

AIT-BUDAPEST



AQUINCUM INSTITUTE OF TECHNOLOGY

Creativity in
Computer Science &
Engineering

COMPUTATIONAL BIOLOGY and MEDICINE

Genetic association studies II.

Andras Falus afalus@gmail.com

Peter Antal antal@mit.bme.hu

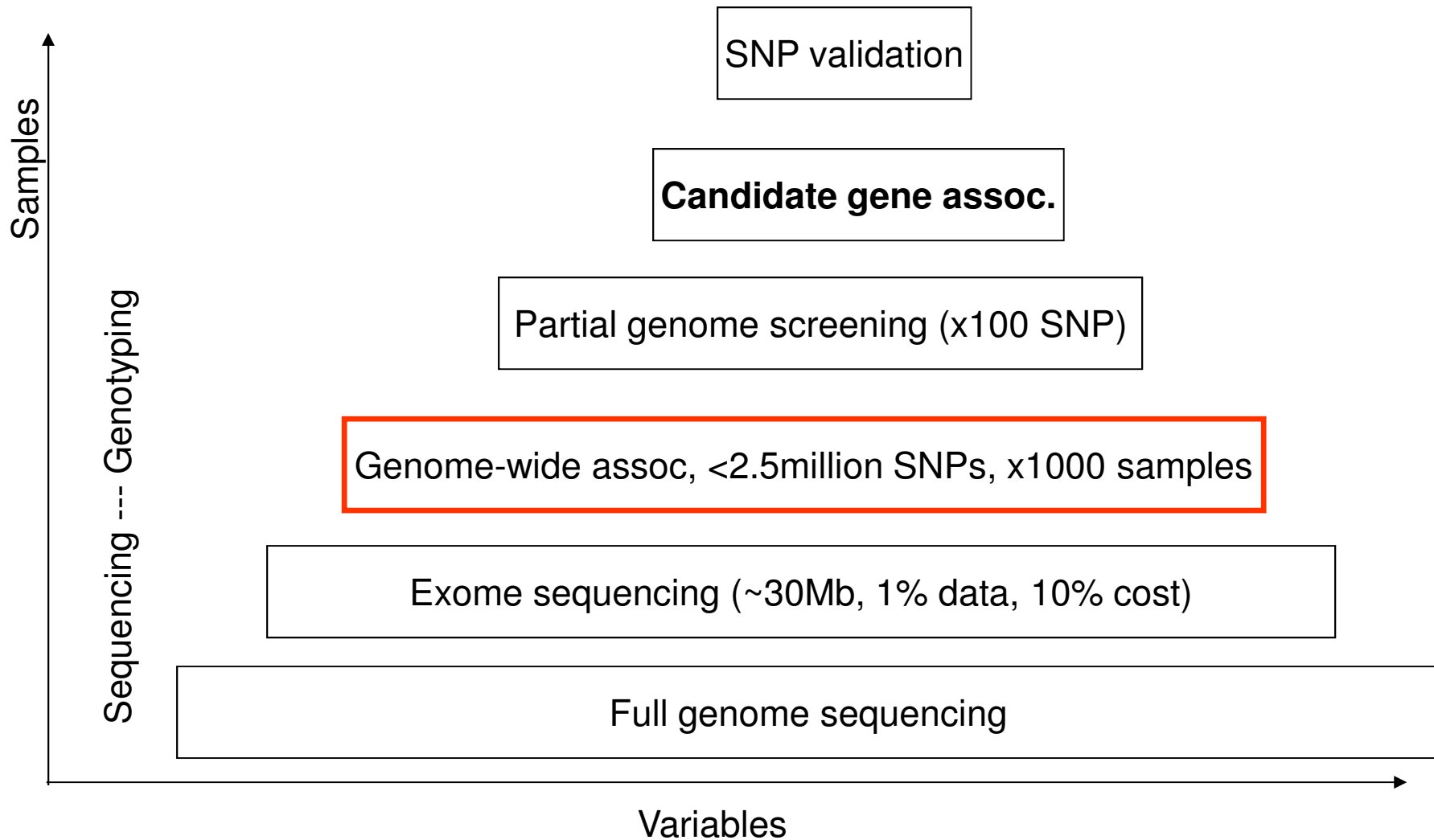
Gábor Csonka csonkagi@gmail.com

AIT, Budapest 2011. fall

Genetic association studies (GAS)

- Data
 - Measurement uncertainty
- Challenges
 - Association vs causation
 - Multiple testing problem
 - Knowledge-intensive statistics
 - Granularity: variations, genes, pathways
 - Pre- or post-aggregation
 - Computation-intensive statistics
 - Permutation, bootstrap

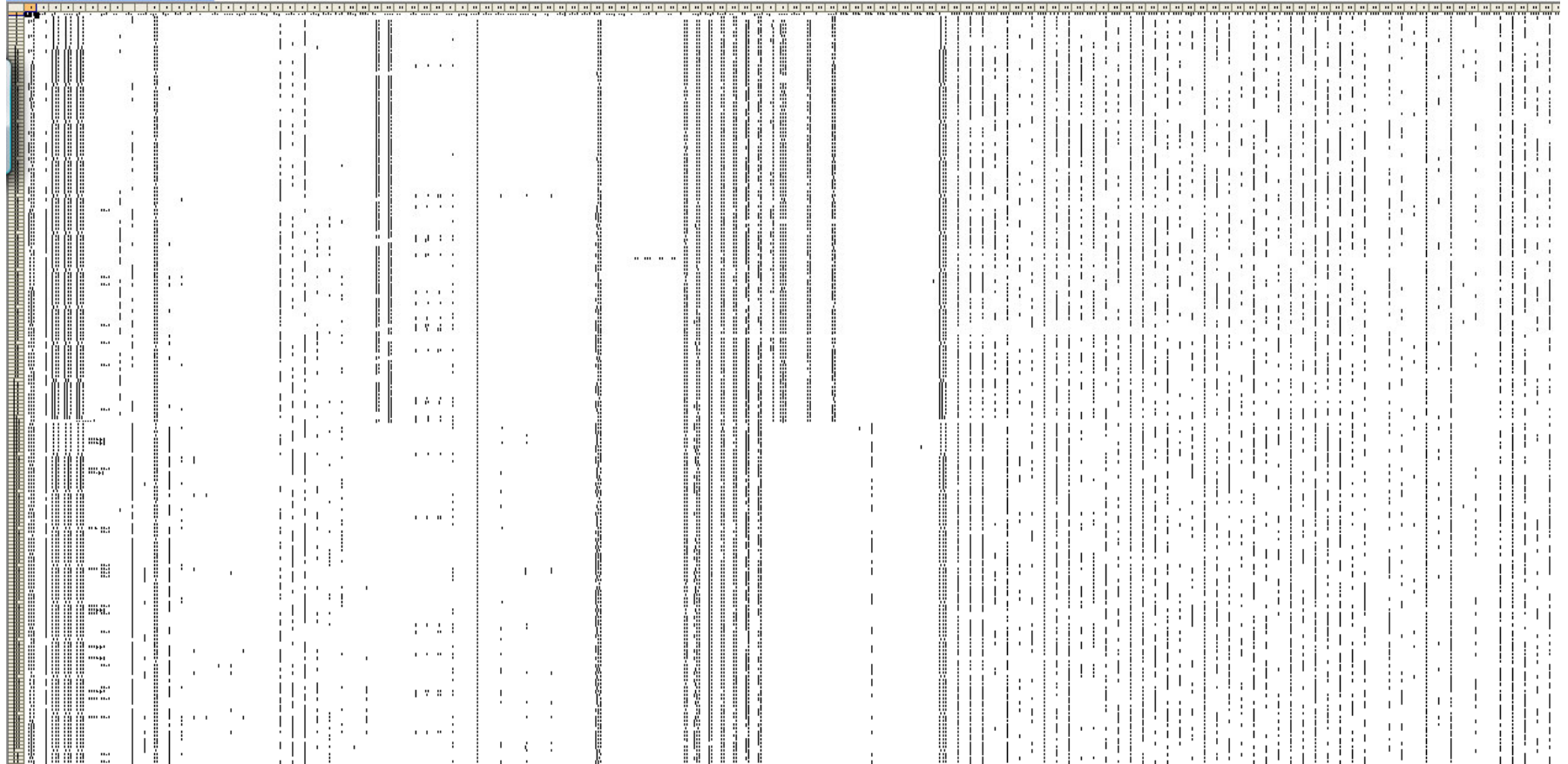
Genetic association data



About biomedical data....

Variables (observations, interventions)

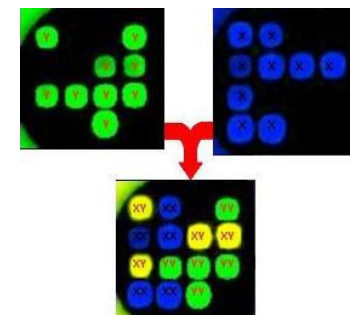
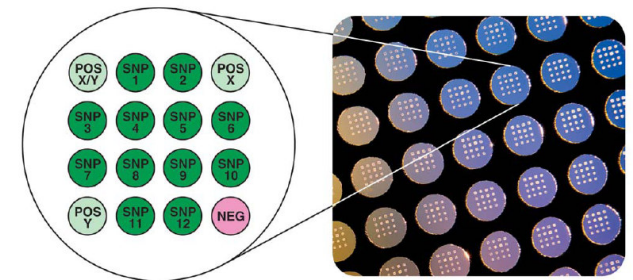
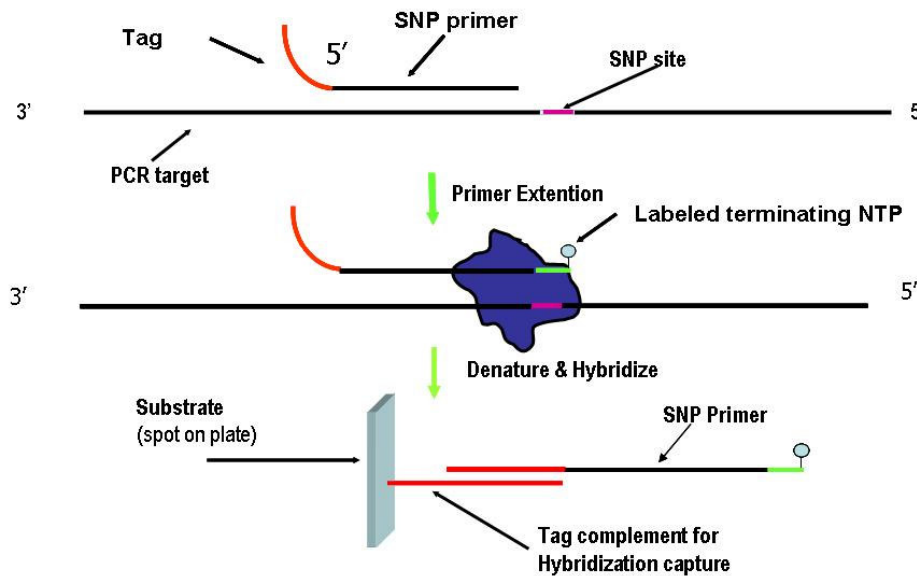
Samples



