The site frequency spectrum of dispensable genes

Franz Baumdicker

Albert-Ludwigs Universität Freiburg

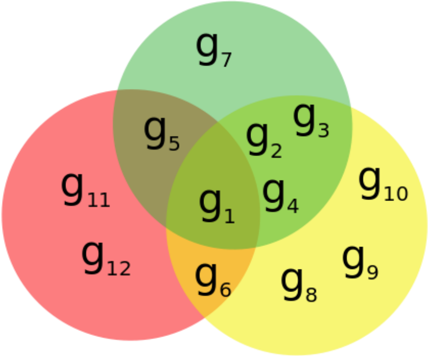
June 15, 2015

Introduction

The distributed genome hypothesis

The set of genes in a population of bacteria is distributed over all individuals.

- individuals of the same population do not have the same set of genes
- no organism contains the full complement of genes of the species
- genes can be gained and get lost again



data structure

- genomes are set of genes
- a gene is either present or absent in each of the genomes
- gene sequences of the same gene are typically not identical between genomes (SNPs)

	gene 1	gene 2	gene 3	gene 4	…	gene m
genom 1	`--A-T-`	✖	`-------T-`	✖		✖
genom 2	`--A---`	`--CT---`	`T--A-----`	✖		`--T`
genom 3	✖	✖	`T-----C--`	✖		`C--`
genom 4	✖	✖	`T--A-----`	`--A---A--`		`C--`
⋮	⋮	⋮	⋮	⋮		⋮
ind. n	`-----C`	✖	`T--A-----`	✖		✖

Tree-indexed Markov chain for gene gain and loss

- $I := [0,1]$ set of all possible genes, which might be gained
- $\mathcal{T}$ Kingman coalescent
- Define the Markov chain $(\mathcal{G}_t)_{t\in\mathcal{T}}$ with state space $\mathcal{N}_f([0,1])$, the space of finite counting measures on $[0,1] = I$.
- $\mathcal{G}_t$ makes transitions forwards in time

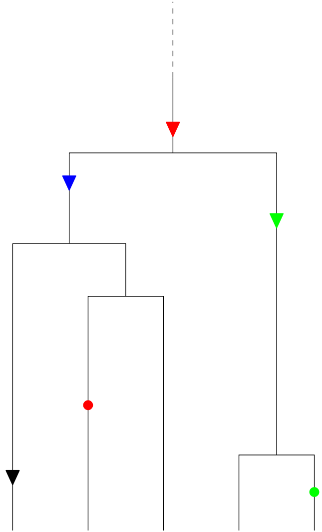
$$\begin{aligned} &\text{from } m \text{ to } m+\delta_u \qquad \text{at rate } \tfrac{\theta_1}{2}\lambda_I(du),\\ &\text{from } m \text{ to } m-\delta_u \qquad \text{at rate } \tfrac{\rho}{2}m(u) \end{aligned}$$

  along $\mathcal{T}$. λ_I is Lebesgue measure on I
- Denote the n leaves of the tree by $1,...,n \in \mathcal{T}$. $\mathcal{G}_1,...,\mathcal{G}_n$ describe the genes present in individuals $1,...,n$.

infinitely many genes model – time measured in N_e

- genealogy is given by Kingman's coalescent
- pairs of lineages coalesce at rate 1
- genes are gained (▼) at rate $\frac{\theta_1}{2}$
- each gene is lost (•) at rate $\frac{\rho}{2}$

	Gene 1	Gene 2	Gene 3	Gene 4
Genome 1	✓	✓	✓	✗
Genome 2	✓	✗	✗	✗
Genome 3	✓	✓	✗	✗
Genome 4	✗	✓	✗	✓
Genome 5	✗	✓	✗	✗



gene frequency spectrum

The *gene frequency spectrum* is given by $G_1, ..., G_n$, where

$$G_k := |\{u \in I : u \in \mathcal{G}_i \text{ for exactly } k \text{ different } i\}|.$$

G_k is the number of genes present in k of n individuals

We can calculate the expected gene frequency spectrum using Hoppe's urn model...

