Introduction to Bayesian Methods - 1

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Course webpage:

http://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/Bayes.html

Conditional probabilities and Bayes' Theorem

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

Joint probability and
conditional probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem

Conditional probabilities and Bayes' Theorem

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

Joint probability and conditional probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem: a purely logical statement

$$P(H|D) = \frac{P(D|H)}{P(D)}P(H)$$

Bayes' theorem again: now as an inferential statement

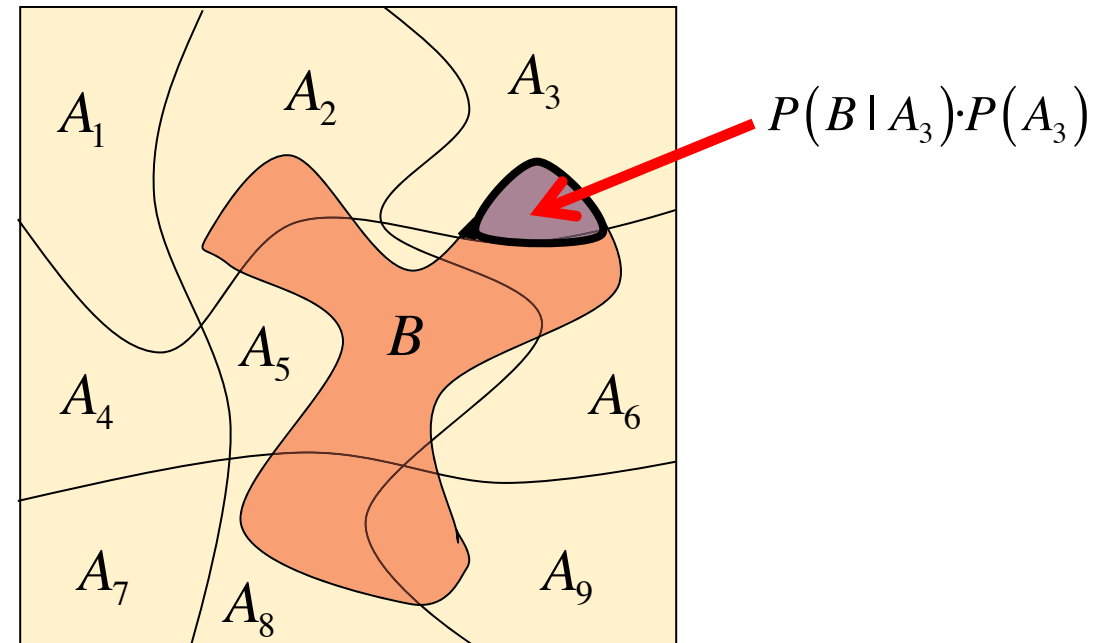$$P(H|D) = \frac{P(D|H)}{P(D)} P(H)$$

Likelihood

Prior distribution

Posterior distribution

Evidence

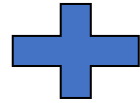$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P(A_k \mid B) = \frac{P(B \mid A_k) \cdot P(A_k)}{P(B)} \qquad k = 1, \dots, N$$

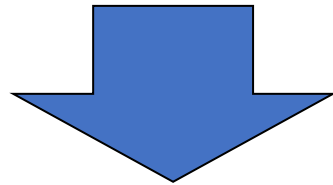if the events A_k are mutually exclusive, and they fill the universe

$$P(B) = \sum_{k=1}^{N} P(B \mid A_k) \cdot P(A_k)$$
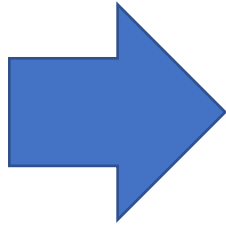

$P(B \mid A_3) \cdot P(A_3)$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P(B) = \sum_{k=1}^{N} P(B \mid A_k) \cdot P(A_k)$$

$$P(A_k \mid B) = \frac{P(B \mid A_k)}{\sum_{j=1}^{N} P(B \mid A_j) P(A_j)} P(A_k)$$

$$P(H_k|D) = \frac{P(D|H_k)}{\sum_j P(D|H_j)P(H_j)} P(H_k)$$

 MAP estimates

Example of MAP inference: the case of the Phoenix virus

Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements

Marie Dewannieux,[1,3] Francis Harper,[2,4] Aurélien Richaud,[1,4] Claire Letzelter,[1] David Ribet,[1] Gérard Pierron,[2] and Thierry Heidmann[1,5]

[1]Unité des Rétrovirus Endogènes et Eléments Rétroïdes des Eucaryotes Supérieurs, UMR 8122 CNRS, Institut Gustave Roussy, 94805 Villejuif Cedex, France; [2]Laboratoire de Réplication de l'ADN et Ultrastructure du Noyau, UPR1983 Institut André Lwoff, 94801 Villejuif Cedex, France

Human Endogenous Retroviruses are expected to be the remnants of ancestral infections of primates by active retroviruses that have thereafter been transmitted in a Mendelian fashion. Here, we derived in silico the sequence of the putative ancestral "progenitor" element of one of the most recently amplified family—the HERV-K family—and constructed it. This element, *Phoenix*, produces viral particles that disclose all of the structural and functional properties of a bona-fide retrovirus, can infect mammalian, including human, cells, and integrate with the exact signature of the presently found endogenous HERV-K progeny. We also show that this element amplifies via an extracellular pathway involving reinfection, at variance with the non-LTR-retrotransposons (LINEs, SINEs) or LTR-retrotransposons, thus recapitulating ex vivo the molecular events responsible for its dissemination in the host genomes. We also show that in vitro recombinations among present-day human *HERV–K* (also known as *ERVK*) loci can similarly generate functional HERV-K elements, indicating that human cells still have the potential to produce infectious retroviruses.

