Global Alliance
for Genomics & Health

# Research platforms that link "omics" to biobank data – opportunities in Ontario

**Speaker:**

Tom Hudson  Ontario Institute for Cancer Research

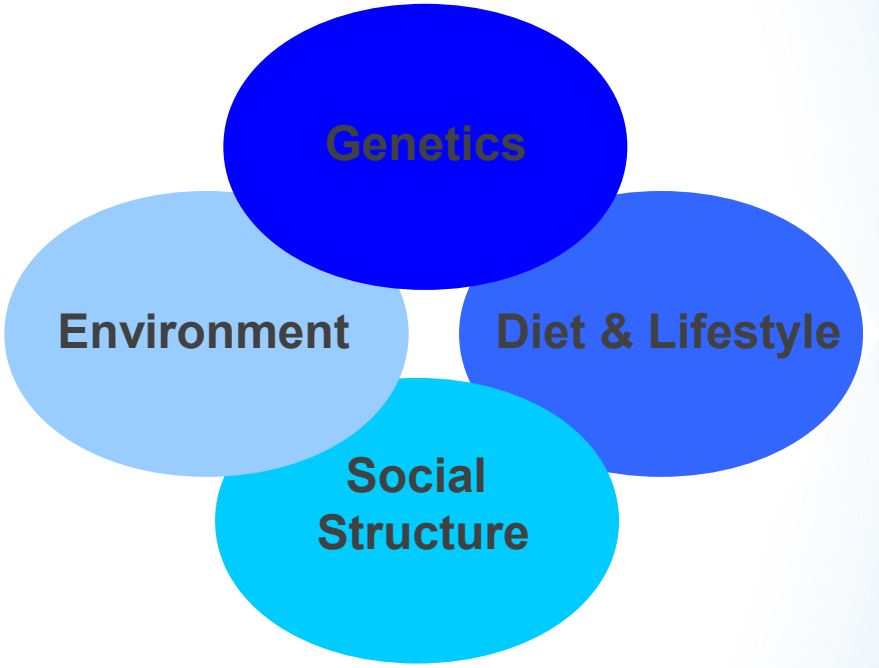
Global Alliance
for Genomics & Health

# How big is “BIG”?

# The Causal Complexity of Chronic Diseases

Global Alliance for Genomics & Health

Diabetes
Asthma
Heart Disease
Schizophrenia
Cancer
Multiple Sclerosis
Obesity
Arthritis




**“webs of causation”**

# For common diseases, effects that we want to measure are small.

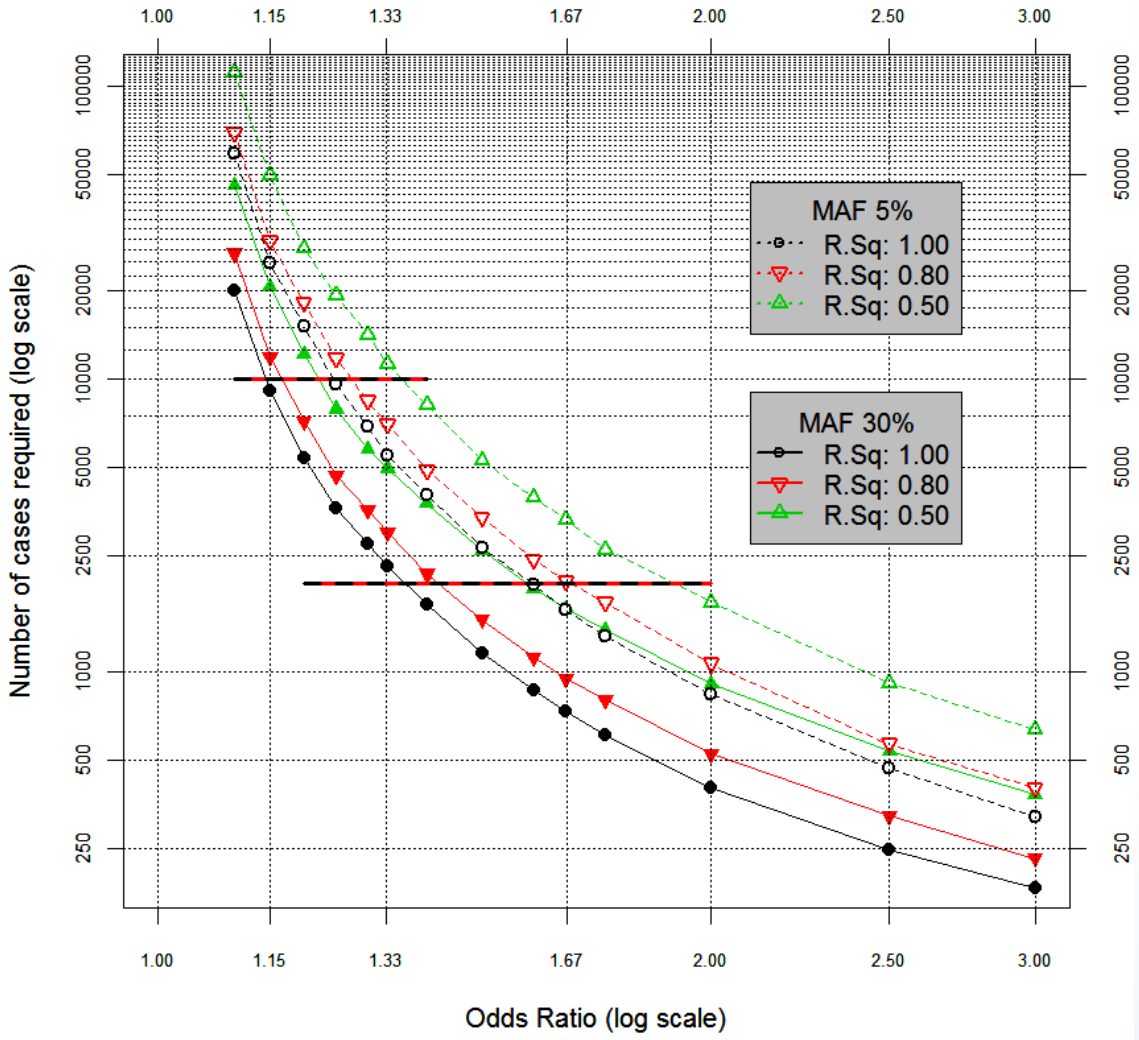
Most GWAS hits have Odds ratios in this range:
1.1 – 1.5

| Disease | Gene | Polymorphism | Approximate frequency of the disease associated allele | Approximate odds ratio for disease associated allele | Ref |
|---|---|---|---|---|---|
| Thrombophilia | *F5* | Leiden Arg506Gln | 0.03 | 4 | 12 |
| Crohn's disease | *CARD15* | 3 SNPs | 0.06(composite) | 4.6 | 67 |
| Alzheimer's disease | *APOE* | ε2/3/4 | 0.15 | 3.3 | 13,68 |
| Osteoporotic fractures | *COL1A1* | Sp1 restriction site | 0.19 | 1.3 | 69,70 |
| Type 2 diabetes | *KCNJ11* | Glu23Lys | 0.36 | 1.23 | 71 |
| Type 1 diabetes | *CTLA4* | Thr17Ala | 0.36 | 1.27 | 72,73 |
| Graves' Disease | *CTLA4* | Thr17Ala | 0.36 | 1.6 | 74 |
| Type 1 diabetes | *INS* | 5' VNTR | 0.67 | 1.2 | 75 |
| Bladder Cancer | *GSTM1* | Null (gene deletion) | 0.70 | 1.28 | 76 |
| Type 2 diabetes | *PPARG* | Pro12Ala | 0.85 | 1.23 | 11 |

