Uso do Programa SuperMASSA para Genotipagem de SNPs em Poliploides

7° Congresso Brasileiro de Melhoramento de Plantas

Antonio Augusto Franco Garcia

http://about.me/augusto.garcia
augusto.garcia@usp.br

Marcelo Mollinari

http://www.mendeley.com/profiles/marcelo-mollinari/
mmollina@usp.br

Departamento de Genética
ESALQ/USP
2013

Outline

Outline

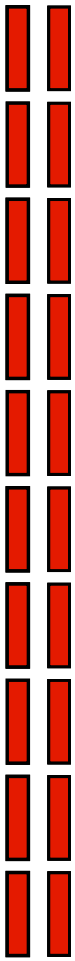
Outline

Paper

Serang, O. R.; Mollinari, M.; Garcia, A. A. F.
Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids
PLoS ONE 7(2), e30906, 2012

Genotyping diploids

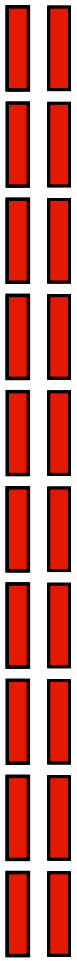
- Diploids: two sets of chromosomes
- Genotyping: measurement of variations (alleles) in homologous chromosomes within a locus
- Molecular markers: access the allelic variation in each of the homologous chromosomes
- Several techniques used to access these variations are qualitative



10 pairs of homologous chromosomes

Genotyping diploids

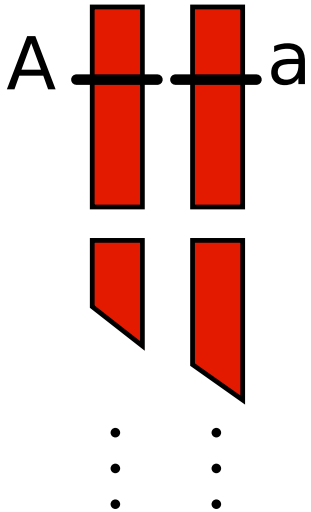
- Diploids: two sets of chromosomes
- Genotyping: measurement of variations (alleles) in homologous chromosomes within a locus
- Molecular markers: access the allelic variation in each of the homologous chromosomes
- Several techniques used to access these variations are qualitative



10 pairs of homologous chromosomes

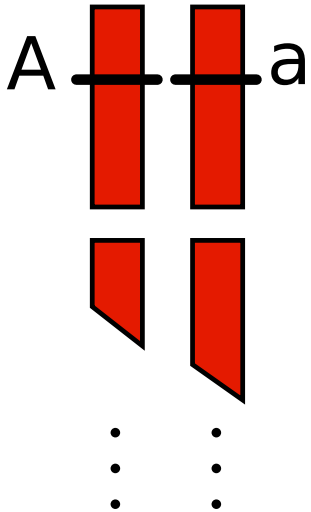
Genotyping diploids

- Diploids: two sets of chromosomes
- Genotyping: measurement of variations (alleles) in homologous chromosomes within a locus
- Molecular markers: access the allelic variation in each of the homologous chromosomes
- Several techniques used to access these variations are qualitative



Genotyping diploids

- Diploids: two sets of chromosomes
- Genotyping: measurement of variations (alleles) in homologous chromosomes within a locus
- Molecular markers: access the allelic variation in each of the homologous chromosomes
- Several techniques used to access these variations are qualitative



Use of molecular markers

- Access the genome via molecular markers and help breeders - Markers assisted selection (QTL mapping, GWS, GWAS)
- It is working pretty well in several diploid species (maize, soybean, etc)
- A saturated genetic map can help the genome assembly of sugarcane
- Genotype/Phenotype/Statistical Genetics

Use of molecular markers

- Access the genome via molecular markers and help breeders - Markers assisted selection (QTL mapping, GWS, GWAS)
- It is working pretty well in several diploid species (maize, soybean, etc)
- A saturated genetic map can help the genome assembly of sugarcane
- Genotype/Phenotype/Statistical Genetics

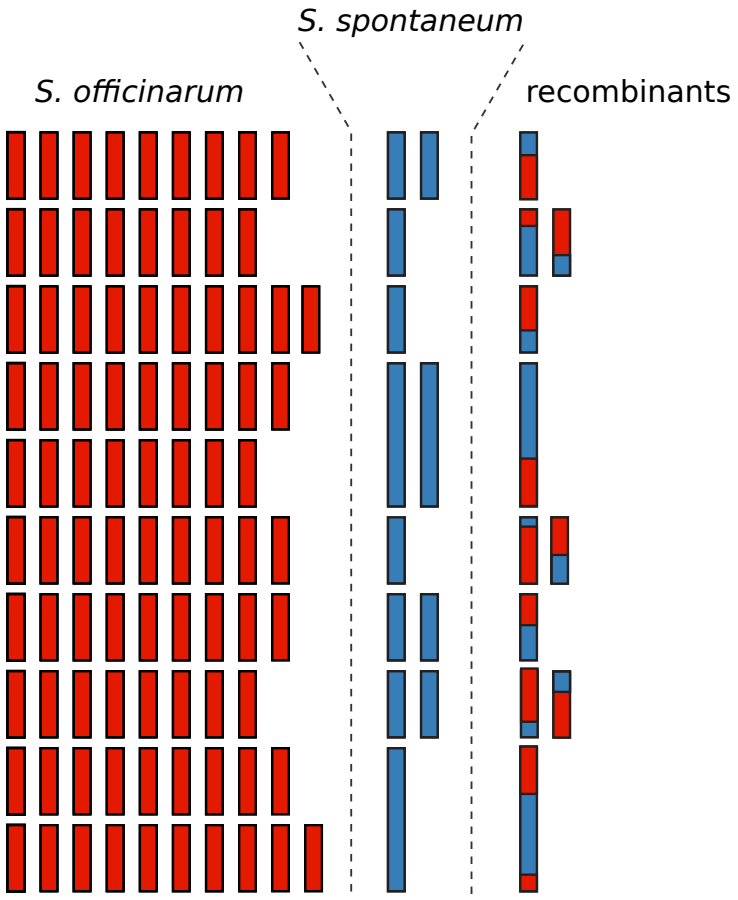
Use of molecular markers

- Access the genome via molecular markers and help breeders - Markers assisted selection (QTL mapping, GWS, GWAS)
- It is working pretty well in several diploid species (maize, soybean, etc)
- A saturated genetic map can help the genome assembly of sugarcane
- Genotype/Phenotype/Statistical Genetics

Use of molecular markers

- Access the genome via molecular markers and help breeders - Markers assisted selection (QTL mapping, GWS, GWAS)
- It is working pretty well in several diploid species (maize, soybean, etc)
- A saturated genetic map can help the genome assembly of sugarcane
- Genotype/Phenotype/Statistical Genetics

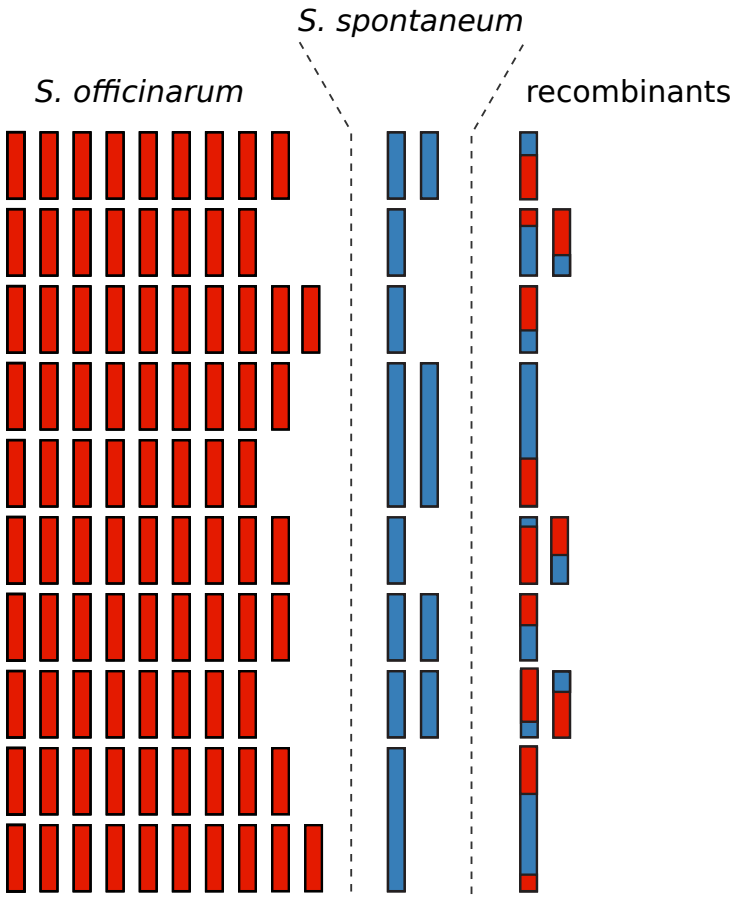
A complex auto(alo)polyploid example



D'Hont, 2005 - Cytogenet Genome Res 109:27–33

How can we use molecular markers in such complex genome?

A complex auto(alo)polyploid example



D'Hont, 2005 - Cytogenet Genome Res 109:27–33

How can we use molecular markers in such complex genome?

Genotyping in polyploids

- Using a biallelic marker, there are up to $p + 1$ possible genotypes.

```
0 - aaaaaaaa
1 - Aaaaaaaa
2 - AAaaaaaa
3 - AAAaaaaa
4 - AAAAaaaa
5 - AAAAAaaa
6 - AAAAAAaa
7 - AAAAAAAa
8 - AAAAAAAA
```

- Problem: using qualitative techniques, e.g. microsatellites, it is impossible distinguish genotypes which have at least one A

- It is important to distinguish genotypes (doses) in order to study association between molecular markers and phenotypes.

Genotyping in polyploids

- Using a biallelic marker, there are up to $p + 1$ possible genotypes.

```
0 - aaaaaaaa
1 - Aaaaaaaa
2 - AAaaaaaa
3 - AAAaaaaa
4 - AAAAaaaa
5 - AAAAAaaa
6 - AAAAAAaa
7 - AAAAAAAa
8 - AAAAAAAA
```

- Problem: using qualitative techniques, e.g. microsatellites, it is impossible distinguish genotypes which have at least one A

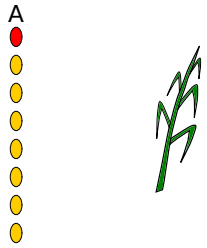
```
A-------
(presence)

aaaaaaaa
(absence)
```

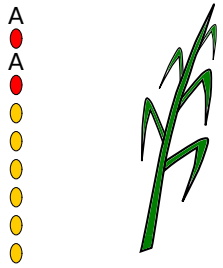
- It is important to distinguish genotypes (doses) in order to study association between molecular markers and phenotypes.

Genotyping in polyploids

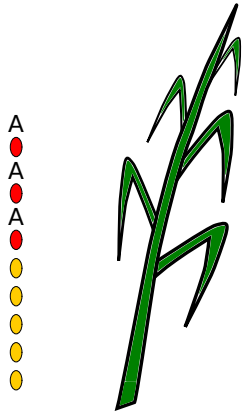
- Using a biallelic marker, there are up to $p + 1$ possible genotypes.



- Problem: using qualitative techniques, e.g. microsatellites, it is impossible distinguish genotypes which have at least one A



- It is important to distinguish genotypes (doses) in order to study association between molecular markers and phenotypes.

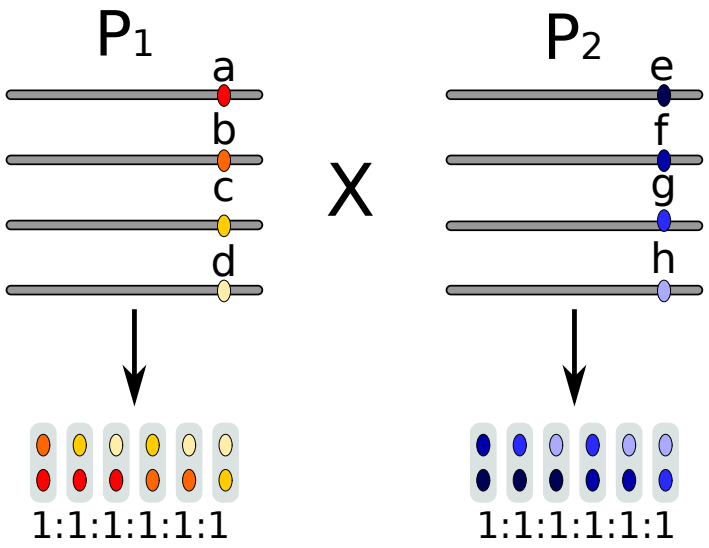


Codominant genotyping in autotetraploids

Tetraploid example - multiallelic

Number of balanced gametes: $\binom{p}{\frac{p}{2}}$

Number of possible combinations: $\binom{p}{\frac{p}{2}}^2$



Complex segregation patterns

Codominant genotyping in autotetraploids

Tetraploid example - multiallelic

a - 9 repeats

```
CTGACTTCAGGTACTTCTTCTTCTTCTTCTTCTTCTTCTTATGCCGTGATTGATC
GACTGAAGTCCATGAAGAAGAAGAAGAAGAAGAAGAAGAATACGGCACTAACTAG
```

b - 8 repeats

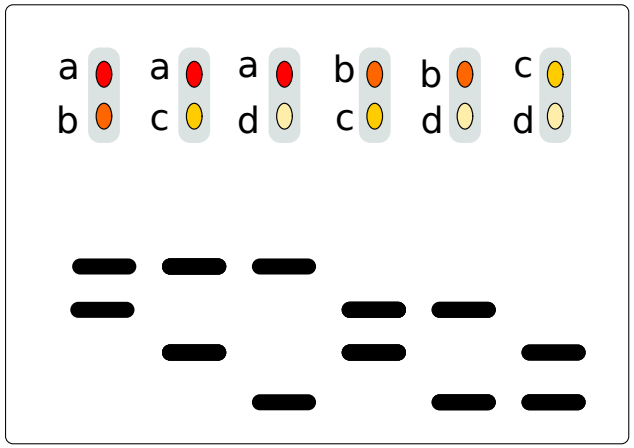
```
CTGACTTCAGGTACTTCTTCTTCTTCTTCTTCTTCTTATGCCGTGATTGATC
GACTGAAGTCCATGAAGAAGAAGAAGAAGAAGAAGAATACGGCACTAACTAG
```

c - 7 repeats

```
CTGACTTCAGGTACTTCTTCTTCTTCTTCTTCTTATGCCGTGATTGATC
GACTGAAGTCCATGAAGAAGAAGAAGAAGAAGAATACGGCACTAACTAG
```

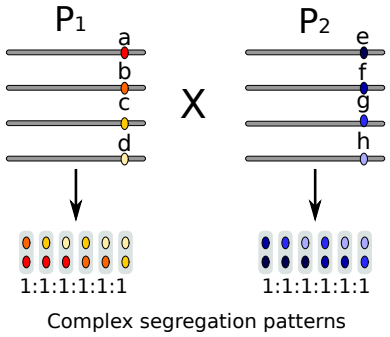
d - 6 repeats

```
CTGACTTCAGGTACTTCTTCTTCTTCTTCTTATGCCGTGATTGATC
GACTGAAGTCCATGAAGAAGAAGAAGAAGAATACGGCACTAACTAG
```



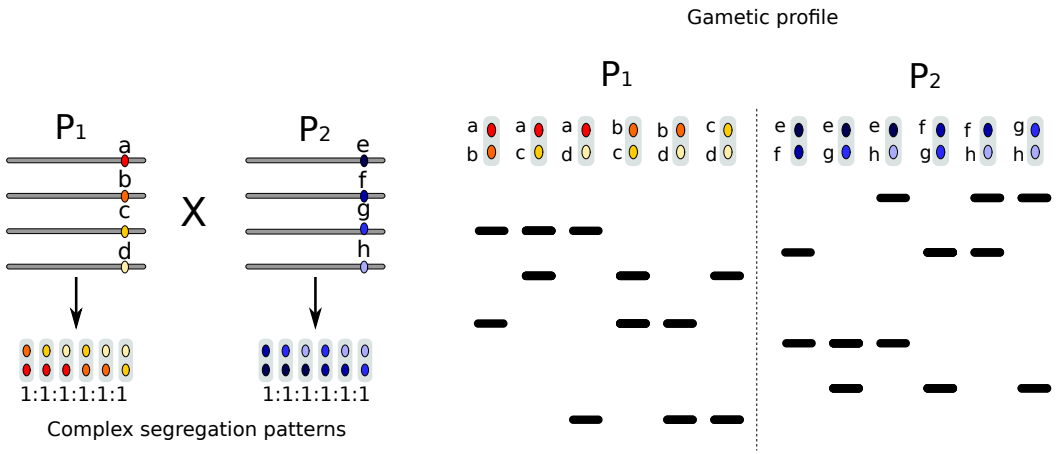
Codominant genotyping in autotetraploids

Tetraploid example - multiallelic



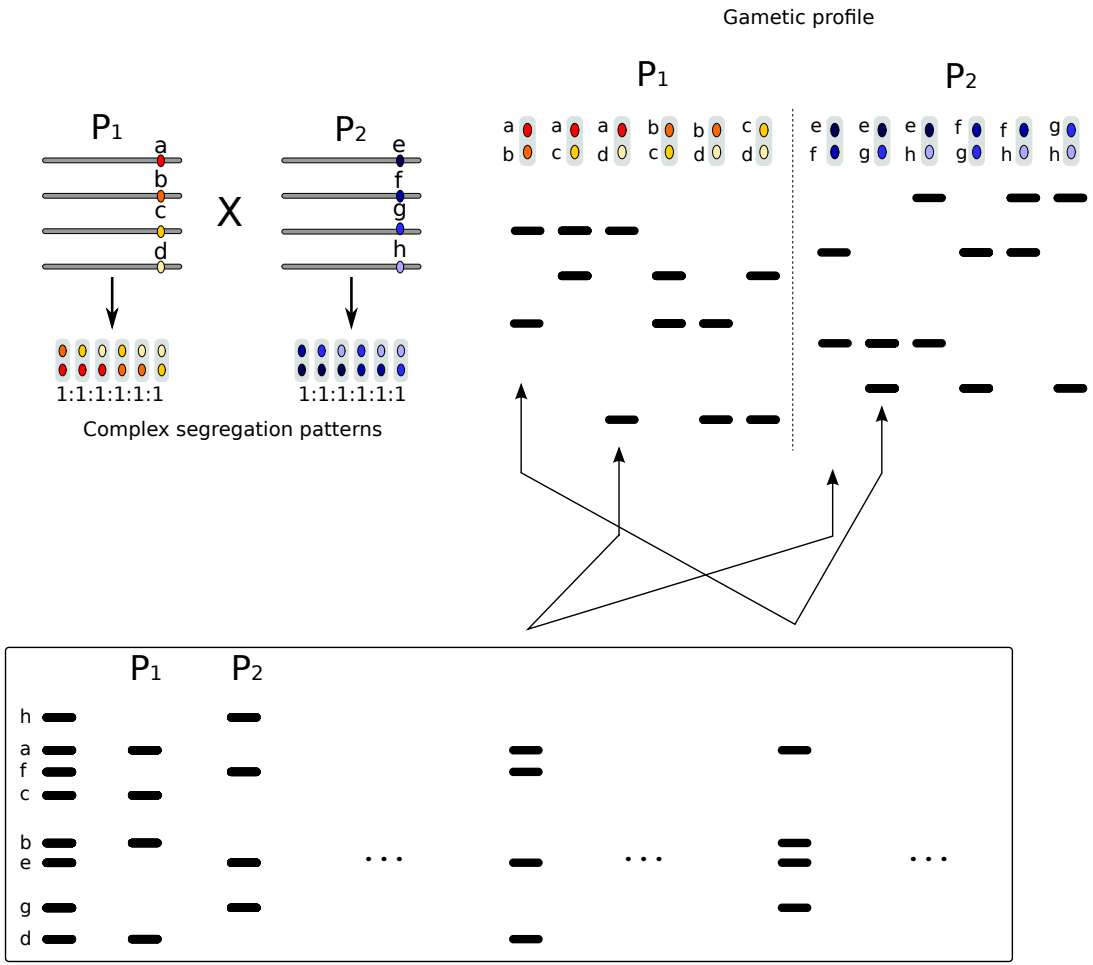
Codominant genotyping in autotetraploids

Tetraploid example - multiallelic



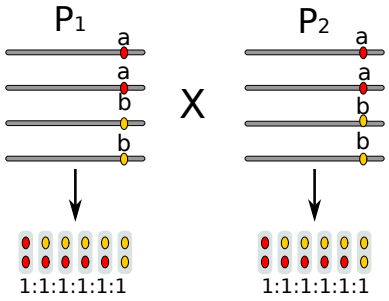
Codominant genotyping in autotetraploids

Tetraploid example - multiallelic



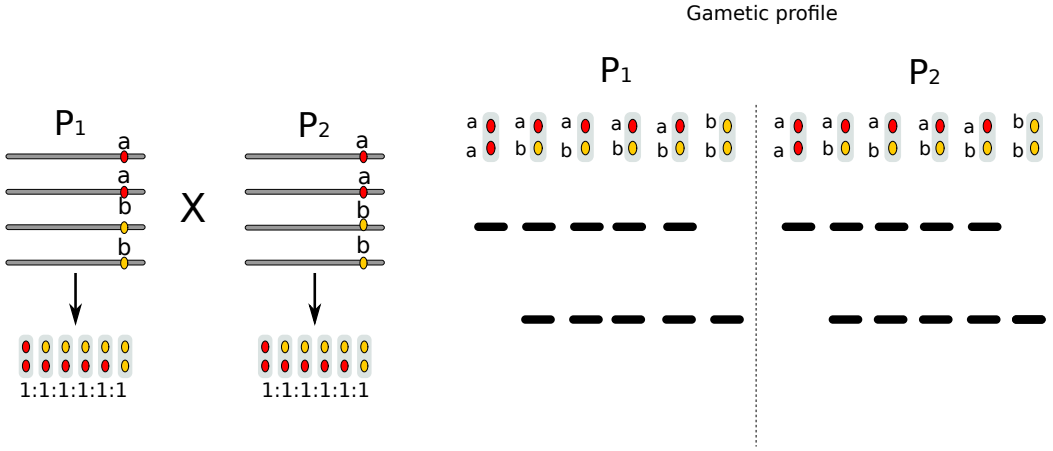
Codominant genotyping in autotetraploids

