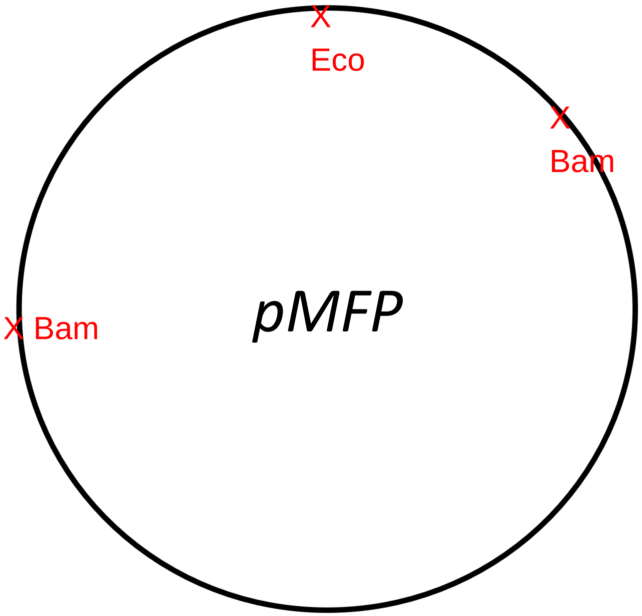
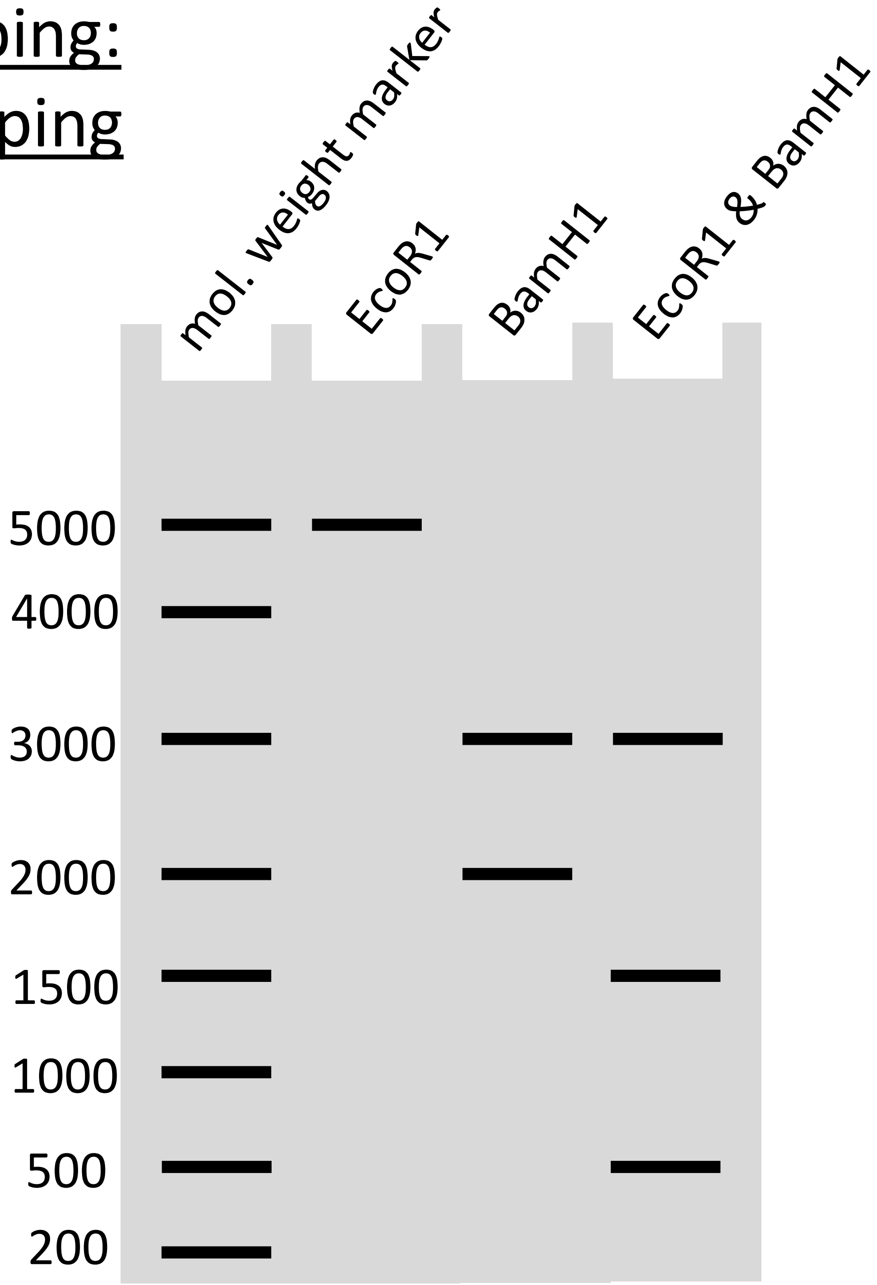
mapping and map-based sequencing

