DNA Sequencing techniques

1. The "**Sanger" Technique**

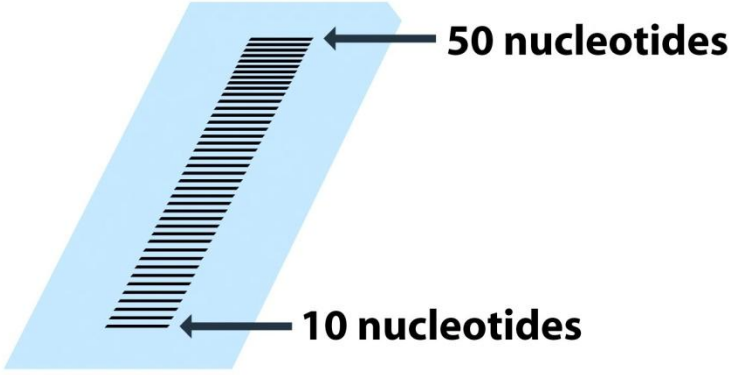
2. **2nd generation** sequencing (Illumina ™, 454 ™, Solid™)

3. **3rd generation** sequencing (Pacific Bio, Oxford Nanopore)

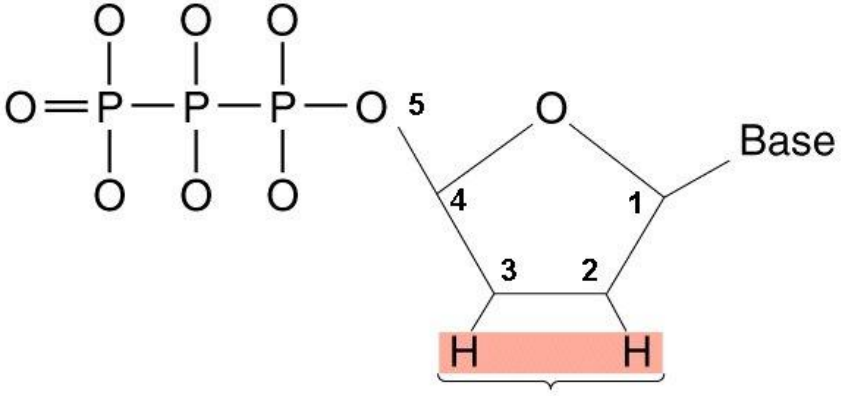
“Sanger” DNA sequencing

Based on two components

1. Electrophoresis on **polyacrylamide gel**



2. Use of **<u>dideo</u>xyribonucleotides** (**dd**NTPs)

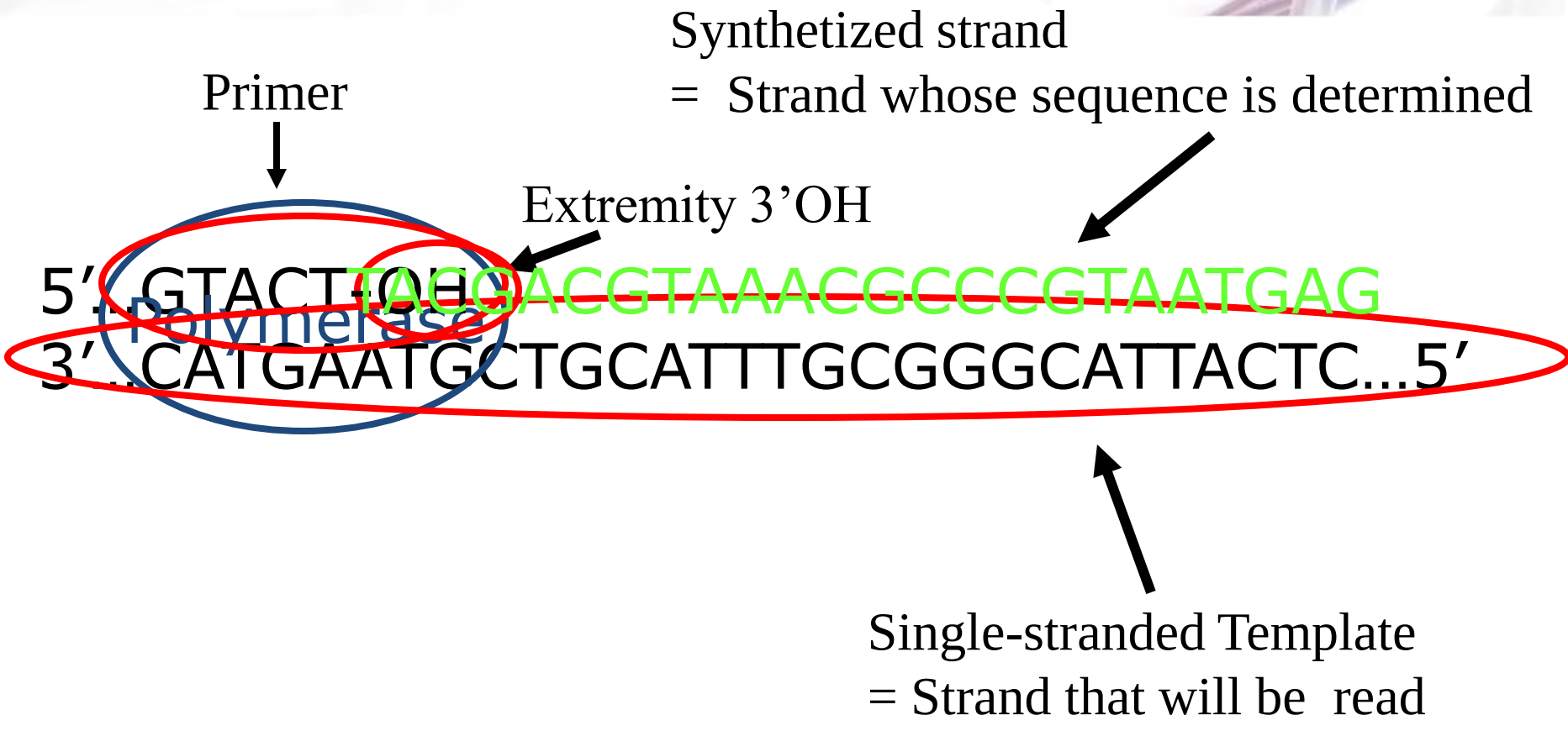


Sanger DNA sequencing: Principles

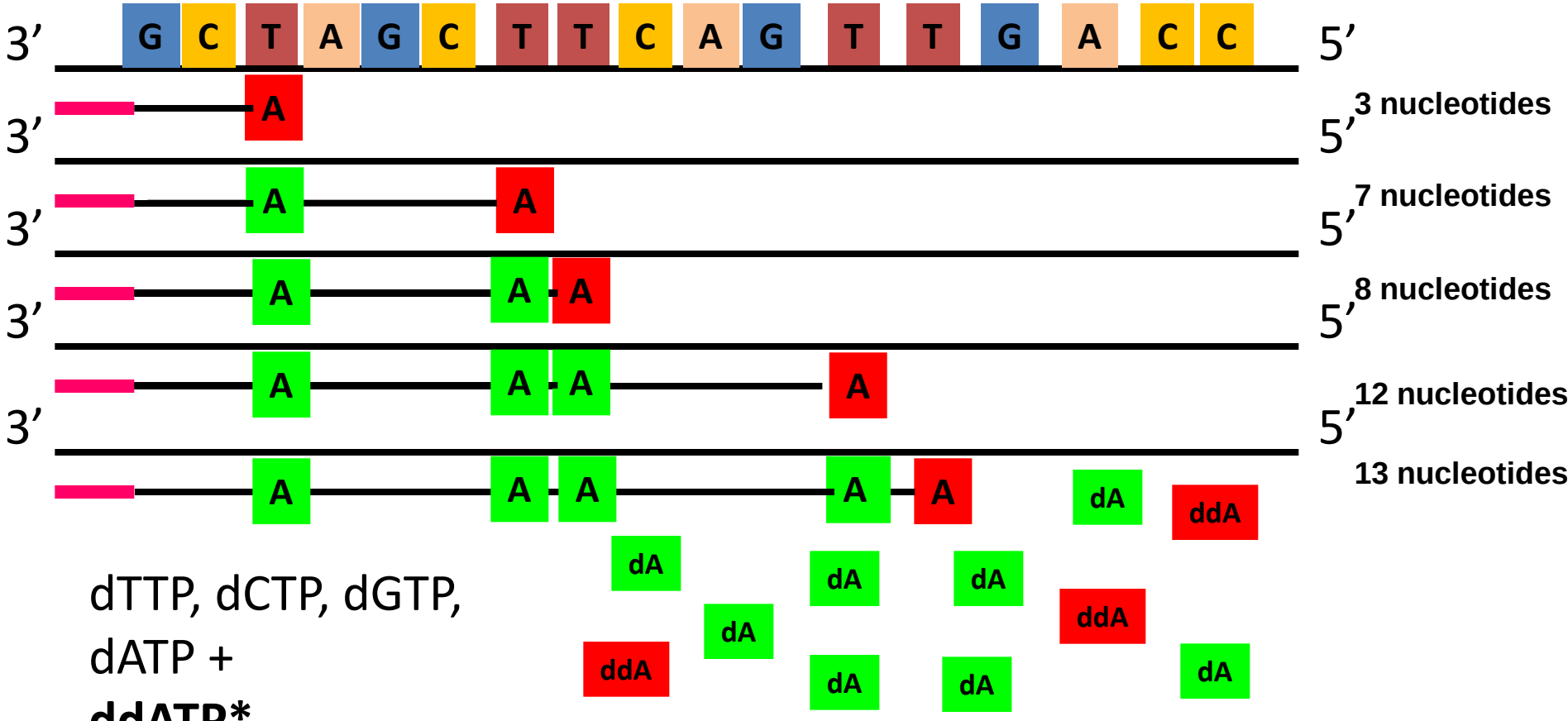
3 major differences with a « normal » PCR :

1. **One single** primer (oligonucleotide)
2. **Robust ADN polymerase** (ex : Sequenase)
3. **NTP mix → dNTPs** (large quantity) and **ddNTPs** (at lower concentration)

