Metodologias de Sequenciamento de Genomas

Profa. Dra. Chirlei Glienke
BIOGEMM-UFPR
GS Treinamentos e Consultoria

Genoma completo

“Draft genome”

Exemplo de um “genoma” difícil de montar...

Genômica

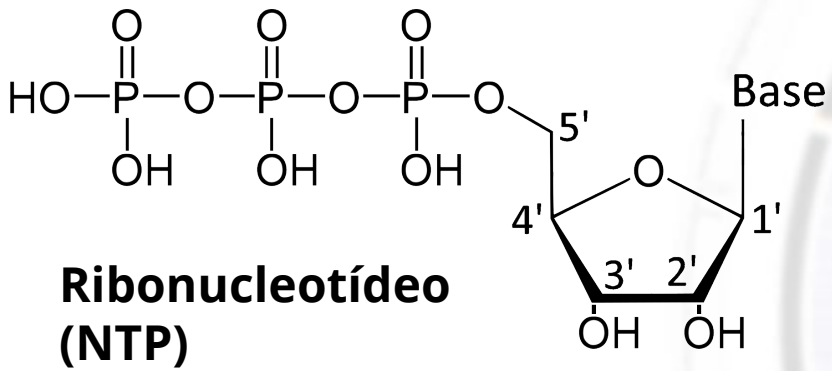
- Refere-se ao estudo do genoma envolvendo
  - Mapeamento
  - Sequenciamento
  - Análise
  - Comparação

Importância

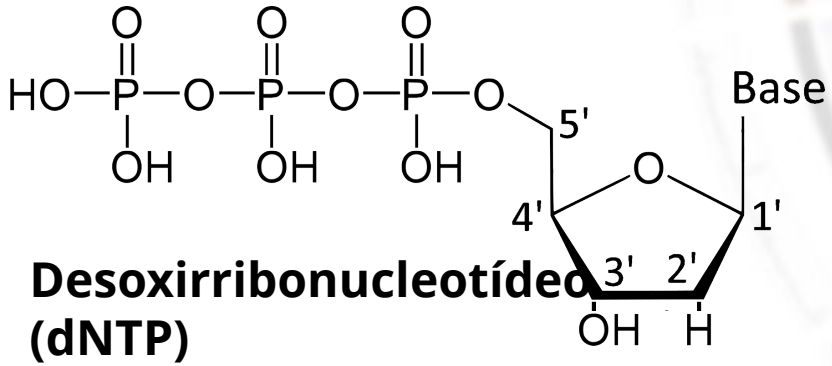
- Revelar os genes presentes no genoma
  - Estudo de patógenos, mineração de gene clusters, etc
- Fornecer informações sobre as funções do organismo
- Fornecer informações sobre a história evolutiva do organismo
- Auxiliar em estudos da expressão gênica

- 1. Sequenciamento de DNA

Relembrando...

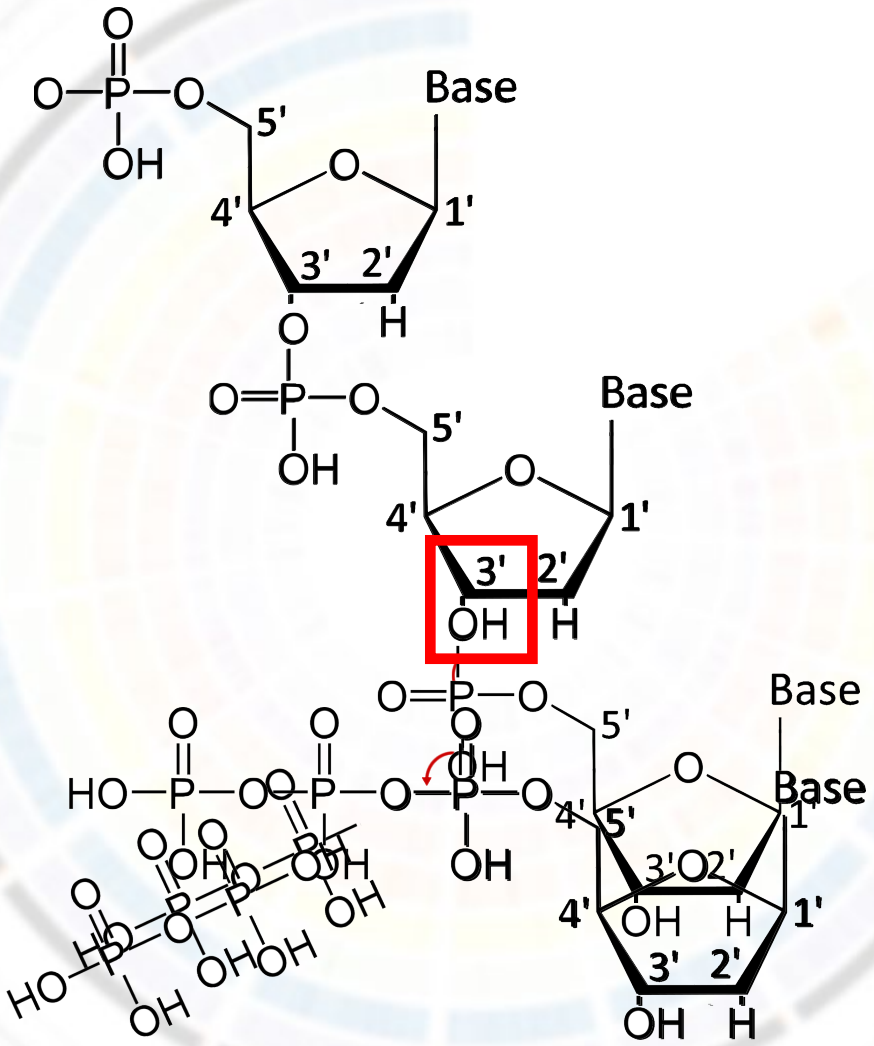


Ribonucleotídeo (NTP)

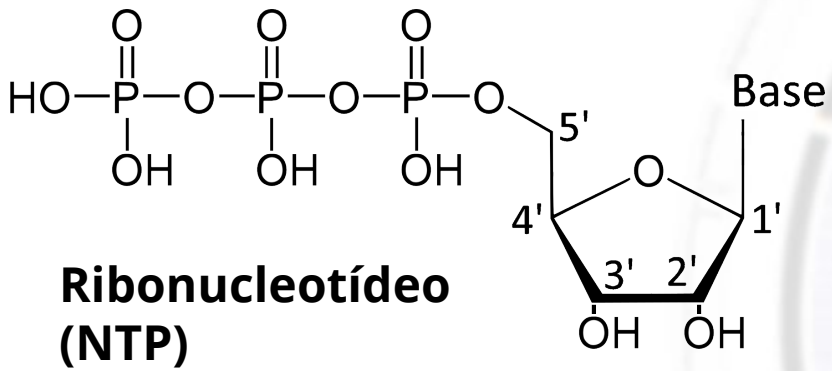


Desoxirribonucleotídeo (dNTP)

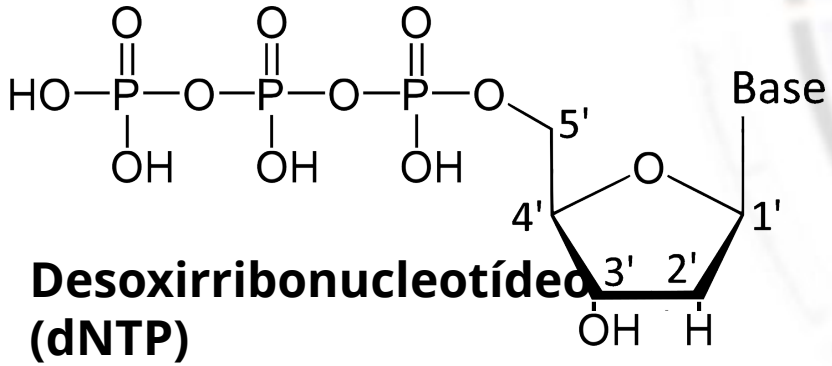
(Brown, 2010)



Relembrando...

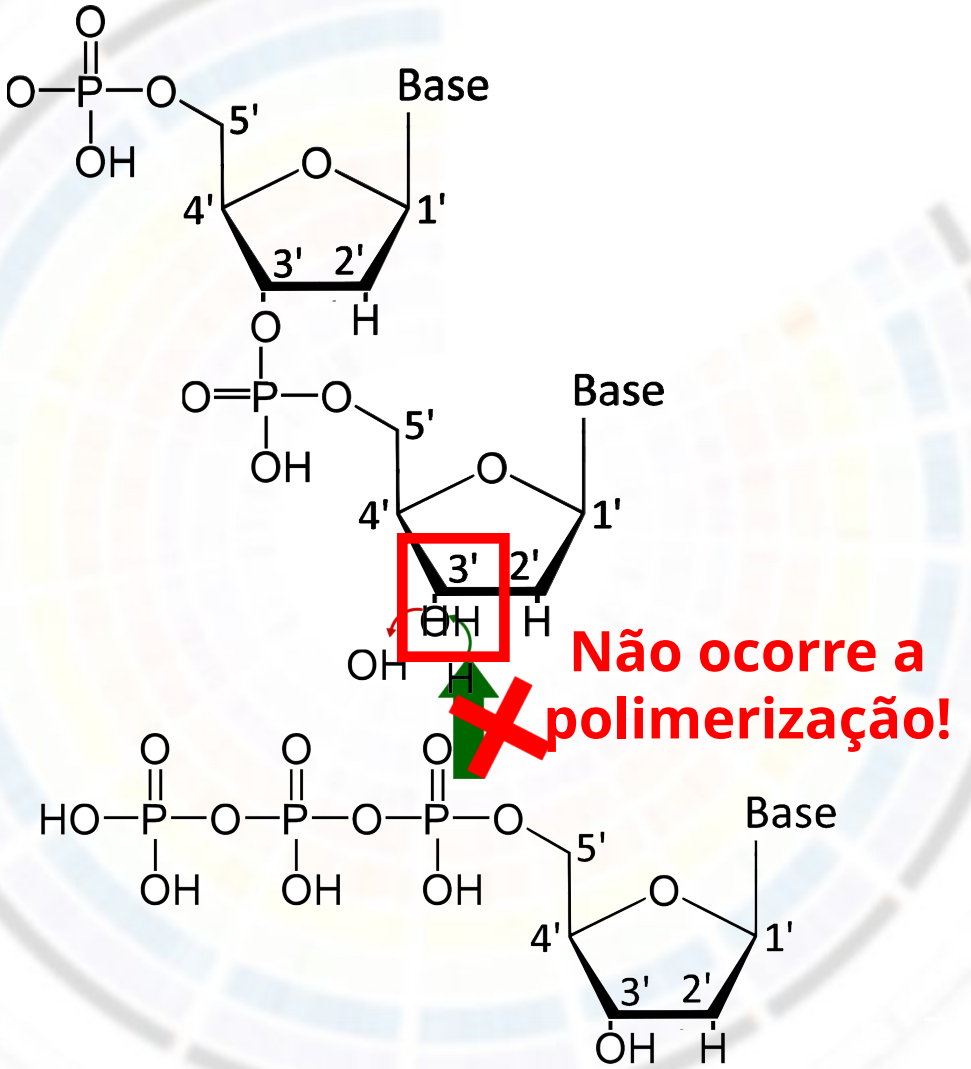


Ribonucleotídeo (NTP)

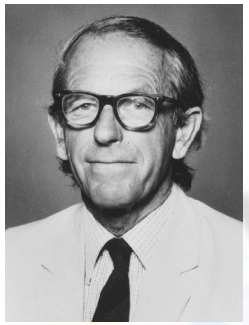


Desoxirribonucleotídeo (dNTP)

(Brown, 2010)



Sequenciamento de DNA



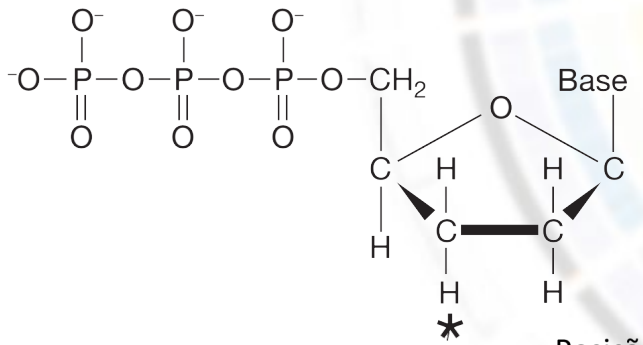
Frederick Sanger

Método de Sanger / Método de terminação em cadeia

Sequenciamento associado com polimerização

Popularizou a técnica e permite hoje sequenciamento de até cerca de 1000 pb

2’,3’–didesoxirribonucleosídeo trifosfatado (ddNTP)



Posição 3’ onde ocorreria
a ligação do próximo nucleotídeo

Posição que diferencia
ribose de desoxirribose

+

Método de Sanger

- Muitos Desoxirribonucleotídeos (dNTP)
- Poucos Didesoxirribonucleotídeos com uma das bases nitrogenadas (ddATP)

Parada da síntese!

Método de Sanger

- Muitos Desoxirribonucleotídeos (dNTP)
- Poucos Didesoxirribonucleotídeos com uma das bases nitrogenadas (ddATP)

Parada da síntese!

Método de Sanger

- Muitos Desoxirribonucleotídeos (dNTP)
- Poucos Didesoxirribonucleotídeos com uma das bases nitrogenadas (ddATP)

Parada da síntese!

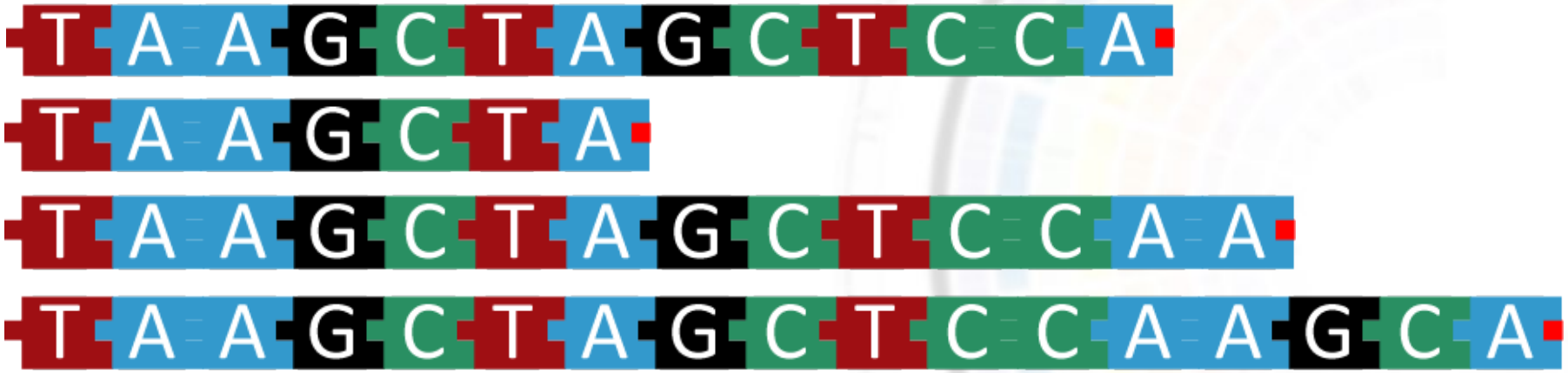
Método de Sanger

- Muitos Desoxirribonucleotídeos (dNTP)
- Poucos Didesoxirribonucleotídeos com uma das bases nitrogenadas (ddATP)

Parada da síntese!

Método de Sanger

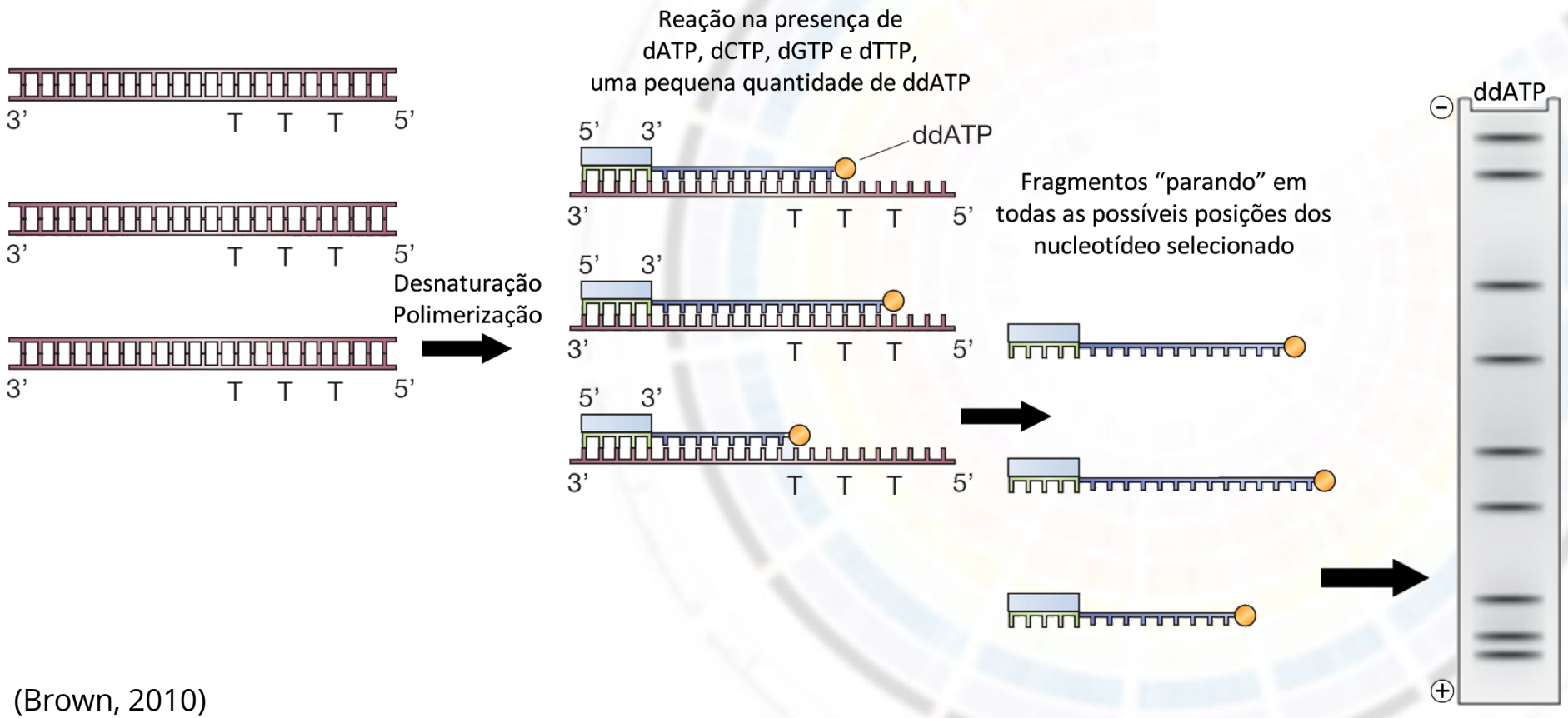
- Muitos Desoxirribonucleotídeos (dNTP)
- Poucos Didesoxirribonucleotídeos com uma das bases nitrogenadas (ddATP)



- Múltiplas fitas simples com diferentes tamanhos
- Diferentes tamanhos = diferentes mobilidade eletroforética

Método de Sanger

Grande abundância de DNA molde



(Brown, 2010)

Método de Sanger

STEP 1 Montar 4 reações de polimerização de DNA com os seguintes componentes:

Fita molde 3' – GCATGATCGG – 5'

Primer (1) 5 ~~OH 3'

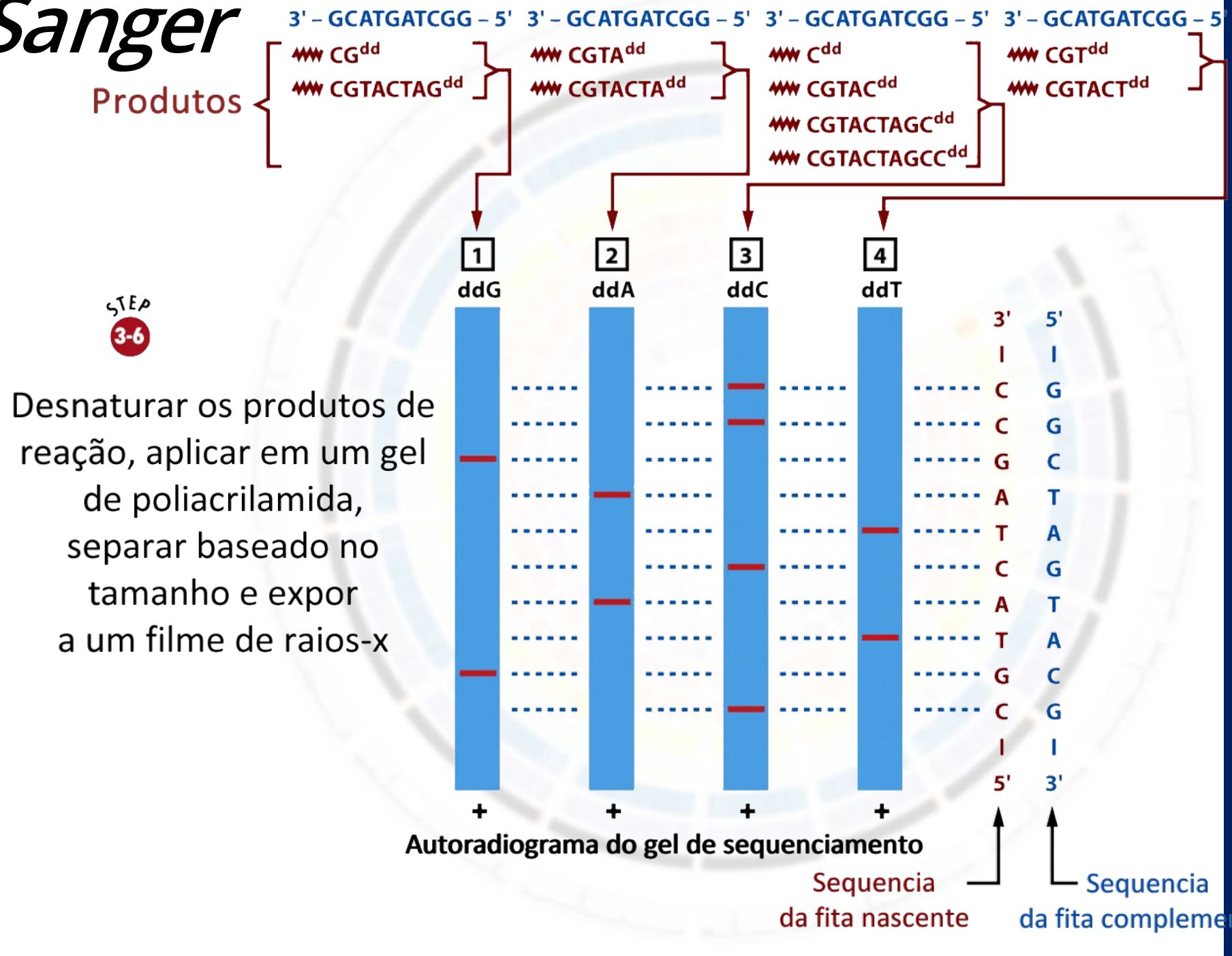
↓ DNA Polimerase
dGTP, dATP, dTTP, ^{32}P-dCTP

STEP 2 Adicione um dos quatro 2',3'-didesoxirribonucleotídeos trifosfatados (terminadores de cadeia) a cada uma das reações

Reação 1 : ddGTP	Reação 2 : ddATP	Reação 3 : ddCTP	Reação 4 : ddTTP
3' – GCATGATCGG – 5'	3' – GCATGATCGG – 5'	3' – GCATGATCGG – 5'	3' – GCATGATCGG – 5'
Produtos: ~~CG^{dd}	~~CGTA^{dd}	~~C^{dd}	~~CGT^{dd}
~~CGTACTAG^{dd}	~~CGTACTA^{dd}	~~CGTAC^{dd}	~~CGTACT^{dd}
		~~CGTACTAGC^{dd}	
		~~CGTACTAGCC^{dd}	

