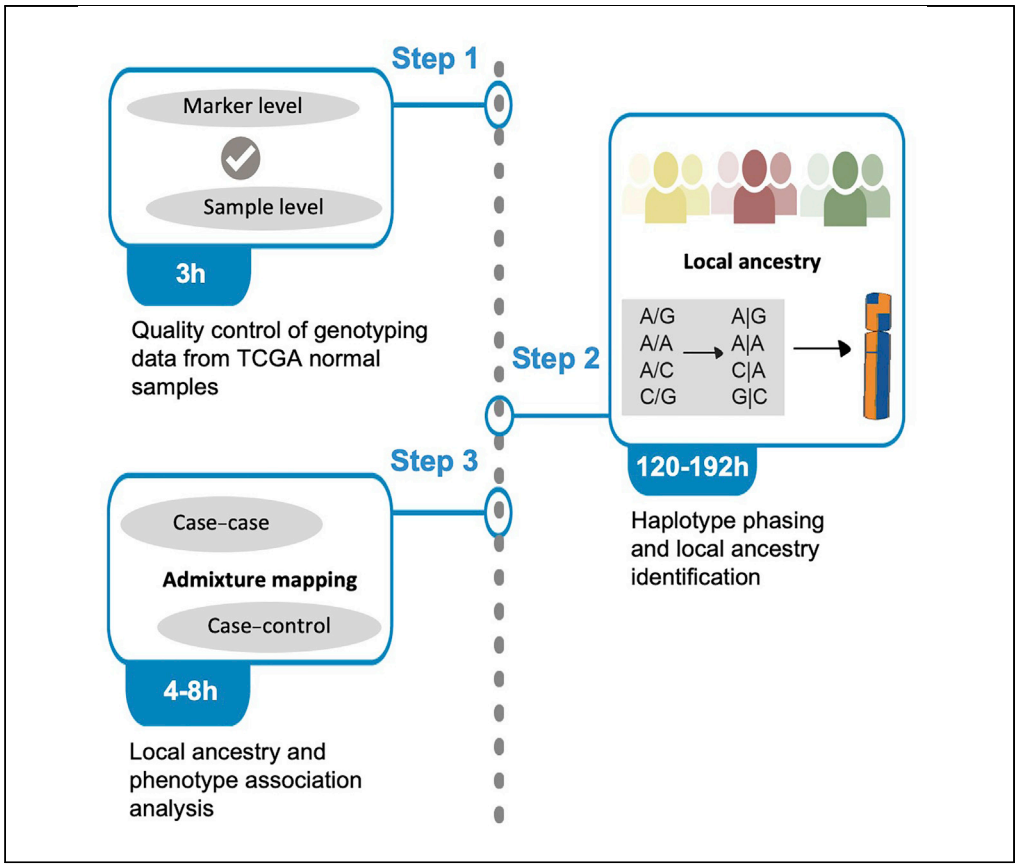
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Analytical protocol to identify local ancestry-associated molecular features in cancer



People of different ancestries vary in cancer risk and outcome, and their molecular differences may indicate sources of these variations. Determining the "local" ancestry composition at each genetic locus across ancestry-admixed populations can suggest causal associations. We present a protocol to identify local ancestry and detect the associated molecular changes, using data from the Cancer Genome Atlas. This workflow can be applied to cancer cohorts with matched tumor and normal data from admixed patients to examine germline contributions to cancer.

Jian Carrot-Zhang, Seunghun Han, Wanding Zhou, ..., Ninad Oak, Andrew D. Cherniack, Rameen Beroukhim

zhangj@broadinstitute.org (J.C.-Z.)
achernia@broadinstitute.org (A.D.C.)
rameen_beroukhim@dfci.harvard.edu (R.B.)

Highlights

Protocols for local ancestry identification using the TCGA data

Detecting local ancestry associated with cancer risk

Statistical analysis associating molecular changes with local ancestry

Understanding the contribution of genetic ancestry in admixed patients

Carrot-Zhang et al., STAR Protocols 2, 100766
December 17, 2021
https://doi.org/10.1016/j.xpro.2021.100766

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Analytical protocol to identify local ancestry-associated molecular features in cancer

Jian Carrot-Zhang,[1,2,3,10,*] Seunghun Han,[2,3] Wanding Zhou,[4,5] Jeffrey S. Damrauer,[6] Anab Kemal,[7] Cancer Genome Atlas Analysis Network, Andrew D. Cherniack,[1,2,3,11,*] and Rameen Beroukhim[1,2,3,8,9,*]

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

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

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

[4]Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, PA, 19104, USA

[5]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

[6]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

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

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

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

[10]Technical contact

[11]Lead contact

*Correspondence: zhangj@broadinstitute.org (J.C.-Z.), achernia@broadinstitute.org (A.D.C.), rameen_beroukhim@dfci.harvard.edu (R.B.)

https://doi.org/10.1016/j.xpro.2021.100766

SUMMARY

People of different ancestries vary in cancer risk and outcome, and their molecular differences may indicate sources of these variations. Determining the "local" ancestry composition at each genetic locus across ancestry-admixed populations can suggest causal associations. We present a protocol to identify local ancestry and detect the associated molecular changes, using data from the Cancer Genome Atlas. This workflow can be applied to cancer cohorts with matched tumor and normal data from admixed patients to examine germline contributions to cancer. For complete details on the use and execution of this protocol, please refer to Carrot-Zhang et al. (2020).

BEFORE YOU BEGIN

Prepare input files

Timing: 2–4 h

This protocol uses genotyping data, ancestry assignments, and molecular features associated with ancestry generated by The Cancer Genome Atlas (TCGA) pan-cancer atlas.