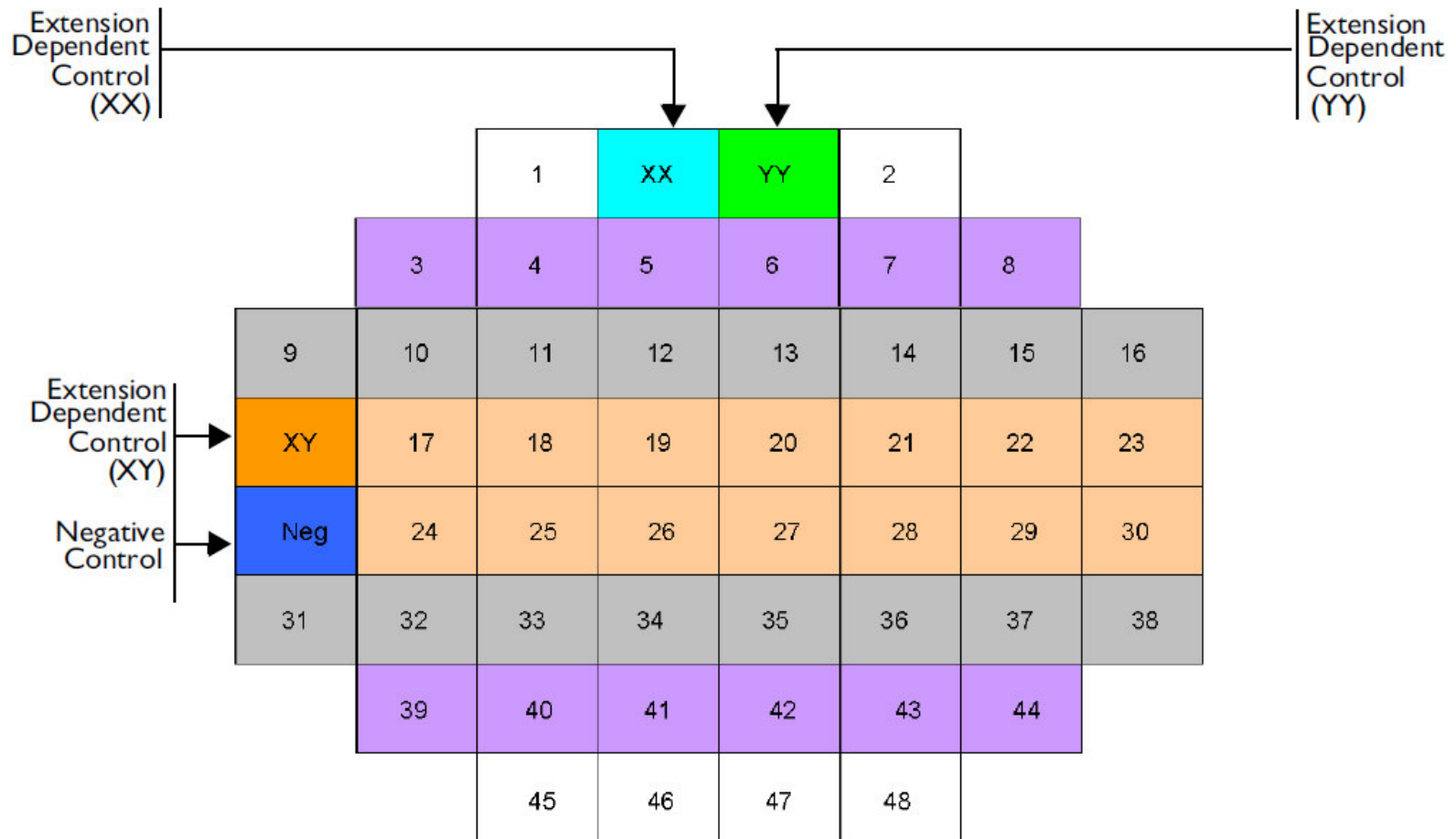
....frequently biomedical data is incomplete, but genotyping/sequencing is “digital”(???)

About “digital” biomedical data: genotyping using tagged single-base extension primer

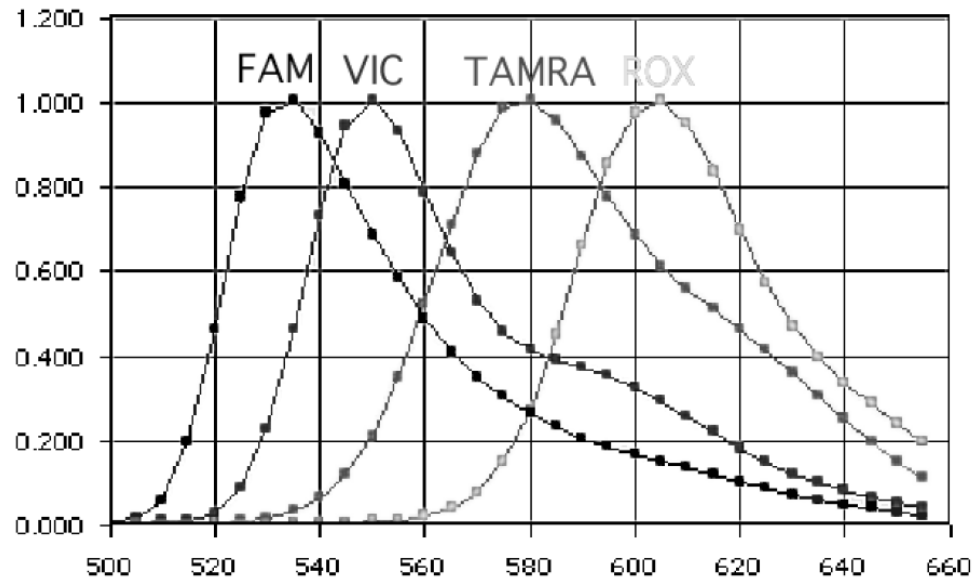
Rs25:CTCTGTGAGCTTCTGCATGCAATCCT[C/T]TGCAATTGGAATTTGATAGTCCTTT



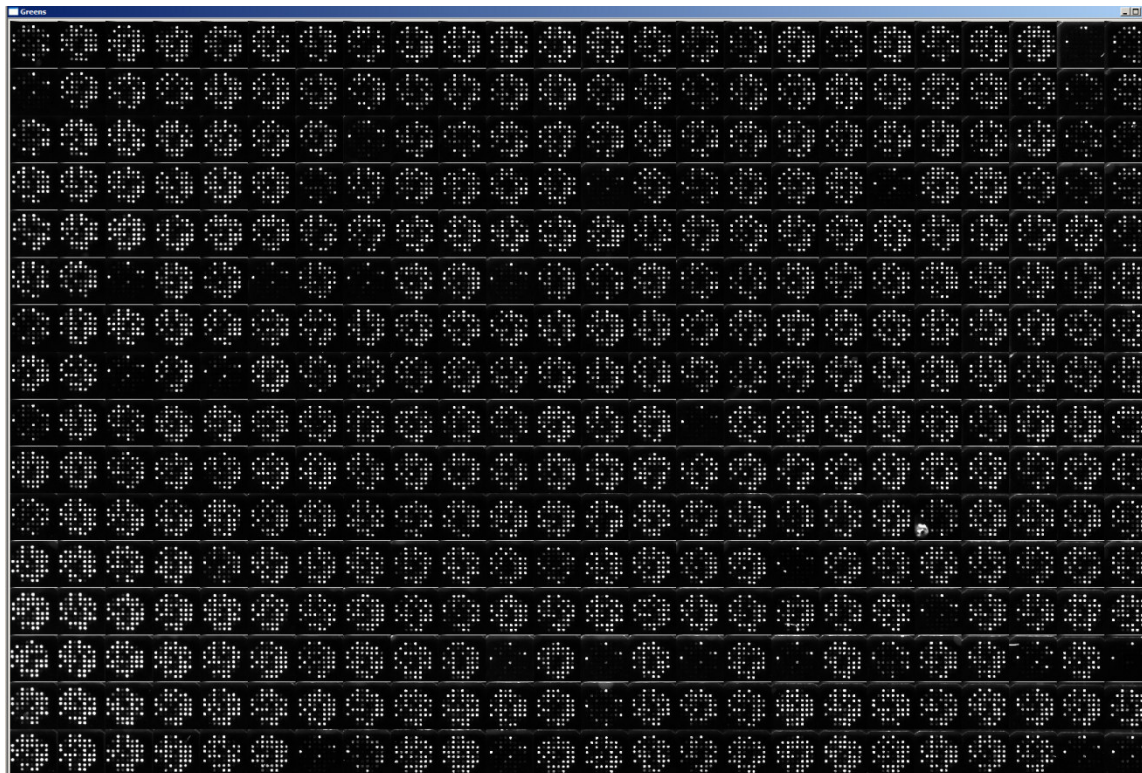
Control spots and spot layout for 48plex plates



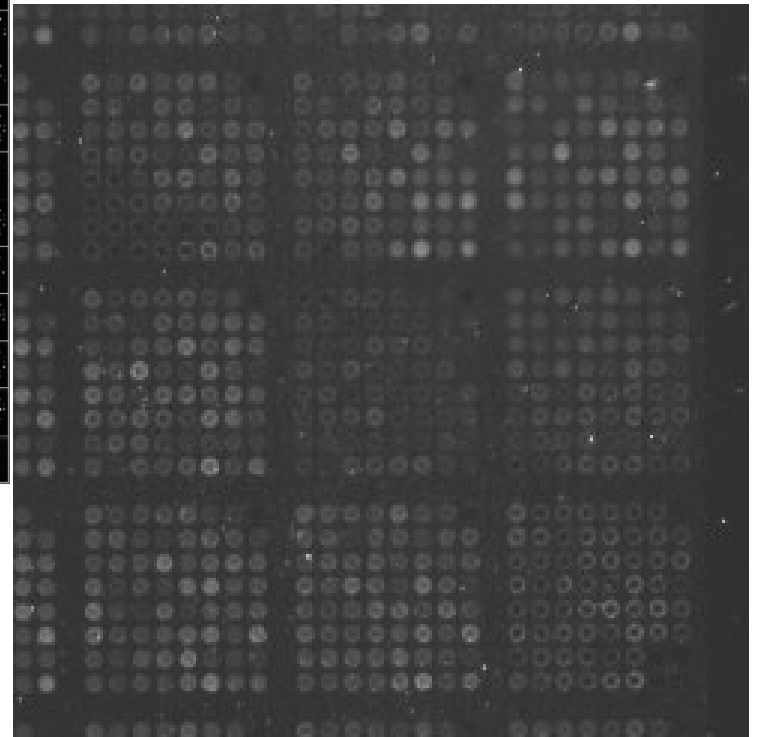
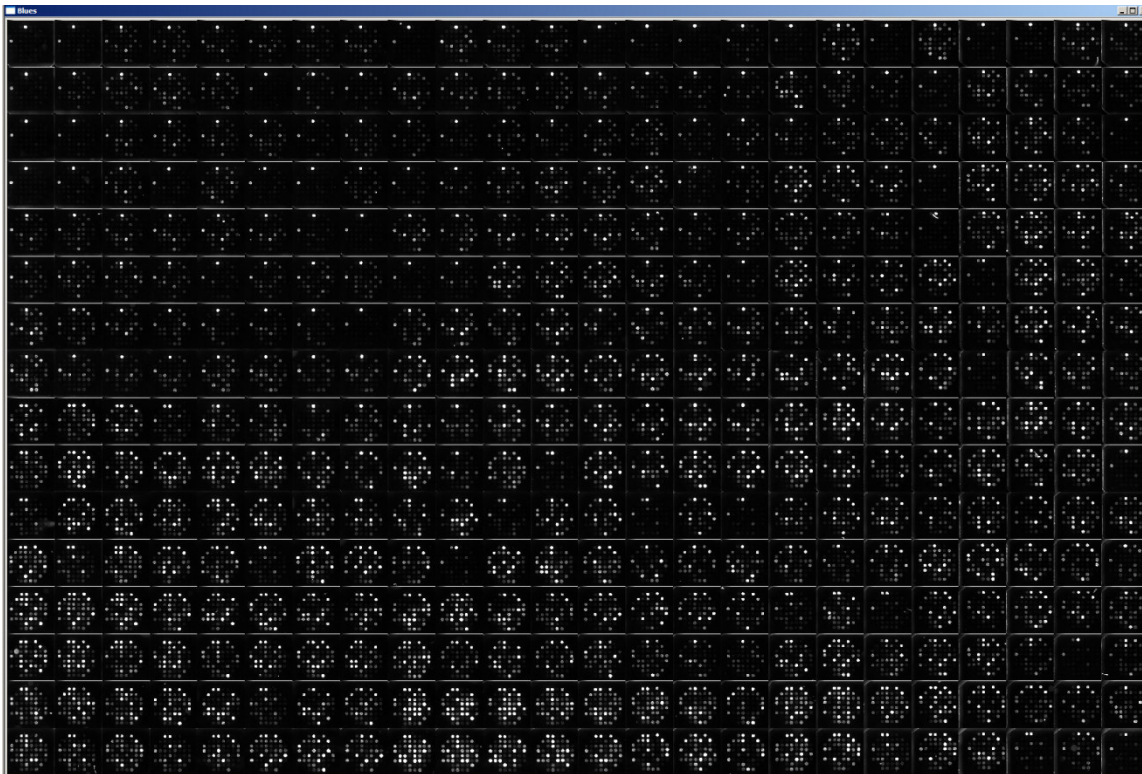
Fluorescent dyes



