Analysis of variants in untranslated and promoter regions and breast cancer risk using whole genome sequencing data.

Naomi Wilcox[1], Jonathan P. Tyrer[1], Leila Dorling[1], Joe Dennis[1], Marc Naven[1], Mustapha Abubakar[2], Thomas U. Ahearn[2], Irene L. Andrulis[3, 4], Antonis C. Antoniou[1], Natalia V. Bogdanova[5-7], Stig E. Bojesen[8-10], Manjeet K. Bolla[1], Hiltrud Brauch[11-13], Nicola J. Camp[14], Jenny Chang-Claude[15, 16], Kamila Czene[17], Thilo Dörk[6], D. Gareth Evans[18, 19], Peter A. Fasching[20], Jonine D. Figueroa[2, 21, 22], Henrik Flyger[23], Eugene J. Gardner[24], Anna González-Neira[25], Pascal Guénel[26], Eric Hahnen[27, 28], Per Hall[17, 29], Mikael Hartman[30-32], Maartje J. Hooning[33], Anna Jakubowska[34, 35], Elza K. Khusnutdinova[36, 37], Vessela N. Kristensen[38, 39], Jingmei Li[40], Annika Lindblom[41, 42], Artitaya Lophatananon[43], Arto Mannermaa[44-46], Siranoush Manoukian[47], Roger L. Milne[48-50], Rocio Nuñez-Torres[51], Nadia Obi[52, 53], Mihalis I. Panayiotidis[54], Sue K. Park[55-57], John R.B. Perry[24, 58], Muhammad U. Rashid[59, 60], Emmanouil Saloustros[61], Elinor J. Sawyer[62], Marjanka K. Schmidt[63-65], Melissa C. Southey[48, 50, 66], Amanda B. Spurdle[67], Diana Torres[59, 68], Qin Wang[1], Jacques Simard[69], Soo Hwang Teo[70, 71], Alison M. Dunning[72], Peter Devilee[73, 74], Douglas F. Easton[1, 72].

[1] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
[2] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA.
[3] Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada.
[4] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.
[5] Department of Radiation Oncology, Hannover Medical School, Hannover, Germany.
[6] Gynaecology Research Unit, Hannover Medical School, Hannover, Germany.
[7] N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus.
[8] Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
[9] Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
[10] Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
[11] Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany.
[12] iFIT-Cluster of Excellence, University of Tübingen, Tübingen, Germany.
[13] German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Partner Site Tübingen, Tübingen, Germany.
[14] Department of Internal Medicine and Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA.
[15] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
[16] Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
[17] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

[18] Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK.
[19] North West Genomics Laboratory Hub, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.
[20] Department of Gynecology and Obstetrics, Comprehensive Cancer Center Erlangen-EMN, Friedrich-Alexander University Erlangen-Nuremberg, University Hospital Erlangen, Erlangen, Germany.
[21] Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK.
[22] Cancer Research UK Edinburgh Centre, The University of Edinburgh, Edinburgh, UK.
[23] Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
[24] MRC Epidemiology Unit, Wellcome-MRC Institute of Metabolic Science, University of Cambridge, Cambridge, UK.
[25] Human Genotyping Unit-CeGen, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
[26] Team 'Exposome and Heredity', CESP, Gustave Roussy, INSERM, University Paris-Saclay, UVSQ, Villejuif, France.
[27] Center for Familial Breast and Ovarian Cancer, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany.
[28] Center for Integrated Oncology (CIO), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany.
[29] Department of Oncology, Södersjukhuset, Stockholm, Sweden.
[30] Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore City, Singapore.
[31] Department of Surgery, National University Health System, Singapore City, Singapore.
[32] Department of Pathology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore City, Singapore.
[33] Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands.
[34] Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland.
[35] International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University in Szczecin, Szczecin, Poland.
[36] Institute of Biochemistry and Genetics of the Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia.
[37] Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia.
[38] Department of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway.
[39] Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway.
[40] Human Genetics Division, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore City, Singapore.
[41] Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden.
[42] Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden.

[43] Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK.
[44] Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland.
[45] Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland.
[46] Biobank of Eastern Finland, Kuopio University Hospital, Kuopio, Finland.
[47] Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy.
[48] Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia.
[49] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia.
[50] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia.
[51] Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
[52] Institute for Occupational and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
[53] Institute for Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
[54] Department of Cancer Genetics, Therapeutics and Ultrastructural Pathology, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus.
[55] Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea.
[56] Integrated Major in Innovative Medical Science, Seoul National University College of Medicine, Seoul, Korea.
[57] Cancer Research Institute, Seoul National University, Seoul, Korea.
[58] Metabolic Research Laboratory, Wellcome-MRC Institute of Metabolic Science, University of Cambridge, Cambridge, UK.
[59] Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany.
[60] Department of Basic Sciences, Shaukat Khanum Memorial Cancer Hospital and Research Centre (SKMCH & RC), Lahore, Pakistan.
[61] Department of Oncology, University Hospital of Larissa, Larissa, Greece.
[62] School of Cancer & Pharmaceutical Sciences, Comprehensive Cancer Centre, Guy's Campus, King's College London, London, UK.
[63] Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands.
[64] Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, the Netherlands.
[65] Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands.
[66] Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia.
[67] Population Health Program, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia.
[68] Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia.

[69] Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval Research Center, Québec City, Québec, Canada.
[70] Breast Cancer Research Programme, Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia.
[71] Department of Surgery, Faculty of Medicine, University of Malaya, UM Cancer Research Institute, Kuala Lumpur, Malaysia.
[72] Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK.
[73] Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands.
[74] Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands.

Abstract
Recent exome-wide association studies have explored the role of coding variants in breast cancer risk, highlighting the role of rare variants in multiple genes including *BRCA1, BRCA2, CHEK2, ATM* and *PALB2*, as well as new susceptibility genes e.g., *MAP3K1*. These genes, however, explain a small proportion of the missing heritability of the disease. Much of the missing heritability likely lies in the non-coding genome. We evaluated the role of rare variants in the 5' and 3' untranslated regions (UTRs) of 18,676 genes, and 35,201 putative promoter regions, using whole-genome sequencing data from UK Biobank on 8,001 women with breast cancer and 92,534 women without breast cancer. Burden tests and SKAT-O tests were performed in UTR and promoter regions. For UTR regions of 35 putative breast cancer susceptibility genes, we additionally performed a meta-analysis with a large breast cancer case-control dataset. Associations for 8 regions at $P<0.0001$ were identified, including several with known roles in tumorigenesis. The strongest evidence of association was for variants in the 5'UTR of *CDK5R1* ($P=8.5\times10^{-7}$). These results highlight the potential role of non-coding regulatory regions in breast cancer susceptibility.

Introduction

Genetic susceptibility to breast cancer is known to be conferred by common variants, identified through genome-wide association studies (GWAS), together with rare coding variants conferring higher disease risks. Protein-truncating variants (PTVs) and rare missense variants identified through linkage and targeting sequencing studies in some genes have been well established, including *ATM, BARD1, BRCA1, BRCA2, CHEK2, RAD51C, RAD51D, PALB2* and *TP53*[1]. Recently, whole exome sequencing (WES) analysis identified additional associations for PTVs and rare missense variants in *MAP3K1, LZTR1, ATRIP* and *CDKN2A*[2]. However, these common and rare variants, in aggregate, only explain ~50% of the familial aggregation of the disease. Much of the "missing" heritability is likely to be due to variants in the non-coding genome[2]. Here we explore the association of variants in non-coding untranslated regions (UTRs) and promoter regions with breast cancer risk.

UTRs are regions upstream (5' UTR) and downstream (3' UTR) of the coding sequence of a gene that are transcribed but not translated into protein[3], while promoters are untranscribed regions upstream of genes where proteins bind to initiate transcription[4].

Variants in these regions do not affect the protein-coding sequence but may regulate gene expression. We used whole genome sequencing (WGS) data from the UK Biobank (UKB), to assess the role of rare variants in promoter and UTR regions in all genes. After quality control (QC; see methods), this dataset comprised 8,001 women with breast cancer and 92,534 women without breast cancer (Supplementary Table 1). For 35 putative breast cancer susceptibility genes, we incorporated additional data from 51,494 women with breast cancer and 43,884 women without breast cancer from the BRIDGES dataset[1].

For promoter regions, we considered all rare variants (minor allele frequency<0.001) in promoter regions defined by Ensembl BioMart[5] excluding variants falling in coding sequences. For UTR regions, we considered all rare variants annotated as a UTR variant according to Ensembl Variant Effect Predictor (VEP)[6].

