From genome-wide association studies to disease relationships

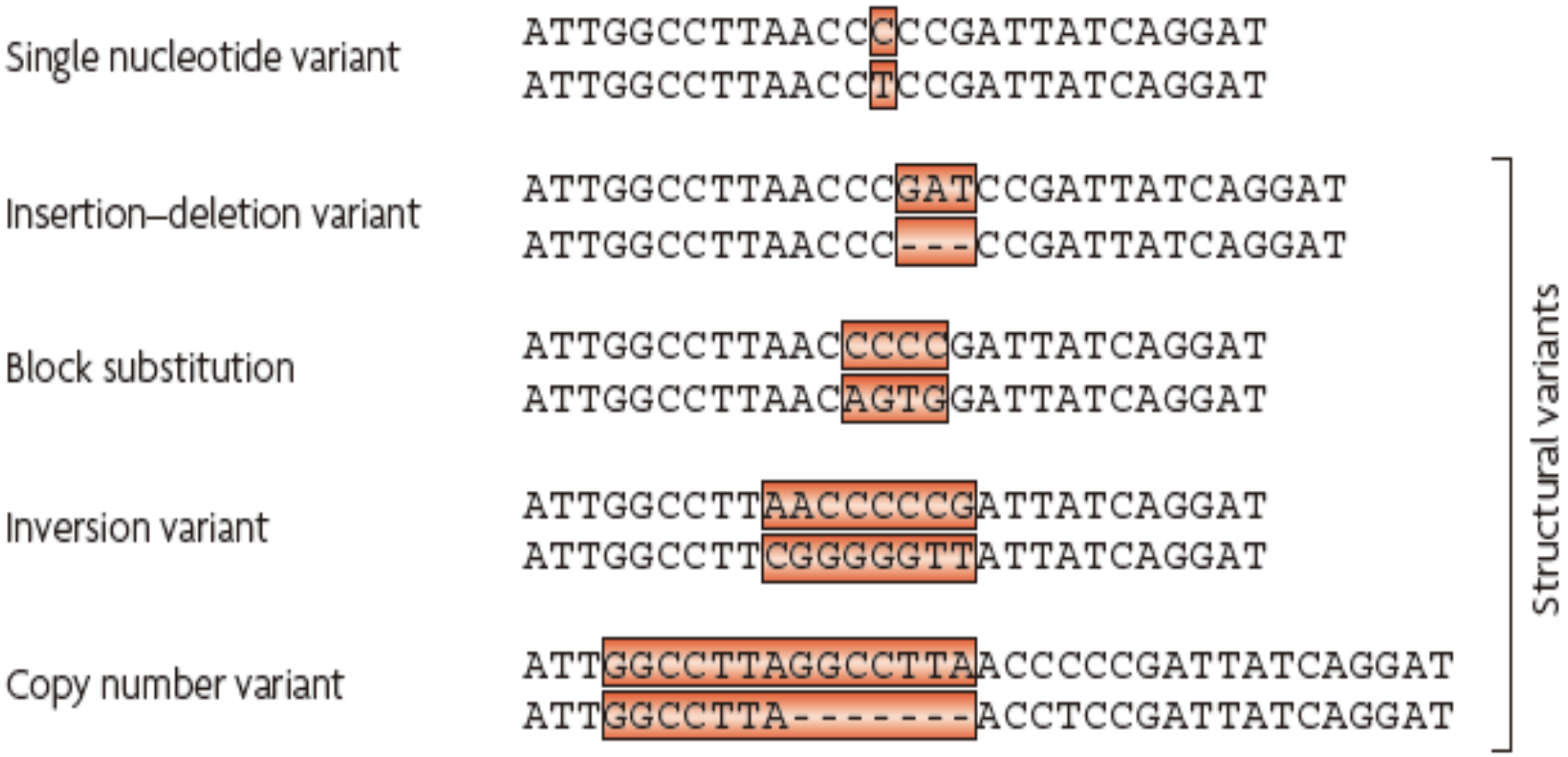
Liqing Zhang
Department of Computer Science
Virginia Tech

Types of variation in the human genome

- SNPs (single nucleotide polymorphisms)
- Insertions
- Deletions
- Duplications
- Rearrangements

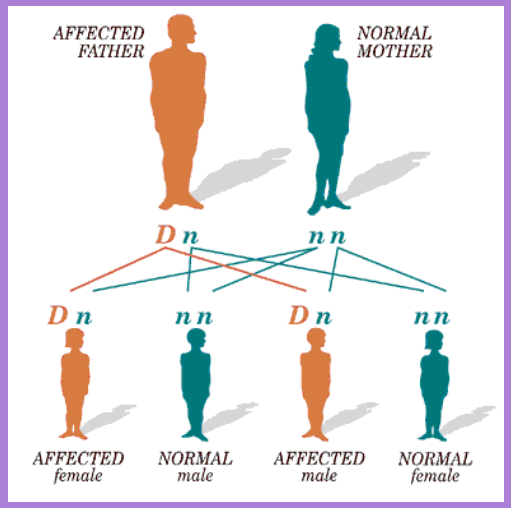
Types of variation in the human genome

- Small- vs. large-scale



Why do we care about variation?

underlie phenotypic differences



cause inherited diseases

allow tracking ancestral human history

SNPs in the human genome

- SNP location with respect to genes
  - ✓ Exonic
    - nonsynonymous: lead to amino acid substitutions
    - synonymous: do not alter amino acid, but they might affect mRNA stability and alter splicing signals
  - ✓ Introns, regulatory and gene-distant regions
    - affect gene regulation
    - splicing
- Effect: likely give rise to the observable phenotypic differences within and between populations, including disease susceptibility and outcome.

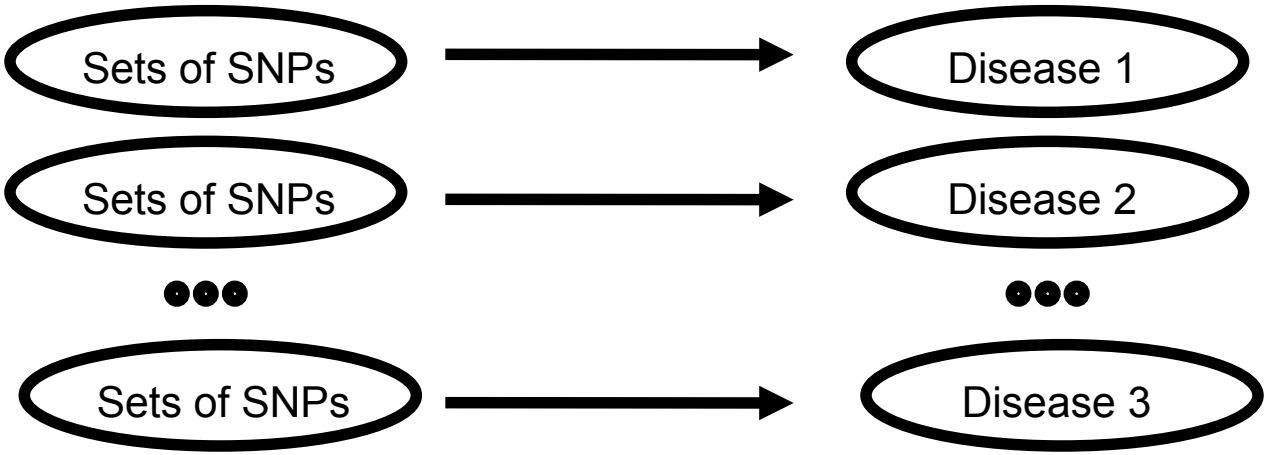
SNPs in the human genome

- SNPs frequency and evolution
  - ✓ Transitions are more common than transversions
  - ✓ New SNPs arise via mutation and with time they either disappear or reach fixation
  - ✓ SNPs that have a selective advantage among members of population may become enriched within that population through positive selection
  - ✓ Pattern of selection in the human genome is not uniform and vary according to gene functions

