Genetic ancestry inference from cancer-derived molecular data across genomic and transcriptomic platforms

Pascal Belleau[1,2], Astrid Deschênes[2,3], Nyasha Chambwe[4], David A. Tuveson[2,3], and Alexander Krasnitz[1,2]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; [2]Cancer Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; [3]Lustgarten Foundation Pancreatic Cancer Research Laboratory, Cold Spring Harbor, New York, USA; [4]Institute of Molecular Medicine, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, New York, USA

Corresponding Author: Alexander Krasnitz, Simons Center for Quantitative Biology, Cold Spring Habor Laboratory, Cold Spring Harbor, New York, USA, 11724; E-mail: krasnitz@cshl.edu; Telephone: 516-3676863

Running Title

Ancestry inference from cancer-derived molecular data

Keywords

GENETICS OF CANCER RISK & OUTCOME/GENETICS OF CANCER RISK & OUTCOME, COMPUTATIONAL METHODS/Algorithms, COMPUTATIONAL METHODS/Sequence analysis, COMPUTATIONAL METHODS/Software, genetic ancestry

Conflicts of interest

The authors declare that they have no competing interests.

Word count

5042

Total number of figures

5

Total number of tables

2

Significance statement

The development of a computational approach that enables accurate and robust ancestry inference from cancer-derived molecular profiles without matching cancer-free data provides a valuable methodology for genetic ancestry-oriented cancer research.

Abstract Genetic ancestry-oriented cancer research requires the ability to perform accurate and robust genetic ancestry inference from existing cancer-derived data, including whole exome sequencing, transcriptome sequencing, and targeted gene panels, very often in the absence of matching cancer-free genomic data. Here we examined the feasibility and accuracy of computational inference of genetic ancestry relying exclusively on cancer-derived data. A data synthesis framework was developed to optimize and assess the performance of the ancestry inference for any given input cancer-derived molecular profile. In its core procedure, the ancestral background of the profiled patient is replaced with one of any number of individuals with known ancestry. The data synthesis framework is applicable to multiple profiling platforms, making it possible to assess the performance of inference specifically for a given molecular profile and separately for each continental-level ancestry; this ability extends to all ancestries, including those without statistically sufficient representation in the existing cancer data. The inference procedure was demonstrated to be accurate and robust in a wide range of sequencing depths. Testing of the approach in four representative cancer types and across three molecular profiling modalities showed that continental-level ancestry of patients can be inferred with high accuracy, as quantified by its agreement with the gold standard of deriving ancestry from matching cancer-free molecular data. This study demonstrates that vast amounts of existing cancer-derived molecular data are potentially amenable to ancestry-oriented studies of the disease without requiring matching cancer-free genomes or patient self-reported ancestry.

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 (1-5). 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 point to differences in the molecular make-up of the disease among groups of different ancestral background and to the need for more molecular data to power discovery of such differences (6-11).

Ancestry annotation of cancer-derived data largely draws on two sources: patient's self-identified race and/or ethnicity (SIRE) and patient's cancer-free genotype. SIRE is often missing, sometimes inaccurate and usually incomplete. As a recent analysis (12) of PubMed database entries since 2010 reveals, patients' SIRE is massively under-reported in genome and exome sequencing studies of cancer, with only 37% of these reporting race, and 17% reporting ethnicity. Furthermore, SIRE is not always consistent with genetic ancestry. Finally, a self-declaring patient is often given a choice from a small number of broad racial or ethnic categories, which fail to capture complete ancestral information, especially in cases of mixed ancestry (13).

A far more accurate and detailed ancestral characterization may be obtained by genotyping a patient's DNA from a cancer-free tissue. Powerful methods exist for ancestry inference from germline DNA sequence (14-17). These methods were recently used to determine ancestry of approximately 10,000 patients profiled by The Cancer Genome Atlas (TCGA) (7,11). However, genotyping of DNA from patient-matched cancer-free specimens is not part of standard clinical practice, where the purpose of DNA profiling is often identification of mutations with known oncogenic effects, such as those in the Catalog Of Somatic Mutations In Cancer (COSMIC) database (18). As a result, it is not performed routinely outside academic clinical centers or major research projects. There also are studies yielding sequence data from tumors, whose purpose does not require germline profiling. RNA sequencing (RNA-seq) for expression quantification is in this category. Finally, peripheral blood is most often the source of germline DNA in the clinic, but this is not always the case for diseases of the hematopoietic system, such as leukemia, wherein cancer cells are massively present in circulation. In summary, matched germline DNA sequence is not universally available for cancer-derived molecular data. In such cases, it is necessary to infer ancestry from the nucleic acid sequence of the tumor itself.

Standard methods of ancestry inference commonly rely on population specificity of germline single-nucleotide variants (SNV). Whole-genome (WGS) or whole-exome sequences (WES), at depths sufficient for reliably calling single-nucleotide variants, and readouts from genotyping microarrays, are therefore data types most suitable for this purpose. However, such detailed DNA profiling is often not performed in molecular studies of cancer. In such cases, it is necessary to infer ancestry from other types of tumor-derived data, including RNA sequence and DNA sequence for a small panel of genes, e.g., FoundationOne® CDx (19).

