**GWAS Data Sources Documentation**

**Dataset Overview**

This documentation describes the GWAS summary statistics used in the fine-mapping analysis of candidate causal cell types in solid cancers using cis-regulatory elements. The datasets span multiple cancer types with varying sample sizes and represent primarily European ancestry populations.

**Public Data Sources**

**Overall Breast Cancer**

* **Reference PMID**: 29059683
* **Release Year**: 2017
* **Sample Size**: 122,977 cases, 105,974 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 11,792,542
* **Imputation Reference**: 1000G Phase 3
* **Availability**: Publicly available summary statistics

**ER+ Breast Cancer**

* **Reference PMID**: 29059683
* **Release Year**: 2017
* **Sample Size**: 69,501 cases, 105,974 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 11,792,542
* **Imputation Reference**: 1000G Phase 3
* **Availability**: Publicly available summary statistics

**ER– Breast Cancer**

* **Reference PMID**: 29059683
* **Release Year**: 2017
* **Sample Size**: 21,468 cases, 105,974 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 11,792,542
* **Imputation Reference**: 1000G Phase 3
* **Availability**: Publicly available summary statistics

**Ovarian Cancer**

* **Reference PMID**: 28346442
* **Release Year**: 2017
* **Sample Size**: 22,406 cases, 40,941 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 9,870,154
* **Imputation Reference**: 1000G Phase 3
* **Availability**: Publicly available summary statistics

**Prostate Cancer**

* **Reference PMID**: 37945903
* **Release Year**: 2023
* **Sample Size**: 122,188 cases, 604,640 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 27,221,027
* **Imputation Reference**: 1000G phase 3, TOPMed freeze 5, HRC, UK10K, or SISu v3
* **Availability**: Publicly available summary statistics

**Colorectal Cancer**

* **Reference PMID**: 36539618
* **Release Year**: 2022
* **Sample Size**: 78,743 cases, 107,143 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 8,782,440
* **Imputation Reference**: 1000G Phase 1/ Phase 3, HRC, UK10K, or SISu
* **Availability**: Publicly available summary statistics

**Esophageal Cancer**

* **Reference PMID**: 27527254
* **Release Year**: 2016
* **Sample Size**: 4,112 cases, 17,159 controls
* **Ancestry**: European
* **SNPs**: 13,031,593
* **Imputation Reference**: 1000G Phase 1
* **Availability**: Publicly available summary statistics

**Lung Cancer**

* **Reference PMID**: 28604730
* **Release Year**: 2017
* **Sample Size**: 29,266 cases, 56,450 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 7,884,164
* **Imputation Reference**: 1000G Phase 1/ Phase 3
* **Availability**: Publicly available summary statistics

**Renal Cancer**

* **Reference PMID**: 38671320
* **Release Year**: 2024
* **Sample Size**: 25,890 cases, 743,585 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 26,781,105
* **Imputation Reference**: TOPMed imputation version R2, HRC, UK10K
* **Availability**: Publicly available summary statistics

**Restricted Access Data Sources**

**Endometrial Cancer**

* **Reference**: Not available (NA)
* **Release Year**: 2024
* **Sample Size**: 16,353 cases, 228,566 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 9,869,688
* **Imputation Reference**: 1000G Phase 3, UK10K, Estonian and Finnish population-specific imputation references
* **Data Sharing**: **Restricted access** - Data shared by Dr. Tracy O’Mara and is currently awaiting publication/release
* **Usage Terms**: Data used under collaborative agreement; not publicly available at time of analysis

**Bladder Cancer**

* **Reference PMID**: 37210288
* **Release Year**: 2023
* **Sample Size**: 13,790 cases, 343,502 controls
* **Ancestry**: European
* **No. SNPs in meta-analysis**: 9,683,336
* **Imputation Reference**: HRC release 1.1, 1000G Phase 3
* **Data Sharing**: **Restricted access** - Data shared by the original author team led by Dr. Stella Koutros at National Cancer Institute (NCI) through direct collaboration
* **Usage Terms**: Data used under collaborative agreement with NCI research team

**Quality Control Standards**

We performed rigorous quality control to clean and harmonize each GWAS dataset. This involved removing sex chromosomal variants, multi-allelic SNPs, and ambiguous variants; filtering out variants with low imputation quality (INFO score < 0.9), rare variants (minor allele frequency [MAF] < 0.005 in the GWAS), variants with extreme effect sizes (|BETA| > 3), and variants with invalid P-values (P-value ≤ 0 or > 1); and removing duplicate entries from the datasets. Additionally, we harmonized all datasets to the hg19 genome build to ensure consistency across analyses. Finally, we aligned the datasets with the 1000G reference panel, excluding variants absent from the panel, variants with MAF < 0.005 in the reference panel, and variants exhibiting significant allele frequency discrepancies (MAF difference > 0.2) between the GWAS and reference panel populations.

**Data Availability**

All processed GWAS summary statistics data following quality control procedures have been uploaded to this repository, except for the two restricted access datasets (endometrial cancer and bladder cancer) which cannot be shared due to collaborative agreements and pending publication status.

**Data Usage and Citation Guidelines**

**Public Datasets**: All publicly available datasets should be cited using their respective PMIDs and original publication details when results are published.

**Restricted Datasets**:

* Endometrial cancer data courtesy of Dr. Tracy O’Mara (pending publication)
* Bladder cancer data courtesy of Dr. Stella Koutros and team at NCI

**Acknowledgments**: Any publications using these datasets must acknowledge the original study investigators and funding sources as specified in the original publications.

**Contact Information**

For questions regarding data usage or access:

* Public datasets: Refer to original publication contact information
* Endometrial cancer data: Contact Dr. Tracy O’Mara
* Bladder cancer data: Contact Dr. Stella Koutros