^n) \\
 = O(2V_c + e^n \cdot n^2 + e^n + e^n \cdot n \cdot V_c^4) \\
 = O(2V_c + e^n \cdot (n^2 + 1 + n \cdot V_c^4)) \\
 = O(e^n \cdot (n^2 + n \cdot V_c^4))
 \end{aligned} \tag{4-19}$$

Im Worst Case entspricht  $n$  genau der Anzahl der Reads in der Zusammenhangskomponente und es gilt  $n = n_{cand} = V_c$ , so dass die Laufzeit dann auf  $O(e^{V_c} \cdot V_c^5)$  geschätzt werden kann (4-20).

$$\begin{aligned}
 O(e^n \cdot (n^2 + n \cdot V_c^4)) &= O(e^{V_c} \cdot (V_c^2 + V_c \cdot V_c^4)) \\
 &= O(e^{V_c} \cdot (V_c^2 + V_c^5)) \\
 &= O(e^{V_c} \cdot V_c^5)
 \end{aligned} \tag{4-20}$$

Die Laufzeit wird also durch die Anzahl der VAFs dominiert und kann bei großen Clustern mit vielen Allelen problematisch werden. Die konfigurierbaren Schwellenwerte `threshold-cluster-size` und `threshold-seq-noise` ermöglichen eine Anpassung solcher großen Cluster, um die Anzahl der Kombinationen und damit den Rechenaufwand flexibel reduzieren zu können.

## 4.3 Laufzeit zur Bestimmung der Loci mit maximaler Likelihood

### Laufzeit zur Bestimmung der möglicher Loci-Kombinationen

Wie bereits in Kap. 4.2 besprochen, benötigt die Generierung aller Allelkombinationen eine Laufzeit von  $O(e^n)$ , wobei  $n = n_{cand}$  die Anzahl der beobachteten Allele der Zusammenhangskomponente ist. Für diese Kombinationen werden jeweils alle  $n!$  Permutationen

gebildet, also mit einer Laufzeit von  $O(n!)$ . Für jede Permutation erfolgt über ihre jeweilige Länge die Gruppierung zu den Loci. Die Länge jeder Permutation kann wegen der Funktion `get_max_parsimony_n_alleles()` (vgl. Kap. 3.4.1) maximal  $n_{alleles} = n_{cand} + \phi - 1$  betragen. Wegen  $\phi = const$  beträgt die Laufzeit hier also  $O(n_{cand}) = O(n)$ . Die Sortierung bei Python erfordert eine Laufzeit von  $O(n \cdot \log(n))$  [46]. Die beiden Sortiervorgänge erfolgen zum einen über jeden Locus mit der Länge  $\phi$  entsprechend seiner Ploidie und über alle Loci innerhalb der Permutation. Die Anzahl der Loci entspricht  $\frac{n}{\phi}$ , dadurch ergeben sich für die Sortiervorgänge Laufzeiten von  $O(\phi \cdot \log(\phi))$  bzw.  $O\left(\frac{n}{\phi} \cdot \log\left(\frac{n}{\phi}\right)\right)$ . Unter Berücksichtigung, dass  $\phi = const$  gilt, beträgt die Laufzeit für `get_candidate_loci()` insgesamt  $O(e^n \cdot n! \cdot n^2 \cdot \log(n))$ .

$$\begin{aligned}
& O(e^n) \cdot O(n!) \cdot O(n) \cdot O(\phi \cdot \log(\phi)) \cdot O\left(\frac{n}{\phi} \cdot \log\left(\frac{n}{\phi}\right)\right) \\
&= O\left(e^n \cdot n! \cdot n \cdot \phi \cdot \log(\phi) \cdot \frac{n}{\phi} \cdot \log\left(\frac{n}{\phi}\right)\right) \\
&= O\left(e^n \cdot n! \cdot n^2 \cdot \log(\phi) \cdot \log\left(\frac{n}{\phi}\right)\right) \\
&\in O(e^n \cdot n! \cdot n^2 \cdot \log(n))
\end{aligned} \tag{4-21}$$

