LEIDEN UNIVERSITY MEDICAL CENTER

Y-STRs and population sampling for forensic reference purposes

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Short tandem repeat (STR) analysis is used for *genetic fingerprinting*.

Introduction

Short tandem repeat (STR) analysis is used for *genetic fingerprinting*.

- Tetra- or penta-nucleotide repeats.
  - Shorter repeats suffer from PCR stutter and preferential amplification.
  - Longer repeats are less likely to survive degradation.

Introduction

Short tandem repeat (STR) analysis is used for *genetic fingerprinting*.

- Tetra- or penta-nucleotide repeats.
  - Shorter repeats suffer from PCR stutter and preferential amplification.
  - Longer repeats are less likely to survive degradation.
- A small number of these loci can already be used for identification.
  - Only 13 are used in the US.
  - Each STR allele is common in 5% to 20% of a population.
  - Chance of a false positive: $1 \cdot 10^{-18}$ (rough estimate).

Introduction

Short tandem repeat (STR) analysis is used for *genetic fingerprinting*.

- Tetra- or penta-nucleotide repeats.
  - Shorter repeats suffer from PCR stutter and preferential amplification.
  - Longer repeats are less likely to survive degradation.
- A small number of these loci can already be used for identification.
  - Only 13 are used in the US.
  - Each STR allele is common in 5% to 20% of a population.
  - Chance of a false positive: $1 \cdot 10^{-18}$ (rough estimate).
    - However, a false positive match was found in 2007, in a database of 30,000 samples.
    - There are also about $12 \cdot 10^{6}$ monozygotic twins.

Introduction

The classical identification procedure goes as follows:

Introduction

The classical identification procedure goes as follows:

Determination by length:

- Extract nuclear DNA.
- Amplify the polymorphic regions with PCR.
- With gel electrophoresis or capillary electrophoresis, determine the size of the amplicon.

Introduction

The classical identification procedure goes as follows:

Determination by length:

- Extract nuclear DNA.
- Amplify the polymorphic regions with PCR.
- With gel electrophoresis or capillary electrophoresis, determine the size of the amplicon.

Determination by sequencing (J.W.F. van der Heijden, K.J. van der Gaag, P. de Knijff, J.F.J. Laros) :

- After amplification, sequence the amplicons.
- With *semi-global alignment*, determine the exact repeat pattern.
    - The length can be determined on the nucleotide level.
    - Diversity within the repeat can be identified.

Estimation of diversity

The classical reference file.

“No family relations allowed.”

Estimation of diversity

The classical reference file.

"No family relations allowed."

- Rare alleles might be overrepresented.

This has never been examined.

Estimation of diversity

The classical reference file.

"No family relations allowed."

- Rare alleles might be overrepresented.

This has never been examined.

While this might be sufficient for autosomal STR markers, Y-STR marker diversity is only caused by mutation.

- The above rule was already established and was copied without considering the consequences.

Estimation of diversity

The classical reference file.

"No family relations allowed."

- Rare alleles might be overrepresented.

This has never been examined.

While this might be sufficient for autosomal STR markers, Y-STR marker diversity is only caused by mutation.

- The above rule was already established and was copied without considering the consequences.

An underestimation of the frequency of commonly shared Y-STR haplotypes can be unfavourable for a suspect.

The Rucphen study

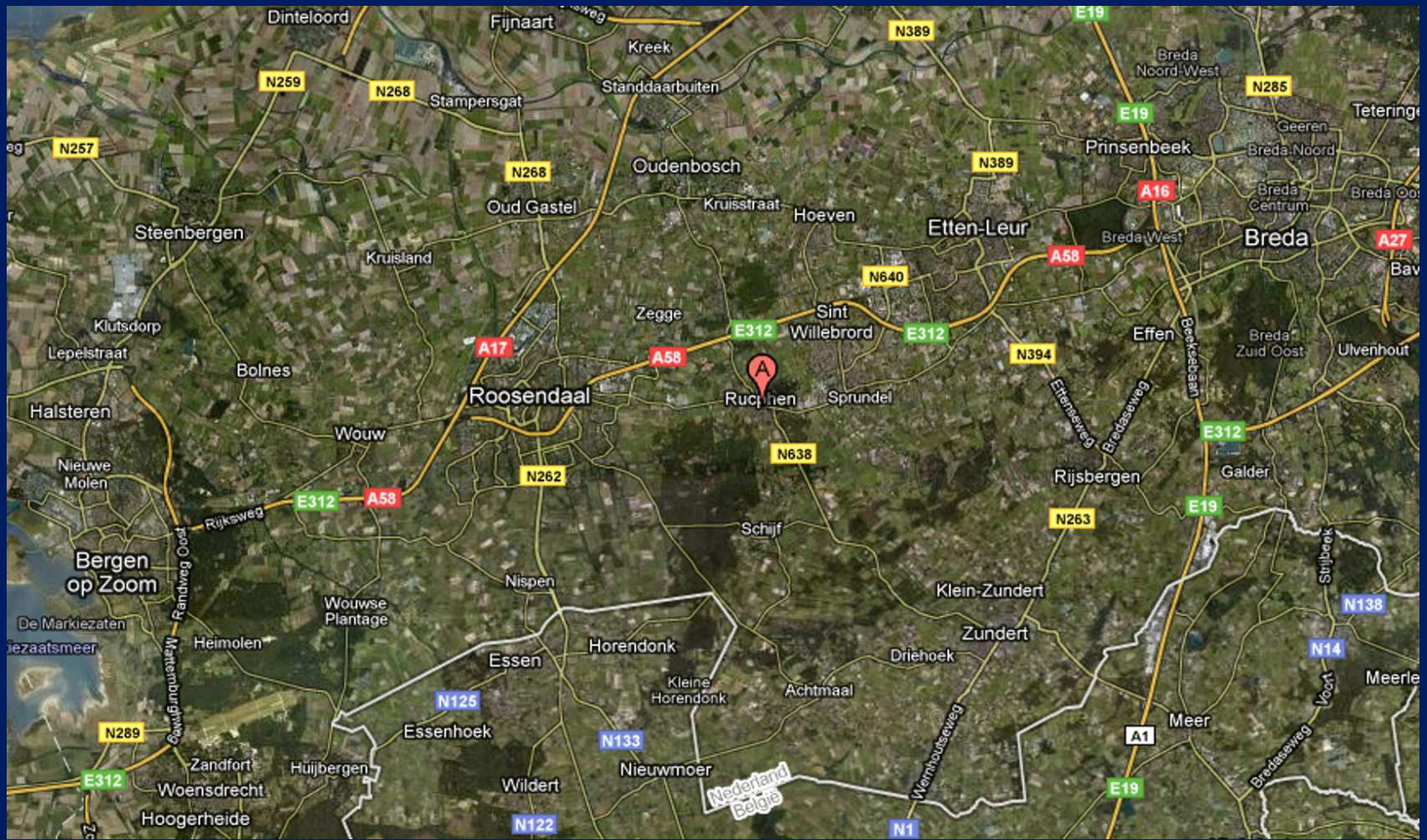


Figure 1: Rucphen, Noord-Brabant.

The Rucphen study

Set up:

- Random selection of 1350 males from Rucphen.
- Genotype for 15 autosomal STR loci, 25 Y-SNPs, 12 Y-STR loci.
- Retrieve paternal genealogies.