We conducted burden tests for all promoter regions or UTR regions in which there was at least one carrier of a variant in either cases or controls (35,201 promoter regions, 16,381 3' UTRs and 17,185 5' UTRs). These tests, in which variants are collapsed together, can be more powerful than single-variant tests if variants have similar effect sizes[7]. We used logistic regression with all rare variants grouped together and further improved power by incorporating data on family history of breast cancer, as described elsewhere[2]. Since the assumption of similar effect sizes may break down for these regulatory regions[8], for burden analyses with P<0.05, we conducted robust SKAT-O tests, which allow for variants to have different effect sizes and directions of association. We denote P-values from the burden test as P_B, P-values from the robust SKAT-O test as P_S, and P-values from the meta-analysis of burden results for the 35 putative breast cancer susceptibility genes as P_{BM}.

Results

5' UTRs

23 5' UTR regions were associated at $P_B<0.001$ (Figures 1 and 2; Supplementary Table 2), slightly more than the number (17) expected by chance. 21 of these corresponded to an increased risk, compared with ~8.5 which would be expected by chance. SKAT-O robust was tested on 861 5' UTR regions with $P_B<0.05$; 14 of these had $P_S<0.001$ (Figure 3, Supplementary Table 3). Associations with $P_B<1x10^{-4}$ or $P_S<1x10^{-4}$ were observed for: *CDK5R1* ($P_B=8.5x10^{-7}$, $P_S=7.2x10^{-7}$), *MVB12A* ($P_B=6.6x10^{-5}$, $P_s=1.6x10^{-4}$) and *SYNE1* ($P_B=8.0x10^{-5}$, 0.014). None of the five most clearly established breast cancer risk genes had P_B or P_S <0.001 (Supplementary Table 4).

The most significant association, meeting $P<10^{-6}$ for both the burden and robust-SKAT-O test, was *CDK5R1*. UTR variants were in two distinct regions approximately 350bp apart, with 69 variants with positions 32486993-32487160 and 36 variants with positions 32487483-32487620 (Figure 4). The effect sizes did not differ significantly between the two regions (Supplementary Table 5). The interval between the UTR regions, 32487172-32487467 was an intron, which had no association with risk (Supplementary Table 5). A table of the UTR variants in *CDK5R1*, sorted by the number of carriers, is provided in Supplementary Table 6. There were 9 variants at position 32487007 with varying repeats of GCC (combined $P_B=0.0030$, $P_S=0.0017$).

3' UTRs

18 3' UTR regions were associated with breast cancer risk at $P_B<0.001$ (Figures 5 and 6; Supplementary Table 7), similar to the number (16) expected by chance. 11 of these corresponded to an increased risk, compared with ~ 8 which would be expected by chance. SKAT-O robust was tested on 821 3' UTR regions with $P_B<0.05$; 17 of these had associations with $P_S<0.001$ (Figure 3, Supplementary Table 8). Associations with $P_B<1x10^{-4}$ or $P_S<1x10^{-4}$ were observed for: *KCNN3* ($P_B=4.6x10^{-5}$, $P_S=5.6x10^{-4}$) and *ZNF821* ($P_B=2.3x10^{-4}$, $P_B=2.2x10^{-5}$). None of the five most clearly established breast cancer risk genes (*ATM, BRCA1, BRCA2, CHEK2,* and *PALB2*) had P_B or P_S <0.001 (Supplementary Table 9).

UTR UK Biobank and BRIDGES meta-analysis

We obtained additional data on UTR regions for 35 putative breast cancer susceptibility genes that were sequenced in the BRIDGES study[1] and performed a meta-analysis of burden associations across the UK Biobank and BRIDGES datasets. No regions were associated with breast cancer risk at $P_{BM}<0.001$ (Supplementary Table 10). The strongest associations were for the *BARD1* 5' UTR ($P_{BM}=0.0076$) and *CHEK2* 3' UTR ($P_{BM}=0.017$).

Promoters

31 promoter regions were associated at $P_B<0.001$ (Figures 7 and 8; Supplementary Table 11), consistent with the number (35.2) expected by chance. However, 23 of the 31 regions at $P_B<0.001$ correspond to an increased risk, compared with ~17.6 which would be expected by chance. SKAT-O robust was tested on 1,756 promoter regions with $P_B<0.05$; 25 of these had $P_S<0.001$ (Figure 9, Supplementary Table 12). For the most significant associations, we identified possible genes associated with each promoter by searching for those with transcription start site (TSS) within 500bp of the promoter end. For each promoter, 1 gene was identified satisfying this condition.

Associations with $P_B<1x10^{-4}$ or $P_S<1x10^{-4}$ were observed for 3 promoter regions: ENSR00001081062 in 4p14 for with nearest downstream gene *NWD2* ($P_B=8.9x10^{-5}$, $P_S=0.038$), ENSR00000016877 in 1q25.3 with nearest downstream gene *ARPC5* ($P_B=9.9x10^{-5}$, $P_s=2.0x10^{-3}$) and ENSR00000189328 in 5q33.2 with nearest downstream gene *SAP30L* ($P_B=0.0049$, $P_S=9.7x10^{-5}$). None of the five most clearly established breast cancer risk genes had P_B or P_S <0.001 (Supplementary Table 13).

Discussion

Of the regulatory genomic regions we examined, the strongest association with breast cancer risk was for rare variants in the 5'UTR of *CDK5R1*, which reached $P<10^{-6}$ using both the simple burden and robust SKAT-O tests. *CDK5R1* shows increased expression in many malignancies and has been associated with poor prognosis, proliferation, and drug resistance in many cancers including breast cancer[9]. Of the other 5'UTRs associated at P<0.0001, SYNE1 is a nuclear envelope protein critical for cellular structure and signalling, which is downregulated in many malignancies. Loss of function mutations in the gene have been found in ovarian cancer patients, and downregulation of the gene has been associated with increased tumour mutation burden and immune cell infiltration[10, 11]. Hypermethylation of the *SYNE1* promoter has also been associated with tumour aggressiveness in breast

cancer[12] and poor outcomes in gastric cancer[13]. MVB12A forms part of the ESCRT-1 complex, which has potential links to cancer[14]. Two 3' UTR associations reached $P_B<1x10^{-4}$ or $P_S<1x10^{-4}$: *KCNN3* and *ZNF821*. *KCNN3* encodes SK3, also known as KCa2.3, a potassium channel. Increased expression of SK3 is associated with greater breast cancer cell migration[15]. Furthermore, decreased expression levels of *KCNN3* have been associated with drug resistance and poor prognosis for ovarian cancer[16]. ZNF821 interacts with ATM, encoded by the known breast cancer susceptibility gene *ATM* involved in DNA damage signalling and repair[17].

Three promoter regions reached $P_B<1x10^{-4}$ or $P_S<1x10^{-4}$: on 1q25.3 (closest likely target *ARPC5*), on 5q23.2 (*SAP30L*) and on 4p14 (*NWD2*), though none reached $P<10^{-6}$. All were high-confidence promoters for these genes, although other genes may be regulated by these promoters. *ARPC5* encodes one of 7 subunits of the Arp2/3 protein complex, which is overexpressed in a variety of cancers including breast cancer[18]. This protein is thought to be involved in the mechanism controlling tumour cell migration, invasion, and metastasis, by mediating actin polymerisation, and therefore to be closely linked to tumour prognosis[18]. CRISPR studies have additionally shown that loss of *ARPC5* can delay the migration of adherent MDA-MB-231 cells (a triple-negative breast cancer cell line)[19]. SAP30L is a SAP30-like protein that, along with SAP30, is a subunit of the SIN3 protein complex, which has suggested roles in breast cancer progression including gene upregulation affecting cell motility, angiogenesis and lymphangiogenesis[20, 21]. *NWD2* is a paralog of *AAMP* which plays a role in angiogenesis and cellular migration. High expression of *AAMP* has been associated with poor prognosis and metastasis of breast cancer[22].

We considered two types of burden analyses: a simple burden analysis in which all rare variants are considered equivalent, and robust SKAT-O which allows for variation in effect across variants. The latter may be more powerful if only a subset of variants are risk-associated or there are effects in opposite directions, while the simpler test is likely to be more powerful if the effects are similar in magnitude, and also has the advantage that power can be improved by taking family history of the disease into account. Notably, the *CDK5R1* 5'UTR association was the strongest using either method.

