「GWAS後」のための遺伝統計解析:
メタ解析と精密マッピング

竹内史比古（たけうち ふみひこ）
国立感染症研究所

2011年9月16日
@理研・CGMセミナー

GWAS(後)に役立つ統計手法

* 関連の強さを定量的に評価する
  1. ゲノムワイド関連解析（GWAS）
  2. GWASのメタ解析
  3. 低頻度多型の関連解析
* GWASで見つかった染色体領域の精密マッピング

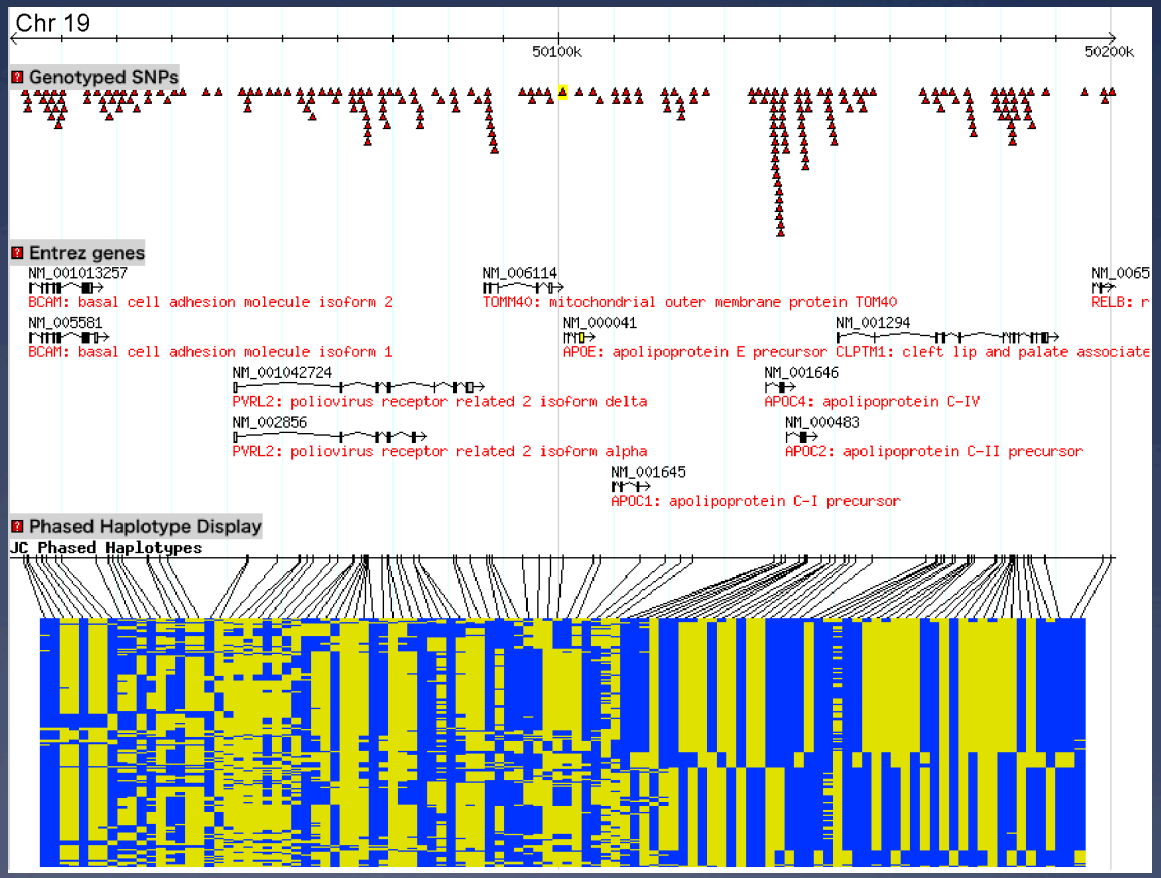
ゲノムワイド関連解析（GWAS）

- 目標: ありふれた（頻度 ≥5%）の一塩基多型（SNP）の全てについて、ありふれた疾患との関連を検定する
- 染色体上で近傍のSNPsは相関しており（連鎖不平衡）、冗長なもの（r^2 >0.8）を省いて、約10^6 SNPs をタイピングできるマイクロアレイを使う
  - ヒトゲノム配列決定、dbSNP、HapMap、アレイ技術により実現
- 10^6 回の多重検定を行うので、擬陽性を抑えるために、有意水準を $0.05/10^6 = 5\times10^{-8}$ と厳しくしないといけない
- 検出力を上げるためには、罹患者・健常者を数千人タイピングする必要がある

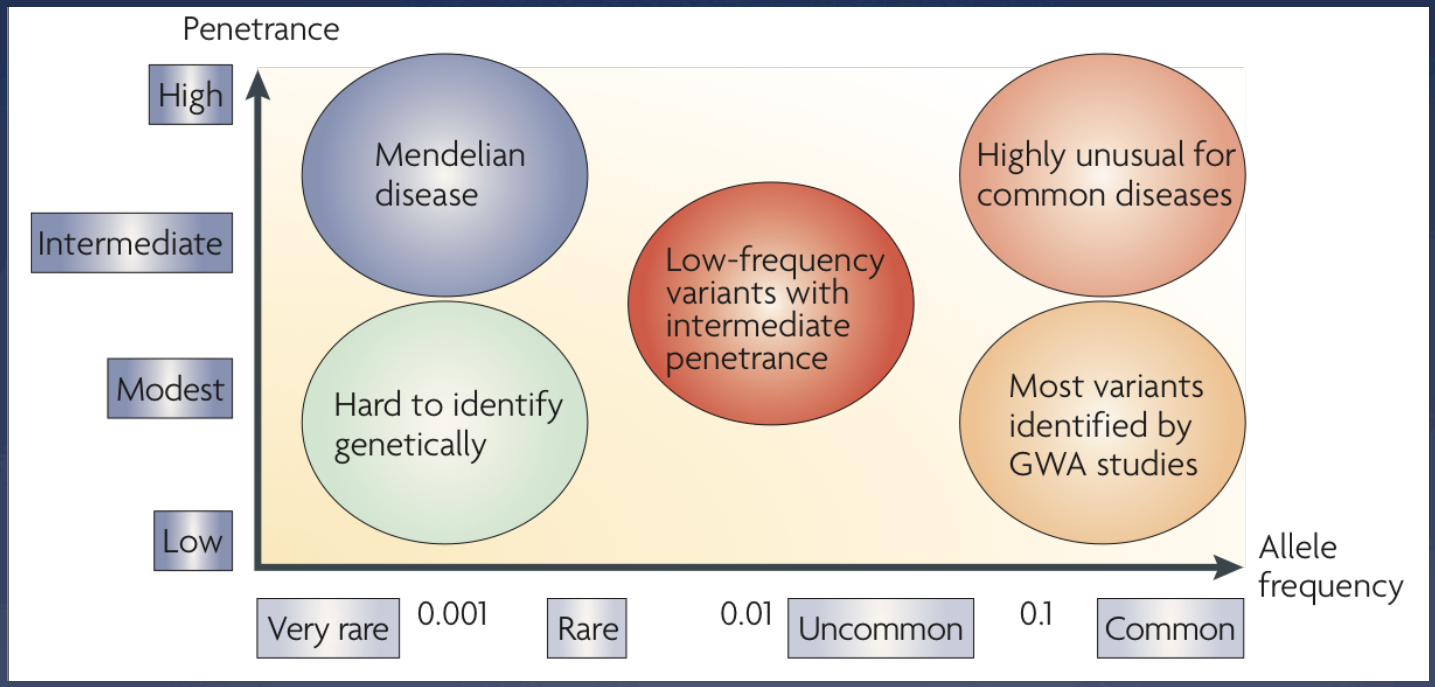
全ゲノムは不要だが、10^6 SNPsは調べる
↓
有意水準を厳しくする
↓
多数のサンプルが必要

