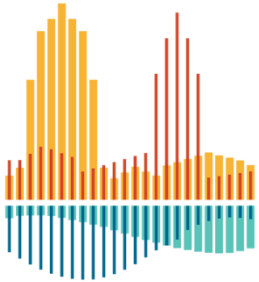
Extracting Country-of-Origin from electronic Health Records for gene-environment studies as part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE)



INSTITUTE FOR
COMPUTATIONAL
BIOLOGY

March 28, 2017

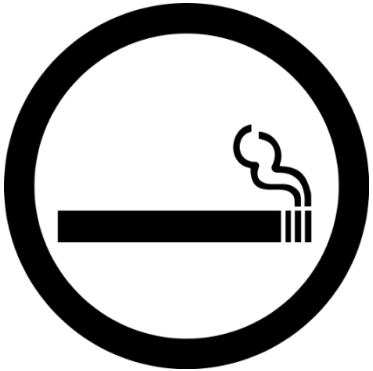
Dana C. Crawford, PhD
Associate Professor
Epidemiology and Biostatistics
Institute for Computational Biology

GxE Studies

- Lifestyle
- Behaviors
- Exposures
- Social context

https://www.niehs.nih.gov/health/topics/science/gene-env/

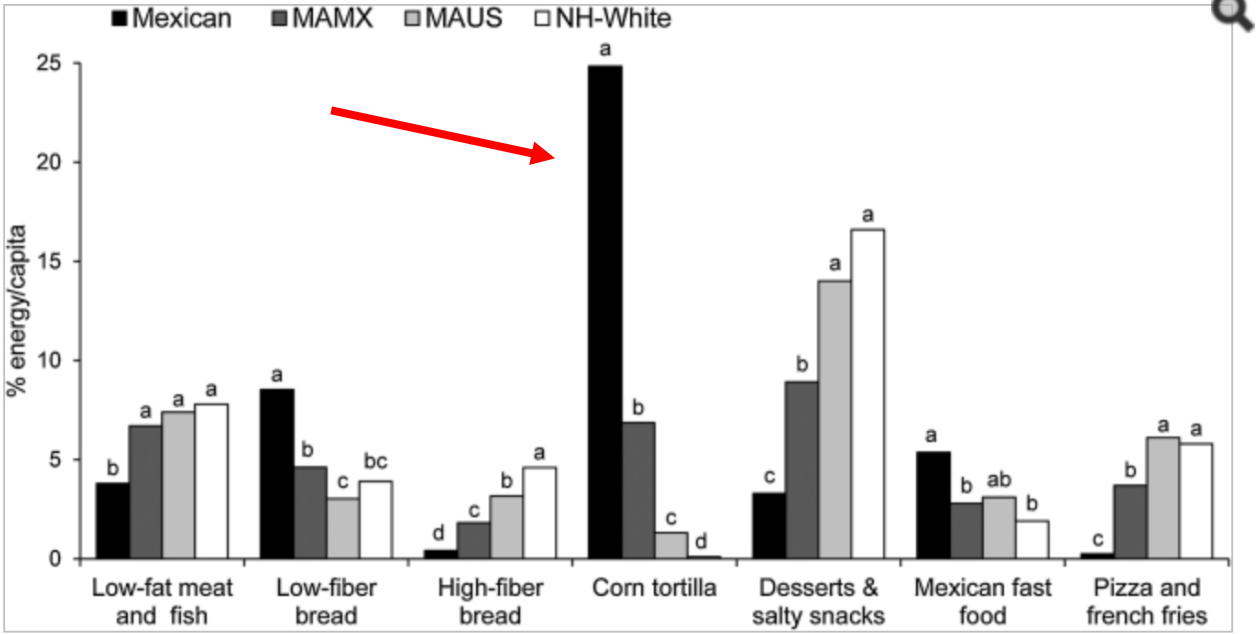
EHRs and Unstructured Data



Acculturation

- Modification of beliefs and/or behaviors by an individual or group when introduced to another group or environment

Batis et al (2011) *J Nutr* 141(10):1898-1906

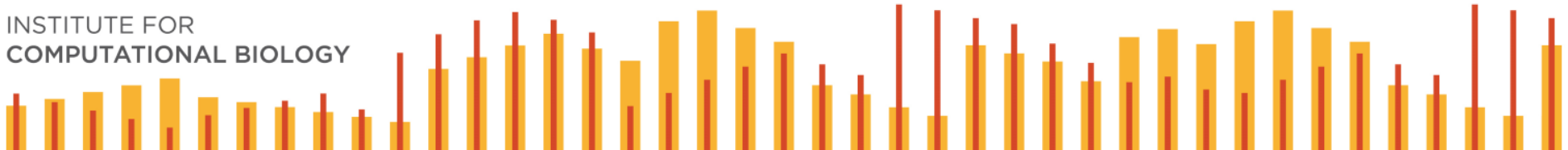


INSTITUTE FOR COMPUTATIONAL BIOLOGY

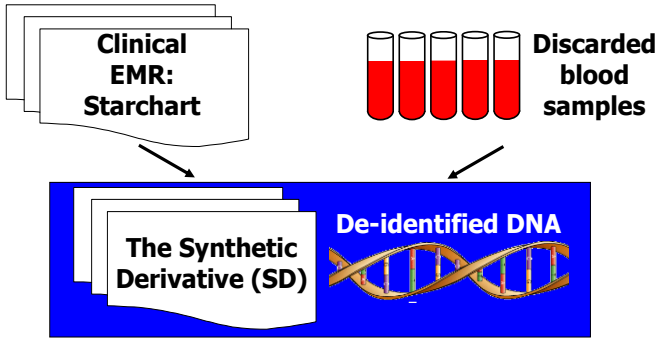
Country-of-Origin

- Current EHR labels available for research are broad racial/ethnic group labels
  - Some are *not* self-identified race/ethnicity
  - Label gives no information on where patient was born

Can we mine clinical text for clues to country-of-origin?



