Duncan Taylor

duncan.taylor@sa.gov.au

Hardy Weinberg equilibrium
Allele frequencies
Genotype probabilities
Confidence intervals
Databases
Linkage

forensic science sa

When a DNA profile obtained from a crime scene and there is a reference DNA profile to compare we need some way to assess the significance of the similar or different allelic information

A statistical weighting

The way that this statistical weighting is generated has a lot of population genetics and statistics behind it

I will just scratch the surface of some of the key ideas

forensic science sa

Hardy Weinberg Equilibrium

forensic science sa

To apply a statistical weighting we must know the 'rarity' of the DNA profile we are examining.

This is governed by the 'rarity' of the components (alleles) that make up the reference DNA profile in the population of interest:

$p^2+2pq+q^2=1$

Works for 'ideal' population i.e. one that is in Hardy Weinberg equilibrium

Genotype frequencies are constant between generations and all frequencies sum to 1

forensic science sa

So for example – we wanted to calculate a profile frequency for a homozygous locus [A,A] where the frequency of [A] is 0.3:

Profile frequency = $p^2 = (0.3)^2 = 0.09$

Or Heterozygous locus [A,B] where the frequency of [A] is 0.3 and [B] is 0.7:

Profile frequency = $2pq = 2(0.3)(0.7) = 0.42$

Why is 2 out the front here for heterozygotes?

forensic science sa

In our population where p = 0.3 and q = 0.7 we would expect genotypes in the following proportions:

	A	B
A	AA (0.3^2)=0.09	AB (0.3x0.7)=0.21
B	AB (0.7x0.3)=0.21	BB (0.7^2)=0.49

A heterozygote individual who contains an [A] and a [B] could actually be [A,B] or [B,A] = pq + qp = 2pq

And $p^2+2pq+q^2=(0.3)^2+2(0.3)(0.7)+(0.7)^2$ = 0.09+0.42+0.49 = 1

forensic science sa

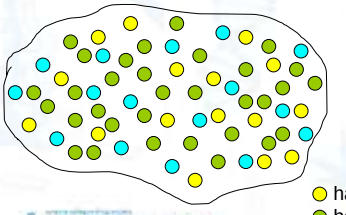
This idea can be extended to multi-allele scenarios, e.g. consider a population with frequencies A=0.2, B=0.3, C=0.2, D=0.2, E=0.05, F=0.05

	A	B	C	D	E	F
A	0.2^2=0.04	0.2x0.3=0.06	0.2x0.2=0.04	0.2x0.2=0.04	0.2x0.05=0.01	0.2x0.05=0.01
B	0.2x0.3=0.06	0.3^2=0.09	0.3x0.2=0.06	0.3x0.2=0.06	0.3x0.05=0.15	0.3x0.05=0.15
C	0.2x0.2=0.04	0.3x0.2=0.06	0.2^2=.0.4	0.2x0.2=0.04	0.2x0.05=0.01	0.2x0.05=0.01
D	0.2x0.2=0.04	0.3x0.2=0.06	0.2x0.2=0.04	0.2^2 =0.04	0.2x0.05=0.01	0.05^2=0.0025
E	0.2x0.05=0.01	0.3x0.05=0.15	0.2x0.05=0.01	0.2x0.05=0.01	0.05^2=0.0025	0.05^2=0.0025
F	0.2x0.05=0.01	0.3x0.05=0.15	0.2x0.05=0.01	0.2x0.05=0.01	0.05^2=0.0025	0.05^2=0.0025

forensic science sa

Hardy Weinberg Island

These genotype proportions are only valid under the assumption of Hardy Weinberg equilibrium. i.e. random mating, large population, no mutation, no selection and no migration.



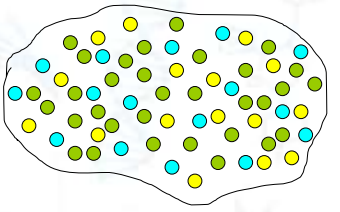
This island population is in Hardy Weinberg equilibrium (imagine there are an infinite number of dots)

- have the genotype [Y,Y]
- have the genotype [B,Y]
- have the genotype [B,B]

forensic science sa

Hardy Weinberg Island

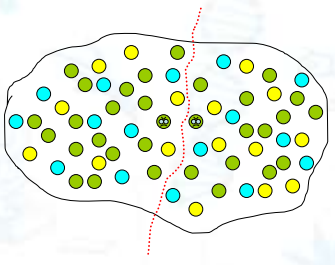
This is the same population after 50 generations



Because the population is in HWE genotype frequencies remain constant and the population remains the same

forensic science sa

Hardy Weinberg Island



Oh No! Someone from West HWI has started a feud with someone from East HWI.

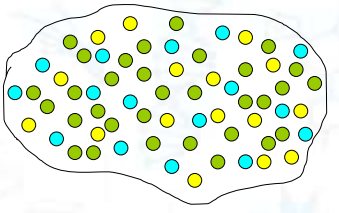
The island is divided

The population sub-structure now means that there is no more random mating (which violates one of the requirements for HWE)

forensic science sa

Hardy Weinberg Island

Move forward by 1 generation



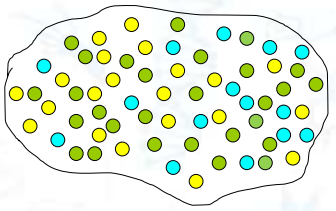
After one generation there would be no difference

So lets speed it up a little and power through a few more generations

forensic science sa

Hardy Weinberg Island

Generation 10



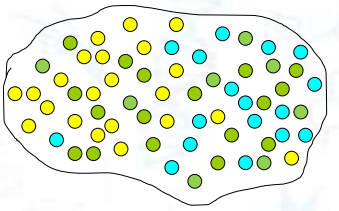
Lets also introduce mutation and genetic drift.

