Strategies for multiple testing

As we saw earlier, in genetic studies such as genome-wide association studies and in gene expression studies, there are often thousands of hypothesis tests performed on the same data set. This can lead to a large number of false positives unless using some method for correcting for *p*-values.

The most straightforward method to correct for *p*-values is to use a Bonferroni adjustment, where for each null hypothesis, you reject the null only if $p < \alpha/k$ where k is the number of hypothesis tests and typically $\alpha = 0.05$. The Bonferroni method makes the probability of rejecting at least one null hypothesis less than or equal to 0.05, and also makes the expected number of false positives equal to 0.05.

The Bonferroni method tends to be very conservative in that the type I error rate is actually less than 0.05 and might make it difficult to detect cases where the null hypothesis should be rejected. Bonferroni correction therefore can decrease power for detecting associations.

Multiple testing

Generally, we can count the decisions of hypothesis tests using the following table

	Decision		
Reality	Don't reject H_0	Reject H_0	total
H_0 true	U	V (False positive)	m_0
H_0 false	W (False negative)	S	m_1
total	$m - R$	R	m

Multiple testing

From the table, we can define some other quantities when the m null hypotheses are true for each hypothesis test:

1. Per-familty error rate (PFER), $E(V)$
2. Per-comparison error rate (PCER) $E(V)/m$
3. Family-wise error rate (FWER) $P(V \leq 1)$
4. False Discovery Rate (FDR), $E(Q)$, $Q = V/R$ for $R > 0$; otherwise $Q = 0$.

Bonferroni controls both the per-comparison error rate and family-wise error rate. The FDR measures the expected proportion of rejections of the null that are false.

Multiple testing

The **sensitivity** (true positive rate) is the probability of rejecting H_0 when H_0 is false, and the **specificity** is the probability of not rejecting H_0 when H_0 is true. In otherwords, sensitivity and specificty are probabilities of making correct decisions when the null is false and true, respectively. It's difficult to remember these. You can associate them with alphabetical order (sensitivity $<$ specificity and false $<$ true).

Multiple testing

To get a little philosophical here, it isn't clear what error rates we should try to control. Suppose I have a research project or paper with one gene expression data set in humans, and for comparison, a gene expression data set in chimpanzees. Suppose each has 20,000 hypothesis tests. Should I test for signficance separately in each data set, using $m = 20,000$? Should I use $m = 40,000$? Suppose I have 20,000 genes and 10,000 for chimpanzees. Should I use different values of m for the data sets?

Suppose the project original had gorilla data also, and we decide to exclude it in the final paper. Should this be adjusted for also?

Suppose you write one paper this year and your advisor writes three papers. Should you control your yearly type 1 error rate? Your paper type 1 error rate? How about your career-wise type 1 error rate?

Multiple testing

Suppose I'm consulting and my client comes in the afternoon and says, "I'm only testing 1 hypothesis, so I don't need to do multiple testing adjustments". I respond, "Well I tested 20 hypotheses before lunch, so I'd like to use Bonferroni on your project....". (Ok, I would never do this....)

Multiple testing

Since for Bonferroni, the FWER is less than α, we might try to use α' such that the FWER is exactly α. The probability of all decisions being correct is $(1-\alpha')^m$ which we set equal to $1-\alpha$.

$$(1-\alpha')^m = 1-\alpha \Rightarrow (1-\alpha') = (1-\alpha)^{1/m} \Rightarrow \alpha' = 1-(1-\alpha)^{1/m}$$

This controls the FWER. How does it fare with the PCER (expected proportion of false positives)? The expected proportion of false positives?

Although the probability that all decisions are correct is $1-\alpha$, and therefore the probability of at least one wrong decision is $1-(1-\alpha) = \alpha$, the expected number of wrong decisions is $m\alpha'$:

$$m\alpha' = 1-(1-\alpha)^{1/m}$$

For $m = 100$, this amounts to 0.05128014. This also means, that we reject H_0 when $p < .0005128014$, which is slightly larger than the Bonferroni cutoff of .0005.

Multiple testing

Thus, the PCER is not quite controlled at level α, but it is quite close. For larger values of m, things also do not get much worse, for example, for 100,000 tests, the PCER is 0.05129328. Overall, this method is not very different from Bonferroni, and is just slightly less conservative.

