Hochdurchsatzdatenanalyse Teil 2 - M-BS2-S4B Projekt Ötzi Dokumentation

Fachbereich:08 Biologie

der Justus-Liebig-Universität

Rabia Ayche Serradj,
Dominik March,
Jan Michel Pfeifer

17. September 2023

Inhaltsverzeichnis

1 Aufgabenstellung

Im Jahr 1991 erfolgte die Entdeckung von Ötzi, einer Gletschermumie, die in den Ötztaler Alpen in Österreich gefunden wurde. Eine komparative Analyse des Genoms von Ötzi durch Keller *et al.* [1], mittels *Next-Generation-Sequencing*, konnte bakterielle Sequenzdaten, die *Borrelia burgdorferi* zugeordnet werden konnten, generieren. Dies gab Anlass für die Vermutung, dass Ötzi an Borreliose erkrankt sei.

Zur Untersuchung dieser Annahme wird im Rahmen dieses Projekts ein *Workflow* implementiert, der aus den folgenden vier Aufgabenstellungen besteht.

- Identifikation bakterieller Sequenzen in den Datensätzen
- Taxonomische Profilierung
- Rekonstruktion des *Borrellia burgdorferi* (Draft)-Genoms
- Rekonstruktion der mtDNA von Ötzi

Die Rekonstruktion des *Borrellia burgdorferi* (Draft)-Genoms sowie der mtDNA beinhaltet folgenden Aufgaben.

- *Mapping* gegen das entsprechende Referenzgenom
- *Consensus Calling*
- Annotation der erzeugten Contigs
- Bestimmung von 5'/3' Substituionsraten
- Phylogenetische Charakterisierung

[1] Keller et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing, 2012

2 Daten

Die verwendeten Daten stammen aus der Sequenzierung durch Keller *et al.* [1], die in der *ENA* Datenbank [2] unter der Projektnummer PRJEB2830 abrufbar sind.

Bei diesen Daten handelt es sich um *paired-end* Daten, im Fastq Format, die mittels der *ABI SOLiD* Technik sequenziert wurden. Insgesamt umfasst dieses Projekt zwölf Datensätze. Einige dieser Datensätze enthalten Sequenzen, welche farbkodiert sind. Diese Daten sind für eine Analyse unbrauchbar und wurden für die eingehende Analyse verworfen.

Für den *Workflow* wurden die restlichen Daten mit den folgenden *Run Accession Number* verwendet.

- ERR069107
- ERR069108
- ERR069109

Datensätze, in denen die *reverse reads* enthalten sind, beinhalten kürzere *reads* (ca. 25 Basenpaare) als die *forward reads* (ca. 50 Basenpaare). Während der Analyse ist festgestellt worden, dass die Qualität der *reverse reads* schlechter verglichen zu den *forward reads* war.

Durch die mangelhafte Qualität der *reverse reads* konnten diese nicht getrimmt werden. Dadurch haben wir uns entschlossen, diese Daten für den *Workflow* nicht zu verwenden. Der finale Datensatz bestand daher aus den drei Datensätzen der *forward reads*, mit den oben gelisteten *Run Accession Number*.

Der in Kapitel 3 implementierte Workflow ist somit für *single-end* Daten konzipiert.

[1] Keller et al, New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing, 2012

[2] European Nucleotide Archive (`https://www.ebi.ac.uk/ena/browser/home`)

3 Workflow

Der implementierte *Workflow* wurde mittels eines Workflow Management Systems, *Snakemake*, umgesetzt.

Entsprechend der Grafik in Abbildung 3.1, wird zu Beginn des *Workflows* die Qualität der in den Datensätzen enthaltenen *reads* mittels *FastQC* bestimmt. Die Ergebnisse von *FastQC* wurden dann mit *MultiQC* visuell zusammengefasst und aufbereitet.

Anschließend wurden die *reads* mit schlechter Qualität und Sequenzieradaptersequenzen mittels *Trim Galore* entfernt. Dazu ruft *Trim Galore* die Anwendung *Cutadapt* auf.

