nature genetics

Supplementary information

https://doi.org/10.1038/s41588-024-01966-6

Gene-based burden tests of rare germline variants identify six cancer susceptibility genes

In the format provided by the authors and unedited

SUPPLEMENTARY NOTES

Supplementary Note 1

In a recent study by Stankovic et al., burden of LOF+missense variants in *SAMHD1* associated with higher age at menopause and increased risk of cancer[1]. *SAMHD1* plays a variety of roles in virology, immunology, and cell biology[2]. In our results, burden of 63 LOF and 322 missense variants in *SAMHD1* associated with increased risk of any cancer diagnosis (OR=1.4, 95% CI=1.3-1.5, $P=2.6\times10^{-20}$, Table 1) as well as cancers in several different organs; where the most significant results are for; breast cancer (OR=1.5, 95% CI=1.3-1.7, $P=3.2\times10^{-9}$), prostate cancer (OR=1.5, 95% CI=1.3-1.7, $P=7.8\times10^{-8}$) and chronic lymphocytic leukemia (OR=2.7, 95% CI=1.8-4.0, $P=1.0\times10^{-6}$, Supplementary Table 13). In comparison, burden of only LOF variants in *SAMHD1* associated with increased risk of cancers with an OR of 1.7 (95% CI=1.3-2.3, $P=2.4\times10^{-4}$).

Supplementary Note 2

The TGC repeat microsatellite in *BIK* has a 15-repeat reference allele and multiple other alleles, ranging in length from five to 25 repeats (Supplementary Table 8). In the UK Biobank dataset, the most common alternative allele contains nine repeats, corresponding to an 18 base-pair deletion (rs759887547), which associated nominally with prostate cancer (MAF=0.23%, OR=1.4, $P=7.5\times10^{-3}$). In the Icelandic dataset, the most common alternative allele contains 19 repeats, corresponding to a 12 base-pair insertion (rs753983450), which associated nominally with prostate cancer (MAF=0.50%, OR=1.4, P=0.014). In the publicly available FinnGen dataset[3], both the five- and nine-repeat alleles, also associate with prostate cancer (rs965427251,

MAF=0.30%, OR=2.1, P=2.6×10^{-8} and rs759887547, MAF=0.56%, OR=1.5, P=8.2×10^{-6}). We note that the five-repeat microsatellite allele in the UK Biobank corresponds to both rs758385246 and rs965427251, resulting in the same sequence. In summary, all non-reference repeat alleles that have at least 100 male carriers, associated with prostate cancer and we see a dosage effect such that the odds ratios increase with both lower and higher repeat length than the 15-repeat reference allele.

Supplementary Note 3

Burden of LOF variants in *CMTR2* did not associate with smoking (ever vs never smokers or smoking quantity) (Supplementary Table 10). Smoking quantity was available for 1,039 lung cancer cases and 23,412 controls in the Icelandic data, and 2,106 lung cancer cases and 94,807 controls in the UK Biobank data. Adjusting for the number of cigarettes smoked per day had no impact on the association between *CMTR2* and lung cancer (adjusted OR=5.3, P=7.6×10^{-6} and unadjusted OR=5.3, P=6.1×10^{-6}, among individuals with smoking quantity information; P-het = 0.99). These results indicate that the association between burden of LOF variants in CMTR2 and lung cancer was not driven by an effect on smoking.

Supplementary Note 4

Members of the DBDS Genomic Consortium

Karina Banasik[1], Jakob Bay[2], Jens Kjærgaard Boldsen[3], Thorsten Brodersen[2], Søren Brunak[1], Alfonso Buil Demur[4], Lea Arregui Nordahl Christoffersen[2], Maria Didriksen[5], Khoa Manh

Dinh[3], Joseph Dowsett[5], Christian Erikstrup[3], Bjarke Feenstra[5], Frank Geller[5], Daniel Gudbjartsson[6], Thomas Folkmann Hansen[7], Dorte Helenius Mikkelsen[4], Lotte Hindhede[3], Henrik Hjalgrim[8], Jakob Hjorth von Stemann[5], Bitten Aagaar Jensen[9], Andrew Joseph Schork[4], Kathrine Kaspersen[3], Bertram Dalskov Kjerulff[3], Mette Kongstad[5], Susan Mikkelsen[3], Christina Mikkelsen[5], Janna Nissen[5], Mette Nyegaard[10], Sisse Rye Ostrowski[5], Ole Birger Pedersen[2], Liam James Elgaard Quinn[2], Thorunn Rafnar[6], Palle Duun Rohde[10], Klaus Rostgaard[8], Michael Schwinn[5], Erik Sørensen[5], Kari Stefansson[6], Hreinn Stefansson[6], Lise Wegner Thørner[5], Unnur Þorsteinsdóttir[6], Mie Topholm Bruun[11], Henrik Ullum[12], Thomas Werge[4], David Westergaard[1]

1. Novo Nordisk Foundation, Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

2. Department of Clinical Immunology, Zealand University Hospital, Køge, Denmark

3. Department of Clinical Immunology, Aarhus University Hospital, Aarhus, Denmark

4. Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen University, Roskilde, Denmark

5. Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark

6. deCODE genetics, Reykjavik, Iceland

7. Danish Headache Center, Department of Neurology, Rigshospitalet, Glostrup, Denmark

8. Danish Cancer Society Research Center, Copenhagen, Denmark

9. Department of Clinical Immunology, Aalborg University Hospital, Aalborg, Denmark

10. Department of Health Science and Technology, Faculty of Medicine, Aalborg Univeristy, Aalborg, Denmark

11. Department of Clinical Immunology, Odense University Hospital, Odense, Denmark

12. Statens Serum Institut, Copenhagen, Denmark

Supplementary Note 5

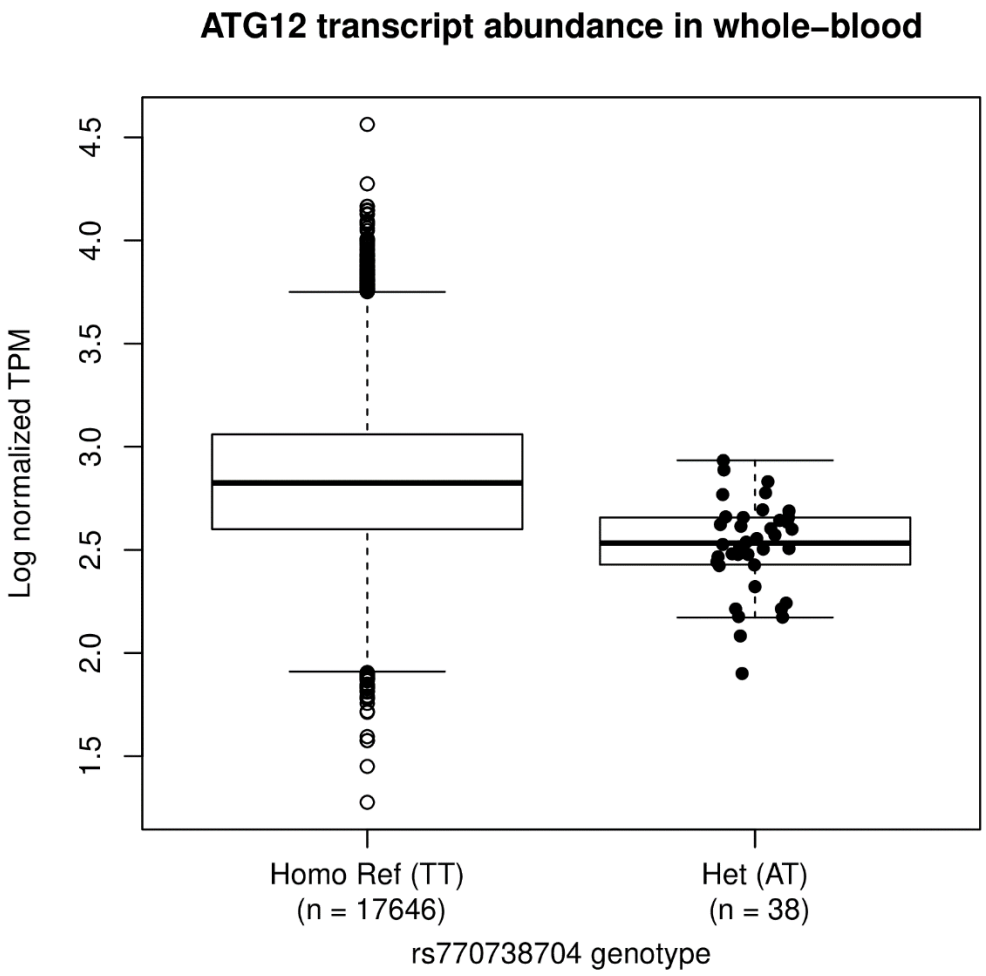
Rare variants that are only present in one or few families, can create false positive associations. However, within the burden framework we collapse rare variants together and consequently reduce the relatedness among the carriers in each test, compared to carriers of a single variant of the same frequency. In the Icelandic dataset, the relatedness is much higher than in the UK Biobank dataset, and in the UK Biobank dataset we are collapsing together substantially more variants per gene than in the Icelandic dataset (median number of variants per gene is 2 in Iceland and 34 in UK Biobank, for burden of LOF variants, and 8 in Iceland and 137 in UK Biobank for burden of LOF+missense variants).

To explore if the structure of our Icelandic dataset is creating false positive associations, we can utilize intronic and intergenic variants to analyze the distribution of test statistics under the null hypothesis. For the traits we are reporting novel associations for, we tested all intronic and intergenic variants in the genome with minor allele frequencies matching the frequencies of the reported burden genotypes. We found no variants associating with any of the traits (Supplementary Figure 7), except a few variants that are correlated with known variants in *BRCA2* and *BRCA1* for breast cancer and *CDKN2A* for melanoma. We therefore removed chromosomes 13 and 17 for breast cancer and chromosome 9 for melanoma in Supplementary Figure 7. This confirms that the relatedness is not producing false positive associations for these traits and frequencies, and that our association test is well calibrated, as the observed distribution of p-values matches the expected distribution under the null hypothesis.

