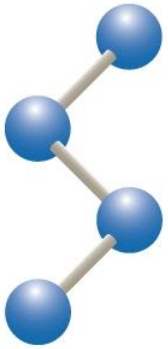


THE SCRIPPS RESEARCH INSTITUTE

Scripps Translational Science Institute

# Genome and Medical Information – Portability, Retrieval and Analysis

Amalio Telenti, MD PhD

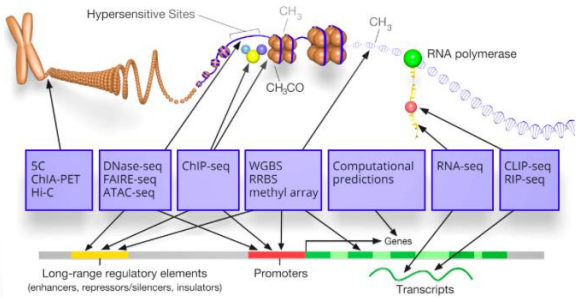
The Scripps Research Institute

- Genomic data
- Information retrieval
- Data analyses
- Data protection

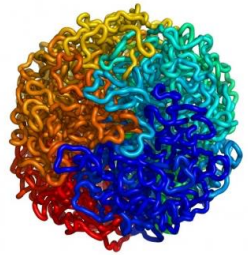
# 2001

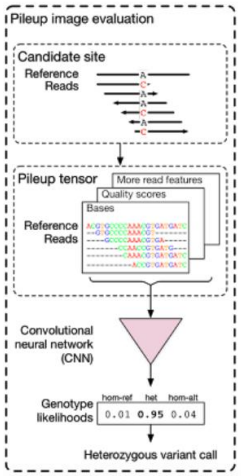
# Gigabyte → Petabyte

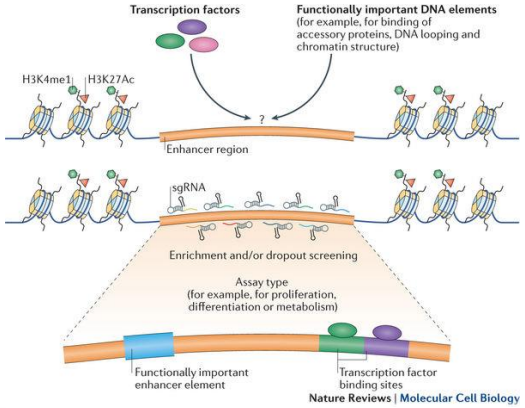
- Human Genome
  2001
- Encode project
  2003
- GTEX
  2008
- Nucleosome 3D genome
  ~2014
- Deep learning in genomics
  - Functional screens (CRISPR and parallel reporters)
  ~2015
- Large scale deep sequencing of the human population
  ~2016




GTEx



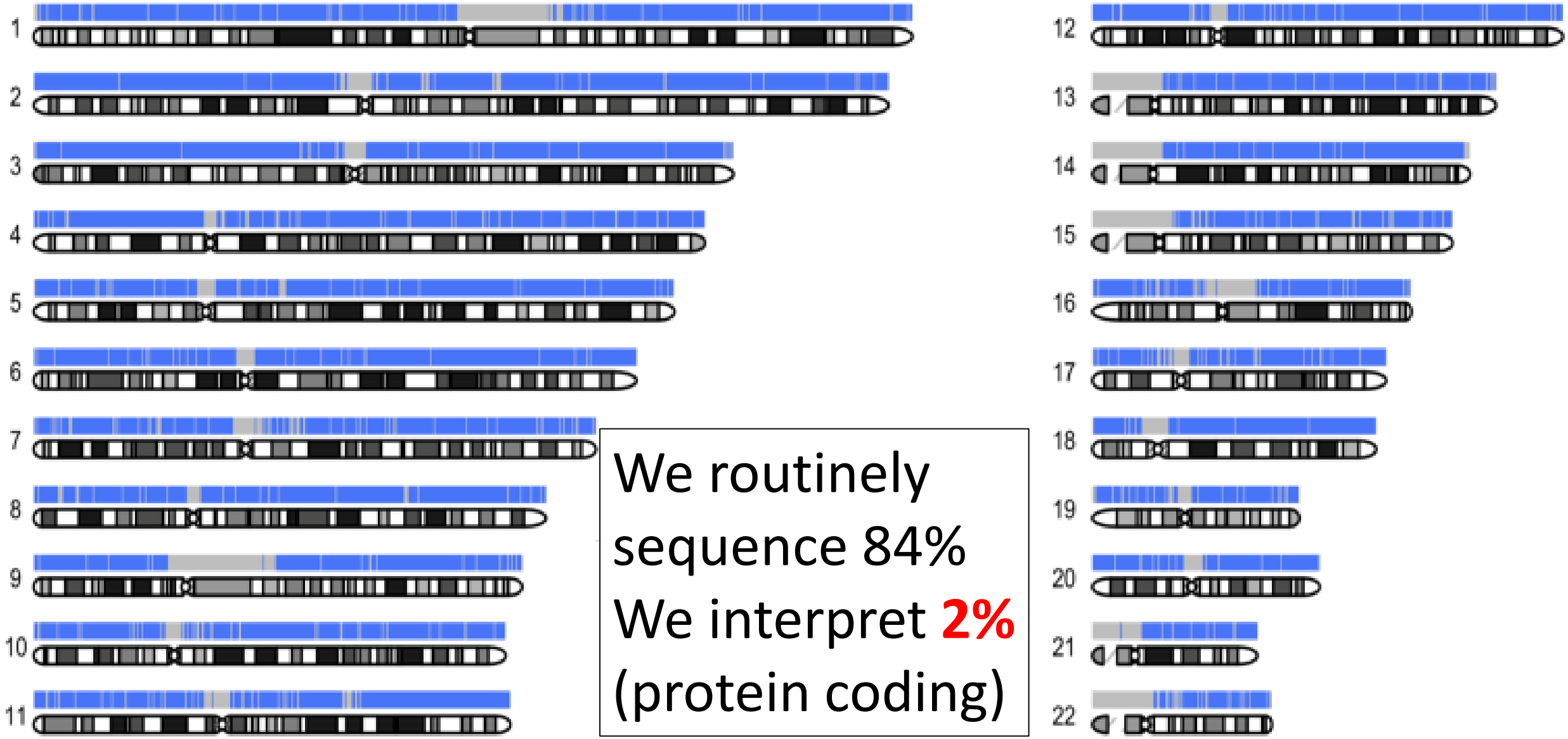




gnomAD Genome Aggregation Database

*Telenti, Lippert, Chang, DePristo. Hum Mol Genet, 2018*




We routinely sequence 84%
We interpret **2%** (protein coding)

