Taq polymerase errors in PCR: Frequency and management

Xavier Foissac, UMR Fruit Biology and Pathology,
INRA and University of Bordeaux

Taq polymerase

- DNA polymerase purified from *Thermus aquaticus* a bacterium living in hot springs
    - replicates DNA by incorporating dNTPs on 3' OH end on a primer hybridized to a DNA matrix
    - Mg^{2+} is a co-factor
    - Optimal temperature for activity 72 °C
- Purified in 1976 by Chien and colleagues
- Nowadays purified from recombinant *Escherichia coli*
- Used in PCR due to its thermostability Saiki, R. K., Scharf, S., Faloona, F., **Mullis, K. B.**, Horn, G. T., Erlich, H. A., and Arnheim, N. 1985. Science
- Replication rate of 35-100 nucleotides per second (Wittver *et al.*, 1991 Biotechniques)

Accuracy of polymerases

- Accuracy of polymerase = number of adequate nucleotide incorporated / total nucleotides
    - Also called fidelity
- Error frequency = number of misincorporated nucleotide / total nucleotides
    - ranges from 10^{-6} (high fidelity) to 10^{-4} (low fidelity) per incorporated nucleotide
    - Measured by reversion of mutants (opal lacZ mutant for example) or sequencing
- Important for cloning error-free coding sequences for heterologous expression or for variability studies

Errors of *Taq* polymerase: frequency and distribution

- Well studied by **Chen and colleagues** (**1991, Mutation Research**)

- Topic: mutations in the human adenine phosphoribosyltransferase genes
  = HPTR deficiency that cause a kidney disease

1. Cloning and sequencing of HPRT gene from human DNA library (reference sequence)
2. Cloning of HPRT PCR products and sequencing of 5 clones per patient from 5 patients

3. The five independent sequences showed discrepancies : 44 for 58 kbp when compared between each others **= errors introduced by PCR**

4. No errors were found in 5 out of 25 clones sequenced, one clone contained 5 errors

Errors of *Taq* polymerase: frequency and distribution

6. As PCR were of 30 cycles, **absolute error frequency was 2.5 10^{-5}** per nucleotide leading to 0.76 errors per kb after 30 cycles, 4 times lower than that reported by Tindall and Kunkel (1988, Biochemistry)

7. Statistical analysis showed that occurrence of errors followed the Poisson distribution (Khi2 = 0.892, *P*=0.05)

8. **In addition errors were randomly distributed**

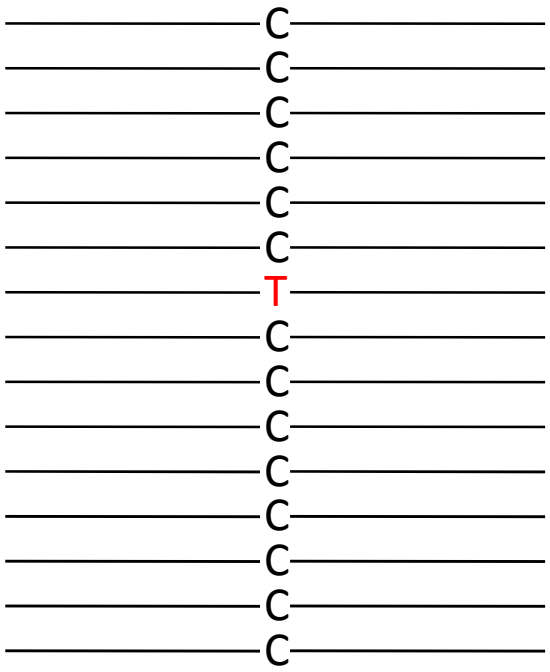
9. All were substitutions: 38/44 were transitions (T→C, A→G or C→T, G→A), 6 were transversions (A →C, A→T, C→A, C→G).

