**Seeding**

The length of the reads, mean of 1000 bp, makes it impossible to map the complete sequence against the genome; therefore, the first step of the alignment is seeding. During this step, the read is fragmented into small parts called seeds. A fix shift of five bp and a fix length were detected when analysing the coverage and the loss of mappings. For simplicity, the read is seeded in the sense he is written in the *fastQ* file and the reverse complement of it, allowing so a mapping on the reverse strand too.

After the fragmentation, the seeds are mapped with Bowtie 2. We use Bowtie 2 because of its quickness and precision in mapping short reads against a long genome. The parameter for the mapping are End-to-End mapping, allowing a maximum of one mismatch, not allowing mapping on the reverse strand and allowing only 100 alignments for each seed.

**Filtering and chaining**

After the Bowtie 2 mapping of the seeds, the mappings must be filtered and grouped. Since not all mapped regions are relevant, only for those with overlapping seeds a further mapping makes sense. The first filtering step is the deletion of unmapped seeds. Then groups of overlapping seeds are identified. For each seed, we search for overlapping mappings, if there is at least one overlapping seed and the seeds are in a correct order the overlap can be extended forming a seed chain. The detection of the right order occurs using the seed ID. Knowing the ID, the seed length and the used shift one can determine if there is an overlap between the two seeds on the read. Only if there is an overlap on the read the seeds are added to the chain, else they are filtered out. The chain will be extended until the last seed of the chain will not have any overlap with other seeds. If a seed has no other overlapping seeds but in between two seeds in the right order, the seed is annotated as well. After filtering out the mappings of the sense or the antisense, due to the number of mappings or, if similar, the length of the chains, all annotated chains are written to a so-called *fastM* file. In this file, the chains are grouped for all chromosomes of a read.

Analysis of seeds

Analysis of chains

Structure and quality of the reads?