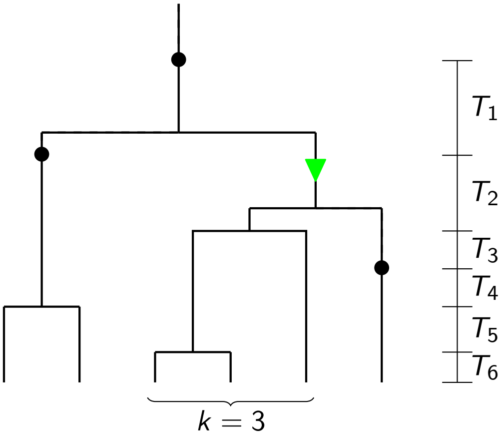
Hoppe's urn

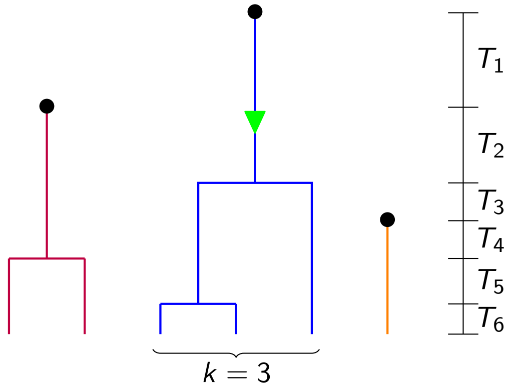
- Start with
  - one black ball with weight ρ and
  - one colored ball with weight 1.
- Draw a ball at random. If the ball is

• $\rightarrow$ •• black: put back the black ball and an additional ball in a new color.

• $\rightarrow$ •• black: put back the colored ball with an additional ball of the same color.

- continue until there are n colored balls in the urn


T_1
T_2
T_3
T_4
T_5
T_6
$k = 3$



Prob. for next event to be a merger/split:

backwards in time (coalescent)

$$\frac{\binom{i}{2}}{\binom{i}{2} + i\frac{\rho}{2}} = \frac{i-1}{i-1+\rho}$$

forwards in time (urn)

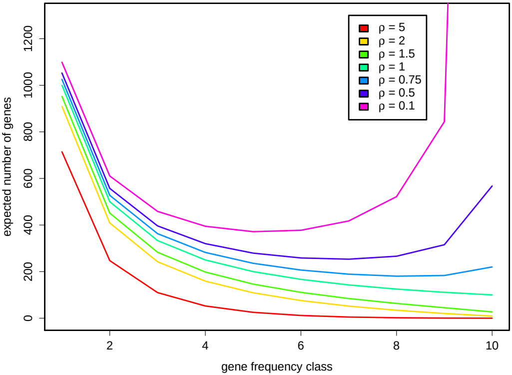
G_k: Number of genes present in k of n individuals

$$
\begin{aligned}
\mathbb{E}[G_k] &= \int_I \mathbb{E}[du \in \mathcal{G}_i \text{ for } k \text{ different } i] \\
&= \sum_{i=1}^{n} \sum_{l=1}^{i} \mathbb{P}[l\text{-th line during } T_i \text{ is of size } k] \\
&\qquad \cdot \int_I \mathbb{P}[\text{gene gain in } du \text{ on } l\text{-th line during } T_i] \\
&= \sum_{i=1}^{n} \sum_{l=1}^{i} \binom{n-i}{k-1} \frac{(k-1)!(i-1+\rho)\cdots(n-k-1+\rho)}{(i+\rho)\cdots(n-1+\rho)} \\
&\qquad \cdot \int_I \frac{\theta_1}{i(i-1+\rho)} du
\end{aligned}
$$

expected gene frequency spectrum

G_k: Number of genes present in k of n individuals

$$\mathbb{E}[G_k] = \frac{\theta_1}{k} \frac{n \cdots (n-k+1)}{(n-1+\rho)\cdots(n-k+\rho)}$$



$\theta = 1000$
$n = 10$

Based on the IMGM we can analyze microbial pangenomes:

- estimate θ_1 and ρ and test the hypothesis of neutral genome evolution based on the observed gene frequency spectrum
- estimate the number of different genes in the population
- forecast number of new genes found in sequencing projects

provide general insights:

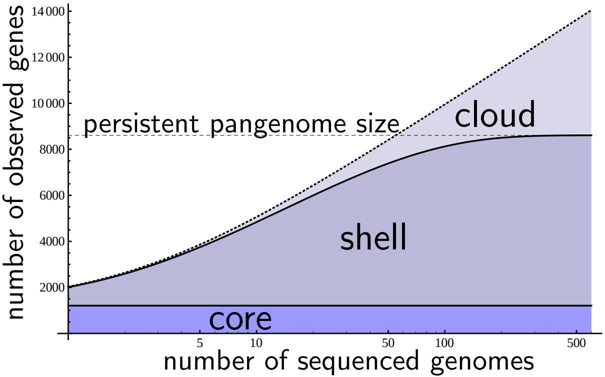
- The expected number of dispensable genes in freq > 0.01, can not exceed the $\sim$28 fold of the average single genome size:

  $\mathbb{E}[G^{0.01}] \leq 28.33 \cdot \mathbb{E}[G]$ (even for strong selection/HGT)
- $\mathbb{E}[G^{0.5}] \leq 1.0 \cdot \mathbb{E}[G]$ (only for neutral genes)
- the pangenome grows like $\theta_1 log(N)$ for large population sizes

"easily" account for additional features

- horizontal gene transfer $\rightarrow$ ancestral gene transfer graph
- **site mutations within the genes**

Diversity of the *Prochlorococcus*-pan-genome



- p-value (neutrality) ≈ 0.630
- pangenome: 57792 genes
- persistant genes: ~ 8500
- sequenced genomes: 41 (11)
- known genes: 9331 (5025)
- 52 genes in 42th genome

- What about the site mutations within the gene sequences?
- How does the site frequency spectrum for dispensable genes look like?

Tree-indexed Markov chain for gene gain, loss and site mutation

- $I := [0,1]$ set of all possible genes, which might be gained
- $J = (0,1]$ set of all sites, which might mutate
- $\mathcal{T}$ Kingman coalescent
- Define the Markov chain $(\mathcal{M}_t)_{t\in\mathcal{T}}$ with state space $\mathcal{N}_f([0,1]^2)$, the space of finite counting measures on $[0,1]^2 = I \times (\{0\} \cup J)$.
- $\mathcal{M}_t$ makes transitions forwards in time

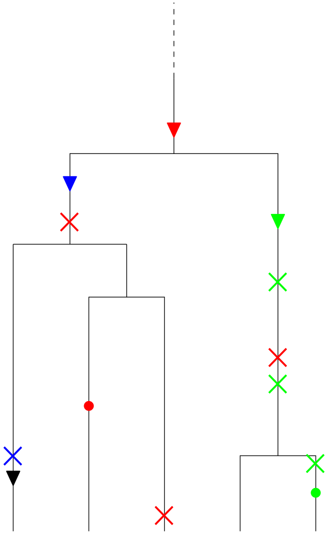
$$\begin{aligned} &\text{from } m \text{ to } m + \delta_{(u,0)} && \text{at rate } \tfrac{\theta_1}{2}\lambda_I(du), \\ &\text{from } m \text{ to } m - m|_{\{u\}\times I} && \text{at rate } \tfrac{\rho}{2}m(u,0), \text{ and} \\ &\text{from } m \text{ to } m + \delta_{(u,v)} && \text{at rate } \tfrac{\theta_2}{2}m(u,0)\lambda_I(dv) \end{aligned}$$

  along $\mathcal{T}$. λ_I is Lebesgue measure on I
- Denote the n leaves of the tree by $1, ..., n \in \mathcal{T}$. $\mathcal{M}_1, ..., \mathcal{M}_n$ describe the genes & site mutations present in individuals $1, ..., n$.

IMG model with site mutations – time measured in N_e

- along Kingman's coalescent
- pairs of lines merge at rate 1
- genes are gained (▼) at rate $\frac{\theta_1}{2}$
- each gene is lost (•) at rate $\frac{\rho}{2}$
- a present gene is hit by a site mutation ($\times$) at rate $\frac{\theta_2}{2}$

	Gene 1	Gene 2	Gene 3	Gene 4
Genome 1	--T--	-A---	----	✖
Genome 2	-----	✖	✖	✖
Genome 3	-----	-AA--	✖	✖
Genome 4	✖	T----	✖	---AC--
Genome 5	✖	T----	✖	✖



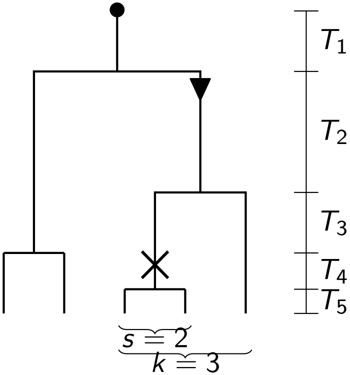
joint gene and site frequency spectrum

The *joint gene and site frequency spectrum* is given by $G_{1,1}, \dots, G_{1,n}, G_{2,1}, \dots, G_{2,n}, \dots, G_{n,n}$, where

$$G_{k,s} := \Big| \Big\{ (u,v) \in I \times I : u \in \mathcal{G}_i \text{ for exactly } k \text{ different } i\ , i_1, \dots, i_k, \\ \text{and } (u,v) \in \mathcal{M}_{i_j} \text{ for exactly } s \text{ different } i_j \text{ with } j \in \{1, ..., k\} \Big\} \Big|$$

$G_{k,s}$ is the number of SNPs present in s of k sequences, where the corresponding gene exists in k of n genomes.

