Estimating the size of the bacterial pan-genome

Lapierre *et al*

I690 Metagenomics

Presenter: Heewook Lee

02/09/10

Genome Plasticity and Evolution

- Availability of many completed genomes has changed our views of genome evolution.
- Traditional view:
  - "The view of stable genomes that function as unchanging information repositories."
- Contemporary View:
  - "Dynamic view in which genomes frequently lose genes and incorporate foreign genetic materials"

What is the ‘pan-genome’?

- Also known as ‘supragenome’
- Denotes the set of all genes present in the genomes of members of a group of organisms, usually a species.

What's in the 'pan-genome'

- Genes present in only one organism
  - ORFans
- Genes in the genomes of few members of the group but not in all.
- Gene that are present in all genomes of the group
  - The core genome

Goal of the paper

- Apply the pan-genome concept to the bacterial branch.
- Evaluate the dynamics of genome and gene family evolution
- Characterizing two modes of evolution
  - Reuse with variation
  - *De novo creation*

Computing the 'pan-genome'

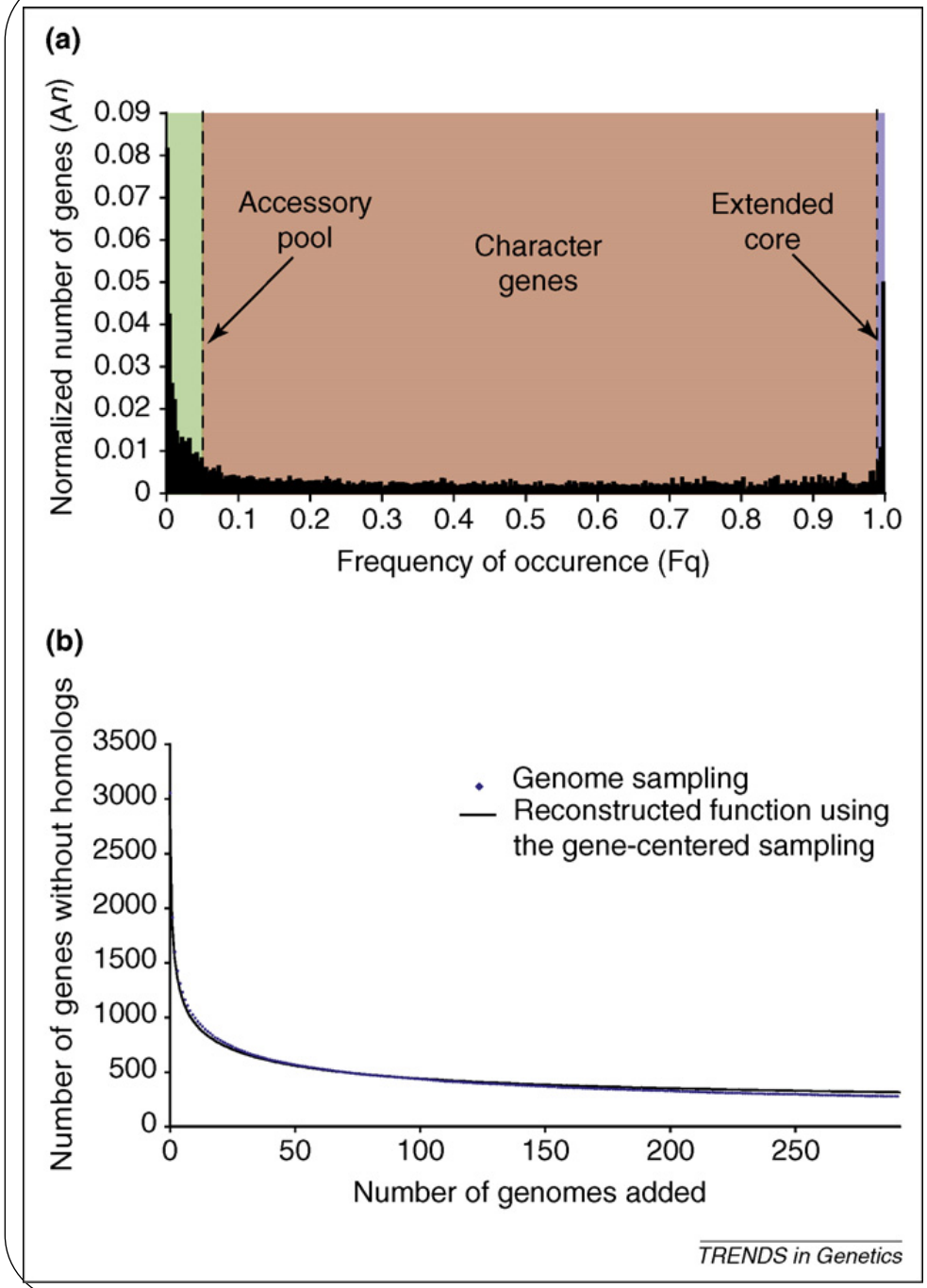
- Genome-oriented method by Tettelin *et al*
  - Define the pan-genome consisted of tracking the number of unique genes among genomes in successive blast searches.
  - Useful when there are a limited number of genomes.
  - Computationally difficult when the sample size grows

