Accurate and robust inference of genetic ancestry from cancer-derived molecular data across genomic platforms

Pascal Belleau^{1,2}, Astrid Deschênes^{2,3}, David A. Tuveson^{2,3}, and Alexander Krasnitz^{1,2*}

*For correspondence:
krasnitz@cshl.edu (AK)

¹Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; ²Cancer Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; ³Lustgarten Foundation Pancreatic Cancer Research Laboratory, Cold Spring Harbor, New York, USA

Abstract Genetic ancestry-oriented cancer research requires the ability to perform accurate and robust ancestry inference from existing cancer-derived data, including whole exomes, transcriptomes and targeted gene panels, very often in the absence of matching cancer-free genomic data. In order to optimize and assess the performance of the ancestry inference for any given input cancer-derived molecular profile, we develop a data synthesis framework. In its core procedure, the ancestral background of the profiled patient is replaced with one of any number of individuals with known ancestry. Data synthesis is applicable to multiple profiling platforms and makes it possible to assess the performance of inference separately for each continental-level ancestry. This ability extends to all ancestries, including those without statistically sufficient representation in the existing cancer data. We further show that our inference procedure is accurate and robust in a wide range of sequencing depths. Testing our approach for three representative cancer types, and across three molecular profiling modalities, we demonstrate that global, continental-level ancestry of the patient can be inferred with high accuracy, as quantified by its agreement with the golden standard of the ancestry derived from matching cancer-free molecular data. Our study demonstrates that vast amounts of existing cancer-derived molecular data potentially are amenable to ancestry-oriented studies of the disease, without recourse to matching cancer-free genomes or patients' self-identification by ancestry.

Keywords

genetic ancestry, cancer, secondary data analysis

Introduction

There is ample epidemiological evidence that race and/or ethnicity are important determinants of incidence, clinical course and outcome in multiple types of cancer (*Siegel et al., 2020; Cronin et al., 2018; Ashktorab et al., 2017; Huang et al., 2019; Tan et al., 2016*). As such, these categories must be taken into account in the analysis of molecular data derived from cancer. A number of recently published large-scale genomic studies of cancer (*Mahal et al., 2020; Carrot-Zhang et al., 2020; Yuan*

39 *et al., 2018; Sinha et al., 2020; Bhatnagar et al., 2021; Carrot-Zhang et al., 2021*) point to differences
40 in the molecular make-up of the disease among groups of different ancestral background and to
41 the need for more molecular data to power discovery of such differences.

42 Ancestry annotation of cancer-derived data largely draws on two sources. One is a patient's
43 self-identified race and/or ethnicity (SIRE). SIRE is often missing, sometimes inaccurate and usually
44 incomplete. As a recent analysis (*Nugent et al., 2019*) of PubMed database entries since 2010
45 reveals, patients' SIRE is massively under-reported in genome and exome sequencing studies of
46 cancer, with only 37% of these reporting race, and 17% reporting ethnicity. Furthermore, SIRE is
47 not always consistent with genetic ancestry. Finally, a self-declaring patient is often given a choice
48 from a small number of broad racial or ethnic categories, which fail to capture complete ancestral
49 information, especially in cases of mixed ancestry (*Mersha and Abebe, 2015*).

50 A far more accurate and detailed ancestral characterization may be obtained by genotyping
51 a patient's DNA from a cancer-free tissue. Powerful methods exist for ancestry inference from
52 germline DNA sequence (*Pritchard et al., 2000; Price et al., 2006; Alexander et al., 2009; Diaz-
53 Papkovich et al., 2019*). These methods were recently used to determine ancestry of approxi-
54 mately 10,000 patients profiled by The Cancer Genome Atlas (TCGA) (*Carrot-Zhang et al., 2020;
55 Yuan et al., 2018*). However, genotyping of DNA from patient-matched cancer-free specimens is
56 not part of standard clinical practice, where the purpose of DNA profiling is often identification
57 of mutations with known oncogenic effects, such as those in the Catalog Of Somatic Mutations In
58 Cancer (COSMIC) database (*Tate et al., 2018*). As a result, it is not performed routinely outside aca-
59 demic clinical centers or major research projects. There also are studies yielding sequence data
60 from tumors, whose purpose does not require germline profiling. RNA sequencing (RNA-seq) for
61 expression quantification is in this category. Finally, peripheral blood is most often the source of
62 germline DNA in the clinic, but this is not always the case for diseases of the hematopoietic system,
63 such as leukemia, wherein cancer cells are massively present in circulation. In summary, matched
64 germline DNA sequence is not universally available for cancer-derived molecular data. In such
65 cases, it is necessary to infer ancestry from the nucleic acid sequence of the tumor itself.

66 Standard methods of ancestry inference commonly rely on population specificity of germline
67 single-nucleotide variants (SNV). Whole-genome (WGS) or whole-exome sequences (WES), at depths
68 sufficient for reliably calling single-nucleotide variants, and readouts from genotyping microarrays,
69 are therefore data types most suitable for this purpose. However, such detailed DNA profiling is
70 often not performed in molecular studies of cancer. In such cases, it is necessary to infer ancestry
71 from other types of tumor-derived data, including RNA sequence and DNA sequence for a small
72 panel of genes, e.g., FoundationOne® CDx (*Frampton et al., 2013*).

73 For all types of tumor-derived sequence, accurate inference of ancestry is a potential challenge.
74 Tumor genome is often replete with somatic alterations, including loss of heterozygosity (LOH),
75 copy number variants (CNV), translocations, microsatellite instabilities and SNV. Of these, struc-
76 tural variants, especially LOH and CNV, are the most likely to affect the genetic ancestry calls, but
77 other types of alterations also are, to various degrees, potential obstacles to accurate ancestry
78 inference. Tumor RNA-seq presents additional challenges, namely, extremely uneven coverage of
79 the transcript due to a broad range of RNA expression levels and distortions due to allele-specific
80 expression. Gene panels represent a very small fraction of the genome, whose sufficiency for an-
81 cestry inference is not clear and may vary from panel to panel. In addition, cancer gene panels are
82 enriched in cancer driver genes, which tend to undergo somatic alteration more frequently than
83 other parts of the genome.

