Genome-wide Association Study of Cancer Risk in UK Biobank

VERSION 1: 23/11/2021

Kimberley Burrows, Caroline J Bull, Tom Dudding, Mark Gormley, Tim Robinson, Vanessa Tan, James Yarmolinsky, Philip C Haycock.
MRC Integrative Epidemiology Unit (IEU), University of Bristol, UK
Integrative Cancer Epidemiology Unit (ICEP), University of Bristol, UK

1. Introduction

This document describes genome-wide association studies (GWAS) of pan-cancer and site-specific cancers in UK Biobank participants. The UK Biobank is a population-based cohort study consisting of approximately 500,000 middle-aged participants, who were recruited between 2006 and 2010 from across the UK (Fry et al. 2017; Bycroft et al. 2017). A full description of the study design, participants and quality control (QC) methods have been described in detail previously (Sudlow et al. 2015). UK Biobank received ethical approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382).

We performed GWAS for pan-cancer and site-specific cancers identified via linkage to the UK Cancer Registry (updated to April 2019). GWAS was performed using BOLT-LMM and summary statistics were deposited in the MRC Integrative Epidemiology Unit (IEU) Open GWAS database (https://gwas.mrcieu.ac.uk/). These are publicly available for use in further analyses.

This work relates to the UK Biobank application 15825: PI Dr Philip C Haycock.

If you have any questions about the suitability of these summary statistics for your analysis, please contact grp-ukbbcanceroutcomes@groups.bristol.ac.uk

2. Defining cancer cases and controls

GWAS was undertaken for pan-cancer or site-specific cancer diagnoses before or after enrolment to UK Biobank.

Cancer cases

The following UK Biobank field codes contain coded data on cancer incidence, obtained through linkage to national cancer registries (40006 [type of cancer: ICD10], 40013 [type of cancer: ICD9], and 40012 [behaviour of cancer tumour]. Cancer cases were recorded according to the International Classification of Diseases (ICD9, ICD10) with data completed to April 2019.

Table 1 describes the ICD codes and case/control sample sizes for each of the cancers included in each GWAS [Version 1 23/11/2021].

Cancer cases were defined using the following parameters:

I. Individuals with a site-specific cancer code (ICD10:C00-C97 and ICD9:140.0-208.9)
II. Site-specific cancer morphology (behaviour) was dealt with using the following rules:
    - Cancer behaviours: "Malignant, primary site", "Malignant, microinvasive", "Malignant, metastatic site", "Malignant, uncertain whether primary or metastatic site" were included in the dataset.
    - Cancer behaviours: “Benign”, “Uncertain whether benign or malignant”, and “Carcinoma in situ” were excluded from the dataset.
III. Individuals with an ICD10: D code but no C code were not included as cases (these are benign or carcinomas in situ).

Cancer controls

Controls were defined using the following parameters:

I. Individuals who do not have any cancer code (ICD9 & ICD10 - **C and D codes**)
II. Individuals who have no self-report of cancer

Special considerations

- Controls for sex specific cancers were further filtered by sex.
- Pan-cancer was analysed twice: 1) cases of any cancer, and 2) cases of any cancer but **excluding** ICD10 code C44 and ICD9 code 173 from cases (these are non-malignant skin cancer).
- Participants with diagnoses of different site-specific cancers are included in each site-specific cancer GWAS (if meeting the above parameters) e.g., participant “A” has a cancer diagnosis of malignant melanoma several years prior to a cancer diagnosis of breast cancer. Participant “A” will be included as a case in both GWAS for malignant melanoma and breast cancer.
- Lung cancer was analysed twice: 1) GWAS was adjusted for array chip, and 2) GWAS was unadjusted for array chip (see below).
- The head and neck cancer cases were selected using specific behaviour and histology codes per request. See Table 1 for details of these inclusions.

3. Undertaking GWAS

GWAS of the UK Biobank cancer phenotypes was performed using an established analysis pipeline described in:

Quality control filtering of the UK Biobank genetic data was conducted by R.Mitchell, G.Hemani, T.Dudding, L.Corbin, S.Harrison, L.Paternoster as described in the published protocol (doi:10.5523/bris.1ovaau5sxunp2cv8rcy88688v).