Tetraploid example - biallelic



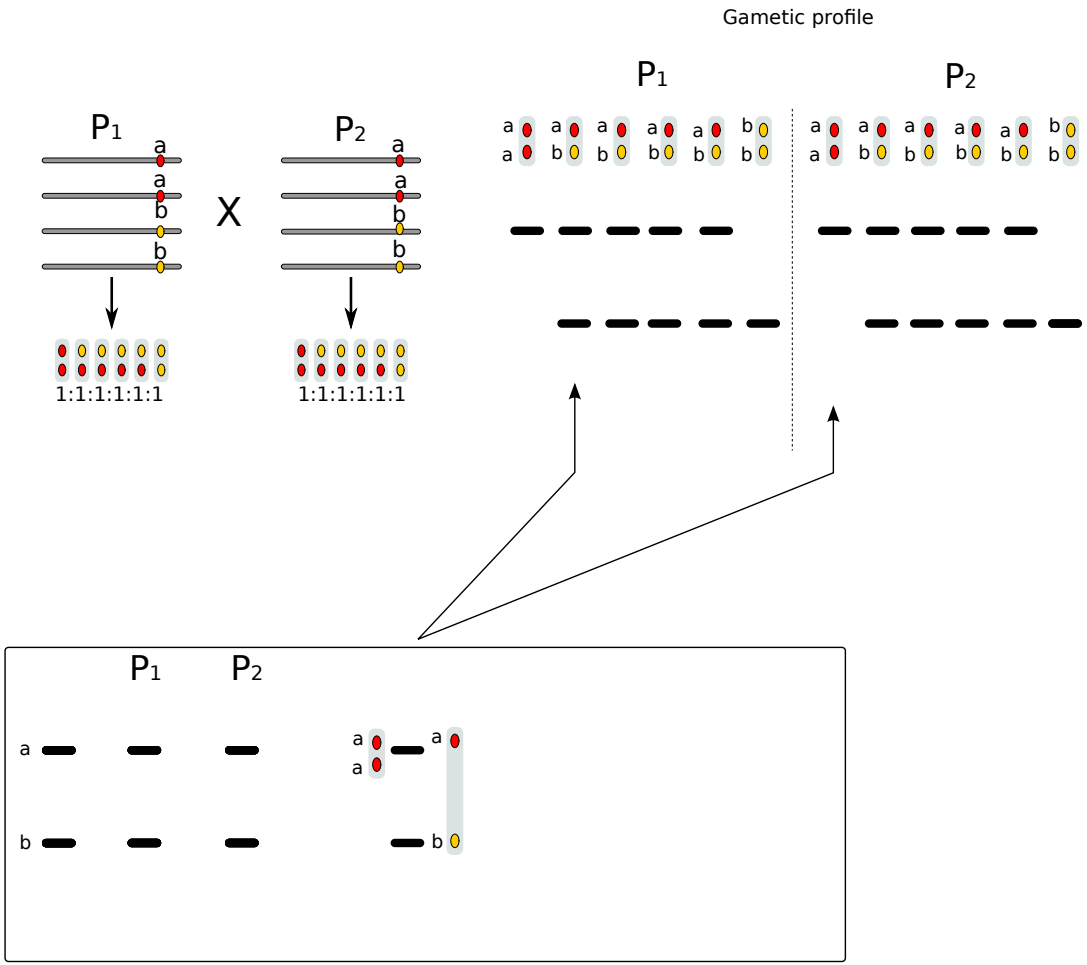
Codominant genotyping in autotetraploids

Tetraploid example - biallelic



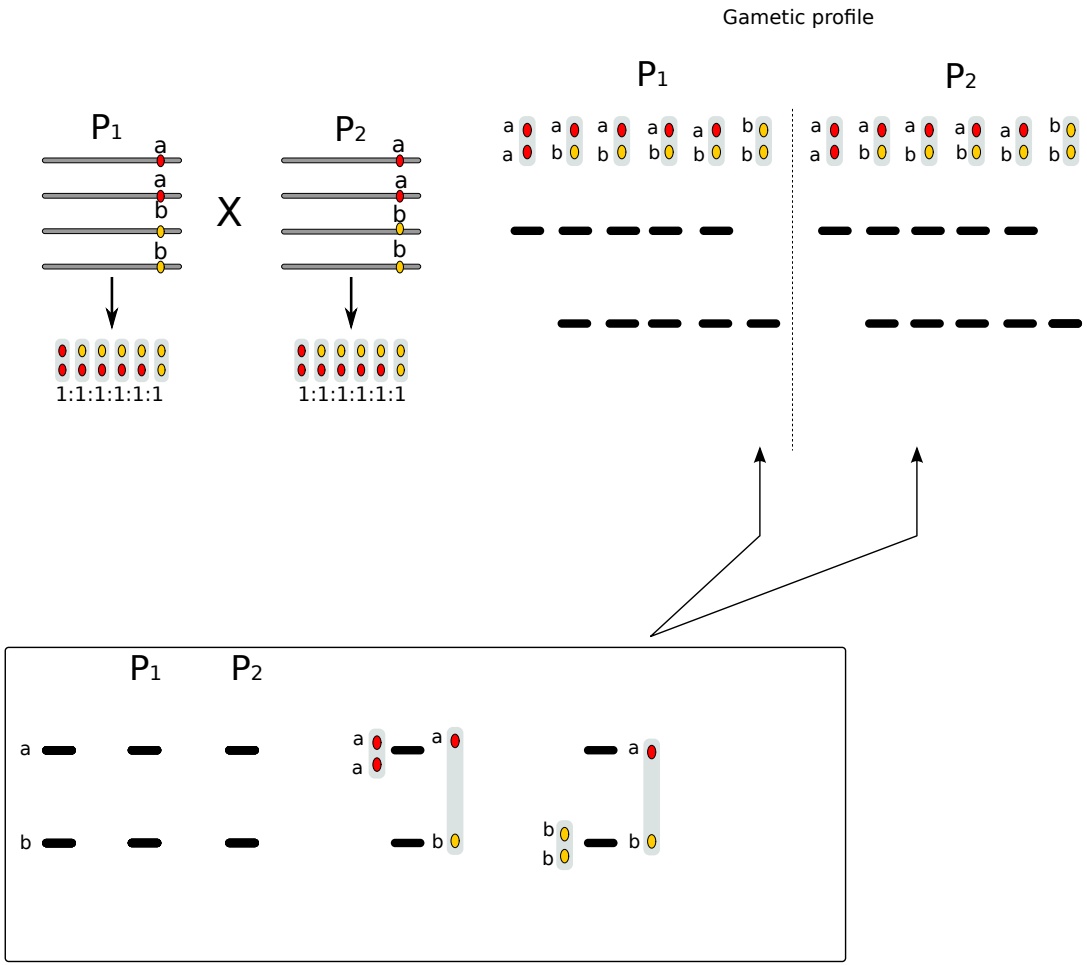
Codominant genotyping in autotetraploids

Tetraploid example - biallelic



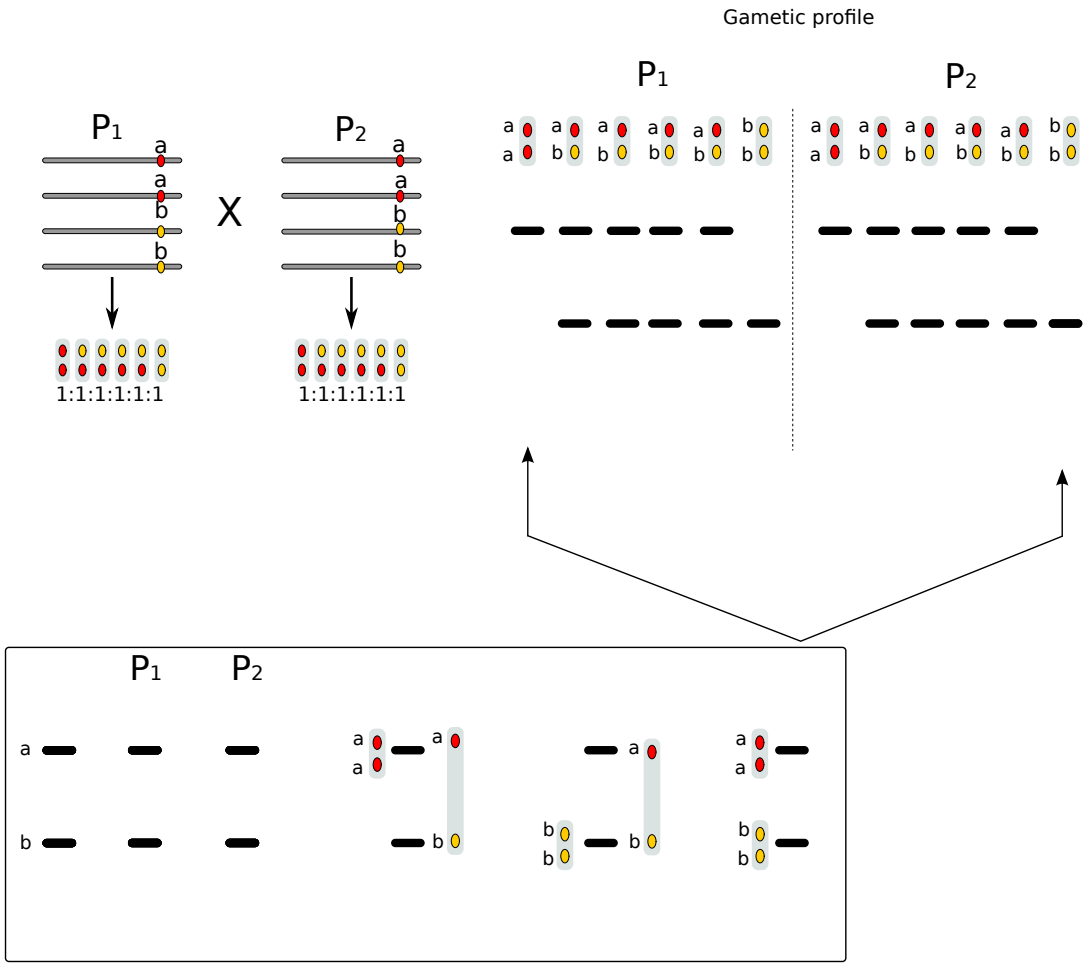
Codominant genotyping in autotetraploids

Tetraploid example - biallelic



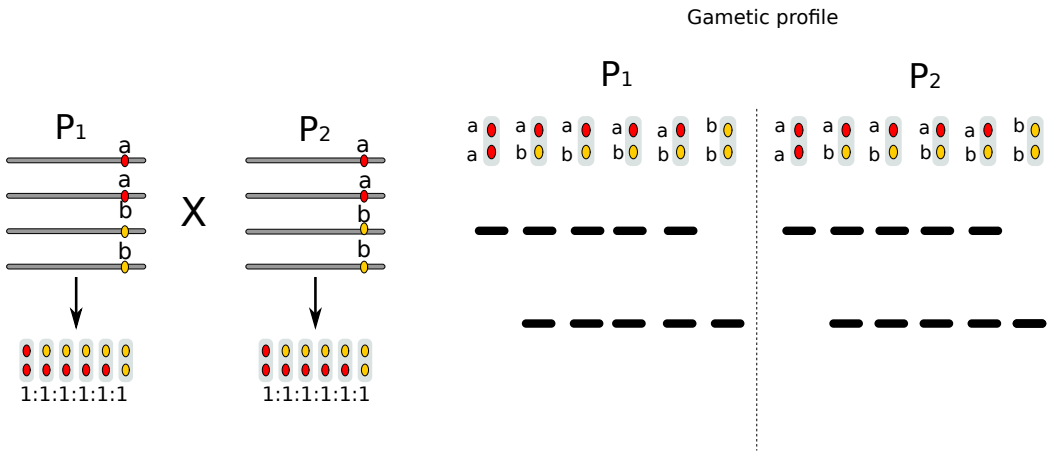
Codominant genotyping in autotetraploids

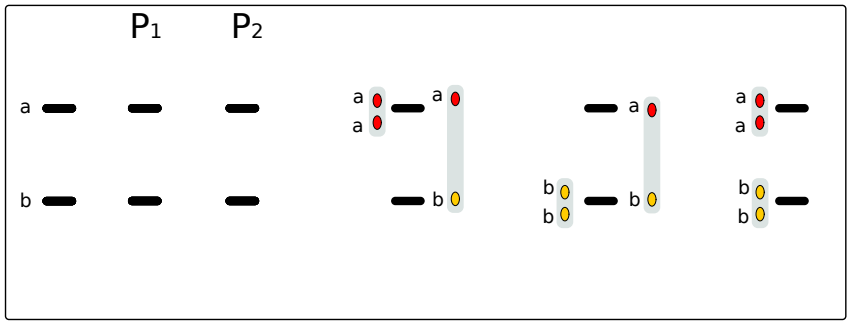
Tetraploid example - biallelic



Codominant genotyping in autotetraploids

Tetraploid example - biallelic





Possible outcomes

aaaa
aaab
aabb
abbb
bbbb

Codominant genotyping in autotetraploids

Question:

How can we use microssatelites (and similar markers) in complex polyploids?

Outline

Quantitative genotyping

- Basic idea: measure the abundance of "A" and "a" in a sample based on:

fluorescence

Mass

- MALDI-TOF Mass spectrometry: more precise than fluorescence-based techniques
- We use Sequenom's MassARRAY® with iPLEX™ biochemistry.

Quantitative genotyping

- Basic idea: measure the abundance of "A" and "a" in a sample based on:

fluorescence

Mass

- MALDI-TOF Mass spectrometry: more precise than fluorescence-based techniques
- We use Sequenom's MassARRAY® with iPLEX™ biochemistry.

Quantitative genotyping

- Basic idea: measure the abundance of "A" and "a" in a sample based on:

fluorescence

Mass

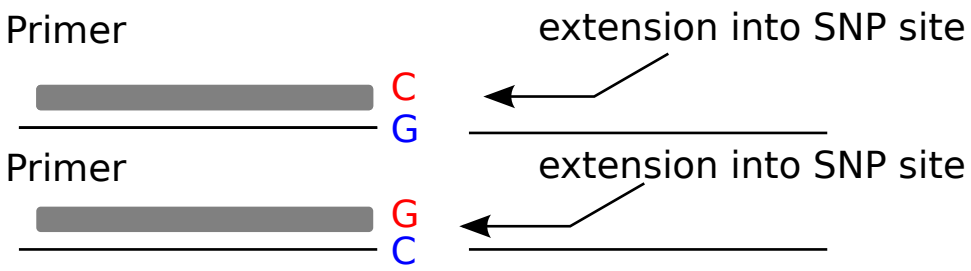
- MALDI-TOF Mass spectrometry: more precise than fluorescence-based techniques
- We use Sequenom's MassARRAY® with iPLEX™ biochemistry.

Sequenom's MassARRAY.

- A first PCR is carried out using a capture primer.
- Specific single base primers are designed.
- Extension reactions are carried out in a allele-specific manner using mass-modified terminator nucleotides so all four bases can be resolved on the basis of their mass.

Sequenom's MassARRAY.

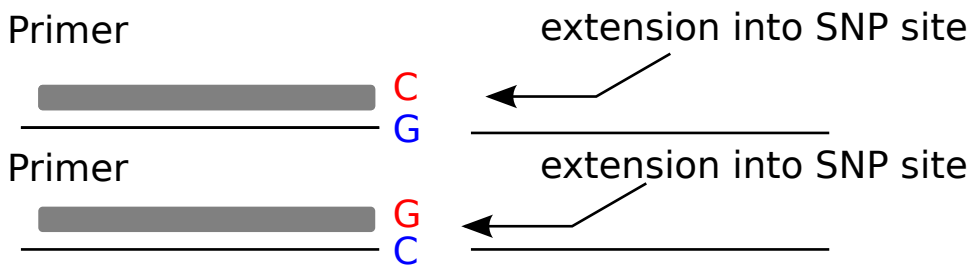
- A first PCR is carried out using a capture primer.
- Specific single base primers are designed.



- Extension reactions are carried out in a allele-specific manner using mass-modified terminator nucleotides so all four bases can be resolved on the basis of their mass.

Sequenom's MassARRAY.

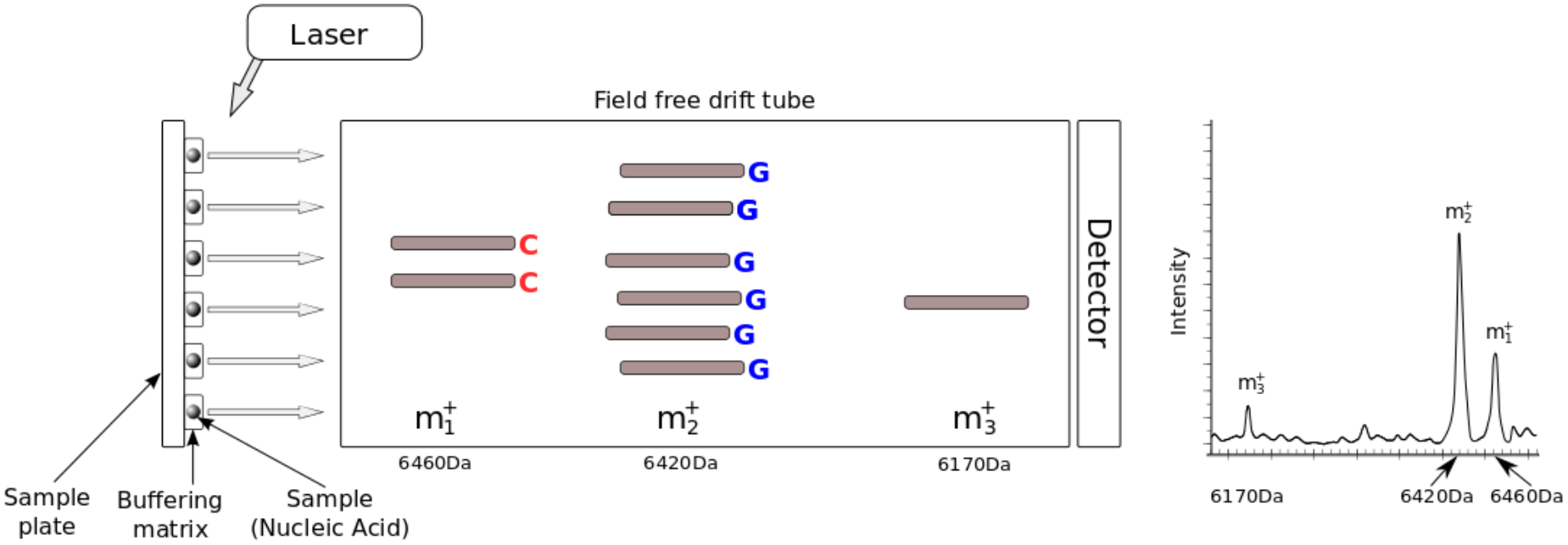
- A first PCR is carried out using a capture primer.
- Specific single base primers are designed.



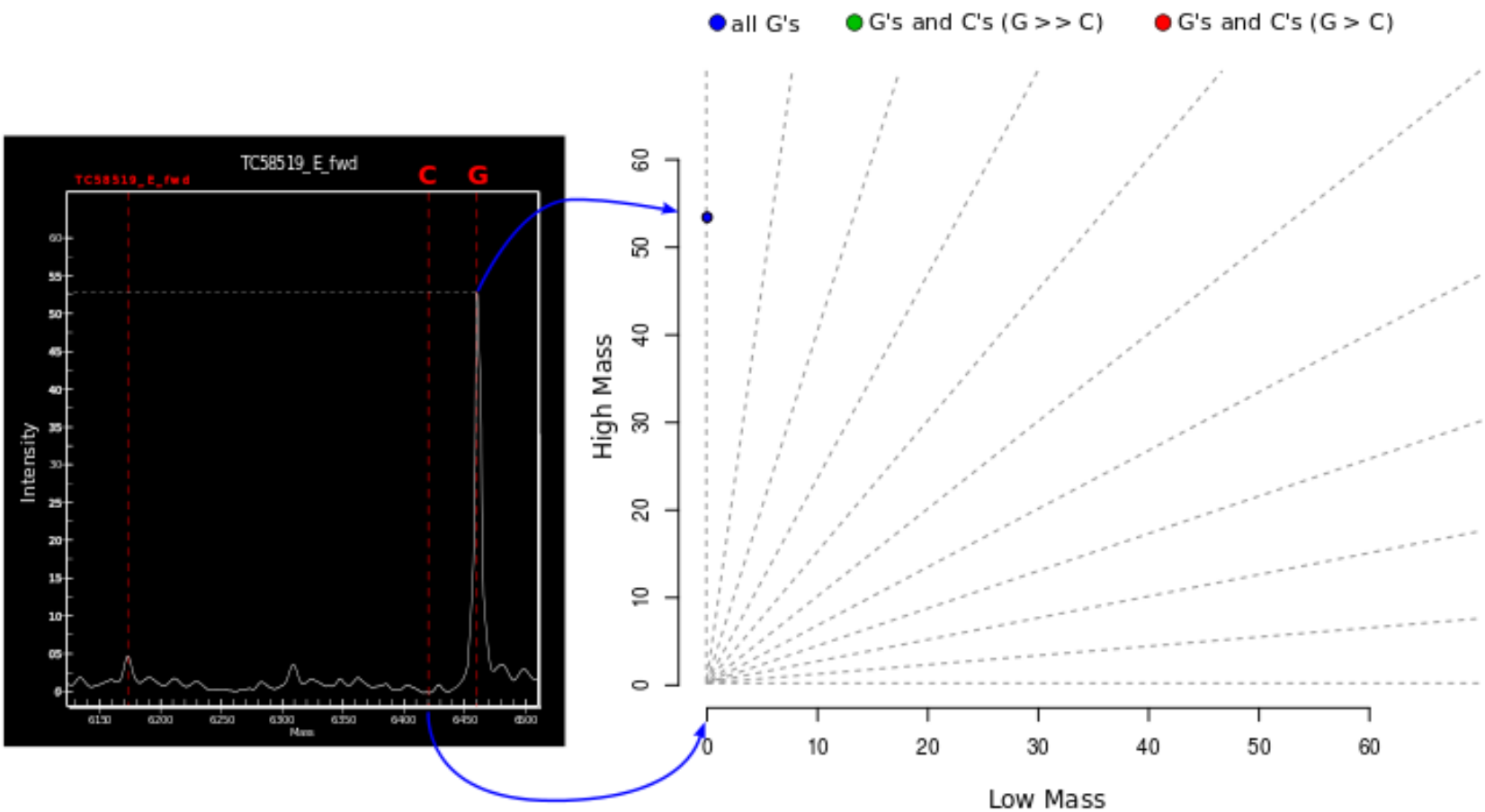
- Extension reactions are carried out in a allele-specific manner using mass-modified terminator nucleotides so all four bases can be resolved on the basis of their mass.