84 Important recent publications on ancestral effects in cancer reported patient ancestry inferred
85 from matching cancer-free DNA (*Carrot-Zhang et al., 2020; Yuan et al., 2018; Carrot-Zhang et al.,
86 2021*). At the same time, there has been much less work on ancestry inference from tumor-derived
87 nucleic acids. A recent analysis of tumor genomes from TCGA and GEO repositories, profiled by
88 SNP microarrays, demonstrated a high degree of coincidence between patient ancestries inferred
89 from these data and those inferred from SNP profiles of matching germline genomes (*Huang and*

90 *Baudis, 2020*). This study did not report inference results from other molecular profiling modalities.
 91 Similar agreement has been found, for a set of over 300 cancer cell lines, between the self-declared
 92 race/ethnicity of the donors and ancestry inferred from the SNP array data (*Yuan et al., 2018*), but
 93 that finding was not validated against matching cancer-free data. Ancestry was also inferred in
 94 two large collections of cancer cell lines using SNP microarray data (*Dutil et al., 2019; Kessler et al.,*
 95 *2019*). In the absence of matching cancer-free genotypes or self-declared ancestry of the donor
 96 the inference accuracy could not be assessed in these two studies. Ancestry inference from RNA
 97 sequences, 174 of which were derived from cancer tissue specimens, was considered in a recent
 98 study (*Barral-Arca et al., 2019*). However, these inferred ancestries were neither compared to
 99 ancestry calls from germline sequence nor to self-declared ancestries for accuracy assessment.
 100 Ancestry has been inferred for a large set of patient cases profiled with the FoundationOne® CDx
 101 gene panel (*Frampton et al., 2013*), but these ancestry calls were neither compared to those from
 102 the germline sequence nor to the patients' SIRE. A more recent study (*Carrot-Zhang et al., 2021*)
 103 compared, with encouraging results, ancestry inference from cancer-derived FoundationOne® CDx
 104 data to matching cancer-free ancestry calls, but this analysis was confined to lung cancer in mixed
 105 American super-population. To our knowledge, no systematic computational framework for an-
 106 cestry inference from cancer-derived molecular data, across assay and cancer types, has been
 107 developed to date. There is presently no ability to assess the inference accuracy specifically for
 108 a given input tumor-derived molecular profile with all its attendant properties, including the data
 109 quality and the depth of coverage. Reliable and accurate ancestry inference from tumor-derived
 110 nucleic acids thus represents an unmet need, which the present work aims to address.

111 For this purpose, we designed an inference procedure having in mind a scenario, likely to occur
 112 in studies of existing data or of archived tissue specimens, with an input molecular profile of a tu-
 113 mor from a single patient, and no matching cancer-free sequence available. The profile in question
 114 may have its unique set of sequence properties. These include the target sequence and uniformity
 115 of its coverage, depth, read length and sequencing quality. These profile-specific properties may be
 116 vastly dissimilar from those in the available public data sets with reliably known genetic ancestry of
 117 the patients. Furthermore, not all ancestries are equally easy to infer: for example, a Mixed Amer-
 118 ican ancestral category is sometimes difficult to distinguish either from African or from European
 119 ancestry. This profile specificity would make it impossible to confidently assess the accuracy of the
 120 inference procedure for the input profile from its performance with the public cancer-derived data
 121 in aggregate. In order to overcome this difficulty, we develop a computational technique, which is
 122 described schematically in *Figure 1* wherein the ancestral background of the patient is supplanted
 123 in the input profile by one of an unrelated individual with known ancestry. We next apply estab-
 124 lished methods of ancestry inference to this synthetic profile and compare the result to that known
 125 ancestry. Generating multiple such synthetic profiles allows us to assess how accurate the ancestry
 126 inference is for the patient, both overall and as a function of the profile's continental-level ancestry.
 127 Furthermore, using synthetic data, we are able optimize the inference procedure with respect to
 128 parameters on which it depends. Importantly, this assessment and optimization procedure does
 129 not require the profile in question to be part of a larger data set from a cohort of patients with a
 130 similar diagnosis. Very often in public cancer-derived data, such cohorts do not provide statistically
 131 meaningful representation of non-European ancestries. This insufficiency is not an impediment to
 132 the application our methodology.

133 In the following, we assess the accuracy of global ancestry calls from tumor exomes, narrowly
 134 targeted gene panels and RNA sequences, in comparison to such calls from matching germline
 135 genotypes, as profiled by exome sequencing or SNP microarrays. We do so for three cancer types,
 136 namely, pancreatic adenocarcinoma (PDAC) and ovarian cystadenocarcinoma (OV) as representa-
 137 tive types of epithelial tumors, and acute myeloid leukemia (AML), as an example of hematopoietic
 138 malignancy. Each of these data sets represents a unique challenge for patients' ancestry inference.
 139 OV is characterized by massive copy number alterations, often spanning much of the genome. Our
 140 PDAC data originate from patient-derived organoid (PDO) models of the disease (*Tiriac et al., 2018*).

141 In PDO, near-100% tumor purity is achieved, exacerbating effects of copy number loss and loss of
142 heterozygosity on the sequence. In AML the peripheral blood, the usual source of cancer-free DNA,
143 may be severely contaminated by the cancer.

144 Results

145 We assessed the performance of genetic ancestry inference from three genomic data types: whole
146 exomes, gene panels targeting exomes of several hundred cancer-related genes each and RNA se-
147 quences. Throughout the study, we used the 1000 Genomes (1KG) data set, with no relatives for
148 the individuals included (*Altshuler et al., 2010; Fairley et al., 2019*), as reference, against which pa-
149 tient molecular data were compared to infer continental-level global ancestry. The latter is defined
150 as a categorical variable taking five values: African (AFR), East Asian (EAS), European (EUR), Mixed
151 American (AMR) and South Asian (SAS). These are called super-populations in the 1KG terminology.
152 Each super-population comprises a number of subcontinental-level populations ((*Fairley et al.,*
153 *2019*)).

154 Our assessment relied on molecular data collected from three patient cohorts, each represent-
155 ing a cancer type, namely, tissue donors to the Cold Spring Harbor Laboratory (CSHL) pancreatic
156 ductal adenocarcinoma (PDAC) library of patient-derived organoids; acute myeloid leukemia (AML)
157 patients enrolled in Beat AML clinical trial; and patients comprising TCGA ovarian cancer cohort
158 (TCGA-OV) (*The Cancer Genome Atlas Research Network, 2011*). In these cohorts, patient molecu-
159 lar data were available from tissue specimens both of cancer and cancer-free. *Figure 2* and Sup-
160 plementary Table S2 contain a summary of molecular data underlying the study.

161 We employed principal-component analysis (PCA) as our inference tool of choice, and applied
162 it as follows (*Figure 3*) (*Alexander et al., 2009*).