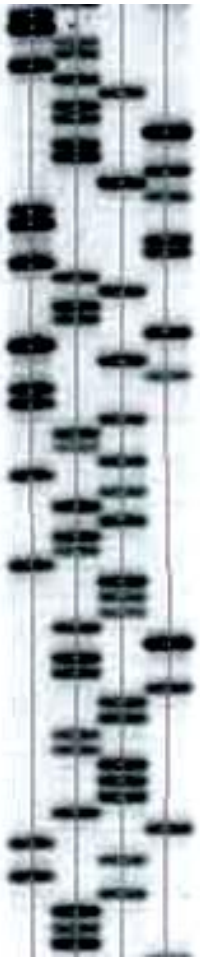
Sanger sequencing: Mechanism



Radioactive “Sanger” sequencing: Mechanism



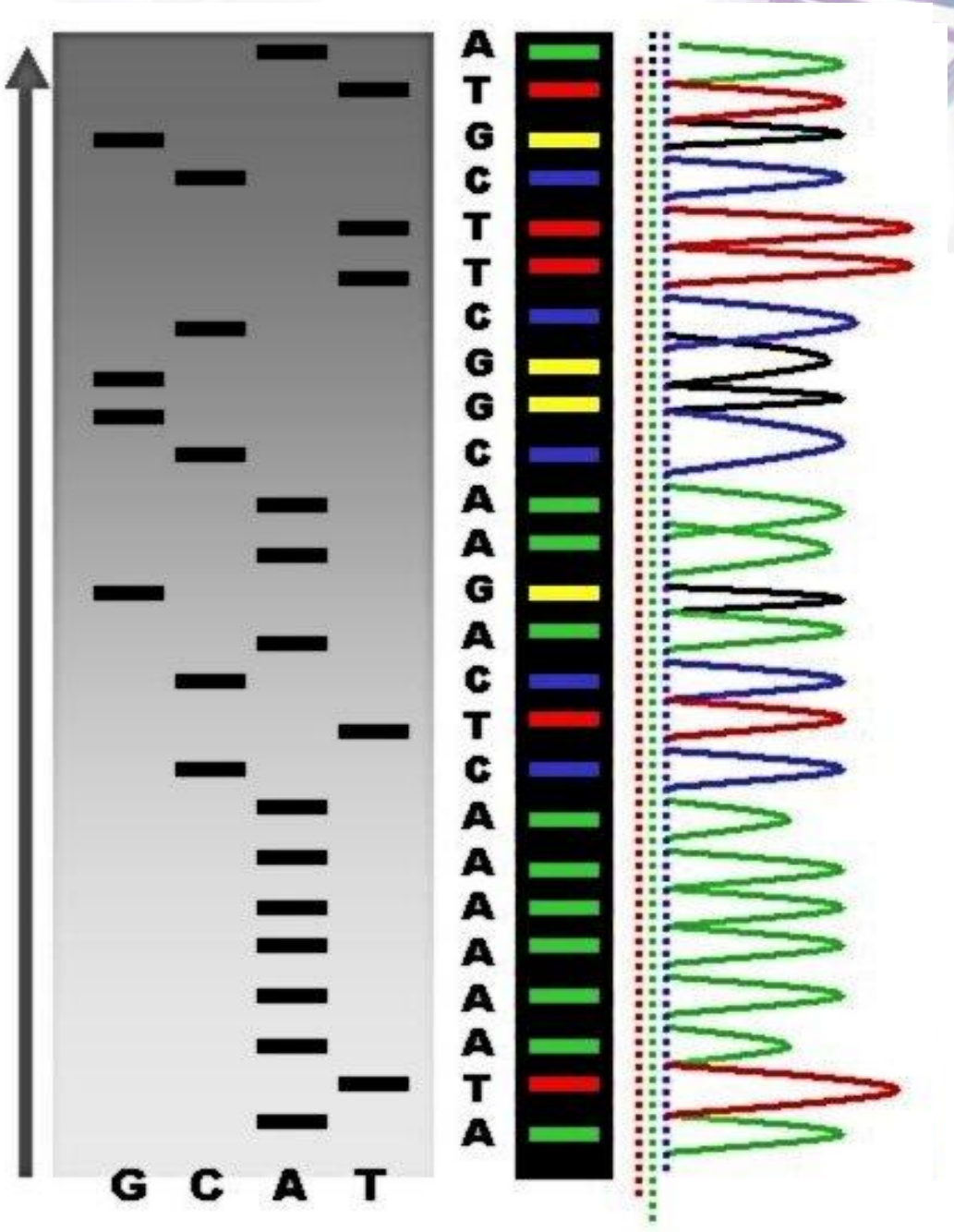
Radioactive “Sanger” DNA sequencing: Mechanism



- Same reaction is repeated for each nucleotide (ddTTP, ddGTP, ddCTP)
  - Each reaction will show the position of the complementary **dideoxyribo**nucleotides along the strand
  - The upper part of the gel represents the extremity of …
    - ? of the template
    - ? of the strand that has been synthetised

Fluorescent "Sanger" sequencing: Mechanism

- Mechanism = the same
  - The ddNTPs are labeled with different fluorophores (NOT the **dNTPs!**)
- Automation! → **DNA "sequencer"**



Alternatives to Sanger sequencing?

2nd generation DNA sequencing

- Pyrosequencing (*454™*)
- Polony sequencing (PCR colonies) (*Illumina™*)

3rd generation sequencing **(Not yet in the market)**
mostly based on single molecule sequencing

Pyrosequencing (454 ™)

Based **on Total DNA** sequencing

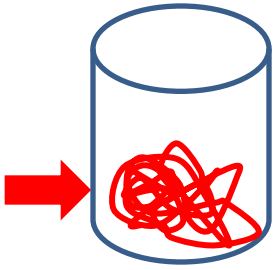
Follows 4 major steps

1. Break Total DNA (or cDNA) in small fragments
2. Ligate adapters to each newly produced fragment
3. Amplification of fragments by PCR
4. Parallel sequencing of each newly produced fragment **by luminescence** (Pyro)

How to prepare a 454 ™ library

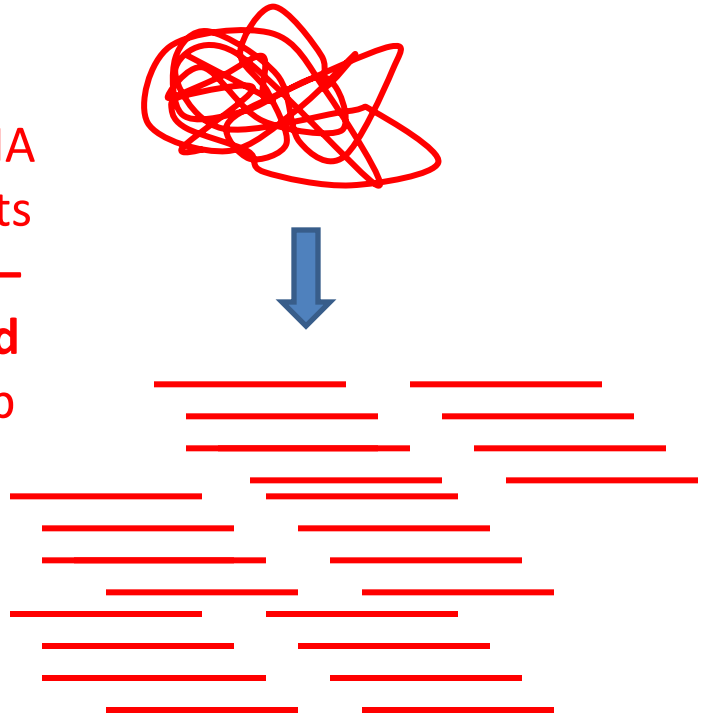
1. Break Total DNA (or cDNA) → Produce small fragments

A. DNA Extraction



B. Nebulization (or Sonication)

Small DNA fragments **Double – stranded** ~ 400 bp



Pyrosequencing (454 ™)