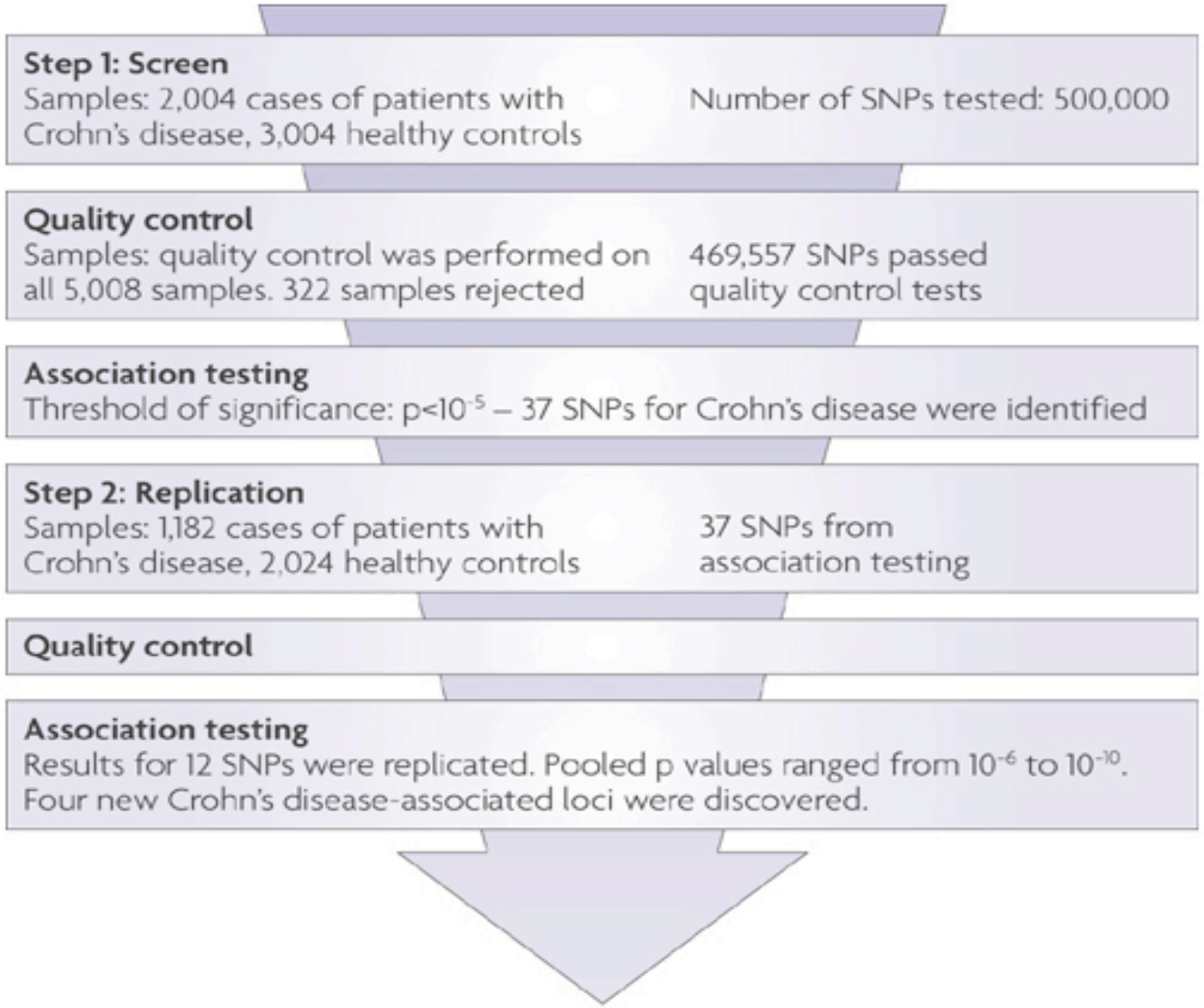
SNPs and disease association

- Identify those SNPs that are strongly (statistically significant) associated with a disease of interest

Normal individual	ATCG**G**TCGGAAGAGTTCCATGGGGGGTACAAAATGAGTAGAGC
Normal individual	ATCGCTCGGAAGAGTTCCAAGGGGGGTACAAAATGAGTAGAGC
Disease individual	ATCGGTCGGAAGAGTTCCATGGGGGGTACAAAACGAGTAGAGC
Disease individual	ATCGGTCGGAAGAGTTCCAAGGGGGGTACAAAACGAGTAGAGC



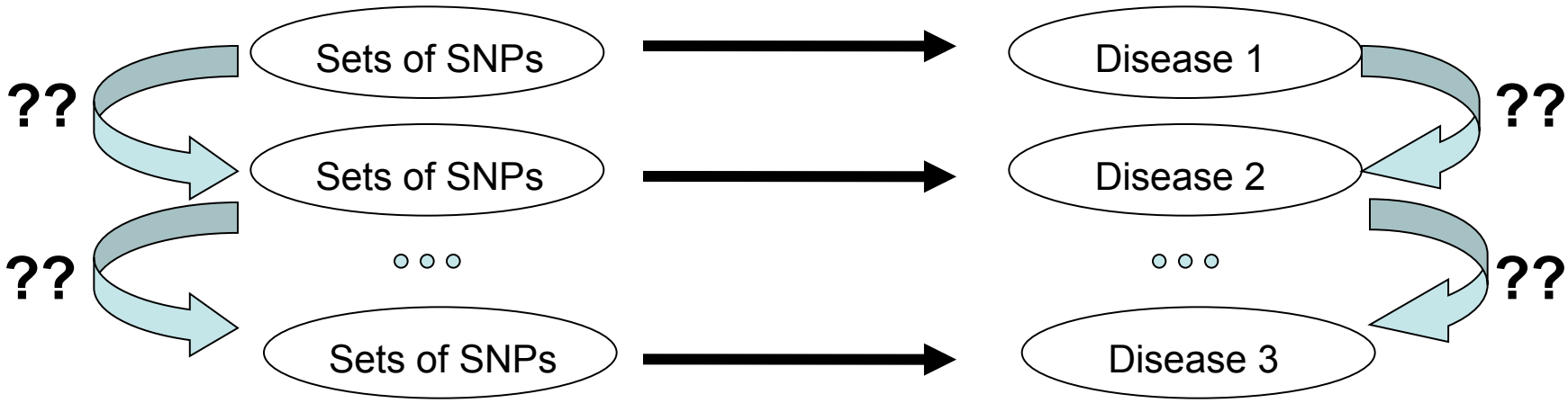
Genome Wide Association Studies



Nature Reviews | Immunology

SNP association and disease association

Different diseases are related.
What is the genetic basis for disease associations?

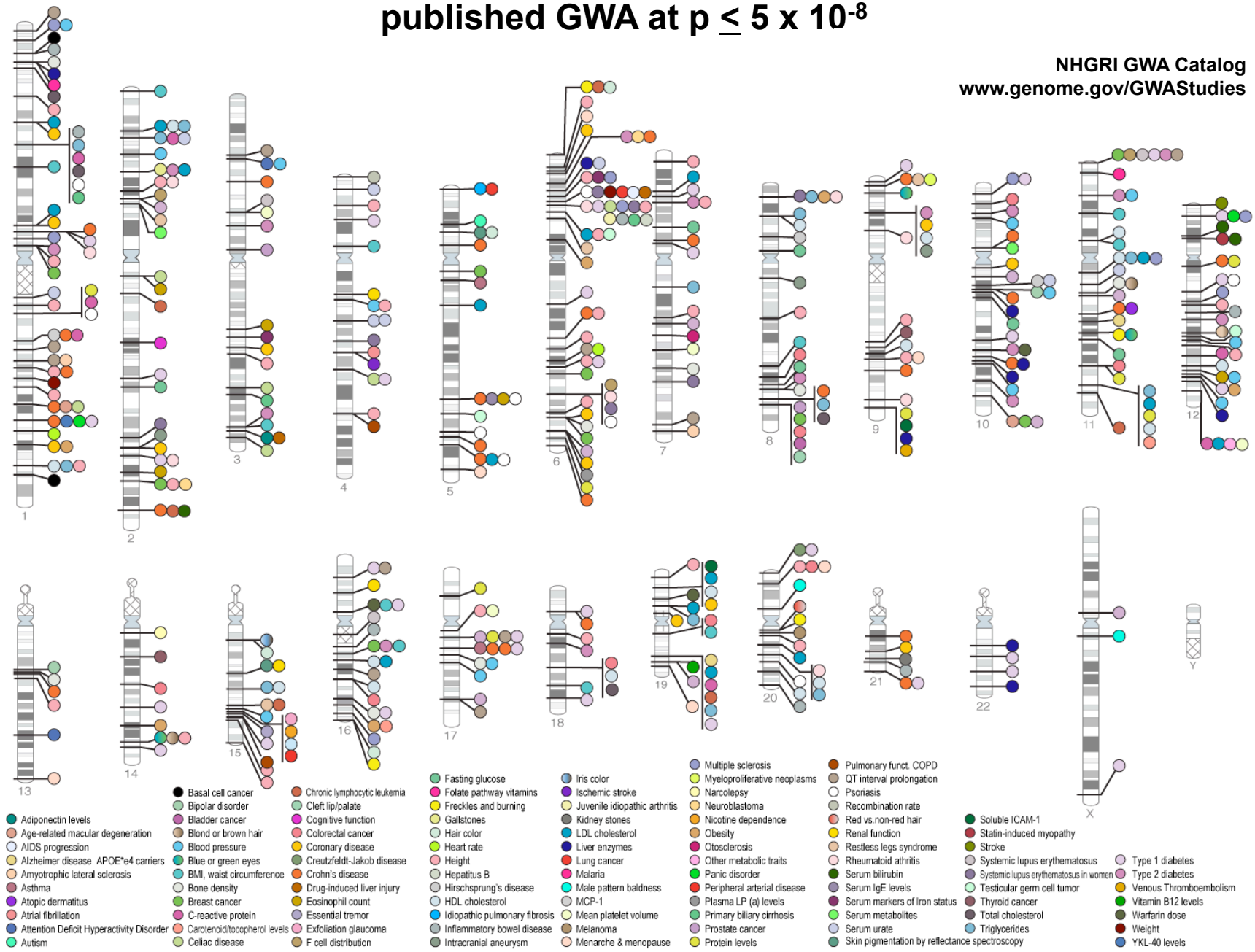


Most studies focus on identifying horizontal causal relationships.
Identifying vertical relationships is very important:

- discover possible hidden relationships between diseases.
- improve therapeutic treatment, disease diagnosis, & better prevention.

