HHS Public Access

Author manuscript

Cancer Cell. Author manuscript; available in PMC 2021 May 11.

Published in final edited form as:

Cancer Cell. 2020 May 11; 37(5): 639–654.e6. doi:10.1016/j.ccell.2020.04.012.

Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer

Jian Carrot-Zhang[1,2,3,24], **Nyasha Chambwe**[4,24], **Jeffrey S. Damrauer**[5,24], **Theo A. Knijnenburg**[4,24], **A. Gordon Robertson**[6,24], **Christina Yau**[7,8,24], **Wanding Zhou**[9,24], **Ashton C. Berger**[1,2,24], **Kuan-lin Huang**[10,24], **Justin Newberg**[11,24], **R. Jay Mashl**[12,25], **Alessandro Romanel**[13,25], **Rosalyn W. Sayaman**[14,15,25], **Francesca Demichelis**[13], **Ina Felau**[16], **Garret Frampton**[11], **Seunghun Han**[2,3], **Katherine A. Hoadley**[5], **Anab Kemal**[16], **Peter W. Laird**[9], **Alexander J. Lazar**[17], **Xiuning Le**[18], **Ninad Oak**[19,20], **Hui Shen**[9], **Christopher K. Wong**[21], **Jean C. Zenklusen**[16], **Elad Ziv**[14], **Cancer Genome Atlas Analysis Network AguetFrancois1DingLi11DemchokJohn A.16MensahMichael K.A.16TarnuzzerRoy16WangZhining16YangLiming16AlfoldiJessica1KarczewskiKonrad J.1MacArthurDaniel G.1MeyersonMatthew1,2,3BenzChristopher7StuartJoshua M.21**, **Andrew D. Cherniack**[1,2,3,26,27,*], **Rameen Beroukhim**[1,2,3,22,23,26,*]

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

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

[6.]British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, V5Z4S6, Canada

[7.]Buck Institute for Research on Aging, Novato, CA, 94945, USA

[8.]Department of Surgery, University of California, San Francisco, San Francisco, CA, 94115, USA

[*]Correspondence: achernia@broadinstitute.org (ADC), Rameen_Beroukhim@dfci.harvard.edu (RB).

Author Contributions

J.C-Z., K-l.,H., S.H., R.J.M, J.N., A.R., R.W.S, F.D., and E.Z. generated ancestry calls. J.C-Z., N.C., A.C.B, J.S.D., K-l.H., T.A.K., A.G.R., C.Y., W.Z., and J.N., analyzed the data. A.K., I.F., J.C.Z. and The Cancer Genome Atlas Research Network provided project administration. A.D.C. and R.B. provided supervision. J.C-Z., N.C., J.S.D., T.A.K., A.G.R., C.Y., W.Z., A.C.B., K-l.H, X.L., K.A.H, P.W.L, A.D.C., and R.B. wrote and all authors reviewed the manuscript.

Declaration of Interests

J.N. and G.F. are employees and shareholders of Roche (Foundation Medicine). X.L. receives a consultant/advisory fee from Eli Lilly, AstraZeneca, EMD Serono, and research funds from Eli Lilly, Boehringer Ingelheim. P.L. is on the Scientific Advisory Boards of AnchorDx and Progenity Inc. A.D.C receives research funding from Bayer. R.B. owns equity in Ampressa Therapeutics and receives research funding from Novartis.

[9.]Van Andel Research Institute, Grand Rapids, MI 49503, USA

[10.]Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY 2129, USA

[11.]Foundation Medicine, Inc., Cambridge, MA, 02141, USA

[12.]Department of Medicine, Washington University in St. Louis, St. Louis, MO, 63110, USA

[13.]Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Via Sommarive 9 Povo (TN) 38123 Italy

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

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

[16.]National Cancer Institute, Bethesda, MD 20892, USA

[17.]Departments of Pathology, Genomic Medicine, and Translational Molecular Pathology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

[18.]Department of Thoracic and Head and Neck Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

[19.]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA

[20.]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA

[21.]Department of Biomolecular Engineering, Center for Biomolecular Sciences and Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

[22.]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

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

[24.]These authors contributed equally

[25.]These authors contributed equally

[26.]Corresponding authors

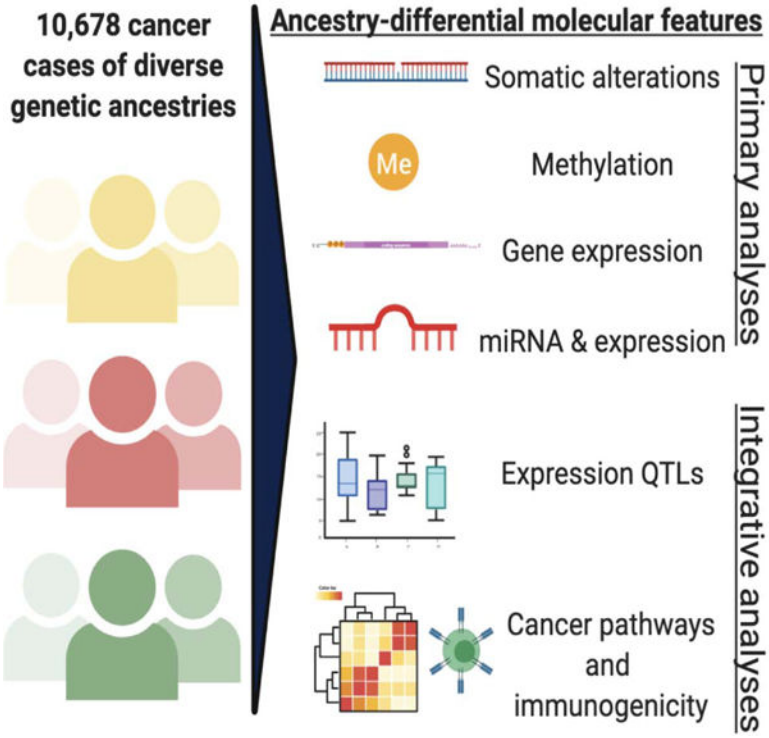
[27.]Lead contact

Summary

We evaluated ancestry effects on mutation rates, DNA methylation, and mRNA and miRNA expression among 10,678 patients across 33 cancer types from The Cancer Genome Atlas. We demonstrated that cancer subtypes and ancestry-related technical artifacts are important confounders that have been insufficiently accounted for. Once accounted for, ancestry-associated differences spanned all molecular features and hundreds of genes. Biologically significant differences were usually tissue-specific but not specific to cancer. However, admixture and pathway analyses suggested some of these differences are causally related to cancer. Specific

findings included an increased *FBXW7* mutations in patients of African origin, decreased *VHL* and *PBRM1* mutations in renal cancer patients of African origin, and decreased immune activity in bladder cancer patients of East Asian origin.

Graphical Abstract



In Brief:

Analyzing mutation rates, gene and miRNA expression, and DNA methylation across tumor types, Carrot-Zhang et al. separate confounders and identify ancestry-related effects that potentially explain cancer etiology and treatment.

Introduction

People of different ancestries exhibit varying germline genetics (Rosenberg et al, 2002; Price et al, 2006) and tend to encounter different exposures, resulting in varying cancer incidence, outcome (Freedman et al, 2006; Yang et al, 2011), and molecular characteristics (Shigematsu et al, 2005). However, a comprehensive accounting of ancestry-associated differences in molecular features has not been performed across cancers or even non-neoplastic tissues. Moreover, analyses of ancestry associations rarely account for varying prevalence of cancer subtypes across ancestries (Sanchez-Vega et al, 2018; Yuan et al, 2018), which can obscure differences within subtypes.

The Cancer Genome Atlas (TCGA) is the largest and most comprehensive multi-omics oncology cohort (Hutter and Zenklusen, 2018), rendering possible the simultaneous

assessment of ancestry associations in mRNA and miRNA expression and DNA methylation and mutation across 33 cancer types. Such analyses can improve our understanding of molecular and cellular effects of ancestry in at least four ways. *First*, by detecting novel ancestry-associated molecular features and cancer types. *Second*, by determining whether ancestry effects are cancer- or tissue-type specific, or common across types. *Third*, by increasing power to detect the common effects, using combined data across cancer types. And *fourth*, by integrating cross-platform analyses. The wealth of molecular data also enables accurate cancer subtype classification (Sanchez-Vega et al, 2018), enabling precise accounting of cancer subtype-ancestry associations.

We sought to exploit these advantages to improve our understanding of the molecular and cellular effects of ancestry across tissue and cancer types.

Results

Determination of genetic ancestry

We determined the ancestry of each TCGA patient using five independent classification methods, each employing SNP array and/or exome sequencing data (Figure 1A). Among 9,257 patients for whom at least three of the methods provided calls, 9,090 (98.1%) exhibited complete agreement, and 99.7% of non-admixed patients exhibited concordance with prior ancestry assignments (Yuan et al, 2018). Discordant calls were primarily differences in the relative degree of ancestry assignments in admixed patients.

The final data spanned 10,678 individuals of primarily European (“EUR”; n=8,836), East Asian (“EAS”; n=669), African (“AFR”; n=651), Native/Latin American (n=41), South Asian (n=27), or at least 20% admixed (n=454) descent (Figure 1B, Table S1) and 33 cancer types, of which 13 were divided into pre-defined subtypes. In several cases, ancestries were associated with different subtypes (Figure 1C). Admixed individuals were further distinguished by their primary ancestries: African-Admixed (n=343), European-Admixed (n=68), South Asian-Admixed (n=24), East Asian-Admixed (n=7) and Undetermined (n=12).

We also determined local ancestry across 70,748 genomic loci in all samples. These local calls appeared to be accurate: for example, the summed local ancestry calls were nearly identical with our estimated global ancestry (Pearson’s r>0.99). Among the 1,076 samples with only EUR and at least 10% AFR ancestry, we also evaluated whether individual loci were enriched for AFR or EUR ancestry relative to their genome-wide levels of admixture. No single locus reached statistically significant levels of enrichment after controlling for multiple hypotheses (Figure 1D, Table S1).

Next, we explored associations between ancestry and molecular data. For each data type (somatic alterations, methylation, mRNA and miRNA expression, and cross-platform data), we performed pan-cancer and tissue-specific multivariate regression analyses controlling for cancer type and subtype (Figure 1C; STAR Methods). Because most samples were EUR, we used them as a reference to which we compared EAS or AFR data.

The sample size provided substantial statistical power. In pan-cancer analyses, we had greater than 90% power to detect ancestry-associated somatic alterations of at least 10% prevalence and an odds ratio (OR) greater than two in the AFR-EUR comparison (controlling for cancer type as a confounder; Figure S1A); and in both AFR and EAS analyses to detect methylation differences that exceed the standard deviation in the data (0.2; Figure S1B) or z-score differences in population means of at least 0.15 in mRNA and miRNA expression (Figure S1C). Tissue-specific analyses had lower power (Figure S1B).

Somatic alterations associated with ancestry

With respect to the overall burden of somatic alteration, we initially observed significant correlations between AFR and aneuploidy (p=0.004) and EAS with tumor mutation burden (TMB; p=0.02) in pan-cancer analyses. However, controlling for cancer subtype eliminated both correlations (Figure S2A–D).

We then evaluated somatic mutation (single nucleotide variant/indel) and copy-number alteration (SCNA) frequencies at the gene level. In the pan-cancer analysis, we initially identified significant differences in three genes in AFR individuals, and one gene in EAS, relative to EUR (Figure 2A, Figure S2E,F). Two of these in the AFR-EUR comparison have been described (Yuan et al, 2018): enriched mutations of *TP53* in AFR samples and of *PIK3CA* in EUR samples. We also observed enriched *CCND1* amplification in EAS samples and *FBXW7* mutation in AFR samples. However, after subtype correction, only the *FBXW7* finding remained significant (FDR q=0.07), highlighting how variations in cancer subtype frequencies can confound ancestry associations. Pan-cancer analyses restricted to cancer types with at least 10 samples in each ancestry produced similar results, though fewer were statistically significant (Figure S2G).

Three additional datasets supported the association between *FBXW7* mutations and AFR ancestry. First, international Cancer Genome Consortium Pan-Cancer Analysis of Whole Genomes (ICGC PCAWG) data (excluding TCGA samples) and MSK-IMPACT data, which primarily included EUR samples (PCAWG also included many EAS samples) exhibited few *FBXW7* mutations (20/1225=1.6% in PCAWG and 360/10336=3.5% in MSK-IMPACT) relative to the 6.7% mutation rate we had observed in AFR samples. This supports but does not formally validate the *FBXW7*-AFR association. However, an independent Foundation Medicine (FMI) cohort of 60,454 tumors from 12 cancer types (Table S2) also exhibited more frequent *FBXW7* mutations in AFR relative to EUR samples (Fisher's exact p=0.01), and specifically in HNSC (16/134 AFR and 117/2268 EUR samples, p=0.007) and UCEC (116/730 AFR and 520/4333 EUR samples, p=0.005).

Within cancer types, we observed four genes with differential mutation rates, with two in a single cancer type: kidney clear-cell carcinomas (KIRC), in which AFR samples lacked *VHL* and *PBRM1* mutations (OR 0.37 and 0.25, respectively; FDR q=0.06 and 0.04). EAS bladder and esophageal cancers were enriched in *HRAS* (OR=6.6; q=0.03), and *NFE2L2* (OR=11.6, q=0.07) mutations, respectively (Figure 2B, Table S2). The finding that only pan-cancer analyses identified differential mutation rates in *FBXW7* indicates that these differences spanned more than one cancer type. However, the finding that most ancestry

associations were identified only in individual cancer types, despite smaller sample numbers and less power, suggests that these ancestry effects tend to be cancer type-specific.

We validated the associations between *VHL*, *PBRM1*, *HRAS* and *NFE2L2* mutations and ancestry in external cohorts. We observed fewer *VHL* and *PBRM1* mutations among AFR KIRC samples in the FMI cohort (OR=0.20 and 0.44 respectively; p<0.01 in each case; Table S2). An independent study (Pena-Llopis et. al. 2012, Krishnan et. al. 2016) also found significantly more *VHL* mutations in EUR (101/125) than AFR KIRC samples (4/10; p=0.008). In this cohort, fewer *PBRM1* mutations were also detected in AFR (4/10) than EUR (62/125) samples, but the association was not significant, possibly due to the small number of AFR samples. We observed enrichment of *HRAS* mutations in EAS (5/89) relative to EUR (64/2482) bladder cancers in the FMI cohort and in two additional datasets (Nassar et al, 2019 and Wu et al, 2019), where 22/448 EUR and 7/69 EAS samples carried those mutations. Neither of these cohorts reached statistical significance on its own, but the combined data did (Fisher's exact p=0.004). For *NFE2L2* in ESCA, a study from East Asia (Chang et. al, 2017) reported 7/94 samples with mutations-a similar rate to the 5/117 EUR samples with the mutation in TCGA. Although prior studies suggested that *NFE2L2* mutations are enriched in EAS esophageal squamous cell carcinoma (Deng et. al, 2017), we conclude that we could not validate *NFE2L2*.