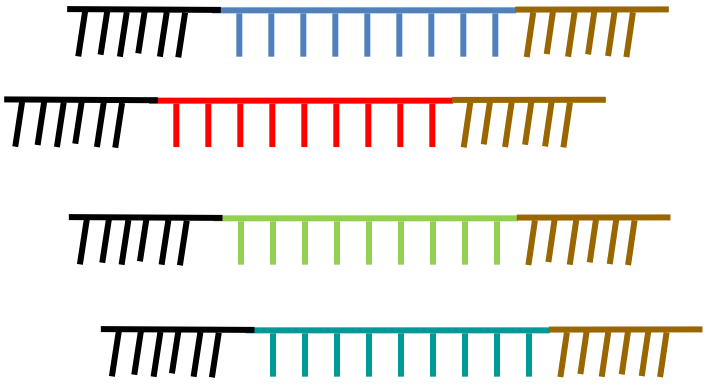
2. Adaptater's ligation

A. Double stranded fragments

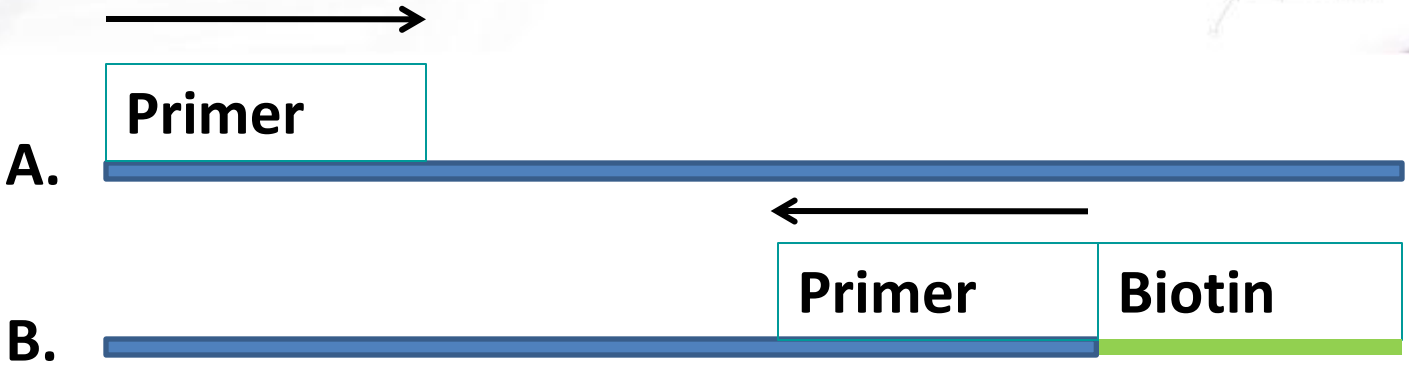
Ligase

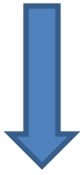
Adapters

B. Two adapters are ligated and denatured



The adapters





Attachment on streptavidin beads

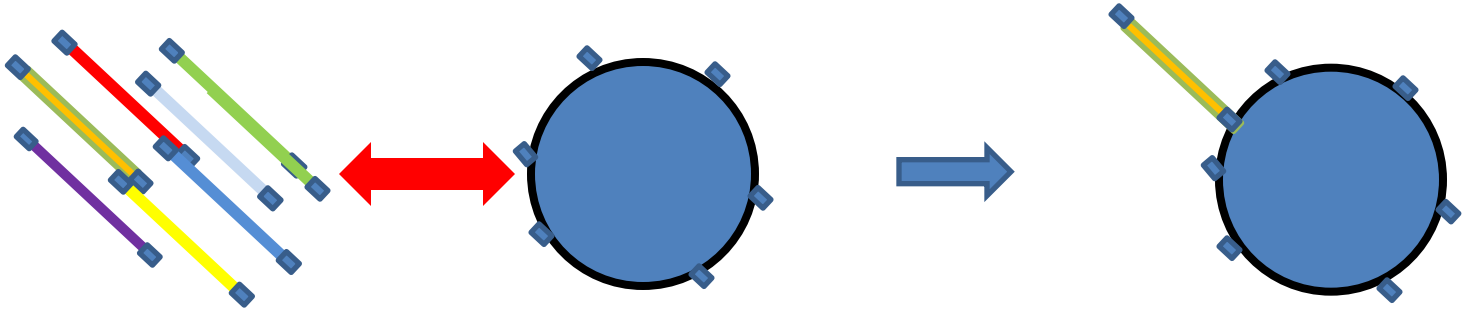
Emulsion PCR

Fragments with adapters

+

Streptavidin beads

One fragment per bead



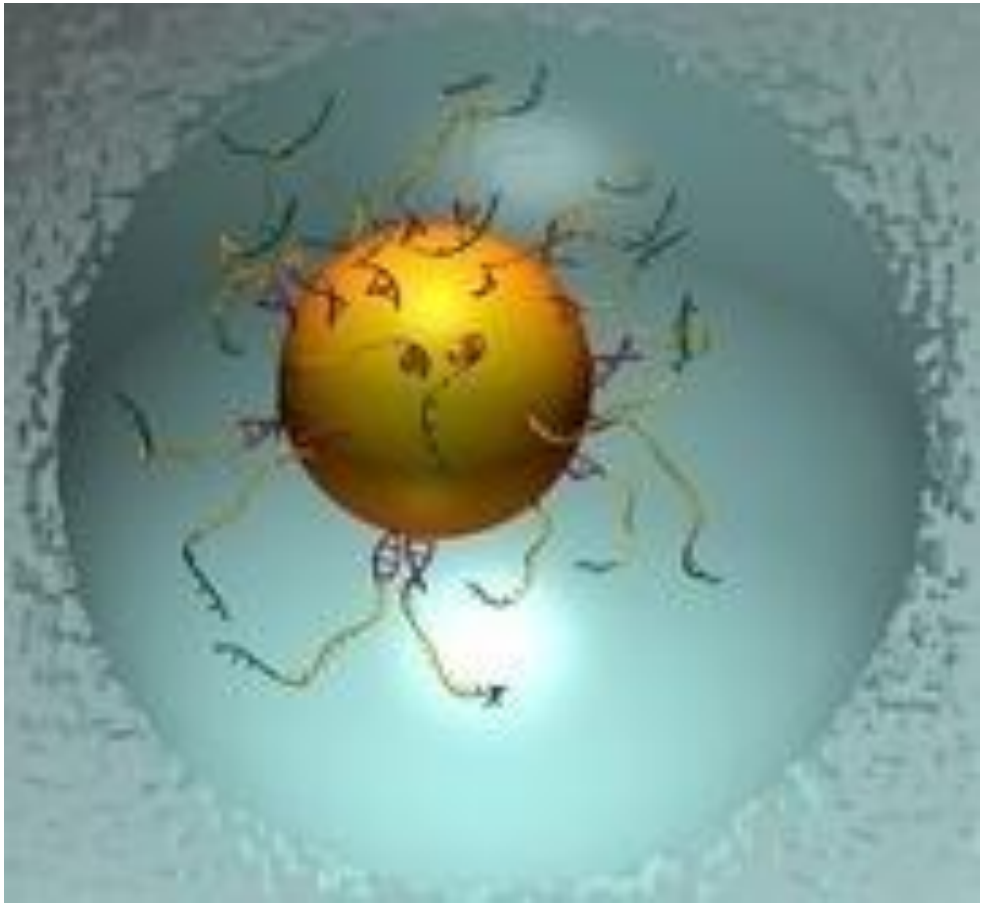
Mixed in a solution that contains Water, Oil, and all the reagents necessary to perform a PCR reaction

Water will form bubbles → PCR MICROREACTORS

Emulsion PCR

Each fragment is amplified on each bead (within each bubble)

Results in the cloning of millions of fragments, all attached on the same bead!

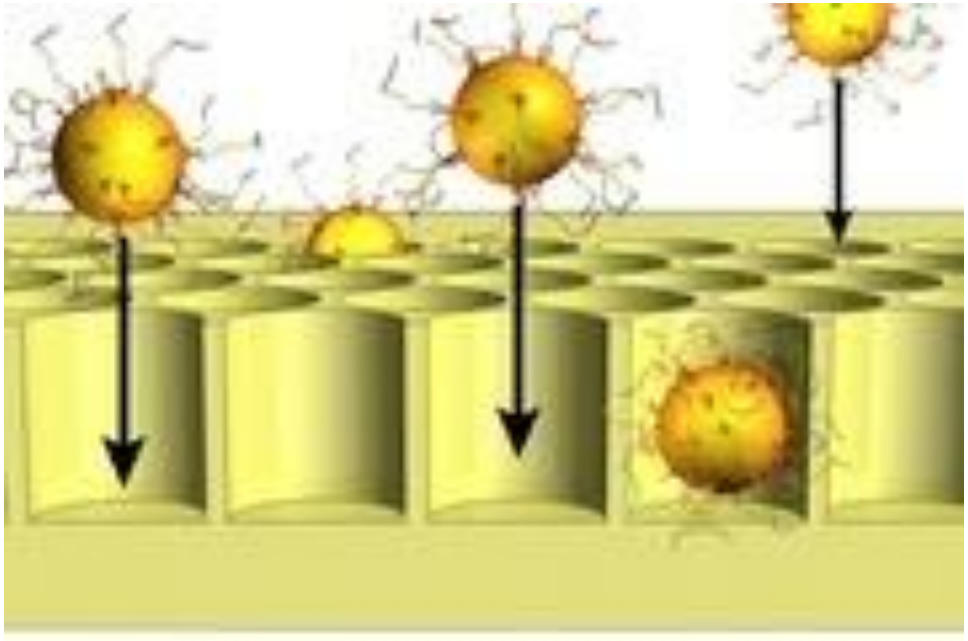


Pyrosequencing

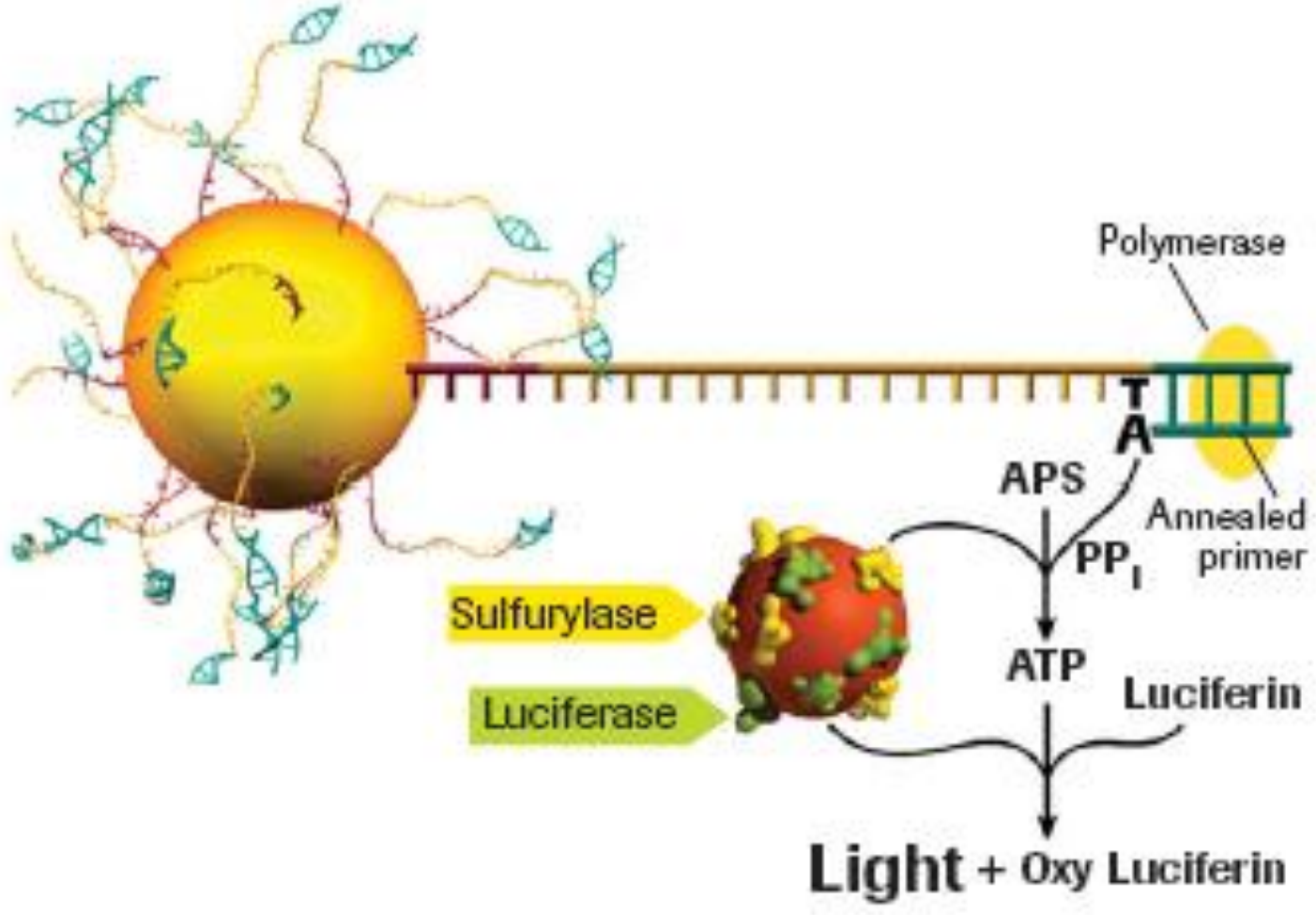
The beads are added into “micro-wells”

One bead per well!

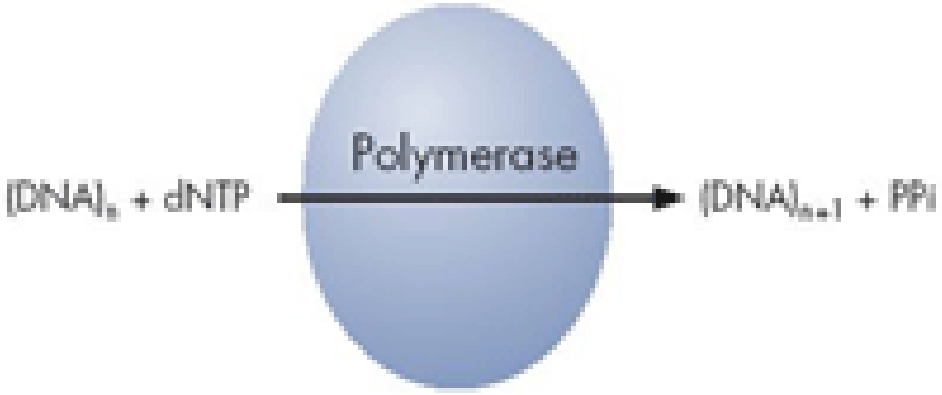
The different beads (i.e. the differents fragments) will be sequenced in parallel