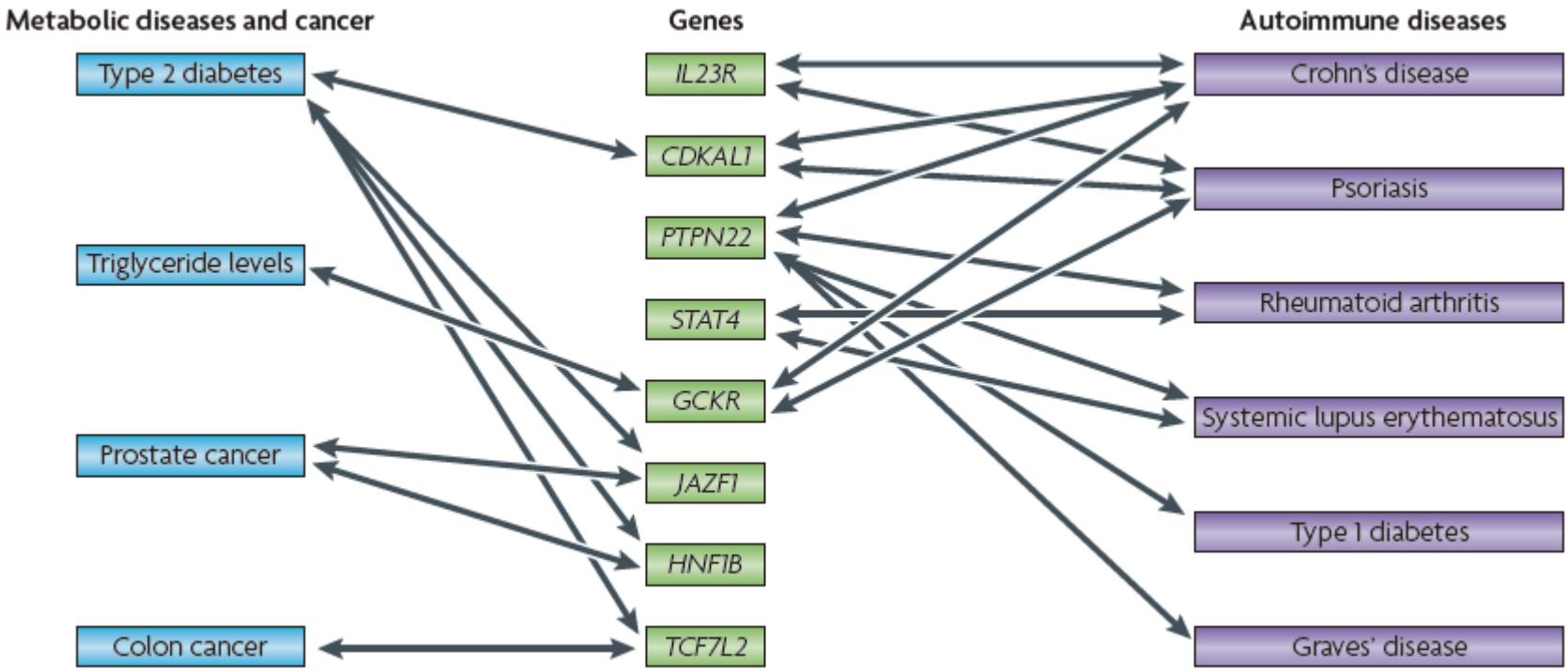
Published Genome-Wide Associations through 6/2009, 439 published GWA at $p \leq 5 \times 10^{-8}$

NHGRI GWA Catalog
www.genome.gov/GWAStudies



Human diseaseome

- Disease relationship through shared genes and pathways



Objectives

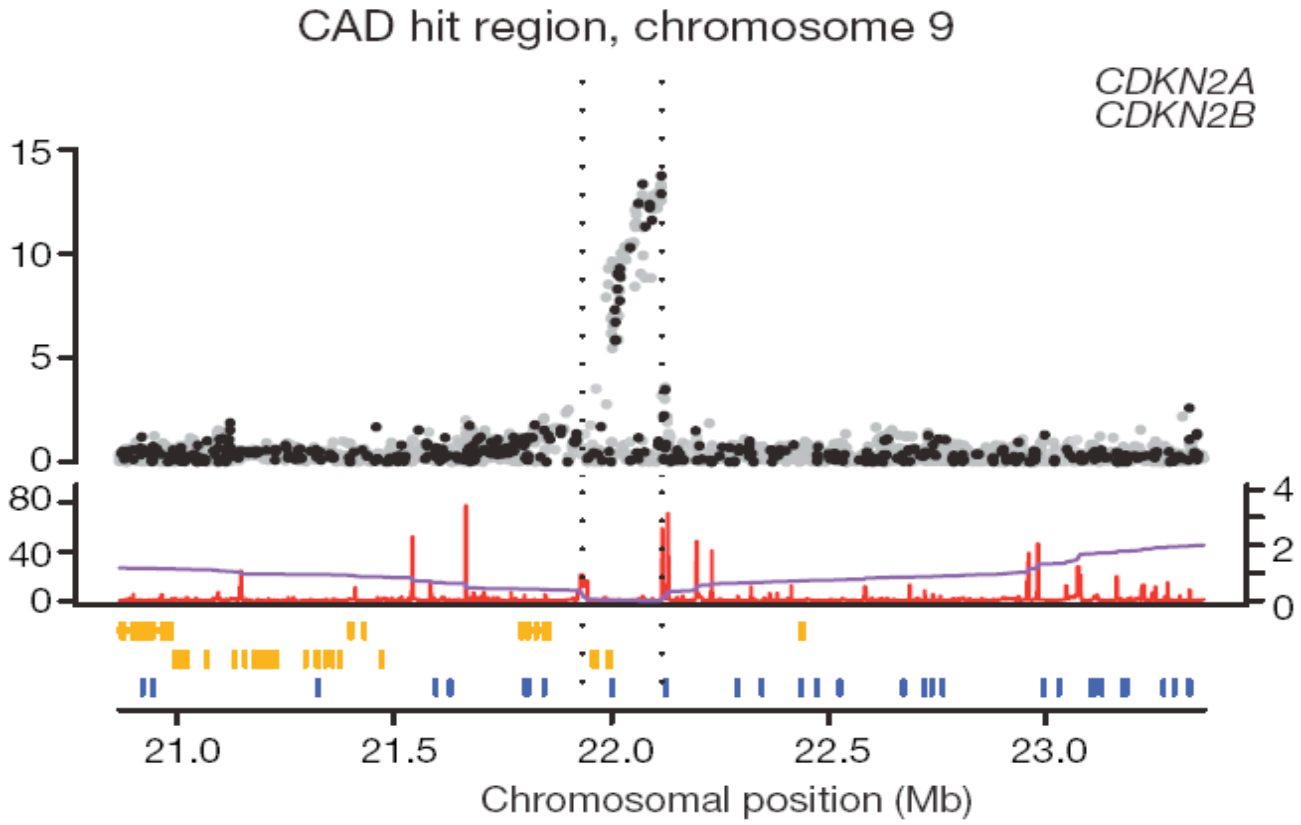
- Develop a framework to identify
  t
  h
  e genetic commonality between diseases.
  - Use available GWA data.
  - Analyze SNPs at four levels:
    - Nucleotide level (SNPs)
    - Gene level
    - Protein level
    - Phenotype level

Data used to identify the genetic basis of disease associations

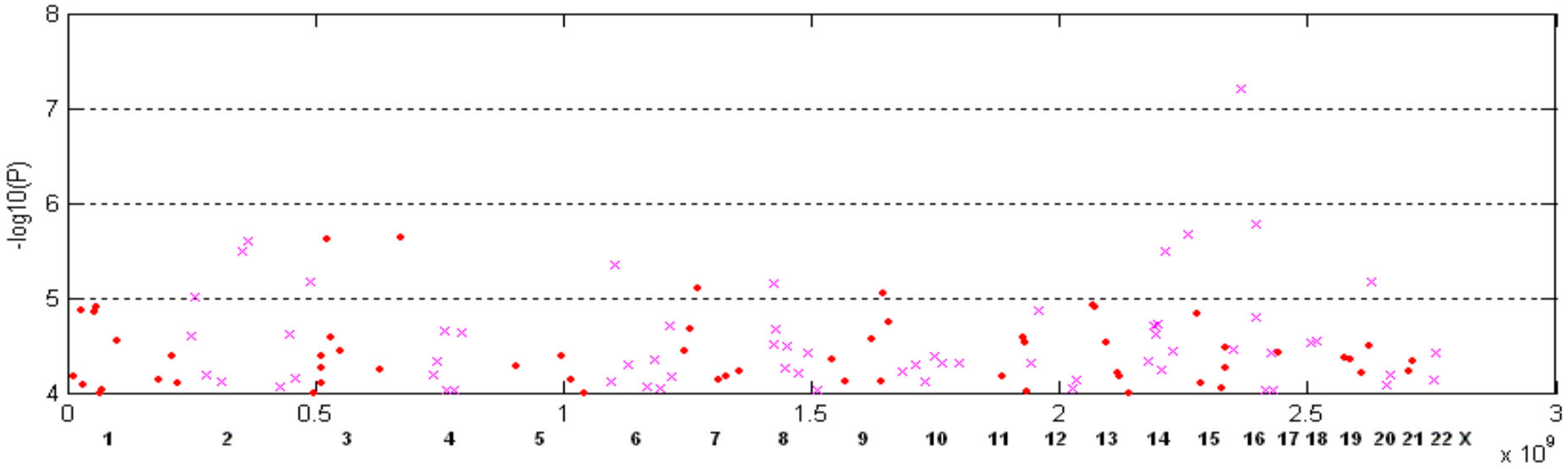
- 7 diseases
  from
  Wellcome Trust Case Control Consortium (WTCCC):
  - bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid
    arthr
    itis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D).
  - about 2,000 humans in the British population
    for
    e
    a
    ch of 7 diseases and a shared set of about 3,000 controls.
- Disease classification:
  - 3 natural groups:
    - **CAD+HT+T2D (metabolic**
      an
      d

SNP level analysis

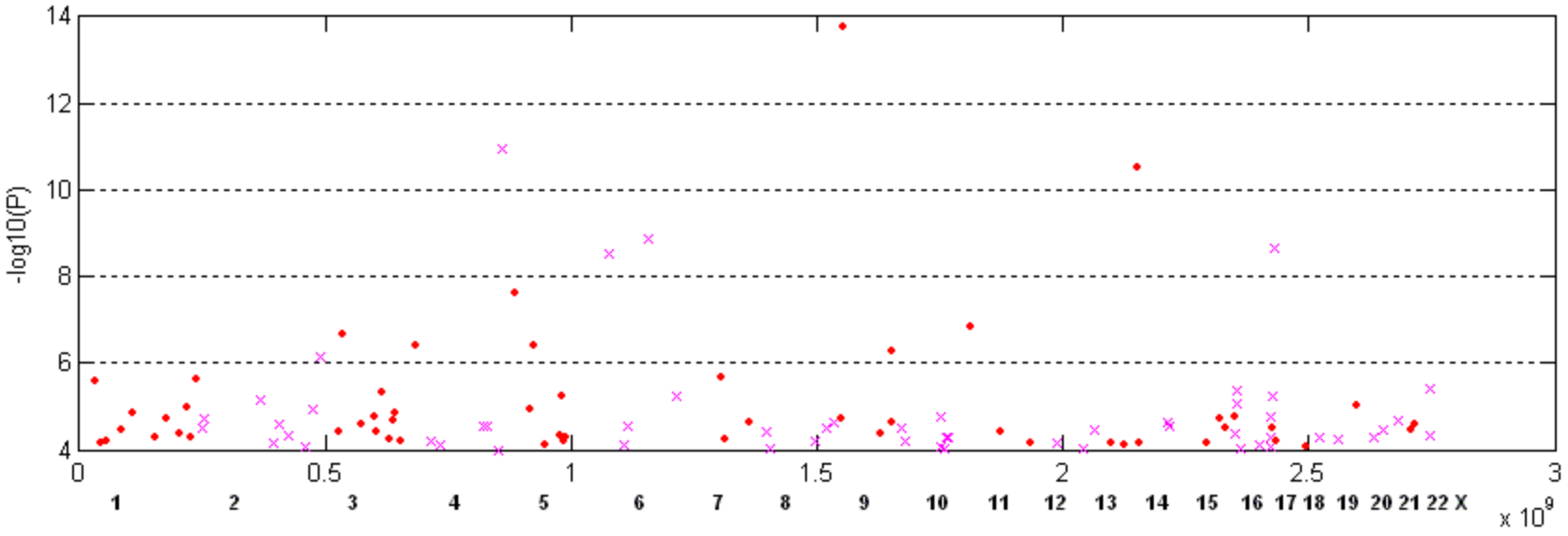
- Strong or moderate associations are associated SNPs which have P-value < 10^{-4}
- These SNPs are clustered into blocks with the distances between blocks being at least one mega base pairs.



