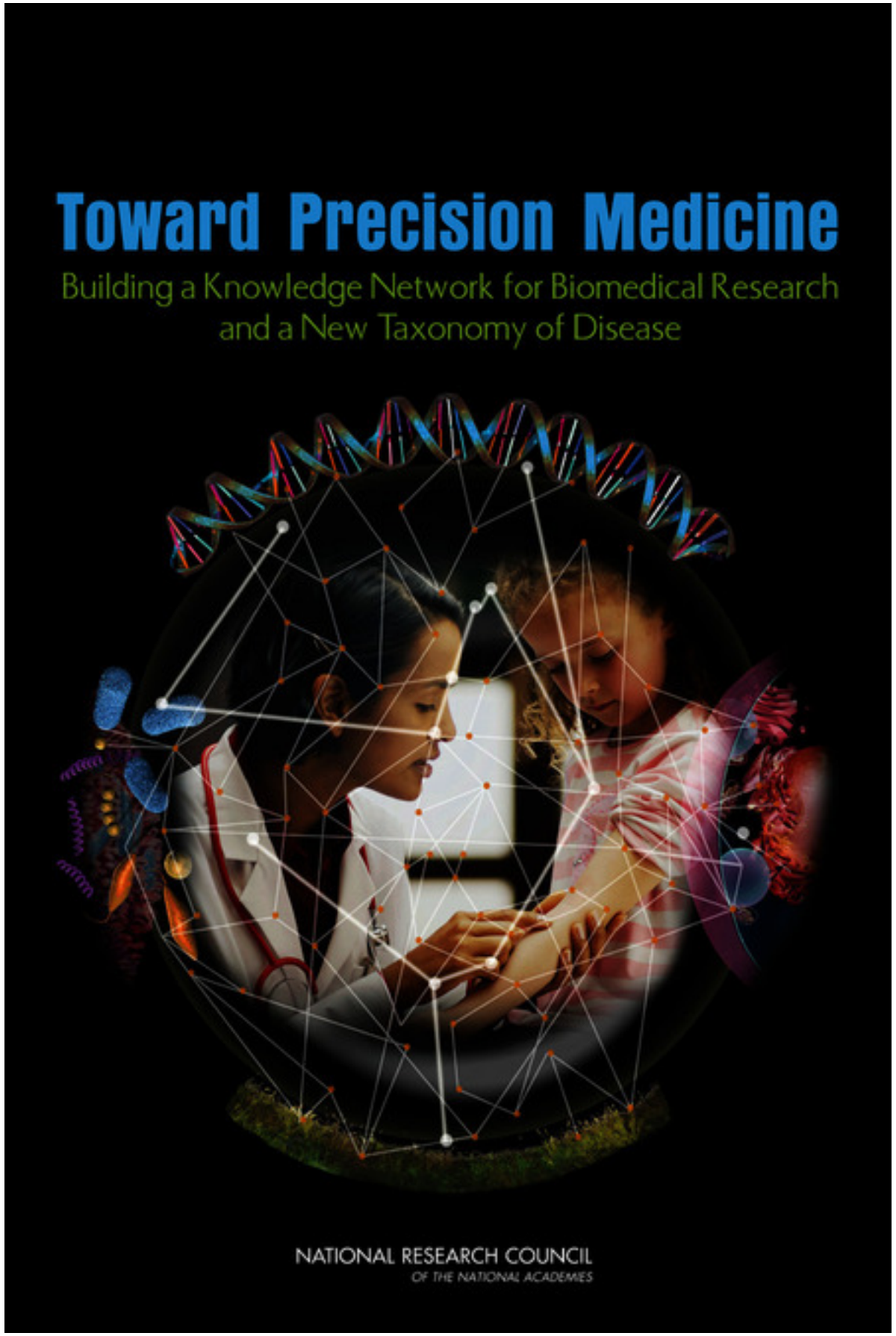
Precision Medicine

April 25, 2019

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.

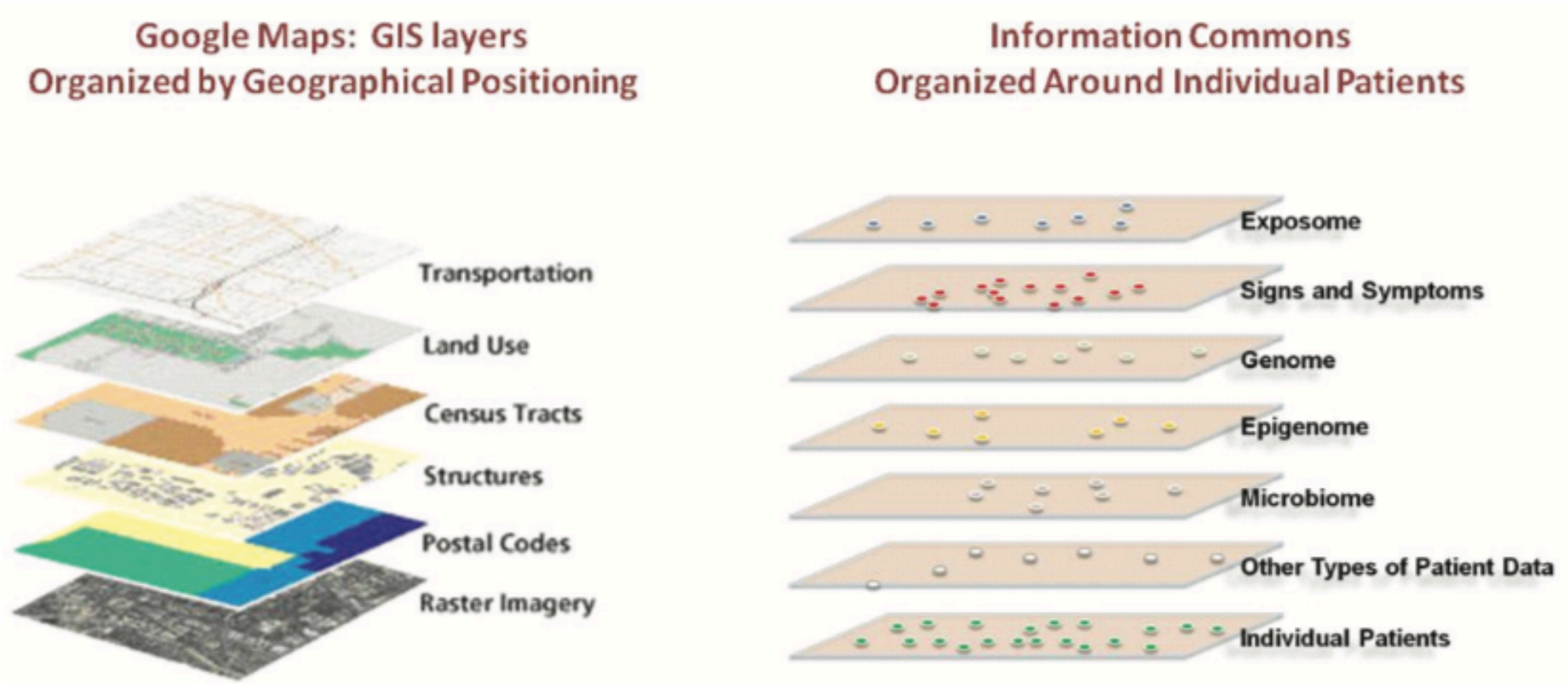


FIGURE 1-2 An Information Commons might use a GIS-type structure.
The proposed, individual-centric Information Commons (right panel) is somewhat analogous to a layered GIS (left panel). In both cases, the bottom layer defines the organization of all the overlays. However, in a GIS, any vertical line 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).

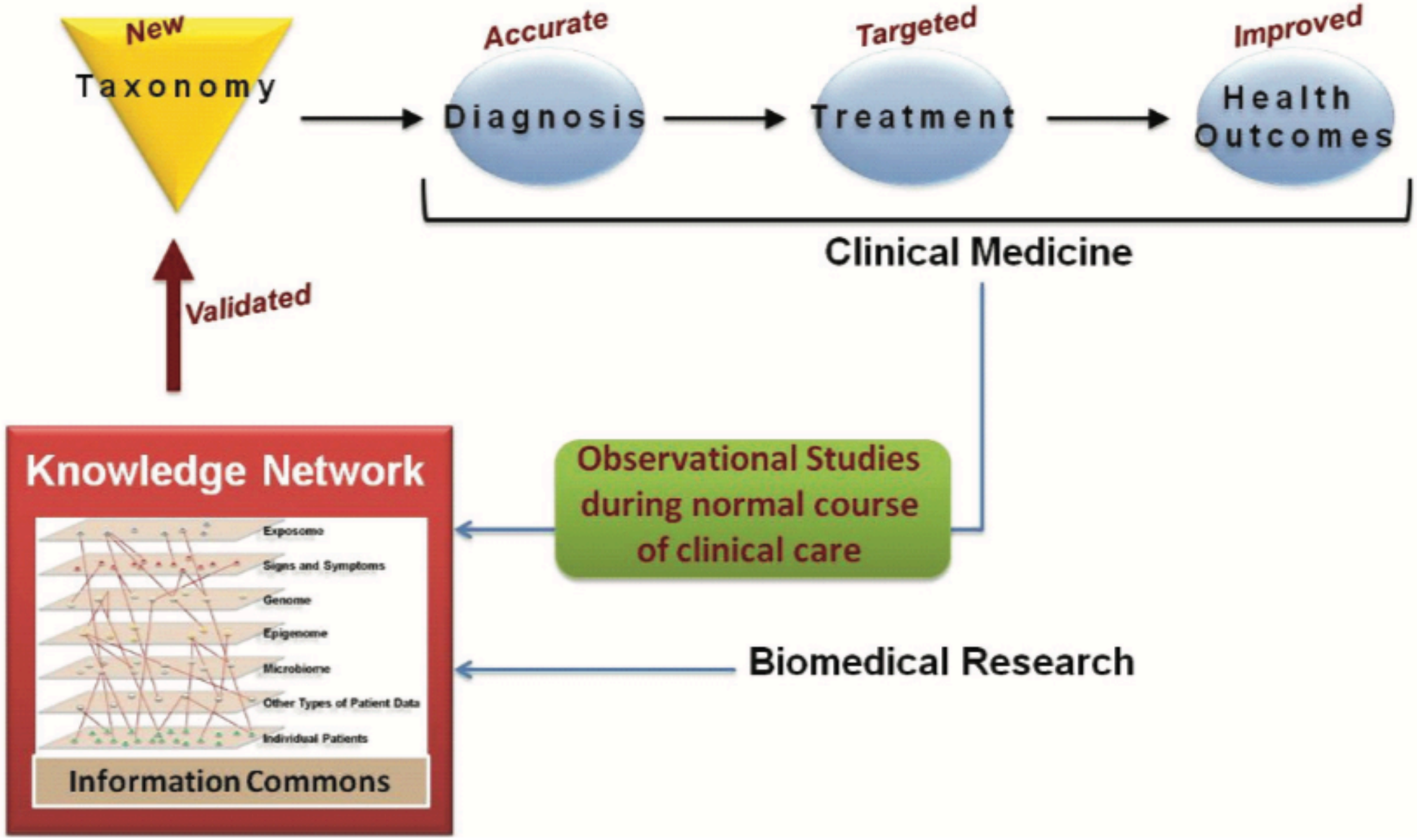
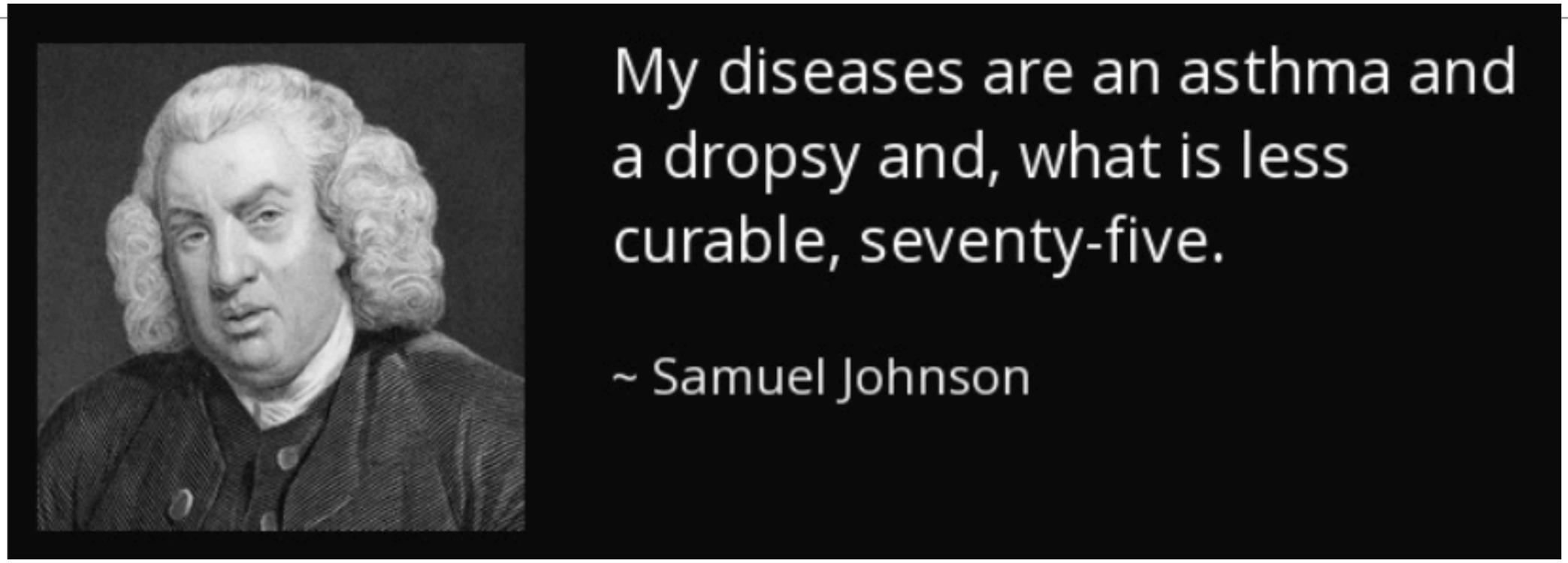


FIGURE 1-3 A knowledge network of disease would enable a new taxonomy. An individual-centric Information Commons, in combination with all extant biological knowledge, will inform a Knowledge Network of Disease, which will capture the exceedingly complex causal influences and pathogenic mechanisms that 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.