We can calculate the expected joint gene and site frequency spectrum using Hoppe's urn model...



2nd line at T_2 is of size 3

$\mathcal{T}(2,3,2) = \{T_3, T_4\}$

A line during T_i is of size k if the ball belonging to this line produces exactly k offspring.

Let $\mathcal{T}(i,k,m)$ be the set of all T_j for $j \in \{i,\dots,n\}$ where m of j colored balls in the urn are marked by a gene gain.

$$\mathbb{E}[G_k] = \int_I \mathbb{E}[du \in \mathcal{G}_i \text{ for } k \text{ different } i]$$

$$\mathbb{E}[G_{k,s}] = \int_I \int_J \mathbb{E}\big[(du, 0) \in \mathcal{M}_i \text{ for exactly } k \text{ different } i \\ \text{and } (du, dv) \in \mathcal{M}_i \text{ for exactly } s \text{ different } i\big]$$

$$= \sum_{i=1}^{n} \sum_{l=1}^{i} \mathbb{P}[l\text{th line during } T_i \text{ is of size } k] \cdot \int_I \mathbb{P}[\text{mark in } du \text{ on } l\text{th line during } T_i]$$

$$\underbrace{\cdot \int_J \mathbb{E}\left[(du \times dv) \in \mathcal{M}_{i_j} \text{ for } s \text{ different } i_j \,\middle|\, \begin{array}{l} l\text{th line during } T_i \text{ is of size } k \\ \text{and has mark in } du \text{ during } T_i \end{array}\right]}_{(\star)}$$

$$(\star) = \sum_{m=1}^{k} \sum_{r=1}^{m} \mathbb{P}[r\text{th line during } \mathcal{T}(i,k,m) \text{ is of size } s] \\ \cdot \int_J \mathbb{P}[\text{ mutation in } dv \text{ on } r\text{th line during } \mathcal{T}(i,k,m)]$$

$$(\star) = \sum_{m=1}^{k}\sum_{r=1}^{m} \mathbb{P}[r\text{th line during } \mathcal{T}(i,k,m) \text{ is of size } s]$$
$$\cdot \int_J \mathbb{P}[\text{ mutation in } dv \text{ on } r\text{th line during } \mathcal{T}(i,k,m)]$$
$$= \sum_{m=1}^{k}\sum_{r=1}^{m} \binom{k-m}{s-1} \frac{(s-1)!(m-1)\cdots(k-s-1)}{(m)\cdots(k-1)}$$
$$\cdot \frac{\theta_2}{2} \sum_{j=i}^{n} \mathbb{P}[T_j \in \mathcal{T}(i,k,m)]\mathbb{E}[T_j]$$

expected joint gene and site frequency spectrum

$$\mathbb{E}[G_{k,s}] = \frac{\theta_1}{k} \frac{(n-k+1)\cdots n}{(n-k+\rho)\cdots(n-1+\rho)} \frac{\theta_2}{s} \frac{k}{n} \binom{n-1}{s}^{-1} \sum_{j=0}^{n-s-1} \frac{j+1}{j+1+\rho} \binom{n-j-2}{s-1}$$

- gene gain rate $\frac{\theta_1}{2}$
- gene loss rate $\frac{\rho}{2}$
- site mutation rate $\frac{\theta_2}{2}$

The site frequency spectrum of gene $u \in [0,1]$ is given by $S_1^u, \dots, S_{F(u)}^u$, where

$$S_s^u := |\{v \in J : v \in \mathcal{M}_i(u,.) \text{ for exactly } s \text{ different } i \text{ with } \mathcal{M}_i(u,0) = 1\}|$$

if $F(u) := |\{i \in \{1,\dots,n\} : \mathcal{M}_i(u,0) = 1\}|$ is frequency of gene u.

