Big data in biology

Data Science Summer School

University of Göttingen

05 August 2019

Johannes Söding

Quantitative and Computational Biology

MPI for Biophysical Chemistry

The summer school is funded by the DAAD with funds of the Federal Foreign Office

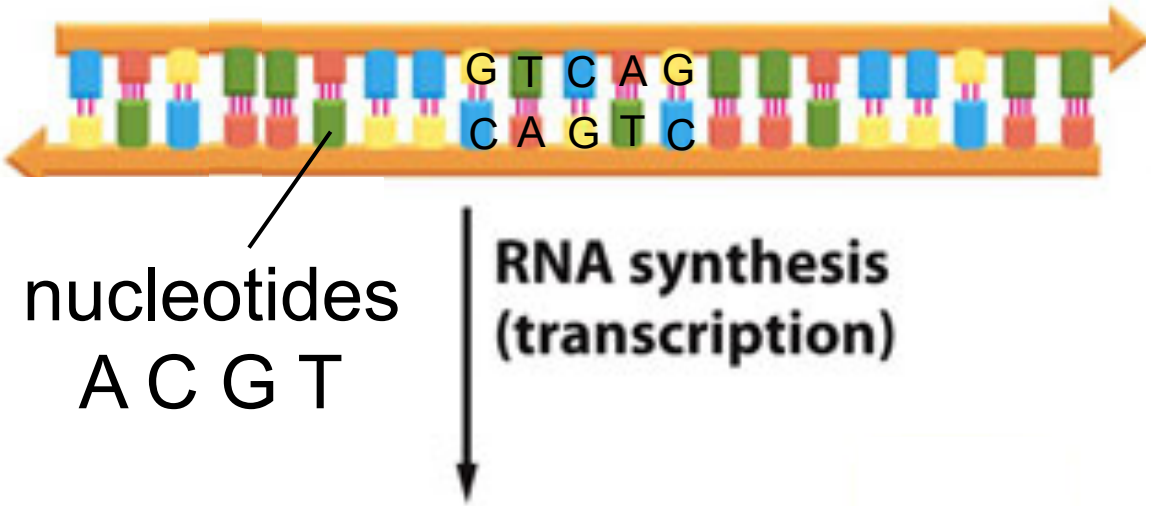
Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

Central dogma of molecular biology

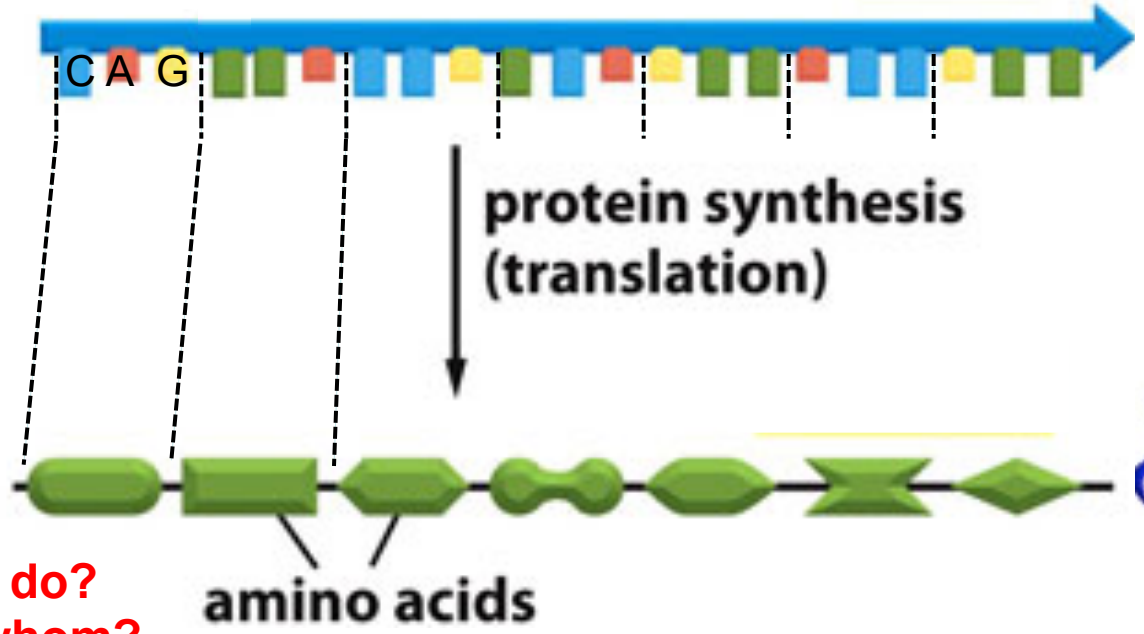
DNA

3 billion base pairs in human genome

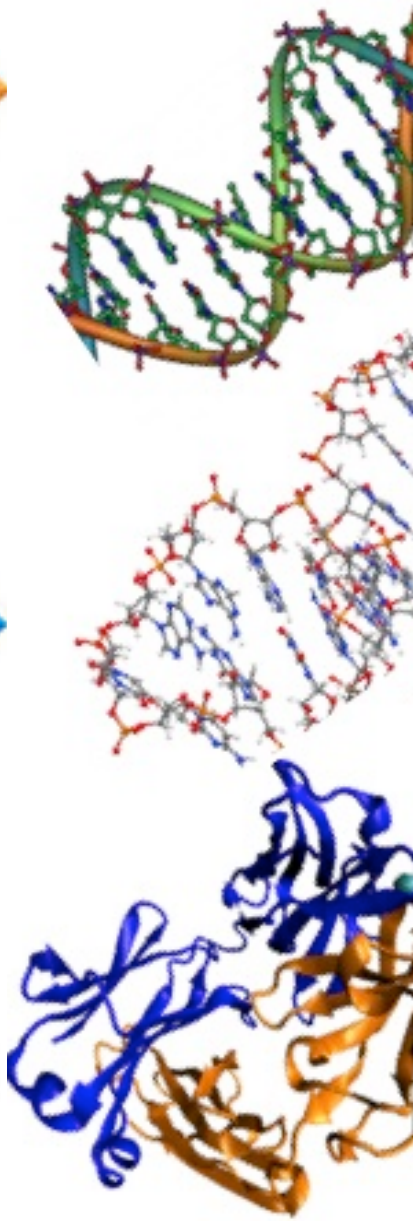


RNA

20 000 protein-coding genes in human genome



Protein



What do they do?
How and with whom?

Bioinformatics = biology ⋒ data science

Analyse / model biological (high-throughput) data

Genome sequences, RNA sequences

Protein sequences (translated from DNA sequences)

Protein 3D structures, RNA structures

Microscopy images (automatically acquired)

Mass spectroscopy data (proteins, small molecules,...)

Concepts and methodologies

- (Bayesian) statistics
- Machine learning, including deep learning
- Computer science: algorithms and data structures

Sequencing costs have fallen 10 000-fold within 8 years!

Oxf. nanopore

Illumina GAIIx

High-throughput sequencing is a transformative technology for biology and medicine

- De-novo assembly of genomes
- Gene expression profiling (RNA-seq,...)
- Genome-wide protein binding (ChIP-seq, …)
- Transcriptome-wide protein binding (CLIP,...)
- Histone maps & modifications (MNase-seq, ChIP-seq,...)
- Open chromatin (ATAC-seq, DNAse-seq,…)
- Genome-wide DNA methylation (bisulfite-seq, meDIP-seq,...)
- DNA 3D contacts (Hi-C,…)
- **Single-cell gene expression profiling**
- **Metagenomics, metatranscriptomics**
- …

The progress in sequencing is leading to a paradigm change in biological research:

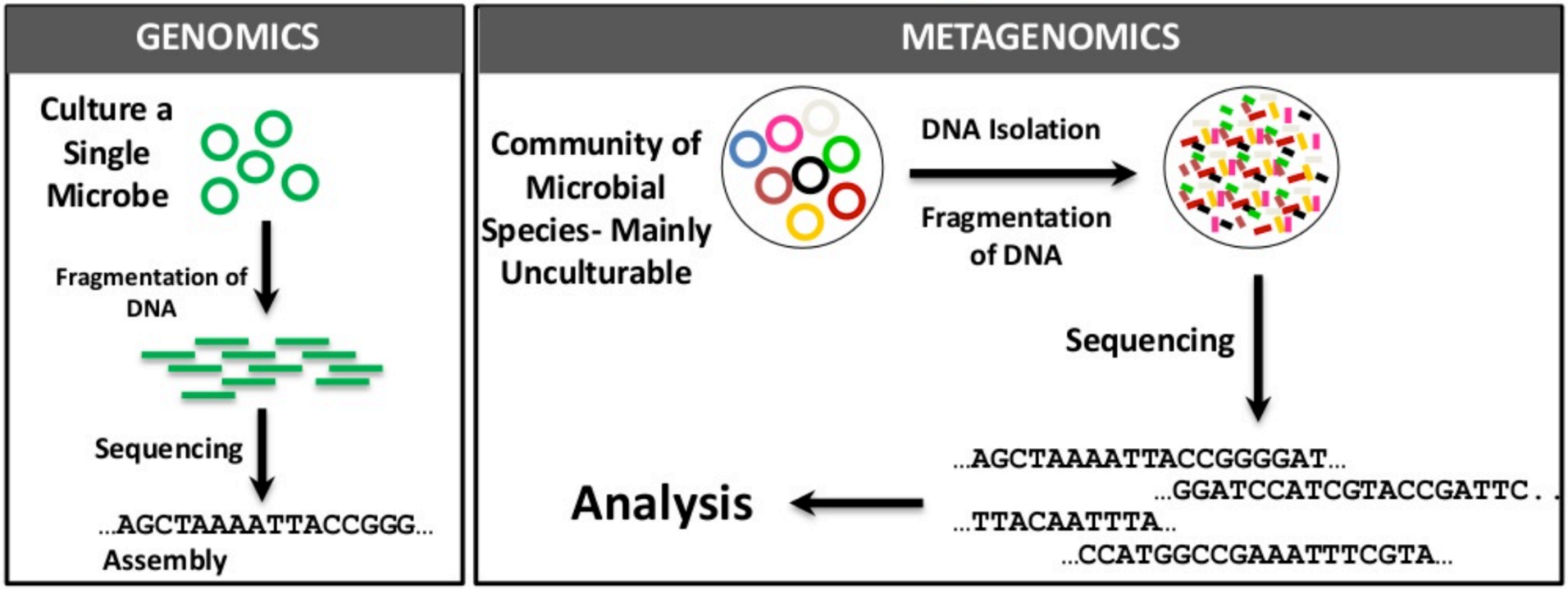
descriptive ⇒ quantitative

hypothesis driven ⇒ data- / analysis-driven

Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values, sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

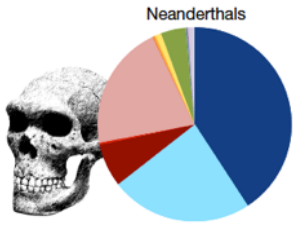
Metagenomics allows us to study the 99% of uncultivatable microbes - by sequencing their DNA directly from the environment



Metagenomics boom: a few recent papers

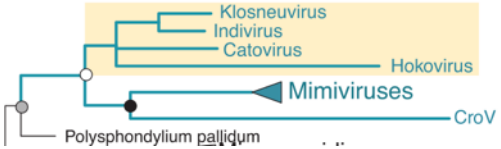
Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus

Nature 2017, Apr 20



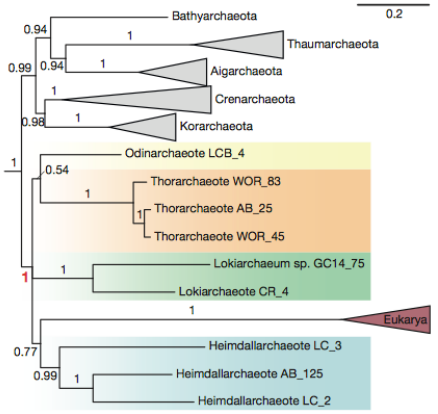
Giant viruses with an expanded complement of translation system components

Science 2017, Apr 7



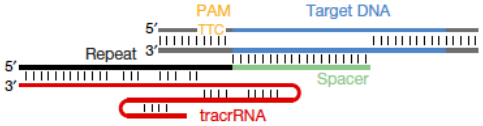
Asgard archaea illuminate the origin of eukaryotic cellular complexity

Nature 2017, Jan 19



New CRISPR–Cas systems from uncultivated microbes

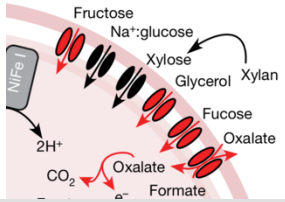
Nature 2017, Feb 09



Metagenomics boom: a few recent papers

Genome-centric view of carbon processing in thawing permafrost

Nature 2018, Aug 02



Structure and fu microbiome

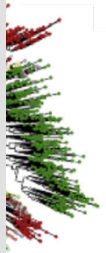
Nature 2018, Aug 09

Applications:

- **Human health (gut, skin, ...)**
- **Ecology & climate**
- **Enzymes for biotechnology**
- **New drugs and natural compounds**
- **Evolution, tree of life**
- **...**

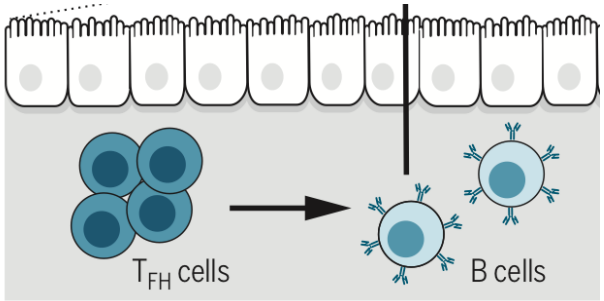
Extensive Une Revealed by O Metagenomes Lifestyle

Cell 2019, Jan 24



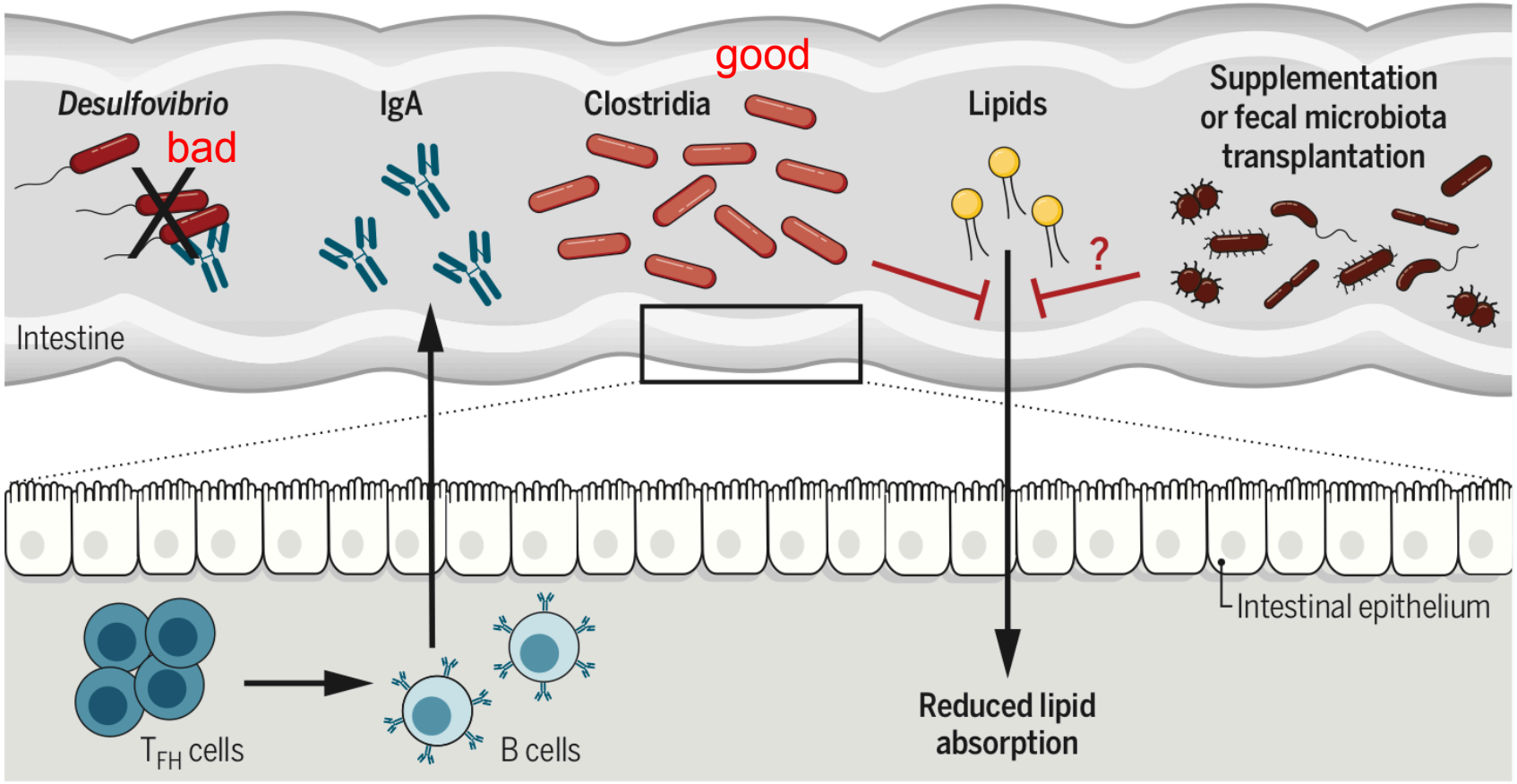
T cell-mediated regulation of the microbiota protects against obesity

Science 2019, Jul 26



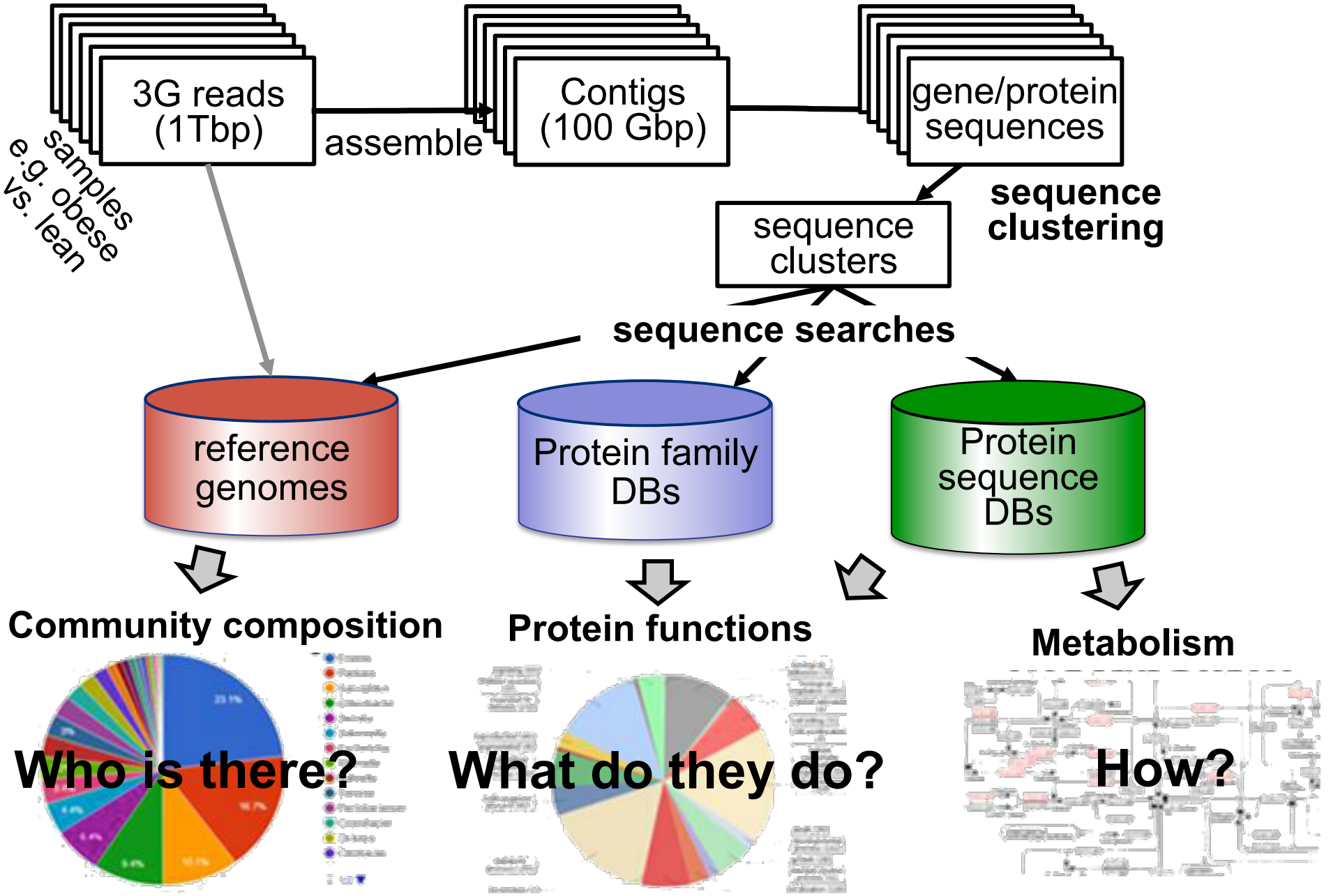
Immune system and gut microbiome are intimately connected

T cell-mediated regulation of the microbiota protects against obesity Science 2019, Jul 26



- Our adaptive immune system shapes our microbiome by suppressing noxious bacteria
- Modern life style kicks microbiomes out of balance ⇒ obesity + autoimmunity