Previous GWAS have identified more than 300 common susceptibility loci for breast cancer[23-25], but few of these have been definitively mapped to UTRs or promoters, with the majority apparently located in more distal regulatory regions. A notable exception is the variant rs78378222 in the 3'UTR of *TP53*, which is associated (in opposite directions) with both ER-negative and ER-positive breast cancer[23]. In this analysis, one association reached $P<10^{-6}$, and the excess of positive associations at $P<0.001$ suggests that additional breast cancer susceptibility variants may be present in these regions. Many of the putative associations are in regions regulating plausible cancer-related genes, but further, and even larger, replication studies will be required to validate these associations and provide reliable risk estimates.

Methods

UK Biobank

UK Biobank is a prospective cohort study of more than 500,000 individuals. Detailed information is given elsewhere[26, 27]. WGS data for 200,000 individuals were released in November 2021 and accessed via the UK Biobank DNA Nexus platform. QC metrics were applied to Variant Call Format files as described by Gardener et al, including genotype level filters for depth and genotype quality[28]. Other filters, including samples with disagreement between genetically determined and self-reported sex and excess relatives, were applied as described elsewhere[2]. Cases were defined by having invasive breast cancer (International Classification of Diseases (ICD)-10 code C50) or carcinoma in situ (D05), as determined by linkage to the National Cancer Registration and Analysis Service (NCRAS), or self-reported breast cancer. Both prevalent and incident cases were included. Only breast cancers that were an individual's first or second diagnosed cancer were included as cases. By this definition, 8,001 female and 38 male cases were included.

BRIDGES

For the meta-analysis of UTR regions, we also performed the analysis in the BRIDGES dataset for 35 genes sequenced on a targeted sequencing panel, as described elsewhere[1]. This dataset included 51,494 women with breast cancer and 43,884 women without breast cancer from 43 studies participating in the Breast Cancer Association Consortium (BCAC). Phenotype data were based on the BCAC database v14. Of these 41,609 women with breast cancer were from cohort or population-based case-control studies and unselected for family history. The remainder were from clinic-based studies with some oversampling for familial cases. Details of the sequencing methodology, variant calling and quality control are described in detail elsewhere[1].

Data preparation

Promoter regions were extracted using the Ensembl BioMart[5] data mining tool web page. These regions do not directly identify a specific target gene. To identify the likely corresponding gene for each significant region, we identified genes with TSS within 500bp of the promoter end. UTR regions for genes were similarly extracted from Ensembl BioMart using the R package biomaRt.

Ensemble Variant Effect Predictor (VEP)[6] v101.0 was used to annotate variants within regions of interest. Annotations included the 1000 genomes phase 3 allele frequency, sequence ontology variant consequences and exon/intron number. For each region, the MANE Select transcript[29] was used, if available, otherwise the RefSeq Select transcript was used[30]. Annotation files were used to identify rare variants in promoter regions and variants in UTR regions. PTVs, missense and other coding variants were excluded.

Rare variant analysis

Association analyses were carried out separately for each promoter or UTR region. The main association analyses were burden tests in which genotypes were collapsed into a 0/1 variable based on whether samples carried a variant in the region. For each region we estimated odds ratios and confidence intervals, and derived Z-scores, and two-sided P-values (P_B), using logistic regression. The method used here also incorporates the family history of breast cancer as a surrogate for disease status with weight ½, as explained in detail elsewhere[2]. This model assumes that all variants are associated with the same effect

size and is expected to be most powerful under this assumption. SKAT is a method that allows for different variants within a region to have different effects or no effect and with different effect directions[31]. SKAT is likely to be more powerful when a smaller proportion of variants are causal, or effects have different directions[32]. SKAT-O is an optimal test that is a linear combination of the burden and SKAT statistic, which optimises the weighting of the two tests for each region[26]. In large biobanks with unbalanced case-control ratios, these methods can suffer from inflated type-1 error rates. A robust SKAT-O method was recently developed that accounts for this using Saddle Point Approximation and Efficient Resampling[33]. This method is more computationally intensive than the simple burden test and we therefore only obtain SKAT-O robust P-values (P_S) for promoter or UTR regions with $P_B<0.05$. Exome-wide significance is defined as $P<2.5x10^{-6}$. Here we define a more stringent level at $P<10^{-6}$ given we consider approximately 35,000 promoter regions and 34,000 UTR regions. However, we additionally consider regions with P<0.001 to be of interest.

For promoter regions, 3' UTR regions and 5' UTR regions, separate Manhattan plots and QQ plots were made for the Z-scores and P_B values from the simple burden tests. For regions with $P_B<0.05$ scatter plots summarising P_S and P_B values were made.

For genes in both the UKB and BRIDGES datasets, we combined Z-scores for the UTR regions in a meta-analysis using an inverse variance weighting approach. The combined Z-score was defined as $Z_M = \frac{\sum_j \frac{1}{se_j} Z_j}{\sqrt{\sum_j \frac{1}{se_j^2}}}$. Here, Z_M is the combined z score, Z_j the z-score for study j and se_j is the standard error of $\beta_j = log(OR_j)$.

Statistics and reproducibility

No statistical method was used to predetermine the sample size. The experiments were not randomized, and we did not use blinding. Some samples were excluded for reasons as described in the methods above, for example, for sex discrepancies, excess relatives, or discrepancies with previous genotyping.

Data Availability:

Individual level data for the BRIDGES data are not publicly available due to ethical review board constraints but are available on request through the BCAC Data Access Co-ordinating Committee (BCAC@medschl.cam.ac.uk). Requests for access to UK Biobank data should be made to the UK Biobank Access Management Team (access@ukbiobank.ac.uk).

Code Availability:

Quality Control filtering of vcf files was performed using vcftools v0.1.15, bcftools v1.9, picard v2.22.2 and plink v1.90b, as outlined in the methods. Variants were annotated using Ensembl Variant Effect Predictor v101 with assembly GRCh38. The code for each software is available at the website of each package. Data manipulation and analysis were performed using R-4.3.3 with packages clusterProfiler (4.2.2), data.table (1.14.2), dplyr (1.0.9), dbplyr (2.5.0), gtools (3.9.5), HGNChelper (0.8.9), SKAT (2.2.5), tibble (3.2.1) and tidyr (1.3.1). Plots were created using additional packages ggplot2 (3.5.1) and ggrepel (0.9.5). The code for each of the R packages can be found in their associated vignettes.

Acknowledgments

The research has been conducted using the UK Biobank Resource under Application Number 28126. N.W. was supported by the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK (C14478/A29329), Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London, and the University of Manchester. J.P.T. was supported by Cancer Research UK (PRCPJT-May21\100006 and G110748). Quality control of the UK Biobank sequencing data has been funded by the Medical Research Council (unit programs: MC_UU_12015/2, MC_UU_00006/2). The BRIDGES project was supported by the European Union Horizon 2020 research and innovation programs BRIDGES (grant number, 634935 to P.D A.G.-N., A,M.D. and D.F.E) and B-CAST (633784 to M.K.S. and D.F.E.), the Wellcome Trust (v203477/Z/16/Z to S.H.T ad D.F.E), and Cancer Research UK (C1287/A16563). Details regarding funding of specific BRIDGES studies are provided in the Supplementary Material.

Author contributions:

D.F.E. supervised this work and directed the overall analysis. N.W. performed the statistical analysis. N.W., J.P.T., L.D., J.D., M.N, E.J.G. and J.R.B.P., developed the bioinformatics and computational pipelines. M.K.B., S.B., R.K. and Q.W. led data management within the BCAC. J.C-C. and M.K.S. led working groups within the BCAC. A.M.D. and P.D. directed the BRIDGES project. M.A., T.U.A., I.L. A., A.C.A., S.E.B., M.K.B., H.B., N.J.C., J.C.-C., K.C., T.D., D.G.E., P.A.F., J.D.F., H.F., A.G.-N., P.G., E.H., P. H., M.H,, M.J.H., A.J., V.N.K., J.L., A.Lindblom,

A.Lophatananon, A.M., S.M., R.L. M., N.O., M.I.P., S.K.P., M.U.R., E.S., E.J.S., M.K.S., M.C.S., A.B.S., D.T., Q.W., J.S. and S.H.T. contributed to the design and conduct of the contributing BCAC studies.

N.W. and D.F.E. drafted the manuscript. All authors reviewed and approved the paper.

Competing interests:

JRBP and EJG are employees of Insmed Innovation UK and holds stock/stock options in Insmed Inc. JRBP also receives research funding from GSK and engages in paid consultancy for WW International Inc.

