

2023年8月25-27日
遺伝統計学・夏の学校 講義実習資料

GenomeDataAnalysis3

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>



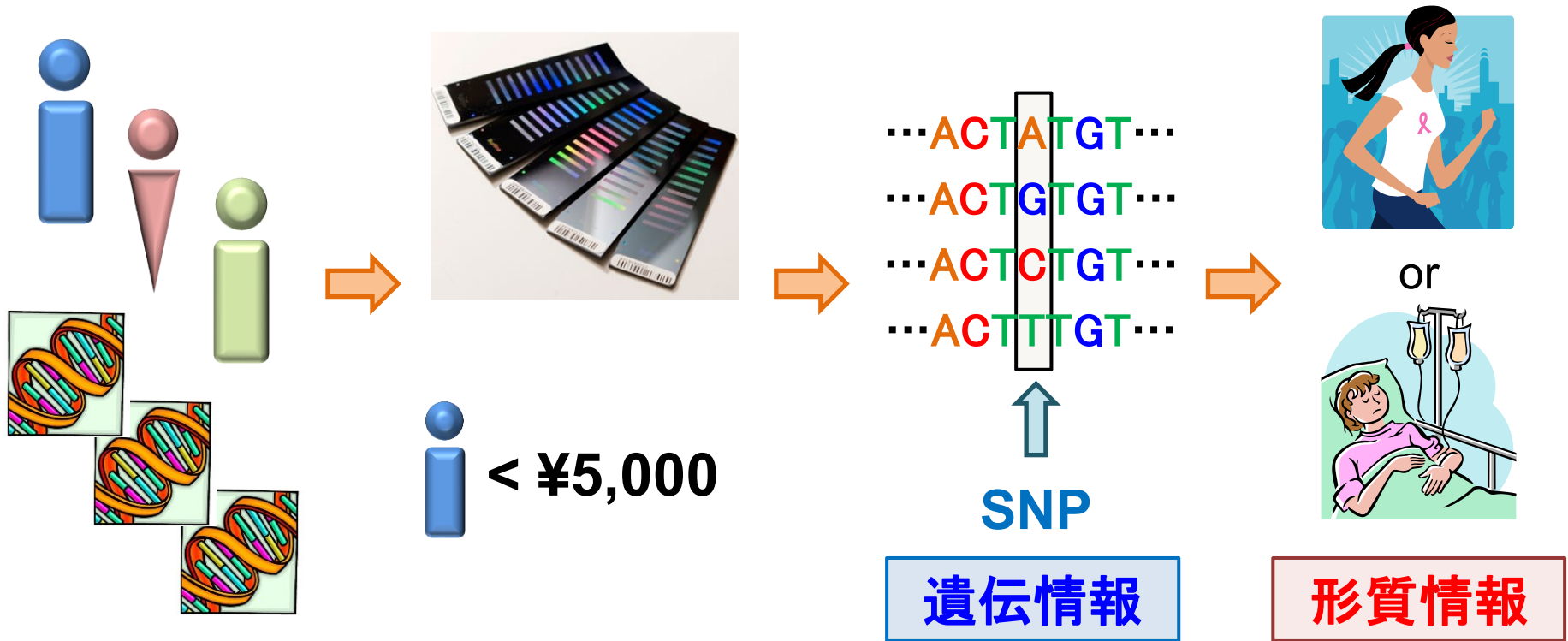
GenomeDataAnalysis3

- ① **SNP genotype imputation**
- ② **HLA imputation法**
- ③ **SNP2HLAを使ったHLA imputation法**

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 SNP genotype imputation

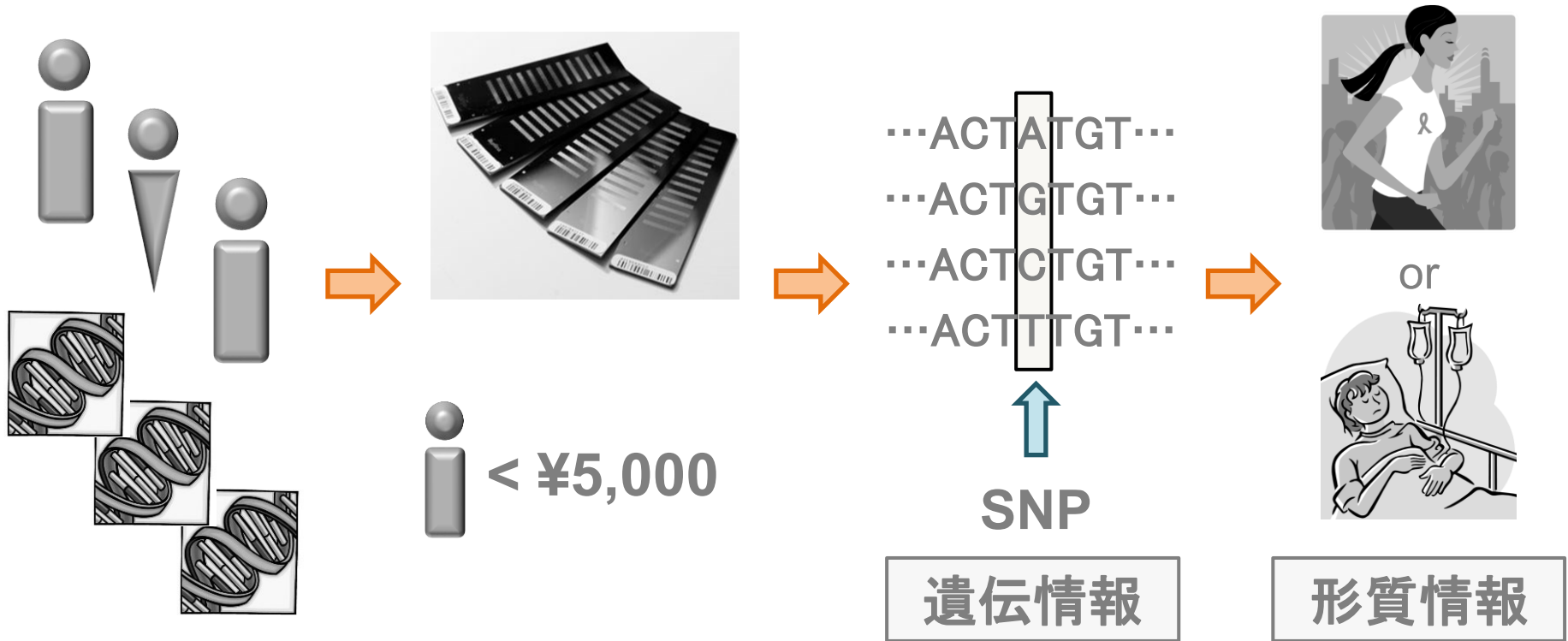
ゲノムワイド関連解析(GWAS)



- 遺伝情報と形質情報との結びつきを評価する遺伝統計学の手法。
- 数十万人を対象に、ヒトゲノム全体を網羅する数千万箇所のSNPのタイピングを実施し、対象形質との関連を評価する手法。
- 2002年に日本の理化学研究所で世界に先駆けて実施された。

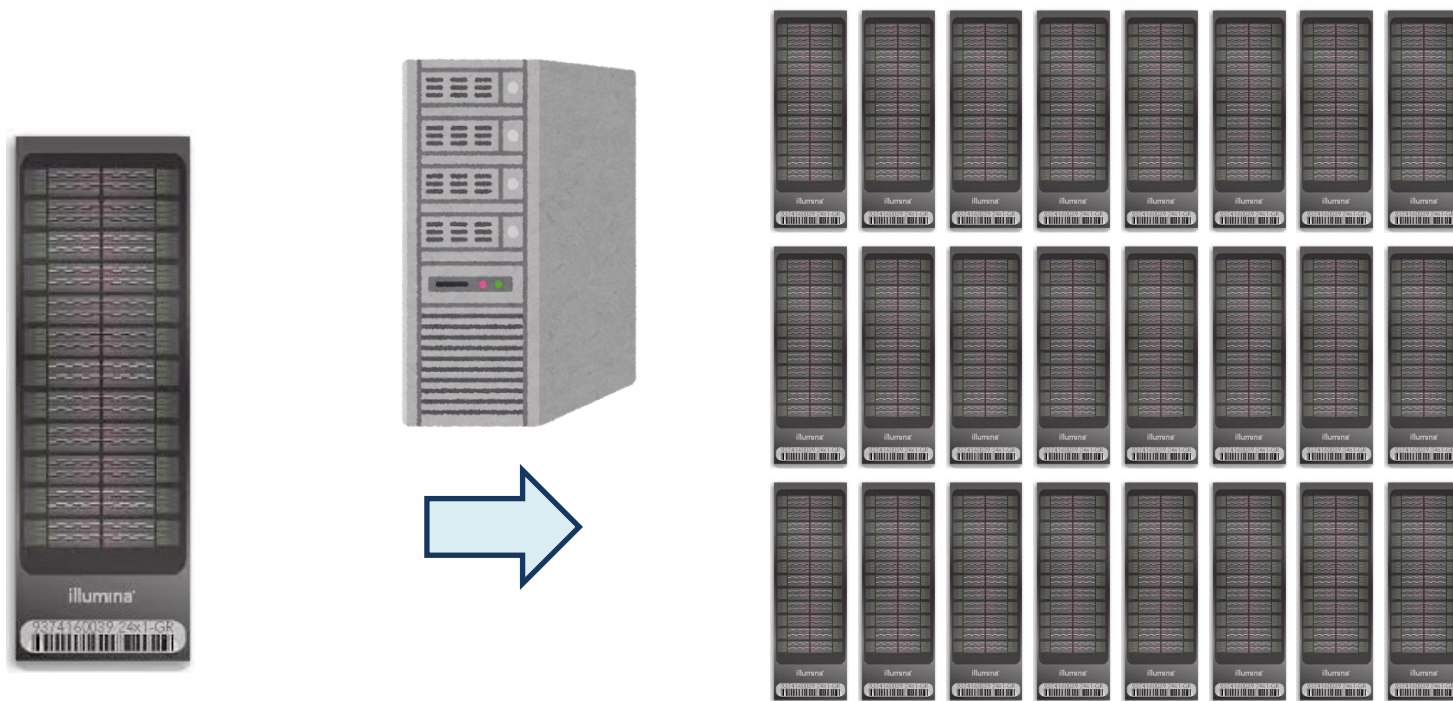
① SNP genotype imputation

ゲノムワイド関連解析(GWAS)



- 遺伝情報と形質情報との結びつきを評価する遺伝統計学の手法。
- 数十万人を対象に、ヒトゲノム全体を網羅する**数千万箇所のSNP**のタイピングを実施し、対象形質との関連を評価する手法。
- 2002年に日本の理化学研究所で世界に先駆けて実施された。

① SNP genotype imputation



- SNPマイクロアレイによるジェノタイピングは、**数十万箇所**が対象です。
- ゲノムワイド関連解析に使用されるのは、**数千万箇所**のSNPです。
- **未観測のSNPジェノタイプをコンピューター上で推測**することで、数十万SNPから数千万SNPの情報を得ています。
- この作業を、“**SNP genotype imputation**”といいます。

① SNP genotype imputation



- SNP genotype imputationは、疾患ゲノム解析の一般的な手法です。
- SNP genotype imputationの普及により、マイクロアレイに搭載する1サンプル分のSNP数が頭打ちになり、価格低下に貢献しました。
- ゲノム全体のSNPを高精度に推定する場合、30-50万程度のSNPで十分と考えられています。

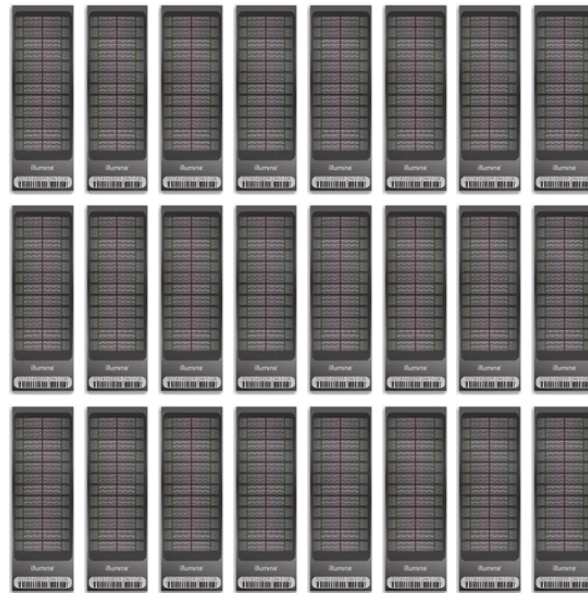
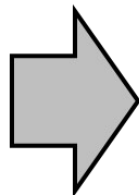
① SNP genotype imputation

Imputationのメリット①：解析コストの削減

Illumina Asian
Screening Array



Genotype
imputation



~500,000 genotyped SNPs
~\$40 per a sample

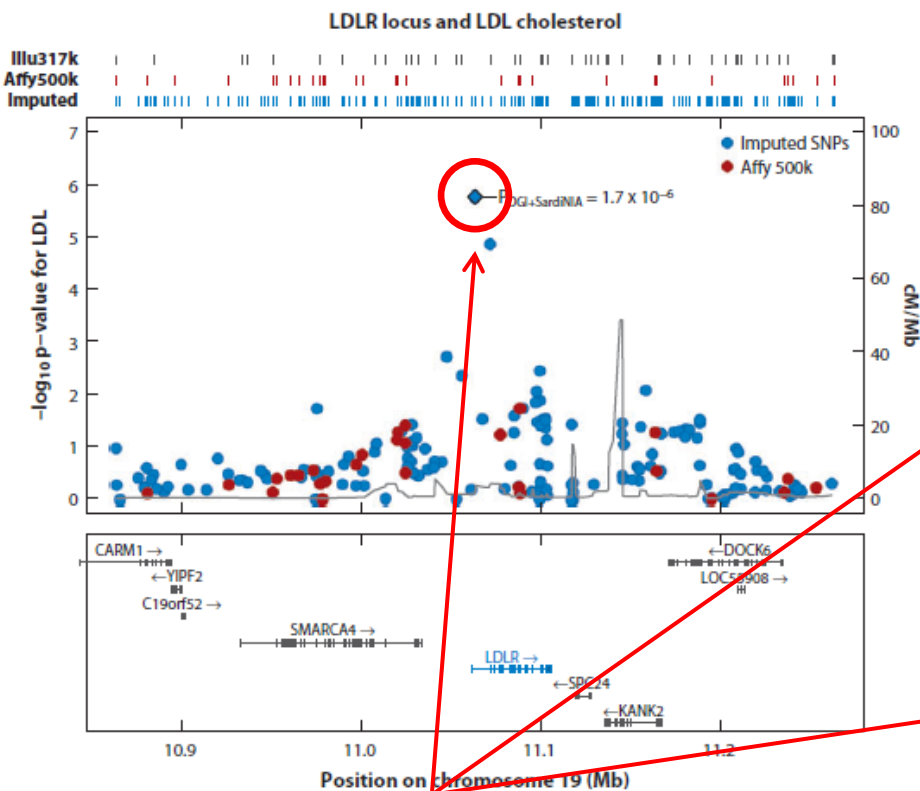
~7,500,000 imputed SNPs
(without additional costs)

- SNP genotype imputationを実施することで、追加コストをかけることなく、多数のSNPの情報を取得することが可能になりました。
- SNP解析のコスト削減と、対象サンプル数の増加に貢献しています。⁷