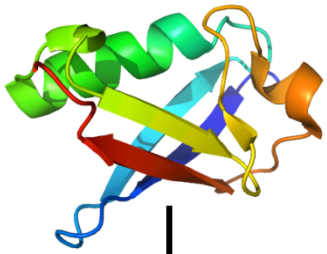
Metagenomics data analysis



Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

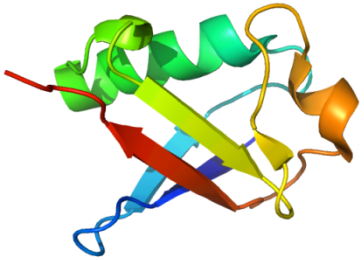
Homologous = descended from common ancenstor



ancestral protein

2B years

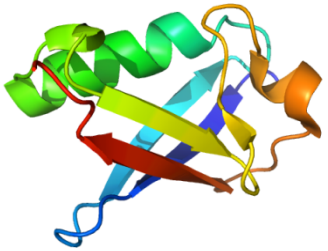
MKAYLGLYTARLETP-ARSLKEKRALIKPALERLKARFPVSAARL



query protein

similar structure
similar function

MLIGVCRVIISIPESFSLKEKRKVKRSIVDKVRSKFNVSIAEV



homologous protein

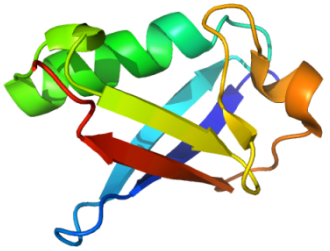
Homology-based inference of protein structure and function

query protein

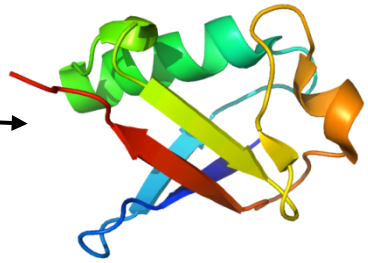
sequence search

Protein sequence DB

predict structure and function of query from those of database match



homologous sequence found with known structure and functions



2B years

Sequence-sequence comparison

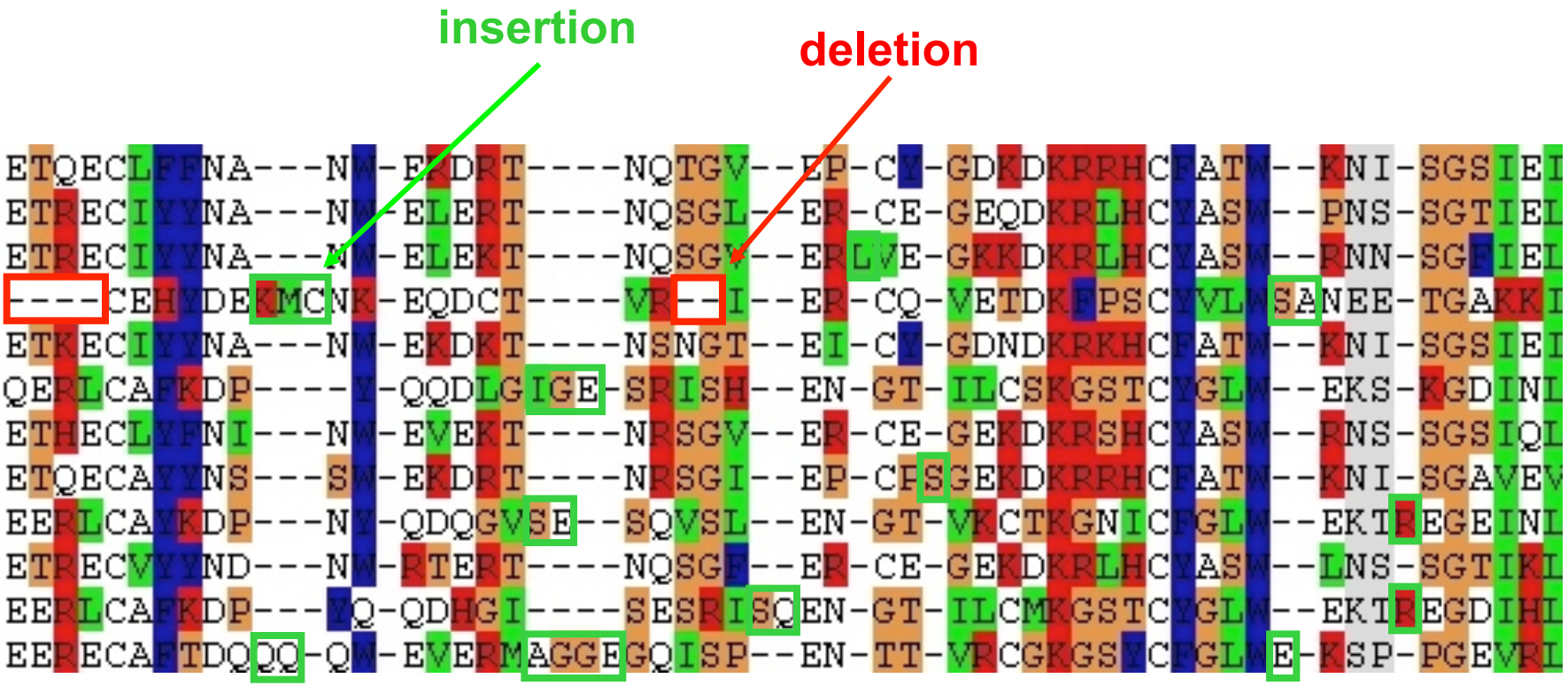
- A sequence alignment groups similar residues into same column. These residues are assumed to occupy homologous positions in the proteins

- Alignment score = sum of **similarity scores** − gap penalties:

  Score = S(V,I)+…+S(V,I)+…+S(E,G)+…+S(G,G) – d – e

- Find alignment with maximum score, rank by score
  → „dynamic programming“ takes ~10× $L_1 L_2$ steps → O($L_1 L_2$)
  (with L_1 , L_2 the sequence lengths)

A multiple sequence alignment



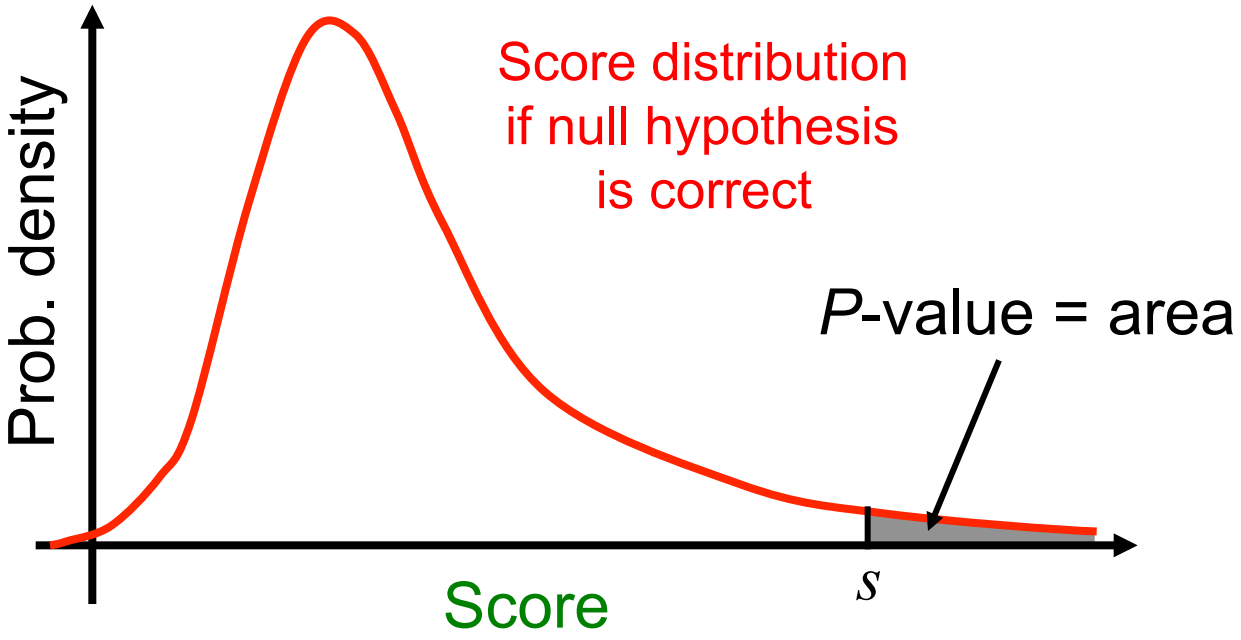
P-values quantify plausability of *null hypothesis*

Given: a *null hypothesis* (boring "hypothesis of randomness") and a score ("test statistic") with known distribution under the null hypothesis

Goal: find interesting cases for which the null hypothesis can be rejected

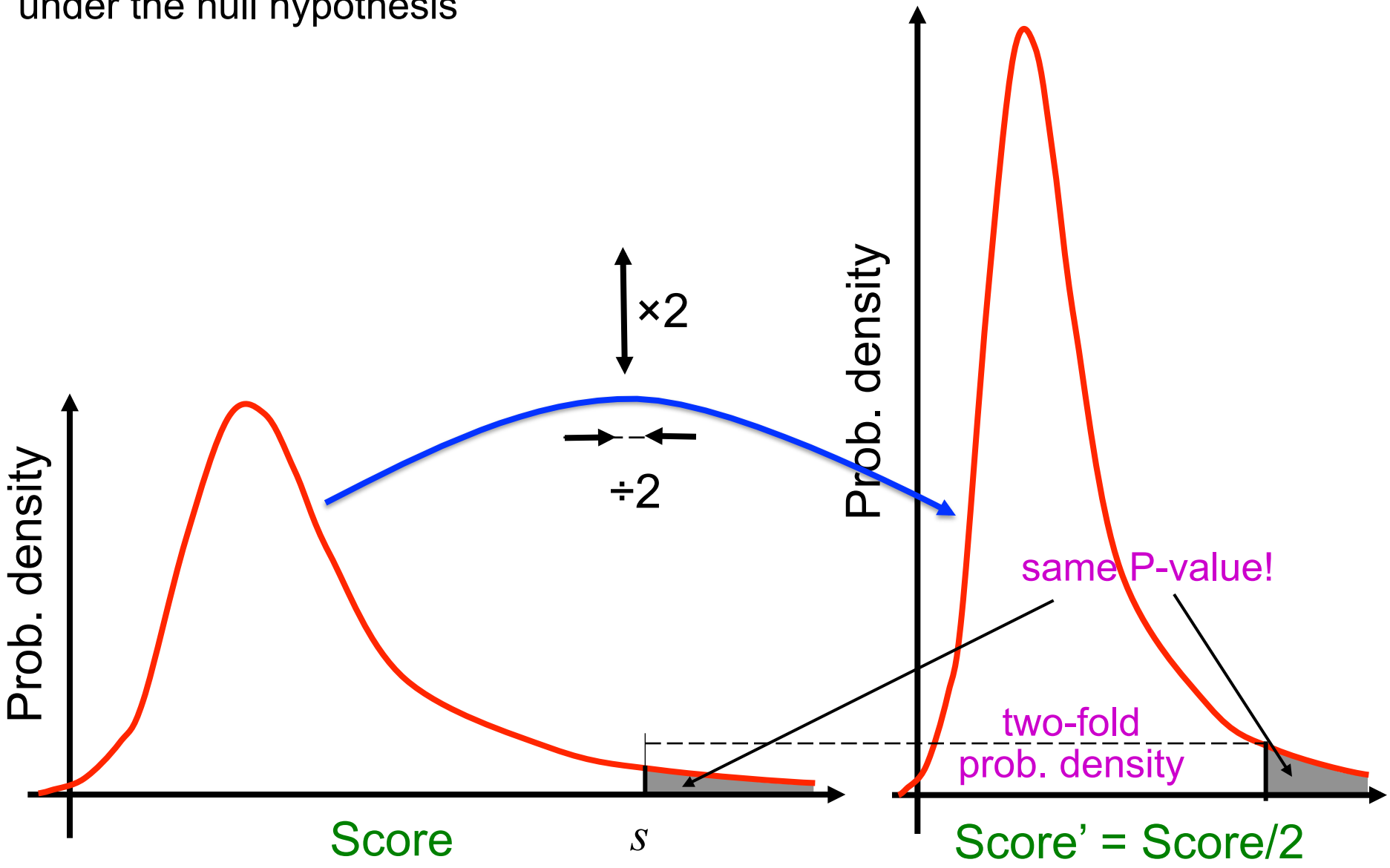
***P*-value** = the probability to obtain a score as observed *or more extreme*, under the null hypothesis.

A small *P*-value (e.g. < 0.01) indicates the null hypothesis can be rejected.



Quizz: why „or more extreme“?

***P*-value** = the probability to obtain a score as observed **or more extreme** under the null hypothesis



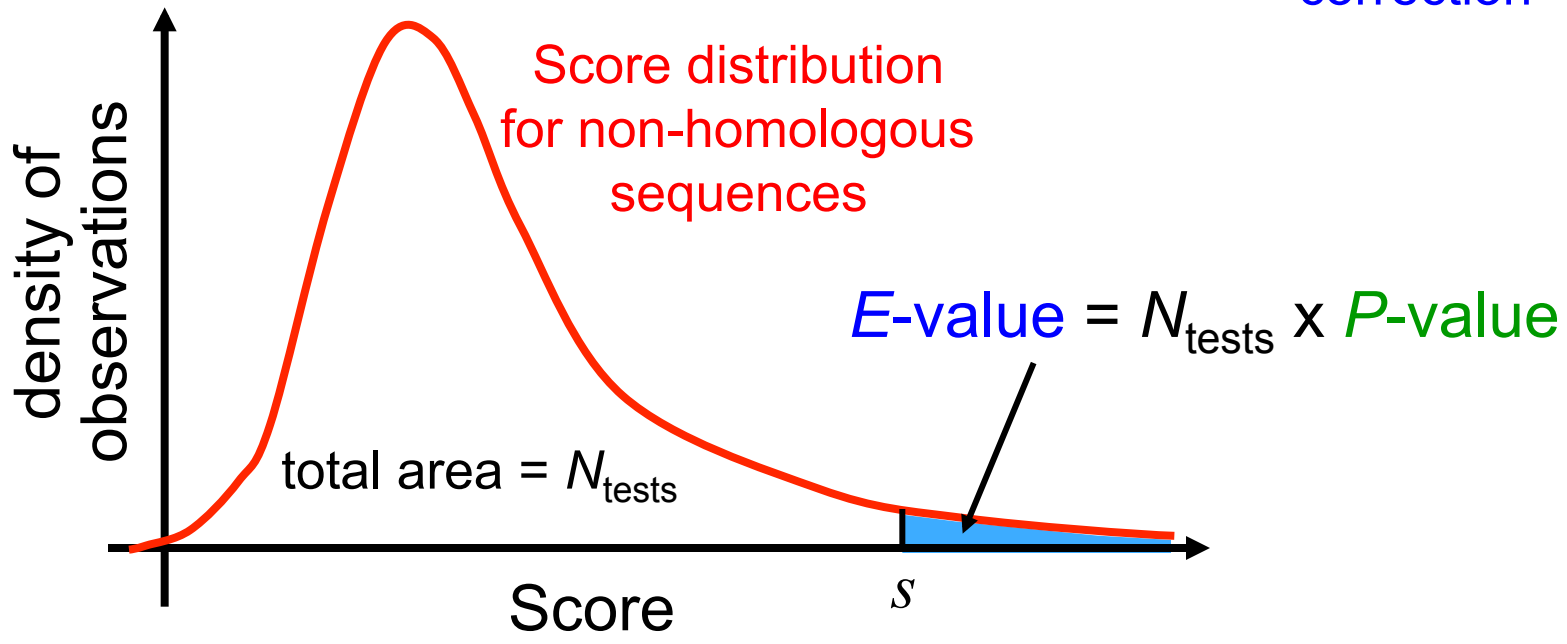
E-value = expected number of observations more extreme than the one observed

① P-value = Probability for event with score ≥ s under the null hypothesis

② E-value = *Expected number of events out of* N_{tests} *trials* with score ≥ S under the null hypothesis

$$E\text{-value} = N_{tests} \times P\text{-value}$$

similar to Bonferroni multiple testing correction



Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

MICROBIOLOGY

nature

Metagenomics

Philip Hugenholtz and Gene W. Tyson

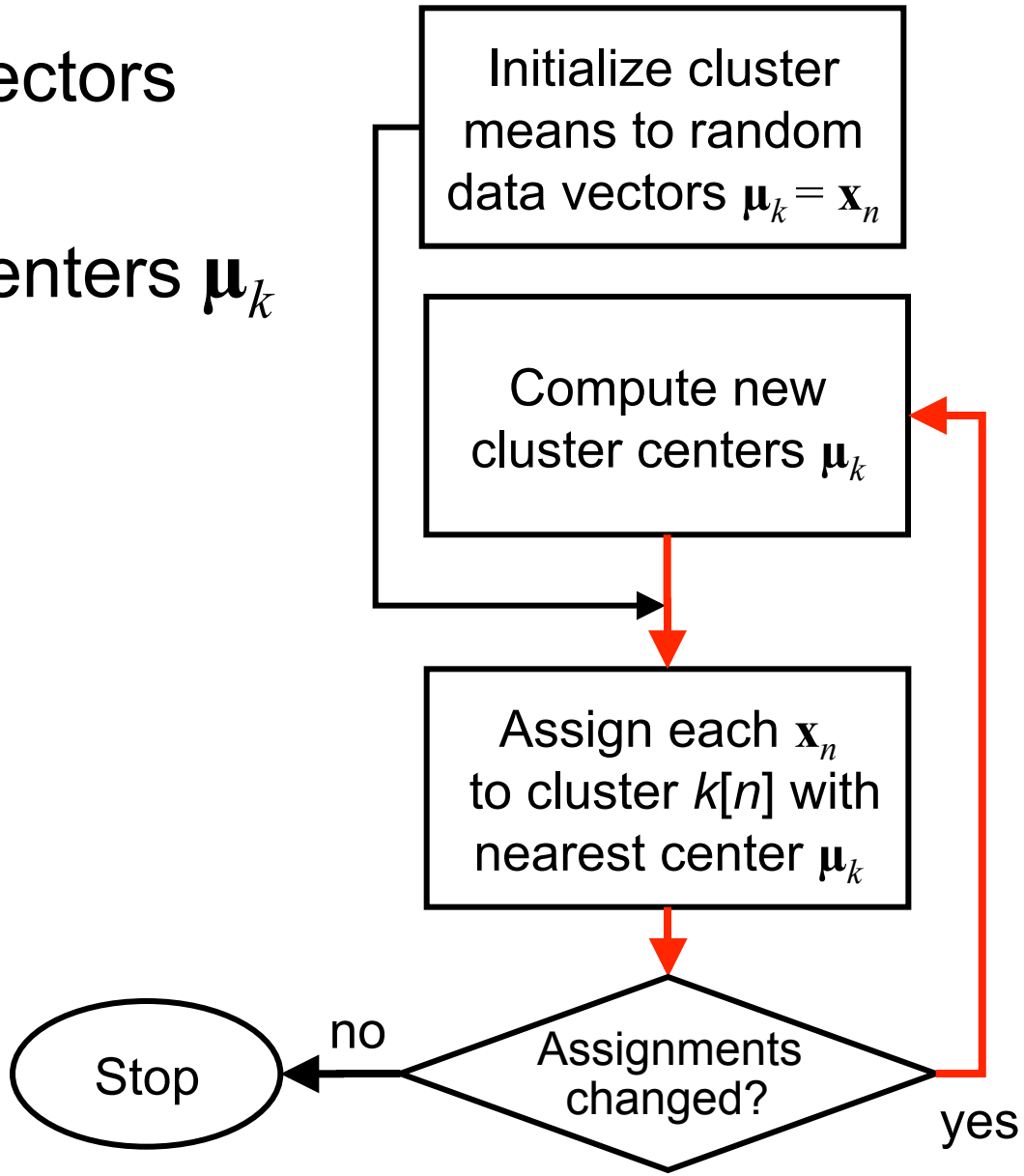
Vol 455|25 September 2008

Will computational capacity keep pace?
That remains to be seen. Sequence data are increasing at a rate higher than increases in computational power. Even more problematic than simple storage of sequence data are the 'all-versus-all' sequence comparisons required to best interpret metagenomes, which raise the computational requirements exponentially. Unless there is a radical breakthrough in computing, for example if quantum computers

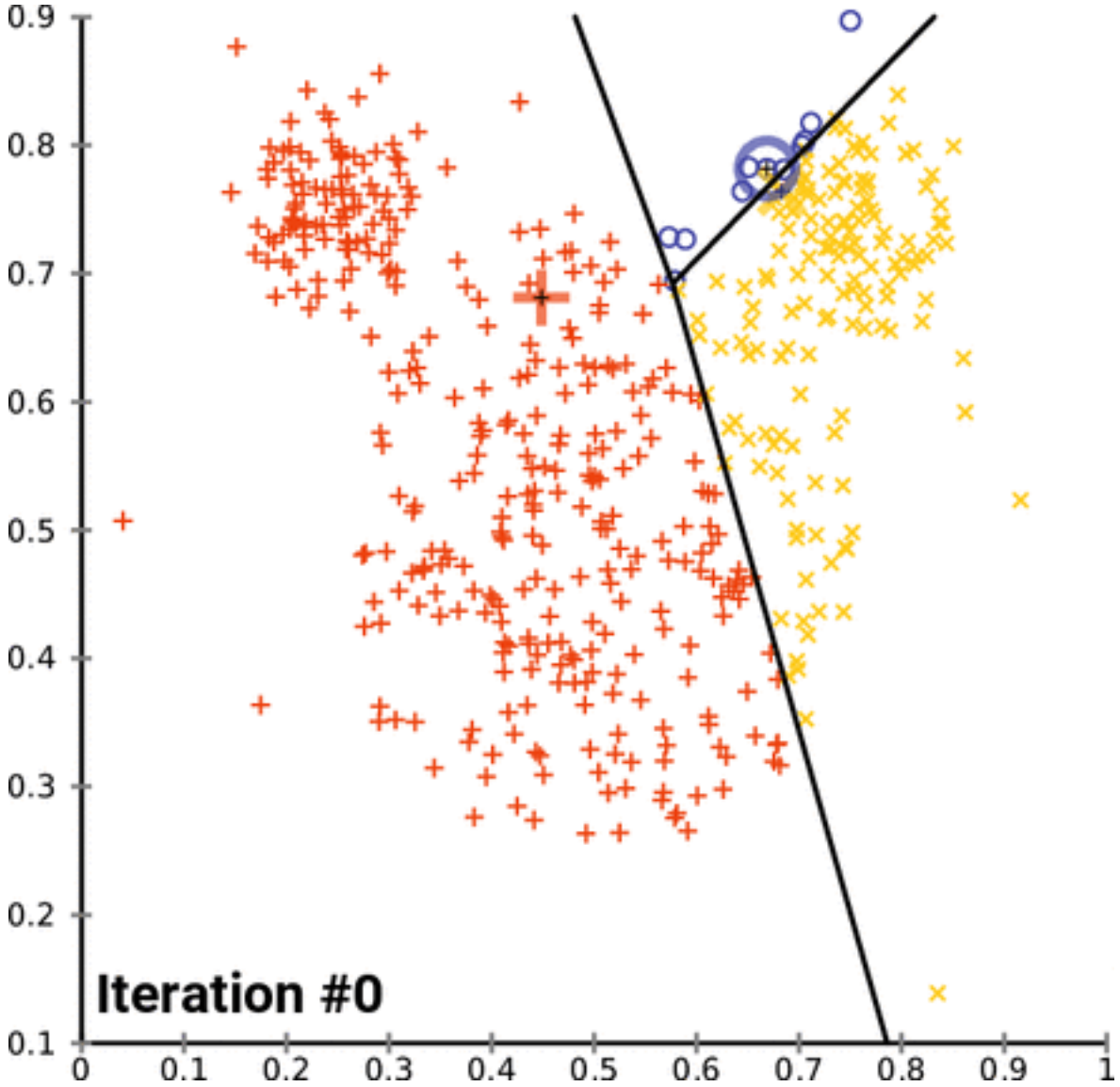
Sequence clustering groups sequences with same functions and structures together, reduces redundancy

K-means clustering

Goal: cluster the N vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathrm{R}^D$ into K clusters with centers $\boldsymbol{\mu}_k$



K-means clustering



https://en.wikipedia.org/wiki/K-means_clustering

K-means clustering − time complexity

Goal: cluster the N vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathrm{R}^D$ into K clusters with centers $\boldsymbol{\mu}_k$

How ma… take? (T… depend on N, K, and D)