Nach dem Trimmen wurde *MultiQC* erneut aufgerufen, um die Qualität der verbleibenen *reads* zu bestimmen.

Die taxonomische Profilierung der Datensätze wurde sowohl mit *kraken2* als auch *kaiju* umgesetzt. Die Visualisierung der Ergebnisse der taxonomischen Klassifikation erfolgte mit *krona*.

Nachfolgend wurden die aufbereiteten *reads* gegen die Referenz der humanen mtDNA, welche von der Seite des *NCBI* [1] stammt, gemappt. Das *mapping* wurde mit *bowtie2* durchgeführt.

Die aus dem *mapping* resultierenden *Sequence Alignment Map (SAM)* Dateien wurden anschließend sortiert, indiziert und in *Binary Alignment Map (BAM)* Dateien umgewandelt.

Reads, die gegen das humane mtDNA Referenzgenom gemappt werden konnten, wurden mit dem Tool *samtools* zu einem Genom zusammengefasst, indem aus den indizierten BAM Dateien eine Konsensussequenz erstellt wird.

Die Annotation des rekonstruierten mitochondrialen Genoms erfolgte mittels *prokka*.

[1] National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov)

Die Bestimmung der 5'/3' Substituionsraten des rekonstruierten mitochondrialen Genoms erfolgte nach der Annotation durch *MapDamage*.

Zuletzt ist eine phylogenetische Charakterisierung des rekonstruierten mitochondrialen Genoms durchgeführt worden. Zunächst ist hierfür die Haplogruppe der mtDNA von Ötzi durch *haplogrep* ermittelt worden.

Aus der *AmtDB* [2] sind Daten entnommen worden, die aus den Epochen, Kupfer-, Bronzezeit und Neolithikum entstammen. Die hierbei gesammelten Daten beschränkten sich lediglich auf Europa.

Anschließend erfolgte die Berechnung eines multiplen Sequenzalignments mit *Clustal Omega* für die gesammelten Daten und das rekonstruierte mitochondirale Genom von Ötzi.

Auf Grundlage des multiplen Sequenzalignments erfolgte schließlich die Erstellung eines phylogenetischen Baums, der auf dem UPGMA [3] Algorithmus basiert.

Alle hierfür verwendeten Tools und die jeweiligen Versionen sind in Kapitel 4 aufgelistet.

[2] Ancient mtDNA database (https://amtdb.org)
[3] Unweighted Pair Group Method with Arithmetic mean