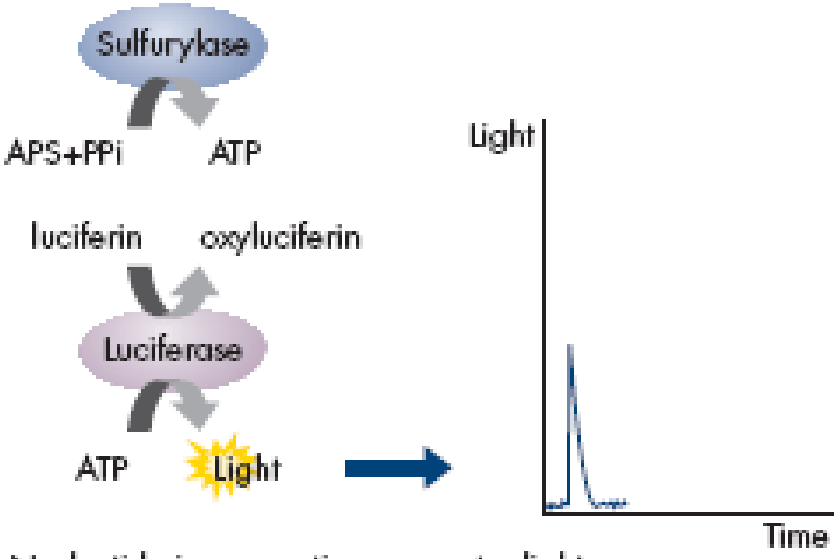


Pyrosequencing : process



Pyrosequençing : process





Nucleotide incorporation generates light seen as a peak in the Pyrogram trace

The DNA polymerase incorporates one complementary nucleotide to the fragments attached to one one bead

→ Release of <u>bi-phosphate</u>

The bi-phosphate is used by the Sulfurylase to produce l'ATP

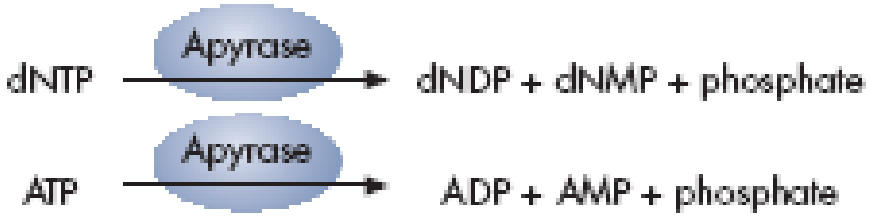
The Luciferase uses ATP to produce LIGHT (PYRO)

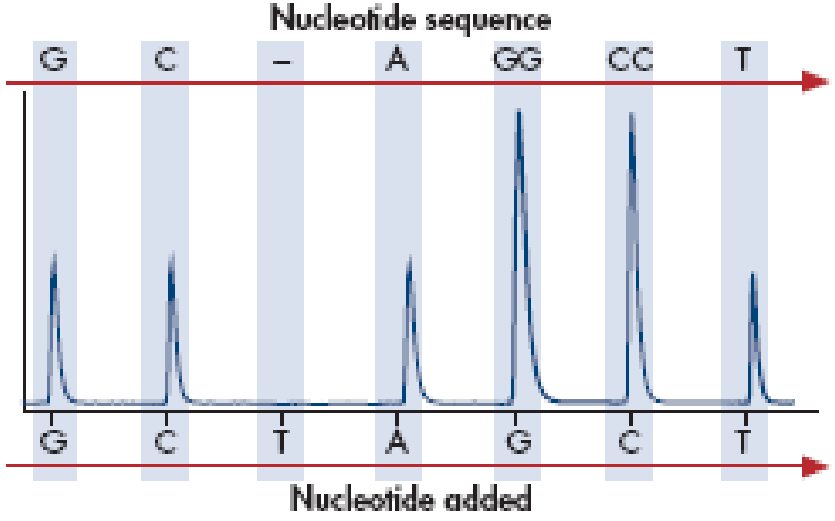
Pyrosequencing : process

The Apyrase degrades the nucleotides that have not been incorporated

The cycle continues, one nucleotide at a time!

The luminescent signal will be proportional to the number of nucleotides that have been incorporated





Pyrosequencing requires highly specialized instrumentation

Genome sequencer GS10 (Roche)

Utility of Pyrosequencing :

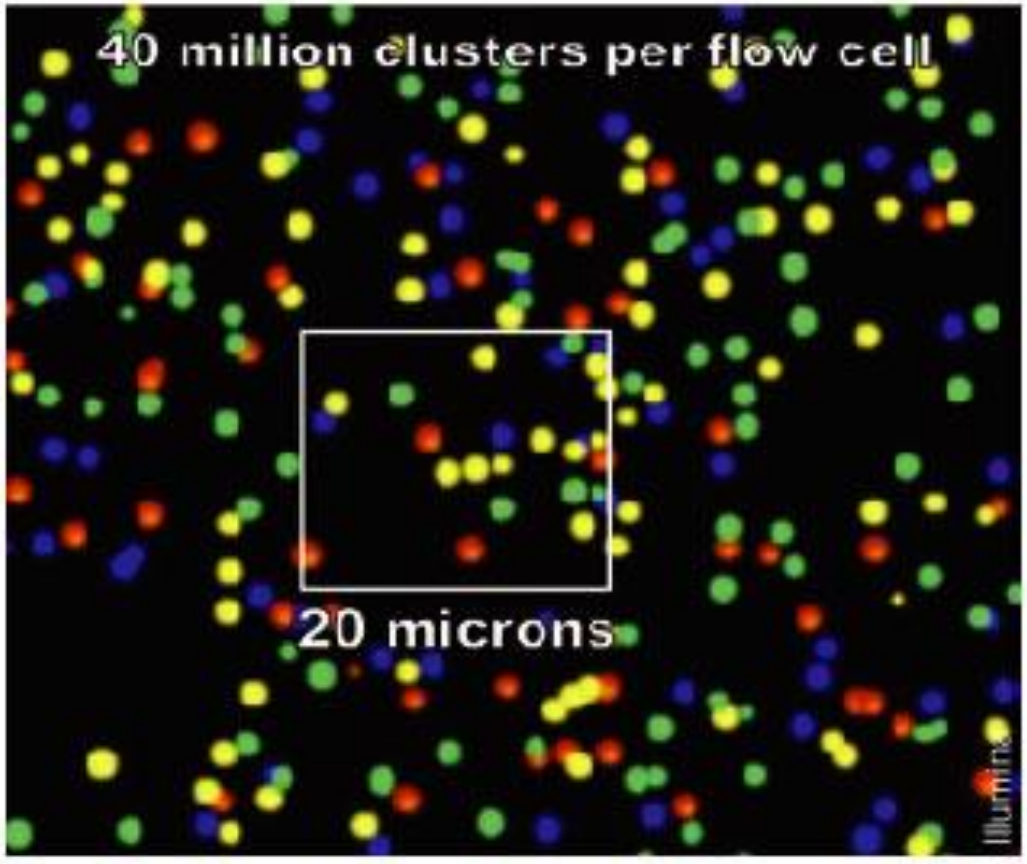
1. **Whole genome “de-novo” sequencing**
2. **Genome re-sequencing**
3. **Transcriptome sequencing**
4. **Parallel genotyping**

PCR colony sequencing (Illumina™)

Based on sequencing of **total DNA**, fragmented *a priori*

Also requires the **ligation of specific adapters**

Sequencing done on a solid surface



PCR colony sequencing (Illumina™)

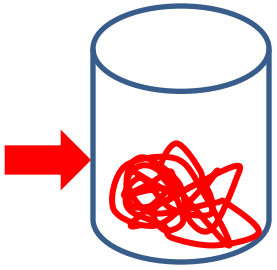
6 major steps

1. Total DNA (or cDNA) broken into very small fragments
2. **Double - Ligation of adapters** to each, newly produced fragment
3. Selection of fragments with a specific size (200bp)
4. Fragments attached on a **solid** surface
5. Solid-phase PCR → production of PCR colonies (polonies)
6. Fragment denaturation and sequencing **by fluorescence**

Illumina™ library production

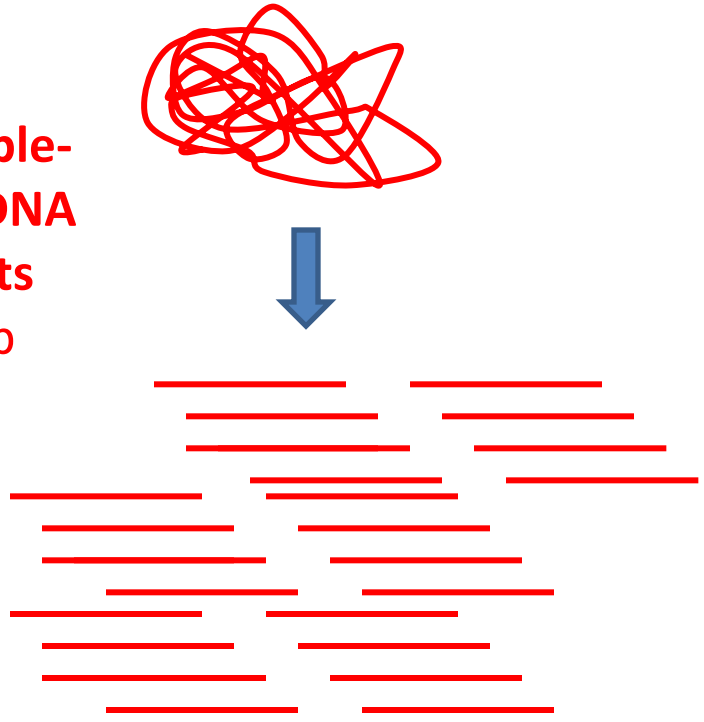
1. Total DNA (or cDNA) broken into small fragments

A. DNA extraction



B. Nebulization (or Sonication)

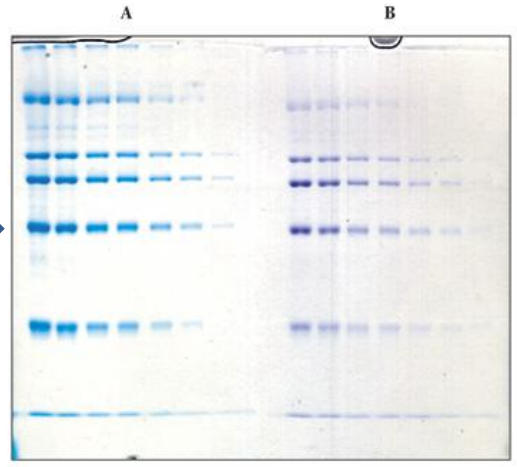
Small **double-stranded DNA fragments**
~ 200 pb



