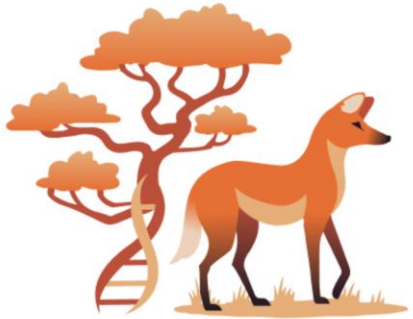
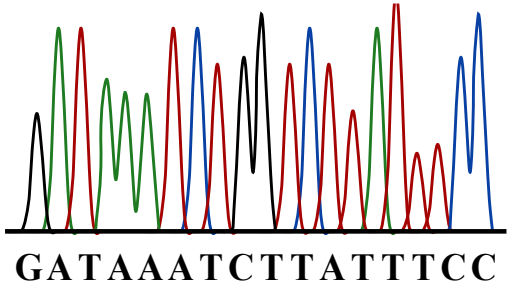
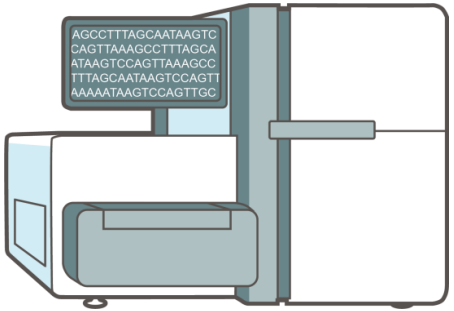
Introdução à Análise de Dados Genômicos

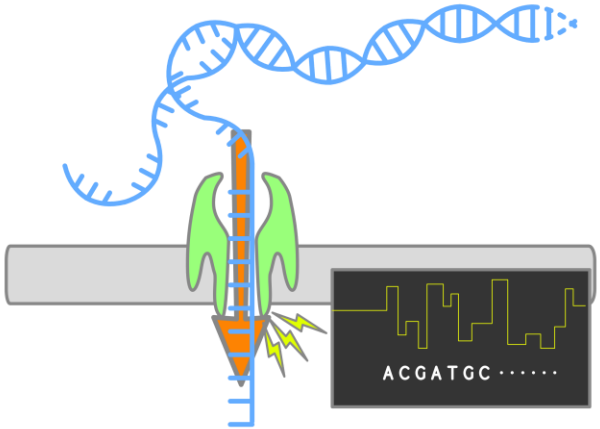


SBBE26

Parte I
O que são “dados genômicos” e como são gerados







E se os nossos craques quisessem sequenciar seus genomas?

Genoma humano tem **6 bi de pares de base**

Anos 2000

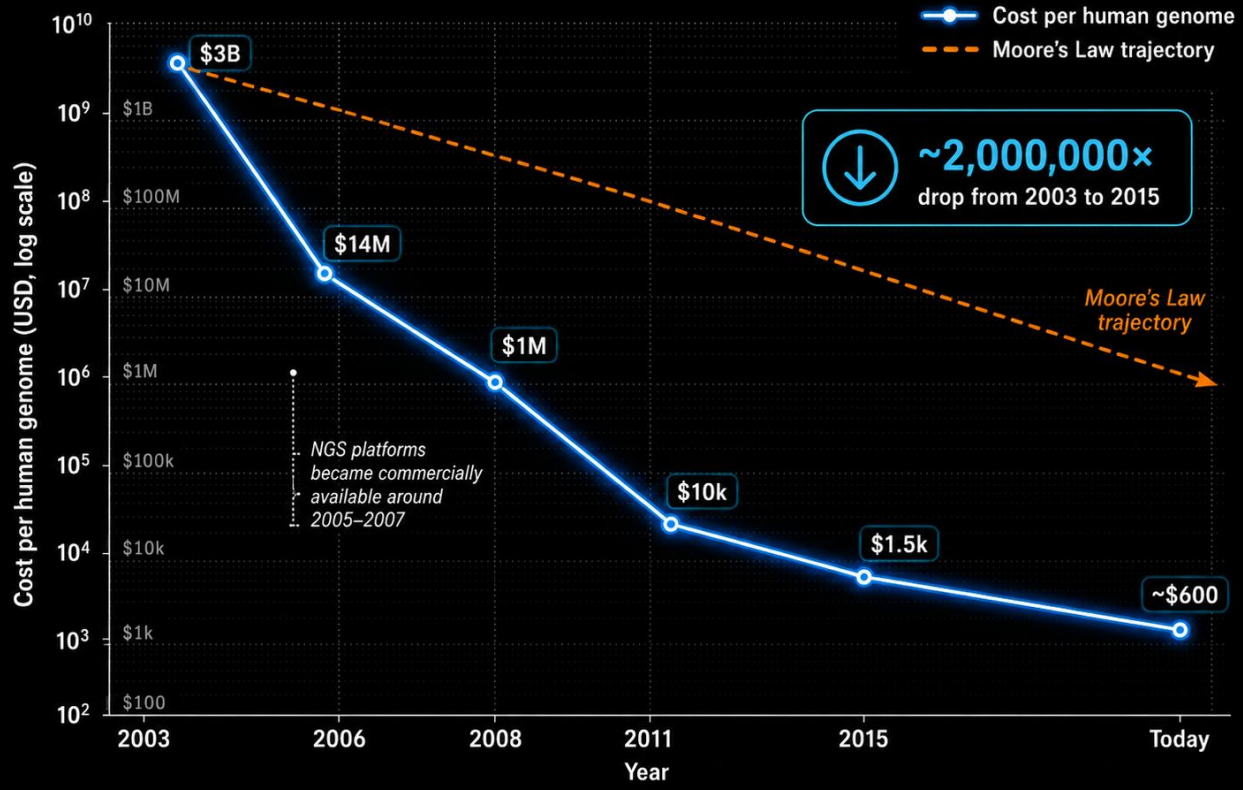
13 anos $ 3 bi

~2015 ao presente

4h $ ~500

The Cost Collapse of Sequencing

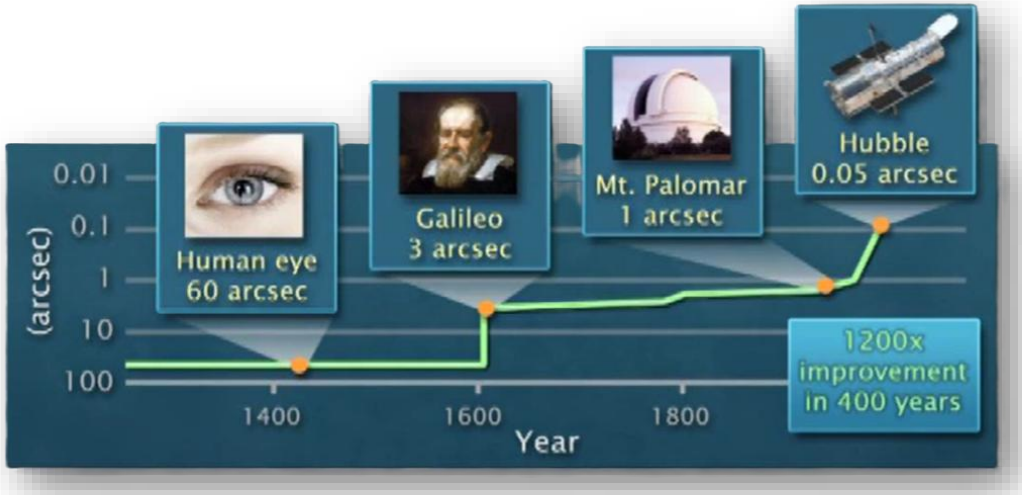
Cost per human genome, 2003–present



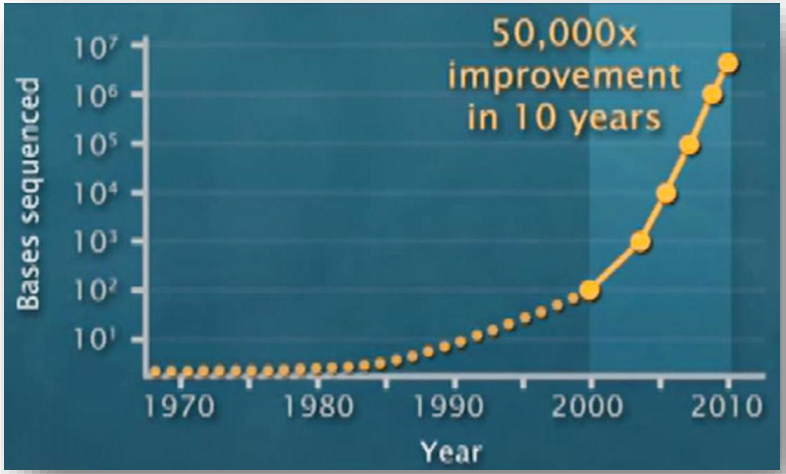
Source: National Human Genome Research Institute (NHGRI) | Data as of the present