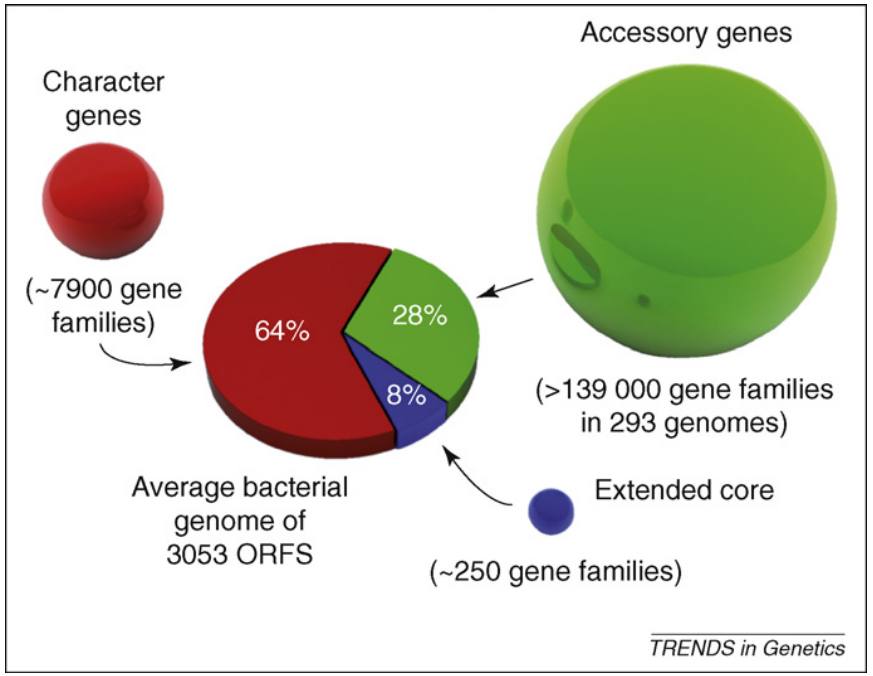
The ‘pan-genome’ via the gene-oriented method

- Sampling of genes
- Calculating the frequency of occurrence of genes in genomes
- Extrapolating back the sampling curve of the the actual pan-genome of the group.

Bacterial pan-genome

- 15,000 ORFs were randomly selected from 293 genomes
- BLAST search(each gene to genomes with bitscore > 50)
- Used both the gene-oriented and the genome-oriented methods to confirm comparability.


(a)
Normalized number of genes (An)
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Accessory pool
Character genes
Extended core
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Frequency of occurence (Fq)
(b)
Number of genes without homologs
3500
3000
2500
2000
1500
1000
500
0
Genome sampling
Reconstructed function using the gene-centered sampling
0 50 100 150 200 250
Number of genomes added
TRENDS in Genetics


Accessory genes
Character genes
(~7900 gene families)
64%
28%
8%
(>139 000 gene families in 293 genomes)
Average bacterial genome of 3053 ORFS
Extended core
(~250 gene families)
TRENDS in Genetics

Breakdown of the bacterial ‘pan-genome

	Genome Centered Approach	Gene Centered Approach			Average Gene Centered
	293 Genomes	293 Genomes	573 Genomes	508 Genomes	
Extended Core	6.8%	8%	5.6%	9.5%	**7.7%**
Character Genes	62.7%	63.7%	64.8%	61%	**63.2%**
Accessory Genes	30.4%	28.2%	29.6%	29.5%	**29.1%**

The Extended Core

- Conservative nature of evolution
- No successful replacement of the core gene within several billion of years of bacterial evolution
- Under high selective pressure for a function that prevents drastic changes.
- This set does not correspond to the minimal set of survival genes. More of a backbone of essential components.

The Character Genes

- Main component of every bacterial genome (~64% of the total genes on average)
- Only consists of ~7900 gene families. Why?
- Flexible in their ability to adapt to new functions.
- Similar on the sequence level, but high diversity of substrate specificity
- Via gene duplications, mutations and a mix and match assembly of modular proteins.

