Properties of bacterial genomes

Matthieu Barret (EmerSys, IRHS)

Projet ABDC

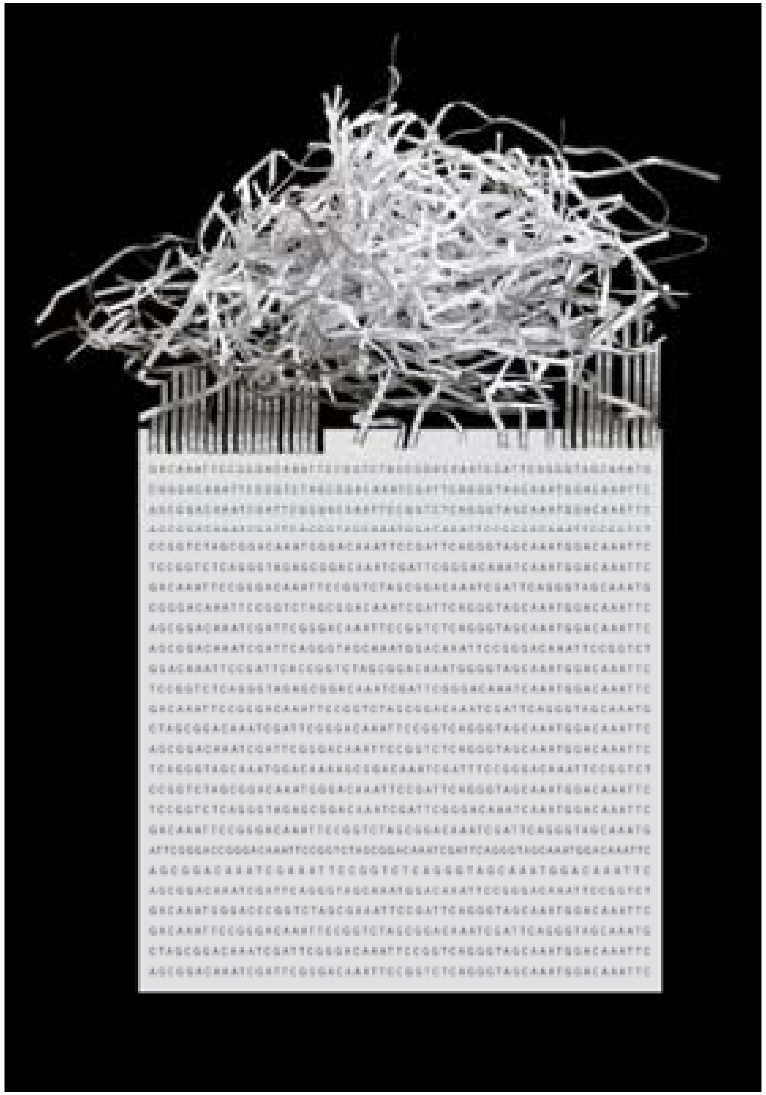
Jeudi 9 avril 2015

General characteristics of bacterial genomes

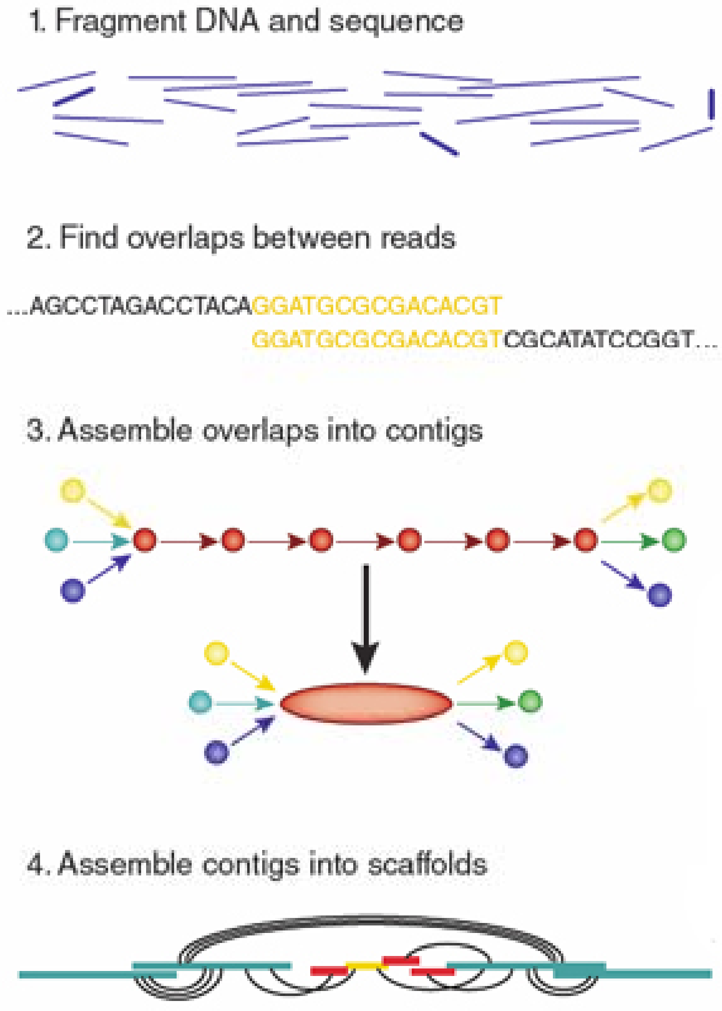
- Bacterial chromosome = large circle of double stranded DNA
- Bacterial plasmid = circle of double stranded DNA 1000 to 10 smaller than chromosome
- Replicon = any genetic entity that controls its own replication (chromosome + plasmid)
- Genome = stable replicon of one organism (most plasmids lost in standard culture media)