Biomedical data: a high-quality plate



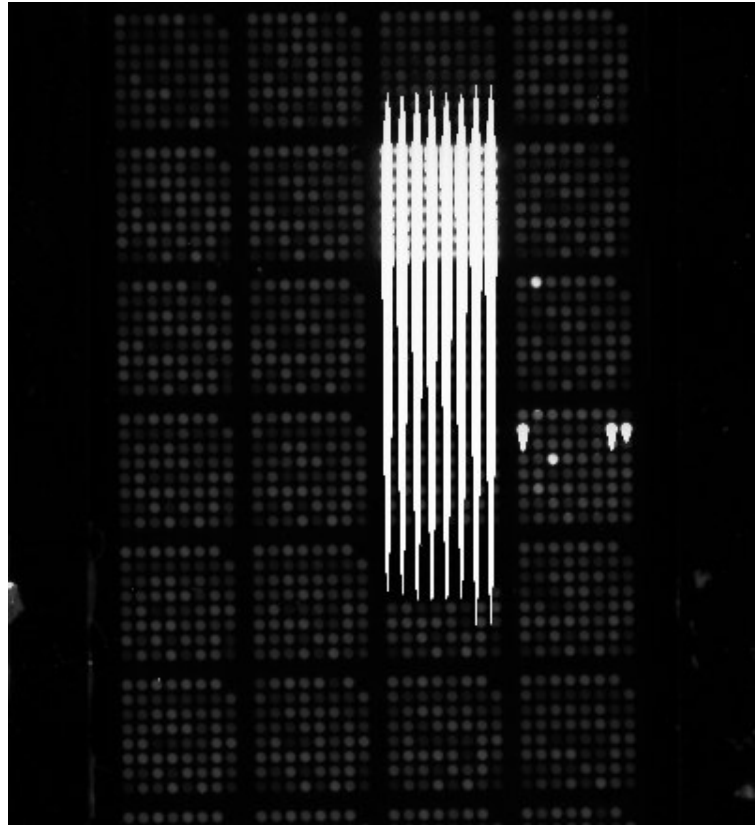
Biomedical data: plates with insufficient reagent



Biomedical data: a plate with blur



Biomedical data: a plate with digital noise

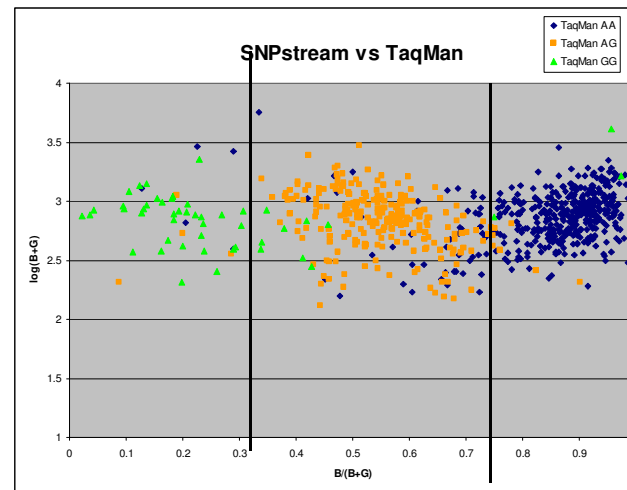
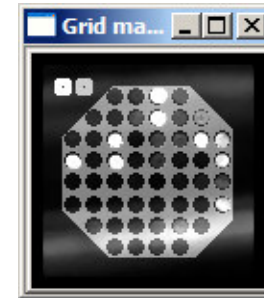


Biomedical data:
a plate with a bubble



From image processing to symbolic data

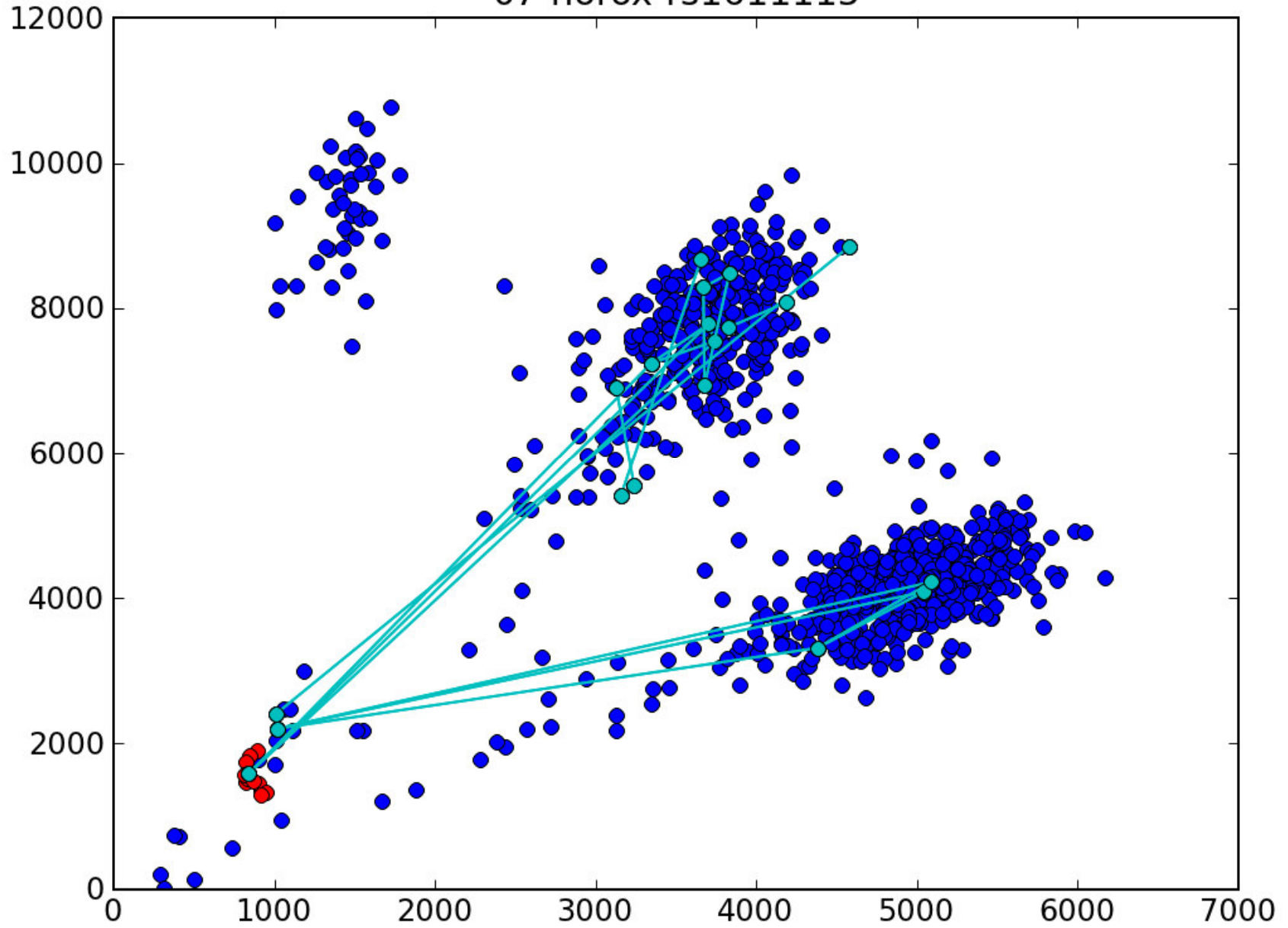
- Image processing (filtering)
- Grid alignment
- Segmentation
- Intensity calculation
- Clustering



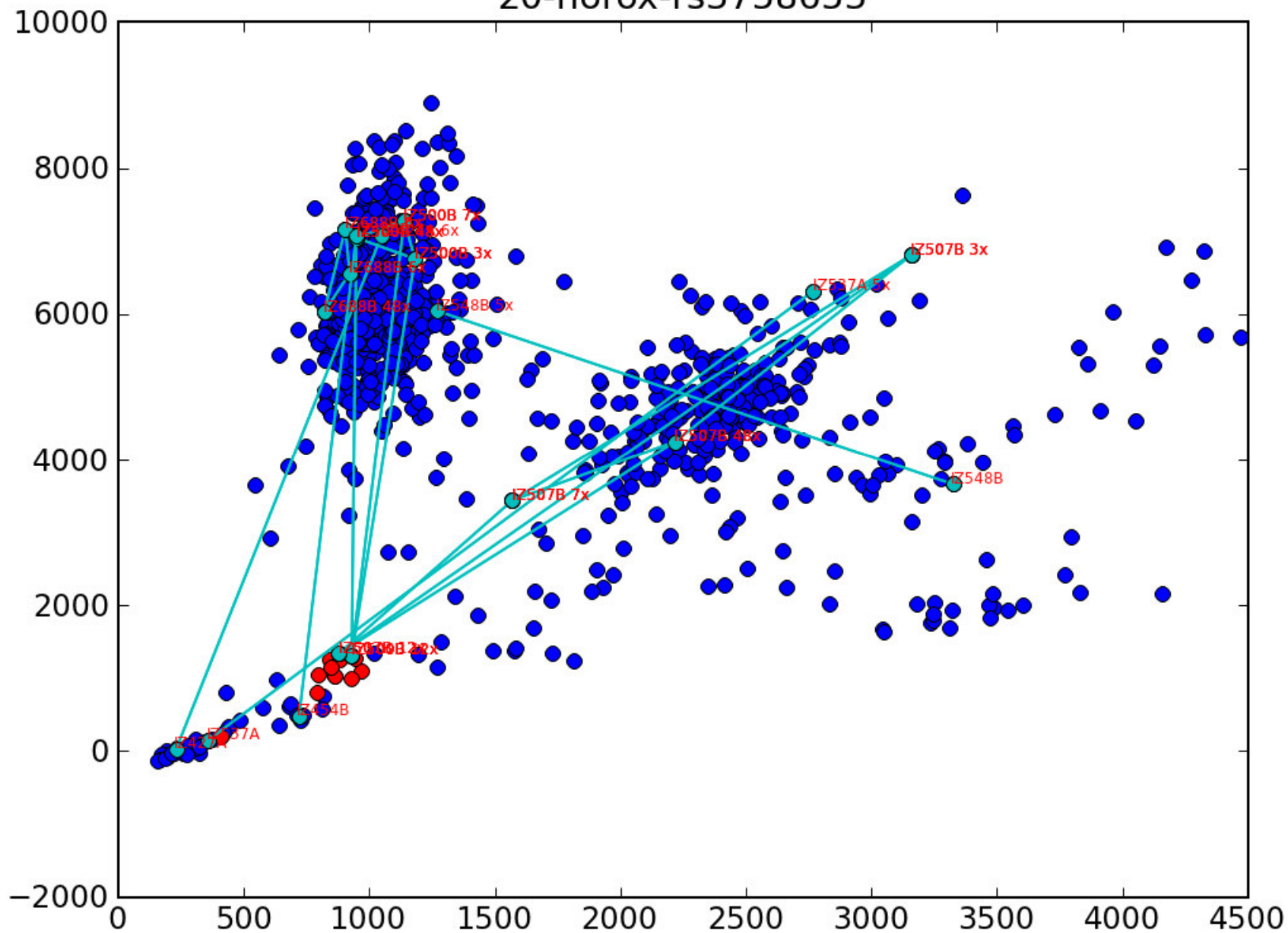
Genotyping: 0, 1 2

.... biomedical data is always uncertain (probabilistic).