Große Cluster sind hier also hinsichtlich der Laufzeit und des Speicherbedarfs noch problematischer als bei der Generierung der Allelkombinationen in der Funktion `get_candidate_vafs()` (Kap. 4.2). Eine sinnvolle Anpassung der Grenzwerte `threshold-cluster-size` und `threshold-seq-noise` ermöglicht diesbezüglich eine Verbesserung. Allerdings ist an dieser Stelle ein effizienterer Ansatz zur Bestimmung der Loci-Kombinationen möglich, der in diesem Prototypen zwar keine Anwendung fand, aber für die spätere Implementierung in Rust von Bedeutung sein könnte. Dieser Ansatz wird im Ausblick in Kap. 6 kurz beschrieben.

### **Laufzeit für die Bestimmung der Loci-Kombination mit maximaler Likelihood**

Die Gesamtlaufzeit ergibt sich wiederum aus den einzelnen Laufzeiten der oben beschriebenen Funktionen. Diese Funktionen werden über allen Permutationen der Allelkombinationen in `noderad_main.py` aufgerufen, wodurch die Laufzeit um den Faktor  $O(e^n \cdot n!)$  steigt (vgl. Kap. 3.5.1). Die Indikatorfunktion prüft die Bedingung über die gesamte Länge einer der Permutationen und benötigt somit eine Laufzeit von  $O(n)$ . Anschließend wird in `calc_loci_likelihoods()` über alle Loci einer Permutation in  $O(\frac{n}{\phi})$  iteriert und die Funktion `get_allele_likelihood_allele()` aufgerufen. Diese ermittelt bei jedem Locus in  $O(\phi)$  die Kombinationen ohne Wiederholung und ohne Beachtung der Reihenfolge. Nach (3-2) und (4-13) erhöhen die Kombinationen dabei die Laufzeit um den Faktor  $O\left(\binom{n}{2}\right) \in O(e^n)$ . Die Bestimmung der CIGAR-Tupel für jede Kombination vergrößert die Laufzeit zusätzlich um den Faktor  $O(V_c^3)$  (siehe Kap. 3.4.2). Über die CIGAR-Tupel wird

schließlich für jede Base die Likelihood aus den Heterozygotiewahrscheinlichkeiten berechnet, analog zur Likelihoodberechnung zweier Reads in Kap. 3.3.2 kann dies mit  $O(\bar{k})$  veranschlagt werden, wobei  $\bar{k} = \text{const}$  gilt. Da für alle Permutationen der Kombinationen die Berechnung Likelihood erfolgt, wird die anschließende Maximumsbestimmung über deren Anzahl also in  $O(e^n \cdot n!)$  durchgeführt.

Nach (4-22) ergibt sich damit für die Berechnung der wahrscheinlichsten Loci-Kombination insgesamt eine Laufzeit von  $O(e^n \cdot n! \cdot V_c^3)$ .

$$\begin{aligned}
& O(e^n \cdot n!) \cdot \left( O(n) + O\left(\frac{n}{\phi}\right) \cdot O(\phi) \cdot O(e^n) \cdot O(V_c^3) \cdot O(\bar{k}) \right) + O(e^n \cdot n!) \\
&= O(e^n \cdot n! \cdot (n + n \cdot e^n \cdot V_c^3) + e^n \cdot n!) \\
&= O(e^n \cdot n! \cdot n + e^{2n} \cdot n! \cdot n \cdot V_c^3) \\
&= O(e^{2n} \cdot n! \cdot n \cdot V_c^3) \\
&\leq O(e^{2n} \cdot n! \cdot (n + 1) \cdot V_c^3) \\
&= O(e^{2n} \cdot (n + 1)! \cdot V_c^3) \\
&\in O(e^n \cdot n! \cdot V_c^3)
\end{aligned} \tag{4-22}$$

Wie bereits in Kap. 4.2 durchgeführt, soll auch hier der Worst Case betrachtet werden, bei dem die Anzahl der Kandidatenallele genau der Anzahl der Reads in der Zusammenhangskomponente entspricht, also  $n = n_{cand} = V_c$ . Dann ergibt sich für die Berechnung der wahrscheinlichsten Loci-Zuordnung nach (4-23) eine maximale Laufzeit von  $O(e^{V_c} \cdot V_c!)$ .