SNPsの相関（連鎖不平衡）

- 染色体19番の200kbの領域中の108 SNPs
- 日本人45人（染色体90本）における遺伝子型



GWASで検出できる関連多型



* あまり定量的ではない

McCarthy et al. (2008) *Nat Rev Genet* 9:356

SNPの関連の検定

- *i*番目の人のSNP 遺伝子型を $x_i = 0, 1, 2$
  - 例、アリルがA/Cのとき、0 (CC), 1 (AC), 2 (AA)
- 連続形質との関連の検定
  - *i*番目の人の形質の値を y_i（例、血糖値）
  - 線形回帰
    - 誤差 $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$
    - 帰無仮説: $\beta = 0$

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- 疾患との関連の検定
  - *i*番目の人の表現型を $y_i = 1$（罹患）, 0（健常）
  - ロジスティック回帰
    - $y_i \sim \text{Bernoulli}(p_i)$
    - 帰無仮説: $\beta = 0$

$$\log \frac{p_i}{1 - p_i} = \alpha + \beta x_i$$

- 尤度を最大化する $\hat{\alpha}, \hat{\beta}$ を求める

SNPの関連の検出力

- y の分散は、x で説明される部分（S_R）と残差平方和（S_E）に分解できる

$$\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{N}\left(\hat{\alpha}+\hat{\beta}x_i-\bar{y}\right)^2+\sum_{i=1}^{N}\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)^2$$
$$= \quad S_R \quad + \quad S_E$$

- 検定に用いる統計量 $S_R/\{S_E/(N-2)\}$ は
  - 関連が無いとき（帰無仮説）は $F_{1,N-2}$ 分布に従う
  - 関連が有るとき（対立仮説）は非心度パラメータ $N R^2/\{1-R^2\}$ の $F_{1,N-2}$ 分布に従う
    - 連続形質 y の分散のうち、SNP遺伝子型 x で説明される割合を R^2 とする（決定係数）。これは相関係数の二乗。
    - N はサンプルの人数

- 有意水準 5×10^{-8} のもとで、検出力が 80% となるのは、非心度パラメータが約40のとき
  - $R^2=0.1$ なら $N=360$
  - $R^2=0.01$ なら $N=4000$（例、日本人での糖尿病に対する *KCNQ1*）
  - $R^2=0.005$ なら $N=8000$（例、同じく *CDKAL1*）
  - $R^2=0.001$ なら $N=40000$
  - ざっくり $N \fallingdotseq 40/R^2$

➔ 弱い関連を検出するには多数のサンプルが必要

R^2 とアリルの頻度・効果の関係

* アリルの頻度が p のとき

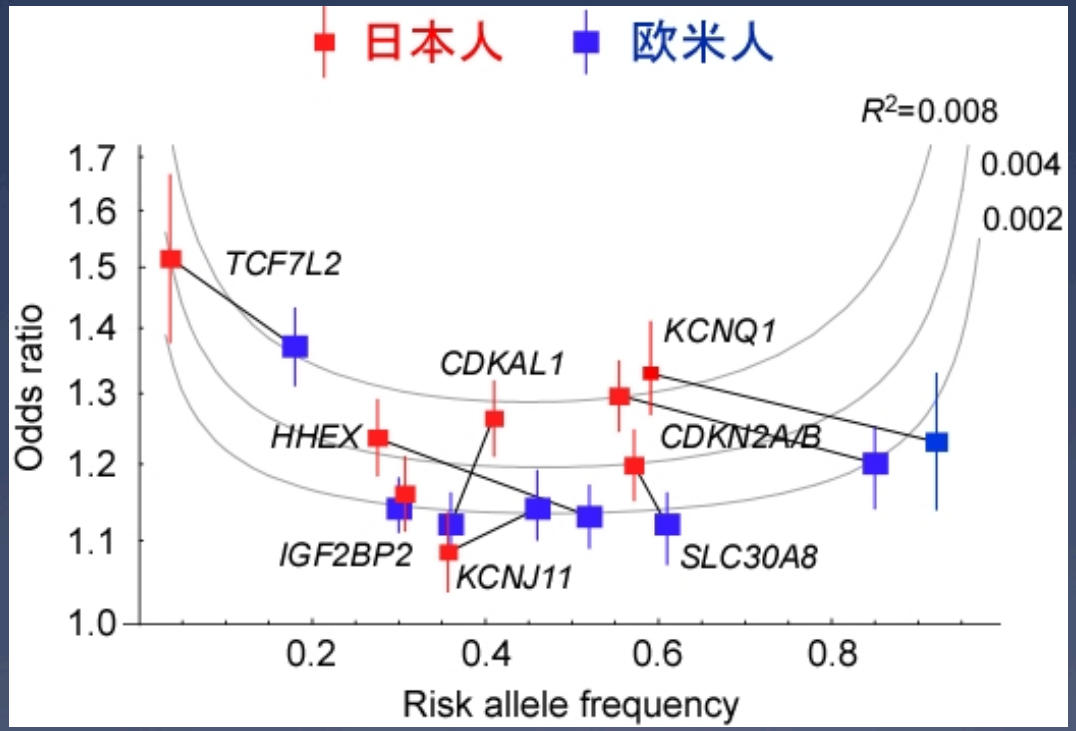
  $R^2 = 2p(1-p)\beta^2$
  * 量的形質の値 y は分散が1になるように標準化しておく

* 疾患との関連については

  $R^2 \fallingdotseq 0.5p(1-p)(\log \mathrm{OR})^2$
  * OR はオッズ比
  * $p > 0.05$, $\mathrm{OR} < 1.3$ のときに使える近似
  * R^2 は相関係数の二乗として定義
  * サンプル中の疾患群と健常群が半々と仮定

糖尿病の初期のGWAS

- ORが大きく、アリル頻度が0.5に近いSNPは関連が強い
  - 日本人では *CDKAL1, CDKN2A/B, KCNQ1*
  - 欧米人では *TCF7L2*



Takeuchi et al. (2009) *Diabetes* 58:1690

GWAS(後)に役立つ統計手法

- 関連の強さを定量的に評価する
  1. ゲノムワイド関連解析（GWAS）
  2. GWASのメタ解析
  3. 低頻度多型の関連解析
- GWASで見つかった染色体領域の精密マッピング

メタ解析による複数研究の統合

- 線形回帰
  - *i*番目の人のSNP 遺伝子型を $x_i = 0, 1, 2$
  - *i*番目の人の連続形質の値を y_i（例、血糖値）
  - 誤差 $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$
  - 連続形質に対するSNPの効果が β