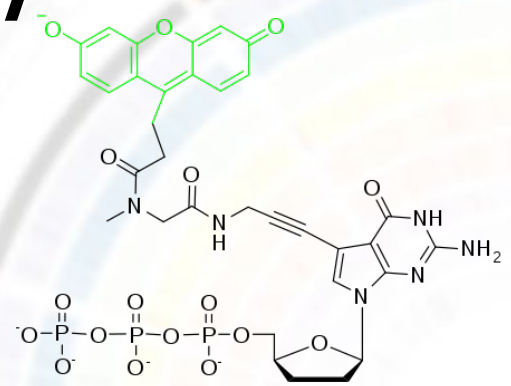
Método de Sanger

- Eletroforese em poliacrilamida (alta resolução para pequenos fragmentos)
- Corrida simultânea de todas as quatro reações

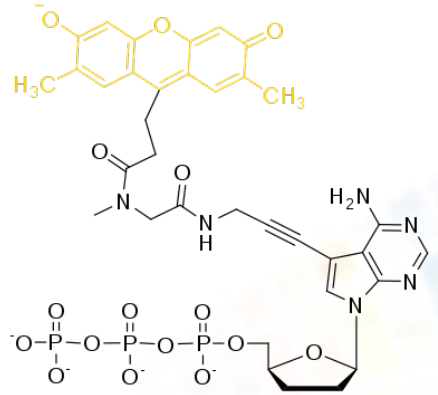


Método de Sanger

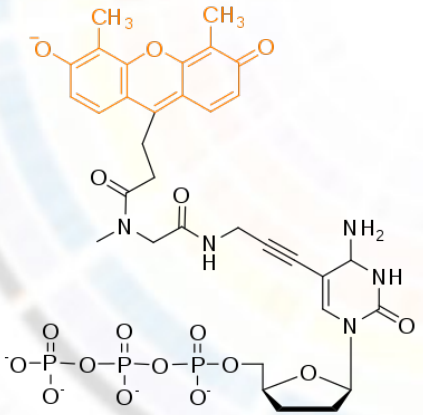
- Utilização de nucleotídeos com modificações fluorescentes
- Permitiu realizar uma reação apenas e não quatro separadas
- Permitiu automatizar o processo aquisição do dado e análise



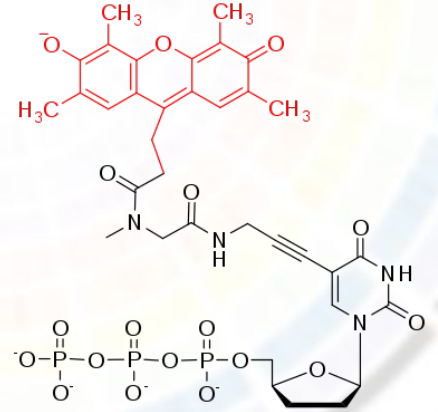
G-505



A-512



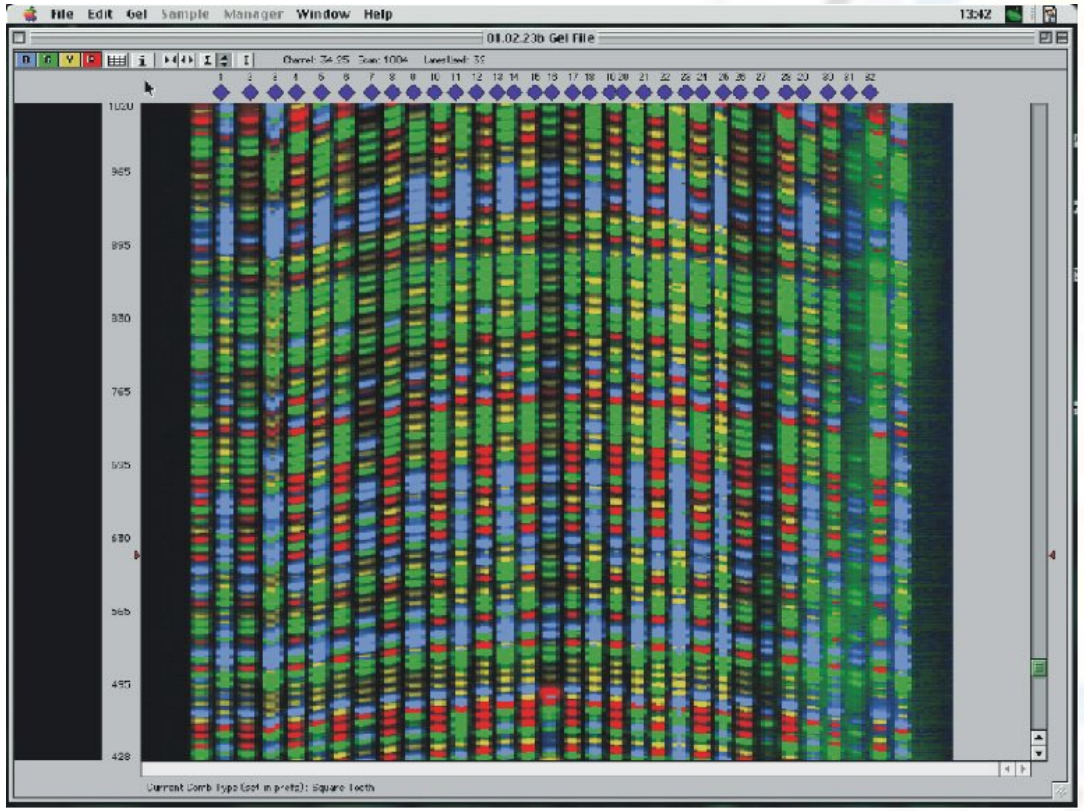
C-519

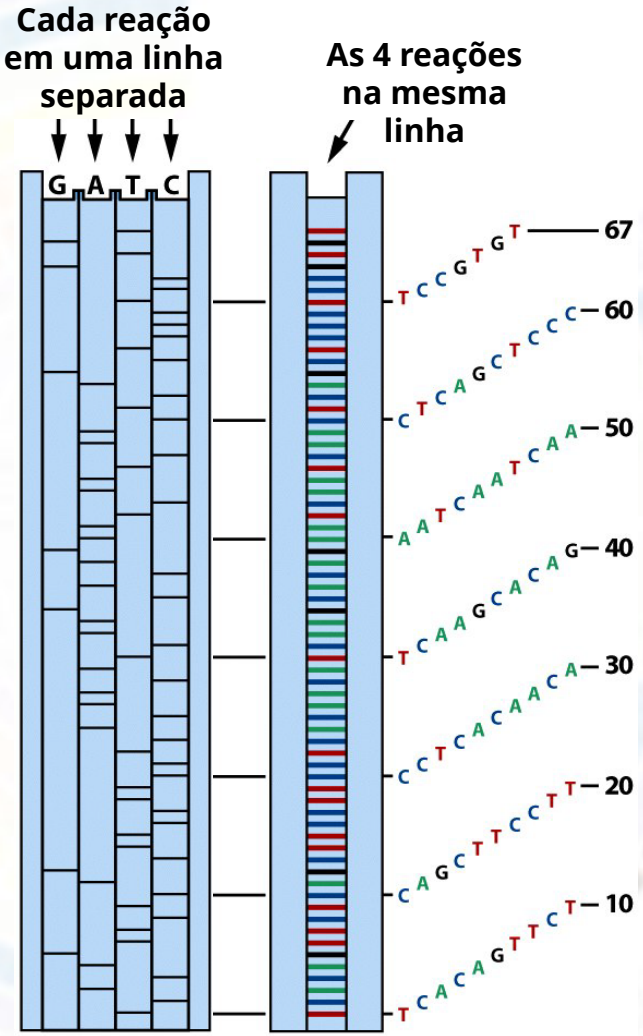


T-526

Método de Sanger

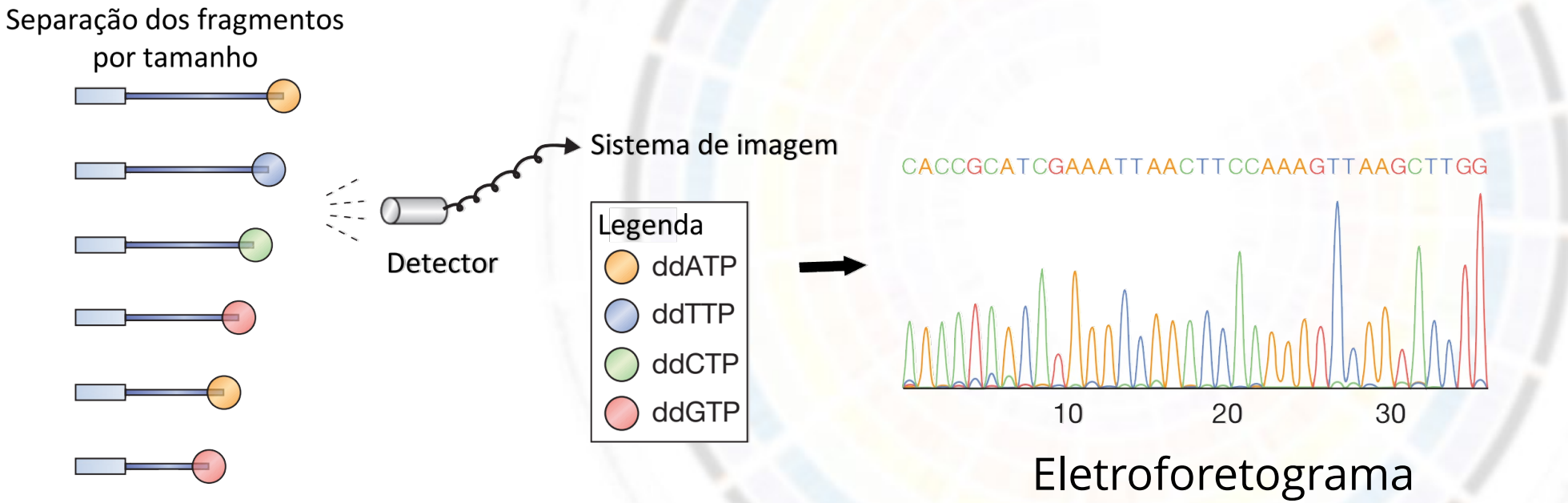
TCAG





Método de Sanger

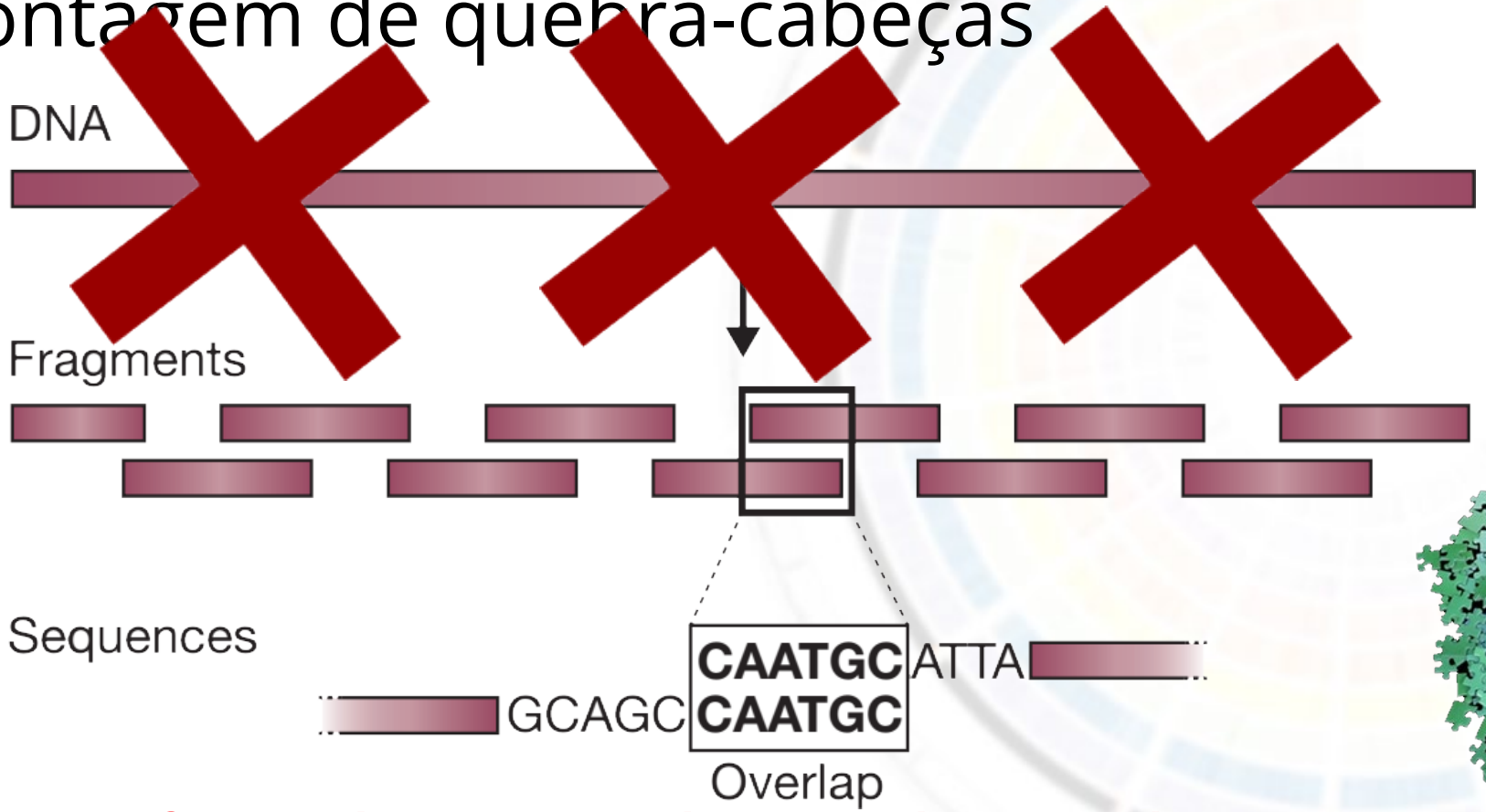
Eletroforese em capilar e automatização

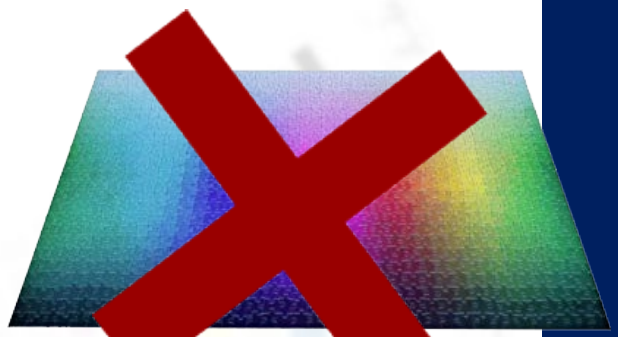


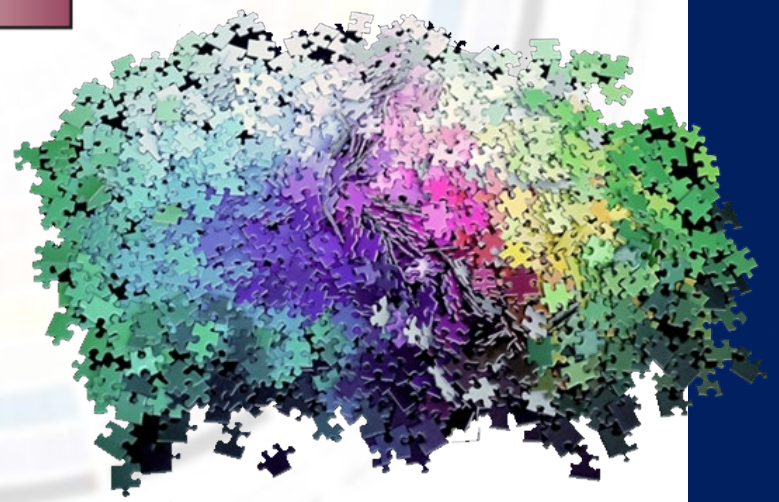
Eletroforetograma

Sequenciamento de Genomas

- Promessa de resolução de todos os problemas biotecnológicos e de saúde
- Montagem de quebra-cabeças





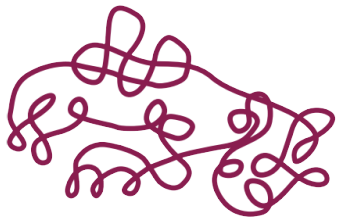


- Sem a foto da caixa do quebra-cabeças

Sequenciamento de Genomas - 1ª Geração

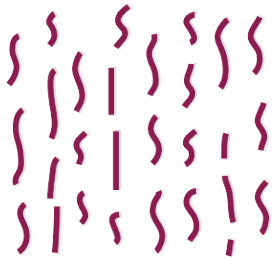
- Estratégia de sequenciamento *shotgun*

DNA genômico



Fragmentação
(Sonicação)

