How adapter and indexes are linked to the DNA, and what parts of them may be left in the sequencing reads

This is assuming that we follow the NEBNext® UltraII library preparation protocol, and it is a compilation of the information given in the following documents:

Index sequences (set 1):
LIBRARY PREPARATION NEBNext® Multiplex Oligos for Illumina® (Dual Index Primers Set 1) Instruction manual
https://www.neb.com/-/media/catalog/datacards-or-manuals/manuale7600.pdf

Sequencing reading process:
Illumina Indexed Sequencing overview guide Document # 15057455 v04
https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/indexed-sequencing-overview-guide-15057455-04.pdf

1. End preparation

DNA fragment = the insert (a few bases are given to keep track of orientation and strand):

```
5' xxxxxxxxxxxxxxxxatgcxxxxxxxxxxxxxxx 3'
3' xxxxxxxxxxxxxxxxtacgxxxxxxxxxxxxxxx 5'
```

The DNA fragment may lack some bases in the 5’ extremities so the first step is to repair these, using the other strand as a template:

```
5'   xxxxxxxxxxxxxxxatgcxxxxxxxxxxxxxxx 3'
3' xxxxxxxxxxxxxxxxtacgxxxxxxxxxx      5'
                     |
                     V
5' xxxxxxxxxxxxxxxxatgcxxxxxxxxxxxxxxx 3'
3' xxxxxxxxxxxxxxxxtacgxxxxxxxxxxxxxxx 5'
```

In addition, during the End Prep step, one a is added to the 3’ extremities:

```
5'  xxxxxxxxxxxxxxxxatgcxxxxxxxxxxxxxxxa 3'
3' axxxxxxxxxxxxxxxxtacgxxxxxxxxxxxxxxx 5'
```

2. Adapter ligation

NEBNext Adapter for Illumina (hairpin shape, the grey and brown parts form the loop of the hairpin):

5′ **GATCGGAAGAGCACACGTCTGAACTCCAGTC/ideoxyU/**ACACTCTTTCCCTACACGAC**GCTCTTCCGATC*T** 3′
=
5′ **GATCGGAAGAGCACACGTCTGAACTCCAGTCU**ACACTCTTTCCCTACACGAC**GCTCTTCCGATC*T** 3′

The adapter is ligated to the insert, and the USER enzyme cuts the hairpin loop between the U and the A:
the grey and brown extremities of the adapter, corresponding to the loop, do not pair

5' ACACTCTTTCCCTACACGAC**GCTCTTCCGATCT**xxxxxxxxxxxxxxxxx**atgc**xxxxxxxxxxxxxxxa**GATCGGAAGAGCACACGTCTGAACTCCAGTCU** 3'
3' ***UCTGACCTCAAGTCTGCACACGAGAAGGCTAG***axxxxxxxxxxxxxxxxx**tacg**xxxxxxxxxxxxxxx***TCTAGCCTTCTCG***CAGCACATCCCTTTCTCACA 5'

Once the adapter is ligated, total size = insert size + 66 bp (33 bp + insert size + 33 bp)

(shades of a same colour represent reverse complements, italics indicate reversed sequences)

3. Primers+indexes ligation (PCR)

All the following assumes that we use the NEB indexes i501 and i701

The primers that we put during the PCR step consist in:

the i5 primer:

NEB universal primer + **index 5** + **the loop** and **the pin** of the 3′ part of the adapter (both together called P5):

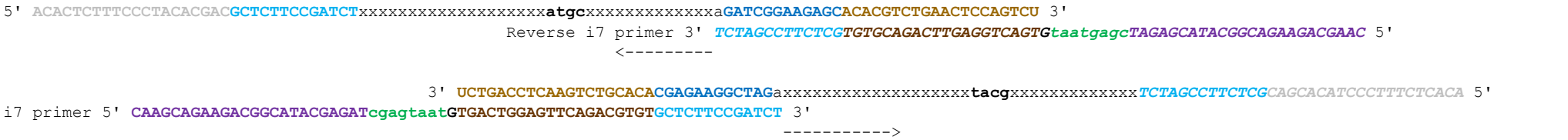
5' **AATGATACGGCGACCACCGAGATCTACAC****tatagcct**ACACTCTTTCCCTACACGAC**GCTCTTCCGATC*T** 3'

and the i7 primer:

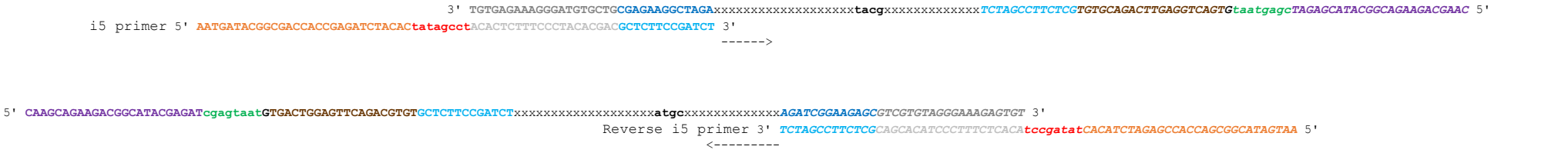
Primer + **index 7** + **"G", the reverse complement of the loop of the 5′ part of the adapter** and **the pin of the 3′ part of the adapter** (the three together called P7):

5' **CAAGCAGAAGACGGCATACGAGAT****cgagtaat****GTGACTGGAGTTCAGACGTGT****GCTCTTCCGATC*T** 3'

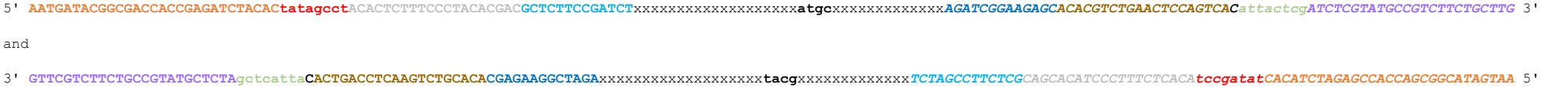
Mixing the DNA-Adapters with the i7 primer:



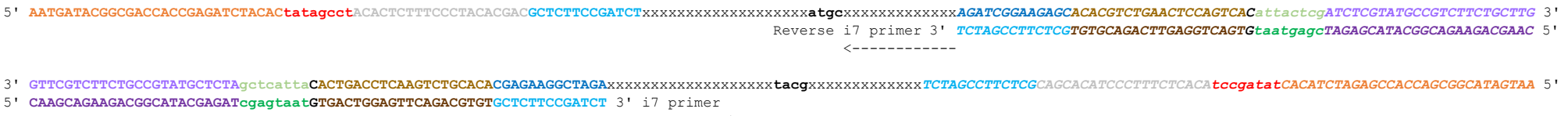
Adding the i5 primer:



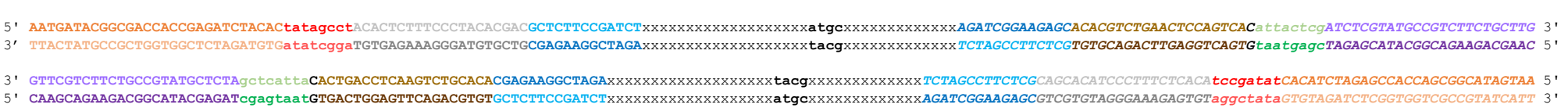
gives:



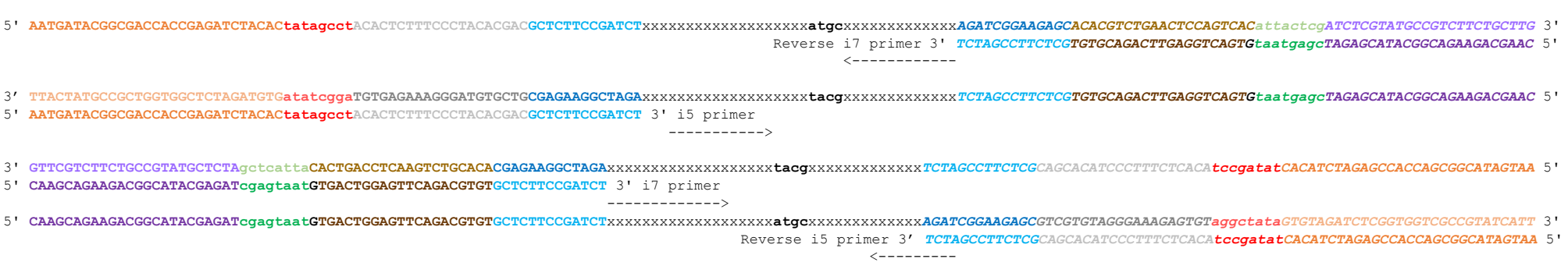
The PCR continues:



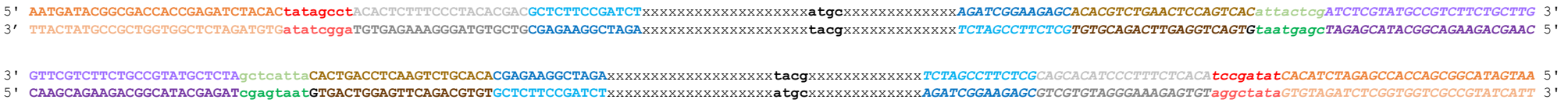
gives:



These products are denaturated and i7 and i5 primers are used many times to amplify them:



This gives a lot of:



(These are not exactly the same because the inserts are in reverse orientation relative to the adapters, see the bases in the middle)

At the end of the PCR, the library fragments consist in 70 bp + insert + 66 bp = insert size + 136 bp

4. Sequencing

(The following is valid for NovaSeq™ 6000, MiSeq™, HiSeq 2500, and HiSeq 2000, but for other configurations, check Workflow B in the pdf cited at the beginning of this document)

The sequencing of a library fragment is performed in five steps:

1. Read 1
2. Index 1 Read (i7 read)
3. Index 2 Read (i5 read)
4. Preparation for read 2
5. Read 2

4.1 Read 1

P5 is added and anneals to the DNA, and the complementary strand of the insert is synthesized.

```
3' TTACTATGCCGCTGGTGGCTCTAGATGTGatatcggaTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAxxxxxxxxxxxxxxxxxxxxtacgxxxxxxxxxxxxxTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGtaatgagcTAGAGCATACGGCAGAAGACGAAC 5'
                          P5 5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
                                                                ---------->
```

The Illumina software reads what is synthesized, so read 1 will consist in:

5' xxxxxxxxxxxx 3'

Or:

5' xxxxxxxxxxxxxxxxxxxx**atgc**xxxxxxxxxxxxx

Or:

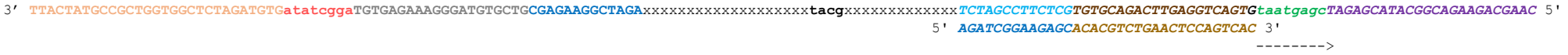
5' xxxxxxxxxxxxxxxxxxxx**atgc**xxxxxxxxxxxxxx*AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC*attactcg*ATCTCGTATGCCGTCTTCTGCTTG* 3'

Read 1 will thus consist in the complement of the insert, either partial, or full, or full + a part of, or the full, reverse-complemented i7 primer.
This will depend on the size of the insert relative to the number of synthesis cycles performed (i.e. the chosen "read length" at the beginning of the sequencing, which is the same for all fragments sequenced in a run)
In bad cases, read 1 may even go through the i7 primer and finish with random bases:

5' xxxxxxxxxxxxxxxxxxxx**atgc**xxxxxxxxxxxxxx*AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC*attactcg*ATCTCGTATGCCGTCTTCTGCTTGGGGGGGGGGGGCCCCCCCGG* 3'

4.2 Index 1 (i7) read

P7 is added and anneals to the same template strand after removal of read 1, and the complementary strand of the index is synthesized:



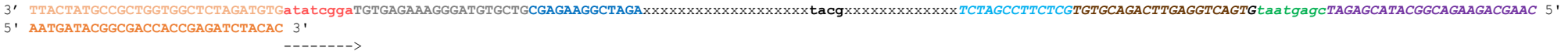
The Illumina software reads what is synthesized, so index 1 read will consist in:

attactcg

Index 1 read will thus consist in the reverse complement of i7.

