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%

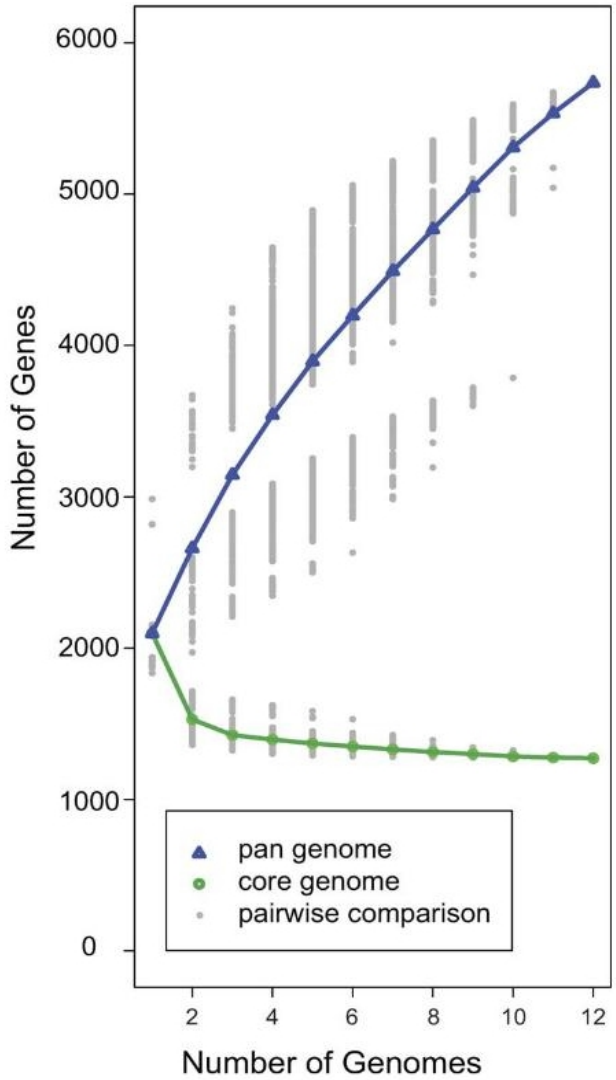
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Individual 1	–	–	–	–	✓	✓	–
Individual 2	✓	✓	✓	–	–	–	–
Individual 3	✓	✓	✓	✓	–	–	✓
Individual 4	✓	✓	✓	✓	–	–	✓

✓ gene present
– gene absent

The pangenome

- **Pangenome**
  total set of genes of a population
- **Core genome**
  genes carried by all individuals
  (selective constraints?)

