Heritability of Complex Human Diseases in the UK Biobank.

M. Muñoz[*], **R. Pong-Wong**[*], **C.S. Haley**[*,†], ***A. Tenesa***[*,†]

[*]The Roslin Institute and R(D)SVS, University of Edinburgh, Midlothian, UK, [†]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

ABSTRACT: Heritabilities for the four most common human cancers were estimated using cancer prevalence among the 502,682 individuals recruited to the UK Biobank (www.ukbiobank.ac.uk/) and their parents. The heritability estimated from parent-offspring regressions ranged from 0.12 for lung cancer to 0.33 for prostate cancer. Our estimates were substantially smaller than previous estimates from twin studies for lung, bowel and prostate cancer whilst were similar for breast cancer, and indicate that part of the missing heritability problem in genome-wide association studies of cancer is due to twin-based heritability estimates being substantially inflated. These estimates obtained from a unique European cohort will inform future directions in cancer gene mapping and risk profiling.

Keywords: UK Biobank; heritability; cancer

Introduction

Cancer is a leading cause of death among both women and men in the Western world. The four most common cancers are those of the lung, bowel, breast and prostate. Together, they account for 46% of all cancer deaths in the UK and represent a major burden to the health system.

The aetiology of cancer is due to the complex interplay of environmental and genetic factors. Genetic risk factors are implied by the increased prevalence of cancer among relatives of cancer patients compared with relatives of random people from the population. However, the increased incidence among the relatives of patients could be due to their genetic similarity or to shared environmental exposures within family (e.g. smoking or drinking habits), therefore it is important to distinguish which proportion of the increased familial risk is due to genetics and which to shared environmental factors. It is also important to estimate the relative contribution of genetics to the overall variation in risk (Curnow (1972); Falconer (1965); Rowe and Tenesa (2012); Tenesa and Haley (2013)). The heritability is defined as the proportion of the phenotypic variation that is due to genetic factors. A heritability larger than zero justifies the search for the susceptibility loci that contribute to the increased risk of disease, their functional characterization and the use of genetics for constructing prediction models of cancer risk based on genetic markers (Dunlop et al. (2013); Tenesa and Dunlop (2009)).

Similarly to other complex human diseases, cancer is recorded as a dichotomous trait and heritability is estimated under the assumption that there is an underlying and normally distributed phenotypic liability to disease. Under this model, only those individuals with a liability phenotype above the threshold determined by the disease prevalence express the disease (Lush et al. (1948)).

Twin-studies have estimated the heritability of the liability to cancer to be moderate to high (Lichtenstein et al. (2000)) and this has prompted large-scale gene mapping studies in what are known as genome-wide association studies (GWAS). The apparent failure of GWAS to identify a substantial proportion of the underlying genetic variation has been coined as 'the missing heritability' problem. In this paper, we explore the idea that at least part of the 'missing heritability' is due to previous overestimation of the heritability as a result of shared environmental factors among relatives or non-additive genetic variation.

We contrast parent-offspring data from the UK Biobank (UKB) with previous estimates of heritability from the largest twin study to date, and make predictions of the likely contribution of additive and non-additive genetic variation to the heritability of liability to breast, bowel, prostate and lung cancer.

Materials and Methods

UK Biobank Data. A total of 502,682 individuals (i.e. the propositi) were recruited across the UK between 2006 and 2010 to take part in one of the largest prospective epidemiological studies on human health. All participants gave informed consent, and completed several questionnaires about their lifestyle, environmental risk factors and medical history (Allen et al. (2014)). Personal history of disease was collected for each propositus. The propositi reported 445 distinct types of non-cancer and 81 cancer illnesses. Table 1 shows the age and gender distribution of the propositi.

Table 1. Age and gender distribution in the propositi of the UK Biobank.

Age group	Male	Female
38-50	59,049	72,302
51-60	77,376	99,124
Over 61	91,571	100,930
Total	227,996	272,356

First-degree family history of disease was collected for twelve broadly defined diseases: heart disease, stroke, chronic bronchitis/emphysema, high blood pressure, diabetes, Alzheimer's disease/dementia, Parkinson's disease, severe depression, lung cancer, bowel cancer, prostate cancer and breast cancer. Family history was available for the blood relatives of 493,853 propositi.

Heritability estimates. The regression of parents on propositi (b) was used to estimate the narrow sense heritability ($h^2 = 2b$) of liability to cancer of the lung, prostate, breast and bowel. Across generation differences in disease prevalence were accounted for by using the appropriate control population for comparison. The regression is estimated by the formula derived by Falconer (1965), that is:

$$b = \frac{p_g(x_c - x_r)}{a_g}$$

where p_g is the prevalence of the cancer in the relevant population (i.e. comparable with the propositi) within the UK Biobank, x_c is the deviation of the threshold of liability that defines cancer status from the mean of relatives of unaffected propositi, x_r is the deviation of the threshold of liability that defines cancer status from the mean of relatives of affected propositi, and a_g is the mean liability deviation of the affected propositi from the mean liability of the relevant population within the UK Biobank. The sampling variance (V_b) of the estimates of b was estimated according to the appendix of Falconer (1965) and by bootstrapping. The two estimates were practically identical and only the approximation of Falconer is presented here.

Four independent estimates of b were obtained from father-son, mother-son, father-daughter and mother-daughter pairings and a weighted mean (b_w) was estimated from the four estimates. The reciprocal of the sampling variance of each regression coefficient was used to weight each regression coefficient. The sampling variance (V_{bw}) of the weighted mean was estimated as the reciprocal of the sum of the weights, and the standard error of the heritability estimates was obtained as the square root of $4V_{bw}$. We assume that larger than second order epistatic variances are negligible. Hence, b_w estimates:

$$h^2_{PO} = 2b_w = V_A + 2V_{Cpo} + \frac{1}{2}V_{AA}$$

where V_{Cpo} is the parent-offspring shared environmental variance, V_{AA} is the additive by additive epistatic variance, and V_P is assumed to be 1.

