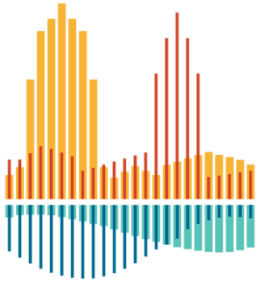
Large-scale Processing of Genomic Data coupled to Electronic Health Records

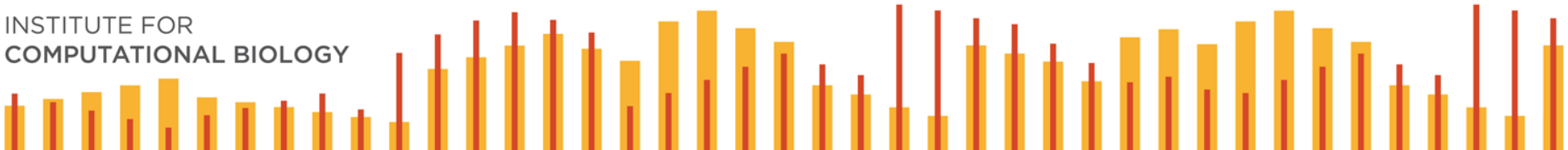


INSTITUTE FOR
COMPUTATIONAL
BIOLOGY

William S. Bush, PhD, MS
Assistant Professor
Department of Epidemiology and Biostatistics
Institute for Computational Biology

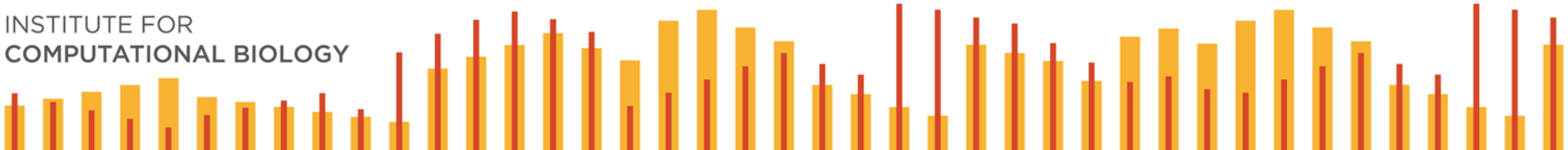
Disclosures

- I have no financial conflict of interests to report
- I liked data before it was big…

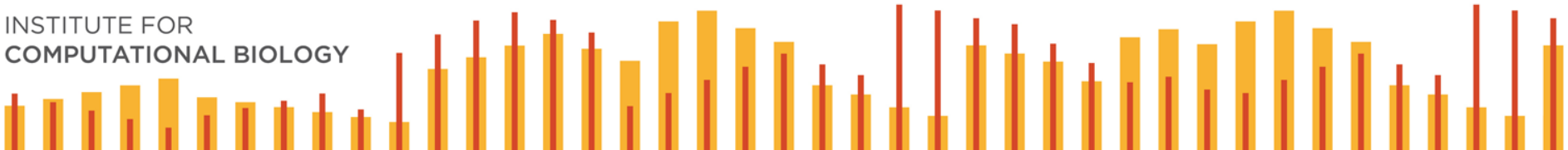


Overview

- Use of Electronic Health Records for Research
- Applications of High Performance Computing to EHR Research
- Large-scale Genomic Analysis



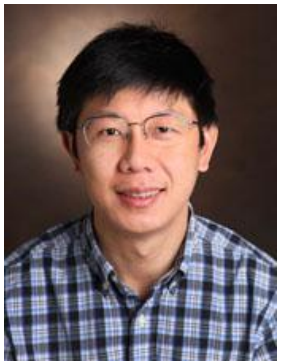
Use of Electronic Health Records for Research



The Institute for Computational Biology


Jonathan Haines


Ricky Chan


Chun Li


Dana Crawford

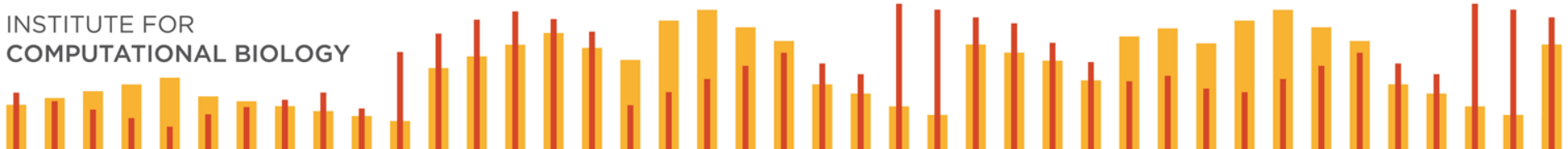

Will Bush


Jill Barnholtz-Sloan

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Electronic Health Records

- Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009
- Incentivizes the nationwide adoption of EHR systems in the US
- While there are standards and commonalities, EHR systems are very heterogeneous
- Epic Systems, Allscripts, Meditech, Cerner, IBM, McKesson, Siemens, GE Healthcare
- Epic is most common at large medical centers