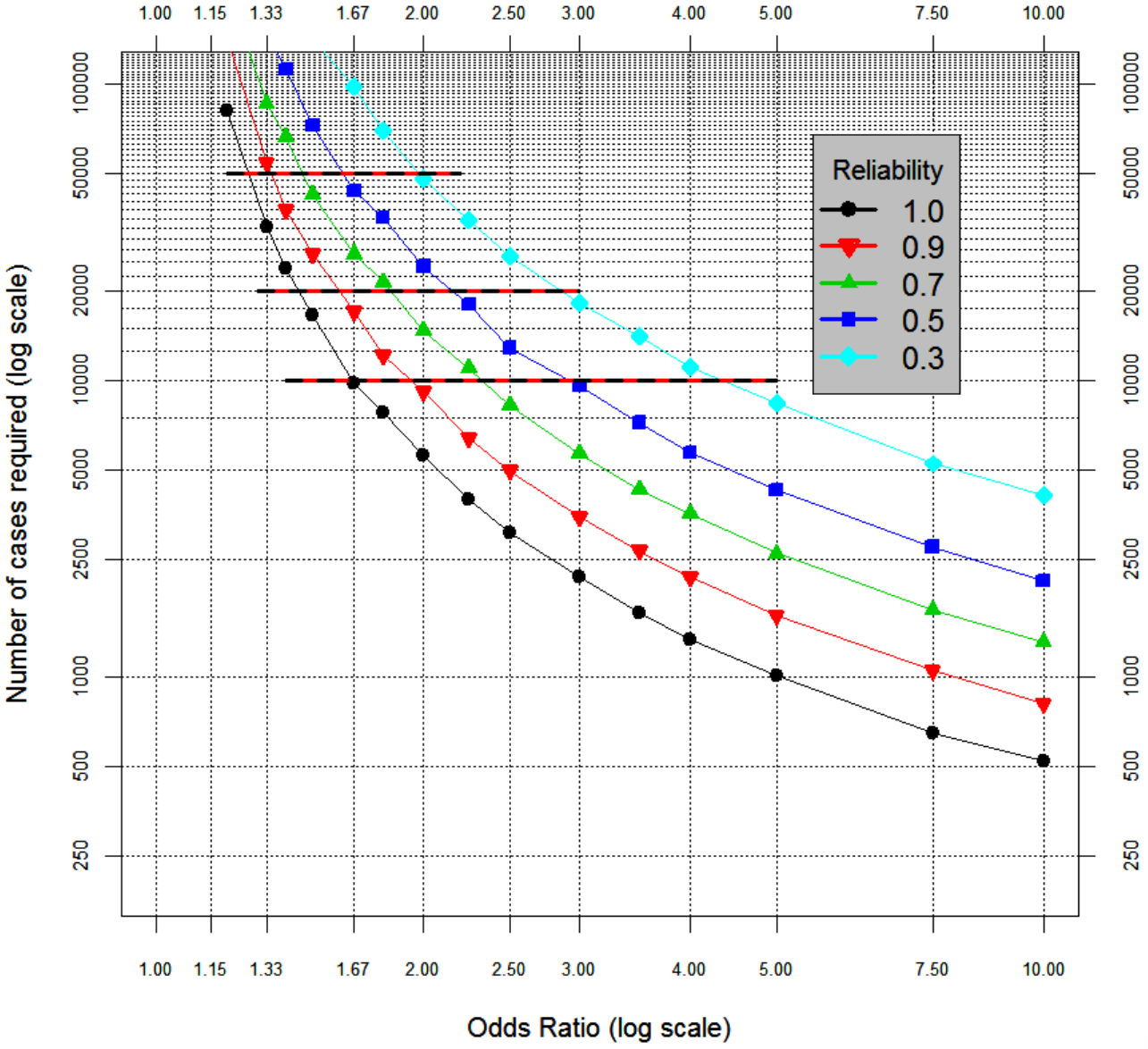
Hattersley AT, McCarthy MI. Lancet 2005;366:1315-1323
Examples of some polymorphisms or haplotypes that have shown consistent association with complex disease