163 As a basis for the analysis, we used genotypes at genomic positions where single-nucleotide se-
164 quence variants occurred with a frequency above a threshold in at least one super-population as
165 sampled by 1KG. This basis was further reduced, for each individual cancer-derived molecular pro-
166 file, to genotypes at positions with high sequence coverage by high-quality reads in the profile. We
167 then computed singular-value decomposition of the reduced 1KG genotype matrix and projected
168 the genotype of the cancer-derived profile onto the first D of the resulting principal components.
169 The ancestry for the profile was determined as that of the majority among the nearest K 1KG neigh-
170 bors of the profile in this D -dimensional space (*Yuan et al., 2018*). For a subset of patients in each
171 cohort we individually assessed the performance of the ancestry inference, as a function of the
172 parameters D and K . This assessment was based, for each patient in the subset, on a large num-
173 ber of synthetic cancer-derived molecular profiles, as outlined in the Introduction, schematically
174 described in *Figure 5* and explained in greater detail in the Methods section. The result was quan-
175 tified, for a given D, K pair of parameters, as the area under receiver operating characteristic (AU-
176 ROC) (*Robin et al., 2011; Sun and Xu, 2014; Hand and Till, 2001*). Both super-population-specific and
177 overall AUROC values were computed in a range of D, K pairs, as illustrated in *Figure 4* for 10 PDAC
178 patients and AMR-specific AUROC (the similar figures for all the cohorts and super-populations are
179 in Figure S1). Optimal D, K pairs maximizing the overall AUROC were chosen. From this subset of
180 patients we observed, for each cancer type considered and for each of the three molecular profil-
181 ing modalities, an optimal range of D and K parameters where the performance of inference was
182 consistently high in the subset and only weakly dependent on these parameters (Figure S1). We
183 then selected and used, for the remainder of the patients with this cancer type and for this profiling
184 modality, a pair D and K values from within the optimal range. As an additional validation of our
185 parameter optimization procedure, we applied it to a set of cancer-free WES profiles of TCGA-OV
186 patients. Comparing the resulting ancestry calls to the consensus calls (C5) by TCGA (*Carrot-Zhang*
187 *et al., 2020*), we find the two to be in excellent agreement Table S3.

188 We also assessed the cohort-wide performance of our ancestry calls from original cancer-derived
189 molecular data, by comparison to the gold standard of ancestry as determined from the match-
190 ing cancer-free genotypes. For Beat AML and TCGA-OV patients, we performed ancestry inference

Figure 2. Summary of the molecular data used in this study. These originate from three patient cohorts: **A)** TCGA ovarian cancer **B)** acute myeloid leukemia and **C)** pancreatic ductal adenocarcinoma library of patient-derived organoids. **D)** The distribution of the patients by SIRE for Beat AML, PDAC and TCGA-OV cohorts. UNK means not reported or unknown.

Figure 3. A flowchart of the inference of genetic ancestry.

Figure 4. Dependence of AMR-specific AUROC on the inference parameters D and K , computed using data synthesis for 10 PDAC patients and the three profiling modalities: WES, RNA-seq and FoundationOne[®] CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

Study	D	K	Accuracy	95% CI	AUROC	95% CI
TCGA-OV WES	5	13	0.998	0.994-1	0.993	0.992-0.994
TCGA-OV Panel	4	12	0.984	0.972-0.996	0.966	0.965-0.967
TCGA-OV RNA-seq	7	12	0.993	0.983-1	0.977	0.975-0.979
BeatAML WES	5	13	0.989	0.978-1	0.978	0.976-0.980
BeatAML Panel	4	13	0.991	0.981-1	0.999	0.999-0.999
BeatAML RNA-seq	4	13	0.992	0.981-1	0.999	0.999-0.999
PDAC WES	8	13	1	NA	NA	NA
PDAC Panel	6	5	0.952	0.861-1	0.958	NA
PDAC RNA-seq	4	13	1	NA	NA	NA

Table 1. Cohort-wide performance measures for super-population calls from cancer-derived molecular data, as compared to the matching cancer-free WES or (in the case of PDAC) WGS. A reliable estimate of the confidence intervals (CI) was not possible in the case of PDAC, due to the small number of cases with matching cancer-free genotypes.

from cancer-free patient exomes, using the same methodology as we did for the cancer-derived sequences of these patients. In the case of PDAC, cancer-free whole-genome sequencing data were available, and used for the same purpose for a portion of the patient cohort. For all three cohorts, we summarize our cohort-wide findings in **Table 1** (we include similar tables for the synthetic data Table S9-S11). Ancestry calls from both microarray- and exome-derived genotypes were recently published by TCGA consortium (*Carrot-Zhang et al., 2020*), and we also used these so-called consensus (C5 in the following) calls in our performance assessment for TCGA-OV (Table S3).

We note that in the three patient cohorts we analyze here the sampling of patients with non-European ancestries is statistically insufficient for a purely cohort-based assessment of performance (**Table 2** and Table S5). We therefore report cohort-wide overall but not super-population specific AUROC values. Using data synthesis, we are able to compensate for this data shortfall in non-European ancestries and estimate super-population specific AUROC, as explained above (Tables S6,S7 and S8 and Figure S1).

The results of our analysis as presented in Tables S6,S7 and S8, lead to the following key observations. First, we demonstrate a consistently high performance of our inference procedure across all cohorts and profiling modalities. Second, the super-population specific performance was the highest for the European and both Asian super populations. The slightly lower accuracy as observed for the African and mixed American super-populations is likely due to a greater genetic variability within the African super-population and to a higher degree of (the predominantly European) admixture in both super-populations. Third, the optimal choice of the *D*, *K* inference parameters, in general, depends on an individual cancer-derived molecular profile, even within the same cancer type and profiling modality (Figure S1 B,G,L).

In order to examine whether our inference procedure is robust against variation in the sequence target coverage, we re-computed the ancestry calls for a subset of ten OV patients, with the cancer-derived whole-exome and RNA sequences of these patients down-sampled to between 75% and 10% of the original coverage. The results, presented in (Figure S2) exhibit no substantial sensitivity of the inference accuracy to the depth of coverage in this range.

Discussion

With this work, we introduce a systematic approach to ancestry inference from cancer-derived molecular data. The approach is rooted in a combination of an established, extensively used PCA-based technique of ancestry inference with a central idea of inference parameter optimization using data synthesized *in silico*. Crucially, this combination permits a statistically rigorous assessment of inference accuracy for an individual cancer-derived molecular profile, with its unique biological

(a) TCGA-OV WES

		Inferred				
		EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	10	0	0	0	0
	EUR	0	378	0	0	0
	AFR	0	0	29	0	0
	AMR	0	1	0	16	0
	SAS	0	0	0	0	7
	UNK	0	2	0	0	0

(b) BeatAML WES

		Inferred				
		EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	11	0	0	0	0
	EUR	0	283	0	6	0
	AFR	0	0	14	0	0
	AMR	0	0	0	27	0
	SAS	0	0	0	0	2
	UNK	0	0	0	0	0

(c) TCGA-OV Panel

		Inferred				
		EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	10	0	0	0	0
	EUR	0	376	0	2	0
	AFR	0	0	28	1	0
	AMR	0	4	0	13	0
	SAS	0	0	0	0	7
	UNK	0	2	0	0	0