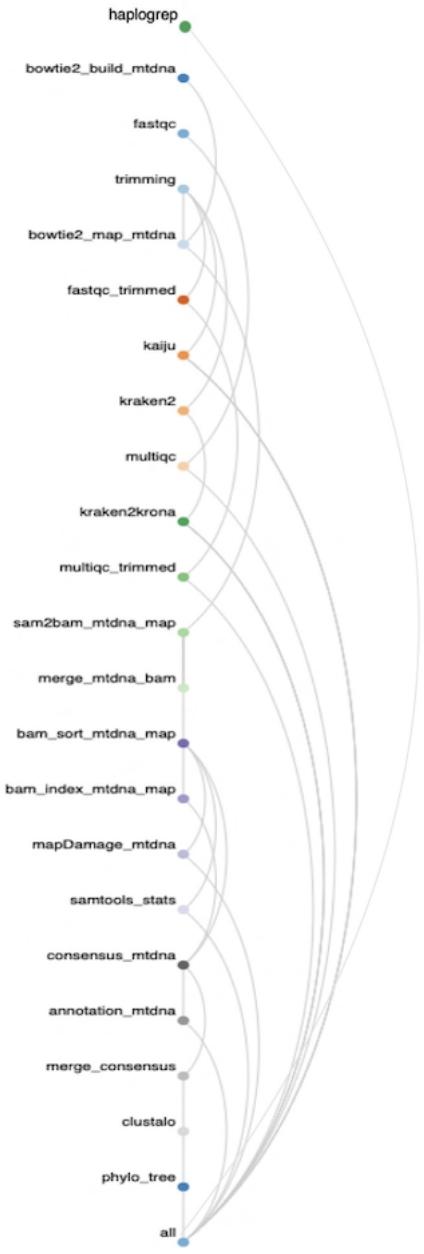

Abbildung 3.1: Ablauf des implementierten Workflows

4 Tools

Die Implementierung des Workflows zur Rekonstruktion und der Analyse sowohl der mtDNA als auch des *Borrelia burgdoferi* (Draft)-Genoms erfolgte mittels des Workflow Management Systems *Snakemake*.
Hierfür wurden verschiede Tools verwendet, die in der nachfolgenden Tabelle 4.1 aufgelistet sind.

Tabelle 4.1: Liste aller verwendeten Tools, Versionen und Links

Tools	**Version**	**Link**
FastQC	0.12.1	https://home.cc.umanitoba.ca/∼psgendb/doc/fastqc.help
MultiQC	1.15	https://multiqc.info/docs/
Trim Galore	0.6.10	https://github.com/FelixKrueger/TrimGalore
Bowtie2	2.5.1	https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
Kraken2	2.1.3	https://github.com/DerrickWood/kraken2
Kaiju	1.9.2	https://github.com/bioinformatics-centre/kaiju
Samtools	1.17	http://www.htslib.org/doc/samtools.html
Prokka	1.14.6	https://github.com/tseemann/prokka
Snakemake	7.32.3	https://snakemake.readthedocs.io/en/stable/
Krona	2.8.1	https://github.com/marbl/Krona/wiki
Clustal Omega	1.2.4	https://www.ebi.ac.uk/Tools/msa/clustalo/
Haplogrep	2.4.0	https://github.com/seppinho/haplogrep-cmd
Python3	3.10.12	https://www.python.org/downloads/
BioPython	1.81	https://https://biopython.org/wiki/Download

5 Ergebnisse

Dieses Kapitel stellt alle Ergebnisse dar, die vom Workflow generiert wurden. Zudem ist eine visuelle Dastellung dieser Ergebnisse mittels einer *HTML*[1] Seite umgesetzt worden.

5.1 Taxonomische Profilierung

Die taxonomische Klassifikation dient dazu, einen Überblick über die in den Daten enthaltenen *reads* zu erhalten. Dazu wurde für jeden der drei Datensätze eine seperate taxonimische Klassifikation durchgeführt. Hierfür wurden jeweils eine Datenbank, welche taxonomische Informationen aus der NCBI enthält, verwendet.

5.1.1 Kraken2

Die taxonomischen Klassifikation erfolgte mittels *kraken2* und basiert auf exakten *k-mer matches*. Die Analyse mit *kraken2* ergab, dass nur ein sehr geringer Teil der generierten Sequenzen *Borrelia burgdorferi* zugeordnet werden konnten (siehe Tabelle 5.1).

Tabelle 5.1: Anteile der Reads, die *Borrelia burgdorferi* zugeordnet werden konnten

Datensatz	**Anteil Borrrelia burgdorferi**
ERR069107_1	0,0003%
ERR069108_1	0,00005%
ERR069109_1	0,0002%

Aufgrund dieser sehr geringen Anteile gehen wir davon aus, dass es sich bei den von *kraken2* klassifizierten *Borrelia burgdorferi* Sequenzen um flasch-positive Ergebnisse handelt.

[1] Hypertext Markup Language

Insgesamt konnten durch *kraken2* etwas 34 % bis 36 % aller *reads* taxonomisch klassifiziert werden. Die restlichen nicht klassifizierten *reads* haben einen humanen Ursprung.

Durch diese Ergebnisse konnte nicht bestätigt werden, dass Ötzi an Borreliose erkrant war.