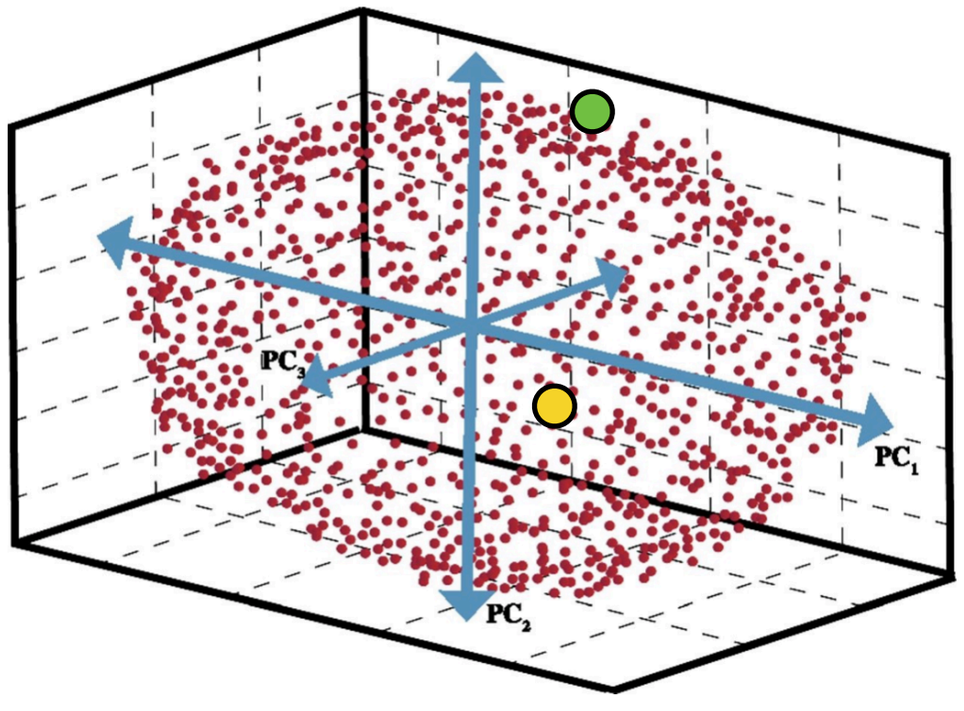
Centrality of Taxonomy (as a *hypothesis*)



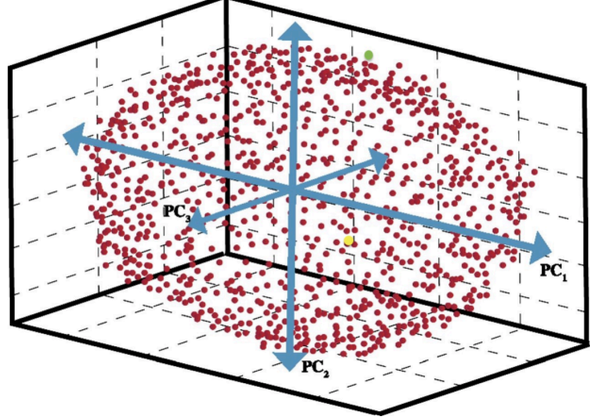
- What is “dropsy”?
  - “water sickness”, “swelling”, “edema”
  - *disease that got Grandma to take to her bed permanently in Victorian dramas*
  - causes: COPD, CHF, CKD, …
  - Last recorded on a death certificate ~1949
- Is “asthma” equally non-specific?

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?
  - …

Central Dogma of Molecular Biology

- Francis Crick, 1958 — *at the time, controversial and tentative*
  - Sequence Hypothesis
    - “the specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and that this sequence is a (simple) code for the amino acid sequence of a particular protein”
  - Central Dogma
    - “the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible”
- … a few Nobel Prizes later:
  - Transcription is regulated by *promoter*, *repressor*, and *enhancer* regions on the genome, to which proteins bind.

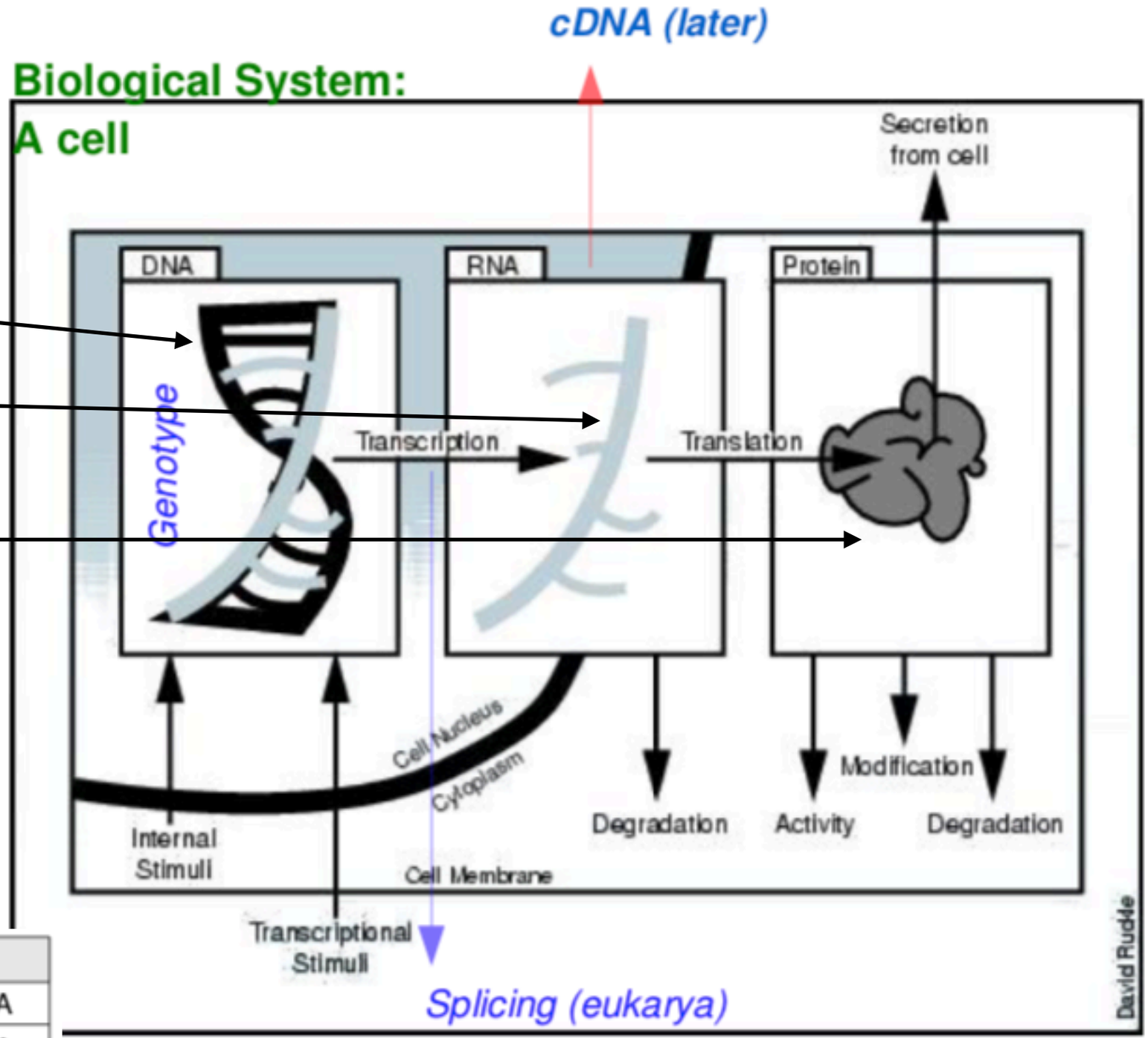
Symp Soc Exp Biol. 1958;12:138-63

Current Interpretation of Central Dogma

DNA: C, G, A, T double strand

RNA: C, G, A, U single strand

Protein: 21 amino acids
(genetic code, codon)

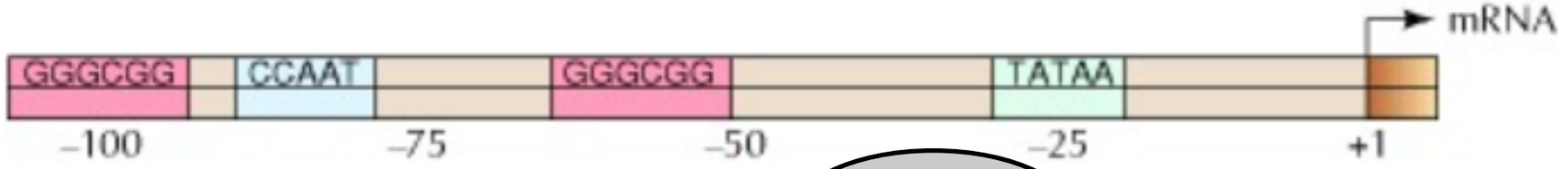


General	Special	Unknown
DNA → DNA	RNA → DNA	protein → DNA
DNA → RNA	RNA → RNA	protein → RNA
RNA → protein	DNA → protein	protein → protein

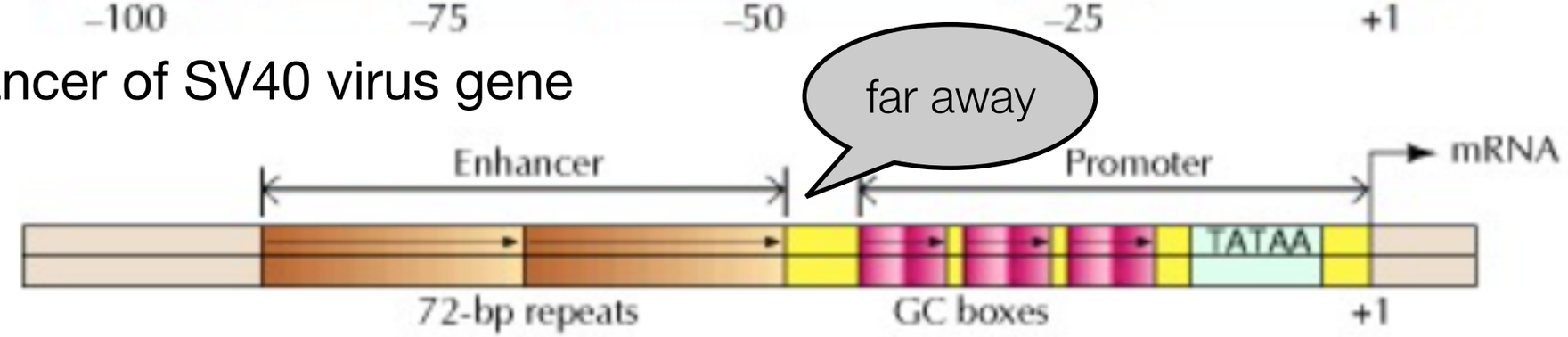
Kohane et al, MIT Press, 2003

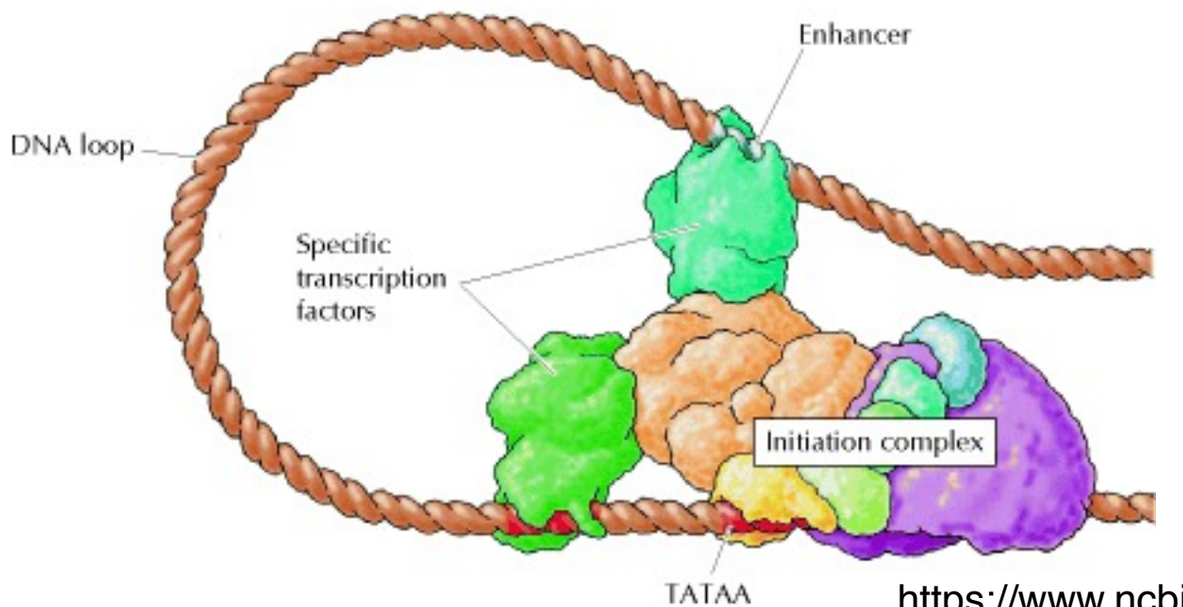
A few Nobel Prizes Later…

- Transcription is regulated by *promoter*, *repressor*, and *enhancer* regions on the genome, to which proteins bind.
  - Promoter of the thymidine kinase gene of herpes simplex virus