# Genetic main effects

Global Alliance for Genomics & Health



Global Alliance for Genomics & Health



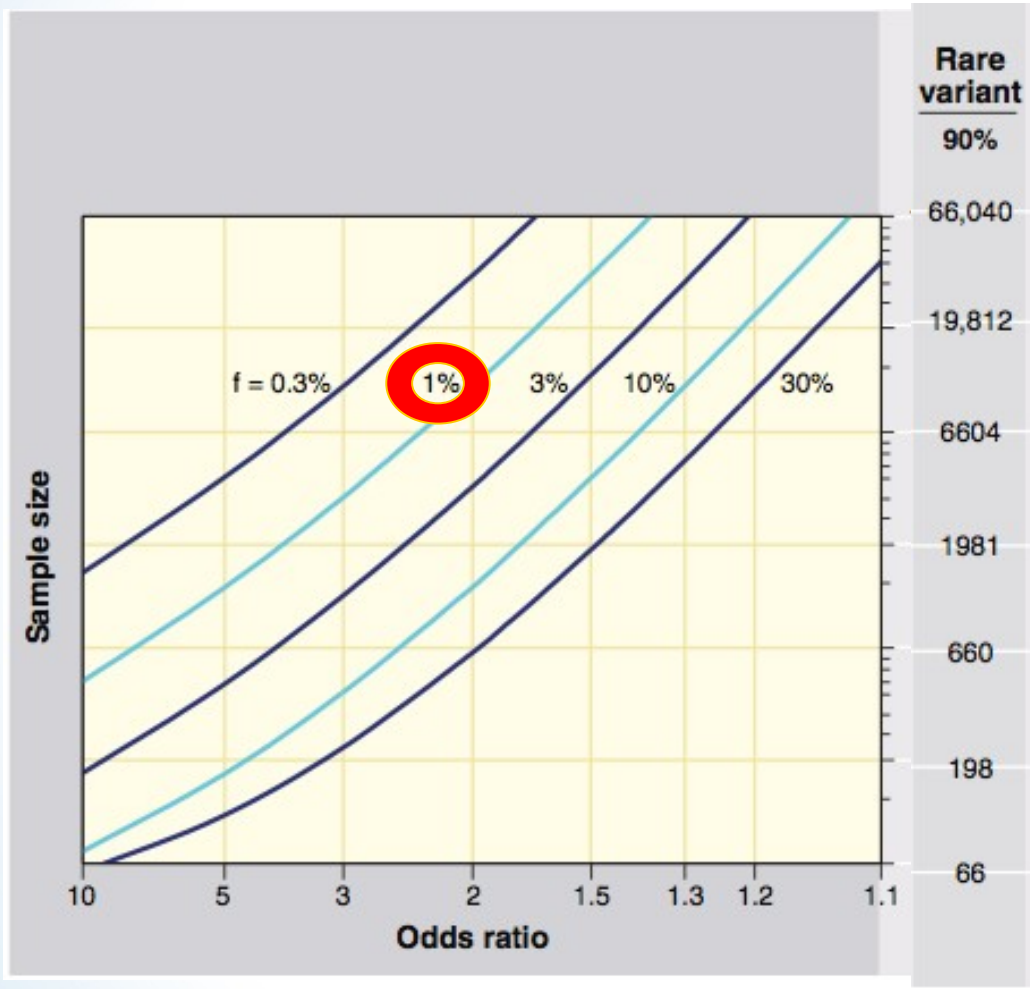

# Gene-lifestyle interactions

MAF for 'at-risk' genotype = 5%
R2=0.8

Prevalence of 'at-risk' life-style factor = 20%

# Detecting rare variants

Global Alliance for Genomics & Health



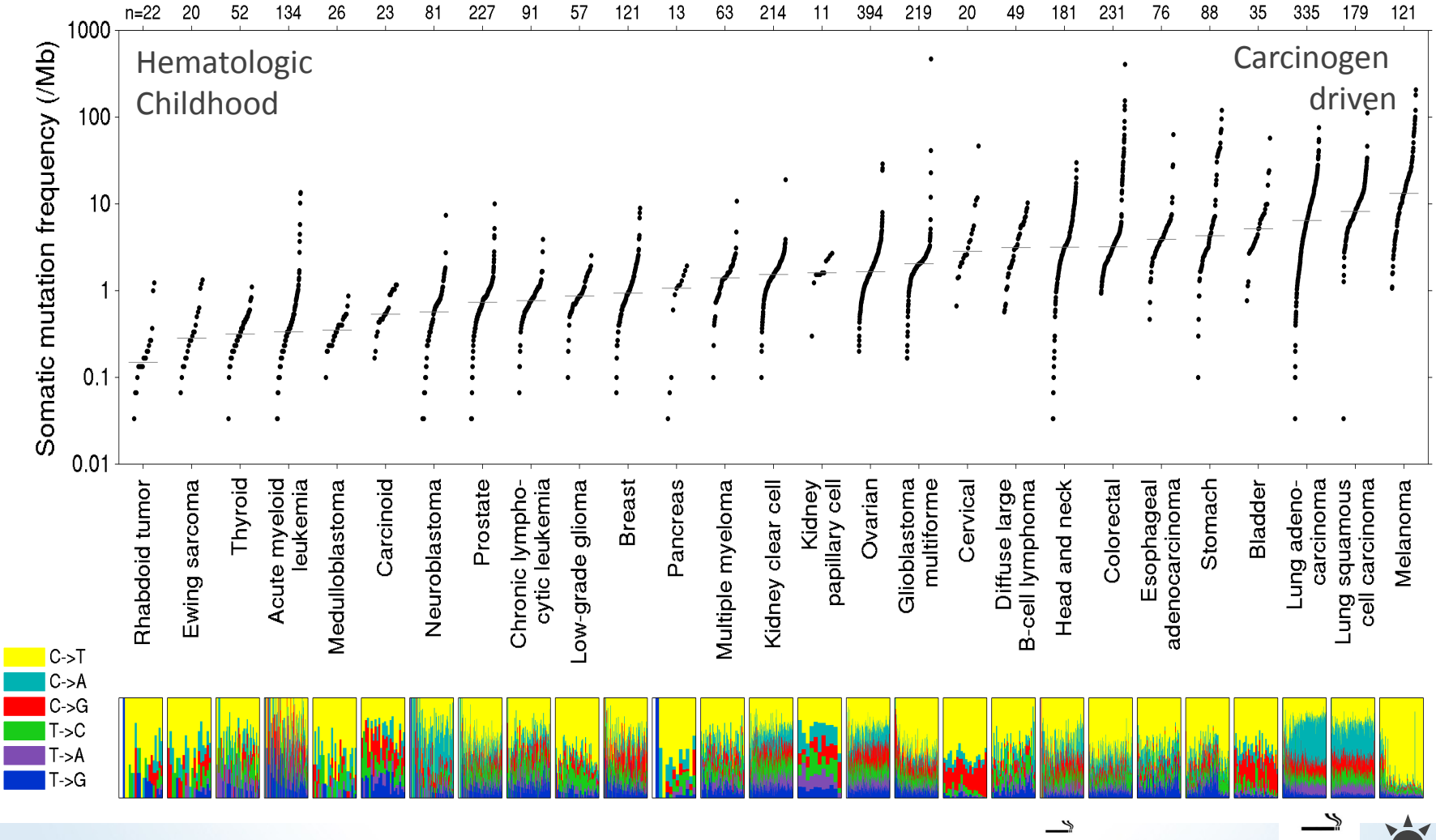

**N ~ 10,000**
**for 90% power to detect variants in 1% of the population with the expected odds ratio of ~2**

**(assuming matched controls, equal sizes, etc)**

Altshuler et al. Science (2008)

# Cancer genomes have a high background mutation rate

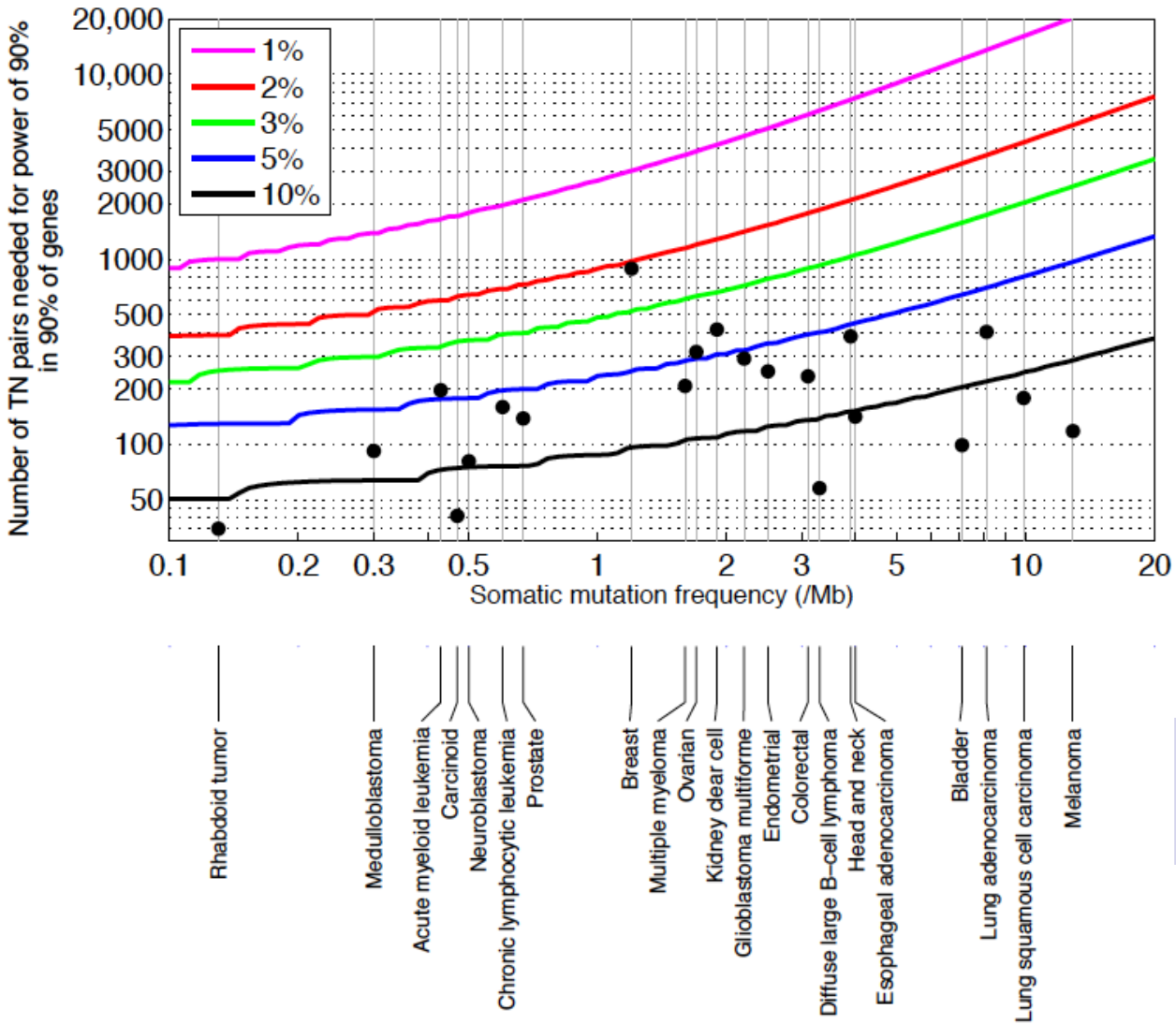
Global Alliance for Genomics & Health




Mike Lawrence, Petar Stojanov, Paz Polak et al. *Nature* (2013)

