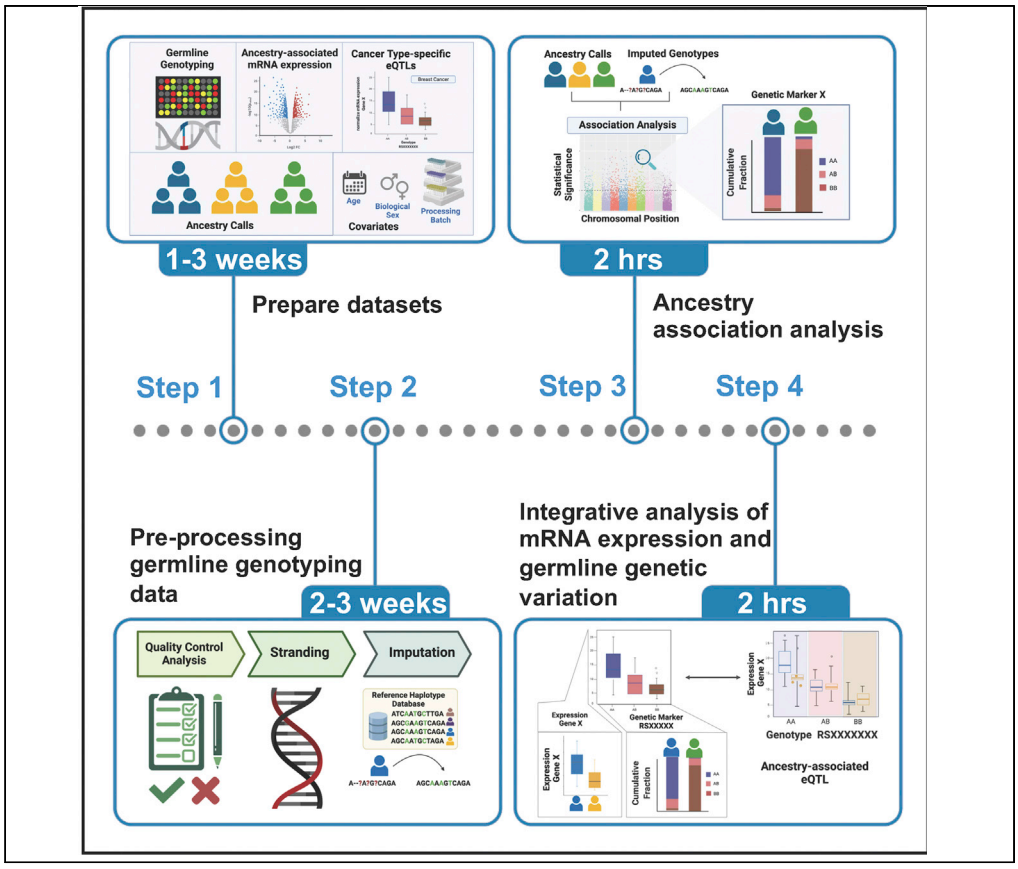
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Analysis of germline-driven ancestry-associated gene expression in cancers



Differential mRNA expression between ancestry groups can be explained by both genetic and environmental factors. We outline a computational workflow to determine the extent to which germline genetic variation explains cancer-specific molecular differences across ancestry groups. Using multi-omics datasets from The Cancer Genome Atlas (TCGA), we enumerate ancestry-informative markers colocalized with cancer-type-specific expression quantitative trait loci (e-QTLs) at ancestry-associated genes. This approach is generalizable to other settings with paired germline genotyping and mRNA expression data for a multi-ethnic cohort.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Nyasha Chambwe, Rosalyn W. Sayaman, Donglei Hu, ..., Elad Ziv, Rameen Beroukhim, Andrew D. Cherniack

rameen_beroukhim@dfci.harvard.edu (R.B.)
achernia@broadinstitute.org (A.D.C.)
rwsayaman@gmail.com (R.W.S.)

Highlights

Protocol for obtaining controlled access TCGA datasets

Protocols for quality control analysis and genotype imputation of TCGA germline data

Statistical analysis for determining ancestry-associated SNPs

Determination of ancestry-associated germline genetic variation driving mRNA expression

Chambwe et al., STAR Protocols 3, 101586 September 16, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.xpro.2022.101586

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Analysis of germline-driven ancestry-associated gene expression in cancers

Nyasha Chambwe,[1,12,13] Rosalyn W. Sayaman,[2,3,4,12,14,*] Donglei Hu,[5] Scott Huntsman,[5] The Cancer Genome Analysis Network, Anab Kemal,[6] Samantha Caesar-Johnson,[6] Jean C. Zenklusen,[6] Elad Ziv,[5] Rameen Beroukhim,[7,8,9,10,11,*] and Andrew D. Cherniack[7,8,9,15,*]

[1]Institute for Systems Biology, Seattle, WA 98109, USA

[2]Department of Laboratory Medicine, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA 94143, USA

[3]Department of Population Sciences, Beckman Research Institute, City of Hope, Duarte, CA 91010, USA

[4]Biological Sciences and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[5]Department of Medicine, Institute for Human Genetics, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA 94143, USA

[6]National Cancer Institute, Bethesda, MD 20892, USA

[7]The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[9]Harvard Medical School, Boston, MA 02115, USA

[10]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

[11]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[12]These authors contributed equally

[13]Present address: Institute of Molecular Medicine, Feinstein Institutes for Medical Research, Manhasset, NY 11030, USA

[14]Technical contact

[15]Lead contact

*Correspondence: rameen_beroukhim@dfci.harvard.edu (R.B.), achernia@broadinstitute.org (A.D.C.), rwsayaman@gmail.com (R.W.S.)
https://doi.org/10.1016/j.xpro.2022.101586

SUMMARY

Differential mRNA expression between ancestry groups can be explained by both genetic and environmental factors. We outline a computational workflow to determine the extent to which germline genetic variation explains cancer-specific molecular differences across ancestry groups. Using multi-omics datasets from The Cancer Genome Atlas (TCGA), we enumerate ancestry-informative markers colocalized with cancer-type-specific expression quantitative trait loci (e-QTLs) at ancestry-associated genes. This approach is generalizable to other settings with paired germline genotyping and mRNA expression data for a multi-ethnic cohort.

For complete details on the use and execution of this protocol, please refer to Carrot-Zhang et al. (2020), Robertson et al. (2021), and Sayaman et al. (2021).

BEFORE YOU BEGIN

The protocol below describes a computational analysis workflow to determine germline genetic variation associated with differences in mRNA expression between groups defined by shared genetic ancestry as shown in (Carrot-Zhang et al., 2020). While this cancer-focused example specifically uses datasets and resources generated by The Cancer Genome Atlas (TCGA) project (Hutter and Zenklusen, 2018), the methods described here are broadly applicable to other disease cohorts for which both germline genotyping and mRNA expression data are assayed for the same individuals.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

In the following sections we describe how to obtain access to and prepare the relevant datasets for integrative analysis. Where appropriate, tools and methods are highlighted to demonstrate how to carry out similar analyses in an independent cohort.

System requirements

This protocol describes workflows that require a high-performance compute environment and data storage capabilities. Ensure that you have adequate computational resources.

Expected run times are dependent on system specifications and availability of computational resources – e.g., communal vs. dedicated resources.

> *Note:* For reference, the "quality control analysis of germline data", "stranding", and "genotype imputation" workflows were performed in the University of California, San Francisco (UCSF) high-performance compute environment which had 8 communal compute nodes and 1 dedicated node, each with 12–64 cores (each node had from 64 to 512 GB of RAM and at least 1.8 TB of fast local disk space). All input and output data were saved in a dedicated storage server with ~200 TB of space. Estimated run times are based on these specifications and the availability of communal nodes.