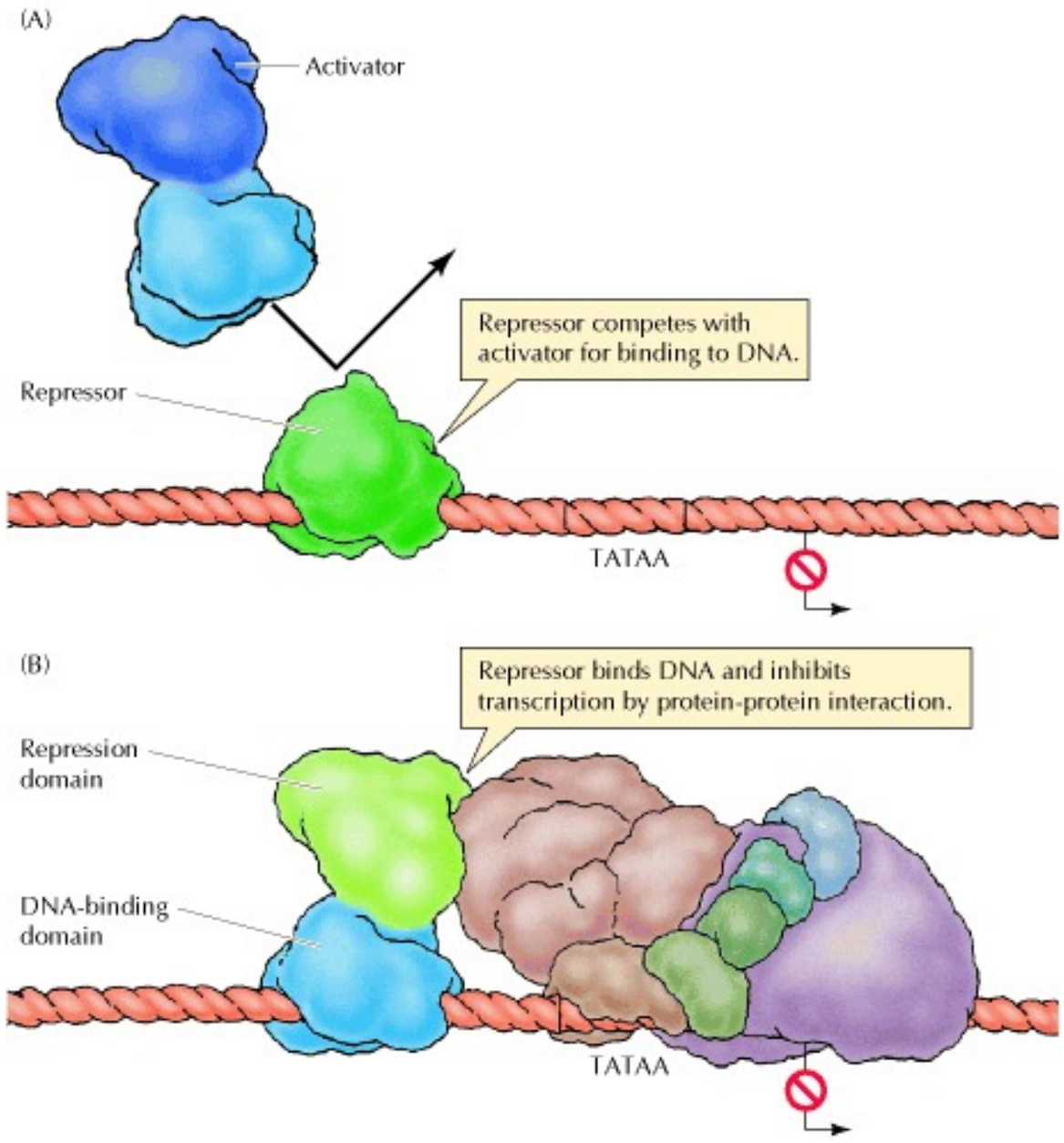
  - Enhancer of SV40 virus gene



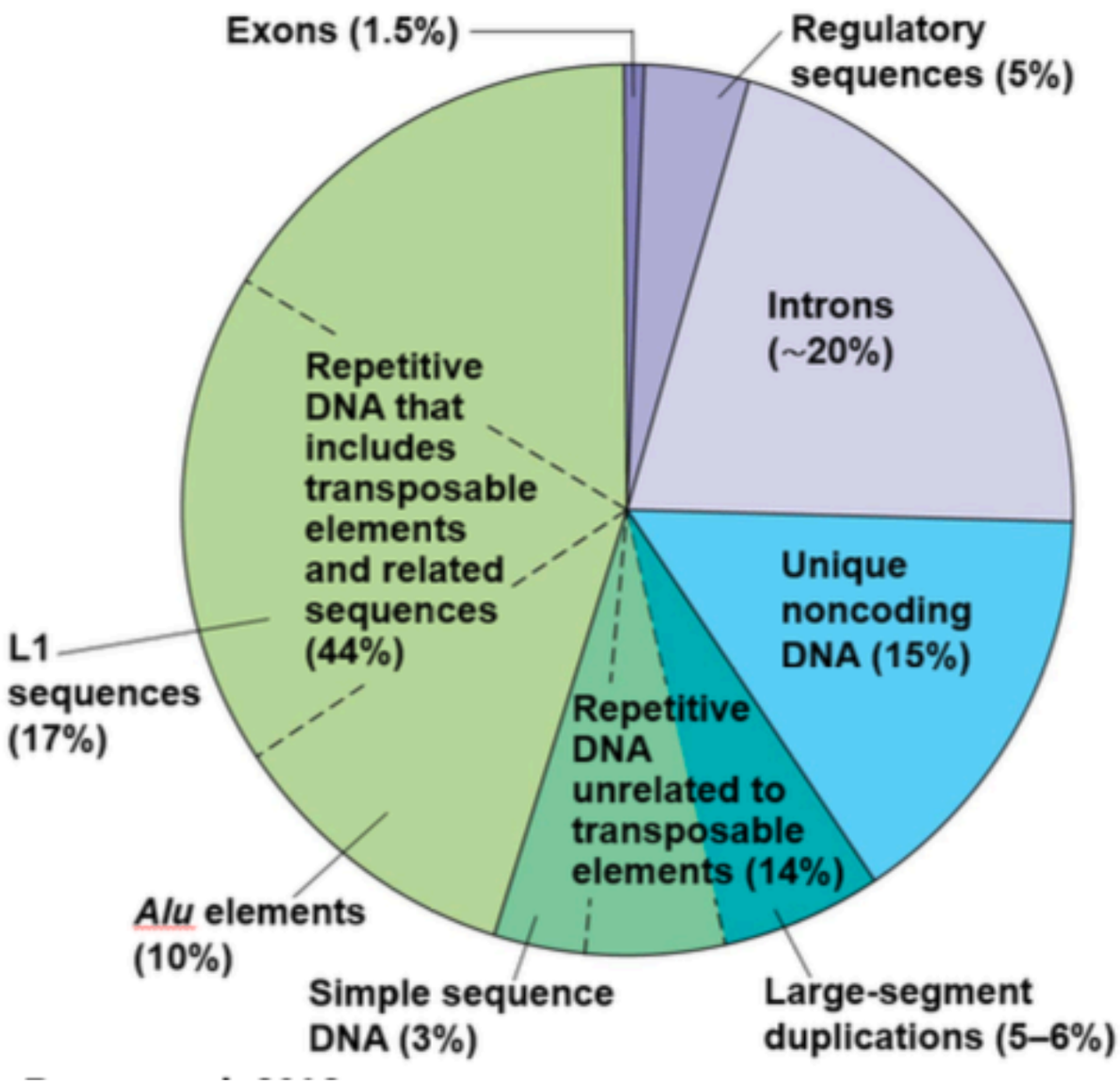


https://www.ncbi.nlm.nih.gov/books/NBK9904/

- Repressor prevents activator from binding or alters activator



Genes are a tiny fraction of the genome!



Reece et al. 2013

FULL SCREEN

MIT Professor Gerald Fink delivers the 2018-2019 James R. Killian Jr. Faculty Achievement Award Lecture, titled, "What is a Gene?"

Photo: Jake Belcher

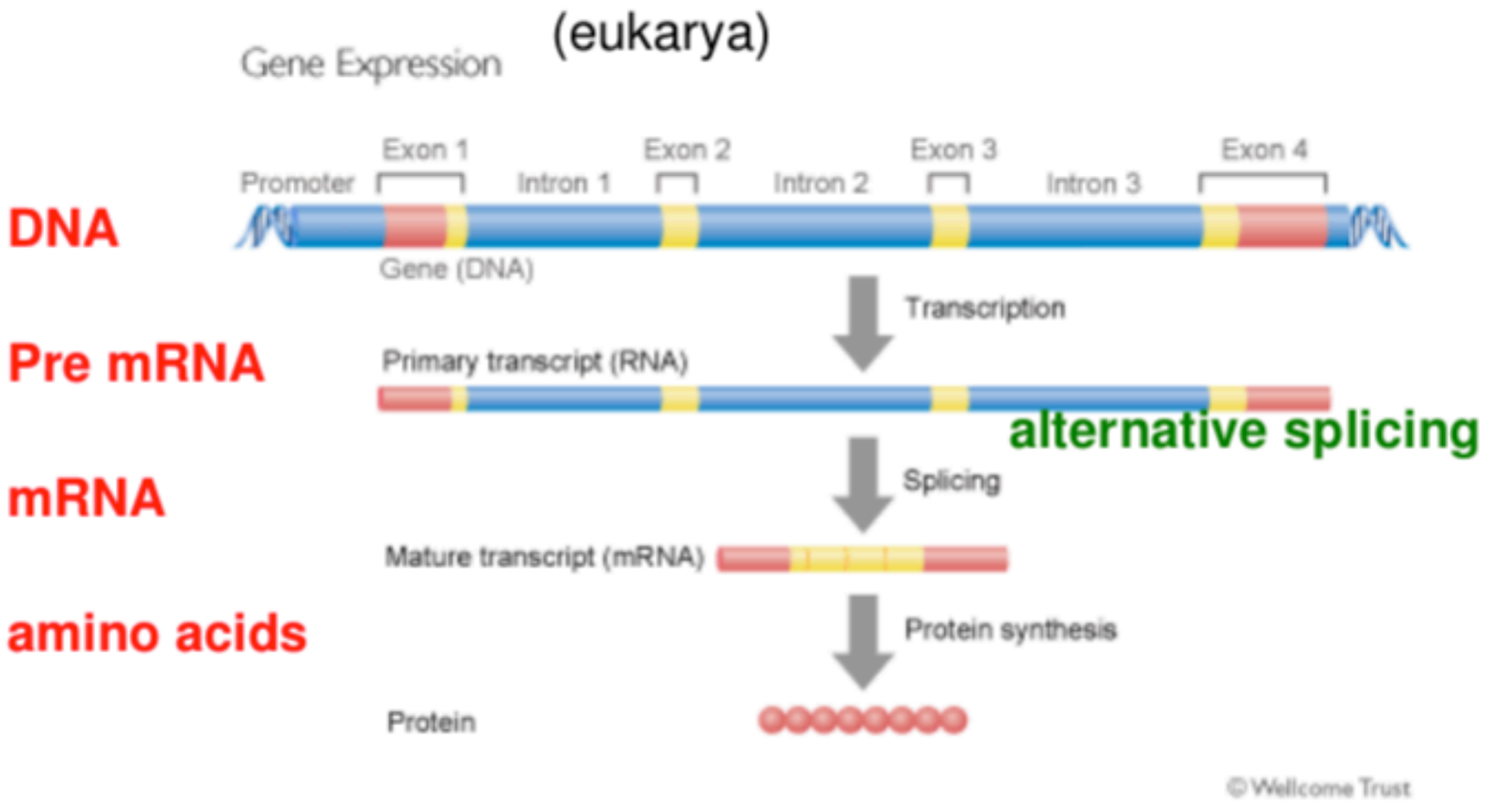
The evolving definition of a gene

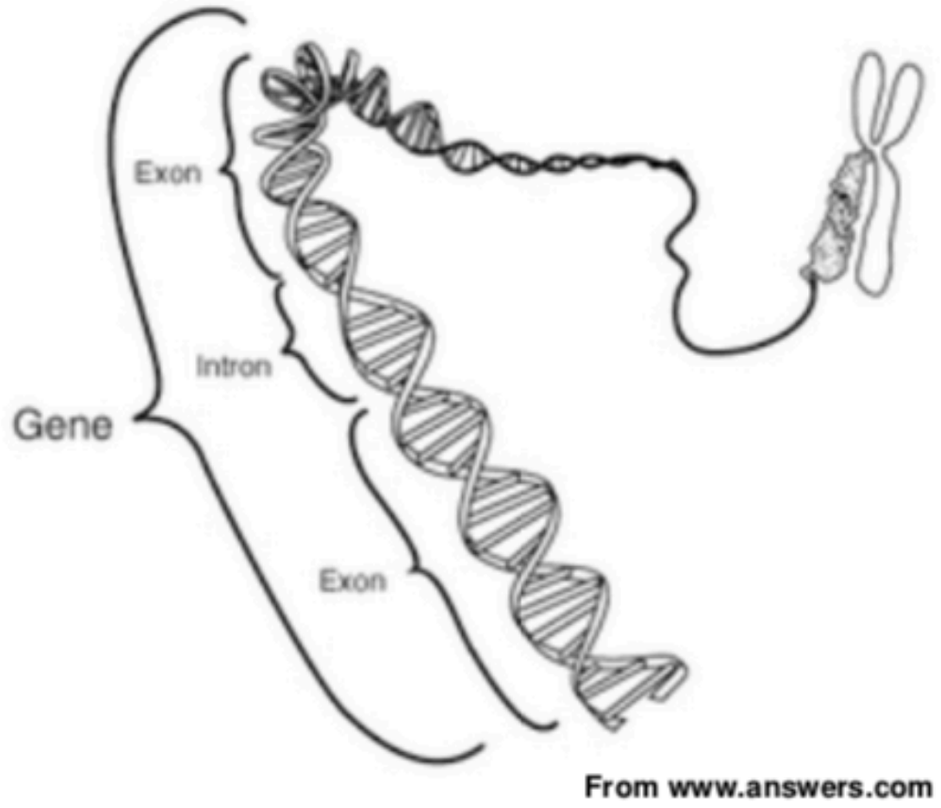
Professor Gerald Fink, a pioneer in the field of genetics, delivers the annual Killian Lecture.

MIT News Office
April 5, 2019

"a gene is any segment of DNA that is transcribed into RNA that has some function"

It's More Complex:
Alternative Splicing, 3-D structure, etc.





From www.answers.com

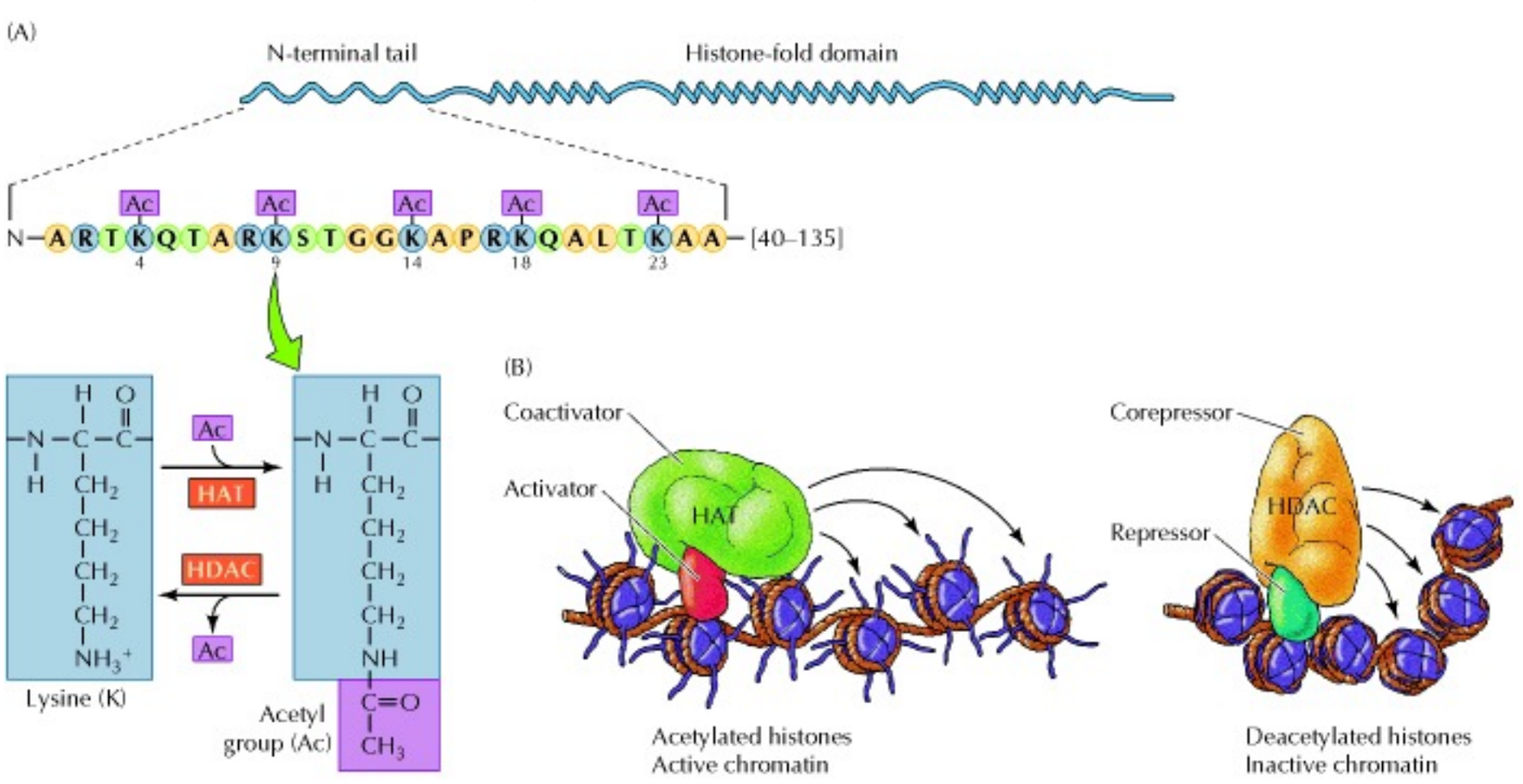
Phenotype = f(Genotype, Environment)