# Completing the catalogue of cancer genes will require 100,000's of cancer genomes

Global Alliance for Genomics & Health




**For 90% power to detect 90% of genes at frequency ≥ 2%:**

**Need mean of ~2000 samples**

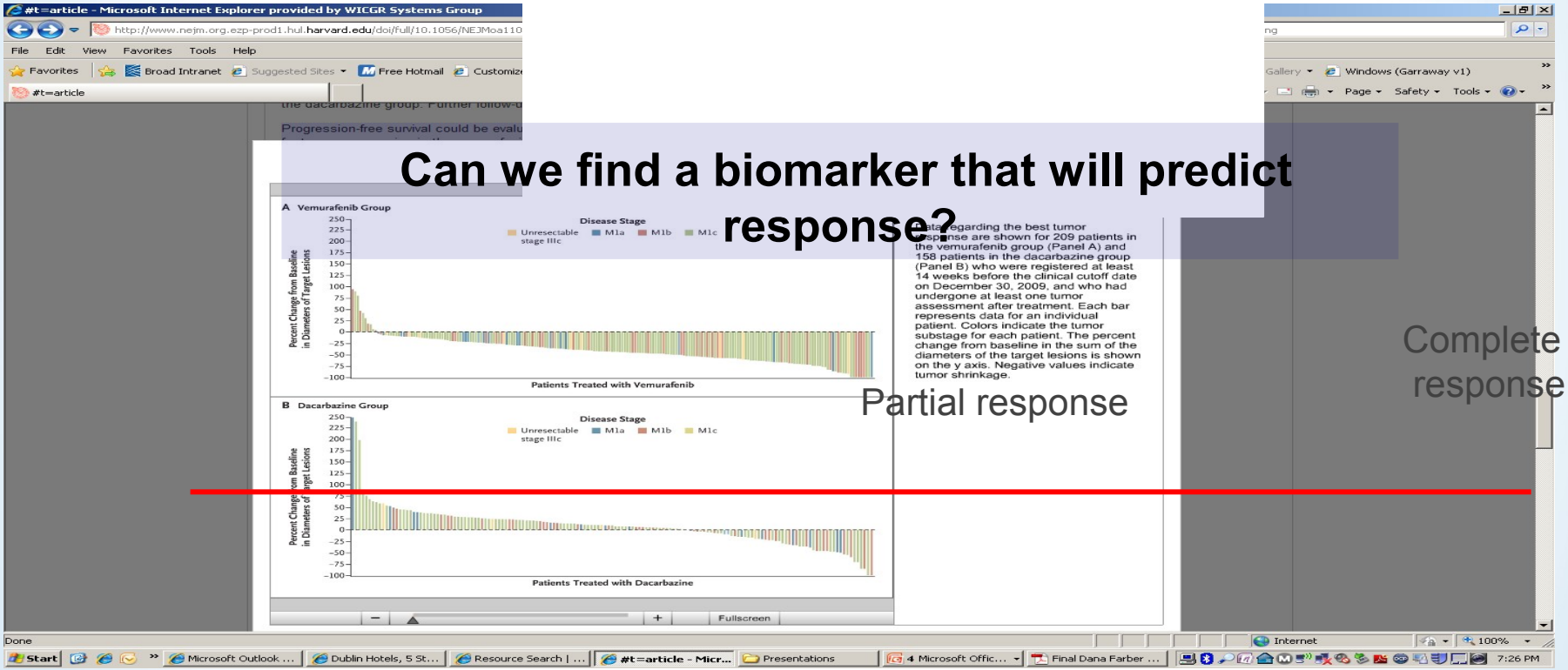
**50 tumor types x 2,000 = 100,000 tumors**

Lawrence et al. *Nature* (2014)

# Detecting Biomarkers that predict drug response

Global Alliance for Genomics & Health

Example: Response to RAF/MEK inhibition in *BRAFV600E* melanoma



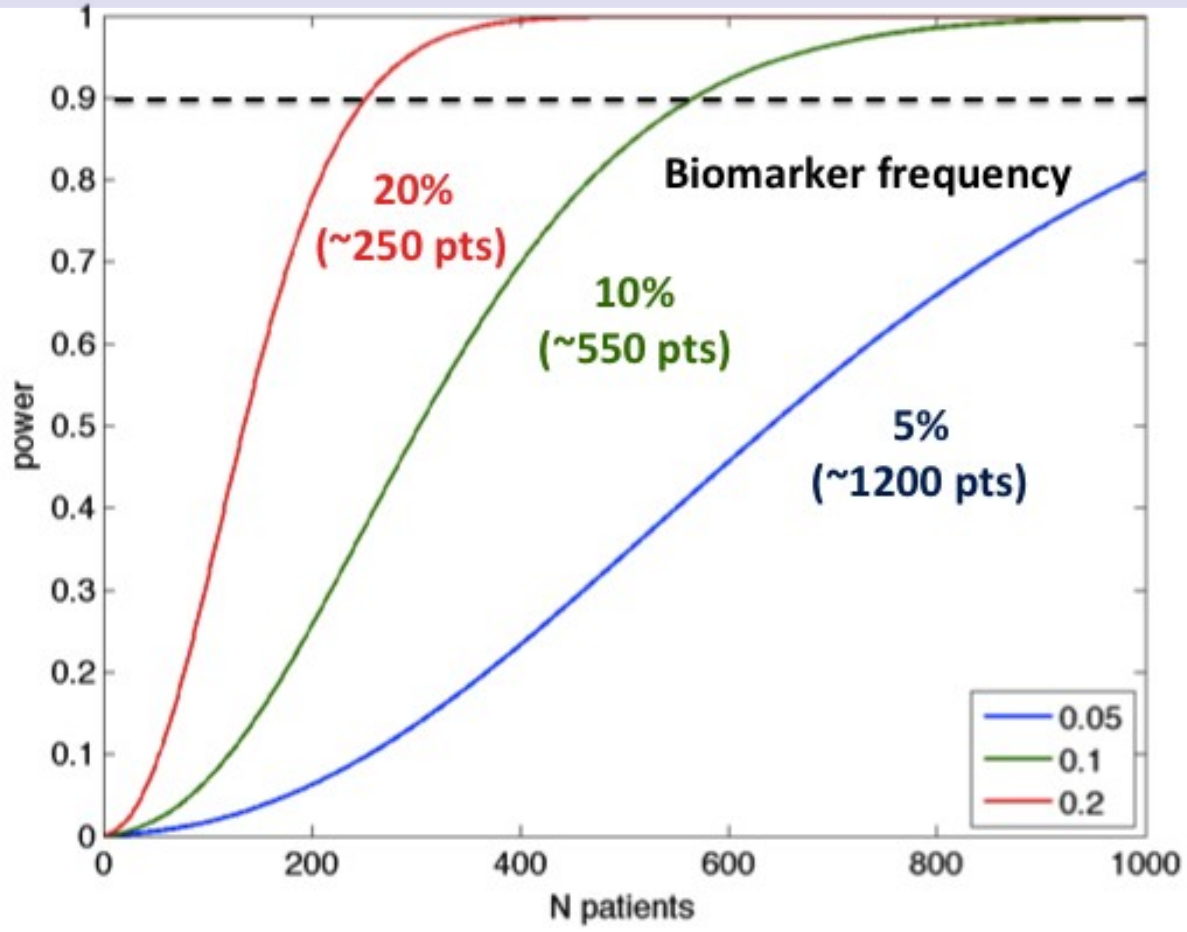

Chapman et al., *NEJM* (2011)

Levi Garraway

# Detecting biomarkers that predict drug response

Global Alliance for Genomics & Health

**Assumptions: 50% of patient respond to a drug.**
**We want to find a biomarker (out of 100 candidates) that predicts 80% response**

