Electronic health records and genomics – a dynamic duo for precision medicine

Marylyn D. Ritchie, PhD

Paul Berg Professor, Biochemistry & Molecular Biology,

The Pennsylvania State University

Director, Biomedical and Translational Informatics,

Geisinger Health System

GEISINGER

REDEFINING BOUNDARIES

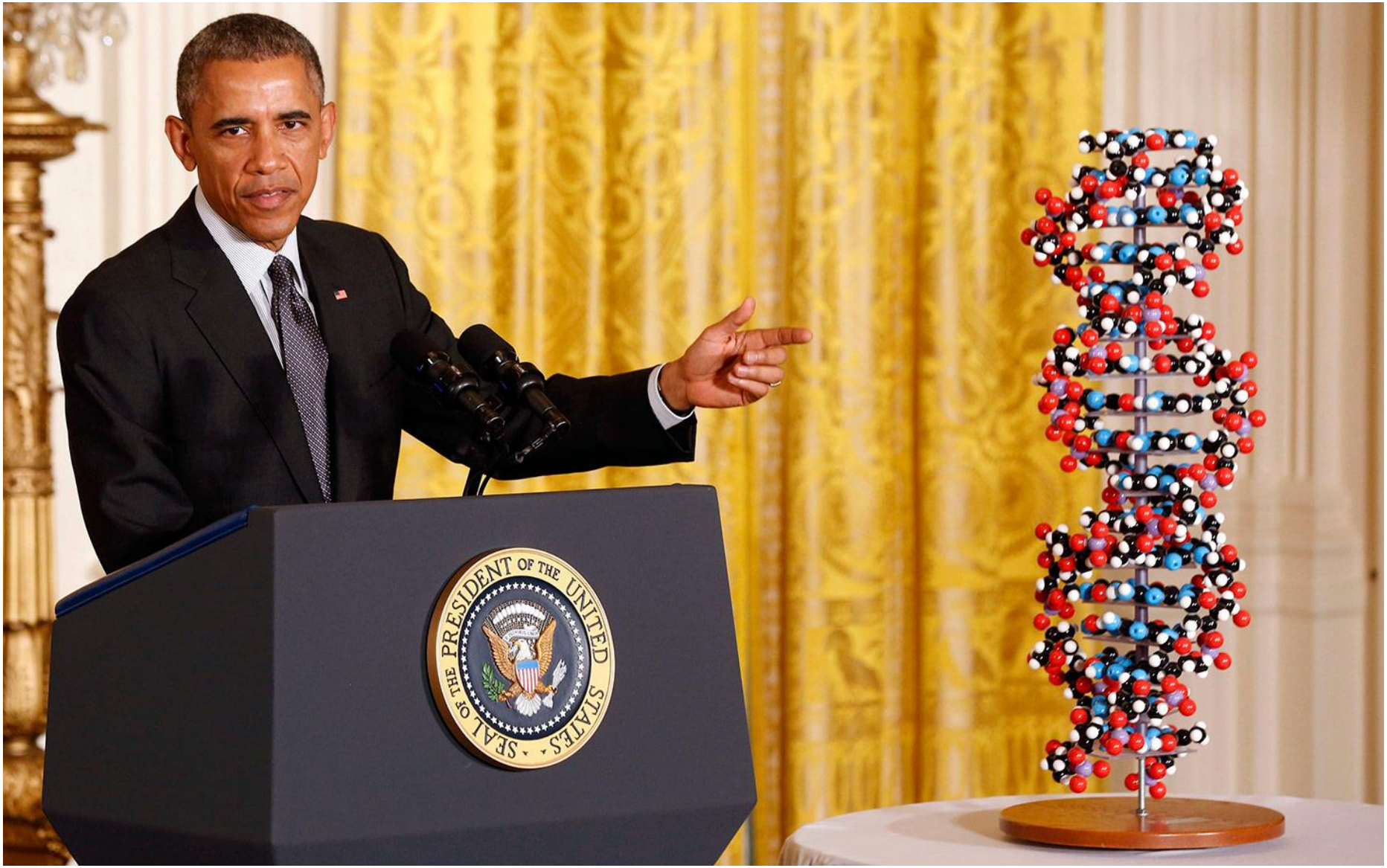
PENNSTATE

1855


peanut
butter


Spock & Kirk (Leonard Nimoy a William Shatner)
http://puvodni.startrek.cz

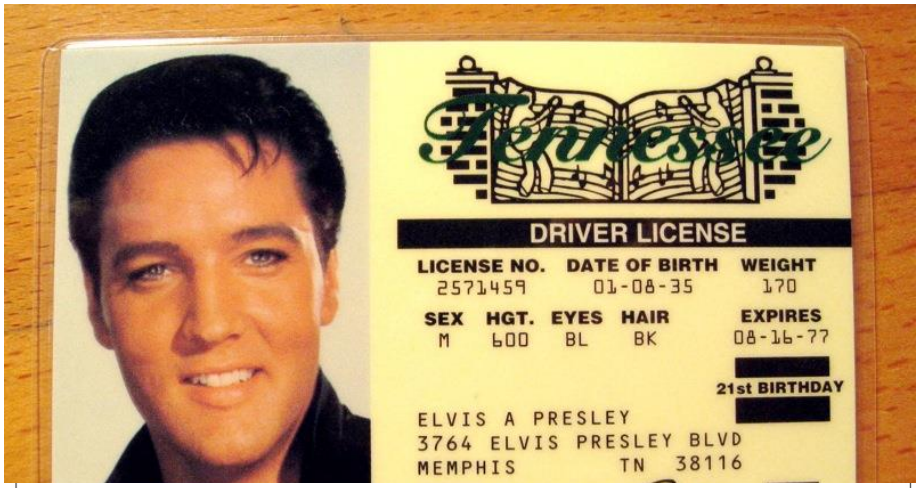
Precision Medicine Initiative

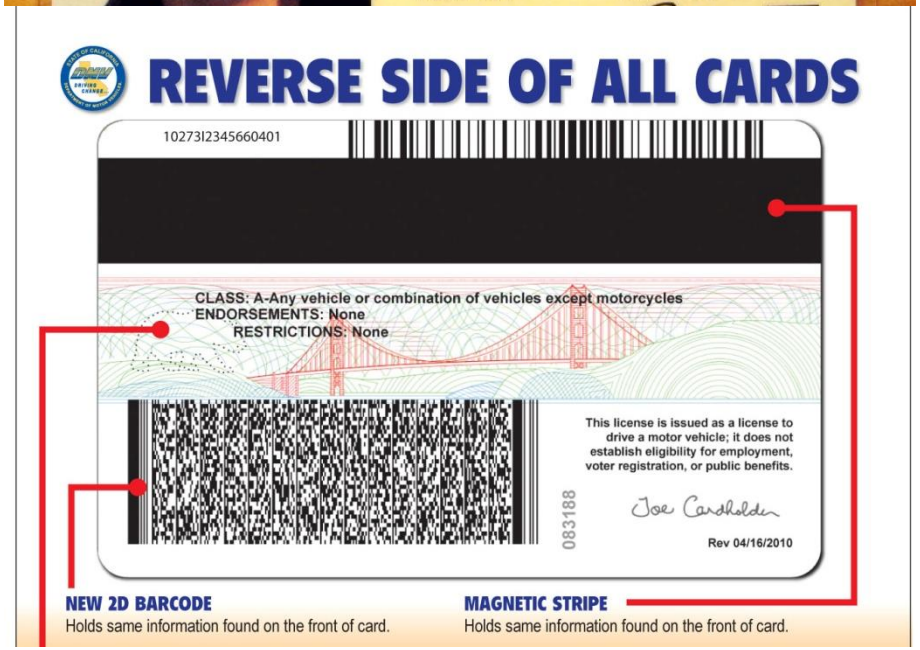


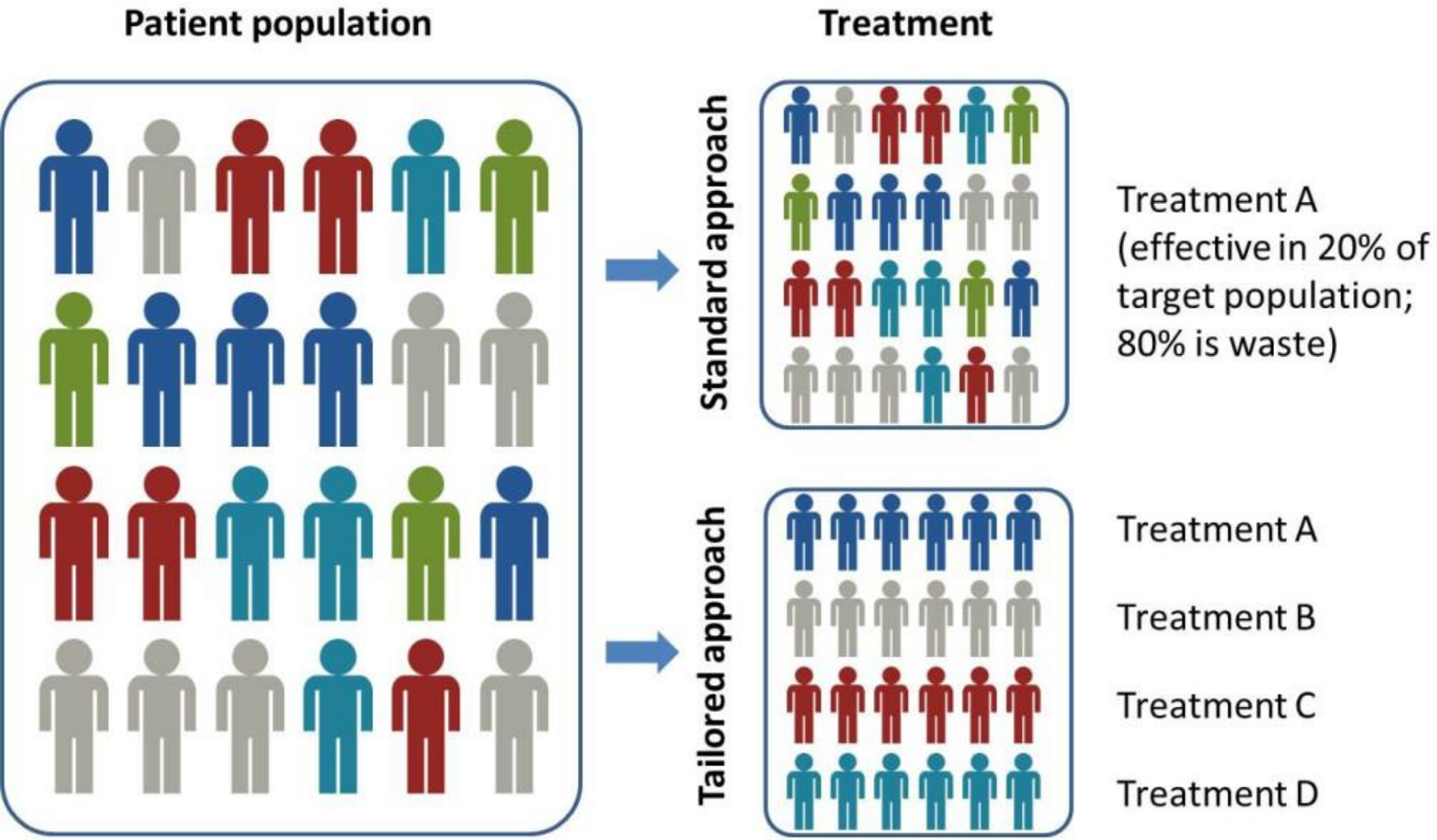
January 30, 2015

Precision/Personalized Medicine

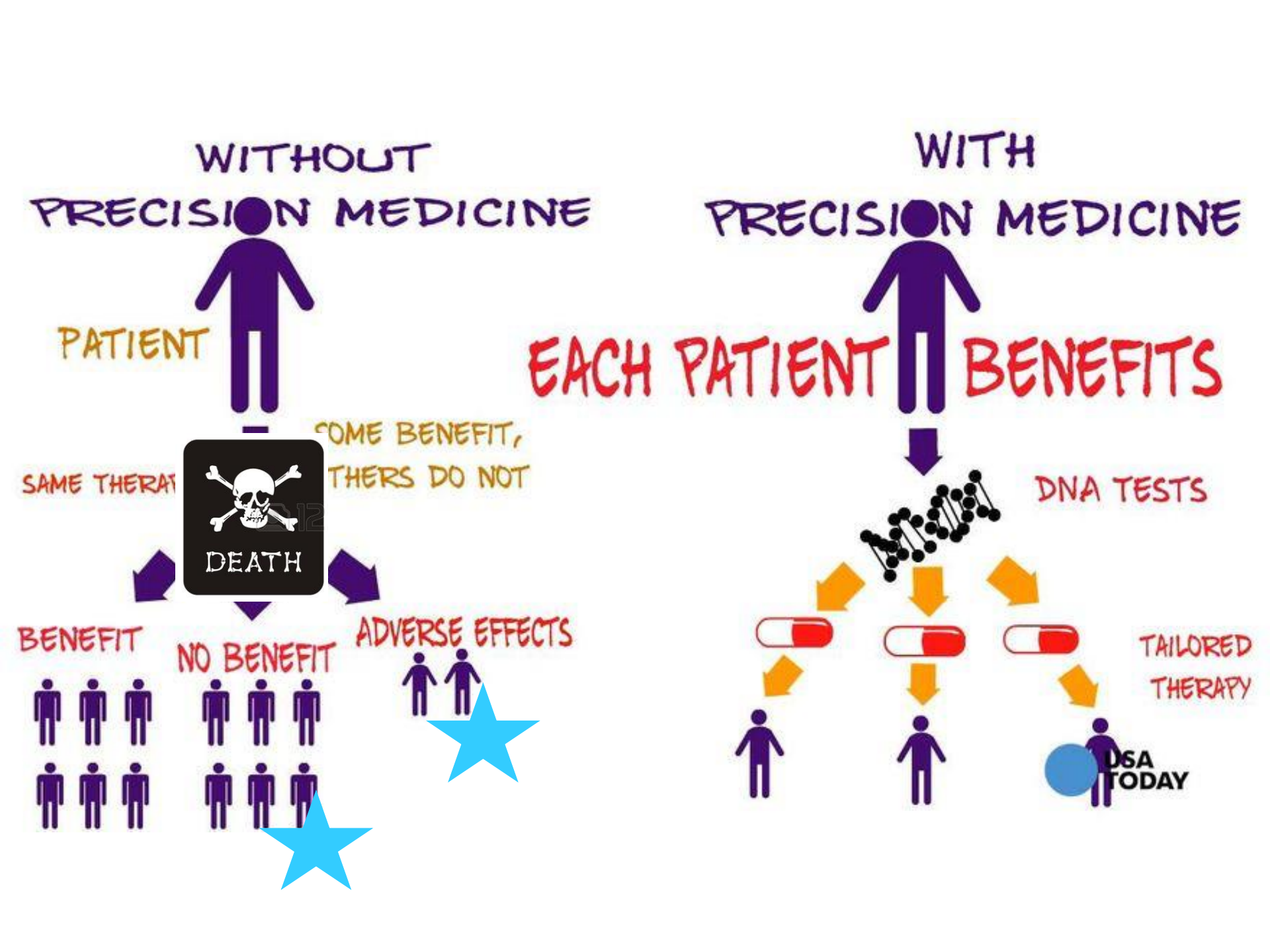
Pharmacogenomics: When medicine gets personal







http://www.cpgr.org.za/precision-medicine-in-south-africa-a-cost-benefit-analysis-framework/


WITHOUT
PRECISION MEDICINE
PATIENT
SAME THERAT
COME BENEFIT,
THERS DO NOT
DEATH
BENEFIT
NO BENEFIT
ADVERSE EFFECTS
WITH
PRECISION MEDICINE
EACH PATIENT
BENEFITS
DNA TESTS
TAILORED
THERAPY
USA
TODAY

News & Events

CYP2C9

FDA NEWS RELEASE

FOR IMMEDIATE RELEASE
August 16, 2007

