2023年8月25-27日
遺伝統計学・夏の学校　講義実習資料

GenomeDataAnalysis4

大阪大学大学院医学系研究科 遺伝統計学
東京大学大学院医学系研究科 遺伝情報学
理化学研究所生命医科学研究センター　システム遺伝学チーム

http://www.sg.med.osaka-u.ac.jp/index.html

講義の概要

GenomeDataAnalysis4

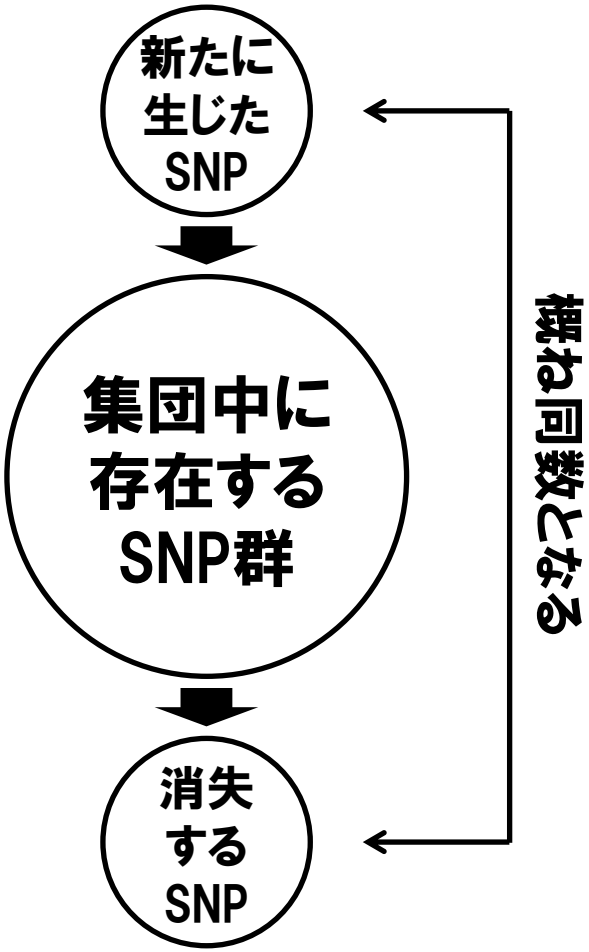
① 選択圧と適応進化

② 全ゲノムシークエンスに基づく日本人の適応進化

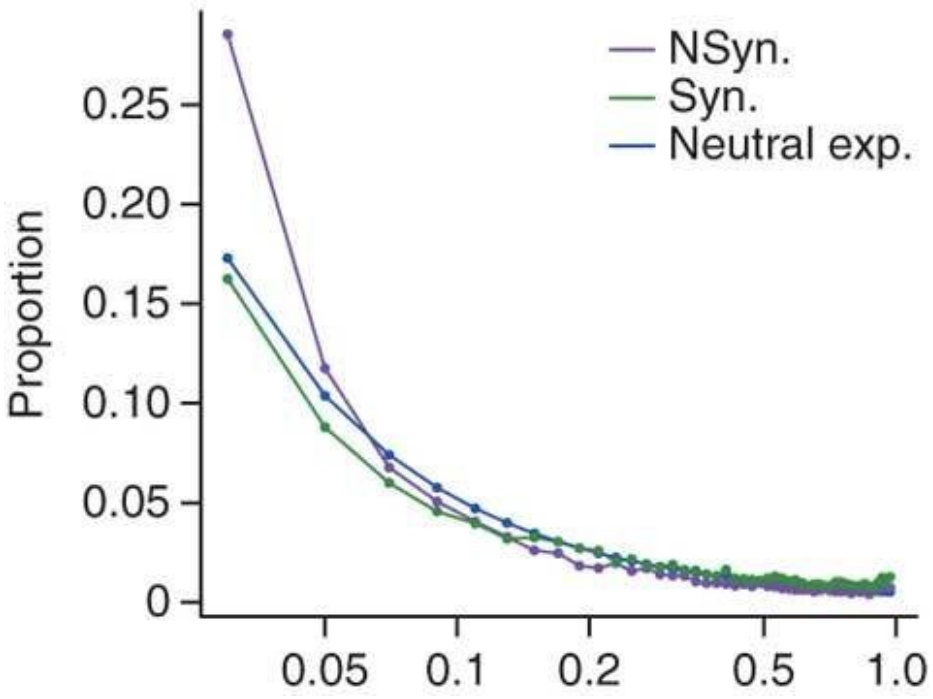
③ selscanを使った選択圧解析

本講義資料は、Windows PC上で
C:¥SummerSchoolにフォルダを配置すること
を想定しています。

① 選択圧と適応進化



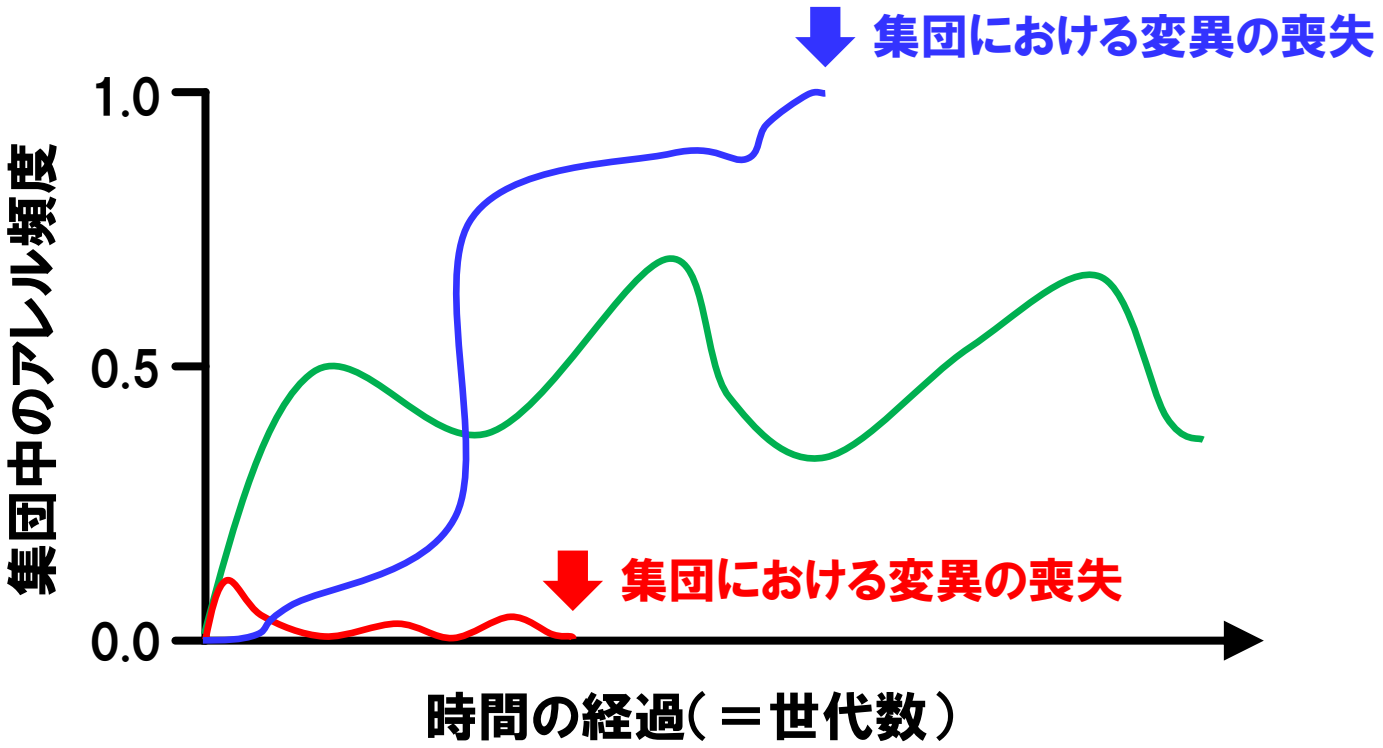
デンマーク集団における
SNPアレル頻度分布



（Li et al. ***Nat Genet*** 2010）

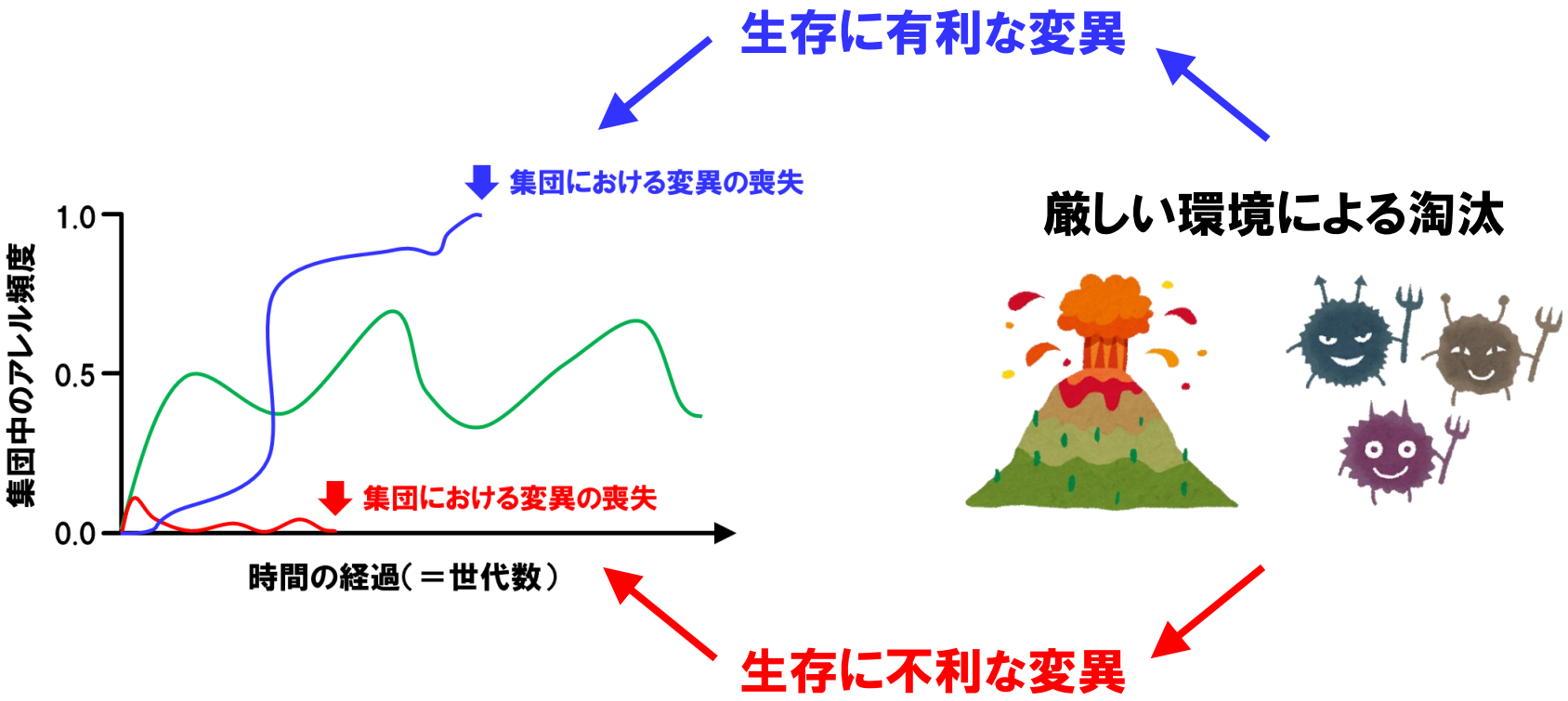
- 一定数のSNPが突然変異により生じ、また子孫に受け継がれずに消失することにより、集団中に存在するSNPの数は概ね保たれています。
- アレル頻度の低いSNPほど多く存在する傾向が知られています。

① 選択圧と適応進化



- 突然変異により集団中に新たに発生した変異は、子孫へと受け継がれていく過程で、集団中でのアレル頻度が経時的に変化していきます。
- 一部の変異は集団中で拡散できずに消失し、一部の変異は集団中で拡散することでアレル頻度を増やしていきます。集団を構成する全個体に変異が拡散すると多型性が消失し、変異の状態を喪失します。

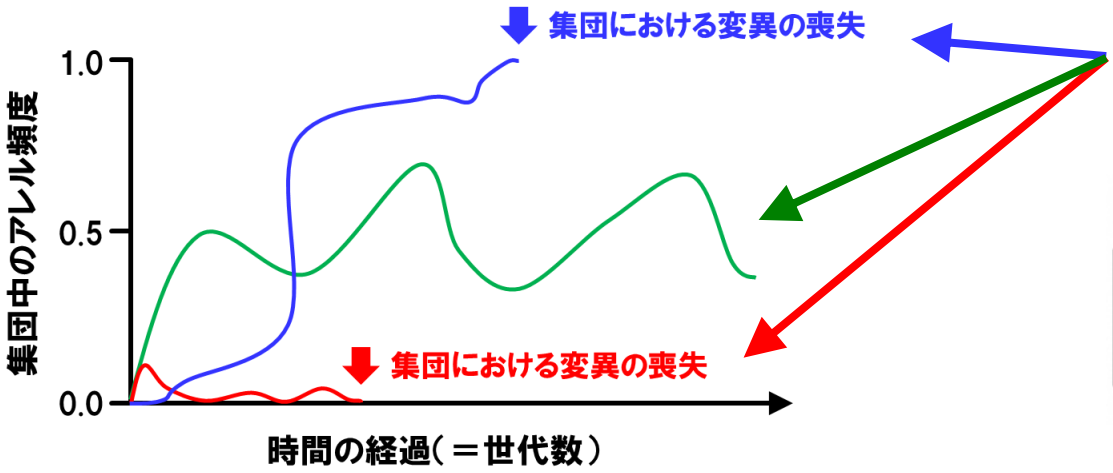
① 選択圧と適応進化



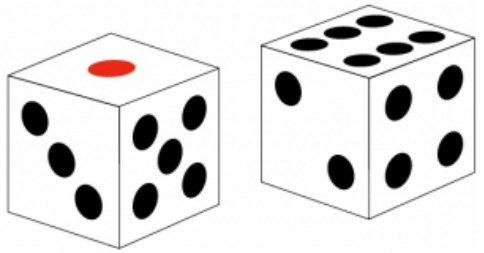
・変異が個体の生存に関わる表現型に影響を及ぼす場合、変異を有する個体が生存に有利（もしくは不利）となるため、集団中における頻度が速いスピードで変化します。これを**選択**と呼びます。

・外的環境への対応過程で変異への**選択圧**（selection pressure）が生じ、遺伝的多様性を生じるのが、**自然選択説**（natural selection）になります。

① 選択圧と適応進化



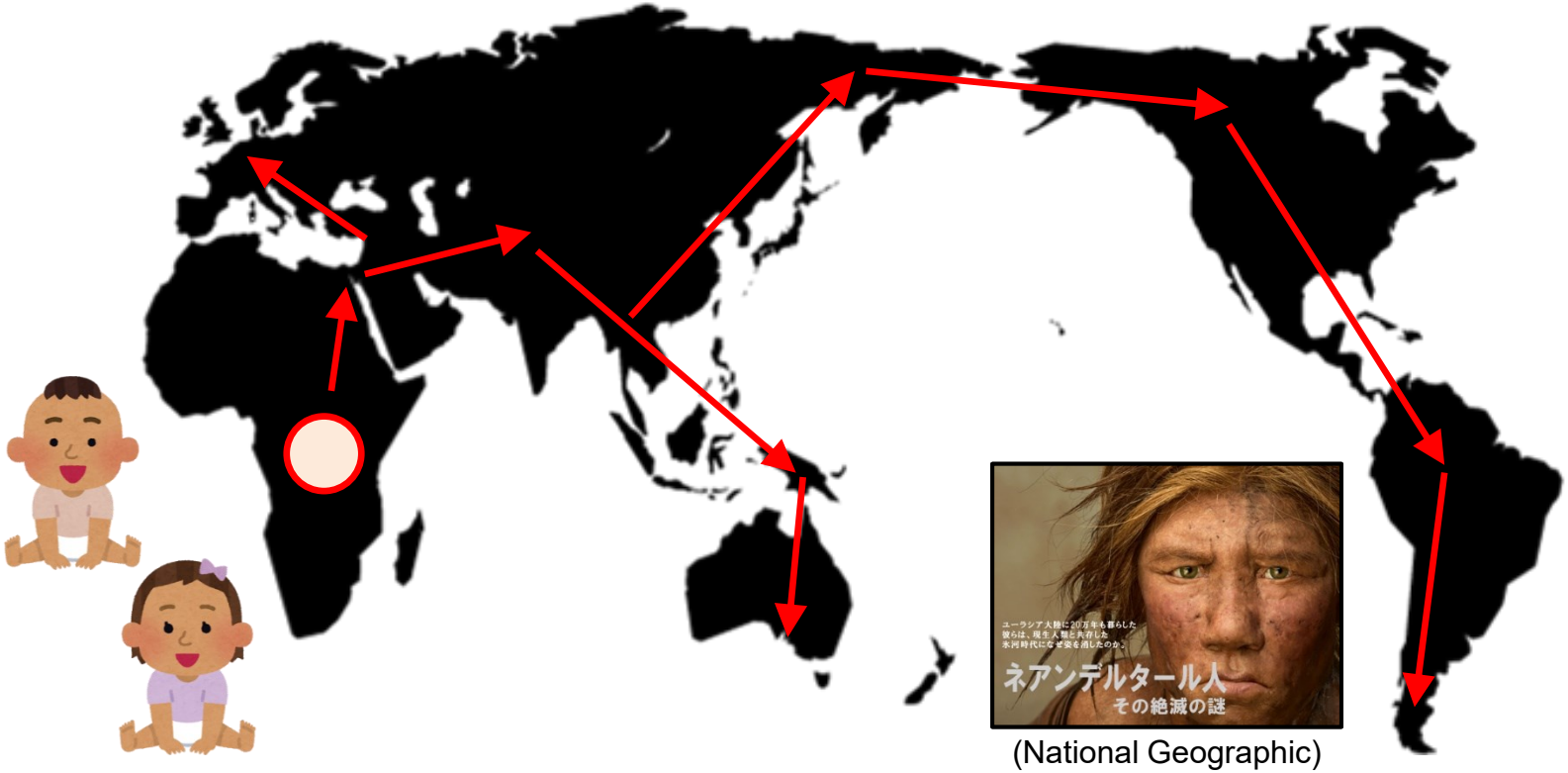
無作為抽出の結果



- 一方で、変異のほとんどは生存に有利でも不利でもなく、集団中での頻度変化は**無作為抽出**（＝遺伝的浮動、genetic drift）によるという考え方もあり、**中立進化説**（neutral theory of molecular evolution）と呼ばれています。
- 中立進化説は日本人の木村資生博士により提唱され、「**木村の中立説**」として有名です。
- 自然選択説と中立進化説の間には長い論争がありますが、現在ではどちらの現象も存在していると考えられています。

① 選択圧と適応進化

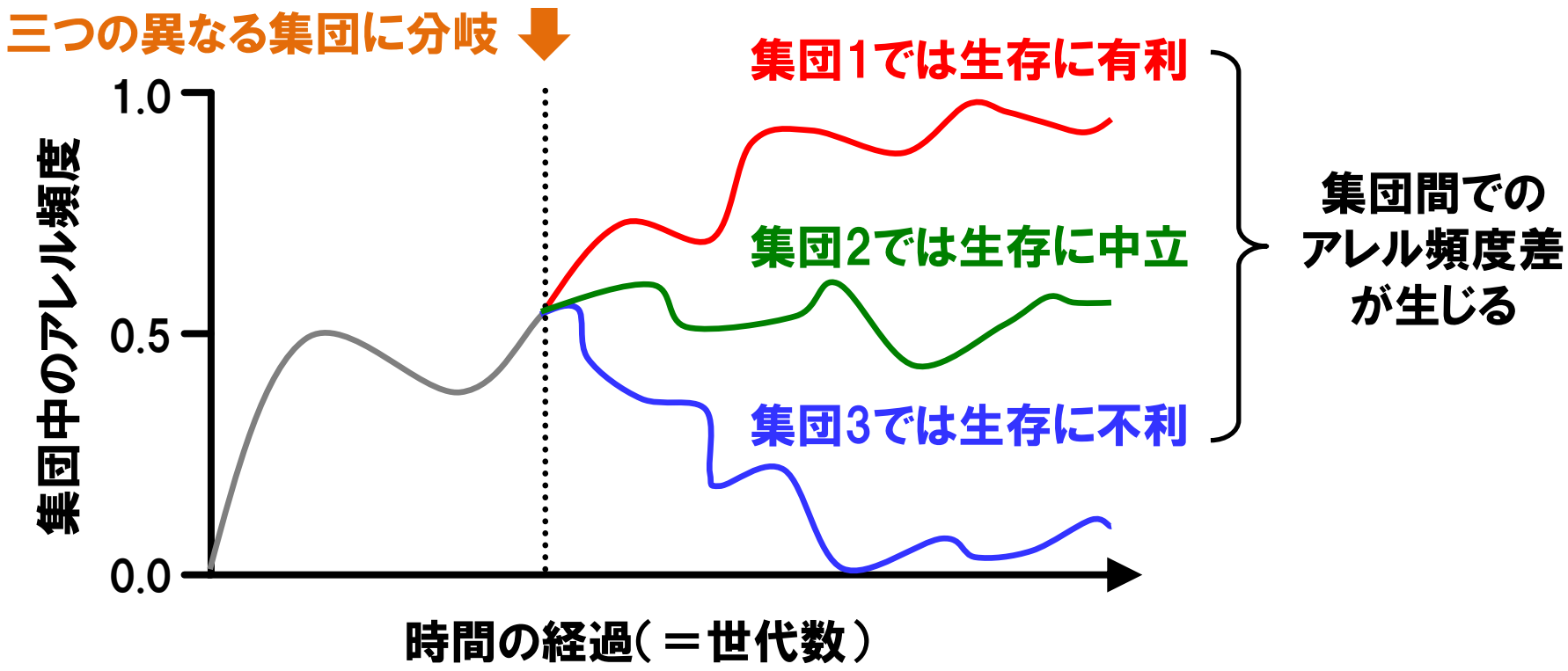
ホモ・サピエンスの移動の歴史



(National Geographic)

- ヒト（ホモ・サピエンス）は15－10万年前にアフリカで出現し、その後、数多くの集団に分岐しながら、世界中に広がったと考えられています。
- その過程では、他のホモ属との交雑もあったと考えられ、現生人類のヒトゲノムの数%程度は、ネアンデルタール人由来と考えられています。

