Introduction of Next-Generation Sequencing

09/06/2024

Dr. Luu Phuc Loi (luu.p.loi@googlemail.com)

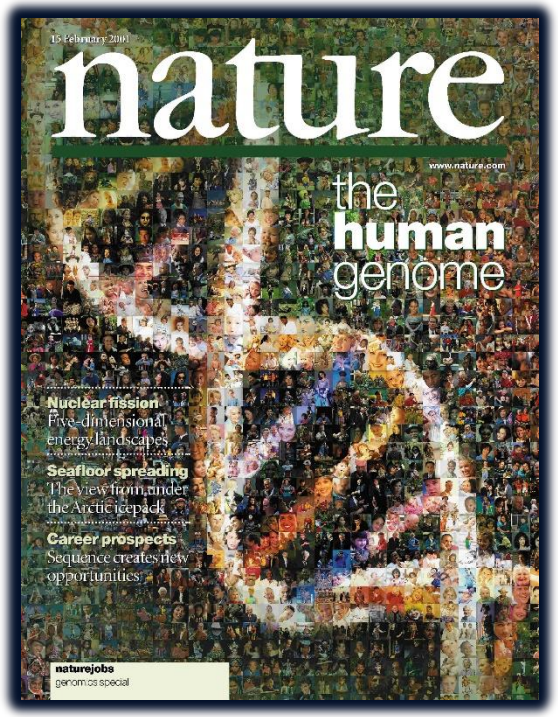
Content

1. Human Genome Project (HGP)
2. Next-Generation Sequencing (NGS)
3. Short read vs Long read
4. NGS Applications

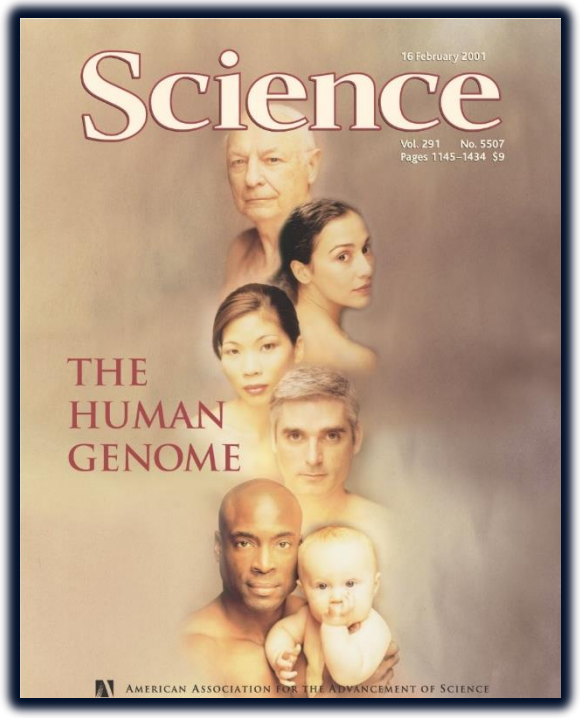
Human Genome Project - HGP (Oct 1990 - April 2003)

1. In 2003, the Human Genome Project produced a genome sequence that accounted for over 90% of the human genome (**~3 GB**).
2. It was as close to complete as the technologies for sequencing DNA allowed at the time.
3. Cost ~3 billion US$

=> Facilitating advancements in **next-generation sequencing (NGS)** technologies