Nenhuma* outra tecnologia **avançou tanto em tão pouco tempo**

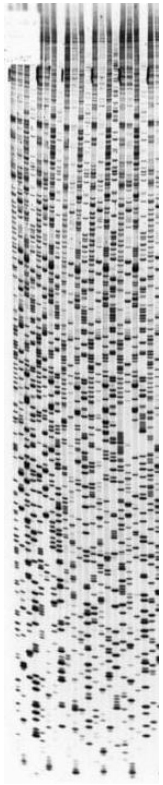
Evolução da resolução angular



Quantidade de bases sequenciadas por 1 dolár

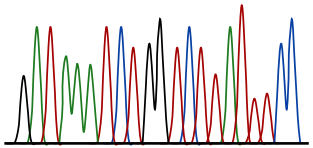


Mas como? A evolução do sequenciamento

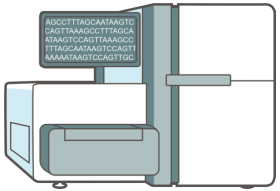


1ª geração de sequenciamento

1977



1986



2005

2ª geração de sequenciamento

1990

Human Genome Project

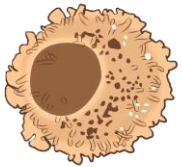
2011

3ª geração de sequenciamento

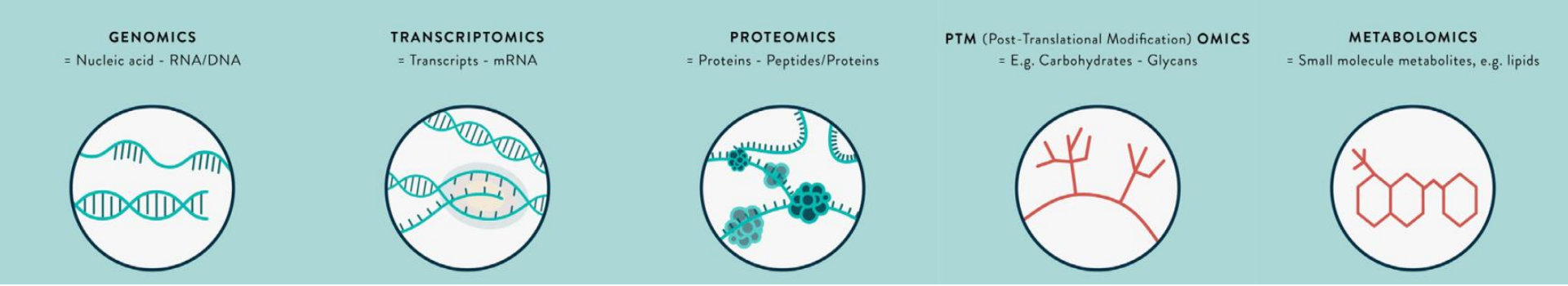
A “era Omics” é agora

NGS

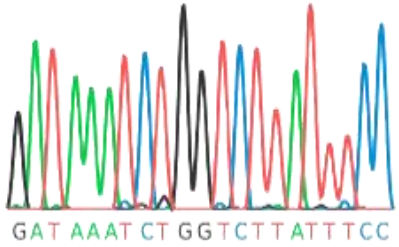
- Genômica
  - WGS
  - Exoma
  - RADseq e variações
- Epigenômica
  - Metilação
  - Cromatina (ATAC-seq)
  - Proteína-DNA (Chip-Seq)
- Transcriptômica
  - RNAseq e variações
- Metagenômica
  - Microbioma
  - Metabarcoding
  - eDNA

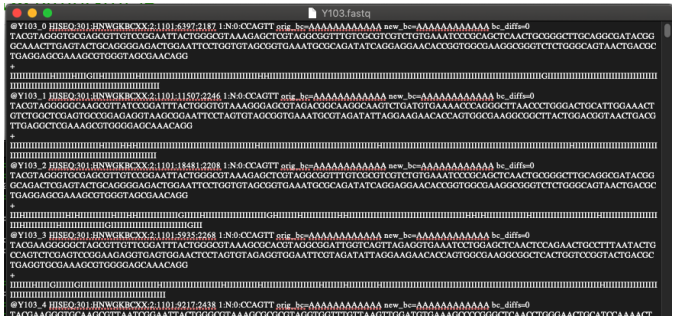


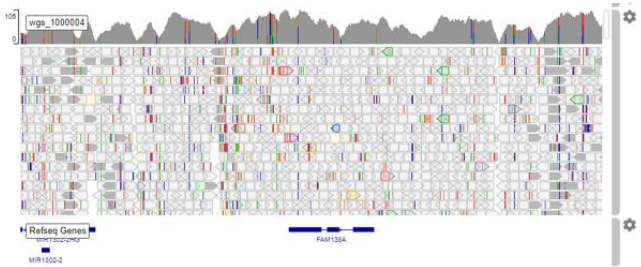
* **Sequenciamento single-cell**



Com a mudança da tecnologia, a **<u>natureza dos dados também</u>** mudou:





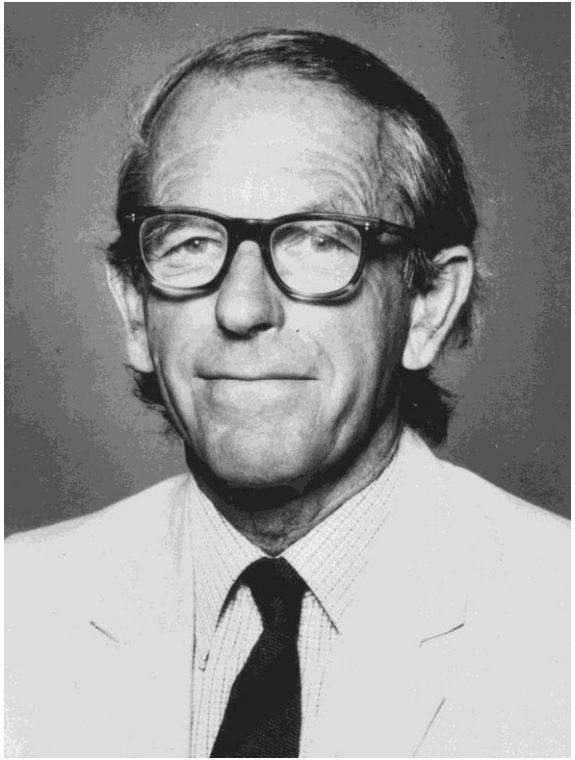


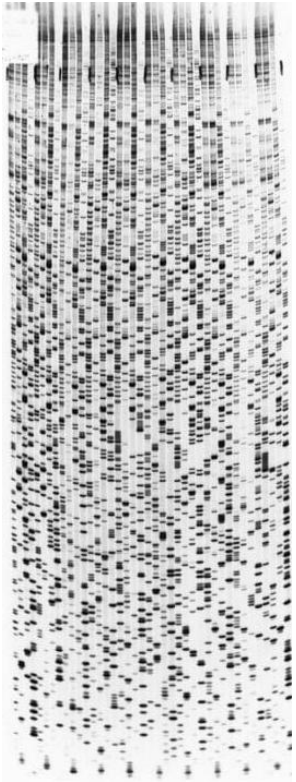
Entender como **<u>os dados gênomicos são gerados é crucial</u>** para o planejamento de projetos e análise de dados.