For all types of tumor-derived sequence, accurate inference of ancestry is a potential challenge. Tumor genome is often replete with somatic alterations, including loss of heterozygosity (LOH), copy number variants (CNV), translocations, microsatellite instabilities and SNV. These alterations interfere with germline genotyping of the patient that is used as input for inference of genetic ancestry. Structural variants, especially LOH and CNV, are the most likely to affect the germline genotyping, and thereby the genetic ancestry calls. This effect is especially clearly seen in the case of LOH, as a result of which heterozygous genotypes are transformed into homozygous, but other types of alterations also are, to various degrees, potential obstacles to accurate ancestry inference. Tumor RNA-seq presents additional challenges, namely, extremely uneven coverage of the transcript due to a broad range of RNA expression levels and distortions due to allele-specific expression. Gene panels represent a very small fraction of the genome, whose sufficiency for ancestry inference is not clear and may vary from panel to panel. In addition, cancer gene panels are enriched in cancer driver genes, which tend to undergo somatic alteration more frequently than other parts of the genome.

Important recent publications on ancestral effects in cancer reported patient ancestry inferred

from matching cancer-free DNA (7,8,11). At the same time, there has been much less work on ancestry inference from tumor-derived nucleic acids (7,11,20-23). Collectively, this work demonstrates the feasibility of accurate genetic ancestry inference from cancer-derived DNA profiled by SNP arrays or by high-coverage gene panels, such as the FoundationOne® CDx gene panel (19). However, to our knowledge, no systematic computational framework for ancestry inference from cancer-derived molecular data, across assay and cancer types, has been developed to date. There is presently no ability to assess the inference accuracy specifically for a given input tumor-derived molecular profile with all its attendant properties, including the data quality and the depth of coverage. Reliable and accurate ancestry inference from tumor-derived nucleic acids thus represents an unmet need, which the present work aims to address.

For this purpose, we designed an inference procedure having in mind a scenario, likely to occur in studies of existing data or of archived tissue specimens, with an input molecular profile of a tumor from a single patient, and no matching cancer-free sequence available. The profile in question may have its unique set of sequence properties. These include the target sequence and uniformity of its coverage depth, read length and sequencing quality. These profile-specific properties may be vastly dissimilar from those in the available public data sets with reliably known genetic ancestry of the patients. Furthermore, not all ancestries are equally easy to infer: for example, an American ancestral category is sometimes difficult to distinguish either from African or from European ancestry. This profile specificity would make it impossible to confidently assess the accuracy of the inference procedure for the input profile from its performance with the public cancer-derived data in aggregate. In order to overcome this difficulty, we developed a computational technique, which is described schematically in ***Figure 1***, wherein the ancestral background of the patient is supplanted in the input profile by one of an unrelated individual with known ancestry. A similar data synthesis procedure was employed in our prior work in a different genomic context (24). We next apply established methods of ancestry inference to this synthetic profile and compare the result to that known ancestry. Generating multiple such synthetic profiles allows us to assess how accurate the ancestry inference is for the patient, both overall and as a function of the profile's continental-level ancestry. Furthermore, using synthetic data, we are able to optimize the inference procedure with respect to parameters on which it depends. Importantly, this assessment and optimization procedure does not require the profile in question to be part of a larger data set from a cohort of patients with a similar diagnosis. Very often in existing cancer-derived data, such cohorts do not provide statistically meaningful representation of non-European ancestries. This insufficiency is not an impediment to the application of our methodology.

In the following, we assess the accuracy of global ancestry calls from tumor exomes, narrowly targeted gene panels and RNA sequences, in comparison to such calls from matching germline genotypes, as profiled by exome sequencing or genotyping microarrays. We do so for four cancer types, namely, pancreatic adenocarcinoma (PDAC), ovarian cystadenocarcinoma (OV) and breast carcinoma (BRCA) as representative types of epithelial tumors, and acute myeloid leukemia (AML), as an example of hematopoietic malignancy. Each of these data sets was chosen because it presents a challenge for patients' ancestry inference and/or an opportunity to test our approach. Specifically, OV is characterized by massive copy number alterations, often spanning much of the genome. Our PDAC data originate from patient-derived organoid (PDO) models of the disease (25). In PDO, near-100% tumor purity is achieved, exacerbating effects of copy number loss and loss of heterozygosity on the sequence. In BRCA, a large patient cohort size makes it possible for us to choose an ancestrally diverse subset of the data for testing our methods. In AML the peripheral blood, the usual source of cancer-free DNA, may be severely contaminated by the cancer.

Methods and Materials

Data sets and pre-processing

The data sets used in this work originate from four sources: TCGA collection for ovarian cystadenocarcinoma (26) (TCGA-OV), an ancestrally diverse subset of TCGA collection for breast carcinoma (27) (TCGA-BRCA), Beat AML clinical trial (28) (Beat AML), and a study of pancreatic ductal adenocarcinoma using patient-derived organoids (25) (PDAC). For all four, the data used are

summarized in the form of Venn diagrams in ***Figure 2***A-D and tabulated in Supplementary Table S1. These data include cancer DNA (whole-exome or whole-genome) sequence, cancer RNA sequence and matching normal DNA (whole-exome or whole-genome) sequence. As explained in the following, genetic ancestry inferred from the latter was used as the ground truth in assessing the performance of ancestry inference from the cancer-derived data cohort-wide for each of the four cohorts. Also available for comparison was the donor SIRE, as depicted in ***Figure 2***E. In addition, published genetic ancestry calls from matching cancer-free genotypes, representing a consensus of five inference pipelines (C5), were available for comparison with our findings for the TCGA-OV and TCGA-BRCA cohorts (7).