(d) BeatAML Panel

		Inferred				
		EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	11	0	0	0	0
	EUR	0	286	0	3	0
	AFR	0	0	14	0	0
	AMR	0	0	0	27	0
	SAS	0	0	0	0	2
	UNK	0	0	0	0	0

(e) TCGA-OV RNA

		Inferred				
		EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	4	0	0	0	0
	EUR	0	242	0	0	0
	AFR	0	0	21	0	0
	AMR	1	1	0	9	0
	SAS	0	0	0	0	4
	UNK	0	1	0	0	0

(f) BeatAML RNA

		Inferred				
		EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	10	0	0	0	0
	EUR	0	210	0	2	0
	AFR	0	0	9	0	0
	AMR	0	0	0	24	0
	SAS	0	0	0	0	1
	UNK	0	0	0	0	0

Table 2. Confusion matrices comparing TCGA-OV or Beat AML patients' super-population calls from the cancer-derived molecular profiles for the three profiling modalities (rows) to those from the matching cancer-free WES.

(e.g. cancer type) and technical (e.g., sequencing depth and quality) properties. Synthetic data here are used as a substitute for a real-world set of molecular profiles sharing these properties and with known ground-truth genetic ancestry. It is unrealistic to expect such a real-world set to be available in all cases. Our tests of the resulting computational methodology on a representative subset of cancer-derived data demonstrate its accurate and robust performance. As we describe in detail in the Methods section, our data synthesis method relies on heuristic components for an estimate of the allele fractions throughout the cancer-derived profile. This estimate can be made more rigorous by using haplotypes in future implementations of the method, but the present version produces allele fractions in good agreement with published allele fractions (ASCAT2 results in (Grossman *et al.*, 2016; NCI, 2021)).

A line of research and development initiated with this work must be extended in several directions. First, the performance of the methods presented must be examined more comprehensively across cancer types, and sequence properties, such as quality and depth. This task is computing-intensive but feasible given extensive, well annotated repositories of cancer-derived data, such as those resulting from TCGA Research Network (Network, 2021) and ICGC (Zhang *et al.*, 2019) projects. For these, the genetic ancestry of the patients either is known or can be readily es-

240 tablished using matching cancer-free molecular data. Second, an extension of our approach to
241 additional profiling modalities should be examined. Chief among these are low-coverage whole-
242 genome sequences commonly used for copy-number analysis and single-molecule, long-read se-
243 quences. Each of these presents unique challenges and opportunities for the ancestry inference:
244 in the former, the sparsity of coverage is compensated by its whole-genome breadth; in the lat-
245 ter, the trade-off is between the high sequence error rate and the long-distance phasing afforded
246 by long reads. Third, while the present work relied on PCA followed by nearest-neighbor classifi-
247 cation for ancestry assessment, alternatives including UMAP for the former and Random Forest
248 or Support Vector Machine for the latter exist and should be evaluated. Third, future method de-
249 velopment should be extended beyond inference of global ancestry to that of local ancestry and
250 ancestral admixture. Such an extension is particularly important in the study of cancer in strongly
251 admixed populations, such as African and Latin Americans and may require more extensive refer-
252 ence data, in addition to the 1KG reference used here. Finally, beyond cancer, our methodology
253 can be applied to inference from genomic data originating in any kind of fragmentary or damaged
254 nucleic-acid specimens, such as those encountered in forensic, archaeological or paleontological
255 contexts.

256 We anticipate the computational approach described here to have a major, two-fold, impact
257 on investigation of links between ancestry and cancer. First, it will become possible to massively
258 boost the statistical power of such studies by leveraging existing tumor-derived molecular data
259 sets without matching germline sequences or ancestry annotation. Our search of the Gene Ex-
260 pression Omnibus (GEO) database alone has identified over 1,250 such data sets, containing RNA
261 expression data for nearly 48,000 cancer tissue specimens. Such resources dwarf those of fully an-
262 notated repositories, such as TCGA and International Cancer Genome Consortium (ICGC) (*Zhang*
263 *et al.*, 2019). Other molecular data repositories are likely to contain resources of this category on a
264 similar order of magnitude. Second, hundreds of thousands of tumor tissue specimens stored at
265 multiple clinical centers constitute another major resource for ancestry-aware molecular studies of
266 cancer. Here again, matching normal tissue specimens are often absent, and so is ethnic or racial
267 annotation for the patients. According to a recent estimate (*Polubriaginof et al.*, 2019) such anno-
268 tation is missing in electronic health records of over 50% of patients. Inferential tools presented
269 here will make these massive resources of archival tissues available for ancestry-oriented cancer
270 research.

271 **Methods and Materials**

272 **Data sets and pre-processing**

273 The data sets used in this work originate from three sources: TCGA collection for ovarian cystadeno-
274 carcinoma (*The Cancer Genome Atlas Research Network*, 2011), Beat AML clinical trial (*Tyner et al.*,
275 2018), and a study of pancreatic ductal adenocarcinoma (PDAC) using patient-derived organoids
276 (*Tiriac et al.*, 2018). For all three, the data used are summarized, in the form of Venn diagrams and
277 included cancer DNA (whole-exome or whole-genome) sequence, cancer RNA sequence and match-
278 ing DNA (whole-exome or whole-genome) sequence. In all cases, read data mapped to the hg38
279 version of the human genome were used. In order to study ancestry inference from targeted pan-
280 els, the cancer-derived whole-exome data were reduced to reads mapping to the FoundationOne®
281 CDx cancer-related gene panel (*INC*, 1999). Reads in the cancer-derived data were filtered for qual-
282 ity using a cutoff phred score of 20. Following this filter, single-nucleotide substitutions were called
283 at all positions with read coverage of at least 10, using Varscan version 2.4.4 (*Koboldt et al.*, 2013).
284 This set of positions is called the high-confidence substitution (HCS) set in the following. From the
285 1000 Genomes (1KG) variant call data in the Variant Call Format (VCF) (*Lowy-Gallego et al.*, 2019),
286 genomic positions where substitution variants occur at a frequency of at least 0.01 in at least one
287 of the super-populations comprising 1KG were selected as a basis for the ancestry inference. This
288 set is referred to as the high-frequency substitution (HFS) set in the following. At the HFS positions

in the cancer-derived profile with the coverage above 10, the genotype was called. This set of positions is referred to as high-confidence genotype (HCG) set in the following. In the HCG set, the total read count and the read counts for the reference and the alternative (according to HFS) alleles were determined. A genotype at an HCG position was considered undetermined if the excess of the total read count over the sum of the reference and alternative counts was inconsistent with the error of 0.001 at the $p = 0.001$ level of significance. The same rule was used to call a heterozygous genotype. The HCG genomic positions were pruned to reduce correlation between neighboring genotypes using Bioconductor SNPRelate package version 1.22.0 (Zheng et al., 2012), resulting in the pruned high-confidence genotype (PHCG) set of positions.