The Rucphen study

Set up:

- Random selection of 1350 males from Rucphen.
- Genotype for 15 autosomal STR loci, 25 Y-SNPs, 12 Y-STR loci.
- Retrieve paternal genealogies.
- Make a 1328×1328 distance matrix.
    - Counting the number of meioses in each pair.

The Rucphen study

Set up:

- Random selection of 1350 males from Rucphen.
- Genotype for 15 autosomal STR loci, 25 Y-SNPs, 12 Y-STR loci.
- Retrieve paternal genealogies.
- Make a 1328×1328 distance matrix.
  - Counting the number of meioses in each pair.
- Select 100 random subsets of 100 males using different criteria:
  - Fully random.
  - Fully unrelated.
  - Random, not allowing 1 generation differences.
  - Random, not allowing 1 or 2 generation differences.
  - Random, not allowing 1..3 generation differences.

The Rucphen study

Set up:

- Random selection of 1350 males from Rucphen.
- Genotype for 15 autosomal STR loci, 25 Y-SNPs, 12 Y-STR loci.
- Retrieve paternal genealogies.
- Make a 1328×1328 distance matrix.
  - Counting the number of meioses in each pair.
- Select 100 random subsets of 100 males using different criteria:
  - Fully random.
  - Fully unrelated.
  - Random, not allowing 1 generation differences.
  - Random, not allowing 1 or 2 generation differences.
  - Random, not allowing 1..3 generation differences.
- Calculate *haplotype diversity* for Y-STRs.
- Calculate *probability of identity* for autosomal STRs.

Paternal family trees

A *linkage pedigree data* (PED) file consists of the parents of a person on each line.

Family id, Person id, Paternal id, Maternal id, Sample id.

0,	1,	0,	0,	1
0,	2,	1,	0,	2
0,	3,	1,	0,	3
0,	4,	2,	0,	4
0,	5,	2,	0,	5
0,	6,	3,	0,	6
0,	7,	3,	0,	7
0,	8,	3,	0,	8
0,	9,	4,	0,	9
0,	10,	5,	0,	10
0,	11,	5,	0,	11
0,	12,	8,	0,	12
0,	13,	8,	0,	13

Table 1: An example PED file.

Paternal family trees

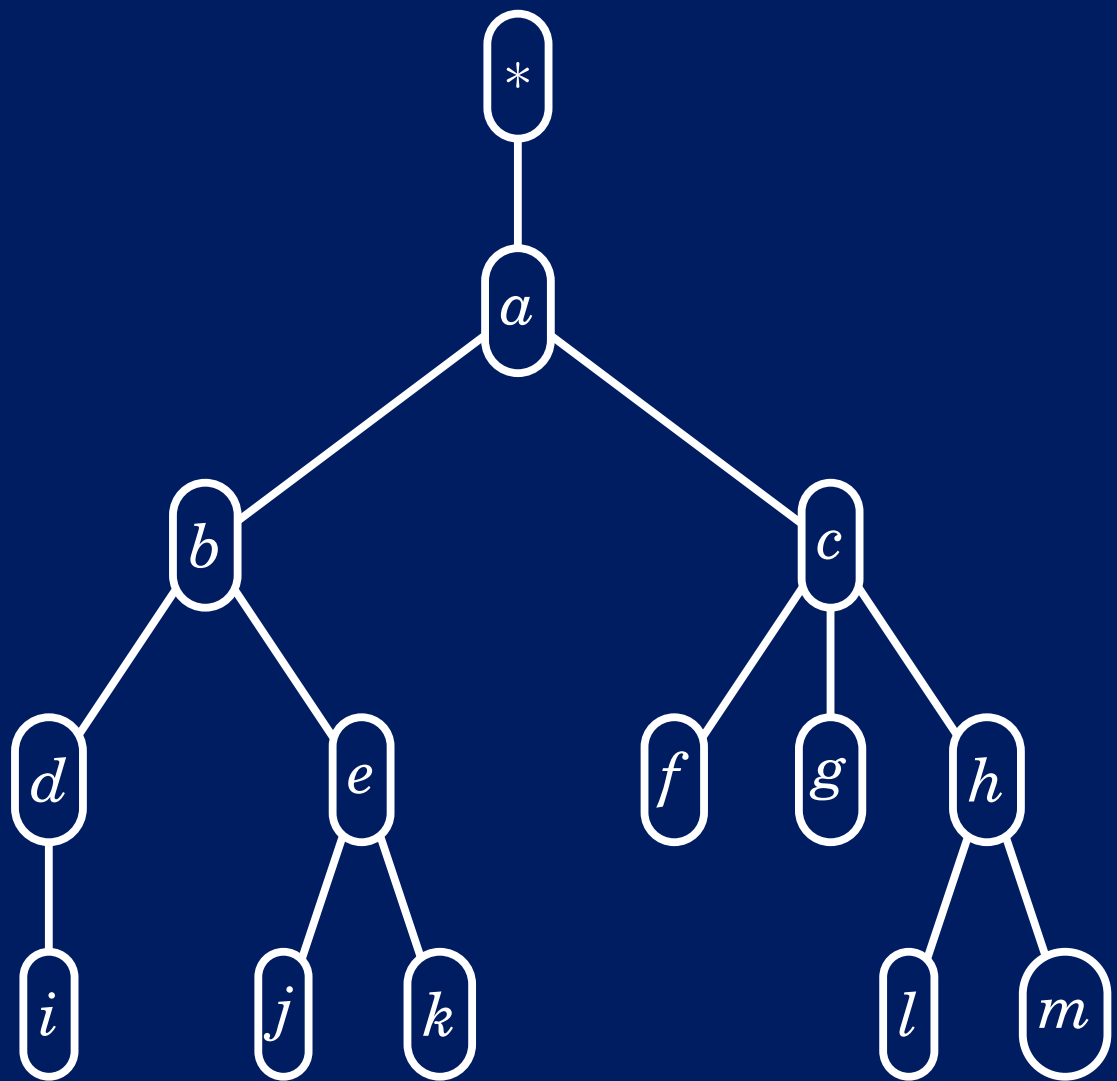


Figure 2: Paternal family tree of Table 1.

