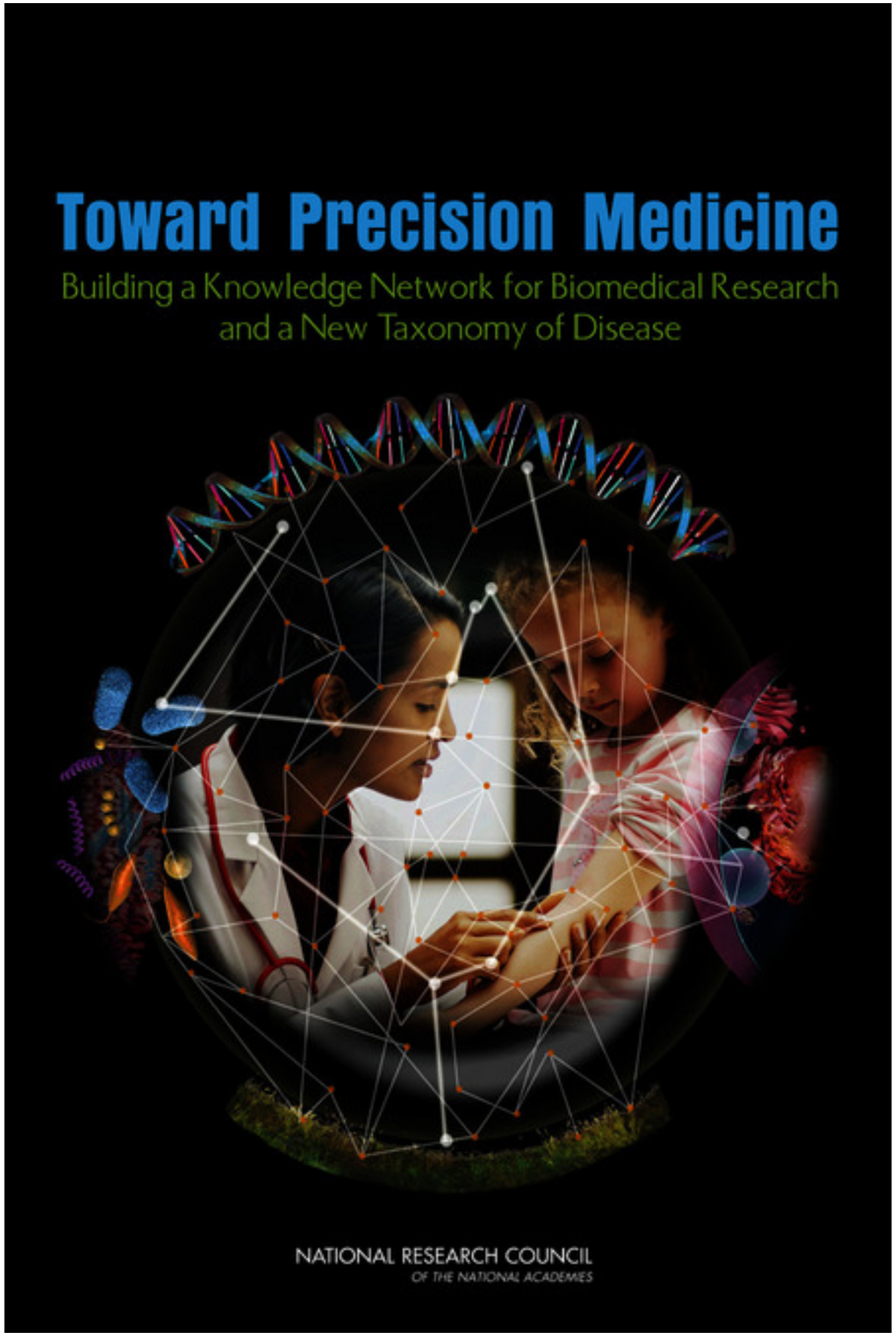
Precision Medicine

Massachusetts Institute of Technology

How precisely can we understand the individual patient?

- Disease subtyping: clustering patients by
  - Demographics
  - Co-morbidities
  - Vital Signs
  - Medications
  - Procedures
  - Disease “trajectories”
  - Image similarities
  - Genetics:
    - SNPs, Exome sequence, Whole genome sequence, RNA-seq, proteomics



Committee on a Framework for Developing a New Taxonomy of Disease. (2017). Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease (pp. 1–143). Washington, D.C.: National Academies Press. http://doi.org/10.17226/13284

Drivers of Change

- New capabilities to compile molecular data on patients on a scale that was unimaginable 20 years ago.

- Increasing success in utilizing molecular information to improve the diagnosis and treatment of disease.

- Advances in information technology, such as the advent of electronic health records, that make it possible to acquire detailed clinical information about large numbers of individual patients and to search for unexpected correlations within enormous datasets.

- A “perfect storm” among stakeholders that has increased receptivity to fundamental changes throughout the biomedical research and healthcare-delivery systems.

- Shifting public attitudes toward molecular data and privacy of healthcare information.

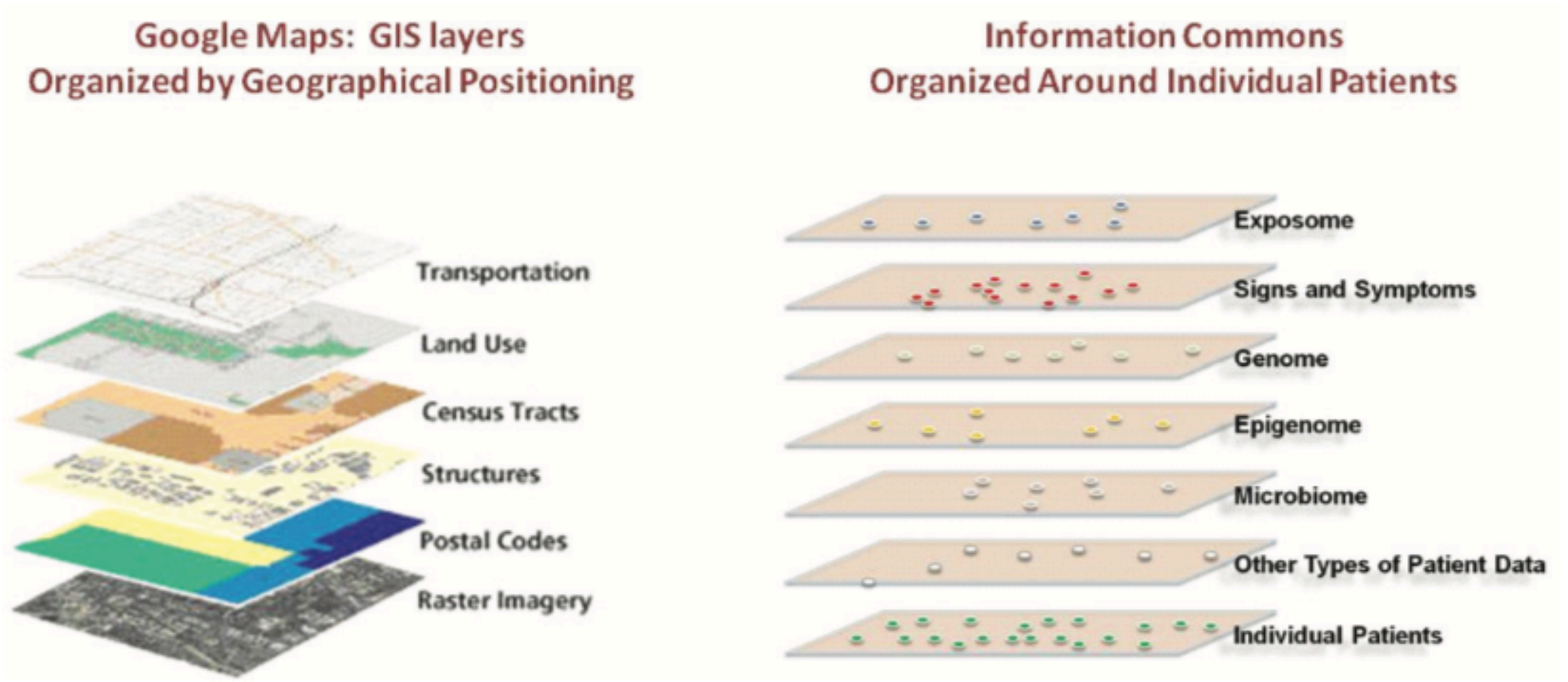


FIGURE 1-2 An Information Commons might use a GIS-type structure.
The proposed, individual-centric Information Commons (right panel) is somewhat analogous to a layered GIS (left panel). In both cases, the bottom layer defines the organization of all the overlays. However, in a GIS, any vertical line through the layers connects related snippets of information since all the layers are organized by geographical position. In contrast, data in each of the higher layers of the Information Commons will overlay on the patient layer in complex ways (e.g., patients with similar microbiomes and symptoms may have very different genome sequences).