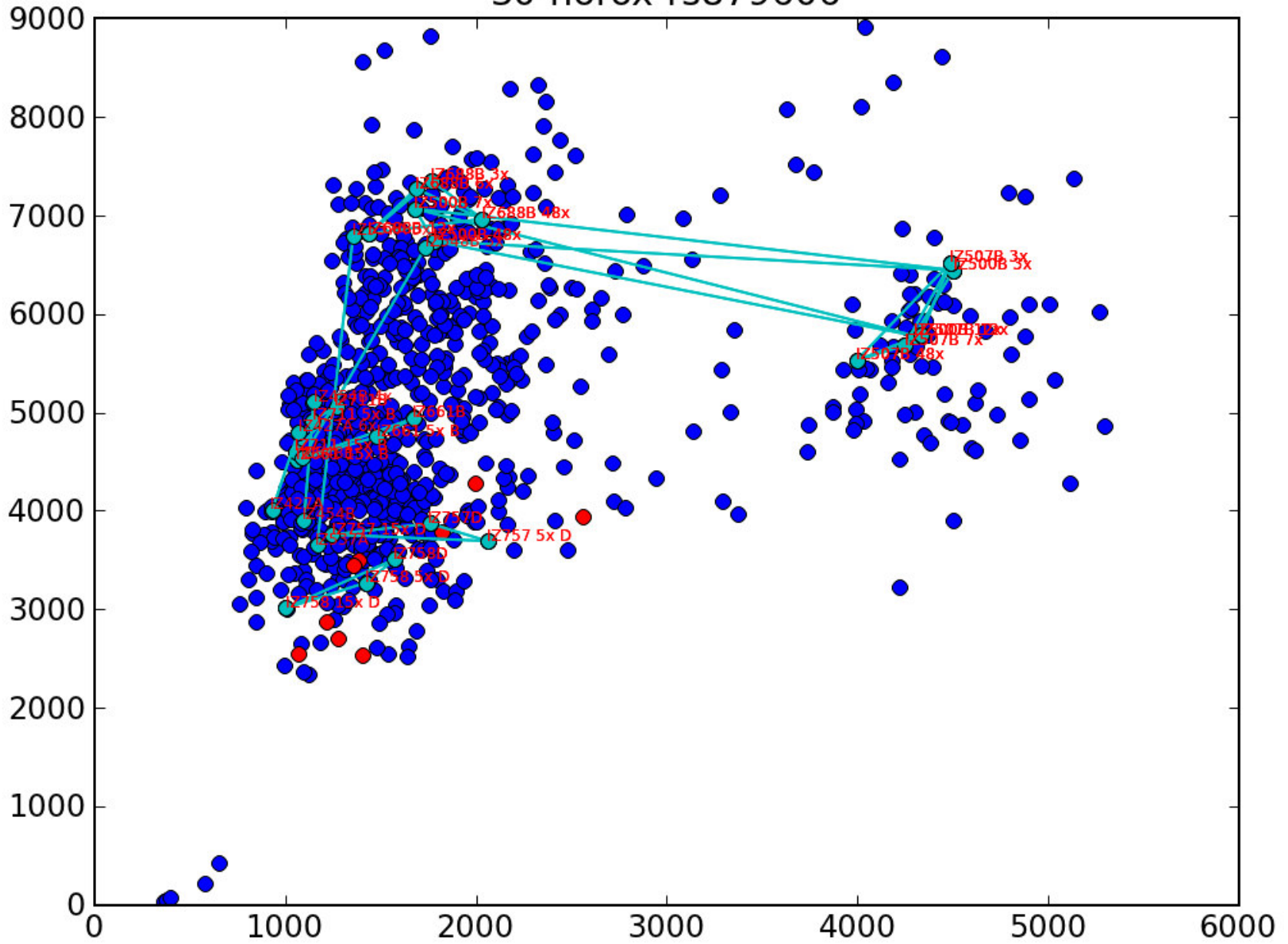
07-norox-rs1611115



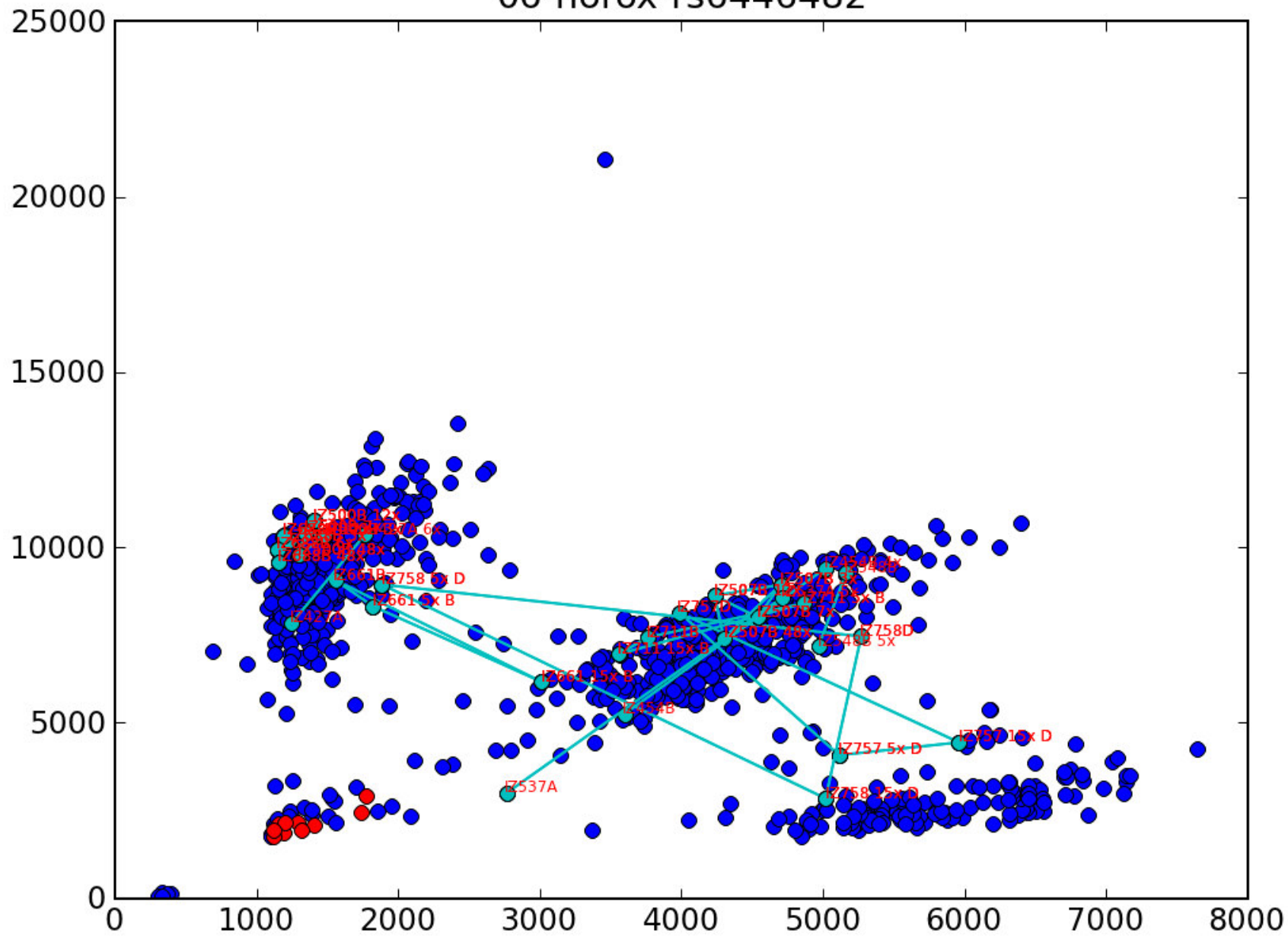
20-norox-rs3758653



30-norox-rs879606



06-norox-rs6446482



Genome-wide associations studies

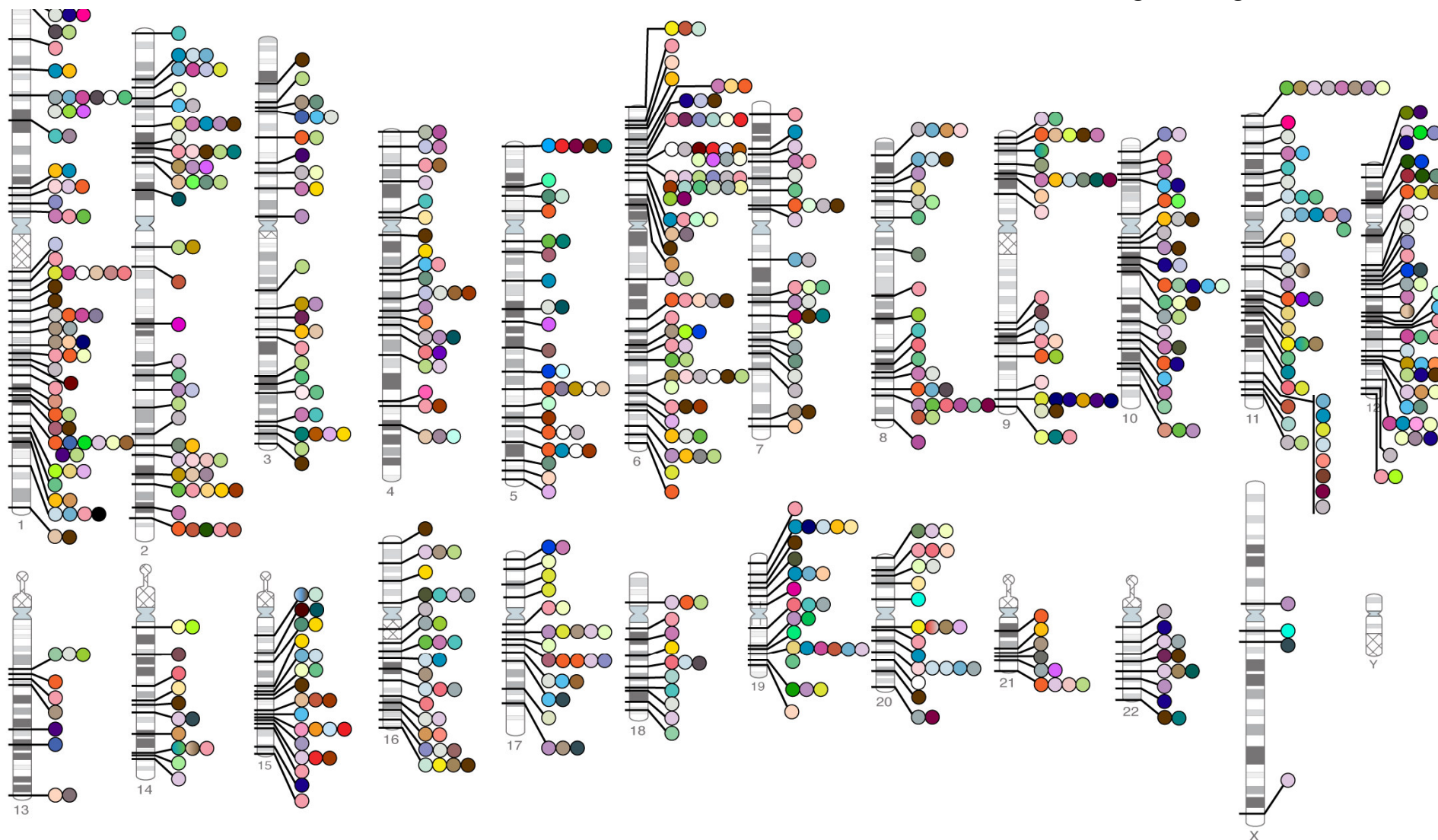
06/2010: 779 studies, $p \leq 5 \times 10^{-8}$ 148 markers

12/2010: 1212 studies, $p \leq 5 \times 10^{-8}$ 210 markers

06/2011: 1449 studies at $p \leq 5 \times 10^{-8}$ for 237 traits

NHGRI GWA Catalog

www.genome.gov/GWASStudies



GWAS catalogue

- www.genome.gov/GWAStudies
 - Catalogue!
 - Test with:
 - Asthma
 - Smoking(!)
 - Intelligence(!??)
 - Autism
 - ...
 - Compare with
 - OMIM
 - SNPedia

