

SHORT REPORT

Empirical assessment of the validity of the 'fundamental theorem of the HapMap' in the light of 'cryptic' tagging of multiple susceptibility loci

Zoltán Bochdanovits^{*,1,2}, Peter Heutink² and Aad van der Vaart³

¹Section Medical Genomics, Department of Clinical Genetics, VUMC, Amsterdam, The Netherlands; ²Center for Neurogenomics and Cognitive Research, VU/VUMC, Amsterdam, The Netherlands; ³Section Stochastics, Department of Mathematics, Faculty of Sciences, Vrije Universiteit, Amsterdam, The Netherlands

Underestimation of the sample size needed to detect genetic association may occur as a result of deviations from the 'fundamental theorem of the HapMap'. A biologically plausible mechanism that might cause this deviation is 'cryptic' tagging of multiple susceptibility loci by the same neutral marker. For complex disorders, the existence of multiple susceptibility loci on the same chromosome is probably the rule rather than the exception. Our results show that conditional on the known haplotype structure of the genome the probability that a tagging SNP that is in linkage disequilibrium (LD) with a susceptibility gene is also in LD with another susceptibility gene is not negligible. Consequently, we were able to estimate the extent and the prevalence of the bias in the necessary sample size to find association induced by 'cryptic' tagging. In general, the underestimation of the necessary sample size is modest: 5% of all association studies will underestimate the sample size by 5–30%. On the basis of our results, a safe bet is to use a sample that is 10% larger than otherwise deemed necessary.

European Journal of Human Genetics (2008) 16, 525–529; doi:10.1038/sj.ejhg.5201984; published online 16 January 2008

Keywords: HapMap; association study; linkage disequilibrium mapping; cryptic tagging

Introduction

Genetic association studies rely by design on the presence of linkage disequilibrium (LD) between the yet unknown susceptibility locus and the neutral markers that have been genotyped. LD describes the nonindependence of alleles segregating at two or more loci. The sample size necessary to detect the susceptibility locus by studying a nearby neutral marker is supposed to be inflated by the inverse $1/r^2$ of the LD r^2 between them. For instance, if r^2 between marker and susceptibility locus is 0.5, then the sample size needed to achieve the same power would be twice as large.¹

This concept of genetic association has recently been named 'the Fundamental Theorem of the HapMap' and received criticism based on theoretical grounds.² The main issue raised is that r^2 is not necessarily multiplicative across multiple loci. It is commonly assumed that given three subsequent loci A , B and C , LD between A and C is the product of LD between AB and BC : $r_{AC} = r_{AB} \times r_{BC}$. Terwilliger and Hiekkalinna (T&H) correctly pointed out that this relationship does not need to hold.

As has been suggested before, genetic heterogeneity is a plausible biological mechanism for the lack of multiplicity.³ If the genotyped marker A is in LD with two different susceptibility loci (B and C), which cause disease D , then the multiplicativity of r^2 will not hold ($r_{AD} = r_{AB} \times r_{BD}$ and $r_{AD} = r_{AC} \times r_{CD}$). Hence, an important issue with respect to the feasibility of LD-based gene mapping in the light of the critique by T&H is to determine

*Correspondence: Dr Z Bochdanovits, VU Medical Center, Van der Boechorstraat 7, 1081 BT Amsterdam, Amsterdam, The Netherlands. Tel: +31205989280; Fax: +31204448285;

E-mail: z.bochdanovits@vumc.nl

Received 20 July 2007; revised 13 November 2007; accepted 20 November 2007; published online 16 January 2008

how often we might expect that any given neutral marker that truly tags a susceptibility locus also tags another. T&H show that underestimation of the necessary sample size for an association study as a result of deviation from the multiplicativity of r^2 is possible. Here, we aim to assess from the HapMap data itself how probable this deviation is, assuming that ‘cryptic’ tagging of multiple susceptibility loci is the predominant biological cause of this effect and an additive model for penetrance.