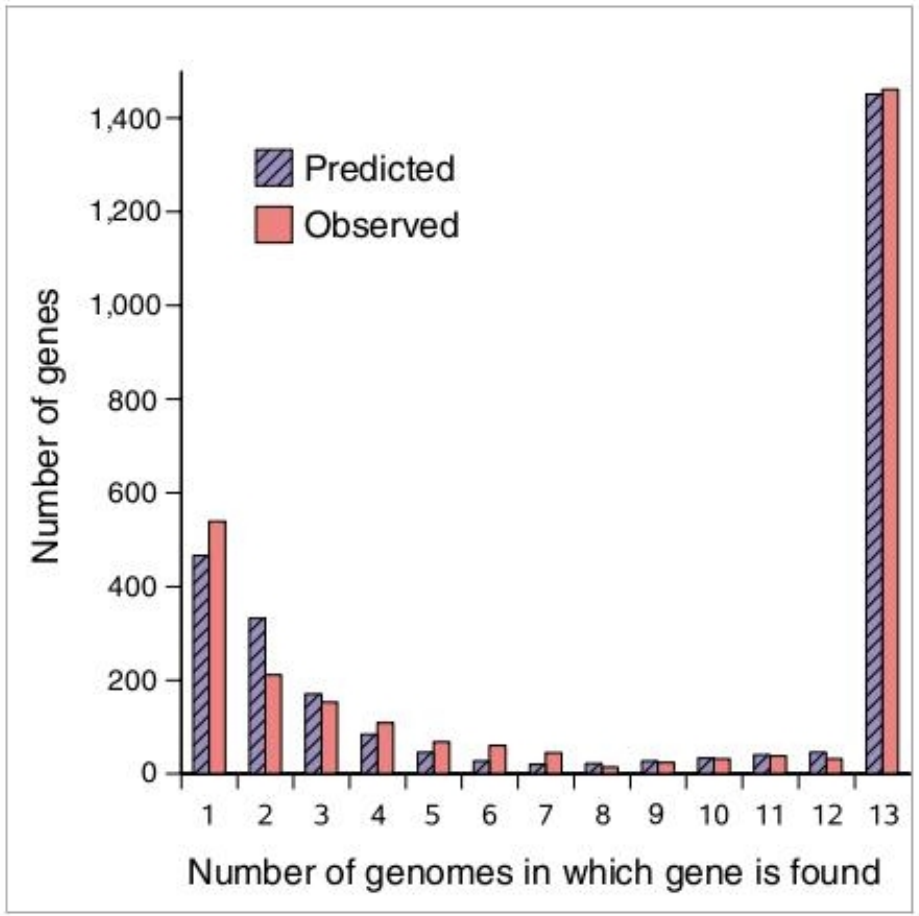
- Data from 12 *Prochlorococcus* strains (Kettler et al 2007)



