Genetik für Studierende der Bioinformatik

Freie Universität Berlin, Institut für Informatik
Peter Robinson
Wintersemester 2014/2015

5. Übungsblatt
Diskussion am 11. Dezember 2014

In dieser Übung werden Sie mit Exom-Daten arbeiten.

Wir werden eine App names fastQC verwenden, um einen typischen Exom-Datensatz zu visualisieren und die Qualität der Daten zu überprüfen.

FASTQC kann von Babraham Bioinformatics heruntergeladen werden.[1] Zur Installation muss das Archiv entpackt werden. Das Program kann dann von der Kommandozeile gestartet werden

```
$ ./fastqc
```

Laden Sie nun Exomdaten von diesem Artikel herunter: Glusman G et al. (2012) Low budget analysis of Direct-To-Consumer genomic testing familial data.[2]

Suchen Sie nach `Son's Exome Files` und laden Sie eine FASTQ (`1.fq`) Datei herunter. Laden Sie auch die VCF-Datei (`Son's VCF file.vcf`) herunter. Als Erstes wollen wir die durchschnittliche Basenqualität der Daten untersuchen.

Das FASTQ-Format sieht wie folgt aus:

```
@My-Illu:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Die vier Zeilen beschreiben eine einzelne Sequent ("Read")

1. Read identifier
2. sequence reported by the machine
3. '+'
4. ASCII-kodierte Qualitätswerte (PHRED)

Der PHRED Qualitätswert wird definiert als:

$$\boxed{Q_{PHRED} = -10 \log_{10} p} \tag{1}$$

wobei p die Wahrscheinlch bezeichnet, dass die angegebene Base falsch ist.

Q_{PHRED}	p	Accuracy
10	10^{-1}	90%
20	10^{-2}	99%
30	10^{-3}	99.9%
40	10^{-4}	99.99%
50	10^{-5}	99.999%

Q_{PHRED}	ASCII
2	B: Special indicator: Trim off rest of read
10	L
20	V
30	‘
40	j
...	...

Tabelle 1: ASCII Codes for Phred Scores

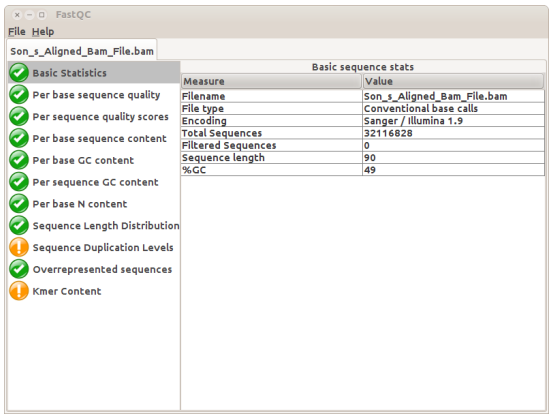

Abbildung 1: fastQC

Die PHRED Qualitätswerte werden zu ASCII Zeichen konvertiert, indem man die folgende Transformation anwendet: Q_{PHRED} +33 (ASCII 0–62).

Aufgabe 1.

Wie viele Reads sind in 1.fq enthalten? (Hinweis, überlegen, welche Shellprogramme geeignet sind...)

Aufgabe 2.

Unter der Annahme, dass Q Werte zwischen 0 undf 41 annimmt, berechnen die den Wertebereich für die Fehlerwahrscheinlichkeiten der Basen? Nehmen Sie eine Sequenz aus der fastq-Datei und berechnen Sie die Qualitätswerte für alle Basen mit einem Perl oder Python-Skript.

Aufgabe 3.

Starten Sie nun das FASTQC Programm und analysieren Sie damit die heruntergeladene FASTQ-Datei. Konsultieren Sie falls nötig die Online-Hilfe von FASTQC.

Die Analyse wird ein paar Minuten dauern und ein Ergebnis wie in Abb. 1 dargestellt erzeugen.

[1] `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

[2] *F1000Research*, **1**:3 (`http://f1000research.com/articles/1-3`).

To get started with FastQC, answer the following questions.

Aufgabe 4.

Wie viele Reads enthält die Datei?

Aufgabe 5.

Wie lang sind die Reads? Wieviel beträgt der beste durchschnittliche Qualitätswert für eine einzelne Basenposition? Was ist die entsprechende Fehlerwahrscheinlichkeit?

Aufgabe 6.

Laden Sie die Applikation Jannovar herunter (http://charite.github.io/jannovar/). Die Applikation wird mit maven erzeugt. Falls Sie maven nicht installiert haben, können SIe die entsprechende Jar-Datei von unserer Homepage herunterladen: (http://compbio.charite.de/contao/index.php/jannovar.jar). Analysieren Sie nun die heruntergeladene VCF-Datei

```
$ java -jar Jannovar.jar Son.vcf
```

Analysieren Sie die Verteilung der Varianten in der Ausgabe-Datei. Wie viele Missense-Mutationen gibt es? Nonsense? Spleißmutation?