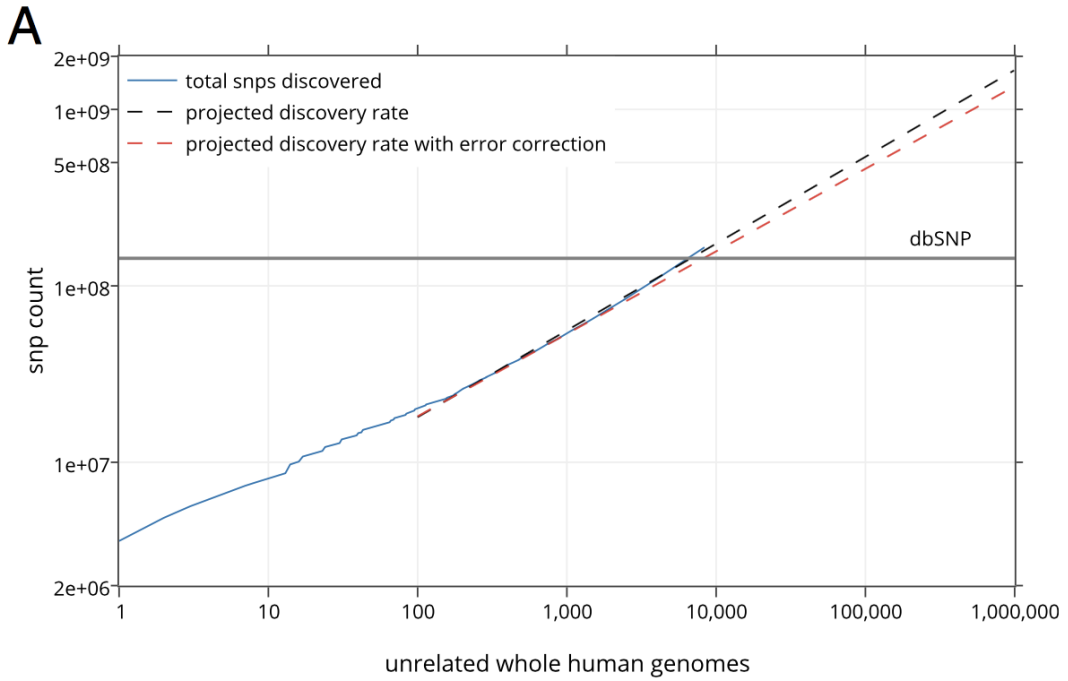
*Telenti PNAS 2016*

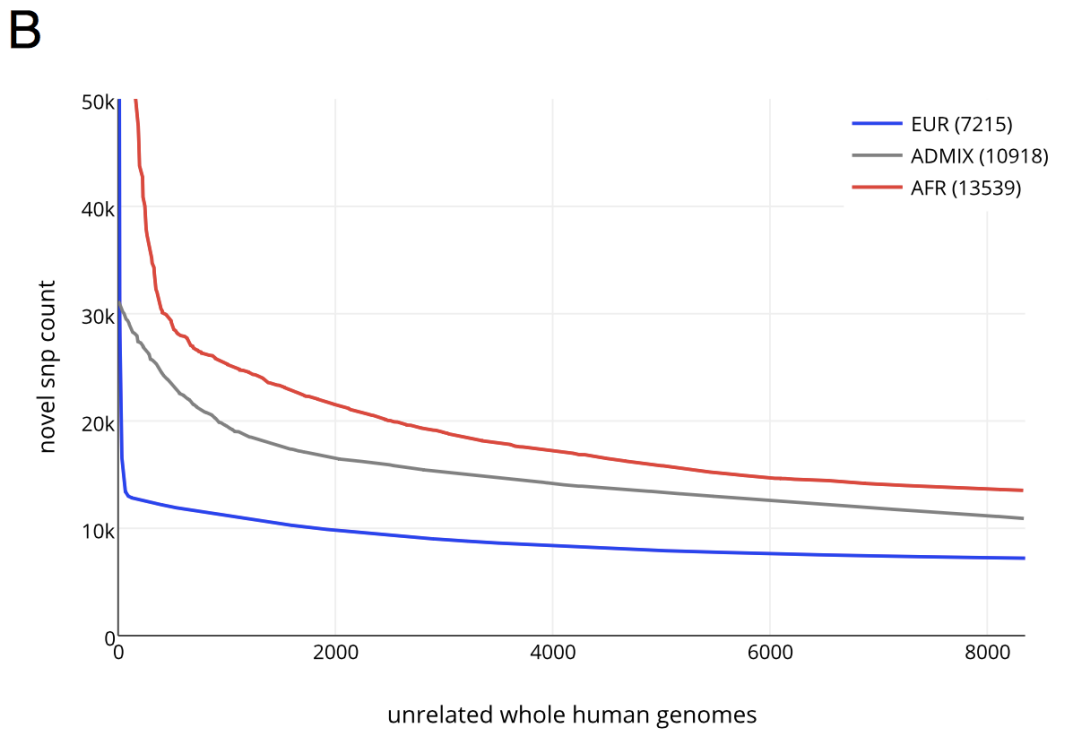
# Genome variation 101

- Human genome = $3.2 \times 10^9$ base pairs.
- Mid-2017:
  - 150 million SNVs in the public database dbSNP
  - 10 million coding variants in ExAC
  - 150 million SNVs in 10,545 deeply-sequenced whole genomes
- The union of these resources = 242 million unique SNVs
- **1 out of every 13 nucleotides in the genome has been observed variant in the population**

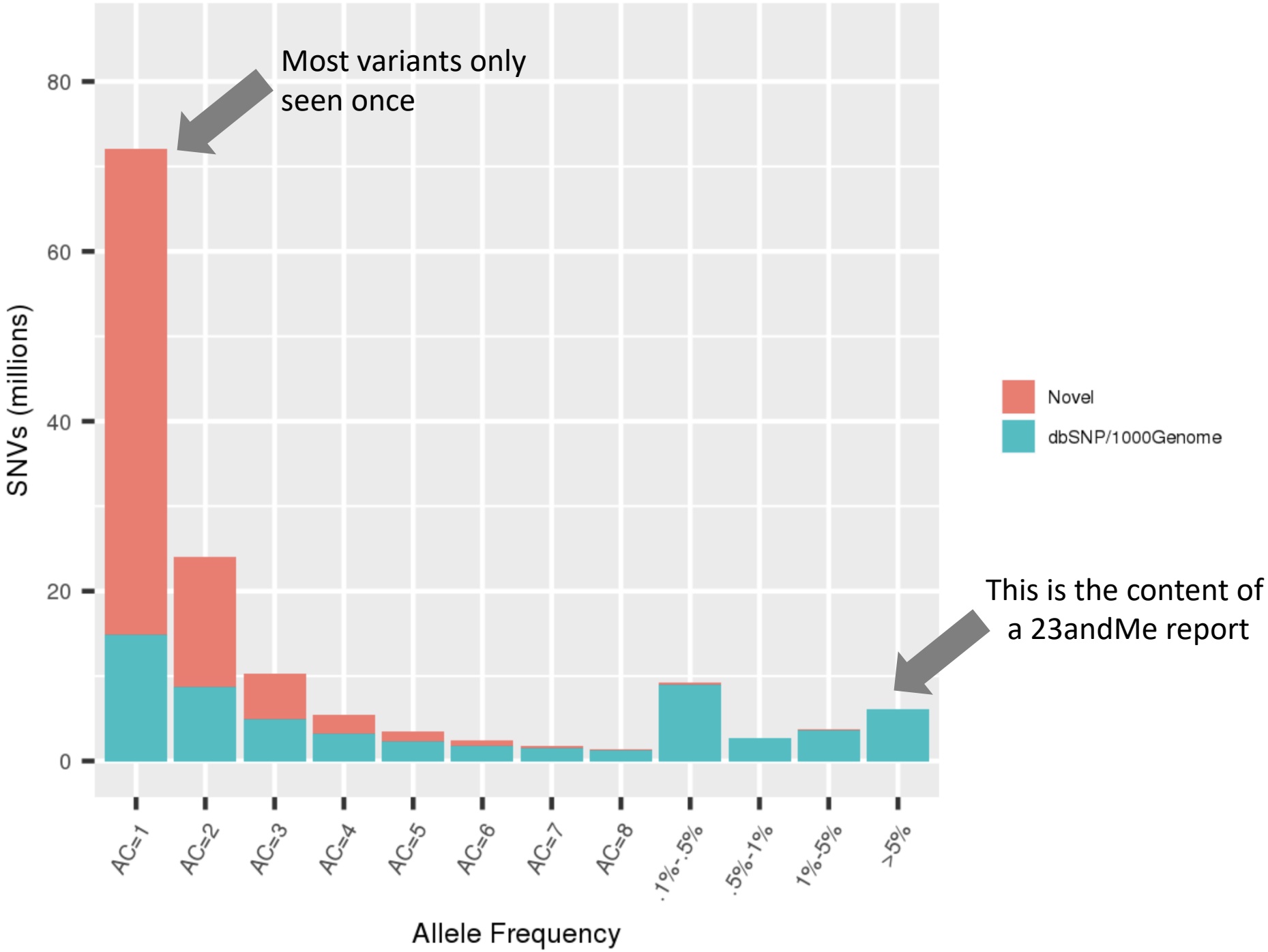
1,000,000 whole genome sequences
= 1 every third nucleotide variant