Some minor differences seen in some individuals on either side of the island

forensic science sa

Hardy Weinberg Island

Generation 30

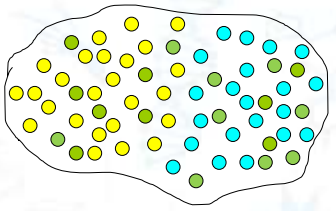


differences increasing

forensic science sa

Hardy Weinberg Island

Generation 50

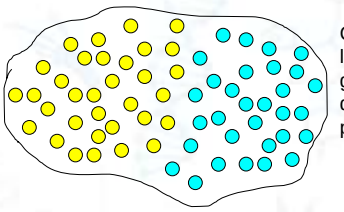


differences increasing

forensic science sa

Hardy Weinberg Island

Generation X

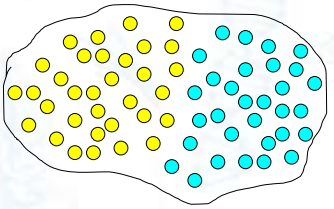


Genetic drift has lead to 2 genetically distinct populations

What would happen if we still assumed HWE ??

forensic science sa

Hardy Weinberg Island



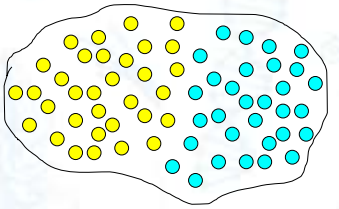
A crime is committed on West HW island and the evidence at the scene is [Y,Y]

The HW police have a suspect who is also [Y,Y]

- [Y,Y]
- [B,B]

forensic science sa

Hardy Weinberg Island



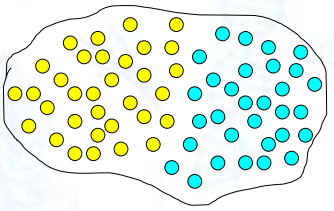
The HW forensic science centre knows that the frequency of [Y] is 0.5 and [B] is 0.5 and conclude that the frequency of [Y,Y] (i.e. a yellow dot person) is:

$p^2 = 0.5^2 = 0.25$

- [Y,Y]
- [B,B]

forensic science sa

Hardy Weinberg Island



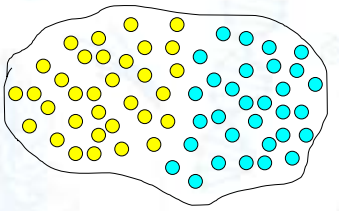
[Y,Y]
[B,B]

Frequency [Y,Y] = 0.25

So the report says "*The chance of seeing this genotype in another unrelated individual is 1 in 4*"

forensic science sa

Hardy Weinberg Island



[Y,Y]
[B,B]

But <u>every</u> individual on west HW island is [Y,Y] and so the chance of seeing this profile again in another east HW island is 1 in 1.

By assuming the population is in HWE the strength of the evidence has been grossly overstated.

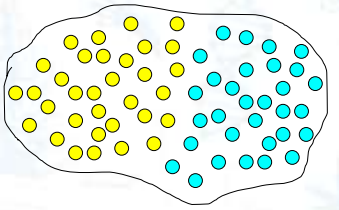
i.e. allele frequencies cannot be used to accurately determine genotype frequencies any longer

The imperfect world

forensic science sa

Reality Island

Ideally we would recognise the fact that the population is not in HWE and we would calculate the statistic in a database of only west HW Island individuals.

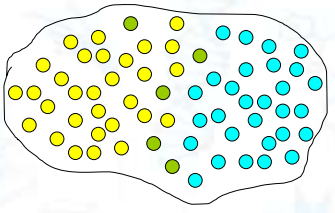


Reality differs from the ideal model in a number of ways.

We will pick up the population from where we left off

forensic science sa

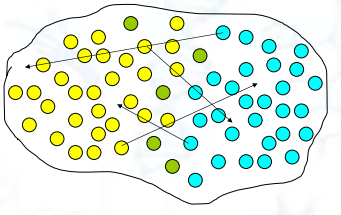
Reality Island



Old boundaries forgotten – there may be some mating between populations at the border

forensic science sa

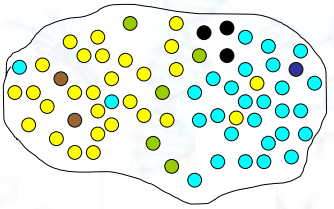
Reality Island



People may migrate over the border and modern technology allows travel from one side of the island to the other

forensic science sa

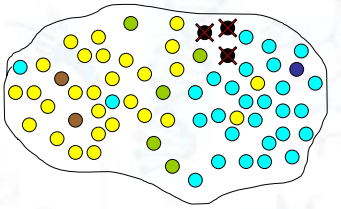
Reality Island



Rare mutations will occur, which will make some individuals very different from the general population

forensic science sa

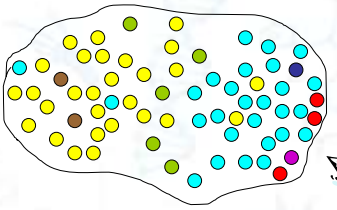
Reality Island



Selection comes into play when disease hits the island and it turns out one of the genotypes is susceptible

forensic science sa

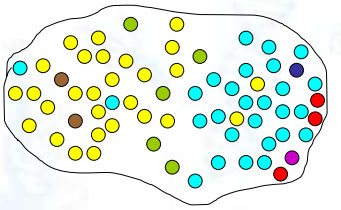
Reality Island



Migrations from far off lands will introduce exotic genotypes into the island

forensic science sa

Reality Island



Now the picture is much more like reality.

Clear boundaries do not exist but general trends can be seen

forensic science sa

Population databases

forensic science sa

In reality we aren't able to obtain the genotypes for every individual in the population.

Practically we can only obtain samples from a small percentage of the population, and then use that to draw inferences about the whole population.

The subsets of individuals are called population databases, and they are used to determine allele frequencies.

forensic science sa

How we generate a population database in practise

Convenience – The sample is not truly random, it is built from a collection of individuals that we have profiled in the course of casework.

