Genotype interpretation, checking HWE

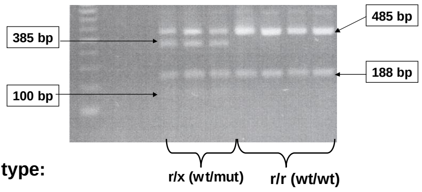
8.11.2006

BfMS, Genetics module

PO

GENOTYPE INTERPRETATION

Gel photo:



Genotype:

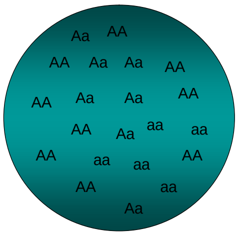
GENE: ACTN3
Amplified region:
RE digestions: Dde1

Concepts

- **Population** = collection of people, or organisms of a particular species, living in a given geographic area
- **Gene pool** = The gene pool is the complete set of alleles (in a locus) found in every living member of that species or population.
- **Allele frequency** = frequency of an allele in a genetic locus in a given population
- **Genotype frequency**= frequency of a genotype in a genetic locus in a given population

Counting genotype frequencies

- Locus where we have alleles *A* and *a*:



$$f(AA) = \frac{n(AA)}{n(AA)+n(Aa)+n(aa)}$$

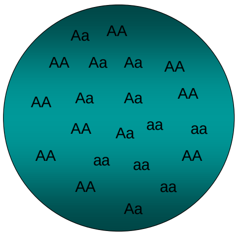
$$f(Aa) = \frac{n(Aa)}{n(AA)+n(Aa)+n(aa)}$$

$$f(aa) = \frac{n(aa)}{n(AA)+n(Aa)+n(aa)}$$

$$f(Aa)+f(Aa)+f(aa)=1$$

Calculating allele frequencies

- Every individual has two alleles in each autosomal locus à number of alleles in the population is 2**x** number of individuals in the population.



$$f(A) = \frac{2[n(AA)]+n(Aa)}{2[n(AA)+n(Aa)+n(aa)]}$$

$$f(a) = \frac{2[n(aa)]+n(Aa)}{2[n(AA)+n(Aa)+n(aa)]}$$

$$f(A)+f(a)=1$$

Statistical analyses: Genotype data

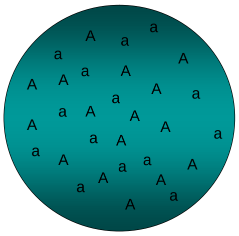
- **Standard checks:**
- Hardy-Weinberg equilibrium – are the alleles distributed in genotypes randomly, as they should, or are some genotypes over-represented at some loci? → Indicates a problem with genotyping, or in the sampling of individuals
- In case you've got families etc. (related individuals) Mendelian inheritance of each and every marker! E.g. with program "pedcheck"

Hardy-Weinberg equilibrium

- Basic principle: at equilibrium, the frequency of each genotype is defined by allele frequencies (here, for allele A, p, and for allele B, q)
  - – AA p^2
  - – AB 2pq
  - – BB q^2

Hardy-Weinberg Equilibrium

- Consider a population, where there are two alleles in a locus:



The probability to pick an allele *A* from the gene pool = allele frequency of *A* in the population f(*A*)

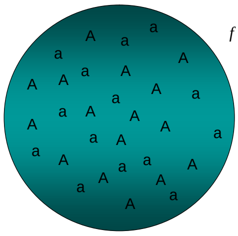
$$f(A) = p$$

$$f(a) = q$$

$$p + q = 1$$

Hardy-Weinberg Equilibrium

- The formulation of genotypes from the gene pool of the population can be considered as a random process; thus basic laws of probability apply



$$f(AA) = p \times p = p^2$$

$$f(Aa) = p \times q + q \times p = 2pq$$

$$f(aa) = q \times q = q^2$$

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

Hardy-Weinberg Equilibrium

- In H-W equilibrium, the genotype frequencies of a population can be derived from allele frequencies. The equilibrium holds from generation to generation*

$$\begin{tikzcd} & p+q=1 & \\ f(A) \arrow[ur] & & f(a) \arrow[ul] \end{tikzcd}$$

$$\begin{tikzcd} & p^2+2pq+q^2=1 & \\ f(AA) \arrow[ur] & F(Aa) \arrow[u] & f(aa) \arrow[ul] \end{tikzcd}$$

HWE applies when there is

- Random mating
- No mutations
- No natural selection
- No migration
- Big population size

- In real world, no population strictly follows these conditions, though
- …vast majority of natural populations can be found to be in or very near to HWE
- HWE is utilized in various research purposes: from validation of genetic markers (to be used in genotyping) to ecological and evolutionary research questions

How to test for HWE?

- Deviations from HWE might tell us important facts of the population, such as
  - Genotyping problems
  - Population substructure
  - Natural selection functioning on the locus
  - We are at or very near to a disease gene which has strong effect on disease susceptibility
- Thus, very important to TEST for HWE in population samples!
- The simplest test type is χ^2 –contingency test
- Here, we are not making tests with pen-and-paper, but it's still useful to know/remember that…

χ^2-contingency test

- A sample of individuals from a population has been genotyped. Is the population in HWE concerning the locus in question?

	A1A1	A1A2	A2A2	Sum
Observed	31	89	122	242

- **How should the genotype frequency look like if the data is in HWE?**
  -> in HWE, f(A1A1)=p^2, f(A1A2)=2pq and f(A2A2)=q^2
  -> thus we'll need estimates for allele frequencies
- We don't have any background info on allele frequencies now, but we can *estimate* them from genotype frequency data: For A1,

$$p = \frac{2\cdot 31 + 89}{2\cdot 242} \approx 0{,}312$$

- Thus, f(A2)=q=1-p=1-0,312=0,688

- Thus in HWE, the following numbers of each genotype are expected:
  - **AA:** **$p^2 \cdot N = 0{,}312^2 \cdot 242 = 23{,}56$**
  - **Aa:** **$2pq \cdot N = 2 \cdot 0{,}312 \cdot 0{,}688 \cdot 242 = 103{,}89$**
  - **aa:** **$q^2 \cdot N = 0{,}688^2 \cdot 242 = 114{,}55$**
- This is called the **null hypothesis,** H_0 distribution
- Let's compare the expected with the observed ones!

	A1A1	A1A2	A2A2	Sum
Observed	31	89	122	242
Expected	23,56	103,89	114,55	242,00

- Do the observed numbers coincide with the expected *well enough*?
- The correspondence is evaluated with a **test statistic,** here with
- **χ^2 -contingency test**
- The test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

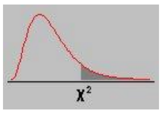
Where k is the number of different classes (genotypes)

- Now,

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i} = \frac{(31-23{,}56)^2}{23{,}56} + \frac{(89-103{,}89)^2}{103{,}89} + \frac{(122-114{,}55)^2}{114{,}55} \approx 4{,}97$$

- What does 4,97 mean?
- It can be shown that χ^2-test statistic asymptotically follows **χ^2-distribution** with *given degrees of freedom*, if H_0 is true
- Degrees of freedom (df) = number of classes – 1 – number of parameters estimated.
- Now, there are 3 classes (3 genotypes) and df=3-1-1=1 (we estimated two allele frequencies which means **one parameter estimated. Note that q=1-p!**)

- Check χ^2-distribution or table of χ^2_p (**critical values**) on different **significance levels p**



- E.g. if H_0 is true and df=1, there is 0.05 probability of getting a test statistic value equal or greater than the critical value 3.841, in other words (χ^2_3>3,841)=0.05
- In our example, the value of the test statistic was 4,97: the probability of getting it if H_0 is true is less than 0.05.
- -> the deviation is said to be **statistically significant**
- -> H_0 is rejected at significance level p (here, 0.05)
- -> in our example, this means that the observed genotypes are **not** in HWE!

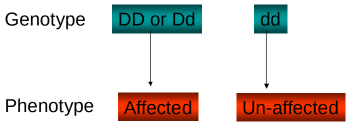
χ^2-distributions: critical values on varying numbers of degrees of freedom and significance levels

df	0,995	0,9500	0,100	0,050	0,025	0,010	0,005
1	0,000	0,004	2,706	3,842	5,024	6,635	7,879
2	0,010	0,103	4,605	5,992	7,378	9,210	10,597
3	0,072	0,352	6,251	7,815	9,348	11,345	12,838
4	0,207	0,711	7,779	9,488	11,143	13,277	14,860
5	0,412	1,146	9,236	11,071	12,833	15,086	16,750
6	0,676	1,635	10,645	12,592	14,449	16,812	18,548
7	0,989	2,167	12,017	14,067	16,013	18,475	20,278
8	1,344	2,733	13,362	15,507	17,535	20,090	21,955
9	1,735	3,325	14,684	16,919	19,023	21,666	23,589
10	2,156	3,940	15,987	18,307	20,483	23,209	25,188
11	2,603	4,575	17,275	19,675	21,920	24,725	26,757
12	3,074	5,226	18,549	21,026	23,337	26,217	28,300
13	3,565	5,892	19,812	22,362	24,736	27,688	29,819
14	4,075	6,571	21,064	23,685	26,119	29,141	31,319
15	4,601	7,261	22,307	24,996	27,488	30,578	32,801

- **Note: restrictions on the use of χ^2-tests**
  - At most, 20 % of expected frequencies <5
  - Every expected frequency >1
- **Note!** χ^2-tests are always performed with absolute frequencies, **not** with relative frequencies!
- Note: test is implemented in most genetic software (e.g.,PEDSTATS, so in reality there is no need to calculate it manually

Estimating genotype frequencies when only phenotype frequencies are known for a dominant trait

- Dominance in the locus, D>d



$f(DD) + f(Dd) = f(Affected)$ $\quad f(dd) = f(Un-affected)$

Solution: Assume that the locus in the population is in HWE

- Note: *f(D)=p, f(d)=q*

$$f(Un-affected) = f(dd) = q^2$$

f(d)= $q = \sqrt{q^2}$

and because $p+q=1$,

$p = 1-\sqrt{q^2}$

- Thus, the genotype frequencies are

$$f(DD) = p^2 = [1-\sqrt{q^2}]^2$$

$$f(Dd) = 2pq = 2\sqrt{q^2}[1-\sqrt{q^2}]$$

- By assuming HWE and knowing one of the genotype frequencies we are able to calculate the rest