The Cleveland Metro Area

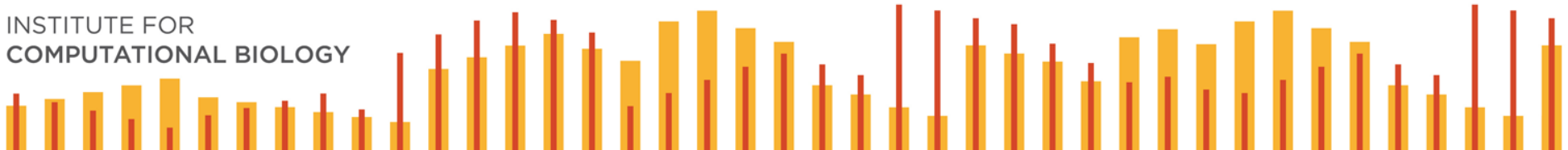
University Hospitals

Cleveland Clinic

MetroHealth

Approaching ~90% of all healthcare in the Cleveland Metropolitan Area

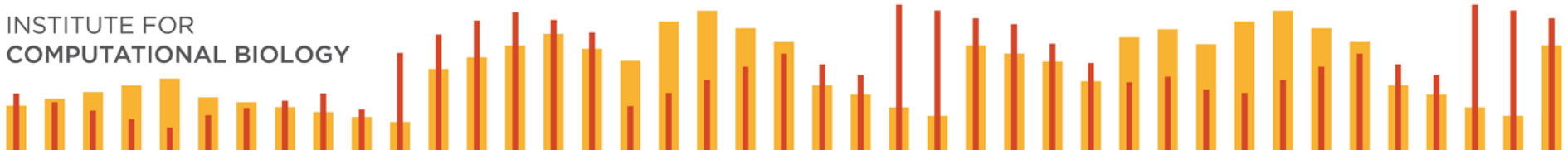
INSTITUTE FOR
COMPUTATIONAL BIOLOGY



Electronic Health Record Components

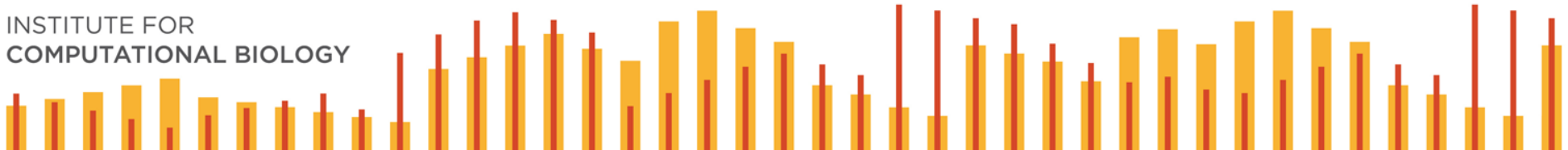
- Structured Elements
- Pseudo-Structured Text
- Unstructured Text

EHR elements can differ between systems and clinics



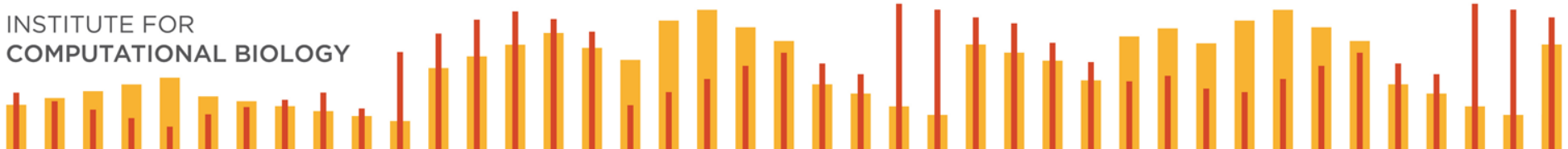
Structured Elements

- Demographic information
  - Date of birth, race/ethnicity, gender
- Vital Signs
  - Heart rate, blood pressure, height, weight, body temperature
- Some Laboratory Values
  - WBC, insulin, glucose, glomerular filtration rate (GFR)
- Billing and Procedure codes
  - ICD9/10, CPT, ICD-O-3



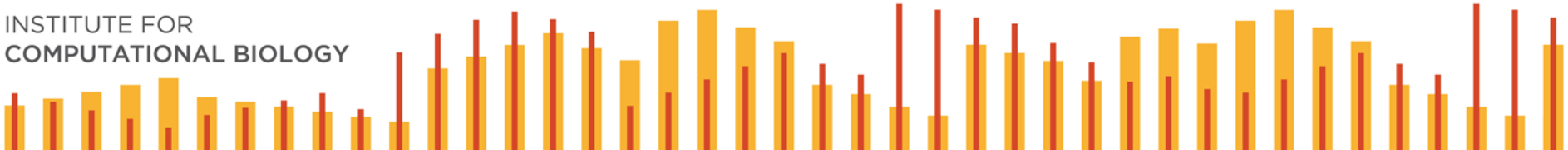
Pseudo-structured Text

- Free text documents with a loosely standardized format
- History and Physical Examination (HPE)
- *“… approximately 10 months status post bilateral bunionectomy with metatarsal head resections 2 through 5 bilaterally. She denies any pain in her feet although she is a little upset that she has had a recurrence of her bilateral hallux valgus, left worse than the right. ”*
- *“She does have some residual hallux valgus, worse on the left then on the right, but this is not uncommon following bunionectomy. We discussed the fact that her bunions were so bad to begin with, that her feet actually look pretty good.“*
- Problem List / Known significant medical conditions and procedures / Allergies



Unstructured Text

- Clinical communications (phone calls, prescription refills, etc)
- Discharge notes
- Clinic-specific notes
- Laboratory / Procedure reports (CT scan, colonoscopy, etc)
- “Medical flotsam”

