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)

↓

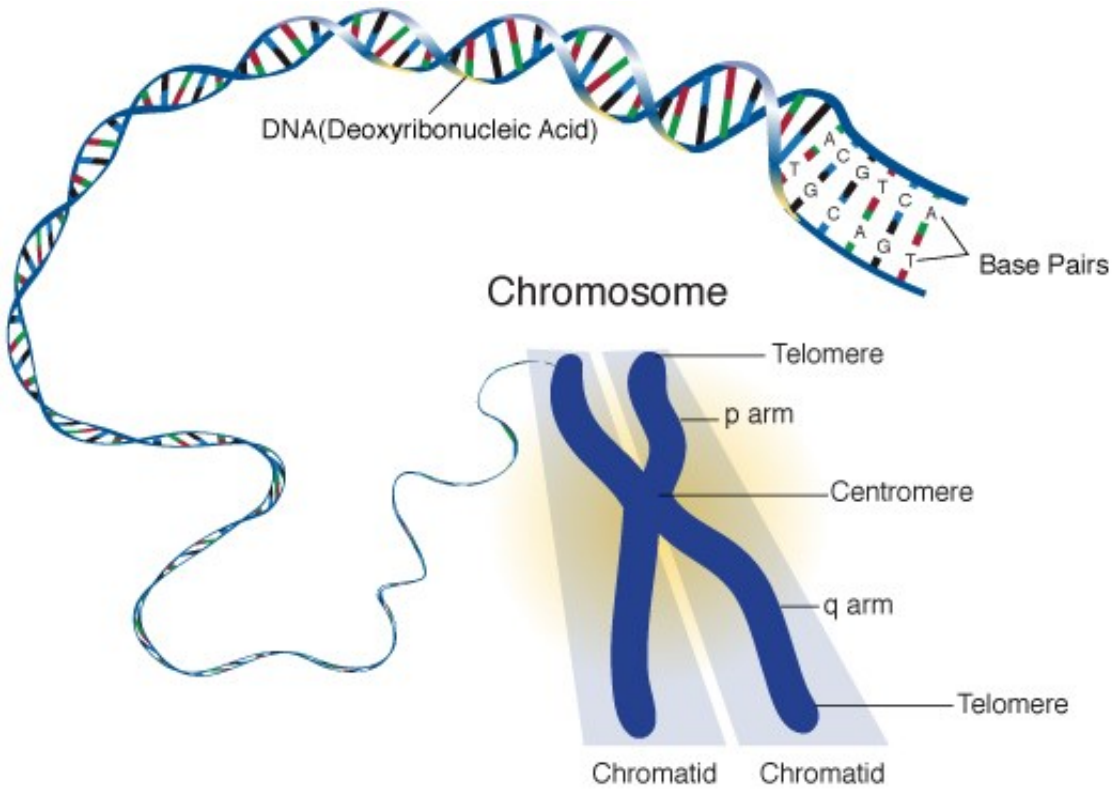
sentences (genes)

↓

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





Sample (blood) → DNA/RNA Extraction → Sequencer → Raw Data (.fastq File) → Quality Control → Mapping (.bam File) → Variants Calling (.vcf File) → Variation Viewer

