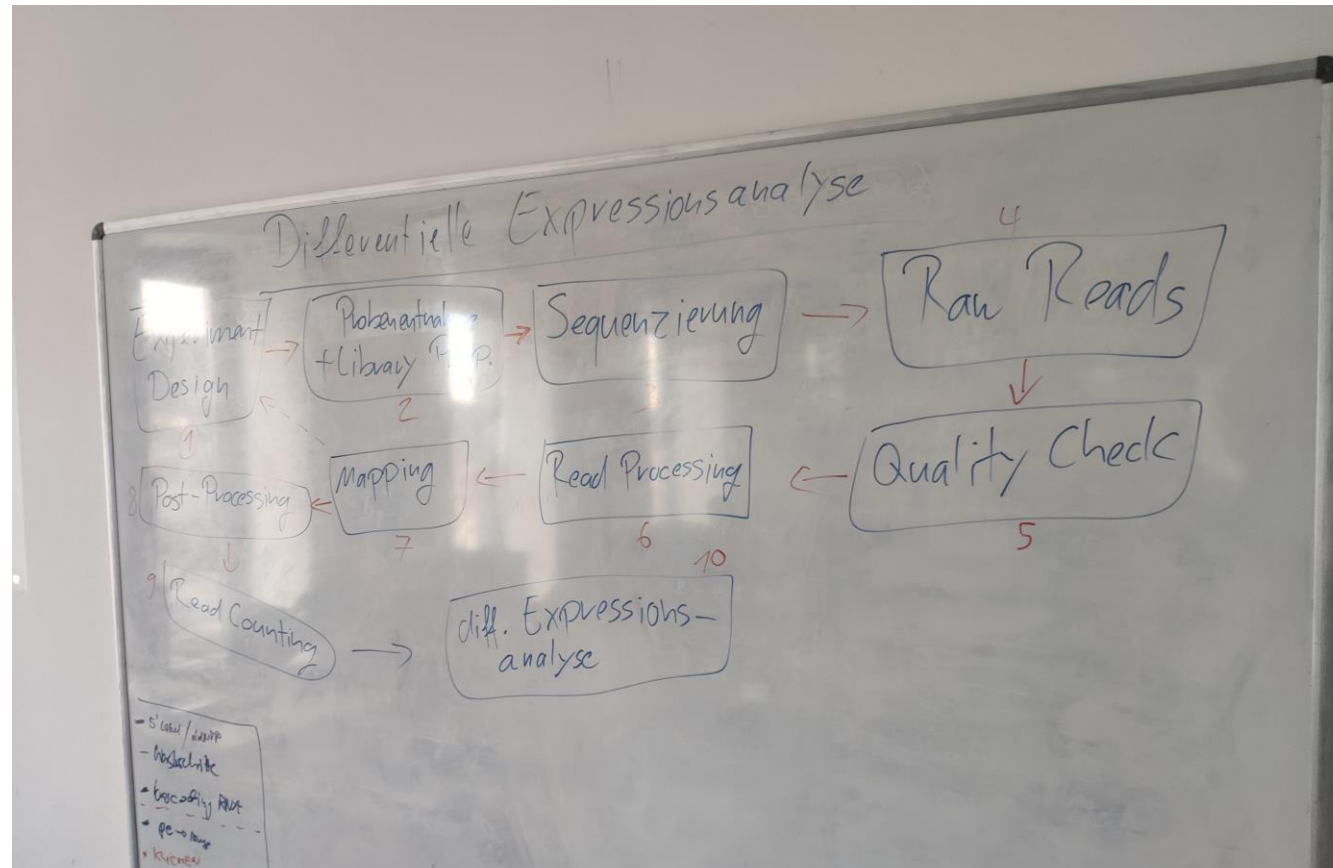
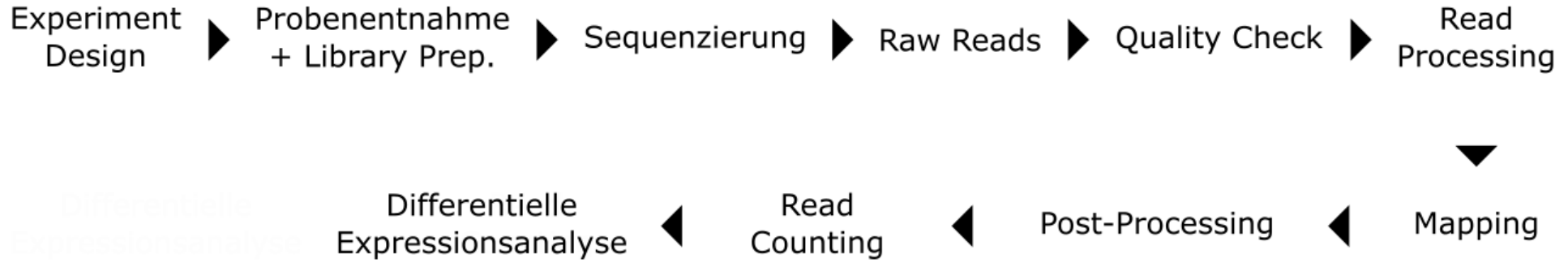


Methoden der Hochdurchsatz- sequenzierung

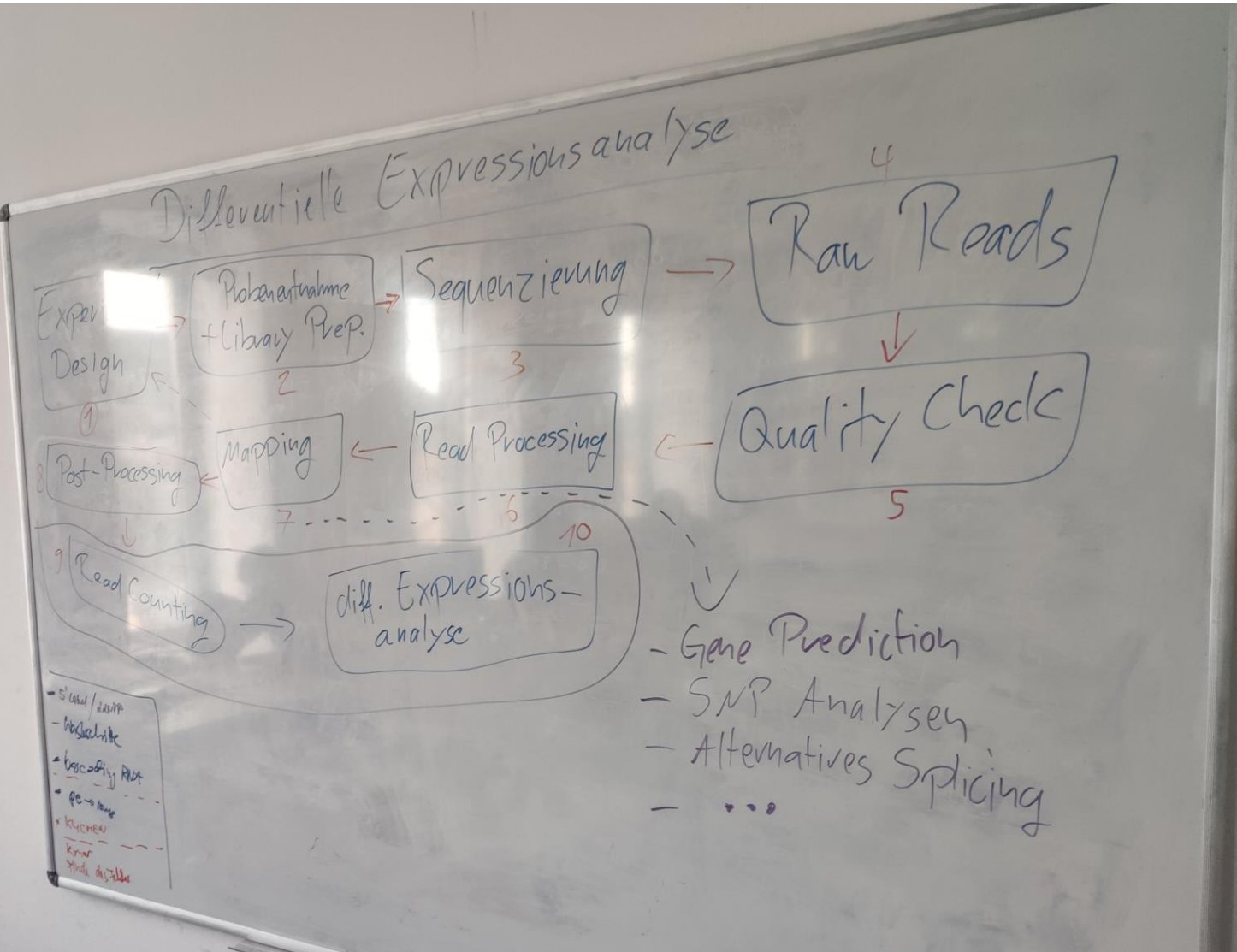
– *Die klassische Genexpressionsanalyse Pipeline* –
Part I

Die klassische Pipeline – “Das 10 Schritte Programme”



Die klassische Pipeline – *But wait... there is more!*

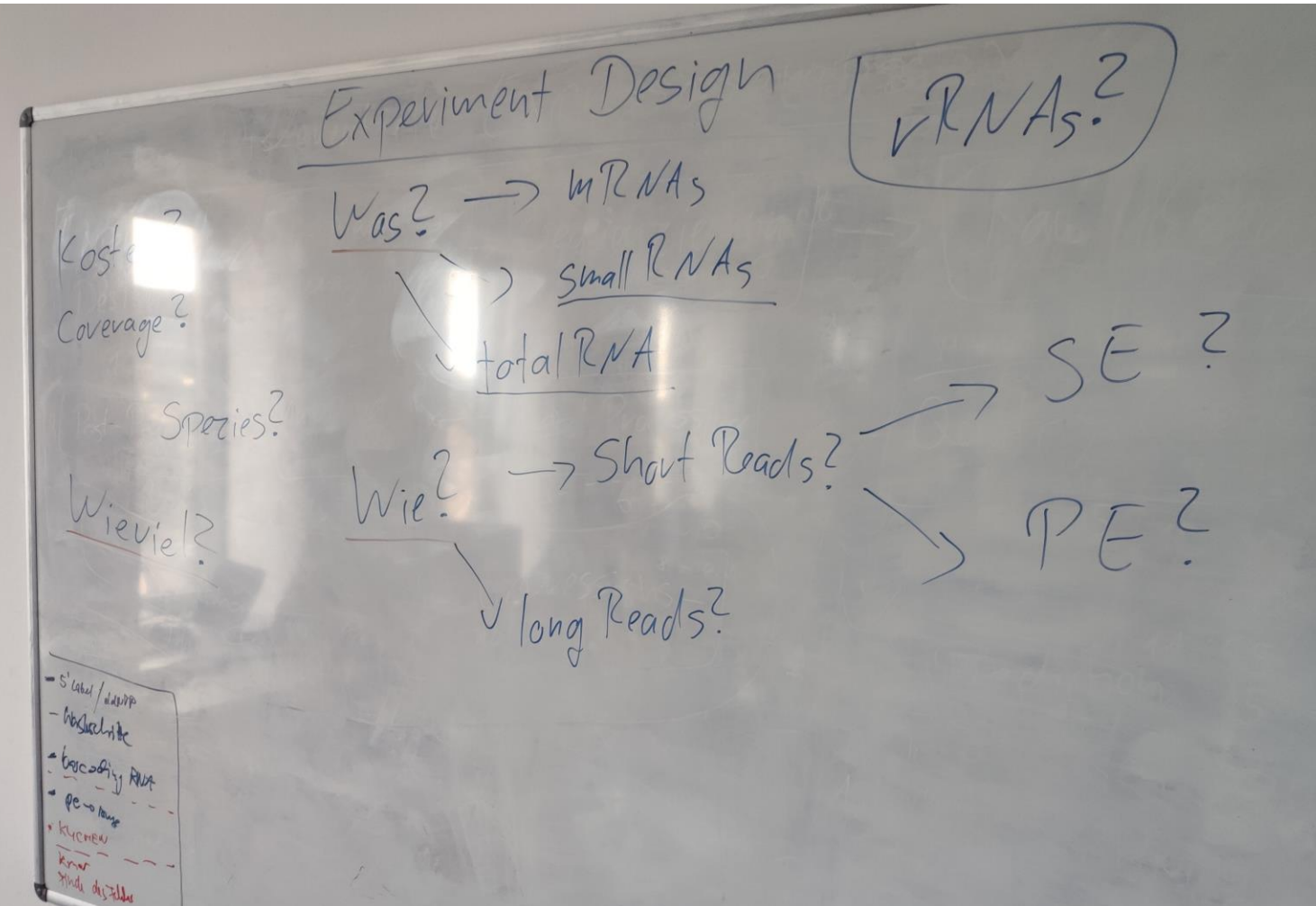
- Am Ende der klassischen Pipeline erhalten wir eine Liste mit potentiell differentiell expremierten genomischen Features.
→ Hier beginnt dann die eigentlich spannende biologische Analyse- und Interpretationsarbeit.



