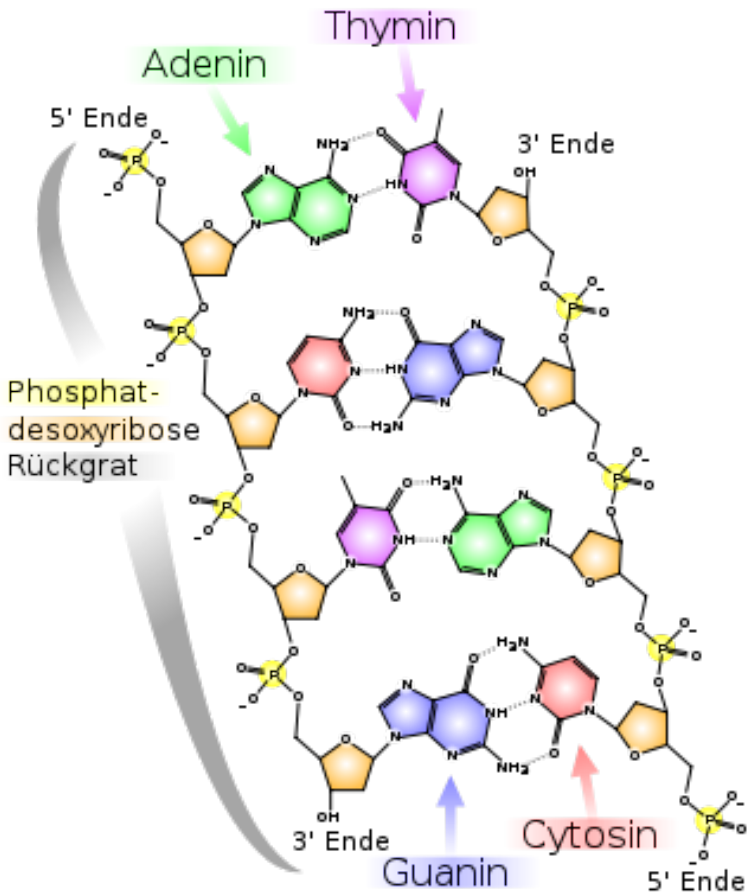
Haindorf Seminar 2012
An effective number of tests in genetic association studies

Jens Stange
Thorsten Dickhaus

10. Februar 2012

DNA

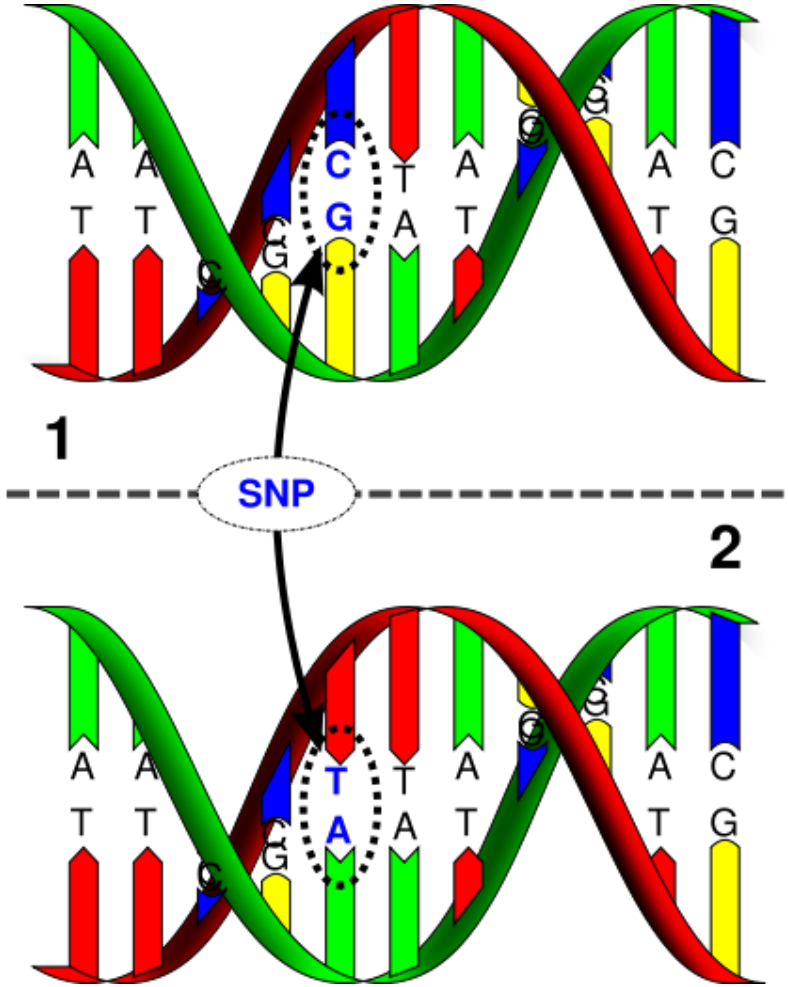
- the entire genetic information is encoded in a sequence of nucleic bases (more than 3 billion bp)
- between individuals there are variations in the sequence on specific loci, one variant is called allele
- human body cells are diploid, so every individual carries two alleles
- if both alleles on one locus are the same, the genotype on this locus is termed homozygot, otherwise heterozygot



