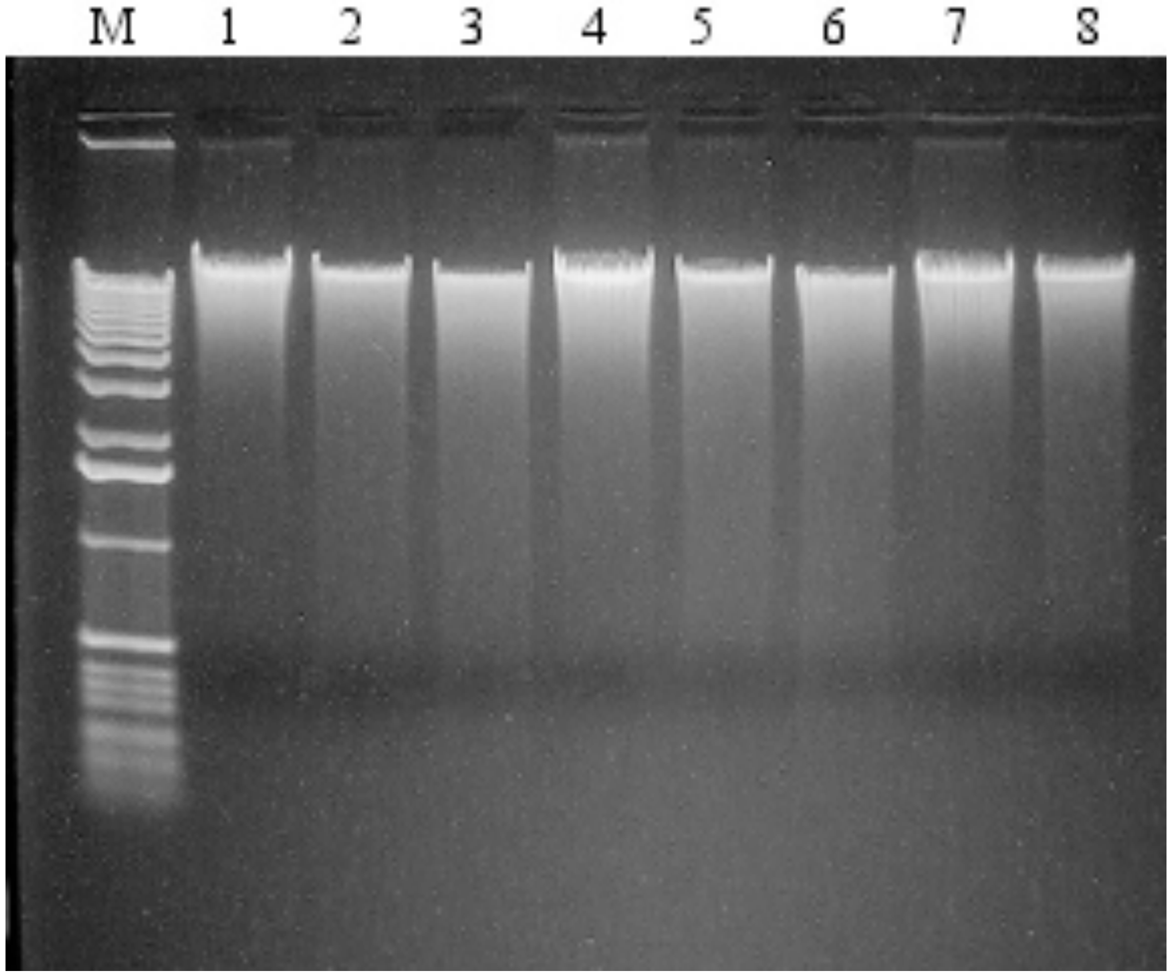
Long reads and hard assembly problems

- How to make an assembly
  - Step 1: Generate high quality genomic DNA.
  - Step 2: Make Illumina sequencing libraries
    - short insert
    - mate pair
    - long reads
  - Step 3: Quality Control and Assembly
    - unitigs
    - Contigs
    - scaffolds