If $m = 2$, then this method results in using $\alpha' = 0.02532$ instead of $\alpha' = 0.25$ using Bonferroni.

Multiple testing

The idea of the False Discovery Rate is to control the expected number of incorrect rejections of the null hypothesis given that you reject the null hypothesis. The idea is that you have a large number of hypotheses, say 10,000, and 100 of them result in rejecting the null hypothesis, then you want a small number of rejections to be false positives. If you set the FDR to 0.10, then you expect 10 out of the 100 to false positives, and 90 out of 100 to be true positives.

In many testing situations, most or many positive test results turn out to be false positive test results, and FDR is a way to control the proportion of positive test results that are false positives.

Bayes example

As a classic example of the problem of false positives, suppose that the proportion of people in the population who are HIV positive is 0.01. Suppose the test correctly identifies 99% of people who are HIV positive, and correctly identifies 98% of HIV-negative people as not having HIV. A test is given to a random individual in this population, and the person tests positive. What is the probability that the person has HIV?

Multiple testing

In this case the probability that they have HIV given that they tested positive is 1/3. So we could say that the FDR for this testing procedure is 2/3, which is quite high. In genetic testing, screening thousands of genes, this FDR rate would mean that you might get 20 false positive associations for every 10 true positive associations. This might or might not be acceptable to you – it might be a reasonable way of determining candidate genes that should be investigated in further studies, for example, where you care more about minimizing false negatives (you don't want to miss genuine associations) than minimizing false positives.

FDR was developed in the 1990s and has gained wide popularity in genetic testing situations.

Multiple testing

The FDR approach is to look at the p-values themselves and order them. If we have m p-values, we can order them by

$$p_{(1)}, p_{(2)}, \ldots, p_{(m)}$$

where $p_{(1)}$ is the smallest p-value. Ties are retained, so p-values don't have to be distinct, but the list is non-decreasing.
If the null hypothesis is true, then what should this list look like? What is the distribution of p-values under the null hypothesis?

Multiple testing

From the idea of order statistics, we can use the distribution of $p_{(1)}$ to get its expected value and compare to the actual value obtained from the experiment. Under the null hypothesis, p-values should be uniformly distributed (from continuous distributions — for discrete distributions, there will be a finite number of possible p-values which still should approximate a uniform distribution. Why is this?

The p-value is a function of the data and is therefore a random variable, just like $\max X, \overline{X}$, etc. To understand it's distribution, we can use the cdf method.

Multiple testing

$$P(p - value \leq \alpha) = \alpha$$

If the test has type I error of α, then the *p*-value is less than α exactly $100 \times \alpha\%$ of the time, so the probability that it is less than α is precisely α. This argument holds for any choice of α. Therefore the cdf of the *p*-value is the cdf of the unifrom distribution, so *p*-values must have a uniform(0,1) distribution. For a sample of size m uniform random variables, the expected value of $X_{(i)}$ is $i/(m+1)$. Plotting the observed ranked *p*-values against their expected values should show a straight line. If there are values below this straight line, then those *p*-values are smaller than expectation.

Multiple testing

The Benjamini-Hochberg procedure is to reject the first k hypotheses $j = 1, \ldots, k$ if k is the **largest** value for which $p_{(k)} \leq \alpha k/m$. Ignoring the difference between m and $m+1$ (which is minor for large m), this procedure is essentially reject hypothesis for which the ordered p-value is much less than expected. For example

```
      pvals   threshold
 [1,] 0.010     0.005         0
 [2,] 0.013     0.010         0
 [3,] 0.014     0.015         1
 [4,] 0.190     0.020         0
 [5,] 0.350     0.025         0
 [6,] 0.500     0.030         0
 [7,] 0.630     0.035         0
 [8,] 0.670     0.040         0
 [9,] 0.750     0.045         0
[10,] 0.810     0.050         0
```

Here we reject the first three null hypothesis, even though the first two p-values are larger than $\alpha j/m$.

Multiple testing

Here is an example where 50% of the tests are generated from a $N(0.25, 1)$ and 50% are generated from a $N(0, 1)$ and we test $H_0 : \mu = 0$ for each case.