Paternal family trees

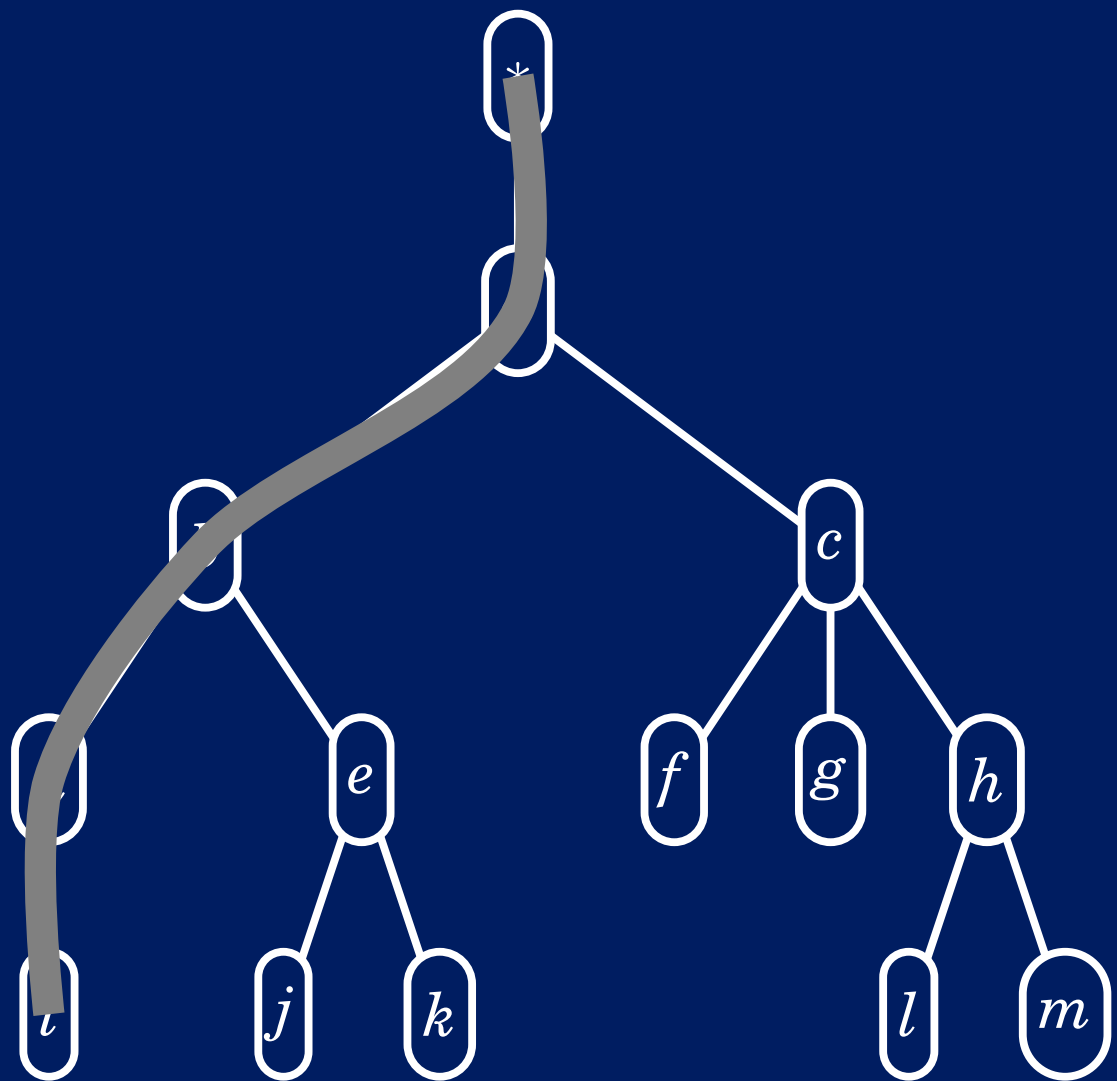


Figure 2: Paternal family tree of Table 1.

Paternal family trees

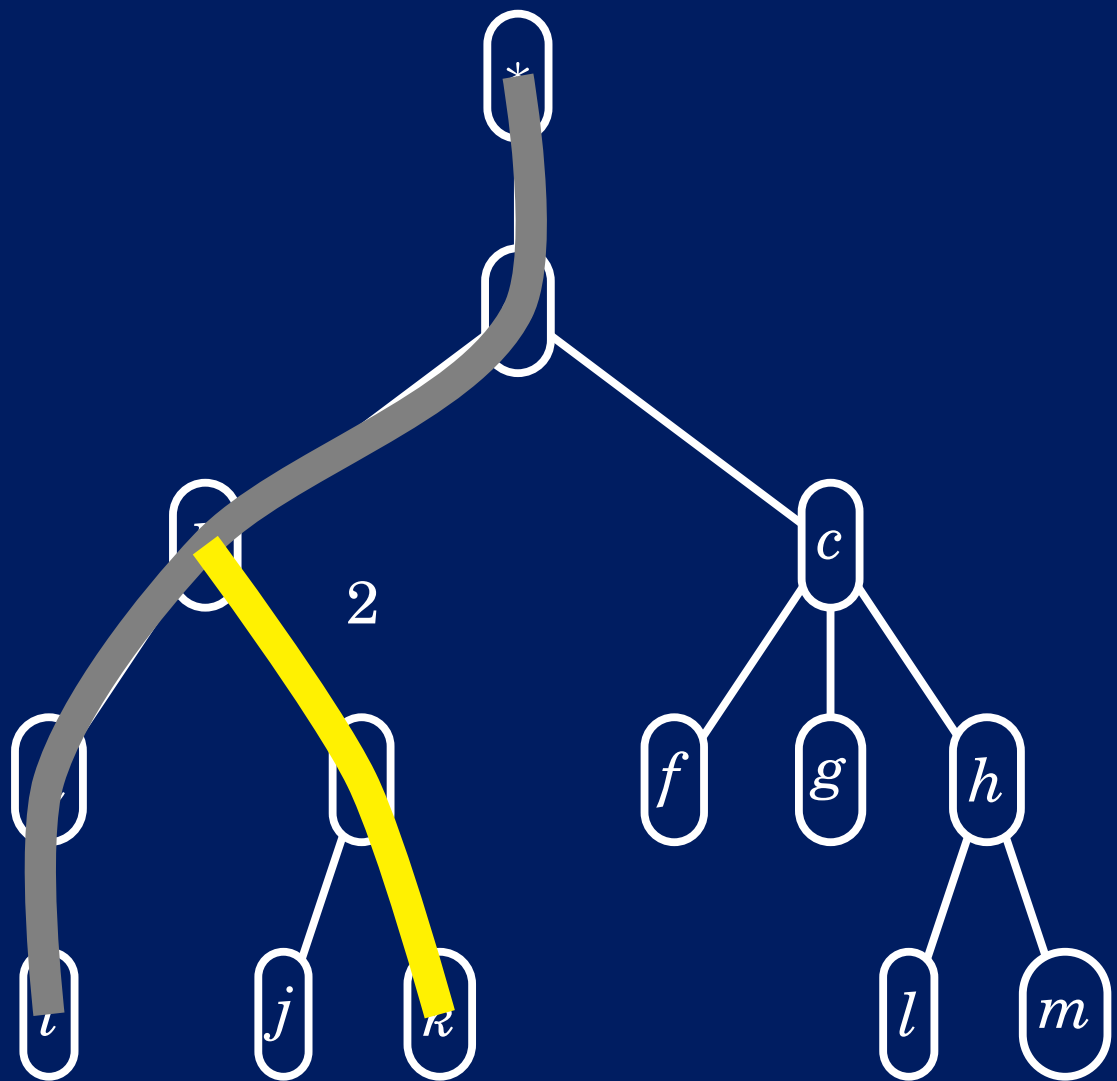


Figure 2: Paternal family tree of Table 1.