- Mit RNA-Seq Daten kann man aber viel mehr als nur differentielle Genexpressionsanalyse untersuchen.
- Ausgehend vom fertigen Read Mapping (Schritt 7) bzw. eventuellen Mapping Post-Processing (Schritt 8), lassen sich eine Menge weiterer Analysen durchführen, bspw.:
 - Gene prediction/annotation
 - SNP und GWAS Analysen
 - Alternative Splicing + Isoform Analysen
 - Promotor Activity
 - Chromatin profiling
 - Conditional profiling
 - ...

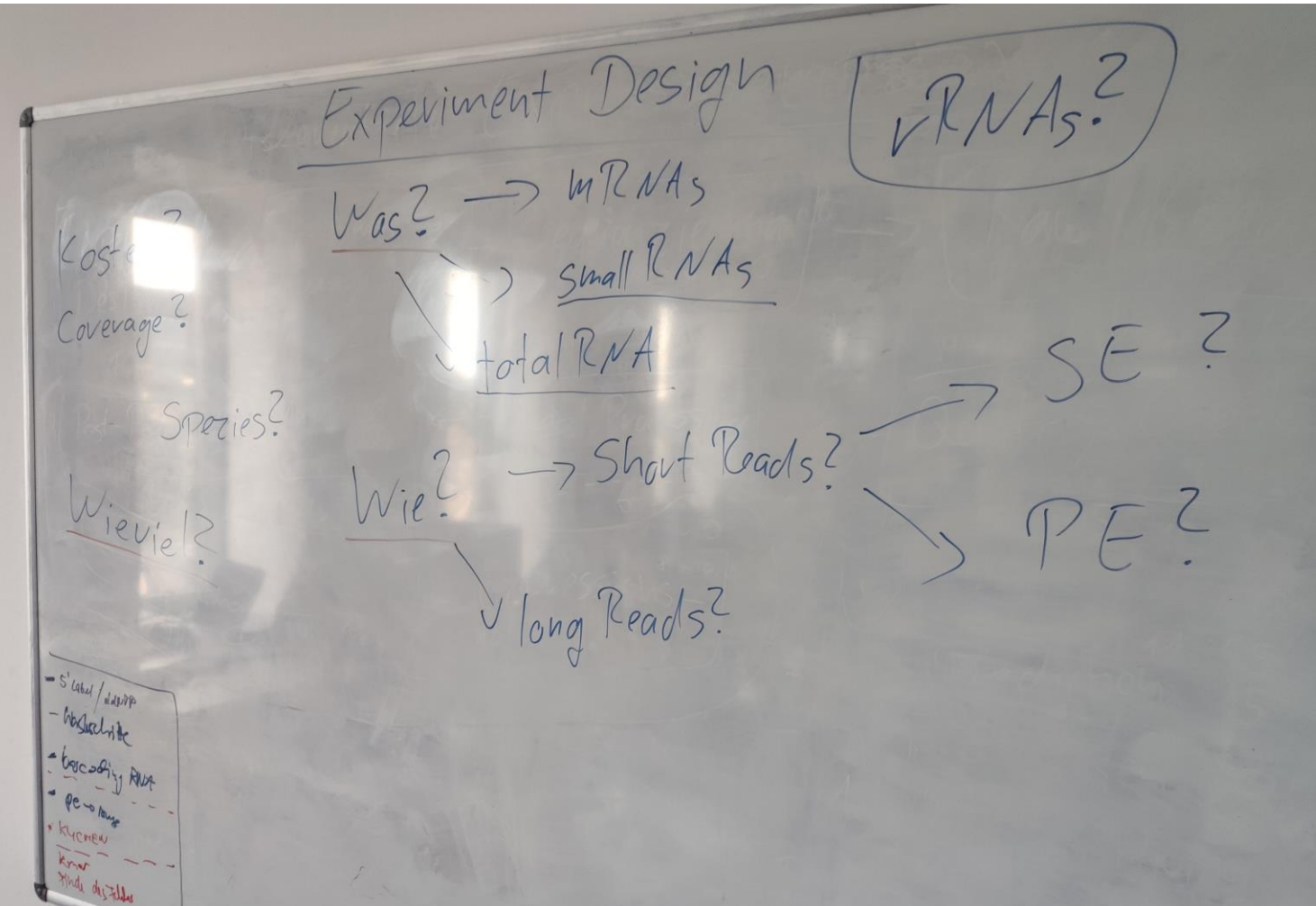
Schritt 1 – Experiment Design

- Wichtige Fragen die genau geklärt sein sollten, bevor auch nur die erste Probe zum Sequenzieren gewonnen wird:
- **Was** genau wollen wir untersuchen? **Was** wollen wir sequenzieren? (polyA-Seq, total RNA-Seq, small RNA-Seq, Clip-Seq, ...)
- **Wie** genau wollen wir sequenzieren? (short Read oder long Read? single-end oder paired-end? single cell oder pooled?)
- **Wieviel** genau wollen wir sequenzieren? (mind.(!) 5 biologische Samples pro Condition, sonst kann man es gleich sein lassen)



Schritt 1 – Experiment Design

- Wichtige Fragen die genau geklärt sein sollten, bevor auch nur die erste Probe zum Sequenzieren gewonnen wird:
- **Was** genau wollen wir untersuchen? **Was** wollen wir sequenzieren? (polyA-Seq, total RNA-Seq, small RNA-Seq, Clip-Seq, ...)
- **Wie** genau wollen wir sequenzieren? (short Read oder long Read? single-end oder paired-end? single cell oder pooled?)
- **Wieviel** genau wollen wir sequenzieren? (mind.(!) 5 biologische Samples pro Condition, sonst kann man es gleich sein lassen)

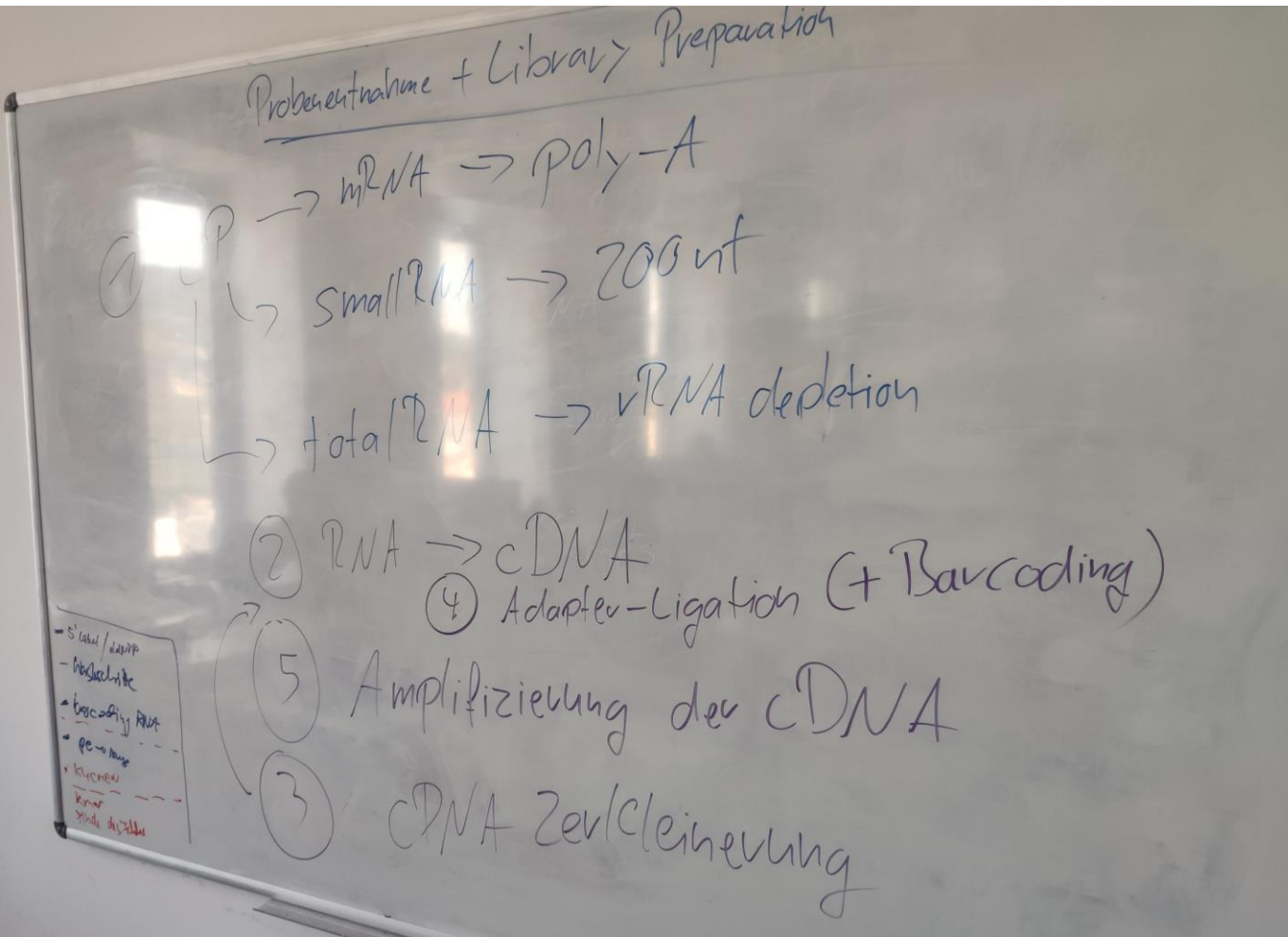


- Weitere wichtige Fragen:
 - Wieviel soll das Ganze kosten?
 - Wie tief wollen wir sequenzieren? (Coverage)
 - Gibt es ein (gutes!) Referenzgenom und –annotation für unsere Spezies?