HGP Paper



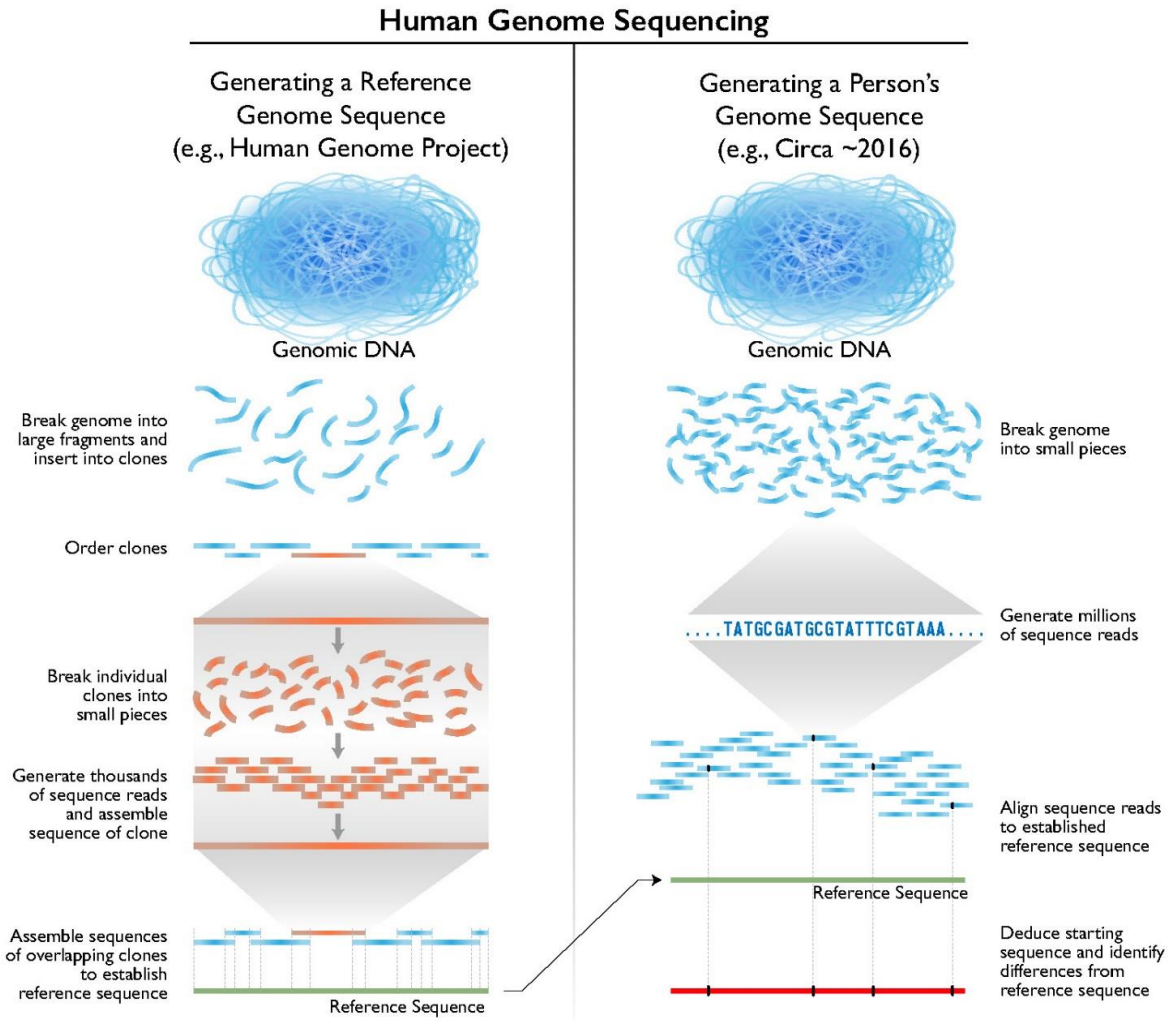
Venter/Celera Paper

https://www.genome.gov/human-genome-project

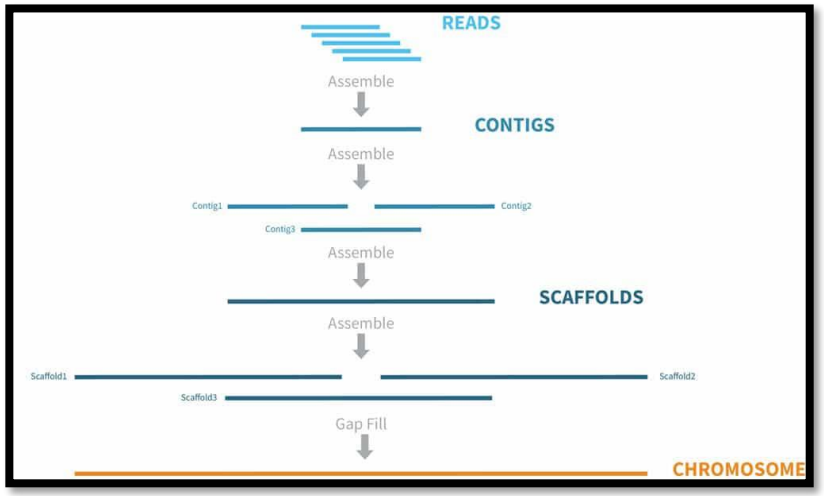
```
48541 agcccttcaa agaaatgttc tcagcaggca tggagcccag gacttgctcc ctttgggtag
48601 agagccgggt tgaaggtgac tgaagtgaaa tgggacagta gaggcggggg gggtggtgag
48661 ttcctggagg tggggggtgt gggaacctgc tgtgtactga gatgcacccc tgccagttct
48721 gcctgaagat ttgaggcggg gggcaggggg gcggagtgaa gtcattttac tggtaagtaa
48781 ttttaaacct tttaatatta aagcaaacgt ggatatgtaa tgaatgaaat tcattctgga
48841 atgaaaaatt cacgtgatgt tgaaaaataa cacggggctt cagagaggac tttctggctg
48901 gcagcagact ccagattccc agggcccctg caccctcctc tgcccacagg gcaccttaat
48961 ggagaaggtg tgggaggaga gccaggccgg agtcagagca cactggtgac tccacatttg
49021 cagcgtgccc tgcctctctc ctgaggcttg gcaacgtgca atatgctaag caaactcccc
49081 ctgtccccgt ccagtttctg aggacaagag ccaccacctg tagcaaataa agacccagca
49141 acccctttgac tcatctttgt gagtctctgg aatcagaggg tagccacatc gctgagaggt
49201 ggagtgaagc actcgggtga aaaggtacaa ggaagtcagg gacaggagtg tggggacatc
49261 acctagacaa tgacagagaa gaggggcaca gccgagtgag gggagagggg ccggcagtcc
49321 tacatcccct ggcctgaagc acgctccagg gcagaaggaa aaacactgtc tttggggtcc
49381 aagagacctg agttcaaatt ctggctccac cactgaccac ctgtgtaacc ttgaactgct
49441 gctgcctgaa cctcaggttc cccttctaaa aatagaggag aaaaggatgc atttctcctt
49501 gcccctgtga gaacgaaatg gtgcaagcac caaggagcct cagcaaaggt cgggcctgcc
49561 cccgcctggc caaacctttc ctcttcagga ggccacggca accgtagttt gacagaagag
49621 cagcaccttg atttaatgct tcccagcatg tgtccttgag caagtcacct aacctctctg
49681 ggctgcttcc tcattgggaa aatatggctg ccagtaaaac ctgccctgtc cacctcctgg
49741 ggcacttggc aaacagcaaa agagtccaaa tgtgcaggct gggccaggcg cagtggctca
49801 tgcctgtaat cccagcaatt taggaagcca aggtgggcgg atcacctgag gtcaggagtt
49861 tgagaccagc ctggccaaca tggtgaaacc ttgtctctac aaaaatacaa aaattagccg
49921 ggcatgatgg cgggtgcctg taatcccagt tactcgggag gctgaggcaa gagaatcgct
49981 tgaacccgga aggggaaggt tgcagtgagc caagattgtg ccactgcact ccagcctggg
50041 caacagagcg agactctgtc tcaaaaaaaa aaaaaaaaaa aaacaatgca gagctggctg
50101 tgtaaaaaac ctgttccact gcagggccca gtgtccacca ggctggggtg caggcctatg
50161 gggtgggggc ccagcatcag cctctcagca gccctgggag gcggggcgca tcccgtgccc
50221 ctcgtggtct ggatgtgttc tagcccaagt cctaggttac acctgccgtc gcctggcctc
50281 tcaggagagg cccagggtga ggaggagcat ggtaaaggtg aagctgattg ggaagtcagc
50341 tgttgggaaa gcaactcctt gcacattgga ggaaccgaga aagactgacc ccgaggacag
50401 cagccagcat ggcccttcct gggagcccat gttgggggat tcctgctgca gccaaggctc
50461 agcccttgtg gtcgcaggtg ctggtcctgg cctcttcccc tcccatgcag gagcacagga
50521 gagatggctt ctgaggacct gttgcagctg tggccctggg aatagatttg ccagggagct
50581 ttaaagcagc tgagtgtgtc atccagctaa gcctggggaa ggagcttggc tcaggtcctg
50641 acaggtgtga cagggatggg gactgggaag taagagatga aaccctggct ggaggctgtg
50701 agcttccaca gccagcgctg gacagggagg gtccagatat acccactagt gccctcacca
```

https://www.ncbi.nlm.nih.gov/nuccore/806904736

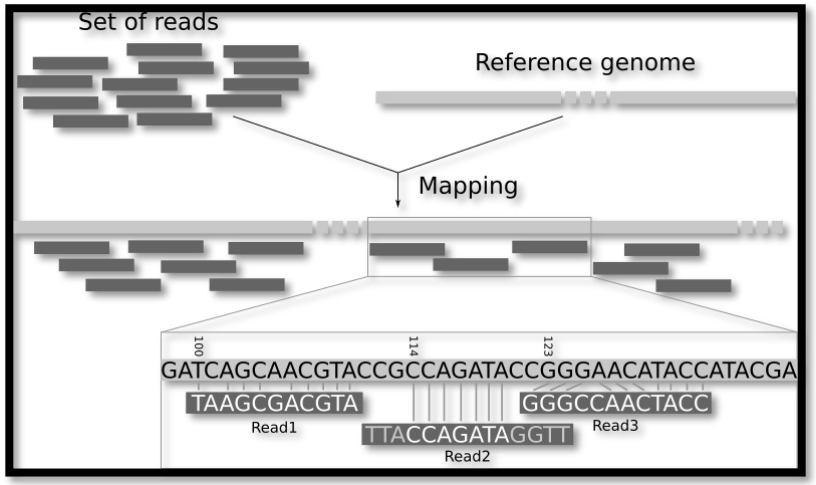
Next-Generation Sequencing (NGS): Reference Genome