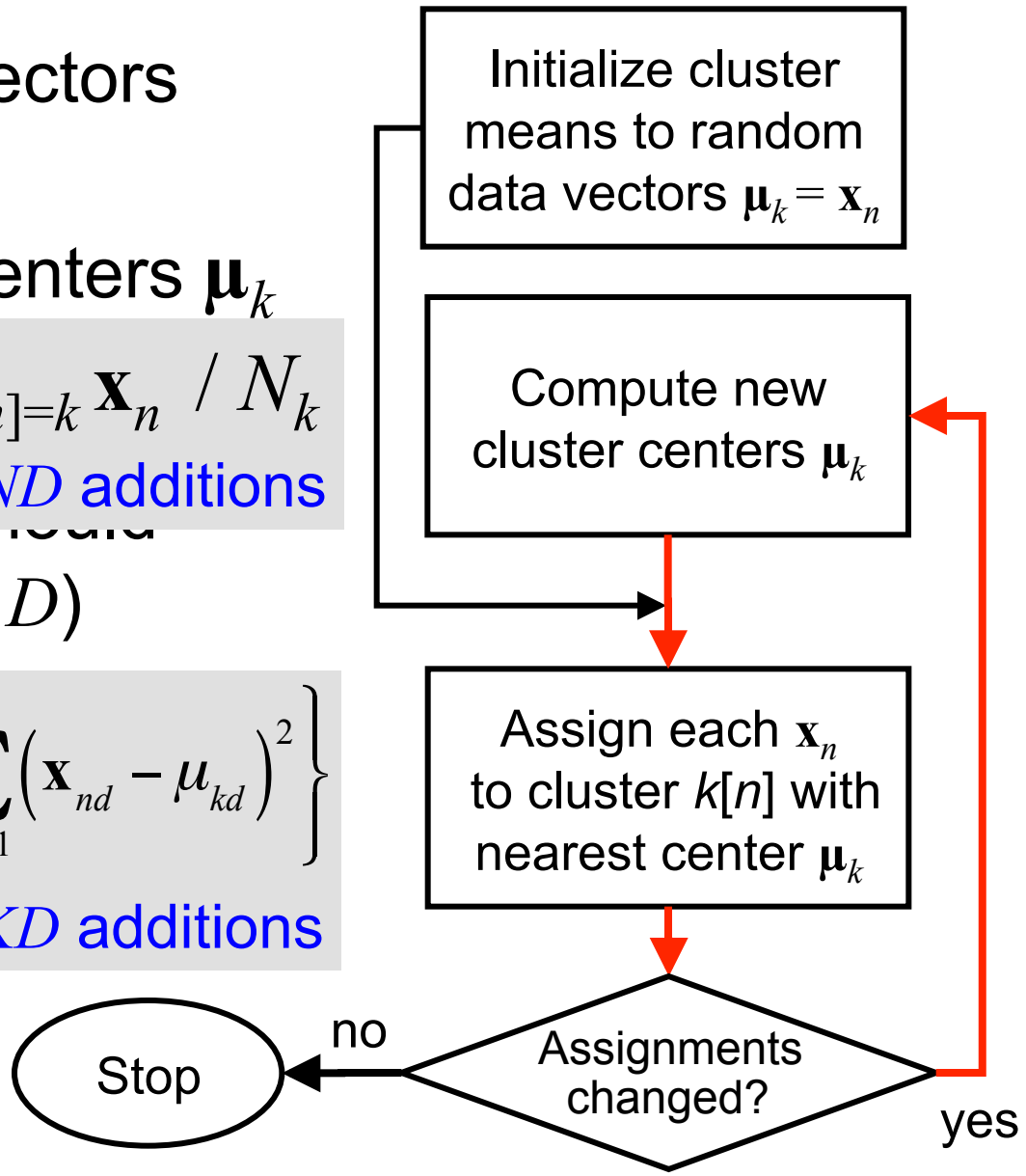
$$\boldsymbol{\mu}_k = \sum_{n:k[n]=k} \mathbf{x}_n \; / \; N_k$$

ND additions

$$k[n] = k \text{ giving } \min \left\{ \sum_{d=1}^{D} \left(\mathbf{x}_{nd} - \mu_{kd} \right)^2 \right\}$$

NKD additions

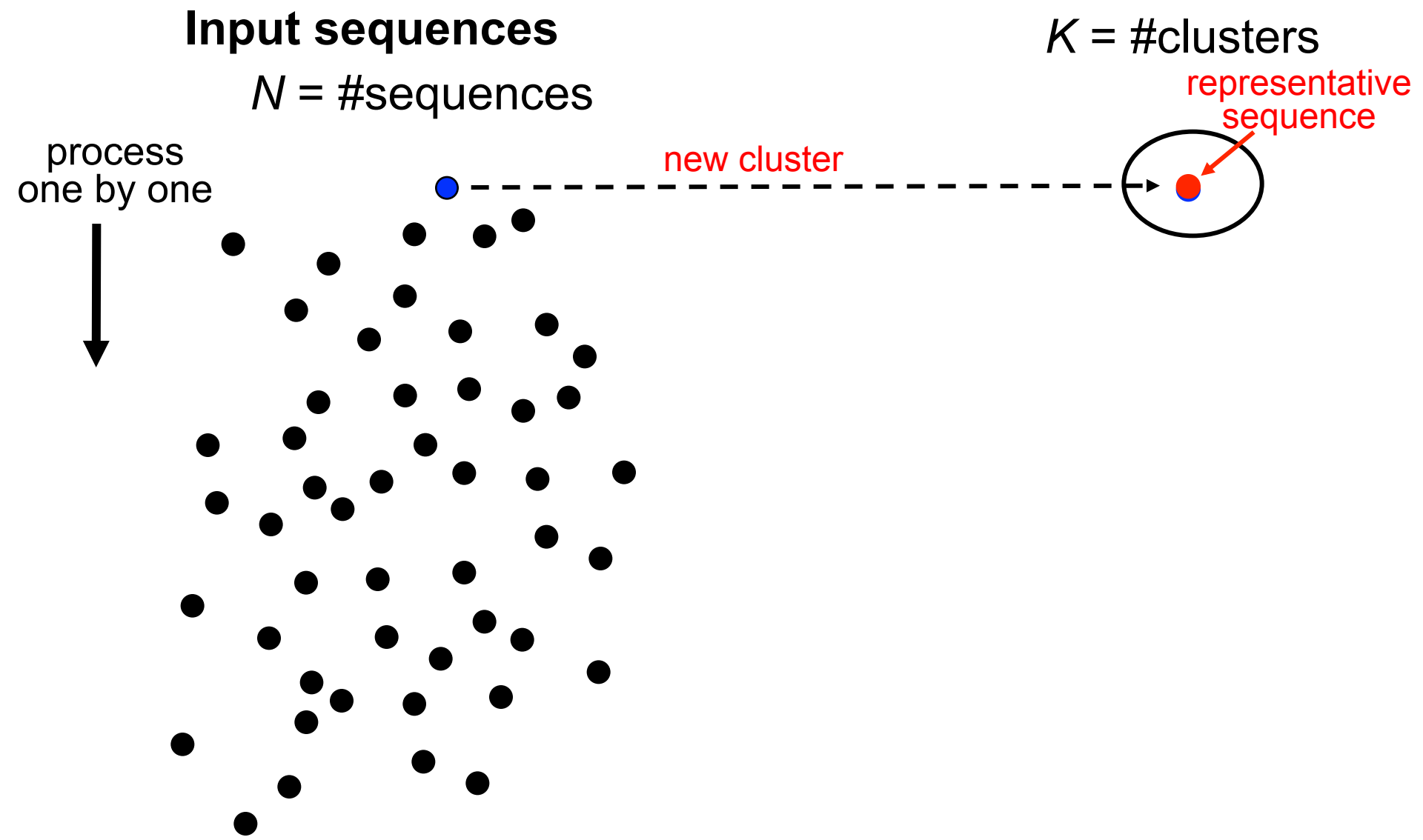
#steps per iteration = $(N + NK)\, D$



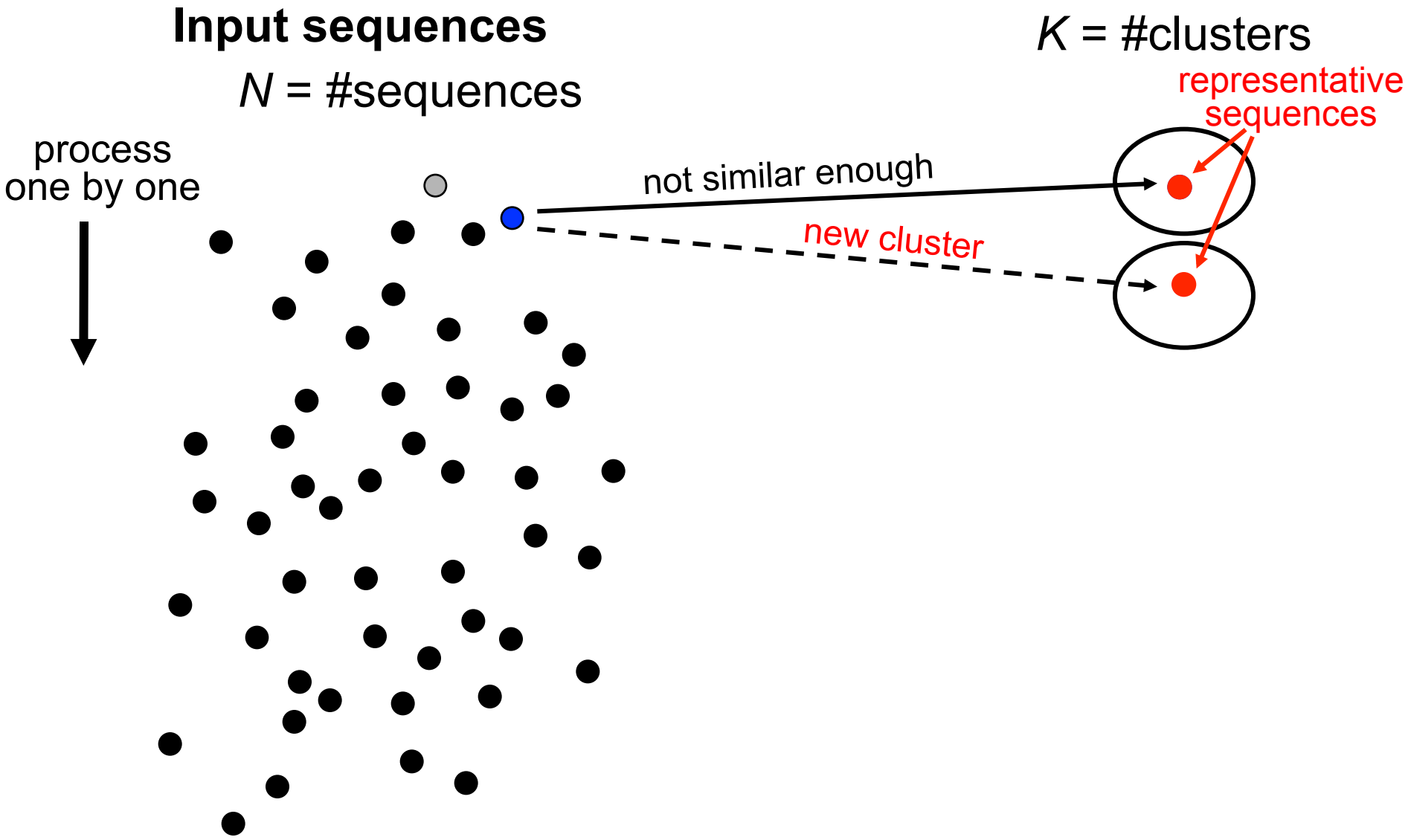
We cannot use K-means for clustering sequences

- We cannot calculate the mean of sequences (→ possible solution: chose as center the sequence most similar to all cluster members)
- The nearest center sequence is often not **homologous** to all its members! ⚡⚡
- We cannot know the right cluster number K

