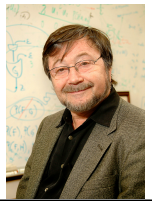
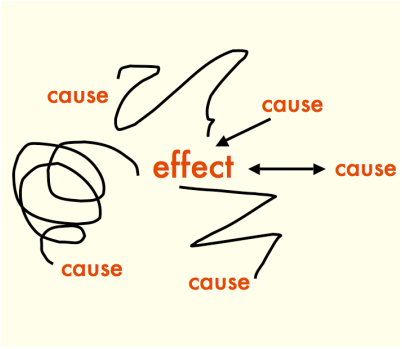
"The research questions that motivate most quantitative studies in the health, social and behavioral sciences are not statistical but causal in nature. For example, what is the efficacy of a given drug in a given population? Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual, in a specific incident? These are causal questions because they require some knowledge of the data-generating process; they cannot be computed from the data alone."



--Judea Pearl

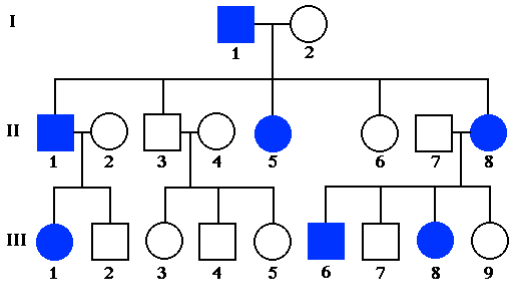
What is causality?

- What do you think when I say:
  - Smoking ***causes*** lung cancer?



Genes and causality

SYNDROMIC / MONOGENIC DISEASE

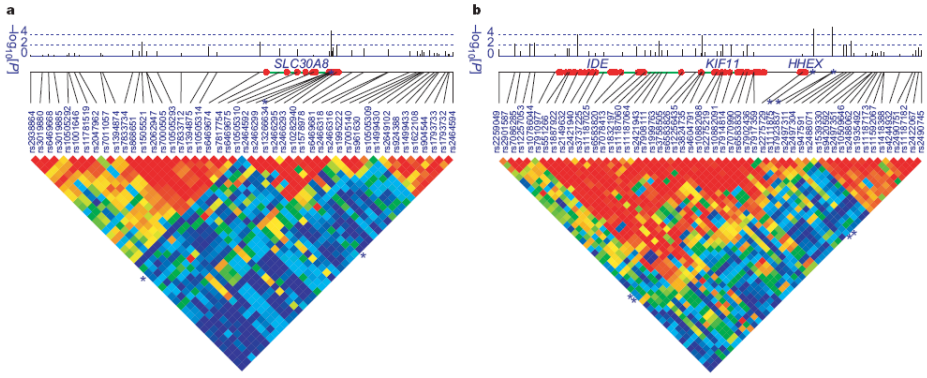


Beyond co-segregation studies, additional arguments are needed to demonstrate the causal role of a mutation in the disease

⇒ **functional genomics**

Genes and causality

POLYGENIC DISEASE (e.g. type 2 diabetes)



Beyond association studies, additional arguments are needed to demonstrate the causal role of a variant / gene in the disease

⇒ **functional genomics**

Sladek et al., *Nature* 2007

Statistical Association

- X and Y are associated:
  - Observing X may change conditional distribution of observed values of Y: $P(Y|X) \neq P(Y)$
  - Knowledge of X provides information on Y
  - Observing X allows you to predict Y (& vice versa)
  - Knowing X changes your belief for the distribution of Y

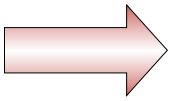
Association is NOT causation

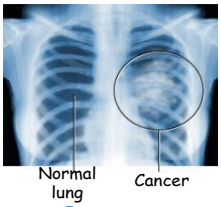
- Yellow teeth and lung cancer are associated
- Can I bleach my teeth and reduce the probability of lung cancer?
- Is smoking really causing lung cancer?

If A and B are correlated:
A causes B (causation),
B causes A (reverse causation),
they share a latent common cause,
or the correlation is coincidental.