① 選択圧と適応進化

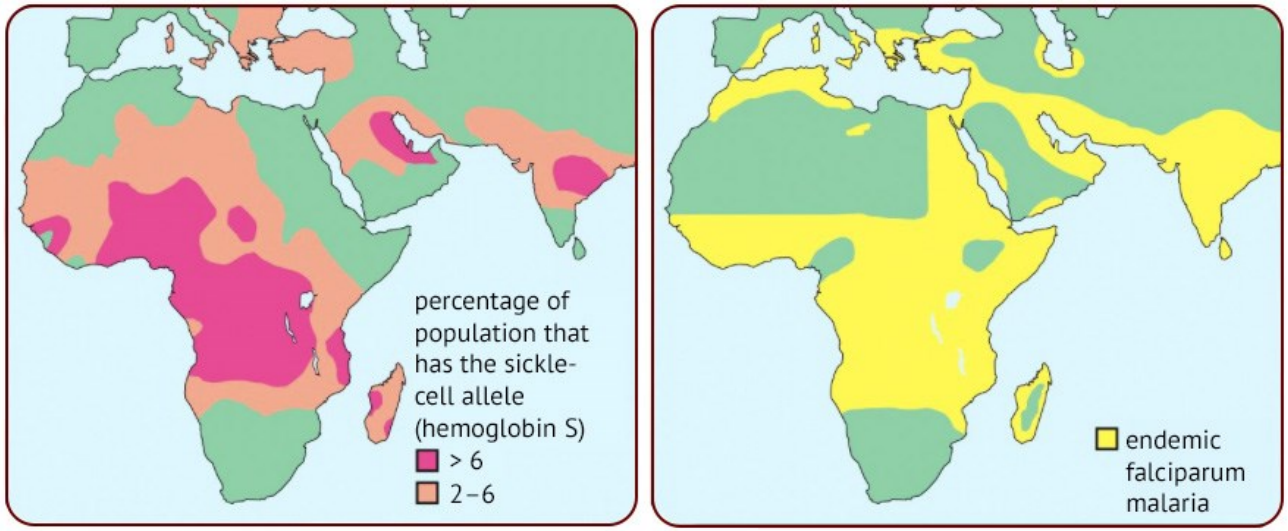


・分岐したヒト集団が、異なる地理的環境で異なる選択圧にさらされると、集団間での遺伝子変異のアレル頻度差が生じます。

・集団間で著しいアレル頻度差を示した遺伝子変異に着目し、その遺伝子変異がどのような環境や表現型に対応しているか調べることで、その集団に特有の選択圧を知ることができます。

① 選択圧と適応進化

鎌状赤血球症の変異分布　　マラリア発生地域分布



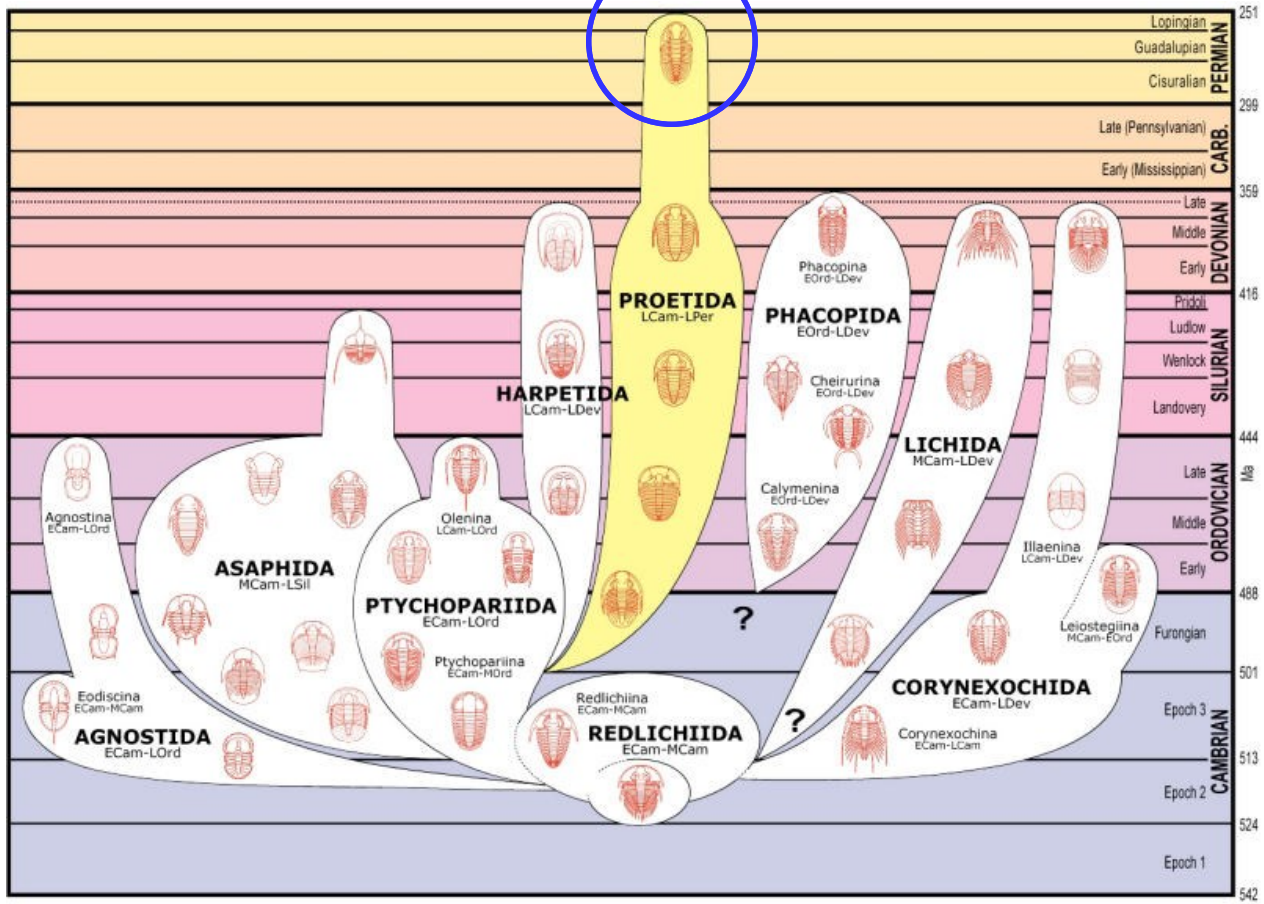
(https://www.philpoteducation.com)

- 各地域に特有の**選択圧**がはたらき、特定の表現型に関わる**遺伝子変異の頻度が集団特異的に変化**してきた例が複数確認されています。
- マラリア蔓延地域に住む集団では、ヘモグロビンβ鎖遺伝子の変異による**鎌状赤血球症**が高い発生率を示します。これは、**マラリア耐性**を獲得するため、遺伝子変異が急速に広まったことに由来します。
- 環境に応じて生物がその性質を変える現象を、**適応進化**といいます。

① 選択圧と適応進化

最後まで生き残ったのは、一番
シンプルな形をした三葉虫だった。

ペルム紀

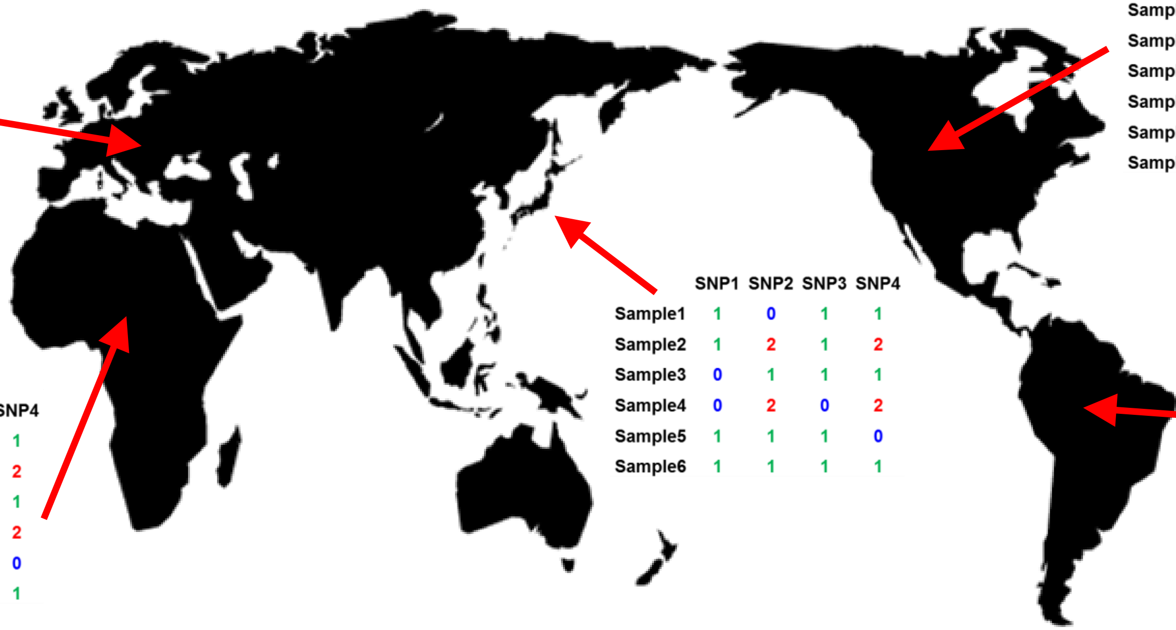


カンブリア紀

- 適応進化は性質が変化する現象を指し、「優れた生き物に変わる」という意味はありません。限定的な環境に適応しすぎたため、結果として絶滅してしまった生き物は数多く知られています。

(http://www.trilobites.info/)

① 選択圧と適応進化



	SNP1	SNP2	SNP3	SNP4
Sample1	1	0	1	1
Sample2	1	2	1	2
Sample3	0	1	1	1
Sample4	0	2	0	2
Sample5	1	1	1	0
Sample6	1	1	1	1

	SNP1	SNP2	SNP3	SNP4
Sample1	1	0	1	1
Sample2	1	2	1	2
Sample3	0	1	1	1
Sample4	0	2	0	2
Sample5	1	1	1	0
Sample6	1	1	1	1

	SNP1	SNP2	SNP3	SNP4
Sample1	1	0	1	1
Sample2	1	2	1	2
Sample3	0	1	1	1
Sample4	0	2	0	2
Sample5	1	1	1	0
Sample6	1	1	1	1

	SNP1	SNP2	SNP3	SNP4
Sample1	1	0	1	1
Sample2	1	2	1	2
Sample3	0	1	1	1
Sample4	0	2	0	2
Sample5	1	1	1	0
Sample6	1	1	1	1

	SNP1	SNP2	SNP3	SNP4
Sample1	1	0	1	1
Sample2	1	2	1	2
Sample3	0	1	1	1
Sample4	0	2	0	2
Sample5	1	1	1	0
Sample6	1	1	1	1

- ヒトゲノム配列の**どの遺伝子領域**が、**どの集団**で、**どの表現型**との関わりで選択圧を受けてきたか検討することが、**選択圧の解析**になります。
- 選択圧の解析は、ホモ・サピエンスの歴史や現代人の疾患の背景を理解する点からも、重要な研究テーマです。
- 複数集団を代表するサンプル群から得られたゲノムワイドな遺伝子変異ジェノタイプデータがあれば、その問いに答えることができます。

① 選択圧と適応進化

対象ゲノムデータによる選択圧解析手法の分類

	各ゲノム領域毎に個別に計算		異なる領域に跨がってヒトゲノム全体で計算
	単一の遺伝子変異が対象	複数の遺伝子変異が対象	複数の遺伝子変異が対象
単一の集団が対象	HWE	iHS SDS ASMC	*f*
複数の集団が対象	F_{ST}	XP-EHH	-

※代表的な解析手法だけを表に載せています。

- これまでに、数多くの選択圧の解析手法が開発されてきました。
- 下記特徴により、選択圧解析手法をおおまかに分類することができます。

①：単一の遺伝子変異が対象か、複数の遺伝子変異が対象か。

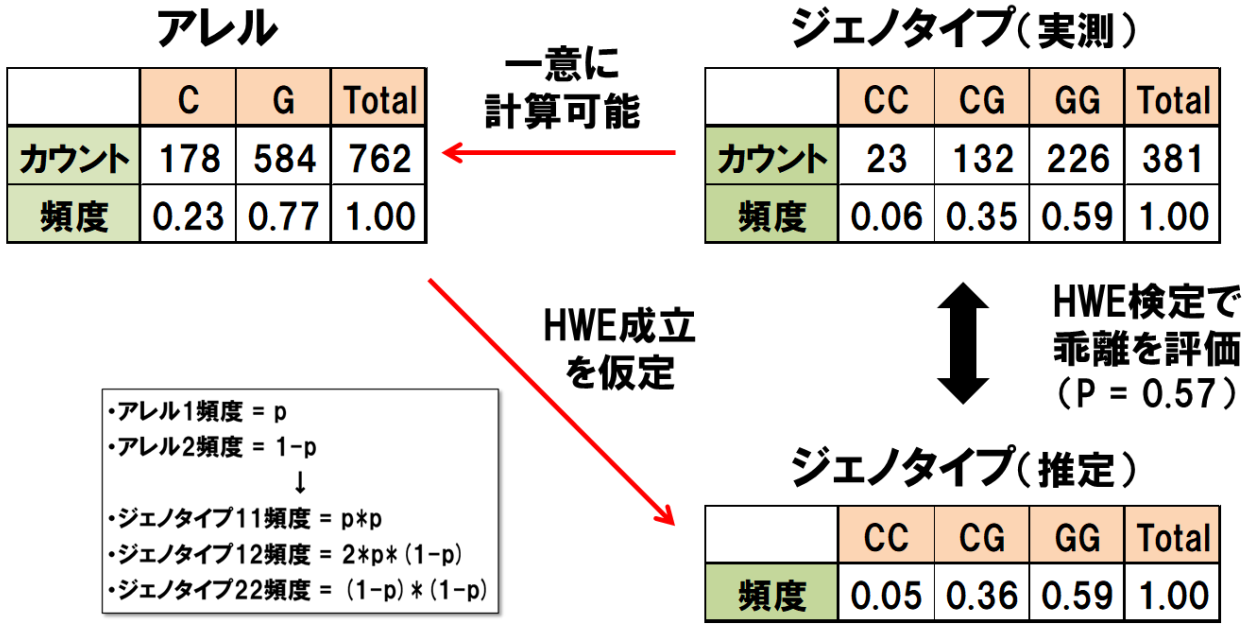
②：各ゲノム領域毎に個別の計算か、ヒトゲノム全体で計算か。

③：単一集団と複数集団の、どちらのゲノムデータが解析対象か。

(岡田　随象. *実験医学増刊 Vol.36* 2018)

① 選択圧と適応進化

選択圧解析手法①：HWE (Hardy-Weinberg equilibrium)

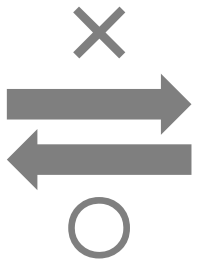


- **単一の多型**を対象に、**単一集団**で検証する解析手法。
- 集団内でのジェノタイプ頻度分布が、アレル頻度分布から理論的に推定される値と乖離しているかを検討する手法です。
- 集団内で有意な乖離が認められる場合、ジェノタイピングエラー以外の原因として、集団構造の階層化や**選択圧**の存在が示唆されます。

① 選択圧と適応進化

アレル

	C	G	Total
カウント	178	584	762
頻度	0.23	0.77	1.00



ジェノタイプ

	CC	CG	GG	Total
カウント	23	132	226	381
頻度	0.06	0.35	0.59	1.00

HWEが成立する条件

- 集団サイズが大きい
- 集団が均一である
- ランダム交配である
- その遺伝子座に自然選択がない
- その遺伝子座に突然変異がない

HWEの法則

- アレル1頻度 = p
- アレル2頻度 = 1−p

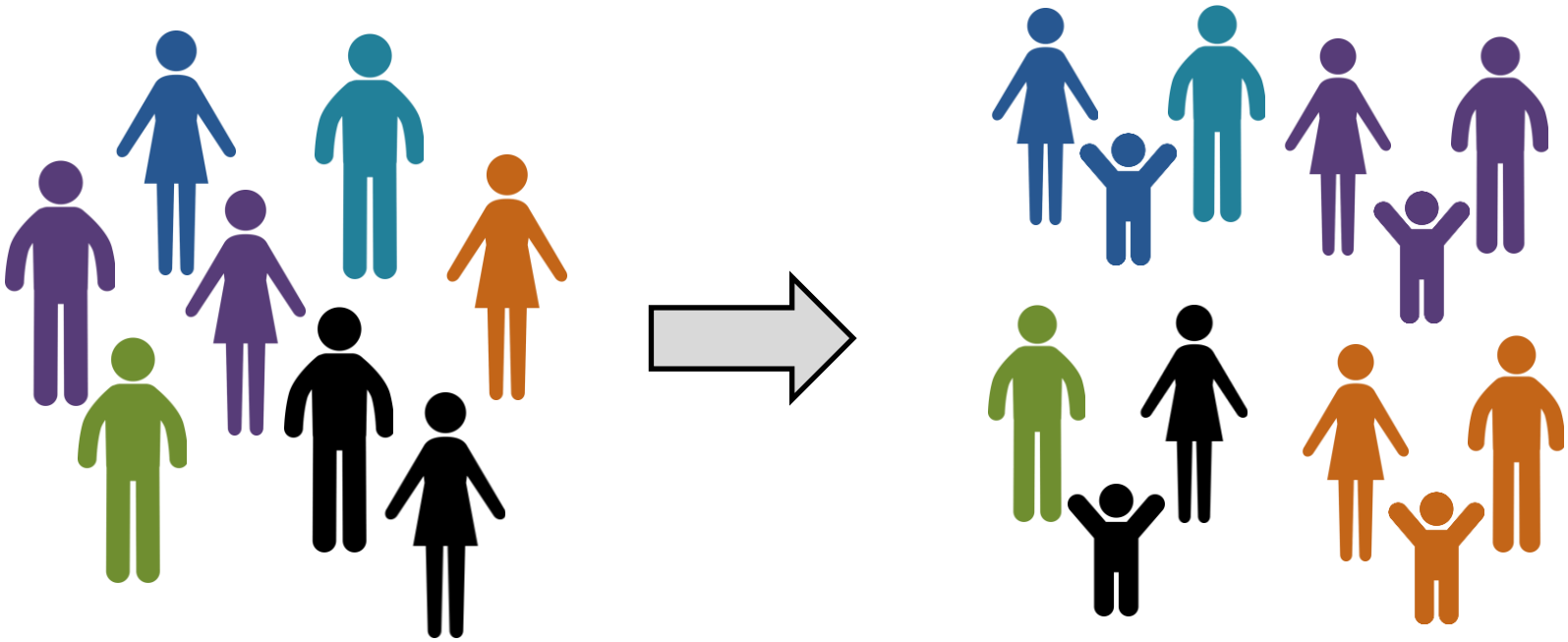
↓

- ジェノタイプ11頻度 = p*p
- ジェノタイプ12頻度 = 2*p*(1−p)
- ジェノタイプ22頻度 = (1−p)*(1−p)

- HWE平衡の成立条件の一つである「自然選択がない」に反する状況の同定を目的に、HWE検定を選択圧解析に用いる、とも解釈できます。
- 「ランダム交配」や「突然変異がない」に反する状況も知られています。

① 選択圧と適応進化

ヒトにおける同類交配と集団ゲノム情報の関わり



- **同類交配**（assortative mating; AM）：類似した表現型を持つ個体同士が任意で予想されるよりも高い確率でペアを形成する現象。
- 集団遺伝学においては、表現型の類似性の増加は**遺伝的類似性**につながり、表現型感受性遺伝子変異の集団中の不均衡分布をもたらす。
- 欧米人集団においては**身長や学歴**に対する同類交配が存在。

(Yengo L et al. ***Nat Hum Behav*** 2018, Yamamoto K et al. ***Nat Hum Behav*** 2023)

① 選択圧と適応進化

Infinite site model と biallelic mutation

Infinite site modelを提唱した
木村資生博士の論文(1969)

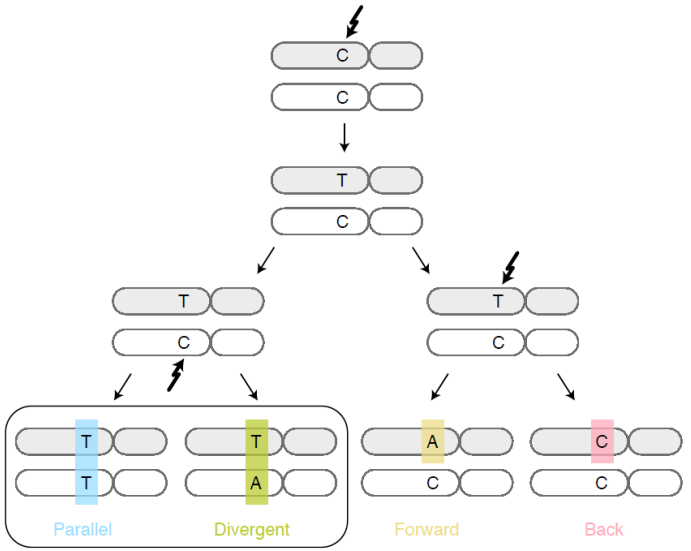
THE NUMBER OF HETEROZYGOUS NUCLEOTIDE SITES MAINTAINED IN A FINITE POPULATION DUE TO STEADY FLUX OF MUTATIONS[1]