Schritt 2 – Probenentnahme + Library Preparation

- Dies hier ist nur eine gaaaaanz grobe Darstellung, genaueres möge man in der [Primärliteratur](#) nachlesen.

0. Probenentnahme (z. B. Gewebeproben, Zellkultur, etc.), anschließend Zellen lysieren und RNA Moleküle isolieren.
1. RNA Moleküle eventuell filtern (rRNAs abreichern oder nur mRNAs am poly(A) rausziehen oder nur small RNAs rausholen).
2. RNA Moleküle in cDNA umschreiben.



3. cDNA in kleinere Fragmente spalten (z. B. mittels Enzyme oder per Ultraschall).

4. Adapter-Ligation (und eventuell Barcoding) an die 5'- und 3'-Enden aller cDNA Fragmente.

5. Amplifizierung der cDNA Fragmente.

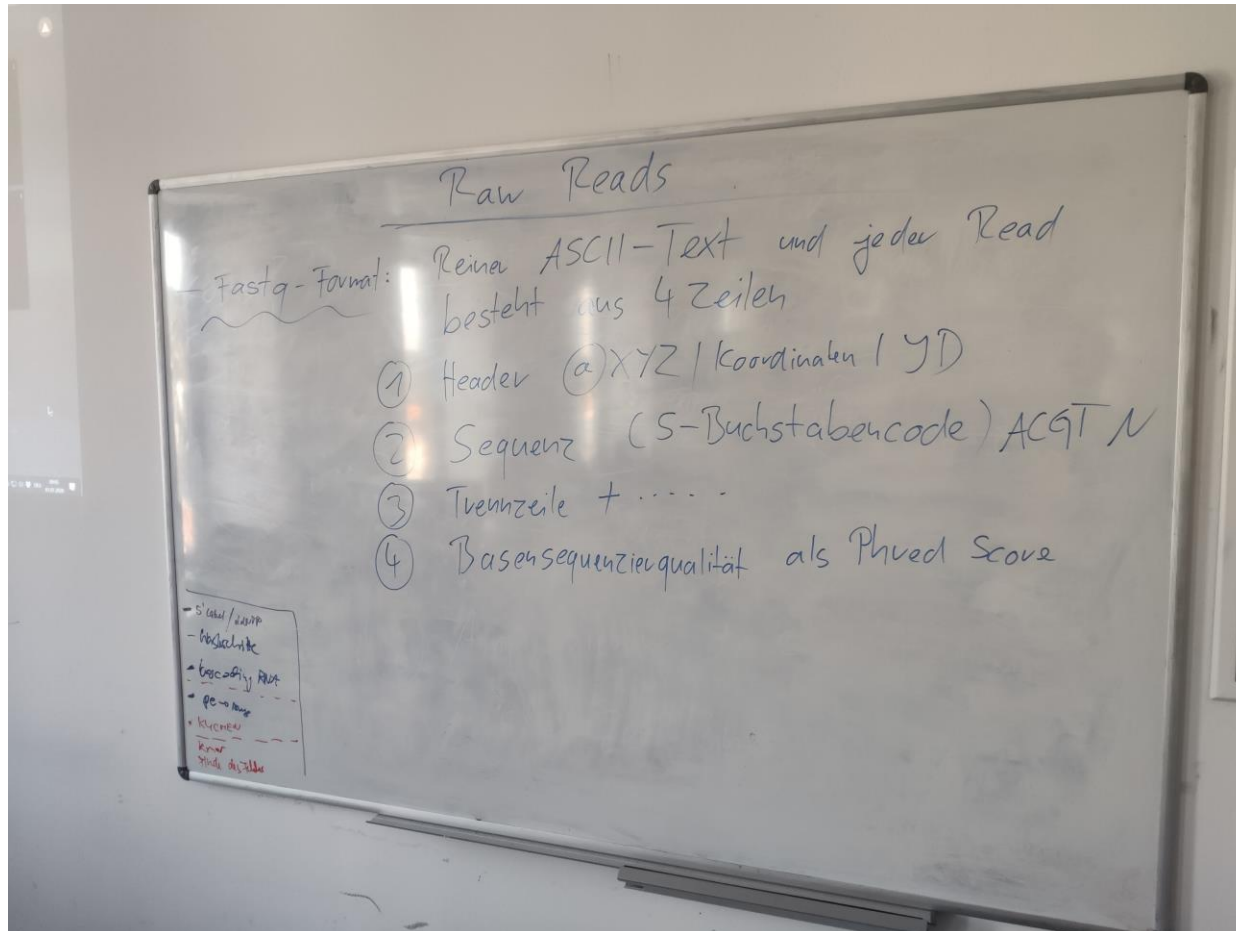
6. Sequenzierung.

Schritt 3 – Sequenzierung

Siehe vorangegangene Vorlesungsstunden.

Schritt 4 – Raw Reads

- Erster Check: Entspricht die Anzahl der sequenzierten Reads den Erwartungen? (Entspricht der Durchsatz den Erwartungen? → #Reads * Readlänge)
- Das Fastq-Format:
 - Reiner ASCII-Text, wobei jeder Read aus 4 Zeilen besteht:

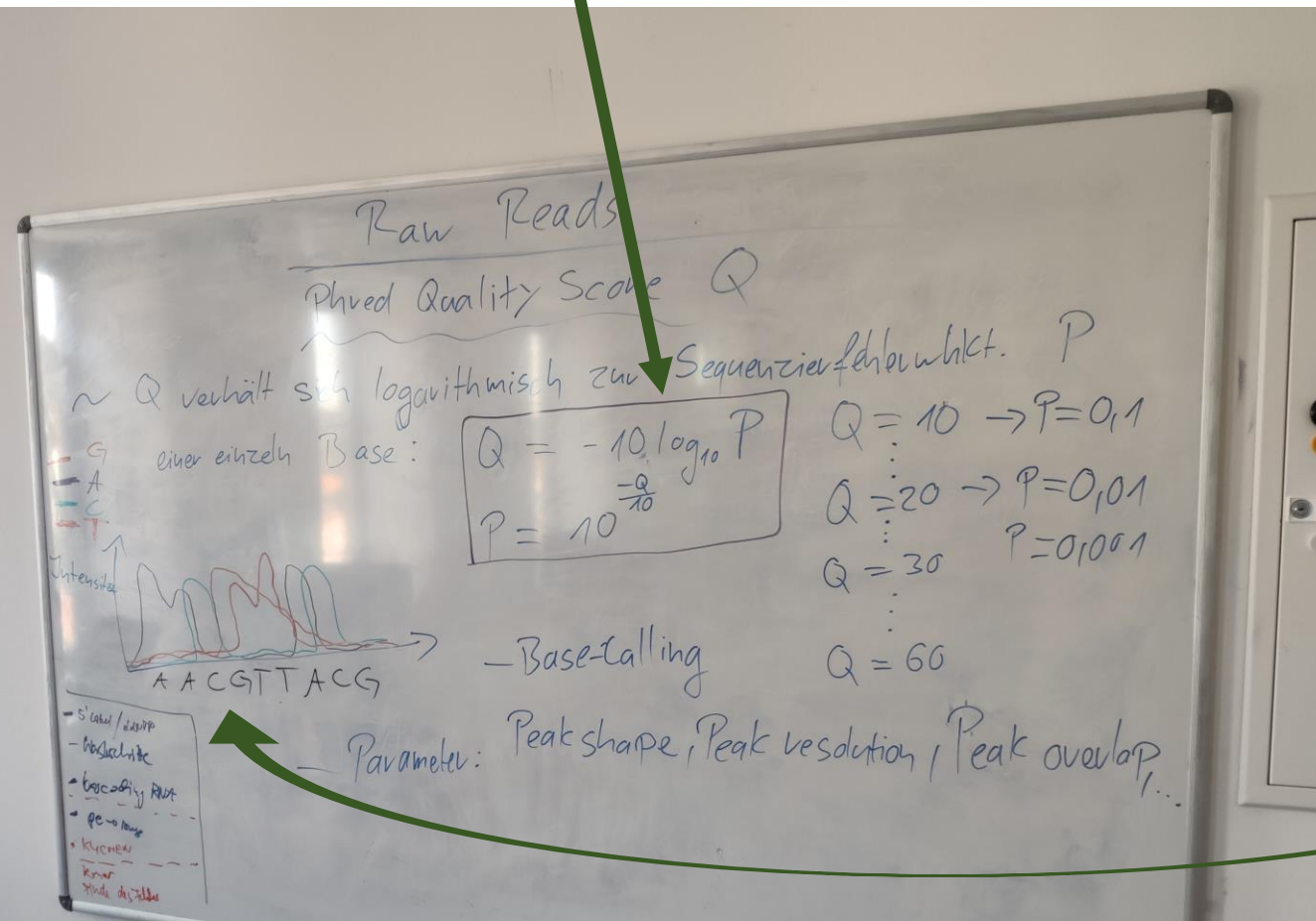


Beispiel:

```
1 @HISEQ2500:386:C65KJACXX:6:1101:1547:2249 1:N:0: TGACCA
2 CGGGTGT TTGAGAAAGAAGAGGAAGATGAGGATGAAGACGAGGAGGAGGA
3 +
4 ???D=ADDDHHHFDGGE>EGHGGHG;GCGGEFIEIIGFGGGGI@FHFGA;
5 @HISEQ2500:386:C65KJACXX:6:1101:2459:2246 1:N:0: TGACCA
6 CACCNTCCGCAGTGAAGTCCACATGACAGTCCTTCTTCCACTAACC
7 +
8 @;@D#2ABF=CAFBEAHHGIIIGCGGGIIIGGGIDHGGGIGGG<AEF
9 @HISEQ2500:386:C65KJACXX:6:1101:2764:2249 1:N:0: TGACCA
10 CTAGCCCTTCCCTAGGCCTGAAGTTAGAAGGGGAGGCTGAGGACAAAGA
11 +
12 @@@FAEDDFDDHHIGBFFHBHF:CFEA?EEEEGHGEHGAH==F?)=C38
13 @HISEQ2500:386:C65KJACXX:6:1101:3833:2245 1:N:0: TGACCA
14 TCTANACTCTTACCCAGCCACCTACCCTACCCTGCCCTGCCCATCTG
15 +
16 CCCF#2ADHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
17 @HISEQ2500:386:C65KJACXX:6:1101:4752:2249 1:N:0: TGACCA
18 CTGCATTGGCCCCAGGAGTCCCCACAGCTTTAGAGGTCATAAGTGCTGCT
19 +
20 CCCCCFFFGFHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

Schritt 4 – Raw Reads

- Phred Quality Score Q:
 - Q verhält sich logarithmisch zur Sequenzierfehlerwahrscheinlichkeit P **einer einzelnen** Base.
 - Wichtige Formel.

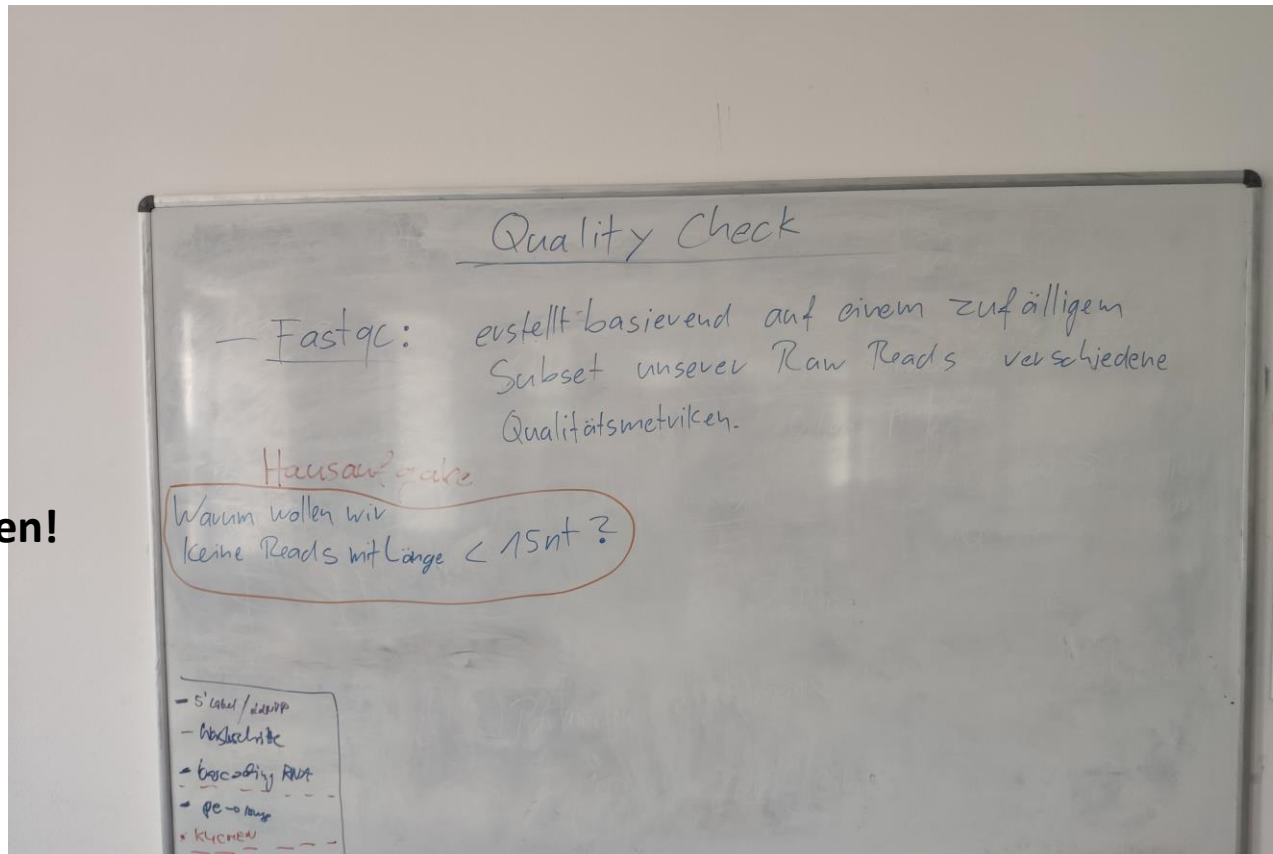


- Phrad ist eigentlich ein *Base-Caller*, d. h. eine Software, welche die Lichtsignale (fluorescence intensity peaks) die beim Sequenzieren der cDNA Fragmente aufgenommen werden, interpretiert und jeweils einer von 5 Basen zuordnet (ACGTN).
- Nach den Werten bestimmter Parameter (wie Form des Peaks, Peakauflösung, Peak-Überlappung, etc.) wird jeder sequenzierten Base ein Phred Score Q zugewiesen, mithilfe dessen sich dann die Sequenzierfehlerwahrscheinlichkeit P für diese Base berechnen lässt.

• (Für modernere Sequenzierverfahren gibt es mittlerweile andere Scores.)

Schritt 5 – Read Quality Check

- Beliebtes Tool: FASTQC
 - Analysiert eine zufällige Teilmenge an Reads eines Fastq-Files und erstellt aus deren Sequenzen und Quality-Scores verschiedene statistische Metriken die einem helfen die Gesamtsequenzierqualität des Fastq-Files abzuschätzen.