Temporality

- This refers to the necessity for the exposure to precede the outcome (effect) in time
- Any claim of causation must involve the cause preceding in time the presumed effect
- Easier to establish in certain study designs
  - Prospective cohort study

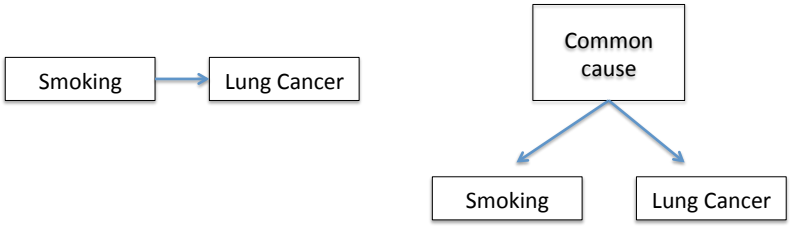




Exposure TIME Outcome

Temporal ordering

- Assume smoking precludes lung cancer:



How to learn causality?

- Take 200 people
- Randomly split them into control and treatment groups
- Force control group to smoke and treatment group to not smoke
- Wait until they are 60 years old
- Measure correlations / statistics.

This is classical randomized controlled trial.

Evidence for causality?

I-TRANS-ETHNIC FINE MAPPING APPROACH

II-EVOLUTIONARY GENETICS

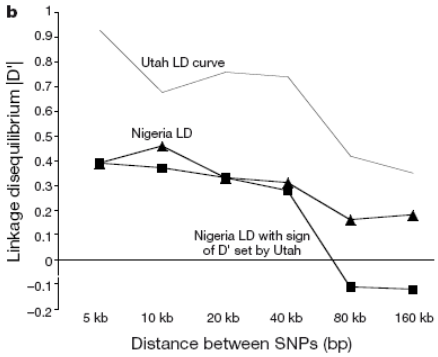
III-GENE VARIANT AND FUNCTION

IV-GENE CANDIDACY

V-STUDY OF ENDOPHENOTYPES

I-TRANS-ETHNIC FINE MAPPING APPROACH

Trans-ethnic fine mapping approach



letters to nature

Linkage disequilibrium in the human genome

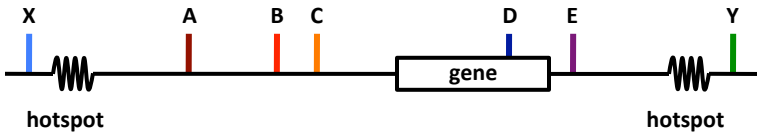
David E. Reich*, Michele Cargill*†, Stacey Bolk*, James Ireland*, Pardis C. Sabeti‡, Daniel J. Richter*, Thomas Lavery*, Rose Kouyoumjian*, Shelli F. Farhadian*, Ryk Ward‡ & Eric S. Lander*§

Linkage disequilibrium is the non-random association of alleles at two or more loci

The human genome is composed of blocks of linkage disequilibrium

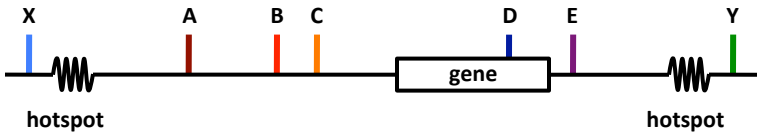
The extent of linkage disequilibrium blocks varies according to the ethnic background

Linkage disequilibrium



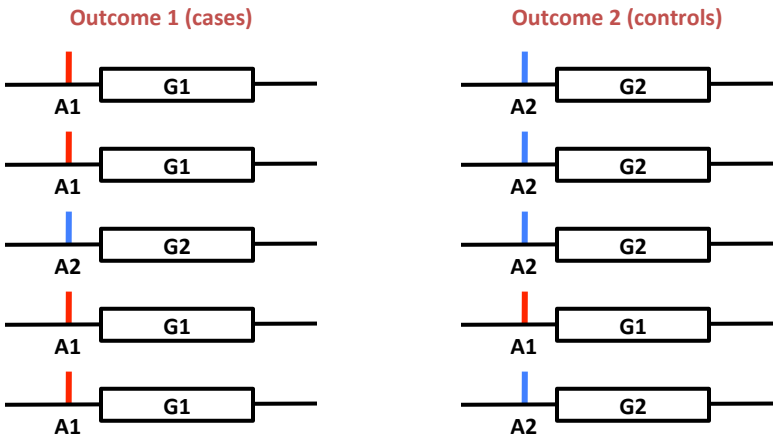
- Groups of SNPs are separated by recombination "hotspots" – the SNPs tend to be passed from parents to children as a unit, called a "linkage disequilibrium" block
- Thus, SNPs A, B, C, D, E tend to stay together as a group – the bases at these SNPs make up a "haplotype" – whereas SNPs X and Y are not linked (not in linkage disequilibrium)
- Because SNPs A–E are all linked, only one of these SNPs is needed to act as a "tag" SNP for the whole group