Sequenom's MassARRAY.

- Primer extension products masses (in daltons), and their allele ratios are resolved with a MALDI-TOF system.

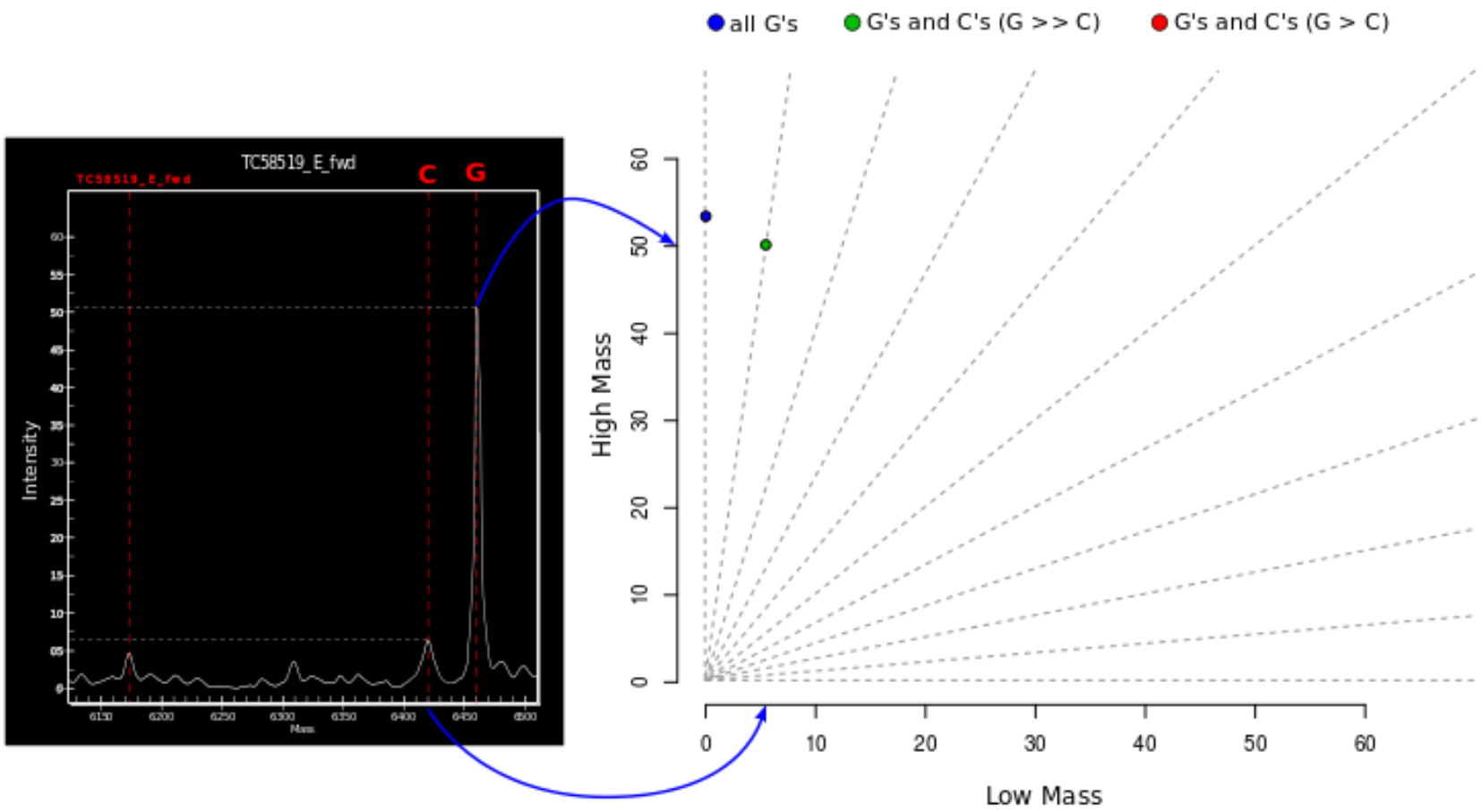


Scatter plot



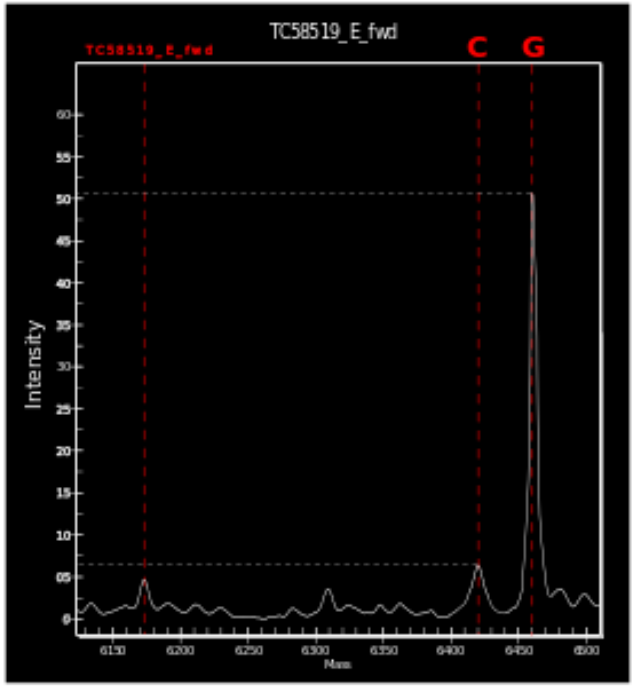
- Ratio between peaks or areas (angles): dosage

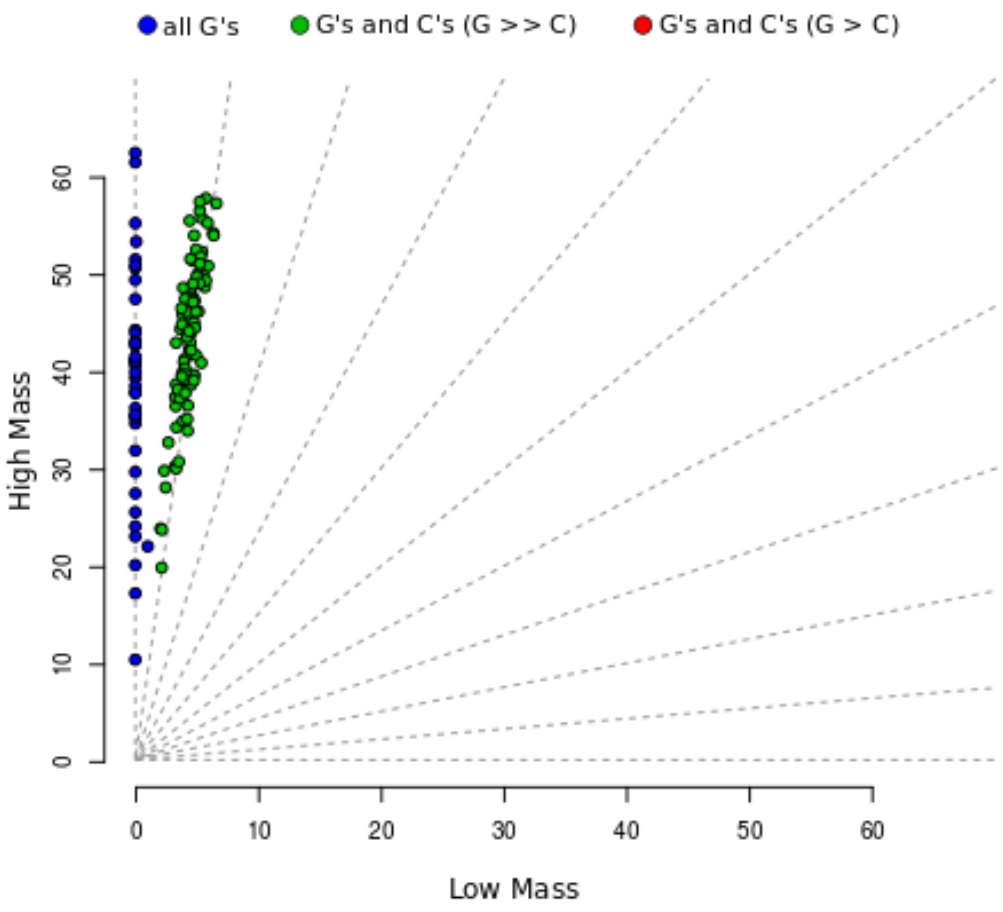
Scatter plot



- Ratio between peaks or areas (angles): dosage

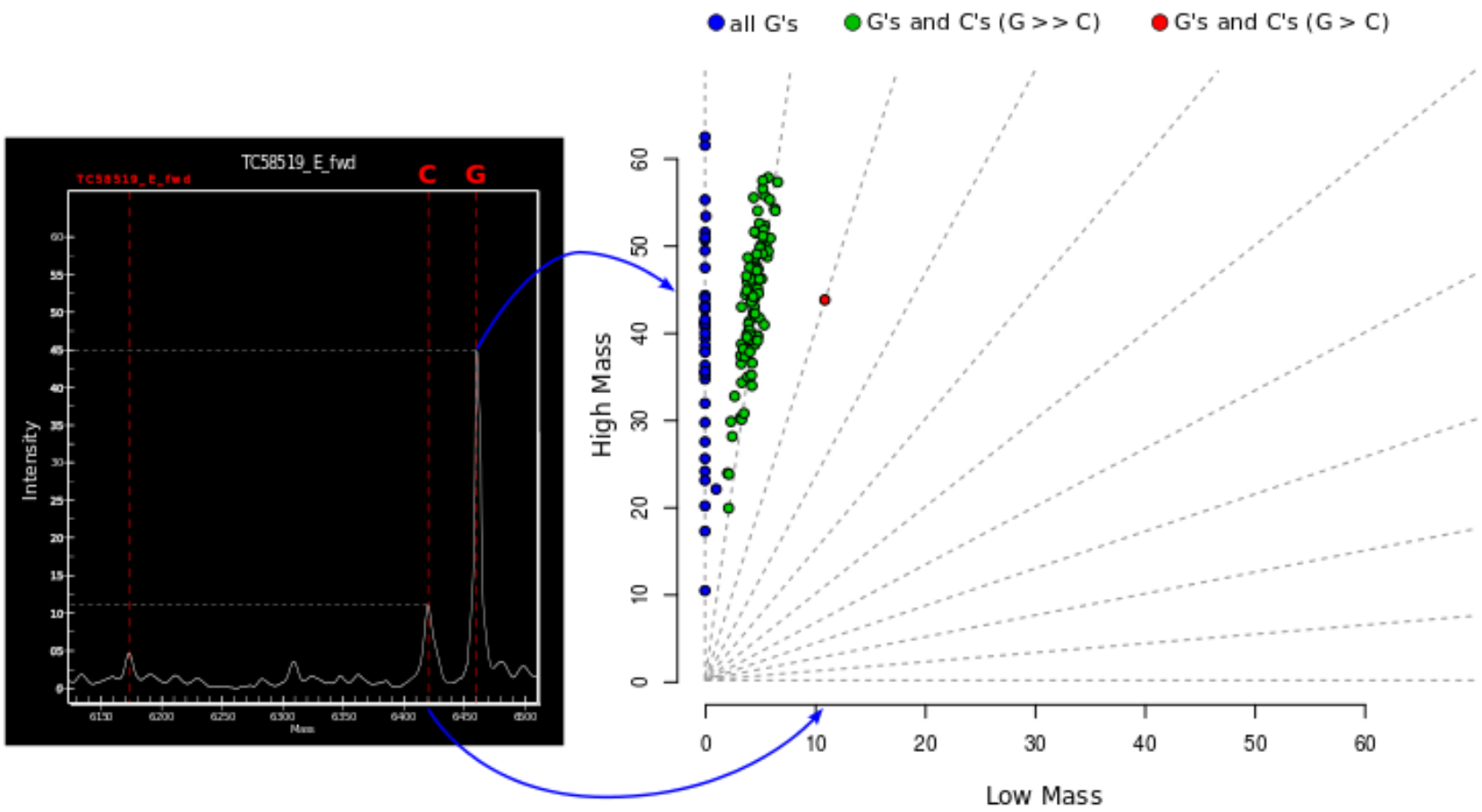
Scatter plot





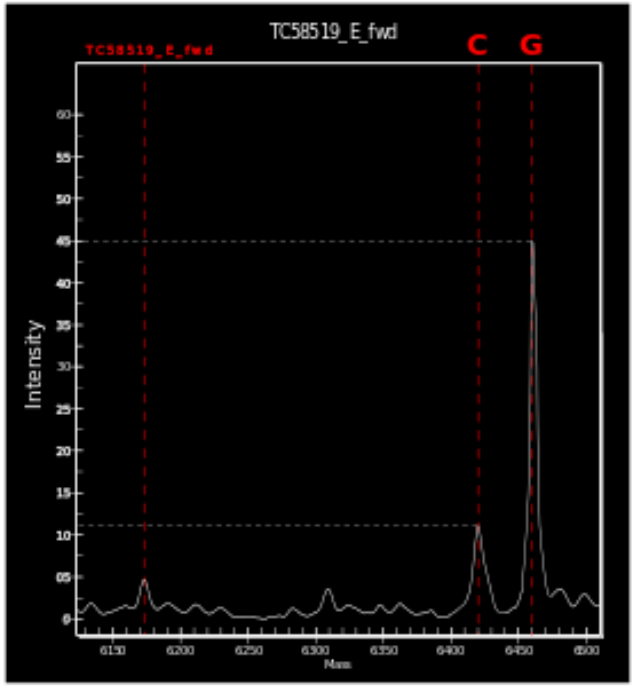
- Ratio between peaks or areas (angles): dosage