 - Wurde eigentlich für Whole-Genome Sequenzierprojekte geschrieben, deswegen sind ein paar der Metriken für RNA-Seq Daten irrelevant und können ignoriert werden.



Hausaufgabe nicht vergessen!

Schritt 5 – Read Quality Check – FASTQC Beispiele

FastQC Report

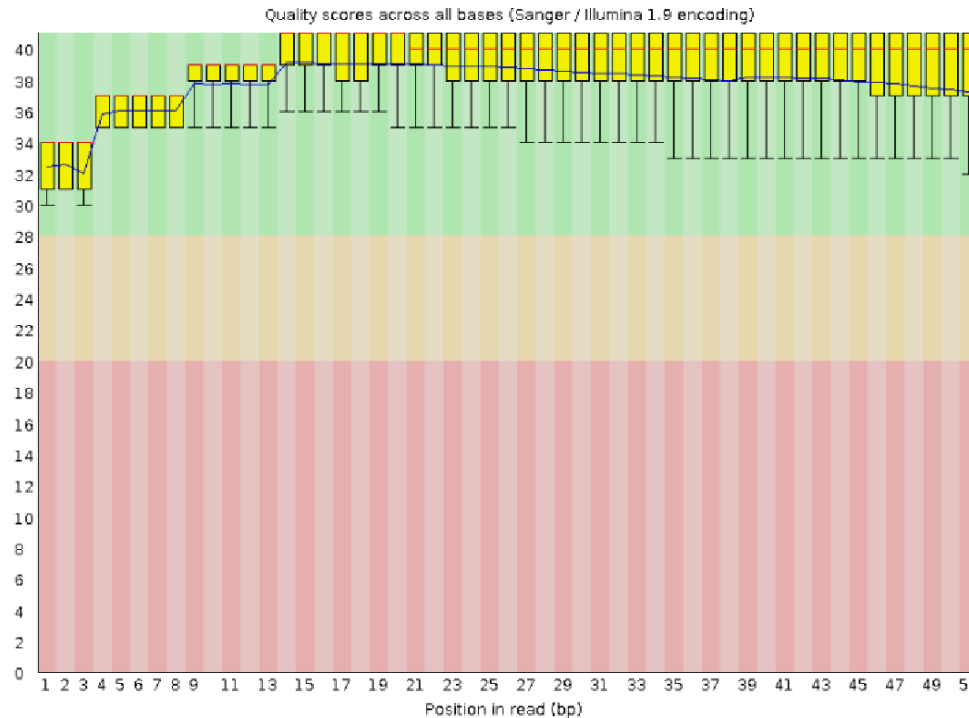
Summary

- ✔ [Basic Statistics](#)
- ✔ [Per base sequence quality](#)
- ⚠ [Per tile sequence quality](#)
- ✔ [Per sequence quality scores](#)
- ✘ [Per base sequence content](#)
- ✔ [Per sequence GC content](#)
- ✔ [Per base N content](#)
- ✔ [Sequence Length Distribution](#)
- ✘ [Sequence Duplication Levels](#)
- ✔ [Overrepresented sequences](#)
- ✔ [Adapter Content](#)
- ✘ [Kmer Content](#)

✔ Basic Statistics

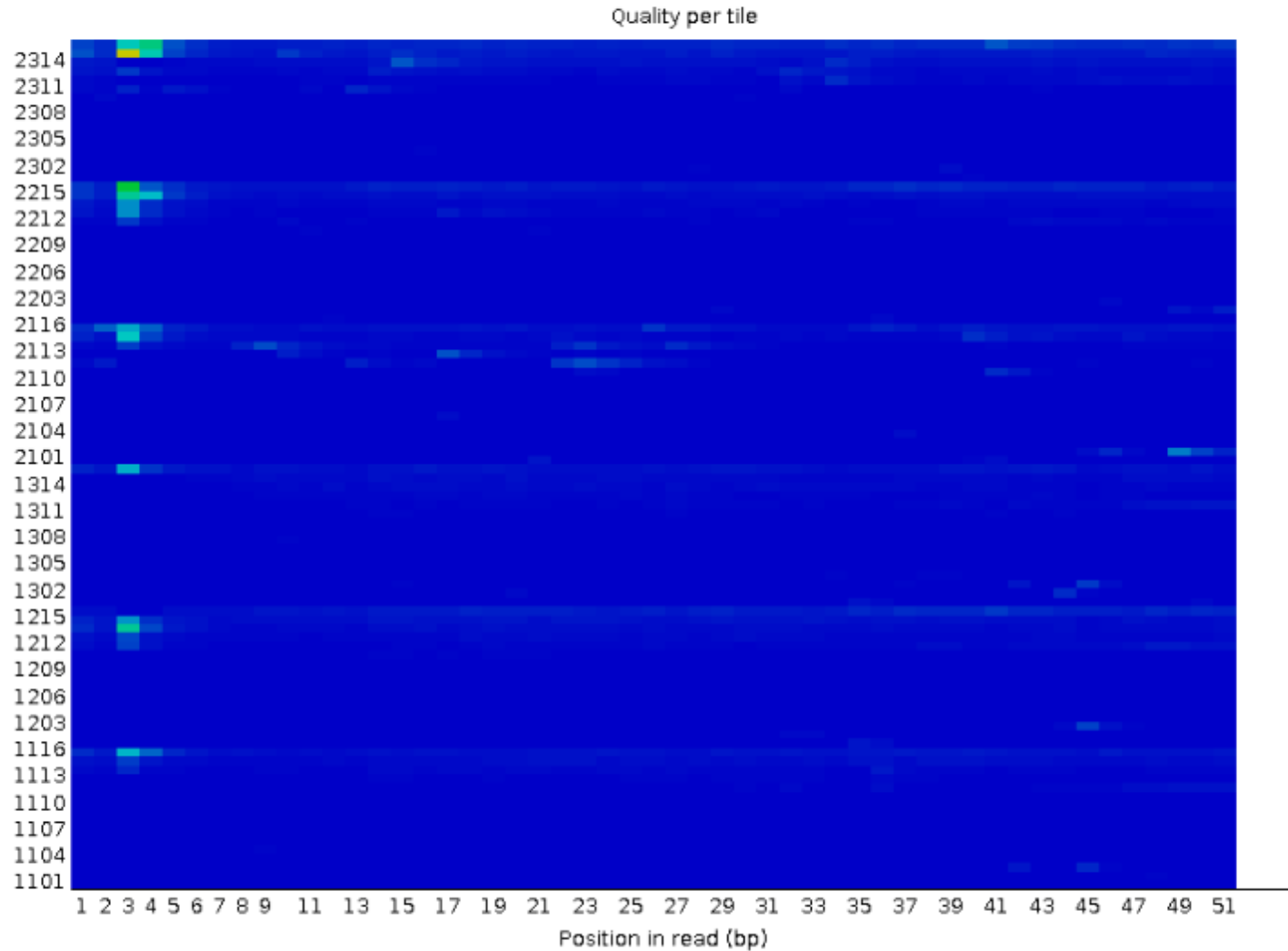
Measure	Value
Filename	15_240214_1HS_AGTCAG_L006_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	50456642
Sequences flagged as poor quality	0
Sequence length	51
%GC	48

✔ Per base sequence quality



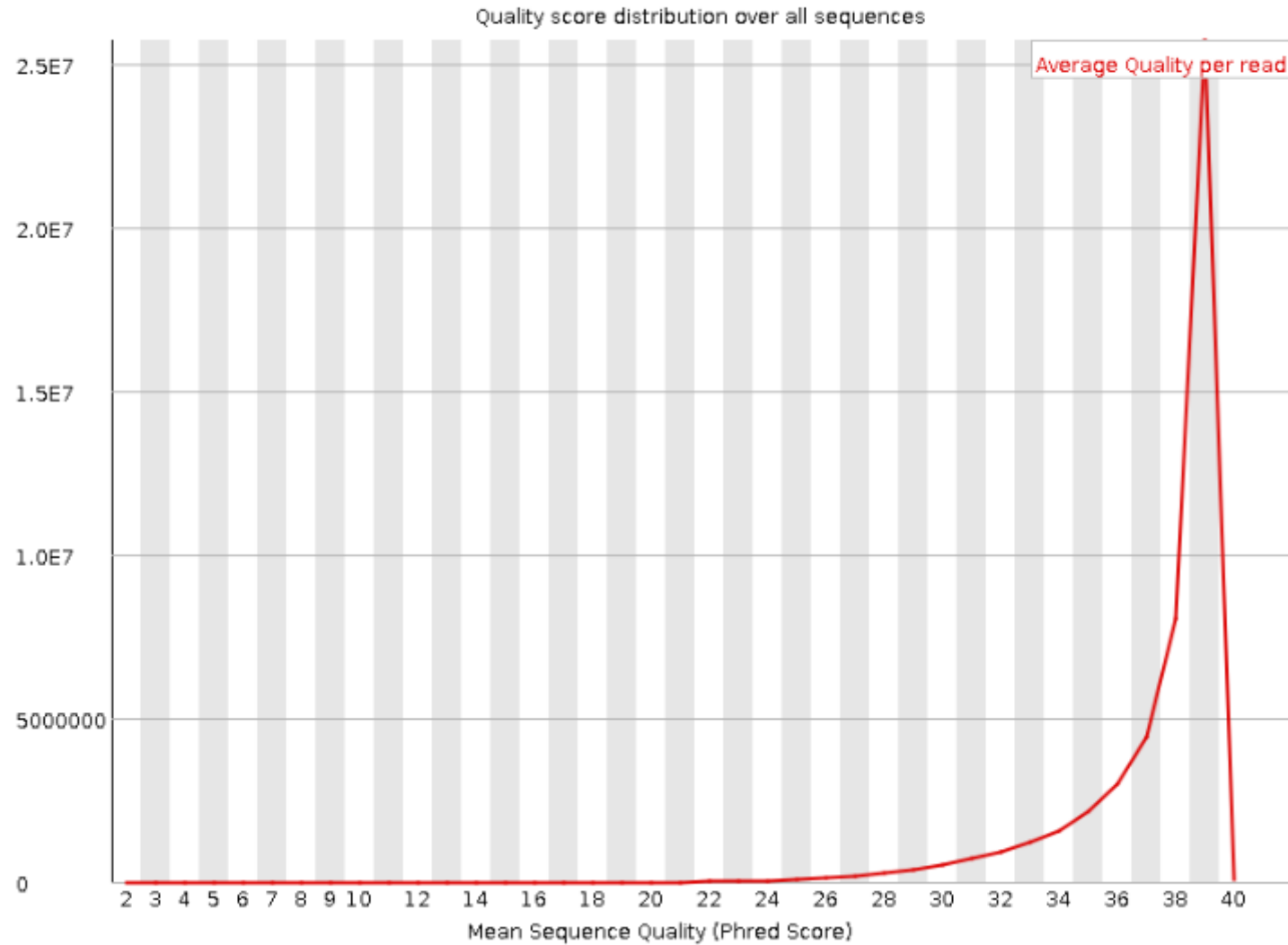
Schritt 5 – Read Quality Check – FASTQC Beispiele

! Per tile sequence quality



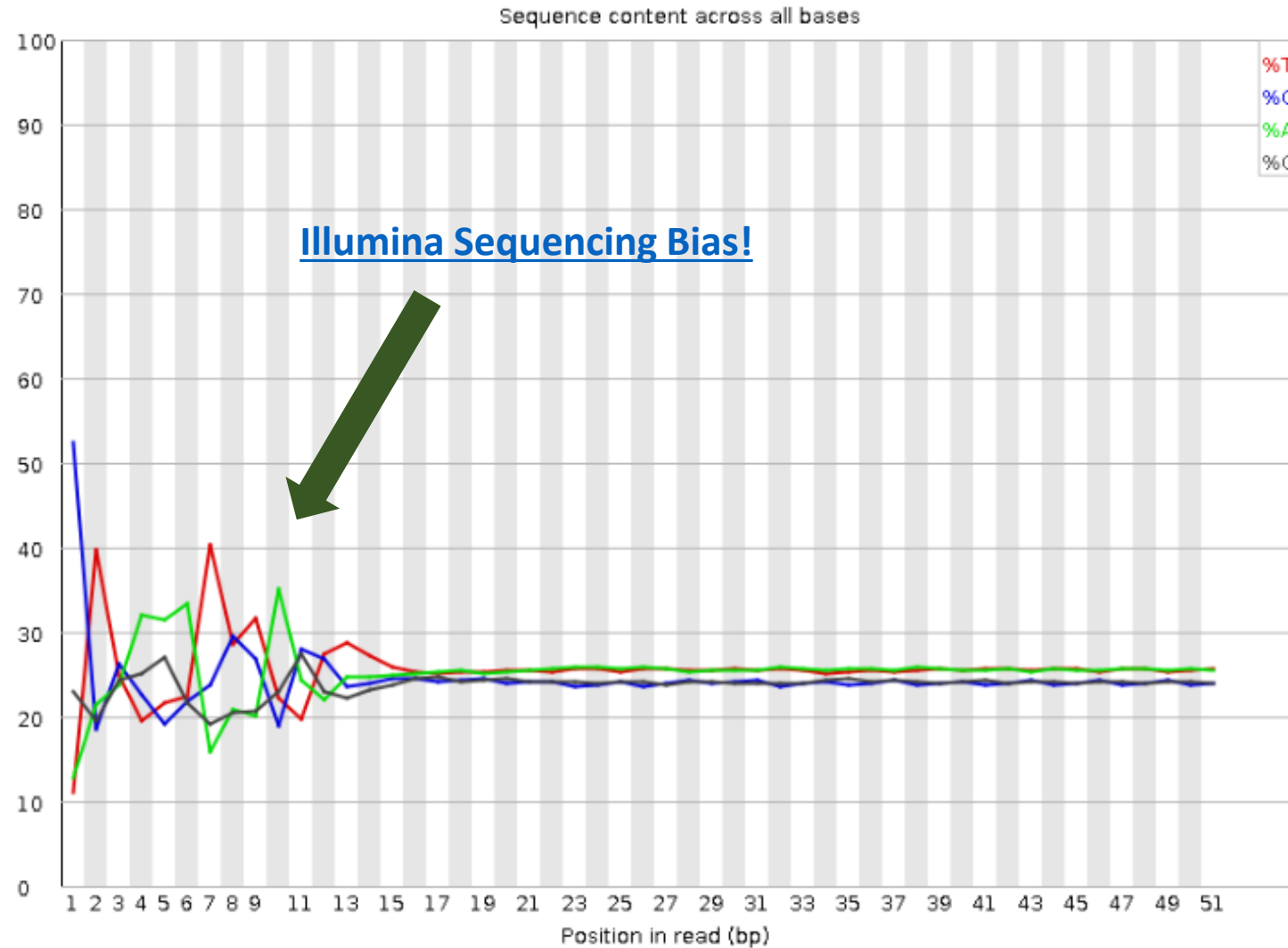
Schritt 5 – Read Quality Check – FASTQC Beispiele

✔ Per sequence quality scores



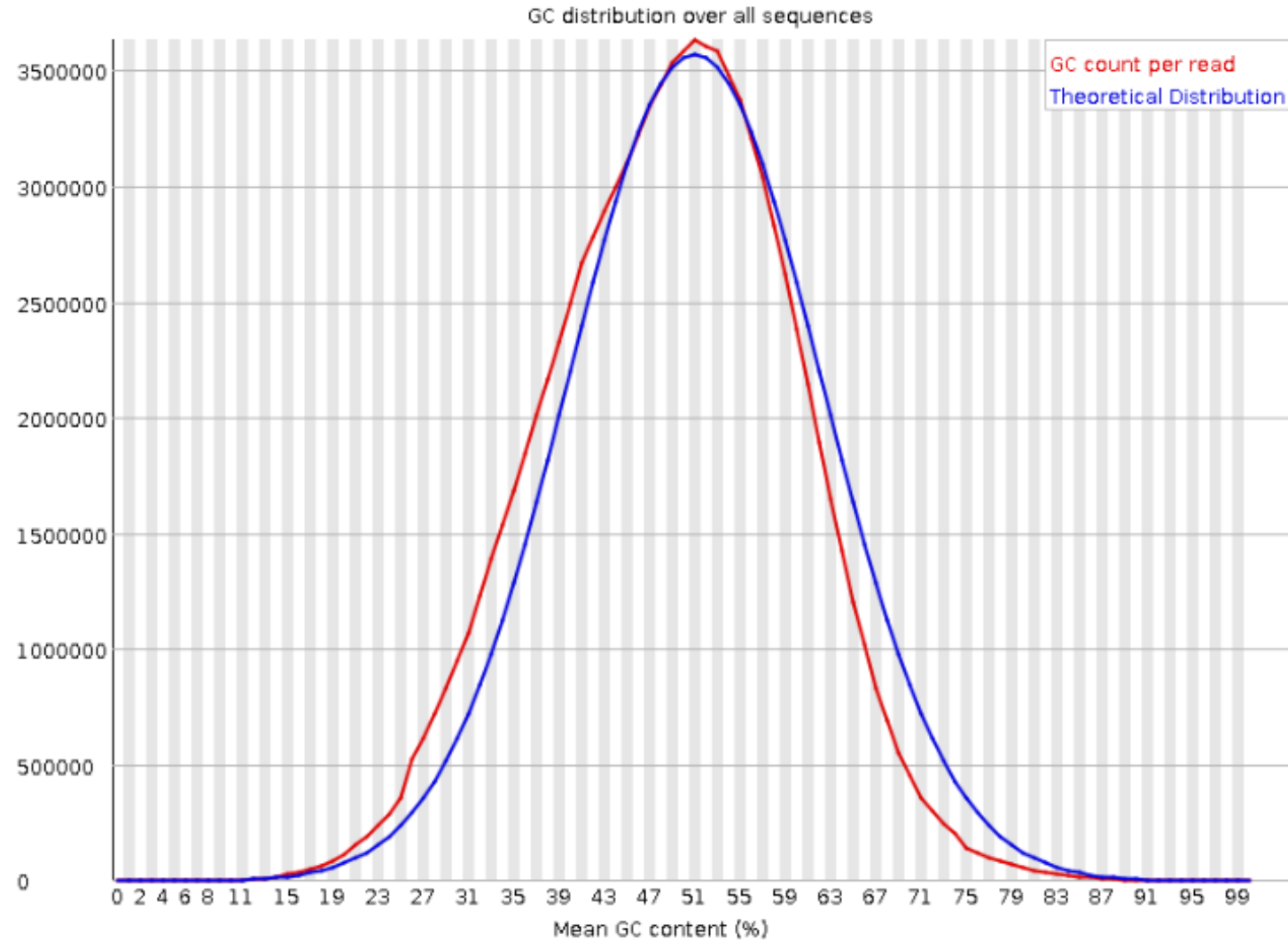
Schritt 5 – Read Quality Check – FASTQC Beispiele

❌ Per base sequence content



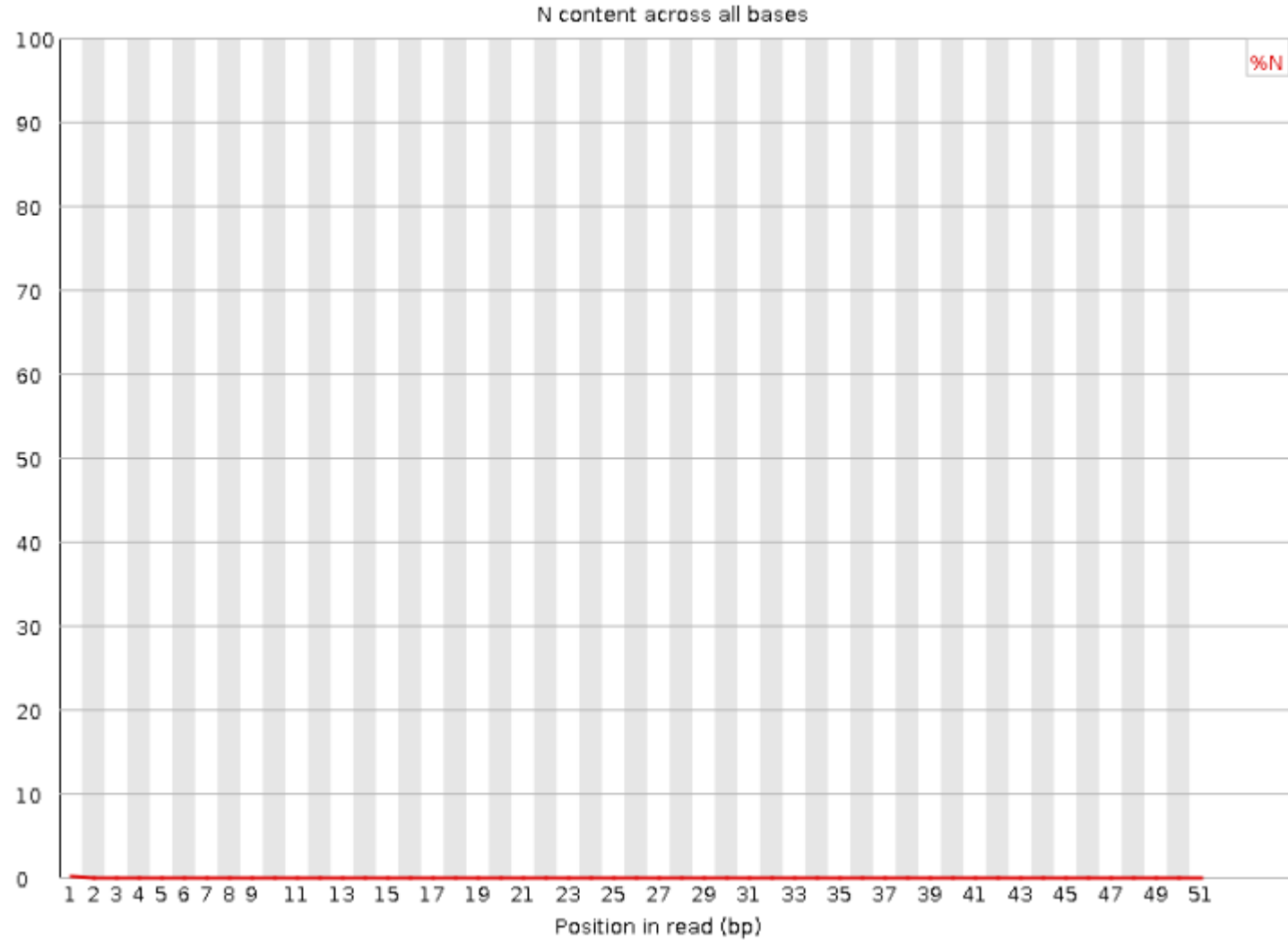
Schritt 5 – Read Quality Check – FASTQC Beispiele

✔ Per sequence GC content



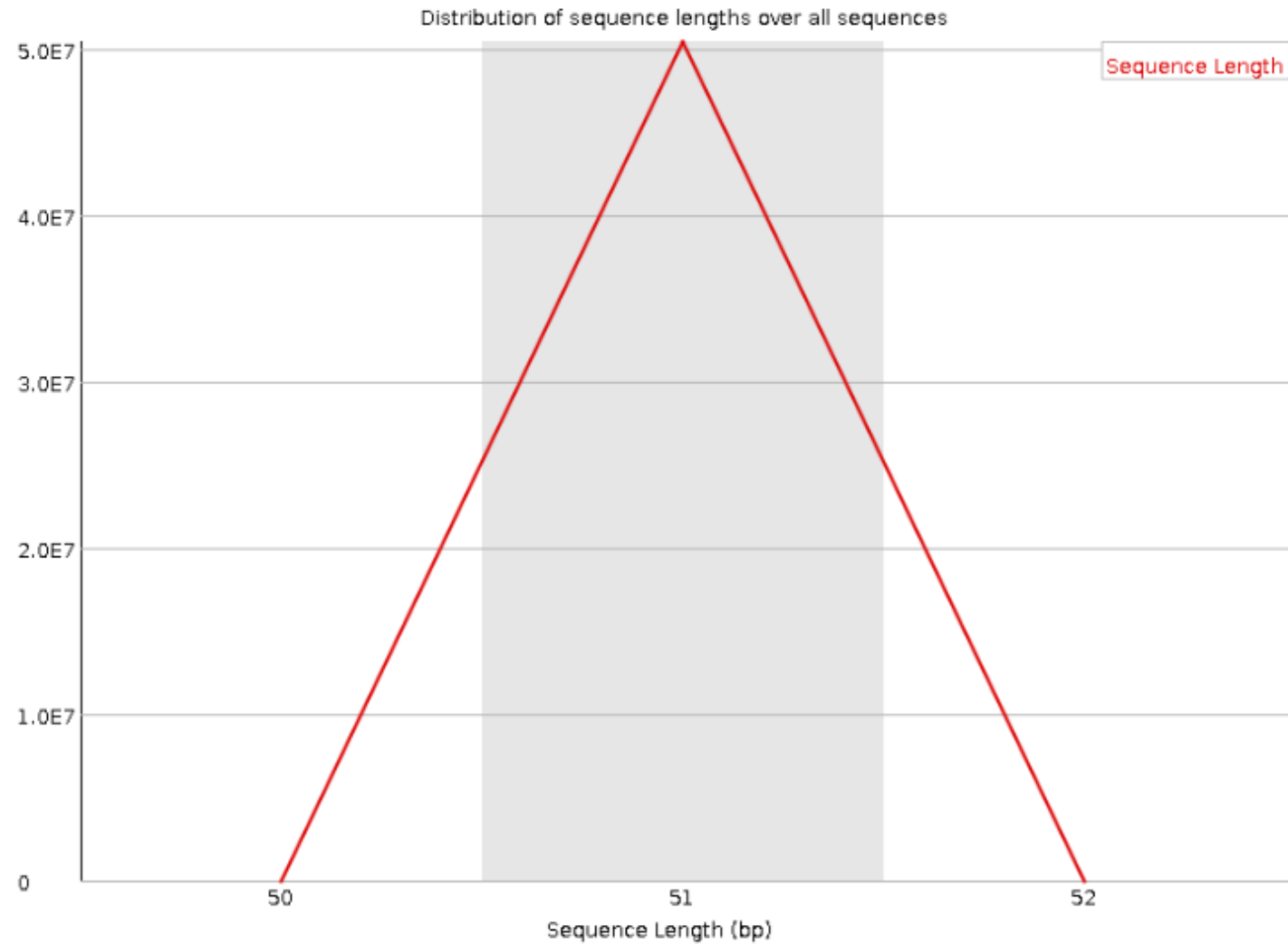
Schritt 5 – Read Quality Check – FASTQC Beispiele

✔ Per base N content



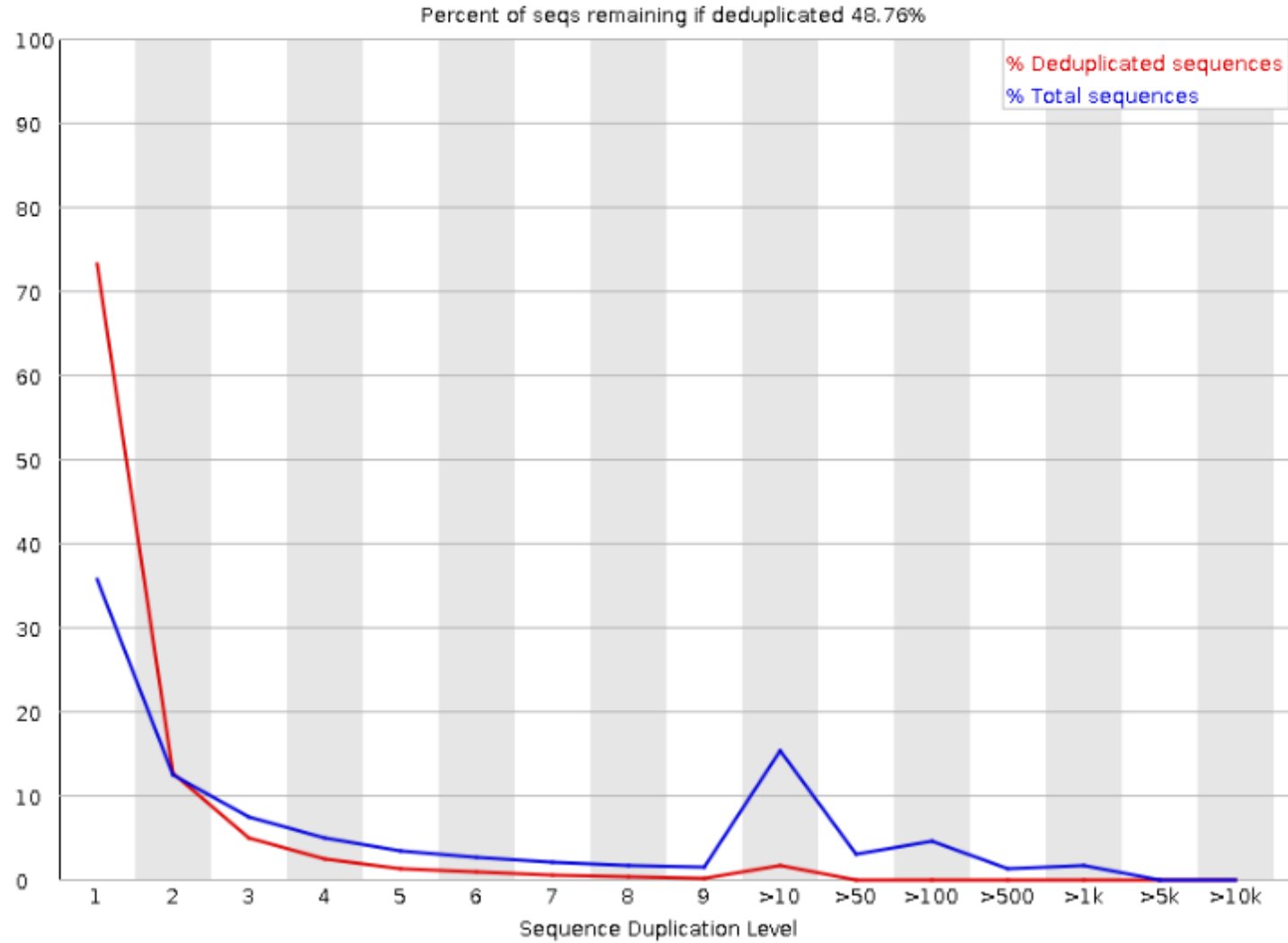
Schritt 5 – Read Quality Check – FASTQC Beispiele

✔ Sequence Length Distribution



Schritt 5 – Read Quality Check – FASTQC Beispiele

❌ Sequence Duplication Levels



Schritt 5 – Read Quality Check – FASTQC Beispiele

FastQC Report

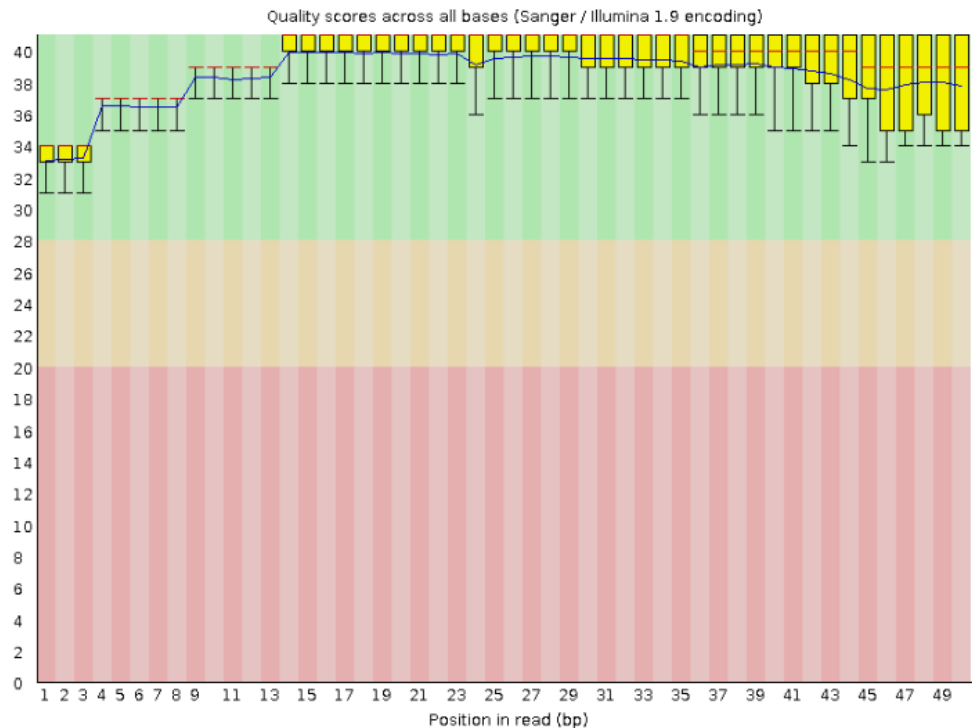
Summary

- ✔ [Basic Statistics](#)
- ✔ [Per base sequence quality](#)
- ✘ [Per tile sequence quality](#)
- ✔ [Per sequence quality scores](#)
- ✘ [Per base sequence content](#)