The MRC IEU UK Biobank GWAS pipeline was developed by B.Elsworth, R.Mitchell, C.Raistrick, L.Paternoster, G.Hemani, T.Gaunt (doi:10.5523/bris.pnoat8cxo0u52p6ynfaekeigi.).

Briefly, GWAS was conducted using a linear mixed model (LMM) association method as implemented in BOLT-LMM (v2.3) (Loh et al. 2015). To model population structure in the sample we used 143,006 directly genotyped SNPs, obtained after filtering on MAF > 0.01; genotyping rate > 0.015; Hardy-Weinberg equilibrium p-value < 0.0001 and LD pruning to an r^2 threshold of 0.1 using PLINKv2.00.

Genotype array and sex were adjusted for in the models. However, for lung cancer, an additional GWAS was performed unadjusted for genotype array chip. The QC document [**doi:10.5523/bris.1ovaau5sxunp2cv8rcy88688v**] states:

“There is evidence of differential array effect on markers scattered across the genome and so you may wish to adjust for genotyping array ('chip') in your analysis. However, if your outcome of interest is likely to affect lung function or smoking behaviour you should be aware that such an adjustment may introduce collider bias (due to UKBiLEVE participants being genotyped on a different array) and so we would recommend performing analyses with and without adjustment for genotyping array as sensitivity analyses.”

BOLT-LMM association statistics are on the linear scale. The effect estimates from this analysis can therefore be interpreted as the change in disease risk per copy of the effect allele. These results can be converted to log odds ratios using a Taylor transformation expansion series (Loh et al. 2018).

Conversion of test statistics to the log odds ratio scale

Betas and SEs can be converted to approximate log odds ratios using the following R code as an example:

```
# Convert BOLT LMM effects to log odds
# formula: log OR = beta / (u(1-u)); where u=ncases/(ncases + ncontrol) REPEAT with SE
# ukbb_all is a data-frame of GWAS summary statistics
ukbb_all$ncase <- 52400
ukbb_all$ncontrol <- 372016
ukbb_all$u <- ukbb_all$ncase/(ukbb_all$ncase + ukbb_all$ncontrol)
ukbb_all$beta <- ukbb_all$beta/ (ukbb_all$u * (1 - ukbb_all$u))
ukbb_all$se <- ukbb_all$se / (ukbb_all$u * (1 - ukbb_all$u))
```

Exclusion of SNPs with unreliable minor allele frequency in cases

It is known that the use of linear models to test genetic associations with binary phenotypes can lead to inflated false positive findings, especially for rare variants in analyses with a small number of cases compared to controls. To mitigate the impact of model misspecification on our results, we therefore removed SNPs with a minor allele frequency less than the following threshold (Howrigan, Abbott, and Palmer 2017):