MOTOO KIMURA

National Institute of Genetics, Mishima, Japan

Received September 10, 1968

がんゲノムにおけるbiallelic mutation生成過程



Biallelic mutation Monoallelic mutation

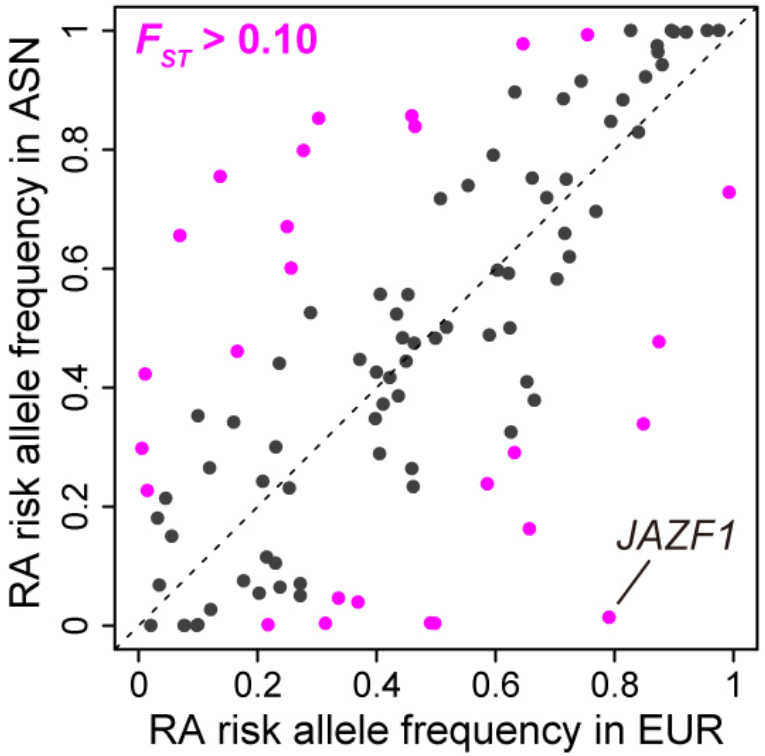
- Infinite site modelは、ゲノム全体に多数の塩基配列があり、突然変異率が低いため、同一塩基箇所には繰り返して突然変異が生じない、という仮説です。集団遺伝学解析の基礎理論となっています。
- 実際には、繰り返し生じた突然変異(=biallelic mutation)が一部存在していると考えられ、突然変異率の高いがんゲノムでは観測されています。

(Kimura M. ***Genetics*** 1969, Demeulemeester J et al. ***Nat Genet*** 2022)

① 選択圧と適応進化

選択圧解析手法②：F_{ST} (*F*-statistics)

関節リウマチ感受性SNPの集団間アレル頻度比較



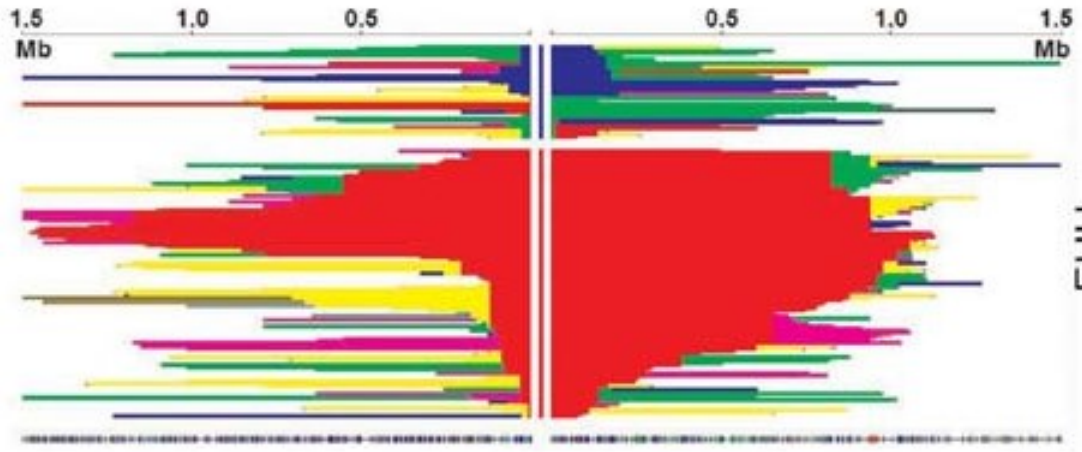
(Okada Y et al. ***Nature*** 2014)

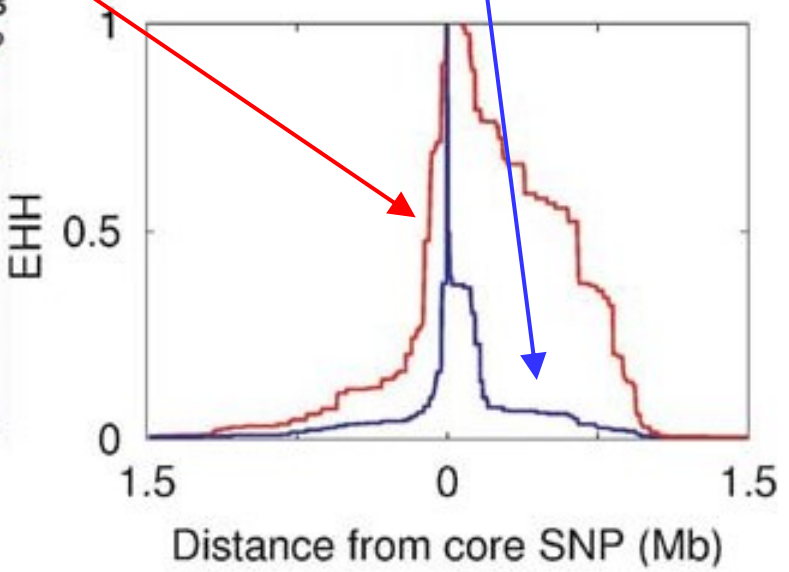
- **単一の多型**を対象に、**複数集団**で検証する解析手法。
- **アレル頻度が集団間でどの程度異なっているか**を定量化した指標です。
- 明確な基準値はありませんが、F_{ST}>0.10が基準値となることがあります。

① 選択圧と適応進化

選択圧解析手法③：iHS (integrated haplotype score)

片方のアレルが、もう片方のアレルより長いハプロタイプを形成している。





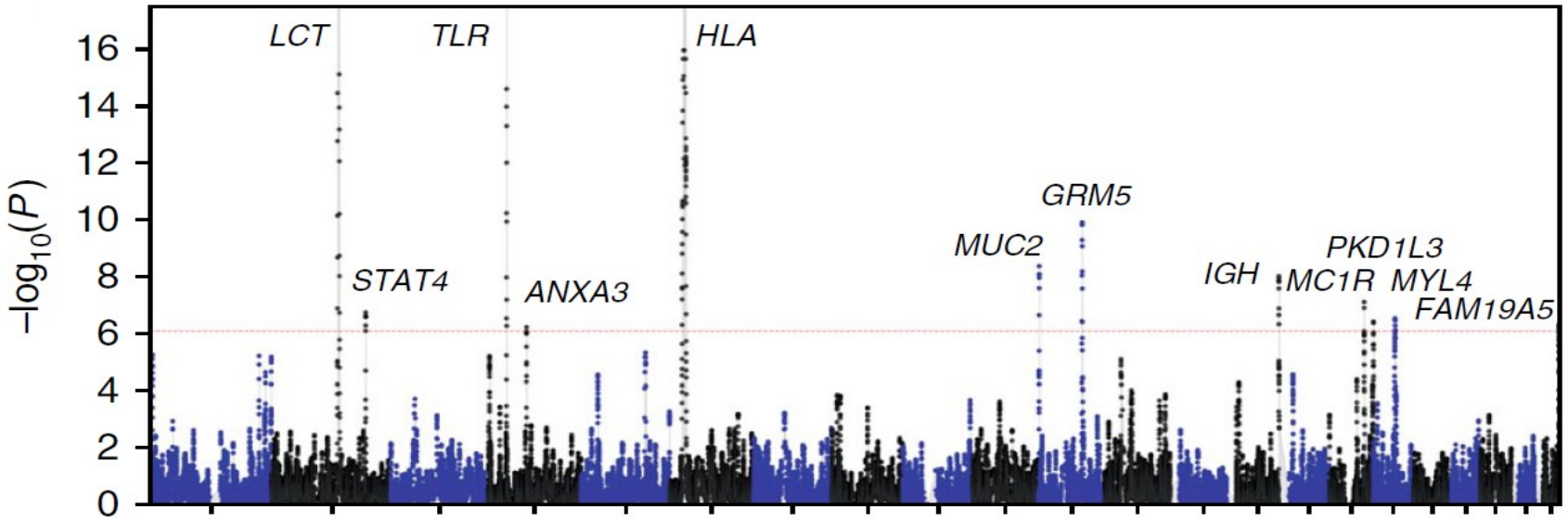
(Voight BF et al. ***PLoS Biol*** 2006)

- 特定領域内の複数の多型を対象に、単一集団で検証する解析手法。
- 特定の遺伝子変異に強い選択圧が働き、集団内でアレル頻度が急速に増えた場合、その近傍に組換えが生じる回数が相対的に小さくなるため、長いハプロタイプが保存されることに注目した手法です。

選択圧解析手法④：ASMC

(Ascertained Sequentially Markovian Coalescent)

UKバイオバンク欧米人集団における選択圧解析結果



(Palamara PF et al. ***Nat Genet*** 2018)

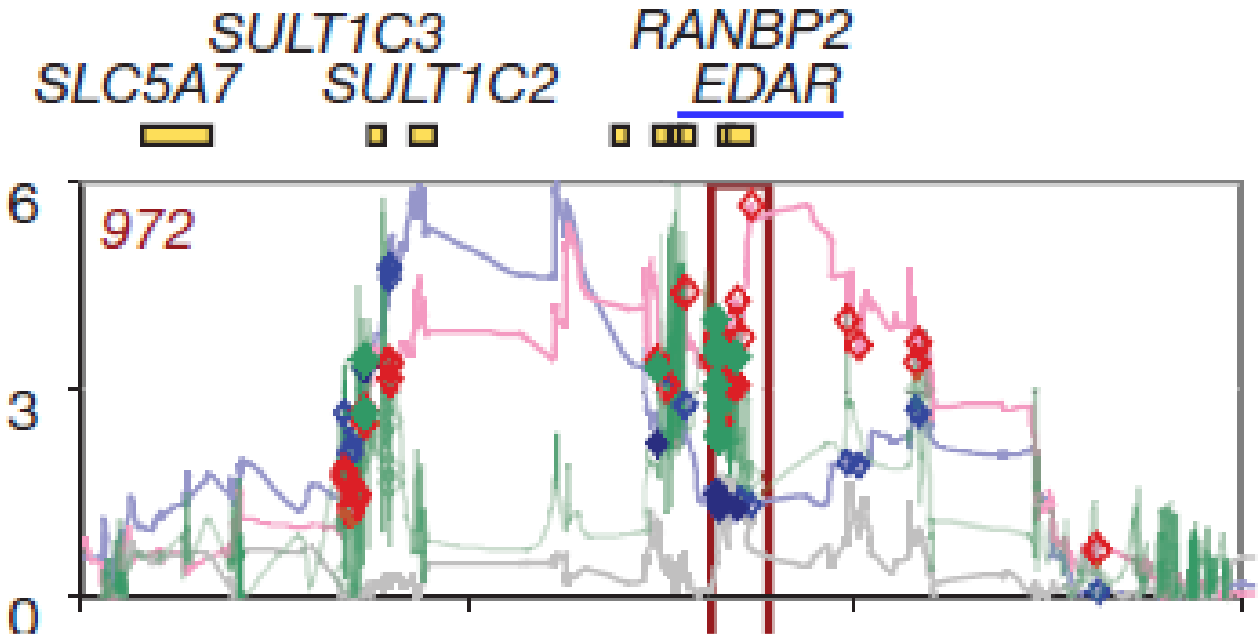
- **特定領域内の複数の多型を対象に、単一集団で検証する解析手法。**
- **近接する遺伝子変異の共通祖先を過去の集団に遡って推定する、合祖理論(coalescence theory)に基づき、選択圧を定量化します。**
- **数十万人規模のゲノム情報でも実施可能な選択圧解析手法です。**

① 選択圧と適応進化

選択圧解析手法⑤：XP-EHH

(cross-population extended haplotype homozygosity)

髪の毛の太さ遺伝子領域(*EDAR*)における選択圧

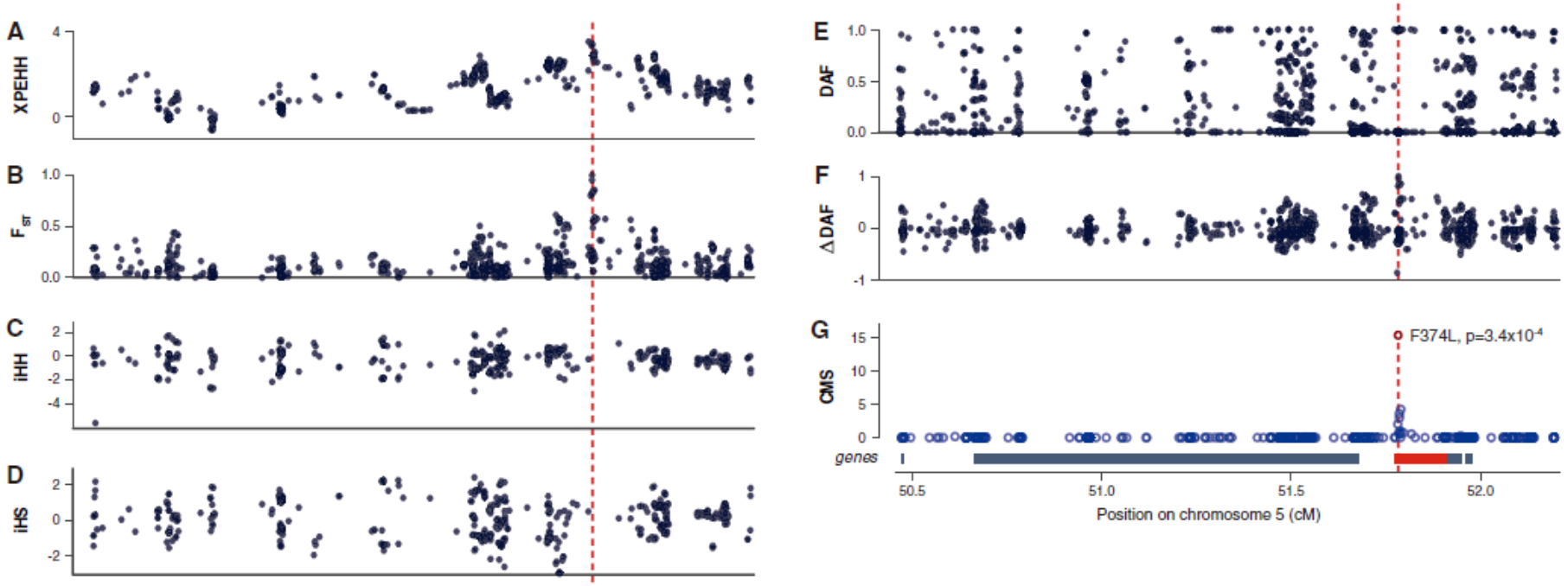


(Sabeti PC et al. ***Nature*** 2007)

- **特定領域内の複数の多型**を対象に、**複数集団**で検証する解析手法。
- iHSの考え方を複数集団間で比較する形で拡張した手法です。
- **選択圧の強い集団と弱い集団が存在する遺伝子領域**を同定できます。

① 選択圧と適応進化

選択圧解析手法⑥：CMS (composite of multiple signals)



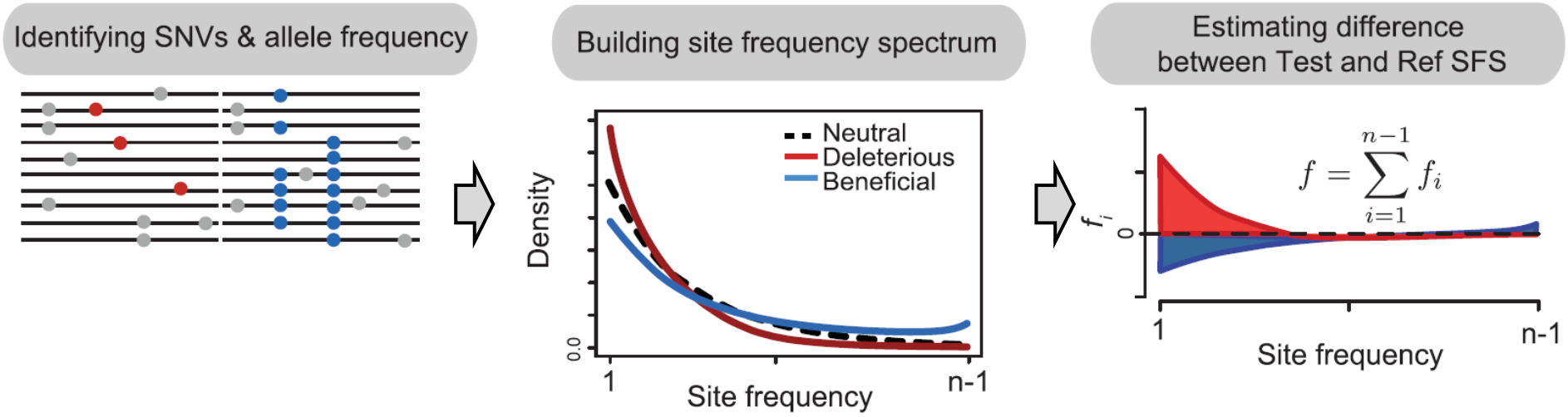
(Grossman SR et al. ***Science*** 2010)

- **複数の選択圧解析手法の結果を統合**する解析手法。
- F_{ST}、iHS、XP-EHH等の複数の解析結果をあわせることで、領域内で**真に選択圧を受けている機能性多型をピンポイントに同定**すること(fine-mapping)を目的とした手法です。

① 選択圧と適応進化

選択圧解析手法⑦：*f*

(fraction of sites under selection)



(Moon S et al. ***Genome Res*** 2016)

- ゲノム領域全体の多型を対象に、単一集団で検証する解析手法。
- 遺伝子変異のアレル頻度分布を、特定の遺伝子変異カテゴリー毎（例：アミノ酸変異SNP）に集計することで、どのカテゴリーに属する変異のどの程度の割合が選択圧を受けているのか、を定量化した指標です。

① 選択圧と適応進化

ゲノムデータ取得方法による選択圧解析手法の適用範囲

		ヒトゲノムデータの取得方法			
		個別SNPのデータ	SNPマイクロアレイ	エクソームシークエンス	全ゲノムシークエンス
実施可能な選択圧解析手法	HWE	○	○	○	○
	F_{ST}	○	○	○	○
	iHS	×	○	×	○
	ASMC	×	○	×	○
	SDS	×	×	×	○
	XP-EHH	×	○	×	○
	f	×	△	○	○
	CMS	×	○	×	○

- ゲノムデータの取得方法によって、適用可能な解析手法が異なります。
- ゲノムワイドな遺伝子変異情報が必要な場合はSNPマイクロアレイ、低頻度の遺伝子変異情報が必要な場合はエクソームシークエンス解析が必要になります。
- **全ゲノムシークエンスデータなら、全ての解析手法が適用可能**です。

(岡田　随象. *実験医学増刊 Vol.36* 2018)

講義の概要

GenomeDataAnalysis4

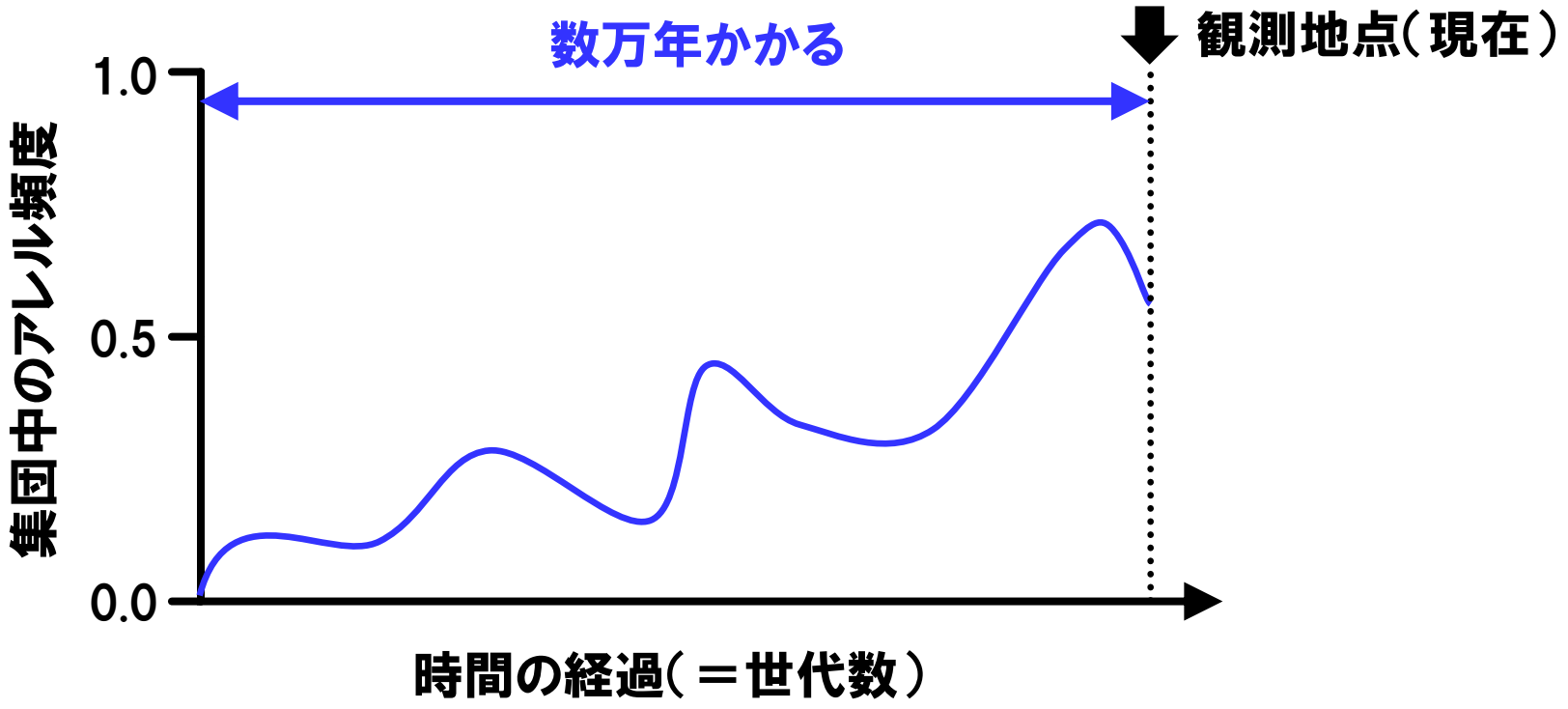
① 選択圧と適応進化

② 全ゲノムシークエンスに基づく日本人の適応進化

③ selscanを使った選択圧解析

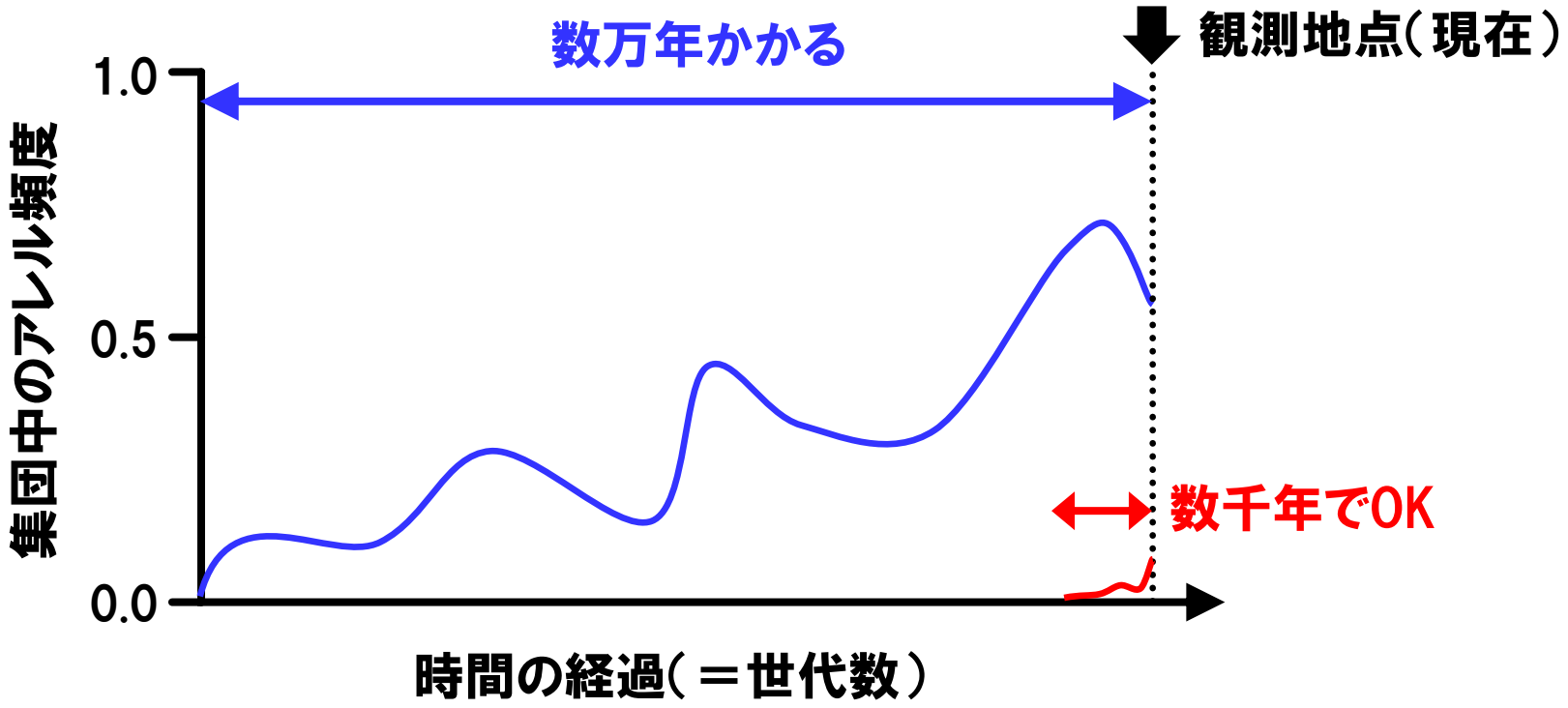
本講義資料は、Windows PC上で
C:¥SummerSchoolにフォルダを配置すること
を想定しています。