① SNP genotype imputation

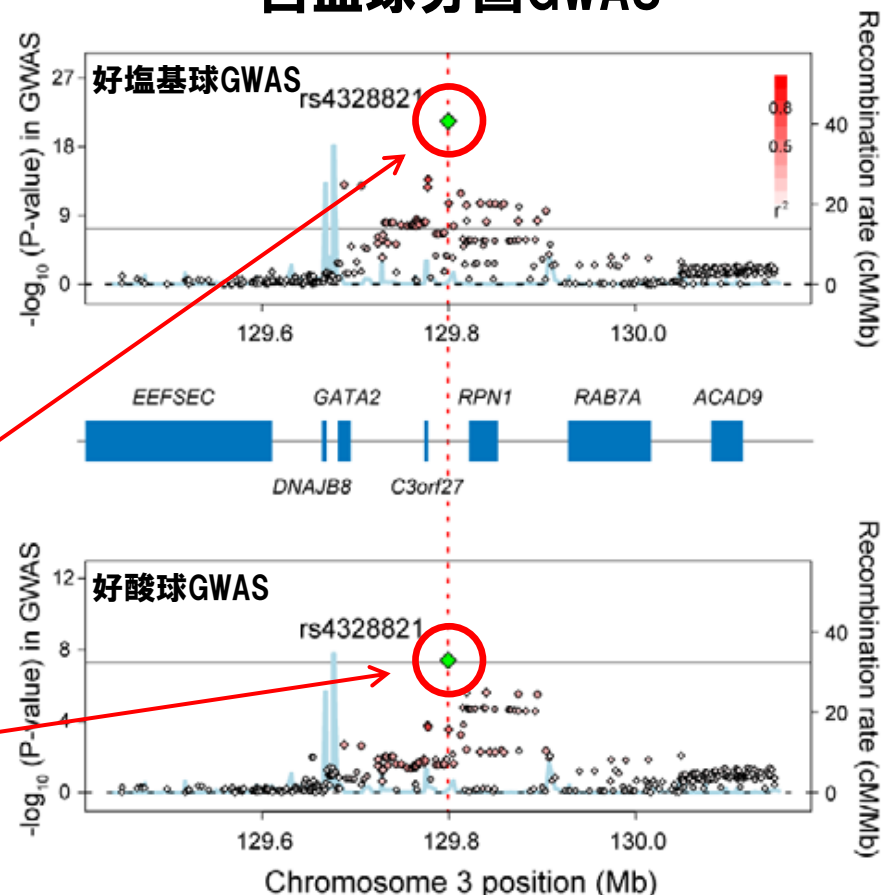
Imputationのメリット②：原因SNPのfine-mapping

LDLコレステロールGWAS



Imputationで新たに得られた原因SNP

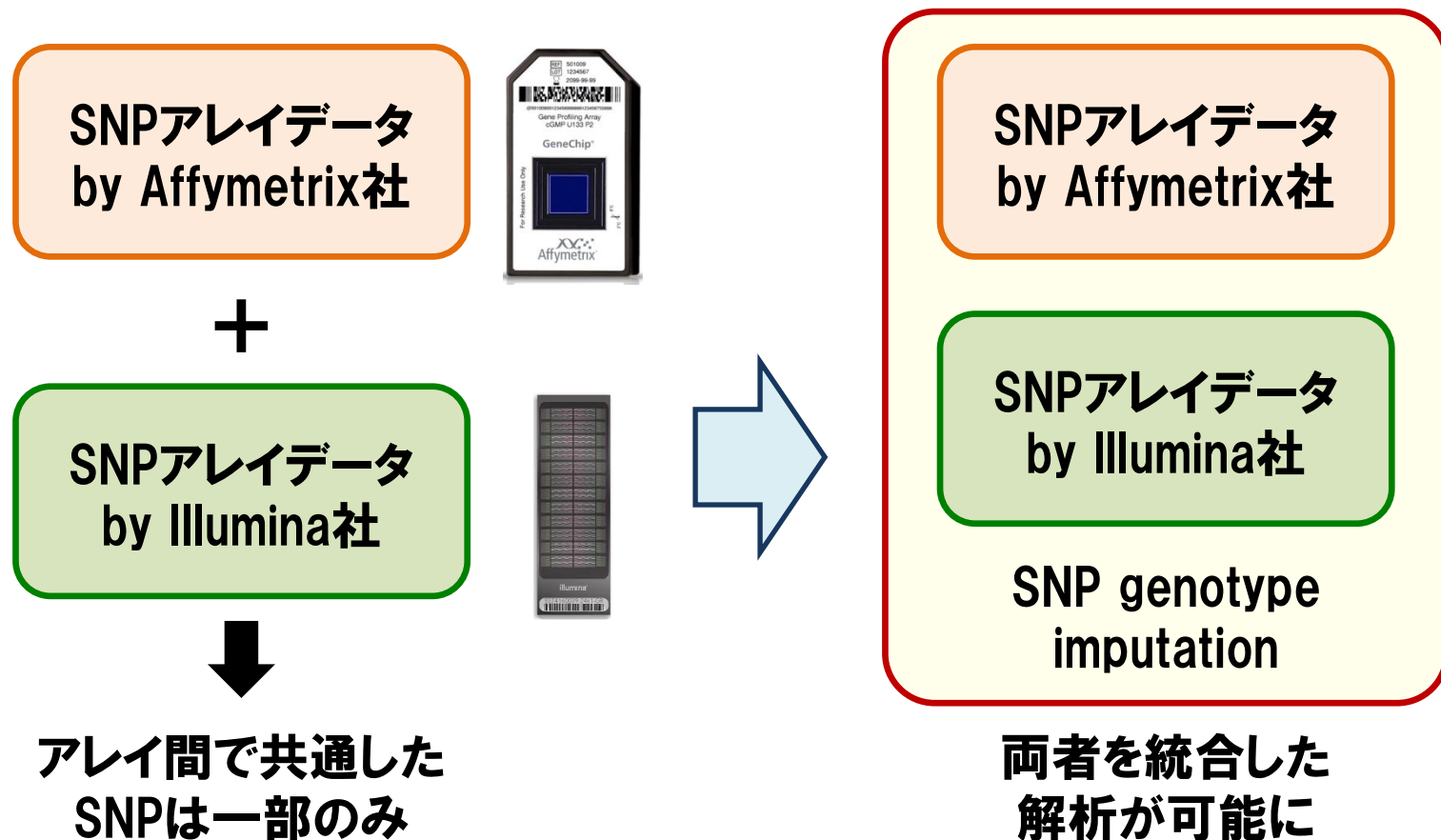
白血球分画GWAS



- Imputationで解析対象SNP数が増えた結果、より有意な関連を示す(傾向のある)原因SNPが同定(fine-mapping)されることがあります。

① SNP genotype imputation

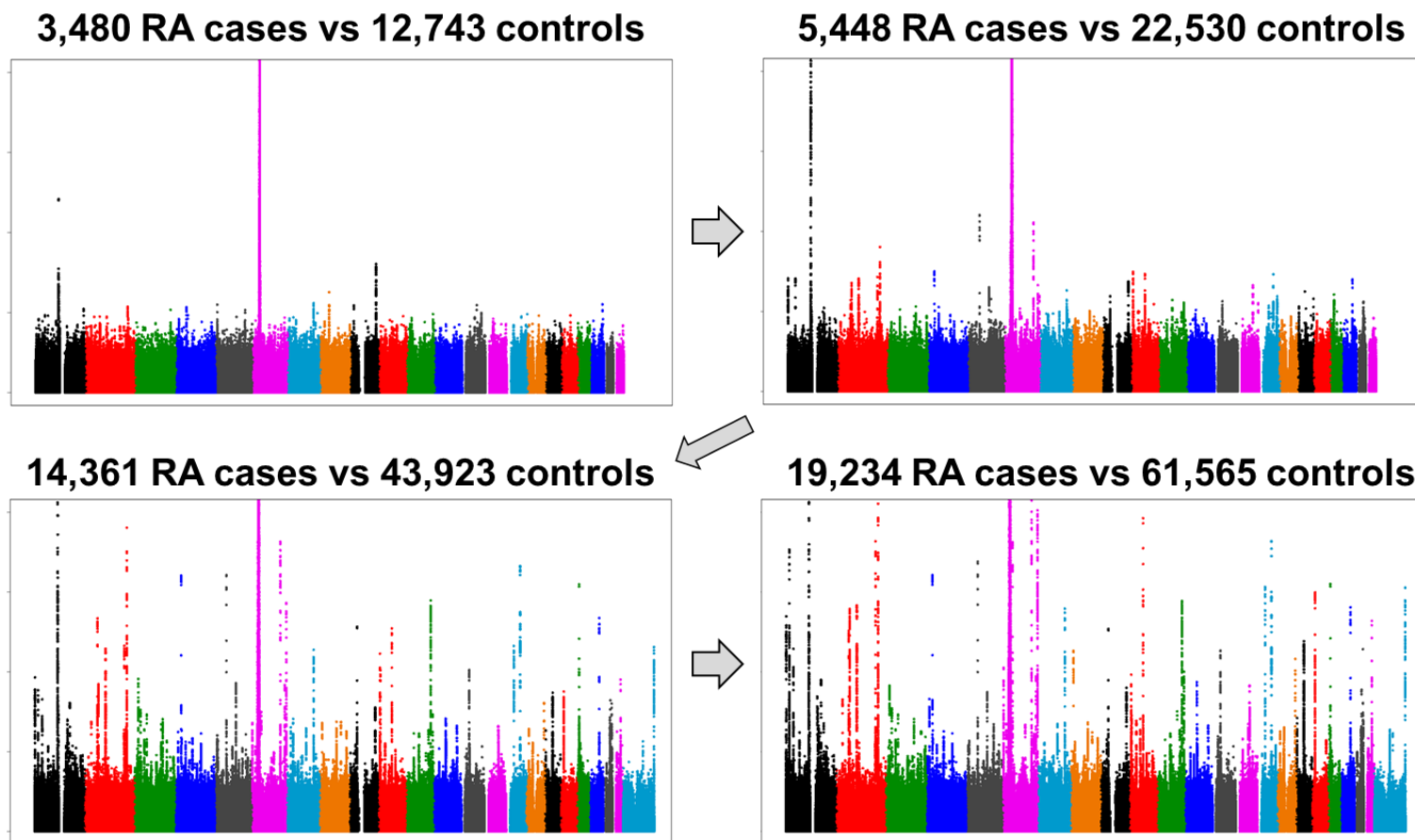
Imputationのメリット③：異なるSNPマイクロアレイデータの統合



- SNP genotype imputationを実施することで、異なるSNPマイクロアレイで得られたGWASデータのメタアナリシスが、可能になります。

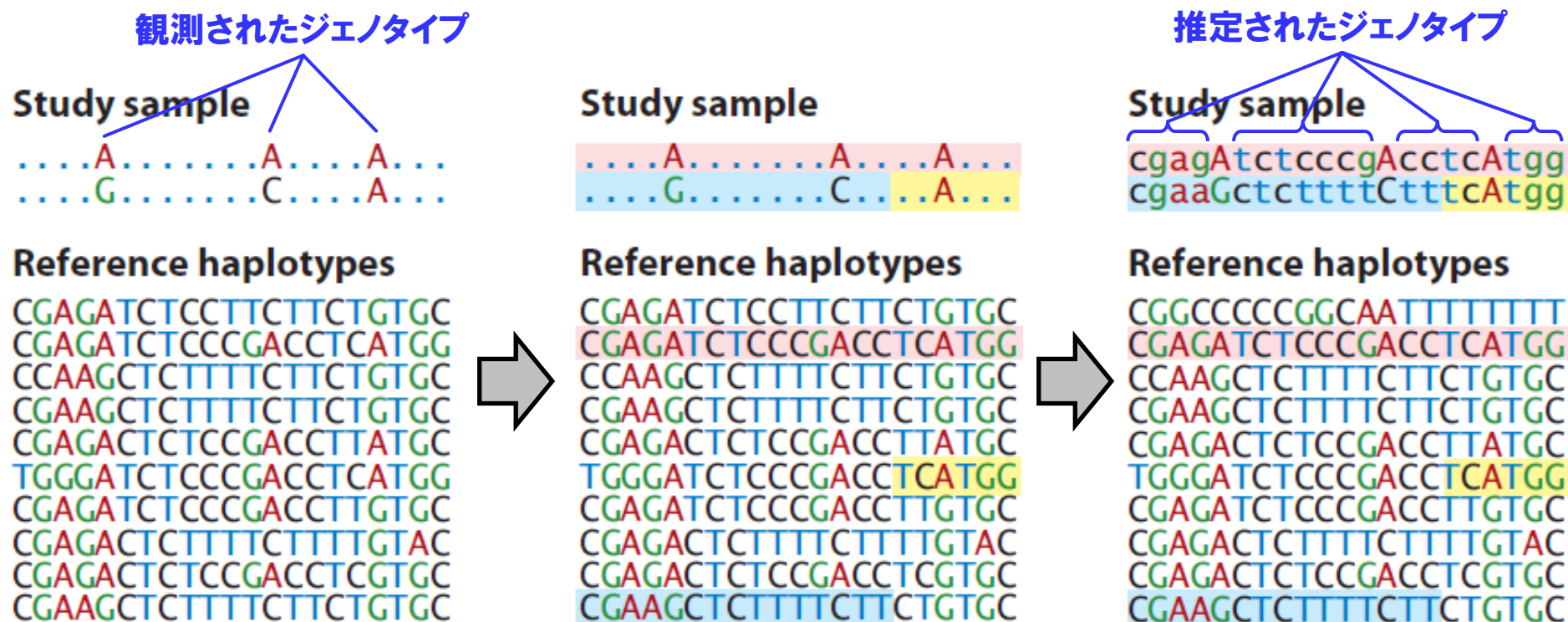
① SNP genotype imputation

Imputationのメリット③：異なるSNPマイクロアレイデータの統合



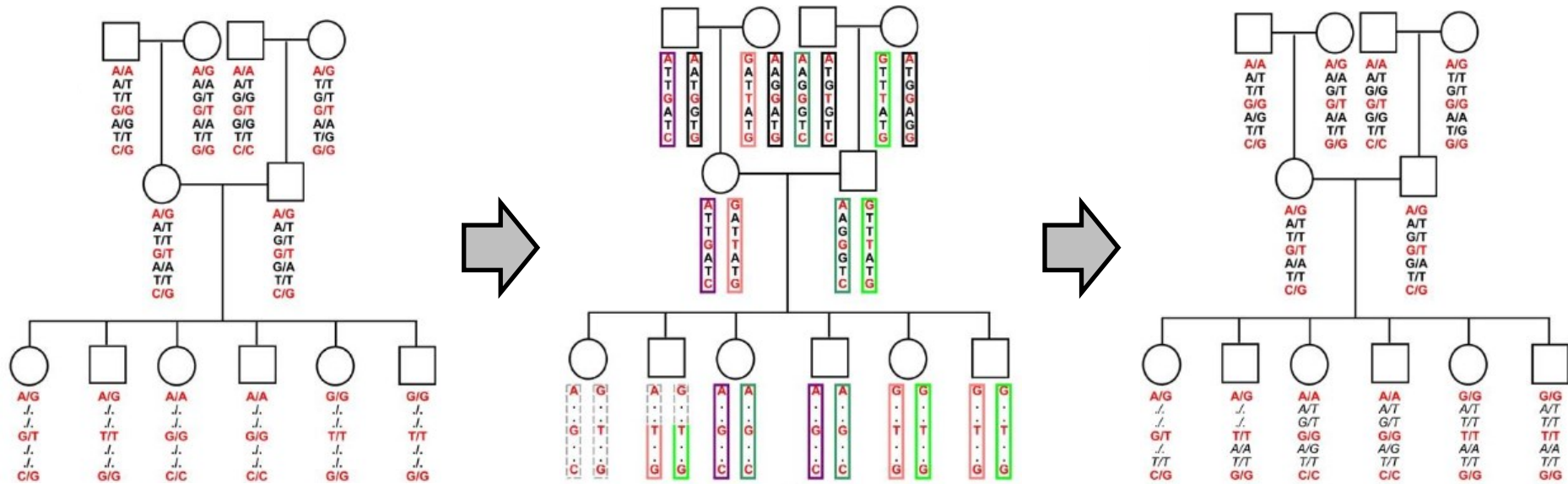
- Imputationを通じて複数のGWASのメタアナリシスを実施することで**検出力が増加**し、多数の疾患感受性遺伝子変異の同定に繋がります。

① SNP genotype imputation



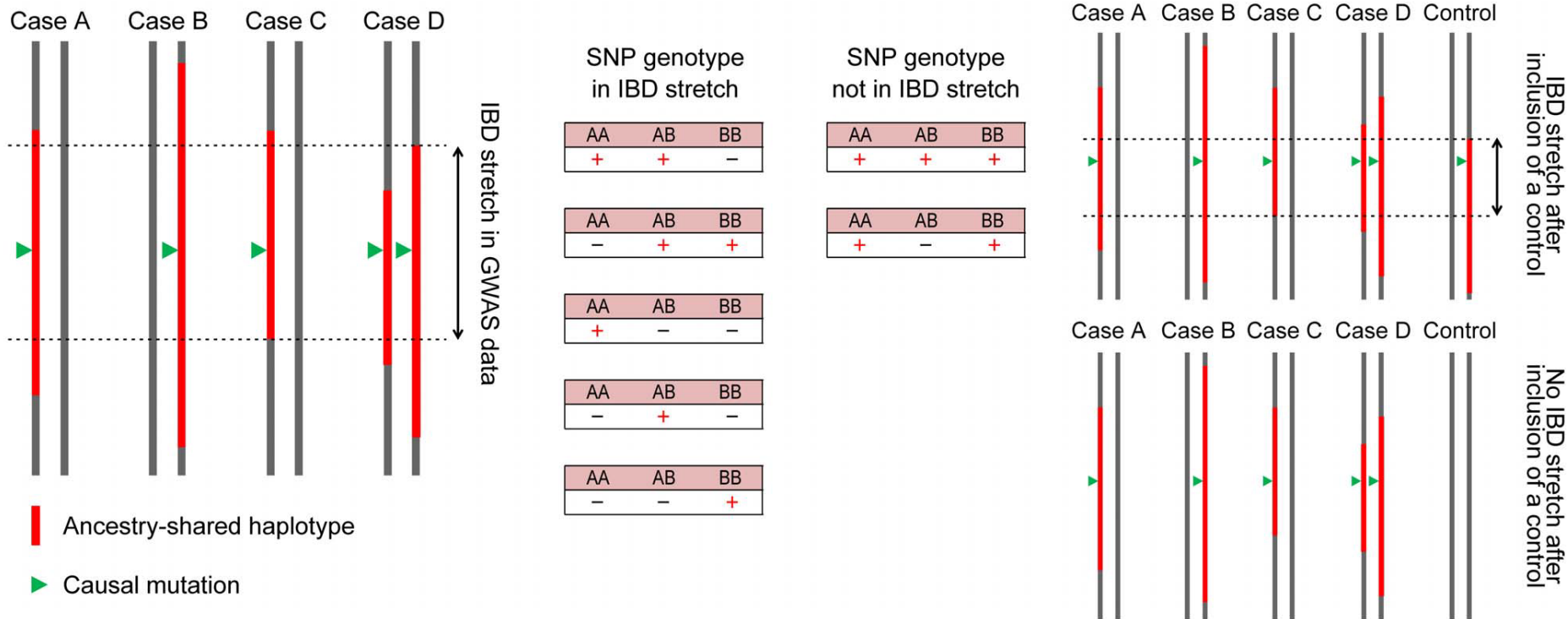
- SNP genotype imputationの実施に際しては、予め高密度のSNPハプロタイプ情報を搭載した**参照データ**(reference panel)が必要です。
- 参照データ中のSNPハプロタイプを、GWASデータのSNP情報に充てていくことで、未観測のSNPジェノタイプを推定します。

① SNP genotype imputation



- SNP genotype imputationは、元々は家系データ解析まで遡ります。
- 実際、SNP genotype imputationのアルゴリズムと、家系ハプロタイプ推定アルゴリズムは、よく似ています。

① SNP genotype imputation



- 通常のSNP genotype imputationは、集団中で頻度の高いコモンバリエーションを対象としていて、レアバリエーションのimputationは苦手です。
- 家系データを対象にすることで、レアバリエーションのimputationに特化した解析方法もあります。