Die folgenden Tabellen 5.2 bis 5.4 listen für jeden Datensatz die häufigsten gefundenen Organismen auf. Den Tabellen ist zu entnehmen, dass *Borrelia burgdorferi* nicht zu diesen gehört. Darüber hinaus ist erkennbar, dass oft die gleichen Organismen klassifiziert werden, jedoch variiert leicht der Anteil der zugeordneten *reads*.

Tabelle 5.2: Die fünf häufigsten gefundenen Organismen im Datensatz ERR069107_1

Organismus	**Anteil der reads**
Salmonella enterica subsp. enterica serovar Dessau	22%
Clostridium tagluense	11%
Bacillus paralicheniformis	4%
Paraglaciecola sp. L1A13	3%
Corynebacterium amycolatum	2%

Tabelle 5.3: Die fünf häufigsten gefundenen Organismen im Datensatz ERR069108_1

Organismus	**Anteil der reads**
Salmonella enterica subsp. enterica serovar Dessau	35%
Bacillus paralicheniformis	8%
Paraglaciecola sp. L1A13	5%
Corynebacterium amycolatum	2%
Pasteurella multocida subsp. multocida	2%

Tabelle 5.4: Die fünf häufigsten gefundenen Organismen im Datensatz ERR069109_1

Organismus	Anteil der reads
Salmonella enterica subsp. enterica serovar Dessau	22%
Clostridium tagluense	13%
Bacillus paralicheniformis	3%
Paraglaciecola sp. L1A13	3%
Corynebacterium amycolatum	2%

5.1.2 Kaiju

Da *kraken2* nur einen geringen Anteil der *reads* als *Borrelia burgdorferi* klassifizieren konnte, erfolgte eine weitere taxnomische Klassifikation mittels *kaiju*. Hier wurde wie in Abschnitt 5.1.1 vorgegangen.

Die durch *kaiju* erlangten Ergebnisse bestätigen, dass Ötzi nicht an Borreliose erkrankt war, da hier keiner der *reads Borrelia burgdorferi* zugeordnet werden konnte.

5.2 Mapping

5.2.1 Borrelia burgdorferi

Da wir ausschließen konnten, dass Ötzi an Borreliose erkrankt war, ergab sich keine Notwendigkeit eines *Mappings* der *reads* gegen das durch Keller *et al.* [2] publizierte *Borrelia burgdorferi* Referenzgenom. Dies hatte zur Folge, dass kein *Borrelia burgdorferi*-(Draft)-Genom rekonstruiert werden konnte.

5.2.2 mtDNA

Alle *reads* der drei Datensätze aus Kapitel 2 wurden gegen das Referenzgenom der mtDNA gemappt. Die Datensätze umfassen insgesamt 65.7674.318 *reads*. 1.253.440 dieser *reads* konnten gegen das humane mitochondriale Referenzgenom gemappt werden. Die Anteile der gemappten *reads* der jeweiligen Datensätze sind in der nachfolgenden Tabelle 5.5 aufgeführt.

[2] Keller et al, New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing, 2012

Tabelle 5.5: Anzahl der gemappten *reads* gegen das humane mitochondriale Referenzgeneom

Datensatz	Anzahl gemappter reads
ERR069107_1	366687 (0.18%)
ERR069108_1	578375 (0.22%)
ERR069109_1	308378 (0.16%)

5.3 Consensus Calling

Mittels der gemappten *reads* konnte das mitochondriale Genom von Ötzi rekonstruiert werden. Das rekonstruierte Genom weist eine Größe von 16568 Basenpaaren auf. Somit weicht das rekonstruierte mitochondriale Genom um ein Basenpaar vom Referenzgenom, das 16596 Basenpaare groß ist, ab.

5.4 Annotation

Die Annotation der mtDNA von Ötzi ergab, dass dieses Genom 33 *coding sequences* enthält. Ein Vergleich dieses Ergebnisses mit der Annotation der Referenz mtDNA aus der *NCBI* ergab, dass weniger *coding sequences* als bei Referenz, welche 37 Gene enthält, bestimmt werden konnten.