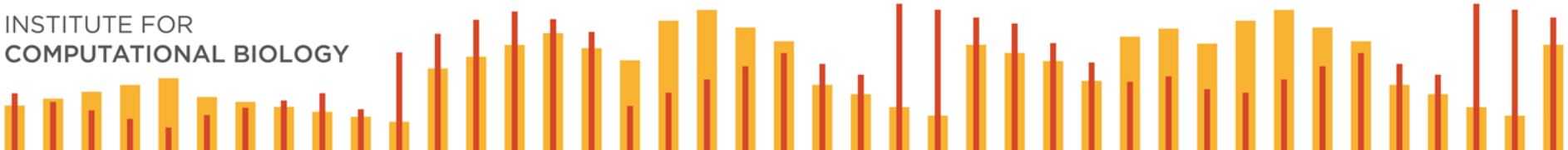


My EHR Existential Crisis

At some point in everyone's life, they realize that…

- Their parents don't know everything.
- Their doctor doesn't know everything.
- Their medical record may be wrong.

EMRs are very useful, but they can be noisy and inaccurate.

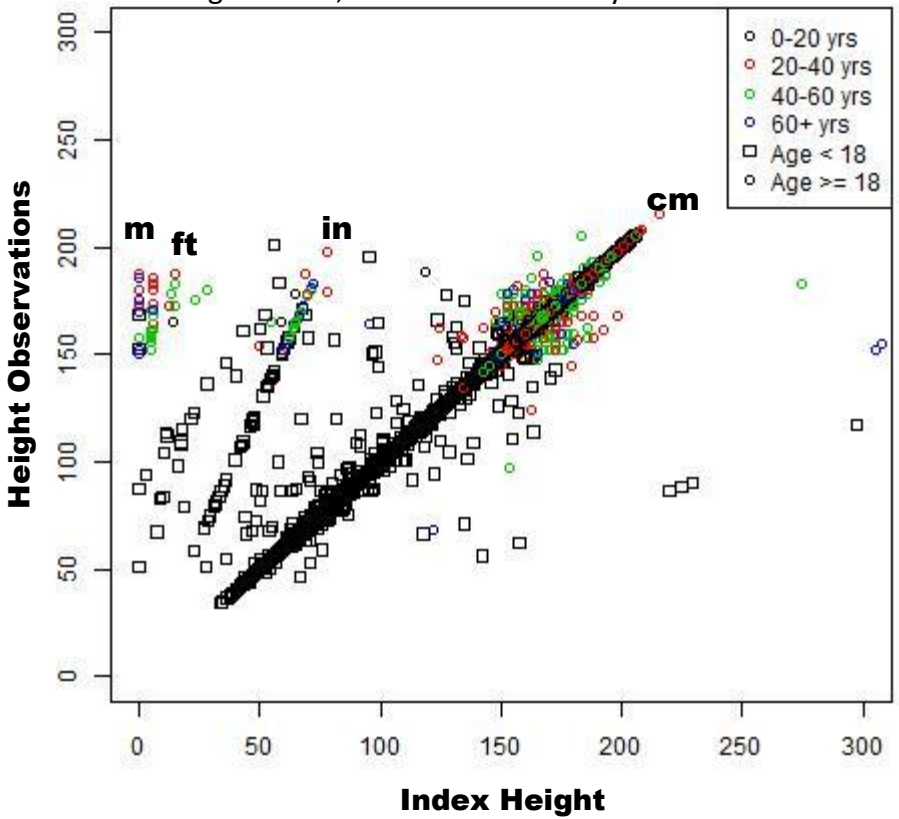


Example: Body Mass Index

Height and Weight are the most ubiquitous measures reported in an EHR

Robert Goodloe

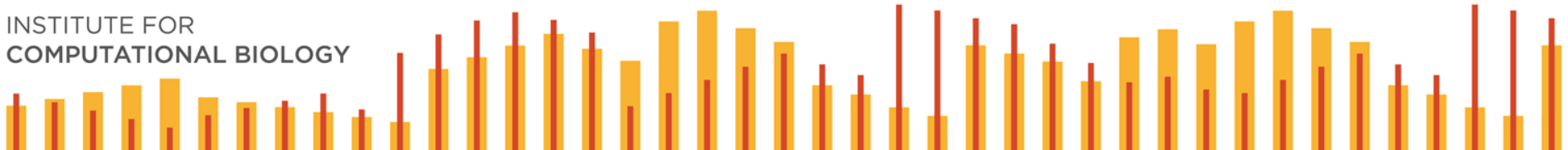
Eagle-BioVU, Vanderbilt University Medical Center



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Clinical Versus Research Use

- In clinical practice, records and values are examined individually
  - Errors are easy to spot and ignore
  - Aggregate data is synthesized by experts in the context of a patient
- In research, records and values are examined in aggregate
  - Errors are not so easy to identify
  - Data points are examined outside the context of the patient



Structuring Data for Research Use

record_id	icd_code	Date
24067	729.5	40909
24067	070.30	40910
24067	783.1	40910
24067	45.23	40911
24067	77	40911
24067	455.6	40912

record_id	vital_sign	value	Date
24067	hr	72	40912
24067	sbp	160	40912
24067	dbp	92	40912
24067	height	150	40912