We also validated the *VHL* and *PBRM* associations in KIRC patients with admixed AFR and EUR ancestry. We hypothesized that these mutations would be observed at rates that are proportional to the fraction of the genome with EUR ancestry. This was indeed the case for *VHL* in both TCGA and FMI cohorts (Wilcoxon p=0.02 and p<0.001, respectively) and for *PBRM1* in TCGA (p=0.007; Figure 2C,D).

We next looked for evidence that germline genetics at the *FBXW7*, *VHL*, and *PBRM1* loci contribute to cancer formation. To that end, we asked whether any of these loci were locally enriched for AFR or EUR ancestry among samples with at least 10% AFR ancestry, after controlling for global EUR and AFR ancestry rates (Figure 1D). In each case, the ancestry with the higher mutation rates was enriched at the gene locus: AFR ancestry at *FBXW7* (OR=1.001) and EUR ancestry at *VHL* (OR=1.822) and *PBRM1* (OR=1.221). However, none were statistically significant. We conclude that the germline features at these loci may not be associated with cancer.

To assess the contribution of environmental exposures to differences in mutation frequency between ancestries, we compared 57 mutational signatures between AFR or EAS samples and EUR samples (Alexandrov et. al. 2019). These signatures, derived from mutational patterns across trinucleotide contexts, often reflect mutagen exposure. We did not observe significant differences in the AFR-EUR pan-cancer comparison, suggesting that mutagen exposures were not major confounders in the differences in *FBXW7*, *VHL*, or *PBRM1* mutation rates. We did find six signature associations (Table S2) in the EAS-EUR pan-cancer analysis. Two signatures frequently observed in liver cancer (Signatures 16 and 24), Signature 9 (related to AID activity), and Signature 26 (related to defective mismatch repair) were enriched in EAS samples, and two signatures related to APOBEC activity (Signatures 2 and 13) were enriched in EUR samples. We observed no significant differences in

signatures between EAS and EUR BLCA samples, however, suggesting that mutagen exposures were not major confounders in the difference in *HRAS* mutation rates between these samples. However, other exposures might still shape mutation rates across ancestries.

We also tested the association of ancestry with chromosome arm-level SCNAs. We observed no such associations in the pan-cancer analysis, but found two in cancer type-specific analyses. First, 3p loss, encompassing both *VHL* and *PBRM1*, was more frequent in EUR than AFR KIRC samples (q=0.02; Figure 2E). This, along with our prior finding that *VHL* and *PBRM1* mutations are enriched in EUR KIRC samples, indicates that these genes are biallelically inactivated more often in EUR KIRC samples. Prior work noted the disparity in *VHL* mutations, but not loss of chromosome 3p, between TCGA AFR and EUR KIRC samples (Krishnan et al, 2016). We also found that chromosome 4q loss was more frequent in AFR than EUR COAD samples.

Pan-cancer analysis uncovers regions of ancestry-differential DNA methylation

In the pan-cancer analysis, we found statistically significant (F-test p value < 0.05) differences in methylation between ancestries in 94,012 of the 482,421 HM450 array CpG sites, comprising 19% of all tested probes (Fig S3A). More ancestry associations were identified in this pan-cancer analysis than in any single cancer type. However, these associations tended to have small effect sizes, predominantly below a change of 0.1 (Fig S3B–C), and therefore unlikely to be biologically significant. We conclude that many of these findings result from the statistical power of our large dataset, rather than representing substantive differences between ancestries.

When restricting to differences that are both significant and large enough to be biologically meaningful (methylation change >= 0.1), we found very little of the cancer genome was differentially methylated across ancestries. In pan-cancer analyses, we initially identified only 3,001 (0.6%) such CpGs. Moreover, 75% of these are likely due to artifact (Figure 3A) caused by SNPs in the five bases at the 3'-end of the DNA methylation probes (Zhou et al, 2017). Only 4.2% of non-ancestry associated CpGs were associated with such SNPs, and this artifact was not overrepresented in probes that exhibit tumor-type-associated methylation differences (Figure S3D). Probes subject to other types of technical artifact were only slightly enriched in ancestry-differential CpGs (Figure S3E). After removing these problematic probes, only 374 CpG sites exhibited significant ancestry associations. These results highlight the importance of controlling for artifacts related to germline variants when comparing methylation data across ancestries.

Similar to somatic genetic alterations, we observed more significant and potentially biologically meaningful methylation changes in cancer type-specific than pan-cancer analyses, despite the decreased power in the former. Across the six cancer types we analyzed, an average of 3,116 sites exhibited ancestry-associations (range 474–12,176), eight times the number in the pan-cancer analysis. However, when we performed the same analysis on the 65 "rs" probes on the array that interrogate SNP variants and would therefore often differ between ancestries, we detected significant differences for 63 probes in the pan-cancer analysis, and only 43 probes on average in cancer-specific analyses, consistent with the greater power in pan-cancer analyses (Figure S3F).

As might be expected, sites that were significant in pan-cancer analyses exhibited similar ancestry associations in most cancer-specific analyses (Figure 3B), but most significant sites in cancer type-specific analyses did not. Some differences might be caused by residual subtype heterogeneity. For instance, when we did not distinguish BLCA luminal and basal subtypes, we detected considerably more ancestry-differential sites (36,926 rather than 16,716), highlighting the need to account for cancer subtype.

Although the methylation differences were largely tissue-specific (Figure S3G), they appeared to be remarkably consistent across cells within those tissues. Ancestry-associations could reflect mixed populations of cells with different methylation patterns, whose composition varies by ancestry, or different methylation signatures within each population. Among the 374 pan-cancer ancestry-differential CpG sites, 204 exhibited multimodal distributions in methylation signal ($p<0.05$, Hartigan's unimodality test; four exemplary cases are shown in Figure S3H), likely representing biallelic lack of methylation, monoallelic methylation, and biallelic methylation. Mixed populations of cells with different signatures would tend to fill in the three modes, making them more uniform.

We considered genes associated with multiple ancestry-differential methylation sites to have the greatest support for ancestry effects (Figure 3C). Forty-one genes were supported by at least two probes, and ten genes were supported by at least four (Table S3). These ten include known methylation quantitative trait loci (meQTLs) such as *SPATC1L* (Heyn et al, 2013) and *PM20D1* (Sanchez-Mut et al, 2018); genes previously recognized as variably methylated such as *HOOK2* (Kraus et al, 2015); and genes for which variation in methylation has not been described, such as *FLJ26850*, *PACS2*, and *FAAP20* (Figure 3D). Among nine of these ten genes, all associated probes exhibited similar ancestry effects. For example, all four *FAAP20* sites were more frequently methylated in AFR samples while all four *HOOK2* sites were less frequently methylated in those samples.

The top gene, *SPATC1L,* uniquely exhibited opposite ancestry associations across its associated probes (Figure 3B). These are related to different functional elements, comprising probes that cluster at either the gene's promoter or its transcription termination site (TTS) (Figure 3C–D). Prior reports also described coordinated differences in methylation at the promoter and TSS sites across haplotypes (Heyn et al, 2013). Promoter methylation of *SPATC1L* (shown by cg12016809) was negatively associated with *SPARC1L* mRNA levels in four ancestry groups (Figure 3E), suggesting an impact on gene transcription, with more methylation and less expression in AFR samples.

The observation that multiple neighboring loci show coordinated DNA methylation patterns suggests that their ancestry-related differences were not due to technical artifact. We attempted to further validate the 374 ancestry-differential sites from the pan-cancer analysis in two independent datasets that assayed 149 non-neoplastic hematopoietic samples (Stunnenberg et al, 2016) and 49 TCGA samples (40 tumors and 9 normals) (Zhou et al, 2018) using the orthogonal technology of whole genome bisulfite sequencing (WGBS). Among the 374 sites, 343 were also probed in the TCGA WGBS dataset, and all exhibit differential methylation. Over 80% (277 sites) had at least one neighboring CpG with highly correlated methylation (Spearman's rho > 0.7), and over 40% had more than 10 such

neighboring CpGs (Figure S3I). Among the 149 hematopoietic samples, the regional differences in ancestry that we had observed in TCGA cancers again appeared as components of larger variably methylated regions (VMRs, consistent with ancestry-associated differentially methylated regions: A-DMRs) that encompassed multiple concordantly methylated CpGs (Figure 3F). Although we focused on regions supported by at least two probes, many loci with only one differentially methylated probe show similar patterns in the WGBS validation data, such as *S100A14* (Figure 3G). These findings indicate that many ancestry-associated methylation differences reflect regional chromatin states, and that ancestry-associated methylation effects in cancers often reflect similar patterns in normal tissue.

We queried four additional datasets to validate the ancestry associations we had detected in TCGA, all representing non-neoplastic tissue. Two represented individuals of AFR, EUR and EAS ancestry, similar to TCGA: one of whole blood (Heyn et al, 2013, GSE36369) and the other of brain tissue (Guintivano et al, 2013, GSE41826). Two represented ancestries not well covered in TCGA: one of umbilical cord blood in three EAS populations (Teh et al, 2014, GSE53816) and another of lymphoblastoid cell lines from five populations of more specific ancestry such as Mozabites and Cambodians (Carja et al, 2017, GSE101431) (Figure S4A–D). We found that 80% and 60% of the ancestry-differential CpG sites in TCGA validated in the whole blood and brain tissue data, respectively (Figure S4A–B, group "++"), with substantially lower rates of significant ancestry differences for random sites without ancestry associations in TCGA data (Figure S4A–B, group "−"). We observed lower validation rates in the two datasets that differ from TCGA in ancestry composition, as expected: 40% and 8% respectively (Figure S4C–D)-but still significantly higher than in randomly selected CpGs. Further validating these ancestry-specific differences, we found strong positive correlations between the magnitudes of methylation preferences in the validation datasets and TCGA (Figure S4E–H). We conclude that most of the ancestry-specific methylation differences we identified can be validated. The fact that the validation sets were mostly disease-free tissues also suggests that most of the differences apply to healthy tissue as well as cancer.

mRNA associations with genetic ancestry

In pan-cancer analyses correcting for batch, tissue type, and subtype, we found significant mRNA expression differences between AFR and EUR samples for 327 genes and between EAS and EUR samples for 654 genes (Figure 4A–B, Table S4). These two lists had 85 genes in common, including 35 with higher expression in EUR samples, 31 with lower expression in EUR samples, and 19 for which expression in EUR samples was between that of AFR and EAS samples (Table S4). Prior GTEx consortium analyses identified 221 protein-coding genes associated with AFR ancestry (Mele et al, 2015), of which 44 overlapped with our analysis (p=2.2e-16) (and exhibited similar effect sizes; r^2=0.84, p=3.4e-18) (Figures 4C–D, Table S4).

We observed fewer significant associations within cancer types, with a maximum of 61 in the EAS-EUR hepatocellular carcinoma (LIHC) analysis (Table S4). Several genes were consistently identified across cancer types. For example, four AFR-associated genes

(*CRYBB2, NOTCH2NL, LOC90784, PPIL3)* and eight EAS-associated genes (*POM121L10P, TSPAN10, THOC3, XKR9, LOC162632, SIRPB2, MGC23270, DDX11L2, TGOLN2)* were significant in at least 33% of the cancer types (Figures 4A–B and 4E–F).

However, as in the methylation analyses above, the effect sizes in the mRNA analyses were much higher in individual cancer types: over 60% of ancestry-associated genes in these analyses had coefficients greater than one, relative to only 2% in the pan-cancer analyses (EAS-EUR 12/654, AFR-EUR 7/327) (Figure S5A–D). We conclude that, although common ancestry associations exist across cancers, the strongest associations are within individual cancer types.

To validate the mRNA results, we compared EAS data from the ICGC PCAWG Japanese liver cancer (LIHC) cohort to TCGA EUR LIHC samples. We selected LIHC because it does not split into subgroups, whereby different subgroup compositions might confound our analyses. Among the 54 genes found to be ancestry-associated within TCGA, we found a significant correlation ($p<0.01$) between beta values calculated within TCGA and using the new PCAWG data and 62% concordance based on directionality (positive or negative) (Figure S5E–F). In contrast, 54 randomly selected genes exhibited only 37% concordance between the datasets. Next, to determine if expression patterns were similar between the two datasets on a per sample basis, we weighted each gene's expression by the absolute value of its beta to account for its ancestry effect size, and then summed across genes in each sample. In both PCAWG and TCGA, the EAS and EUR samples were enriched with genes with positive and negative beta values, respectively, as expected. and the differences were significant (t-test $p<1e-6$ in all cases; Figure S5G–J). We conclude that our ancestry associations largely validated.

Two of the pan-cancer hits, Glutathione S-Transferase Theta 1 (*GSTM1*) and Crystallin Beta B2 (*CRYBB2*), were previously associated with both ancestry and susceptibility to cancer (White et al, 2008; Zhang et al, 2011; Bin et al, 2013; Mo et al, 2009; Faruque et al, 2015). *CRYBB2*, specifically, has high expression among African-American women with Luminal A breast tumors (D'Arcy et al, 2015). These genes exhibited consistent ancestry associations across different tissue types, though with varying magnitude (Figures 4A and 4E).