Ancestry inference

Figure 3 lays out the workflow for ancestry inference. For a given cancer-derived profile, principal component analysis of the 1KG genotypes reduced to the PHCG was performed, and D top principal components retained. The patient genotype reduced to PHCG was projected onto the subspace spanned by these D components. Within this subspace, the patient's ancestry was called as that of the 1KG super-population with the highest number of 1KG individuals among K nearest neighbors of the patient's genotype, using Euclidean distance in the D -dimensional subspace. If two or more super-populations were found tied in the nearest-neighbor count, no ancestry call was made for the patient. Only two such ties were observed in this work.

Measures of performance

We evaluate the performance of the ancestry inference by comparison to the ancestry inferred from the matching cancer-free data, wherever the latter are available. This is the case for the entirety of Beat AML and the OV data. For both, we infer the ancestry from the matching cancer-free exome profiles. In the case of OV data, we also compare the results to the consensus ancestry call (Carrot-Zhang et al., 2020). In the case of PDAC matching cancer-free WGS data are available for 22 patient cases (Figure 2), and our assessment of accuracy is based on this subset of the data. We compute, for each dataset, the 5×5 confusion matrix (CM) for the 1KG superpopulation calls from the cancer-derived and cancer-free data sources. From the CM, the call accuracy is computed as the sum of the diagonal terms divided by that of the whole CM. Since the ancestral composition of all data sets considered here is heavily skewed towards the European super-population, we also compute the multi-class version of the area under the receiver operating characteristic curve (AUROC) (Hand and Till, 2001). AUROC is a measure of the call quality which compensates for the asymmetry in the class sizes. We use an R package pROC (CRAN version 1.16.2) (Robin et al., 2011) for this purpose, and compute both the class-specific AUROC for each super-population and the 5-class AUROC. In the class-specific case, we use a version DeLong's algorithm Sun and Xu (2014); DeLong et al. (1988) as implemented in the pROC package to compute the AUROC confidence intervals. In the 5-class case the confidence intervals are computed using bootstrap with 100-fold sampling.

Data synthesis

Data synthesis is defined here as replacement of the sequence variants detected in a cancer-derived profile P by those found in the genome of an unrelated individual U . Ingredients required for this procedure are: (a) allele fraction (AF) estimates in P and (b) the haplotype of U in the portion of the genome covered by P . With this knowledge, the procedure, depicted in Figure 5, consists of the following steps. First, sequence reads comprising P are distributed at random among the alleles with probabilities equal to the observed allele fractions. Second, in each haplotype block in the genome of U that is covered by P , allele assignment is made at random, yielding variant and reference read counts for each substitution in the genome of U within the scope of P .

Figure 5. A schematic overview of the data synthesis process.

Inference parameter optimization using synthetic data

In order to optimize ancestry inference parameters D and K for a given cancer-derived molecular profile, we generate a synthetic data set by repeatedly pairing the profile with 1KG genomes. A subset of 780 1KG genomes is set aside for this purpose by drawing at random 30 genomes from each of the 26 ancestral populations represented in 1KG. Genetic ancestry is then inferred for each of the 780 synthetic profiles following the procedure described in the Ancestry Inference subsection, each time with the 1KG genome used for synthesis removed from the reference data set. The inference performance is then assessed as the 5-class AUROC, as explained in the Measures of Performance subsection. AUROC is computed for the D, K pairs in a range of values of these parameters, and the optimal D, K pairs yielding the highest accuracy are identified. Throughout this work, AUROC was computed for all D and K in the rectangle $3 \leq D \leq 11; 3 \leq K \leq 15$. For all combinations of data sources and profiling modalities considered, a set of D, K pairs was found where the performance was optimal or differed from the optimum by no more than 3% (Figure 4).

Determination of allele fractions

As the Data Synthesis subsection makes clear, knowledge of allele fractions (AF) in a cancer-derived profile is a prerequisite for data synthesis. We describe a 3-step AF estimate procedure which relies exclusively on the cancer-derived molecular profile, in the absence of a matching cancer-free genotype from the patient, as would be the case for the intended application of our methods. First (step 1), the loss-of-heterozygosity (LOH) regions are delineated. Next (step 2), the regions of allele imbalance where AF differs significantly from 1/2 are identified. Finally (step 3), AF are computed throughout the regions of allele imbalance. These steps are implemented differently, depending on whether the profile originates in the cancer DNA or RNA. We now discuss these steps, in turn for the DNA- and the RNA-derived profiles (Figure S3).

For the DNA-derived profiles, the LOH regions (step 1) are detected as follows. An LOH region in P must fit into a gap G between any two consecutive HCS positions, where all the observed genotypes are consistent with homozygosity. Any region within G is then considered an LOH region (see Figure S3 b) if it contains k_1 PHCG positions with $k_1 \geq k_{min}$ and for which the 1KG frequencies F_i , $1 \leq i \leq k_1$ of the alleles observed in the cancer-derived profile P satisfy

$$\log_{10} \left(\prod_{i=1}^{k_1} \frac{F_i^2}{\max [F_i^2, (1 - F_i)^2, 2F_i(1 - F_i)]} \right) < \lambda.$$

PHCG positions only are used for this purpose, to reduce correlations due to linkage. The values of k_{min} and λ were chosen so as to maximize, in TCGA OV data set, the overlap between the regions found to be LOH by these criteria and the published LOH regions ASCAT2 files from NCI's

361 Genomic Data Commons ((*Grossman et al., 2016; NCI, 2021*)). The latter were determined with full
362 knowledge of the patient's cancer-free genotype. The optimal values were found to be $k_{min} = 3$ and
363 $\lambda = -3$.

Step 2 is based on the notion of an "empty box" (see Figure S3 b). By this, we mean a contiguous region where the allele fraction of 1/2 is inconsistent with the read counts for the reference and alternative alleles at the HCS positions it contains. An empty box is constructed as follows. First, we consider sliding windows, each encompassing k_2 consecutive HCS positions not separated by an LOH region. A window is called asymmetric if (a) for no less than $k_2 - 1$ of the positions the minor allele count is outside the inner-quartile range (IQR) of the binomial distribution with the minor AF of $f_0 = 1/2$ and (b) satisfy

$$\log_{10} \left(\prod_{i=1}^{k_2} \frac{2P_i}{(1 - 2P_i)} \right) < \lambda.$$

364 where $P_i = P(X_i \leq \text{number of reads covering the minor allele at position } i)$ and, X_i is the binomial
365 distribution with the number of trials equals the coverage at the position i and the probability of
366 success $\rho = 1/2$. In this work, $\lambda = -3$. A polymorphic position is called asymmetric if it belongs
367 to at least one asymmetric window. An empty box is a region with no less than k_2 polymorphic
368 positions, all of which are asymmetric. We used $k_2 = 10$ throughout this work.