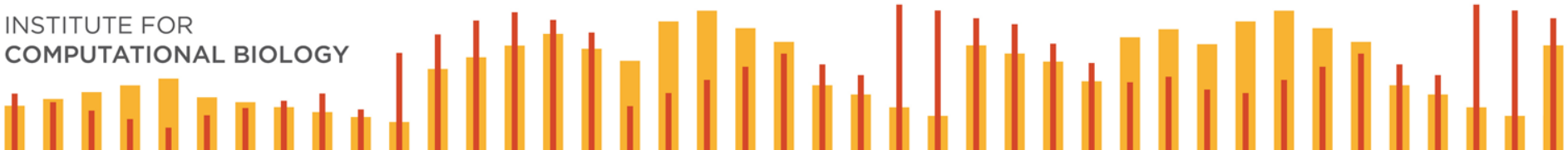
"patient started Simvastatin 800 mg"

"42 year old male, overweight"

OfficeEMR by iSALUS Healthcare

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Applications of High Performance Computing to EHR Research



Example: Calcineurin-inhibitor Toxicity

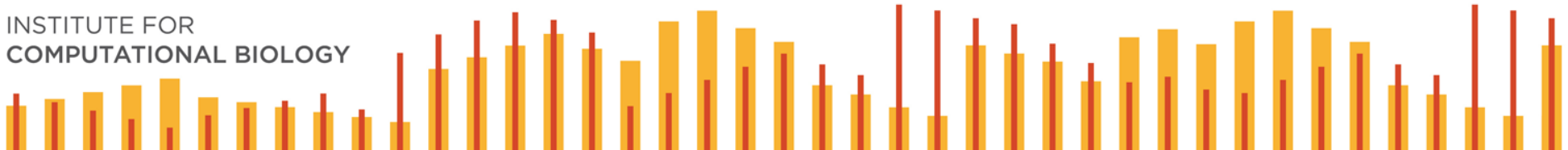
- Given post-transplant to prevent organ rejection
  - Tacrolimus and cyclosporine
- Narrow therapeutic window
- Nephrotoxicity is a serious and common complication
- Serum creatinine and glomerular filtration rates (GFR) are monitored post-transplant to assess kidney function

Matt Oetjens

Dana Crawford

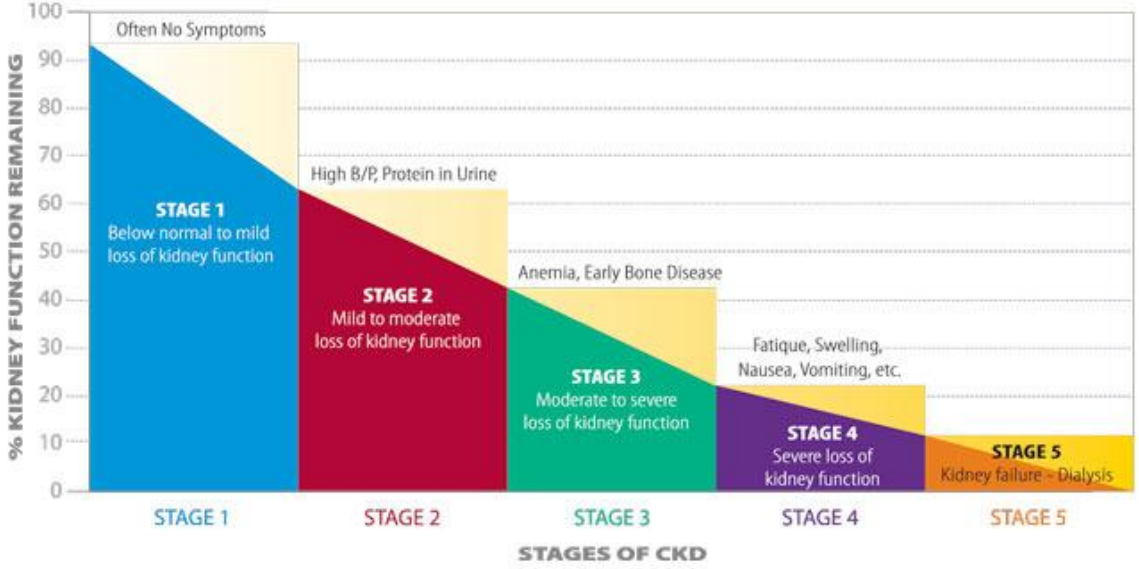
Study Design

- Patients identified having:
  - Heart transplant documented with >= 3 ICD9 Code V42.1 (heart replaced by transplant) and/or one CPT Code 33945 (cardiectomy with heart transplant)
  - One or more mention of an immunosuppressant
  - Age > 15 at date of transplant
  - Available DNA



Chronic Kidney Disease

Chronic kidney disease is classified in 5 stages of severity (determined by estimated GFR)



Outcome for this Study

- eGFR is calculated

$$186 \times \text{Serum Creatinine}^{-1.154} \times \text{Age}^{-0.203} \times [1.212 \text{ if Black}] \times [0.742 \text{ if Female}]$$

