HARDY-WEINBERG EQUILIBRIUM

Normal-based Tests

Hardy-Weinberg Law

For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

These are also the results of setting the inbreeding coefficient f to zero.

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i}p_{A_j} \end{aligned}$$

Why would HWE not hold?

- Natural selection.
- LD with trait in trait-only sample.
- Population Structure/Admixture.
- Problems with data.
- etc.

Problems with Data

A SNP with genotype counts $40, 0, 60$ for AA, Aa, aa is likely to cause HWE rejection. What about $4, 0, 6$?

Typing systems may report heterozygotes as homozygotes, as was the likely explanation for
"To justify applying the classical formulas of population genetics in the Castro case, the Hispanic population must be in Hardy-Weinberg equilibrium. In fact, Lifecodes' own data show that it is not. ... Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium: 17 per cent observed homozygotes at D2S44 and 13 per cent observed homozygotes at D17S79 compared with only 4 per cent expected at each locus, indicating, perhaps not surprisingly, the presence of genetically distinct subgroups within the Hispanic sample."

Lander ES. 1989. DNA fingerprinting on trial. Nature 339:501-505.

Population Structure

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the *Wahlund effect*.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Admixture: Departures from HWE

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marriages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Population 2 is in HWE, but Population 1 has 51% heterozygotes instead of the expected 49.8%.

Inference about HWE

If $\widehat{f}$ is the MLE of the within-population inbreeding coefficient f, it has a normal distribution for large sample sizes n (and for large np). It can be transformed into a standard normal variable z by

$$z = \frac{\widehat{f} - f}{\sqrt{\mathrm{Var}(\widehat{f})}}$$

If the true value f is zero, then $\mathrm{Var}(\widehat{f}) = 1/n$, and $X^2 = z^2$ has a chi-square distribution with one degree of freedom:

$$X^2 = \left(\frac{\widehat{f} - 0}{\sqrt{1/n}}\right)^2 = n\widehat{f}^2 \sim \chi 2_{(1)}$$

The HWE hypothesis is rejected at the 5% significance level if $X^2 > 3.84$.

Aside: Inference about HWE

Departures from HWE can be described by the within-population inbreeding coefficient f. This has an MLE that can be written as

$$\hat{f} = 1 - \frac{\tilde{P}_{AB}}{2\tilde{p}_A\tilde{p}_B} = \frac{4n_{AA}n_{BB} - n_{AB}^2}{(2n_{AA} + n_{AB})(2n_{BB} + n_{AB})}$$

and we can use "Delta method" to find

$$\begin{aligned} \mathcal{E}(\hat{f}) &= f \\ \text{Var}(\hat{f}) &\approx \frac{1}{2np_Ap_B}(1-f)[2p_Ap_B(1-f)(1-2f) + f(2-f)] \end{aligned}$$

If $\hat{f}$ is assumed to be normally distributed then, $(\hat{f}-f)/\sqrt{\text{Var}(\hat{f})} \sim N(0,1)$. When H_0 is true, the square of this quantity has a chi-square distribution.

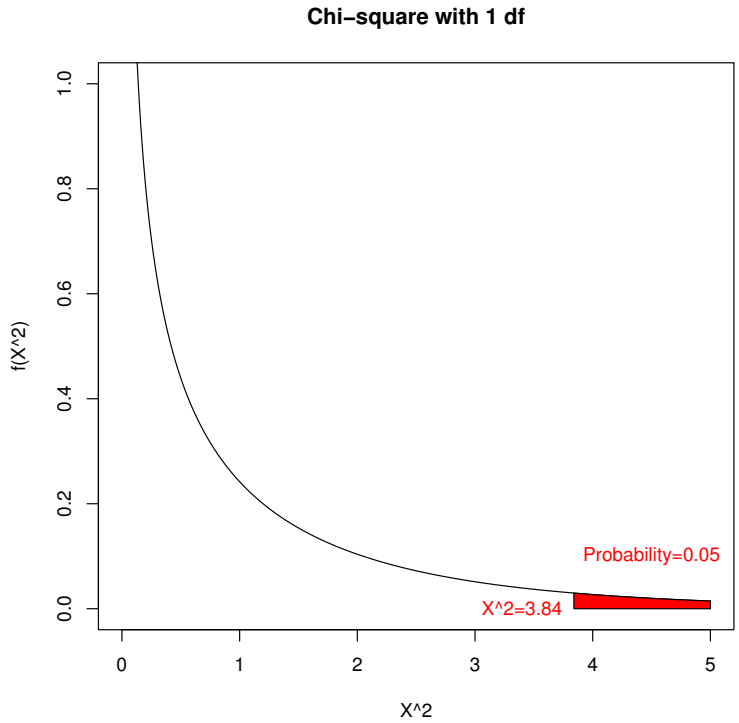
Aside: Inference about HWE

Since $\text{Var}(\widehat{f}) = 1/n$ when $f = 0$:

$$\begin{aligned} X^2 &= \left(\frac{\widehat{f} - f}{\sqrt{\text{Var}(\widehat{f})}} \right)^2 \\ &= \frac{\widehat{f}^2}{1/n} \\ &= n\widehat{f}^2 \end{aligned}$$

is appropriate for testing $H_0 : f = 0$. When H_0 is true, $X^2 \sim \chi^2_{(1)}$. Reject HWE if $X^2 > 3.84$.

Significance level of HWE test



The area under the chi-square curve to the right of $X^2 = 3.84$ is the probability of rejecting HWE when HWE is true. This is the significance level of the test.

Goodness-of-fit Test

An alternative, but equivalent, test is the goodness-of-fit test.

Genotype	Observed	Expected	$\frac{(\text{Obs.}-\text{Exp.})^2}{\text{Exp.}}$
AA	n_{AA}	$n\tilde{p}_A^2$	$n\tilde{p}_a^2\tilde{f}^2$
Aa	n_{Aa}	$2n\tilde{p}_A\tilde{p}_a$	$2n\tilde{p}_A\tilde{p}_a\tilde{f}^2$
aa	n_{aa}	$n\tilde{p}_a^2$	$n\tilde{p}_A^2\tilde{f}^2$

The test statistic is

$$X^2 = \sum \frac{(\text{Obs.} - \text{Exp})^2}{\text{Exp.}} = n\tilde{f}^2$$

Goodness-of-fit Test

Does a sample of 6 AA, 3 Aa, 1 aa support Hardy-Weinberg?

First need to estimate allele frequencies:

$$\tilde{p}_A = \tilde{P}_{AA} + \frac{1}{2}\tilde{P}_{Aa} = 0.75$$

$$\tilde{p}_a = \tilde{P}_{aa} + \frac{1}{2}\tilde{P}_{Aa} = 0.25$$

Then form "expected" counts:

$$\begin{aligned} n_{AA} &= n(\tilde{p}_A)^2 = 5.625 \\ n_{Aa} &= 2n\tilde{p}_A\tilde{p}_a = 3.750 \\ n_{aa} &= n(\tilde{p}_a)^2 = 0.625 \end{aligned}$$

Goodness-of-fit Test

Perform the chi-square test:

Genotype	Observed	Expected	(Obs. $-$ Exp.)^2/Exp.
AA	6	5.625	0.025
Aa	3	3.750	0.150
aa	1	0.625	0.225
Total	10	10	0.400

Note that $\widehat{f} = 1 - 0.3/(2 \times 0.75 \times 0.25) = 0.2$ and $X^2 = n\widehat{f}^2$.

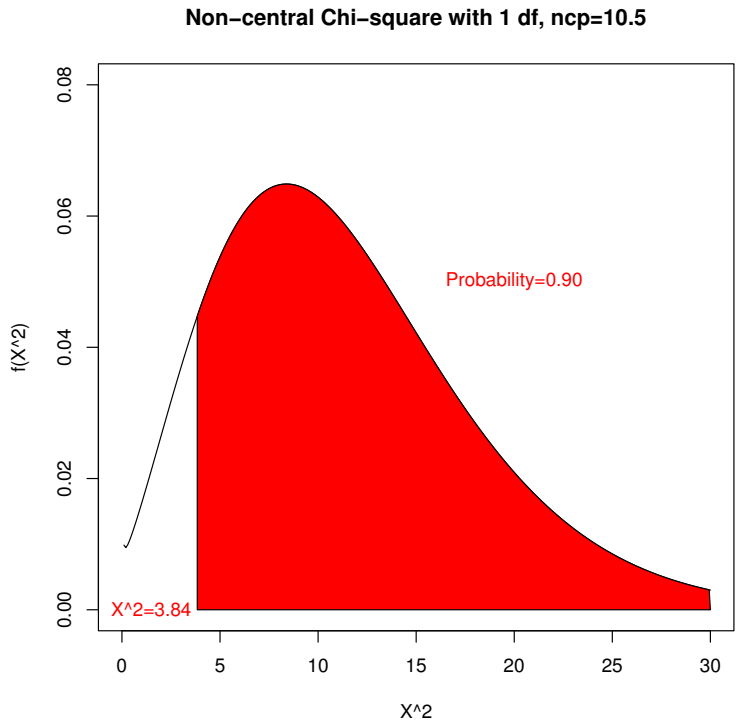
Sample size determination

Although Fisher's exact test (below) is generally preferred for small samples, the normal or chi-square test has the advantage of simplifying power calculations.

When the Hardy-Weinberg hypothesis is not true, the test statistic $n\widehat{f}^2$ has a non-central chi-square distribution with one degree of freedom (df) and non-centrality parameter $\lambda = nf^2$. To reach 90% power with a 5% significance level, for example, it is necessary that $\lambda \geq 10.51$.

```
> pchisq(3.84,1,0)
[1] 0.9499565
> pchisq(3.84,1,10.51)
[1] 0.09986489
> qchisq(0.95,1,0)
[1] 3.841459
> qchisq(0.10,1,10.51)
[1] 3.843019
```

Power of HWE test



The area under the non-central chi-square curve to the right of $X^2 = 3.84$ is the probability of rejecting HWE when HWE is false. This is the power of the test. In this plot, the non-centrality parameter is $\lambda = 10.5$.

Sample size determination

To achieve 90% power to reject HWE at the 5% significance level when the true inbreeding coefficient is f, need sample size n to make $nf^2 \geq 10.51$.

For $f = 0.01$, need $n \geq 10.51/(0.01)^2 = 105,100$.

For $f = 0.05$, need $n \geq 10.51/(0.05)^2 = 4,204$.

For $f = 0.10$, need $n \geq 10.51/(0.10)^2 = 1,051$.