Every additional sequenced genome
= 8,000 new variants



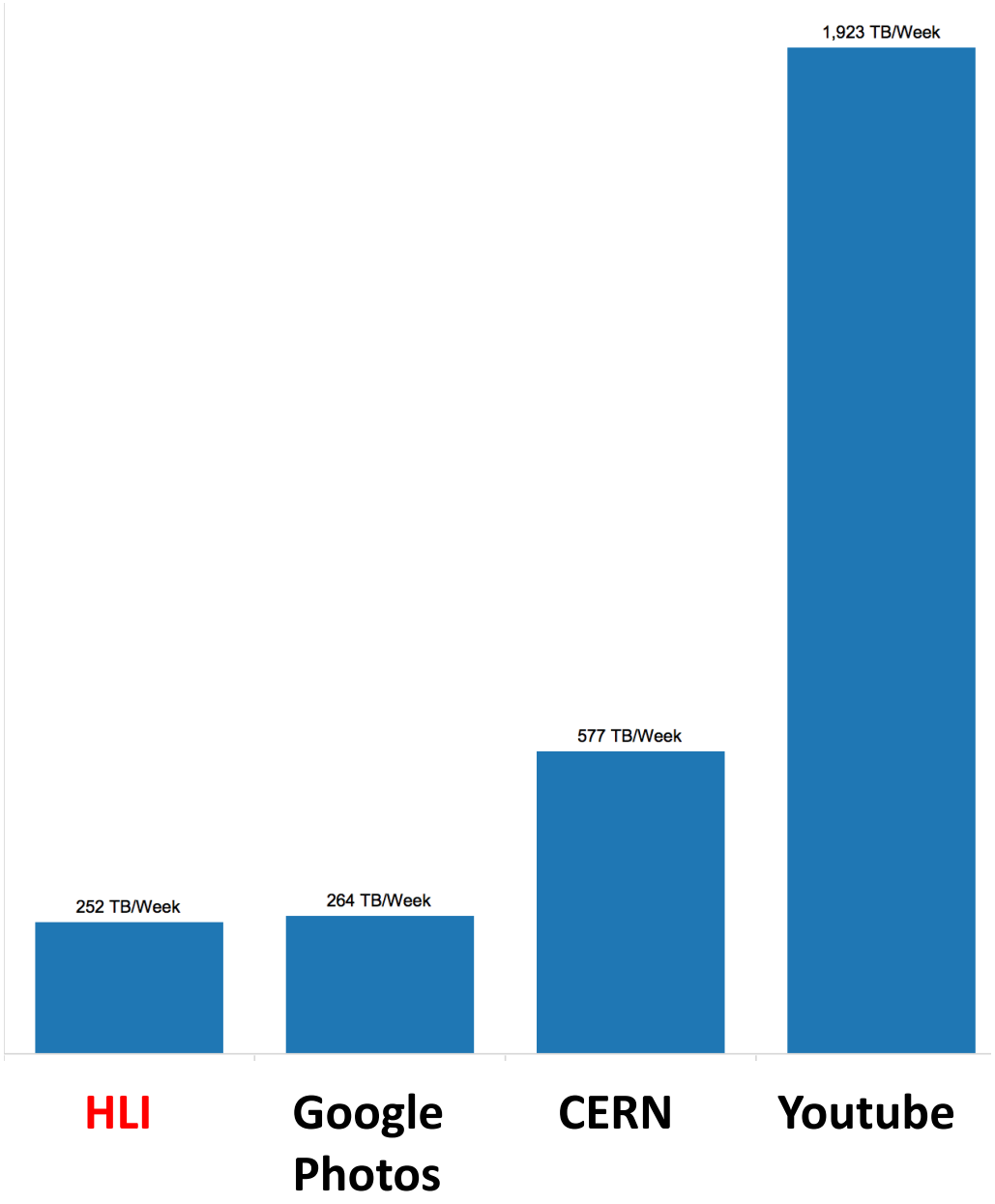




*Telenti et al. PNAS, 2016*




*Telenti et al. PNAS, 2016*

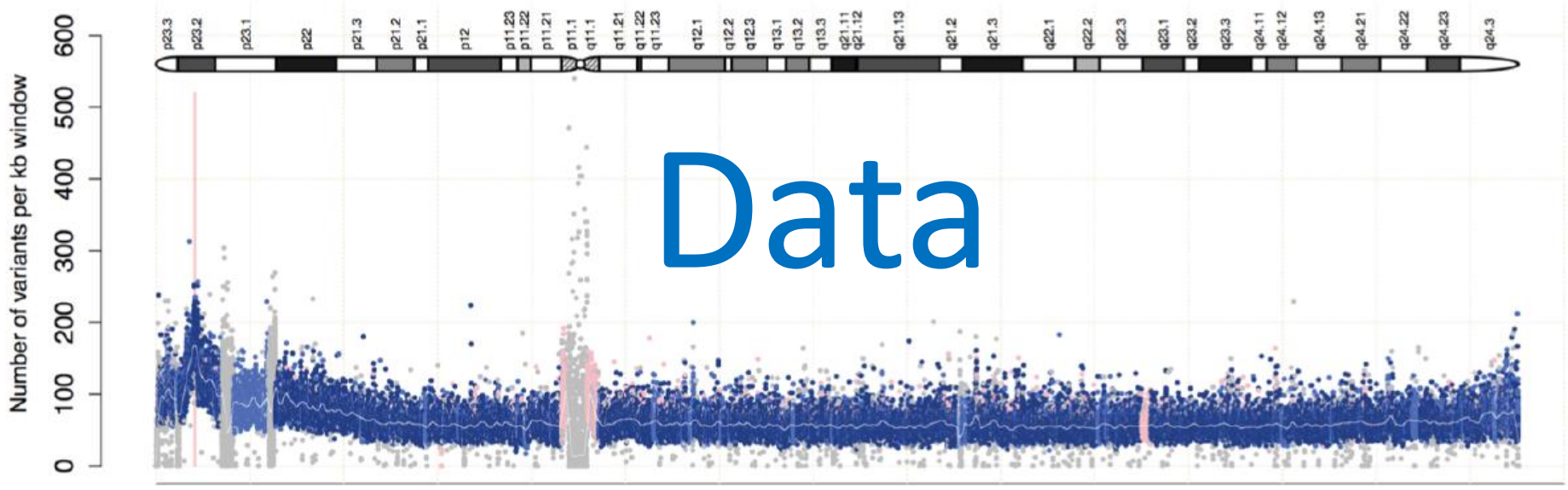
# HLI Cloud Data Volume per Week




## Human Longevity, Inc.

- In the top 1% of all AWS customers
  - The only customer to ever hit 1% in under 2 years
- Close to 1,000,000 EC2 (elastic cloud compute) instance hours used per month
  - 11 years of core hours used per day to just to analyze raw sequence data
- 8PB of data per year and growing
- Over 20PB of genomics data stored today
  - 1 Million WGS will require nearly 1 Exabyte of data

*Source: Human Longevity Inc.*




# Data

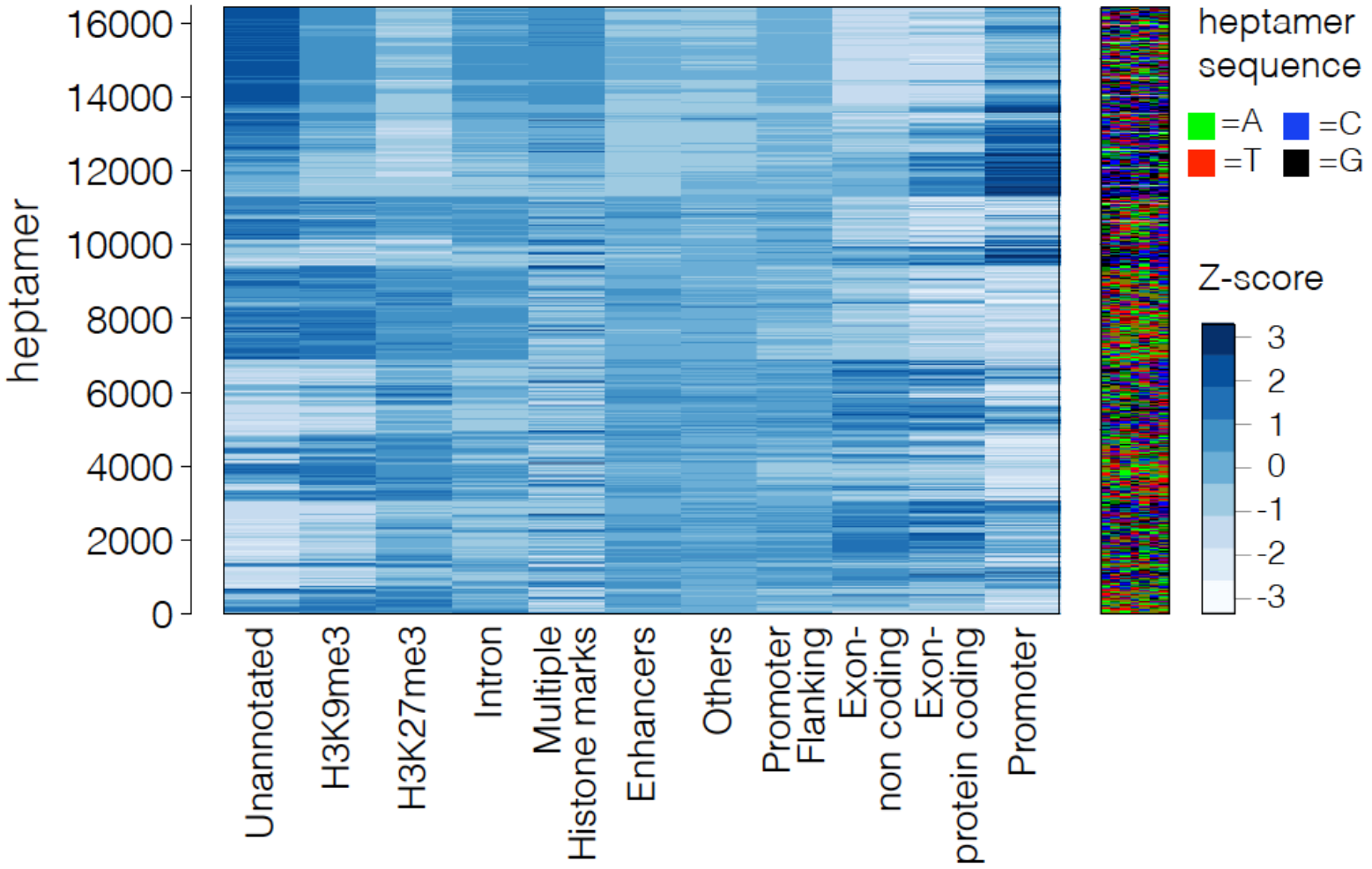
# Patterns

## New Genomic elements

## Conservation of function

## Essentiality

# Patterns *Your Genome in 16,384 Bricks*




## Heptamer composition:

**Most uncommon** = TACGCGA
n=1,691 x 10,000 genomes = 16,910,000 datapoints

**Most common** = AAAAAAA
n=4,346,493 x 10,000 genomes = 43,464,930,000 datapoints

*Di Iulio et al. Nat Genet 2018*

# Take home message - 1

- Current pace of generation of genome and other OMICs data is heading towards the Exabyte.
- The burden arises from the raw sequencing data
- The computing cost associated with moving data and from specific analyses that use the raw data

- Genomic data
- Information retrieval
- Data analyses