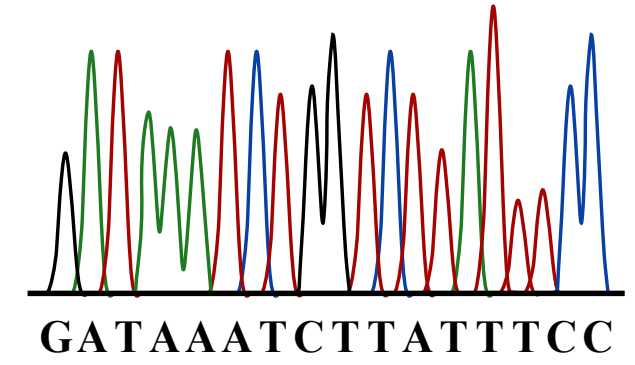
Fluxo da aula

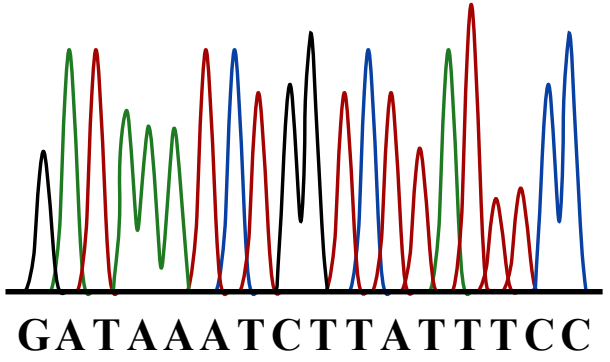
1. O contexto histórico do sequenciamento de DNA;
2. A transição de tecnologias de baixa escala para sequenciamento paralelo massivo;
3. Como diferentes plataformas de sequenciamento funcionam;
4. Prós e contras de cada tecnologia e seu uso para asdiferentes perguntas biológicas.