Theory

Testing for association is equivalent to testing if the correlation between the tagging SNP (T) and disease status (Ca) is equal to 0. Under the standard assumption that T is independent of Ca conditional on a disease locus D with one ‘healthy’ and susceptibility allele, the correlation between T and Ca is given by the familiar expression

$$r_{T,Ca} = r_{T,D}r_{D,Ca} \tag{1}$$

Here, we explore a biologically plausible scenario that could cause a deviation from this ‘multiplicativity of the correlation’: the situation where multiple correlated disease loci affect the phenotypic outcome.

Assume that there are $k \geq 1$ disease loci, and let D denote a haplotype at these loci. Assume that all disease loci are biallelic, with one variant a disease allele and the other a healthy allele, and let $\#D$ denote the number of loci among the k loci in D that carry the disease variant. Assume that, for some numbers α and β ,

$$p_{Ca|D} \approx \alpha + (\#D)\beta \tag{2}$$

This corresponds to an ‘additive model’ in which every disease allele adds an amount β to the penetrance. According to this formula, the prevalence is

$$p_{Ca} = \sum_D p_{Ca|D} p_D = \alpha + \beta \sum_D (\#D) p_D = \alpha + \beta E(\#D) \tag{3}$$

for $E(\#D)$ the average number of disease alleles of an individual in the population. Provided the second term $\beta E(\#D)$ is small, the parameter α can be approximately interpreted as the prevalence.

Let D_i be the event that an individual has the disease allele at the disease locus i (with the other loci unspecified, so that D_i is a union of certain haplotypes D), and correspondingly let p_{D_i} be the marginal frequency of the disease allele at locus i . We prove in the appendix that, under the additive model (2)

$$r_{T,Ca} \approx \sum_{i=1}^k r_{Ca,D_i} r_{T,D_i} \frac{1}{1 + \sum_{j=i}^k r_{D_i,D_j} \sqrt{p_{D_j}(1-p_{D_j})/p_{D_i}(1-p_{D_i})}} \tag{4}$$

This formula is exact if the additive model (2) is exact. If the allele frequencies p_{D_i} are equal, then the square root in the formula is 1 and disappears, and, moreover, (under (2))

the correlations r_{Ca,D_i} are equal. The formula then simplifies to

$$r_{T,Ca} \approx r_{Ca,D_1} \sum_{i=1}^k r_{T,D_i} \frac{1}{1 + \sum_{j=i}^k r_{D_i,D_j}}$$

The correlations r_{D_i,D_j} here refer to the association of the disease loci in the population (and not linkage). For two disease loci ($k=2$), the formula becomes

$$r_{T,Ca} \approx r_{Ca,D_1} \frac{r_{T,D_1} + r_{T,D_2}}{1 + r_{D_1,D_2}}$$

In comparison with formula (1) for the one-locus model, this exhibits the additional multiplication factor

$$\Delta = \frac{1 + r_{T,D_2}/r_{T,D_1}}{1 + r_{D_1,D_2}}$$

This factor quantifies the bias introduced in the correlation between T and Ca , hence in the necessary sample size to detect association, due to cryptic tagging. When the two disease loci are not correlated and T does not tag D_2 , this factor reduces to 1. However, when the two disease loci are indeed correlated and both are tagged by T , Δ can still reduce to 1 if the multiplicativity of the correlation coefficients, which is not assumed, does hold: if $r_{T,D_2} = r_{T,D_1} \times r_{D_1,D_2}$, then $\Delta = 1$. In all other cases, cryptic tagging will introduce a bias in the necessary sample size to detect association equal to the inverse of Δ^2 .

HapMap data

Δ is a function of the three pairwise correlation coefficients between tagging SNP and both disease loci. The frequency distribution of these correlation coefficients can be estimated from the phased genotype data available from the HapMap project (<http://www.hapmap.org/downloads/index.html>). For this analysis, we used data from the CEU population.

