Supplementary Methods

Study population - age, BMI

An age and BMI-matched common control set was selected for all non-sex specific cancers from MyCode participants who did not have any ICD9/ICD10 code in their encounters diagnoses, inpatient hospitalization-discharge diagnoses or problem-list. The controls for sex-specific cancers including breast, uterine and prostate cancer were selected separately to be sex-matched and having the same number of controls as the common control set.

Age was defined as:

Cases: age at diagnosis

Controls: current age if alive or age at death.

BMI was defined as:

Cases: median value of BMI values recorded in the EHR from a year before diagnosis date

Controls: median value of BMI within recent one year for deceased patients and median value of BMI from a year before current date for alive patients.

TCGA dataset – merge with discovery controls

The TCGA germline dataset was downloaded from https://gdc.cancer.gov/about-data/publications/PanCanAtlas-Germline-AWG. The dataset consisted of 10,389 samples that passed QC criteria. All the patients were part of BLCA (bladder), BRCA (breast), COADREAD (colorectal), KIPAN (kidney), LUAD (lung), LUSC (lung), PRAD (prostate), SCKM (melanoma), THCA (thyroid) and UCEC (uterine) cohorts were retained. Further, only patients

with European ancestry were retained, as all patients in the discovery and replication cohort were of European ancestry. Finally, 4004 samples in 9 cancers with European ancestry were extracted.

The TCGA dataset was aligned to GRCh37, but the DiscovEHR discovery/replication data in this was aligned to GRCh38. So, first all the positions in the TCGA vcf were lifted over to GRCh38. Followed by liftover, the vcf was normalized and multi-allelic variants were split into multiple lines using bcftools ("bcftools norm -m"). The vcf only contained genotype calls where minor alleles were called, all other genotype calls were coded as './.'. We assumed all calls coded as './.' as reference allele and modified the calls to '0/0'. From the 'moderate' discovery dataset, all the controls (common, male only and female only) were extracted. The chromosome, position and alleles were matched (allele flips were checked) between TCGA dataset and the control dataset. Both the datasets were then subset to only retain the variants common between them. Further, wherever necessary the alleles were flipped in the TCGA dataset and both datasets were merged to generate the TCGA moderate dataset. Further, the PLP variants were subset from the TCGA moderate dataset as PLP variants to generate the TCGA PLP dataset. The number of variants in the final datasets is shown in Table S4.

Sequencing and Quality control

In Phase 1, a set of ~60,000 samples were sequenced using NimbleGen probe target-capture (SeqCap VCRome). In phase 2, another set of ~30,000 (the validation dataset) samples were sent to Regeneron genetics center and they were sequenced using a slightly modified version of xGen capture (Integrated DNA Technologies), which had supplemental probes added to capture

regions of the genome well-covered by VCRome capture reagent but poorly covered by xGen. Briefly, the reads were aligned to GRCh38 genome reference using BWA-mem, duplicate reads were identified and flagged using Picard MarkDuplicates tool, followed by variant calling using GATK. The INDEL-realigned and duplicate-marked reads were processed using GATK HaplotypeCaller to generate gVCFs. The joint calling was done in batches of 200 single-sample gVCFs and all pVCF files generated were merged. Further, variants were filtered for QualityByDepth (QD) score <3 and depth <7, and indels for QD <5 and depth <10. SNP sites that did not carry an alternate Allele Balance (AB) >= 15% and indels with AB >= 20% in at least one sample were filtered out. Further, the markers and samples with a call rate below 90% were filtered out. All related patients showing up to 3rd degree relatedness corresponding to IBD > 0.125 were removed.

Power Analysis

The power for this study was calculated using "Power_Logistic_R" function provided in the SKAT package in R.

```
Power_Logistic_R(SubRegion.Length=lengths, N.Sample.ALL = N.Sample.ALL,
        Case.Prop = Case.Prop,
        Causal.MAF.Cutoff = 0.01,
        Causal.Percent = x , N.Sim=100,
        alpha = c(0.05, 5*10^-5, 5*10^-8),
        Prevalence = y, Negative.Percent = 20)
```

The input to SubRegion.Length was vector of gene lengths that were binned in each cancer. The lengths were calculated separately for each cancer. The analysis was run by using varying values for Causal.Percent - 10%, 20%,30% and 50%. Prevalence was calculated for each cancer type

using data from https://seer.cancer.gov/statfacts/. The Negative.Percent, which is the percentage of coefficients of causal variants that are negative was assumed to be 20% for all calculations.