Apply for dbGaP authorization

Timing: 1–3 weeks

⚠ CRITICAL: dbGAP authorization is necessary to download controlled access TCGA germline data. While the application process is relatively straightforward, the review process can take some time. We recommend applying as soon as is feasible.

1. Verify that your institution has an account. If not, apply for an institutional dbGap account with the relevant institutional officers.
2. Apply for dbGaP authorization to access TCGA controlled access data. See instructions here: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login.
3. Prepare a data access request: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=GeneralAAInstructions.pdf.

Prepare cohort ancestry calls

Timing: 30 min–1 h

Determine the reference ancestral population according to genetic similarity for each individual in the study cohort.

4. Download the TCGA cohort ancestry calls from 'Table S1. Admixture and Ethnicity Calls' from (Carrot-Zhang et al., 2020).
5. Filter out admixed individuals (any individual whose 'consensus_ancestry' matches '<pop>_admix') as they will be excluded from downstream analysis.

> *Note:* For further details on how to determine ancestral population proportions in an independent cohort please refer to (Carrot-Zhang et al., 2020).

> *Note:* The UCSF Ancestry Calls were used for genotype data pre-processing.

Prepare mRNA expression dataset

Timing: 30 min–1 h

Determine mRNA expression associations with ancestry. Perform differential expression analysis between the reference and comparison populations using genetic ancestry labels, taking into account any relevant confounding factors such as cancer type, subtype, age and biological sex. A detailed protocol detailing how to perform this analysis can be found in (Robertson et al., 2021).

6. Download mRNA expression associations with ancestry from 'Table S4. mRNA Associations' from (Carrot-Zhang et al., 2020).
7. Download the sample-level batch-corrected normalized mRNA expression dataset from the NCI Genomic Data Commons (GDC) Pan-cancer Atlas Publications Page (https://gdc.cancer.gov/about-data/publications/pancanatlas) (See key resources table).

Note: Confounding variables include both technical artifacts and other biological or clinical factors that influence the dependent and independent variables under consideration leading to spurious associations (see troubleshooting section, Problem 5). Technical artifacts, such as batch effects, associated with mRNA expression data in TCGA have been addressed through sample-level batch adjustment in the Level 3 normalized processed dataset release.

Prepare cancer expression quantitative trait locus (eQTL) dataset

Timing: 30 min–1 h

8. Download the full pan-cancer atlas cis- and trans-eQTL dataset from PancanQTL (Gong et al., 2018), a comprehensive resource of tumor-type specific eQTLs for all 33 TCGA cancer types.
   a. Each individual cancer-type dataset can be downloaded individually from the PancanQTL portal) from the 'Download' page (http://gong_lab.hzau.edu.cn/PancanQTL).
   b. Alternatively, download the full PancanQTL dataset from synapse (https://www.synapse.org) using one of their programmatic clients (Python/R/Java/Command line)
      i. cis-eQTLs: 'syn12169709'.
      ii. trans-eQTLs: 'syn12169715'.

Prepare germline genetic variation dataset

Timing: Approximately 1–3 weeks. Dependent on server capabilities

Note: The raw TCGA germline genotyping data are distributed via the controlled access mechanism. Instructions on how to obtain controlled access datasets are available at: https://gdc.cancer.gov/access-data/obtaining-access-controlled-data. It may take several weeks to obtain TCGA controlled data access so prepare accordingly.

9. Download Affymetrix Genome Wide SNP 6.0 birdseed genotyping files from normal samples (peripheral blood or normal tissue) and corresponding metadata from the Genomic Data Commons (GDC) TCGA Legacy archive (https://portal.gdc.cancer.gov/legacy-archive)

Note: At sometime after August 2022, TCGA Affymetrix genotyping data will be migrated to the GDC Portal (https://portal.gdc.cancer.gov/).

   a. Select the 'Cases' tab.
      i. Select 'TCGA' for Cancer Program.

ii. Click on 'Add a Case/Biospecimen Filter', and select 'samples.sample_type_id'. Enter sample type codes corresponding to normal samples: 10 (Blood Derived Normal), 11 (Solid Tissue Normal), 12 (Buccal Cell Normal), 14 (Bone Marrow Normal) (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes).

b. Select the 'Files' tab.
   i. Select 'Simple nucleotide variation' from Data Category.
   ii. Select 'Genotypes' from Data Type.
   iii. Select 'Genotyping array' from Experimental Strategy.
   iv. Select 'TXT' from Data Format.
   v. Select 'Affymetrix SNP Array 6.0' from Platform.
   vi. Select 'controlled' from Access Level.
c. Select 'Add all files to the Cart'.

Note: Maximum number of items to add to the 'Cart' is 10,000. For files >10,000, use the 'Case/Biospecimen Filter' to download sample types in batches.

d. Go to the current 'Cart', select 'Metadata' and download the metadata JSON file.
e. From the current 'Cart', select 'Download' and download the Manifest text file.
f. Using the Manifest file, download genotyping files using the GDC Data Transfer Tool (https://gdc.cancer.gov/access-data/gdc-data-transfer-tool).
   i. Follow instructions for downloading the GDC Data Transfer Tool Client.
   ii. Follow instructions from the GDC Data Transfer Tool User's Guide. See "Preparing for Data Download and Upload" and "Data Transfer Tool Command Line Documentation" (https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Getting_Started).

10. Download the Affymetrix SNP Array 6.0 (release 35) annotation file (http://www.affymetrix.com/support/technical/byproduct.affx?product=genomewidesnp_6).

Optional: Alternatively, you can skip this section and request access to the TCGA controlled access pre-processed and Haplotype Reference Consortium (HRC)-imputed datasets through dbGap. Upon approval, you can download the final pre-processed genotyping data (pre- or post-imputation) generated by (Sayaman et al., 2021) from the GDC publication page associated with (Carrot-Zhang et al., 2020). See step 11.

11. Access the "Supplemental Data Files" section of the "TCGA QC HRC Imputed Genotyping Data" generated by (Sayaman et al., 2021) and provided by the AIM AWG from (Carrot-Zhang et al., 2020) (https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020).
   a. Download and read the associated metadata file "READ_ME.txt" for a description of the composition of the genotyping files.
   b. Download the associated "Map_TCGAPatientID_BirdseedFileID.txt" file which describes the file mapping of TCGA Patient IDs to corresponding Birdseed genotyping files.
   c. Choose the genotyping pre-processing level to download genotyping files for the 10,128 unique individuals that passed the pre-imputation QC protocol:
      i. Download the "QC Unimputed Genotyping Data" to access the 838,948 autosomal chromosome variants that passed the pre-imputation QC protocol and its associated "READ_ME_1.txt" file.
      ii. Download the "HRC Stranded Genotyping Data" to access the 680,389 correctly matched Haplotype Consortium Reference (HRC) variants that remain after pre-imputation QC, removal of palindromic SNPs and stranding to the HRC (v1.1) panel (Carrot-Zhang et al., 2020) and its associated "READ_ME_2.txt" file.
      iii. Download the "HRC Imputed Genotyping Data" to access the 39,127,678 SNPs imputed to the HRC panel (v1.1) and its associated "READ_ME_4.txt" file.

Optional: To impute to the 1000 Genomes Project reference panel (1000G) (1000 Genomes Project Consortium et al., 2015) rather than the HRC panel used in (Gong et al., 2018), download the "1000G Stranded Genotyping Data" to access the 678,304 correctly matched 1000G variants that remain after pre-imputation QC, removal of palindromic SNPs and stranding to the 1000G Phase 3 (version 5) panel and its associated "READ_ME_3.txt" file.

d. To download the controlled access data, follow instructions under the "Instructions for Data Download" for "Controlled Access Data".
   i. The necessary manifest files are found under the "Data in the GDC" section for "Controlled Access Data".