In reality the markers that we examine are not selective with respect to crime and so a convenience database can be considered effectively random

Self-declared – the ethnicity of the individuals are on the basis of self declaration

Once again this isn't perfect as there is no checking of the claims of the suspects, however this type of collection has been shown to produce databases that are for use in a forensic setting

forensic science sa

How we generate a population database in practise

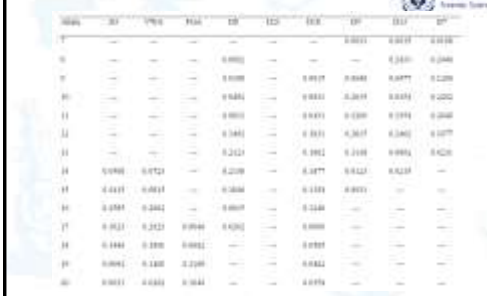
Once the individuals have been chosen their profiles are compiled and the number of occurrences of each allele counted and divided by the total number of alleles to determine allele frequencies

These allele frequencies are listed in a table in such a way that the data is easily accesable

Using the example:
Taylor DA, Henry JM, Walsh SJ. *South Australian Aboriginal sub-population data for the nine AMPFlSTR Profiler Plus short tandem repeat (STR) loci.* Forensic Sci Int Genet. 2008 Mar;2(2):e27-30.

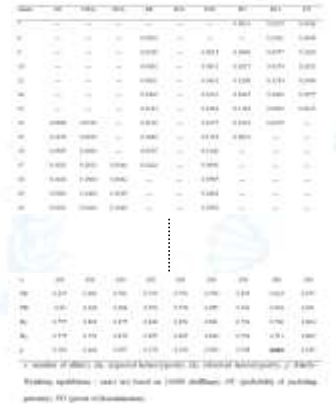
forensic science sa

How we generate a population database in practise



A number of tests are then performed on the data

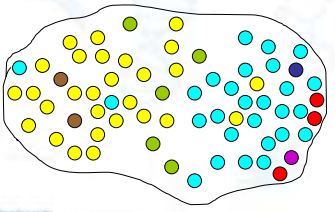
forensic science sa



- count of the number of alleles at each locus
- The expected and observed heterozygosities
- The Fischer's exact tests for adherence to HWE
- The probability of excluding paternity
- The probability of discrimination

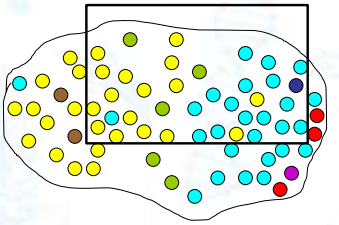
Back to our earlier example

We will go through some of the population database construction and considerations



forensic science sa

Population Databases



The box shows the individuals sampled in our population database

- 16 [Y,Y]
- 17 [B,B]
- 3 [B,Y]
- 1 [B,D]

37 In total

forensic science sa

Population Databases – Allele frequencies

16 [Y,Y]	32 [Y] alleles
17 [B,B]	34 [B] alleles
3 [B,Y]	3 [B] alleles and 3 [Y] alleles
1 [B,D]	1 [B] allele and 1 [D] allele
37 In total	

74 alleles in total

Allele frequencies can be calculated by: $\frac{\text{\# alleles}}{\text{Total \# alleles}}$

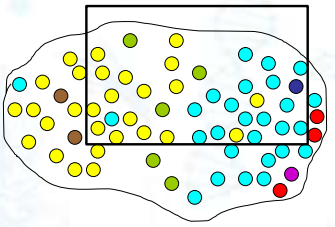
e.g. for [Y] frequency (p_Y)= $\frac{32+3}{74}$

= 0.473 Or 47.3%

Similarly p_B= 51.4% and p_D = 1.3%

forensic science sa

Population Databases



There is an obvious problem with this database

It spans two populations and so contains substructure

But how could you tell just from the data that was collected ?

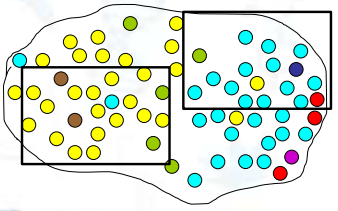
We will come back to this a bit later

forensic science sa

Population Databases

A better approach would be, once it is recognised that there is some structure in the population, to sample the two populations separately e.g.

Population Database 1



Population Database 2

forensic science sa

Population Databases

These databases still aren't in HWE, but will give closer estimates of genotype frequency from the allele frequencies.

They may or may not show departures from HWE when tested for departures from equilibrium using the Fisher's Exact test.

The Fisher's Exact test is quite weak in its ability to detect these departures

forensic science sa

Population Databases

We expect human populations to depart from HWE (even if we don't detect any)

Why is this ? Because we violate the assumptions required for HWE, i.e.

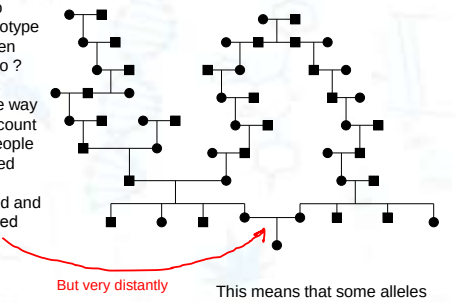
- We don't randomly mate
- We immigrate and emigrate
- Mutations occur
- There may be some selective pressures
- Our population size isn't infinite

forensic science sa

Population Databases

So if we can't use the HWE formulae (p^2 and 2pq) to determine genotype frequencies then what can we do ?

