

# Big Data & Bioinformatics .

By Elsayed Hegazy  
Research Assistant, Nile University.

---

WHY?

HOW?

Why?

---

# Big Data

The power of ~~petabytes~~ Exabyte



1000 terabytes



1 000 000 terabytes

# Data Sources in Life science.

---

- Medical labs.
- Scanning centers.
- IoT and Smart watches.
- EHR (Electronic Health Record).
- Genetic Sequencing (**Genome**).
- Microbial Sequencing.
- Proteomic Sequencing.
- And many more . . .

# Genome like a book

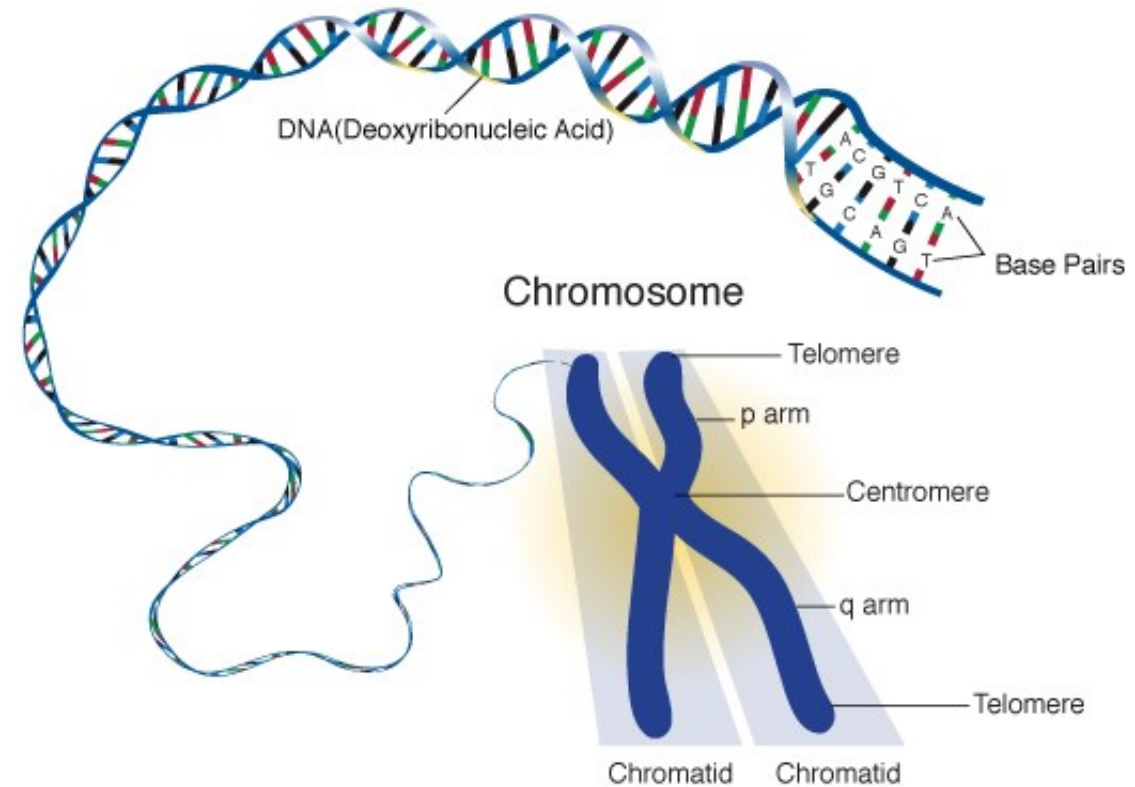
Genome

organized into chapters (chromosomes)

sentences (genes)

