Ancestry inference from targeted NGS tests to enable precision medicine and improve racial/ethnic representation in clinical trials

Francisco M. De La Vega[1], Brooke Rhead[1], and Sean Irvine[2]

[1]Tempus Labs, Chicago, IL; [2]Real Time Genomics, Ltd., Hamilton, New Zealand

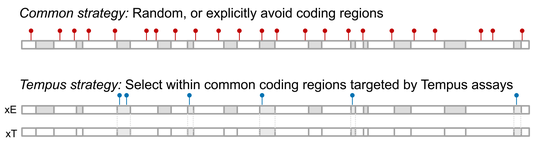
INTRODUCTION

There are well-established racial and ethnic disparities in cancer incidence and outcomes, in part due to structural, socioeconomic, environmental, and behavioral factors. Some of these differences can be attributed to biological factors, such as cancer mutation frequencies that vary by ancestry. It is well known that diversity in clinical trials is low, with Blacks and Hispanics consistently underrepresented compared to their cancer incidence, and race and ethnicity is missing in up to 50% of patient medical records and genomic profiling test orders. Moreover, self-reported race/ethnicity does not accurately reflect genetic ancestry, disproportionately affecting admixed patients. Rather than relying on self-reported race/ethnicity labels when investigating genetic effects and accounting for diversity, ancestry can be inferred directly from sequencing data collected during tumor profiling and other tests. Inferred ancestry can be used to improve representative participation in clinical trials and enable the assessment of biological differences that may determine differential efficacy of drugs for oncology and other indications.

METHODS

Ancestry is usually inferred from genome-wide data, either array or whole-genome sequencing (WGS), using unlinked random markers and clustering methods. However, this approach is inappropriate for targeted next-generation sequencing (NGS) gene panels or even whole-exome sequencing (WES) data. Instead, we selected 654 and 6,711 ancestry informative markers (AIMs) overlapping the regions targeted by the Tempus xT (exons of 648 cancer genes) and xE (WES) NGS assays, respectively .

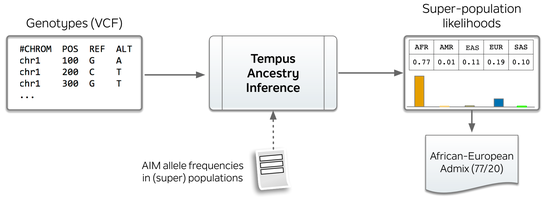
Figure 1. AIM selection strategy



We selected synonymous single-nucleotide variants in well-covered regions targeted by the assays, ensuring no substantial linkage disequilibrium by maintaining a minimum physical separation, and using an entropy measure to ensure maximum informativeness among pre-selected populations.

We developed a supervised global ancestry inference algorithm, akin to the ADMIXTURE[1] method, that calculates likelihoods for 5 continental groups defined in the 1000 genomes project (1KP)[2] – African (AFR), Americas (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) – using the selected AIMs and their allele frequencies in these populations compiled from different sources.

Figure 2. Ancestry inference workflow



The Tempus Ancestry inference software implements the ADMIXTURE[1] algorithm in a supervised fashion for pre-defined populations.

RESULTS

Validation

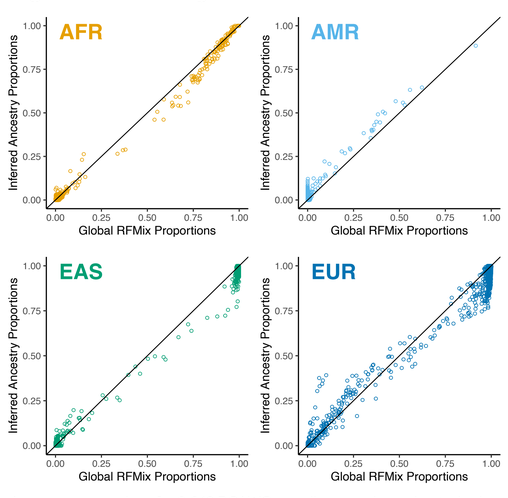
We validated our methods by comparing our results with labels of unadmixed donors from the 1KP and local ancestry inference previously derived with the RFMix software[3] for admixed individuals from the 1KP and the ICGC PCAWG[4] projects. To more easily compare performance in admixed samples where some ancestries are present at substantially lower fractions than others, we calculated a normalized mean squared error considering ancestry proportions inferred with RFMix. Results for PCAWG-project admixed donors are shown in **Table 1**, and graphical comparisons of our estimates to RFMix for the same samples are shown in **Figure 3**.

Table 1. Ancestry inference performance

	Normalized mean squared error (MSE)					
Marker set	AFR	EAS	EUR	AMR	SAS	Average MSE
xT-654	0.0084	0.0138	0.0072	0.3058	0.2716	0.1214
xE-6711	0.0034	0.0020	0.0018	0.0875	0.0615	0.0312

Normalized MSE: The MSE for each population divided by the sum of the RFMix proportions for that population. A value of 0 represents no difference between our results and RFMix. analysis using Tempus AIMs.