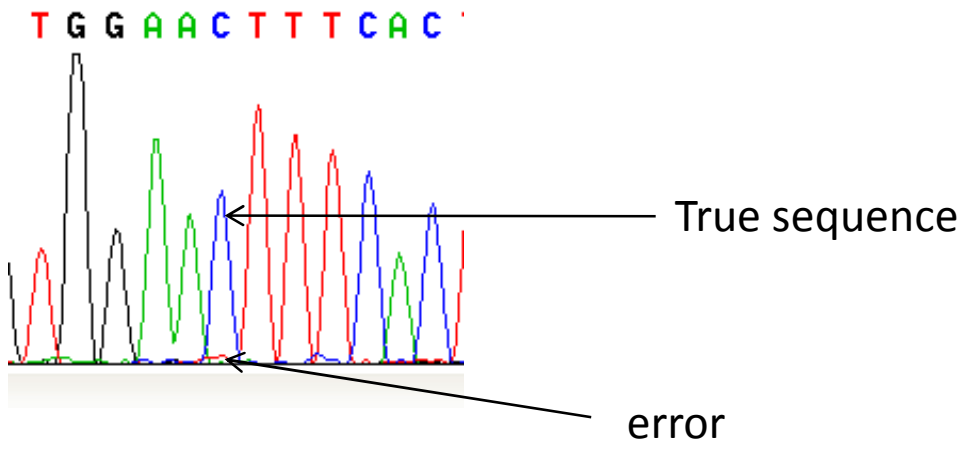
10. No insertion or deletion observed but reported in other studies at low frequency

Management of Taq polymerase errors

1. Solution : direct sequencing of PCR products

- Errors randomly distributed : for a given base the majority of nucleotides correspond to the true sequence → errors are in the background of sequencing

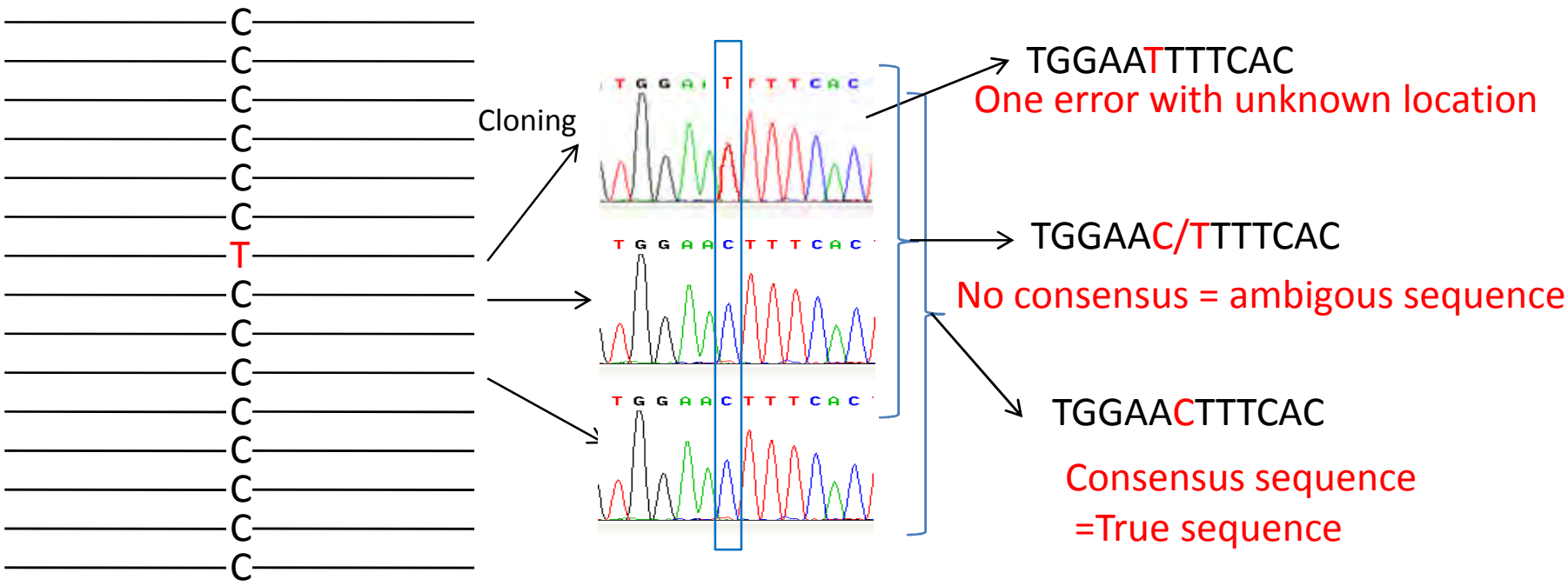




Management of Taq polymerase errors

2. Due to bad sequencing or double pics in sequence (mixte infection) : cloning is unavoidable

- Random cloning : each insert of about 1,000 bp will contain on average one error
- The consensus of n sequences will constitute the true sequence (n >2)



Management of Taq polymerase errors

Practical training : what is the true sequence of the following insert sequences obtained after nested PCR and cloning into an *E. coli* plasmid ?