De novo assembly

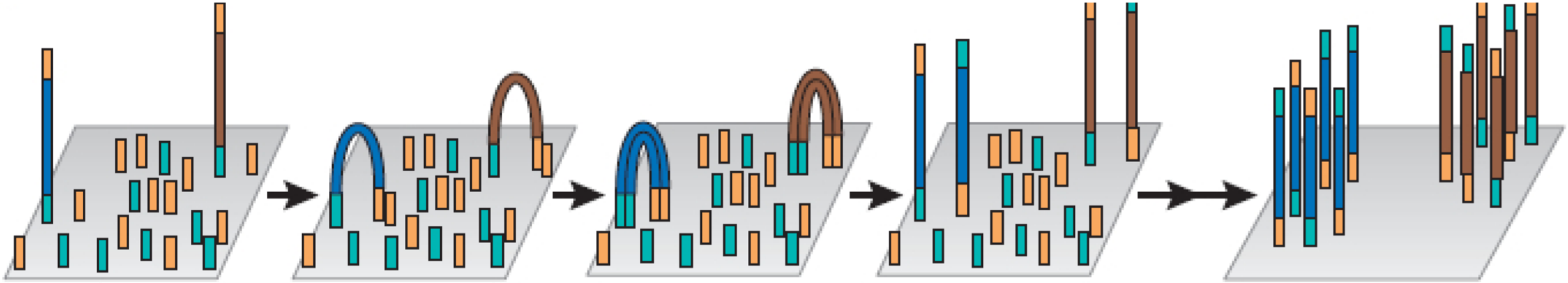


Mapping to reference



https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

NGS: 1) Parallel Sequencing



10.3390/medsci2020098

NGS: 2) Reference Genome

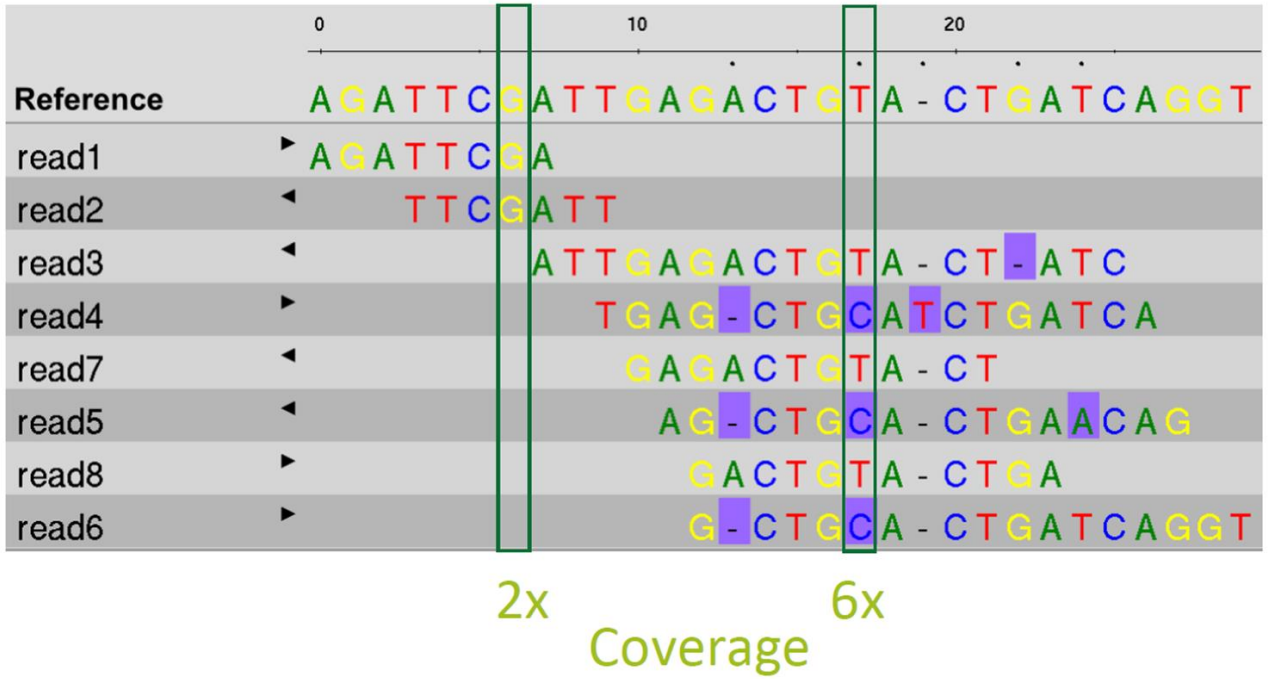
De novo assembly

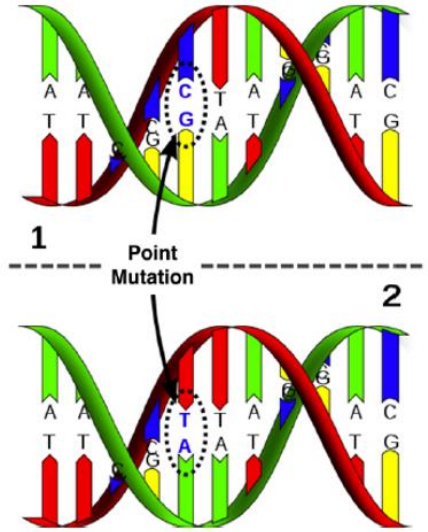
Mapping to reference

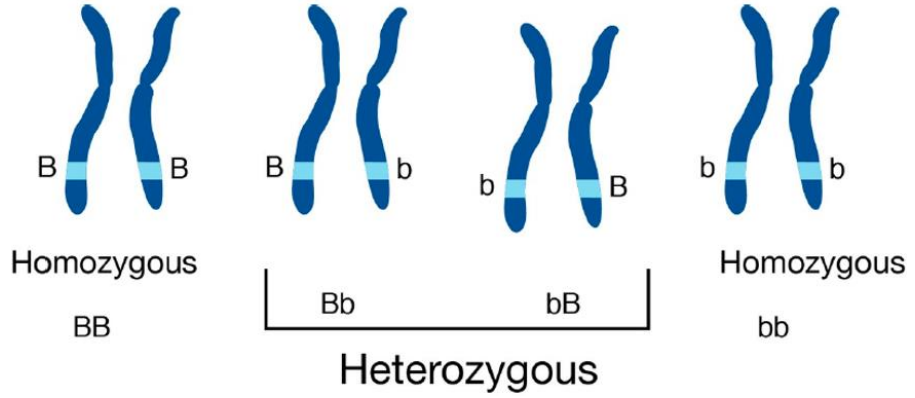
https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Human Genome Variation

- Human genomes are > 99% similar by sequence
- A typical human genome has ~ 5 million variants with 3-4 million single nucleotide variants
- Humans are diploid