DNAs de diversos tamanhos

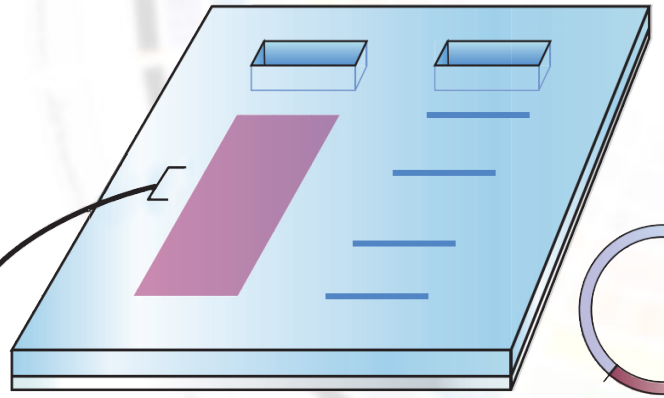


Purificar o DNA do gel

Eletroforese em gel de agarose

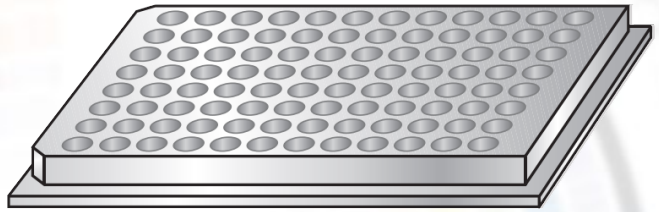
DNA Fragmentado

Marcador



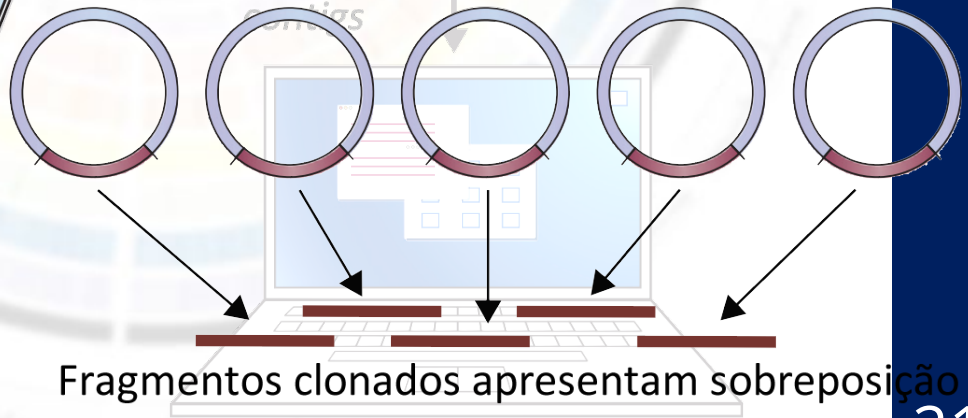
Fragmentos de DNA 1.6 - 2.0 kb

Preparo da biblioteca de clones



Obtenção de sequências dos finais dos insertos de DNA

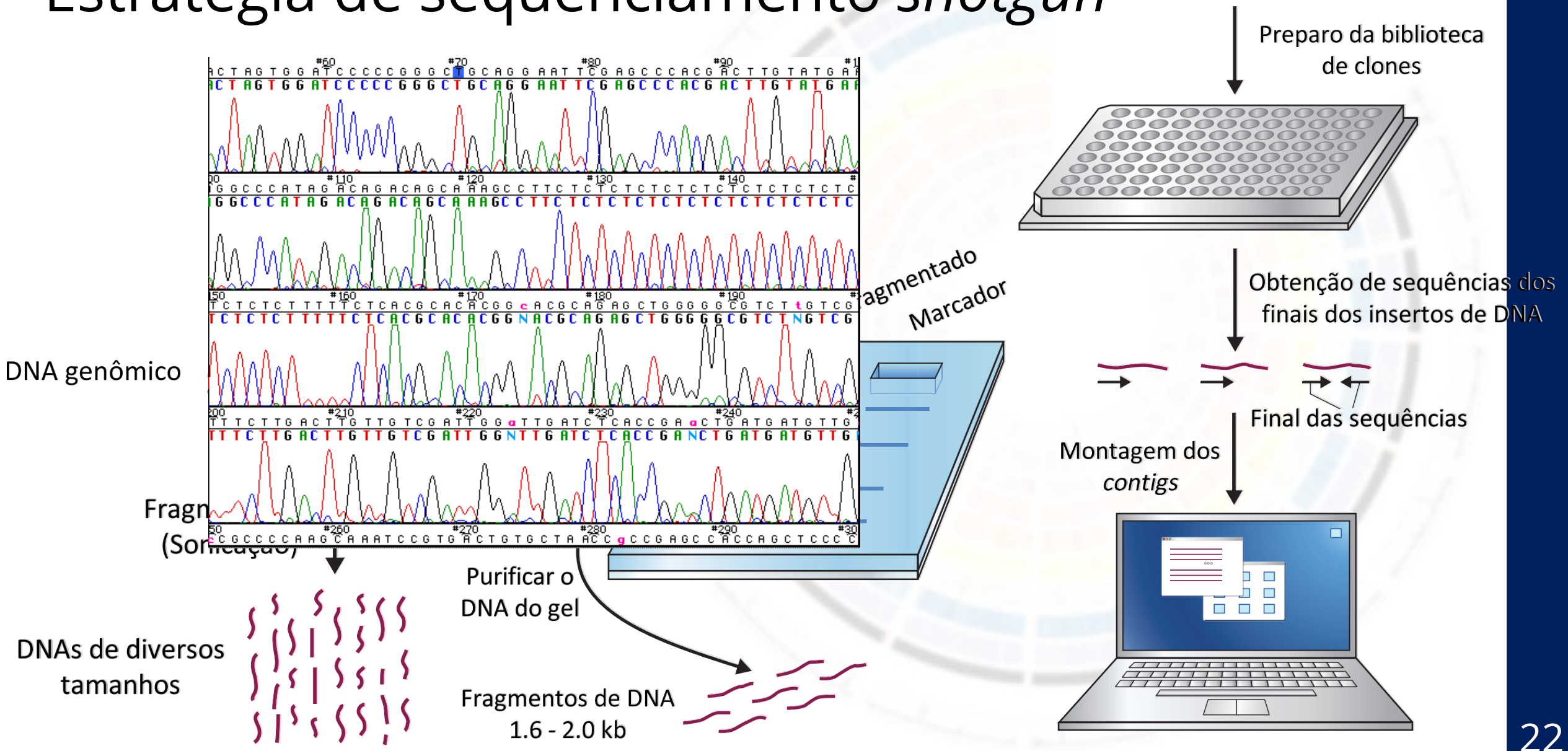
Biblioteca de clones



Fragmentos clonados apresentam sobreposição

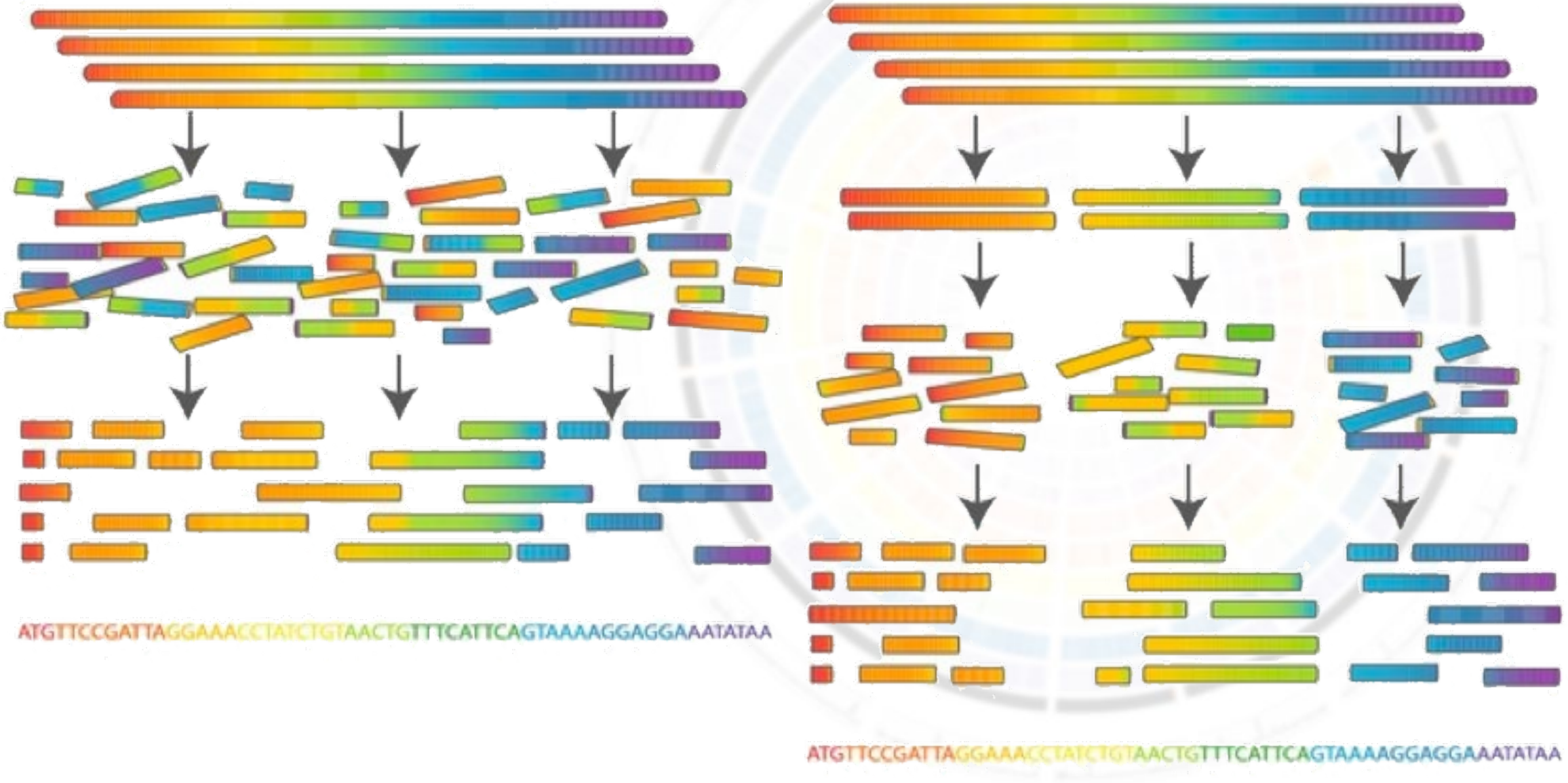
Sequenciamento de Genomas - 1ª Geração

- Estratégia de sequenciamento *shotgun*



Sequenciamento de Genomas - 1ª Geração

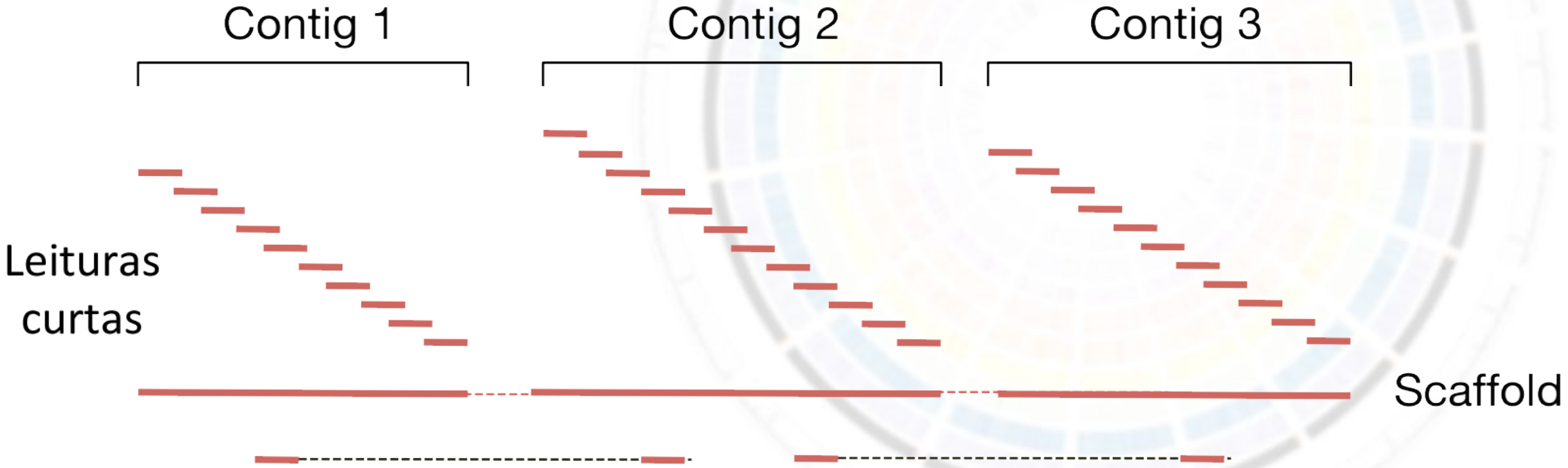
Shotgun vs *Shotgun* hierárquico



Sequenciamento de Genomas - 1ª Geração

Contigs vs *Scaffolds*

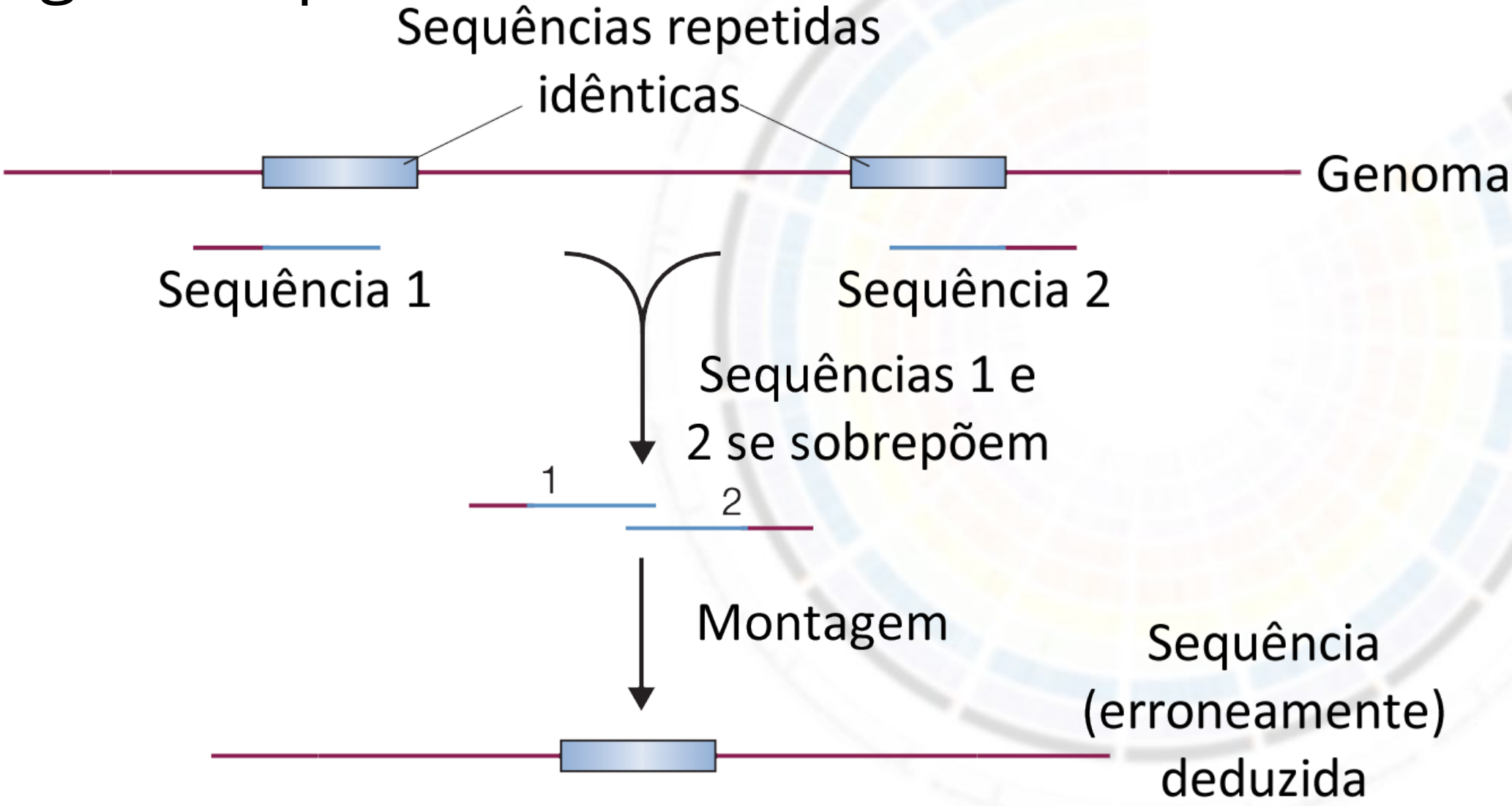
Importância da leitura de “pontas” de fragmentos mais longos



Leituras Paired-end

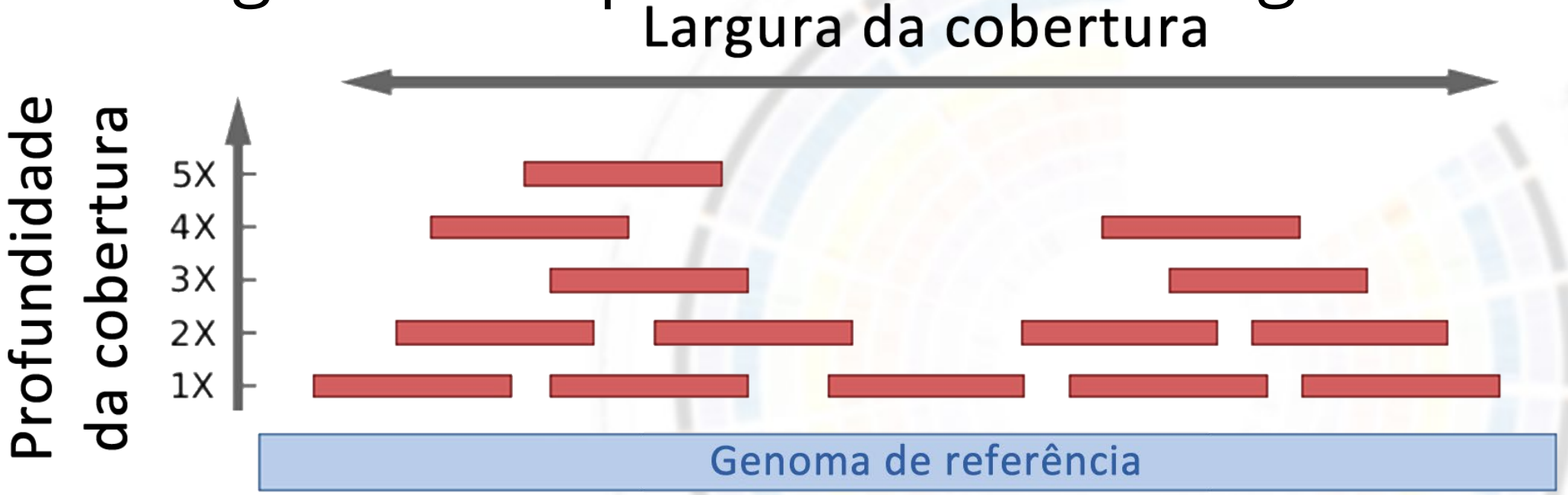
Sequenciamento de Genomas

- Regiões repetitivas



Sequenciamento de Genomas

- *Cobertura* garante a qualidade da montagem final

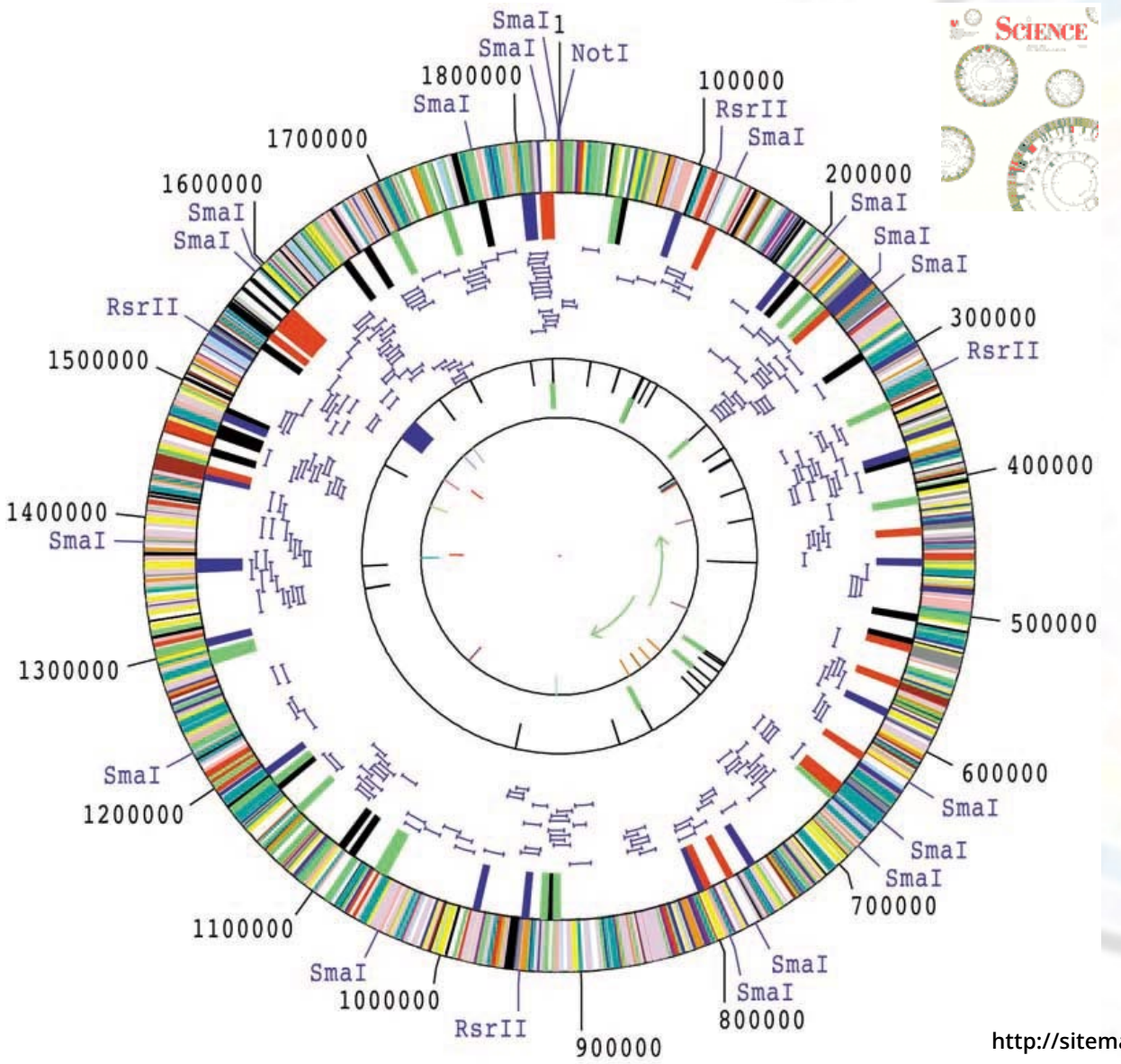



```
                                AGCATCGTAGCTTCAGTATGATGATGCTAG   Read 1
ATGATCGTAGCTAGCATCGTAGCTAGC                                      Read 2
   ATCGTAGCTAGCATCGTAGCTAGCATCGTAGCTT                            Read 3
                                TTGTAGCTTCAGTATGATGATGCTAG        Read 4
              GCATCGTAGCTAGCATCGTAGCTTCAGT                       Read 5
ATGATCGTAGCTAGCATCGTA                                            Read 6
ATGATCGTAGCTAGCATCGTAGCTAGCATCGTAGCTTCAGTATGATGATGCTAG           Deduced sequence
```

Sequenciamento de Genomas - 1ª Geração

- (1977) - Primeiro genoma completo: vírus phiX174
- (1990) - Início do projeto genoma humano
- (1995) - *Haemophilus influenza*: primeiro genoma bacteriano
- (1996) - *Saccharomyces cerevisiae*
- (1999) - *Caenorhabditis elegans*
- (2000) - *Drosophila melanogaster*
- (2001) - Rascunho do genoma humano
- (2002) - *Mus musculus*: primeiro genoma de mamífero
- (2003) - Genoma humano "finalizado"
- (2010) - Genoma Neanderthal

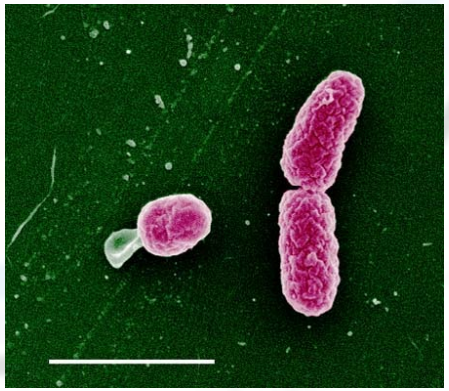
Genoma da bactéria Haemophilus influenzae Rd KW20



Primeiro genoma de um organismo celular

Publicado em 1995

Bactéria *Haemophilus influenzae*
Tamanho de 1.830.137pb



http://sitemaker.umich.edu/mc13/bacterial_meningitis_causative_organism

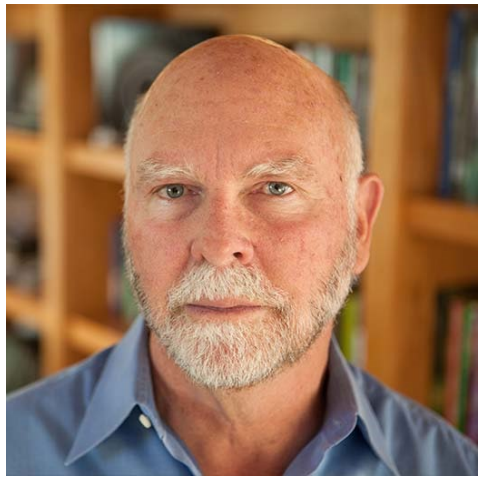
Sequenciamento de Genomas - 1ª Geração

- (1977) - Primeiro genoma completo: vírus phiX174
- (1990) - Início do projeto genoma humano
- (1995) - *Haemophilus influenza*: primeiro genoma bacteriano
- (1996) - *Saccharomyces cerevisiae*
- (1999) - *Caenorhabditis elegans*
- (2000) - *Drosophila melanogaster*
- (2001) - Rascunho do genoma humano
- (2002) - *Mus musculus*: primeiro genoma de mamífero
- (2003) - Genoma humano "finalizado"
- (2010) - Genoma Neanderthal

Sequenciamento de Genomas - 1ª Geração

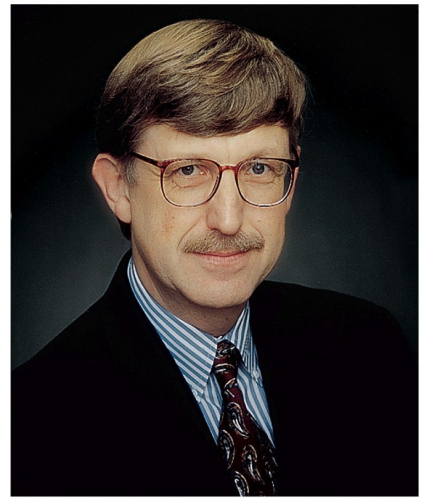
- "Corrida" pelo genoma humano.
- *Shotgun* vs *Shotgun hierárquico*

Celera genomics



Craig Venter

DOE

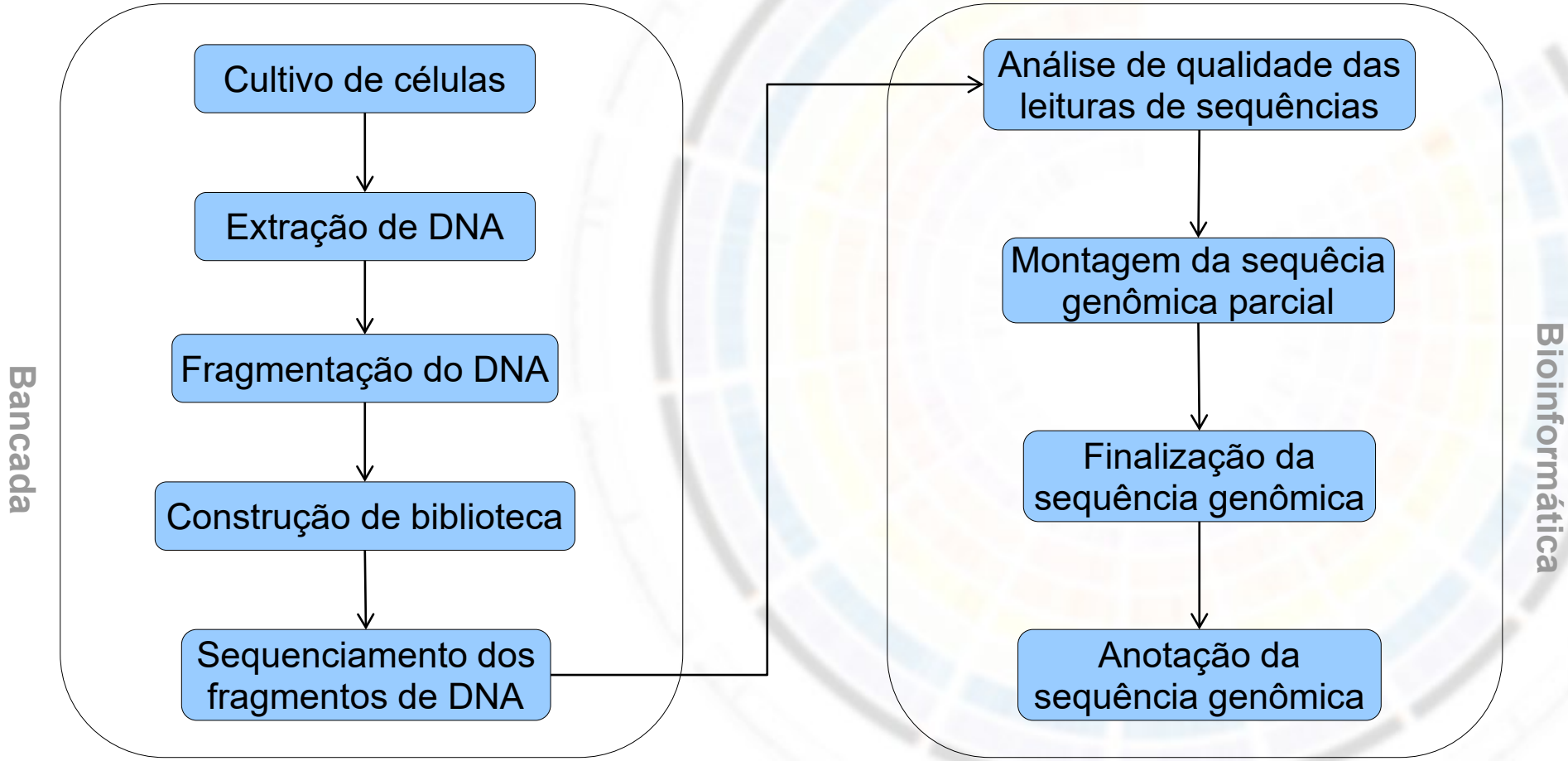


Francis Collins

Sequenciamento de Genomas - 1ª Geração

- (1977) - Primeiro genoma completo: vírus phiX174
- (1990) - Início do projeto genoma humano
- (1995) - *Haemophilus influenza*: primeiro genoma bacteriano
- (1996) - *Saccharomyces cerevisiae*
- (1999) - *Caenorhabditis elegans*
- (2000) - *Drosophila melanogaster*
- (2001) - Rascunho do genoma humano
- (2002) - *Mus musculus*: primeiro genoma de mamífero
- (2003) - Genoma humano "finalizado"
- (2010) - Genoma Neanderthal