TGTTAATCAGATACCTAGGGATACTACAGTT
TTTAAATCAGATACCTAGGGATACTAGAGTT
TTTTAATCAGATTCCTAGGGATACTAGAGTT
ATTTAATCAGATACCTAGGGATACTAGAGTT
TTTTAATCAAATACCTAGGGATACTAGAGTT
TTTTAATCAGATACCTAGGGATACTAGAGTT
TTTTACTCAGATACCTAGGGATACTAGAGTT
TTTTAATCGGGTACCTAGGGATACTAGAGTT
TTCTAATCAGATACTTAGGGATACTAGAGTC
TTTTAATCAGATACCTAGAGATACTAGAGTT
TTTTAATCAGATACCTAGGGATACTAGAGTT
GTTTGATCAGATACCTAGGGATACTAGAGTT
TTTTAATCAGATACCTAGGGATACTAGATTT

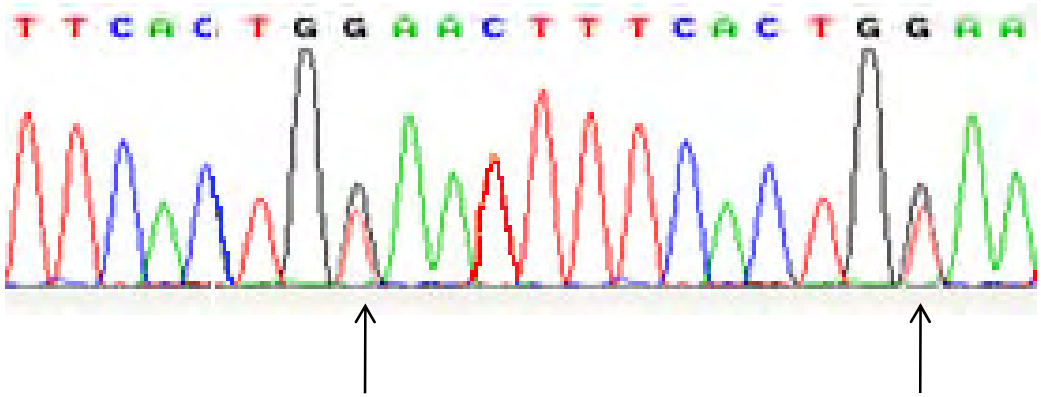
Management of Taq polymerase errors in population studies

- Mixed infections or two copies of a gene, like divergent 16S rDNA genes

One population of 2 variants
with relative incidence of 50 % each

Errors of *Taq* polymerase

Direct sequencing

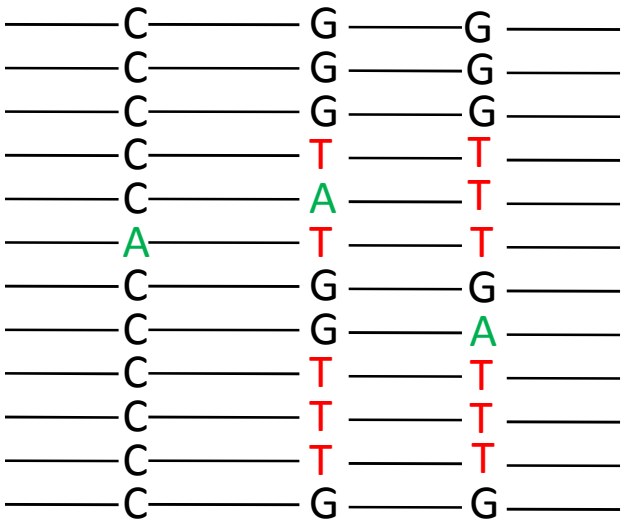


Double pics → Mixed infection

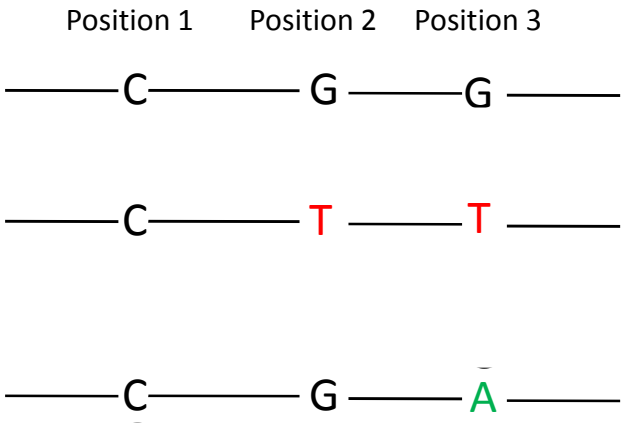
Management of Taq polymerase errors in population studies