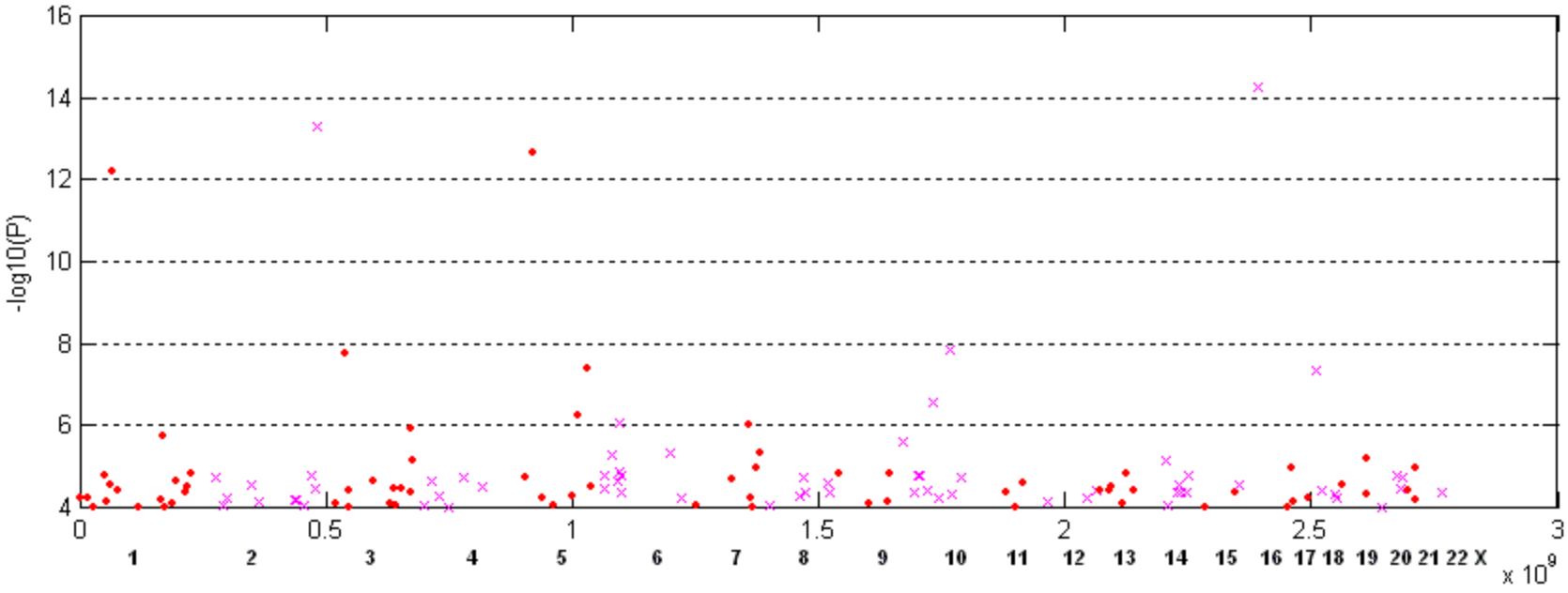
- Clustering result of BD



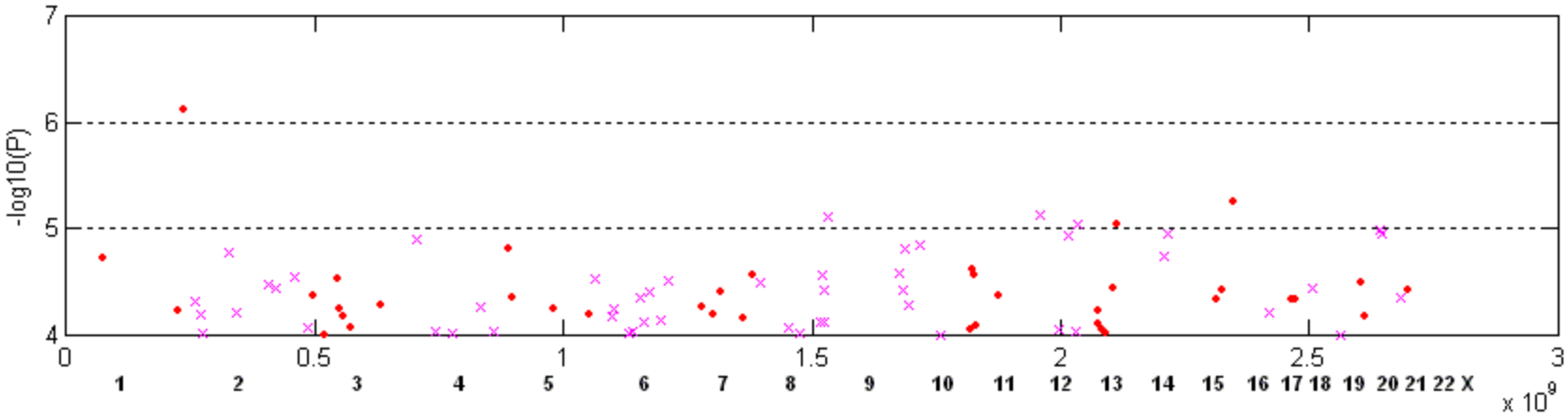
- Clustering result of CAD



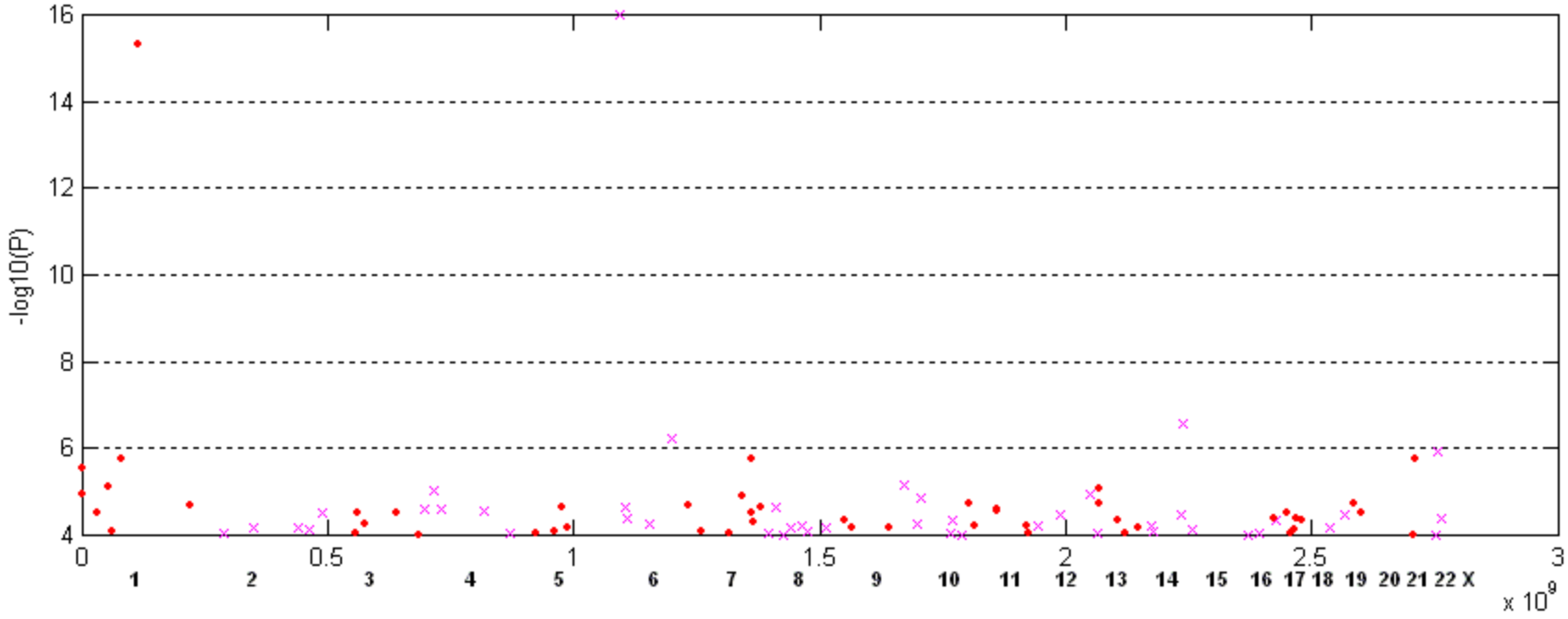
- Clustering result of CD



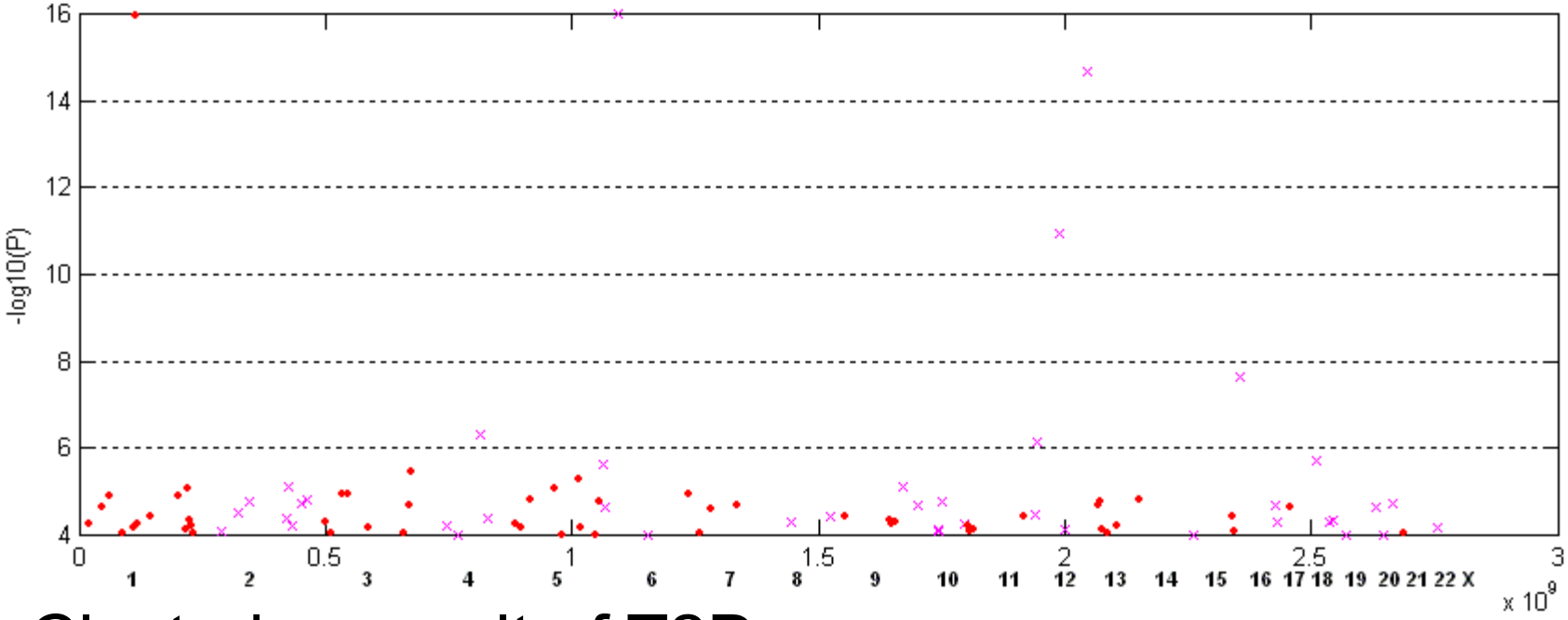
- Clustering result of HT



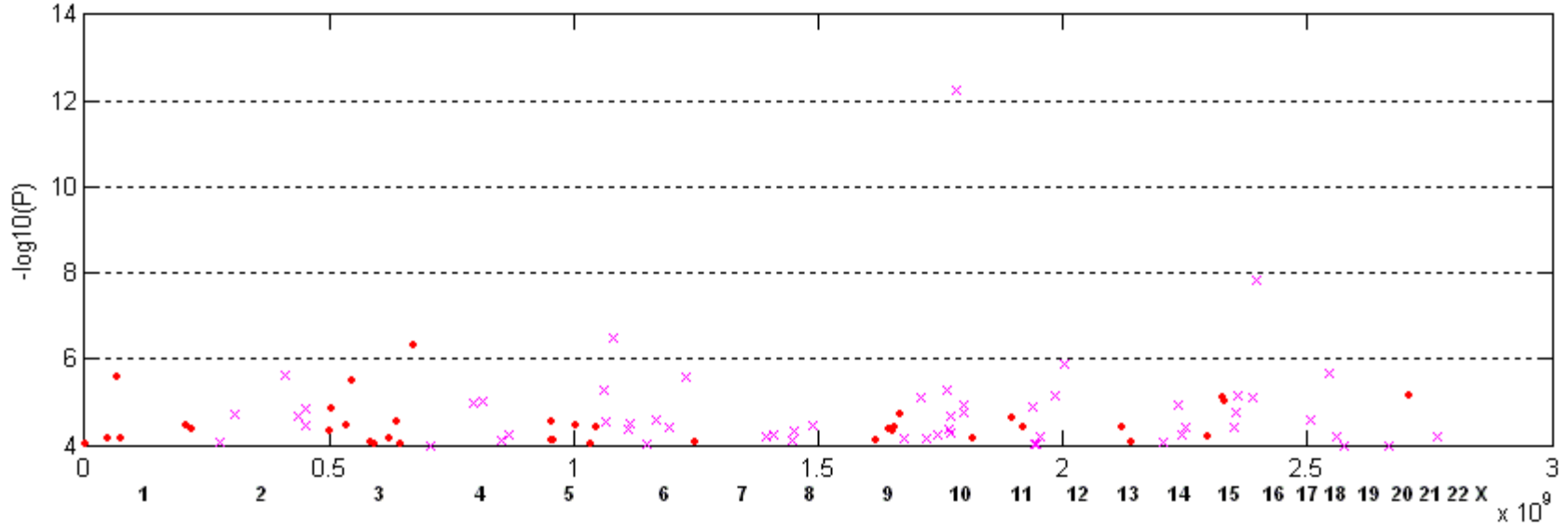