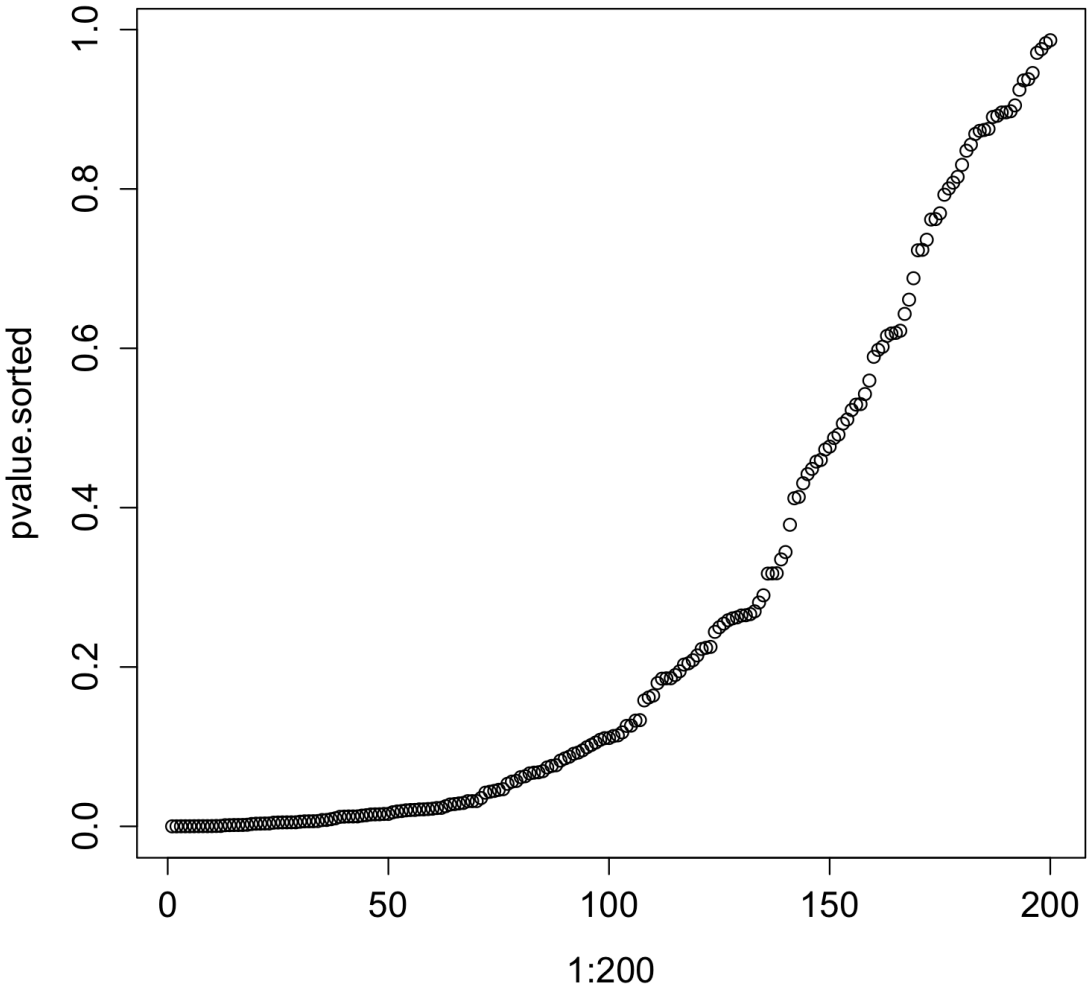
```
> pvalue <- NULL
> for(i in 1:100) {
+ x <- rnorm(100,mean=.25)
+ y <- rnorm(100,mean=0)
+ pvalue <- c(pvalue,t.test(x)$p.value)
+ pvalue <- c(pvalue,t.test(y)$p.value)
+ }
> pvalue.sorted <- sort(pvalue,decreasing=F)
> plot(1:200,pvalue.sorted,cex.lab=1.3,cex.axis=1.3)
> k <- 1:200
> as.numeric(pvalue.sorted <= .05*k/200)