- Mixed infections or two copies of a gene (variants), like divergent 16S rDNA genes

One population of 2 variants
with relative incidence of 50 % each



Cloning



Errors of *Taq* polymerase

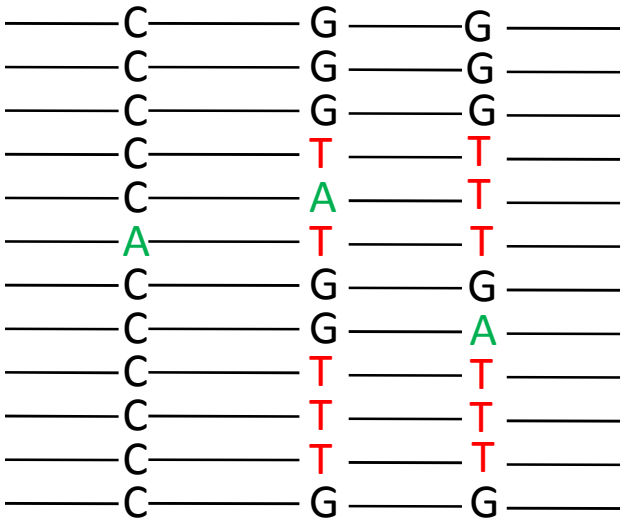
Conclusion ?

Position 1 : C =true
Position 2: G= true, T = error or variant ?
Position 3: G/T/A error or variant ?

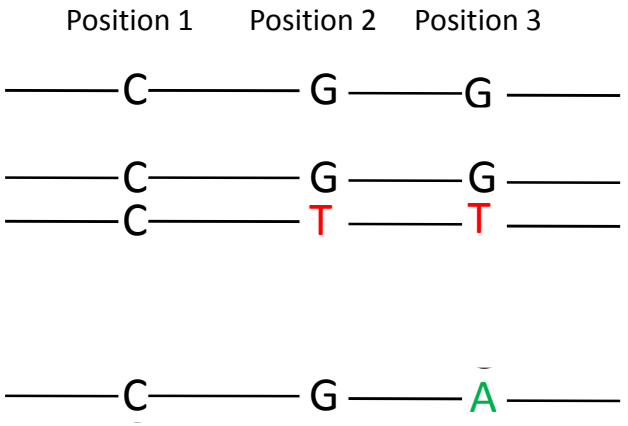
Management of Taq polymerase errors in population studies

- Mixed infections or two copies of a gene, like divergent 16S rDNA genes

One population of 2 variants
with relative incidence of 50 % each



Cloning



Errors of *Taq* polymerase

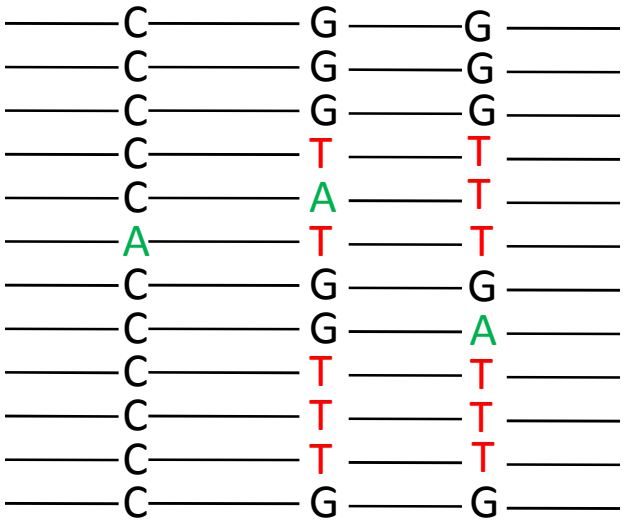
Conclusion ?

Position 1 : C =true
Position 2: G= true variant, T = error or variant ?
Position 3: G= true variant, T/A error or variant ?

