Mathematical constraints on genetic differentiation statistics: theory, ecological, and oncological applications

N. Alcala

Rare Cancers Genomics Team

February 2022

International Agency for Research on Cancer

World Health Organization

RARE CANCERS GENOMICS

Plan

Introduction: measuring genetic differentiation

1. F_{ST} and genetic differentiation controversy
2. Alternative measures G'_{ST}, and D
3. Comparisons between their values

Results

1. Maximal F_{ST}, G'_{ST}, and D for a multiallelic marker in arbitrarily many populations
2. Application to Humans and species of conservation importance
3. Application to cancer

Measuring genetic differentiation

The controversy

Measuring genetic differentiation | F_{ST}

Genetic differentiation is often estimated using Wright's fixation index F_{ST}

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

where $H_T = 1 - \sum_{i=1}^{I} \left(\frac{1}{K} \sum_{k=1}^{K} p_{k,i} \right)^2$ is the total diversity and $H_S = 1 - \frac{1}{K} \sum_{i=1}^{I} \sum_{k=1}^{K} p_{k,i}^2$ is the subpopulation diversity, *K* is the number of subpopulations, *I* the number of alleles, and $p_{k,i}$ the frequency of allele i in subpopulation *k*.

Measuring genetic differentiation | F_{ST}

Genetic differentiation is often estimated using Wright's fixation index F_{ST}

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

where $H_T = 1 - \sum_{i=1}^{I} \left(\frac{1}{K} \sum_{k=1}^{K} p_{k,i} \right)^2$ is the total diversity and $H_S = 1 - \frac{1}{K} \sum_{i=1}^{I} \sum_{k=1}^{K} p_{k,i}^2$ is the subpopulation diversity, K is the number of subpopulations, I the number of alleles, and $p_{k,i}$ the frequency of allele i in subpopulation k.

Interpretation [edit]

This comparison of genetic variability within and between populations is frequently used in applied population genetics. The values range from 0 to 1. A zero value implies complete panmixis; that is, that the two populations are interbreeding freely. A value of one implies that all genetic variation is explained by the population structure, and that the two populations do not share any genetic diversity.

Wikipedia (accessed February 2022)

Measuring genetic differentiation | F_{ST}

Genetic differentiation is often estimated using Wright's fixation index F_{ST}

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

where $H_T = 1 - \sum_{i=1}^{I} \left(\frac{1}{K} \sum_{k=1}^{K} p_{k,i} \right)^2$ is the total diversity and $H_S = 1 - \frac{1}{K} \sum_{i=1}^{I} \sum_{k=1}^{K} p_{k,i}^2$ is the subpopulation diversity, K is the number of subpopulations, I the number of alleles, and $p_{k,i}$ the frequency of allele i in subpopulation k.

Interpretation [edit]

This comparison of genetic variability within and between populations is frequently used in applied population genetics. The values range from 0 to 1. A zero value implies complete panmixis; that is, that the two populations are interbreeding freely. A value of one implies that all genetic variation is explained by the population structure, and that the two populations do not share any genetic diversity.

Wikipedia (accessed February 2022)

Pros: related to many evolutionary models (coalescence in structured populations, probability of fixation of an allele, simple dependence to migration and mutation rates under an island model of migration)

Measuring genetic differentiation | *Issue with F_{ST}*

F_{ST} was shown to be constrained by the value of H_S (Hedrick 1999),

$$F_{ST} \leq \frac{(K-1)(1-H_S)}{K-1+H_S},$$

and to be smaller for highly variable markers (e.g., microsatellites).

Measuring genetic differentiation | *Issue with F_{ST}*

F_{ST} was shown to be constrained by the value of H_S (Hedrick 1999),

$$F_{ST} \leq \frac{(K-1)(1-H_S)}{K-1+H_S},$$

and to be smaller for highly variable markers (e.g., microsatellites).

Example (Balloux et al. 2000): in contact zone of common shrew, F_{ST}=0.19 for haplotypes and 0.10 for microsatellites

Measuring genetic differentiation | *Solution 1: normalization*

Hedrick (2005) proposed to normalize F_{ST} by its maximal value given H_S,

$$G'_{ST} = \frac{F_{ST}}{\frac{(K-1)(1-H_S)}{K-1+H_S}}.$$

Pros: can reach 1 for any H_S
Cons: unrelated to any evolutionary model parameters

Measuring genetic differentiation | *Solution 2: novel statistics*

Jost (2008) criticized F_{ST} as a measure of the differences in genetic composition of populations

	Species A		Species B		Species C	
Allele	Subpop. 1	Subpop. 2	Subpop. 1	Subpop. 2	Subpop. 1	Subpop. 2
1	0.5	0.5	0.2	0.8	0.095	0
2	0.5	0.5	0.8	0.2	0.08	0
3	0	0	0	0	0.11	0
4	0	0	0	0	0.08	0
5	0	0	0	0	0.095	0
6	0	0	0	0	0.06	0
7	0	0	0	0	0.07	0
8	0	0	0	0	0.096	0
9	0	0	0	0	0.094	0
10	0	0	0	0	0.08	0
11	0	0	0	0	0.03	0
12	0	0	0	0	0.06	0
13	0	0	0	0	0.05	0
14	0	0	0	0	0	0.15
15	0	0	0	0	0	0.16
16	0	0	0	0	0	0.12
17	0	0	0	0	0	0.13
18	0	0	0	0	0	0.17
19	0	0	0	0	0	0.14
20	0	0	0	0	0	0.13

Measures of differentiation; should increase with increasing differentiation:

	Species A	Species B	Species C
D_{ST}	0	0.18	0.06(!)
G_{ST}	0	0.36	0.06(!)
H_{ST}	0	0.26	0.5
Δ_{ST}	1.00	1.36	2.00
D	0	0.53	1.00

Measures of similarity; should decrease with increasing differentiation:

	Species A	Species B	Species C
H_S/H_T	1.00	0.64	0.94(!)
Δ_S/Δ_T	1.00	0.74	0.50

Illustrative cases highlighting misconceptions in interpretation of F_{ST} (here called G_{ST} following Nei…)

Measuring genetic differentiation | *Solution 2: novel statistics*

Jost (2008) laid out a series of desirable properties that a measure of genetic differentiation useful for conservation should have

Motivating question: given 20 independent, equally diverse subpopulations (100 equi-frequent alleles), how many should I save to preserve 50% of the species' diversity?

Using H, H_T=0.9995 and H_S=0.99 in each subpopulation, so a single subpopulation preserves 99% of H

Measuring genetic differentiation | *Solution 2: novel statistics*

Jost (2008) laid out a series of desirable properties that a measure of genetic differentiation useful for conservation should have

Motivating question: given 20 independent, equally diverse subpopulations (100 equi-frequent alleles), how many should I save to preserve 50% of the species' diversity?

Using H, H_T=0.9995 and H_S=0.99 in each subpopulation, so a single subpopulation preserves 99% of H

Jost's interpretation: H is not well-equipped to use as ratios. Need an important property:

- Linear metric with respects to subpopulation pooling (Hill 1973): *if we merge 2 equally diverse subpopulations, the resulting population should have twice its initial diversity*

Measuring genetic differentiation | *Solution 2: novel statistics*

Jost (2008) laid out a series of desirable properties that a measure of genetic differentiation useful for conservation should have

Motivating question: given 20 independent, equally diverse subpopulations (100 equi-frequent alleles), how many should I save to preserve 50% of the species' diversity?

Using H, H_T=0.9995 and H_S=0.99 in each subpopulation, so a single subpopulation preserves 99% of H

Jost's interpretation: H is not well-equipped to use as ratios. Need an important property:

- Linear metric with respects to subpopulation pooling (Hill 1973): *if we merge 2 equally diverse subpopulations, the resulting population should have twice its initial diversity.*

  General formula for this family of metrics: $\Delta_q = \left(\sum_{i=1}^{I} p_i^q\right)^{1/(1-q)}$, where q determines the sensitivity to allele frequencies. Interestingly, when q tends to 1, Δ_q tends to the exponential of Shannon entropy, and when q=2, $\Delta_q = 1/(1-H)$. Note that Δ_q is not bounded by 1.

Measuring genetic differentiation | *Solution 2: novel statistics*

Jost (2008) proposed novel diversity and differentiation statistics satisfying this linearity principal, and using a multiplicative partitioning of total diversity into within- and between-population components:

$$\Delta_S = \frac{1}{1-H_S}, \Delta_T = \frac{1}{1-H_T}, \Delta_T = \Delta_{ST}\Delta_S, D = \frac{K}{K-1}\frac{\Delta_S}{\Delta_T},$$

Nevertheless, this statistic can also be expressed in terms of heterozygosities

$$D = \frac{K}{K-1}\frac{H_T - H_S}{1 - H_S},$$

Pros: can reach 1 for any H_S
Cons: relationship with evolutionary model parameters not obvious

Measuring genetic differentiation | *Not so novel statistics*

Measures of diversity are all related

$$D = \frac{K}{K-1}\frac{F_{ST}H_T}{1-H_S}, D = \frac{KH_T}{K-1+H_S}G'_{ST}$$

Especially when *K* is large

$$D \rightarrow_{K\rightarrow\infty} \frac{F_{ST}H_T}{1-H_S}, D \rightarrow_{K\rightarrow\infty} H_T G'_{ST}$$

Measuring genetic differentiation | *Not so novel statistics*

Measures of diversity are all related

$$D = \frac{K}{K-1}\frac{F_{ST}H_T}{1-H_S}, D = \frac{KH_T}{K-1+H_S}G'_{ST}$$

Especially when *K* is large

$$D \rightarrow_{K\rightarrow\infty} \frac{F_{ST}H_T}{1-H_S}, D \rightarrow_{K\rightarrow\infty} H_T G'_{ST}$$

And can be interpreted as different normalizations of Nei's absolute differentiation $D_{ST} = H_T - H_S$