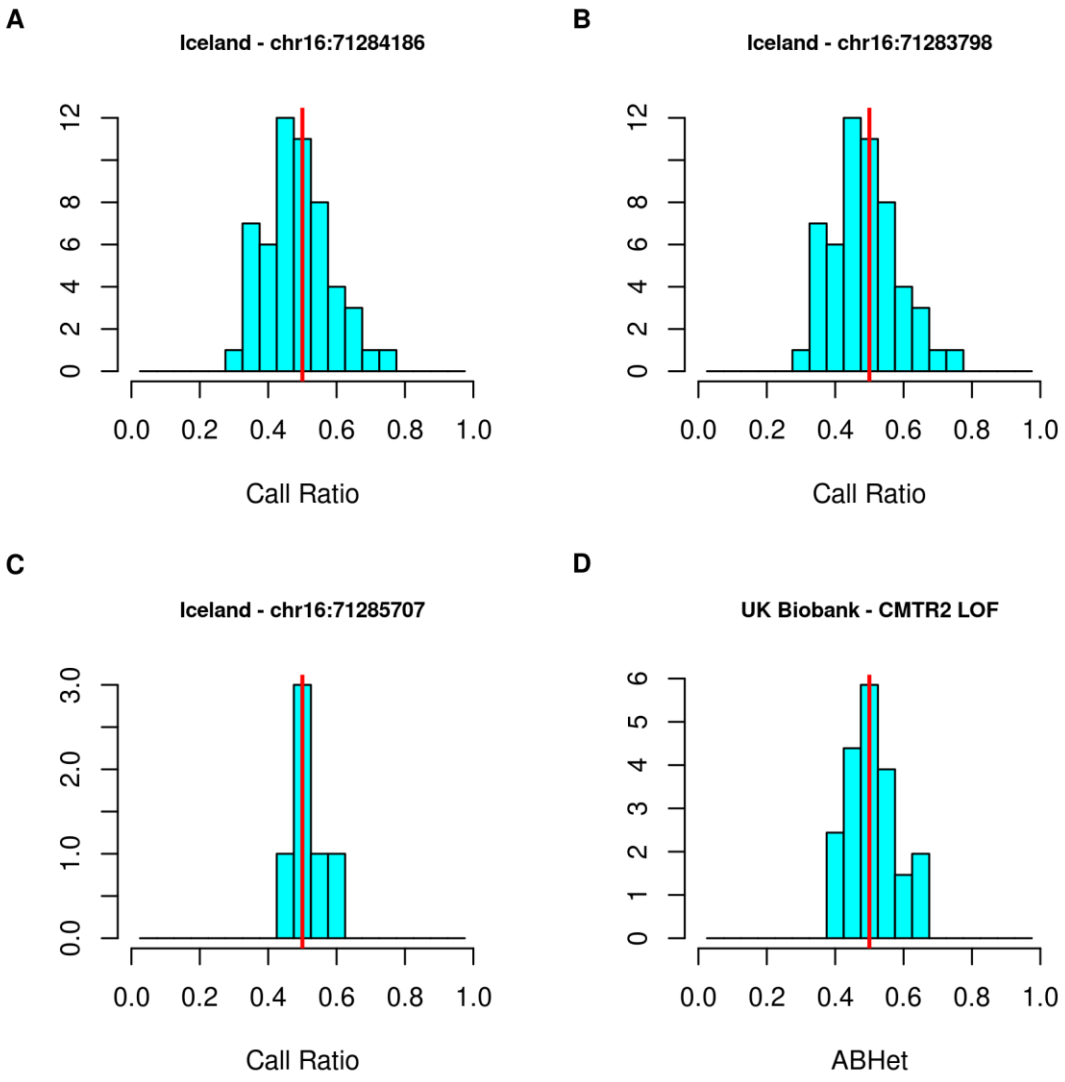
To further validate our novel associations, we re-ran the association analysis using SAIGE[4], which is a generalized mixed model that effectively controls for relatedness. For technical reasons, we can only use SAIGE on our chip-typed dataset, i.e. we can not use our family

imputation to gain statistical power. We therefore, ran our association test using only chip-typed individuals and compared with the results from SAIGE. For the novel genes there is very little difference in the P-values, as is evident in Supplementary Figures 8-9. This further confirms that our significant association results are not due to excessive relatedness.