Significance Levels and p-values

The *significance level* α of a test is the probability of a false rejection. It is specified by the user, and along with the null hypothesis, it determines the rejection region. The specified, or "nominal" value may not be achieved for an actual test.

Once the test has been conducted on a data set, the probability of the observed test statistic, *or a more extreme value*, if the null hypothesis is true is the p-*value*. The chi-square and normal tests shown above give approximate p-values because they use a continuous distribution for discrete data (and because the sample allele frequencies are not normally distributed).

An alternative class of tests, "exact tests," use a discrete distribution for discrete data and provide accurate p-values. It may be difficult to construct an exact test with a particular nominal significance level.

Exact Tests

HWE Exact Test

If the counts of genotypes AA, Aa, aa are n_{AA}, n_{Aa}, n_{aa} in a sample of n individuals, and if the sample allele counts are $n_A = 2n_{AA} + n_{Aa}$ and $n_a = 2n_{aa} + n_{Aa}$, then the probability of the genotypic data *conditional on the allele counts* if there is HWE is

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a) = \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

HWE is rejected if this probability is amongst the smallest probabilities for all possible sets of genotype counts for those allele counts.

The p-value for the dataset is this probability plus probabilities for other possible sets of genotype counts that are smaller than this probability.

Aside: Exact HWE Test

The preferred test for HWE is an exact one. The test rests on the assumption that individuals are sampled randomly from a population so that genotype counts have a multinomial distribution:

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{n_{AA}}(P_{Aa})^{n_{Aa}}(P_{aa})^{n_{aa}}$$

This equation is always true, and when there is HWE ($P_{AA} = p_A^2$ etc.) there is the additional result that the allele counts have a binomial distribution:

$$\Pr(n_A, n_a) = \frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a}$$

Aside: Exact HWE Test

Putting these together gives the conditional probability

$$\begin{aligned}
\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a) &= \frac{\Pr(n_{AA}, n_{Aa}, n_{aa} \text{ and } n_A, n_a)}{\Pr(n_A, n_a)} \\
&= \frac{\frac{n!}{n_{AA}! n_{Aa}! n_{aa}!}(p_A^2)^{n_{AA}}(2p_A p_a)^{n_{Aa}}(p_a^2)^{n_{aa}}}{\frac{(2n)!}{n_A! n_a!}(p_A)^{n_A}(p_a)^{n_a}} \\
&= \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!}\frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}
\end{aligned}$$

Reject the Hardy-Weinberg hypothesis if this quantity, the probability of the genotypic array conditional on the allelic array, is considered too small to allow that outcome if HWE holds. Is the probability for the data among the smallest of its possible values?

Exact HWE Test Example

For convenience, write the probability of the genotypic array, conditional on the allelic array and HWE, as $\Pr(n_{Aa}|n, n_A)$. Reject the HWE hypothesis for a data set if this value is among the smallest probabilities.

As an example, consider $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$. The allele counts are $(n_A = 2, n_a = 98)$ and there are only two possible genotype arrays:

AA	Aa	aa	$\Pr(n_{Aa}\|n, n_A)$
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$

The p-value is 1/99=0.01 and HWE is rejected at the 5% level.

Exact HWE Test Example

In this example, $\widehat{f} = 0$ and the chi-square test statistic is $X^2 = 50$. The resulting p-value is 1.54×10^{-12}, substantially different from the exact value of 0.01.

```
> 1-pchisq(50,1,0)
[1] 1.537437e-12
```

Exact HWE Test Example

As another example, the sample with $n_{AA} = 6, n_{Aa} = 3, n_{aa} = 1$ has allele counts $n_A = 15, n_a = 5$. There are two other sets of genotype counts possible and the probabilities of each set for a HWE population are:

n_{AA}	n_{Aa}	n_{aa}	n_A	n_a	$\Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a)$
7	1	2	15	5	$\frac{10!}{7!1!2!}\frac{2^1 15!5!}{20!} = \frac{15}{323} = 0.047$
6	3	1	15	5	$\frac{10!}{6!3!1!}\frac{2^3 15!5!}{20!} = \frac{140}{323} = 0.433$
5	5	0	15	5	$\frac{10!}{5!5!0!}\frac{2^5 15!5!}{20!} = \frac{168}{323} = 0.520$

The p-value is $0.433 + 0.047 = 0.480$. Compare this to the chi-square p-value for $X^2 = 0.40$:

```
> pchisq(0.4,1)
[1] 0.4729107
```

Exact HWE Test Example

For a sample of size $n = 100$ with minor allele frequency of 0.07, there are 8 sets of possible genotype counts:

			Exact		Chi-square	
n_{AA}	n_{Aa}	n_{aa}	Prob.	p value	X^2	p value
93	0	7	0.0000	0.0000*	100.00	0.0000*
92	2	6	0.0000	0.0000*	71.64	0.0000*
91	4	5	0.0000	0.0000*	47.99	0.0000*
90	6	4	0.0002	0.0002*	29.07	0.0000*
89	8	3	**0.0051**	**0.0053***	14.87	0.0001*
88	10	2	0.0602	0.0655	**5.38**	**0.0204***
87	12	1	0.3209	0.3864	0.61	0.4348
86	14	0	0.6136	1.0000	0.57	0.4503

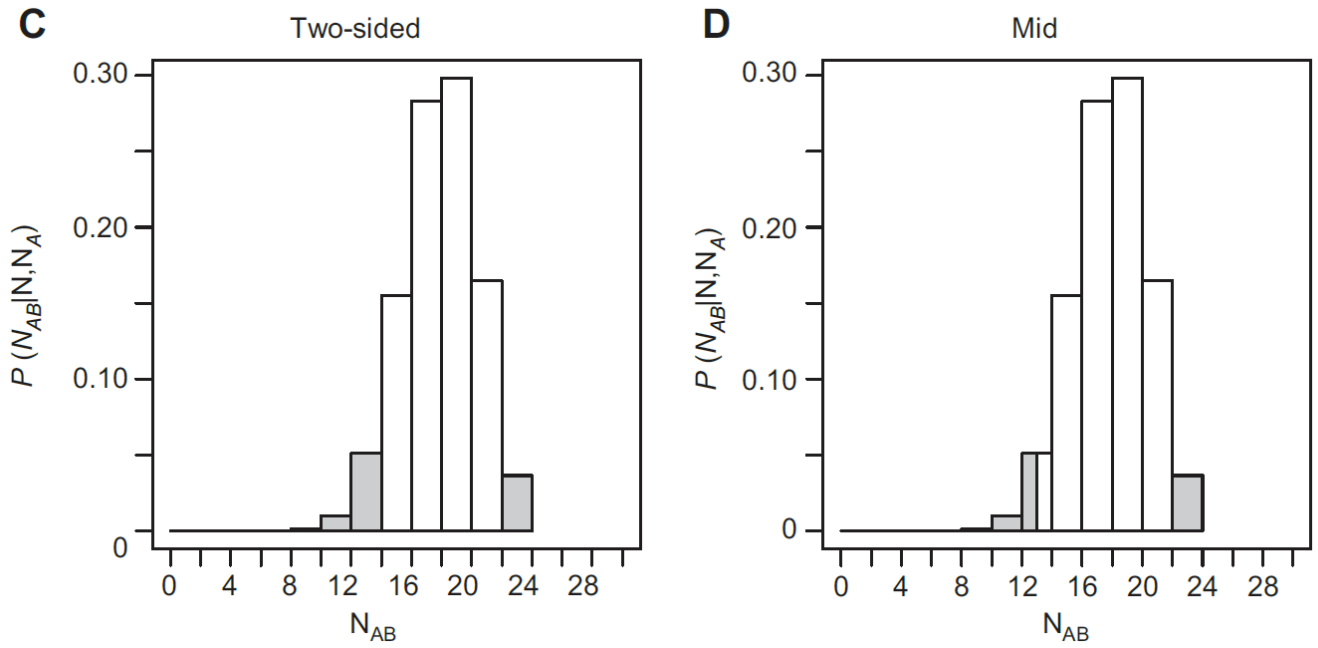
So, for a nominal 5% significance level, the actual significance level is 0.0053 for an exact test that rejects when $n_{Aa} \leq 8$ and is 0.0204 for an exact test that rejects when $n_{AB} \leq 10$.

Modified Exact HWE Test

Traditionally, the p-value is the probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true and are conditional on the observed allele frequencies. More recently Graffelman and Moreno showed that the test has a significance value closer to the nominal value if the p-value is half the probability of the data plus the probabilities of all datasets that are less probably under the null hypothesis. For the $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$ example then, the p-value is 1/198.

Graffelman J, Moreno V. 2013. Statistical Applications in Genetics and Molecular Biology 12:433-448

Graffelman and Moreno, 2013



Computation of the p-value in an exact test for HWP, for a sample of 50 individuals with a minor allele count of 23, for which 13 heterozygotes were observed. (C) Standard two-sided p-value, (D) Mid p-value based on half the probability of the observed sample.

Usual vs Mid p values

				p value	
AA	Aa	aa	$\Pr(n_{Aa}\|n, n_A)$	Usual	Mid
5	5	0	0.520	1.000	0.740
6	3	1	0.433	0.480	0.287
7	1	2	0.047	0.047	0.023

Modified Exact HWE Test Example

For a sample of size $n = 100$ with minor allele frequency of 0.07, there are 8 sets of possible genotype counts:

			Exact		Chi-square	
n_{AA}	n_{Aa}	n_{aa}	Prob.	Mid p value	X^2	p value
93	0	7	0.0000	0.0000*	100.00	0.0000*
92	2	6	0.0000	0.0000*	71.64	0.0000*
91	4	5	0.0000	0.0000*	47.99	0.0000*
90	6	4	0.0002	0.0002*	29.07	0.0000*
89	8	3	0.0051	0.0028*	14.87	0.0001*
88	10	2	**0.0602**	**0.0353***	**5.38**	**0.0204***
87	12	1	0.3209	0.2262	0.61	0.4348
86	14	0	0.6136	0.6832	0.57	0.4503