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- 複数の研究で推定された効果を統合する
  - j番目の研究での効果の推定値が β_j、標準誤差が s_j
  - $1/s_j^2$ で重み付け
  - 全体での効果の推定値 β、標準誤差が s
  - 利点：個人の遺伝子情報は外に出さずに済む

$$\beta = \frac{\sum_j \frac{\beta_j}{s_j^2}}{\sum_j \frac{1}{s_j^2}}$$

$$s = \sqrt{\frac{1}{\sum_j \frac{1}{s_j^2}}}$$

メタ解析の実際

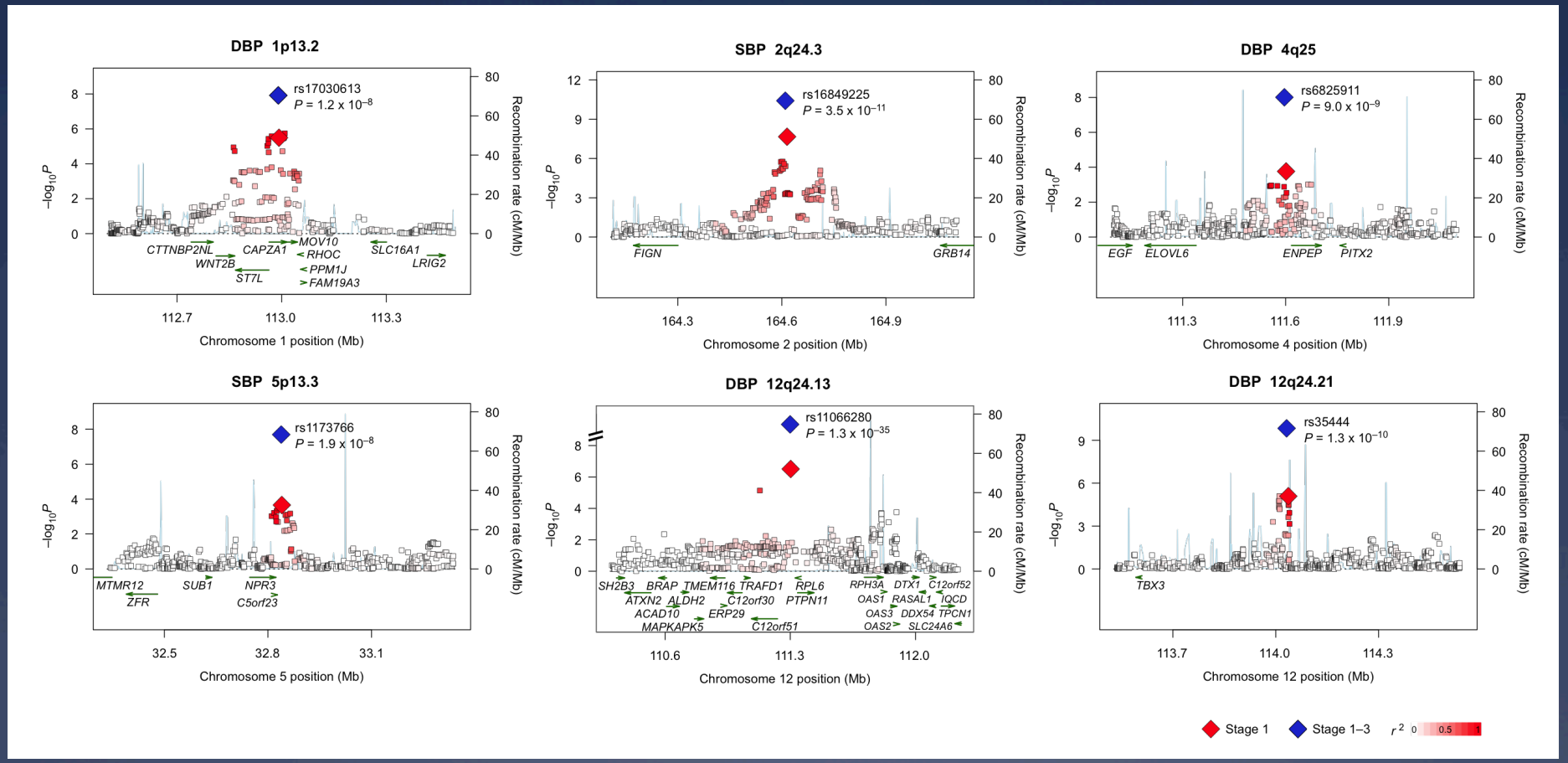
- 実際の解析はMETALなどのソフトウェアを使えば簡単
- QCが肝要
  - 遺伝子型 imputation は正確に行われているか
    - タイピングに用いたマイクロアレイに搭載されていないSNPsの遺伝子型を推測
  - 各studyが報告している効果 β は「どちらのアリル」のものか
    - “coding allele” “effect allele” を理解していない人は多い
    - Minor, major, VIC, FAM, illumina A/B
    - 間違いを含んでいるstudy。それを加えると
      - アリル頻度のばらつきが大きくなる
      - 効果のheterogeneityが大きくなる

東アジア人大規模GWASメタ解析

- 目的
  - 東アジア人で影響の強い高血圧関連遺伝子座の探索
- 方法
  - 1次スクリーニング
    - 東アジア人（日本・韓国・中国・台湾・シンガポール）ゲノム疫学コンソーシアム（AGEN）の約2万人を対象としたGWAS結果をメタ解析する。
    - ゲノムワイドに ~200万 SNPs を検定
  - 2次スクリーニング
    - 日本人1万人をタイピング
  - 追試
    - 日本人2万人をタイピング

Kato, Takeuchi et al. (2011) *Nature Genet* 6:531

血圧感受性領域の新規同定



Kato, Takeuchi et al. (2011) *Nature Genet* 6:531

GWASメタ解析で見つかった感受性多型の強さ

- R^2=0.0005～0.0031
- 検出力80%となるのは、*N*=13,000～80,000

Chr	SNP ID (pos Build 36.3)	Coded/ Other allele	Nearby Gene(s)	*N*	Coded allele freq.	SBP			DBP		
						Beta (SE), mm Hg	*P*	R^2	Beta (SE), mm Hg	*P*	R^2
1	rs17030613	C/A	*ST7L*	49,952	0.49	0.49 (0.11)	8.4E-06	0.0003	**0.38 (0.07)**	**1.2E-08**	0.0006
	(112,971,190)		*CAPZA1*								
2	rs16849225	C/T	*FIGN*	49,511	0.61	**0.75 (0.11)**	**3.5E-11**	0.0007	0.29 (0.07)	2.7E-05	0.0003
	(164,615,066)		*GRB14*								
4	rs6825911	C/T	*ENPEP*	49,515	0.51	0.60 (0.11)	7.3E-08	0.0005	**0.39 (0.07)**	**9.0E-09**	0.0006
	(111,601,087)										
5	rs1173766	C/T	*NPR3*	49,970	0.60	**0.63 (0.11)**	**1.9E-08**	0.0005	0.36 (0.07)	1.2E-07	0.0005
	(32,840,285)										
12	rs11066280	T/A	*PTPN11*	46,957	0.75	**1.56 (0.13)**	**7.9E-31**	0.0024	**1.01 (0.08)**	**1.3E-35**	0.0031
	(111,302,166)		*ALDH2*								
12	rs35444	A/G	*TBX3*	49,984	0.75	0.63 (0.13)	7.5E-07	0.0004	**0.50 (0.08)**	**1.3E-10**	0.0008
	(114,036,820)										