$$\begin{aligned}
O(e^n \cdot n! \cdot V_c^3) &= O(e^{V_c} \cdot V_c! \cdot V_c^3) \\
&\leq O(e^{V_c} \cdot V_c! \cdot (V_c + 1) \cdot (V_c + 2) \cdot (V_c + 3)) \\
&= O(e^{V_c} \cdot (V_c + 3)!) \\
&\in O(e^{V_c} \cdot V_c!)
\end{aligned} \tag{4-23}$$

Die Anzahl der Permutationen dominiert durch  $O(V_c!)$  die Laufzeit noch stärker als bei der Ermittlung der Allelkombinationen in Kap. 4.2. Dies ist für große Cluster sehr problematisch. Die Anpassung der Schwellenwerte `threshold-cluster-size` und `threshold-seq-noise` können dies zwar in einem gewissen Rahmen für den Prototyp kompensieren. Bei der späteren Implementierung in Rust sollte hier jedoch ein effizienterer Ansatz gewählt werden. Eine effizientere Vorgehensweise diesbezüglich wird in Kap. 6 vorgestellt.

## 4.4 Abschließende Laufzeitanalyse des gesamten Algorithmus

Für jeden Konstruktions- und Berechnungsschritt wurden die Laufzeiten bereits an entsprechender Stelle angesprochen. Zusammenfassend ergab sich dabei für die einmalige Konstruktion des Graphen und die Extraktion seiner Zusammenhangskomponenten  $C$  eine

Laufzeit von  $O(V \cdot (V + E))$  (Kap. 4.1). Für jede Zusammenhangskomponente erfolgte die Berechnung der wahrscheinlichsten Allelkombination in  $O(e_c^V \cdot V_c^5)$  (Kap. 4.2) und die Berechnung der wahrscheinlichsten Loci-Zuordnung in  $(e_c^V \cdot V_c!)$  (Kap. 4.3). Die Laufzeiten können also wie folgt zusammengefasst werden:

$$\begin{aligned}
& O(V \cdot (V + E)) + O\left(\sum_{c=1}^{|C|} (e^{V_c} \cdot V_c^5 + e^{V_c} \cdot V_c!)\right) \\
&= O\left(V \cdot (V + E) + \sum_{c=1}^{|C|} (e^{V_c} \cdot V_c^5) + \sum_{c=1}^{|C|} (e^{V_c} \cdot V_c!)\right) \quad (4-24) \\
&\in \left(O \sum_{c=1}^{|C|} (e^{V_c} \cdot V_c!)\right)
\end{aligned}$$

Besitzt im Worst Case der Graph nur eine einzige Zusammenhangskomponente, welche alle Knoten des Graphen enthält, so dass gilt  $|C| = 1$  und  $V_c = V$  so würde sich hieraus eine maximale Laufzeit von  $O(e^V \cdot V!)$  ergeben.

## Kapitel 5

# Evaluation an simulierten Datensätzen

### 5.1

#### 5.1.1





## Kapitel 6

# Zusammenfassung und Ausblick

var callen

verschiedene Individuen mit einander vergleichen bei gleichem Ansatz, aber nun mit locus seq. aus den vcf's f mehrere individuen als eingabe -> vergleich der individuen

Diversität -> je diverser, desto besser geht es der Population

SE-Reads besser geeignet für Analyse: bei mutation im bereich einer der rev-schnittstelle würde bei PE nicht geschnitten werden und der read würde nicht geclustert und ausgewertet werden -> heterozygotie nicht mehr sichtbar, bei SE sind aber die reads unabhängig von einander, die Heterozygotie an dieser stelle würde erfasst werden

eventuell Vgl mit ILP:

-> ILP wird sehr groß, die Lösung wird 2 schrittig gefunden, aber dadurch kein globales Optimum, dass für jede optimale Zuordnung diese durch erwartete Sequenziertiefe erklärt -> es könnte ebenso eine nicht optimale Lösung im ILP eine optimale Lösung in zusammenschau mit der seq-tiefe werden