VUMC BioVU



- Opt-out model (2007-2015)
- DNA collected from discarded blood after routine clinical testing has been completed
  - Matched with clinical and demographic data within de-identified EHR (“Synthetic Derivative” database)
  - >225,000 DNA samples

INSTITUTE FOR COMPUTATIONAL BIOLOGY

VUMC EHR

VUMC in Nashville, TN

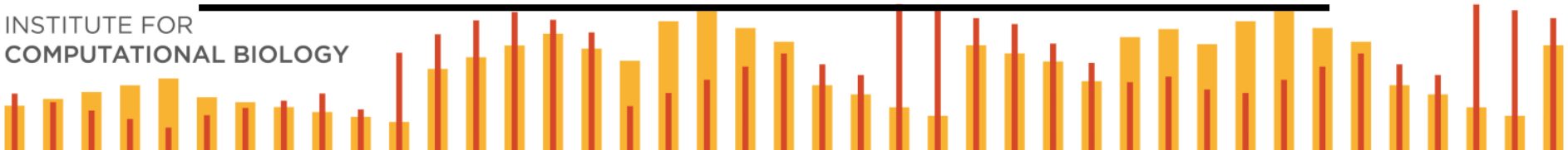
- StarChart
  - Designed, built, and maintained by faculty-led teams
  - Being replaced with EPIC (2018)
- >2 million records, including order entry data on inpatients since 1994
- A document-centric architecture
  - Structured (ex. ICD-9-CM and ICD-10-CM codes)
  - Unstructured (ex. clinical notes)

EAGLE BioVU

	n=15,863
Female (%)	10,050 (63.35)
Median (SD) age	37 (20.46)
African American (%)	11,521 (73.06)
Hispanic (%)	1,714 (10.87)
Asian (%)	1,122 (7.12)
Other (%)	1,412 (8.95)

EAGLE
Epidemiologic Architecture for Genes Linked to Environment

INSTITUTE FOR
COMPUTATIONAL BIOLOGY



Country-of-Origin Extraction

- Text search for 231 world countries, including independent sovereign states, dependent areas, and disputed territories (2013)
- Include common misspellings
- Output 30 characters of flanking text

Guatemala

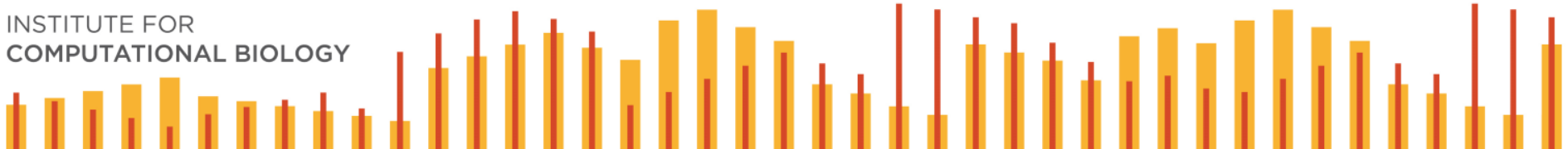
quatimala

guatamala

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Country-of-Origin Extraction

- Manual review of initial output to determine rules of filtering process
  - Average of 9.7 spellings per country
  - 77.5% countries on list had at least one mention
  - 52 countries not mentioned
    - Small (population) territories and countries
    - Large (North Korea, Burkina Faso, United Arab Emirates, Tajikistan, Kyrgyzstan, and Turkmenistan)



Country-of-Origin Extraction

Before filtering:

- 99.4% of patients were assigned a country
- Average 11.3 countries assigned per patient
- >93% of patients assigned to India, Italy, Chile, Greece, and China

Country-of-Origin Extraction

Why so many in TN from India, Italy, Chile, Greece, and China?

"Indi" misspelling for India

*indi*cated, contra*indi*cated, f*indi*ngs, *indi*vidual, etc.

"Ital" misspelling for Italy

hosp*ital*, dig*ital*ly, v*ital*, mar*ital*, gen*ital*, etc.

"Chil" misspelling for Chile

*chil*dren, *chil*ls, A*chil*les, etc.

"Gree" misspelling for Greece

Wal*gree*ns, a*gree*ment, de*gree*s, etc.

"Hia" misspelling for China

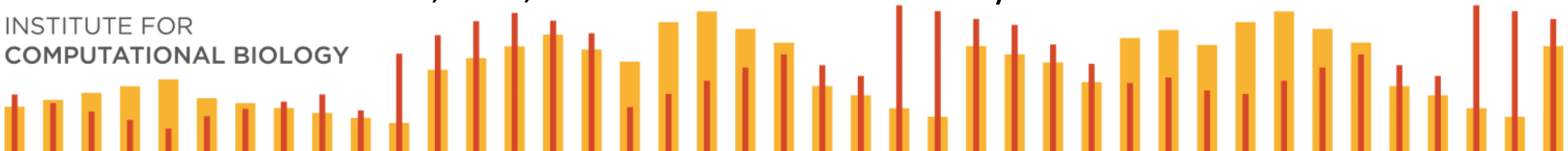
hydrochlorot*hia*zide, psych*hia*tric, brac*hia*l, etc.