Exceptions to oversimplified Central Dogma

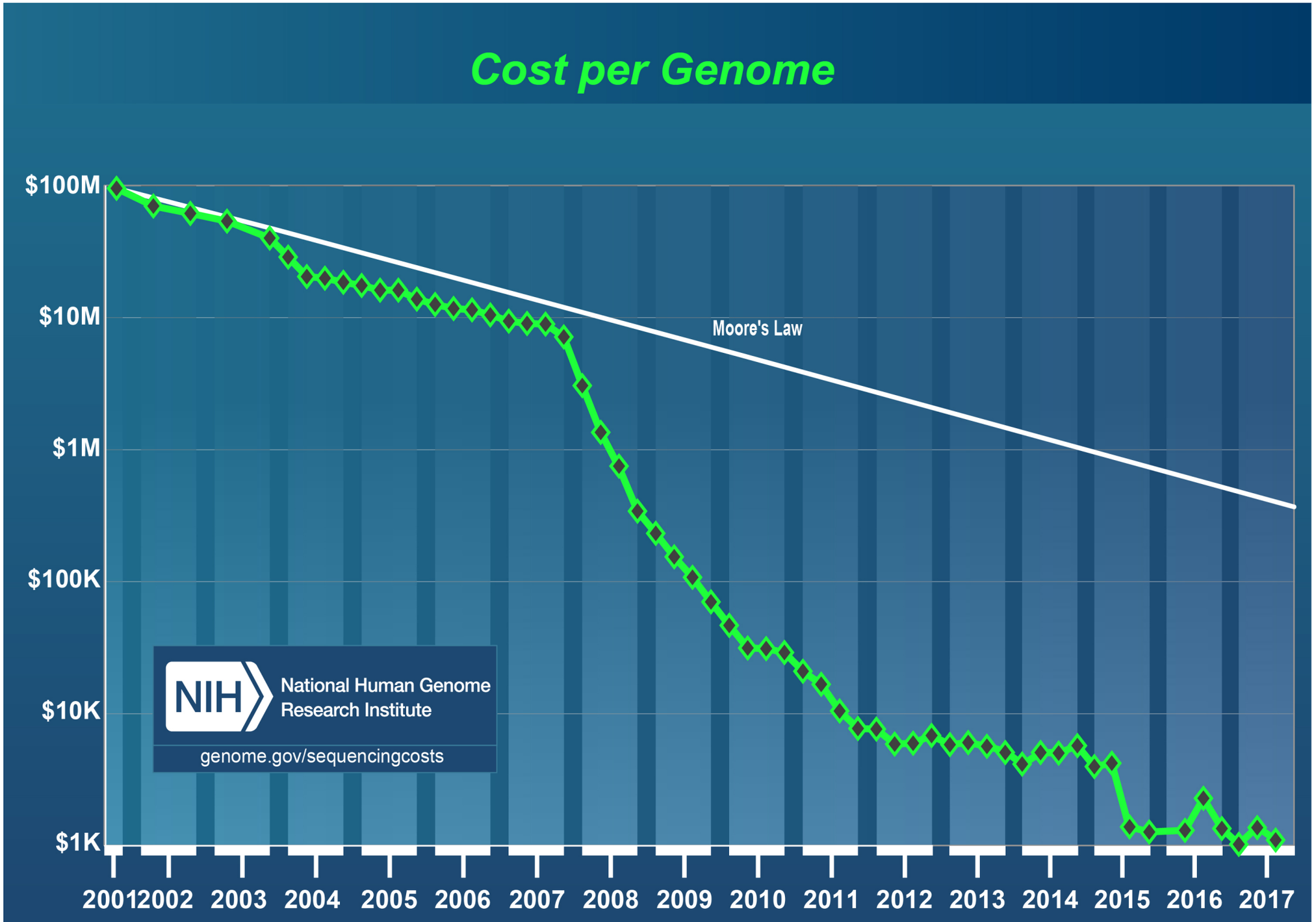
- Retroviruses
  - RNA into DNA via reverse transcriptase: E.g., avian sarcoma/leukosis viruses, mouse leukemia viruses, human immunodeficiency virus (HIV)
  - RNA (virus) > DNA (host) > RNA (virus) > Protein (virus)
- Primitive RNA viruses
  - Error-prone RNA replication. E.g., hepatitis B, rabies, Dengue, Ebola, flu
  - Genetic RNA > Intermediate RNA > Protein
- Prions
  - Self-replicating proteins. E.g., Creuzfeldt-Jakob, “mad cow”, kuru
  - Protein > Protein
- DNA-modifying proteins
  - DNA-repair proteins: MCM (Minichromosome maintenance) family
  - CRISPR-CAS9
- Retrotransposons
  - Mobile DNA (genetic) segments in eukarya. Esp. plants, >90% wheat genome.
  - Retrotransposon DNA > RNA > DNA

More Complexity

- Long non-coding RNA (lncRNA) participate in gene regulation
- RNA Interference (miRNA) prevent post-transcriptional function of mRNA
- Protein degradation mechanisms alter post-translational population of proteins
- Chromatin packages DNA into compact forms, but accessible to transcription only by histone modifications
- Methylation, genomic imprinting



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/

Early Efforts to Characterize Disease Subtypes using Gene Expression Microarrays

- mRNA -> cDNA
- Amplification
- Mark with red fluorescent dye
- Flow over microarray with thousands of spots/wells holding complementary single-stranded DNA fragments, which are distinct parts of genes
- Measure fluorescence at each spot to determine expression level of each gene
- Alternative: Mark “normal” tissue with green fluorescence, flow both over microarray, and measure ratio of red to green at each spot
- Cluster samples by nearness in gene expression space, genes by expression similarity across samples (bi-clustering)

Journal of Pathology
J Pathol 2001; **195**: 41–52.
DOI: 10.1002/path.889

Review Article

Towards a novel classification of human malignancies based on gene expression patterns

Ash A. Alizadeh[1], Douglas T. Ross[2], Charles M. Perou[3] and Matt van de Rijn[4]*

Typical Expression Microarray Experiment

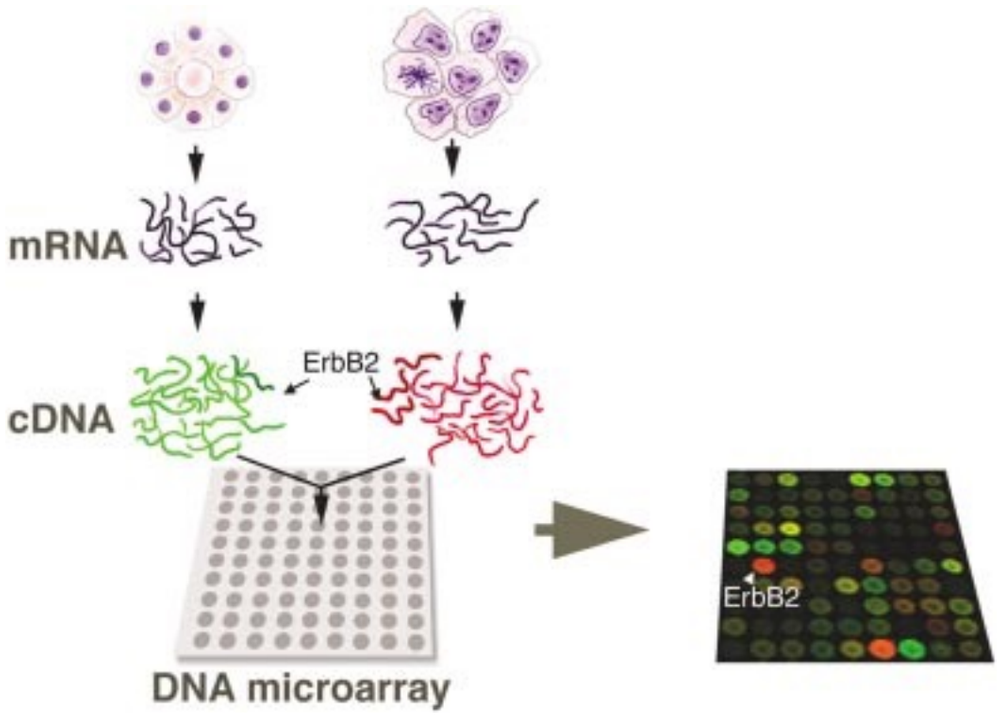


Figure1. Schematic representation of a DNA microarray hybridization comparing gene expression of a malignant epithelial cancer with its normal tissue counterpart

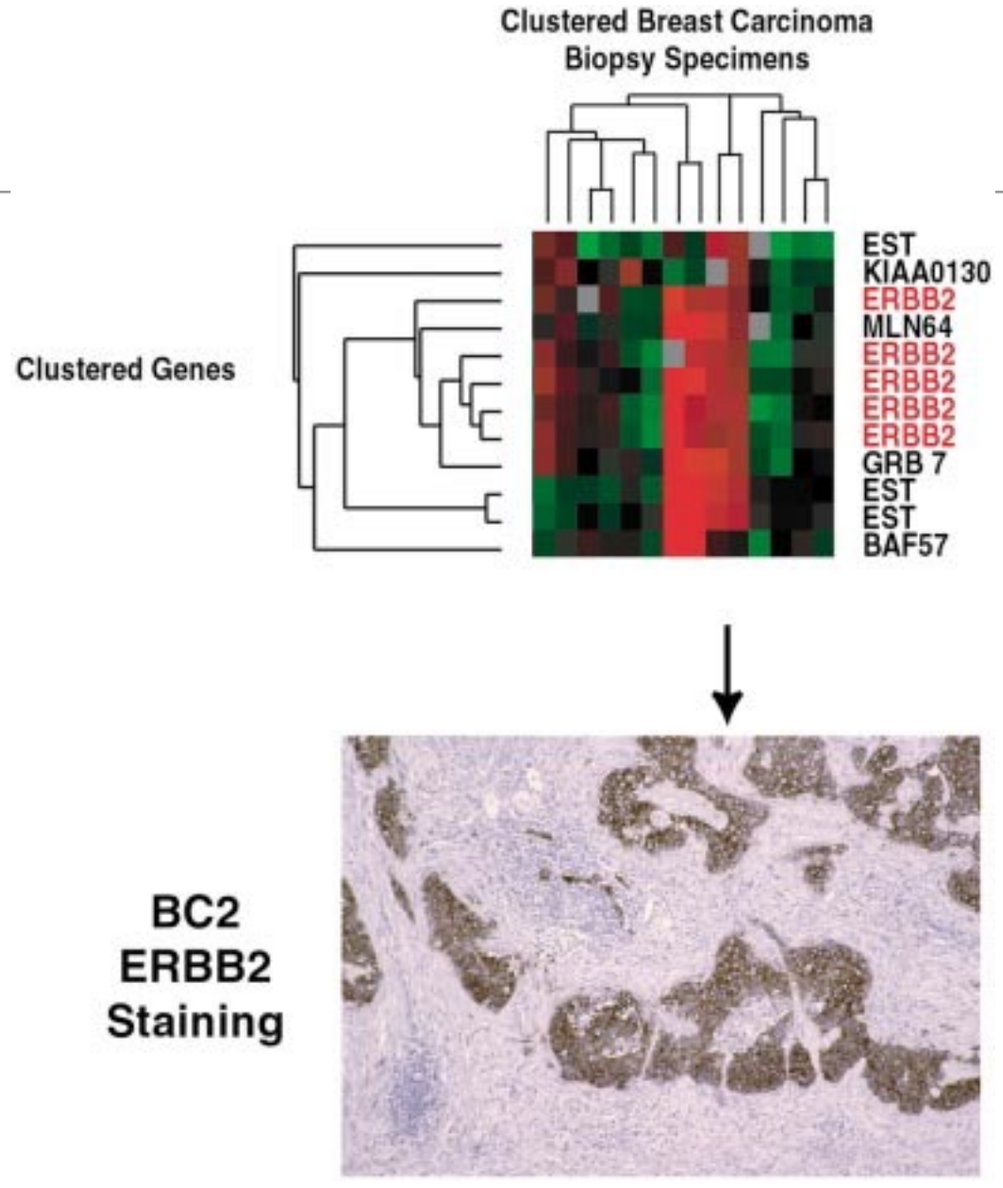


Figure 2. Example of data clustering. This small sample of array data was copied from a much larger data set, similar to the one shown in Figure 3. Note how all five different cDNA clones specific for ERBB2 on the array cluster tightly together. The immunostaining for ERBB2 on one of the breast samples (column indicated by an arrow) is shown in the lower panel