Throughout the study, we used the 1000 Genomes (1KG) data set, with no relatives for the individuals included (29-31), as reference, against which patient molecular data were compared to infer continental-level global ancestry. The latter is defined as a categorical variable taking five values: African (AFR), East Asian (EAS), European (EUR), American (AMR) and South Asian (SAS). These are called super-populations in the 1KG terminology. Each super-population comprises a number of subcontinental-level populations, as explained in the 1000 Genomes consortium publications (31). The composition of the 1KG data, as used in this study, is summarized in Supplementary Table S2.

In all cases, read data mapped to the hg38 version of the human genome were used. In order to study ancestry inference from targeted panels, the cancer-derived whole-exome data were reduced to reads mapping to the FoundationOne® CDx cancer-related gene panel (19). The pre-processing is illustrated in the first part of the ***Figure 3***. Reads in the cancer patient-derived data were filtered for quality using a cutoff phred score of 20. Following this filter, single-nucleotide substitutions were called at all positions with read coverage of at least 10, using snp-pileup in FACETS (32) and Varscan version 2.4.4 (33). This set of positions is called the high-confidence substitution (HCS) set in the following. From the 1000 Genomes (1KG) variant call data in the Variant Call Format (VCF) (34), genomic positions where substitution variants occur at a frequency of at least 0.01 in at least one of the super-populations comprising 1KG were selected as a basis for the ancestry inference. This set is referred to as the high-frequency substitution (HFS) set in the following. The genotype was called at the HFS positions in the cancer-derived profile with the coverage above 10. This subset of the HFS 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 (35), 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, TCGA-OV and TCGA-BRCA data. For all three, we infer the ancestry from the matching cancer-free exome profiles. In the case of TCGA-OV and TCGA-BRCA data, we also compare the results to the consensus ancestry calls (7).

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 except TCGA-BRCA 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) (36). 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) (37) for this purpose, and compute both the class-specific AUROC for each super-population and the 5-class overall AUROC. In the class-specific case, we use a version DeLong's algorithm (38,39) as implemented in the pROC package to compute the AUROC confidence intervals. In the overall 5-class case the confidence intervals are computed using bootstrap with 100-fold sampling.

Data synthesis

Data synthesis is defined here as replacement of PHCG genotypes 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, as explained in detail in the Supplementary Methods and illustrated in Figure S1; and (b) the haplotype of U in the portion of the genome covered by P. With this knowledge, the procedure, depicted in ***Figure 4***, 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 PHCG substitution in the genome of U within the scope of P.

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 5***).

Down-sampling of sequence data

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

Software used in making figures

All diagrams were made using draw.io version 15.7.3 (http://www.diagrams.net). The Venn diagrams in ***Figure 2*** were produced with CRAN packages VennDiagram version 1.7.3 (40) and multipanelfigure version 2.1.2 (41). The bar plot in ***Figure 2*** and the plots in ***Figure 5*** were made using packages ggplot2 (version 3.3.6, RRID: SCR_014601) and cowplot (version 1.1.1, RRID: SCR_018081).

Software and data availability

Ancestry inference methods introduced in this work are implemented in an R language package RAIDS (Robust Ancestry Inference using Data Synthesis) is publicly available, under the Apache-2.0

license, at https://github.com/KrasnitzLab/RAIDS. Documentation for this software is available at https://krasnitzlab.github.io/RAIDS/. The data analyzed in this study were obtained from the National Center for Biotechnology (NCBI) database of Genotypes and Phenotypes (dbGaP) archive under accession numbers phs001611.v1.p1, phs001657.v1.p1 and phs000178.v11.p8.

Results

We assessed the performance of genetic ancestry inference from three genomic data types: whole exomes, gene panels targeting exomes of several hundred cancer-related genes each and RNA sequences. Our assessment relied on molecular data collected from four patient cohorts, each representing a cancer type, namely, tissue donors to the Cold Spring Harbor Laboratory (CSHL) pancreatic ductal adenocarcinoma (PDAC) library of patient-derived organoids; acute myeloid leukemia (AML) patients enrolled in Beat AML clinical trial; patients comprising TCGA ovarian cancer cohort (TCGA-OV) (26) and a subset of TCGA breast cancer cohort (TCGA-BRCA). Throughout the study we used the 1000 Genomes (1KG) genotype collection as our population reference.

As explained in detail in the Methods and Materials section, for inference of genetic ancestry we employed principal-component analysis (PCA) in combination with K-nearest-neighbor classification. For a subset of patients in each cohort we individually assessed the performance of the ancestry inference, as a function of the parameters K and D, the number of principal dimensions retained. We relied on data synthesis for this assessment. Both super-population-specific and overall AUROC values were computed in a range of D, K pairs, as illustrated in ***Figure 5*** for 10 PDAC patients and AMR-specific AUROC and in Figure S2 for all other cohorts and super-populations. Optimal D, K pairs maximizing the overall AUROC were chosen. From this subset of patients we observed, for each cancer type considered and for each of the three molecular profiling modalities, an optimal range of D and K parameters where the performance of inference was consistently high in the subset and only weakly dependent on these parameters (Figure S2). For all four tumor types, our overall performance findings using data synthesis are summarized in Tables S3-S6. We then selected and used, for the remainder of the patients with this cancer type and for this profiling modality, a pair D and K values from within the optimal range. As an additional validation of our parameter optimization procedure, we applied it to cancer-free WES profiles of TCGA-OV and TCGA-BRCA patients included in this study. Comparing the resulting ancestry calls to the consensus calls (C5) by TCGA (7), we find the two to be in good agreement (Tables S7-S10).