What are SNPs?

- SNP=single nucleotide polymorphism
- about 15Mio. SNP-loci refer to 90% of variation in DNA
- mostly biallelic=two possible nucleic bases at one SNP-locus
- modern sequencing-technologies allow to determine the genotypes of SNP-loci across the whole DNA (=genome-wide)

$\rightsquigarrow$ SNPs are suitable indicators for genetic association



genetic association studies

Search for candidate positions on the DNA, which may cause a specific phenotype (eg. risk preference).

genetic association studies

Search for candidate positions on the DNA, which may cause a specific phenotype (eg. risk preference).

The participants of such a study are genotyped at M SNP-loci.
Let the SNPs be enumerated by $\{1, \ldots, M\}$

For each SNP $m \in \{1, \ldots, M\}$ state the hypothesis:

$$H_m = \{\text{“SNP m is not associated with the phenotype“}\}$$

There are several possibilities to test these hypotheses, which may depend on specific genetic/inheritance models.

One way is to compare the distribution of the alleles/genotypes within a target population with the distribution of the alleles/genotypes in the whole population, which is done in case-control studies.

If there is a significant difference between the distributions, we can reject the hypothesis and have found a candidate.

On any locus denote the two possible alleles „0“ and „1“
three possible genotypes “00”, “01”, “11”.
The counts for each SNP may be aggregated in contingency tables:

(i) the allele frequencies in a 2×2-table:

	0	1	$\sum$
cases	ν_{10}	ν_{11}	$2n_{1\cdot}$
controls	ν_{20}	ν_{21}	$2n_{2\cdot}$
$\sum$	$\nu_{\cdot 0}$	$\nu_{\cdot 1}$	$2N$

marginal frequencies: $p_0 = \frac{\nu_{\cdot 0}}{2N}$, $p_1 = \frac{\nu_{\cdot 1}}{2N}$

(ii) the genotype frequencies in a 2×3-table:

	00	01	11	$\sum$
cases	n_{10}	n_{11}	n_{12}	$n_{1\cdot}$
controls	n_{20}	n_{21}	n_{22}	$n_{2\cdot}$
$\sum$	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot 2}$	N

marginal frequencies: $p_0 = \frac{n_{\cdot 0}}{N}$, $p_1 = \frac{n_{\cdot 1}}{N}$, $p_2 = \frac{n_{\cdot 2}}{N}$

Pearson's χ^2-test for goodness of fit

We have the following test statistics for each SNP m

(i) for the alleles

$$T^{(m)} = \frac{(\nu_{10}^{(m)} - 2n_{1\cdot}p_0^{(m)})^2}{2n_{1\cdot}p_0^{(m)}} + \frac{(\nu_{11}^{(m)} - 2n_{1\cdot}p_1^{(m)})^2}{2n_{1\cdot}p_1^{(m)}}$$

(ii) for the genotypes

$$T^{(m)} = \frac{(n_{10}^{(m)} - n_{1\cdot}p_0^{(m)})^2}{n_{1\cdot}p_0^{(m)}} + \frac{(n_{11}^{(m)} - n_{1\cdot}p_1^{(m)})^2}{n_{1\cdot}p_1^{(m)}} + \frac{(n_{12}^{(m)} - n_{1\cdot}p_2^{(m)})^2}{n_{1\cdot}p_2^{(m)}}$$

If the sample size is sufficiently large the test at significance level α is given by:

$$\varphi_m = \begin{cases} 0 & \text{if } T^{(m)} \leq c_\alpha \\ 1 & \text{if } T^{(m)} > c_\alpha \end{cases}$$

where c_α is $(1-\alpha)$-quantile from χ^2-distribution with

(i) one degree of freedom in the allelic case .