$$MAF\ threshold = \frac{25}{2 * case\ sample\ size}$$

4. Data and availability

Table 1 Cases and controls for UK Biobank cancers

CANCER	ICD9	ICD10	TOTAL	CASES	CONTROLS	NUMBER OF SNPS[3]
PAN-CANCER	1400-2089	C00.0-C97.9	442239	70223	372016	12321875
PAN-CANCER EXCLUDING NON-MELANOMA SKIN CANCER[1]	1400-2089	C00.0-C97.9	422659	50643	372016	12321875
SKIN - NON-MELANOMA	1730, 1731, 1732, 1733, 1734, 1735, 1736, 1737, 1738, 1739	C44.0, C44.1, C44.2, C44.3, C44.4, C44.5, C44.6, C44.7, C44.8, C44.9	395710	23694	372016	12321875
BREAST CANCER	1740, 1741, 1742, 1743, 1744, 1745, 1746, 1747, 1748, 1749	C50.0, C50.1, C50.2, C50.3, C50.4, C50.5, C50.6, C50.8, C50.9	212402	13879	198523	12321854
PROSTATE CANCER	185	C61	182625	9132	173493	12099538
COLORECTAL CANCER	1530, 1531, 1532, 1533, 1534, 1535, 1536, 1537, 1538, 1539	C18.0, C18.1, C18.2, C18.3, C18.4, C18.5, C18.6, C18.7, C18.8, C18.9, C19, C20	377673	5657	372016	11743334
HAEMATOLOGICAL MALIGNANCIES	2000, 2001, 2008, 2014, 2015, 2016, 2017, 2019, 2020, 2021, 2022, 2024, 2028, 2030, 2040, 2041, 2049, 2050, 2051, 2059, 2061, 207, 2081, 2089	C81.0, C81.1, C81.2, C81.3, C81.7, C81.9, C82.0, C82.1, C82.2, C82.7, C82.9, C83.0, C83.1, C83.2, C83.3, C83.4, C83.5, C83.7, C83.8, C83.9, C84.0, C84.1, C84.2, C84.3, C84.4, C84.5, C85.0, C85.1, C85.7, C85.9, C86.2, C88.0, C88.4, C88.9, C90.0, C90.1, C90.2, C90.3, C91.0, C91.1, C91.3, C91.4, C91.5, C91.9, C92.0, C92.1, C92.3, C92.4, C92.5, C92.7, C92.9, C93.0, C93.1, C94.0, C94.4, C94.6, C95.0, C95.1, C95.7, C95.9, C96.1, C96.2, C96.3, C96.7, C96.8, C96.9	376568	4552	372016	11568275
SKIN - MALIGNANT MELANOMA	1720, 1721, 1722, 1723, 1724, 1725, 1726, 1727, 1728, 1729	C43.0, C43.1, C43.2, C43.3, C43.4, C43.5, C43.6, C43.7, C43.8, C43.9	375767	3751	372016	11402537

CANCER	ICD9	ICD10	TOTAL	CASES	CONTROLS	NUMBER OF SNPS[3]
LUNG CANCER[2]	1622, 1623, 1624, 1625, 1628, 1629	C34.0, C34.1, C34.2, C34.3, C34.8, C34.9	374687	2671	372016	11085930
BLADDER CANCER	1880, 1882, 1884, 1886, 1888, 1889	C67.0, C67.1, C67.2, C67.3, C67.4, C67.5, C67.6, C67.7, C67.8, C67.9	373295	1279	372016	9914976
ALL LEUKAEMIA	207, 2040, 2041, 2049, 2050, 2051, 2059, 2061, 2081, 2089	C91.0, C91.1, C91.3, C91.4, C91.5, C91.9, C92.0, C92.1, C92.3, C92.4, C92.5, C92.7, C92.9, C93.0, C93.1, C94.0, C94.4, C94.6, C95.0, C95.1, C95.7, C95.9	373276	1260	372016	9890971
OVARIAN CANCER	1830	C56	199741	1218	198523	9832397
HEAD AND NECK CANCER[4]	1410, 1412, 1413, 1419, 1431, 1449, 1450, 1451, 1452, 1453, 1455, 1460, 1461, 1610	C00.3, C00.4, C00.5, C00.6, C00.9, C01, C02.0, C02.1, C02.2, C02.3, C02.4, C02.8, C02.9, C03.0, C03.1, C03.9, C04.0, C04.1, C04.8, C04.9, C05.0, C05.1, C05.2, C05.8, C05.9, C06.0, C06.1, C06.2, C06.8, C06.9, C09.0, C09.1, C09.8, C09.9, C10.0, C10.1, C10.2, C10.3, C10.4, C10.8, C10.9, C12, C13.0, C13.1, C13.2, C13.9, C14.0, C32.0, C32.1, C32.2, C32.3, C32.8, C32.9	373122	1106	372016	9665502
ORAL AND OROPHARYNGEAL CANCER[4]	1412, 1413, 1419, 1431, 1449, 1450, 1451, 1452, 1410, 1453, 1455, 1460, 1461	C00.3, C00.4, C00.5, C00.6, C00.9, C0.1, C02.0, C02.1, C02.2, C02.3, C02.4, C02.8, C02.9, C03.0, C03.1, C03.9, C04.0, C04.1, C04.8, C04.9, C05.0, C05.1, C05.2, C05.8, C05.9, C06.0, C06.1, C06.2, C06.8, C06.9, C09.0, C09.1, C09.8, C09.9, C10.0, C10.1, C10.2, C10.3, C10.4, C10.8, C10.9, C12, C13.0, C13.1, C13.2, C13.9, C14.0	372855	839	372016	9196331
LYMPHOID LEUKAEMIA	2040, 2041, 2049	C91.0, C91.1, C91.3, C91.4, C91.5, C91.9	372776	760	372016	9026368
MALIGNANT NEOPLASMS OF OESOPHASGUS	150, 1505, 1509	C15.0, C15.1, C15.2, C15.3, C15.4, C15.5, C1.58, C15.9	372756	740	372016	8981825
MALIGNANT NEOPLASM OF BRAIN	1910, 1911, 1912, 1914, 1916, 1917, 1918, 1919	C71.0, C71.1, C71.2, C71.3, C71.4, C71.5, C71.6, C71.7, C71.8, C71.9	372622	606	372016	8640796
MULTIPLE MYELOMA	2030	C90.0	372617	601	372016	8627432
CERVICAL CANCER	1800, 1801, 1808, 1809	C53.0, C53.1, C53.8, C53.9	199086	563	198523	8518071
OROPHARYNGEAL CANCER[4]	1410, 1453, 1455, 1460, 1461	C01, C02.4, C05.1, C05.2, C05.8, C05.9, C09.0, C09.1, C09.8, C09.9, C10.0, C10.1, C10.2, C10.3, C10.4, C10.8, C10.9, C12, C13.0, C13.1, C13.2, C13.9, C14.0	372510	494	372016	8295888