Haploblock structure



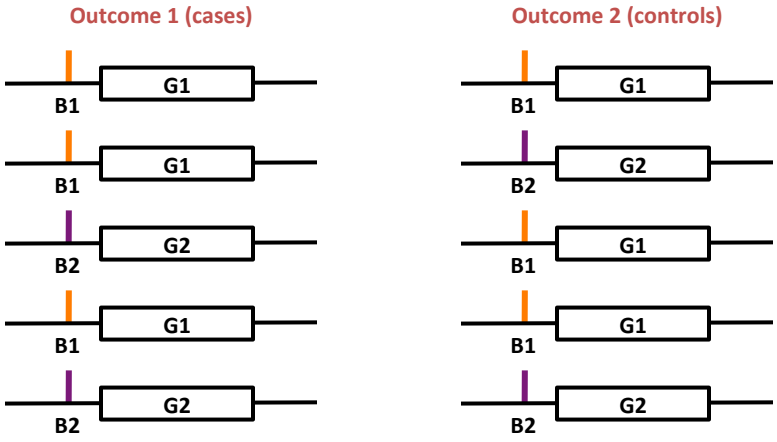
- Even though (in this example) SNPs A, B, C, and E are not in the gene, each is in "linkage disequilibrium" with the gene and remains associated with the gene as it is passed from parents to children
- If the gene causes a particular outcome (e.g., higher risk for a disease), SNPs A–E will be associated with that outcome
- This is the basis for genome wide association studies

GWAS to Find Outcome-Associated SNPs



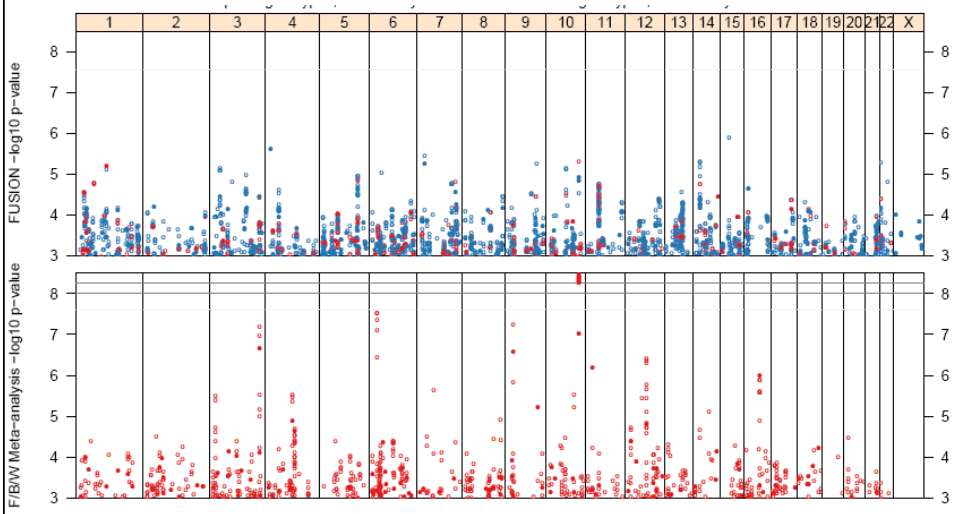
Association between SNP variant A1 / gene variant G1 and outcome 1