Illumina™ library production

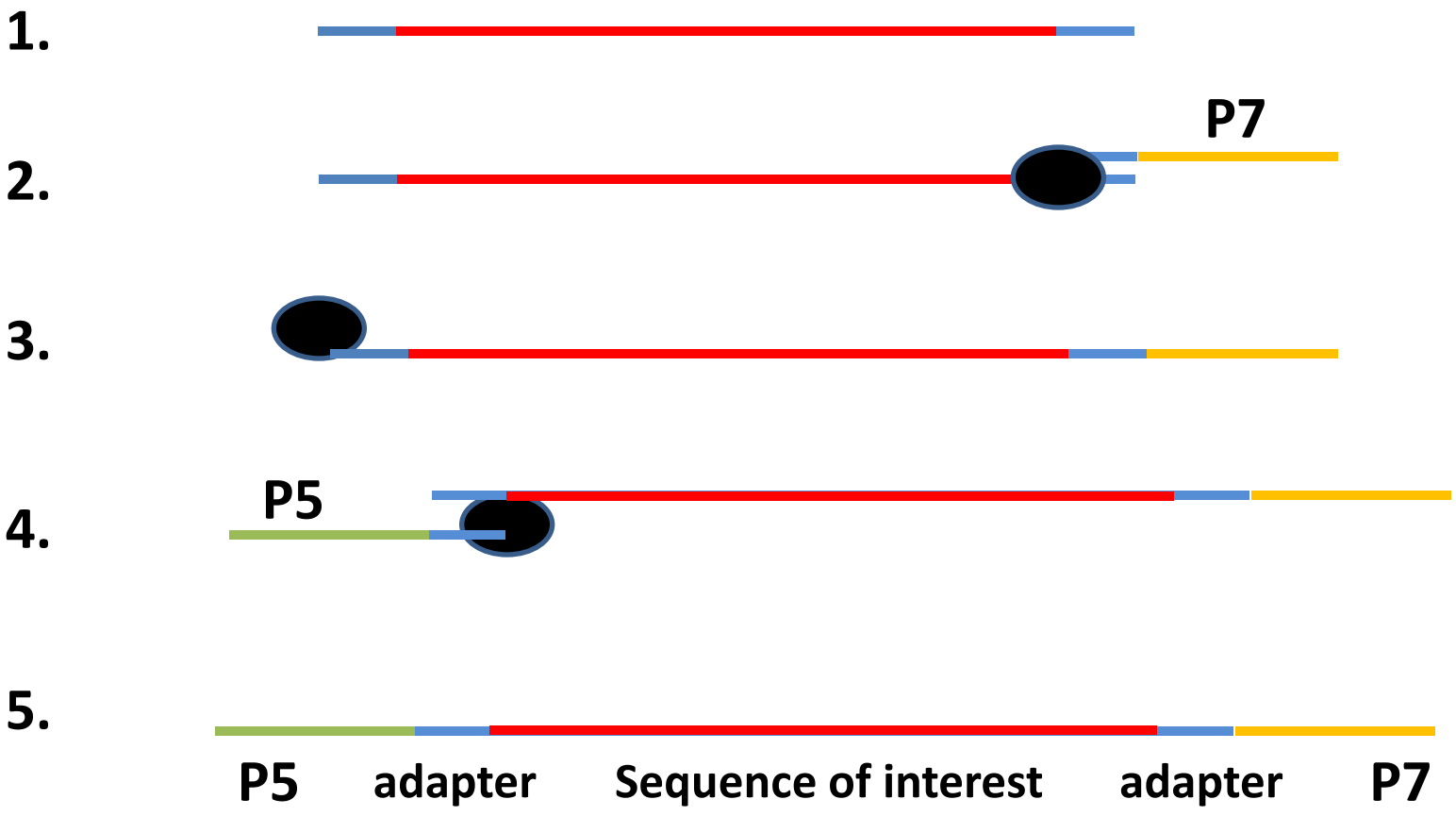
2. Ligation of “UNIVERSAL” **Illumina adapters**

3. Ligated fragments of a certain size
are selected selected

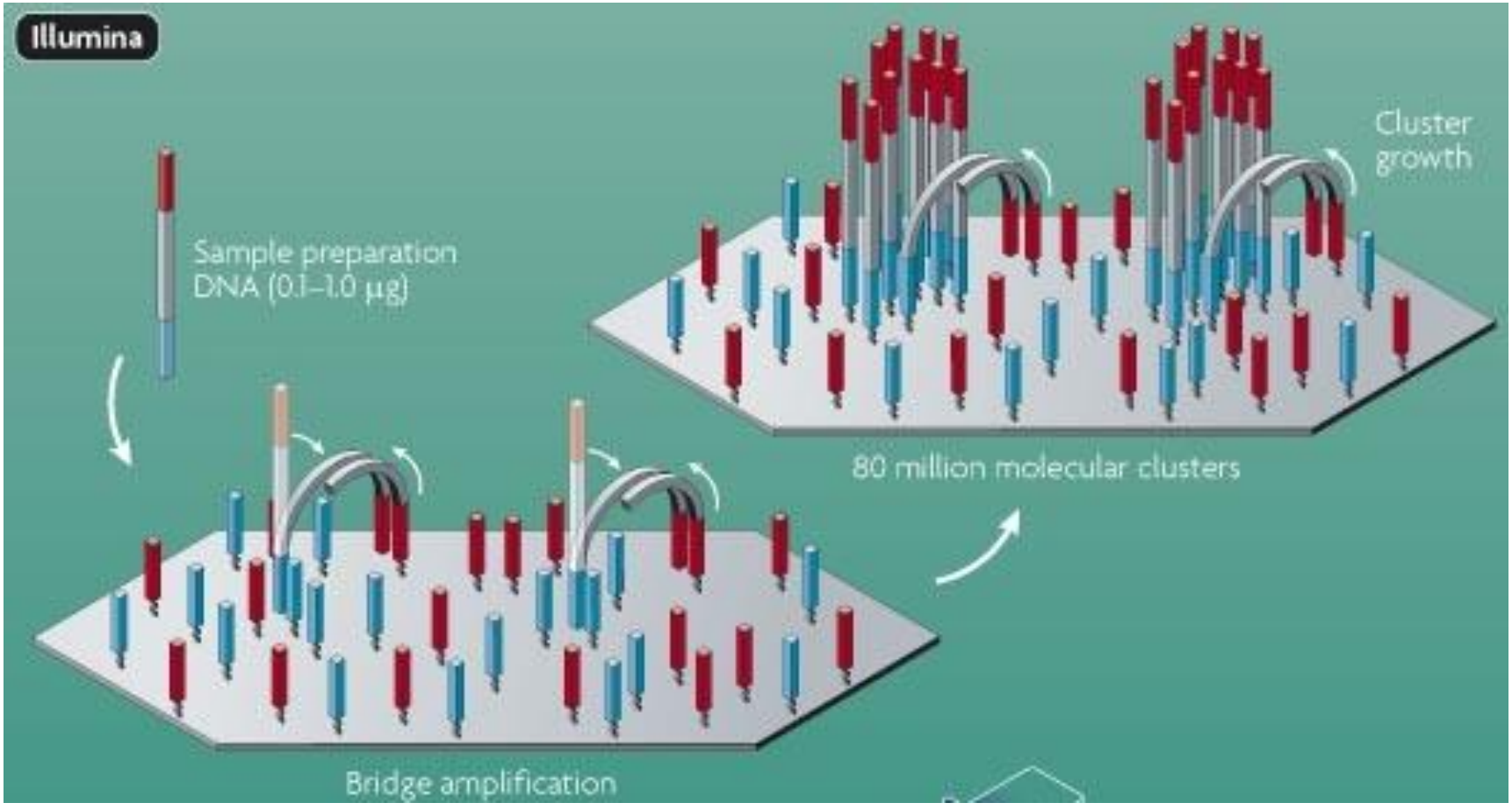


4. Addition of **2 new adapters** by PCR

New adapters added by PCR

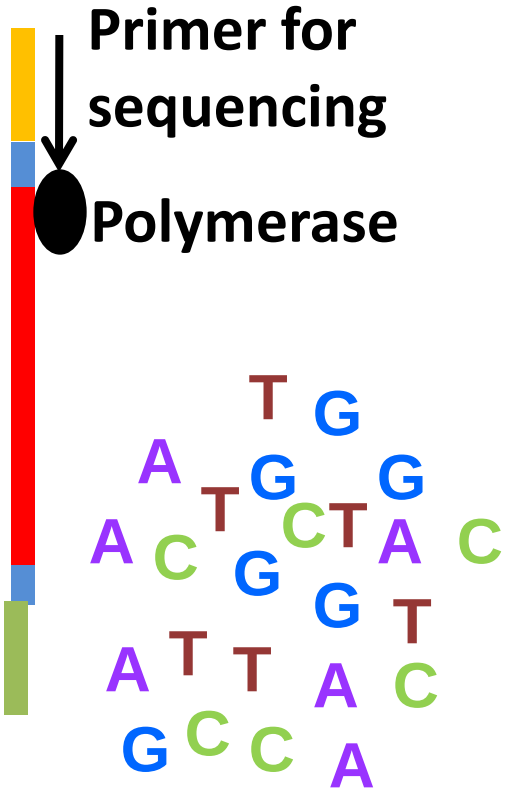


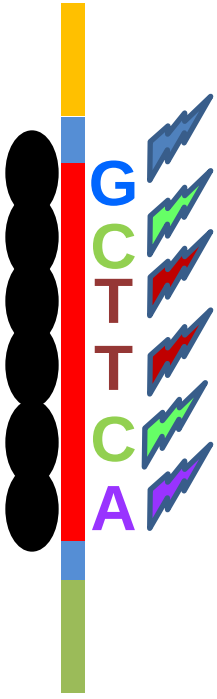
Attachment of fragments on solid surface (Glass)

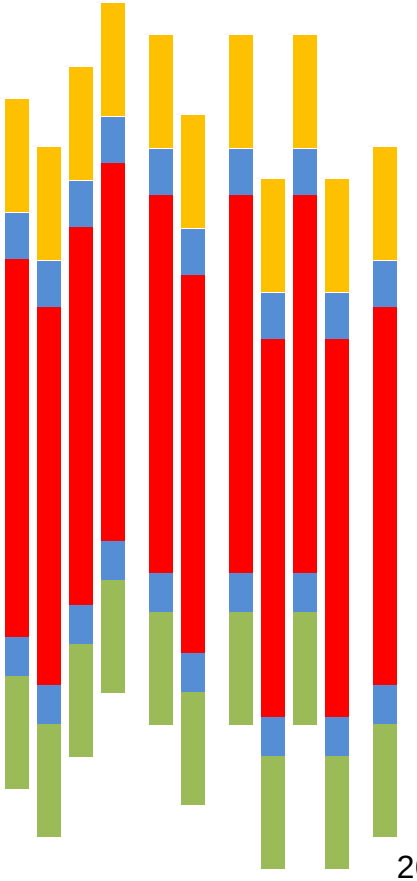


Illumina™ Sequencing

Clusters amplifies the Fluorescent signal

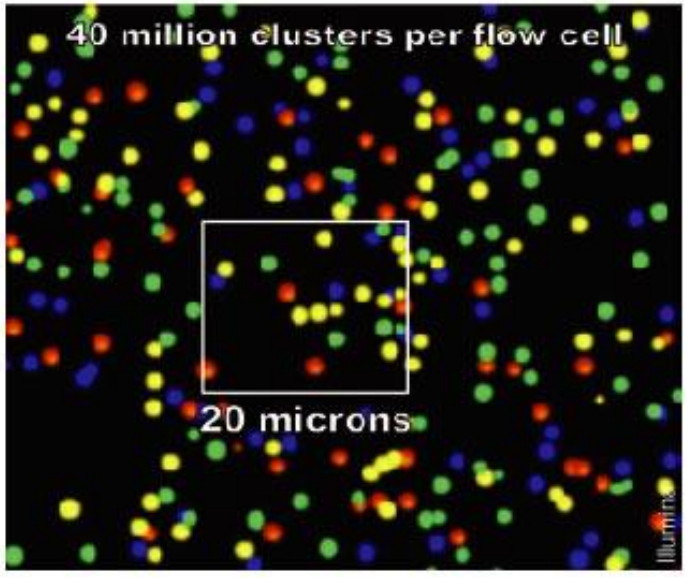






Illumina™sequencing

One color for each nucleotide incorporated in the different “clusters”



Requires highly specific equipment, reagents and software

Comparing sequencing technologies that are readily available

"Sanger" sequencing :

→Maximum sequence length = 1000 nucleotides

→ One single DNA fragment sequenced at a time

→ Requires prior knowledge of the sequence to design primers

Pyrosequencing (454™):

→Maximum sequence length = 400-600 nucleotides

→ Hundreds of thousands of fragments sequenced in parallel

→ Useful to sequence large chunks of the genome or trascriptomes

→ Doesn't require prior knowledge about the genome sequence, content and structure

PCR Colony sequencing (Illumina™):

→Maximum sequence length = 100 nucleotides

→ Millions of fragments sequenced in parallel

→ Useful to sequence large chunks of the genome or trascriptomes

→Doesn't require prior knowledge about the genome sequence, content and structure

Other available Technologies...