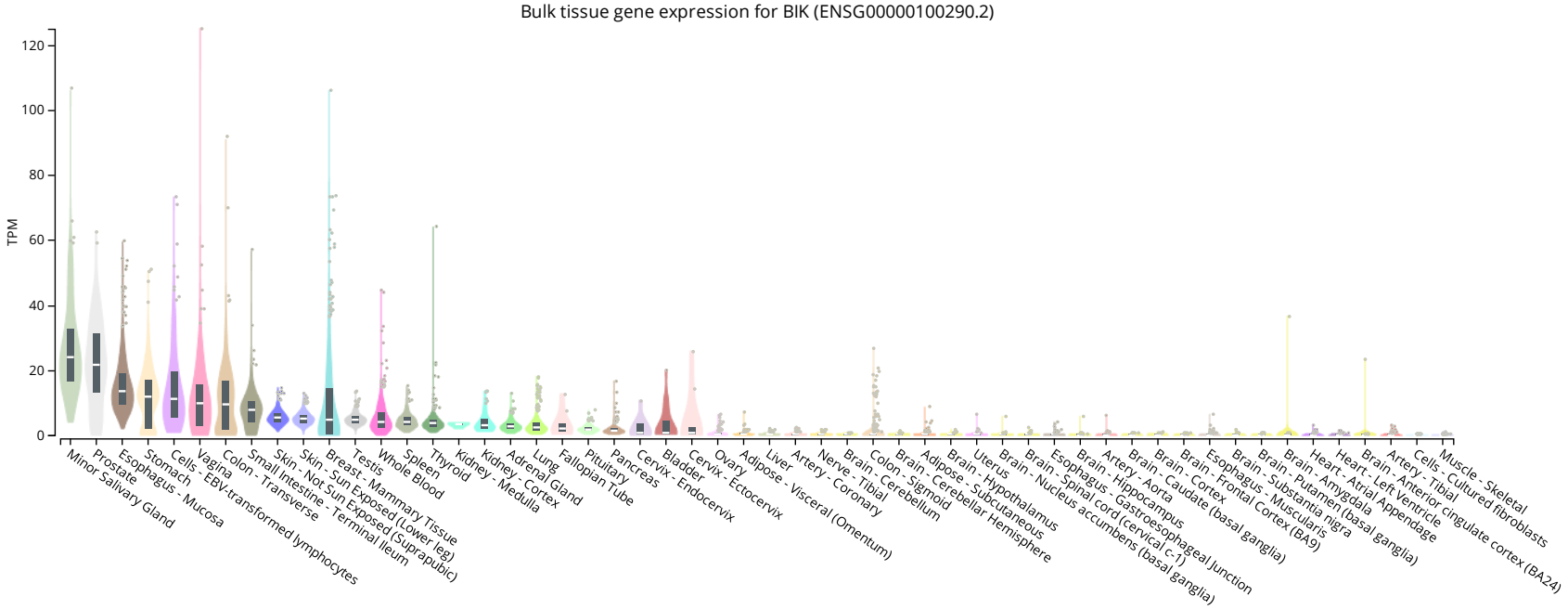
SUPPLEMENTARY FIGURES



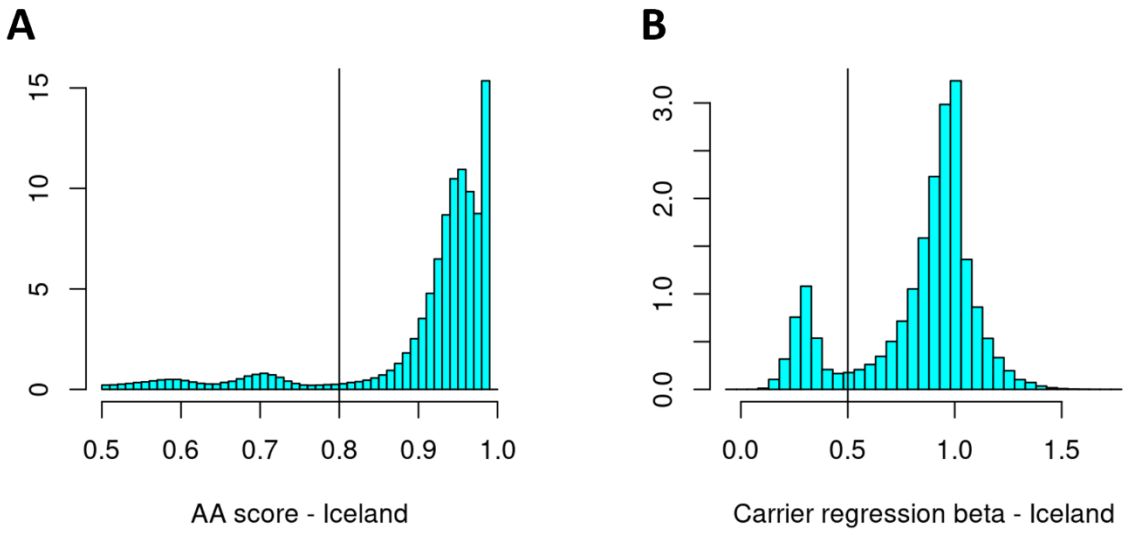
Supplementary Figure 1. Transcript abundance of *ATG12* from RNA sequenced whole blood from carriers of p.Met1Leu (n=38) and non-carriers (n=17,656). Box plots are shown as median and 25th and 75th percentiles. Points are displayed as outliers if they are above or below 1.5 times the interquartile range for non-carriers. All points are displayed for carriers.