Paternal family trees

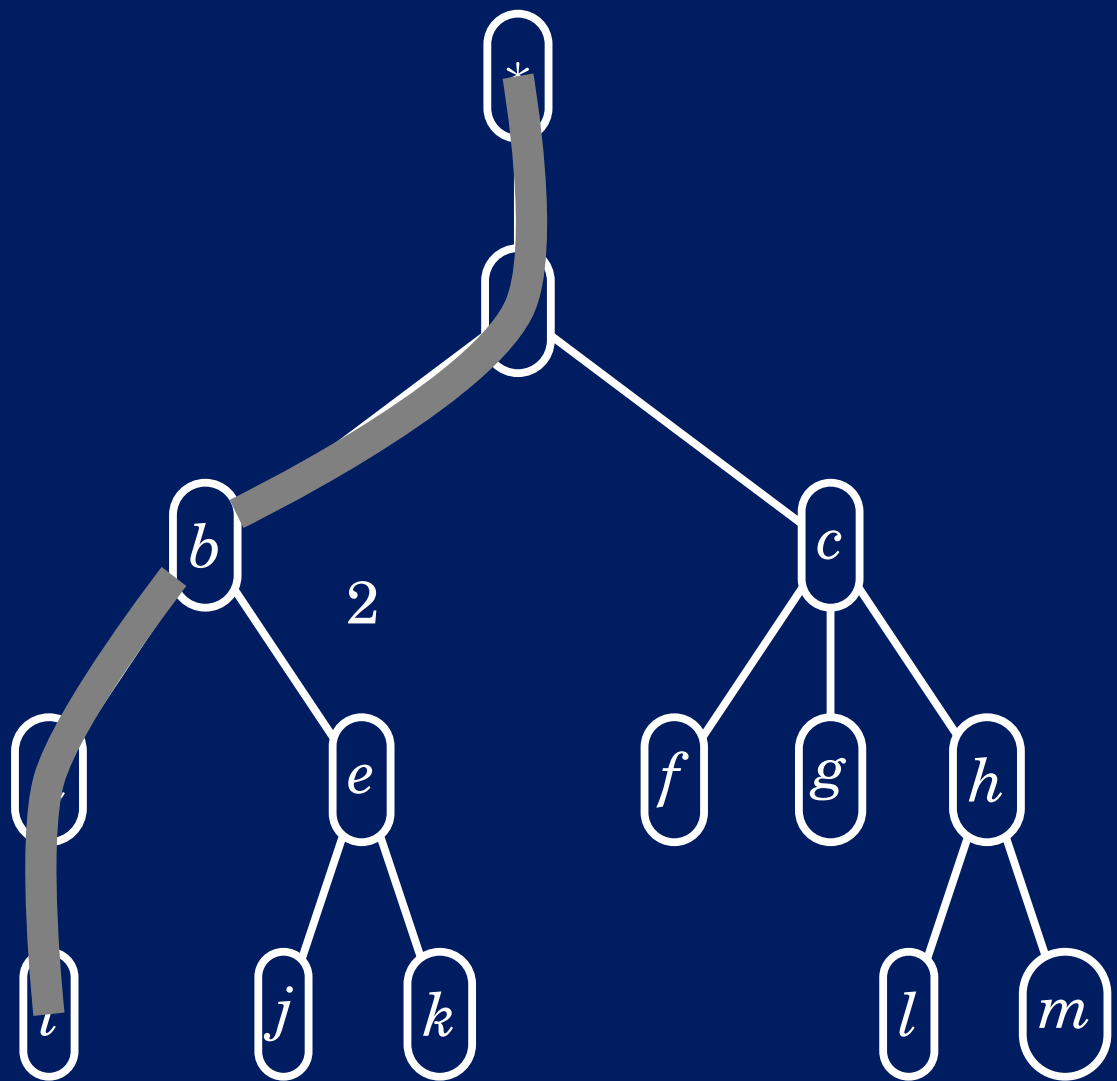


Figure 2: Paternal family tree of Table 1.

Paternal family trees

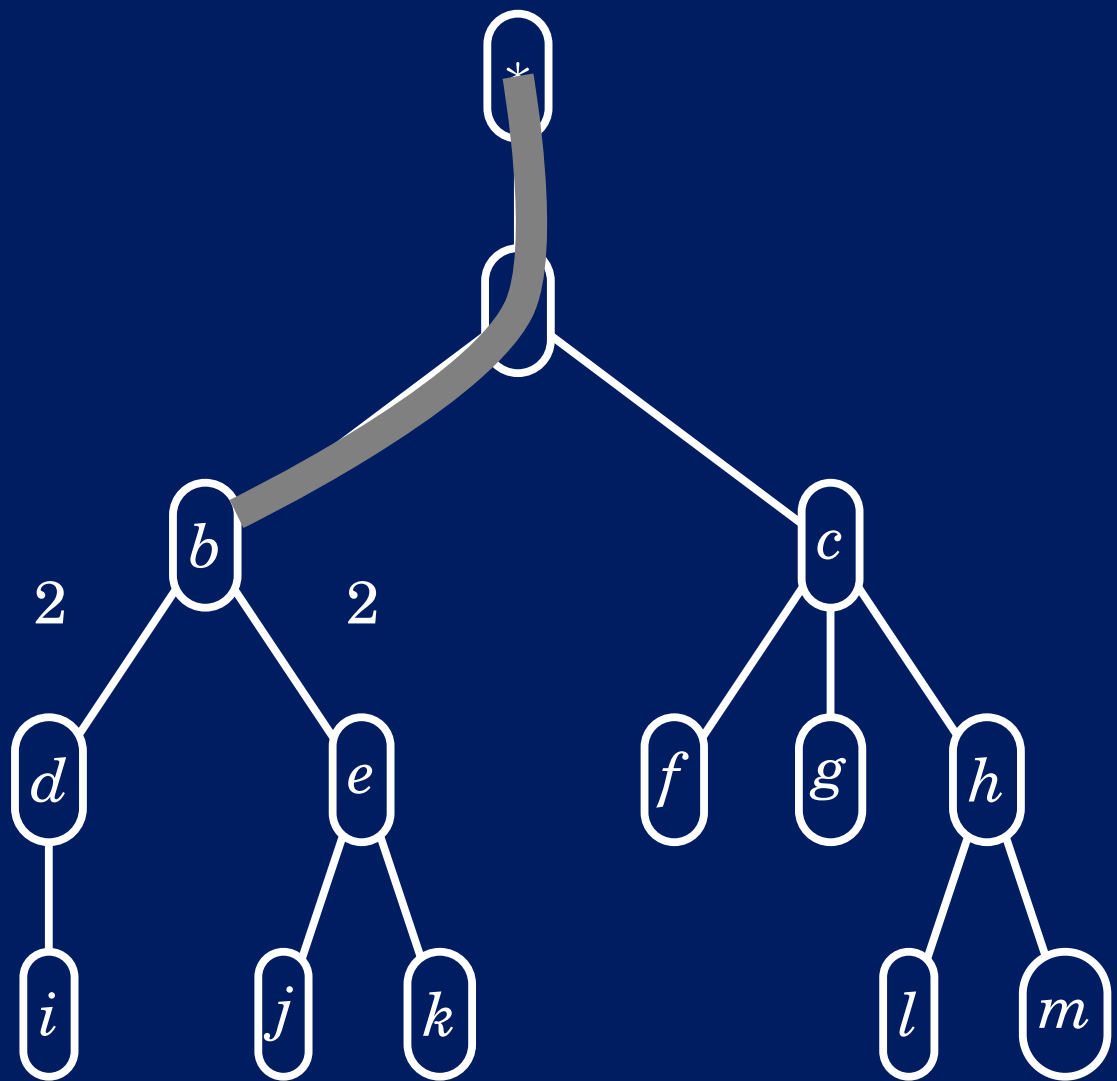


Figure 2: Paternal family tree of Table 1.