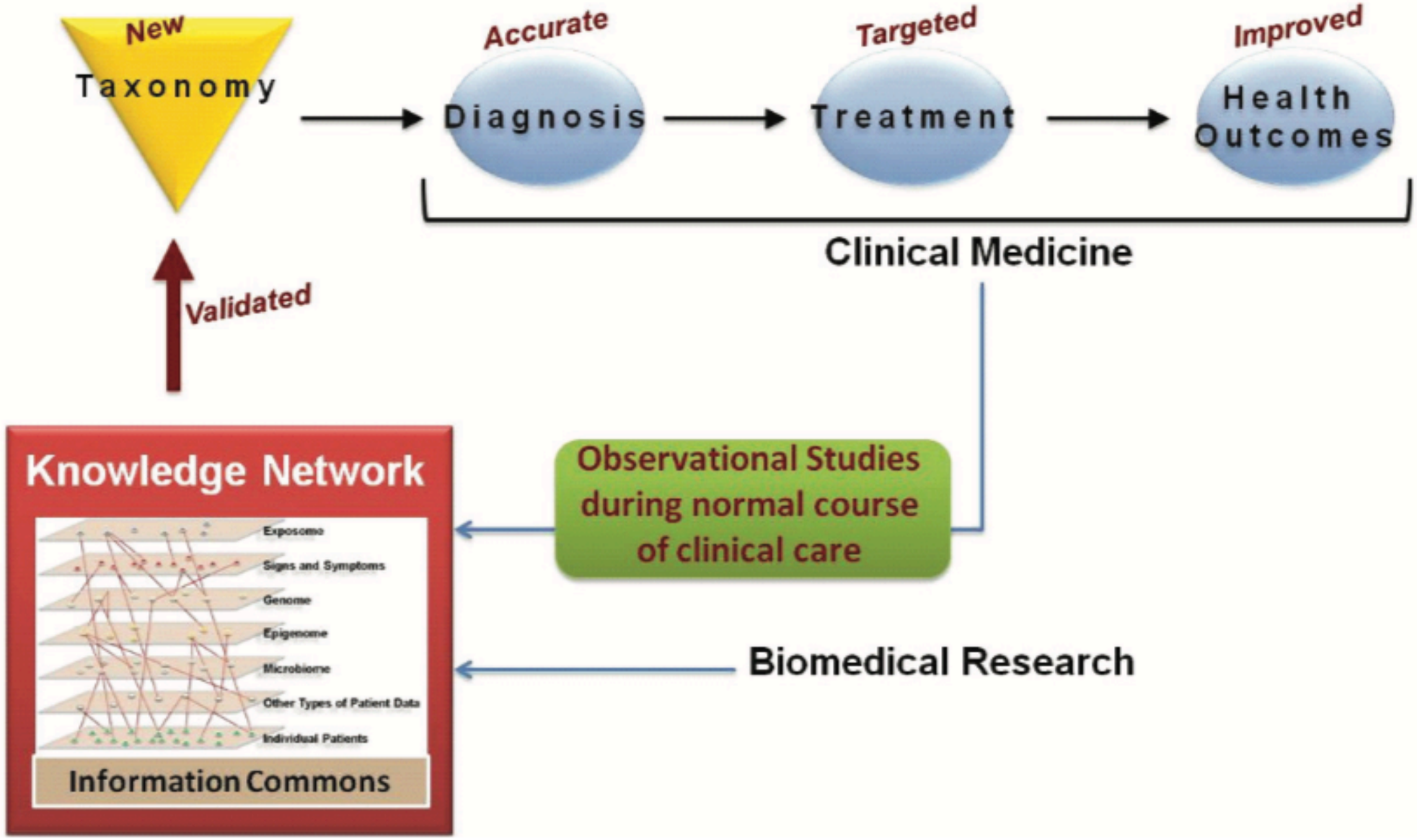
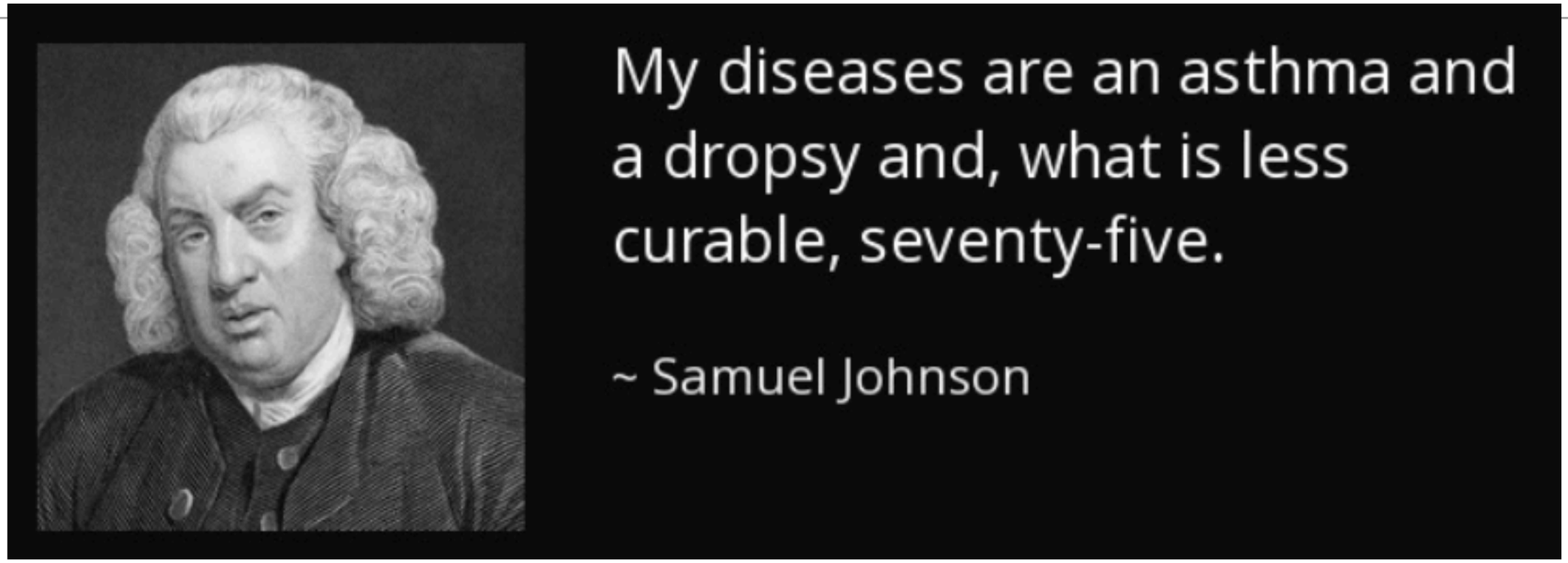


FIGURE 1-3 A knowledge network of disease would enable a new taxonomy. An individual-centric Information Commons, in combination with all extant biological knowledge, will inform a Knowledge Network of Disease, which will capture the exceedingly complex causal influences and pathogenic mechanisms that determine an individual's health. The Knowledge Network of Disease would allow researchers to hypothesize new intralayer cluster and interlayer connections. Validated findings that emerge from the Knowledge Network, such as those which define new diseases or subtypes of diseases that are clinically relevant (e.g., which have implications for patient prognosis or therapy) would be incorporated into the New Taxonomy to improve diagnosis and treatment.

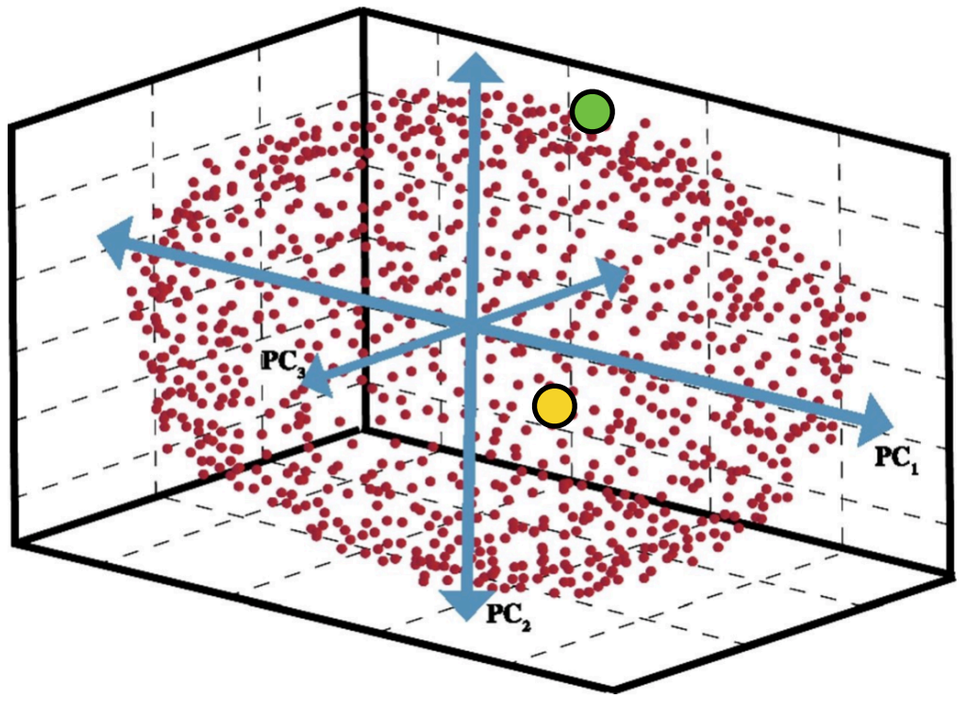
Centrality of Taxonomy (as a *hypothesis*)



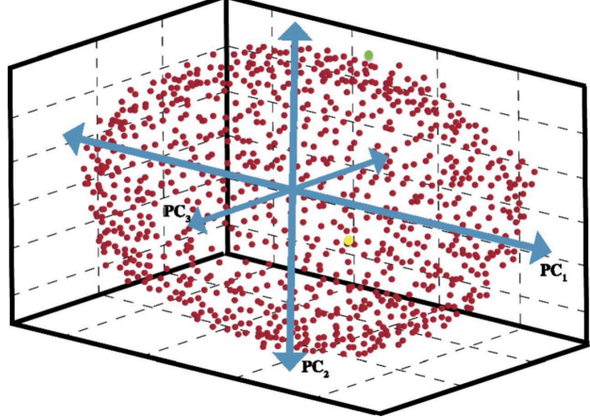
- What is “dropsy”?
  - “water sickness”, “swelling”, “edema”
  - *disease that got Grandma to take to her bed permanently in Victorian dramas*
  - causes: COPD, CHF, CKD, …
  - Last recorded on a death certificate ~1949
- Is “asthma” equally non-specific?