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Multiple testing

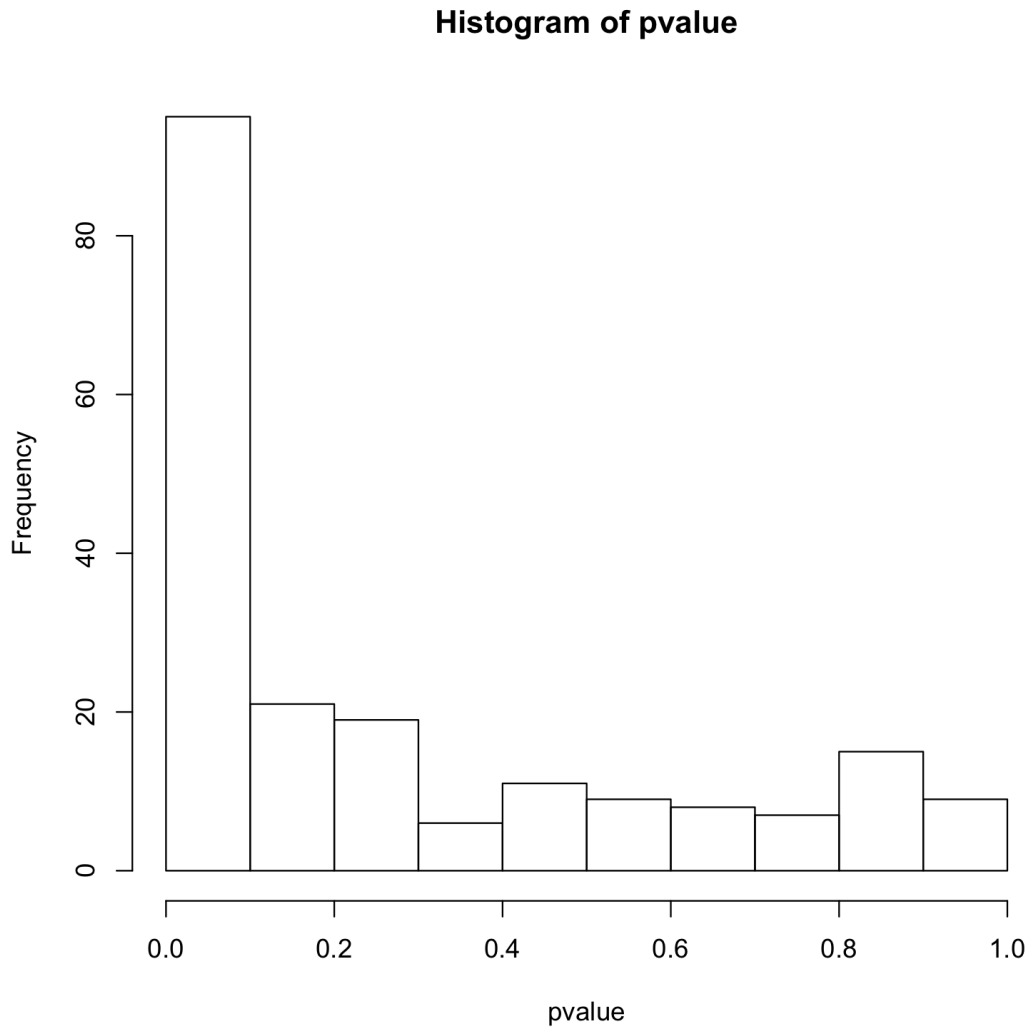
In the previous example, we reject the 37 hypotheses that had the smallest p-values. The larget pvalue that lead to a rejection was 0.00897, and was the only false positive. Using Bonferroni, we would have only rejected 8 hypotheses, and the largest pvalue leading to a rejection would be .00025.