CANCER	ICD9	ICD10	TOTAL	CASES	CONTROLS	NUMBER OF SNPS[3]
MYELOID LEUKAEMIA	2050, 2051, 2059	C92.0, C92.1, C92.3, C92.4, C92.5, C92.7, C92.9	372478	462	372016	8183381
ORAL CAVITY CANCER[4]	1412, 1413, 1419, 1431, 1449, 1450, 1451, 1452	C00.3, C00.4, C00.5, C00.6, C00.9, C02.0, C02.1, C02.2, C02.3, C02.8, C02.9, C03.0, C03.1, C03.9, C04.0, C04.1, C04.8, C04.9, C05.0, C06.0, C06.1, C06.2, C06.8, C06.9	372373	357	372016	7735567
MALIGNANT NEOPLASMS OF LIVER AND INTRAHEPATIC BILE DUCTS	1550	C22.0, C22.1, C22.3, C22.4, C22.7, C22.9	372366	350	372016	7700172
LARYNGEAL CANCER[4]	1610	C32.0, C32.1, C32.2, C32.3, C32.8, C32.9	372289	273	372016	7252199
LIVER CELL CARCINOMA	1550	C22.0	372184	168	372016	6316692

1 Pan-cancer excluded cases of non-melanoma skin cancers (ICD10:C44 and ICD9:173)

2 Lung cancer was additionally GWAS'd without adding chip as a covariate. The QC document [**doi:10.5523/bris.1ovaau5sxunp2cv8rcy88688v**] states:
"There is evidence of differential array effect on markers scattered across the genome and so you may wish to adjust for genotyping array ('chip') in your analysis. However, if your outcome of interest is likely to affect lung function or smoking behaviour you should be aware that such an adjustment may introduce collider bias (due to UKBiLEVE participants being genotyped on a different array) and so we would recommend performing analyses with and without adjustment for genotyping array as sensitivity analyses."