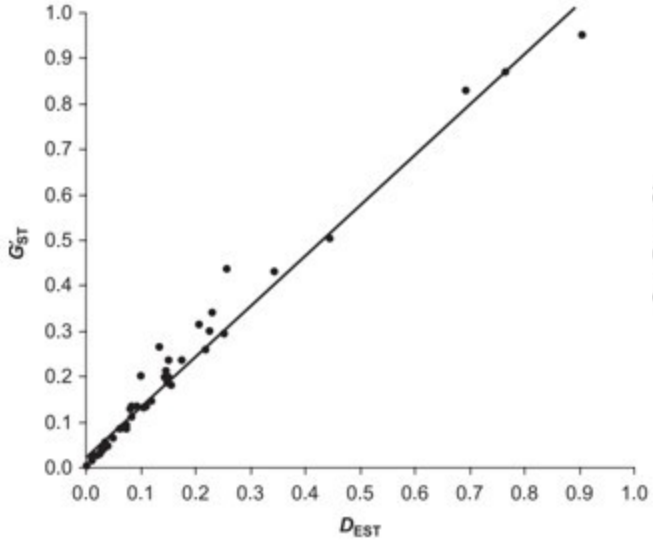
$F_{ST} = \frac{D_{ST}}{H_T}$, a normalization of D_{ST} by its maximal value given H_T

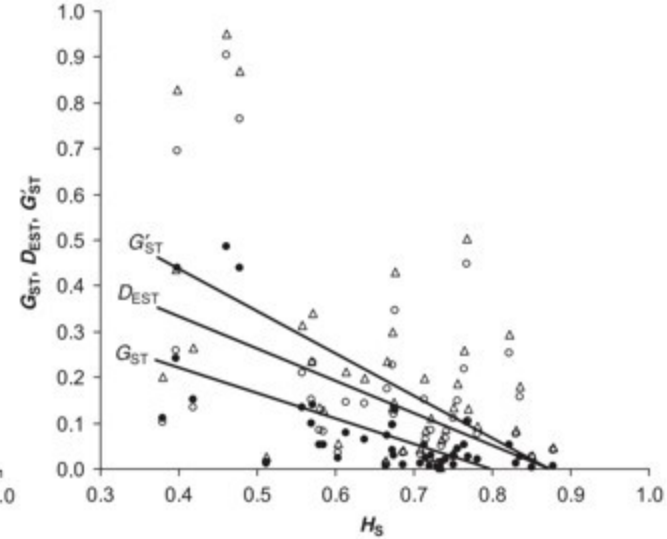
$D = \frac{K}{K-1}\frac{D_{ST}}{1-H_S}$, a normalization of D_{ST} by its maximal value given H_S

$G'_{ST} = \frac{\frac{D_{ST}}{H_T}}{\frac{(K-1)(1-H_S)}{K-1+H_S}}$, a double normalization of D_{ST} by its maximal values given H_T and then by that given H_S

Measuring genetic differentiation | *Not so novel statistics*

Empirically (Heller and Siegismund 2009), F_{ST} does appear generally smaller than G'_{ST} and D, but G'_{ST} and D are usually quite close





Comparison of the different statistics in 34 studies published in journal Molecular Ecology

Measuring genetic differentiation | *The neverending story*

This new knowledge on bounds on F_{ST} as a function of H_S and novel statistics closed the issue of the effect of H_S…

Measuring genetic differentiation | *The neverending story*

This new knowledge on bounds on F_{ST} as a function of H_S and novel statistics closed the issue of the effect of H_S…

… but opened new questions regarding the dependencies to other features of the allele frequency distribution, that the lab has now thoroughly explored.

Motivating example (Alcala and Rosenberg, In Press): using 246 microsatellites in Humans and Chimpanzee, between Humans and Chimp F_{ST}=0.10, but between the 6 Chimp subpopulations F_{ST}=0.16, although Humans and Chimp have similar H_S (0.3--0.4).

Measuring genetic differentiation | *The neverending story*

This new knowledge on bounds on F_{ST} as a function of H_S and novel statistics closed the issue of the effect of H_S…

… but opened new questions regarding the dependencies to other features of the allele frequency distribution, that the lab has now thoroughly explored.

Motivating example (Alcala and Rosenberg, In Press): using 246 microsatellites in Humans and Chimpanzee, between Humans and Chimp F_{ST}=0.10, but between the 6 Chimp subpopulations F_{ST}=0.16, although Humans and Chimp have similar H_S (0.3--0.4).
⇒ **something else is constraining F_{ST} values and impairing our interpretation**

Focus on total heterozygosity H_T , and the allele frequency distribution themselves, and in particular the frequency of the most frequent allele in the total population, $M = (1/K)\sum_{k=1}^{K} p_{k,1}$.

Measuring genetic differentiation | *A decade of studies*

Reference	# alleles	# pops	Statistic	Variable
Long and Kittles 2003	Unspecified value>1	Fixed finite >1	F_{ST}	H_S
Hedrick 2005	Unspecified value>1	Fixed finite >1	F_{ST}	H_S
Maruki et al. 2012	2	2	F_{ST}	H_S, M
Jakobsson et al. 2013	Unspecified value >1	2	F_{ST}	H_T, M
Edge and Rosenberg 2014	Fixed finite>1	2	F_{ST}	H_T, M
Alcala and Rosenberg 2017	2	Fixed finite>1	F_{ST}	M
Alcala and Rosenberg 2019	2	Fixed finite>1	G'_{ST}, D	M
Alcala and Rosenberg In Press	Unspecified value >1	Fixed finite>1	F_{ST}	M
Alcala and Rosenberg In prep.	Unspecified value >1	Fixed finite>1	G'_{ST}, D	M

Constraints on F_{ST}, G'_{ST}, and D in terms of the frequency of the most frequent allele

Constraints on differentiation statistics | *Multiallelic markers*

Deriving the maximal F_{ST} , G'_{ST}, and D values as functions of M

Proof sketch:

1. Express F_{ST} , G'_{ST} , and D in terms of the most frequent allele M and sums of other allele frequencies
2. Maximize the 3 statistics for M in $(0,1/K]$
3. Consider the case when KM is integer
4. Consider the case when M is in $(1/K,1)$ and KM is not integer

Constraints on differentiation statistics | *Multiallelic markers*

1. Express F_{ST}, G'_{ST}, and D in terms of the most frequent allele M and sums of other allele frequencies

We denote $\sigma_i = \sum_{k=1}^{K} p_{k,i}$, and $S = \sum_{i=1}^{I} \sum_{k=1}^{K} p_{k,i}^2$. We assume that alleles are ordered from most frequent to least frequent in the total population, so $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_I$, and denote by $M = \frac{\sigma_1}{K}$ the frequency of the most frequent allele, and by $S_1 = \sum_{k=1}^{K} p_{k,1}^2$ the sum of squared allele frequencies of allele 1.

Constraints on differentiation statistics | *Multiallelic markers*

1. Express F_{ST}, G'_{ST}, and D in terms of the most frequent allele M and sums of other allele frequencies

We denote $\sigma_i = \sum_{k=1}^{K} p_{k,i}$, and $S = \sum_{i=1}^{I} \sum_{k=1}^{K} p_{k,i}^2$. We assume that alleles are ordered from most frequent to least frequent in the total population, so $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_I$, and denote by $M = \frac{\sigma_1}{K}$ the frequency of the most frequent allele, and by $S_1 = \sum_{k=1}^{K} p_{k,1}^2$ the sum of squared allele frequencies of allele 1. Then, $H_S = 1 - S/K$, $H_T = 1 - S/K^2 + S_1/K^2 - \sigma_1^2/K^2 - (\frac{2}{K^2}) \sum_{i=2}^{I} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} p_{k,i} p_{l,i}$

Constraints on differentiation statistics | *Multiallelic markers*

1. Express F_{ST}, G'_{ST}, and D in terms of the most frequent allele M and sums of other allele frequencies

We denote $\sigma_i = \sum_{k=1}^{K} p_{k,i}$, and $S = \sum_{i=1}^{I} \sum_{k=1}^{K} p_{k,i}^2$. We assume that alleles are ordered from most frequent to least frequent in the total population, so $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_I$, and denote by $M = \frac{\sigma_1}{K}$ the frequency of the most frequent allele, and by $S_1 = \sum_{k=1}^{K} p_{k,1}^2$ the sum of squared allele frequencies of allele 1. Then, $H_S = 1 - S/K$, $H_T = 1 - S/K^2 + S_1/K^2 - \sigma_1^2/K^2 - (\frac{2}{K^2}) \sum_{i=2}^{I} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} p_{k,i} p_{l,i}$, and we can write the statistics as:

$$F_{ST} = \frac{(K-1)S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{I} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} p_{k,i} p_{l,i}}{K^2 - S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{I} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} p_{k,i} p_{l,i}},$$

$$G'_{ST} = \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D = 1 - \frac{\sigma_1^2 - S_1 + 2\sum_{i=2}^{I} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} p_{k,i} p_{l,i}}{K(K-1)S}.$$

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

All statistics decrease as function of $2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}$, thus they are maximal when this sum is 0, which happens when for each allele *i*, there is no pair of subpopulations where alleles are non-zero, i.e., all alleles are private.

$$F_{ST} \leq \frac{(K-1)S + S_1 - \sigma_1^2}{K^2 - S + S_1 - \sigma_1^2},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S}.$$

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

All statistics decrease as function of $2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}$, thus they are maximal when this sum is 0, which happens when for each allele *i*, there is no pair of subpopulations where alleles are non-zero, i.e., all alleles are private.

$$F_{ST} \leq \frac{(K-1)S + S_1 - \sigma_1^2}{K^2 - S + S_1 - \sigma_1^2},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S}.$$

In addition, all statistics decrease as a function of $\sigma_1^2 - S_1 = \sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,1}p_{l,1}$. For *M* in (0,1/*K*], this term can equal 0 when allele 1 is present in a single subpopulation.

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

Then $\sigma_1 = S_1$, and the expressions simplify

$$F_{ST} \leq \frac{(K-1)S}{K^2 - S},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST} = 1,$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S} = 1.$$

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

Then $\sigma_1 = S_1$, and the expressions simplify

$$F_{ST} \leq \frac{(K-1)S}{K^2 - S},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST} = 1,$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S} = 1.$$

Unconstrained in this range of *M* values

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

Then $\sigma_1 = S_1$, and the expressions simplify

$$F_{ST} \leq \frac{(K-1)S}{K^2 - S},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST} = 1,$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S} = 1.$$

Unconstrained in this range of *M* values

This just leaves F_{ST}.

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

Then $\sigma_1 = S_1$, and the expressions simplify

$$F_{ST} \leq \frac{(K-1)S}{K^2 - S},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST} = 1,$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S} = 1.$$

Unconstrained in this range of *M* values

This just leaves F_{ST}. We note that F_{ST} decreases as a function of S, so maximizing F_{ST} can be reduced to maximizing $S = \sum_{i=1}^{I} \sum_{k=1}^{K} p_{k,i}^2$.