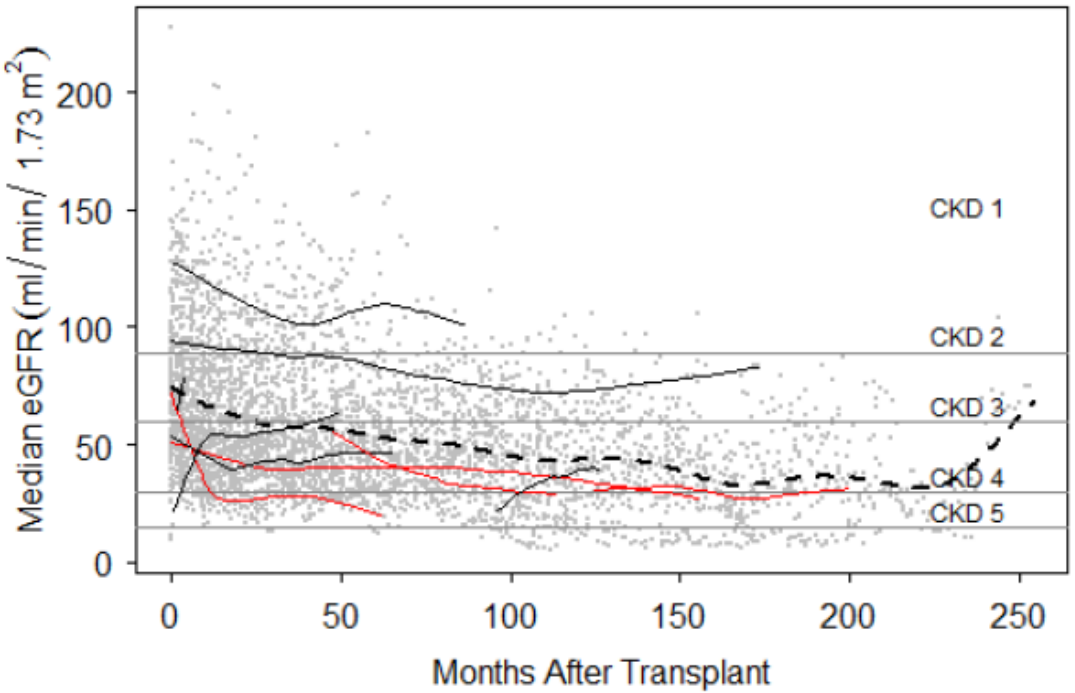
- Severe Kidney Disease: Post-transplant eGFR < 30 mL/min/1.73m^2 for 3 consecutive months
- Time to development of severe nephrotoxicity clinically attributed to calcinurin inhibitor toxicity

Clinical Trajectories

A functional variant, rs1801265 in the gene *DPYD* associated to Cyclosporine induced nephrotoxicity (p = 0.03)

We can make new discoveries in EHR data, but it takes some work

Vanderbilt University Medical Center



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Exploring "Phenomes"

Human Genome

Genome-Wide Association Study (GWAS)

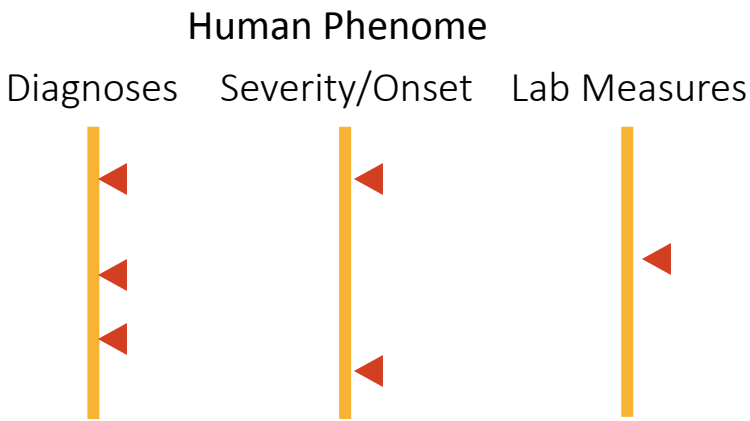
PheWAS

- Can we explore the
  n e way
  ol e?

Marylyn Ritchie
Geisinger/PSU

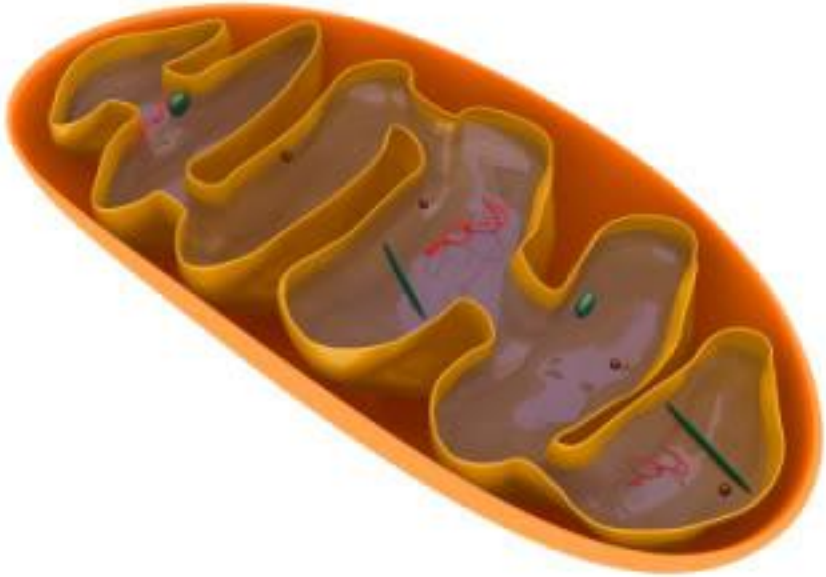
Sarah Pendergrass
Geisinger

Dana Crawford
CWRU



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

The Mitochondria



- *Power plants* of all living cells
- Have their own genome
- 26,000 base pairs versus 3 billion

Mitochondrial PheWAS

- Multiple quantitative traits for cardiovascular disease
- A series of mitochondrial genetic variants