Primeira geração - Sanger sequencing

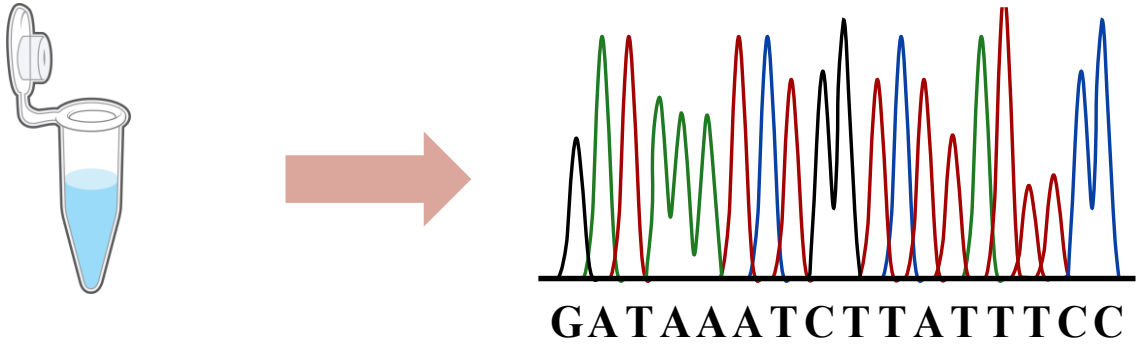


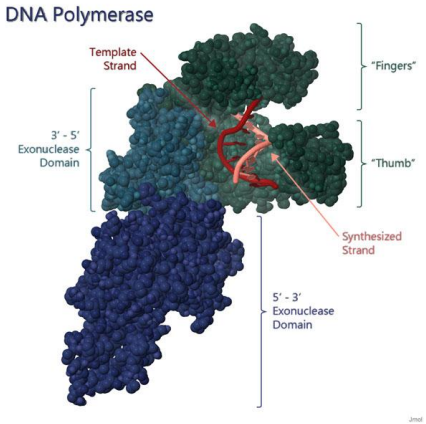


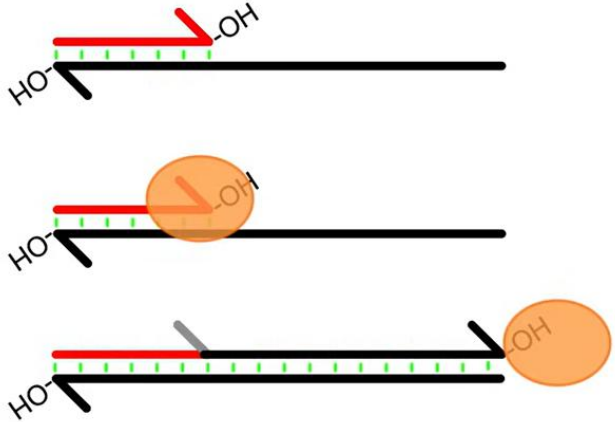




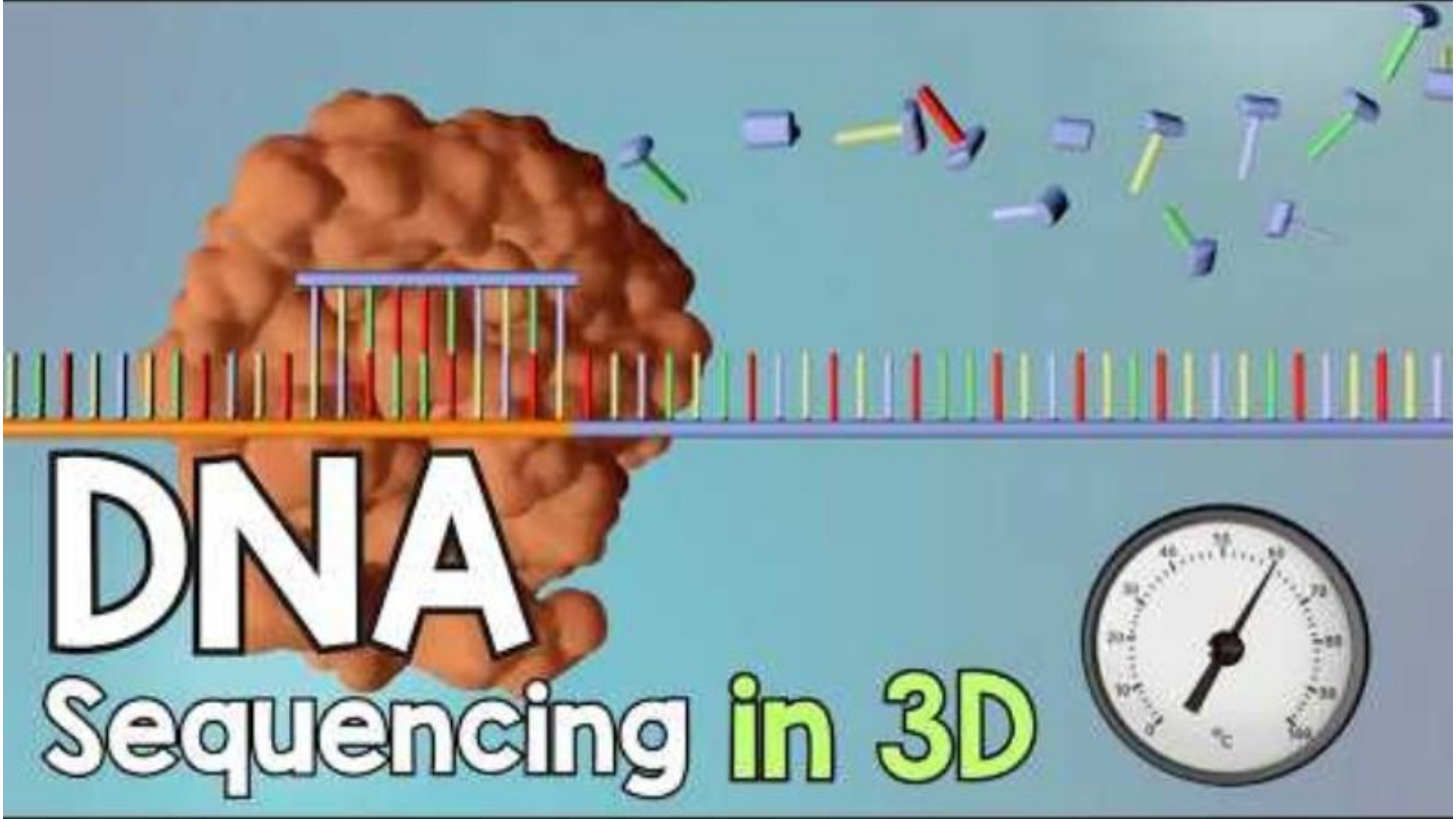
Primeira geração - Sanger sequencing

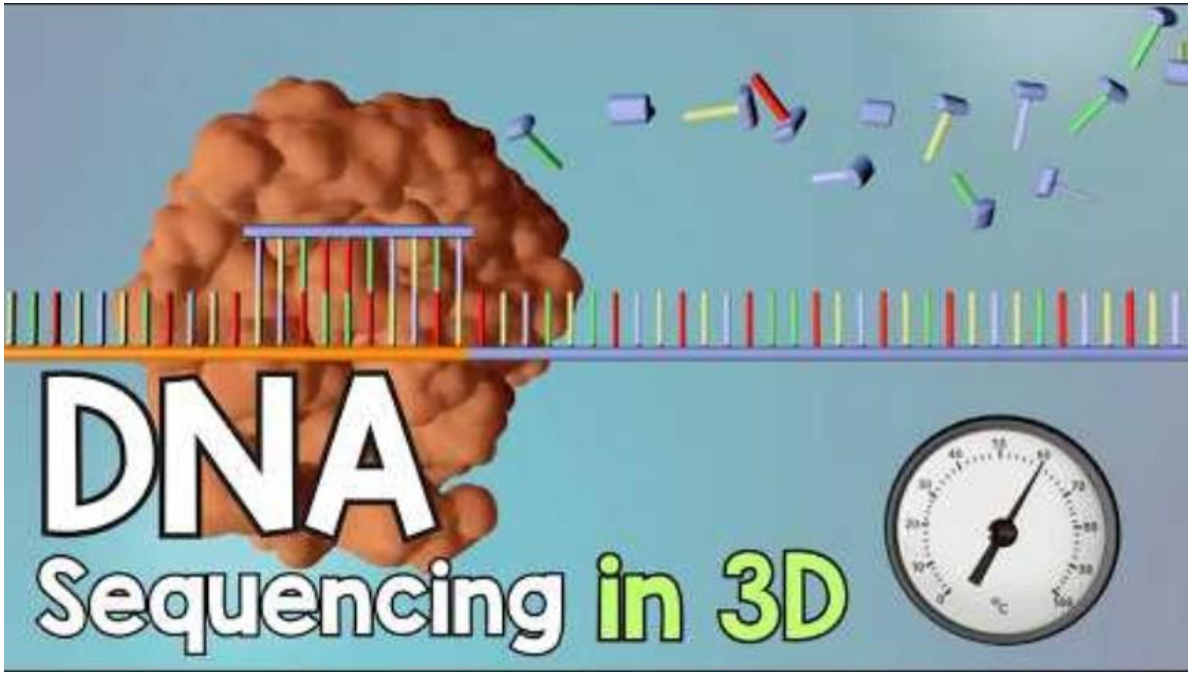




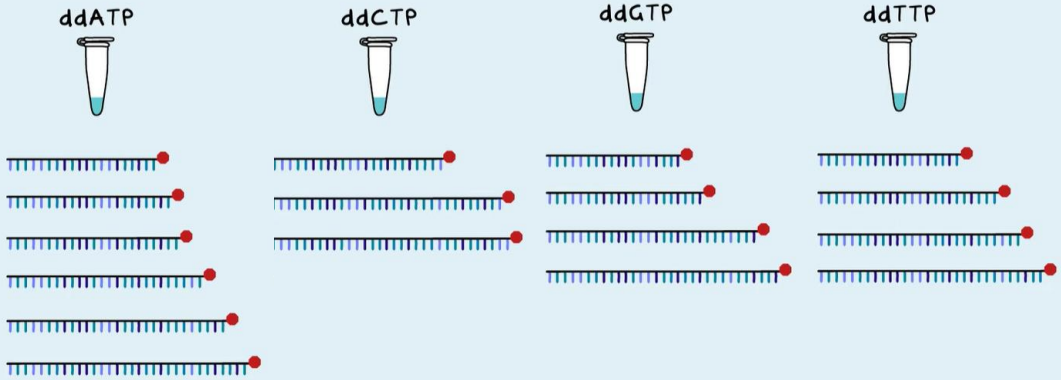


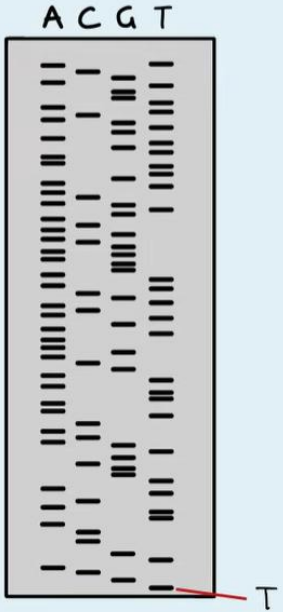
Sanger usa uma propriedade básica das polimerases de DNA
A necessidade de um grupo hidroxil (-OH) na ponta 3' do DNA


DNA
Sequencing in 3D

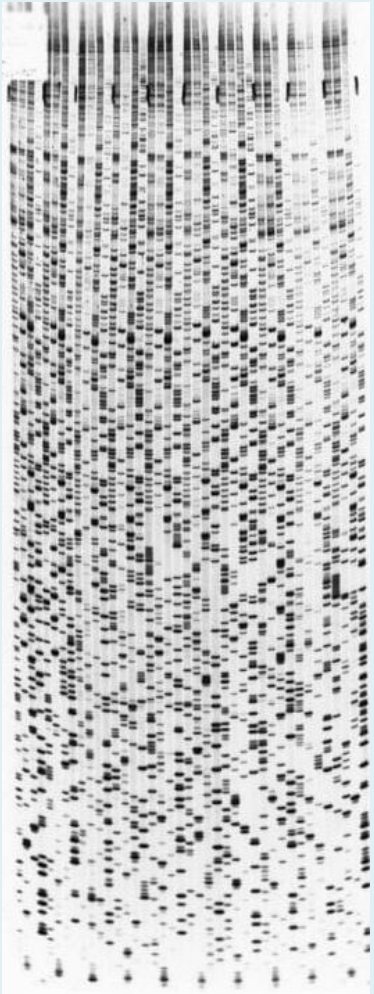


Link para vídeo aqui

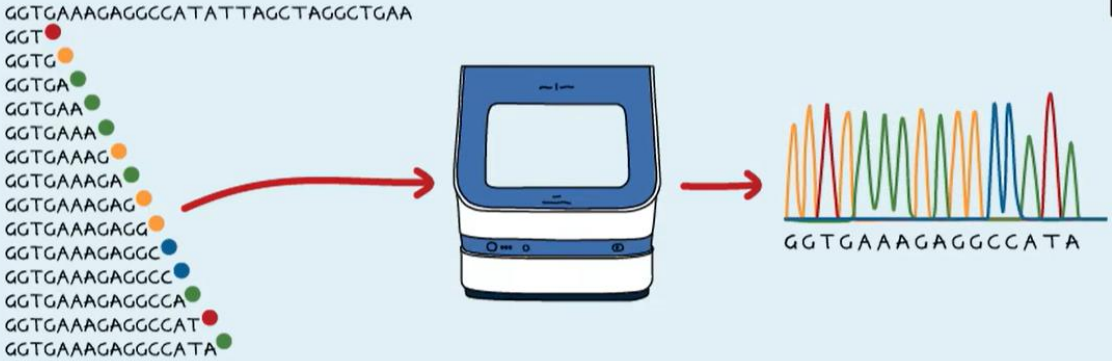




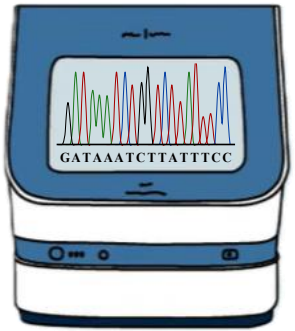
Original



Atual

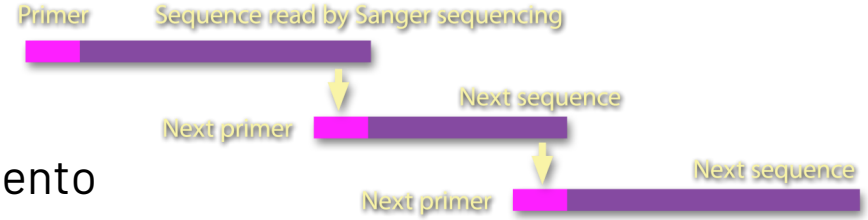


Projeto Genoma humano foi um projeto *baseado em Sanger*

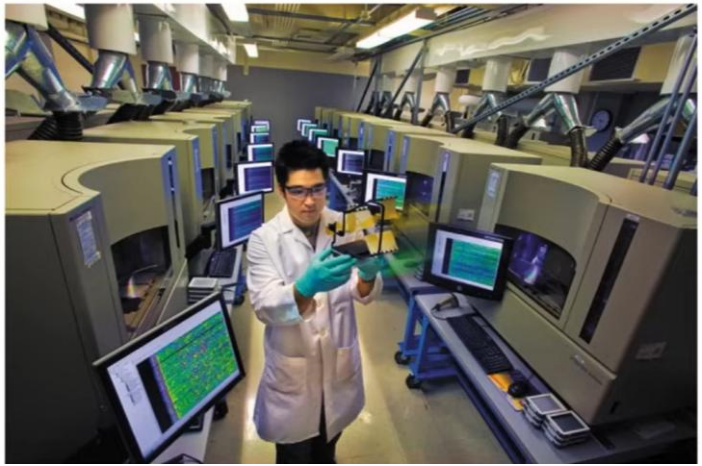


ABI 370A

- Máximo de 384 amostras
- ~700 pb confiáveis por fragmento
- No máx 1gb por dia



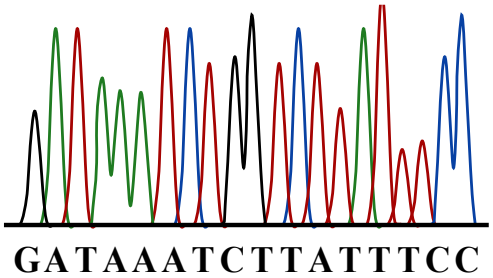
Para sequenciar um genoma humano inteiro (6bi de pares de base a 7x) com Sanger em apenas uma máquina, **levaria aproximadamente 100 anos**.