② 全ゲノムシークエンスに基づく適応進化の解明



- **観察された選択圧が何年位前に生じたイベントの結果**か、という時間軸を「**時相**」といいます。
- 集団中で高いアレル頻度を持つSNPを用いた場合、**数万年前という遠い過去**の選択圧しか解析できませんでした。SNPが高いアレル頻度を獲得し、集団間で異なる頻度を示すまでに、長い時間がかかるのです。

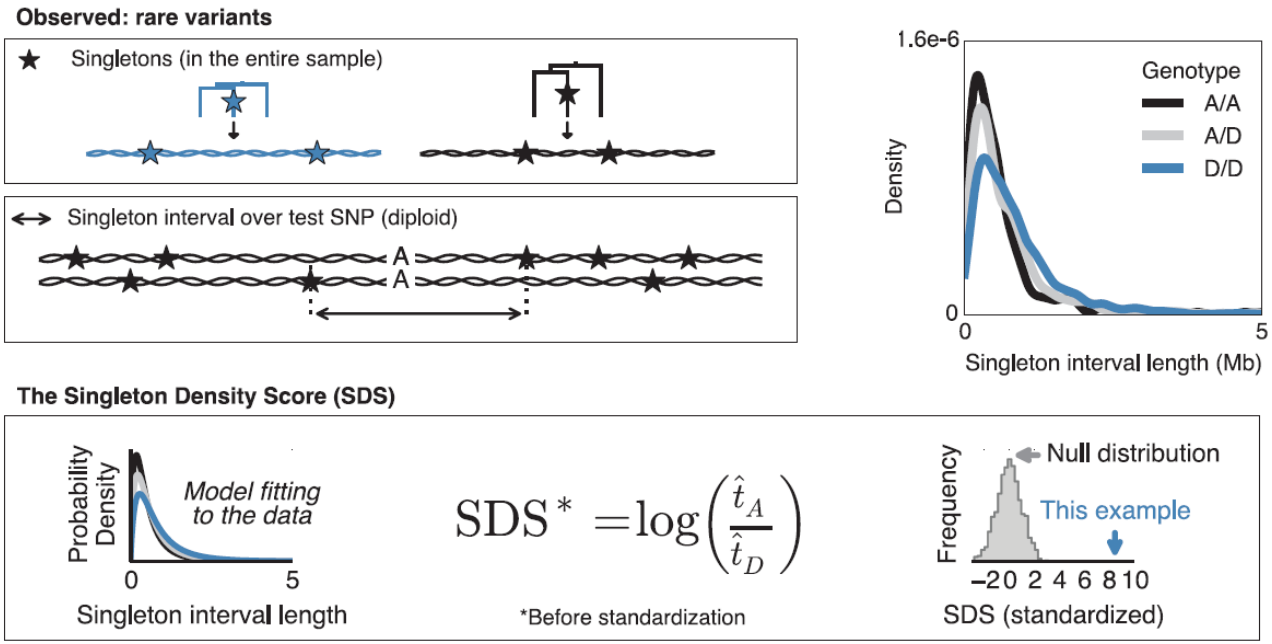
② 全ゲノムシークエンスに基づく適応進化の解明



- 一方、集団中で極めて低い頻度を持つ遺伝子変異に注目すると、数千年前という近い過去の時相における選択圧解析が可能になります。
- 極めて低頻度の変異の観測には、次世代シークエンス解析が必要です。
- 低頻度の変異の例として、対象集団内の1サンプルでしか観測されていない変異（=singleton）が挙げられます。

② 全ゲノムシークエンスに基づく適応進化の解明

選択圧解析手法⑦：SDS (singleton density score)



(Field Y et al. ***Science*** 2016)

- 全ゲノムシークエンスで観測されたsingletonの集団中での分布に着目することで、数千年前という近い過去の時相における選択圧解析を可能にした手法として、SDSが挙げられます。
- SDSの開発により、世界各地の集団がその地域に定住していく最中で生じた、最近かつ集団特異的な選択圧の検討が可能になりました。

② 全ゲノムシークエンスに基づく適応進化の解明

選択圧解析手法⑦：SDS (singleton density score)

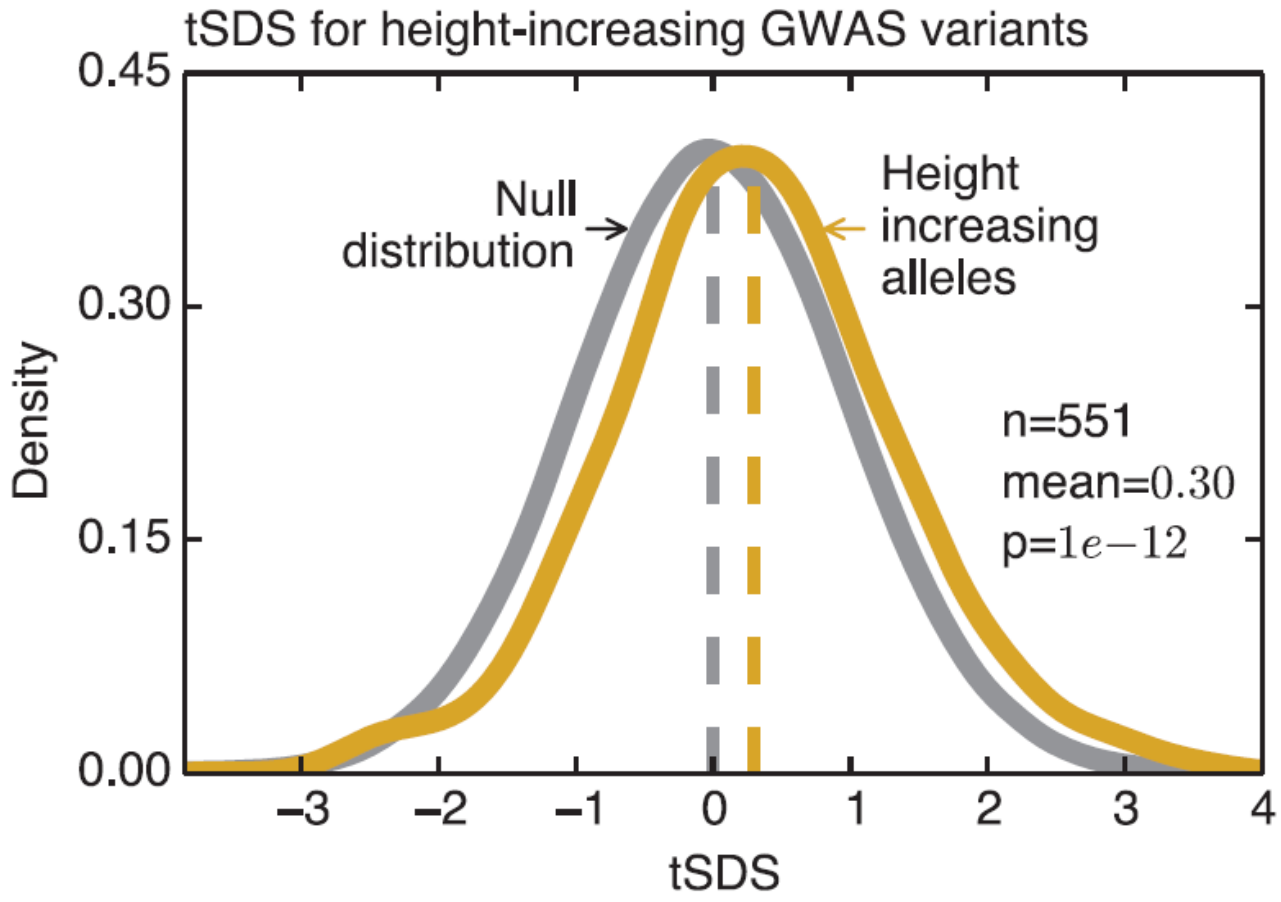
選択圧解析手法における時相の分類

		選択圧解析における時相	
		～数万年前	～数千年前
選択圧解析手法	HWE	○	×
	F_{ST}	○	×
	iHS	○	×
	ASMC	○	×
	SDS	×	○
	XP-EHH	○	×
	f	○	×
	CMS	○	×

- SDS解析を実施するためには、単一集団で数千人規模の全ゲノムシークエンスデータが必要となります。
- しかし、近い過去における選択圧を検討できる数少ない解析手法として、幅広い集団における適用が期待されています。

(岡田　随象. *実験医学増刊 Vol.36* 2018)

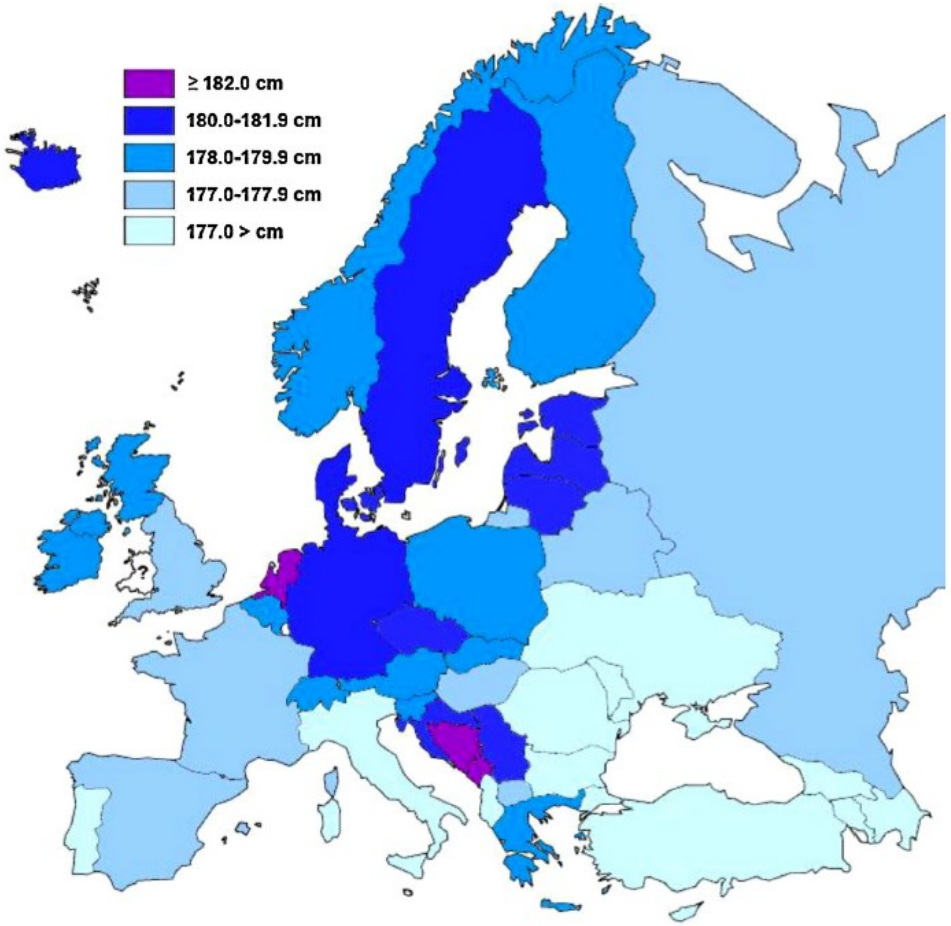
② 全ゲノムシークエンスに基づく適応進化の解明



・欧米人集団3000人の全ゲノムシークエンスデータにSDSを適用したところ、過去数千年間で、**身長を高くする遺伝子変異**が**正の選択**を強く受けていたこと(=アレル頻度が急速に増えていたこと)が明らかとなりました。

(Field Y et al. ***Science*** 2016)

② 全ゲノムシークエンスに基づく適応進化の解明



体が大きいほど、体積と表面積の比率が変わり、寒さに強くなる

⬇

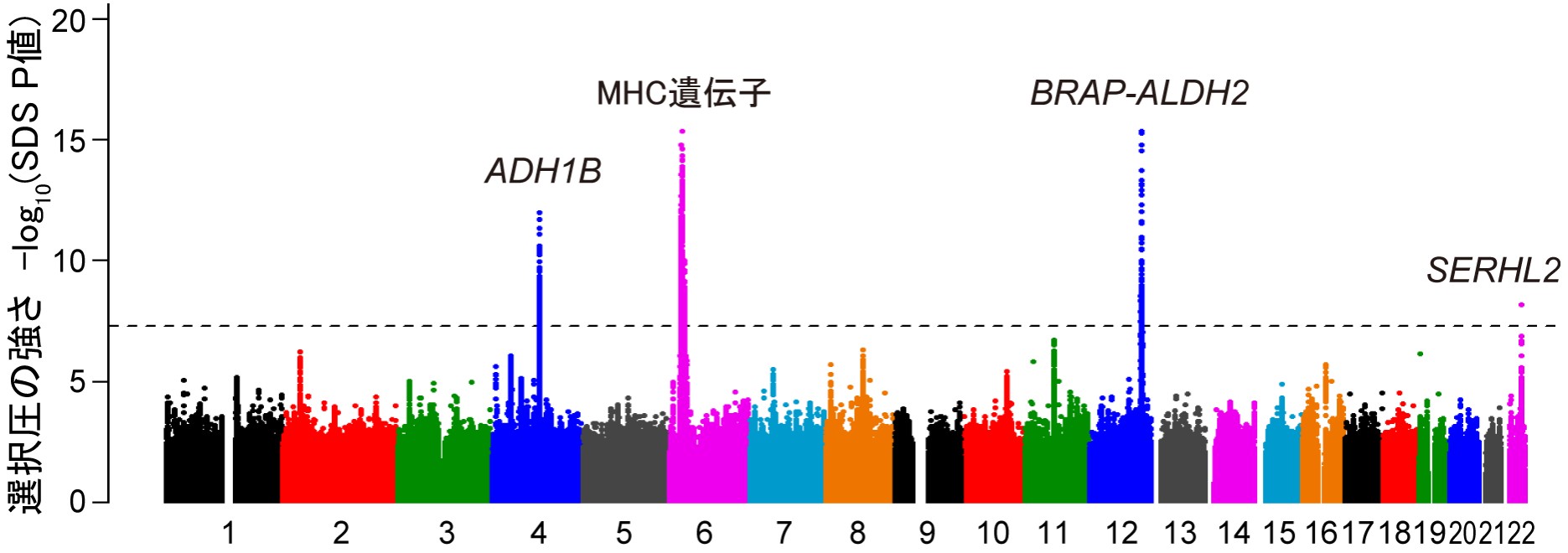
ベルクマンの法則：
同じ種でも寒冷な地域に生息するものほど体が大きくなる。

- 欧米人集団における身長への選択圧は、**北方適応**と考えられています。
- **背が高いほど寒さに強く**、北方の寒冷な環境に適応できるようです。
- 日本人集団の選択圧は、どのような形質と関わっているのでしょうか？

(Grasgruber P et al. ***Econ Hum Biol*** 2014)

② 全ゲノムシークエンスに基づく適応進化の解明

日本人集団高深度全ゲノムシークエンスデータ（2,234名）を活用した、過去数千年における選択圧解析（SDS）

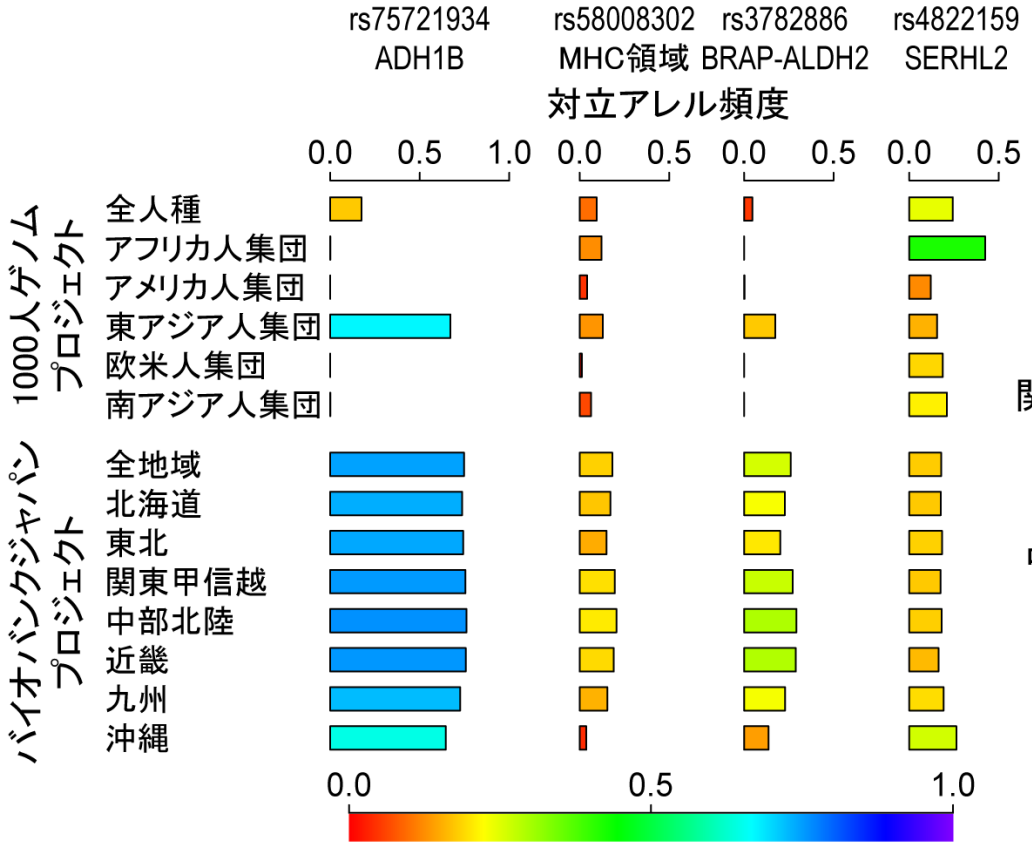


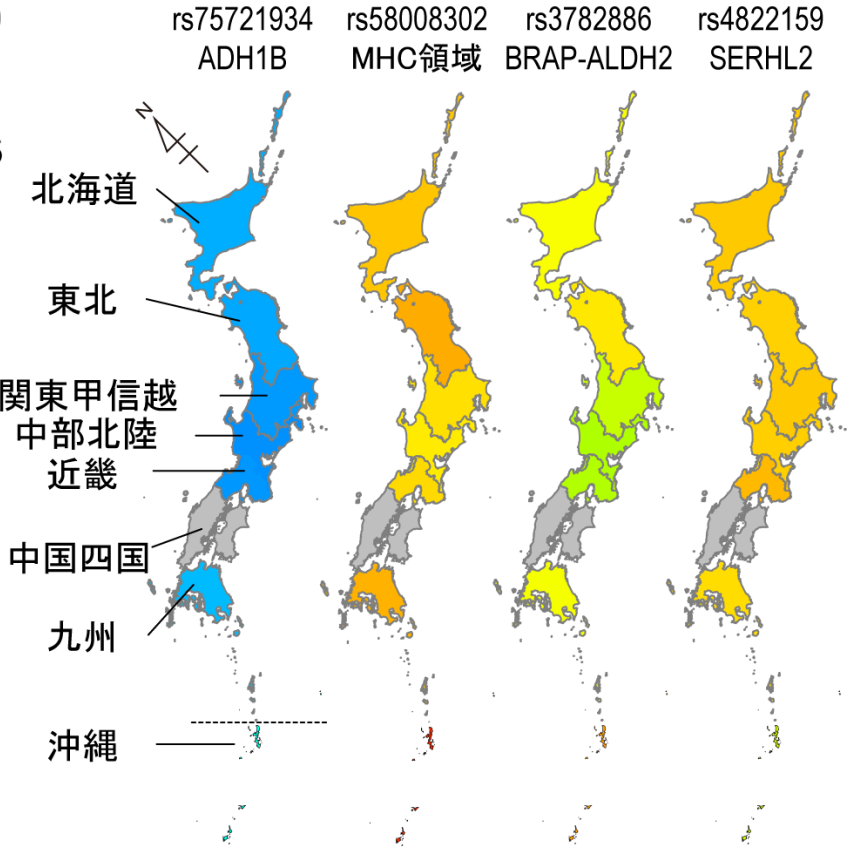
- 日本人集団2,234名の高深度全ゲノムシークエンス解析を実施し、日本人集団の近い過去（約3000年前）における選択圧を測定しました。
- 複数の遺伝子領域に、ゲノムワイド水準を満たす強い選択圧が働いていたことが明らかになりました。