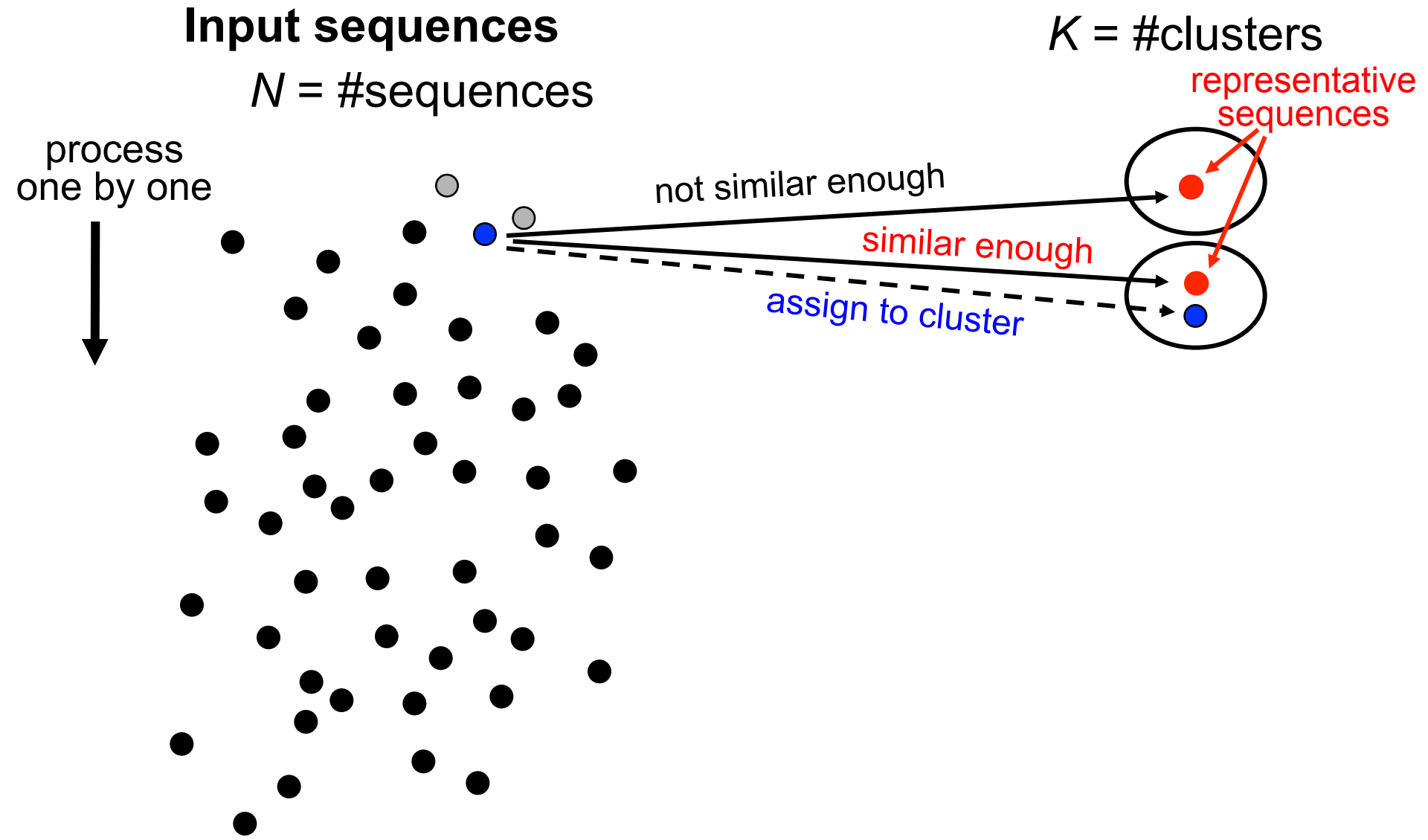
Sequence clustering in O(*NK*) time



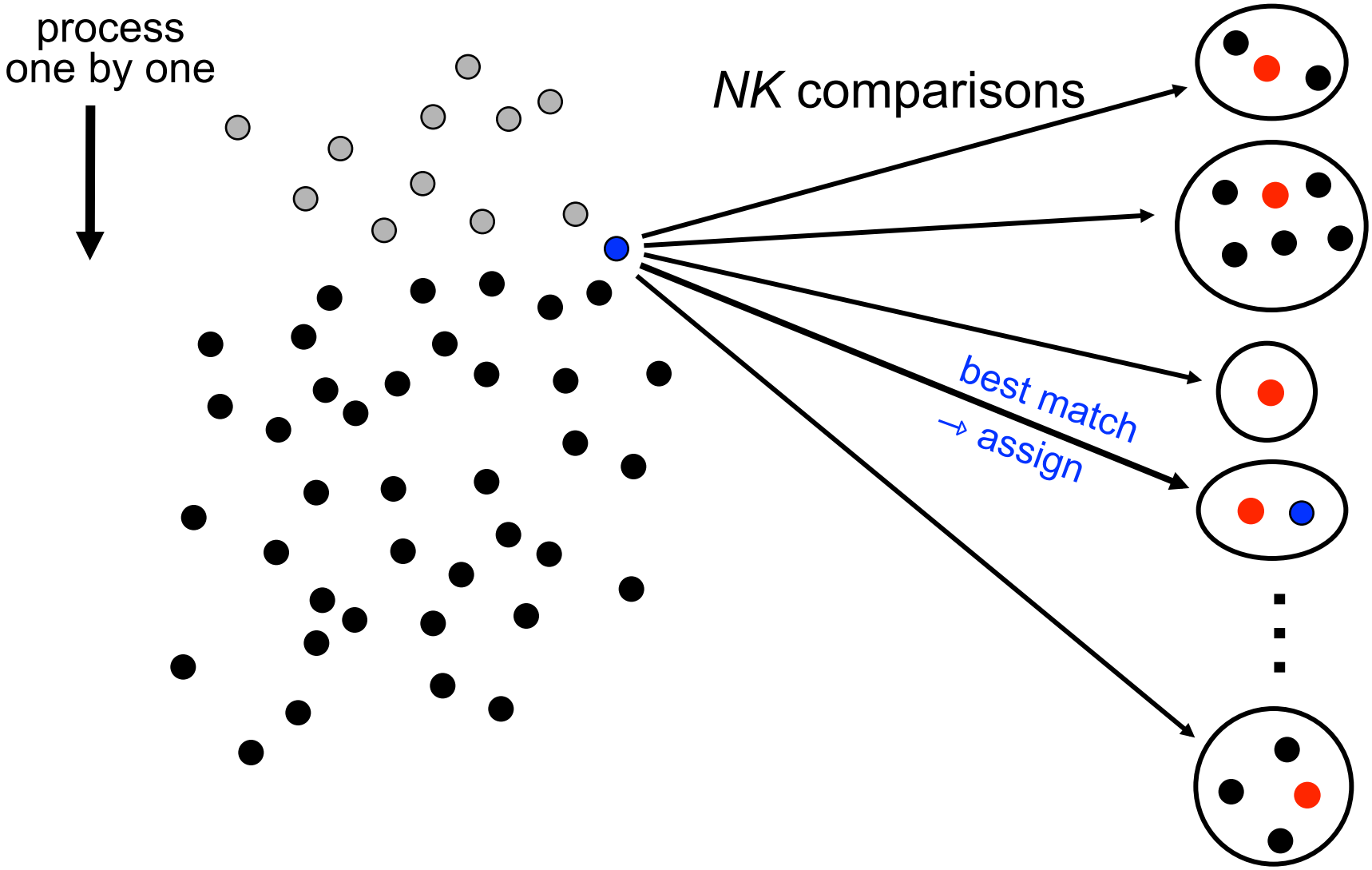
Sequence clustering in O(*NK*) time



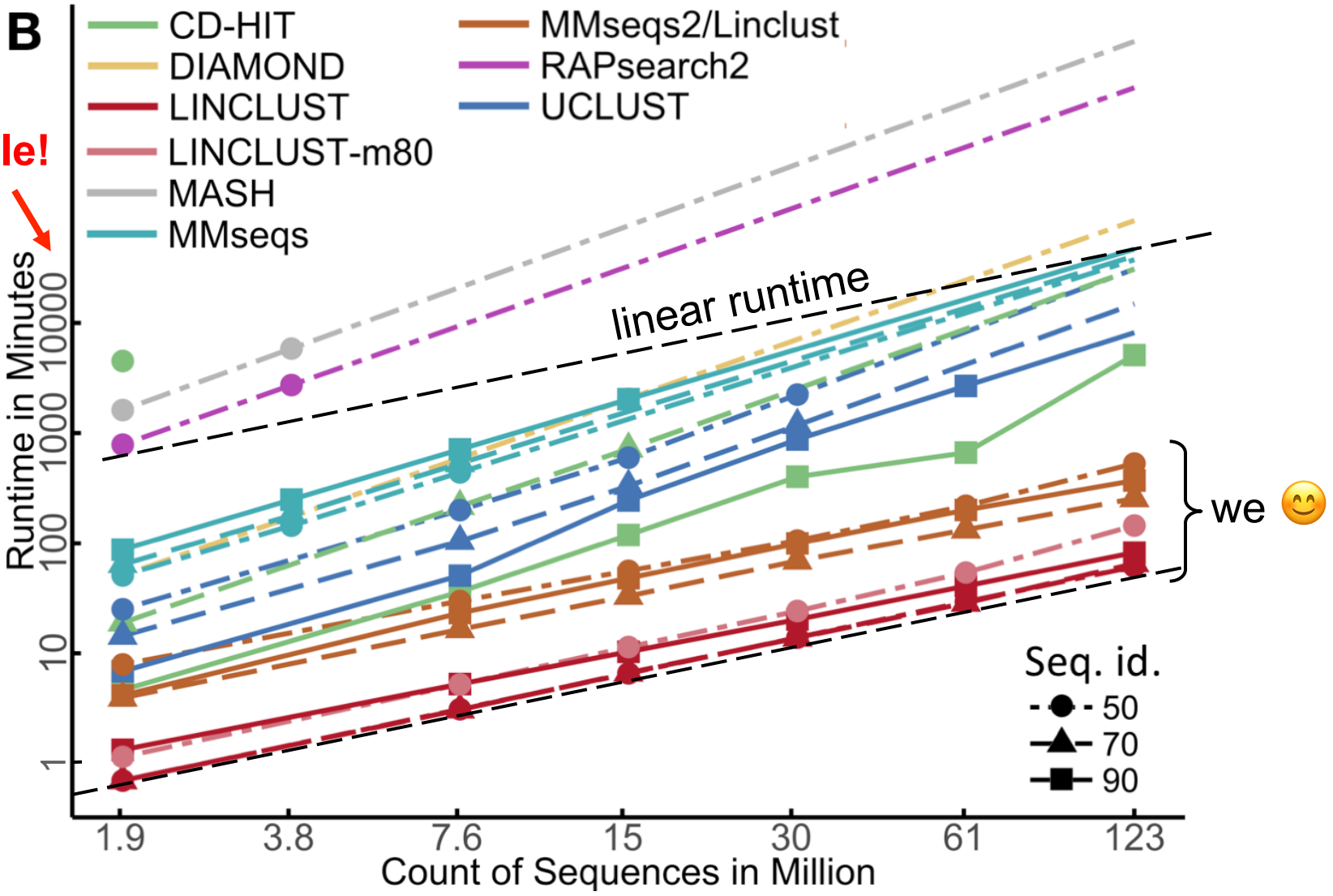
Sequence clustering in O(*NK*) time



Sequence clustering in O(*NK*) time

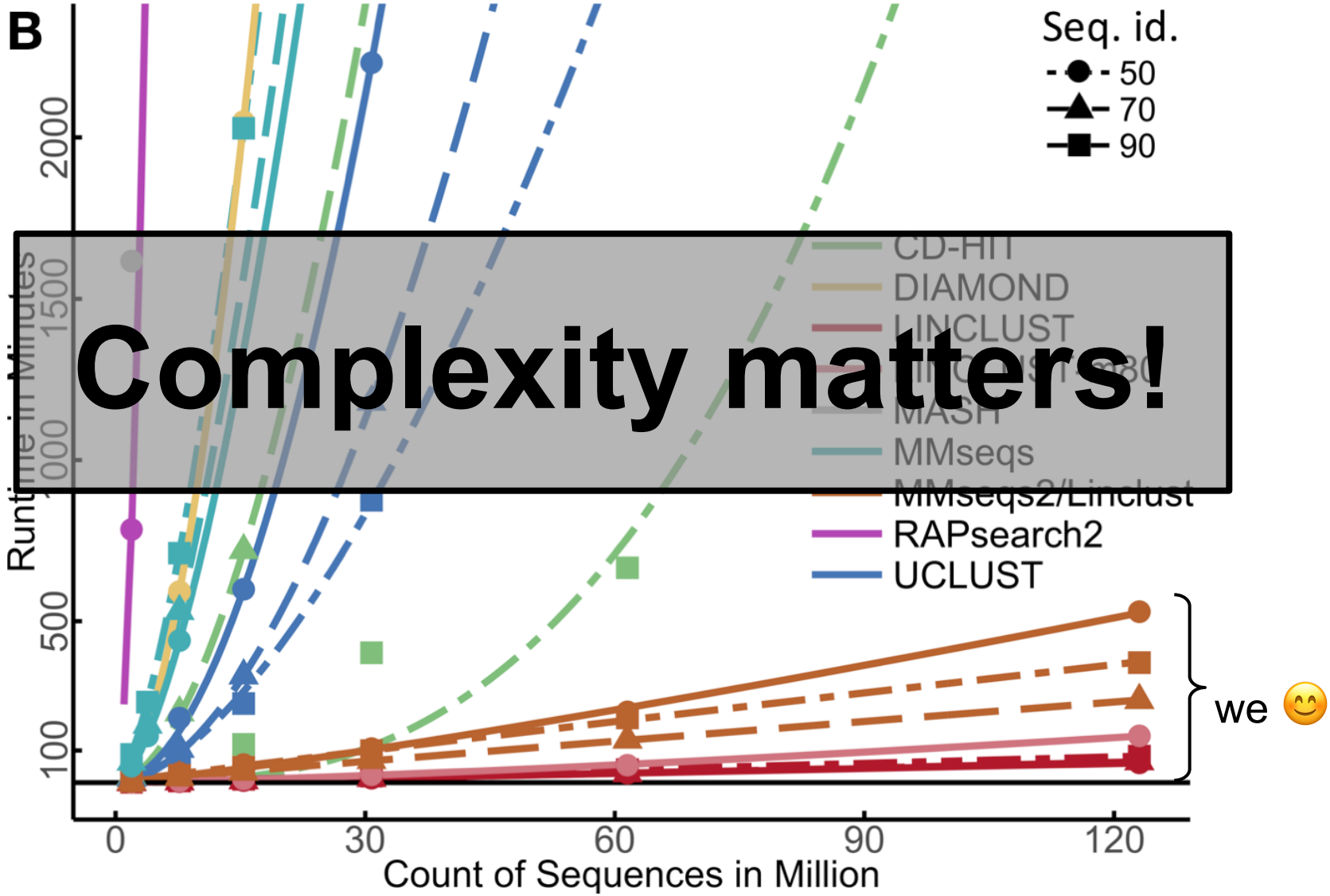


Linclust clustering takes O(*N*) time!



Clustering 10^9 sequences requires only 10^9 × **20** comparisons!

Linclust clustering takes O(*N*) time!

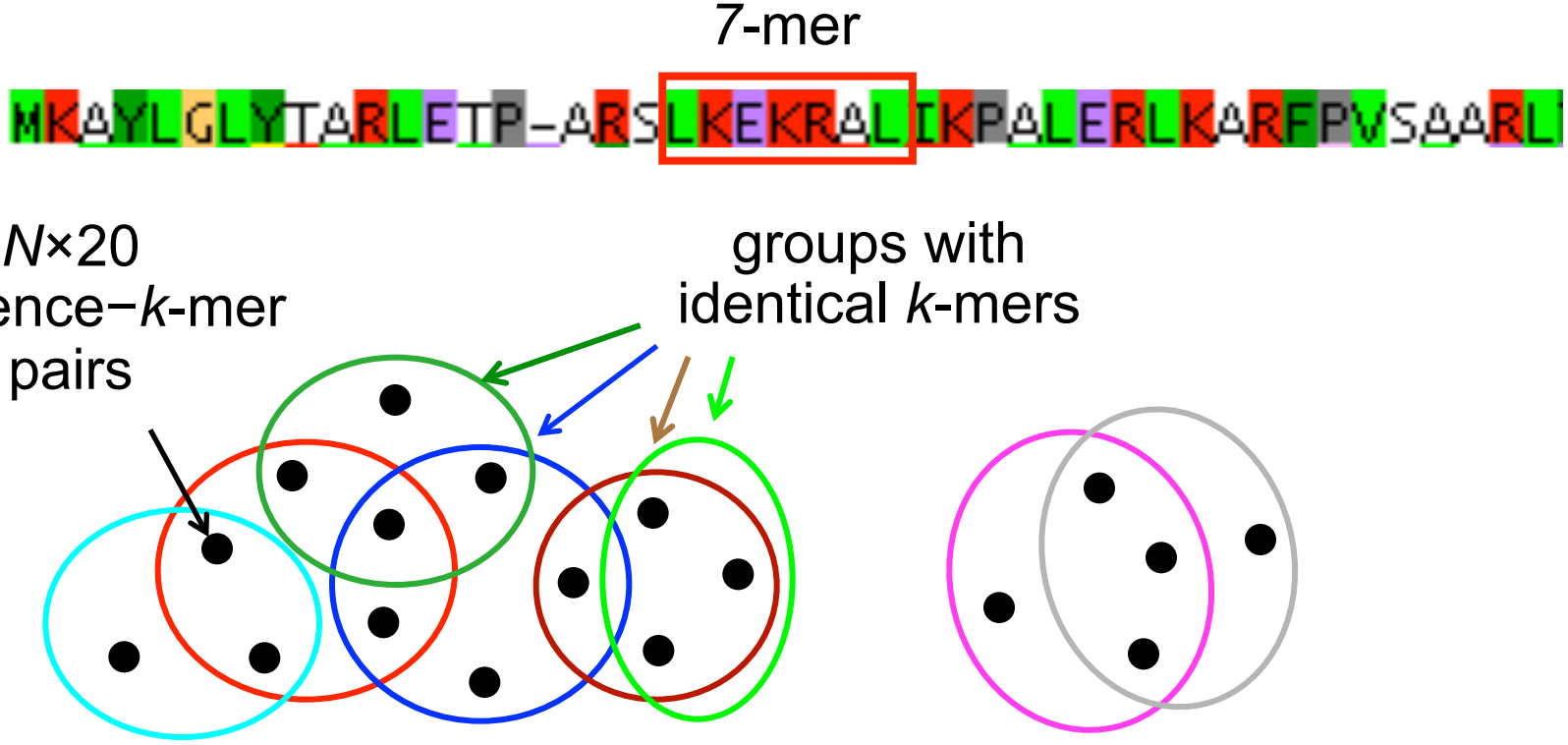


Complexity matters!

Clustering 10^9 sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

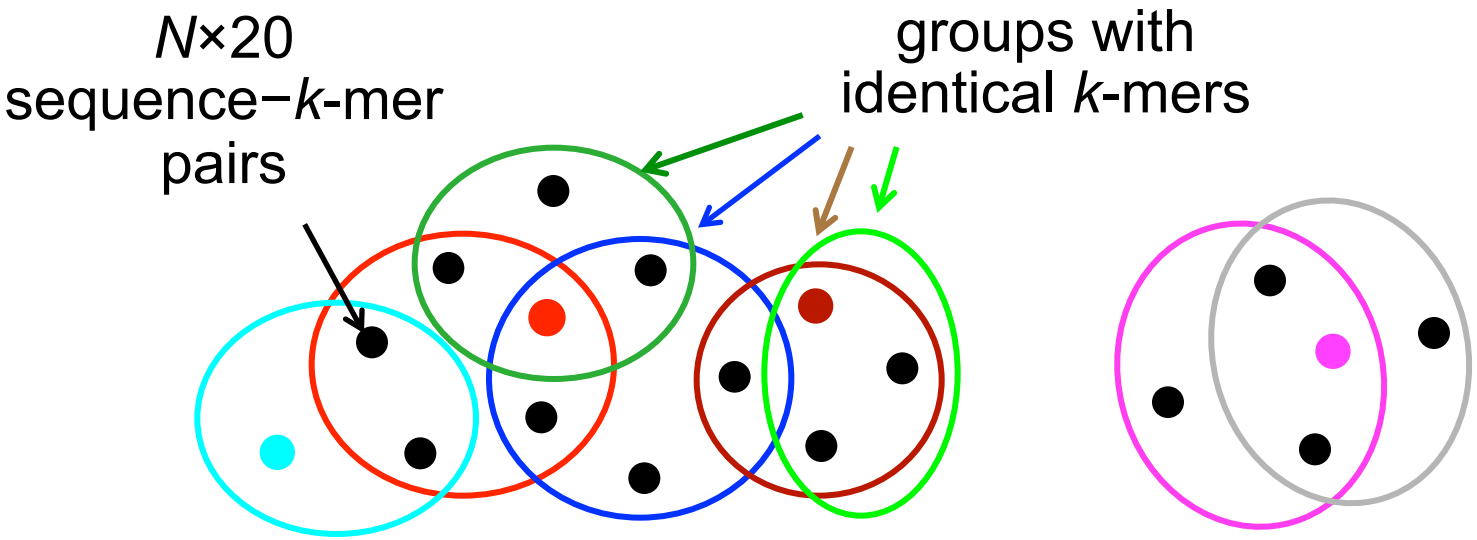
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer (by numerically sorting seq-*k*mer pairs)



Clustering N=10^9 sequences requires only 10^9 × **20** comparisons!

How does Linclust achieve linear runtime?

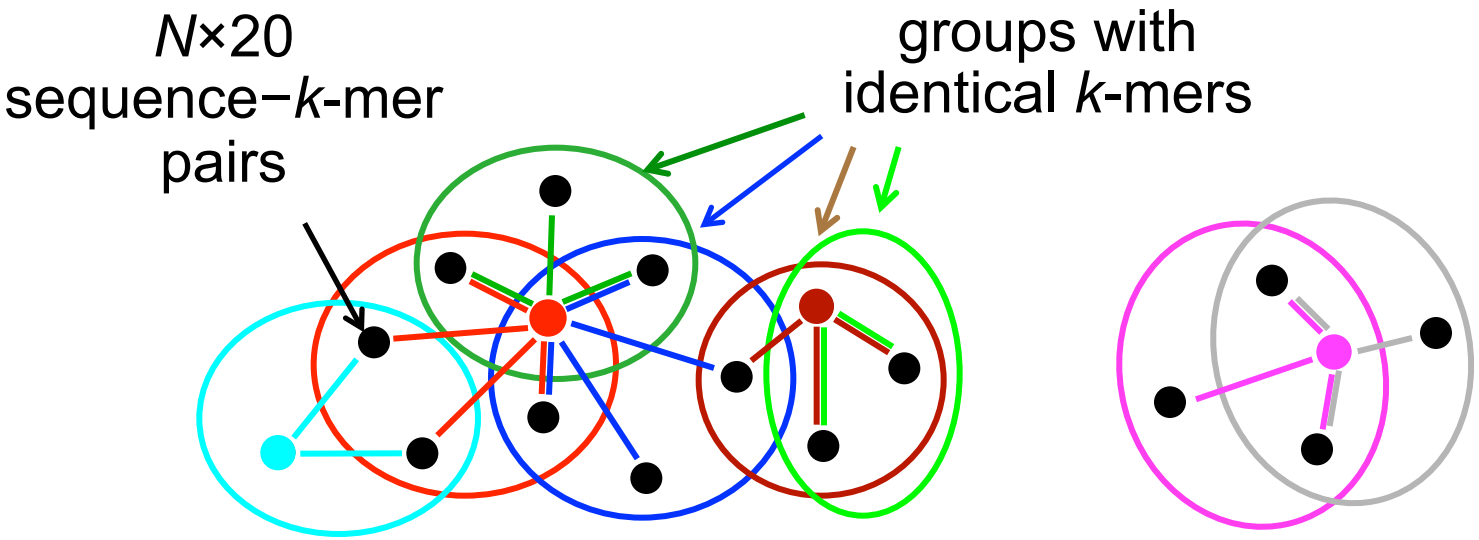
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group (the longest)



Clustering $N=10^9$ sequences requires only $10^9 \times$ **20** comparisons!

How does Linclust achieve linear runtime?

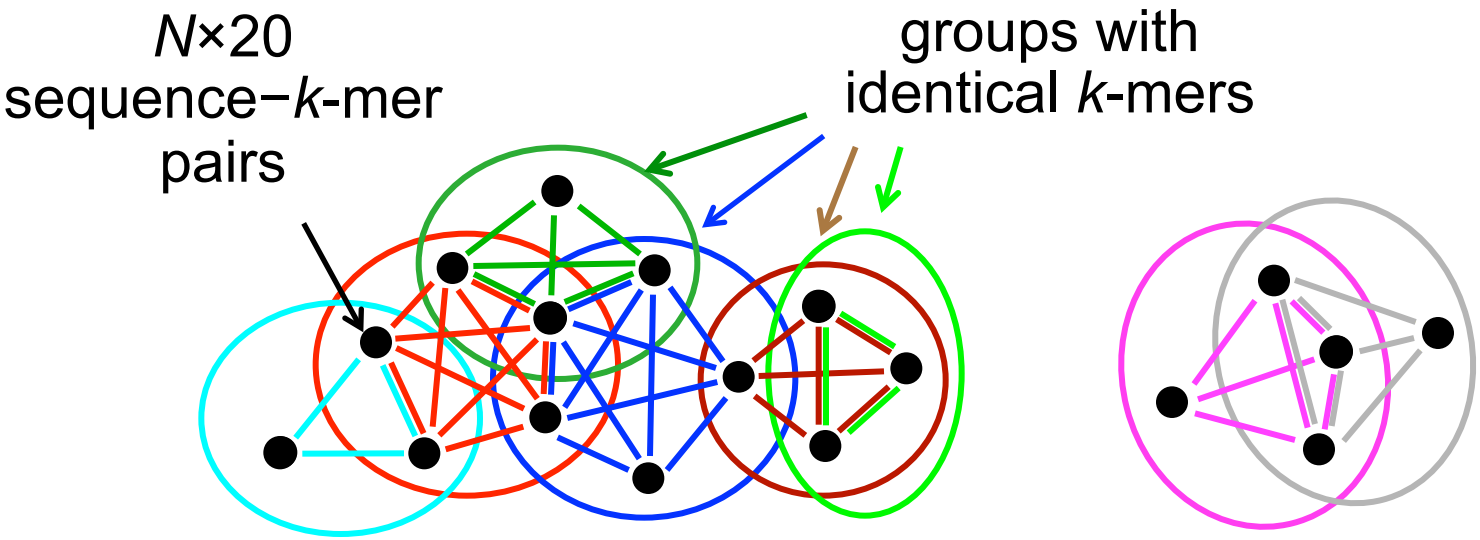
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group **only with the centre sequence**



Clustering $N=10^9$ sequences requires only $10^9 \times$ **20** comparisons!

How does Linclust achieve linear runtime?

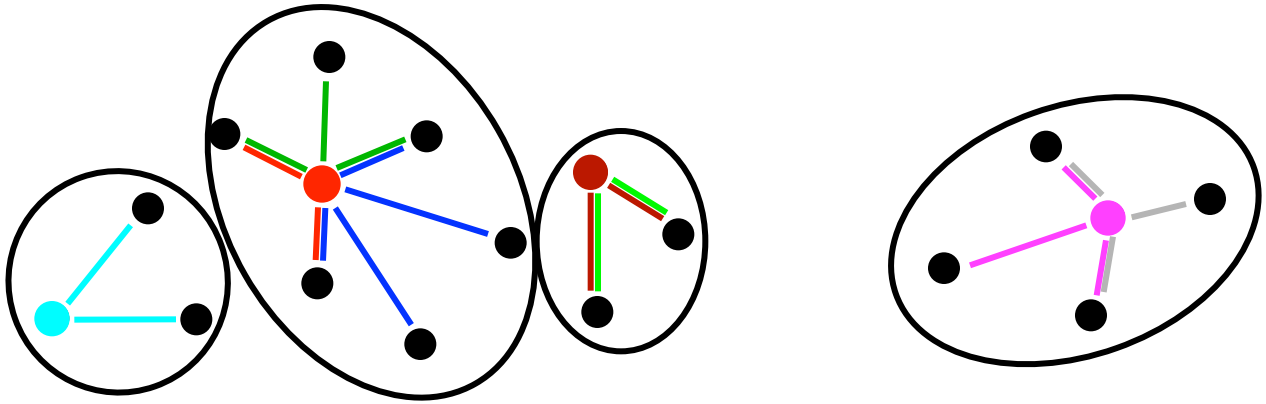
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group **only with the centre sequence**, instead of with all sequences in the group



Clustering $N=10^9$ sequences requires only $10^9 \times$ **20** comparisons!

How does Linclust achieve linear runtime?

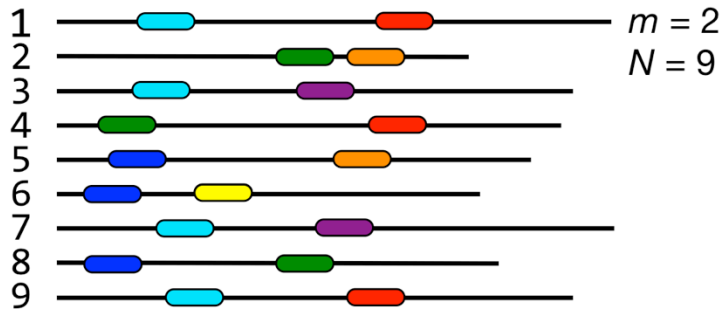
1. Linclust selects **20 *k*-mers** per sequence and finds groups of sequences sharing a *k*-mer.
2. It selects one centre sequence  per group
3. It aligns each sequence in the group **only with the centre sequence**, instead of with all sequences in the group
4. Sequences similar enough to a centre sequence will form a cluster



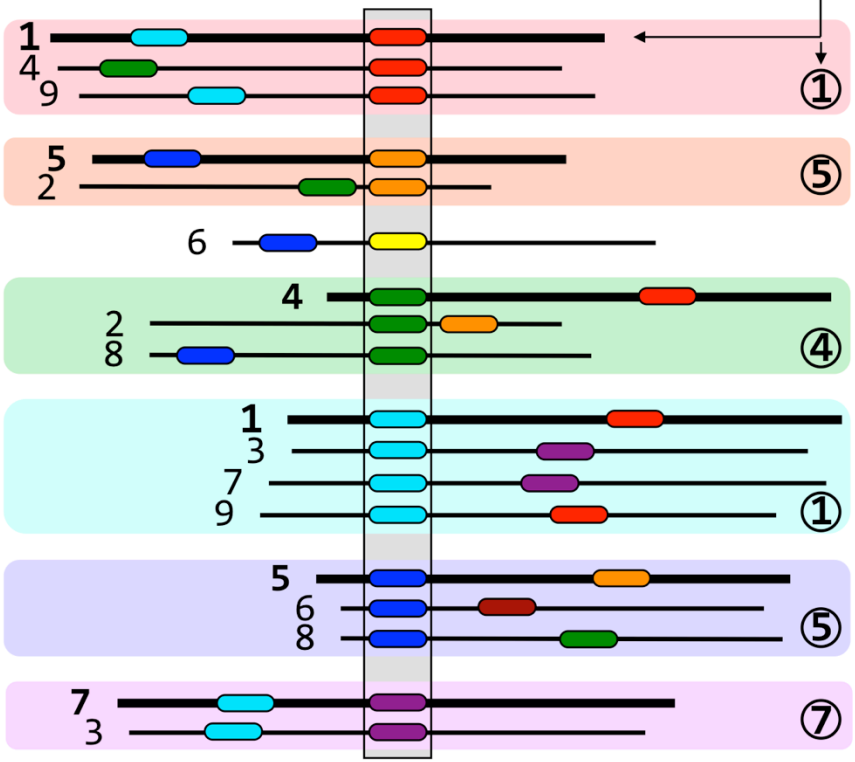
Clustering $N=10^9$ sequences requires only $10^9 \times$ **20** comparisons!

(1) Select *m* *k*-mers with lowest hash values in each of *N* sequences;

Generate table of *m*×*N* lines, 1 per *k*-mer (*k*-mer; sequence ID, *k*-mer position);

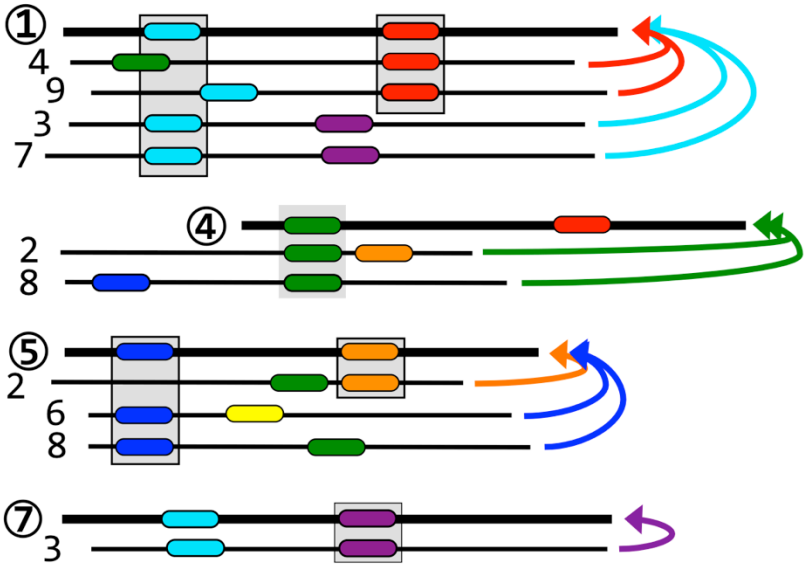


(2) Sort table and select longest sequence per *k*-mer group as **centre sequence**

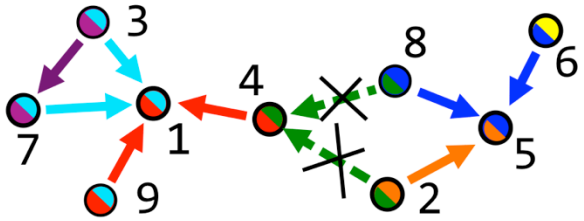


(3) Merge groups by centre sequence;

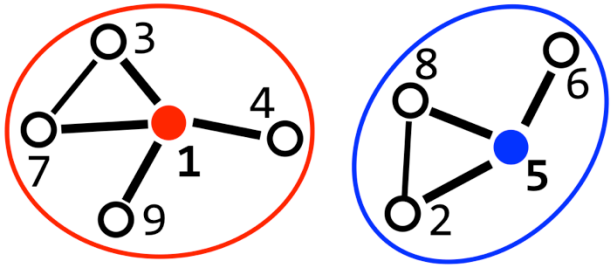
Align each sequence *without gaps* to its centre sequence (< *m*×*N* alignments!)