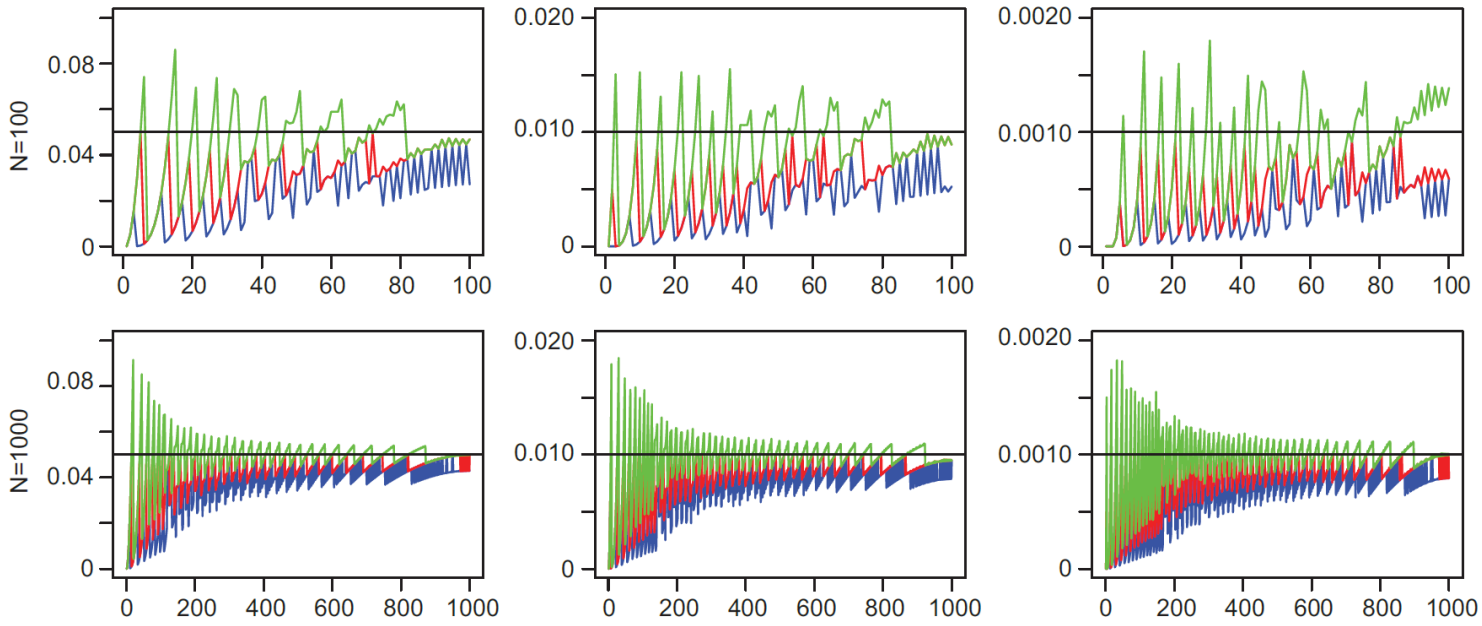
So, for a nominal 5% significance level, the actual significance level is 0.0353 for an exact test that rejects when $n_{Aa} \leq 10$ and is 0.0204 for a chi-square test that also rejects when $n_{AB} \leq 10$.

Effect of Minor Allele Frequency

Even though the nominal significance level for a HWE test may be set at 0.05, for example, the actual significance level can be quite different. (e.g. 0.0353 vs 0.05 on the previous slide.)

The difference between nominal and actual values depends on the sample size and the minor allele frequency, as shown on the next slide.

Graffelman and Moreno, 2013



Type I error rate against minor allele count for sample sizes 100 and 1000 and significance levels (0.05, 0.01, and 0.001) for exact tests with standard two-sided (red), doubled one-sided (blue) and mid p-values (green).

Power of Exact Test

Calculating the power of an HWE test is easy for the chi-square test statistic as it follows from the non-central chi-square distribution.

It is more complicated for the exact test, and the power depends on the quantity $\psi = P_{Aa}/(\sqrt{P_{AA}P_{aa}})$, involving the genotype probabilities in the population. This quantity depends on both the inbreeding coefficient f and the allele probabilities p_A, p_a in the population.

Aside: Power of exact test

If there is not HWE:

$$\begin{aligned}
\Pr(n_{Aa}|n_A, n_a) &= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{n_{AA}}(P_{Aa})^{n_{Aa}}(P_{aa})^{n_{aa}} \\
&= \frac{n!}{n_{AA}!n_{Aa}!n_{AA}!}(P_{AA})^{\frac{n_A - n_{Aa}}{2}}(P_{Aa})^{n_{Aa}}(P_{aa})^{\frac{n_a - n_{Aa}}{2}} \\
&= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}\sqrt{P_{AA}^{n_A}}\sqrt{P_{aa}^{n_a}}\left(\frac{P_{Aa}}{\sqrt{P_{AA}P_{aa}}}\right)^{n_{Aa}} \\
&= \frac{C\psi^{n_{Aa}}}{n_{AA}!n_{Aa}!n_{aa}!}
\end{aligned}$$

where $\psi = P_{Aa}/(\sqrt{P_{AA}P_{aa}})$ measures the departure from HWE. The constant C makes the probabilities sum to one over all possible n_{Aa} values: $C = 1/[\sum_{n_{Aa}} \psi^{n_{Aa}}/(n_{AA}!n_{Aa}!n_{aa}!)]$.

Power of Exact Test

Once the rejection region has been determined, the power of the test (the probability of rejecting) can be found by adding these probabilities for all sets of genotype counts in the region. HWE corresponds to $\psi = 2$. What is the power to detect HWE when $\psi = 1 (f > 0)$, the sample size is $n = 10$ and the sample allele frequencies are $\tilde{p}_A = 0.75, \tilde{p}_a = 0.25$?

			$\Pr(n_{Aa} \mid n_A, n)$	
n_{AA}	n_{Aa}	n_{aa}	$\psi = 2$	$\psi = 1$
7	1	2	0.047	0.374
6	3	1	0.433	0.364
5	5	0	0.520	0.262

The $\psi = 2$ column shows that the rejection region is $n_{Aa} = 1$, and significance level is 4.7%.

The $\psi = 1$ column shows that the power (the probability $n_{Aa} = 1$ when $\psi = 1$) is 37.4%.

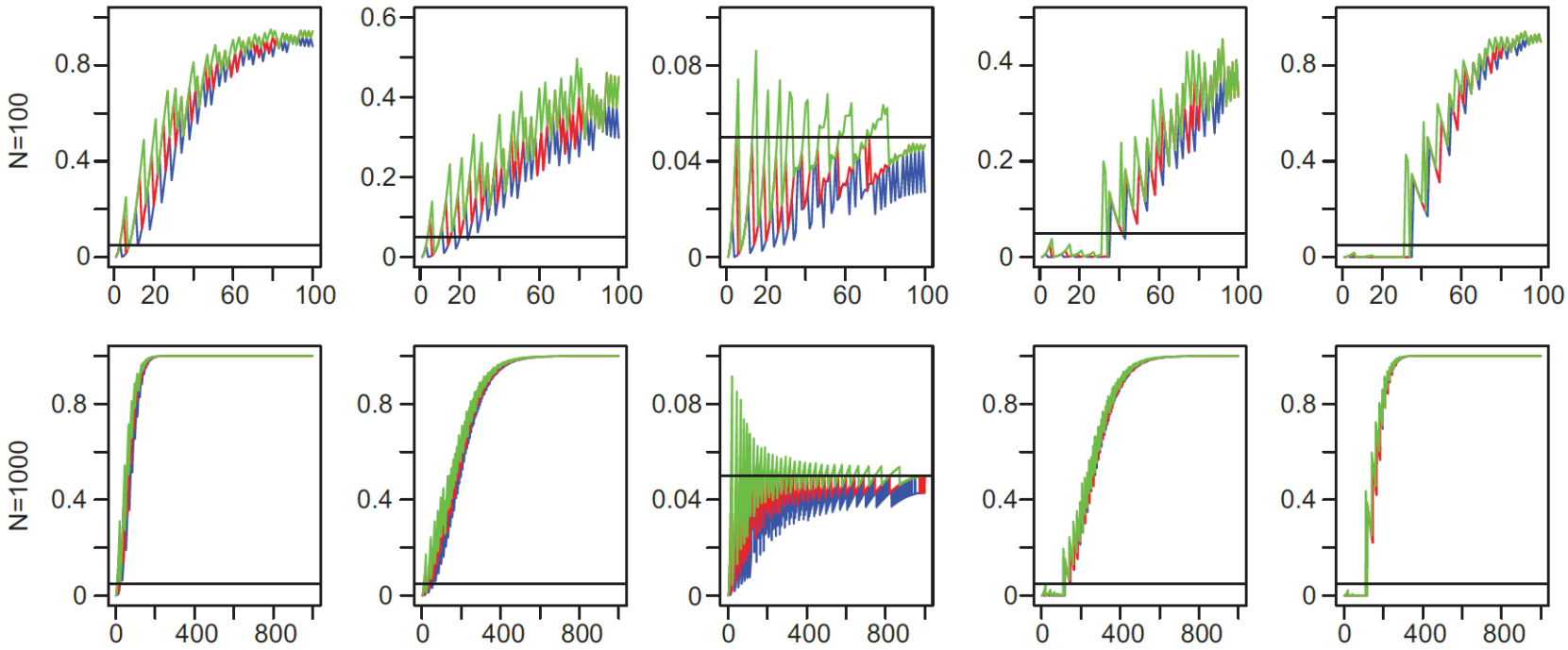
Power Examples

For given values of n, n_a, the rejection region is determined from null hypothesis and the power is determined from the multinomial distribution.

		$\Pr(n_{Aa} \mid n_a = 16, n = 100)$						
	ψ	.250	.500	1.000	2.000	4.000	8.000	16.000
n_{Aa}	f	.631	.398	.157	.000	−.062	−.081	−.085
0		.0042	.0000	.0000	.0000	.0000	.0000	.0000
2		.0956	.0026	.0000	.0000	.0000	.0000	.0000
4		.3172	.0349	.0003	.0000	.0000	.0000	.0000
6		.3568	.1569	.0056	.0000	.0000	.0000	.0000
8		.1772	.3116	.0441	.0008	.0000	.0000	.0000
10		.0433	.3047	.1725	.0123	.0003	.0000	.0000
12		.0054	.1506	.3411	.0974	.0098	.0007	.0000
14		.0003	.0356	.3223	.3681	.1485	.0422	.0109
16		.0000	.0032	.1142	.5214	.8414	.9571	.9890
	Power*	.9943	.8107	.2225	.0131	.0003	.0000	.0000

* $\Pr(n_{Aa} \leq 10)$.

Graffelman and Moreno, 2013



Power ofmHWE exact tests against minor allele count for sample sizes 100 and 1000 and disequilibria 1,2,4,8,16. Standard two-sided (red), double one-sided (blue) and mid p-values (green).

Permutation Test

For large sample sizes and many alleles per locus, there are too many genotypic arrays for a complete enumeration and a determination of which are the least probable 5% arrays.