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

From Lemma 3 of Jakobsson et al., $S \leq K[1-\sigma_1(J-1)(2-J\sigma_1)]$, with $J = [\sigma_1^{-1}]$. The maximum is reached when each population has *J*-1 alleles at frequency σ_1 and 1 allele at frequency $1-(J-1)\sigma_1$.

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

From Lemma 3 of Jakobsson et al., $S \leq K[1-\sigma_1(J-1)(2-J\sigma_1)]$, with $J = [\sigma_1^{-1}]$. The maximum is reached when each population has *J*-1 alleles at frequency σ_1 and 1 allele at frequency $1-(J-1)\sigma_1$. Then

$$F_{ST} \leq \frac{(K-1)[1-\sigma_1(J-1)(2-J\sigma_1)]}{K-[1-\sigma_1(J-1)(2-J\sigma_1)]}.$$

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

From Lemma 3 of Jakobsson et al., $S \leq K[1-\sigma_1(J-1)(2-J\sigma_1)]$, with $J = [\sigma_1^{-1}]$. The maximum is reached when each population has *J*-1 alleles at frequency σ_1 and 1 allele at frequency $1-(J-1)\sigma_1$. Then

$$F_{ST} \leq \frac{(K-1)[1-\sigma_1(J-1)(2-J\sigma_1)]}{K-[1-\sigma_1(J-1)(2-J\sigma_1)]}.$$

Nota Bene: the condition maximizing F_{ST} in this domain encompasses that maximizing G'_{ST} and D, but are more restrictive. For F_{ST}, all allele frequencies are completely specified, while for *G'_{ST} and D*, there is an infinity of configurations satisfying the conditions to reach the maximum.

Constraints on differentiation statistics | *Multiallelic markers*

2. Maximize the 3 statistics for *M* in (0,1/*K*]

From Lemma 3 of Jakobsson et al., $S \leq K[1-\sigma_1(J-1)(2-J\sigma_1)]$, with $J = [\sigma_1^{-1}]$. The maximum is reached when each population has *J*-1 alleles at frequency σ_1 and 1 allele at frequency $1-(J-1)\sigma_1$. Then

$$F_{ST} \leq \frac{(K-1)[1-\sigma_1(J-1)(2-J\sigma_1)]}{K-[1-\sigma_1(J-1)(2-J\sigma_1)]}.$$

Nota Bene: the condition maximizing F_{ST} in this domain encompasses that maximizing G'_{ST} and D, but are more restrictive. For F_{ST}, all allele frequencies are completely specified, while for *G'_{ST} and D*, there is an infinity of configurations satisfying the conditions to reach the maximum.
In other words, having completely different sets of alleles in each population is sufficient to maximize D and G'_{ST} , but F_{ST} further requires within-subpopulation diversity to be minimal

Constraints on differentiation statistics | *Multiallelic markers*

3. Consider the case when *KM* is integer

Coming back to the expressions of the 3 statistics from (1), we have:

$$F_{ST} = \frac{(K-1)S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}}{K^2 - S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}},$$

$$G'_{ST} = \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D = 1 - \frac{\sigma_1^2 - S_1 + 2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}}{K(K-1)S}.$$

We note that F_{ST}=1 iff $S = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{k,i}^2 = K$. Then we also have G'_{ST}=1, but not necessarily D=1.

Constraints on differentiation statistics | *Multiallelic markers*

3. Consider the case when *KM* is integer

Coming back to the expressions of the 3 statistics from (1), we have:

$$F_{ST} = \frac{(K-1)S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}}{K^2 - S + S_1 - \sigma_1^2 - 2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}},$$

$$G'_{ST} = \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D = 1 - \frac{\sigma_1^2 - S_1 + 2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}}{K(K-1)S}.$$

We note that F_{ST}=1 iff $S = \sum_{i=1}^{I}\sum_{k=1}^{K} p_{k,i}^2 = K$. Then we also have G'_{ST}=1, but not necessarily D=1. S=K iff each population has an allele *i* at frequency 1 (i.e., complete fixation of all alleles). This requires $\sigma_1 = \sum_{k=1}^{K} p_{k,1}$ to be an integer, corresponding to the number of populations where the most frequent allele is fixed.

Constraints on differentiation statistics | *Multiallelic markers*

3. Consider the case when *KM* is integer

Thus for *KM* integer,

$$F_{ST} \leq 1$$

$$G'_{ST} \leq 1.$$

Under these conditions, F_{ST} and G'_{ST} have the same behavior.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when *M* is in (1/*K*,1) and *KM* is not integer

Under these conditions, all allèles except the most frequent allele can be present in a single population, so the term $2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}$ can be 0. The upper bound on the statistics shown previously applies:

$$F_{ST} \leq \frac{(K-1)S + S_1 - \sigma_1^2}{K^2 - S + S_1 - \sigma_1^2},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S}.$$

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when *M* is in (1/*K*,1) and *KM* is not integer

Under these conditions, all allèles except the most frequent allele can be present in a single population, so the term $2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}$ can be 0. The upper bound on the statistics shown previously applies:

$$F_{ST} \leq \frac{(K-1)S + S_1 - \sigma_1^2}{K^2 - S + S_1 - \sigma_1^2},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S}.$$

Nevertheless, allele 1 is necessarily present in at least 2 subpopulations, so $\sigma_1^2 \neq S_1$.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when *M* is in (1/*K*,1) and *KM* is not integer

Under these conditions, all allèles except the most frequent allele can be present in a single population, so the term $2\sum_{i=2}^{I}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} p_{k,i}p_{l,i}$ can be 0. The upper bound on the statistics shown previously applies:

$$F_{ST} \leq \frac{(K-1)S + S_1 - \sigma_1^2}{K^2 - S + S_1 - \sigma_1^2},$$

$$G'_{ST} \leq \left[\frac{K^2 - S}{(K-1)S}\right] F_{ST},$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)S}.$$

Nevertheless, allele 1 is necessarily present in at least 2 subpopulations, so $\sigma_1^2 \neq S_1$. We further split *S* into S_1 and $S^*=S-S_1$ to separate the case of the most frequent allele and that of the others.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when M is in $(1/K,1)$ and KM is not integer

The maximal values the become:

$$F_{ST} \leq \frac{KS_1 + (K-1)S^* - \sigma_1^2}{K^2 - S^* - \sigma_1^2},$$

$$G'_{ST} \leq \left[\frac{K^2 - S^* - S_1}{(K-1)(S^* + S_1)}\right] F_{ST},$$

$$D \leq 1 - \frac{\sigma_1^2 - S_1}{K(K-1)(S^* + S_1)}.$$

All statistics increase with S^* and S_1. Thus, they are maximized under the same conditions, when both S^* and S_1 are maximum.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when M is in $(1/K,1)$ and KM is not integer

We first maximize S_1. From theorem 1 from Alcala and Rosenberg 2017, this maximum is reached when allele 1 is fixed in $\lfloor \sigma_1 \rfloor$ populations, at frequency $\{\sigma_1\} = \sigma_1 - \lfloor \sigma_1 \rfloor$ in a single population, and at frequency 0 elsewhere.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when *M* is in (1/*K*,1) and *KM* is not integer

We first maximize S_1. From theorem 1 from Alcala and Rosenberg 2017, this maximum is reached when allele 1 is fixed in $\lfloor \sigma_1 \rfloor$ populations, at frequency $\{\sigma_1\} = \sigma_1 - \lfloor \sigma_1 \rfloor$ in a single population, and at frequency 0 elsewhere. Then $S_1 = \sum_{k=1}^{K} p_{k,1}^2 = \lfloor \sigma_1 \rfloor + \{\sigma_1\}^2$.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when M is in $(1/K,1)$ and KM is not integer

We then maximize S^*. Because we considered all alleles except the first one to be private, for all alleles except the first, σ_i is equal to the frequency of the only population where it is non null, and it is thus lower than or equal to 1.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when *M* is in (1/*K*,1) and *KM* is not integer

We then maximize *S**. Because we considered all alleles except the first one to be private, for all alleles except the first, σ_i is equal to the frequency of the only population where it is non null, and it is thus lower than or equal to 1. In addition, because $\sum_{i=1}^{I} \sigma_i = K$, $\sum_{i=2}^{I} \sigma_i = K - \sigma_1$. Thus maximizing *S** is equivalent to maximizing $\sum_{i=2}^{I} \sigma_i^2$ with the constraint that $\sum_{i=2}^{I} \sigma_i = K - \sigma_1$ and σ_i in [0,1].

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when *M* is in (1/*K*,1) and *KM* is not integer

We then maximize *S**. Because we considered all alleles except the first one to be private, for all alleles except the first, σ_i is equal to the frequency of the only population where it is non null, and it is thus lower than or equal to 1. In addition, because $\sum_{i=1}^{I} \sigma_i = K$, $\sum_{i=2}^{I} \sigma_i = K - \sigma_1$. Thus maximizing *S** is equivalent to maximizing $\sum_{i=2}^{I} \sigma_i^2$ with the constraint that $\sum_{i=2}^{I} \sigma_i = K - \sigma_1$ and σ_i in [0,1].

This problem was solved by Rosenberg and Jakobsson (2008), lemma 3. The maximum is reached when σ_i=1 for $K - \lfloor \sigma_1 \rfloor$ alleles, $\sigma_j = 1 - \{\sigma_1\}$ for a single allele, and all other allèles have frequency 0. Then $S^* = (1 - \{\sigma_1\})^2 + (K - \lfloor \sigma_1 \rfloor - 1)$.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when M is in $(1/K,1)$ and KM is not integer

The conditions maximizing S_1 and S^* require that as many populations as possible have allele 1 fixed, a single population has 2 alleles, at frequencies $\{\sigma_1\}$ and $1-\{\sigma_1\}$, and all other populations have a different allele fixed.

These conditions are the same for the three measures, and indeed is a trade-off between minimizing overlap, and maximizing fixation.

Constraints on differentiation statistics | *Multiallelic markers*

4. Consider the case when M is in $(1/K,1)$ and KM is not integer

The conditions maximizing S_1 and S^* require that as many populations as possible have allele 1 fixed, a single population has 2 alleles, at frequencies $\{\sigma_1\}$ and $1-\{\sigma_1\}$, and all other populations have a different allele fixed.

These conditions are the same for the three measures, and indeed is a trade-off between minimizing overlap, and maximizing fixation.

Substituting all maximal values into the previous equations leads to the upper bounds.