INSTITUTE FOR COMPUTATIONAL BIOLOGY

Country-of-Origin Extractions

Other false positives

- Misspellings
  - “chile” (a misspelling for “child”)
  - “benign” (such as “benin”, “bening”, and “beningn”)
- Local Places
  - Lebanon, TN
  - Columbia, TN; Columbia University



Country-of-Origin Extractions

Other false positives

- Food, animals, and common phrases
  - Turkey sandwich
  - Guinea pig
  - Cooking grease
  - Is hungry

www.wikipedia.org

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Country-of-Origin Extraction

- Manual review of initial output to determine rules of filtering process
- Identified key words to help filter output to assign probable country-of-origin

Country-of-Origin Extraction

Grew

Up

Born

Live

From

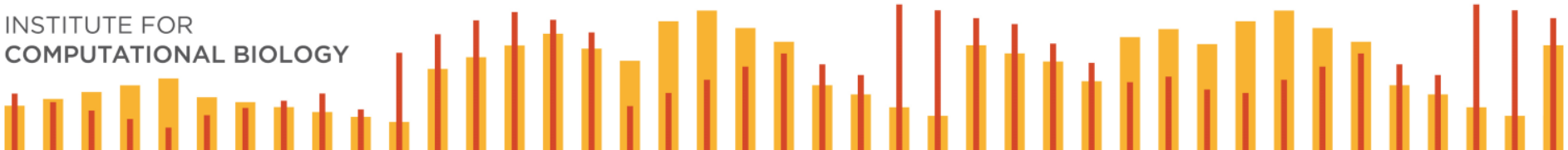
Orig

Grow

Raised

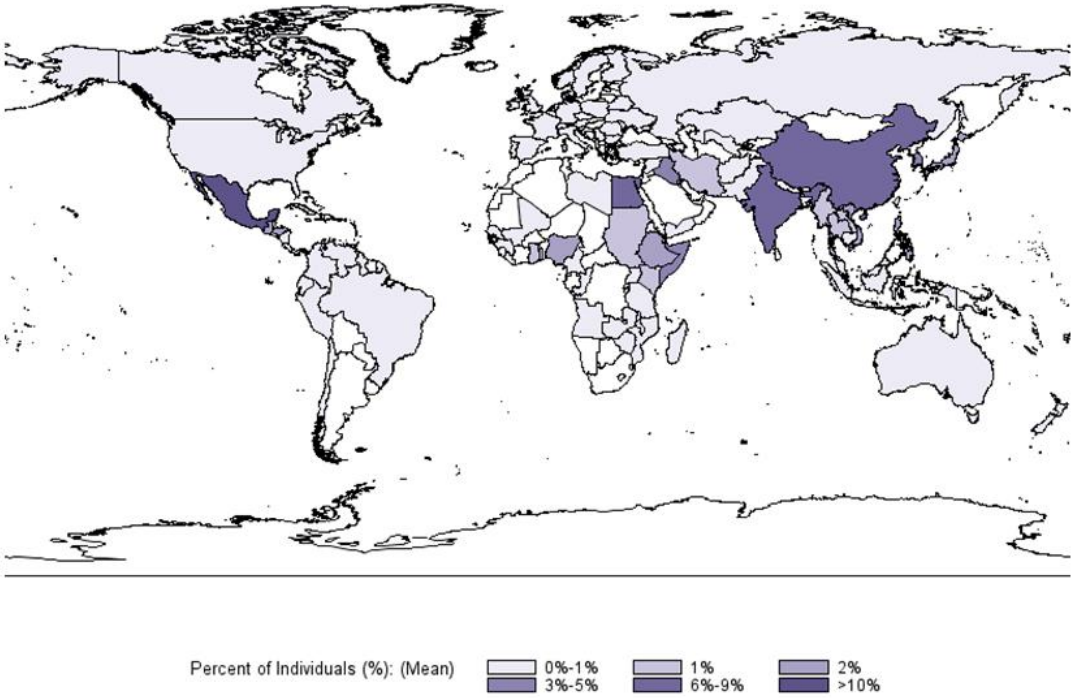
Home

Resided



Country-of-Origin Results

- After filtering, 10.9% had a country assigned
- US was default



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

African Americans (n=353)	Hispanics (n=438)	Asians (n=522)	Indians (n=60)	Others (n=539)
Nigeria (13.0%)	Mexico (55.9%)	China (28.9%)	India (46.7%)	Egypt (25.0%)
Somalia (10.5%)	Honduras (10.3%)	South Korea (11.7%)	Egypt (8.3%)	India (9.9%)
Ethiopia (7.0%)	Guatemala (9.6%)	Vietnam (8.0%)	Bangladesh (6.7%)	Iraq (9.7%)
Ghana (5.7%)	Cuba (3.7%)	India (6.5%)	Mexico (5.0%)	Somalia (8.2%)
Haiti (4.8%)	Peru (2.7%)	Japan (5.7%)	Afghanistan (3.3%)	Iran (4.0%)
Jamaica (4.8%)	Ecuador (2.5%)	Myanmar (5.6%)	Ethiopia (3.3%)	Ethiopia (4.0%)
Kenya (4.5%)	Nicaragua (2.5%)	Laos (5.4%)	Iraq (3.3%)	Sudan (3.0%)

Evaluation of Approach

- Manually reviewed notes of 16 patients with assigned country outside of US
  - 13 were clearly born outside the US
    - “originally from the Bahamas”
    - “He immigrated from Mexico”

www.wikipedia.org

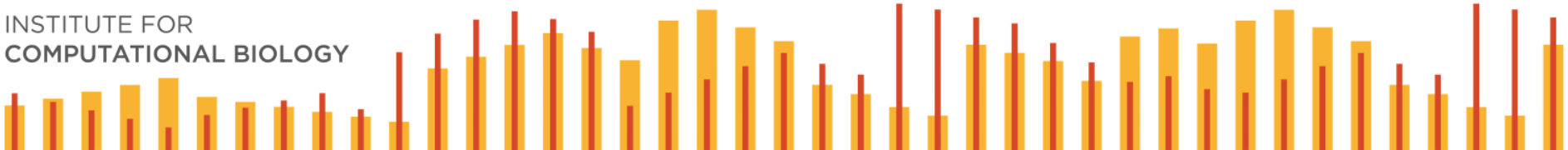
Evaluation of Approach

- Manually reviewed notes of 16 patients with assigned country outside of US
  - 1 probably a false positive
    - “Just returned from Australia and New Zealand”
    - Related to travel?

www.wikipedia.org

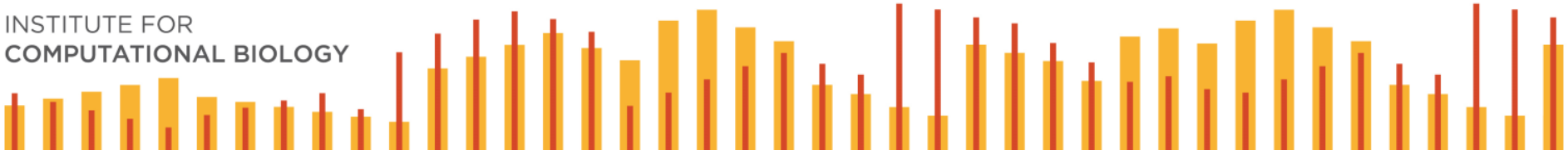
Evaluation of Approach

- Manually reviewed notes of 16 patients with assigned country outside of US
  - Another probably false positive?
    - “Foreign language ALBANIAN”
    - But is this an acceptable proxy for country-of-origin?