(4) Remove links below cut-off ; validate remaining links using gapped alignment



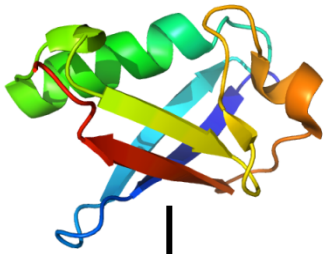
(5) Cluster with greedy incremental algorithm



Test

What does it mean when two proteins are homologous?

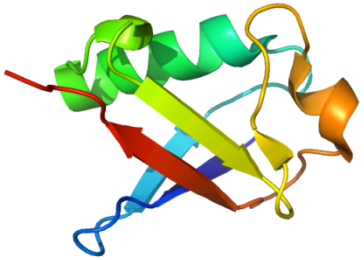
Homologous = descended from common ancenstor



ancestral protein

2B years

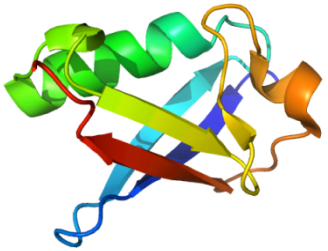
MKAYLGLYTARLETP-ARSLKEKRALIKPALERLKARFPVSAARL



query protein

similar structure
similar function

MLIGVCRVIISIPESFSLKEKRKVKRSIVDKVRSKFNVSIAEV



homologous protein

What is the goal of sequence searching?

1. Find a homologous sequence that with known functions or structure
2. Predict functions and structure from homologous protein

How do we proceed to test if two proteins are homologous?

1. Compute sequence alignment => score
2. Compute P-value for score
3. Significant?

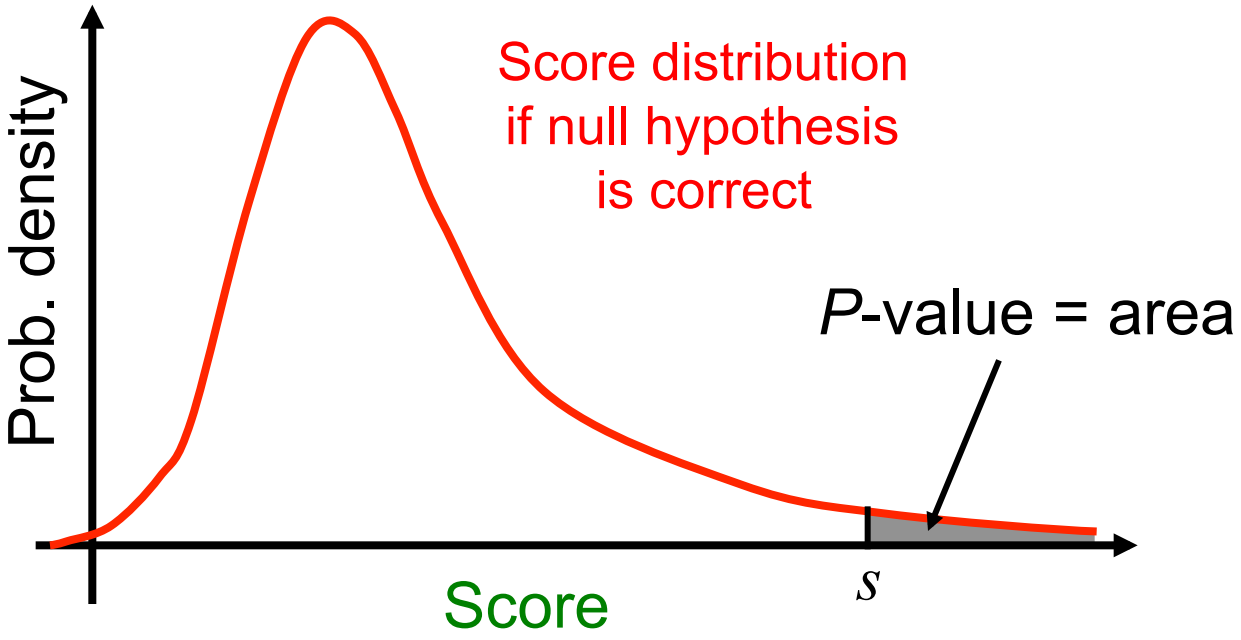
What is the P-value for score S ?

Given: a *null hypothesis* (boring "hypothesis of randomness") and a score ("test statistic") with known distribution under the null hypothesis

Can null hypothesis can be rejected for score S?

***P*-value** = the probability to obtain a score as observed *or more extreme*, under the null hypothesis.

A small *P*-value (e.g. < 0.01) indicates the null hypothesis can be rejected.



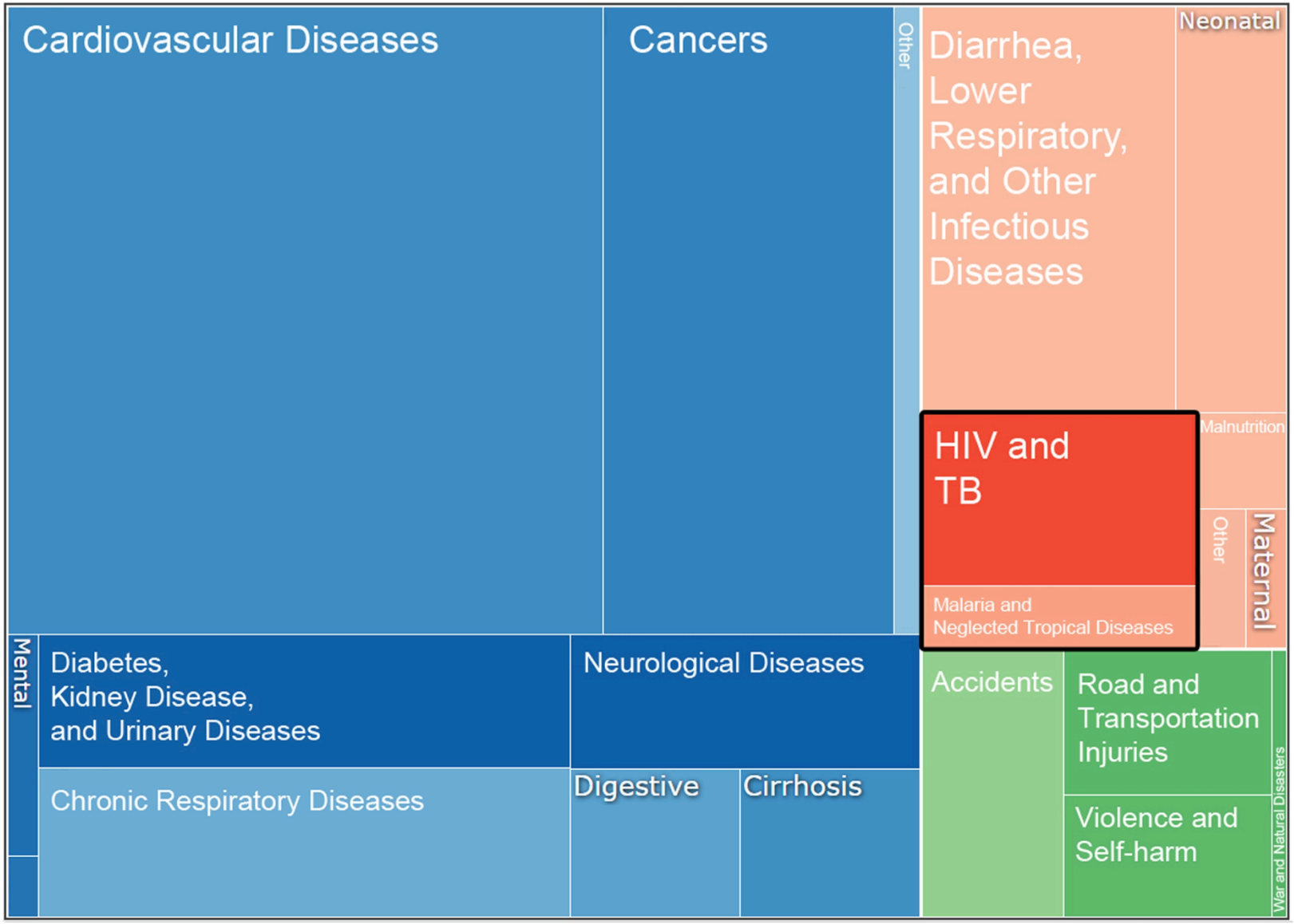
Suppose the time complexity of your algorithm is $O(N^3)$ and it takes 1 min to run on N=100 data points.

How long will it run on 1000 points?

Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, GWAS, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

Cardiovascular diseases responsible for 1/3 of all human deaths in 2015

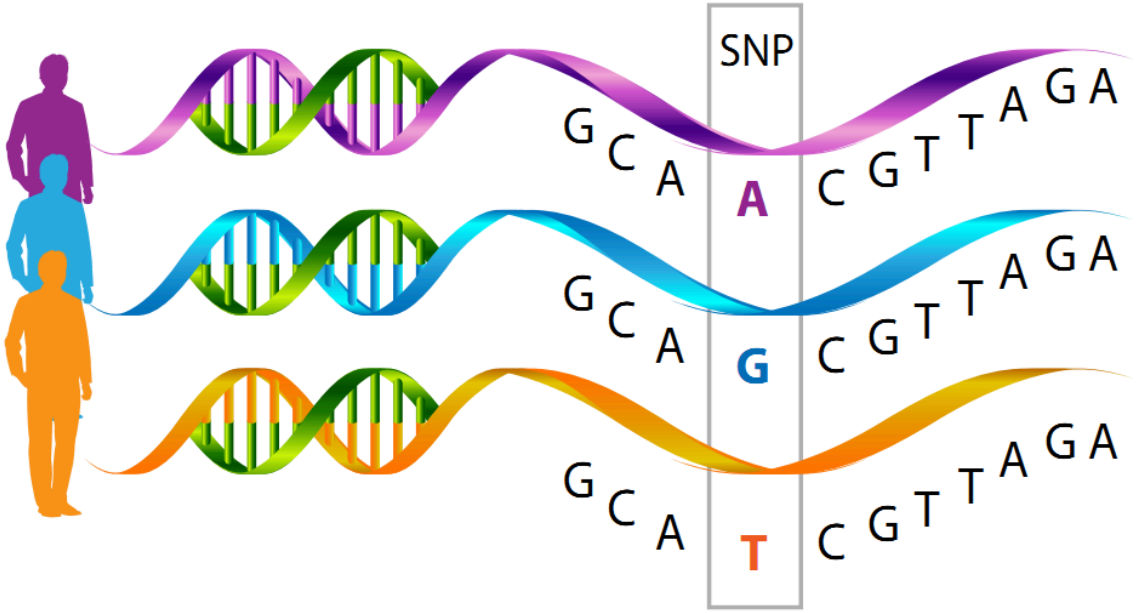


Rare versus complex diseases

Rare (Mendelian) diseases	Complex diseases
	(Coronary artery disease, stroke, diabetes, Alzheimer‘s, depression, Parkinson‘s, MS, Crohn’s, ...)
Monogenic: usually mutation in coding region of a single gene	Polygenic: 10s – 1000s genes involved
Mutation has high effect size	Genetic variants have low effect sizes
Highly heritable	Low heritability
Environment / lifestyle has little effect	Environment / lifestyle has strong influence on risk
Affect young children	Affects older adults
Rare	Frequent
	Main cause of death and disability!

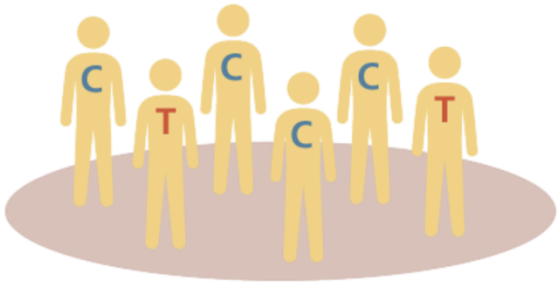
SNP = single-nucleotide polymorphism: position in genome with ≥2 variants

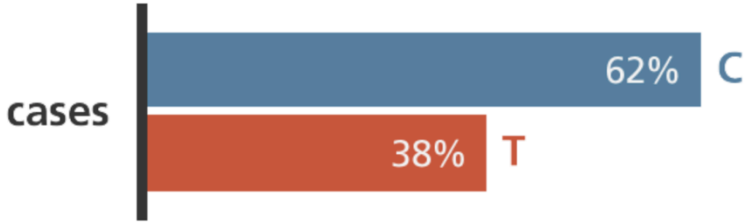
poly-morphism = many-forms



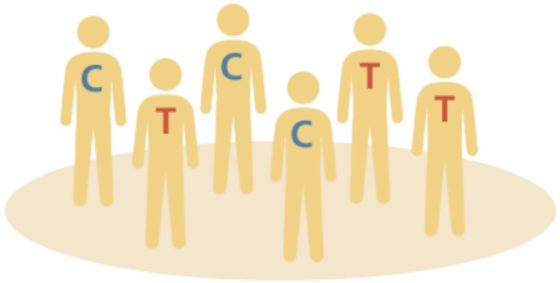
- Humans differ on average at a few million (0.1%) positions.
- About 1 million SNPs occur with >5% in one of major populations.
- The set of nucleotides at these 1 million positions is a summary of a person's genome ⇒ **genotype**

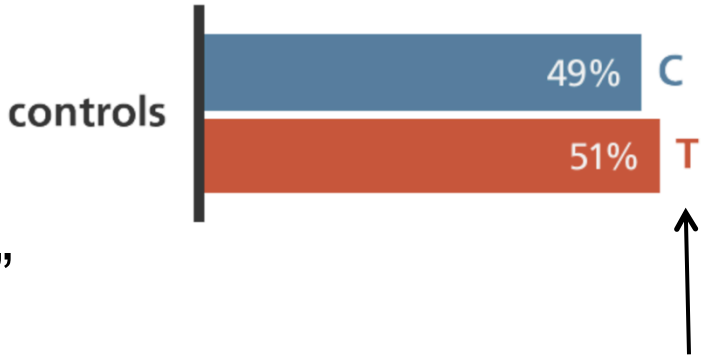
Genome-wide association studies (GWAS): study origin of complex diseases





1000 patients with disease



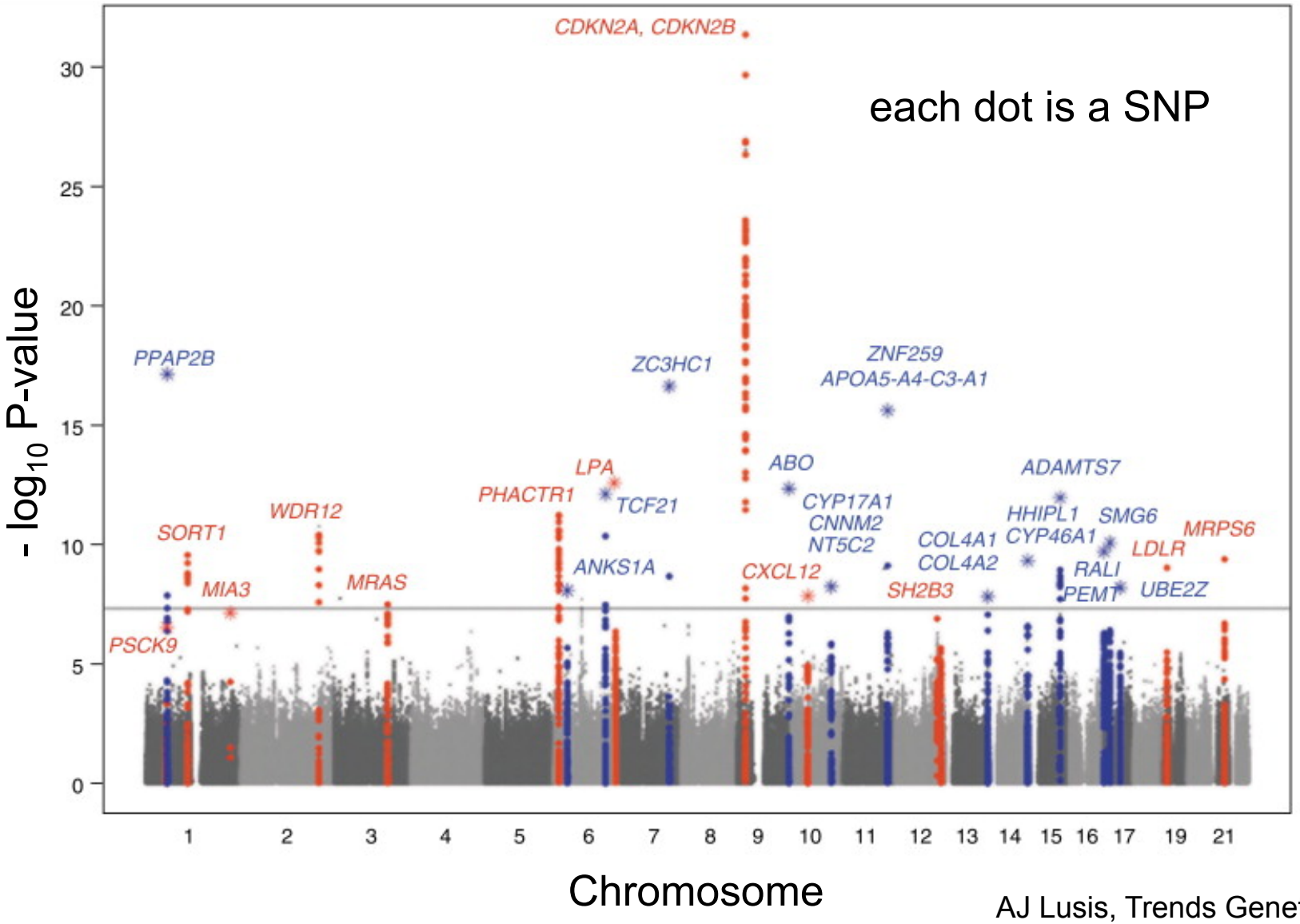


1000 healthy “controls”

This particular SNP is statistically associated with disease risk.
Can it tell us something about how/why the disease develops?

Major challenge: correlation versus causation!

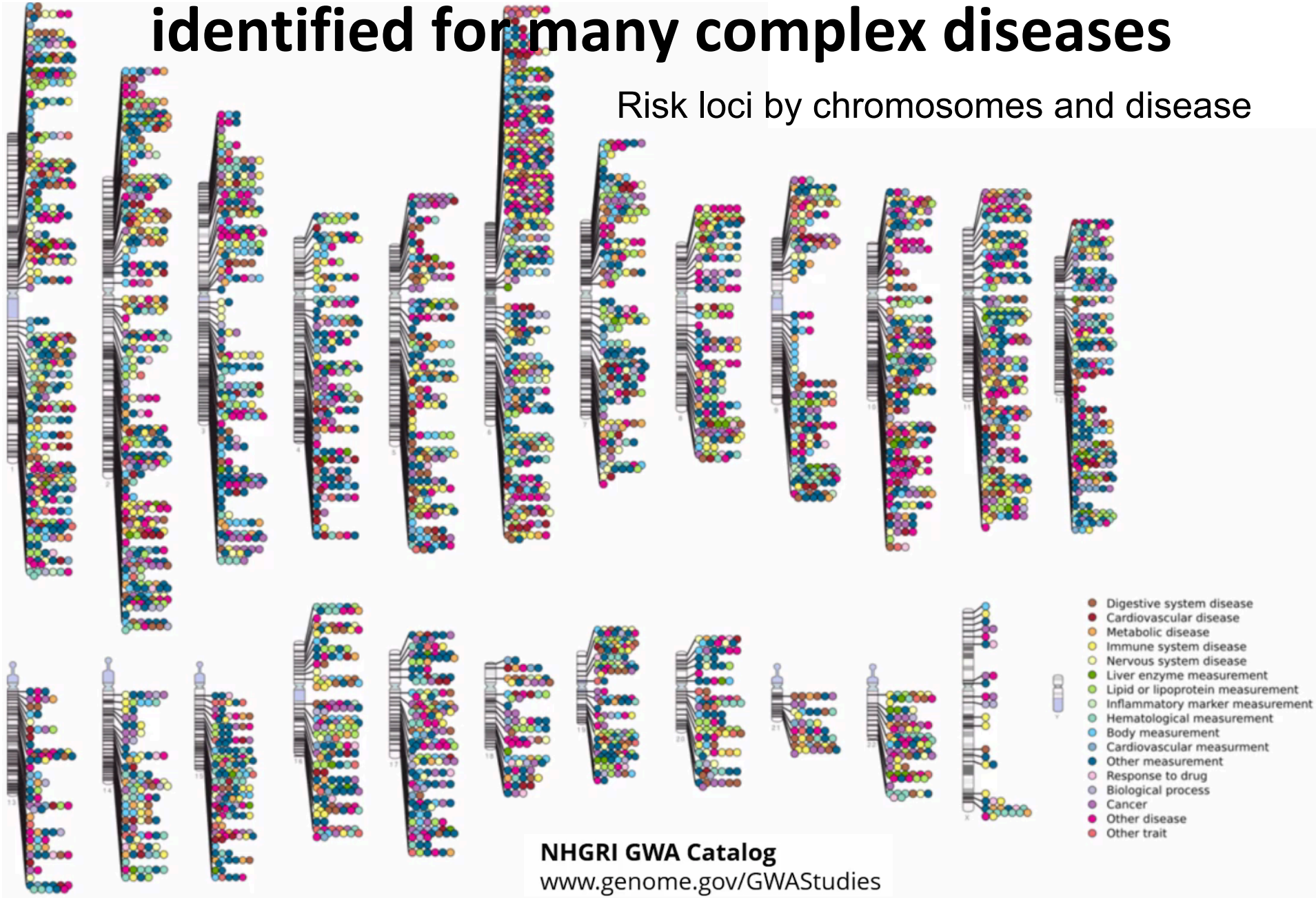
GWAS compute for each SNP a P-value for association with the disease (case/control)