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA Pan-cancer Atlas normalized mRNA expression	NCI Genomic Data Commons (GDC), Pancan Atlas Portal	http://api.gdc.cancer.gov/data/3586c0da-64d0-4b74-a449-5ff4d9136611
PancanQTL - Pan-cancer eQTL Database	(Gong et al., 2018) Synapse download cis-eQTLs: 'syn12169709' trans-eQTLs: 'syn12169715'	http://gong_lab.hzau.edu.cn/PancanQTL/ https://www.synapse.org/#!Synapse:syn12169709 (Carrot-Zhang et al., 2020)
TCGA Pan-cancer Atlas Ancestry Calls	Table S1 from (Carrot-Zhang et al., 2020)	https://ars.els-cdn.com/content/image/1-s2.0-S1535610820302117-mmc2.xlsx
TCGA mRNA associations with ancestry	Table S4 from (Carrot-Zhang et al., 2020)	https://ars.els-cdn.com/content/image/1-s2.0-S1535610820302117-mmc5.xlsx
TCGA Germline Whitelisted Samples	Table S1 from (Sayaman et al., 2021)	https://www.cell.com/cms/10.1016/j.immuni.2021.01.011/attachment/edb228c3-a345-4292-9f60-8b18f2852bbe/mmc2.xlsx
TCGA Germline Data - Affymetrix Genome-wide SNP 6.0 array	Genomic Data Commons Legacy Archive	(Carrot-Zhang et al., 2020; Sayaman et al., 2021)
TCGA QC'ed and HRC-Imputed Data	(Carrot-Zhang et al., 2020; Sayaman et al., 2021)	https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020
Haplotype Reference Consortium Reference Dataset	Haplotype Reference Consortium (McCarthy et al., 2016)	https://www.haplotype-reference-consortium.org
Software and algorithms		
Hail v0.2	N/A	https://hail.is/docs/0.2/getting_started.html
PLINK v1.9	(Chang et al., 2015; Purcell et al., 2007)	http://www.cog-genomics.org/plink/1.9/
bcftools 1.9	(Danecek et al., 2021)	https://samtools.github.io/bcftools/
McCarthy Group Tools	N/A	https://www.well.ox.ac.uk/~wrayner/tools/
Michigan Imputation Server	(Das et al., 2016)	https://imputationserver.sph.umich.edu
Eagle v2.3	(Loh et al., 2016)	https://data.broadinstitute.org/alkesgroup/Eagle
Minimac3	(Das et al., 2016; Fuchsberger et al., 2015; Howie et al., 2012)	https://genome.sph.umich.edu/wiki/Minimac3
NCI Genomic Data Commons (GDC) Data Transfer Tool	NCI Genomic Data Commons (GDC)	(Sayaman et al., 2021)
Other		
Custom scripts	(Sayaman et al., 2021)	https://github.com/rwsayaman/TCGA_PanCancer_Genotyping_Imputation https://doi.org/10.5281/zenodo.6658317

MATERIALS AND EQUIPMENT

Software installation

Timing: 1 h

This protocol describes the manipulation of large-scale genomic datasets on the order of millions of genetic markers in thousands of samples. We used Hail (Hail v0.2. https://github.com/hail-is/hail),

an open-source Python library that supports scalable analyses of very large datasets, to perform genetic association testing.

- Install the Hail framework and the appropriate dependent libraries according to the guidelines provided in the library documentation (https://hail.is/docs/0.2/getting_started.html).
- Install the PLINK software (version 1.9 or current version) (Chang et al., 2015; Purcell et al., 2007).
- Install the BCFtools software (version 1.9 or current version) (Danecek et al., 2021) (https://samtools.github.io/bcftools/).
- Install the R programming software (version 3.5.0 or current version) (https://www.r-project.org/).

STEP-BY-STEP METHOD DETAILS

Quality control analysis of germline data

Timing: Approximately 1–2 weeks. Dependent on server capabilities

This section describes quality control (QC) assessment of the TCGA Affymetrix Genome-Wide SNP 6.0 germline genotyping data (Figure 1) using PLINK to generate a high-quality set of SNPs for all whitelisted TCGA samples (i.e., a list of platform-specific samples verified to be appropriate for use). See key resources table.

Review other resources for suitable QC steps based on the study design (Anderson, 2011; Anderson et al., 2010; Aron and Choudhury, 2015).

Note: Original QC steps were performed in PLINK version 1.9. QC analysis requires a high-performance compute cluster.

Note: To skip this step, download the controlled access "QC Unimputed Genotyping Data" generated from (Sayaman et al., 2021), as described in step 11 of the "prepare germline genetic variation dataset" section of this protocol OR proceed with the QC protocol steps provided below.

Note: Scripts used in this section are available at: https://github.com/rwsayaman/TCGA_PanCancer_Genotyping_Imputation.

1. Map birdseed genotyping file names to corresponding TCGA aliquot barcode using the download annotation JSON file from GDC TCGA Legacy archive.
2. Verify sample list for inclusion in the analysis and filter out samples which are not represented in the whitelist, and which do not pass the analyte code filter.
   a. Cross-reference sample set with whitelisted germline samples from GDC PanCanAtlas Publications page (https://gdc.cancer.gov/about-data/publications/pancanatlas). Non-whitelisted samples have since been flagged for withdrawal in the various TCGA projects.
      i. Download the Merged Sample Quality Annotations file (merged_sample_quality_annotations.tsv).
      ii. To select whitelisted samples, filter for samples with "platform" column set to "Genome_Wide_SNP_6" and the "Do not use"' column set to "FALSE".
   b. Based on established TCGA barcode identifiers, ensure all whitelisted samples have Analyte code "D" (DNA). Exclude samples with other Analyte codes.

Note: The final TCGA whitelisted samples used in this analysis are available from (Sayaman et al., 2021), Table S1. The GDC Genome Wide SNP 6.0 platform whitelisted files included samples with TCGA analyte barcode identifiers annotated "D" (DNA) or "G" (Whole Genome Amplification). Samples with analyte barcode identifier "G" were excluded from our analysis.