- Acute lymphoblastic leukemia
- Adhesion molecules
- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alcohol dependence
- Alzheimer disease
- Amyotrophic lateral sclerosis
- Angiotensin-converting enzyme activity
- Ankylosing spondylitis
- Arterial stiffness
- Asthma
- Atherosclerosis in HIV
- Atrial fibrillation
- Attention deficit hyperactivity disorder
- Autism
- Basal cell cancer
- Bipolar disorder
- Bilirubin
- Bladder cancer
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- C-reactive protein
- Cardiac structure/function
- Carnitine levels
- Carotenoid/tocopherol levels
- Celiac disease
- Chronic lymphocytic leukemia
- Cleft lip/palate
- Cognitive function
- Colorectal cancer
- Coronary disease
- Creutzfeldt-Jakob disease
- Crohn's disease
- Cutaneous nevi
- Dermatitis
- Drug-induced liver injury
- Eosinophil count
- Eosinophilic esophagitis
- Erythrocyte parameters
- Esophageal cancer
- Essential tremor
- Exfoliation glaucoma
- F cell distribution
- Fibrinogen levels
- Folate pathway vitamins
- Freckles and burning
- Gallstones
- Glioma
- Glycemic traits
- Hair color
- Hair morphology
- HDL cholesterol
- Heart rate
- Height
- Hemostasis parameters
- Hepatitis
- Hirschsprung's disease
- HIV-1 control
- Homocysteine levels
- Idiopathic pulmonary fibrosis
- IgE levels
- Inflammatory bowel disease
- Intracranial aneurysm
- Iris color
- Iron status markers
- Ischemic stroke
- Juvenile idiopathic arthritis
- Kidney stones
- LDL cholesterol
- Leprosy
- Leptin receptor levels
- Liver enzymes
- LP (a) levels
- Lung cancer
- Major mood disorders
- Malaria
- Male pattern baldness
- Matrix metalloproteinase levels
- MCP-1
- Melanoma
- Menarche & menopause
- Multiple sclerosis
- Myeloproliferative neoplasms
- Narcolepsy
- Nasopharyngeal cancer
- Neuroblastoma
- Nicotine dependence
- Obesity
- Open personality
- Osteoarthritis
- Osteoporosis
- Otosclerosis
- Other metabolic traits
- Ovarian cancer
- Pain
- Pancreatic cancer
- Panic disorder
- Parkinson's disease
- Periodontitis
- Peripheral arterial disease
- Phosphatidylcholine levels
- Platelet count
- Primary biliary cirrhosis
- PR interval
- Prostate cancer
- Protein levels
- Psoriasis
- Pulmonary funct. COPD
- QRS interval
- QT interval
- Quantitative traits
- Recombination rate
- Red vs.non-red hair
- Renal function
- Response to antipsychotic therapy
- Response to hepatitis C treatment
- Response to statin therapy
- Restless legs syndrome
- Rheumatoid arthritis
- Schizophrenia
- Serum metabolites
- Skin pigmentation
- Speech perception
- Sphingolipid levels
- Statin-induced myopathy
- Stroke
- Systemic lupus erythematosus
- Telomere length
- Testicular germ cell tumor
- Thyroid cancer
- Tooth development
- Total cholesterol
- Triglycerides
- Type 1 diabetes
- Type 2 diabetes
- Ulcerative colitis
- Urate
- Venous thromboembolism
- Vitamin B12 levels
- Warfarin dose
- Weight
- White cell count
- YKL-40 levels

Table 1 | Estimates of heritability and number of loci for several complex traits

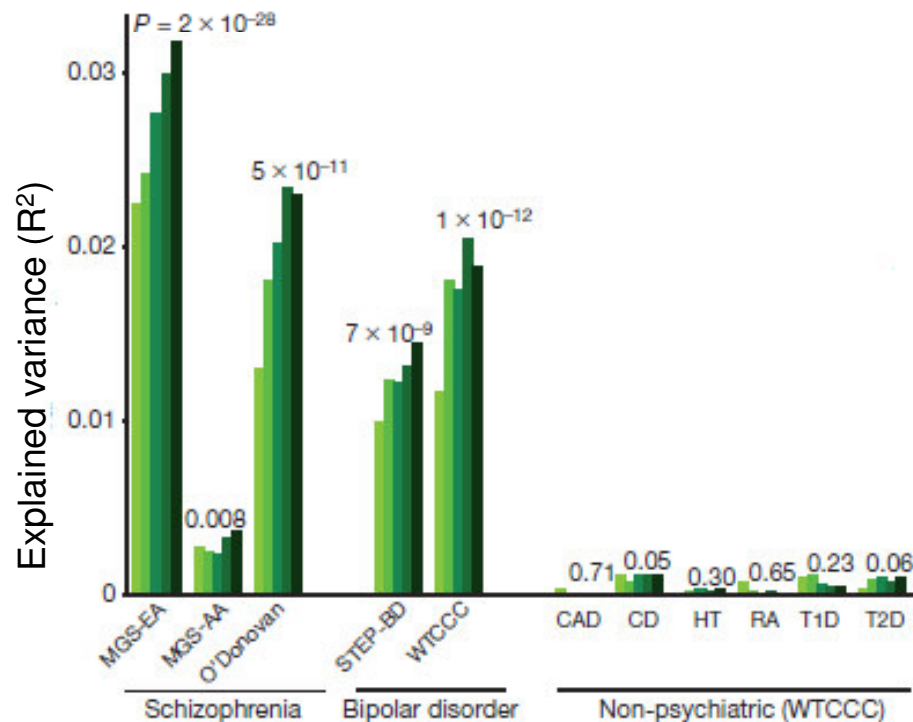
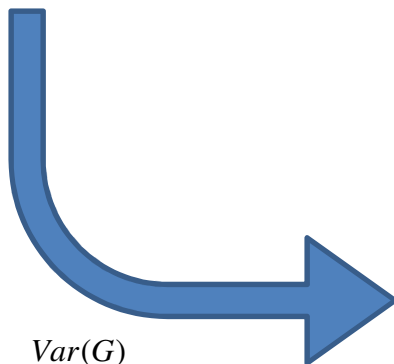
Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration ⁷²	5	50%	Sibling recurrence risk
Crohn's disease ²¹	32	20%	Genetic risk (liability)
Systemic lupus erythematosus ⁷³	6	15%	Sibling recurrence risk
Type 2 diabetes ⁷⁴	18	6%	Sibling recurrence risk
HDL cholesterol ⁷⁵	7	5.2%	Residual* phenotypic variance
Height ¹⁵	40	5%	Phenotypic variance
Early onset myocardial infarction ⁷⁶	9	2.8%	Phenotypic variance
Fasting glucose ⁷⁷	4	1.5%	Phenotypic variance

* Residual is after adjustment for age, gender, diabetes.

The „missing heritability” I.

Disease	Loci	Explained heritance
T2 diabetes	18	6%
HDL cholesterol	7	5.20%
Height	40	5%
Schizophrenia	5	3%

$$H^2 = \frac{Var(G)}{Var(G) + Var(E) + 2Cov(G, E)}$$



The „missing heritability” II.

- **B.Maher: „The case of the missing heritability”, Nature, 2008**
 - „Right under everyone’s noses”: rare variants (RVs)
 - „Out of sight”: many, small effects
 - „In the architecture”: structural variations
 - „In underground networks”: epistatic
 - „The great beyond”: Epigenetics
 - „Lost in diagnosis”: phenome
- „Rare Variants Create Synthetic Genome-Wide Associations”, PLoS, 2009
- „Finding the missing heritability of complex diseases”, Nature, 2009
- McClellan&King, Cell, 2010:
 - **The CD-CV hypothesis (and the corresponding GWAS) has failed.**

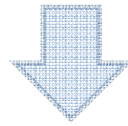
Missing heritability: issues

- Better outcome variable
 - „Lost in diagnosis”: phenome
- Better and more complete set of predictor variables
 - „Right under everyone’s noses”: rare variants (RVs)
 - „The great beyond”: Epigenetics, environment
- Better statistical models
 - „In the architecture”: structural variations
 - „Out of sight”: many, small effects
 - „In underground networks”: epistatic interactions
- **Causation (confounding)**
- **Statistical significance („multiple testing problem”)**
- **Complex models: interactions, epistasis**
- **Interpretation**

GAS challenges

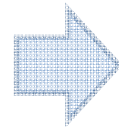
Genetic factors

- Broader scale
- Aggregation: functional, pathway



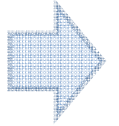
Omic

- Cost
- Quality
- Broader



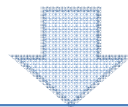
Environment – life style

- Standardization



Modeling

- Multivariate
- Systems biology
- Model averaging
- Fusion
- Causation

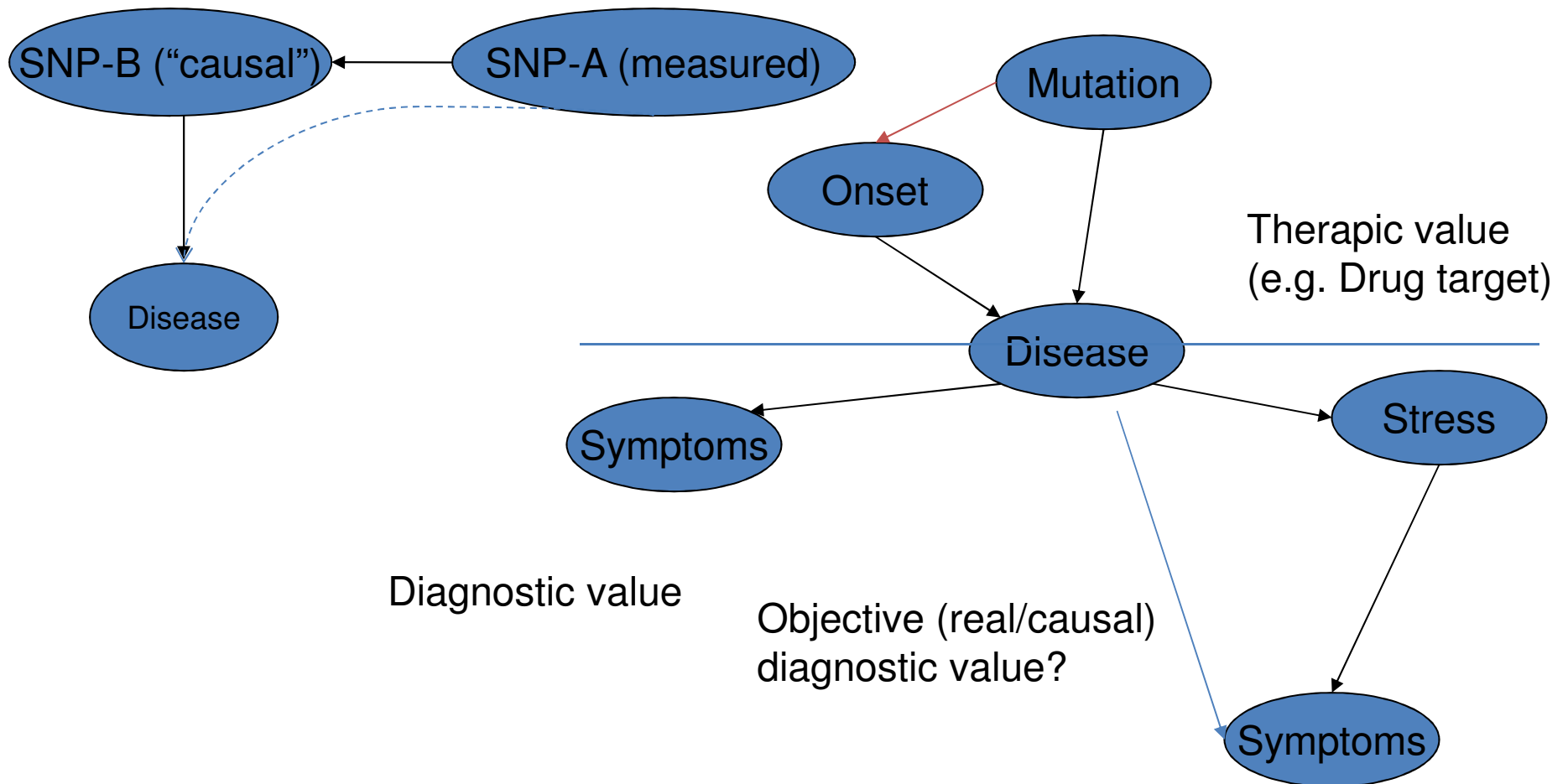


Phenotypes

- Refinement
- Standardization

Causal vs. diagnostic markers

Direct \neq Causal

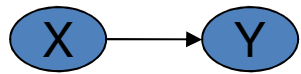


Principles of causality

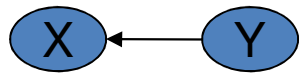
- Principles for a causal relation between $X \rightarrow Y$:
 - **Probabilistic association**,
 - **Temporal asymmetry**: X precedes temporally Y,
 - ~~(Physical locality)~~
 - Quantitative effect of interventions: **dose-effect relation**
 - necessity (i.e., if the cause is removed, effect is decreased)
 - sufficiency (if exposure to cause is increased, effect is increased)
 - **Counterfactuals**:
 - Y would not have been occurred with that much probability if Y hadn't been present
 - Y would have been occurred with larger probability if X had been present
 - **Bounded context-sensitivity** (\sim context-free): relevant on average
 - Plausible **explanation** (no alternative based on confounding).