369 At step 3, in the case of DNA, we consider contiguous genome regions of allele asymmetry iden-
370 tified at step 2. Each of these may consist of sub-regions with differing allele fractions. To detect
371 these sub-regions, we "seed" the first sub-region with k_3 HCS positions at the region's boundary
372 and, in this window, estimate the minor allele fraction. We consider the adjacent window W of
373 k_3 HCS positions $k_3 + 1$ through $2k_3$ and apply to it the empty box criteria as described for step 2,
374 with f_0 set to the estimated minor allele fraction of the first window. If the criteria are satisfied, W
375 becomes the seed of the next sub-region, and the process is repeated. Otherwise, HCS position
376 $k_3 + 1$ is added the first sub-region and W is shifted to start at $k_3 + 2$, etc.

377 In the case of a cancer-derived RNA profile, the expressed allele fractions are, in general, gene
378 specific. Therefore the steps 1 and 2 (condition b), as described above, are performed separately
379 for each gene, assuming the minor allele fraction to be constant throughout the gene. Step 3 is
380 then reduced to an empirical estimate of the minor allele fraction using read counts from all HCS
381 positions within the gene.

382 Down-sampling of sequence data

383 In order to down-sample the sequence data to a desired fraction f of the original coverage, we sam-
384 pled reads from the original patient profile P with the Bernoulli probability f without replacement.
385 The ancestry inference procedure was then performed with the resulting sample of reads.

386 Schematic overviews and figures

387 All schematic overviews have been generated with draw.io version 15.7.3 (<http://www.diagrams.net>).

388 The Venn diagrams in *Figure 2* have been generated with CRAN packages VennDiagram version
389 1.6.20 (*Chen, 2018*) and multipanelfigure version 2.1.2 (*Graumann and Cotton, 2018*).

390 The bar plot graph in *Figure 2* has been generated with CRAN package ggplot2 version 3.3.5
391 (*Wickham, 2016*).

392 The AUROC graphs in *Figure 4* have been generated with CRAN packages ggplot2 version 3.3.5
393 (*Wickham, 2016*) and cowplot version 1.1.1 (*Wilke, 2020*).

394 Acknowledgments

395 DAT is a distinguished scholar of the Lustgarten Foundation and Director of the Lustgarten Foundation-
396 designated Laboratory of Pancreatic Cancer Research. DAT is also supported by the Cold Spring
397 Harbor Laboratory Association, the New York Genome Center Polyethnic 1000 Project, the V Foun-
398 dation (T2016-010), the Thompson Foundation, the Simons Foundation (552716), and the NIH

(P30CA45508, P20CA192996, U10CA180944, U01CA224013, U01CA210240, R01CA188134, R01CA249002, and R01CA229699). AK's work was supported by the New York Genome Center Polyethnic-1000 Major Grant, Simons Foundation award # 519054, Lustgarten Foundation OPT2 project award and by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. This work was performed with assistance from the US National Institutes of Health Grant S10OD028632-01. The results published here are in part based upon data generated by TCGA Research Network: <https://www.cancer.gov/tcga>. We thank Adam Siepel, Lloyd Trotman, Jeffrey Boyd, W. Richard McCombie, Thomas Gingeras, Justin Kinney, Camila dos Santos, Michael Schatz, Louis Staudt, Michael Berger, David Solit and Samuel Aparicio for illuminating discussions.

Authors' contributions

PB and AK conceived the study. PB performed the data analysis with support of AK and AD. PB and AK wrote the manuscript with contributions from AD. PB and AD generated the figures. AK and DAT supervised the work and secured funding. All authors reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

- Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009; 19(9):1655–64. <https://www.ncbi.nlm.nih.gov/pubmed/19648217>, doi: 10.1101/gr.094052.109.
- Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, De La Vega FM, Donnelly P, Egholm M, Flicek P, Gabriel SB, Gibbs RA, Knoppers BM, Lander ES, Lehrach H, Mardis ER, McVean GA, Nickerson DA, Peltonen L, et al. A map of human genome variation from population-scale sequencing. *Nature.* 2010; 467(7319):1061–1073. doi: 10.1038/nature09534.
- Ashktorab H, Kupfer SS, Brim H, Carethers JM. Racial Disparity in Gastrointestinal Cancer Risk. *Gastroenterology.* 2017; 153(4):910–923. <https://www.ncbi.nlm.nih.gov/pubmed/28807841>, doi: 10.1053/j.gastro.2017.08.018.
- Barral-Arca R, Pardo-Seco J, Bello X, Martinon-Torres F, Salas A. Ancestry patterns inferred from massive RNA-seq data. *RNA.* 2019; 25(7):857–868. <https://www.ncbi.nlm.nih.gov/pubmed/31010885>, doi: 10.1261/rna.070052.118.
- Bhatnagar B, Kohlschmidt J, Mrozek K, Zhao Q, Fisher JL, Nicolet D, Walker CJ, Mims AS, Oakes C, Giacomelli B, Orwick S, Boateng I, Blachly JS, Maharry SE, Carroll AJ, Powell BL, Kolitz JE, Stone RM, Byrd JC, Paskett ED, et al. Poor Survival and Differential Impact of Genetic Features of Black Patients with Acute Myeloid Leukemia. *Cancer Discov.* 2021; 11(3):626–637. <https://www.ncbi.nlm.nih.gov/pubmed/33277314>, doi: 10.1158/2159-8290.CD-20-1579.
- Carrot-Zhang J, Chambwe N, Damrauer JS, Knijnenburg TA, Robertson AG, Yau C, Zhou W, Berger AC, Huang KL, Newberg JY, Mashl RJ, Romanel A, Sayaman RW, Demichelis F, Felau I, Frampton GM, Han S, Hoadley KA, Kemal A, Laird PW, et al. Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell.* 2020; 37(5):639–654 e6. <https://www.ncbi.nlm.nih.gov/pubmed/32396860>, doi: 10.1016/j.ccell.2020.04.012.
- Carrot-Zhang J, Soca-Chafre G, Patterson N, Thorner AR, Nag A, Watson J, Genovese G, Rodriguez J, Gelbard MK, Corrales-Rodriguez L, Mitsuishi Y, Ha G, Campbell JD, Oxnard GR, Arrieta O, Cardona AF, Gusev A, Meyerson M. Genetic Ancestry Contributes to Somatic Mutations in Lung Cancers from Admixed Latin American Populations. *Cancer Discov.* 2021; 11(3):591–598. <https://www.ncbi.nlm.nih.gov/pubmed/33268447>, doi: 10.1158/2159-8290.CD-20-1165.
- Chen H. VennDiagram: Generate High-Resolution Venn and Euler Plots; 2018, <https://CRAN.R-project.org/package=VennDiagram>, r package version 1.6.20.
- Cronin KA, Lake AJ, Scott S, Sherman RL, Noone AM, Howlader N, Henley SJ, Anderson RN, Firth AU, Ma J, Kohler BA, Jemal A. Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. *Cancer.* 2018; 124(13):2785–2800. <https://www.ncbi.nlm.nih.gov/pubmed/29786848>, doi: 10.1002/cncr.31551.

