

GenomeDataAnalysis2

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>

GenomeDataAnalysis2

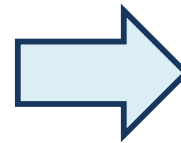
- ① 遺伝統計学における関連解析
- ② PLINKを使ったゲノムワイド関連解析
- ③ 遺伝子発現量を対象としたeQTL解析

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。

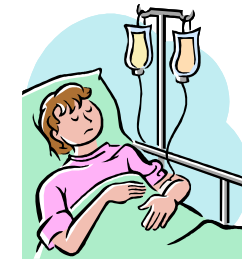
① 遺伝統計学における関連解析



?

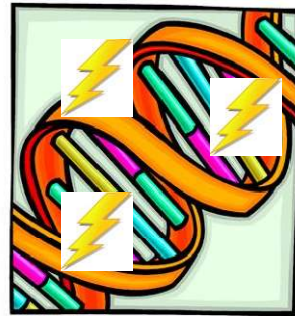


or

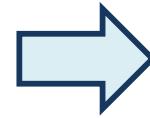


- **何故、人は病気になるのでしょうか。**
- **怪我、加齢、生活習慣、食事、色々な事象が原因で病気になります。**
- **一生の間、一度も病気にならない人はいないと思われれます。**
- **「病気になる」ことは必ずしも異常な状況ではなく、個性(個人間の形質の違いの一つ)、という捉え方もできます。**

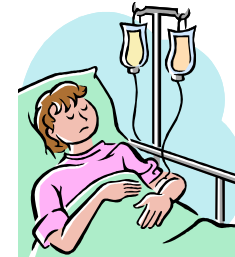
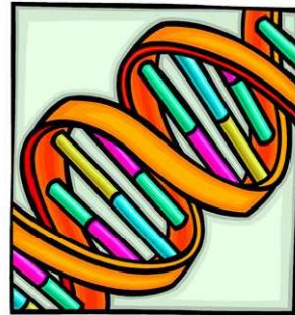
① 遺伝統計学における関連解析



?