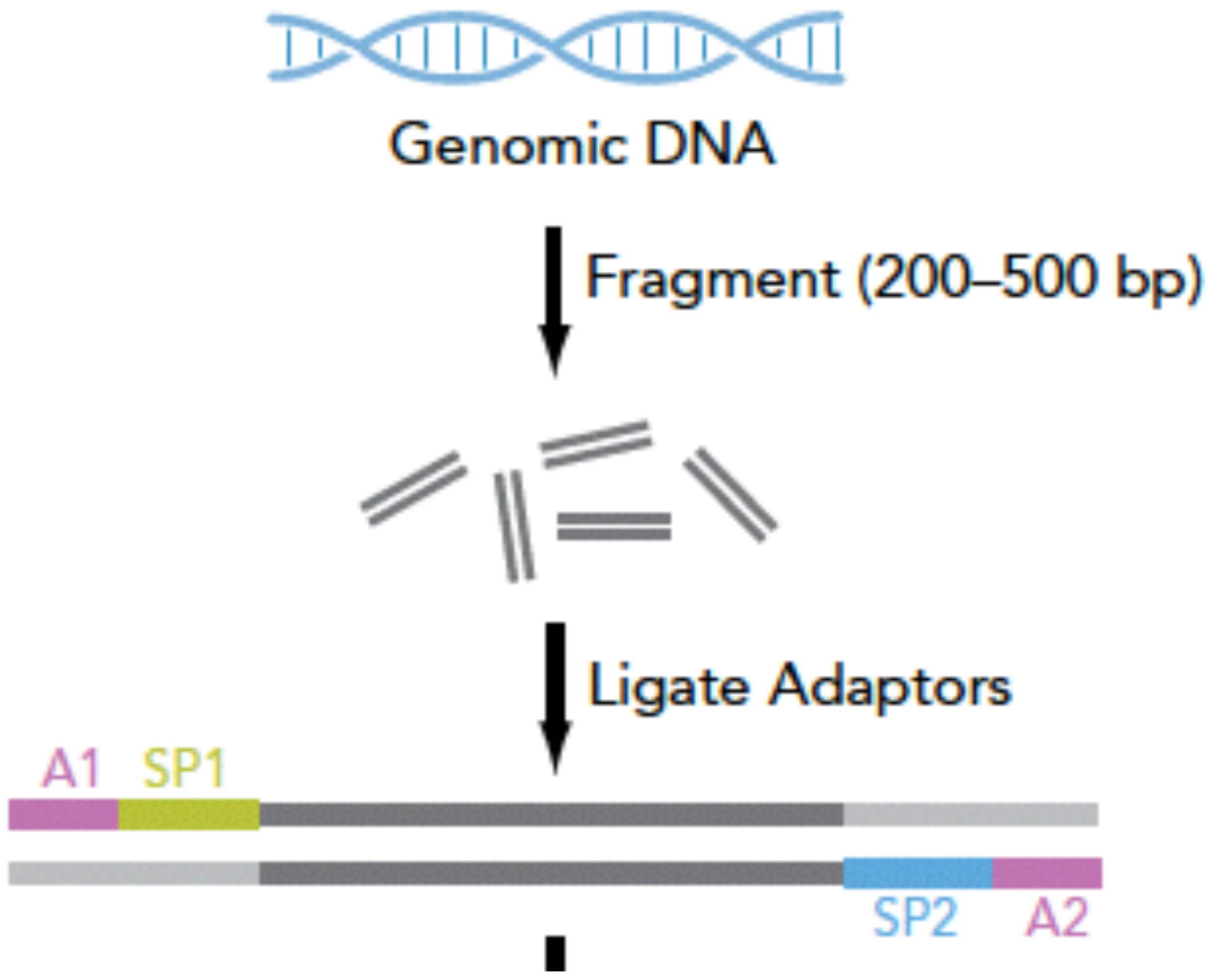
Long reads and hard assembly problems

- Step 1



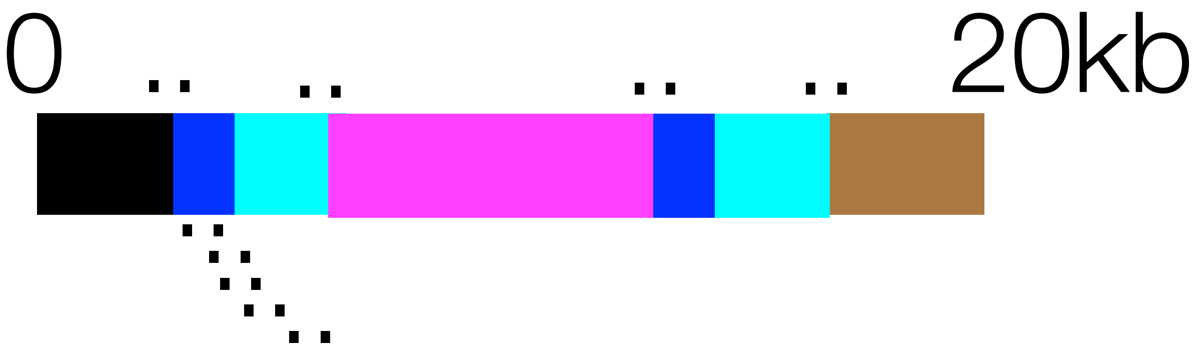
Long reads and hard assembly problems

- Step 2: Short Insert



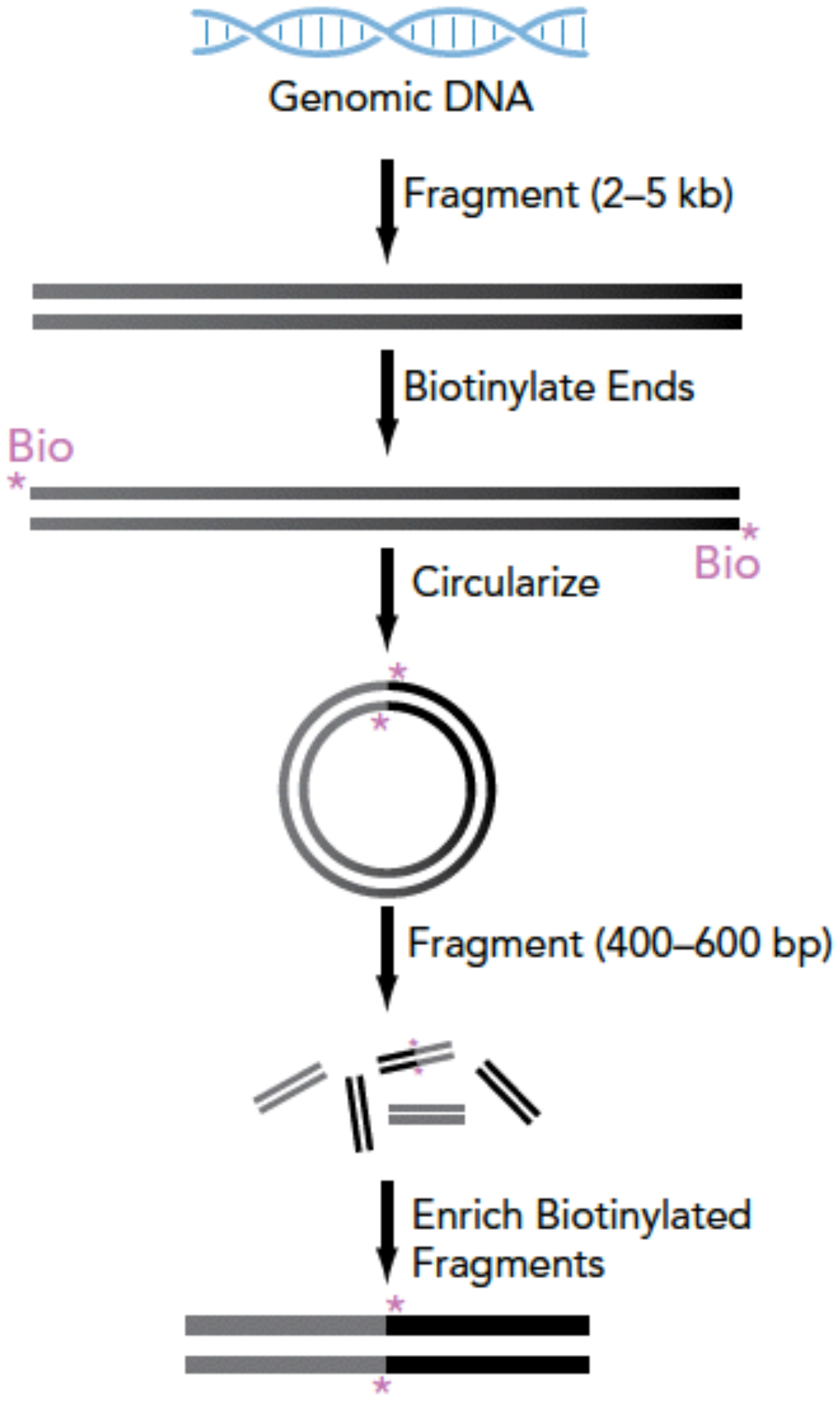
Long reads and hard assembly problems

- Step 2: Short Insert



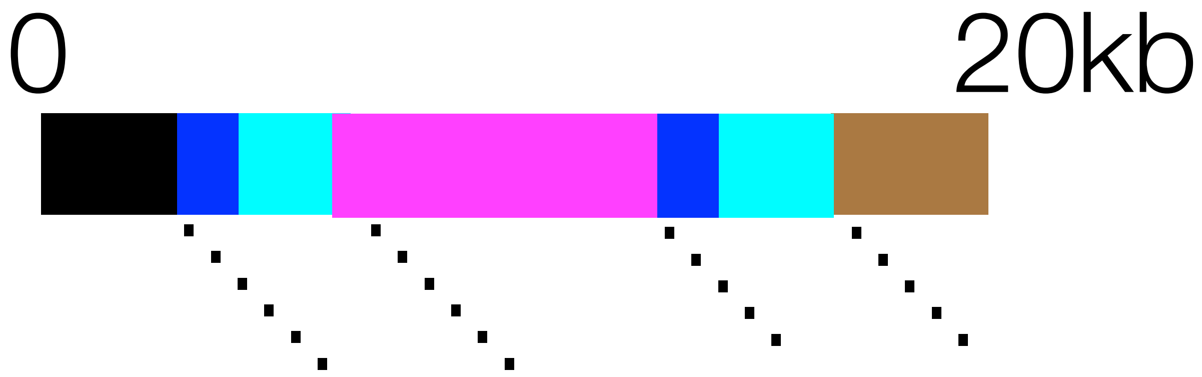
Long reads and hard assembly problems

- Step 2: Mate Pair



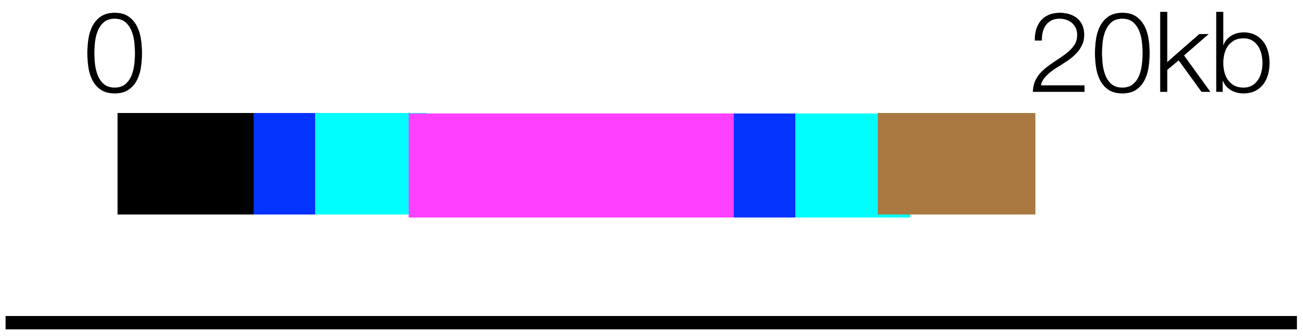
Long reads and hard assembly problems

- Step 2: Mate Pair



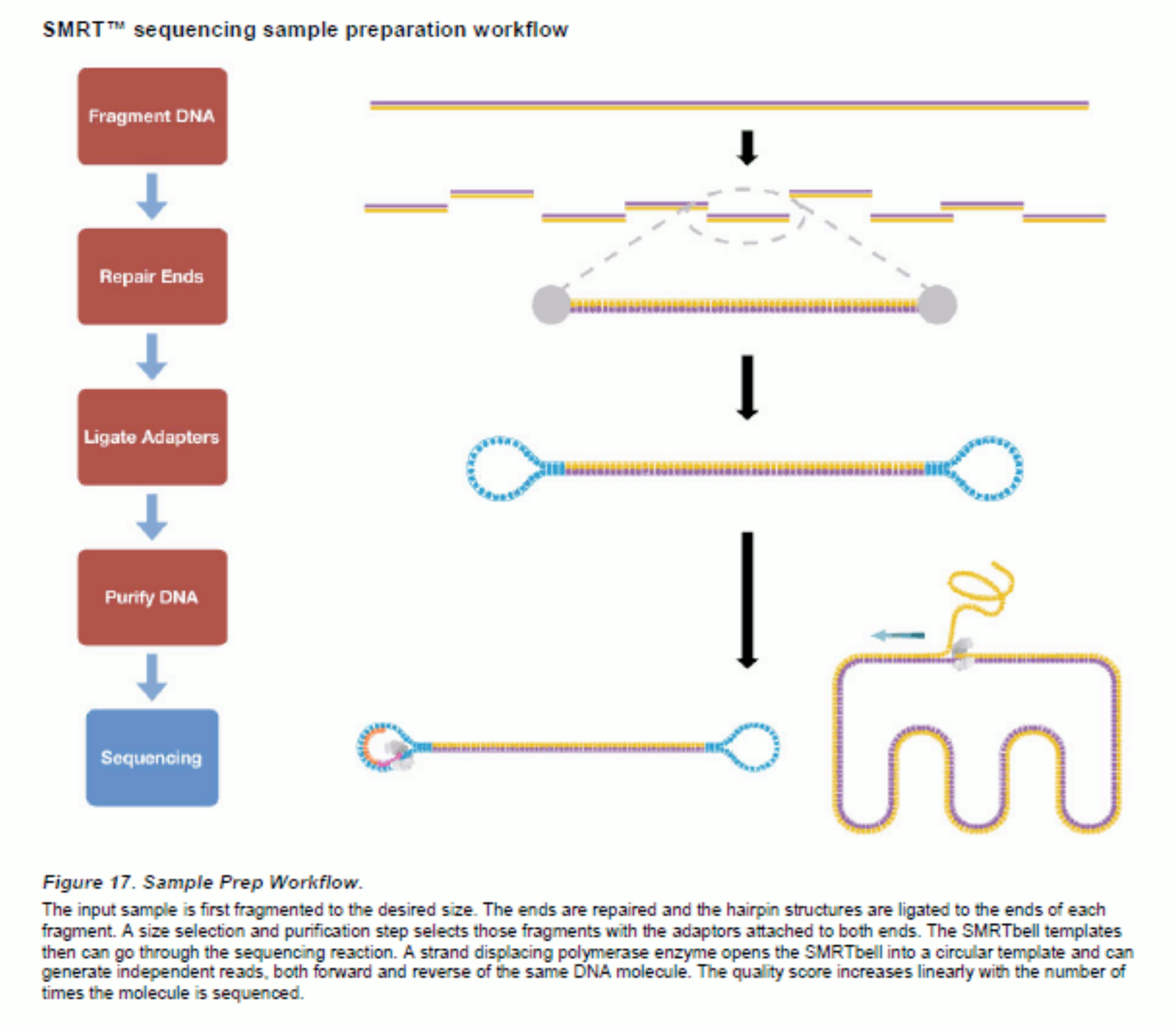