Very simple example for mapping: Restriction-digest-based mapping





- How big is the plasmid (bp)?

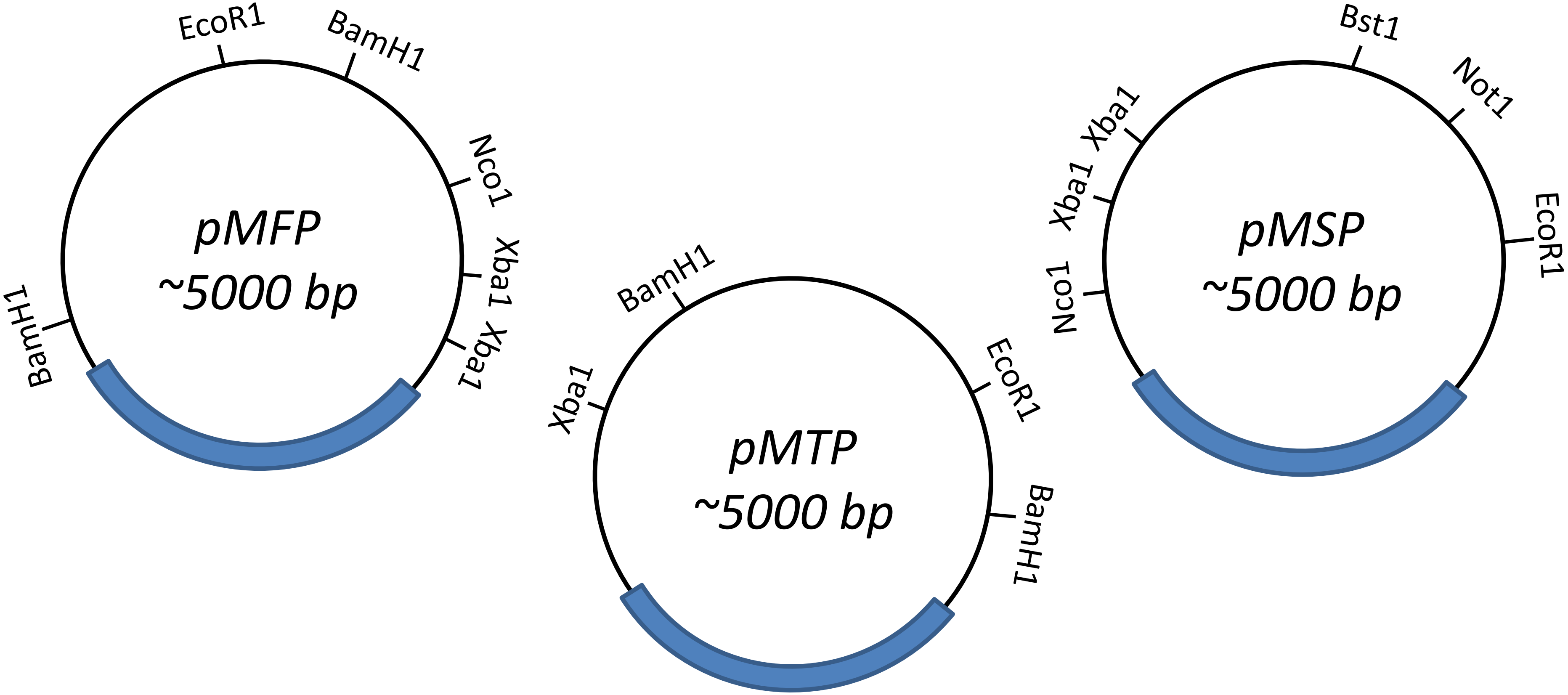
5k bp

- Where are the restriction sites for the enzymes BamH1 and EcoR1?

EcoR1: somehwat random: at 0bp

BamH1: at 2k bp and at 3k bp

Using restriction maps of fragments in plasmids to generate map of original DNA sequence



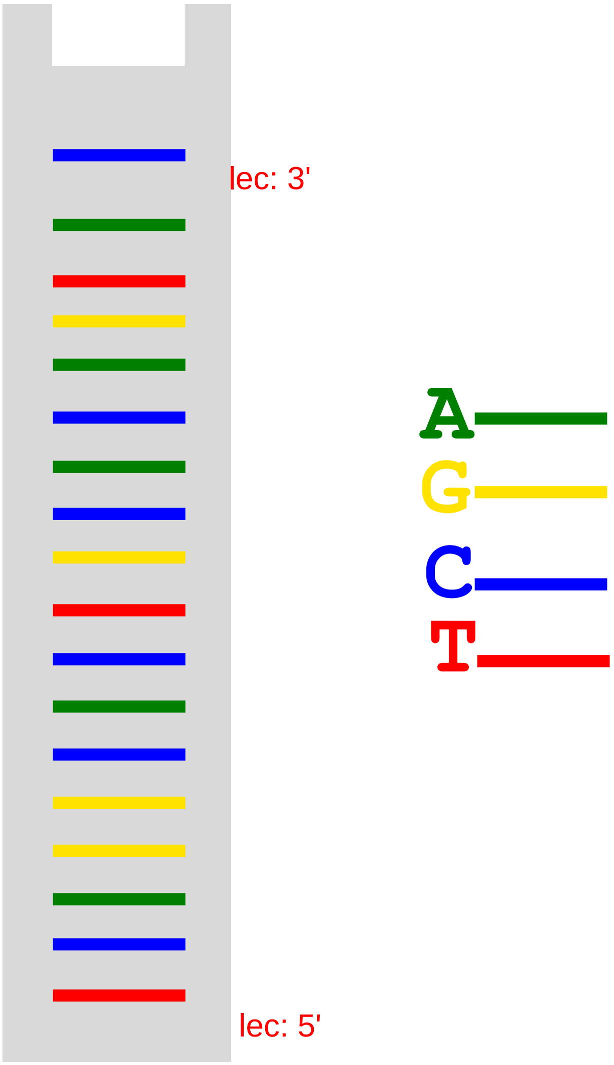
Multiple copies of an original linear DNA sequence were broken into fragments and then cloned into plasmids (plasmid backbone in blue). You have generated restriction maps of your plasmids.

Can you generate a map of the original DNA sequence?

Xba1-BamH1-EcoR1-BamH1-Nco1-Xba1-Xba1-Bst1-Not1-EcoR1

Sanger sequencing

Sanger Sequencing



- Given the sequencing gel on the left, what is the sequence in 5’ to 3’ order?
- How long would be a real Sanger sequence read?
- What would be the error rate?