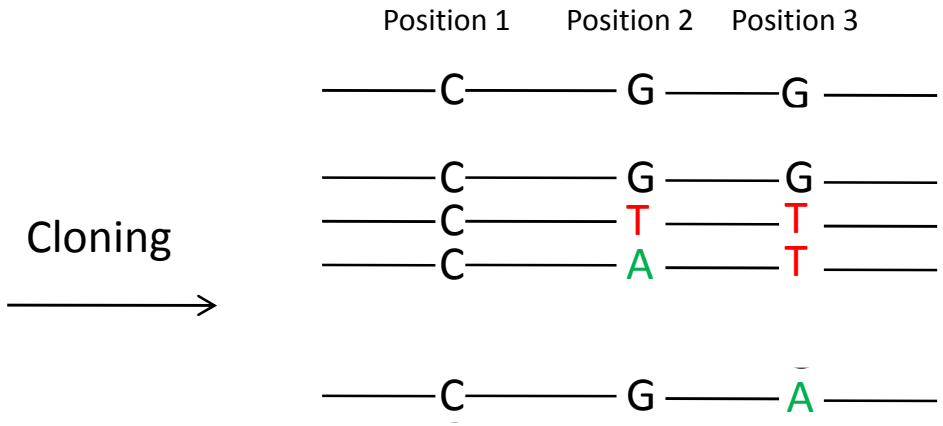
Management of Taq polymerase errors in population studies

- Mixed infections or two copies of a gene, like divergent 16S rDNA genes

One population of 2 variants
with relative incidence of 50 % each



Cloning



Errors of *Taq* polymerase

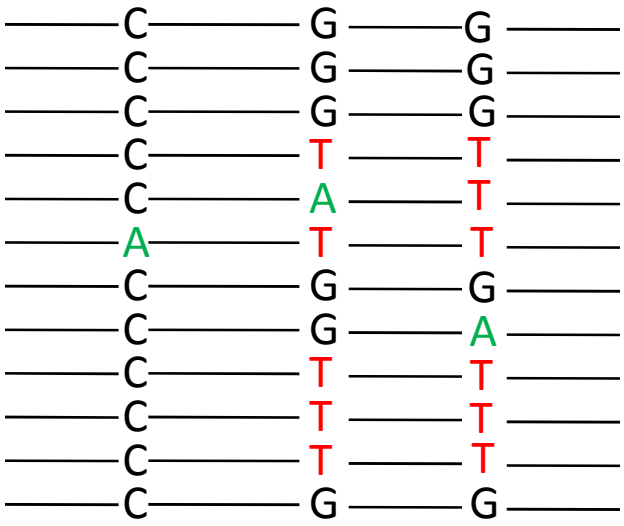
Conclusion ?

Position 1 : C =true
Position 2: G= true variant, T/A = error or variant ?
Position 3: G and T = true variants, A = error

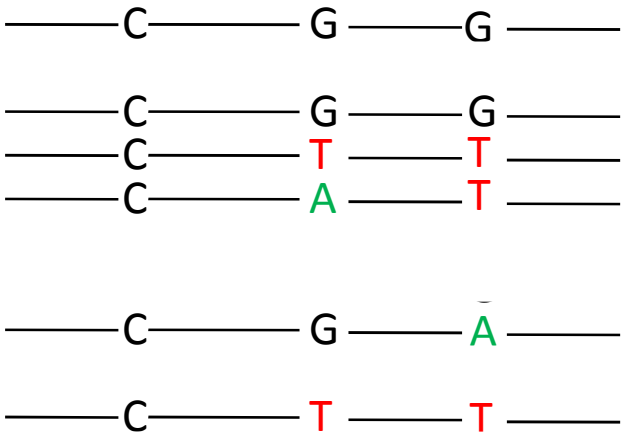
Management of Taq polymerase errors in population studies

- Mixed infections or two copies of a gene, like divergent 16S rDNA genes

One population of 2 variants
with relative incidence of 50 % each



Cloning



Errors of *Taq* polymerase

Conclusion ?

Position 1 : C=true
Position 2: G and T= true variants, A=error
Position 3: G and T= true variants , A=error

Management of Taq polymerase errors in complex populations

- If n variants of similar relative incidence in the population, probability to randomly sequence each of the variant follow the statistical law : Poisson distribution

→ $P_o = E^{-m}$

P_o is the probability that a variant sequence is missed
m is the number of times the variability needs to be covered

If P_o is 0.05 → 5 % chance to miss one variant, m=3 : so 3n sequences are required

- As 3 inserts need to be sequenced to resolve *Taq* polymerase errors,

→ 9 n sequences are needed to resolve all error-free variant sequences at P_O=0.05

- If relative incidence are not similar and a minor variant represent x % of the population, then n can be considered 100/x

Factors influencing the error frequency

1. Low $MgCl_2$ and dNTP concentrations reduce the error frequency
2. Increase of cycle number increase final error frequency :
   - errors in nested PCR are twice than PCR (approx. 1-2 errors per kbp)
3. Number of templates at the initial stage of PCR :
   - High number of template gives a reduced number of final number of errors
4. Presence or absence of a proofreading activity in the polymerase or in additional polymerase added to the reaction