Tables:

NA in the main text

Figure legends:

Figure 1: Manhattan Plot of Z-scores from assessing the association between rare variant carriers in 5' UTR regions and breast cancer risk with the burden test. The x-axis is the chromosomal position, and the y-axis is the Z-score from testing H0 of no association. The blue lines correspond to Z=± 3.29, *P*=0.001, and the red lines correspond to Z=± 4.71, *P*=2.5x10^{-6}. All labelled genes are those with P<0.001. *P*-values are unadjusted for multiple testing.

Figure 2: Quantile-Quantile Plot of P-values from assessing the association between rare variant carriers in 5' UTR regions and breast cancer risk using the burden test. The x-axis is the expected $\log_{10}$ *P*-values from the null hypothesis, the y-axis is the observed $\log_{10}$ P-values. Blue dots correspond to regions associated with increased risk and red dots correspond to regions associated with decreased risk. All P-values are unadjusted for multiple testing.

Figure 3: Scatter Plot comparing P-values from the burden test and the robust SKAT-O test for 3' and 5' UTR regions for regions with burden test P-value<0.05. The x-axis is the -ve $\log_{10}$ *P*-values from the robust SKAT-O test (P_S) and the y-axis is the -ve $\log_{10}$ P-values from the burden test (P_B). Genes with P-value<0.001 for either the burden or robust SKAT-O test are labelled. Blue dots correspond to $(-log_{10}(P_B))^2 + (-log_{10}(P_S))^2 \leq 9$, orange corresponds to $9 < (-log_{10}(P_B))^2 + (-log_{10}(P_S))^2 \leq 16$ and red correspond to $(-log_{10}(P_B))^2 + (-log_{10}(P_S))^2 \geq 16$. All P-values are unadjusted for multiple testing.

Figure 4: Lollipop Plot of the frequency and position of 5' UTR variants for *CDK5R1*. The x-axis is the chromosomal position, and the y-axis is the frequency for cases (red; above y=0) and controls (blue; below y=0). The frequency is calculated for females in the dataset only.

Figure 5: Manhattan Plot of Z-scores from assessing the association between rare variant carriers in 3' UTR regions and breast cancer risk with the burden test. The x-axis is the chromosomal position, and the y-axis is the Z-score from testing H0 of no association. The blue lines correspond to Z=± 3.29, *P*=0.001, and the red lines correspond to Z=± 4.71, *P*=2.5x10^{-6}. All labelled genes are those with P<0.001. *P*-values are unadjusted for multiple testing.

Figure 6: Quantile-Quantile Plot of P-values from assessing the association between rare variant carriers in 3' UTR regions and breast cancer risk with the burden test. The x-axis is the expected $\log_{10}$ *P*-values from the null hypothesis, the y-axis is the observed $\log_{10}$ P-values. Blue dots correspond to regions associated with increased risk and red dots correspond to regions associated with decreased risk. All P-values are unadjusted for multiple testing.

Figure 7: Manhattan Plot of Z-scores from assessing the association between rare variant carriers in promoter regions and breast cancer risk with the burden test. The x-axis is the chromosomal position, and the y-axis is the Z-score from testing H0 of no association. The blue lines correspond to Z=± 3.29, *P*=0.001, and the red lines correspond to Z=± 4.71, *P*=2.5x10^{-6}. All red dots are regions with P<0.001. *P*-values are unadjusted for multiple testing.

Figure 8: Quantile-Quantile Plot of P-values from assessing the association between rare variant carriers promoter regions and breast cancer risk with the burden test). The x-axis is the expected $\log_{10}$ *P*-values from the null hypothesis, the y-axis is the observed $\log_{10}$ P-values. Blue dots

correspond to regions associated with increased risk and red dots correspond to regions associated with decreased risk. All P-values are unadjusted for multiple testing.

Figure 9: Scatter Plot comparing P-values from burden tests and the robust SKAT-O test for promoter regions for regions with burden test P-value<0.05. The x-axis is the -ve $\log_{10}$ *P*-values from the robust SKAT-O test (P_S) and the y-axis is the -ve $\log_{10}$ P-values from the burden test (P_B). Genes with P-value<0.001 for either the burden or robust SKAT-O test are labelled. Blue dots correspond to $(-log_{10}(P_B))^2 + (-log_{10}(P_S))^2 \leq 9$, orange correspond to $9 < (-log_{10}(P_B))^2 + (-log_{10}(P_S))^2 \leq 16$ and red correspond to $(-log_{10}(P_B))^2 + (-log_{10}(P_S))^2 \geq 16$. All P-values are unadjusted for multiple testing.

References:

1. Dorling L*, et al.* Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women. *New England Journal of Medicine* **384**, 428-439 (2021).

2. Wilcox N*, et al.* Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk. *Nature Genetics* **55**, 1435-1439 (2023).

3. Steri M, Idda ML, Whalen MB, Orrù V. Genetic variants in mRNA untranslated regions. *WIREs RNA* **9**, e1474 (2018).

4. Khambata-Ford S*, et al.* Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. *Genome Res* **13**, 1765-1774 (2003).

5. Martin FJ*, et al.* Ensembl 2023. *Nucleic Acids Research* **51**, D933-D941 (2022).

6. McLaren W*, et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

7. Lee S, Gonçalo, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* **95**, 5-23 (2014).

8. Li Z*, et al.* A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods* **19**, 1599-1611 (2022).

9. Dastjerdi S*, et al.* Elevated CDK5R1 expression associated with poor prognosis, proliferation, and drug resistance in colorectal and breast malignancies: CDK5R1 as an oncogene in cancers. *Chemico-Biological Interactions* **368**, 110190 (2022).

10. Harbin LM, Lin N, Ueland FR, Kolesar JM. SYNE1 Mutation Is Associated with Increased Tumor Mutation Burden and Immune Cell Infiltration in Ovarian Cancer. *International Journal of Molecular Sciences* **24**, 14212 (2023).

11. Doherty JA*, et al.* ESR1/SYNE1 polymorphism and invasive epithelial ovarian cancer risk: an Ovarian Cancer Association Consortium study. *Cancer Epidemiol Biomarkers Prev* **19**, 245-250 (2010).

12. Rimner A*, et al.* Syne1 Promoter Hypermethylation as a Predictor of Tumor Aggressiveness in Primary Breast Cancer. *International Journal of Radiation Oncology*Biology*Physics* **72**, S681-S682 (2008).

13. Qu Y, Gao N, Wu T. Expression and clinical significance of SYNE1 and MAGI2 gene promoter methylation in gastric cancer. *Medicine (Baltimore)* **100**, e23788 (2021).

14. Gregor L, Stock S, Kobold S. ESCRT machinery: role of membrane repair mechanisms in escaping cell death. *Signal Transduction and Targeted Therapy* **7**, (2022).

15. Potier M*, et al.* Identification of SK3 channel as a new mediator of breast cancer cell migration. *Molecular Cancer Therapeutics* **5**, 2946-2953 (2006).

16. Liu X, Wei L, Zhao B, Cai X, Dong C, Yin F. Low expression of KCNN3 may affect drug resistance in ovarian cancer. *Molecular Medicine Reports*, (2018).

17. Lee JH, Paull TT. Cellular functions of the protein kinase ATM and their relevance to human disease. *Nat Rev Mol Cell Biol* **22**, 796-814 (2021).

18. Zheng S*, et al.* Role and mechanism of actin-related protein 2/3 complex signaling in cancer invasion and metastasis: A review. *Medicine* **102**, (2023).

19. Sindram E*, et al.* ARPC5 deficiency leads to severe early-onset systemic inflammation and mortality. *Disease Models & Mechanisms* **16**, (2023).

20. Lewis MJ, Liu J, Libby EF, Lee M, Crawford NPS, Hurst DR. SIN3A and SIN3B differentially regulate breast cancer metastasis. *Oncotarget* **7**, 78713-78725 (2016).

21. Bao L*, et al.* SAP30 promotes breast tumor progression by bridging the transcriptional corepressor SIN3 complex and MLL1. *Journal of Clinical Investigation* **133**, (2023).

22. Yukun YIN, Andrew J S, Wen G J. The Impact of Angio-associated Migratory Cell Protein (AAMP) on Breast Cancer Cells In Vitro and Its Clinical Significance. *Anticancer Research* **33**, 1499 (2013).