As the distribution of LD may differ between chromosomes and between SNPs of different minor allele frequency (MAF), we considered each chromosome individually and used five different bins of MAF in the analysis. SNPs with MAF of 5–10, 10–20, 20–30, 30–40 and 40–50% were considered separately. All SNPs on a chromosome that fall in a given MAF bin were ascertained and two were chosen randomly. These represent the susceptibility loci D_1 and D_2 . Next, the extent of LD was determined between the D_1 locus and all other SNPs on the chromosome. One SNP was randomly picked conditional on it being in LD with the D_1 locus with an r^2 larger than 0.8. This is the tagging SNP T , which tags D_1 and may or may not, depending on the haplotype structure of the genome, tag D_2 . With the three SNPs D_1 , D_2 and T chosen in this manner, the three pairwise correlation coefficients were calculated and Δ determined. This procedure was repeated

1000 times for each chromosome and MAF–bin combination to generate a genome-wide distribution of Δ^2 .

We did not find any systematic differences in the locations of the distribution of Δ^2 between chromosomes (Kruskal–Wallis test), but the more rare SNPs showed a visible concentration around 1 (Figure 1) (shape difference confirmed by Kolmogorov–Smirnov test, $P < 0.001$). The observed values of Δ^2 ranged between 0.00018 and 1.63.

This large range is in line with the theoretical prediction of T&H that the upper limit for the necessary sample size to detect association when multiplicity of r^2 is not assumed includes infinity. However, the percentiles of the distribution of Δ^2 show that the extreme values are rare. In 95% of our data, Δ^2 lies between 0.92 and 1.09 and except for one case Δ^2 did not reach below ≈ 0.7 . (The 0.1, 0.5, 1, 2, 2.5, 3, 4, 5, and 97.5 percentiles were equal to 0.79, 0.86,