.FASTQ File (raw data)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).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
```

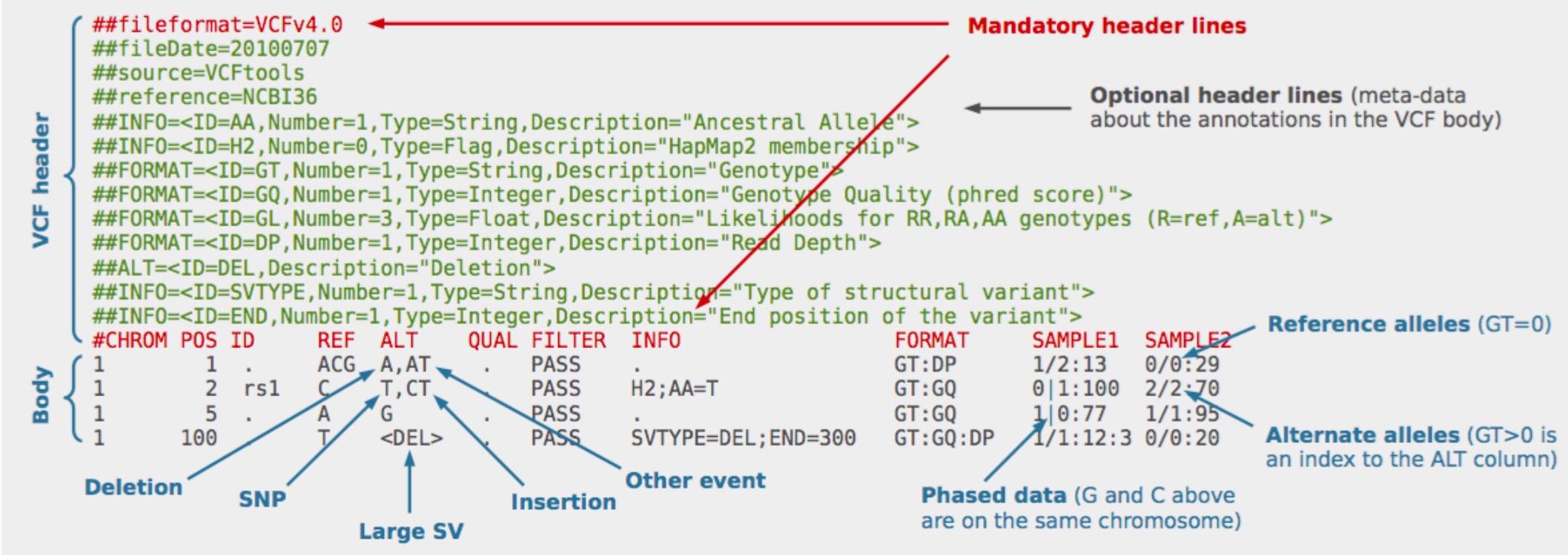
Accuracy of sequence file.

Reference	CCGTTAGAGTTACAATTCGA
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

> 3 Billion letter

Variant Calling (.vcf file)

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



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 . . .

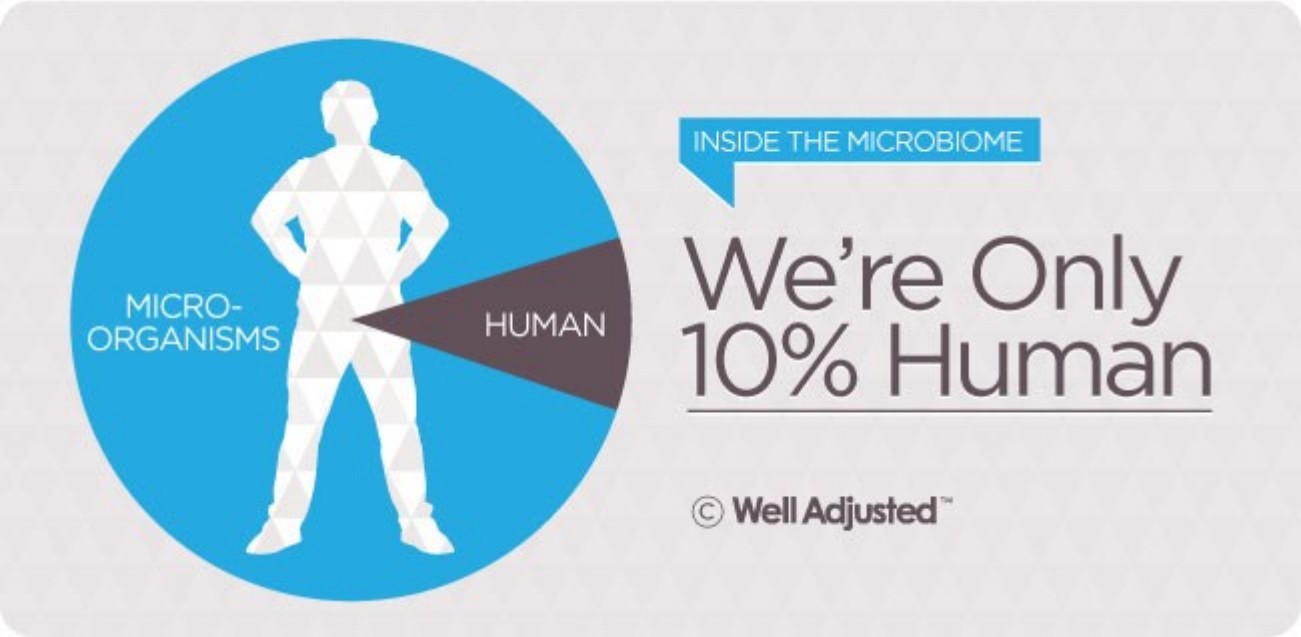