1. Download TCGA Affymetrix SNP 6.0 microarray data from Genomic Data Commons Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/) for each matched normal sample. Select the Birdseed outputs (TXT under Data format), which are the genotyping files processed from raw SNP array CEL files.

Note: Please refer to the following link to request access to controlled germline genotyping data: https://gdc.cancer.gov/access-data/obtaining-access-controlled-data

This protocol runs local ancestry identification by cancer type. It is recommended to select samples of a specific cancer type and copy their Birdseed genotyping files to a separate folder.

2. Download ancestry assignments from Table S1 of Carrot-Zhang et al., 2020 (10.1016/j.ccell.2020.04.012), which includes percentage of EUR, EAS or AFR ancestry for each individual.
3. Download a table mapping TCGA SNP array ID to Sample and Patient ID from Genomic Data Commons Legacy Archive or download an example file:(https://github.com/jcarrotzhang/ancestry-from-panel/blob/master/GDAN_AIM/geno_sample_map.txt)
4. Download somatic mutations associated with ancestry from Table S2 of Carrot-Zhang et al. 2020 (10.1016/j.ccell.2020.04.012).
5. Download methylation differences associated with ancestry from Table S3 from Carrot-Zhang et al. 2020 (10.1016/j.ccell.2020.04.012).
6. Download mRNA expression associated with ancestry from Table S4 from Carrot-Zhang et al. 2020 (10.1016/j.ccell.2020.04.012).

This protocol uses 1000 Genomes samples as a reference panel to identify haplotypes and assign ancestry to genomic loci by chromosome (local ancestry identification).

7. Download 1000 Genomes reference panel version 5a (http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a/b37.vcf/) and convert it into PLINK binary PED format using the PLINK –make-bed option, and join all chromosomes into one file, using the PLINK –merge-list option. Next, remove duplicate markers and markers exceeding missingness in the 1000 Genomes samples.
   a. Identify duplicate germline variants (markers) in the *.bim file and write them to a separate file. Then, exclude those duplicate markers using the PLINK –exclude option.
   b. Identify markers excessed missingness (a minimum frequency of 5% genotype missing rate allowed) using the PLINK –freq option, and then, exclude markers excessed missingness in the *frq.counts file using the PLINK –exclude option. After filtering, 32,161,998 markers are left.
8. To convert Birdseed genotyping files into VCF format, download hg19 human genome reference: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz Unzip and index the fasta file using Samtools faidx. Then, download the genome annotation file for Affymetrix SNP array 6.0: https://software.broadinstitute.org/cancer/cga/sites/default/files/data/tools/contest/GenomeWideSNP_6.na30.annot.hg19.csv.pickle.gz
9. Download the population information for 1000 Genomes samples from: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel
10. Download the 1000 Genomes phased haplotypes (*.hap.gz and *.legend.gz): https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
11. Download the genetic map file (genetic_map*) containing at least three columns (chromosome, physical position in bp and genetic position in cM):https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html

Software installation

Timing: 3–5h

Because some algorithms take time and memory, this protocol recommends installing the following software in a high-performance compute cluster.

12. Install the most recent version of Samtools, Bcftools and Bedtools as described in key resources table.
13. Install PLINK v1.9, Shapeit v2 and RFMix v1.5.4 as described in key resources table.
14. Install Python 2.7 library to support downstream analyses.

15. To convert Birdseed genotyping files into VCF format, install pyfasta (https://pypi.org/project/pyfasta) and add the module to your Python classpath.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA Affymetrix SNP 6.0 microarray data	Genomic Data Commons Legacy Archive	https://portal.gdc.cancer.gov/legacy-archive/
TCGA cancer type and subtype	(Sanchez-Vega et al., 2018)	https://ars.els-cdn.com/content/image/1-s2.0-S0092867418303593-mmc1.xlsx
TCGA patient ancestry call	(Carrot-Zhang et al., 2020)	Table S1 of 10.1016/j.ccell.2020.04.012
Somatic mutation, methylation and mRNA expression associated with ancestry in TCGA	(Carrot-Zhang et al., 2020)	Tables S2–S4 of 10.1016/j.ccell.2020.04.012
TCGA SNP array and sample ID map	(Carrot-Zhang et al., 2020)	https://github.com/jcarrotzhang/ancestry-from-panel/blob/master/GDAN_AIM/geno_sample_map.txt
TCGA patient gender and age	(Liu et al., 2018)	https://api.gdc.cancer.gov/data/1b5f413e-a8d1-4d10-92eb-7c4ae739ed81
Resource website for the local ancestry calls	(Carrot-Zhang et al., 2020)	https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020
Software and Algorithms		
PLINK v1.9	(Purcell et al., 2007)	https://www.cog-genomics.org/plink/1.9/
Samtools	(Li et al., 2009)	http://www.htslib.org/
Bcftools	n/a	http://samtools.github.io/bcftools/bcftools.html
Bedtools	(Quinlan and Hall, 2010)	https://bedtools.readthedocs.io/en/latest/
SHAPEIT v2	(Delaneau et al., 2011)	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
RFMIX v1.5.4	(Maples et al., 2013)	https://sites.google.com/site/rfmixlocalancestryinference/
birdseed2vcf	n/a	https://github.com/ding-lab/birdseed2vcf
Local ancestry identification	(Martin et al., 2017)	https://github.com/armartin/ancestry_pipeline
Local ancestry association analysis	(Carrot-Zhang et al., 2020)	https://github.com/jcarrotzhang/ancestry-from-panel/tree/master/GDAN_AIM

STEP-BY-STEP METHOD DETAILS

Local ancestry identification

Timing: 3h for step 1 and step 2; 72–120h for step 3; 48–72h for step 4; 100–180h for step 5 (alternative step)

This step phases the TCGA samples into haplotypes, produces local ancestry calls that assign African (AFR), European (EUR) and East Asian (EAS) ancestry to each genomic locus (Figure 1), and estimates global ancestry proportions to ensure accuracy of local ancestry calls.

