**Global ancestry estimation and data repository for the TCGA pan-cancer resource**

Jordan Creed and Travis Gerke

**Background:** The Cancer Genome Atlas (TCGA) is a vital resource in molecular cancer research. Opportunities to conduct cancer health disparities research from this resource are currently limited by incomplete data capture for self-reported race. Moreover, self-reported measures have known limitations, such as binning mixed race individuals into a single racial group which may not reflect their genetic make-up and thus risk. Therefore, we estimated global ancestry for all available TCGA samples according to standardized populations from 1000 Genomes. We provide this information in a publicly accessible data resource.

**Methods:** For all available sample types (primary solid tumor, blood derived normal or other), genotypes were downloaded from TCGA’s Legacy Archive. ADMIXTURE software was used to estimate ancestral proportions from each of the five 1000 Genomes global super populations. Phase 3 samples from 1000 Genomes (n = 2504) were used as reference. Ancestry estimation was based on approximately 700,000 variants that overlapped between TCGA and 1000 Genomes. Supervised ADMIXTURE clustering (K = 5) was run with 200 bootstrapped replicates.

**Results:** From the 11127 TCGA participants, 22963 samples were examined spanning 9 sample types (10769 (46.90%) primary solid tumor, 9157 (39.88%) blood derived normal and 2356 (10.26%) solid tissue normal, the three most common). According to the dominant estimated genetic super population, Europeans were the most common with 18757 (81.68%) samples, followed by African (2077; 9.04%), East Asian (1361; 5.93%), admixed American (611; 2.66), and South Asian (157; 0.68%). Most participants (10853; 97.54%) had genotypes assessed from multiple samples, 1030 (9.49%) of which differed in dominant super population between tissues. This was predominantly observed as a difference in European and African classification from blood derived normal versus primary solid tumor.

We created a public resource (https://github.com/GerkeLab/TCGAancestry), which contains global ancestry estimates to enable other researchers to incorporate this information into their own analyses.

**Conclusions:** Our release of global ancestry estimates impactfully expands the TCGA data resource and may accelerate health disparities related discoveries in the molecular cancer domain.