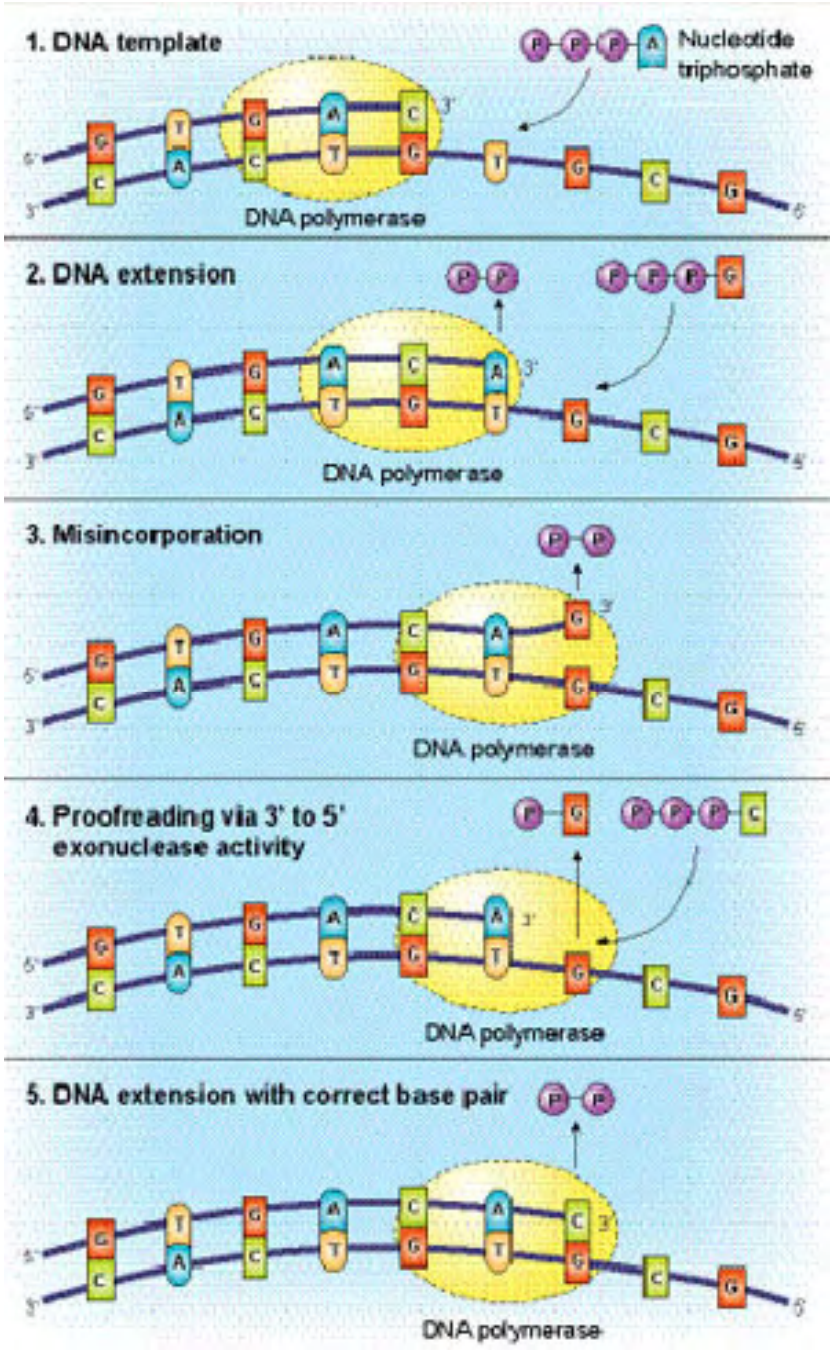
Proofreading activity: 3'→5' exonuclease

1. At stage 1 & 2 : normal 5' →3' polymerization

2. At stage 3 a wrong nucleotide is incorporated

3. At stage 4 the non hybridized nucleotide (mismatch) is recognized by 3'→5' exonuclease activity and the misincorporated nucleotide is removed