adapted from Mrazek and Summers, 2008

De novo genome assembly



Kelly Howe, Lawrence Berkeley Laboratory



1. Fragment DNA and sequence

2. Find overlaps between reads

...AGCCTAGACCTACAGGATGCGCGACACGT

GGATGCGCGACACGTCGCATATCCGGT...

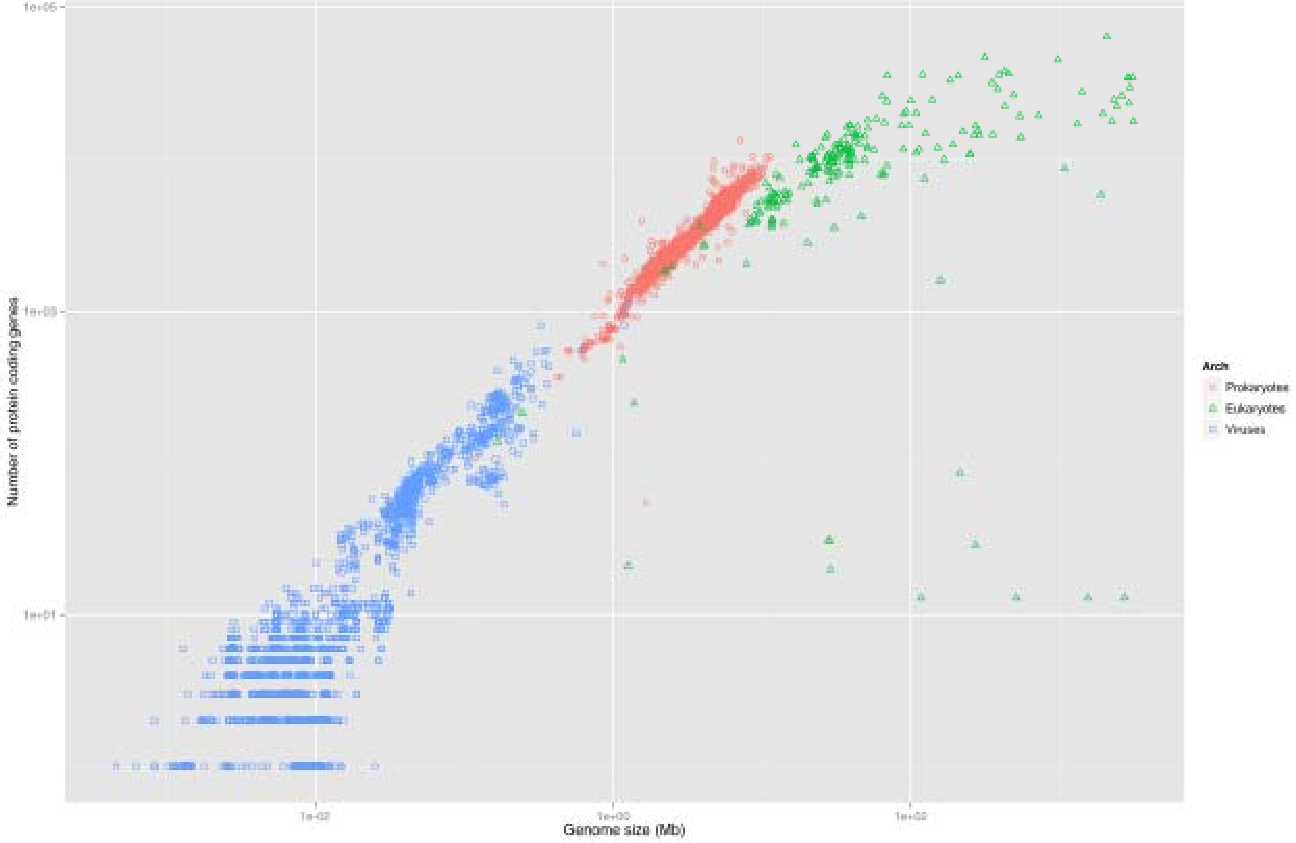
3. Assemble overlaps into contigs

4. Assemble contigs into scaffolds

Michael Schatz, Cold Spring Harbor

Properties of bacterial genomes

- Chromosome sizes (0.5-10 Mb) correlate with the number of genes
- Protein coding regions occupy about 90% of a prokaryotic genome
- Average gene density is a one gene per 1 kb

