Benefit from using Big Data to Enhance Genomic and Cancer Health Disparities Research

2019 Professional Development Workshop and Mock Review

June 3 - 4, 2019

NIH Natcher Conference Center | Bethesda, MD

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H, M.S.

Assistant Professor

USC Institute Of Translational Genomics

Keck Medicine of USC

OUTLINE

My Timeline:

-Big Data
-Bioinformatics
-ML / AI

Benefit of learning Big Data:

-Suggested Tools/Software

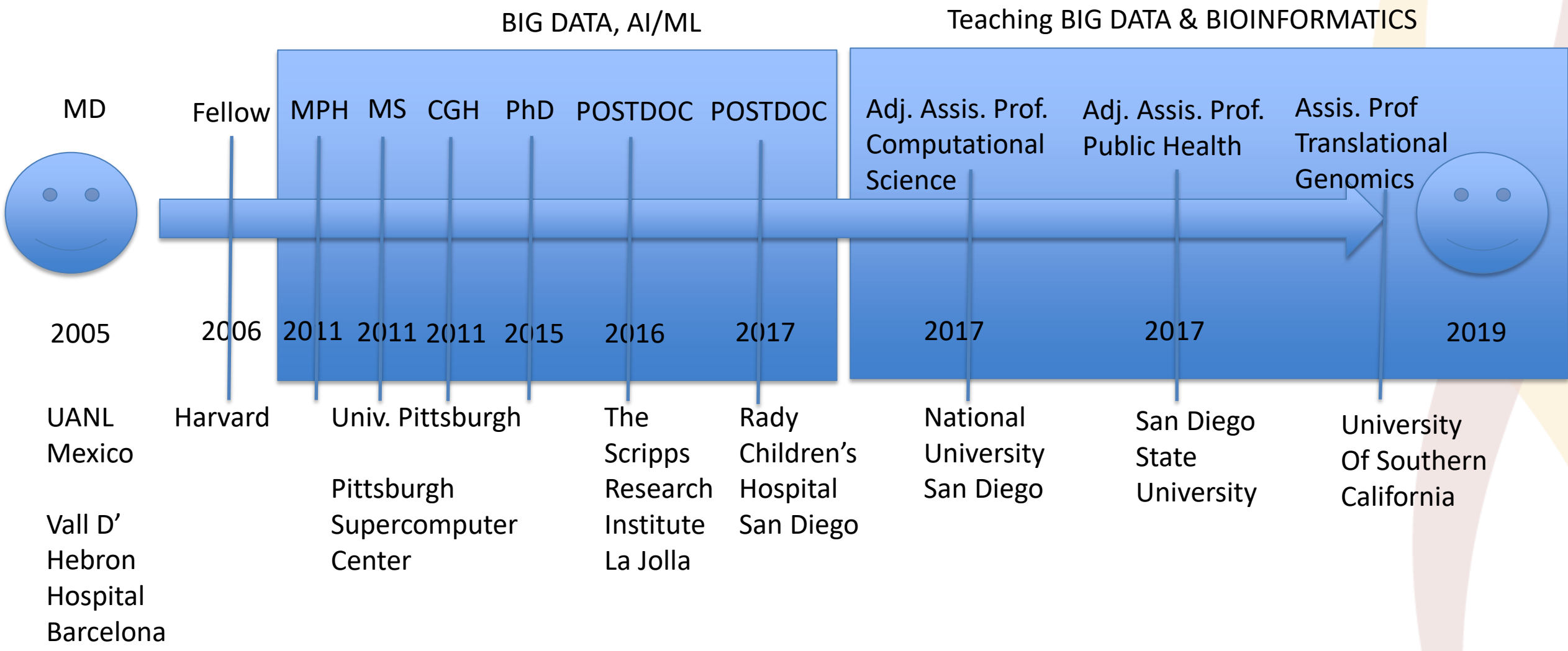
Benefit of using Big Data in Genomic and CHD Research:

-Endometrial Cancer
-Bulk, Single Cell & Spatial Sequencing
- Data Integration

Conclusions:

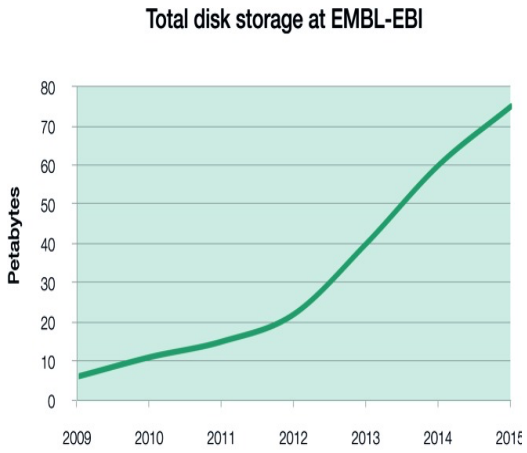
Big Data Enhances Genomic and Cancer Health Disparities Research