1. First, for each cancer type, merge all samples into PLINK input format (binary PED files):
   a. Exclude variants with genotype call confidence greater than 0.1 in the Birdseed files downloaded from the Genomic Data Commons Legacy Archive.
   b. Convert the Birdseed genotyping files into VCF format, using birdseed2vcf (https://github.com/ding-lab/birdseed2vcf) following the instructions.
   c. Select samples of a specific cancer type and copy the VCF files to a folder named for that cancer type.
   d. Compress and index the VCF files using bgzip and tabix from Samtools. Prepare a list of VCF files for each cancer type, then, use the Bcftools –merge option to merge all samples in each cancer type into one VCF file.

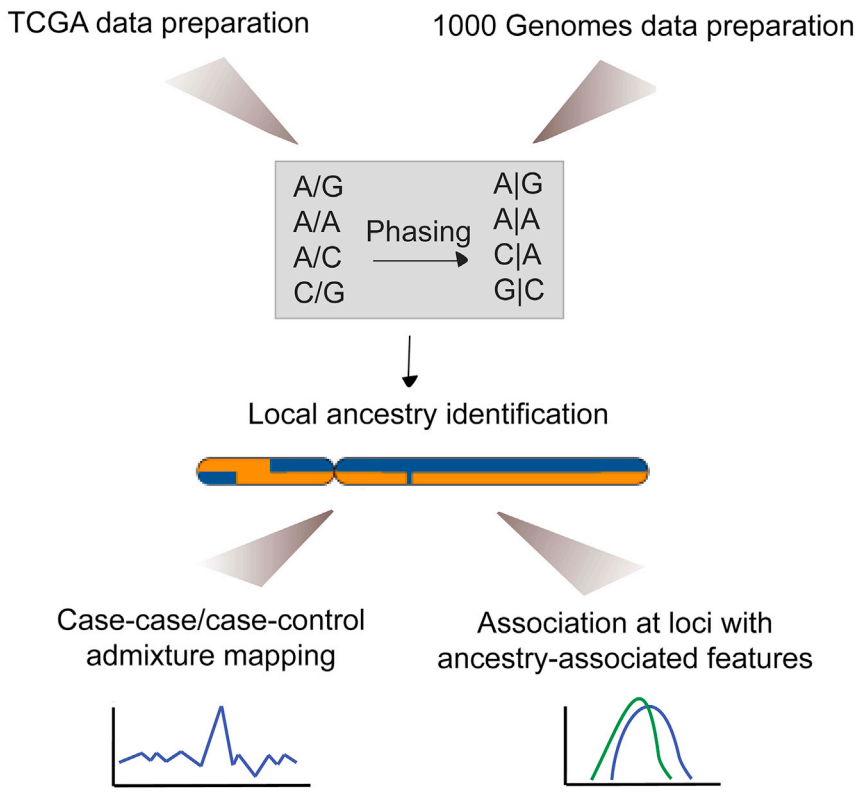


Figure 1. Analytical workflow for identifying local ancestry-associated molecular features in cancer
Part 1: Preparation of TCGA and 1000 Genomes genotyping data; Part 2: Haplotype estimation and local ancestry assignments; Part 3: Examples of correlation analyses utilizing local ancestry information.

An example Unix bash command:

```
bcftools merge -l $vcflist.txt -Oz > $OUT.vcf.gz
```

e. For each merged file, use PLINK to convert VCF into format to binary PED files (BED/BIM/FAM) using the PLINK –make-bed option.

⚠ CRITICAL: It is recommended to merge TCGA samples by cancer type and run local ancestry identification for each cancer type in parallel to reduce the computational runtime.

2. To prepare for phasing, perform quality control for genome-wide SNP genotyping data and all TCGA germline samples, and then merge TCGA samples from the same cancer type with 1000 Genomes reference samples into a single PLINK file.
   a. Perform basic quality control of markers using PLINK:
      i. Exclude markers that are not on autosomes using the PLINK –not-chr option.
      ii. Identify markers excessed missingness (a minimum frequency of 5% genotype missing rate allowed) using the –freq option, and exclude markers excessed missingness in the *frq.counts file using the PLINK –exclude option.
      iii. Calculate Hardy Weinberg equilibrium using the PLINK –hardy option, and identify markers failing Hardy Weinberg equilibrium test (P<1e-6) due to excess heterogeneousness in the *.hwe file using the PLINK –exclude option.
   b. Perform basic quality control of samples:
      i. Compute pairwise relationships using the PLINK –genome option.
      ii. Exclude samples due to duplicates, with the PI_HAT column in the *genome file greater than or equal to 0.7.

iii. Exclude samples due to relatedness (e.g., sibling), with the PI_HAT column in the *genome file between 0.35 and 0.7, and the Z1 column (IBD=1) less than 0.7.

c. Merge the TCGA samples with 1000 Genomes project phase 3 samples:

i. Identify markers with minor allele frequency (MAF) of at least 1% in 1000 Genomes project phase 3 samples, using the PLINK –maf option.

ii. Extract markers that passed quality control from both the TCGA data set identified in step 2a using the PLINK –extract option.

iii. Merge the two datasets using the PLINK –bmerge option.