Example of adaptability

- ABC Transporters
  - Substitution in periplasmic binding subunit
- Type I polyketide-synthase(PKSs)
  - Large modular proteins
  - Combination of different protein sub-domains to achieve different functions
  - Spreading in genomes through gene duplications and transfers

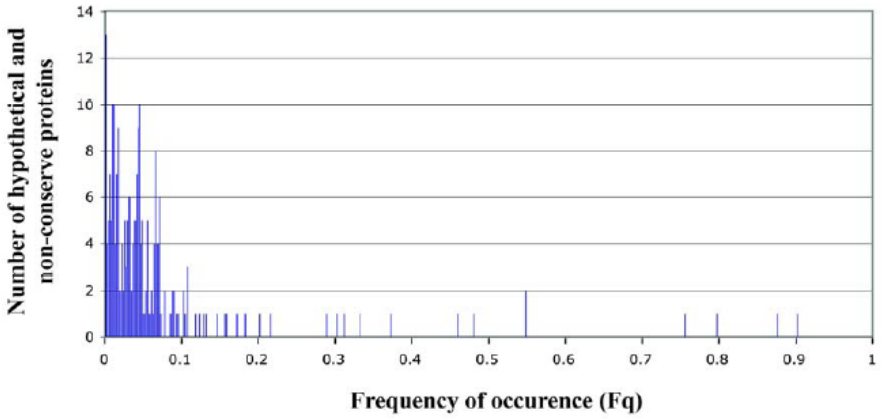
De novo creation of new proteins
The accessory pool

- Many of these genes are ORFans

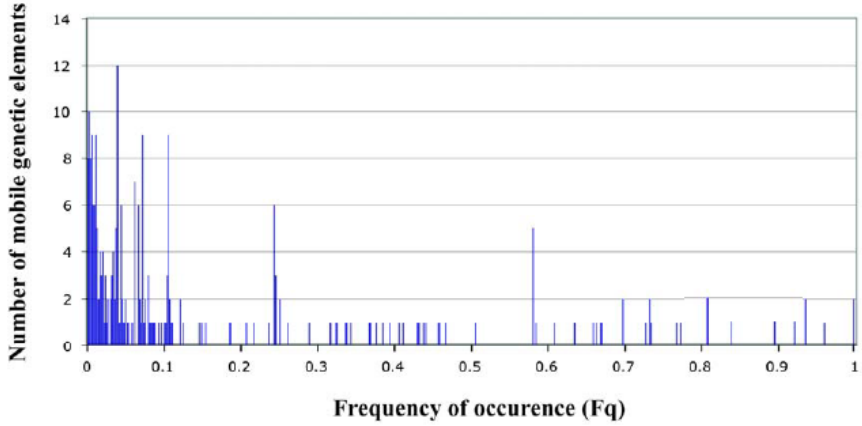
- A closer analysis of the accessory genes found in *E. coli* K12 → greater AT%, tend to be shorter, many composed of IS element and prophage sequences.

E. Coli K-12

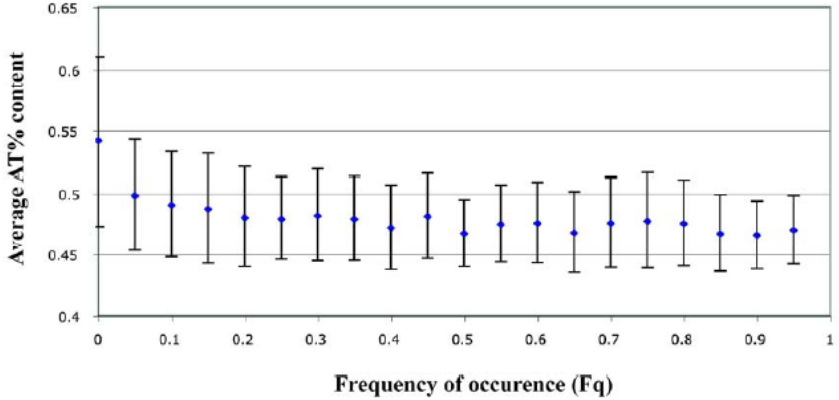
A.