A large number of the possible arrays is generated by permuting the alleles among genotypes, and calculating the proportion of these permuted genotypic arrays that have a smaller conditional probability than the original data. If this proportion is small, the Hardy-Weinberg hypothesis is rejected.

This procedure is not needed for SNPs with only 2 alleles. The number of possible arrays is always less than about half the sample size.

Multiple Testing

When multiple tests are performed, each at significance level α, a proportion α of the tests are expected to cause rejection even if all the hypotheses are true.

Bonferroni correction makes the overall (experimentwise) significance level equal to α by adjusting the level for each individual test to α'. If α is the probability that at least one of the L tests causes rejection, it is also 1 minus the probability that none of the tests causes rejection:

$$\begin{aligned} \alpha &= 1 - (1 - \alpha')^L \\ &\approx L\alpha' \end{aligned}$$

provided the L tests are independent.

If $L = 10^6$, the "genome-wide significance level" is 5×10^{-8} in order for $\alpha = 0.05$.

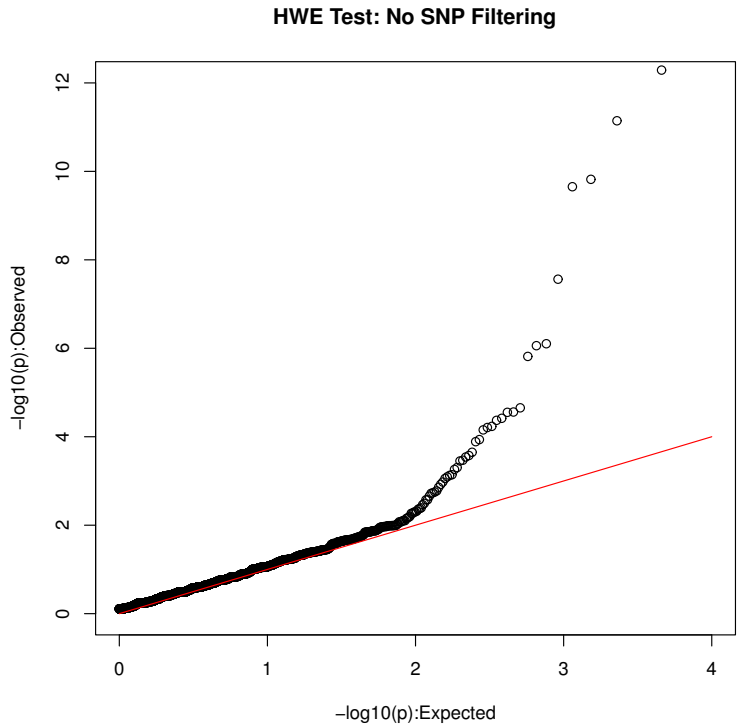
QQ-Plots

An alternative approach to considering multiple-testing issues is to use QQ-plots. If all the hypotheses being tested are true then the resulting p-values are uniformly distributed between 0 and 1.

For a set of n tests, the n p-values are expected to be evenly spread p values between 0 and 1 e.g. $1/2n, 3/2n, \ldots, (2n-1)/2n$. The observed p-values can be plotted against these expected values: the smallest against $1/2n$ and the largest against $(2n-1)/2n$. It is more convenient to transform to $-\log_{10}(p)$ to accentuate the extremely small p values. The point at which the observed values start departing from the expected values is an indication of "significant" values in a way that takes into account the number of tests.

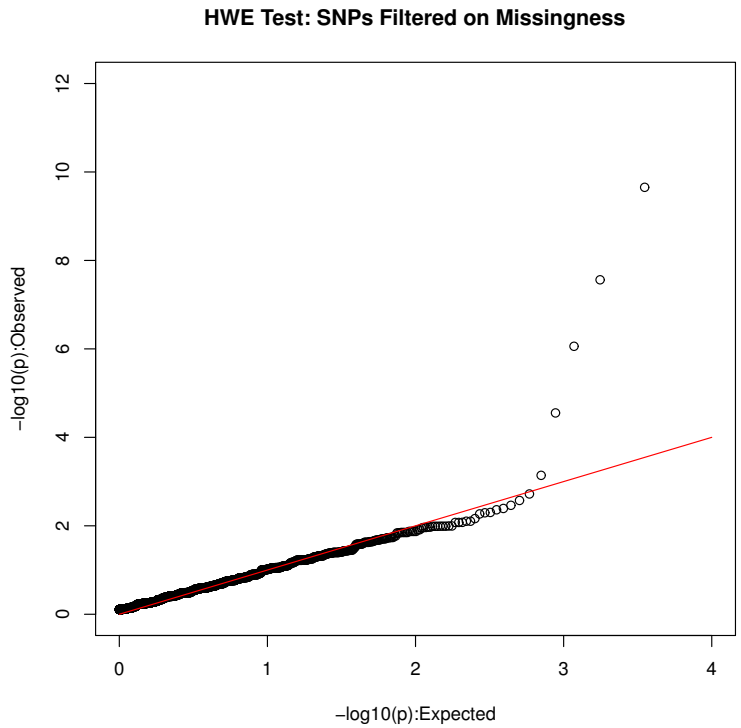
A useful diagnostic for QQ-plots is the "genomic control" quantity λ. This is the ratio of the median of the observed p-values to the median of the expected values. If the expected p-values have a uniform distribution on [0,1], under the null hypothesis of HWE, the median is 0.5. The λ ratio should be 1.

QQ-Plots



The results for 9208 SNPs on human chromosome 1 for 50 AMD controls ($\lambda = 0.86$). Bonferroni would suggest rejecting HWE when $p \leq 0.05/9208 = 5.4 \times 10^{-6}$ or $-\log_{10}(p) \geq 5.3$.

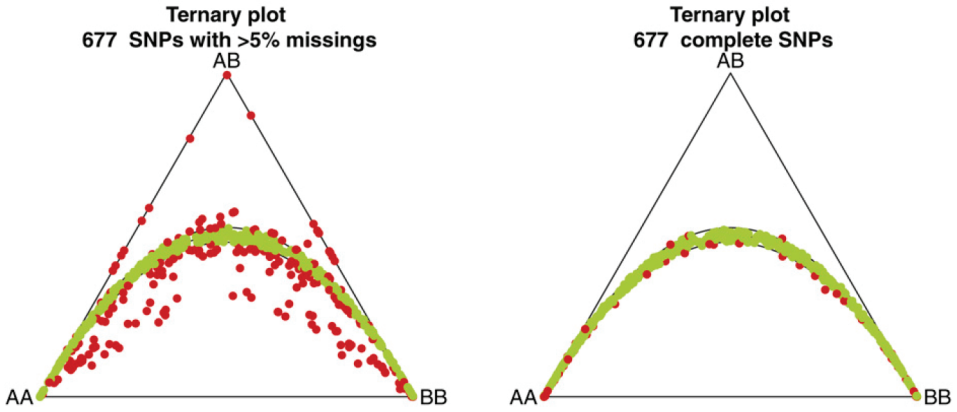
QQ-Plots



The same set of results as on the previous slide except now that any SNP with any missing data was excluded ($\lambda = 1.035$, closer to 1 than for all the SNPs). Now 7446 SNPs and Bonferroni would reject if $-\log_{10}(p) \geq 5.2$. All five outliers had zero counts for the minor allele homozygote and at least 32 heterozygotes in a sample of size 50.

Imputing Missing Data

Instead of discarding an individual for any SNP when there is no genotype call, it may be preferable to use neighboring SNPs to impute the missing values. Graffelman applied this procedure to a study on pre-term birth:

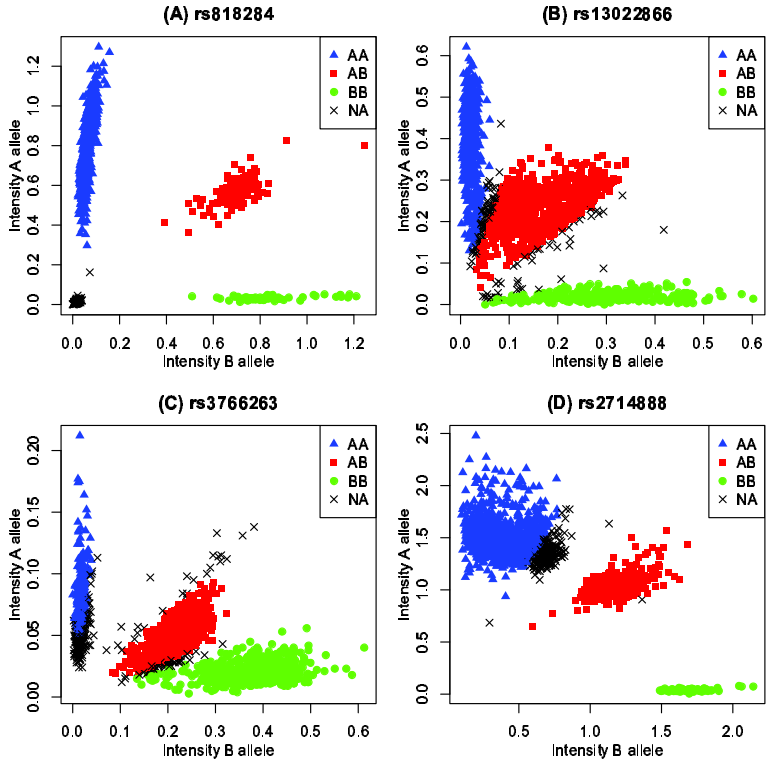


Significant markers are red and non-significant markers are green ($\alpha = 0.05$).

Ternary plot: distance of point to side of triangle is frequency of genotype shown on opposite vertex.

Graffelman J, et al. 2015, G3 (Genes, Genomes, Genetics) 5:2365-2373.

Imputing Missing Data



	p-value		
SNP	Discard	Impute	Comment
rs818284	0.000	0.000	Null alleles
rs13022866	0.046	0.571	Het deficiency
rs3766263	0.020	0.539	Het excess
rs2714888	0.192	0.007	Hom deficiency

Separate Sexes

HWE Test for X-linked Markers

It is usual to test HWE for X-linked markers using only females.

Under HWE allele frequencies for SNPs in males and females, on the X chromosome, should be the same. Should examine the difference allele frequencies when testing for HWE.

If a sample has n_m males and n_f females, and if the males have m_A, m_a alleles of types A, a, and if females have f_{AA}, f_{Aa}, f_{aa} genotypes AA, Aa, aa, then the probability of the data, under HWE, is

$$\frac{n_A! n_a! n_m! n_f!}{m_A! m_a! f_{AA}! f_{Aa}! f_{aa}! n_t!} 2^{f_{Aa}}$$

where $n_t = n_m + 2n_f$.