Big Data Demand

- Storage ? (Think about Personal genome for all Egyptians)
- Applications (Apache Spark)

Human Microbiome Project.

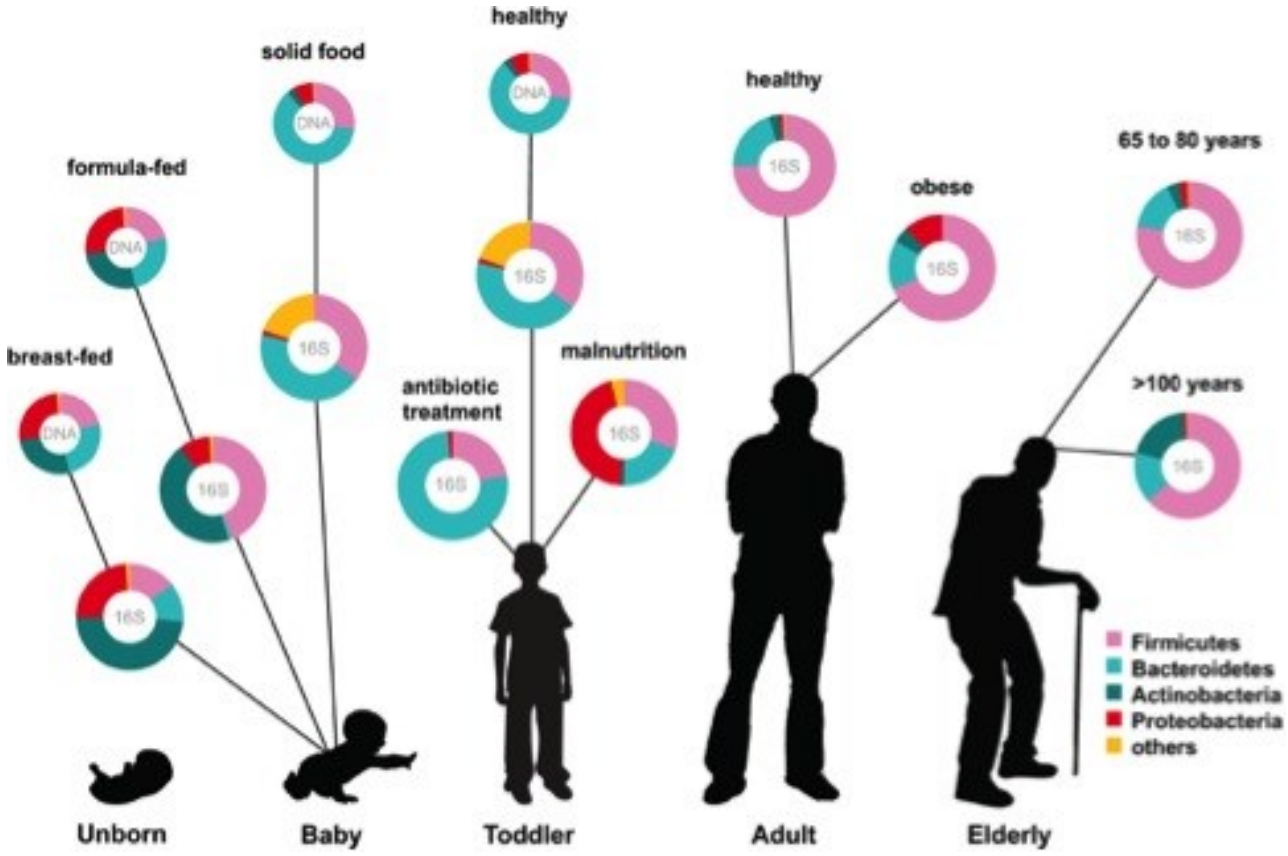
The Human Microbiome: Our Other Genome

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

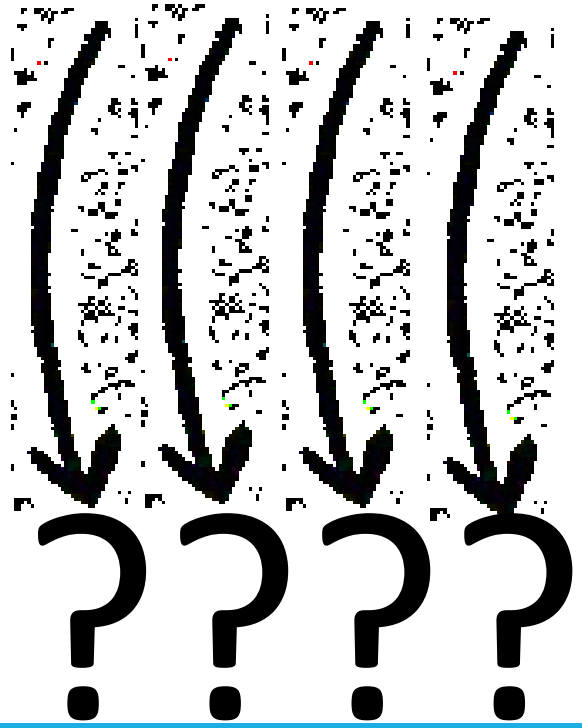


Human Microbiome Project.

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

s.hegazy@nu.edu.eg

0100 6 99 88 36

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

@elsayedhejazy