Detection of ancestry-associated mRNA expression is heavily dependent upon accurately controlling for cancer subtype distributions across ancestries. For example, breast cancer subtypes differ widely in expression profiles (Cancer Genome Atlas Network, 2012) and associate with ancestry (Huo et al, 2017; Troester et al, 2018). As a result, when not controlling for subtype, over 2,000 genes appeared to be associated with AFR ancestry in BRCA (Table S4) and 1,427 genes appeared to be associated with EAS ancestry in esophageal cancers. After controlling for subtype and batch, however, these numbers diminished to 59 and 0, respectively (Figure 4A; Table S4). Our cohort included nine cancer types that have been further subclassified and five that have not (Table S4). On average, prior to subtype correction, we detected 358 ancestry-associated mRNAs in the former group and only 9 in the latter. After subtype correction, however, we detected an average of 9–10 associations in both groups. Many cancer subtypes are defined by differences in methylation, miRNA, and somatic genetic profiles in addition to mRNA (Hoadley et al,

2018), indicating the value of generating comprehensive molecular data to conduct properly controlled comparisons.

miRNA associations with genetic ancestry

An important consideration in miRNA analysis across ancestries is the possibility of artifacts associated with germline variants. In generating miRNA mature strand (miR) expression data, TCGA considered only exact-match read alignments (Chu et al, 2016). If an ancestry is enriched for a variant within a miR, sequence reads for that mature strand will be undercounted. To mitigate this artifact, we ignored 41 miRs that contain ancestry-specific SNPs (Table S5). Among the remaining miRs, 149 miRs exhibited ancestry-differential expression passing a significance threshold of FDR $q<0.001$ in the pan-cancer analysis (Figure 5A,B, Table S5). Fewer miRs exhibited associations within individual cancer types or subtypes, with a maximum of 54 in BRCA. Thus, similar to the methylation and mRNA results, there was sufficient commonality across cancer types that the increased power in the pan-cancer analysis identified larger numbers of associations. However, as with methylation and mRNAs, the associations in the pan-cancer miRNA analysis represented only small differences between ancestries; none had even a two-fold change in expression. Conversely, associations within cancer types often represented greater differences, sometimes over four-fold (Figure 5B).

We therefore performed a separate analysis to focus on miRs with at least two-fold differences in expression between ancestries. Using this stringent effect size threshold in combination with a more relaxed significance threshold (FDR<0.05), we detected 117 associations across all cancer types and subtypes, and none in the pan-cancer analysis. These ancestry associations were distributed across 71 miRs and 22 tumor types and subtypes: 89 associations for AFR, 27 for EAS, 1 for AMR and 0 for SAS (Table S5; the 17 miRs with the largest effect sizes are indicated in Figure S6A). Most miRs were associated with ancestry in only one tumor type. Differences in the numbers of observed associations reflected the number of samples from different ancestry groups in a cancer type, and therefore statistical power.

We next determined the extent to which the miR associations could be explained by differential expression of the genes that host them. Most miRNAs (74%) overlap a ‘host’ gene on the same strand, and tend to be expressed with that gene (Figure S6B, Table S5, STAR Methods: Resources for miRNA annotation). However, none of the 117 significant miR-ancestry associations that reached our effect size threshold corresponded to an mRNA that was similarly differentially expressed between the same ancestries in the same cancer type.

Two contributors to this disparity are that some differentially expressed miRs did not have same-strand host genes, and others had host genes (particularly non-coding transcripts) whose expression was not assessed by TCGA. Among the 71 differentially expressed miRs, 66 (93%) were hosted, compared to only 74% of miRs overall ($p=0.0004$, test of proportions). However, for approximately half of these 66 hosted miRs, expression of the host gene was not assessed by TCGA.

A third reason is that most correlations between hosted miRs and host genes were modest, due to the diverse genomic contexts in which miRNAs occur and to the many factors that can influence correlations between expression of a host gene and its hosted miR(s) (Figures 5C–D, S6B–J). For example, for four TCGA cohorts, few Spearman correlations between hosted miRs and host genes had large positive values (medians 0.31 to 0.36, Figure 5E, see also Figure S6K).

We also found that few miRNA host genes were themselves ancestry-associated. Of 62 AFR-associated mRNAs in four cancer types (BRCA, COAD, HNSC, and UCEC), only *CLN8* was identified as a host gene (for hsa-miR-3674, which was not among the 743 miRs available for analysis). Of the 56 EAS-associated mRNAs in BLCA, BRCA, and STAD, none was a miRNA host gene. As a result, hosted miRs were not enriched among the 71 ancestry-associated miRNAs.

Taken together, we found that ancestry-associated differences with large effect sizes in miRNA expression were largely specific to individual cancer types, as was the case with methylation and mRNA. However, expression correlations between hosted miRs and host genes were generally weak, and few ancestry-associated mRNAs were miRNA host genes. At the same time, approximately 80% of ancestry-associated miRs with large effect sizes were within host genes, which suggests that work to identify (epi)genetic causes of a miR being ancestry-associated will often need to account for different or more subtle effects of those same factors on a host gene.

Relation between ancestry associations and germline genetics

The ancestry associations we observed in DNA methylation and RNA expression raise two questions: *first*, how are these associated with germline genetic differences, and *second*, are these differences associated with cancer? We attempted to address the first question by identifying ancestry-associated expression quantitative trait loci (eQTLs) associated with the mRNA differences we had identified. We attempted to address the second question by determining whether loci encoding ancestry-associated genes are non-randomly distributed between ancestries in cancer.

Ancestry associations in mRNA expression can be due to differences in underlying genetics or in the environments experienced by different ancestry groups. In the former case, we might expect to identify eQTLs. We therefore assessed the extent to which both cis- and trans-eQTLs might account for ancestry-differential mRNA expression. We first identified ancestry-associated genetic polymorphisms by testing for the association between SNP genotype proportions and ancestry using TCGA cohort matched normal samples (n=10,678). Approximately 85% of tested SNPs were associated with ancestry. We then integrated these data with a cancer cis- and trans-eQTLs catalog, PancanQTL (Gong et al, 2018), to find ancestry-associated SNPs that overlap cancer-type specific eQTLs.

Focusing on cancer types with at least 10 minority population samples, we observed varying numbers of cancer type-specific cis-eQTLs associated with ancestry-specific gene expression: from 2,760 in UCEC to 26,089 in BRCA in AFR-EUR comparisons, and from 3,311 in UCEC to 44,640 in THCA in EAS-EUR comparisons (Table S6). We found support

for ancestry-associated genetic variation for 64 to 90% of these cis-eQTLs (Figures 6A–B). We detected fewer such associations among trans-eQTLs, possibly due to the more stringent corrections required for multiple hypotheses (Figures S7A–B, Table S6). Much of the ancestry-associated differences in expression linked to cis- or trans-eQTLs can be explained by differences in genotype frequencies underlying these loci (Figures 6C, S7C). We conclude that germline genetic variation can partially explain differences in mRNA expression across populations in a cancer type-specific manner.

To assess whether the ancestry-associations that we identified in methylation and mRNA expression might be associated with the development of cancer, we evaluated whether one ancestry was enriched at each of the involved genomic loci in TCGA subjects with admixed ancestry. We focused on the AFR-EUR analysis because admixed populations were best represented in our dataset for this comparison. In Figure 1D, we found that no single locus was significantly enriched with AFR or EUR ancestry when considering all loci as independent hypotheses. Here, we considered whether focusing on loci with ancestry-associated differences in methylation or expression might provide greater resolution.

In aggregate, both the 191 loci with pan-cancer ancestry-associated differences in mRNA expression (excluding genes that spanned ancestry blocks and therefore had ambiguous results) and the 176 loci with pan-cancer ancestry-associated differences in methylation were modestly enriched for one ancestry-most often AFR-relative to what would be expected by chance (Wilcoxon $p<0.001$ in both cases; Figures S7D–E). These findings support the hypothesis that pan-cancer differences in methylation and mRNA expression between AFR and EUR ancestries contribute to cancer, with a modest bias towards association between AFR ancestry at these loci with cancer. Evaluation of additional cohorts will be necessary to validate this finding.

Integrated ancestry-associated pathways and cell states

We integrated across our molecular data to answer two questions: *first*, do ancestry-associated differences in methylation account for ancestry-associated differences in mRNA expression, and *second*, do all these molecular features, when taken together, indicate consistent differences in activation of specific molecular pathways? For the first question, we found that, among the 251 genes with ancestry-associated methylation, 27 were also among the 806 genes with annotated CpGs that exhibited ancestry-associated mRNA expression, constituting a strong association ($p<0.001$, OR=7; Figure S7F). This association was strongest for genes with the most differentially methylated CpGs (Figures S7G–H). However, differential methylation could account for only 3.3% of differentially expressed genes.

For the second question, we used PARADIGM (Vaske et al, 2010; Sedgewick et al, 2013) to infer the activation of ~19K pathway features between ancestry groups within tumor types. We observed significant differential features between AFR and EUR groups in eight tumor types (Table S7). Only BRCA, however, provided significant differences in “key” regulatory nodes with at least 10 differential downstream targets, and only three-*ATM*, a known breast cancer susceptibility gene, *SP1*, and *MAPK14*-remained significant in subtype-adjusted analyses (within the Luminal A subtype) (Figures 7A–B). Similarly, significant EAS-EUR

differential features were observed in seven tumor types, but key regulatory nodes were identified only in BLCA, BRCA, and ESCA (Figure 7C). None of these remained significant in subtype-adjusted analyses, suggesting that they reflect subtype enrichments.

We also assessed whether known cancer pathways and driver genes (Knijnenburg et al, 2018; Sanchez-Vega et al, 2018; Bailey et al, 2018) are overrepresented among genes with differential PARADIGM-inferred IPLs from subtype-adjusted analyses, using a hypergeometric test with Benjamini-Hochberg multiple testing corrections. Significantly enriched pathways included DNA repair, HIPPO, RTK-RAS, p53, NRF2 and Notch pathways in the BRCA AFR-EUR comparison and the WNT pathway in the BLCA EAS-EUR comparison (Figure 7D, Table S7). These analyses suggest contributions of ancestry to cancer-related pathway activity.

We also hypothesized that the cellular composition of individual cancers might differ between populations. Indeed, in EAS BLCA were depleted for immune infiltrates as estimated from mRNA expression after controlling for age, gender, subtype, TMB, and aneuploidy (Figure 7E), and higher inferred pathway activities of immune-related features were found in the EUR group (Figure 7A). The mRNA expression of *CD274* that encodes PD-L1 was also significantly lower in EAS relative to EUR and AFR samples (Figure 7F). These results are consistent with a prior orthogonal analysis that found lower lymphocyte infiltrates in EAS samples (Thorsson et al, 2018), and suggest that ancestry should be taken into account when evaluating immunotherapy response.

Discussion

This comprehensive analysis of molecular features associated with ancestry across a range of tumor types has implications for both cancer and normal tissue, given the limited prior analyses available. An analysis of mRNA profiles from several dozen non-neoplastic tissues found that differences between AFR and EUR tended to be shared across tissue types (Mele et al, 2015). We obtained a similar result when considering all statistically significant associations. However, when considering associations whose effect sizes are biologically meaningful, we found that ancestry associations tend to be tissue-specific, regardless of whether those associations reflect rates of somatic alteration, degrees of CpG methylation, or levels of mRNA or miRNA expression

The sources of these ancestry associations are unclear. Our eQTL analyses indicated that germline genetic differences could explain much of the differences in mRNA expression, but varied environmental exposures may also be a major contributor. Although more than two thirds of TCGA donors were from the United States, most EAS donors were likely from other countries, and the collected samples may not represent the entire cancer patient population in any country. TCGA samples were largely collected from academic centers whose patients are often from different socioeconomic strata than the general population. Separating ancestry from the effects of social and environmental factors, and comparing ancestral groups across regions and countries, requires greater study (Gomez et al, 2015).

Although some ancestry-associated differences in methylation and RNA expression are likely to be somatic, we found evidence that many are shared by normal tissue. A robust distinction of cancer-specific ancestral association will require profiling more tissue samples across ancestries, especially normal tissue samples due to the limitations of current data. However, we did observe modest evidence that AFR-EUR differences in methylation and mRNA expression are causally related to cancer, and that these differences were enriched in several cancer-related and immune pathways. These findings may inform cancer prevention and treatment across ancestral groups.

In the process of detecting ancestry associations, we found that uneven distributions of ancestry groups between cancer subtypes was a major confounder. The causes of these ancestry associations with cancer subtypes-possibly disparities in cancer incidence or sample collection biases-are not understood and deserve further exploration, but by controlling for subtypes we aimed to detect ancestry-associated differences within subtypes as well. We were aided by TCGA-derived molecularly-defined subtypes (Sanchez-Vega et al, 2018). This ability to control for subtypes, for example, allowed us to focus on *FBXW7* mutations as associated with ancestry, where prior analyses had instead identified *TP53* and *PIK3CA* mutations (Yuan et al, 2018) that we found to disappear with subtype controls.

A particularly robust finding was enrichment of *VHL* and *PBRM1* mutations and chromosome 3p loss, on which these genes reside, in EUR over AFR KIRC samples. Given the centrality of *VHL* and 3p loss to KIRC, samples without these alterations might represent a different cancer subtype, one which is more prevalent in AFR patients. Many of the *VHL* wild-type cases not only differ transcriptomically from *VHL* mutants, but have very disparate expression profiles themselves (Beroukhim et al, 2009). *VHL* wild-type cases may therefore represent more than one subtype, or their limited transcriptomic similarities may be particularly important. The indication that EAS BLCA samples exhibit less immune infiltration than EUR samples might suggest differing responses to immunotherapies such as Bacillus Calmette–Guérin (BCG) and immune checkpoint inhibition, which form mainstays of BLCA treatment (Marchioni, et al, 2018). Perhaps the greatest analytic obstacle we faced was the small number of non-EUR TCGA samples. Only 17% were at least partially non-EUR, as opposed to ~40% of the U.S. general population (https://www.census.gov/quickfacts/fact/table/US/PST045218) and cancer population (https://seer.cancer.gov/csr/1975_2016/results_merged/topic_race_ethnicity.pdf). Even fewer non-EUR samples would have been included if AFR BRCA collection had not been prioritized (Huo et al, 2017). Additional comprehensively characterized non-EUR cancers would be of great value.

STAR Methods

RESOURCE AVAILABILITY

Lead Contact—Further information and requests for data or code generated in this study should be directed to and will be fulfilled by Lead Contact, Andrew D. Cherniack (achernia@broadinstitute.org).

Materials Availability—This study did not generate new unique reagents.

Data and Code Availability—The raw ancestry assignments and local ancestry calls generated during this study can be found at https://portal.gdc.cancer.gov. Code generated for local ancestry-related analyses is available from the following URL https://github.com/jcarrotzhang/ancestry-from-panel/tree/master/GDAN_AIM.

METHOD DETAILS

Ancestry assignment—We assigned ancestries using five separate approaches, including three (Broad Institute, Washington University, and University of California San Francisco methods) that are based upon SNP 6.0 array genotyping calls (https://portal.gdc.cancer.gov/legacy-archive/) and two (University of Trento and ExAC/Broad methods) based upon exome sequencing data (https://portal.gdc.cancer.gov/). Details on each method are as follows.