- Clustering result of RA



- Clustering result of T1D

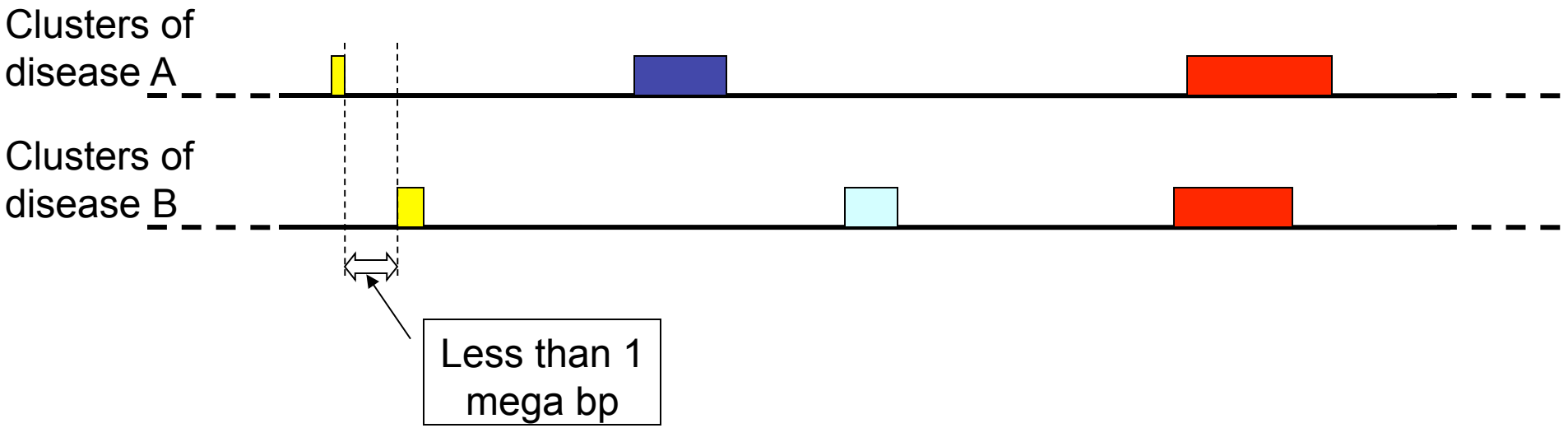


- Clustering result of T2D



Comparing cluster patterns of different diseases

- Paired (common) clusters:
  - Clusters from different diseases are compared to see whether they are overlapping or very close.