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

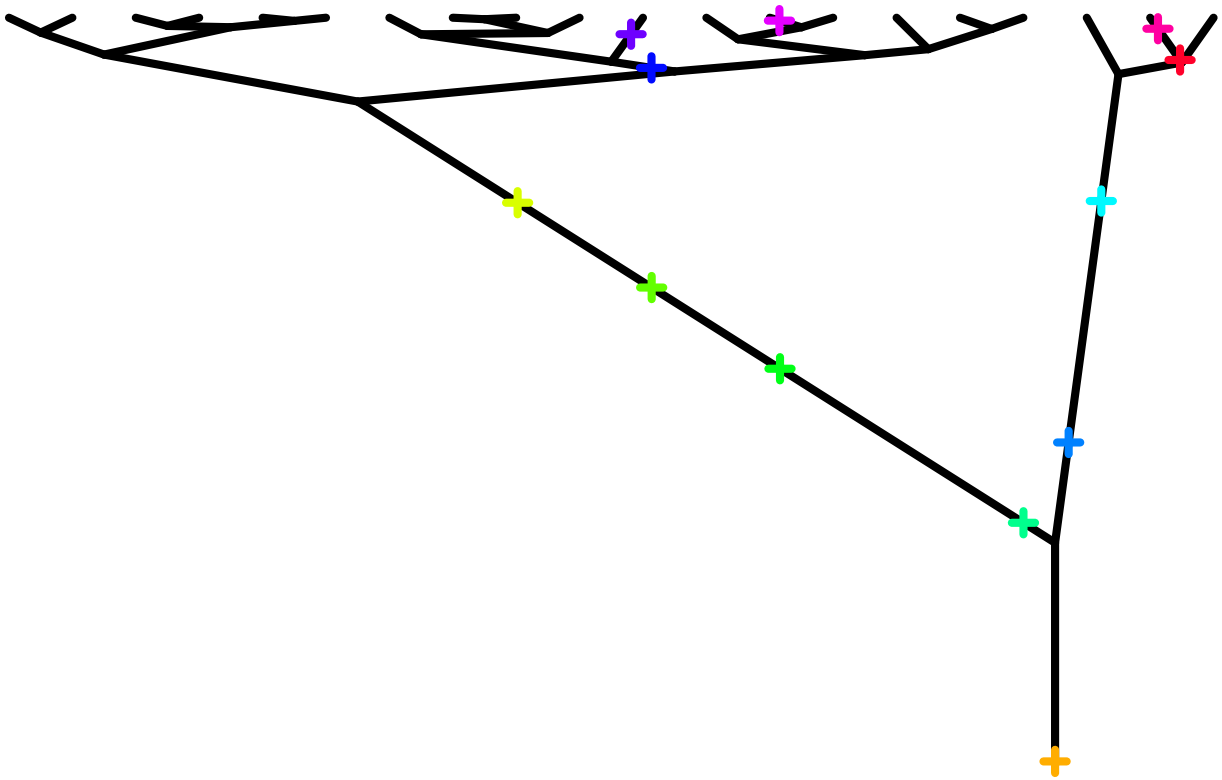
Daniel H. Huson[1,] and Mike Steel[2]*

- **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

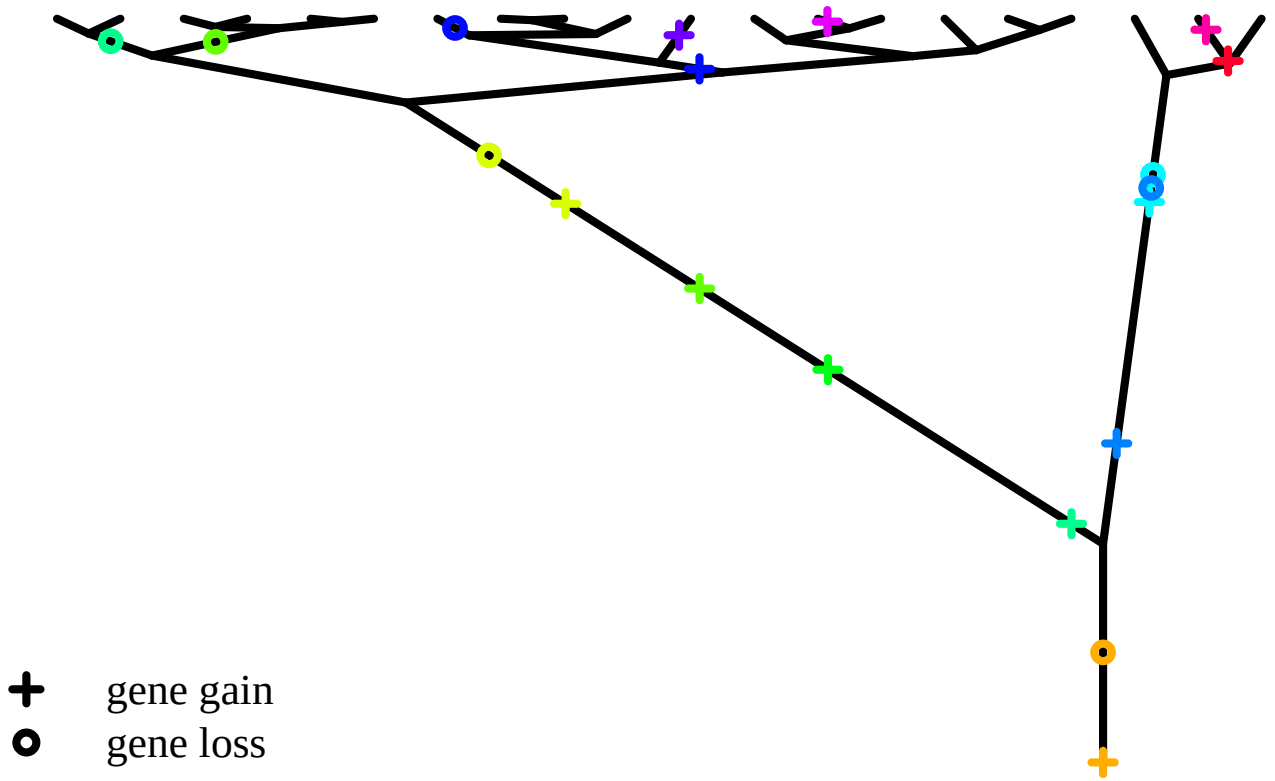
The infinitely many genes model

Gene gain at rate $\frac{\theta}{2}$ along ancestral lines



The infinitely many genes model

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

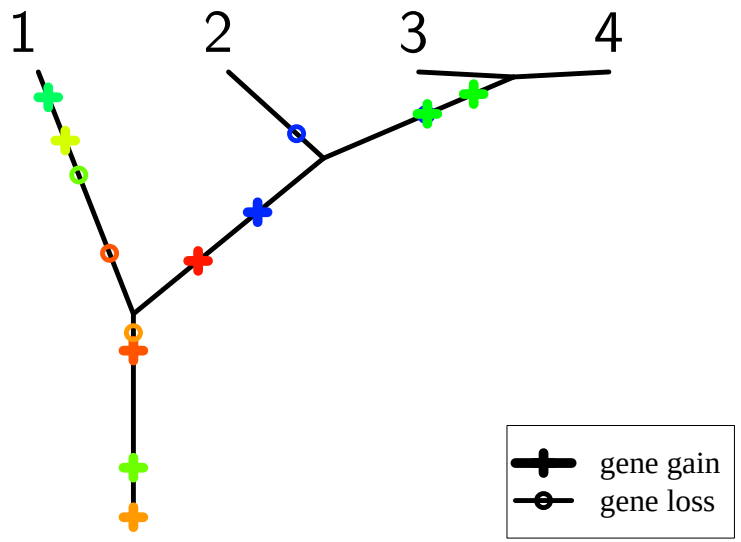


Data, Genealogies and Mutations

Data

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Individual 1	–	–	–	–	✓	✓	–
Individual 2	✓	✓	✓	–	–	–	–
Individual 3	✓	✓	✓	✓	–	–	✓
Individual 4	✓	✓	✓	✓	–	–	✓

Model



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 nth individual?
- What does the **gene frequency spectrum** look like?

Questions (on the dispensable genome)

Let $\mathcal{G}_i$ be the set of genes carried by individual i

- How many **genes** does a single individual carry?
  What is $|\mathcal{G}_i|$?
- How many **different genes** are there in the sample?
  What is $|\mathcal{G}|$ for $\mathcal{G} = \bigcup_{i=1}^{n} \mathcal{G}_i$?
- How many **new genes** are there in the nth individual?
  What is $\left|\mathcal{G}_n \setminus \left(\bigcup_{i=1}^{n-1} \mathcal{G}_i\right)\right|$
- What does the **gene frequency spectrum** look like?
  What is $G_i := |\{u \in \mathcal{G} : u \in \mathcal{G}_i \text{ for exactly } k \text{ different } \mathcal{G}_i\}|$?

Single individual

- $|\mathcal{G}_i|$: **number of genes** in individual i
- **Lemma**

$$|\mathcal{G}_i| \sim \text{Poi}\left(\frac{\theta}{\rho}\right)$$

- 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

- $|\mathcal{G}|$: **number of genes** in a sample of size n

Theorem

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

Corollary

$$\mathbb{E}[\textbf{new genes in } (n+1)\textbf{st individual}]$$
$$= \mathbb{E}\Big[\Big|\mathcal{G}_n \setminus \Big(\bigcup_{i=1}^{n-1} \mathcal{G}_i\Big)\Big|\Big] = \frac{\theta}{n+\rho}.$$