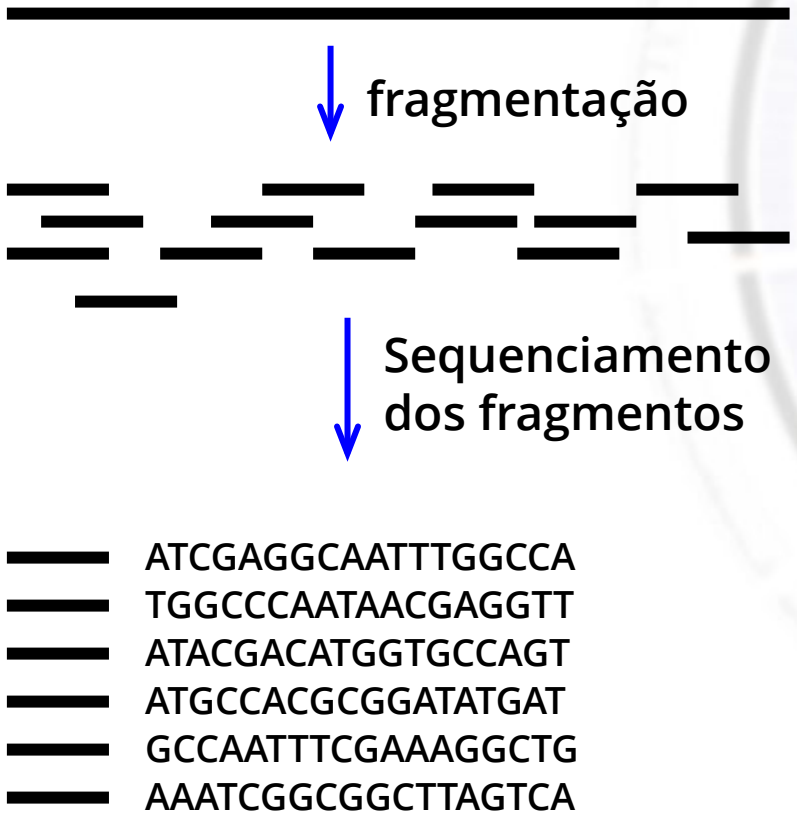
Etapas para o sequenciamento genômico



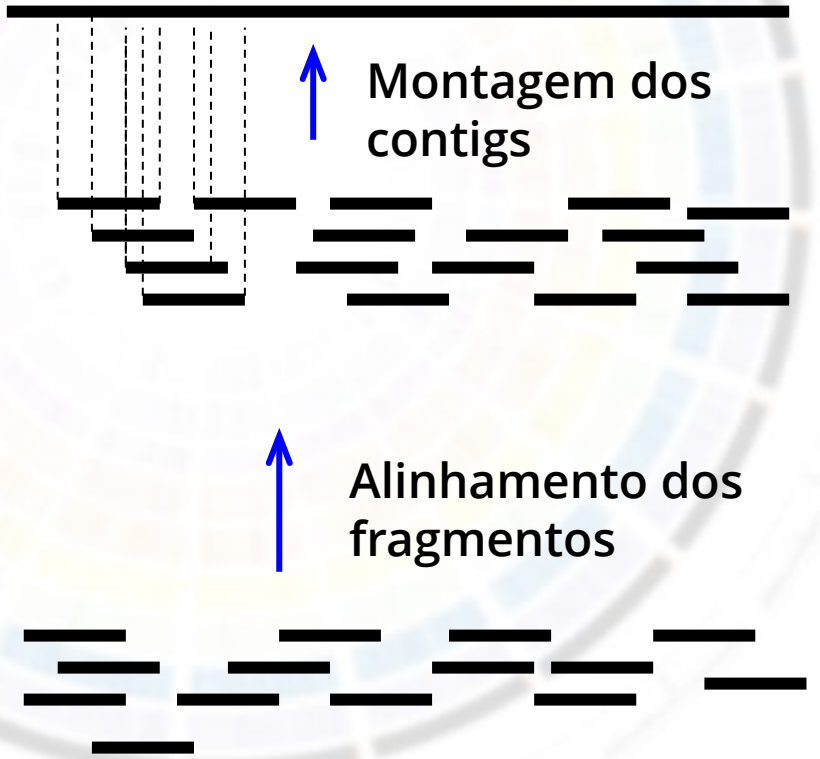
Genômica: sequenciamento do genoma

Biologia Molecular ⇄ Bioinformática

Para o sequenciamento de DNA



Para a montagem da sequência



Sequenciamento de nova geração

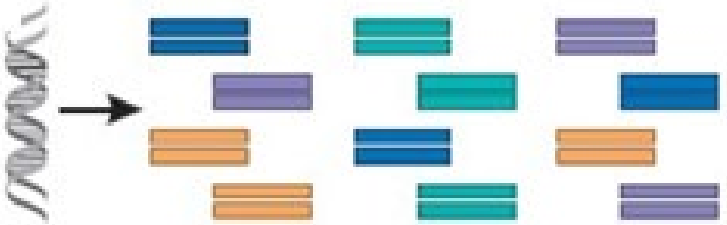
- Disputas entre tecnologias concorrentes
- Várias delas já estão completamente *falecidas*

Plataform	Read lenght (bp)	Throughput	Reads	Runtime
SOLiD	75	320 Gb	1.4 B	6 - 10 d
Ilumina	150	500 Gb	2 – 4 B	1 – 11 d
454	600 - 1000	700 Gb	1 M	23 h
Ion Proton	200 – 400	15 Gb	80 M	4 h
Pacific BioSciences	20 kb	1 Gb	55.000	4h

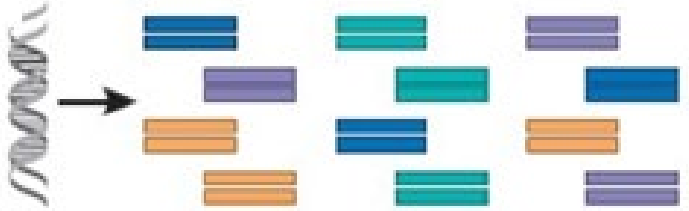
Sequenciamento de nova geração

- Principal diferença em relação a geração anterior: necessidade de clonagem

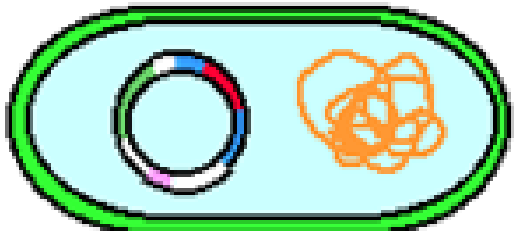
1ª Geração



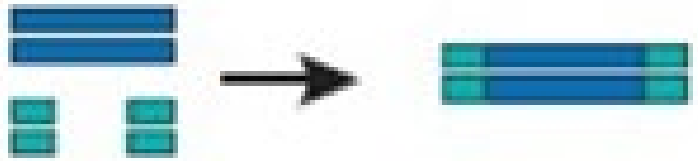
2ª Geração



Clonagem e amplificação



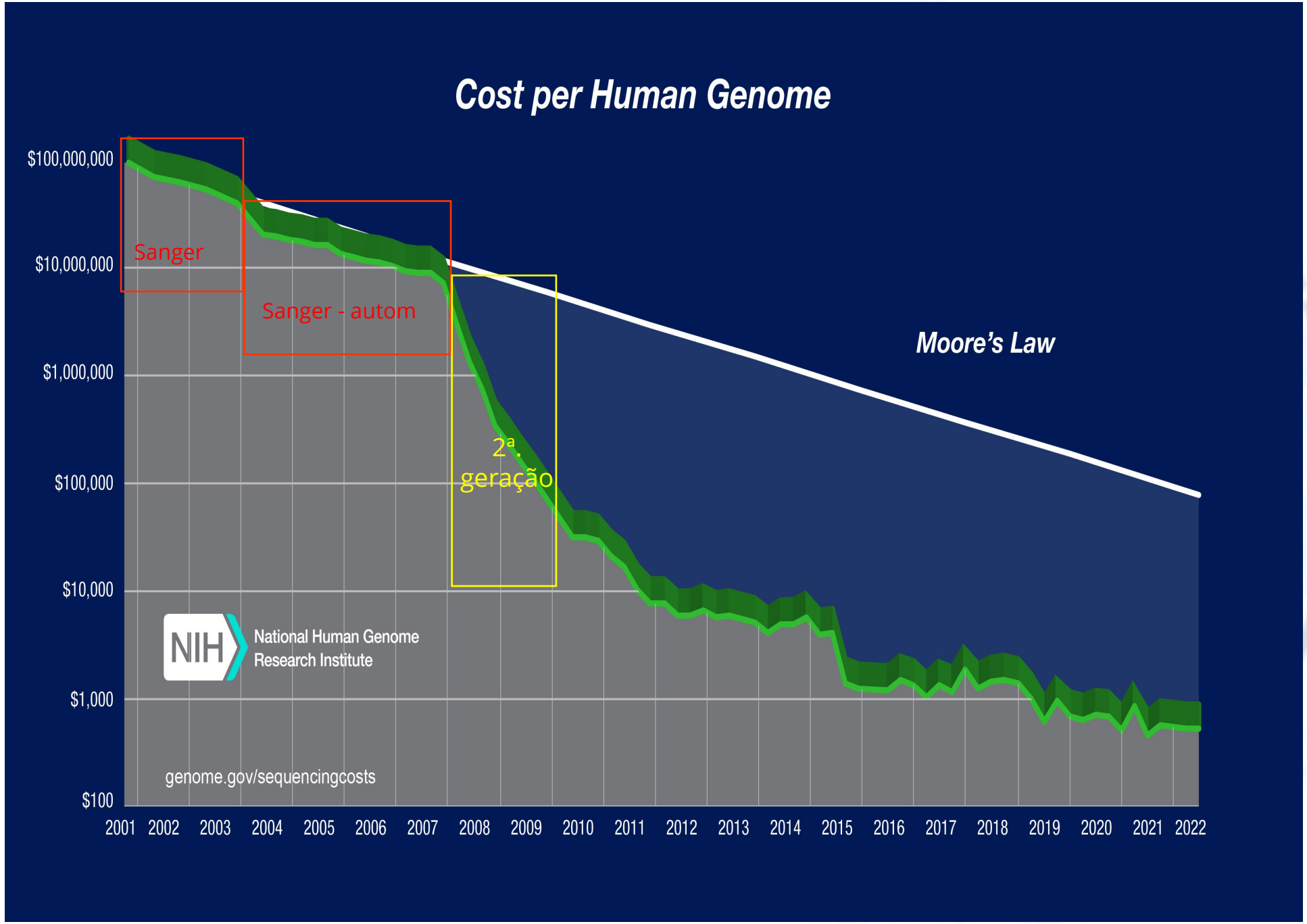
Ligação de adaptadores e amplificação



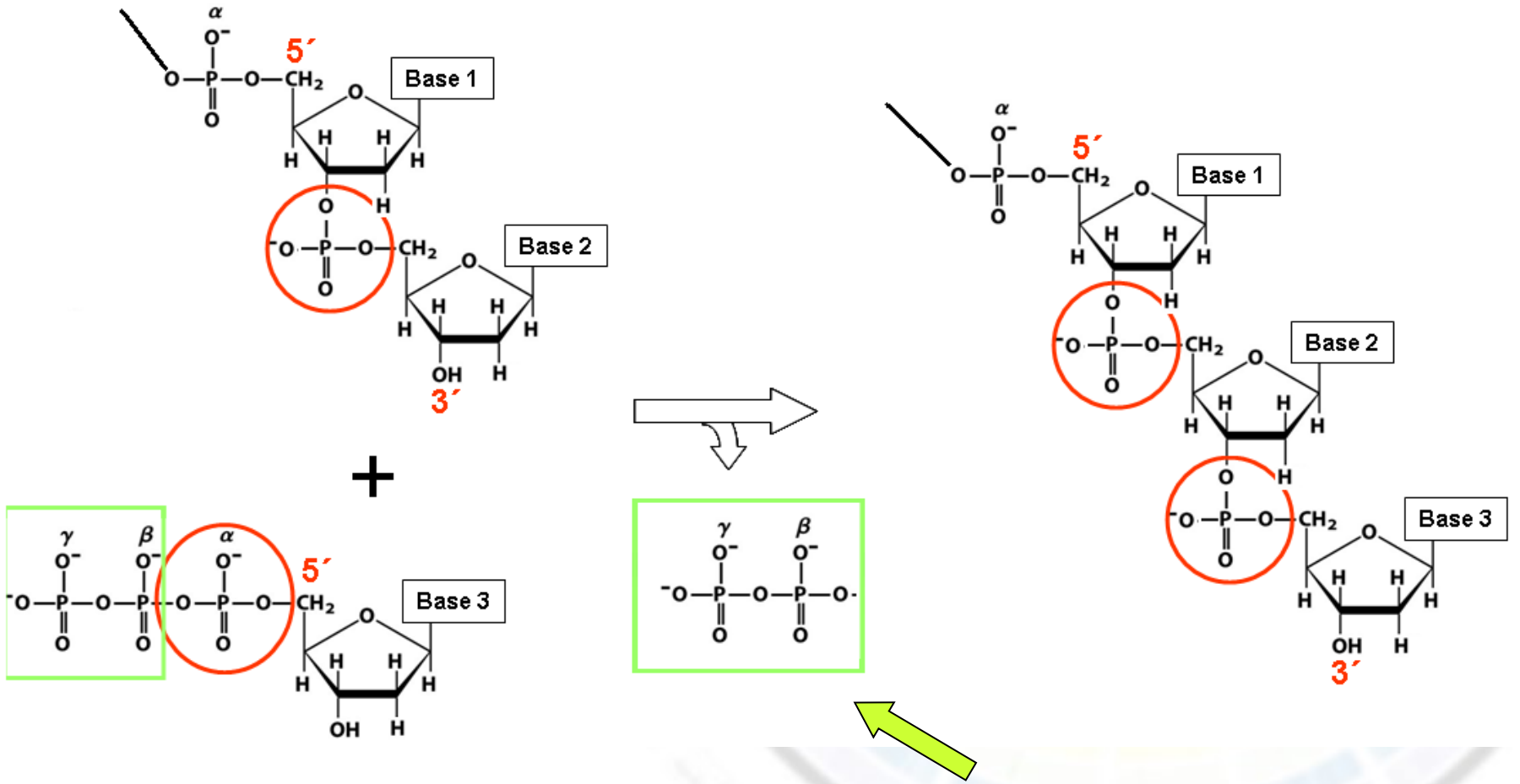
Sequenciamento de DNA de nova geração (“Next-Generation DNA Sequencing”)

Obtenção de uma grande quantidade de dados
(“high throught put sequencing”)

http://www.youtube.com/watch?v=_ApDinCBt8g


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Sanger
Sanger - autom
2ª. geração
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022

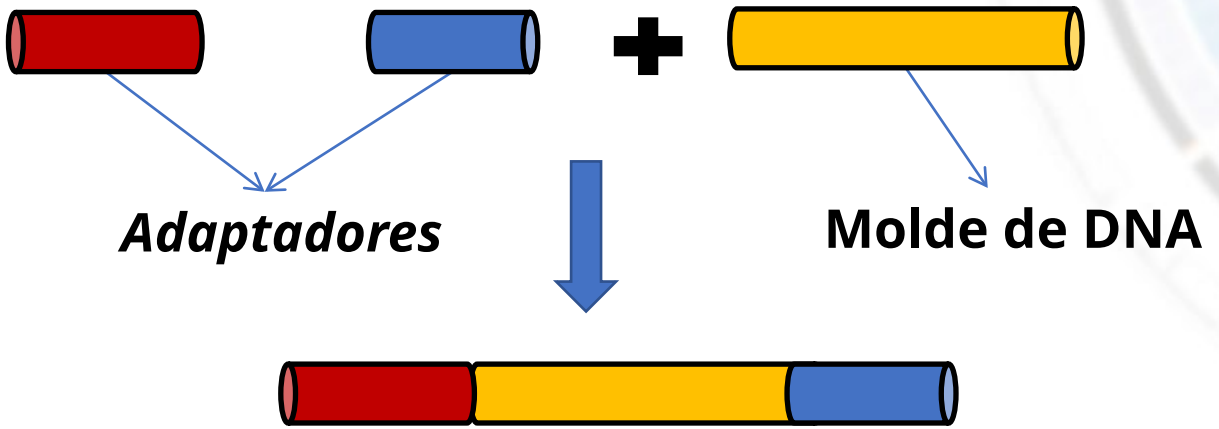
⇒ Pirosequenciamento – sistema 454-Roche® (2004)

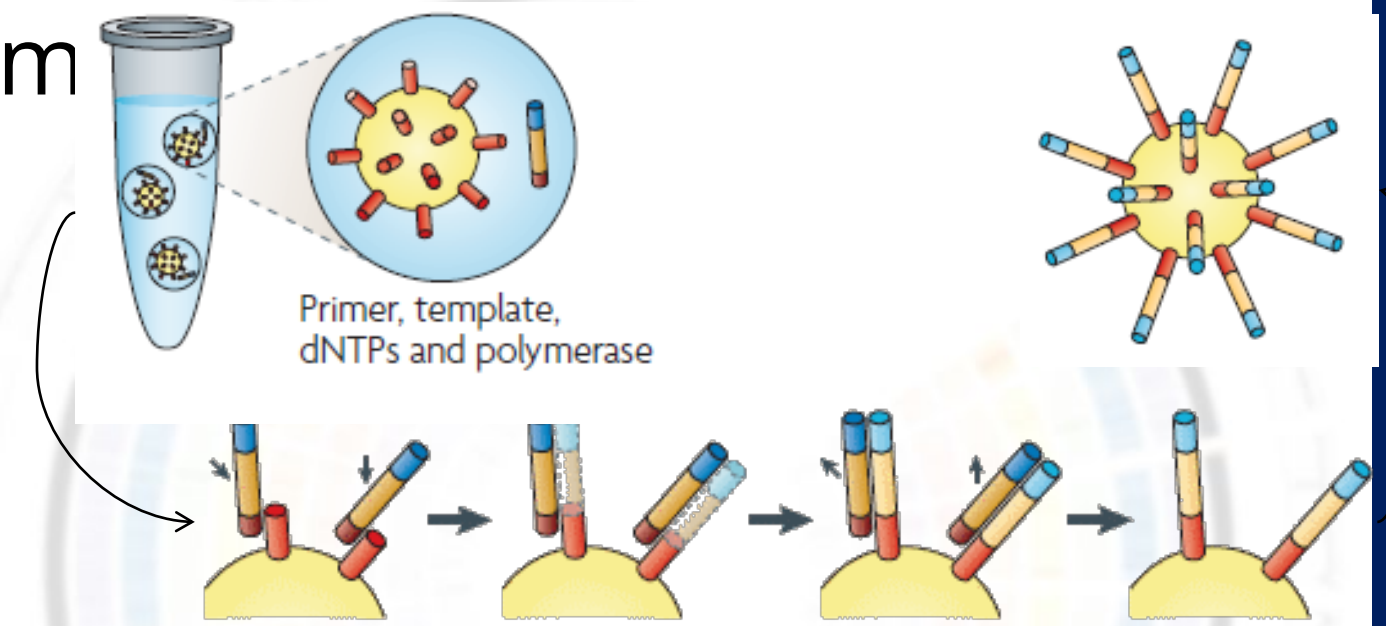


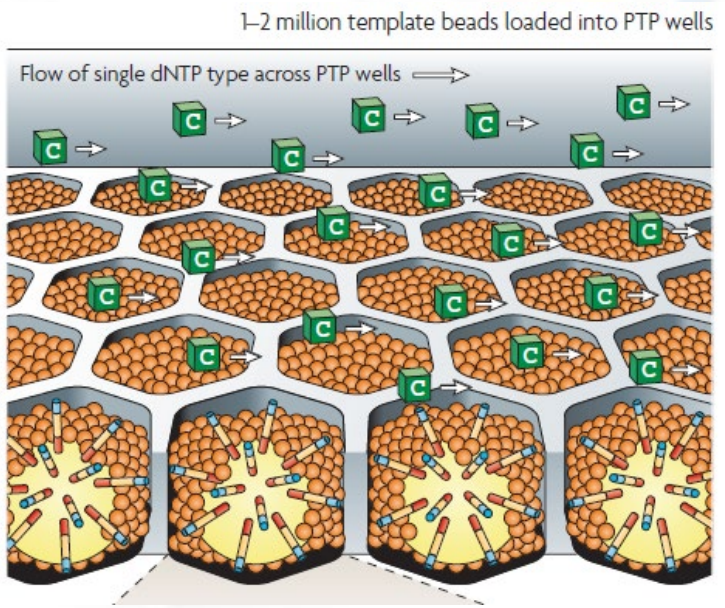
http://www.youtube.com/watch?v=bFNjxKHP8Jc
http://www.youtube.com/watch?v=JNqXgLKOzKU

Pirossequenciamento (454/Roche)

- Já foi descontinuada em 2016
- 1. Ligação de adaptadores
- 2. PCR em emulsão
- 3. Sequenciamento