Paternal family trees

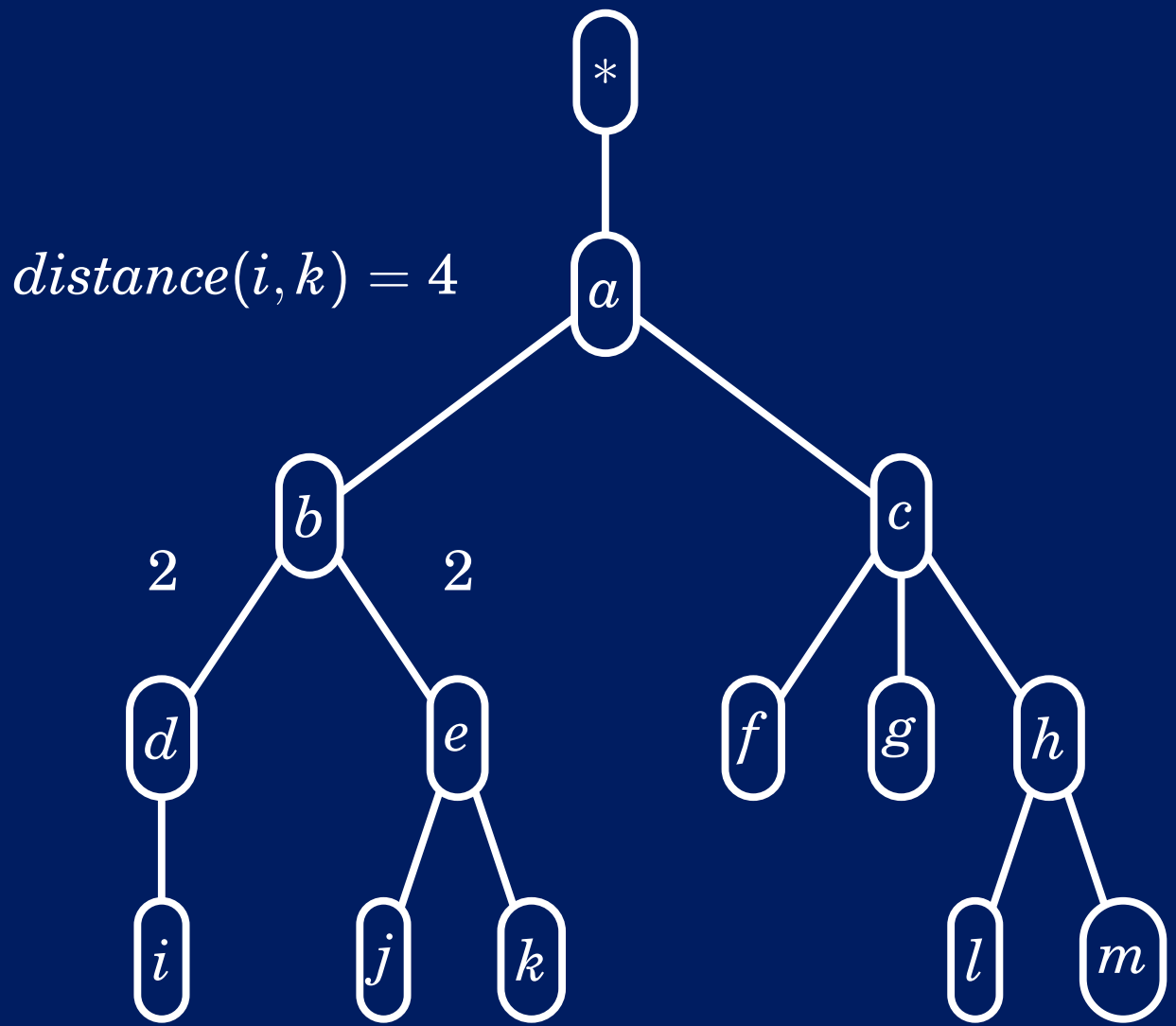


Figure 2: Paternal family tree of Table 1.

Paternal family trees

Implementation note:

A node is represented by an object which has various attributes.

- IDs (own, paternal, maternal, etc.).
- A flag to mark a visited node.
- A link (pointer) to an other object (the father).

Paternal family trees

Implementation note:

A node is represented by an object which has various attributes.

- IDs (own, paternal, maternal, etc.).
- A flag to mark a visited node.
- A link (pointer) to an other object (the father).

Initially, these objects are put in a *dictionary* (or hash table), then this collection of objects is traversed once to make the links.

Paternal family trees

Implementation note:

A node is represented by an object which has various attributes.

- IDs (own, paternal, maternal, etc.).
- A flag to mark a visited node.
- A link (pointer) to an other object (the father).

Initially, these objects are put in a *dictionary* (or hash table), then this collection of objects is traversed once to make the links.

This dual data structure allows us to quickly access any node in the tree and determine the path to the root of the tree easily.

Paternal family trees

dist(x,y)	a	b	c	d	e	f	g	h	i	j	k	l	m
a	0	1	1	2	2	2	2	2	3	3	3	3	3
b	1	0	2	1	1	3	3	3	2	2	2	4	4
c	1	2	0	3	3	1	1	1	4	4	4	2	2
d	2	1	3	0	2	4	4	4	1	3	3	5	5
e	2	1	3	2	0	4	4	4	3	1	1	5	5
f	2	3	1	4	4	0	2	2	5	5	5	3	3
g	2	3	1	4	4	2	0	2	5	5	5	3	3
h	2	3	1	4	4	2	2	0	5	5	5	1	1
i	3	2	4	1	3	5	5	5	0	4	4	6	6
j	3	2	4	3	1	5	5	5	4	0	2	6	6
k	3	2	4	3	1	5	5	5	4	2	0	6	6
l	3	4	2	5	5	3	3	1	6	6	6	0	2
m	3	4	2	5	5	3	3	1	6	6	6	2	0

Table 2: Distance of all persons in the tree of Figure 4.

More than one family in the same tree

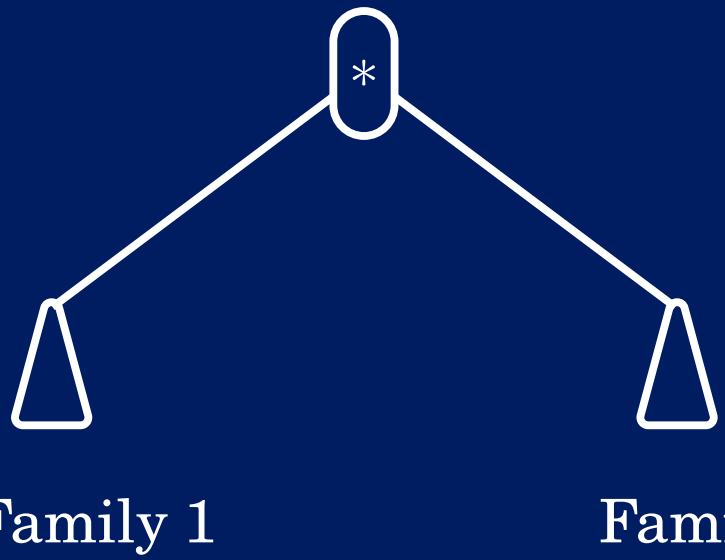
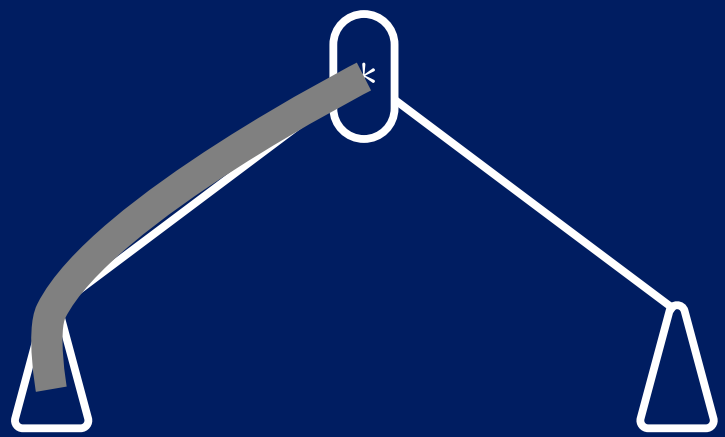


Figure 3: Combining families in one tree.