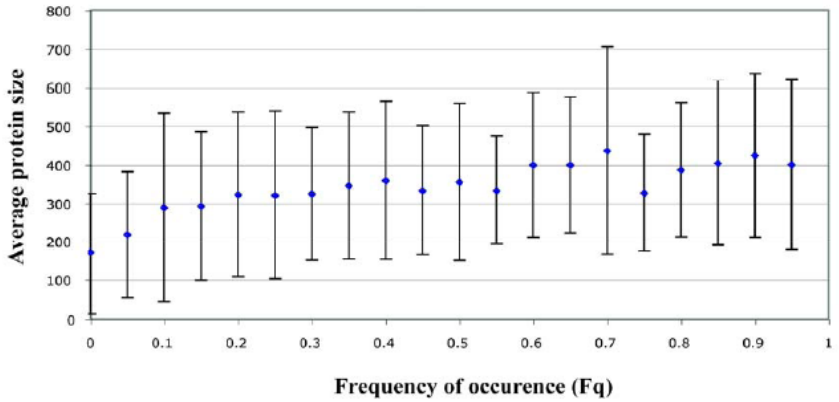
B.



A.



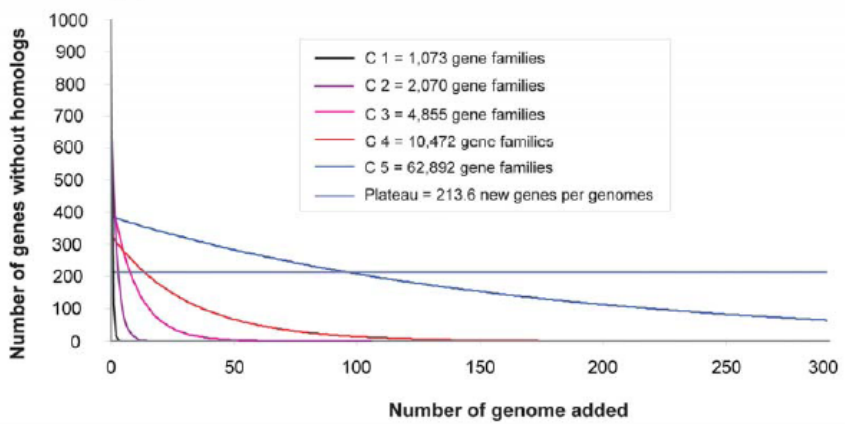
B.

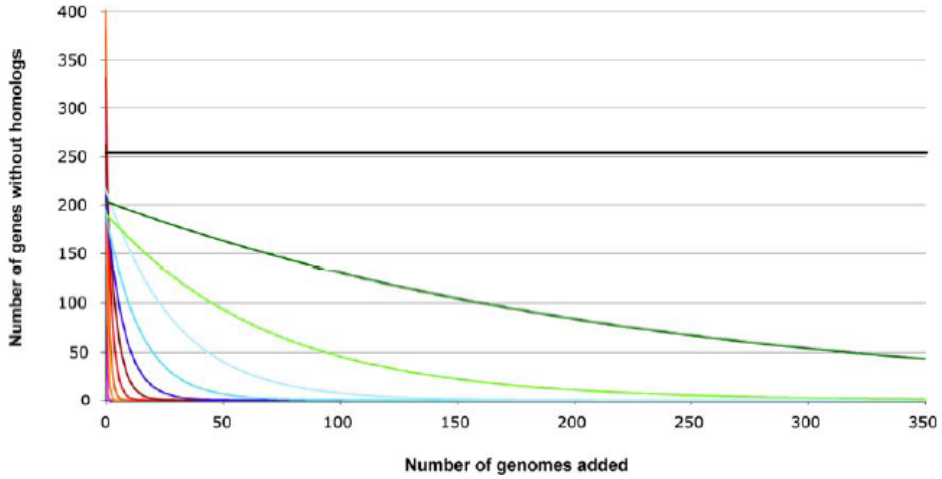


An Open Pan-Genome

- >139,000 rare gene families scattered in this study.

- Fitted exponential function approaches a plateau indicates an open pan-genome.





The accessory pool

- Doesn't seem tightly bound to a particular organismal lineage.
- Low level conservation might indicate processes that can create new proteins.
- Not subject to strong selective pressures and high turnover rates in genomes.
- Seems to represent an ongoing gene creation process different from domain shuffling.
- Annotation artifacts caused ORFs?
  - But the accessory pool is ~28% of the genome.

Conclusion

- Genomes function like collecting bins(+,-)
- Proteins in 3 categories are evolving under different constraints and rules.
  - The extended core
  - The character genes
  - The accessory pool
- How is protein space explored in biological evolution?
  - The result of selection and rearrangements of already existing proteins
  - Protein evolution is an ongoing process in which new proteins evolve as the exploration of the protein landscape continues