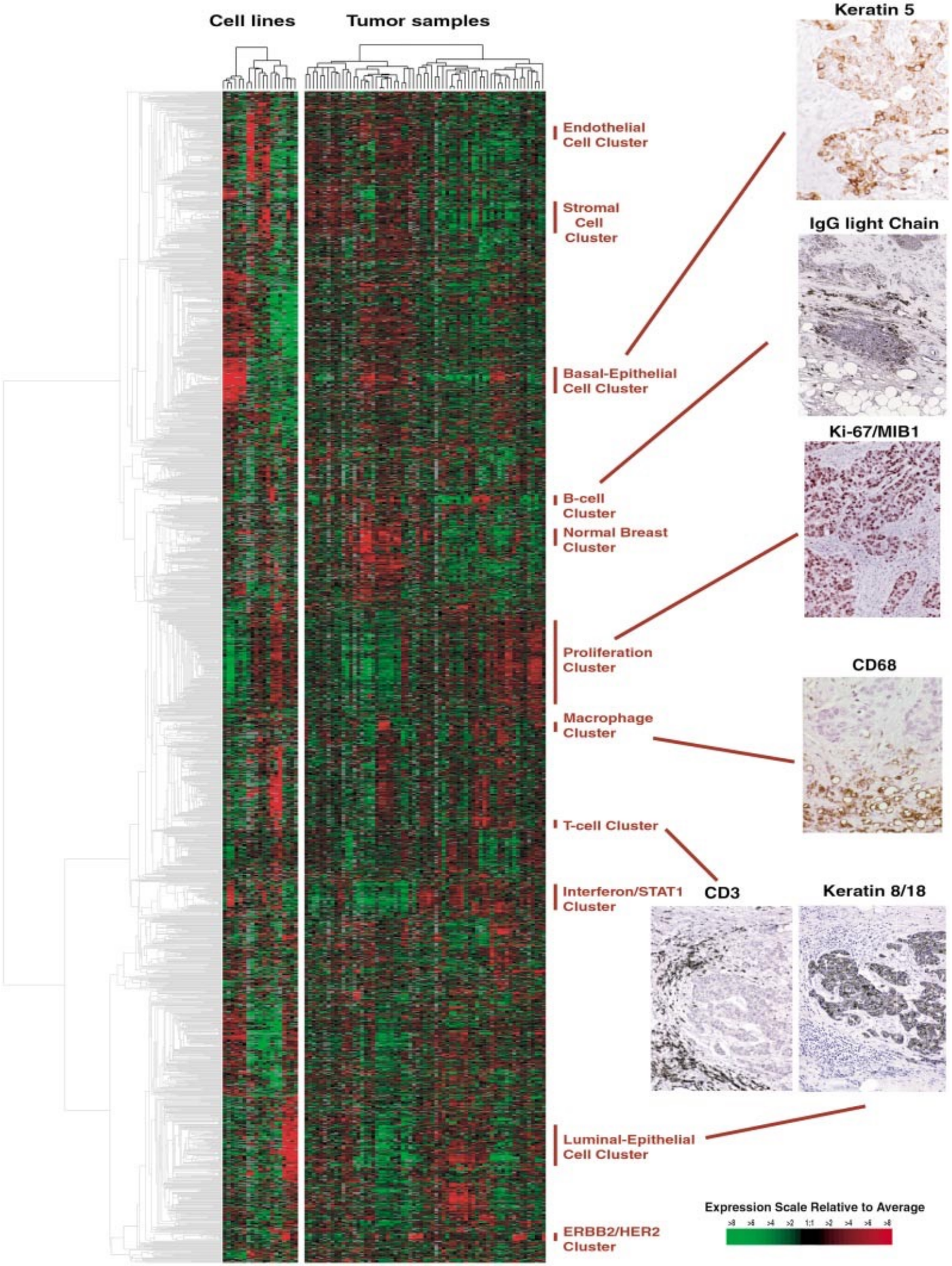


Figure 3. Cluster analysis of 19 cell lines and 65 breast tumour samples showing how different host cell populations can be identified in the tumour samples

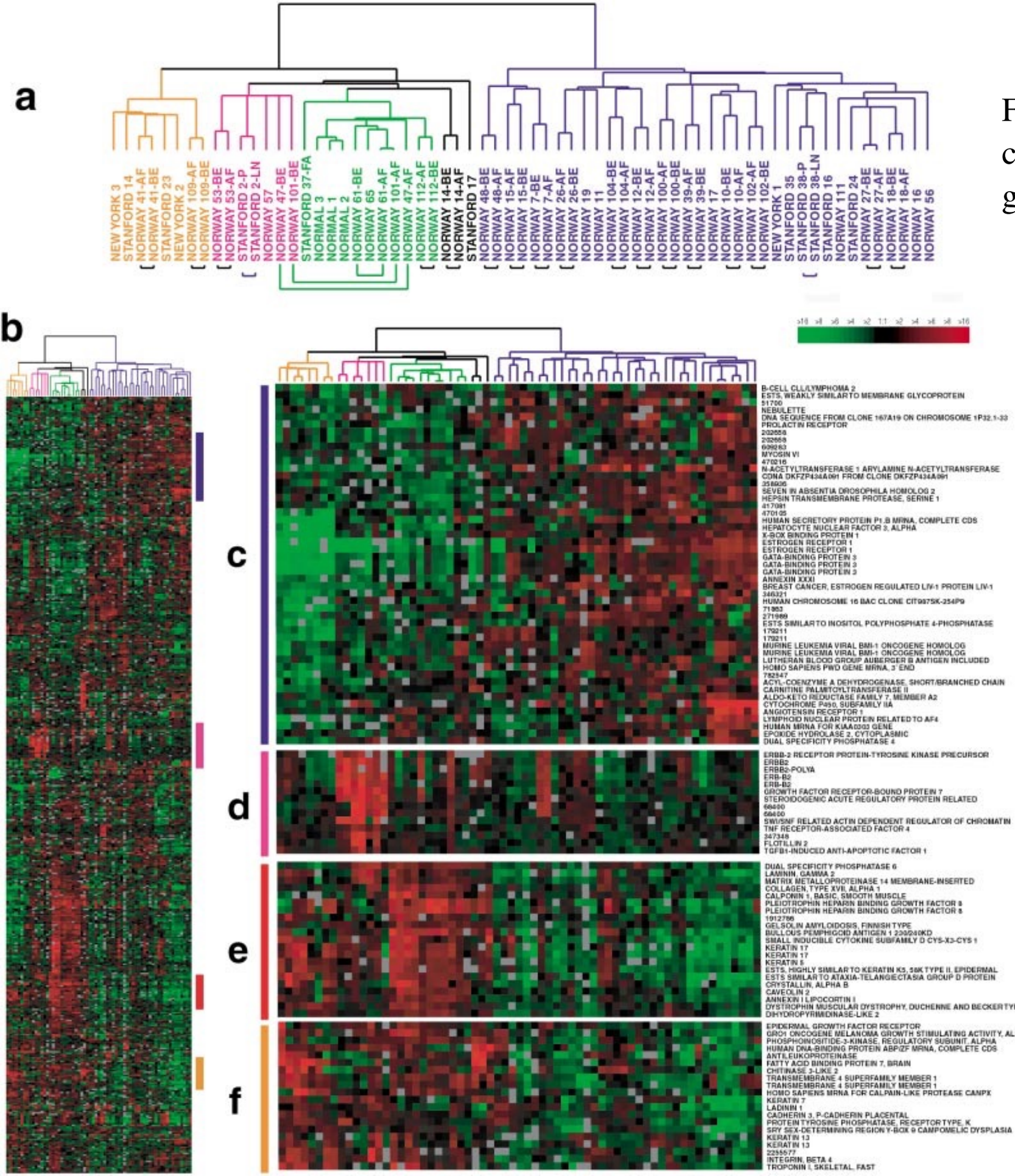


Figure 5. Cluster analysis on 65 breast carcinoma samples, using the 'intrinsic' gene list

[T]he branching pattern of the dendrogram clustered with this 'intrinsic' gene list identified four major groups of breast tumours:

- (c) luminal-epithelial/ER+
- (d) ERBB2 and other associated genes
- (e) normal breast
- (f) high-level expression of two clusters of genes that are characteristic of normal breast basal epithelial cells

… found to be statistically significantly associated with differences in overall patient survival and relapse-free survival

Survival of Different Subgroups of Breast Cancer Patients

- from a similar (later) analysis of a different breast cancer cohort, they identified five subgroups

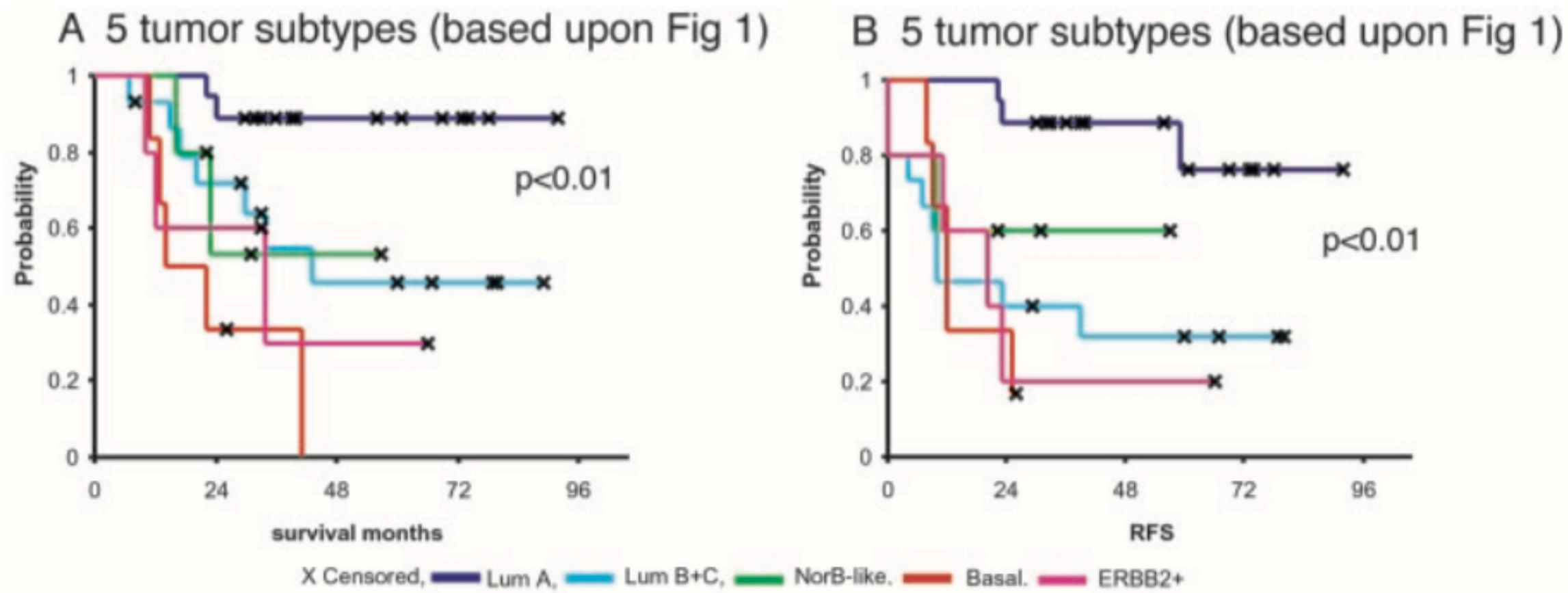


Fig. 3. Overall and relapse-free survival analysis of the 49 breast cancer patients, uniformly treated in a prospective study

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS, 98(19), 10869–10874. http://doi.org/10.1073/pnas.191367098

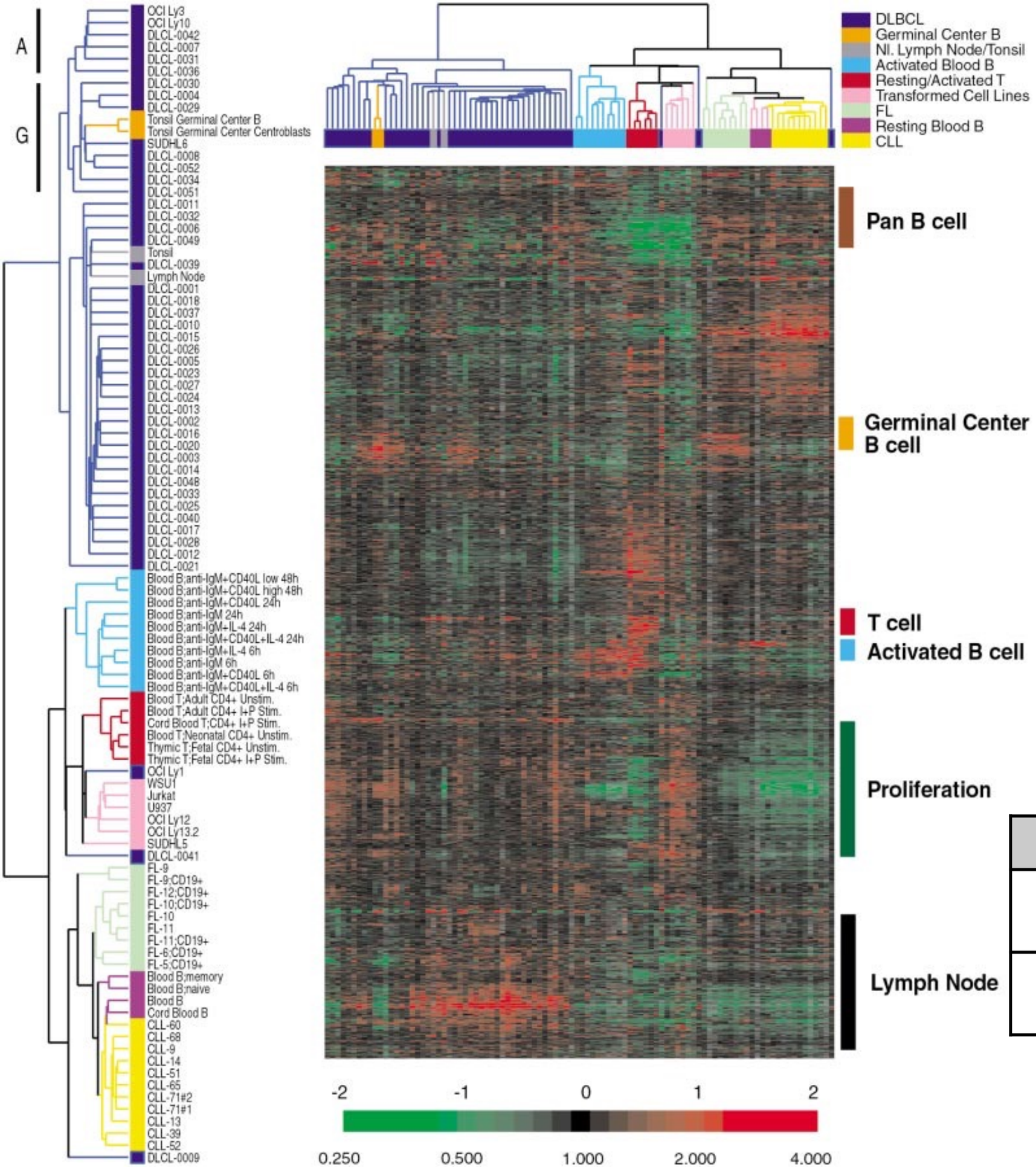
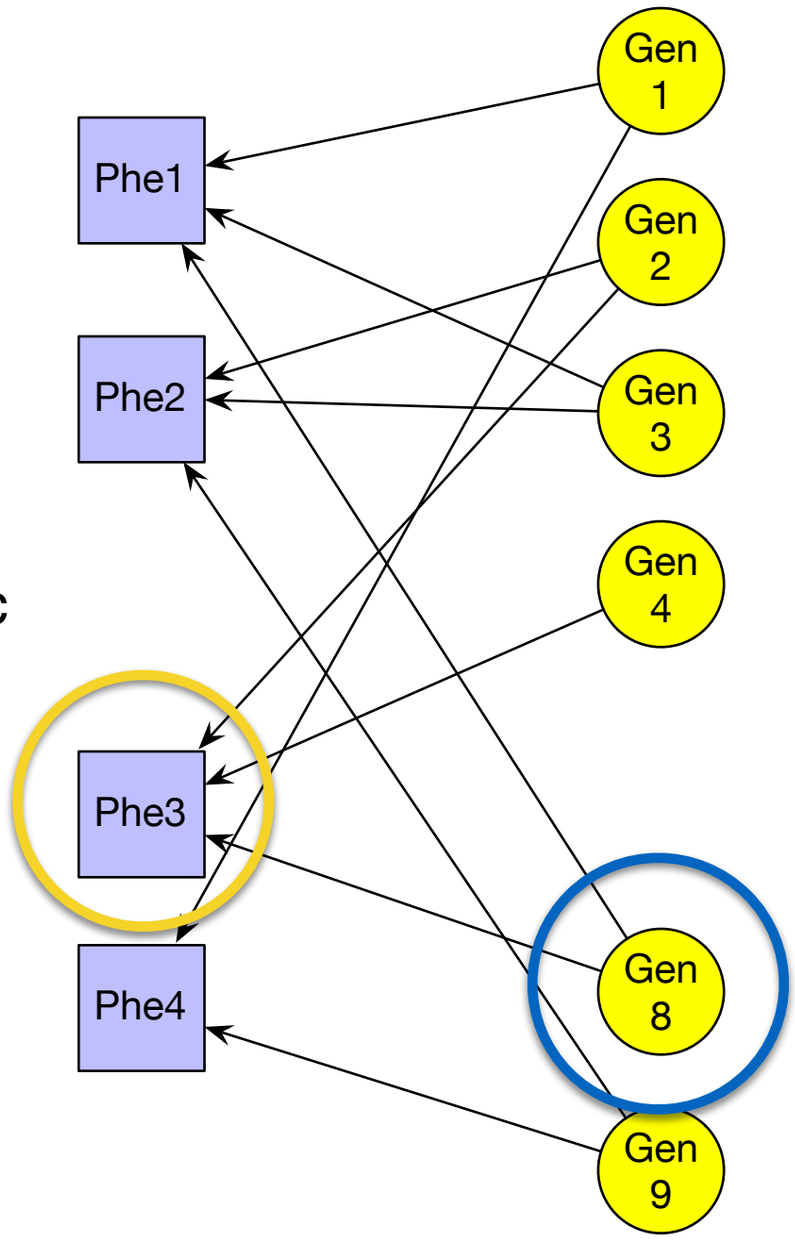