(Okada Y et al. ***Nat Commun*** 2018)

② 全ゲノムシークエンスに基づく適応進化の解明

日本人集団で強い選択圧を受けた遺伝子変異のアレル頻度分布



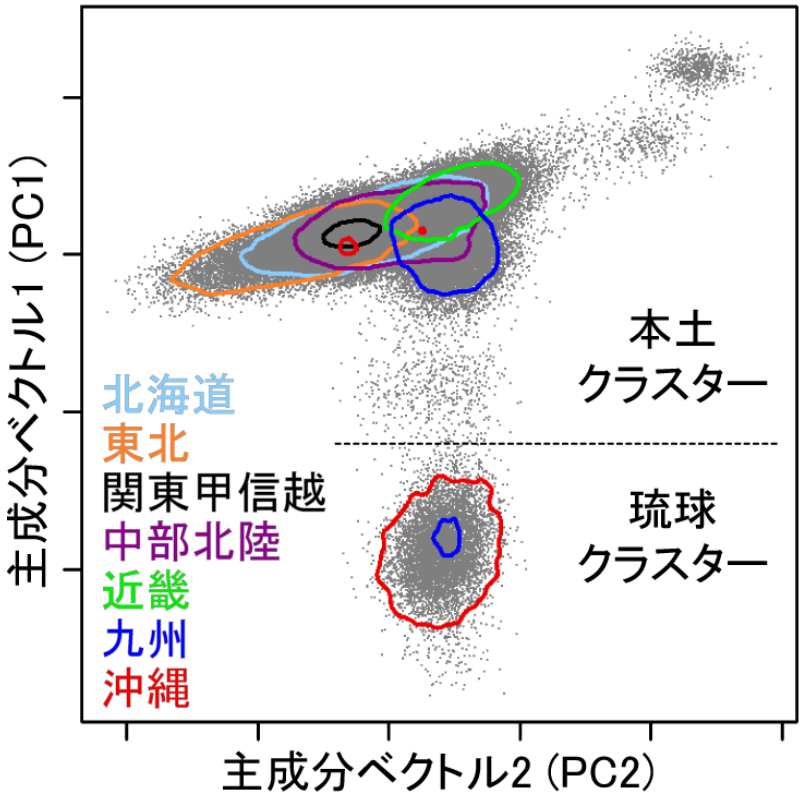


- 日本人集団で強い選択圧が働いた遺伝子変異は、日本国内でも地理的に異なる分布を示し、主に**沖縄居住者でアレル頻度が最も変化**していることが判明しました。

(Okada Y et al. ***Nat Commun*** 2018)

② 全ゲノムシークエンスに基づく適応進化の解明

ゲノム情報に基づく日本人集団の集団構造



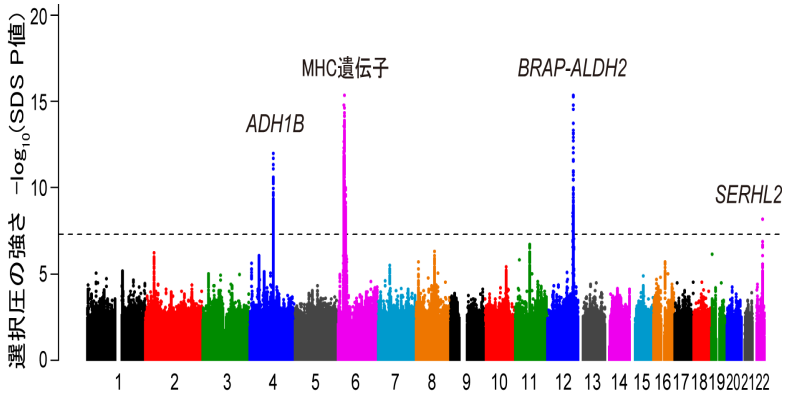
- GWASデータに対する主成分分析を用いて日本人集団の集団構造を解析すると、本州居住者（本土クラスター）と沖縄居住者（琉球クラスター）で、おおまかに二つに分かれることが確認されています。

(Okada Y et al. ***Nat Commun*** 2018)

② 全ゲノムシークエンスに基づく適応進化の解明

日本人集団における
既知の疾患リスク遺伝子変異

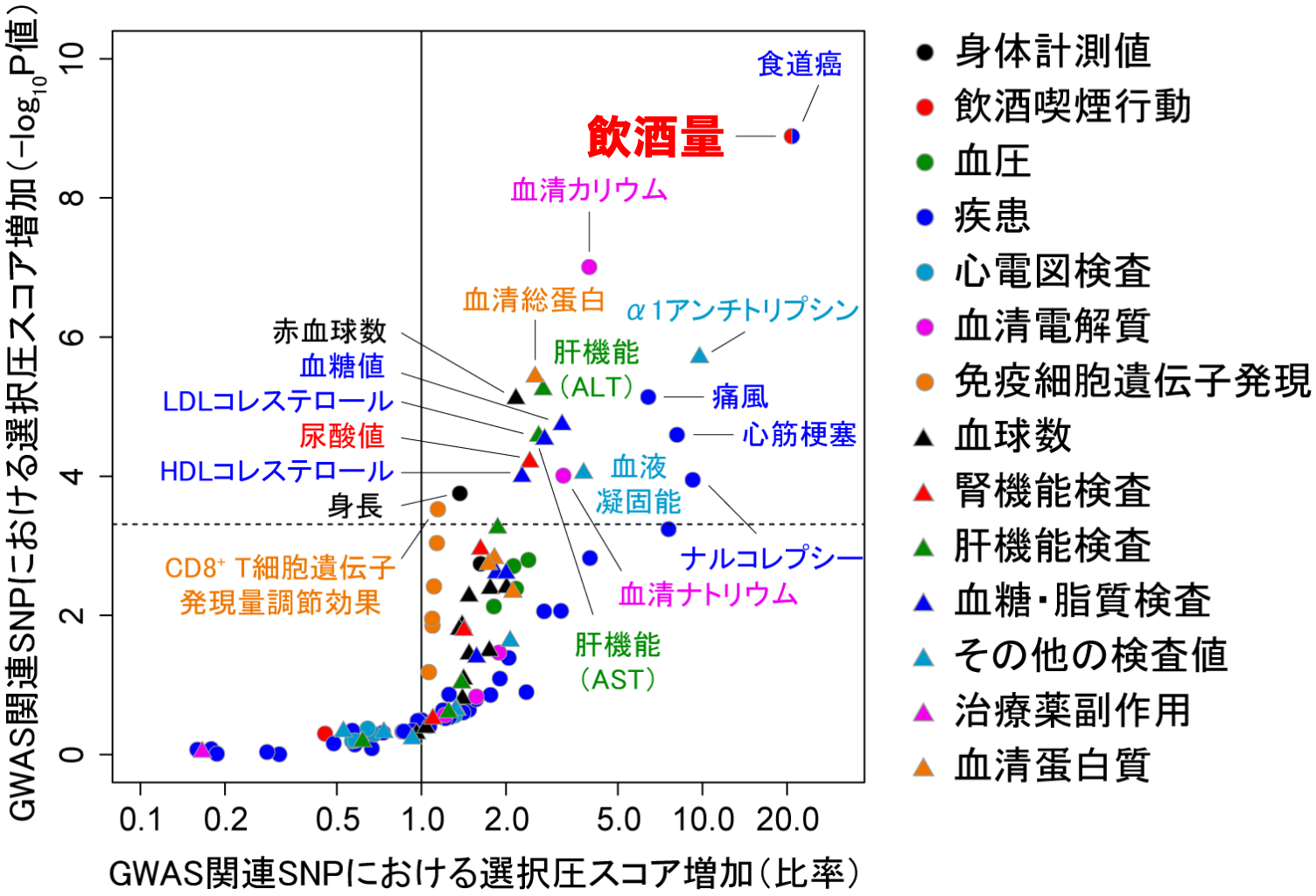
日本人集団における
ゲノムワイドな選択圧



- 日本人集団における既知の疾患リスク遺伝子変異について、今回観測した選択圧を調べることで、**各疾患に対する選択圧の強さ**を定量的に検討してみました。

② 全ゲノムシークエンスに基づく適応進化の解明

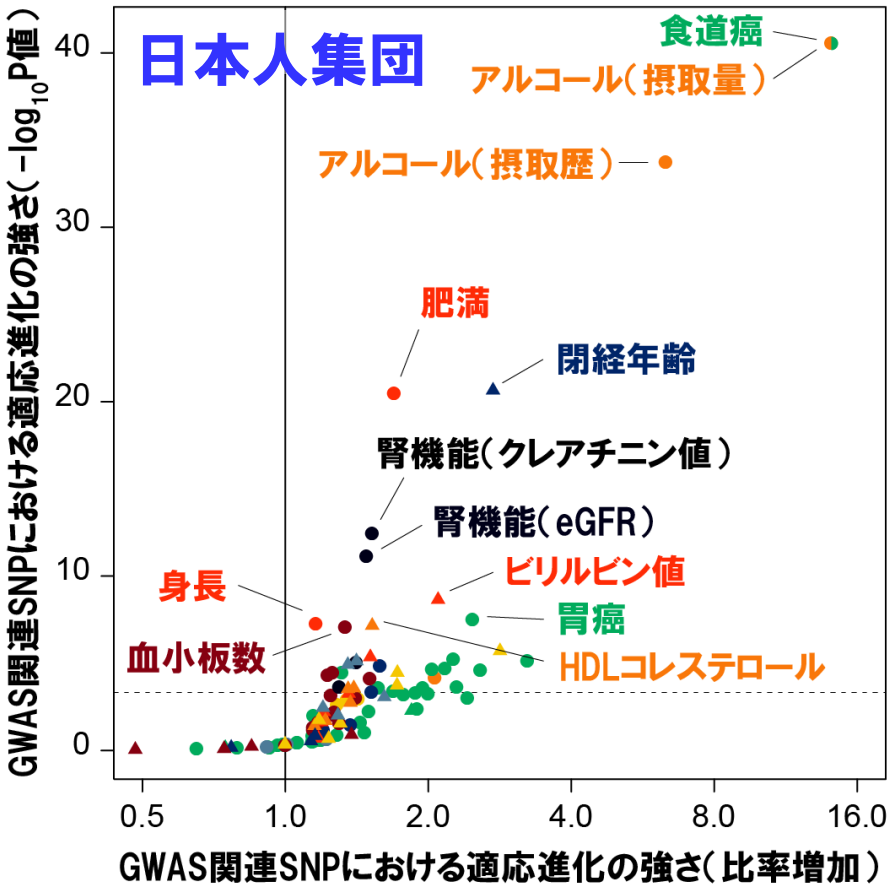
日本人集団において強い選択圧が働いた形質

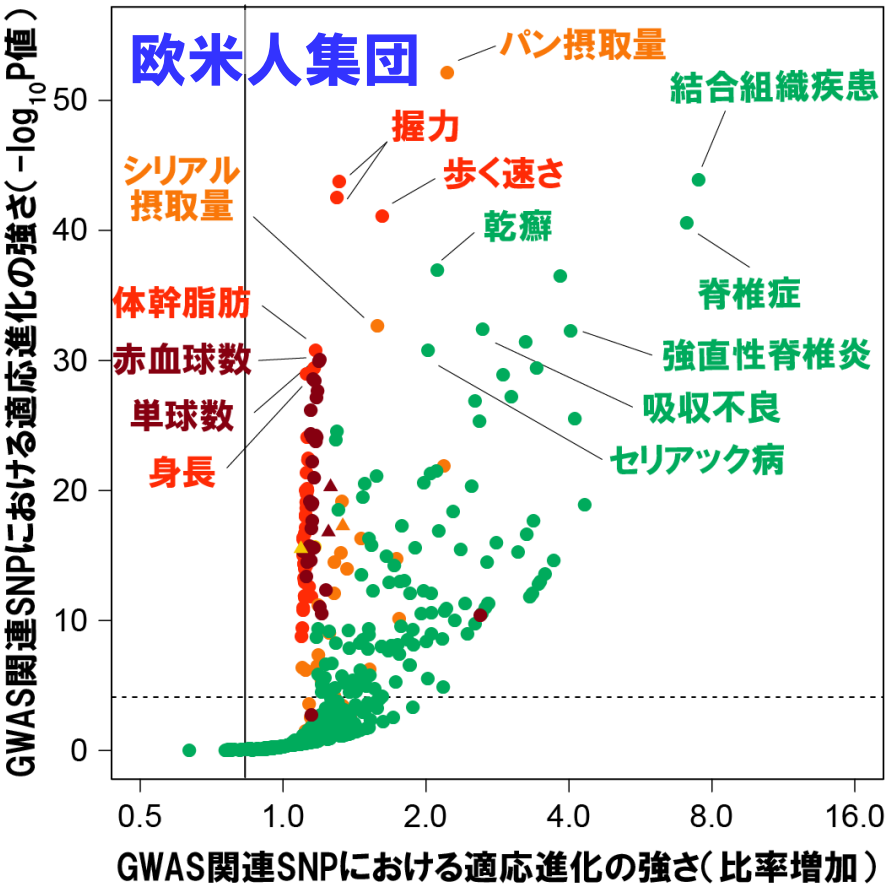


・日本人集団では、アルコール代謝および栄養に関わる病気や臨床検査値に、強い選択圧が働いていたことが明らかになりました。

日本人・欧米人集団で強い選択圧が働いた形質

ASMC選択圧解析と形質感受性遺伝子変異の統合解析結果

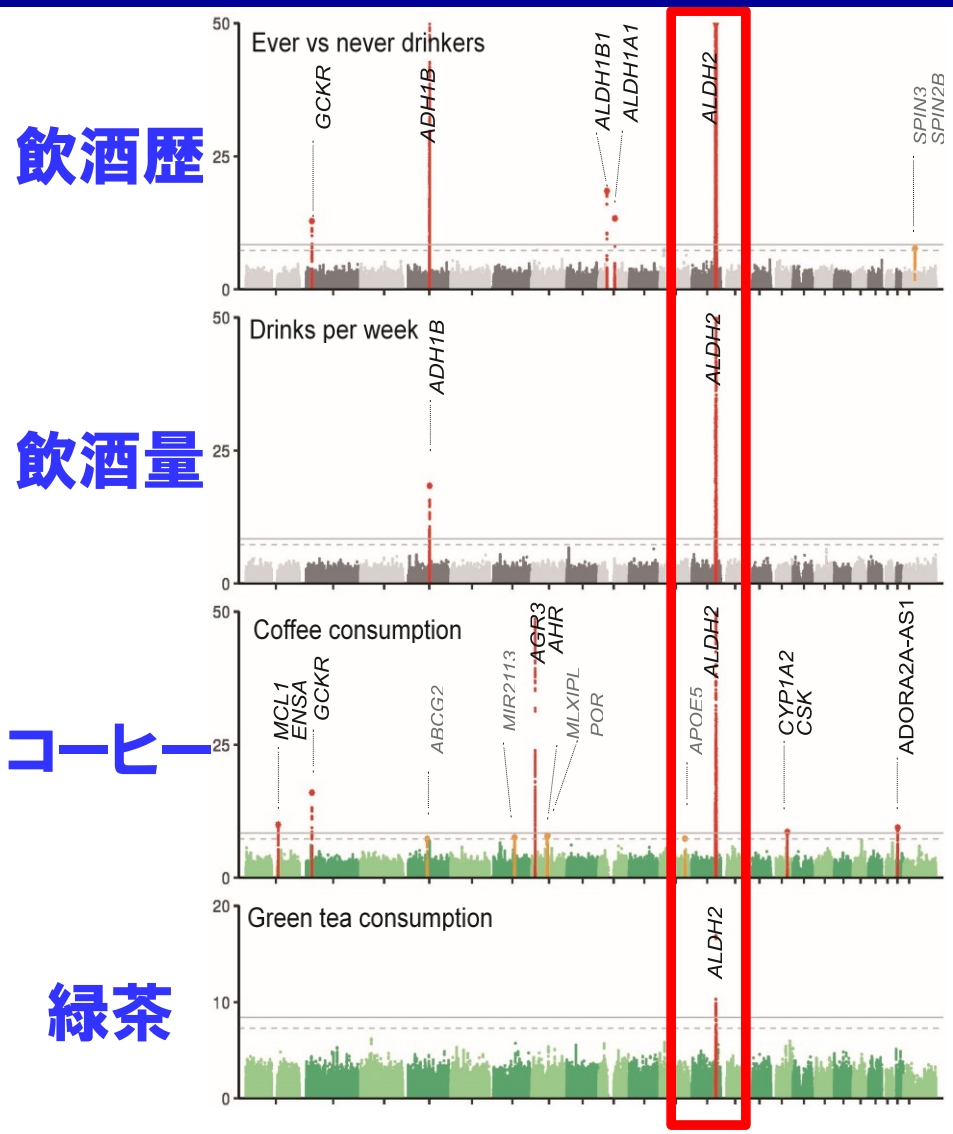


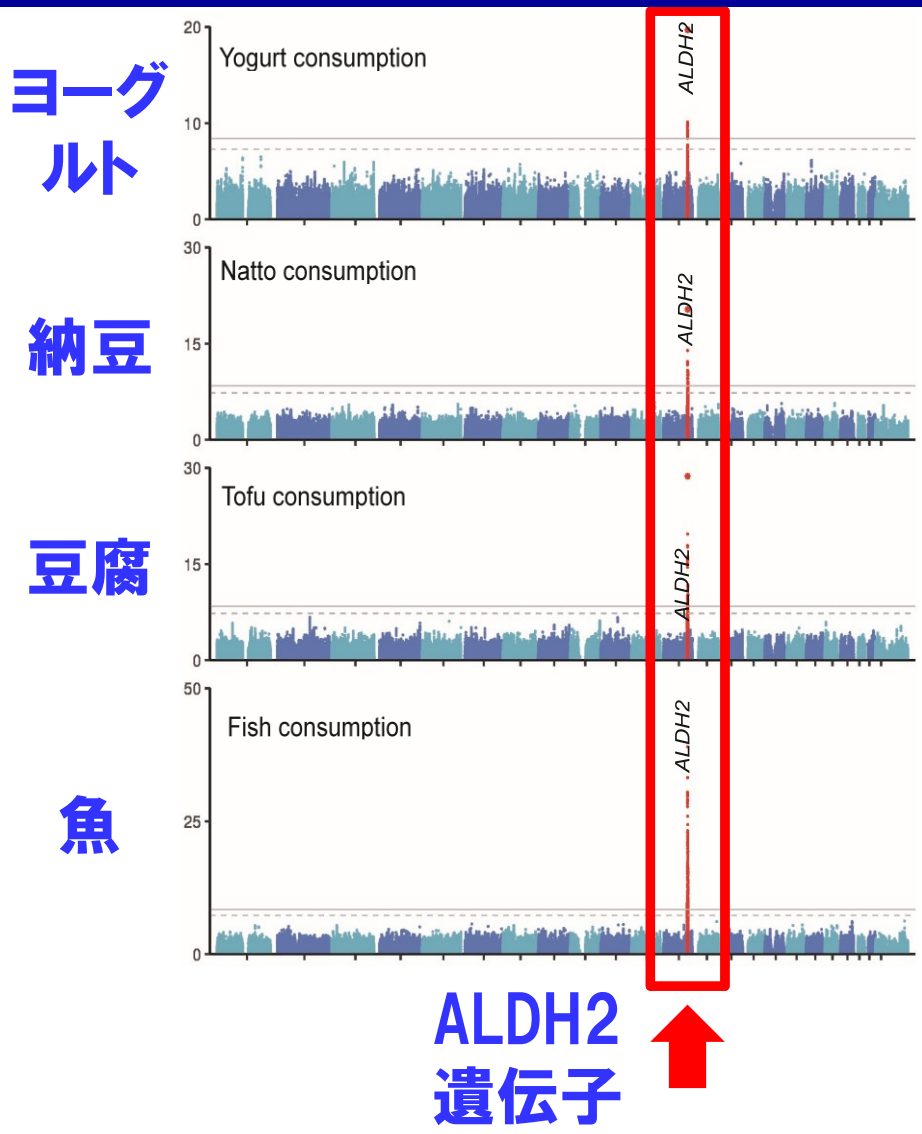


- ASMCの導入により、バイオバンク規模の選択圧解析が可能になった。
- 日本人集団では、**飲酒・肥満・腎機能**、欧米人集団では、**パン摂取量・握力・身長・歩測・免疫関節疾患**に、強い選択圧が働いていた。

(Yasumizu Y and Sakaue S et al. ***Mol Biol Evol*** 2020)