Jaccard Value Matrix

$$J_{A-B} = \frac{A \cap B}{A \cup B}$$

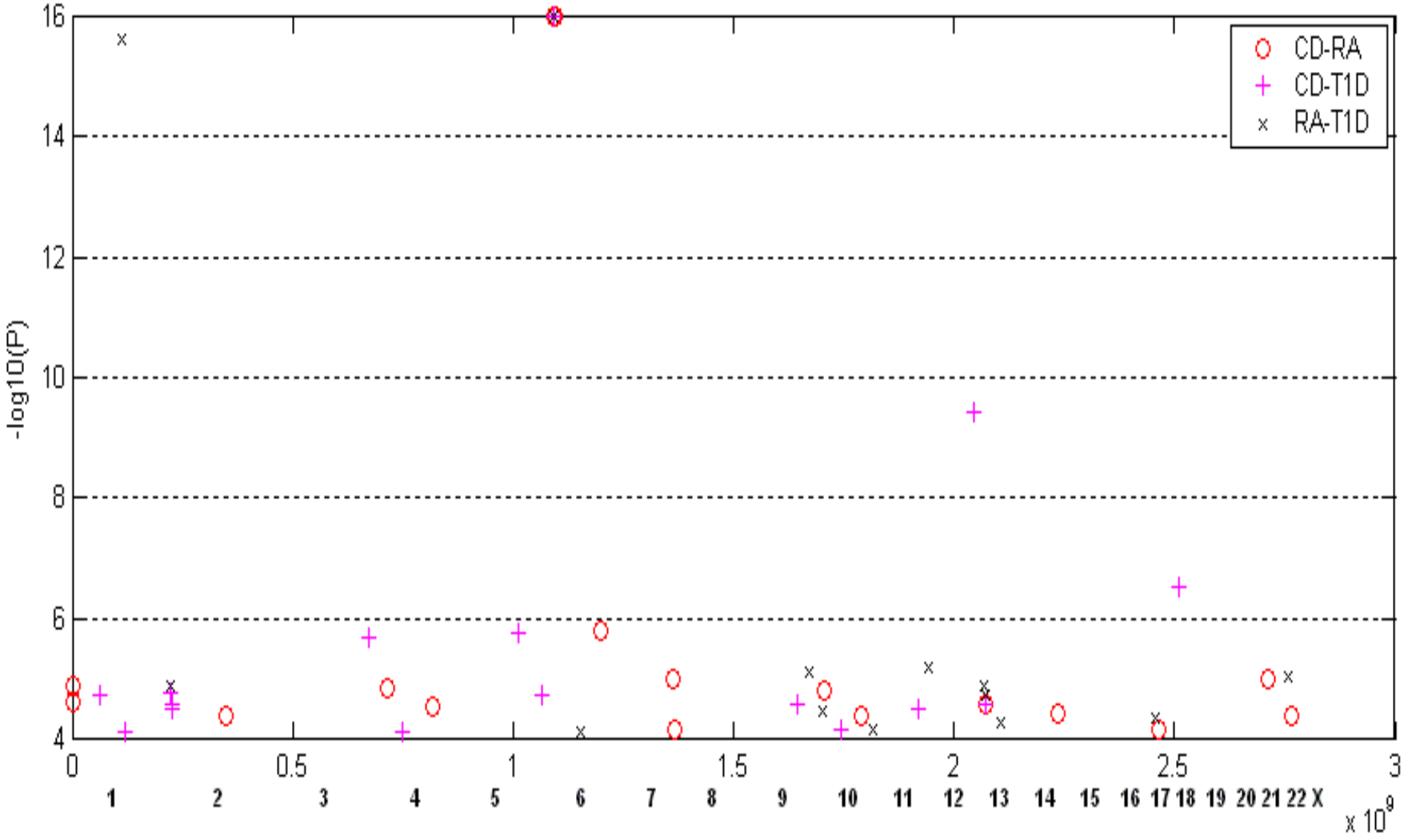
	BD	CAD	CD	HT	RA	T1D	T2D
BD		0.0563	0.0547	0.0569	0.0648	0.0645	0.0787
CAD	0.0563		0.0494	0.0556	0.0591	0.0637	0.0443
CD	0.0547	0.0494		0.0493	0.0789	0.0699	0.0658
HT	0.0569	0.0556	0.0493		0.0328	0.0380	0.0109
RA	0.0648	0.0591	0.0789	0.0328		0.0688	0.0319
T1D	0.0645	0.0637	0.0699	0.0380	0.0688		0.1005
T2D	0.0787	0.0443	0.0658	0.0109	0.0319	0.1005	

Total Length Matrix

- Total length = sum of all paired clusters' lengths for two diseases.
- Motivation
  - proportional to number of paired clusters
  - proportional to cluster length, which indicates the reliability of the association.

	BD	CAD	CD	HT	RA	T1D	T2D
BD		0.6883	0.7004	2.0869	4.4758	9.2207	0.7883
CAD	0.6883		1.8142	0.6240	0.3076	1.7135	1.2430
CD	0.7004	1.8142		1.2177	10.4997	15.2214	1.6777
HT	2.0869	0.6240	1.2177		0.6605	0.0456	0.1341
RA	4.4758	0.3076	10.4997	0.6605		16.2455	1.3246
T1D	9.2207	1.7135	15.2214	0.0456	16.2455		1.7031
T2D	0.7883	1.2430	1.6777	0.1341	1.3246	1.7031	

Distribution of cluster pairs for CD, RA and T1D



From SNPs to Genes

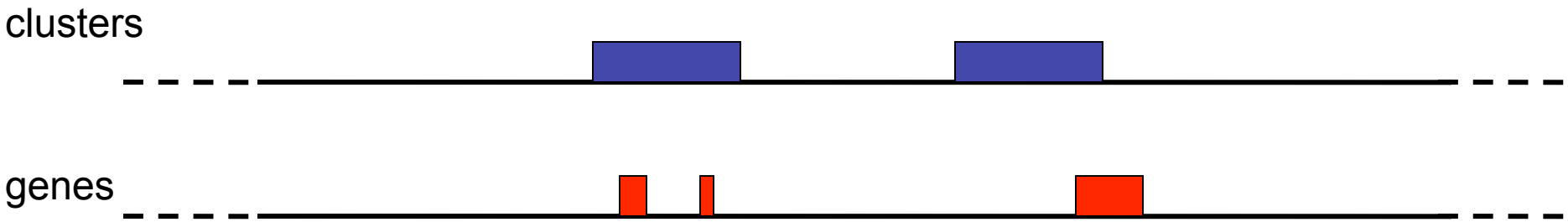
- Convert SNP clusters to genes
  - Assumptions:
    - Each SNP cluster may participate in regulating one or more genes.
    - Biological functions are expressed by genes.

From SNPs to Genes

- Goal 1
  - To explore disease association on the level of genes.
  - Motivation
    - Mutated or disregulated genes are the causes of many diseases
- Goal 2
  - To make possible our further investigation on the level of protein networks.
  - Motivation
    - Genes related to these diseases may indicate protein counterparts in patients' body.

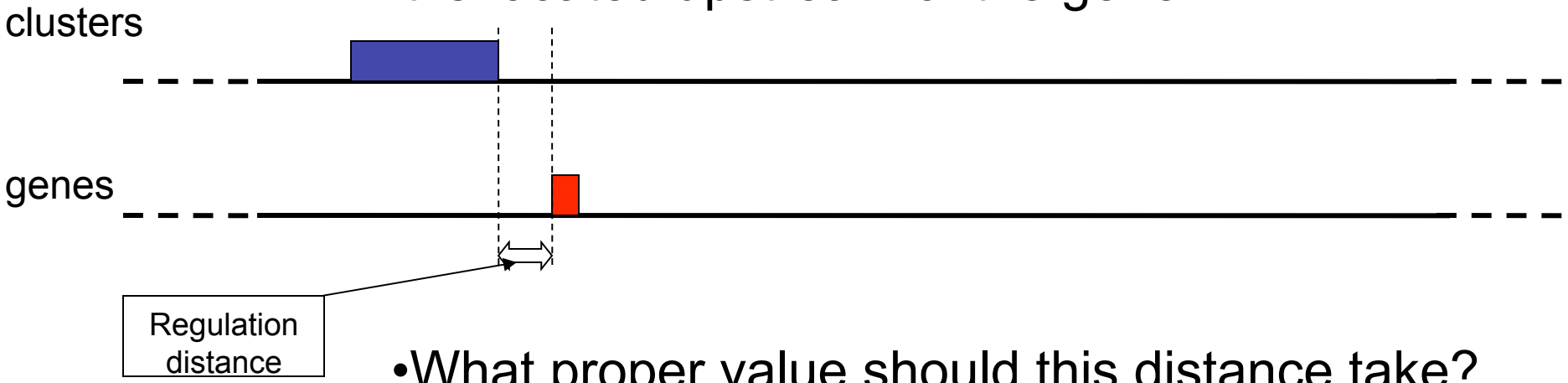
From SNPs to Genes

- Decide regulation distance from upstream
  - Straightforward conversion:
    - Convert a SNP cluster to a gene that shares an overlapping region.



From SNPs to Genes

- Decide regulation distance from upstream
  - Our concern:
    - Some SNP cluster may still activate a gene if it is located upstream of the gene.

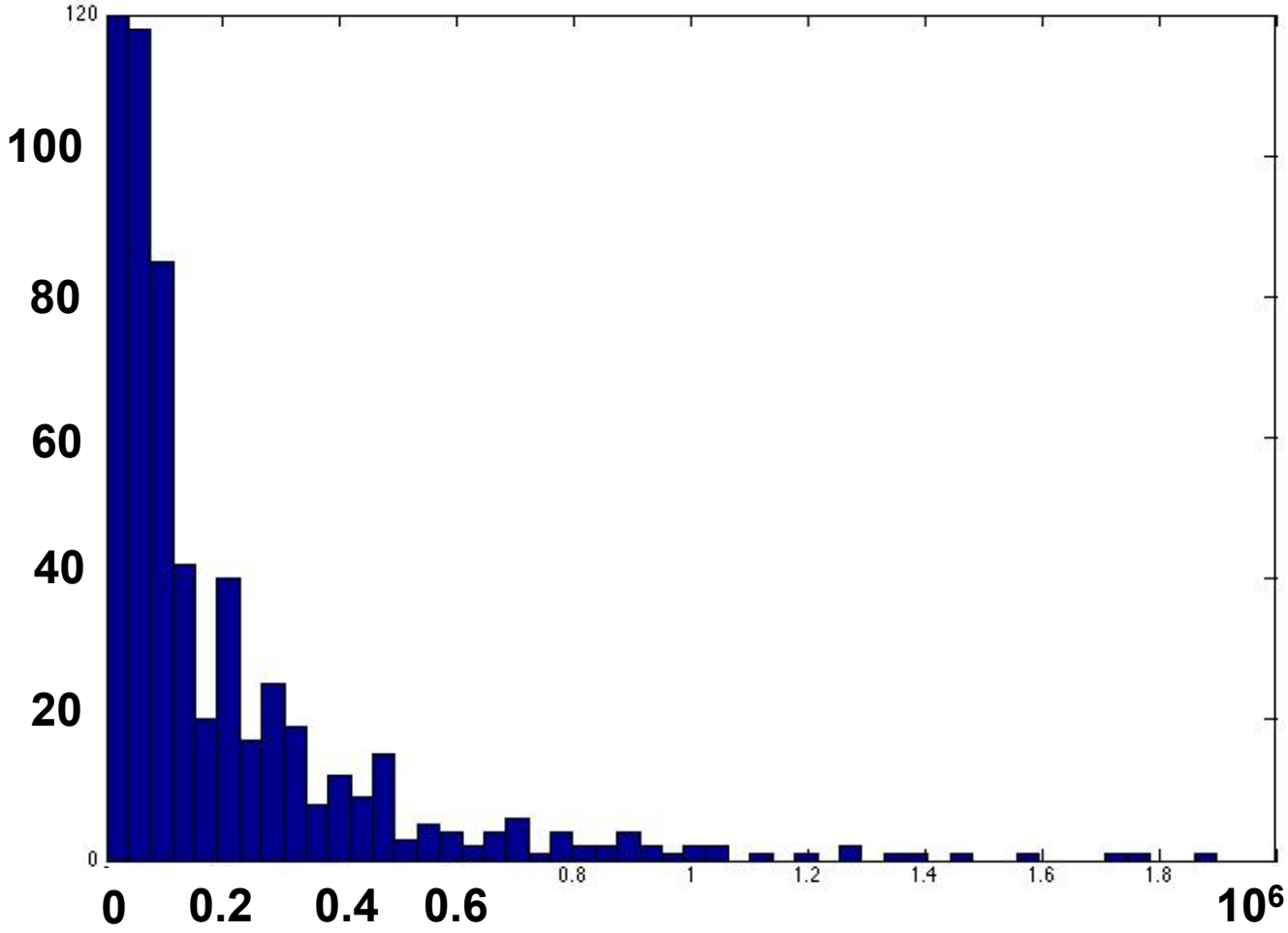


- What proper value should this distance take?