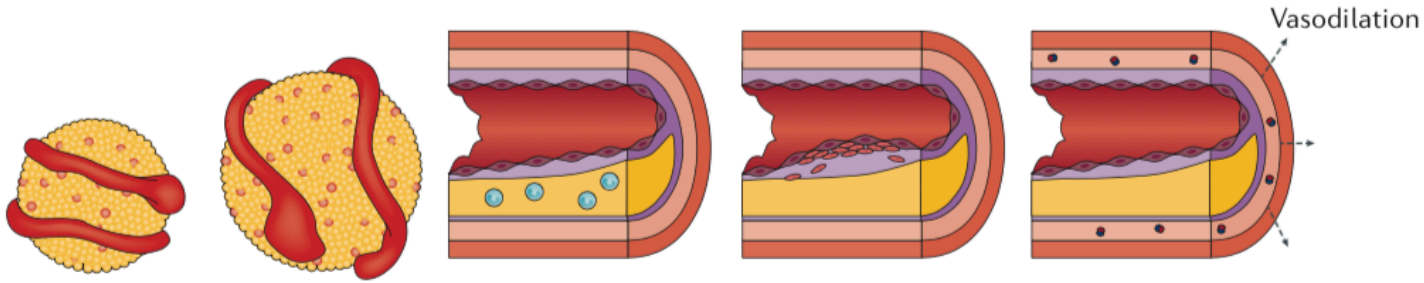
AJ Lusis, Trends Genet. 2012

20 - 100 risk loci per disease have been identified for many complex diseases

Risk loci by chromosomes and disease



50% of genes near risk SNPs for cardiovascular disease have no known mechanism of action



LDL cholesterol and lipoprotein(a)	Triglyceride-rich lipoproteins	Inflammation	Cellular proliferation and vascular remodelling	Vascular tone and nitric oxide signalling	Pathway remains uncertain	
				hypertension		
• *SORT1* • *PCSK9* • *APOB* • *ABCG5–ABCG8* • *LPA* • *LIPA* • *LDLR* • *APOE*	• *LPL* • *TRIB1* • *APOA5–APOA4–APOC4–APOA1* • *ANGPTL4*	• *IL6R* • *CXCL12*	• *COL4A1–COL4A2* • *MIA3* • *REST–NOA1* • *ZC3HC1* • *9p21* • *PDGFD* • *SWAP70* • *KSR2* • *ADAMTS7* • *BCAS3* • *FLT1* • *SMAD3*	• *GUCY1A3* • *EDNRA* • *NOS3*	• *PPAP2B* • *WDR12* • *VAMP5–VAMP8–GGCX* • *ZEB2* • *AK097927* • *MRAS* • *SLC22A4–SLC22A5* • *ANKS1A* • *PHACTR1* • *TCF21* • *KCNK5* • *PLG* • *HDAC9* • *ABO* • *SVEP1* • *SH2B3*	• *CYP17A1–CNNM2–NT5C2* • *KIAA1462* • *ATP2B1* • *HHIPL1* • *MFGE8–ABHD2–SMG6–SRR* • *RASD1–SMCR3–PEMT* • *UBE2Z–GIP–ATP5G1–SNF8* • *PMAIP1–MC4R* • *ZNF507* • *SLC5A3–MRPS6–KCNE2* • *POM121L9P–ADORA2A* • *FURIN–FES*

statins

Aspirin

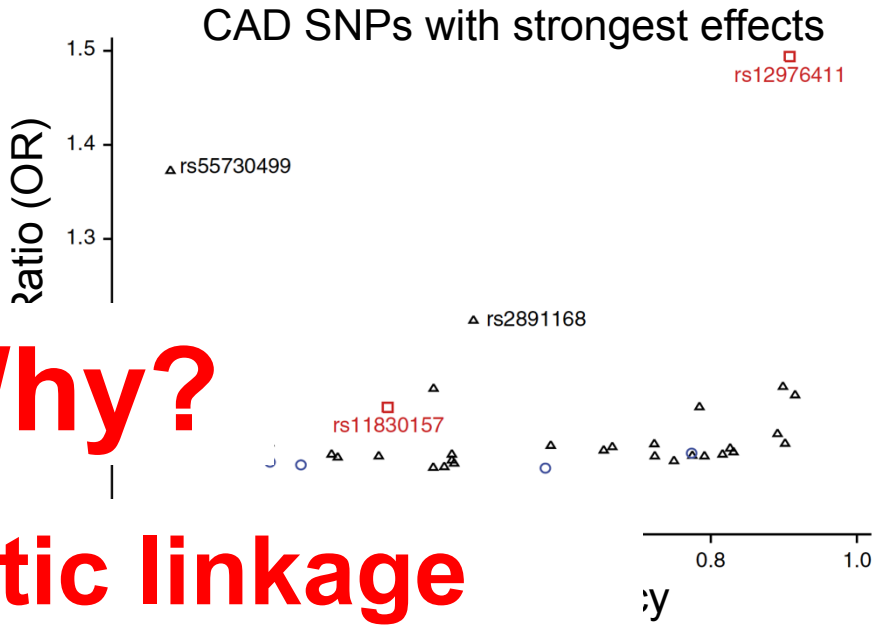
ACB inhibitors
ARBs

Preventive drugs

?

GWAS have discovered many new loci... but fewer <u>new pathways</u> than hoped for

CAD GWAS identified ~100 statistically significant risk loci in unbiased way![1]



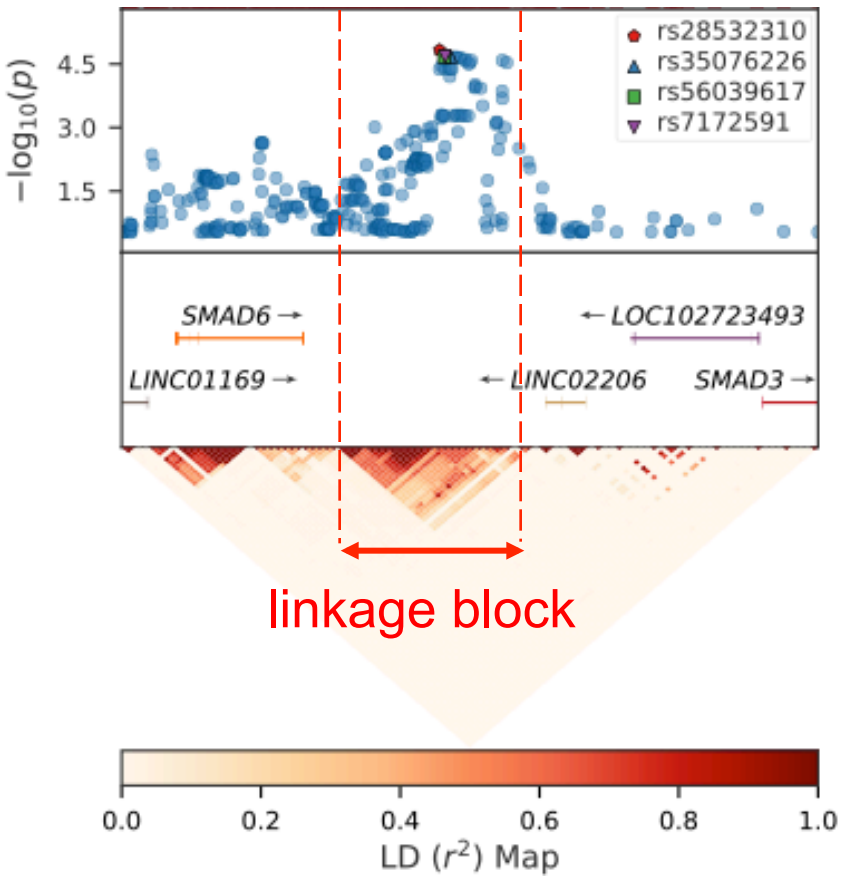
Why?

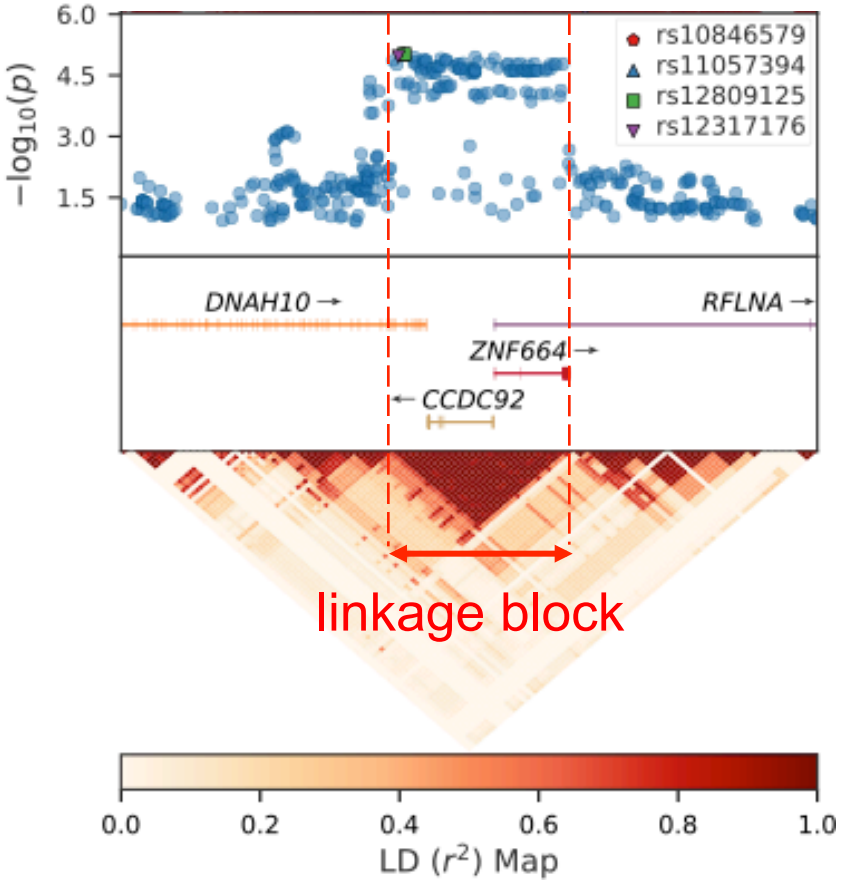
1. Genetic linkage

2. Which gene's regulation is affected?

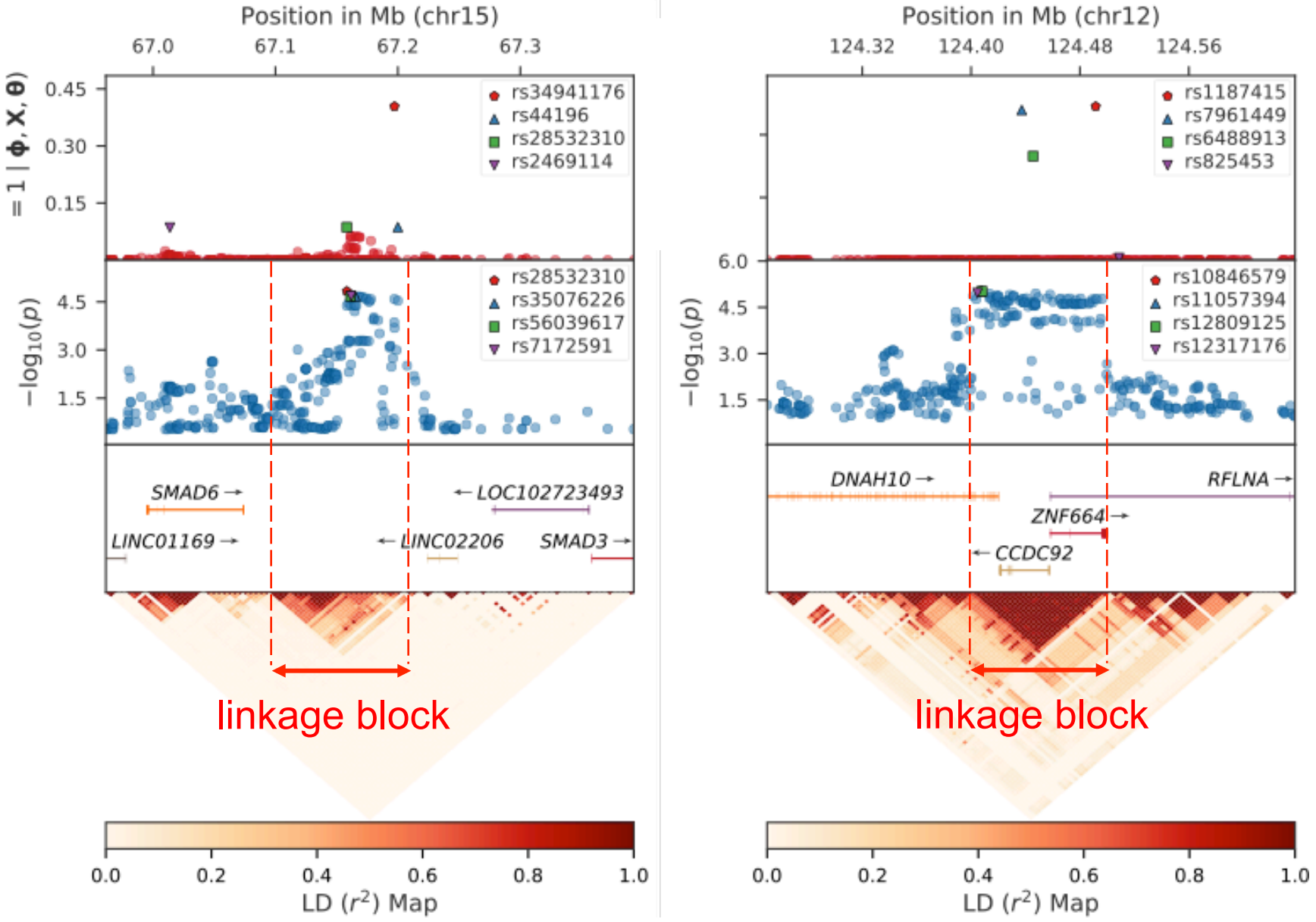
3. Weak effect sizes

Genetic linkage: most SNPs with low P-values not causal (correlation ⇏ causation)





Multiple regression can explain away correlation of most noncausal SNPs



Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, GWAS, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting and regularization
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

Logistic regression is a popular approach for identifying associated SNPs

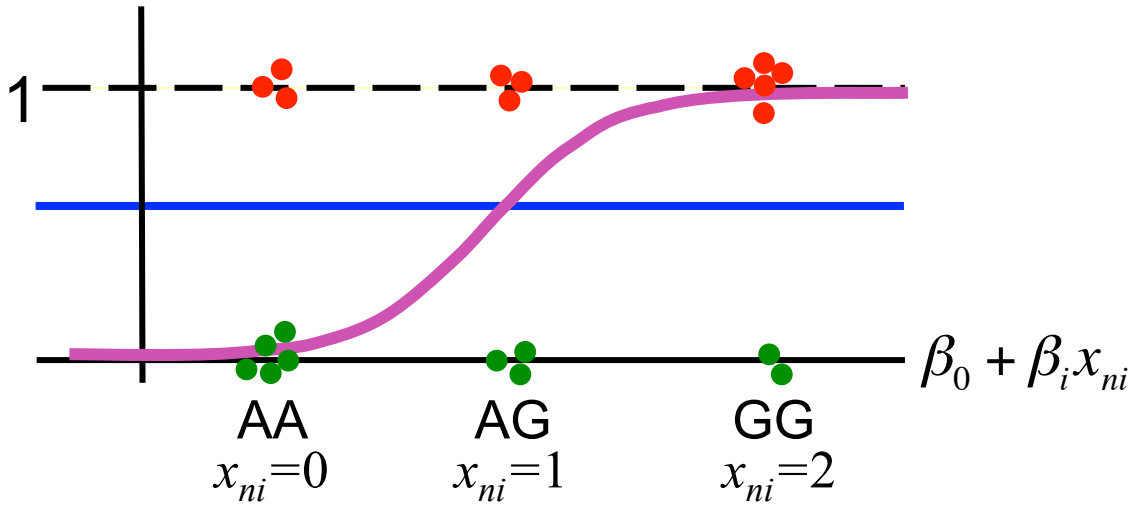
Patient n:
diseased $\varphi_n = 1$
healthy $\varphi_n = 0$

minor alleles of SNP i in patient n

logistic function

effect of SNP i on disease risk

$$p(\phi_n = 1 \mid x_{ni}) = \mathrm{lf}(\beta_o + \beta_i x_{ni})$$



Classical statistics:

Do we need to set $\beta_i \neq 0$…

or is the data compatible with the null hypothesis $\beta_i = 0$?

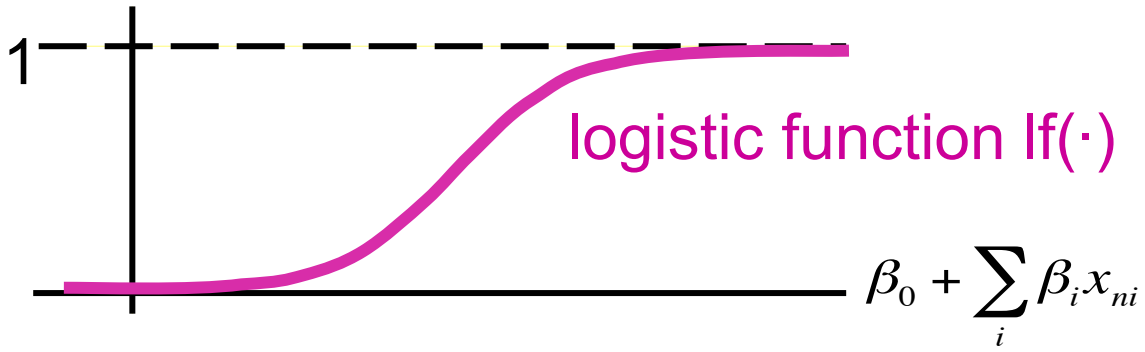
Joint treatment of SNPs automatically takes account of linkage structure

$$p(\phi_n = 1 \mid x_{ni}) = \mathrm{lf}(\beta_o + \beta_i x_{ni})$$

independent SNPs

$$p(\phi_n = 1 \mid \mathbf{x}_n) = \mathrm{lf}\left(\beta_0 + \sum_{\text{SNP } i} \beta_i x_{ni}\right)$$

joint analysis of SNPs



Most SNPs are only associated with disease through the correlation with a causal SNP. Their association will be "**explained away**" by the causal SNP ⇒ their β_i will become 0!

Big data in biomedicine & key concepts

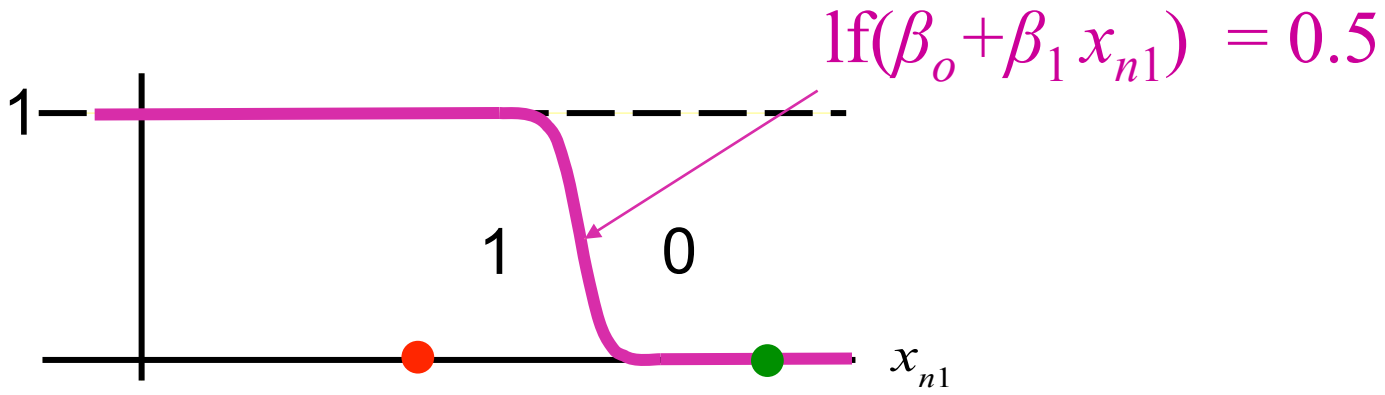
- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, GWAS, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

The $p > N$ problem and overtraining

p = number of model parameters (β_i)
N = number of training data points (patients)

When $p > N$, the training data can be perfectly fit. However, no sensible model has been learned and performance on unseen data will be abysmal.

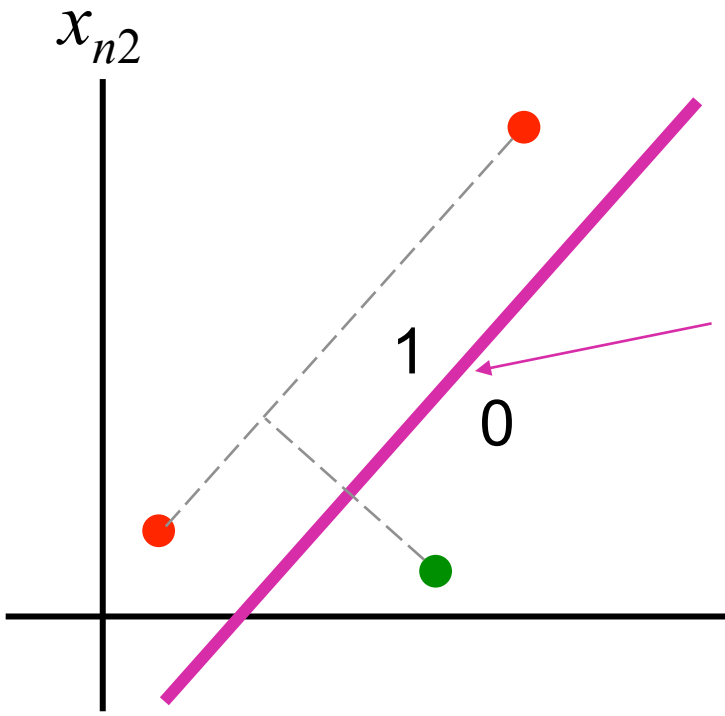
Two points can be perfectly fit with two parameters (β_0 and β_1)



The $p > N$ problem and overtraining

p = number of model parameters (β_i)
N = number of training data points (patients)

When $p > N$, the training data can be perfectly fit. However, no sensible model has been learned and performance on unseen data will be abysmal.