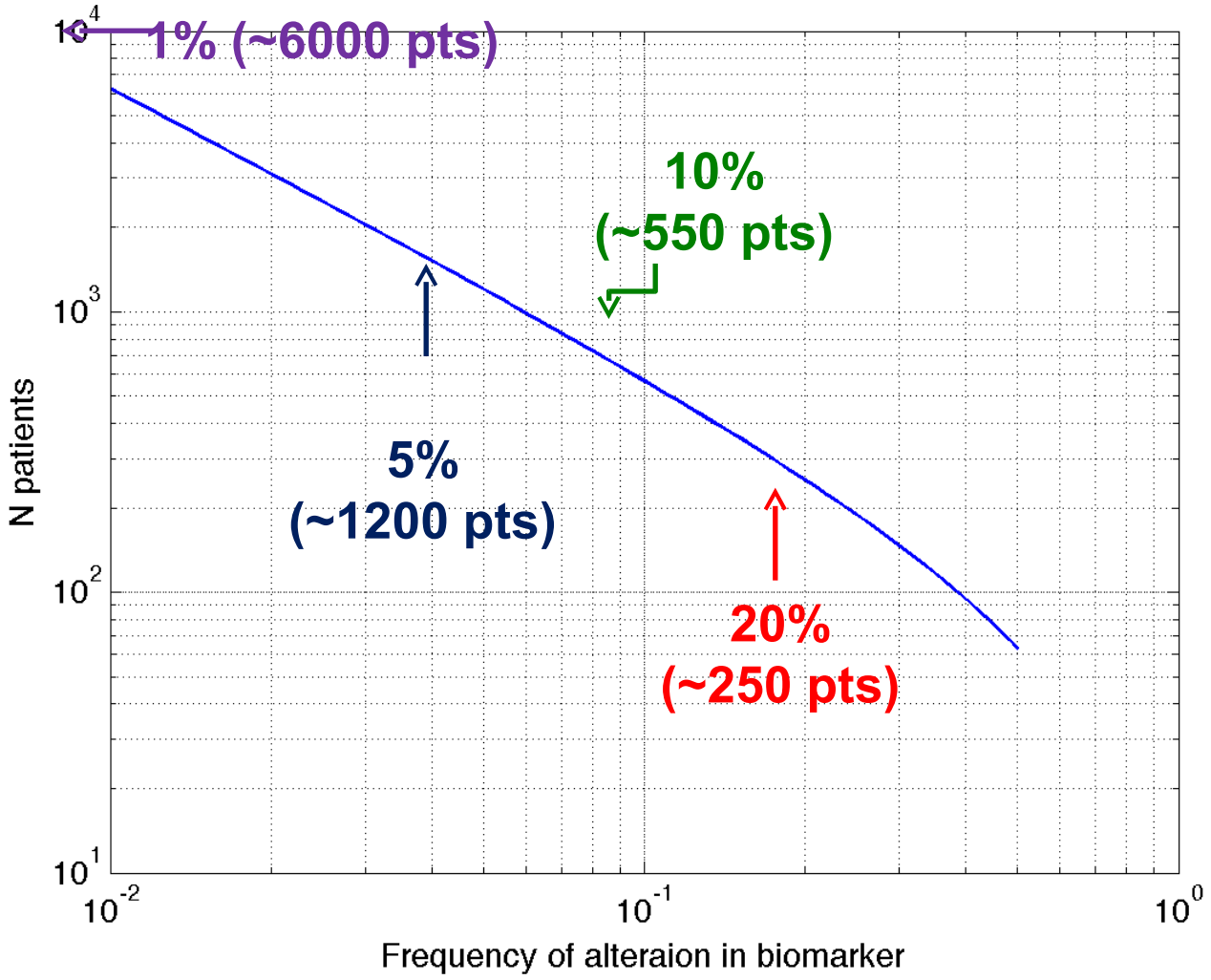



Charles Sawyers

# Detecting biomarkers that predict drug response

Global Alliance for Genomics & Health

## Number of patients required versus biomarker frequency




Charles Sawyers

# Summary for "How big is "BIG"?"

Global Alliance for Genomics & Health

We need to aggregate large datasets with genomic and clinical data to obtain sufficient power to:

1. Find germline risk alleles (10,000s / tumor type)
2. Complete the catalog of cancer genes and pathways (>2% of patients) (1000s / tumor)
3. Detect biomarkers for response (100s to 1000s / tumor type / drug)

**We must share GENOMIC and CLINICAL DATA from hundreds of thousands to MILLIONS of subjects!**

We need to make harmonized data and results easily available to researchers/tool developers, clinicians and patients

Global Alliance for Genomics & Health

# Overview of the Ontario Health Study

# What is the Ontario Health Study?

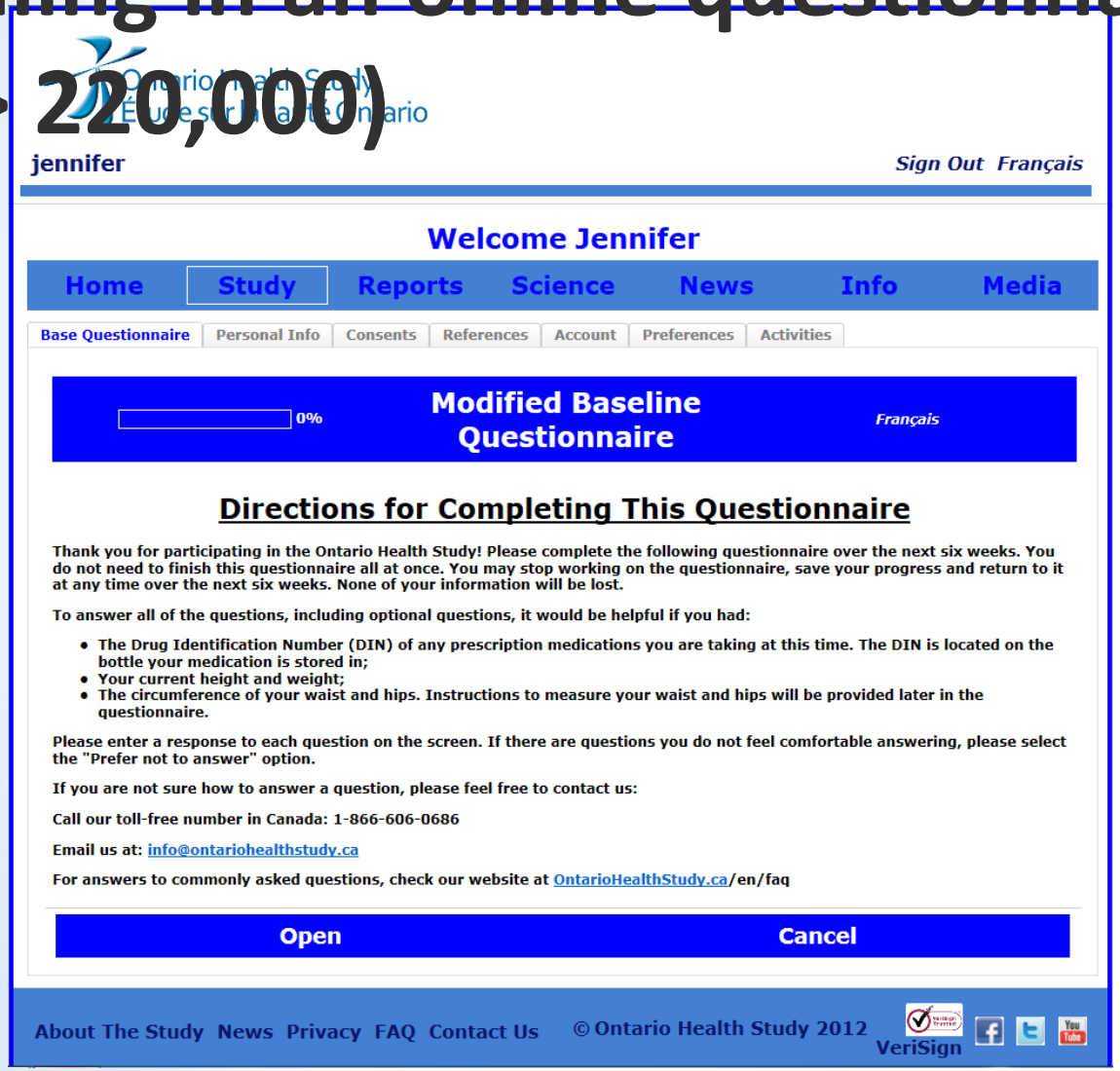
Global Alliance for Genomics & Health

A large innovative prospective cohort in Ontario that will serve as an integrated platform for investigating the complex interplay of environmental, lifestyle and genetic factors that increase individual and community risk of developing cancer, heart disease, diabetes, asthma, depression and other common adult diseases

The Study is one of five regional initiatives being conducted across Canada for the Canadian Partnership for Tomorrow Project

- Ontario Health Study
- Atlantic Partnership for Tomorrow's Health
- Alberta Tomorrow Project
- CartaGene Quebec
- BC Generations project

# Most patients participated by filling in an online questionnaires (> 220,000)




Data linkages at ICES are underway

# Biospecimen Collection (currently 20,000)

Global Alliance for Genomics & Health

## OHS Toronto Assessment Centre

- Operating at 790 Bay Street from July 2012 through March 2014, the OHS Toronto Assessment Centre completed physical assessments and collected blood and urine samples from over 4,800 Study participants.

## Blood Collection Program

- Through a Partnership with LifeLabs Medical Laboratory Services, nearly 7,000 participants have volunteered to provide a small blood sample at one of over 120 LifeLabs locations in Ontario. This program was launched in November 2012.