Figure 6. Hierarchical clustering of gene expression data depicting relationships between 96 samples of normal and malignant lymphocytes [19]. The dendrogram on the left lists the samples studied and provides a measure of the relatedness of gene expression in each sample. The dendrogram is colour-coded according to the category of mRNA sample studied (see upper right key)

sub-clusters of DLBCL	5yr survival
germinal centre B-like	76%
activated B-like	16%

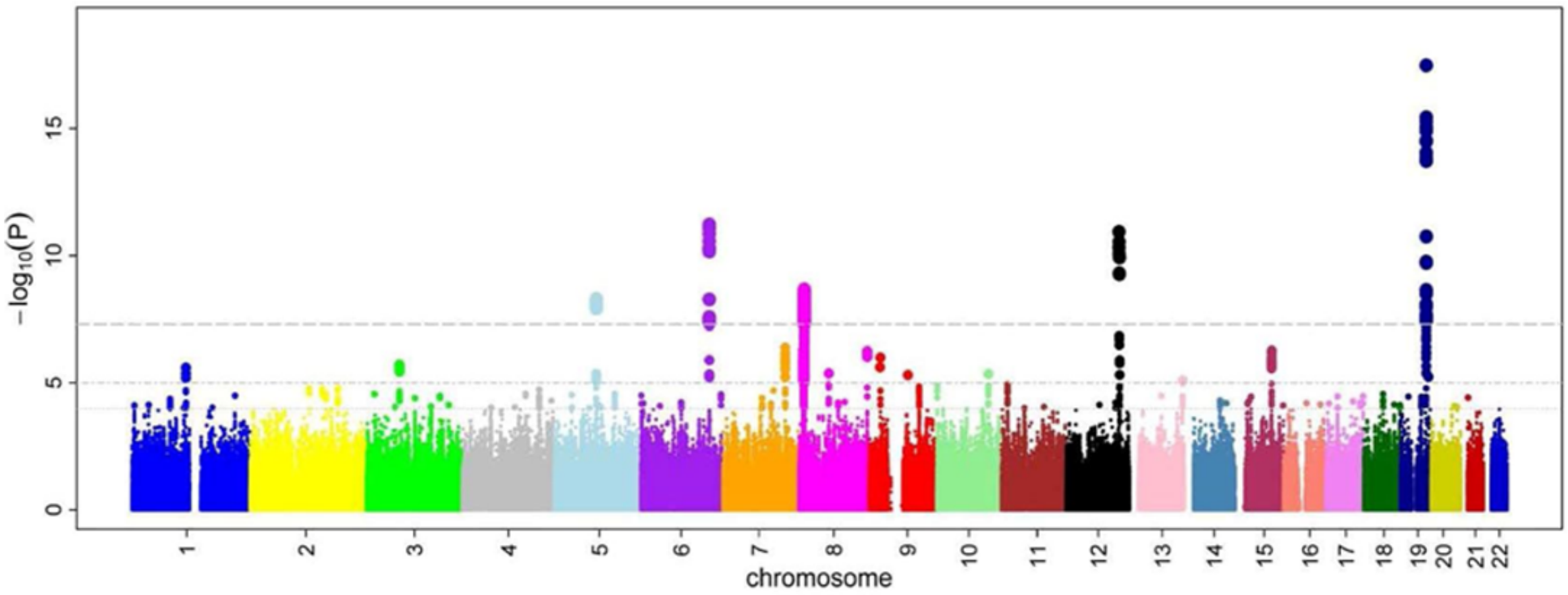
Relationships between Genotype and Phenotype

- 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.

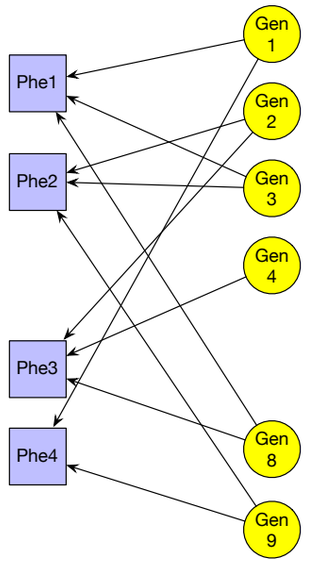


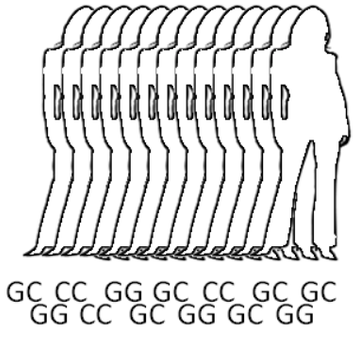
An illustration of a Manhattan plot depicting several strongly associated risk loci. Each dot represents a SNP, with the X-axis showing genomic location and Y-axis showing association level. This example is taken from a GWA study investigating microcirculation, so the tops indicates genetic variants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

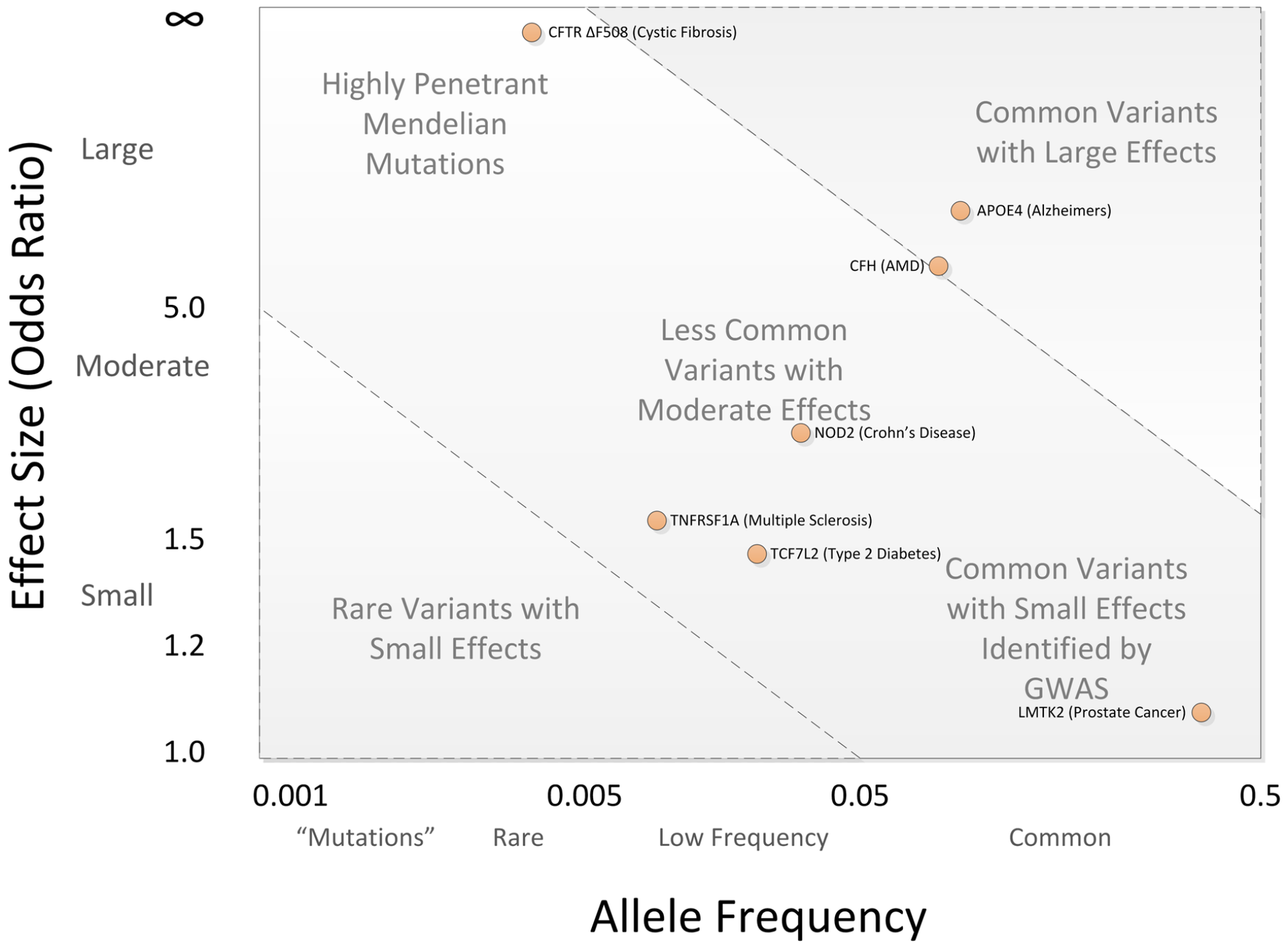




	SNP1	*SNP2*	*SNP…*
	Cases	**Cases**	*Repeat for all SNPs*
	Count of G: 2104 of 4000	Count of G: 1648 of 4000	
	Frequency of G: 52.6%	Frequency of G: 41.2%	
	Controls	**Controls**	
	Count of G: 2676 of 6000	Count of G: 2532 of 6000	
	Frequency of G: 44.6%	Frequency of G: 42.2%	
	P-value: $5.0 \cdot 10^{-15}$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

Bayesian Non-Negative Matrix Factorization (bNMF)

$$\boldsymbol{X} = \boldsymbol{AB} + \boldsymbol{E}$$

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i,j} \mathcal{N}(X_{i,j}; (\boldsymbol{AB}_{i,j}), \sigma^2)$$

where $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{B}, \sigma^2\}$ are all parameters of the model, and

$$\mathcal{N}(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp(-(x-\mu)^2/(2\sigma^2))$$

Assume $\boldsymbol{A}$ and $\boldsymbol{B}$ are independently exponentially distributed, with scales $\alpha_{i,n}$ and $\beta_{n,j}$. Then

$$p(\boldsymbol{A}) = \prod_{i,n} \alpha_{i,n} \exp(-\alpha_{i,n} A_{i,n}) u(A_{i,n})$$

and

$$p(\boldsymbol{B}) = \prod_{n,j} \beta_{n,j} \exp(-\beta_{n,j} B_{n,j}) u(B_{n,j})$$

where $u(x)$ is the unit step function.
The prior for the noise term is chosen as an inverse gamma density with shape k and scale θ:

$$p(\sigma^2) = \mathcal{G}^{-1}(\sigma^2; k, \theta) = \frac{\theta^k}{\Gamma(k)} (\sigma^2)^{-k-1} \exp(-\frac{\theta}{\sigma^2})$$

Schmidt, M. N., Winther, O., & Hansen, L. K. (2009). Bayesian non-negative matrix factorization. Presented at the Independent Component Analysis and Signal Separation.

T2D GWAS

- Association matrix (47x94) of traits x variants
  - traits doubled: one set inverted where z-score was negative, the other positive
    - maintains non-negativity of matrix
- NNMF to factor X ~ WH
  - W is (47 x K), H^T is (94 x K), K optimized by bNMF:
    - maximizing $p(\boldsymbol{X})$ for different K lets this technique estimate the right number of factors
    - loss function is $||X-WH||^2$ + L_2(W and H, coupled by relevance weights)
- MCMC: Gibbs sampling + tricks to compute estimates of $p(\boldsymbol{X})$

- Data about 17K people from four different studies, all “European ancestry”
  - Metabolic Syndrome in Men Study; Diabetes Genes in Founder Populations (Ashkenazi) study; The Partners Biobank; The UK Biobank
  - Individual-level analyses of individuals with T2D from all four data sets

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

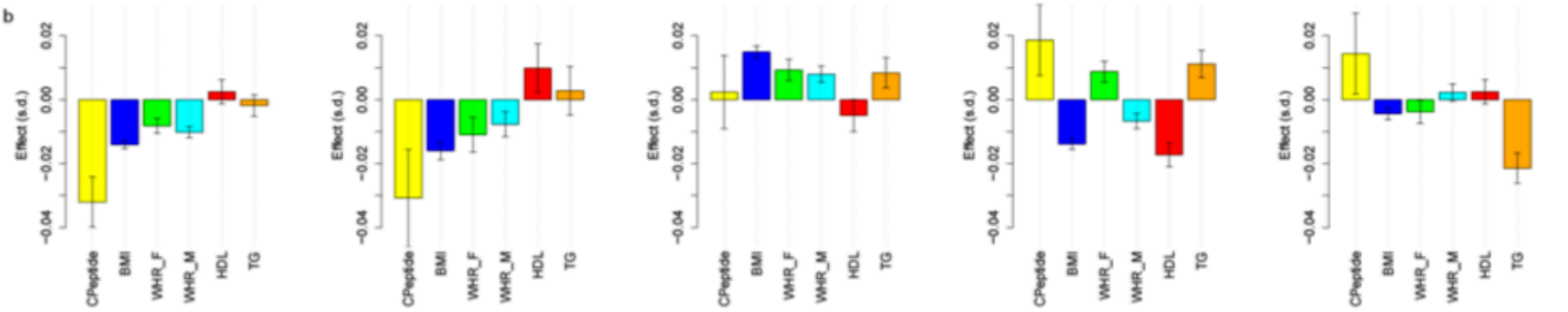
Results

- 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

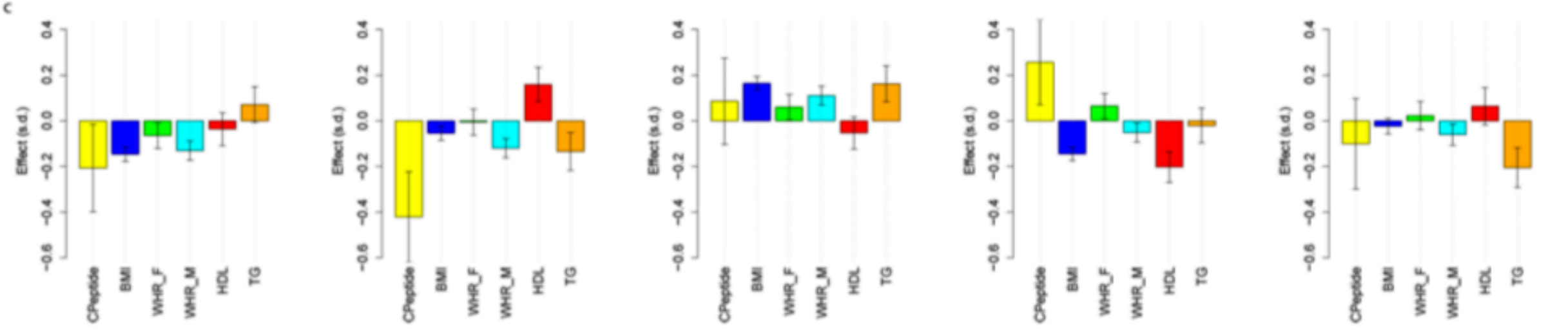

Effect (s.d.)
0.4
0.2
0.0
−0.2
−0.4
−0.6
CPeptide
BMI
WHR_F
WHR_M
HDL
TG

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.