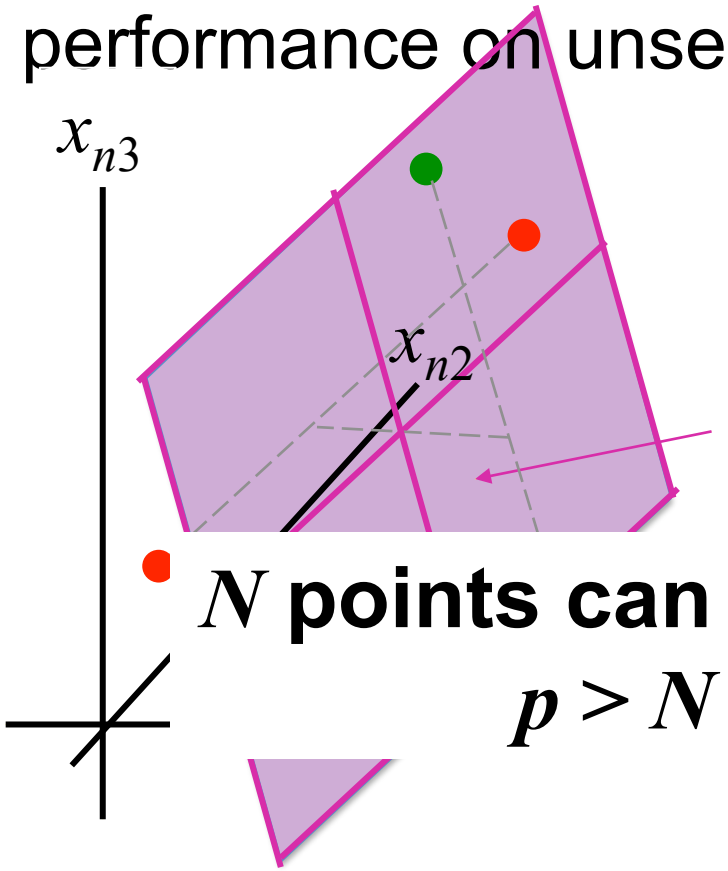
Three points can be perfectly fit with three parameters (β_0, β_1 and β_2)

$\text{lf}(\beta_o + \beta_1 x_{n1}) = 0.5$

The $p > N$ problem and overtraining

p = number of model parameters (β_i)
N = number of training data points (patients)

When $p > N$, the training data can be perfectly fit. However, no sensible model has been learned and performance on unseen data will be abysmal.

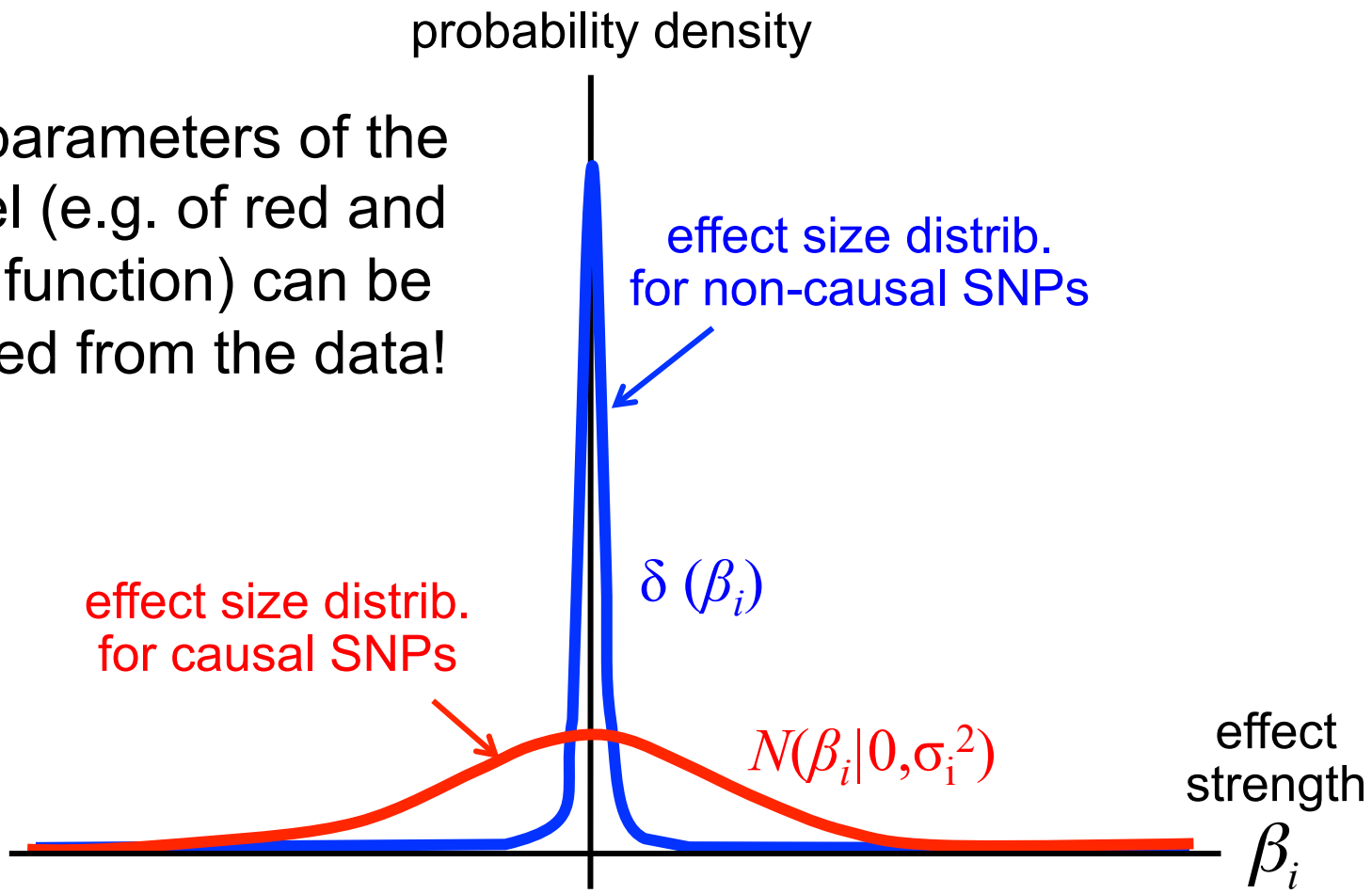


Four points can be perfectly fit with four parameters ($\beta_0, \beta_1, \beta_2$ and β_3)

N points can be perfectly fit with $p > N$ parameters

We can penalize SNPs with non-zero effects ($\beta_i > 0$) to keep their numbers low

The parameters of the model (e.g. of red and blue function) can be learned from the data!



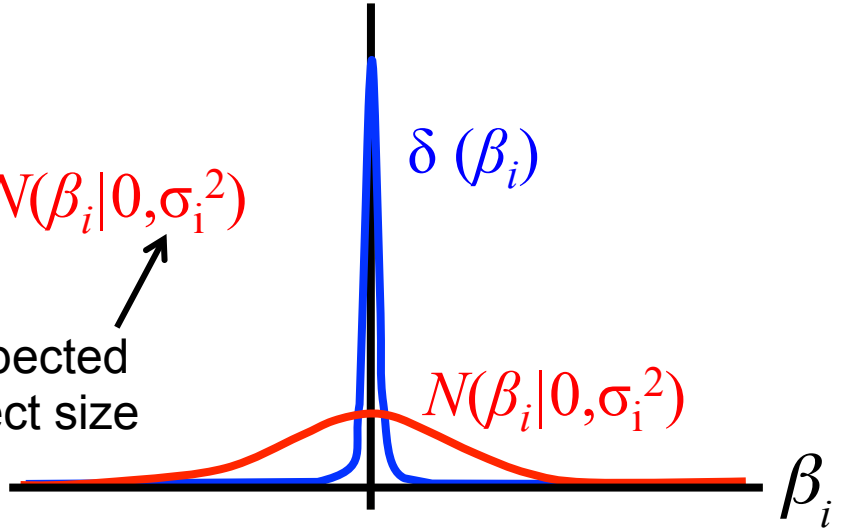
We can make prior probability for a SNP to be causal depend on external info ξ_{if}

Prior distribution:

$$p(\beta_i|\,\pi_i,\,\sigma_i) = (1-\pi_i)\,\delta(\beta_i) + \pi_i\,N(\beta_i|0,\sigma_i^2)$$

prior prob. for causality

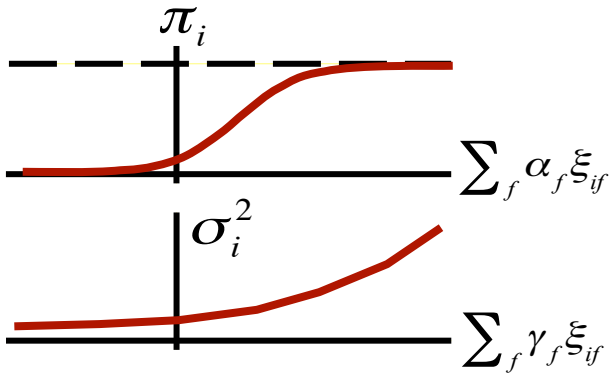
expected effect size



Dependence of parameters on external information ξ_{if}:

$$\pi_i = \mathrm{lf}\left(\sum_f \alpha_f \xi_{if}\right)$$

$$\sigma_i^2 = \exp\left(\sum_f \gamma_f \xi_{if}\right)$$



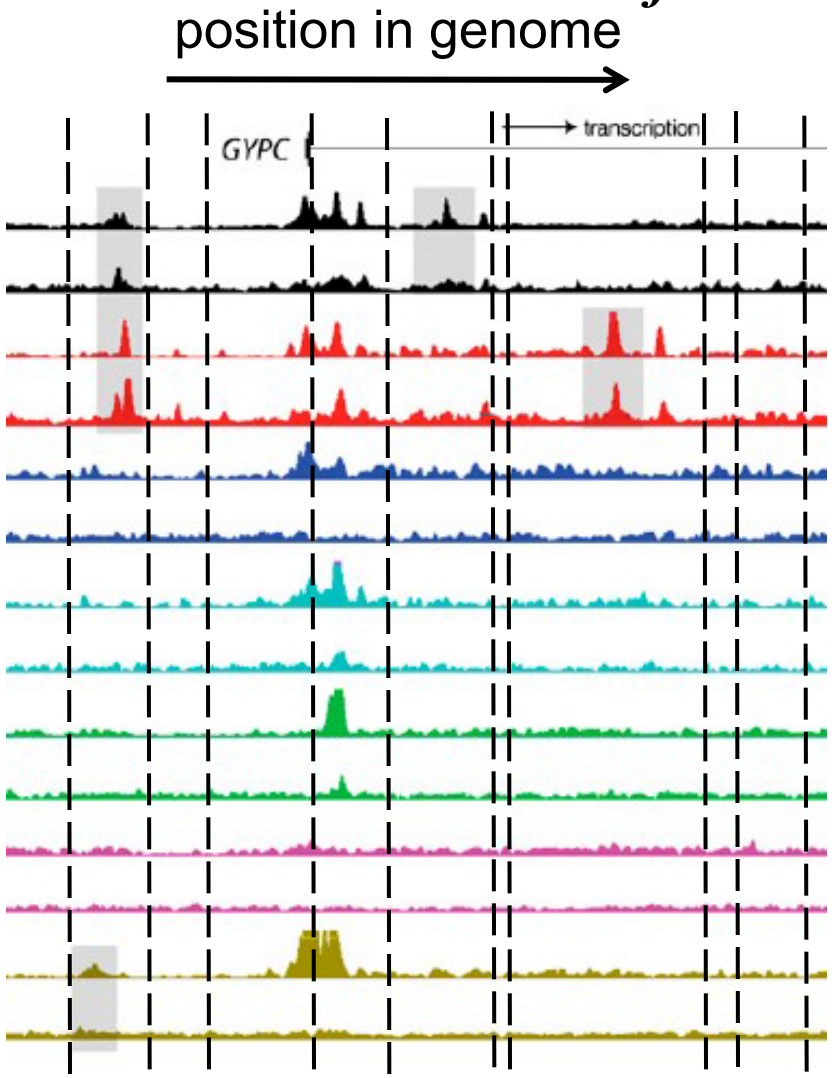
Hyperparameters α_f, γ_f can be learnt from the data!

We can make prior probability for a SNP to be causal depend on external info ξ_{if}

Genome-wide measurements from ENCODE, Roadmap Epigenomics Project, ...

- chromatin state (1-9)
- DNA accessibility in 100s of tissues
- Enhancer predictions in 100s of tissues
- Predicted impact of SNP on DNA accessibility etc.
- Methylation state in 100s of tissues
- …

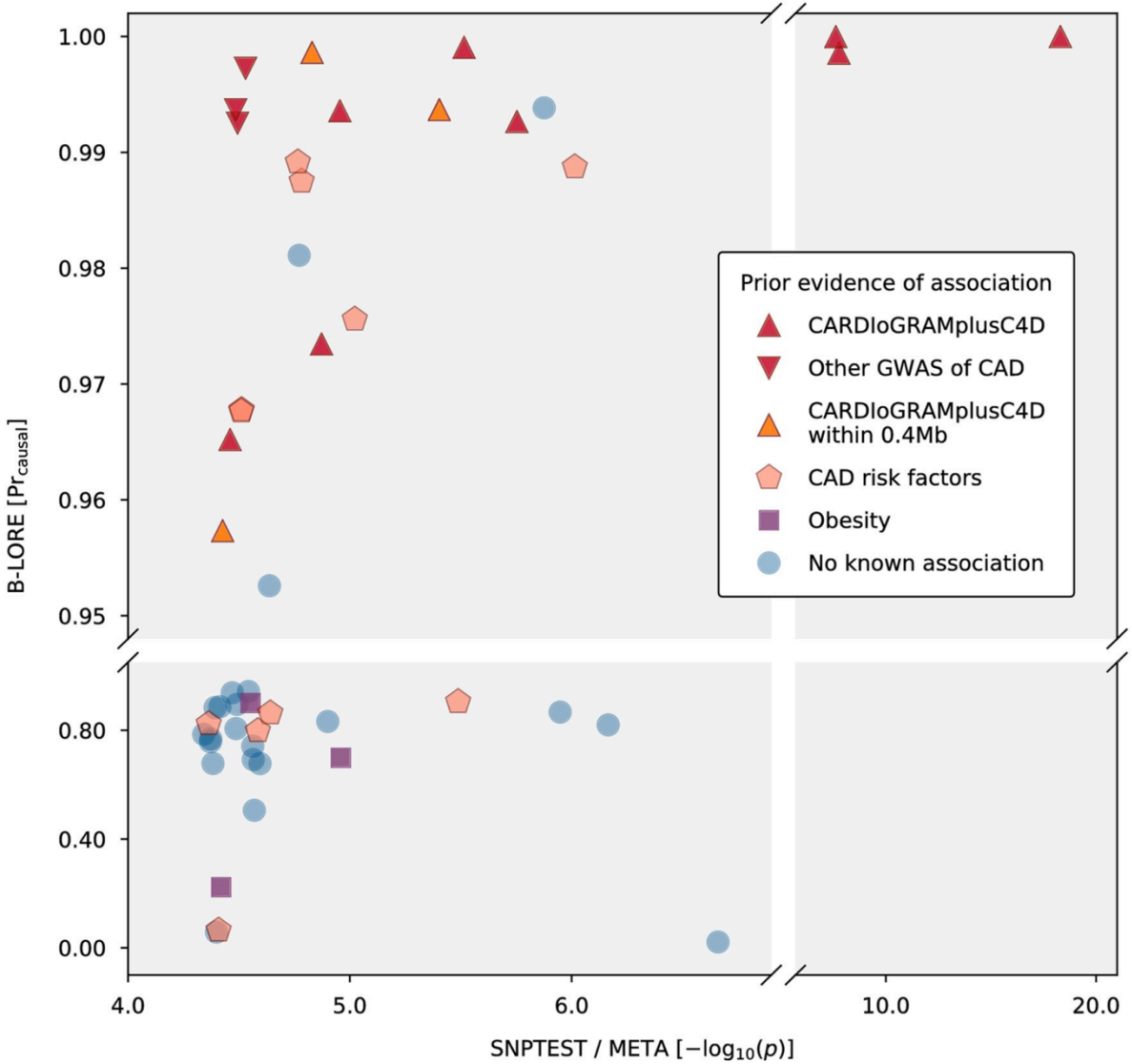
f



Learn which features are predictive of causality of a SNP

Bayesian multiple regression leads to drastically improved prediction of risk loci

using small GerMIFS test dataset with 13k cases/controls



(180k patients)

Banerjee & Soeding et al, PloS Genet (Dec. 2018)

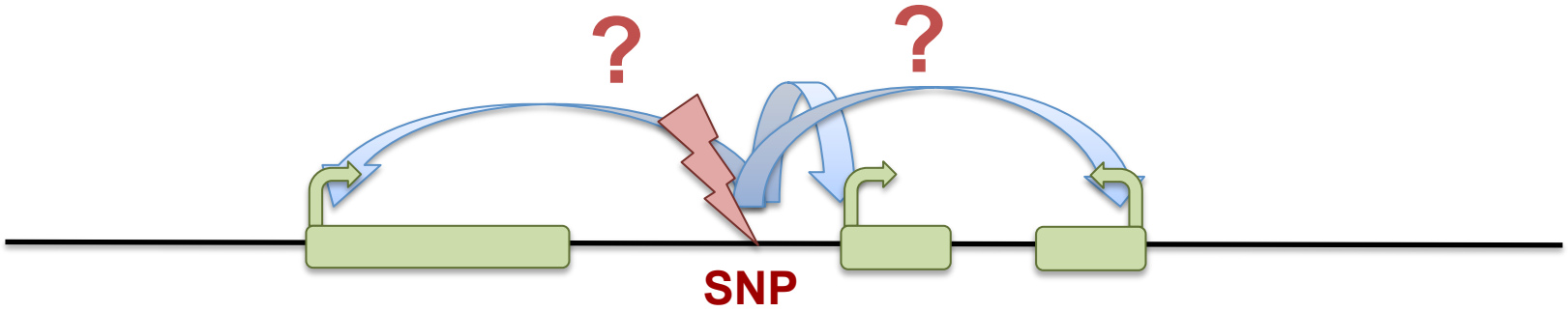
Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, GWAS, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

Most causal SNPs have regulatory effect on a gene often not so nearby

- Previous assumption: affected gene will be nearest to most significan SNP ⇒ most flagged genes probably wrong!

  Why?
- >90% of disease-/trait-associated variants lie within non-coding sequence, **not** in gene
- SNPs can affect expression of genes
- These genes can be far away, need not be nearest gene



- How do we find the true target genes?

We don‘t want SNPs, we want genes and pathways!

Goal: To predict those groups of genes, whose dysregulation in certain tissues predisposes to a higher risk to develop the disease.

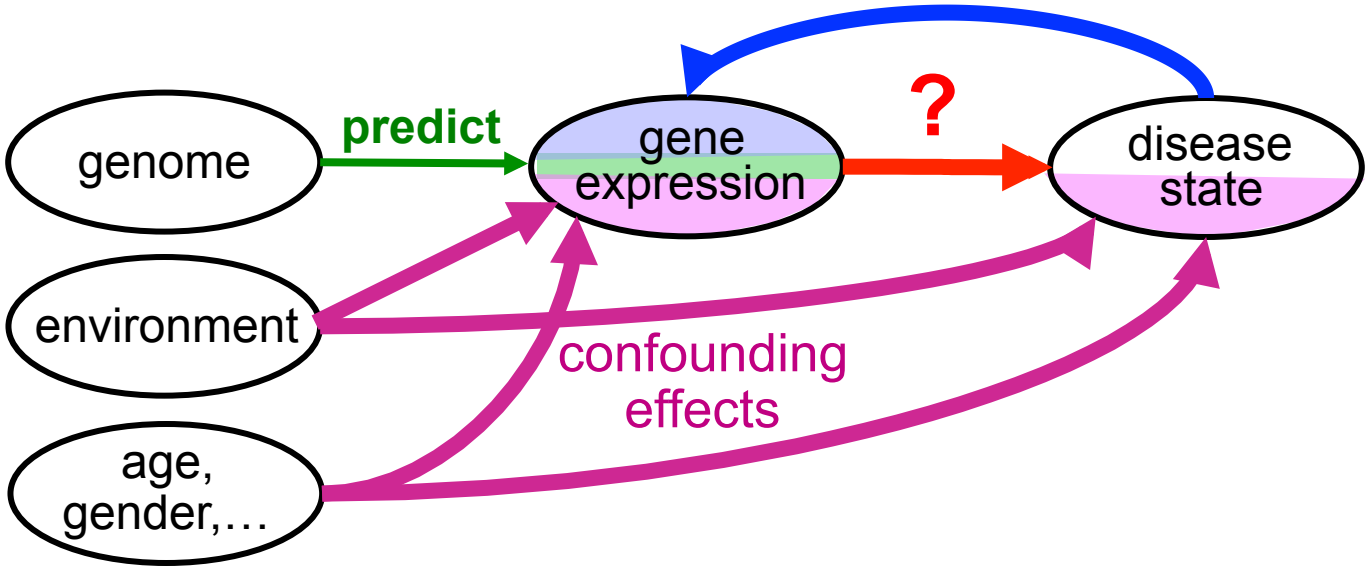
⇓

Understanding is power!

Genes whose upregulation increase risk are prime targets for pharmacological intervention

Why not measure gene expression for hundreds of patients to find out which ones are higher in disease patients?

- **Back-causation** 😳: Gene expression strongly affected by disease state, e.g. drug treatment, reaction of body to disease,...



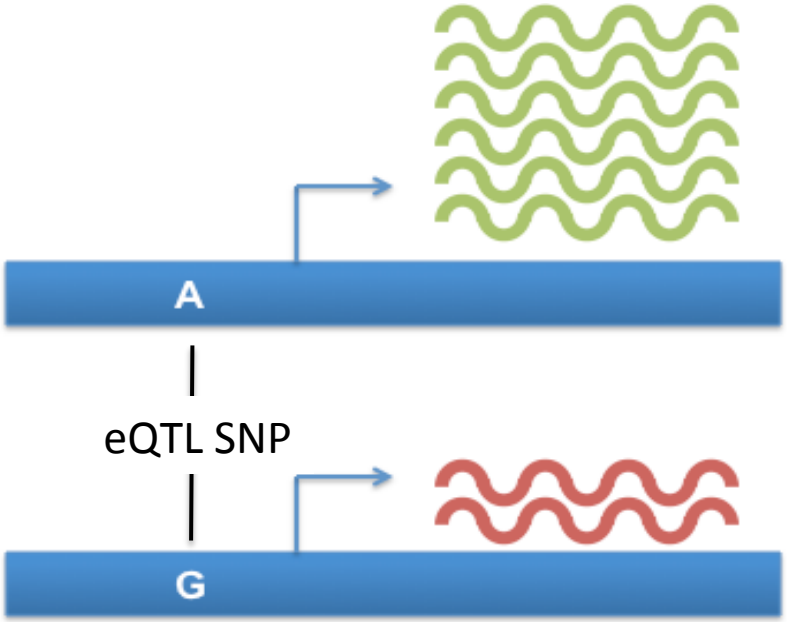
- **Counfounding effects** 😳: Gene expression & disease risk depend strongly on age, gender, lifestyle etc. This generates correlations w/o causation, e.g. genes expressed in old patients would look like risk genes.