Media Inquiries:
Karen Riley, 301-827-6242
Consumer Inquiries:
888-INFO-FDA

FDA Approves Updated Warfarin (Coumadin) Prescribing Information

New Genetic Information May Help Providers Improve Initial Dosing Estimates of the Anticoagulant for Individual Patients

The U.S. Food and Drug Administration announced today the approval of updated labeling for the widely used blood-thinning drug, Coumadin, to explain that people's genetic makeup may influence how they respond to the drug.

Manufacturers of warfarin, the generic version of Coumadin, are to add similar information to their products' labeling, FDA said.

The labeling change highlights the opportunity for healthcare providers to use genetic tests to improve their initial estimate of what is a reasonable warfarin dose for individual patients. Testing may help optimize the use of warfarin and lower the risk of bleeding complications from the drug.

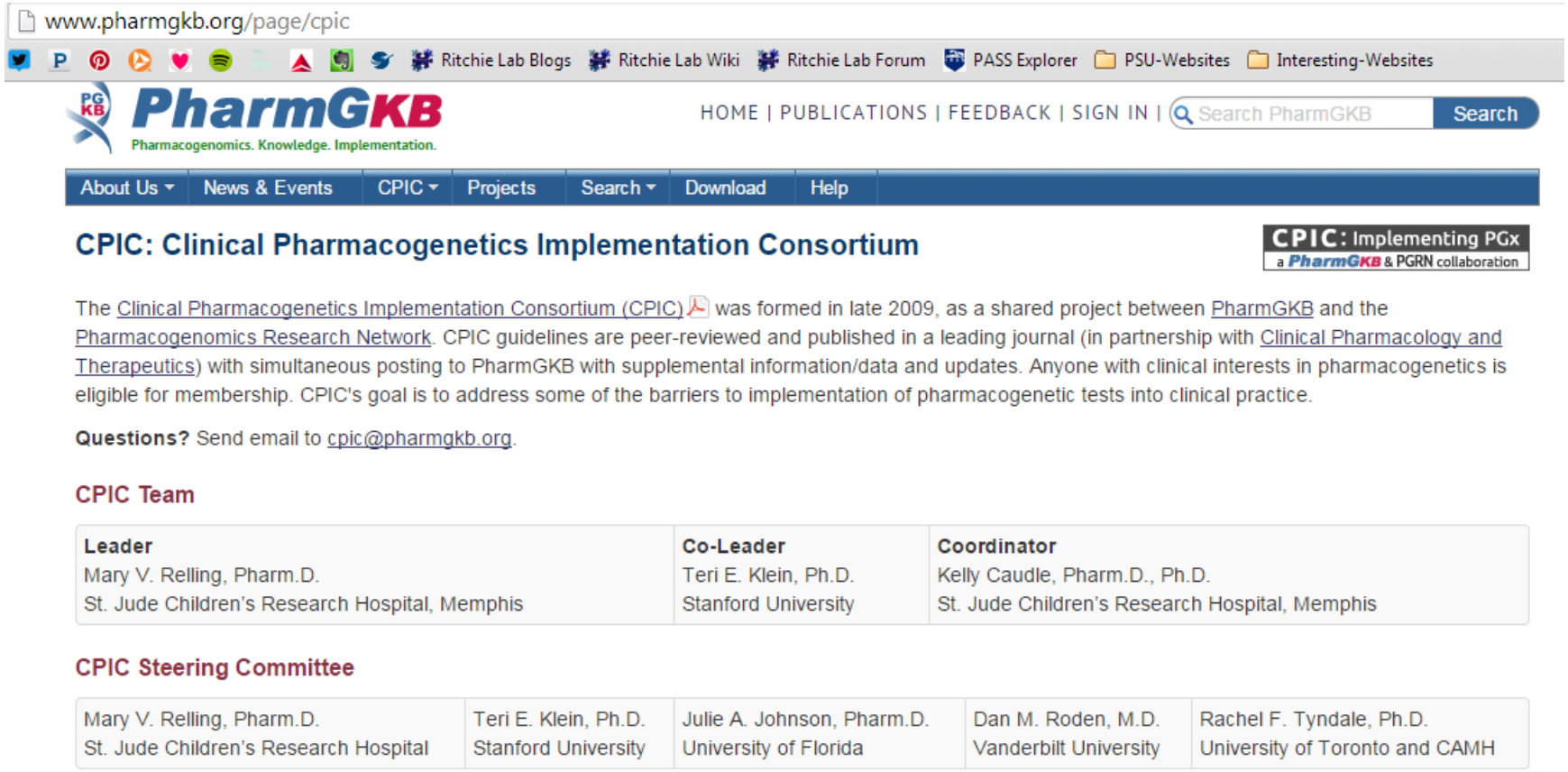
These labeling updates are based on an analysis of recent studies that found people respond to the drug differently based, in part, on whether they have variations of certain genes.