Evaluation of Approach

- Manually reviewed notes of 16 patients with assigned country outside of US
  - Another probably false positive?
    - “Mom is from Angola, speaks portugese" and “Maternal race: Angola”
    - Relevant, but where was patient born (same as mom)?



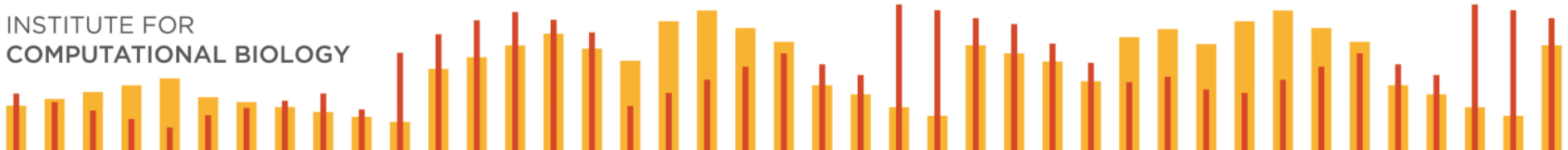
Evaluation of Approach

- Manually reviewed notes of 16 patients with assigned country outside of US

Positive Predictive Value (PPV):

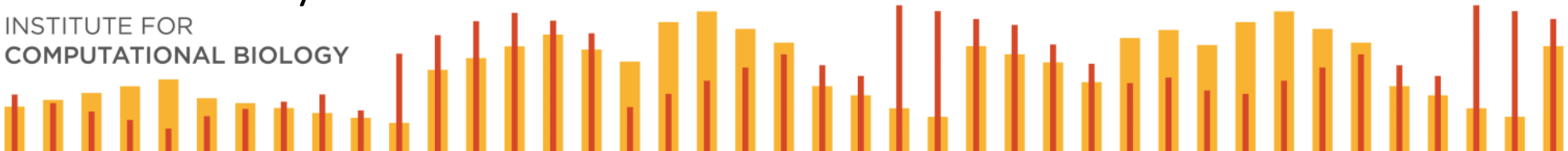
81.25 – 93.75%,

depending on how the three possible mismatches are classified



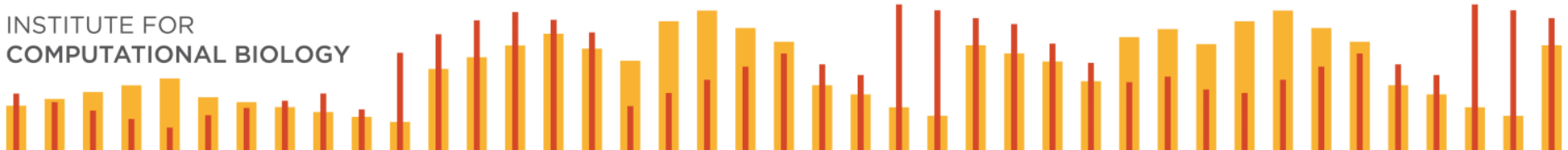
General Observations

- Only ~11% of patients had mention of a country
  - Physician prompted by an accent?
- Details vary by note taker
  - Few offer duration in US related to age of patient
  - May limit acculturation variable for research

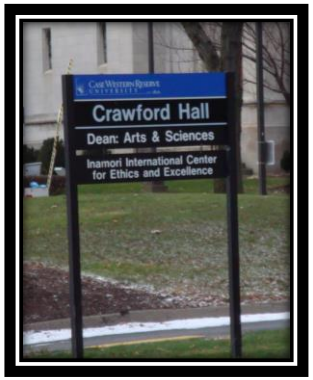


Conclusions and Future

- Simple extraction approach seems to work OK
- Can reduce number of records for manual review
- Can be run in parallel
- Needs to be applied to another set



Crawford Lab and Collaborators



Robert Goodloe, MS

Williams S. Bush, PhD, MS

Dana C. Crawford, PhD
Brittany Hollister, PhD candidate

NIH/NHGRI HG004798 (EAGLE)
NCATS 2 UL1 TR000445-06 (BioVU)

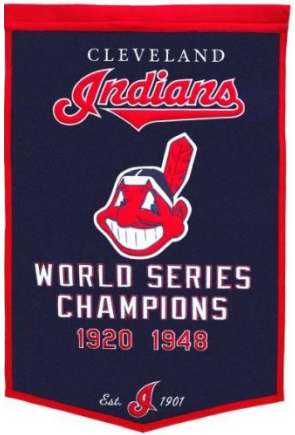
Eric Farber-Eger Jonathan Boston

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

CWRU and Cleveland

Biomedical Data Science open rank position available!

http://epbiwww.case.edu/



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

EHRs and Research

But…..

insight commentary

The case for a US prospective cohort study of genes and environment

Francis S. Collins

- Already several existing cohorts

- Very expensive

Collins (2004) *Nature* 429:475-477

EHRs and Research

American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2012.

Vol. 175, No. 9
DOI: 10.1093/aje/kwr453
Advance Access publication:
March 12, 2012

Commentary

New Models for Large Prospective Studies: Is There a Better Way?