When we encounter the $*$ node with the yellow line, we know people are unrelated.

More than one family in the same tree



Family 1 Family 2

Figure 3: Combining families in one tree.

When we encounter the $*$ node with the yellow line, we know people are unrelated.

More than one family in the same tree

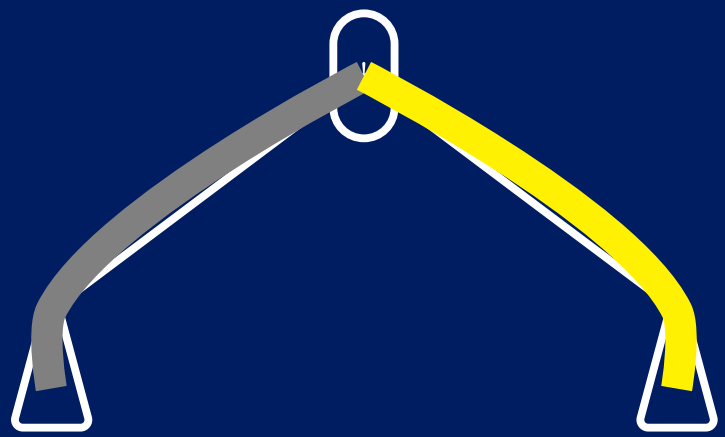


Figure 3: Combining families in one tree.

When we encounter the $*$ node with the yellow line, we know people are unrelated.

Note that this is not very efficient, but trivial to implement.

Relatedness in our selection

The resulting 1328×1328 matrix gives the relatedness of each pair of persons in our selection.

Generations	Count	Generations	Count
1	343	13	409
2	376	14	98
3	440	15	70
4	591	16	42
5	562	17	21
6	824	18	184
7	905	19	438
8	822	20	370
9	916	21	116
10	1100	22	12
11	1131	23..99	0
12	852	100	870506

Table 3: Summary generations matrix.

Note that we use 100 to indicate unrelatedness.

Selecting people on relatedness

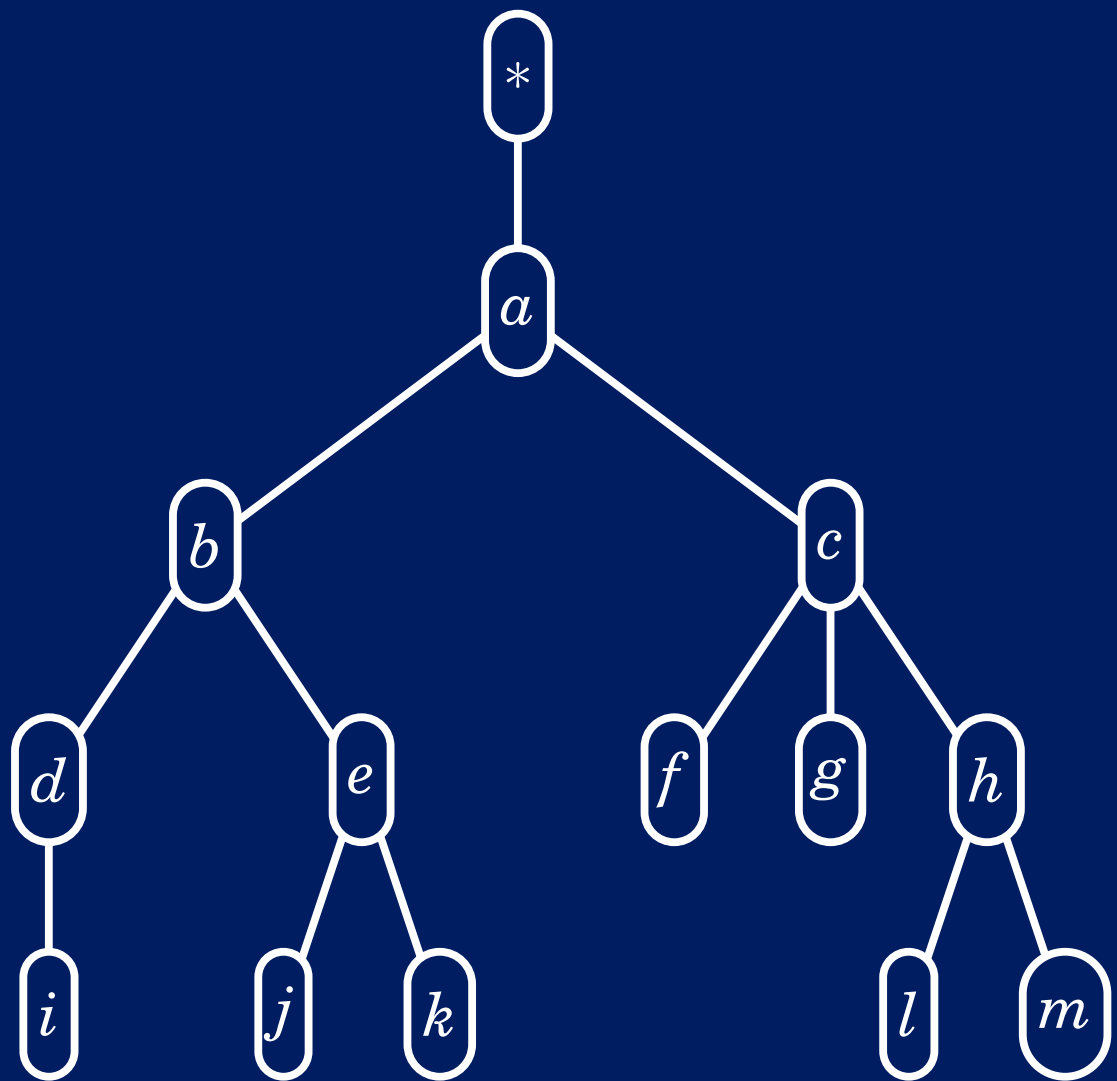


Figure 4: Paternal family tree of Table 1.

Selecting people on relatedness

Person i and everyone within distance 3 removed.