23. Zhang H*, et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet* **52**, 572-581 (2020).

24. Shu X*, et al.* Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nat Commun* **11**, 1217 (2020).

25. Michailidou K*, et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94 (2017).

26. Collins R. What makes UK Biobank special? *The Lancet* **379**, 1173-1174 (2012).

27. Sudlow C*, et al.* 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**, e1001779 (2015).

28. Gardner EJ*, et al.* Damaging missense variants in IGF1R implicate a role for IGF-1 resistance in the etiology of type 2 diabetes. *Cell Genom* **2**, None (2022).

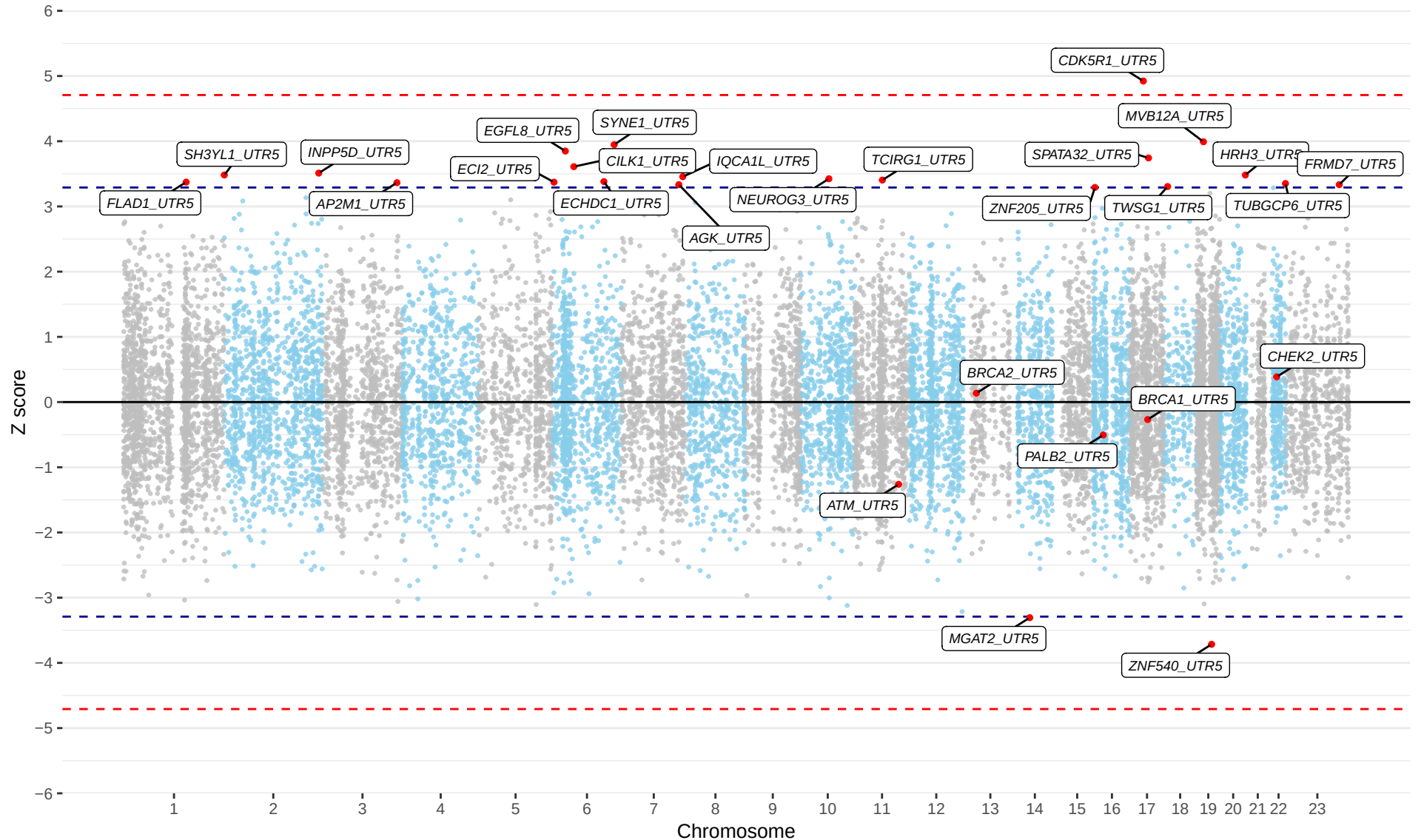
29. Yates AD*, et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682-d688 (2020).

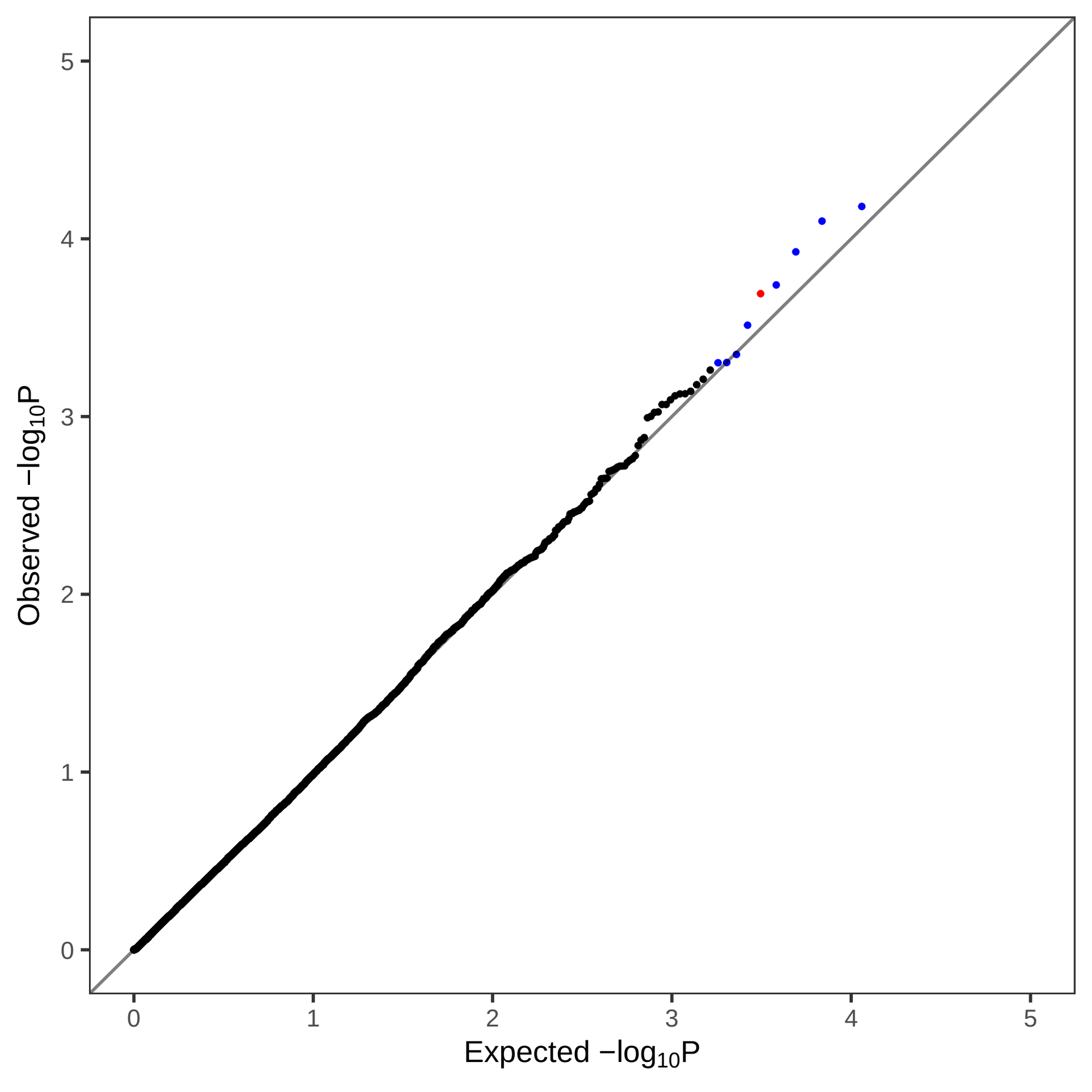
30. O'Leary NA*, et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).

31. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).