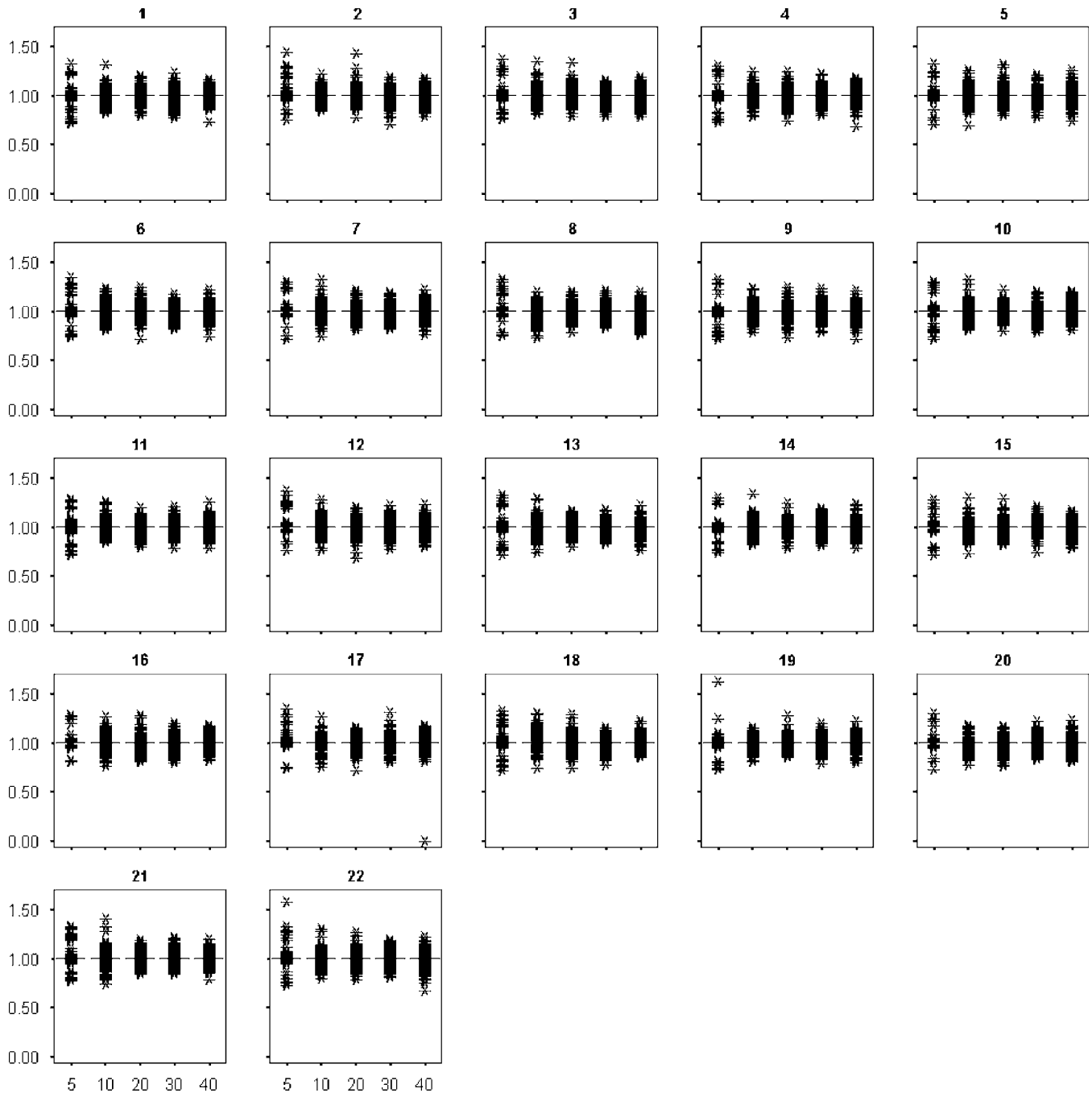


Figure 1 Distribution of the bias parameter Δ^2 . The value of Δ^2 is given on the vertical scale. Shown are scatterplots for the five MAF classes described in the text for each of the 22 autosomes.

0.88, 0.91, 0.92, 0.93, 0.94, 0.95, and 1.09, respectively.) We also explored other scenarios (with the two disease loci chosen either randomly with respect to minor allele frequency or ascertained such that one would have $MAF \approx 0.1$ and the other $MAF \approx 0.4$), but the distribution of Δ^2 was similar to the one described above (data not shown). In conclusion, more than 5% of all association studies would need 5–30% larger sample sizes to achieve the same power.

Incidentally, the lowest value of Δ^2 involved a pair of common alleles on chromosome 17. The correlation coefficients were $r_{T,D_1} = -0.9$, $r_{T,D_2} = 0.9$ and $r_{D_1,D_2} = -0.83$, representing a situation where the susceptibility allele at one locus is in fairly strong LD with the healthy allele at the other locus while the tagging SNP tags both loci equally. In such a rare, worst case scenario, detecting association through tagging is virtually impossible, even under our additive model (2).

Discussion

In this paper, we empirically assess the deviation from the HapMap theorem induced by cryptic tagging of multiple susceptibility loci by a neutral SNP. This scenario seems the most likely biological mechanism that might result in the nonmultiplicativity of r^2 . Conditional on the haplotype structure of the genome, a tagging SNP in LD with one susceptibility locus might also exhibit high levels of LD with another susceptibility locus. In fact, this paper identifies this distribution of r^2 values between a tagging SNP and a randomly placed second ‘susceptibility’ locus on the basis of the CEU HapMap data, under an additive model for penetrance. In agreement with T&H, we find that nonmultiplicativity of r^2 can indeed decrease the power of an association study, but show that this bias introduced by cryptic tagging is relatively modest. We did observe one instance where cryptic tagging would have completely abolished the power of an association test through a tagging SNP, as predicted by T&H. However, this scenario is extremely rare as long as susceptibility loci do not tend to be colocalized in the genome. If that were the case, underestimation of the necessary sample size owing to cryptic tagging would be much more severe. As also other mechanisms than the one analyzed here might account for the nonmultiplicativity of r^2 , some caution in designing genome-wide association studies seems to be in place. On the basis of these results, a safe bet is to use a sample size that is 10% larger than otherwise deemed necessary.