① SNP genotype imputation

直接観測された SNPデータ

	AA	AG	GG
Sample A	1	0	0
Sample B	0	1	0
Sample C	0	0	1

	A	G
Sample A	2	0
Sample B	1	1
Sample C	0	2

Imputationで 得られた SNPデータ

	AA	AG	GG
Sample A	0.95	0.04	0.01
Sample B	0.05	0.90	0.05
Sample C	0.05	0.10	0.85

	A	G
Sample A	1.94	0.06
Sample B	1.00	1.00
Sample C	0.20	1.80

- SNP genotype imputationは、各サンプルにおいて、各ジェノタイプ毎の存在確率および各アレル毎の存在確率を、統計的に推定します。
- ジェノタイプやアレルを一意に決定することは難しく、結果として、ジェノタイプやアレルの本数が、整数ではなく小数で得られます。

① SNP genotype imputation

精度の高い Imputation SNPデータ

	AA	AG	GG	Rsq
Sample A	0.95	0.04	0.01	0.90
Sample B	0.05	0.90	0.05	
Sample C	0.05	0.10	0.85	

	A	G	Rsq
Sample A	1.94	0.06	0.90
Sample B	1.00	1.00	
Sample C	0.20	1.80	

精度の低い Imputation SNPデータ

	CC	CT	TT	Rsq
Sample A	0.50	0.25	0.25	0.30
Sample B	0.40	0.40	0.20	
Sample C	0.35	0.05	0.60	

	C	T	Rsq
Sample A	1.25	0.75	0.30
Sample B	1.20	0.80	
Sample C	0.75	1.25	

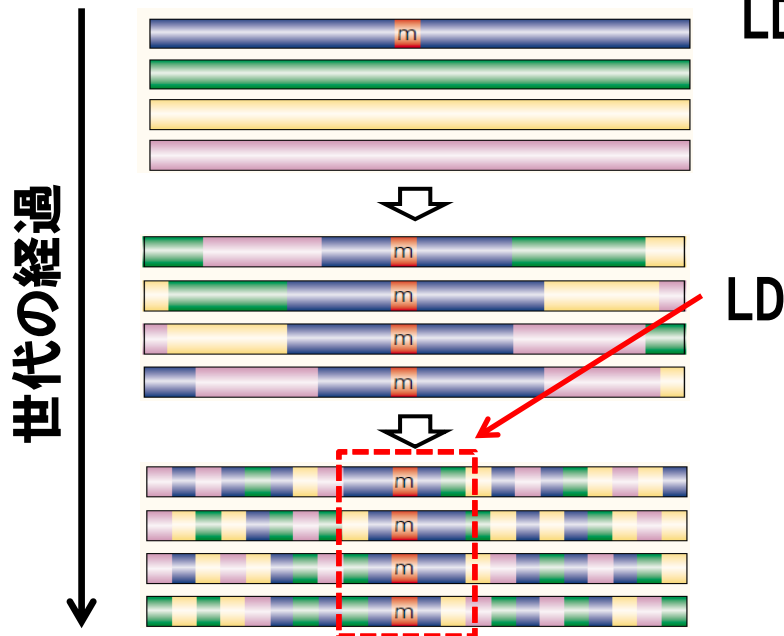
※0~1.0の範囲をとり、大きい値が高い推定精度を示す。

※Imputation精度の指標が一定閾値以下の場合、解析対象から除外。

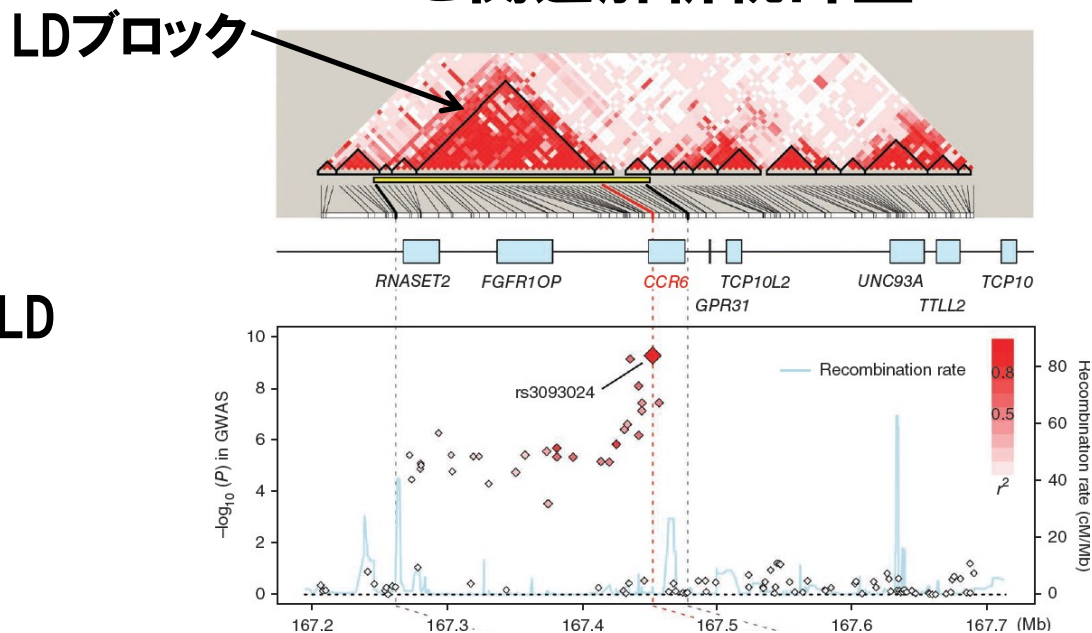
- 推定精度はSNP毎に異なり、うまくいくことも、いかないこともあります。
- 各SNPに対して、**imputation結果の精度を表す指標**が出力されます。
- 例：真のジェノタイプと推定ジェノタイプの相関の決定係数(=Rsq)。
- **推定精度が低いSNPは、解析の対象外とします。**

① SNP genotype imputation

染色体組換えとLD

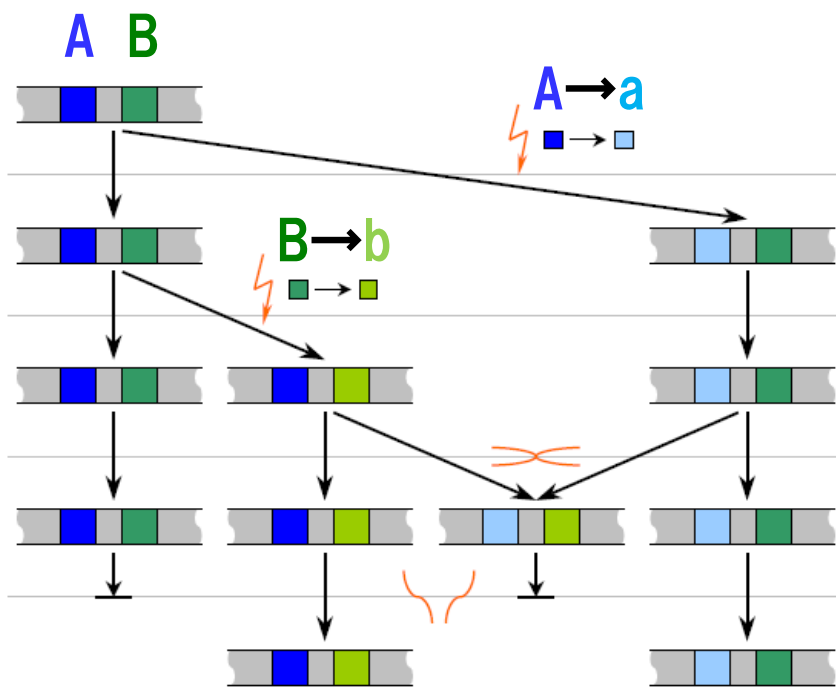


LDと関連解析統計量



- 近接する複数のSNPのジェノタイプは、集団中で非独立の分布をとることが多く、連鎖不平衡(Linkage Disequilibrium: LD)と呼ばれます。
- 近接するSNPのジェノタイプの分布が似通っている状態、ということです。
- LD関係にあるSNP同士は、関連解析の統計量も似ています。
- 参照データで観測されたSNP間のLD関係を、GWASデータに適用することで、imputationが実施されます。

① SNP genotype imputation



LD指標の計算式

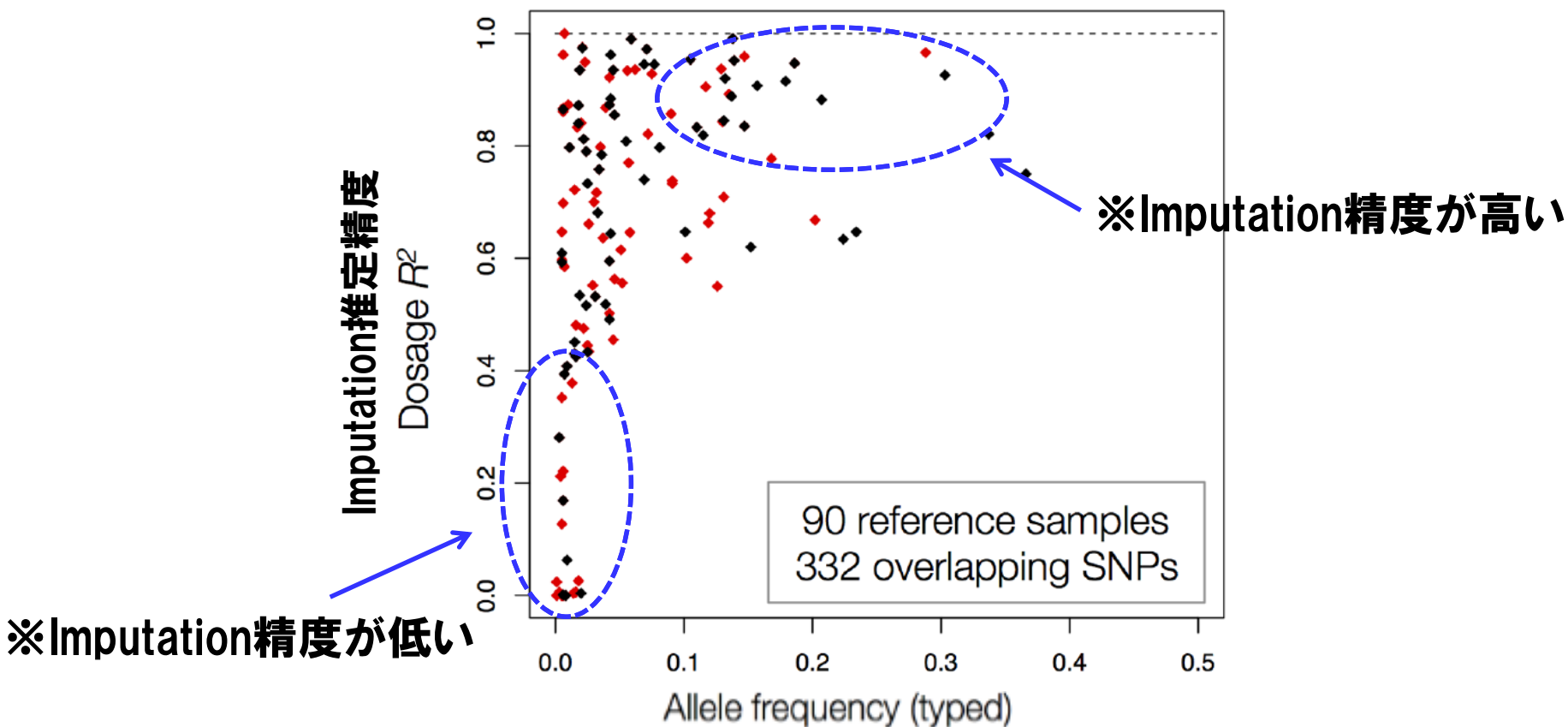
$$D = h_{AB} - p_A p_B = h_{AB} h_{ab} - h_{Ab} h_{aB}.$$

$$D' = \frac{D}{D_{\max}} = \frac{D}{\min(h_{AB} + h_{Ab}, h_{AB} + h_{aB}) - p_A p_B}.$$

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}.$$

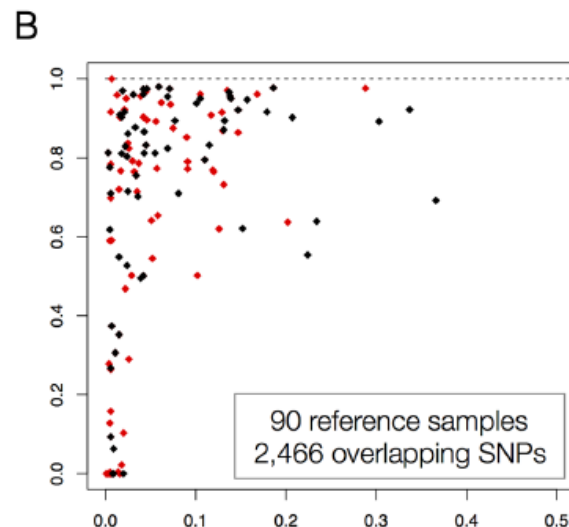
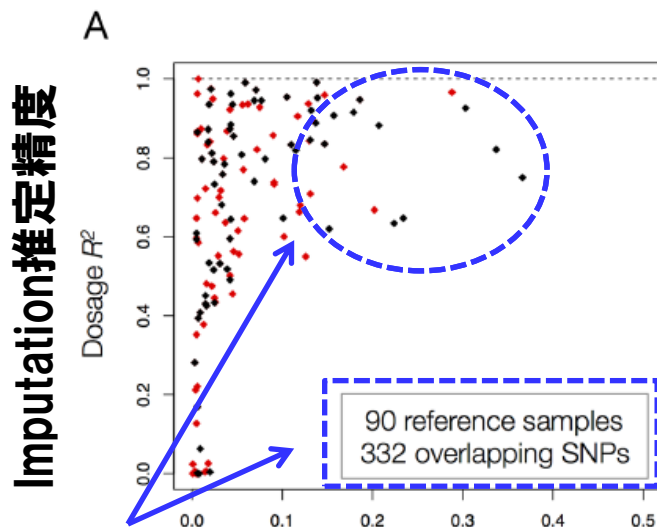
- LD = ”ハプロタイプ頻度が構成アレルの頻度の積と異なる状態”
- 2SNP間のハプロタイプ頻度とアレル頻度から、LDの程度を表す指標(r^2 、 D 、 D')を計算することが可能です。
- LD指標の一つ「 r^2 」は、0~1.0の値をとり、2SNP間のハプロタイプ分布(≡ジェノタイプ分布)の相関の決定係数や関連統計量の比に相当します。

① SNP genotype imputation

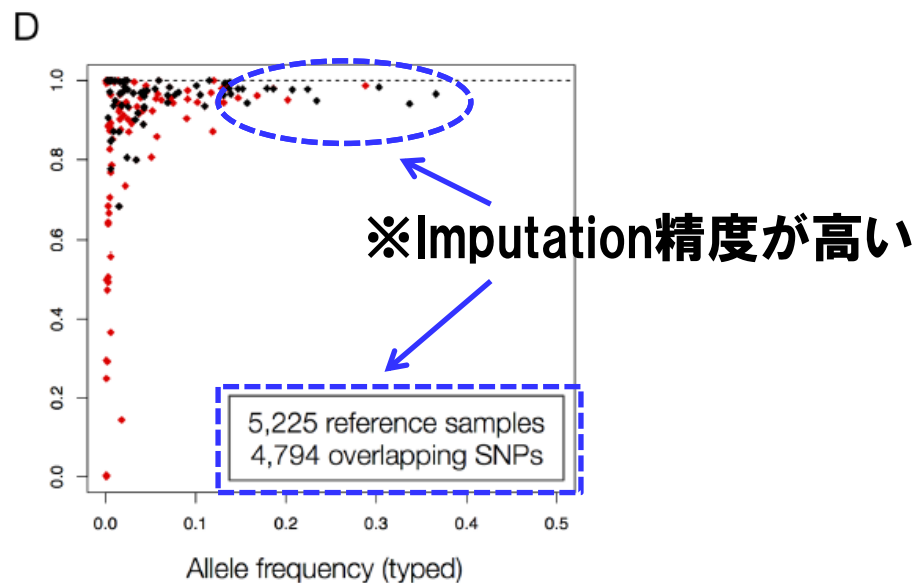
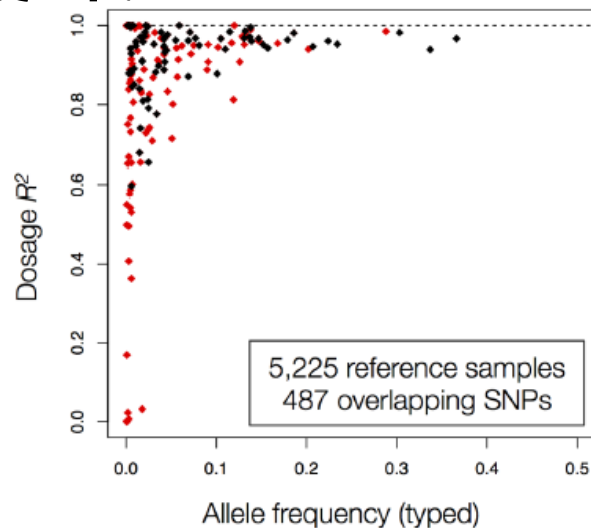


- 一般に、アレル頻度が大きく、周辺のSNPと強い連鎖不平衡関係にあるSNPは、imputation推定精度が高くなります。
- 逆に、アレル頻度が小さく、周辺のSNPと連鎖不平衡関係に乏しいSNPは、imputation推定精度が低くなります。