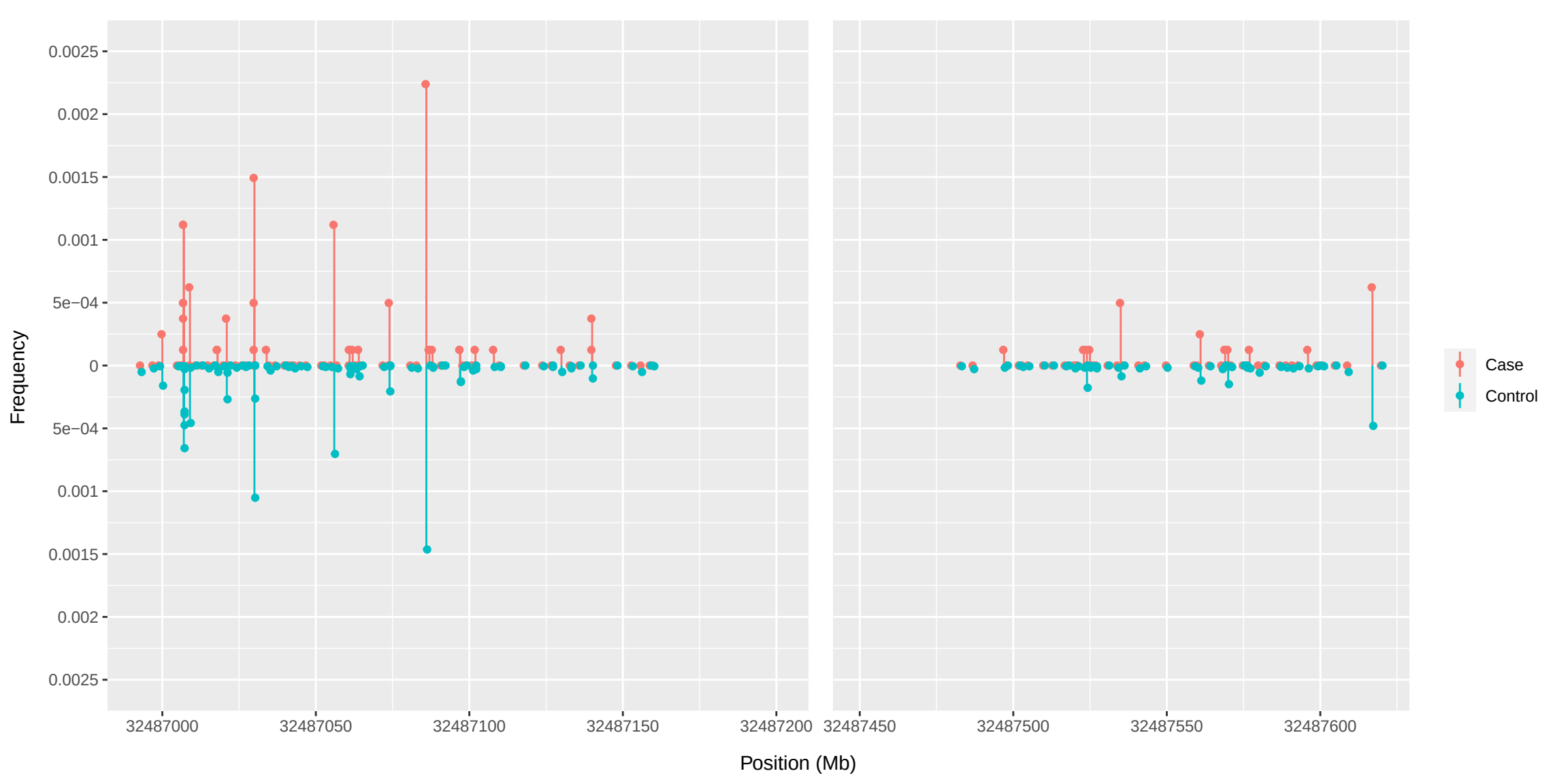
32. Lee S*, et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-237 (2012).

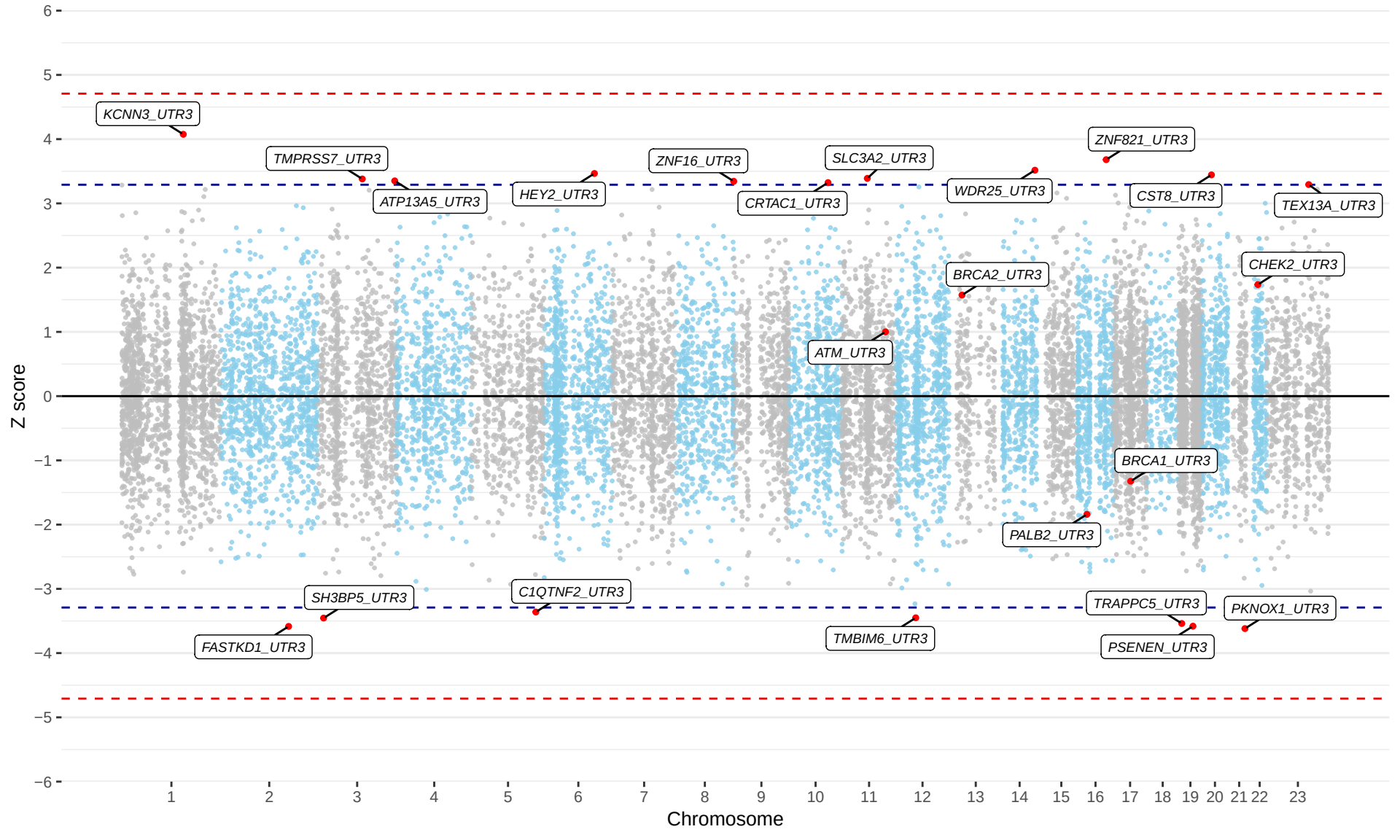
33. Zhao Z, Bi W, Zhou W, VandeHaar P, Fritsche LG, Lee S. UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test. *The American Journal of Human Genetics* **106**, 3-12 (2020).