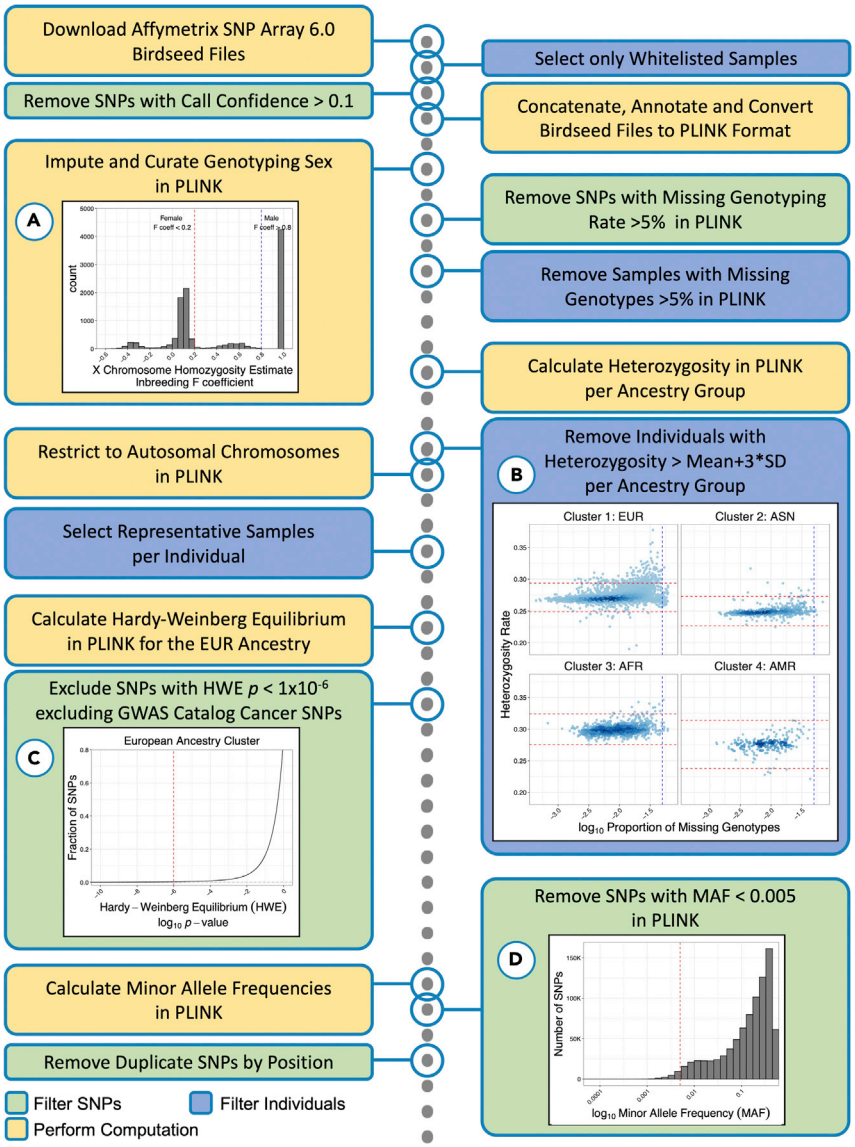


Figure 1. Schematic overview of genotype quality control workflow

Stepwise description of pre-processing steps taken to generate clean quality-controlled germline genotyping data for stranding and imputation. The protocol requires specific calculations to be performed (yellow), and steps to filter SNPs (green) or individuals (blue).

(A) Histogram of X chromosome homozygosity estimate (XHE) inbreeding F coefficient. F coeff thresholds at 0.2 and 0.8 are shown.

(B) Heterozygosity rate vs. $\log_{10}$ of the proportion of missing genotypes per ancestry group. Thresholds for the proportion of missing genotypes at $\log_{10}(0.05)$ and mean heterozygosity +/- 3*standard deviations per ancestry group are shown.

(C) Empirical cumulative distribution function of HWE $\log_{10}$ p-value for the European ancestry group. HWE threshold at $p=10^{-6}$ is shown.

(D) Histogram of $\log_{10}$ MAF. MAF threshold at $\log_{10}(0.005)$ is shown.

3. Load and concatenate individual whitelisted genotyping birdseed files using custom scripts, selecting SNPs with call confidence values $\leq$ 0.1. Annotate variants and generate PLINK files.
   a. To take advantage of parallel processing, concatenate and filter birdseed text files in batches.
   b. Read each birdseed text file as a tab delimited table with 906,600 SNPs as rows and three columns containing the following information: (See page 1, http://tools.thermofisher.com/content/sfs/brochures/genome_wide_snp6_sample_dataset_readme.pdf).

i. Composite Element REF: the probeset ID.
ii. Call: the genotype call with values of {-1, 0, 1, 2} corresponding to {NoCall, AA, AB, BB}.
iii. Confidence: the call confidence with values ranging from [0,1] with lower values corresponding to greater confidence.

c. Pre-filter to exclude SNPs with lower call confidence and set the "Call" value to NA for SNPs with "Confidence" > 0.1 prior to concatenation.

d. Iteratively concatenate each call column, generating a table with SNPs as rows, samples as columns, and call values as elements of the matrix.
i. Check that probeset IDs match prior to concatenating a genotyping call; if not, exclude and log the mismatched birdseed file.

e. Using a custom script, convert batch concatenated birdseed files into PLINK standard input transposed text format files.
i. Using the Affymetrix SNP Array 6.0 (release 35) annotation file, convert concatenated data into PLINK transposed text genotype tables (.tped) with allele calls (See .tped file format specification: https://www.cog-genomics.org/plink2/formats#tped).
ii. Create corresponding PLINK sample information files (.tfam) (See .tfam file format specification: https://www.cog-genomics.org/plink2/formats#tfam).

4. Import whitelisted germline data into PLINK for QC. Convert PLINK standard input transposed text files (`–tfile`) to standard input binary files (`–bfile`).
https://www.cog-genomics.org/plink2/input.

a. Import the tfile set (`–tfile`) into PLINK and create a bfile set (`–make-bed –out`) that generates corresponding PLINK binary biallelic genotype tables (.bed), PLINK extended MAP files (.bim) and PLINK sample information files (.fam). See file format specifications:
https://www.cog-genomics.org/plink2/formats#bed.
https://www.cog-genomics.org/plink2/formats#bim.
https://www.cog-genomics.org/plink2/formats#fam.

5. Impute the genotyping sex associated with each sample by calculating the X chromosome homozygosity estimate (XHE): https://www.cog-genomics.org/plink/1.9/basic_stats#check_sex.

Note: To minimize loss of TCGA samples when no self-reported sex is available and sex information is needed as a covariate in the analysis, sex can be imputed based on the XHE (F or inbreeding coefficient).

a. Split off the X chromosome's pseudo-autosomal region (`–split-x`) which is treated by PLINK as a separated XY chromosome. Indicate the proper build code.
b. Perform LD pruning (`–indep-pairphrase`).
c. Run check sex (`–check-sex`) which compares reported sex assignments with those imputed from X chromosome F coefficients.
d. Plot a histogram of the XHE F coefficients (F coeff). See Figure 1A.
i. A very tight distribution of F coeff around 1 is expected for males, and a more spread distribution of F coeff centered around zero is expected for females.
ii. In PLINK, F estimates < 0.2 are by default assigned female and F estimates > 0.8 assigned male. However, when (i) is observed and there is a clear gap between the two distributions, F coeff thresholds can be loosened and adjusted to correspond to the empirical gap. See "`–check-sex`" implementation and notes on TCGA sex assignment below.
e. Impute sex (`–impute-sex`) based on the XHE F coefficient.
f. Curate imputed sex assignments as needed and update sex assignments (`–update-sex`).

Note: Not all TCGA samples have self-reported sex information and we imputed sex based XHE. However, we found cases where self-reported and imputed sex were discordant; sex assignments were curated depending on whether F coefficients fall within the expected range (F coeff < 0.2 for females and > 0.8 for males) or F coefficients fall out of the expected range (F coeff > 0.2 and < 0.8) (see troubleshooting section, Problem 4). These

imputed/curated sex assignments for TCGA germline samples are available in Table S1 from (Sayaman et al., 2021).

6. Exclude SNPs and individuals with greater than 5% missingness.
   a. Filter variants (`–geno`) to include only SNPs with 95% genotyping rate (5% missing).
   b. Filter samples (`–mind`) to exclude individuals with more than 5% missing genotypes.
7. Calculate heterozygosity within each ancestry cluster, and filter samples with excess heterozygosity. https://www.cog-genomics.org/plink/1.9/basic_stats#ibc.
   a. Calculate heterozygosity (`–het`) vs. missingness (`–missing`) rates.
   b. Using downloaded UCSF ancestry assignments, calculate heterozygosity means and standard deviations within each of the European (EUR), African (AFR), East Asian (EAS) and Admixed American (AMR) ancestry clusters.
   c. Plot the $\log_{10}$ proportion of missing genotypes against heterozygosity rates with mean +/- 3*SD for each ancestry cluster for QC. See Figure 1B.
   d. Flag samples with heterozygosity >3*SD above the mean for each ancestry cluster; remove individuals as part of 8b sample filtering.

   Note: Samples with low heterozygosity are expected for certain ancestry groups and are not removed.

   Note: Not all TCGA samples have self-reported race and ethnicity data. Initial ancestry cluster assignments can be calculated based on principal component analysis (PCA) of germline data (`–pca`). In (Sayaman et al., 2021) initial ancestry calls were made based on Partition Around Medoids (PAM) clustering with k=4 using the first 3 principal components as described in (Sayaman et al., 2021), (Carrot-Zhang et al., 2020).