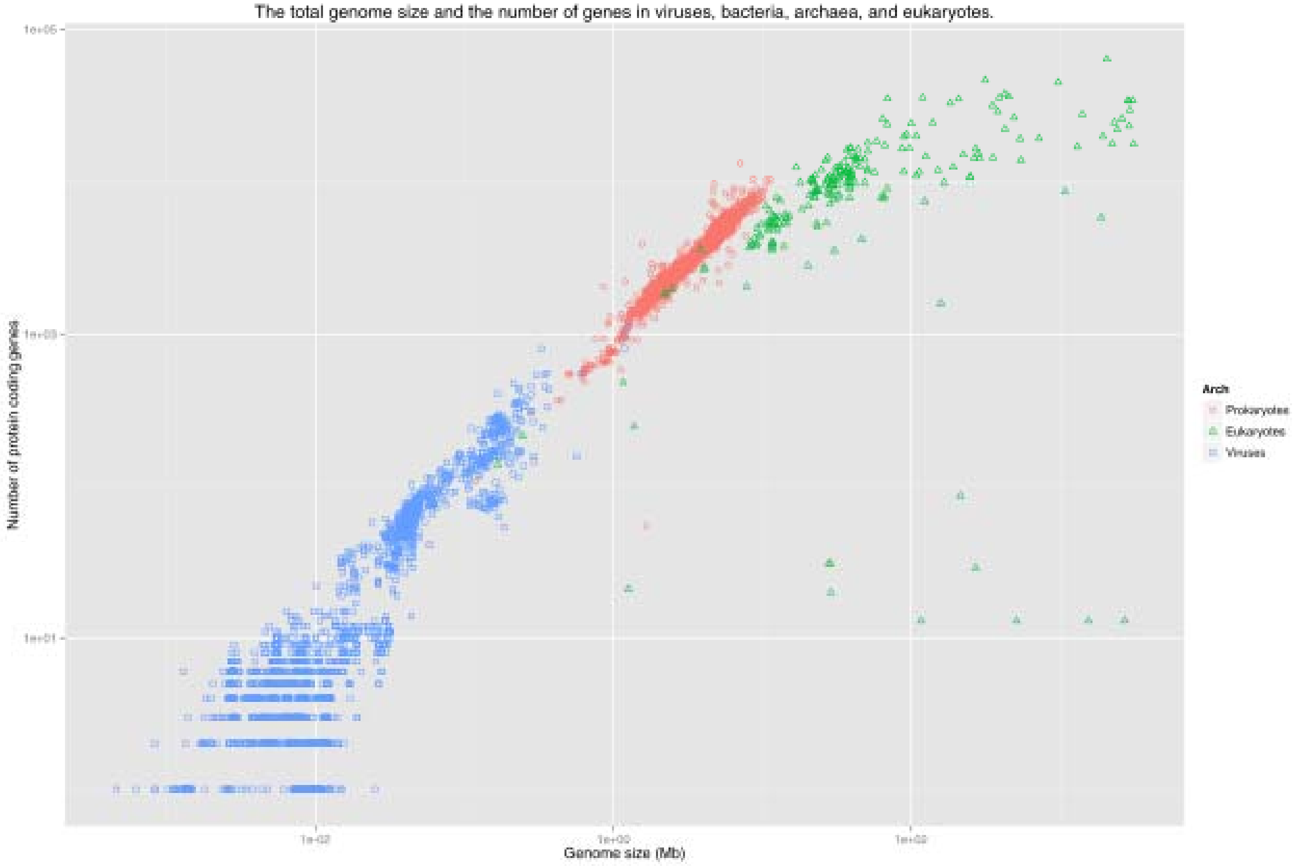


The total genome size and the number of genes in viruses, bacteria, archaea, and eukaryotes.

adapted from Koonin and Eugene, 2011

Properties of eukaryotic genomes

- Eukaryotic genomes vary widely in size (10 Mb – 100 Gb)
- Protein coding regions also vary (3-70%) of a prokaryotic genome
- Presence of introns within immature mRNA