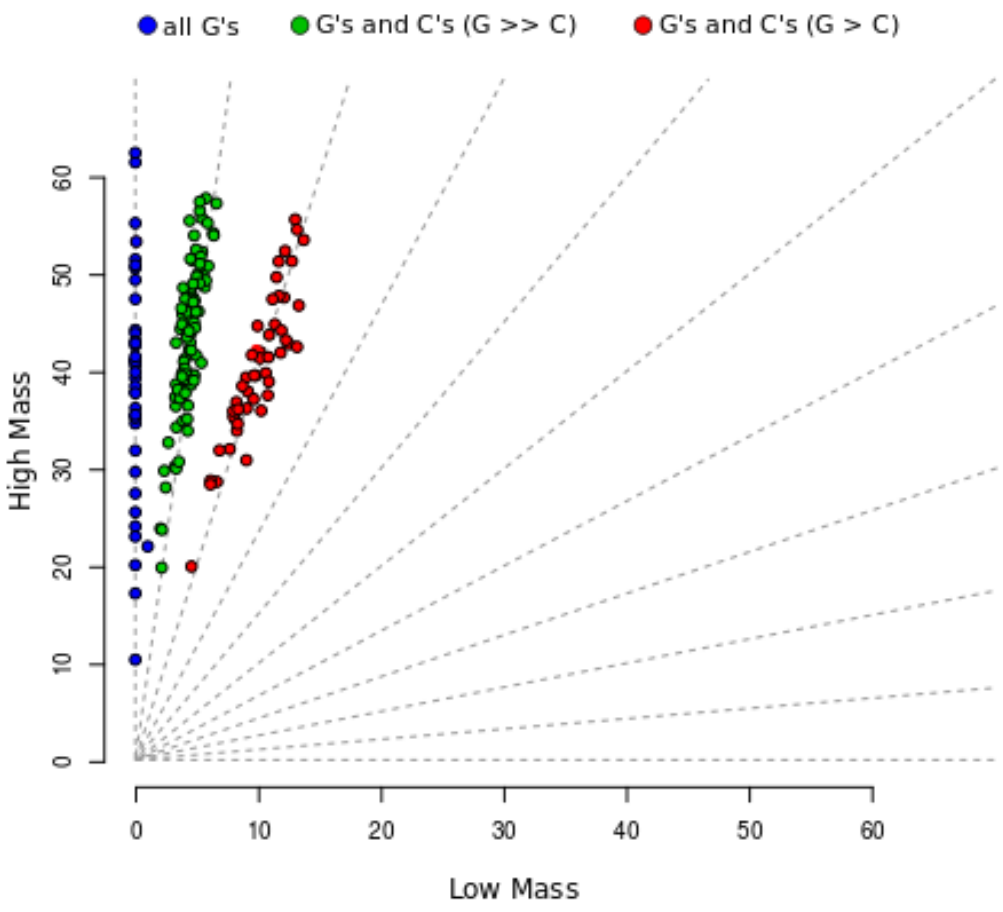
Scatter plot



- Ratio between peaks or areas (angles): dosage

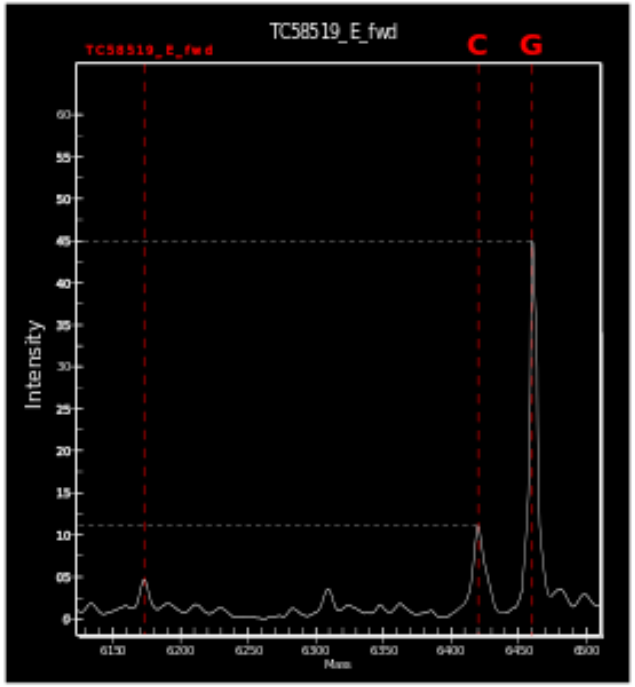
Scatter plot

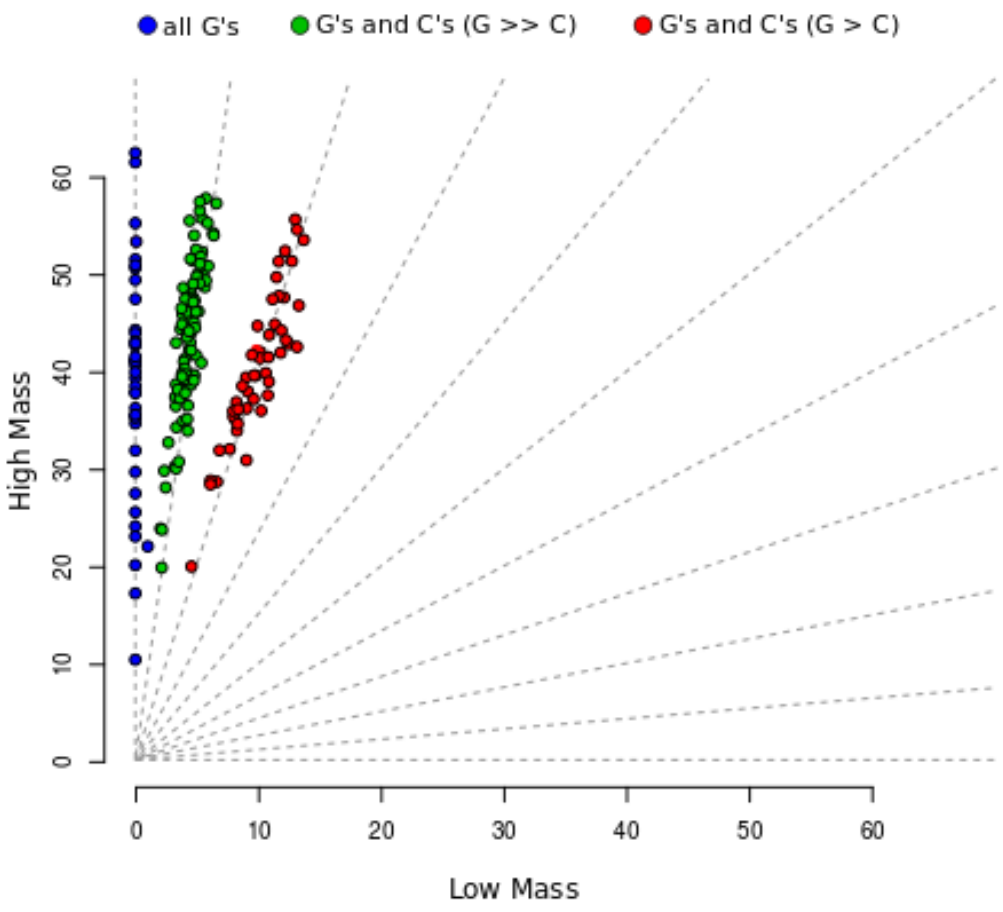




- Ratio between peaks or areas (angles): dosage

Scatter plot

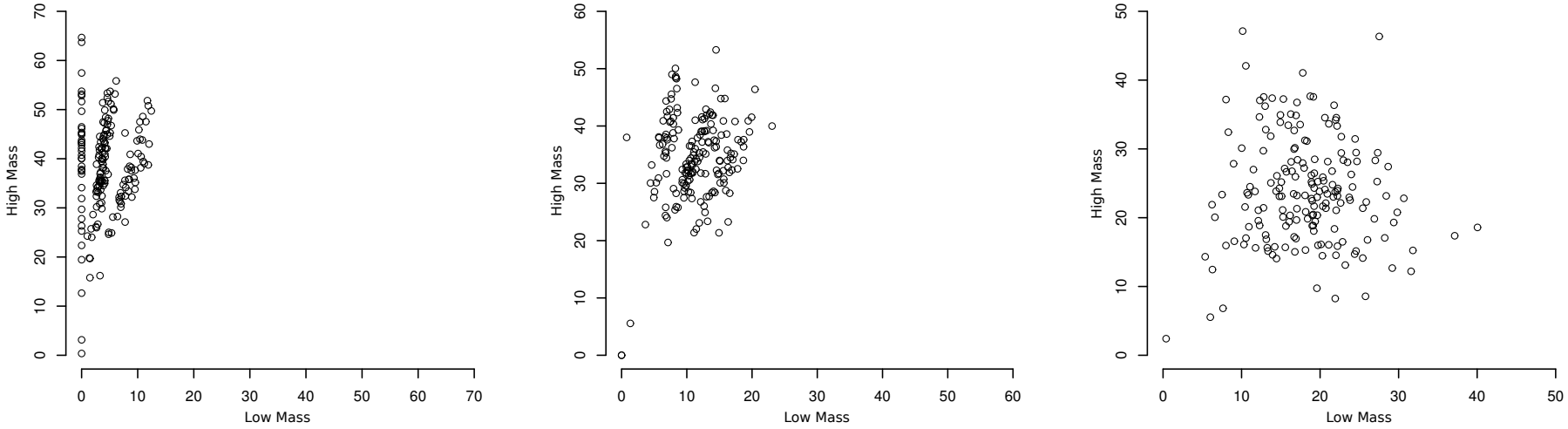




- Ratio between peaks or areas (angles): dosage

Real Data, F_1

- How many clusters?
- What are the angles (dosages) and proportions?
- How to allocate the individuals?

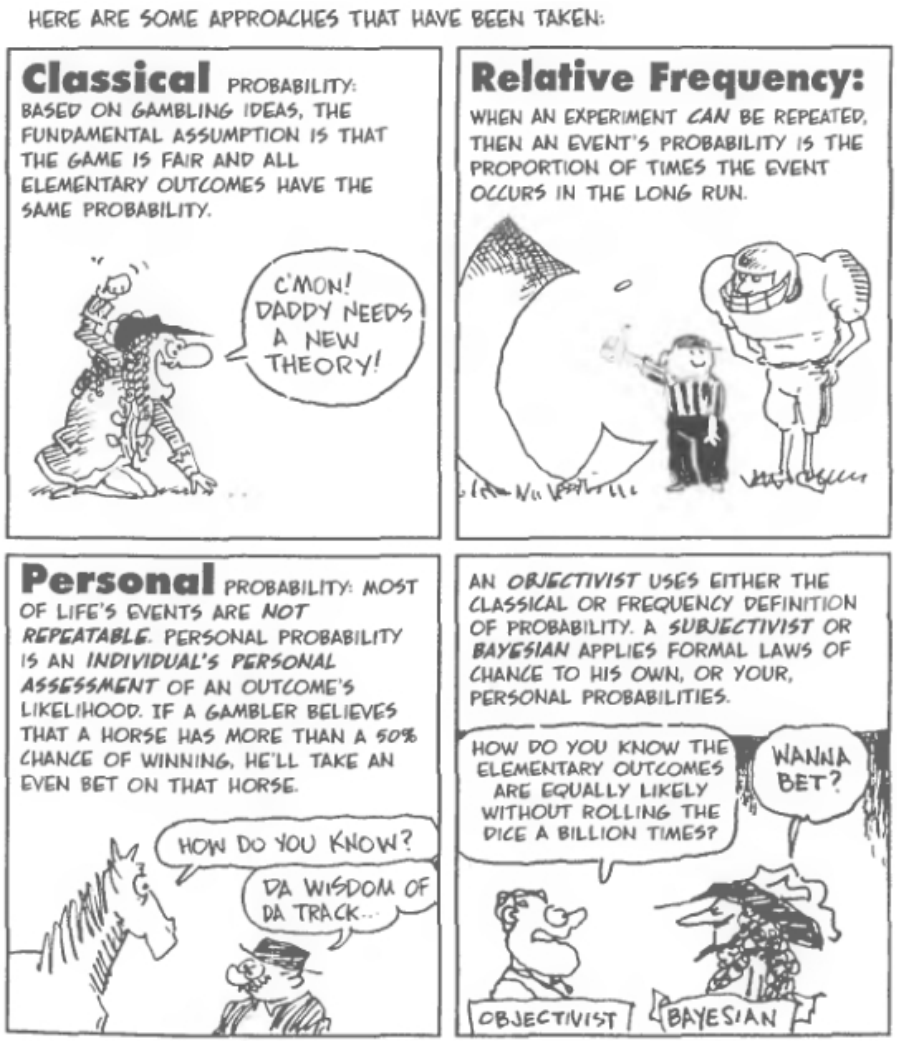


Problem

- Functions of ploidy and dosage, that are unknown!

Outline

Definições


HERE ARE SOME APPROACHES THAT HAVE BEEN TAKEN:
Classical PROBABILITY: BASED ON GAMBLING IDEAS, THE FUNDAMENTAL ASSUMPTION IS THAT THE GAME IS FAIR AND ALL ELEMENTARY OUTCOMES HAVE THE SAME PROBABILITY.
C'MON! DADDY NEEDS A NEW THEORY!
Relative Frequency: WHEN AN EXPERIMENT CAN BE REPEATED, THEN AN EVENT'S PROBABILITY IS THE PROPORTION OF TIMES THE EVENT OCCURS IN THE LONG RUN.
Personal PROBABILITY: MOST OF LIFE'S EVENTS ARE NOT REPEATABLE. PERSONAL PROBABILITY IS AN INDIVIDUAL'S PERSONAL ASSESSMENT OF AN OUTCOME'S LIKELIHOOD. IF A GAMBLER BELIEVES THAT A HORSE HAS MORE THAN A 50% CHANCE OF WINNING, HE'LL TAKE AN EVEN BET ON THAT HORSE.
HOW DO YOU KNOW?
DA WISDOM OF DA TRACK...
AN OBJECTIVIST USES EITHER THE CLASSICAL OR FREQUENCY DEFINITION OF PROBABILITY. A SUBJECTIVIST OR BAYESIAN APPLIES FORMAL LAWS OF CHANCE TO HIS OWN, OR YOUR, PERSONAL PROBABILITIES.
HOW DO YOU KNOW THE ELEMENTARY OUTCOMES ARE EQUALLY LIKELY WITHOUT ROLLING THE DICE A BILLION TIMES?
WANNA BET?
OBJECTIVIST
BAYESIAN

Regras

- Adição

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

- Adição (eventos mutuamente exclusivos)

$$P(A \text{ ou } B) = P(A) + P(B)$$

- Subtração

$$P(A) = 1 - P(\text{não } A)$$

- Multiplicação

$$P(A \text{ e } B) = P(A) \times P(B|A)$$

- Multiplicação (A e B independentes)

$$P(A \text{ e } B) = P(A) \times P(B)$$

- Notação: $P(A \text{ e } B) = P(A \cap B) = P(A, B)$

Regras

- Adição

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

- Adição (eventos mutuamente exclusivos)

$$P(A \text{ ou } B) = P(A) + P(B)$$

- Subtração

$$P(A) = 1 - P(\text{não } A)$$

- Multiplicação

$$P(A \text{ e } B) = P(A) \times P(B|A)$$

- Multiplicação (A e B independentes)

$$P(A \text{ e } B) = P(A) \times P(B)$$

- Notação: $P(A \text{ e } B) = P(A \cap B) = P(A, B)$

Regras

- Adição

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

- Adição (eventos mutuamente exclusivos)

$$P(A \text{ ou } B) = P(A) + P(B)$$

- Subtração

$$P(A) = 1 - P(\text{não } A)$$

- Multiplicação

$$P(A \text{ e } B) = P(A) \times P(B|A)$$

- Multiplicação (A e B independentes)

$$P(A \text{ e } B) = P(A) \times P(B)$$

- Notação: $P(A \text{ e } B) = P(A \cap B) = P(A, B)$

Regras

- Adição

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

- Adição (eventos mutuamente exclusivos)

$$P(A \text{ ou } B) = P(A) + P(B)$$

- Subtração

$$P(A) = 1 - P(\text{não } A)$$

- Multiplicação

$$P(A \text{ e } B) = P(A) \times P(B|A)$$

- Multiplicação (A e B independentes)

$$P(A \text{ e } B) = P(A) \times P(B)$$

- Notação: $P(A \text{ e } B) = P(A \cap B) = P(A, B)$

Regras

- Adição

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

- Adição (eventos mutuamente exclusivos)

$$P(A \text{ ou } B) = P(A) + P(B)$$

- Subtração

$$P(A) = 1 - P(\text{não } A)$$

- Multiplicação

$$P(A \text{ e } B) = P(A) \times P(B|A)$$

- Multiplicação (A e B independentes)

$$P(A \text{ e } B) = P(A) \times P(B)$$

- Notação: $P(A \text{ e } B) = P(A \cap B) = P(A, B)$

Regras

- Adição

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

- Adição (eventos mutuamente exclusivos)

$$P(A \text{ ou } B) = P(A) + P(B)$$

- Subtração

$$P(A) = 1 - P(\text{não } A)$$

- Multiplicação

$$P(A \text{ e } B) = P(A) \times P(B|A)$$

- Multiplicação (A e B independentes)

$$P(A \text{ e } B) = P(A) \times P(B)$$

- Notação: $P(A \text{ e } B) = P(A \cap B) = P(A, B)$

Probabilidade Condicional

Dois dados com cores diferentes

- Se eu jogar os dois dados simultaneamente, qual é a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: $6 \times 6 = 36$
  - $\#$ resultados com soma 3: 2 ($\{1, 2\}$, $\{2, 1\}$)
  - Resp: $P(\text{soma } 3) = 2/36$

Probabilidade Condicional

Dois dados com cores diferentes

- Se eu jogar os dois dados simultaneamente, qual é a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: $6 \times 6 = 36$
  - $\#$ resultados com soma 3: 2 ($\{1, 2\}$, $\{2, 1\}$)
  - Resp: $P(\text{soma } 3) = 2/36$

Probabilidade Condicional

Dois dados com cores diferentes

- Se eu jogar os dois dados simultaneamente, qual é a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: $6 \times 6 = 36$
  - $\#$ resultados com soma 3: 2 ($\{1, 2\}$, $\{2, 1\}$)
  - Resp: $P(\text{soma } 3) = 2/36$

Probabilidade Condicional

Dois dados com cores diferentes

- Se eu jogar os dois dados simultaneamente, qual é a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: $6 \times 6 = 36$
  - $\#$ resultados com soma 3: 2 ($\{1, 2\}$, $\{2, 1\}$)
  - Resp: $P(\text{soma } 3) = 2/36$

Probabilidade Condicional

Dois dados com cores diferentes

- Suponha agora que um dos dois dados foi jogado antes, e o resultado foi 1
- Qual a probabilidade de obter soma 3?
  - Quantos possíveis
  - Quantos são soma 3 (1,2)
  - Logo P(soma 3 dado que um dos dados é 1) = 1/6

Probabilidade Condicional

Dois dados com cores diferentes

- Suponha agora que um dos dois dados foi jogado antes, e o resultado foi 1
- Qual a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: 6
  - $\#$ resultados com soma 3: 1 ($\{1, 2\}$)
  - Resp: $P(\text{soma } 3 | \text{valor } 1 \text{ em um dos dados}) = 1/6$

Probabilidade Condicional

Dois dados com cores diferentes

- Suponha agora que um dos dois dados foi jogado antes, e o resultado foi 1
- Qual a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: 6
  - $\#$ resultados com soma 3: 1 ($\{1, 2\}$)
  - Resp: $P(\text{soma } 3 | \text{valor } 1 \text{ em um dos dados}) = 1/6$

Probabilidade Condicional

Dois dados com cores diferentes

- Suponha agora que um dos dois dados foi jogado antes, e o resultado foi 1
- Qual a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: 6
  - $\#$ resultados com soma 3: 1 ($\{1, 2\}$)
  - Resp: $P(\text{soma } 3 | \text{valor } 1 \text{ em um dos dados}) = 1/6$

Probabilidade Condicional

Dois dados com cores diferentes

- Suponha agora que um dos dois dados foi jogado antes, e o resultado foi 1
- Qual a probabilidade de obter soma 3?
  - $\#$ resultados possíveis: 6
  - $\#$ resultados com soma 3: 1 ($\{1, 2\}$)
  - Resp: $P(\text{soma } 3 | \text{valor 1 em um dos dados}) = 1/6$

Probabilidade Condicional

$P(A|B)$

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

Atenção

- Note a relação entre probab. condicional e a regra da multiplicação
- O que significam $P(A|B) = 1$ e $P(A|B) = 0$?
- Eventos independentes: $P(A,B) = P(A) \times P(B)$

Exemplo anterior

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

Probabilidade Condicional

$P(A|B)$

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

Atenção

- Note a relação entre probab. condicional e a regra da multiplicação
- O que significam $P(A|B) = 1$ e $P(A|B) = 0$?
- Eventos independentes: $P(A,B) = P(A) \times P(B)$

Exemplo anterior

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

Probabilidade Condicional

$P(A|B)$

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

Atenção

- Note a relação entre probab. condicional e a regra da multiplicação
- O que significam $P(A|B) = 1$ e $P(A|B) = 0$?
- Eventos independentes: $P(A,B) = P(A) \times P(B)$

Exemplo anterior

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

Eventos independentes

Moeda "honesta"

- Qual a probabilidade de obter uma sequência de 4 caras?
- Resp: $\left(\frac{1}{2}\right)^4$

Eventos independentes

Moeda "honesta"

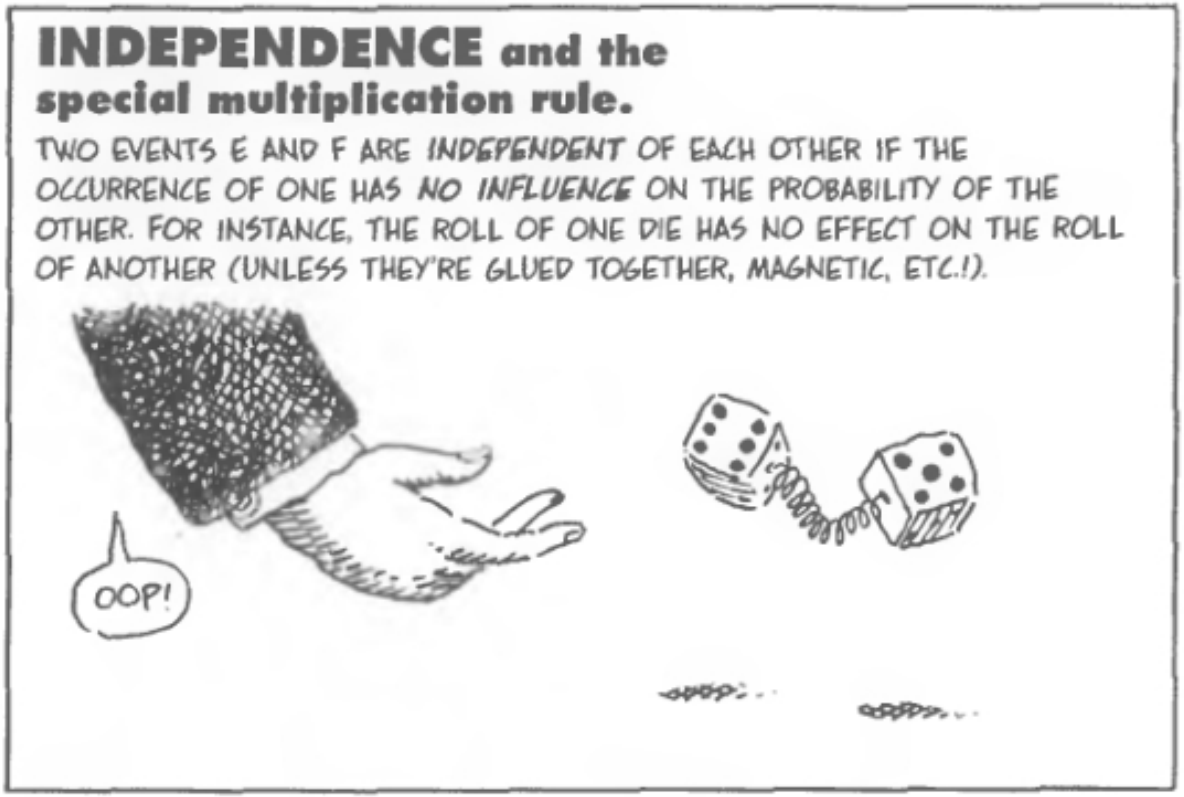
- Qual a probabilidade de obter uma sequência de 4 caras?
- Resp: $\left(\frac{1}{2}\right)^4$

Eventos independentes

Moeda “honesta”

Qual a probabilidade de obter uma sequência de 4 caras?

Eventos independentes



Um caso simples

Doença, Genótipo

	mm	Mm	MM	
R	0.10	0.21	0.47	0.78
S	0.05	0.09	0.08	0.22
	0.15	0.30	0.55	1

- $P(D = R) = 0.78$
- $P(G = Mm) = 0.30$
- $P(D = R|G = MM) = \frac{P(D=R,G=MM)}{P(G=MM)} = \frac{0.47}{0.55} = 0.85$
- $P(D = R, G = MM) = P(D = R)\, P(G = MM|D = R) = 0.78 \times \frac{0.47}{0.78} = 0.47$
- Note que $P(D = R).P(G = MM) = 0.78 \times 0.55 = 0.429$

Um caso simples

Doença, Genótipo

	mm	Mm	MM	
R	0.10	0.21	0.47	0.78
S	0.05	0.09	0.08	0.22
	0.15	0.30	0.55	1

- $P(D = R) = 0.78$
- $P(G = Mm) = 0.30$
- $P(D = R|G = MM) = \frac{P(D=R,G=MM)}{P(G=MM)} = \frac{0.47}{0.55} = 0.85$
- $P(D = R, G = MM) = P(D = R)\, P(G = MM|D = R) = 0.78 \times \frac{0.47}{0.78} = 0.47$
- Note que $P(D = R).P(G = MM) = 0.78 \times 0.55 = 0.429$

Um caso simples

Doença, Genótipo

	mm	Mm	MM	
R	0.10	0.21	0.47	0.78
S	0.05	0.09	0.08	0.22
	0.15	0.30	0.55	1

- $P(D = R) = 0.78$
- $P(G = Mm) = 0.30$
- $P(D = R|G = MM) = \frac{P(D=R,G=MM)}{P(G=MM)} = \frac{0.47}{0.55} = 0.85$
- $P(D = R, G = MM) = P(D = R)\, P(G = MM|D = R) = 0.78 \times \frac{0.47}{0.78} = 0.47$
- Note que $P(D = R).P(G = MM) = 0.78 \times 0.55 = 0.429$

Um caso simples

Doença, Genótipo

	mm	Mm	MM	
R	0.10	0.21	0.47	0.78
S	0.05	0.09	0.08	0.22
	0.15	0.30	0.55	1

- $P(D = R) = 0.78$
- $P(G = Mm) = 0.30$
- $P(D = R|G = MM) = \frac{P(D=R,G=MM)}{P(G=MM)} = \frac{0.47}{0.55} = 0.85$
- $P(D = R, G = MM) = P(D = R)\, P(G = MM|D = R) = 0.78 \times \frac{0.47}{0.78} = 0.47$
- Note que $P(D = R).P(G = MM) = 0.78 \times 0.55 = 0.429$

Um caso simples

Doença, Genótipo

	mm	Mm	MM	
R	0.10	0.21	0.47	0.78
S	0.05	0.09	0.08	0.22
	0.15	0.30	0.55	1

- $P(D = R) = 0.78$
- $P(G = Mm) = 0.30$
- $P(D = R|G = MM) = \frac{P(D=R,G=MM)}{P(G=MM)} = \frac{0.47}{0.55} = 0.85$
- $P(D = R, G = MM) = P(D = R)\, P(G = MM|D = R) = 0.78 \times \frac{0.47}{0.78} = 0.47$
- Note que $P(D = R).P(G = MM) = 0.78 \times 0.55 = 0.429$