## Local Study Centres

- Beginning in 2014, the OHS will pilot and launch a series of Local Study

# OHS Demographics

Global Alliance for Genomics & Health

**Gender**: 60.9% female

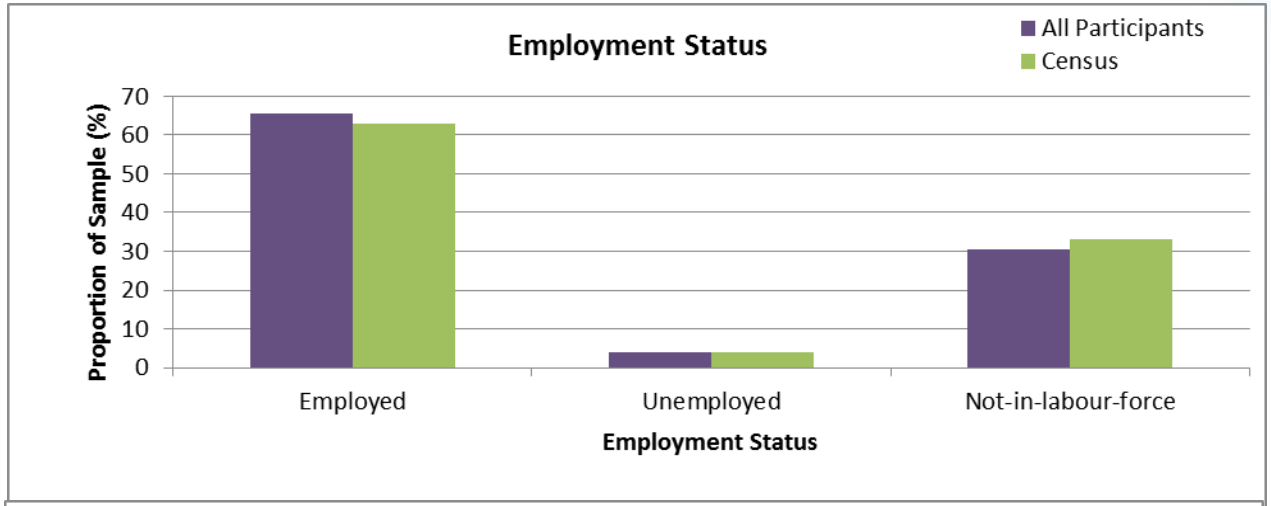
**Age**: mean: 46.5 years
median: 47.2 years

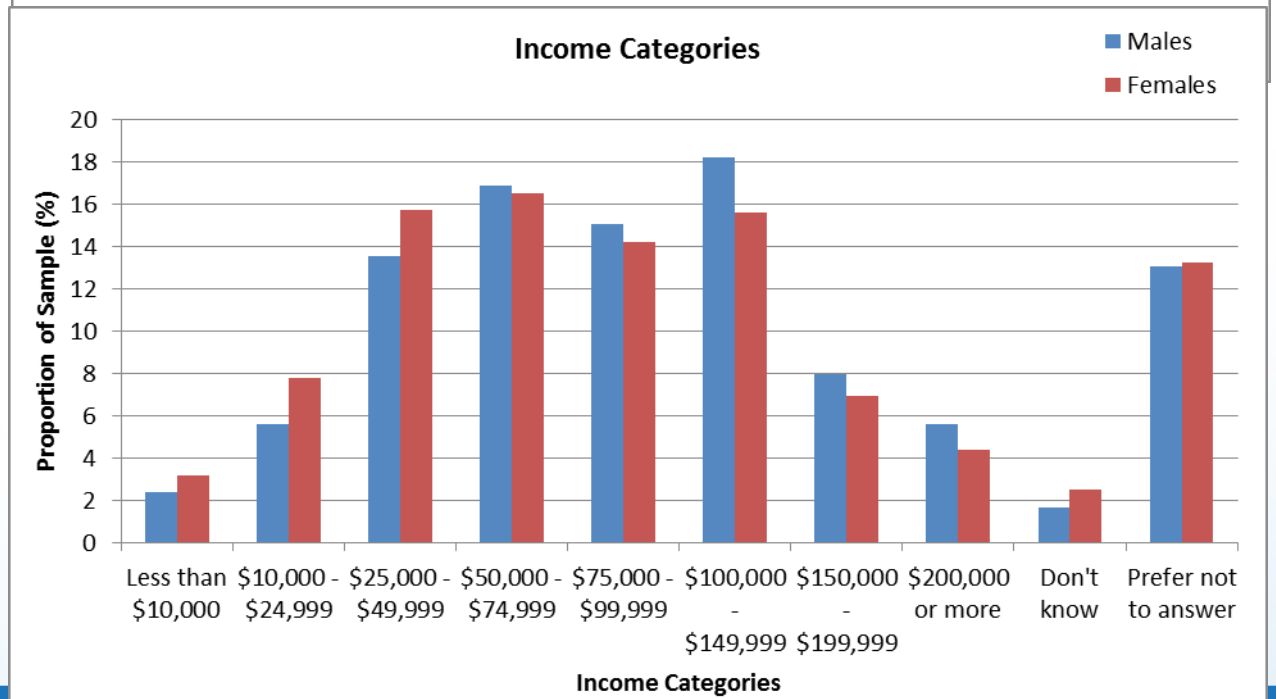
**Ethnicity**

| | OHS Participants | 2006 Census |
|---|---|---|
| Aboriginal | 2.7% | 1.9% |
| Black | 2.0% | 3.7% |
| Chinese | 4.2% | 4.5% |
| South Asian | 3.7% | 6.2% |
| White | 76.5% | 72.1% |

# OHS Demographics

Global Alliance for Genomics & Health



# Prevalence of Major Chronic Diseases in OHS Participants

Global Alliance for Genomics & Health

| | All Participants (n= 188,015) | Females (n= 112,927) | Males (n= 75,088) |
|---|---|---|---|
| Hypertension | 20.1% (37,509) | 16.2% (18,128) | 26.0% (19,381) |
| Heart Disease | 3.3% (5,460) | 1.8% (1,1721) | 5.5% (3,739) |
| Diabetes | 6.1% (11,306) | 4.7% (5,268) | 8.1% (6,038) |
| Arthritis | 19.4% (35,999) | 20.9% (23,271) | 17.1% (12,728) |
| Cancer | 8.4% (15,565) | 8.2% (9,159) | 8.6% (6,406) |
| Major depression | 10.3% (19,137) | 12.9% (14,424) | 6.3% (4,713) |

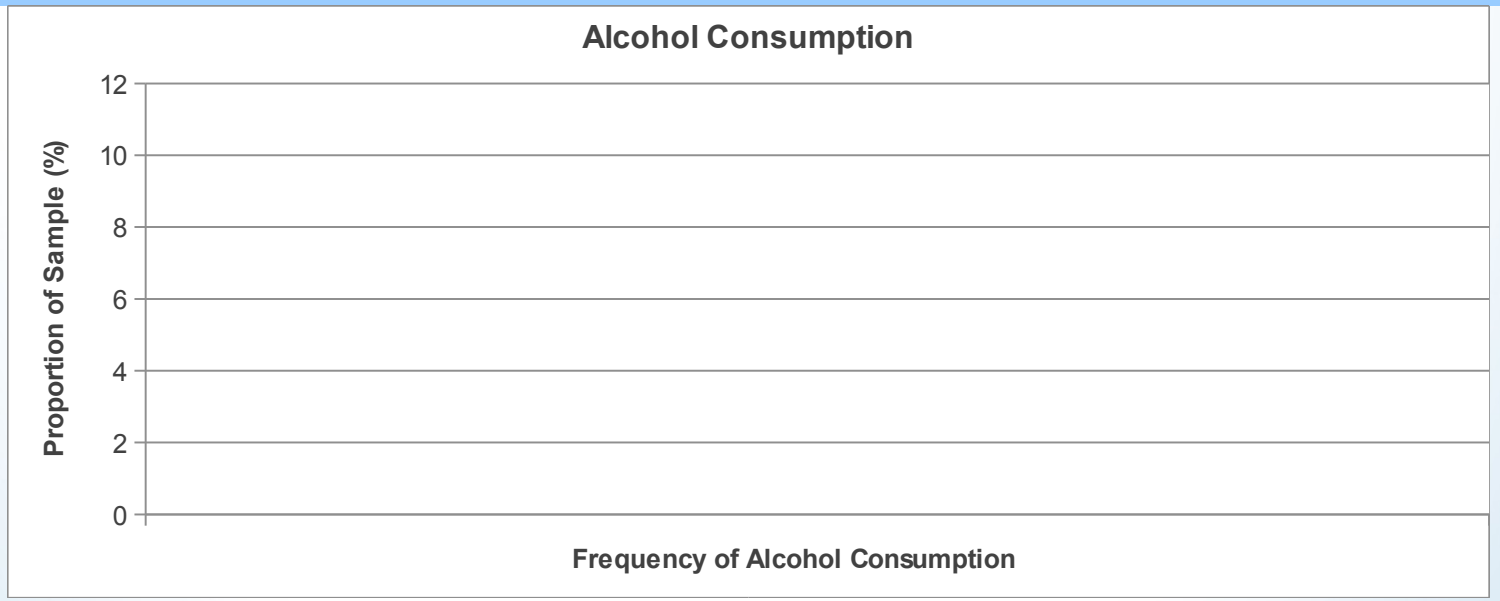
- *Preliminary OHS findings from participants completing the first version of the baseline questionnaire.*