⚠ CRITICAL: It is recommended to merge TCGA samples with reference samples into one PLINK binary PED file and analyze them together to avoid technical errors, such as strand flip in the phasing step.

3. To prepare for local ancestry identification, for each TCGA germline sample, phase alleles to haplotypes using Shapeit2 by chromosome.

a. Split the genome into 22 files corresponding to each autosomal chromosome using the –chr option in PLINK

b. Perform phasing using Shapeit2 for each chromosome by running the following bash command line as an example:

i. Marker inspection to exclude SNPs that are incompatible (e.g., exceeding missing data rate; strand error etc.) for phasing using Shapeit2. SNPs to be excluded are shown in the output file (*.mendel.snp.strand.exclude).

```
for i in {1..22}; do
shapeit -check –input-ref 1000GP_Phase3_chr${chr}.hap.gz \
1000GP_Phase3_chr${chr}.legend.gz \
1000GP_Phase3.sample \
-B BRCA/chr${chr} \
--input-map genetic_map_chr${chr}_combined_b37.txt –input-thr 1 \
--output-log BRCA/chr${chr}.mendel;
done
```

ii. Phasing algorithm using Shapeit2. The output files include *.haps and * sample files.

```
for i in {1..22}; do

shapeit –input-ref 1000GP_Phase3_chr${chr}.hap.gz \

1000GP_Phase3_chr${chr}.legend.gz 1000GP_Phase3.sample \

-B BRCA/chr${chr} \

--duohmm –input-map genetic_map_chr${chr}_combined_b37.txt \

--output-max BRCA/chr${chr}.haps.gz BRCA}/chr${chr}.sample \

--exclude-snp BRCA/chr${chr}.mendel.snp.strand.exclude \

--thread 4;

done
```

⚠ CRITICAL: Phased 1000 Genomes samples (1000GP_Phase3_chr${chr}.hap.gz) are used as a reference panel for phasing the TCGA samples (BRCA/chr${chr}). This step takes time, so

it is recommended to parallelize the tasks by chromosome. In the following, four cores are requested for each phasing job.

Imputation is not required in this step, as the number of germline alleles used to capture local ancestry is sufficient, as a linkage disequilibrium generated from continental-scale admixtures that occurred recently (e.g., African American and Latin American populations) is expected to span many centimorgans.

Note: Other tools, such as Eagle2 (Loh et al. 2016), Beagle (Browning and Browning 2007) etc. can also be used for phasing larger datasets. Moreover, a larger reference panel, such as Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/home) consisting of 64,976 haplotypes, may improve phasing accuracy. One may use the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html#!pages/home) to run Eagle phasing with Haplotype Reference Consortium as the reference panel. One may write a custom script to convert Eagle output files into RFMix input files for step 4.

4. Perform local ancestry identification to assign AFR, EUR or EAS ancestry to each genomic region by chromosome using RFMix.
   a. Choose parental populations from 1000 Genomes samples to use for local ancestry identification. An example file with 1668 samples from AFR, EUR and EAS populations can be found: https://github.com/jcarrotzhang/ancestry-from-panel/blob/master/GDAN_AIM/BRCA.ref

Note: It is recommended by RFMix to balance the size of continental-scale ancestries in the reference samples to avoid biases caused by skewed reference sample sizes. For example, one can randomly choose between 400 to 500 1000 Genomes samples from AFR, EUR, or EAS population.

   b. Read in the ancestry assignment from Table S1 of Carrot-Zhang et al., 2020, exclude samples with South Asian (SAS) or American (AMR) ancestry greater than 20%, and then generate a list of TCGA samples by cancer type (for example, BRCA.notref).

Note: Because most TCGA samples were from AFR, EUR or EAS populations, the identification of local SAS or AMR ancestry is not performed in this protocol.

   c. For each chromosome, convert phased shapeit outputs into RFMix input files, and prepare other RFMix input files (*.alleles; *.classes; *.snp_locations). Detailed information on how to prepare the RFMix input files can be found in the manual: https://www.dropbox.com/s/cmq4saduh9gozi9/RFMix_v1.5.4.zip?file_subpath=%2FRFMix_v1.5.4%2FManual.pdf
   An example script (shapeit2rfmix.py) can be found in the following Github repository: https://github.com/armartin/ancestry_pipeline
   This script requires *haps.gz files, a list of TCGA samples (https://github.com/jcarrotzhang/ancestry-from-panel/blob/master/GDAN_AIM/BRCA.notref), a list of reference samples (https://github.com/jcarrotzhang/ancestry-from-panel/blob/master/GDAN_AIM/BRCA.ref), and genetic_map* files downloaded in before you begin section step 11.

```
for i in {1..22}; do
python ~bin/ancestry_pipeline-master/shapeit2rfmix.py \
--shapeit_hap_ref BRCA/chr${i}.haps.gz \
--shapeit_hap_admixed BRCA/chr${i}.haps.gz \
--shapeit_sample_ref BRCA/chr${i}.sample \
--shapeit_sample_admixed BRCA/chr${i}.sample \
--ref_keep BRCA.ref \
--admixed_keep BRCA.notref \
```

```
--chr ${i} \
--genetic_map 1000GP_Phase3/genetic_map_chr${i}_combined_b37.txt \
--out BRCA/rfmix;
done
```

d. Run RFMix, with a minimum window size of 0.2 cM and a node size of 5 for random forest trees, using the following bash command as an example:

```
for i in {1..22}; do
python RunRFMix.py –forward-backward -w 0.2 -n 5 \
--num-threads 4 \
PopPhased BRCA/rfmix_chr${i}.alleles \
BRCA_rfmix.classes \
BRCA/rfmix_chr${i}.snp_locations \
-o BRCA/outputPopPhased.chr.${i}.withref;
done
```