Amplification + genetic drift

Infection of primate ancestor germ cell

Present-day human genome
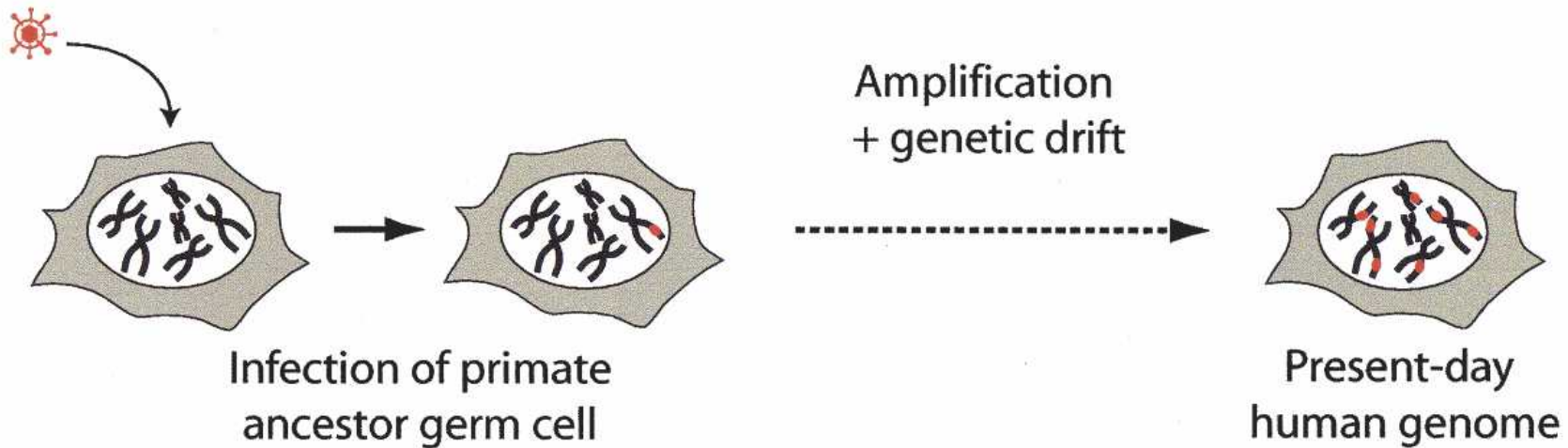
Phoenix, the ancestral HERV-K(HML2) retrovirus

To construct a consensus HERV-K(HML2) provirus, we assembled all of the complete copies of the 9.4-kb proviruses that are human specific (excluding those with the 292-nt deletion at the beginning of the _env_ gene) and aligned their nucleotide sequence to generate the consensus in silico, taking for each position the most frequent nucleotide. The resulting provirus sequence contains, as expected, ORFs for all of the HERV-K(HML2)-encoded proteins (Gag, Pro, Pol, Env, and the accessory Rec protein), with _gag_, _pro_, and _pol_ separated by -1 frameshifts. Noteworthily, this consensus provirus is distinct from each of the sequences used to generate it, with at least 20 amino acid changes on the overall sequences (Fig. 1).

* provirus = virus genome integrated into DNA of host cell

Notes:

- Nearly 8% of the human genome is composed of sequences of retroviral origin.
- HERV = Human Endogenous RetroVirus
- ORF = Open Reading Frame. The part of the reading frame that has the potential to be translated, a continuous stretch of codons that does not contain the stop codon.
- The three major proteins encoded within the retroviral genome: Gag, Pol, and Env:
  o Gag is a polyprotein and is an acronym for Group Antigens (ag).
  o Pol is the reverse transcriptase.
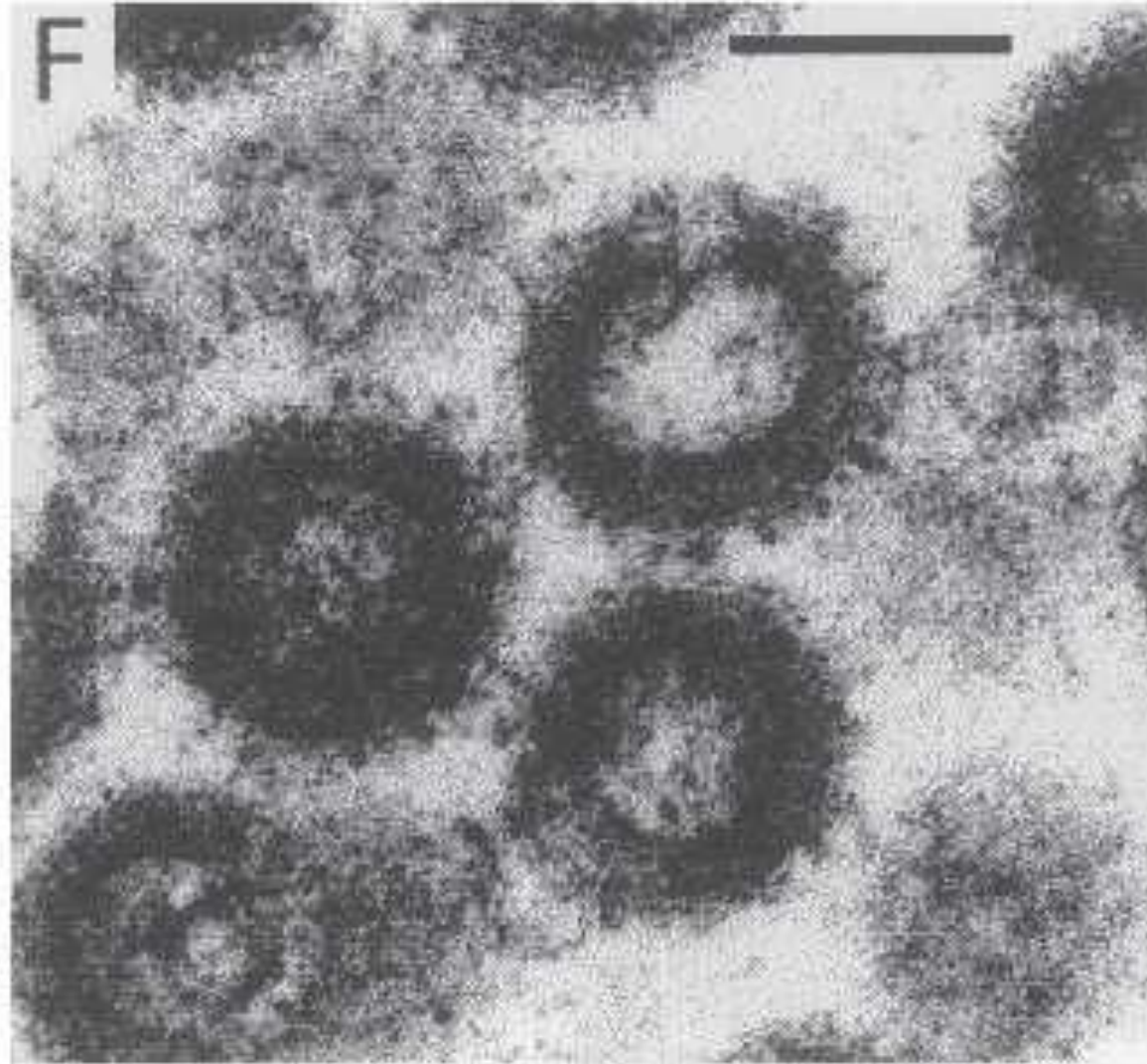 o Env in the envelope protein.

Image of representative particles obtained after transfection with an expression vector for the *Phoenix pro* mutant. Scale bar 100 nm.

A problem of male twins (Efron, 2003)

Pregnant with twins:
fraternal or identical?

Fraternal: 2/3 of all cases

Identical: 1/3 of all cases

What is the probability of identical twins IF both boys in sonogram?