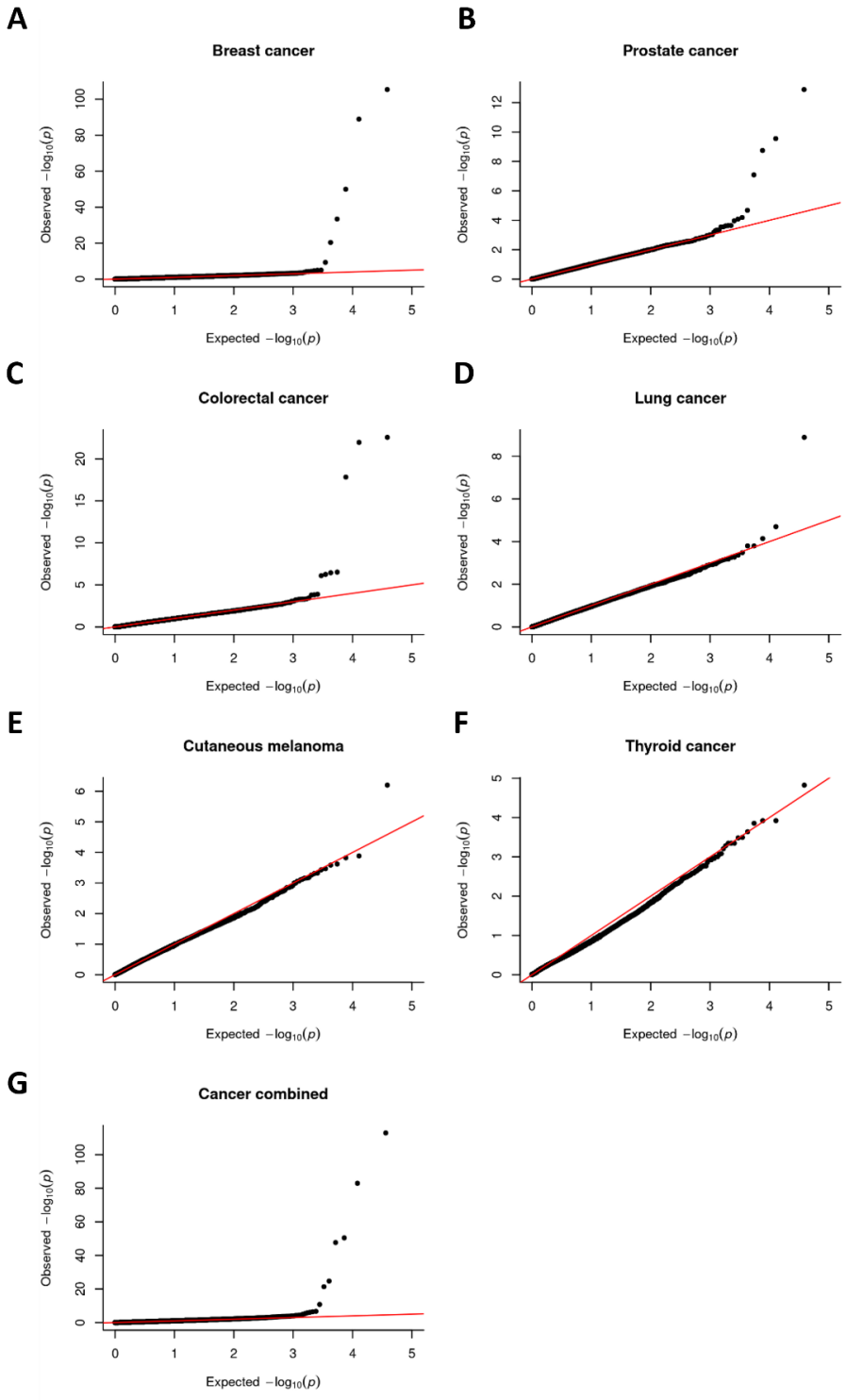
Supplementary Figure 2. The allelic-balance or call ratio distributions for the three Icelandic *CMTR2* LOF variants **A** chr16:71284186, **B** chr16:71283798 and **C** chr16:71285707. For all three variants, the allelic-balance distribution is centered around 0.5, as is expected for germline variants. **D** shows the distribution for the average call ratio for the 53 variants combined for the *CMTR2* LOF burden in the UK Biobank dataset (same as ABHet from Graphtyper), which is centered around 0.5, as is expected for germline variants.



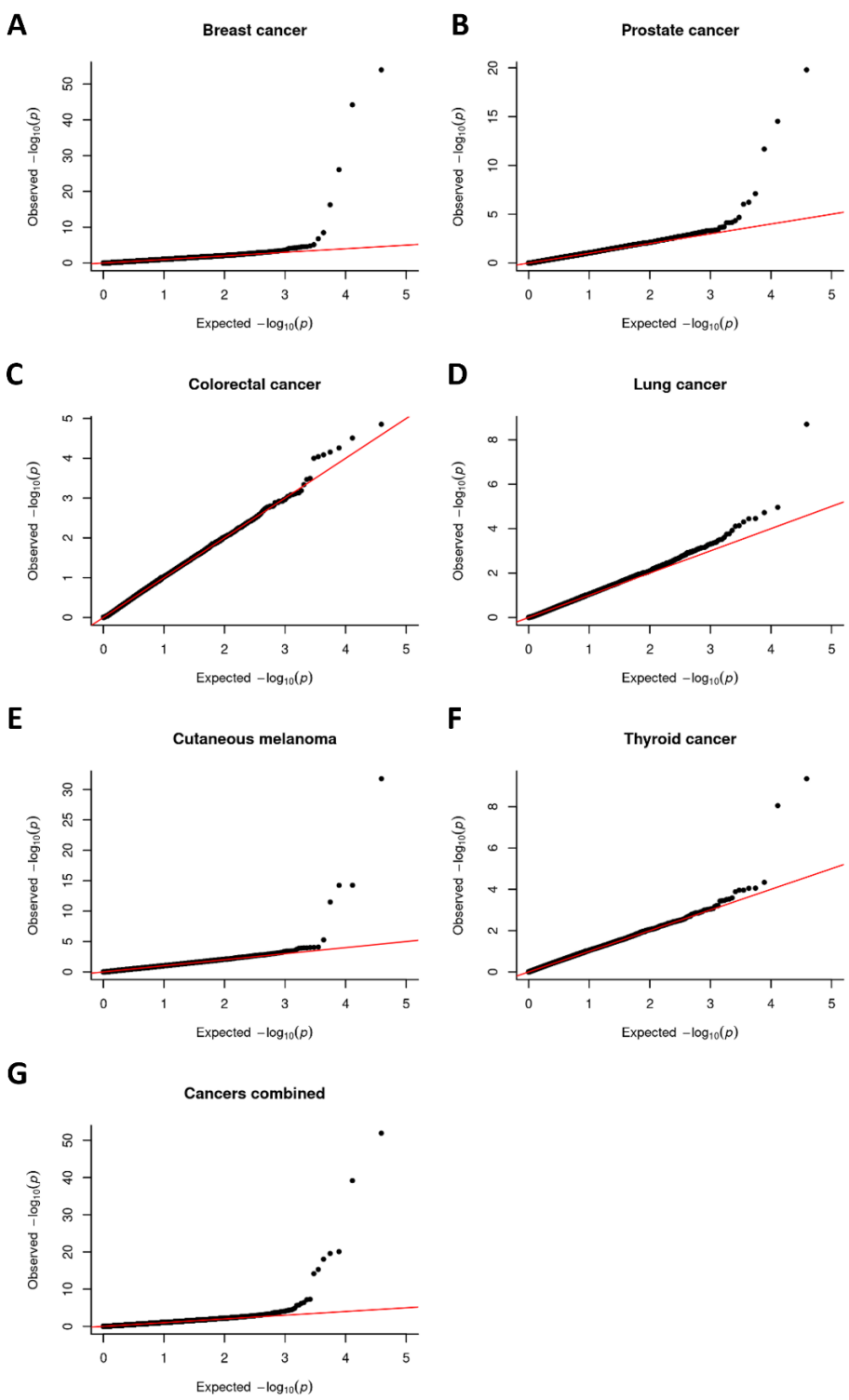
Supplementary Figure 3. Tissue expression for *BIK* obtained from the GTEx database. Expression values are shown in TPM (Transcripts Per Million). Box plots are shown as median and 25th and 75th percentiles. Points are displayed as outliers if they are above or below 1.5 times the interquartile range.