GWAS to Find Outcome-Associated SNPs



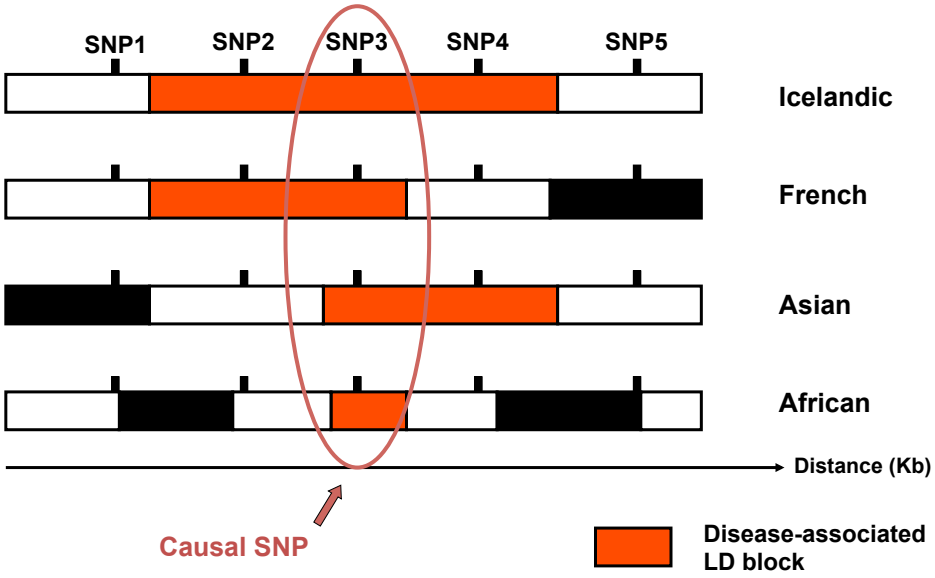
No association between SNP variant B1 / gene variant G1, and outcome 1

Results of Genome-Wide Association of Type 2 Diabetes with 317,503 SNPs



Stage 2 – FUSION + DGI + WTCCC
(4549 cases + 5579 controls)

Trans-ethnic fine mapping approach



Trans-ethnic fine mapping approach

ORIGINAL ARTICLE

Resequencing and Analysis of Variation in the *TCF7L2* Gene in African Americans Suggests That SNP rs7903146 Is the Causal Diabetes Susceptibility Variant

Nicholette D. Palmer,[1,2,3] Jessica M. Hester,[2,3,4] S. Sandy An,[1,2,3] Adebowale Adeyemo,[5] Charles Rotimi,[5] Carl D. Langefeld,[6] Barry I. Freedman,[7] Maggie C.Y. Ng,[2,3,8] and Donald W. Bowden[1,2,3,9]

Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution

Agnar Helgason[1], Snæbjörn Pálsson[1,2], Gudmar Thorleifsson[1], Struan F A Grant[1,13], Valur Emilsson[1], Steinunn Gunnarsdottir[1], Adebowale Adeyemo[3], Yuanxiu Chen[3], Guanjie Chen[3], Inga Reynisdottir[1], Rafn Benediktsson[4,5], Anke Hinney[6], Torben Hansen[7], Gitte Andersen[7], Knut Borch-Johnsen[7,8], Torben Jorgensen[9], Helmut Schäfer[10], Mezbah Faruque[3], Ayo Doumatey[3], Jie Zhou[3], Robert L Wilensky[11], Muredach P Reilly[11], Daniel J Rader[11], Yu Bagger[12], Claus Christiansen[12], Gunnar Sigurdsson[4,5], Johannes Hebebrand[6], Oluf Pedersen[7,8], Unnur Thorsteinsdottir[1], Jeffrey R Gulcher[1], Augustine Kong[1], Charles Rotimi[3] & Kári Stefánsson[1]

Large-scale resequencing and case control association studies in Icelandic, Danish, West African and American African subjects identified the rs903146 as the likely causal type 2 diabetes-associated SNP

II-EVOLUTIONARY GENETICS

Evolutionary genetics

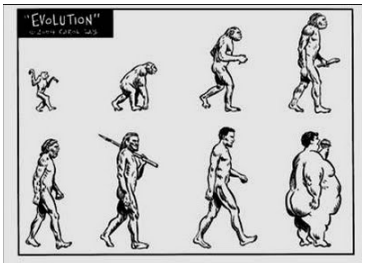
Natural selection is the gradual, non-random process by which biological traits become either more or less common in a population as a function of differential reproduction of their bearers. It is a key mechanism of evolution. The term "natural selection" was popularized by **Charles Darwin**.