① SNP genotype imputation



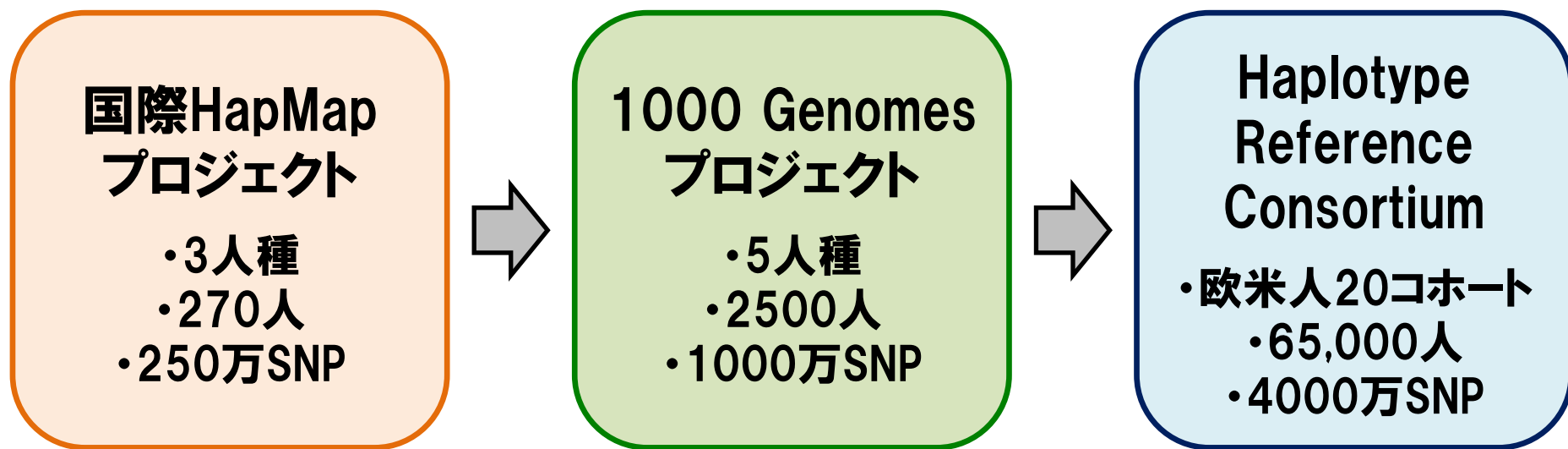
※Imputation精度が低い



• 参照データのSNP密度が高く、サンプル数が多いほど、imputation推定精度が高くなります。

① SNP genotype imputation

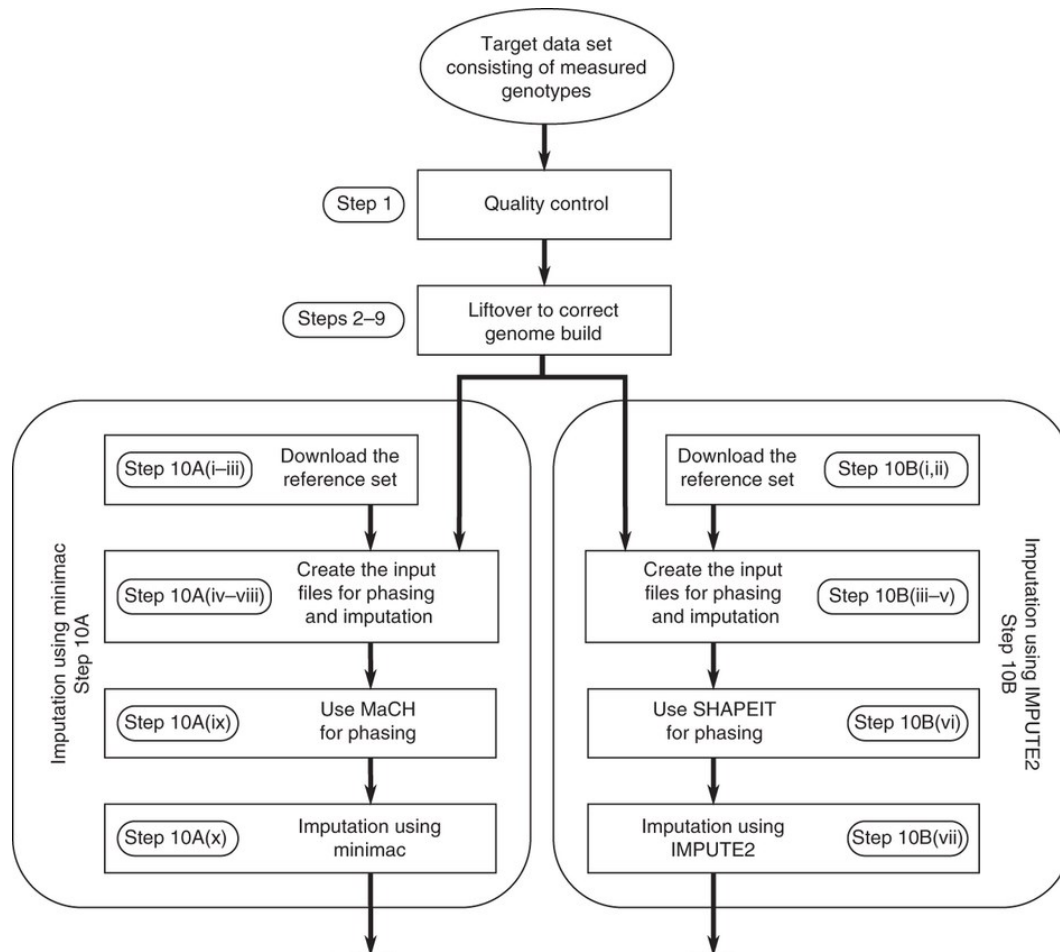
Imputation参照データの変遷



- ・SNP genotype imputationで推奨される参照データも、**高密度化、多サンプル化、多国籍化**、が進んでいます。
- ・日本人集団GWASについては、日本人集団が含まれる1000 Genomesプロジェクトを参照データに使うのが入門的です。

① SNP genotype imputation

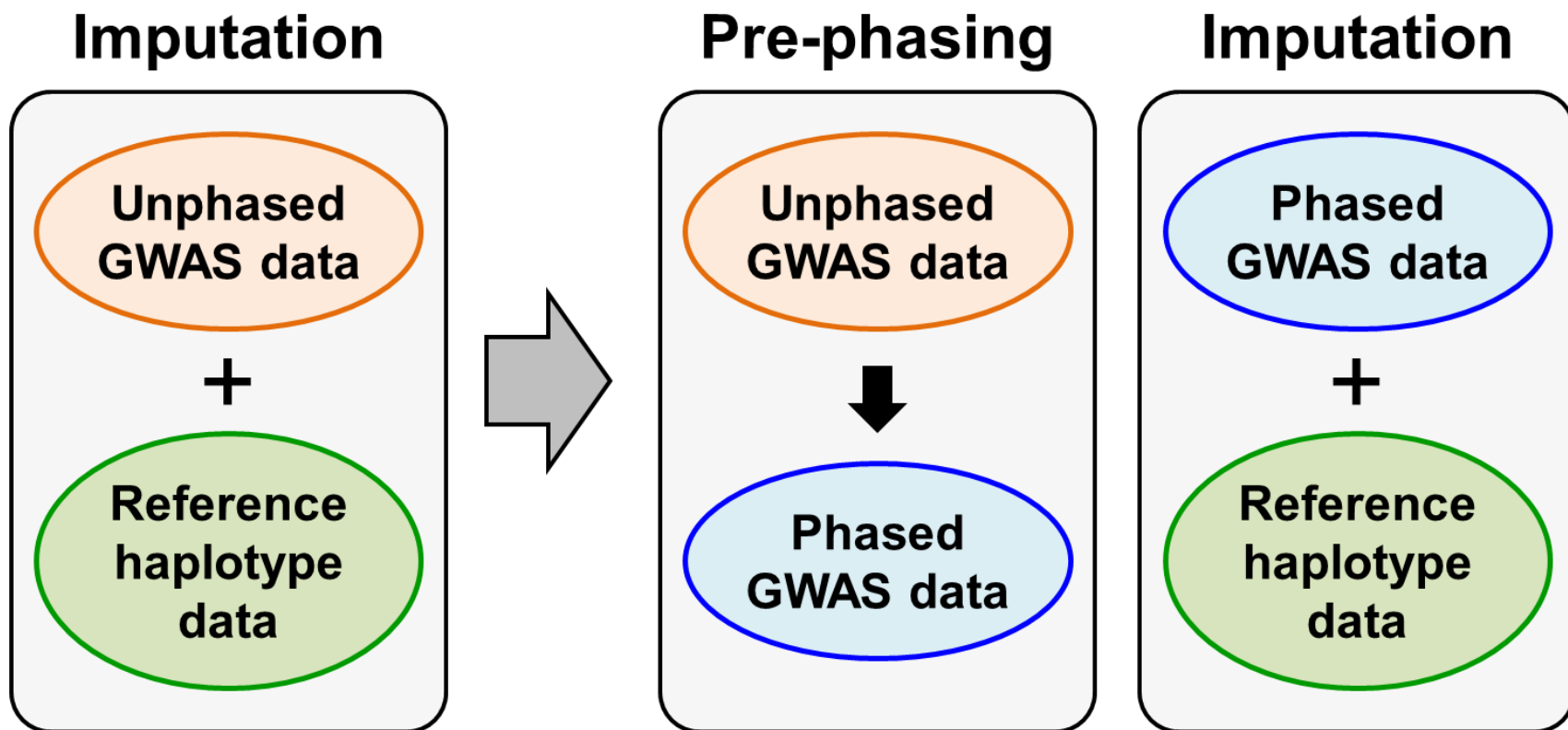
SNP genotype imputationの実施方法



• Imputationの実施方法については、各種プロトコルを参照して下さい。

① SNP genotype imputation

Imputation方法の変遷



- 当初、GWASデータのimputationは1ステップで実施されていました。
- 数十万人規模のGWASデータを扱う必要性から、**GWASデータの phasing → imputation**と、**2ステップの作業**へと変化しています。

① SNP genotype imputation




TopMed Imputation Server

TOPMed Imputation Server
Free Next-Generation Genotype Imputation Service

[Sign up now](#) [Login](#)

20.3M Imputed Genomes	1768 Registered Users	3 Running Jobs
--------------------------	--------------------------	-------------------

The easiest way to impute genotypes

- 
Upload your genotypes to our secured service.
- 
Choose a reference panel. We will take care of pre-phasing and imputation.
- 
Download the results.
All results are encrypted with a one-time password. After 7 days, all results are deleted from our server.

<https://imputation.biodatacatalyst.nhlbi.nih.gov/>

•大規模参照データに基づく**Imputation server**を構築し、各自がGWASデータをアップロードしてimputationするシステムも構築されています。

(Taliun D et al. *Nature* 2021)²³

① SNP genotype imputation

SNP genotype imputationのツール

○: GWAS data pre-phasing

- **Beagle5** <https://faculty.washington.edu/browning/beagle/beagle.html>
(Browning BL et al. *Am J Hum Genet* 2018)
- **SHAPEIT5** <https://odelaneau.github.io/shapeit5/>
(Hofmeister RJ et al. *Nat Genet* 2023)
- **Eagle** <https://alkesgroup.broadinstitute.org/Eagle/>
(Loh PR et al. *Nat Genet* 2016)

○: Imputation

- **Minimac4** <http://genome.sph.umich.edu/wiki/Minimac4>
(Howie B et al. *Nat Genet* 2012)
- **Impute5** <https://jmarchini.org/software/#impute-5>
(Bycroft C et al. *Nature* 2018)

• SNP genotype imputationの一般的なツールです。

① SNP genotype imputation



Imputationを実施したら、**凄い(=有意な)結果**が出てきました！



Imputationソフトを動かすことと、**imputationを正しく実施することは、別です**。凄い結果が本当に凄いのか、間違いなのか、確認することが大事です。

- Imputationに限らず、ソフトウェアを実行すると解析結果が得られます。しかし、得られた結果が正しいことを意味するわけではありません。
- しばしば、**間違っただけの結果ほど凄い結果に見える**、ことがあります。
- 正しく解析を実施できたか、別の観点から確認する必要があります²⁵。

① SNP genotype imputation

SNP genotype imputationのTips

- SNP表記のstrand(+/-)を、GWASと参照データで一致させる。
- GWASと参照データでアレル頻度が(理由なく)著しく異なるSNPを、予め除外。
- Quality Controlを実施後のGWASデータで、imputationを実施。
- ケース群、コントロール群を一緒にimputationする。
- Imputation実施後に、推定精度の低いSNPを除外。
- Imputationの前後で、直接観測されたSNPのジェノタイプが変化する(=修正される)ことがあります。
- Imputation後の関連解析には、ジェノタイプ/アレルの存在確率を使用。
- Imputation後のデータ解析で有意な関連を示すSNPが同定された場合、imputationの精度が低いことが原因でないか確認する。

• Imputationを実施時の**チェックポイントの確認**を、忘れずに。

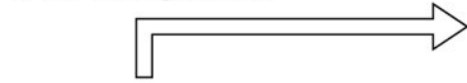
GenomeDataAnalysis3

- ① SNP genotype imputation
- ② HLA imputation法
- ③ SNP2HLAを使ったHLA imputation法

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 HLA imputation法

MHC領域



古典的HLA遺伝子 (クラスI)

HLA-A
HLA-B
HLA-C

古典的HLA遺伝子 (クラスII)

HLA-DRB1
HLA-DQA1/DQB1
HLA-DPA1/DPB1

非古典的HLA遺伝子

HLA-DM/DO

HLA様遺伝子

MICA, MICB

発症
リスク



自己免疫疾患

関節リウマチ
炎症性腸疾患
バセドウ病/橋本病

アレルギー性疾患

喘息、薬剤性皮疹

悪性腫瘍

血液腫瘍、肺癌

精神疾患

統合失調症、双極性障害

感染症

腸チフス、HIV

ゲノムワイド関連解析

- 6番染色体短腕の**主要組織適合遺伝子複合体**(major histocompatibility complex; MHC)領域は、免疫関連疾患、悪性腫瘍、精神疾患、感染症等の多彩なヒト疾患に対するリスクとの強い関連を有します。
- MHC領域内は構造が複雑であり、複数の**ヒト白血球抗原**(human leukocyte antigen; HLA)遺伝子が存在するため、**感受性遺伝子変異の同定**(fine-mapping)が困難でした。

② HLA imputation法

クラスI HLA遺伝子:

HLA-A
HLA-C
HLA-B

クラスII HLA遺伝子:

HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1

HLA様遺伝子:

MICA, MICB

2-digit alleles:

HLA-DRB1*04
HLA-DRB1*09

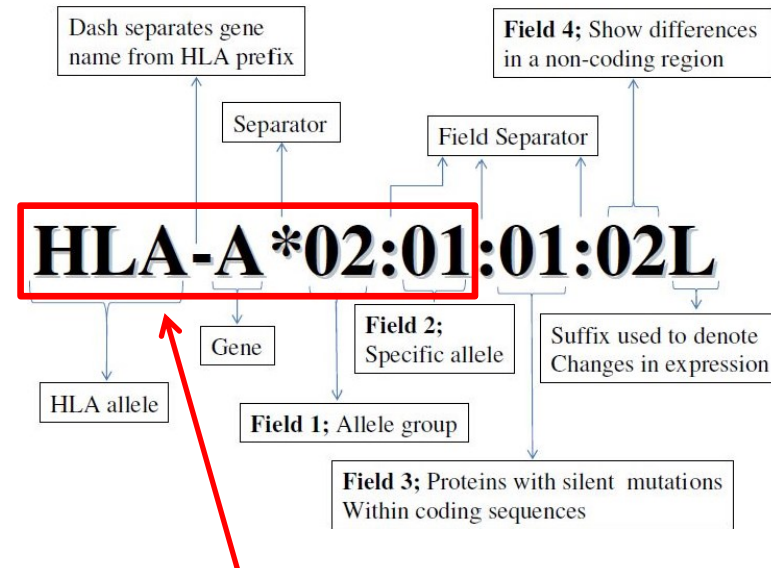
4-digit alleles:

HLA-DRB1*04:01
HLA-DRB1*04:05
HLA-DRB1*04:10
HLA-DRB1*09:01

Amino acid sequences:

...FL**E**Q**V**K**H**ECHF...
...FL**E**Q**V**K**H**ECHF...
...FL**E**Q**V**K**H**ECHF...
...FL**K**Q**D**K**F**ECHF...

HLA遺伝子多型の命名方法



※4桁表記がアミノ酸配列に対応

- MHC領域には、白血球の血液型を決めるHLA遺伝子が複数存在し、領域内の疾患リスクを説明すると考えられています。
- 各HLA遺伝子が、多数のHLAアレル(2-digit/4-digit classical alleles)やHLAアミノ酸配列多型を持つことが、解析の障壁となっていました。²⁹

② HLA imputation法

2-digitアレル

	日本人集団でのアレル頻度
HLA-B*07	0.057
HLA-B*13	0.009
HLA-B*15	0.102
HLA-B*27	0.002
⋮	⋮

4-digitアレル

	日本人集団でのアレル頻度
HLA-B*15:01	0.069
HLA-B*15:07	0.010
HLA-B*15:11	0.008
HLA-B*15:18	0.014
⋮	⋮

アミノ酸配列多型

MRVTAP...TLQ**R**MYG...

MRVTAP...TLQ**S**MYG...