② 全ゲノムシークエンスに基づく適応進化の解明



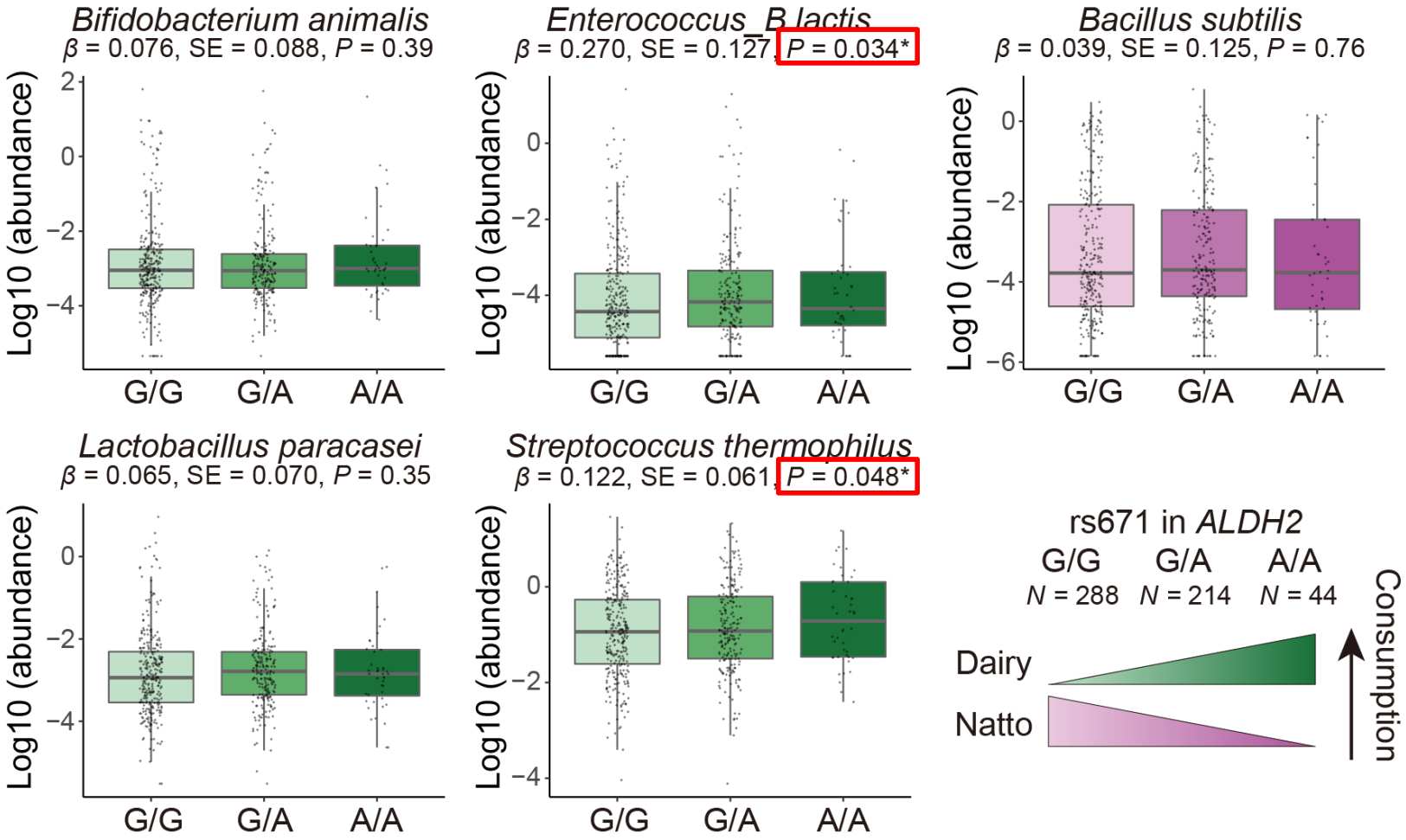


- ALDH2遺伝子変異は、飲酒を含む日本人集団の食生活を広範に規定しているようです。

(Matoba N et al. ***Nat Hum Behav*** 2020)

② 全ゲノムシークエンスに基づく適応進化の解明

ALDH2 rs671の日本人集団腸内細菌叢への影響



- ALDH2 rs671変異の日本人集団腸内細菌叢への影響は限定的であるが、食事関連細菌の一部に関連が認められた。

(Tomofuji Y et al. ***Cell Genom*** 2022)

② 全ゲノムシークエンスに基づく適応進化の解明

日本人集団でアルコール代謝に選択圧が働いた理由？

①：飲酒行動が、各集団におけるコミュニティ形成や生存競争に大事だったから？

②：古代日本人（縄文人・弥生人）の地理的な移住パターンを反映している？

③：寄生虫への防御反応に、アルデヒド代謝が重要であった？

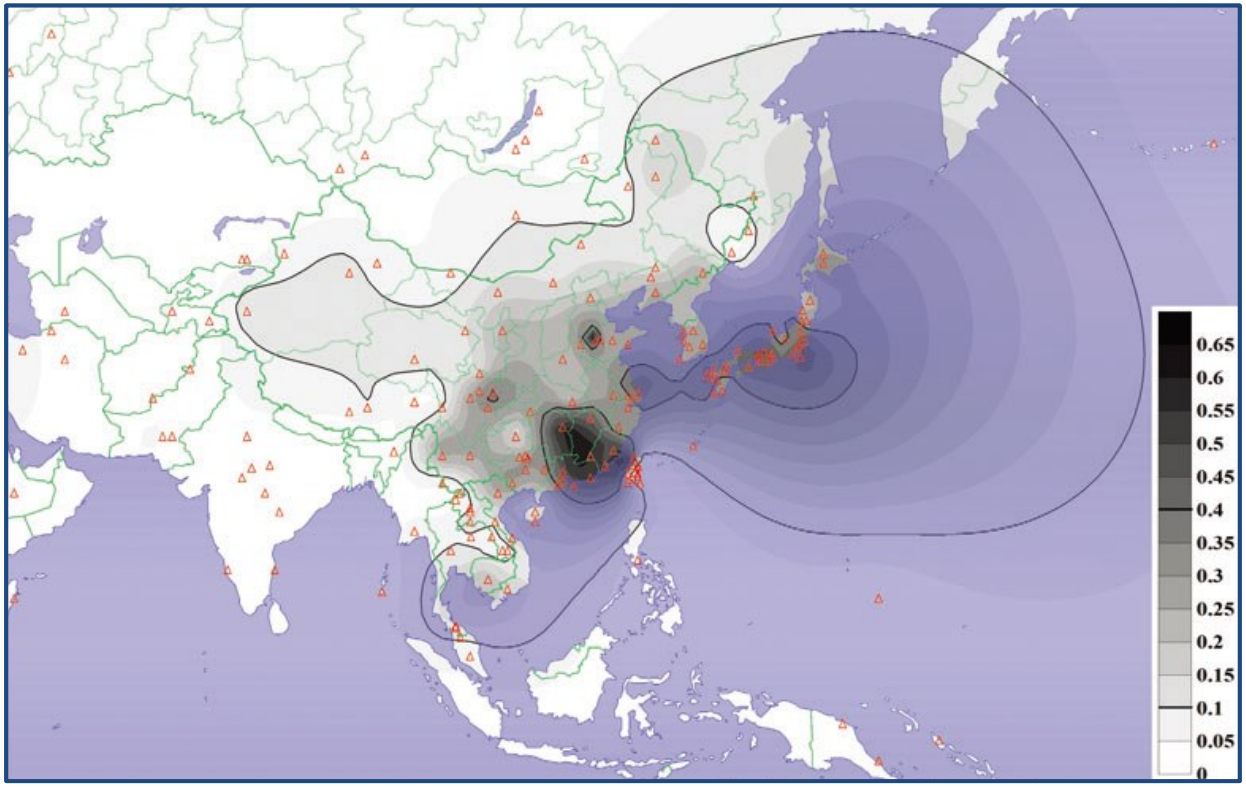
④：熱帯雨林と違い日本では自然発生するアルコールがなく、お酒が飲めなくても問題なかったから？

・何故、日本人集団で、アルコール代謝に強い選択圧が働いていたのか、諸説ありますが、**本当の理由はよくわかっていません**。

・今後、より多くのサンプルを用いた選択圧解析を行うことで、**日本人集団の適応進化の経緯**がより詳しくわかると期待されます。

② 全ゲノムシークエンスに基づく適応進化の解明

東アジア地域におけるALDH2 rs671の頻度分布

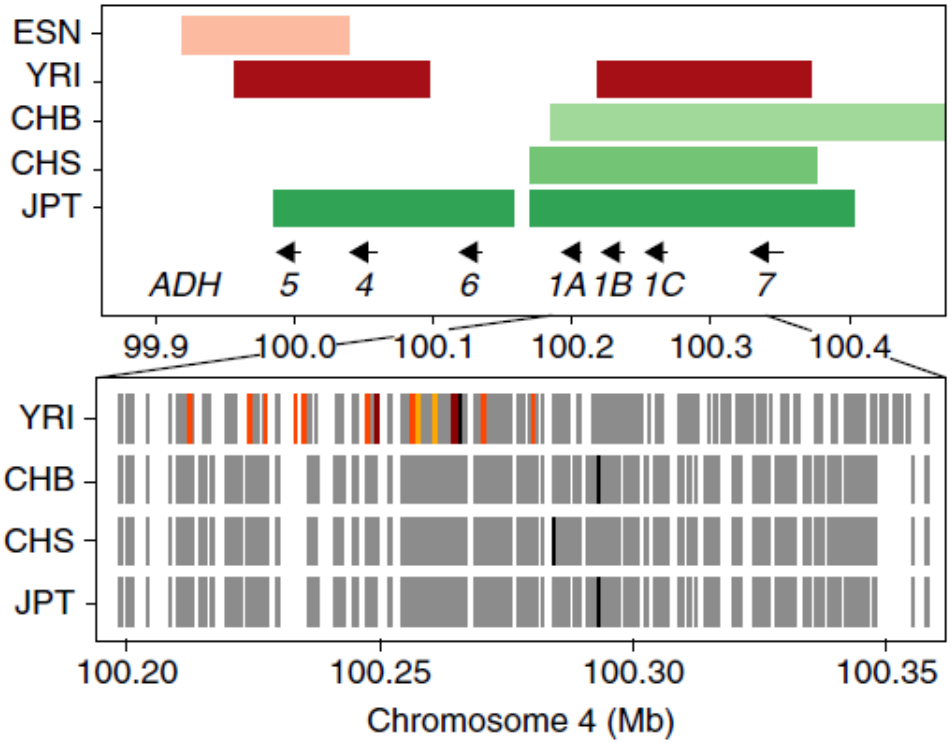


- お酒が弱くなるADH1BやALDH2のSNPは、日本人集団だけでなく、稲作を営んできた東アジア人集団においても高い頻度で観測されます。
- おそらく、これらの東アジア人集団のゲノムデータを対象に適応進化の解析を実施しても、近い結果が得られると予想されます。

(Li H et al. ***Ann Hum Genet*** 2009)

② 全ゲノムシークエンスに基づく適応進化の解明

世界中の集団でアルコール代謝に選択圧が働いている？

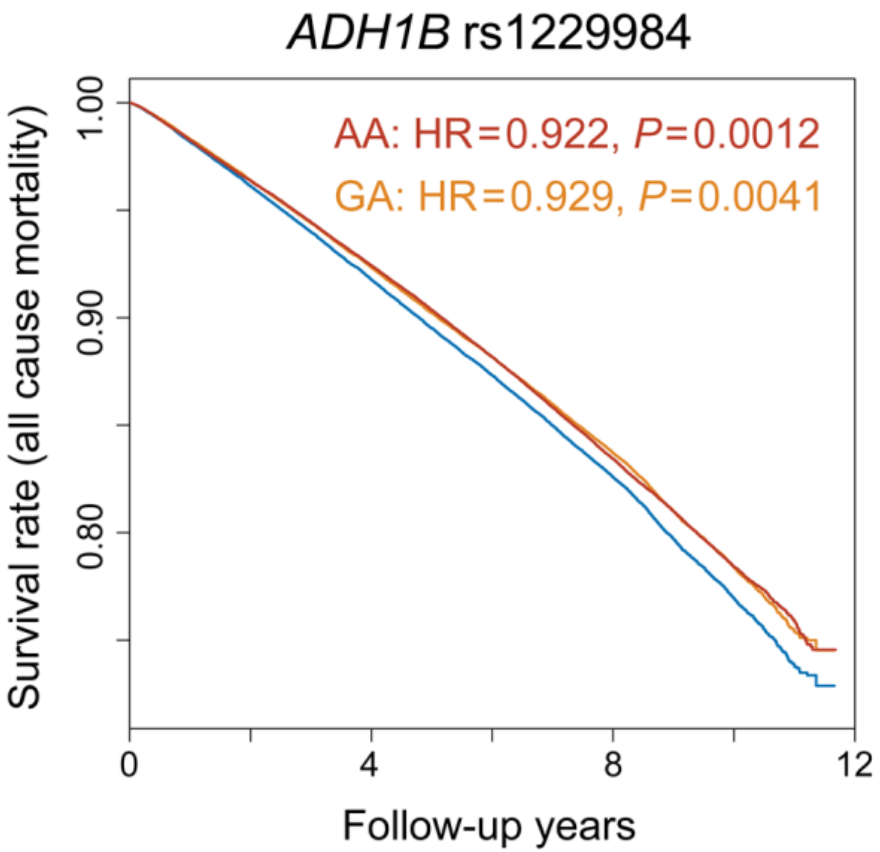


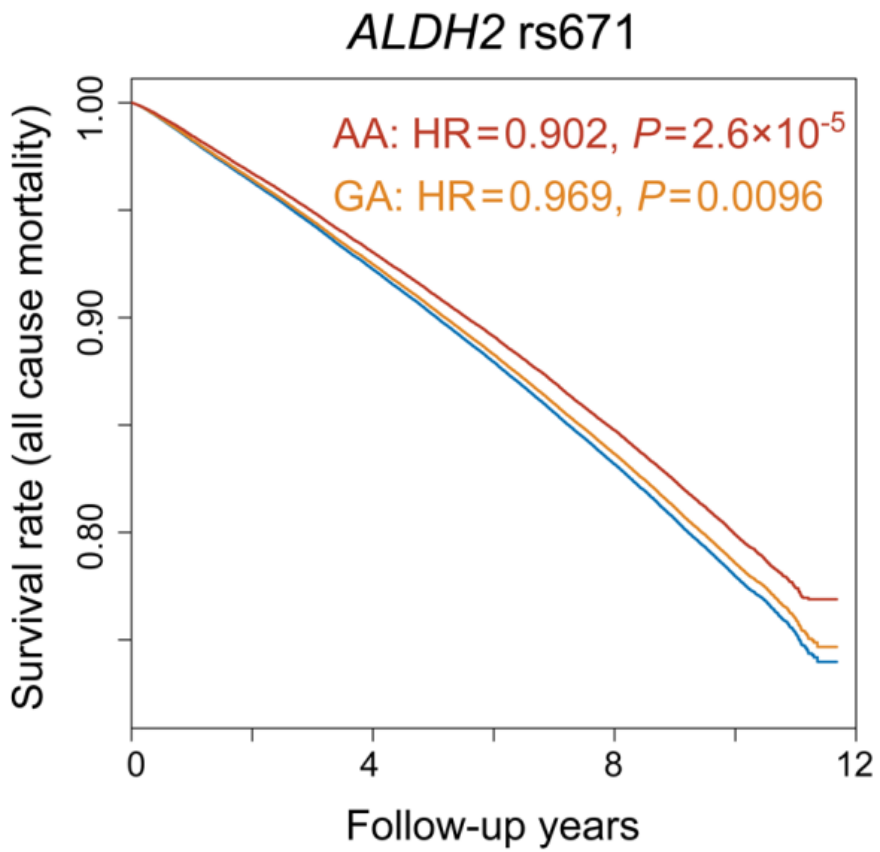
- 最近の研究では、程度の差はあるものの、世界中の複数の集団でアルコール代謝への選択圧が働いていたことが指摘されています。
- アルコール摂取量を減らす遺伝子変異が正の選択を受けてます。
- 「酒は百薬の長」という言葉は、近い将来見直されるかもしれません。

(Johnson KE et al. ***Nat Ecol Evol*** 2018)

② 全ゲノムシークエンスに基づく適応進化の解明

ADH1B/ALDH2変異による死亡率への影響



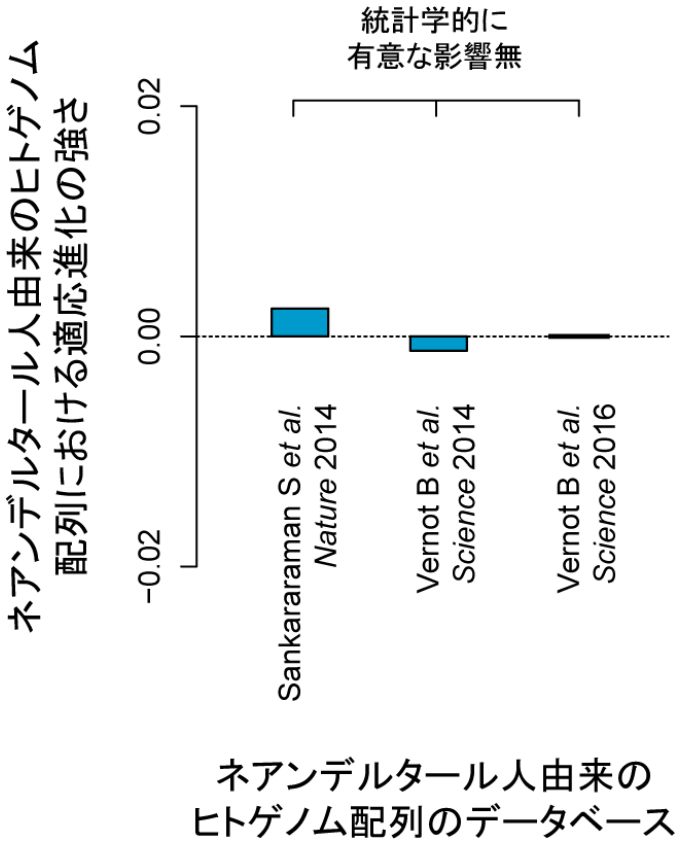


- ADH1B変異(rs1229984)とALDH2変異(rs671)は、**現代日本人集団の死亡率**に影響し、**お酒に弱いアレルと長生きの関連**が明らかに。
- 「酒は百薬の長」という言葉は、近い将来見直されるかもしれません。

(Sakaue S et al. ***Eur J Hum Genet*** 2019)

② 全ゲノムシークエンスに基づく適応進化の解明

ネアンデルタール人由来ゲノム配列における選択圧



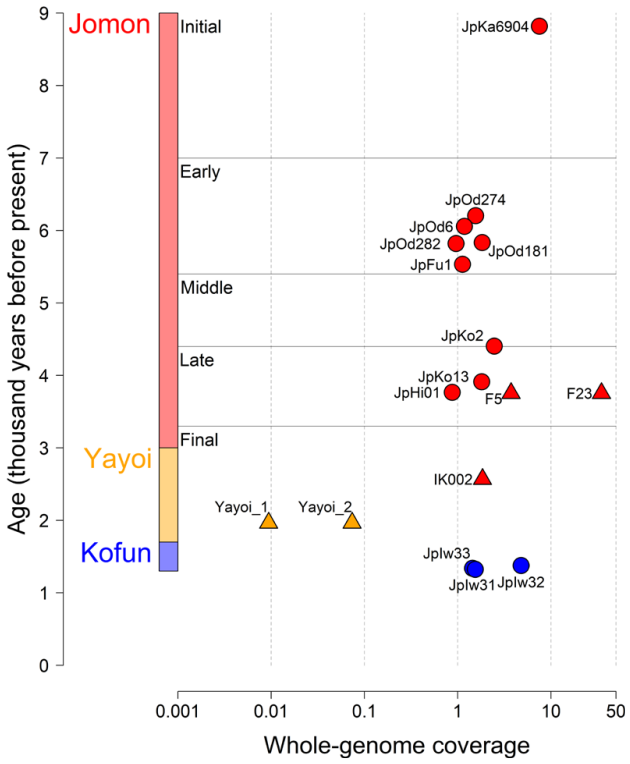
- 一方、日本人集団における**ネアンデルタール人由来のヒトゲノム配列**においては、**有意な選択圧が働いていた痕跡は確認できません**でした。
- ネアンデルタール人由来ゲノム配列が、選択圧を介して現生人類の疾患発症に寄与しているという学説とは必ずしも一致しない結果でした。

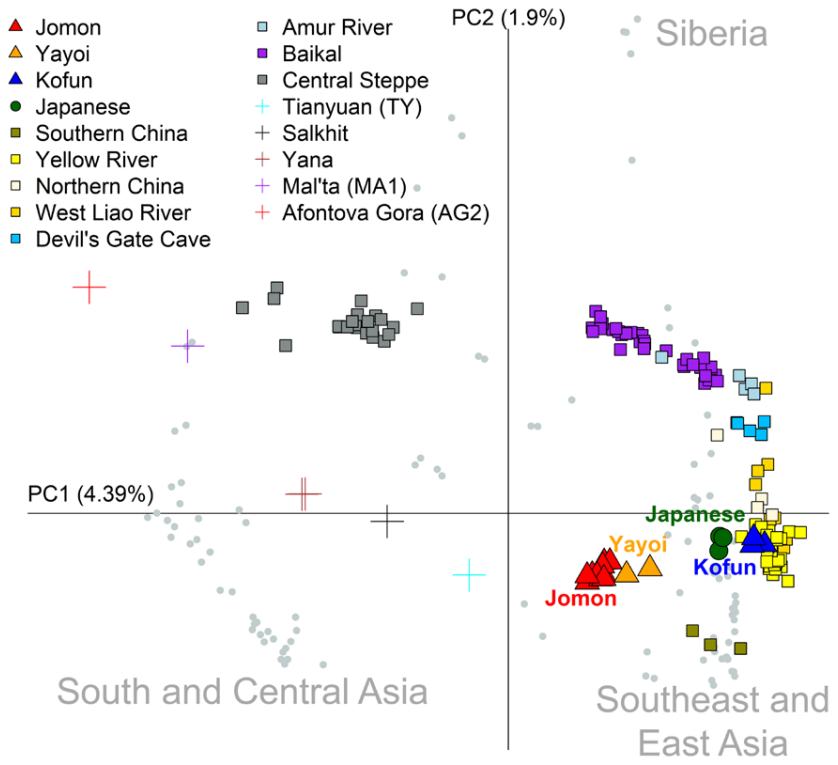
② 全ゲノムシークエンスに基づく適応進化の解明

古代日本人ゲノムと現代日本人ゲノムの比較

縄文人：20,000−15,000年前に大陸由来
弥生人：2,000−3,000年前に北東アジア由来
古墳人：1,000−2,000年前に東アジア由来

➡ **日本人集団三重構造説**





- 縄文人、弥生人、古墳人のゲノム解読に基づき、現代日本人集団のゲノム情報に3祖先が含まれていうという、**日本人集団三重構造**を提唱されています。

(Cooke N et al. ***Sci Adv*** 2021)