- ✔ [Per sequence GC content](#)
- ✔ [Per base N content](#)
- ⚠ [Sequence Length Distribution](#)
- ✘ [Sequence Duplication Levels](#)
- ✘ [Overrepresented sequences](#)
- ✔ [Adapter Content](#)
- ✘ [Kmer Content](#)

✔ Basic Statistics

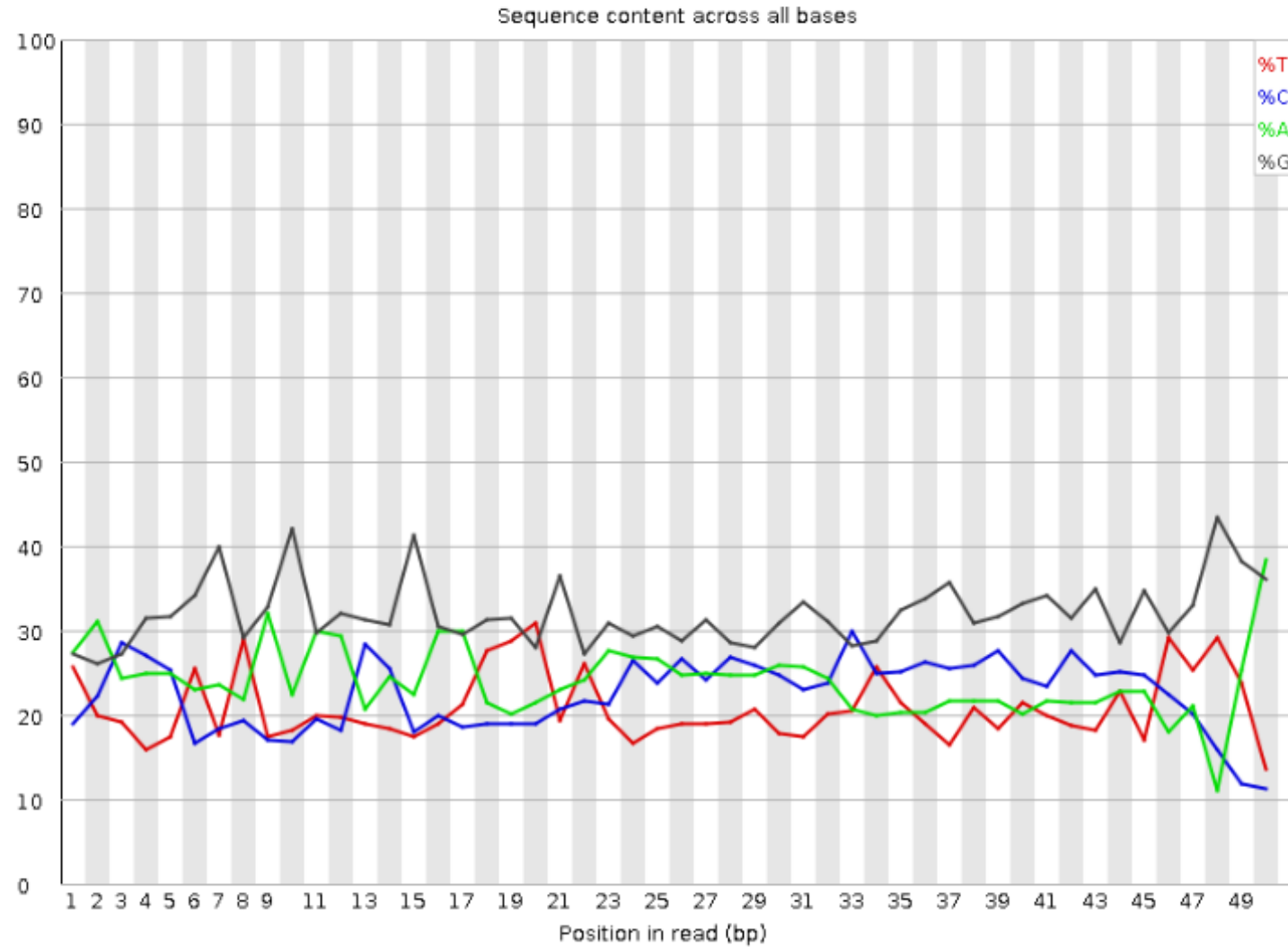
Measure	Value
Filename	Pool12_Index9_GATCAG_L006_R1_001_prinseq_good_KXW5.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	13104468
Sequences flagged as poor quality	0
Sequence length	15-50
%GC	53

✔ Per base sequence quality



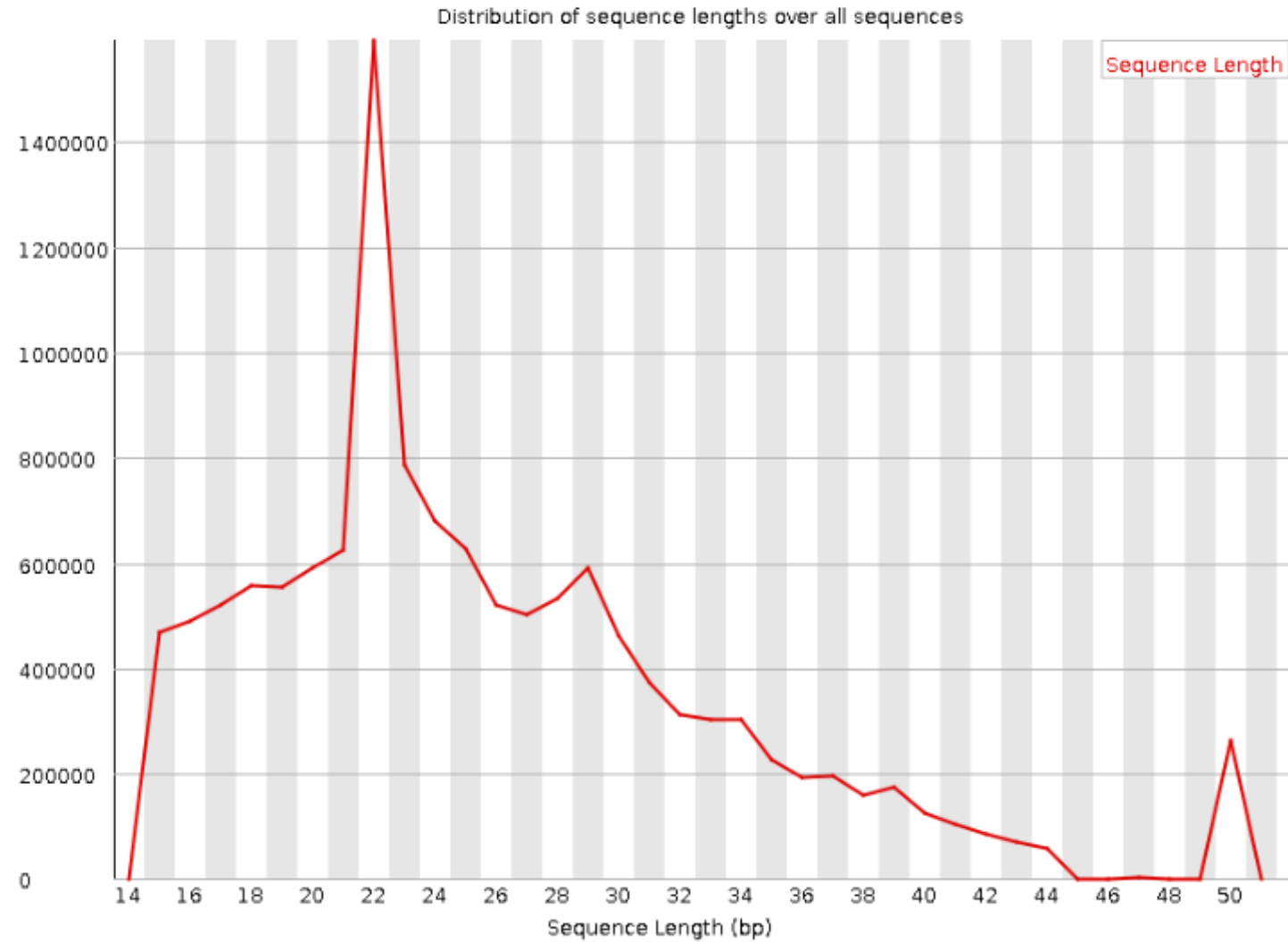
Schritt 5 – Read Quality Check – FASTQC Beispiele

❌ Per base sequence content



Schritt 5 – Read Quality Check – FASTQC Beispiele

🚨 Sequence Length Distribution



Schritt 5 – Read Quality Check – FASTQC Beispiele

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TACCTGTAGAACCGAATTTGT	543712	4.149058168557472	No Hit
TACCTGTAGAACCGAATGTGT	169634	1.2944745257876933	No Hit
TACCTGTAGAACCGAATTTGCG	123943	0.9458071857629017	No Hit
TACCTGTAGAACCGAATTTGC	94925	0.7243712602449791	No Hit
GGCAATACCGGGTTGTAGGACTA	91590	0.6989219249495668	No Hit
GCAATACCGGGTTGTAGGACTACA	80869	0.6171101337345399	No Hit
GTTACGGCAATACCGGGTTGTAGGACTA	51617	0.3938885577041357	No Hit
TTACGGCAATACCGGGTTGTAGGACTA	50409	0.38467032770807635	No Hit
ACCTGTAGAACCGAATTTGCG	41319	0.3153046731847489	No Hit
GTTACGGCAATACCGGGTTGTAGGACTAC	36158	0.2759211591039026	No Hit
TTACGGCAATACCGGGTTGTAGGACTAC	31160	0.2377814955937166	No Hit
GAAGAGTCGTTCCGAGACCAGGACGTTGATAGGCTGGGTG	31060	0.23701839708410904	No Hit
GGCAATACCGGGTTGTAGGACTAC	30505	0.23278320035578703	No Hit
CAATACCGGGTTGTAGGACTACA	30040	0.22923479228611188	No Hit
ACCGGGTTGTAGGACTACA	29758	0.22708285448901855	No Hit
GCAATACCGGGTTGTAGGACTA	22892	0.1746885108193633	No Hit
ACCGGGTTGTAGGACTAC	22330	0.1703998971953688	No Hit
TAAGTGCTAAGTTGGGGTAG	21705	0.16563053151032153	No Hit
TGTTACGGCAATACCGGGTTGTAGGACTA	20622	0.15736617465127162	No Hit
ACCTGTAGAACCGAATTTGTGT	20009	0.15268838078737726	No Hit
ACCGGGTTGTAGGACTA	19548	0.1491704966580864	No Hit
