Computational inference of genetic ancestry in cancer

Jenea Adams

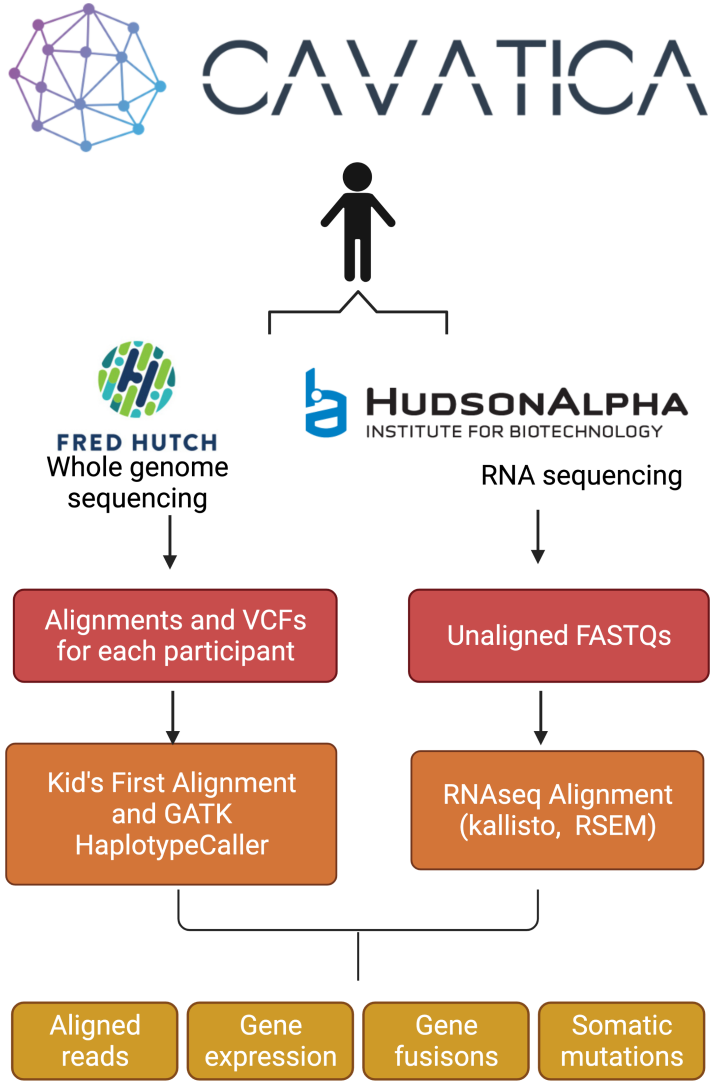
Xing Lab Roundtable

October 12, 2021

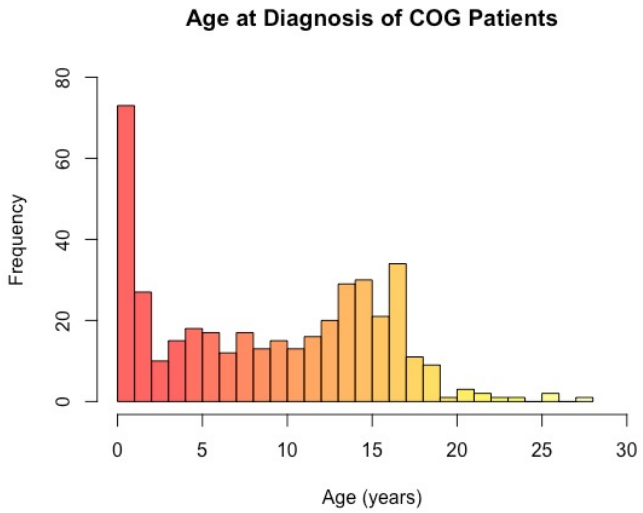
CAVATICA: CHOP data-analysis platform of raw and harmonized multiomic data