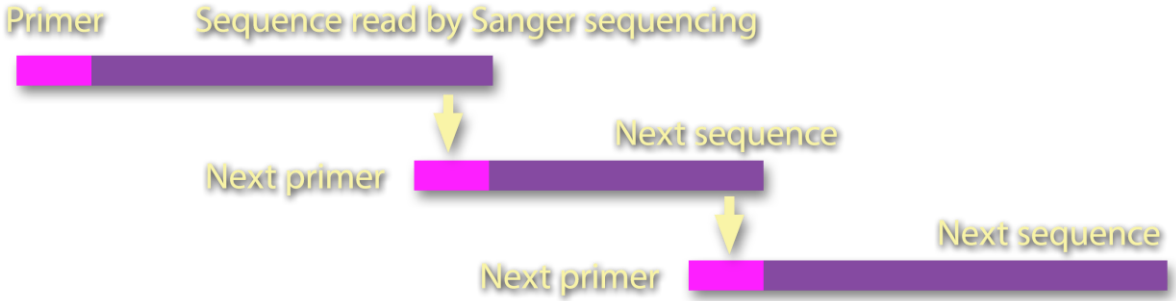


Nature 526, 29-31

Sanger - Eficiente porém custoso

Sanger ***ainda é o gold standard*** para poucos genes, porém:

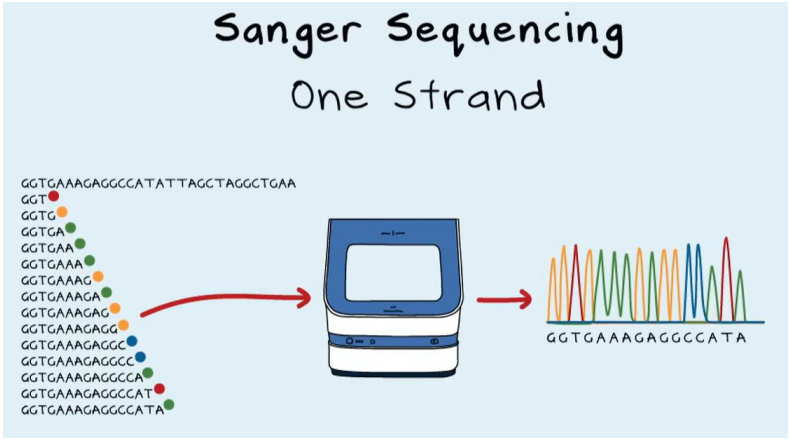


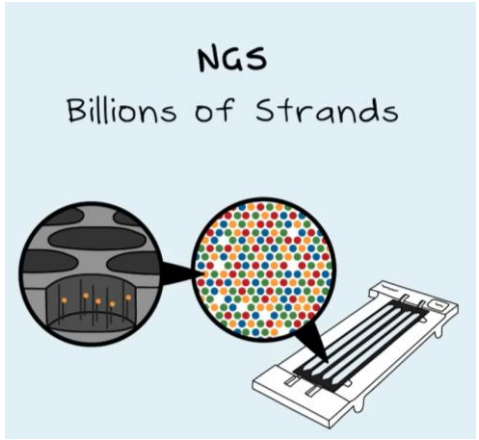


- ***Limite máximo de ~1000 pb***
- Alta precisão, ***mas alto custo por par de base***.
- ***Uma região por amostra***. 300 amostras = 300 PCRs para a mesma região.

lih, mas meu projeto involve muitos genes e amostras...

Next-Generation Sequencing é a grande virada de chave

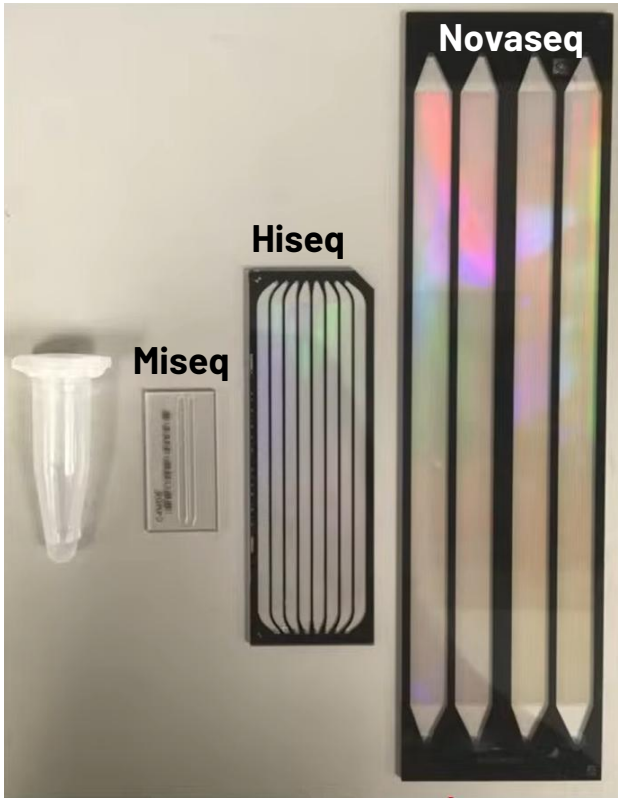




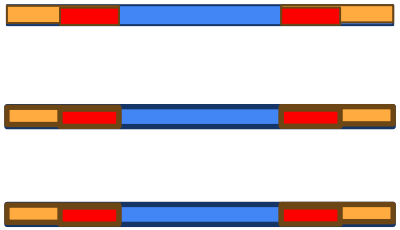
- A grande sacada do NGS é a **paralelização em massa**.
- As tecnologias de segunda geração **sequenciam milhões ou bilhões de fragmentos simultaneamente.**

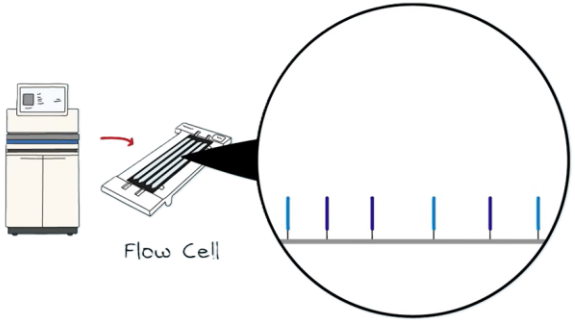
Illumina sequencers

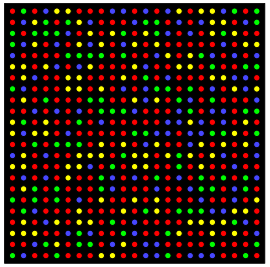
	MiSeq	HiSeq	NovaSeq	Sanger
Reads (millions)	30	3,000	13,000	0.0004
Gigabases/day	7	500	4000	0.001