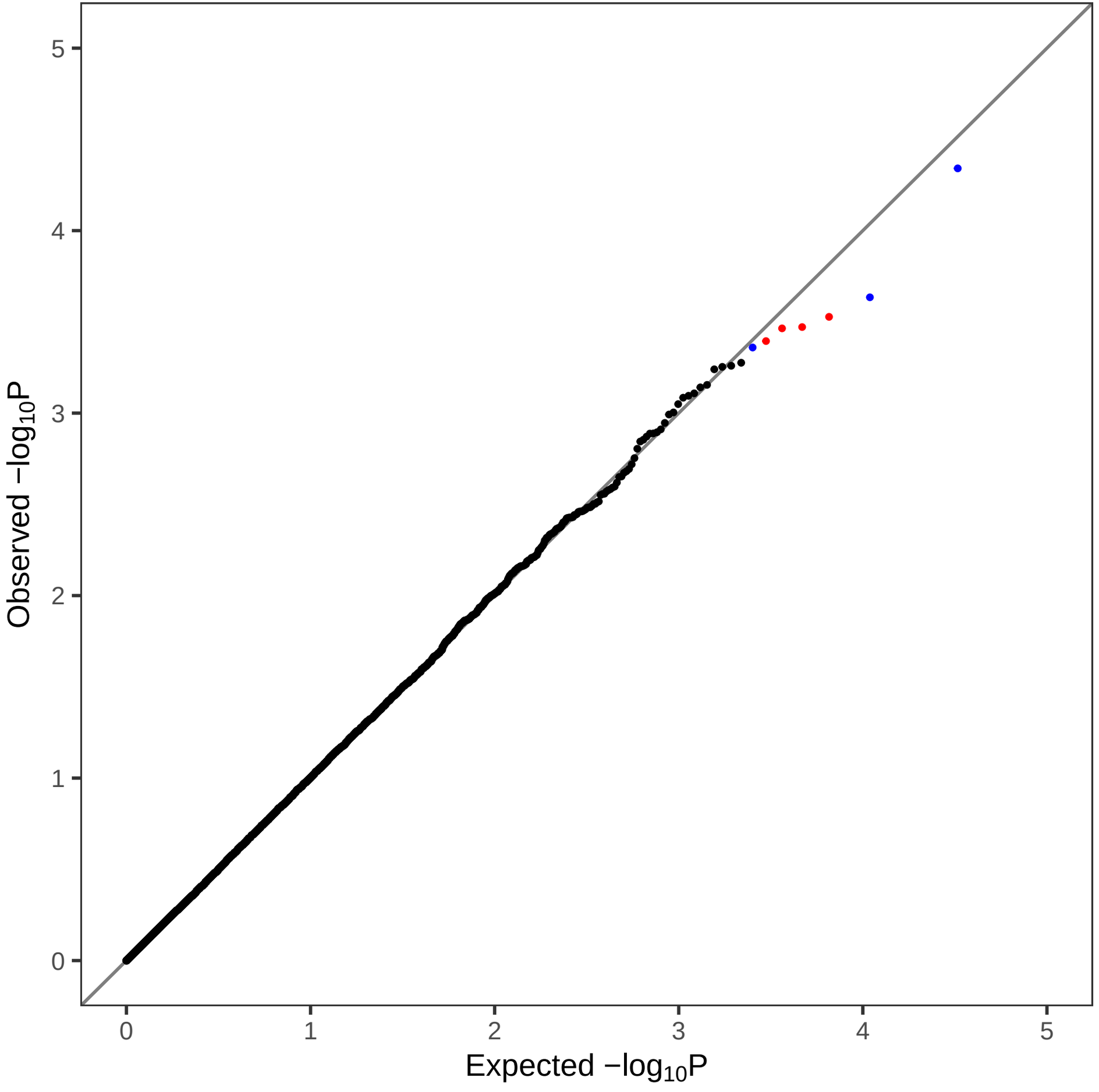

CDK5R1_UTR5
MVB12A_UTR5
SYNE1_UTR5
EGFL8_UTR5
SH3YL1_UTR5
INPP5D_UTR5
CILK1_UTR5
IQCA1L_UTR5
TCIRG1_UTR5
SPATA32_UTR5
HRH3_UTR5
FRMD7_UTR5
ECI2_UTR5
FLAD1_UTR5
AP2M1_UTR5
ECHDC1_UTR5
NEUROG3_UTR5
ZNF205_UTR5
TWSG1_UTR5
TUBGCP6_UTR5
AGK_UTR5
CHEK2_UTR5
BRCA2_UTR5
BRCA1_UTR5
PALB2_UTR5
ATM_UTR5
MGAT2_UTR5
ZNF540_UTR5
Z score
Chromosome
6
5
4
3
2
1
0
−1
−2
−3
−4
−5
−6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