```
> as.numeric(pvalue.sorted <= .05/200) # Significant by Bonferroni
  [1] 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> pvalue.sorted[37] #largest pvalue for Benjamini-Hochberg
[1] 0.008973795
> pvalue.sorted[8] # largest pvalue for Bonferroni
[1] 0.0002489277
> pvalue.y <- pvalue[2*(1:100)] # extract pvalues when H0 is true
> sort(pvalue.y)
  [1] 0.00897 0.01399 0.01561 0.01921 0.02746 0.02797
```

Multiple testing: pvalues far from uniform



Multiple testing: pvalues far from uniform



Multiple testing

Sometimes instead of testing at level α/m when using Bonferroni or other methods, another approach is to use adjusted p-values, which might be easier to interpret. The idea is to multiply the p-value by m, and then test at the $\alpha = .05$ level (or other α level). If the result is above 1, you can call the adjusted p-value 1. The Bonferroni adjusted p-value is therefore

$$p^* = \min(mp, 1)$$

Multiple testing

Adjusted p-values can also be used for other methods. For the Bejamini-Hochberg FDR approach, you can use

$$p^*_{(i)} = \min_{k \in \{i,\ldots,m\}} \{\min((m-k+1)p_{(k)}, 1)\}$$

A more intuitive approach is to multiply the pvalue by m/k, keeping in mind that you only reject $p_{(j)}$ for the smallest j until some test doesn't lead to a rejection. Multiplying the p-values by m/k might make the sequence

$$p_{(1)} \cdot 1/m, p_{(2)} \cdot 2/m, \ldots, p_{(m)} \cdot m/m$$

nonmonotonic, so there might be values in the sequence less than α for which you don't reject the null hypothesis.

Multiple testing

The Benjamini-Hochberg approach to controlling FDR is part of a more general class of procedures called step-down procedures where the required α is adjusted for each test and made larger for each test, with increments in successive αs decreasing at each step. There are a number of variations on the Benjamini-Hochberg approach, including some that account for tests being not independent.

Multistage designs

Another strategy is to use multistage designs. The idea here is to partition the data set, for example, into two halves. For case-control data, you might separately partition the cases into two halves and then partition the controls into two halves (how this is done might depend on whether the controls are matched to the cases based on age, sex, etc. For genetic association studies, controls are simply people who don't have the disease/condition, and are not more specifically matched).

This can be done randomly. The first half of the data is used to find candidate genes. Using typical procedures, such as Bonferroni or Benjamini-Hochberg. Since only candidate genes are desired, it might be acceptable to have a high false positive rate here, so you might use $\alpha = 0.10$ (which is then adjusted for multiple comparisons) in order to get more candidate genes.

Multistage designs

As a result of this procedure, if you start with 20,000 genes, you might end up with, say, 500 candidate genes. Then the 500 genes are tested in the remaining half of the data. Since the other half of the data is independent of the first set, you're now testing only 500 times, so you could use $\alpha' = .05/500 = 0.0001$ (using Bonferroni), which is much less stringent than usual cutoffs for GWAs. Or again, you could use a step-down method for the replication data set as well.