8. Select a representative sample for each individual with more than one sample. Conduct final filtering steps for all autosomal SNPs across the set of unique individuals.
   a. Restrict to autosomal chromosomes by excluding all unplaced and non-autosomal SNPs (`–autosome`).
   b. Create a final list of samples to include in the study (`–keep`).
      i. Exclude samples flagged in 7d for excess heterozygosity. ii. For individuals with more than one sample, preferentially select blood-derived normal samples; for those with more than one blood-derived sample, retain the samples with higher call rates.

   Note: All individuals and selected representative sample aliquots from TCGA germline data are listed in Table S1 from (Sayaman et al., 2021).

9. Calculate Hardy-Weinberg Equilibrium (HWE) within the largest ancestry cluster (EUR ancestry cluster). https://www.cog-genomics.org/plink/1.9/basic_stats#hardy.
   a. Subset for samples in EUR ancestry cluster. Calculate HWE (`–hardy`) across autosomal chromosomes.
   b. Plot the -$\log_{10}$ HWE p-value distribution for QC. See Figure 1C.
   c. Exclude SNPs (`–exclude`) that deviate from the expectation under HWE ($p < 1 \times 10^{-6}$) within the EUR ancestry cluster with the exception of SNPs previously associated with any cancer as reported in the GWAS catalog ($p < 5 \times 10^{-8}$) (Rashkin et al., 2020) since they may deviate from HWE in cancer patients.
10. Calculate Minor allele frequency (MAF) and exclude SNPs with MAF less than 0.5%. https://www.cog-genomics.org/plink/1.9/filter#maf.
    a. Calculate SNP MAFs (`–freq`).
    b. Plot the MAF cumulative distribution and histogram of -$\log_{10}$ MAF for QC. See Figure 1D.
    c. Filter out SNPs (`–maf`) with MAF < 0.005.
11. Remove duplicate SNPs with identical genomic first position.

a. Using a custom script, find SNPs with duplicate genomic first positions in the .bim file or alternatively identify SNPs sharing the same bp coordinate and allele codes in PLINK (`--list-duplicate-vars`).
b. Filter out duplicate SNPs (`--exclude`).

Note: The final QC'd list of sample (.fam) and SNP (.bim) files are available as part of the "Quality-controlled unimputed genotyping data plink files - QC_Unimputed_plink.zip" file under the "QC Unimputed Genotyping Data" sub-section of "TCGA QC HRC Imputed Genotyping Data used by the AIM AWG (from Sayaman et al.)" section of the "Supplemental Data Files": https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020.

Stranding

Timing: Approximately <1 day. Dependent on server capabilities

This section describes the stranding of the QC'ed genotyping data to the Haplotype Reference Consortium (HRC) prior to imputation.

Note: To skip this step, download the controlled access "HRC Stranded Genotyping Data" generated from (Sayaman et al., 2021), as described in step 11 of the "prepare germline genetic variation dataset" section of this protocol OR proceed with the Stranding protocol steps provided below.

Note: Scripts used in this section are available at: https://github.com/rwsayaman/TCGA_PanCancer_Genotyping_Imputation.

12. Prior to stranding, identify and remove all palindromic SNPs (A/T or G/C) (`--extract`).
13. Perform stranding to the Haplotype Reference Consortium using the McCarthy Group tools (https://www.well.ox.ac.uk/~wrayner/tools/; see section "HRC or 1000G Imputation preparation and checking").
    a. Download and unzip the tab delimited HRC reference file (version v1.1 HRC.r1-1.GRCh37.wgs.mac5.sites.tab or current version) from the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/site).
    b. Perform stranding of the quality-controlled genotyping file against the HRC reference panel using the high performance cluster version of the script (HRC-1000G-check-bim-v4.2.13-NoReadKey.zip), which compares genotyping alleles to the corresponding SNP alleles from HRC.
    c. Provide the .bim file, the calculated allele frequencies (`--freq`) and the reference panel as inputs (See "Usage with HRC reference panel").

Note: The McCarthy Group tools (https://www.well.ox.ac.uk/~wrayner/tools/) stranding script removes SNPs with differing alleles, SNPs with > 0.2 allele frequency difference, and SNPs not in the reference panel. The McCarthy Group stranding script would also remove A/T & G/C palindromic SNPs with MAF > 0.4, however we chose to remove all palindromic SNPs in the preceding step to remove ambiguity.

Genotype imputation

Timing: Approximately 1 week. Dependent on imputation server availability

This section describes generation of Haplotype Reference Consortium (HRC) imputed genotyping files from the stranded and QC'ed data (Figure 2A).

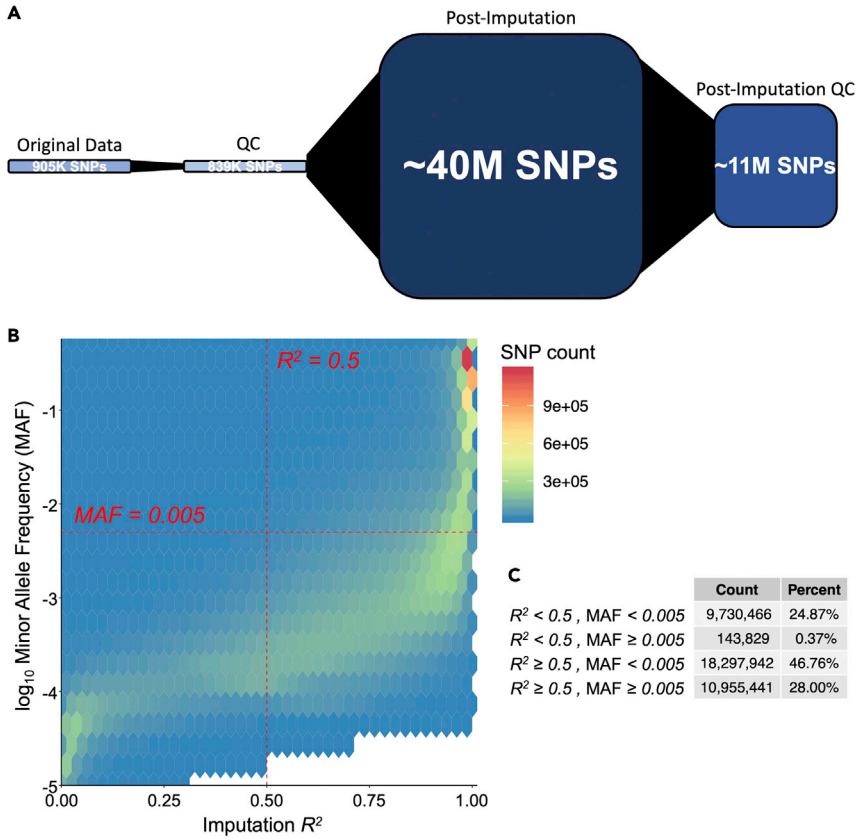


Figure 2. Expected distributions of imputation R^2 and MAF values

(A) Schematic of the number of SNPs (i) originally downloaded, (ii) after QC, (iii) after imputation, and (iv) after imputation QC.

(B) Hexagonal heatmap of 2d bin counts of the number SNPs post-imputation, showing the distribution of SNP HRC Imputation R^2 (x-axis) against the $\log_{10}$ Minor Allele Frequency (MAF) values across all autosomal chromosomes (y-axis). (c) Table showing the number and percent of SNPs below and above the suggested threshold levels of $R^2 \geq 0.5$ and MAF ≥ 0.005.

Note: To skip this step, download the controlled access "HRC Imputed Genotyping Data" generated from (Sayaman et al., 2021) as described in step 11 of the "prepare germline genetic variation dataset" section of this protocol OR proceed with phasing and imputation protocol steps provided below.

Note: Scripts used in this section are available at: https://github.com/rwsayaman/TCGA_PanCancer_Genotyping_Imputation https://github.com/rwsayaman/TCGA_PanCancer_Immune_Genetics.

