The infinitely many genes model for genomic diversity in bacteria

Peter Pfaffelhuber

joint with Franz Baumdicker, Wolfgang Hess

University of Freiburg

Gothenburg, August 2010
Genomic bacterial data

- **Observation**
  Not all bacteria of one population carry the same genes

- **Difference** in number of genes: up to 25%

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Individual 2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Individual 3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Individual 4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

✓ gene present
- gene absent
The pangenome

- **Pangenome**
  total set of genes of a population

- **Core genome**
  genes carried by all individuals (selective constraints?)

- Data from 12 *Prochlorococcus* strains (Kettler et al 2007)
The supragenome

- **Supragenome** = Gene frequency spectrum
- **Predicted** using a test dataset of 8 individuals
- Data from 13 *Haemophilus influenzae* strains (Hogg et al. 2007)
Modelling genomic diversity

- **Goal**: describe diversity of genes in a bacterial population

- **Genealogy**: given by Kingman’s coalescent
Phylogenetic trees based on gene content

Daniel H. Huson¹,* and Mike Steel²

- New genes taken from the environment at rate $\frac{\theta}{2}$
- Present genes lost at rate $\frac{\rho}{2}$
- A set of core genes must not be lost
The infinitely many genes model
The infinitely many genes model

**Gene gain** at rate $\frac{\theta}{2}$ along ancestral lines
The infinitely many genes model

Present genes lost at rate $\frac{\rho}{2}$

+ gene gain
○ gene loss
Data, Genealogies and Mutations

Is the genomic pattern compatible with neutral evolution?
Questions (on the dispensable genome)

- How many genes does a single individual carry?
- How many different genes are there in the sample?
- How many new genes are there in the $n$th individual?
- What does the gene frequency spectrum look like?
Questions (on the dispensable genome)

Let $G_i$ be the set of genes carried by individual $i$

- How many \textit{genes} does a single individual carry?
  What is $|G_i|$?

- How many \textit{different genes} are there in the sample?
  What is $|G|$ for $G = \bigcup_{i=1}^{n} G_i$?

- How many \textit{new genes} are there in the $n$th individual?
  What is $|G_n \setminus \left( \bigcup_{i=1}^{n-1} G_i \right)|$?

- What does the \textit{gene frequency spectrum} look like?
  What is $G_i := |\{u \in G : u \in G_i \text{ for exactly } k \text{ different } G_i\}|$?
Single individual

- $|G_i|$: number of genes in individual $i$

- Lemma
  \[ |G_i| \sim \text{Poi}\left( \frac{\theta}{\rho} \right) \]

- Reason:
  \( \frac{\theta}{2} dt \): average number of genes gained a time $t$ in the past
  \( e^{-\frac{\rho}{2} t} \): probability that a gene gained at time $t$ not lost

Summing up all $t$,

\[ \int_0^\infty \frac{\theta}{2} e^{-\frac{\rho}{2} t} dt = \frac{\theta}{\rho}. \]
Size of the pangenome

- \(|G|\): **number of genes** in a sample of size \(n\)

**Theorem**

\[
\mathbb{E}[|G|] = \theta \sum_{k=0}^{n-1} \frac{1}{k + \rho}
\]

**Corollary**

\[
\mathbb{E}[\text{new genes in } (n + 1)\text{st individual}] = \mathbb{E}\left[|G_n \setminus \left( \bigcup_{i=1}^{n-1} G_i \right)\right] = \frac{\theta}{n + \rho}.
\]
Size of the pangenome

- $\mathcal{T}$: coalescent
- $\frac{\theta}{2} dt$: average number of genes gained at $x \in \mathcal{T}$
- $p_T(x)$: probability that a gene gained at $x \in \mathcal{T}$ is not lost

$$
\mathbb{E}[|\mathcal{G}|] = \mathbb{E}\left[\frac{\theta}{2} \int_{\mathcal{T}} p_T(x) \, dx\right] \\
= \mathbb{E}\left[\frac{\theta}{2} \int_{\mathcal{T}} \mathbb{1}(\text{gene gained at } x \text{ not lost}) \, dx\right] \\
= \frac{\theta}{2} \mathbb{E}[\text{length of unlost lines in } \mathcal{T}] \\
= \frac{\theta}{2} \sum_{k=1}^{n} \frac{k}{\binom{n}{2}} + \frac{\rho}{2} k = \theta \sum_{k=1}^{n} \frac{1}{k - 1 + \rho} = \theta \sum_{k=0}^{n-1} \frac{1}{k + \rho}
$$
Size of the pangenome