3 SNPs were filtered post-GWAS according to the MAF threshold set for unreliable minor allele frequency in cases

4 These cancer sub-sites included the following behaviours only: "Malignant, primary site" (8010/3) & "Carcinoma in situ" (8010/2). The cases also only contained squamous cell carcinomas as identified using the histology codes 8070-8078

GWAS summary statistics have been deposited in the IEU OpenGWAS database [https://gwas.mrcieu.ac.uk/]

5. Acknowledgement

If you use these summary statistics in your analyses, please reference this documentation (stating the DOI) in the methods and use the following citation:

Genome-wide association study of cancer risk in UK Biobank. Kimberley Burrows, Caroline J Bull, Tom Dudding, Mark Gormley, Timothy Robinson, Vanessa Tan, James Yarmolinsky, Philip C Haycock as described in the published protocol

These GWAS results are accessible through The MRC IEU OpenGWAS database. Please cite the following:

The MRC IEU OpenGWAS data infrastructure. Ben Elsworth, Matthew Lyon, Tessa Alexander, Yi Liu, Peter Matthews, Jon Hallett, Phil Bates, Tom Palmer, Valeriia Haberland, George Davey Smith, Jie Zheng, Philip Haycock, Tom R Gaunt, Gibran Hemani. bioRxiv 2020.08.10.244293v1. doi: 10.1101/2020.08.10.244293 - (Elsworth et al. 2020)

Please also cite the QC and GWAS pipeline documents:

Quality control filtering of the UK Biobank genetic data was conducted by R.Mitchell, G.Hemani, T.Dudding, L.Corbin, S.Harrison, L.Paternoster as described in the published protocol (doi:10.5523/bris.1ovaau5sxunp2cv8rcy88688v) - (Ruth Mitchell, Gibran Hemani, Tom Dudding, Laura Corbin, Sean Harrison 2019)

The MRC IEU UK Biobank GWAS pipeline was developed by B.Elsworth, R.Mitchell, C.Raistrick, L.Paternoster, G.Hemani, T.Gaunt (doi:10.5523/bris.pnoat8cxo0u52p6ynfaekeigi.) - (Ruth Mitchell, Elsworth, BL, Mitchell, R, Raistrick, CA, Paternoster, L, Hemani, G, Gaunt 2019)

References

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2017. “Genome-Wide Genetic Data on ~500,000 UK Biobank Participants.” *BioRxiv*, July, 166298. https://doi.org/10.1101/166298.

Elsworth, Ben, Matthew Lyon, Tessa Alexander, Yi Liu, Peter Matthews, Jon Hallett, Phil Bates, et al. 2020. “The MRC IEU OpenGWAS Data Infrastructure.” *BioRxiv*, August, 2020.08.10.244293. https://doi.org/10.1101/2020.08.10.244293.

Fry, Anna, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. 2017. “Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population.” *American Journal of Epidemiology* 186 (9): 1026–34. https://doi.org/10.1093/aje/kwx246.

Howrigan, Daniel, Liam Abbott, and Duncan Palmer. 2017. “Details and Considerations of the UK Biobank GWAS.” Neale Lab. 2017. http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas.

Loh, Po Ru, Gleb Kichaev, Steven Gazal, Armin P. Schoech, and Alkes L. Price. 2018. “Mixed-Model Association for Biobank-Scale Datasets.” *Nature Genetics*. Nature Publishing Group.

https://doi.org/10.1038/s41588-018-0144-6.

Loh, Po Ru, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjálmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, et al. 2015. “Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts.” *Nature Genetics* 47 (3): 284–90. https://doi.org/10.1038/ng.3190.

Ruth Mitchell, Elsworth, BL, Mitchell, R, Raistrick, CA, Paternoster, L, Hemani, G, Gaunt, TR. 2019. “MRC IEU UK Biobank GWAS Pipeline Version 2.” https://doi.org/10.5523/bris.pnoat8cxo0u52p6ynfaekeigi.

Ruth Mitchell, Gibran Hemani, Tom Dudding, Laura Corbin, Sean Harrison, Lavinia Paternoster. 2019. “UK Biobank Genetic Data: MRC-IEU Quality Control, Version 2.” https://doi.org/10.5523/bris.1ovaau5sxunp2cv8rcy88688v.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.” *PLOS Medicine* 12 (3): e1001779. https://doi.org/10.1371/journal.pmed.1001779.