- 例えば、HLA-B遺伝子では、日本人集団で数十種類の4-digitアレルが報告され、各々が特有のアミノ酸配列多型に対応しています。
- 既報のHLA-B遺伝子アレルは、数千種類にもなります。
- HLAアレルの公式情報は、IMGT/HLAデータベースに登録されています。

② HLA imputation法

マルチアレル 多型の表記例

	HLA-Bアレル	HLA-Cアレル
Sample A	15:01/56:01	01:02/04:01
Sample B	13:02/13:02	03:04/06:02
Sample C	40:01/58:01	03:04/07:02
Sample D	13:02/15:02	06:02/08:01
Sample E	51:01/58:01	03:02/14:02

マルチアレル 多型のHWE検定

A_1	f_{11}			
A_2	f_{21}	f_{22}		
\vdots	\dots	\dots	\dots	
A_m	f_{m1}	f_{m2}	\dots	f_{mm}
	A_1	A_2	\dots	A_m

$$\Pr(\mathbf{f}) = \frac{n! \prod_{i=1}^m f_i!}{(2n)! \prod_{j>i} f_{ij}!} 2^{\sum_{j>i} f_{ij}}$$

$$P = \sum_{\mathbf{g} \in \mathcal{S}} \Pr(\mathbf{g}),$$

- ゲノムデータ解析の観点からは、HLA遺伝子は、3種類以上のアレルの組み合わせでジェノタイプが決まる、**マルチアレル多型**といえます。
- マルチアレル多型も、SNPに代表される**バイアレル多型**と同じく、**HWE検定、連鎖解析、関連解析等の実施が可能です。**
(ごく一部のSNPでは、マルチアレルとなっている例があります)

② HLA imputation法

薬剤副作用関連 HLA遺伝子型

・アロプリノール	B*58:01
・カルバマゼピン	A*31:01
・感冒薬	A*02:06
・クロザピン	B*59:01
・アバカビル	B*57:01
・サニルブジン	B*40:01
・チアマゾール	B*27:05

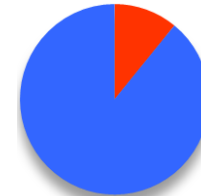
高尿酸血症治療剤
日本薬局方

アロプリノール錠
ザイロリック[®]錠50
ザイロリック[®]錠100

51例中全ての症例がHLA-B*5801保有者
報告がある³⁾。また、別の研究では、
ルにより皮膚粘膜眼症候群及び中毒性
発症した日本人及びヨーロッパ人にお
て10例中4例(40%)、27例中15例(55%)
01保有者であったとの報告もある^{4),5)}。

B*58:01頻度情報

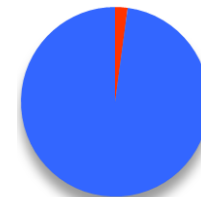
台湾



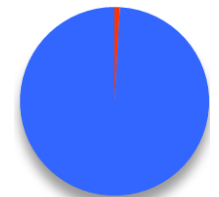
中国



欧米



日本



- ・一部のHLA遺伝子型は、重篤な疾患(例:薬剤重症副作用)の発症リスクを有し、薬剤添付文書に記載されるなど、**個別化医療**の観点からも重要。
- ・リスクHLA遺伝子型は集団間で頻度が異なる例が多く、**各集団におけるHLA遺伝子型頻度分布の把握**が重要になります。

② HLA imputation法

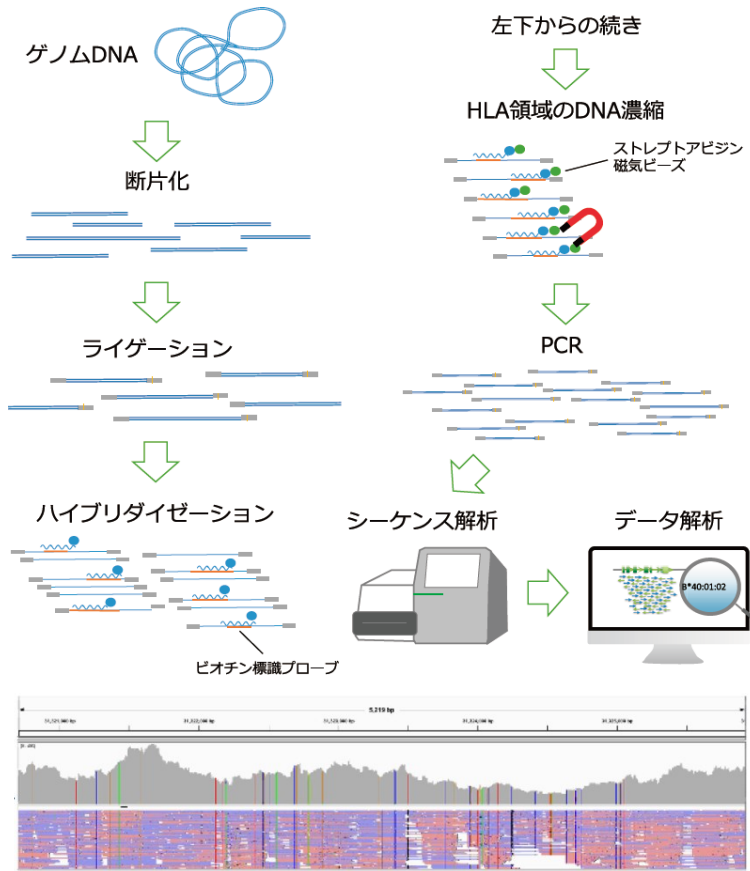


Luminex法の手順
(ジェノダイブファーマ
株式会社のHPより)

- HLA遺伝子型の測定方法は、複数種類あります。
- 4-digit HLAアレルの決定では、**Luminex(PCR-SSO)法**が有名です。
- 大規模疾患サンプルの解析において、全HLA遺伝子座位を測定する場合、アッセイコストが高額になることがボトルネックでした。

② HLA imputation法

NGSリードからのHLA遺伝子型推定



推定結果の精度比較

Tool	Accuracy (Success)
optitype ⁺	35% (71%)
hlavbseq	52% (52%)
hlaminer assembly	17% (36%)
hlaminer alignment	15% (26%)
phlat	38% (46%)
seq2hla [*]	7% (12%)
optitype ⁺	49% (98%)
hlavbseq	68% (68%)
hlaminer assembly	43% (49%)
hlaminer alignment	26% (27%)
phlat	73% (73%)
seq2hla [*]	60% (61%)
optitype ⁺	50% (99%)
hlavbseq [*]	67% (67%)
hlaminer assembly	52% (61%)
hlaminer alignment	20% (20%)
phlat	81% (81%)
seq2hla	79% (79%)

(Bauer DC et al. *Brief Bioinform* 2016)

- WGS/WES等のNGSリード情報からのHLA遺伝子型推定も技術的に可能になっています。潜在的な有用性は高いものの、データ解析の専門性の高さや解析結果の精度が課題となっています。

② HLA imputation法

HLA-DRB1*14:01 と DRB1*14:54 に関するアナウンスメント

日本組織適合性学会
標準化委員会

経緯:

HLA-DRB1*14:54 は 2005 年に報告された比較的新しいアリルであるが、欧米の調査で既にタイプされている HLA-DRB1*14:01 の多くが DRB1*14:54 であることが判明してきた。

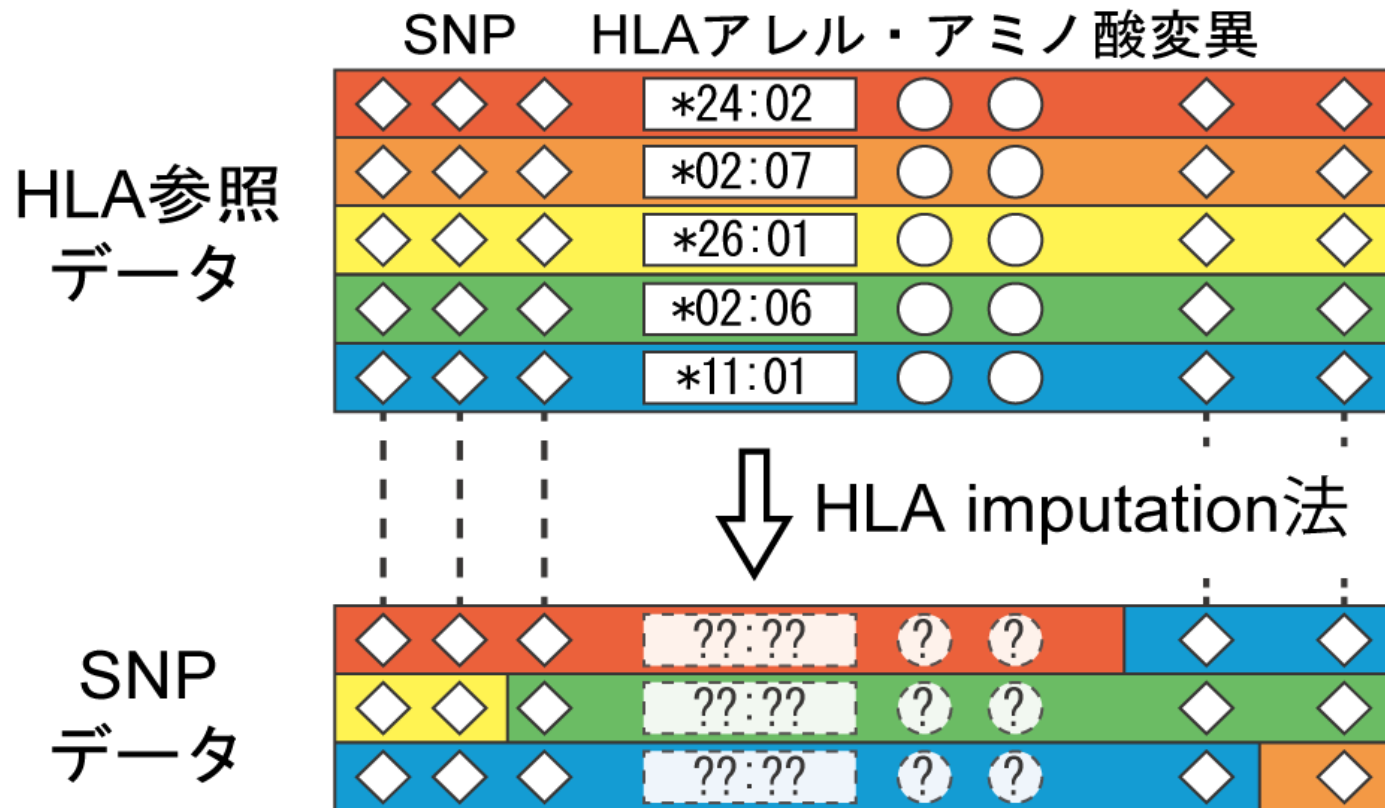
近年、韓国、台湾などアジア地域の調査結果が公表され、ほぼ 100%HLA-DRB1*14:54 であった。日本国内でも、いくつかの施設で調査が行われており、本学会の呼びかけで集計したところ 6 施設 728 検体の全検体で HLA-DRB1*14:54 と判定された。

両アリルの塩基配列の違いは HLA-DRB1 座の Exon3 に存在しており、現在使用されている多くの HLA-DRB1 座タイピング試薬が Exon2 の変異を検出していることから、両者の分布を明確にすることが出来ず、現在の混乱を招いている。

- NGS法などの解析手段の発達により、従来の方法で判定されていた HLA 遺伝子型推定に誤りがあったことも判明しています。
- 例えば、アジア人集団で報告されていた「HLA-DRB1*14:01」は、「HLA-DRB1*14:54」の誤りであったと考えられています。

② HLA imputation法

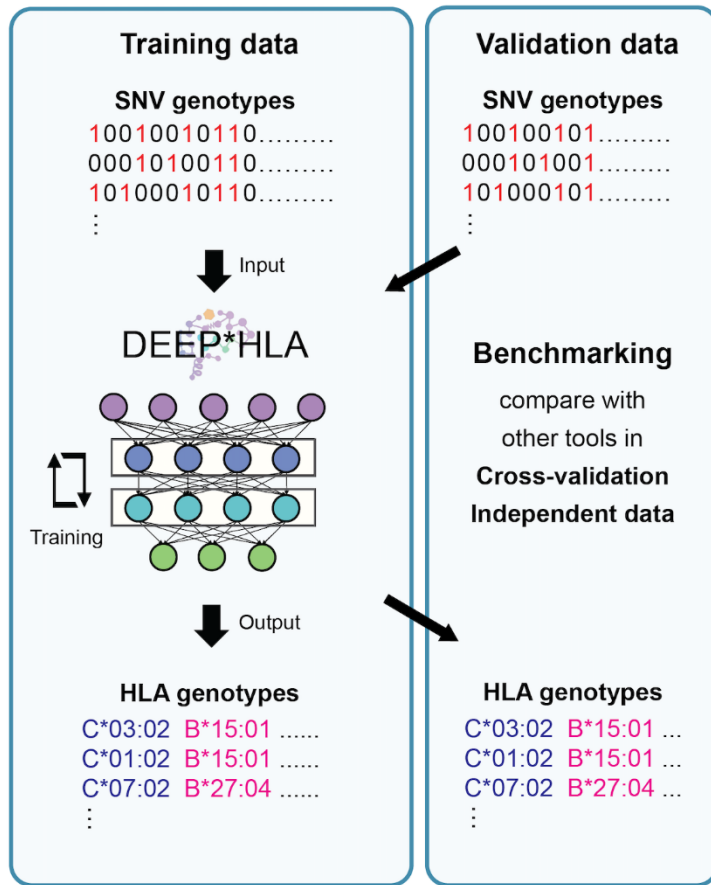
HLA imputation法によるHLA遺伝子多型推定



- HLA imputation法により、SNPデータから「追加費用なし」で、HLA遺伝子多型をコンピューター上で高精度で推定可能になりました。
- 既に数百万人規模で存在するSNPデータを対象に、HLA遺伝子型の網羅的な解析が可能になり、数多くの知見が報告されています。

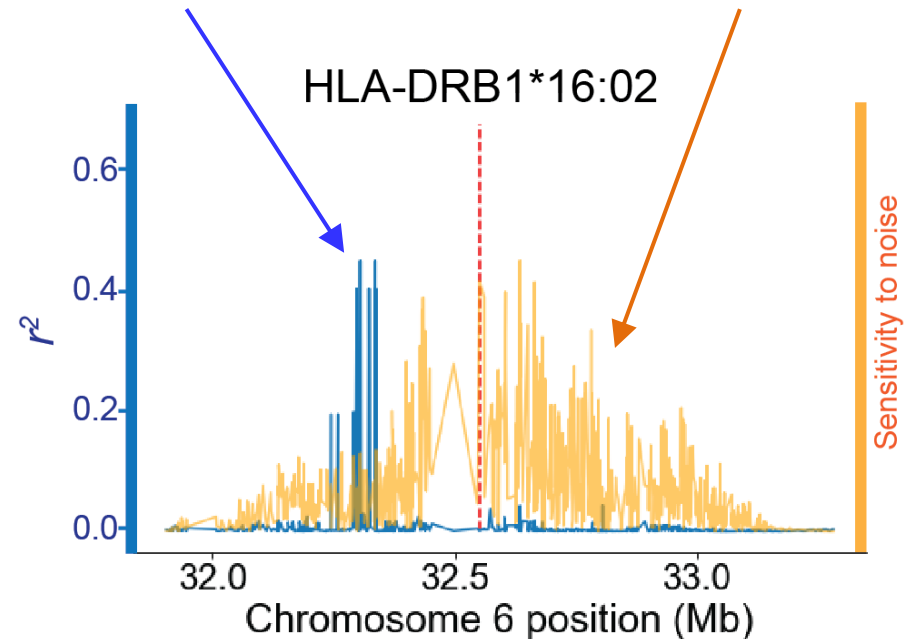
② HLA imputation法

ヒト集団ゲノムデータ行列への深層学習の適用



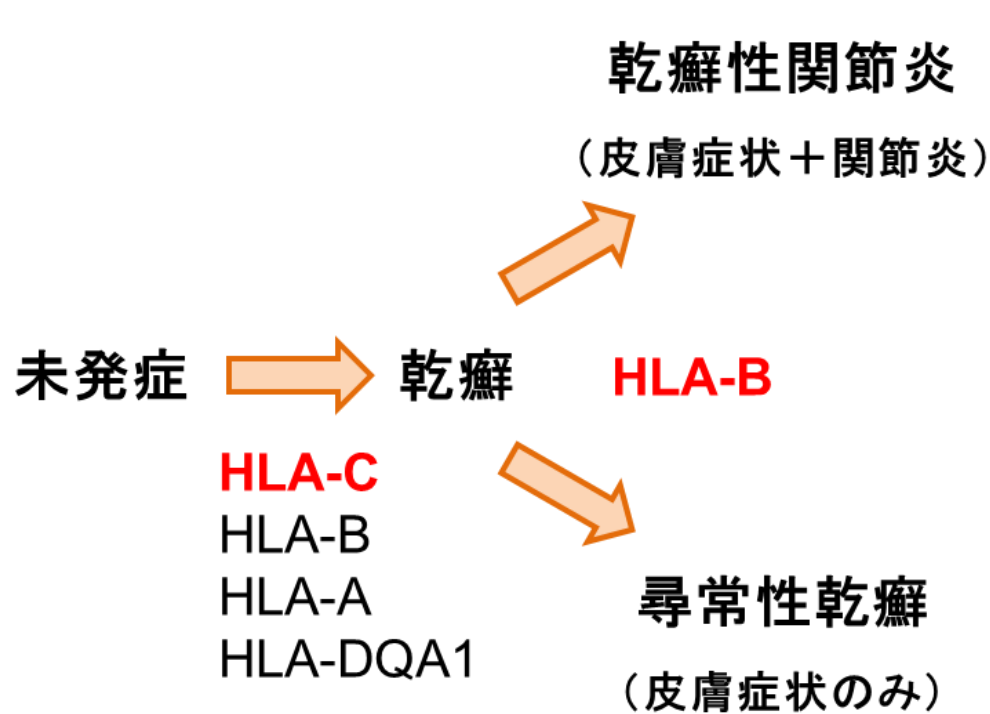
既存の機械学習が
参考にする箇所

深層学習が
参考にする箇所

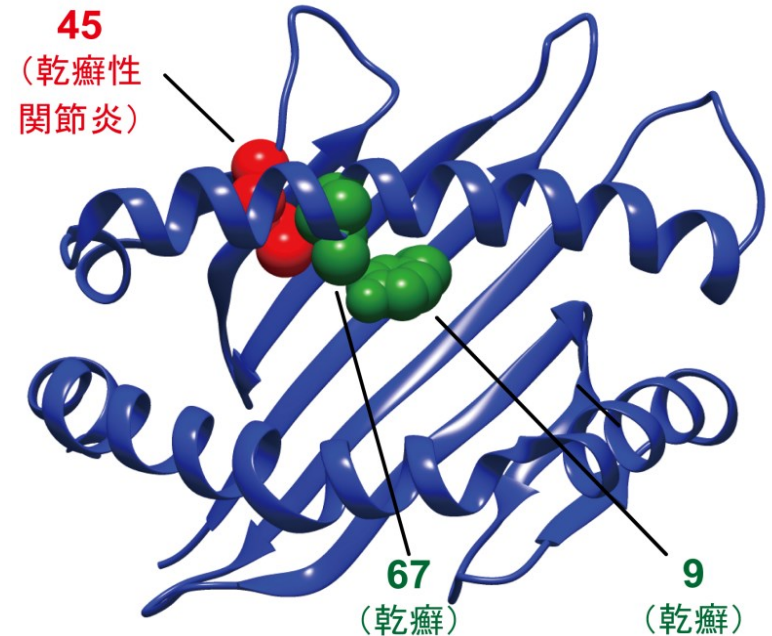


- 深層学習によるHLA imputation法(DEEP*HLA)を開発。MHC領域内のヒト集団ゲノム行列を画像変換することで深層学習の適用を可能にした。
- 従来の機械学習(例:マルコフ連鎖)と比較して、**希なHLA遺伝子型**の推定精度が改善。