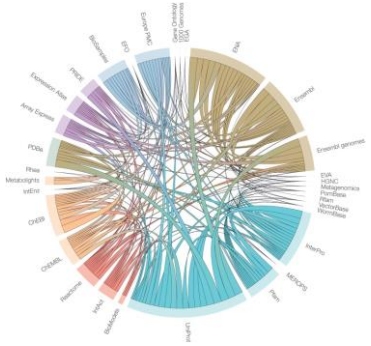
My Timeline

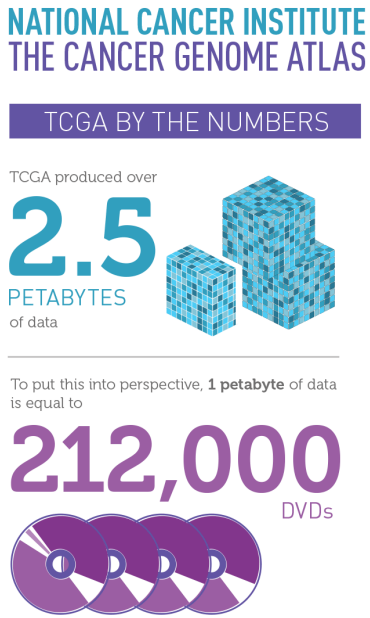

BIG DATA, AI/ML
Teaching BIG DATA & BIOINFORMATICS
MD
Fellow
MPH
MS
CGH
PhD
POSTDOC
POSTDOC
Adj. Assis. Prof. Computational Science
Adj. Assis. Prof. Public Health
Assis. Prof Translational Genomics
2005
2006
2011
2011
2011
2015
2016
2017
2017
2017
2019
UANL Mexico
Vall D' Hebron Hospital Barcelona
Harvard
Univ. Pittsburgh
Pittsburgh Supercomputer Center
The Scripps Research Institute La Jolla
Rady Children's Hospital San Diego
National University San Diego
San Diego State University
University Of Southern California

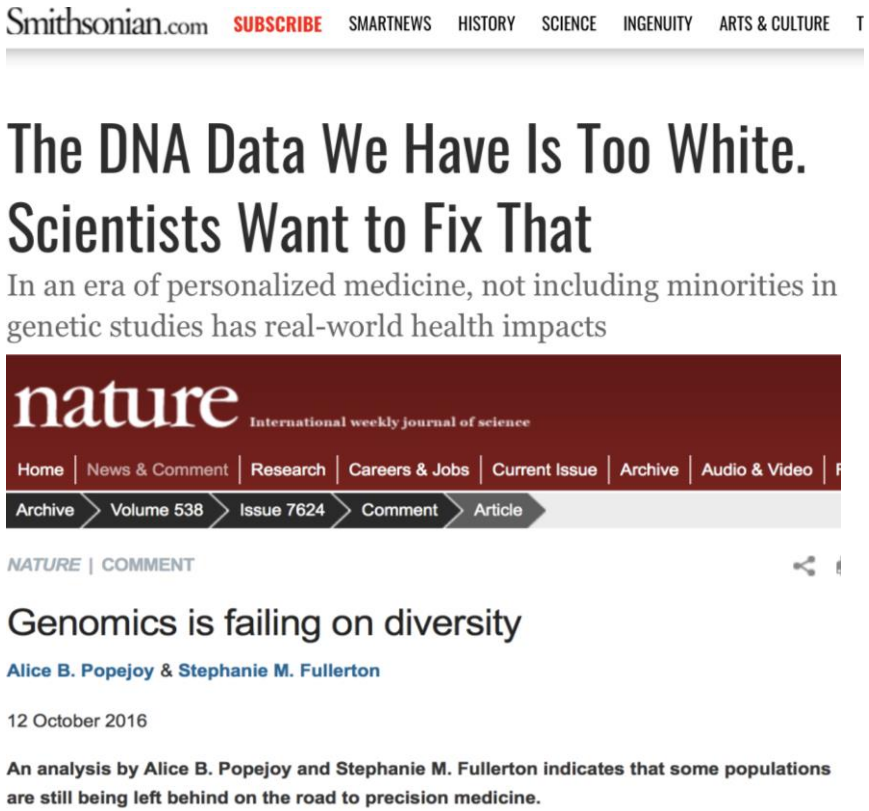
Big Data

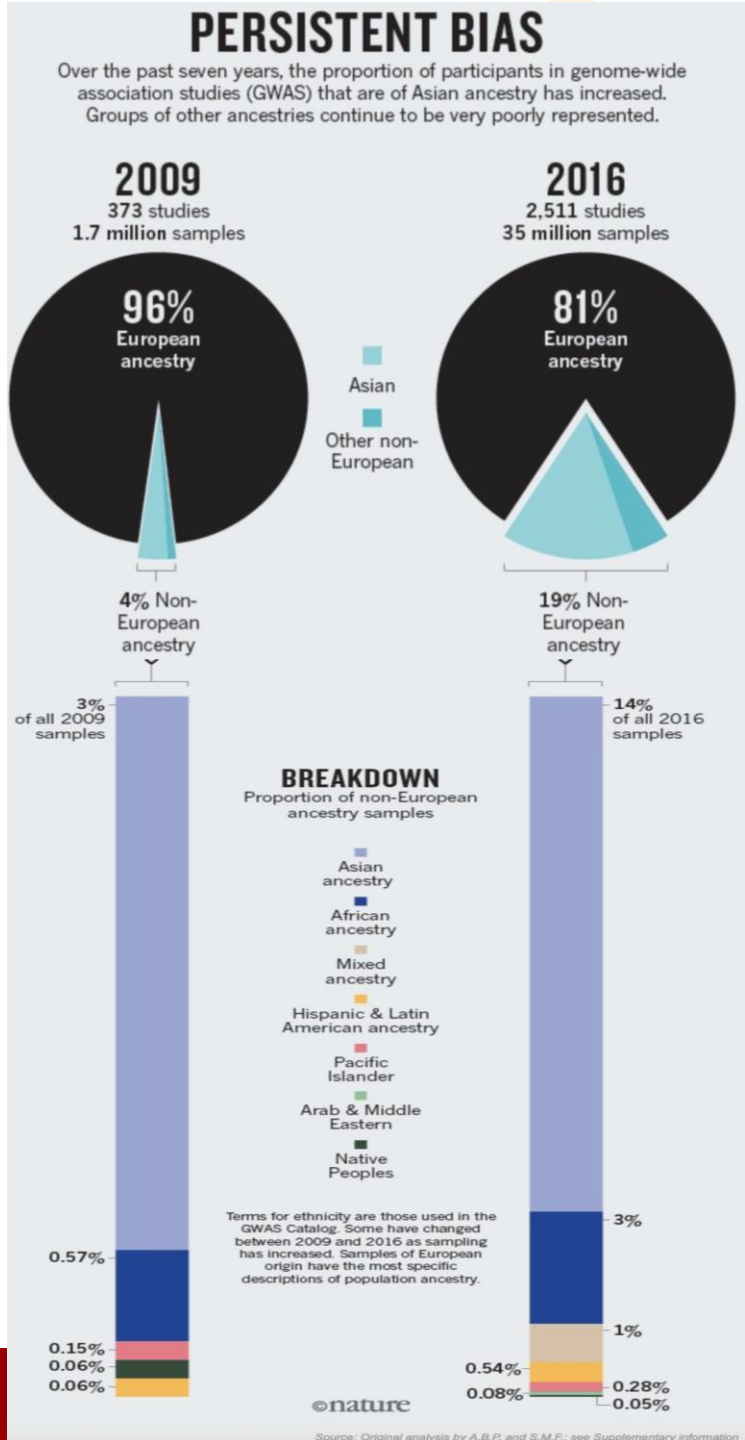
- Improvements in medical and genomic tech. have dramatically increased the production of electronic data over in the 21ST Century