Note: Other parameters in RFMix can be considered. For example, RFMix can perform expectation-maximization (EM) optimization of the model, using the -e option. Although it is also recommended to use -e to increase the accuracy of local ancestry identification, the -e option will significantly increase computational complexity.

e. Collapsing RFMix outputs (*.Viterbi.txt) into bed files. This step generates two *.bed files for each allele (A.bed and B.bed) for all samples, with the first four columns as chromosome, start position, end position and ancestry assignment, respectively.
An example script (collapse_ancestry.py) can be found in the following Github repository: https://github.com/armartin/ancestry_pipeline

f. Visualize the local ancestry calls using karyogram plots for inspection of technical artifacts (Figure 2A).

⚠ CRITICAL: It is recommended to visually inspect the reference samples to avoid upstream errors, such as mislabeling population assignments of reference samples in the class file etc. If the –forward-backward option is used in RFMix, local ancestry calls can be filtered with the posterior probability of a given ancestry less than 0.9.

g. Apply a custom script to calculate global ancestry (the percentage of AFR, EUR and EAS ancestry per individual) based on RFMix generated local ancestry calls. This step is to ensure that the local ancestry assignment can accurately reflect the global ancestry generated by a different algorithm.
An example script (lai_global.py) can be found in the Github repository: https://github.com/armartin/ancestry_pipeline

h. Exclude samples with ADMIXTURE estimated global SAS or AMR ancestry greater than 20% from Table S1. in Carrot-Zhang et al., 2020, and compare the global ancestry estimated from RFMix to ADMIXTURE estimated global ancestry using a Pearson correlation.

Note: Because most TCGA samples were from AFR, EUR or EAS populations, only samples in the 1000 Genomes Project from AFR, EUR or EAS populations were used for local ancestry identification in Carrot-Zhang et al., 2020. However, other parental populations should be used when analyzing other populations with mixed ancestries. For example, 1000 Genomes samples from Peru and Mexico can be used as reference samples when analyzing cancer samples from the Latin American populations.

5. **Alternative:** It is now recommended to use RFMix2 (https://github.com/slowkoni/rfmix) - a more user friendly version. RFMix2 takes phased VCF files as input directly, without requiring users to generate *.alleles and *.snp_locations. To use RFMix2, it is recommended to use Beagle for phasing, which outputs phased variants in VCF format that can be used as an RFMix2 input