SOLID™ sequencing (Sequencing by Oligonucleotide Ligation and Detection - Applied Biosystems):

→Library production = in-between 454 and Illumina
→ Very “special” way of sequencing each fragment
→ Unique way of reading the sequencing results.

HELICOS™ sequencing:

→ No PCR involved!
→ True, single DNA molecule sequencing. Each fragment is directly sequenced in parallel, without the need to clone it

3rd (4th??) generation DNA sequencing

Pacific Biosiences

→Single molecule DNA sequencing
→One DNA molecule + One polymerase!

Oxford Nanopore

→Single molecule DNA sequencing
→One DNA molecule + One polymerase!

Genome Assembly

Shotgun method

A. DNA extraction

B. DNA breakage

C. Genomic library

Plasmidic or chromosomic vector or with phage vector

Library on beads

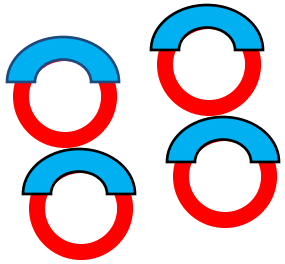
Library on solid phase

ADN Total

The « Shotgun » method and vector library

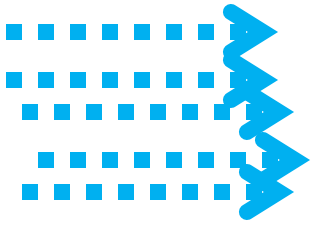
Complete genome:
2 Mb (very small)

DNA broken into
1.0-2.0 kb fragments



Insertion into
plasmids

Library including
overlapping clones



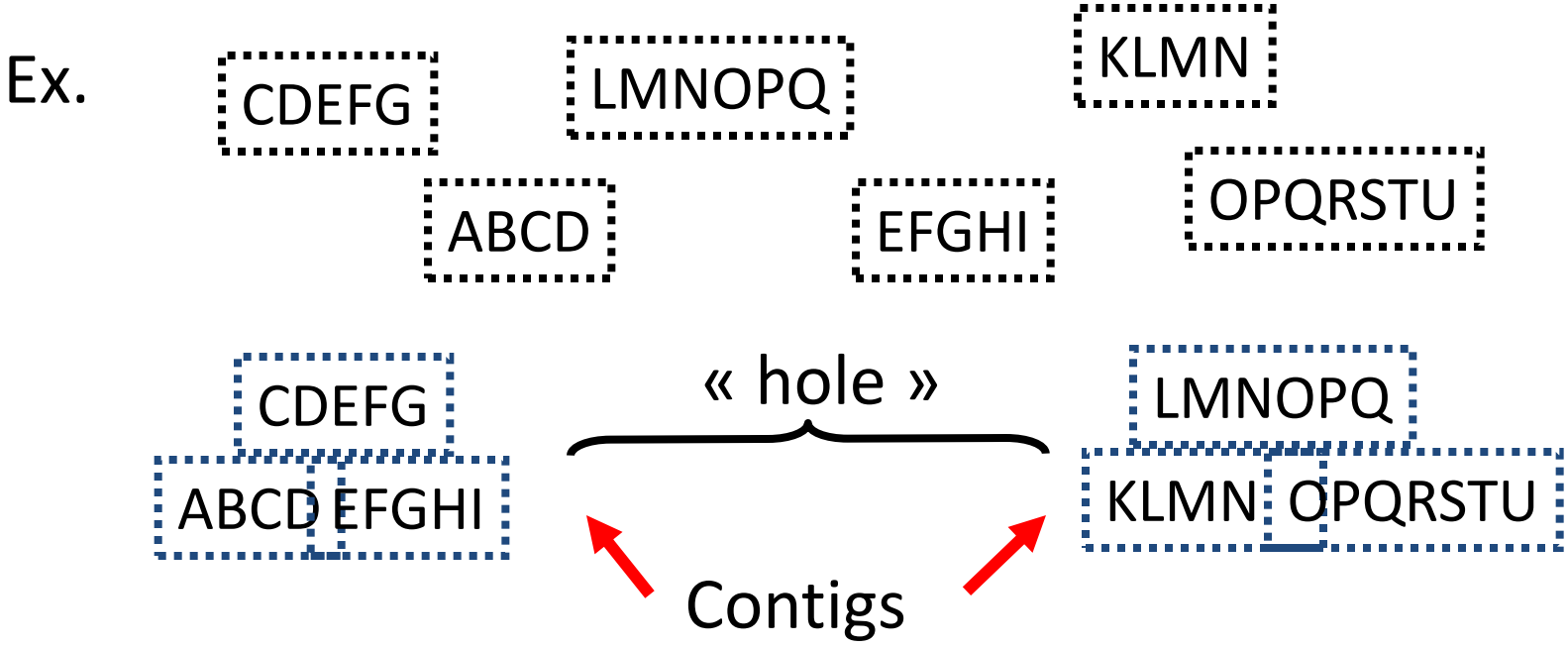
Random sequencing of clones
with universal primers

Sequence assembly

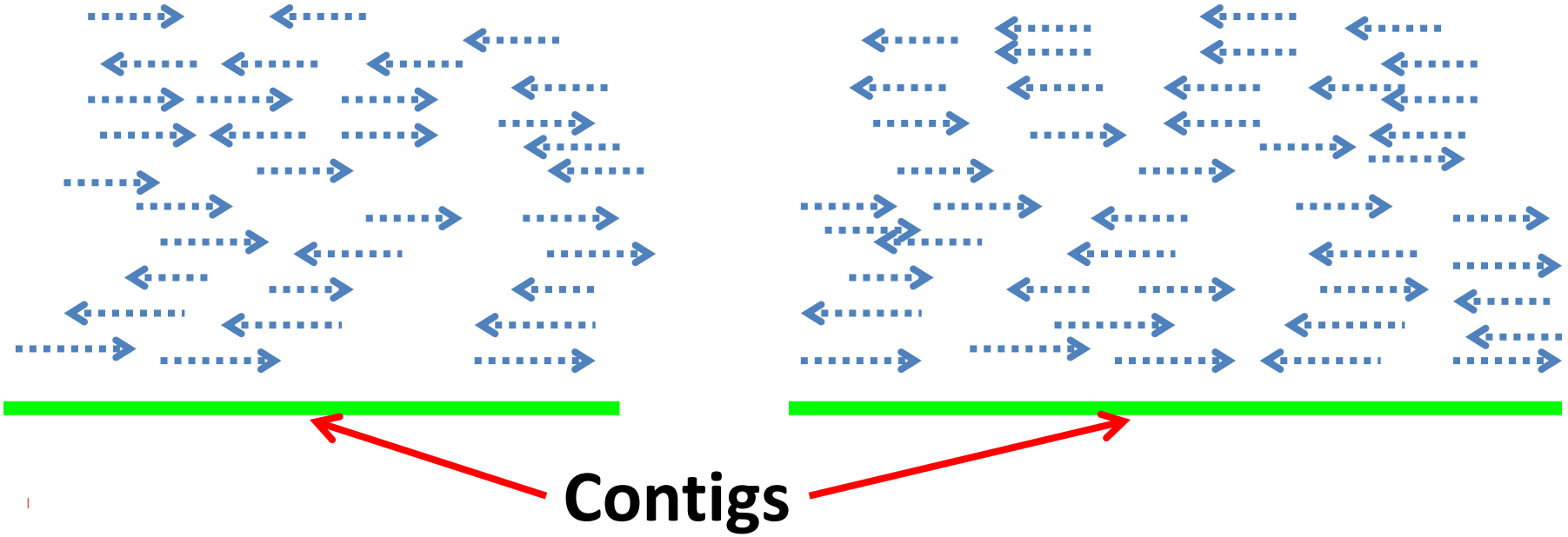
- The different sequences obtained have to be assembled into larger fragments

→ We use the « overlap » in sequence between different clones

Ex.

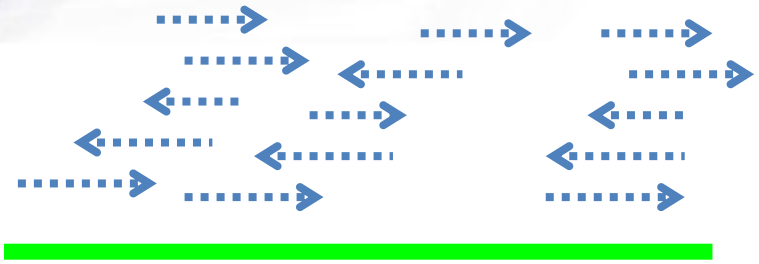


Genome assembly – the contigs



- The assembly of sequence fragments into contigs is done using **bioinformatics (*in silico*)**
- Contigs = same concept for all sequencing technologies

Genome assembly – the contigs



Contig 1 = 13Kb
= region X of the genome

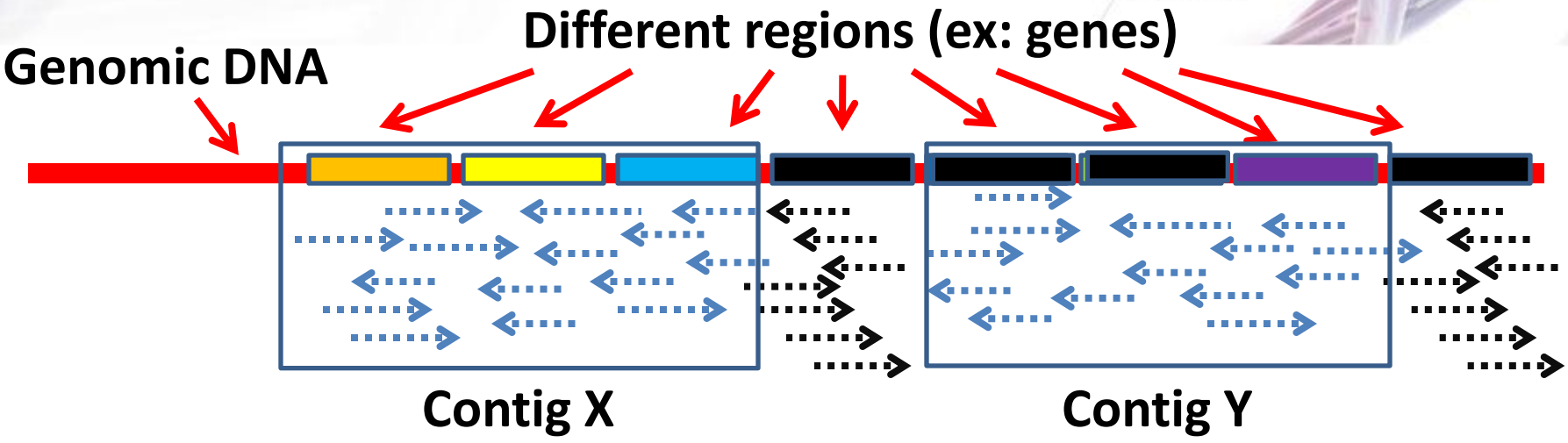
Contig 2 = 50 kb
= region Y of the genome

Why the contigs dont include entire chromosomes?