5.5 Phylogenetische Chrakterisierung

Die Analyse der Haplogruppe der mtDNA von Ötzi durch *haplogrep* ergab, dass Ötzi der Haplogruppe *K1f* zugeordnet werden konnte. Der phylogenetische Baum (siehe Abbildung 5.1) zur Bestimmung von Ötzis Herkunft zeigt, dass dieser aus der Alpenregion stammt.

Die rekonstruierte mtDNA ist im phylogenetischen Baum mit der Kennung *NC_012920.1* gekennzeichnet. Die nächsten Nachbarn sind folgende Einträge aus der AmtDB:

- MX209
- MX210
- MX204
- Iceman

Die Einträge, MX209, MX210 und MX204 stammen aus der Schweiz, während der Eintrag *Iceman* seinen Ursprung im heutigem Norditalien hat.

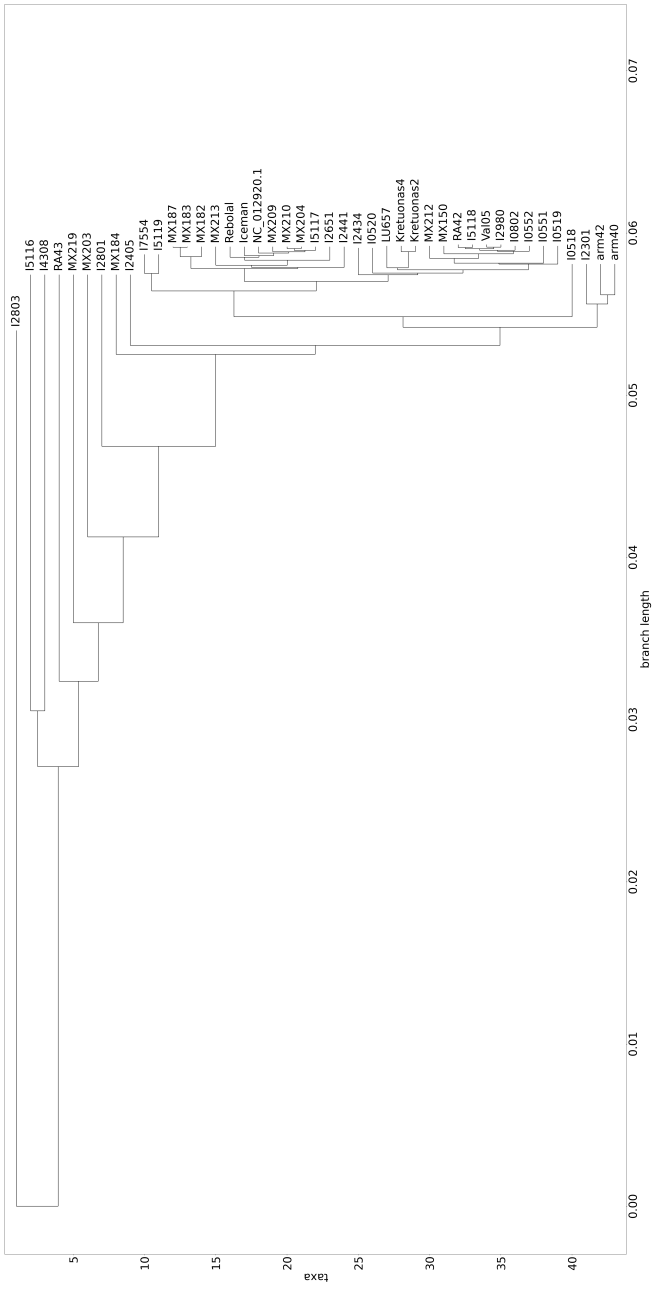

Abbildung 5.1: Viseuelle Darstellung des phylogenetischen Baum zur Herkunfsbestimmung von Ötzi

5.6 MapDamage

Zur Bestimmug der 3'/5' Substitutionsraten der mtDNA von Ötzi wurde das rekonstruierte mitochondriale Genom durch *MapDamage* analysiert.