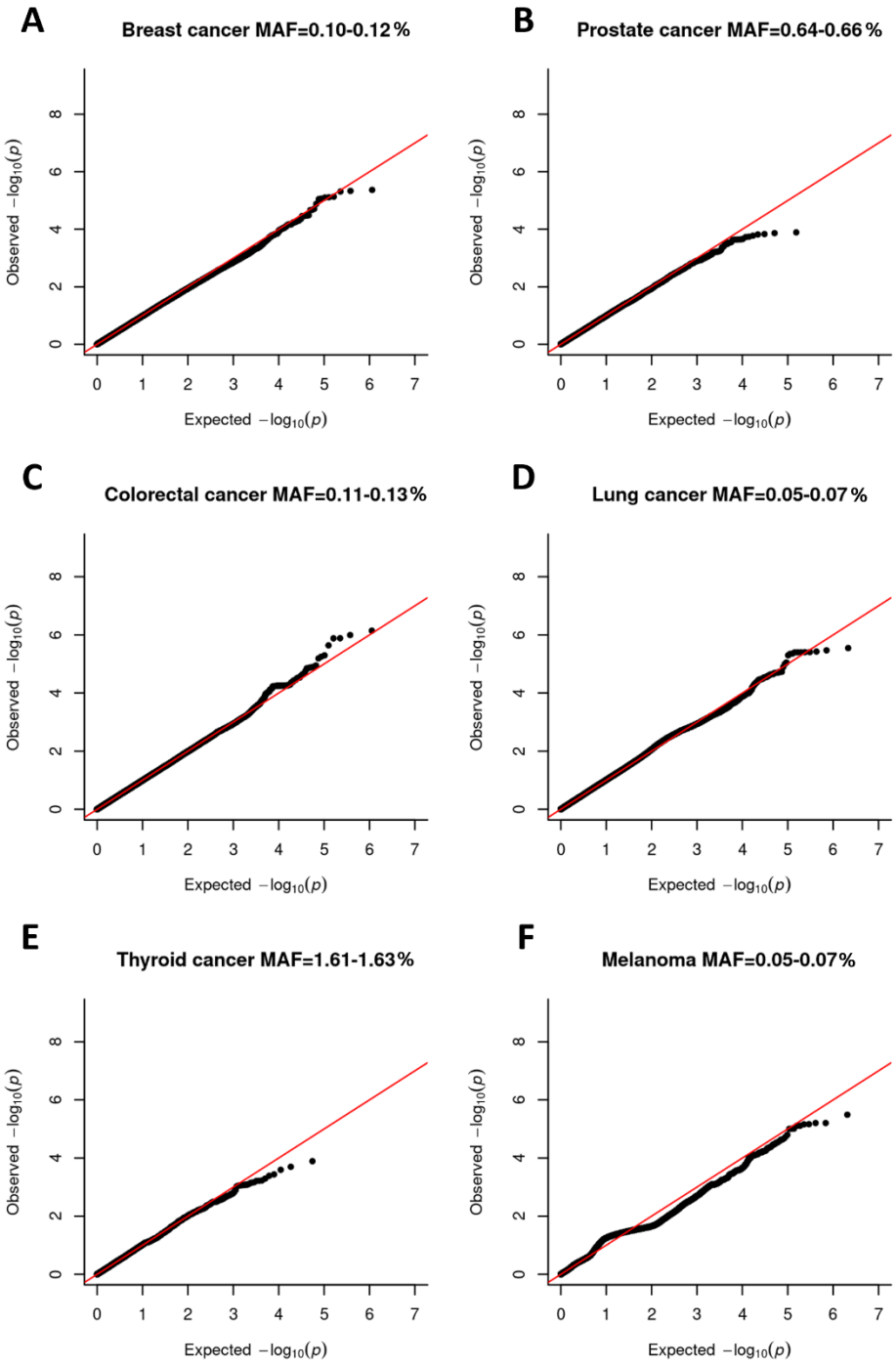
Supplementary Figure 4. Quality control filtering for sequence variants. **A.** The distribution of the AA score, obtained from GraphTyper, for all moderate or high impact sequence variants in the Icelandic data. For Burden analysis we only used variants with AA score>0.8. **B.** The distribution of the carrier regression beta, obtained from GraphTyper, for all moderate or high impact sequence variants in the Icelandic data. For Burden analysis we only used variants with beta>0.5 (Methods).



