Unraveling the Genetic Basis of Disease: A Journey Through Bioinformatics Methods

竹内史比古
Baker Heart and Diabetes Institute
国立国際医療研究センター

2024.10.11
日本人類遺伝学会 第69回大会

https://www.fumihiko.takeuchi.name

日本人類遺伝学会第69回大会
利益相反状態の開示

発表者名：竹内史比古

発表演題に関連し、発表者らに開示すべき
利益相反状態はありません。

疾患遺伝子解析とバイオインフォマティクスについて

疾患遺伝子同定の歴史の中で、バイオインフォマティクスが果たした役割を振り返る

1.連鎖解析(1990年代)

2.トリオの全エキソーム解析(2010年代〜)

3.ゲノムワイド関連解析(2007年〜)

4.マルチオミックス(2017年〜)

5.大規模バイオバンク研究(2017年〜)

連鎖解析

- 対象疾患: メンデル遺伝性疾患
  - *HTT* ハンチントン病
  - *DMD* Duchenne型筋ジストロフィー
  - *RB1* 網膜芽細胞腫
  - *CFTR* 嚢胞性線維症
  - *BRCA1, BRCA2* 乳癌・卵巣癌
  - など多数
- 特徴: 稀なバリアント、高い浸透率
- 被験者: 大規模罹患家系
- アッセイ: DNA多型マーカー(約200個)
- バイオインフォマティクスに課されたお題
  - 疾患原因バリアントの染色体上でのおおよその位置(1cM)を見つける
  - そこから疾患原因バリアントにまで至るのは、当時は手作業

連鎖解析

- バイオインフォマティクスに課されたお題
  - 疾患原因バリアントの染色体上でのおおよその位置(1cM)を見つける
- アルゴリズム
  - DNA多型マーカー(約200個)と発症の相関を検定する [Botstein 1980]
  - Elston-Stewart
  - Lander-Green
  - EM法
- データベース
  - 多型マーカーの連鎖地図

Am J Hum Genet 32:314–331, 1980

Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms

DAVID BOTSTEIN,[1] RAYMOND L. WHITE,[2] MARK SKOLNICK,[3] AND RONALD W. DAVIS[4]

SUMMARY

We describe a new basis for the construction of a genetic linkage map of the human genome. The basic principle of the mapping scheme is to develop, by recombinant DNA techniques, random single-copy DNA probes capable of detecting DNA sequence polymorphisms, when hybridized to restriction digests of an individual's DNA. Each of these probes will define a locus. Loci can be expanded or contracted to include more or less polymorphism by further application of recombinant DNA technology. Suitably polymorphic loci can be tested for linkage relationships in human pedigrees by established methods; and loci can be arranged into linkage groups to form a true genetic map of "DNA marker loci." Pedigrees in which inherited traits are known to be segregating can then be analyzed, making possible the mapping of the gene(s) responsible for the trait with respect to the DNA marker loci, without requiring direct access to a specified gene's DNA. For inherited diseases mapped in this way, linked DNA marker loci can be used predictively for genetic counseling.