Lec. solution: 5'-TCAGGCACTGCACAGTAC-3'

3'-CATGACACGTCACGGACT-5'
solution:
5'-GTACTGTGCAGTGCCTGA-3'

ca. 1k bp

error rate: 10**-5

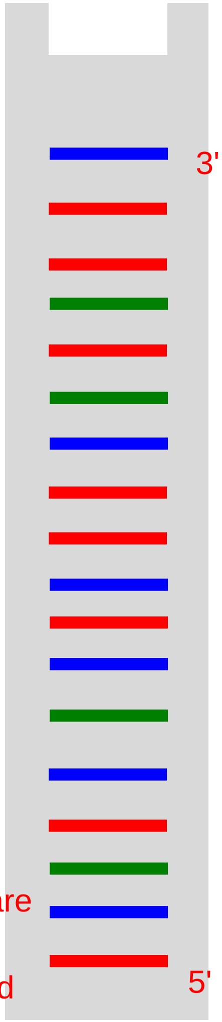
Using Sanger Sequencing to check the success of a cloning experiment

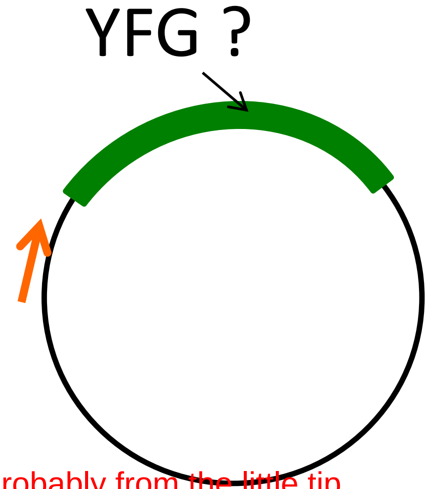
You are trying to clone your favorite Gene (YFG) for protein expression (sequence see below). You finally have obtained a plasmid vector that might contain YFG? Your Sanger sequencing reaction with the orange sequencing primer gives you the gel shown on the right.

What does the data tell you, did you clone YFG or some other random stretch of DNA?

If you want to continue sequencing what would be the sequence of the next sequencing primer (5'-3')?

A 🟩
G 🟨
C 🟦
T 🟥





first three are from plasmid

sequenced:
5'-TCATCACTCTTCATATTC-3'
delete TCA to get the right overlap - the frist three are probably from the little tip right before yfg

```
            M  P  R  L  L  S  A  G  A  L  H  E  Y  E  E  stp
YFG seq  5'-ATGCCTAGATTACTCAGCGCAGGTGCGCTCCATGAATATGAAGAGTGA-3'
         3'-TACGGATCTAATGAGTCGCGTCCACGCGAGGTACTTATACTTCTCACT-5'
                                         CTTATACTTCTCACTACT
```

for the next primer: just take the end of the current sequence and use it as the starting (binding) site for the primer and dna poly

sequencing by synthesis & shot-gun sequencing

Assembly of a DNA sequence from short read sequencing data (easy example) (1)

TAGATGACCT

GAGGCATGGA

TCTATTCCCA

ATGGACGTTG

TCCCATAAGT

GACCTTCTAT

CGTTGGATAT

ACTACTAGAT

CTGGGACTAC

TAAGTGAGGC

1. Can you find the original sequence?
2. How long is the sequence?
3. What is the average coverage?
4. Is this a realistic example? Why?

Assembly of a DNA sequence from short read sequencing data: Example 2

TAGATGACCT

TTCTATTCC

TTGCTAGTTA

TTACTGGGAC

ACGTTGCTAG

GTTGCTAGTT

CTGGGACTAC

CCTTCTATTC

TCTATTCCC

TCTATTCCC

ACTGGGACTA

CTAGTTACTG

CTAGTTACT

CTAGTTACTG

GGGACTACT

GGACTACTAG

GGACTACTAG

solution: ACGTTGCTAGTTACTGGGACTACTAGATGACCTTCTATTCCC

- Can you find the original sequence?
- How long is the sequence? 42BP
- What is the average coverage? 3.93
- Is average coverage a good indicator of data quality? C = 30 for human genome is good