FDA estimates that 2 million persons start taking warfarin in the United States every year to prevent blood clots, heart attacks and stroke. Warfarin is a difficult drug to use because the optimal dose varies and depends on many risk factors including a patient's diet, age, and the use of other medications.

Patients who take a dose larger than they can tolerate are at risk of life-threatening bleeding. Those who receive too low a dose are at risk of equally dangerous blood clots. Dosing is particularly important at the beginning of therapy, when problems in adjusting the dose can lead to complications such as bleeding.

Warfarin is the second most common drug — after insulin — implicated in emergency room visits for adverse drug events.

Physicians and other health care professionals who prescribe warfarin regularly check to see if the drug is working properly by ordering a test called the PT or prothrombin



- 28 CPIC guidelines for gene-drug pairs
- ~100 genetic variants

NIH National Institutes of Health
Turning Discovery Into Health

Search

For Employees | Staff Directory | En Español

Health Information | Grants & Funding | News & Events | Research & Training | Institutes at NIH | About NIH

NIH Home > Research & Training > Precision Medicine Initiative

PRECISION MEDICINE INITIATIVE

Precision Medicine Initiative

- What are the near-term goals?
- What are the longer-term goals?
- How is it different?
- Who will participate?
- NIH Workshop

NIH Workshop on Building a Precision Medicine Research Cohort

On February 11-12, 2015, NIH hosted a workshop to discuss the opportunities and challenges around building a large research cohort focused on precision medicine and heard from several leading experts from many disciplines and sectors. More than 2,000 people watched on videocast and more than 500 people engaged through WebEx, submitting comments and questions to the workshop panelists. The workshop panelists also took comments and questions from Twitter based on a lively discussion from the hashtag #PMINetwork.

Watch the Videocast

- Day 1 Videocast Information
- Day 2 Videocast Information

Email Updates

To sign up for updates please enter your e-mail address.

Submit

Related Links

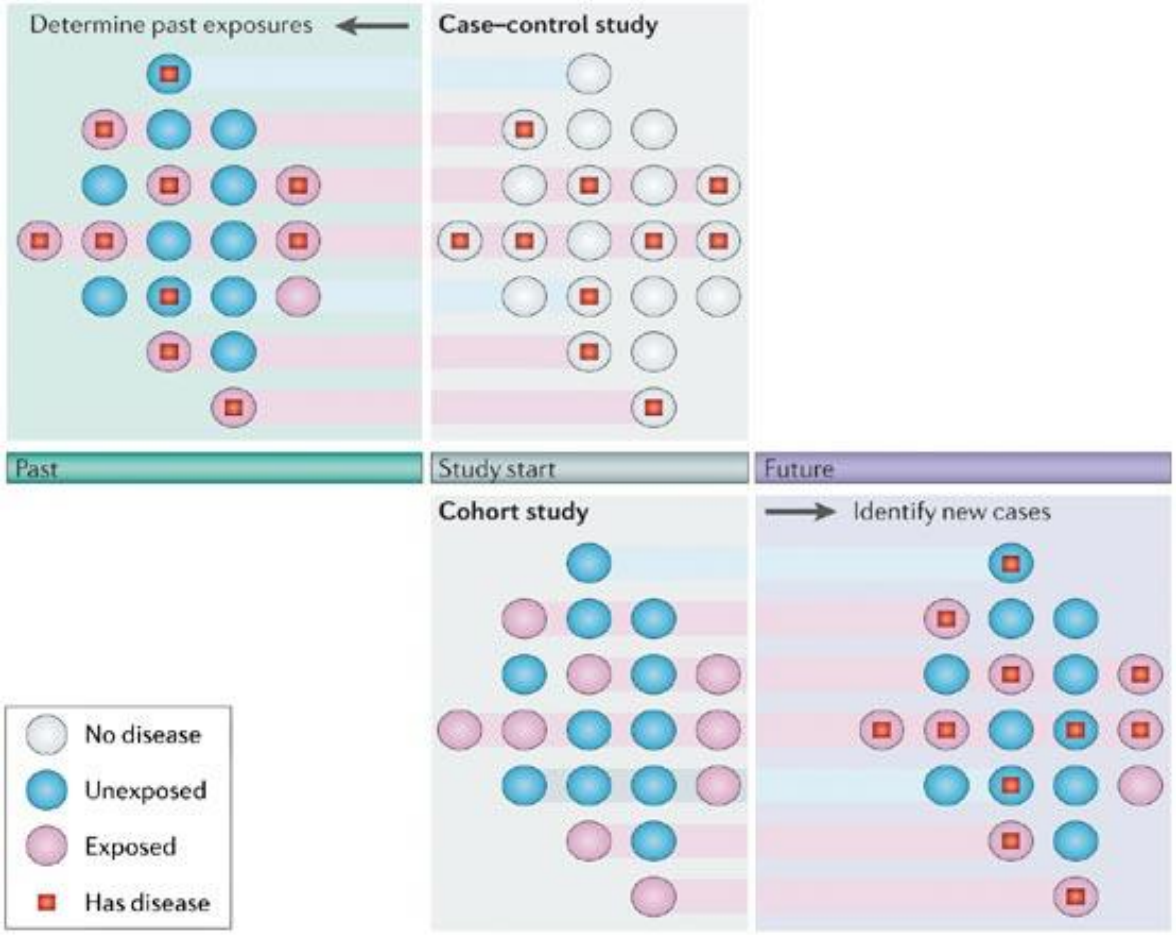
NIH Workshop

- NEJM Perspective: A New Initiative on Precision Medicine
- White House Precision Medicine Web Page
- White House Fact Sheet: President Obama's Precision Medicine Initiative

What study design should be used for the 1M project?

What analysis strategies should be employed?

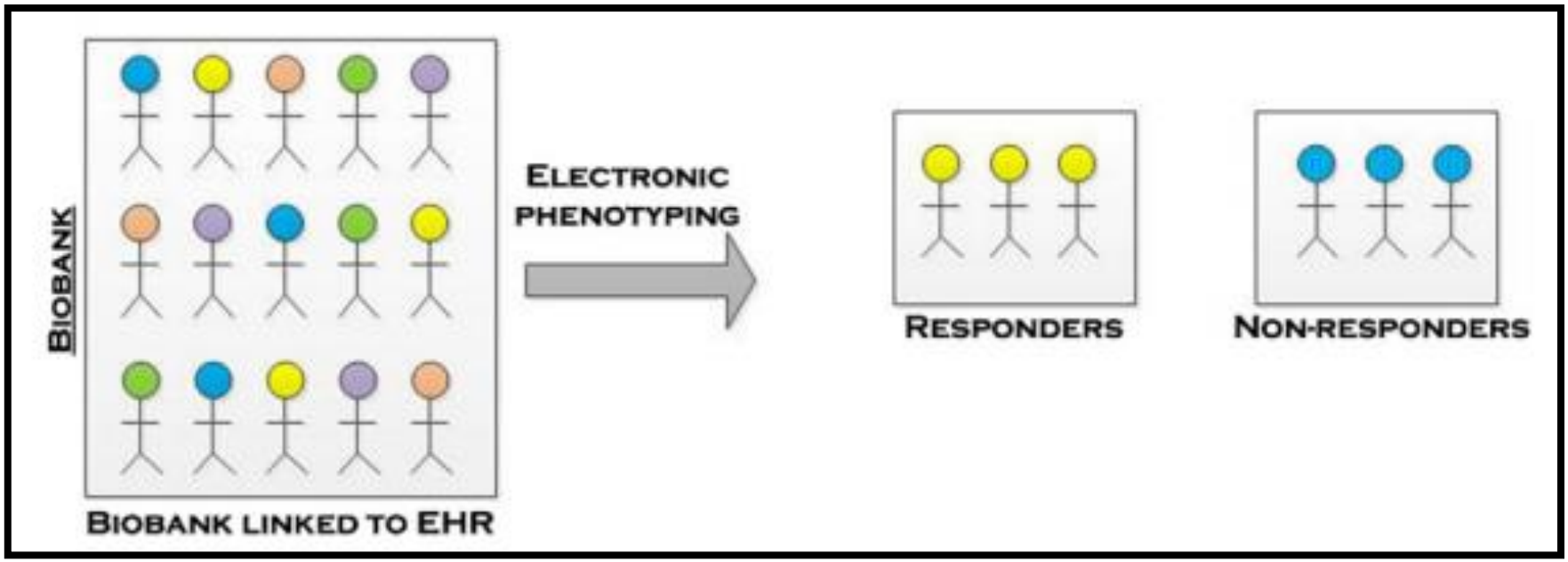
Study Designs for 1M Person Project



Manolio *et al. Nature Reviews Genetics* **7**, 812–820 (October 2006)