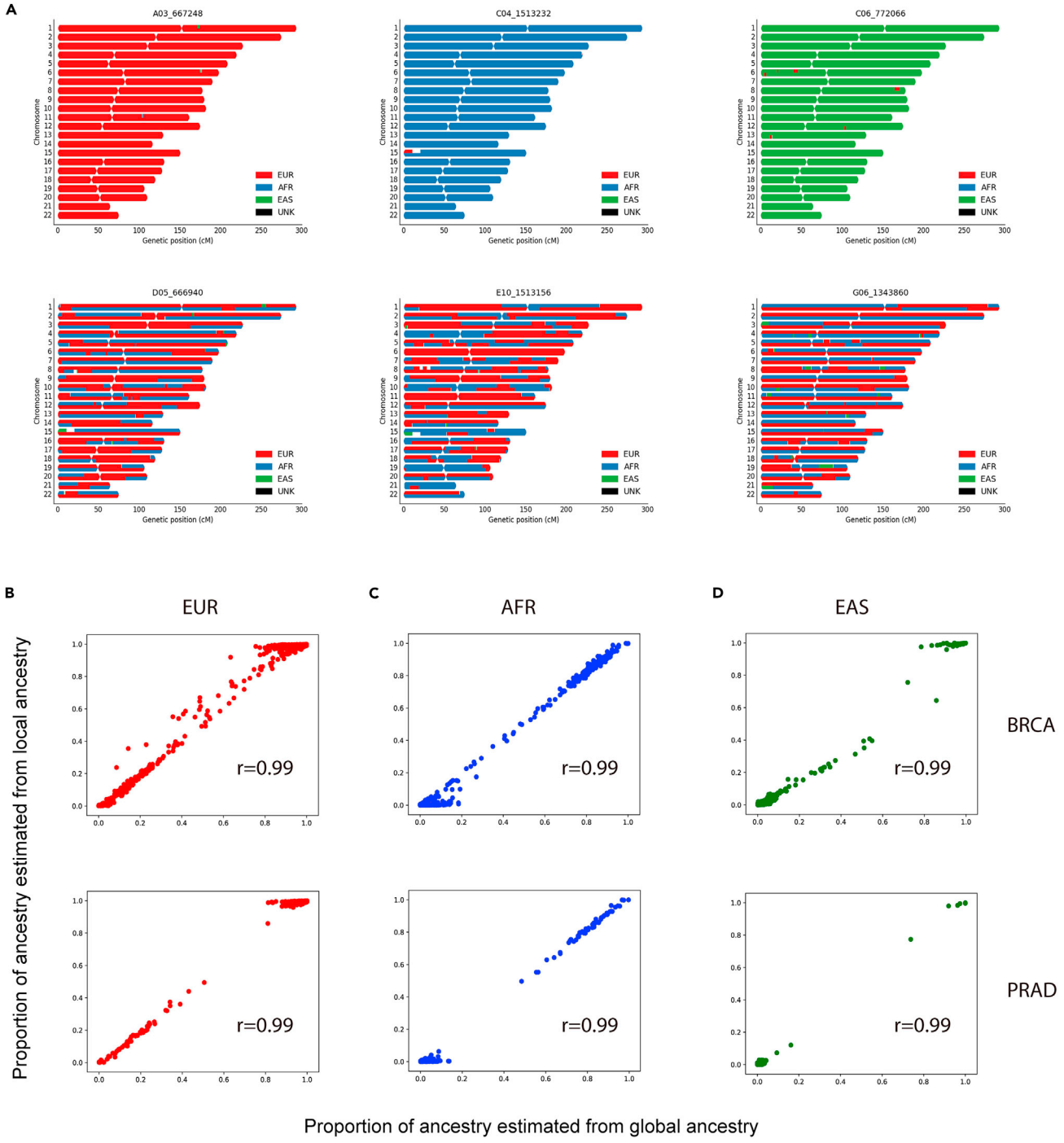


Figure 2. Accuracy assessment of local ancestry identification

(A) Karyograms of six individuals from the TCGA cohort. Upper panel: none-admixed individuals from EUR, AFR and EAS populations, respectively. Lower panel: individuals with mixed ancestries.

(B–D) (B) Correlation of global ancestry indicated by the proportions of EUR, (C) AFR and (D) EAS estimated from RFMix and ADMIXTURE. Upper panel: TCGA breast cancer cohort; Lower panel: TCGA prostate cancer cohort. r values are calculated from Pearson's correlation.

file. RFMix2 also reports global ancestry estimated based on local ancestries identified by their algorithm in their standard output file (*results.rfmix.Q) corresponding to the .Q output files from ADMIXTURE that can be compared to ADMIXTURE generated global ancestry directly.

a. To run Beagle, the following reference files are needed:
   i. Beagle reference vcf: http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/
   ii. Beagle reference map: http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a
b. An example bash command to run RFMix2 on Beagle phased VCF file:

```
for i in {1..22}; do
java -Xmx10 g -jar ~bin/beagle.08Jun17.d8b.jar \
gt=BRCA/chr${chr}.vcf.gz \
ref=~beagle/chr${chr}.1 kg.phase3.v5a.vcf.gz \
out=BRCA/$chr.beagle_phased \
map=~reference/beagle/plink.chr${chr}.GRCh37.map \
chrom=${chr} \
impute=false;
done
```

   For more information about phasing using Beagle, please refer to http://faculty.washington.edu/browning/beagle/beagle.html
c. An example bash command to run RFMix2 on Beagle phased VCF file:

```
for i in {1..22}; do

~bin/rfmix-master/rfmix -f BRCA/$chr.beagle_phased.vcf \

-r ~reference/beagle/chr${chr}.1 kg.phase3.v5a.vcf.gz \

-g genetic_map_chr${chr}_combined_b37.txt \

-m BRCA.ref \

-o $BRCA/chr${chr}.results –chromosome=${chr};

done
```

Association of local ancestry with cancer features and risk

Timing: 4–8h

This step demonstrates two types of analyses leveraging local ancestry calls to investigate germline influence on cancer development (Figure 1). The first analysis identifies genomic loci with enrichment of certain ancestry in the pan-cancer cohort, to investigate the contributions of ancestry at those loci to cancer risk. In the second analysis, ancestries are compared between loci harboring genes whose mRNA expression (or methylation) levels are associated with ancestry and those that do not, in order to evaluate whether ancestry at these specific loci–loci with identified molecular correlates of ancestry–contributes to cancer risk. These analyses should be focused on an admixed population, for example, the African American population with mixed AFR and EUR ancestry that has a relatively larger sample size compared to other admixed populations in the TCGA cohort.

For the genome-wide, pan-cancer analysis, perform the following:

6. Read in global and local ancestries for all TCGA samples:
   a. Read in the ancestry assignment from Table S1 of Carrot-Zhang et al., 2020. Select samples with AFR ancestry between 20%–80% for the downstream analyses.
   b. Read in bed files converted from RFMix outputs, and output regions with local ancestry (48,939 genomic loci) across all samples.

An example script (find_local_ancestry_count_blocks.py) can be found in the following Github repository: https://github.com/jcarrotzhang/ancestry-from-panel/tree/master/GDAN_AIM

7. Code regions identified from step 4 with both alleles from AFR ancestry as 2, with one AFR allele as 1, and loci with both alleles from EUR as 0.
8. Compute a z-score for each region with unique local ancestry identified from step 4. In Python, one can store coded ancestry of each locus in an array and then use the scipy.stats.zscore module to compute the z-score in each locus relative to the genome. This step again, only uses samples with AFR ancestry between 20%–80%.

An example script (local_ancestry_enrichment.py) can be found in the following Github repository: https://github.com/jcarrotzhang/ancestry-from-panel/tree/master/GDAN_AIM

9. Plot each locus and its corresponding z-score of local ancestry across the genome. Apply FDR correction to identify regions passed multiple testing corrections.

   Note: In theory, certain loci might be enriched for one ancestry over others across the admixed population so that the observed enriched regions may not be specific to cancer risks. Therefore, filtering out known regions associated with ancestry in healthy individuals, or a case-control study, is necessary when applicable. In practice, populations that have mixed within the last several hundred years are unlikely to exhibit such biases.

