

ゲノムワイド関連解析 (GWAS) による高血圧遺伝子の解明

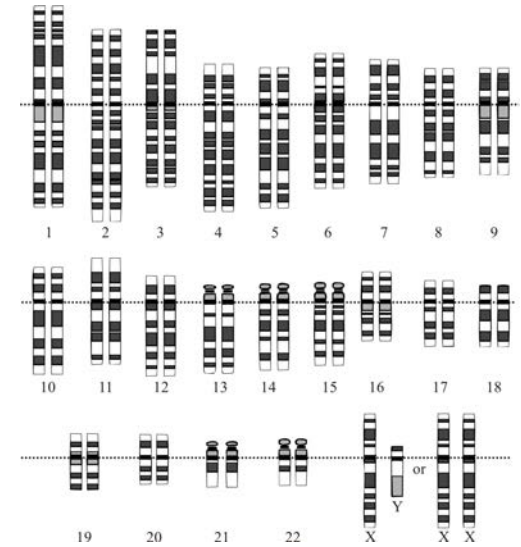
竹内史比古
国立国際医療研究センター研究所

2016年統計関連学会連合大会
2016.09.06 @金沢大

パワーポイント:
<http://fumihiko.takeuchi.name>

- 
1. **イントロ**
 2. 統計モデル
 3. 解析結果
 4. 細かい点をいくつか

DNA多型とは？



- 各人のヒトゲノム

- 染色体

- 23本(父由来)+ 23本(母由来)

- DNA

- A, C, G, Tの並び
- 3×10^9 塩基対(父由来)
+ 3×10^9 塩基対(母由来)

- DNA多型

- DNAで個人差がある箇所
- 日本人集団中での頻度 $\geq 1\%$
- 6×10^6 箇所

- 太郎

- ...ACT **G**AA GTG... (父由来)
- ...ACT **G**AA GTG... (母由来)

- 花子

- ...ACT **G**AA GTG... (父由来)
- ...ACT **A**AA GTG... (母由来)

- 大輔

- ...ACT **G**AA GTG... (父由来)
- ...ACT **A**AA GTG... (母由来)

G/AがDNA多型

Aの頻度が $2/6=33\%$

DNA多型と体質

- アルデヒド脱水素酵素2遺伝子
- DNA多型
 - c.1510G>A (p.Glu504Lys) rs671
 - …ACT **G**AA GTG…
 - ↓
 - …ACT **A**AA GTG…
- 遺伝型**GG**の人
 - お酒飲める
- 遺伝型**AG**の人
 - 酵素活性が1/16
 - お酒を飲むと赤くなる
- 遺伝型**AA**の人
 - 酵素活性ない
 - お酒が飲めない



エタノール



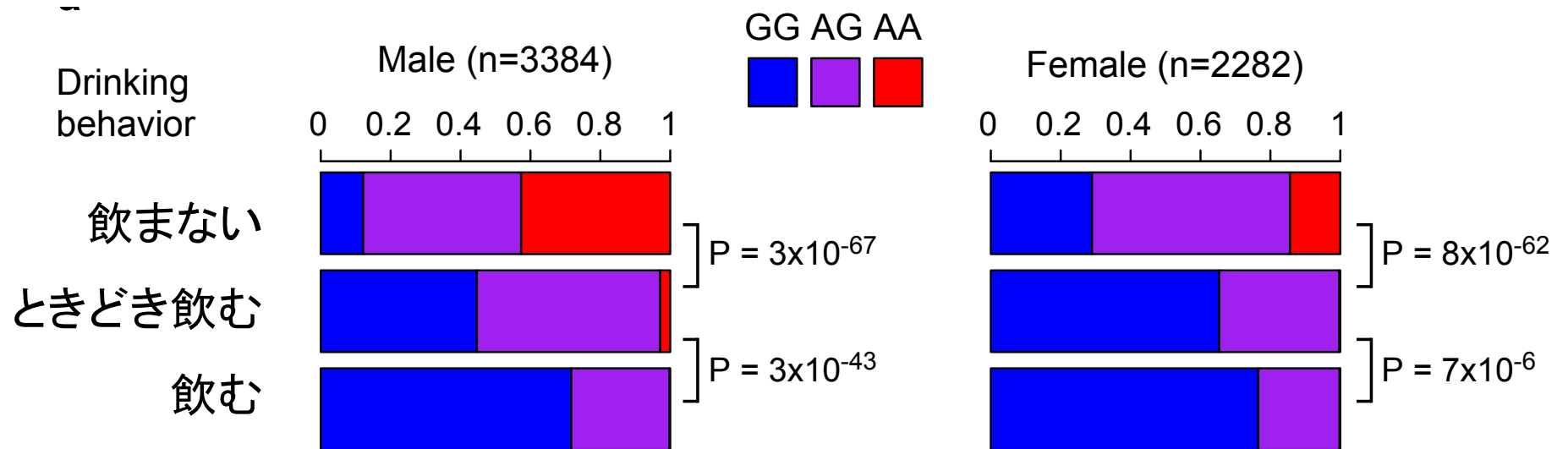
アセトアルデヒド→毒性

↓アルデヒド脱水素酵素2

酢酸

DNA多型と飲酒行動

遺伝型(アルデヒド脱水素酵素2 rs671)



遺伝型と飲酒行動が明確に関連している。

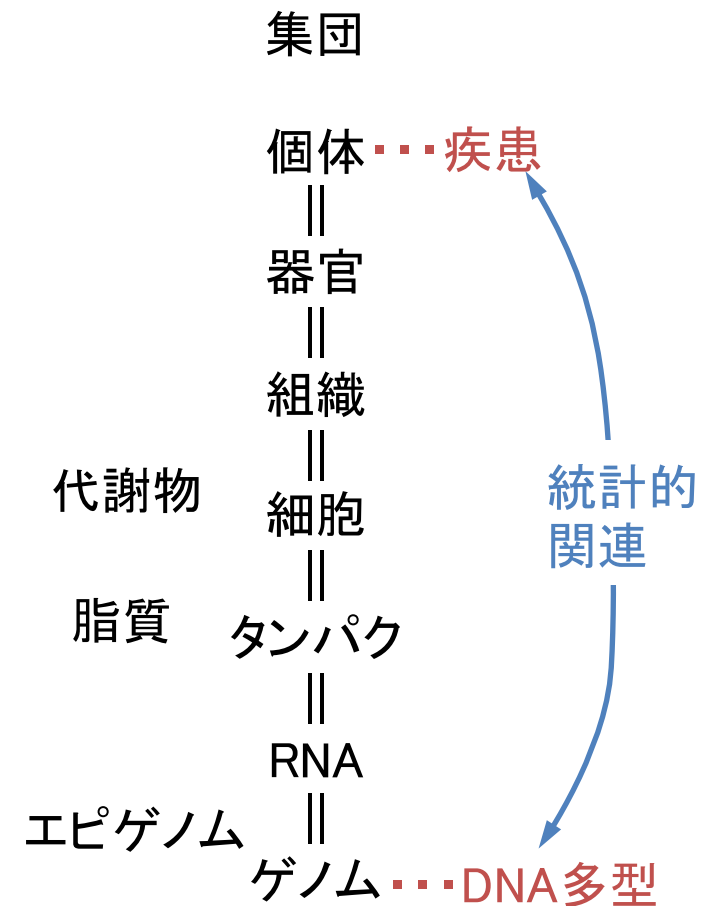
DNA多型と病気

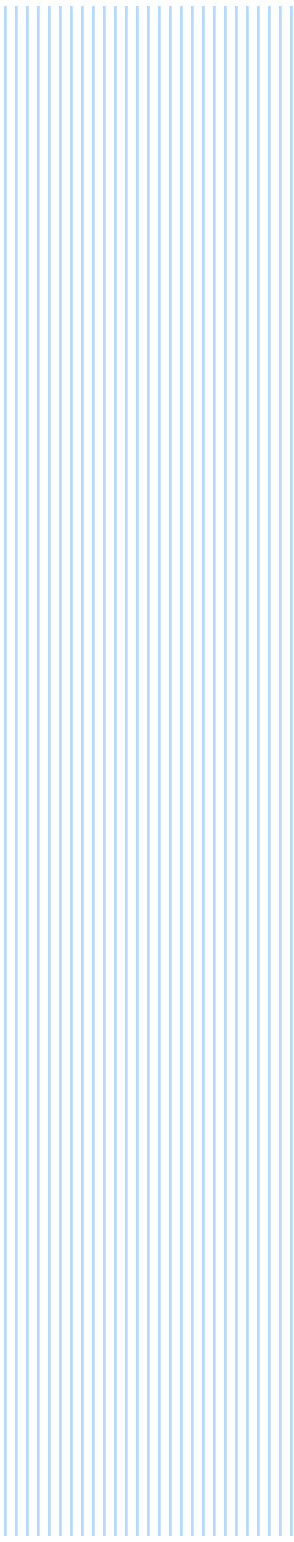
- 疾患感受性遺伝子とは
 - DNA多型により、病気の罹り易さ(感受性)が変わる遺伝子
- 疾患感受性遺伝子を見つける意義
 - 病気の仕組みの解明
 - 創薬のターゲットになる
 - 個人の発症予測・至適治療法の選択(個別化医療)
- 疾患感受性遺伝子がそもそも存在するか？
 - 疾患感受性の素因は、遺伝と環境(食事など)
 - 家族集積性から遺伝が占める割合(遺伝率)が分かる
 - 糖尿病 0.5
 - 身長 0.8
- こういうのをごっそり見つけよう→ゲノムワイド関連解析(GWAS)

DNA多型と疾患の関連解析

- DNA多型と疾患

- 生体階層構造の両端に離れている
- 関連をゲノムワイドに検定するのが、ゲノムワイド関連解析 (GWAS)
- 統計的関連が、ヒトでの因果関係を示唆する
- 中間は、ブラックボックスとしてよい
- 遺伝統計学は疾患解明・治療法開発の強力な手段の一つ



- 
1. イントロ
 2. **統計モデル**
 3. 解析結果
 4. 細かい点をいくつか

ゲノムワイド関連解析 (GWAS)

- 目標
 - ゲノムワイドに、DNA多型の全てについて疾患との関連を検定する
- DNA多型は 6×10^6 個あるが、染色体上で近傍のものは相関している(連鎖不平衡)ので、独立なもののは正味 10^6 個
- 約 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 $0.05/10^6 = 5 \times 10^{-8}$ 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・健常者を数千調べる必要がある

ゲノムワイドに網羅的に調べる



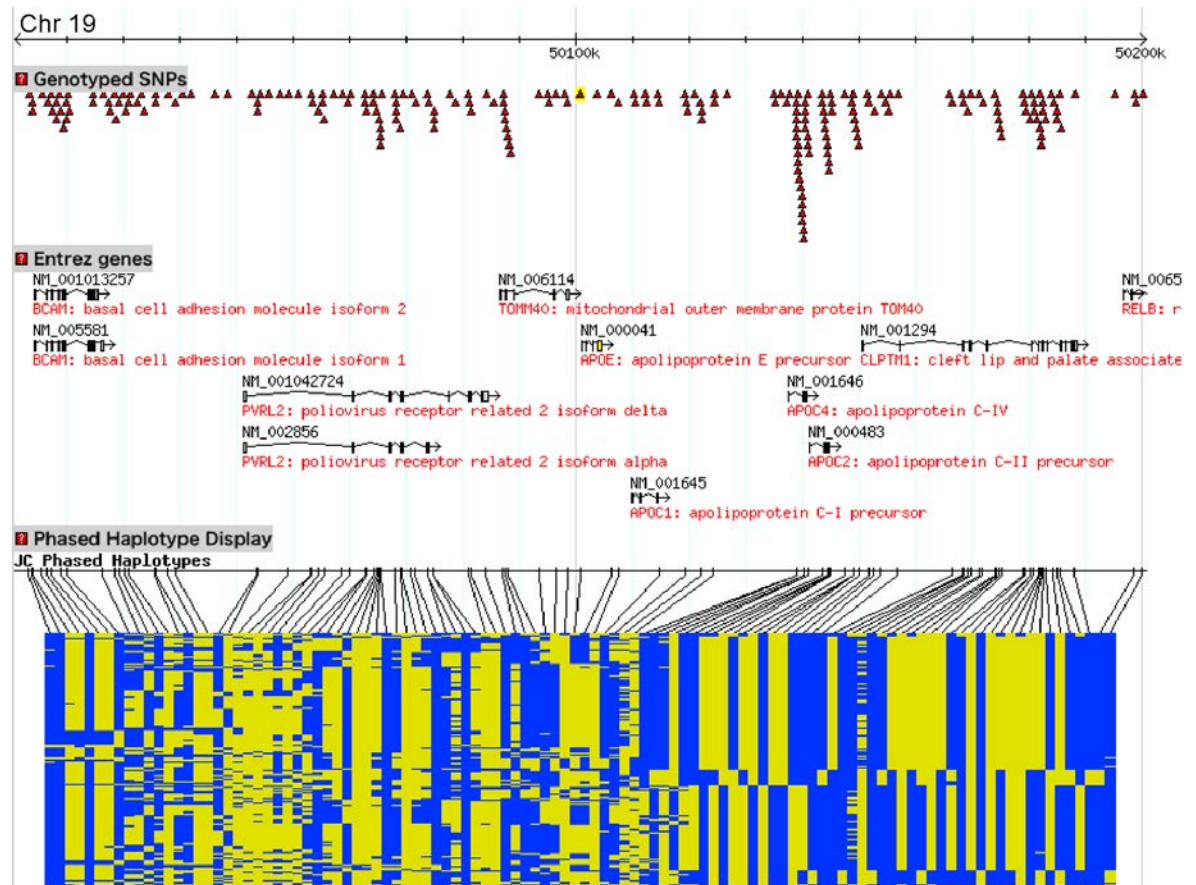
有意水準を厳しくする



多数のサンプルが必要

DNA多型同士の相関(連鎖不平衡)

- 染色体19番の 200×10^3 塩基対の領域中の108 DNA多型
- 日本人45人(染色体90本)



DNA多型と疾患の関連の検定

- i 番目の人の遺伝子型を $x_i = 0, 1, 2$
 - 例、DNA多型がG/Aのとき、0 (GG), 1 (AG), 2 (AA)
- 連続形質(血圧など)との関連の検定
 - i 番目の人の形質の値を y_i
 - 線形回帰
 - 誤差 $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$
 - 帰無仮説: $\beta = 0$
- 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: $\beta = 0$
- 尤度を最大化する $\hat{\alpha}, \hat{\beta}$ を求める

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

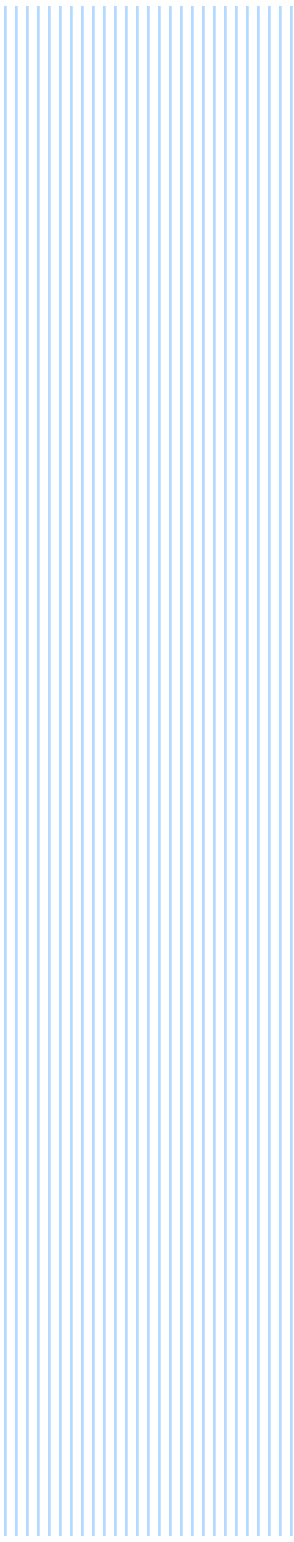
$$\log \frac{p_i}{1 - p_i} = \alpha + \beta x_i$$

関連検定の検出力

- 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

$$\begin{aligned}\sum_{i=1}^N (y_i - \bar{y})^2 &= \sum_{i=1}^N (\hat{\alpha} + \hat{\beta}x_i - \bar{y})^2 + \sum_{i=1}^N (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \\ &= S_R + S_E\end{aligned}$$

- 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、DNA多型 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 $N \doteq 40/R^2$
- ➔ 弱い関連を検出するには多数のサンプルが必要

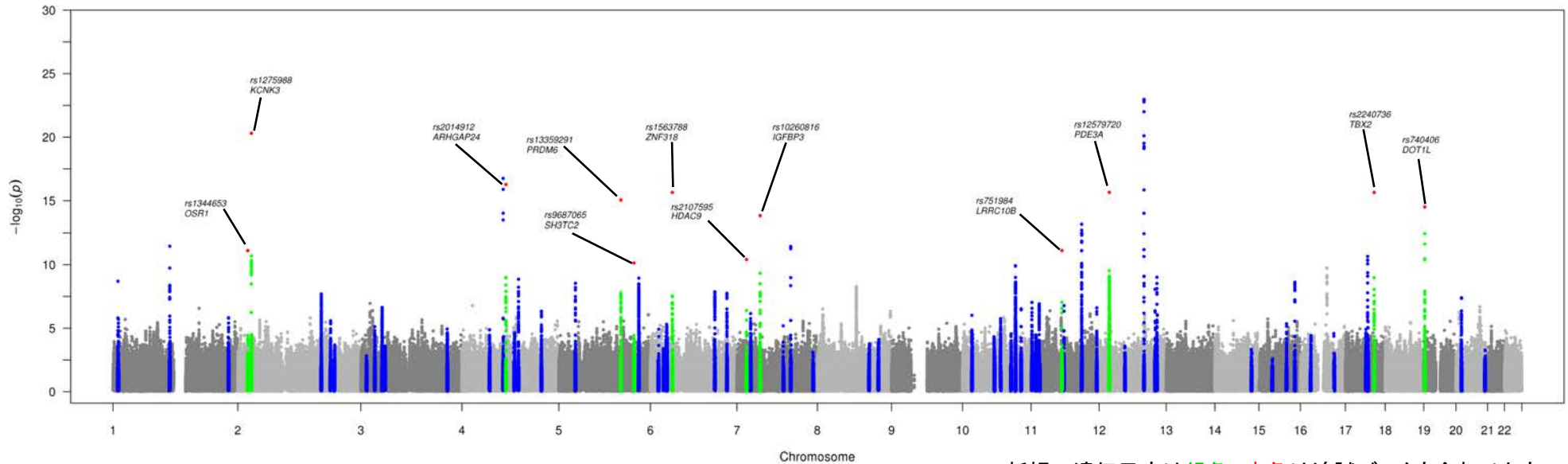
- 
1. イントロ
 2. 統計モデル
 - 3. 解析結果**
 4. 細かい点をいくつか

(高) 血圧の大規模GWAS

Study	Publication	人数	ゲノムワイド有意なDNA多型の数	うち新規のもの
WTCCC	Nature 447:661, 2007	英国人5000	0	0
Global BPgen	Nat Genet 41:666, 2009	欧米人34433 + 追試	8	8
CHARGE	Nat Genet 41:677, 2009	欧米人29136 + 追試	8	8
AGEN-BP	Nat Genet 43:531, 2011	東アジア人19608 + 追試	7(+2)	5
ICBP	Nature 478:103, 2011	欧米人69395 + 追試	29	16
iGEN-BP	Nat Genet 47:1282, 2015	東アジア人31516 + 欧米人35352 + 南アジア人33126 + 追試	35	12

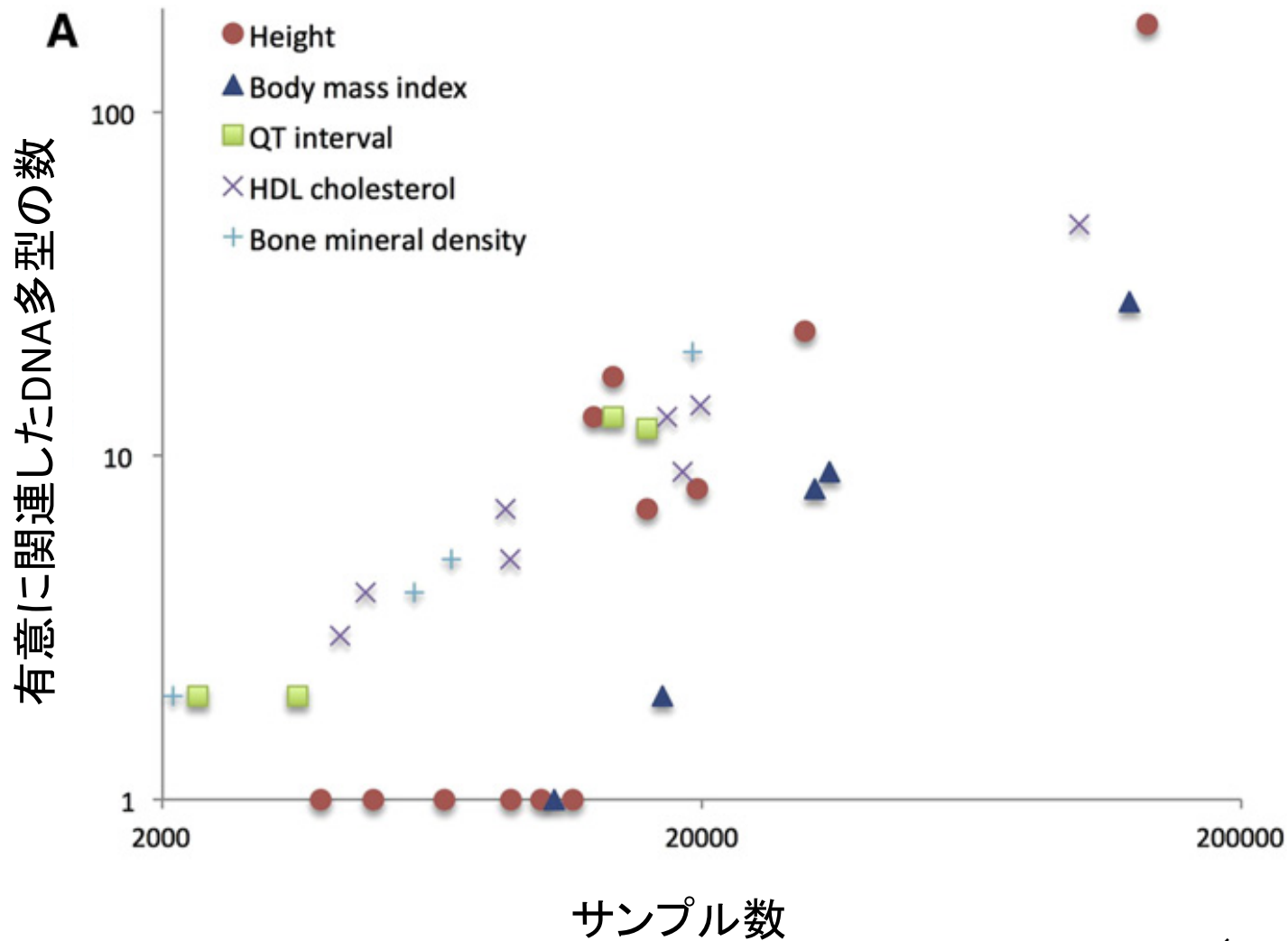
GWASで見つかった血圧関連DNA多型

新規のもの 12カ所を含む、計 52カ所のDNA多型を同定・確認



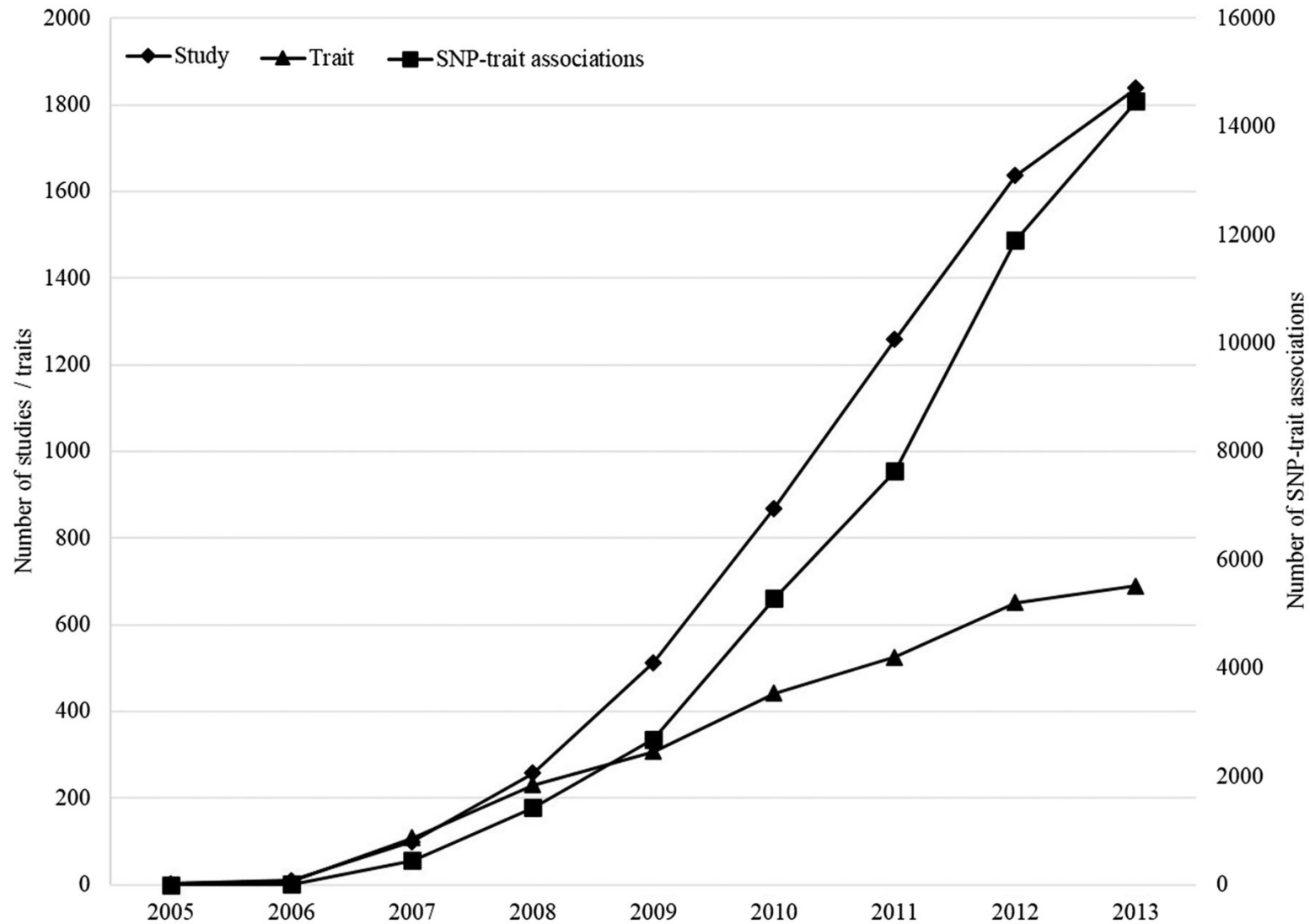
新規の遺伝子座は緑色、赤色は追試データも合わせたもの

GWASの大規模化による検出力向上



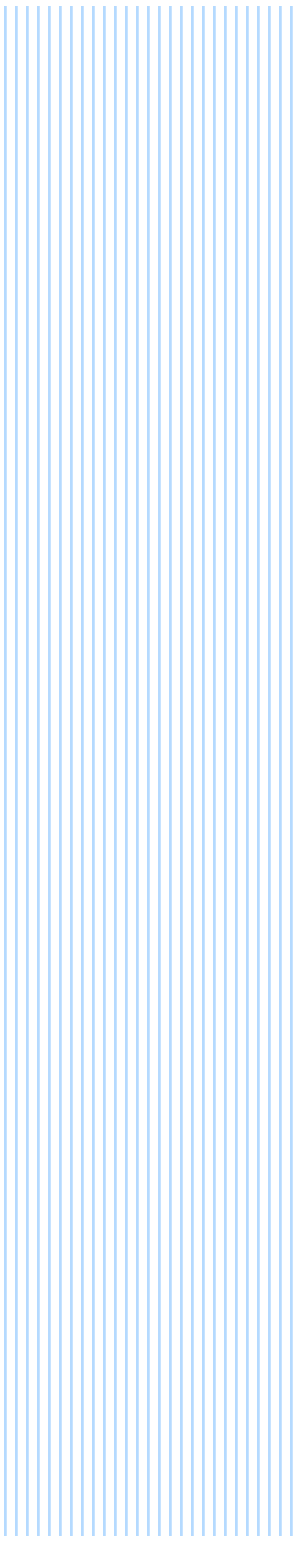
GWASの成功

<http://www.ebi.ac.uk/gwas/>



2007年頃から普及

Welter et al. NAR (2014) 42:D1001

- 
1. イントロ
 2. 統計モデル
 3. 解析結果
 4. 細かい点をいくつか

GWASのメタ解析

- 1施設のGWASではサンプル数に限りがあり、複数のGWASをメタ解析するのが、今は主流

- 連続形質 y_i は、例えば血圧

- 個別GWASで、DNA多型の効果を推測

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

- i 番目の人のDNA多型遺伝子型を $x_i = 0, 1, 2$
- i 番目の人の連続形質の値を y_i
- 誤差 $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$
- 連続形質に対するDNA多型の効果 β を線形回帰で推定

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

- 複数GWASで推定された効果をメタ解析で統合

- j 番目の研究における効果の推定値が β_j 、標準誤差が s_j
- $1/s_j^2$ で重み付けした平均
- 全体での効果の推定値 β 、標準誤差が s
- メタ解析では、個人情報(遺伝型、形質)は不要

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

さらなる大規模化で疾患感受性遺伝子をもっと見つけよう

- 身長・BMIについては、ありふれたDNA多型(1000人ゲノムでimputeできる多型)、遺伝率のほとんどを説明できる

形質	DNA多型で説明できる分散	家族研究で推定される遺伝率
身長	56%	60-70%
体重	27%	30-40%

- ありふれた形質については、恐らく、**サンプル数**を増やして検出力を上げれば、関連が弱い遺伝子も見つかってくる

Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index

Nat Genet (2015) 47:1114

Jian Yang^{1,2,24}, Andrew Bakshi¹, Zhihong Zhu¹, Gibran Hemani^{1,3}, Anna A E Vinkhuyzen¹, Sang Hong Lee^{1,4}, Matthew R Robinson¹, John R B Perry⁵, Ilja M Nolte⁶, Jana V van Vliet-Ostaptchouk^{6,7}, Harold Snieder⁶, The LifeLines Cohort Study⁸, Tonu Esko⁹⁻¹², Lili Milani⁹, Reedik Mägi⁹, Andres Metspalu^{9,13}, Anders Hamsten¹⁴, Patrik K E Magnusson¹⁵, Nancy L Pedersen¹⁵, Erik Ingelsson^{16,17}, Nicole Soranzo^{18,19}, Matthew C Keller^{20,21}, Naomi R Wray¹, Michael E Goddard^{22,23} & Peter M Visscher^{1,2,24}

ただし異論もある

PNAS (2016) 113:E61

PNAS (2016) 113:E4579

PNAS (2016) 113:E4581

Limitations of GCTA as a solution to the missing heritability problem

Siddharth Krishna Kumar^{a,1}, Marcus W. Feldman^a, David H. Rehkopf^b, and Shripad Tuljapurkar^a

まとめ

- ゲノムワイド関連解析(GWAS)では、多数の罹患者と健常者についてDNA多型をゲノム全体に渡って測定し、両グループで有意に頻度が異なるDNA多型を探索する。
- これまでに数百の疾患や形質についてGWASが行われ、万以上のDNA多型との関連が同定された。
- 高血圧などの生活習慣と関連する個々のDNA多型は(本物ではあるものの)関連が極めて弱いことが分かってきた。
- 検出力を上げるために大規模なサンプルが必要であり、複数のGWASを統合するメタ解析、多人種を統合するメタ解析が行われている。
- 今後の方向性
 - DNA多型と分子的形質の関連解析→熊坂先生講演
 - 疾患感受性DNA多型を利用した罹患予測→八谷先生講演
- パワーポイント <http://fumihiko.takeuchi.name>