We also assessed the cohort-wide performance of our ancestry calls from the original cancer-derived molecular data, by comparison to the gold standard of ancestry as determined from the matching cancer-free genotypes. For Beat AML, TCGA-OV and TCGA-BRCA patients, we performed ancestry inference 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 four cohorts, we summarize our cohort-wide findings in ***Table 1***. We also used the C5 ancestry calls (7) in our performance assessment for TCGA-OV and TCGA-BRCA and found close agreement for both these cohorts (Tables S7-S10).

We note that in all patient cohorts we analyze here except TCGA-BRCA (***Table 2*** and Table S11) the sampling of patients with non-European ancestries is statistically insufficient for a purely cohort-based assessment of performance (Table S12-S14). We therefore report cohort-wide overall but not super-population specific AUROC values for Beat AML, TCGA-OV and TCGA-BRCA. 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 S15-S18 and Figure S2). We do report super-population-specific AUROC for TCGA-BRCA and for the aggregate of all four cohorts.

The results of our analysis as presented in Tables S15-S18, 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 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 S2 B,G,L). Full results of our inferential analysis for the patients in all four cohorts are compiled in Table S19.

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 TCGA-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 S3) 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 (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 (42,43)).

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 (44) and International Cancer Genome Consortium (ICGC) (45) projects. For these, the genetic ancestry of the patients either is known or can be readily established using matching cancer-free molecular data. Second, an extension of our approach to additional profiling modalities should be examined. Chief among these are low-coverage whole-genome sequences commonly used for copy-number analysis, single-molecule, long-read sequences, chromatin-accessibility profiles (ATAC-seq) and cytosine-converted sequences used for methylation profiling. Each of these presents unique challenges and opportunities for the ancestry inference. For example, in the low-coverage whole-genome profiles the sparsity of coverage is compensated by its whole-genome breadth, whereas in the long-read sequences the trade-off is between the high sequence error rate and the long-distance phasing afforded by the read length. Third, while the present work relied on PCA followed by nearest-neighbor classification for ancestry assessment, alternatives including UMAP for the former and Random Forest or Support Vector Machine for the latter exist and should be evaluated. Third, future method development should be extended beyond inference of global ancestry to that of local ancestry and ancestral admixture. Such an extension is particularly important in the study of cancer in strongly admixed super-populations, such as AFR and AMR, and may require more extensive reference data, in addition to the 1KG reference used here. Finally, beyond cancer, our methodology can be applied to any molecular data from which ancestry inference is challenging. Examples include RNA-seq of non-cancer origin and sequences originating in any kind of fragmentary or damaged nucleic-acid specimens, such as those encountered in forensic, archaeological or paleontological contexts.

We anticipate the computational approach described here to have a major, two-fold, impact on investigation of links between ancestry and cancer. First, it will become possible to massively boost the statistical power of such studies by leveraging existing tumor-derived molecular data sets without

matching germline sequences or ancestry annotation. Our search of the Gene Expression Omnibus (GEO) database alone has identified over 1,250 such data sets, containing RNA expression data for nearly 48,000 cancer tissue specimens. Such resources dwarf those of fully annotated repositories, such as TCGA (44) and ICGC (45). Other molecular data repositories are likely to contain resources of this category on a similar order of magnitude. Second, hundreds of thousands of tumor tissue specimens stored at multiple clinical centers constitute another major resource for ancestry-aware molecular studies of cancer. Here again, matching normal tissue specimens are often absent, and so is ethnic or racial annotation for the patients. According to a recent estimate (46), such annotation is missing in electronic health records (EHR) of over 50% of patients. Where the donor SIRE is provided by the EHR, it can be used to guide the initial specimen collection for a study of ancestral effects in cancer, with a subsequent genetic ancestry validation using methods developed in this work. In summary, inferential tools presented here will make massive resources of archival tissues available for ancestry-oriented cancer research.

Multiple directions of exploratory and correlative analysis are open to pursuit with the accurate ancestry annotation made possible by the methods described here, even in the absence of matching cancer-free molecular data. Single-nucleotide and other small-scale somatic alterations may be identified in cancer-only exomes, both whole and restricted to specialized gene panels, using methods developed for this purpose (47) alongside databases of frequent somatic variants in cancer (18) and of frequent germline variants like gnomAD (48) and 1KG (31). Copy number variants and losses of heterozygosity in cancer exomes are overwhelmingly somatic and may be determined computationally (49,50). Cancer RNA expression quantification is feasible in the absence of the germline genotype of the patient, including alllele- and isoform-specific analysis. These and similar genomic and transcriptional properties may be explored for associations with ancestral background of the patients.