14. Perform phasing and imputation using the Haplotype Reference Consortium (HRC) (Loh et al., 2016; McCarthy et al., 2016).
    a. To reduce the run time, divide the HRC stranded PLINK file into 22 files corresponding to individual autosomal chromosomes, recode to VCF files and compress as .vcf.gz files.
    b. Conduct phasing and imputation using a standard pipeline on the Michigan Imputation Server (MIS).
    c. Perform phasing using Eagle (version v2.3 or current version) on the variant call file (VCF) (Loh et al., 2016). By default, Eagle restricts analysis to bi-allelic variants that exist in both the target and reference data.

d. Run Minimac3 (Das et al., 2016) for imputation. For each of the 22 VCF files, the MIS breaks the dataset into non-overlapping chunks prior to imputation. For HRC imputation, select the HRC reference panel (version r1.1.2016 or current version) using mixed population for QC.

15. Download the HRC imputed germline files for each chromosome ("chr*.zip) from the MIS.
    a. Unzip each file using the provided password.
    b. Each unzipped folder contains 3 files:
        i. .dose.vcf.gz - imputed genotypes with dosage information.
        ii. .dose.vcf.gz.tbi - index file of the .vcf.gz file.
        iii. .info.gz file - information for each variant including quality and frequency (For Minimac3 info file, see: https://genome.sph.umich.edu/wiki/Minimac3_Info_File).
16. Filter to exclude SNPs with imputation $R^2 < 0.5$ using bcftools, see Figure 2B. The imputation R^2 is the estimated value of the squared correlation between imputed genotypes and true, unobserved genotypes.
    a. Filter "chr*.dose.vcf.gz" files for $R^2 \geq 0.5$ and index. Generate filtered "chr*.rsq0.5.dose.vcf.gz" and "chr*.rsq0.5.dose.vcf.gz.tbi "files.
    b. Generate new filtered "chr*.info.rsq0.5.gz" files.
17. Convert VCF files to PLINK files. Filter to exclude SNPs with MAF < 0.005, see Figure 2B.
    a. Convert VCF "chr*.rsq0.5.dose.vcf.gz" files to PLINK "tcga_imputed_hrc1.1_rsq0.5_chr*.bed" files (`–double-id –vcf`).
    b. Filter out SNPs (`–maf`) with MAF < 0.005 in PLINK.

Note: If you plan to analyze only a subset of the samples, recalculate the MAF in PLINK (`–freq`) for the population of interest. Filter SNPs based on the recalculated frequency.

Determination of ancestry-associated SNPs

Timing: 2 h

This section describes the association analysis between inferred genetic ancestry and SNP genotypes using the logistic regression implementation in the Hail framework.

18. Load imputed genotype data into the Hail framework.
    a. Import multi-sample .vcf files for each chromosome into Hail to create a 'matrix table' object.
    b. Load sample metadata (described above) and annotate matrix table object using the hail 'annotate_cols' function.
19. Perform sample quality control analysis using the hail 'sample_qc' function by filtering out samples that do not meet the following criteria:
    a. Sample call rate, the proportion of non-missing or filtered genotype calls $\geq$ 95%.
    b. Non-admixed samples according to the consensus ancestry call annotation.
20. Perform variant quality control analysis by filtering out SNPs that:
    a. Deviate from Hardy-Weinberg Equilibrium (HWE test $p < 1 \times 10^{-6}$).
    b. Global allele frequency < 1%.
21. For each comparison (EUR-AFR and EUR-EAS), test the association between genetic ancestry and SNP genotypes using logistic regression (Hail function 'logistic_regression_rows').
    a. binary response variable: ancestry (encoding: 0 EUR and 1 AFR/EAS).
    b. explanatory variable: number of alternate alleles per sample (Ref: 0, Homozygous:1, Homozygous Alternative: 2).
    c. covariates: biological sex, age.

Detecting ancestry-associated quantitative expression trait loci

Timing: 2 h

Lastly, we integrate the SNP genotype associations with ancestry and cancer-specific eQTLs to determine the extent to which germline genetic variation explains differential expression between ancestries.

22. Extract significant cancer-specific eQTLs from the PancanQTL database.
    a. Keep eGene-eSNP pairs for which a given eGene is in the set of genes with significant ancestry-association expression.
23. Perform table joins between the filtered eQTL results table and SNP genotype associations by SNP identity (dbSNP identifier).

Note: The underlying datasets may be derived from different versions of dbSNP. Alternatively, you can consider joining tables by the chromosome name, genomic position, reference allele and alternate allele, assuming that both datasets are derived from the same version of the reference genome.

24. For each gene with demonstrated ancestry-differential expression, determine whether it has at least one ancestry-associated eSNP.
25. Calculate summary statistics and visualize representative loci by cancer type as shown in Figure 6 of (Carrot-Zhang et al., 2020).

EXPECTED OUTCOMES

Good quality TCGA germline imputation calls

If this protocol is carried out as described here, you can expect to identify a total of 838,948 autosomal chromosome variants for 10,128 unique individuals that pass the QC filters. After removal of palindromic SNPs and stranding to the HRC panel, 680,389 correctly matched variants remain. These are submitted to the MIS which returns 39,127,678 SNPs for 10,128 unique individuals (Figure 2A). Subsequent quality control analysis and filtering based on imputation quality ($R^2 \geq 0.5$) and minor allele frequency (MAF ≥ 0.005) thresholds yields 10,955,441 SNPs (Figure 2A and 2B).

Ancestry-associated SNPs

Ancestry-associated SNPs should show strong differences in allele frequency distribution across populations. These findings are difficult to validate but comparison of allele frequency distributions in population genetic reference catalogs can provide a sanity check to ensure that results are plausible. A non-parametric two-sample Kolmogorov-Smirnov (KS) test can be performed to compare the equality of two distributions (the null hypothesis being that the two samples are drawn from the same distribution).

For instance, we compared the minor allele frequencies between European and African samples in the 1000 genomes catalog and show that ancestry associated SNPs have a significantly different MAF distribution according to the reference population (Figure 3), with KS test $p < 0.05$.

Ancestry-associated eQTLs

See (Carrot-Zhang et al., 2020) Figure 6 for expected outcomes of ancestry-associated eQTLs.

LIMITATIONS

Data availability

The analysis outlined in this protocol requires paired germline genotypes and mRNA expression data in an ancestrally-diverse cohort to determine e-QTL loci that are driven by germline genetic differences between ancestry groups. A key limitation of the type of analysis described here is the availability of a substantial amount of data in a tumor type of interest with sufficient numbers of samples in each ancestry group to have enough statistical power to make statistical inference possible. As outlined in (Carrot-Zhang et al., 2020), our molecular analyses were limited to tumor types with at least 10 samples in the minority population for a given comparison. While our results provide a general view of the trends

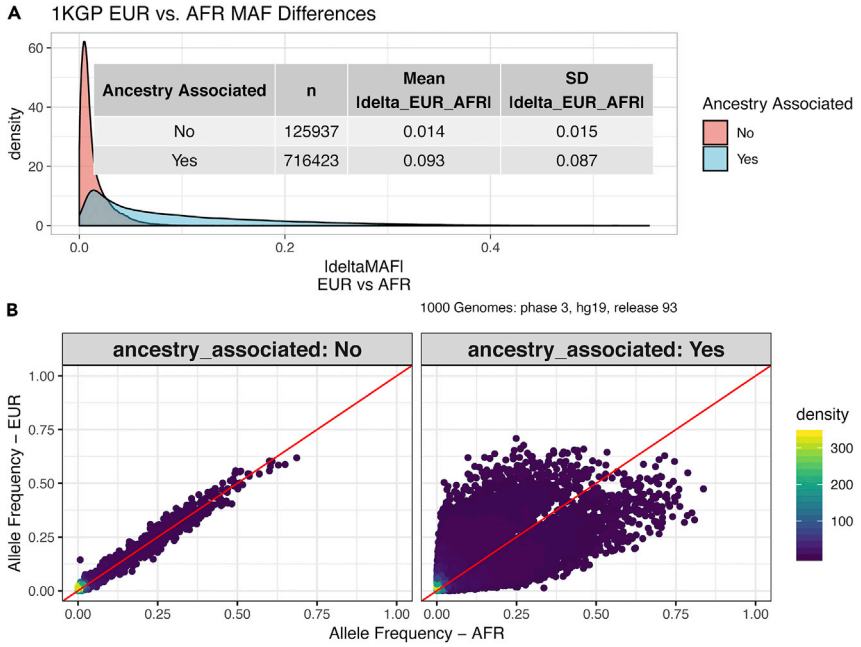


Figure 3. 1000 Genomes allele frequency distributions for ancestry associated SNPs in European and African populations

(A) Delta Minor Allele Frequency (dMAF) distributions for imputed SNPs by ancestry association status as class determined by logistic regression in the 1000 Genomes reference populations. Kolmogorov-Smirnov (KS) test $p < 0.05$.