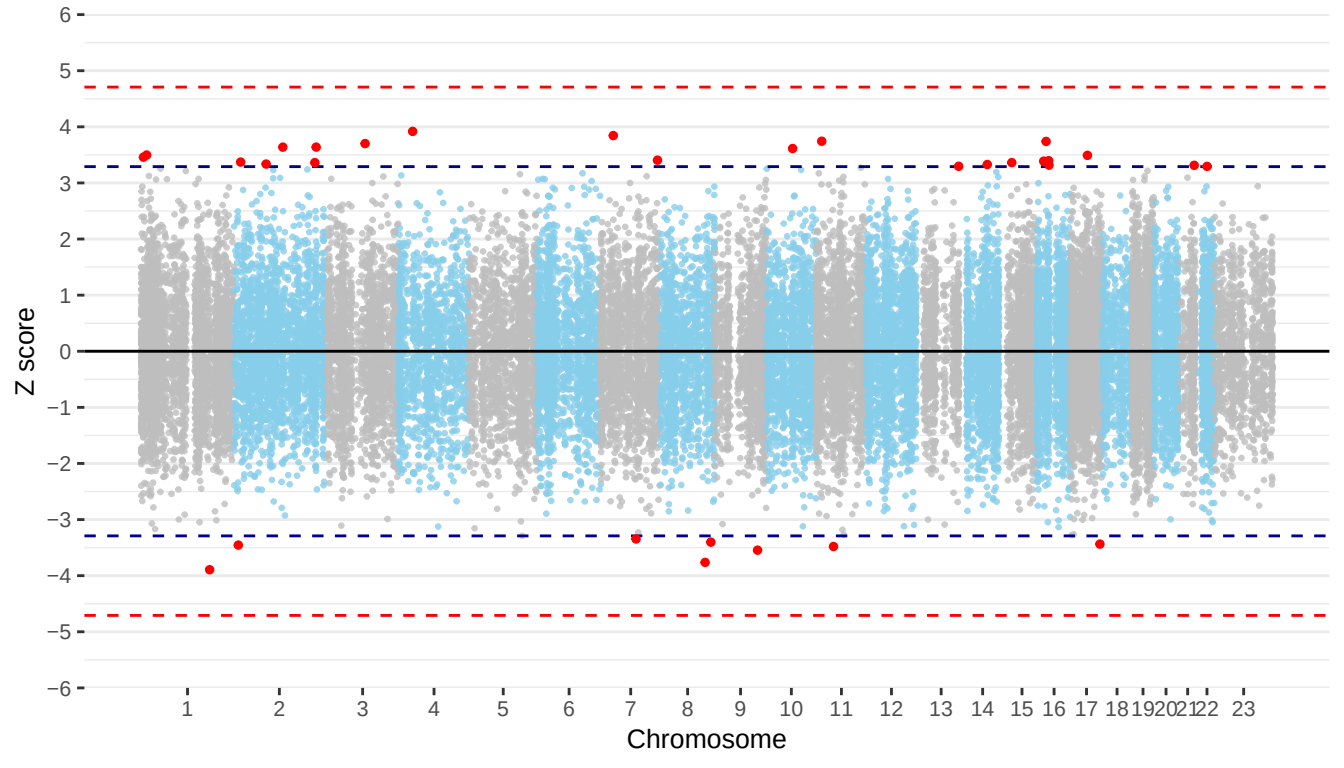

Observed −log10P
Expected −log10P
0
1
2
3
4
5

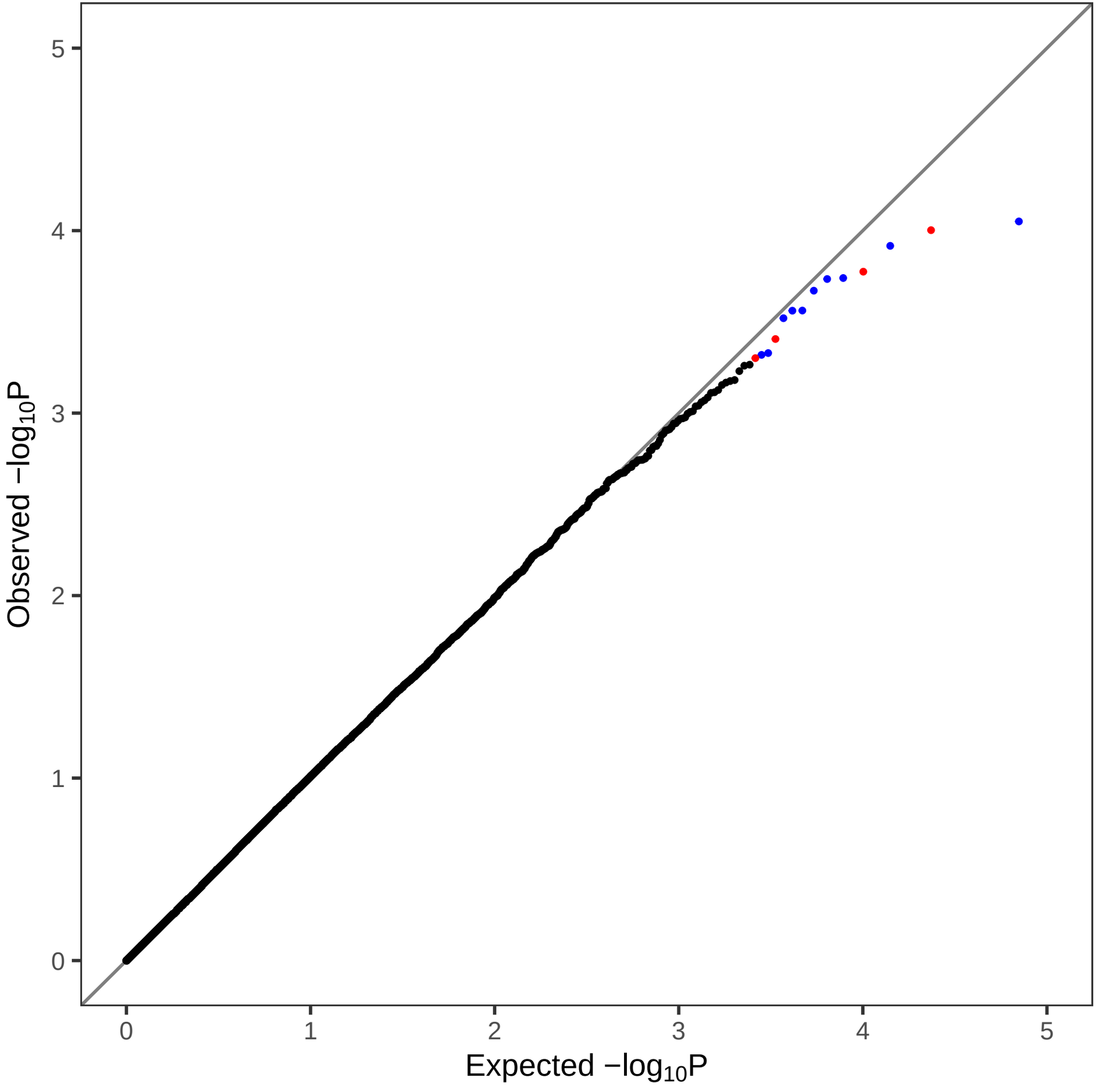

UTR3
UTR5
KCNN3
HEY2
PSENEN
TRAPPC5
SLC3A2
ZNF821
TMPRSS7
CRTAC1
G0S2
CNBP
COL3A1
EPB41L3
CD177
CCL4
TSPAN11
ZGPAT
CDK5R1
SYNE1
EGFL8
MVB12A
ZNF540
IQCA1L
SPATA32
CILK1
AP2M1
ECI2
FRMD7
ZNF74
FLAD1
ZNF205
ZNF106
ZNF546
−log10(p−value)Burden
−log10(p−value)SKATOrobust
0
1
2
3
4
5
6
7

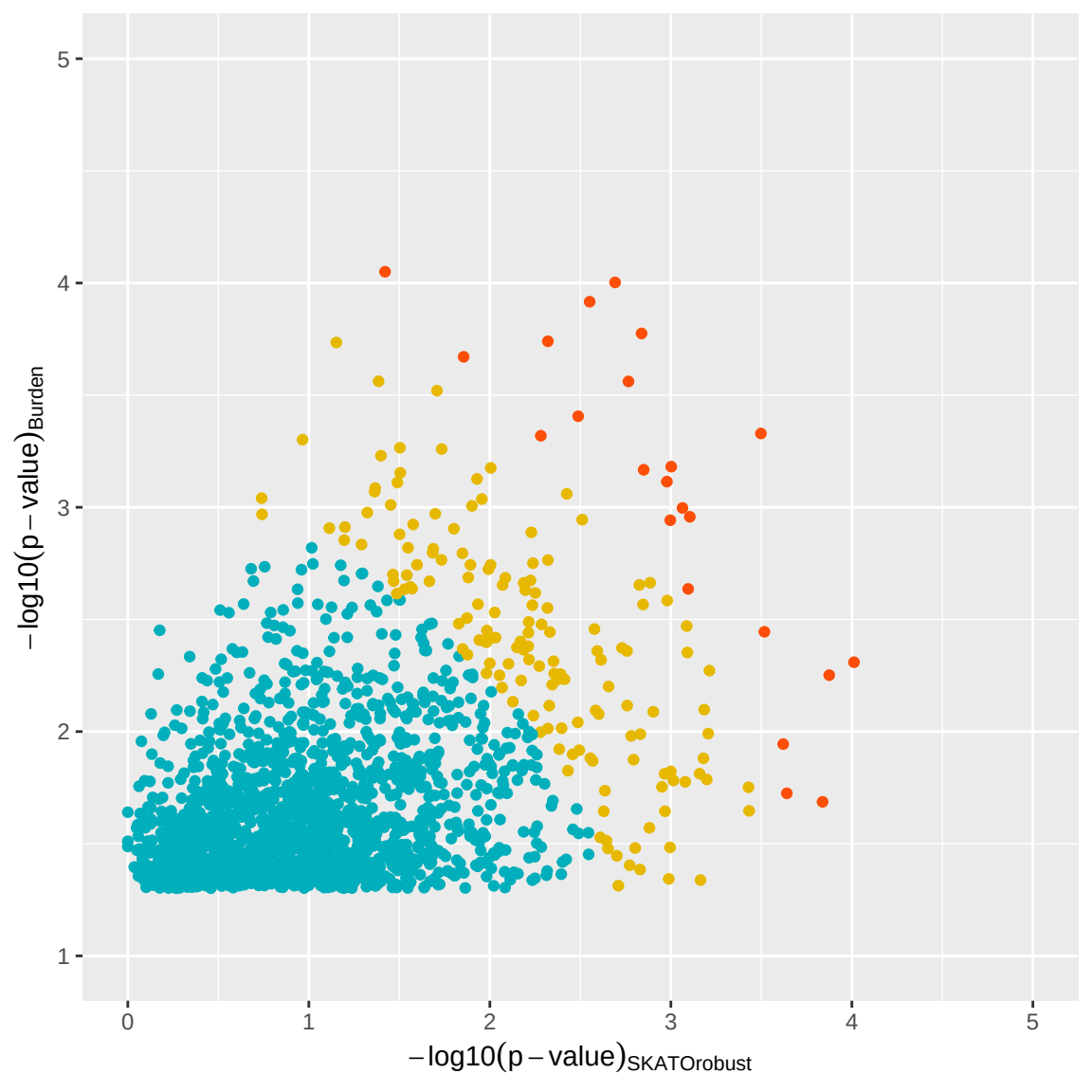

Frequency
0.0025
0.002
0.0015
0.001
5e−04
0
5e−04
0.001
0.0015
0.002
0.0025
32487000
32487050
32487100
32487150
32487200
32487450
32487500
32487550
32487600
Position (Mb)
Case
Control


KCNN3_UTR3
TMPRSS7_UTR3
ATP13A5_UTR3
HEY2_UTR3
ZNF16_UTR3
CRTAC1_UTR3
SLC3A2_UTR3
ATM_UTR3
WDR25_UTR3
BRCA2_UTR3
ZNF821_UTR3
CST8_UTR3
CHEK2_UTR3
TEX13A_UTR3
BRCA1_UTR3
PALB2_UTR3
SH3BP5_UTR3
FASTKD1_UTR3
C1QTNF2_UTR3
TMBIM6_UTR3
TRAPPC5_UTR3
PSENEN_UTR3
PKNOX1_UTR3
Z score
Chromosome


Observed −log10P
Expected −log10P
0
1
2
3
4
5


Z score
6
5
4
3
2
1
0
−1
−2
−3
−4
−5
−6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Chromosome


Observed −log10P
Expected −log10P
0
1
2
3
4
5


5
4
3
2
1
0
1
2
3
4
5
$-\log10(p-value)_{Burden}$
$-\log10(p-value)_{SKATOrobust}$