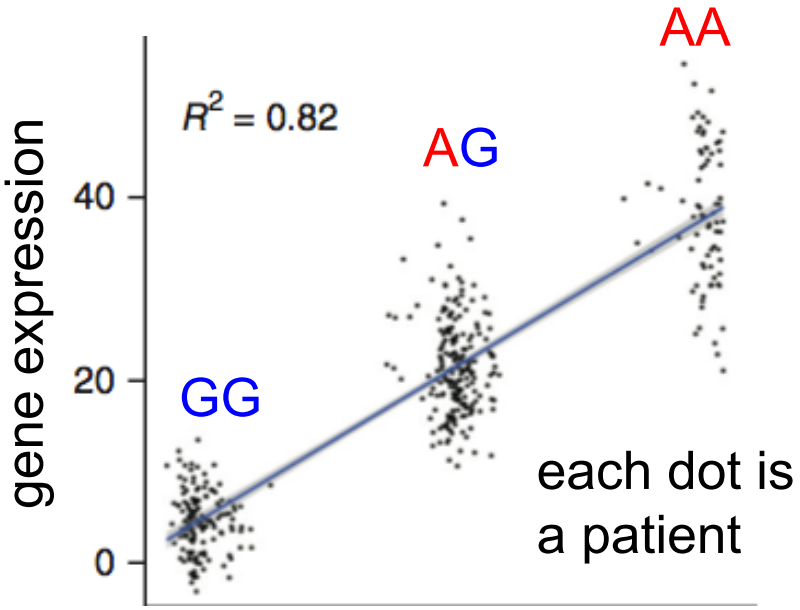
- **Costs + numbers:** can only get gene expression data from <1000 patients. Too few to have enough cases for all common diseases.

Big data in biomedicine & key concepts

- Biology is becoming a quantitative science
- Metagenomics is revolutionizing study of human microbiome and natural environments
- *P*-values: sequence searching & homology inference
- Time complexity can be crucial: sequence clustering
- Quantitative medicine to study origin of complex diseases
- SNPs, GWAS, linkage, and role of regulatory mutations
- Logistic regression, multivariate regression
- p > N and overfitting
- Correlation versus causation, backcausation
- eQTLs and integrative GWAS/eQTL modeling

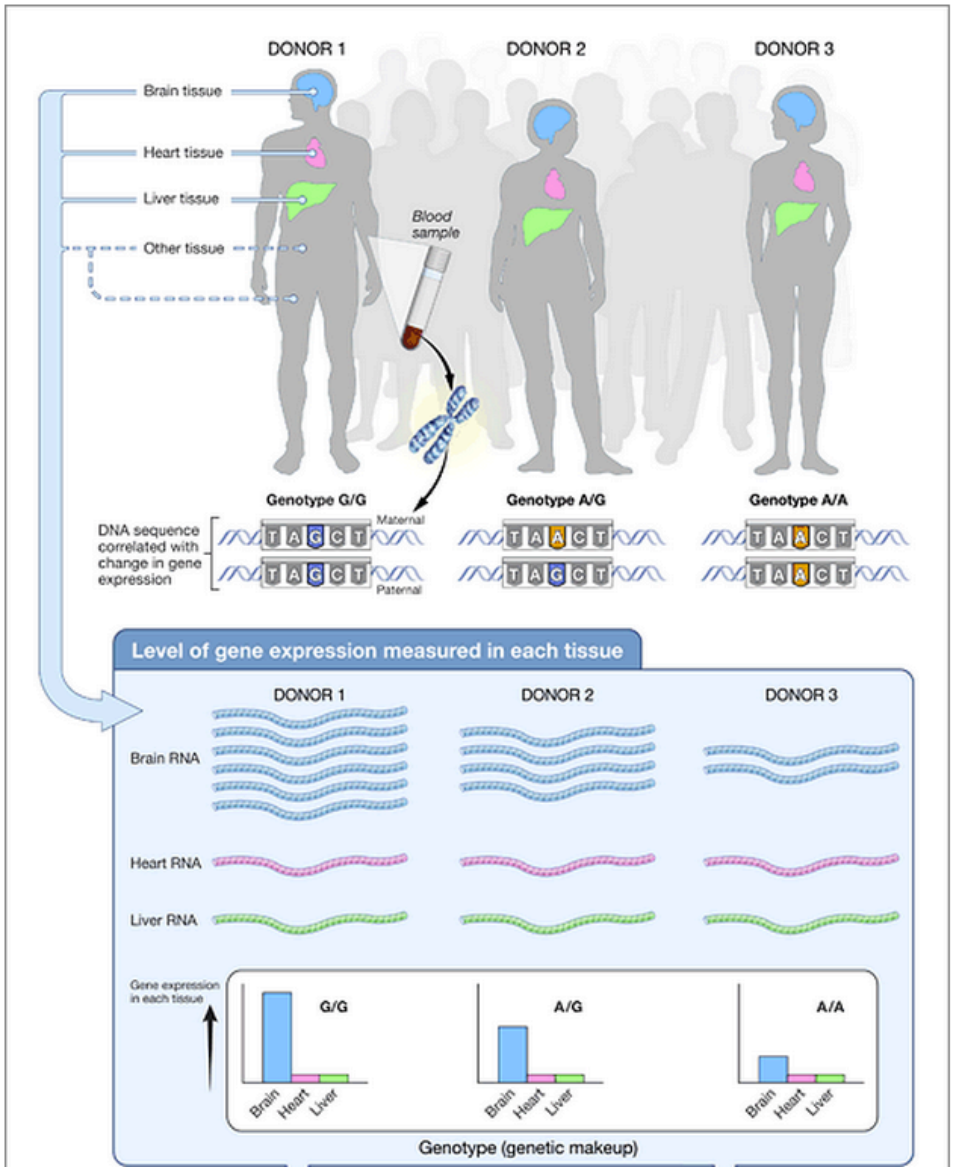
Expression quantitative trait locus (eQTL) SNP affects the expression of a target gene

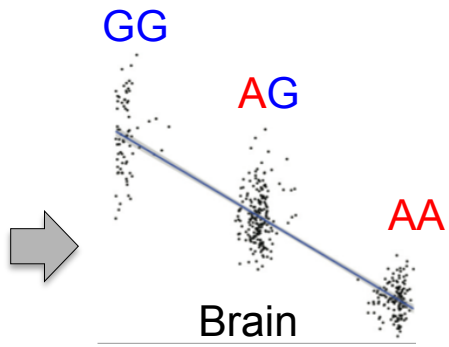




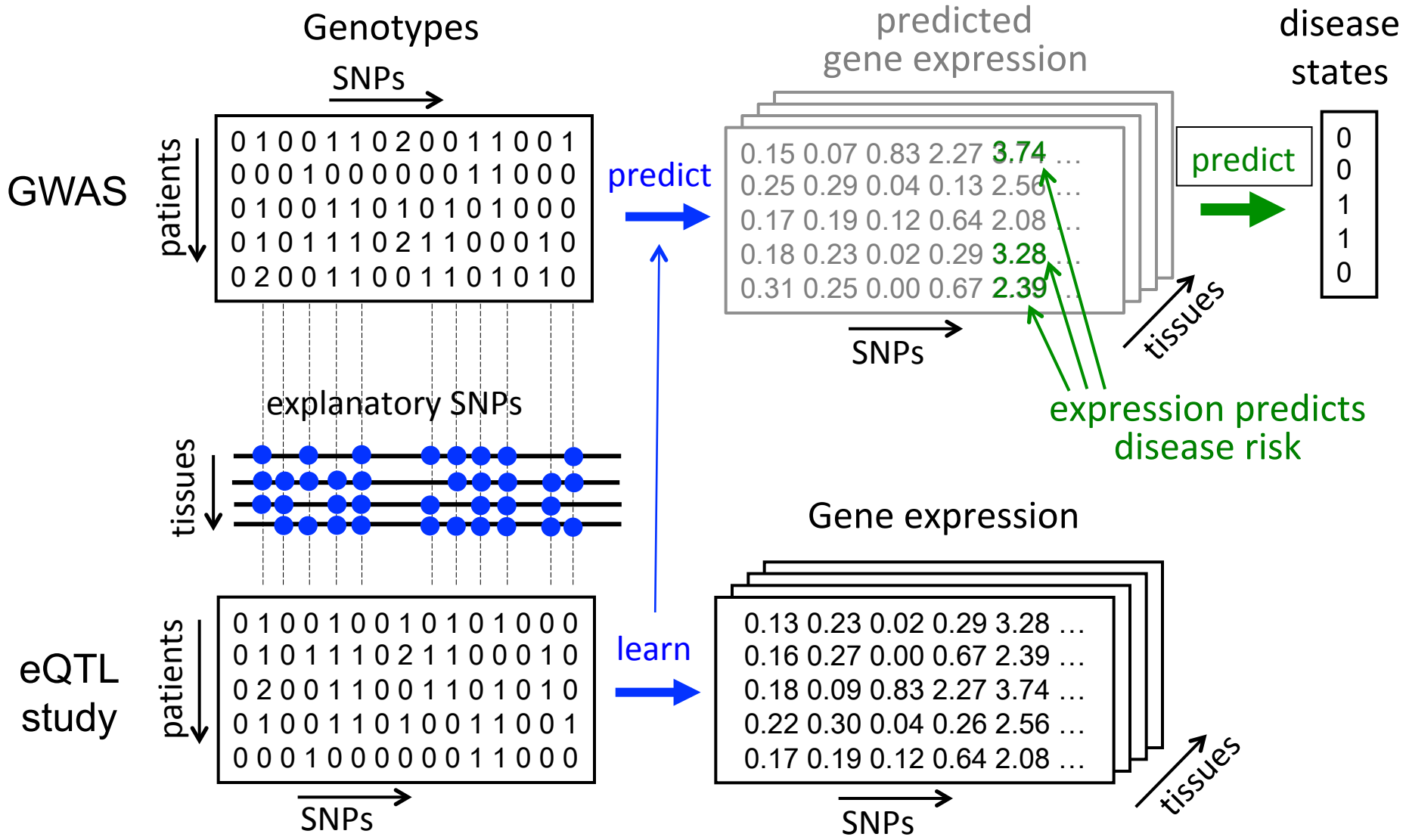
- eQTL SNPs can be identified genome-wide by measuring genotypes and gene expressions for many patients
- SNP may only have effect in particular tissue or cell type
- Correlation ≠ causation

GTEx: ~500 patients, each with gene expression measurements from ~20 tissues

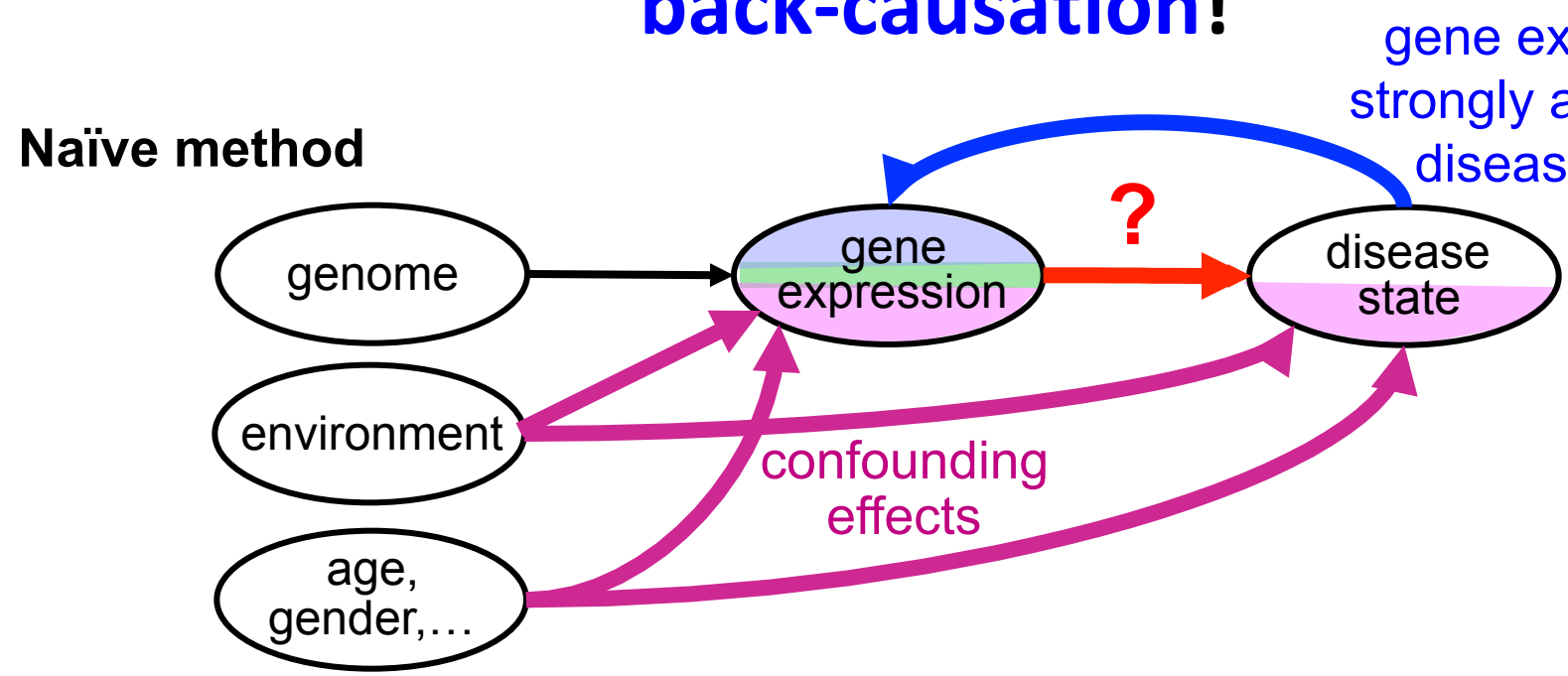


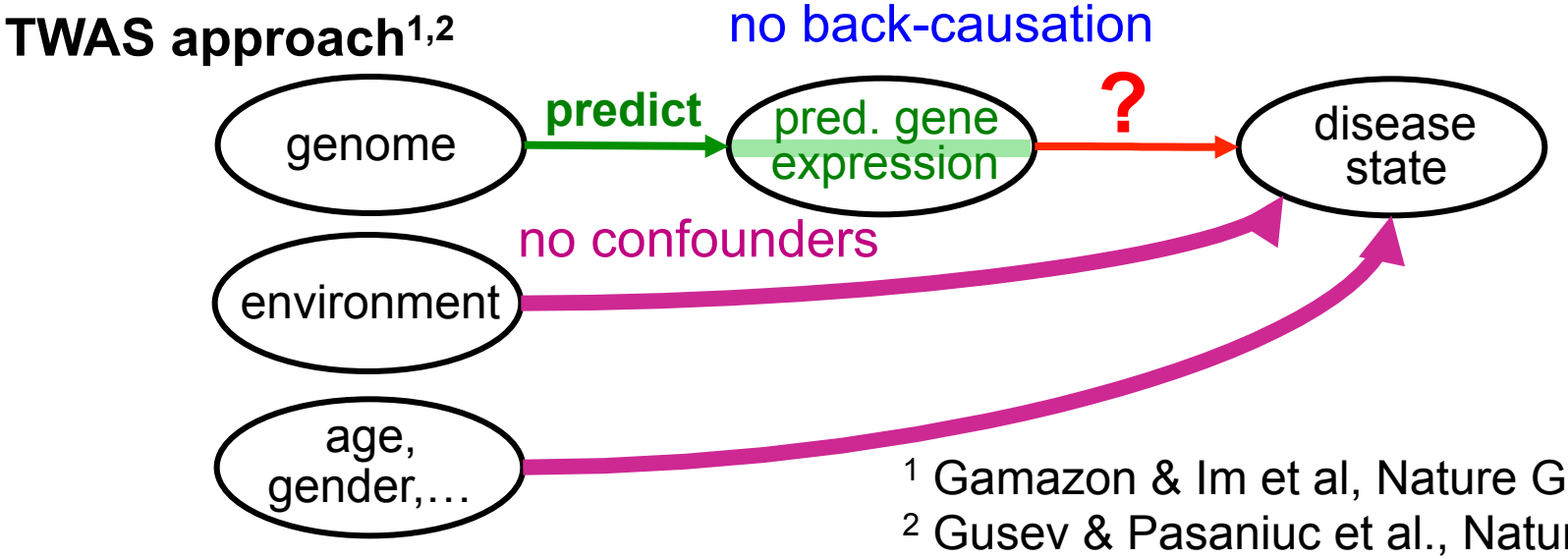


Joint GWAS-eQTL regression analysis aggregates signal from all SNPs



Joint GWAS-eQTL regression analysis eliminates back-causation!





[1] Gamazon & Im et al, Nature Genet. 2015
[2] Gusev & Pasaniuc et al., Nature Gent. 2016

What is a complex disease?

What is a single-nucleotide polymorphism?

What is a genome-wide association study?

What is genetic linkage?

Why is it a problem for genome-wide association studies?

Why does multiple regression help against genetic linkage?

What is overtraining?

What can we do against it?

Many thanks to my team...

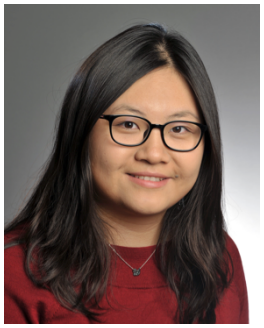
Tools for metagenomics, protein structure & function

Martin Steinegger (− Sept 2018)

Milot Mirdita

Annika Seidel

Dr. Eli Levy Karin



Rushi Zhang

Christian Roth

Transcription / transcriptomics / quant. medicine

Wanwan Ge

Salma Sohrabi-Jahromi

Niko Papadopoulos

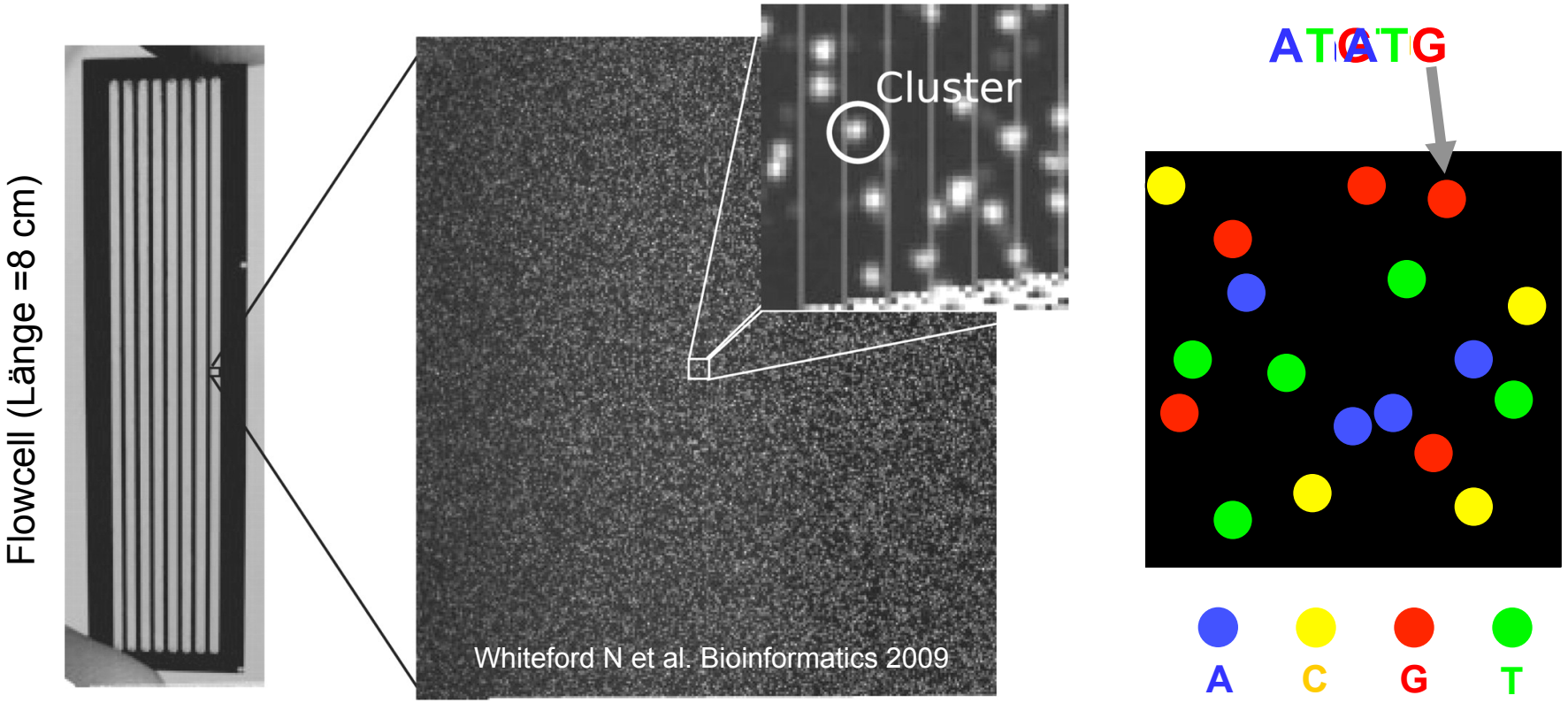
Dr. Franco Simonetti

Dr. Saikat Banerjee

... and to you for your attention!

Principle of Illumina sequencing

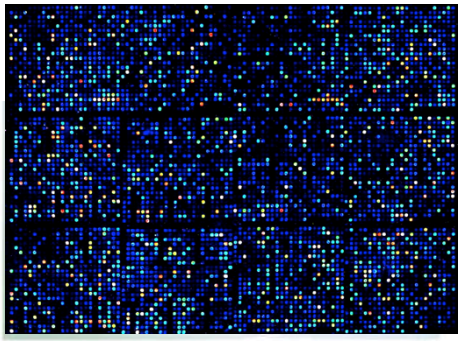
- DNA fragments of ~400 base pairs length are attached to a glass plate. Each fragment is amplified ~1000-fold into a cluster of single-strand DNAs
- Per cycle, nucleotides marked with 4 different fluorescence dyes are flowed over the clusters. In each cluster, the nucleotide that matches for extending the single strang attaches to the cluster.
- Read-out of the color at each cluster identifies the nucleotide in the DNA

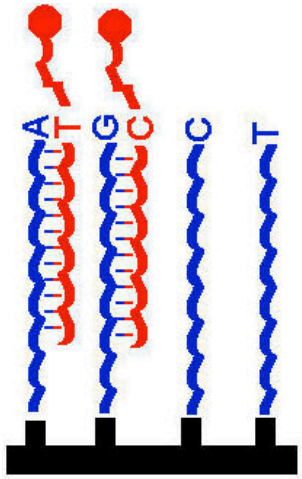


Patients can be genotyped at 1 million common SNPs with SNP arrays

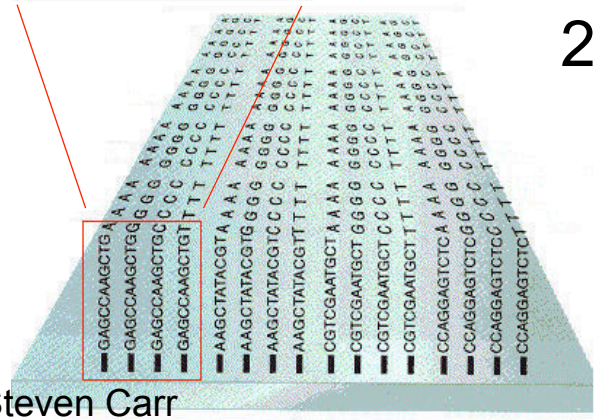
≥ 5% minor allele frequency

4. Patient's variants are read off from the fluorescence image.





3. Fragments pair up only with perfectly complementary probes..



2. Patient DNA is fragmented and fluorescence labeled.

1. For each SNP, the array has oligo-DNA probes for all SNP variants (A, C, G, T).