Association vs. Causation

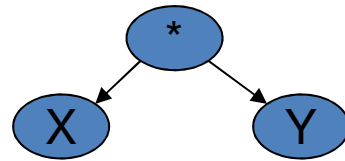
Causal models:



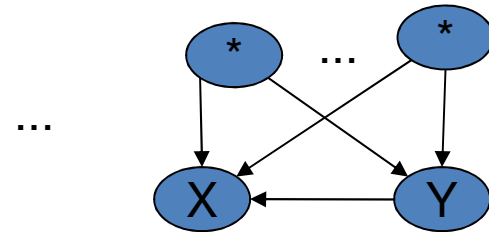
X causes Y



Y causes X

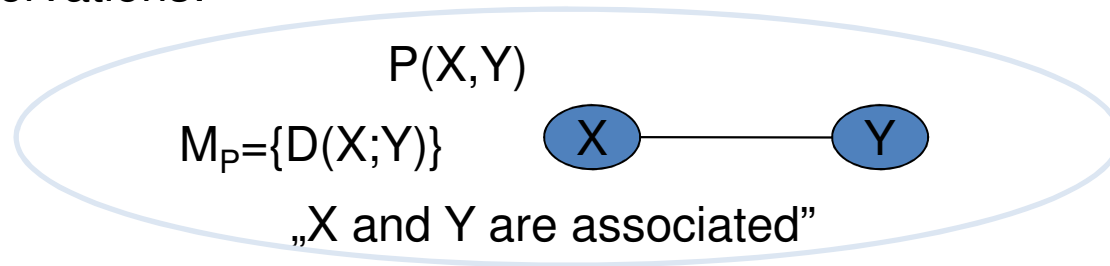


There is a common cause
(pure confounding)



Causal effect of Y on X
is confounded by many
factors

From passive observations:



Reichenbach's Common Cause Principle:

a correlation between events X and Y indicates either that X causes Y , or that Y causes X , or that X and Y have a common cause.

Statistical vs. Causal

- What is the advantage of causal knowledge?

Assume X (directly) causes Y, $P(X,Y)$

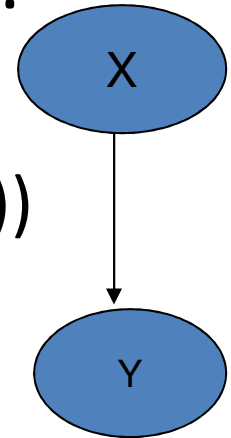
– Setting X will influence Y: $P(Y | X=x) = P(Y | \text{do}(X=x))$

– But setting Y will NOT influence X:

$$P(X | Y=y) = P(X)$$

– (Directness) Fixing all contextual variables will not screen off X.

- How can we identify a causal relation?



Intervention

Randomized experiment (C.S.Pierce.,R.A.Fisher):

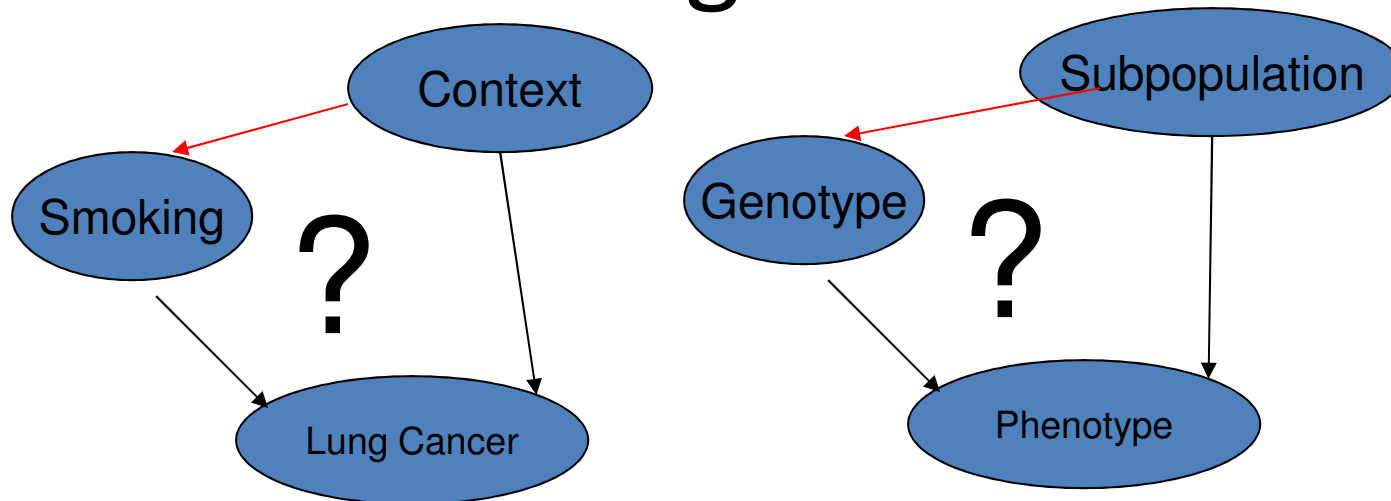
In context c randomly setting X to estimate its average effect on Y :

$$P(Y | \text{do}(X=x),c) = \sum_c P(Y | x,c)p(c) ,$$

known as adjusting, controlling, keeping constant.

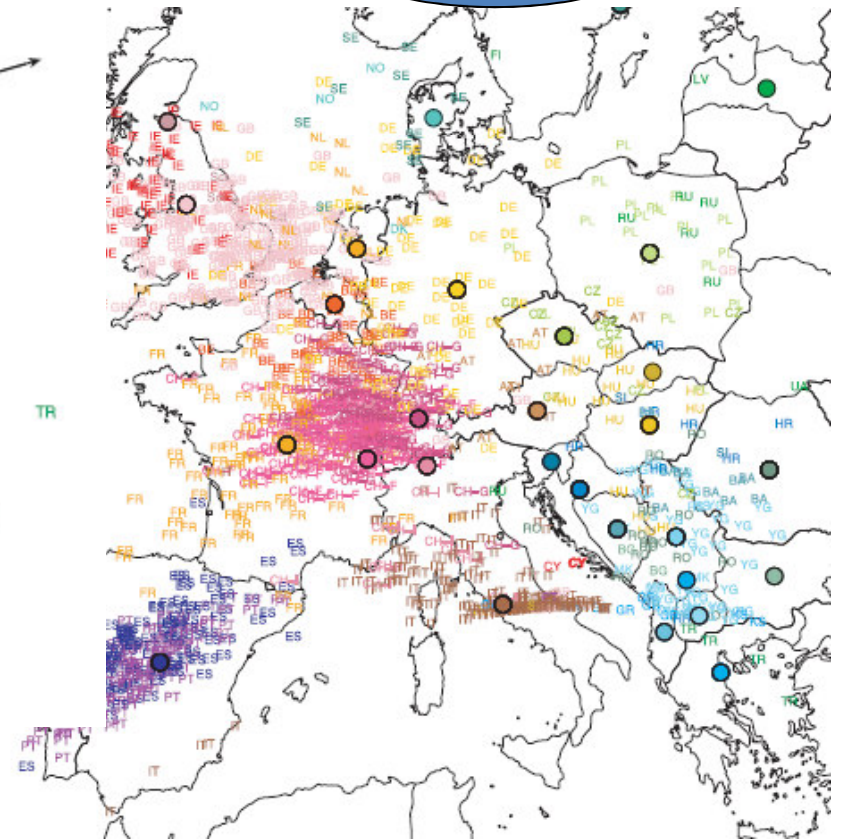
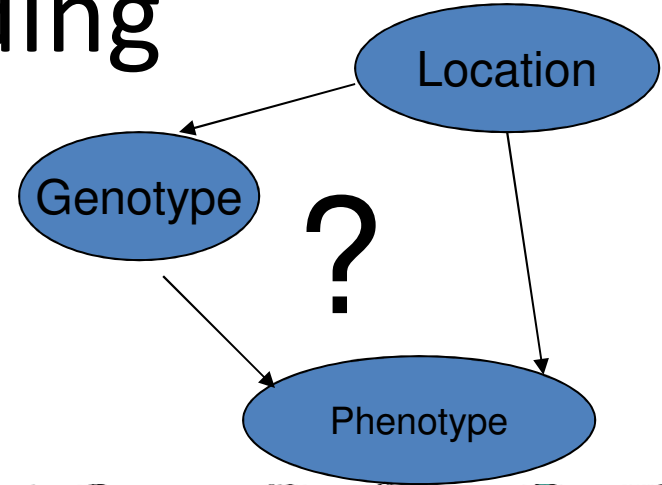
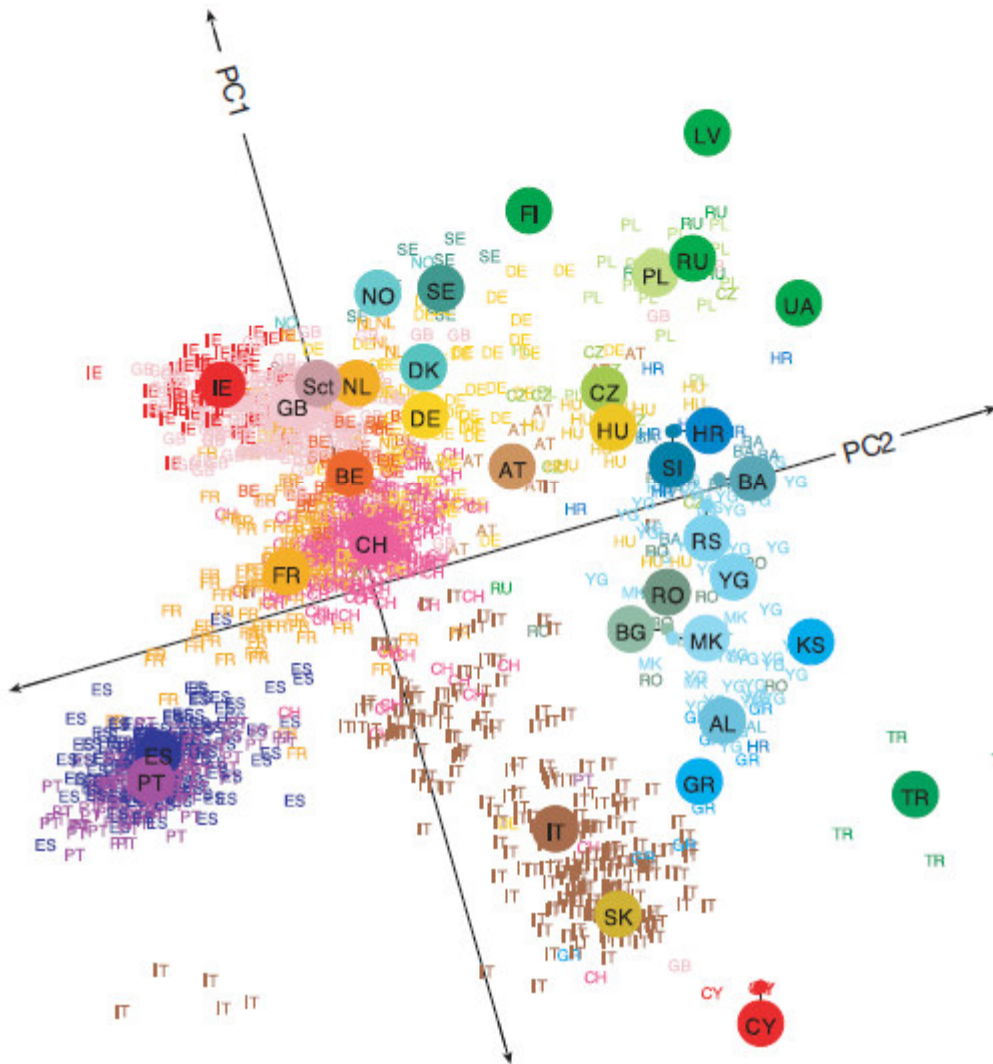
A causal relation is an asymmetric and autonomous (modular and transportable) **mechanism**.