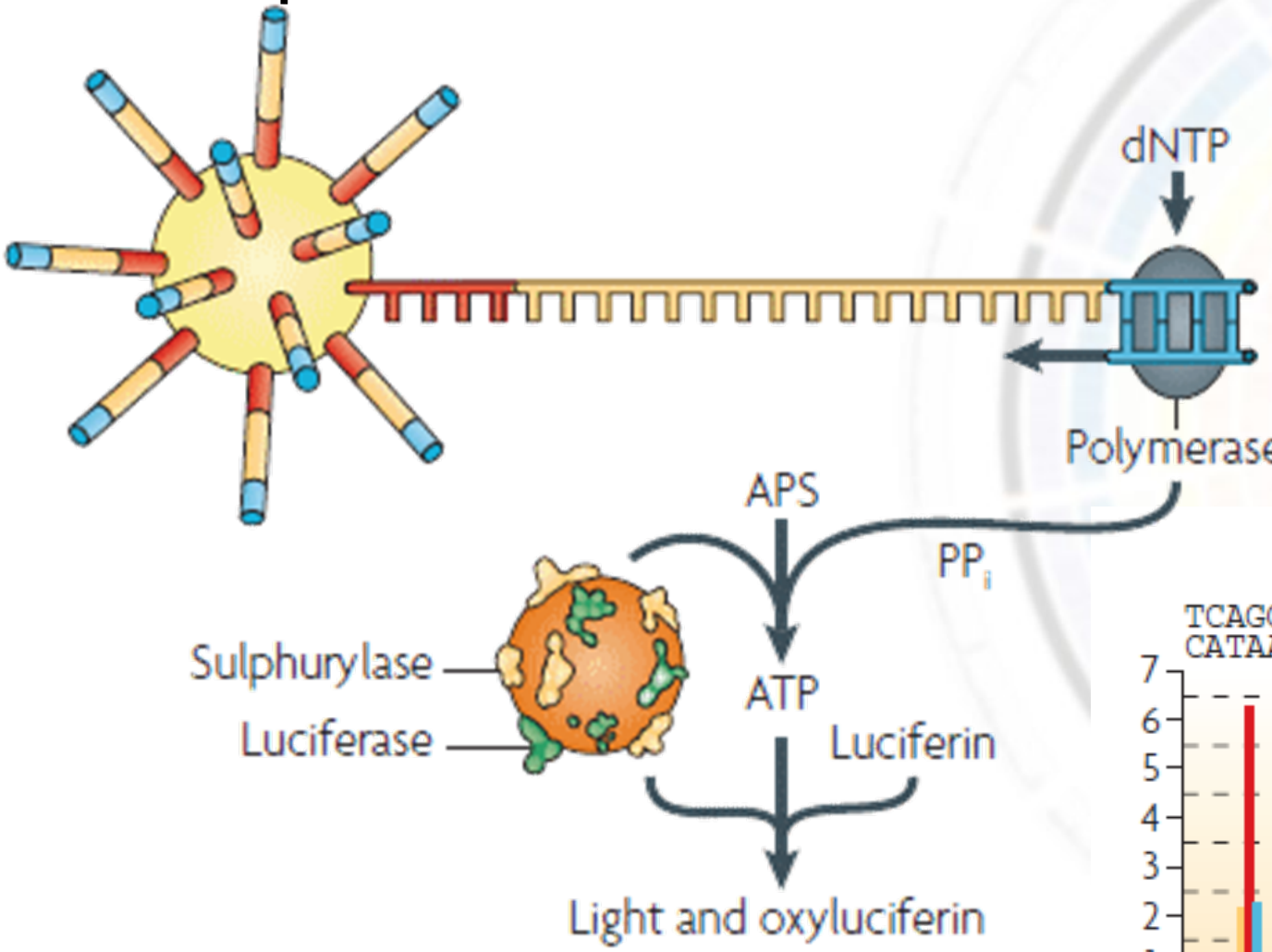


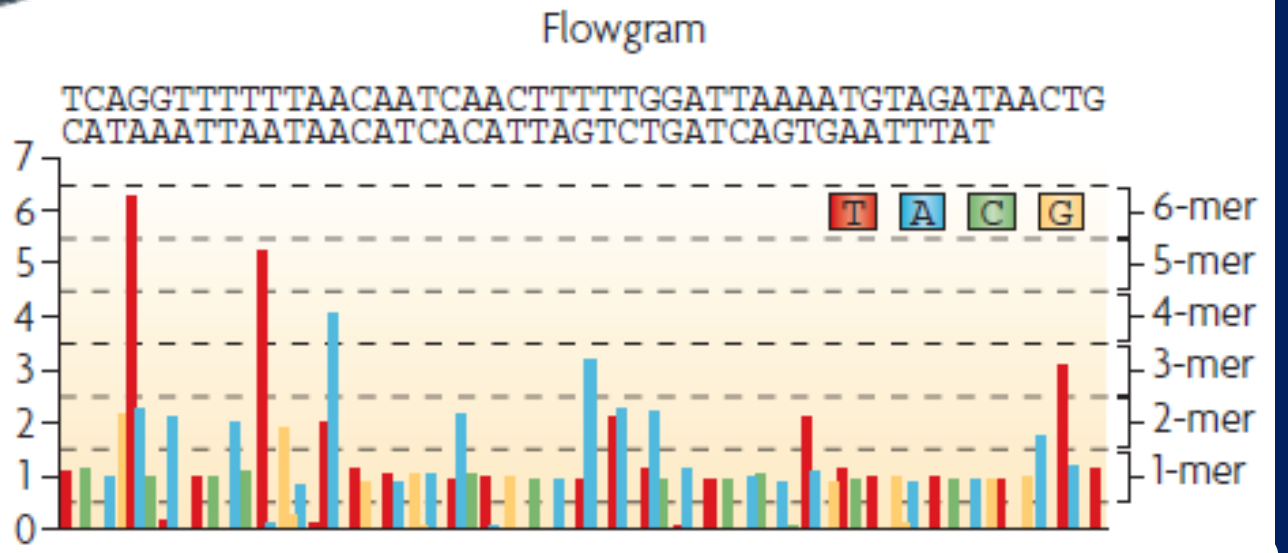




Pirossequenciamento (454/Roch

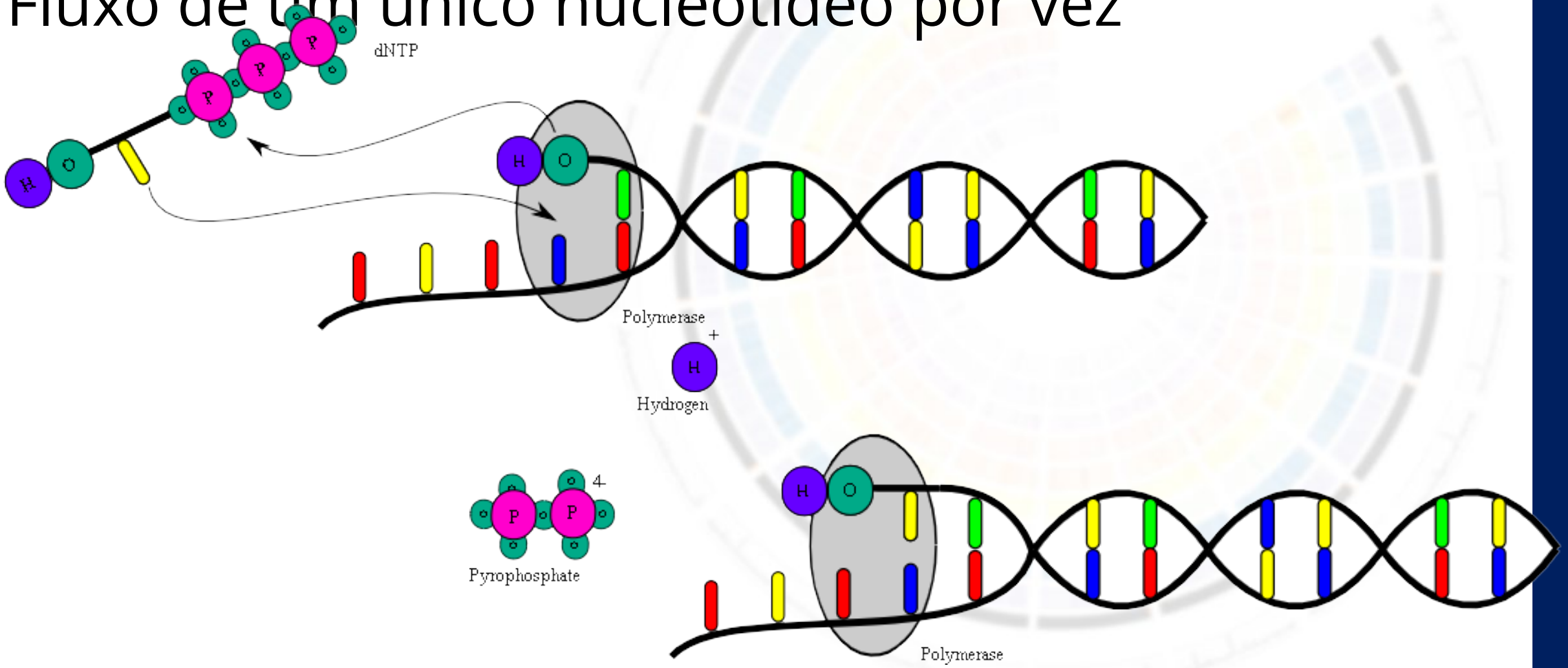
- 3. Sequenciamento





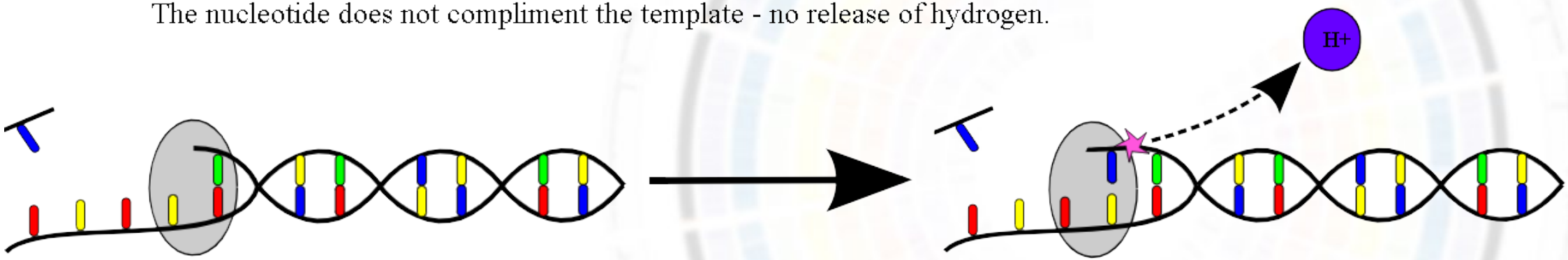
Ion Torrent/Proton/PGM

- Sequenciamento por alteração de pH
- Fluxo de um único nucleotídeo por vez

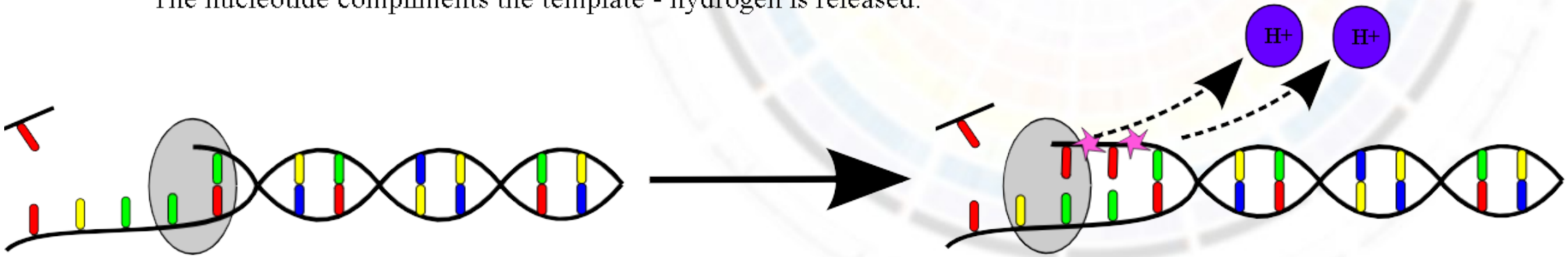


Ion

The nucleotide does not compliment the template - no release of hydrogen.

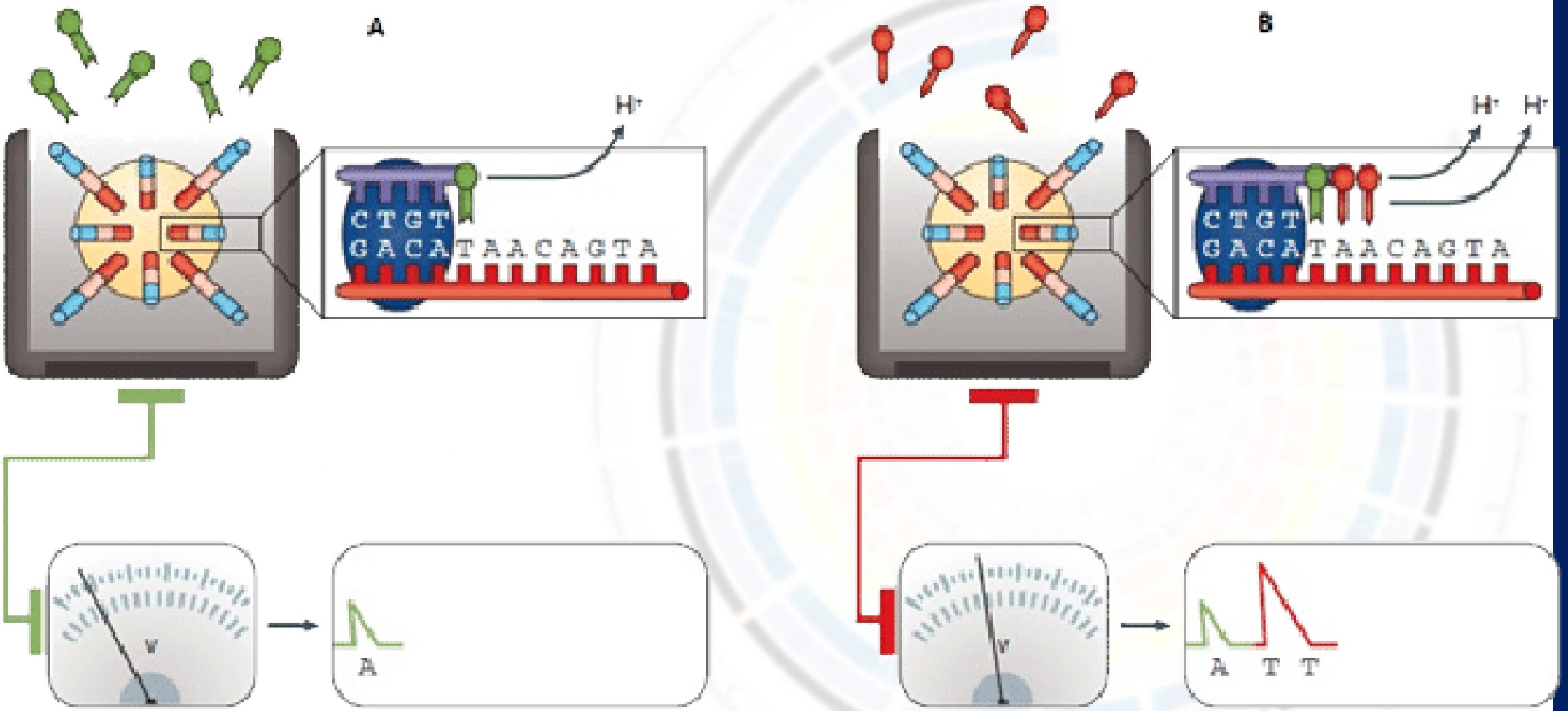


The nucleotide compliments the template - hydrogen is released.



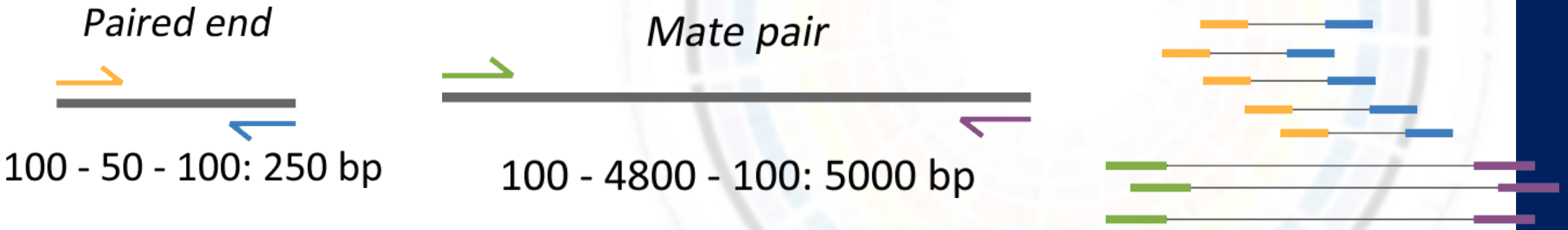
The nucleotide compliments several bases in a row - multiple hydrogen ions are released.

Ion



Capacidades atuais das duas principais tecnologias

- Ion PGM
  - ~50 gb (50 bilhões de bases)
  - 130 milhões de fragmentos
  - Sequências pequenas (até 400 pb)
- **Illumina HiSeq 2500**
  - 9 gb – 1024gb (9 bilhões a 1 trilhão de bases)
  - Até 4 bilhões de fragmentos lidos
  - Sequências pequenas (até 2x250bp)



- **PacBio SMRT**
  - Output com qualidade: até 36 bilhões de bases
  - Até 8 milhões de fragmentos lidos
  - Sequências longas (30-100 mil bp)

Solexa-Illumina (2006)

Etapas:

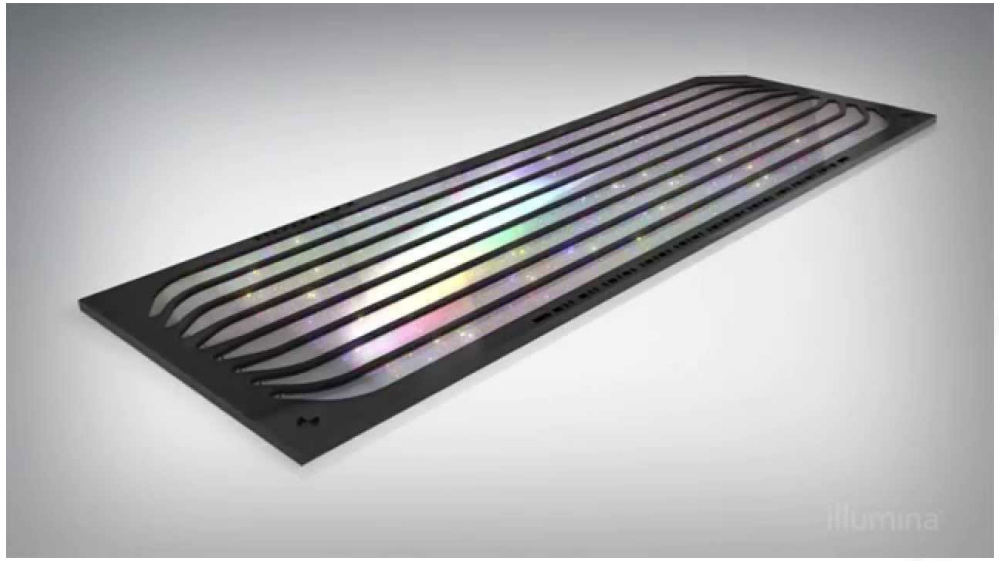
1) Ligação de adaptadores

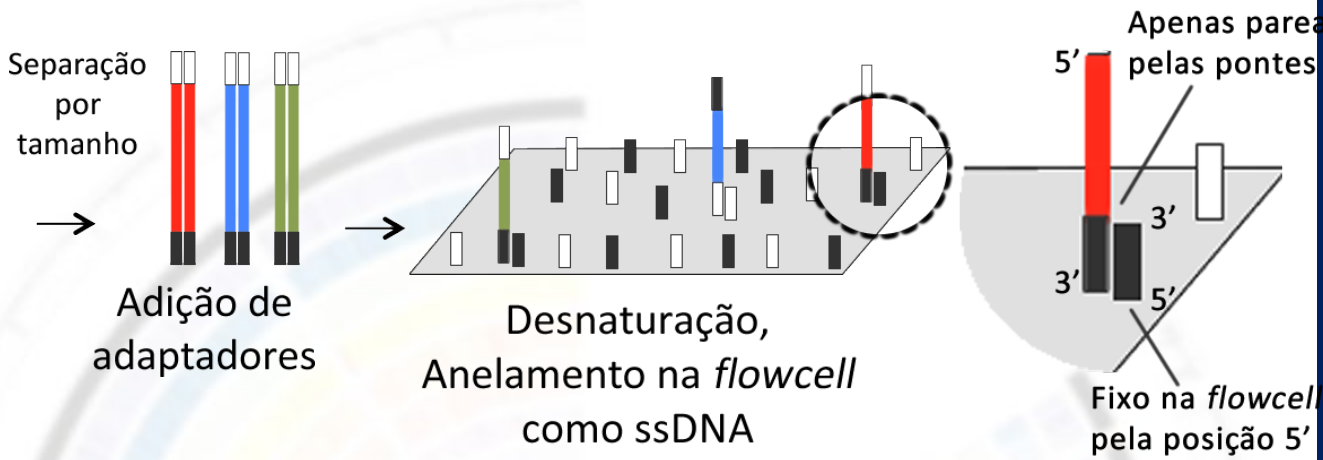
2) PCR em ponte (formação de clusters)

3) Sequenciamento

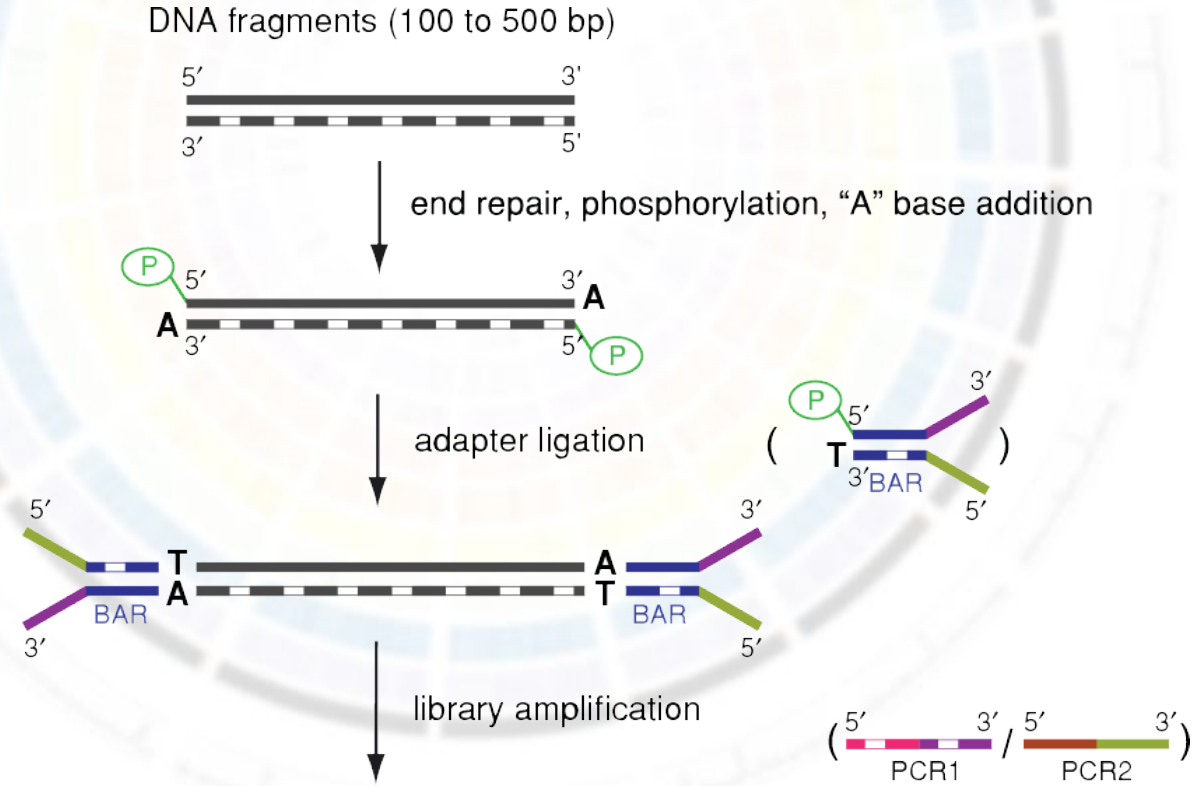
https://www.youtube.com/watch?v=womKfikWlxM
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina "*Sequencing by synthesis*"

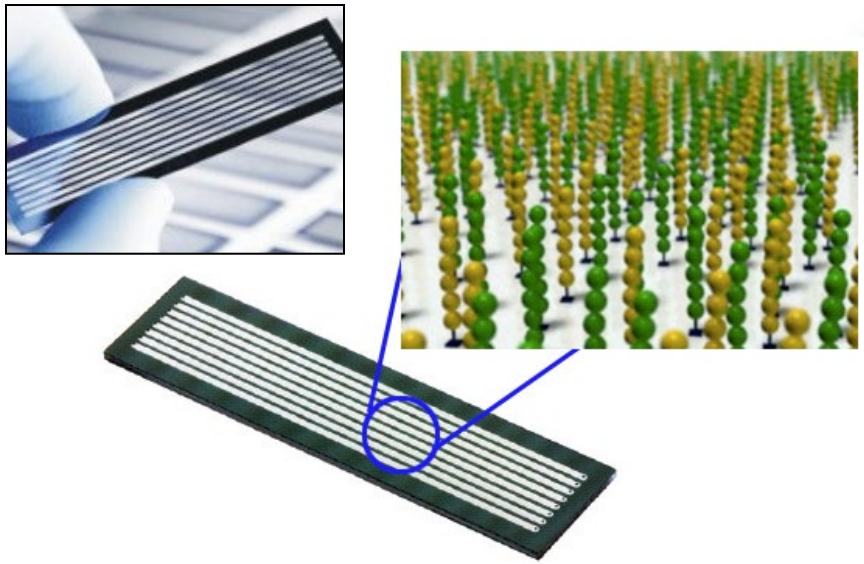


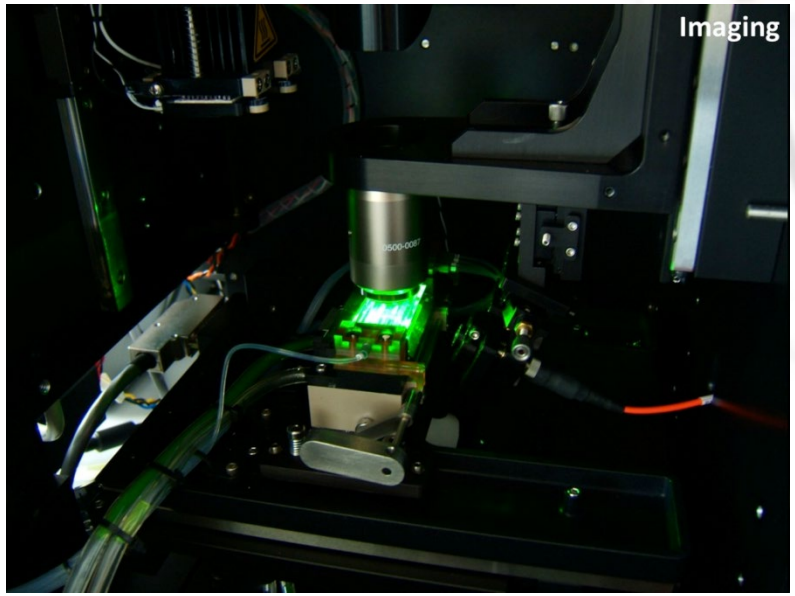


- *Flowcell* dividida em 8 canais (modelos com 2)



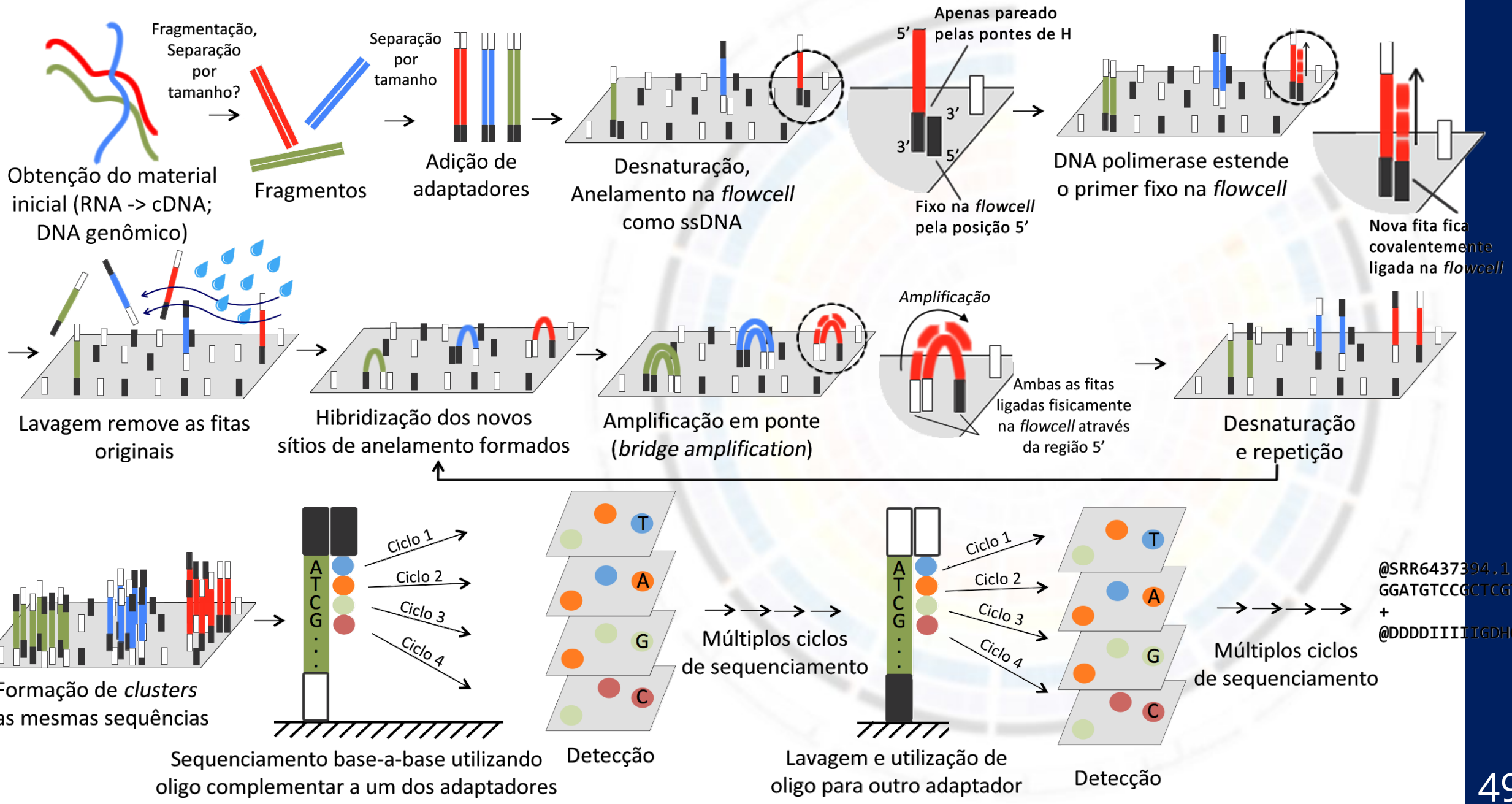
A amostra é aplicada em uma lâmina aonde haverá amplificação do fragmento seguido de sequenciamento

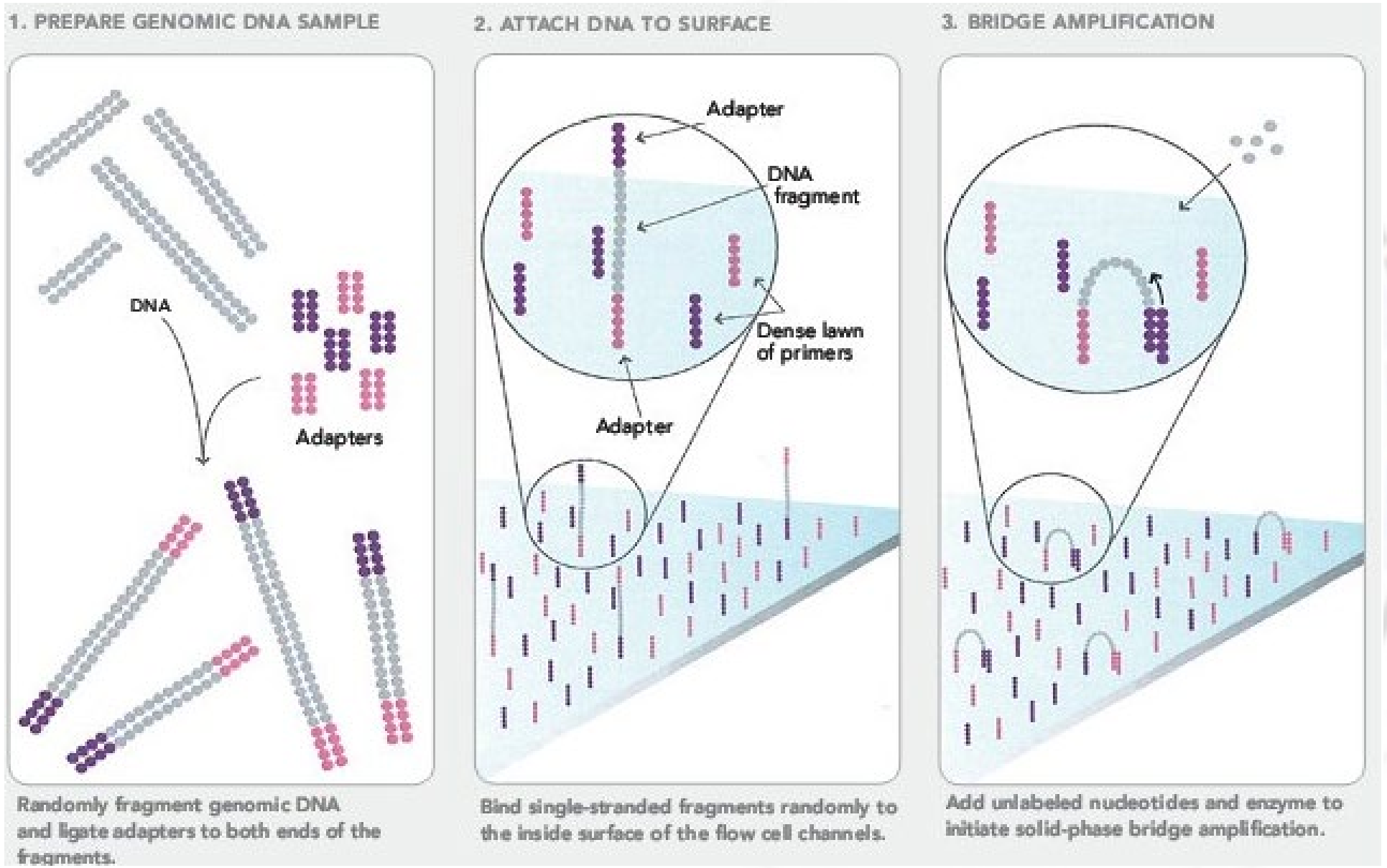


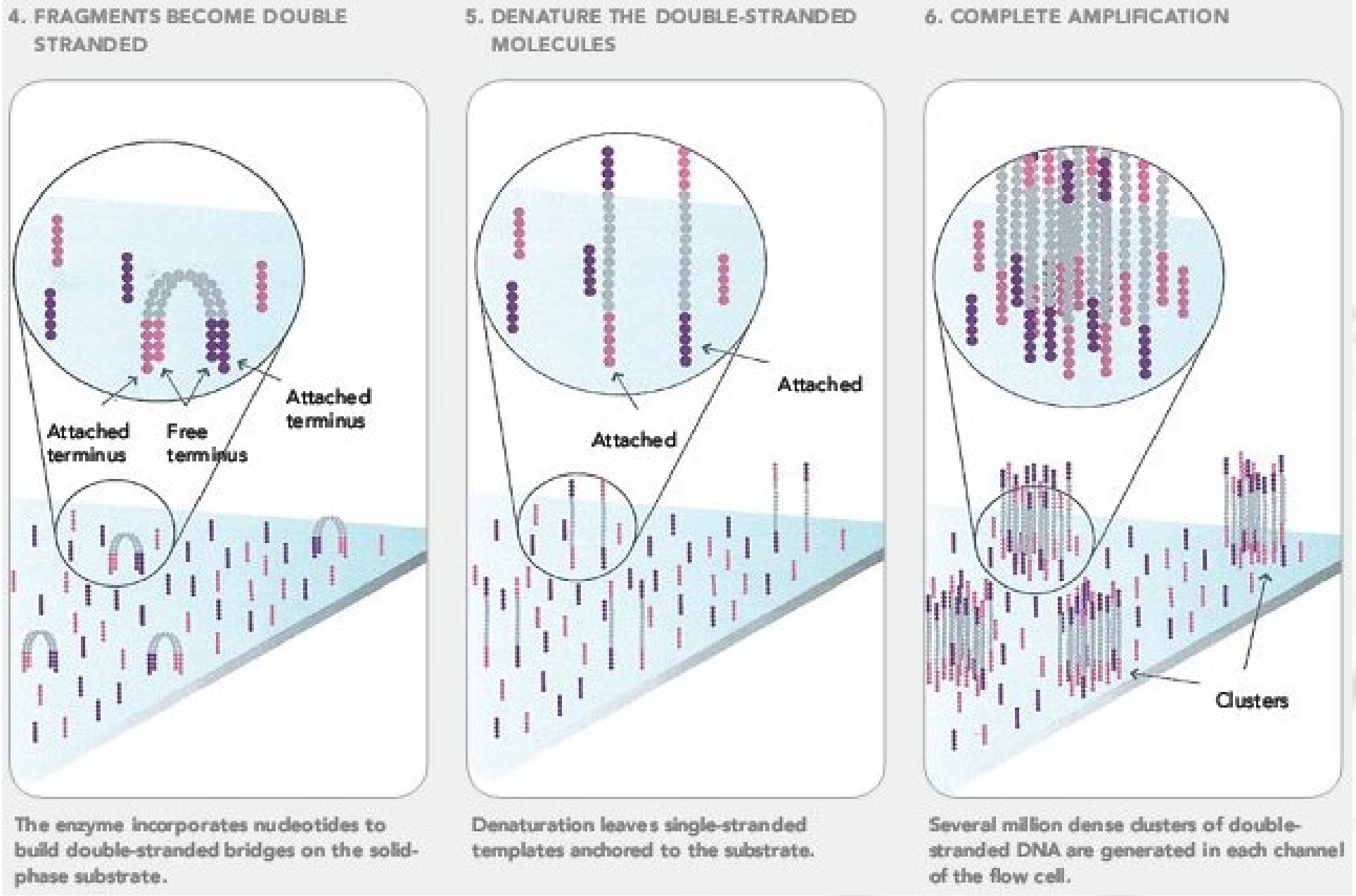


Todo o processo de sequenciamento é realizado na lâmina acoplada ao aparelho que irá detectar os sinais de fluorescência correspondente a cada base

Illumina *flowcell* & *Sequencing by synthesis* (SBS)

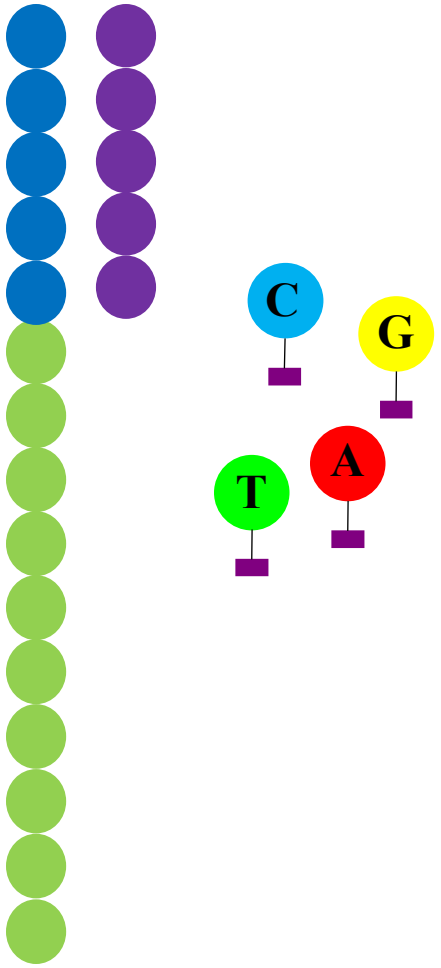



1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

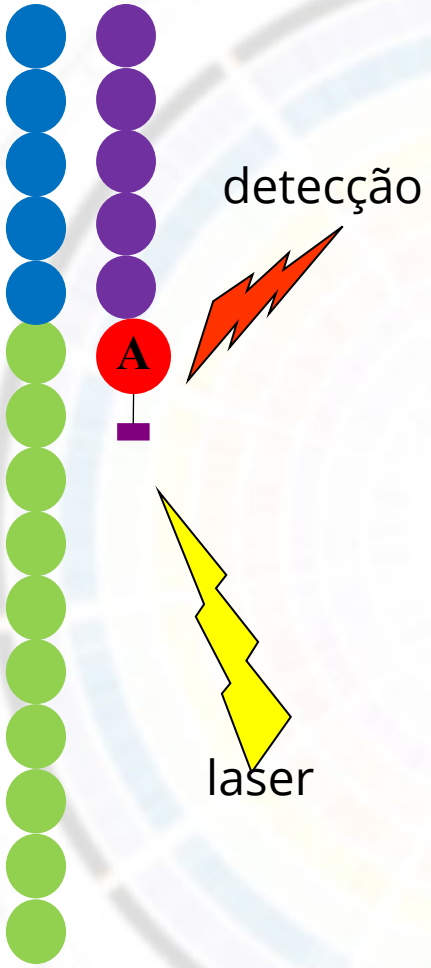


Obs.: após o processo de amplificação "em ponte" e formação do "cluster", somente uma das fitas de DNA permanece ligado à lâmina

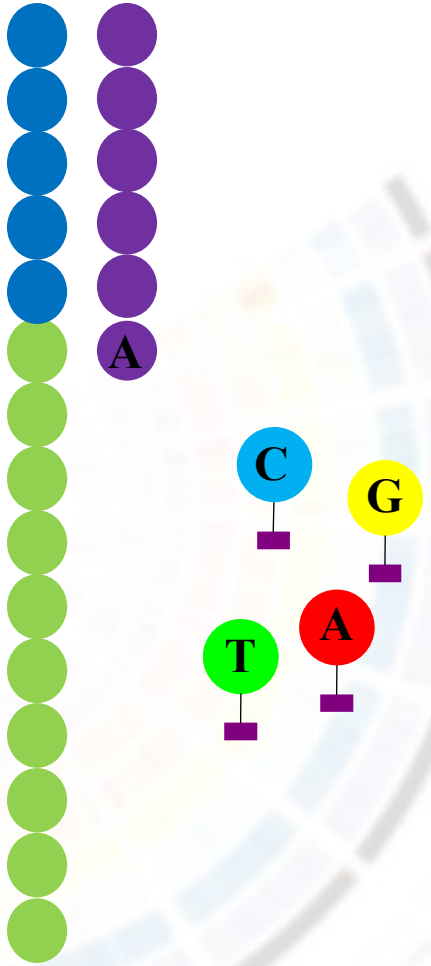
Sequenciamento



A partir do "primer" ocorre a incorporação de nucleotídeo marcado



O nucleotídeo incorporado é detectado

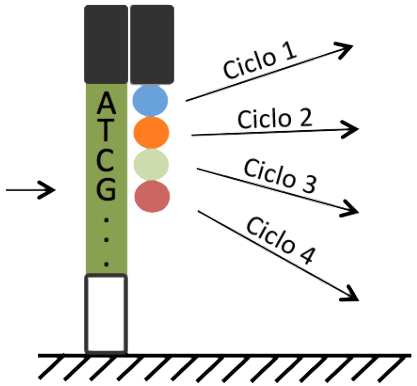


O terminal do nucleotídeo incorporado é modificado permitindo a adição do próximo nucleotídeo

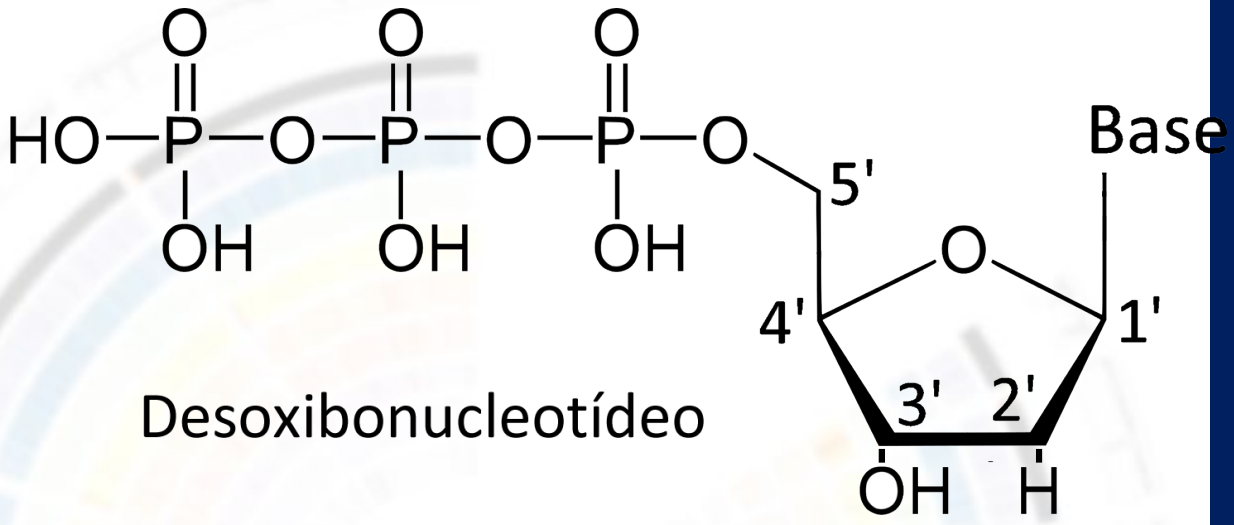
Illumina *"Sequencing by synthesis"*

- ddNTP com bloqueio fluorescente na posição 3’ OH

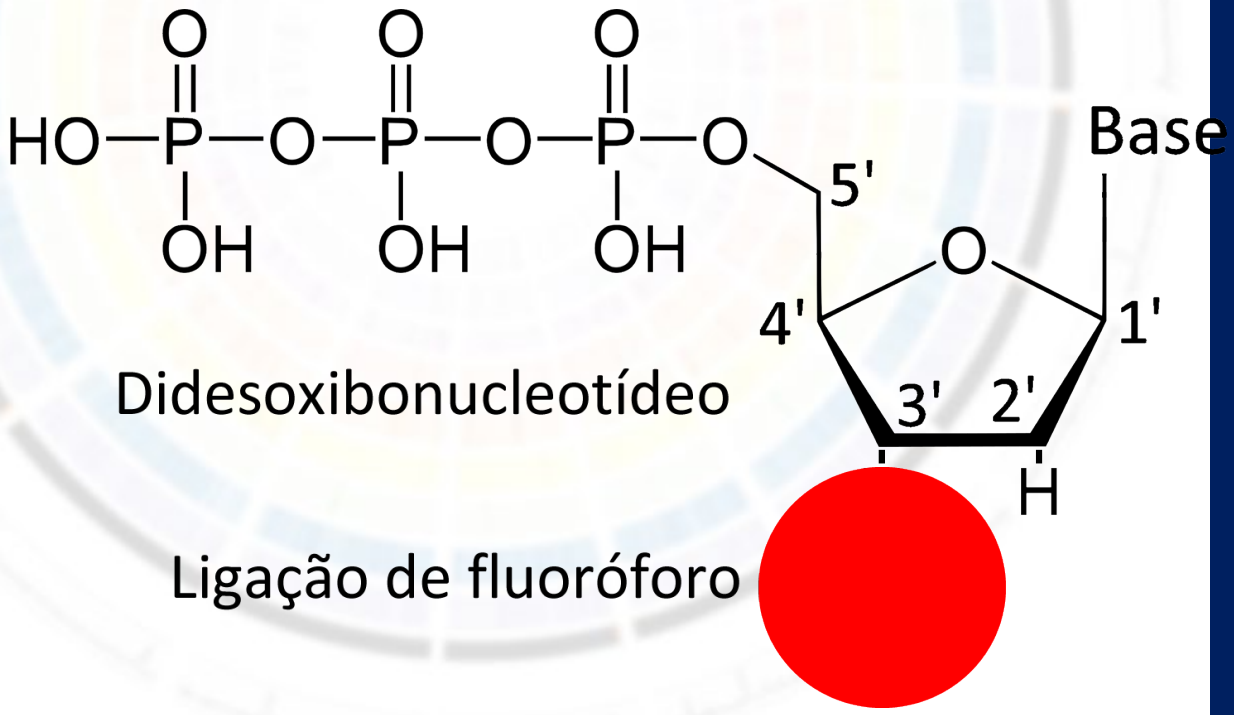
- Ciclos de adição -> leitura -> desbloqueio -> adição



Sequenciamento base-a-base utilizando oligo complementar a um dos adaptadores



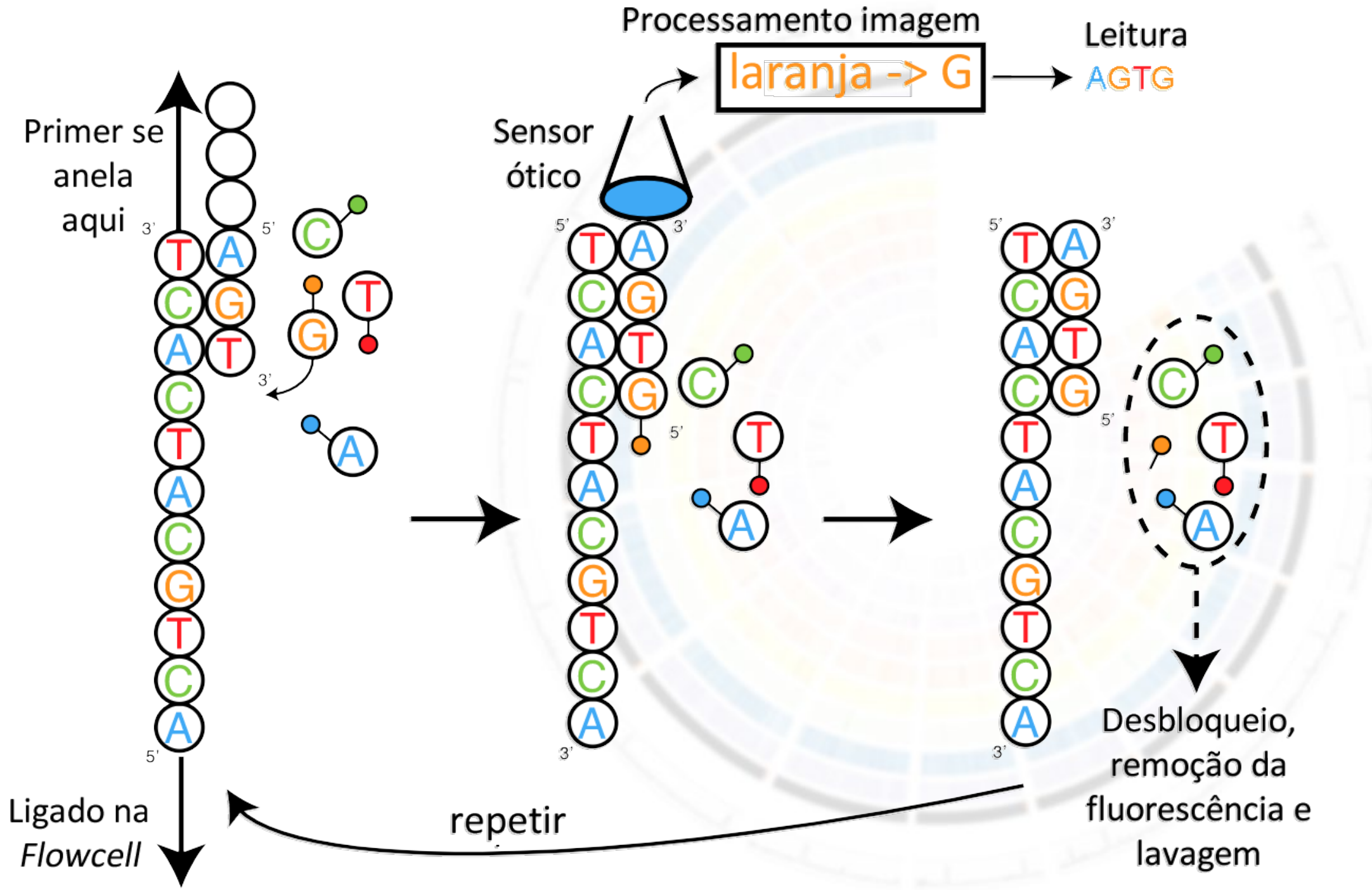
Desoxibonucleotídeo



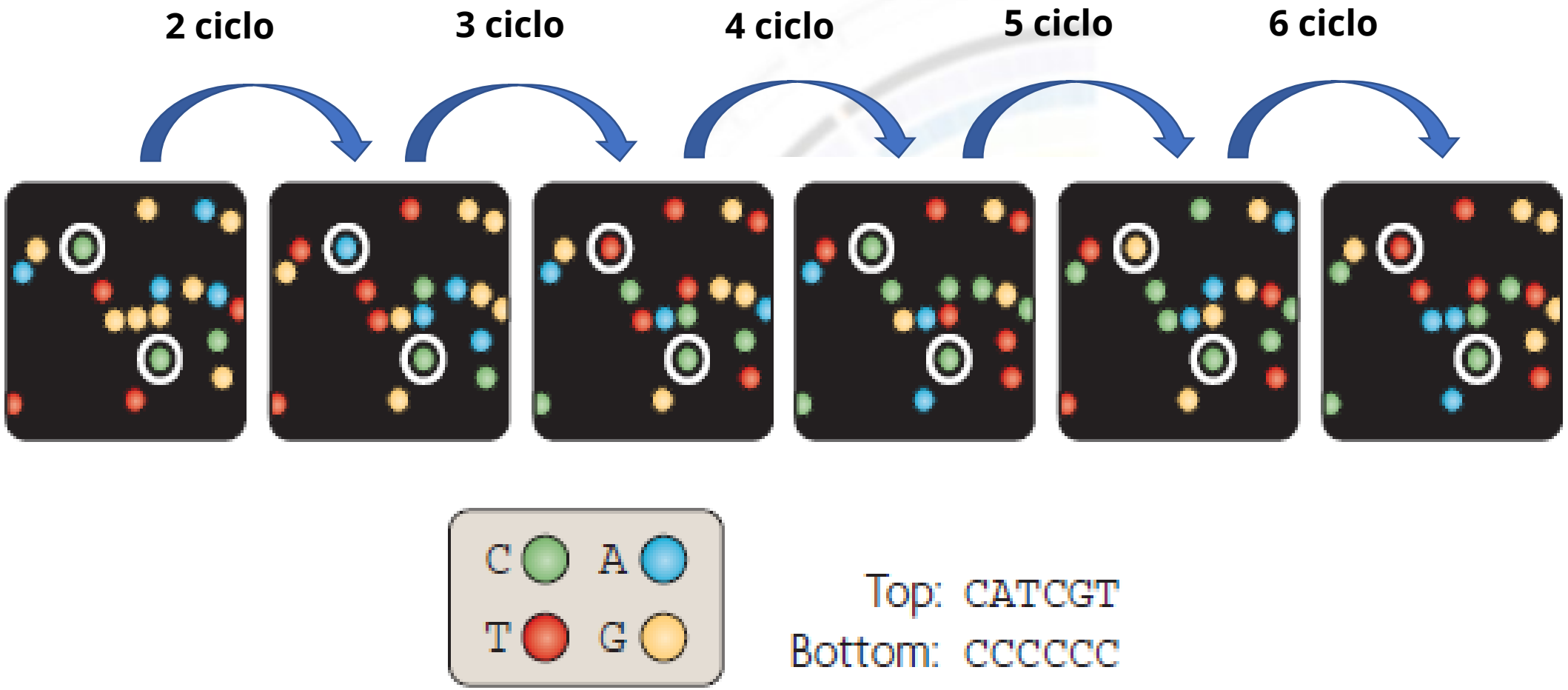
Didesoxibonucleotídeo

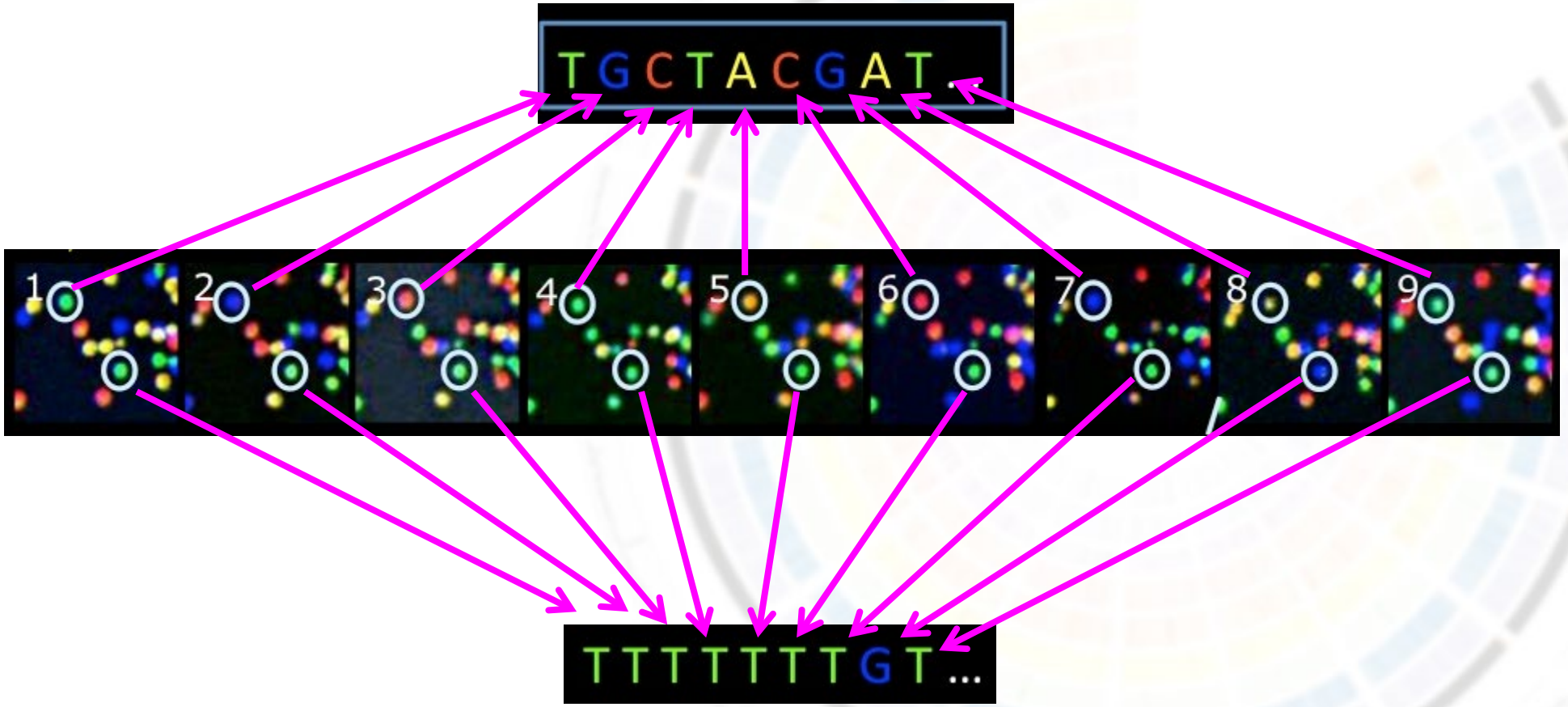
Ligação de fluoróforo

Illumina "*Sequencing by synthesis*"



Illumina *"Sequencing by synthesis"*



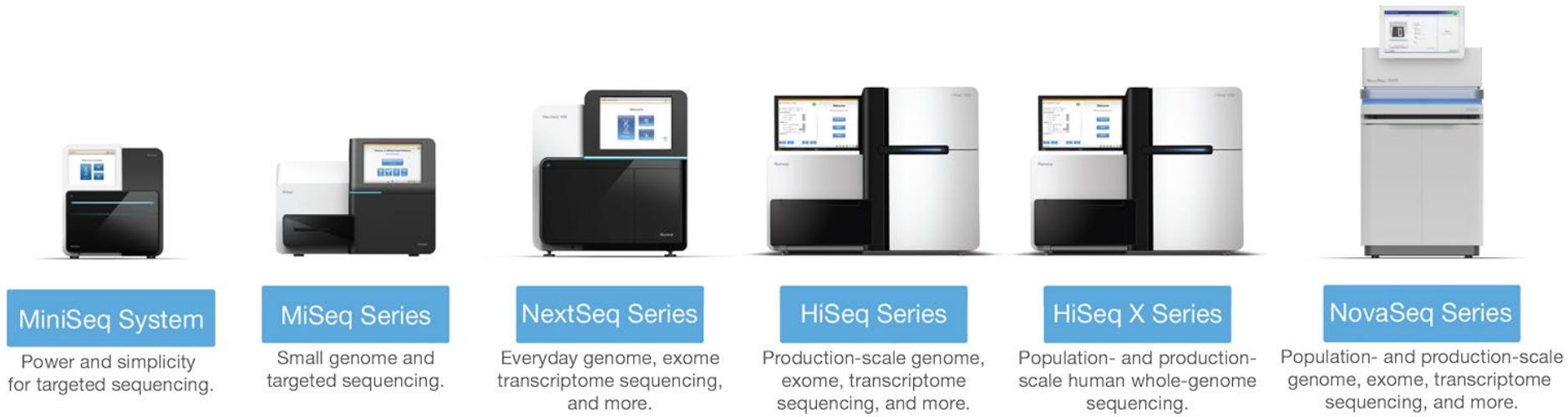


Sequência de incorporação de nucleotídeos

Ilumina

MiSeq
NextSeq

Reads max
15 Gb x 400 Gb
Tamanho fragmentos
2x 300 – 600 pb



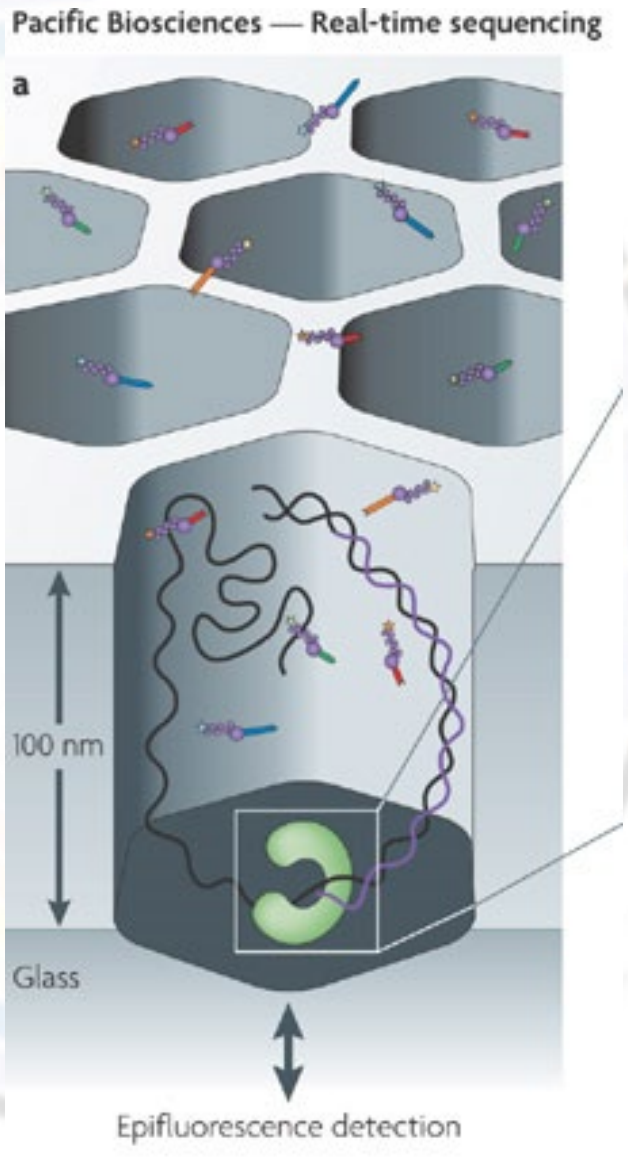
Sequenciamento de nova geração (larga escala, 3ª geração)

- Capacidade de sequenciamento de uma única molécula de DNA
- Sem necessidade de amplificação
- Altíssima capacidade de geração de um grande volume de dados em curto período de tempo: um genoma humano pode ser sequenciado em uma única corrida

Pacific Biosciences SMRT Sequencing (Single Molecule Real Time Sequencing)

- Janela de observação em nano-escala (ZMW, *zero-mode waveguide*), com um volume extremamente reduzido, suficiente para visualizar a incorporação de um único nucleotídeo pela DNA polimerase

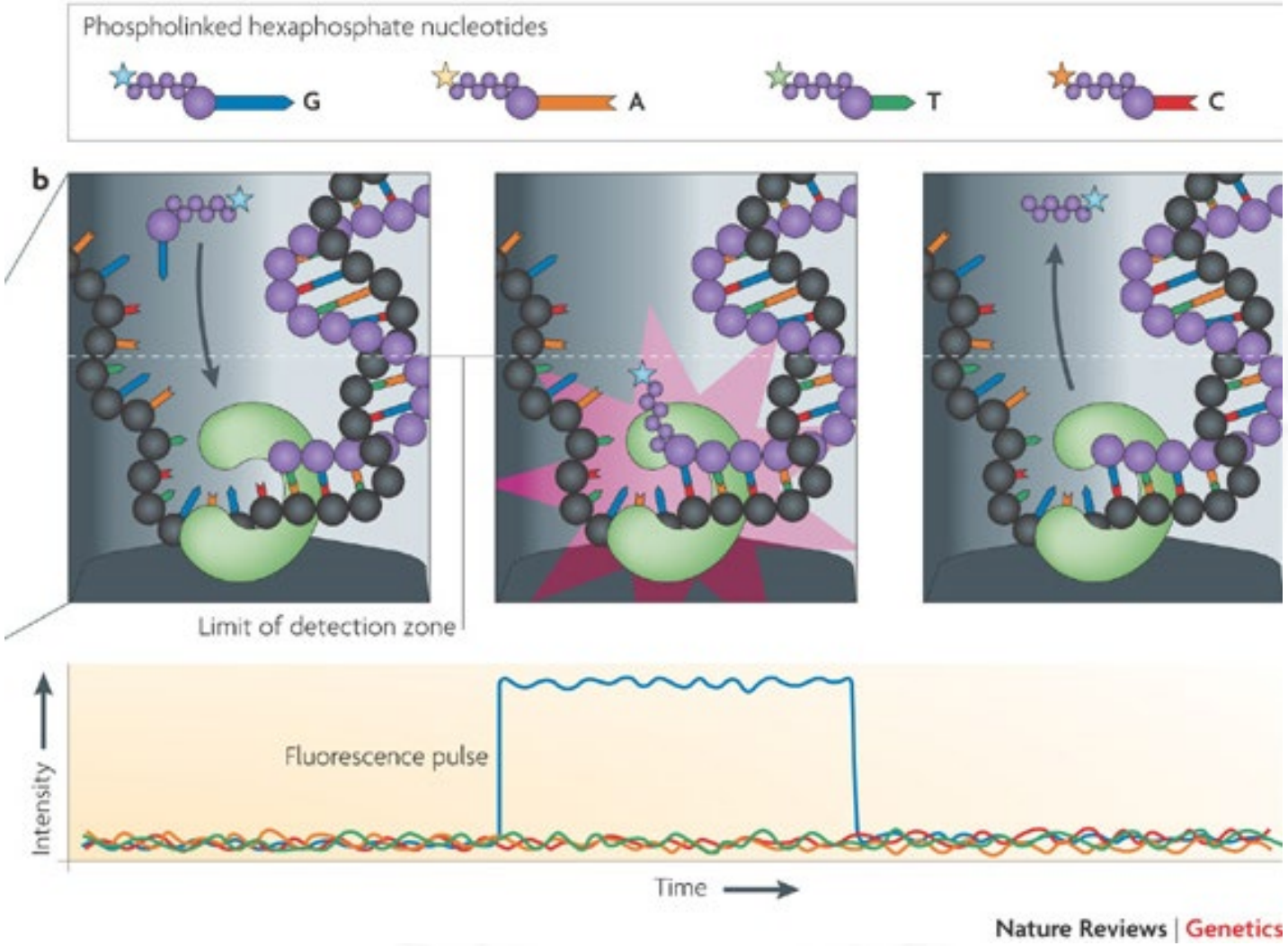
- Uma única DNA polimerase contendo uma única molécula de DNA molde é fixada no fundo da ZMW



Adaptado de:
METZKER, 2009. **Nature Reviews Genetics.** DOI: 10.1038/nrg2626

Pacific Biosciences SMRT Sequencing (Single Molecule Real Time Sequencing)

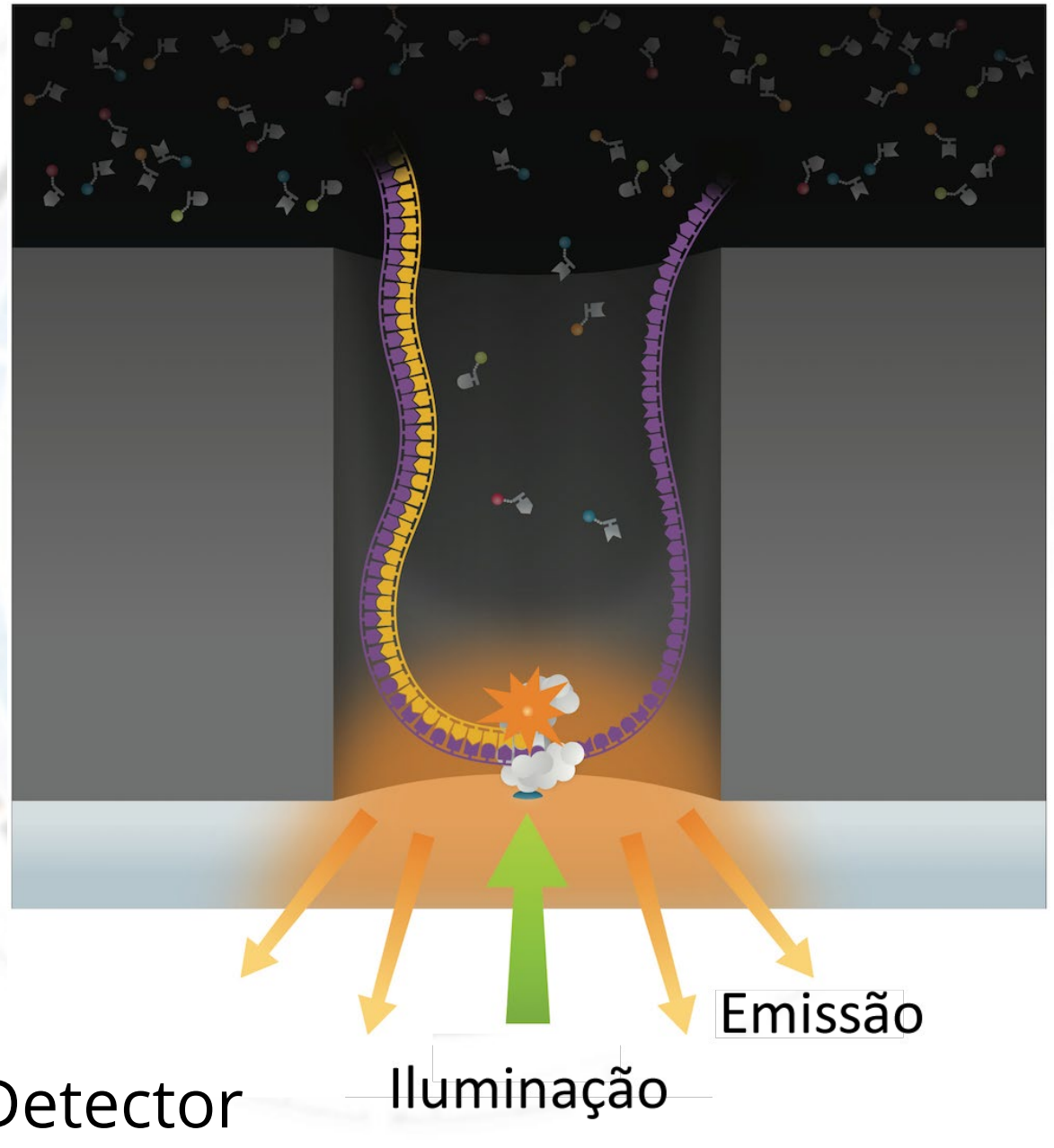
- Nucleotídeos com fluoróforos são utilizados, e quando um nucleotídeo é incorporado a marcação fluorescente é clivada, emitindo luz, que é detectada e transformada em dado de sequência



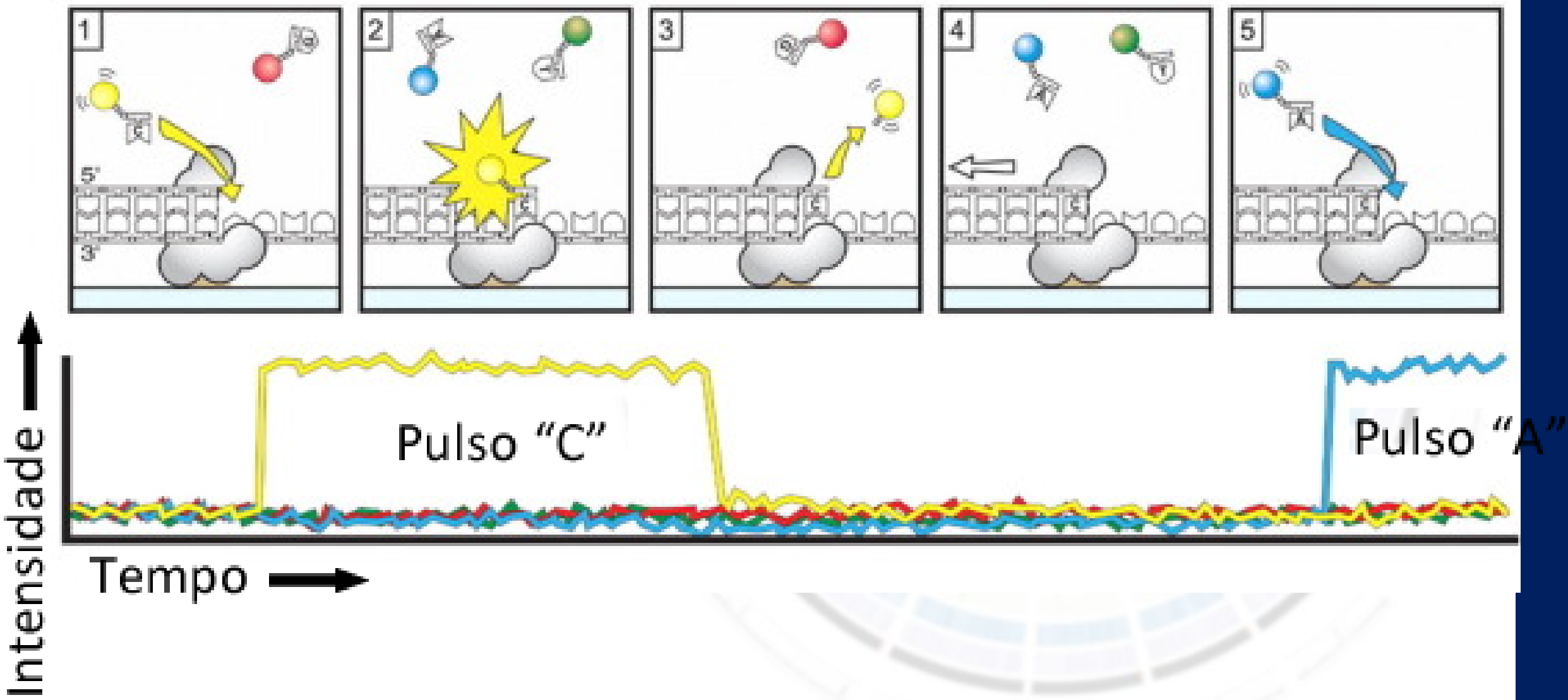
Adaptado de:
METZKER, 2009. **Nature Reviews Genetics.** DOI: 10.1038/nrg2626