Acknowledgments

DAT is a distinguished scholar of the Lustgarten Foundation and Director of the Lustgarten Foundation-designated Laboratory of Pancreatic Cancer Research. DAT is also supported by the Cold Spring Harbor Laboratory Association, the New York Genome Center Polyethnic 1000 Project, the Simons Foundation (552716), and the NIH (P30CA45508, P20CA192996, U01CA224013, U01CA210240, R01CA188134, R01CA249002, and R01CA229699). DAT also acknowledges support from The Pershing Square Foundation, William Ackman, and Neri Oxman. AK's work is supported by the New York Genome Center Polyethnic-1000 Project, Simons Foundation award # 519054 and by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory and by the Lustgarten Foundation. 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.

Tables

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
TCGA-BRCA WES	4	9	1	NA	NA	NA
TCGA-BRCA Panel	4	9	0.995	0.984-1	0.995	0.994-0.996
TCGA-BRCA RNA-seq	4	9	0.995	0.984-1	0.995	0.994-0.996
Aggregate WES	-	-	0.993	0.981-1	0.997	0.997-0.998
Aggregate Panel	-	-	0.988	0.972-1	0.987	0.986-0.988
Aggregate RNA-seq	-	-	0.993	0.981-1	0.993	0.993-0.994

Table 1. Overall 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. The *D* and *K* values shown provide consistently high performance in each respective data set.

(a) TCGA-BRCA WES

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	47	0	0	0	0
	EUR	0	56	0	0	0
	AFR	0	0	51	0	0
	AMR	0	0	0	25	0
	SAS	0	0	0	0	4

(b) Aggregate WES

	Inferred				
pop	EAS	EUR	AFR	AMR	SAS
EAS	69	0	0	0	0
EUR	0	732	0	6	0
AFR	0	0	96	0	0
AMR	0	1	0	70	0
SAS	0	0	0	0	14

(c) TCGA-BRCA Panel

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	47	0	0	0	0
	EUR	0	56	0	0	0
	AFR	0	0	51	0	0
	AMR	0	0	0	24	1
	SAS	0	0	0	0	4

(b) Aggregate Panel

	Inferred				
pop	EAS	EUR	AFR	AMR	SAS
EAS	69	0	0	0	0
EUR	0	733	0	5	0
AFR	0	0	95	1	0
AMR	0	5	0	65	1
SAS	0	0	0	0	14

(a) TCGA-BRCA RNA

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	47	0	0	0	0
	EUR	0	56	0	0	0
	AFR	0	0	51	0	0
	AMR	0	0	0	24	1
	SAS	0	0	0	0	4

(b) Aggregate RNA

	Inferred				
pop	EAS	EUR	AFR	AMR	SAS
EAS	62	0	0	0	0
EUR	0	521	0	2	0
AFR	0	0	83	0	0
AMR	1	1	0	59	1
SAS	0	0	0	0	10

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

References

1. Ashktorab H, Kupfer SS, Brim H, Carethers JM. Racial Disparity in Gastrointestinal Cancer Risk. Gastroenterology **2017**;153:910-23
2. Cronin KA, Lake AJ, Scott S, Sherman RL, Noone AM, Howlader N*, et al.* Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. Cancer **2018**;124:2785-800
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin **2020**;70:7-30
4. Tan DS, Mok TS, Rebbeck TR. Cancer Genomics: Diversity and Disparity Across Ethnicity and Geography. J Clin Oncol **2016**;34:91-101
5. Huang BZ, Stram DO, Le Marchand L, Haiman CA, Wilkens LR, Pandol SJ*, et al.* Interethnic differences in pancreatic cancer incidence and risk factors: The Multiethnic Cohort. Cancer Med **2019**;8:3592-603
6. Bhatnagar B, Kohlschmidt J, Mrozek K, Zhao Q, Fisher JL, Nicolet D*, et al.* Poor Survival and Differential Impact of Genetic Features of Black Patients with Acute Myeloid Leukemia. Cancer Discov **2021**;11:626-37
7. Carrot-Zhang J, Chambwe N, Damrauer JS, Knijnenburg TA, Robertson AG, Yau C*, et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. Cancer Cell **2020**;37:639-54 e6
8. Carrot-Zhang J, Soca-Chafre G, Patterson N, Thorner AR, Nag A, Watson J*, et al.* Genetic Ancestry Contributes to Somatic Mutations in Lung Cancers from Admixed Latin American Populations. Cancer Discov **2021**;11:591-8
9. Mahal BA, Alshalalfa M, Kensler KH, Chowdhury-Paulino I, Kantoff P, Mucci LA*, et al.* Racial Differences in Genomic Profiling of Prostate Cancer. N Engl J Med **2020**;383:1083-5
10. Sinha S, Mitchell KA, Zingone A, Bowman E, Sinha N, Schäffer AA*, et al.* Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans. Nature Cancer **2020**;1:112-21
11. Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J*, et al.* Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. Cancer Cell **2018**;34:549-60 e9
12. Nugent A, Conatser KR, Turner LL, Nugent JT, Sarino EMB, Ricks-Santi LJ. Reporting of race in genome and exome sequencing studies of cancer: a scoping review of the literature. Genet Med **2019**;21:2676-80
13. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum Genomics **2015**;9:1
14. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res **2009**;19:1655-64
15. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLOS Genetics **2019**;15:e1008432
16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet **2006**;38:904-9
17. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics **2000**;155:945-59
18. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N*, et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res **2019**;47:D941-d7
19. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J*, et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol **2013**;31:1023-31
20. Dutil J, Chen Z, Monteiro AN, Teer JK, Eschrich SA. An Interactive Resource to Probe Genetic Diversity and Estimated Ancestry in Cancer Cell Lines. Cancer Res **2019**;79:1263-73
21. Huang Q, Baudis M. Enabling population assignment from cancer genomes with SNP2pop. Sci Rep **2020**;10:4846
22. Kessler MD, Bateman NW, Conrads TP, Maxwell GL, Dunning Hotopp JC, O'Connor TD. Ancestral characterization of 1018 cancer cell lines highlights disparities and reveals gene expression and mutational differences. Cancer **2019**;125:2076-88
23. Arora K, Tran TN, Kemel Y, Mehine M, Liu YL, Nandakumar S*, et al.* Genetic Ancestry Correlates with Somatic Differences in a Real-World Clinical Cancer Sequencing Cohort. Cancer Discov **2022**
24. Krasnitz A, Kendall J, Alexander J, Levy D, Wigler M. Early Detection of Cancer in Blood Using Single-Cell Analysis: A Proposal. Trends Mol Med **2017**;23:594-603
25. Tiriac H, Belleau P, Engle DD, Plenker D, Deschênes A, Somerville TDD*, et al.* Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer. Cancer Discov **2018**;8:1112-29
26. Integrated genomic analyses of ovarian carcinoma. Nature **2011**;474:609-15
27. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF*, et al.* Comprehensive molecular portraits of human breast tumours. Nature **2012**;490:61-70
28. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL*, et al.* Functional genomic landscape of acute myeloid leukaemia. Nature **2018**;562:526-31
29. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA*, et al.* A map of human genome variation from population-scale sequencing. Nature **2010**;467:1061-73

30. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv **2021**:2021.02.06.430068
31. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic acids research **2020**;48:D941-D7
32. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res **2016**;44:e131
33. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics **2009**;25:2283-5
34. Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. Wellcome Open Res **2019**;4:50
35. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics **2012**;28:3326-8
36. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn **2001**;45:171-86
37. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **2011**;12:77
38. Delong ER, Delong DM, Clarkepearson DI. Comparing the Areas under 2 or More Correlated Receiver Operating Characteristic Curves - a Nonparametric Approach. Biometrics **1988**;44:837-45
39. Sun X, Xu WC. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. Ieee Signal Proc Let **2014**;21:1389-93
40. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. Bmc Bioinformatics **2011**;12
41. Graumann J, Cotton R. multipanelfigure: Simple Assembly of Multiple Plots and Images into a Compound Figure. J Stat Softw **2018**;84:1-10
42. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, *et al.* Toward a Shared Vision for Cancer Genomic Data. N Engl J Med **2016**;375:1109-12
43. Heath AP, Ferretti V, Agrawal S, An M, Angelakos JC, Arya R, *et al.* The NCI Genomic Data Commons. Nat Genet **2021**;53:257-62
44. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, *et al.* Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. Cell Syst **2019**;9:24-34.e10
45. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, *et al.* The International Cancer Genome Consortium Data Portal. Nat Biotechnol **2019**;37:367-9
46. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, *et al.* Challenges with quality of race and ethnicity data in observational databases. J Am Med Inform Assoc **2019**;26:730-6
47. Sun JX, He Y, Sanford E, Montesion M, Frampton GM, Vignot S, *et al.* A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput Biol **2018**;14:e1005965
48. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. Nature **2020**;581:434-43
49. Oh S, Geistlinger L, Ramos M, Morgan M, Waldron L, Riester M. Reliable Analysis of Clinical Tumor-Only Whole-Exome Sequencing Data. JCO Clin Cancer Inform **2020**;4:321-35
50. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, *et al.* PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med **2016**;11:13

Figure Legends

Figure 1. An overview of genetic ancestry inference from cancer-derived molecular data using data synthesis.

Figure 2. Summary of the molecular data used in this study. These originate from four patient cohorts: **A)**donors to TCGA ovarian cancer collection **B)** Beat AML clinical trial **C)** pancreatic ductal adenocarcinoma patients donating to CSHL patient-derived organoid collection **D)** a subset of donors to TCGA breast cancer collection. **E)** SIRE composition for the TCGA-OV, Beat AML, PDAC and TCGA-BRCA cohorts and in aggregate over all four cohorts. UNK means not reported or unknown.

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

Figure 4. An overview of the data synthesis.

Figure 5. 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.

Figure 1, Belleau et al.