TTACGGCAATACCGGGTTGTAGGACTACA	18821	0.1436227704932394	No Hit
CCCGTGTAAAGTAGGACATCGTCAGGCT	18679	0.14253917060959667	No Hit
ACCTGTAGAACCGAATTTGTG	18475	0.14098244964999723	No Hit
GTTACGGCAATACCGGGTTGTAGGACTACA	17558	0.1339848363168959	No Hit
ACTTGAAGGGTTGATCCTGGCTCAGAACGAACGTTGGCGCGTGGATTG	16194	0.12357617264584872	No Hit
TATTGCACCTGTCCCGCCTGT	15669	0.11956990547040902	No Hit
ACCTGTAGAACCGAATTTGT	15626	0.11924177311127776	No Hit
GCAATACCGGGTTGTAGGACTAC	15502	0.1182955309593644	No Hit
GCGGGTGTAGCTCAGTTGGTTAGAGTGCC	15296	0.11672354802957281	No Hit
TACGGCAATACCGGGTTGTAGGACTA	15171	0.11576967489256337	No Hit
TACCTGTAGATCCGGATTGT	14407	0.1099396022791616	No Hit
ACCGGGTTGTAGGACT	14183	0.10823026161764064	No Hit

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TACCTGTAGAACCGAATTTGTTGGAATTCCTGGGTGCCAAGGAAGTCCA	524116	3.4482164232198875	RNA PCR Primer, Index 1 (100% over 28bp)
TGGAATTCCTGGGTGCCAAGGAAGTCCAAGTACGATCAGATCTCGTATGC	235155	1.5471104354804521	RNA PCR Primer, Index 9 (100% over 50bp)
TACCTGTAGAACCGAATGTGTTGGAATTCCTGGGTGCCAAGGAAGTCCA	164281	1.0808226465572244	RNA PCR Primer, Index 1 (100% over 28bp)
TACCTGTAGAACCGAATTTGCGTGAATTCCTGGGTGCCAAGGAAGTCCA	119653	0.7872101589868067	RNA PCR Primer, Index 1 (100% over 27bp)
TACCTGTAGAACCGAATTTGCTGGAATTCCTGGGTGCCAAGGAAGTCCA	91454	0.6016858572704354	RNA PCR Primer, Index 1 (100% over 28bp)
GGCAATACCGGGTTGTAGGACTATGGAATTCCTGGGTGCCAAGGAAGTCCA	90833	0.5976002304267224	RNA PCR Primer, Index 1 (100% over 26bp)
GCAATACCGGGTTGTAGGACTACATGGAATTCCTGGGTGCCAAGGAAGTCCA	78755	0.5181377489156641	RNA PCR Primer, Index 1 (100% over 25bp)
GTTACGGCAATACCGGGTTGTAGGACTATGGAATTCCTGGGTGCCAAGG	49939	0.32855413679257633	Illumina Small RNA Adapter 2 (100% over 21bp)
TTACGGCAATACCGGGTTGTAGGACTATGGAATTCCTGGGTGCCAAGG	48749	0.32072499678610517	RNA PCR Primer, Index 1 (100% over 22bp)
ACCTGTAGAACCGAATTTGCGTGAATTCCTGGGTGCCAAGGAAGTCCA	39871	0.2623156648722804	RNA PCR Primer, Index 1 (100% over 28bp)
GTTACGGCAATACCGGGTTGTAGGACTACTGGAATTCCTGGGTGCCAAG	35149	0.23124911099786272	No Hit