4.3 Index 2(i5) read

The index 1 read is removed. The NEB universal primer is added and anneals to the same template strand (which forms a bridge to attach on the flow cell)**, and the complementary strand of the index is synthesized:**



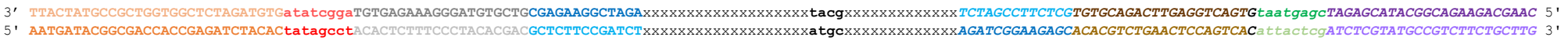
The Illumina software reads what is synthesized, so index 2 read will consist in:

tatagcct

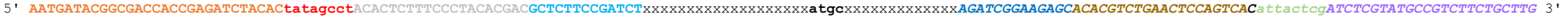
Index 2 read will thus consist in i5.

4.4 Preparation for read 2

The index 2 read is removed and the complementary strand of the same template is synthesized

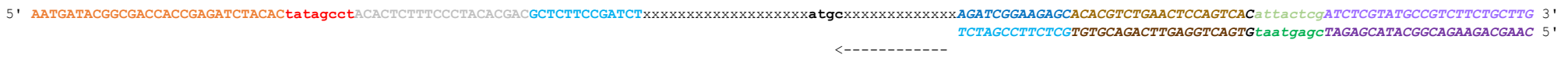


Then the original template strand is removed, leaving the complementary strand ready for read 2:



4.5 Read 2:

P7 is added and anneals to the DNA, and the complementary strand of the insert is synthesized.



The Illumina software reads what is synthesized, so read 2 will consist in:

3' xxxxxxxxxxx 5'

Or

3' xxxxxxxxxxxxxxxxxxxtacgxxxxxxxxxxxxxx 5'

Or

3' TTACTATGCCGCTGGTGGCTCTAGATGTGatatcggaTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAxxxxxxxxxxxxxxtacgxxxxxxxxxxxxxxxxxx 5'

The read is in fact in the other direction, starting by the end, giving thus:

5' xxxxxxxxxxx 3'

Or

5' xxxxxxxxxxxxxxgcatxxxxxxxxxxxxxxxxxx 3'

Or

5' xxxxxxxxxxxxxxgcatxxxxxxxxxxxxxxxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTaggctataGTGTAGATCTCGGTGGTCGCCGTATCATT 3'

Read 2 will thus consist in the complement of the insert, either partial, or full, or full and a part of, or the full, reverse-complemented i5 primer.
In bad cases, read 1 may even go through the i5 primer and finish with random bases:

5' xxxxxxxxxxxxxxgcatxxxxxxxxxxxxxxxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTaggctataGTGTAGATCTCGGTGGTCGCCGTATCATTGGGGGGGGGGGGGGG 3'

5. Adapters in the data:

At the end of the sequencing we thus potentially have:

- Read 1 (and possibly part of the reverse complement of i7 primer) in the file with R1 in the name:

5' xxxxxxxxxxxxxxxxxxx**atgc**xxxxxxxxxxxxxx***AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC*attactcg*ATCTCGTATGCCGTCTTCTGCTTG*** 3'

- Read 2 (and possibly part of the reverse complement of i5 primer) in the file with R2 in the name:

5' xxxxxxxxxxxxxx**gcat**xxxxxxxxxxxxxxxxxx***AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTaggctataGTGTAGATCTCGGTGGTCGCCGTATCATT*** 3'

Read 1 aligns directly on the forward strand of the reference (see DNA sequence at the beginning of this document), whereas Read 2 has to be reverse complemented to align with it.

- Index 1 read (i7) in the file with I1 in the name:

attactcg (same as in sample sheet, but reverse complement of the sequence given in the indexes pdf)

- Index 2 read (i5) in the file with I2 in the name:

tatagcct (same as in sample sheet, same as the sequence given in the indexes pdf)

Greping all in their respective R1, R2 and I1, I2 files should show this.

The beginnings of the read through the i7 or i5 primers fit with the sequences PE2_rc (Read 1) and PE1_rc (Read 2) given in the TruSeq3-PE-2.fa file from Trimmomatic (Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170).

For the MiniSeq, Hiseq3000, Hiseq4000 and NextSeq, I think that the only difference is that the Index 2 read (i5) is the reverse complement of the sequence as given in the pdf, and thus maybe should be given as reverse-complemented in the sample sheet?
Need to be tested when get the data.
(see workflow B, p. 5-6 of pdf for sequencing, reference at the beginning of this document)