From SNPs to Genes

- Decide regulation distance from upstream
  - Assumption:
    - Every SNP cluster has a major regulatory effect on its nearest gene.
  - Method:
    - Collect distances from all SNP clusters to their nearest genes, and study the statistics.

From SNPs to Genes



Histogram for the distances from the nearest genes.
Max Distance = 1,899,079 base pairs

From SNPs to Genes

- Convert clusters to genes

Numbers of genes found

BD	CAD	CD	HT	RA	T1D	T2D
105	111	610	40	808	1095	33

Regulation distance = 0

BD	CAD	CD	HT	RA	T1D	T2D
5182	4370	6505	2812	4904	5304	3389

Regulation distance = 1,899,079

From SNPs to Genes

- Find shared genes between pairs of diseases

Number of shared genes

	BD	CAD	CD	HT	RA	T1D	T2D
BD		0	11	0	0	0	0
CAD	0		0	0	0	0	0
CD	11	0		0	223	240	0
HT	0	0	0		0	0	0
RA	0	0	223	0		686	0
T1D	0	0	240	0	686		0
T2D	0	0	0	0	0	0	

Regulation distance = 0

From SNPs to Genes

- Find shared genes between pairs of diseases

Number of shared genes

	BD	CAD	CD	HT	RA	T1D	T2D
BD		492	837	460	1005	712	598
CAD	492		354	333	648	522	261
CD	837	354		604	1164	1171	584
HT	460	333	604		151	251	139
RA	1005	648	1164	151		1412	281
T1D	712	522	1171	251	1412		702
T2D	598	261	584	139	281	702	

Regulation distance = 1,899,709

From SNPs to Genes

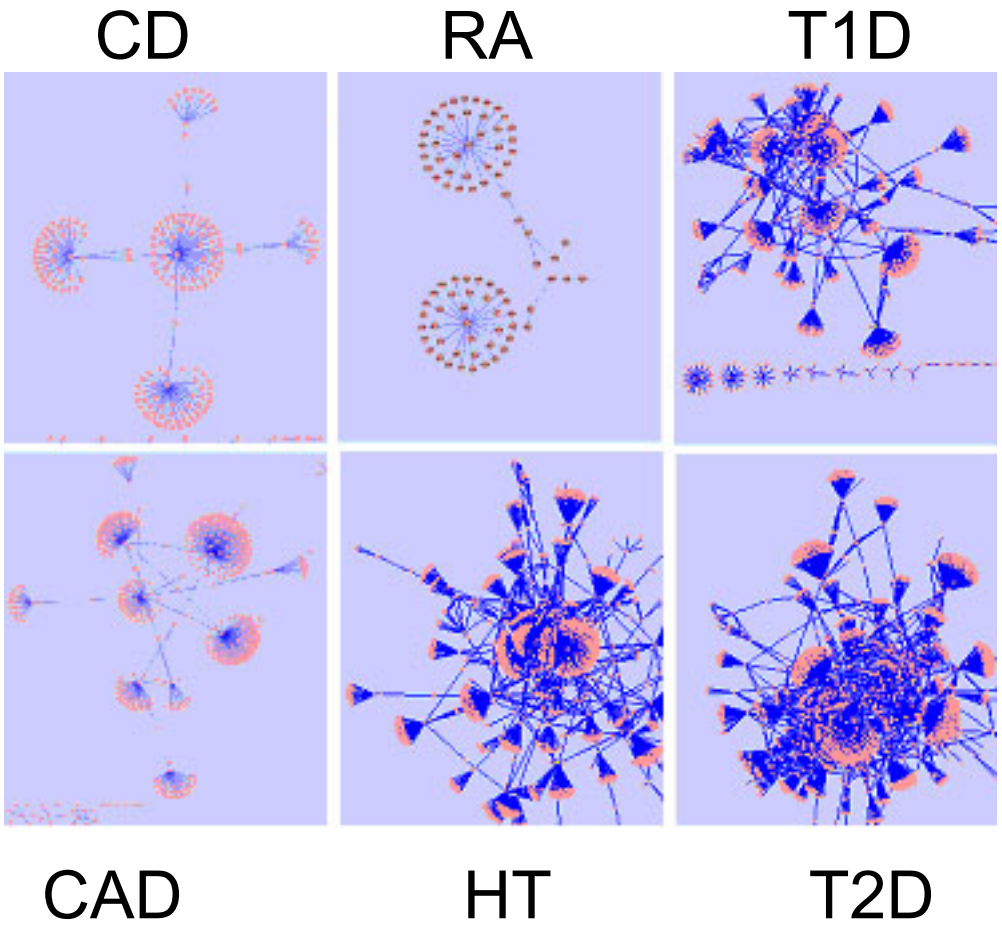
- Conclusion
  - Under different regulation distances from upstream, associated patterns among CD, RA and T1D are observed, at the level of genes.
  - no significant relation is found among CAD, HT and T2D.
  - Further investigation on protein networks is required.

From Genes to protein-protein interaction (PPI) network

- Convert genes to proteins

- Construct the PPI network using the STRING database

From Genes to PPI network



From Genes to PPI network

- Hub proteins:
  - nodes with >=15 degrees in the network.
  - form the backbone of a network
  - an important measurement for the similarity between protein interaction networks

From Genes to PPI network

number of hub proteins identified for each disease

Disease	Number of hub proteins
BD	2
CAD	11
CD	21
HT	2
RA	17
T1D	25
T2D	0

From Genes to protein-protein interaction network

the number of shared hub proteins in disease networks

Diseases	Number of hub proteins
CD, RA, and T1D	1
RA and T1D	5

The function of five hub proteins shared by RA and T1D and supporting publications

Protein name	Function(GO)	Association with RA	Association with T1D	Pubmed id
HLA class I histocompatibility antigen	Immune response	Yes	Yes	11369787
Radiation-inducible immediate-early gene IEX-1	NOT found in GO	Yes	Yes	16368886, 14630199
Collagen alpha	Integral to membrane;Beta-amyloid binding; Heparin binding	Yes	No	8816431
Death domain-associated protein 6	Nucleus; Protein binding; Protein homodimerization activity; Transcription factor binding	No	No	n/a
Mediator of DNA damage checkpoint protein 1	Nucleus; Protein binding	Yes	No	17913746

Comparison of disease phenotypes

- MimMiner to score the degree of similarity of various diseases to the diseases of our interest; The higher the score is, the more similar the two diseases are in terms of their phenotypes.

Comparison of disease phenotypes

Rank	id	Score	Disease Name
Crohn's disease and its phenotype hits.			
1	266600	1.0000	inflammatory bowel disease 1
2	191390	0.6624	ulcerative colitis
3	605225	0.6402	inflammatory bowel disease 7
4	**180300**	**0.4314**	**rheumatoid arthritis**
5	177900	0.4256	soriasis susceptibility
6	301000	0.4220	wiskott-aldrich syndrome
7	**222100**	**0.4065**	**diabetes mellitus, insulin-dependent**
8	249100	0.4054	familial mediterranean fever
9	232220	0.4048	glycogen storage disease ib
10	219700	0.4043	cystic fibrosis

Comparison of disease phenotypes

rheumatoid arthritis and its phenotype hits.			
1	180300	1.0000	rheumatoid arthritis
2	180350	0.4476	rheumatoid nodulosis
3	**266600**	**0.4314**	**inflammatory bowel disease 1**
4	**222100**	**0.4117**	**diabetes mellitus**
5	106300	0.3881	ankylosing spondylitis
6	191390	0.3861	ulcerative colitis
7	606044	0.3816	sjogren syndrome
8	254500	0.3736	myeloma, multiple
9	109100	0.3681	autoimmune disease
10	300310	0.3665	agammaglobulinemia

Comparison of disease phenotypes

type 1 diabetes and its phenotype hits.			
1	125480	1.0000	diabetes mellitus
2	275000	0.4704	graves disease
3	601318	0.4478	diabetes mellitus
4	270150	0.4458	sjogren syndrome
5	600496	0.4415	maturity-onset diabetes of the young
6	217000	0.4141	complement component 2 deficiency
7	**180300**	**0.4117**	**rheumatoid arthritis**
8	**266600**	**0.4065**	**inflammatory bowel disease 1**
9	137100	0.4049	immunoglobulin a defi1
10	125850	0.4042	maturity-onset diabetes

Conclusions and Discussion

- We analysed SNP associations between 7 diseases.
- For one group of diseases (CD, RA, T1D), strong genetic associations are found for ALL levels of analysis.
- For another group of diseases (CAD, HT, T2D), no genetic association is found.
- Negative result could be due to inappropriate grouping of diseases, or primitive analysis approach, or wrong assumption.

Acknowledgements

- Wenhui Huang
- Pengyuan Wang
- Zhen Liu
- Deng Pan

National Science Foundation (NSF).

AdvanceVT grant.