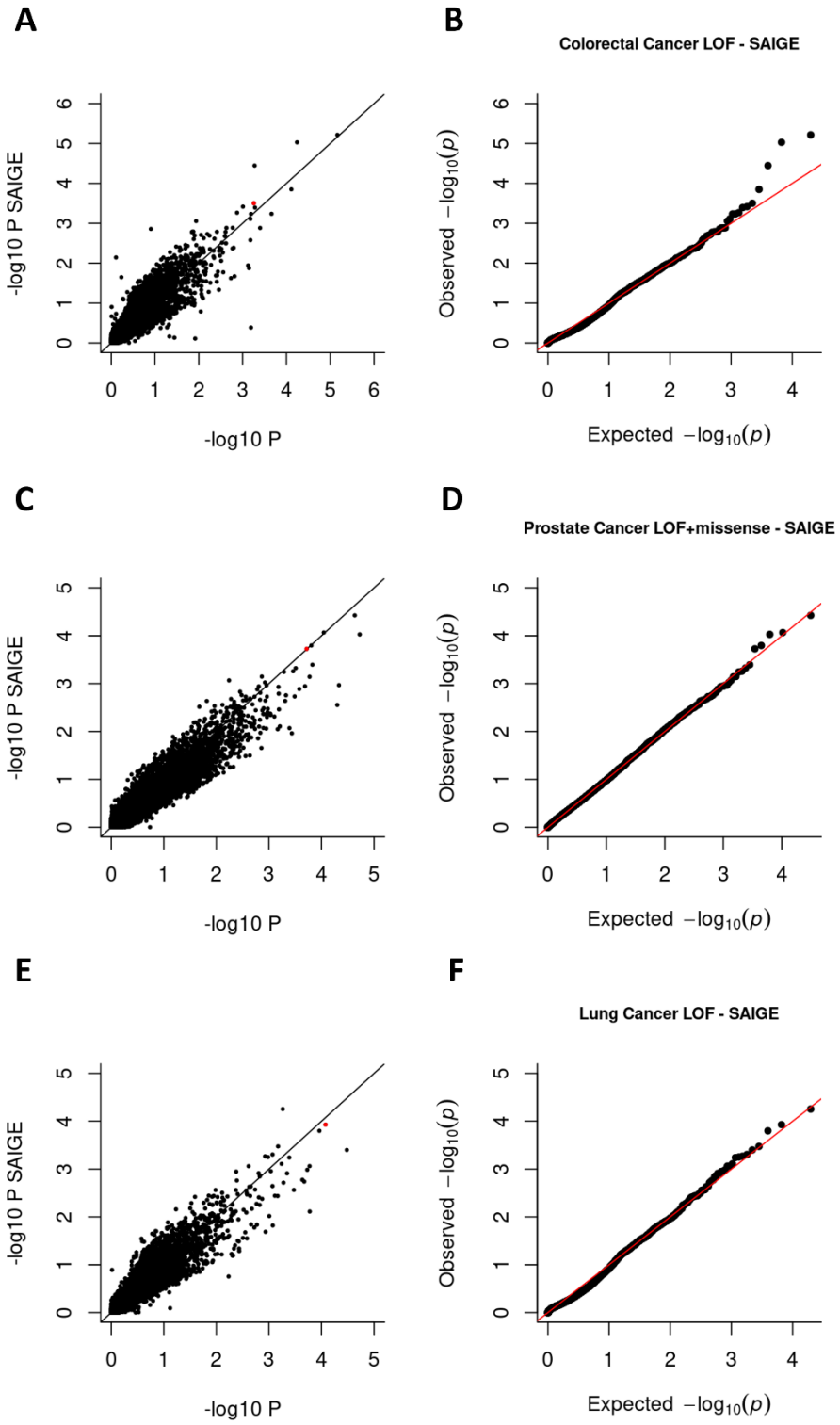
Supplementary Figure 5. QQ plots for the association results from the gene-based LOF burden test for; (**A**) breast cancer, (**B**) prostate cancer, (**C**) colorectal cancer, (**D**) lung cancer, (**E**) cutaneous melanoma, (**F**) thyroid cancer and (**G**) cancers combined. Logistic regression and two-sided likelihood ratio tests were used to test for associations.