(ii) two degrees of freedom in the genotypic case.

Pearson's χ^2-test for goodness of fit

We have the following test statistics for each SNP m

(i) for the alleles

$$T^{(m)} = \frac{(\nu_{10}^{(m)} - 2n_{1\cdot}p_0^{(m)})^2}{2n_{1\cdot}p_0^{(m)}} + \frac{(\nu_{11}^{(m)} - 2n_{1\cdot}p_1^{(m)})^2}{2n_{1\cdot}p_1^{(m)}}$$

(ii) for the genotypes

$$T^{(m)} = \frac{(n_{10}^{(m)} - n_{1\cdot}p_0^{(m)})^2}{n_{1\cdot}p_0^{(m)}} + \frac{(n_{11}^{(m)} - n_{1\cdot}p_1^{(m)})^2}{n_{1\cdot}p_1^{(m)}} + \frac{(n_{12}^{(m)} - n_{1\cdot}p_2^{(m)})^2}{n_{1\cdot}p_2^{(m)}}$$

If the sample size is sufficiently large the test at significance level α is given by:

$$\varphi_m = \begin{cases} 0 & \text{if } T^{(m)} \leq c_\alpha \\ 1 & \text{if } T^{(m)} > c_\alpha \end{cases}$$

where c_α is $(1-\alpha)$-quantile from χ^2-distribution with

(i) one degree of freedom in the allelic case .

(ii) two degrees of freedom in the genotypic case.

$\rightsquigarrow$ Question: How to choose the significance level α?

multiple testing problem

Let $(\mathcal{X}, \mathcal{F}, \mathbb{P}_{\vartheta \in \Theta})$ be a statistical experiment.
Let $\varphi = (\varphi_1, \ldots, \varphi_M)$ be a multiple test for a set of hyptheses $H_1, \ldots, H_M \subset \Theta$.
Every single hypothesis H_m is simultaneously tested at a local significance level α_{loc} against the alternative $K_m = \Theta \setminus H_m$, i.e:

$$\forall m = 1, \ldots, M : \vartheta \in H_m \Rightarrow \mathbb{P}_\vartheta(\varphi_m = 1) \leq \alpha_{loc}$$

A classical type-I-error quantity for multiple tests is the probablity that at least one of the hypotheses is falsely rejected, the so called

familywise error rate

$$\mathrm{FWER}_\vartheta(\varphi) := \mathbb{P}_\vartheta\Big(\bigcup_{I_0(\vartheta)} \{\varphi_m = 1\}\Big) \quad \vartheta \in \Theta, \; I_0(\vartheta) = \{m \in \{1, \ldots, M\} \mid \vartheta \in H_m\}$$

A multiple test φ is termed to control the FWER at a global significance level α_G iff.

$$\sup_{\vartheta \in \Theta} \mathrm{FWER}_\vartheta(\varphi) \leq \alpha_G$$

Typically one sets the global significance level α_G (e.g.: 0.05) to derive the local significance level.

For example:

a) Bonferroni-correction:

$$\alpha_{loc} = \frac{\alpha_G}{M}$$

since:

$$\mathbb{P}_\vartheta\left(\bigcup_{I_0(\vartheta)}\{\varphi_m = 1\}\right) \leq \mathbb{P}_\vartheta\left(\bigcup_{m=1}^{M}\{\varphi_m = 1\}\right) \leq \sum_{m=1}^{M}\mathbb{P}_\vartheta(\{\varphi_m = 1\}) \leq M\alpha_{loc} = \alpha_G$$

b) Šidák-correction:

Suppose all the tests $(\varphi_1, \ldots, \varphi_M)$ stochastically independent:

$$\alpha_{loc} = 1 - (1 - \alpha_G)^{\frac{1}{M}}$$

since:

$$\mathbb{P}_\vartheta\left(\bigcup_{I_0(\vartheta)}\{\varphi_m = 1\}\right) \leq \mathbb{P}_\vartheta\left(\bigcup_{m=1}^{M}\{\varphi_m = 1\}\right) = 1 - \mathbb{P}_\vartheta\left(\bigcap_{m=1}^{M}\{\varphi_m = 0\}\right)$$

$$= 1 - \prod_{m=1}^{M}\mathbb{P}_\vartheta(\{\varphi_m = 0\}) \leq 1 - (1 - \alpha_{loc})^M = \alpha_G$$

Motivation for an effective number of tests:

Imagine we had the following situation:
The tests $(\varphi_1, \ldots, \varphi_M)$ can be partitioned in M_0 classes $C_1, \ldots, C_{M_0}$, such that all tests within one class C_j are equivalent.
If additionally the classes were stochastically independent, the correction

$$\alpha_{loc} = 1 - (1 - \alpha_G)^{\frac{1}{M_0}}$$

would be valid and M_0 can be interpreted as an effective number of tests.