講義の概要

GenomeDataAnalysis4

① 選択圧と適応進化

② 全ゲノムシークエンスに基づく日本人の適応進化

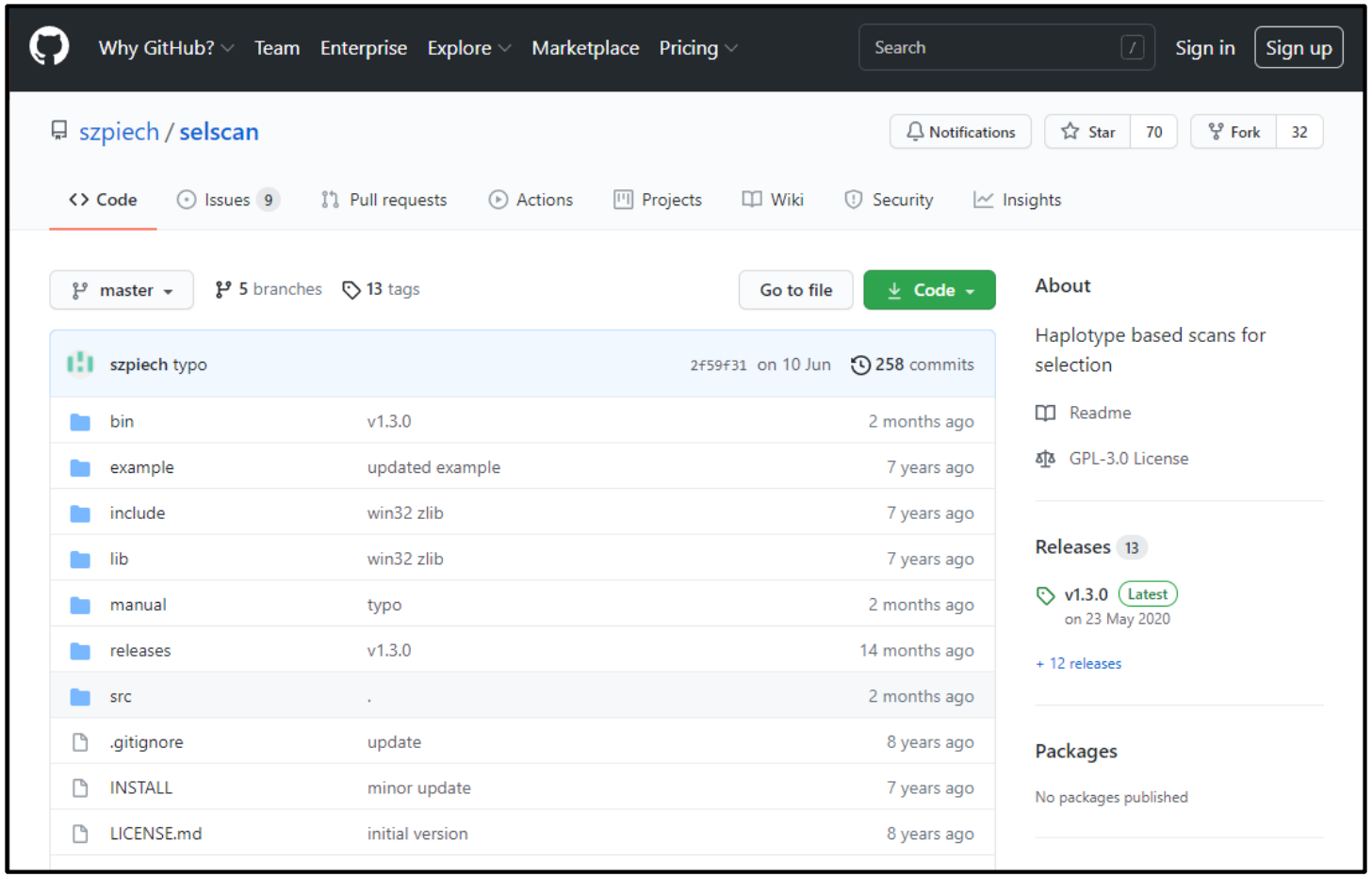
③ selscanを使った選択圧解析

本講義資料は、Windows PC上で
C:¥SummerSchoolにフォルダを配置すること
を想定しています。

③ selscanを使った選択圧解析

selscan

https://github.com/szpiech/selscan

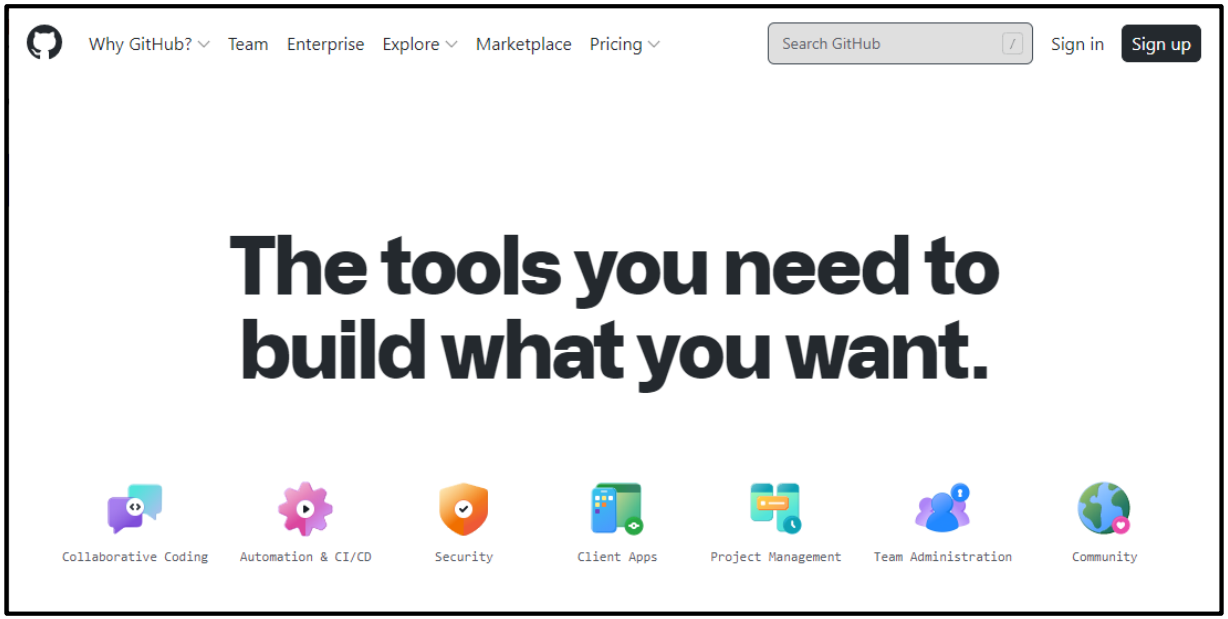


- 1000 Genomes Projectゲノムデータに対して、selscanという遺伝統計解析ソフトを使って、選択圧の解析を実施してみましょう。

③ selscanを使った選択圧解析

Github

https://github.com/features



- selscanは、GithubというWeb上のソフトウェア開発プラットフォームを使って公開されています。
- Githubでは、ソースコードを共有したり、複数人で共同してプロジェクトを進めることが可能です。
- 遺伝統計解析ソフトも、Githubで公開する例が増えています。

③ selscanを使った選択圧解析

selscanで計算可能な選択圧指標

selscanで計算可能な選択圧指標
iHS
nSL
iHH12
EHH
XP-EHH
π

マルチスレッド計算機能による計算時間短縮

Table 1. Runtime Performance (in seconds) of `ihs`, `rehh`, and `selscan` for Calculating Unstandardized iHS for Various Data Sets.

Data Set	ihs	rehh[a]	selscan				
			Threads = 1	2	4	8	16
IHS250	19,275	563	618	306	162	84	58
IHS500	45,547	1,652	1,554	782	399	220	150
IHS1000	>100,000	4,834	4,018	2,019	1,040	566	380
IHS2000	>100,000	12,652	7,054	3,633	1,869	1,046	752
CEU22	19,434	588	353	182	93	50	33

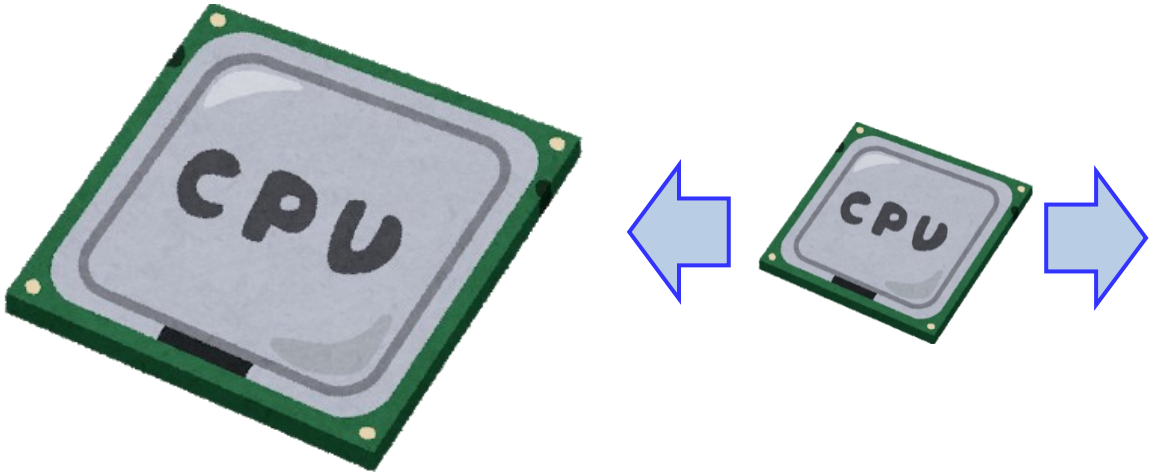
- selscanでは、手持ちのゲノムデータを対象に、幾つかの選択圧指標を計算することができます。
- **マルチスレッド計算機能**を実装したことで、時間のかかる選択圧指標計算を、短時間で実施できるようになりました。

(Szpiech ZA et al. ***Mol Biol Evol*** 2014)

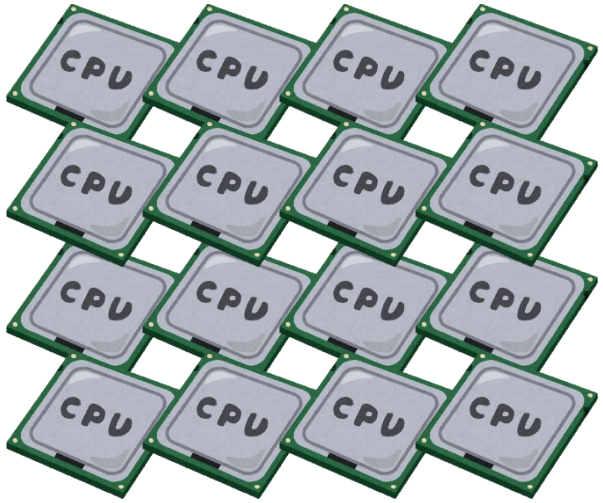
③ selscanを使った選択圧解析

計算速度を上げるための二つの方法

CPUを高速化して計算



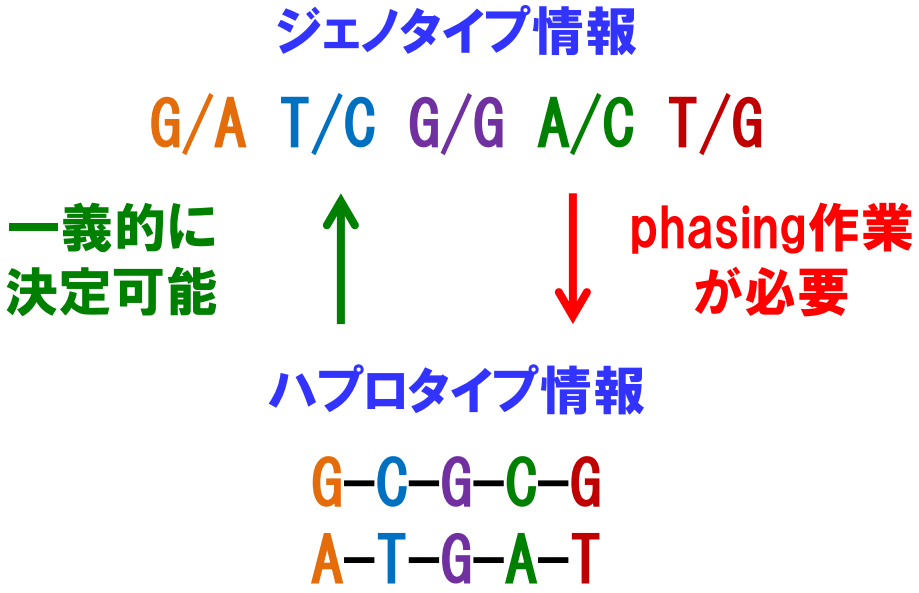
複数のCPUで並列して計算



- コンピューターの計算速度を上げる方法として、①：CPUの高速化、②：複数のCPUで並列して計算、の二つが上げられます。
- CPUの計算速度（クロック数）の高速化が技術的観点から頭打ちとなり、ソフトウェア開発においても**並列計算（＝マルチスレッド計算）の導入による高速化**の重要性が高まっています。

③ selscanを使った選択圧解析

ジェノタイプとハプロタイプの関係



phasingソフトウェア

phasing ソフトウェア
IMPUTE
Beagle
Eagle
MaCH
SHAPEIT
PHASE

- 選択圧の理論が減数分裂時のハプロタイプ組み替えに関わることから、phasing済みのハプロタイプ情報がselscanの入力データになります。
- 個人のジェノタイプからハプロタイプを推定するには、phasing作業が必要になり、複数のソフトウェアが作られています。
- ハプロタイプphasingとジェノタイプimputationは密接な関係があり、両者の機能が実装されたソフトウェアも多く存在しています。

③ selscanを使った選択圧解析

1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines (prefixed with "##"), a header line (prefixed with "#"), and data lines each containing information about a position in the genome and genotype information on samples for each position (text fields separated by tabs). Zero length fields are not allowed, a dot (".") must be used instead. In order to ensure interoperability across platforms, VCF compliant implementations must support both LF ("\n") and CR+LF ("\r\n") newline conventions.

1.1 An example

付帯情報の説明

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A      3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .      47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
```

各行が各SNPに対応

各サンプル

- selscanの入力ファイルは、"vcfファイル"と"mapファイル"になります。
- vcfファイル上のハプロタイプ情報と、mapファイル上の各SNPの位置情報を使用します。

(https://samtools.github.io/hts-specs/VCFv4.3.pdf)

③ selscanを使った選択圧解析

- GT (String): Genotype, encoded as allele values separated by either of / or |. The allele values are 0 for the reference allele (what is in the REF field), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on. For diploid calls examples could be 0/1, 1 | 0, or 1/2, etc. Haploid calls, e.g. on Y, male non-pseudoautosomal X, or mitochondrion, are indicated by having only one allele value. A triploid call might look like 0/0/1. If a call cannot be made for a sample at a given locus, '.' must be specified for each missing allele in the GT field (for example './.' for a diploid genotype and '.' for haploid genotype). The meanings of the separators are as follows (see the PS field below for more details on incorporating phasing information into the genotypes):
  - / : genotype unphased
  - | : genotype phased
- GL (Float): Genotype likelihoods comprised of comma separated floating point log_{10}-scaled likelihoods for all possible genotypes given the set of alleles defined in the REF and ALT fields. In presence of the GT field the same ploidy is expected; without GT field, diploidy is assumed.

  GENOTYPE ORDERING. In general case of ploidy P and N alternate alleles (0 is the REF and $1 \ldots N$ the alternate alleles), the ordering of genotypes for the likelihoods can be expressed by the following pseudocode with as many nested loops as ploidy: †