Die oberen vier Abbildugen in Abbildung 5.2 visualisieren die Fragmentierung der sequenzierten DNA-Fragmente. Dazu ist die Frequenz der vier Nukleotide an den Positionen -10 bis +10 relativ zum 5'- und 3'-Ende der sequenzierten Reads aufgetragen.

Für Guanin ist ein gleichmäßiges Fragmentierungsmuster zu erkennen, während das Fragmentierungsmuster für die restlichen Nukleotide unregelmäßiger ist. Die größte Schwankung des Fragmentierungsmusters ist für Thymin zu beobachten. Ein unregelmäßiges Fragmentierungsmuster kann auf Schäden der Sequenzen hinweisen.

Die untere Abbildung in Abbildung 5.2 visualisiert die positionsspezifischen Substitutionen für die ersten 25 Positionen vom 5'- (links) und vom 3'-Ende (rechts).

Der rote Verlauf stellt dabei die Cytosin zu Thymin Substitution dar, die zum Ende der Sequenzen häufiger wird, während die Guanin zu Adenin Substitution, die zu Beginn der Sequenzen öfter auftritt, durch den blauen Verlauf veranschaulicht wird.

Die Deletionen relativ zum Referenzgenom sind durch den grünen Verlauf gekennzeichnet. Währenddessen repräsentiert der violette Verlauf entsprechende Insertionen im Vergleich zur Referenz. Alle Substitutionen sind durch die graue Linie dargestellt.