Constraints on differentiation statistics | *Multiallelic markers*

Final expressions

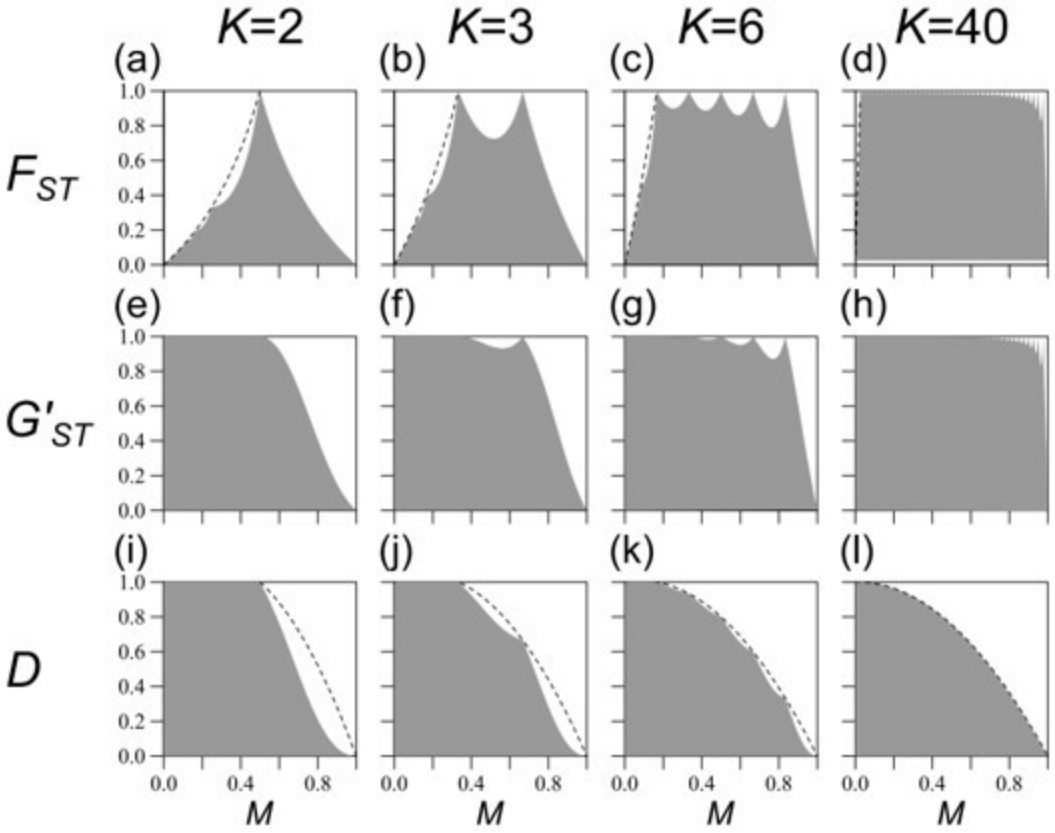
$$F'_{ST} \leq 1, \qquad \sigma_1 = 1,2,\ldots,K-1$$
$$\leq \frac{(K-1)[1-\sigma_1(J-1)(2-J\sigma_1)]}{K-[1-\sigma_1(J-1)(2-J\sigma_1)]}, 0 < \sigma_1 < 1$$
$$\leq \frac{[K(K-1)-\sigma_1^2+\lfloor\sigma_1\rfloor-2(K-1)\{\sigma_1\}+(2K-1)\{\sigma_1\}^2]}{[K(K-1)-\sigma_1^2-\lfloor\sigma_1\rfloor+2\sigma_1-\{\sigma_1\}^2]}, \text{non integer } 1 < \sigma_1 < K$$

$$G'_{ST} \leq 1, \qquad \sigma_1 < 1 \text{ or } \sigma_1 = 1,2,\ldots,K-1$$
$$\leq \frac{[K(K-1)+1-\{\sigma_1\}^2-(1-\{\sigma_1\})^2]}{[(K-1)(K-1+\{\sigma_1\}^2+(1-\{\sigma_1\})^2)]} \times$$
$$\frac{[K(K-1)-\sigma_1^2+\lfloor\sigma_1\rfloor-2(K-1)\{\sigma_1\}+(2K-1)\{\sigma_1\}^2]}{[K(K-1)-\sigma_1^2-\lfloor\sigma_1\rfloor+2\sigma_1-\{\sigma_1\}^2]}, \text{non integer } 1 < \sigma_1 < K$$

$$D \leq 1, \qquad \sigma_1 \leq 1,$$
$$\leq 1-\frac{\sigma_1^2-\{\sigma_1\}^2-\lfloor\sigma_1\rfloor}{(K-1)[K-1+\{\sigma_1\}^2+(1-\{\sigma_1\})^2]}, \qquad 1 < \sigma_1 < K$$

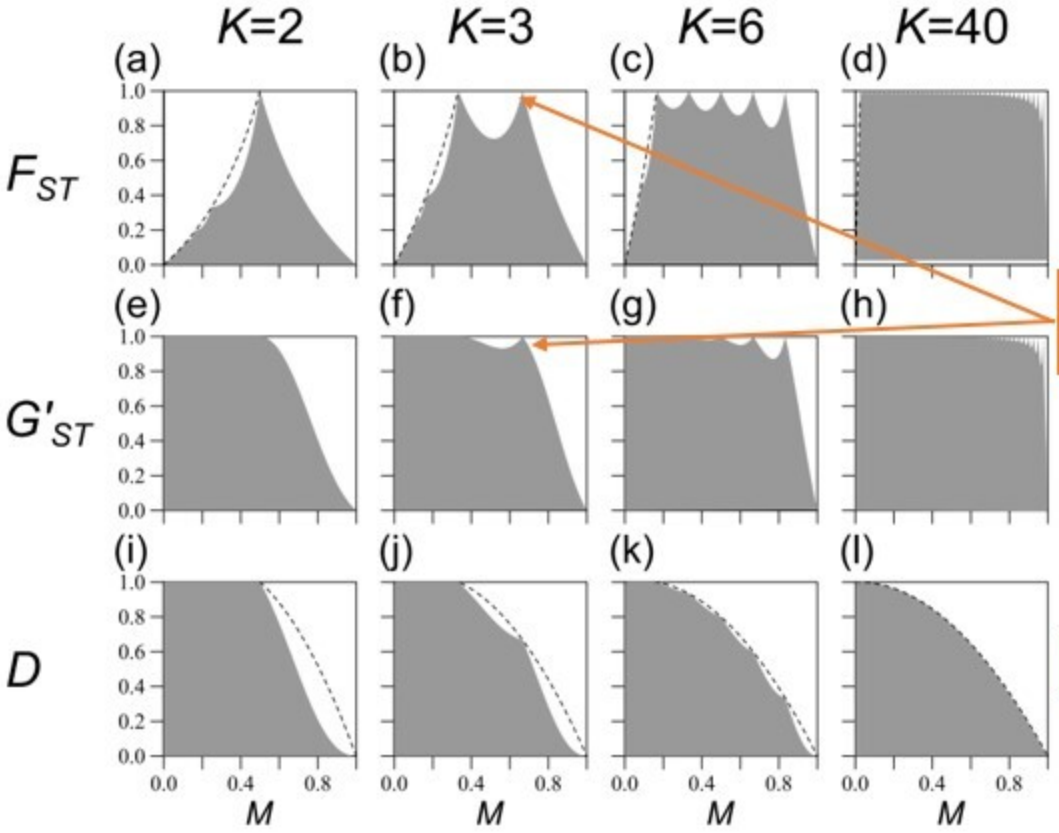
Constraints on differentiation statistics | *Multiallelic markers*