- Data protection

# Volume of data currently in use:
# 10 times the library of Congress

*<u>One</u> library is:*

- 838 miles of bookshelves
- 155 million items
- Three buildings on Capitol hill

Thomas Jefferson Building and part of the John Adams Building (upper-right)

John Adams Building

Madison Building

Search history: [clear]
10:01 cancer (35,464)
10:01 trim5 (3,412)

# OPEN SEARCH BETA

cancer

200,000,000+ distinct variants in 10,000+ genomes

found 35,464 variants (16/104ms)

<0.1% **chr1**:146018516 **A→G** | p.Ile281Thr | c.842T>C [ucsc]
missense Hemochromatosis type 2A [clinvar]
HFE2 Significance: Pathogenic, variation to disease, no assertion criteria provided.
rs74315326 Origin: germline. MedGen:C1865614. Orphanet:79230. OMIM:602390. :.
Allele frequency: admix:<0.1%

<0.1% **chr3**:37012044 **C→T** | p.Pro208Ser | c.622C>T [ucsc]
missense not specified [clinvar]
MLH1 Significance: Uncertain significance, variation to disease, criteria provided, single
rs587781509 submitter. Origin: germline. MedGen:CN169374. :. :. :.
Allele frequency: eur:<0.1%

<0.1% **chr5**:112838439 **A→G** | p.Met949Val | c.2845A>G [ucsc]
missense Hereditary cancer-predisposing syndrome [clinvar]
APC Significance: Uncertain significance, variation to disease, criteria provided, single
rs587781348 submitter. Origin: germline. MedGen:C0027672. :. :. :.
Allele frequency: eur:<0.1%