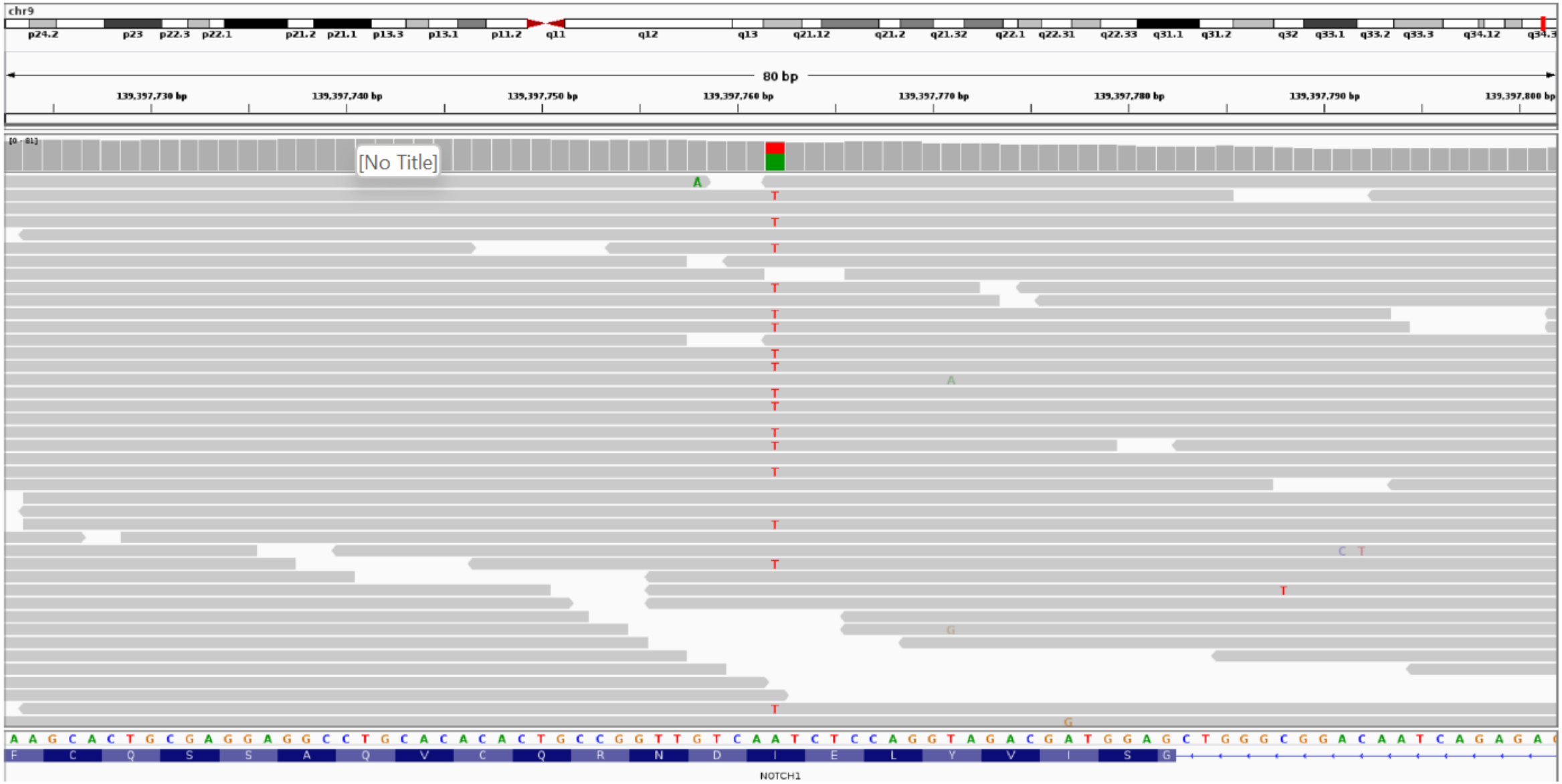




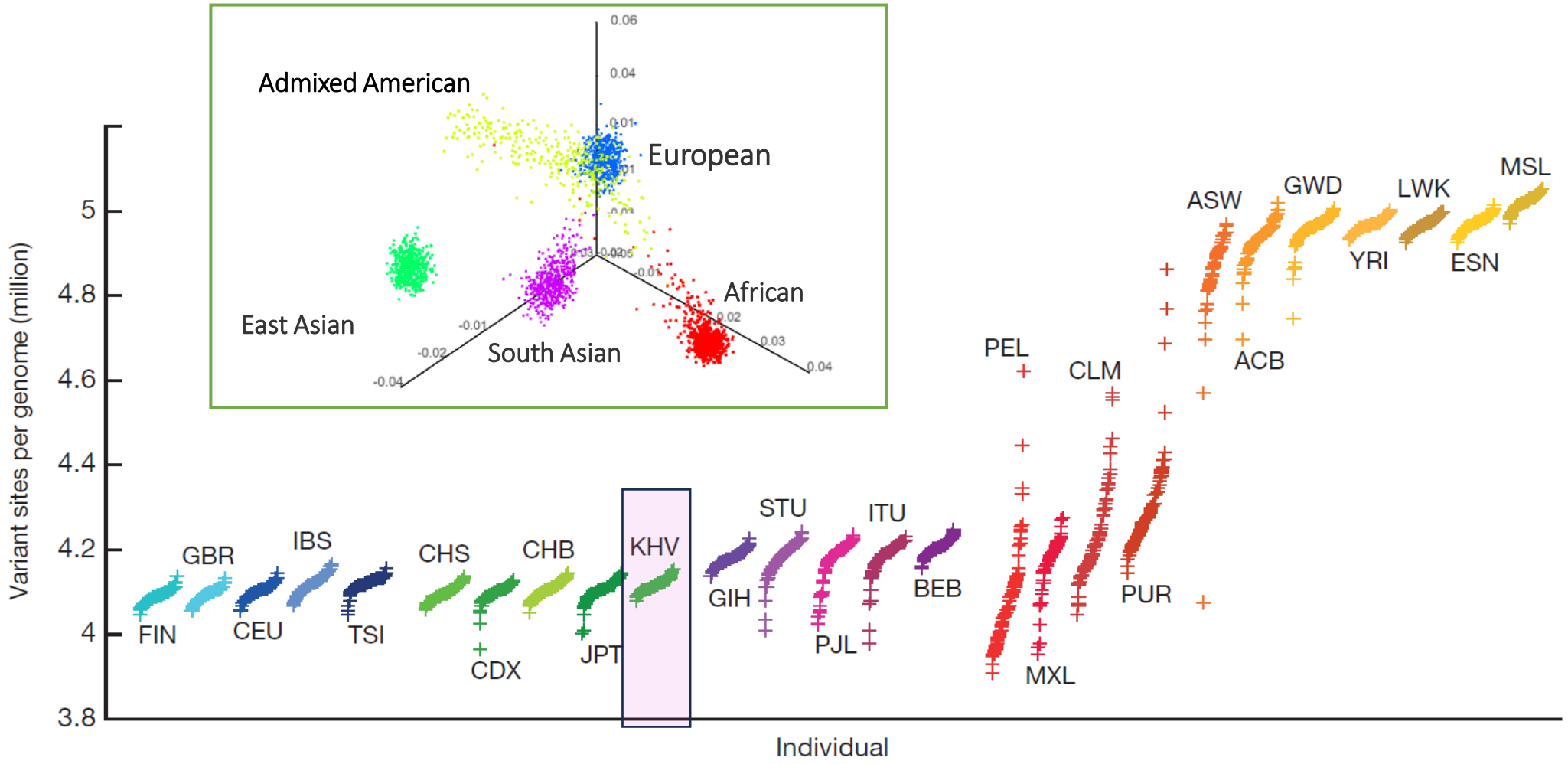


https://rosalind.info/media/point_mutation.png, https://en.wikipedia.org/wiki/Zygosity

Alignment and variant viewers



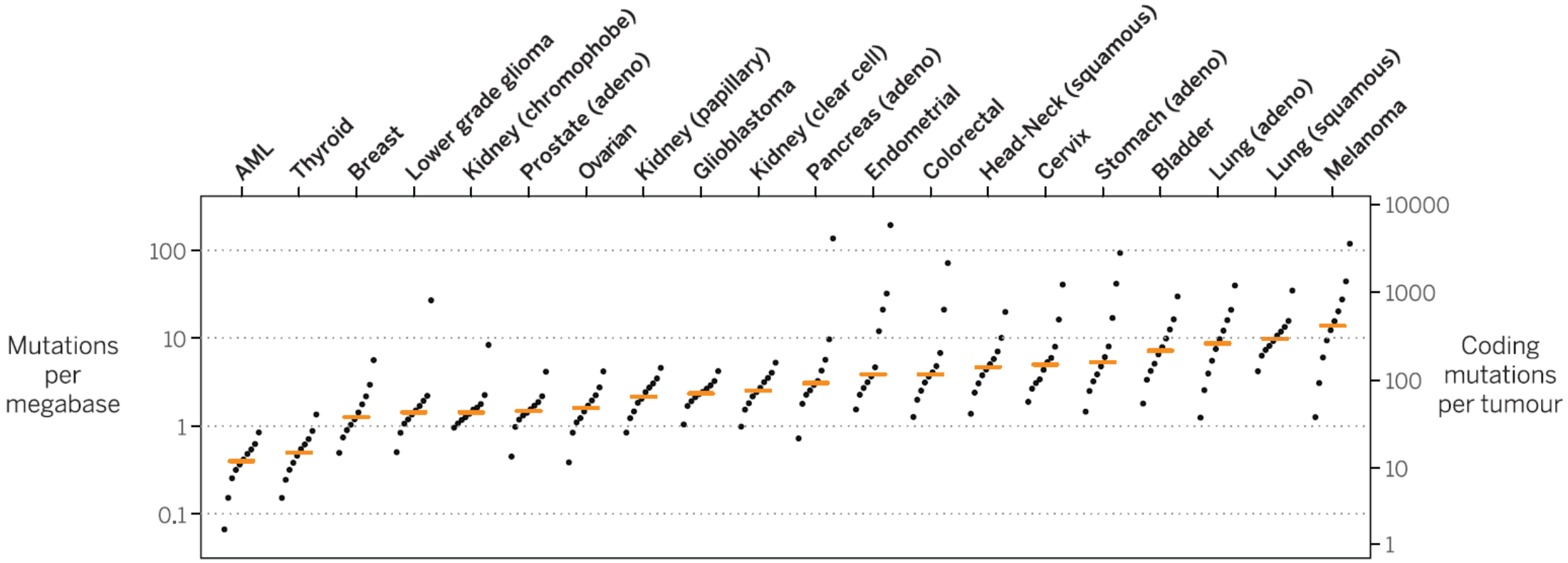
Human Genome Variation: 1000 Genomes Project