Broad Institute - SNP and exome based calls: We merged the genotype files of all TCGA normal samples with reference samples from the 1000 genome project. We excluded TCGA variants with excess missingness of 5% or failed the Hardy Weinberg equilibrium test. EIGENSOFT smartpca version 9102 (Price et al, 2006) was used to remove outliers and 9,040 samples were kept after quality control. We used 90,4099 markers with minor allele frequency greater than 1% in the 1000 genome cohort for global ancestry identification of TCGA samples using smartpca. We defined the AFR, EUR and EAS ancestral groups using the 1000 genome samples based on smartpca generated PC1 and PC2, and AMR and SAS ancestral groups based on PC2 and PC3. We then used ADMIXTURE version 1.23 (Alexander et al, 2009) to estimate the percentage of global ancestry of AFR, EUR, EAS, AMR and SAS (k=5) for each sample. Samples with the proportion of the secondary ancestry greater than 20% were considered as admixed samples.

The Exome based ancestry assignments for 8,066 TCGA patients were provided by the Exome Aggregation Consortium (ExAC). The ExAC dataset ancestry calls were created by using principal component analysis (PCA) of 5400 common exome SNPs to stratify the exomic data into principal components and identify major clusters of continental ancestry (Lek et al, 2016).

Washington University - SNP based calls: Birdseed genotype files were converted to individual VCF files and then merged into a combined VCFs containing all 11,459 samples and 522,606 variants. We conducted PCA as implemented by PLINK 1.9 (Purcell et al, 2007). Specifically, we retained 298,004 variants with MAF > 15% for population structure analysis. The resulting eigen values and eigen vectors were then recorded. PC1 and PC2 accounted for 51.6% and 29.2% of the variations across the first 20 PCs and none of the trailing PCs accounted for more than 3.2%. We then visually examined samples and their self-reported ethnicity based on PC1 and PC2. We defined (1) EUR as samples that self-reported as white and with PC1 < 0.01 and PC2 < 0.02, (2) EAS as samples that self-reported as EAS and with PC1 > 0.01 and PC2 > 0.02, and (3) AFR as samples that self-reported as black or AFR and with PC1 > 0.01 and PC2 < 0.

University of California San Francisco - SNP based calls: Ancestry calls were computed based on partition around medoids (PAM) clustering of principal components (PC's) 1–3 generated from quality controlled genotyping files of 10,128 individuals. PCA without LD pruning was computed in PLINK 1.9 (Purcell et al, 2007), and visual examination of the principal component plots annotated by self-reported race and ethnicity reveal the first 3–4 PCs capture population structure information, while PC 5–6 capture outliers. PCA-based initial ancestry clusters were determined by performing both k-means and PAM clustering on either the first three or first four PCs. We computed gap statistics and average silhouette widths iteratively for number of clusters, k=1 to 10 for k-means and PAM methods respectively to find the optimal number of clusters for each method. Four clusters were found to be optimal based on average silhouette width statistics computed iteratively for number of clusters, k=1 to k=10 (GDC Publication Page Figure S1–A https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). The four PCA-based ancestry clusters show high concordance with the self-reported race/ethnicity of the individuals (GDC Publication Page Figure S1–B,C,D https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). The four ancestry cluster are as follows: (1) PAM ancestry cluster 1 is concordant with EUR ancestry, capturing 97.27% of individuals self-reporting as White, as well as 82.16% of individuals with self-reported non-Hispanic/non-Latino ancestry and 45.96% with self-reported Hispanic/Latino ancestry; (2) ancestry cluster 2 with AFR ancestry, capturing of 97.53% of individuals self-reporting as Black/African-American race; (3) ancestry cluster 3 with EAS ancestry, capturing 90.88% of individuals self-reporting as EAS and 88.89% self-reporting as Native Hawaiian/Pacific Islander; and (4) ancestry cluster 4 with a subgroup of individuals with AMR ancestry capturing 60% of individuals self-reporting as American Indian /Alaska Native and 47.2% with self-reported Hispanic/Latino ethnicity (GDC Publication Page Figure S1–B https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). PC's 1–7 show further population sub-structure in the EAS and EUR ancestry clusters (GDC Publication Page Figure S2 https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). PAM ancestry sub-clusters were computed using PC's 1–7 for individuals within the EAS ancestry cluster which yielded two optimal sub-clusters (GDC Publication Page Figure S2–A https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020), and within the EUR ancestry cluster which yielded three optimal sub-clusters (GDC Publication Page Figure S2–B https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). Of note, 72.46% of EUR sub-cluster 3 self-reports as EAS (15.94% have no race reported). Ancestry clusters, sub-clusters, self-reported race and ethnicity and PC's 1–7 are provided for each individual (GDC Publication UCSF_Ancestry_Calls.csv https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). For individuals represented by more than one sample, blood-derived normal samples were preferentially selected; for those with more than one blood-derived samples, samples with higher call rates were retained leaving 10,128 unique individuals.

University of Trento - whole-exome sequencing based calls: Ancestry analysis was performed by means of EthSEQ (Romanel et al, 2017). First, by combining 1,000 Genome Project and Ashkenazi (Carmi et al, 2014) genotype data, a reference model including 70,415 common (MAF>1%) exonic SNPs and representing 6 main ethnic groups (EUR, ASH, AMR, AFR, SAS, EAS) was built. Then, genotypes of all considered SNPs for 9,666

TCGA individuals were inferred from WES data using ASEQ (Romanel et al, 2015) and a target model for each tumor tissue (N=24) was built. Genotype calls from WES required depth of coverage >=10X and read/base mapping qualities >=20. EthSEQ PCA-based analysis was then performed for each tumor tissue on aggregated target and reference models genotype data considering only SNPs with appropriate overall call rate (99% threshold was applied). Three-dimensional Euclidean space defined by the first three PCA components ("3D" EthSEQ option) was used to first generate the smallest convex sets identifying reference ethnic groups and then to assign an ancestry all TCGA individuals (GDC Publication Page Figure S3a https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). EthSEQ refinement analysis was used to better characterize spatially close AMR and EUR groups (GDC Publication Page Figure S3b https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). Individuals projection inside a reference ethnic group set were annotated with the corresponding ancestry and INSIDE label; individuals outside ethnic group sets were annotated with the nearest (Euclidean distance) ethnic group and CLOSEST label. All reference models were built using 1,000 Genome Project genotype data, including 246,164 common (MAF>1%) exonic SNPs (103,471 and 142,693 even or odd chromosomes, respectively) and representing 5 main ethnic groups (EUR, AMR, AFR, SAS, EAS). Overall concordance of EthSEQ ancestry calls between the three analyses is shown in GDC Publication Page Figure S3a, while fraction of calls preserved using only even/odd chromosomes is shown in GDC Publication Page Figure S4 (https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020).

Consensus ancestry calls for the TCGA cohort: After ancestries were independently determined using these five methods, consensus calls were created based on the ancestral population that received the majority of assignments for each patient. Ancestry assignments are in Table S1.

Ancestry calls for the Foundation Medicine cohort: Comprehensive genomic profiling was performed in a Clinical Laboratory Improvement Amendments (CLIA)-certified, CAP (College of American Pathologists)-accredited laboratory (Foundation Medicine Inc., Cambridge, MA, USA) on de-identified, consented-for-research samples using the FoundationOne test. For each sample, genome-wide ancestry-calling and chromosome-level ancestry-calling were performed. Ancestry callers were trained on 1000 Genomes samples to recognize five ancestral groups (AFR, AMR, EAS, EUR, SAS), using SNPs cataloged by both the 1000 Genomes Project and captured by FoundationOne. These SNPs were projected using principal components analysis, and the top N resulting features were used to train a random forest classifier (with N=5 for genome-wide calling, and N=100 for chromosome-level calling). Samples with > 80% chromosome-level consensus calls that matched the genome-wide call were considered in this study. Next, disease ontology terms in the Foundation Medicine dataset were harmonized with TCGA datasets (Table S2) to ensure proper comparisons. Finally, Fisher's Exact test was used to determine statistically significant differences in alteration rates in different ancestral groups. Admixture analysis was also performed, wherein ADMIXTURE was run on the 1000 Genomes AFR, EAS, and EUR samples with K=3 to learn admixture groups, and then ADMIXTURE was rerun on the

Foundation Medicine data in projection mode using the groups learned on the 1000 Genomes data.

Local ancestry assignment: We performed local ancestry identification on 10,366 samples based on the SNP array genotype data. We used SHAPEIT v2 (Delaneau et al, 2011) to phase the SNPs and then RFMix version 1.5.4 (Maples et al, 2013) to infer local, AFR, EUR or EAS ancestry by chromosomes, using 1,668 AFR, EUR, or EAS samples from the 1000 genome Phase 3v5 reference panel as the reference panel. For each sample, we collapsed nearby SNPs with the same ancestry into regions that were used for association analyses (Martin et al, 2017).

TCGA tumor subtypes: TCGA subtypes for all tumor types except bladder cancer were published by the TCGA Pancancer Atlas (https://gdc.cancer.gov/about-data/publications/pancanatlas, Sanchez-Vega et al, 2018). Bladder cancer (BLCA) mRNA subtypes were obtained from Robertson et al, 2017.

Imputation: Birdseed files were read in R using an in-house tool (courtesy of Donglei Hu), and 905,422 variants were loaded and analyzed in PLINK 1.9 using the SNP Array 6.0 (release 35) annotation file. A total of 861,351 variants passed missingness thresholds of 5% maximum per variant, and 10,917 samples passed missingness thresholds of 5% per sample. Hardy-Weinberg equilibrium (HWE) was calculated in PLINK for autosomal chromosome variants in the largest UCSF ancestry cluster (European ancestry cluster 1), and SNPs out of HWE (<10^−6) were flagged. Flagged SNPs were cross-referenced against cancer risk SNPs, and non-risk SNPs with HWE <10^−6 were removed. Minor allele frequency (MAF) was then calculated and variants with MAF <0.5% were excluded. Duplicate SNPs with identical genomic first positions were removed. A total of 838,948 autosomal chromosome variants for 10,128 individuals passed QC (clean file). PCA was performed on the clean genotyping file and final PAM-ancestry clusters were computed for the 10,128 individuals for optimal k=4 (see UCSF ancestry calls). We found very high concordance of initial and final ancestry assignments (99.98% matching, the 2 samples varying between initial and final ancestry cluster computation assigned to NA).

The cleaned genotyping file was then stranded and imputed against two reference panels: Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/) and 1000 Genomes (http://www.internationalgenome.org/). For Haplotype Reference Consortium, all palindromic SNPs were removed and stranding was done using the McCarthy Group tools (HRC-1000G-check-bim-v4.29), which compares genotyping alleles to reference SNP list from Haplotype Reference Consortium (v1.1 HRC.r1–1.GRCh37.wgs.mac5.sites.tab) leaving 680,389 correctly matched variants for imputation. For 1000 Genomes, all palindromic SNPs were removed and stranding was done using an in-house tool (courtesy of Scott Huntsman), which compares genotyping alleles to stated alleles from 1000 Genomes (Phase 3v5) legend files leaving 678,304 correctly matched variants for imputation.

Imputation and phasing were performed using a standard pipeline on the Michigan Imputation Server (https://imputationserver.sph.umich.edu). The process of phasing involved

running on the WGS variant call file (VCF). To reduce the run time, the VCF file was split-up into 22 files corresponding to individual autosomes. By default, Eagle (Loh et al, 2016) restricts analysis to bi-allelic variants that exist in both the target and reference. Minimac 3 (Howie et al, 2012) was used to run the imputation. For Haplotype Reference Consortium, the HRC r1.1.2016 reference panel was selected using mixed population for QC, with a total of 39,127,678 SNPs were returned after imputation. For 1000G, the 1000G Phase 3v5 reference panel was selected using mixed population, with a total of 43,826,430 SNPs and 3,233,367 INDELs were obtained (1000G imputation courtesy of Younes Mokrab).

QUANTIFICATION AND STATISTICAL ANALYSIS

Ancestry and molecular features—A multivariate regression model was generated for each data type to determine the effects of ancestry on the molecular features while controlling for potential confounders such as cancer types, subtypes, age and gender. For each data type, two main regression tests were performed: *first*, a pan-cancer analysis using all cancer types, and *second*, cancer type-specific analyses. For each analysis, p values were corrected for multiple hypotheses using the Benjamini-hochberg procedure (Benjamini and Hochberg, 1995).

Somatic alteration: Pan-cancer mutation and copy number data were used (https://gdc.cancer.gov/about-data/publications/pancanatlas). ICGC PCAWG data was obtained from https://dcc.icgc.org/pcawg and MSK-IMPACT data was obtained from https://www.cbioportal.org/. Mutational signature data were obtained from Alexandrov et. al. (2018). We used signatures 1 to 45 in our analysis. Arm-level SCNA calls, TMB, genome doubling, immune infiltration score represented by leukocyte fraction, and aneuploidy scores were downloaded from previously published work (Taylor et al, 2018).

For the pan-cancer analyses, we counted somatic SNVs, indels and focal CNAs for all significantly mutated genes in all cancer types. For cancer-specific analyses, we counted somatic SNVs, indels and focal CNAs for significantly mutated genes in that cancer type. Focal CNA was defined by a $\log_2$ copy number ratio > 1 or < −1. Multivariate logistic regression was applied to test the association of somatic alteration for each gene with ancestry, while coding AFR or EAS ancestry as 1, and EUR ancestry as 0, with controlling for age, gender, and cancer type and subtype when applicable. Genes with FDR adjusted q values < 0.1 were considered as candidates for validation. In the Foundation Medicine cohort, gene alterations (short variants, copy alterations, and rearrangements) were detected and alteration status was related to genetic admixture proportions using binomial logistic regression. For admixture validation of *FBXW7*, *VHL* and *PBRM1*, we collapsed SNPs in the locus of each gene into blocks, and correlated the local ancestry of the blocks with the somatic mutation status in the gene, using logistic regression controlling for the global ancestry of individuals (somatic alteration ~ local ancestry + percentage of EUR ancestry + percentage of AFR ancestry).

To test if ancestry is associated with arm-level SCNA, multivariate logistic regression was used with controlling for age, gender, and cancer type and subtype when applicable, as well as aneuploidy and genome doubling (ancestry ~ arm-level SCNA + aneuploidy + genome

doubling + age + gender + subtype). Similarly, to test if ancestry is associated with genomic/molecular features including TMB, aneuploidy and IMS, regression was performed as ancestry ~ aneuploidy + TMB + IMS + age + gender + subtype.