Teri A. Manolio*, Brenda K. Weis, Catherine C. Cowie, Robert N. Hoover, Kathy Hudson, Barnett S. Kramer, Chris Berg, Rory Collins, Wendy Ewart, J. Michael Gaziano, Steven Hirschfeld, Pamela M. Marcus, Daniel Masys, Catherine A. McCarty, John McLaughlin, Alpa V. Patel, Tim Peakman, Nancy L. Pedersen, Catherine Schaefer, Joan A. Scott, Timothy Sprosen, Mark Walport, and Francis S. Collins

biobank uk
Improving the health of future generations

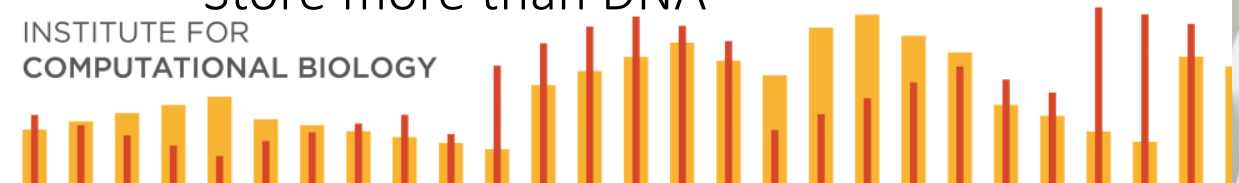
Manolio et al (2012) *Am J Epidemiol* 175:859-66

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

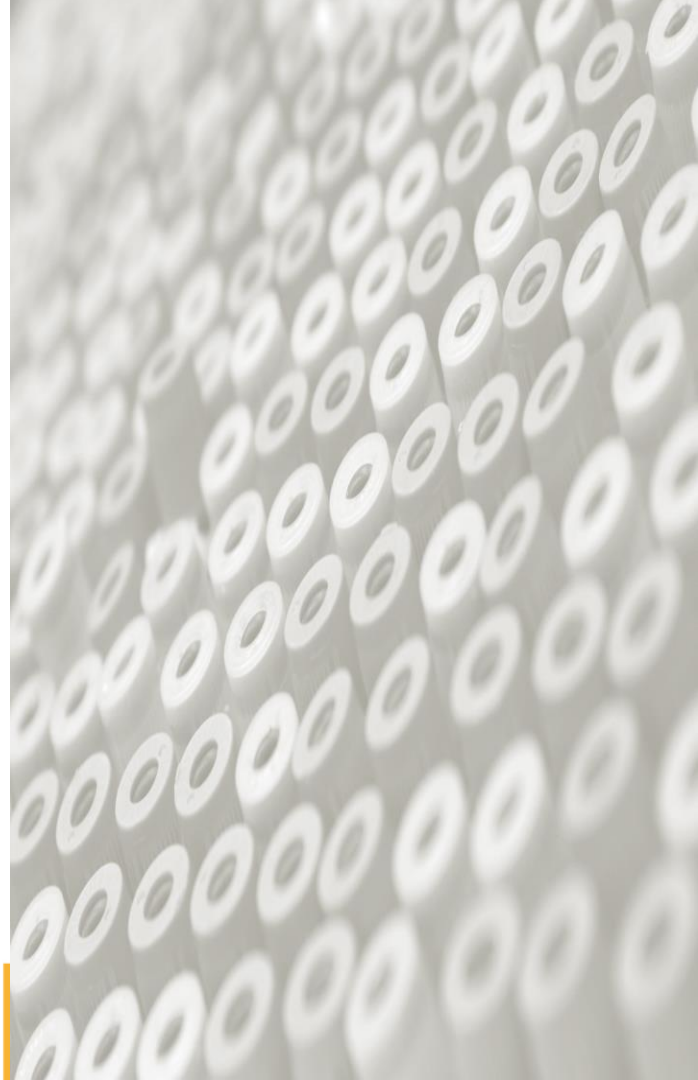
EHRs and Research

Biobanks linked to EHRs

- Large
- Relatively inexpensive compared with cohorts
- Some are prospective; population-based
- Store more than DNA