A criticism of the multistage design approach is that it is less powerful than using all of the data at once by using only half of the data. However, it is essentially a cross-validation approach, so that using the first half of the data to generate a candidate list, and the second half to see how many of these are confirmed to be significant, you could also reverse the roles of the two halves of the data set and check to see if the same genes show up in both analyses. You will generally have more confidence in genes that show up under different partitionings of the data.

Multistage designs

A similar approach that has been applied in Family-Based Association Studies is to first estimate the power for rejecting the null hypothesis of no association for each gene. The power depends on an estimated effect size (so the proportion with the disease for the two different SNPs), the number of families involved, the genotypes of the parents (for example two heterozygous parents might be more informative than one heterozygous parent), and the allele frequencies.

The idea is then to only test genes that have high power. This reduces the number of hypothesis tests done and makes the Bonferroni (or other) correction less severe. The technique has used a type of weighted Bonferroni correction, where the adjusted p-value for test i is $m_i p_i$ (where the pvalues are not ordered), and m_i is proportional to the estimated power in such a way that $\sum_i m_i = m$. Thus, the average α level used is α/m, but different tests are given different α levels depending on the power of the test. This way genes with less estimated power can have a higher chance of getting a significant association, and the overall α level is still controlled.

Multistage designs

Unfortunately, power is higher for moderate allele frequencies (e.g., if the two SNPs have frequencies .4 and .6 versus .1 and .9). This means that if a rare SNP is associated with a disase then there is likely to be less power to detect an association. Filtering genes before testing based on power might mean that genes with rare variants don't get tested at all.

The idea of only looking at genes that have high power has been given the analogy of someone looking for their keys in the street at night. The person is looking only near a lamppost. Someone asks, "Why are you looking near the lamppost? Is that where you lost your keys?" Answer: "I don't know if I lost my keys there, but that it is the only place light enough to see them."

Multistage designs

Although an association is more convincing if it can replicated within a data set using cross-validation, this is not as good as replication with a completely new study using a different population, different researchers etc. Often genetic associations are not replicated in new data sets, in which case there might have been other causes for the assocation, such as population structure, ascertainment bias (where individuals that are selected in the study not at random but for other reasons associated with the disease).

Ascertainment bias essentially means that not all members of a population are equally likely to be sampled. This is generally the same idea as sampling bias in statistics. There can be many reasons for sampling bias. In genetic studies, SNP panels represent a subset of possible SNPs in the human genome. For example, the company 23 and Me, uses a SNP panel of 700,000 SNPs, although there are a few million SNPs in the human genome. Using a smaller SNP panel is cheaper than using whole genome sequences.

However, the SNPs chosen in the SNP panel are based on SNPs that have been found in previous studies. A major source of data for SNP panels has been the HapMap project, which included 30 Yoruba trios from Nigeria, 30 trios from Utah of Northern and Western European acestry, 45 "unrelated" Han Chinese individuals from Beijing, and 44 "unrelated" Japenese from Tokyo. Variation from these populations is therefore represented in many SNP panels, but individuals from other populations might go undetected since it might include SNPs not present in those samples.