Answer provided by Bayes theorem

$$P(\text{Identical}|\text{Both boys}) = \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})}P(\text{Identical})$$

$$P(\text{Identical}) = 1/3$$
$$P(\text{Fraternal}) = 2/3$$

Prior probabilities

$$P(\text{Both boys}|\text{Identical}) = 1/2$$
$$P(\text{Both boys}|\text{Fraternal}) = 1/4$$

Conditional probabilities from simple counting argument

$$P(\text{Both boys}) = P(\text{Both boys}|\text{Identical})P(\text{Identical})$$
$$+ P(\text{Both boys}|\text{Fraternal})P(\text{Fraternal})$$
$$= (1/2)(1/3) + (1/4)(2/3) = 1/3$$

$$P(\text{Identical}|\text{Both boys}) = \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})}P(\text{Identical})$$
$$= \frac{(1/2)}{(1/3)}(1/3) = 1/2$$

A simple application to medical tests (HIV testing)

$$P(\text{positive}|\text{infected}) = 1; \qquad P(\text{positive}|\text{not infected}) = 0.015$$

what is the probability *P(infected|positive)* ?

A common answer is 98.5% ... and it is wrong!

Let's use Bayes' theorem ...
$$P(A_k \mid B) = \frac{P(B \mid A_k) \cdot P(A_k)}{\displaystyle\sum_{k=1}^{N} P(B \mid A_k) \cdot P(A_k)}$$

$$P(\text{infected}|\text{positive}) = \frac{P(\text{positive}|\text{infected}) \times P(\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})}$$

$$= \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \right] \times P(\text{infected})$$

The estimate depends on the size of the infected population

i.e., on the probabilities

$$P(\text{infected}) \qquad\qquad P(\text{not infected})$$

$$P(\text{infected}|\text{positive}) = \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \right] \times P(\text{infected})$$

The posterior estimate strongly depends on prior probability

Example: AIDS testing

$$P(\text{infected}|\text{positive}) = \left[\frac{P(\text{positive}|\text{infected})}{P(\text{positive}|\text{infected}) \times P(\text{infected}) + P(\text{positive}|\text{not infected}) \times P(\text{not infected})} \right] \times P(\text{infected})$$
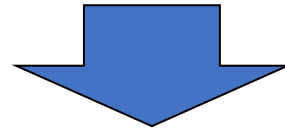
$$P_{\text{Italy}}(\text{infected}|\text{positive}) = \frac{1}{1 \times 0.003 + 0.015 \times 0.997} \times 0.003 \approx 16.7\%$$

$$P_{\text{South Africa}}(\text{infected}|\text{positive}) = \frac{1}{1 \times 0.173 + 0.015 \times 0.827} \times 0.173 \approx 93.3\%$$

The large number of false positives and the small probability of finding a sick person mean that the probability of being infected if positive is not actually very high.

Repeating measurements changes the reference population.

We incorporate a new positive result in a repeated measurement by using the previous posterior as the new prior:

$$P_{\text{Italy}}(\text{infected}|\text{positive}, \text{positive}) = \frac{1}{1 \times 0.167 + 0.015 \times 0.833} \times 0.167 \approx 93.0\%$$

$$P_{\text{South Africa}}(\text{infected}|\text{positive}, \text{positive}) = \frac{1}{1 \times 0.933 + 0.015 \times 0.067} \times 0.933 \approx 99.9\%$$

The first test changes the reference population, and the second test, if positive, gives a significant result.

Prosecutor's fallacy & Defendant's fallacy

Two common mistakes, associated with the wrong reference population

Consider a case where the probability of finding a given DNA subsequence – detected on a crime scene – is 0.00014 in the whole population: what is the probability that an individual who is found to have this rare subsequence in his/her DNA is guilty ???

A common answer is 1- 0.00014 = 0.99986 ... but this is WRONG !

$$P(innocent \,|\, DNA\, compatible)$$

this is what we actually want!

DNA classification - 1: alleles

allele: one of two or more alternative forms of the same gene, at the same position in a chromosome.
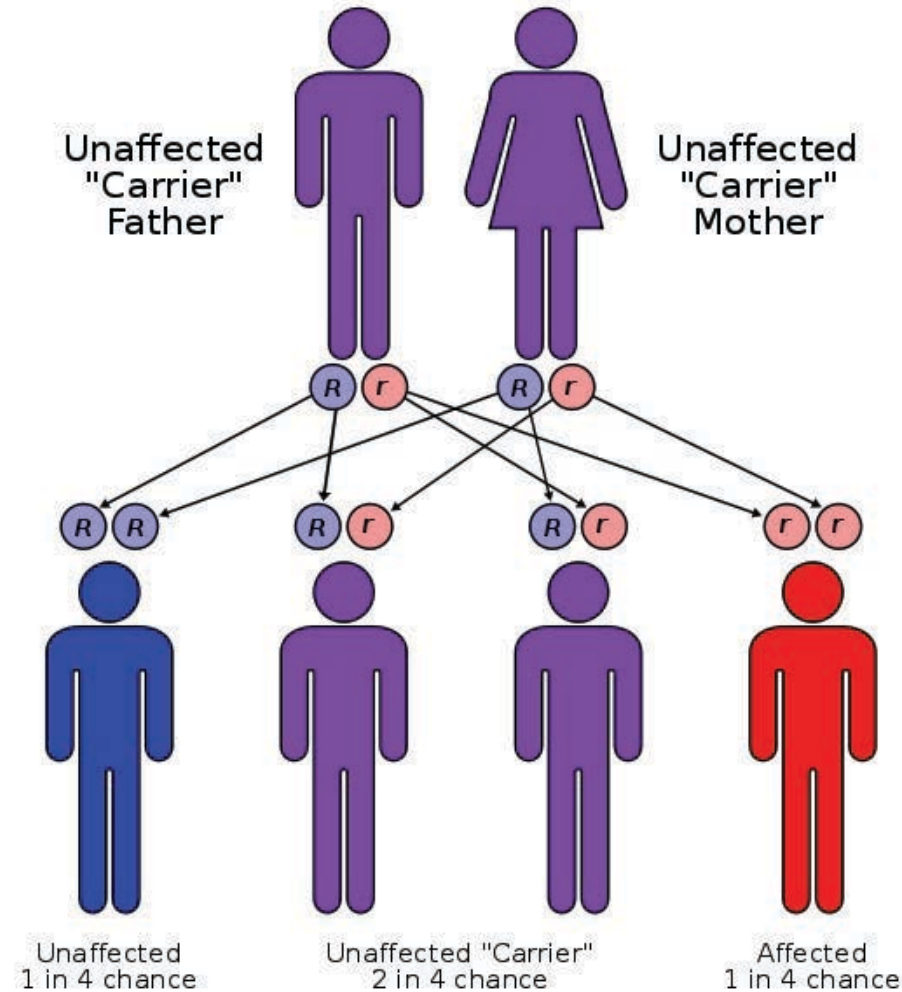
example: sickle cell anemia