Teorema de Bayes

Thomas Bayes, 1701–1761

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

- $P(A)$: "priori"
- $P(A|B)$: "posteriori"
- $P(B|A)/P(B)$: suporte que B fornece para A

Um exemplo mais completo

Doença, Genótipo, Temperatura, Umidade

- Temperatura T: A, B (alta, baixa)
- Umidade U: S, U (seco, úmido)
- Genótipo G: mm, Mm, MM
- Doença D: R, Su (resistente, suscetível)
- É relevante calcular

$$P(T, U, G, D) = P(T)\, P(U|T)\, P(G|T, U)\, P(D|T, U, G)$$

- (Regra da cadeia)
- 23 parâmetros

Um exemplo mais completo

Doença, Genótipo, Temperatura, Umidade

- Temperatura T: A, B (alta, baixa)
- Umidade U: S, U (seco, úmido)
- Genótipo G: mm, Mm, MM
- Doença D: R, Su (resistente, suscetível)
- É relevante calcular

$$P(T, U, G, D) = P(T)\, P(U|T)\, P(G|T, U)\, P(D|T, U, G)$$

- (Regra da cadeia)
- 23 parâmetros

Um exemplo mais completo

Doença, Genótipo, Temperatura, Umidade

- Temperatura T: A, B (alta, baixa)
- Umidade U: S, U (seco, úmido)
- Genótipo G: mm, Mm, MM
- Doença D: R, Su (resistente, suscetível)
- É relevante calcular

$$P(T, U, G, D) = P(T)\, P(U|T)\, P(G|T, U)\, P(D|T, U, G)$$

- (Regra da cadeia)
- 23 parâmetros

Um exemplo mais completo

Doença, Genótipo, Temperatura, Umidade

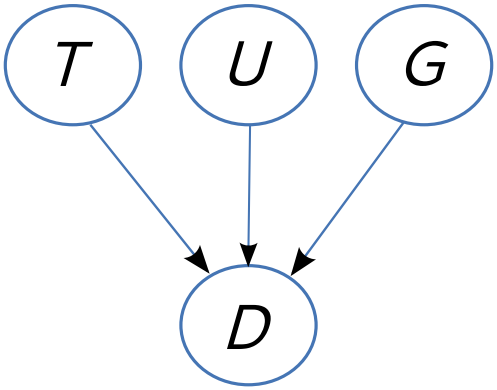
- Temperatura T: A, B (alta, baixa)
- Umidade U: S, U (seco, úmido)
- Genótipo G: mm, Mm, MM
- Doença D: R, Su (resistente, suscetível)
- É relevante calcular

$$P(T, U, G, D) = P(T)\, P(U|T)\, P(G|T, U)\, P(D|T, U, G)$$

- (Regra da cadeia)
- 23 parâmetros

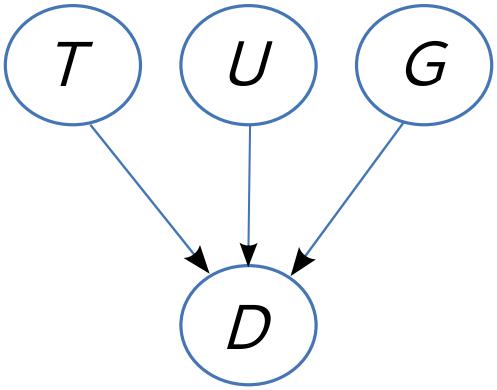
Modelo Gráfico Probabilístico

Rede bayesiana



Modelo Gráfico Probabilístico

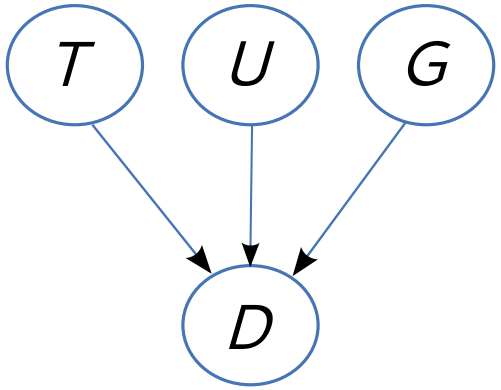
Rede bayesiana



- $P(T, U, G, D) = P(T)\, P(U)\, P(G)\, P(D|T, U, G)$
- 16 parâmetros
- Posso calcular o valor mais provável de um dado parâmetro, dadas as evidências (realizações de variáveis aleatórias)
- Como ficaria o modelo se incluíssemos Irrigação (S/N)?

Modelo Gráfico Probabilístico

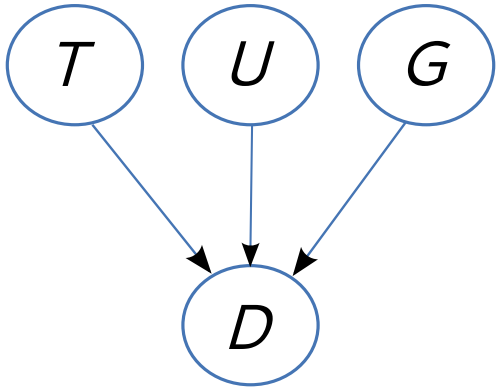
Rede bayesiana



- $P(T, U, G, D) = P(T)\, P(U)\, P(G)\, P(D|T, U, G)$
- 16 parâmetros
- Posso calcular o valor mais provável de um dado parâmetro, dadas as evidências (realizações de variáveis aleatórias)
- Como ficaria o modelo se incluíssemos Irrigação (S/N)?

Modelo Gráfico Probabilístico

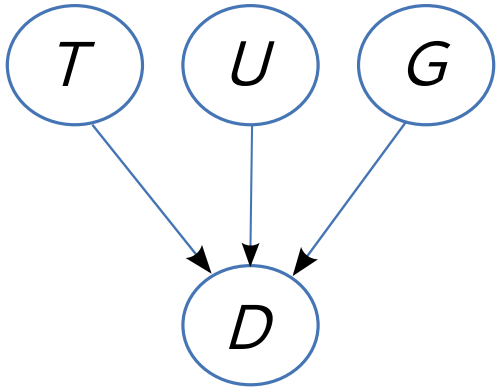
Rede bayesiana



- $P(T, U, G, D) = P(T)\, P(U)\, P(G)\, P(D|T, U, G)$
- 16 parâmetros
- Posso calcular o valor mais provável de um dado parâmetro, dadas as evidências (realizações de variáveis aleatórias)
- Como ficaria o modelo se incluíssemos Irrigação (S/N)?

Modelo Gráfico Probabilístico

Rede bayesiana



- $P(T, U, G, D) = P(T)\, P(U)\, P(G)\, P(D|T, U, G)$
- 16 parâmetros
- Posso calcular o valor mais provável de um dado parâmetro, dadas as evidências (realizações de variáveis aleatórias)
- Como ficaria o modelo se incluíssemos Irrigação (S/N)?

Outline

Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

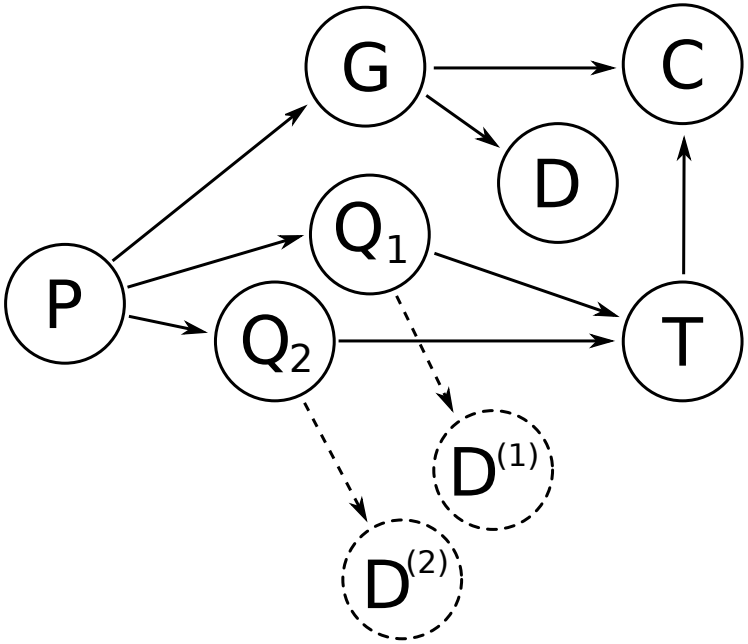
Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

Scenario - F1 population

- Observations
  - D: observed data
  - Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)
  - C: observed distribution (histogram of genotypes)
- Theory
  - P: ploidy
  - G: genotype of all individuals
  - T: expected distribution (theoretical distribution of genotypes)

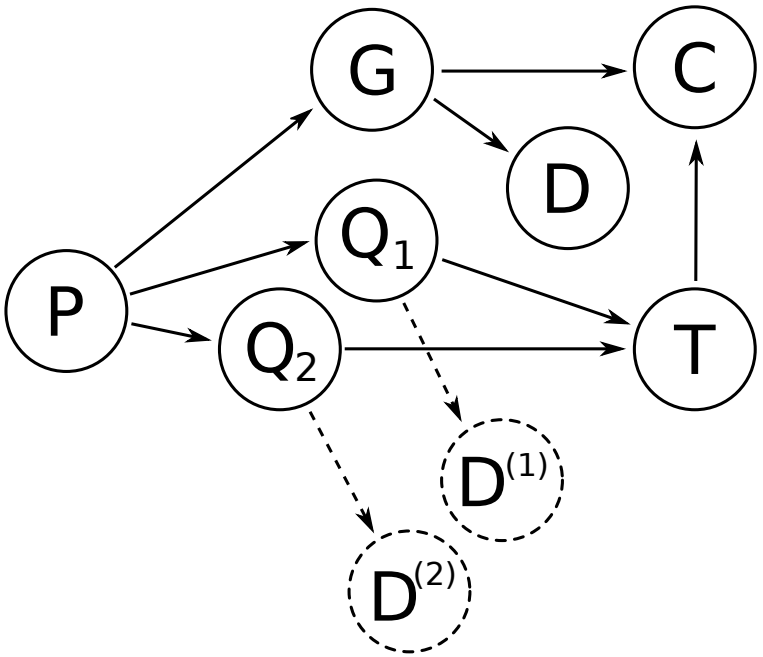
Probabilistic Graphical Model



$$\Pr(P, G, D, T, C, Q_1, Q_2, D_1, D_2)$$

- P: ploidy
- G: genotype of all individuals
- D: observed data
- T: theoretical distribution of genotypes
- C: histogram of genotypes
- Q_1 and Q_2: parent genotypes, with data D_1 and D_2 (if available)

Probabilistic graphical models



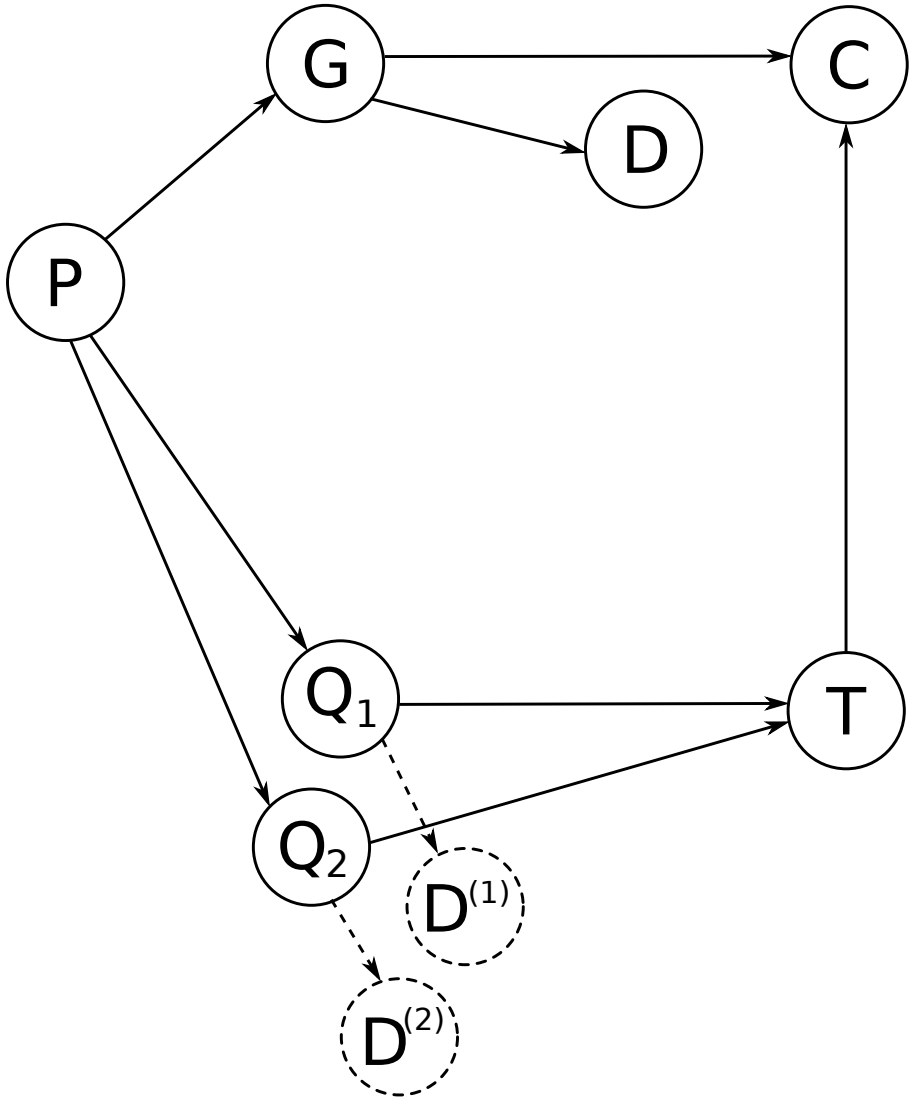
$$
\begin{aligned}
P(P,G,D,T,C,Q_1,Q_2,D_1,D_2) \quad = \quad & P(P)P(G|P)P(D|G) \\
& P(Q_1|P)P(Q_2|P) \\
& P(D_1|Q_1)P(D_2|Q_2) \\
& P(T|Q_1Q_2)P(C|G,T)
\end{aligned}
$$

Maximum a posteriori (MAP) solution

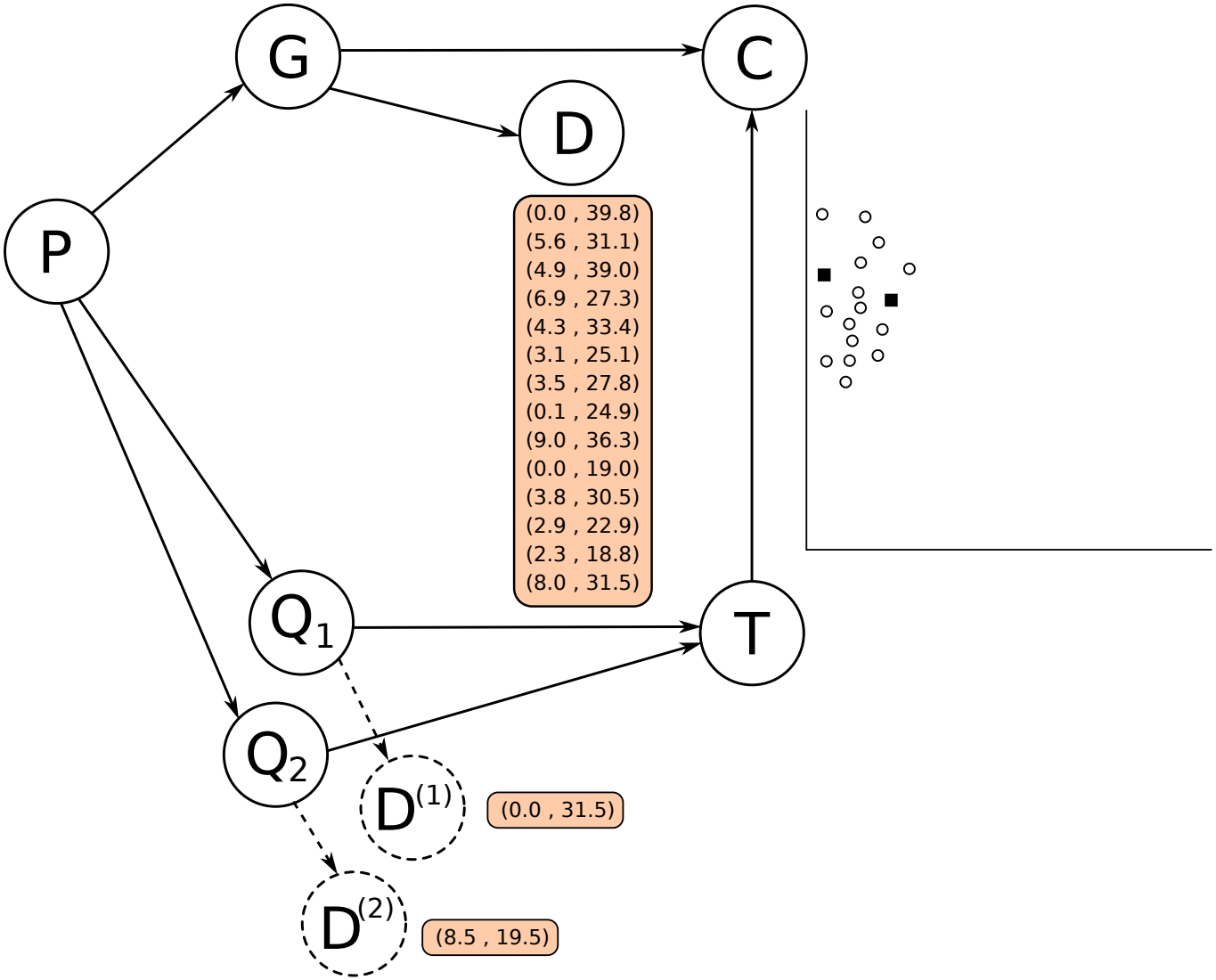
Maximum a posteriori configuration

It is the configuration (set of assignments) that maximizes the joint probability of the network, given the evidences (observed data)