DNA methylation: IDAT files for Infinium HumanMethylation450 (HM450) arrays were downloaded from GDC (https://portal.gdc.cancer.gov/legacy-archive/) and preprocessed using the openSeSAMe pipeline (Zhou et al, 2018b). We did not explicitly mask for design issues, to enable associations between SNP artifacts and ancestry-differential probes. Samples with mixed ancestry background or undetermined purity estimate were excluded, leaving 6,264 tumor samples. The HM450 array included 485,577 probes, 65 of which were nested SNP probes ('rs' probes) that reflect the sample genotype rather than DNA methylation. Whole-Genome Bisulfite Sequencing (WGBS) data for sorted blood cells from 149 nonmalignant samples were downloaded from the BLUEPRINT project (Schuyler et al, 2016) for orthogonal validation of HM450 array results. The 49 TCGA WGBS dataset (Zhou et al, 2018a) was downloaded from Genomic Data Commons (https://portal.gdc.cancer.gov/). Four additional validation HM450 data sets were downloaded from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/, GSE36369; GSE41826; GSE53816; GSE101431). These include two data sets that studied DNA methylation in individuals of AFR, EUR and EAS ancestry, similar to the TCGA cohort: one of whole blood (Heyn et al, unpublished, GSE36369) and the other of brain tissue (Guintivano et al, 2013, GSE41826). In order to assess the validity of our ancestry-differential methylation calls in ancestries not well covered in TCGA, we also studied a data set (Teh et al, 2014, GSE53816) that assayed umbilical cord blood in three EAS populations, and another data set (Carja et al, 2017, GSE101431) obtained from lymphoblastoid cell lines from five human populations of more specific ancestry origins such as Mozabites and Cambodians.

We performed a multivariate linear regression to identify probes that were differentially methylated between ancestry groups. The model explicitly adjusts for tumor type and subtype (Sanchez-Vega, 2018), age, gender, and tumor purity. We performed model fitting on both the entire TCGA cohort (pan-cancer analysis) as well as within each individual cancer type, with and without subtype as a covariate. From the pan-cancer analysis, cancer subtypes known for global methylation alterations such as TGCT with hypomethylation and IDH mutant with hypermethylation are seen with a higher number of corresponding cancer-type and subtype differential probes, validating our model fitting (Figure S3H). The significance of ancestry differences was measured by an F-test contrasting a full model with a model without adjusting for ancestry differences (Table S3). After regression, the mean slope coefficient of methylation from all ancestry groups was re-centered to zero for each probe. The effect size of a probe is calculated by the range of the slope coefficients from different ancestry groups. Probes with effect size greater than or equal to 0.1 and adjusted p value less than 0.05 are considered ancestry-differentially methylated. We also tested alternative regression models including ones that depend on the iCluster assignment (instead of tumor types and subtypes) and self-reported ancestry (instead of consensus inferred ancestry). The top ancestry-differentially methylated genes were robust across alternative regression models (Table S3).

The 374 ancestry-differentially methylated sites that were not identified with design artifacts were projected and displayed using Uniform Manifold Approximation and Projection (McInnes et al, 2018) based on their ancestry methylation bias. A probe was considered to be associated with a gene when the interrogated CpG was located from 1500bp upstream of the TSS until the TTS of any isoform of the gene. Some probes may be associated with more than one gene, but only the one gene symbol is shown in the plot for brevity. Alternative gene names are shown in Table S3.

To estimate the power of detecting DNA methylation differences, we modeled DNA methylation measurements as a binomial distribution with the number of experiments N equal to 40 and the mean equal to true methylation, with assigned ancestry-specific methylation differences. N was set at 40 to model the average bead number for each probe in the Infinium DNA methylation microarray. To imitate the distribution of gender-specific DNA methylation, we introduced a gender difference modeled by a beta distribution parameterized by a=1 and b=5 for the two shape parameters. We also introduced subtype-specific DNA methylation modeled using a normal distribution with a mean of 0.1 and standard deviation of 0.2 with caps at 1 and 0 to keep values between 0 and 1. For each given methylation difference, we performed 1000 simulations and the same regression analysis we performed on the real data. We then computed the fraction of significant ancestry-specific differences and plotted against known methylation differences.

mRNA expression: Pan-cancer mRNA normalized data (https://gdc.cancer.gov/about-data/publications/pancanatlas) was filtered to retain samples with an admixture of <20% for EUR, EAS, and AFR ancestry. The samples were then split to compare EAS (n=532) vs EUR (n=5,901) and AFR (n=380) vs EUR (n=5,901). The dataset was $\log_2$ transformed and filtered for genes present in >80% of samples. The filtered genes (n=16,269) were determined to be significantly associated with EAS vs EUR (FDR q<0.001) and AFR vs EUR (FDR q<0.001) by linear regression correcting for TCGA plate ID and tumor specific subtype. The same linear model was applied both across all cancer types and on a per cancer type basis for tumor types with more than ten samples of the minor ancestry, corrected for tumor subtype where appropriate. For *GSTM1/CRYBB2* and *PPIL3/FBLL1*, the median expression per tumor type and ancestry was plotted to highlight within-ancestry variance.

miRNA: We obtained the TCGA miRNA data prepared for the TCGA Pan-Cancer Atlas (https://gdc.cancer.gov/about-data/publications/pancanatlas). This dataset includes expression levels of 743 miRNA mature strands (miRs) for 10,824 TCGA samples, which we batch-corrected to enable pan-cancer analyses (pancanMiRs_EBadjOnProtocolPlatformWithoutRepsWithUnCorrectMiRs_08_04_16.csv). The consolidated dataset included 8,180 samples across 32 tumor types for which both miRNA expression data and ancestry calls were available (Table S6). No miRNA data were available for GBM. Twelve of the 32 tumor types had subtype annotation.

Before statistical tests, negative miRNA reads per million mapped reads (RPM) values (introduced due to the batch correction procedure) were set to zero, and miRNA RPM values were then log transformed using $y = \log_2(x+1)$, where x values are the RPM values and y values are the log-transformed values used for statistical analyses and visualization.

To determine ancestry associations, we applied a linear regression model with a binary design matrix based on the subtype calls as predictors to explain the normalized expression across the samples of each tumor type and subtype. We then performed Wilcoxon rank sum tests against ancestry calls on the output of this model. We applied the following pre-filtering criteria:

1. The sample size of both groups to be tested should be 5 or larger.
2. The coefficient-of-variation (CoV) across the expression levels of the union of samples of both groups was 0.1 or larger.
3. There were at least 5 samples among the union of samples of both groups with an RPM value of 25 or higher (4.7 in logarithmic space).
4. miRNAs that were flagged as having ancestry-specific SNPs are discarded (n=3). See below for more details about the ancestry-specific SNPs.

Since in TCGA miRNA-seq data only exact-match reads (to the hg19 reference genome) were counted towards expression (Chu et al, 2016), TCGA samples with SNPs in miRNAs will report artificially low (or zero) expression levels for these miRNAs. Ancestry-specific SNPs in miRNAs will thus lead to spurious relationships of differential miRNA expression between ancestry groups. We therefore discarded miRNAs with ancestry-specific SNPs. We merged miRNA annotation from miRBase (http://www.mirbase.org/, v21 released June 2014) with 1000 Genomes Phase 3 information from Ensembl, which contains ancestry-specific SNP allele frequencies, and called SNPs “ancestry-specific” if the difference between the maximum and minimum AAFs of the SNP among the five superpopulations (AFR, AMR, EAS, EUR, SAS) was 0.25 or larger. Table S6 lists the SNPs in miRNAs along with their ancestry-specific allele frequencies. Process details are also provided in our GDC publication page (https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020).

To identify miRNA mature strands (miRs) that were overlapped by and on the same strand as ‘host’ genes, we compiled a general miR-gene resource from GRCh38/hg38 GFF3 files for Ensembl v94 genes and miRBase v22.1 miRNAs, using the rtracklayer v1.42.2 R package (Lawrence et al, 2009), in R v3.5.3. In using GRCh38 annotations, we recognized that the GRCh37/hg19 TCGA RSEM gene expression data would be unavailable for some Ensembl genes, but we prioritized using current miRNA annotations and (largely) current gene annotations and biotypes (https://gdc.cancer.gov/about-data/publications/CCG-AIM-2020). To report Spearman correlations between hosted miRs and host genes, we used Pan-cancer, batch-corrected, normalized GRCh37 expression data for 20,531 RSEM genes and 743 expressed mature strands (https://gdc.cancer.gov/about-data/publications/pancanatlas). Given our GRCh38-based annotation/overlap resource, these data supported calculating correlations for 203 host genes and 331 hosted miRs.

Ancestry associated SNPs and eQTLs—Association analyses were carried out using the Hail framework (https://github.com/hail-is/hail). We transformed the post-imputation vcf files from the Michigan imputation server into Hail matrix format to speed up downstream analysis. Using the Haplotype Reference Consortium imputation calls, we performed further quality control analyses to filter out low-imputation confidence variants. Common variants

(MAF>0.05) that had an imputation confidence score (r^2>=0.5) and did not violate Hardy-Weinberg equilibrium assumption (Hail HWE test p>0.05) were kept for association analysis. We also filtered out samples with call rates<0.95 and admixed samples, leaving 8,696 samples and 8,551,986 SNPs for testing.

To identify SNPs associations with ancestry, we performed logistic regression analysis for EUR versus AFR and EUR versus EAS (Hail Wald's test implementation), including patient age and gender as covariates. For the set of genes with pan-cancer differential mRNA expression associated with ancestry, we extracted pancanQTL cis- and trans-eQTLs from the pancanQTL resource (Gong et al, 2018), and determined the number of eQTL pairs for each cancer-type in which the underlying SNP also showed a significant association with ancestry (Wald's test FDR q<0.05) by merging tables across dbSNP identifiers.

Ancestry and pathways—The PARADIGM algorithm was used to integrate platform-corrected expression, gene-level copy number, and pathway interaction data for 9829 TCGA Pan-Cancer samples to infer the activities of ~19K pathway features (Vaske et al, 2010; Sedgewick et al, 2013). The inferred activities, termed integrated pathway levels (IPLs), reflect the log likelihood of the probability that a given feature is activated (vs. inactivated). Only samples with admixture proportions ≤20% were included, yielding 9046 evaluable samples.

Among the 33 tumor types, the 24 with ≥5 patients not of EUR ancestry were considered. Within each tumor type, we identified pathway features with differential inferred activities between each ancestry group with ≥5 patients and the EUR group using t-tests and Wilcoxon Rank sum tests. Three initial minimum variation filters were applied prior to statistical testing: *first*, at least 1 sample with absolute activity > 0.05; *second*, at least 10% of samples have non-zero activity; and *third*, standard deviation of activity > 0.05. Features deemed significant (q<0.05) by both tests and showing an absolute difference in group means >0.05 were selected. The selected pathway features were assessed for interconnectivity; and regulatory nodes with differential IPLs that also had at least 10 differential downstream regulatory targets were identified. We also evaluated whether DNA repair genes (Knijnenburg et al, 2018) and known cancer pathway or driver genes (Sanchez-Vega et al, Cell 2018 and Bailey et al, Cell 2018) were enriched among the selected differential features by comparing the proportion of pathway genes selected against the proportion of total genes selected, using a hypergeometric test with a Benjamini-Hochberg correction. Pathway gene sets were considered significantly enriched if there were at least two members that were differential at q<0.05.

Among the 24 included tumor types, 10 have subtype annotation; the above described analyses were also performed within tumor subtypes. In addition, for each tumor type, we conducted a subtype-adjusted analysis by first fitting a linear model of each IPL as a function of a binary matrix of subtype membership. The resulting residuals were then compared to identify 'subtype-adjusted' differential features and key regulatory nodes as described above.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank 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 Banting fellowship. We are grateful for advice from numerous colleagues, TCGA and GDAN collaborators, and the GDC technical support team.

References

Alexander DH, Novembre J, and Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19, 1655–1664. [PubMed: 19648217]

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. (2019). The Repertoire of Mutational Signatures in Human Cancer. Nature 578, 94–101.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173, 371–385 e318. [PubMed: 29625053]

Benjamini Y, and Hochberg Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Statist Soc B 57, 289–300.