② HLA imputation法



HLA-B 遺伝子上で乾癬および乾癬性関節炎の発症リスクを有するアミノ酸配列

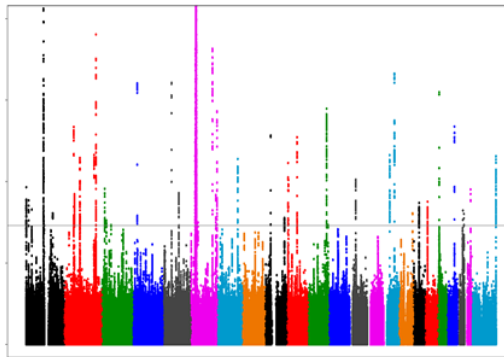


- HLA imputation法を、欧米人集団における乾癬(psoriasis)GWASに適用
(患者群:9,247名、対照群: 13,589名)。
- 乾癬の発症にはHLA-Cを含む複数のHLA遺伝子の関与を同定。
- 乾癬発症後の病態進展にはHLA-Bの関与を同定。
- HLA様遺伝子(MICA)の関与は明らかでなかった。

② HLA imputation法

日本人集団におけるHLA imputation法の実装

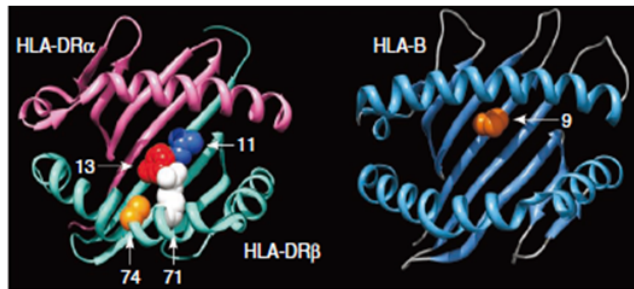
欧米人集団GWAS



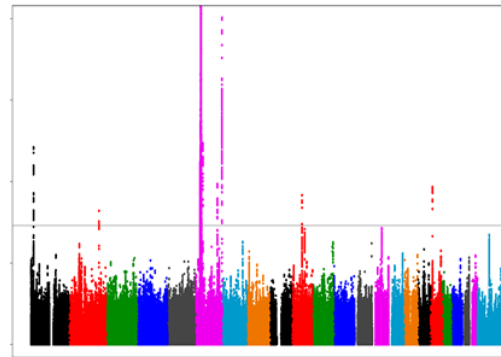
+

欧米人
集団HLA
参照データ

HLA imputation ↓



日本人集団GWAS



+

日本人
集団HLA
参照データ

HLA imputation ↓

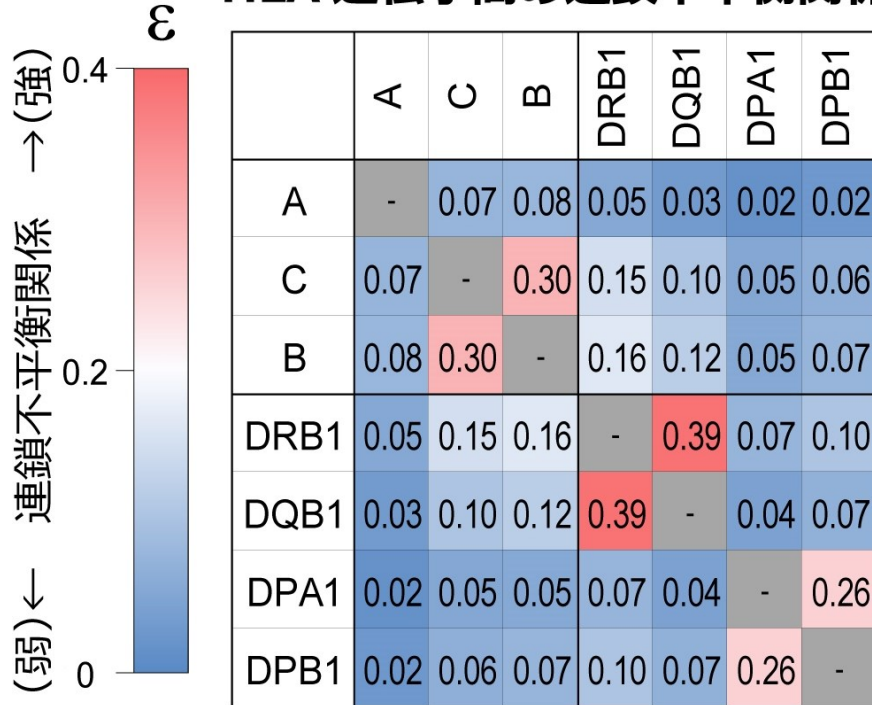
?

- 日本人集団においてHLA imputation法を実装するため、**日本人集団を対象とした参照データ**を新たに作成しました($n = 908$)。
- 高精度なHLA imputationを実施可能に(4-digitアレル一致率 $\geq 96\%$)。

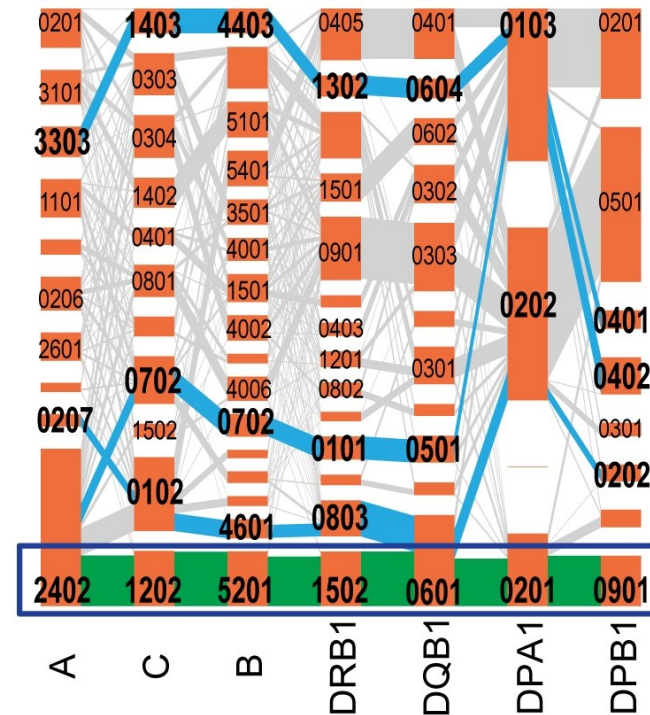
② HLA imputation法

高次元ビッグデータ可視化手法の開発

HLA 遺伝子間の連鎖不平衡関係



HLA 遺伝子のハプロタイプ構造



← 日本人集団に特異的なハプロタイプ

- 情報量エントロピーの正規化と高次元データ圧縮技術を用いて、HLA遺伝子配列構造における人種特異性の可視化に成功。
- 日本人集団に特異的なHLAハプロタイプの存在が明らかに。

② HLA imputation法

次世代シーケンサーによるHLA解析の新展開

クラスI HLA遺伝子

HLA-A
HLA-B
HLA-C

クラスII HLA遺伝子

HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DPA1
HLA-DPB1

非古典的HLA遺伝子

HLA-DOA/DOB
HLA-DMA/DMB
HLA-E/F/G
HLA-V/H/K/J/L
HLA-DRB2/6/7/8/9

HLA様遺伝子

MICA, MICB
TAP1, TAP2

Long PCR
+
Target Capture
+
Long Read NGS



• NGSの活用で、これまで注目されてこなかった、**マイナーなHLA遺伝子**の配列が解読可能になり、日本人集団1,150名を対象に、**非古典的HLA遺伝子・HLA様遺伝子・偽HLA遺伝子**に対してもHLA imputation法の適用を拡大することができました。

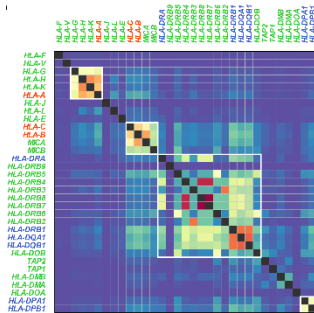
② HLA imputation法

機械学習による白血球血液型の分類

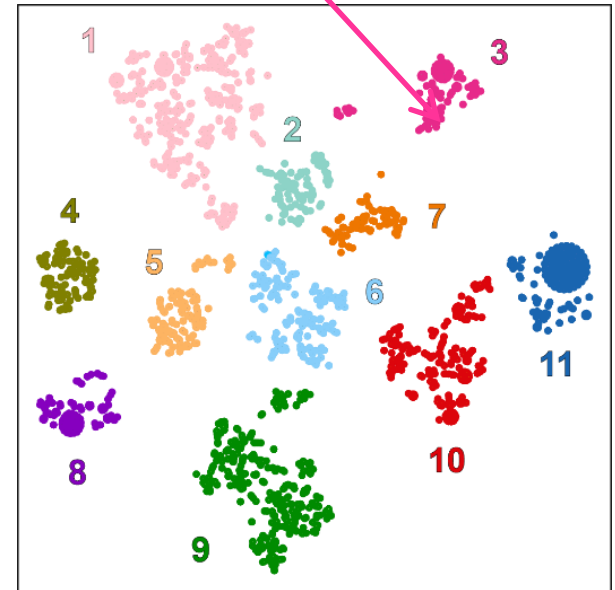
星の数より多い
白血球の血液型



機械学習による
白血球の
血液型分類

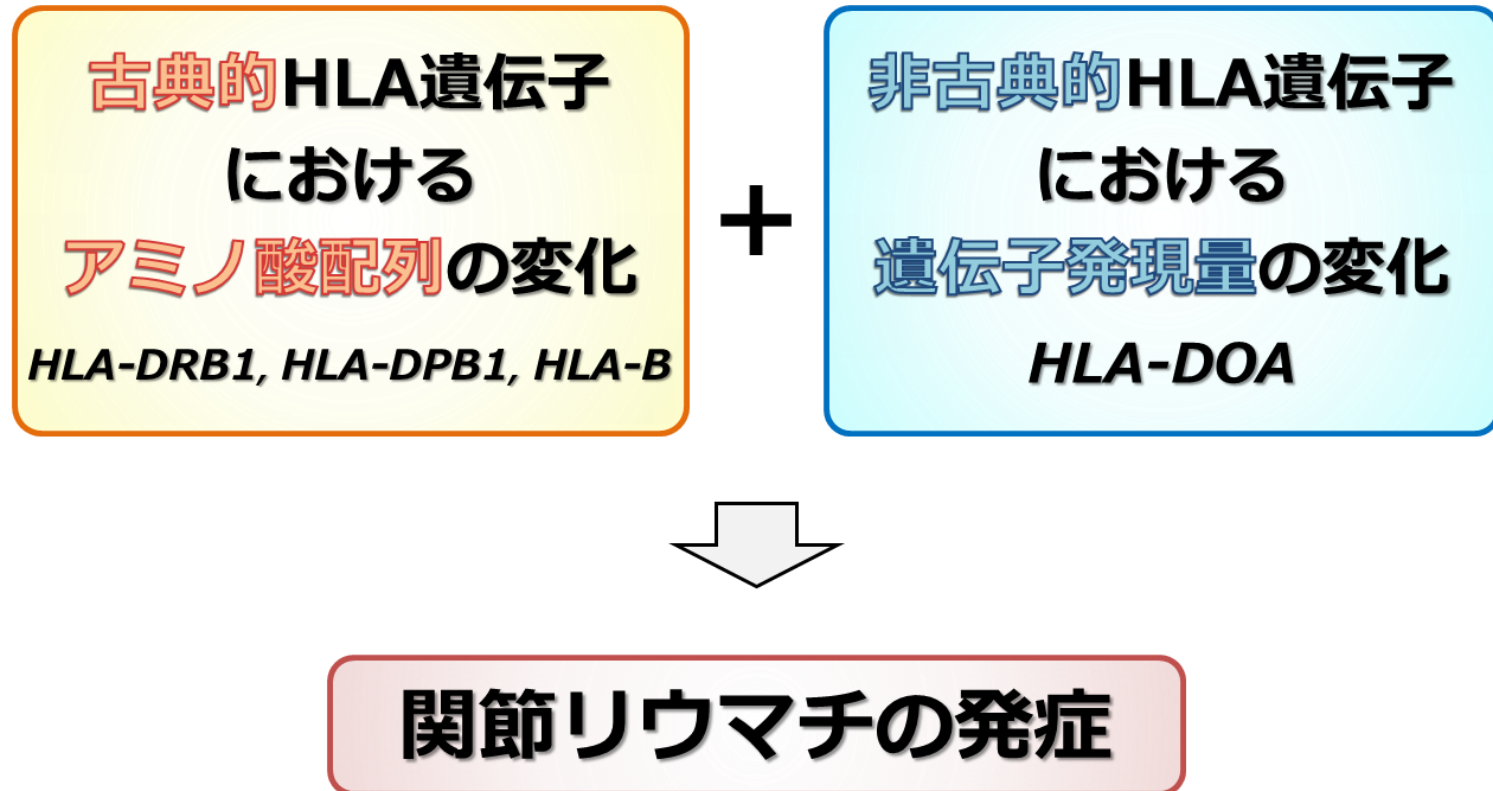


機械学習により
分類された血液型パターン



- 300種類以上のHLA遺伝子型の組み合わせは、10の24乗パターンに。
- 機械学習手法t-SNEを適用することで、日本人集団の白血球の血液型が11パターンの組み合わせで分類されることが明らかになりました。
- ヒトゲノム研究分野における機械学習の応用例の一つと考えられます。

② HLA imputation法



- HLA遺伝子配列データベースに基づく解析の結果、従来の**古典的HLA遺伝子**に加えて、**非古典的HLA遺伝子**においても疾患リスクを有することが判明しつつあります。

② HLA imputation法

PheWASが同定したMHC領域内遺伝子変異に関連した形質

表現型カテゴリ	表現型名	表現型カテゴリ	表現型名	表現型カテゴリ	表現型名
アレルギー疾患	アトピー性皮膚炎 喘息 花粉症	血液検査値	赤血球数 ヘモグロビン濃度 平均赤血球容積 平均赤血球ヘモグロビン値 平均赤血球ヘモグロビン濃度 白血球数 好中球数 好酸球数 好塩基球数 単球数 リンパ球数 血小板数	生化学検査値	アルブミン 非アルブミン蛋白 アルブミン/グロブリン比 血清クレアチニン 推定糸球体濾過量 尿酸 カリウム 無機リン 総ビリルビン アスパラギン酸アミノ基転移酵素 アラニンアミノ基転移酵素 アルカリフォスファターゼ クレアチンキナーゼ 乳酸脱水素酵素
自己免疫疾患	関節リウマチ バセドウ病 1型糖尿病			生理検査結果	収縮期血圧 平均血圧
感染症	B型肝炎 C型肝炎	生化学検査値	総コレステロール HDLコレステロール 中性脂肪 血糖 ヘモグロビンA1c 総蛋白		
心血管障害	心筋梗塞 安定狭心症				
生活習慣病	2型糖尿病 高脂血症				
悪性腫瘍	肺癌 肝臓癌				
臓器疾患	肝硬変 ネフローゼ症候群				
身体測定値	身長 肥満				

- 日本人集団17万人で、**100以上の表現型**とMHC領域内多型との関連をPhenome-wide association study(PheWAS)で網羅的に検討。
- **約半数の52の形質**で、MHC領域内遺伝子多型との関連を同定。
- MHC領域内の**非HLA遺伝子**のリスクも複数確認されました。

GenomeDataAnalysis3

- ① SNP genotype imputation
- ② HLA imputation法
- ③ SNP2HLAを使ったHLA imputation法

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 SNP2HLAを使ったHLA imputation法

HLA imputation法の解析ソフトウェア

ソフトウェア	URL	引用文献
SNP2HLA	https://www.broadinstitute.org/mpg/snp2hla/	Jia X et al. <i>PLoS One</i> 2013
HLA*IMP2	https://oxfordhla.well.ox.ac.uk/hla/	Dilthey AT et al. <i>Bioinformatics</i> 2011
HIBAG	http://www.biostat.washington.edu/~bsweir/HIBAG/	Zheng X et al. <i>Pharmacogenomics J</i> 2014
CookHLA	https://github.com/WansonChoi/CookHLA	Cook S et al. <i>Nat Commun</i> 2021
DEEP*HLA	https://github.com/tatsuhikonaito/DEEP-HLA	Naito T et al. <i>Nat Commun</i> 2021
HLA-TAPAS	https://github.com/immunogenomics/HLA-TAPAS	Luo Y et al. <i>Nat Genet</i> 2021

- HLA imputation法を実施するソフトウェアは、複数あります。
- Imputation精度は、ソフトウェア間であまり差がないと報告されています。
- 本実習では、下記の理由から**SNP2HLA**を使った演習を行います。

①: ~~元上司が作ったから。~~

②: 使いやすい。

③: 参照データと共に公開されている。

④: アレルだけでなく、**アミノ酸配列多型のimputationも可能。**

③ SNP2HLAを使ったHLA imputation法

HLA imputation at Michigan Imputation Server

The screenshot shows the Michigan Imputation Server website navigation menu. At the top is a search bar labeled 'Search docs'. Below it are links for 'Home', 'Getting Started', 'Data Preparation', and 'Reference Panels'. A 'Pipeline Overview' section is highlighted, containing links for 'Pipeline Overview', 'Quality Control', 'Phasing', 'Imputation', 'HLA Imputation Pipeline', 'Compression and Encryption', and 'Chromosome X Pipeline'. Below this are links for 'Security', 'FAQ', 'Developer Documentation', 'API', 'Docker', 'Create Reference Panels', and 'Workshops'.