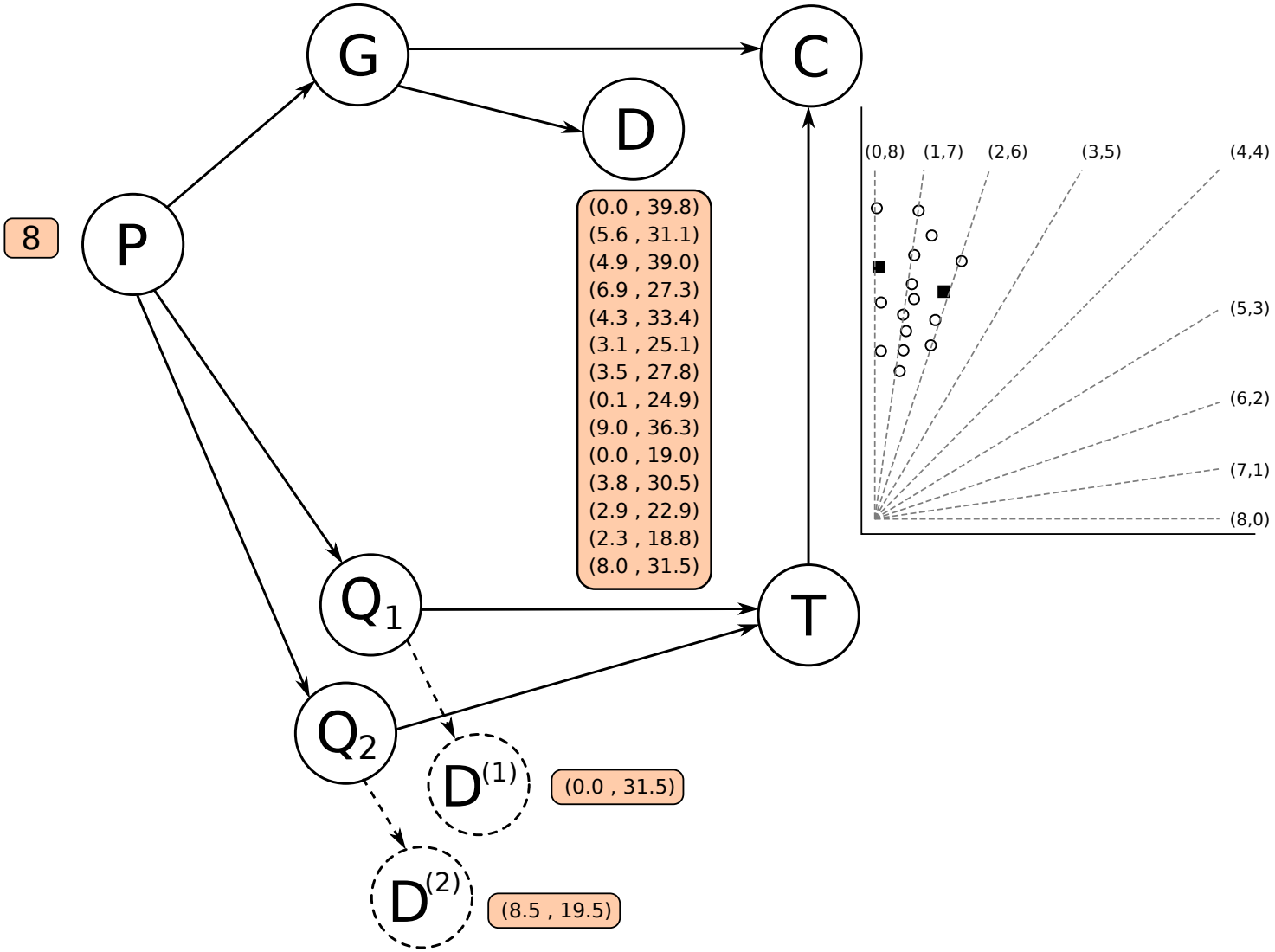
F_1 population



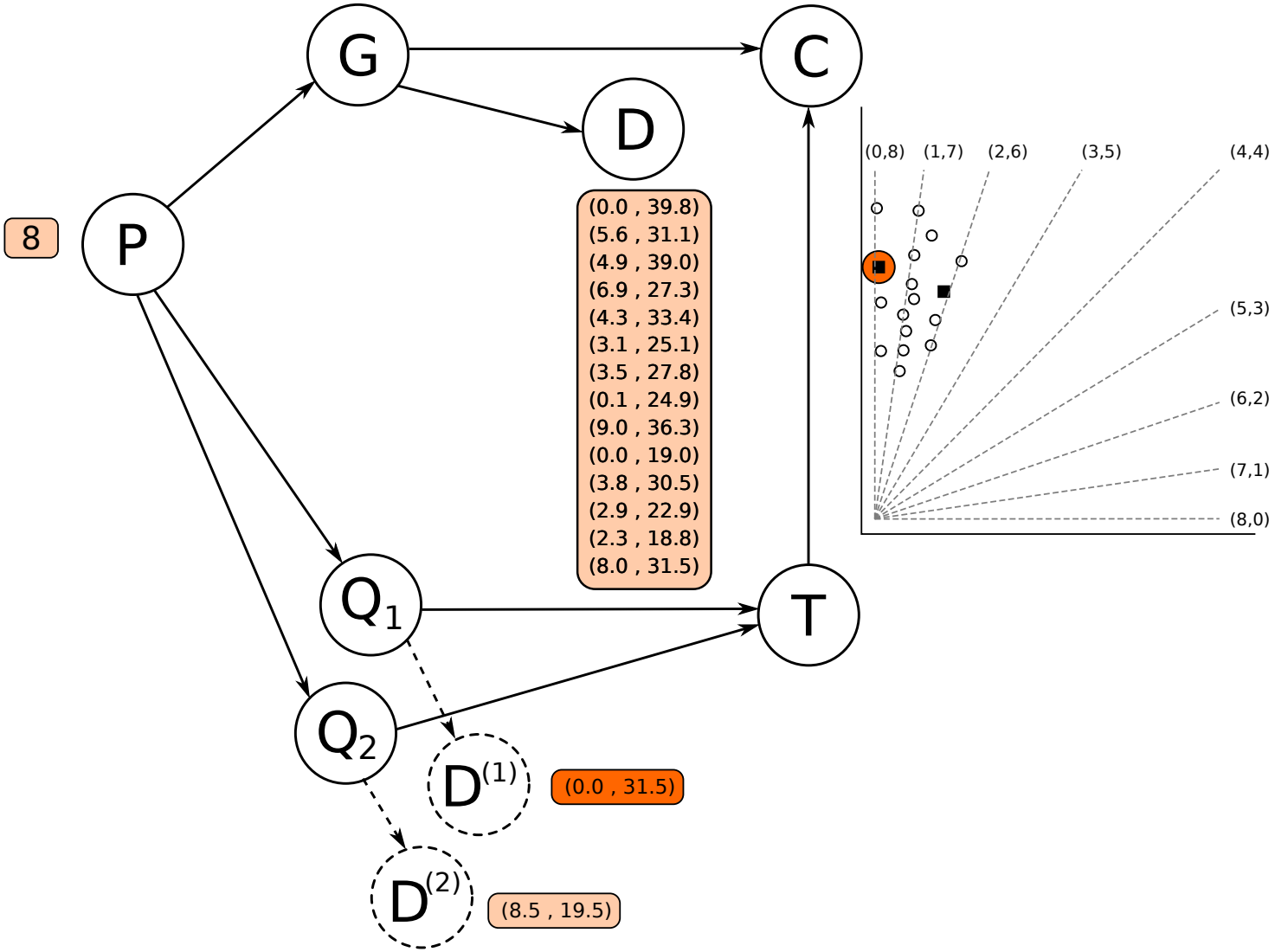
F_1 population



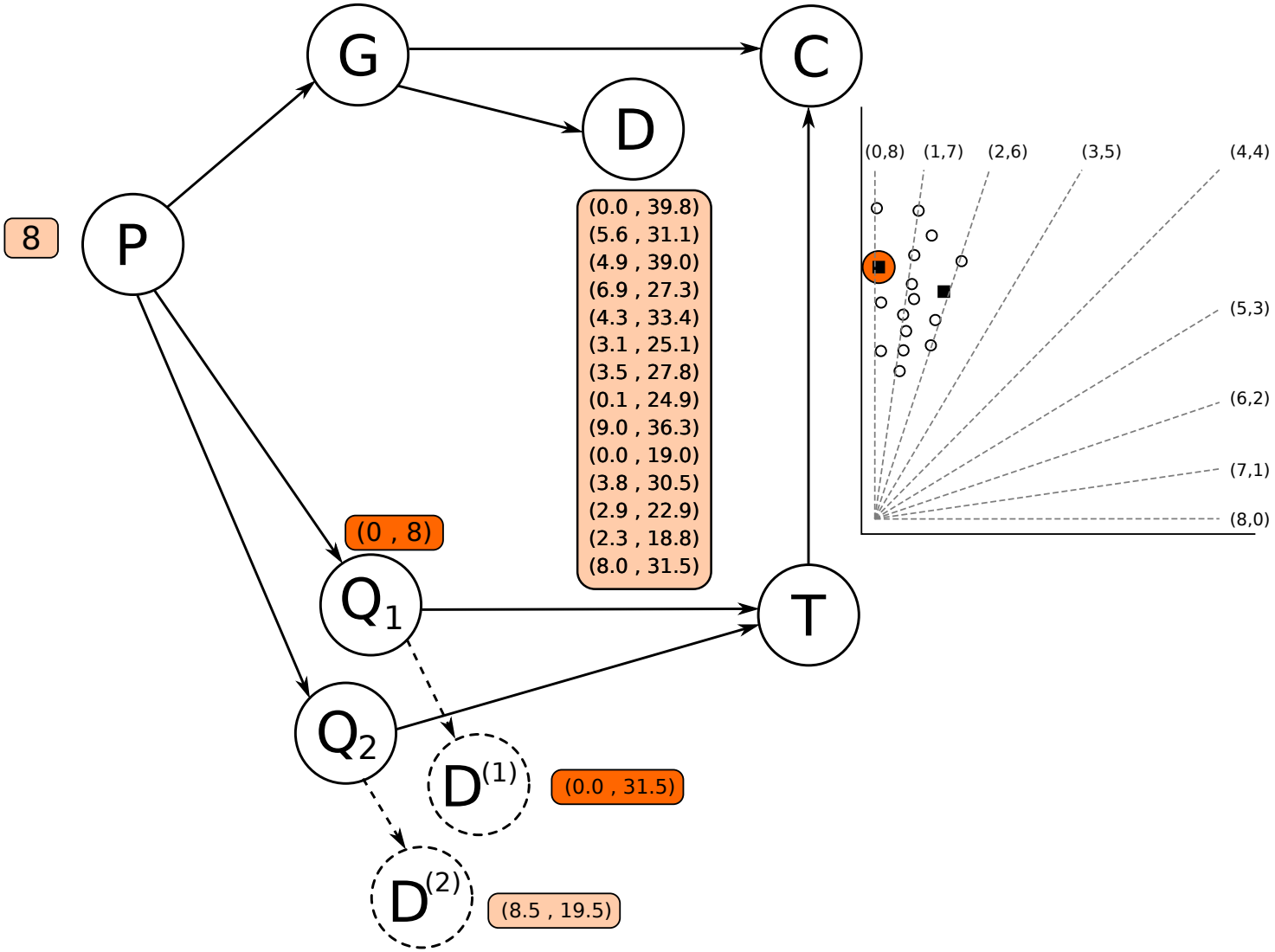
F_1 population



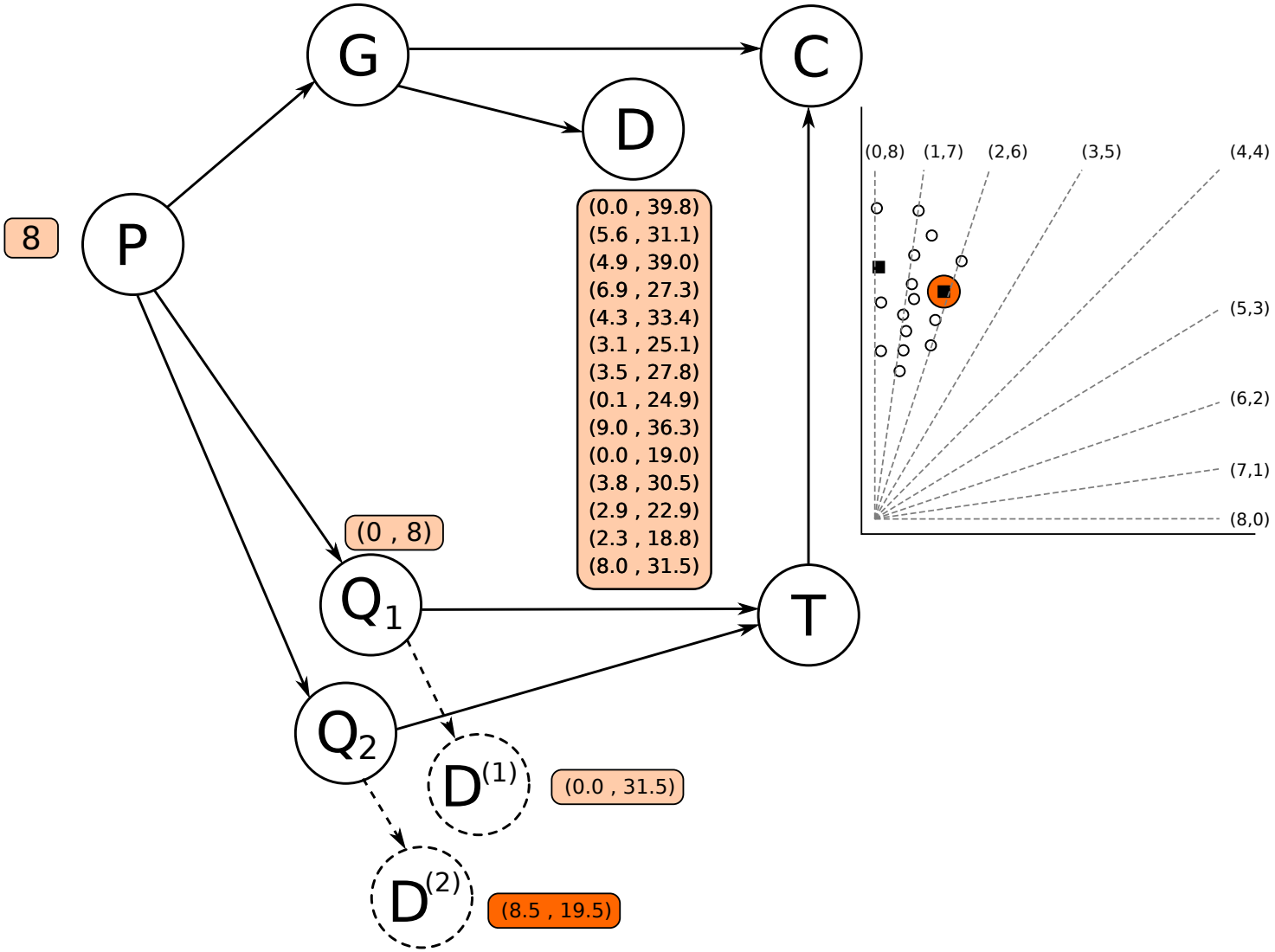
F_1 population



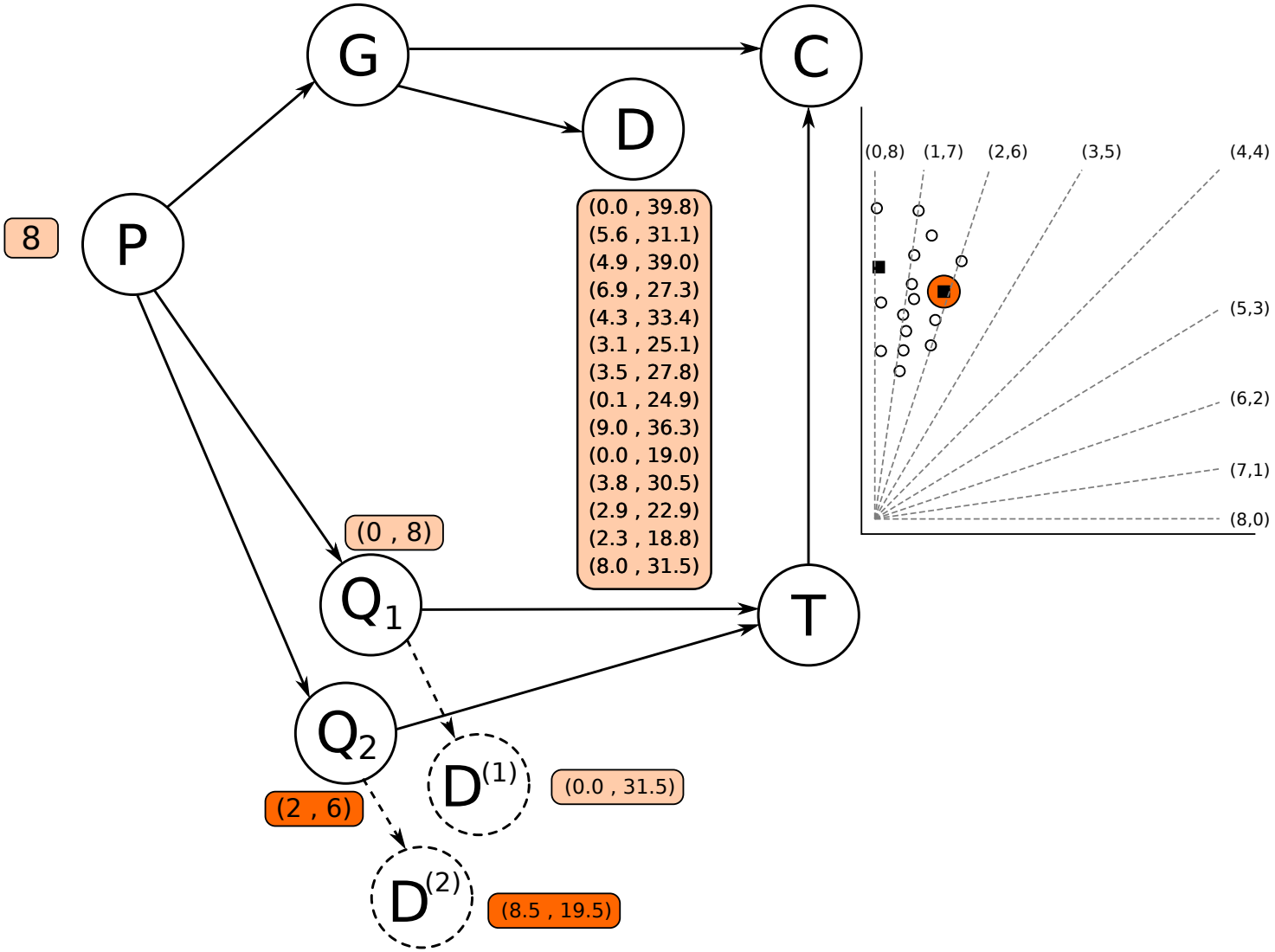
F_1 population



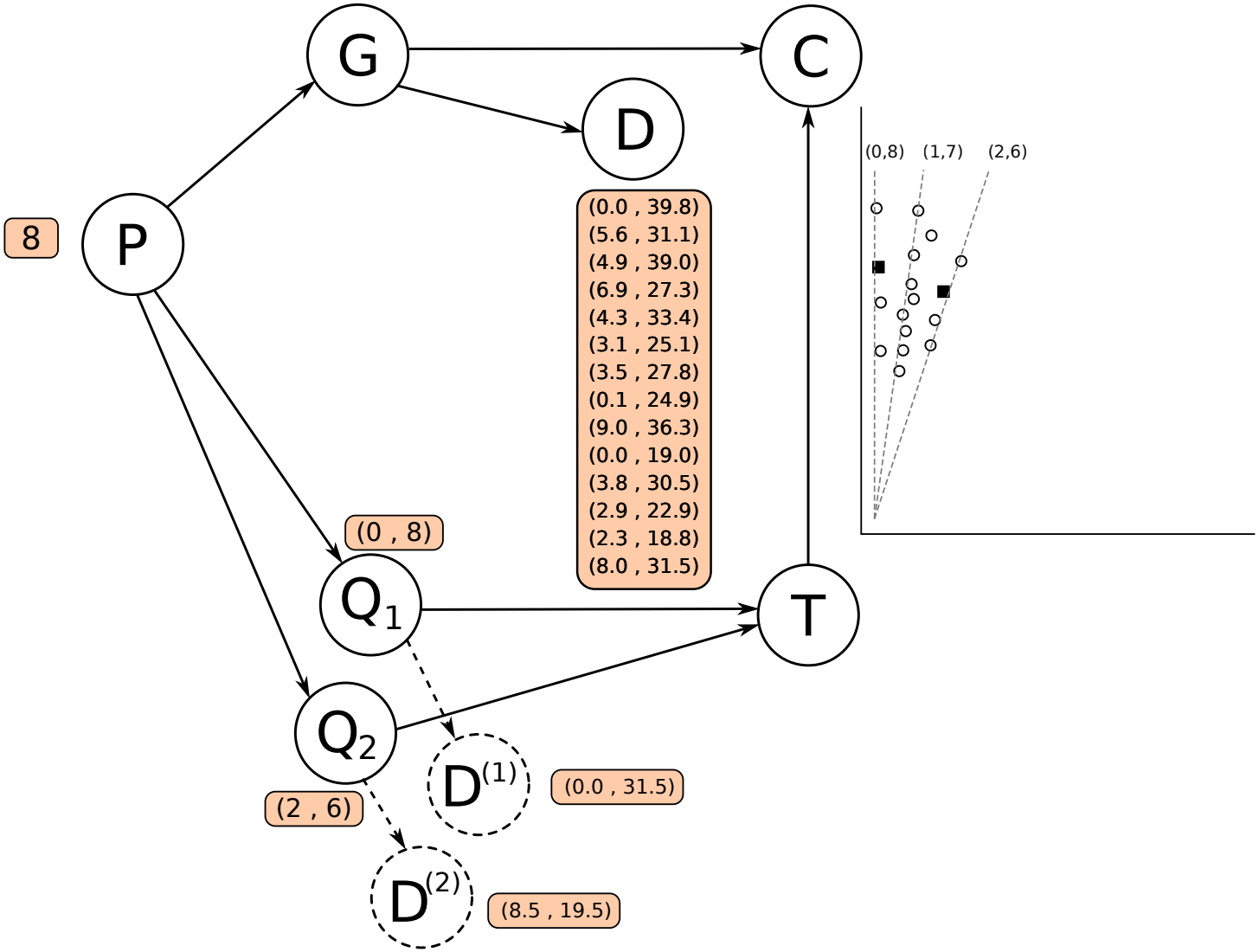
F_1 population



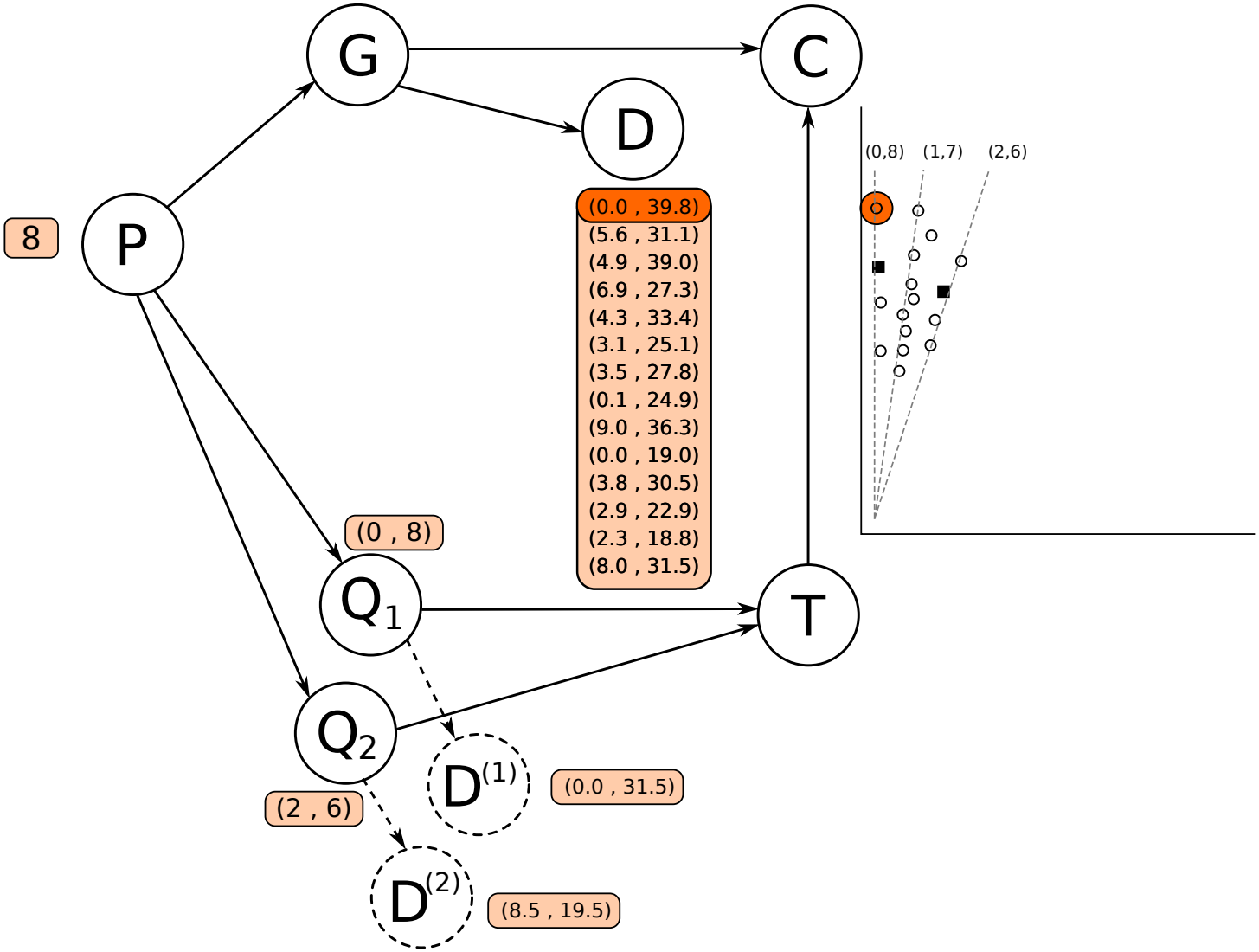
F_1 population



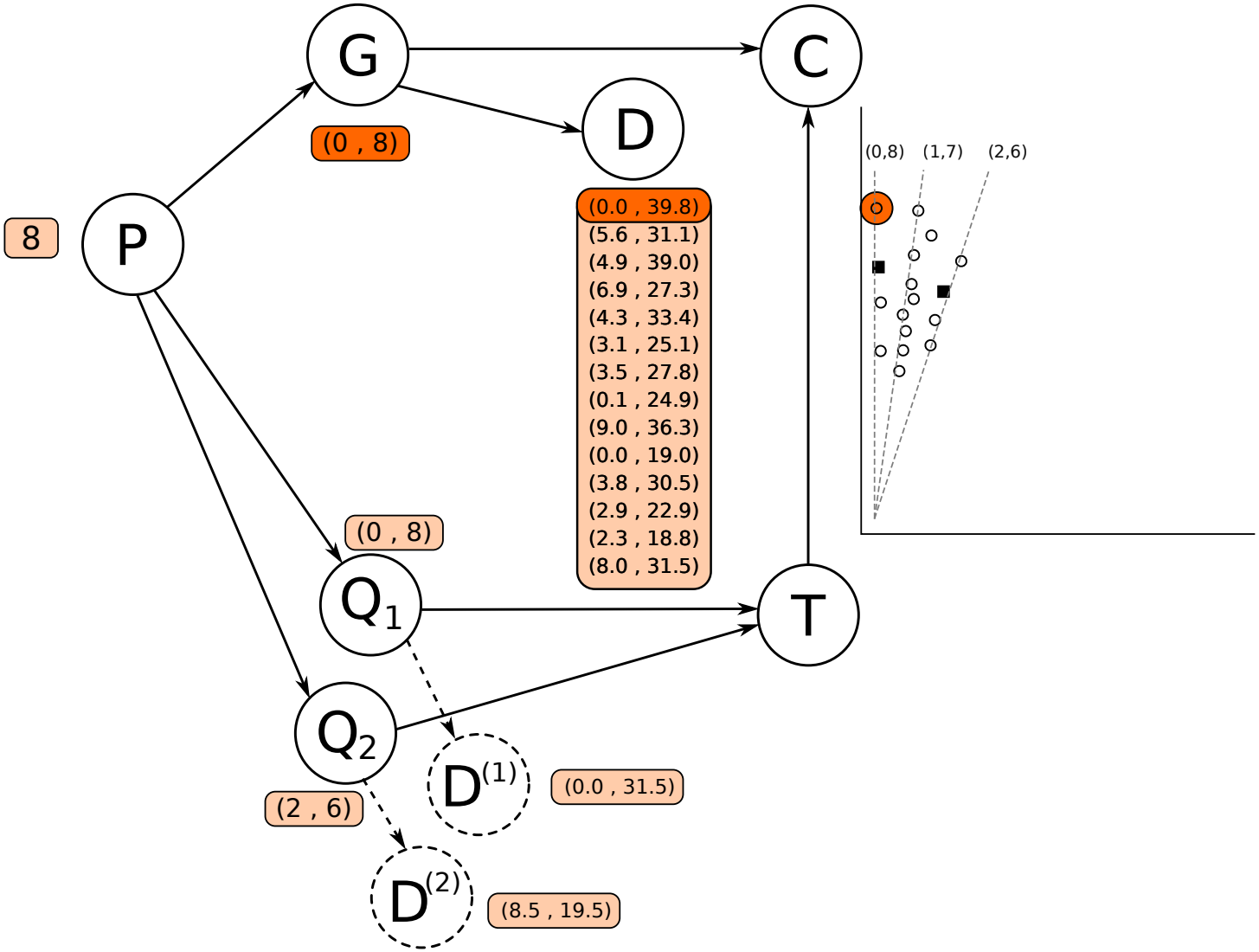
F_1 population



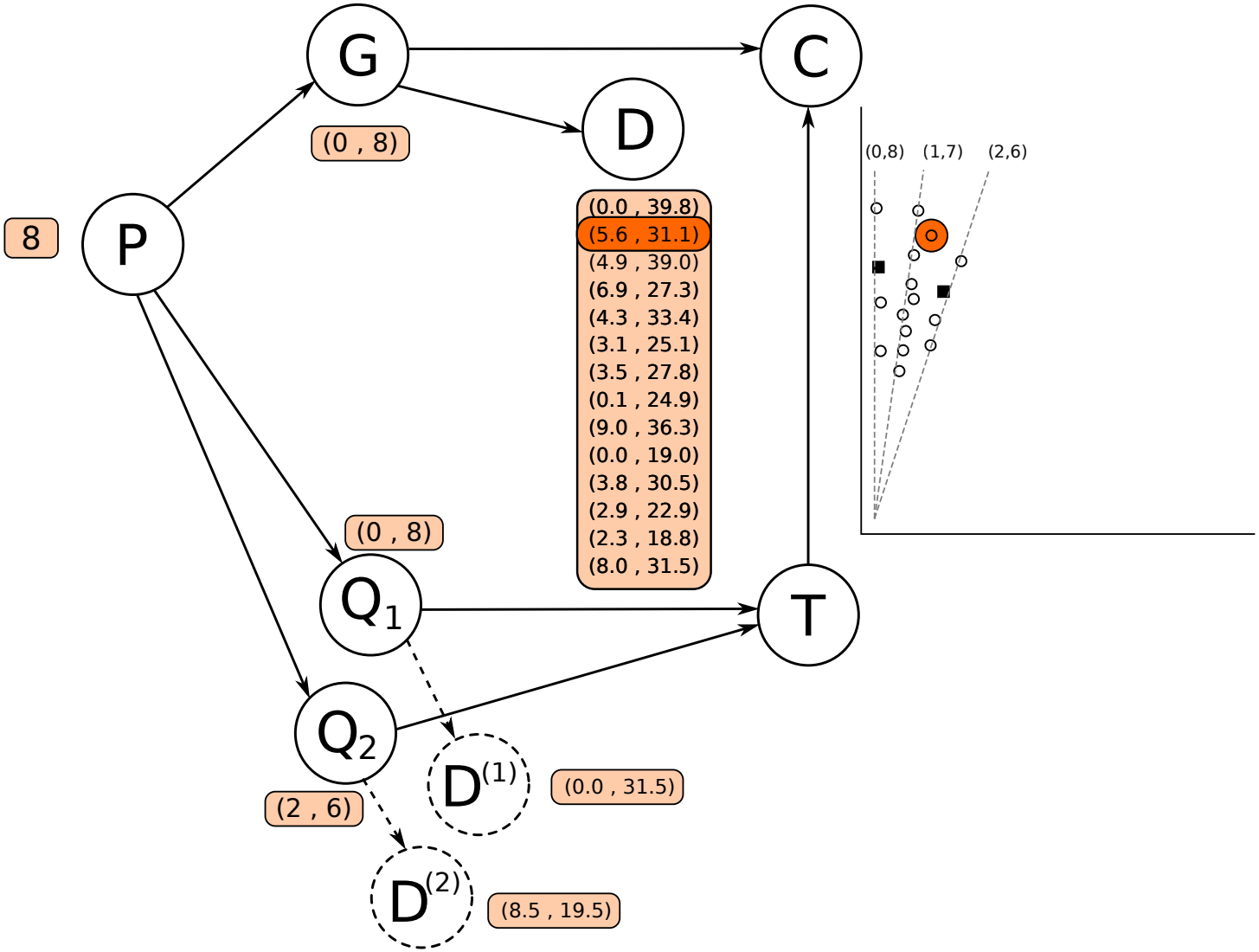
F_1 population



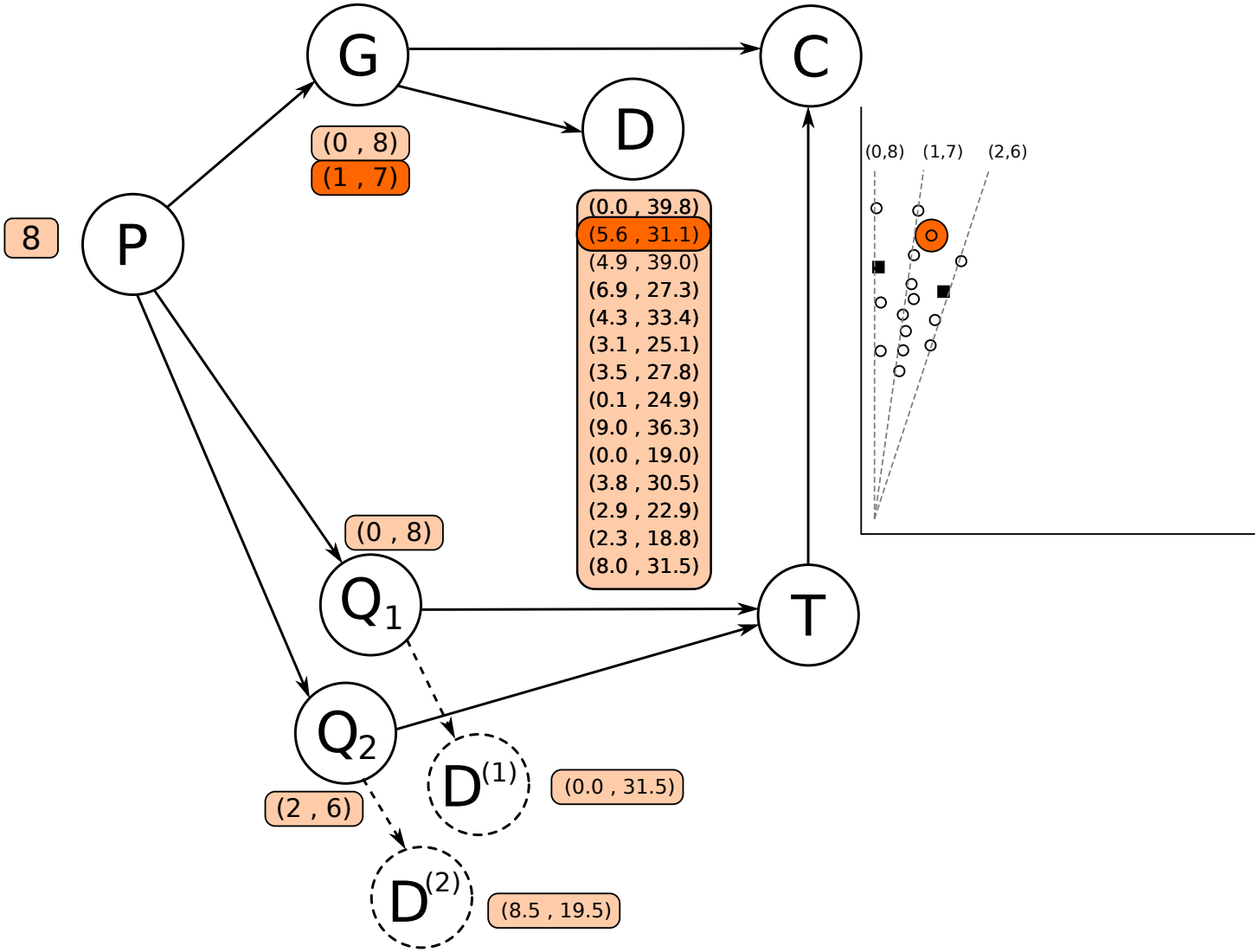
F_1 population



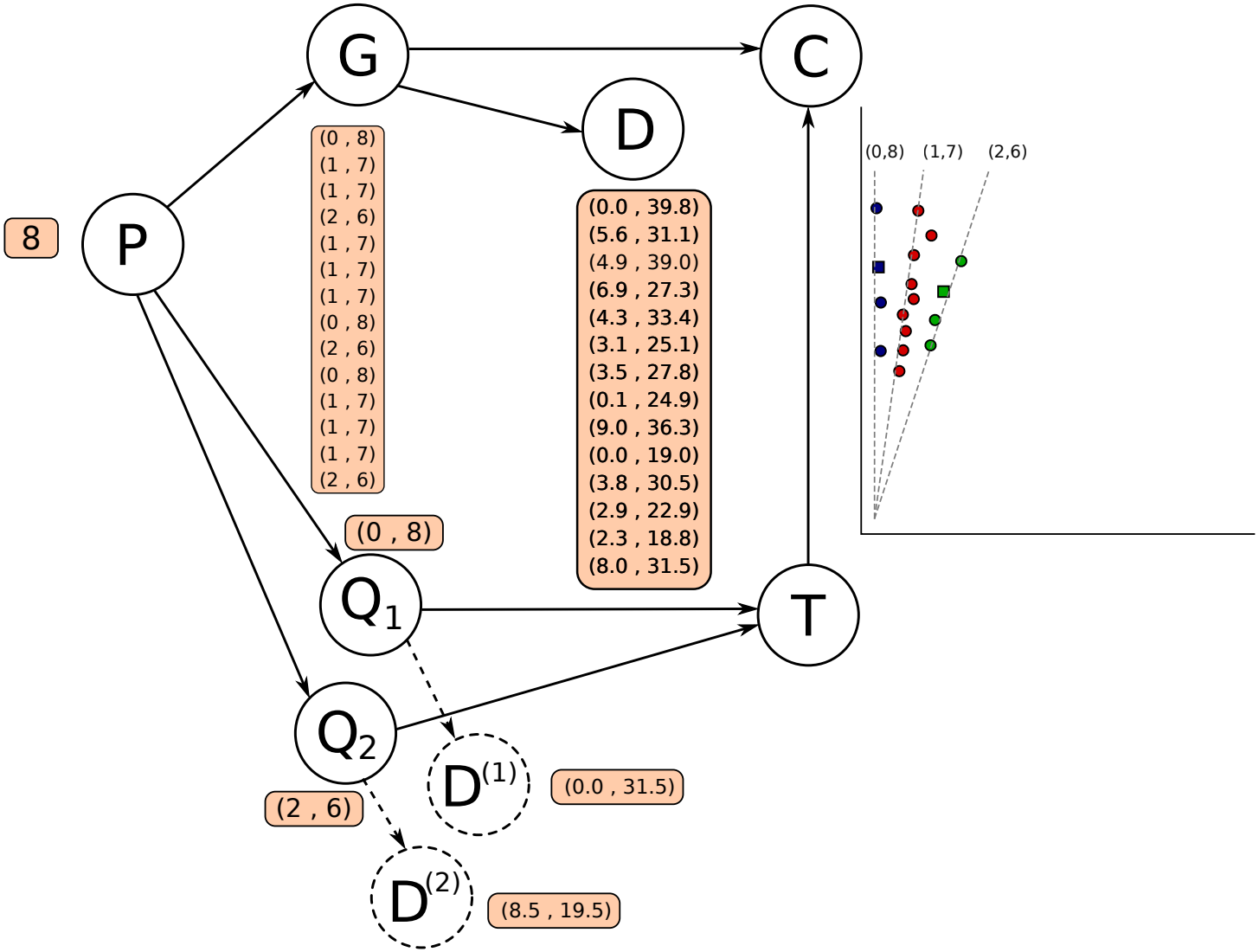
F_1 population



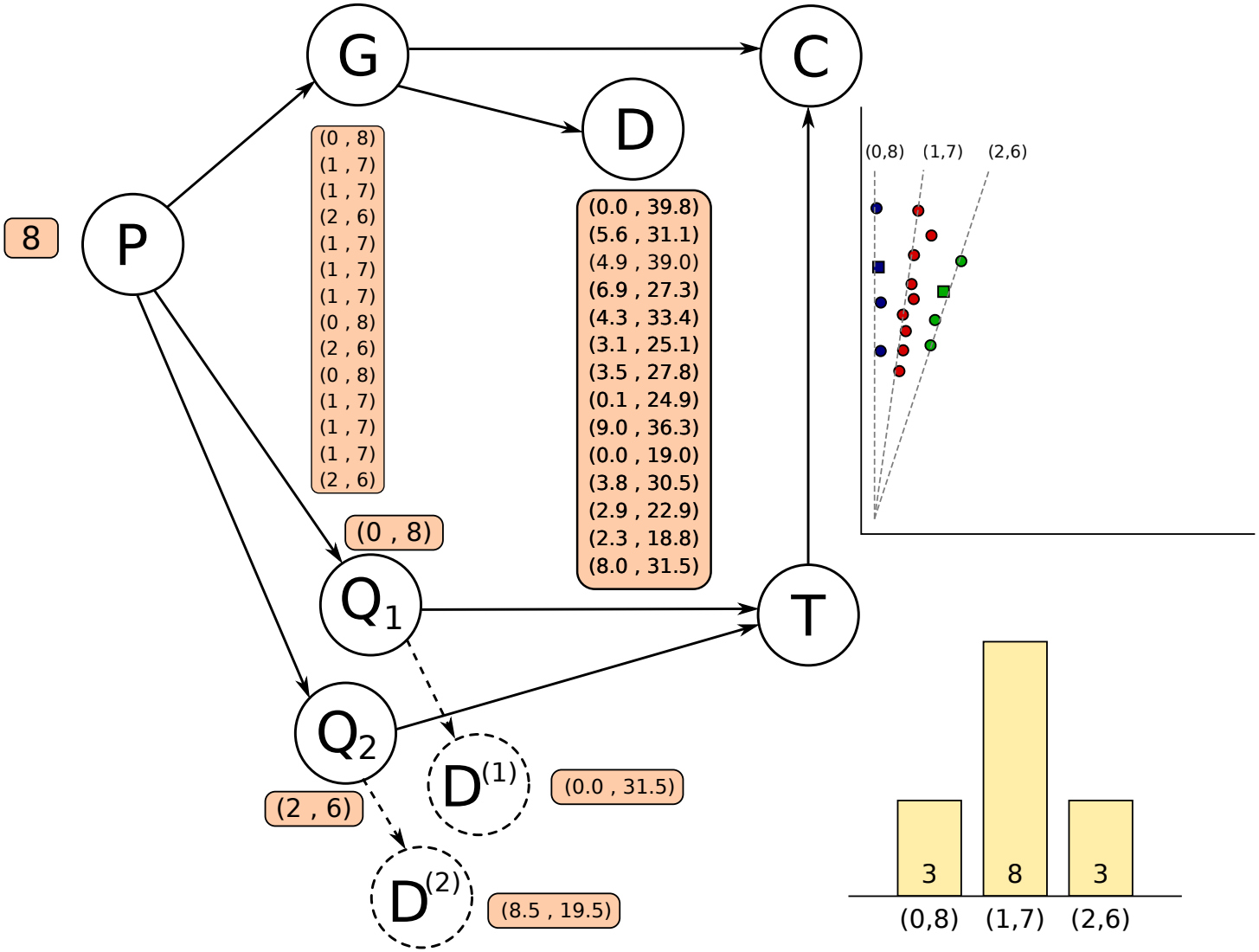
F_1 population