Maximal F_{ST} , G'_{ST}, and D values strongly differ but share some features



Constraints on differentiation statistics | *Multiallelic markers*

Maximal F_{ST} , G'_{ST}, and D values strongly differ but share some features

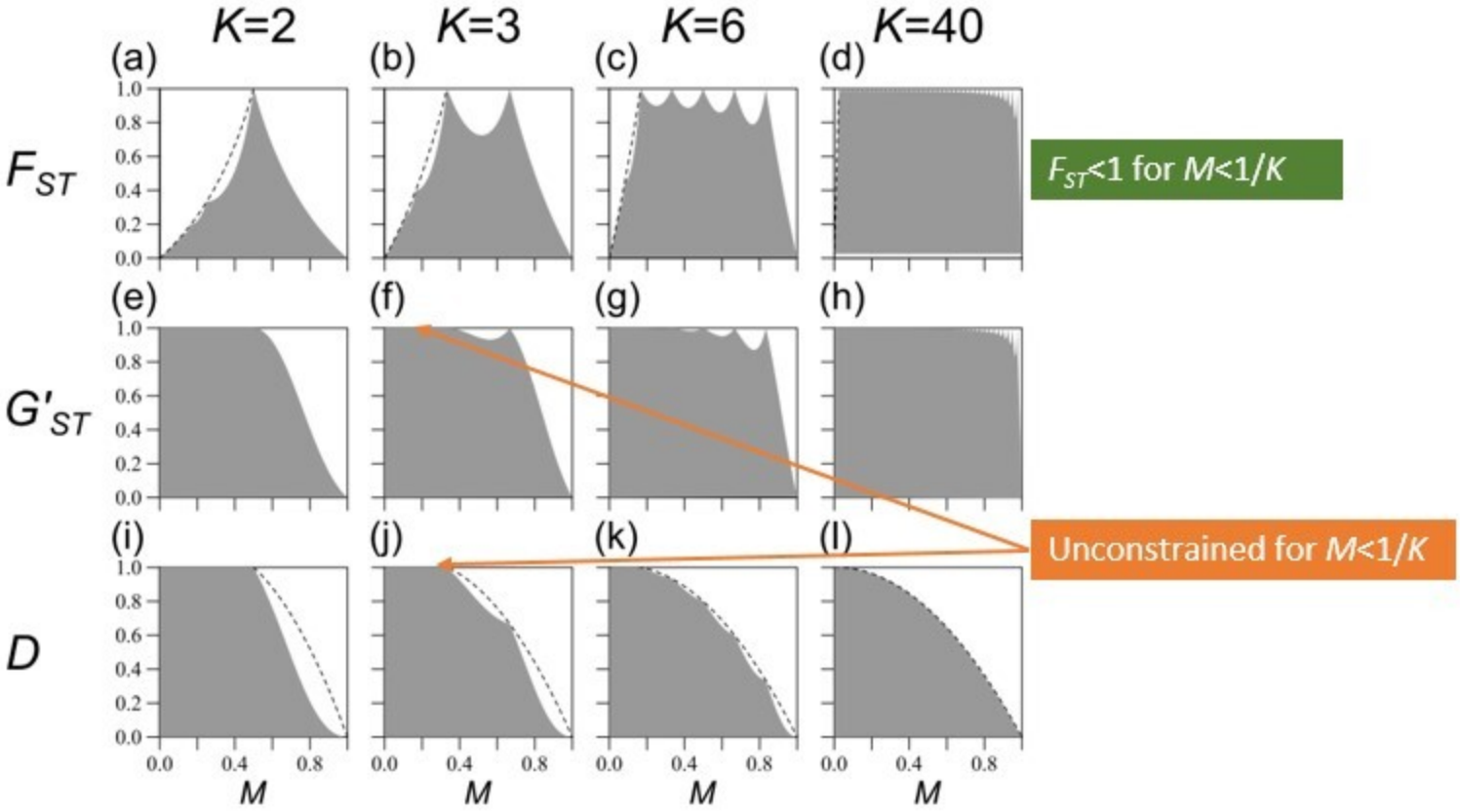


Peaks at i/K where maximum is 1

D<1 at i/K , for i>1

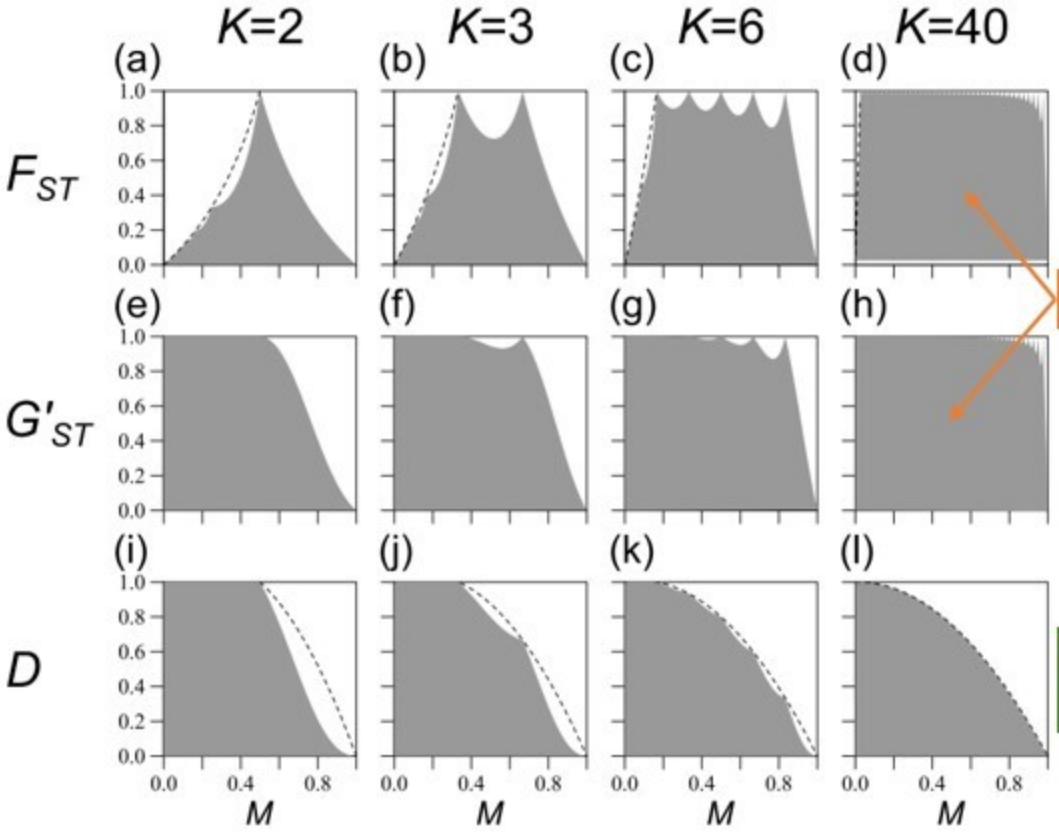
Constraints on differentiation statistics | *Multiallelic markers*

Maximal F_{ST} , G'_{ST}, and D values strongly differ but share some features



Constraints on differentiation statistics | *Multiallelic markers*

Maximal F_{ST} , G'_{ST}, and D values strongly differ but share some features



Ranges seem to increase with *K*

Range seems stable with *K*, but always large constraints

Constraints on differentiation statistics | *Multiallelic markers*

The ranges of F_{ST} , G'_{ST}, and D are differently impacted by the number of subpopulations K

Method: compute the range of possible values of each statistic across *M* values, i.e., the integral of the upper bound on each statistic.
This can be interpreted as the mean statistic assuming a uniform distribution of *M* values

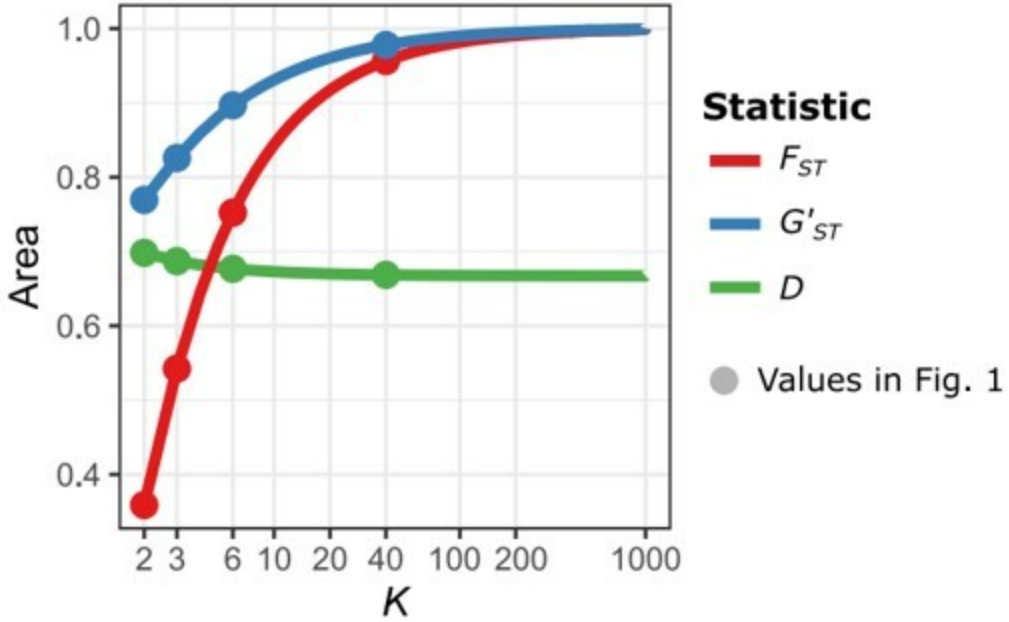
$$A_F(K) = \int_0^1 F_{\mathrm{ST,max}}(K, M)dM,$$

$$A_{G'}(K) = \int_0^1 G'_{\mathrm{ST,max}}(K, M)dM,$$

$$A_D(K) = \int_0^1 D_{\mathrm{max}}(K, M)dM$$

Constraints on differentiation statistics | *Multiallelic markers*

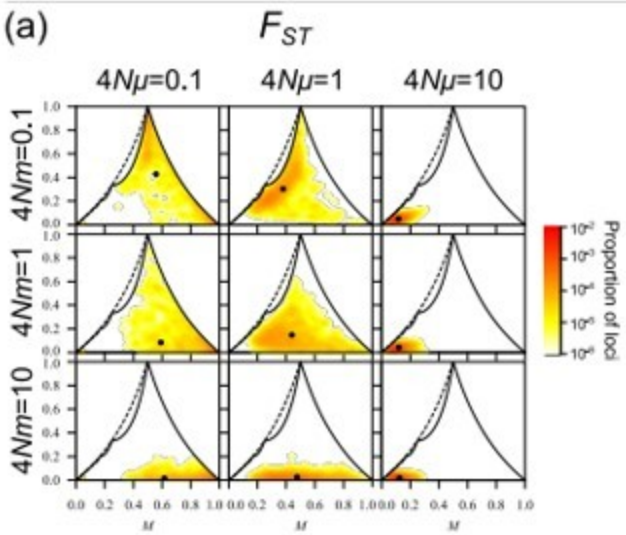
The ranges of F_{ST} , G'_{ST}, and D are differently impacted by the number of subpopulations K



The ranges of F_{ST} and G'_{ST} indeed tend to the full unit interval while D is restricted to approximately 2/3.