HLA Imputation Pipeline

In addition to intergenic SNPs, HLA imputation outputs five different types of markers: (1) binary marker for classical HLA alleles; (2) binary marker for the presence/absence of a specific amino acid residue; (3) HLA intragenic SNPs, and (4) binary markers for insertion/deletions, as described in the typical output below. The goal is to minimize prior assumption on which types of variations will be causal and test all types of variations simultaneously in an unbiased fashion. However, the users are always free to restrict analyses to specific marker subsets.

Note

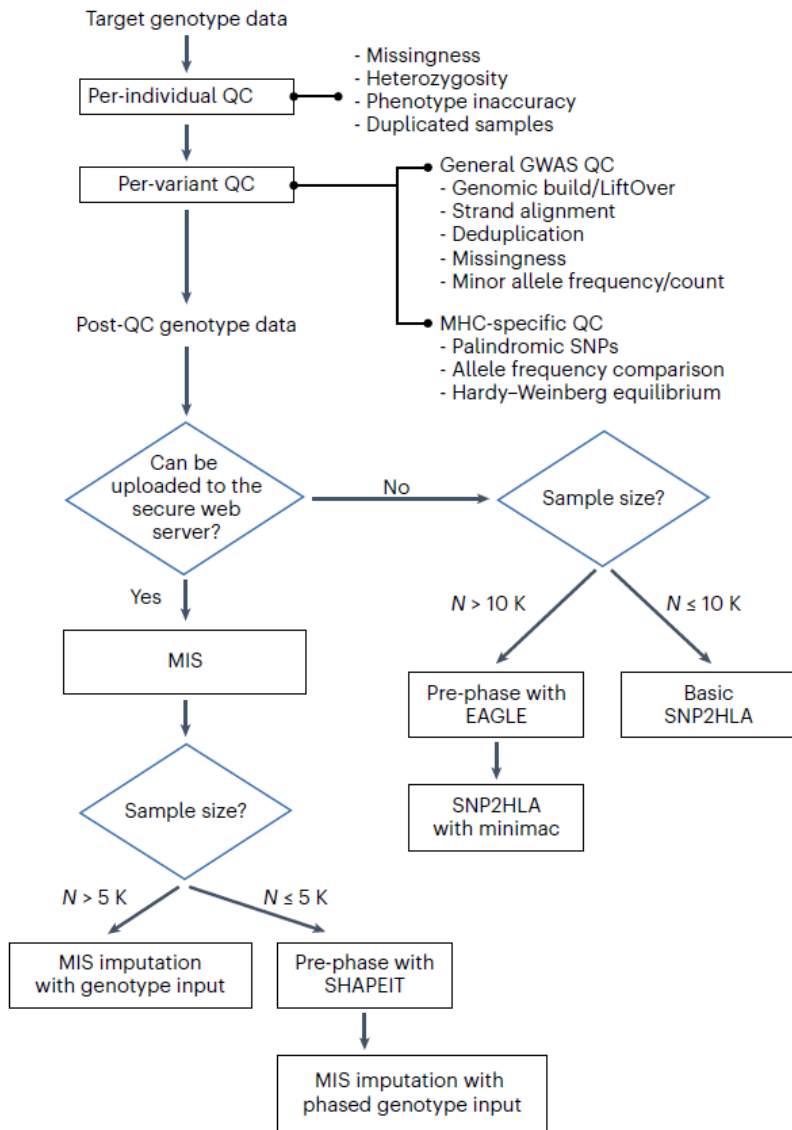
For binary encodings, A = Absent, T = Present.

Type	Format	Example
Classical HLA alleles	HLA_[GENE]*[ALLELE]	HLA_A*01:02 (two-field allele) HLA_A*02 (one-field allele)
HLA amino acids	AA_[GENE]_[AMINO ACID POSITION]_[GENOMIC POSITION]_[EXON]_[RESIDUE]	AA_B_97_31324201_exon3_V (amino acid position 97 in HLA-B, genomic position 31324201 (GrCh37) in exon 3, residue = V (Val))
HLA intragenic SNPs	SNPS_[GENE]_[GENE POSITION]_[GENOMIC POSITION]_[EXON/INTRON]	SNPS_C_2666_31237183_intron6 (SNP at position 2666 of the gene body, genomic position 31237183 in intron 6)
Insertions/deletions	INDEL_[TYPE]_[GENE]_[POSITION]	INDEL_AA_C_300x301_31237792 (Indel between amino acids 300 and 301 in HLA-C, at genomic position 31237792)

<https://imputationserver.readthedocs.io/en/latest/pipeline/#hla-imputation-pipeline>

・サーバーに各自がGWASデータをアップロードして、HLA imputationを実施するシステムも構築されています。

③ SNP2HLAを使ったHLA imputation法

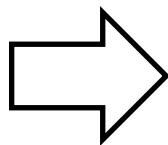
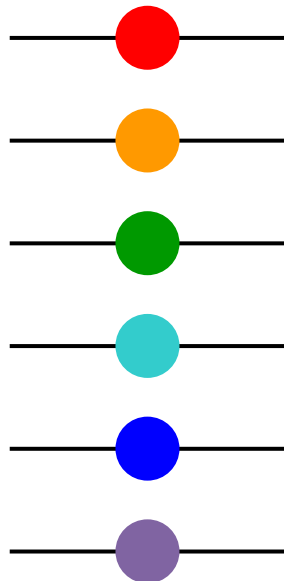


• HLA imputation法やHLA遺伝子型関連解析手順の protocols 論文です。
 (Sakaue S et al. *Nat Protoc* 2023)

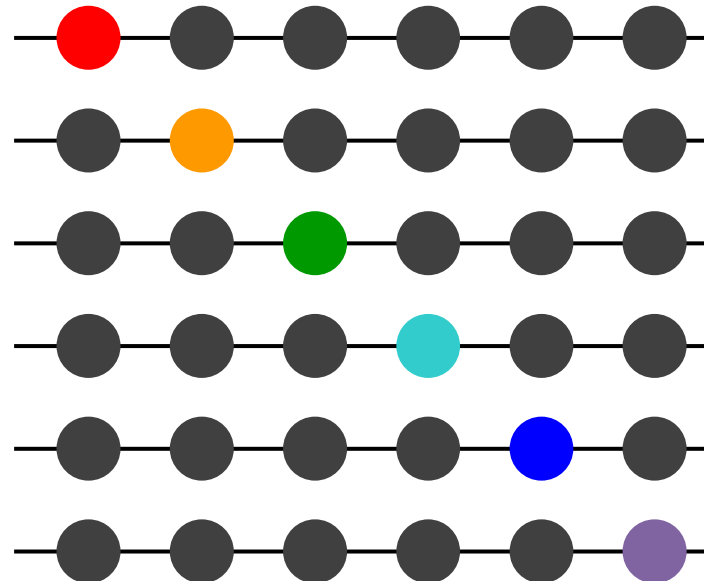
③ SNP2HLAを使ったHLA imputation法

SNP2HLAにおけるマルチアレル多型の取り扱い

一つのmultivariate変数



複数のbinary変数



- HLA imputationでは、**マルチアレル多型**の推定がネックとなります。
- SNP2HLAでは、マルチアレルなHLA遺伝子多型を「**一つのmultivariate変数**」ではなく「**複数のbinary変数**」として扱った結果、高精度のimputationと、HLAアミノ酸多型への適用拡大が可能になりました。

③ SNP2HLAを使ったHLA imputation法

Strandによるrs671(ALDH2)の表記方法

Positive (+) strand → GGCATACACTGAAGTGAAAAC
|||||
Negative (-) strand → CCGTATGTGACTTCACTTTTG



Positive strand表記: G>A GGCATACACTAAAGTGAAAAC
|||||
Negative strand表記: C>T CCGTATGTGATTTCACTTTTG

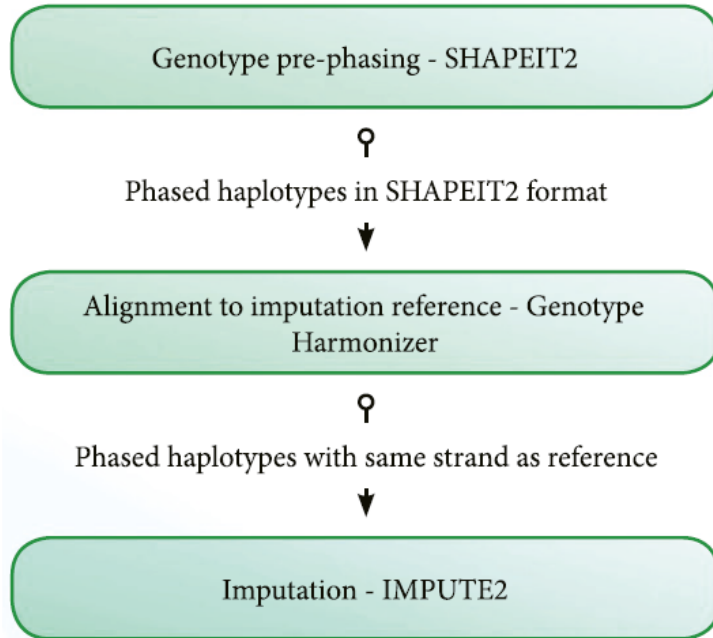
*A>T、G>CタイプのSNPは、strandを逆転してもT>A、C>Gとなるため、見た目ではstrandを判別できません。

- SNPのアレル表記は、標準ゲノム配列の2重鎖のどちら側から読むか(=strand)で変わるため、imputation実施前に、GWASデータと参照データで、共通SNPのstrandのマッチングを行う必要があります。
- 特に、A/TおよびG/CタイプのSNP(=palindromic SNP)は、strandマッチングが困難です。
- SNP2HLAは、共通SNPのstrandマッチングを自動的に実施します。

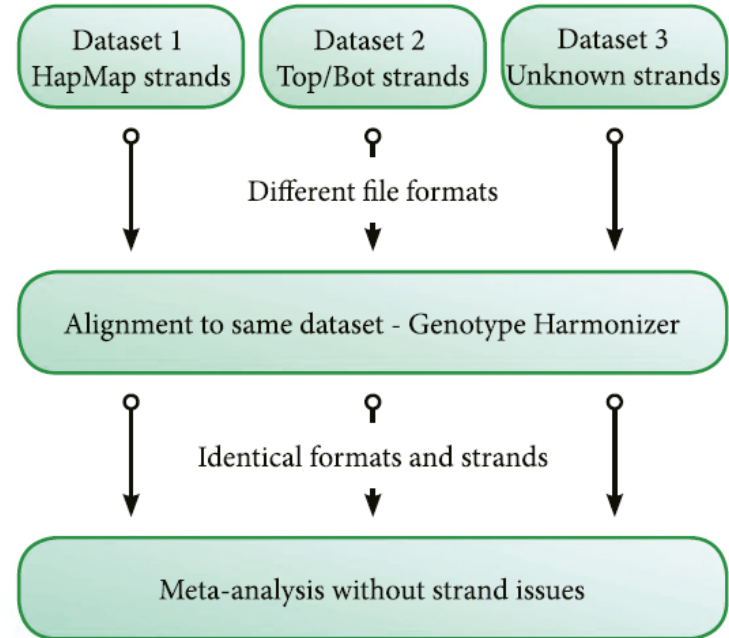
③ SNP2HLAを使ったHLA imputation法

Genotype Harmonizer

Usage in genotype imputation workflow



Usage in meta-analysis workflow



- GWASデータと参照データで、共通SNPのstrandのマッチングは、通常のゲノムワイドのSNP imputationにおいても必要な作業です。
 - GWASデータと参照パネル間のstrandマッチングを実施するソフトウェアも開発が進んでおり、Genotype Harmonizer等があります。
 - 全SNPについて正確にマッチングしない例があり、別途確認が必要です。
- (<https://github.com/molgenis/systemsgenetics/wiki/Genotype-Harmonizer>; Deelen P et al. *BMC Res Notes* 2014)

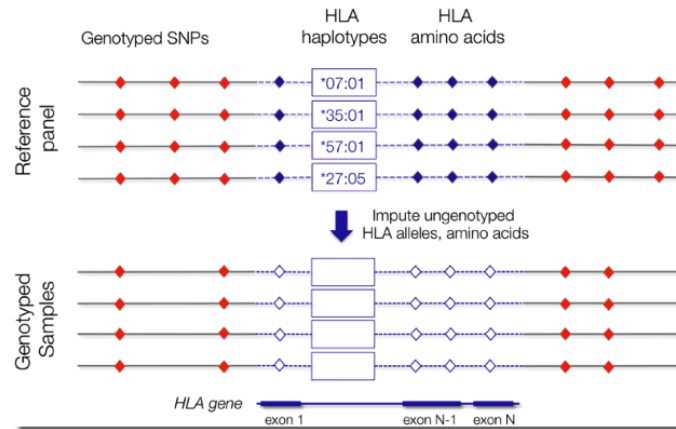
③ SNP2HLAを使ったHLA imputation法

SNP2HLA


<http://software.broadinstitute.org/mpg/snp2hla/>

SNP2HLA: Imputation of Amino Acid Polymorphisms in Human Leukocyte Antigens

SNP2HLA is a tool to impute amino acid polymorphisms and single nucleotide polymorphisms in human leukocyte antigens (HLA) within the major histocompatibility complex (MHC) region in chromosome 6.



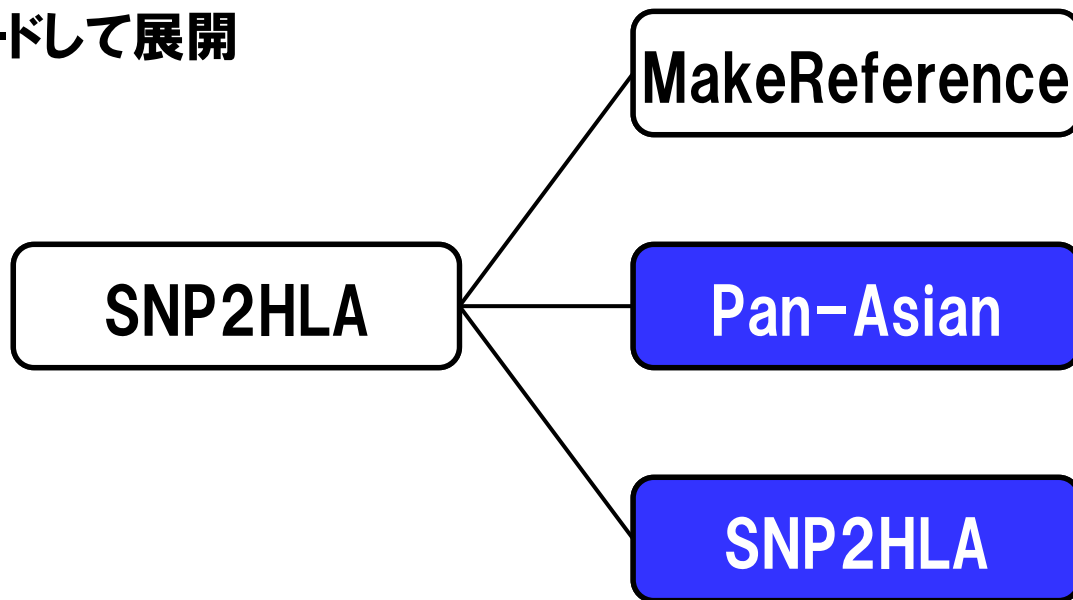
Links

- [Download: SNP2HLA v1.0.3 package including SNP2HLA, MakeReference, and Pan Asian Reference Panel \(tarball\)](#) 
- *NEWS1*: Pan-Asian reference panel is now included in the package V1.0.3! (7/10/14)
- *NEWS2*: T1DGC reference panel is now removed from the package V1.0.3 due to security issues relating to individual-level genotype data (3/10/15). If you are a researcher interested in obtaining access to this reference panel, please contact snp2hla@broadinstitute.org
- [Manual: SNP2HLA](#)
- [Manual: MakeReference](#)

• SNP2HLAのソフトウェアは、参照データと共に公開されています。

③ SNP2HLAを使ったHLA imputation法

※SNP2HLA_package_v1.0.3.tar.gz
をダウンロードして展開



• SNP2HLAソフトウェアは、下記の3つで構成されています。

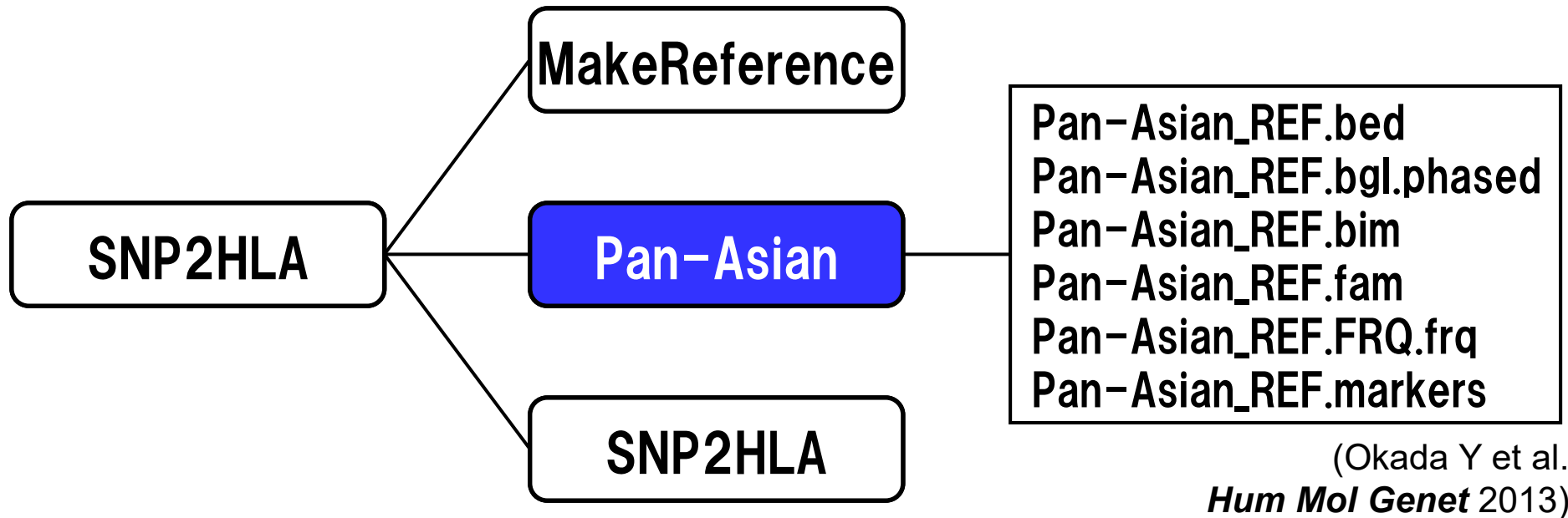
①: **MakeReferene** → 参照データを作るパッケージ

②: **Pan-Asian** → 東アジア人集団の参照データ

③: **SNP2HLA** → HLA imputationを実施するパッケージ

• 今回は、②と③を使ってHLA imputationを実施します。

③ SNP2HLAを使ったHLA imputation法



• Pan-Asianは、**東アジア人集団530名分の参照データ**です。

HapMap JPT+CHB : $n = 89$ Indian : $n = 119$ Malaysian : $n = 120$
Chinese : $n = 111$ Singapore Chinese : $n = 91$

• PLINK形式のファイル(xxx.bed/bim/fam/FRQ.frq)と、imputationソフト
Beagle形式のファイル(xxx.bgl.phased/markers)で構成されています。

• 古典的HLA遺伝子の、**2-digitアレル**、**4-digitアレル**、**アミノ酸配列多型**が、**MHC領域内SNPデータ**と共に公開されています。

③ SNP2HLAを使ったHLA imputation法

The screenshot shows the Human data NBDC database interface. At the top, there is a navigation bar with 'Human data NBDCヒトデータベース', 'English', and a search box. Below the navigation bar, there are tabs for 'ホーム', 'データの利用', 'データの提供', 'ガイドライン', '機関外サーバ', 'NBDCヒトデータ審査委員会', '成果発表', 'お問い合わせ', and 'FAQ'. The main content area displays 'NBDC Research ID: hum0028.v1' and a note that the latest version is 'こちらです'. Under the heading '研究内容の概要', there is a '目的' (Objective) section stating the goal is to create reference data for HLA genotype imputation in Japanese healthy individuals. The '方法' (Method) section describes the use of Illumina BeadChips and WAKFlow HLA typing kit. The '対象' (Subjects) section mentions 908 Japanese healthy individuals. A URL is provided: <http://biobankjp.org/index.html>. At the bottom, a table lists the data ID, content, access restrictions, and publication date.