The number of variant sites per genome of 1K human genomes project (2015)
Kinh in Ho Chi Minh City, Vietnam (KHV)

doi:10.1038/nature15393

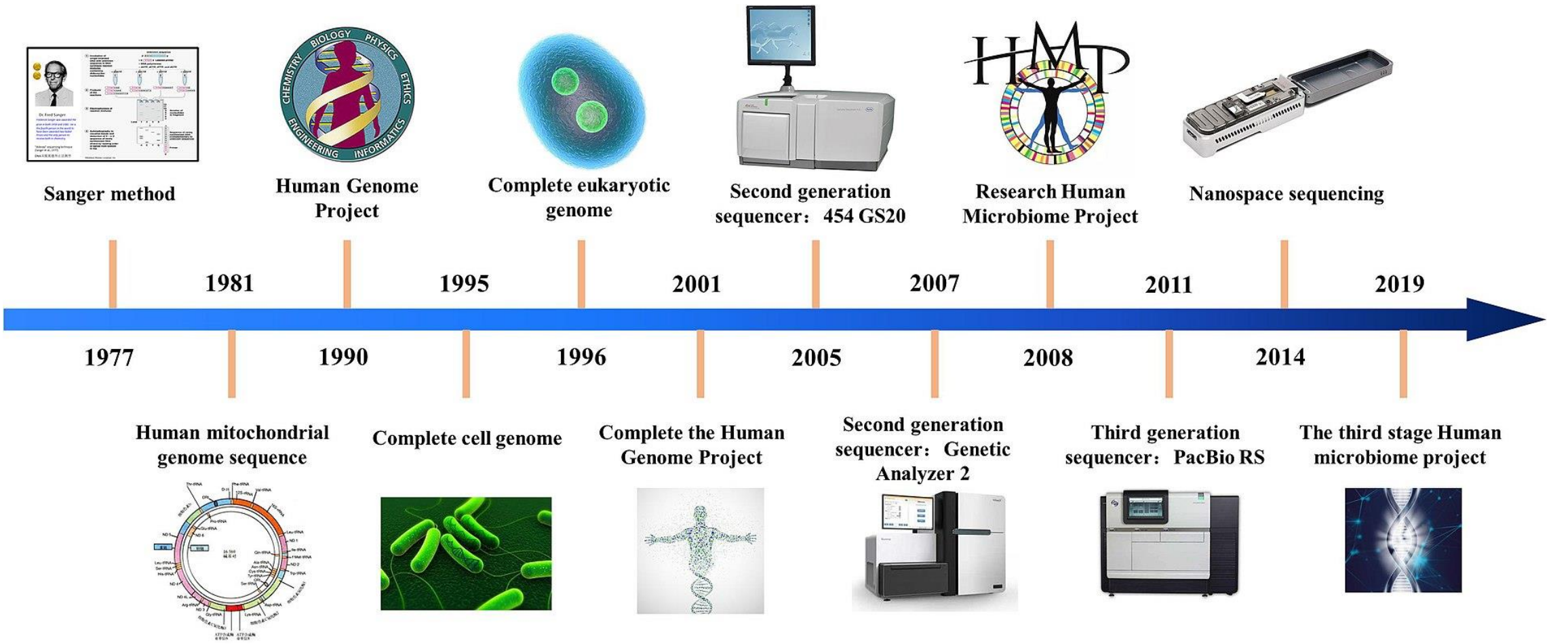
Cancer Genome Somatic Variation



Mutation burden in 20 tumor types and relative contribution of different mutational processes.
For each tumor type, samples were divided into deciles on the basis of their mutation burden. (2015)