=> Seq-tiefe ist nur sehr bedingt als Maß geeignet, z.B. Cutting-Enzym schneidet nicht richtig -> seq-tiefe und häufigkeit ändern sich

prototyp, zunächst für ploidie 2 -> bei höherer ploidie multiples Sequenzalignment [47]  
[25]

## 6.1

### 6.1.1



# Abbildungsverzeichnis

2.1	Übersicht der Prozesse des Workflows . . . . .	14
3.1	Ausschnitt einiger Zusammenhangskomponenten des Gesamtgraphen . . . .	26
3.2	Beispiele von einzelnen Zusammenhangskomponenten unterschiedlicher Größe	27



# Algorithmenverzeichnis

3.1	Suche eines Alignments zwischen Read und Kandidatenallel . . . . .	32
3.2	Berechnung der Likelihood zwischen Read und Kandidatenallel . . . . .	34
3.3	Bestimmung der Likelihood der Allele innerhalb einer Loci-Kombination . .	42
3.4	Bestimmung der Likelihood zwischen zwei Kandidatenallelen hinsichtlich der Heterozygotiewahrscheinlichkeiten . . . . .	43
3.5	CIGAR-Tupel bestimmen . . . . .	45



# 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] DUNHAM, IAN, ANSHUL KUNDAJE, SHELLEY F ALDRED, PATRICK J COLLINS, CARRIE A DAVIS, FRANCIS DOYLE, CHARLES B EPSTEIN, and SETH FRIETZE: *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 ETTEN, 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] GROUP, THE SAM/BAM FORMAT SPECIFICATION WORKING: *Sequence Alignment/Map Format Specification*. source: <https://samtools.github.io/hts-specs/SAMv1.pdf>.
- [21] LI, HENG, BOB HANDSAKER, ALEC WYSOKER, TIM FENNELL, JUE RUAN, NILS HOMER, GABOR MARTH, GONCALO ABECASIS, and RICHARD DURBIN: *The sequence alignment/map format and samtools*. Bioinformatics (Oxford, England), 25:2078–9, Aug 2009.
- [22] DANECEK, PETR, ADAM AUTON, GONCALO ABECASIS, CORNELIS A. ALBERS, ERIC BANKS, MARK A. DEPRISTO, ROBERT E. HANDSAKER, GERTON LUNTER, GABOR T. MARTH, STEPHEN T. SHERRY, GILEAN MCVEAN, and RICHARD DURBIN: *The variant call format and vcftools*. Bioinformatics (Oxford, England), 27:2156–8, Aug 2011.
- [23] CATCHEN, JULIAN, PAUL A. HOHENLOHE, SUSAN BASSHAM, ANGEL AMORES, and WILLIAM A. CRESKO: *Stacks: an analysis tool set for population genomics*. Molecular ecology, 22:3124–40, Jun 2013.
- [24] LI, HENG: *Minimap2: pairwise alignment for nucleotide sequences*. Bioinformatics, 34(18):3094–3100, 05 2018.
- [25] DURBIN, RICHARD, SEAN R. EDDY, ANDERS KROGH, and GRAEME MITCHISON: *Markov chains and hidden markov models*. In *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, pages 47–80. Cambridge University Press, 1998.
- [26] YOON, BYUNG-JUN: *Hidden markov models and their applications in biological sequence analysis*. Current genomics, 10:402–15, Sep 2009.
- [27] KNUDSEN, BJARNE and MICHAEL M. MIYAMOTO: *Sequence alignments and pair hidden markov models using evolutionary history*. Journal of molecular biology, 333:453–60, Oct 2003.
- [28] KUHNER, MARY K. and JAMES MCGILL: *Correcting for sequencing error in maximum likelihood phylogeny inference*. G3 (Bethesda, Md.), 4:2545–52, Nov 2014.