4. At stage 5 the adequate nucleotide is incorporated



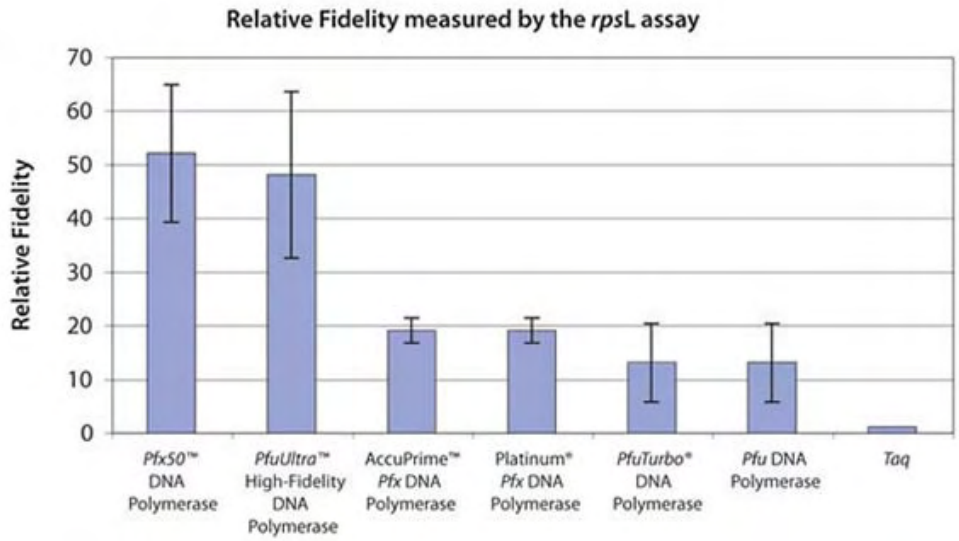
Examples of polymerases with proofreading activity

- DyNAzyme™ EXT DNA Polymerase (FINNZYME) : error rate 6 x10^{-7} per base

- *Thermococcus litoralis Vent* polymerase (INVITROGEN): error rate 6 x10^{-6} per base

- *Pyrococcus furiosus Pfu* polymerase : error rate 1.6 x10^{-6} per base (NEB)
  (Lundberg *et al.*, 1991, Gene)

- *Pfx 50* polymerase (PROMEGA): error rate 4 x10^{-7} per base



Case study: genetic diversity of alder yellow phytoplasmas in Europe (Malembic-Maher *et al.*, 2008, IOM congress)

Alder Yellows (AldY) phytoplasma. AldY is a frequent disease of *Alnus glutinosa* in Europe. It is transmitted by the leafhopper *Oncopsis alni*.