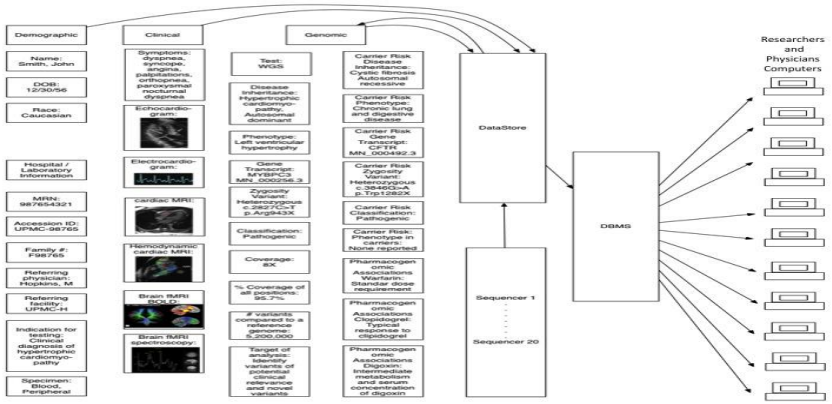


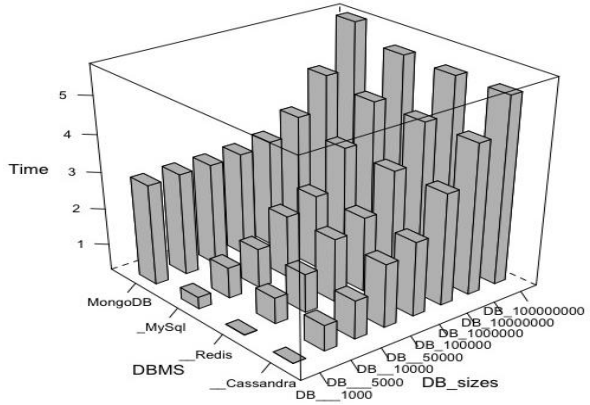




DATA SCIENCE

- Data management and data analysis is becoming essential in Cancer research.