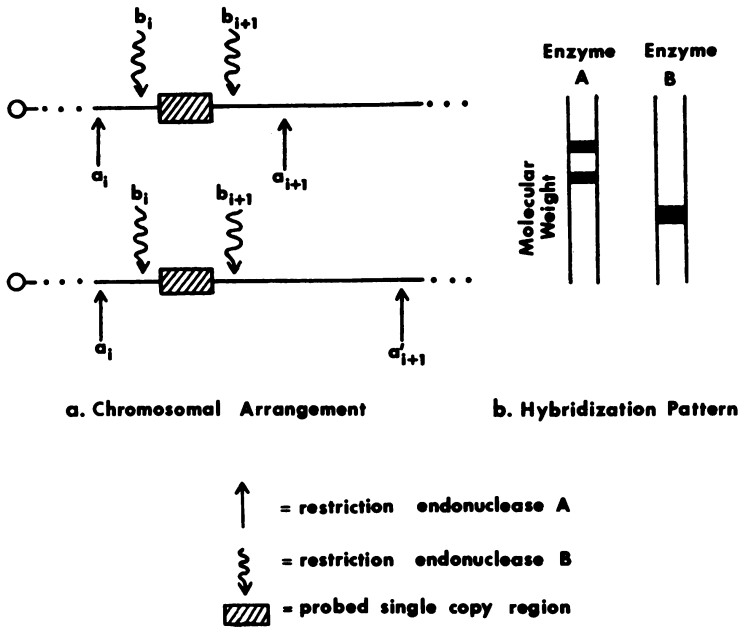


FIG. 1.—*a*, Cuts made in pair of homologous chromosomes by enzyme A and enzyme B; *b*, hybridization pattern of enzymes A and B given cuts of *a*.

A polymorphic DNA marker genetically linked to Huntington's disease

James F. Gusella*, Nancy S. Wexler†‖, P. Michael Conneally†, Susan L. Naylor§, Mary Anne Anderson*, Rudolph E. Tanzi*, Paul C. Watkins*¶, Kathleen Ottina*, Margaret R. Wallace‡, Alan Y. Sakaguchi§, Anne B. Young‖, Ira Shoulson‖, Ernesto Bonilla‖ & Joseph B. Martin*

* Neurology Department and Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA
† Hereditary Disease Foundation, 9701 Wilshire Blvd, Beverley Hills, California 90212, USA
‡ Department of Medical Genetics, Indiana University Medical Center, Indianapolis, Indiana 46223, USA
§ Department of Human Genetics, Roswell Park Memorial Institute, Buffalo, New York 14263, USA
‖ Venezuela Collaborative Huntington's Disease Project#

Family studies show that the Huntington's disease gene is linked to a polymorphic DNA marker that maps to human chromosome 4. The chromosomal localization of the Huntington's disease gene is the first step in using recombinant DNA technology to identify the primary genetic defect in this disorder.

原因遺伝子*HTT*が同定されるのは10年後

Table 2 lod scores

		Recombination fraction (θ)					
		0.0	0.05	0.1	0.2	0.3	0.4
Huntington's disease against G8	A	1.81	1.59	1.36	0.90	0.48	0.16
	V	6.72	5.96	5.16	3.46	1.71	0.33
	T	8.53	7.55	6.52	4.36	2.19	0.49
Huntington's disease against MNS		$-\infty$	−3.22	−1.70	−0.43	−0.01	0.07
Huntington's disease against GC		$-\infty$	−2.27	−1.20	−0.32	0.00	0.07
G8 against MNS		$-\infty$	−8.38	−3.97	−0.55	0.45	0.37
G8 against GC		$-\infty$	−2.73	−1.17	−0.08	0.14	0.08

A, American pedigree; V, Venezuelan pedigree; T, total.

5世代40人　7世代65人

lod score = $\log_{10}$ [連鎖している確率 ／連鎖していない確率]

1.連鎖解析(1990年代)

2.トリオの全エキソーム解析(2010年代〜)

3.ゲノムワイド関連解析(2007年〜)

4.マルチオミックス(2017年〜)

5.大規模バイオバンク研究(2017年〜)

トリオの全エキソーム解析

- 対象疾患: 連鎖解析では解明できなかったメンデル遺伝性疾患
  - *DHODH* Miller症候群
  - *KMT2D* 歌舞伎症候群
  - など多数
- 特徴
  - 突然変異でもOK
  - 臨床診断が明確でなくてもOK
- 被験者: 罹患者と両親
- アッセイ: 全エキソーム解読
- バイオインフォマティクスに課されたお題
  - 疾患原因バリアントを見つける

トリオの全エキソーム解析

- バイオインフォマティクスに課されたお題
  - 疾患原因バリアントを見つける
- アルゴリズム
  - バリアントのフィルタリングと関連解析
- データベース

