Adapter Considerations

1. **Wie lang sind die Fragements/ inserts, die wir sequenzieren**

Hierfür hab ich mir von NgoG4_Azi_15min_S1_1.fastq den ersten read angeschaut und mit NgoG4_Azi_15min_S1_2.fastq verglichen:

Frage, wie lang sind die Fragemente, die ausgelesen werden?
Hierzu aus NgoG4_Azi_15min_S1_1.fastq & NgoG4_Azi_15min_S1_2.fastq

aus _1.fastq (@A00685:115:HYG3TDSXY:4:1101:2230:1000 1:N:0:CCATCCGC+TACGCCTT)
NCCACCTGTGTCGGTTTGCGGTACGGTTCGATTCAAACTGAAGCTTAGTGGCTTTTCCTGGAAGCGT GGTATCGGTTGCTTCGTGTCCGTAGACACTCGTC

aus _2.fastq (@A00685:115:HYG3TDSXY:4:1101:2230:1000 2:N:0:CCATCCGC+TACGCCTT)
CCGGGTTTTCTTAACACCGAGAAGTGATGACGAGTGTCTACGGACACGAAGCAACCGATACCACGC TTCCAGGAAAAGCCACTAAGCTTCAGTTTGAATCG
-> das reverse Kompliment dazu machen -->_2_rc
CGATTCAAACTGAAGCTTAGTGGCTTTTCCTGGAAGCGTGGTATCGGTTGCTTCGTGTCCGTAGACA CTCGTCATCACTTCTCGGTGTTAAGAAAACCCGGt

- die orangenen Sequenzen sind gleich
- es erscheint so, dass
  - ... die reverse reads auch tatsächlich noch in ihrer reversen form angegeben (also muss man vermutlich alle adapter auch als reverse compliments angeben)
  - unsere Fragmente (also die Länge der DNA Stücke, an die Adapter vorne und hinten angeheftet werden) haben eine Länge von 130bp, weil ist 30 + 70 + 30 = 130. Es gibt eine Überlappung von 70bp, also hat jeder read noch 30bp individuelle Information

TODO: Skript schreiben, dass die Länge der Fragemente ausfindet. Oder kann man das an den .sam Dateien ablesen?

2. **Welche Adapter brauchen wir für das Trimmen?**

Für Illumina Daten von eurofins benutzen wir immer die angegebenen TruSeq3-PE-2 files:

>>>TruSeq3-PE-2.fa

>PrefixPE/1
TACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PrefixPE/2
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
>PE1
TACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PE1_rc
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA

>PE2
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
>PE2_rc
AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

>>>Email CCG
Read1:
>A1_1 rc
AGATCGGAAGAGCACACGTCTGAACTCCAGTCA .TGACTGGAGTTCAGACGTGTGCT
CTTCCGATCT
>A1_2
AGATCGGAAGAGCACACGTCTGAAC …….GTTCAGACGTGTGCTCTTCCGATCT
>A1_3
TGGAATTCTCGGGTGCCAAGG CCTTGGCACCCGAGAATTCCA
>A1_4
AGATCGGAAGAGCACACGTCT……... ……...AGACGTGTGCTCTTCCGATCT
>A1_5
CTGTCTCTTATACACATCT AGATGTGTATAAGAGACAG
>A1_6
AGATGTGTATAAGAGACAG CTGTCTCTTATACACATCT

Read2:
>A2_1
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
>A2_2
AGATCGGAAGAGCGTCGTGTAGGGA TCCCTACACGACGCTCTTCCGATCT
>A2_3
TGGAATTCTCGGGTGCCAAGG CCTTGGCACCCGAGAATTCCA
>A2_4
AGATCGGAAGAGCACACGTCT……... ……...AGACGTGTGCTCTTCCGATCT
>A2_5
CTGTCTCTTATACACATCT AGATGTGTATAAGAGACAG
>A2_6
AGATGTGTATAAGAGACAG CTGTCTCTTATACACATCT

Result:

- the grey and yellow sequences are not in the TruSeq file
- do they appear in the overrepresented sequences in Fastqc?

Orginal email von Ccg
No quality filtering or adapter trimming was applied to the FastQ files.
When you run adapter trimming, use the following sequences:
Read1:
- AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
- AGATCGGAAGAGCACACGTCTGAAC
- TGGAATTCTCGGGTGCCAAGG
- AGATCGGAAGAGCACACGTCT
- CTGTCTCTTATACACATCT
- AGATGTGTATAAGAGACAG

Read2:
- AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
- AGATCGGAAGAGCGTCGTGTAGGGA
- TGGAATTCTCGGGTGCCAAGG
- AGATCGGAAGAGCACACGTCT
- CTGTCTCTTATACACATCT
- AGATGTGTATAAGAGACAG