We need some way to take into account the fact that people in a finitely sized population are distantly related and hence……inbred



But very distantly

This means that some alleles are more common (and by the inverse others would be rarer) than under HWE estimates

forensic science sa

Inbreeding
Coancestry
Substructure

forensic science sa

Theta

DNA profiles from current forensic profiling kits can have frequencies in the order of 1 in 10^{22}

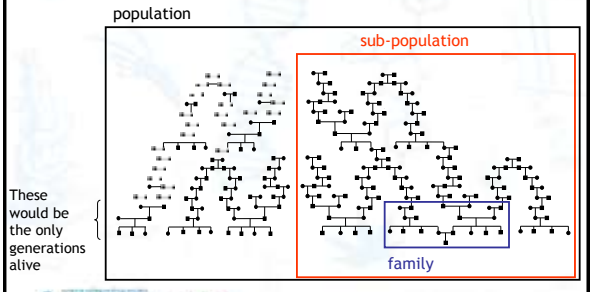
This means that it each DNA profile is incredibly rare

However given that we have seen a profile already, it makes it more likely to see that profile again

This is because populations are not infinitely large and the choice of mate is not random (violations of the HWE model)

forensic science sa

Theta

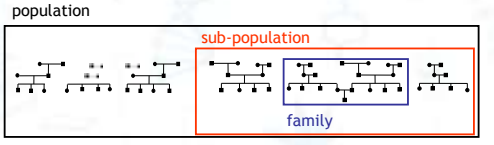


forensic science sa

Theta

Looking at the population that is alive at the present day we would see a picture more like that shown below

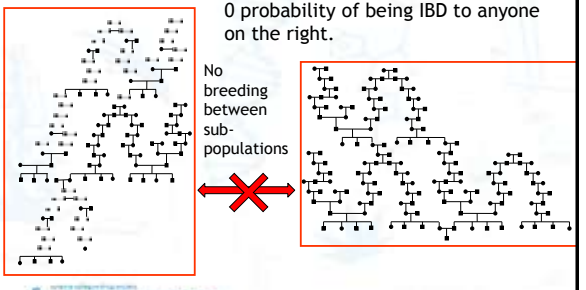
Regardless of which individuals breed within a subpopulation there is going to be a distant level of relatedness between them (theta)



forensic science sa

Theta

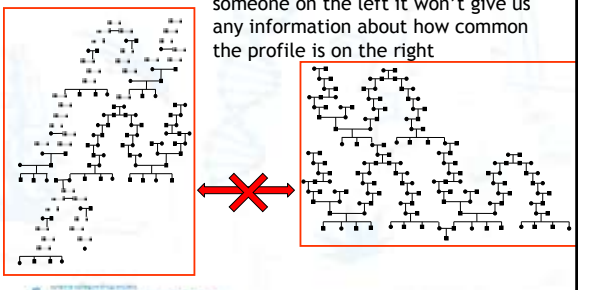
An allele from anyone on the left has 0 probability of being IBD to anyone on the right.



forensic science sa

Theta

This means if we have a profile from someone on the left it won't give us any information about how common the profile is on the right

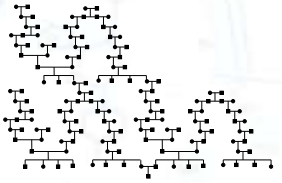


forensic science sa

Theta

Within a sub-population there are only a finite number of individuals, and so eventually a level of inbreeding within the sub-population will build up

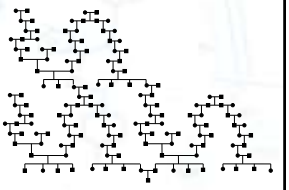
Inbreeding reduces genetic diversity and drives loci towards fixation (only containing one allele)



forensic science sa

Theta

This effect is balanced out by mutation, which introduces new alleles into the population



The effect on 'normal' sized human populations is that some alleles are more common in certain sub-populations than in others.

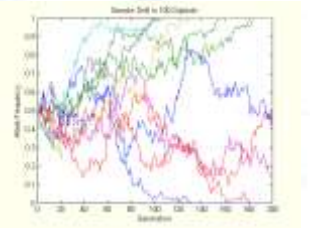
forensic science sa

Genetic drift causes allele frequencies to change over the generations.

Some alleles will drift to become more common

Others will drift to be lost in the population

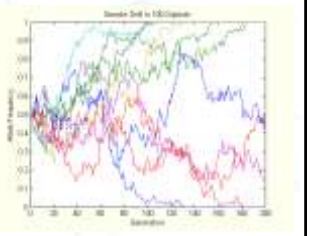
If only genetic drift was acting then eventually all loci would become fixed for an allele



forensic science sa

However mutation counteracts drift by introducing new alleles into the population

When two sub-populations are reproductively isolated then genetic drift will act independently on both and cause allele frequencies to differ



forensic science sa

Sub-populations

The smaller the isolated population, the quicker the effects of genetic drift will be

Inbreeding magnifies the effect of genetic drift by reducing the effective population size

forensic science sa

Lets look at smallest possible populations to show how inbreeding causes populations to develop different allele frequencies

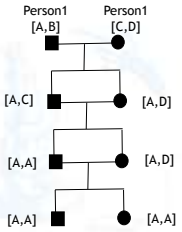
2 men and 2 women who form a very small population

Person1	Person2	Person3	Person4
[A,B]	[C,D]	[A,B]	[C,D]
■	●	■	●

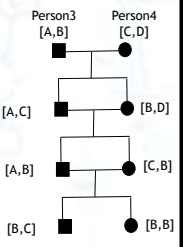
Person1 says – I only want to breed with person2
Person2 – ok
Person3 – Well then I only want to breed with person4
Person4 – ok

forensic science sa

p_A=0.25, p_B=0.25, p_C=0.25, p_D=0.25 For both sub-pops



Alleles randomly passed on to offspring cause allele frequencies between the two subpopulations to differ



p_A=1, p_B=0, p_C=0, p_D=0

p_A=0, p_B=0.75, p_C=0.25, p_D=0

forensic science sa

1. Common theta used in forensic calculations:
   - Caucasian θ = 0.01 (1%)
   - Aboriginal θ = 0.03 (3%)
2. Thetas that correspond to familial relationships
   - First cousins θ = 0.063 (6.3%)
   - Second cousins θ = 0.016 (1.6%)