Twin estimates. Heritability estimates based on 44,788 pairs of twins registered in the Swedish, Danish, and Finnish twin registries were obtained from Lichtenstein et al. (2000). These estimates were obtained assuming a liability model of disease and a genetic model where all genetic variation was assumed to be additive (i.e. a model that contains additive, shared and non-shared environmental effects). We assume the intraclass correlations among monozygotic and dizygotic twins estimate:

$$t_{mz} = V_A + V_D + V_{AA} + V_{Cmz}$$
$$t_{dz} = \frac{1}{2}V_A + \frac{1}{4}V_D + \frac{1}{4}V_{AA} + V_{Cdz}$$

Assuming $V_{Ctwin} = V_{Cdz} = V_{Cmz}$ and $V_P = 1$, one can estimate the heritability from twin correlations as:

$$h^2_{twin} = 2(t_{mz} - t_{dz}) = V_A + \frac{3}{2}V_D + \frac{3}{2}V_{AA}$$

Additive and non-additive genetic variation. The twin-based estimates of narrow sense heritability will be inflated by non-additive genetic variation unless $V_D = V_{AA} = 0$. We partition the total genetic variation estimated in the twin-study into additive and non-additive by removing the additive component estimated from the parent-offspring regressions in the UK Biobank after correction by the common environment estimated in the twin study (i.e. we assumed $V_{Cpo} = V_{Ctwin}$). Under this model, $V_{non\text{-}A}$ is estimated as:

$$V_{non-A} = h^2_{twin} - (h^2_{PO} - 2V_{Ctwin}) = \frac{3}{2}V_D + V_{AA}$$

Results and Discussion

Table 2 shows the prevalence of the four cancers in the UK Biobank and the estimated heritability from parent-offspring regressions. All heritabilities were significantly different from zero. Prostate cancer showed the largest genetic contribution followed by breast, bowel and lung cancer. Despite being significantly larger than zero, due to the large sample size, the heritability of lung cancer was very modest. Comparison of our estimates with the estimates of narrow sense heritability in the twin study (Table 3) shows that estimates obtained from parent-offspring regressions were at most of the same magnitude as those estimated in the twin-study and often smaller despite being inflated by twice the shared environmental component. Indeed, the shared environment is likely to be very important to explain the familial aggregation of lung cancer since the parent-offspring correlation could be completely explained by the shared environment among twins. Non-additive sources of genetic variance could potentially contribute to ~55%, ~21% and ~37% of the heritability estimates obtained in the twin study for bowel, prostate and breast cancer.

Table 2. Prevalence of cancer in the UK Biobank and heritability estimates from the parent-offspring regression.

Cancer	q (SE)	h^2_{po} (SE)
Lung	0.001 ($5.08\text{x}10^{-05}$)	0.120 (0.034)
Bowel	0.006 ($1.02\text{x}10^{-04}$)	0.258 (0.015)
Prostate	0.015 ($2.54\text{x}10^{-04}$)	0.333 (0.021)
Breast	0.041 ($4.06\text{x}10^{-04}$)	0.291 (0.014)

q: prevalence; SE: Standard error; h^2_{po}: heritability estimated from parent-offspring regression

One could consider an alternative model in which the common environment contributes to the sibling correlation but not the parent-offspring correlation (i.e. $V_{Cpo} = 0$). In this case, the overestimation of the narrow sense heritability would be somehow smaller, but still important for lung (116%), bowel (35%), and prostate (27%).

Table 3. Partition of the total genetic variance in additive and non-additive variance.

	Twin study		UKB		
	h^2_{twin}	V_{Ctwin}	h^2_{po}	V_A	V_{non-A}
Lung	0.26	0.12	0.12	-0.12	0.38
Bowel	0.35	0.05	0.26	0.16	0.19
Prostate	0.42	0.00	0.33	0.33	0.09
Breast	0.27	0.06	0.29	0.17	0.10

h^2_{twin}: narrow sense heritability estimated from twins; V_{Ctwin}: common environment estimated from twins; h^2_{po}: narrow sense heritability estimated from parent-offspring regression; V_A: additive variance; V_{non-A}: non-additive variance; $h^2_{po} = h^2 + V_{Cpo}$;
V_{Cpo} is assumed to be equal to V_{Ctwin}; $h^2_{twin} = V_A + V_{non-A}$

Conclusion

We have shown preliminary results that suggest that heritability estimates of liability to cancer from twin-studies are larger than from parent-offspring correlations and this discrepancy could potentially be explained by non-additive sources of genetic variation.

Acknowledgments

This research has been conducted using the UK Biobank Resource.

Literature Cited

Allen, N. E., Sudlow, C., Peakman, T., et al. (2014). Sci. Transl. Med. 6**:** 224ed224.

Curnow, R. N. (1972). Biometrics. 28**:** 931-946.

Dunlop, M. G., Tenesa, A., Farrington, S. M., et al. (2013). Gut. 62**:** 871-881.

Falconer, D. S. (1965). Ann. Hum. Genet. 29**:** 51-76.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., et al. (2000). New Engl. J. Med. 343**:** 78-85.

Lush, J. L., Lamoreux, W. F., and Hazel, L. N. (1948). Poultry Sci. 27**:** 375-388.

Rowe, S. J., and Tenesa, A. (2012). Curr. Genomics. 13**:** 213-224.

Tenesa, A., and Dunlop, M. G. (2009). Nat. Rev. Genet. 10**:** 353-358.

Tenesa, A., and Haley, C. S. (2013). Nat. Rev. Genet. 14**:** 139-149.