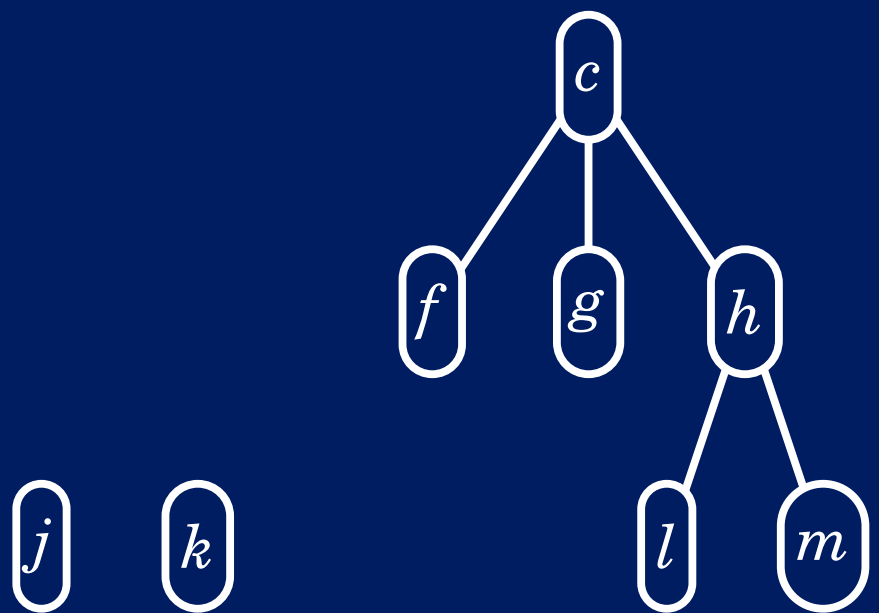


Figure 4: Paternal family tree of Table 1.

Selecting people on relatedness

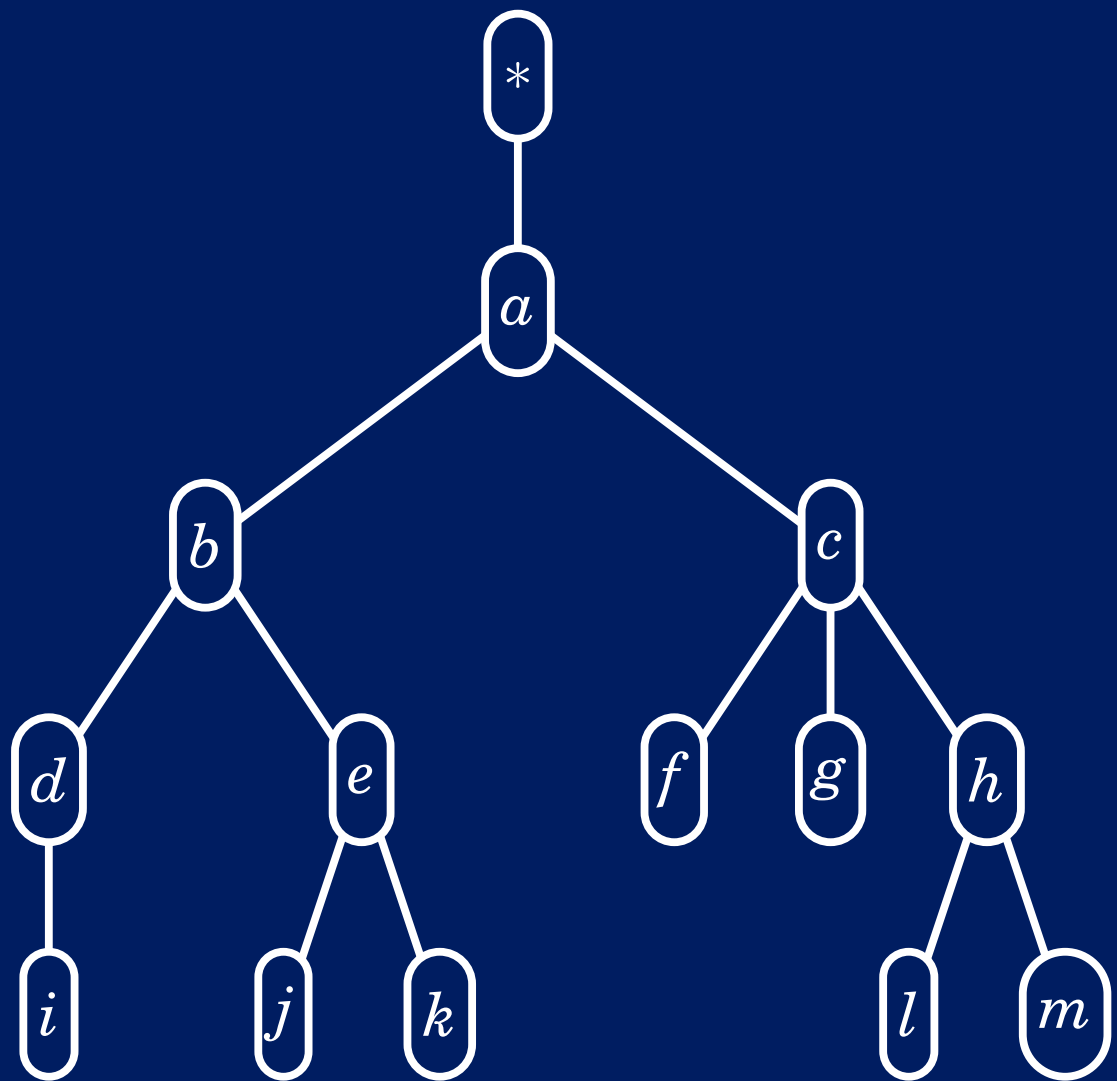


Figure 4: Paternal family tree of Table 1.

Selecting people on relatedness

Person a and everyone within distance 3 removed.

Figure 4: Paternal family tree of Table 1.

Selecting people on relatedness

We make a list of persons that are within a certain distance.

Person	Family			
a	b	c		
b	a	d	e	
c	a	f	g	h
d	b	i		
e	b	j	k	
f	c			
g	c			
h	c	l	m	
i	d			
j	e			
k	e			
l	h			
m	h			

Table 4: Family members within distance 1.

This can easily be extracted from the distance matrix.

Selecting people on relatedness

We make a list of persons that are within a certain distance.

Person	Family			
c	a	f	g	h
f	c			
g	c			
h	c	l	m	
i	d			
j	e			
k	e			
l	h			
m	h			

Table 5: Table 4 after selecting person b.

Removed persons can still have a reference.