- [29] KÖSTER, JOHANNES and SVEN RAHMANN: *Snakemake—a scalable bioinformatics workflow engine*. *Bioinformatics*, 28(19):2520–2522, 08 2012.
- [30] 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 (OASISs)*, pages 49–56, Dagstuhl, Germany, 2012. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- [31] MARTIN, MARCEL: *Cutadapt removes adapter sequences from high-throughput sequencing reads*. *EMBnet.journal*, 17(1):10–12, 2011.
- [32] ANDREWS, SIMON, FELIX KRUEGER, ANNE SEGONDS-PICHON, LAURA BIGGINS, CHRISTEL KRUEGER, and STEVEN WINGETT: *FastQC*. Babraham Institute, January 2012. source: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- [33] 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.
- [34] PEIXOTO, TIAGO P.: *The graph-tool python library*. figshare, 2014.
- [35] COCK, PETER J. A., TIAGO ANTAO, JEFFREY T. CHANG, BRAD A. CHAPMAN, CYMON J. COX, ANDREW DALKE, IDDO FRIEDBERG, THOMAS HAMELRYCK, FRANK KAUFF, BARTEK WILCZYNSKI, and MICHIEL J. L. DE HOON: *Biopython: freely available Python tools for computational molecular biology and bioinformatics*. *Bioinformatics*, 25(11):1422–1423, 03 2009.
- [36] COCK, PETER J. A., CHRISTOPHER J. FIELDS, NAOHISA GOTO, MICHAEL L. HEUER, and PETER M. RICE: *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants*. *Nucleic Acids Research*, 38(6):1767–1771, 12 2009.
- [37] HEGER, ANDREAS and KEVIN JACOBS: *FastQC*. source: <https://github.com/pysam-developers/pysam>, documentation: <https://pysam.readthedocs.io/en/latest/index.html>.
- [38] PEIXOTO, TIAGO DE PAULA: *graph-tool 2.35 documentation: graph\_tool - efficient graph analysis and manipulation*. source: [https://graph-tool.skewed.de/static/doc/graph\\_tool.html](https://graph-tool.skewed.de/static/doc/graph_tool.html).
- [39] VOSS, W, SAUERBIER T: *Taschenbuch der Statistik*, chapter 8. Kombinatorik, pages 281–281. Fachbuchverlag Leipzig, 2004.

- [40] BRONSTEIN, I. N. and K.A. SEMENDJAJEW: *Taschenbuch der Mathematik*, 22. Auflage, chapter 2.2. Kombinatorik, pages 103–105. Verlag Nauka, Moskau and BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 1985.
- [41] PYTHON SOFTWARE FOUNDATION, VAN ROSSUM, GUIDO: *itertools — functions creating iterators for efficient looping*. source: <https://docs.python.org/3/library/itertools.html>.
- [42] EWING, B. and P. GREEN: *Base-calling of automated sequencer traces using phred. ii. error probabilities*. Genome research, 8:186–94, Mar 1998.
- [43] SCHIRMER, MELANIE, ROSALINDA D’AMORE, UMER Z. IJAZ, NEIL HALL, and CHRISTOPHER QUINCE: *Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data*. BMC Bioinformatics, 17(1):125, March 2016.
- [44] PAULA PEIXOTO, TIAGO DE: *Mail conversation/graph-tool: problem building weighted graph*, September 2019. source: <https://www.mail-archive.com/graph-tool@skewed.de/msg03486.html>.
- [45] WITTENBURG, KIM: *Algorithmen und datenstrukturen*, 2017. source: [https://www2.informatik.uni-hamburg.de/fachschaft/wiki/images/3/3e/AD-Skript\\_\(WS16-17\).pdf](https://www2.informatik.uni-hamburg.de/fachschaft/wiki/images/3/3e/AD-Skript_(WS16-17).pdf), pages 156-157.
- [46] FOUNDATION, PYTHON SOFTWARE and DANIEL STUTZBACH: *Python — time-complexity of various operations*. source: <https://wiki.python.org/moin/TimeComplexity>.
- [47] LIU, YONGCHAO, BERTIL SCHMIDT, and DOUGLAS L. MASKELL: *Msaprobs: multiple sequence alignment based on pair hidden markov models and partition function posterior probabilities*. Bioinformatics (Oxford, England), 26:1958–64, Aug 2010.



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 16. Januar 2021

Muster Mustermann