Kato, Takeuchi et al. (2011) *Nature Genet* 6:531

GWAS(後)に役立つ統計手法

- 関連の強さを定量的に評価する
  1. ゲノムワイド関連解析（GWAS）
  2. GWASのメタ解析
  3. 低頻度多型の関連解析
- GWASで見つかった染色体領域の精密マッピング

低頻度変異の関連解析の現状

- ありふれた疾患についての、低頻度変異(0.5～5%)関連解析
  - 全ゲノムシーケンス(~4x)はせいぜい500人まで
  - Exome シーケンスはせいぜい1000人
- 見つかる変異の数が多い
  - そのままでは、サンプルサイズが小さいのに有意水準を厳しくしないといけない！
- ヒットが報告は少ない。クローン病、脂質で報告有り

原因変異の頻度スペクトラム

- 原因変異のうち、ありふれたもの・低頻度のものはどのくらいあり、関連の強さはどのくらいか？
- 対象形質が選択圧を受けてきたか
  - 選択圧を受けている形質
    - あまりに効果の強いアリルは、頻度が高くなれない
    - ありふれた疾患の多く
  - 選択圧を受けていない形質
    - 頻度が大きくなりうる
    - 薬剤反応。ワルファリン服用量など
    - 現代の環境（高齢化・飽食）のみで問題になるもの？
- 集団サイズの歴史

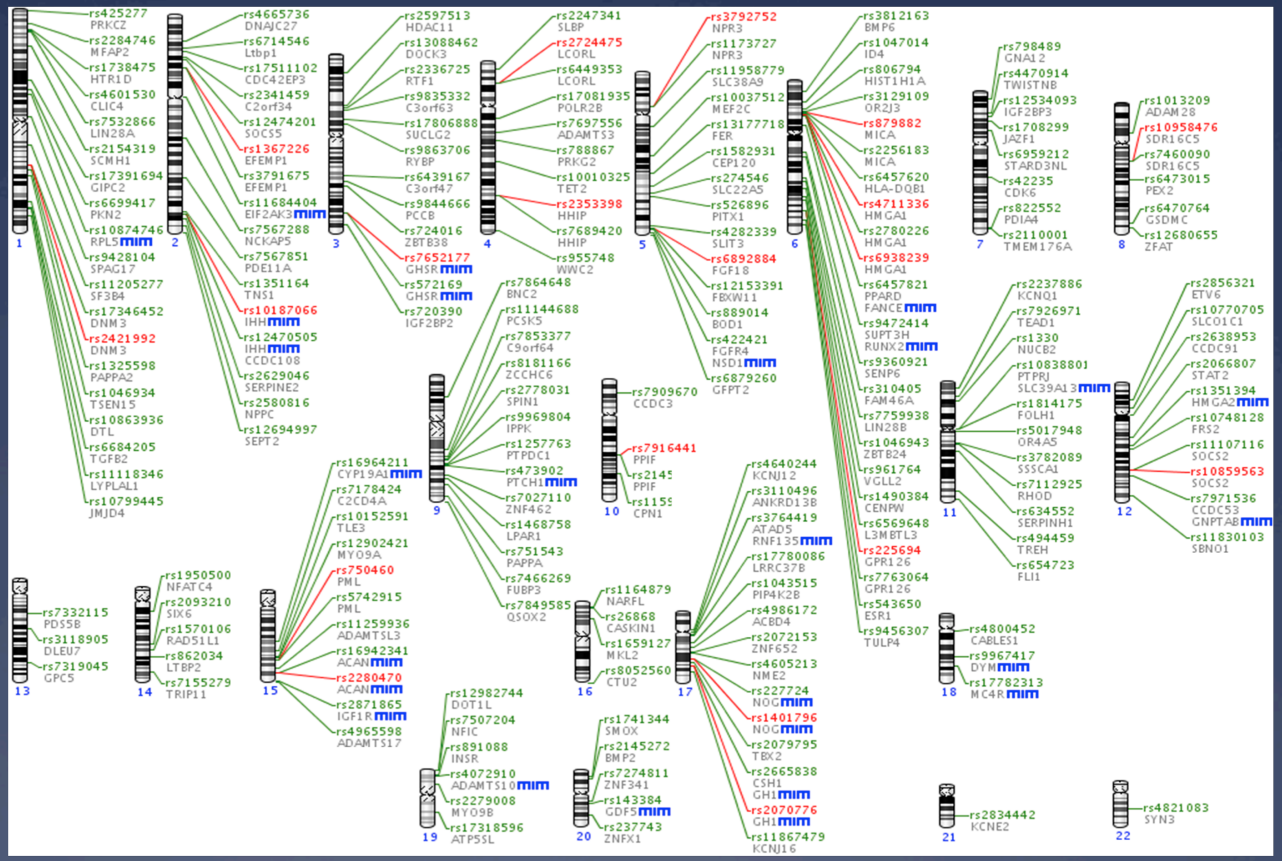
頻度スペクトラム

- 原因変異のうち、低頻度(0.5～5%)のものの関連の強さはどのくらいか
  - ありふれた変異(≥5%)から推測(外挿)する
    - 身長
    - 脂質
  - ありふれた変異については、頻度と関連の強さ(R^2)の大小は関係ない
  - ➔ 頻度に依らず、従来のGWASと同程度のサンプルサイズが必要(有意水準と検出力を揃えたとき)
- 変異の数(疾患と関連しないもの含めた全体)は頻度が下がるに従って次第に増える
- 低頻度領域で原因変異が容易に見つかることは無さそう