nature REVIEWS GENETICS

Study Designs for 1M Person Project



Human Genetics (2012) 131: 1615-1626

The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine

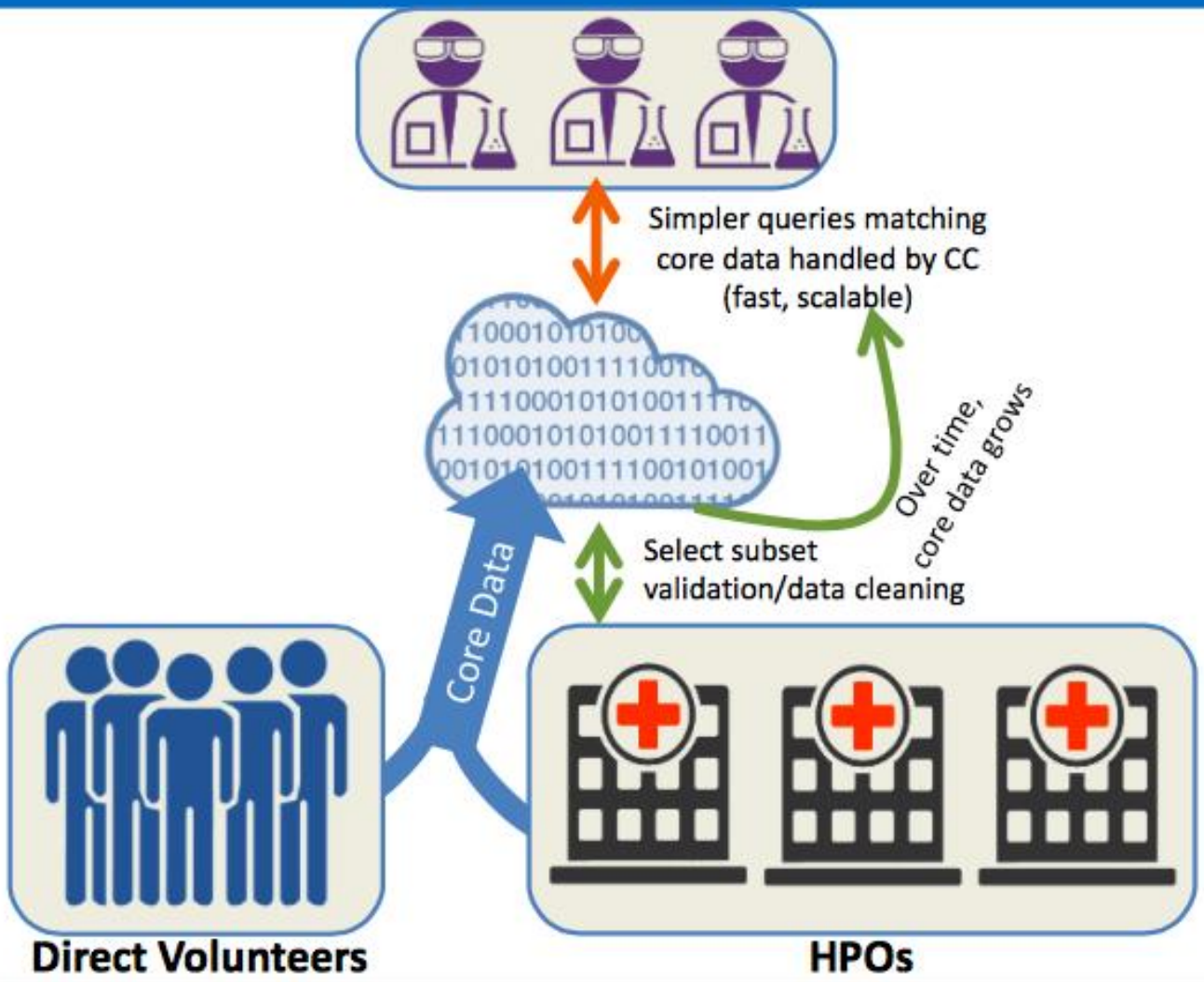
Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director

September 17, 2015

Kathy Hudson, PhD (NIH)
Rick Lifton, MD, PhD (Yale)
Bray Patrick-Lake, MFS (Duke)
Josh Denny, MD, MS (Vanderbilt)

http://acd.od.nih.gov/presentations/PMI_WG_report_2015-09-17_ACD-NN.pdf

Data Flow Between Coordinating Center (CC) and Participant Sites



http://acd.od.nih.gov/presentations/PMI_WG_report_2015-09-17_ACD-NN.pdf

ARTICLE

AJHG 2010 Apr 9;86(4):560-72

Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record

Marylyn D. Ritchie,[2,7,9] Joshua C. Denny,[5,6,9] Dana C. Crawford,[2,7] Andrea H. Ramirez,[6] Justin B. Weiner,[6] Jill M. Pulley,[3] Melissa A. Basford,[1,3] Kristin Brown-Gentry,[2] Jeffrey R. Balser,[3,4,8] Daniel R. Masys,[5] Jonathan L. Haines,[2,7] and Dan M. Roden[1,6,8,*]

Large-scale DNA databanks linked to electronic medical record (EMR) systems have been proposed as an approach for rapidly generating large, diverse cohorts for discovery and replication of genotype-phenotype associations. However, the extent to which such resources are capable of delivering on this promise is unknown. We studied whether an EMR-linked DNA biorepository can be used to detect known genotype-phenotype associations for five diseases. Twenty-one SNPs previously implicated as common variants predisposing to atrial fibrillation, Crohn disease, multiple sclerosis, rheumatoid arthritis, or type 2 diabetes were successfully genotyped in 9483 samples accrued over 4 mo into BioVU, the Vanderbilt University Medical Center DNA biobank. Previously reported odds ratios (OR_{PR}) ranged from 1.14 to 2.36. For each phenotype, natural language processing techniques and billing-code queries were used to identify cases (n = 70–698) and controls (n = 808–3818) from deidentified health records. Each of the 21 tests of association yielded point estimates in the expected direction. Previous genotype-phenotype associations were replicated ($p < 0.05$) in 8/14 cases when the OR_{PR} was > 1.25, and in 0/7 with lower OR_{PR}. Statistically significant associations were detected in all analyses that were adequately powered. In each of the five diseases studied, at least one previously reported association was replicated. These data demonstrate that phenotypes representing clinical diagnoses can be extracted from EMR systems, and they support the use of DNA resources coupled to EMR systems as tools for rapid generation of large data sets required for replication of associations found in research cohorts and for discovery in genome science.

eMERGE has demonstrated quality of phenotypes derived from EHR

https://emerge.mc.vanderbilt.edu/about-emerge/

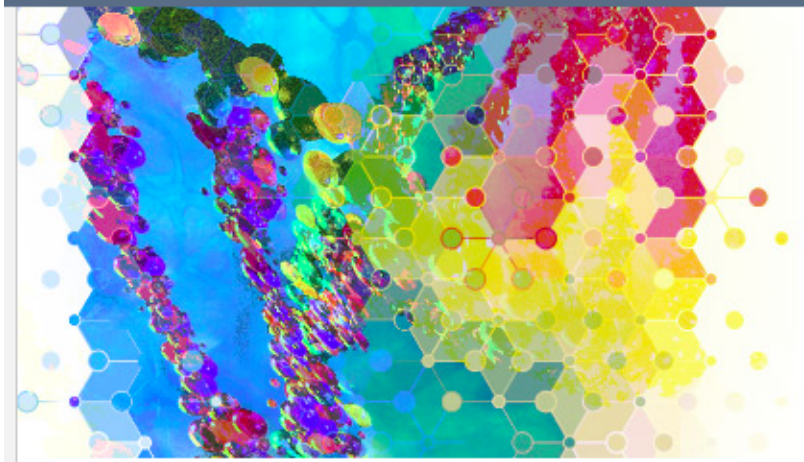
EHR biobank for genomics

- Cost is lower than de novo cohort collection
  - Add-on to routine clinical care
  - All of the clinical visit data is available
  - Longitudinal data
- Limitations
  - Usually little/no environmental exposure data

    Supplement with survey tools and geocoding
  - Usually little/no behavioral data

    Supplement with survey tools, questionnaires, apps
  - Biased to clinic populations → potential inference issues

    To what population do you want to make inferences?



mycode

MyCode® Community Health Initiative

The MyCode® Community Health Initiative includes a Geisinger system-wide biobank designed to store blood and other samples for research use by Geisinger and Geisinger collaborators. A biobank is like a bank, but instead of securely storing money, it securely stores your blood or saliva sample and information, along with the samples and information from thousands of other Geisinger patients. Samples and information in the biobank are used to do health research.

Ultimately, our goal is to find ways to make health care better - for you, your family, your community and individuals around the world.