Constraints on differentiation statistics | *Multiallelic markers*

The 3 statistics have *different* behaviors as functions of the migration rate

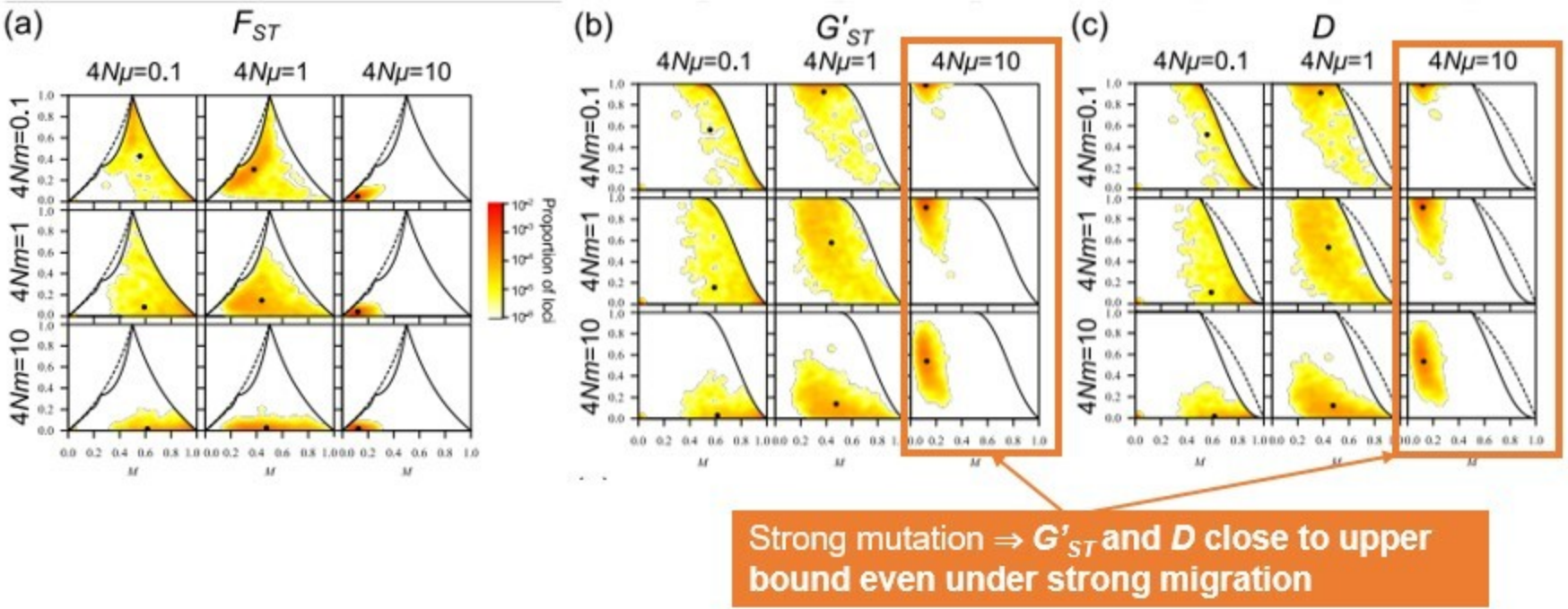


Weak migration ⇒ **close to upper bound**

Strong migration ⇒ **close to 0**

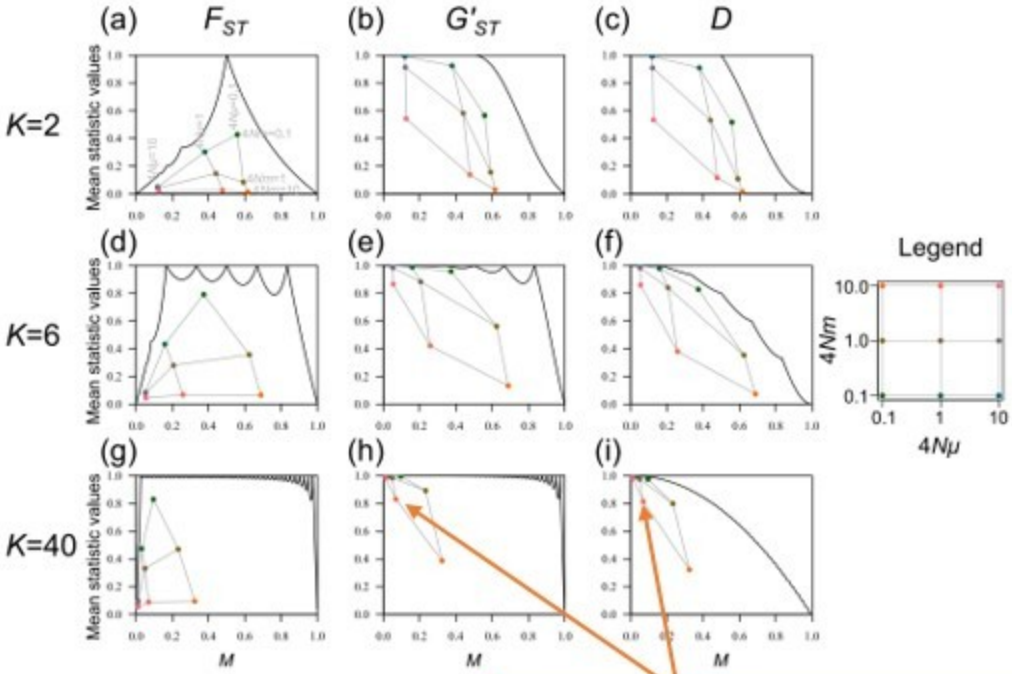
Constraints on differentiation statistics | *Multiallelic markers*

The 3 statistics have *different* behaviors as functions of the migration rate



Constraints on differentiation statistics | *Multiallelic markers*

The 3 statistics have *different* behaviors as functions of the migration rate



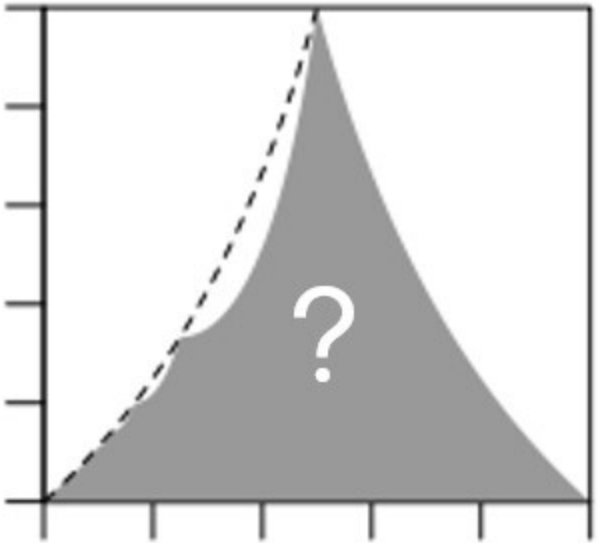
For large K, G'_{ST} and D even closer to upper bound under moderate and strong mutation even under strong migration

Constraints on differentiation statistics | *Multiallelic markers*

Take-home message

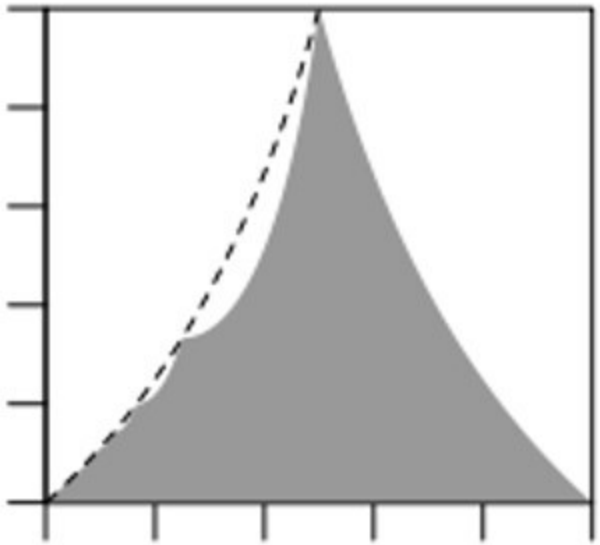
- F_{ST}, G'_{ST}, and D are differently constrained by the most frequent allele and thus total diversity
- Although their maximal values are not *always* reached under the same allelic configuration, they are reached under the same allelic configuration for a large range of allele frequencies M
- The number of subpopulations K strongly influences both the maximal values and the behavior of the statistics within their permissible range
- **Thus, visualizing the values of the statistics with respect to their maximal value given the most frequent allele *M* enables to separate differences due to their mathematical constraints from differences due to different dependencies on evolutionary forces**

A little game: guess my bound



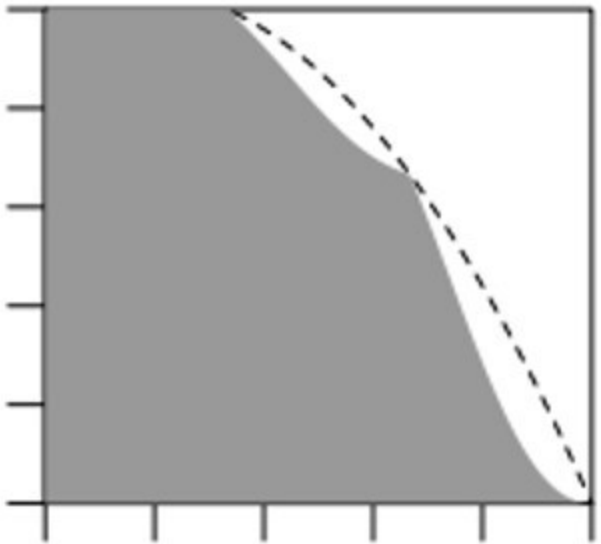
A little game: guess my bound

What statistic? What K?



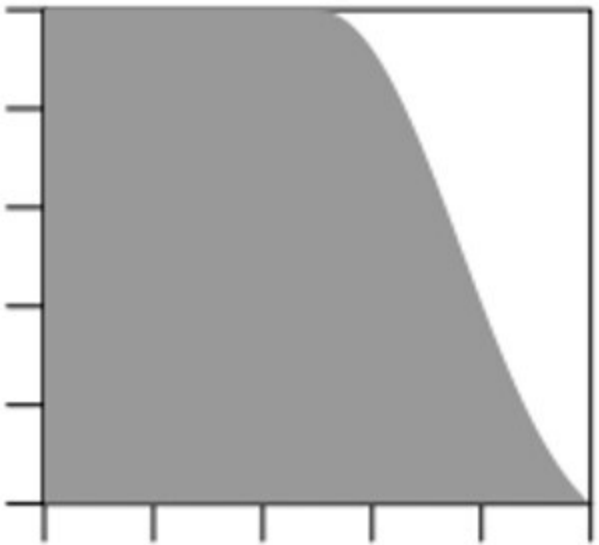
A little game: guess my bound

What statistic? What K?



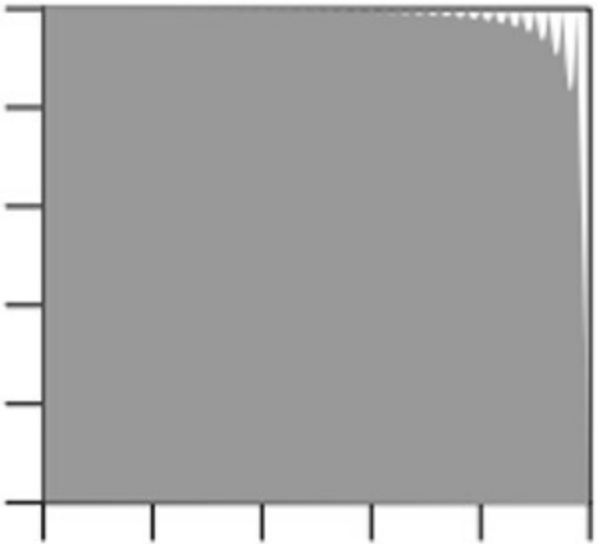
A little game: guess my bound

What statistic? What K?



A little game: guess my bound

What statistic? What K?