身長と関連する遺伝子多型

* 2010年には、欧米人183,727人のGWASメタ解析で、180の遺伝子領域が同定された

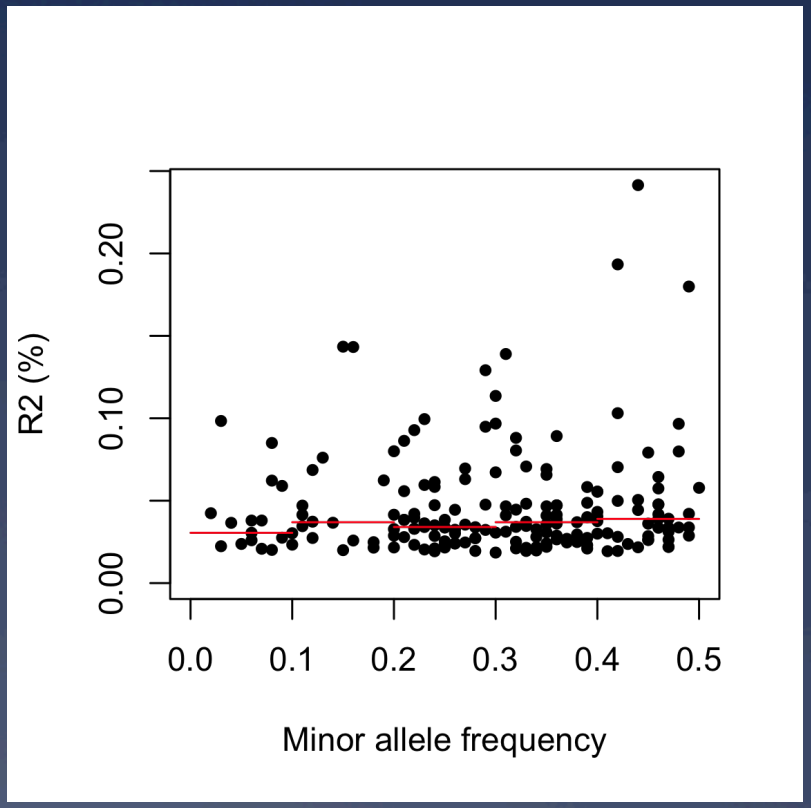
Supplementary Figure 2. 199 loci associated with adult height variation. Karyogram displaying the genome location of the 180 height SNPs identified from the primary meta-analysis (green) and the 19 secondary signals (red) discovered in the conditional analysis to be associated with height. The closest genes to the SNPs (gray) are followed by a MIM (blue) label if the gene underlies a skeletal growth-related Mendelian disorder described in OMIM. The plot was created using Affyrmation (http://genepipe.ngc.sinica.edu.tw/affyrmation/).



Allen et al. (2010) *Nature* 467:832

頻度と関連の強さ: 身長関連SNPs

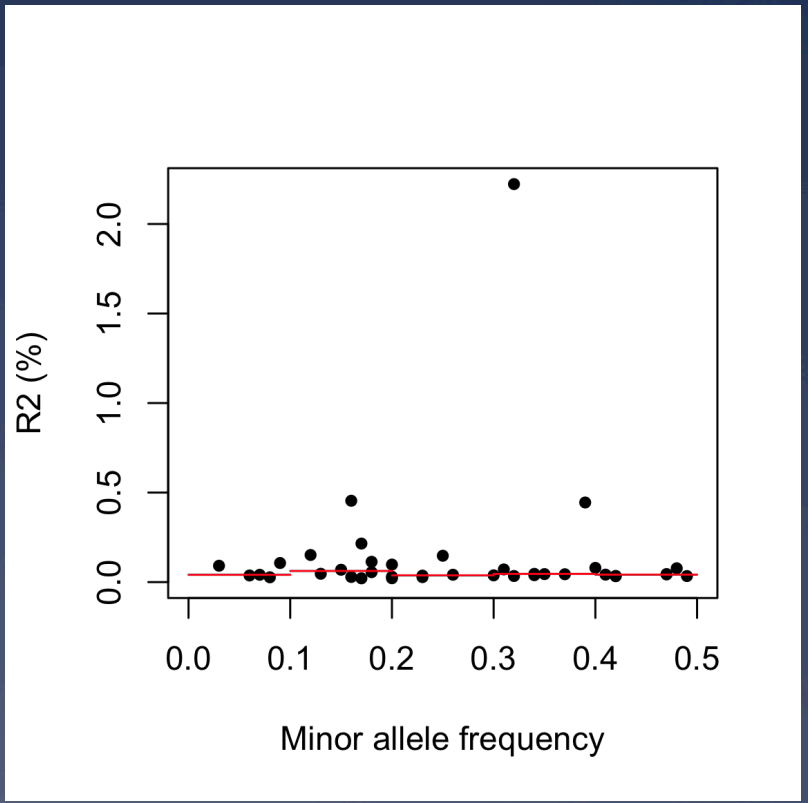
- 欧米人 183,727人の GWASメタ解析
- 身長と関連する 180 SNPs
- 赤線は、頻度10%ごとの R^2の中間値
- R^2の分布はアリル頻度に依らず一定している



Allen et al. (2010) *Nature* 467:832

頻度と関連の強さ: 脂質関連SNPs

- 欧米人 >100,000人のGWASメタ解析
- HDLとの関連が強い 38 SNPs
- 赤線は、頻度10%ごとのR^2の中間値
- R^2の分布はアリル頻度に依らず一定している



Teslovich et al. (2010) *Nature* 466:707

第1部のまとめ:

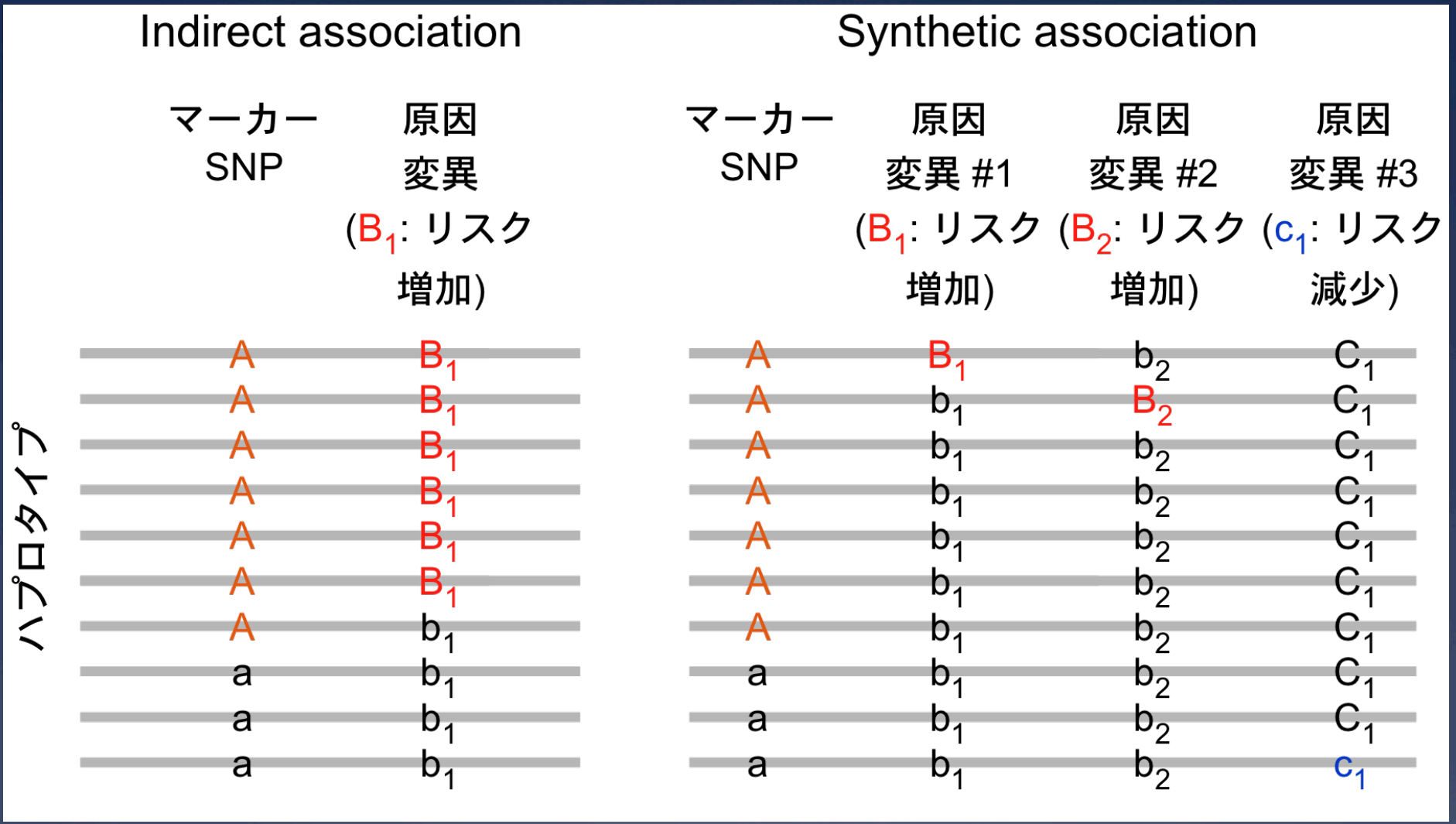
- 連続形質や疾患と関連するSNPについては、アリルの頻度と効果の強さ(β かオッズ比)から、関連の強さ R^2 を計算できる
- このパラメータに基づいて、関連解析の検出力や必要なサンプルサイズが評価できる
- GWASからGWASメタ解析へとサンプルサイズを増やすことにより、R^2 が小さい関連多型も検出できるようになった
- 低頻度多型についても、ありふれた多型と R^2 の分布は似ており、同程度のサンプルサイズが必要かもしれない