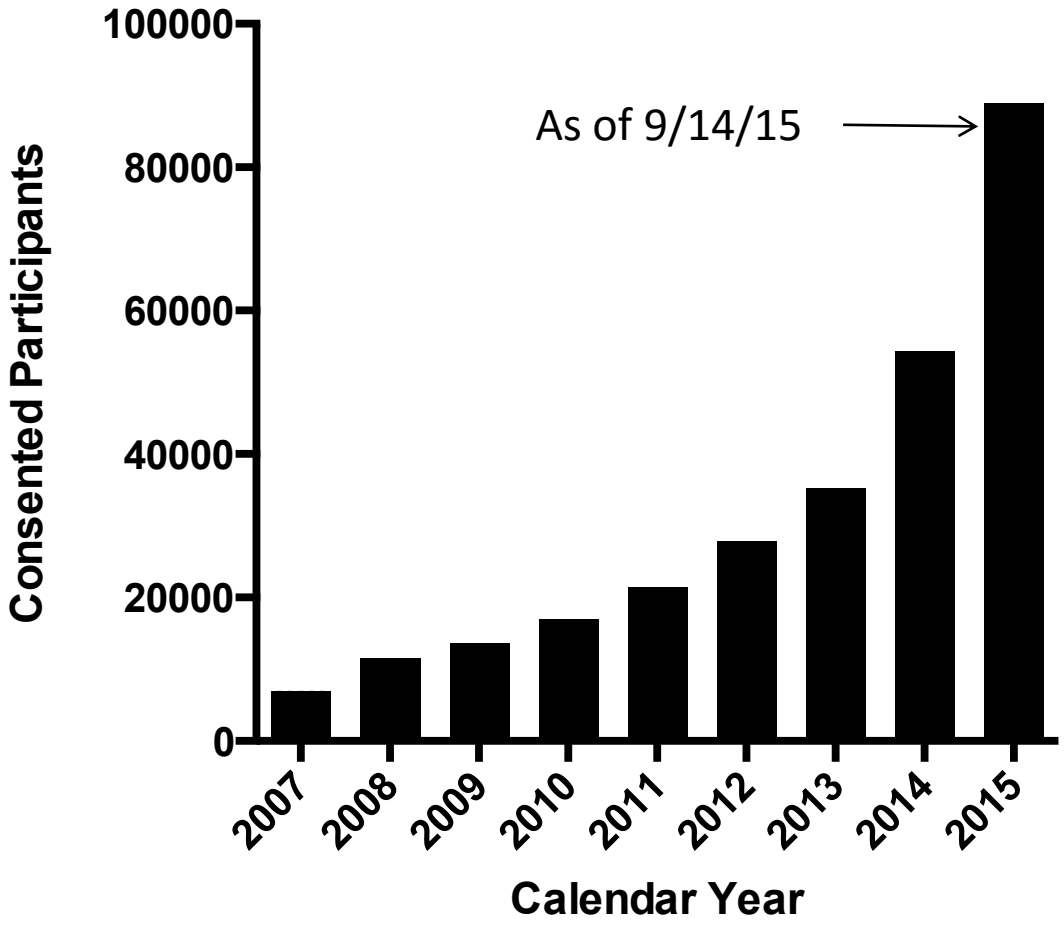
Partnering with Patients

About
Clinical Trials
Donate to Research
Family History Project
MyCode® Community Health Initiative

MyCode Community Health Initiative

- Broad based biobank
- Fully consented
  - Access to EHR
  - Data sharing (collaborators and public)
  - Return of results
  - Re-contact
    - New biospecimens, samples, information, data
- Implementing
  - Electronic consent
  - Exposome, social, behavioral surveys
  - mHealth

MyCode Enrollment



~50,000 participants whole exome sequenced

~30,000 GWAS genotyped

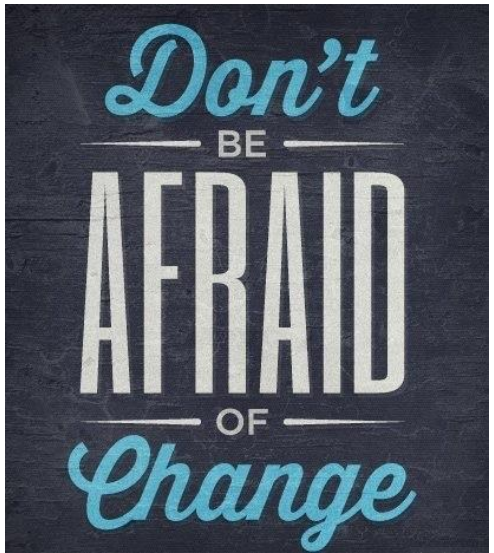
mycode

Which analysis strategies should be employed?


TIME FOR CHANGE

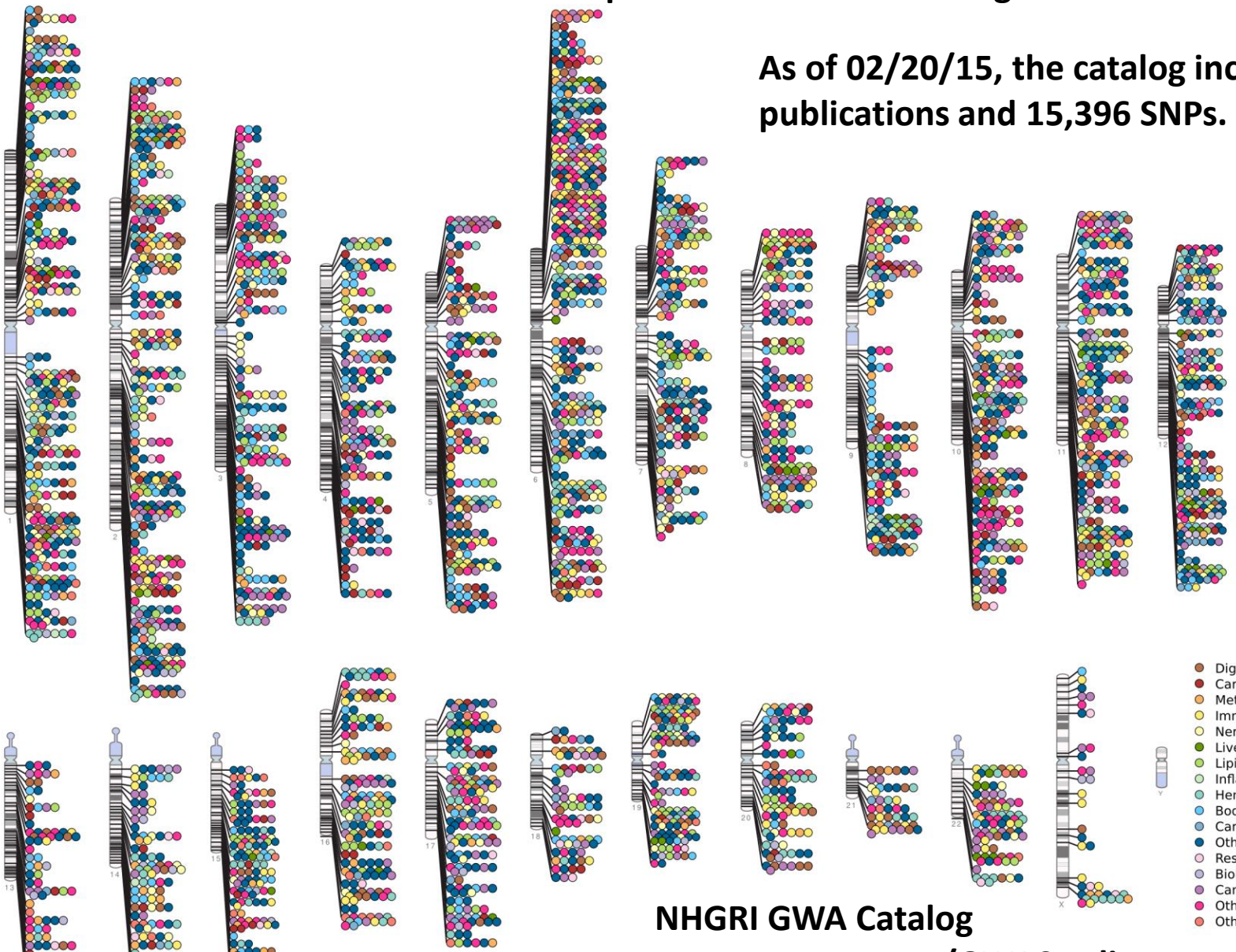

KEEP
CALM
IT'S TIME
FOR
CHANGE


Changes
NEXT EXIT


Don't
BE
AFRAID
OF
Change

Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

As of 02/20/15, the catalog includes 2,111 publications and 15,396 SNPs.