Precision Medicine Modality Space (PMMS)
(Isaac Kohane)

- Very high dimensionality
  - All of the characteristics of the NRC information commons
    - Many of these individually have high dimension
  - Time
  - Claims
  - Response to therapy
- Lumpy, corresponding to different highly specific diseases



The Vision
(Isaac Kohane)



A 13 year old boy presented with a recurrence of abdominal pain, hourly diarrhea and blood per rectum.

10 years earlier, he had been diagnosed with ulcerative colitis. At 3 years of age he was treated with a mild anti-inflammatory drug and had been doing very well until this most recent presentation.

On this occasion, despite the use of the full armamentarium of therapies: antimetabolites, antibiotics, glucocorticoids, immunosuppressants, first and second generation monoclonal antibody-based therapies, he continued to have pain and bloody diarrhea and was scheduled to have his colon removed. This is often but not always curative but has its own risks and consequences. After the fact, he and his parents had their exomes sequenced, which revealed rare mutations affecting specific cytokines (inflammation mediators/signalling mechanisms).

If we had plotted his position in PMMS by his proximity in clinical presentation at age 3, he would have been well within the cloud of points (each patient is a point in the above diagram) like the yellow point. If we had included the mutational profile of his cytokines he would have been identified as an outlier, like the green point. Also, if we had included his later course, where he was refractory to all therapies, he would have also been an outlier. But only if we had included the ***short*** duration (< 6 months) over which he was refractory because for a large minority of ulcerative colitis patients they become refractory to multiple medical treatments but of many years.

How to Classify this Patient?

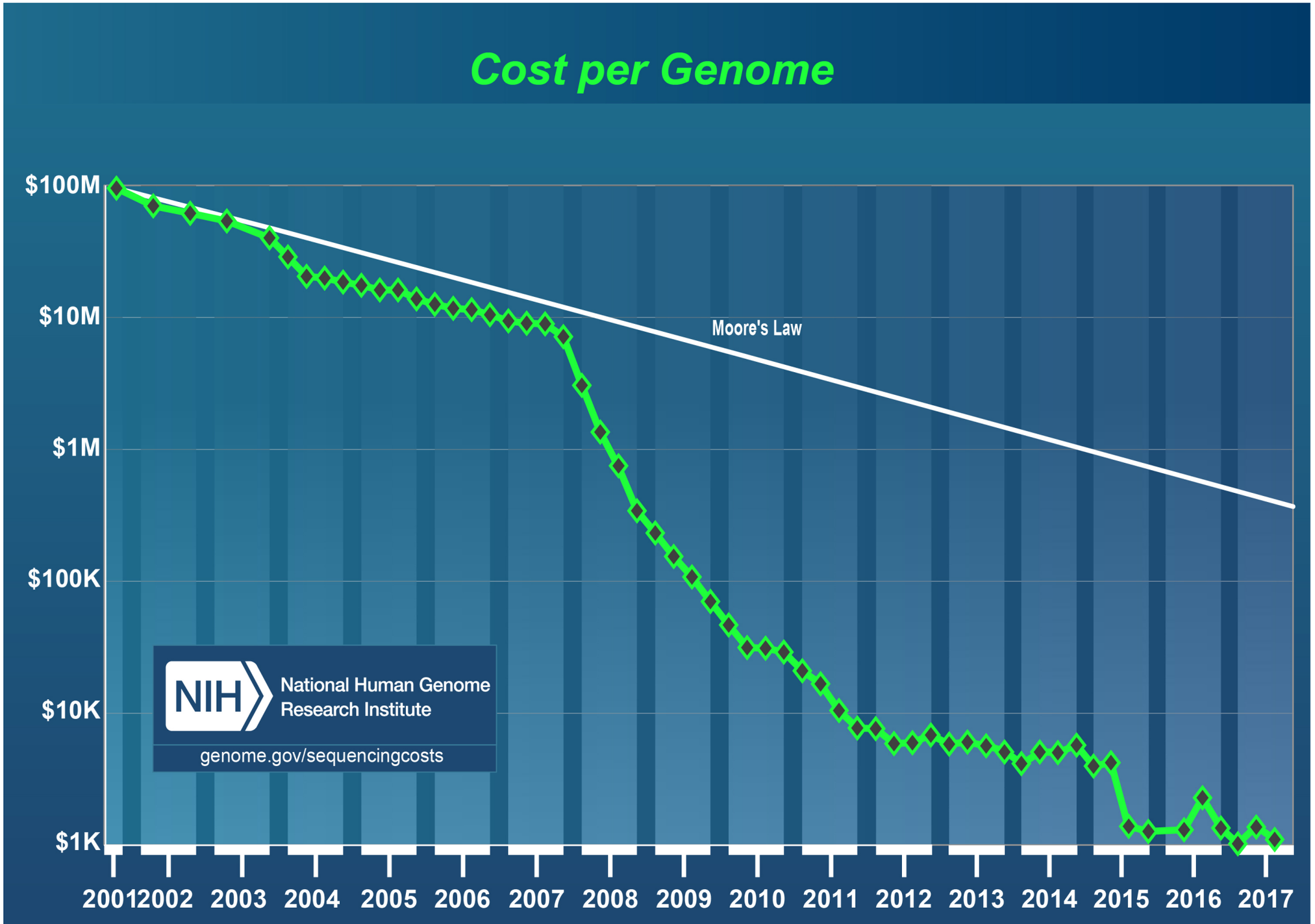
- Perhaps there are 3 main groups of Ulcerative Colitis patients:
    1. life-long remission after treatment with a commonly used monoclonal antibody
    2. initially have a multiyear remission but over the decades become refractory one after the other to each treatment and have to undergo colectomy
    3. initially have a remission but then no standard therapy works
- Could we have identified this patient as belonging to group 3 long before his crisis?
- Machine learning challenges:
    - Defining closeness to centrality of a specified population in PMMS: a distance function
    - Defining outliers in PMMS. Distance function may change the results considerably but it’s driven by the question you are asking.
    - Which is the best PMMS representation for time varying data?
    - What is the optimal weighting/normalization of dimensions in a PMMS? Is it task specific and if so how are the task-specific metrics determined.
    - How best to find the most specific neighborhood for a patient? What is a minimal size for such a neighborhood from the information theoretic perspective and from the practical “it makes no difference to be more precise” perspective?

A shallow dive into genetics

(following a lecture by Alvin Kho, Boston Children's Hospital)

- "Biology is the science of exceptions." — O. Pagan
- Children inherit traits from parents; how?
  - Gregor Mendel (~1854): discrete factors of inheritance, called "genes"
  - Johann Miescher (~1869): "nuclein", a compound in cell nuclei, now called DNA
  - Alfred Hershey & Martha Chase (1952): DNA, not protein, carries genetic info
  - James Watson, Francis Crick and Rosalind Franklin (1953): DNA is a double helix
- Gene:
  - "A fundamental physical and functional unit of heredity that is a DNA sequence located on a specific site on a chromosome which encodes a specific functional product (RNA, protein)." (From NCBI)
- Remaining mysteries
  - Still hard to find what parts of DNA code genes
  - What does the rest (vast majority) of DNA do? Control structure?
  - How does geometry affect this mechanism?
  - …

What Makes All This Possible?