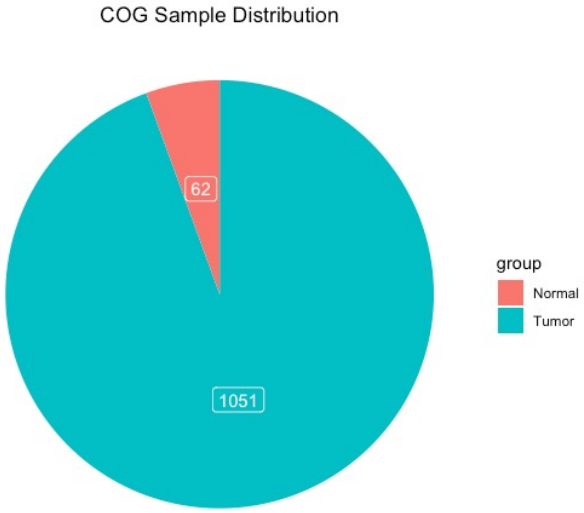
Overview of data used for this analysis

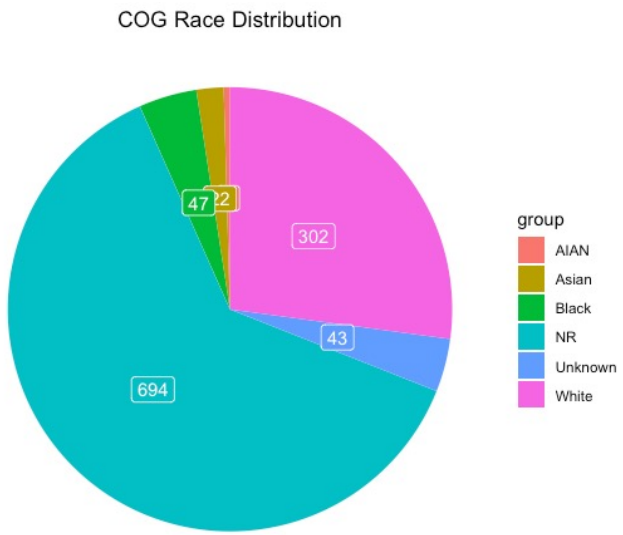
- De-novo AML, DS-AML, APL-AML
- Data from Children's Oncology Group Clinical Trial (No. AAML1031)
- 1,113 RNA-seq files (aligned and quantified)
  - Both kallisto and RSEM were used
  - Gene fusions also quantified
- 408 whole-genome sequencing (WGS) files