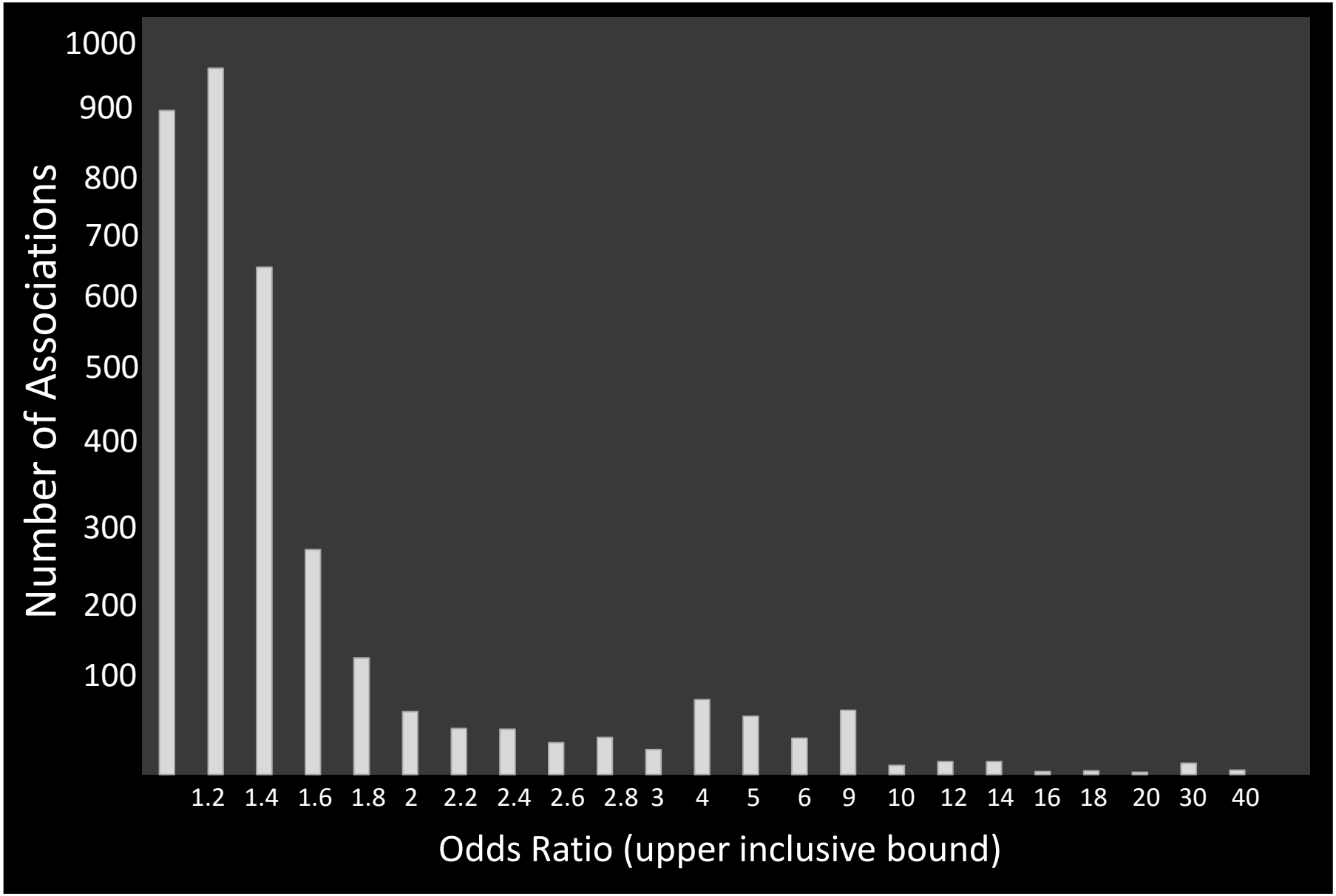


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

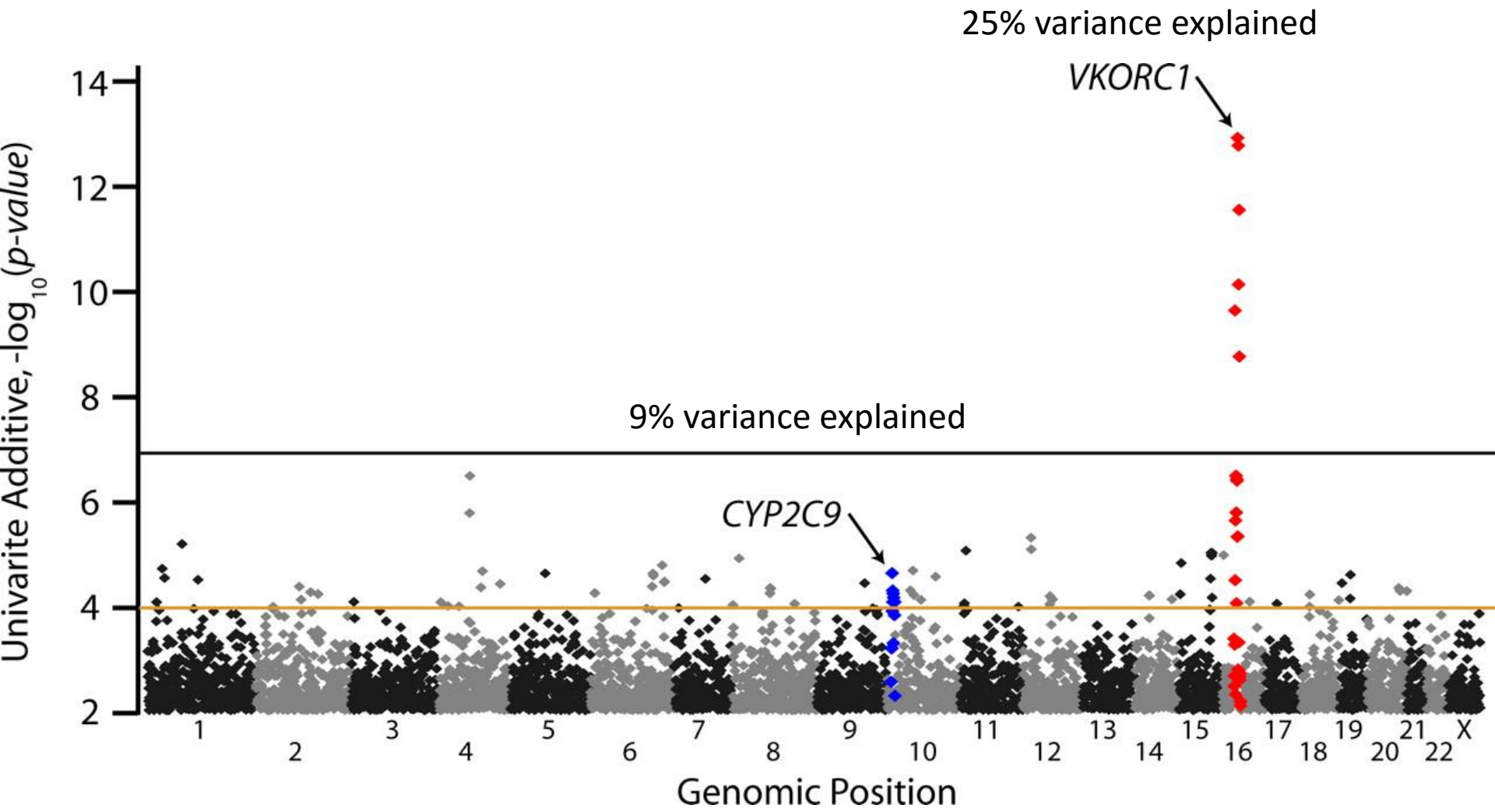
National Human Genome Research Institute

EMBL-EBI

Distribution of Effects

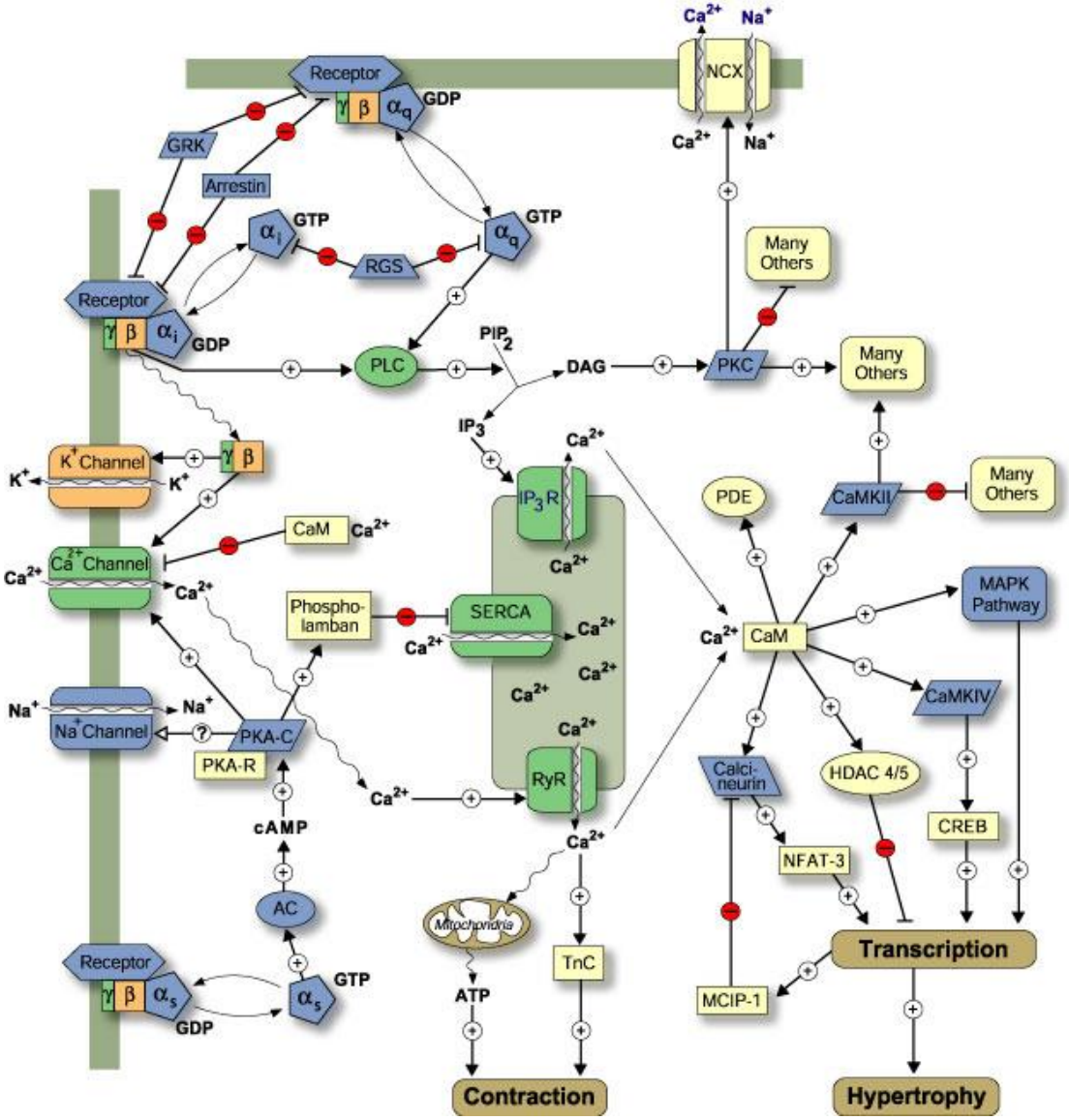


Marylyn Ritchie, Jan 2014



A multivariate model, including age, gender, treatment with amiodarone, treatment with losartan, *VKORC1* genotype (rs9923231), *CYP2C9* carrier status (either *2 or *3), and patient weight (n = 507) increased the total predicted dose variance to approximately 47%.