Jake Hall

Sabrina Mitchell

Mitchell *et al. BioData Mining* 2014, **7**:6
http://www.biodatamining.org/content/7/1/6

BioData Mining

RESEARCH Open Access

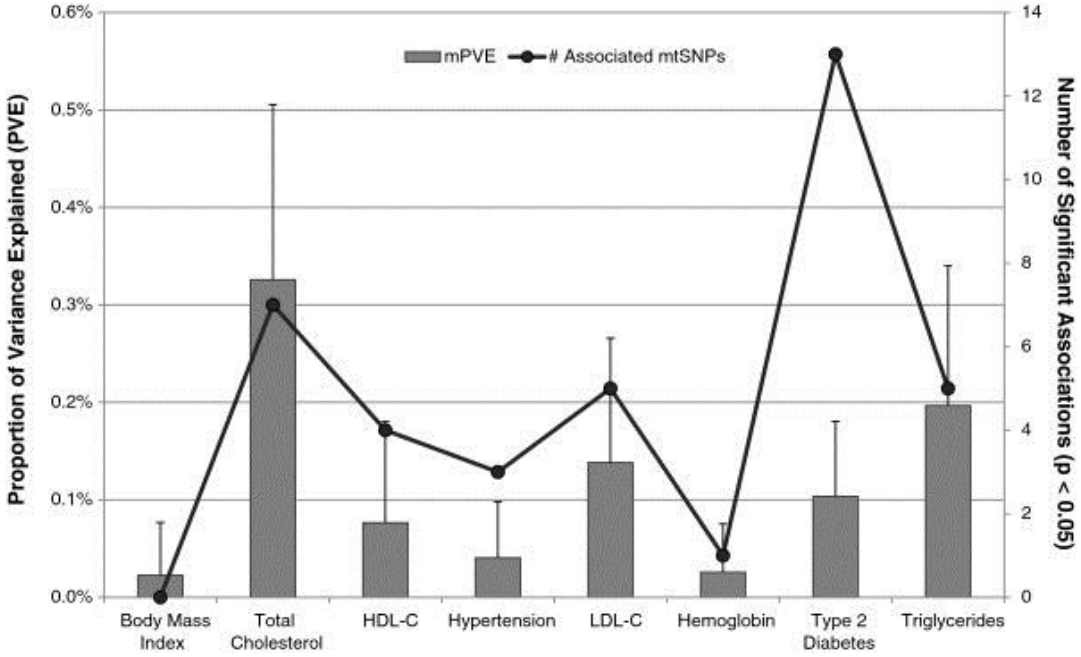
Investigating the relationship between mitochondrial genetic variation and cardiovascular-related traits to develop a framework for mitochondrial phenome-wide association studies

Sabrina L Mitchell[1,2†], Jacob B Hall[1†], Robert J Goodloe[1], Jonathan Boston[1], Eric Farber-Eger[1], Sarah A Pendergrass[3], William S Bush[1,4] and Dana C Crawford[1,2*]

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

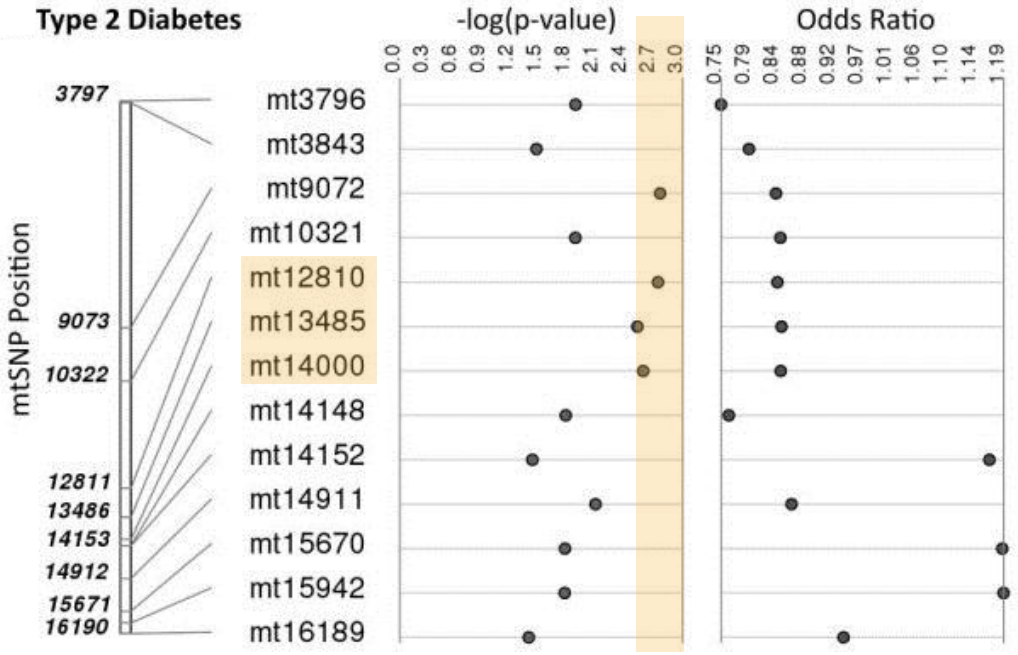
Mitochondrial PheWAS

- Total Cholesterol and Type II Diabetes show a significant amount of risk explained