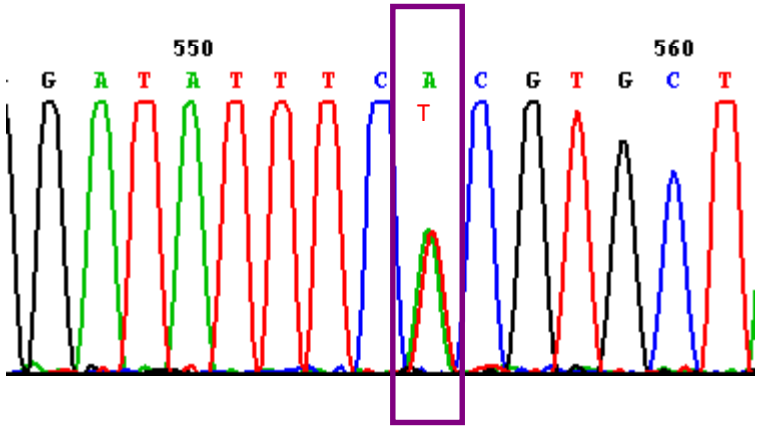
AldY

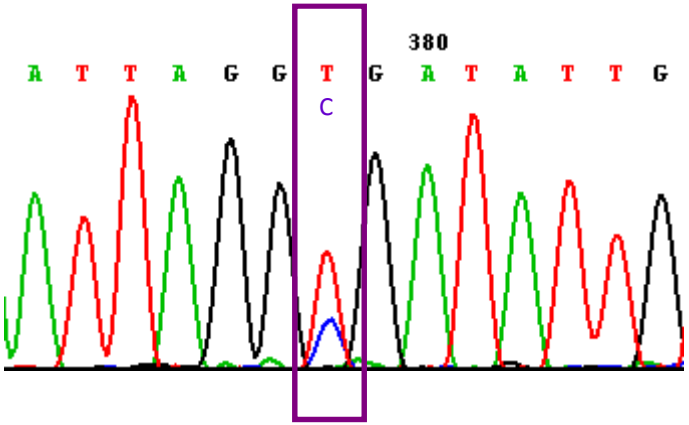
PGY

Characterization by nested PCR and sequencing of the *map* gene

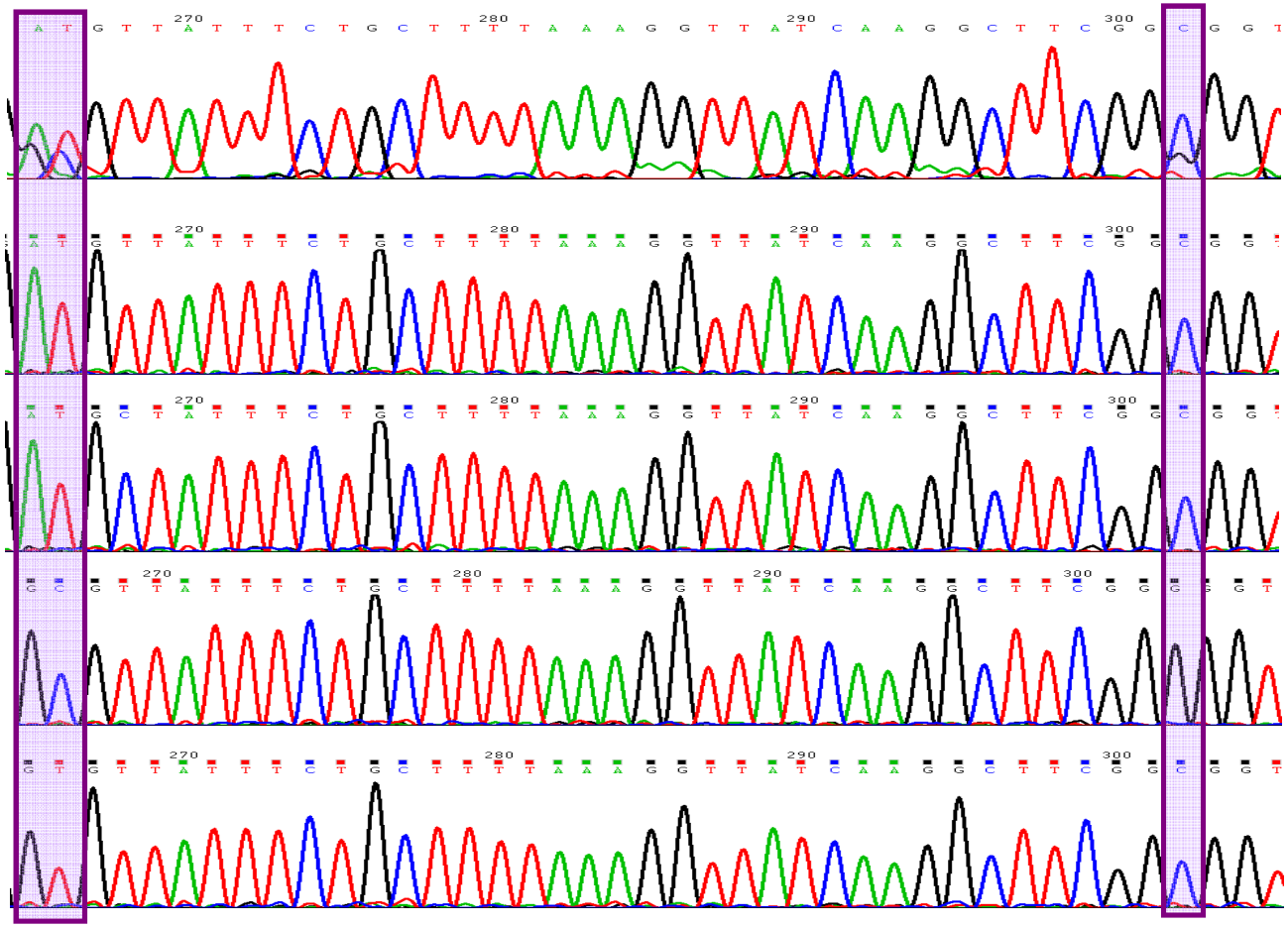
32 alder samples out of 55 with sequence ambiguities on chromatograms, superposition of nucleotide pics (up to 10 sites).

→ Reflects a mix of bacterial strains inside one tree.





- Amplification of gene ***map*** (673 bp) <u>in PCR of 25 cycles</u> with a proof reading polymerase (DyNAzyme™ EXT : error rate 6 x10^{-7} per base), cloning and sequencing of 4 clones



<u>Question:</u> what is the probability of having a polymerase error in a clone sequence ?

<u>Answer:</u> 6 x10^{-7} x 25 cycles x 673 bases = 0.01 1 fragment out 100 will contain an error