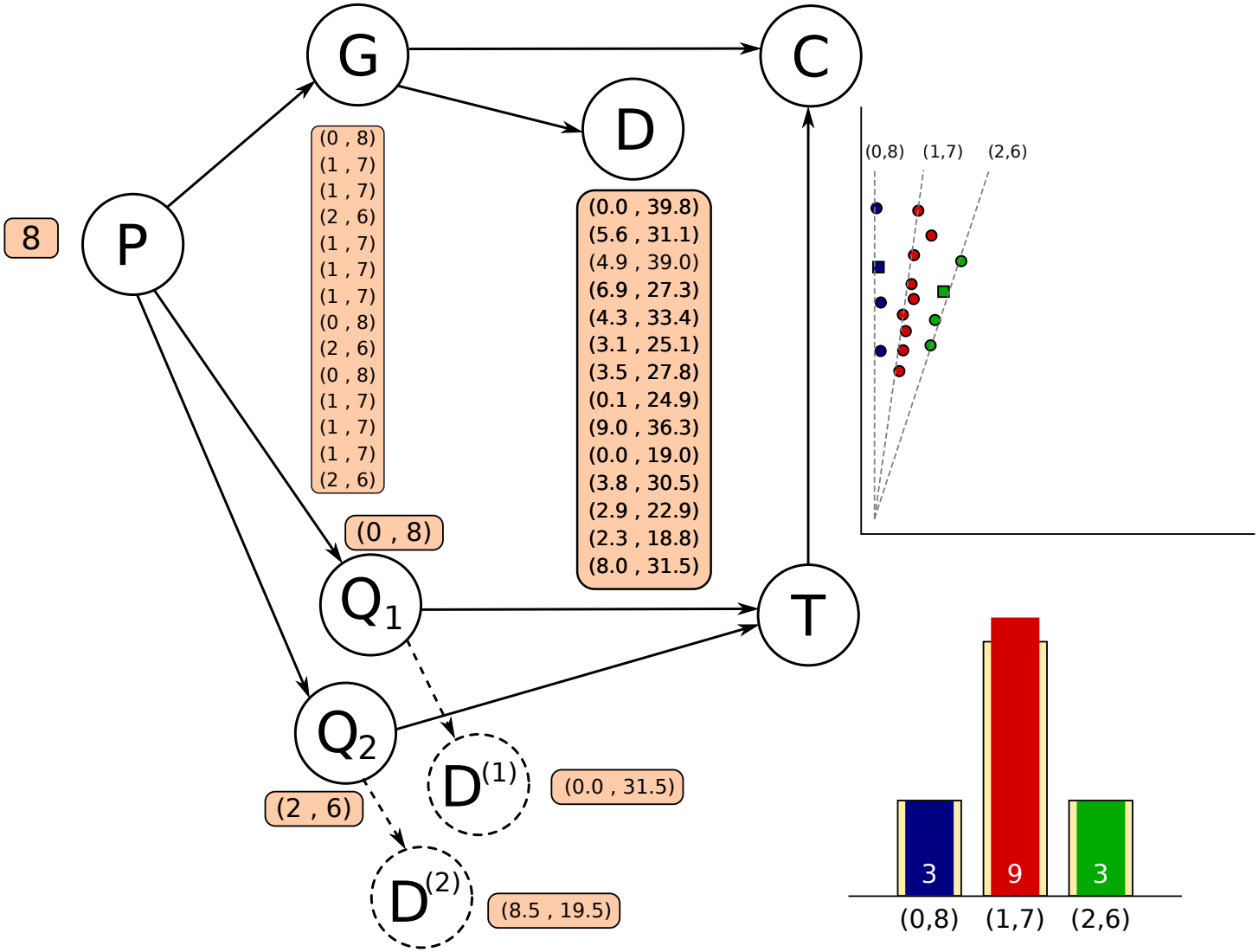
F_1 population



F_1 population



F_1 population



Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Elements of the model

- Ploidy P: even number, say, from 2 to 14
- *Genotype configuration* $G = (G_1, G_2, \ldots G_n)$
  - Example: $G_1 = (1, 7)$: 1 dose A, 7 T; $G_2 = (0, 8)$: 0 A, 8 T, etc.
- Set of possible genotype outcomes for a given ploidy as $\mu(P) = \{\mu_0, \mu_1, \ldots \mu_P\}$
  - Example: $\mu(P = 2) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$
- Data: $D = (D_1, D_2, \ldots D_n)$, each of which comprises an (x, y) (scatterplot)
- Same ideas for Q_1, Q_2, D^1 and D^2
- Theoretical distribution T: Mendelian segregation
- Actual distribution C: histogram