GWAS(後)に役立つ統計手法

- 関連の強さを定量的に評価する
  1. ゲノムワイド関連解析（GWAS）
  2. GWASのメタ解析
  3. 低頻度多型の関連解析
- GWASで見つかった染色体領域の精密マッピング

GWASで見つかった染色体領域の精密マッピング

- ゲノムワイド関連解析により、形質と関連するcommon（頻度≥5%）な一塩基多型（SNP）が多数見つかったが、これらはマーカーであり、関連の元となる原因変異は同定できていない
  - 関連SNPの周辺の遺伝子のどれが原因か絞り込めないことも多い
- マーカーSNPと原因変異の関係は？
  - Indirect association: マーカーSNPと頻度が同等で、強い相関を示す（連鎖不平衡係数 $r^2 \approx 1$）原因変異が1つ有る
  - Synthetic association : マーカーSNPよりも頻度が低く、相関が強くはない原因変異が（複数）有る [Dickson 他 PLoS Biol 8:e1000294]
- 原因変異の同定には、大規模な塩基配列再解読・タイピングが必要で、手間が大変！
- 遺伝統計から、少し見当をつけられないか…


Indirect association
マーカー SNP
原因変異 (B_1: リスク増加)
ハプロタイプ
A B1
A B1
A B1
A B1
A B1
A B1
A b1
a b1
a b1
a b1
Synthetic association
マーカー SNP
原因変異 #1 (B_1: リスク増加)
原因変異 #2 (B_2: リスク増加)
原因変異 #3 (c_1: リスク減少)
A B1 b2 C1
A b1 B2 C1
A b1 b2 C1
A b1 b2 C1
A b1 b2 C1
A b1 b2 C1
A b1 b2 C1
a b1 b2 C1
a b1 b2 C1
a b1 b2 c1

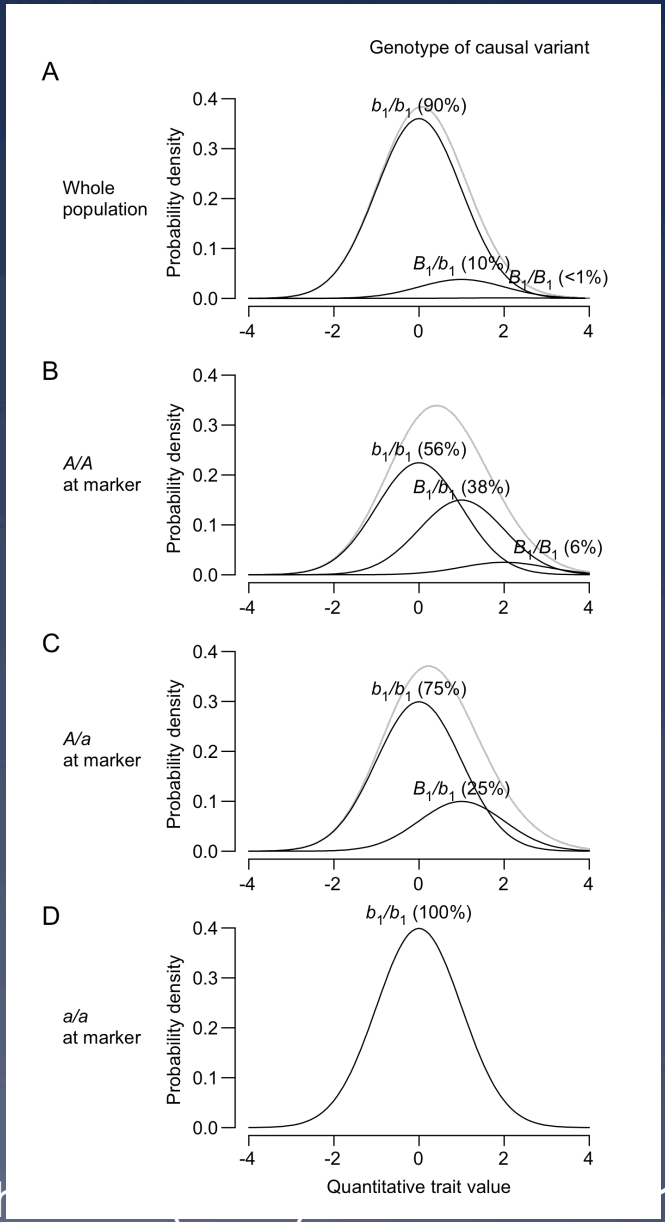
提案1: 遺伝子型間での異分散性から synthetic associationを検出する

- マーカーSNPと量的形質(QT)の関連がsyntheticかを統計的に検出する
  - マーカーSNPの3つの遺伝子型間で、QTの分散が不均一ならば、syntheticと判定する
  - 異分散性は Bartlett's test [Bartlett 1937] で検定する
  - QT全体の分布は予め rank-based inverse normal transformation [Blom 1958] で正規化しておく
  - 原因変異が未知でも検出できる！