Passos do sequenciamento com Illumina







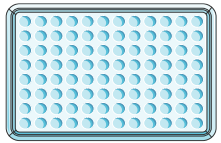
1. Preparação da biblioteca

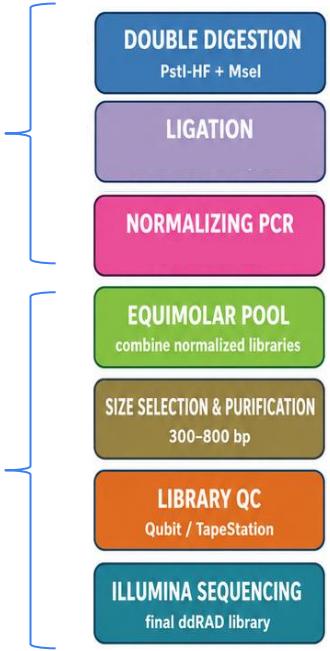
2. Amplificação local

3. Sequenciamento via síntese

1. Preparação da biblioteca – Passos gerais

Passos e Objetivos





1. Digestão enzimática ou sonificação para fragmentação do DNA.

2. Ligação dos adaptadores Illumina para sequenciamento.

3. PCR para aumentar as réplicas de fragmentos e ligação dos index com barcode.

4. Quantificação, equimolaridade e combinação **em um tubo (multiplex).**

5. Seleção de tamanho de fragmentos e limpeza.

6. Qualidade, concentração e distribuição de leituras.

7. Envio para o sequenciamento **e pode rezar!**

Passos *especialmente* relevantes para a bioinformática

1. Fragmentação do DNA

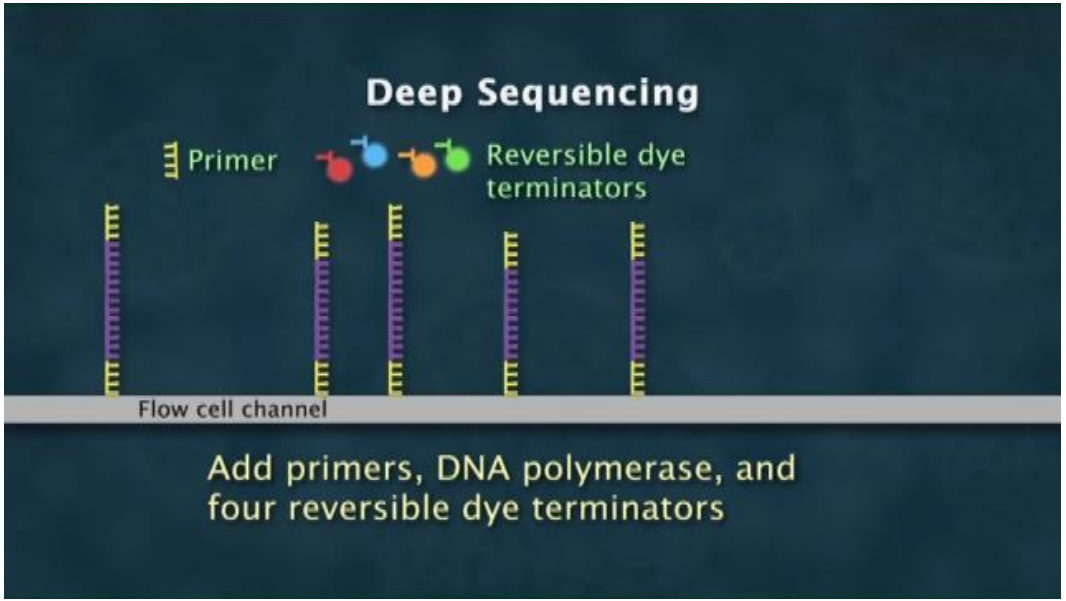
2. Ligação de adaptadores

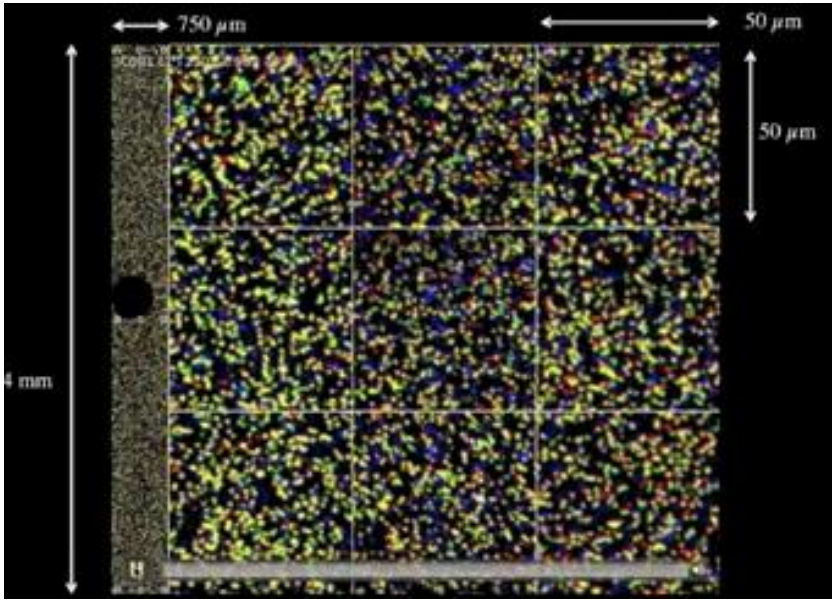
3. PCR para indexação dos barcodes e aumento do número de leituras

2. Amplificação local (de onde suas leituras vêm)

Link para vídeo aqui

3. Sequenciamento via síntese (como as reads são sequenciadas)

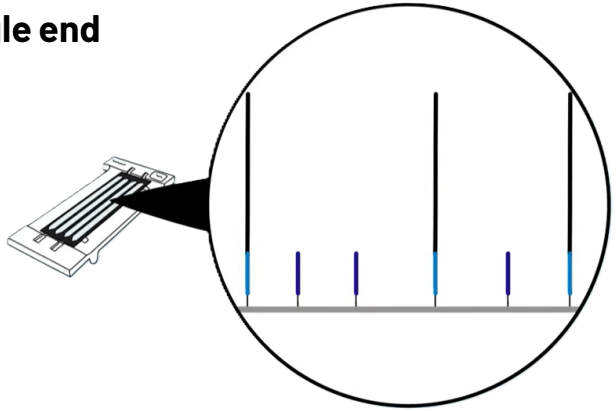




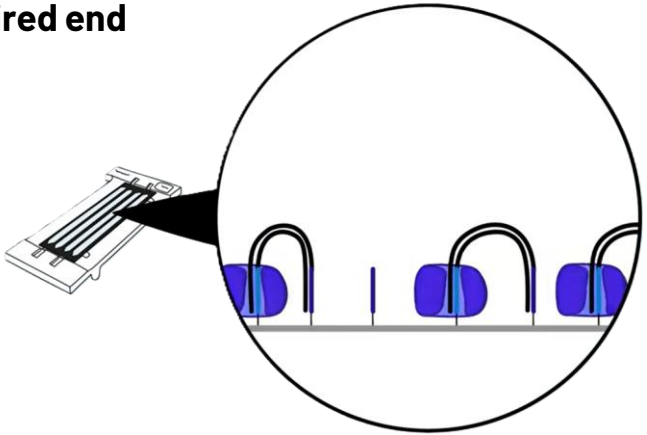
Link aqui

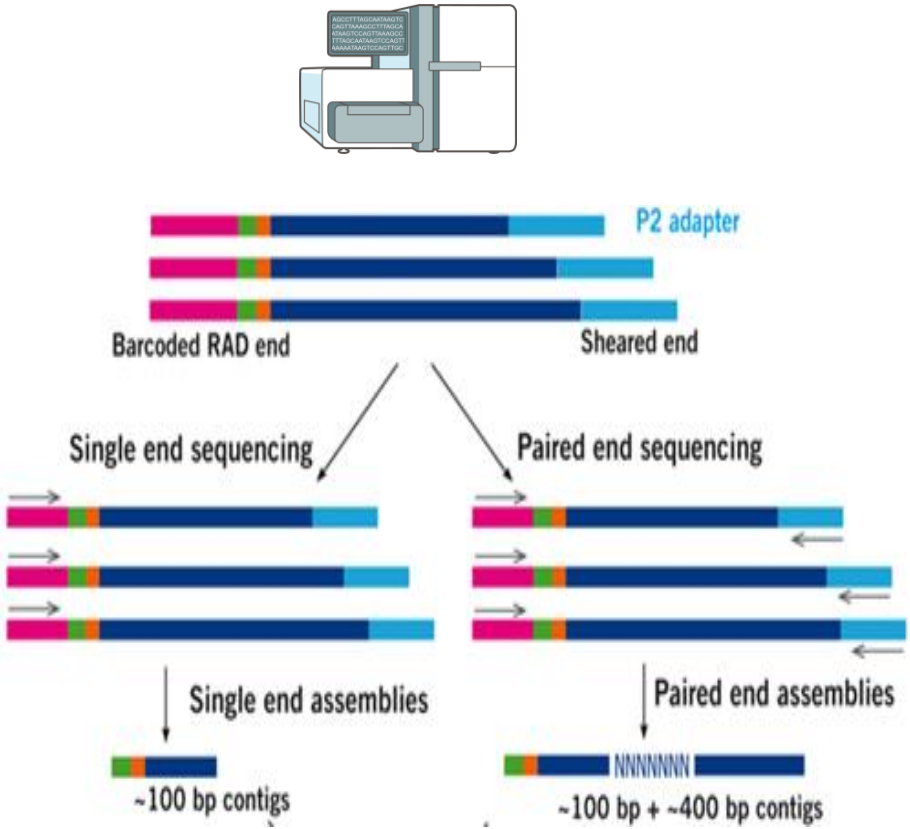
Single end vs Paired end (single está em extinção)