This also means that SNPs that were relatively rare in those populations and more common in other populations would be less likely to be represented in the SNP panel.

Assignment of population ancestry

The idea for the companies like ancestry.com and 23andMe is to assign proportions of ancestry for individuals. Individual genes or SNPs usually carry very little information about the ancestry of an individual. But you can make a best guess for a person's ancestry using likelihood. Here is a rough idea

Location	SNP1=0	SNP2=0	SNP3=0	SNP4=0	SNP5=0
1	.25	.30	.20	.45	.35
2	.25	.35	.25	.25	.45
3	.20	.40	.30	.30	.40

Here each individual has either a 0 or 1 at each SNP, and the probabilities of a SNP being a 0 or 1 are given for each location. If someone's SNPs are say 0, 1, 1, 0, 0 for these data, it is very difficult to tell just by looking which location they belong to. However, we can do a likelihood calculation.

Assignment of population ancestry

What is the best guess for the location that the person is from? We can get likelilood values and see what the maximum likelihood location is:

$$L(1) = (0.25)(0.70)(0.80)(0.45)(0.35) = 0.02205$$

$$L(2) = (0.25)(0.65)(0.75)(0.25)(0.45) = 0.01371094$$

$$L(3) = (0.20)(0.60)(0.70)(0.30)(0.40) = 0.01008$$

Therefore population 1 has a considerably higher likelihood. The magnitude of the likelihoods doesn't matter, only their relative values, and population 1 has nearly twice the likelihood as the other locations. So even though all of these population are fairly similar in terms of their allele frequencies (for all of them allele 0 is more rare than allele 1), there is some signal in the data that can give a reasonable guess. In a data set with 700,000 SNPs you can be more certain about assignment of ancestry. However, ancestry often changes at different points in the genome if your ancestors are from different locations.

Assignment of population ancestry

If you want probabilities, you can use Bayes Theorem if you have prior information about the probability that a person's ancestry is from different locations. In the absence of information, you might assume that all of the locations are equally likeliy. In this case

$$P(1|0,1,1,0,0) = \frac{P(0,1,1,0,0|1)P(1)}{P(0,1,1,0,0)|1)P(1) + P(0,1,1,0,0)|2)P(2) + P(0,1,1,0,0)|1)P(3)} = 0.4810$$

$$P(2|0,1,1,0,0) = \frac{P(0,1,1,0,0|2)P(2)}{P(0,1,1,0,0)|1)P(1) + P(0,1,1,0,0)|2)P(2) + P(0,1,1,0,0)|1)P(3)} = 0.4187$$

$$P(3|0,1,1,0,0) = \frac{P(0,1,1,0,0|3)P(3)}{P(0,1,1,0,0)|1)P(1) + P(0,1,1,0,0)|2)P(2) + P(0,1,1,0,0)|1)P(3)} = 0.2198$$

However, assignment of population ancestry might be more accurate if ancestry is close to populations well represented in the SNP panel due to ascertainment bias.

Multiple testing

ABO and Rh blood type distribution by country (population averages)

Country	Population[1]	O+	A+	B+	AB+	O−	A−	B−	AB−
Australia[2]	21,262,641	40.0%	31.0%	8.0%	2.0%	9.0%	7.0%	2.0%	1.0%
Austria[3]	8,210,281	30.0%	33.0%	12.0%	6.0%	7.0%	8.0%	3.0%	1.0%
Belgium[4]	10,414,336	38.0%	34.0%	8.5%	4.1%	7.0%	6.0%	1.5%	0.8%
Brazil[5]	198,739,269	36.0%	34.0%	8.0%	2.5%	9.0%	8.0%	2.0%	0.5%
Canada[6]	33,487,208	39.0%	36.0%	7.6%	2.5%	7.0%	6.0%	1.4%	0.5%
China[7]	1,339,724,852	47.7%	27.8%	18.9%	5.0%	0.3%	0.2%	0.1%	0.03%
Czech Republic[8]	10,532,770	27.0%	36.0%	15.0%	7.0%	5.0%	6.0%	3.0%	1.0%