Size of the pangenome

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

$$\mathbb{E}[|\mathcal{G}|] = \mathbb{E}\left[\frac{\theta}{2}\int_{\mathcal{T}} p_{\mathcal{T}}(x)dx\right]$$

$$= \mathbb{E}\left[\frac{\theta}{2}\int_{\mathcal{T}} \mathbb{1}(\text{gene gained at } x \text{ not lost})dx\right]$$

$$= \frac{\theta}{2}\mathbb{E}\left[\text{length of unlost lines in } \mathcal{T}\right]$$

$$= \frac{\theta}{2}\sum_{k=1}^{n}\frac{k}{\binom{k}{2}+\frac{\rho}{2}k} = \theta\sum_{k=1}^{n}\frac{1}{k-1+\rho} = \theta\sum_{k=0}^{n-1}\frac{1}{k+\rho}$$

Size of the pangenome

- $|\mathcal{G}|$: **number of genes** in a sample of size n

Theorem

$$\mathbb{V}[|\mathcal{G}|] = \theta \sum_{k=0}^{n-1} \frac{1}{\rho + i} - \theta^2 \Big(\sum_{k=0}^{n-1} \frac{1}{\rho + i} \Big)^2 + \frac{\theta^2}{4} g_{(n,0,0)}$$

where $g_{(k_1,k_2,k_3)}$ can be defined recursively. In particular,

$$\mathbb{V}_{n=2}[|\mathcal{G}|] = \theta \frac{1+2\rho}{\rho(1+\rho)} + \theta^2 \frac{1}{(1+\rho)^2(1+2\rho)},$$

$$\mathbb{V}_{n=3}[|\mathcal{G}|] = \frac{\theta}{\rho} + \frac{\theta}{1+\rho} + \frac{\theta}{2+\rho} + \theta^2 \frac{90+249\rho+275\rho^2+145\rho^3+30\rho^4}{(1+\rho)^2(2+\rho)^2(1+2\rho)(3+2\rho)(6+5\rho)}$$

The gene frequency spectrum

- G_i: Number of genes present in i individuals
- **Theorem** For the **gene frequency spectrum**

$$\mathbf{E}[G_i] = \frac{\theta}{i} \frac{n \cdots (n-i+1)}{(n-1+\rho) \cdots (n-i+\rho)}$$

- **Corollary**

$$\mathbf{E}[G_n] = \frac{\theta}{\rho} \frac{(n-1)!}{(n-1+\rho) \cdots (1+\rho)}$$

The random core genome

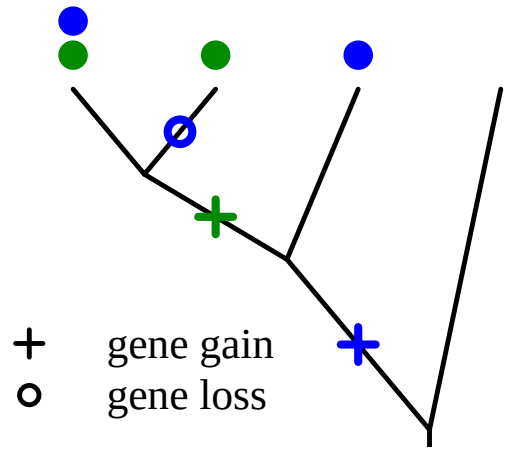
- L: length of genealogy

$$
\begin{aligned}
&\mathbf{E}[\textbf{no of genes present in } n \textbf{ individuals}] \\
&\quad = \mathbf{E}\big[\mathbf{E}[\text{number of genes present in } n \text{ individuals}|L]\big] \\
&\quad = \frac{\theta}{\rho}\mathbf{P}[\text{genealogy of length } L \text{ not hit by a gene loss}] \\
&\quad = \frac{\theta}{\rho}\mathbf{E}\big[e^{-\frac{\rho}{2}L}\big] \\
&\quad = \frac{\theta}{\rho}\frac{(n-1)!}{(n-1+\rho)\cdots(1+\rho)}
\end{aligned}
$$

Incongruent pairs

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

	Gene 1	Gene 2
Individual 1	✓	✓
Individual 2	✓	–
Individual 3	–	✓
Individual 4	–	–