# Smoking and Alcohol Use

Global Alliance for Genomics & Health

Smoking Status

| | All participants (n= 106,427) | Females (n= 62,511) | Males (n= 43,916) |
|---|---|---|---|
| Current Smoker | 23.7% (25,178) | 24.9% (15,554) | 21.9% (9,624) |
| Former Smoker | 48.9% (52,079) | 47.8% (29,849) | 50.6% (22,230) |
| Non-Smoker | 27.4% (29,170) | 27.4% (17,108) | 27.5% (12,062) |




- *Preliminary OHS findings from participants completing the first version of the baseline questionnaire.*

# Physical Activity

Global Alliance for Genomics & Health

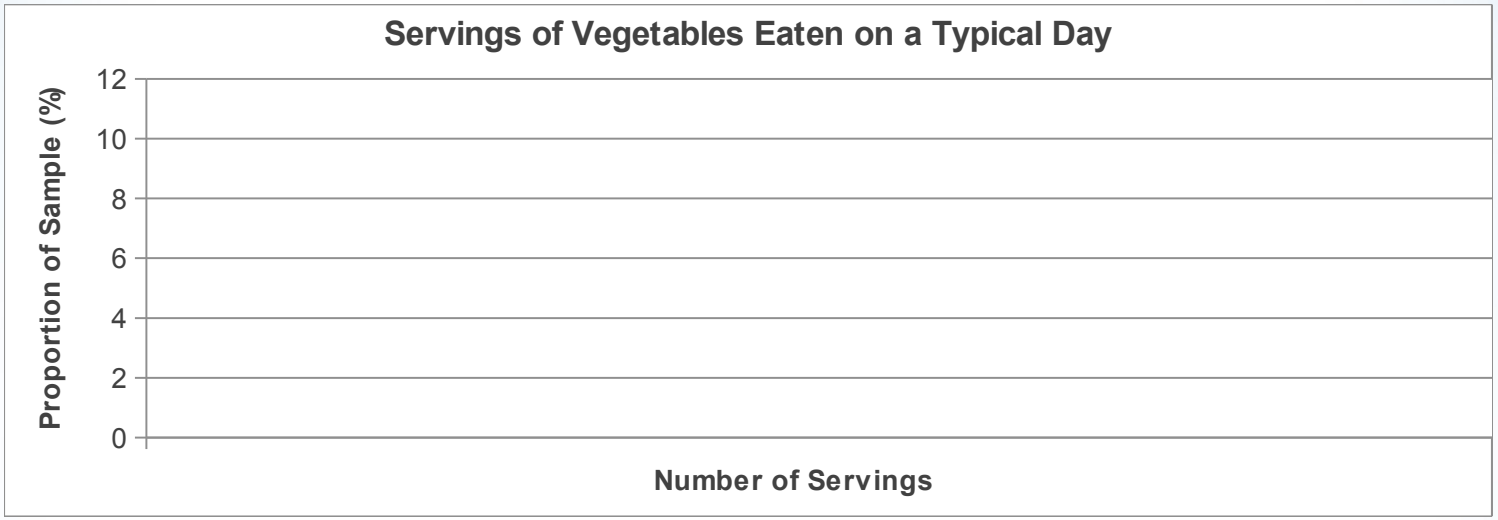
## Physical Activity Level

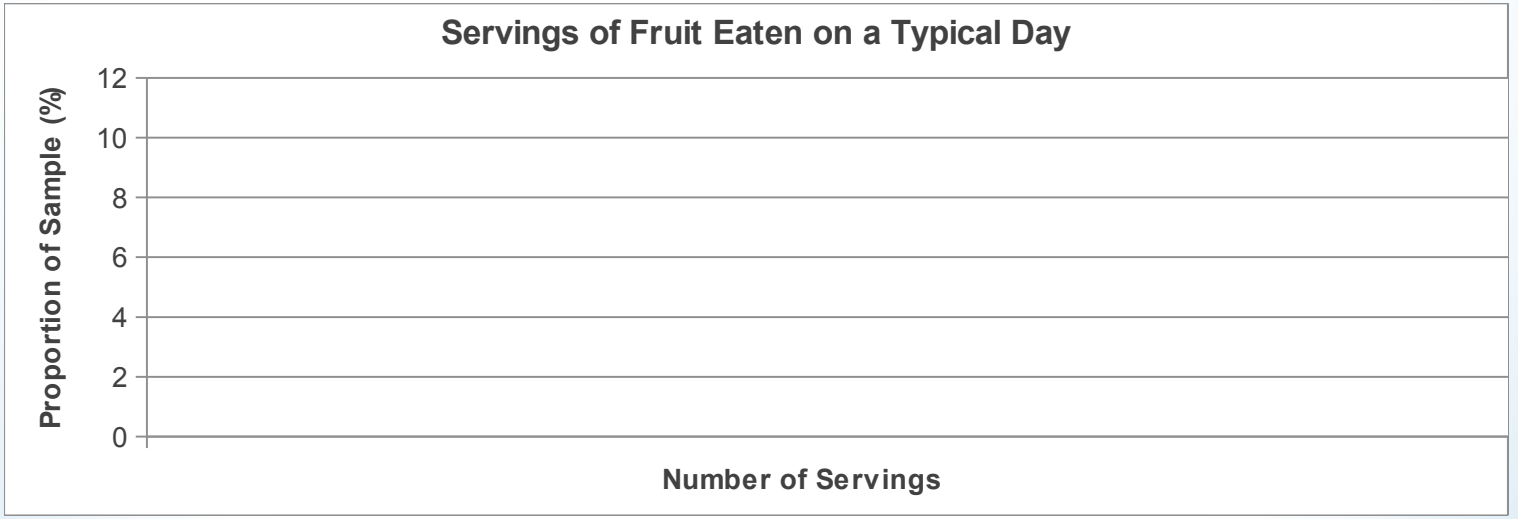
| | All Participants (n= 188,015) | Females (n= 112,927) | Males (n= 75,088) |
|---|---|---|---|
| Low | 4.9% (9,296) | 4.7% (5,289) | 5.3% (4,007) |
| Medium | 29.1% (54,643) | 30.0% (33,818) | 27.7% (20,825) |
| High | 23.5% (44,209) | 22.2% (25,034) | 25.5% (19,175) |
| Unknown | 42.5% (79,867) | 43.2% (48,786) | 41.4% (31,081) |

- *Preliminary OHS findings from participants completing the first version of the baseline questionnaire.*

# Nutrition

Global Alliance for Genomics & Health




- *Preliminary OHS findings from participants completing the first version of the baseline questionnaire.*

Global Alliance for Genomics & Health

# Challenges in handling BIG data

Global Alliance
for Genomics & Health

# GA4GH Data Working Group

Developing solutions for genome and health datasets for millions of research participants