We are interested in

$$\mathbb{E}[S_s^u \,|F(u) = k]$$

the expected site frequency spectrum of dispensable genes present in k of n individuals

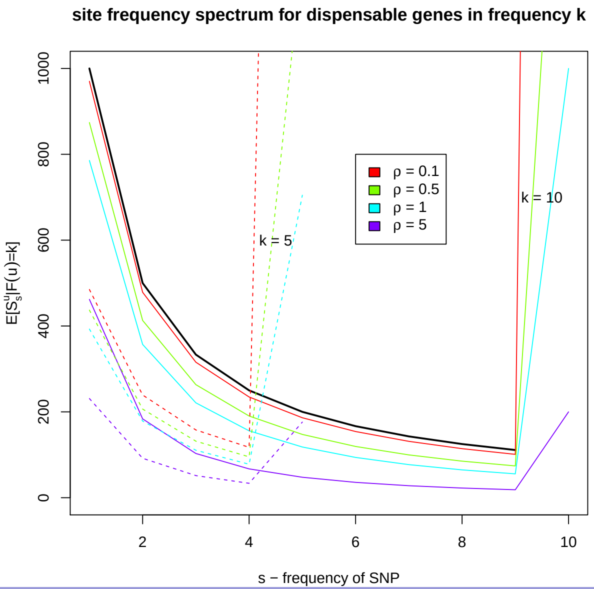
site frequency spectrum: classic vs. dispensable

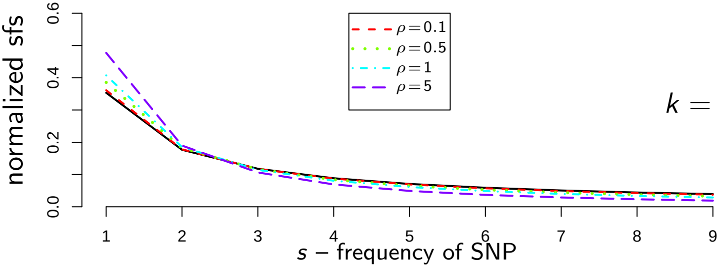
The site frequency spectrum in dispensable genes present in k out of n individuals is given for $s < k$ by

$$\mathbb{E}[S_s^u \,|F(u) = k] = \frac{\mathbb{E}[G_{k,s}]}{\mathbb{E}[G_k]}$$

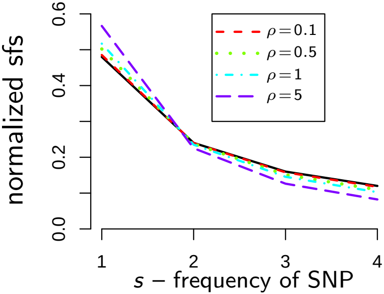
$$= \frac{\theta_2}{s}\frac{k}{n}\binom{n-1}{s}^{-1}\sum_{j=0}^{n-s-1}\frac{j+1}{j+1+\rho}\binom{n-j-2}{s-1}$$

$$\leq \frac{\theta_2}{s}\frac{k}{n}$$


site frequency spectrum for dispensable genes in frequency k
E[S^u_s|F(u)=k]
0
200
400
600
800
1000
2
4
6
8
10
s − frequency of SNP
ρ = 0.1
ρ = 0.5
ρ = 1
ρ = 5
k = 5
k = 10


normalized sfs
0.0
0.2
0.4
0.6
ρ = 0.1
ρ = 0.5
ρ = 1
ρ = 5
1
2
3
4
5
6
7
8
9
s − frequency of SNP

$k = 10$


normalized sfs
0.0
0.2
0.4
0.6
ρ = 0.1
ρ = 0.5
ρ = 1
ρ = 5
1
2
3
4
s − frequency of SNP

$k = 5$

estimators for the scaled site mutation rate θ_2

Watterson's estimator

$\widehat{\theta}_W := \frac{S_{\text{seg}}}{\sum_{s=1}^{k-1} \frac{1}{s}}$ $\qquad \mathbb{E}[S_{\text{seg}}] = \sum_{s=1}^{k-1} \mathbb{E}[C_s^u] = \sum_{s=1}^{k-1} \frac{\theta_2}{s}$

$\mathbb{E}[\widehat{\theta}_W] = \theta_2$

S_{seg}: total number of segregating sites in the sample of size k.

Tajimas estimator

$\widehat{\pi} := \sum_{i<j}^{n} \pi_{ij}$ $\qquad \mathbb{E}[\widehat{\pi}] = \sum_{s=1}^{k-1} \mathbb{E}[C_s^u] \frac{s(k-s)}{\binom{k}{2}} = \theta_2$

π_{ij} : number of sites which differ between individual i and individual j.