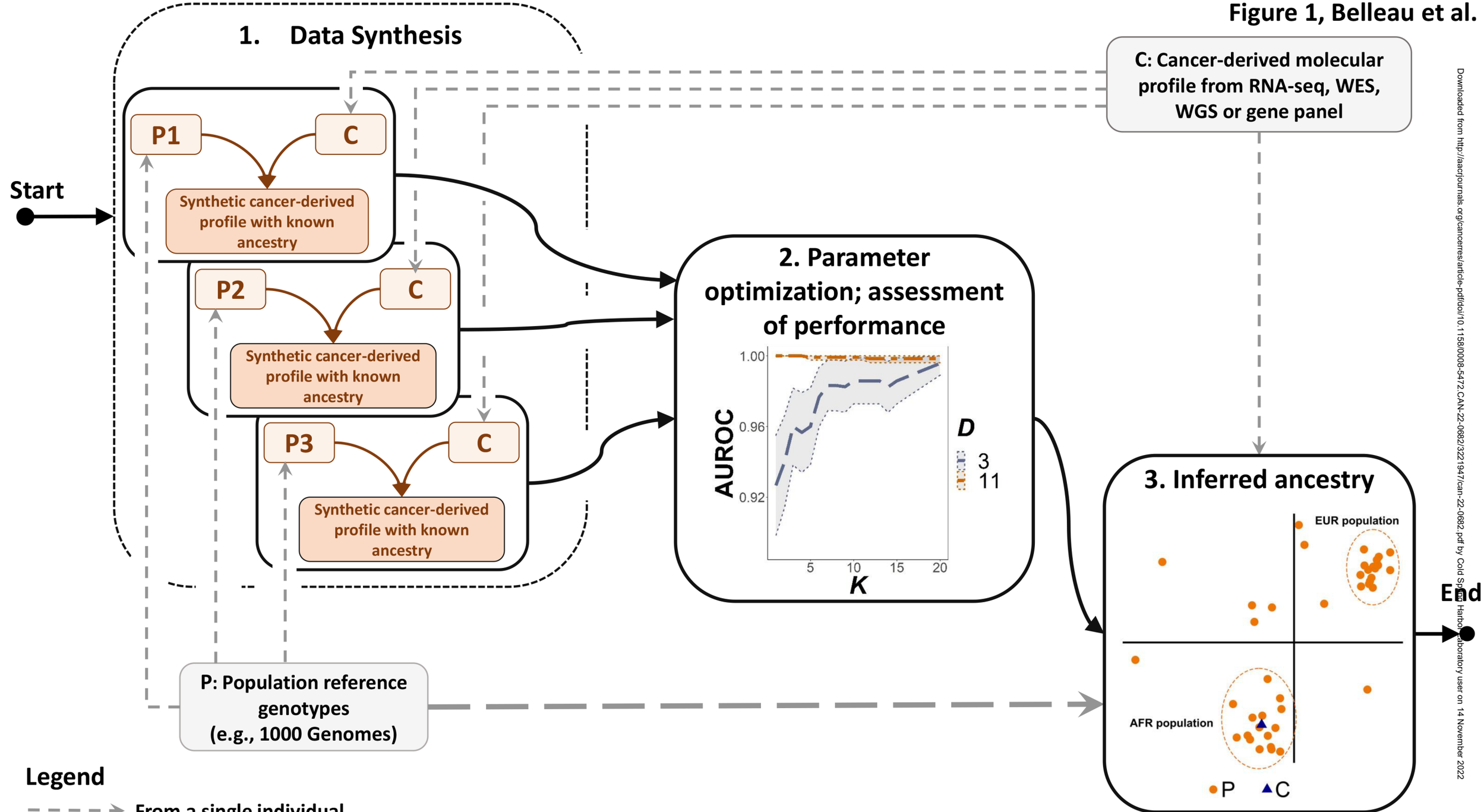

Figure 2, Belleau et al.