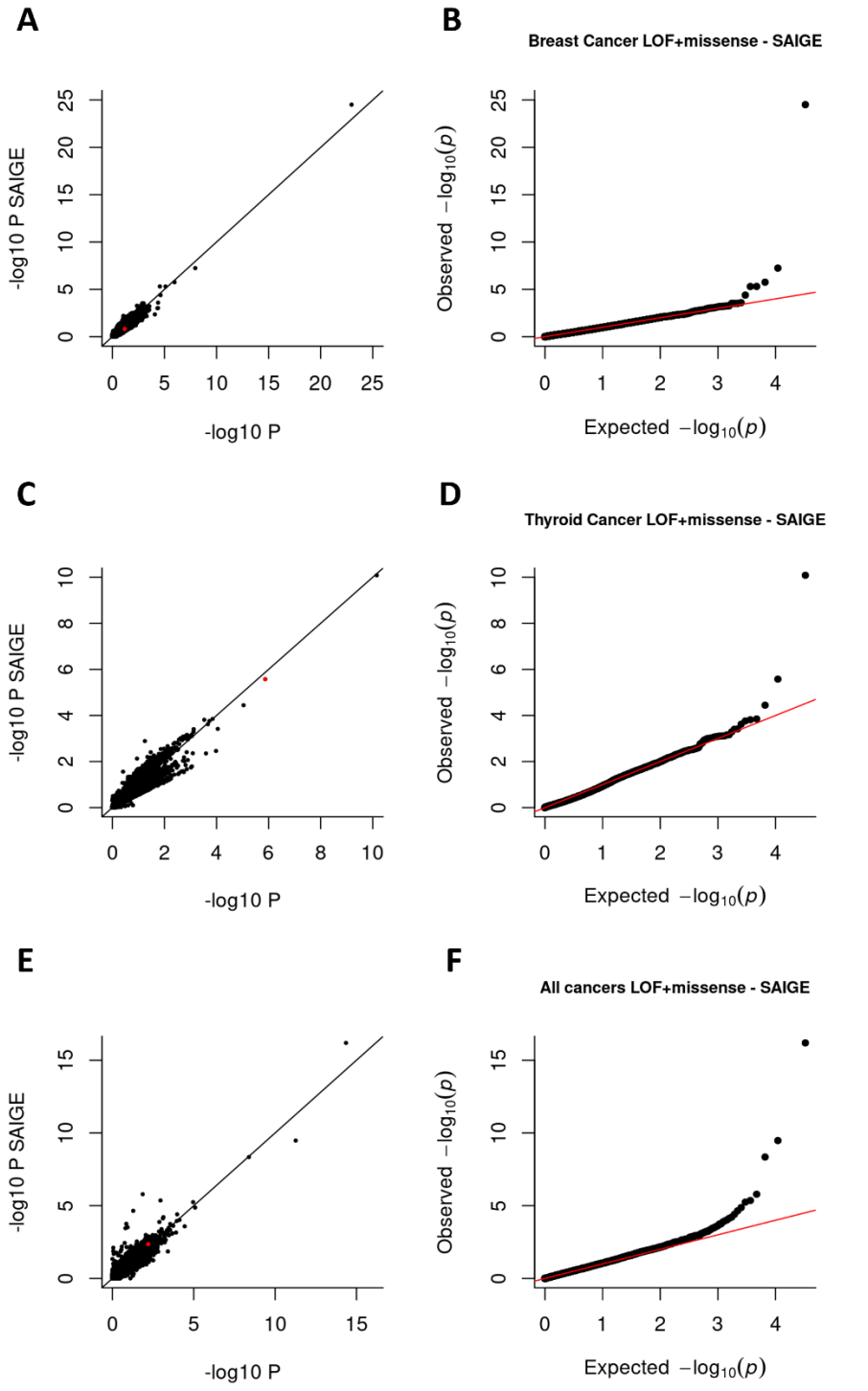
Supplementary Figure 6. QQ plots for the association results from the gene-based LOF+missense burden test for; (**A**) breast cancer, (**B**) prostate cancer, (**C**) colorectal cancer, (**D**) lung cancer, (**E**) cutaneous melanoma, (**F**) thyroid cancer and (**G**) cancers combined. Logistic regression and two-sided likelihood ratio tests were used to test for associations.



Supplementary Figure 7. QQ plots for the association results in the Icelandic dataset, for all intronic and intergenic variants in the genome with minor allele frequencies matching the frequencies of the reported burden genotypes for; (**A**) breast cancer, (**B**) prostate cancer, (**C**) colorectal cancer, (**D**) lung cancer, (**E**) cutaneous melanoma, (**F**) thyroid cancer and (**G**) cancers combined. Logistic regression and two-sided likelihood ratio tests were used to test for associations.



Supplementary Figure 8. Comparing P-values from our likelihood ratio tests against P-values using SAIGE, using chip-typed individuals in Iceland for **A** colorectal-, **C** prostate- and **E** lung cancer. The red dots represent *ATG12* in **A**, *BIK* in **C** and *CMTR2* in **E**. Figures **B**, **D** and **F** are QQ plots for P-values obtained using SAIGE.



Supplementary Figure 9. Comparing P-values from our likelihood ratio tests against P-values using SAIGE, using chip-typed individuals in Iceland for **A** breast-, **C** thyroid- and **E** all cancers. The red dots represent *PPP1R15A* in **A**, *TG* in **C** and *AURKB* in **E**. Figures **B**, **D** and **F** are QQ plots for P-values obtained using SAIGE.

SUPPLEMENTARY REFERENCES

1. Stankovic, S. *et al.* Genetic links between ovarian ageing, cancer risk and de novo mutation rates. *Nature* 1–7 (2024) doi:10.1038/s41586-024-07931-x.
2. Coggins, S. A., Mahboubi, B., Schinazi, R. F. & Kim, B. SAMHD1 Functions and Human Diseases. *Viruses* **12**, 382 (2020).
3. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
4. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).