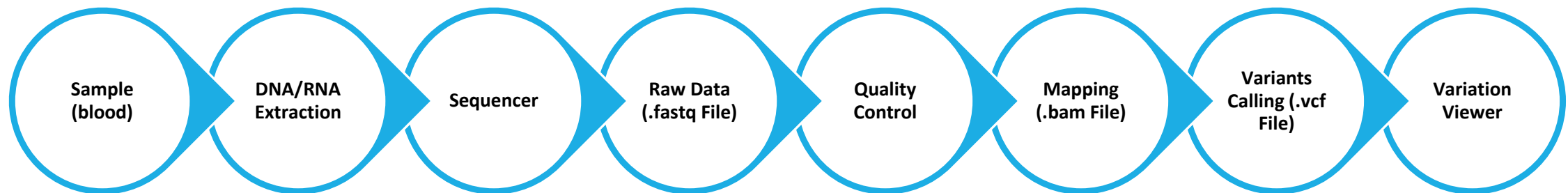
words (codons = 3 letters = Amino acid)

DNA is composed of letters (A,T,C and G)



MIT Professor Eric Lander says "Genome bought the book hard to read"

# Next Generation Sequencing workflow



# .FASTQ File (raw data)

---

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%)++)(%%%) .1***-+*'')**55CCF>>>>>>CCCCCCC65
```

The character '!' represents the lowest quality while '~' is the highest. Here are the quality value characters in left-to-right increasing order of quality (ASCII).

# .BAM & .SAM files (Aligned file) Alignment

A **BAM file** (.bam) is the binary version of a **SAM file**. A **SAM file** (.sam) is a tab-delimited text **file** that contains sequence alignment data.

For example:

```

RefPos:      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read:  ACTAGAATGGCT
  
```

Aligning these two:

```

RefPos:      1  2  3  4  5  6  7      8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T      G  A  A  C  T  G  A  C  T  A  A  C
Read:           A  C  T  A  G  A  A      T  G  G  C  T
  
```

# Accuracy of sequence file.

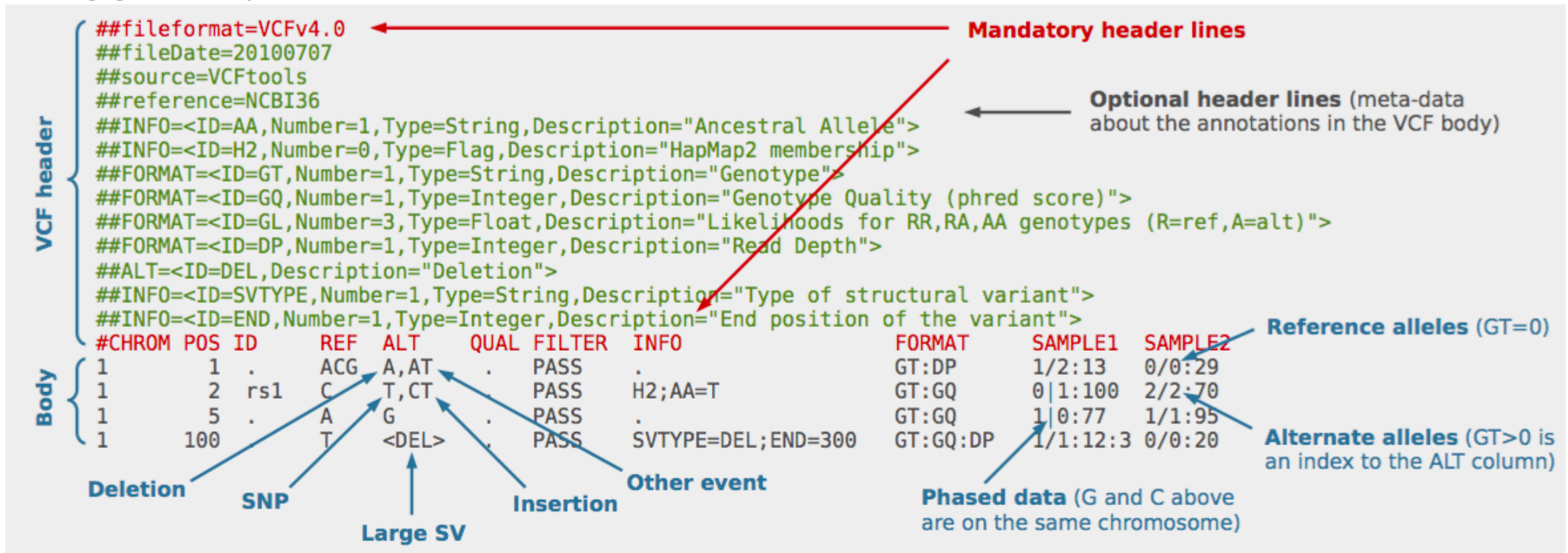
---

Reference	CCGTTAGAGTTACAATTCGA	> 3 Billion letter
Read 1	CCGTTAGAGTTACAATTCGA	
Read 2	CCGTTAGAGTAACAATTCGA	
Read 3	CCGTTAGAGTTACAATTCGA	
Read 4	CCGTTAGAGTTACAATTCGA	
Read 5	CCGTTAGAGTAACAATTCGA	
Read 6	CCGTTAGAGTAACAATTCGA	
Read 7	CCGTTAGAGTTACAATTCGA	
Read 8	CCGTTAGAGTTACAATTCGA	
Read 9	CCGTTAGAGTTACAATTCGA	



# Variant Calling (.vcf file)

The **Variant Call Format (VCF)** specifies the format of a text file used in bioinformatics for storing gene sequence variations.



The diagram illustrates the structure of a VCF file, divided into a **VCF header** and a **Body**.

**VCF header:** Contains meta-information. The first line, `##fileformat=VCFv4.0`, is labeled as a **Mandatory header line**. Subsequent lines, such as `##fileDate=20100707`, `##source=VCFtools`, and `##reference=NCBI36`, are labeled as **Optional header lines (meta-data about the annotations in the VCF body)**. The header also includes `##INFO` and `##FORMAT` lines that describe the fields used in the body.

**Body:** Contains variant records. Each record is a line of text with columns: `#CHROM`, `POS`, `ID`, `REF`, `ALT`, `QUAL`, `FILTER`, `INFO`, `FORMAT`, `SAMPLE1`, and `SAMPLE2`.

Annotations for the body include:

- Reference alleles (GT=0):** Points to the first allele in the ALT field (e.g., 'A' in 'A,AT').
- Alternate alleles (GT>0 is an index to the ALT column):** Points to the second allele in the ALT field (e.g., 'AT' in 'A,AT').
- Deletion:** Points to the `<DEL>` symbol in the ALT field.
- SNP:** Points to the 'A' and 'T' alleles in the ALT field.
- Large SV:** Points to the 'G' allele in the ALT field.
- Insertion:** Points to the 'CT' alleles in the ALT field.
- Other event:** Points to the 'G' allele in the ALT field.
- Phased data:** Points to the vertical bar in the FORMAT field (e.g., '0|1:100').

# Personal Genome applications.

---

- Personal customization of drugs.
- Personal customization of food for better lifestyle.
- Prediction of future disease (eg: Cancer.)
- Detecting of microbial infection.
- Classify disease stage and type.
- Detecting pathogenic mutation.
- Drug Docking.
- And many more . . .



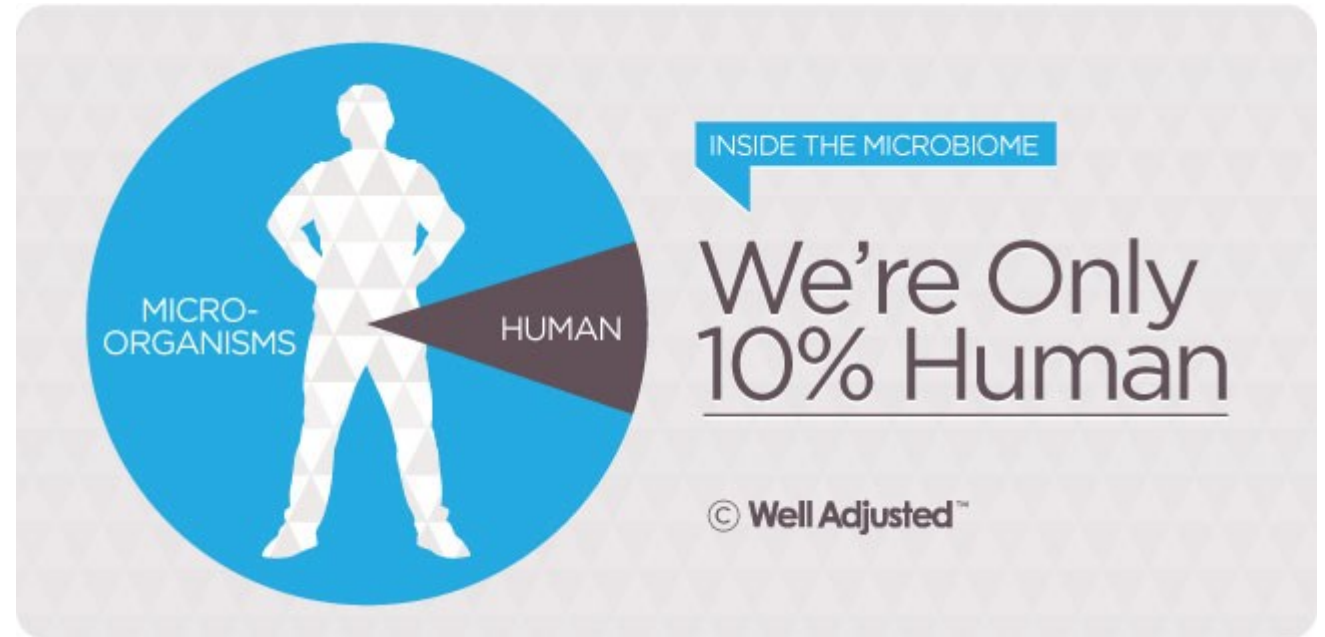


# Human Microbiome Project.

---

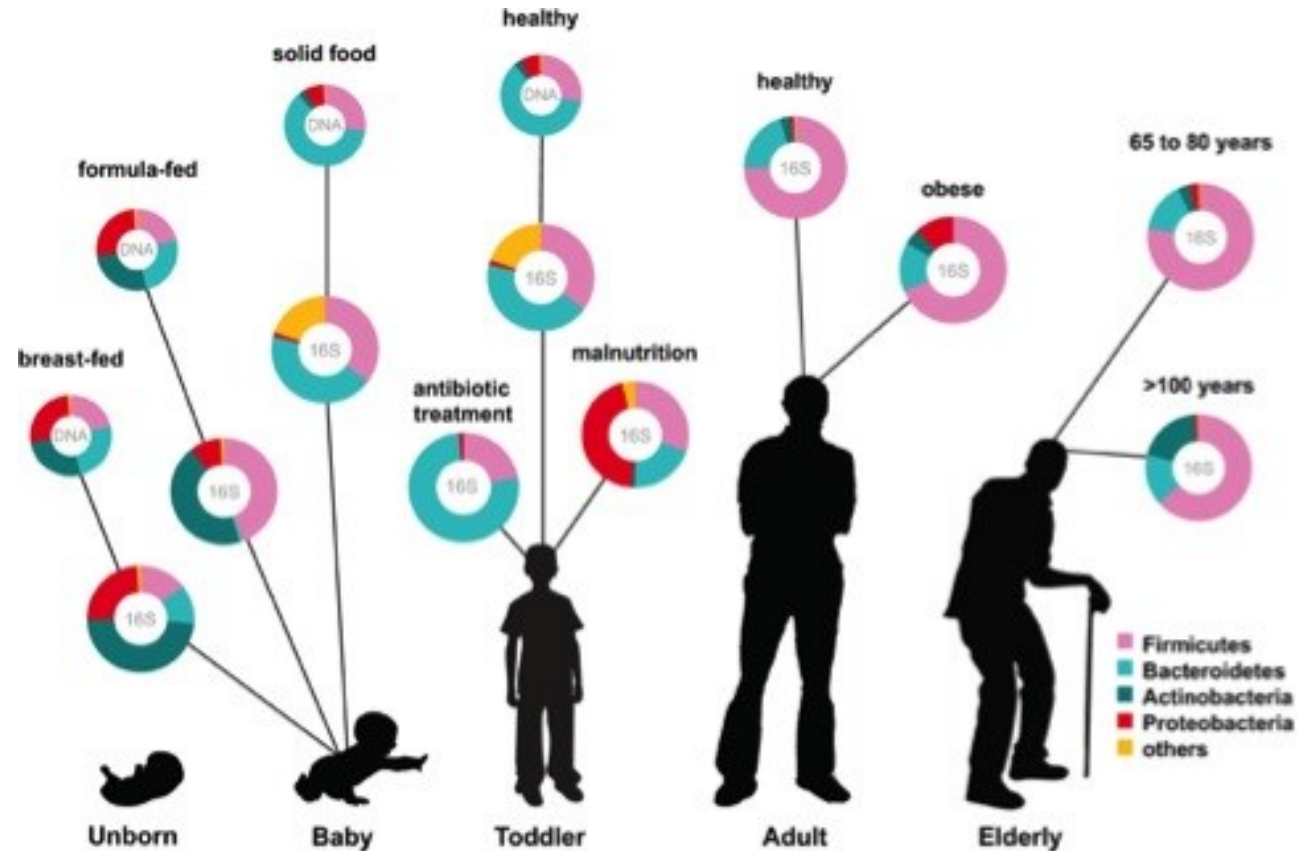
## The Human Microbiome: Our Other Genome

its the full collection of genomes of all the microbes in a Human.

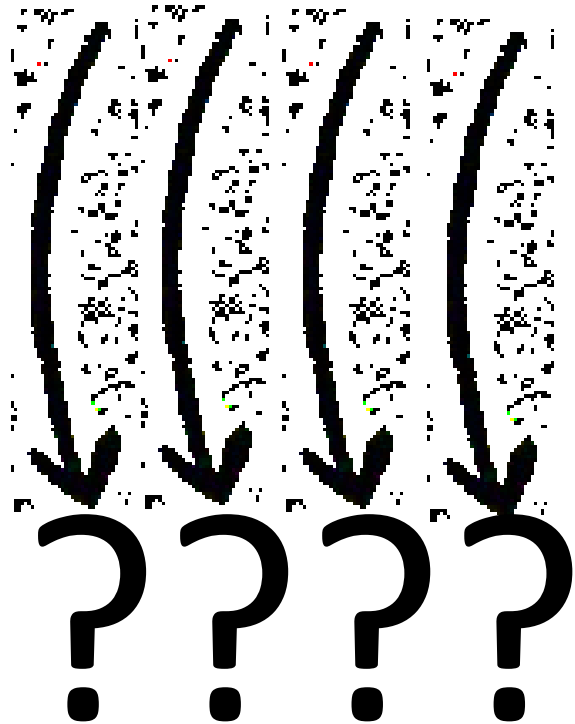


# Human Microbiome Project.

The characteristics of human microbiota change over time in response to varying environmental conditions and life stages



# Ask Me A Question




**It's Free**



---

 [elsayedhegazy@live.com](mailto:elsayedhegazy@live.com)

 [s.hegazy@nu.edu.eg](mailto:s.hegazy@nu.edu.eg)

 0100 6 99 88 36

 <https://eg.linkedin.com/in/elsayedhegazy>

 [@elsayedhejazy](https://twitter.com/elsayedhejazy)