Long reads and hard assembly problems

- Step 2: Long reads



Long Reads

- Types of long reads
    - PacBio



SMRT™ sequencing sample preparation workflow

Figure 17. Sample Prep Workflow.

The input sample is first fragmented to the desired size. The ends are repaired and the hairpin structures are ligated to the ends of each fragment. A size selection and purification step selects those fragments with the adaptors attached to both ends. The SMRTbell templates then can go through the sequencing reaction. A strand displacing polymerase enzyme opens the SMRTbell into a circular template and can generate independent reads, both forward and reverse of the same DNA molecule. The quality score increases linearly with the number of times the molecule is sequenced.

Long Reads

- PacBio
    - Single molecules
    - 600Mb per SMRT Cell ($330 per)
    - Mean read length ~10-15kb
    - Max read length 40kb
    - Error as much as 20%

Long Reads

- PacBio
  - Use for primary assembly (bacterial, bigger genomes)
    - HGAP/Quiver/wgs-assembler/falcon
  - Or for gap filling Illumina assemblies.
    - PBJelly (http://sourceforge.net/p/pb-jelly/wiki/Home/?#058c)
  - Hybrid assembly
    - AllPathsLG (http://www.broadinstitute.org/software/allpaths-lg/blog/)
    - Mira (http://www.chevreux.org/projects_mira.html)
    - ABySS (https://github.com/bcgsc/abyss)

Long Reads

- PacBio:Error Correction
    - Auto-correction (requires high coverage)
    - Correct with short read data
        - Non trivial, lots of time, RAM, IO
            - PBcR (http://wgs-assembler.sourceforge.net/wiki/index.php/PBcR)
            - LSC (http://www.healthcare.uiowa.edu/labs/au/LSC/)

Long Reads

- Tutorial