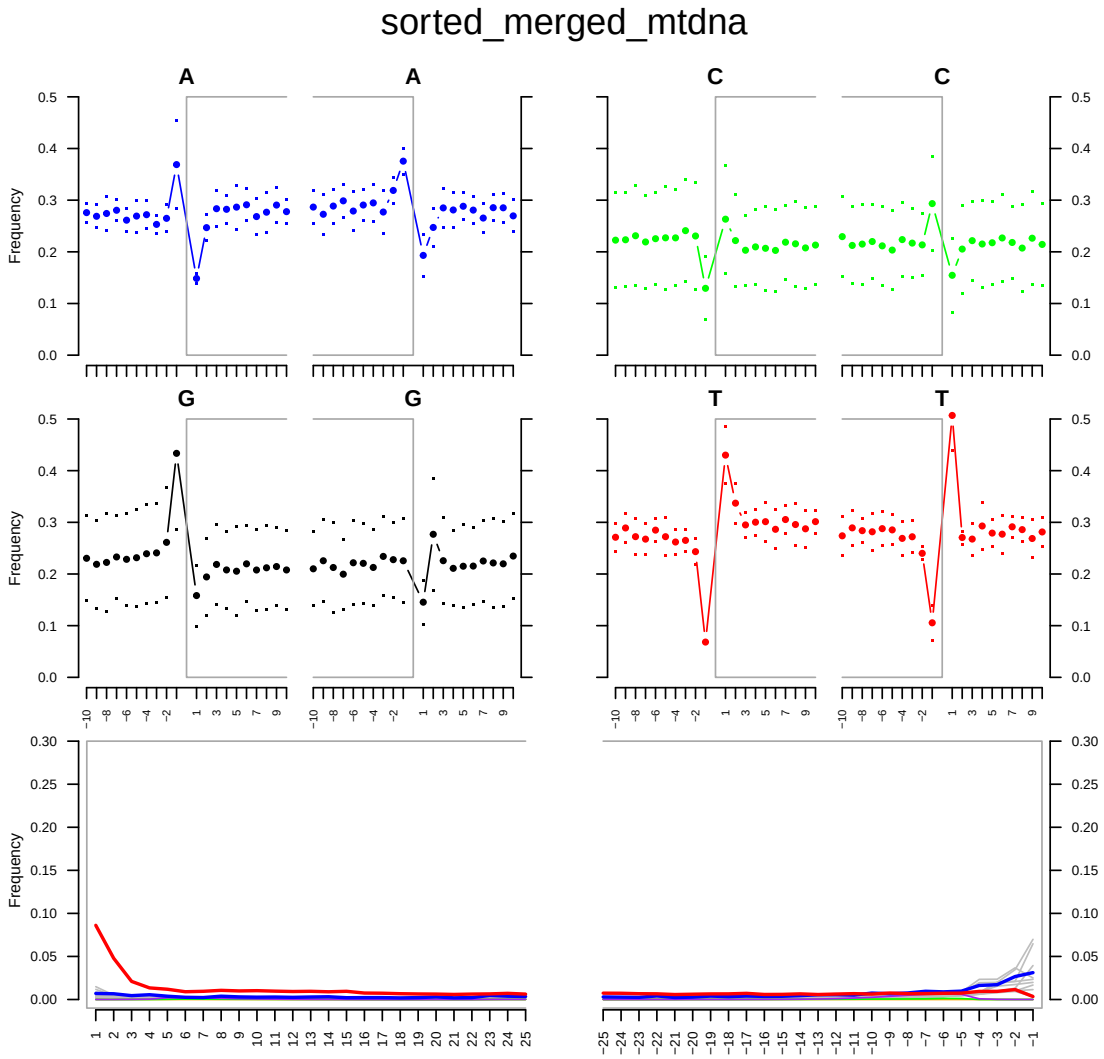

Abbildung 5.2: Visualisieung der Substitutionsraten der mtDNA von Ötzi plot

Abbidung 5.3 enthält vier verschiedene Grafiken. Die oberen zwei Abbildungen visualisieren die Längen der DNA-Fragmente in Form eines Histogramms. Die linke Abbildung zeigt hierbei die Verteilung aller Sequenzen an, während in der rechten Abbildung die Verteilungen für die einzelnen Stränge seperat dargestellt werden.

Die unteren Abbildungen repräsentieren zum einem die kumulative Häufigkeit von Basenveränderungen, bei denen ein Cytosin in Thymin substituiert wird, an den 5'-Enden der DNA-Fragmente, sowie die kumulative Frequenzen von G->A bei 3'-Enden, bei dem ein Guanin in Adenin umgewandelt wird.