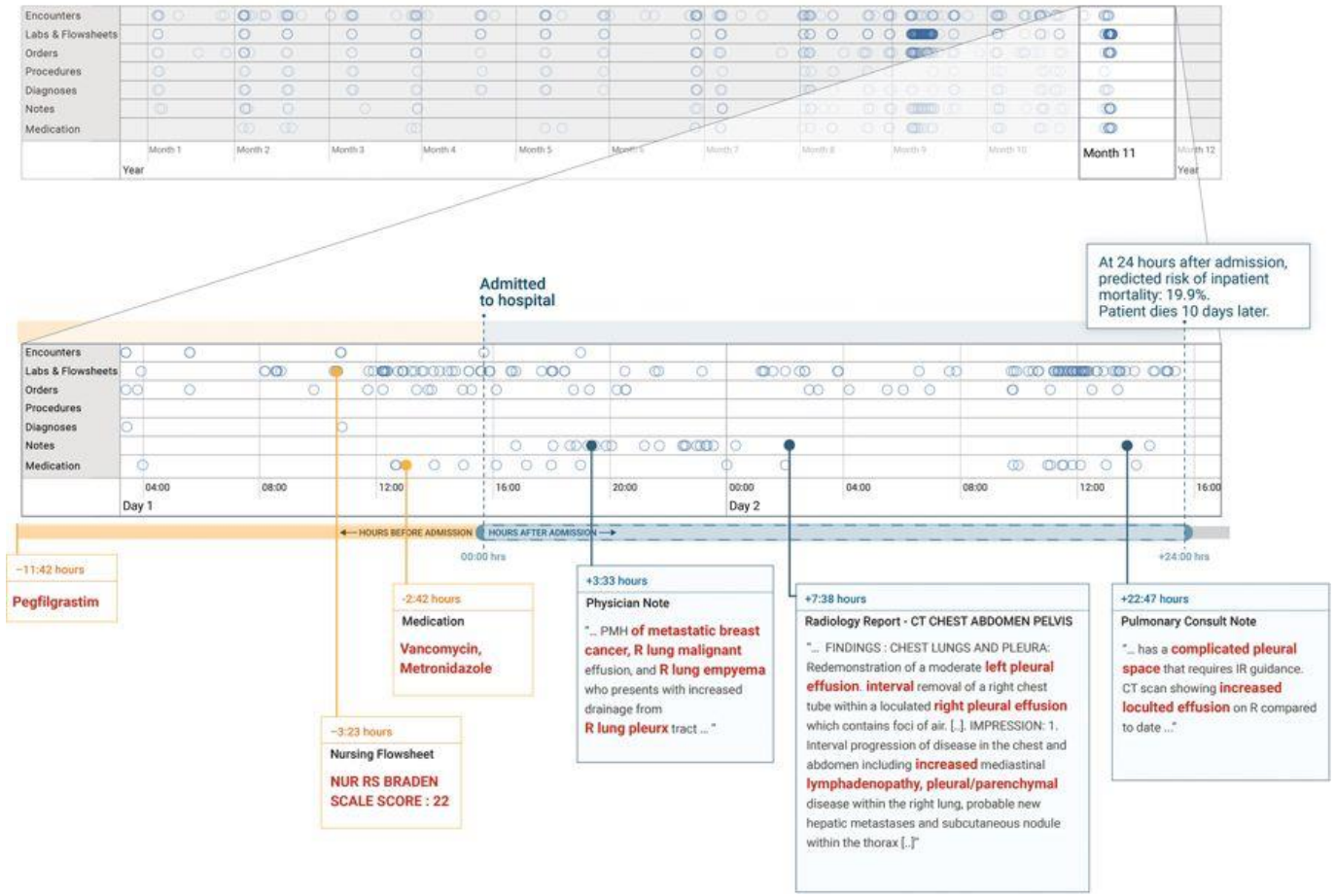


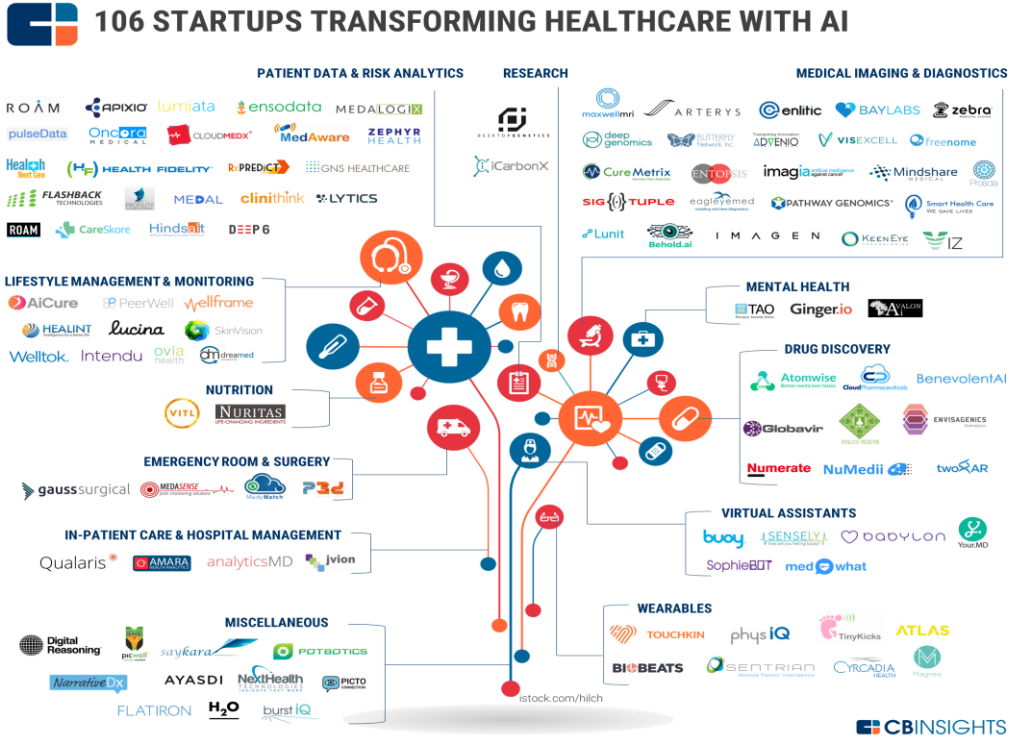
npj | Digital Medicine

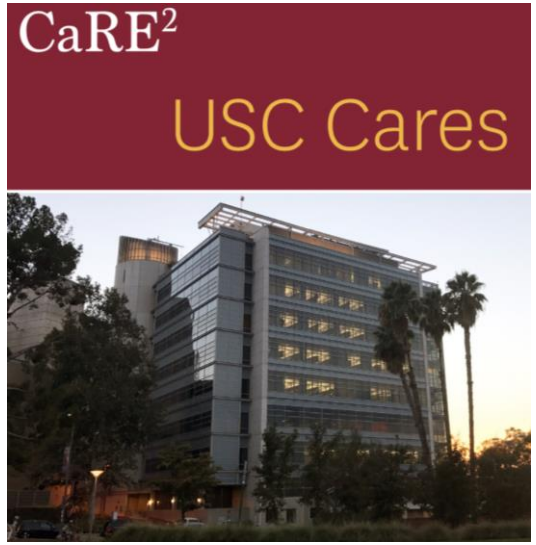
www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records







Bioinformatics Statistical and Methodological core:

- HPC
  - 21,000 cores
  - 64 terabytes of RAM
  - 2 petabytes of disk storage

Maximum speed of 157 teraflops = 157 trillion floating point operations

SUGGESTED BIG DATA RELATED TOOLS

Introduction to object oriented programming R Python	Introduction to terminal Linux/Unix	Introduction to databases SQL	Introduction to open source software (BI) Bioconductor (R) - TCGAbiolinks
Introduction to open source software (BI) Bioconductor (R) - TCGAWorkflowData	Introduction to open source software (BI) Biopython (Python)	Introduction to building pipelines (BI) BWA, SAMtools,TopHat, FreeBayes, CuffLinks	Introduction to web services Amazon, Google Cloud Comp Services
Introduction to open source software (ML) R Caret Package	Introduction to AI resources Google AI	Introduction to databases NoSQL	Introduction to AI resources: Watson IBM