Biology is complex



Biology is complex

To explore genetic architecture we need…

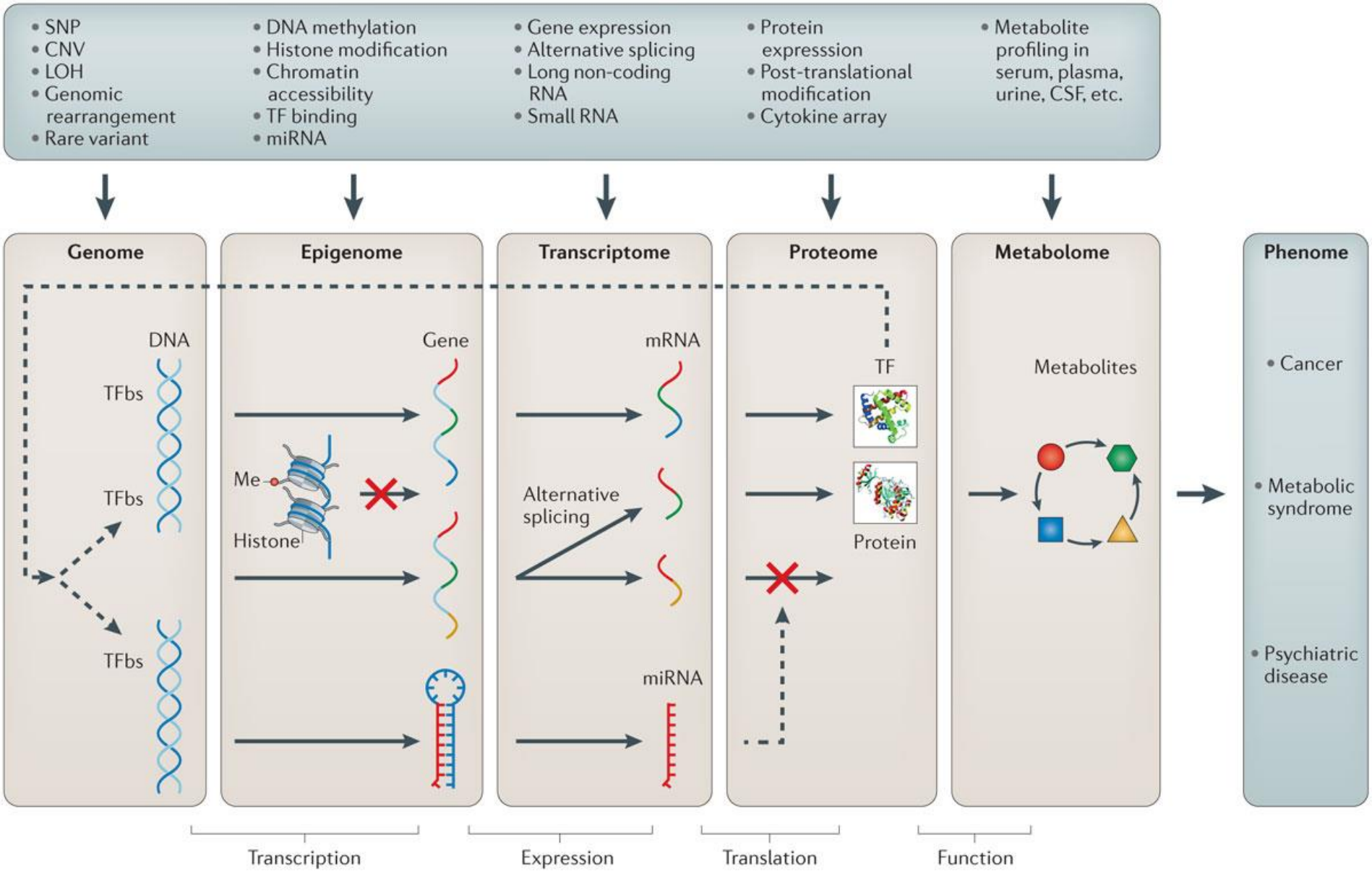
- Large sample size
  - Not simply large because we want to detect OR ~1.1
  - Large so that we can subset into clinical meaningful sets and explore complex genomic effects (GxG, GxE, meta-dimensional)
- Rich, longitudinal phenotypic data
  - Many different phenotypes
- Comprehensive genomic data
  - DNA sequence/genotyping, transcriptome, metabolome, etc.
- Environmental and behavioral data
  - Surveyed through health system tools along with geocoding, mobile apps, etc.
- Powerful analytic tools
  - Holistic data-driven approaches

Methods of integrating data to uncover genotype–phenotype interactions

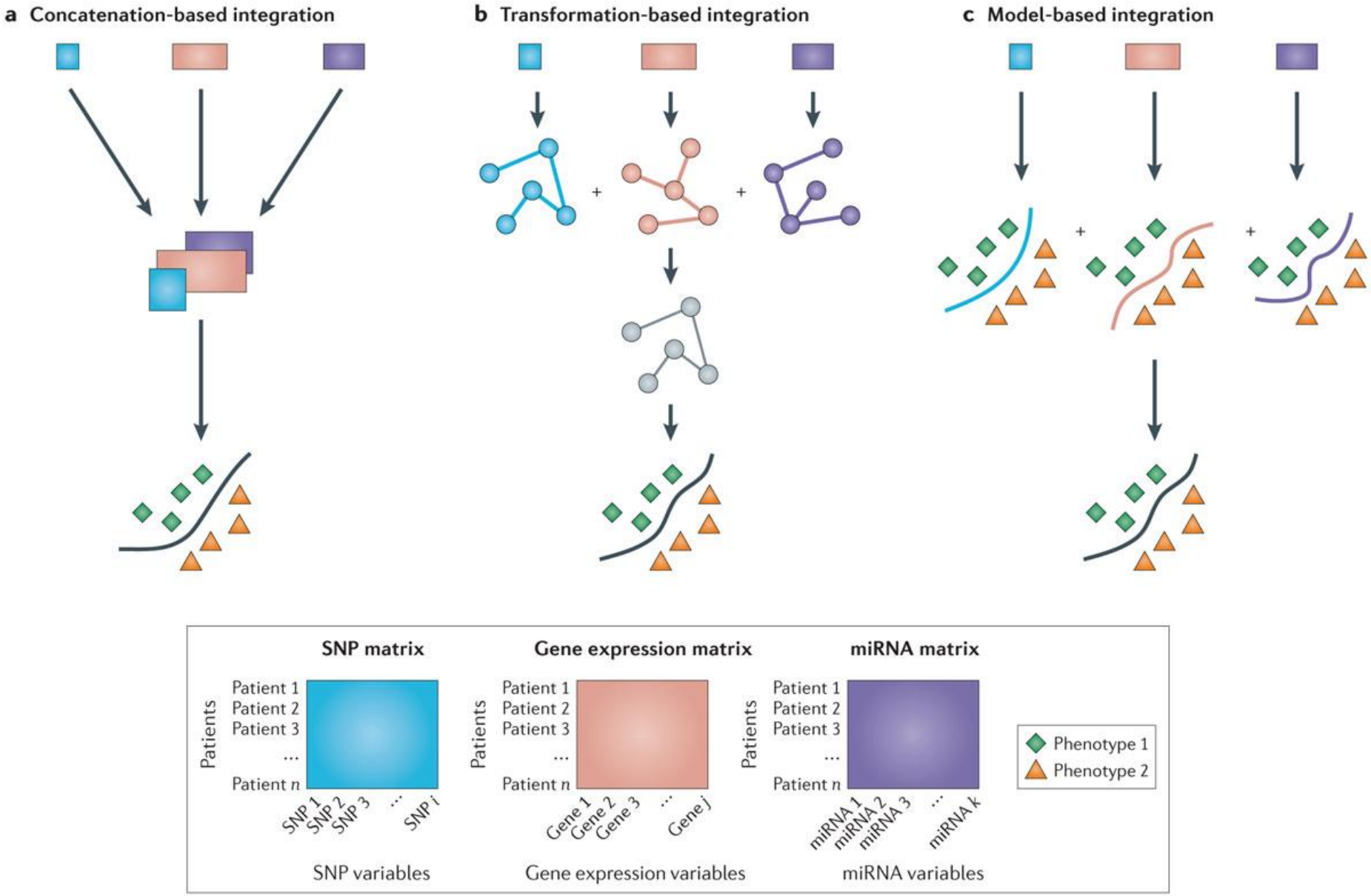
Marylyn D. Ritchie[1], Emily R. Holzinger[2], Ruowang Li[1], Sarah A. Pendergrass[1] and Dokyoon Kim[1]

Abstract | Recent technological advances have expanded the breadth of available omic data, from whole-genome sequencing data, to extensive transcriptomic, methylomic and metabolomic data. A key goal of analyses of these data is the identification of effective models that predict phenotypic traits and outcomes, elucidating important biomarkers and generating important insights into the genetic underpinnings of the heritability of complex traits. There is still a need for powerful and advanced analysis strategies to fully harness the utility of these comprehensive high-throughput data, identifying true associations and reducing the number of false associations. In this Review, we explore the emerging approaches for data integration — including meta-dimensional and multi-staged analyses — which aim to deepen our understanding of the role of genetics and genomics in complex outcomes. With the use and further development of these approaches, an improved understanding of the relationship between genomic variation and human phenotypes may be revealed.