<0.1% **chr5**:132642204 **G→A** | p.Arg1260His | c.3779G>A [ucsc]
missense Hereditary cancer-predisposing syndrome [clinvar]

# Differences in performance between **information retrieval** through indexing vs **conventional databases**

## Considerations

- Data to store (text, numeric, structured)
- Queries  to run (lookup, Boolean, free-text)
- Other options  (incremental updates, fault tolerance, atomicity, ability to delete records, access control)
- A conventional database provides more features and easier setup, but is slower and more resource-hungry that an IR index
- The difference in speed becomes more pronounced as queries become longer and as data becomes less-structured
- IR indices are also easier to scale to PB-sized data

Sample comparison over Wikipedia (textual data, biased to favor IR):

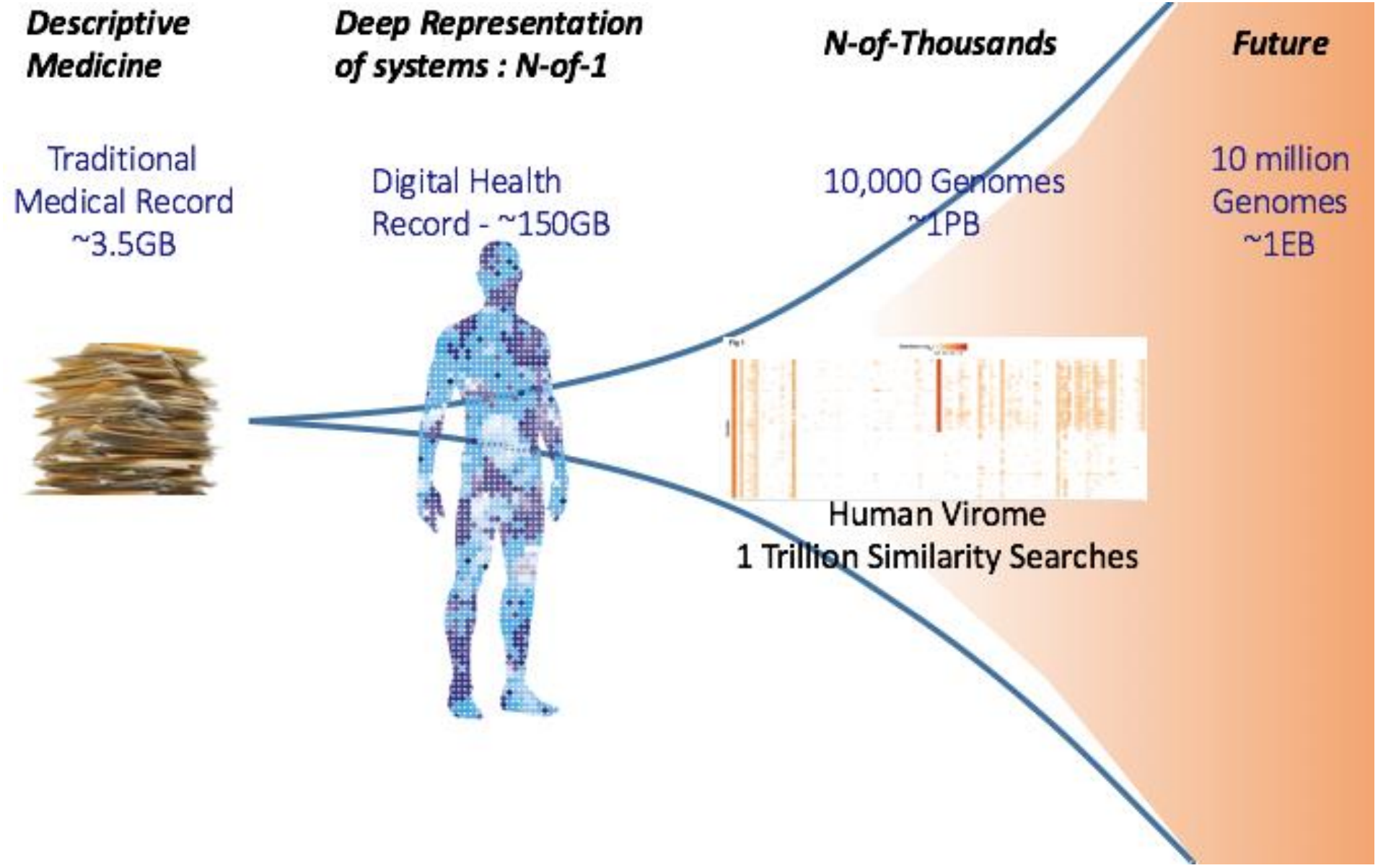
- - time to index: 84min for DB vs. 18min for IR
- - index footprint: 3GB for DB, 2.9GB for IR
- - free-text query: 286ms for DB, 22ms for IR
- - Boolean query: 24ms for DB, 13ms for IR
- - phrase query: 3692ms for DB, 21ms for IR

Sources:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.636.5266&rep=rep1&type=pdf
http://ceur-ws.org/Vol-1226/paper36.pdf
https://www.programmierer-forum.de/files/2009/01/eurooscon2006_high_performance_fulltext_search.pdf
https://pdfs.semanticscholar.org/509f/e1a5c08012268ac57ea7f7915a910002fccc.pdf