Confounding



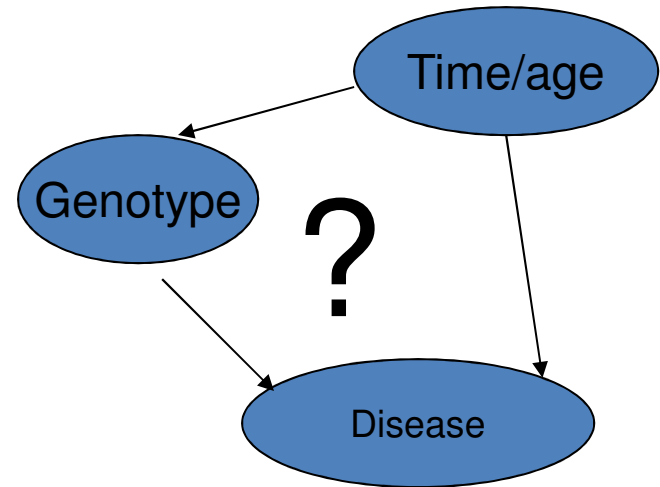
- The effect of Smoking on LungCancer:
 - $P(LC|S,Context=c)$ vs. $P(LC|\neg S,Context=c)$ in all context c (context dependent).
- Solution: average the effect in all context
 - Stratification: collect complete observation about context, and analyze them separately.
 - Nature set the cause depending on the context.
 - Intervention: collect data only about cause-effect and analyze them jointly.
 - We set the cause and eradicate the effect of context.
- Identification of causal effect?
 - Computability of a „causal” quantity $P(Y|do(X))$ from „statistical” quantities such as $P(Y|X)$?

Geographic confounding



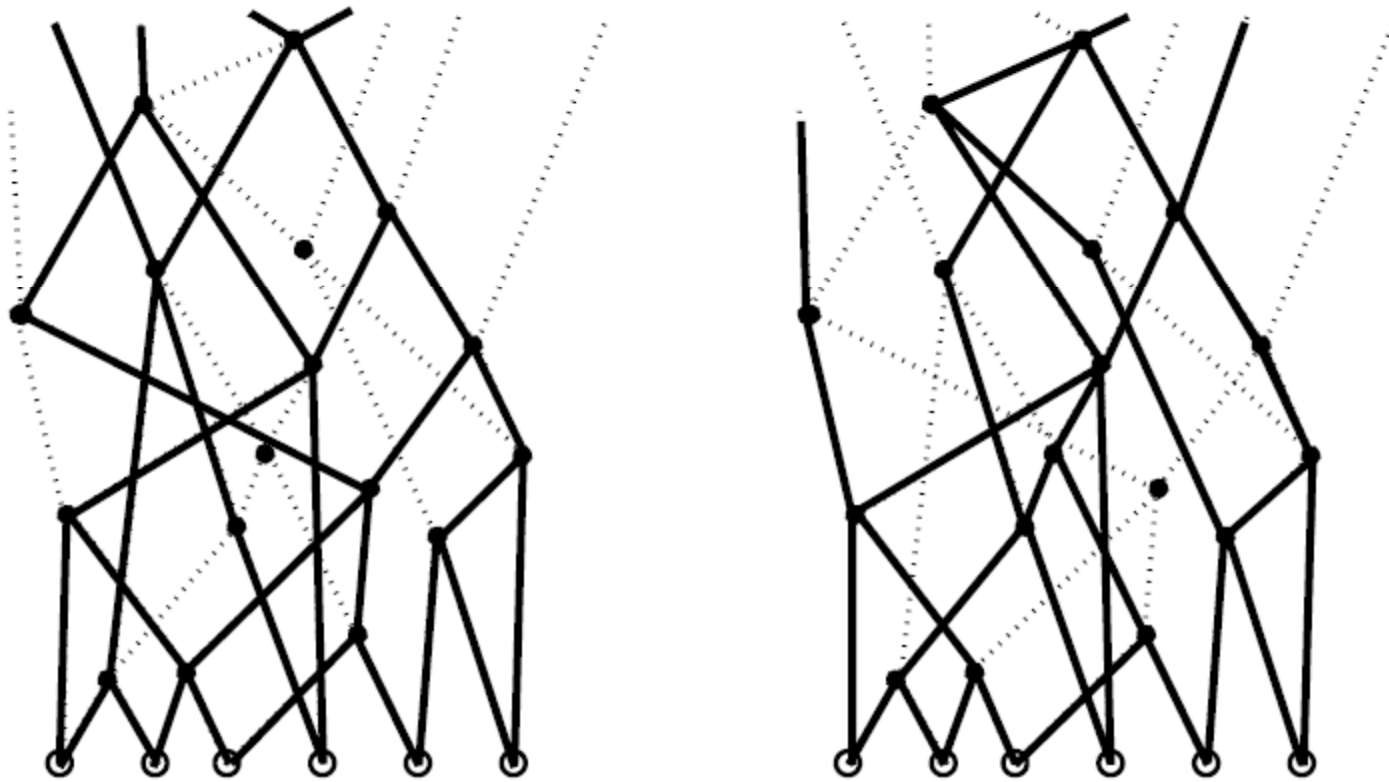
Confounding by „time/age”

- Rapid/structured change in the population (genotype) can create dependency $p(\text{Genotype} | \text{Time})$.
- If dependency $p(\text{Disease} | \text{Onset})$ exists, then the relation $p(D | G)$ is confounded.



Confounding:

Population substructure + cryptic relatedness



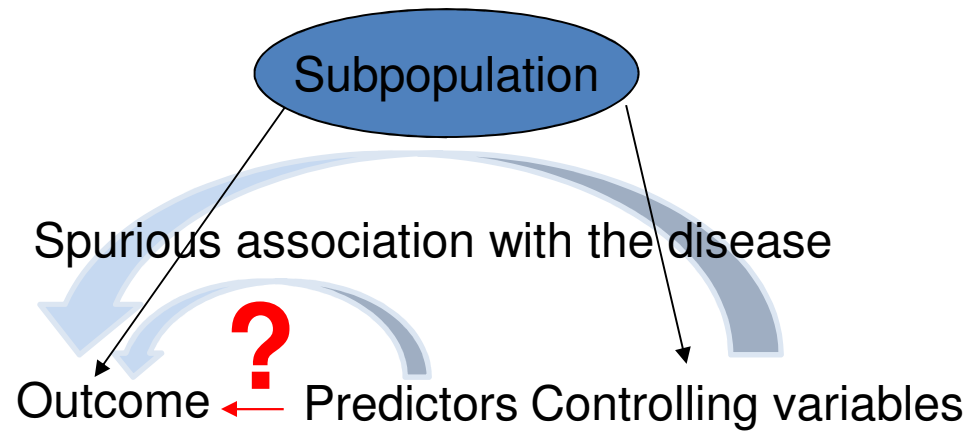
Genomwide

and

allelic pedigrees.

Balding, 2010: Allowing for population structure and cryptic relatedness in GAS

Confounding

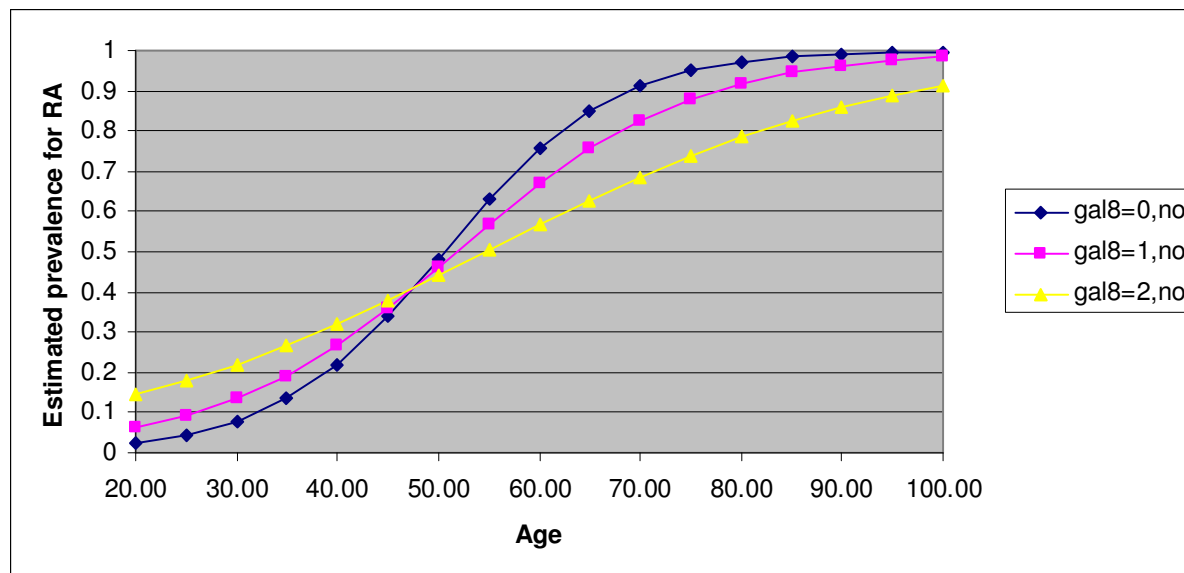


Y	X ₁					X _n
 	 					
 	 					
 	 					
 	 					
 	 					

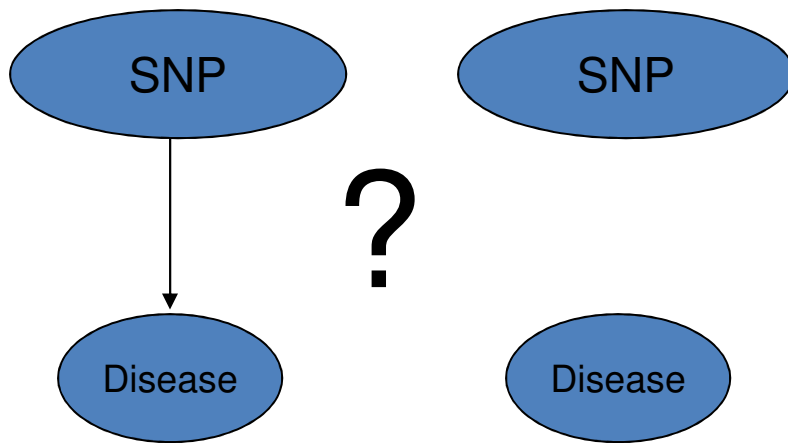
Samples

Pleiotropy

- Pleiotropy: different phenotypic effects
- Antagonistic pleiotropy: opposite effects
- E.g.: a SNP can be protective for young people, but risk for old people.



Testing genetic associations



Genotype	D	$\neg D$	
0	N_{11}	N_{12}	$N_{1.}$
1	N_{21}	N_{22}	$N_{2.}$
2	N_{31}	N_{32}	$N_{3.}$

Number of alleles („high risk” allele is A): $aa=0, aA=1, AA=2$

Allele frequency:

$p(A | \neg D)$ vs. $p(A | D)$

Hardy-Weinberg equilibriums for $p(G | D)$.

Disease frequencies:

$p(D | G=0)$ vs. $p(D | G=1)$ vs. $p(D | G=2)$

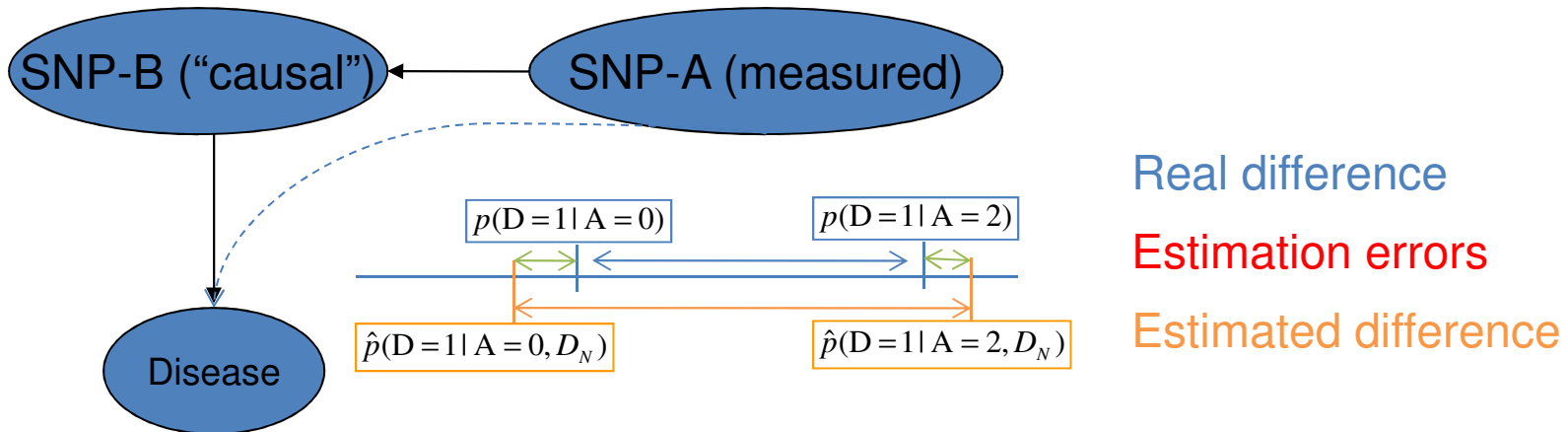
Recessive model:

$p(D | G=0)$ vs. $p(D | G=1 \text{ or } G=2)$

Dominant model:

$p(D | G=0 \text{ or } G=1)$ vs. $p(D | G=2)$

Fundamental questions in statistics



Estimation error because of finite data D_N : $\hat{p}(D=1|A=0, D_N) - p(D=1|A=0)$

Inequalities for finite(!) data (ϵ accuracy, δ confidence)
 sample complexity: $N_{\epsilon, \delta}$ $p(D_{N_{\epsilon, \delta}} : \epsilon < |\hat{p}(D=1|A, D_{N_{\epsilon, \delta}}) - p(D=1|A)| < \delta$

The hypothesis testing framework

- Terminology:

- False/true x positive/negative
- Null hypothesis: independence

reported	Ref.:0/N	Ref.1/P
0/N	TN	FN
1/P	FP	TP

- Type I error/error of the first kind/ α error/FP: $p(\neg H_0 | \underline{H}_0)$
 - Specificity: $p(H_0 | \underline{H}_0) = 1 - \alpha$
 - Significance: α
 - p-value: „probability of more extreme observations in repeated experiments”
- Type II error/error of the second kind/ β error/FN: $p(H_0 | \neg \underline{H}_0)$:
 - Power or sensitivity: $p(\neg H_0 | \neg \underline{H}_0) = 1 - \beta$

reported	Ref. \underline{H}_0	Ref.: $\neg \underline{H}_0$
H_0		Type II
$\neg H_0$	Type I („false rejection”)	

Genetic power calculator I.