Evolutionary genetics (Huxley 1942)

-advantageous mutations have been positively selected in human populations during recent evolution

-disadvantageous mutations have been negatively selected in human populations during recent evolution

Evolutionary genetics



THRIFTY GENOTYPE HYPOTHESIS: the 'thrifty' genotype would have been advantageous for hunter-gatherer populations, especially child-bearing women, because it would allow them to fatten more quickly during times of abundance. Fatter individuals carrying the thrifty genes would thus better survive times of food scarcity.

⇒ Obesity and type 2 diabetes predisposing mutations may show evidence of positive signature of evolution

Evolutionary genetics

Evidence of Still-Ongoing Convergence Evolution of the Lactase Persistence T_{-13910} Alleles in Humans

Nabil Sabri Enattah, Aimee Trudeau, Ville Pimenoff, Luigi Maiuri, Salvatore Auricchio, Luigi Greco, Mauro Rossi, Michael Lentze, J. K. Seo, Soheila Rahgozar, Insaf Khalil, Michael Alifrangis, Sirajedin Natah, Leif Groop, Nael Shaat, Andrew Kozlov, Galina Verschubskaya, David Comas, Kazima Bulayeva, S. Qasim Mehdi, Joseph D. Terwilliger, Timo Sahi, Erkki Savilahti, Markus Perola, Antti Sajantila, Irma Järvelä, and Leena Peltonen

The *LCT* rs4988235 T variant confers lactase persistence

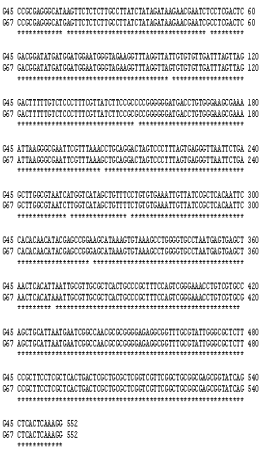
The *LCT* rs4988235 T variant is associated with more milk / dairy products consumption and increased body mass index

The *LCT* rs4988235 T variant has a selective advantage in milk-producing dairy farming populations and has been submitted to positive selection in relation with events of cattle domestication

The *LCT* rs4988235 T allele frequency is more frequent in Northern (MAF: 0.7) than in Southern Europe (MAF: 0.1)

III-GENE VARIANT AND FUNCTION

Gene variant and function



⇒ **Missense, nonsense, frameshift (indels) coding mutations: altered protein function**

⇒ **Intron / exon mutations: exon skipping**

⇒ **Copy Number Variants (CNV): modulation of gene expression, haplo-insufficiency**

⇒ **gene variant in the promoter (Transcription Factor Biding Site): change in gene expression**

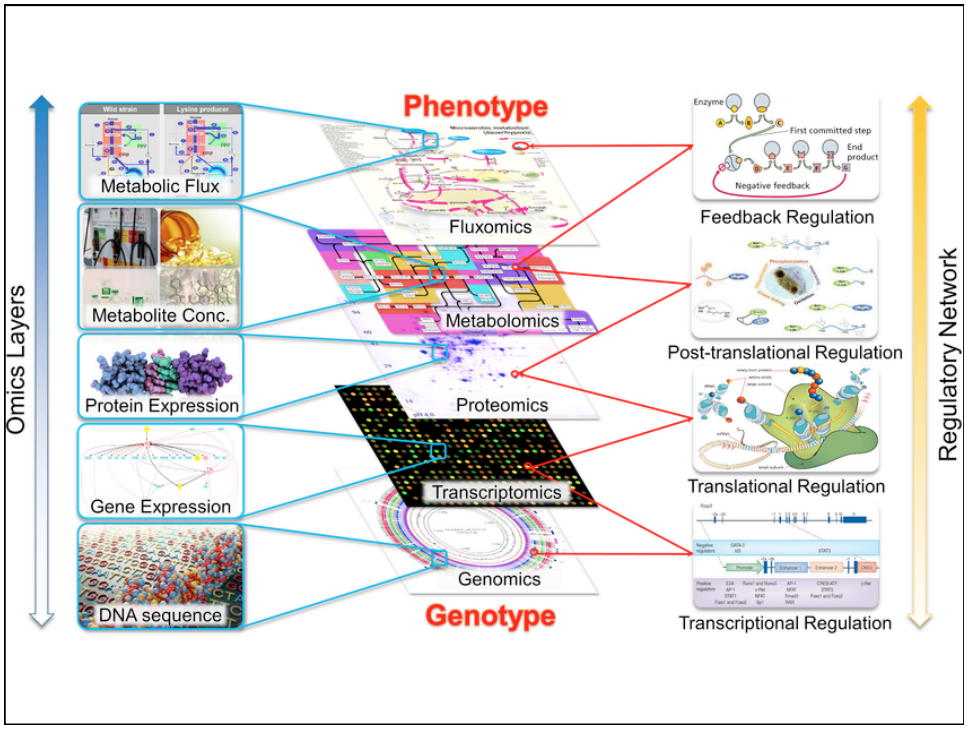
⇒ **gene variant in 3'UTR: altered mRNA stability**