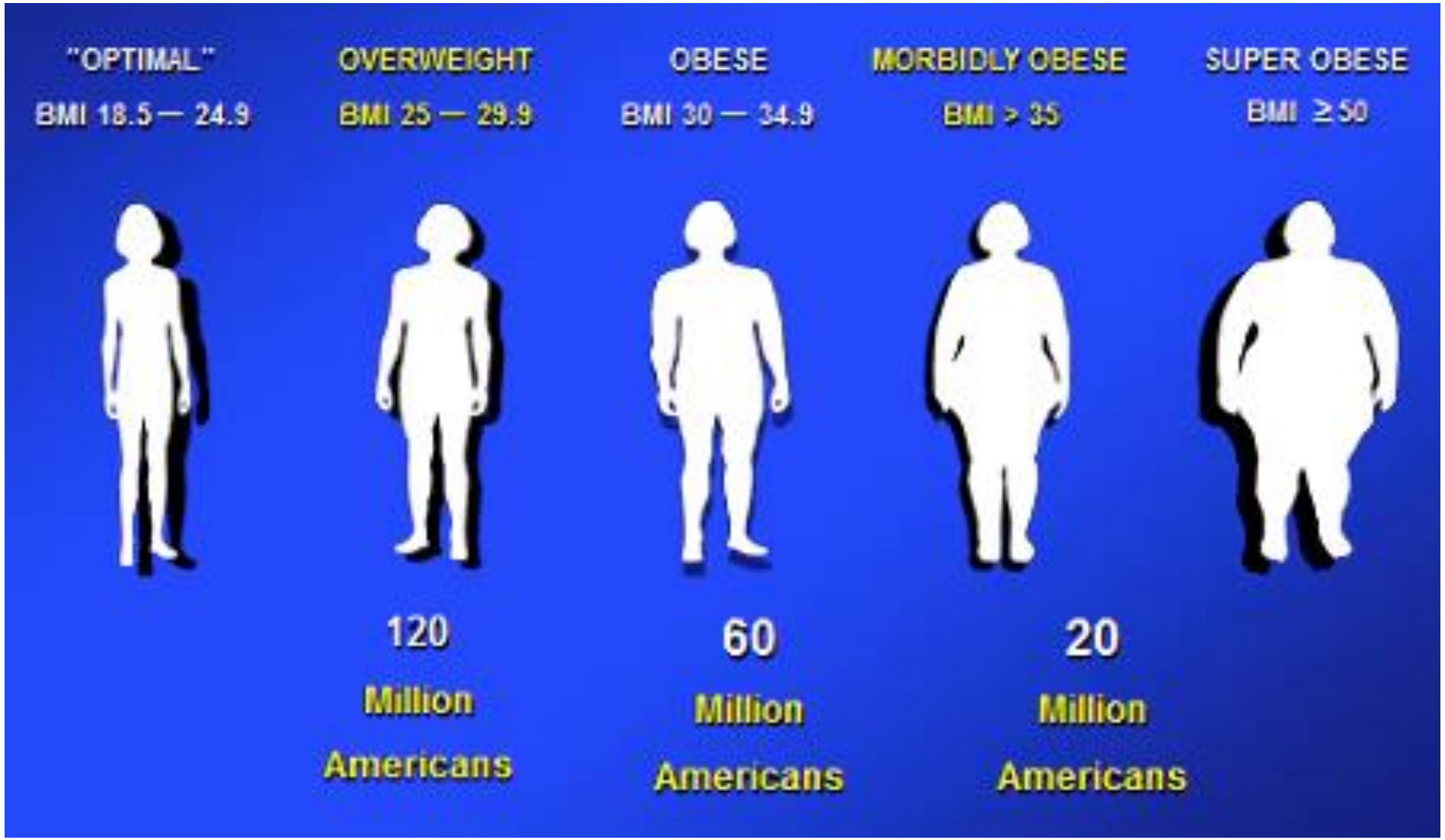
The Scripps Research Institute

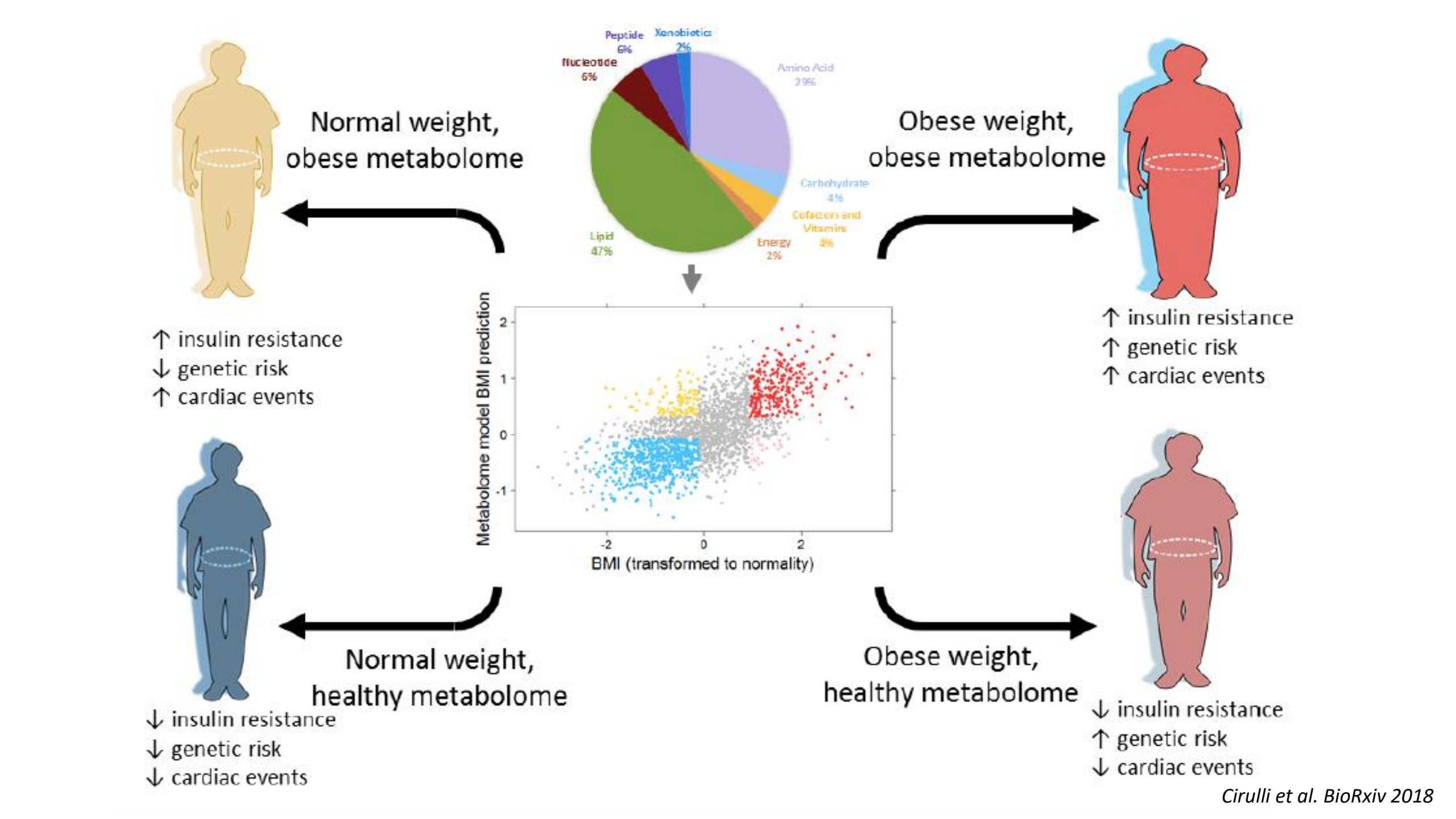
- Genomic data
- Information retrieval
- Data analyses
- Data protection

# Medicine as Data Science



# Case scenario – Morbid Obesity






*Cirulli et al. BioRxiv 2018*

# “Deep learning needs deep data”

*Eric Topol*

*Telenti, Steinhubl, Topol. Lancet , March 2018*

# Deep Learning

## DNA Binding predictions

DeepBind and DeeperBind (CNN-LSTM)

Basset (CNN)

DanQ (CNN)

DeepMotif (CNN)