For the gene-specific analysis in pan-cancers, perform the following:

10. Similarly, the correlation of somatic alterations in a specific gene and the local ancestry of the locus harboring the gene of interest, while controlling for global ancestry, can be tested, using the following logistic regression model in Python:

mutation status ~ local ancestry + global EUR ancestry + global AFR ancestry

Global ancestries can be obtained from Table S1 of Carrot-Zhang et al., 2020. A significant p-value indicates a *cis* effect of germline variation to somatic alteration in the gene of interest.

An example script (find_local_ancestry_by gene.py) can be found in the following Github repository: https://github.com/jcarrotzhang/ancestry-from-panel/tree/master/GDAN_AIM

This script selects samples with AFR ancestry between 20%–80%, finds the local ancestry of the locus of interest in each sample, and correlates the local ancestry with mutation status of the gene of interest, using the regression model described above. The local ancestry is coded as 0 for homozygous EUR ancestry, 1 for heterozygous AFR ancestry and 2 for homozygous AFR ancestry.

⚠ CRITICAL: Because local ancestry can be significantly correlated with global ancestry, when testing the association between local ancestry and somatic mutation status, global ancestry should be added as a covariate to the regression model. A significant local ancestry–phenotype correlation independent of global ancestry may suggest germline influence on the phenotype, as local ancestry is not expected to be associated with ancestry-associated environmental exposure.

This analysis is performed in a pan-cancer fashion. Due to limited sample size, we were unable to perform tissue-type specific analysis.

For the second analysis, restricted to loci with known molecular correlates of ancestry (in this case, mRNA expression in *cis*), perform the following:

11. Keep loci associated with AFR ancestry (z-score >3) or EUR ancestry (z-score <−3) calculated from step 8, or downloaded from Table S1 Carrot-Zhang et al., 2020.
12. For the remaining loci from step 11, use the Bedtools intersect option to locate loci with differentially expressed genes between populations from Table S4 of Carrot-Zhang et al., 2020. Label those loci as "ancestry-associated expression" loci. Label other loci as "non-ancestry-associated expression" loci.
13. Compare z-scores between "ancestry-associated loci" and "non-ancestry-associated loci", using Wilcoxon's rank test.

Step 11 to 13 can also be applied to other loci with ancestry-associated molecular features, such as methylation differences, to assess their association with cancer risk.

EXPECTED OUTCOMES

Local ancestry identification accuracy

When inspecting the karyograms of TCGA samples (Figure 2A) from self-reported African American patients, one should expect varying admixture contributions from AFR and EUR ancestries, but less contribution from EAS ancestry. If the local ancestry calls are filtered by the posterior probability, one would expect few regions with unknown ancestry assignment, especially for regions close to the ancestry switch locations.

One should expect a high correlation (Pearson's $r > 0.98$) between global ancestry estimated from ADMIXTURE (Alexander and Lange, 2011) and the genome-wide average of local ancestry-based determinations obtained from RFMix (Figure 2B), suggesting an agreement of the ancestry estimates based on two different algorithms. The local-global ancestry comparison can be performed using global ancestry inferred from other methods (i.e., principal component analysis-based approaches).

Association of local ancestry with cancer risk

See Carrot-Zhang et al., 2020 Figure 1D for expected outcomes of genome-wide pan-cancer association of local ancestry with cancer risk.

A case-case admixture mapping can be used to identify a marker locus associated with a certain cancer phenotype, by comparing the observed frequency of the risk ancestry at a marker locus to the rest of the genomes of the cancer cases. Sample size may limit the detection of a specific locus that is significantly enriched in AFR or EUR ancestry in TCGA samples from admixed patients, even in the pan-cancer analysis. A power calculation, considering risk allele frequency, penetrance, and degree of admixture in the population may be helpful prior to this type of analysis.

A case-control admixture mapping compares the observed frequency of the risk ancestry at a marker locus from the cases to the frequency observed in the controls. The controls can be samples without the molecular feature of interest. For example, to test the association of local ancestry with a certain somatic mutation, samples from the same cancer type without carrying the mutation can be used as controls. However, none of the TCGA cancer types provide sufficient statistical power for a genome-wide case-control comparison, and therefore, one would not expect to identify a significant local ancestry-somatic alteration association using the TCGA data set.

Association of local ancestry at loci known ancestry-associated molecular features, with cancer risk

See Carrot-Zhang et al., 2020 Figure S7D and S7E for expected outcomes.

By restricting analyses of ancestry-associated cancer risk to loci for which ancestry-associated molecular features have been detected, one would hope to reduce the number of hypotheses tested while enriching for signal, thereby improving power. Indeed, this analysis detected an enrichment of AFR ancestry in loci with ancestry-associated mRNA expression (Carrot-Zhang et al., 2020). Notably,

datasets such as TCGA provide methylation and gene expression data for tumor samples; those data types are not available for matched normal samples. However, most ancestry-associated molecular features appear to be shared between tumor and normal tissue (Carrot-Zhang et al., 2020).

LIMITATIONS

Studying the differences between cancer genomes from patients with different ancestries, and more broadly germline effects on somatic alterations, can be particularly challenging because it requires both germline and somatic data from the same patient. The TCGA cohort of over 10,000 cases across 33 cancer types with matched germline samples is such a data set for joint genomic and ancestry analysis. However, it is important to note that population diversity for most TCGA cancer types is poor, especially for patients from admixed populations. For example, we were unable to perform case-case admixture mapping analysis in tissue-type specific fashion. Moreover, although we identified several ancestry-specific somatic alterations in cancers using the TCGA cohort (e.g., *FBXW7* in multiple cancer types), we did not identify a genomic region harboring a germline allele associated with *FBXW7*-mutant cancers (either in case-case studies or using the case-control approach) in the pan-cancer analysis, possibly due to limited sample size of admixed samples with AFR and EUR ancestries.