Table 1. Associations of cluster genetic risk scores and selected GWAS traits.

	Beta Cell *N* Loci = 30		Proinsulin *N* Loci = 7		Obesity *N* Loci = 5		Lipodystrophy *N* Loci = 20		Liver/Lipid *N* Loci = 5	
Trait	**beta**	***P*-value**	**beta**	***P*-value**	**beta**	***P*-value**	**beta**	***P*-value**	**beta**	***P*-value**
Adiponectin	−0.0005	0.55	−0.0019	0.37	−0.0007	0.74	**−0.0114**	**$3.34E^{-23}$**	−0.0007	0.77
BMI	**−0.0026**	**6.0×10^{-5}**	**−0.0080**	**3.1×10^{-8}**	**0.0396**	**9.7×10^{-157}**	**−0.0079**	**$1.81E^{-21}$**	0.0001	0.94
Bodyfat	−0.0016	0.11	−0.0061	4.5×10^{-3}	**0.0247**	**2.1×10^{-25}**	**−0.0120**	**$9.04E^{-22}$**	−0.0031	0.26
CIR	**−0.0584**	**7.1×10^{-43}**	−0.0234	0.014	−0.0010	0.92	0.0087	0.10	−0.0021	0.85
DI	**−0.0543**	**6.6×10^{-37}**	−0.0080	0.40	−0.0086	0.40	−0.0102	0.05	−0.0115	0.30
2hrGlu adj BMI	**0.0288**	**2.0×10^{-13}**	0.0204	0.02	0.0064	0.49	**0.0292**	**$2.26E^{-9}$**	−0.0257	0.01
FI	**−0.0033**	**4.4×10^{-7}**	−0.0054	2.2×10^{-4}	**0.0087**	**6.1×10^{-8}**	**0.0068**	**$1.96E^{-16}$**	**0.0071**	**3.8×10^{-5}**
FI adj BMI	**−0.0026**	**4.4×10^{-6}**	−0.0040	1.3×10^{-3}	−0.0008	0.57	**0.0082**	**$3.01E^{-31}$**	**0.0082**	**2.1×10^{-8}**
HDL	−0.0008	0.51	−0.0031	0.27	−0.0059	0.05	**−0.0191**	**$1.96E^{-33}$**	0.0069	0.038
Height	0.0009	0.12	**−0.0058**	**3.3×10^{-5}**	**−0.0033**	**1.9×10^{-5}**	**0.0061**	**$4.44E^{-05}$**	−0.0005	0.77
HC	**−0.0031**	**3.5×10^{-5}**	**−0.0113**	**1.0×10^{-11}**	**0.0345**	**9.1×10^{-79}**	**−0.0116**	**$9.69E^{-34}$**	−0.0007	0.73
HOMA-B	**−0.0066**	**1.9×10^{-21}**	**−0.0103**	**2.6×10^{-11}**	**0.0066**	**8.0×10^{-5}**	0.0019	0.03	0.0019	0.30
HOMA-IR	−0.0011	0.21	−0.0041	0.03	**0.0108**	**9.0×10^{-8}**	**0.0066**	**2.2×10^{-10}**	**0.0093**	**2.6×10^{-5}**
Incr30	**−0.0398**	**6.9×10^{-19}**	−0.0239	0.02	−0.0053	0.63	0.0198	4.8×10^{-4}	0.0102	0.38
Ins30 adj BMI	**−0.0503**	**1.8×10^{-28}**	−0.0310	1.8×10^{-3}	0.0027	0.81	0.0163	3.9×10^{-3}	0.0054	0.64
ISI adj BMI	−0.0039	0.06	−0.0020	0.67	0.0045	0.37	**−0.0213**	**1.3×10^{-13}**	−0.0086	0.12
Leptin	0.0009	0.50	−0.0067	0.03	**0.0197**	**1.0×10^{-9}**	**−0.0245**	**8.9×10^{-21}**	**0.0147**	**3.1×10^{-5}**
Linoleic acid	0.0093	0.29	−0.0232	0.25	0.0027	0.90	−0.0024	0.83	**0.1330**	**$1.31\text{x}10^{-8}$**
Palmitoleic	0.0002	0.74	0.0024	0.11	0.0034	0.03	−0.0020	0.02	**−0.0104**	**$5.50\text{x}10^{-9}$**
Proinsulin	**0.0097**	**1.2×10^{-10}**	−0.0297	**1.4×10^{-18}**	0.0047	0.18	0.0059	1.3×10^{-3}	0.0059	0.13
Total Chol	0.0023	0.06	−0.0055	0.04	−0.0023	0.45	0.0046	3.2×10^{-3}	**−0.0182**	**3.1×10^{-8}**
Triglycerides	0.0022	0.07	−0.0027	0.33	0.0066	0.03	**0.0194**	**1.8×10^{-34}**	**−0.0416**	**1.0×10^{-35}**
Urate	−0.0007	0.51	−0.0045	0.084	**0.0165**	**1.4×10^{-9}**	**0.0090**	**2.2×10^{-10}**	**−0.0260**	**2.6×10^{-18}**
WC	−0.0020	5.23×10^{-3}	**−0.0096**	**1.5×10^{-9}**	**0.0379**	**1.0×10^{-102}**	**−0.0058**	**3.6×10^{-10}**	−0.0005	0.80
WC female	−0.0010	0.30	−0.0073	3.9×10^{-4}	**0.0376**	**1.4×10^{-60}**	−0.0022	0.07	0.0000	0.99
WC male	−0.0031	1.6×10^{-3}	**−0.0128**	**5.4×10^{-9}**	**0.0374**	**1.1×10^{-51}**	**−0.0102**	**2.8×10^{-15}**	0.0007	0.80
WHR	0.0014	0.05	−0.0016	0.30	**0.0229**	**3.7×10^{-39}**	**0.0051**	**1.6×10^{-8}**	0.0016	0.43
WHR female	0.0027	4.0×10^{-3}	0.0010	0.62	**0.0221**	**2.8×10^{-22}**	**0.0140**	**5.6×10^{-32}**	0.0027	0.31
WHR male	0.0003	0.74	−0.0049	0.03	**0.0242**	**1.4×10^{-21}**	**−0.0059**	**7.6×10^{-6}**	0.0003	0.92

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

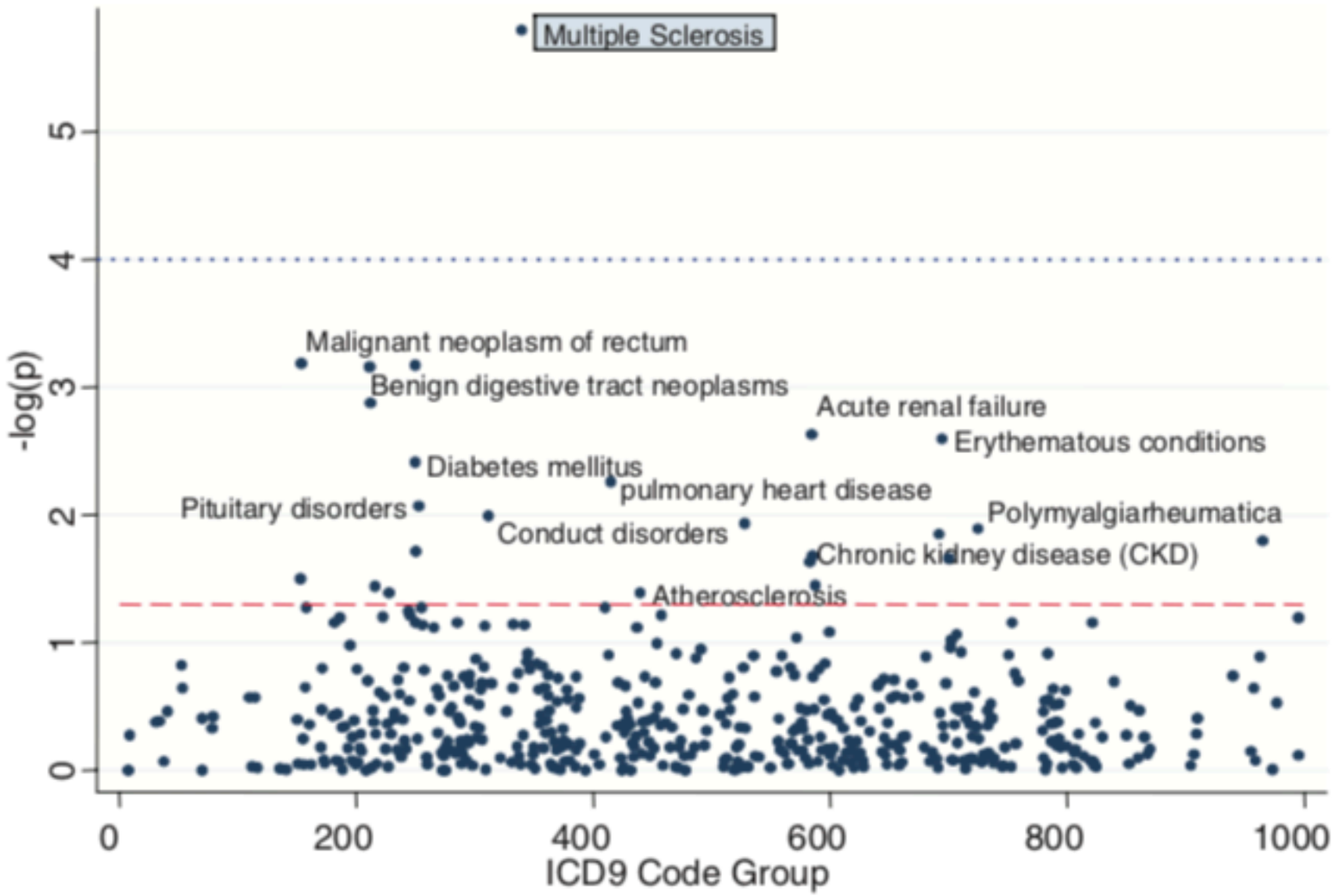


Fig. 1. Phenome-wide scan for association with rs3135388. MS is replicated from prior analyses. The dashed line represents the $P = 0.05$; the dotted line represents the Bonferroni correction.

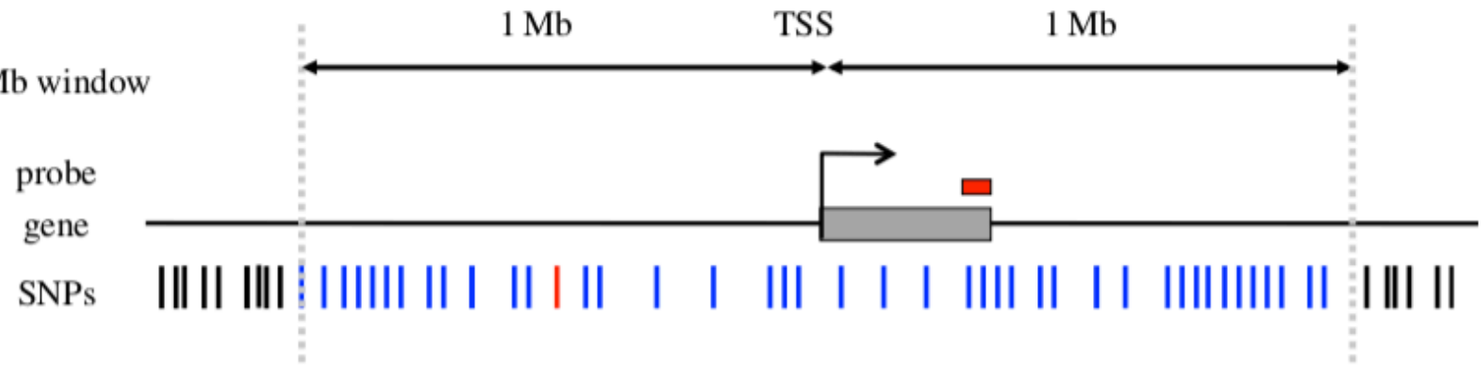
You Don’t Always Get What You Expect

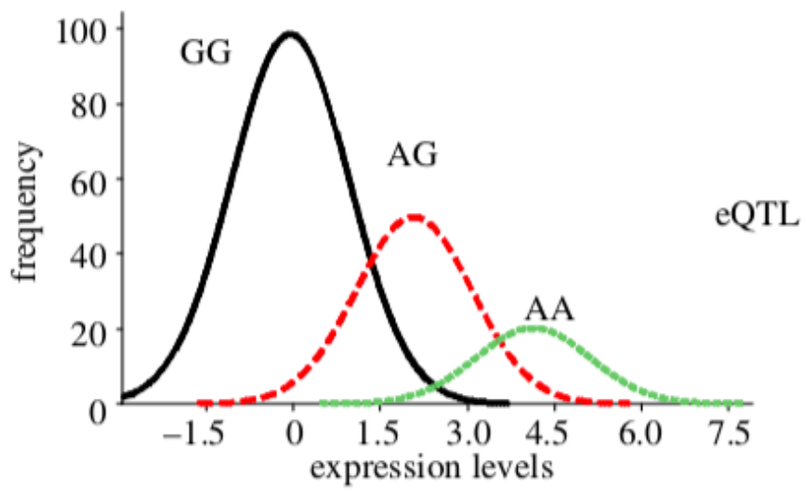
Table 2. Diseases previously associated with the five SNP studied and current PheWAS ORs