# Data Working Group Members

Global Alliance for Genomics & Health

| Name | Institution |
|---|---|
| Richard Durbin (Co-Chair) | Wellcome Trust Sanger Institute, Cambridgeshire, United Kingdom |
| David Haussler (Co-Chair) | University of California, Santa Cruz, United States |
| Ewan Birney | European Bioinformatics Institute, Cambridgeshire, United Kingdom |
| Gaddy Getz | Broad Institute and Massachusetts General Hospital, Boston, United States |
| Heng Li | Broad Institute, Boston, United States |
| Gil McVean | University of Oxford, Oxford, United Kingdom |
| Nicola Mulder | University of Cape Town, Cape Town, South Africa |
| David Patterson | University of California, Berkeley, Berkeley, United States |
| Anthony Philippakis | Genome Bridge LLC, Cambridge, United States |
| Lincoln Stein | Ontario Institute for Cancer Research, Toronto, Canada |
| Michael Baudis | Swiss Institute of Bioinformatics, Zurich, Switzerland |

# Big data problems to overcome

Global Alliance for Genomics & Health

- Existing open source bioinformatics software is unprofessional, large medical centers are making the problem worse
- Major medical centers are separately hiring software engineers or using postdocs to build custom genomics pipelines
- Creates Balkanized, incompatible, inadequate systems
- Reinforces barriers to data sharing

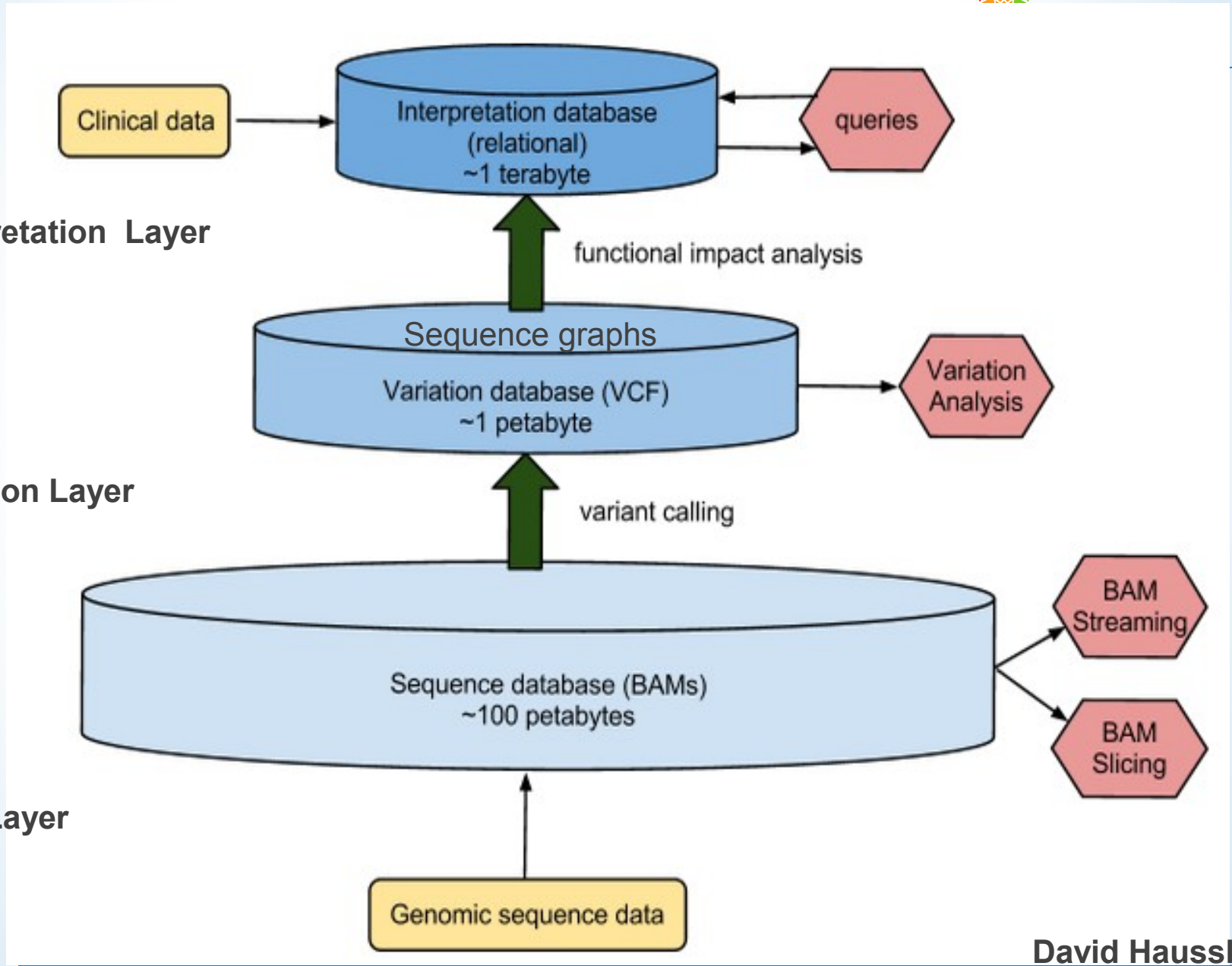
# Different Requirements for 1M Genomes

Global Alliance for Genomics & Health

- Different types of data interactions:
  - Support both research and clinical practice
  - Compute within a provided cloud
  - Separately URIed, metadata-tagged parts of a single patient file supporting 3rd party mashups and tools
- New consents models, sample donor trusts the security provided
- APIs, not file formats.
- Benchmarking so all can use system to improve methods, e.g. SMaSH, somatic variant calling DREAM competition

Dave Patterson, www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html

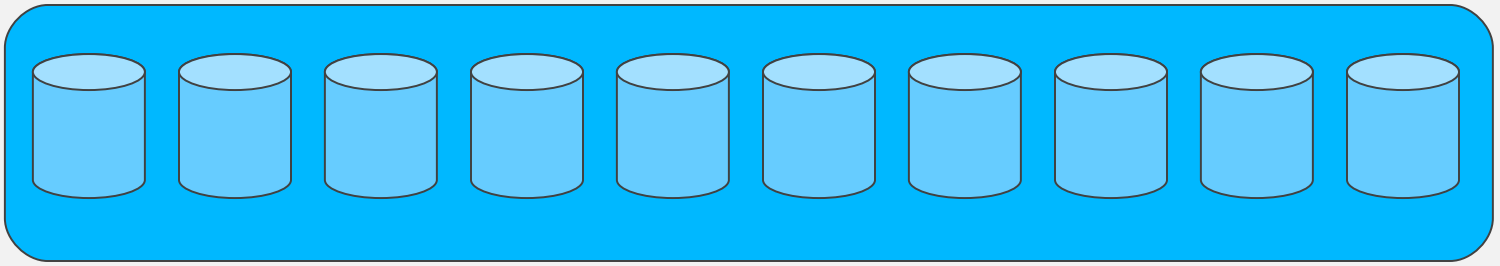
# Possible Genome Commons Architecture



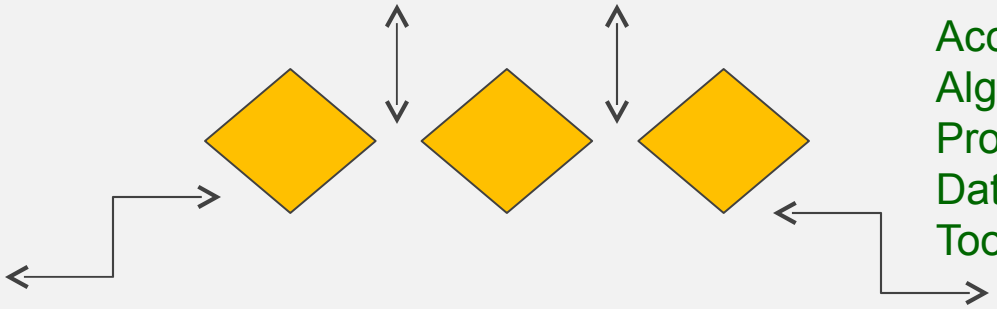

David Haussler

# Ontario Commons Database

Global Alliance for Genomics & Health



Ontario Data Sets



Access control
Algorithm development
Programmer APIs
Data browsers
Toolkits

Virtual Machines

# Conclusion

Global Alliance for Genomics & Health

- Very large datasets are needed to answer clinically relevant questions related to common diseases and cancer
- New technologies are needed to store, share and analyze large amounts of data to enable learning rules and patterns
- International standards need to be developed for data to be shared responsibly with researchers, clinicians, and public health organizations to accelerate progress and provide benefits to patients. [Discussed by Peter Goodhand]
- Ontario has significant clinical and population research