内容	代表的なもの
ヒト標準ゲノム配列	
ゲノムブラウザ	UCSC, Ensembl
RNA, タンパク発現量のアトラス	GTEx, Human Protein Atlas
バリアントの機能予測	SnpEff, PolyPhen-2
バリアントの一般集団での頻度	dbSNP, gnomAD
バリアントの罹患者での頻度	DECIPHER, Matchmaker Exchange, GeneMatcher
バリアントのキュレーション	ClinVar

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng[1,10], Kati J Buckingham[2,10], Choli Lee[1], Abigail W Bigham[2], Holly K Tabor[2,3], Karin M Dent[4], Chad D Huff[5], Paul T Shannon[6], Ethylin Wang Jabs[7,8], Deborah A Nickerson[1], Jay Shendure[1] & Michael J Bamshad[1,2,9]

We demonstrate the first successful application of exome sequencing to discover the gene for a rare mendelian disorder of unknown cause, Miller syndrome (MIM%263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40× and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and eight HapMap exomes for genes with two previously unknown variants in each of the four individuals identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes underlying rare mendelian disorders and will likely transform the genetic analysis of monogenic traits.

Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)		Kindreds 1+2		Kindreds 1+2+3	
Filter	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362	3,099	1,810	2,654	1,525
Not in dbSNP129	641	102	647	114	369	53	105	25	63	21
Not in HapMap 8	898	123	923	128	506	46	117	7	38	4
Not in either	456	31	464	33	228	9	26	1*	8	1*
Predicted damaging	204	6	204	12	83	1	5	0	2	0

Each cell indicates the number of genes with nonsynonymous (NS) variants, splice acceptor and donor site mutations (SS) and coding indels (I). Filtering either by requiring the presence of NS/SS/I variants in siblings (kindred 1 (A+B)) or of multiple unrelated individuals (columns) or by excluding annotated variants (rows) identifies 26 and 8 candidate genes under a dominant model and only a single candidate gene, *DHODH*, under a recessive model (light gray cells). Exclusion of mutations predicted to be benign using PolyPhen (row 5) increases sensitivity under a dominant model but excludes *DHODH* under a recessive model because a variant in kindred 1 is predicted to be benign. A single candidate gene is identified in kindred 1 under a recessive model and excluding benign mutations (dark gray cell), but this candidate is excluded in comparisons with unrelated cases of Miller syndrome. Mutations in this candidate, *DNAH5*, were found to cause a primary ciliary dyskinesia in kindred 1. The asterisk indicates that a second gene, *CDC27*, was also identified as a candidate gene, but this is due to the presence of multiple copies of a processed pseudogene that recurrently gave rise to a false positive signal in exome analyses.

1.連鎖解析(1990年代)

2.トリオの全エキソーム解析(2010年代〜)

3.ゲノムワイド関連解析(2007年〜)

4.マルチオミックス(2017年〜)

5.大規模バイオバンク研究(2017年〜)

ゲノムワイド関連解析(GWAS)

- 対象疾患: ありふれた多因子疾患
  - 血圧 2103遺伝子座
    10.1038/s41588-024-01714-w
  - 血中脂質 1750遺伝子座
    10.1016/j.ajhg.2022.06.012
  - 2型糖尿病 1289遺伝子座
    10.1038/s41586-024-07019-6
- 特徴
  - 遺伝子と環境が関与する
  - ゲノム全体に渡って多くの疾患感受性多型がある
  - 多型と疾患形質の間がブラックボックスで残った
- 被験者: 罹患群、対照群
- アッセイ: ありふれた多型のゲノムワイドマイクロアレイ（100万個のSNP）
- バイオインフォマティクスに課されたお題
  - 疾患感受性多型を見つける

ゲノムワイド関連解析(GWAS)

- バイオインフォマティクスに課されたお題
  - 疾患感受性多型を見つける
- アルゴリズム
  - GWAS: 各SNPと疾患の相関を検定するだけ
  - GWASの有意水準と検出力 [Risch 1996]
  - GWASのソフトウェア
    - PLINK, Regenie
  - マイクロアレイに搭載されていないSNPの推定
    - IMPUTE, MACH, BEAGLE
- データベース
  - ハプロタイプ地図
    - HapMap, 千人ゲノム