Beroukhim R, Brunet JP, Di Napoli A, Mertz KD, Seeley A, Pires MM, Linhart D, Worrell RA, Moch H, Rubin MA, et al. (2009). Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. Cancer Res 69, 4674–4681. [PubMed: 19470766]

Bin Q, and Luo J (2013). Role of polymorphisms of GSTM1, GSTT1 and GSTP1 Ile105Val in Hodgkin and non-Hodgkin lymphoma risk: a Human Genome Epidemiology (HuGE) review. Leuk Lymphoma 54, 14–20. [PubMed: 22734843]

Cancer Genome Atlas N (2012). Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70. [PubMed: 23000897]

Carja O, MacIsaac JL, Mah SM, Henn BM, Kobor MS, Feldman MW, and Fraser HB (2017). Worldwide patterns of human epigenetic variation. Nat Ecol Evol 1, 1577–1583. [PubMed: 29185505]

Carmi S, Hui KY, Kochav E, Liu X, Xue J, Grady F, Guha S, Upadhyay K, Ben-Avraham D, Mukherjee S, et al. (2014). Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. Nat Commun 5, 4835. [PubMed: 25203624]

Chang J, Tan W, Ling Z, Xi R, Shao M, Chen M, Luo Y, Zhao Y, Liu Y, Huang X, et al. (2017). Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. Nat Commun 8, 15290. [PubMed: 28548104]

Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, Ma Y, Jones S, and Marra MA (2016). Large-scale profiling of microRNAs for The Cancer Genome Atlas. Nucleic Acids Res 44, e3. [PubMed: 26271990]

D'Arcy M, Fleming J, Robinson WR, Kirk EL, Perou CM, and Troester MA (2015). Race-associated biological differences among Luminal A breast tumors. Breast Cancer Res Treat 152, 437–448. [PubMed: 26109344]

Delaneau O, Marchini J, and Zagury JF (2011). A linear complexity phasing method for thousands of genomes. Nat Methods 9, 179–181. [PubMed: 22138821]

Deng J, Chen H, Zhou D, Zhang J, Chen Y, Liu Q, Ai D, Zhu H, Chu L, Ren W, et al. (2017). Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. Nat Commun 8, 1533. [PubMed: 29142225]

Faruque MU, Paul R, Ricks-Santi L, Jingwi EY, Ahaghotu CA, and Dunston GM (2015). Analyzing the Association of Polymorphisms in the CRYBB2 Gene with Prostate Cancer Risk in African Americans. Anticancer Res 35, 2565–2570. [PubMed: 25964531]

Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, Penney K, Steen RG, Ardlie K, John EM, et al. (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 103, 14068–14073. [PubMed: 16945910]

Gomez SL, Shariff-Marco S, DeRouen M, Keegan TH, Yen IH, Mujahid M, Satariano WA, and Glaser SL (2015). The impact of neighborhood social and built environment factors across the cancer continuum: Current research, methodological considerations, and future directions. Cancer 121, 2314–2330. [PubMed: 25847484]

Gong J, Mei S, Liu C, Xiang Y, Ye Y, Zhang Z, Feng J, Liu R, Diao L, Guo AY, et al. (2018). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res 46, D971–D976. [PubMed: 29036324]

Guintivano J, Aryee MJ, and Kaminsky ZA (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics 8, 290–302. [PubMed: 23426267]

Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L, and Esteller M (2013). DNA methylation contributes to natural human variation. Genome Res 23, 1363–1372. [PubMed: 23908385]

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 173, 291–304 e296. [PubMed: 29625048]

Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, Yoshimatsu TF, Pitt JJ, Hoadley KA, Troester M, et al. (2017). Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol 3, 1654–1662. [PubMed: 28472234]

Hutter C, and Zenklusen JC (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 173, 283–285. [PubMed: 29625045]

Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, et al. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep 23, 239–254.e236. [PubMed: 29617664]

Kraus WE, Muoio DM, Stevens R, Craig D, Bain JR, Grass E, Haynes C, Kwee L, Qin X, Slentz DH, et al. (2015). Metabolomic Quantitative Trait Loci (mQTL) Mapping Implicates the Ubiquitin Proteasome System in Cardiovascular Disease Pathogenesis. PLoS Genet 11, e1005553. [PubMed: 26540294]

Krishnan B, Rose TL, Kardos J, Milowsky MI, and Kim WY (2016). Intrinsic Genomic Differences Between African American and White Patients With Clear Cell Renal Cell Carcinoma. JAMA Oncol 2, 664–667. [PubMed: 27010573]

Lawrence M, Gentleman R, and Carey V (2009) rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25, 1841–1842. [PubMed: 19468054]

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291. [PubMed: 27535533]

Maples BK, Gravel S, Kenny EE, and Bustamante CD (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet 93, 278–288. [PubMed: 23910464]

Marchioni M, Nazzani S, Preisser F, Bandini M, and Karakiewicz PI (2018). Therapeutic strategies for organ-confined and non-organ-confined bladder cancer after radical cystectomy. Expert Rev Anticancer Ther 18, 377–387. [PubMed: 29429376]

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, and Kenny EE (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 100, 635–649. [PubMed: 28366442]

McInnes L, Healy J, and Melville J (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. In arXiv e-prints.

Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. (2015). Human genomics. The human transcriptome across tissues and individuals. Science 348, 660–665. [PubMed: 25954002]

Mo Z, Gao Y, Cao Y, Gao F, and Jian L (2009). An updating meta-analysis of the GSTM1, GSTT1, and GSTP1 polymorphisms and prostate cancer: a HuGE review. Prostate 69, 662–688. [PubMed: 19143011]

Nassar AH, Umeton R, Kim J, Lundgren K, Harshman L, Van Allen EM, Preston M, Dong F, Bellmunt J, Mouw KW, et al. (2019). Mutational Analysis of 472 Urothelial Carcinoma Across Grades and Anatomic Sites. Clin Cancer Res 25, 2458–2470. [PubMed: 30593515]

Pena-Llopis S, Vega-Rubin-de-Celis S, Liao A, Leng N, Pavia-Jimenez A, Wang S, Yamasaki T, Zhrebker L, Sivanand S, Spence P, et al. (2012). BAP1 loss defines a new class of renal cell carcinoma. Nat Genet 44, 751–759. [PubMed: 22683710]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–909. [PubMed: 16862161]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and Sham PC (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559–575. [PubMed: 17701901]

Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, Hinoue T, Laird PW, Hoadley KA, Akbani R, et al. (2017). Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. Cell 171, 540–556 e525. [PubMed: 28988769]

Romanel A, Lago S, Prandi D, Sboner A, and Demichelis F (2015). ASEQ: fast allele-specific studies from next-generation sequencing data. BMC Med Genomics 8, 9. [PubMed: 25889339]

Romanel A, Zhang T, Elemento O, and Demichelis F (2017). EthSEQ: ethnicity annotation from whole exome sequencing data. Bioinformatics 33, 2402–2404. [PubMed: 28369222]

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, and Feldman MW (2002). Genetic structure of human populations. Science 298, 2381–2385. [PubMed: 12493913]

Sanchez-Mut JV, Heyn H, Silva BA, Dixsaut L, Garcia-Esparcia P, Vidal E, Sayols S, Glauser L, Monteagudo-Sanchez A, Perez-Tur J, et al. (2018). PM20D1 is a quantitative trait locus associated with Alzheimer's disease. Nat Med 24, 598–603. [PubMed: 29736028]

Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell 173, 321–337 e310. [PubMed: 29625050]

Schuyler RP, Merkel A, Raineri E, Altucci L, Vellenga E, Martens JHA, Pourfarzad F, Kuijpers TW, Burden F, Farrow S, et al. (2016). Distinct Trends of DNA Methylation Patterning in the Innate and Adaptive Immune Systems. Cell Rep 17, 2101–2111. [PubMed: 27851971]

Sedgewick AJ, Benz SC, Rabizadeh S, Soon-Shiong P, and Vaske CJ (2013). Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM. Bioinformatics 29, i62–70. [PubMed: 23813010]

Shigematsu H, Lin L, Takahashi T, Nomura M, Suzuki M, Wistuba II, Fong KM, Lee H, Toyooka S, Shimizu N, et al. (2005). Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. J Natl Cancer Inst 97, 339–346. [PubMed: 15741570]

Stunnenberg HG, International Human Epigenome C, and Hirst M (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell 167, 1145–1149. [PubMed: 27863232]

Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, et al. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell 33, 676–689 e673. [PubMed: 29622463]

Teh AL, Pan H, Chen L, Ong ML, Dogra S, Wong J, MacIsaac JL, Mah SM, McEwen LM, Saw SM, et al. (2014). The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res 24, 1064–1074. [PubMed: 24709820]

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, et al. (2018). The Immune Landscape of Cancer. Immunity 48, 812–830 e814. [PubMed: 29628290]

Troester MA, Sun X, Allott EH, Geradts J, Cohen SM, Tse CK, Kirk EL, Thorne LB, Mathews M, Li Y, et al. (2018). Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. J Natl Cancer Inst 110.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, and Stuart JM (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, i237–245. [PubMed: 20529912]

White DL, Li D, Nurgalieva Z, and El-Serag HB (2008). Genetic variants of glutathione S-transferase as possible risk factors for hepatocellular carcinoma: a HuGE systematic review and meta-analysis. Am J Epidemiol 167, 377–389. [PubMed: 18065725]

Wu S, Ou T, Xing N, Lu J, Wan S, Wang C, Zhang X, Yang F, Huang Y, and Cai Z (2019). Whole-genome sequencing identifies ADGRG6 enhancer mutations and FRS2 duplications as angiogenesis-related drivers in bladder cancer. Nat Commun 10, 720. [PubMed: 30755618]

Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D, Yang W, Neale G, Cox NJ, Scheet P, et al. (2011). Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nat Genet 43, 237–241. [PubMed: 21297632]

Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, Hu X, Zhang Y, Wang Y, Jiang J, et al. (2018). Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. Cancer Cell 34, 549–560 e549. [PubMed: 30300578]

Zhang ZJ, Hao K, Shi R, Zhao G, Jiang GX, Song Y, Xu X, and Ma J (2011). Glutathione S-transferase M1 (GSTM1) and glutathione S-transferase T1 (GSTT1) null polymorphisms, smoking, and their interaction in oral cancer: a HuGE review and meta-analysis. Am J Epidemiol 173, 847–857. [PubMed: 21436184]

Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, Laird PW, and Berman BP (2018). DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet 50, 591–602. [PubMed: 29610480]

Zhou W, Laird PW, and Shen H (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res 45, e22. [PubMed: 27924034]

Zhou W, Triche TJ Jr., Laird PW, and Shen H (2018). SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res 46, e123. [PubMed: 30085201]

Significance

We conducted a comprehensive analysis of the molecular effects of ancestry across cancer or normal tissues. We found that, though many ancestry effects were shared by normal tissues, they were profoundly tissue-specific, suggesting ancestry effects have to be considered primarily on a per-tissue basis both among cancers and non-cancer tissues. In tissue-specific analyses of normal tissue especially, more samples from diverse ancestries are required for comprehensive ancestry analyses, and we identified important controls for confounders and artifacts that need to be applied in such studies. Differences between African, European, and East Asian groups in renal and bladder cancers in particular suggest that ancestry should be taken into account when considering routes to disease and response to immunotherapies.

Highlights

- This large analysis identified ancestry correlates in cancer
- Ancestry-associated artifacts and confounders were identified
- Ancestry effects are profoundly tissue-specific
- Rates of *FBXW7*, *VHL*, and *PBRM1* mutations and immune activity vary by ancestry

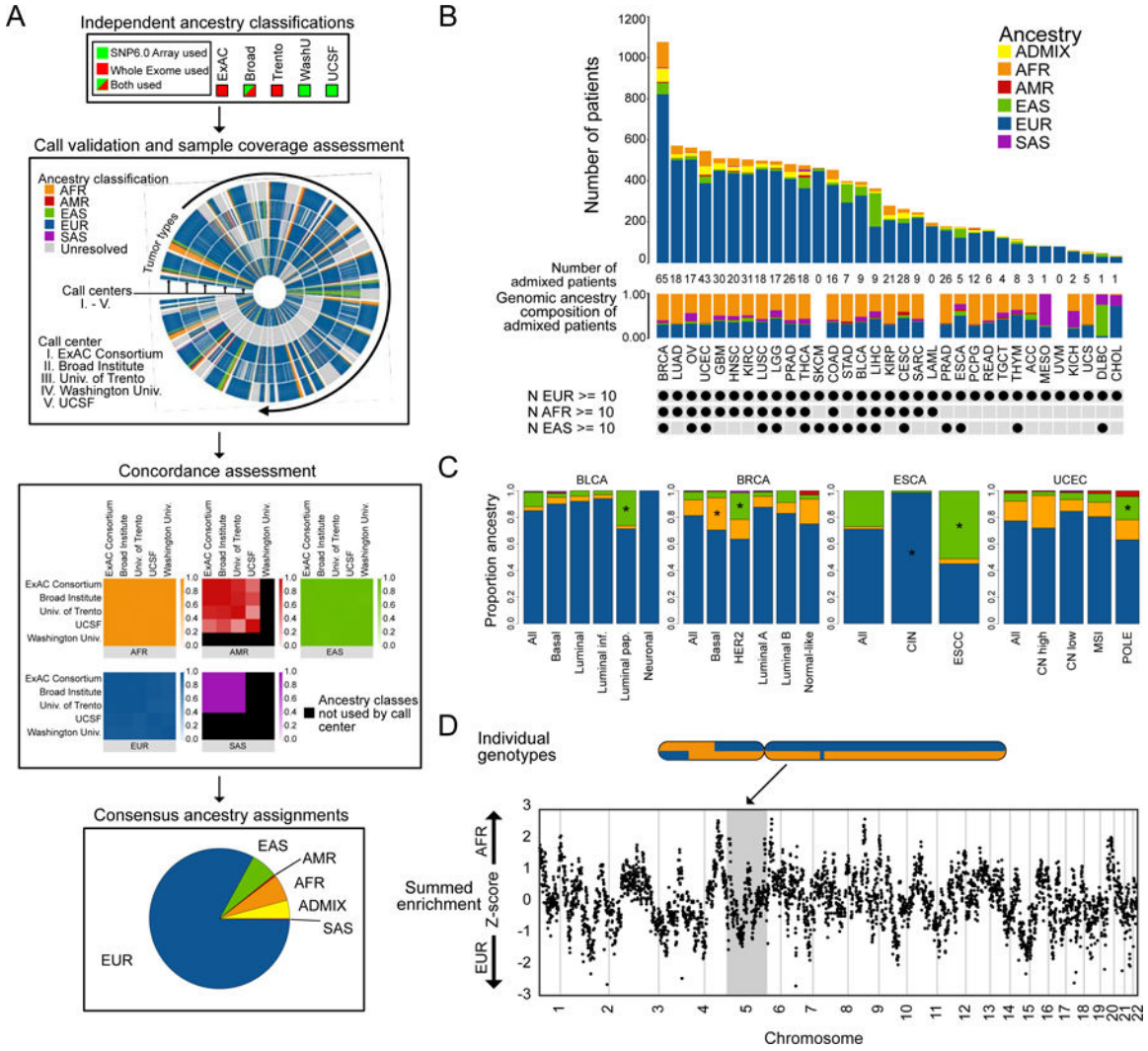


Figure 1. TCGA donor ancestries

(A) Ancestries were called as the consensus between five independent methods based upon SNP array and/or whole exome sequencing. (B) Ancestry representation in each disease type (upper plot), aggregate fractions of each ancestry among admixed individuals (middle panel), and cancer types with at least 10 individuals of the indicated ancestries (black dots; lower panel). (C) Ancestry representation across tumor subtypes with non-random ancestry distributions. (D) Example local ancestry calls (top) and summary enrichment scores for AFR or EUR ancestry (vertical axis), plotted against genomic location (horizontal axis). See also Figure S1 and Table S1.