~1¢/megabase

Whole Exome Sequencing Cost

- 25-day turnaround
- Advanced Analysis
  - Monogenic
  - Complex/multifactorial disorders
  - Cancer (for tumor-normal pair samples)

https://en.novogene.com/next-generation-sequencing-services/human-genome/whole-exome-sequencing-service/

RNA-Seq

- Measuring the transcriptome (gene expression levels, i.e., which genes are active, and to what degree)

- Single-Cell RNA Sequencing analyzes gene expression at the single-cell level for heterogeneous samples.
  - “The SMART-Seq HT Kit is designed for the synthesis of high-quality cDNA directly from 1–100 intact cells or ultra-low amounts of total RNA (10–1,000 pg).”
  - $360.00

https://www.genewiz.com/en/Public/Services/Next-Generation-Sequencing/RNA-Seq
https://www.takarabio.com/products/next-generation-sequencing/single-cell-rna-and-dna-seq/smart-seq-ht-for-streamlined-mrna-seq

Advanced Analysis

Monogenic disorders

1. Variant filtering
2. Analysis under dominant/recessive model (Pedigree information is needed)
   - 2.1. Analysis under dominant model
   - 2.2. Analysis under recessive model
3. Functional annotation of candidate genes
4. Pathway enrichment analysis of candidate genes
5. Linkage analysis
6. Regions of homozygosity (ROH) analysis

Complex/multifactorial disorders

All monogenic analyses plus…

1. De novo mutation analysis (Trio/Quartet)
   - 1.1. De novo SNP/ InDel detection
   - 1.2. Calculation of de novo mutation rates
2. Protein-protein interaction (PPI) analysis
3. Association analysis of candidate genes (at least 20 trios or case/ control pairs)

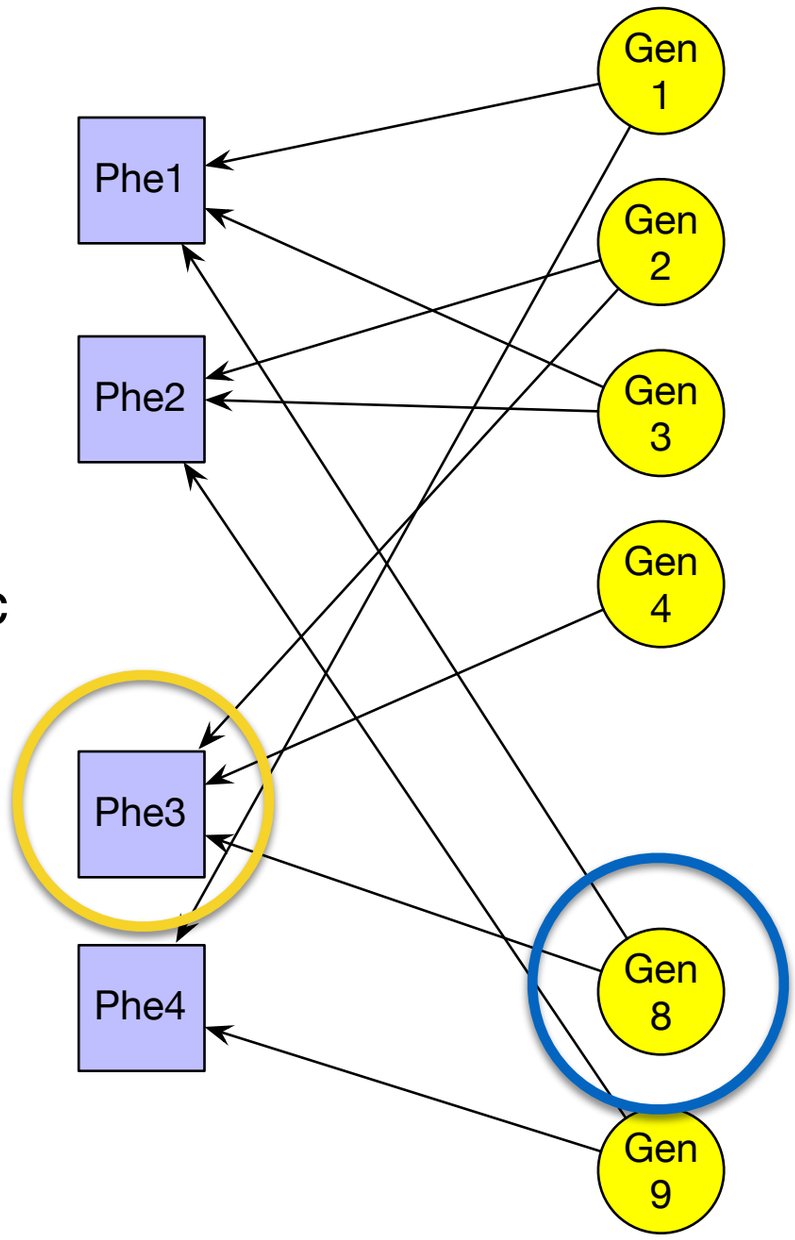
Cancer (for tumor-normal pair samples)

1. Screening for predisposing genes
2. Mutation spectrum & mutation signature analyses
3. Screening for known driver genes
4. Analyses of tumor significantly mutated genes
5. Analysis of copy number variations (CNV)
   - 5.1. distribution
   - 5.2. recurrence
6. Fusion gene detection
7. Purity & ploidy analyses of tumor samples
8. Tumor heterogeneity analyses
9. Tumor evolution analysis
10. Display of genomic variants with Circos

https://en.novogene.com/next-generation-sequencing-services/human-genome/whole-exome-sequencing-service/

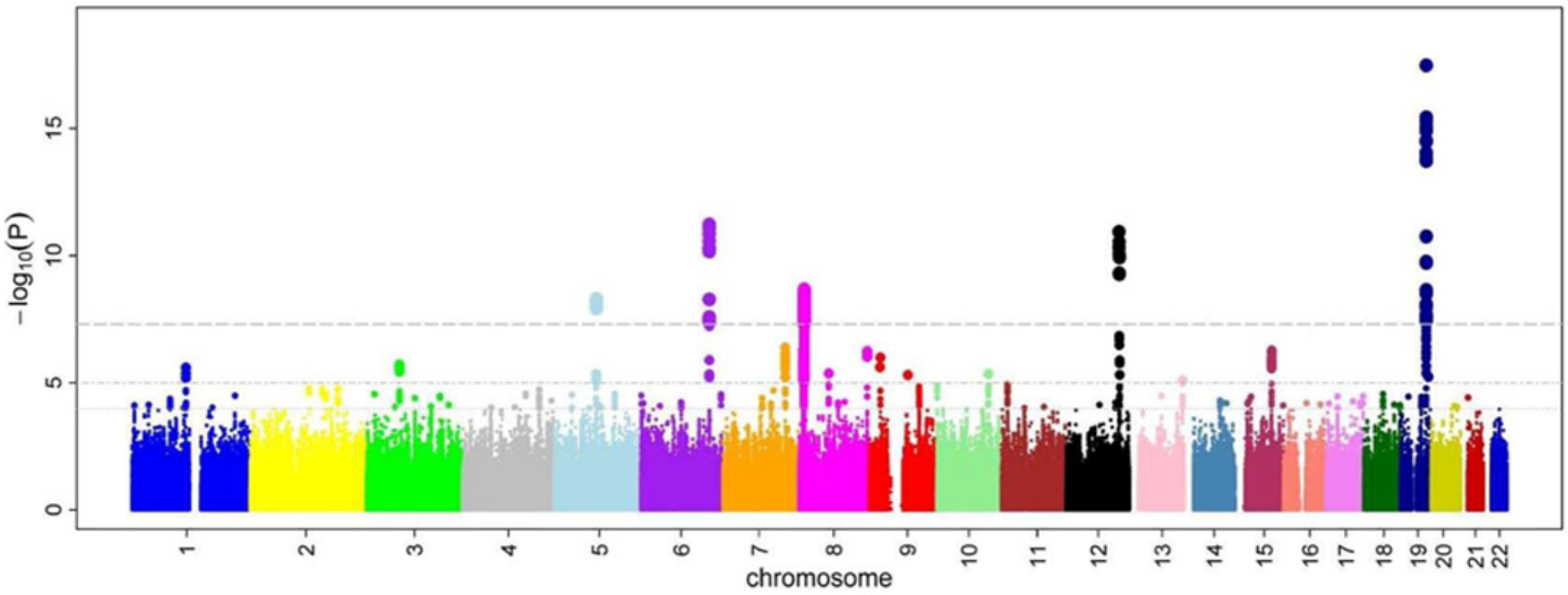
Relationships between Genotype and Phenotype

- What is a Phenotype?
  - Disease (e.g., breast cancer or normal; type of lymphoma)
  - Qualitative or quantitative traits (e.g., eye color, weight)
  - Behavior
  - …
- Gene-wide Association Studies (GWAS) look for genetic differences that correspond to specific phenotypic differences
  - Single-nucleotide polymorphisms (SNP) (n>1M)
  - Copy Number Variations (CNV)
  - Gene expression levels
  - *Looks at **all** genes, not a selected set*
- Phenome-wide Association Studies (PheWAS) look for phenotypic variations that correspond to specific genetic feature variations



GWAS

- Find gene variants associated with phenotype differences
- As of 2017, over 3,000 human GWA studies have examined over 1,800 diseases and traits, and thousands of SNP associations have been found.