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





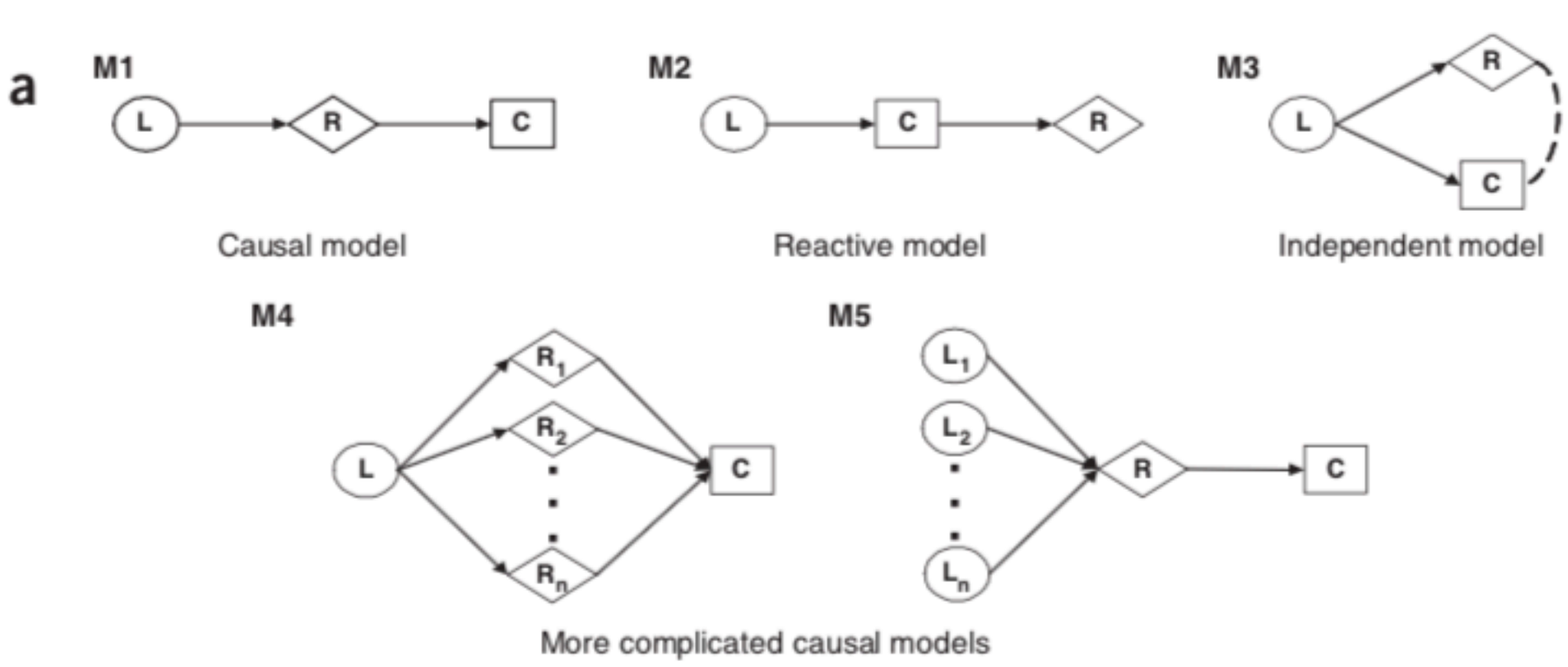
Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1620), 20120362–20120362. http://doi.org/10.1098/rstb.2012.0362

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."

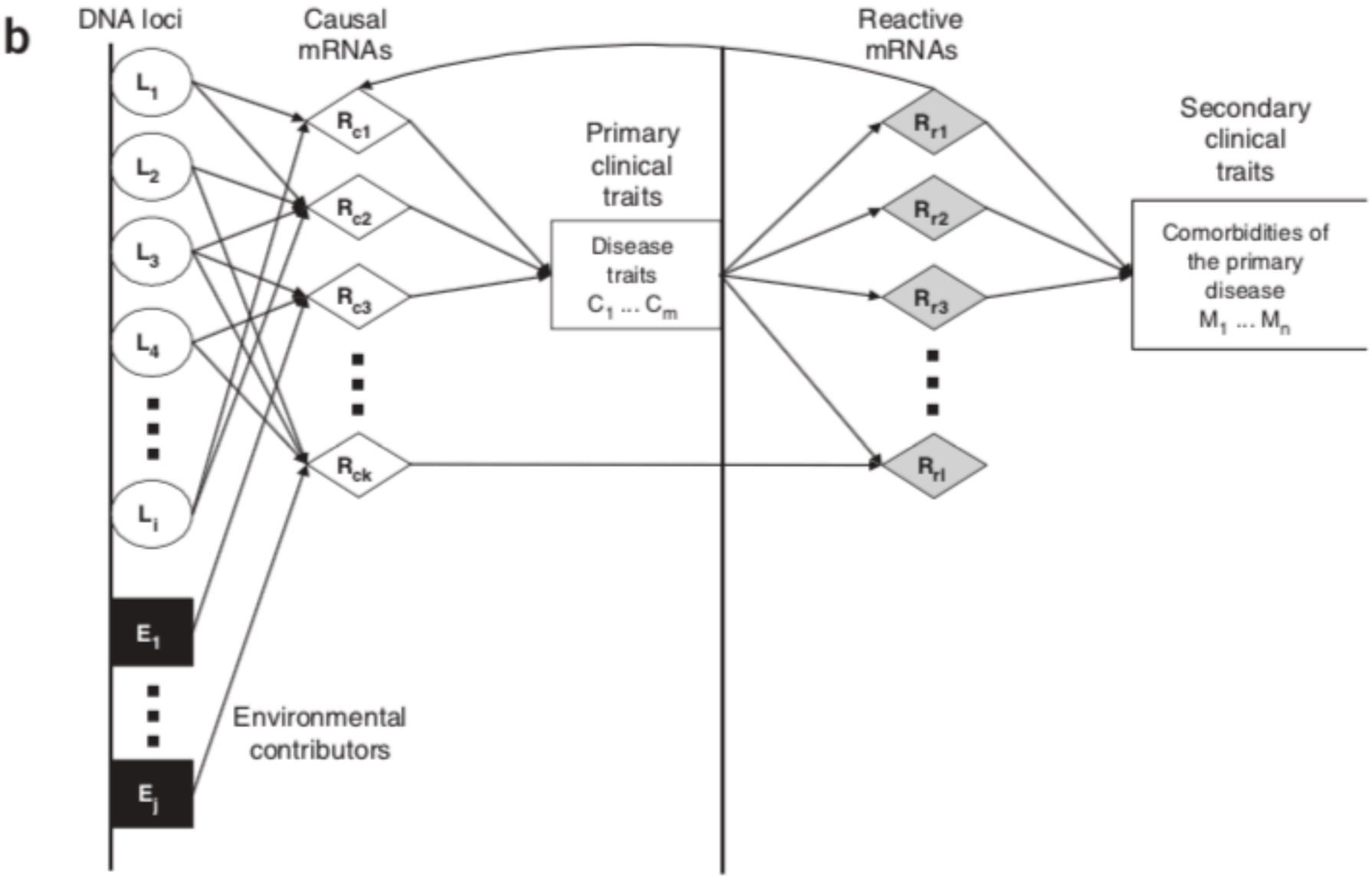
QTL, eQTL, & Disease Traits

- 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

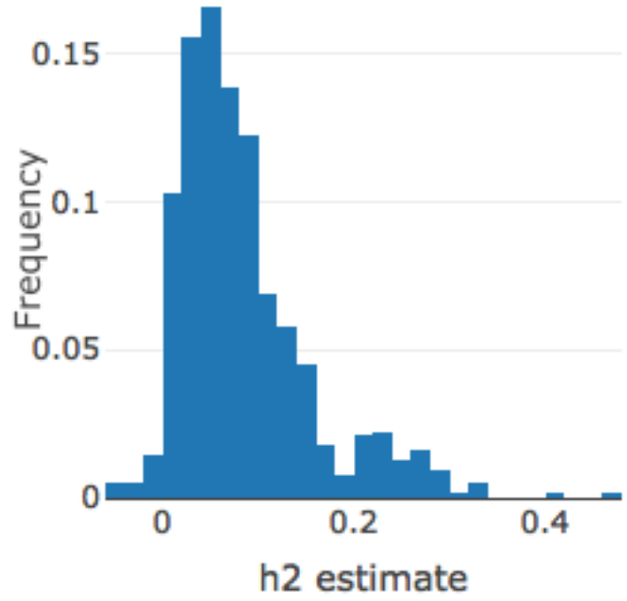
- UK Biobank collects data on ~.5M de-identified individuals
  - everyone will have full exome sequencing (50K so far)
  - 100K have worn 24-hour activity monitor for a week, 20K have had repeat measurements
  - on-line questionnaires: diet, cognitive function, work history, digestive health
  - 100K 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

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

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

Gene Set Enrichment Analysis (GSEA)

- Problems with genome-wide expression analysis
  - No gene may pass multiple hypothesis testing because of weak signals
  - Many genes may pass, but with no coherent understanding of their relationships
  - Single-gene analyses fail to account for pathway interactions
  - Little overlap among genes identified by multiple studies
- Therefore, consider gene sets (defined by biological knowledge of pathways)
  - Broad published 1,325 biologically defined gene sets (2005) [17,810 today]
  - “genes involved in oxidative phosphorylation [in muscle tissue] show reduced expression in diabetics, although the average decrease per gene is only 20%”

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. http://doi.org/10.1073/pnas.0506580102

GSEA

- Consider L, the list of rank-ordered genes by differential expression between cases and controls; the top and bottom ranked genes are the ones of interest
- Given genes in a set S, are they randomly distributed in L, or concentrated?

- "Random walk" proportional to correlation of gene expression to phenotype
- Statistical significance computed by random permutation test on phenotype; adjust for multiple hypotheses

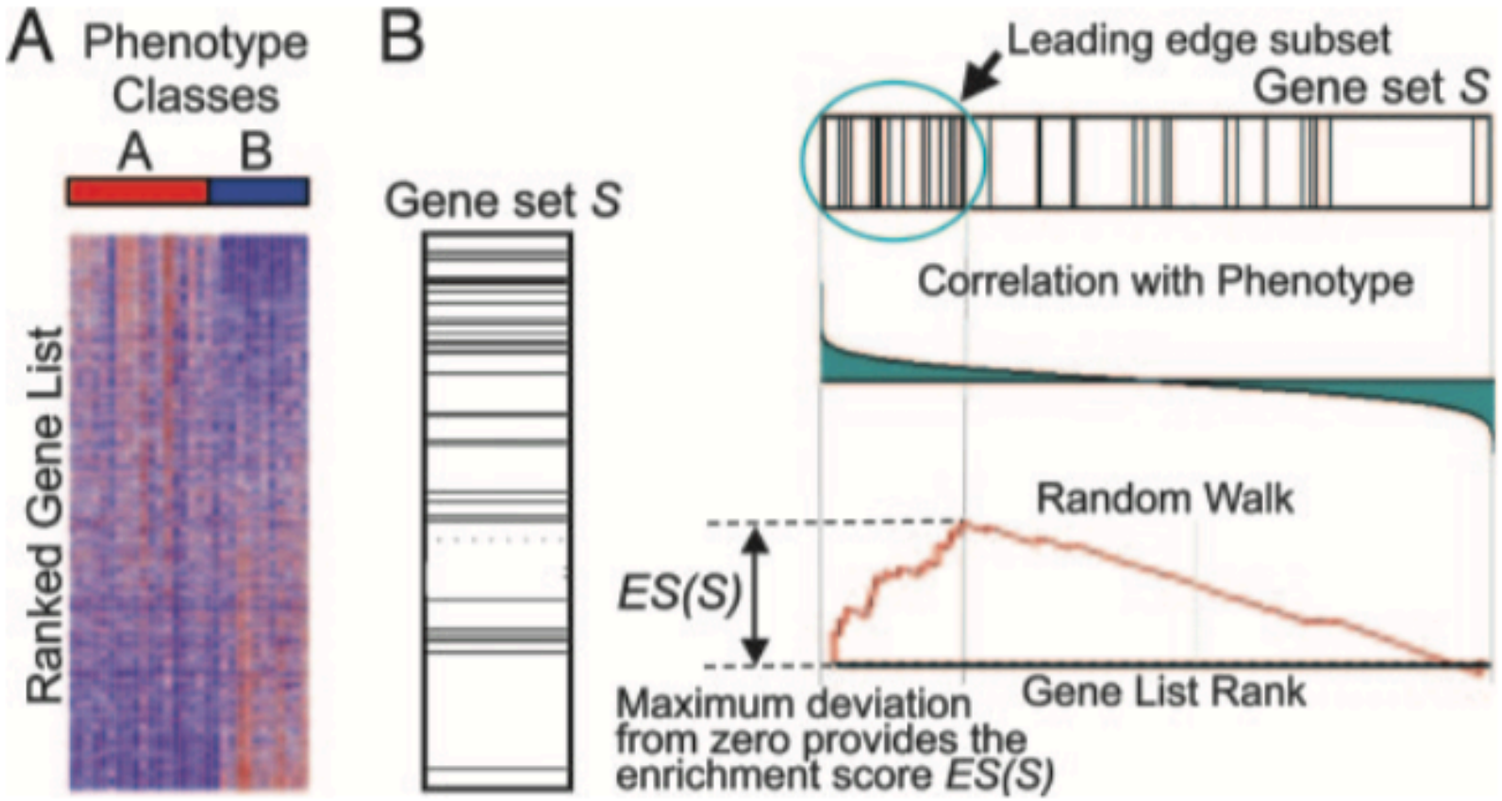


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Early GSEA Successes

Data set: Lymphoblast cell lines	
Enriched in males	
chrY	<0.001
chrYp11	<0.001
chrYq11	<0.001
Testis expressed genes	0.012
Enriched in females	
X inactivation genes	<0.001
Female reproductive tissue expressed genes	0.045
Data set: p53 status in NCI-60 cell lines	
Enriched in p53 mutant	
Ras signaling pathway	0.171
Enriched in p53 wild type	
Hypoxia and p53 in the cardiovascular system	<0.001
Stress induction of HSP regulation	<0.001
p53 signaling pathway	<0.001
p53 up-regulated genes	0.013
Radiation sensitivity genes	0.078
Data set: Acute leukemias	
Enriched in ALL	
chr6q21	0.011
chr5q31	0.046
chr13q14	0.057
chr14q32	0.082
chr17q23	0.071

Data set: Lung cancer outcome, Boston study	
Enriched in poor outcome	
Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216
Data set: Lung cancer outcome, Michigan study	
Enriched in poor outcome	
Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

- Consider pathways, not just gene sets
  - e.g., AND/OR graphs, or circuits

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.