⇒ **gene variant in microRNAs: change in expression**

⇒ **gene variant in a long-range enhancer: change in expression of another gene**

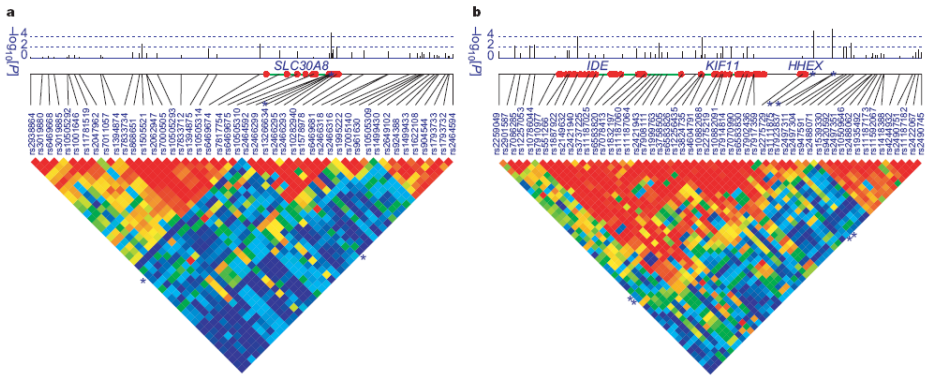
⇒ **gene variant in a CpG methylation site: change in DNA methylation pattern**

How to prove causality between a genetic variant and a biological effect?



IV-GENE CANDIDACY

Gene candidacy



Sometimes several genes are included in a same linkage disequilibrium block ⇒ how can we identify the causal gene(s)?

Sladek et al., Nature 2007

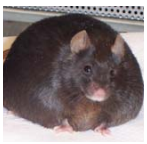
Gene candidacy

Are genes in the disease-associated LD block involved in syndromic / monogenic forms of the same disease?

-loci associated with polygenic obesity: *MC4R, BDNF, POMC, PCSK1, SIM1*

-GWAS for complex traits: 20% of the GWAS loci include genes involved in mendelian disorders for the same trait

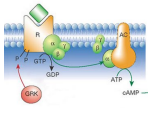
Are genes in the disease-associated LD block involved in a corresponding phenotype in animal models (KO, Tg, SiRNA)?



-loci associated with polygenic obesity: *MC4R, BDNF, POMC, PCSK1, SIM1, FTO, GIPR, NPC1, SH2B1, TBC1D1, NEGR1*

- 170 genes induce a phenotype of severe obesity in genetic mice models

Gene function, biology



-function related to energy metabolism

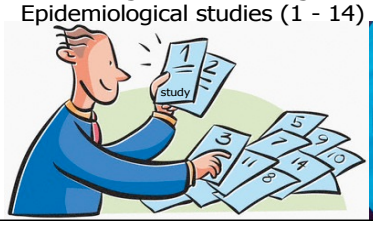
Consistency

- Repeated observation of an association in studies conducted on different populations under different circumstances
- If studies conducted by….
  - different researchers
  - at different times
  - in different settings
  - on different populations
  - using different study designs