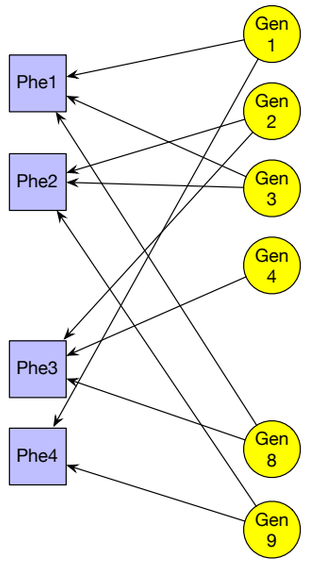


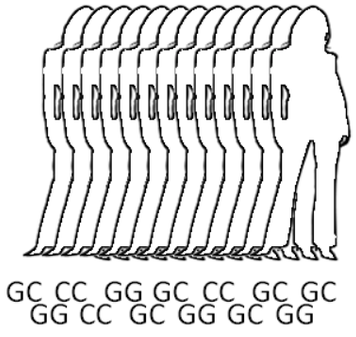
An illustration of a Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. This example is taken from a GWA study investigating microcirculation, so the tops indicates genetic variants that more often are found in individuals with constrictions in small blood vessels.

https://en.wikipedia.org/wiki/Genome-wide_association_study

GWAS

- Genotype a cohort of cases and controls, typically identifying >1M SNPs
- For each SNP, compute odds of disease given the SNP [O(D|S)] and odds of disease given no SNP [O(D|~S)]
- Odds ratio, O(D|S) / O(D|~S) is measure of association between this SNP and the phenotype; if different from 1, indicates association

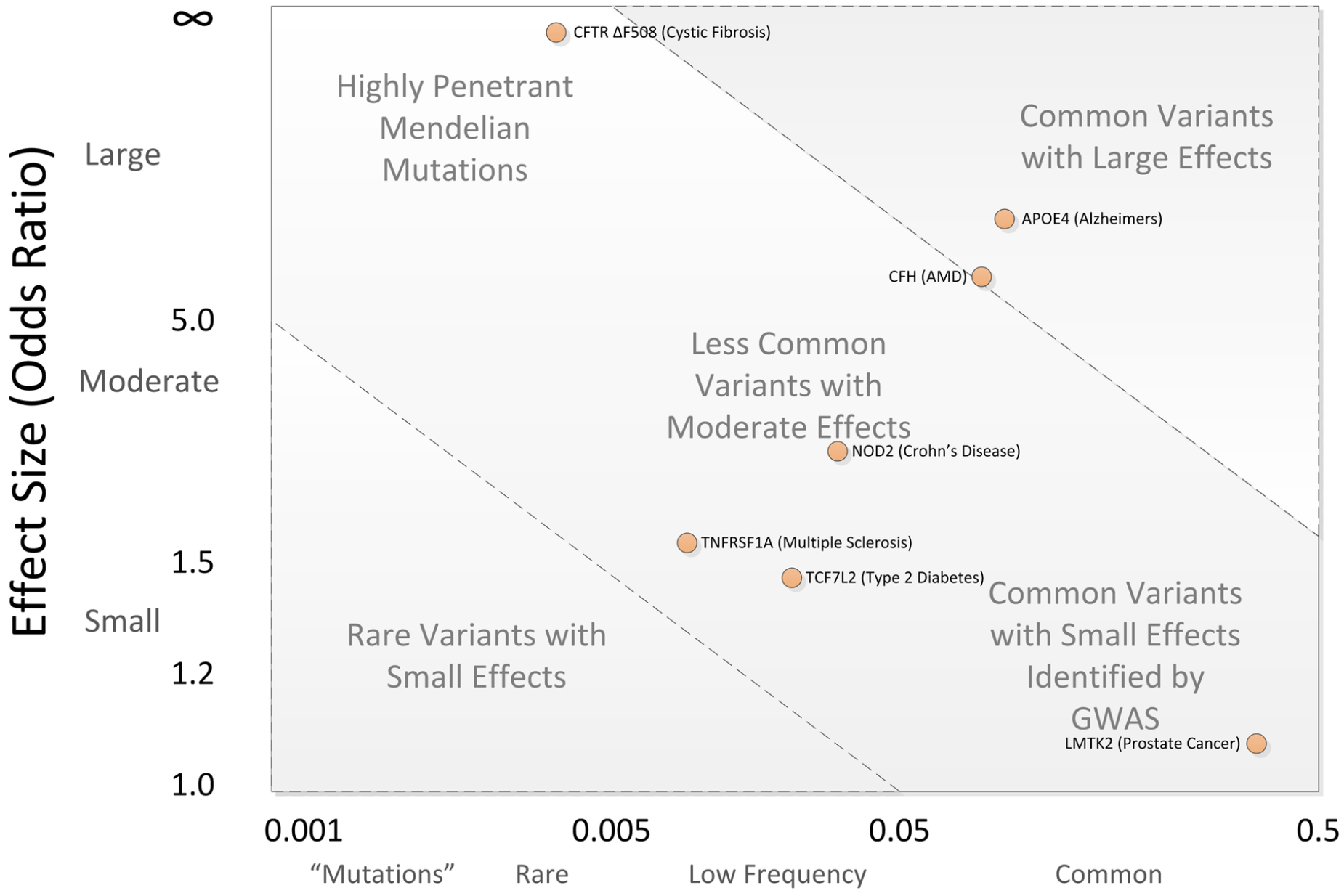




	SNP1	*SNP2*	*SNP…*
	Cases	**Cases**	*Repeat for all SNPs*
	Count of G: 2104 of 4000	Count of G: 1648 of 4000	
	Frequency of G: 52.6%	Frequency of G: 41.2%	
	Controls	**Controls**	
	Count of G: 2676 of 6000	Count of G: 2532 of 6000	
	Frequency of G: 44.6%	Frequency of G: 42.2%	
	P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

https://upload.wikimedia.org/wikipedia/commons/1/1e/Method_example_for_GWA_study_designs.png

“GWA studies typically identify common variants with small effect sizes (lower right).”



https://en.wikipedia.org/wiki/Genome-wide_association_study#/media/File:GWAS_Disease_allele_effects.png

Udler, M. S., Kim, J., Grotthuss, von, M., Bonàs-Guarch, S., Cole, J. B., Chiou, J., et al. (2018). Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Medicine, 15(9), e1002654–23. http://doi.org/10.1371/journal.pmed.1002654

Example: GWAS of Type-II Diabetes

- Goal: identify “soft” clusters of genetic loci to suggest subtypes of T2D and possible mechanisms
- “Over the past decade, genome-wide association stud- ies (GWAS) and other large-scale genomic studies have identified over 100 loci associated with T2D, causing modest increases in disease risk (odds ratios generally <1.2)”
- Data selected from multiple previous studies:
  - 94 T2D-associated variants
  - glycemic traits — fasting insulin, fasting glucose, fasting insulin adjusted for BMI, 2-hour glucose on oral glucose tolerance test [OGTT] adjusted for BMI [2hrGlu adj BMI], glycated hemoglobin [HbA1c], homeostatic model assess- ments of beta cell function [HOMA-B] and insulin resistance [HOMA-IR], incremental insu- lin response at 30 minutes on OGTT [Incr30], insulin secretion at 30 minutes on OGTT [Ins30], fasting proinsulin adjusted for fasting insulin, corrected insulin response [CIR], disposition index [DI], and insulin sensitivity index [ISI]
  - BMI, height, waist circumference [WC] with and without adjustment for BMI, and waist-hip ratio [WHR] with and without adjustment for BMI; birth weight and length; % body fat, HR
  - lipid levels (HDL cholesterol, low-density lipoprotein [LDL] cholesterol, total cholesterol, triglycerides), leptin with and without BMI adjustment, adiponectin adjusted for BMI, urate [35], Omega-3 fatty acids, Omega-6-fatty acids, plasma phospholipid fatty acids in the de novo lipogenesis pathway, and very long-chain saturated fatty acids
  - Associations with: ischemic stroke, coronary artery disease, renal function (eGFR), urine albumin-creatinine ratio (UACR); chronic kidney disease (CKD); and systolic (SBP) and diastolic blood pressure (DBP)

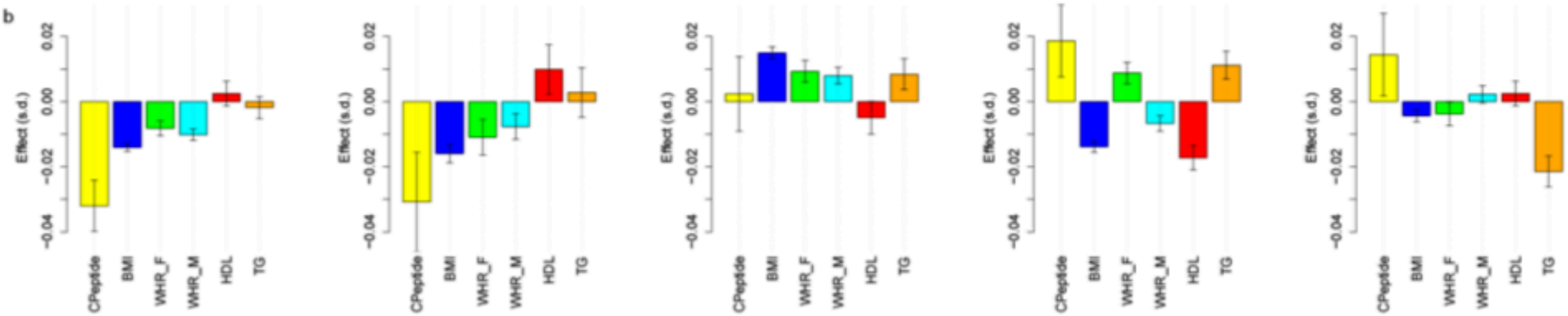
Results

- Five subtypes of T2D (“identification of five robust clusters present on 82.3% of iterations”), with their interpretations:
  - Beta-cell
  - Proinsulin
  - Obesity
  - Lypodistrophy
  - Liver/Lipid

