Genetically regulated tumor gene expression in the Carolina Breast Cancer Study

November 5, 2019

—

Michael Love
Department of Biostatistics
Department of Genetics
UNC-Chapel Hill

UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

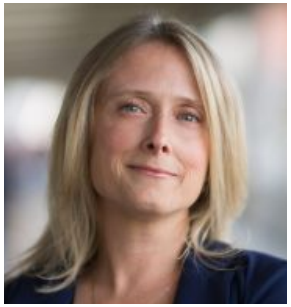
bit.ly/bc-twas

Disclosure statement

- The following personal financial relationships with commercial interests relevant to this presentation existed during the past 12 months:
  - C.M.P. is an equity stockholder in and consultant for BioClassifier LLC; C.M.P. is also listed as an inventor on patent applications on the Breast PAM50 Subtyping assay. The other authors declare that they have no competing interests.

Research Team at UNC

- Arjun Bhattacharya – Biostatistics Ph.D. candidate
- Melissa Troester – PI of Carolina Breast Cancer Study
- Andy Olshan – co-PI of CBCS
- Charles Perou - Molecular Oncology, Genetics

UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

Carolina Breast Cancer Study

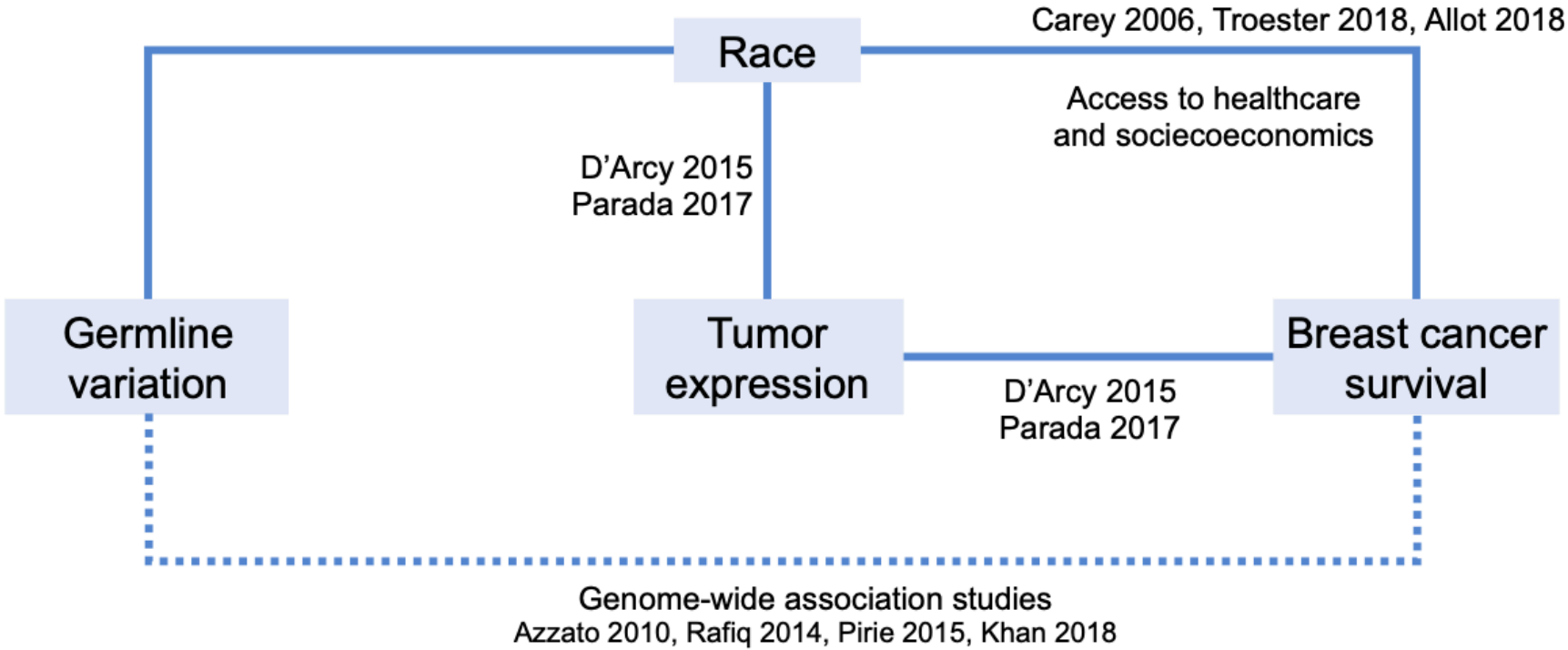
- population-based study among NC women (1993-2013)
- aims of the study at outset:
  - integration of epidemiology and molecular biology
  - identify causes of BC among African-American women (AA) and women of European ancestries (WW)
  - associations between environmental & behavioral factors and BC in relation to specific molecular alterations (germline and tumor)
- young women (20-49 yrs) and self-identified African American women oversampled using randomized recruitment