Notice theta use in forensic setting are high compared to what you would might expect (i.e. the theta of 1% for Caucasians suggests that we might be regularly inbreeding to a level close to second cousins)

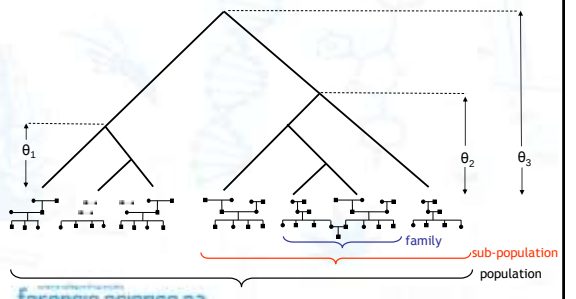
forensic science sa

- We always try to concede as much doubt to the defendant as is reasonable- i.e. the higher the level of theta, the more potential inbreeding we are accounting for and this will make the profile of interest more likely to be seen again in that population

- Theta's use is two-fold. As well as a co-ancestry coefficient it can also be used as a measure of the genetic distance between populations. So we can use a higher value to take into account the fact that relevant database might be different to our allele freq database

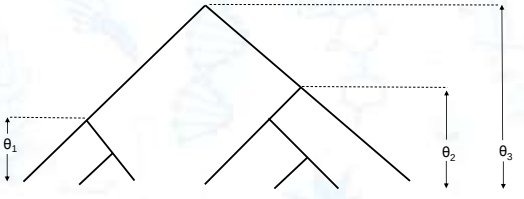
forensic science sa

Theta as a measure of distance between populations also explains why theta is higher in Aboriginal groups than for Caucasian groups.

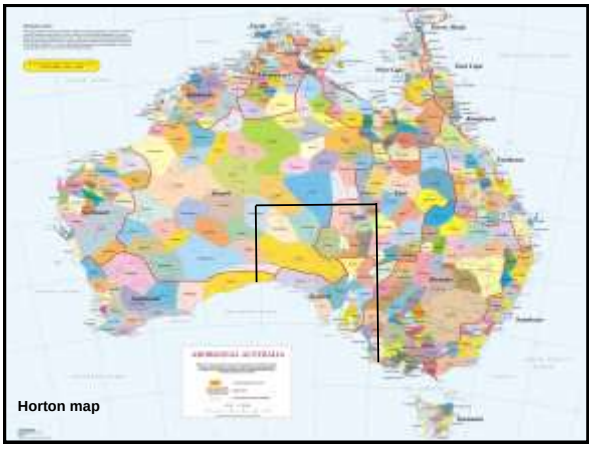


forensic science sa

θ_1 is needed to describe the diversity within subpopulation 1
θ_2 is needed to describe the diversity within subpopulation 2
θ_3 is needed to describe the diversity within the entire population

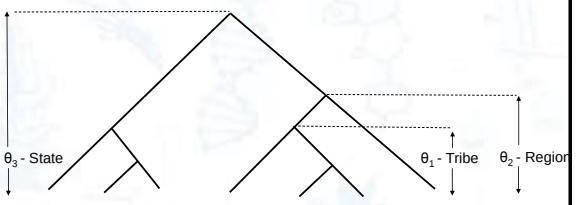


Now we need to look at the population structure of Aboriginal and Caucasian Australians



Horton map

θ_1 is needed to describe the diversity within a tribe
θ_2 is needed to describe the diversity within a traditional regional group
θ_3 is needed to describe the diversity within the entire state



forensic science sa

Caucasian are a lot more boring…

Tend to be the same all over the world, with very little geographic substructure

This means that a smaller theta can be used to cover the genetic diversity within Caucasians

forensic science sa

Match Probability

forensic science sa

Matching statistics – Match Probability

We need some way of taking this inbreeding into account, and we do this with the use of θ, a co-ancestry coefficient (also called theta, or F_{ST})

θ was most knowingly incorporated into a matching statistic in a 1994 paper by Balding and Nichols (Forensic Science International 64:125-140, 1994)

Strictly speaking the definition of theta is "*The proportion of times that two alleles randomly chosen in a population will be Identical By Descent (IBD)*"

forensic science sa

Matching statistics – Match Probability

IBD alleles are when two alleles of the same designation have originated from a common ancestor, rather than by mutation

So we can now move away from profile frequencies and the Hardy Weinberg formulae (known as the product rule formulae) and onto a conditional probability called a Match Probability that incorporates θ

More about this in a bit later

forensic science sa

Matching statistics – Match Probability

The Match Probability does not make the assumption of HWE and so is a more appropriate matching statistic to use for human populations.

It tells us the probability of seeing a profile a second time given that we have already seen that profile once

i.e. if the suspect is [A,A] (and we are assuming that the suspect is not the offender) then we want to know the probability of seeing [A,A] again (in the true offender). This is written as Pr(AA|AA) when the "|" means '*given that we've seen*'

forensic science sa

Matching statistics - Match Probability

For homozygote [A,A]

$$\Pr(AA \mid AA) = \frac{[2\theta + (1-\theta)p_A][3\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

For heterozygote [A,B]

$$\Pr(AB \mid AB) = \frac{2[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}$$

forensic science sa

Matching statistics - Match Probability

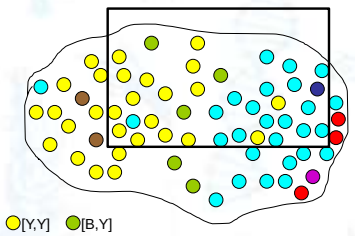
If you assume that there are no dependencies in the data and set θ = 0 then the match probability formulae cancel down to the HW product rule formulae:

A^2
2AB

However if we choose a value for θ of 0.01 (the approximate value that would correspond to a typical Caucasian population) then our frequency estimates from the earlier example would change:

forensic science sa

Population Databases



$p_Y = 0.473$

$p_B = 0.514$

$p_D = 0.013$

forensic science sa

Matching statistics - Match Probability

In our dataset we have $p_Y = 0.473$, $p_B = 0.514$ and $p_D = 0.013$

Assuming HWE
genotype frequency of [B,Y] = 2pq = 2(0.514)(0.473) = 0.486

Using theta
match probability of Pr(BY|BY) =

$$\frac{2[0.01 + (1-0.01)0.514][0.01 + (1-0.01)0.486]}{(1+0.01)(1+2(0.01))}$$

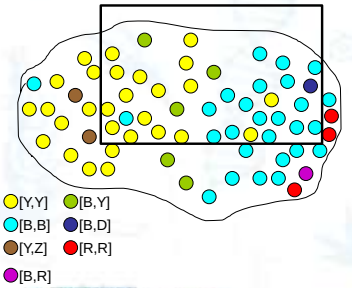
= 0.495

forensic science sa

Confidence Intervals

forensic science sa

Population databases - confidence intervals



forensic science sa

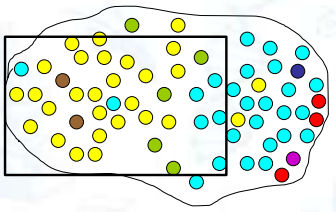
Population database frequencies:
p_Y=0.473
p_B= 0.514
p_D = 0.013

If we profiled everyone on the island we would see the true frequencies are:
p_Y=0.468
p_B= 0.452
p_D = 0.008
p_R = 0.056
p_Z = 0.016

Close but not the same

Population databases - confidence intervals

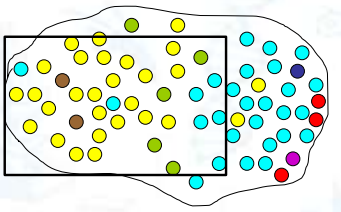
Image that the group of individuals we chose for the database were, by chance, different than those we actually chose



The allele frequencies we will differ from the allele frequencies calculated using the last database (which will both be different from the true allele frequencies in the population.

forensic science sa

Population databases - confidence intervals



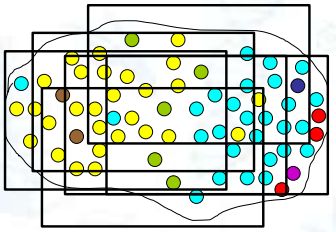
Take Allele 'Z' for example. We know that in the population $p_Z = 0.016$

In our first database we did not observe 'Z' so: $p_Z = 0$

In the database seen here: $p_Z = 0.031$

forensic science sa

Population databases - confidence intervals

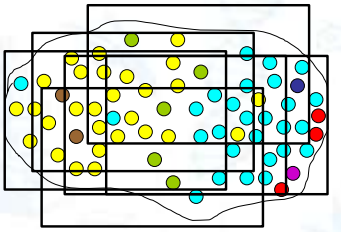


Now imagine databases being randomly chosen from the population many times

frequencies of alleles will vary but will be distributed around the true value in the whole population.

forensic science sa

Population databases - confidence intervals



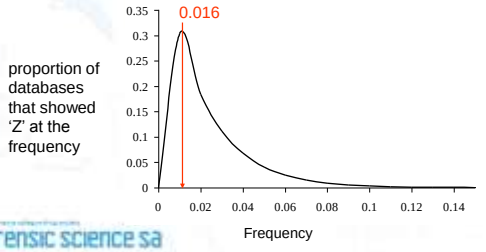
Using allele 'Z' as an example we would expect a distribution of allele frequencies for 'Z' that would centre around 0.016 (the true value in the population) but will deviate slightly either side.

This is called sampling variation

forensic science sa

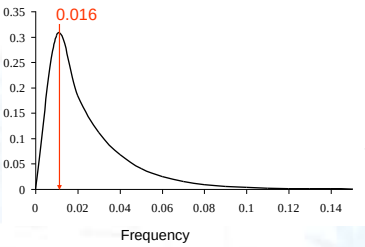
Population databases - confidence intervals

The distribution below shows the spread of allele frequencies that repeated sampling of databases has produced for allele 'Z'.



forensic science sa

Population databases - confidence intervals



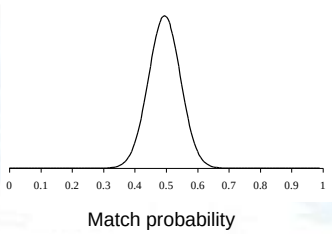
Our match probability calculation: Pr(BY|BY)=0.495

But now take into account that sampling variation means that allele frequencies will differ depending on how the database is chosen

forensic science sa

Population databases - confidence intervals

Using the distribution of allele frequencies (caused by sampling variation) we calculate a distribution for the match probability.



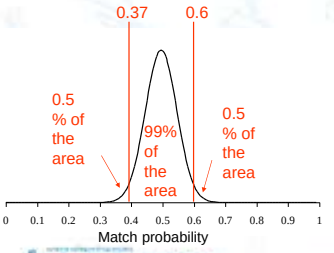
The Match probability point estimate: Pr(BY|BY)=0.495 which is the apex of the graph.

We choose a percentage of this curve to determine the confidence intervals that takes sampling variation into account

forensic science sa

Population databases - confidence intervals

For example a 99% confidence interval takes the inner 99% of the area under this curve:



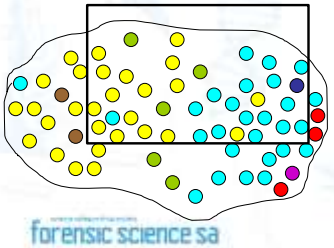
Typically report the MP i.e. 0.6, as that is conceding doubt to the defendant

NOTE: Using the allele frequencies of the whole population (P_B=0.452 & P_Y=0.468) we would obtain: Pr(BY|BY)=0.42 (which is captured within the 99% confidence intervals)

forensic science sa

Population databases - confidence intervals

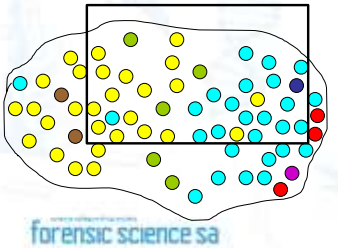
From the database shown we have the following:
frequency [B,Y] =0.486 (assuming HWE)
Pr(BY|BY)=0.495 (taking co-ancestry into account)
99% Pr(BY|BY)=0.37 to 0.6 (with sampling variation)



forensic science sa

Population databases - confidence intervals

But the true rarity of [B,Y] (from a simple count of the green dots) is 0.074
Which is very different from any of our calculated values.



no amount of mathematical adjustments are going to overcome a poorly constructed database (like the one seen here, which spans multiple populations).