a


Beta-cell
Proinsulin
Obesity
Lipodystrophy
Liver/Lipid
HOMA-B
ProIns
TG
HDL
WC
BMI
FastIns
HOMA-IR
0

b


Effect (s.d.)
0.02
0.00
−0.02
−0.04
CPeptide
BMI
WHR_F
WHR_M
HDL
TG

c

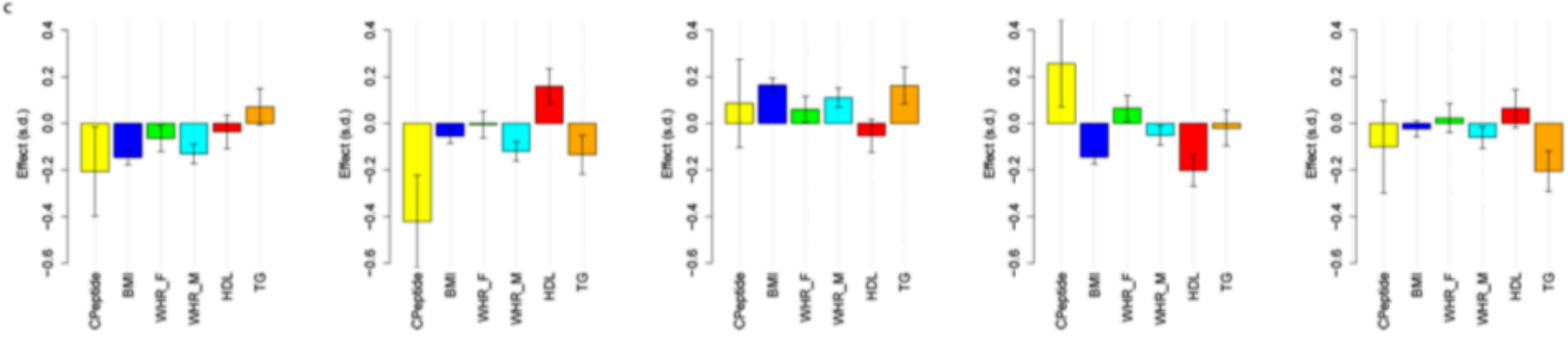

Effect (s.d.)
0.4
0.2
0.0
−0.2
−0.4
−0.6
CPeptide
BMI
WHR_F
WHR_M
HDL
TG

Fig 1. Cluster-defining characteristics. (A) Standardized effect sizes of cluster GRS-trait associations derived from GWAS summary statistics shown in spider plot. The middle of the three concentric octagons is labeled "0," representing no association between the cluster GRS and trait. A subset of discriminatory traits are displayed. Points falling outside the middle octagon represent positive cluster-trait associations, whereas those inside it represent negative cluster-trait associations. (B) Associations of GRSs in individuals with T2D with various traits. Results are from four studies (METSIM, Ashkenazi, Partners Biobank, and UK Biobank) meta-analyzed together. Effect sizes are scaled by the raw trait standard deviation. (C) Differences in trait effect sizes between individuals with T2D having GRSs in the highest decile of a given cluster versus all other individuals with T2D. Results are from the same four studies meta-analyzed together. Effect sizes are scaled by the raw trait standard deviation. BMI, body mass index; Fastins, fasting insulin; GRS, genetic risk score; GWAS, genome-wide association study; HDL, high-density lipoprotein; HOMA-B, homeostatic model assessment of beta cell function; HOMA-IR, homeostatic model assessment of insulin resistance; METSIM, Metabolic Syndrome in Men Study; ProIns, fasting proinsulin adjusted for fasting insulin; TG, serum triglycerides; T2D, type 2 diabetes; WC, waist circumference; WHR-F, waist-hip ratio in females; WHR-M, waist-hip ratio in males.

PheWAS = “reverse GWAS”

- GWAS studies generalized from one to multiple phenotypes
- Unlike SNPs, phenotypes were not well characterized
  - Billing codes, EHR data, temporal progression
- Vanderbilt example:
  - (2010) biobank held 25,769 samples
  - first 6,000 European-Americans with samples; no other criteria
  - five SNPs:
    - rs1333049 [coronary artery disease (CAD) and carotid artery stenosis (CAS)],
    - rs2200733 [atrial fibrillation (AF)],
    - rs3135388 [multiple sclerosis (MS) and systemic lupus erythematosus (SLE)],
    - rs6457620 [rheumatoid arthritis (RA)],
    - rs17234657 [Crohn’s disease (CD)]
- Defined PheWAS code table, cleaning up ICD-9-CM to 744 case groups
  - https://phewascatalog.org/phecodes
  - E.g., tuberculosis = {010-018 (TB in various organs), 137 (late effects of tuberculosis), 647.3 (tuberculosis complicating the peripartum period)}
  - (2015) 1866 PheWAS codes, with 1-496 ICD codes grouped [TB is the one with 496!]

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics (Oxford, England), 26(9), 1205–1210. http://doi.org/10.1093/bioinformatics/btq126

Diseases Associated with SNP rs3135388

- Expected MS, SLE

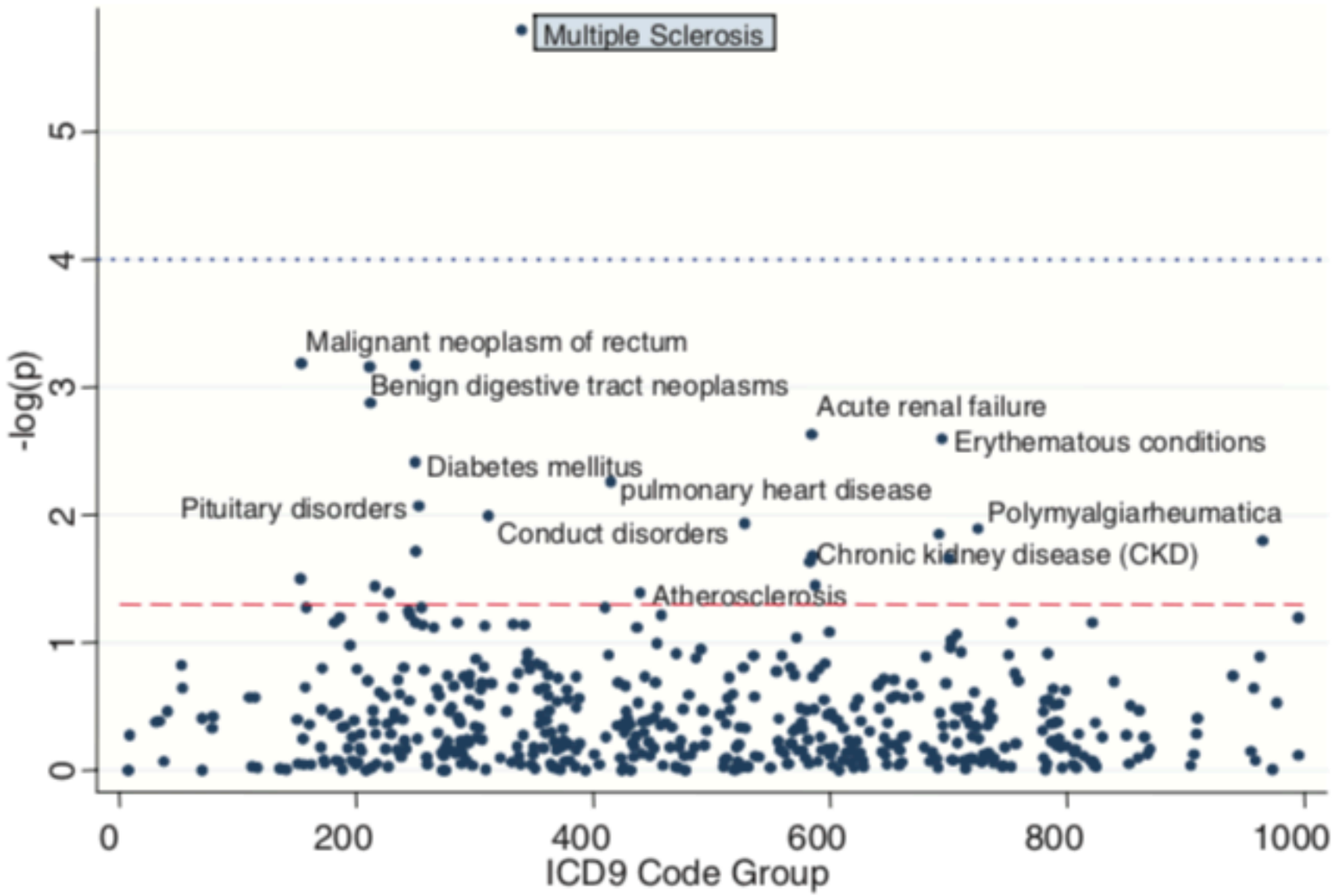


Fig. 1. Phenome-wide scan for association with rs3135388. MS is replicated from prior analyses. The dashed line represents the $P = 0.05$; the dotted line represents the Bonferroni correction.