## Variant callers, pathogenicity scores and identification of genomic elements

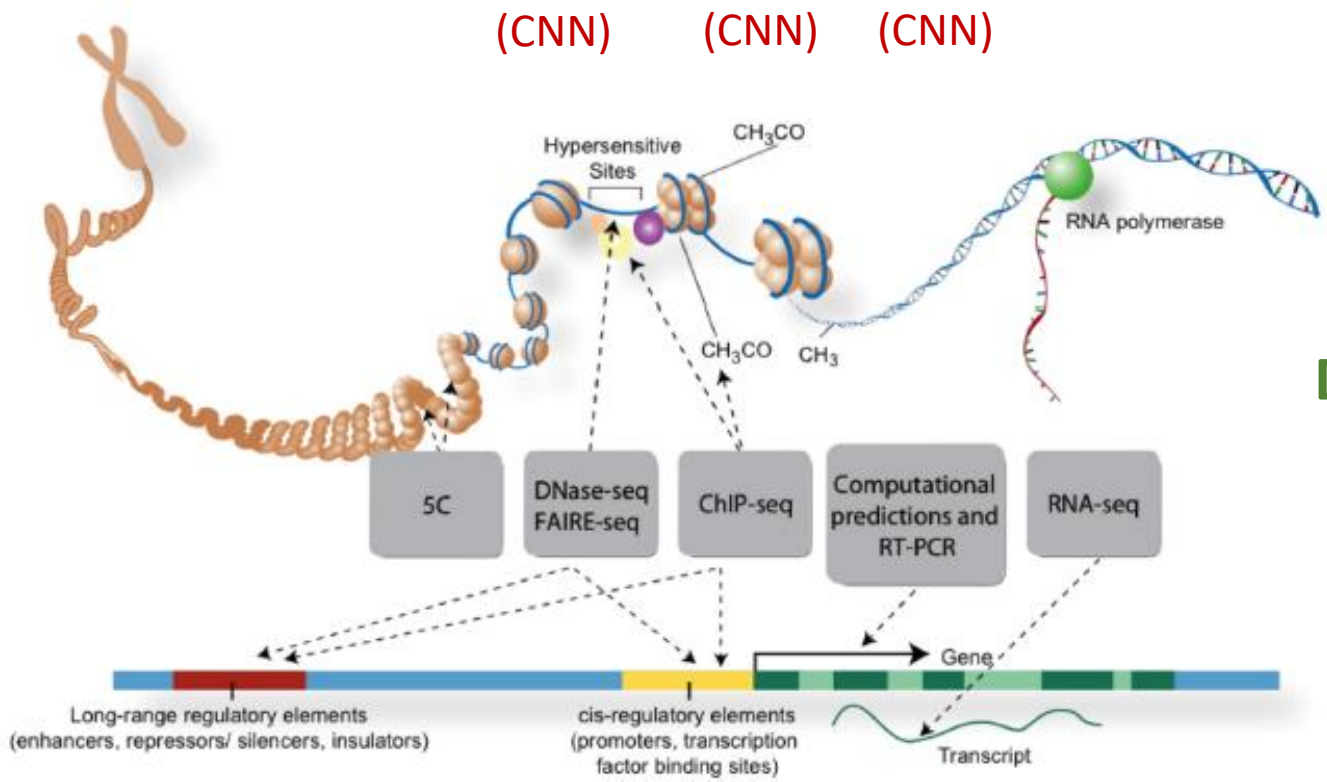
DeepVariant (DCNN)

DeepLNC (DNN)

evoNet (evolved DNN)

DANN (DNN)

DeepSEA (DCNN)




## DNA Methylation

DeepCpG (DNN)

PEDLA (DL-HMM Hybrid)

DEEP (CNN)

FIDDLE (CNN)

## Predicting Enhancers, 3D interactions and *cis*-regulatory regions

*Telenti, Lippert, Chang, DePristo. Hum Mol Genet, 2018*

The Scripps Research Institute

- Genomic data
- Information retrieval
- Data analyses
- Data protection

# Identification of individuals by trait prediction using whole-genome sequencing data

Christoph Lippert[a,1], Riccardo Sabatini[a], M. Cyrus Maher[a], Eun Yong Kang[a], Seunghak Lee[a], Okan Arikan[a], Alena Harley[a], Axel Bernal[a], Peter Garst[a], Victor Lavrenko[a], Ken Yocum[a], Theodore Wong[a], Mingfu Zhu[a], Wen-Yun Yang[a], Chris Chang[a], Tim Lu[b], Charlie W. H. Lee[b], Barry Hicks[a], Smriti Ramakrishnan[a], Haibao Tang[a], Chao Xie[c], Jason Piper[c], Suzanne Brewerton[c], Yaron Turpaz[b,c], Amalio Telenti[b], Rhonda K. Roby[b,d,2], Franz J. Och[a], and J. Craig Venter[b,d,1]

[a]Human Longevity, Inc., Mountain View, CA 94303; [b]Human Longevity, Inc., San Diego, CA 92121; [c]Human Longevity Singapore, Pte. Ltd., Singapore 138542; and [d]J. Craig Venter Institute, La Jolla, CA 92037

## Genetics of the human face: Identification of large-effect single gene variants

Daniel J. M. Crouch, Bruce Winney, Willem P. Koppen, William J. Christmas, Katarzyna Hutnik, Tammy Day, Devendra Meena, Abdelhamid Boumertit, Pirro Hysi, Ayrun Nessa, Tim D. Spector, Josef Kittler, and Walter F. Bodmer

PNAS 2018 January, 115 (4) E676-E685. https://doi.org/10.1073/pnas.1708207114


Contributed by Walter F. Bodmer, October 25, 2017 (sent for review May 18, 2017; reviewed by Marcus W. Feldman and Benjamin M. Neale)



## Modeling 3D Facial Shape from DNA

Peter Claes, Denise K. Liberton, Katleen Daniels, Kerri Matthes Rosana, Ellen E. Quillen, Laurel N. Pearson, Brian McEvoy, Marc Bauchet, Arslan A. Zaidi, Wei Yao, Hua Tang, Gregory S. Barsh, Devin M. Absher, [ ··· ], Mark D. Shriver [ view all ]

Published: March 20, 2014 • https://doi.org/10.1371/journal.pgen.1004224

Scripps Translational Science Institute

Institutions

Data brokers

Cohorts

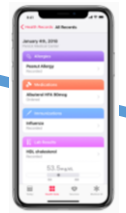
Genotypes

*Zenome, CoverUS, LunaDNA, Doc.ai, Medicalchain, Proof Work, Nebula Genomics*

Phenotypes

Corporate

EHR



Owners

*Telenti, Steinhubl, Topol. Lancet , March 2018*

# Protection

Genetics in Medicine | ORIGINAL RESEARCH ARTICLE

Official journal of the American College of Medical Genetics and Genomics

Open

## Privacy-preserving genomic testing in the clinic: a model using HIV treatment

Paul J. McLaren, PhD[1,2], Jean Louis Raisaro, MSc[3], Manel Aouri, Pharm D, PhD[4], Margalida Rotger, Pharm D, PhD[5], Erman Ayday, PhD[6], István Bartha, PhD[1,2], Maria B. Delgado, PhD[5], Yannick Vallet, MSc[7], Huldrych F. Günthard, MD[8,9], Matthias Cavassini, MD[10], Hansjakob Furrer, MD[11], Thanh Doco-Lecompte, MD[12], Catia Marzolini, Pharm D, PhD[13], Patrick Schmid, MD[14], Caroline Di Benedetto, MD[15], Laurent A. Decosterd, PhD[4], Jacques Fellay, MD, PhD[1,2], Jean-Pierre Hubaux, Dr-Eng[3], Amalio Telenti, MD, PhD[16]; the Swiss HIV Cohort Study