448 **DeLong ER**, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver
449 Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. 1988 jan; 44(3):837–845. <http://www.jstor.org/stable/2531595>, doi: 10.2307/2531595.

451 **Diaz-Papkovich A**, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and
452 phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*. 2019; 15(11):1–24. doi: 10.1371/jour-
453 nal.pgen.1008432.

454 **Dutil J**, Chen Z, Monteiro AN, Teer JK, Eschrich SA. An Interactive Resource to Probe Genetic Diversity and
455 Estimated Ancestry in Cancer Cell Lines. *Cancer Res*. 2019; 79(7):1263–1273. <https://www.ncbi.nlm.nih.gov/pubmed/30894373>, doi: 10.1158/0008-5472.CAN-18-2747.

457 **Fairley S**, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of
458 open human genomic variation resources. *Nucleic Acids Research*. 2019 10; 48(D1):D941–D947. <https://doi.org/10.1093/nar/gkz836>, doi: 10.1093/nar/gkz836.

460 **Frampton GM**, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, Schnall-Levin M, White J, Sanford EM, An P,
461 Sun J, Juhn F, Brennan K, Iwanik K, Maillet A, Buell J, White E, Zhao M, Balasubramanian S, Terzic S, et al.
462 Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA
463 sequencing. *Nat Biotechnol*. 2013; 31(11):1023–31. <https://www.ncbi.nlm.nih.gov/pubmed/24142049>, doi:
464 10.1038/nbt.2696.

465 **Graumann J**, Cotton R. multipanelfigure: Simple Assembly of Multiple Plots and Images into a Compound
466 Figure. *Journal of Statistical Software, Code Snippets*. 2018; 84(3):1–10. doi: 10.18637/jss.v084.c03.

467 **Grossman RL**, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a Shared Vision for Can-
468 cer Genomic Data. *The New England journal of medicine*. 2016 sep; 375(12):1109–1112. <https://pubmed.ncbi.nlm.nih.gov/27653561><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6309165/>, doi: 10.1056/NEJMp1607591.

470 **Hand DJ**, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification
471 Problems. *Machine Learning*. 2001; 45(2):171–186. doi: 10.1023/A:1010920819831.

472 **Huang BZ**, Stram DO, Le Marchand L, Haiman CA, Wilkens LR, Pandol SJ, Zhang ZF, Monroe KR, Setiawan VW.
473 Interethnic differences in pancreatic cancer incidence and risk factors: The Multiethnic Cohort. *Cancer Med*.
474 2019; 8(7):3592–3603. <https://www.ncbi.nlm.nih.gov/pubmed/31066497>, doi: 10.1002/cam4.2209.

475 **Huang Q**, Baudis M. Enabling population assignment from cancer genomes with SNP2pop. *Sci Rep*. 2020;
476 10(1):4846. <https://www.ncbi.nlm.nih.gov/pubmed/32179800>, doi: 10.1038/s41598-020-61854-x.

477 **INC FM**, FoundationOne Liquid CDx; 1999. [https://assets.ctfassets.net/w98cd481qyp0/](https://assets.ctfassets.net/w98cd481qyp0/YqqKHaqQmFec5ueQk48w/c35460768c3a76ef738dcf88f8219524/F1CDx_Tech_Specs_072021.pdf)
478 [YqqKHaqQmFec5ueQk48w/c35460768c3a76ef738dcf88f8219524/F1CDx_Tech_Specs_072021.pdf](https://assets.ctfassets.net/w98cd481qyp0/YqqKHaqQmFec5ueQk48w/c35460768c3a76ef738dcf88f8219524/F1CDx_Tech_Specs_072021.pdf).

479 **Kessler MD**, Bateman NW, Conrads TP, Maxwell GL, Dunning Hotopp JC, O'Connor TD. Ancestral characteriza-
480 tion of 1018 cancer cell lines highlights disparities and reveals gene expression and mutational differences.
481 *Cancer*. 2019; 125(12):2076–2088. <https://www.ncbi.nlm.nih.gov/pubmed/30865299>, doi: 10.1002/cncr.32020.

482 **Koboldt DC**, Larson DE, Wilson RK. Using VarScan 2 for Germline Variant Calling and Somatic Mutation De-
483 tection. *Current Protocols in Bioinformatics*. 2013 dec; 44(1). [https://onlinelibrary.wiley.com/doi/10.1002/](https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1504s44)
484 [0471250953.bi1504s44](https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi1504s44), doi: 10.1002/0471250953.bi1504s44.

485 **Lowy-Gallego E**, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P. Variant calling on the GRCh38 as-
486 sembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Research*. 2019 dec;
487 4:50. <https://pubmed.ncbi.nlm.nih.gov/32175479><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7059836/>
488 <https://wellcomeopenresearch.org/articles/4-50/v2>, doi: 10.12688/wellcomeopenres.15126.2.

489 **Mahal BA**, Alshalalfa M, Kensler KH, Chowdhury-Paulino I, Kantoff P, Mucci LA, Schaeffer EM, Spratt D, Yamoah
490 K, Nguyen PL, Rebbeck TR. Racial Differences in Genomic Profiling of Prostate Cancer. *N Engl J Med*. 2020;
491 383(11):1083–1085. <https://www.ncbi.nlm.nih.gov/pubmed/32905685>, doi: 10.1056/NEJMc2000069.

492 **Mersha TB**, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on
493 understanding health disparities. *Human Genomics*. 2015; 9(1):1. <https://doi.org/10.1186/s40246-014-0023-x>,
494 doi: 10.1186/s40246-014-0023-x.

495 **NCI**, National Cancer Institute, Genomic Data Commons; 2021. <https://gdc.cancer.gov/>.

496 **Network TR**, TCGA Research Network; 2021. [https://www.cancer.gov/about-nci/organization/ccg/research/](https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga)
497 [structural-genomics/tcga](https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga).

498 **Nugent A**, Conatser KR, Turner LL, Nugent JT, Sarino EMB, Ricks-Santi LJ. Reporting of race in genome and
499 exome sequencing studies of cancer: a scoping review of the literature. *Genet Med*. 2019; 21(12):2676–2680.
500 <https://www.ncbi.nlm.nih.gov/pubmed/31160752>, doi: 10.1038/s41436-019-0558-2.