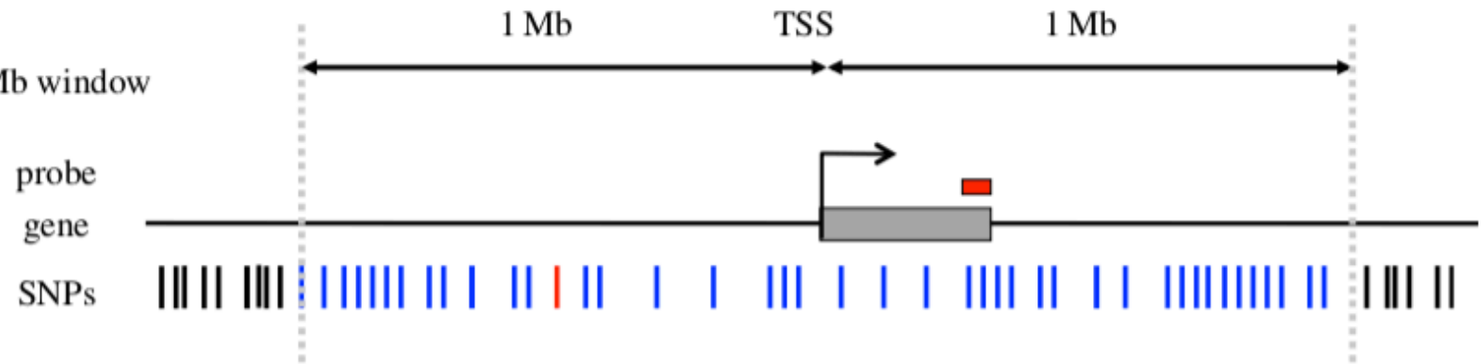
You Don’t Always Get What You Expect

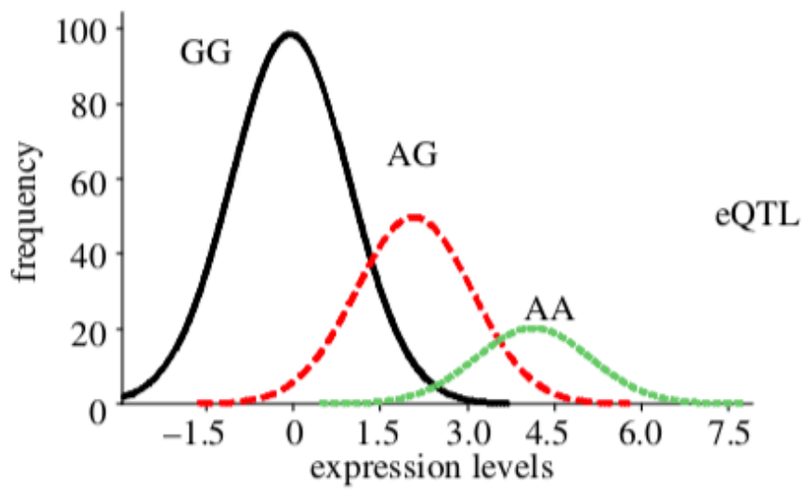
Table 2. Diseases previously associated with the five SNP studied and current PheWAS ORs

SNP	Gene/region	Disease	Cases	Previous OR	PheWAS *P*-value	PheWAS OR
rs3135388	DRB1*1501	MS	89	1.99[a]	2.77×10^{-6}	2.24 (1.56–3.16)
		SLE	141	2.06[b]	0.51	1.13 (0.79–1.58)
rs17234657	Chr. 5	CD	200	1.54[c]	0.00080	1.57 (1.19–2.04)
rs2200733	Chr. 4q25	AF and flutter	606	1.75[d]	0.14	1.15 (0.95–1.39)
rs1333049	Chr. 9p21	CAD	1181	1.20–1.47[e]	0.011	1.13 (1.03–1.23)
		Carotid atherosclerosis	333	1.46[f]	0.82	0.98 (0.84–1.15)
rs6457620	Chr. 6	RA[g]	392	2.36[c]	0.0002	1.35 (1.15–1.58)

Expression Quantitative Trait Loci (eQTLs)

- Genetic variants that explain quantitative expression levels
  - i.e., use expression levels to define phenotype
  - no need for clinical knowledge, human judgment
  - potential to explain genetic mechanisms





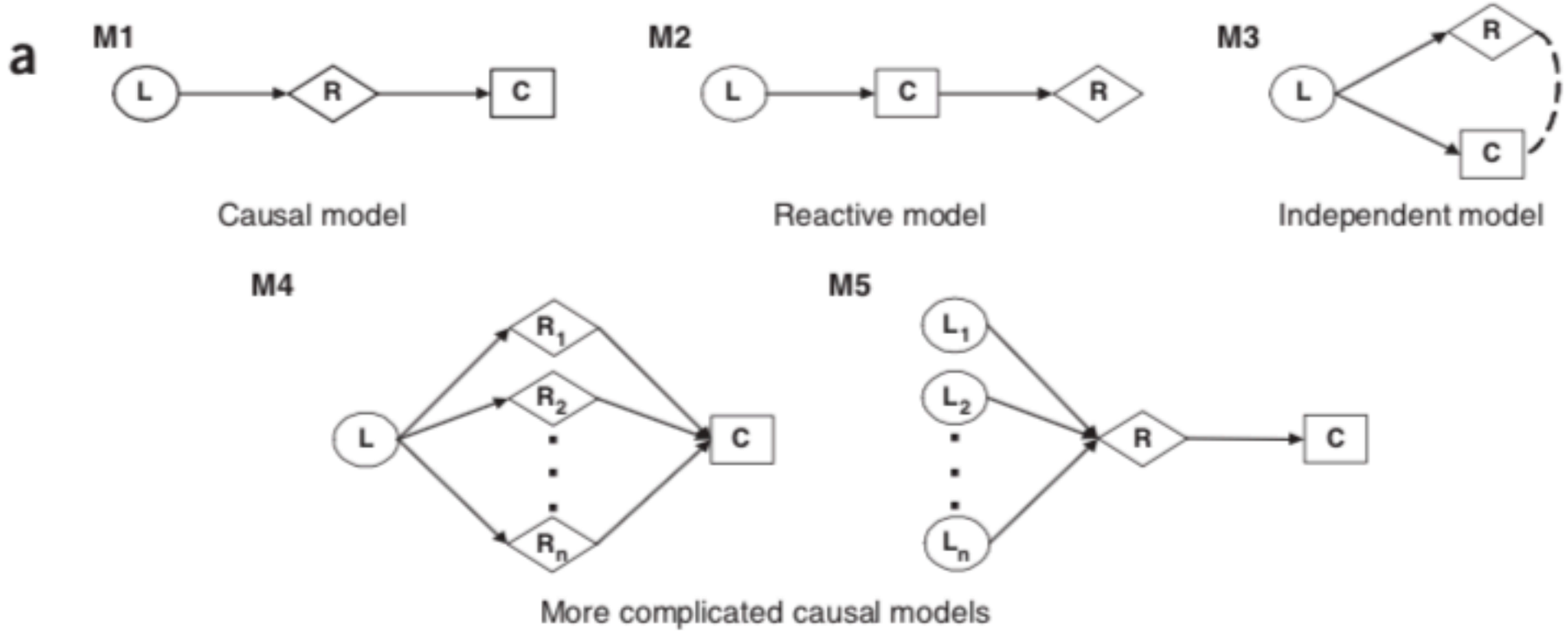
Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1620), 20120362–20120362. http://doi.org/10.1098/rstb.2012.0362

Differential Expression in Different Populations

- European — African: 17% of genes in small sample (16 people)
- European — Asian: 1097/4197 = 26%
- 4 populations from HapMap sample of 270 people: 17-29% different expression levels
- But:
  - Some effect may be environmental
  - Large differences between different tissues (most early studies used only blood)
  - Limited correlation to disease phenotypes
- Nevertheless:
  - Evidence for suspect causative genes in various diseases: asthma, Crohn's
- "The large-scale disease studies performed so far have uncovered multiple variants of low-effect sizes affecting multiple genes. This suggests that common forms of disease are most probably not the result of single gene changes with a single outcome, but rather the outcome of perturbations of gene networks which are affected by complex genetic and environmental interactions."