The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

Has the genetic study of complex disorders reached its limits? The persistent lack of replicability of these reports of linkage between various loci and complex diseases might imply that it has. We argue below that the method that has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach (association studies) that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

		Linkage			Association			
					Singletons		Sib pairs	
Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)	Probability of transmitting disease allele A P(tr-A)	Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

血圧のGWAS

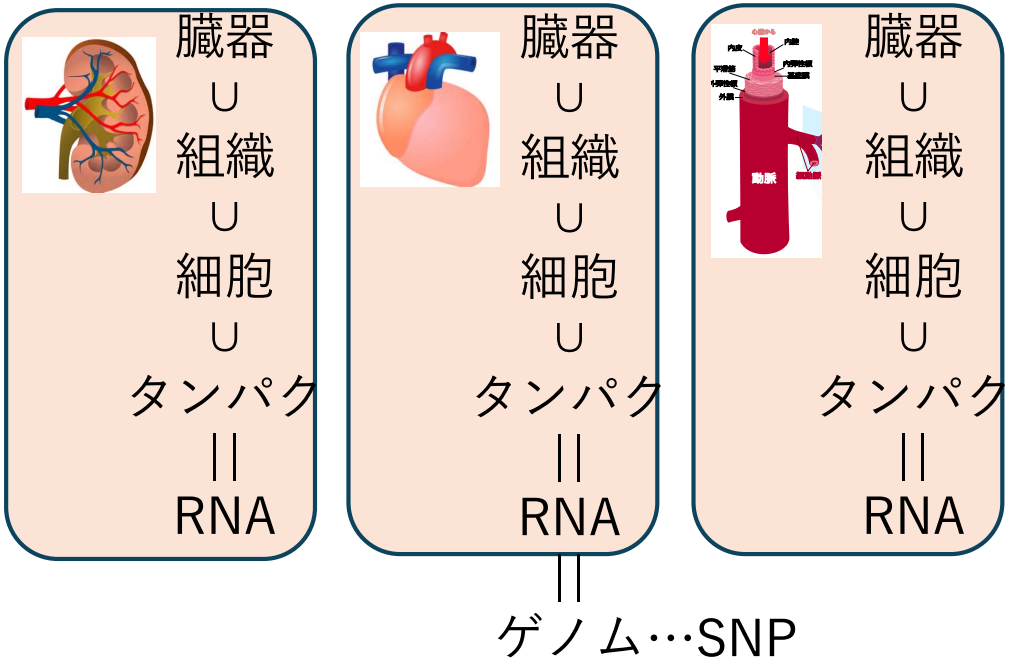
Study	Publication	年	スクリーニング症例数 [万人]				追試症例数 [万人]	ゲノムワイド有意なSNPsの数	新規SNPsの数
			欧米	東アジア	南アジア	アフリカ			
WTCCC	Nature 447:661	2007	0.5					0	0
Global BPgen	Nat Genet 41:666	2009	3				11	8	8
CHARGE	Nat Genet 41:677	2009	3				3	8	8
AGEN-BP	Nat Genet 43:531	2011		2			3	10	5
ICBP	Nature 478:103	2011	7				13	29	16
COGENT	Am J Hum Genet 93:545	2013				3	10	5	3
iGEN-BP	Nat Genet 47:1282	2015	4	3	3		22	35	12
CHD Exome+, ExomeBP, GoT2D	Nat Genet 48:1151	2016	17		3		16	51	30
CHARGE+ Exome	Nat Genet 48:1162	2016	12			2	18	70	31
Cardio-Metabochip	Nat Genet 48:1171	2016	20				14	66	17
GERA	Nat Genet 49:54	2017	9	0.7		0.3		75	39
UK Biobank	Nat Genet 49:403	2017	14				19	107	32
AGEN-BP	Nat Commun 9:5052	2018		13			16	92	19
UK Biobank	Nat Genet 50:1412	2018	76				25	535	535
MVP	Nat Genet 51:51	2019	46				47	505	261
Keaton et al.	Nat Genet 56:778	2024	103					2103	113