GGCAATACCGGGTTGTAGGACTACTGGAATTCCTGGGTGCCAAGGAAGTCCA	30360	0.19974175680375295	RNA PCR Primer, Index 1 (100% over 25bp)
TTACGGCAATACCGGGTTGTAGGACTACTGGAATTCCTGGGTGCCAAGG	30159	0.1984193558446767	Illumina Small RNA Adapter 2 (100% over 21bp)
CAATACCGGGTTGTAGGACTACATGGAATTCCTGGGTGCCAAGGAAGTCCA	29877	0.1965640470364205	RNA PCR Primer, Index 1 (100% over 26bp)
ACCGGGTTGTAGGACTACATGGAATTCCTGGGTGCCAAGGAAGTCCAAGT	28387	0.18676117425520866	RNA PCR Primer, Index 1 (100% over 30bp)
GAAGAGTCGTTCCGAGACCAGGACGTTGATAGGCTGGGTGGAATGGAAT	24543	0.1614710783015319	No Hit
TAAGTGCTAAGTTGGGGTAGTGAATTCCTGGGTGCCAAGGAAGTCCA	21286	0.14004291947709765	RNA PCR Primer, Index 1 (100% over 28bp)
ACCGGGTTGTAGGACTACTGGAATTCCTGGGTGCCAAGGAAGTCCAAGT	21140	0.13908236952672387	RNA PCR Primer, Index 1 (100% over 31bp)
GCAATACCGGGTTGTAGGACTATGGAATTCCTGGGTGCCAAGGAAGTCCA	20350	0.13388487321990686	RNA PCR Primer, Index 1 (100% over 27bp)
TGTTACGGCAATACCGGGTTGTAGGACTATGGAATTCCTGGGTGCCAAG	20055	0.13194403599141188	No Hit
ACCTGTAGAACCGAATTTGTTGGAATTCCTGGGTGCCAAGGAAGTCCA	19356	0.1273452386262662	RNA PCR Primer, Index 1 (100% over 27bp)
ACCGGGTTGTAGGACTATGGAATTCCTGGGTGCCAAGGAAGTCCAAGTCA	18412	0.12113455949508231	RNA PCR Primer, Index 1 (100% over 32bp)
TTACGGCAATACCGGGTTGTAGGACTACATGGAATTCCTGGGTGCCAAG	18304	0.12042401569617568	No Hit
CCCGTGTAAAGTAGGACATCGTCAGGCTTGAATTCCTGGGTGCCAAGG	18097	0.11906214008160464	Illumina Small RNA Adapter 2 (100% over 21bp)
GTAAACGGCGCCTGGAATTCCTGGGTGCCAAGGAAGTCCAAGTACGATC	18065	0.1188516085856323	RNA PCR Primer, Index 9 (100% over 37bp)
ACCTGTAGAACCGAATTTGTGGAATTCCTGGGTGCCAAGGAAGTCCA	17881	0.11764105248379136	RNA PCR Primer, Index 1 (100% over 28bp)
GTTACGGCAATACCGGGTTGTAGGACTACATGGAATTCCTGGGTGCCAAG	17066	0.11227907844574597	No Hit
AGTAAACGGCGCCTGGAATTCCTGGGTGCCAAGGAAGTCCAAGTACGAT	16837	0.11077246242769394	RNA PCR Primer, Index 9 (100% over 36bp)
ACTTGAAGGGTTGATCCTGGCTCAGAACGAACGTTGGCGCGTGGATTG	16379	0.1077592303915899	No Hit

Schritt 5 – Read Quality Check – FASTQC Beispiele

✘ Adapter Content