Big Data in Genomic and CHD Research

Females		
Breast	24,000	29%
Thyroid	6,800	8%
Uterine corpus	6,700	8%
Colon & rectum	6,500	8%
Lung & bronchus	5,000	6%
Non-Hodgkin lymphoma	3,500	4%
Kidney & renal pelvis	3,200	4%
Leukemia	2,500	3%
Ovary	2,500	3%
Uterine cervix	2,400	3%
All sites	81,700	100%

- Bulk & Single Cell Sequencing

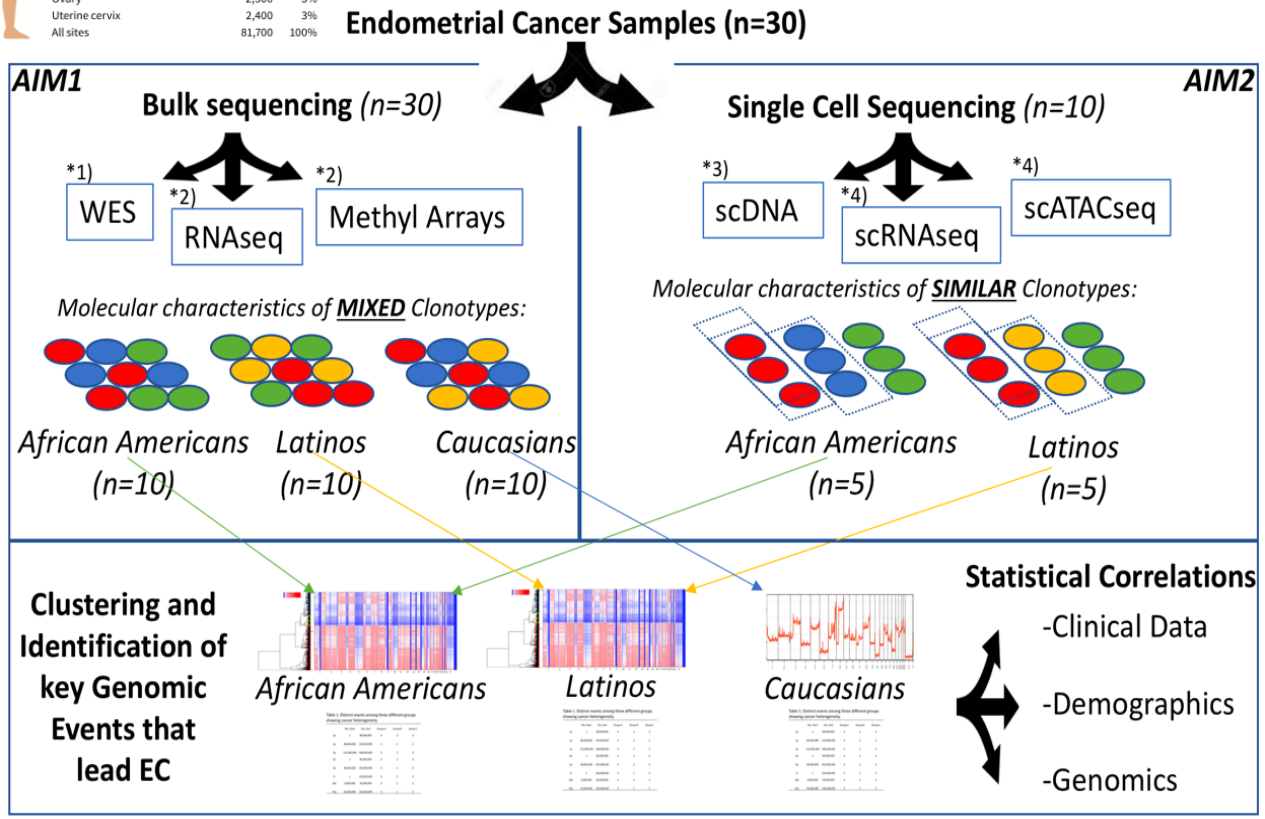


Figure 1. Structurally and functionally multi-genomics bulk and single-cell characterization of EC. At bulk genomic sequencing resolution (**AIM 1**)(n=30), it will be identified, *1) Structurally: genomic variations using *WES*, and, *2) Functionally, differential gene expression (DGE) through *RNAseq* and activity of DNA segments using *DNA Methylation arrays*. At single-cell level resolution (**AIM 2**)(n=10), it will be identified, *3) Structurally, genome heterogeneity and clonal evolution using *scDNA-CNV*, and, *4) Functionally, DGE using *scRNAseq* and identified accessible DNA regions, signatures, of the DNA-binding proteins through *scATACseq*. Clustering and correlations will be performed accordingly.