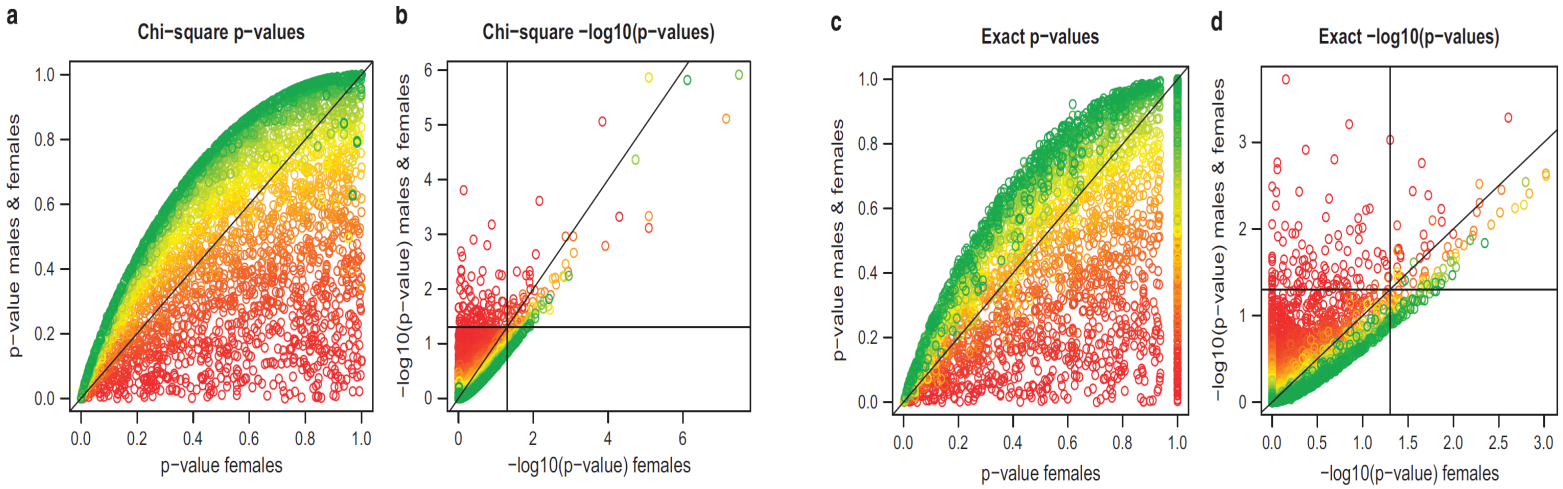
Graffelman and Weir, 2016, Heredity 116:558-568.

Example: 10 males, 10 females, 6 A alleles

If there are six A alleles in a sample that has 10 males and 10 females, there are 16 possible datasets:

Set	m_A	m_a	f_{AA}	f_{Aa}	f_{aa}	Probability
1	0	10	3	0	7	0.0002
2	6	4	0	0	10	0.0004
3	0	10	0	6	4	0.0026
4	2	8	2	0	8	0.0034
5	4	6	1	0	9	0.0035
6	5	5	0	1	9	0.0085
7	0	10	2	2	6	0.0085
8	1	9	2	1	7	0.0121
9	0	10	1	4	5	0.0340
10	3	7	1	1	8	0.0364
11	4	6	0	2	8	0.0637
12	2	8	1	2	7	0.1091
13	1	9	1	3	6	0.1132
14	1	9	0	5	5	0.1358
15	3	7	0	3	7	0.1940
16	2	8	0	4	6	0.2546

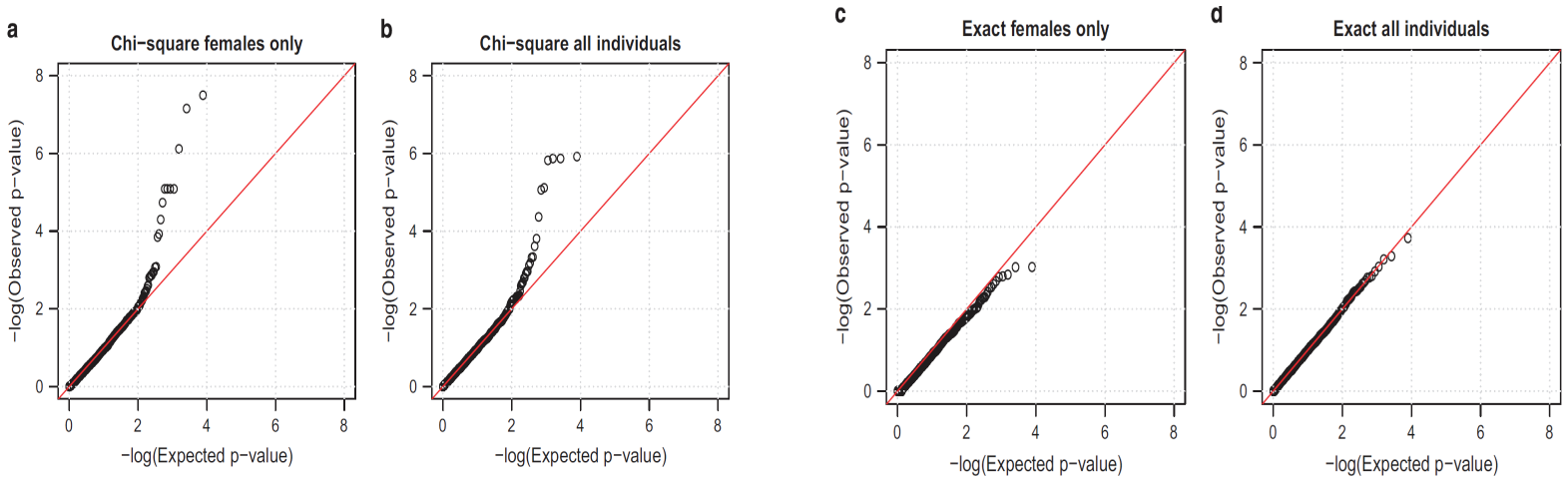
X-linked Markers: Real Data



Scatter plots of P-values in original and -log10 scale for chi-square tests (a, b) and exact tests (c, d) for HWE using females only and using both males and females for 4158 SNPs at the X chromosome of the venous thrombosis database. The horizontal and vertical black lines in (b) and (d) correspond to a significance level of 5%. Points colored according to their significance level in Fisher's test for equality of allele frequencies (range 0-1 from red to green).

Graffelman J, Weir BS. 2016. Heredity 116:558-568.

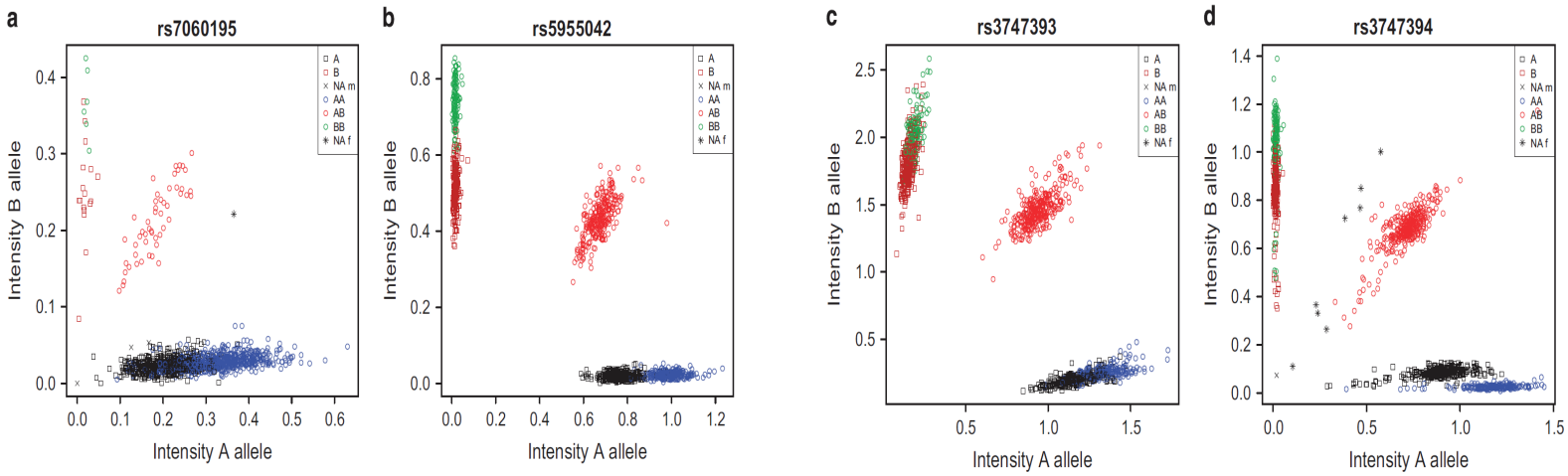
X-linked Markers: Real Data



QQ plots of - log10 transformed P-values of Chi-square and exact tests for HWE for 4158 SNPs of the venous thrombosis database. (a, c) Females only and (b, d) all individuals.

Graffelman J, Weir BS. 2016. Heredity 116:558-568.

X-linked Markers: Real Data

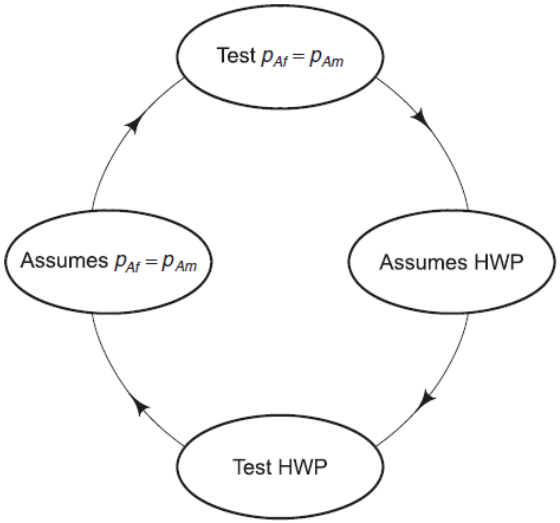


Cluster plots of allele intensities of four SNPs of the venous thrombosis database. (a and b) are significant in both the female-only (P=0.0025, P=0.0010) and all-individual test (P=0.0005, P=0.0023). (c) is non-significant in the female-only test (P=0.4261) but highly significant in the all-individual test (P=0.0012). (d) is non-significant in the female-only test (P=0.8732) and close to significant in the all-individual test (P=0.0914).