Nat Rev Genet. 2015 Feb;16(2):85-97.



Nat Rev Genet. 2015 Feb;16(2):85-97.



Nature Reviews | Genetics

Nat Rev Genet. 2015 Feb;16(2):85-97.

Journal of Biomedical Informatics 56 (2015) 220–228

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer

Dokyoon Kim [a], Ruowang Li [a], Scott M. Dudek [a], Marylyn D. Ritchie [a,b,*]

[a] Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA
[b] Geisinger Health System, Danville, PA, USA

Multi-omics Data
Copy number variation
Gene Expression
Methylation
Protein expression

Outcome
Breast Cancer Survival

Journal of Biomedical Informatics 56 (2015) 220–228

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer

Dokyoon Kim [a], Ruowang Li [a], Scott M. Dudek [a], Marylyn D. Ritchie [a,b,*]

[a] Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA
[b] Geisinger Health System, Danville, PA, USA

Performance comparison between the model from single dimensional genomic data and integration model. Performance was measured from the validation dataset.

Data type	1- MAD
CNA	0.63
Methylation	0.63
Gene expression	0.69
Protein expression	0.64
Integration	0.73

Kim et al. BioData Mining 2014, 7:20
http://www.biodatamining.org/content/7/1/20

BioData Mining

RESEARCH Open Access

Knowledge-driven genomic interactions: an application in ovarian cancer

Dokyoon Kim, Ruowang Li, Scott M Dudek, Alex T Frase, Sarah A Pendergrass and Marylyn D Ritchie*

Kim *et al. BioData Mining* 2014, **7**:20
http://www.biodatamining.org/content/7/1/20

BioData Mining

RESEARCH Open Access

Knowledge-driven genomic interactions: an application in ovarian cancer

Dokyoon Kim, Ruowang Li, Scott M Dudek, Alex T Frase, Sarah A Pendergrass and Marylyn D Ritchie*

Performance comparison between the model with gene expression data alone and models identified using knowledge-based matrices

Data type	Balanced accuracy	AUC
Gene expression	0.6957	0.7103
Pathway	0.7451	0.7457
GO	0.6991	0.7275
Pfam	0.7046	0.7335
Integration	0.7882	0.8108

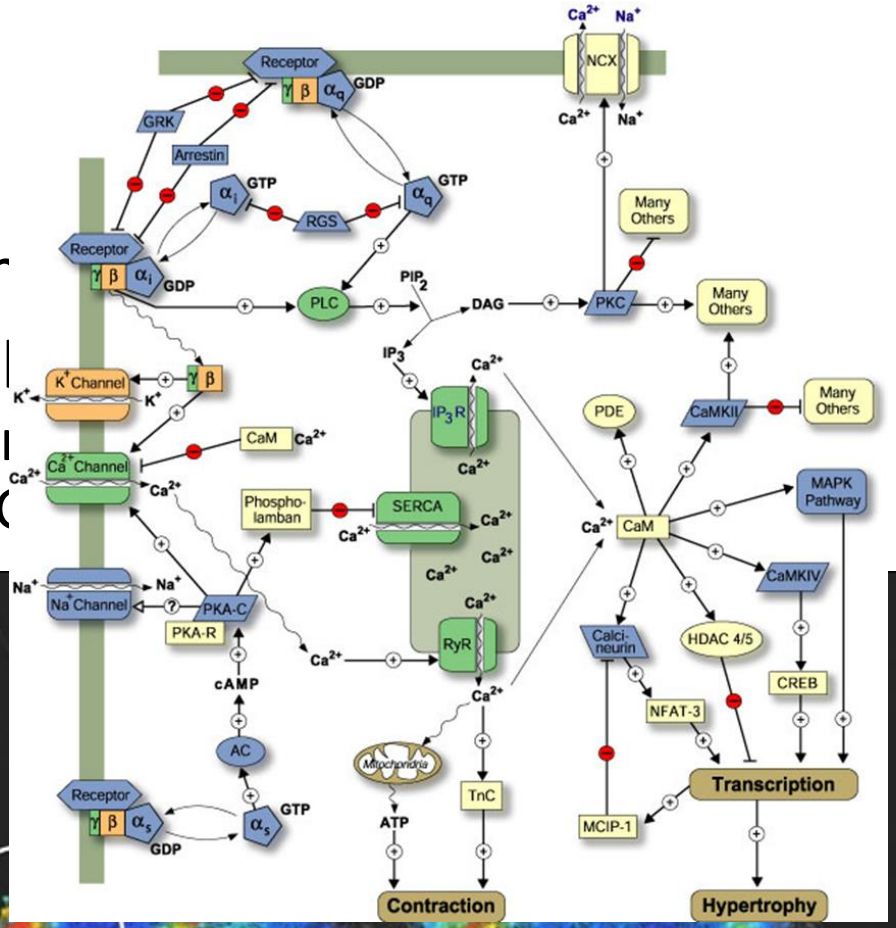
Potential limitations

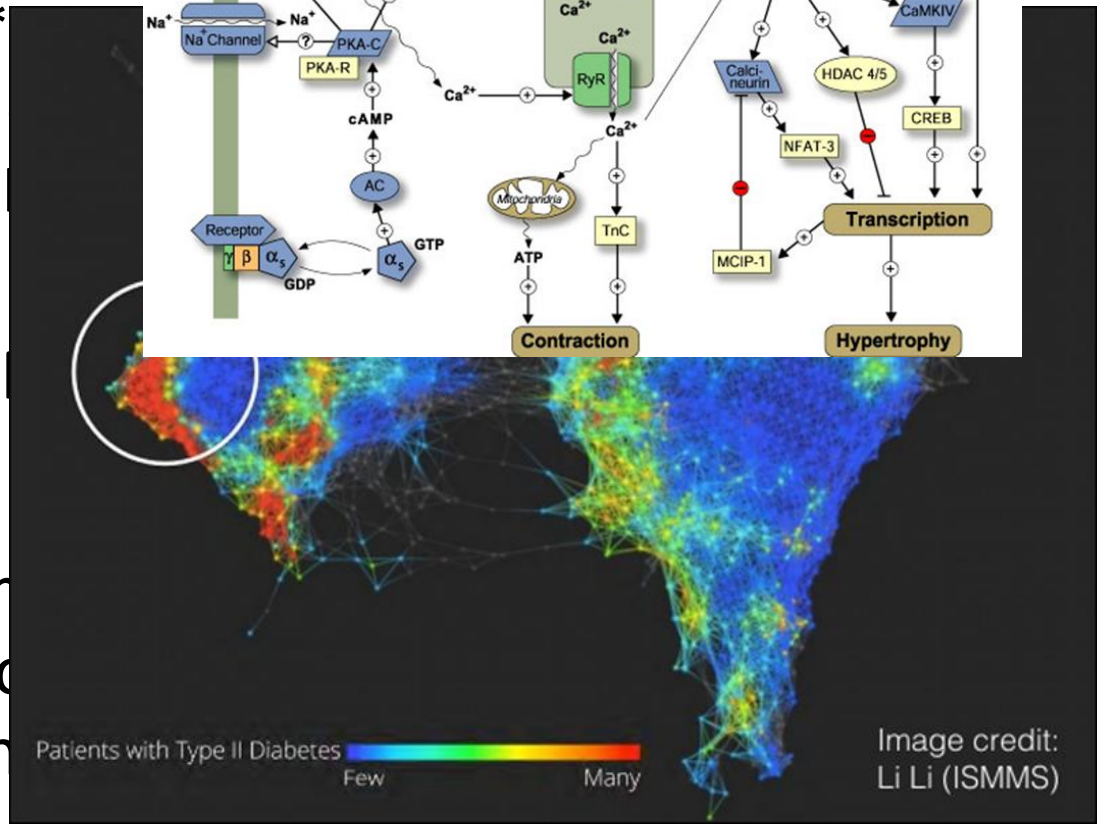
- Data are incomplete
- Biological knowledge is incomplete
- Network connections may be incomplete
- Topology may be incomplete/incorrect

Do we need to know everything and have every data point to make inferences and learn new biology?

Summary

- Biobanks linked to electron
  study designs for genome-
  - Large sample size for creati
    exploring complex effects (
  - Rich phenotypic data*
  - Comprehensive 'omic
  - Environmental and be
- Going beyond single S
- Systems genomics or
  enable better predictio
  treatment, and preven





Precision Medicine Initiative

Acknowledgements

Yuki Bradford, bioinformatics analyst
Marta Byrska-Bishop, Post doc
Scott Dudek, software developer
Alex Frase, software developer
Molly Hall, PhD student
Dokyoon Kim, Post-doc
Ruowang Li, PhD student
Anastasia Lucas, bioinformatics analyst
Donna McMinn, administrative assistant
Anna Okula, PhD student
Suzy Unger, program coordinator
Anurag Verma, bioinformatics programmer
Shefali Verma, bioinformatics analyst
John Wallace, software developer

Acknowledgements

Will Bush

Dana Crawford

Jonathan Haines

David Carey

David Ledbetter

Sarah Pendergrass

Just because we have not found it yet, doesn't mean it's not there…..

www.genetic-programming.org

- marylyn.ritchie@psu.edu
- http://ritchielab.psu.edu