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

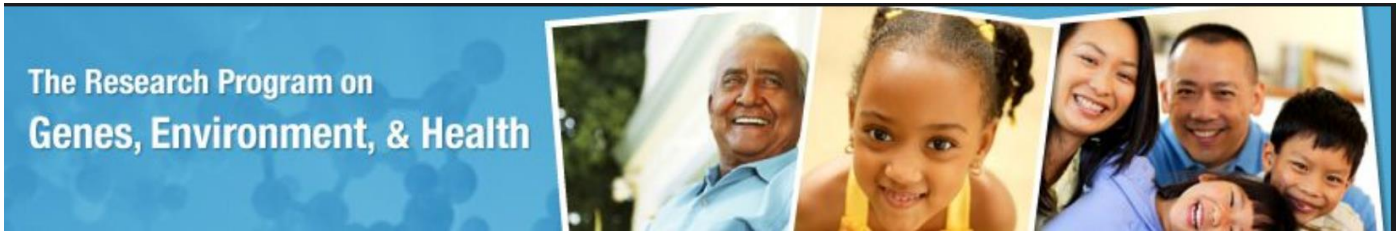


EHRs and Research

eMERGE Network
electronic medical records & genomics

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

EHRs and Research



KAISER PERMANENTE®

DiscovEHR

Geisinger

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

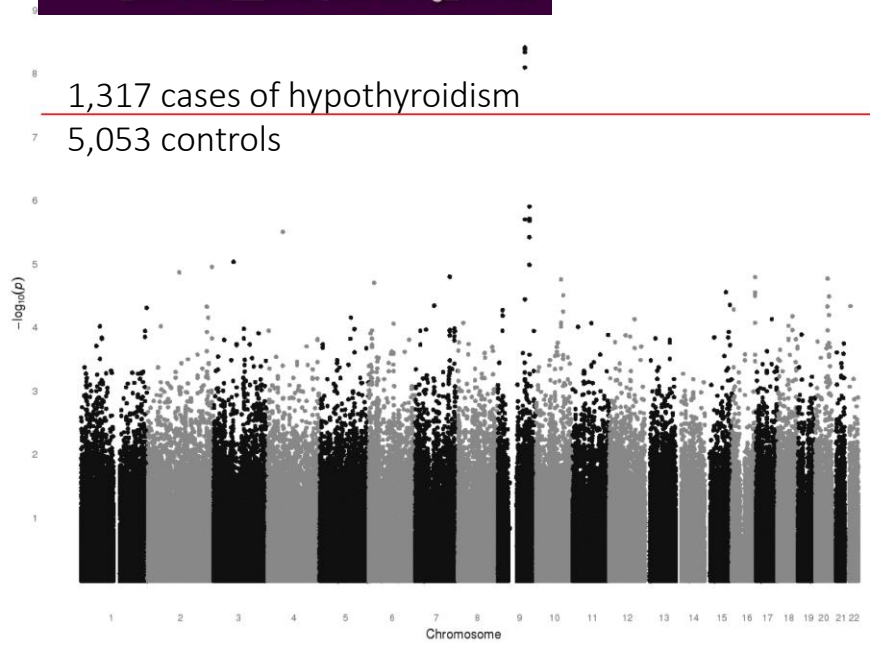
REGENERON™
GENETICS CENTER

EHR Structured Data and Genomics

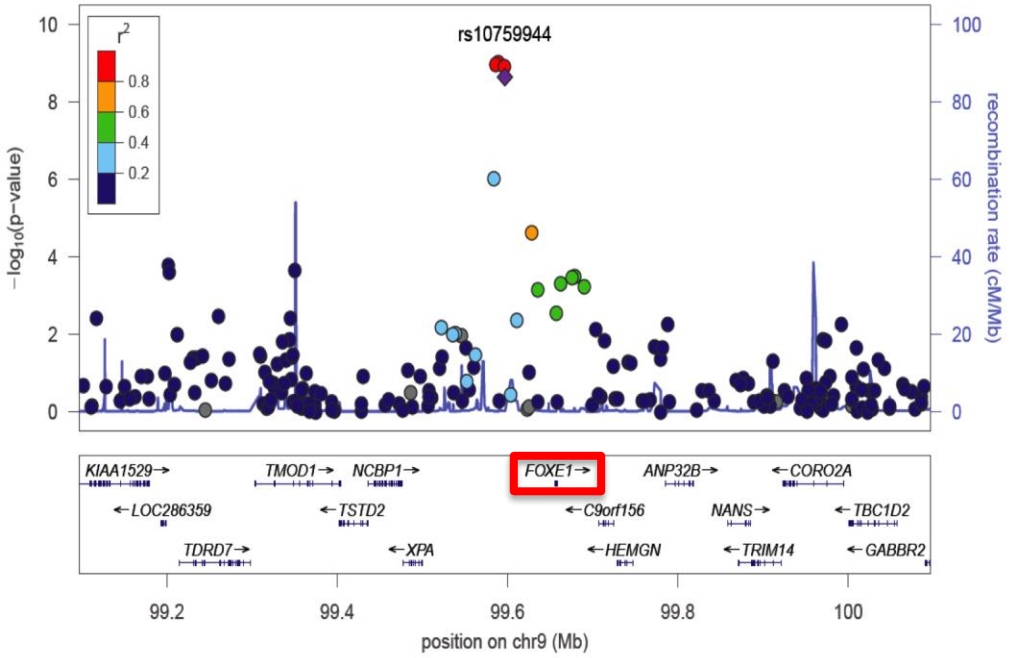
Billing codes (ICD-9-CM) for case/control definition

eMERGE Network
electronic medical records & genomics

1,317 cases of hypothyroidism
5,053 controls

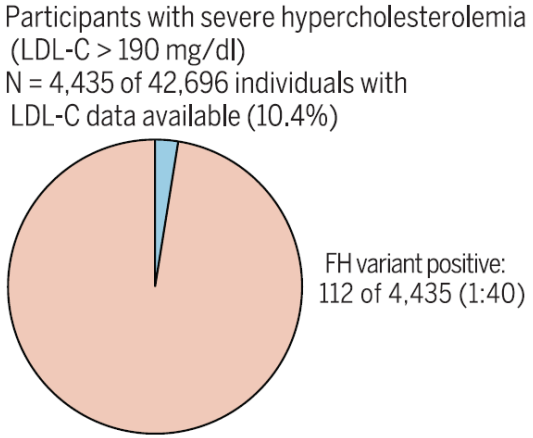


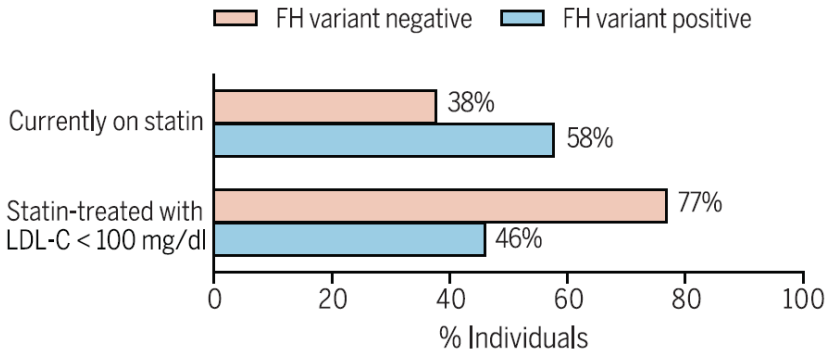
Denny*, Crawford*, et al (2011) *Am J Hum Genet* 89(4):529-42



EHR Structured Data and Genomics

Labs





Abul-Husn et al (2016) *Science* 354(6319)