GWASで解明できたこと

個体 …疾患
∪
臓器
∪
組織
∪
細胞
∪
タンパク
||
RNA
||
ゲノム…SNP

ゲノムワイド関連解析

次に解明したいこと



*APOE*から
血中脂質
冠動脈疾患
アルツハイマー病

疾患A, B, …の因果関係？

どの組織・細胞種でSNPが影響？

1.連鎖解析(1990年代)

2.トリオの全エキソーム解析(2010年代〜)

3.ゲノムワイド関連解析(2007年〜)

4.マルチオミックス(2017年〜)

5.大規模バイオバンク研究(2017年〜)

マルチオミックス

- 研究対象
  - DNA多型→中間形質→疾患の解明
  - 組織・細胞種ごとの
    - RNA発現
    - タンパク発現
    - ヒストン修飾
    - DNAメチル化
    - クロマチンの開き
- 被験者: 中規模(100～1000)の一般集団ないしは罹患群
- アッセイ: ゲノムに加えて、多くの組織のトランスクリプトームなどの中間形質
- バイオインフォマティクスに課されたお題
  - 多型がどの組織で、どのような分子機序で機能しているか

マルチオミックス

- バイオインフォマティクスに課されたお題
  - 多型がどの組織で、どのような分子機序で機能しているか

- アルゴリズム
  - 多型と中間形質の相関の検定
    - eQTL　RNA発現
    - pQTL　タンパク発現
    - mQTL　DNAメチル化
  - 疾患と中間形質の相関の検定
    - TWAS　RNA発現
- データベース
  - RNA発現
    - GTEx
  - エピゲノム
    - RegulomeDB

1.連鎖解析(1990年代)

2.トリオの全エキソーム解析(2010年代〜)

3.ゲノムワイド関連解析(2007年〜)

4.マルチオミックス(2017年〜)

5.大規模バイオバンク研究(2017年〜)

大規模バイオバンク研究

- 研究対象
  - 遺伝型-表現型の包括的な関係
  - 複数の疾患形質の関係
  - 浸透率の精度向上

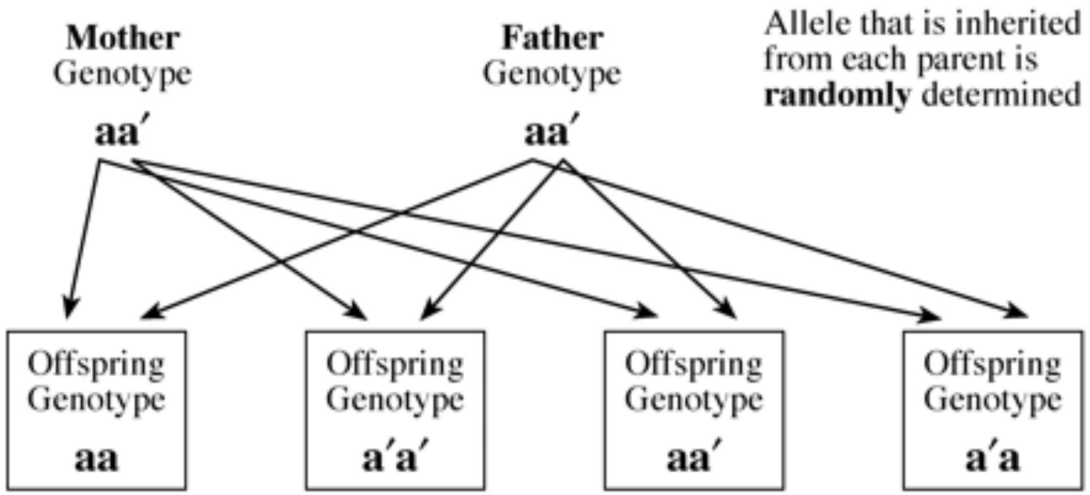
- 被験者: 大規模な一般集団の前向きコホート
  - UK Biobank, ToMMo
- アッセイ: ゲノムに加えて、多くの表現型
- バイオインフォマティクスに課されたお題
  - 多型と複数の表現型の因果関係を明らかにする