Takeuchi et al. (2011) *Genome Res* 21:1122

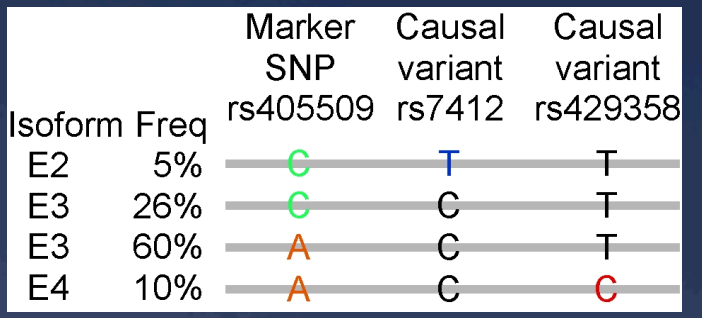
低頻度原因変異に因る異分散性：モデルケース

- SNPs and haplotypes
  - marker SNP with alleles *A* and *a*
  - causal variant with alleles B_1 and b_1
  - allele B_1 (5% in frequency) is always linked to allele *A* (20% in frequency); thus existing haplotype classes are AB_1, Ab_1 and ab_1
- QT distribution
  - Normally distributed with the unit variance and the mean equal to 2, 1 and 0 within a subgroup of individuals having genotype B_1/B_1, B_1/b_1 and b_1/b_1, respectively
  - **A**: In the whole population, a mixture of the normal distributions combined according to the frequency of genotypes B_1/B_1, B_1/b_1 and b_1/b_1
  - **B**: Individuals with *A*/*A* genotype at the marker SNP are enriched with the genotypes of B_1/B_1 and B_1/b_1 at the causal variant, which are minor in the whole population. 分散が大きい
  - **C**: Individuals with *A*/*a* genotype
  - **D**: All individuals with *a*/*a* genotype have b_1/b_1 genotype at the causal variant. 分散 = 1



低頻度原因変異に因る異分散性：*APOE*遺伝子とLDLコレステロール

- *APOE* 遺伝子は LDL コレステロール (LDL-C) 量と関連する
  - 3つの isoform が2つの原因変異 (MAF <10%) でコードされている
  - E3 (一般的) に対して、E2 (rs7412, Arg158Cys) では LDL-C 減少、E4 (rs429358, Cys112Arg) では LDL-C 増加



- マーカー SNP rs405509 (MAF >30%) では異分散性が有意（4990人で検定）

Testing heteroscedasticity of SNPs in the *APOE* locus associated with LDL-C.

SNP	Genotype	Number of individuals	Distribution of LDL-C level		Association with LDL-C level			Heteroscedasticity
			Mean	Variance	Beta	p-value	R^2	p-value
rs405509	C/C	462	-0.153	1.182	-0.117	1.0E-07	0.006	0.019
(GWAS SNP)	C/A	2035	-0.050	0.976				
	A/A	2343	0.073	0.978				
rs377702	T/T	32	-0.487	1.231	-0.191	5.1E-07	0.005	0.583
(GWAS SNP)	T/C	677	-0.149	1.025				
	C/C	4131	0.028	0.991				
rs7412	T/T	12	-1.302	1.079	-0.651	2.0E-44	0.040	0.92
(causal variant)	T/C	452	-0.584	0.981				
	C/C	4376	0.064	0.960				
rs429358	T/T	3954	-0.042	0.987	-0.212	1.4E-09	0.008	0.73
(causal variant)	T/C	850	0.185	1.023				
	C/C	36	0.214	1.104				

We first adjusted LDL-C level for body mass index and categories by sex and age (≤40, 41–50, 51–60, ≥61 years), and then applied rank-based inverse normal transformation. Individuals under lipid treatment were excluded. Data is shown for 4840 individuals with complete observation from the Amagasaki study in (Takeuchi et al. 2010).

21:1122

シミュレーションによる検出力の評価

- ゲノムワイド関連解析で同定されるマーカーSNPを仮定する（寄与率 R^2=0.00592; 有意水準 $5x10^{-8}$のもと、5000人で関連を検定したときの検出力が0.5）
- Synthetic association について4つのモデルを仮定（次ページ）
- シミュレーション（1000回）
  - マーカーSNPと原因変異の遺伝子型とQTをランダムに生成（5000人）
  - Synthetic associationを検定して、検出力を評価

Takeuchi et al. (2011) *Genome Res* 21:1122

Synthetic association のモデル

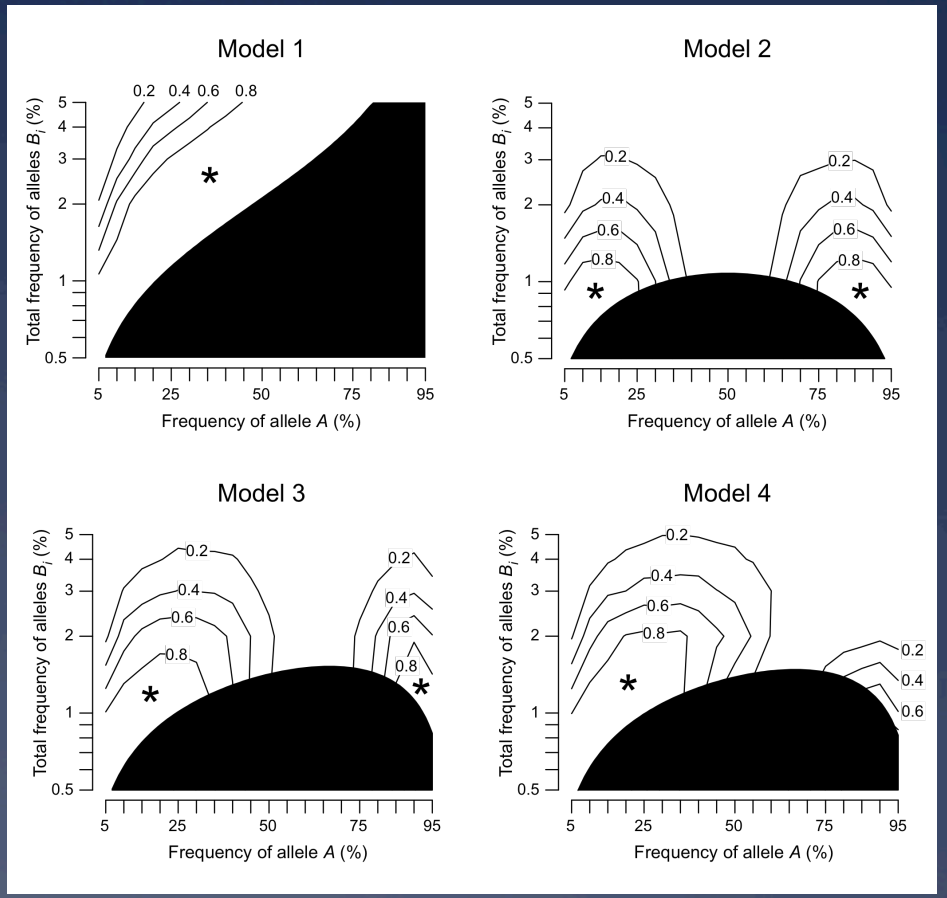
- SNPs and alleles
  - Marker SNP has alleles *A* and *a*
  - *l* causal variants each have alleles B_1 and b_1, B_2 and b_2, up to B_l and b_l, where the *causal* (low-frequency) allele B_j is linked to *marker* allele *A*
  - *m* other causal variants each have alleles C_1 and c_1, C_2 and c_2, up to C_m and c_m, where the *causal* (low-frequency) allele c_j is linked to *marker* allele *a*
- Model 1 (マーカーの片方のアリルに、全ての低頻度変異アリルが連鎖している)
  - All *causal* alleles linked to *marker* allele *A* have identical effect-size
  - No *causal* allele linked to allele *a*
- Model 2 (マーカーの両アリルに、低頻度変異アリルが均等に連鎖している)
  - Effect-size is uniform
  - Cumulative frequencies equal between *causal* alleles B_j and *causal* alleles c_j
- Model 3
  - Effect-size of *causal* alleles is uniform
  - Cumulative frequency of the *causal* alleles B_j is twice the cumulative frequency of *causal* alleles c_j
- Model 4
  - Cumulative frequencies are equal between *causal* alleles linked to the two *marker* alleles
  - Effect-size of *causal* alleles B_j is twice the effect-size of *causal* alleles c_j