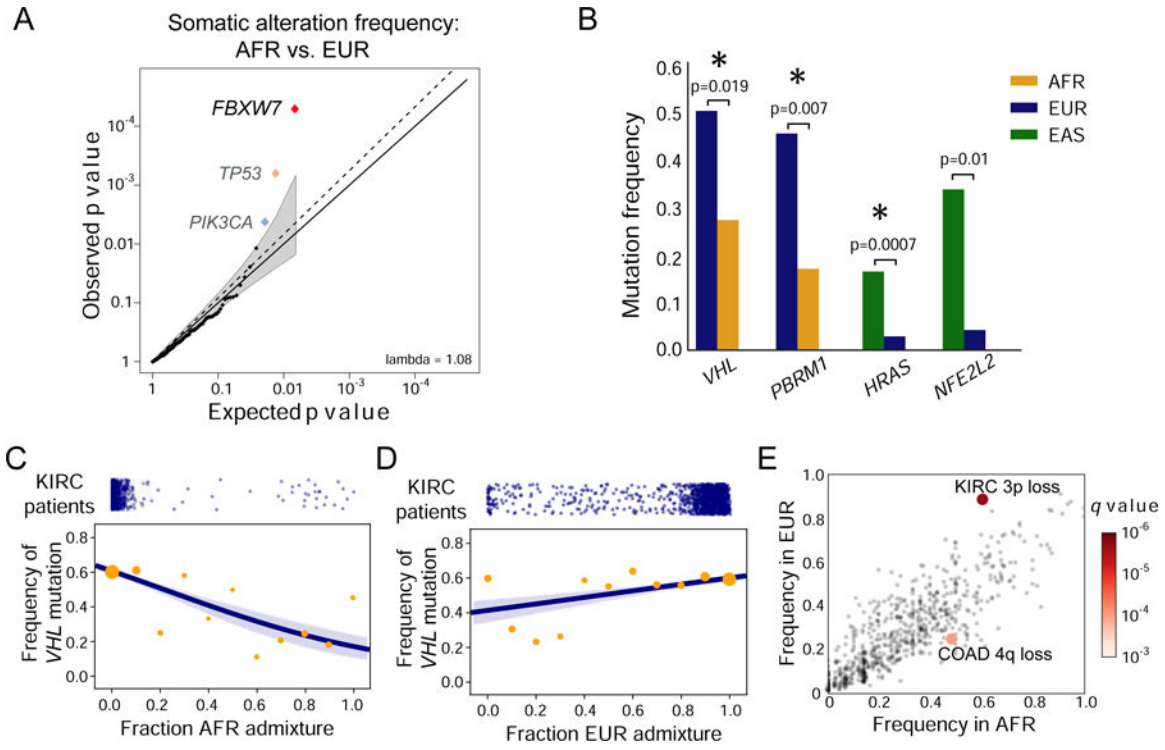


Figure 2. Ancestry-associated somatic genetic alterations.

(A) QQ plot showing genes whose pan-cancer mutation rates were significantly associated with AFR vs EUR ancestry after controlling cancer type but not subtype. Red and blue respectively indicate higher and lower frequencies in AFR (FDR q<0.1). Cancer subtype adjustment removed *TP53* and *PIK3CA* associations, shown in gray; *FBXW7* retained significance. (B) Cancer-specific mutation frequencies in EUR and either AFR (*VHL* and *PBRM1*; KIRC) or EAS (*HRAS* and *NFE2L2*; BLCA and ESCA respectively) TCGA cohorts. p values represent analyses controlled for cancer subtype. Stars represent genes that validated in external cohorts. (C-D) *VHL* mutation frequency (vertical axis) plotted against level of admixture (horizontal axis) of (C) AFR and (D) EUR ancestry in KIRC FMI patients. Individual patient admixture levels are indicated by the blue dots at the top of each panel. Yellow dots represent frequencies at each decile of admixture; dot sizes correspond to the patient numbers in each decile. Blue profiles and shadows represent binomial logistic regression ($p < 0.001$ for *VHL* in AFR and EUR) and confidence intervals, respectively. (E) Arm-level SCNA frequencies in EUR (vertical axis) and AFR (horizontal axis) cohorts, across all diseases and chromosome arms. Chromosomes 3p and 4q had significantly different rates of loss among KIRC and COAD patients respectively.

See also Figure S2 and Table S2.

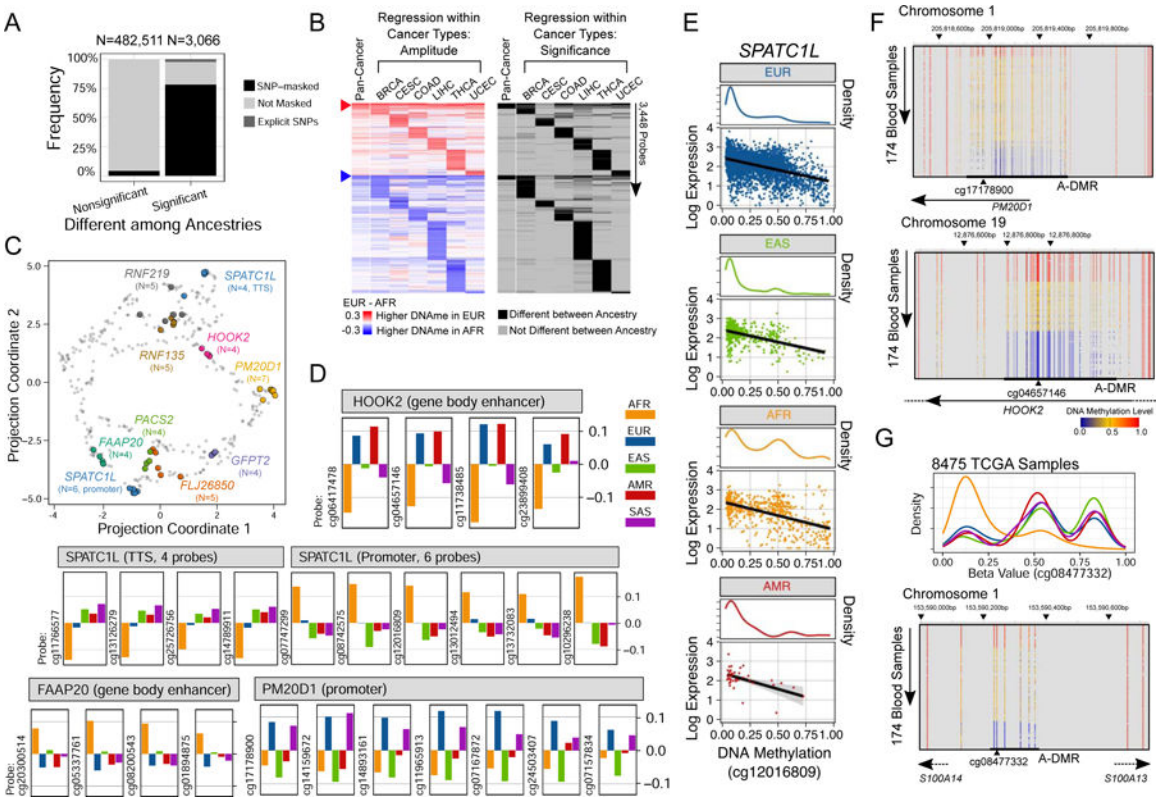


Figure 3. Ancestry-differential DNA methylation

(A) Number of positive control 65 rs probes ('Explicit SNPs'), probes with measurements directly influenced by SNPs ('SNP masked', excluded from later analyses), and all other probes ('Not Masked'), among probes found to be significant or non-significant in ancestry testing. (B) Left: Regression coefficients between AFR and EUR samples, pan-cancer and in six cancer types. Right: the statistical significance of these differences. (C) Concordant ancestry bias across probes (dots) for the same genes. Genes with at least four ancestry-differential probes are colored. (D) Ancestry bias (vertical axis), computed as the slope (beta) in the regression model, across ancestries in example genes. (E) Methylation at the *SPATC1L* promoter (cg12016809 beta value, horizontal axis) is associated with reduced gene expression (vertical axis). Beta value distributions are shown as smoothed density plots above scatter plots. (F) Ancestry-associated differentially methylated regions (A-DMRs) detected in 149 whole-genome bisulfite sequenced samples. The *PM20D1* promoter and *HOOK2* gene body enhancer loci in panels C and D are shown. (G) Isolated probes can also be part of A-DMRs. Top: Probe cg08477332, between *S100A14* and *S100A13,* displays preferential lack of methylation in AFR samples. Bottom: At least six contiguous CpGs neighboring cg08477332 display concordant methylation, a potential A-DMR.

See also Figures S3, S4 and Table S3.

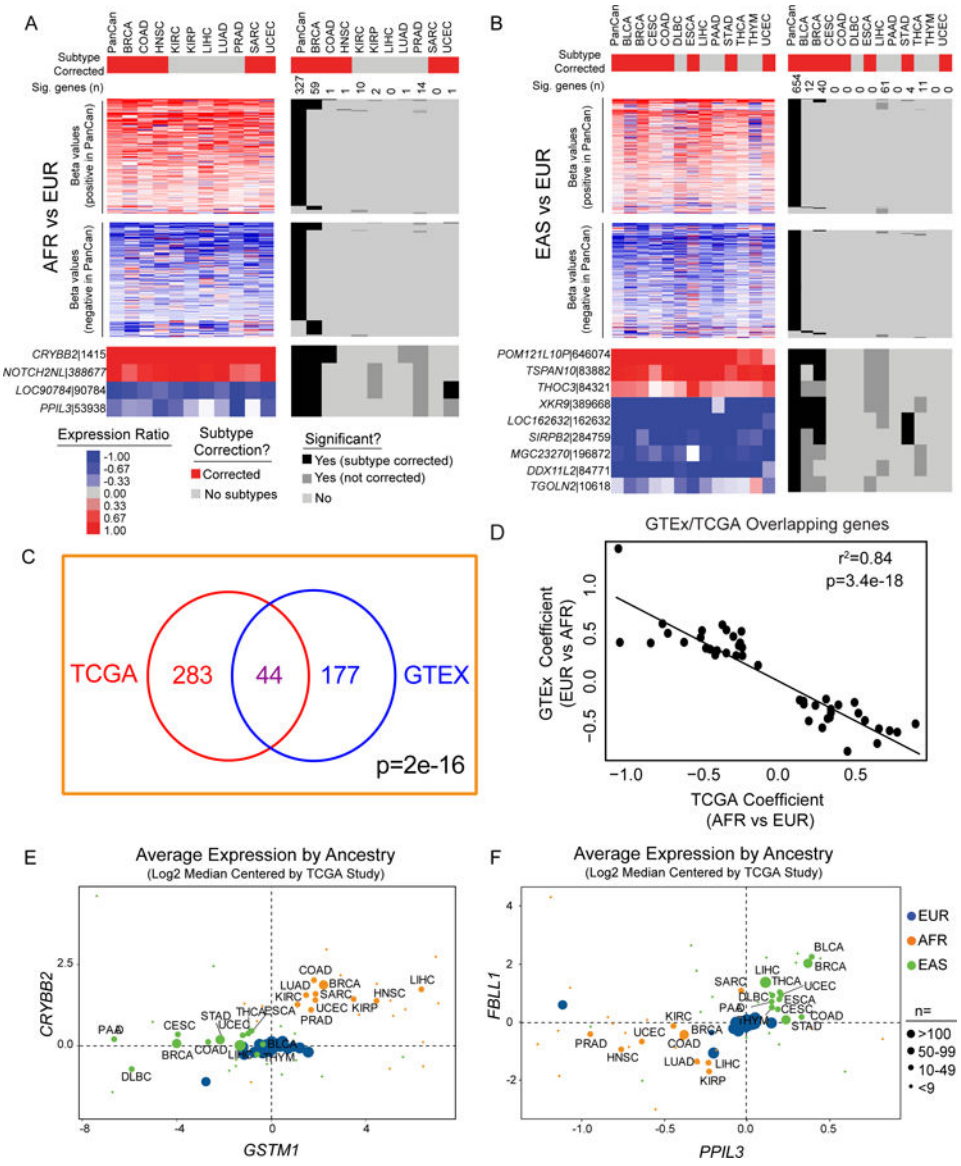


Figure 4. Ancestry-associated mRNAs.

Genes associated with ancestry (FDR q<0.001) after correcting for either batch alone or batch and cancer subtype. Expression ratios and significance levels were plotted for AFR (A) and EAS (B) associated genes. Genes that were significant in 33% of tumor types are highlighted (bottom). (C) Overlap of mRNAs associated with AFR or EUR ancestry, identified by either TCGA or GTEx. (D) Effect sizes (as regression coefficients) from TCGA (horizontal axis) and GTEx (vertical axis) analyses, for ancestry-associated mRNAs identified in both analyses. (E-F) Median levels per tumor type of the ancestry associated genes (E) *GSTM1* and *CRYBB2* (F) *PPIL3* and *FBLL1*. Dot sizes indicate sample sizes and colors indicate ancestry.

See also Figure S5 and Table S4.

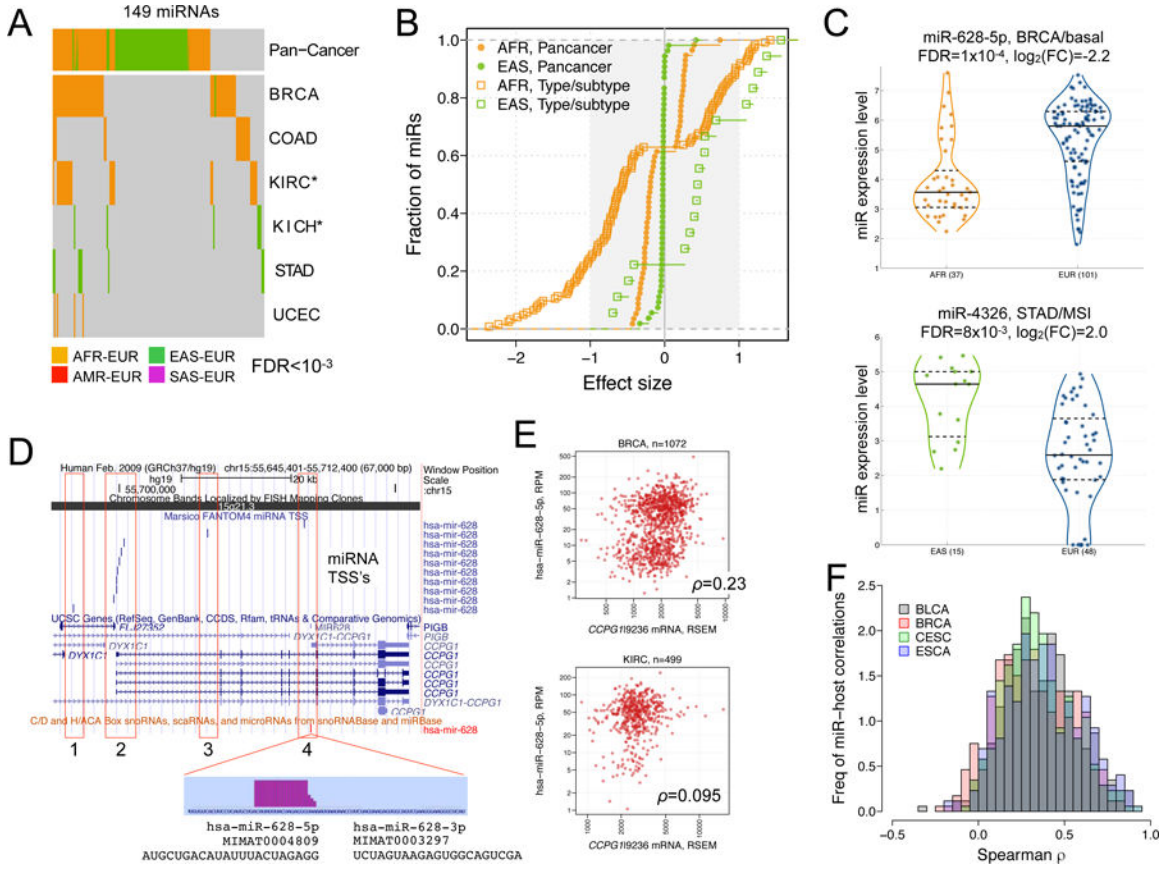


Figure 5. Ancestry-associated miRNA mature strands.