Graffelman J, Weir BS. 2016. Heredity 116:558-568.

Separate Male and Female Autosomal Counts

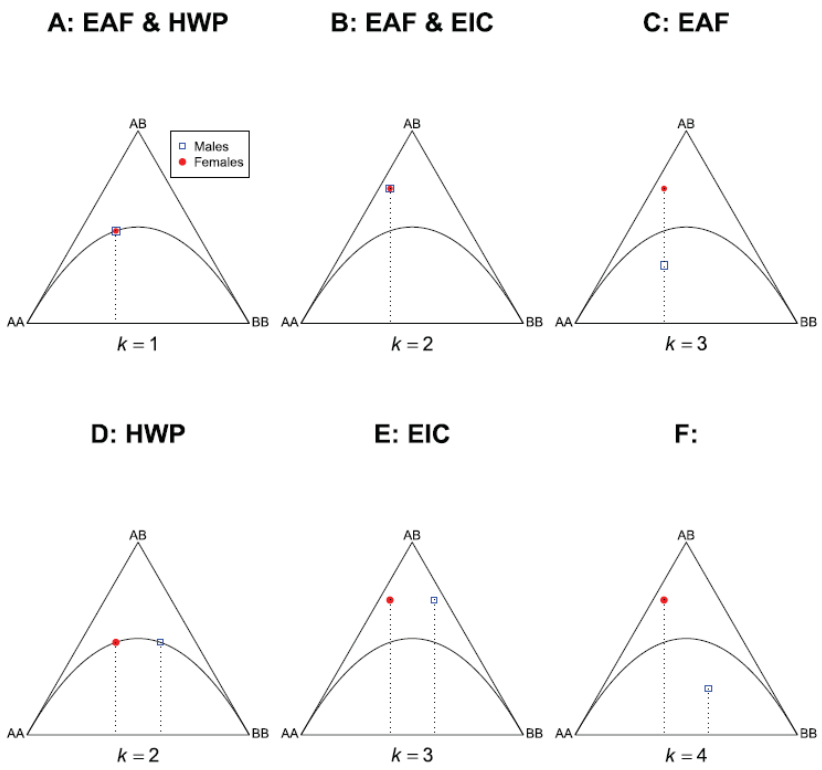
The X-linked test can be extended to autosomal markers when genotype counts are recorded separately for males and females.



Vicious testing circle: mutual dependency of a test for EAF in males and females and a test for HWP Notes: A allele frequencies in males and females are represented by p_{A_m} and p_{A_f}, respectively.

Graffeleman J, Weir BS. 2018. Genetic Epidemiology 42:24-48.

Separate M&F Counts: Scenarios



A) HWP and EAF. (B) Equality of inbreeding coefficients, EAF, and both sexes out of HWP. (C) Unequal inbreeding coefficients, both sexes out of equilibrium but with equal allele frequencies. (D) Both sexes in HWP but with different allele frequencies. (E) Each sex out of equilibrium with identical inbreeding coefficients and different allele frequencies. (F) Both sexes out of equilibrium, with different inbreeding coefficients and different allele frequencies.

Aside: Separate M&F Counts: Joint Exact Test

To test for both Equal Allele Frequencies (EAF) and Hardy-Weinberg Proportions (HWP):

$$\Pr(m_{Aa}, f_{Aa} | n, n_A, n_m) = \frac{n_A! n_a! n_m! n_f! 2^{m_{Aa}+f_{Aa}}}{m_{AA}! m_{Aa}! m_{aa}! f_{AA}! f_{Aa}! f_{aa}! (2n)!}$$

m_{AA}, m_{Aa}, m_{aa}	genotype counts in males
f_{AA}, f_{Aa}, f_{aa}	genotype counts in females
$n_m = m_{AA} + m_{Aa} + m_{aa}$	number of males
$n_f = f_{AA} + f_{Aa} + f_{aa}$	number of females
$n = n_m + n_f$	total sample size
$m_A = 2m_{AA} + m_{Aa}, m_a = 2m_{aa} + m_{Aa}$	numbers of A, a alleles in males
$f_A = 2f_{AA} + f_{Aa}, f_a = 2f_{aa} + f_{Aa}$	numbers of A, a alleles in females
$n_A = m_A + f_A, n_a = m_a + f_a$	total numbers of A, a alleles

Aside: Separate M&F Counts: HWP Exact Test

To test for HWP:

$$\Pr(n_{Aa}|n, n_A) = \frac{n_A! n_a! n! 2^{n_{Aa}}}{n_{AA}! n_{Aa}! n_{aa}!}$$

n_{AA}, n_{Aa}, n_{aa}	total genotype counts in males and females
$n = n_{AA} + n_{Aa} + n_{aa}$	total sample size
$n_A = 2n_{AA} + n_{Aa}, n_a = 2n_{aa} + n_{Aa}$	total numbers of A, a alleles

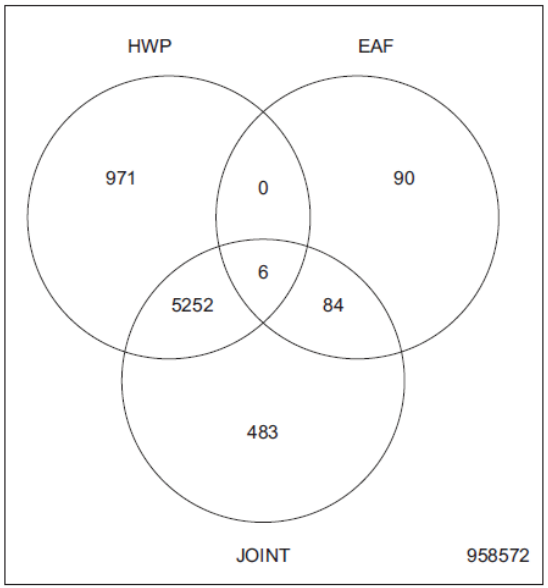
Aside: Separate M&F Counts: EAF Exact Test

To test for EAF:

$$\Pr(n_A|n, m_A) = \frac{n_A!n_a!n_m!n_f!}{m_A!m_a!f_A!f_a!}$$

m_A, m_a	numbers of A, a alleles in males
f_A, f_a	numbers of A, a alleles in females
$n_m = m_A + m_a$	total number of male alleles
$n_f = f_A + f_a$	total number of female alleles
$n_A = m_A + f_A, n_a = m_a + f_a$	total numbers of A, a alleles
$n = n_m + n_f = n_A + n_a$	total number of alleles in males and females

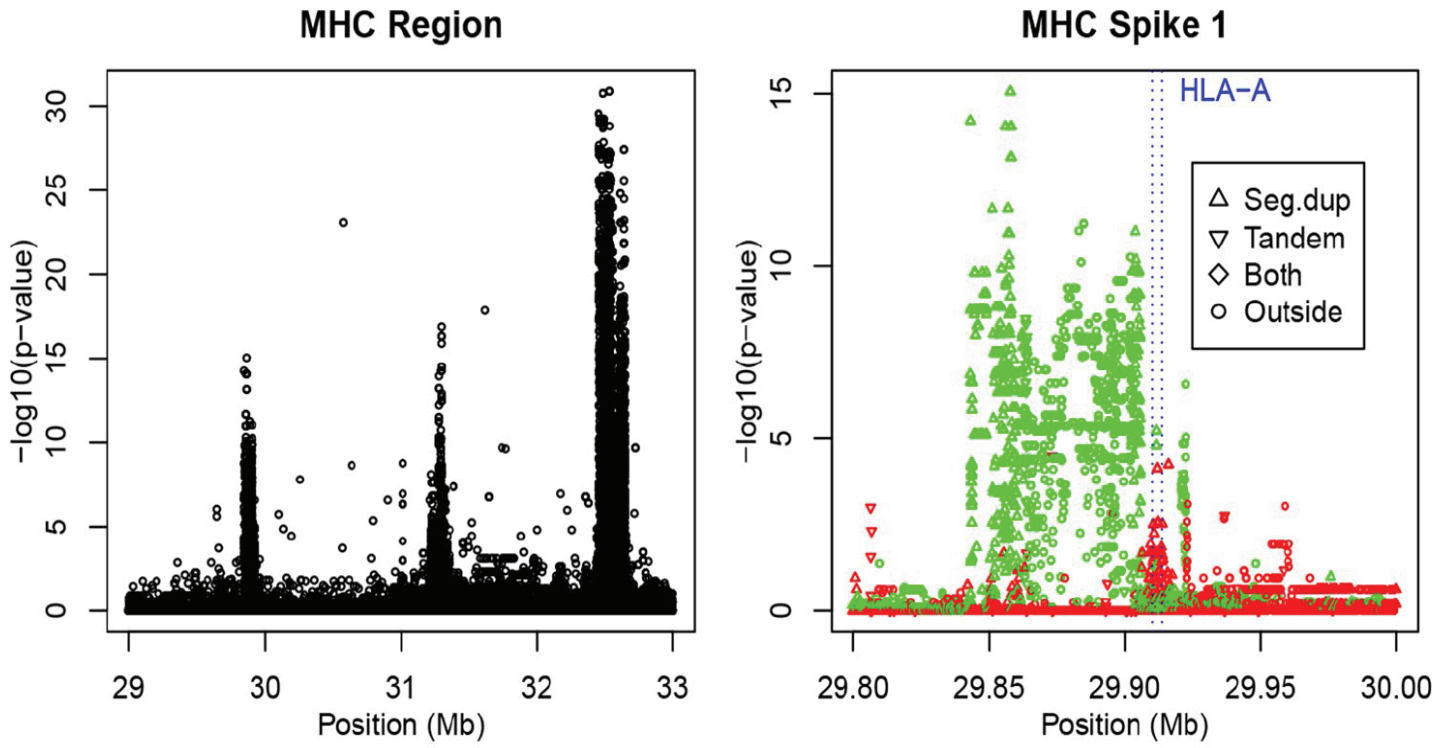
Separate M&F Counts: 1000 Genomes Result



Venn diagrams of HWP, EAF, and joint exact test results for all nonmonomorphic complete SNPs on chromosome 1 of the JPT sample Notes: Circles enclose the number of significant SNPs (at $\alpha = 0.001$) for the different tests.

Graffeleman J, Weir BS. 2018. Genetic Epidemiology 42:24-48.

MHC Region HWE Tests



Green: heterozygote deficiency. Red: heterozygote excess.

Graffelman J, Jain D, Weir B. 2017. Human Genetics 136:727-741.

Linkage Disequilibrium

Linkage Disequilibrium

This term reserved for association between pairs of alleles – one at each of two loci.