PacBio SMRT Sequencing

- *Single-molecule real-time sequencing* (SMRT)
- *Zero-mode waveguide*
- ~70 nm no diâmetro e ~100 nm na profundidade
- A luz se comporta de maneira distinta em volumes pequenos
- Leituras muito longas, mas com alta taxa de erro



PacBio SMRT Sequencing

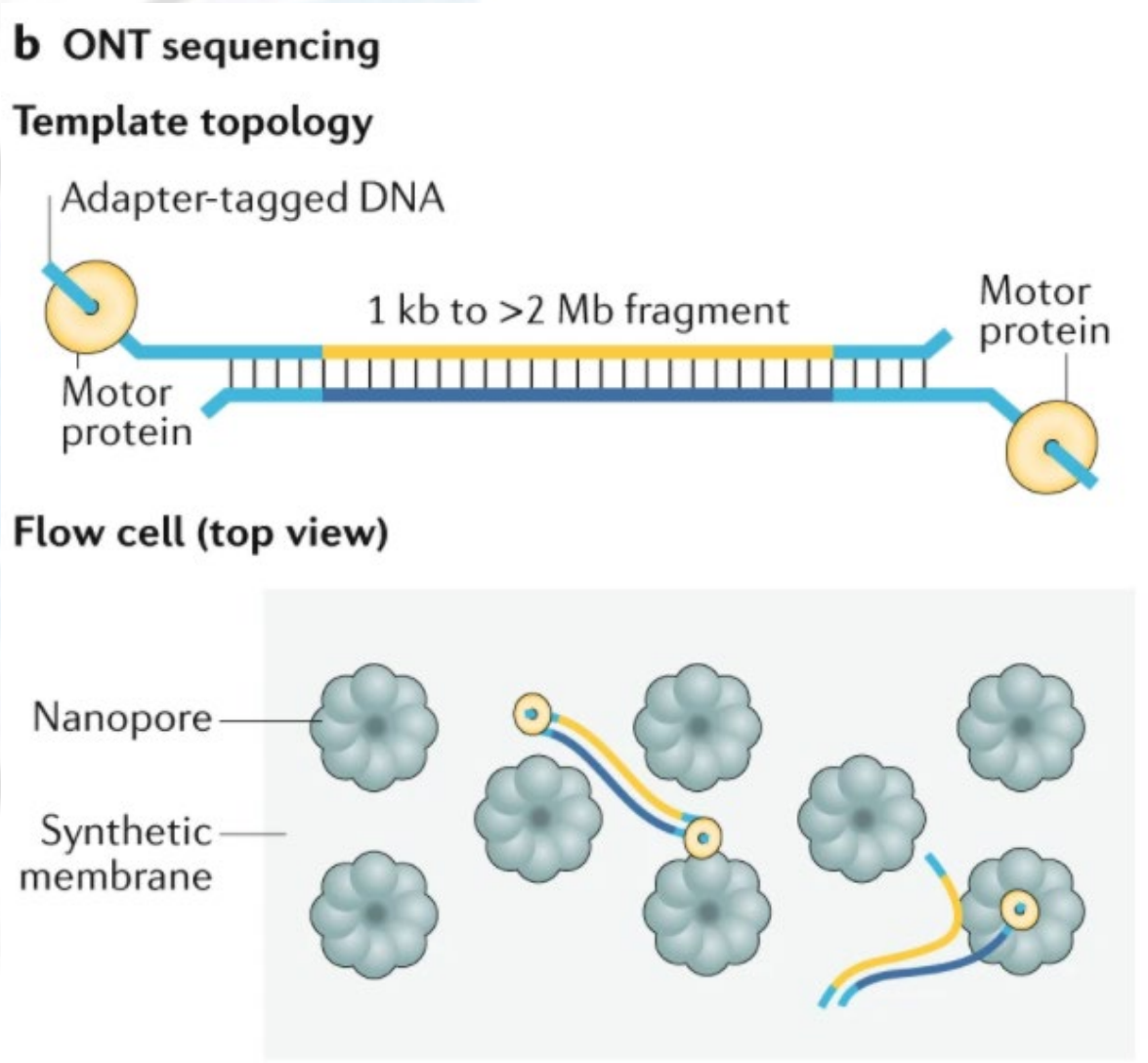


PacBio SMRT Sequencing

Oxford Nanopore Technologies

- DNA é marcado com adaptadores com proteínas motoras em uma ou ambas as extremidades e é combinado à proteínas carregadoras, que o direcionarão aos nanoporos

- A plataforma de sequenciamento contém milhares de nanoporos de proteicos associados à uma membrana sintética

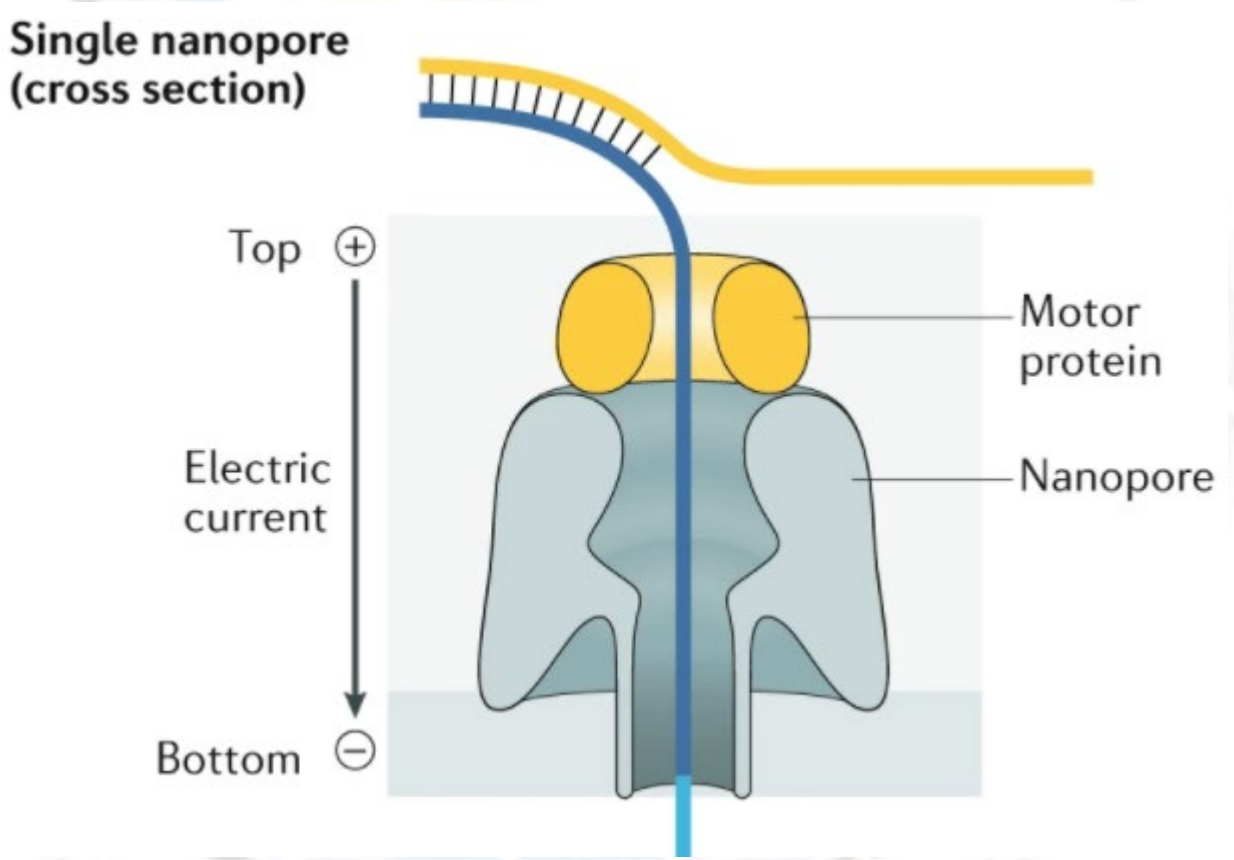


Adaptado de:
Logsdon et al. 2020. **Nature Reviews Genetics.** DOI: 10.1038/s41576-020-0236-x

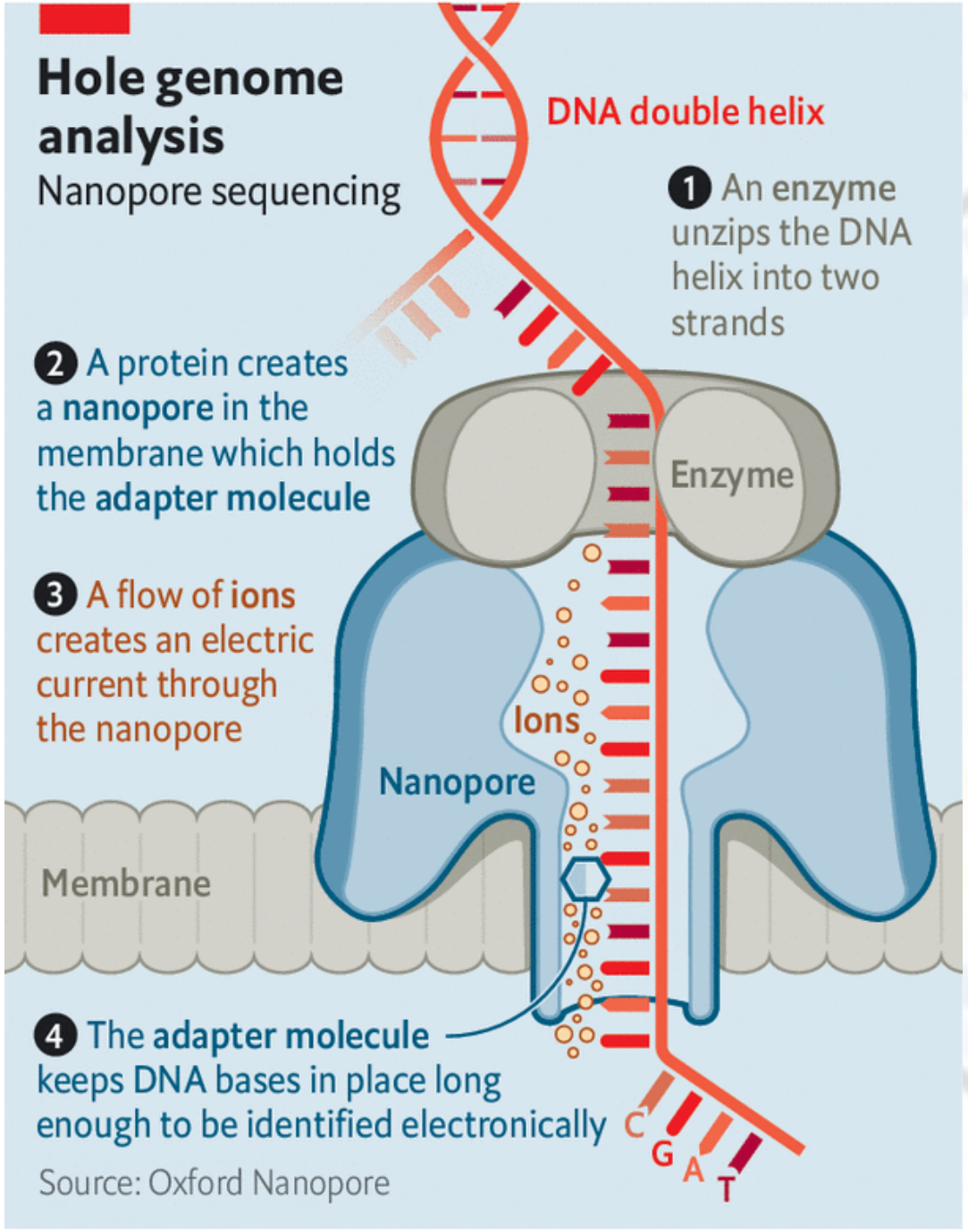
Oxford Nanopore Technologies

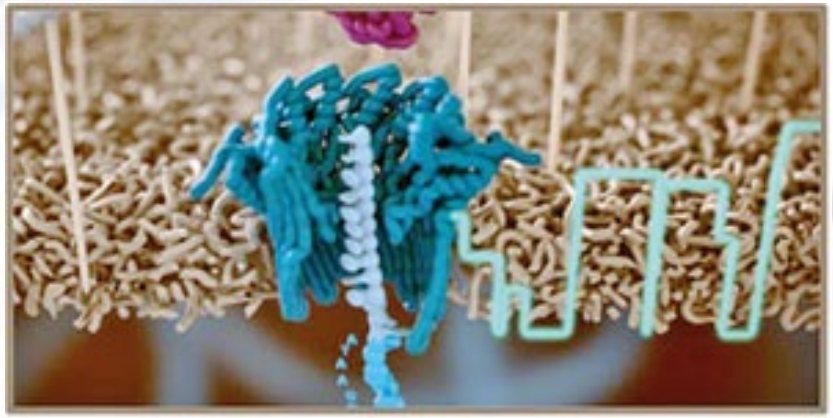
- O adaptador se insere na abertura do nanoporo, e a proteína motora começa a separar as fitas do DNA

- Uma corrente elétrica é aplicada e, em conjunto com a proteína motora, conduz o DNA carregado negativamente através do poro numa velocidade de ~450 bases por segundo



Adaptado de:
Logsdon et al. 2020. **Nature Reviews Genetics.** DOI: 10.1038/s41576-020-0236-x



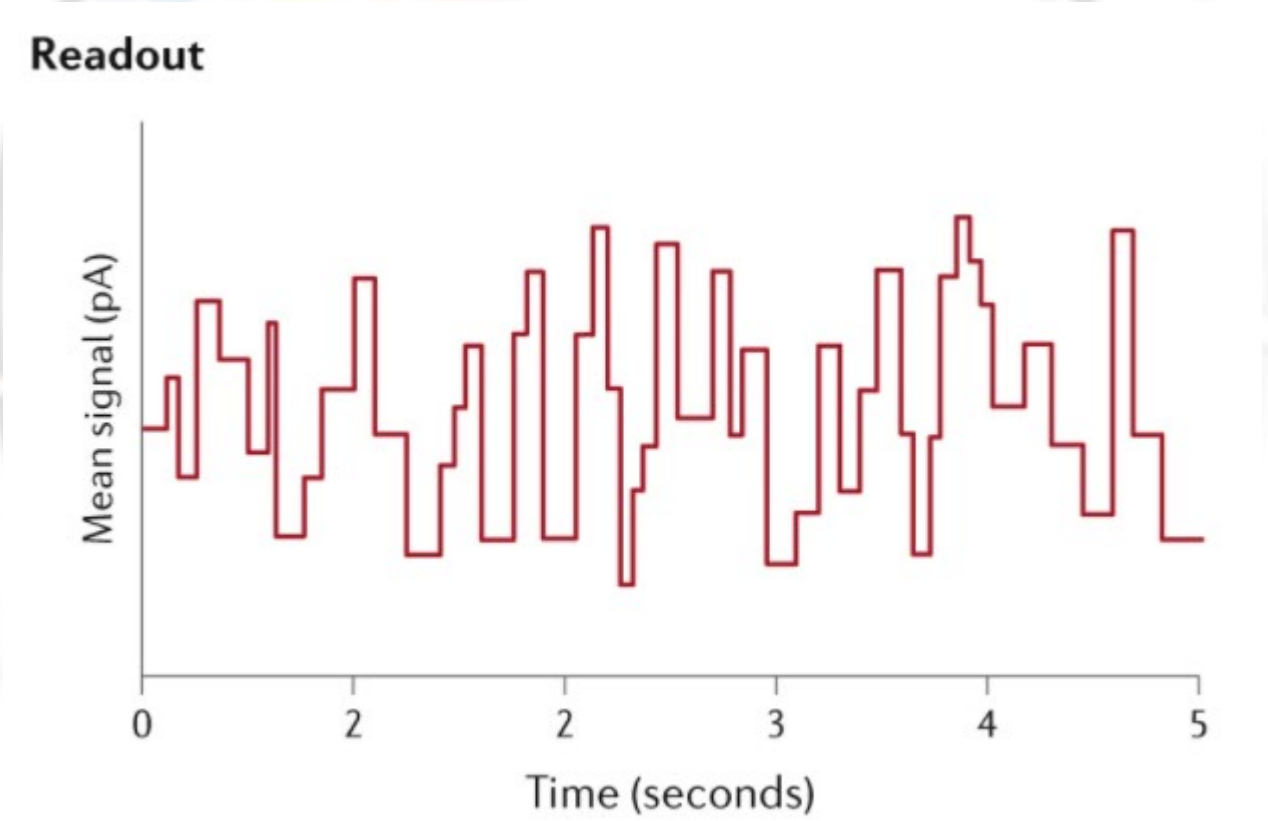


The Economist

Oxford Nanopore Technologies

- À medida que o DNA se move pelo poro, causa perturbações à corrente elétrica, as quais são específicas para cada um dos nucleotídeos

- O perfil de mudanças na corrente elétrica pode ser utilizado para identificar a sequência de bases da molécula



Adaptado de:
Logsdon et al. 2020. **Nature Reviews Genetics.** DOI: 10.1038/s41576-020-0236-x


Oxford
NANOPORE
Technologies

Comparativo entre metodologias

Método	Sanger	454	Illumina	PacBio	Nanopore
Comprimento dos reads	400 - 900 pb	700 bp	100 – 300 pb	10 – 100 kb	Variável (até 1000 kb)
Taxa de erro	0.01 %	0.1 %	0.1%	5 – 15%*	5 – 20%*
Eficiência (bases por corrida)	1.9 - 84 Kb	1 Mb	200 – 600 Gb	10 – 20 Gb	5 – 10 Gb
Tempo de corrida	20 min – 3 horas	24 horas	1 – 3 dias	~ 30 horas	1 minuto até 72 horas
Prós	Alta confiabilidade	Velocidade	Alta confiabilidade e custo baixo	Reads longos, velocidade e alta eficiência	Reads longos, velocidade e alta eficiência
Contras	Baixa eficiência	Baixa eficiência e alto custo	Reads curtos, velocidade	Taxa de erro elevada e alto custo	Taxa de erro elevada e alto custo

Adaptado de:

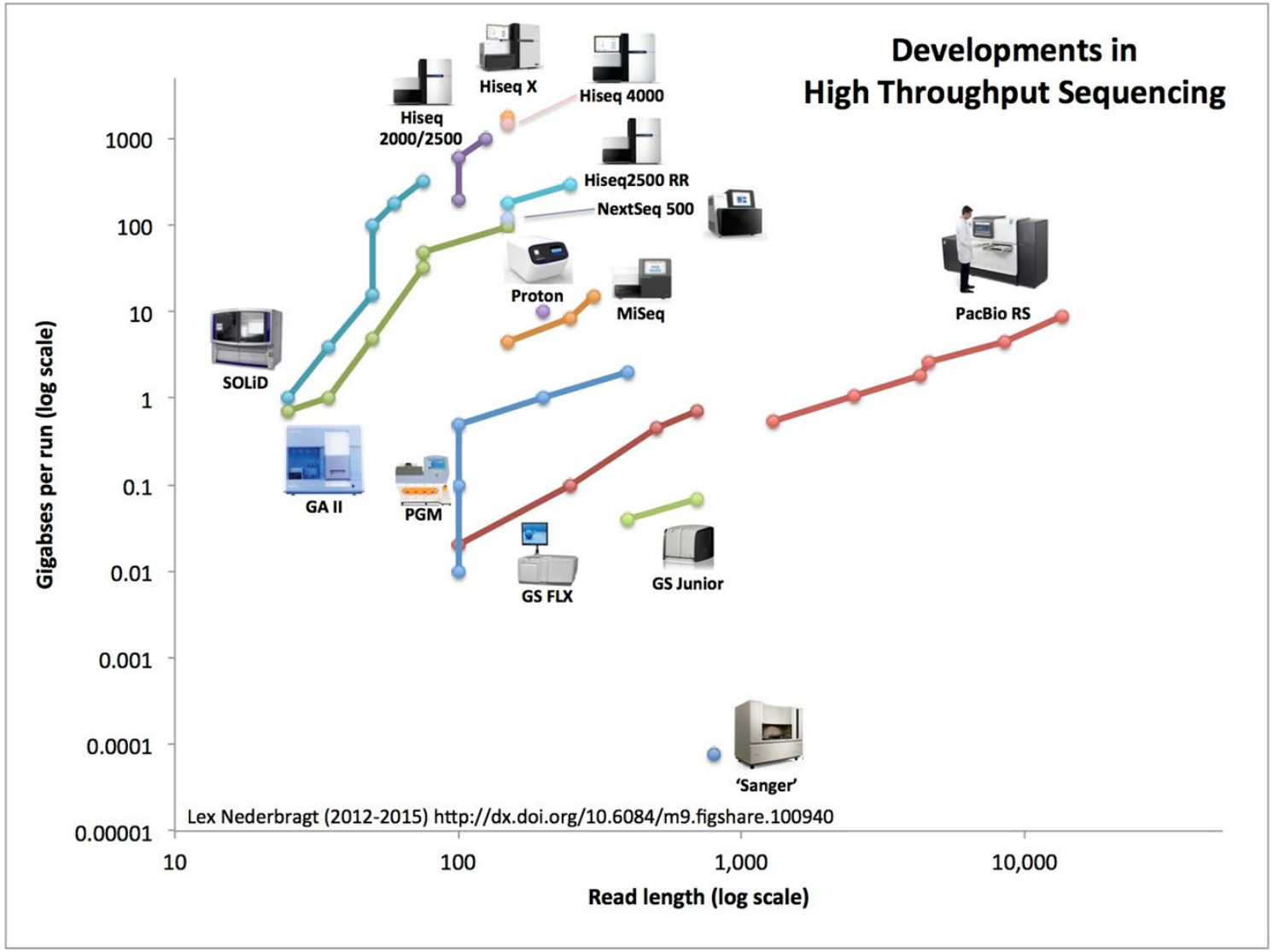
Basal & Boucher et al. 2019 **iScience**. DOI: 10.1016/j.isci.2019.06.035
Liu et al. 2012 **BioMed Research International**. DOI: 10.1155/2012/251364
https://www.pacb.com/products-and-services/sequel-system/

* Em baixa cobertura

Combinar pode ser a melhor estratégia

- PacBio + Illumina
- Muitos reads (grande profundidade de sequenciamento) e PacBio auxilia no processo de organização dos *contigs*

Desenvolvimento da tecnologia next-gen



http://figshare.com/articles/developments_in_NGS/100940

Platform	Instrument	Year	Reads per run	Read length (mode or average)	Bases per run (gigabases)	Source
ABI Sanger	3730xl	2002	96	800	0.0000768	0
454	GS20	2005	200000	100	0.02	
454	GS FLX	2007	400000	250	0.1	
454	GS FLX Titanium	2009	1000000	500	0.45	
454	GS FLX+	2011	1000000	700	0.7	1
454	GS Junior	2010	100000	400	0.04	2
454	GS Junior+	2014	100000	700	0.07	16
IonTorrent	PGM 314 chip	2011	100000	100	0.01	3
IonTorrent	PGM 316	2011	1000000	100	0.1	3
Illumina	HiSeq 2000/2500	2014	4000000000	125	1000	17
Illumina	HiSeq 2500 RR	2012	600000000	150	180	13
Illumina	HiSeq 2500 RR	2014	600000000	250	300	13
Illumina	HiSeq 4000	2015	5000000000	150	1500	19
Illumina	HiSeq X	2014	6000000000	150	1800	18
Illumina	NextSeq 500	2014	400000000	150	120	14
Illumina	MiSeq	2011	30000000	150	4.5	
Illumina	MiSeq	2012	30000000	250	8.5	11
Illumina	MiSeq	2013	30000000	300	15	14
SOLiD	1	2007	40000000	25	1	
SOLiD	2	2008	115000000	35	4	
SOLiD	3	2009	320000000	50	16	