(A) Number of ancestry-associated miRs (FDR q<0.001), pan-cancer and in six cancer types. (B) Distributions of $\log_2$ fold changes for the associations in (A). (C) Ancestry-differential expression of miR-628–5p in basal BRCA, and miR-4326 in MSI STAD. Violin plot widths reflect kernel density estimates; solid and dashed lines reflect median and interquartile range. (D) Genomic neighborhood of hsa-mir-628, modified from the UCSC browser. The miRNA is within an intron of and on the same strand as host gene *CCPG1.* Red boxes are TSS loci (Marsico et al. 2013). The pale blue box at the bottom is a miRBase v22.1 read pileup on the miRNA's stem-loop sequence. (E) Expression of miR-628–5p and *CCPG1* in BRCA (left) and KIRC (right) samples, with Spearman rho values. (F) Distribution of rho values between hosted mature strands and host genes in BLCA, BRCA, CESC and ESCA.

See also Figure S6 and Table S5.

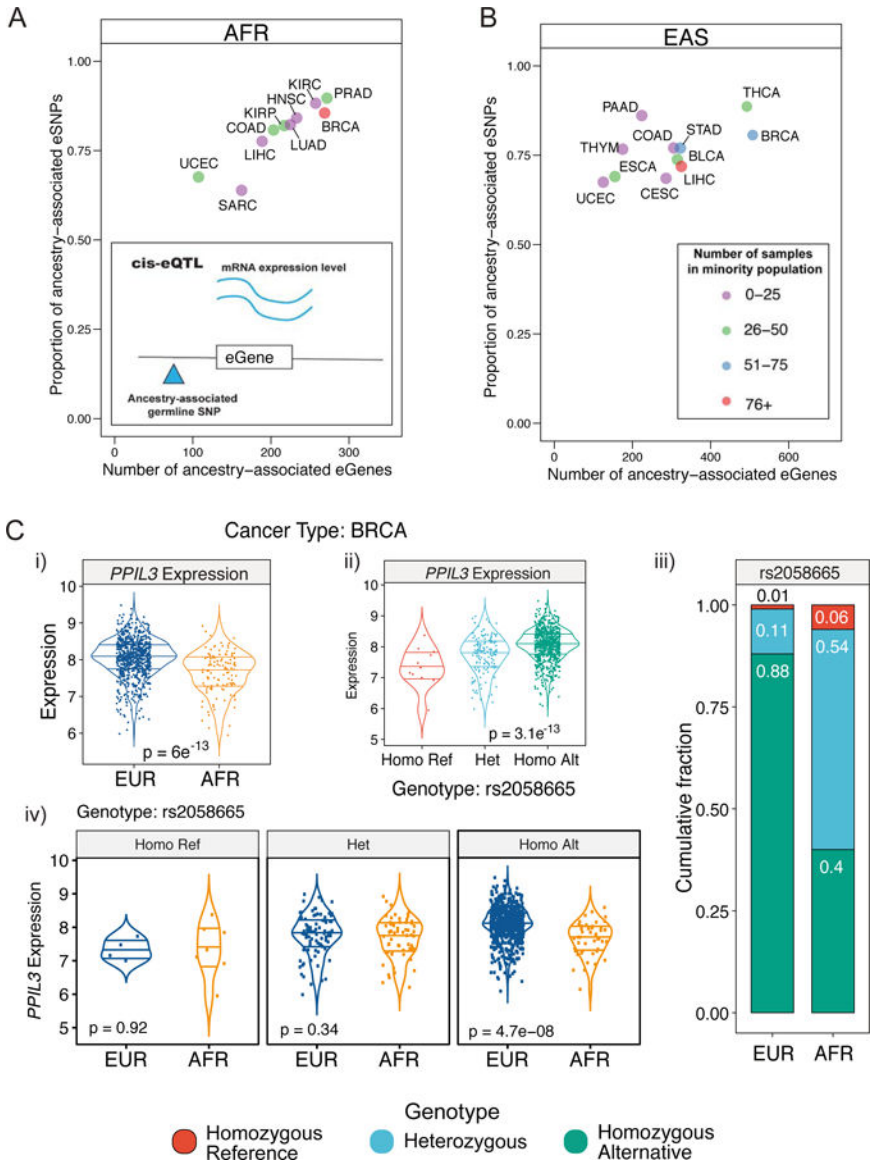


Figure 6. Ancestry-associated eQTLs.

Ancestry-associated germline variation *cis*-effects on expression in (A) AFR-EUR and (B) EAS-EUR comparisons. Dots, representing cancer type and colored by the number of samples in the minority population, are plotted against the number of PancanQTL eGenes with at least one ancestry-associated SNP (horizontal axis), and the proportion of ancestry-associated eSNPs (vertical axis). (C) Representative cis-eQTL rs2058665-PPIL3 in BRCA. (i-ii) *PPIL3* expression by (i) ancestry and (ii) SNP genotype. (iii) Proportions of samples with each genotype, by ancestry (Wald's association test, FDR q<0.01). (iv) *PPIL3* expression by genotype between EUR and AFR samples (p values: Wilcoxon test). Violin plot widths reflect kernel density estimates. Lines show median and interquartile ranges. See also Figure S7 and Table S6.

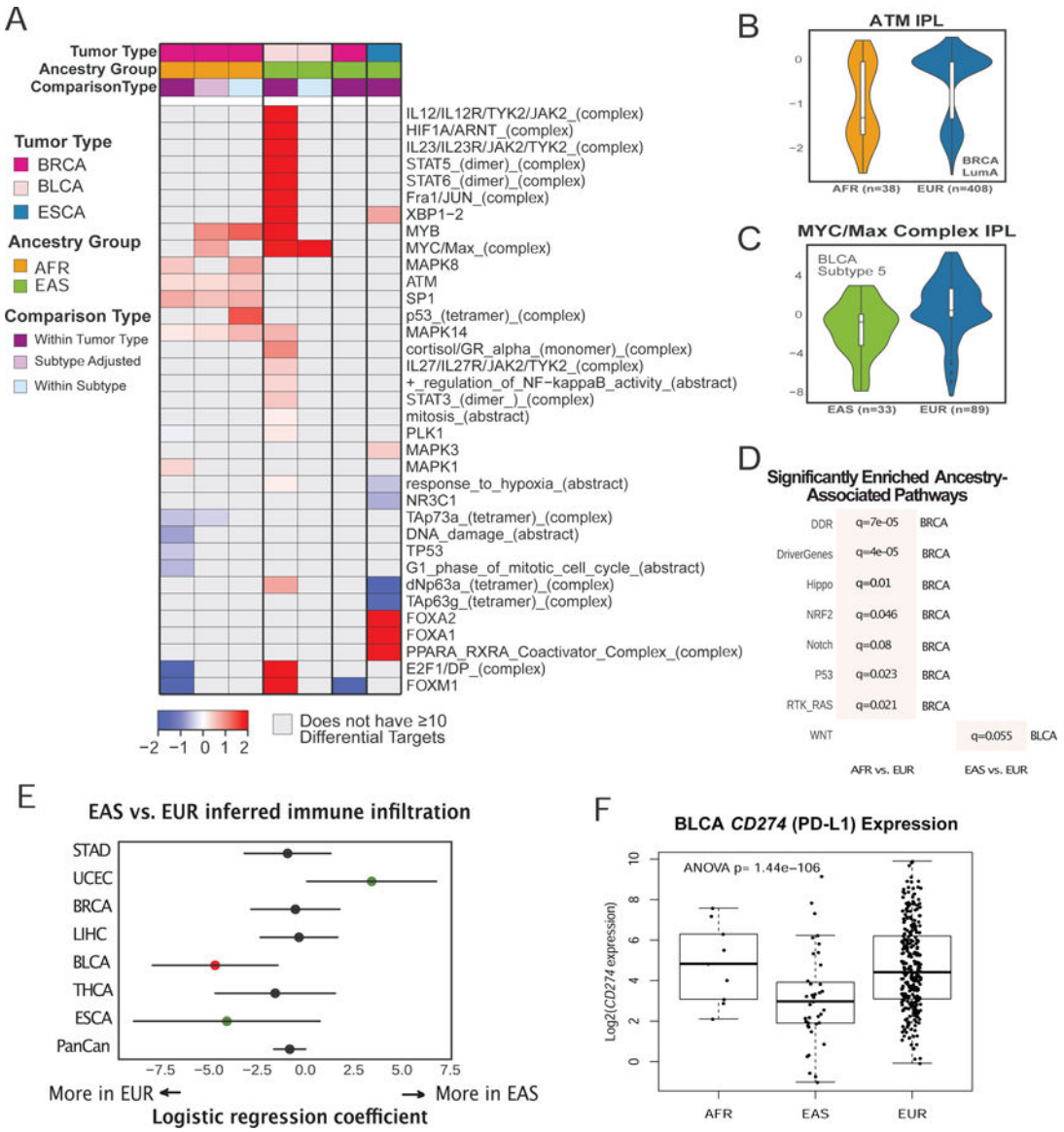


Figure 7. Ancestry-differential pathway features.

(A) Mean differences (red: higher in EUR; blue: lower in EUR) in PARADIGM-inferred integrated pathway levels (IPLs) of regulatory nodes with ≥10 ancestry-differential downstream targets, by tumor type. Gray denotes regulatory nodes that are not differential or have <10 differential downstream targets. (B) ATM IPLs of AFR and EUR Luminal A BRCA samples. (C) MYC/Max complex IPLs of EAS and EUR BLCA subtype 5 samples. In B and C, the violin plot widths reflect kernel density estimates and internal boxplots show median, interquartile range, and 1.5 times the interquartile range. (D) Cancer-associated genes and pathways enriched among differential pathway features between ancestry groups, from subtype-adjusted analyses. (E) Association of EAS ancestry with immune infiltration score. Coefficients from a multivariate logistic regression are shown on the horizontal axis. Red and green dots indicate correlations with FDR q < 0.05 and < 0.25, respectively. (F) Expression of *CD274*, which encodes PD-L1, in AFR, EAS, and EUR ancestries across all

cancers with at least 10 samples from the minority cohort. Boxeplots show median, interquartile range and 1.5 times the interquartile range.
See also Table S7.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
TCGA SNP 6.0 array and HumanMethylation450 array	TCGA legacy archive	https://portal.gdc.cancer.gov/legacy-archive/
TCGA whole-exome sequencing	Genomic Data Commons	https://portal.gdc.cancer.gov/
Haplotype Reference Consortium reference data	Haplotype Reference Consortium	http://www.haplotype-reference-consortium.org/
1000 Genomes project data	1000 Genomes project	http://www.internationalgenome.org/
Somatic mutation, copy number and other genomic/immune features	Genomic Data Commons; Taylor et al. 2018	https://gdc.cancer.gov/about-data/publications/pancanatlas
Mutational signature	Alexandrov et. al. (2018)	
Whole genome bisulfite sequencing	Genomic Data Commons; Schuyler et al, 2016	https://portal.gdc.cancer.gov/
TCGA mRNA normalized and miRNA data	Genomic Data Commons	https://gdc.cancer.gov/about-data/publications/pancanatlas
TCGA clinical and subtype data	Genomic Data Commons; Sanchez-Vega et al, 2018; Robertson et al, 2017	https://gdc.cancer.gov/about-data/publications/pancanatlas
ICGC PCAWG data	International Cancer Genome Consortium	https://dcc.icgc.org/pcawg
MSK-IMPACT data	cBioPortal	https://www.cbioportal.org/
Additional somatic validation datasets	Pena-Llopis et. al. 2012; Krishnan et. al. 2016; Nassar et al, 2019; Wu et al, 2019; Chang et. al, 2017	
Additional methylation validation datasets	Gene expression omnibus https://www.ncbi.nlm.nih.gov/geo/	GSE36369; GSE41826; GSE53816; GSE101431
miRBase	http://www.mirbase.org/	
Software and Algorithms		
AIM local ancestry	https://github.com/jcarrotzhang/ancestry-from-panel/tree/master/AIM_local_ancestry	
EIGENSOFT smartpca v9102	Price et al, 2002	https://github.com/chrchang/eigensoft/
ADMIXTURE v1.23	Alexander et al, 2009	http://software.genetics.ucla.edu/admixture/
PLINK v1.9	Purcell et al, 2007	https://www.cog-genomics.org/plink/1.9/
EthSEQ	Romanel et al, 2017	https://github.com/cibiobcg/EthSEQ
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://github.com/slowkoni/rfmix
McCarthy Group tools	https://www.well.ox.ac.uk/~wrayner/tools/	
Michigan Imputation Server	https://imputationserver.sph.umich.edu	
Eagle v2.3	Loh et al, 2016	https://data.broadinstitute.org/alkesgroup/Eagle/
Minimac 3	Howie et al, 2012	https://genome.sph.umich.edu/wiki/Minimac3
Hail framework	https://github.com/hail-is/hail	
openSeSAMe	Zhou et al, 2018b	https://bioconductor.org/packages/devel/bioc/vignettes/sesame/inst/doc/sesame.html
rtracklayer	Lawrence et al, 2009	https://www.bioconductor.org/packages/release/bioc/html/rtracklayer.html

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PARADIGM	Vaske et al, 2010	http://sbenz.github.io/Paradigm/