This is where database validation and Fisher's exact test become VERY important

forensic science sa

Validating Databases

forensic science sa

Validating databases

There are tests that we subject our population database to prior to using them.

This is to ensure they are fit for forensic use.

There are many many many forensic papers that describe population databases for countries and groups all over the world, all of which will have had some analyses undertaken on them.

forensic science sa

Validating databases

As mentioned earlier, the simplest of these tests is to examine the database for outlying or unusual entries

For example in our database the single 'D' allele might be checked to make sure it is a legitimate allele.

There are other, more biologically focused, tests we perform to assess the data.

forensic science sa

Validating databases - Fisher's Exact Test

Used to test for departures from HWE either through dependencies in alleles within a locus (Hardy Weinberg disequilibrium) or between loci (linkage disequilibrium)

Determines the probability of obtaining the genotype frequencies given the observed allele frequencies

For those of you who like formulae:

$$P_C = \frac{n!2^H}{\prod_g n_g!}\prod_l \frac{\prod_j n_{lj}!}{(2n)!}$$

P_C = the conditional probability of the genotype counts
n_g = the genotype counts
n_{lj} = the allelic counts
n = the number of alleles
H = the total number of heterozygotic loci in the sample

forensic science sa

Validating databases - Fisher's Exact Test

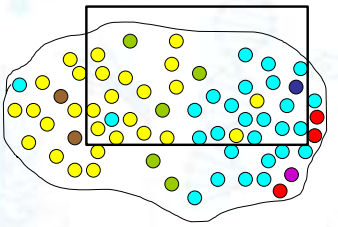
This probability is then compared to all permutations of the data and the probabilities of all values lower than the computed one are added together to give the *p*-value. Looking at [A] and [B] simplifies the formulae to:

$$\Pr(n_{AA}, n_{AB}, n_{BB} \mid n_A, n_B) = \frac{n! n_A n_B 2^{n\,AB}}{(2n)! n_{AA}! n_{AB}! n_{BB}!}$$

If the p-value is less than your cutoff (e.g.5% or 0.05) then your dataset shows some signs of departure from HWE

forensic science sa

Validating databases - Fisher's Exact Test



Think back to this example of a poorly constructed database

Observed genotypes

- 16 [Y,Y]
- 17 [B,B]
- 3 [B,Y]
- 1 [B,D]

forensic science sa

Validating databases - Fisher's Exact Test

In our dataset we have $p_Y = 0.473$, $p_B = 0.514$ and $p_D = 0.013$ so we would expect:

Genotype	Expected	Observed
[Y,Y]	8	16
[B,B]	10	17
[B,Y]	18	3
[B,D]	0.5	1

Without carrying out the complete Fisher's test, allele frequencies are not estimating genotype frequencies very accurately.

But could this just be by chance ?

forensic science sa

Multi-testing problem

Some comparisons may show significant departures from equilibrium

This does not necessarily mean that these dependencies exist

For a database in perfect equilibrium we would expect that the p-values would be evenly spread over the range 0 to 1 (as the exact tests are based on random reshufflings of the data)

forensic science sa

Multi-testing problem

For a p-value cut-off of 0.05 we would expect that if no dependencies existed then 5% of comparisons would show a significant departure for equilibrium by chance alone.

The more tests we do the more significant p-values we will get and this is the multi-testing problem.

With 15 Identifiler loci there end up being 120 comparisons being done so we would expect approximately 6 to have p-values < 0.05

Needs to be taken into account when assessing HWE/LE departures

forensic science sa

Multi-testing problem

Two ways of dealing with multi-testing problem:
- A graphical representation
- A truncated product method

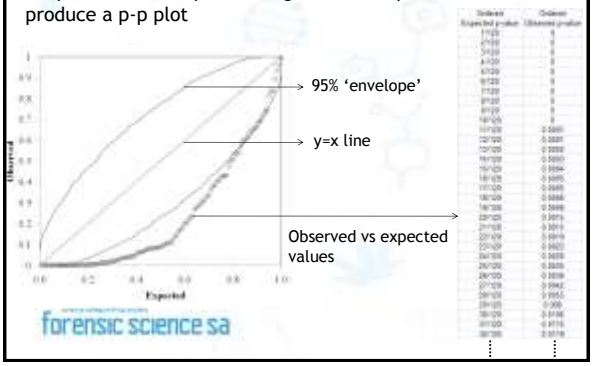
The graphical method is appealing and visually easy to understand. It is based on the idea that for multiple Fisher's exact tests the p-values (for a database with no dependencies) should be evenly spread over the range 0 to 1

This means that if all the p-values were ordered in ascending order then they should fall on a line with the equation y=x

forensic science sa

Multi-testing problem

Graph the ordered p-values against the expected values to produce a p-p plot



forensic science sa

Multi-testing problem

Databases that do not contain dependencies will fall within the 95% envelope, databases that do contain dependencies will fall outside

Below are two pan-Australian databases for Aboriginal and Caucasian database

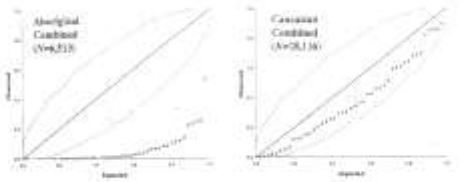


Figure 2. p-p plots for the combined Aboriginal and Caucasian datasets.