Mitochondrial PheWAS

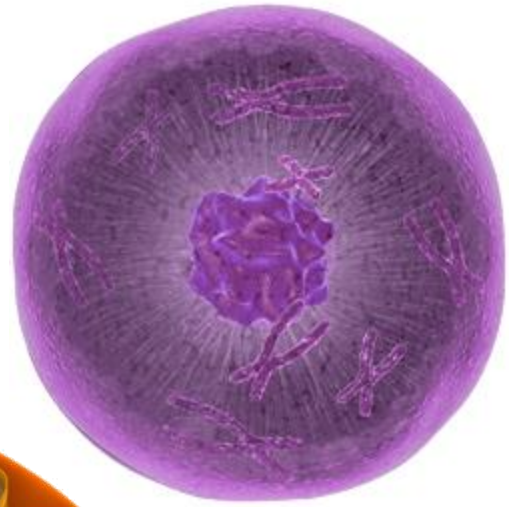
- A series of mitochondrial variants are protective for T2D risk



Mitochondrial Nuclear Interactions

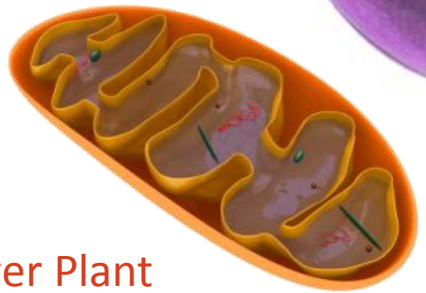
- There are complex chemical signals between the nucleus of the cell and mitochondria

Google Data Center



Nucleus

Mitochondria



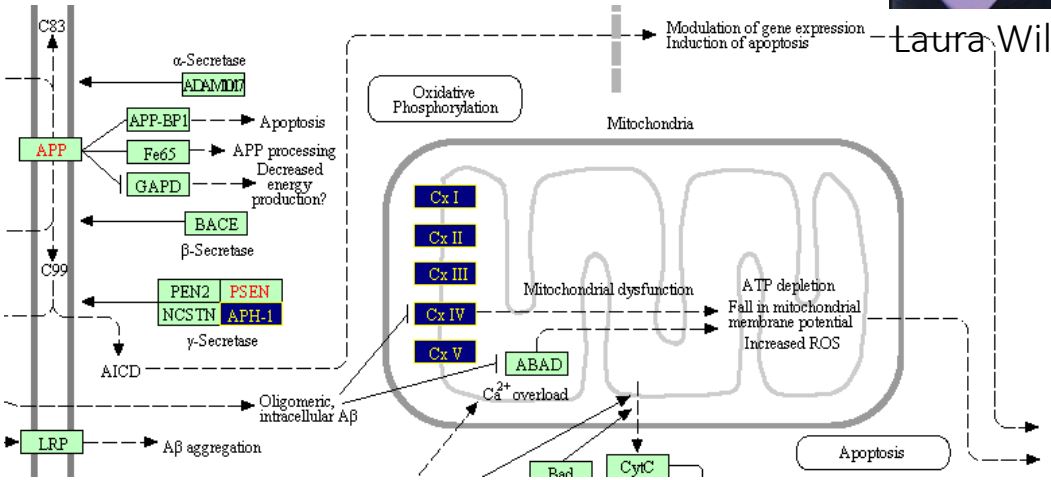
Power Plant

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Mitochondrial Nuclear Interactions