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

DiscovEHR

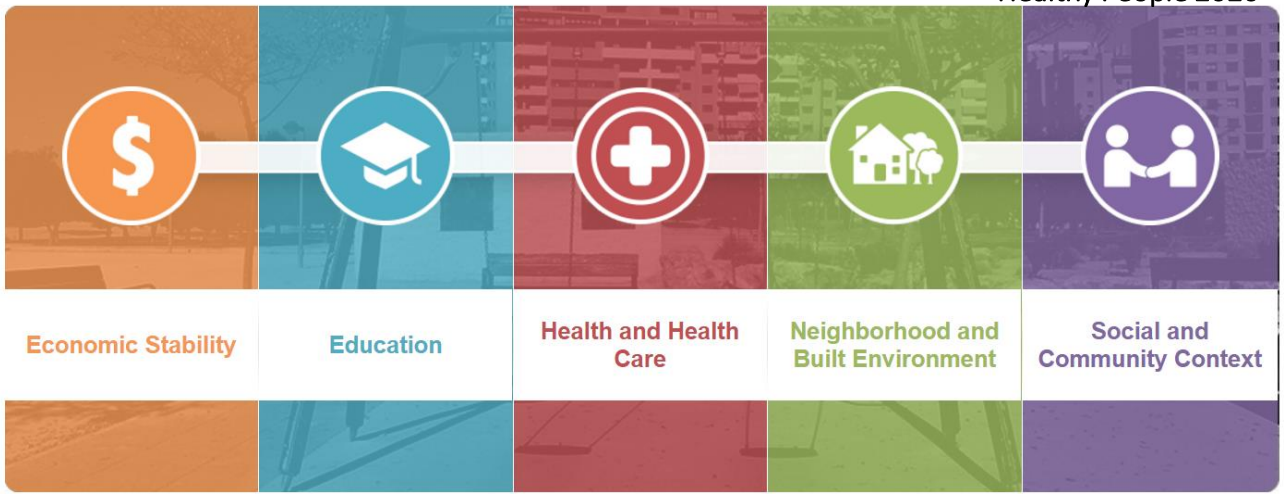
Geisinger

REGENERON GENETICS CENTER

EHR and Unstructured Data

Social Environmental Variables?

Healthy People 2020



Brittany Hollister

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

VUMC BioVU is clinic-based

	Davidson County (n=626,684)	BioVU (n=162,716)
% female	51.55	51.93
% adults 18-64 years	68.06	57.66
% adults ≥65 years	10.23	24.83
% European American	60.48	81.07
% African American	28.43	8.65
% Hispanic	10.04	1.32
% Asian	3.10	0.83

Over-represents European-descent and elderly

Crawford et al (2015) *Hum Hered* 79(3-4):137-46

INSTITUTE FOR COMPUTATIONAL BIOLOGY

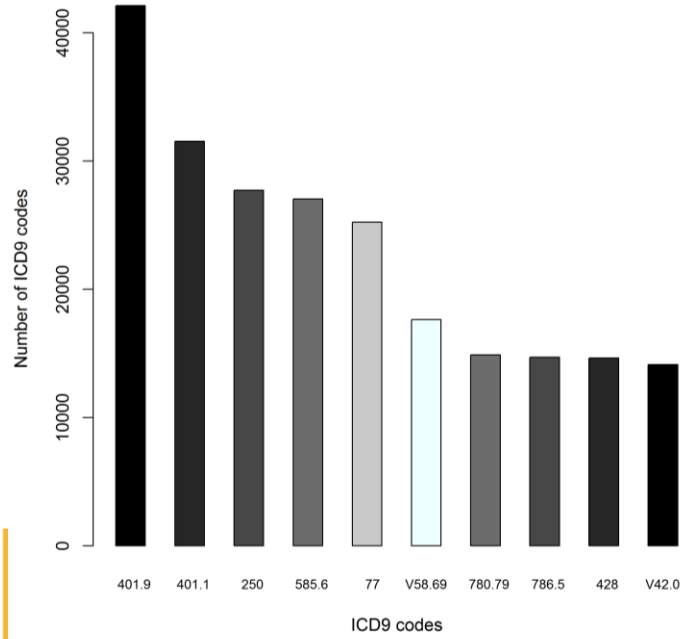
EAGLE BioVU Common Codes

Top 10 codes for African American adults

Hypertension (401.9, 401.1)

Diabetes Mellitus (250)

End-stage renal disease 9585.6)



Crawford et al (2015) *Hum Hered* 79(3-4):137-46

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

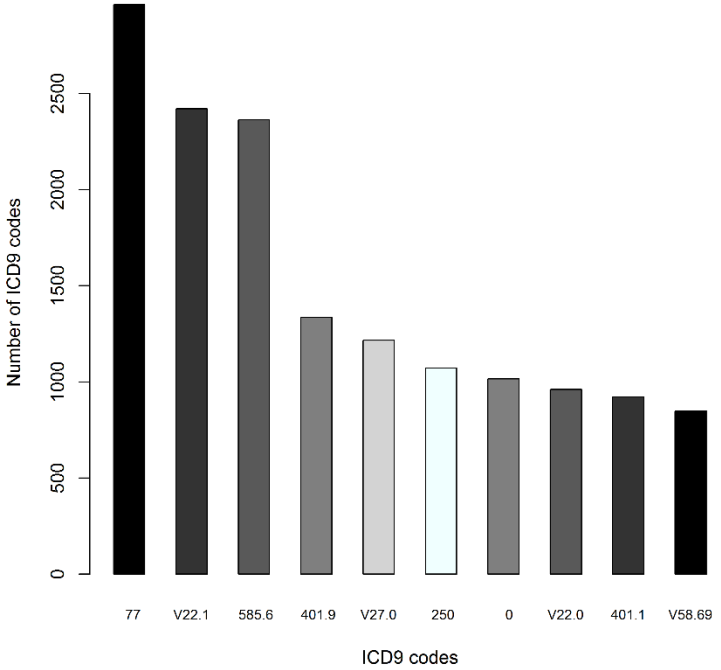
EAGLE BioVU Common Codes

Top 10 codes for Mexican American adults

Sequestrectomy (77)

Supervision of other normal pregnancy (v22.1)

End-stage renal disease (585.6)



Crawford et al (2015) *Hum Hered* 79(3-4):137-46

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Country-of-Origin Extraction

Why so many in TN from India, Italy, Chile, Greece, and China?