Results

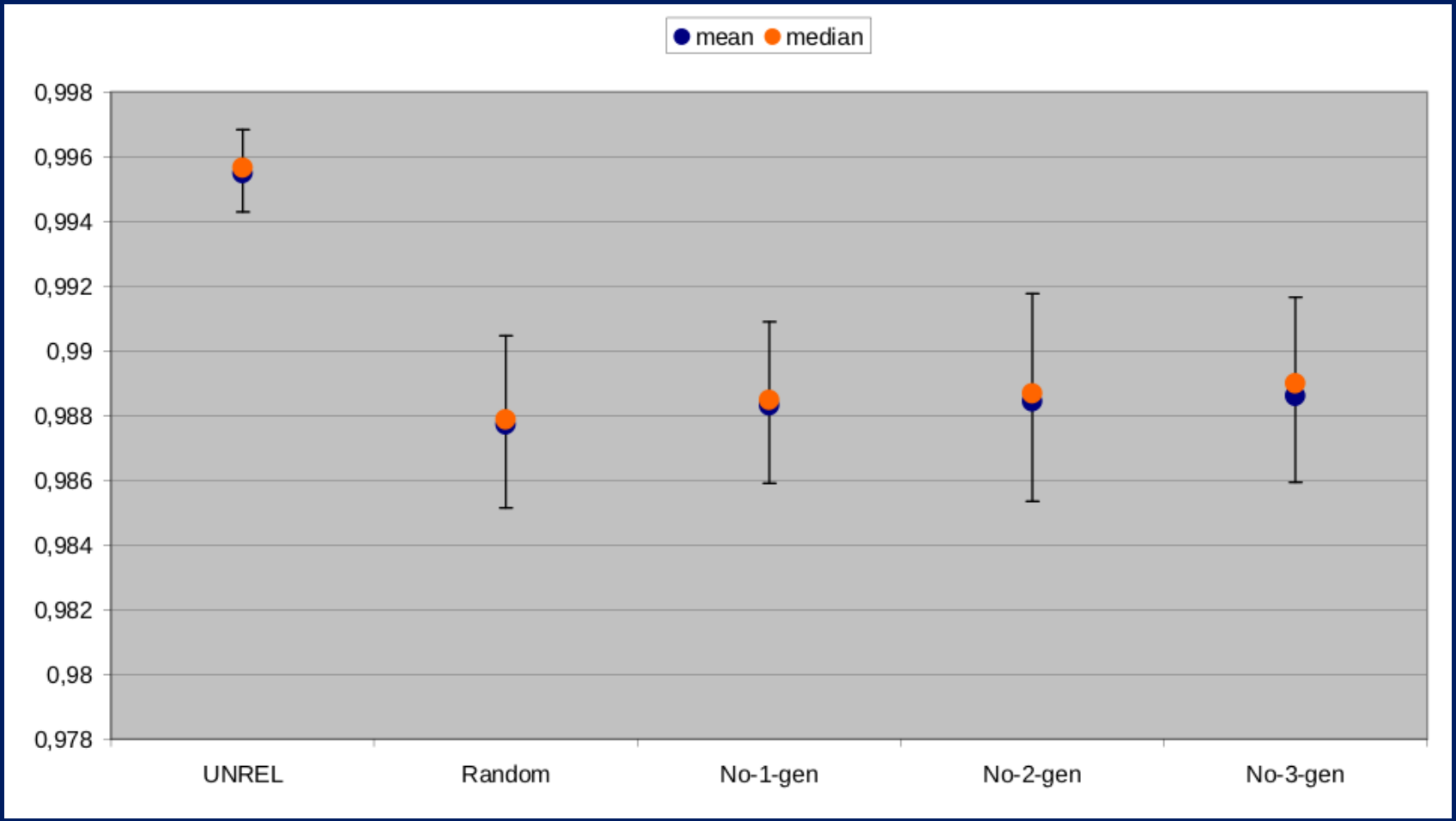


Figure 5: Chromosome-Y haplotype diversity.

Results

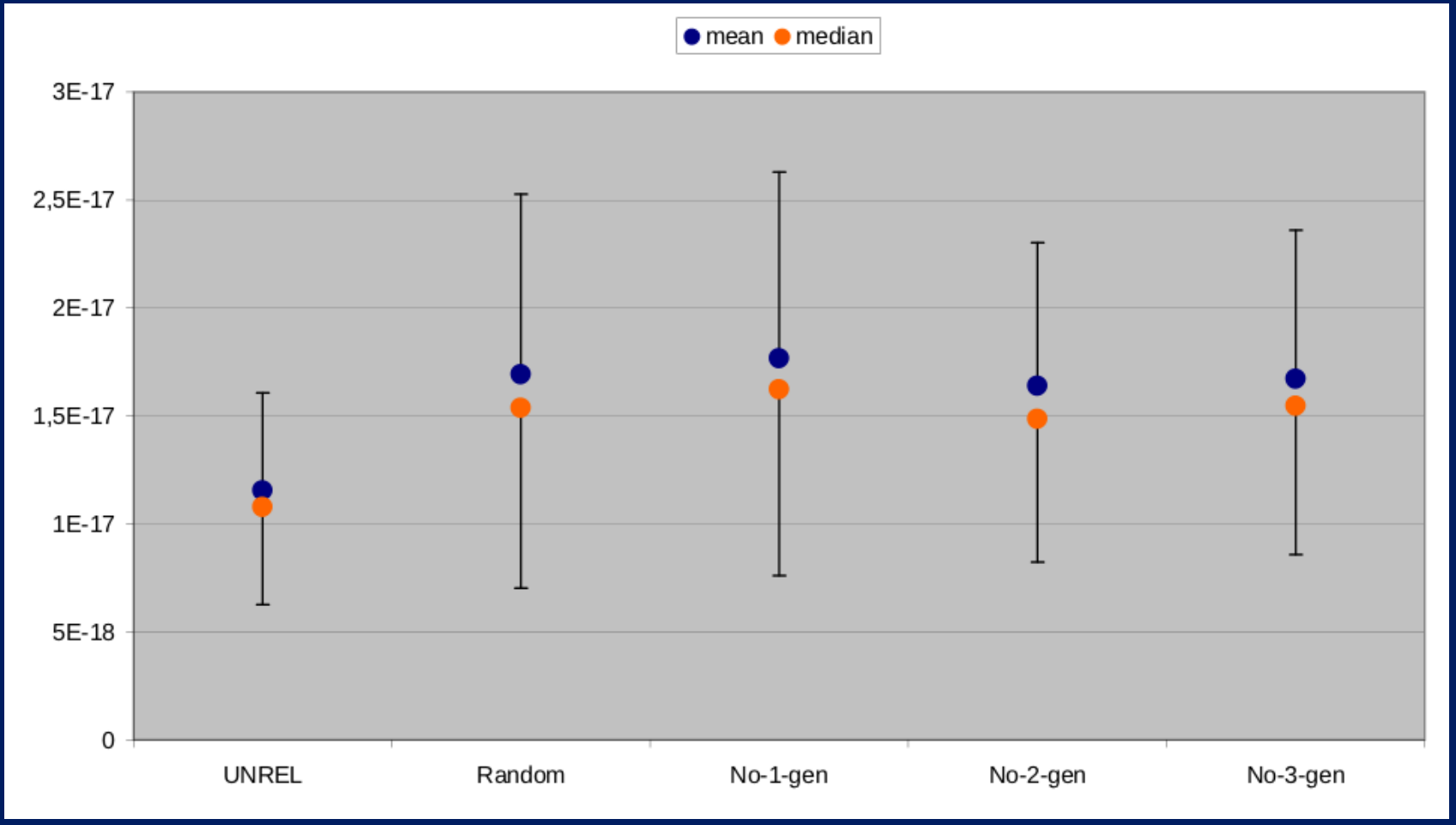


Figure 6: Probability of identity (autosomal STRs only).

Conclusions and further research

We have empirically proven the following:

- Drawing a random sample from a population using the classical selection rule does not result in a truly random subset.
    - This has little effect when using autosomal STR markers.
    - The effect is significant when using Y-STR markers.

Conclusions and further research

We have empirically proven the following:

- Drawing a random sample from a population using the classical selection rule does not result in a truly random subset.
  - This has little effect when using autosomal STR markers.
  - The effect is significant when using Y-STR markers.

Possible solution:

- Use rapidly mutating Y-STRs.

Conclusions and further research

We have empirically proven the following:

- Drawing a random sample from a population using the classical selection rule does not result in a truly random subset.
  - This has little effect when using autosomal STR markers.
  - The effect is significant when using Y-STR markers.

Possible solution:

- Use rapidly mutating Y-STRs.

Questions:

- Are the most abundant haplotype frequencies underestimated?

Questions?

Acknowledgements:

Thirsa Kraayenbrink
Manfred Kayser
Peter de Knijff

https://www.mutalyzer.nl/svn/generations/