Indeed there is a natural correlation structure between the SNPs, the so called Linkage Disequilibrium (LD).
One can use the LD to estimate an effective number of tests.

Let $\varphi = (\varphi_1, \ldots, \varphi_M)$ the χ^2-tests, and denote

$$O_m = \{\varphi_m = 0\} = \{T^{(m)} \leq c_\alpha\}$$

the event that the result of the test statistic for SNP m is non-significant. The test statistics $(T^{(m)})_{m=1,\ldots,M}$ are positively dependent, so it holds for any $\vartheta \in \Theta$:

$$\mathbb{P}_\vartheta(O_1 \cap \ldots \cap O_j) \geq \prod_{i=1}^{j} \mathbb{P}_\vartheta(O_i) \quad j = 2, \ldots, M$$

moreover we have $\forall\ j = 2, \ldots, M,\ \forall\ k < j$:

$$\mathbb{P}_\vartheta(O_j | O_{j-1} \cap \ldots \cap O_1) \geq \mathbb{P}_\vartheta(O_j | O_k)$$

$$\mathrm{FWER}_\vartheta \leq 1 - \mathbb{P}_\vartheta(O_1 \cap \ldots \cap O_M)$$

$$= 1 - \mathbb{P}_\vartheta(O_1)\mathbb{P}_\vartheta(O_2|O_1)\mathbb{P}_\vartheta(O_3|O_2 \cap O_1)\ldots\mathbb{P}_\vartheta(O_M|O_{M-1} \cap \ldots \cap O_1)$$

$$\leq 1 - \mathbb{P}_\vartheta(O_1)\prod_{j=2}^{M}\max_{k<j}\mathbb{P}_\vartheta(O_j|O_k)$$

Let $\zeta_j = \max_{k<j}\mathbb{P}_\vartheta(O_j|O_k)$
by putting:

$$1-(1-\alpha_{loc})^{M_{eff}} = 1-(1-\alpha_{loc})\prod_{j=2}^{M}\zeta_j \iff M_{eff} = 1+\sum_{j=2}^{M}\frac{\log(\zeta_j)}{\log(1-\alpha_{loc})}$$

gives an estimate for an effective number of tests.

Moskvina/Schmidt (2008):
in the allelic situation with χ^2-tests on 2×2-tables:

$$M_{eff} = 1 + \sum_{j=2}^{M} \kappa_j$$

with

$$\kappa_j = \frac{1}{\log(1-\alpha_{loc})} \log\left(1 - \frac{1}{1-\alpha_{loc}} \sqrt{\frac{2}{\pi}} \int_{-q_{\alpha_{loc}}}^{q_{\alpha_{loc}}} e^{-\frac{x^2}{2}} \Phi\left(\frac{r_j x - q_{\alpha_{loc}}}{\sqrt{1-r_j^2}}\right) dx\right)$$

where

- $r_j := \max_{k<j} |r_{jk}|$, r_{jk} the haplotypic Pearsons correlation coefficient of two SNPs j, k,
  a well-known LD-measure,
  available from databases
- $q_{\alpha_{loc}}$ $(1 - \frac{\alpha_{loc}}{2})$-quantile, Φ c.d.f of $\mathcal{N}(0,1)$

numerical approximation for $\alpha_{loc} \leq 0.01$:

$$\kappa_j \approx \sqrt{1 - r_j^{-1.31 \log_{10}(\alpha_{loc})}}$$

They show that the test statistic for each SNP m, can be written as:

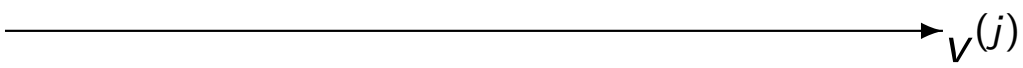
$$T^{(m)} = (\langle X, v^{(m)} \rangle)^2$$

where X is a Gaussian random vector on a hyperplane and $v^{(m)}$ is a unit vector, which may be interpreted as the test direction.