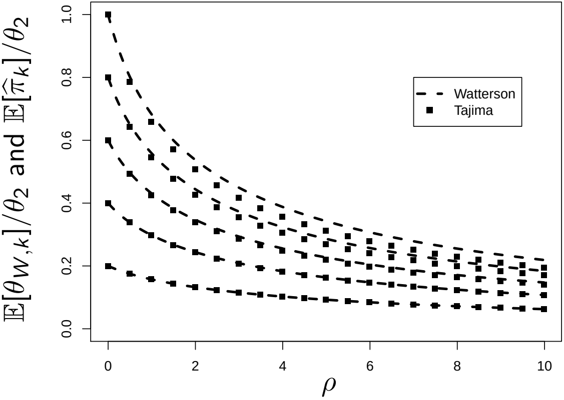
estimators for the site mutation rate of disp. genes

u dispensable gene, which appears in k out of n individuals

$$\mathbb{E}[S_{\text{seg}}] = \sum_{s=1}^{k-1} \mathbb{E}[S_s^u | F(u) = k]$$

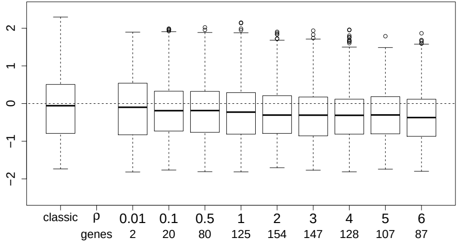
$$\mathbb{E}_{disp}[\widehat{\theta}_W] = \theta_2 \frac{\frac{k}{n} \sum_{s=1}^{k-1} \frac{1}{s} \binom{n-1}{s}^{-1} \sum_{j=0}^{n-s-1} \frac{j+1}{j+1+\rho} \binom{n-j-2}{s-1}}{\sum_{s=1}^{k-1} \frac{1}{s}} \leq \frac{k}{n}\theta_2$$

$$\begin{aligned} \mathbb{E}_{disp}[\widehat{\pi}] &= \sum_{s=1}^{k-1} \mathbb{E}[S_s^u | F(u) = k] \frac{s(k-s)}{\binom{k}{2}} \\ &= \theta_2 \frac{2}{k(k-1)} \frac{k}{n} \sum_{s=1}^{k-1} (k-s) \binom{n-1}{s}^{-1} \sum_{j=0}^{n-s-1} \frac{j+1}{j+1+\rho} \binom{n-j-2}{s-1} \\ &\leq \frac{k}{n}\theta_2 \end{aligned}$$

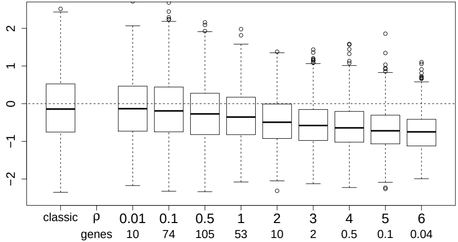


$\mathbb{E}[\widehat{\theta}_{W,k}]/\theta_2$ and $\mathbb{E}[\widehat{\pi}_k]/\theta_2$
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
ρ
Watterson
Tajima

Tajima's D for dispensable genes present in 8 of 20 individuals



Tajima's D for dispensable genes present in 19 of 20 individuals



conclusion

- bacterial genes can be gained and lost
- the site frequency spectrum of dispensable genes differs from the classical site frequency spectrum
- frequency spectra can be calculated using Hoppe's urn
- uncorrected standard estimates for the site mutation rate θ_2 are biased for a dispensable gene present in k of n genomes
- $\mathbb{E}[\widehat{\theta}] \leq \frac{k}{n}\theta_2, \quad \mathbb{E}[\widehat{\pi}] \leq \frac{k}{n}\theta_2$
- $\mathbb{E}[\widehat{\theta}] \neq \mathbb{E}[\widehat{\pi}]$ Tajima's D tends to be negative for disp. genes

Thank you for your attention

Publications

(a) Baumdicker, F., W. R. Hess, and P. Pfaffelhuber.
The diversity of a distributed genome in bacterial populations.
The Annals of Applied Probability (2010)

(b) Baumdicker, F., W. R. Hess, and P. Pfaffelhuber.
The infinitely many genes model for the distributed genome of bacteria.
Genome Biology and Evolution (2012)

(c) Baumdicker, F. and P. Pfaffelhuber.
The infinitely many genes model with horizontal gene transfer.
Electronic Journal of Probability (2014)

(d) **Baumdicker, F.**
The site frequency spectrum of dispensable genes.
Theoretical Population Biology (2015)