1. Some regions are absent from our library – they have not been included in the vector (probability)

2. Some regions are highly repeated (99% of the cases) = PUBLIC ENEMY NUMBER ONE (...for assemblers!)

Genome assembly and repeated regions



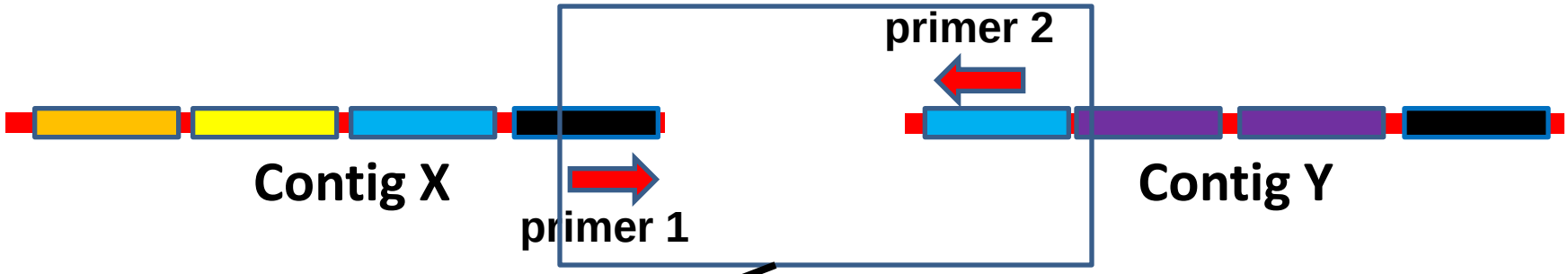
Repeated regions create identical contigs!

Software for assemblies don't know where to place them

Many regions are repeated in TANDEM → even worse than normal repetitions!

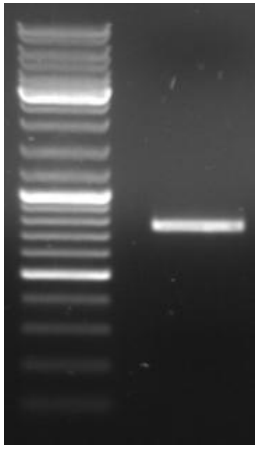
Genome assembly and repeated regions

1. PCR → with all possible combinations of primers



PCR

The PCR product is sequenced (Sanger)



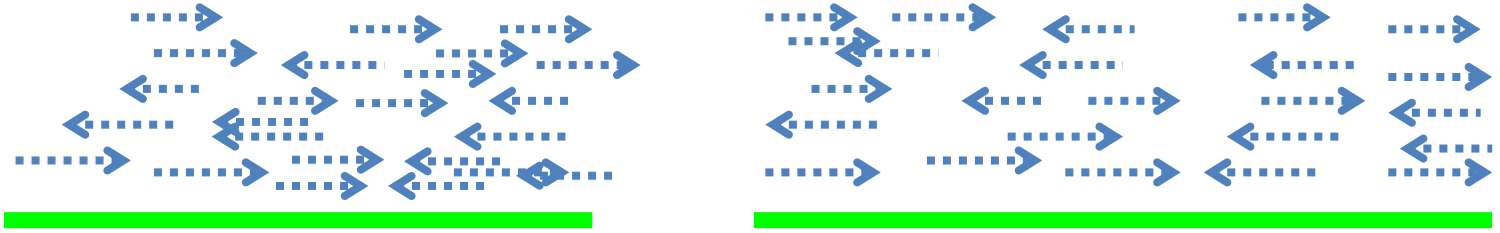
The sequence that links the contigs is amplified

A new, larger contig is produced

Genome assembly and repeated regions

2. Clone and sequence much longer regions of the genome
→Increase the probability that a repeated region is already included in the clone

→ Increase the chances of having an overlap



Contig 1 = 13kb
= region X of the genome

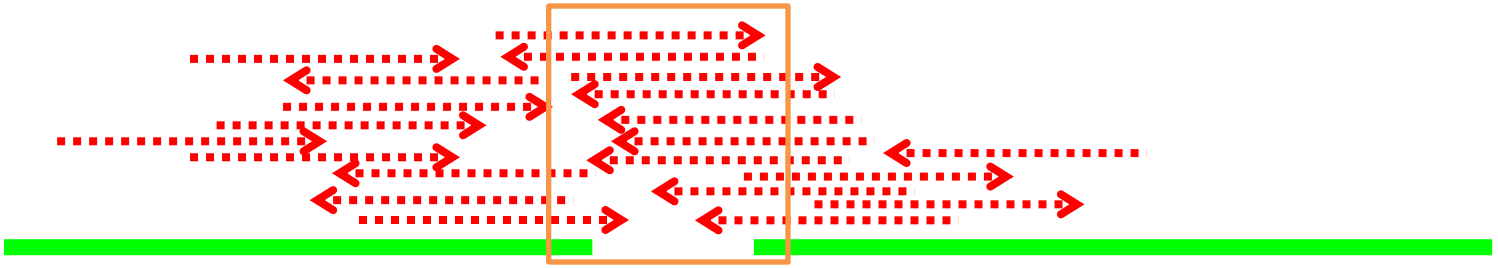
Contig 2 = 50 kb
= region Y of the genome

= 2 kpbs = not long enough to link the contigs *in silico*

Genome assembly and repeated regions

2. Clone and sequence much longer regions of the genome
→Increase the probability that a repeated region is already included in the clone

→ Increase the chances of having an overlap



Contig 1 = 13kb
= region X of the genome

Contig 2 = 50 kb
= region Y of the genome

= 6 kpbs = long enough to link the contigs *in silico*

Genome assembly and short reads (454 ™, Illumina ™)

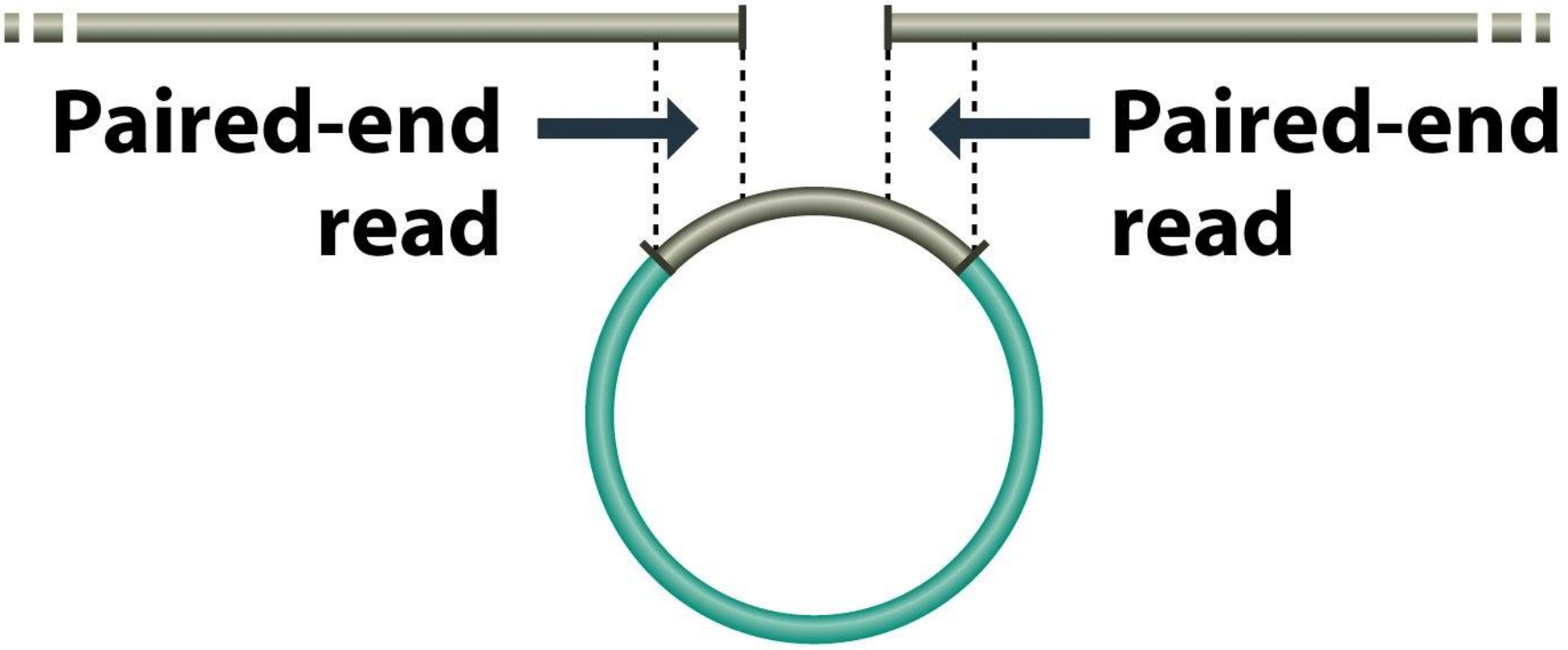
3. Paired-Ends or Mate-Pairs:

→ Gives information on the physical distance between **2 sequences** of 2nd génération (454 ou Illumina)

The different fragments are “tagged” and sequenced from both extremities.