References

1 Gabriel SB, Schaffner SE, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.

- 2 Terwilliger JD, Hiekkalinna T: An utter refutation of the ‘fundamental theorem of the HapMap’. *Eur J Hum Genet* 2006; **14**: 426–437.
3 Thomas DC, Stram DO: An utter refutation of the ‘fundamental theorem of the HapMap’ by Terwilliger and Hiekkalinna. *Eur J Hum Genet* 2006; **14**: 1238–1239.

Appendix

Derivation of formula (4)

For two events A and B , let p_A , p_B and p_{AB} be the probabilities of A , B and $A \cap B$, and let r_{AB} be the correlation between the indicators 1_A and 1_B (defined as 1 or 0 whether the event occurs or not), that is,

$$r_{AB} = \frac{p_{AB} - p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

We assume that the allele at the tagging locus T of a randomly selected individual is conditionally independent of the case–control status Ca given the haplotypes (or genotypes) at the disease loci. Let D denote all possible haplotypic alleles at the joint disease loci, and let $D \in D$ denote the event that a random individual possesses allele D (possibly a multi-locus haplotype). Then, it follows from the general result, listed as Lemma 1 at the end of this section, that

$$r_{T,Ca} = \sum_{D \in D} (1 - p_D) r_{T,D} r_{D,Ca} \quad (A.1)$$

This exhibits the root-noncentrality parameter as a linear combination of the root-noncentrality parameters $r_{D,Ca}$ of the tests of the 2×2 tables that would score case–control status *versus* causal haplotype, for each haplotypic allele $D \in D$ in turn. In its generality, formula (A.1) is only mildly interesting. However, under special assumptions, it turns into easily interpretable formulas.

As a first application, if there are only two possible alleles at the disease loci, say D and d , then sum (A.1) has two terms, and the products of correlations in the two terms are equal ($r_{T,D} r_{D,Ca} = r_{T,d} r_{d,Ca}$), because both correlations change by a minus sign upon replacing D by d . Then, the formula reduces to the multiplicity (1).

Secondly, we derive (4) from (A.1) under assumptions (2) and (3). It follows from the latter pair of assumptions that

$$r_{D,Ca} = \frac{(p_{Ca|D} - p_{Ca})p_D}{\sqrt{p_{Ca}p_{Co}p_D(1-p_D)}} \approx \frac{(\#D - E(\#D))\beta p_D}{\sqrt{p_{Ca}p_{Co}p_D(1-p_D)}}$$

Substitution in formula (A.1) yields

$$\begin{aligned} r_{T,Ca} &\approx \sum_D \frac{(1-p_D)r_{T,D}(\#D - E(\#D))\beta p_D}{\sqrt{p_{Ca}p_{Co}p_D(1-p_D)}} \\ &= \sum_D \frac{\text{cov}(1_T, 1_D)(\#D - E(\#D))\beta}{\sqrt{p_T p_i p_{Ca} p_{Co}}} \\ &= \frac{\beta}{\sqrt{p_T p_i p_{Ca} p_{Co}}} \text{cov}(1_T, \sum_D 1_D(\#D - E(\#D))) \end{aligned}$$

Here $E(\#D)$ can be deleted, because $\sum_D 1_D = 1$ and hence is uncorrelated with any variable, and $\sum_D (\#D)1_D$ can be

rewritten as $\sum_{i=1}^k 1_{D_i}$ for D_i the event that an individual has the disease allele at the disease locus i (with the other loci unspecified). Thus, we obtain

$$r_{T, Ca} \approx \frac{\beta}{\sqrt{p_{Ca}p_{Co}}} \sum_{i=1}^k r_{T, D_i} \sqrt{p_{D_i}(1-p_{D_i})} \quad (A.2)$$