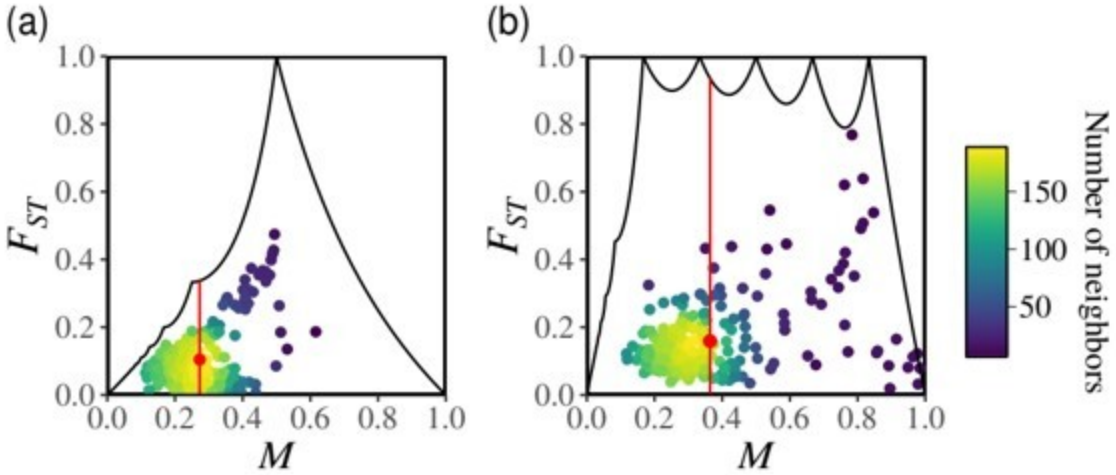


Applications

Application I: understanding human diversity

Application | *Human genetics*

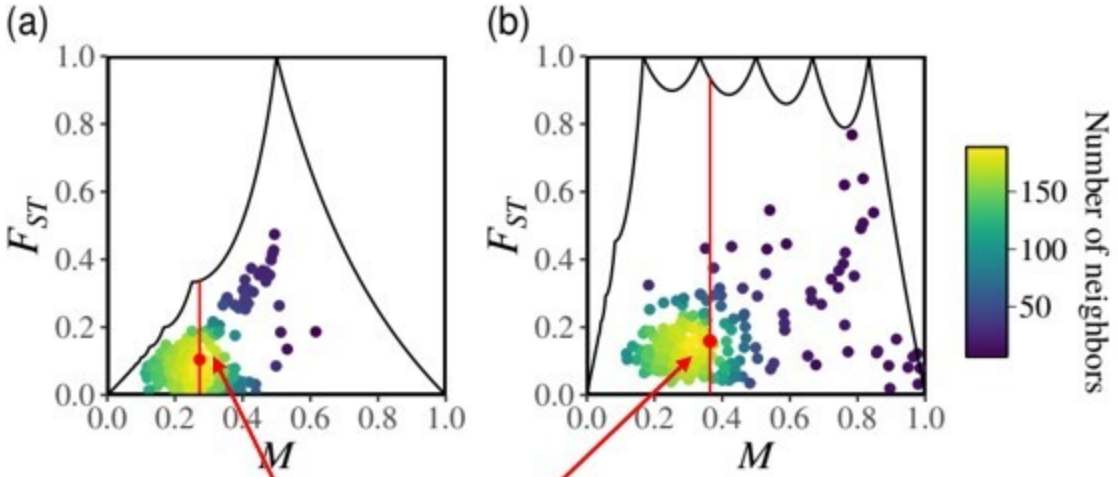
Back to our motivating examples: Humans and Chimps (Alcala and Rosenberg In press)



F_{ST} estimated from microsatellite data (246 loci) in 5795 Humans and 84 Chimpanzees from 6 subpopulations.

Application | *Human genetics*

Back to our motivating examples: Humans and Chimps (Alcala and Rosenberg In press)

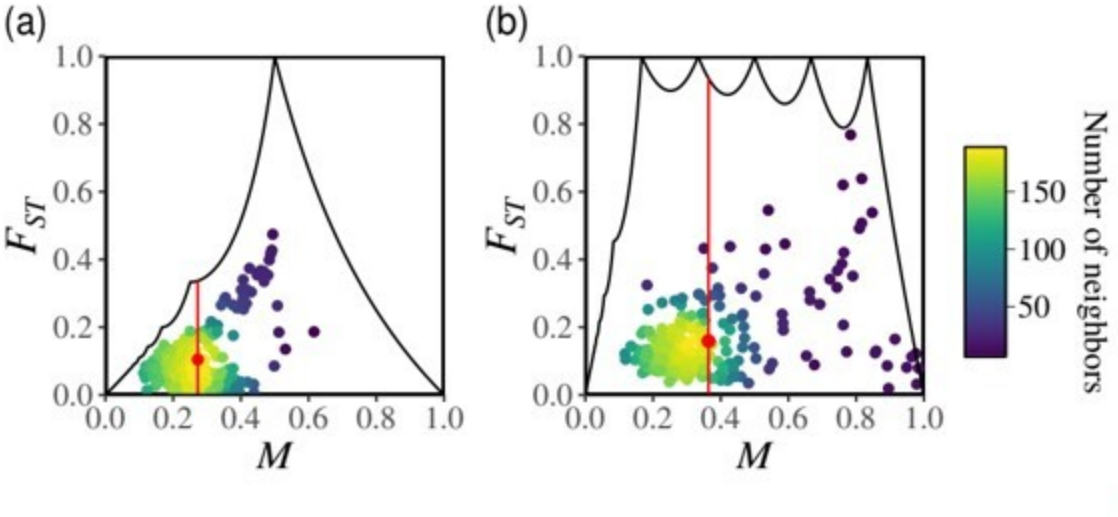


F_{ST} estimated from microsatellite data (246 loci) in 5795 Humans and 84 Chimpanzees from 6 subpopulations.

Stronger constraints in inter-species comparisons (K=2) than in intra-species multi-population comparisons (K=6) explain the paradox.

Application | *Human genetics*

Back to our motivating examples: Humans and Chimps (Alcala and Rosenberg In press)



F_{ST} estimated from microsatellite data (246 loci) in 5795 Humans and 84 Chimpanzees from 6 subpopulations.

Stronger constraints in inter-species comparisons (K=2) than in intra-species multi-population comparisons (K=6) explain the paradox.

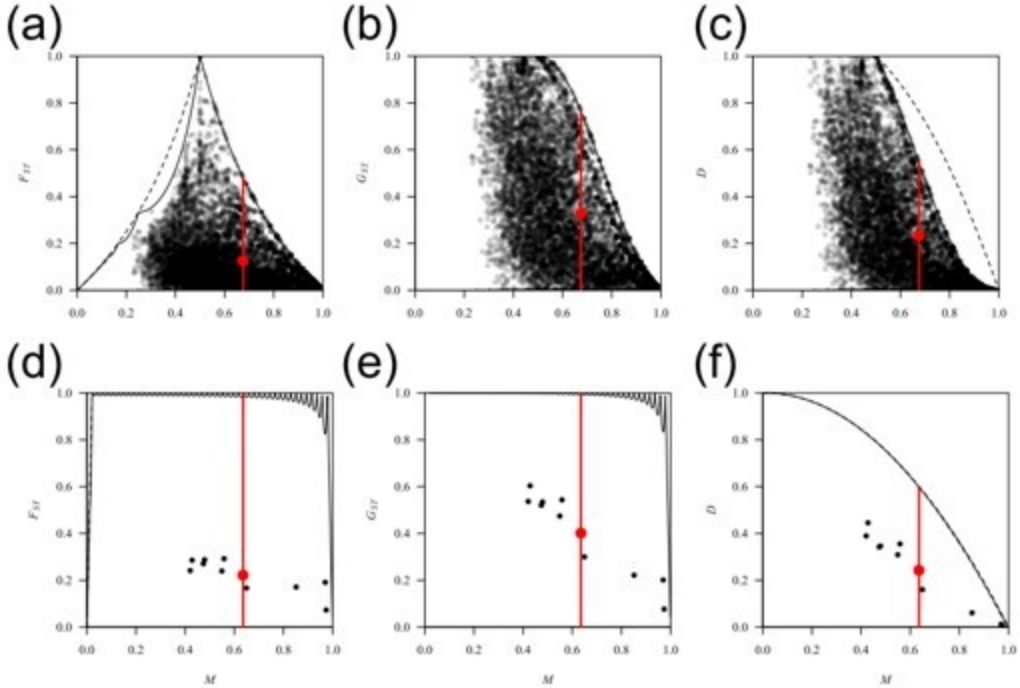
Normalization of F_{ST} by its maximal value given M provides a more intuitive interpretation.

Applications

Application II: monitoring diversity in species of conservation interest

Application | *Conservation*

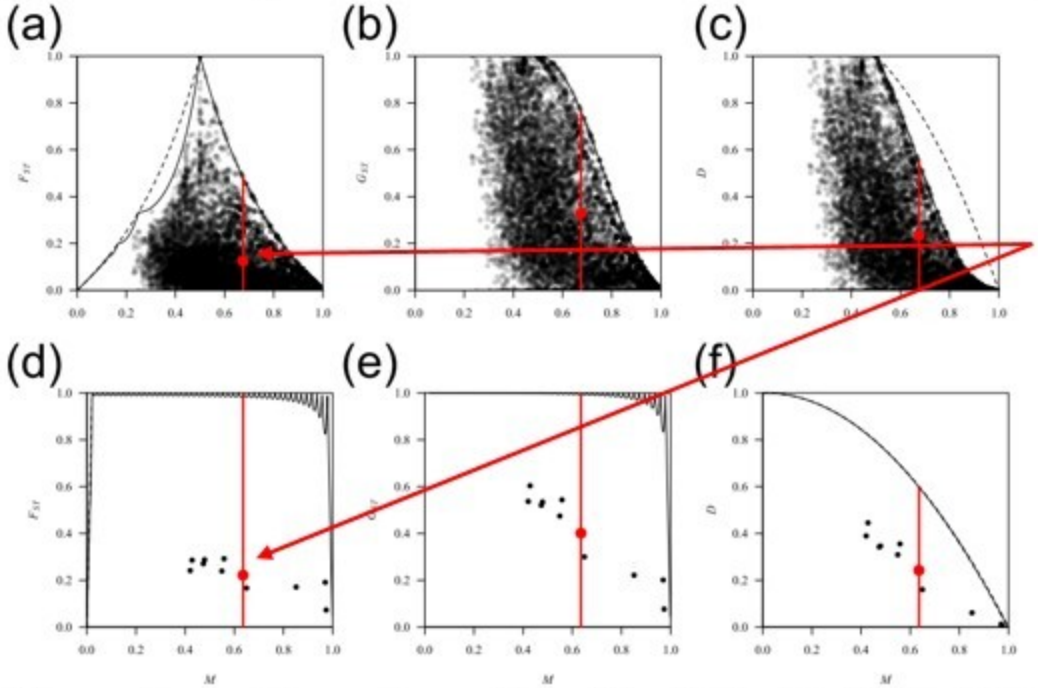
Application to yellow-bellied toad emphasizes the importance of *K*



Statistics estimated from microsatellite data (10 loci) in 885 individuals from 47 subpopulations.
Pairwise (a)-(c) and global (d)-(f) comparisons for each statistic.