Conditional distributions

$$
\begin{aligned}
P(P, G, D, T, C, Q_1, Q_2, D_1, D_2) &= P(P)P(G|P)P(D|G) \\
&\quad P(Q_1|P)P(Q_2|P) \\
&\quad P(D_1|Q_1)P(D_2|Q_2) \\
&\quad P(T|Q_1Q_2)P(C|G,T)
\end{aligned}
$$

- $P(P)$: prior (uniform)
- $P(G|P) = \frac{1}{p+1}$ (uniform)

Conditional distributions

$$
\begin{aligned}
P(P, G, D, T, C, Q_1, Q_2, D_1, D_2) &= P(P)P(G|P)P(D|G) \\
&\quad P(Q_1|P)P(Q_2|P) \\
&\quad P(D_1|Q_1)P(D_2|Q_2) \\
&\quad P(T|Q_1Q_2)P(C|G,T)
\end{aligned}
$$

- $P(P)$: prior (uniform)
- $P(G|P) = \frac{1}{p+1}$ (uniform)

Conditional distributions

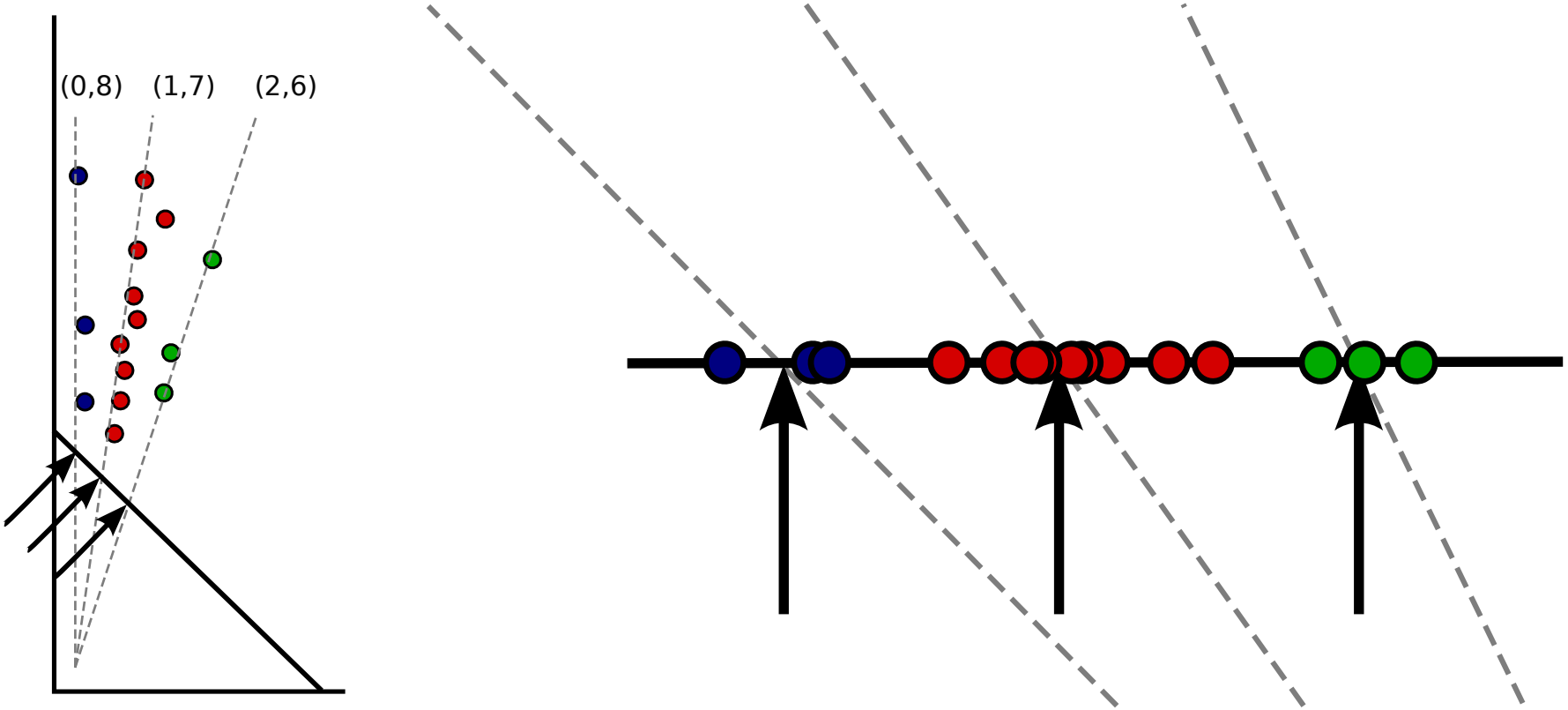
$$
\begin{aligned}
P(P, G, D, T, C, Q_1, Q_2, D_1, D_2) &= P(P)P(G|P)P(D|G) \\
&\quad P(Q_1|P)P(Q_2|P) \\
&\quad P(D_1|Q_1)P(D_2|Q_2) \\
&\quad P(T|Q_1Q_2)P(C|G,T)
\end{aligned}
$$

- $P(D|G)$: Likelihood of any genotype configuration $G = g$ can be written as a product over individuals:

$$\Pr(D|G = g) = \prod_i \Pr(D_i|G_i = g_i)$$

Conditional distributions

- $P(D|G)$



Elements of the model

- Likelihood proportional to $\Pr(D_i|G_i = g_i)$ using a normal distribution

$$\Pr(D_i|G_i = g_i) \propto \frac{e^{\frac{-\|\widehat{D_i} - \widehat{g_i}\|_2^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

where the operator $\widehat{u} = \frac{u}{\|u\|_1}$ is used to perform L_1 normalization on D_i and g_i.

Conditional distributions

$$
\begin{aligned}
P(P,G,D,T,C,Q_1,Q_2,D_1,D_2) &= P(P)P(G|P)P(D|G) \\
&\quad P(Q_1|P)P(Q_2|P) \\
&\quad P(D_1|Q_1)P(D_2|Q_2) \\
&\quad P(T|Q_1Q_2)P(C|G,T)
\end{aligned}
$$

- $P(Q_1|P)$, $P(Q_2|P)$, $P(D_1|Q_1)P$ and $(D_2|Q_2)$: same ideas

Conditional distributions

$$
\begin{aligned}
P(P,G,D,T,C,Q_1,Q_2,D_1,D_2) &= P(P)P(G|P)P(D|G) \\
&\quad P(Q_1|P)P(Q_2|P) \\
&\quad P(D_1|Q_1)P(D_2|Q_2) \\
&\quad P(T|Q_1Q_2)P(C|G,T)
\end{aligned}
$$

- $P(T|Q_1Q_2)$: Mendelian distribution (Hypergeometric distribution)

Conditional distributions

One informative parent

- m: Ploidy level
- k: Dosage in one parent
- d: Dosage in the gamete
- $P(d)$: Probability of observing a gamete with dosage d given ploidy m, dosage in one parent k and dosage in another parent 0

Conditional distributions

One informative parent

- m: Ploidy level
- k: Dosage in one parent
- d: Dosage in the gamete
- $P(d)$: Probability of observing a gamete with dosage d given ploidy m, dosage in one parent k and dosage in another parent 0

Probability of a given dosage (d) in the gamete, one parent

$$P(d) = \frac{\binom{k}{d}\binom{m-k}{\frac{m}{2}-d}}{\binom{m}{\frac{m}{2}}}$$

- $\binom{k}{d}$: k copies in one parent choose d
- $\binom{m-k}{\frac{m}{2}-d}$: chromosomes which DO NOT have copies ($m-k$), choose $\frac{m}{2}-d$ (necessary number to complete a gamete)
- $\binom{m}{\frac{m}{2}}$: all possible gametes

Probability of a given dosage (d) in the gamete, one parent

$$P(d) = \frac{\binom{k}{d}\binom{m-k}{\frac{m}{2}-d}}{\binom{m}{\frac{m}{2}}}$$

- $\binom{k}{d}$: k copies in one parent choose d
- $\binom{m-k}{\frac{m}{2}-d}$: chromosomes which DO NOT have copies ($m-k$), choose $\frac{m}{2}-d$ (necessary number to complete a gamete)
- $\binom{m}{\frac{m}{2}}$: all possible gametes

Probability of a given dosage (d) in the gamete, one parent

$$P(d) = \frac{\binom{k}{d}\binom{m-k}{\frac{m}{2}-d}}{\binom{m}{\frac{m}{2}}}$$

- $\binom{k}{d}$: k copies in one parent choose d
- $\binom{m-k}{\frac{m}{2}-d}$: chromosomes which DO NOT have copies ($m-k$), choose $\frac{m}{2}-d$ (necessary number to complete a gamete)
- $\binom{m}{\frac{m}{2}}$: all possible gametes

Theoretical Distribution

Two informative parents

$$\mathbf{S} = \mathbf{p}'\mathbf{p}$$

Example: Octaploid with two doses in both parents

$$\begin{array}{c|ccc|}
 & \text{P(AAaa)} & \text{P(Aaaa)} & \text{P(aaaa)} \\
\text{P(AAaa)} & \text{P(AAAAaaaa)} & \text{P(AAAaaaaa)} & \text{P(AAaaaaaa)} \\
\text{P(Aaaa)} & \text{P(AAAaaaaa)} & \text{P(AAaaaaaa)} & \text{P(Aaaaaaaa)} \\
\text{P(aaaa)} & \text{P(AAaaaaaa)} & \text{P(Aaaaaaaa)} & \text{P(aaaaaaaa)}
\end{array}$$

Probabilistic Inference

- Parameters: $\theta = (P, Q_1, Q_2)$ (and σ)
- For each θ: compute the MAP genotype configuration g^*_θ
- Compute $P(g^*_\theta, \theta \mid D)$ posterior belief that the MAP parameter and genotype configuration is correct

Probabilistic Inference

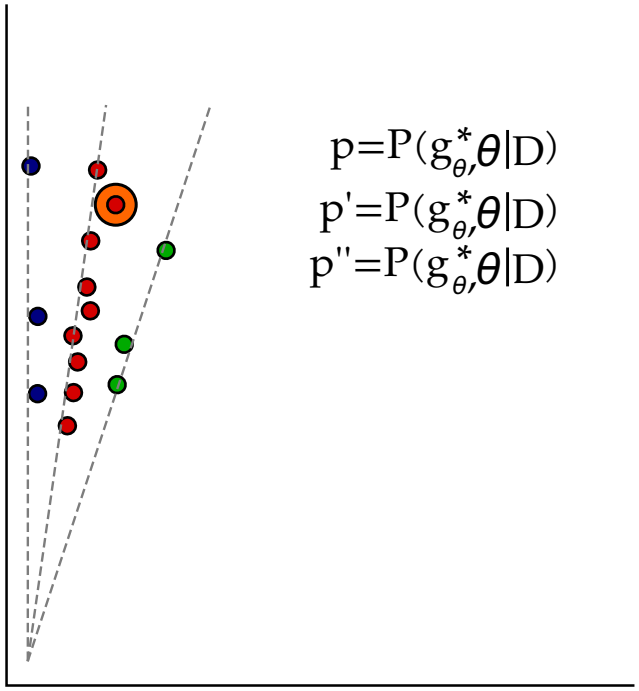
- Parameters: $\theta = (P, Q_1, Q_2)$ (and σ)
- For each θ: compute the MAP genotype configuration g^*_θ
- Compute $P(g^*_\theta, \theta \mid D)$ posterior belief that the MAP parameter and genotype configuration is correct

Probabilistic Inference

- Parameters: $\theta = (P, Q_1, Q_2)$ (and σ)
- For each θ: compute the MAP genotype configuration g_θ^*
- Compute $P(g_\theta^*, \theta \mid D)$ posterior belief that the MAP parameter and genotype configuration is correct

Individual Probabilities

- Posterior estimates that each individual is assigned to the correct genotype



Probabilistic Inference

- Exact MAP computation

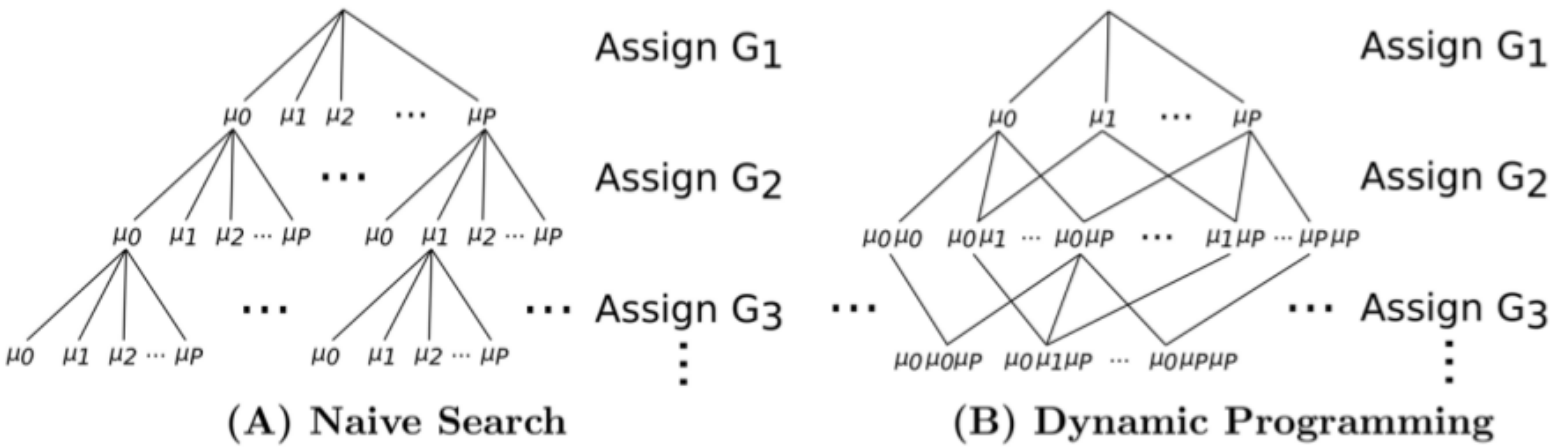
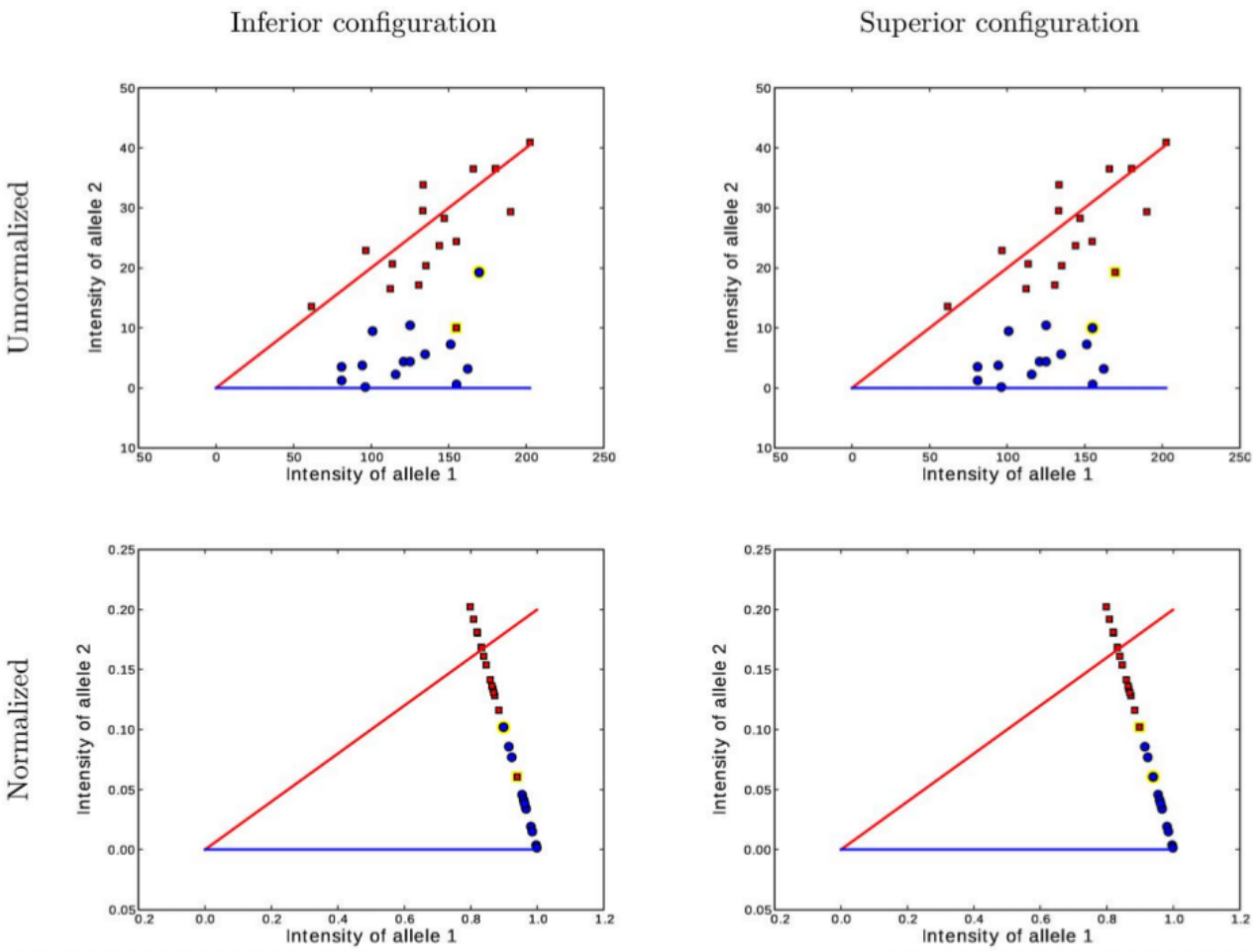


Figure 3. Illustration of Exact Inference. Exact MAP computation can be performed by enumerating all possible genotype configurations.

Efficient Exact Inference

- Optimal genotype configuration can be achieved (contiguous blocks)



Publications

Serang, O. R.; Mollinari, M.; Garcia, A. A. F.
Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids
PLoS ONE 7(2), e30906, 2012

- Free online software to implement the analysis (`SuperMASSA`)

AAF Garcia, M Mollinari, TG Marconi, OR Serang, RR Silva, MLC Vieira, R Vicentini, EA Costa, MC Mancini, MOS Garcia, MM Pastina, R Gazaffi, ERF Martins, N Dahmer, DA Sforça, CBC Silva, P Bundock, RJ Henry, GM Souza, M Sluys, MGA Landell, MS Carneiro, MAG Vincentz, LR Pinto, R Vencovsky, AP Souza
SNP genotyping allows an in-depth characterization of the genome of sugarcane and other complex autopolyploids
Scientific Reports (under review)

People and Funding

- State University of Campinas (UNICAMP)
  - Anete Pereira de Souza
  - Thiago G Marconi
  - Melina C Mancini
  - Estela A Costa
- University of São Paulo (ESALQ/USP)
  - Antonio Augusto F Garcia
  - Marcelo Mollinari
  - Oliver R Serang
- Agronomic Institute of Campinas (IAC)
  - Luciana R Pinto
  - Marcos Landell
- BIOEN/FAPESP
- INCT (Instituto Nacional de Ciência e Tecnologia do Bioetanol)