Assembly against a reference sequence: Distinguishing **S**ingle **N**ucleotide **P**olymorphisms (SNPs) from sequencing errors

Reference Sequence

```
GTTTCAACCACGTTGCTAGTTACTGGGACTACTAGATGACCTTCTATTGTATCAACCT
```

Sequencing data from individual

```
GTTTCAACCACGTT      TACTGGGACTACTGG
  TTCAACCACGTTGCTAG  ACTGGGACTACTGGAT        ATTGTATCAACCT
      GCCACGCTGCTAGT   TAGGACTAATGGA       CTATTGTAGCAACC
      ACCACGTTGCTAGT ACTGGGACTACTG       TTATATTGTATCA
          CGTTGCTAGTTACTGGG   ACTGGATGACCTTCTA
  TTCAACCACGCTGC         GGACTACTGGATGACC      TGTATCAACCT
   TCAACCACGCTG             CTACTGGATGACCTTC  GTGTATCAACCT
 TTTCAACCACGC GCTAGTTACTGGGAC ACTCGATGACCTTCT   GTATCAACCA
```

heterozygous individual

this last one can be misalligned at the end or at the beginning

- How many nucleotides in the aligned sequence reads differ from the reference? 19
- What type of deviations do we see? substitions
- What other deviations would be possible? inversion, deletion, insertion
- Are those deviations sequencing errors or what other explanation could there be?
- Can you imagine a situation where sequencing errors and alignment choice interact?
- How could you estimate the sequencing error rate from the data you have?

unaccounted devation/ nt sequenced = 2%

Here are the short-read data of two patients with neurological problems assembled to the reference sequence of the HTT gene. Do you find evidence for a possible mutation?

human reference genome

```
GTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCCGCCACCG
```

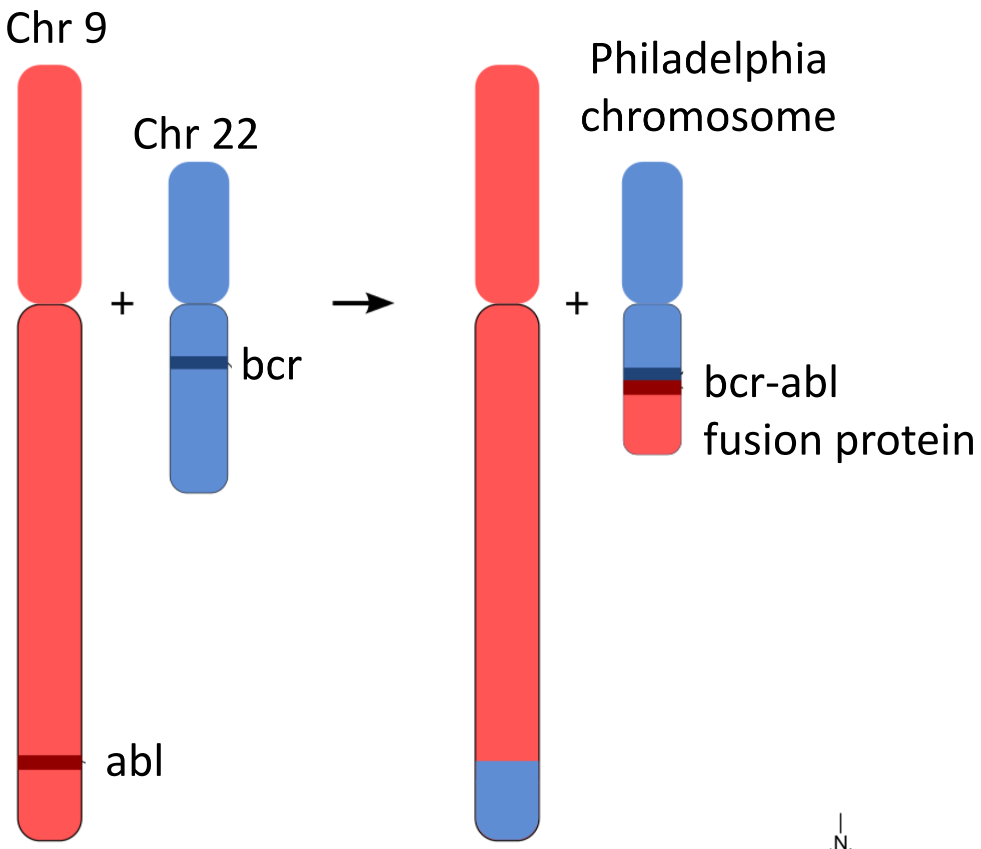
patient A

```
            GCAGCAGCAGC
  CCTTCCAGCAGCA
           AGCAGCAGCAGCAGCA
     TCCAGCAGCA     AGCAGCAGCAGCAG AGCAGCAGCAGCA             CAGCAGCAGCCGCCA
GTCCTTCCAG              GCAGCAGCAGCAGCAGC                  AGCAGCAGCAGC
GTCCTTCCAGC GCAGCAGCAGCAGCA    CAGCAGCAGCAGCAGCA CAGCAGCAGCAGCA     AGCCGCCACCG
                       AGCAGCAGCAGCAGCA  AGCAGCAGCAGCAG         CAGCAGCCGCCAC
           AGCAGCAGCAGCAGCAG AGCAGCAGCAGCAGCAGCA CAGCAGCAGCAG
                                                      AGCAGCAGCAGCA
                                                          GCAGCAGCAGC
```