Figure 3. Ancestry inference correlations

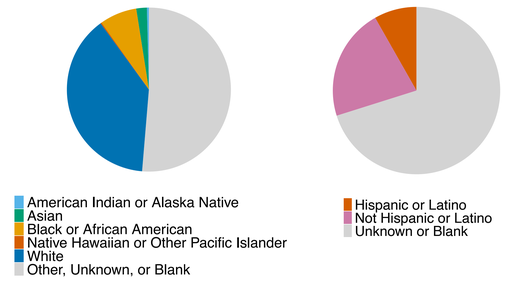


Ancestry proportions for 2,642 PCAWG germline samples estimated using the Tempus xE marker set (y-axis) or RFMix (x-axis). Comparisons are shown separately for AFR, AMR, EAS, and EUR ancestry. Similar results were obtained for SAS (not shown).

A case study: early onset colorectal cancer

The incidence of early onset colorectal cancer (EOCRC), defined as colorectal cancer diagnosed prior to age 50, is rising in the United States, and disproportionately rising in Black and Hispanic/Latino groups.[5] The reasons for these racial/ethnic disparities are unknown and likely include social, environmental, and genetic factors. We examined race and ethnicity metadata obtained from order forms or by abstraction of clinical documents of 1,775 de-identified EOCRC cases sequenced with the Tempus xT assay (**Figure 4**).

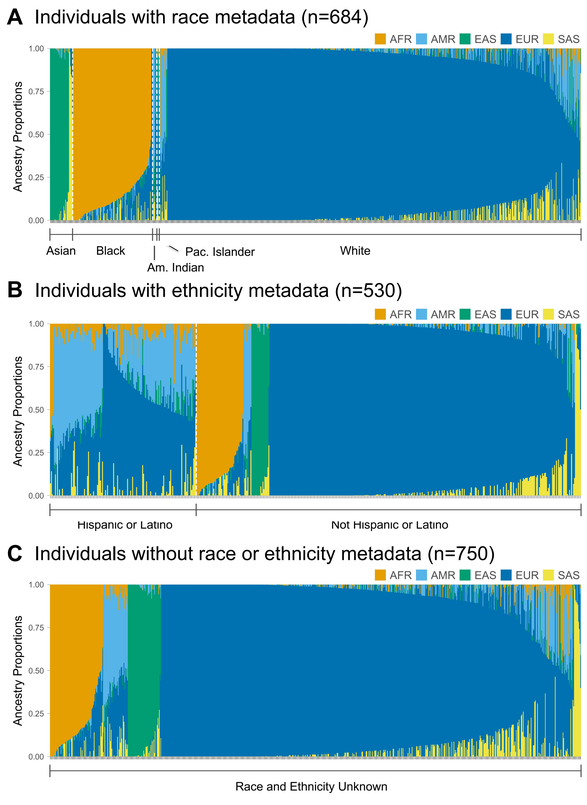
Figure 4. Early-onset colorectal cancer cohort: race and ethnicity metadata



Race and ethnicity metadata for 1,775 EOCRC cases with Tempus xT assay results. Race and ethnicity labels were not available for over half of patients.

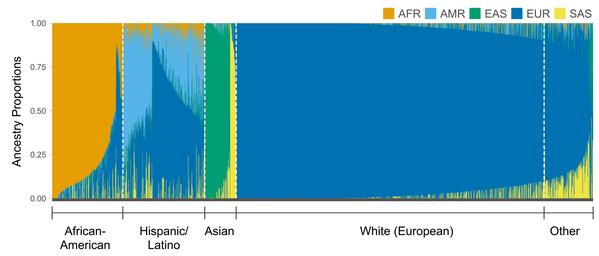
We estimated ancestry proportions from Tempus xT NGS assay data for the EOCRC cohort using the Tempus xT-654 AIM set for patients with available self-described race (**Figure 5A**), or ethnicity (**Figure 5B**), and for patients with neither race nor ethnicity information (**Figure 5C**). Data was from normal tissue when available and from tumor tissue otherwise.

Figure 5. Early-onset colorectal cancer cohort: inferred ancestry proportions



While genetic ancestry is not equivalent to race or ethnicity, it can be used to approximately impute race/ethnicity labels. We used ancestry proportions to match subjects in the EOCRC cohort to race/ethnicity labels (**Figure 6**). By doing this, we observed an approximately 80% increase in the total number of Black and Hispanic/Latino patients that could potentially be included in studies of EOCRC disparities.

Figure 6. Early-onset colorectal cancer cohort: imputed race/ethnicity groups



CONCLUSIONS

- We show that continental ancestry admixture proportions can be robustly inferred from AIMs present on Tempus xE and xT assays
- Inferred ancestry aligns well with race and ethnicity abstracted from test orders and associated clinical documents
- Imputed race/ethnicity labels can be used to identify diverse sets of patients for inclusion in studies and clinical trials
- Future work includes expanding the set of reference populations used for inference

REFERENCES

[1]Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655–1664..

[2]Auton, A. et al. (2015). A global reference for human genetic variation. Nature 526, 68 74.

[3]Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. AJHG, 93:278-288

[4]Campbell, P. J. et al. (2020). Pan-cancer analysis of whole genomes. Nature 578, 82–93.

[5]Muller, C., Ihionkhan, E., Stoffel, E. M., & Kupfer, S. S. (2021). Disparities in Early-Onset Colorectal Cancer. Cells, 10(5).

TEMPUS