A Normal red blood cells

Normal red blood cell (RBC)

RBCs flow freely within blood vessel

Cross-section of RBC

Normal hemoglobin
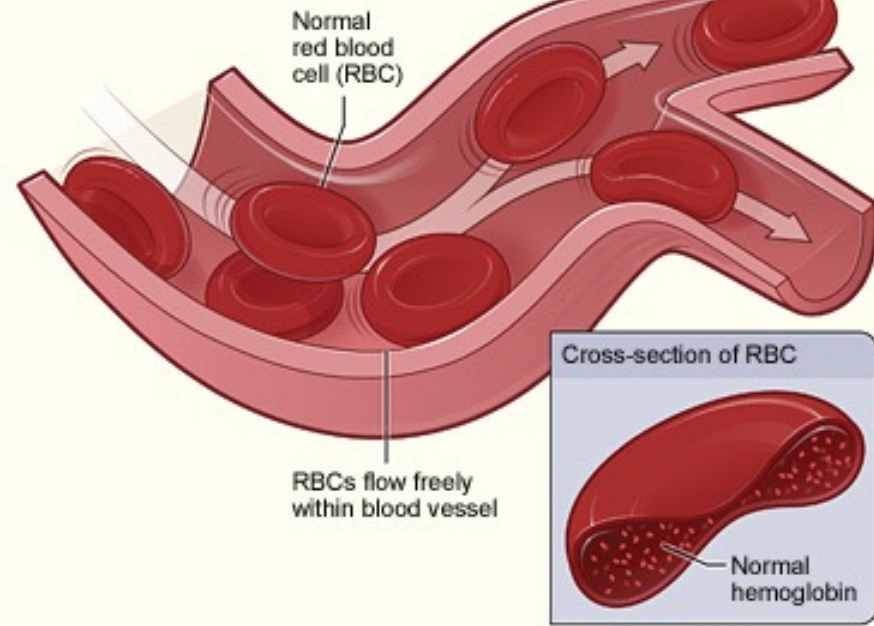
B Abnormal, sickled, red blood cells (sickle cells)

Sickle cells blocking blood flow

Sticky sickle cells

Cross-section of sickle cell

Abnormal hemoglobin form strands that cause sickle shape

b

HbS allele frequency (%)
- 0 – 0.51
- 0.52 – 2.02
- 2.03 – 4.04
- 4.05 – 6.06
- 6.07 – 8.08
- 8.09 – 9.60
- 9.61 – 11.11
- 11.12 – 12.63
- 12.64 – 14.65
- 14.66 – 18.18

