

ゲノムワイド関連解析 (GWAS) : 多人種大規模化の成功、 地域差の心配

竹内史比古

国立国際医療研究センター研究所

第70回インシリコ・メガバンク研究会

2016.01.15 @ToMMo

遺伝子変異と体質: 例1

- *ALDH2*, アルデヒド脱水素酵素2
遺伝子

- 変異

- c.1510G>A (p.Glu504Lys) rs671
- *ALDH2**2アリル

- 変異ホモ接合

- 酵素活性ない
- お酒が飲めない

- 変異ヘテロ接合

- *ALDH2*(4量体)活性が1/16
- お酒を飲むと赤くなる

エタノール



↓

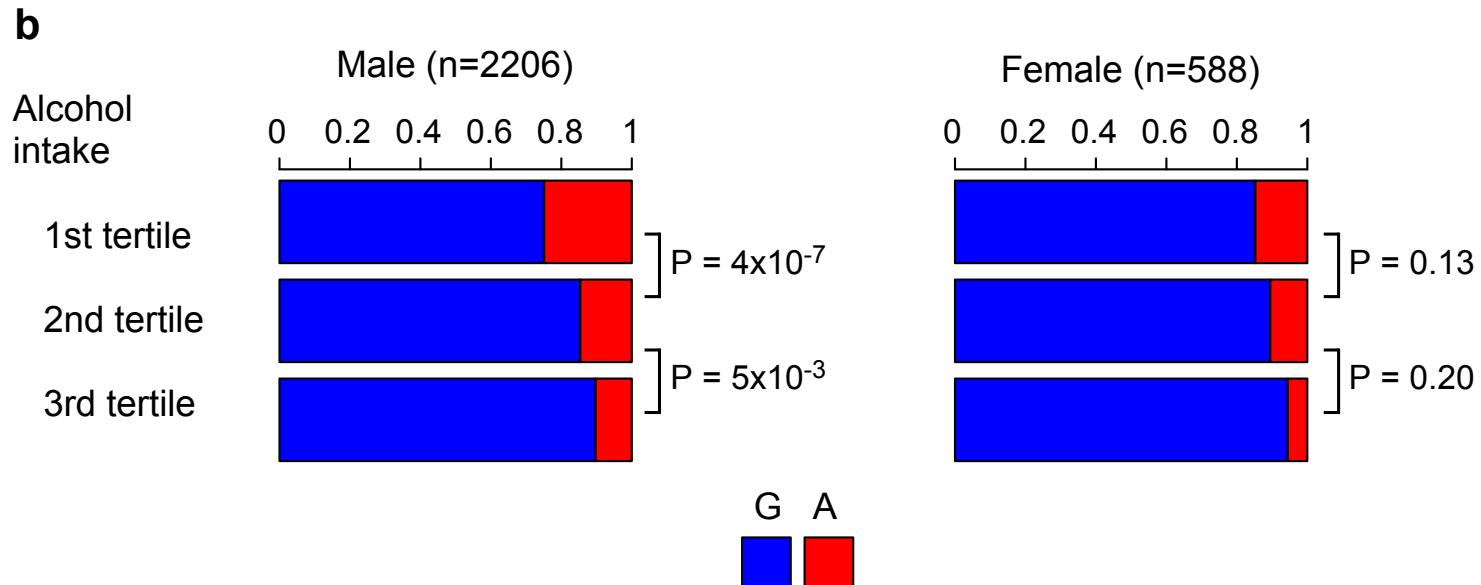
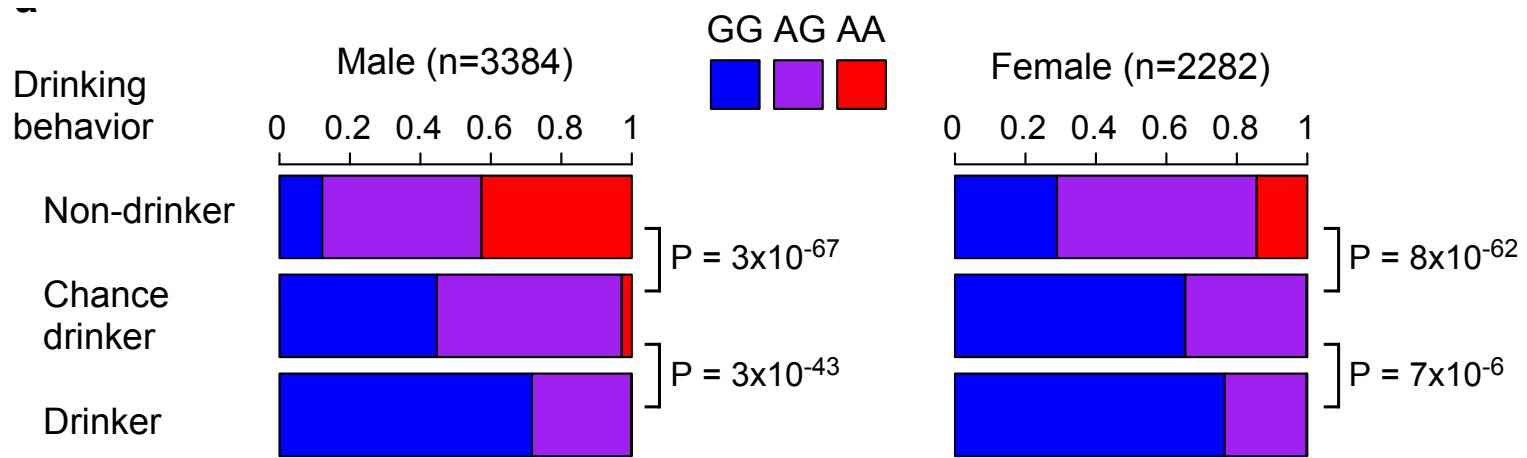
アセトアルデヒド→毒性

↓ *ALDH2*

酢酸

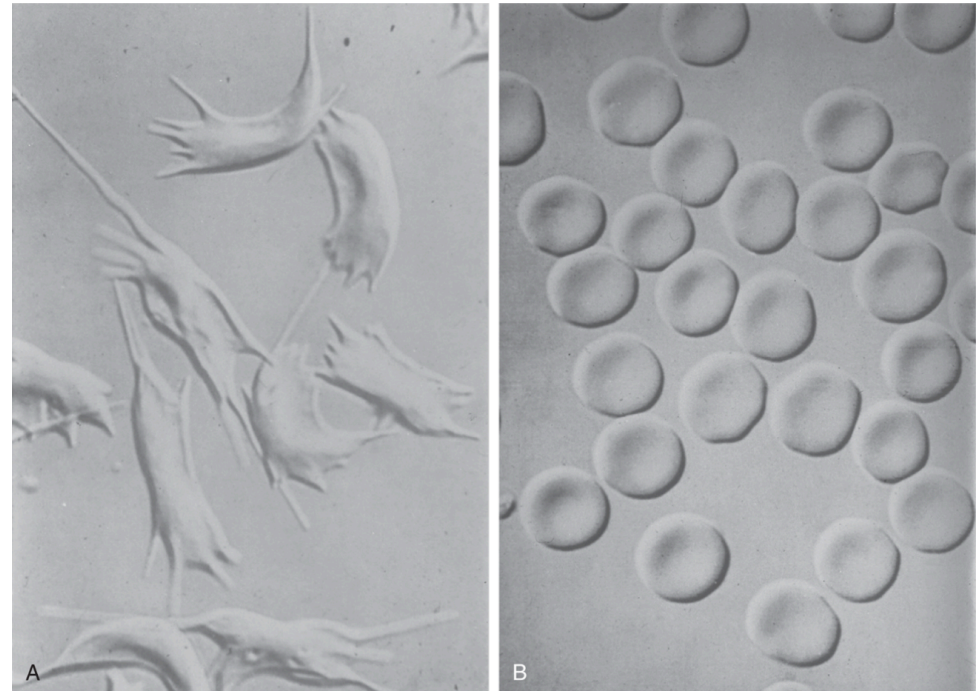


遺伝子型と飲酒行動



遺伝子変異と体質: 例2

- *HBB*, ヘモグロビン β 遺伝子
- 変異
 - c.20A>T (p.Glu7Val) rs334
 - β^S アリル
- 変異ホモ接合
 - ヘモグロビンが凝集
 - 赤血球が鎌状になる
 - 赤血球が壊され貧血に
- 変異ヘテロ接合
 - 貧血ない
 - マラリア原虫に感染しにくい

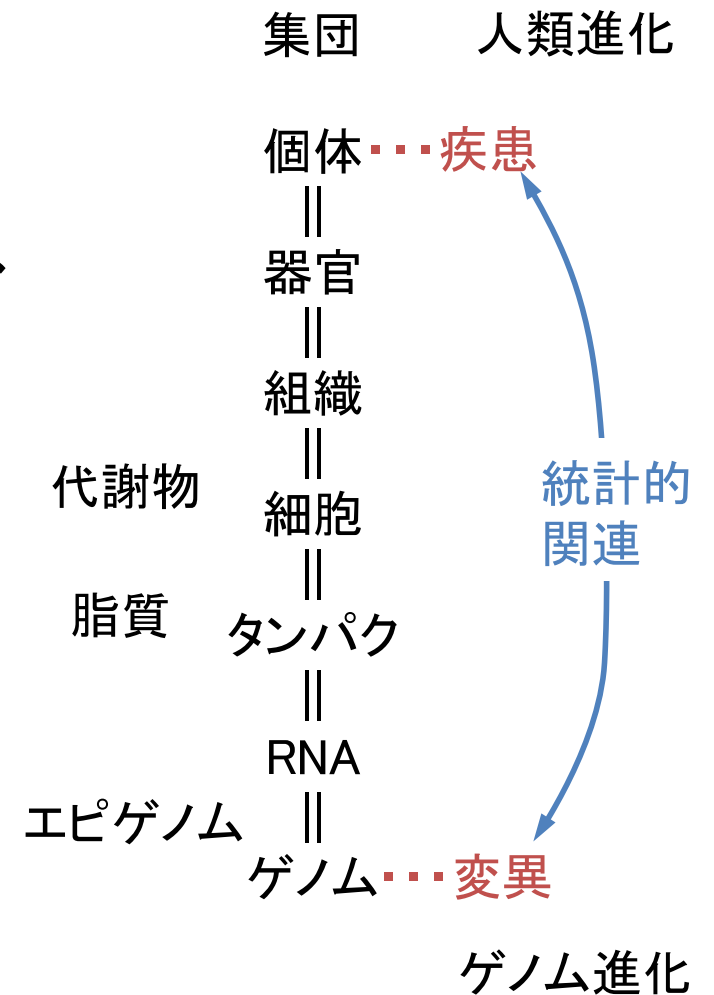


遺伝子変異と病気： 疾患感受性遺伝子とは？

- 遺伝子の型により、病気の罹り易さ(感受性)が変わる遺伝子
 - dystrophin (*DMD*; Xp21.1)
 - 変異による短いタンパク質 → Duchenne 型筋ジストロフィー
 - *ABO* (9q34.2)
 - O型 → 静脈血栓塞栓症に罹りにくい (OR=0.33)
- 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
 - 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

関連解析のコンセプト

- ゲノム変異と疾患
 - 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
- ゲノム変異
 - 安価に大規模に測定できるようになった



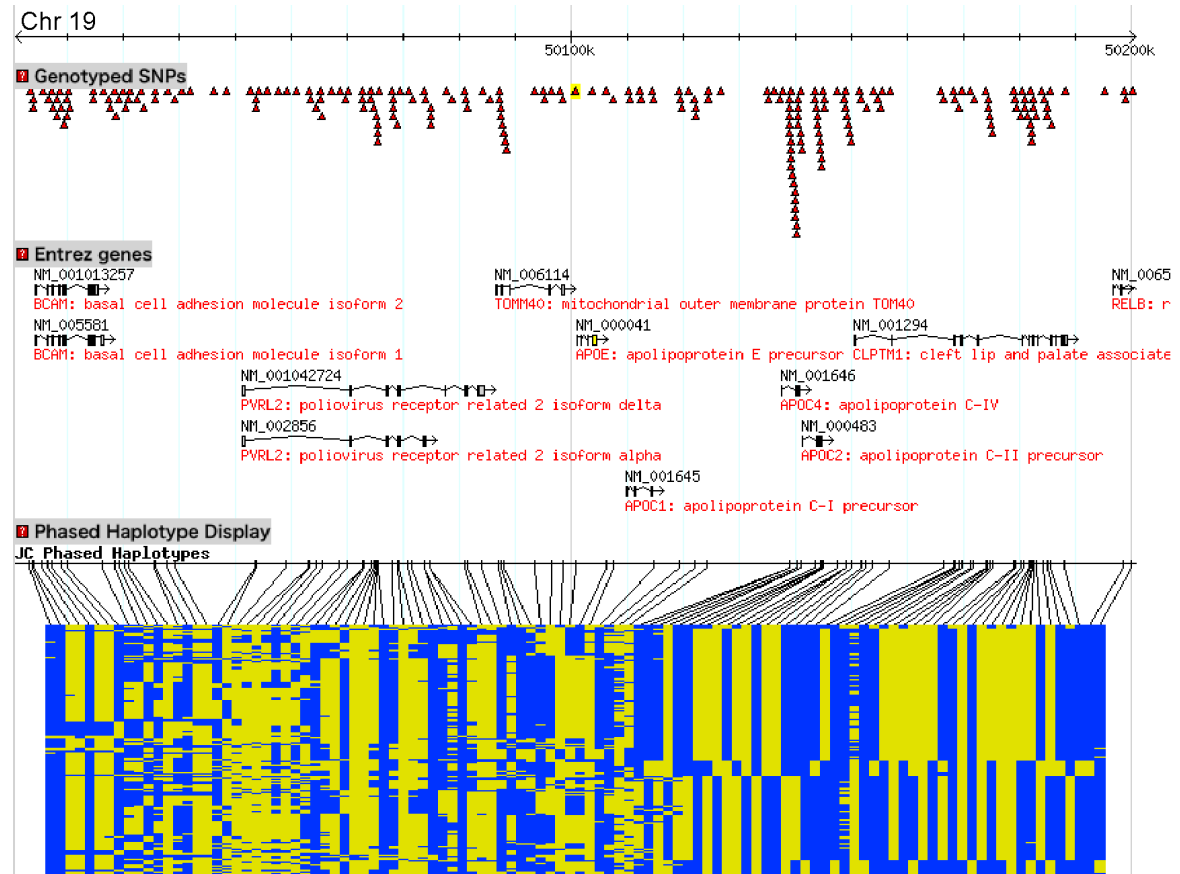
ゲノムワイド関連解析 (GWAS)

- 3万個の遺伝子を網羅的に調べ、疾患感受性遺伝子を統計的に絞り込む
 - 全ての遺伝子型について、疾患との相関を調べる
 - 罹患者はA型が多い、健常者はB型が多い
- 目標: ありふれた(頻度 $\geq 5\%$)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
- 全ゲノム配列決定は、多数の人ではまだできない
 - 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
 - マイクロアレイに搭載されてないSNPの情報は推測(imputation)できる
- 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・健常者を数千人タイピングする必要がある

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

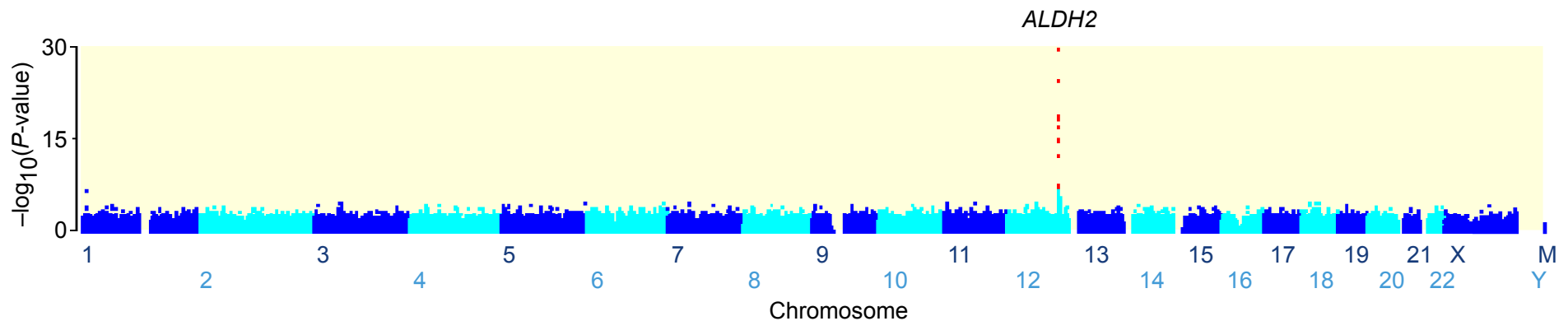
SNPsの相関(連鎖不平衡)

- 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



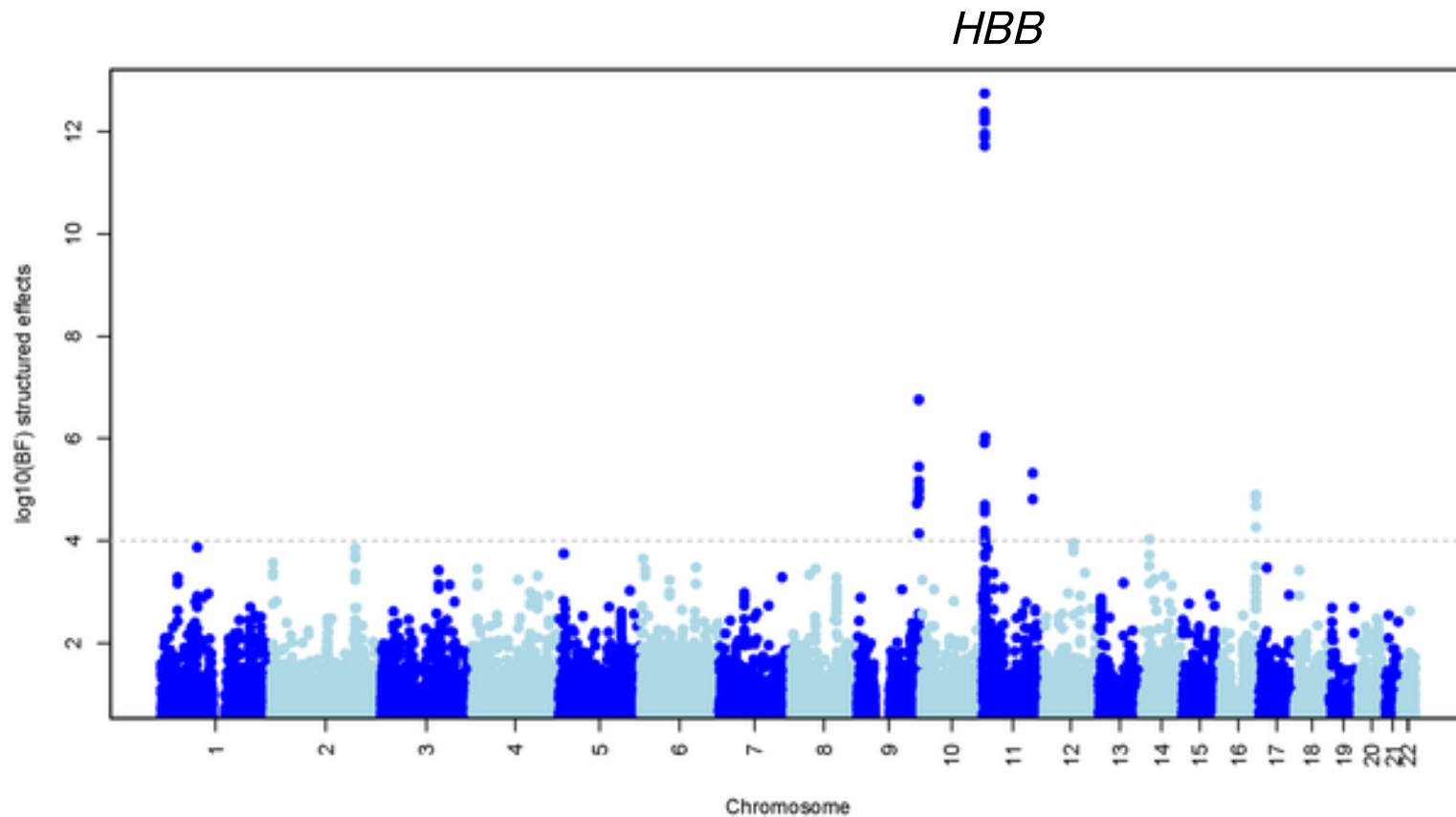
飲酒行動のGWAS

- 飲む 733名、飲まない 729名
- *ALDH2* rs671 でオッズ比0.31、 $P=1.8 \times 10^{-30}$



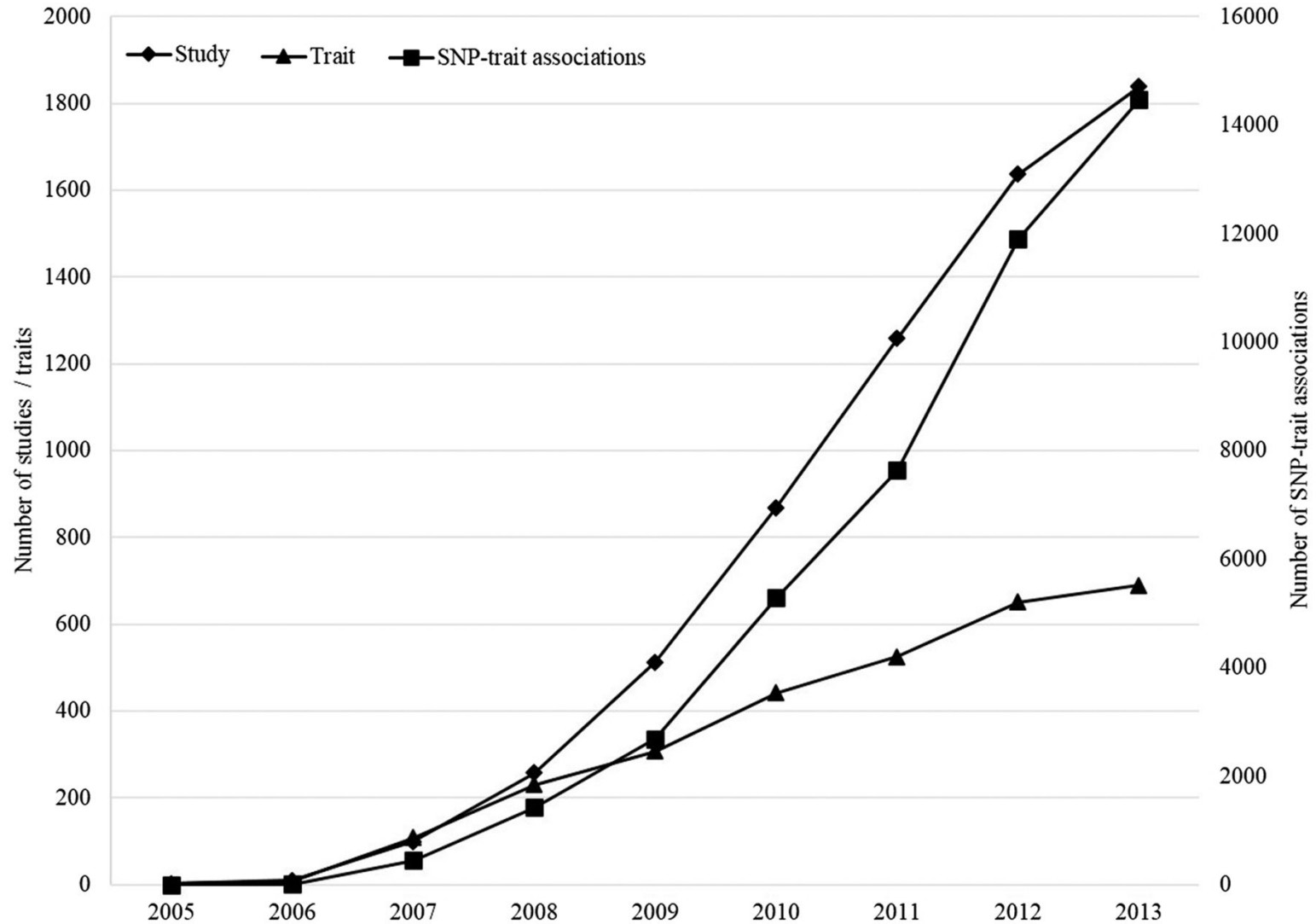
マラリア感受性のGWAS

- 重症マラリア5000名、一般集団7000名



GWASの成功

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

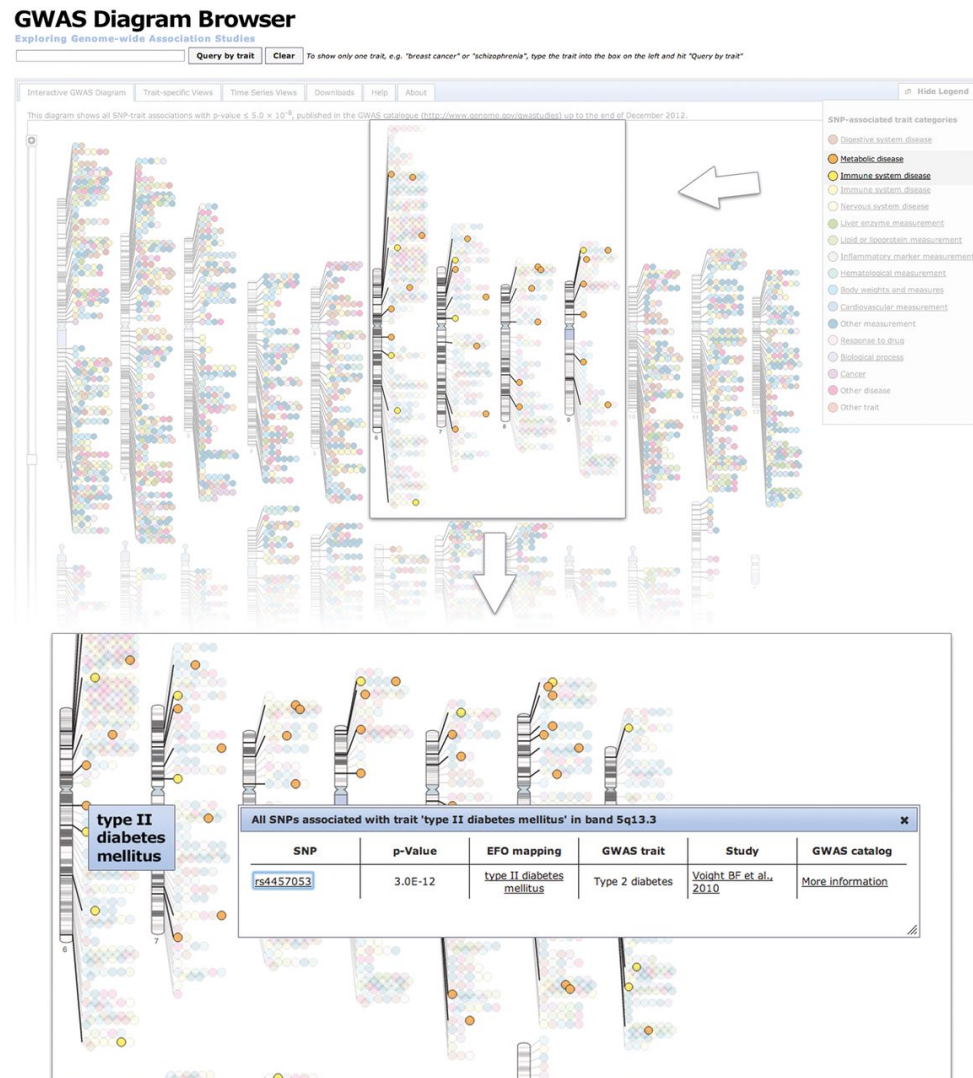


2007年頃から普及

Welter et al. NAR (2014) 42:D1001

GWASの成功

The interactive GWAS diagram is a visualization of all SNP-trait associations with $P < 5 \times 10^{-8}$, mapped to the SNP's cytogenetic band.



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

Welter et al. NAR (2014) 42:D1001

数式のスライドをいくつか...

SNPと形質の関連の検定

- i 番目の人のSNP 遺伝子型を $x_i = 0, 1, 2$
 - 例、アレルが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、血圧)
- 線形回帰

- 誤差 $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$

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

- 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

- $y_i \sim \text{Bernoulli}(p_i)$

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

- 帰無仮説: $\beta = 0$

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

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、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 $N \doteq 40/R^2$

→ 弱い関連を検出するには多数のサンプルが必要
→ 複数の GWAS をメタ解析しよう!

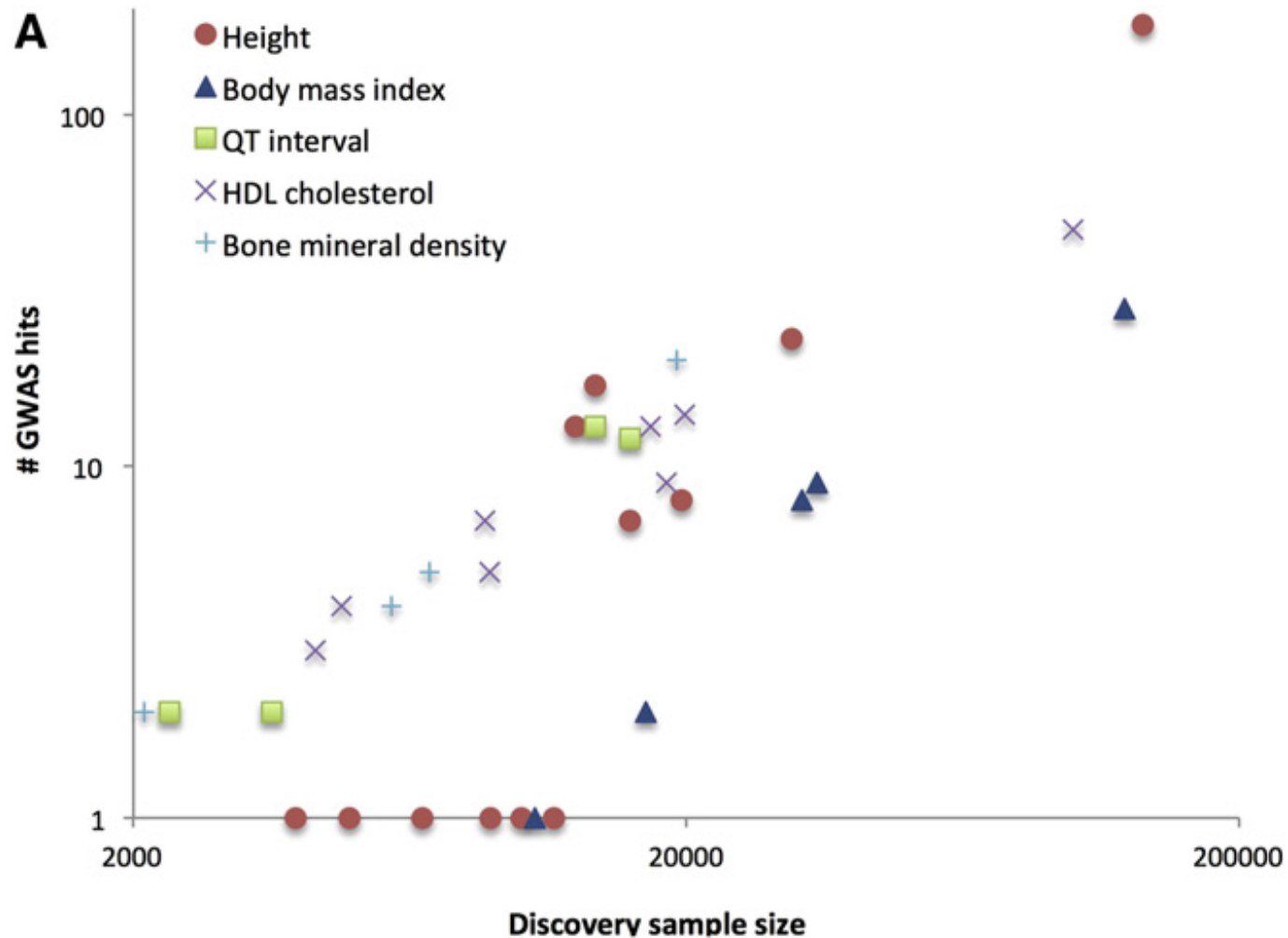
GWASのメタ解析

- 連続形質 y_i は、例えば血圧
- 個別GWASで、SNPの効果을推測
 - 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 $y_i = \alpha + \beta x_i + \varepsilon_i$
 - 誤差 $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$
 - 連続形質に対するSNPの効果 β を線形回帰で推定
- 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

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

GWASの大規模化による検出力向上

Number of Loci against Sample Size



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、1000人ゲノムでimputeできる多型で、遺伝率のほとんどを説明できる

形質	imputeされた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%

- 多型はほぼ網羅している。より低頻度まで調べなくてよい
- ありふれた形質について効率的なのは、恐らく、resequenceするより、アレイでタイピングしてサンプル数を増やす
→ 検出力を上げて関連が弱い遺伝子も見つける

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

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 Esko^{9–12}, 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}

Nat Genet (2015) 47:1114



血圧に関する 多人種大規模GWAS

iGEN-BPコンソーシアム

Kato et al. (2015) Nat Genet 47:1282

(高)血圧の大規模GWAS

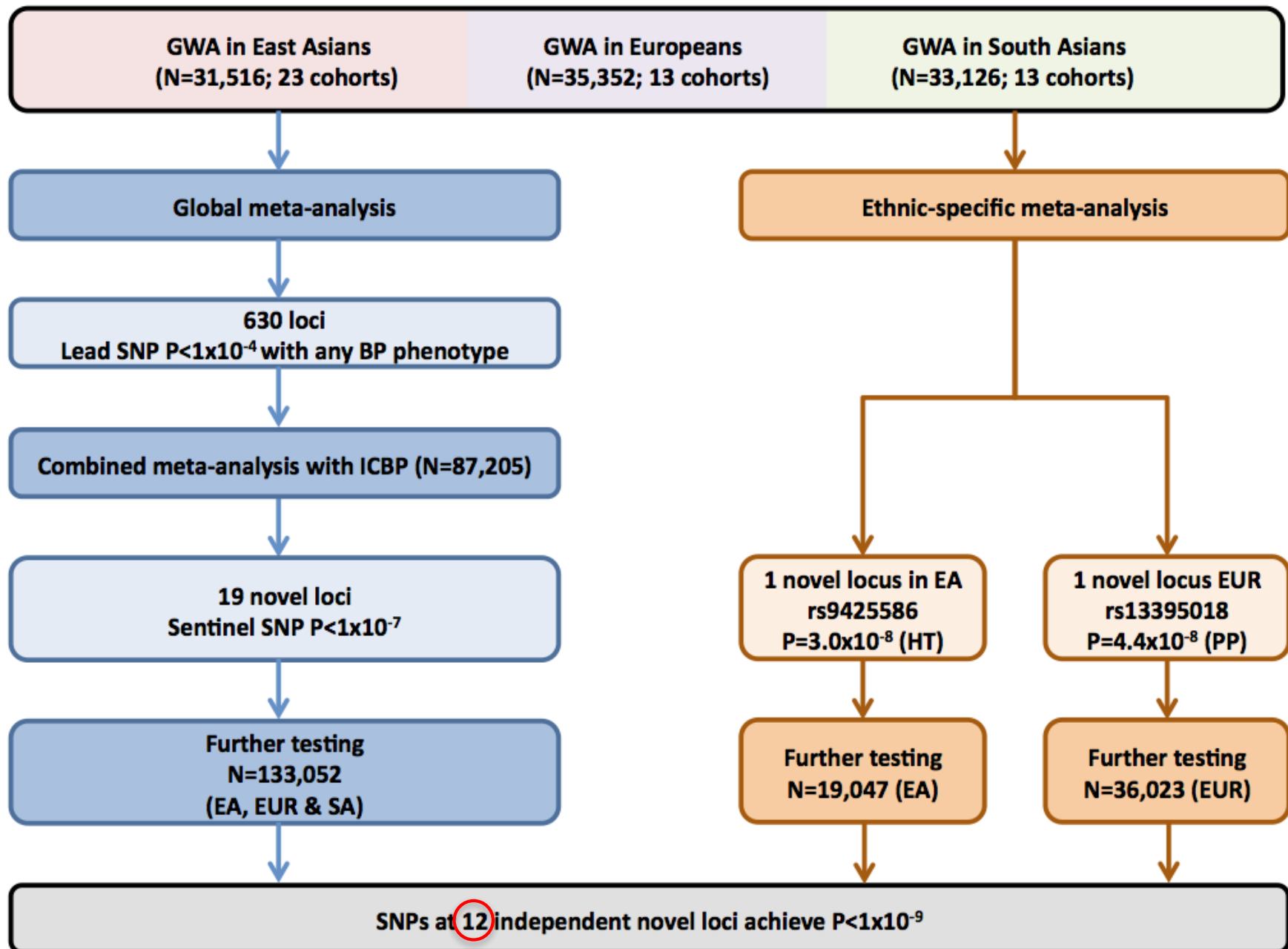
Study	Publication	人数	ゲノムワイド有意なSNPsの数	新規SNPsの数
WTCCC	Nature 447:661, 2007	英国人5000	0	0
Global BPgen	Nat Genet 41:666, 2009	欧米人34433 + 追試	8	8
CHARGE	Nat Genet 41:677, 2009	欧米人29136 + 追試	8	8
AGEN-BP	Nat Genet 43:531, 2011	東アジア人19608 + 追試	7(+2)	5
ICBP	Nature 478:103, 2011	欧米人69395 + 追試	29	16
iGEN-BP	Nat Genet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

結果概要

- 血圧の大規模・多人種ゲノムワイド関連解析を行い、12の新規関連遺伝子座を同定した
- 血圧関連SNPsの多くで、SNPと関連するCpGサイトが近傍にあった
- 血圧関連SNPsはおしなべて、CpGサイトと関連する傾向があった
 - SNP→DNAメチル化→血圧の機序が示唆される

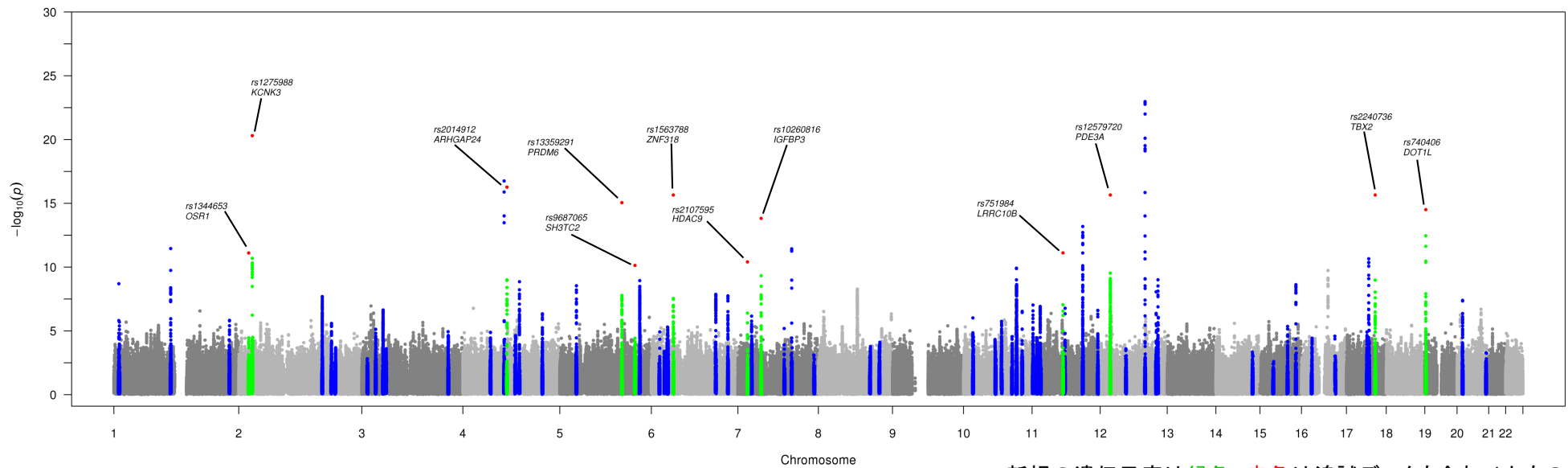
方法: discovery 3人種10万人 + 追試20万人

2.1M
SNPs



GWASで見つかった血圧関連遺伝子座

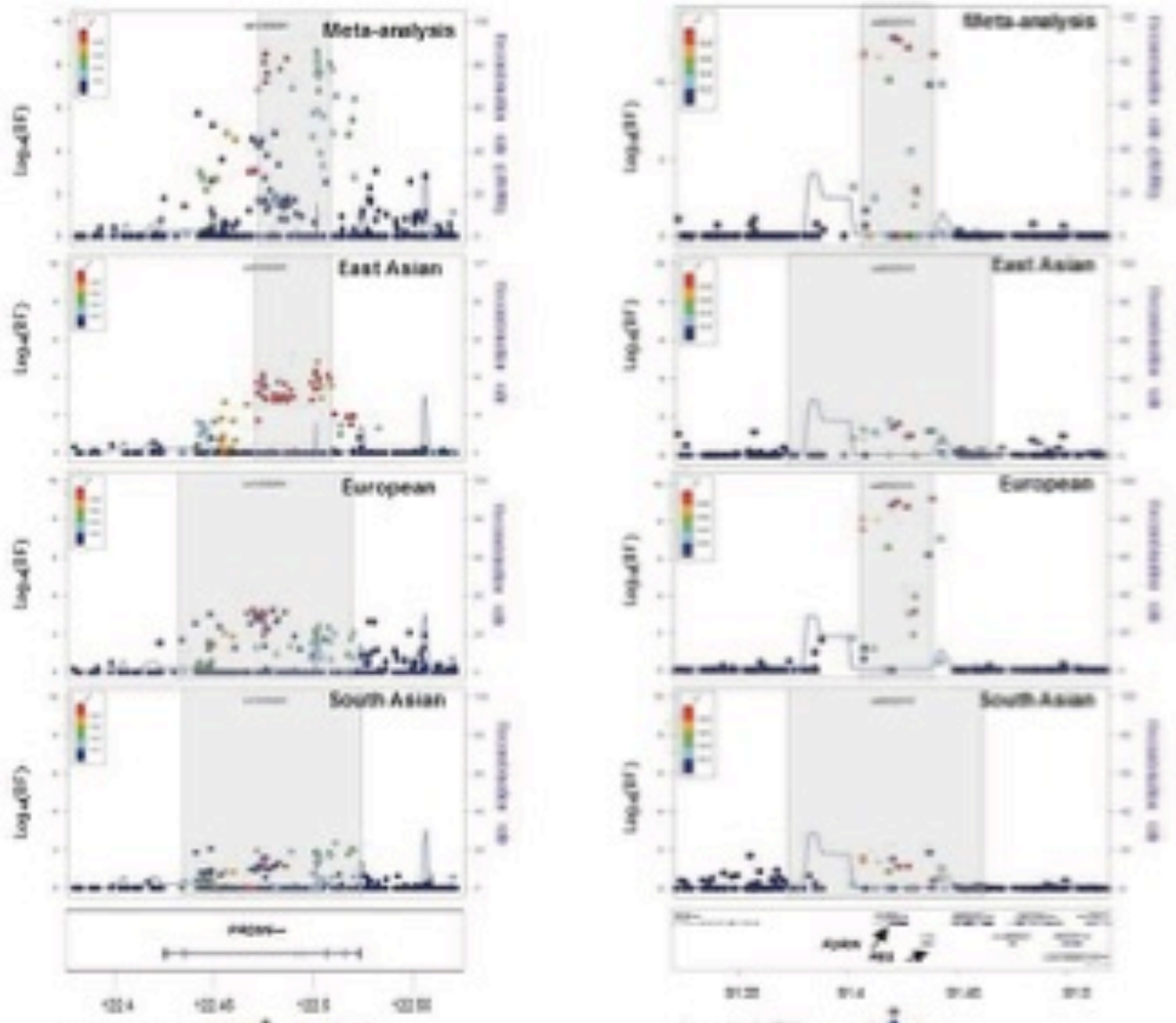
新規のもの **12カ所**を含む、計 52 遺伝子座を同定・確認



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

多人種を利用した、原因変異の絞り込み

- 人種ごとに連鎖不平衡のパターンが異なるの
- 多人種を組み合わせて、原因変異を絞り込める

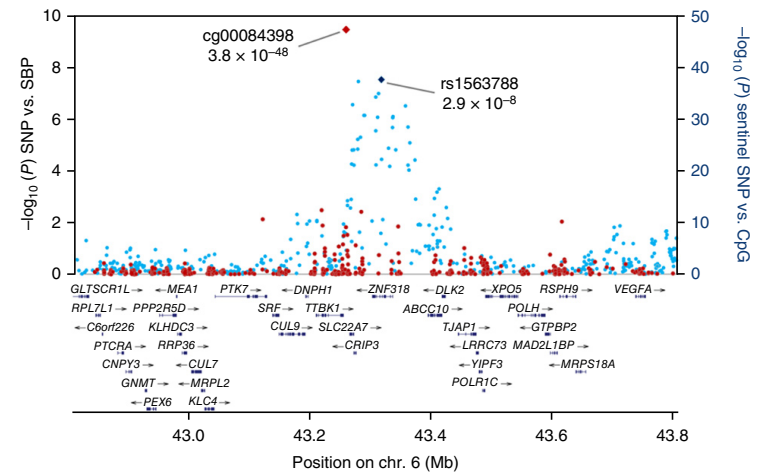
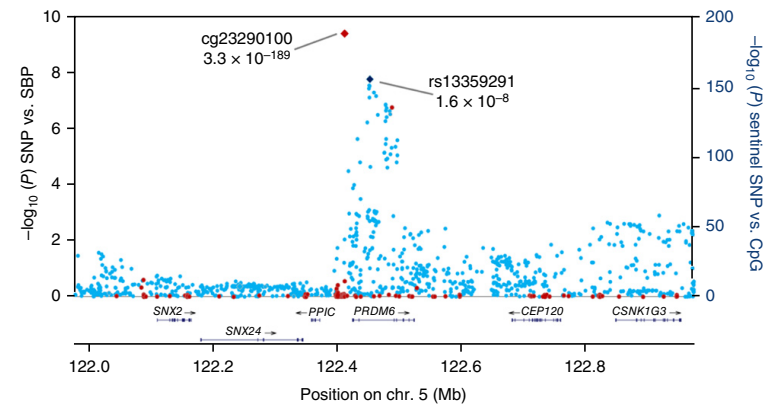
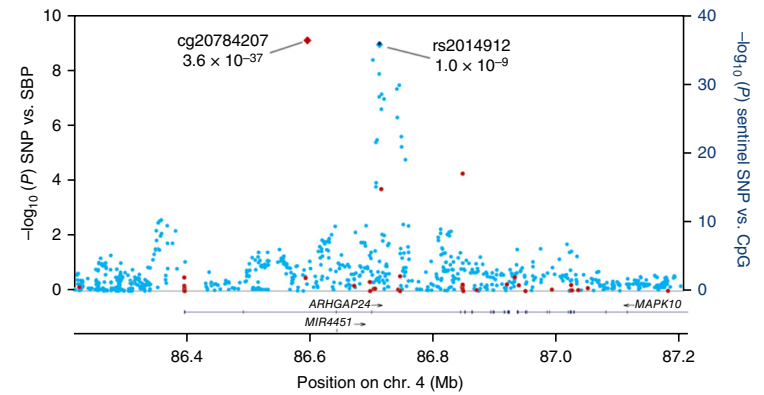


方法: SNPと関連するCpGサイトの探索

- ゲノムワイド有意な**35**遺伝子座(うち新規12)の各々で、P値最小の「標識SNP」を選出
- 標識SNPの近傍1MbのCpGサイトで、**メチル化度**と標識SNP**遺伝子型**が関連するものを探索
 - DNAメチル化450Kアレイを使用
 - 有意水準 $P < 3.8 \times 10^{-6}$ (多重検定補正 $P < 0.05$)
 - discovery: 1904名(南アジア人)
 - **28** SNPs に有意に関連するCpGサイト有り
 - SNPとCpG(P値最小)の距離は1kb~244kb
 - replication: 4780名(欧米人、南アジア人)
 - 全て再現される

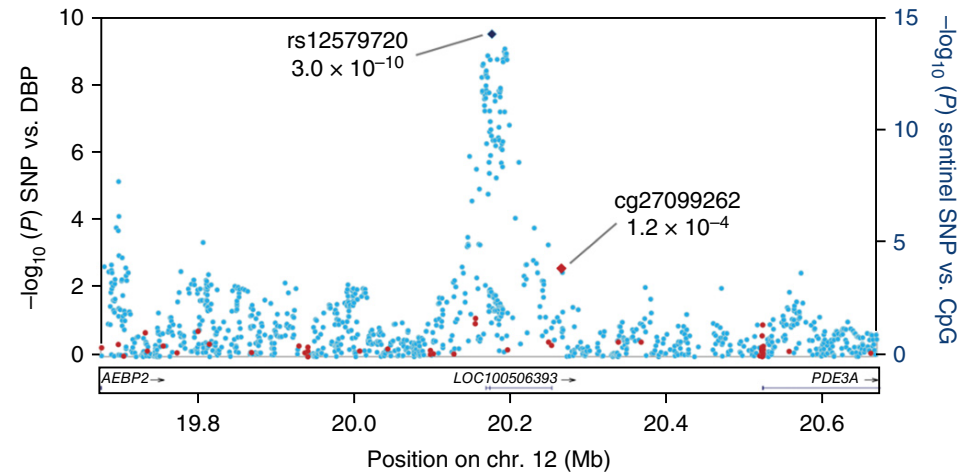
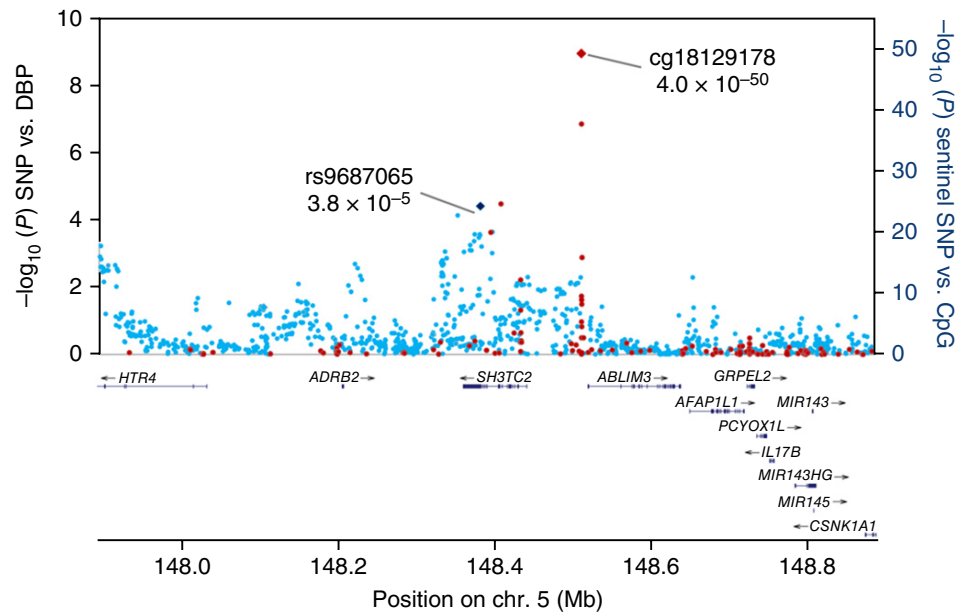
SBP関連遺伝子座

- SNP-血圧形質との遺伝的関連 (●)
- 標識SNP-CpG siteとの関連 (●)



DBP関連遺伝子座

- SNP-血圧形質との遺伝的関連 (●)
- 標識SNP-CpG siteとの関連 (●)



PP関連遺伝子座

- SNP-血圧形質との遺伝的関連 (●)
- 標識SNP-CpG siteとの関連 (●)

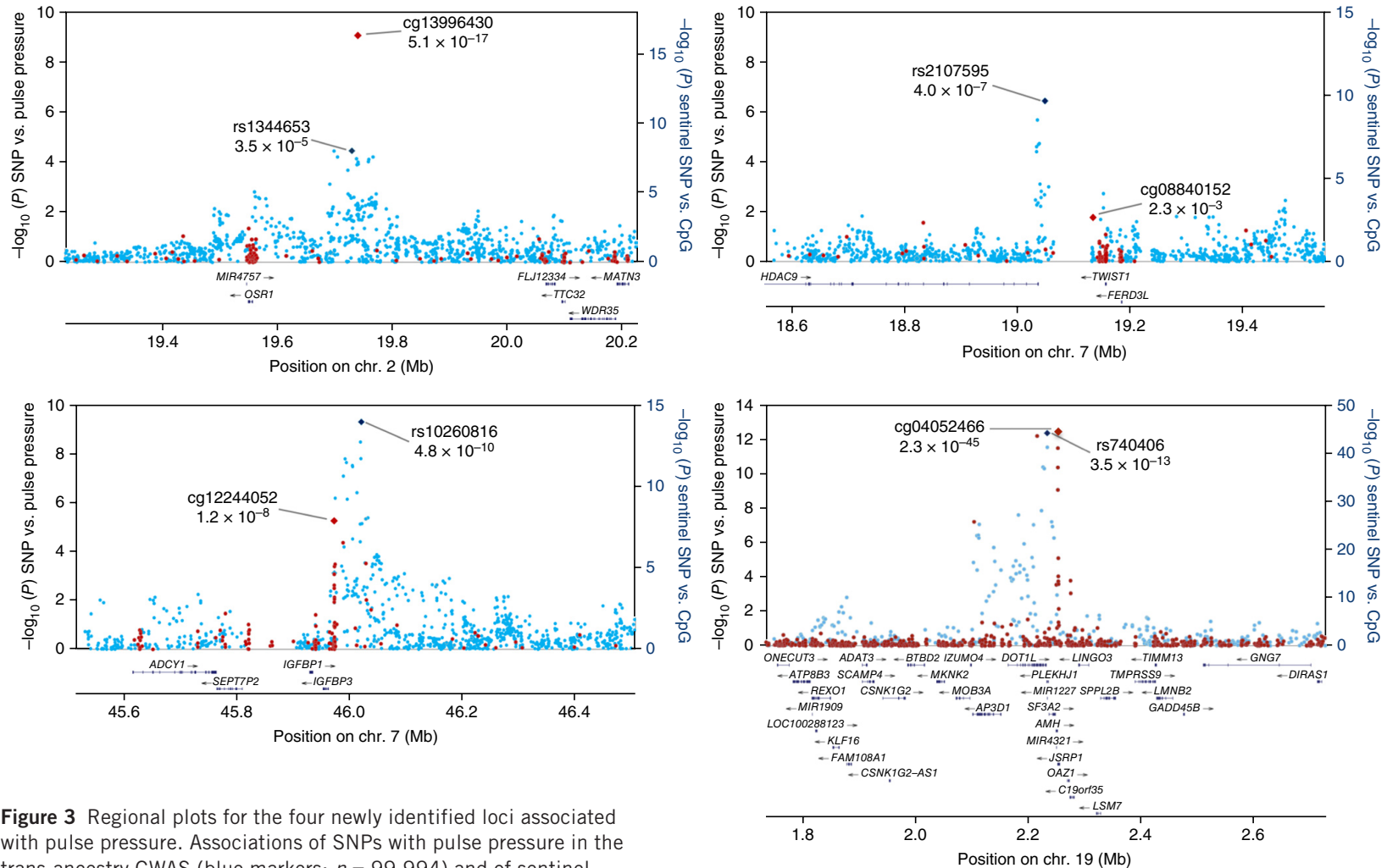
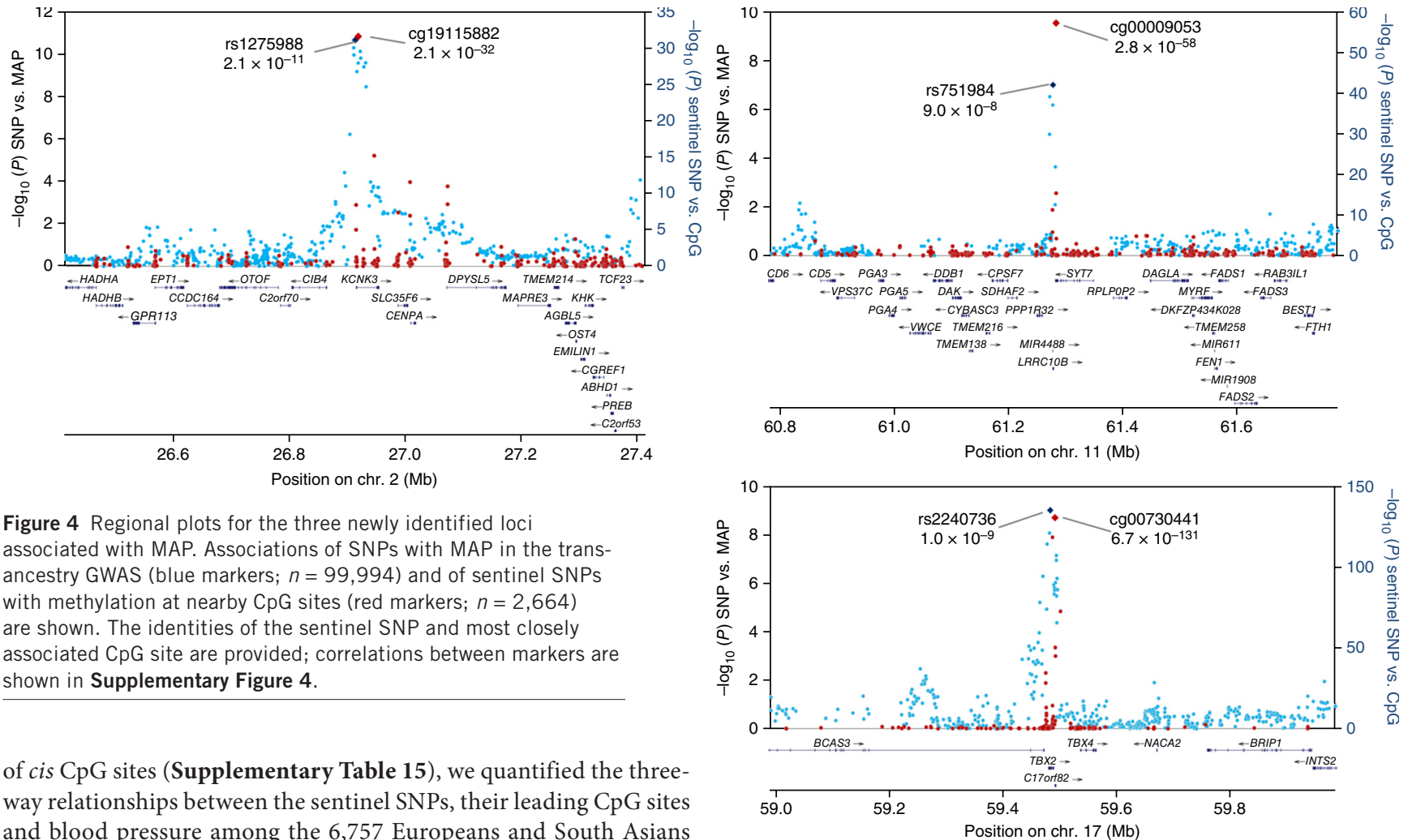


Figure 3 Regional plots for the four newly identified loci associated with pulse pressure. Associations of SNPs with pulse pressure in the trans-ancestry GWAS (blue markers; $n = 99,994$) and of sentinel

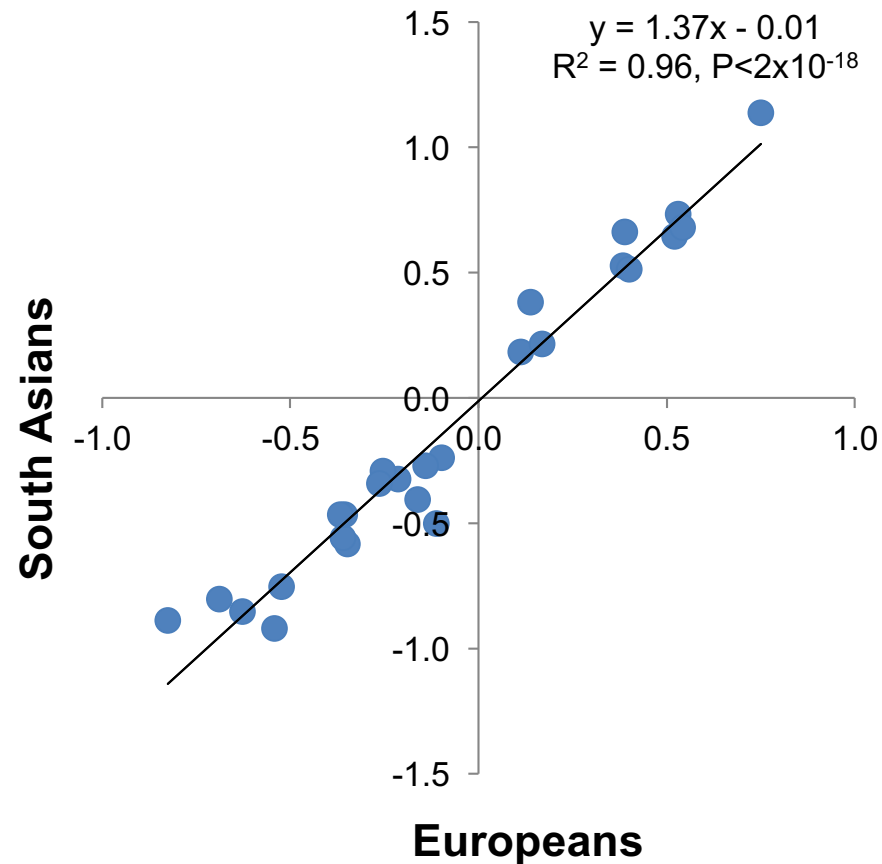
MAP関連遺伝子座

- SNP-血圧形質との遺伝的関連 (●)
- 標識SNP-CpG siteとの関連 (●)



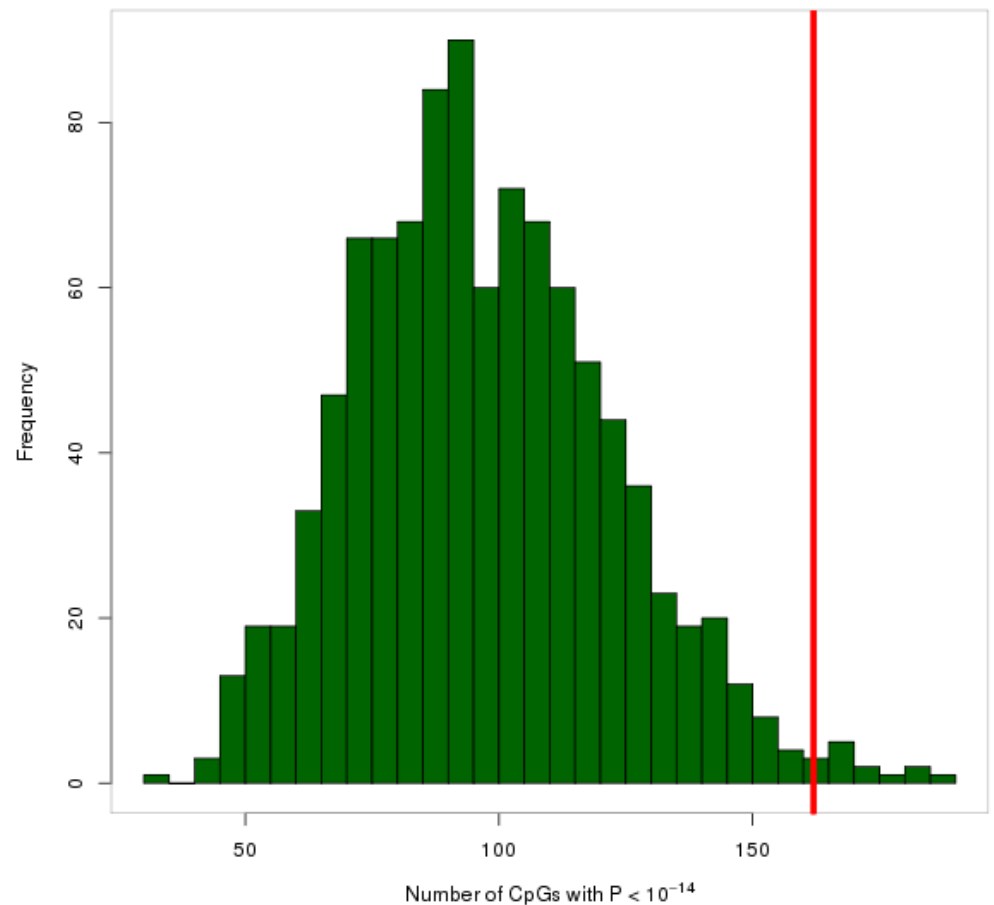
異人種間でのSNP-DNAメチル化関連の再現性

28 CpG sites (most strongly associated with BP)での再現性をプロット

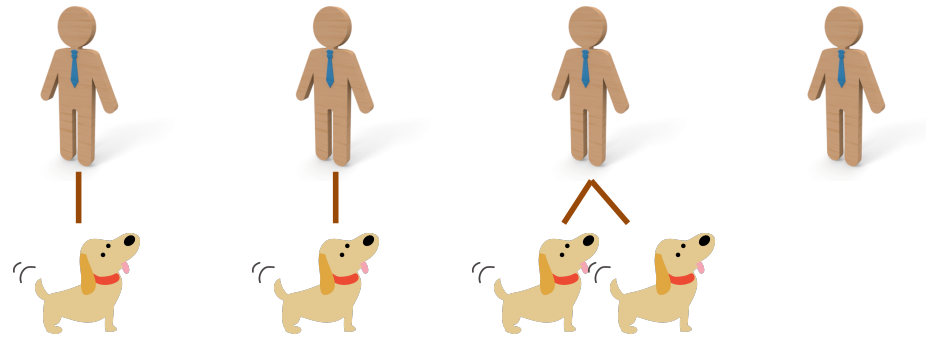


血圧関連SNPsは、近傍のDNAメチル化と関連する傾向がある

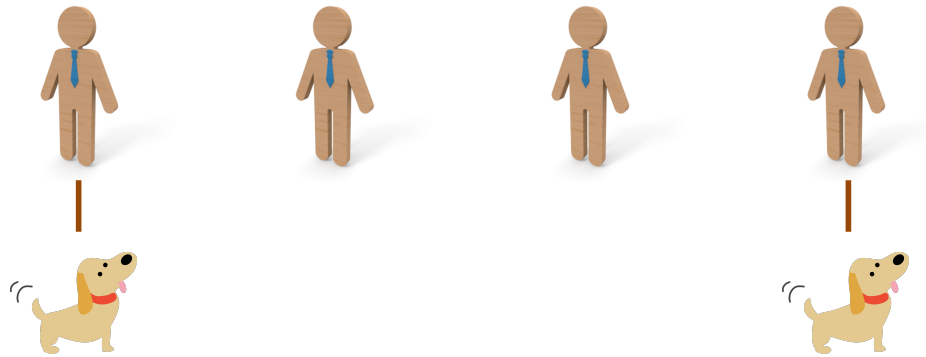
- 35の標識SNPの近傍1MbのCpGサイトで、メチル化度と標識SNP遺伝子型が関連するものを勘定
- ゲノムからランダムに選んだコントロールSNPでも勘定
- 血圧関連SNPsでは2倍enrichしている (P=0.01)



• 早起きの人達



• ランダムにサンプルされた人達



早起き

犬を飼う

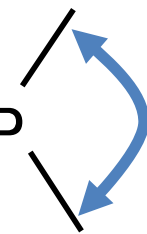


傾向が合う
因果関係は分からない

血圧

SNP

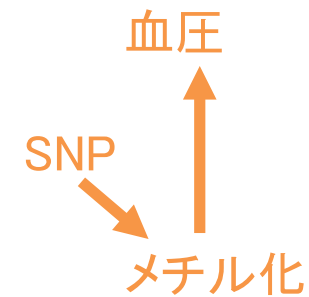
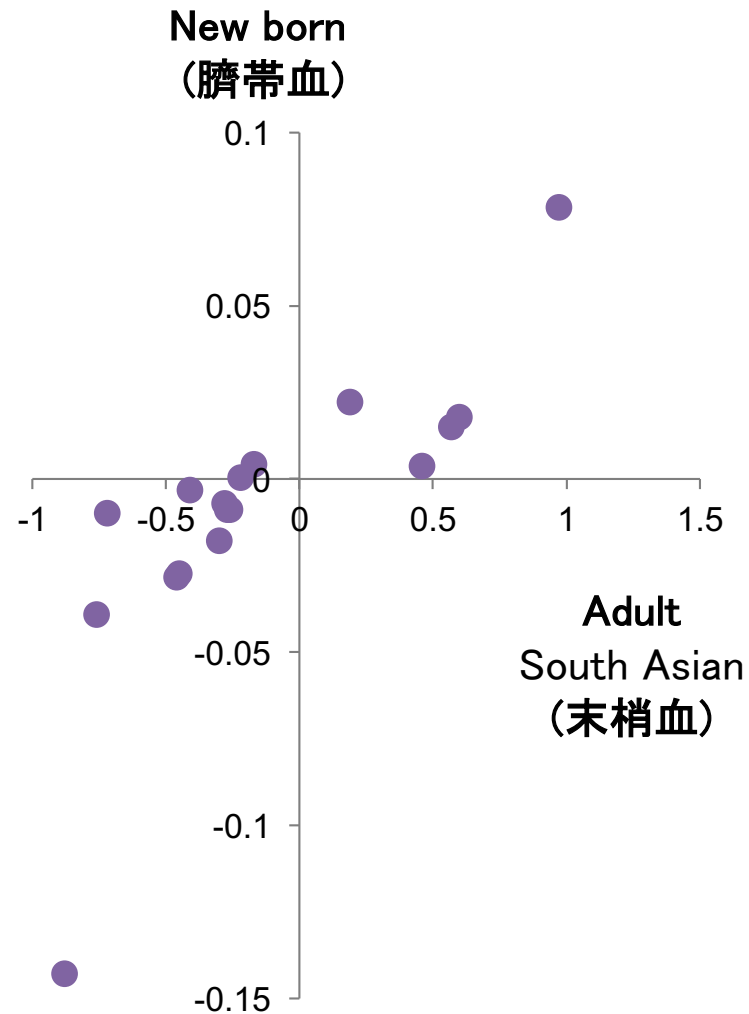
メチル化



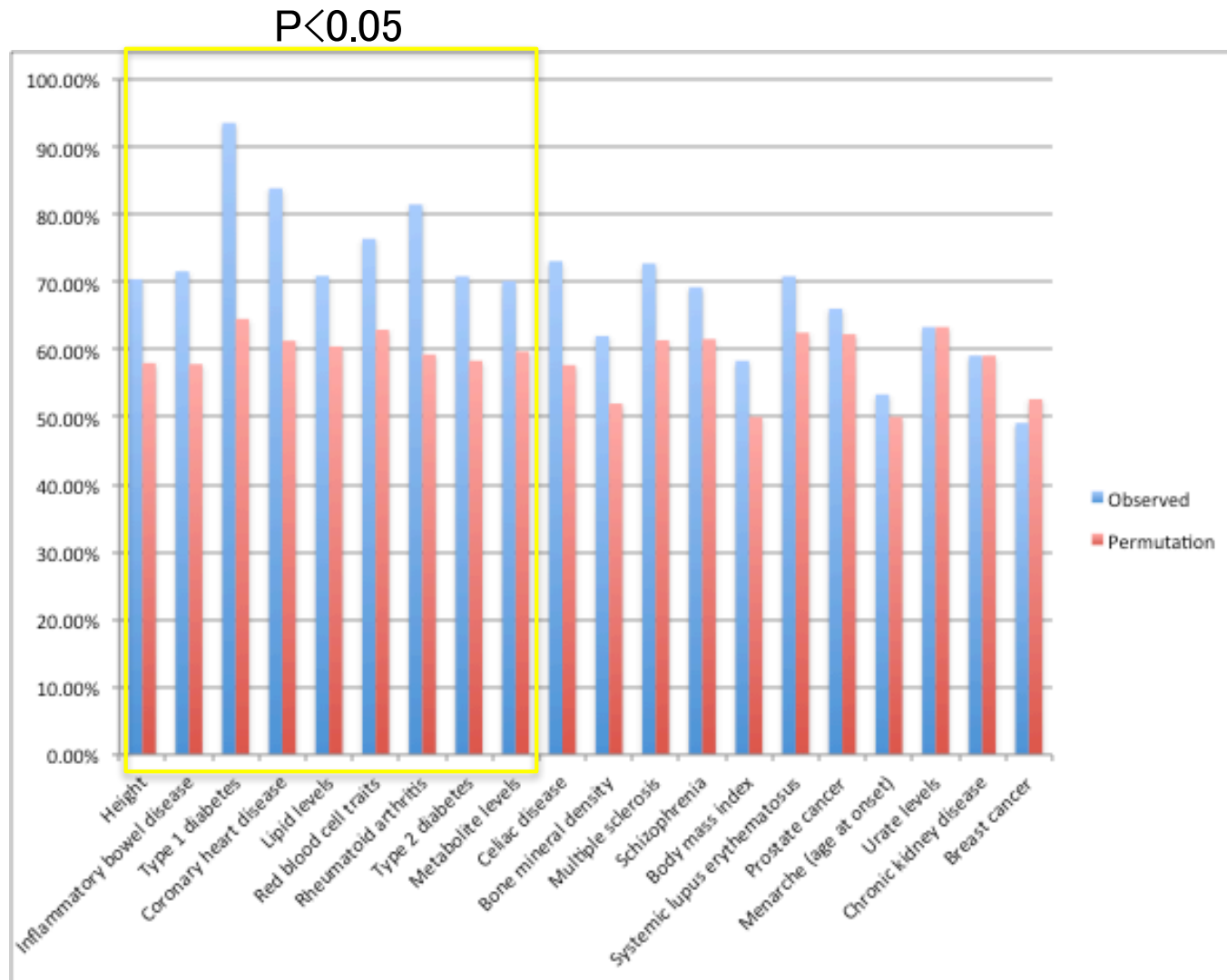
因果関係は
分からない

胎児血でのSNP-DNAメチル化関連の再現性

28 CpG sites (most strongly associated with BP)での再現性をプロット



他疾患でも、関連SNPsが近傍のDNAメチル化と関連する傾向がある



概要

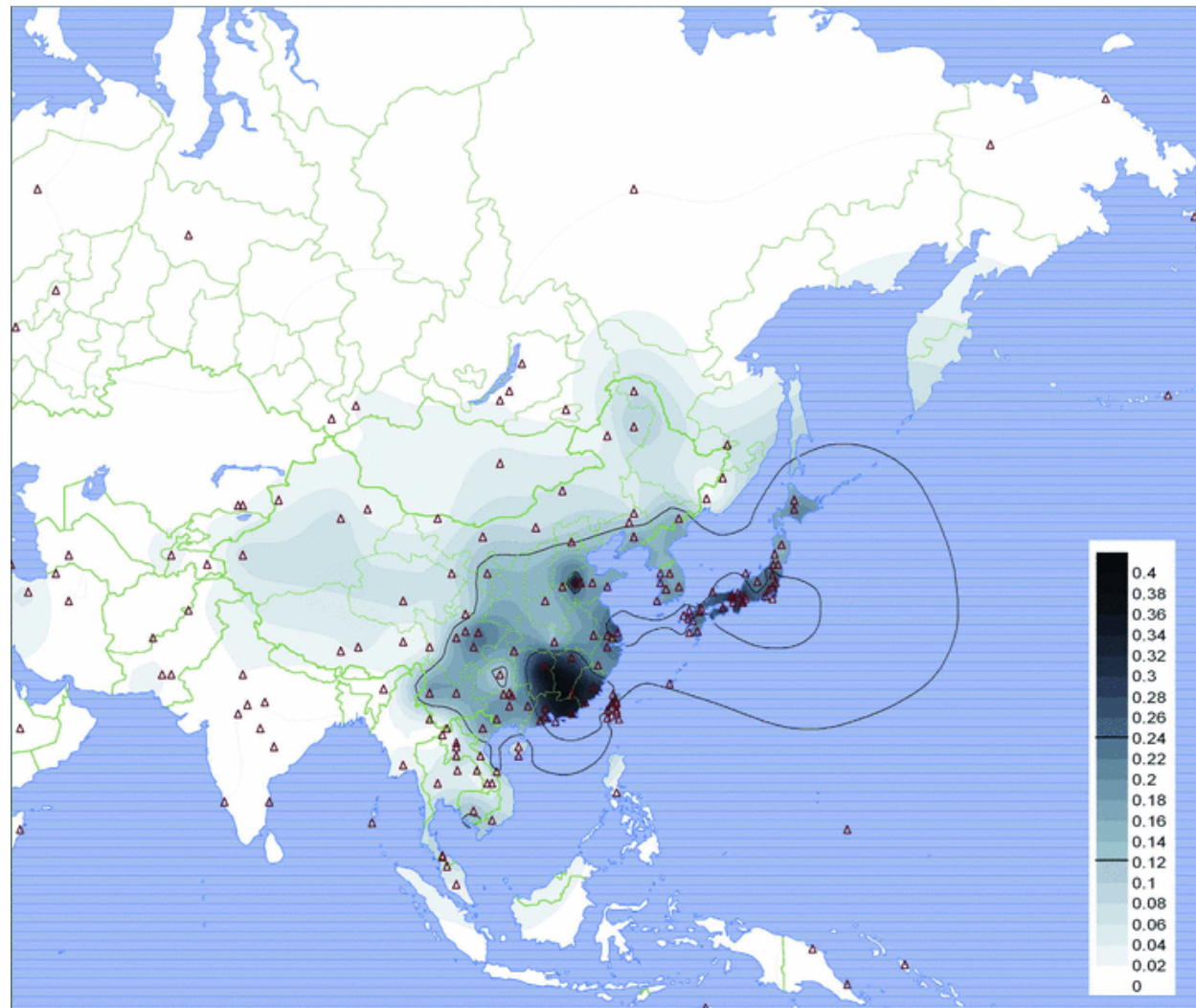
- 血圧(高血圧)に関して、最初の、3人種横断的な大規模ゲノムワイド関連解析を実施し、新規のもの 12 カ所を含む、計 52 遺伝子座を同定・確認した。
- 血圧(高血圧)との関連を示す遺伝子座のSNPは、同時に DNA メチル化とも関連を示し得ることを見出し、DNA メチル化が、SNPと血圧(高血圧)の個人差とをつなぐ分子機序の一部を成す可能性が示唆された。
- 新規の 12 遺伝子座を詳細に調べた結果、その多くは未注目のものであったが、血管平滑筋や腎臓の働きに関係する遺伝子群が血圧(高血圧)の個人差の原因となっている可能性を強めた。

アジアにおける日本人の遺伝的地域差

竹内史比古(国際医療研究センター)、勝谷友宏(阪大)、
木村亮介(琉球大)、並河徹(島根大)、
大久保孝義(帝京大)、田原康玄(京大)、
山本健(久留米大)、横田充弘(愛知学院大)、
SAW, Woei-Yuh、TEO, Yik-Ying(シンガポール大)、
加藤規弘(国際医療研究センター)

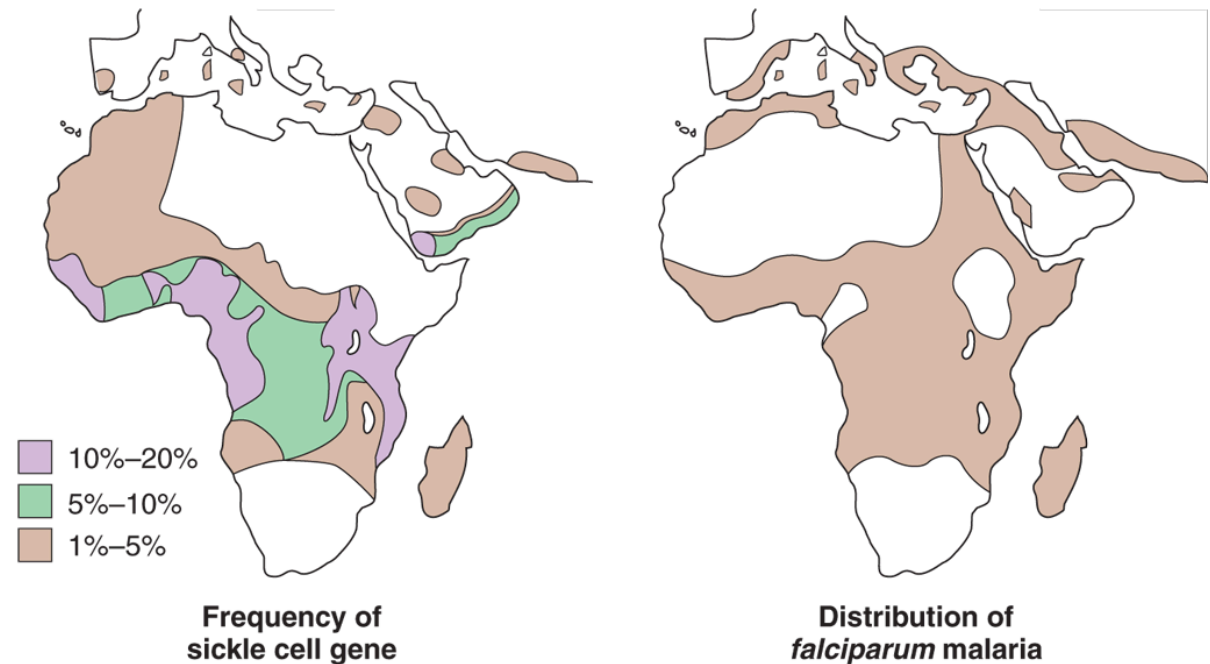
SNPアレル頻度の地域差: 例1

- *ALDH2*
rs671
Glu504Lys
- お酒が飲めなくなるアレル
- 12q24は自然選択を受けているが、原因変異・表現型は不明



SNPアレル頻度の地域差: 例2

- 鎌状赤血球貧血
- β^S アレル=貧血=マラリア耐性は、マラリア発生地域で高頻度
- 平衡選択を受けている



結果概要

- 全ゲノムSNPs情報のみから、日本人をクラスターに分離できる
- そのクラスターは国内の地域と対応する
- 疾患関連解析ではクラスターの影響の有無を適宜確認する必要がある
- 日本国内のクラスターごとに、由来するアジア大陸の地域が若干異なる

背景(日本人の遺伝的地域差)

- 我々はこれまで、ゲノムワイド関連解析により、罹患者と健常者でSNPsのアリル頻度を比較し、疾患感受性遺伝子を多数同定してきた
- 地域間の遺伝的差異の、疾患関連解析への影響の有無は集団の層別化として注目されてきた
- 先行研究では、沖縄と本土はクラスター分けできたが、本土内ではクラスター分けできていない
[Yamaguchi-Kabata et al. (2008) AJHG 83:445]
- 考古学的には、日本人は、縄文人と弥生人が混合して形成されたと推定されている

背景（遺伝的差異の解析法）

- 連鎖していないSNPsを用いる統計解析が、これまで主流だった
 - 各祖先集団におけるアリル頻度をMCMC推定
 - STRUCTURE [Pritchard et al. (2000) Genetics 155:945]
 - SNP一致度によるサンプル間距離の主成分分析
 - EIGENSOFT [Reich et al. (2008) Nat Genet 40:491]
- （連鎖したSNPsを含む）染色体断片のハプロタイプの一致度を用いた統計解析が、最近提案された
 - fineSTRUCTURE [Lawson et al. (2012) PLoS Genet 8:e1002453]
 - 英国内の地域が遺伝的クラスターと綺麗に対応する [Leslie et al. (2015) Nature 519:309]

方法(データセット)

- データセット1
 - 日本の8集団1600名。花巻、東京、北名古屋、尼崎、島根、愛媛、福岡、沖縄から200名ずつ。108,611 SNPs
- データセット2
 - 島根県内の詳細な地域差を検討するために、島根のみ428名に拡大
- データセット3
 - データセット1にAsian Diversity Panel, HapMap3, HGDPのアジア人を追加。34集団3928名。30,579 SNPs
- データセット4
 - データセット1に漢民族(CHB, CHD, CHS)を264名追加。106,266 SNPs

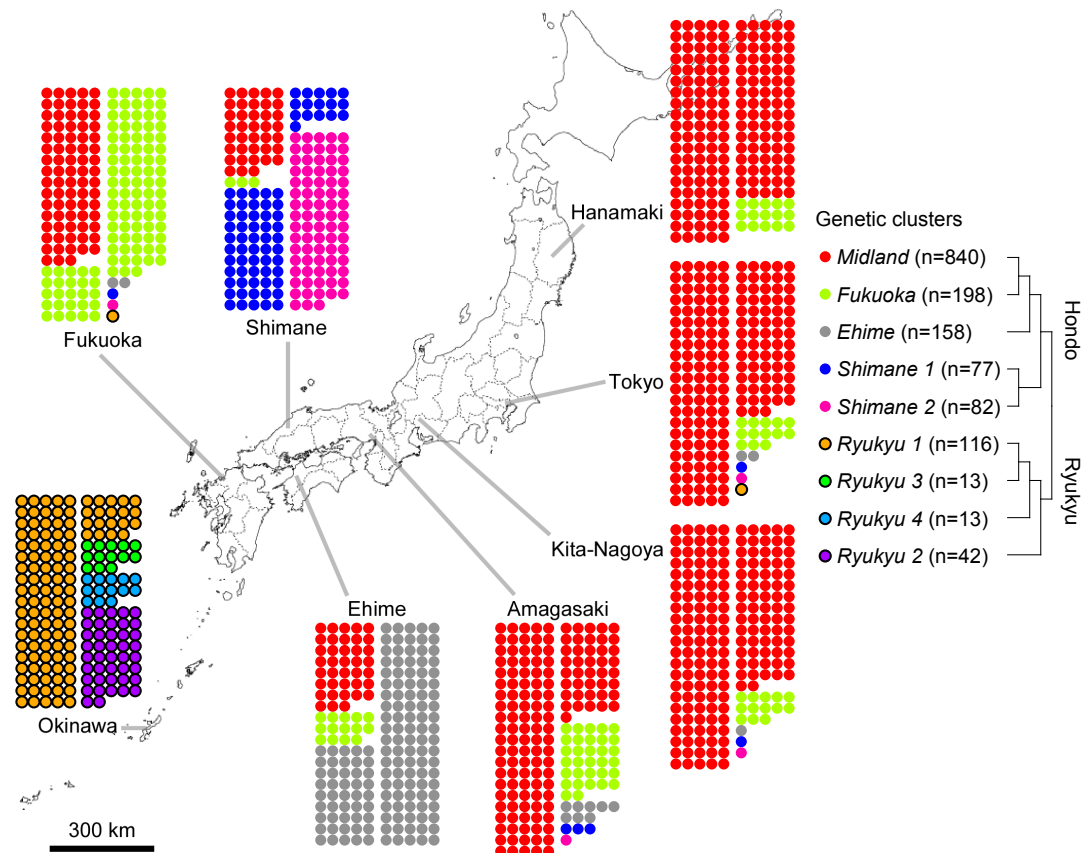
方法（クラスタリング）

- SNPs遺伝情報によるクラスタリング
 - Illumina, Affymetrix アレイにより全ゲノムSNPsタイピング
 - SHAPEITによりphase
 - サンプル対ごとに、共有するハプロタイプ断片の数を表す coancestry matrixをChromoPainterにより計算
 - coancestry matrixに合う遺伝的クラスタリングをfineSTRUCTUREにより推定

日本人の遺传的地域差

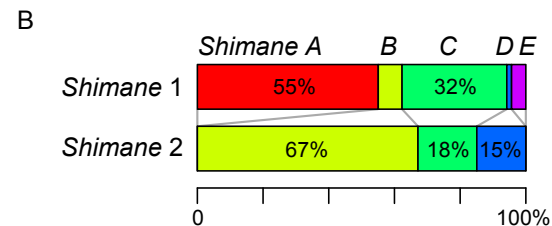
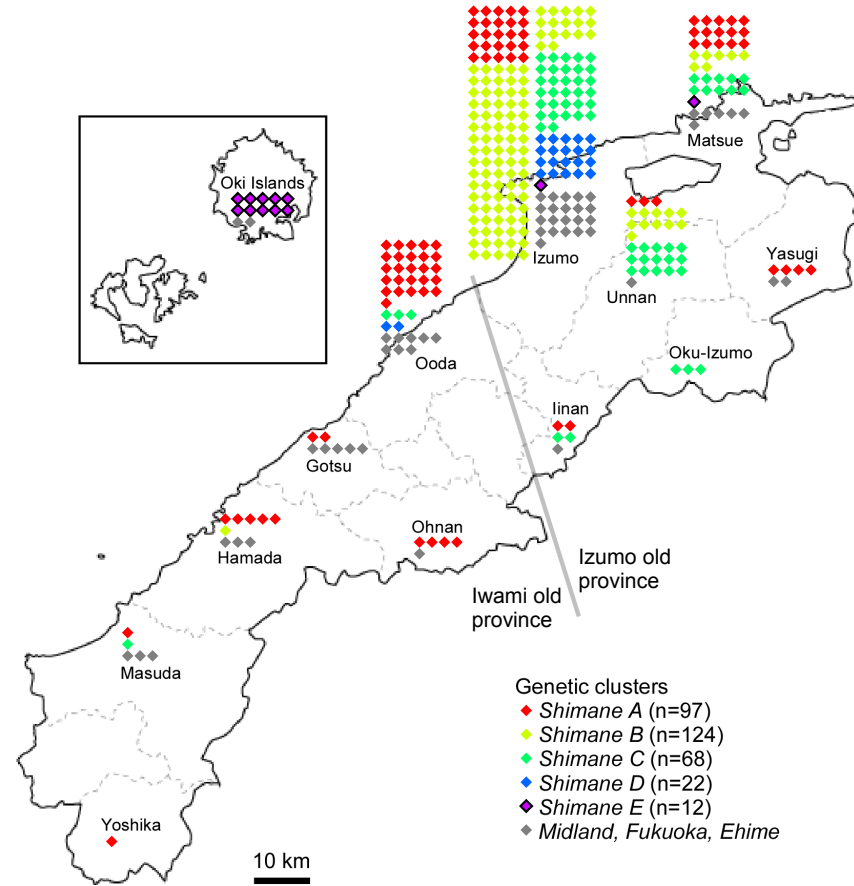
日本人の遺伝的クラスター

- 遺伝情報のみに基づき、日本人は9個のクラスターに分離できた
- クラスターは地域(地理的分布)と対応している



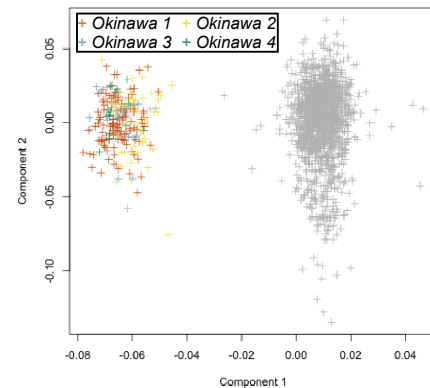
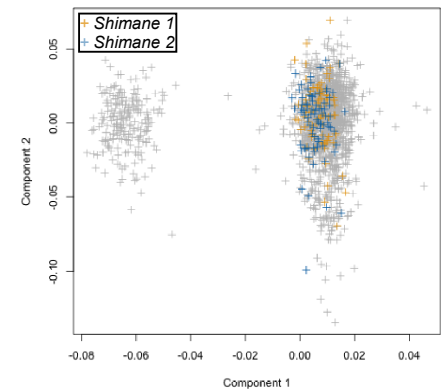
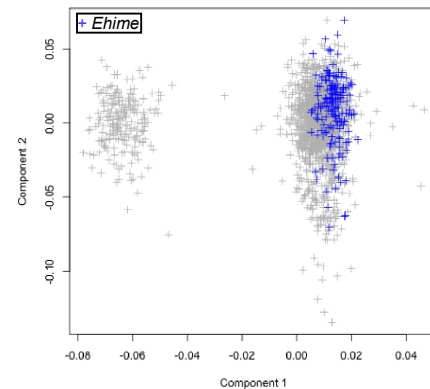
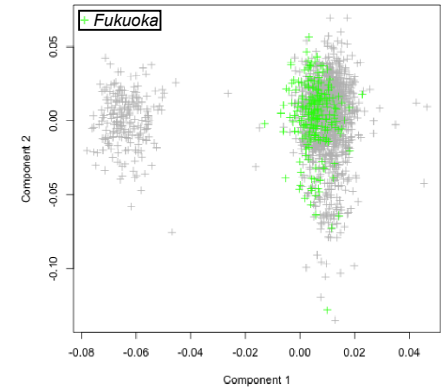
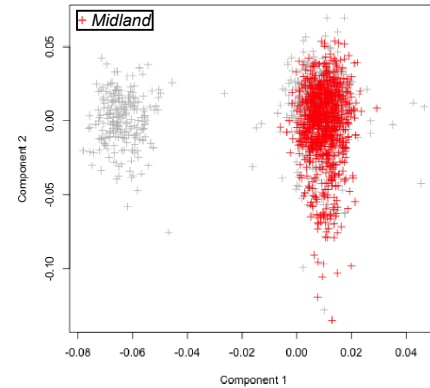
島根の詳細なクラスター

- Subclusterに分離した島根のサンプルを428名に増やして再解析すると、より詳細な5クラスターに分けられる
- 各クラスターが島根県内の地域に対応する



従来法との比較

- 先行研究[Yamaguchi-Kabata et al. (2008) AJHG 83:445]で用いられていたEIGENSOFTでは、沖縄と本土のみがクラスター分けされる
- 本土内の各地域は、分布がずれるものの、分離しない



遺伝的地域差と疾患ゲノム解析

- 各人の遺伝的距離(つまり2人の近さ)はIdentity by state (IBS)一致度で測定できる
 - $[IBS=2 \text{のSNPs数} + 0.5 * IBS=1 \text{のSNPs数}] / \text{全SNPs数}$
 - 沖縄クラスターは本土クラスターから、僅かだが明確に遠い
- クラスター集団間の距離はgenomic controlの λ で測定できる
 - ゲノムワイド関連解析(GWAS)での有意性の膨張の目安に使われる
 - 本土クラスターの違いでもGWASで問題になりうるレベル($\lambda > 1.1$)

→ GWASの被験者選出では地域差に注意する必要あり

SNP IBS一致度。mean (SD)

	Midland	Fukuoka	Ehime	Shimane 1-2	Okinawa 1-4
Midland	0.7372 (0.0016)	0.7370 (0.0016)	0.7370 (0.0016)	0.7372 (0.0016)	0.7358 (0.0016)
Fukuoka		0.7369 (0.0016)	0.7369 (0.0016)	0.7370 (0.0016)	0.7359 (0.0016)
Ehime			0.7370 (0.0016)	0.7370 (0.0016)	0.7355 (0.0016)
Shimane 1-2				0.7374 (0.0017)	0.7360 (0.0016)
Okinawa 1-4					0.7390 (0.0016)

Genomic controlの λ

	Fukuoka	Ehime	Shimane 1-2	Okinawa 1-4
Midland	1.05	1.12	1.13	4.22
Fukuoka		1.10	1.10	2.81
Ehime			1.15	3.03
Shimane 1-2				2.78

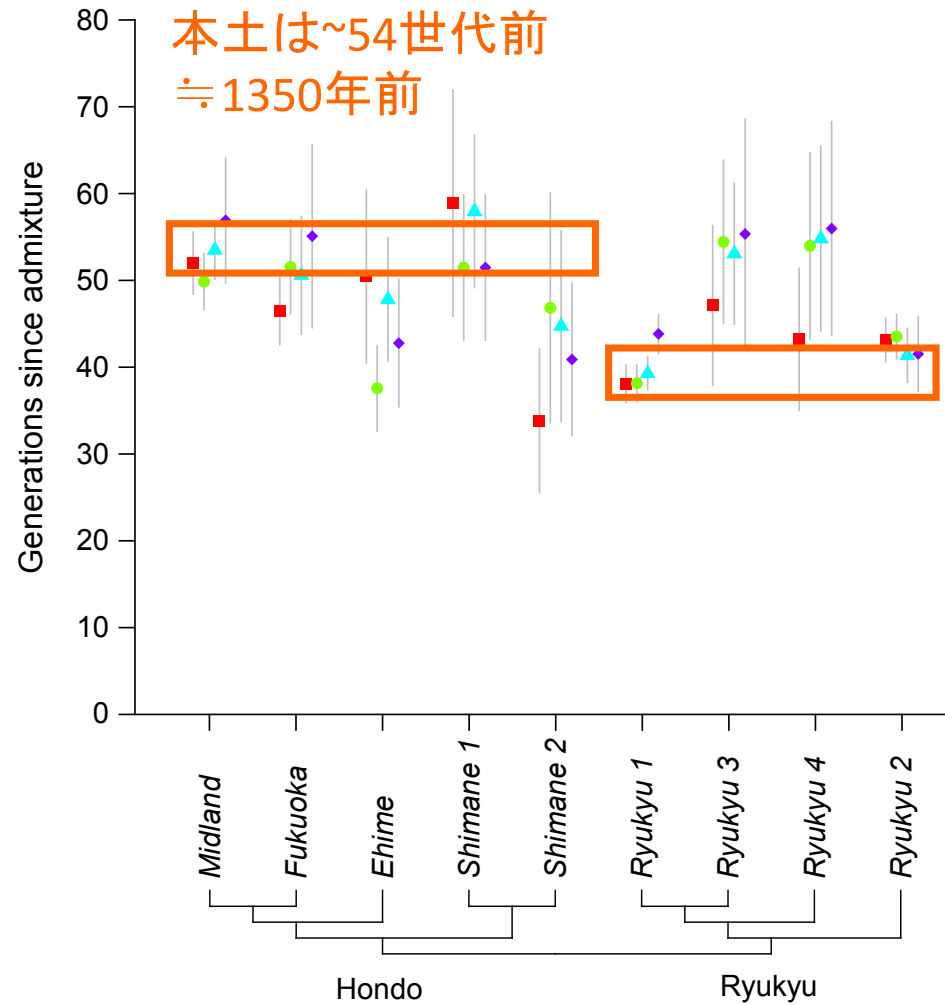
アジア人の中での遺伝的由来

方法（祖先型プロフィール）

- 日本人の祖先型プロフィール
 - 日本人と（日本人以外の）アジア人のcoancestry matrixを計算し、各クラスターについて祖先型プロフィールを計算
 - 祖先型プロフィールについて、各日本人クラスターのものを、アジア人クラスターの混合で近似する
 - 「祖先型」といっても、実際には現代人の混合比
- Admixture（混血）の年代推定
 - 集団AとBが混血して集団Cを形成。SNP x, y の頻度がA, Bと異なっていた
 - 集団Cにおいて、SNP x, y 間に連鎖不平衡が混血時に生じるが、世代を経て弱まる
 - 連鎖不平衡から混血を検定し、年代を推定
[Loh et al. (2013) Genetics 193:1233]

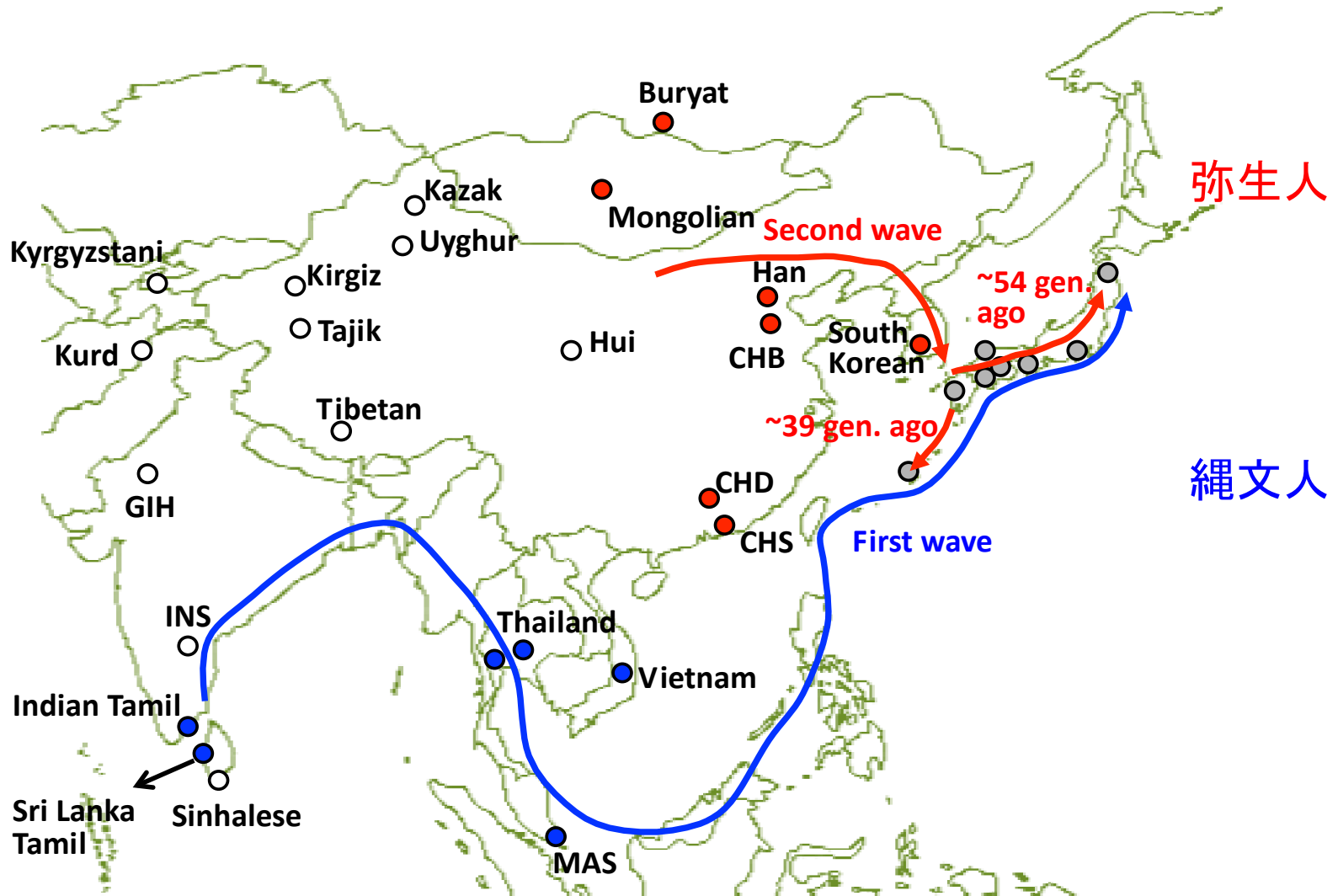
日本人における混血

- C282 (Korean)
- C284 (Han Chinese, North)
- ▲ C291 (Han Chinese, South)
- ◆ C280 (Southeast Asian)



日本人における混血

○ Yakut

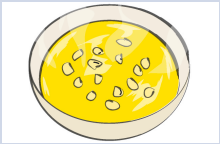


結果概要

- 全ゲノムSNPs情報のみから、日本人をクラスターに分離できる
- そのクラスターは国内の地域と対応する →
- 疾患関連解析ではクラスターの影響の有無を適宜確認する必要がある
 - imputationには殆ど問題ないはず →
- 日本国内のクラスターごとに、由来するアジア大陸の地域が若干異なる

多数のSNPsから、遺伝的クラスターが推定できた

- 「SNPs」を「食事」で置き換えると...

	昼食		夕食
月曜	スパゲッティ 	コーンスープ 	
火曜			
水曜			
木曜			


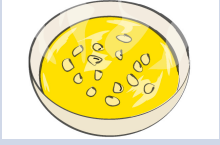

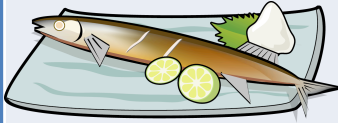







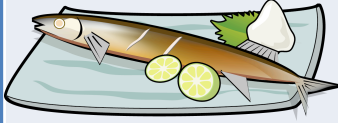



多数のSNPsから、遺伝的クラスターが推定できた

- 「SNPs」を「食事」で置き換えると...

	昼食	夕食
月曜	スパゲッティ 	コーンスープ 
火曜	サンマ塩焼き 	味噌汁 
水曜	ハンバーガー 	ミネストローネ 
木曜	サンマ塩焼き 	味噌汁 

多数のSNPsから、遺伝的クラスターが推定できた


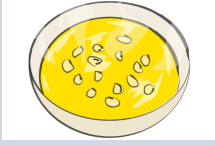




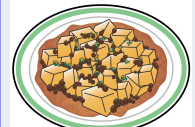








- 「SNPs」を「食事」で置き換えると...

	昼食		夕食	
月曜	スパゲッティ 	コーンスープ 	牛タン焼き 	テールスープ 
火曜	サンマ塩焼き 	味噌汁 	麻婆豆腐 	卵スープ 
水曜	ハンバーガー 	ミネストローネ 	牛タン焼き 	テールスープ 
木曜	サンマ塩焼き 	味噌汁 	鮭塩焼き 	味噌汁 



食事の例でのimputation

-

	既知	未知	既知	未知
	昼食		夕食	
月曜	スパゲッティ 	コーンスープ 	牛タン焼き 	テールスープ 
火曜	サンマ塩焼き 	味噌汁 	麻婆豆腐 	卵スープ 
水曜	ハンバーガー 	ミネストローネ 	牛タン焼き 	テールスープ 
木曜	サンマ塩焼き 	味噌汁 	鮭塩焼き 	味噌汁 