(B) Scatterplot depicting density of allele frequencies for all imputed SNPs that passed QC and were tested for ancestry association in the African (x-axis) and European (y-axis) populations.

and patterns of germline effects on ancestry-associated molecular traits in cancer, specific claims were limited to the cancer types with sufficient sample sizes that allowed performance of valid statistical tests.

Limitations of imputation

Ideally, ancestry associated germline variants would be determined from deep coverage whole genome sequencing datasets. However, this approach can be prohibitively expensive and infeasible. We have used the Affymetrix Genome Wide SNP 6.0 genotyping data and imputed genotypes using the Haplotype Reference Consortium (HRC) imputation panel (version r1.1.2016) that represents human haplotypes determined from 38,821 individuals across 20 diverse genomic cohorts. While this panel represents one of the largest datasets of reference human genomes to capture of human genetic variation, the composition of the reference individuals is predominantly of European ancestry. However, HRC is limited for imputing rare variants, particularly in non-European ancestry populations. There may be rare haplotypes with high impact functional genetic variation in the TCGA non-European samples that are not captured in our analysis. This points to the importance of more ethnically diverse study cohorts in cancer genomics. In addition, HRC does not include short insertion/deletion variants. Imputation to other reference datasets such as the 1000 Genomes may be helpful for capturing insertion/deletion variants. In addition, imputation to the newest and largest reference dataset, the TOPMed dataset, should improve the ability to find additional rare variants.

TROUBLESHOOTING

Problem 1

Issues accessing controlled access data from the GDC Portal or the GDC publication page associated with (Carrot-Zhang et al., 2020) (step 6 of before you begin).

Potential solution

All TCGA germline data are controlled access. Ensure that you have followed all steps required by the GDC to obtain controlled access data including authentication through eRA Commons and dbGaP authorization. Step by step instructions for obtaining access are outlined in the 'before you begin' section above and further details can be found here: https://gdc.cancer.gov/access-data/obtaining-access-controlled-data.

Problem 2

Issues or errors running commands on the high- performance compute server or implementing available code from GitHub. (quality control analysis of germline data (steps 1–9); stranding (steps 12 and 13); genotype imputation (steps 14–17)).

Potential solution

Ensure the proper software, libraries and dependencies are installed. Software implementation may be version specific, the versions used in the protocol are provided to ensure reproducibility. The provided GitHub code for pre-processing genotyping data was optimized for the specifications of the TIPCC high-performance compute (HPC) environment at University of California, San Francisco (UCSF) (which had 8 communal compute nodes and 1 dedicated node, each with 12–64 cores, 64–512 GB of RAM and at least 1.8 TB of fast local disk space) employing Portable Batch System (PBS) job scheduling. Consult your system administrator to adapt the provided code to your system.

Problem 3

Computation run times are much slower than expected. Inadequate computational power to run workflows. (quality control analysis of germline data (steps 1–9); stranding (steps 12 and 13); genotype imputation (steps 14–17); determination of ancestry-associated SNPs (steps 18–21)).

Potential solution

Some sections of this protocol require intensive computation that requires high-performance computing as outlined in the 'Before Your Begin' section. Note that run times are dependent on the specifications of the compute environment. Before running analysis of the whole genome, consider first benchmarking performance on a single chromosome – e.g., the smallest chromosome, chr21 or largest chromosome, chr1, to estimate run times and assess if you have compute resources with adequate specifications. If you encounter difficulties running the protocol such as inadequate disk storage or slow processing speeds such that the protocol takes a prohibitively longer time than is useful, you can consider alternatives such as using an adequately provisioned system in a cloud-computing environment. Work with the relevant vendors to create virtual instances that will meet both the need for data security and processing power to run the workflows described in this protocol.

Problem 4

Sex imputation based on XHE yields discordant self-reported and imputed sex in TCGA (step 5).

Potential solution

Not all TCGA samples have self-reported sex information and we imputed sex based XHE. However, we found cases where self-reported and imputed sex were discordant. Moreover, we found no clear empirical gap in the distribution between those self-reporting as female and male in the F coeff range > 0.2 and < 0.8 (Figure 1A). In (Sayaman et al., 2021), we used both imputed and self-reported sex to curate sex assignments in TCGA.

For all individuals falling within the expected XHE distributions, individuals with F coefficients < 0.2 were assigned female and those with F coefficients > 0.8 were assigned male, these include: (i) individuals with concordant imputed and self-reported sex (Figures 4A and 4F),

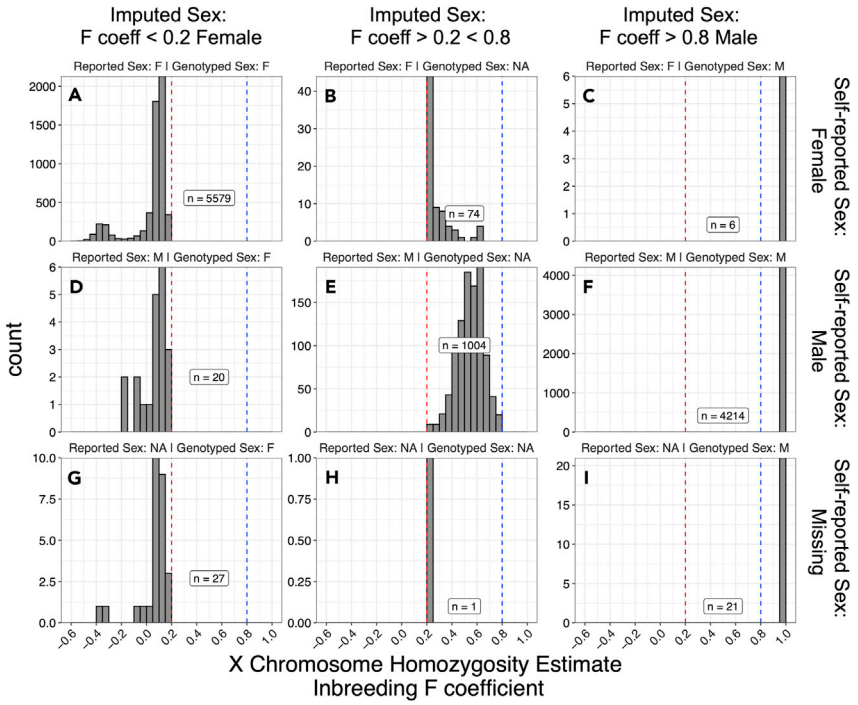


Figure 4. Curation of TCGA sex assignments

(A–I) Histograms of XHE inbreeding F coefficient faceted by imputed genotyped sex and self-reported sex. Number of individuals within each category are annotated. (Note, y-axes are scaled within each category for readability.).