→Distance between the sequences obtained is known, because we select fragments of a given size before sequencing (ex: 200pbs)

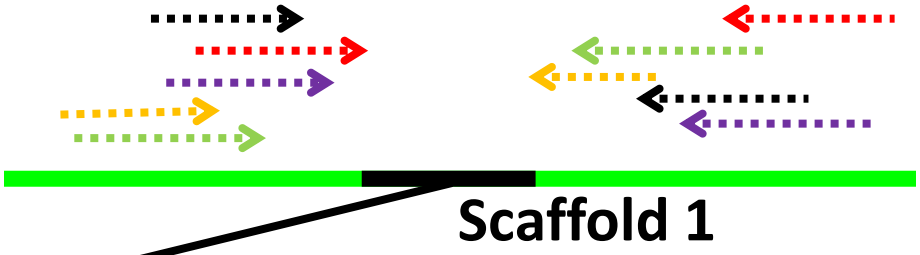
Closing a sequence gap



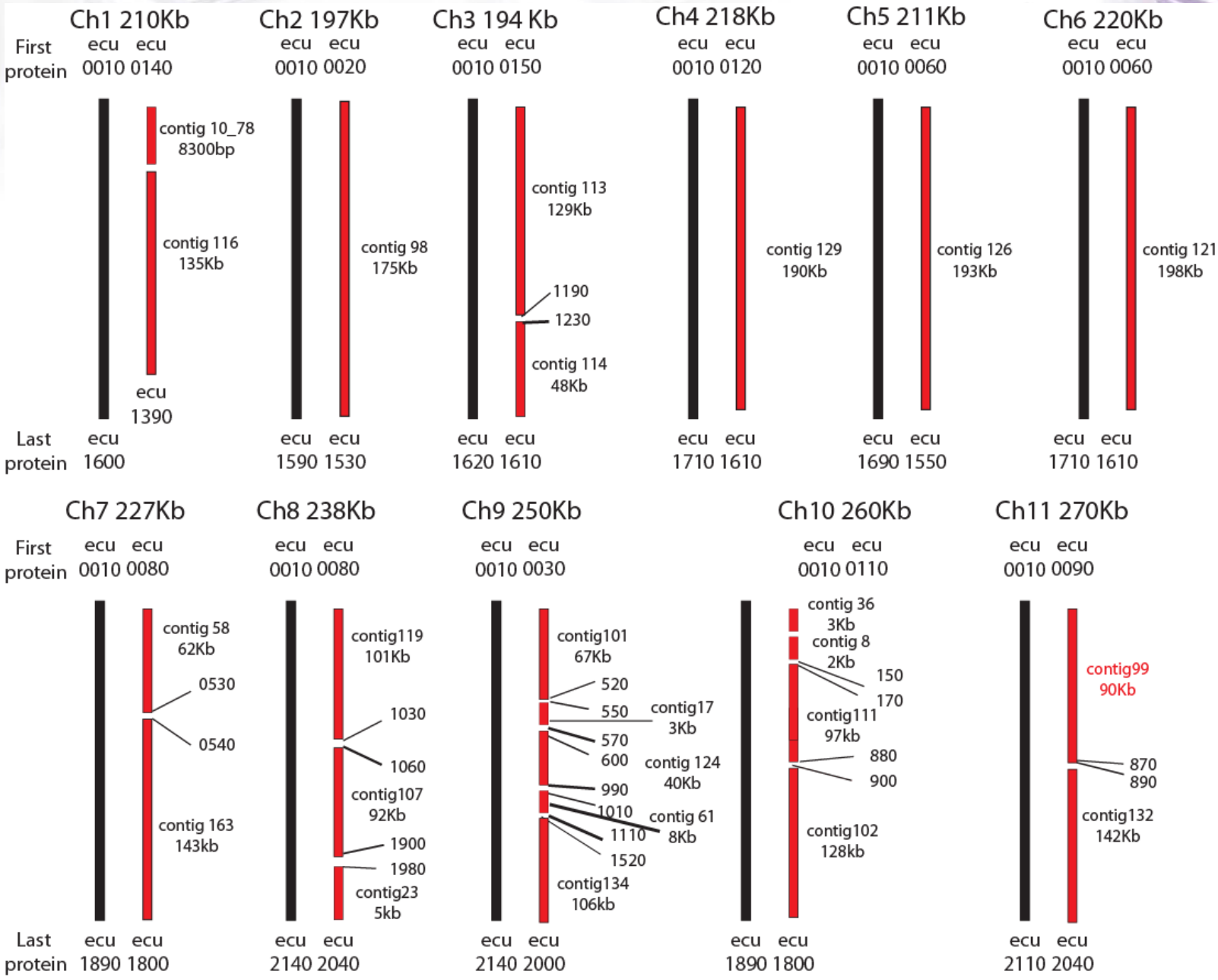
Fragments from different types of libraries (plasmid, bacteriophage, 454, Illumina, etc....)

Genome assembly and short reads (454 ™, Illumina ™)

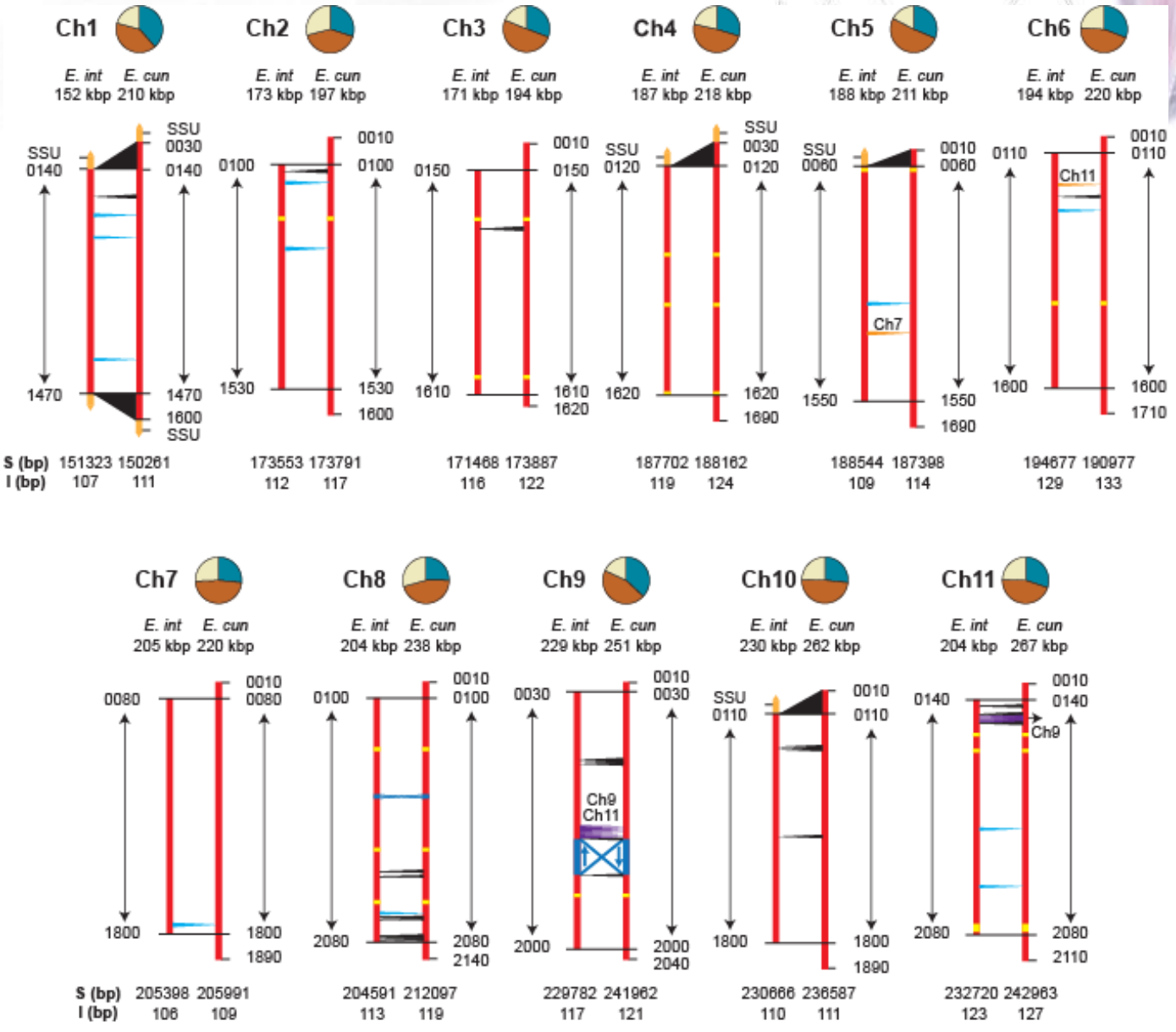
We know these sequences are separated by X base paires



Represents an unknown sequence (NNNN), that we know is located at this position in the genome


Ch1 210Kb
First protein
ecu 0010
ecu 0140
contig 10_78 8300bp
contig 116 135Kb
ecu 1390
Last protein
ecu 1600
Ch2 197Kb
ecu 0010
ecu 0020
contig 98 175Kb
ecu 1590
ecu 1530
Ch3 194 Kb
ecu 0010
ecu 0150
contig 113 129Kb
1190
1230
contig 114 48Kb
ecu 1620
ecu 1610
Ch4 218Kb
ecu 0010
ecu 0120
contig 129 190Kb
ecu 1710
ecu 1610
Ch5 211Kb
ecu 0010
ecu 0060
contig 126 193Kb
ecu 1690
ecu 1550
Ch6 220Kb
ecu 0010
ecu 0060
contig 121 198Kb
ecu 1710
ecu 1610
Ch7 227Kb
First protein
ecu 0010
ecu 0080
contig 58 62Kb
0530
0540
contig 163 143kb
Last protein
ecu 1890
ecu 1800
Ch8 238Kb
ecu 0010
ecu 0080
contig119 101Kb
1030
1060
contig107 92Kb
1900
1980
contig23 5kb
ecu 2140
ecu 2040
Ch9 250Kb
ecu 0010
ecu 0030
contig101 67Kb
520
550
contig17 3Kb
570
600
contig 124 40Kb
990
1010
contig 61 8Kb
1110
1520
contig134 106kb
ecu 2140
ecu 2000
Ch10 260Kb
ecu 0010
ecu 0110
contig 36 3Kb
contig 8 2Kb
150
170
contig111 97kb
880
900
contig102 128kb
ecu 1890
ecu 1800
Ch11 270Kb
ecu 0010
ecu 0090
contig99 90Kb
870
890
contig132 142Kb
ecu 2110
ecu 2040

2000 PCRs later... (+ analyses *in silico*)



OPEN ACCESS Freely available online

PLoS PATHOGENS

Genomic Survey of the Non-Cultivatable Opportunistic Human Pathogen, *Enterocytozoon bieneusi*

Donna E. Akiyoshi[1☯*], Hilary G. Morrison[2☯], Shi Lei[1], Xiaochuan Feng[1], Quanshun Zhang[1], Nicolas Corradi[5], Harriet Mayanja[3], James K. Tumwine[4], Patrick J. Keeling[5], Louis M. Weiss[6], Saul Tzipori[1]

Research

Open Access

Draft genome sequence of the *Daphnia* pathogen *Octosporea bayeri*: insights into the gene content of a large microsporidian genome and a model for host-parasite interactions

Nicolas Corradi*, Karen L Haag†‡, Jean-François Pombert*, Dieter Ebert¤† and Patrick J Keeling¤*

ARTICLE

Received 1 Jun 2010 | Accepted 25 Aug 2010 | Published xx xxx 2010

DOI: 10.1038/ncomms1082

The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*

Nicolas Corradi[1,*,†], Jean-François Pombert[1,*], Laurent Farinelli[2], Elizabeth S. Didier[3] & Patrick J. Keeling[1]

Complexity of the assembly