大規模バイオバンク研究

- バイオインフォマティクスに課されたお題
  - 多型と複数の表現型の因果関係を明らかにする
- アルゴリズム
  - 多型と複数の表現型の関係（多面的効果）
    - Phenome-wide association study (PheWAS)
      - HuGeAMP
  - 複数の表現型の間の因果関係
    - メンデル無作為化法
      - MR-base

メンデル無作為化法における ランダム割り付け

- 原因と仮定している形質Xのみに影響するSNPを選ぶ
- SNPの遺伝子型で、個々人をランダム割り付けできる
  - 個々人の遺伝子型は、集団でのアリル頻度のもとでランダム
  - ランダムなので、生活習慣・健康状態・環境などの「交絡因子」に影響されない

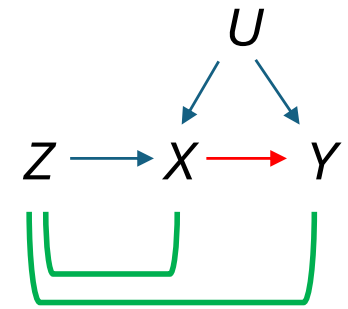


Int J Epidemiol (2003) 32:1

竹内 (2021) メンデル無作為化法 遺伝子医学 36号 83-87頁

メンデル無作為化法における因果効果の推定

	ランダム化変数 Z	検証したい原因 X	結果 Y
ランダム化比較試験	割り付け	薬の服用	疾患
メンデル無作為化法の例	LDL-C代謝遺伝子のSNP	血中LDL-C（曝露）	心筋梗塞（疾患）



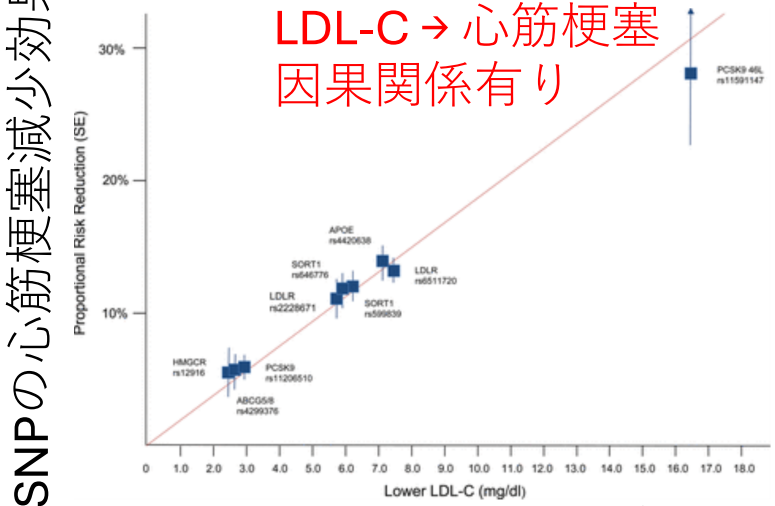
矢印は因果関係

比例している？

ZはXのみに影響

$$X\text{から}Y\text{への効果} = \frac{Z\text{から}Y\text{への効果}}{Z\text{から}X\text{への効果}} = \frac{Z\text{と}Y\text{の相関}}{Z\text{と}X\text{の相関}}$$

Zは他に影響されない



比例しているので
LDL-C → 心筋梗塞
因果関係有り

SNP Zが複数あると精度が上がる

Hum Mol Genet (2014) 23:R89

まとめ: バイオインフォマティクスが疾患遺伝子同定に果たした役割

1. 連鎖解析
   - 疾患原因バリアントの染色体上でのおおよその位置を見つける
2. トリオの全エキソーム解析
   - 疾患原因バリアントを見つける
3. ゲノムワイド関連解析
   - 疾患感受性多型を見つける
4. マルチオミックス
   - 多型がどの組織で、どのような分子機序で機能しているか
5. 大規模バイオバンク研究
   - 多型と複数の表現型の因果関係を明らかにする