"Ital" misspelling for Italy

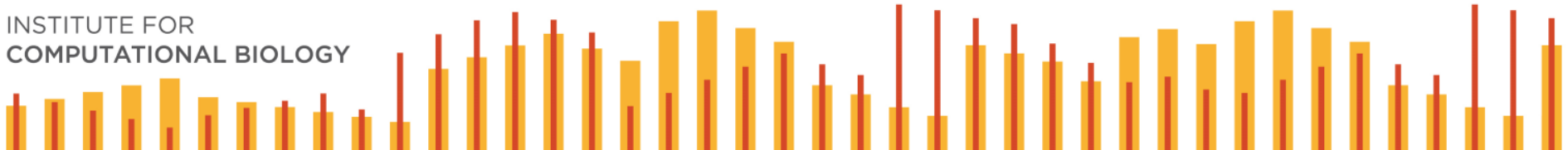
hosp*ital*, dig*ital*ly, v*ital*, mar*ital*, gen*ital*, etc.

Country-of-Origin Extraction

Why so many in TN from India, Italy, Chile, Greece, and China?

"Chil" misspelling for Chile

*chil*dren, *chil*ls, A*chil*les, etc.

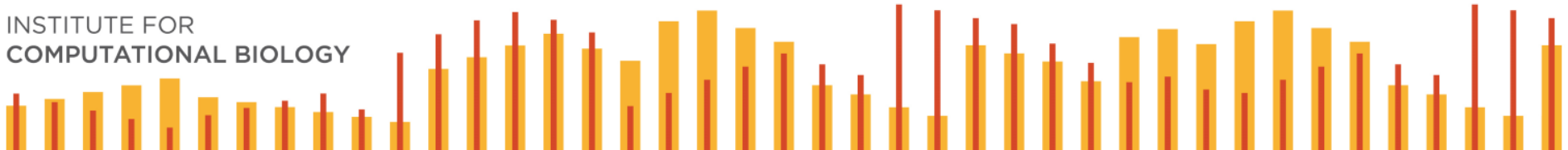


Country-of-Origin Extraction

Why so many in TN from India, Italy, Chile, Greece, and China?

"Gree" misspelling for Greece

Wal*gree*ns, a*gree*ment, de*gree*s, etc.

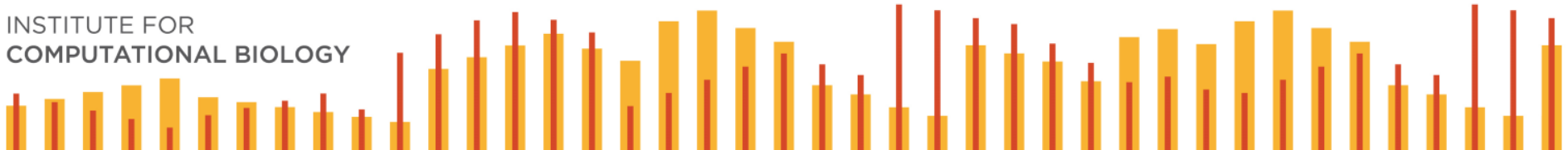


Country-of-Origin Extraction

Why so many in TN from India, Italy, Chile, Greece, and China?

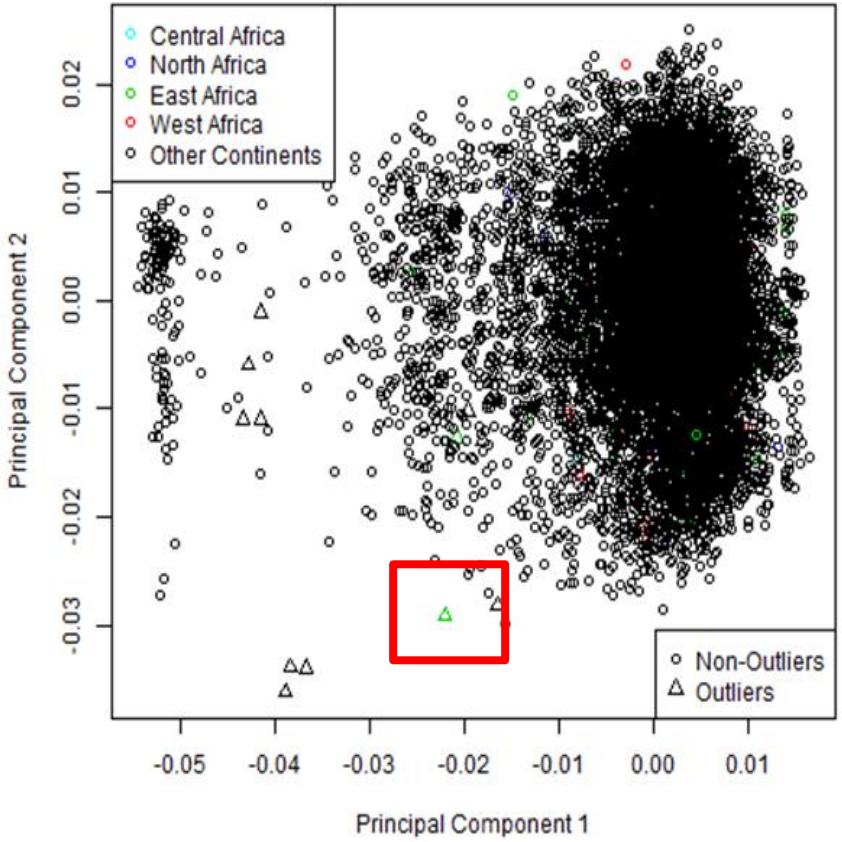
"Hia" misspelling for China

hydrochlorot*hia*zide, psyc*hia*tric, brac*hia*l, etc.



Comparison with Genetic Ancestry

- Only a few outliers assumed African American were actually from East Africa



INSTITUTE FOR
COMPUTATIONAL BIOLOGY