When gametic data are available, could refer to gametic disequilibrium.

When genotypic data are available, but gametes can be inferred, can make inferences about gametic and non-gametic pairs of alleles.

When genotypic data are available, but gametes cannot be inferred, can work with composite measures of disequilibrium.

Linkage Disequilibrium

For alleles A and B at two loci, the usual measure of linkage disequilibrium is

$$D_{AB} = P_{AB} - p_A p_B$$

Whether or not this is zero does not provide a direct statement about linkage between the two loci. For example, consider marker YFM and disease DTD:

		A	N	Total
YFM	+	1	24	25
	−	0	75	75
Total		1	99	100

$$D_{A+} = \frac{1}{100} - \frac{1}{100}\frac{25}{100} = 0.0075, \text{ (maximum possible value)}$$

Aside: Gametic Linkage Disequilibrium

For loci **A, B** define indicator variables x, y that take the value 1 for allele A, B and 0 for any other alleles. If gametes within individuals are indexed by j, $j = 1, 2$ then for expectations over samples from the same population

$$\begin{aligned}
\mathcal{E}(x_j) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_j^2) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j^2) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_1 x_2) = P_{AA} \;\; &, \;\; \mathcal{E}(y_1 y_2) = P_{BB} \\
\mathcal{E}(x_1 y_1) = P_{AB} \;\; &, \;\; \mathcal{E}(x_2 y_2) = P_{AB}
\end{aligned}$$

The variances of x_j, y_j are $p_A(1-p_A), p_B(1-p_B)$ for $j = 1, 2$ and the covariance and correlation coefficients for x and y are

$$\mathrm{Cov}(x_1, y_1) = \mathrm{Cov}(x_2, y_2) \;\; = \;\; P_{AB} - p_A p_B = D_{AB}$$

$$\mathrm{Corr}(x_1, y_1) = \mathrm{Corr}(x_2, y_2) \;\; = \;\; D_{AB} / \sqrt{[p_A(1-p_A)p_B(1-p_B)]} = \rho_{AB}$$

Estimation of LD

With random sampling of gametes, gamete counts have a multinomial distribution:

$$\Pr(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) = \frac{n!(P_{AB})^{n_{AB}}(P_{Ab})^{n_{Ab}}(P_{aB})^{n_{aB}}(P_{ab})^{n_{ab}}}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!}$$

The data are the counts of four gamete types, so there are three degrees of freedom. There are three parameters: p_A, p_B, D_{AB} so Bailey's method leads directly to MLE's:

$$\hat{D}_{AB} = \tilde{P}_{AB} - \tilde{p}_A\tilde{p}_B$$

$$\hat{\rho}_{AB} = r_{AB} = \frac{\hat{D}_{AB}}{\sqrt{\tilde{p}_A\tilde{p}_a\tilde{p}_B\tilde{p}_b}}$$

Testing LD

The MLE of D_{AB} is

$$\hat{D}_{AB} = \tilde{P}_{AB} - \tilde{p}_A\tilde{p}_B = \frac{1}{n^2}(n_{AB}n_{ab} - n_{Ab}n_{aB})$$

where n is the number of gametes in the sample. For large n, this estimate is normally distributed about the parametric value D_{AB}, so if $D_{AB} = 0$

$$X^2_{AB} = \frac{\hat{D}^2_{AB}}{\text{Var}(\hat{D}_{AB})} \sim \chi^2_{(1)}$$

When $D_{AB} = 0$, $\text{Var}(\hat{D}_{AB}) = p_A(1-p_A)p_B(1-p_B)/n$ and the test statistic is calculated as

$$X^2_{AB} = \frac{n\hat{D}^2_{AB}}{\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)}$$

This can be written as $X^2_{AB} = nr^2_{AB}$, by analogy to the test statistic $X^2 = n\tilde{f}^2$ for Hardy-Weinberg equilibrium.

Aside: Testing LD

Writing the MLE of D_{AB} as

$$\hat{D}_{AB} = \frac{1}{n^2}(n_{AB}n_{ab} - n_{Ab}n_{aB})$$

where n is the number of gametes in the sample, allows the use of the "Delta method" to find

$$\begin{aligned}\text{Var}(\hat{D}_{AB}) &\approx \frac{1}{n}[p_A(1-p_A)p_B(1-p_B) \\ &\quad + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2]\end{aligned}$$

When $D_{AB} = 0$, $\text{Var}(\hat{D}_{AB}) = p_A(1-p_A)p_B(1-p_B)/n$.

If $\hat{D}_{AB}$ is assumed to be normally distributed then

$$X_{AB}^2 = \frac{\hat{D}_{AB}^2}{\text{Var}(\hat{D}_{AB})} = n\hat{\rho}_{AB}^2 = nr_{AB}^2$$

is appropriate for testing $H_0 : D_{AB} = 0$. When H_0 is true, $X_{AB}^2 \sim \chi_{(1)}^2$. Note the analogy to the test statistic for Hardy-Weinberg equilibrium: $X^2 = n\hat{f}^2$.

Goodness-of-fit Test

The test statistic for the 2×2 table

n_{AB}	n_{Ab}	n_A
n_{aB}	n_{ab}	n_a
n_B	n_b	n

has the value

$$X^2 = \frac{n(n_{AB}n_{ab} - n_{Ab}n_{aB})^2}{n_A n_a n_B n_b}$$

$$= \frac{n\hat{D}_{AB}^2}{\tilde{p}_A \tilde{p}_a \tilde{p}_B \tilde{p}_b}$$

For DTD/YFM example, $X^2 = 3.03$. This is not statistically significant, even though disequilibrium was maximal.

Composite Disequilibrium

When genotypes are scored, it is often not possible to distinguish between the two double heterozygotes AB/ab and Ab/aB, so that gametic frequencies cannot be inferred.

Under the assumption of random mating, in which genotypic frequencies are assumed to be the products of gametic frequencies, it is possible to estimate gametic frequencies with the EM algorithm. To avoid making the random-mating assumption, however, it is possible to work with a set of composite disequilibrium coefficients.

Composite Disequilibrium

Although the separate digenic frequencies p_{AB} (one gamete) and $p_{A,B}$ (two gametes) cannot be observed, their sum can be since

$$\begin{aligned} p_{AB} &= P_{AB}^{AB} + \frac{1}{2}P_{Ab}^{AB} + \frac{1}{2}P_{aB}^{AB} + \frac{1}{2}P_{ab}^{AB} \\ p_{A,B} &= P_{AB}^{AB} + \frac{1}{2}P_{Ab}^{AB} + \frac{1}{2}P_{aB}^{AB} + \frac{1}{2}P_{aB}^{Ab} \\ p_{AB} + p_{A,B} &= 2P_{AB}^{AB} + P_{Ab}^{AB} + P_{aB}^{AB} + \frac{P_{ab}^{AB} + P_{aB}^{Ab}}{2} \end{aligned}$$

Allele-pair disequilibrium can be measured with a composite measure Δ_{AB} defined as

$$\begin{aligned} \Delta_{AB} &= p_{AB} + p_{A,B} - 2p_A p_B \\ &= D_{AB} + D_{A,B} \end{aligned}$$

which is the sum of the gametic ($D_{AB} = p_{AB} - p_A p_B$) and nongametic ($D_{A,B} = p_{A,B} - p_A p_B$) coefficients.

Composite Disequilibrium

If the counts of the nine genotypic classes are

	BB	Bb	bb
AA	n_1	n_2	n_3
Aa	n_4	n_5	n_6
aa	n_7	n_8	n_9

the count for pairs of alleles in an individual being A and B, whether received from the same or different parents, is

$$n_{AB} = 2n_1 + n_2 + n_4 + \frac{1}{2}n_5$$

and the MLE for Δ is

$$\hat{\Delta}_{AB} = \frac{1}{n}n_{AB} - 2\tilde{p}_A\tilde{p}_B$$

Composite LD and Allele Dosage

The allele dosage for a SNP is the number of copies of the (say) the reference allele carried by an individual. If A is the reference allele for SNP **A**, then genotypes AA, Aa, aa have dosages X_A of 2,1,0.

The covariance of allele dosages X_A, X_B for loci **A, B** is

$$\text{Cov}(X_A, X_B) \;=\; 2\Delta_{AB}$$

By analogy to the tests for within-population inbreeding and for gametic linkage disequilibrium, a test statistic for composite LD is

$$X^2_{AB_c} \;=\; nr^2_{AB_c}$$

where r_{AB_c} is the sample correlation coefficient for allele dosages at the two loci over the n individuals in a sample.

Example

A sample of size 15 has these two-locus genotypes and allele dosages:

		X_A	X_A^2	X_B	X_B^2	X_AX_B
1	$AAbb$	2	4	0	0	0
2	$AAbb$	2	4	0	0	0
3	$AaBB$	1	1	2	4	2
4	$AaBb$	1	1	1	1	1
5	$AaBb$	1	1	1	1	1
6	$AaBb$	1	1	1	1	1
7	$Aabb$	1	1	0	0	0
8	$Aabb$	1	1	0	0	0
9	$Aabb$	1	1	0	0	0
10	$Aabb$	1	1	0	0	0
11	$aaBb$	0	0	1	1	0
12	$aabb$	0	0	0	0	0
13	$aabb$	0	0	0	0	0
14	$aabb$	0	0	0	0	0
15	$aabb$	0	0	0	0	0
Sum		$S_A = 12$	$S_{AA} = 16$	$S_B = 6$	$S_{BB} = 8$	$S_{AB} = 5$

Example (contd.)

The sample means, variances, covariance and correlation of dosages X_A, X_B are:

means: $\bar{X}_A = S_A/n = 12/15; \bar{X}_B = S_B/n = 6/15$

variances: $s_A^2 = (S_{AA} - S_A^2/n)/(n-1) = (16 - 144/15)/14$;
$s_B^2 = (S_{BB} - S_B^2/n)/(n-1) = (8 - 36/15)/14$

covariance: $s_{AB} = (S_{AB} - S_A S_B/n)/(n-1) = (5 - 72/15)/14$

correlation: $r_{AB_c}^2 = s_{AB}^2/s_A^2 s_B^2 = 1/(32 * 28)$

test statistic: $X_{AB_c}^2 = n r_{AB_c}^2 = 0.0168$

The hypothesis of no *composite* LD is not rejected. If there is HWE this is the same as testing for LD.