c

Malaria endemicity
- Malaria free
- Epidemic
- Hypoendemic
- Mesoendemic
- Hyperendemic
- Holoendemic
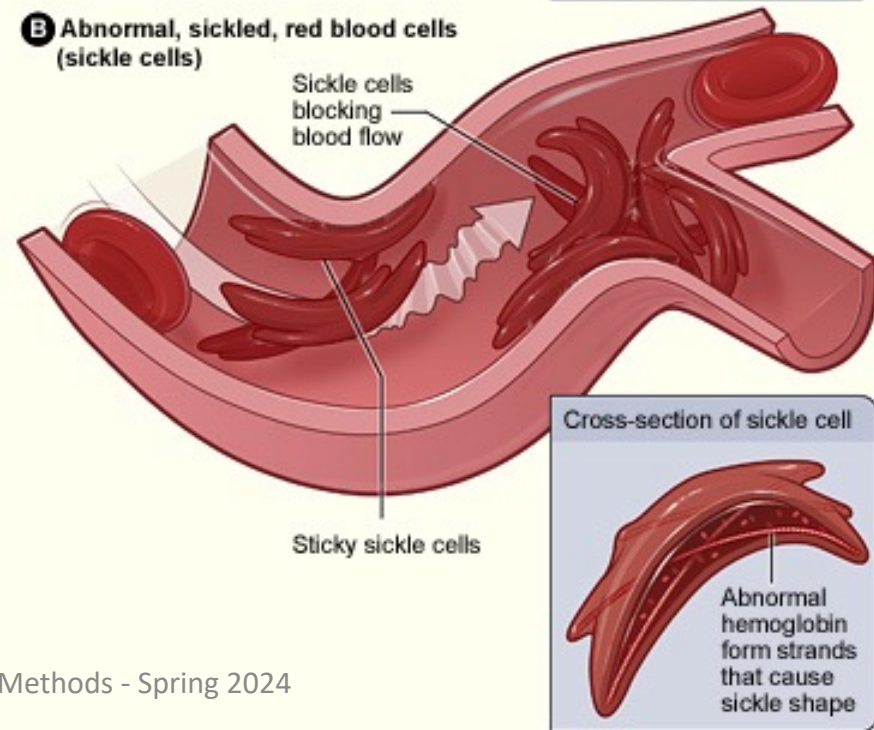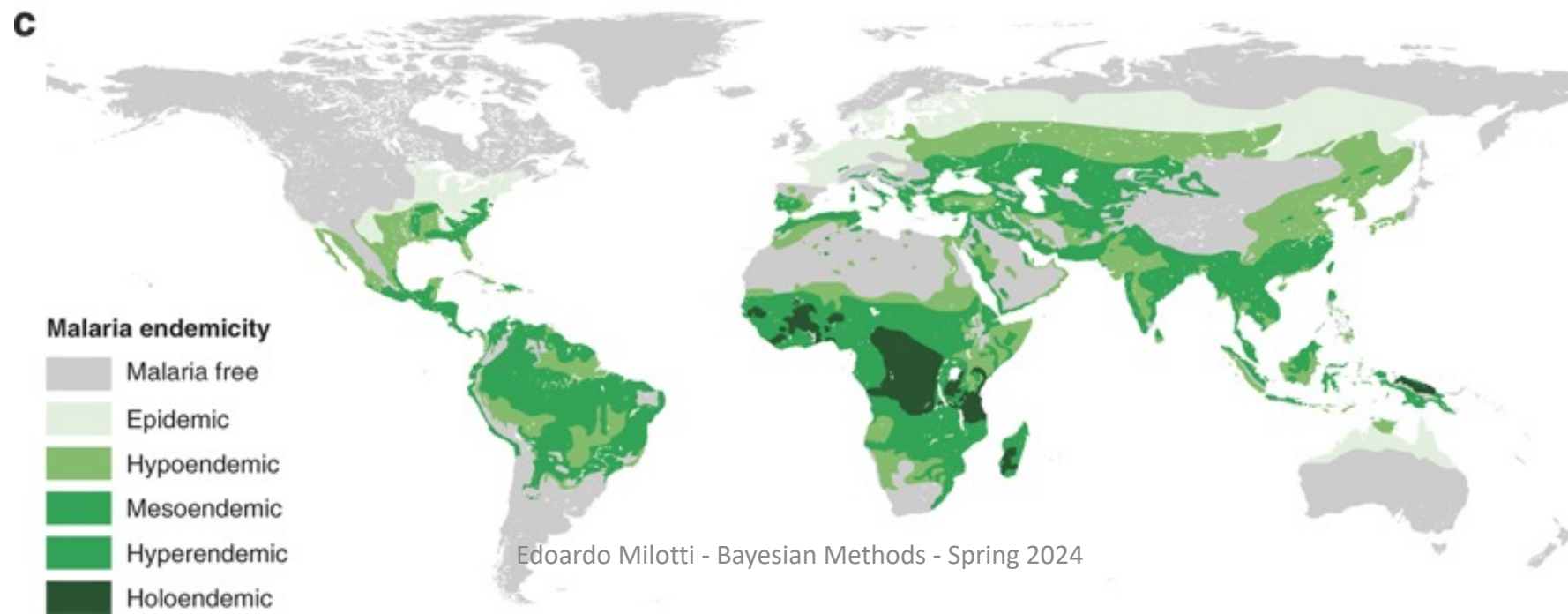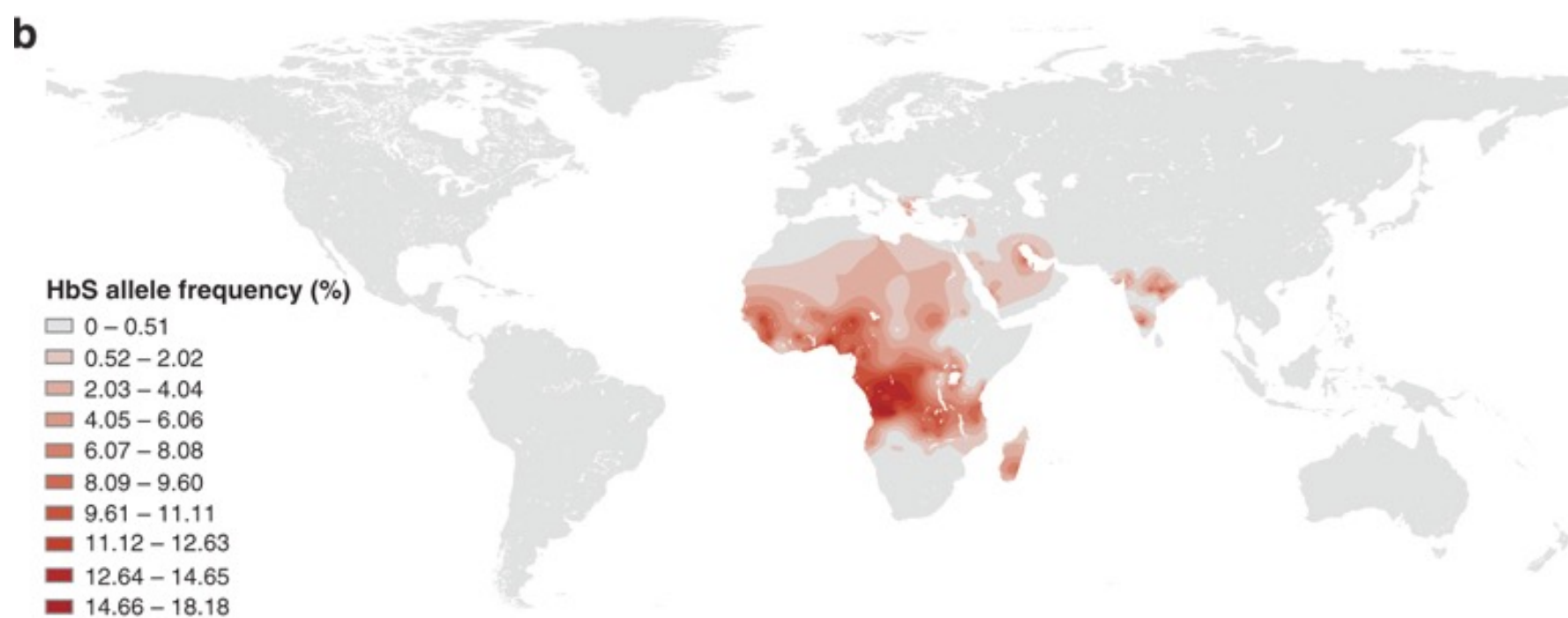
from F. B. Piel et al., Nature Comm. 1 (2010) 104

Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools

Faviel F. Gonzalez-Galarza[1,*,†], Antony McCabe[2,†], Eduardo J. Melo dos Santos[3], James Jones[4], Louise Takeshita[2], Nestor D. Ortega-Rivera[1], Glenda M. Del Cid-Pavon[1], Kerry Ramsbottom[4], Gurpreet Ghattaoraya[5], Ana Alfirevic[5], Derek Middleton[4,*] and Andrew R. Jones[4,*]

[1]Center for Biomedical Research, Faculty of Medicine, Autonomous University of Coahuila, Torreon, Mexico, [2]Computational Biology Facility, University of Liverpool, Biosciences building, Crown Street, Liverpool, L69 7ZB, UK, [3]Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Brazil, [4]Institute of Integrative Biology, University of Liverpool, Biosciences building, Crown Street, Liverpool, L69 7ZB, UK and [5]Department of Molecular and Clinical Pharmacology, Institute of Translational Medicine, University of Liverpool, Liverpool, UK

ABSTRACT

The Allele Frequency Net Database (AFND, www.allelefrequencies.net) provides the scientific community with a freely available repository for the storage of frequency data (alleles, genes, haplotypes and genotypes) related to human leukocyte antigens (HLA), killer-cell immunoglobulin-like receptors (KIR), major histocompatibility complex Class I chain related genes (MIC) and a number of cytokine gene polymorphisms in worldwide populations. In the last five years, AFND has become more popular in terms of clinical and scientific usage, with a recent increase in genotyping data as a necessary component of Short Population Report article submissions to another scientific journal. In addition, we have developed a user-friendly desktop application for HLA and KIR genotype/population data submissions. We have also focused on classification of existing and new data into 'gold–silver–bronze' criteria, allowing users to filter and query depending on their needs. Moreover, we have also continued to expand other features, for example focussed on HLA associations with adverse drug reactions. At present, AFND contains >1600 populations from >10 million healthy individuals, making AFND a valuable resource for the analysis of some of the most polymorphic regions in the human genome.
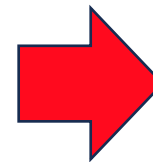
https://www.allelefrequencies.net/

Allele Frequency Net Database

Home | Populations | HLA | KIR | Other polymorphisms | HLA-ADR | KDDB

EUROSTAM

News

www.hlacovid19.org website

Our friends (who are always very supportive of the AFND website, Jill A. Hollenbach @UCSF, Steven J. Mack @UCSF and Martin Maiers @NMDP) have formed the COVID-19 HLA & Immunogenetics Consortium to unite the global community of HLA and immunogenetics experts and leaders in support of these efforts. They invite you to share your expertise as they launch this endeavour by joining the consortium. A project for the next international Histocompatibility Workshop is also planned.