Single end



Paired end



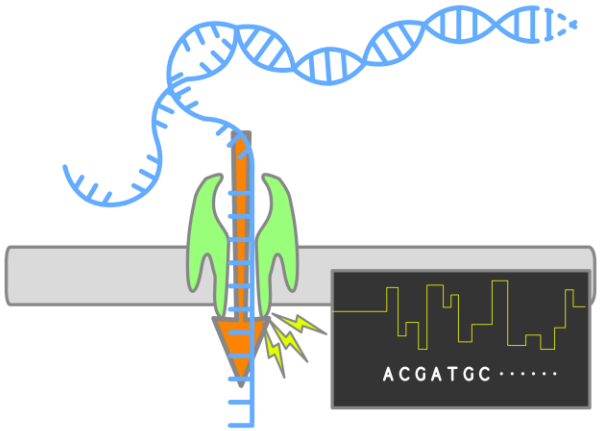


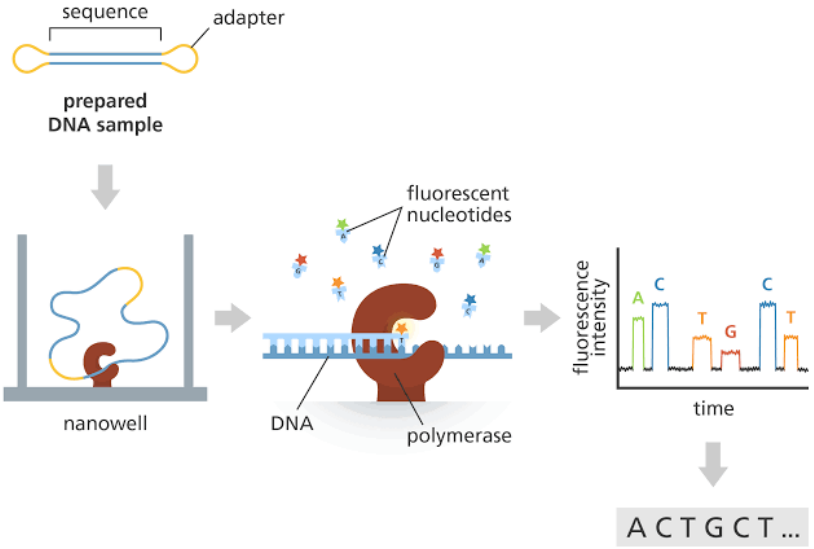
Floragenex website

Sequenciamento de terceira geração – Leitura longa

Oxford NANOPORE Technologies

PacBio

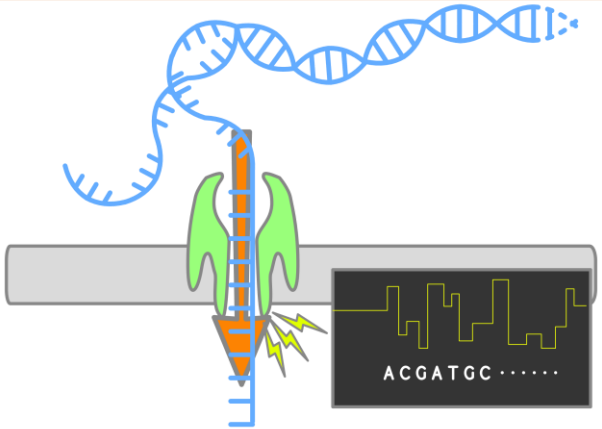


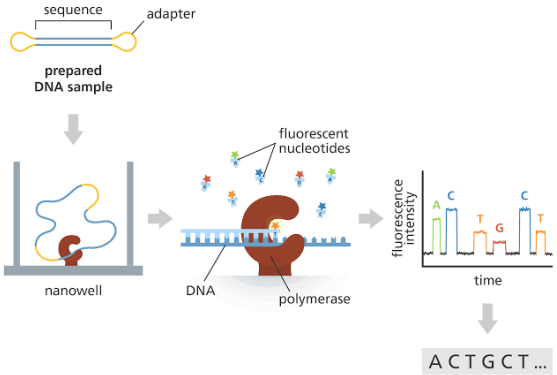


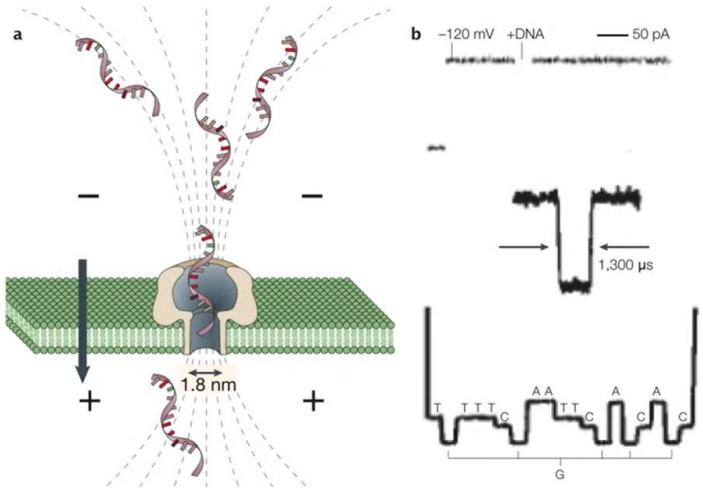
Sequenciamento de terceira geração – Leitura longa

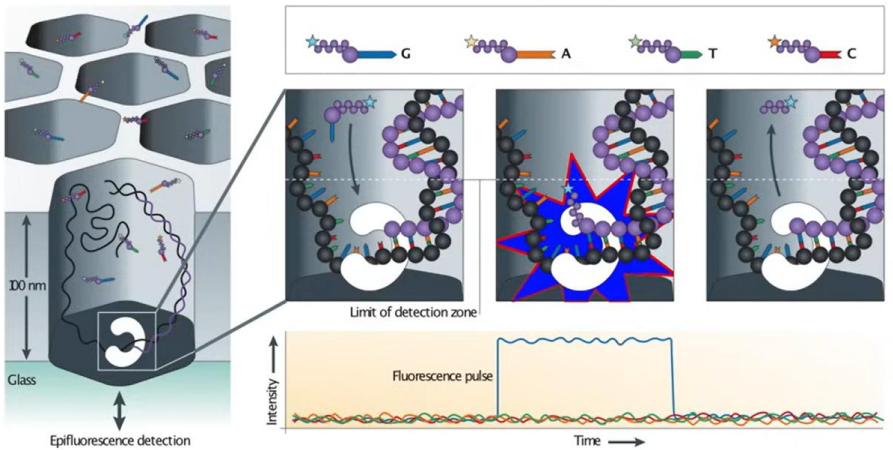
As tecnologias de **terceira geração** vieram para preencher algumas limitações das tecnologias de leitura curta. Com elas:

1. Sequenciar **regiões maiores ou difíceis de alinhar**.
2. Identificar variações estruturais com mais precisão.
3. Dividir os cromossomos por haplótipo.
4. Porém, ainda são custosos e taxa erro alta (pior no Nanopore, mas melhorando!)