Aside: Composite Linkage Disequilibrium

For loci **A, B** define indicator variables x, y that take the value 1 for allele A, B and 0 for any other alleles. If gametes within individuals are indexed by j, $j = 1, 2$ then for expectations over samples from the same population

$$\begin{aligned}
\mathcal{E}(x_j) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_j^2) = p_A, \;\; j = 1,2 \;\; &, \;\; \mathcal{E}(y_j) = p_B \;\; j = 1,2 \\
\mathcal{E}(x_1 x_2) = P_{AA} \;\; &, \;\; \mathcal{E}(y_1 y_2) = P_{BB} \\
\mathcal{E}(x_1 y_1) = P_{AB} \;\; &, \;\; \mathcal{E}(x_2 y_2) = P_{AB} \\
\mathcal{E}(x_1 y_2) = P_{A,B} \;\; &, \;\; \mathcal{E}(x_2 y_1) = P_{A,B}
\end{aligned}$$

Write

$$\begin{aligned}
D_A = P_{AA} - p_A^2 \;\; &, \;\; D_B = P_{BB} - p_B^2 \\
D_{AB} = P_{AB} - p_A p_B \;\; &, \;\; D_{A,B} = P_{A,B} - p_A p_B \\
\Delta_{AB} = D_{AB} + D_{A,B} &
\end{aligned}$$

Aside: Composite LD and Allele Dosage

Now set $X = x_1 + x_2, Y = y_1 + y_2$, the allelic dosages at each locus, to get

$$\mathcal{E}(X) = 2p_A \quad , \quad \mathcal{E}(Y) = 2p_B$$
$$\mathcal{E}(X^2) = 2(p_A + P_{AA}) \quad , \quad \mathcal{E}(Y^2) = 2(p_B + P_{BB})$$
$$\mathrm{Var}(X) = 2p_A(1-p_A)(1+f_A) \quad , \quad \mathrm{Var}(Y) = 2p_B(1-p_B)(1+f_B)$$

and

$$\begin{aligned}
\mathcal{E}(XY) &= 2(P_{AB} + P_{A,B}) \\
\mathrm{Cov}(X,Y) &= 2(P_{AB} - p_A p_B) + 2(P_{A,B} - p_A p_B) \\
&= 2(D_{AB} + D_{A,B}) = 2\Delta_{AB} \\
\mathrm{Corr}(X,Y) &= \frac{\Delta_{AB}}{\sqrt{p_A(1-p_A)(1+f_A)p_B(1-p_B)(1+f_B)}}
\end{aligned}$$

Aside: Composite Linkage Disequilibrium Test

$$\hat{\Delta}_{AB} = n_{AB}/n - 2\tilde{p}_A\tilde{p}_B$$

where

$$n_{AB} = 2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb}$$

This does not require phased data.

By analogy to the gametic linkage disequilibrium result, a test statistic for $\Delta_{AB} = 0$ is

$$X^2_{AB} = \frac{n\hat{\Delta}^2_{AB}}{\tilde{p}_A(1-\tilde{p}_A)(1+\hat{f}_A)\tilde{p}_B(1-\tilde{p}_B)(1+\hat{f}_B)}$$

This is assumed to be approximately $\chi^2_{(1)}$ under the null hypothesis. The approximation rests on ignoring disequilibria between three and four alleles of the two **A** and two **B** alleles.

Aside: Example

For the data shown on earlier:

	BB	Bb	bb	Total
AA	$n_{AABB} = 0$	$n_{AABb} = 0$	$n_{AAbb} = 2$	$n_{AA} = 2$
Aa	$n_{AaBB} = 1$	$n_{AaBb} = 3$	$n_{Aabb} = 4$	$n_{Aa} = 8$
aa	$n_{aaBB} = 0$	$n_{aaBb} = 1$	$n_{aabb} = 4$	$n_{aa} = 5$
Total	$n_{BB} = 1$	$n_{Bb} = 4$	$n_{bb} = 10$	$n = 15$

$$
\begin{aligned}
n_{AB} &= 2 \times 0 + 0 + 1 + \frac{1}{2}(3) = 2.5 \\
n_A &= 12, \;\; \tilde{p}_A = 0.4 \\
n_B &= 6, \;\; \tilde{p}_B = 0.2 \\
\widehat{f}_A &= 1 - \frac{8/15}{0.48} = -0.11 \\
\widehat{f}_B &= 1 - \frac{4/15}{0.32} = 0.17
\end{aligned}
$$

Aside: Example

The estimated composite disequilibrium coefficient is

$$\hat{\Delta}_{AB} = \frac{2.5}{15} - 2(0.4)(0.2) = 0.0067$$

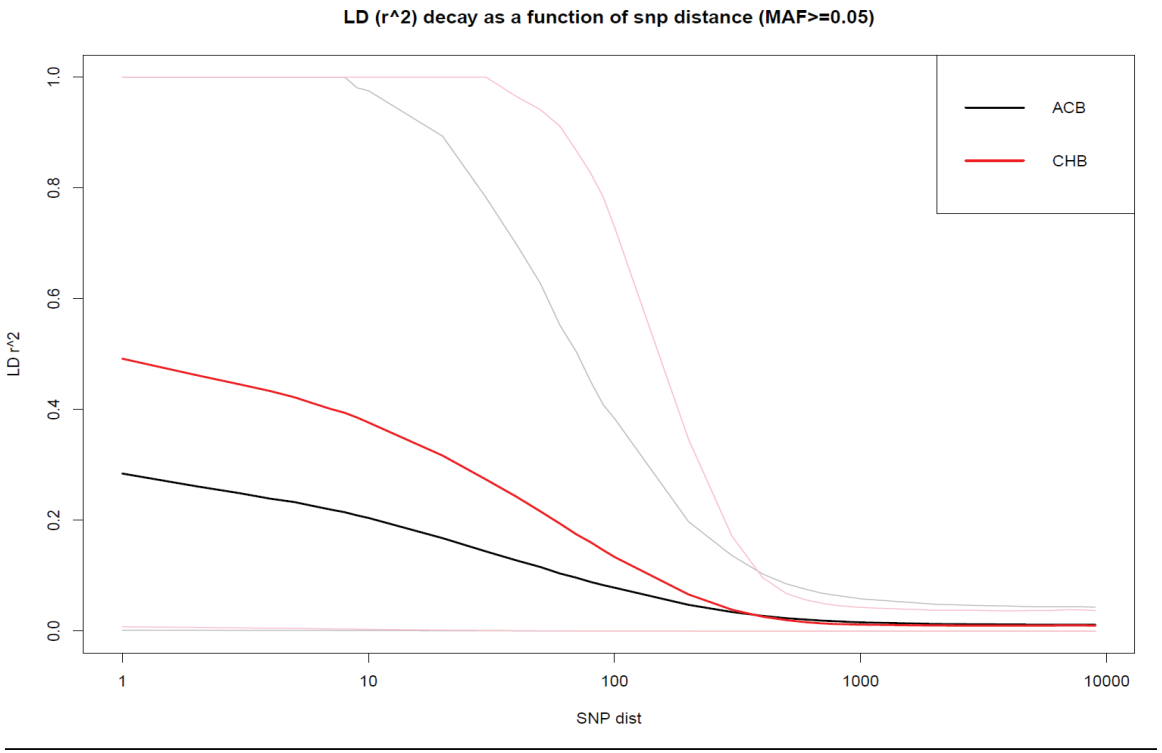
The test statistic is

$$X^2 = \frac{15 \times (0.0067)^2}{0.24 \times 0.89 \times 0.16 \times 1.17} = 0.02$$

Previous work on EM algorithm, assuming HWE, estimated p_{AB} as 0.0893 so

$$\begin{aligned} \hat{D}_{AB} &= 0.0893 - 0.4 \times 0.2 = 0.0093 \\ X^2 &= \frac{30 \times (0.0093)^2}{0.4 \times 0.6 \times 0.2 \times 0.8} = 0.07 \end{aligned}$$

1000 Genomes Example



Allele dosage squared correlations for pairs of SNPs on chromosomes 21 and 22 of the 1000 Genomes ACB and populations. Heavy lines: means. Light lines: 5th and 95th percentiles.

Aside: Single-locus Entropy

It is difficult to describe associations among alleles at several loci. One approach is based on information theory.

For a locus with sample frequencies $\tilde{p}_u$ for alleles A_u the entropy is

$$H_A = -\sum_u \tilde{p}_u \ln(\tilde{p}_u)$$

If there is only allele in the sample, one $\tilde{p}$ is 1 with $n(\tilde{p}) = 0$, and the rest are zero. The entropy is zero.

If there are m equally-frequent alleles, $\tilde{p} = 1/m$ for all alleles and the entropy is maximized at $\ln(m)$.

Aside: Multi-locus Entropy

For two loci with alleles A_u, B_v, the entropy is

$$H_{AB} = -\sum_u \sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv})$$

In the absence of linkage disequilibrium $\tilde{P}_{uv} = \tilde{p}_u \tilde{p}_v$ so

$$\begin{aligned} H_{AB} &= -\sum_u \sum_v \tilde{p}_u \tilde{p}_v [\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] \\ &= H_A + H_B \end{aligned}$$

so if $H_{AB} \neq H_A + H_B$ there is evidence of dependence. This extends to multiple loci.

Aside: Conditional Entropy

If the entropy for a multi-locus profile A is H_A then the conditional probability of another locus B, given A, is $H_{B|A} = H_{AB} - H_A$.

In performing meaningful calculations for Y-STR profiles, this suggests choosing a set of loci by an iterative procedure. First choose locus L_1 with the highest entropy. Then choose locus L_2 with the largest conditional entropy $H(L_2|L_1)$. Then choose L_3 with the highest conditional entropy with the haplotype L_1L_2, and so on.

Aside: Conditional Entropy for Y-STR Data

Added Marker	Entropy		
	Single	Multi	Cond.
DYS385ab	4.750	4.750	4.750
DYS481	2.962	6.972	2.222
DYS570	2.554	8.447	1.474
DYS576	2.493	9.318	0.871
DYS458	2.220	9.741	0.423
DYS389II	2.329	9.906	0.165
DYS549	1.719	9.999	0.093
DYS635	2.136	10.05	0.053
DYS19	2.112	10.08	0.028
DYS439	1.637	10.10	0.024
DYS533	1.433	10.11	0.010
DYS456	1.691	10.12	0.006
GATAH4	1.512	10.12	0.005
DYS393	1.654	10.13	0.003
DYS448	1.858	10.13	0.002
DYS643	2.456	10.13	0.002
DYS390	1.844	10.13	0.002
DYS391	1.058	10.13	0.002

Most-discriminating loci may not contribute to the most-discriminating haplotypes. No additional discriminating power beyond 10 loci.