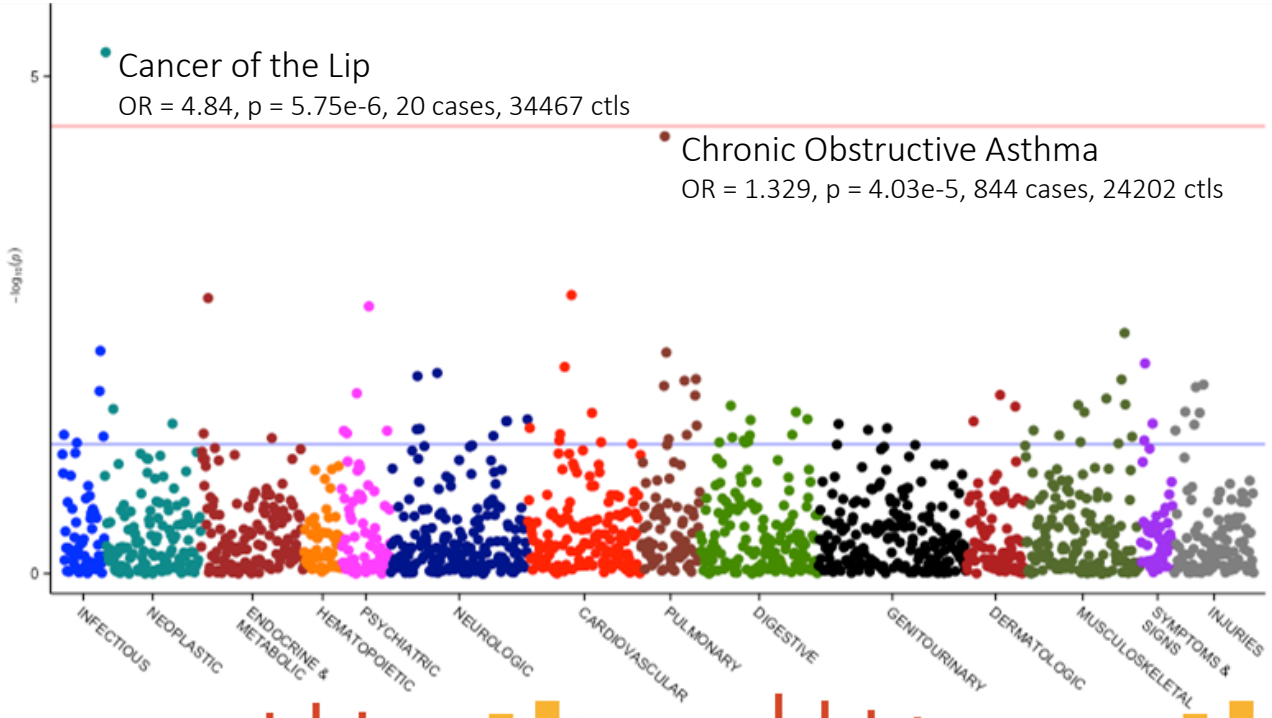
Laura Wiley

- We discovered a nuclear genetic change that influences expression of mitochondrial genes



INSTITUTE FOR COMPUTATIONAL BIOLOGY

FUNCTIONAL SNP PHEWAS



INSTITUTE FOR
COMPUTATIONAL BIOLOGY

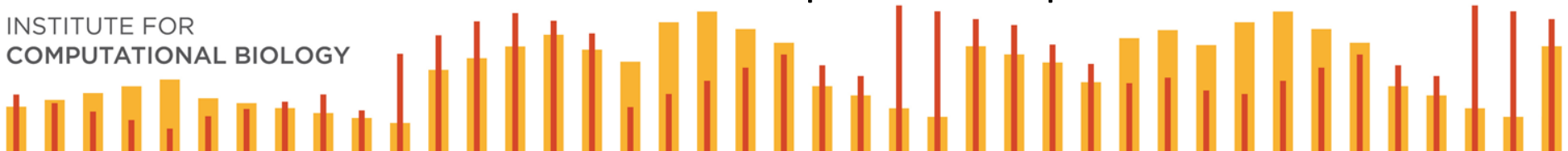
Large Scale Genomic Analysis

Jake Hall

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Pathway-based Analyses

- Individual genetic changes contribute very little to common diseases
- Examining changes in aggregate can inform biology as well (by mechanism)
- Combinations of genetic changes may contribute more risk than their independent parts



Mixed-Model Analyses

- Estimation of a genetic relationship matrix
- Produces an estimate of the risk explained by genetic sharing
- Assumes each genetic change contributes independently

$$A_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(covar.j)(covar.k)}{(var)}$$

46% of Type II Diabetes is explained

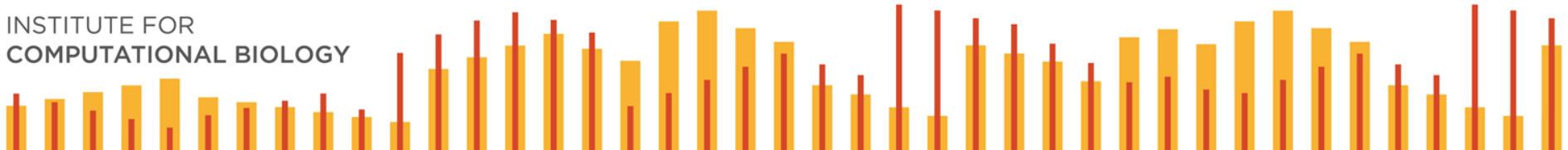
Nathan Morris

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Expanded Genetic Relationship Matrix

- Can we model the effects of *interactions* among variants?

58% of Type II Diabetes is explained
46% due to independent effects
12% due to interaction effects



Expanded Genetic Relationship Matrix

Variance

Component	Equation
A × D Interaction	$\frac{2xy}{(xy)^2}$
D × A Interaction	$\frac{2xy}{(uv)^2}$
D × D Interaction	$\frac{1}{(uvxy)^2}$

Covariance

Component	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
Interaction	f_{*2}	f_{*1}	f_{*0}	f_{*2}	f_{*1}	f_{*0}	f_{*2}	f_{*1}	f_{*0}
D × A Interaction	$\frac{2y}{f_{*2}}$	$\frac{(y-x)}{f_{*2}}$	$\frac{2x}{f_{*2}}$	$\frac{-4y}{f_{*1}}$	$\frac{-2(y-x)1}{f_{*1}}$	$\frac{4x}{f_{*1}}$	$\frac{2y}{f_{*0}}$	$\frac{(y-x)}{f_{*0}}$	$\frac{-2x}{f_{*0}}$
D × D Interaction	$\frac{1}{f_{22}}$	$\frac{-2}{f_{21}}$	$\frac{1}{f_{20}}$	$\frac{-2}{f_{12}}$	$\frac{4}{f_{11}}$	$\frac{-2}{f_{10}}$	$\frac{1}{f_{02}}$	$\frac{-2}{f_{01}}$	$\frac{1}{f_{00}}$

Lots of Computation

INSTITUTE FOR
COMPUTATIONAL BIOLOGY

Acknowledgments

The Bush Lab

- Jake Hall
- Mike Sivley
- Alexandra Fish
- Jeremy Fondran

The Institute for Computational Biology

- Jonathan Haines
- Dana Crawford
- Chun LI
- Jill Barnholtz-Sloan
- Ricky Chan

Grant Support

NIMH 095621
Asha Kallianpur/Todd Hulgan

NHGRI 004798
Dana Crawford

NIA 07133
Farrer, Schellenberg, Haines