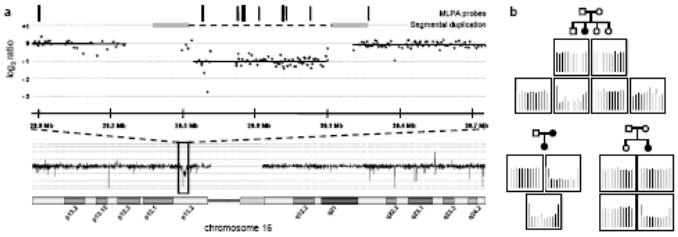
  ……all produce consistent results,

  this strengthens the argument for causation

Epidemiological studies (1 - 14)



Distinct lines of evidence adds support



⇒ a 600kb heterozygous deletion (~30 genes) on chromosome 16p11.2 explains 0.7% of morbid hyperphagic obesity and is associated with developmental delays

⇒ duplications in the same chromosomal region are associated with underweight and eating restrictive disorders

⇒ *SH2B1*, a key modulator of the response to the satiety hormone leptin, and a Mendelian hyperphagic obesity gene, is located in the deleted interval

Walters et al., *Nature* 2010; Jacquemont et al., *Nature* 2012

Model organisms

LETTER

doi:10.1038/nature13138

Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*

Scott Smemo[1]*, Juan J. Tena[2]*, Kyoung-Han Kim[3]*, Eric R. Gamazon[4], Noboru J. Sakabe[1], Carlos Gómez-Marín[2], Ivy Aneas[1], Flavia L. Credidio[1], Débora R. Sobreira[1], Nora F. Wasserman[1], Ju Hee Lee[3], Vijitha Puviindran[3], Davis Tam[3], Michael Shen[1], Joe Eun Son[5], Niki Alizadeh Vakili[3], Hoon-Ki Sung[5], Silvia Naranjo[2], Rafael D. Acemel[2], Miguel Manzanares[6], Andras Nagy[5], Nancy J. Cox[1,4], Chi-Chung Hui[3], Jose Luis Gomez-Skarmeta[2] & Marcelo A. Nóbrega[1]

The obesity-associated *FTO* intron 1 region directly interacts with the promoter of *IRX3* gene (580 Kb downstream of *FTO*)

The intron 1 SNP in *FTO* modulates *IRX3* (but not *FTO*) expression

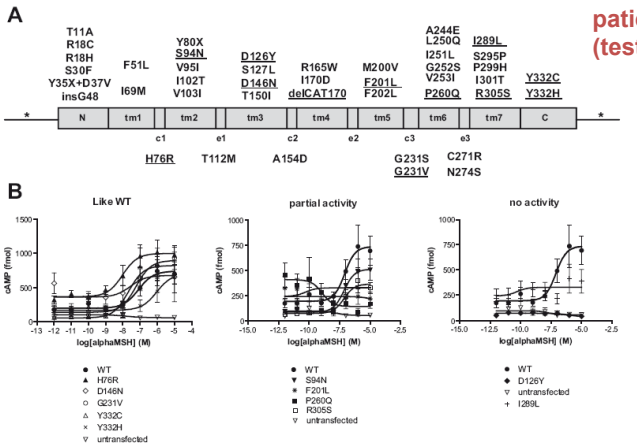
Irx3-deficient mice display a leanness phenotype

Smemo *et al.*, Nature 2014

In vitro functional studies

Prevalence of Melanocortin-4 Receptor Deficiency in Europeans and Their Age-Dependent Penetrance in Multigenerational Pedigrees

Fanny Stutzmann,[1] Karen Tan,[2] Vincent Vatin,[1] Christian Dina,[1] Béatrice Jouret,[3] Jean Tichet,[4] Beverley Balkau,[5] Natascha Potoczna,[6] Fritz Horber,[6] Stephen O'Rahilly,[2] I. Sadaf Farooqi,[2] Philippe Froguel,[1,7] and David Meyre[1]



68% of non-synonymous mutations found in obese patients are deleterious (test alpha-MSH)