Introduction

The **Allele Frequency Net Database** (AFND) provides the scientific community with a freely available repository for the storage of immune gene frequencies in different worldwide populations. Users can contribute the results of their work into one common database and can perform database searches on information already available. We have currently collected data in allele, haplotype and genotype format. However, the success of this website will depend on you to contribute your data.

Does any of this interest you?

- Would you like to publish your population frequency data on HLA and KIR? (In collaboration with Human immunology) See more information in *Hum Immunol*, 76, 393-394
- New search available for low resolution data when you cannot find the high resolution allele data you want

Login

You are logged as: guest
Click here to use your account

SUBMIT YOUR DATA

- Add HLA genotype raw data for SPR
- Add HLA frequency data
- Add non-HLA frequency data

Sponsors

- BAG Health Care
- Fujirebio Europe
- CareDx
- Omixon
- One Lambda
- Immucor
- GenDx

HLA = Human Leucocite Antigen. Human leukocyte antigens (HLA) are genes in major histocompatibility complexes (MHC) that help code for proteins that differentiate between self and non-self. They play a significant role in disease and immune defense.

Line	Allele	Population	% of individuals that have the allele	Allele Frequency (in_decimals)	Sample Size	IMGT/HLA[1] Database	Distribution[2]	Haplotype[3] Association	Notes[a]
1	C*01:02	Italy North		0.0200	277	See			
2	C*02:02	Italy North		0.0270	277	See			
3	C*03:02	Italy North		0.0070	277	See			
4	C*03:03	Italy North		0.0230	277	See			
5	C*03:04	Italy North		0.0110	277	See			
6	C*04:01	Italy North		0.1480	277	See			
7	C*05:01	Italy North		0.0450	277	See			
8	C*06:02	Italy North		0.0920	277	See			
9	C*07:01	Italy North		0.1440	277	See			
10	C*07:02	Italy North		0.0600	277	See			
11	C*07:04	Italy North		0.0130	277	See			
12	C*08:02	Italy North		0.0220	277	See			
13	C*12:02	Italy North		0.0470	277	See			
14	C*12:03	Italy North		0.1500	277	See			
15	C*14:02	Italy North		0.0490	277	See			
16	C*15:02	Italy North		0.0690	277	See			
17	C*16:01	Italy North		0.0430	277	See			
18	C*16:02	Italy North		0.0230	277	See			
19	C*16:04	Italy North		0.0020	277	See			
20	C*17:01	Italy North		0.0050	277	See			

$$P(innocent \mid DNA\,compatible, I) = \frac{P(DNA\,compatible \mid innocent, I)}{P(DNA\,compatible, I)} P(innocent \mid I)$$

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{P(\text{given allele sequence}|\text{innocent}, I)}{P(\text{given allele sequence}, I)} P(\text{innocent}|I)$$

where

$$
\begin{aligned}
P(\text{given allele sequence}, I) \;=\; & P(\text{given allele sequence}|\text{innocent}, I)P(\text{innocent}|I) \\
& + P(\text{given allele sequence}|\text{guilty}, I)P(\text{guilty}|I)
\end{aligned}
$$

Since the test has a very low error probability, i.e.,

$$P(\text{given allele sequence}|\text{guilty}, I) \approx 1$$

we find

$$P(\text{given allele sequence}, I) = 0.00014 \times P(\text{innocent}|I) + 1 \times P(\text{guilty}|I)$$

Once again, just like in the previous example, we see that it is all-important to determine the prior probabilities $P(\text{innocent}|I)$ and $P(\text{guilty}|I)$. For instance, if we pick a suspect at random in a large population, e.g., in a city with 1 million inhabitants, then

$$P(\text{innocent}|I) = 1 - 10^{-6} = 0.999999; \quad P(\text{guilty}|I) = 10^{-6} = 0.000001$$

$$P(\text{given allele sequence}, I) = 0.00014 \times (1 - 10^{-6}) + 1 \times 10^{-6} \approx 0.000141$$

and finally

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.000141}(1 - 10^{-6}) \approx 0.992907$$

This last result shows that the DNA test is quite inconclusive in this case, because it decreases the probability that the suspect is innocent from 0.999999 to 0.992907, only. How can it be? The reason is that in this case the number of random matches is not small, indeed in this city there are on average $1000000/7000 \approx 143$ people that randomly match the given allele sequence.

The argument can be turned upside down by a cunning lawyer, who might claim that since there are so many random matches, the DNA test is not relevant. However it is not so, and this claim is the "defendant's fallacy". Indeed, the problem that we met above was that the starting population was far too large. Other evidence might considerably reduce the number of possible suspects, for instance a surveillance camera might help identify all the people who entered a building and who had a chance to commit the crime, and thus reduce the starting population to, say, 100 people. When we repeat the relevant calculations, we find

$$P(\text{innocent}|I) = 1 - 1/100 = 0.99; \quad P(\text{guilty}|I) = 1/100 = 0.01$$

$$P(\text{given allele sequence}, I) = 0.00014 \times 0.99 + 1 \times 0.01 \approx 0.01014$$

and finally

$$P(\text{innocent}|\text{given allele sequence}, I) = \frac{0.00014}{0.01014}(1 - 10^{-2}) \approx 0.0137$$

We see that the new situation is drastically different, the reason being that on average only $100/7000 \approx 0.0143$ people can randomly match the given allele sequence.