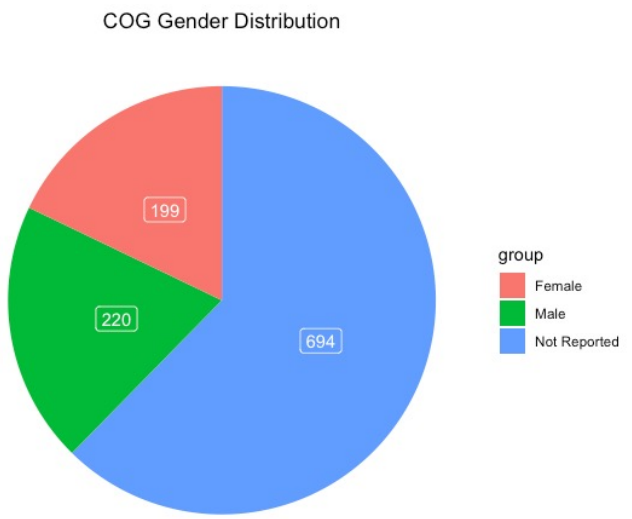


CAVATICA allows easy metadata extraction



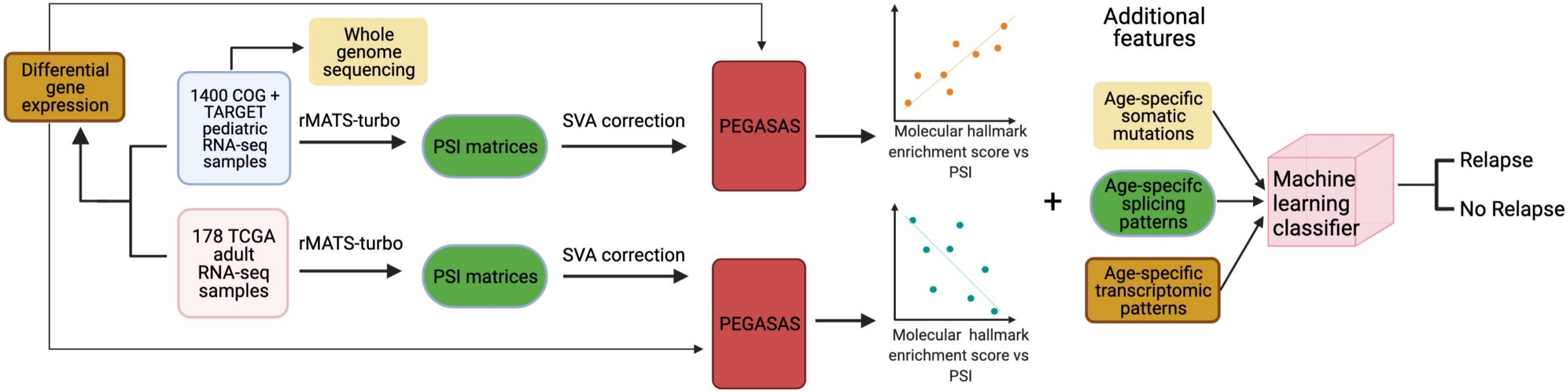






In Progress

- Analysis: Sourcing metadata
  - **Treatment history of COG patients**
  - Use of whole genome seq data for ancestry association with molecular subtypes and disease progression
  - RNA-seq data for gene expression analysis of molecular subtypes



Genetic ancestry can be used to analyze factors affecting clinical outcomes

- Ancestries influence germline genetics
- Ancestries tend to carry different disease exposures
- Multi-omic features can contribute to the prevalence of relapse in pediatric AML, inclucing ancestry