データID	内容	制限	公開日
JGAS000018	HLAアリルおよびHLA領域内SNPの遺伝子型決定	制限公開 (Type I)	2015/06/02

*SNP2HLAの開発バージョンの都合上、参照データ中の位置情報は、(やや古い)Build 36に準拠しています。

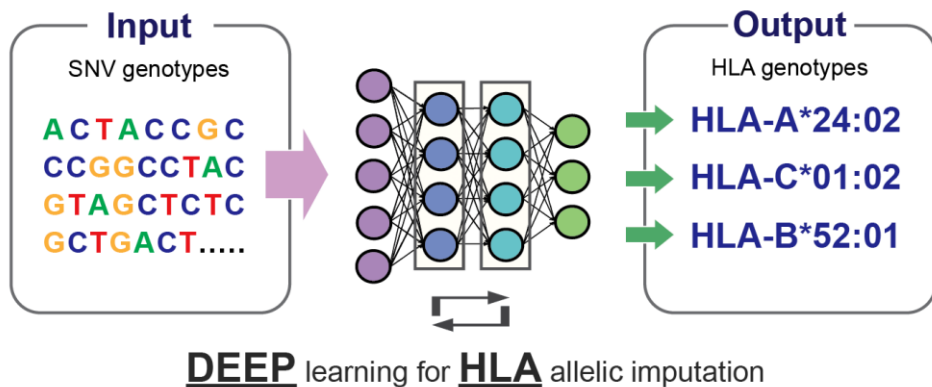
<https://humandbs.biosciencedbc.jp/hum0028-v1>

- HLA imputation法においても、参照データは、**サンプル数が大きく、遺伝的背景がGWASデータに近いほど、推定精度が高くなります。**
- **日本人集団**においては、**908名の参照データ**が、NBDCデータベース上で公開されており、所定の手続きを経て入手することができます。

③ SNP2HLAを使ったHLA imputation法

githubを通じた日本人集団HLA imputation推定モデルの一般公開

DEEP*HLA



DEEP*HLA

DOI: [10.5281/zenodo.4478902](https://doi.org/10.5281/zenodo.4478902)

DEEP*HLA is an HLA allelic imputation method based on a multi-task convolutional neural network implemented in Python.

DEEP*HLA receives pre-phased SNV data and outputs genotype dosages of binary HLA alleles.

In DEEP*HLA, HLA imputation is performed in two processes:

- (1) model training with a HLA reference panel
- (2) imputation with a trained model.

Publication/Citation

The study of DEEP*HLA is described in the manuscript.

- Naito, T. et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. Nat. Commun. 12, 1639 (2021). doi.org/10.1038/s41467-021-21975-x

Please cite this paper if you use DEEP*HLA or any material in this repository.

Requirements

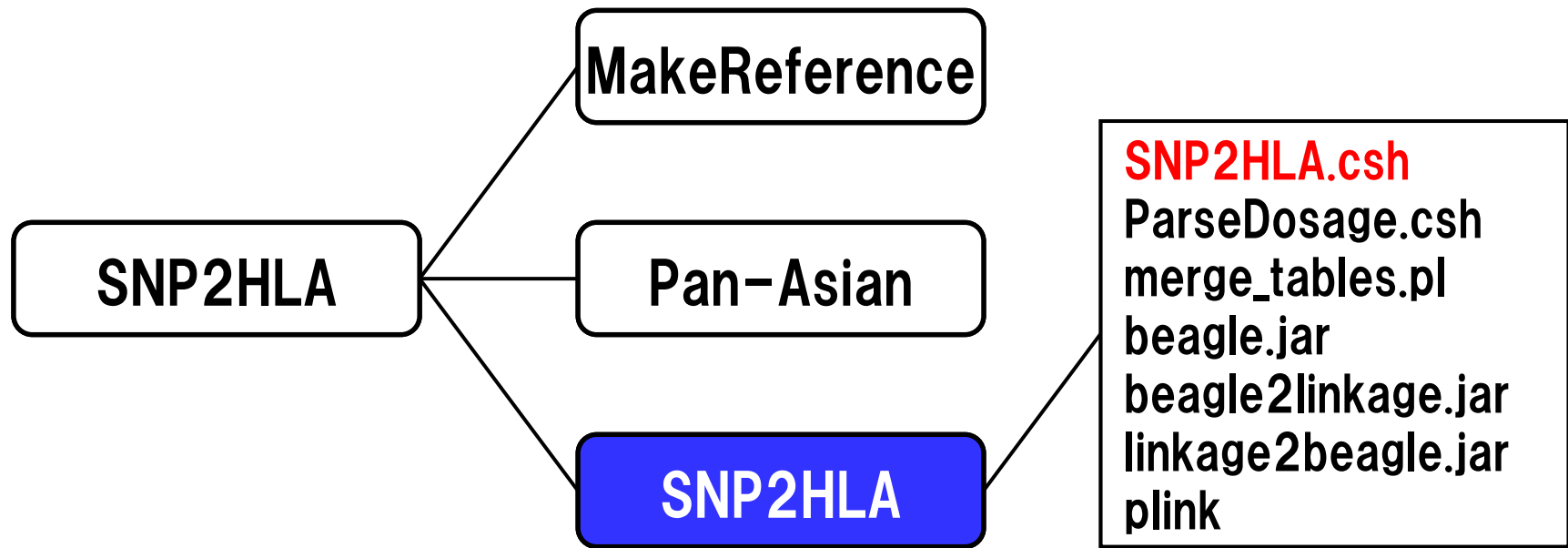
- Python 3.x (3.7.4)
- Pytorch (1.4.0)
- Numpy (1.17.2)
- Pandas (0.25.1)
- Scipy (1.3.1)
- Argparse (1.4.0)

DEEP*HLA was tested on the versions in parentheses, so we do not guarantee that it will work on different versions.

Installation

- 参照データは個人別ジェノタイプを含む個人情報に相当しますが、参照データに基づき学習されたHLA推定モデルは個人情報に相当しません。
- 推定モデルを一般公開することで、個人情報に抵触することなくHLA imputation法の参照データの提供が可能になります。

③ SNP2HLAを使ったHLA imputation法



- SNP2HLAは、いくつかのソースコードで構成されています。
- **SNP2HLA.csh**がメインのソースコードで、SNP2HLA.cshの内部で他のソースコードを呼び出して実行する、という仕組みになっています。
- SNP2HLAは、PLINKやSNP genotype imputationソフトであるBeagleを内部から呼び出して、imputation作業を実施しています。

③ SNP2HLAを使ったHLA imputation法

```
statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis3/SNP2HLA/GWAS
```

```
$ ls
```

```
HapMap3_MHC_EAS.bed HapMap3_MHC_EAS.fam HapMap3_MHC_EAS.bim
```

```
$ wc *fam
```

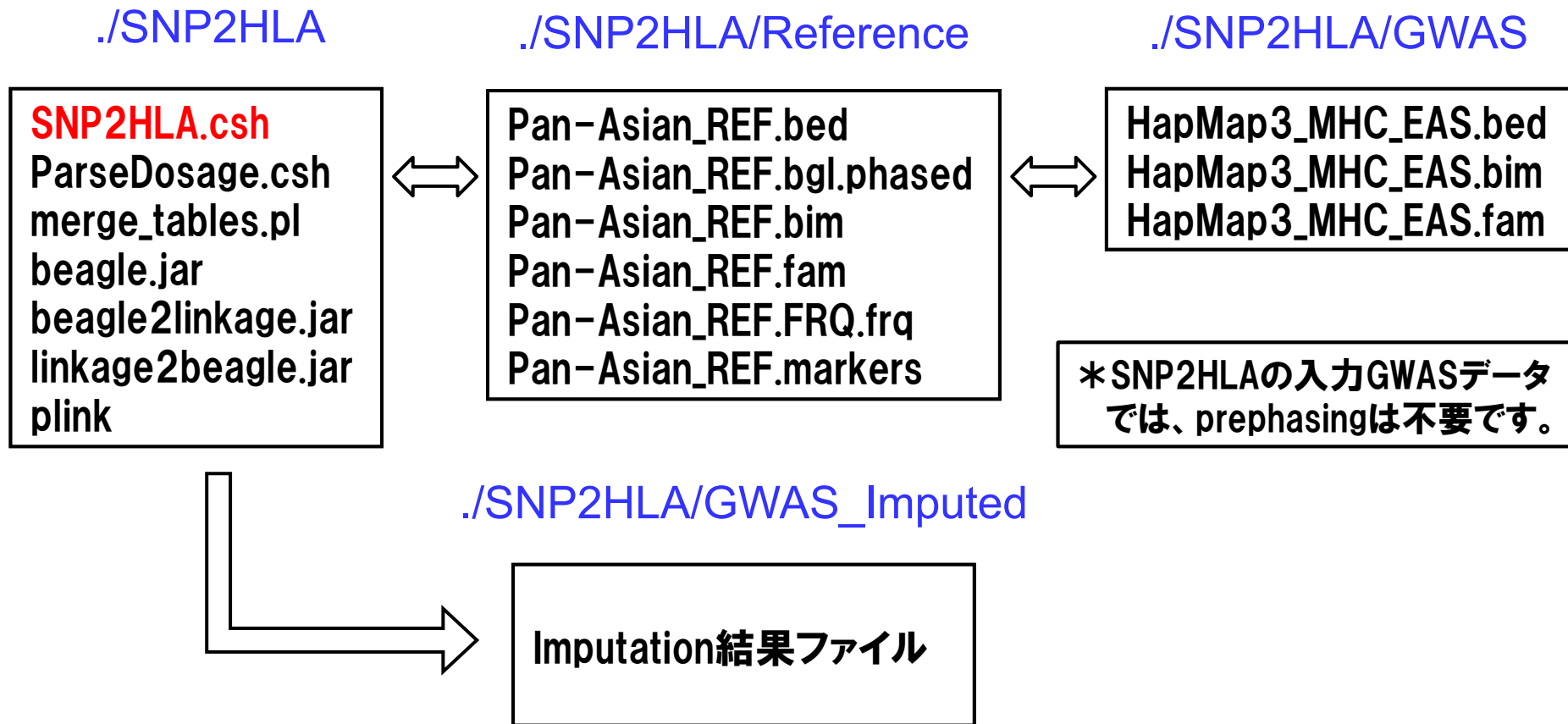
```
170 1020 4250 Hap3_EAS.fam
```

```
$ wc *bim
```

```
7843 47058 219565 Hap3_EAS.bim
```

- GWASデータとして、HapMap Phase3データの東アジア人170名のSNPデータを取得しました。
- MHC領域内(6番染色体:24Mb-36Mb)の7,800SNPを対象としています。(Pan-Asian参照データと共通したHapMapサンプルは除外しています)

③ SNP2HLAを使ったHLA imputation法



- HLA imputation法の実施には、複数のファイル群を扱う必要があります。
- 各ファイル群を異なるフォルダに配置し、お互いを参照しながら解析を行うことで、ファイルの整理が容易になります。

③ SNP2HLAを使ったHLA imputation法

statgen@statgen-PC: ~

\$ cd /mnt/c/SummerSchool/GenomeDataAnalysis3/SNP2HLA/

※Cygwinの場合/mnt/を/cygdrive/に変えてください

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis3/SNP2HLA

\$./SNP2HLA.csh ./GWAS/Hap3_EAS ./Reference/Pan-

Asian_REF ./GWAS_Imputed/Hap3_EAS_MHC ./plink 1000

※ファイル”SNP2HLA_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、”plink_mac_20210606.zip”を解凍して、Mac OS用のPLINK実行ファイルに置き換えて実行してください。

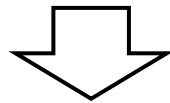
※Macユーザーの方は、演習ファイルを置いたディレクトリを適宜指定してください。

- SNP2HLAは、”./SNP2HLA.csh (GWASデータ名) (参照データ名) (Imputationデータ名) (plink実行ファイル名) (使用メモリ)”という形で実行します。
- 各データの名前は、カレントディレクトリからの相対パスで表記可能です。

③ SNP2HLAを使ったHLA imputation法

```
statgen@statgen-PC:/mnt/c/SummerSchool/GenomeDataAnalysis3/SNP2HLA
$ ./SNP2HLA.csh ./GWAS/Hap3_EAS ./Reference/Pan-Asian_REF ./GWAS_Imputed/Hap3_EAS_MHC ./plink 1000

SNP2HLA: Performing HLA imputation for dataset ./GWAS/Hap3_EAS
- Java memory = 1000Mb
- Beagle window size = 1000 markers
[1] Extracting SNPs from the MHC.
[2] Performing SNP quality control.
[3] Converting data to beagle format.
[4] Performing HLA imputation (see ./GWAS_Imputed/Hap3_EAS_MHC.bgl.log for progress).
```



計算時間:30分程度

```
[5] Converting posterior probabilities to PLINK dosage format.
[6] Converting imputation genotypes to PLINK .ped format.
DONE!
```

- Imputationは、**計算コスト(CPU計算時間、メモリ使用量)の高い作業**です。
- 今回の対象データをノートPCで計算すると、30分ほどかかります。
- 対象サンプル数、対象SNP数の増加に伴い、計算コストも上昇します。
- 月単位で解析の計画を立てることも、あります。

③ SNP2HLAを使ったHLA imputation法

./SNP2HLA/GWAS_Imputed

- **Imputed dosage**のファイル
Hap3_EAS_MHC.bgl.gprobs
Hap3_EAS_MHC.dosage
- **Best guess genotype**のファイル
Hap3_EAS_MHC.bgl.phased
Hap3_EAS_MHC.bed
Hap3_EAS_MHC.bim
Hap3_EAS_MHC.fam
- **Imputation精度**のファイル
Hap3_EAS_MHC.bgl.r2

• SNP2HLAの結果ファイルは、3種類に分別されます。

- ①: ジェノタイプ毎の存在確率(imputed dosage: 小数)
- ②: 最も存在確率の高いジェノタイプ(best guess genotype: 整数)
- ③: 各変異ごとの推定精度

③ SNP2HLAを使ったHLA imputation法

Analysis: imputed SNPs

Analysis procedure on imputed SNPs

Accounts for genotype uncertainty?

Includes correction for population stratification?

Includes additional risk factors?

Analysis performed on imputed SNPs?

Removal of poorly imputed SNPs based on MACH R^2 or SNPTEST criteria?

Genomic inflation factor estimated?

P -values corrected for inflation?

Exchange file prepared?

Rs identifier

Chromosomal position

Strand orientation of allele (+/-)

Coded and noncoded allele

Allele frequency of the coded allele

Odds ratio

Beta and SE (for regression modeling)

Test statistic and P -value

- Imputation作業の不確かさを考慮して、imputation後ジェノタイプデータを用いた**関連解析にはimputed dosage(小数)の使用が推奨**されます。
- Imputed dosageに対応した関連解析ソフトも、増えてきています。
(最新版のPLINK形式では、imputed dosage/best-guess genotypeの両者を扱うことができます)

終わりに

- SNP genotype imputationおよびHLA imputation法について、簡単に
なぞってみました。
- Imputationの開発により、未観測のジェノタイプデータを推定可能になっ
たことで、疾患ゲノム解析の幅が大きく広がりました。
- より高精度のimputationを実現するために、多サンプルかつ高密度の
参照データを、各集団において構築する研究が進んでいます。
- Imputationの過程では確率的な情報を扱うため、解析プロトコルが間
違っていると、現実と乖離した結果が得られることがあります。
- ツールの特性を把握しつつ、正しいimputationの実施を心掛けて下さい。