Another mixed patient population, Latin American patients, is severely under-represented in the TCGA cohort. We were therefore unable to evaluate molecular features and cancer risk that correlate with Native American (NAT) vs. other ancestries (Martin et al. 2017; Carrot-Zhang et al. 2021; Rodriguez et al. 2020). The NAT ancestry is also under-represented in 1000 Genomes reference samples. A larger data set from diverse populations will enable the discovery of germline alleles associated with cancer molecular features and ancestry-associated risks.

TROUBLESHOOTING

Problem 1

Step 3–4

A lack of local ancestry association analysis for admixed patients with NAT ancestry.

Potential solution

To include AMR and SAS patients in the TCGA cohort for the association analysis with local ancestry, local ancestry identification can be performed by adding AMR and SAS samples to the reference panel, and then assigning five (AFR, EUR, NAT, SAS and EAS) ancestries to each patient (Martin et al. 2017; Carrot-Zhang et al. 2020; Rodriguez et al. 2020). Because the AMR population is an admixed population, one may want to only include samples with inferred NAT ancestry greater than 20% in the reference panel.

Problem 2

Step 3–4

The reference panel for local ancestry identification contains samples from multiple sub-populations.

Potential solution

To avoid local ancestry identification biases caused by skewed reference sample sizes from sub-populations, it is recommended to use a minimum node size of 5 (-n 5) in RFMix, as shown in step 4d.

Problem 3

Step 3–4

The reference panel for local ancestry identification contains samples with mixed ancestries.

Potential solution

If the reference panel contains samples with mixed ancestries, one can use the –use-reference-panels-in-EM in RFMix v1.5.4, or –reanalyze-reference in RFMix2, coupled with the EM optimization (-e) to update the ancestry assignment for reference haplotypes. Including the reference panel in the EM can also improve the results when the reference panel is small.

Problem 4

Step 10

A lack of genome-wide association analysis of local ancestry to somatic alterations in ancestry-associated genes.

Potential solution

If the sample size is sufficient for a genome-wide association analysis, step 10 can be extended to test any region to identify *trans* alleles associated with somatic alterations in certain ancestry-associated cancer genes. Multiple testing correction (e.g., Bonferroni) should be performed in the genome-wide analysis.

Problem 5

Step 10

A lack of matched normal samples in the cancer cohort for local ancestry and somatic association analysis.

Potential solution

It was suggested that tumor purity and somatic alterations had minimal influence on the accuracy of inferring germline haplotype and genetic ancestry (Carrot-Zhang et al. 2021; Gusev et al. 2021). It is possible to accurately infer both global and local ancestry from tumor-only DNA, and correlate ancestry with somatic phenotypes, when genotyping data from matched germline samples is not available. One may consider controlling for certain somatic features (e.g., tumor purity, somatic copy number alterations) in the local ancestry and somatic association analysis.

RESOURCE AVAILABILITY

Lead contact

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

Materials availability

This study did not generate new unique materials or reagents.

Data and code availability

Data used and generated are listed in the key resources table, and the GDC Publication Website (https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). Code generated for local ancestry-related analyses is available from the following URL https://github.com/jcarrotzhang/ancestry-from-panel/tree/master/GDAN_AIM.

CONSORTIA

Ashton C. Berger, Matthew Meyerson, Katherine A. Hoadley, Ina Felau, Samantha Caesar-Johnson, John A. Demchok, Michael K.A. Mensah, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Nyasha Chambwe, Theo A. Knijnenburg, A. Gordon Robertson, Christina Yau, Christopher Benz, Kuan-lin Huang, Justin Newberg, Garret Frampton, R. Jay Mashl, Li Ding, Alessandro Romanel, Francesca Demichelis, Rosalyn W. Sayaman, Elad Ziv, Peter W. Laird, Hui Shen, Christopher K. Wong, Joshua M. Stuart, Alexander J. Lazar, Xiuning Le, Ninad Oak

ACKNOWLEDGMENTS

We thank the Cancer Genome Atlas Research Network, the U.S. National Cancer Institute 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 NCI K99 award. We thank Anab Kemal for project administration. We are grateful for advice from numerous colleagues, TCGA and GDAN collaborators, and the GDC technical support team.

AUTHOR CONTRIBUTIONS

J.C.-Z., A.D.C., and R.B. wrote the manuscript. J.C.-Z., S.H., J.S.D., and W.Z. provided data and performed the analyses. A.K. provided project administration.

DECLARATION OF INTERESTS

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

REFERENCES

Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics *12*, 246.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.

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.

Carrot-Zhang, J., Soca-Chafre, G., Patterson, N., Thorner, A.R., Nag, A., Watson, J., Genovese, G., Rodriguez, J., Gelbard, M.K., Corrales-Rodriguez, L., et al. (2021). Genetic ancestry contributes to somatic mutations in lung cancers from admixed Latin American populations. Cancer Discov. *11*, 591–598.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Gusev, A., Groha, S.M., Taraszka, K., Semenov, Y.R., and Zaitlen, N. (2021). Constructing germline research cohorts from the discarded reads of clinical tumor sequences. medRxiv. https://doi.org/10.1101/2021.04.09.21255197.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell *173*, 400–416.e11.

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.

Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.

Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.

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., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rodriguez, D.A., Sanchez, M.I., Decatur, C.L., Correa, Z.M., Martin, E.R., and Harbour, J.W. (2020). Impact of genetic ancestry on prognostic biomarkers in uveal melanoma. Cancers *12*, 3208.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. Cell *173*, 321–337.e10.