Cancer Cell

CellPress

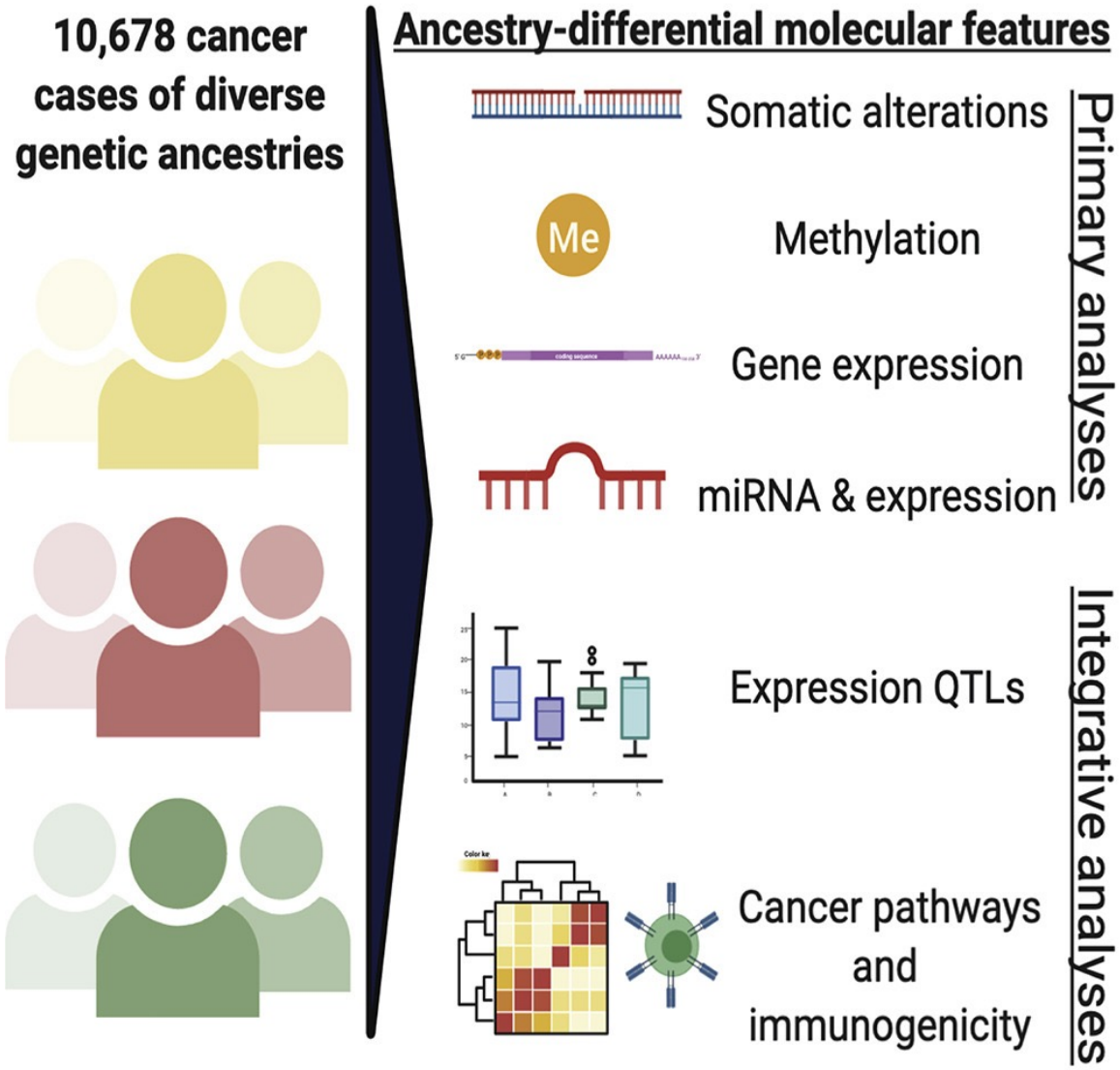
Article

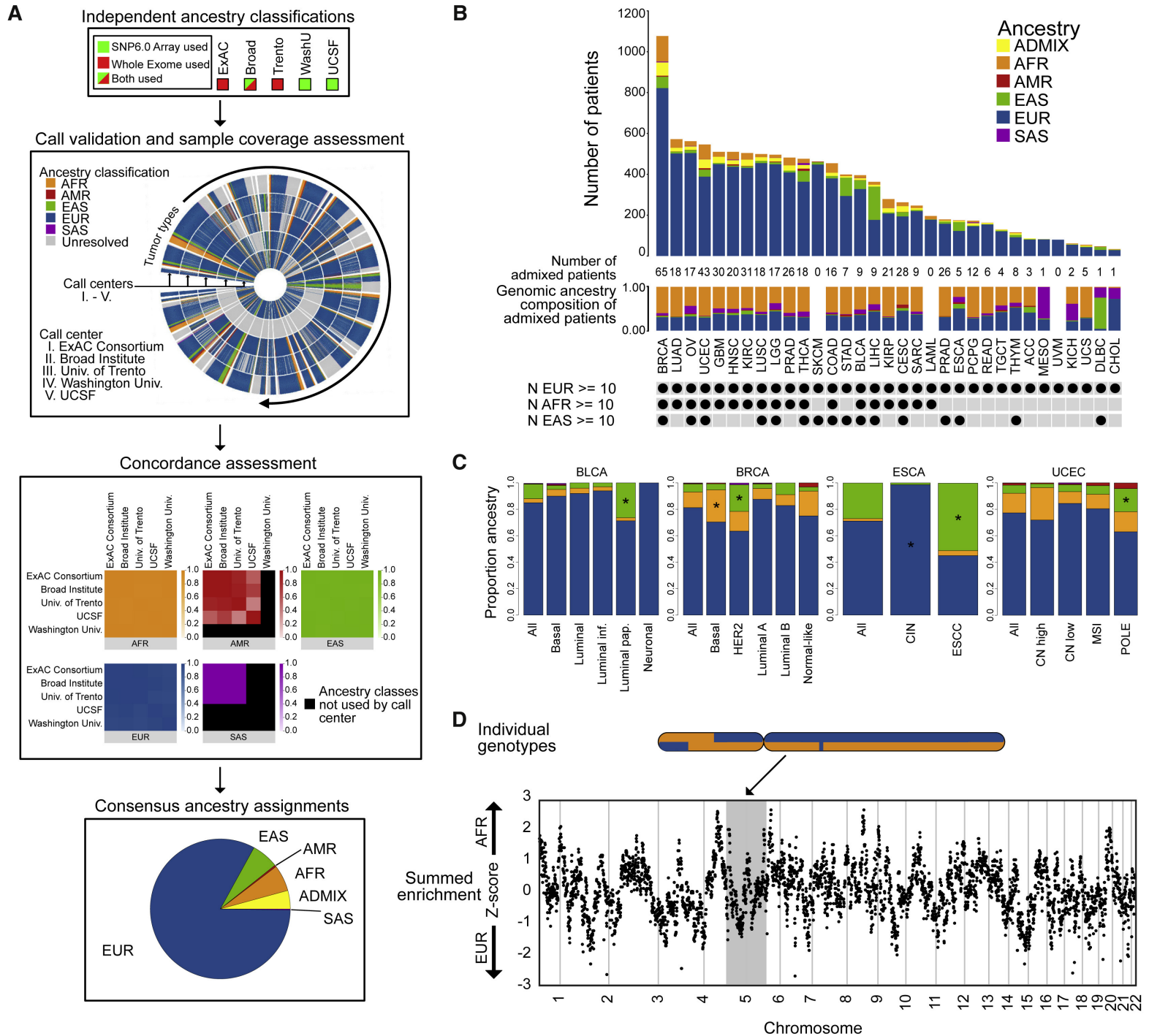
Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer

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

Paper outline

- Determining a biological basis of health disparities in cancer
- Central question: *if you quantify ancestry, could you see differences in molecular characteristics from their tumors?*




A
Independent ancestry classifications
SNP6.0 Array used
Whole Exome used
Both used
ExAC
Broad
Trento
WashU
UCSF
Call validation and sample coverage assessment
Ancestry classification
AFR
AMR
EAS
EUR
SAS
Unresolved
Tumor types
Call centers
I. - V.
Call center
I. ExAC Consortium
II. Broad Institute
III. Univ. of Trento
IV. Washington Univ.
V. UCSF
Concordance assessment
ExAC Consortium
Broad Institute
Univ. of Trento
UCSF
Washington Univ.
AFR
AMR
EAS
EUR
SAS
Ancestry classes not used by call center
Consensus ancestry assignments
EAS
AMR
AFR
ADMIX
SAS
EUR
B
Number of patients
Ancestry
ADMIX
AFR
AMR
EAS
EUR
SAS
Number of admixed patients
65 18 17 43 30 20 31 18 17 26 18 0 16 7 9 9 21 28 9 0 26 5 12 6 4 8 3 1 0 2 5 1 1
Genomic ancestry composition of admixed patients
BRCA LUAD OV UCEC GBM HNSC KIRC LUSC LGG PRAD THCA SKCM COAD STAD BLCA LIHC KIRP CESC SARC LAML PRAD ESCA PCPG READ TGCT THYM ACC MESO UVM KICH UCS DLBC CHOL
N EUR >= 10
N AFR >= 10
N EAS >= 10
C
Proportion ancestry
BLCA
All
Basal
Luminal
Luminal inf.
Luminal pap.
Neuronal
BRCA
All
Basal
HER2
Luminal A
Luminal B
Normal-like
ESCA
All
CIN
ESCC
UCEC
All
CN high
CN low
MSI
POLE
D
Individual genotypes
Summed enrichment Z-score
AFR
EUR
Chromosome

(Some) methods used for ancestry inference

- Broad – SNP and Exome based calls
  - SNP → smartpca with AFR, EUR, EAS, AMR, SAS in first 3 PCs
  - Exome → from ExAC, PCA on 5400 SNPs to cluster continental ancestry
- WashU – SNP based calls
  - PCA in PLINK – use self-reported as reference within first 2 PCs
- UCSF – SNP based calls
  - PAM and k means clustering of PCs with PLINK (w/o LD pruning)

Limitations and questions

- Reference populations
  - We need a diverse reference population (besides 1000 genomes)
    - What other sources exist?
- Most examples are from consortia-level science, what does this mean for this project?
- This paper concluded that biologically differences between ancestries were tissue-specific, not cancer specific