$|G|$: number of genes in a sample of size $n$

**Theorem**

$$\nabla[|G|] = \theta \sum_{k=0}^{n-1} \frac{1}{\rho + i} - \theta^2 \left( \sum_{k=0}^{n-1} \frac{1}{\rho + i} \right)^2 + \frac{\theta^2}{4} g(n,0,0)$$

where $g(k_1,k_2,k_3)$ can be defined recursively. In particular,

$$\nabla_{n=2}[|G|] = \frac{\theta}{\rho(1 + \rho)} + \theta^2 \frac{1}{(1 + \rho)^2(1 + 2\rho)}$$

$$\nabla_{n=3}[|G|] = \frac{\theta}{\rho} + \frac{\theta}{1 + \rho} + \frac{\theta}{2 + \rho} + \theta^2 \frac{90 + 249\rho + 275\rho^2 + 145\rho^3 + 30\rho^4}{(1 + \rho)^2(2 + \rho)^2(1 + 2\rho)(3 + 2\rho)(6 + 5\rho)}$$
The gene frequency spectrum

- $G_i$: Number of genes present in $i$ individuals
- **Theorem** For the gene frequency spectrum
  
  $$E[G_i] = \frac{\theta}{i} \frac{n \cdots (n - i + 1)}{(n - 1 + \rho) \cdots (n - i + \rho)}$$

- **Corollary**
  
  $$E[G_n] = \frac{\theta}{\rho} \frac{(n - 1)!}{(n - 1 + \rho) \cdots (1 + \rho)}$$
The random core genome

- $L$: length of genealogy

\[
\mathbb{E}[\text{no of genes present in } n \text{ individuals}] = \mathbb{E} \left[ \mathbb{E}[\text{number of genes present in } n \text{ individuals} | L] \right]
\]
\[
= \frac{\theta}{\rho} \mathbb{P}[\text{genealogy of length } L \text{ not hit by a gene loss}]
\]
\[
= \frac{\theta}{\rho} \mathbb{E} \left[ e^{-\frac{\rho}{2}L} \right]
\]
\[
= \frac{\theta}{\rho} \frac{(n - 1)!}{\rho(n - 1 + \rho) \cdots (1 + \rho)}
\]
Incongruent pairs

- A pair of genes is **incongruent**, if

<table>
<thead>
<tr>
<th>Individual 1</th>
<th>Individual 2</th>
<th>Individual 3</th>
<th>Individual 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>✓</td>
<td>–</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>–</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

- **Theorem**

Let $P$ be the number of pairs of incongruent genes

$$
\mathbb{E}[P] = \frac{\theta^2 \rho}{4^2} \frac{18 + 117 \rho^2 + 203 \rho^3 + 105 \rho^4}{(1 + \rho^2)^2 (1 + 2 \rho^2)(1 + 4 \rho^2)(3 + 4 \rho^2)(3 + 5 \rho^2)(6 + 5 \rho^2)(6 + 7 \rho^2)}.
$$
Prochlorococcus

- **Tiny:** length $\sim 0.6\mu m$, Genome size 2Mbp
- Smallest known *photosynthetic* bacterium
- **Abundant:** $\sim 10^5$ cells per ml (in the ocean)
- **Structure:** by water depth
- **Recently discovered:** first isolated in 1988
Fit of model and data

Estimates

\[ \hat{\theta} = 1135.27, \quad \hat{\rho} = 1.94, \quad \text{number of core genes} = 1268. \]
Outlook

- All quantities of interest can be computed (different genes in the sample, incongruent pairs of genes, new genes in next individual,...)

- Biologically interesting: how many genes are out there?

- Current project: understand the effect of horizontal gene transfer