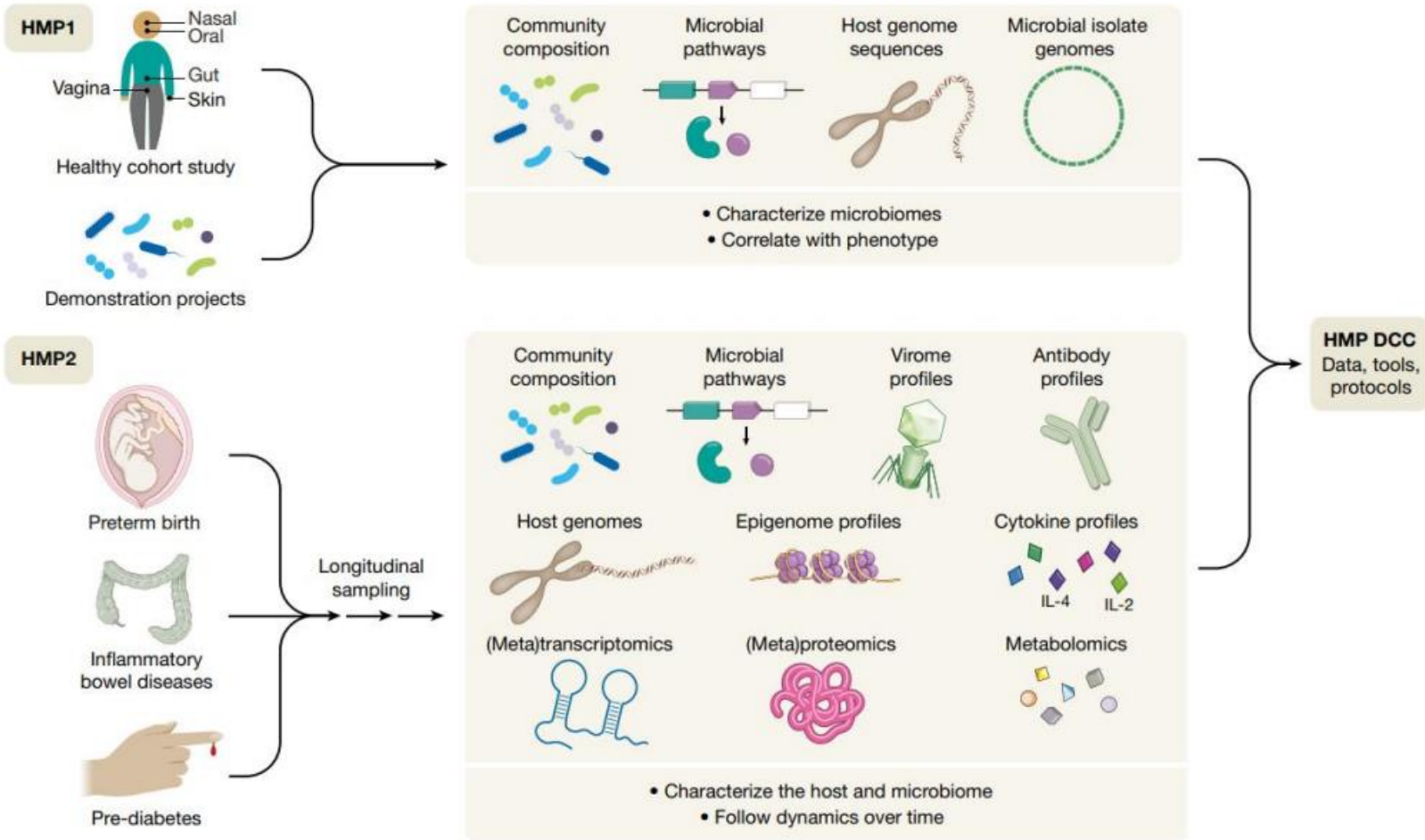
DOI: 10.1126/science.aab4082

High-throughput sequencing (HTS) methods



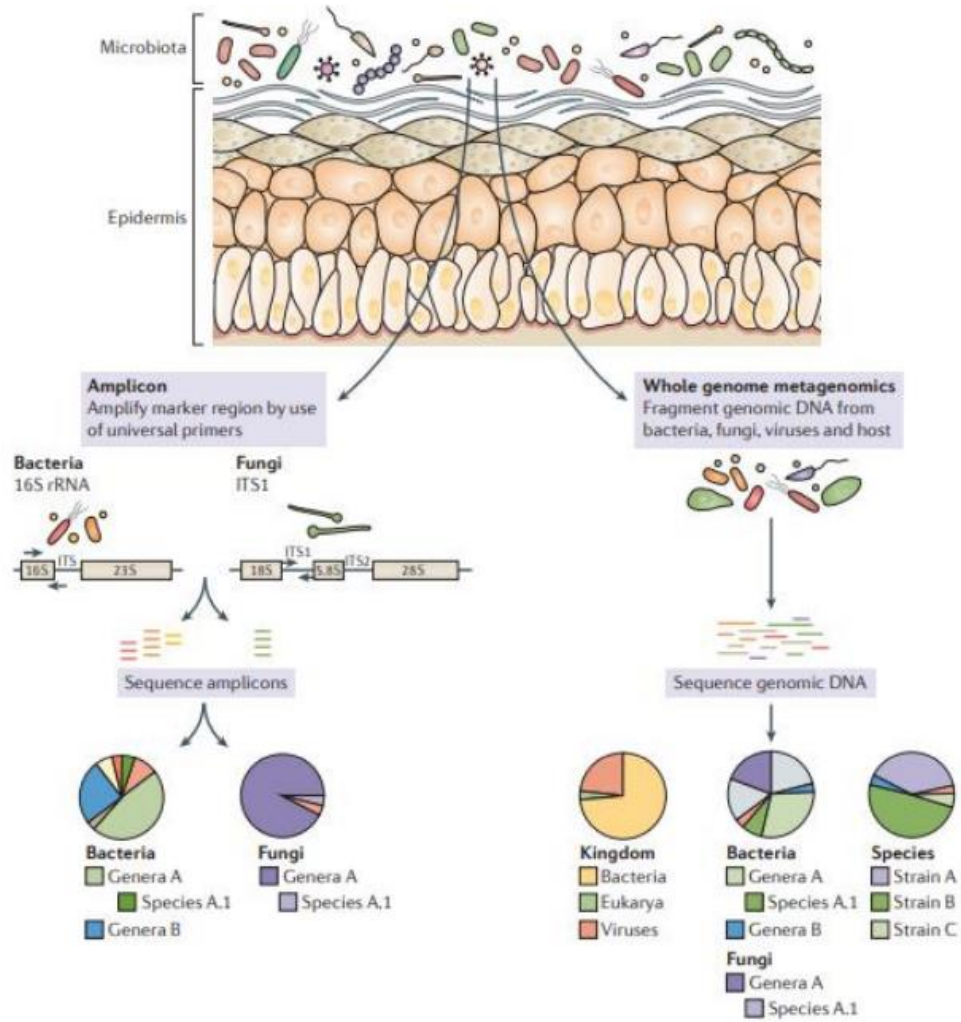
https://en.wikipedia.org/wiki/DNA_sequencing