Carolina Breast Cancer Study

- key considerations for the study:
  - heterogeneity of disease (histopathology, genomics)
  - known racial disparities in incidence and survival in the US
  - small portion of incidence explained by germline susceptibility
  - to what degree does observed variability in disease reflect underlying etiologic heterogeneity?

Carolina Breast Cancer Study

- eligibility criteria:
  - female,
  - first diagnosis of invasive or *in situ* breast cancer,
  - aged 20–74 years at diagnosis,
  - residence in specified NC counties
- this work: 3,828 women with BC (1,865 AA + 1,963 WW) from phases 1-3 with relevant survival, clinical, genomic variables

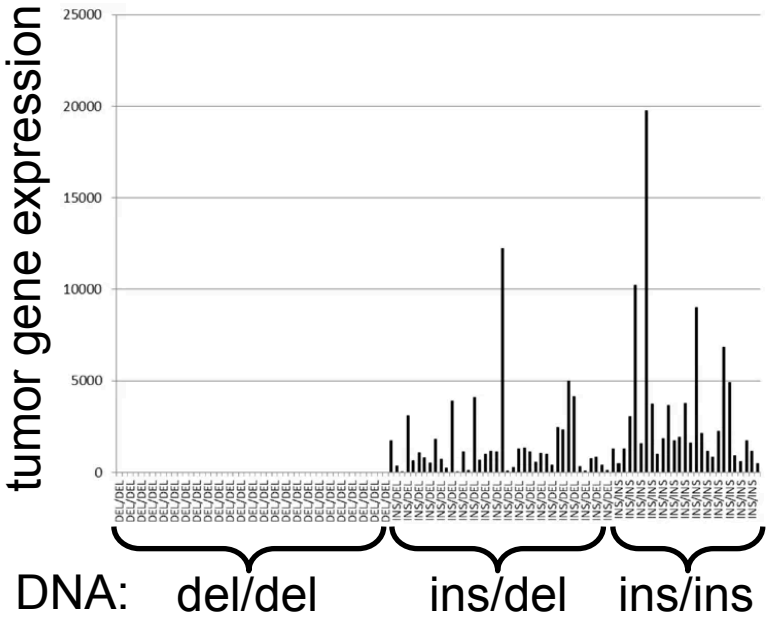

Race
Carey 2006, Troester 2018, Allot 2018
Access to healthcare
and sociecoeconomics
D'Arcy 2015
Parada 2017
Germline
variation
Tumor
expression
Breast cancer
survival
D'Arcy 2015
Parada 2017
Genome-wide association studies
Azzato 2010, Rafiq 2014, Pirie 2015, Khan 2018

How can genetics inform our questions?

- Just because a gene is differentially expressed (DE) across self-reported race, does not make it a good candidate
- Rummel et al 2014: *"PSPHL and breast cancer in African American women: causative gene or population stratification?"* 10.1186/1471-2156-15-38
- *PSPHL* is a pseudogene, expressed in tumor at different levels across race
- Promoter and first three exons are in 30 kb of DNA not in the reference genome

Rummel *et al* 2014: *PSPHL*

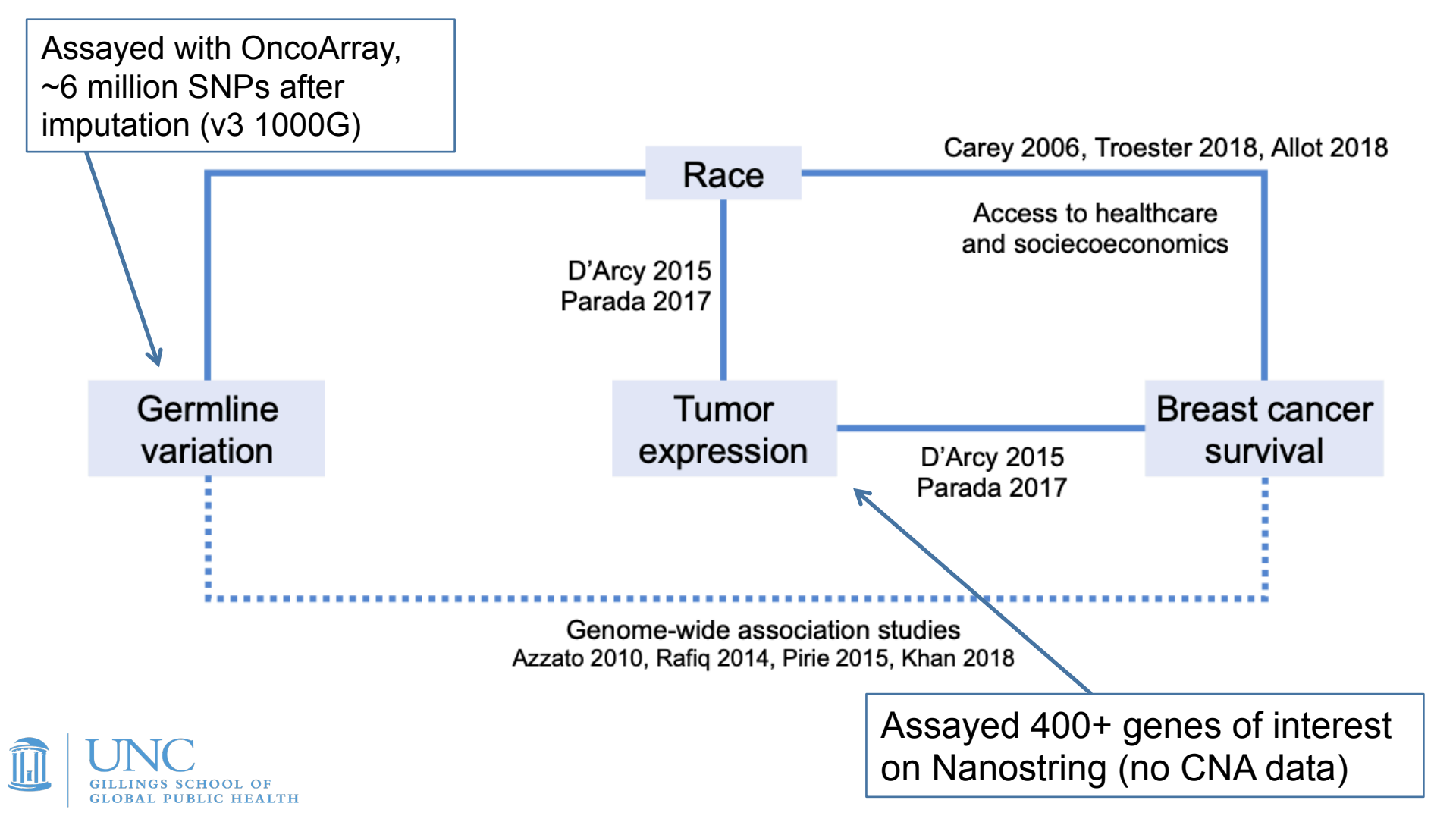
tumor expression **QTL**



	del/del	ins/del	ins/ins
AA	<5%	1/3	2/3
WW	2/3	1/3	<5%

		del/del	ins/del	ins/ins
AA	case	.06	.35	.59
	control	.04	.31	.65
WW	case	.62	.34	.04
	control	.64	.34	.02

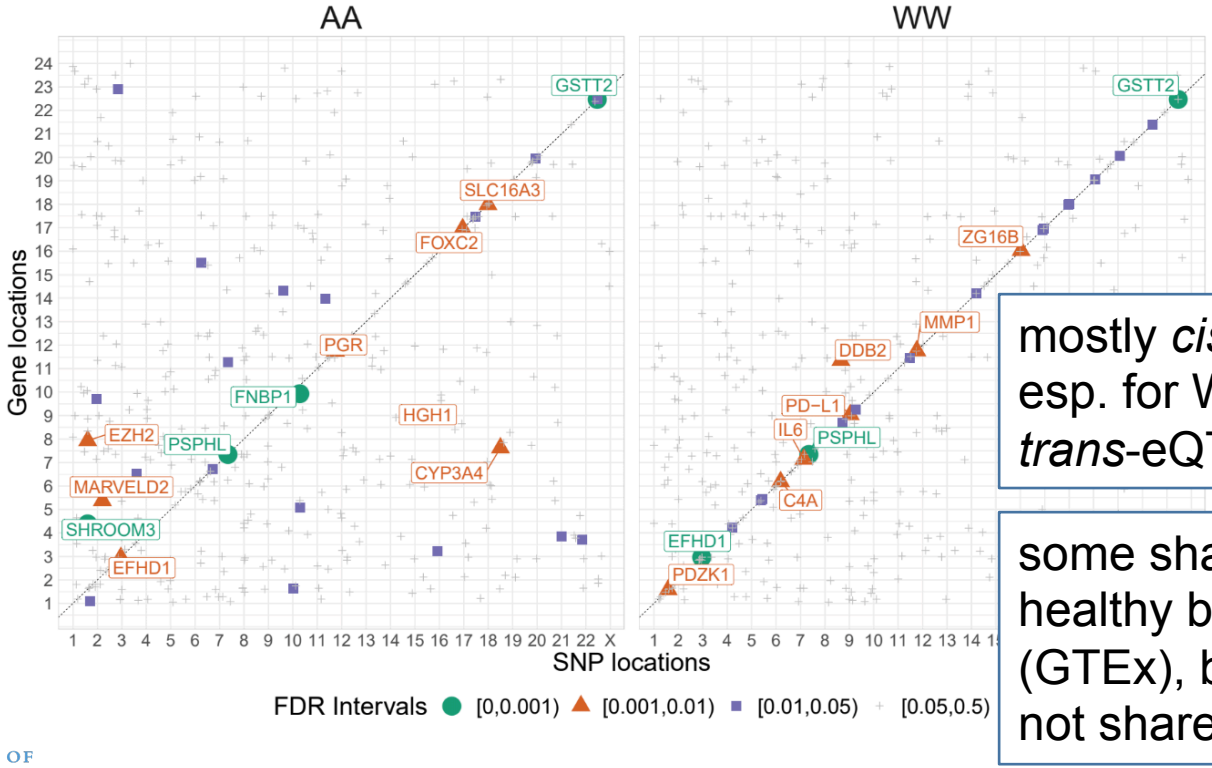
Similar story for stage, ER/HER2 status, grade, size, lymph node status, though small sub-groups



UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Tumor expression QTL

Race-stratified tumor expression analysis of 400+ genes from germline genotype

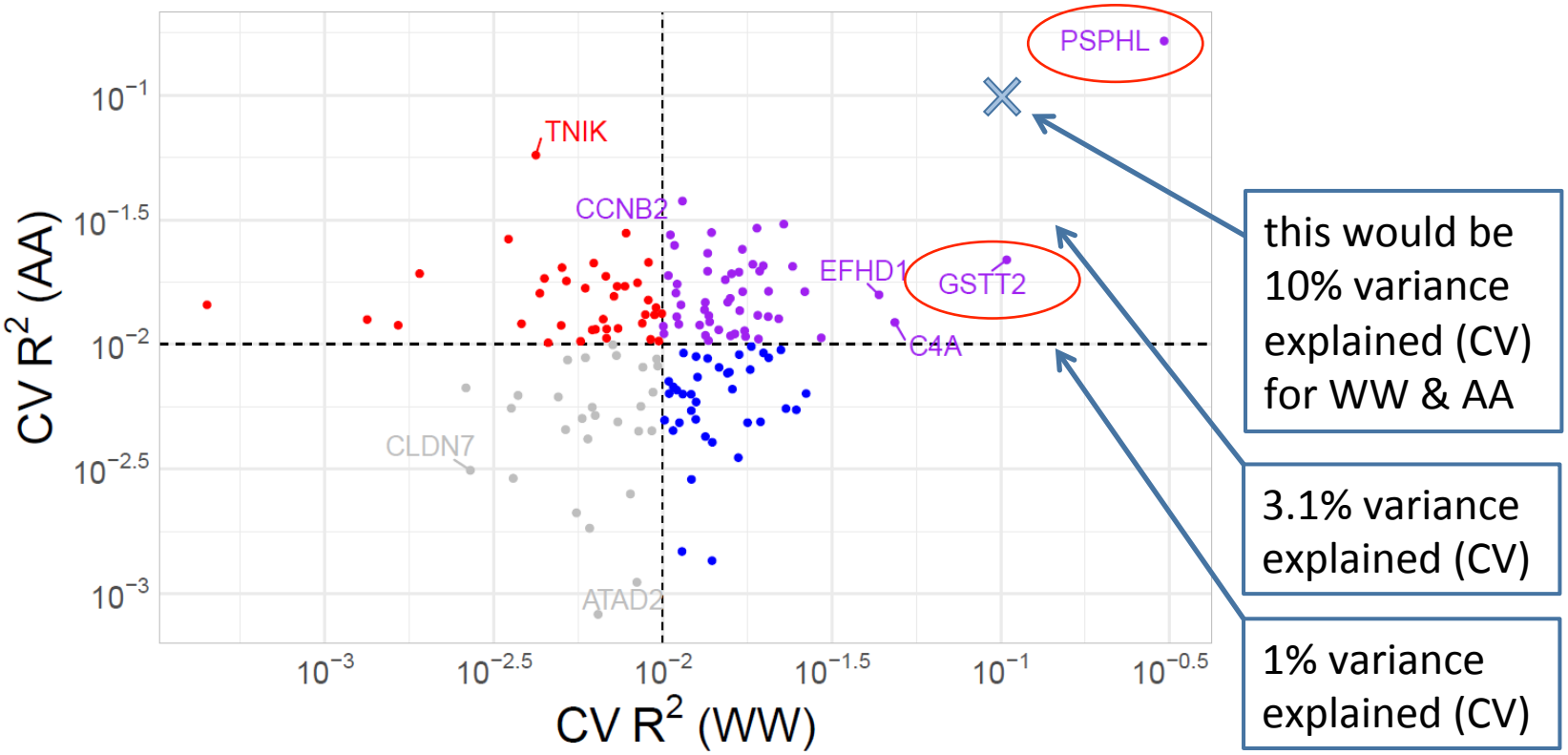


mostly *cis*-eQTL, esp. for WW, some *trans*-eQTL for AA

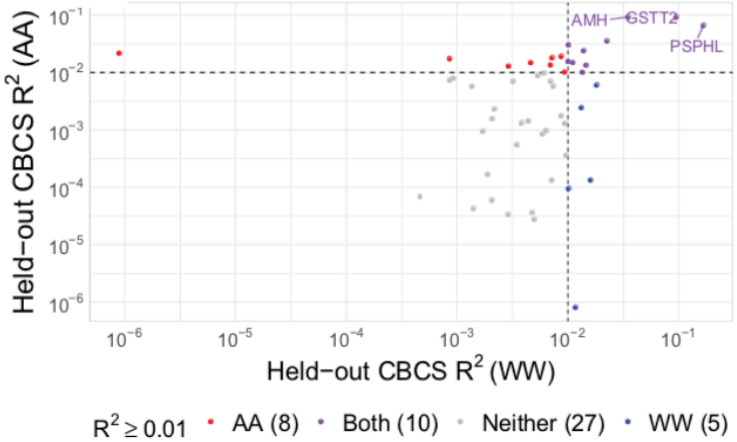
some shared with healthy breast (GTEx), but many not shared

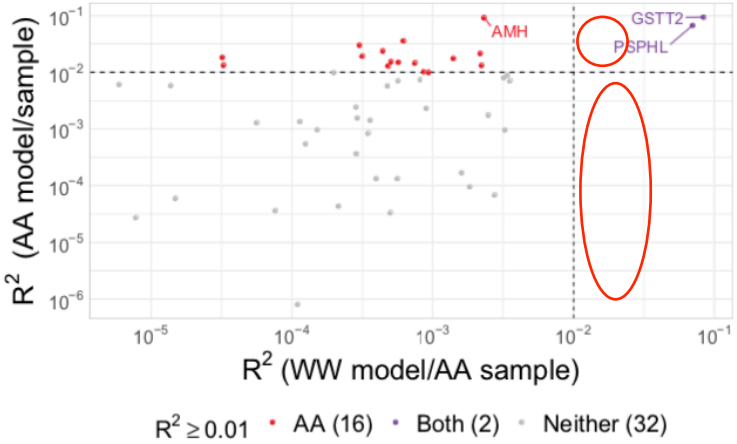
UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

(Germline) genetically regulated tumor gene expression



Expression models weren't generally applicable across race



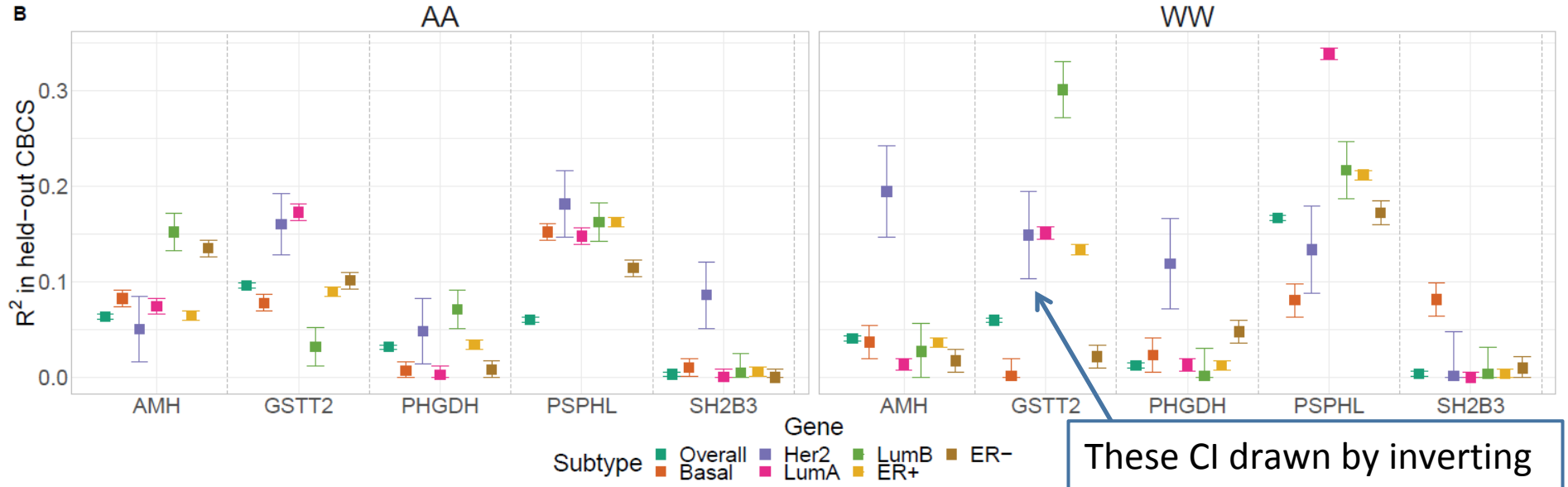


Exceptions are *PSPHL* and *GSTT2*

This is part of a much larger thread in genetics:

- Mogil et al (2018) *"Genetic architecture of gene expression traits across diverse populations"* 10.1371/journal.pgen.1007586
- Wojcik et al (2019) *"Genetic analyses of diverse populations improves discovery for complex traits"* 10.1038/s41586-019-1310-4

Predictive accuracy varies by subtype



These CI drawn by inverting permutation test. Due to heterogeneous sub-group size, critical to include variability here.

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Back to *PSPHL* example

- Genetically-driven associations after stratifying
- Survival analysis for 46 / 57 genes (AA / WW) in CBCS (admittedly small *n* for genetic associations)
- Race-stratified cause-specific hazard model on genetically regulated tumor expression (CV imputed)
- Controlling for age at diagnosis, ER status at diagnosis, tumor stage at diagnosis, study phase

Transcriptome-wide association (TWAS): four BC-mortality-associated loci in AA

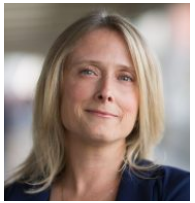
Region	Gene	Hazard Ratio (90% CI - FDR adj)	Z-Statistic[a]	P-value[a]	GReX R^{2}[b]
20q13.2	*AURKA*	0.83 (0.73, 0.95)	-2.52	1.5×10^{-3}	0.021
2p23.1	*CAPN13*	1.22 (1.07, 1.41)	2.76	5.4×10^{-4}	0.011
~~*3q26.32*~~	~~*PIK3CA*~~	~~0.85 (0.74, 0.97)~~	~~-2.34~~	~~3.2×10^{-3}~~	~~0.013~~
18q21.33	*SERPINB5*	0.82 (0.72, 0.93)	-2.85	3.4×10^{-4}	0.010

Collider bias should be a concern: None of these four with genetically regulated tumor expression associated with cancer incidence in AA women available from BCAC using the iCOGs dataset and additional GWAS.

Next steps and CBCS questions

- Genetically regulated tumor expression
  - Collaborate to apply tumor expression models to larger cohorts
  - Local ancestry - better expression models or associations?
- Etiology heterogeneity
  - Why does molecular subtype incidence differ?
  - Subtype-specific risk papers:
    Ahearn et al (2019) 10.1101/733402v1, Zhang et al (2019) 10.1101/778605v2
    Begg & Zabor (2012) 10.1093/aje/kws128, Begg et al (2015) 10.1002/cam4.456
  - Outcome disparities within subtype

Acknowledgments

- Arjun Bhattacharya – bhattacharya-a-bt.github.io
- Melissa Troester (UNC) and Troester group
- Andy Olshan (UNC)
- Charles Perou (UNC)
- Montserrat García-Closas (NCI)
- CBCS patients and participants
- CBCS staff

Susan G. Komen® provided financial support for CBCS study infrastructure.

Funding was provided by the National Institutes of Health, National Cancer Institute **P01-CA151135, P50-CA05822,** and **U01-CA179715** to A.F.O, C.M.P. and M.A.T.

M.I.L. is supported by **R01-HG009937, R01-MH118349, P01-CA142538,** and **P30-ES010126.**

The Translational Genomics Laboratory is supported in part by grants from the National Cancer Institute (**3P30CA016086**) and the University of North Carolina at Chapel Hill **University Cancer Research Fund**.

Genotyping was done at the DCEG Cancer Genomics Research Laboratory using funds from the **NCI Intramural Research Program**.

bioRxiv preprint: bit.ly/bc-twas

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

eQTL analyses

- We conducted all eQTL analyses stratified by race.
- Age, BMI, postmenopausal status, and the first 5 principal components of the joint AA and WW genotype matrix were included in the models as covariates in C.
- Estimated tumor purity was also included as a covariate to assess its impact on strength and location of eQTLs.

Predictive models

- Gene expression residualized for the covariate matrix
- We estimate w_g with the best predictive of three schemes:
  - (1) elastic-net regularized regression with mixing parameter a = 0.5 and penalty parameter tuned over 5-fold cross-validation,
  - (2) linear mixed modeling where the genotype matrix X_g is treated as a matrix of random effects and $\hat{w}_g$ is taken as the best linear unbiased predictor (BLUP) of w_g, using rrBLUP, and
  - (3) multivariate linear mixed modeling as described above, estimated using GEMMA v.0.97.

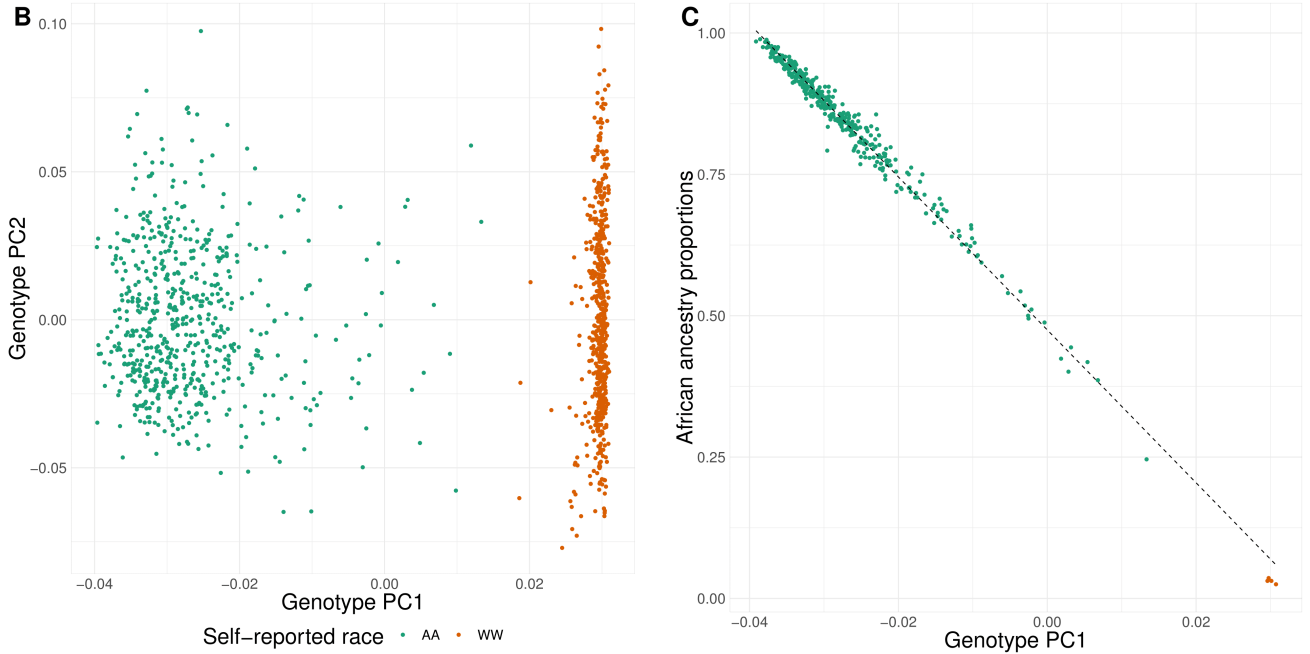
Survival modeling

- We defined a relevant event as a death due to breast cancer. We aggregated all deaths not due to breast cancer as a competing risk. Any subjects lost to follow-up were treated as right-censored observations.
- We estimated the association of GReX with breast cancer survival by modeling the race-stratified cause-specific hazard function of breast cancer-specific mortality, stratifying on race.
- For a given gene g, the model has form

$$\lambda_k(t) = \lambda_{0k}(t)e^{GReX_g\beta_g + Z_C\beta_C},$$

  where β_g is the effect size of GReX on the hazard of breast cancer-specific mortality, Z_C represents the matrix of covariates (age at diagnosis, estrogen-receptor status at diagnosis, tumor stage at diagnosis, and study phase), and β_C are the effect sizes of these covariates on survival. $\lambda_{k(t)}$ is the hazard function specific to breast cancer mortality, and $\lambda_{ok(t)}$ is the baseline hazard function.
- We test H_o: β_g=0 for each gene g with Wald-type tests, as in a traditional Cox proportional hazards model. We correct for genomic inflation and bias using bacon, a method that constructs an empirical null distribution using a Gibbs sampling algorithm by fitting a three-component normal mixture on Z-statistics from TWAS tests of association.

Self-reported race and ancestry PCs in CBCS



This is PC1, linear combination with most variance in *this* dataset.

"Human population structure is not race" - Birney, Raff, Rutherford, Scally [bit.ly/36aId0j](bit.ly/36aId0j)

Adam Rutherford, "*A Brief History of Everyone Who Ever Lived*"

Angela Saini, "*Superior: The Return of Race Science*"