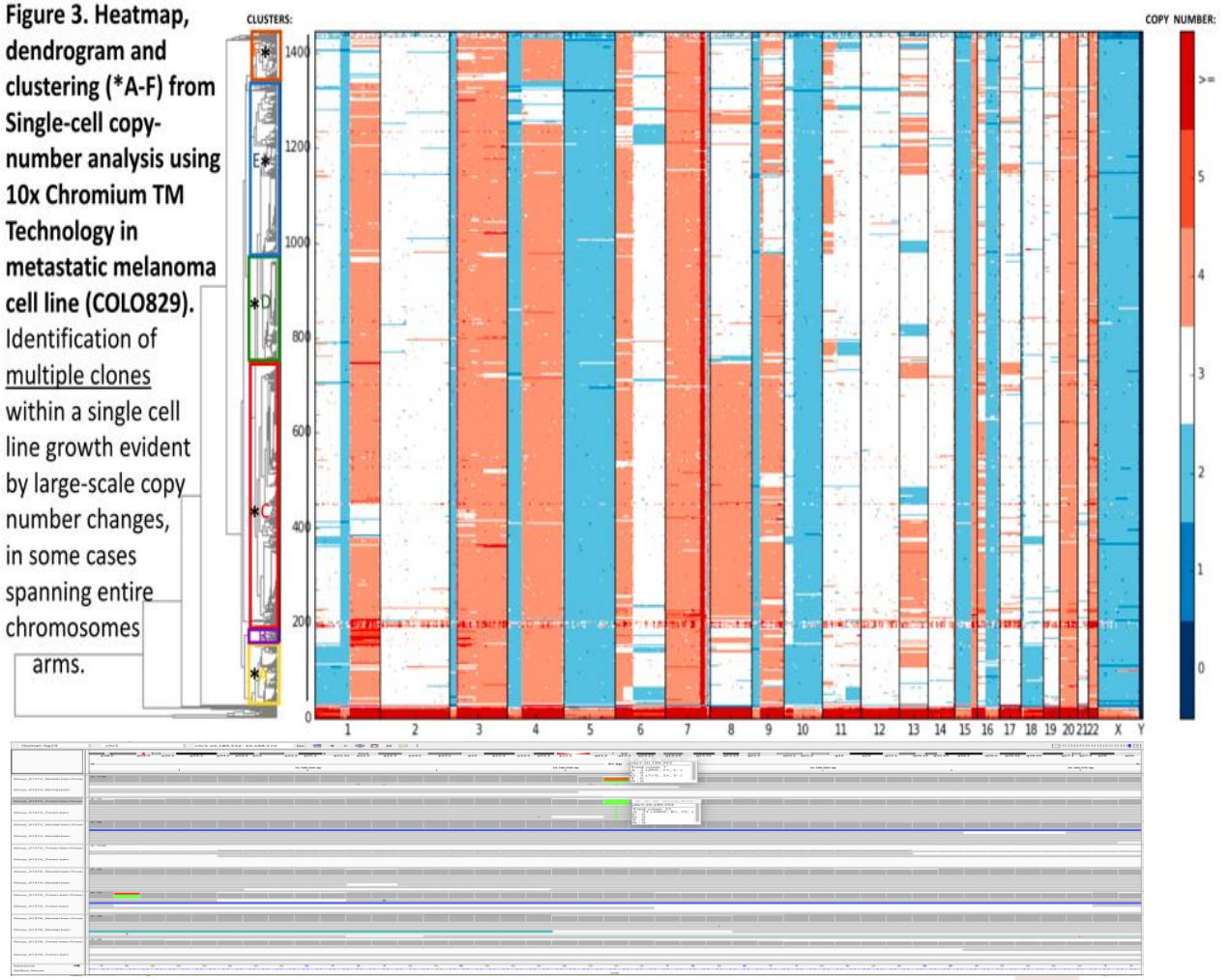


Figure 3. Heatmap, dendrogram and clustering (*A-F) from Single-cell copy-number analysis using 10x Chromium TM Technology in metastatic melanoma cell line (COLO829). Identification of multiple clones within a single cell line growth evident by large-scale copy number changes, in some cases spanning entire chromosomes arms.

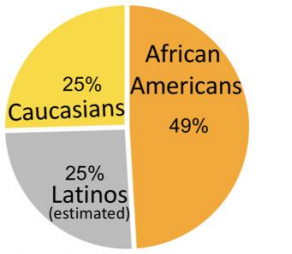
Quick Time video not available online.

BIG DATA IN GENOMIC AND CHD RESEARCH

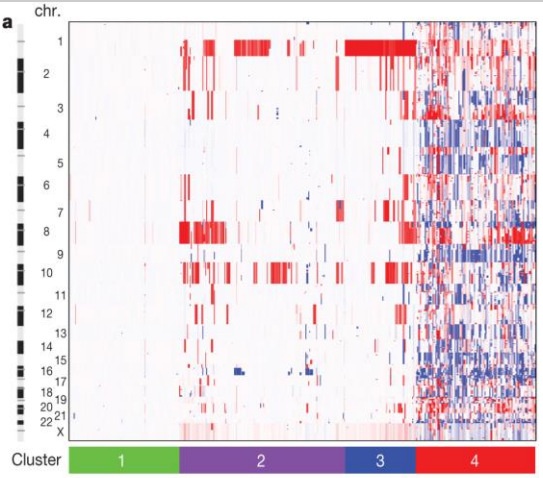
Table 1. Genomic, phenotypic, demographic and clinical characteristics of EC tumors. Histological subtypes include endometrioid adenocarcinoma (EAc), mucinous adenocarcinoma (MAc), serous cell type (Ser) and clear cell type (Clr)

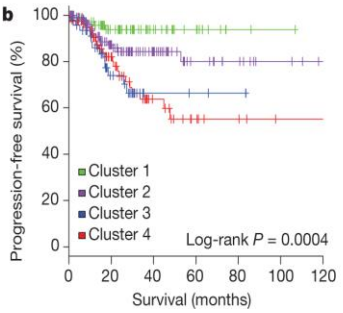
Gene/Signatures	Function	Histology	Prevalence	Age	Prognosis	Reference
p53/high CNV	*Tumor suppressor*	*Ser & Clr*	*Higher in AA*	*Older*	*Poor*	*12,13*
KRAS/low CNV	Cell proliferator	EAc, MAc	Higher in Caucasians	Young	Good	12,14
PTEN/low CNV	Tumor suppressor	EAc, MAc	Higher in Caucasians	Young	Good	12,14
PIK3CA/low CNV	Cell proliferator	EAc, MAc	Higher in Caucasians	Young	Good	12,14
PI3K AKT/low CNV	Cell cycle regulator	EAc, MAc	Higher in Caucasians	Young	Good	12,14
N/A	N/A	EAc, MAc	Higher in Latinos	Young	Good	15