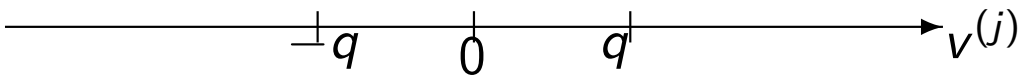
The scalar product of two directions $v^{(j)}$, $v^{(k)}$ gives exactly the haplotypic Pearson's correlation coefficient.

Geometrically this can be seen as the cosine of an angle δ between the two test directions.

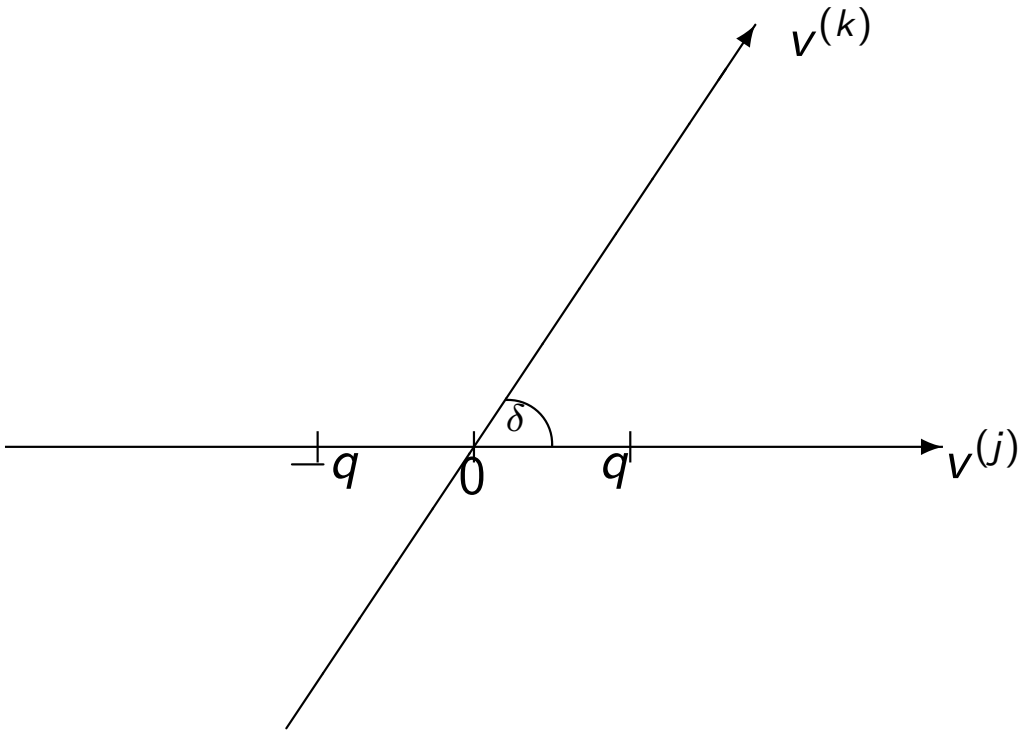
in picture:



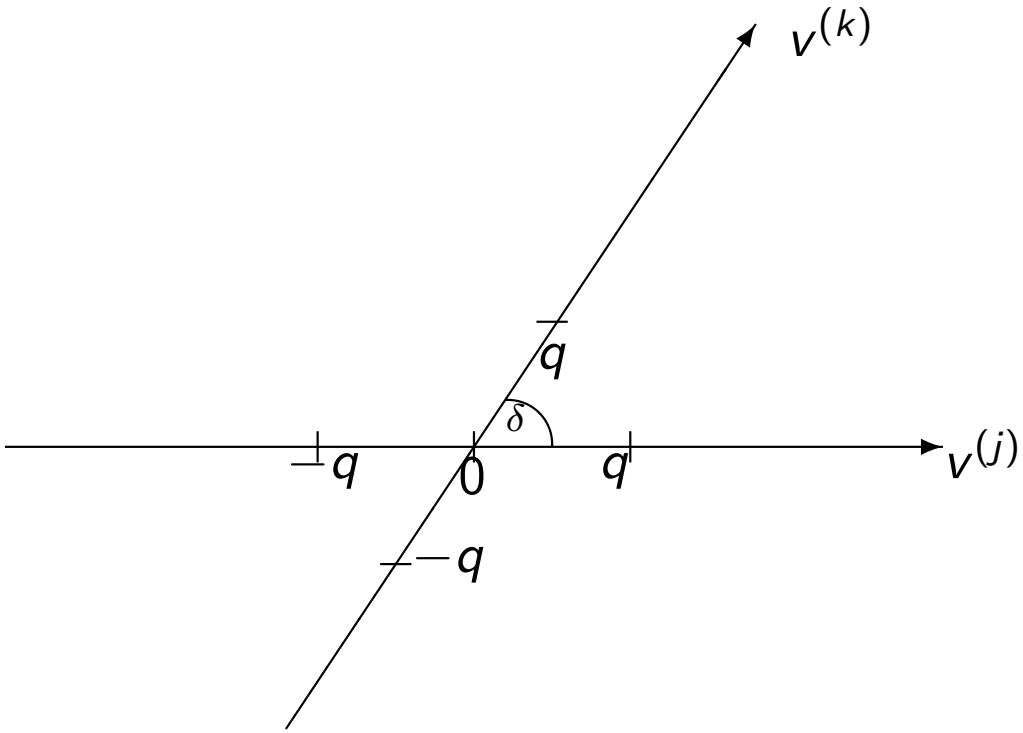
in picture:



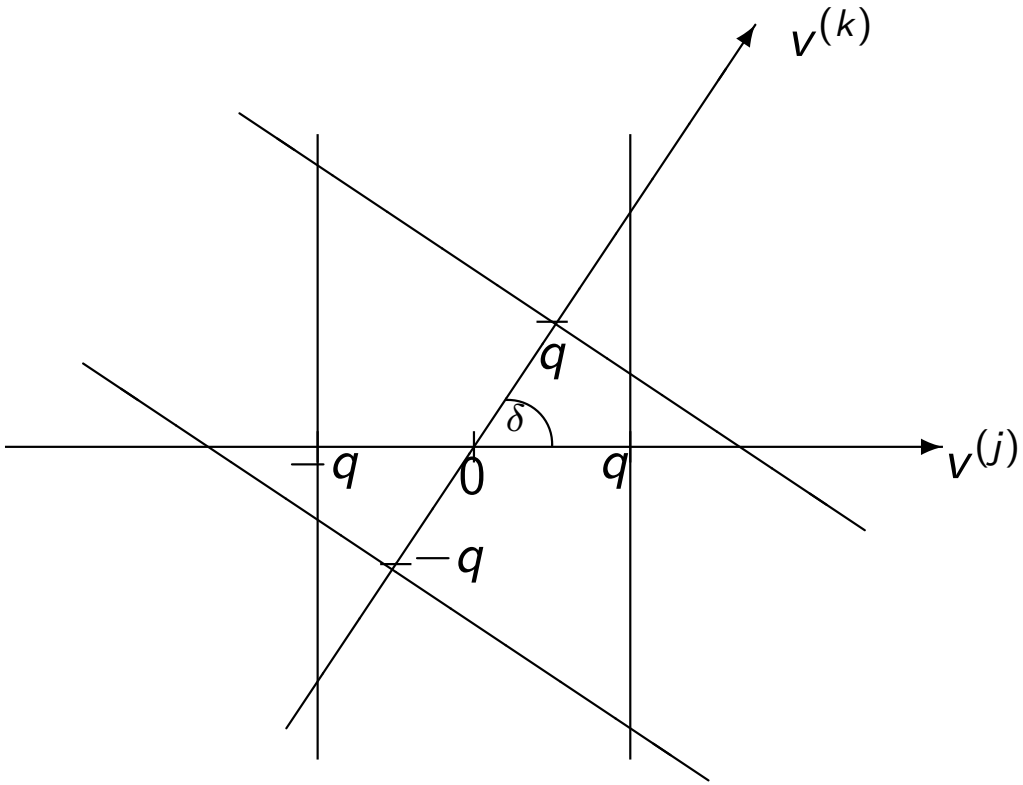
in picture:



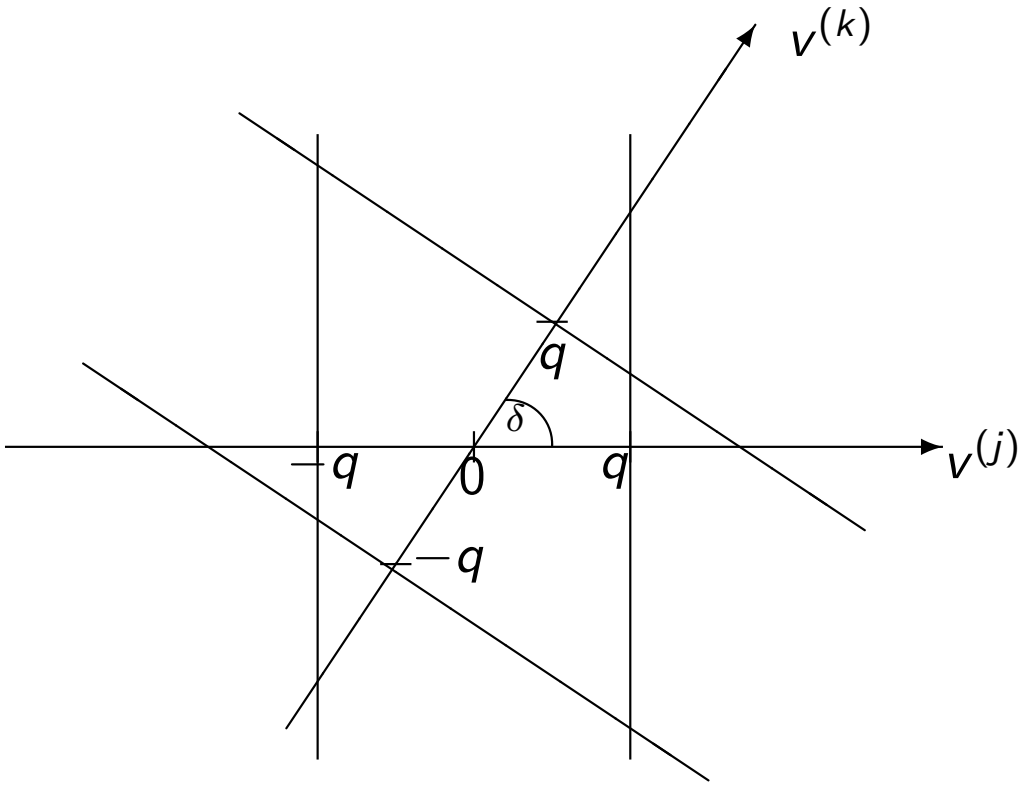
in picture:



in picture:



in picture:



Integration of standard normal distribution over the parellepiped yields:

$$\mathbb{P}_\vartheta(O_j \cap O_k) = 1 - \alpha_{loc} - \sqrt{\frac{2}{\pi}} \int_{-q_{\alpha_{loc}}}^{q_{\alpha_{loc}}} e^{-\frac{x^2}{2}} \Phi\left(\frac{r_{jk}x - q_{\alpha_{loc}}}{\sqrt{1 - r_{jk}^2}}\right) dx$$

Does it work?

In their publication some results were provided:
For a case-control-study 1,868 cases and 2,938 controls were genotyped at $M = 459,446$ SNP-loci.

They computed an effective number of tests:

$$M_{eff} = 298,518.6$$

For comparison they performed a permutation test, which yields an empirical estimate for the *FWER* for given local significance level $\alpha_{loc} (\approx 10^{-6})$

$$M_{eff,perm} = 276,666$$

Own contribution: Extension to the genotypic situation

Here the test statistic for a SNP m, may be represented as

$$T^{(m)} = (\langle X, v_1^{(m)} \rangle)^2 + (\langle X, v_2^{(m)} \rangle)^2$$

where X is standard normal distributed on a hyperplane $\mathcal{H}$,
$v_1^{(m)}, v_2^{(m)}$ are mutually orthogonal unit-vectors in $\mathcal{H}$.
The event $O_j \cap O_k$, that the results of two test statistics $T^{(j)}, T^{(k)}$ are not significant can be written as:

$$\{x \in \mathcal{H} | \langle x, v_1^{(j)} \rangle^2 + \langle x, v_2^{(j)} \rangle^2 \leq c_\alpha, \langle x, v_1^{(k)} \rangle^2 + \langle x, v_2^{(k)} \rangle^2 \leq c_\alpha\}$$

One could imagine an intersection of two cylinders on the hyperplane $\mathcal{H}$.
Approximation of $\mathbb{P}_\vartheta(O_j \cap O_k)$ will be done by MonteCarlo-Integration.