Stutzmann et al., *Diabetes* 2008

V-STUDY OF ENDOPHENOTYPES

Study of endophenotypes

Common genetic variation near *MC4R* is associated with eating behaviour patterns in European populations

F Stutzmann[1], S Cauchi[1], E Durand[1], C Calvacanti-Proença[1], M Pigeyre[2], A-L Hartikainen[3], U Sovio[4], J Tichet[5], M Marre[6], J Weill[7], B Balkau[8], N Potoczna[9], J Laitinen[10], P Elliott[11,12], M-R Järvelin[4,11,12], F Horber[9], D Meyre[1] and P Froguel[1,13]

. **Rs17782313 near *MC4R* has been associated with BMI by GWAS**

. **Deleterious coding mutations in *MC4R* are the commonest form of monogenic obesity with hyperphagia and increased stature**

. **If the SNP modulates the expression / function of MC4R, we can predict associations with the same traits in an appropriate direction**

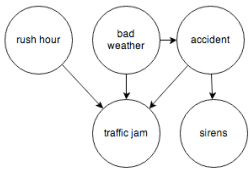
. **The SNP rs17782313 obesity predisposing allele is associated with more snacking and overeating and increased stature**

⇒ ***MC4R* is a highly relevant candidate gene at this locus**

Stutzmann et al., *Int J Obes* 2009

But can we INFER causality computationally?

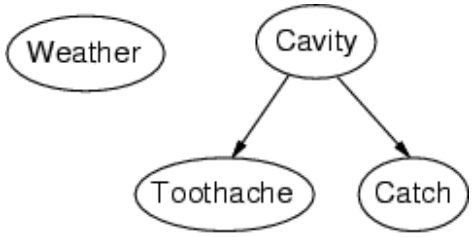
Bayesian networks



- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

    $\mathbf{P}(X_i \mid \text{Parents}(X_i))$
- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Topology of network encodes assertions:

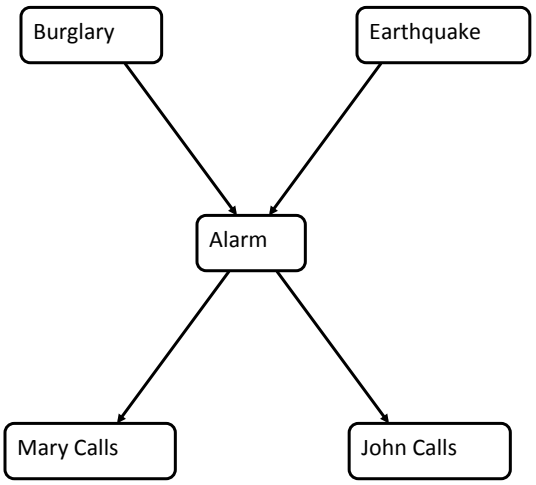


- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Example

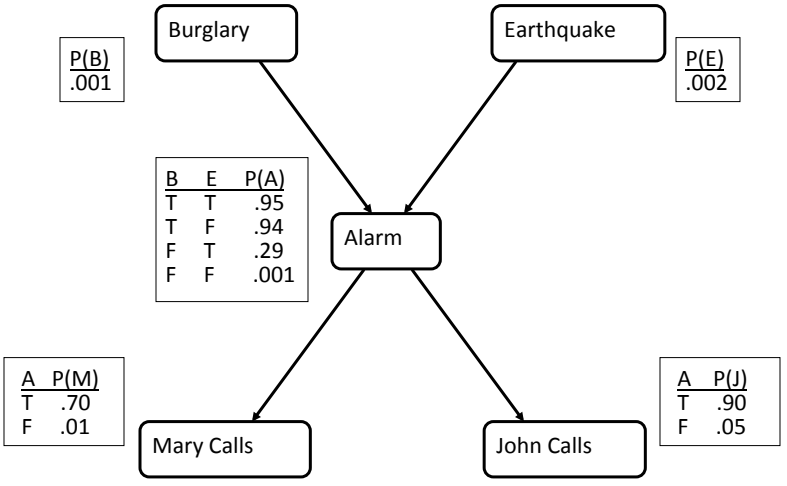
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Bayesian Networks



41

Framework for capturing conditional probabilities.



42