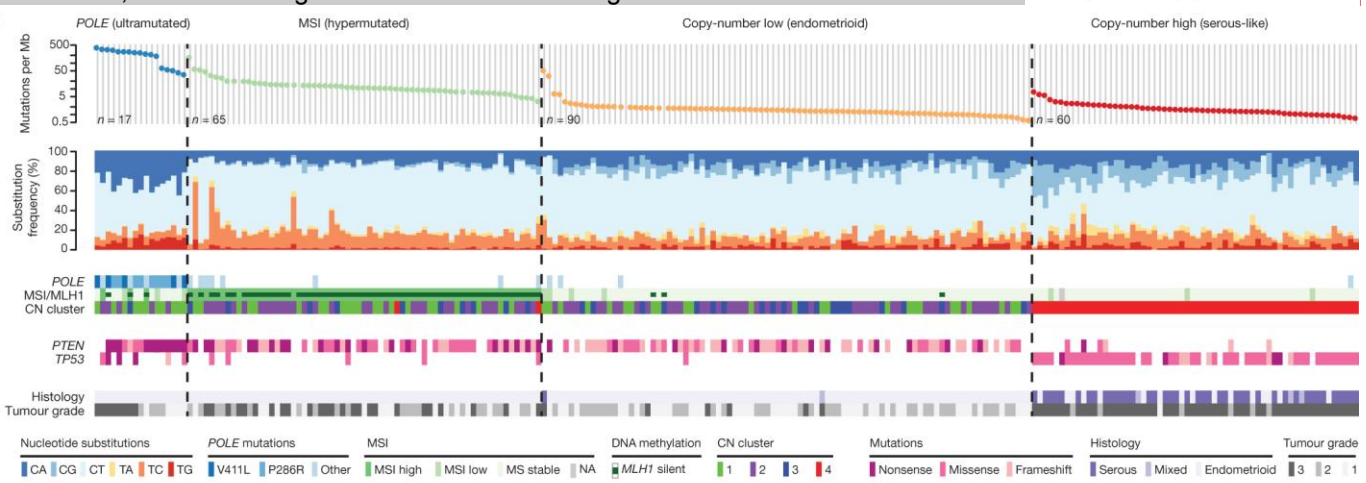
Figure 2. Death Rates by Race

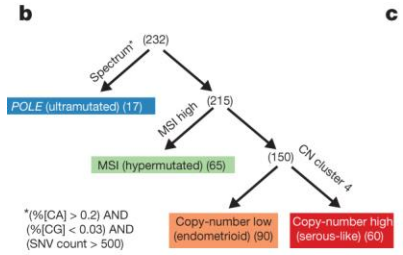


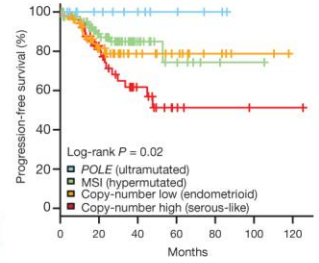
*per 100,000 population in the US

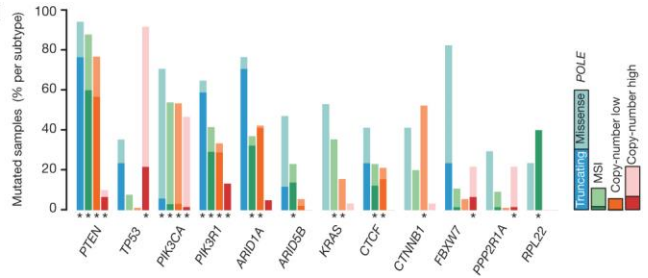












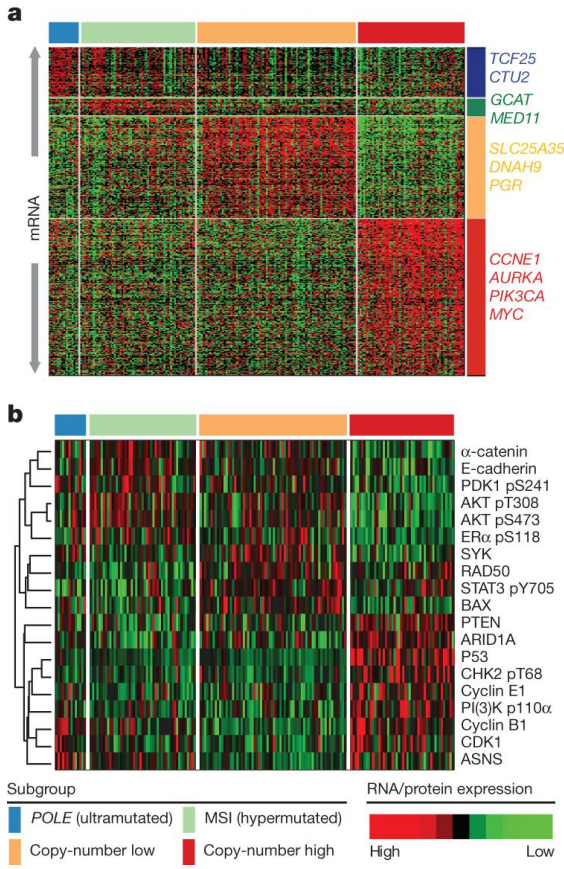
CHD-Research:

Tumor- associated genetic signatures are known to be associated with poorer prognosis of patients with **endometrial cancer** among African Americans, Latino and Caucasians.