Next we eliminate β from this formula by expressing this in the correlations r_{Ca, D_i} . We have

$$p_{Ca|D_i} = \sum_D p_{Ca|D} p_{D|D_i} = \alpha + \sum_D \beta(\#D) p_{D|D_i} = \alpha + \beta E(\#D|D_i)$$

for $E(\#D|D_i)$ the expected total number of disease alleles in an arbitrary individual carrying the disease allele at locus i . Combining this with formula (3) for the prevalence, we see

$$r_{Ca, D_i} = \frac{(p_{Ca|D_i} - p_{D_i})}{\sqrt{p_{Ca}p_{Co}p_{D_i}(1-p_{D_i})}} = \beta \frac{(E(\#D|D_i) - E(\#D))\sqrt{p_{D_i}}}{\sqrt{p_{Ca}p_{Co}(1-p_{D_i})}}$$

Solving for β and substituting the solution in (A.2), we find that

$$r_{T, Ca} \approx \sum_{i=1}^k r_{Ca, D_i} r_{T, D_i} \frac{1-p_{D_i}}{E(\#D|D_i) - E(\#D)} \quad (A.3)$$

The total number of disease alleles in a random individual can be written $\#D = \sum_D (\#D) 1_D = \sum_{j=1}^k 1_{D_j}$. (The curious first equality is a consequence of our abuse of notation: as a random variable the total number of disease alleles $\#D$ in an arbitrary individual is denoted by $\#D$ if the event D occurs.) This gives

$$E(\#D) = E \sum_{j=1}^k 1_{D_j} = \sum_{j=1}^k p_{D_j}$$

$$E(\#D|D_i) = \sum_{j=1}^k E(1_{D_j} | D_i) = 1 + \sum_{j \neq i} p_{D_j | D_i}$$

Thus,

$$\begin{aligned} \frac{E(\#D|D_i) - E(\#D)}{1-p_{D_i}} &= \frac{1-p_{D_i} + \sum_{j \neq i} (p_{D_j|D_i} - p_{D_j})}{1-p_{D_i}} \\ &= 1 + \sum_{j \neq i} r_{D_i, D_j} \sqrt{\frac{p_{D_j}(1-p_{D_j})}{p_{D_i}(1-p_{D_i})}} \end{aligned}$$

We conclude the derivation of (4) by substituting this in (A.3).

Lemma 1. *If events A and B are conditionally independent given a partition D of the outcome space, then*

$$r_{AB} = \sum_{D \in \mathcal{D}} r_{AD} r_{DB} (1 - P(D))$$

Proof Because A and B are conditionally independent, they are conditionally uncorrelated, that is $\text{cov}(1_A, 1_B|D) = 0$ almost surely. Therefore, the usual conditioning rule for covariances gives

$$\begin{aligned} \text{cov}(A, B) &= E \text{cov}(1_A, 1_B|D) + \text{cov}(E(1_A|D), E(1_B|D)) \\ &= 0 + E(E(1_A|D) - E1_A)(E(1_B|D) - E1_B) \end{aligned}$$

Here on the event D the variable $E(1_A|D) - E1_A$ is equal to

$$\begin{aligned} P(AD) - P(A) &= \frac{P(A \cap D) - P(A)P(D)}{P(D)} \\ &= r_{AD} \sqrt{\frac{P(A)P(A^c)P(D^c)}{P(D)}} \end{aligned}$$

Substituting this and the corresponding formula for $P(BD) - P(B)$ in the preceding display gives

$$\begin{aligned} \text{cov}(1_A, 1_B) &= \sum_{D \in \mathcal{D}} r_{AD} \sqrt{\frac{P(A)P(A^c)P(D^c)}{P(D)}} \\ &\quad r_{BD} \sqrt{\frac{P(B)P(B^c)P(D^c)}{P(D)}} \end{aligned}$$

This can be rearranged to give the assertion.