Multi-testing problem

The second method for assessing departures from HWE and LE is the truncated product method.

This method considered all the p-values together to see whether there is evidence that the results from the multiple tests, as a whole, show evidence for significance

forensic science sa

Multi-testing problem

The truncated product method states that the sum of *-2ln(p-values)* from '*t*' independent tests should have a chi-squared distribution with *2t* degrees of freedom

If you are interested in reading about why this is then read the paper:

Zaykin, D., Zhivotovsky, L. A. and Weir, B. S. (2002) Truncated product method for combining p-values. *Genetics and Epidemiology* 22: 170-185.

forensic science sa

Multi-testing problem

Locus	*p*-value	-2ln(*p*) HWE
CSF	0.53	1.26
D12	0.10	4.68
D13	0.26	2.72
D16	0.51	1.37
D18	0.25	2.74
⋮		
FGA	**0.01**	9.09
TPOX	0.35	2.13
vWA	1.00	0.00
Sum[-2ln(*p*)]		50.66
***p*-value**		0.12

This method is most easily carried out in Excel

FGA is showing a significant p-value

But overall data has no significant disequilibrium

=Chidist(50.66, 40)

Sum[-2ln(*p*)] — *2t* d.o.f.

forensic science sa

Delving further into Fisher

You can delve a little further into dependencies by looking at the observed numbers of genotypes against the expected number at each locus (based on allele frequencies)

Counts of alleles are shown in the right hand column and bottom row.

Combined with the total these can be used to generate expected genotype frequencies

forensic science sa

Delving further into Fisher

If in equilibrium these values should adhere to a chi-squared distribution with 1 degree of freedom, so for a 95% confidence interval the 1 d.o.f. chi-squared critical value is 3.84

This means that if we calculate the value:

$$\frac{(Observed - Expected)^2}{Expected}$$

Any values > 3.84 are a significant departure

forensic science sa

Delving further into Fisher

Doing this can give you some further information that the p-values alone would.

It can tell you that disequilibrium is being caused by a few rare genotypes

Also if significant values are on the diagonal (indicating an excess of homozygotes) then this can indicate the database has substructure

This excess homozygosity from grouped populations is known as the Wahlund effect

forensic science sa

A quick word about Linkage

forensic science sa

Linkage

Gregor Johann Mendel (1822 – 1884)
Austrian Augustinian monk and scientist

Studied inheritance of certain traits in pea plants

The law of segregation – Each individual has two 'factors' controlling a given characteristic, one being a copy of a corresponding factor in the father of the individual and one being a copy of the corresponding factor in the mother of the individual. Further, a copy of randomly selected one of the two factors is copied to each child, independently for different children and independently of the factor contributed by the spouse.

forensic science sa

Linkage

Thomas Hunt Morgan (1866 – 1945)
American evolutionary biologist, geneticist and embryologist

Found that Mendel's laws did not always hold true.

He found that some characteristics in Drosophila did not randomly assort as expected

Morgan had shown that some genes were linked

forensic science sa

Linkage

Crossing over is when genetic material is shared between homologous chromosomes at points of recombination

When two points on a chromosome are separated by a large distance then the chance that recombination will occur between them is high

i.e. If these two distant points were originally in phase (i.e. originally on the same chromosome) then they are equally likely to be end up on the same chromosome as different chromosomes at the end of meiosis phase I

Linkage

When the sites on the chromosome are not distant from each other then it is less likely that a recombination will occur between them during meiosis

This means that the two loci will be inherited together as one unit

If two loci that are initially in phase are both inherited more often than is expected by random assortment then they are considered to be linked

Linkage

Fisher's Exact does not actually detect linked loci purely because they are linked

It detected linkage disequilibrium in a population

A population that is many generations removed from an 'evolutionary event' may have partially linked loci that do not show any signs of linkage disequilibrium.

Linkage disequilibrium is often detected in partially linked loci usually because they take a longer time to re-equilibrate.

forensic science sa

Linkage

When a population undergoes an evolutionary event, all loci (linked or not) will be in linkage disequilibrium.

Examples of an evolutionary event include:

- **Bottleneck**: When a population is reduced to less than half its size.

- **Founder Effect**: A population that has arisen from a small group of 'founders'

forensic science sa

Linkage

- **Gene flow**: The exchange of genes between populations (this does not require the physical movement of individuals, only their genes)

- **Selective sweep**: When a gene spread throughout a population due to some positive selection.

Often native populations will have undergone these events and so be in linkage disequilibrium

forensic science sa

Linkage

Linked loci will take longer to reach a level of equilibrium than unlinked loci.

The amount of linkage disequilibrium (D) in a population, '*n*' generations after an evolutionary event can be determined by:

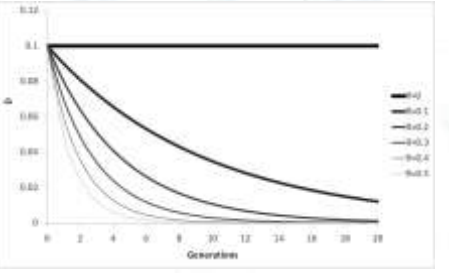
$$D_n = (1-R)^n D_0$$

D_0 = the level of linkage disequilibrium caused by the evolutionary event.

forensic science sa

Linkage

The graph below shows the re-equilibration with an initial level of linkage disequilibrium, D = 0.1 with various levels of Recombination (*R)*.



Linkage

Completely linked loci (i.e. those where recombination never occurs) will never be able to recover from the evolutionary event

Completely unlinked loci half the amount of linkage disequilibrium each generation – But will still show some linkage disequilibrium

This is the reason that linkage equilibrium has the addition required assumption (beyond HWE assumptions) that an infinite number of generations has elapsed since any disturbing force.

forensic science sa