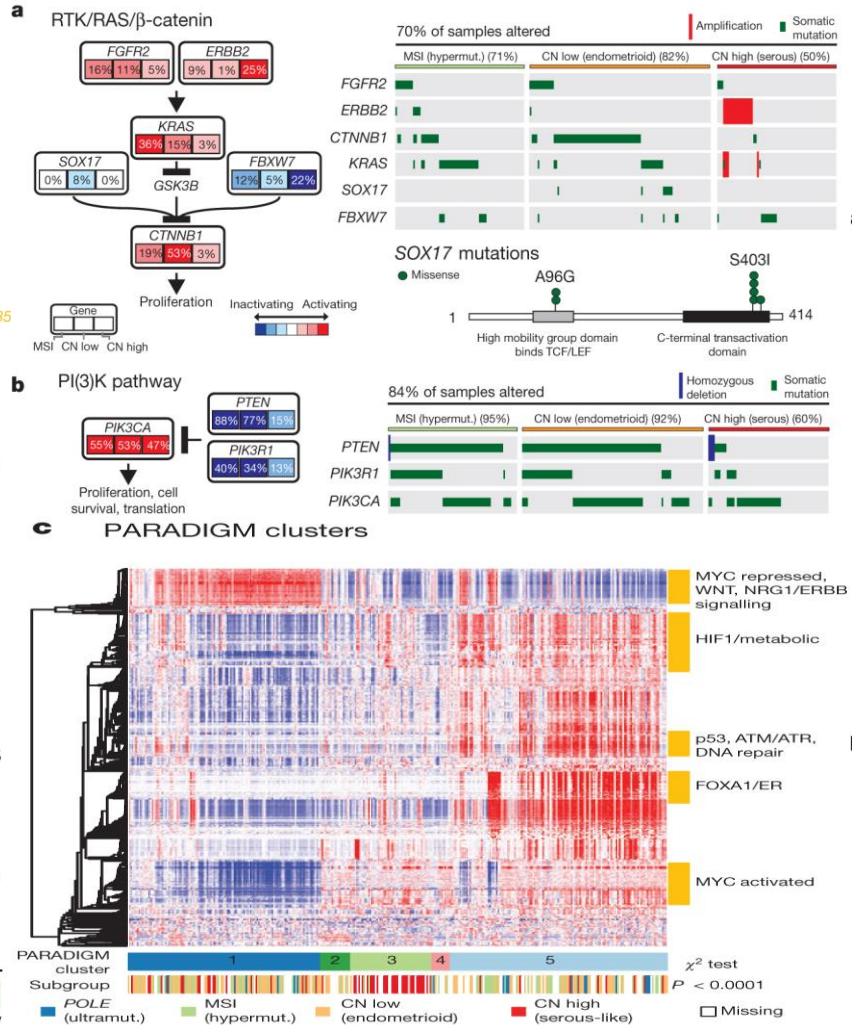
African Americans have **double the mortality** of Caucasians and probably Latinos and their tumors tend to be of higher grade; they also have worse survival comparing to Caucasians and Latinos.

BIG DATA IN GENOMIC AND CHD RESEARCH

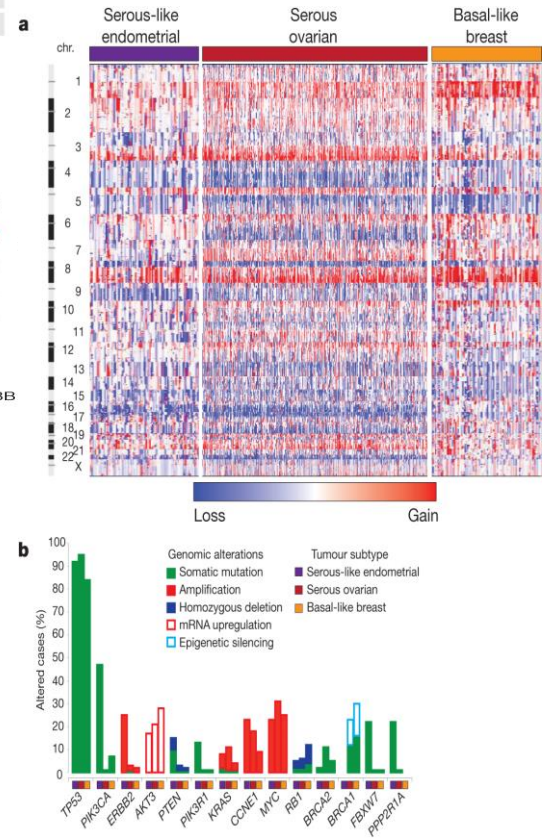
Gene expression across integrated subtypes in endometrial carcinomas:



Pathway alterations in endometrial carcinomas:



Genomic relationships between endometrial serous-like, ovarian serous, and basal-like breast carcinomas:



Little is known regarding the molecular characterization of EC among racial groups.

The most frequent gene alterations in African Americans are enrichments of p53 mutations with **high CNV**; whereas KRAS, PTEN, PIK3CA and PI3K/AKT mutations with **low CNV** are more frequent in Caucasians.

CONCLUSIONS

- Big Data benefit in:
  - Better understanding big picture in CHD research through data integration
  - Generating novel hypothesis through hypothesis-driven analysis
  - Improving current research by increasing the load of information and running more complex and accurate data analyses

THANK YOU!

USC Department Of Translational Genomics

Keck Medicine of USC