Denmark[9]	5,500,510	35.0%	37.0%	8.0%	4.0%	6.0%	7.0%	2.0%	1.0%
Estonia[10]	1,315,819	29.5%	30.8%	20.7%	6.3%	4.3%	4.5%	3.0%	0.9%
Finland[11]	5,250,275	27.0%	38.0%	15.0%	7.0%	4.0%	6.0%	2.0%	1.0%
France[12]	62,150,775	36.0%	37.0%	9.0%	3.0%	6.0%	7.0%	1.0%	1.0%
Germany	82,329,758	35.0%	37.0%	9.0%	4.0%	6.0%	6.0%	2.0%	1.0%
Iceland[13]	306,694	47.6%	26.4%	9.3%	1.6%	8.4%	4.6%	1.7%	0.4%
Ireland[14]	4,203,200	47.0%	26.0%	9.0%	2.0%	8.0%	5.0%	2.0%	1.0%
Israel[15]	7,233,701	32.0%	34.0%	17.0%	7.0%	3.0%	4.0%	2.0%	1.0%
Japan[16]	127,368,088	29.9%	39.8%	19.9%	9.9%	0.15%	0.2%	0.1%	0.05%

Ascertainment bias

Since the Hap Map, which had several phases of data sets published as technology improved (2005, 2007, 2009), there have been more data sets from more populations world wide, so this should lessen ascertainment bias regarding SNPs.

Ascertainment bias

Another example of ascertainment bias occurs when people die early from an illness. In this case, they are not sampled, and this can bias your estimate for how long individuals survive.

A striking example is from this 2014 paper in *American Journal of Human Genetics* criticising a 2013 article published in *PLOS ONE*. The PLOS paper described families with Creutzfeldt-Jakob disease (CJD) caused by a mutation with the memorable name of *PRNP* E200K. The study found that the children with the disease died 12 years earlier (on average) than a parent who also had the disease.

Ascertainment bias

This phenomenon occurs for some diseases, and is called *anticipation*. An example is Huntington's Disease, in which a 3-letter sequence (CAG) is repeated in a certain location on the short arm of chromsome 4. The number of repeats can increase over generations. Huntington's disease can occur if the number of repeats is 36 or larger, and apparently always occurs if the number of repeats is 40 or more. Since the number of repeats tends to increase over generations, children are more likely to be affected than their parents, and hence can have shorter lifetimes.

Multiple testing:

ARTICLE

Ascertainment Bias Causes False Signal of Anticipation in Genetic Prion Disease

Eric Vallabh Minikel,[1,2,3,*] Inga Zerr,[4,5] Steven J. Collins,[6] Claudia Ponto,[4] Alison Boyd,[6] Genevieve Klug,[6] André Karch,[4] Joanna Kenny,[7] John Collinge,[7] Leonel T. Takada,[8] Sven Forner,[8] Jamie C. Fong,[8] Simon Mead,[7,9] and Michael D. Geschwind[8,9]

PLOS ONE | Subject Areas | For Authors | About Us

OPEN ACCESS PEER-REVIEWED

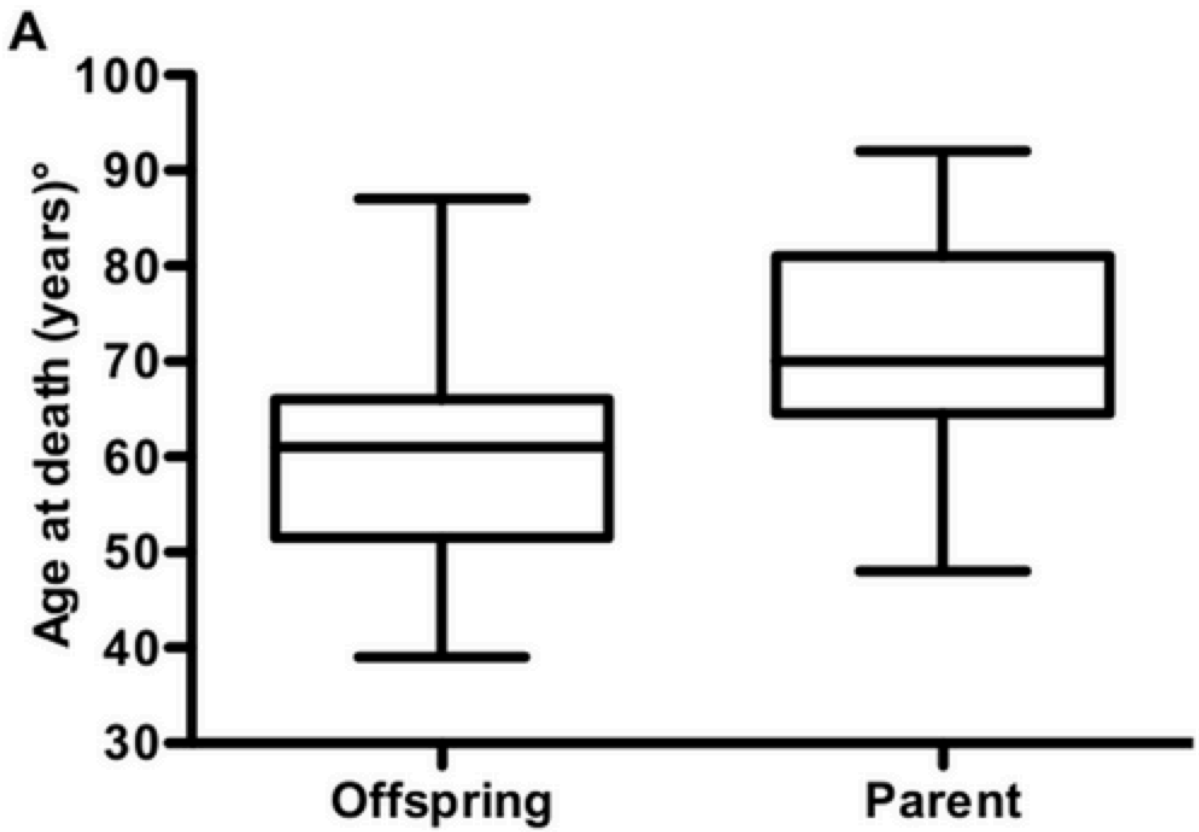
RESEARCH ARTICLE

Age at Death of Creutzfeldt-Jakob Disease in Subsequent Family Generation Carrying the E200K Mutation of the Prion Protein Gene

Maurizio Pocchiari, Anna Poleggi, Maria Puopolo, Marco D'Alessandro, Dorina Tiple, Anna Ladogana

Published: April 2, 2013 • DOI: 10.1371/journal.pone.0060376

Multiple testing:



Ascertainment bias

The study was looking at parent-child pairs where both had been diagnosed with the disease and looking at the age of death for both individuals. There might have been right-censoring. There was also left-truncation because the disease was not studied before 1989. As a result, in cases where the parent died earlier, they were not included in the study. Another problem is that parents might have died of another problem (cancer, heart failure) before they had a chance to develop the disease. In this case, parents with the disease are less likely to be included in the study if their age of onset is later; or to put it another way, they are more likely to be included in the study if their age of onset is earlier.

Multiple testing: