Big Data Analysis to Characterize Triple-Negative Breast Cancer

Jin Liu, PhD

Department of Pharmaceutical Sciences

UNTHSC

Triple-negative breast cancer (TNBC) is a health disparity disease

What is TNBC?

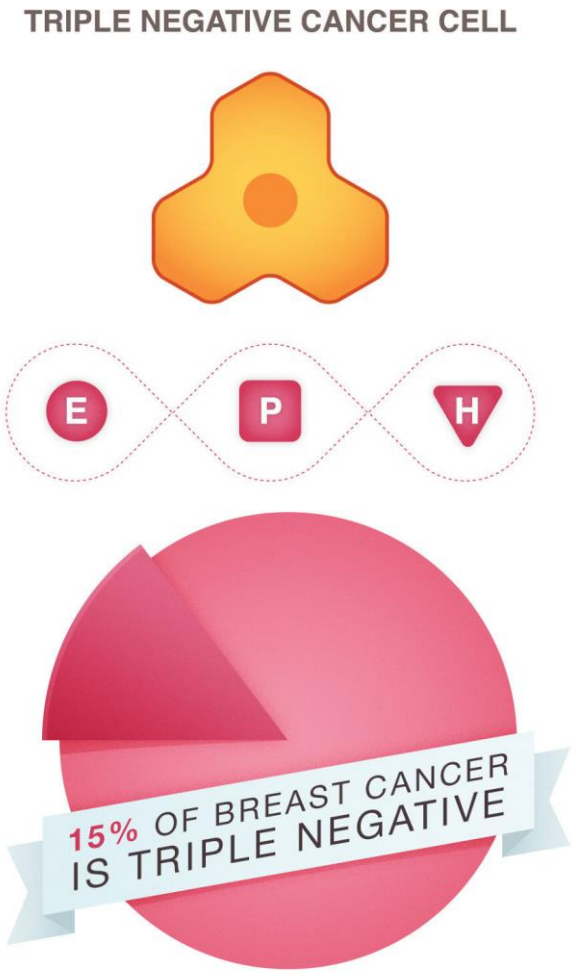
A subtype of breast cancer that is negative for 3 receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor 2 receptor ((Her2)

What is the problem of TNBC?

- Most breast cancer therapies targeting those 3 receptors are ineffective for TNBC patients
- An aggressive subtype: worse prognosis, early relapse, a high frequency of metastasis to lung, liver and brain, and a low overall survival rate

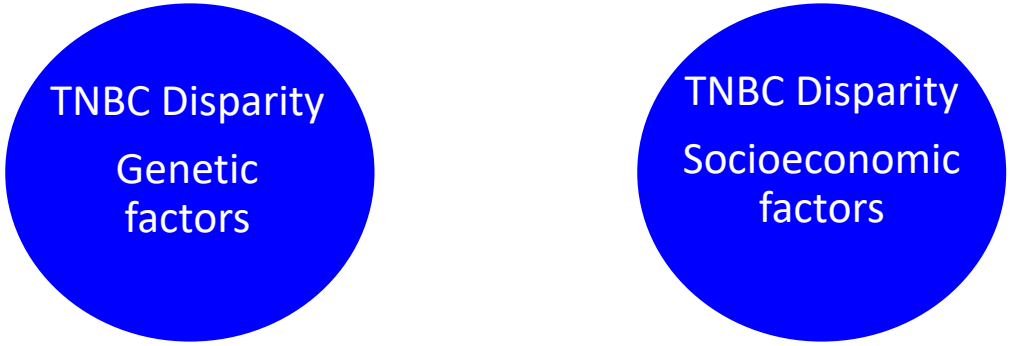
Higher prevalence in African American women

- Twice incidence than Whites
- 42% higher mortality rate than Whites

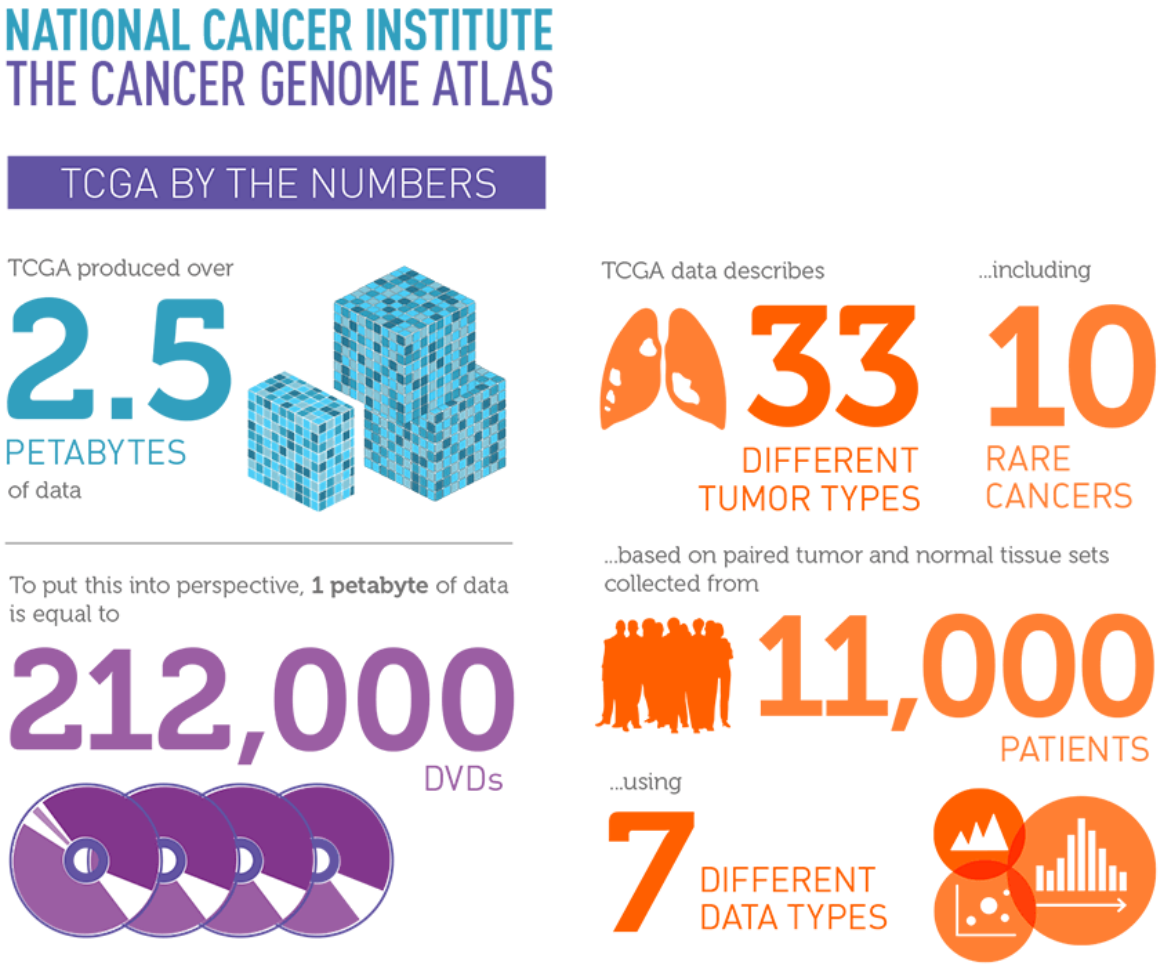


Objectives of this project

- Identify key features contributing to TNBC disparity
- Discover actionable drug targets for TNBC
- Advance the application of big data and artificial intelligence algorithms in TNBC research



Genetic Factors: Data Preparation



cBioPortal FOR CANCER GENOMICS

- Clinical data for 1085 female breast cancer patients in TCGA-BRCA project were downloaded through cBioPortal.

Data Preparation

- 101 TNBC patients
- 9450 gene mutation data for each patient
- >63,000 gene expression data for each patients

Classification models to classify patients based on race

Feature selection models to identify key genes

Gene set enrichment analysis to validate identified genes

What genes are uniquely mutated for African Americans?

Gene Name	Weight
MCF2L2	0.0507
HSPG2	0.0458
LYST	0.0458
APOB	0.0398
CFAP47	0.0398
COL18A1	0.0398
CREBBP	0.0398
FCGBP	0.0398
PXDNL	0.0398
SI	0.0398
USP34	0.0398
POTEG	0.0374

Gene Name	Weight
ALMS1	0.0339
COL6A6	0.0339
CSMD2	0.0339
DCHS2	0.0339
DMD	0.0339
FER1L5	0.0339
MYO18B	0.0339
PIK3CA	0.0339
PRX	0.0339
TDRD5	0.0339
TNIK	0.0339

MCF2L2 and **HSPG2** are top two genes that are uniquely highly mutated in African American TNBC patients

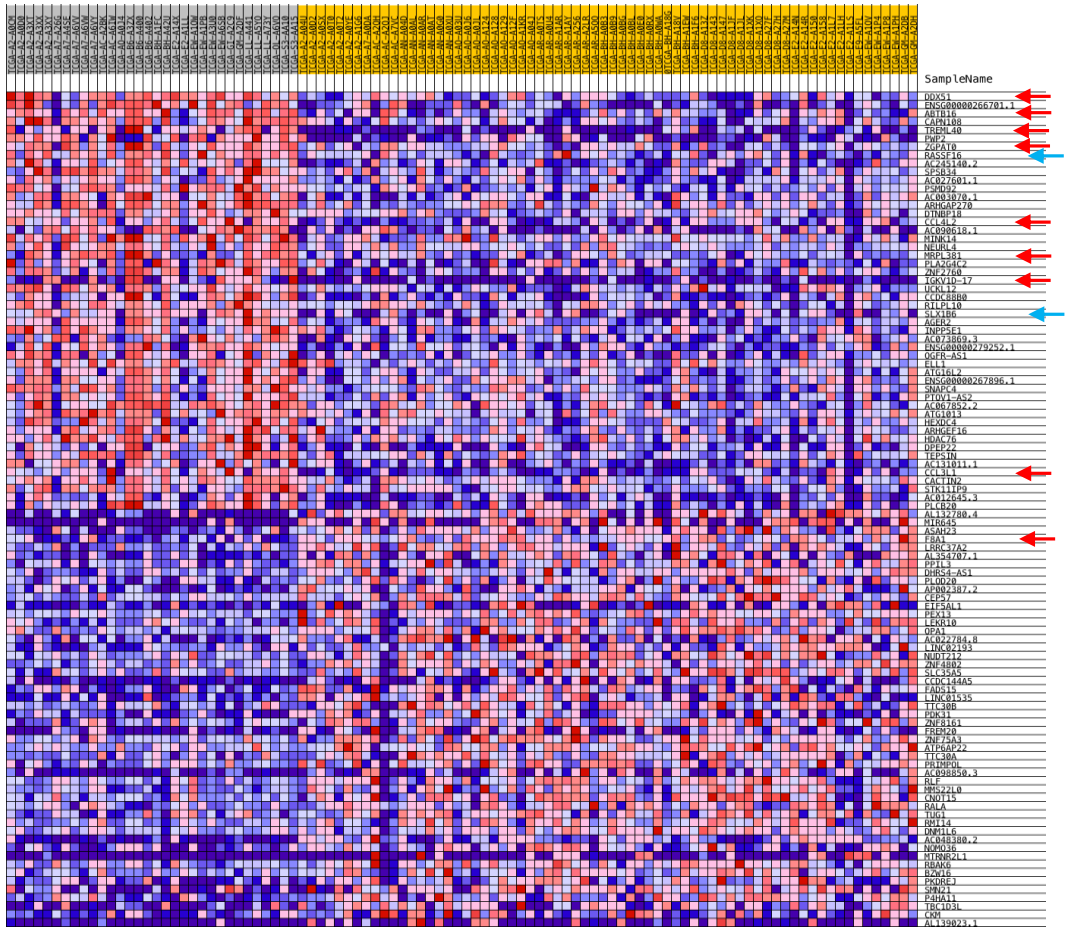
What genes are uniquely overexpressed for African Americans?

- Top ten weighted features by Information Gain (IG), Information Gain Ratio (IGR), Chi-Square (CS) and Gini Index (GI) for a total of 23 features

Gene Expression Attributes Weighted by Information Gain							
Gene ID	Gene Name	Gene Description	Gene Type	IG	IGR	CS	GI
ENSG00000142330.18	CAPN10	calpain 10 [Source:HGNC Symbol;Acc:HGNC:1477]	protein_coding	1	2	3	1
ENSG00000276085.1	CCL3L1	C-C motif chemokine ligand 3 like 1 [Source:HGNC Symbol;Acc:HGNC:10628]	protein_coding	2			3
ENSG00000260005.5	AC027601.1	uncharacterized LOC 105371925 ncRNA [Source:NCBI gene;Acc:105371925]	antisense	3			2
ENSG00000277203.1	F8A1	coagulation factor VIII associated 1 [Source:HGNC Symbol;Acc:HGNC:3547]	protein_coding	4	1		6
ENSG00000280195.1	AC245140.2	novel transcript, antisense to RPL 10	antisense	5			
ENSG00000188056.10	TREML4	triggering receptor expressed on myeloid cells like 4 [Source:HGNC Symbol;Acc:HGNC:30807]	protein_coding	6			4
ENSG00000279656.1	AL132780.4	uncategorized gene, and is affiliated with the lncRNA class	TEC	7		7	7
ENSG00000185163.8	DDX51	DEAD-box helicase 51 [Source:HGNC Symbol;Acc:HGNC:20082]	protein_coding	8	9	2	5
ENSG00000177989.12	ODF3B	outer dense fiber protein 3B	protein_coding	9			9
ENSG00000197114.10	ZGPAT	zinc finger CCCH-type and G-patch domain containing [Source:HGNC Symbol;Acc:HGNC:15948]	protein_coding	10	10		8
ENSG00000226806.1	AC011893.1	uncharacterized LOC100507600, ncRNA [Source:NCBI gene;Acc:100507600]	antisense		3		
ENSG00000213999.14	MEF2B	myocyte enhancer factor 2B [Source:HGNC Symbol;Acc:HGNC:6995]	protein_coding		4		
ENSG00000204316.11	MRPL38	mitochondrial ribosomal protein L38 [Source:HGNC Symbol;Acc:HGNC:14033]	protein_coding		5		10
ENSG00000108405.3	P2RX1	purinergic receptor P2X 1 [Source:HGNC Symbol;Acc:HGNC:8533]	protein_coding		6		
ENSG00000104918.6	RETN	resistin [Source:HGNC Symbol;Acc:HGNC:20389]	protein_coding		7		
ENSG00000242766.1	IGKV1D-17	immunoglobulin kappa variable 1D-17 [Source:HGNC Symbol;Acc:HGNC:5749]	IG_V_gene		8		
ENSG00000116871.14	MAP7D1	MAP7 domain containing 1 [Source:HGNC Symbol;Acc:HGNC:25514]	protein_coding			1	
ENSG00000114626.16	ABTB1	ankyrin repeat and BTB domain containing 1 [Source:HGNC Symbol;Acc:HGNC:18275]	protein_coding			4	
ENSG00000266701.1	AC005702.4	uncharacterized	MiRNA			5	
ENSG00000228157.4	AC007952.2	uncharacterized	processed_transcript			6	
ENSG00000139631.17	CSAD	cysteine sulfinic acid decarboxylase [Source:HGNC Symbol;Acc:HGNC:18966]	protein_coding			8	
ENSG00000241945.6	PWP2	PWP2, small subunit processome component [Source:HGNC Symbol;Acc:HGNC:9711]	protein_coding			9	
ENSG00000122490.17	PQLC1	PQ loop repeat containing 1 [Source:HGNC Symbol;Acc:HGNC:26188]	protein_coding			10	

CAPN10 and **CCL3L1** are top two genes that are uniquely overexpressed for African American TNBC patients

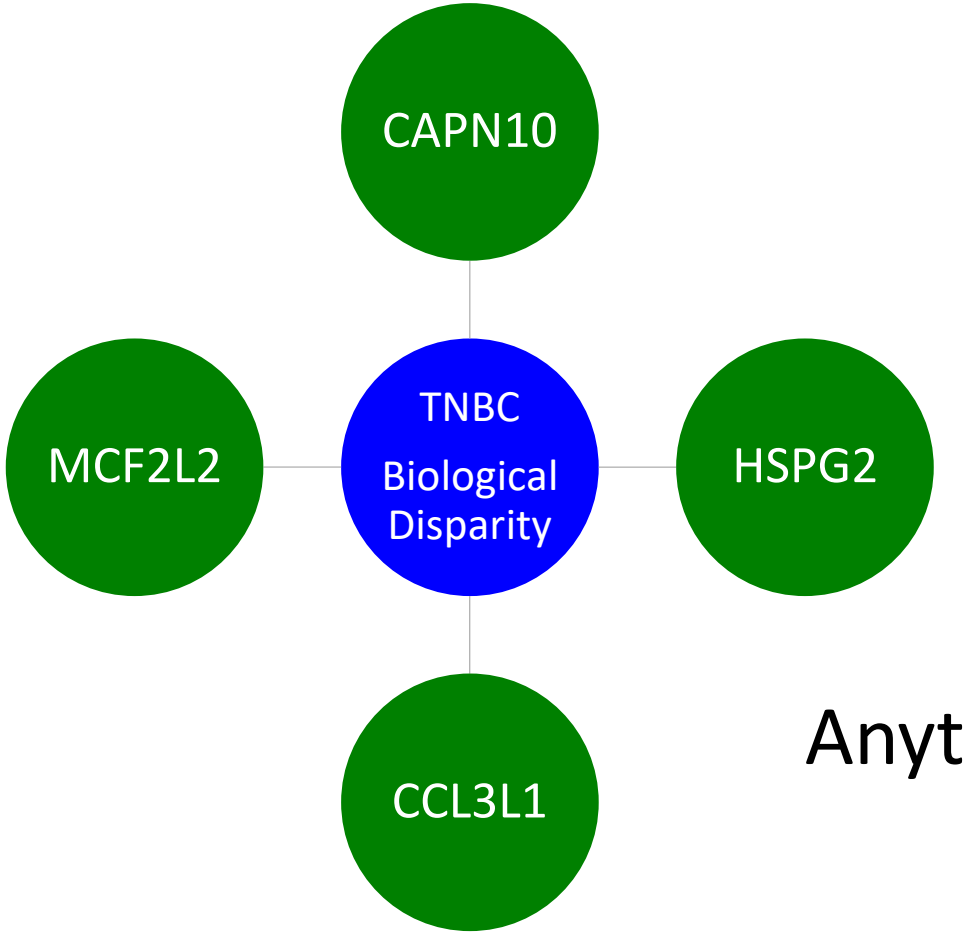
Gene set enrichment study showed the similar results



- Very highly ranked by weighting methods
- Fairly highly ranked by some weighting methods

CAPN10 and **CCL3L1** are among top genes that are identified by GSEA as overexpressed for African American TNBC patients

What do we get so far?



Anything in common for these 4 genes?

They all linked to **diabetes!**

All the identified top 4 genes are linked to diabetes

- MCF2L2 gene mutation was found to be associated with the development of nephropathy in type 1 diabetes mellitus

  Zhang et al, BMC Med Genet. 2010; 11: 116.

- HSPG2 gene mutation was found to affect the diabetes mellitus in type 2 diabetes patients

  Kurnaz et al, Cell Mol Biol. 2016; 62(8):35-9..

- Overexpression of CAPN10 is associated with Type 2 diabetes

  Ridderstråle et al, Curr Hypertens Rep. 2008 Feb;10(1):19-24.

- Overexpression of CCL3L1 is linked to Type 1 diabetes

  McKinny et al, Ann Rheum Dis. 2008;67(3):409-13.

More than top 4 genes are linked to diabetes

Rank	Top mutated genes	Top Over/uderexpressed genes
1	MCF2L2	CAPN10
2	HSPG2	CCL3L1
3	LYST	F8A1*
4	APOB	TREML4
5	CFAP47	DDX51
6	COL18A1	ODF3B
7	CREBBP	ZGPAT
8	FCGBP	MEF2B
9	PXDNL	MRPL38
10	SI	P2RX1
11	USP34	RETN
12	POTEG	IGKV1D-17
13	ALMS1	MAP7D1
14	COL6A6	ABTB1
15	CSMD2	CSAD

All red genes are linked to **diabetes!**

Pathways involved

Rank	Top mutated genes	Top Over/uderexpressed genes
1	MCF2L2	CAPN10
2	HSPG2	CCL3L1
3	LYST	F8A1*
4	APOB	TREML4
5	CFAP47	DDX51
6	COL18A1	ODF3B
7	CREBBP	ZGPAT
8	FCGBP	MEF2B
9	PXDNL	MRPL38
10	SI	P2RX1
11	USP34	RETN
12	POTEG	IGKV1D-17
13	ALMS1	MAP7D1
14	COL6A6	ABTB1
15	CSMD2	CSAD

- Extracellular matrix organization and degradation pathways: HSPG2, CAPN10, COL18A1
- Immune pathways: CCL3L2, RETN, APOB, CREBBP
- Signaling transduction pathways: MCF2L2, HSPG2, APOB, CREBBP
- Metabolism pathways: APOB, CREBBP, HSPG2

African American higher rate of TNBC is linked to diabetes?

- Any reports on it?
  - YES!
  - An observational study showed type 2 diabetes increased the risk for ER-negative breast cancer in African-American women by more than 40 percent

Palmer et al, Cancer Res 2017;77(22):6462-6469

- Do African American women have higher rate on diabetes?
  - YES!
  - The risk of diabetes is 77% higher among African Americans than among non-Hispanic white Americans.

http://www.diabetes.org/diabetes-basics/statistics/

- Could diabetes drugs help with TNBC?
  - YES!
  - Metformin, a first-line drug for type 2 diabetes mellitus, suppresses triple-negative breast cancer stem cells

Shi et al, Cell Discov. 2017; 3: 17010

Take home message #1

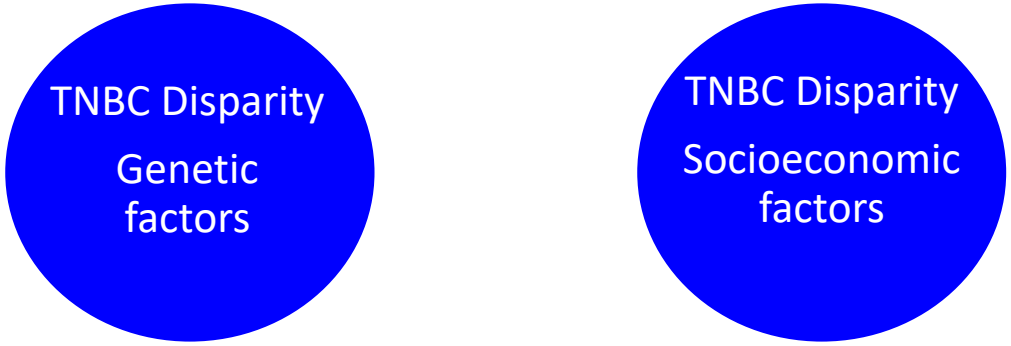
- We identified top mutated and overexpressed genes for African American TNBC patients

- The top genes are all linked to diabetes

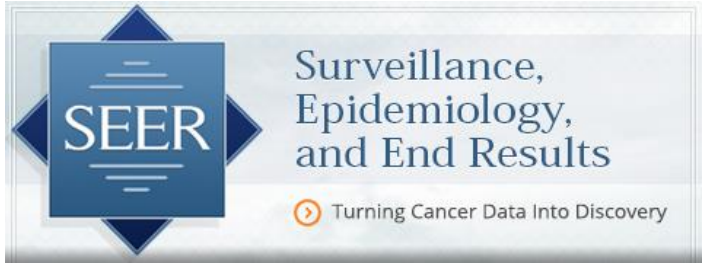
Diabetes may be linked to health disparity in triple negative breast cancer

Objectives of this project

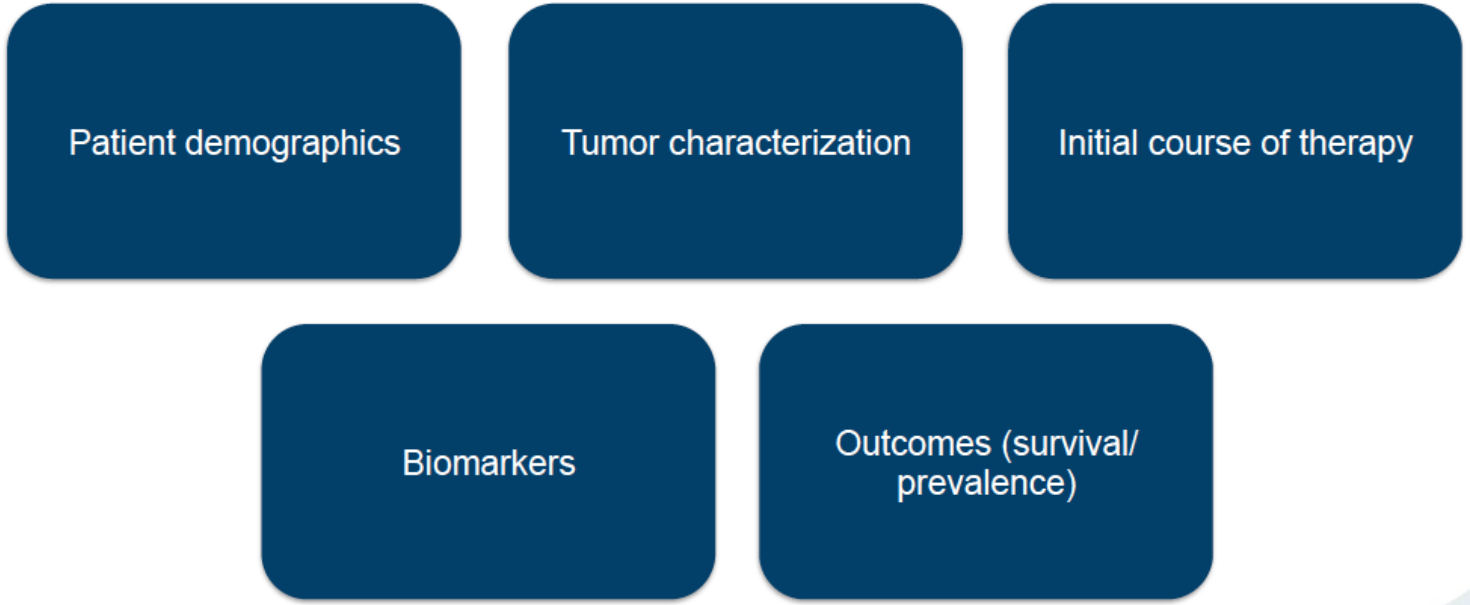
- Identify key features contributing to TNBC disparity
- Discover actionable drug targets for TNBC
- Advance the application of big data and artificial intelligence algorithms in TNBC research



SES Factors: Data Preparation



- National Cancer Institute's longitudinal data repository
- Source of data on cancer incidence, treatment and surviv
- Designed to support research
- Population-based registries covering 28% of U.S. populat
- >400,000 incident cases reported annually



Data Preparation

35,976 TNBC patients (2010-2015)

441 attributes for each patient

Classification models to

classify patients based on race

Feature selection models to identify key attributes

What attributes are unique for African Americans?
Machine Learning Results

Top Attribute	Desciption	Mode Value	Weight
Rural - Urban Coninuum Code 2013	Categorizes metropolitan and nonmetropolitan counties by population size	Counties in metropolitan areas ge 1 million pop	0.054
% Moved Past Year from Outside US ACS 2011-15	Probability of persons moved from outside of the United States from 2011-2015 (0.00-0.609%)*	0.073	0.014
Marital Status at Diagnosis	A persons marital status at time of diagnosis (single, divorced, married, widowed, separated, domestic partner, unknown)	Married (including common law)	0.078
Health Service Area	Categoried county or counties that are related by hospital care	Los Angeles, CA - Orange, CA	0.092
Insurance Recode (2007+)	Persons status and/or type of insurance (insured, uninsured, any medicaid, unknown, insured/no specifics)	Insured	0.11
Normalized cost-of-living index 2004	An approximated cost over time to refect the required amount a person needs to live ($7,190-$15,280)	$11,170	0.029
Age at Diagnosis (years)	Persons age at time of diagnosis	53	0.024
Breast - Tumor Size (1998+)	Tumor size in millimeters taken the breast	15mm	0.039

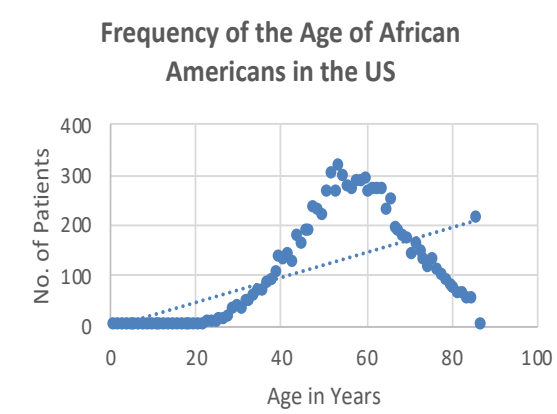
Urban, recent immigrant, single, lack of private insurance, lower household income and younger age are top attributes that are unique for African American TNBC patients

What attributes are unique for African Americans?

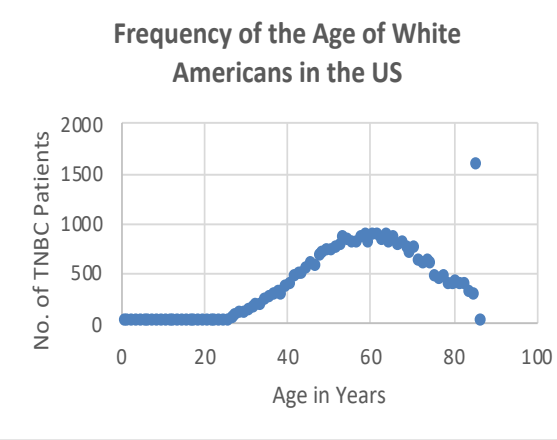
Statistical validation

Age at diagnosis

AA: **53** years old

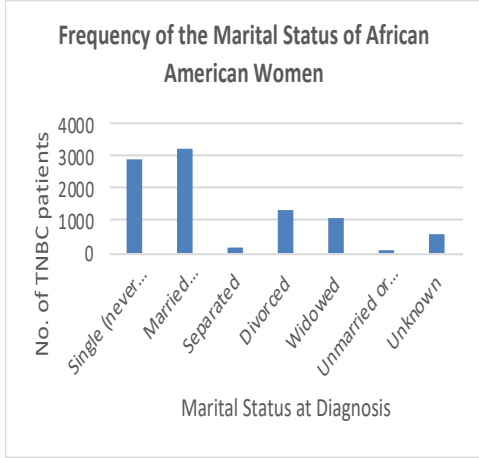


WA: **60** years old

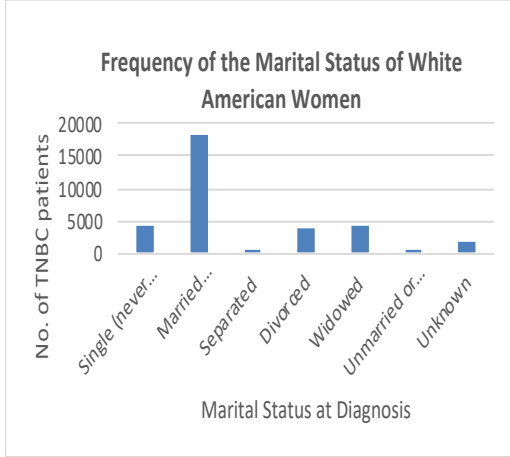


Marital Status at diagnosis

AA: **31%** single



WA: **10%** single



Take home message #2

- **Urban**
- **Recent immigrant in past 5 years**
- **Single**
- **Lack of private insurance**
- **Lower household income**
- **Younger age**

The above factors may be linked to health disparity in triple negative breast cancer

How could our findings help reduce the TNBC disparities?

- African American women would benefit from screening for TNBC beginning at a younger age
- Special attention should be given to the following African American women
  - With diabetes
  - Urban
  - Recent immigrant
  - Single
  - Uninsured
  - Low household income
- Genes identified in this project could serve as novel targets for personalize TNBC treatment

Acknowledgements

Students:
Charlene Radler
Michelle Del Valle
Alexa Calcagno
Emilio Ahuactzin

Postdoctoral Fellows:
Hamed Hayatshahi
Zhicheng Zuo

Collaborators:
Chao Xing (UT Southwestern)
Amin Morid (U of Utah)
Shouyi Wang (UT Arilington)

UNT | HEALTH SCIENCE CENTER
TEXAS CENTER FOR HEALTH DISPARITIES

NIH National Institute on Minority Health and Health Disparities