The first and second phases of the NIH Human Microbiome Project

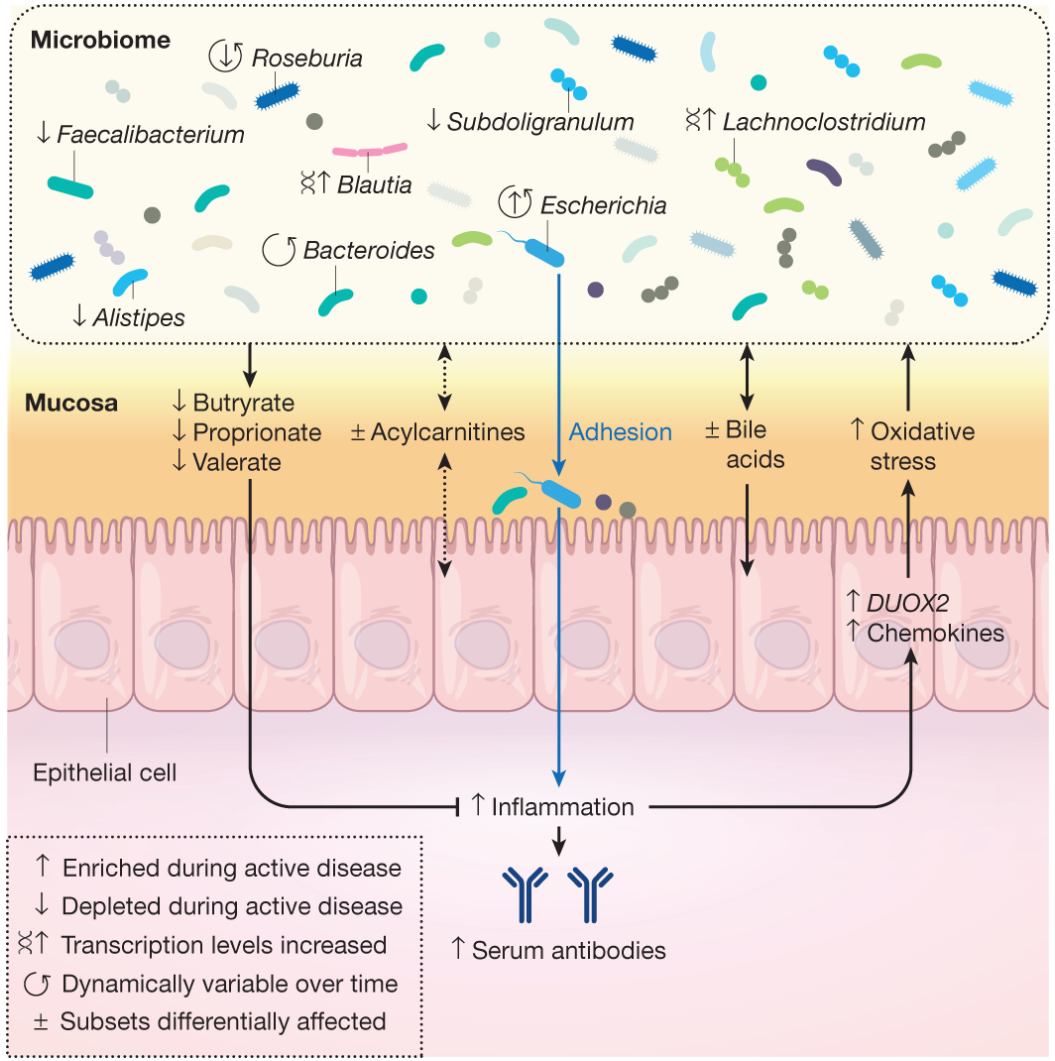


https://www.nature.com/articles/s41586-019-1238-8

Human Skin Microbiome Project

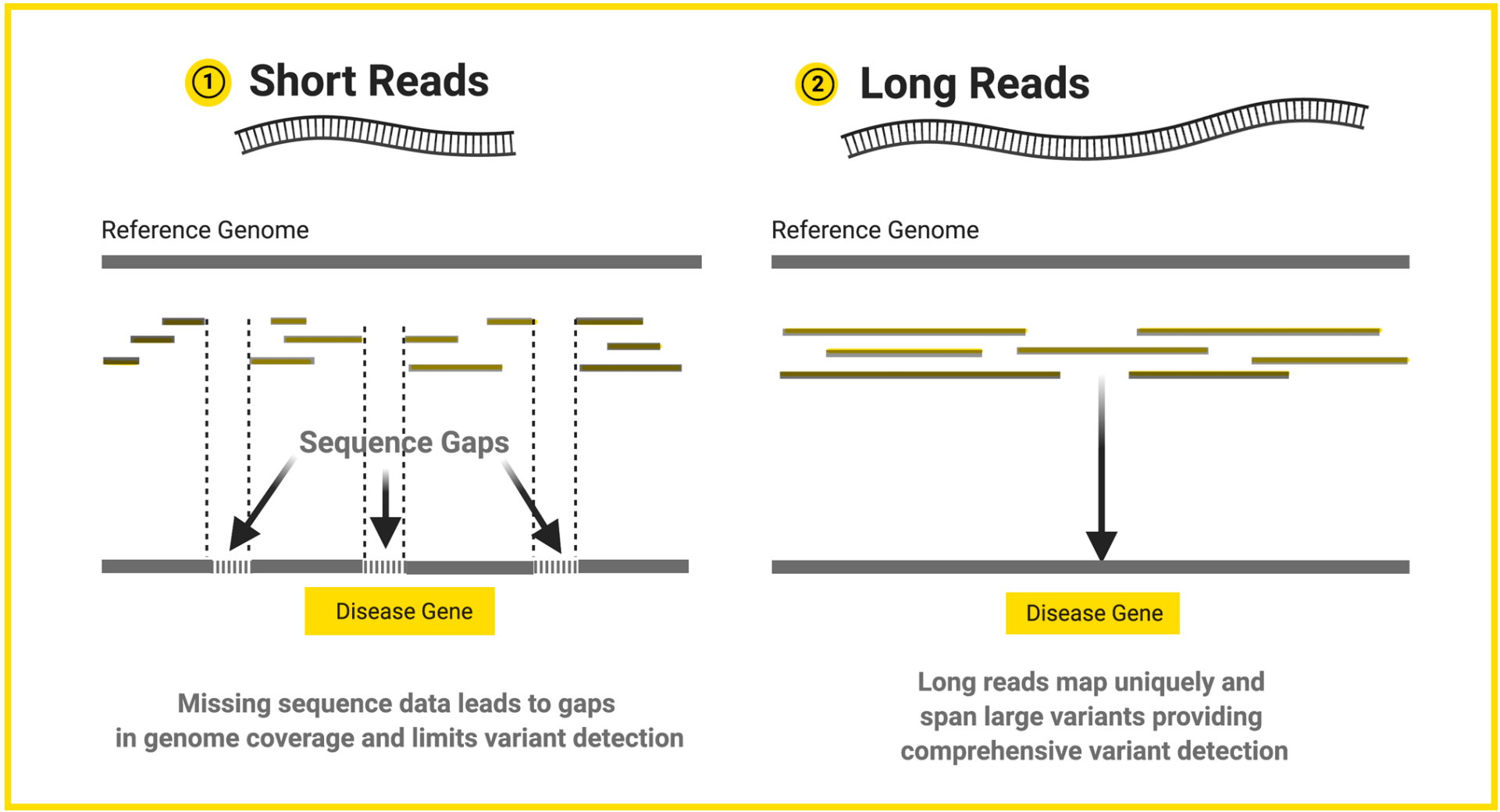


Host–microbiome dynamics in inflammatory bowel diseases



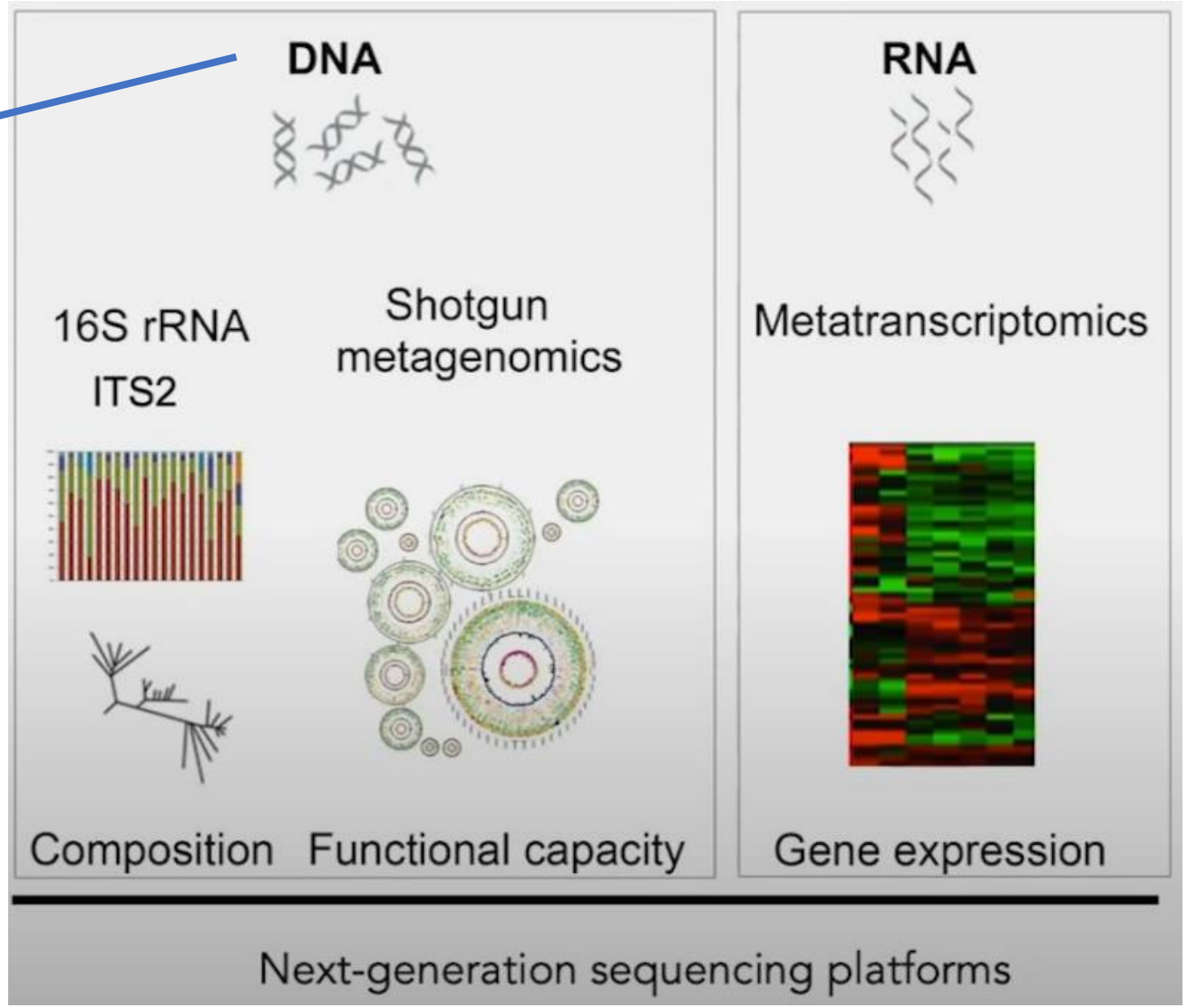
https://www.nature.com/articles/s41586-019-1238-8