An extremely short history of early Bayesianism

- Rev. Thomas Bayes discovered an early form of Bayes' theorem (second half of 18th century)

- Price discovered the theorem inside Bayes' unpublished notes (end 18th century)

- Laplace reinvented a version of the theorem and later expanded it after studying the Bayes' notes (around 1800)

- Laplace successfully applied the theorem to many experimental data analysis problems (until about 1820)

- Laplace was sometimes ridiculed by people who did not understand some of his approaches

- Laplace discovered the basic version of the Central Limit Theorem and in his later life he abandoned the Bayes theorem in favor of frequency-based methods (until about 1830)

- After the death of Laplace, Bayes' theorem was nearly forgotten and cornered to the darkest parts of statistics (crossing the desert …)

Bayesian inference

$$P\left(A_k \mid B\right) = \frac{P\left(B \mid A_k\right) \cdot P\left(A_k\right)}{\displaystyle\sum_{k=1}^{N} P\left(B \mid A_k\right) \cdot P\left(A_k\right)}$$

$$= \frac{P\left(B \mid A_k\right)}{\displaystyle\sum_{k=1}^{N} P\left(B \mid A_k\right) \cdot P\left(A_k\right)} \cdot P\left(A_k\right)$$

$$P\left(H_k \mid D, I\right) = \frac{P\left(D \mid H_k, I\right)}{\displaystyle\sum_{k=1}^{N} P\left(D \mid H_k, I\right) \cdot P\left(H_k \mid I\right)} \cdot P\left(H_k \mid I\right)$$

$$P\left(H_k \mid D, I\right) = \frac{P\left(D \mid H_k, I\right)}{\displaystyle\sum_{k=1}^{N} P\left(D \mid H_k, I\right) \cdot P\left(H_k \mid I\right)} \cdot P\left(H_k \mid I\right)$$

(Posterior probability that *k*-th hypothesis is true, when we observe data D, with prior information I)

=

(Probability of observing data D, given the *k*-th hypothesis) / Normalization

.

(Prior probability that *k*-th hypothesis is true)

$$P(H_k \mid D, I) = \frac{P(D \mid H_k, I)}{P(D \mid I)} \cdot P(H_k \mid I)$$

$$= \frac{P(D \mid H_k, I)}{\sum_{k=1}^{N} P(D \mid H_k, I) \cdot P(H_k \mid I)} \cdot P(H_k \mid I)$$

prior distribution $P(H_k, I)$

posterior distribution $P(H_k \mid D, I)$

likelihood or *sampling distribution* $P(D \mid H_k, I)$

evidence (normalizing factor) $P(D \mid I) = \sum_{k=1}^{N} P(D \mid H_k, I) \cdot P(H_k \mid I)$

Testing hypotheses

$$P(H_k \mid D, I) = \frac{P(D \mid H_k, I)}{P(D \mid I)} \cdot P(H_k \mid I)$$

Odds ratio

$$\frac{P(H_k \mid D, I)}{P(H_n \mid D, I)} = \left(\frac{P(D \mid H_k, I)}{P(D \mid H_n, I)} \right) \cdot \left(\frac{P(H_k \mid I)}{P(H_n \mid I)} \right)$$

Bayes' factor

When prior probabilities are the same (equally probable hypotheses), the posterior probability ratio depends only on the Bayes' factor:

Odds ratio

$$\frac{P\left(H_k|D,I\right)}{P\left(H_n|D,I\right)} = \frac{P\left(D|H_k,I\right)}{P\left(D|H_n,I\right)} \times \frac{P\left(H_k|I\right)}{P\left(H_n|I\right)} = \frac{P\left(D|H_k,I\right)}{P\left(D|H_n,I\right)}$$

Bayes' factor

Bayes' factor

Uniform priors

From discrete sets of hypothesis to the continuum.
The Bayes' theorem in the context of parameter estimation.

$$P\left(H_k \mid D, I\right) = \frac{P\left(D \mid H_k, I\right)}{P\left(D \mid I\right)} \cdot P\left(H_k \mid I\right) = \frac{P\left(D \mid H_k, I\right)}{\sum_{k=1}^{N} P\left(D \mid H_k, I\right) \cdot P\left(H_k \mid I\right)} \cdot P\left(H_k \mid I\right)$$

$$p\left(\boldsymbol{\theta} \mid D, I\right) = \frac{P\left(D \mid \boldsymbol{\theta}, I\right)}{\int_{\Theta} P\left(D \mid \boldsymbol{\theta}', I\right) p\left(\boldsymbol{\theta}' \mid I\right) d\boldsymbol{\theta}'} \times p\left(\boldsymbol{\theta} \mid I\right)$$

What if we "measure" a mathematical constant instead of a physical parameter?
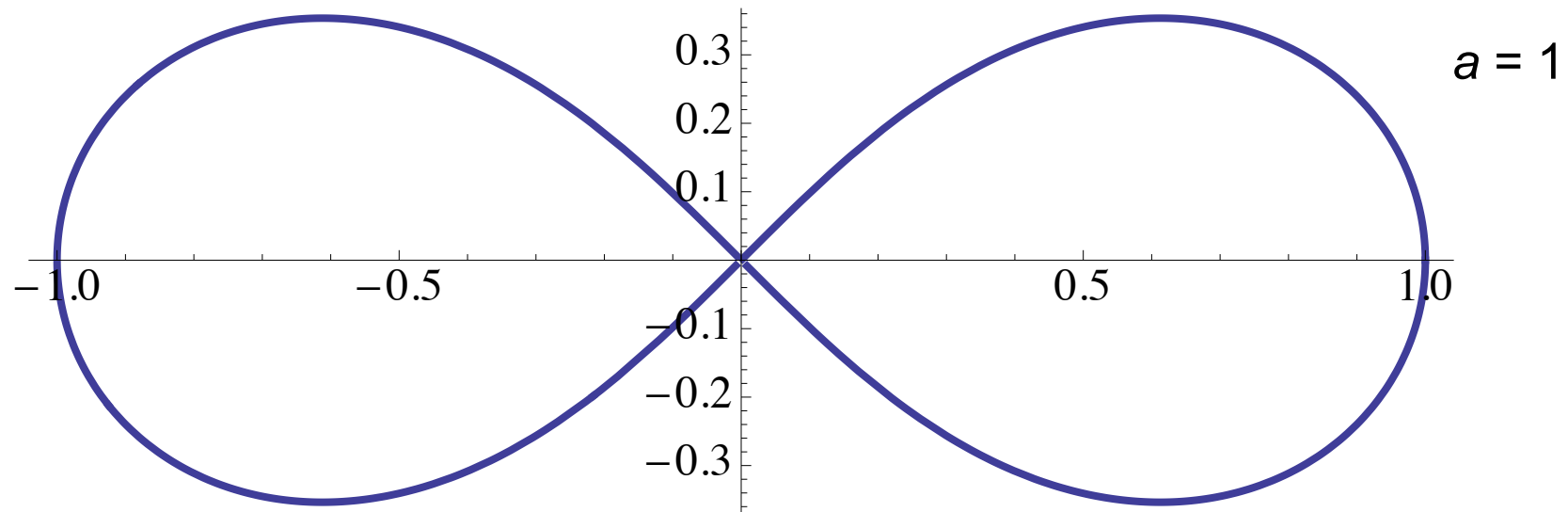
Example:

area of Bernoulli's lemniscate obtained with a Monte Carlo simulation.
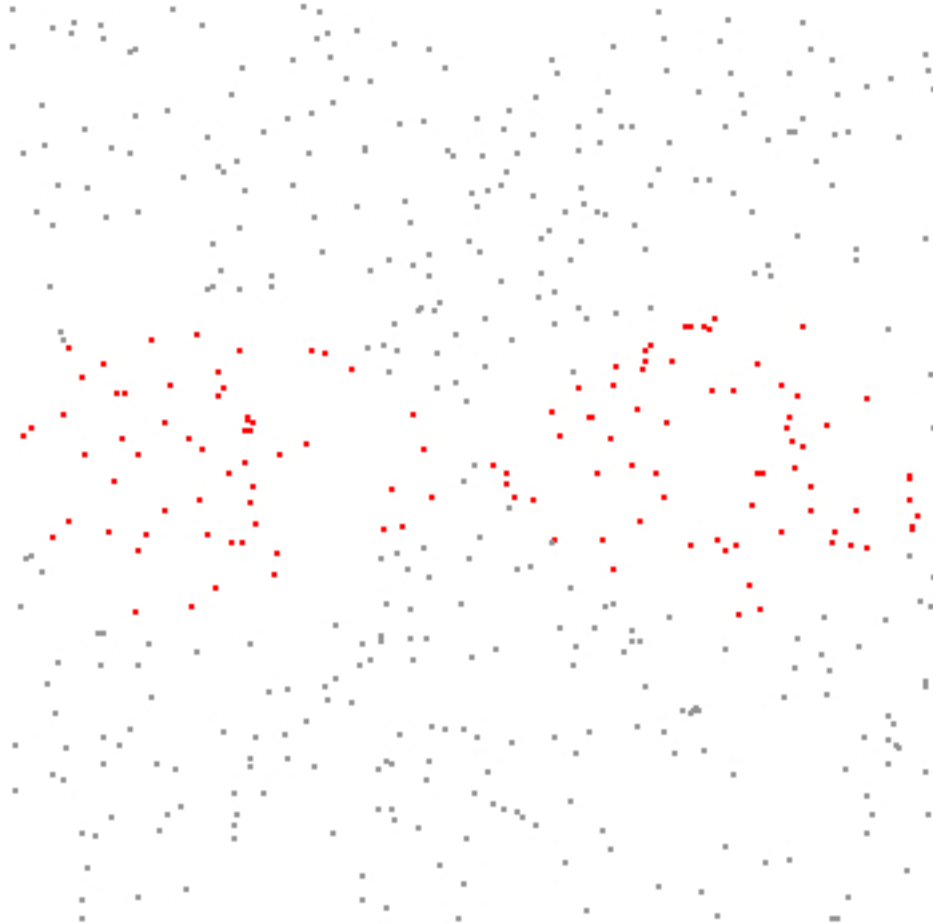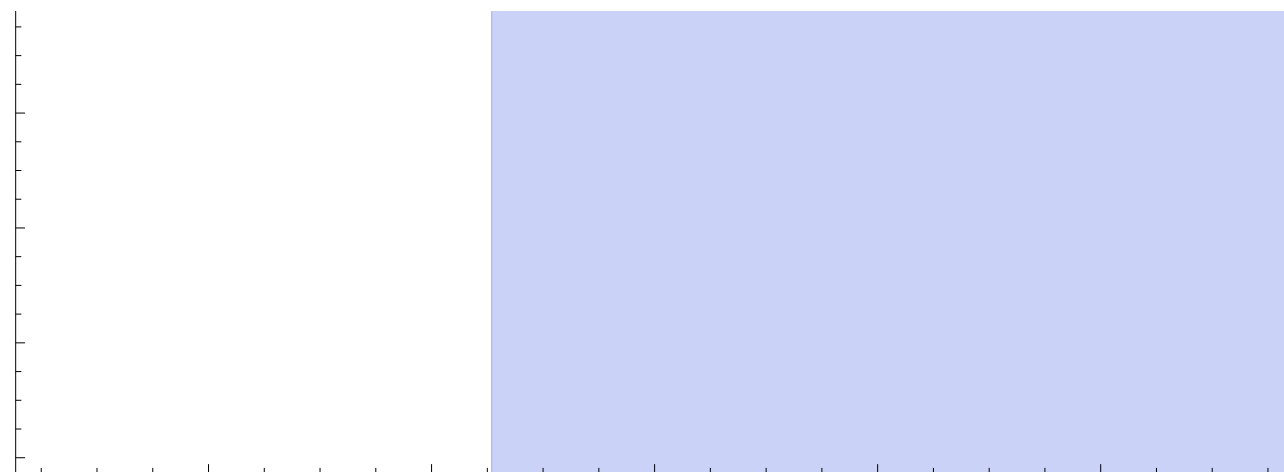
Parametric equation of Bernoulli's lemniscate

$$r = a\sqrt{\cos 2\theta}$$

a = 1

What is its area?

Empirical Monte Carlo distribution of the area estimate
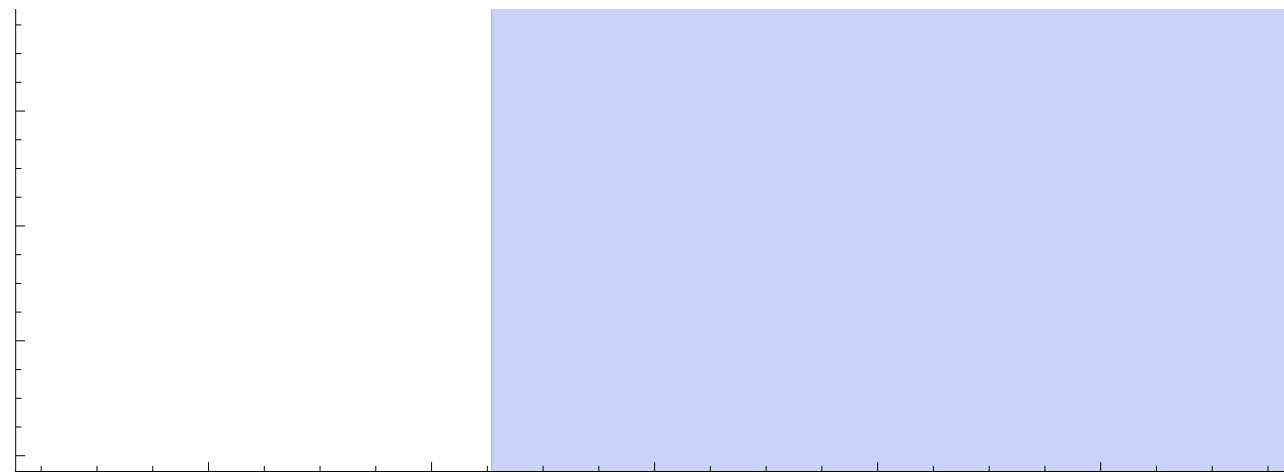
a = 1

1.02

area estimate

Empirical Monte Carlo distribution of the area estimate

a = 1

a probability distribution of
a mathematical constant???

1.02

area estimate

Question:
Are we asking a real scientific question?