Volume 18 | Number 8 | August 2016 | GENETICS in MEDICINE

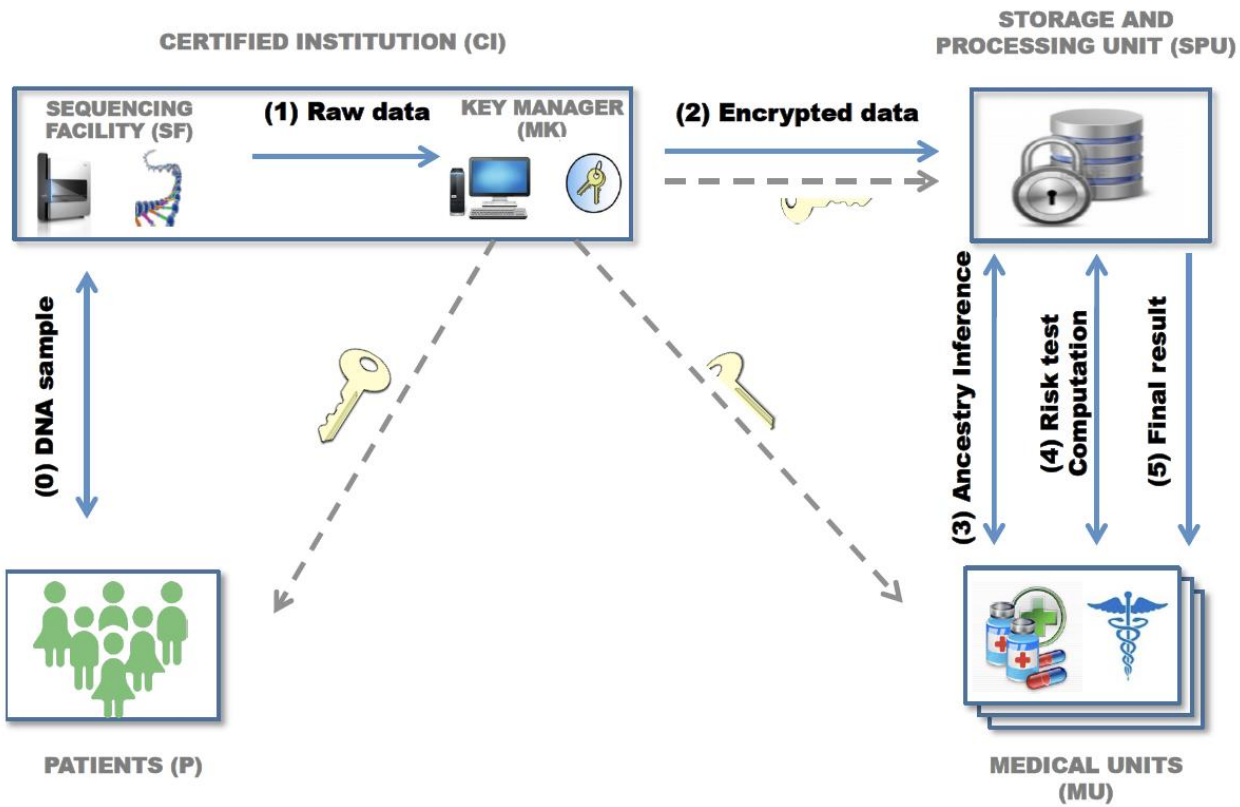
*RESEARCH*

HUMAN GENETICS

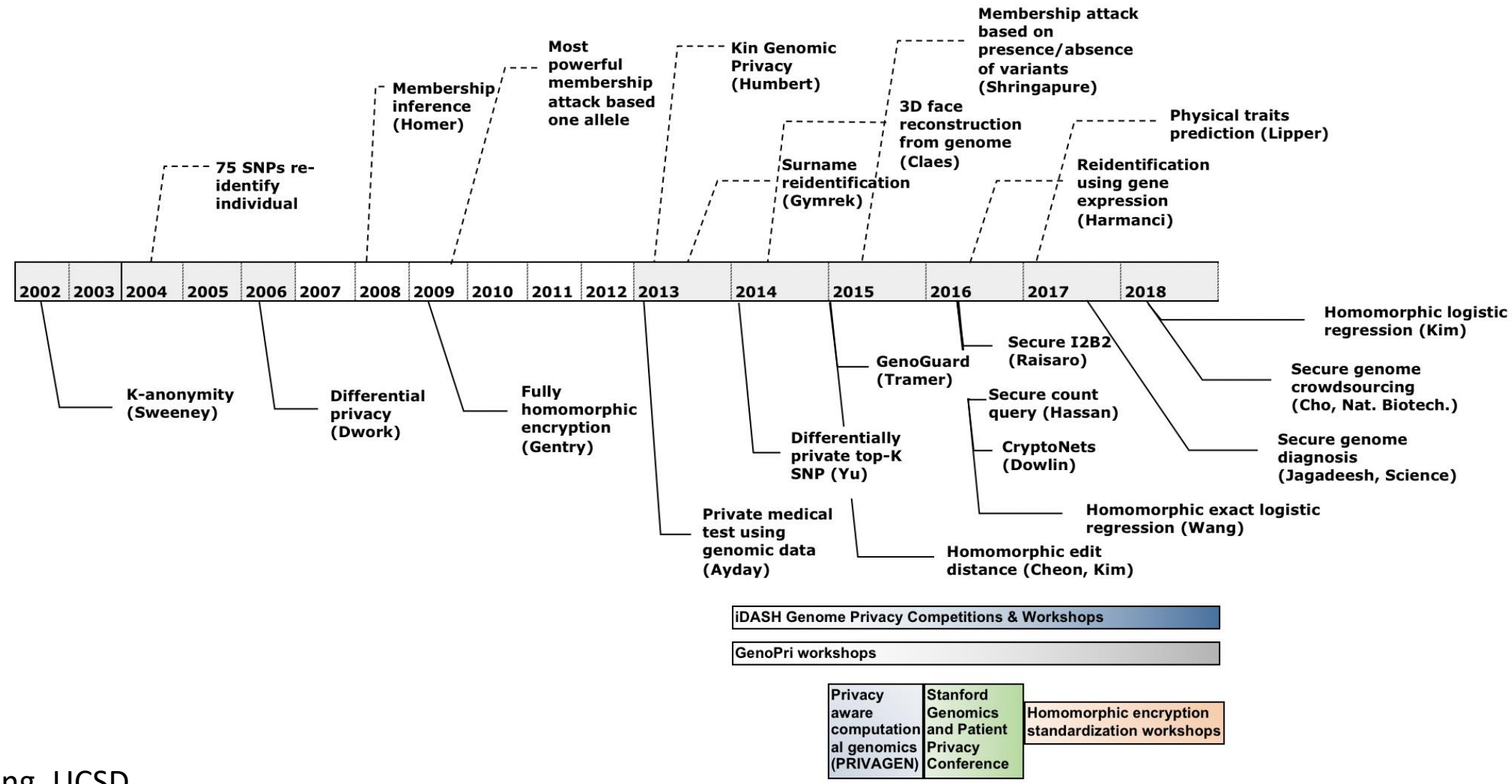
## Deriving genomic diagnoses without revealing patient genomes

Karthik A. Jagadeesh,[1*] David J. Wu,[1*] Johannes A. Birgmeier,[1] Dan Boneh,[1,2†] Gill Bejerano[1,3,4†]

*Science* **357**, 692–695 (2017)



# Recent progress in Genome Privacy




Xioqian Jiang, UCSD

# Summary

- Large scale genomics is happening
- Data management is both a practical and an economic challenge.
- There is an increasing interest in returning data ownership to the individual

# Acknowledgments


- Julia di Iulio (Scripps)
- Alex Wells (Stanford)
- Pejman Mohammadi (Scripps)
- Michael Hicks (Human Longevity)
- Emily Wong (Takeda)
- Xiaoqian Jiang (UCSD)
- Human Longevity Inc.