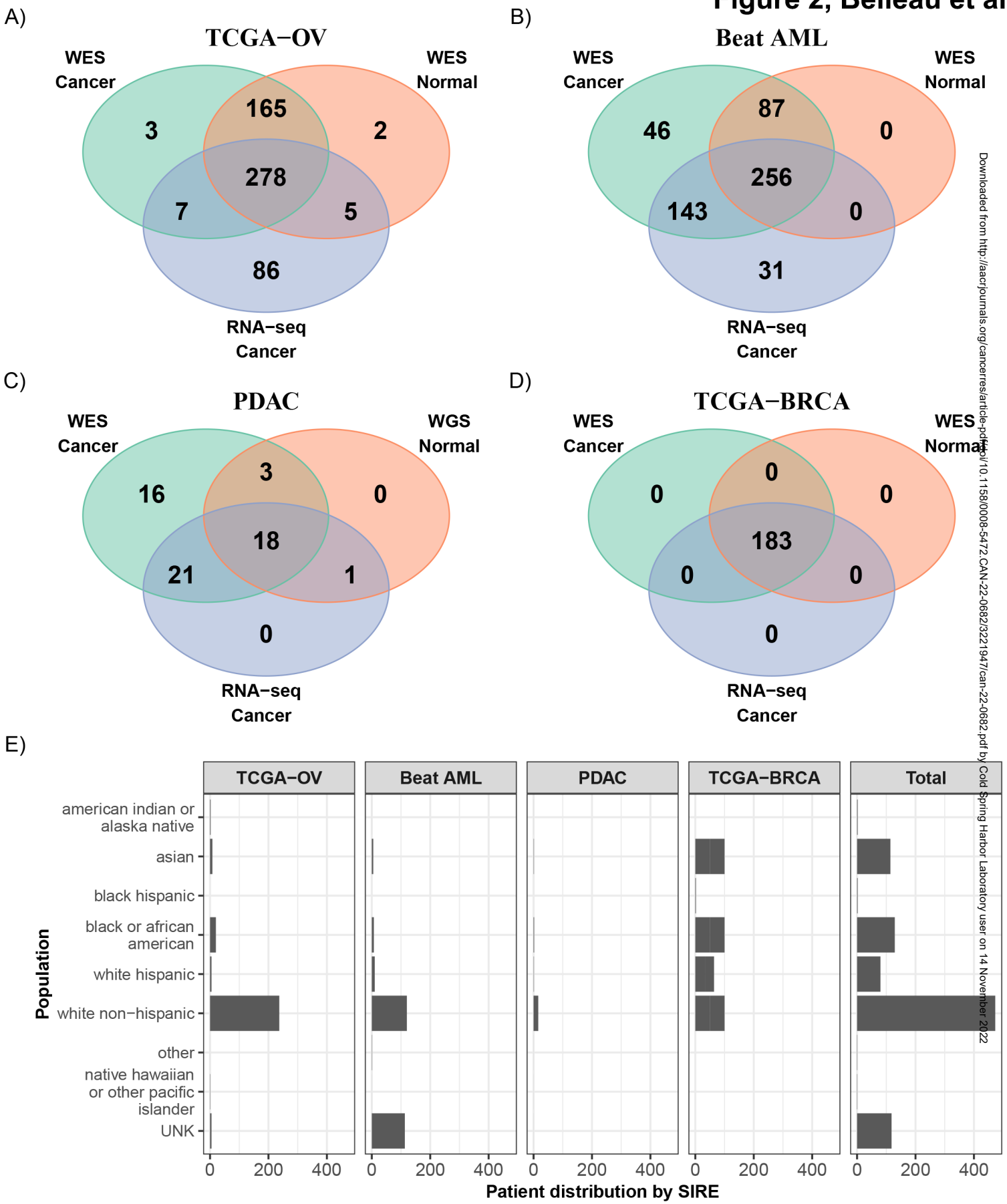

Figure 3, Belleau et al.

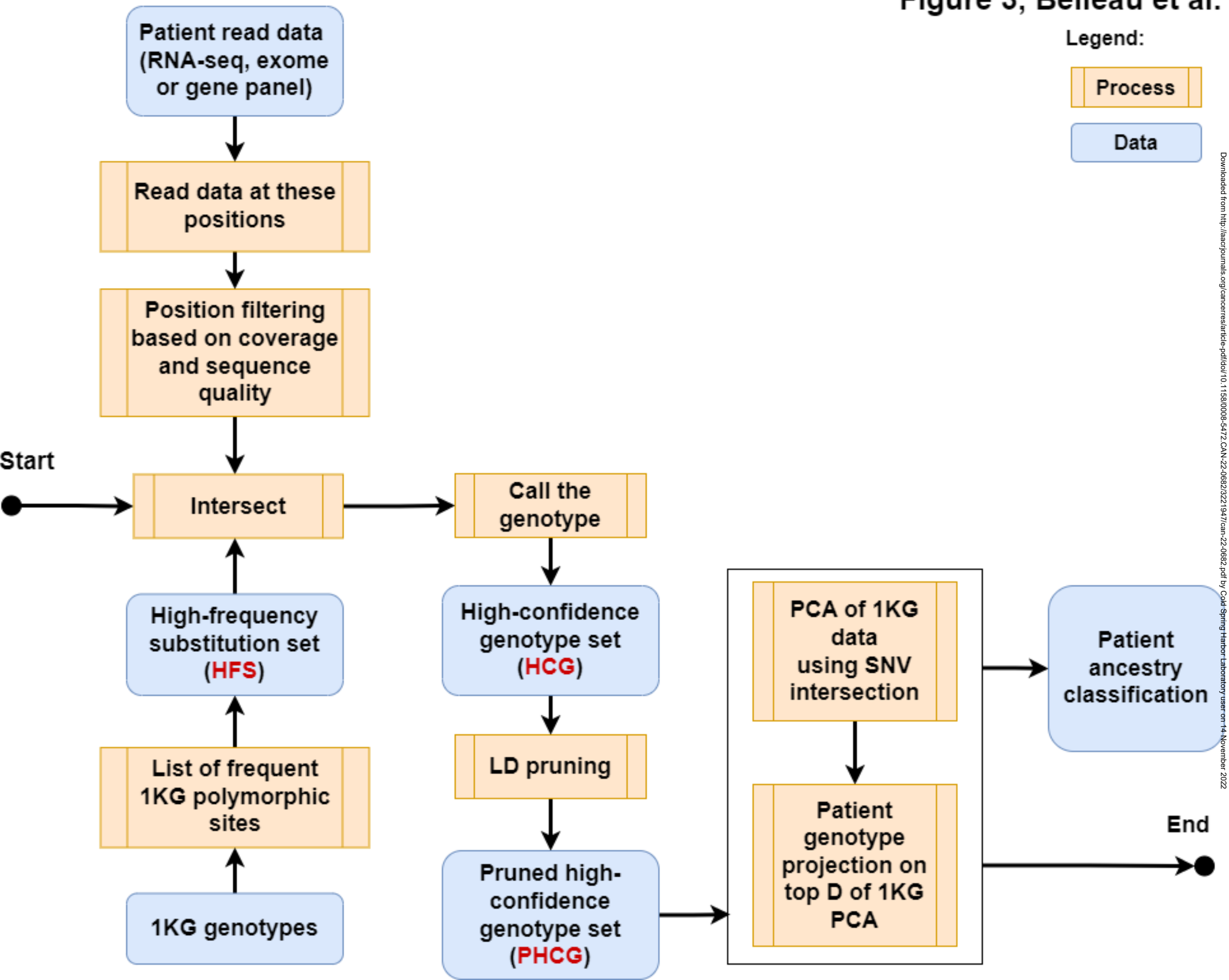

Figure 4, Belleau et al.