Short read vs Long read

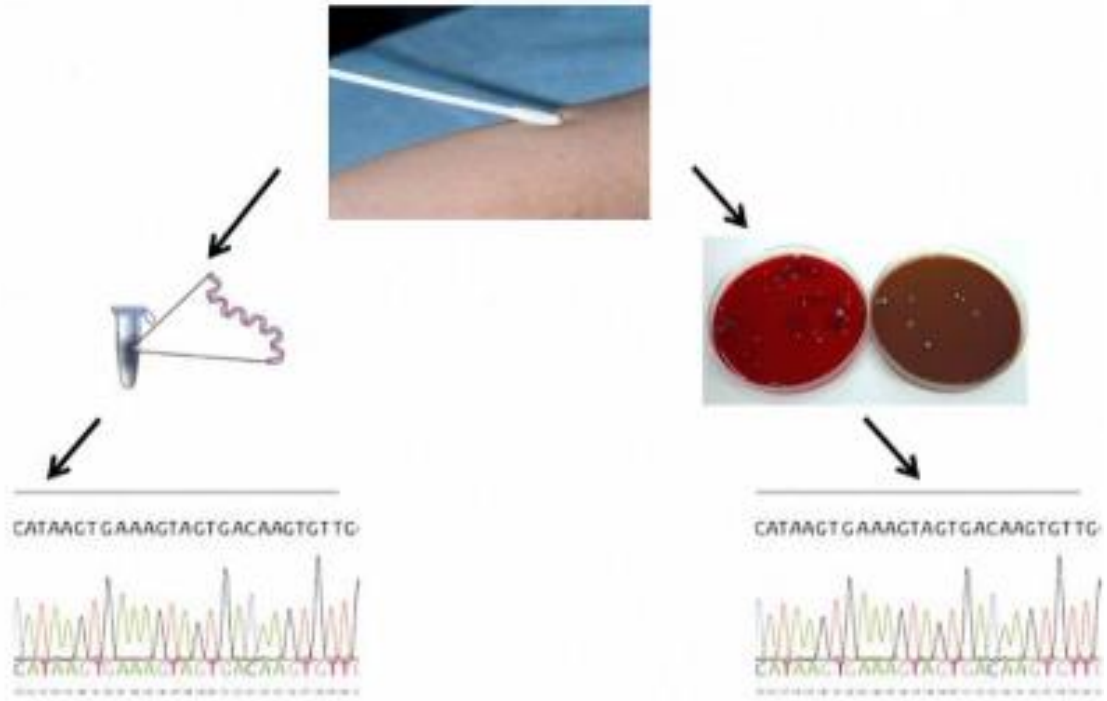


NGS Applications in microorganism research

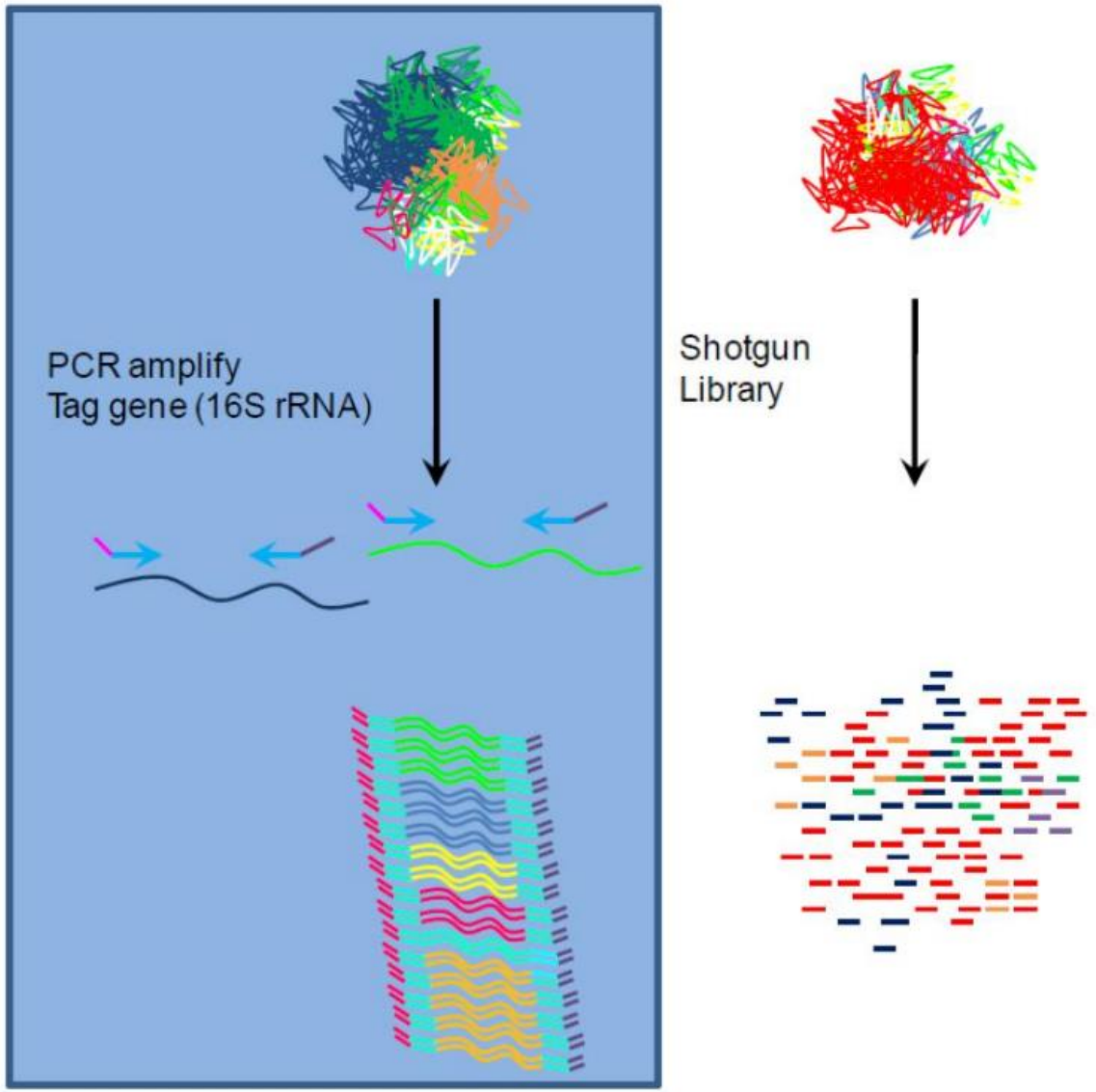
Whole Genome Sequencing (WGS)



Direct sequencing vs. culture-based methods



Target Gene vs Metagenomics



NGS Applications in microorganism research

- **DNA-seq**:
  1. Target sequencing (some genes): serotyping and genotyping
  2. Whole Genome Sequencing (WGS): multi locus genotyping, genetic diversity and evolution study
  3. Amplicon metagenomics (16S/ITS): identification/classification/ surveillance study
  4. Shotgun metagenomics: resistome
- **RNA-seq**:
  1. Transcriptomics: gene expression for a isolate
  2. Meta-transcriptomics: gene expression for a bulk sample
  3. Single cell RNA-seq: gene expression for individual cells

Thank you for your attention!