```
for aP = 0...N
  for aP-1 = 0...aP
      ...
      for a1 = 0...a2
          println a1a2...aP
```

・vcfファイル上のジェノタイプデータが、phasing済みのハプロタイプ情報なのか未実施なのかは、ジェノタイプの区切り文字を見るとわかります。

・"|"で区切られていたphasing実施済み、"/"なら未実施になります。

(https://samtools.github.io/hts-specs/VCFv4.3.pdf)

③ selscanを使った選択圧解析

example.map

1	SNP1	0	10000
1	SNP2	0	20000
1	SNP3	0	30000
1	SNP4	0	40000

example.tped

1	SNP1	0	10000	A	A	A	G	G	G	A	G	G	A	G	G
1	SNP2	0	20000	C	C	T	C	T	T	C	T	C	T	T	T
1	SNP3	0	30000	C	A	C	C	A	A	C	C	C	C	C	A
1	SNP4	0	40000	T	G	T	T	G	T	G	G	T	T	T	G

example.ped

Family1	Sample1	0	0	1	1	A	A	C	C	A	C	T	G
Family2	Sample2	0	0	2	1	A	G	C	T	C	C	T	T
Family3	Sample3	0	0	2	1	G	G	T	T	A	A	T	G
Family4	Sample4	0	0	2	1	A	G	C	T	C	C	G	G
Family5	Sample5	0	0	1	1	A	G	C	T	C	C	T	T
Family6	Sample6	0	0	2	1	G	G	T	T	A	C	T	G

example.tfam

Family1	Sample1	0	0	1	1
Family2	Sample2	0	0	2	1
Family3	Sample3	0	0	2	1
Family4	Sample4	0	0	2	1
Family5	Sample5	0	0	1	1
Family6	Sample6	0	0	2	1

- その他、”tpedファイル”というPLINKファイル形式でも実行可能です。
- ”行がサンプル・列がSNP“であったpedファイル形式と行列が入れ替わり、tpedファイル形式では”行がSNP・列がサンプル”に対応します。
- 一般に、列数が多いより行数が多い方が扱いやすいため、サンプル数よりSNP数が多い昨今のゲノムデータに対応した形式ともいえます。

③ selscanを使った選択圧解析

example.tped

1	SNP1	0	10000	A	A	A	G	G	G	A	G	G	A	G	G
1	SNP2	0	20000	C	C	T	C	T	T	C	T	C	T	T	T
1	SNP3	0	30000	C	A	C	C	A	A	C	C	C	C	C	A
1	SNP4	0	40000	T	G	T	T	G	T	G	G	T	T	T	G

selscan用に
0/1アレル表記
へと変換

G-C-C-T
ハプロタイプ
に対応

1	SNP1	0	10000	0	0	0	1	1	1	0	1	1	0	1	1
1	SNP2	0	20000	1	1	0	1	0	0	1	0	1	0	0	0
1	SNP3	0	30000	1	0	1	1	0	0	1	1	1	1	1	0
1	SNP4	0	40000	0	1	0	0	1	0	1	1	0	0	0	1

- tpedファイル形式では、各列をハプロタイプに対応させることで、vcfファイル形式のように、phasing後のハプロタイプ情報の記録が可能です。
- なお、selscanの仕様上、各SNPのアレルをA/T/G/C表記から0/1表記に変換する必要があります。

③ selscanを使った選択圧解析

statgen@statgen-PC: ~

$ cd /mnt/c/SummerSchool/GenomeDataAnalysis4/1KG_EUR

※Cygwinの場合 /mnt/を/cygdrive/に変えてください

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis4/1KG_EUR

$ ls *vcf.gz

1KG_EUR_chr2Q_MAF005.phased.vcf.gz

$ ls *map

1KG_EUR_chr2Q_MAF005.phased.map

$ wc *map

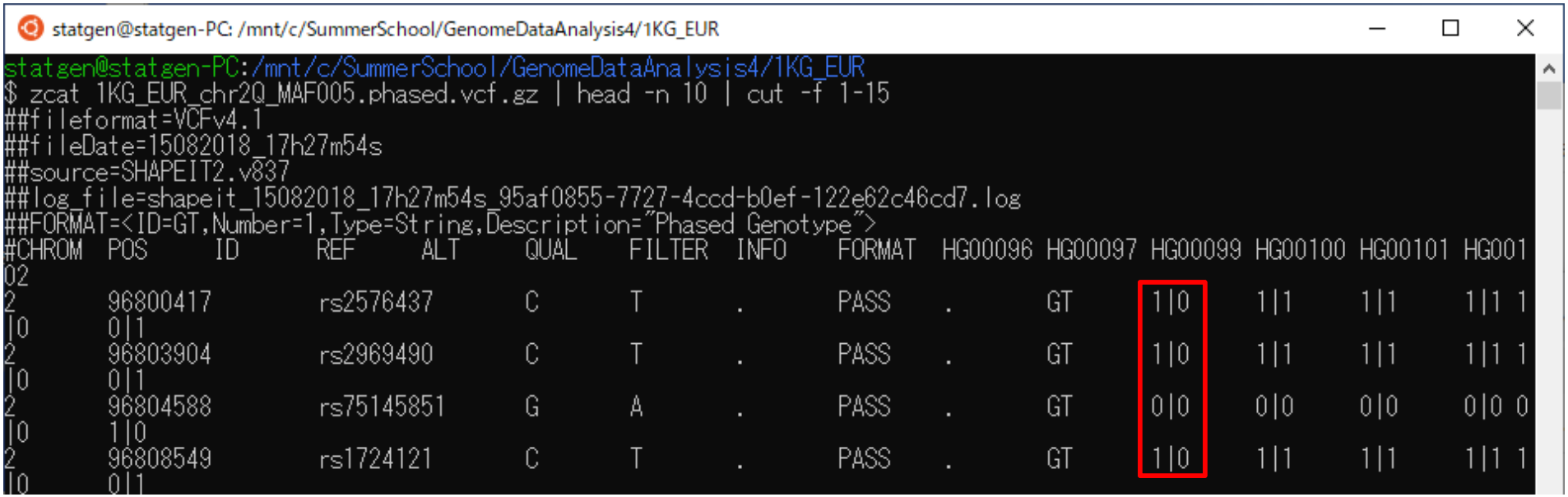
332573 1330292 10162669 1KG_EUR_chr2Q_MAF005.phased.map

- **GWASデータとして、1000 Genomes Project Phase3データの欧米人365名のSNPデータを取得しました。**
- **2番染色体の長腕(2q)においてマイナーアレル頻度≧0.05を示した332,573SNPを対象としています。**

③ selscanを使った選択圧解析

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis4/1KG_EUR

$ zcat 1KG_EUR_chr2Q_MAF005.phased.vcf.gz | head -n 10 | cut -f 1-15



- vcfファイルの中身を見てみましょう。
- gzファイル形式で圧縮されているので、zcatコマンドで一時的に解凍し、最初の10行および15列を、パイプ(|)を使って切り出します。
- "|"で区切られた、phasing済みハプロタイプが確認できました。

③ selscanを使った選択圧解析

statgen@statgen-PC: ~

$ cd /mnt/c/SummerSchool/GenomeDataAnalysis4/1KG_EUR/

※Cygwinの場合 /mnt/を/cygdrive/に変えてください

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis4/1KG_EUR

$ ./selscan --ihs --vcf 1KG_EUR_chr2Q_MAF005.phased.vcf.gz --map 1KG_EUR_chr2Q_MAF005.phased.map --out 1KG_EUR_chr2Q_MAF005.phased --maf 0.05 --threads 3 --cutoff 0.05 --trunc-ok --max-extend 5000000

※ファイル"selscan_Command.txt"を開いて、内容をShellにコピー＆ペーストして下さい。

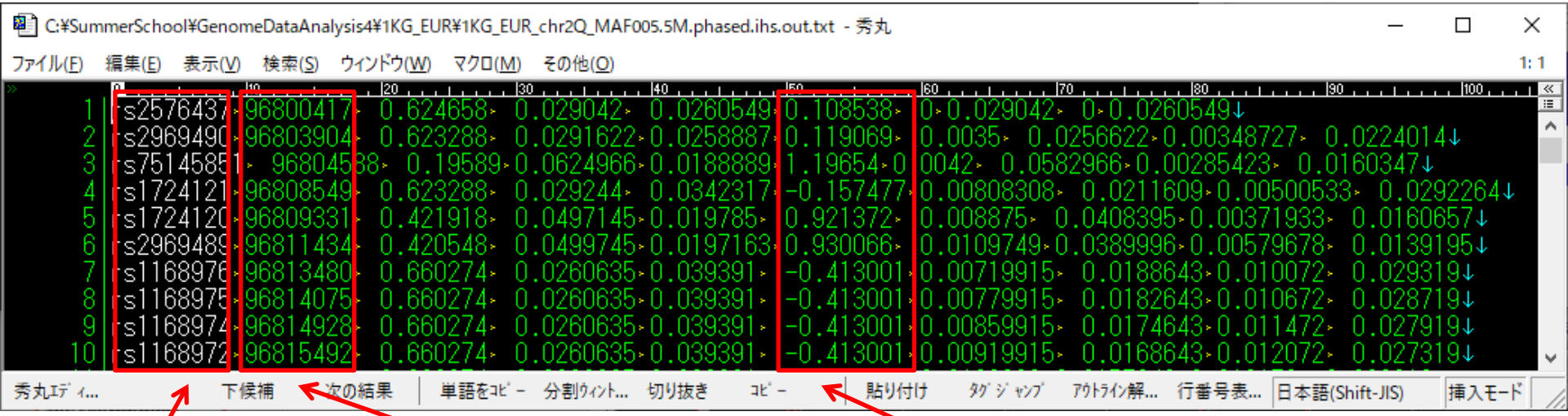
※Macユーザーの方は、"selscan_for_mac"の中にある実行ファイル("selscan")に置き換えて実行してください。

- selscanに実装された選択圧解析指標のうち、iHSを計算してみます。
- iHS計算は、"./selscan --ihs --vcf（vcfファイル）--map（mapファイル）--out（出力ファイル名）（その他のコマンド）"という形で実行します。

③ selscanを使った選択圧解析

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis4/1KG_EUR

$ cat 1KG_EUR_chr2Q_MAF005.5M.phased.ihs.out.txt

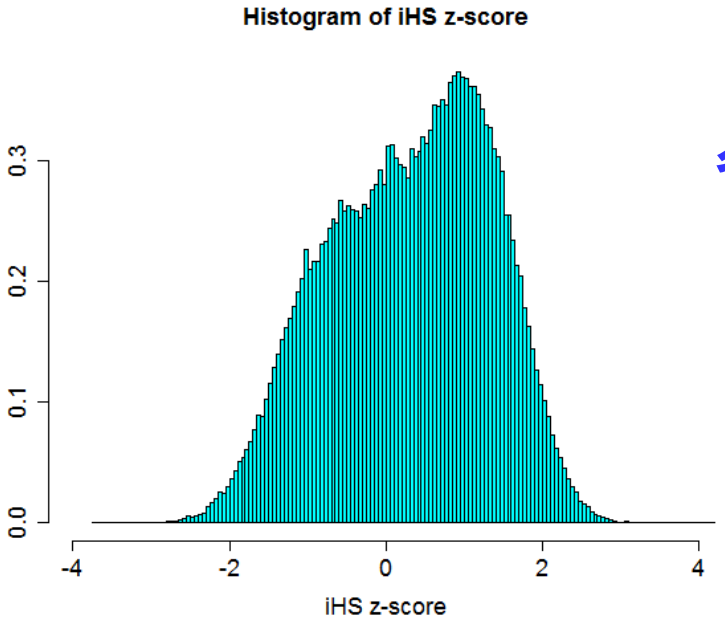


$1:SNP ID　$2:染色体上の位置　　$6:iHSの値

- **iHSの結果ファイルでは、各行が各SNPにおける選択圧を表しています。**
- **1列目がSNP ID、2列目が染色体上の位置を表します。**
- **6列目がiHSの値になります。**
- **詳細は、同梱したマニュアル(selscan-manual.pdf)を参照してください。**

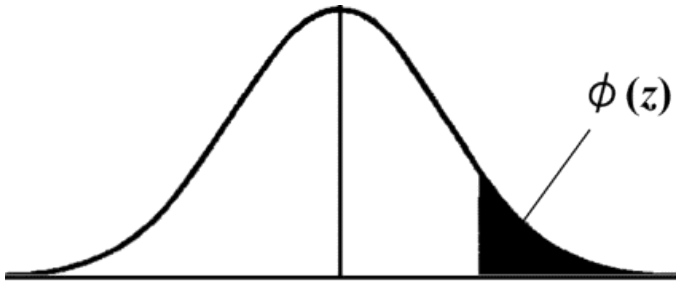
③ selscanを使った選択圧解析

正規化前のiHS Z値の分布



各種パラメーター
を用いた正規化

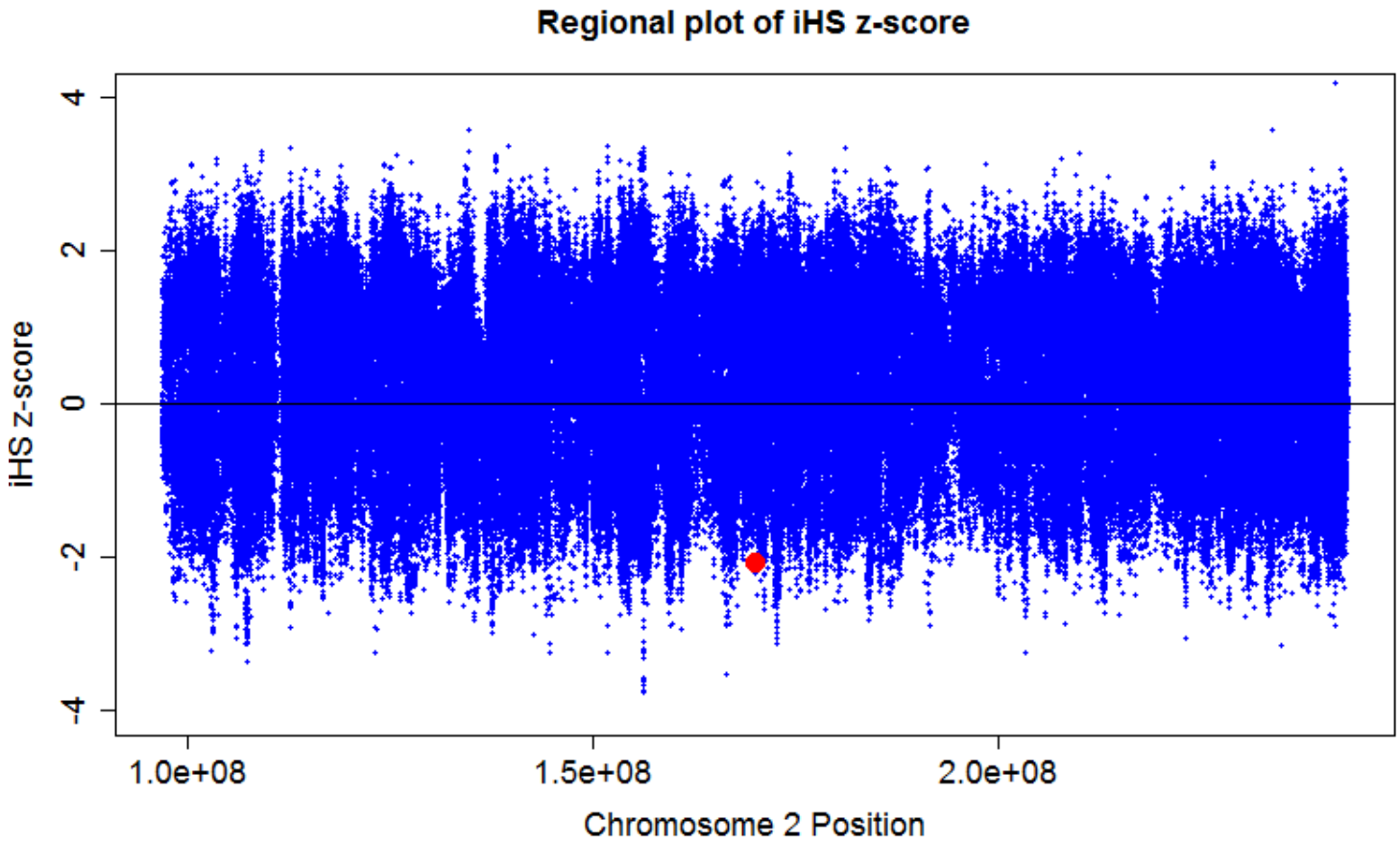
正規化後のiHS Z値の分布



※ファイル"iHS_Plot.R"を開いて、改変の上、Rにコピー＆ペーストして下さい。

- iHSの値は、帰無仮説下で正規分布に従うZ値として計算されます。
- 実際には、各SNPのアレル頻度や、各SNP位置での減数分裂組換率（recombination rate）に依存して偏った値をとるため、追加の正規化作業が必要となります（夏の学校では、正規化手順については説明しません）。

③ selscanを使った選択圧解析



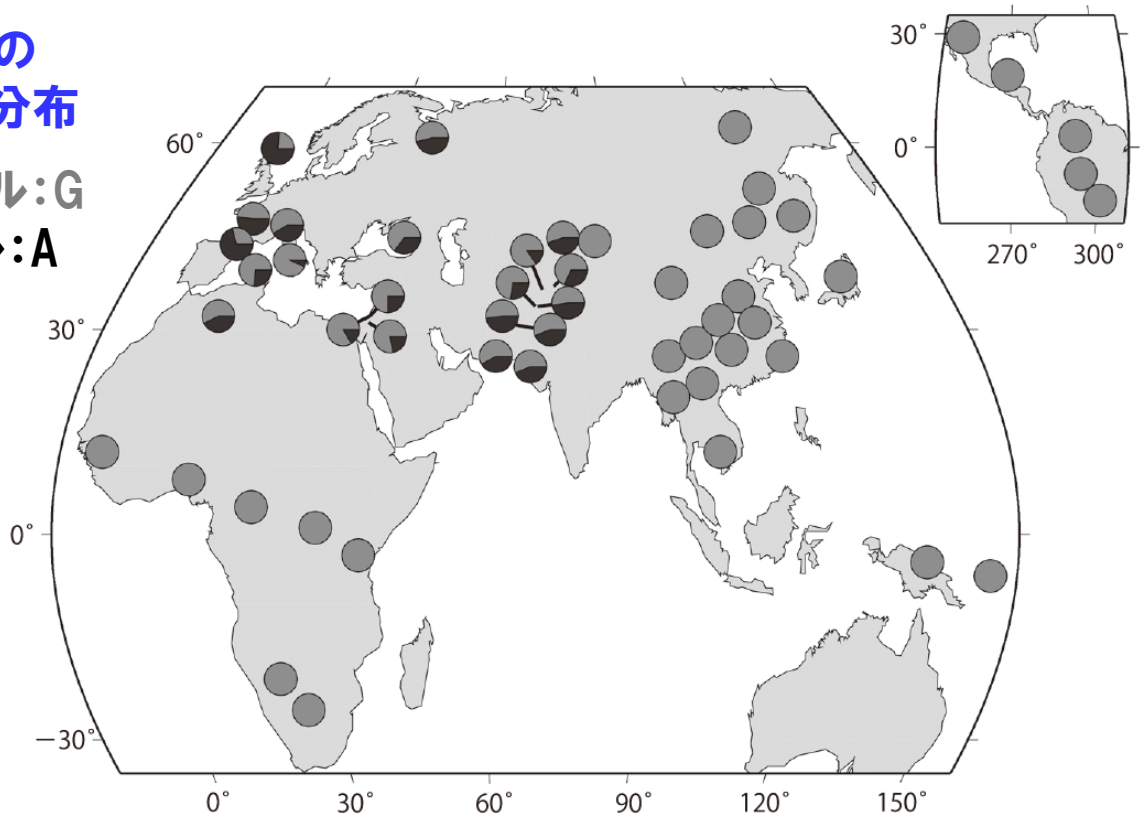
※ファイル"iHS_Plot.R"を開いて、改変の上、Rにコピー＆ペーストして下さい。

- iHSのZ値を、染色体上の位置に沿ってプロットしてみましょう。
- 高いZ値（もしくは低いZ値）を示すSNPに、相対的に強い選択圧が働いていたと考えられます。

③ selscanを使った選択圧解析

rs49788235の
集団アレル頻度分布

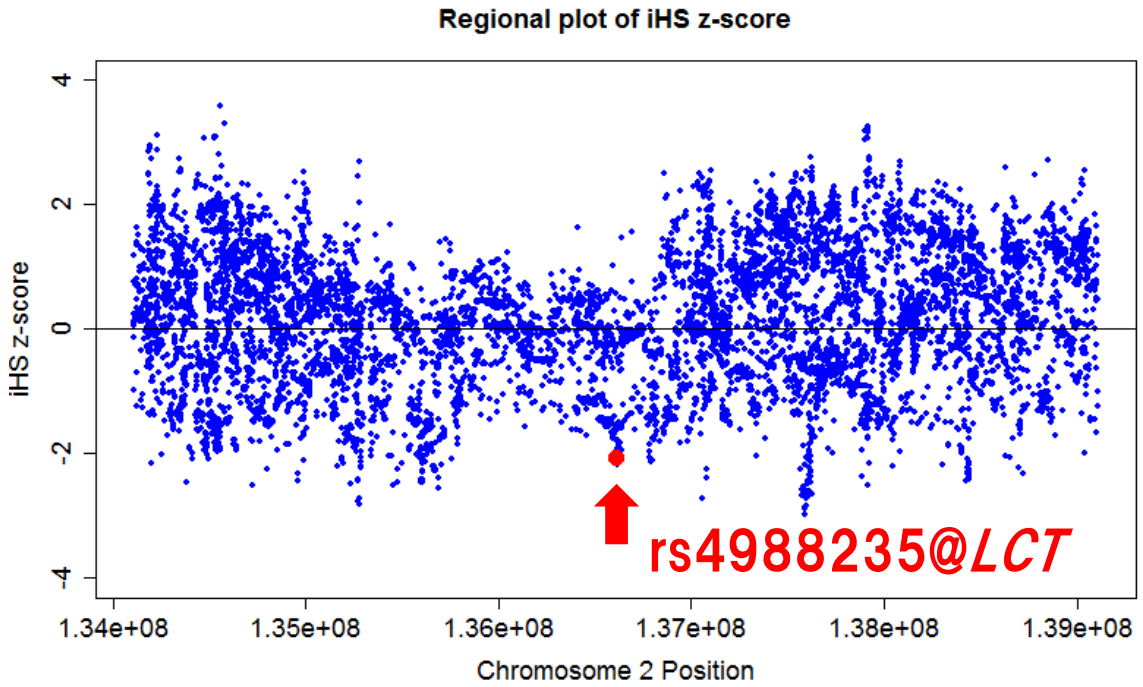
乳糖不耐性アレル：G
乳糖耐性アレル：A



- 乳糖分解酵素ラクターゼ遺伝子(*LCT*)近傍のZ値を見てみましょう。
- *LCT*遺伝子の発現は、近傍SNPであるrs4988235で制御されます。
- 乳糖耐性を示すrs4988235-Aアレルは、顕著な集団間のアレル頻度差を示し、特に家畜乳を栄養源としていた欧米人集団で強い選択圧を受けてきたことが知られています。

(中山一大. ***Anthropol Sci (Jpn S)*** 2015)

③ selscanを使った選択圧解析

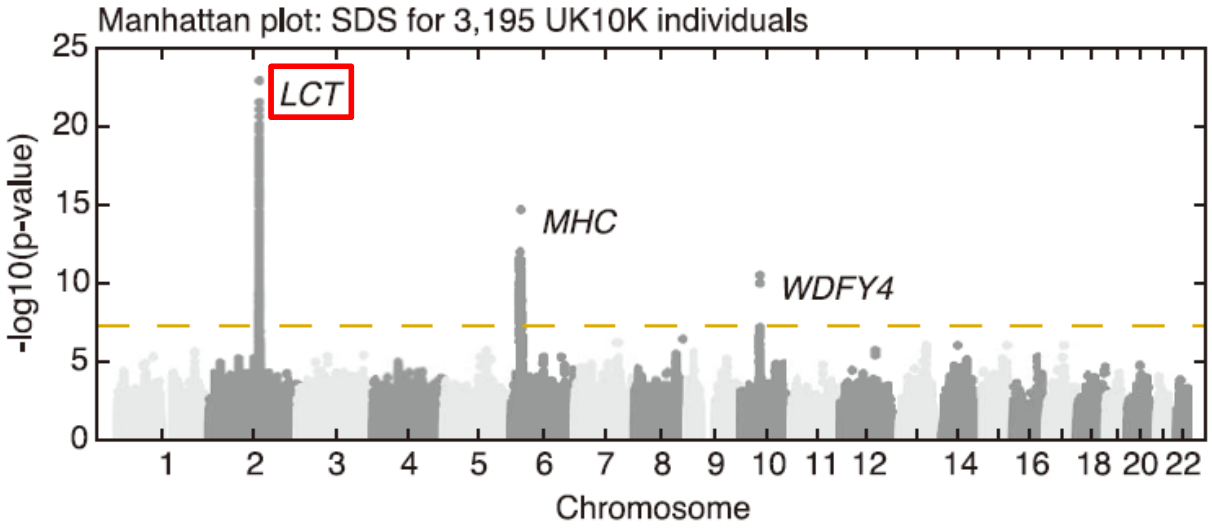


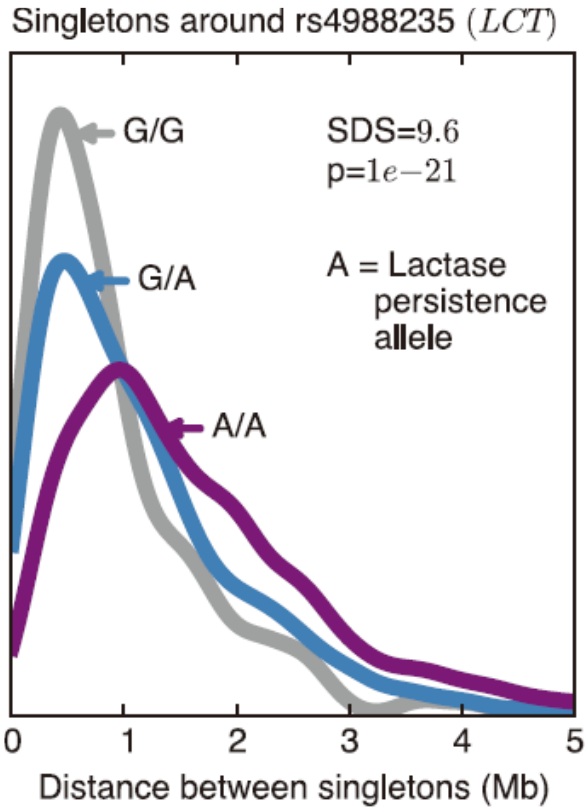
※ファイル"iHS_Plot.R"を開いて、改変の上、Rにコピー＆ペーストして下さい。

- 同SNPのiHS Z値は小さなピークを示すことが確認されました。
- しかし、集団特異的な選択圧を受けた代表的なSNPの割には、そこまで顕著なZ値を示したとはいいにくい結果です。
- これは、iHSが発表当時は画期的な手法だったものの、選択圧の検出力については十分でなかったいう事情に起因します。

③ selscanを使った選択圧解析

欧米人集団全ゲノムシークエンスデータ（3,195名）を活用した、過去数千年における選択圧解析（SDS）





- 検出力を高める目的で、様々な解析手法の開発へとつながりました。
- 現在では、*LCT*遺伝子領域は欧米人集団で最も強い選択圧を受けた領域の一つであることが確認されています。

(Field Y et al. ***Science*** 2016)

終わりに

・選択圧や適応進化について、実際のデータ解析手法の紹介とともに、簡単になぞってみました。

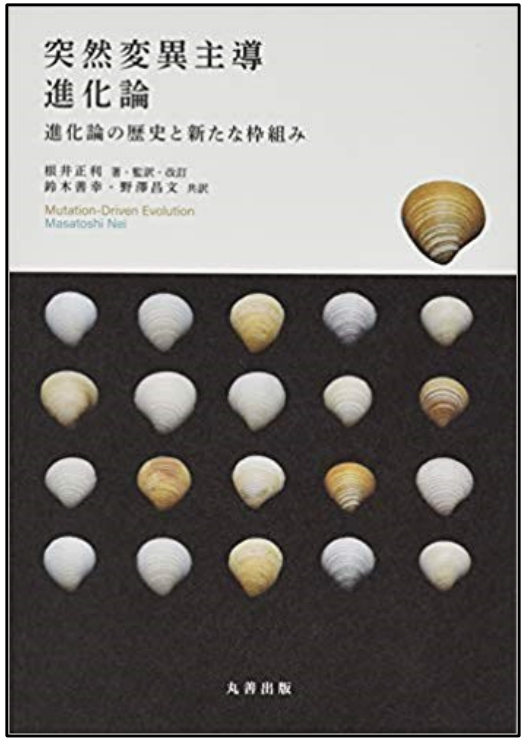
・遺伝子変異が集団内でどのように生じ、またどのように消えていくかを学ぶことは、ヒト集団の遺伝的背景の理解において重要なステップです。

・また、現代人のゲノムデータを使いながら、過去の出来事を間接的に推察するのは、なんとも楽しい作業です。

・全ゲノムシークエンスやバイオバンク規模GWASデータの活用など、最新のゲノムデータに即した選択圧解析手法が開発され、新たな局面を迎えています。

・皆さんのヒトゲノム研究においても、知見を活かしてみてください。

終わりに



突然変異主導進化論
－進化論の歴史と新たな枠組み－

根井正利

丸善出版

- 「集団遺伝学や適応進化について述べられた、お薦めの教科書はありませんか？」と、よく訊かれます。
- 同分野の大家である根井正利先生が執筆された、上記の本がお薦めです。