Research areas

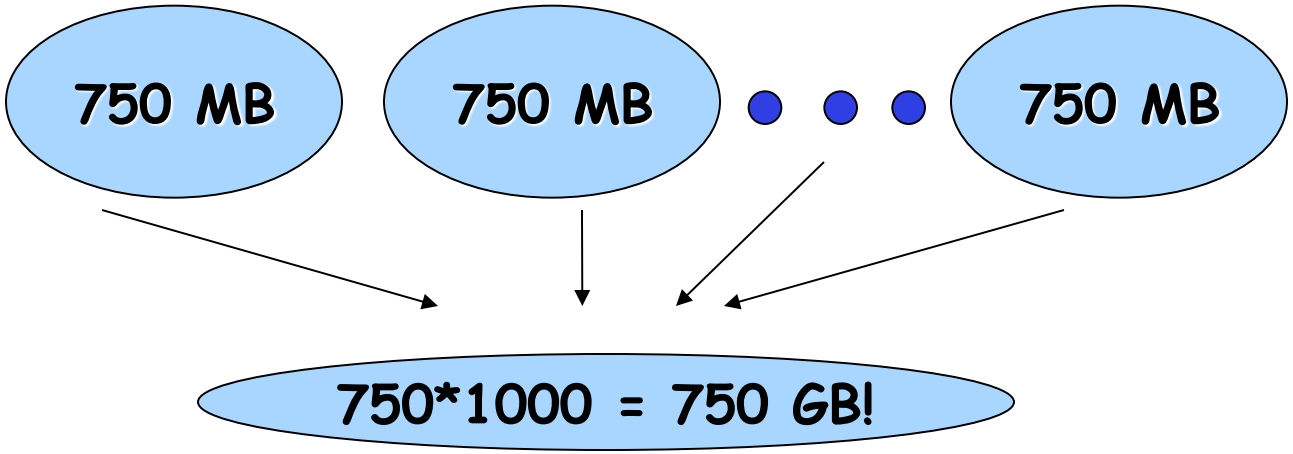
- Systems biology to understand disease relationships
  - Microarray data, protein interaction data
  - Genome wide association studies
- Large-scale data processing software
  - Short reads mapping project
- Data mining
  - Rare class prediction algorithm
- Algorithms in genomics
  - Genome compression project
- Comparative genomics

Several ongoing projects

- Genome compression project
- Rare class prediction problem
- Parallelization of next generation short-reads mapping programs
  - Mapping of the short reads of RNAs to the mosquito genome, determine the patterns of the distribution, infected with virus vs. the uninfected
- Microarray data analysis for lung cancer

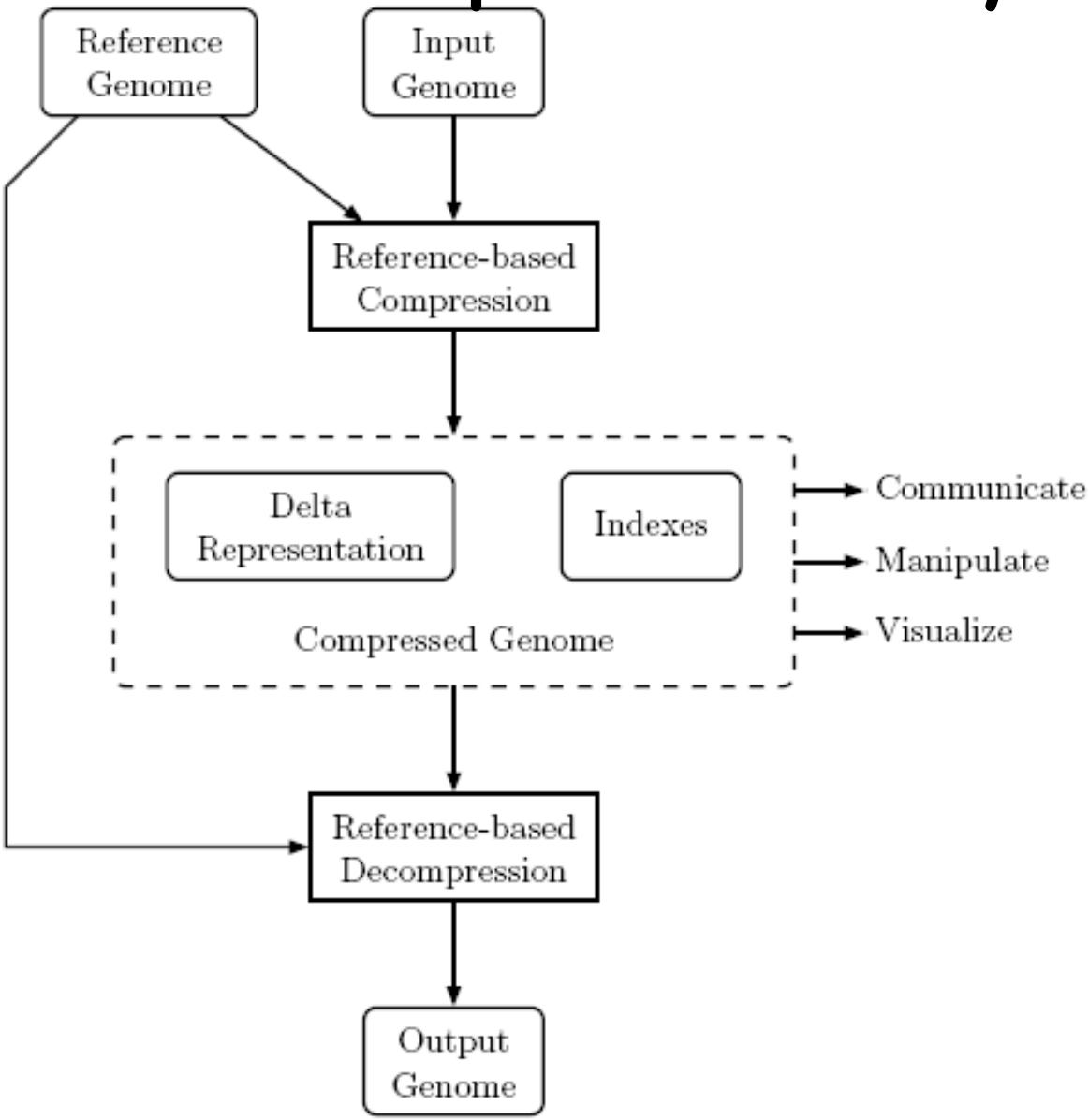
Genome compression

- Human genome: 3 billion bps, takes up about 750 Mb to store.



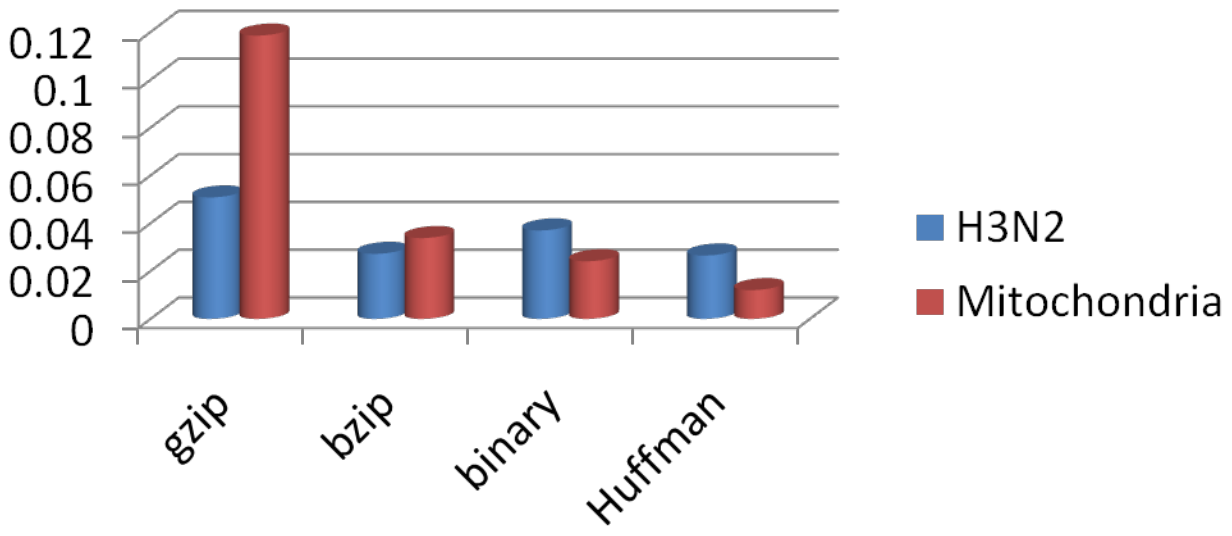
Storage, communication, manipulation challenges!!

Genome compression system



Compression Ratio

5473 Mitochondria DNA (Primitive size 91,590,508 bytes)				
	gzip	bzip2	Binary	Huffman
Size	10,860,887	3,096,944	2,211,120	1,101,628
Ratio	0.1185	0.0338	0.0241	0.01202
1455 H3N2 Virus genome segments (Primitive size 1,995,960 bytes)				
Size	101,494	54,314	73,901	52,927
Ratio	0.0508	0.0272	0.0370	0.0265



The Rare-Class Data Problem

- Occurs when one class is far outweighed by other classes
  - In a binary dataset: minority and majority class
  - Minority class is typically the class of interest

Rare class problem is common

Examples

- D
  etecting a hacker/cracker illegally using a server
  - Majority: normal users
  - Minority: illegal users
- Identifying flaws in manufacturing
  - Majority: correctly working products
  - Minority: flawed products
- Internet search
  - Majority: Irrelevant websites
  - Minority: Relevant website(s)

Biological Examples

- Disease Diagnosis
  - Majority: Do not have disease
  - Minority: Have disease
- Gene prediction on a DNA strand
  - Majority: “Junk” DNA
  - Minority: Genes
- Detecting cancer in a tissue
  - Majority: Healthy tissue
  - Minority: Cancerous tissue

Ongoing projects and directions

- Comparative genomics of the turkey genome with the chicken genome
  - Structural variation in the turkey genome
- Molecular evolution of genes involved in RNAi pathways of different mosquito species.
- Cow comparative genomics with other mammals.
- Dog group.
- Compensatory mutations in protein evolution
- Gene family evolution and gene conversion prediction program.
- Mining of GWA studies -- SNPs and diseases