patient B

```
            GCAGCAGCAGC
  CCTTCCAGCAGCA
           AGCAGCAGCAGCAGCA
     TCCAGCAGCA     AGCAGCAGCAGCAG AGCAGCAGCAGCA             CAGCAGCAGCCGCCA
GTCCTTCCAG              GCAGCAGCAGCAGCAGC                  AGCAGCAGCAGC
GTCCTTCCAGC GCAGCAGCAGCAGCA    CAGCAGCAGCAGCAGCA CAGCAGCAGCAGCA     AGCCGCCACCG
                       AGCAGCAGCAGCAGCA  AGCAGCAGCAGCAG         CAGCAGCCGCCAC
           AGCAGCAGCAGCAGCAG AGCAGCAGCAGCAGCAGCA CAGCAGCAGCAG
                 AGCAGCAGCAGCAG AGCAGCAGCAGCA         AGCAGCAGCAGCA
                     GCAGCAGCAGCAGCAGC                    GCAGCAGCAGC
         GCAGCAGCAGCAGCA    CAGCAGCAGCAGCAGCA CAGCAGCAGCAGCA
                    AGCAGCAGCAGCAGCA   AGCAGCAGCAGCAG
                          AGCAGCAGCAGCAGCAGCA CAGCAGCAGCAG
                           AGCAGCAGCAGCAG AGCAGCAGCAGCA
                               GCAGCAGCAGCAGCAGC
                   GCAGCAGCAGCAGCA    CAGCAGCAGCAGCAGCA CAGCAGCAGCAGCA
                              AGCAGCAGCAGCAGCA  AGCAGCAGCAGCAG
                                    AGCAGCAGCAGCAGCAGCA CAGCAGCAGCAG
```

Detecting large chromosomal rearrangements

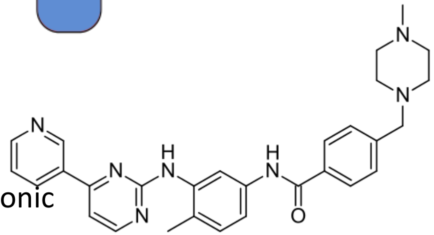


Chromosome translocation creates the new fusion protein *bcr-abl* – a permanently activated tyrosine kinase that causes cancer (chronic myelogenous leukemia).

How would we see such a rearrangement in 2nd gen sequencing data?

Why might we miss it?

What type of sequencing data, or other data could we use to detect this rearrangement?



Gleevec (drug to treat chronic myelogenous)leukemia