- **Theorem**
  Let P be the number of pairs of incongruent genes

$$\mathbb{E}[P] = \frac{\theta^2\rho}{4}\,\frac{18+117\frac{\rho}{2}+203\frac{\rho^2}{4}+105\frac{\rho^3}{8}}{(1+\frac{\rho}{2})^2(1+2\frac{\rho}{2})(1+4\frac{\rho}{2})(3+4\frac{\rho}{2})(3+5\frac{\rho}{2})(6+5\frac{\rho}{2})(6+7\frac{\rho}{2})}.$$

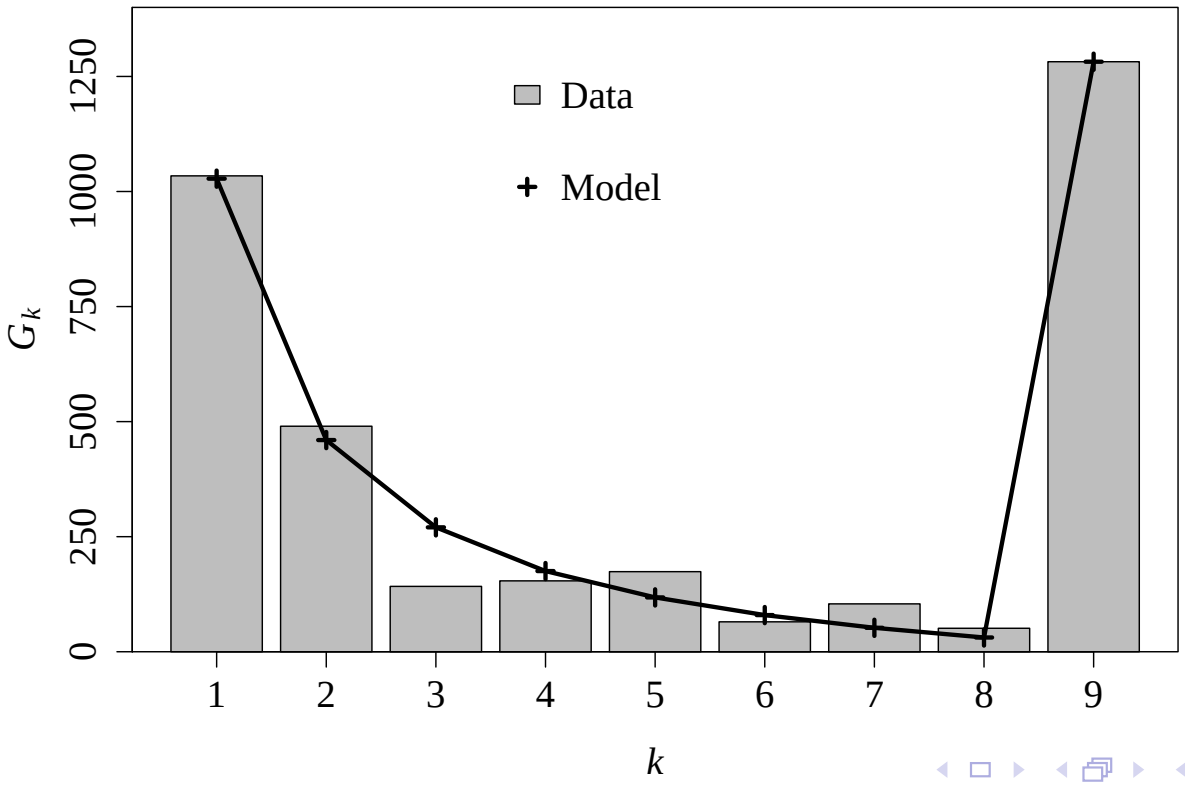
Prochlorococcus

- **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

$\widehat{\theta} = 1135.27, \qquad \widehat{\rho} = 1.94, \qquad$ 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**