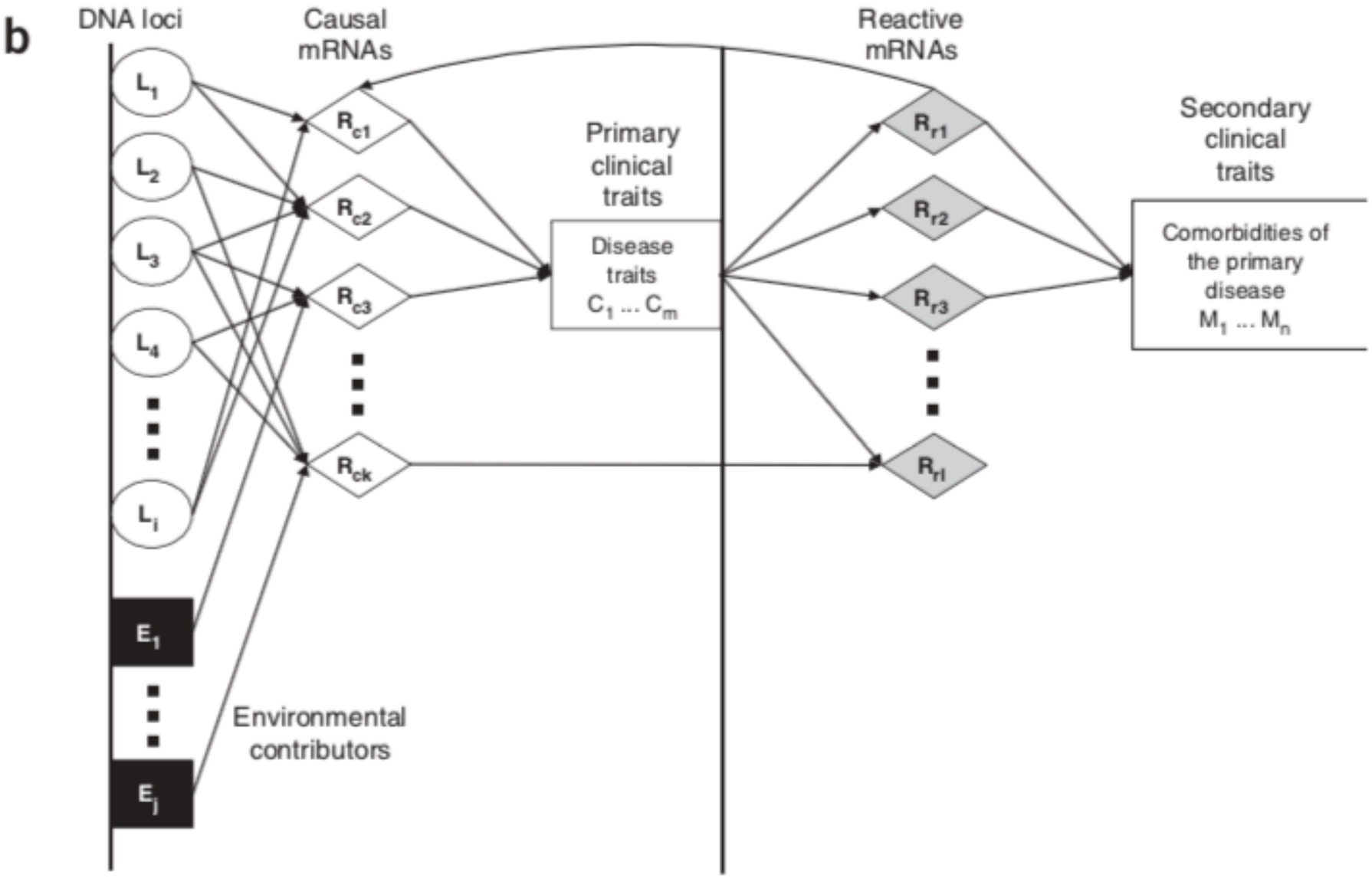
QTL, eQTL, & Disease Traits

- L = QTL
- R = RNA expression level (eQTL)
- C = complex trait
- Model that best fits data is most likely



Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics, 37(7), 710–717. http://doi.org/10.1038/ng1589

A More Complex Story



Scaling Up Gene-Phene Association Studies

- UK Biobank collects data on ~.5M de-identified individuals
  - everyone will have full exome sequencing (50K so far)
  - 100K have worn 24-hour activity monitor for a week, 20K have had repeat measurements
  - on-line questionnaires: diet, cognitive function, work history, digestive health
  - 100K will have imaging: brain, heart, abdomen, bones, carotid artery
  - linking to EHR: death, cancer, hospital episodes, GP, blood biochemistry
  - developing more accurate phenotyping
- Ongoing stream of results
  - April 18th, 2019: Genetic variants that protect against obesity and type 2 diabetes discovered
  - April 17th, 2019: Moderate meat eaters at risk of bowel cancer
  - April 8th, 2019: Research identifies genetic causes of poor sleep

https://www.ukbiobank.ac.uk

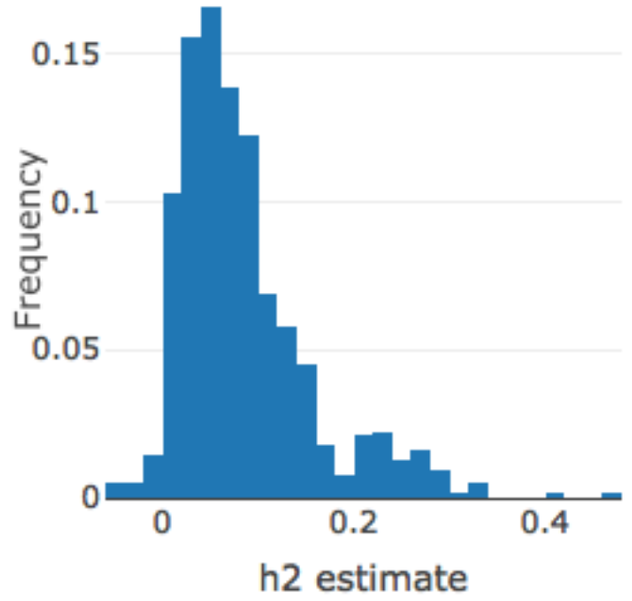
UK Biobank GWAS

- Users; e.g., Neale Lab @ MGH & Broad
  - Phenome scan in UK Biobank (https://github.com/MRCIEU/PHESANT)
    - PHESANT “traits": 2891 total (274 continuous / 271 ordinal / 2346 binary)
    - Aug 2018: 4,203 phenotypes
  - ICD10: 633 binary
  - FinnGen curated: 559
  - imputed-v3 model (“a ‘quick-and-dirty’ analysis that strives to provide a reliable, albeit imperfect, insight into the UK Biobank data”)
    - Linear regression model in Hail (linreg)
    - Three GWAS per phenotype
      - Both sexes
      - Female only
      - Male only
    - Covariates: 1st 20 PCs + sex + age + age^2 + sexage + sexage2
    - Sex-specific covariates: 1st 20 PCs + age + age^2

https://www.ukbiobank.ac.uk

Heritability

- Most heritable traits look genetic for large sample sizes
  - Height (h^2 =.46, p=7.5e-109)
  - College degree (h^2 =.28, p=6.6e-195)
  - TV watching (h^2 =.096, p=2.8e-114)
- How much insight does this convey?



Distribution of LDSR SNP-heritability estimates for phenotypes with N_{eff} > 10,000.

http://www.nealelab.is/blog/2017/9/20/insights-from-estimates-of-snp-heritability-for-2000-traits-and-disorders-in-uk-biobank

Table 1. Deep Learning Architectures and Approaches for Omics Analysis

Method	*Key features*	*Input data and applications*
CNN	Hierarchical architecture commonly used for image classification Includes convolution and pooling layers (Miotto et al., 2017) Detection of locally and globally consistent features in the data (Min et al., 2017a) Strength: established architectures useful for encoding complex local and global interactions (e.g., relationships between DNA motifs) (Angermueller et al., 2016)	Multidimensional arrays such as DNA-seq, DNase-seq, protein-binding microarrays, and ChIP-seq Prediction of binding site, nucleosome positioning, and DNA accessibility (Alipanahi et al., 2015; Kelley et al., 2016; Min et al., 2017b; Zhang et al., 2018)
RNN	Sequential architecture useful for text and time series data (Wenpeng et al., 2017) Cyclic connections share information from previous and current state (Min et al., 2017a) Strength: identification of latent relationships in sequential (Angermueller et al., 2016)	Sequential data such as genomic sequences or natural language Prediction of protein structure, gene expression regulation, protein homology, and DNA methylation (Angermueller et al., 2017; Li et al., 2017a; Seunghyun et al., 2016; Søren and Ole, 2014)
AE	Unsupervised learning Combination of encoder and decoder is used to predict the input data and is useful for detecting consistent patterns in the data (Miotto et al., 2017) Strength: nonsupervised identification of major patterns in the data (Ching et al., 2018)	Genome-scale omics data such as gene expression data Identification of informative features (Ding et al., 2018; Gupta et al., 2015)
DNN-MDA (Date and Kikuchi, 2018)	Application of DNN for construction of classification and regression models, and estimation of variable importance by an MDA Strength: estimation of variable importance	NMR-based metabolite profiling Identification of biomarkers
DeepNovo (Tran et al., 2017)	Integrating CNN and LSTM RNN Strength: combining useful features from CNN and RNN	Tandem mass spectra of proteomics data Prediction of novel peptide sequence

AE, autoencoder; CNN, convolutional neural network; DNN, deep neural network; LSTM, long short-term memory; MDA, mean decrease accuracy; NMR, nuclear magnetic resonance; RNN, recurrent neural network.

Grapov, D., Fahrmann, J., Wanichthanarak, K., & Khoomrung, S. (2018). Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. OMICS: a Journal of Integrative Biology, 22(10), 630–636.