Takeuchi et al. (2011) *Genome Res* 21:1122

結果1:遺伝子型間での異分散性から synthetic associationを検出する

検出力をシミュレーションで評価した

* Model 1
  * マーカーの片方のアリルに、全ての低頻度変異アリルが連鎖している場合
  * （マーカーアリルの頻度）≥ 45% なら検出可能
  * （マーカーアリルの頻度）= 25% で、（低頻度変異アリルの合計頻度）< 3% なら検出可能
* Model 2
  * マーカーの両アリルに、低頻度変異アリルが均等に連鎖している場合
  * マーカーアリルの頻度～50%では検出できない！



•星（＊）の領域で synthetic association が検出可能（検出力 >0.8）
•黒塗りの領域は原因変異が存在し得ないので無視

Takeuchi et al. (2011) *Genome Res* 21:1122

提案2: 異分散性と歪度から synthetic associationを検出する

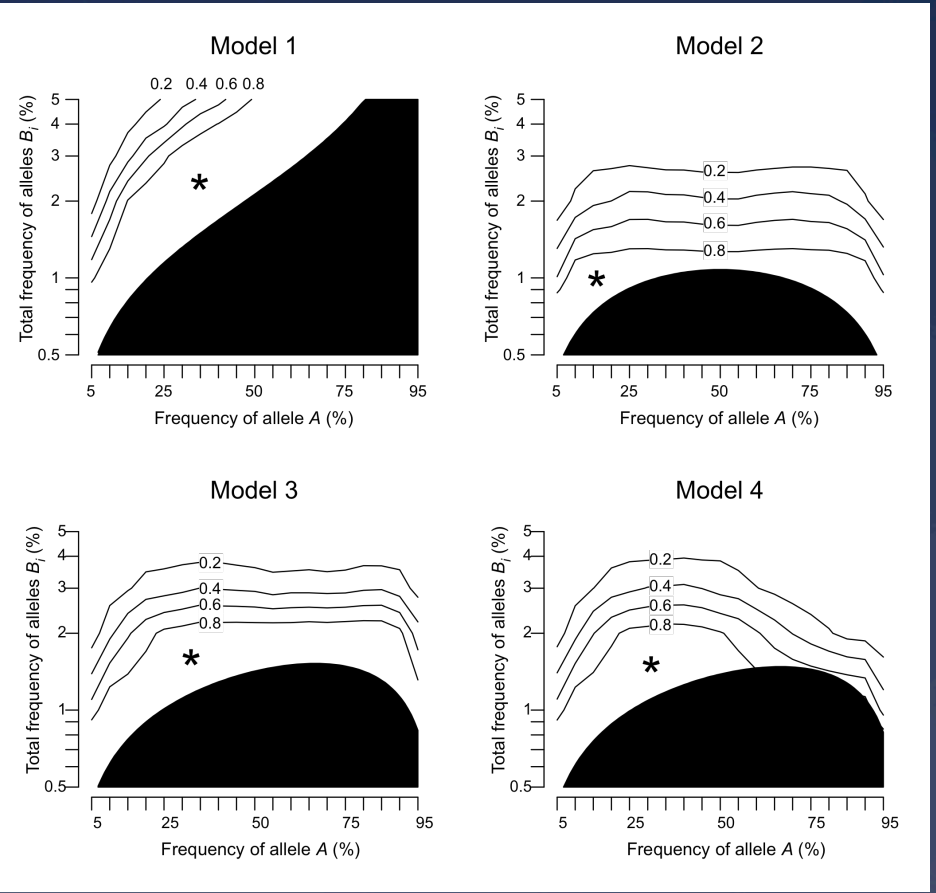
- Model 2（マーカーの両アリルに、低頻度変異アリルが均等に連鎖している）については、
  - 異分散性が生じないが
  - 各遺伝子型でのQT分布は歪んでいる
- 各遺伝子型でのQT分布の歪度もsynthetic associationの検出指標になる
- 提案2：異分散性と歪度を（Fisherの方法で）組合せてsynthetic associationを検定する

Takeuchi et al. (2011) *Genome Res* 21:1122

結果2:異分散性と歪度から synthetic associationを検出する

検出力をシミュレーションで評価した

- Models 1, 3, 4
  - （マーカーアリルの一方に連鎖する変異アリルの合計頻度）≤ 2% なら検出可能
- Model 2
  - マーカーの両アリルに、低頻度変異アリルが均等に連鎖している場合
  - （マーカーアリルの一方に連鎖する変異アリルの合計頻度）≤ 1% なら検出可能
- いずれの場合も
  - （変異アリルの合計頻度）≤ 3% なら検出可能



- 星（＊）の領域で synthetic association が検出可能（検出力 >0.8）
- 黒塗りの領域は原因変異が存在し得ないので無視

Takeuchi et al. (2011) *Genome Res* 21:1122

疾患との関連への応用

- Case-control 解析でも、似たような考えが使える
- 疾患関連マーカーのアリルを A/a とする
- サンプルを遺伝子型 AA, Aa, aa に層別し、各群について近傍の SNPs の関連を調べる
- synthetic association により、アリル A の一部に原因変異が乗っている場合
  - ➔ AA, Aa の群では近傍の SNPs が関連を示すが、aa では示さない
- Indirect association の場合
  - ➔ AA, Aa, aa いずれの群でも、近傍の SNPs は関連を示さない

Takeuchi et al. (2011) *Genome Res* 21:1122

第2部のまとめ

- ゲノムワイド関連解析で見つかったSNPの関連が低頻度変異に由来する(synthetic)かは、QT分布の遺伝子型間の異分散性と歪度により統計的に検出できる
- 低頻度原因変異の合計頻度 < 3% の場合、検出力 >80%(有意水準5%、5000人で検定)
- synthetic な関連が検出されたときは、低頻度多型の探索により、原因変異の同定が期待できる
  ➔ 精密マッピングの方針決定に役立つ

Takeuchi et al. (2011) *Genome Res* 21:1122

今日のまとめ

1. 連続形質や疾患と関連するSNPについて、関連の強さを R^2 で定量できる
   * このパラメータに基づいて、関連解析の検出力や必要なサンプルサイズが評価できる
   * GWASにおいては、10^6 SNPs を多重検定することから、真の関連を見つけ出すには多数のサンプルが必要になる
   * サンプルサイズを大きくするためには、多数のGWASのメタ解析が有効である

2. GWASあるいはそのメタ解析で同定される関連多型はあくまでもマーカーであり、原因多型を同定するための精密マッピングが必要
   * そのための新たな統計手法が必要。その一つとして、遺伝子型ごとの表現型の分散を調べる手法が有効である