(ii) individuals with discordant imputed and self-reported sex, where imputed sex is curatedly assigned (Figures 4C and 4D), and (iii) individuals with no self-reported sex (4G, 4I).

For individuals with an uncharacteristic distribution of F coeff > 0.2 < 0.8 (Figures 4B and 4E) with an unexpectedly significant proportion of individuals self-reporting as male with F coeff < 0.8 (Figure 4E), we elected to keep the self-reported sex assignment. We reasoned that since the distribution of F coeff for those self-reporting as female is still centered around 0 just with larger spread (Figure 4B), the uncharacteristic distribution of F coeff for those self-reporting as male (Figure 4E) may be due to array quality.

For the single individual with F coeff > 0.2 < 0.8 and no self-reported sex, the individual was assigned female based on distribution of F coeff (Figure 1H). These imputed/curated sex assignments for TCGA germline samples are available in Table S1 from (Sayaman et al., 2021).

Problem 5

Modest or few associations with ancestry found at any level (step 21).

Potential solution

Statistical modeling may not have accounted for all possible confounders. Confounding variables include both (1) technical artifacts – such as batch effects arising from differences in sample collection, handling or preparation between individuals, laboratories or institutions or across reagent lots, as well as processing artifacts arising from differences between sequencing runs or microarray plates; and (2) other biological or clinical factors – such as cancer type (or cancer-specific subtypes in per cancer analysis), age, sex etc., that influence the dependent and independent variables under consideration leading to spurious associations. Exploratory data analysis such as Principal Component Analysis (PCA) or clustering methods can reveal correlations between principal components or cluster membership and confounding variables, and can guide selection of variables to include.

Confounding variables can be modeled as covariates; however missing values or missing annotations present a challenge and pose limitations on the selection of confounding variables to be included in the model.

Problem 6

Logistic regression statistical models for the association between ancestry groups and germline genotypes do not run properly or fail converge (step 21).

Potential solution

Logistic regression models may fail to converge when maximum likelihood estimates cannot be estimated accurately. In most cases, maximum likelihood estimates do not exist due to complete or quasi-complete separation – e.g., when the outcome variable separates a predictor variable completely or almost completely (Allison, 2004). For our purposes, where we look at genome-wide patterns, models that did not converge were excluded. Users should consult statistical treatments on how to handle complete or quasi-complete separation.

Similarly, logistic regression models cannot reasonably handle missing values. Incomplete cases can be excluded, or missing values imputed; however, caution should be exercised with imputation – e.g., missingness is not random (Kang, 2013). See Hail methods for logistic regression: https://hail.is/docs/0.2/methods/stats.html#hail.methods.logistic_regression_rows.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Andrew D. Cherniack (achernia@broadinstitute.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The primary data used in this protocol are publicly available on the NCI Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov) or the relevant publication pages from (Carrot-Zhang et al., 2020) at (https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). Controlled access to the TCGA original birdseed and pre-processed quality-controlled genotyping data imputed to the Haplotype Reference Consortium (HRC) (Sayaman et al., 2021) requires dbGAP permission approval. The quality-controlled and HRC imputed genotyping data are accessible at the GDC publication page (https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020; See: "Sayaman et al. TCGA QC HRC Imputed Genotyping Data" section under "Supplemental Data Files"). All other data sources are indicated in the key resources table.

CONSORTIA

The members of the Cancer Genome Analysis Network are: Jian Carrot-Zhang, Ashton C. Berger, Seunghun Han, Matthew Meyerson, Jeffrey S. Damrauer, Katherine A. Hoadley, Ina Felau, John A. Demchok, Michael K.A. Mensah, Roy Tarnuzzer, Zhining Wang, Liming Yang, Theo A. Knijnenburg, A. Gordon Robertson, Christina Yau, Christopher Benz, Kuan-lin Huang, Justin Y. Newberg, Garrett M. Frampton, R. Jay Mashl, Li Ding, Alessandro Romanel, Francesca Demichelis, Wanding Zhou, Peter W. Laird, Hui Shen, Christopher K. Wong, Joshua M. Stuart, Alexander J. Lazar, Xiuning Le, Ninad Oak.

ACKNOWLEDGMENTS

We are grateful for advice from numerous colleagues, TCGA and Genomic Data Analysis Network collaborators, and the GDC technical support team. We thank the Cancer Genome Atlas Research

Network, the National Cancer Institute, United States for funding through U24 grants CA210999, CA210974, CA211006, CA210949, CA210978, CA210952, CA210989, CA210957, CA210990, CA211000, CA210950, CA210969, CA210988, and K24CA169004 and R01CA1845851. J.C.-Z. holds a Banting fellowship. R.W.S. was supported by the NCI Cancer Metabolism Training Program Postdoctoral Fellowship (T32CA221709). E.Z. is supported by the National Institutes of Health R01CA227466 and K24CA169004.

AUTHOR CONTRIBUTIONS

N.C., R.W.S., D.H., and S.H. wrote code and performed analysis. A.K., S.C., J.C.Z., and The Cancer Genome Atlas Analysis Network provided project administration. A.D.C. and R.B. provided supervision and N.C., R.W.S., E.Z., A.D.C., and R.B. wrote, and all authors reviewed the manuscript.

DECLARATION OF INTERESTS

A.D.C. receives research funding from Bayer. R.B. owns equity in and consults for Scorpion Therapeutics and receives research funding from Novartis.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Allison, P. (2004). Convergence problems in logistic regression. In Numerical Issues in Statistical Computing for the Social Scientist (John Wiley & Sons, Inc.), pp. 238–252.

Anderson, C.A. (2011). Chapter 7 - data quality control. In Analysis of Complex Disease Association Studies, E. Zeggini and A. Morris, eds. (Academic Press), pp. 95–108.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat. Protoc. *5*, 1564–1573.

Aron, S., and Choudhury, A. (2015). GWAS QC - theory and steps. Lecture Notes, Medical population genetics and GWAS for complex diseases 19th April – 22nd April, 2015. In H3ABioNet Pan African Bioinformatics Network for H3Africa (Johannesburg: University of the Witwatersrand).

Carrot-Zhang, J., Chambwe, N., Damrauer, J.S., Knijnenburg, T.A., Robertson, A.G., Yau, C., Zhou, W., Berger, A.C., Huang, K.-L., Newberg, J.Y., et al. (2020). Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. Cancer Cell *37*, 639–654.e6.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve Years of SAMtools and BCFtools. GigaScience *10*. https://doi.org/10.1093/gigascience/giab008.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782–784.

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A.-Y., et al. (2018). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res. *46*, D971–D976.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell *173*, 283–285. https://doi.org/10.1016/j.cell.2018.03.042.

Kang, H. (2013). The prevention and handling of the missing data. Korean J Anesthesiol *64*, 402–406.

Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the haplotype reference Consortium panel. Nat. Genet. *48*, 1443–1448.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of *64*, 976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Rashkin, S.R., Graff, R.E., Kachuri, L., Thai, K.K., Alexeeff, S.E., Blatchins, M.A., Cavazos, T.B., Corley, D.A., Emami, N.C., Hoffman, J.D., et al. (2020). Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat. Commun. *11*, 4423.

Robertson, A.G., Yau, C., Carrot-Zhang, J., Damrauer, J.S., Knijnenburg, T.A., Chambwe, N., Hoadley, K.A., Kemal, A., Zenklusen, J.C., Cherniack, A.D., et al. (2021). Integrative modeling identifies genetic ancestry-associated molecular correlates in human cancer. STAR Protoc. *2*, 100483.

Sayaman, R.W., Saad, M., Thorsson, V., Hu, D., Hendrickx, W., Roelands, J., Porta-Pardo, E., Mokrab, Y., Farshidfar, F., Kirchhoff, T., et al. (2021). Germline genetic contribution to the immune landscape of cancer. Immunity *54*, 367–386.e8.