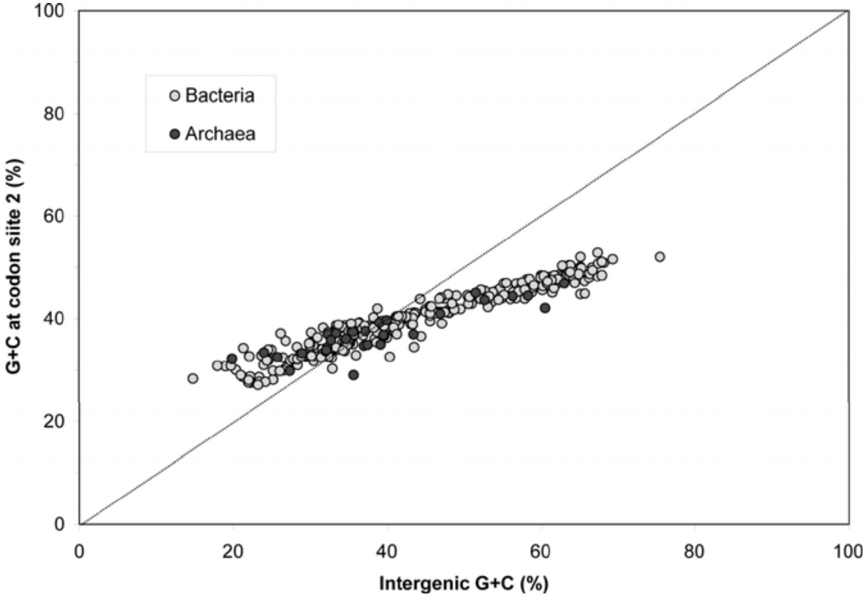
adapted from Koonin and Eugene, 2011

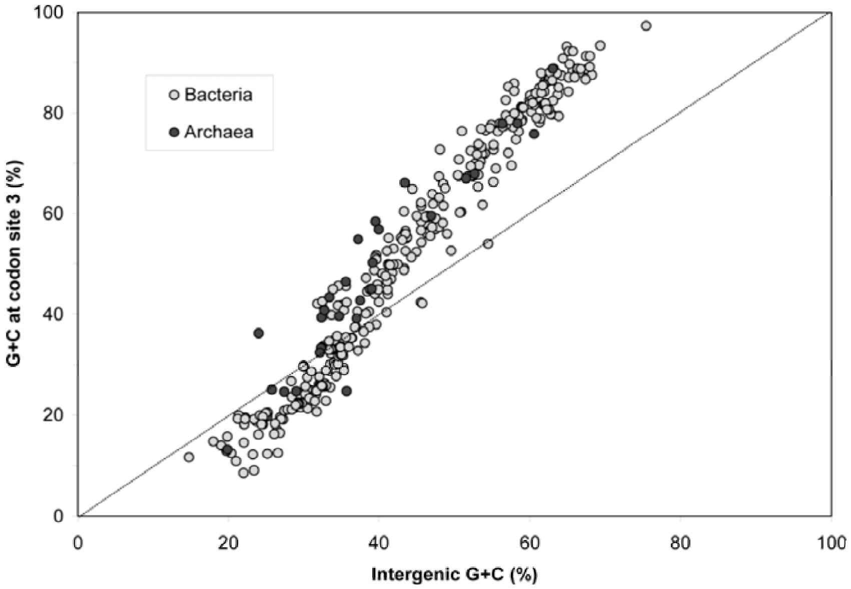
G + C content

- Bacterial genomes variable in their G + C content (from 25 to 75 %)

- Protein-coding genes have higher G+C content (about 10 % on average)

- G + C content varies significantly among the three codon position





adapted from Mrazek and Summers, 2008

Oligonucleotide composition

- If 2 nucleotides X and Y occur independently in a DNA sequence then

$$f\mathrm{XY} = f\mathrm{X}f\mathrm{Y}$$

- Significant deviation of *f*XY from *f*X*f*Y (dinucleotide frequency)

- Dinucleotide relative abundance vary significantly among genomes but remarkably stable within a genome

adapted from Mrazek and Summers, 2008

Synonymous codon usage

- Synonymous codons in genes are not used with equal frequency

- Synonymous codon usage differs significantly between genomes

- However differences exists in synonymous codon usage even among genes from the same genome (relates to gene expression level)

adapted from Mrazek and Summers, 2008

Repeats in bacterial genomes

- Longest expected sequence occurring at least twice is about 26 bp length (random)

- However large repeats often identified in bacterial genomes (duplicated rRNA genes and transposons)

adapted from Mrazek and Summers, 2008

Differences between prokaryotic and eukaryotic genomes

Properties	Prokaryote	Eukaryote
Chromosome	circular	linear
Size	0.5-10 Mb	10 Mb – 100 Gb
Coding regions	90%	3-70%
GC%	25-75	35-45
Dinucleotide frequency	stable within genome	variable between chromosome