501 **Polubriagino FCG**, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, Hripcsak G, Smith S, Tatonetti
502 NP, Vawdrey DK. Challenges with quality of race and ethnicity data in observational databases. *Journal of*
503 *the American Medical Informatics Association*. 2019 aug; 26(8-9):730–736. [https://academic.oup.com/jamia/](https://academic.oup.com/jamia/article/26/8-9/730/5542028)
504 [article/26/8-9/730/5542028](https://academic.oup.com/jamia/article/26/8-9/730/5542028), doi: 10.1093/jamia/ocz113.

505 **Price AL**, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects
506 for stratification in genome-wide association studies. *Nat Genet*. 2006; 38(8):904–9. [https://www.ncbi.nlm.](https://www.ncbi.nlm.nih.gov/pubmed/16862161)
507 [nih.gov/pubmed/16862161](https://www.ncbi.nlm.nih.gov/pubmed/16862161), doi: 10.1038/ng1847.

508 **Pritchard JK**, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Ge-*
509 *netics*. 2000; 155(2):945–59. <https://www.ncbi.nlm.nih.gov/pubmed/10835412>.

510 **Robin X**, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and
511 S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011 dec; 12(1):77. [https://bmcbioinformatics.](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-77)
512 [biomedcentral.com/articles/10.1186/1471-2105-12-77](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-77), doi: 10.1186/1471-2105-12-77.

513 **Siegel RL**, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020; 70(1):7–30. [https://www.ncbi.nlm.](https://www.ncbi.nlm.nih.gov/pubmed/31912902)
514 [nih.gov/pubmed/31912902](https://www.ncbi.nlm.nih.gov/pubmed/31912902), doi: 10.3322/caac.21590.

515 **Sinha S**, Mitchell KA, Zingone A, Bowman E, Sinha N, Schäffer AA, Lee JS, Ruppini E, Ryan BM. Higher prevalence
516 of homologous recombination deficiency in tumors from African Americans versus European Americans.
517 *Nature Cancer*. 2020; 1(1):112–121. <https://doi.org/10.1038/s43018-019-0009-7>, doi: 10.1038/s43018-019-
518 0009-7.

519 **Sun X**, Xu W. Fast Implementation of DeLong’s Algorithm for Comparing the Areas Under Correlated Re-
520 ceiver Operating Characteristic Curves. *IEEE Signal Processing Letters*. 2014 nov; 21(11):1389–1393. [http:](http://ieeexplore.ieee.org/document/6851192/)
521 [//ieeexplore.ieee.org/document/6851192/](http://ieeexplore.ieee.org/document/6851192/), doi: 10.1109/LSP.2014.2337313.

522 **Tan DS**, Mok TS, Rebbeck TR. Cancer Genomics: Diversity and Disparity Across Ethnicity and Geography. *J Clin*
523 *Oncol*. 2016; 34(1):91–101. <https://www.ncbi.nlm.nih.gov/pubmed/26578615>, doi: 10.1200/JCO.2015.62.0096.

524 **Tate JG**, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E,
525 Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, et al.
526 COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*. 2018 10; 47(D1):D941–D947.
527 <https://doi.org/10.1093/nar/gky1015>, doi: 10.1093/nar/gky1015.

528 **The Cancer Genome Atlas Research Network**. Integrated genomic analyses of ovarian carcinoma. *Nature*.
529 2011 jun; 474(7353):609–615. <http://www.nature.com/articles/nature10166>, doi: 10.1038/nature10166.

530 **Tiriac H**, Belleau P, Engle DD, Plenker D, Deschênes A, Somerville TDD, Froeling FEM, Burkhart RA, Denroche
531 RE, Jang GH, Miyabayashi K, Young CM, Patel H, Ma M, LaComb JF, Palmira RLD, Javed AA, Huynh JC, John-
532 son M, Arora K, et al. Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic
533 Cancer. *Cancer Discovery*. 2018 sep; 8(9):1112–1129. <https://www.ncbi.nlm.nih.gov/pubmed/29853643>
534 <https://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-18-0349>, doi: 10.1158/2159-8290.CD-18-
535 0349.

536 **Tyner JW**, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, Long N, Schultz AR, Traer E, Abel M, Agarwal A,
537 Blucher A, Borate U, Bryant J, Burke R, Carlos A, Carpenter R, Carroll J, Chang BH, Coblenz C, et al. Functional
538 genomic landscape of acute myeloid leukaemia. *Nature*. 2018; 562(7728):526–531. [https://www.ncbi.nlm.nih.](https://www.ncbi.nlm.nih.gov/pubmed/30333627)
539 [gov/pubmed/30333627](https://www.ncbi.nlm.nih.gov/pubmed/30333627), doi: 10.1038/s41586-018-0623-z.

540 **Wickham H**. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. [https://ggplot2.](https://ggplot2.tidyverse.org)
541 [tidyverse.org](https://ggplot2.tidyverse.org).

542 **Wilke CO**. cowplot: Streamlined Plot Theme and Plot Annotations for ‘ggplot2’; 2020, [https://CRAN.R-project.](https://CRAN.R-project.org/package=cowplot)
543 [org/package=cowplot](https://CRAN.R-project.org/package=cowplot), r package version 1.1.1.

544 **Yuan J**, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, Hu X, Zhang Y, Wang Y, Jiang J, Li C, Zhong X, Montone KT,
545 Guan G, Tanyi JL, Fan Y, Xu X, Morgan MA, Long M, Zhang Y, et al. Integrated Analysis of Genetic Ancestry
546 and Genomic Alterations across Cancers. *Cancer Cell*. 2018; 34(4):549–560 e9. [https://www.ncbi.nlm.nih.gov/](https://www.ncbi.nlm.nih.gov/pubmed/30300578)
547 [pubmed/30300578](https://www.ncbi.nlm.nih.gov/pubmed/30300578), doi: 10.1016/j.ccell.2018.08.019.

- 548 **Zhang J**, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, Stein LD, Ferretti V. The International Cancer
549 Genome Consortium Data Portal. *Nature Biotechnology*. 2019 apr; 37(4):367–369. [http://www.nature.com/](http://www.nature.com/articles/s41587-019-0055-9)
550 [articles/s41587-019-0055-9](http://www.nature.com/articles/s41587-019-0055-9), doi: 10.1038/s41587-019-0055-9.
- 551 **Zheng X**, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for
552 relatedness and principal component analysis of SNP data. *Bioinformatics*. 2012 10; 28(24):3326–3328.
553 <https://doi.org/10.1093/bioinformatics/bts606>, doi: 10.1093/bioinformatics/bts606.