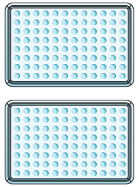




Lavan et al. (2002) Nat Rev Drug Disc, 77-84

	Sanger	Illumina	PacBio	Oxford Nanopore
Geração	**1ª**	**2ª**	**3ª**	**3ª**
Princípio	Terminação de cadeia	Sequenciamento por síntese	Molécula única em tempo real	Corrente elétrica em nanoporo
Tipo de leitura	Média/longa	Curta	Longa	Longa/ultralonga
Rendimento por sequenciamento	Baixo	Muito alto	Alto	Alto, variável
Precisão	Muito alta	Muito alta	Alta/muito alta com consenso	Média mas melhorando
Custo por base	Alto	Baixo	Médio/alto	Médio, variável
Principal vantagem	Gold-standard para regiões específicas	Muitos dados relativamente baratos e precisos	Genomas, SVs, haplótipos	Portabilidade, tempo real, reads ultralongos
Principal limitação	Baixo rendimento	Reads curtos	DNA de alta qualidade/custo	Perfil de erro/DNA de alta qualidade/custo

Algumas limitações práticas no Brasil

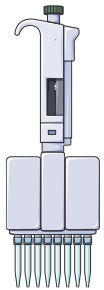


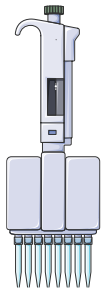
Custo para 192 amostras

Orçamento aproximado ddRAD em 2024

Preparação da library, controle de qualidade **($4000)**	Trabalho do técnico na library **($3000)**	Custo do sequencimanento **($2000)**

Esse custo pode ser **SUBSTANCIALMENTE** menor se for feito no laboratório, porém...





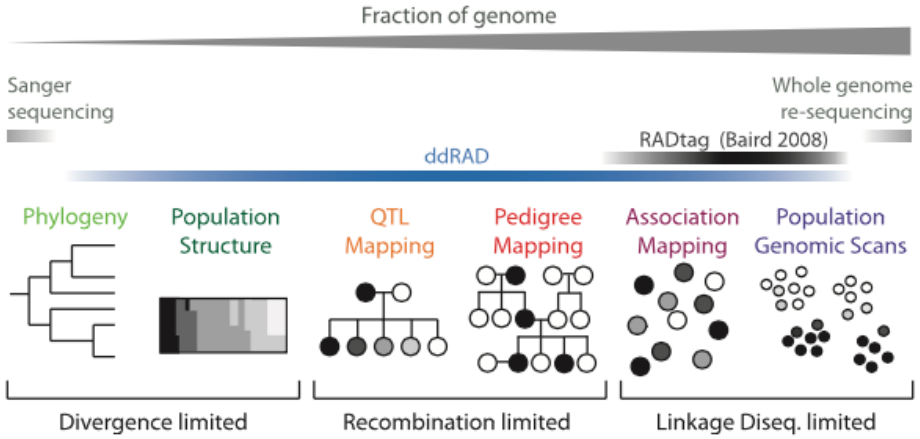
Além de outras limitações, claro:

- Poucos sequenciadores disponíveis
- Reagentes
- Dificuldade de licença e envios

A escolha da tecnologia *já é uma decisão* científica

A escolha do sequenciamento já afeta:

- As regiões que serão bem representadas;
- Os tipos de variantes detectáveis;
- O custo;
- O número de amostras;
- A complexidade da análise;
- **Mais importante:** as perguntas a serem respondidas.



Peterson et al. 2012

Alguns parâmetros *importantes* a se considerar (1)

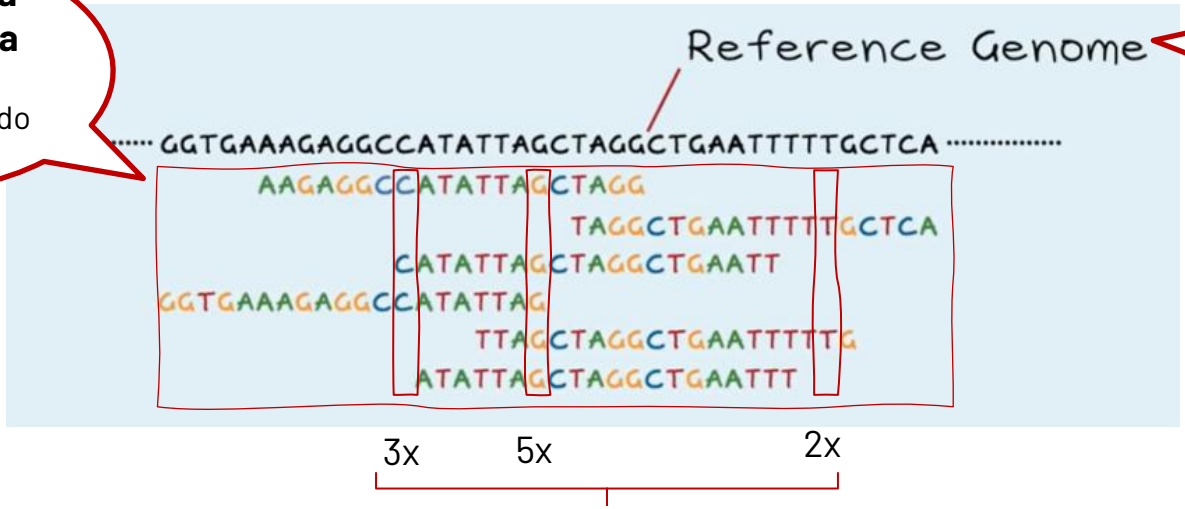
A ampla variação do tamanho/complexidade dos genomas vai interferir na quantidade de dados para a sua pergunta

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

Fonte

Alguns parâmetros *importantes* a se considerar (2)

Abrangência de cobertura
Quanto eu consigo cobrir do genoma?



Genoma de referência
Tenho um bom genoma de referência?

Cobertura média
Quantas reads em média suportam as minhas conclusões?

Missing data
Como a quantidade de dados faltantes afetam as minhas conclusões?

Relação Objetivo x Custo x Qualidade x Quantidade

Objetivo

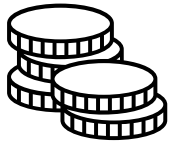
Qual o melhor modo de responder melhor à minha pergunta?

Qualidade

Qual a qualidade dos dados necessária?

Quantidade

De quantas amostras preciso para responder à minha pergunta?



Custo

Quanto custaria essa brincadeira? Como usar melhor os recursos disponíveis?

Show! Tenho meus dados NGS, e agora?

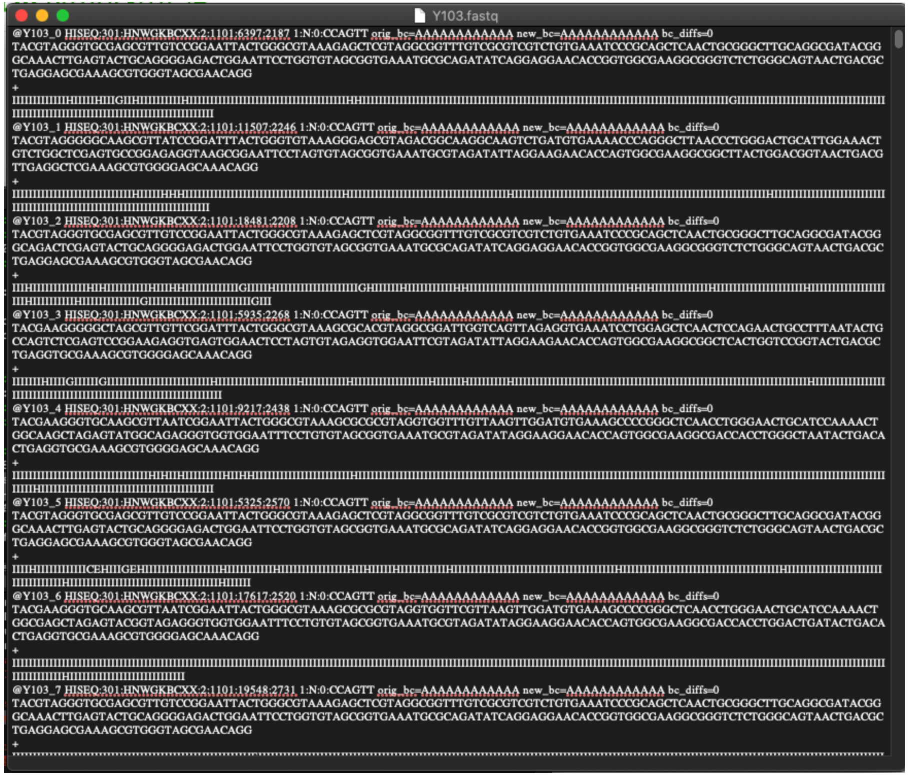
A esse ponto *temos* **alguns problemas**:

1. Nossos adaptadores estão ligados aos reads.

2. As leituras estão todas juntas. Qual read é de qual amostra?

3. Posso confiar no meu sequenciamento? Como acessar a qualidade?

4. Como fazer ***sentido biológico desses gigabytes de letrinhas?***



Parte II – Bioinformática:
A tradução da genômica para a biologia (George)