Application | *Conservation*

Application to yellow-bellied toad emphasizes the importance of *K*

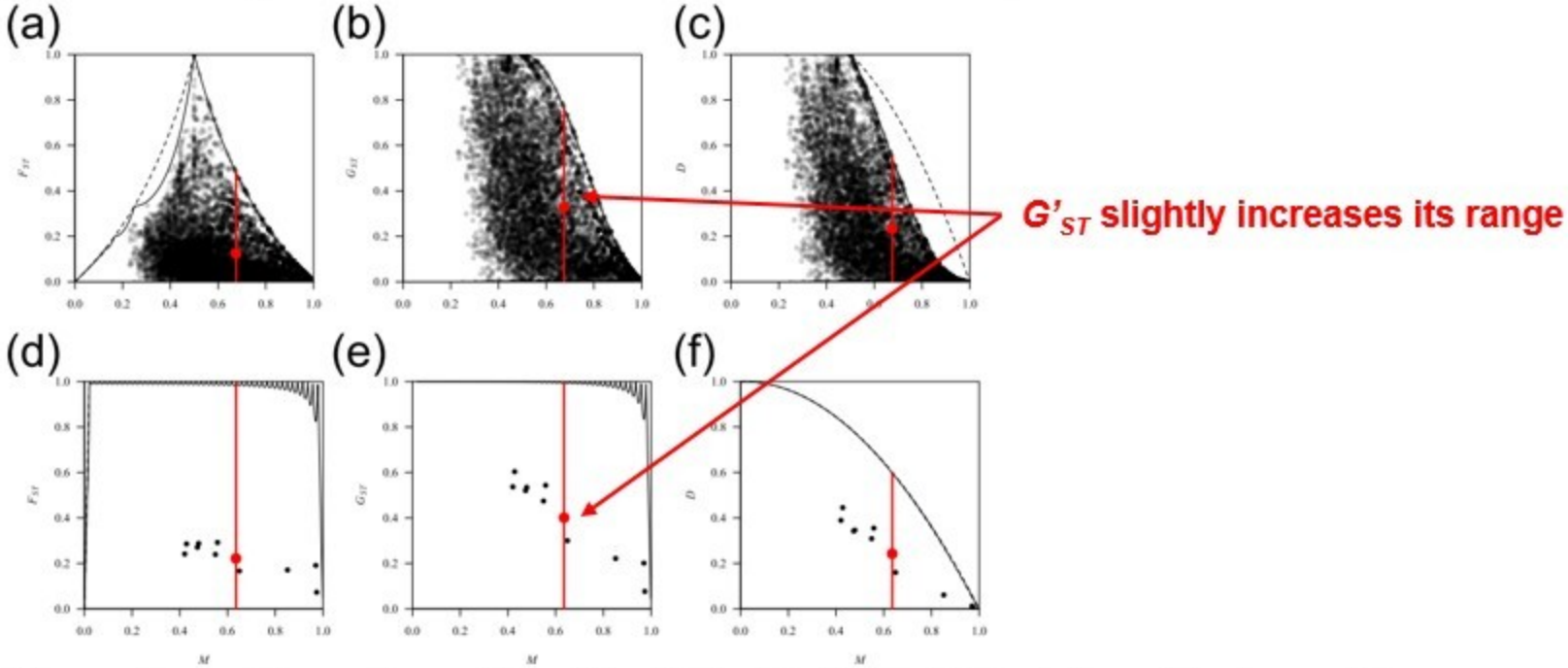


***F_{ST}* indeed doubles its range when K increases**

Statistics estimated from microsatellite data (10 loci) in 885 individuals from 47 subpopulations.
Pairwise (a)-(c) and global (d)-(f) comparisons for each statistic.

Application | *Conservation*

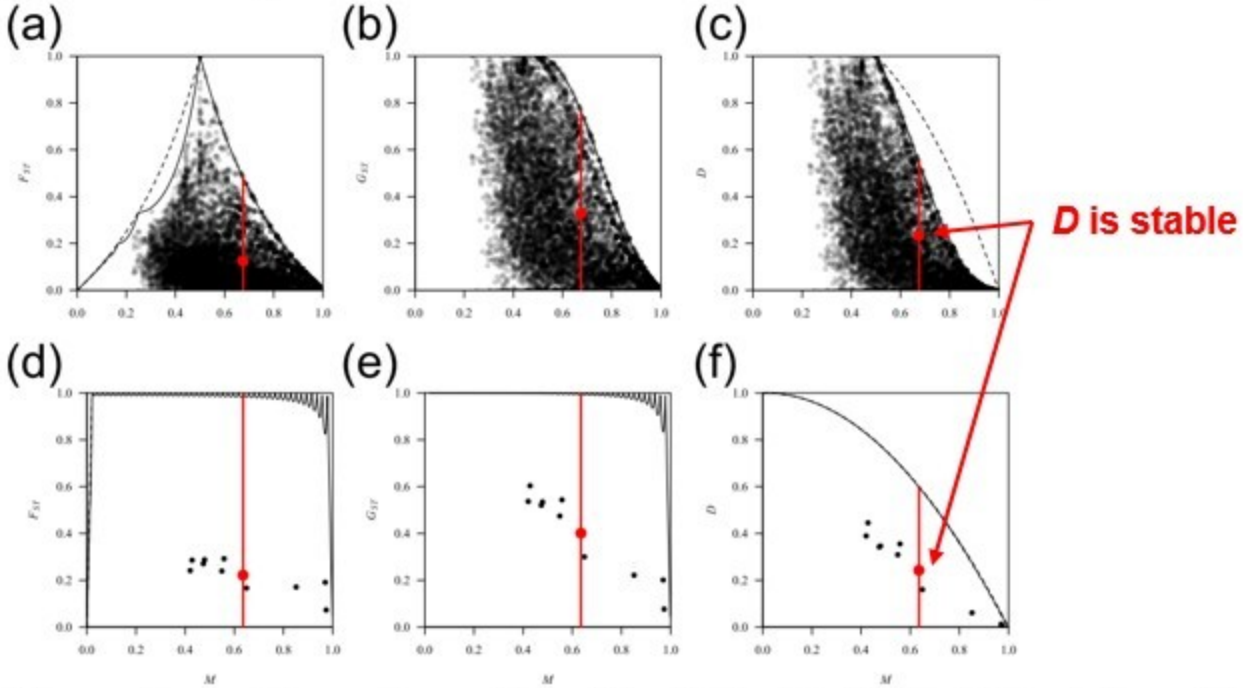
Application to yellow-bellied toad emphasizes the importance of *K*



Statistics estimated from microsatellite data (10 loci) in 885 individuals from 47 subpopulations.
Pairwise (a)-(c) and global (d)-(f) comparisons for each statistic.

Application | *Conservation*

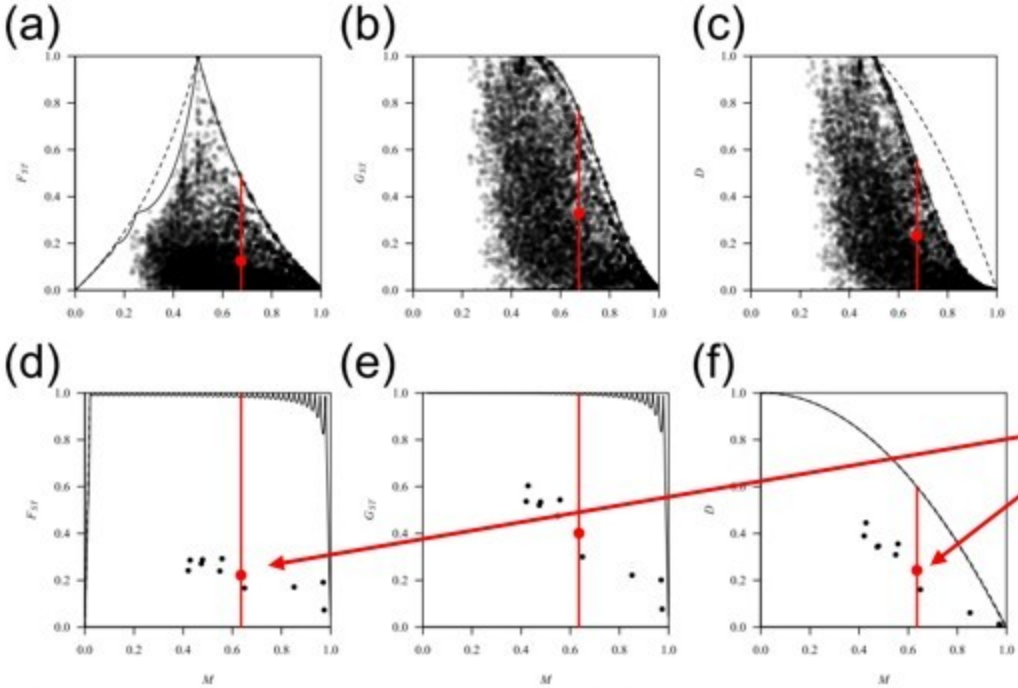
Application to yellow-bellied toad emphasizes the importance of *K*



Statistics estimated from microsatellite data (10 loci) in 885 individuals from 47 subpopulations.
Pairwise (a)-(c) and global (d)-(f) comparisons for each statistic.

Application | *Conservation*

Application to yellow-bellied toad emphasizes the importance of *K*



Even in cases where statistics have no constraints, differences subsist between them, because of different relationships with evolutionary parameters

Statistics estimated from microsatellite data (10 loci) in 885 individuals from 47 subpopulations.
Pairwise (a)-(c) and global (d)-(f) comparisons for each statistic.