High risk allele frequency (A) : (0 - 1)

Prevalence : (0.0001 - 0.9999)

Genotype relative risk Aa : (>1)

Genotype relative risk AA : (>1)

D-prime : (0 - 1)

Marker allele frequency (B) : (0 - 1)

Number of cases : (0 - 10000000)

Control : case ratio : (>0)
(1 = equal number of cases and controls)

Unselected controls? (* see below)

User-defined type I error rate : (0.00000001 - 0.5)

User-defined power: determine N : (0 - 1)
(1 - type II error rate)

- <http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html>
- <http://pngu.mgh.harvard.edu/~purcell/gpc/>

Multiple testing problem (MTP)

- If we perform N tests and our goal is
 - $p(\text{FalseRejection}_1 \text{ or } \dots \text{ or } \text{FalseRejection}_N) < \alpha$
- then we have to ensure, e.g. that
 - for all $p(\text{FalseRejection}_i) < \alpha/N$

➔ loss of power!

E.g. in a GWA study $N=100,000$, so huge amount of data is necessary....(but high-dimensional data is only relatively cheap!)

Solutions for MTP

- Corrections
- Permutation tests
 - Generate perturbed data sets under the null hypothesis: permute predictors and outcome.
- False discovery rate, q-value
- Bayesian approach

Corrections for multiple testing

A. Bonferroni correction

The p-value of each gene is multiplied by the number of genes in the gene list. If the corrected p-value is still below the error rate, the gene will be significant:

$$\text{Corrected P-value} = p\text{-value} * n \text{ (number of genes in test)} < 0.05$$

Bonferroni
Bonferroni Step-Down
Westfall and Young Permutation
Benjamini and Hochberg False Discovery Rate
None



D. Benjamini and Hochberg False Discovery Rate

This correction is the least stringent of all 4 options, and therefore tolerates more false positives. There will be also less false negative genes. Here is how it works:

- 1) The p-values of each gene are ranked from the smallest to the largest.
- 2) The largest p-value remains as it is.
- 3) The second largest p-value is multiplied by the total number of genes in gene list divided by its rank. If less than 0.05, it is significant.

$$\text{Corrected p-value} = p\text{-value} * (n/n-1) < 0.05, \text{ if so, gene is significant.}$$

- 4) The third p-value is multiplied as in step 3:

$$\text{Corrected p-value} = p\text{-value} * (n/n-2) < 0.05, \text{ if so, gene is significant.}$$

And so on.

Corrections for multiple testing

I have 1,000,000 hypotheses that are not mutually exclusive.

1. I test them all.

Correction?

2. I plan to test them all, but I run out of resources after testing only one of them.

Correction?

3. I test one of them, and a year later test the others.

Correction? If so, when?

4. I only test the first one because that is the one I suspect.

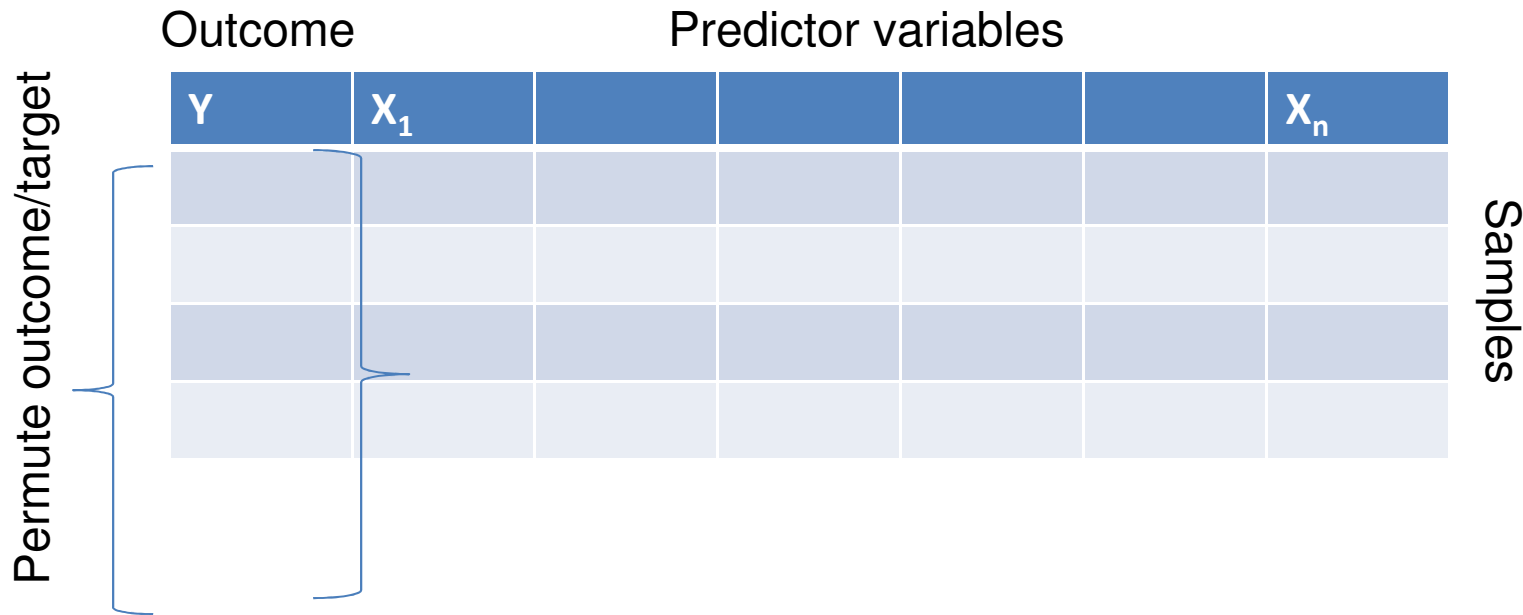
Correction?

5. I run an algorithm that prunes unlikely hypotheses, keeping only 100,000.

Correction for 100,000 or for 1,000,000 hypotheses?

(R.Neapolitan, 2010)

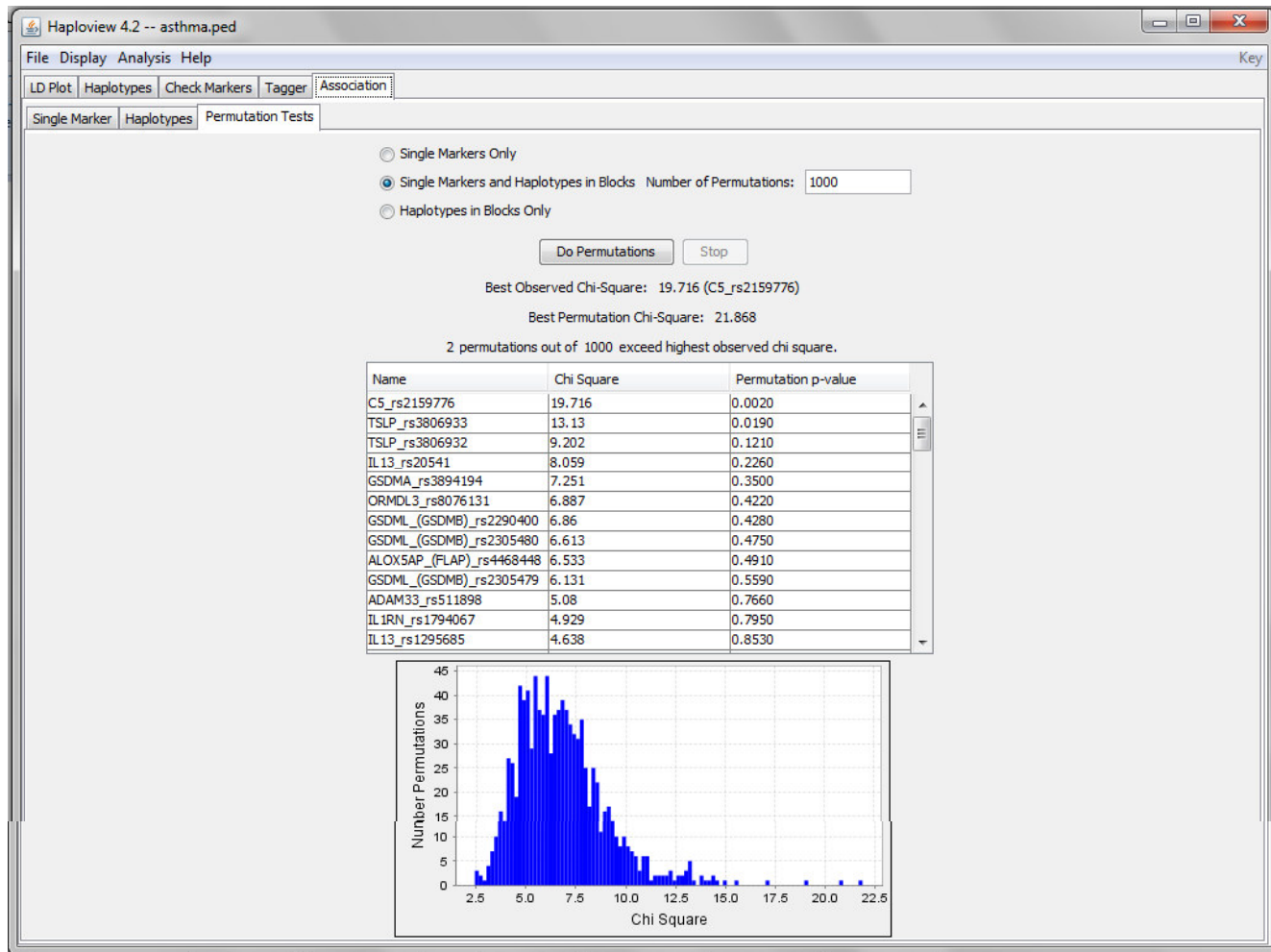
Permutation testing



- A random permutation guarantees the independency of the outcome Y.
- A random permutation corresponds to an artificial data set from the null model. → "direct" estimation of the p-value: the probability of observing a more extreme data set from the null model with the same sample size.

$$p(D_N^{perm} : IncompatibilityWithNull(D_N^{real}) < IncompatibilityWithNull(D_N^{perm}))$$

Permutation testing in Haploview



File|Open|Linkage Data: <http://home.mit.bme.hu/~antal/AIT/asthma.ped>
Options: Do association test, Case/Control

False discovery rate (FDR)

Another aspect of multiple hypothesis testing:

- the probability of Type I. error for any tests
- the expected number of Type I. errors at a given significance level (False discovery rate, FDR)
- q-value: the minimum FDR at which the test may be called significant.

Summary

- The problem of missing heritability
 - Potential explanations/solutions
- The problem of confounding
 - Population substructure
 - Solutions
- The multiple hypothesis testing problem
 - Concept of permutation test (permutation p-value)
 - False discovery rate, q-value