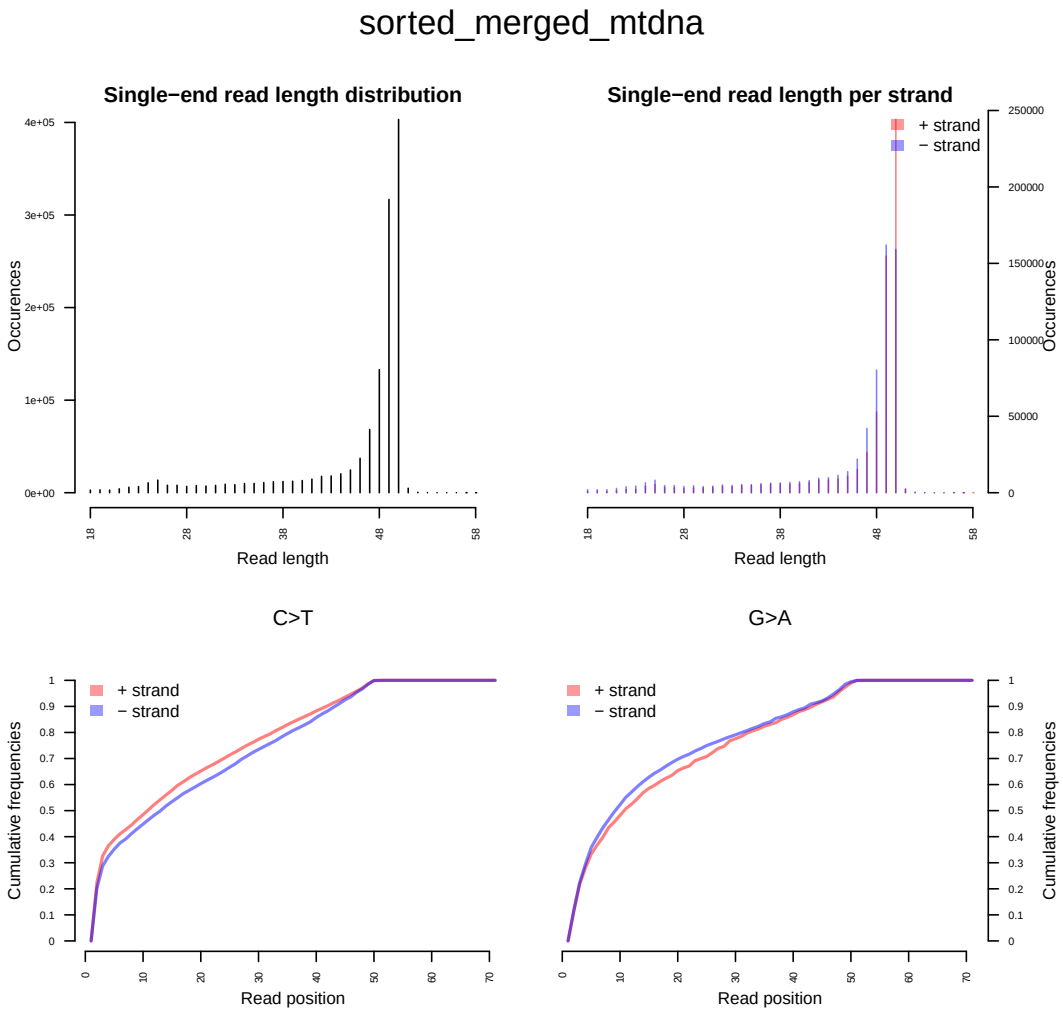

Abbildung 5.3: Längenverteilung der *singletons*

6 Technischer Report

Der technische Report beihaltet die Laufzeitstatistik der einzelnen *rules* des Workflows. Abbilung 6.1 stellt für jede *rule* die jeweilige Laufzeitstatistik dar. Auf der X-Achse ist die Laufzeit in Sekunden aufgetragen, während auf der Y-Achse die einzelnen *rules* aufgelistet sind.

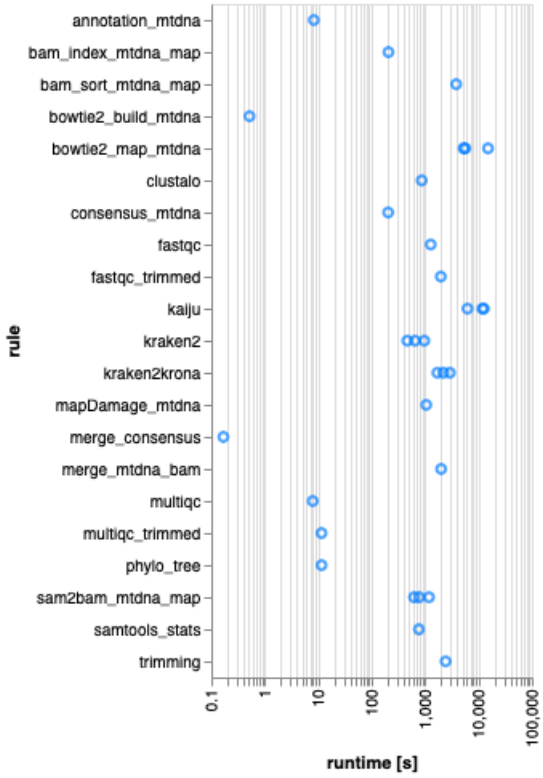

Abbildung 6.1: Laufzeitstatistik des *Workflows*