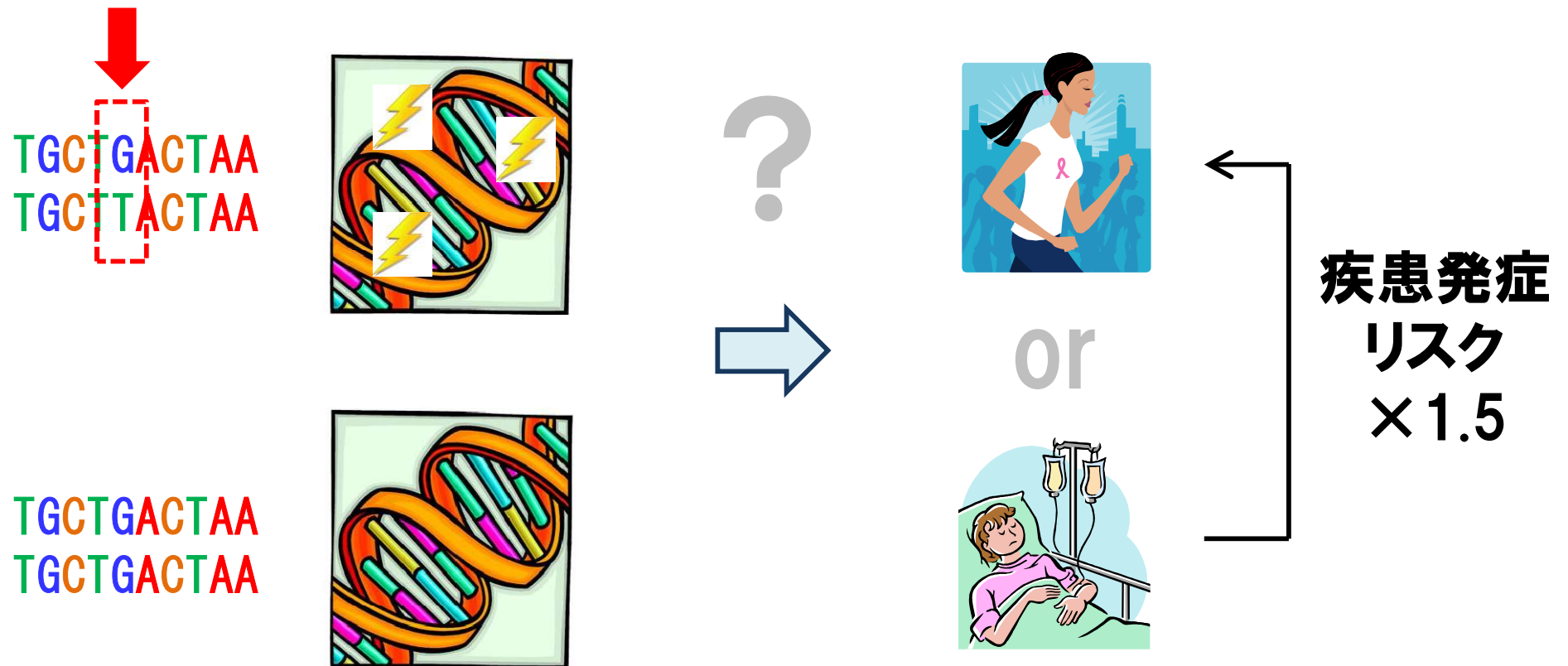


or



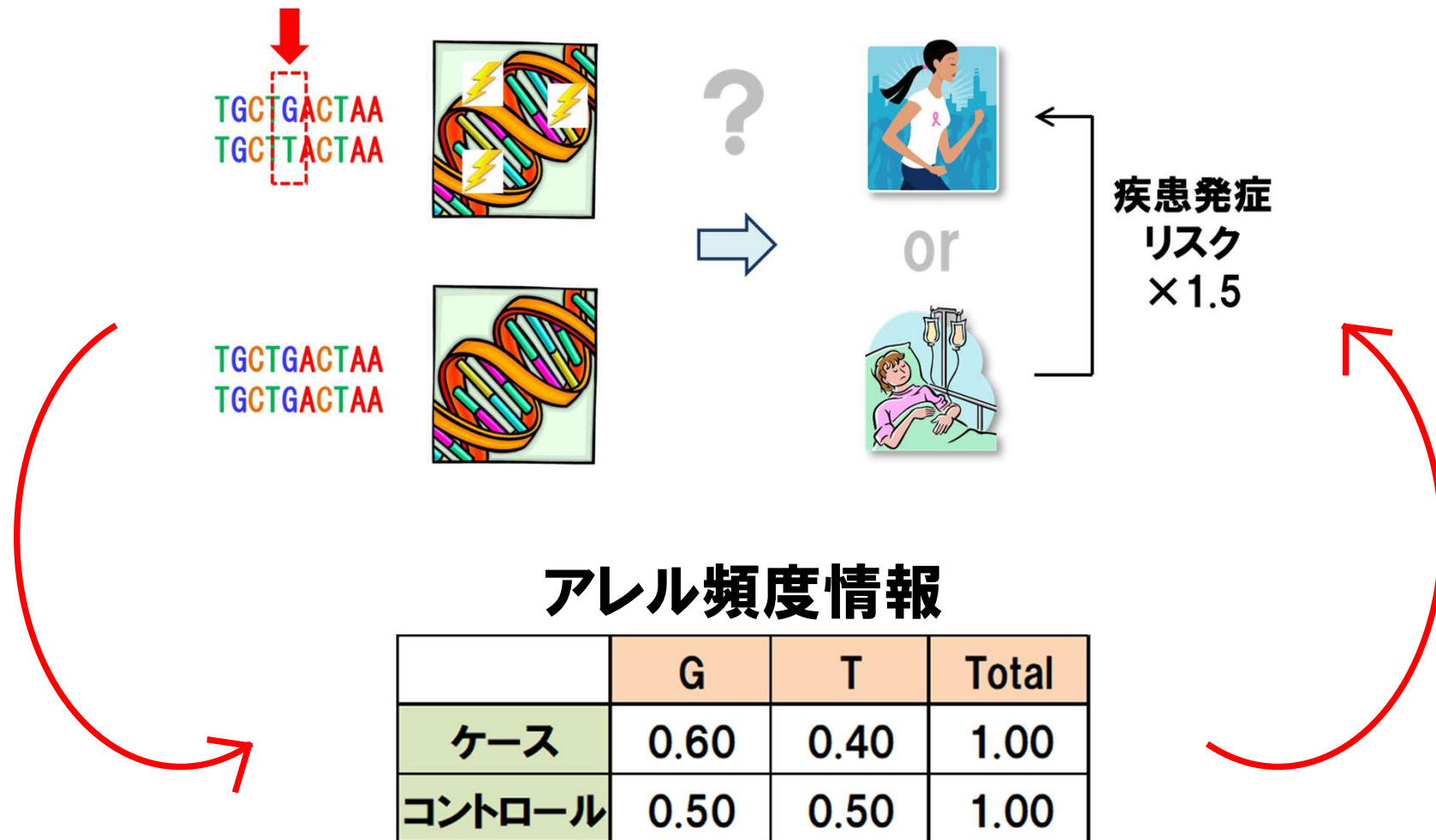
- 一方で、「生まれながらにして、特定の病気(形質)へのなりやすさが存在し、個人間で違いがある」ということを、私達は経験的に知っています。
- 生まれながらにして個人間で異なる現象は、個人の遺伝的背景です。
- つまり、ゲノム配列の個人差により、疾患発症リスクが異なります。

① 遺伝統計学における関連解析



- ゲノム配列の個人差により疾患発症リスクがどの程度変化するのか、を具体的に定量化することで評価を行うのが、遺伝統計学における関連解析 (Association Study) です。

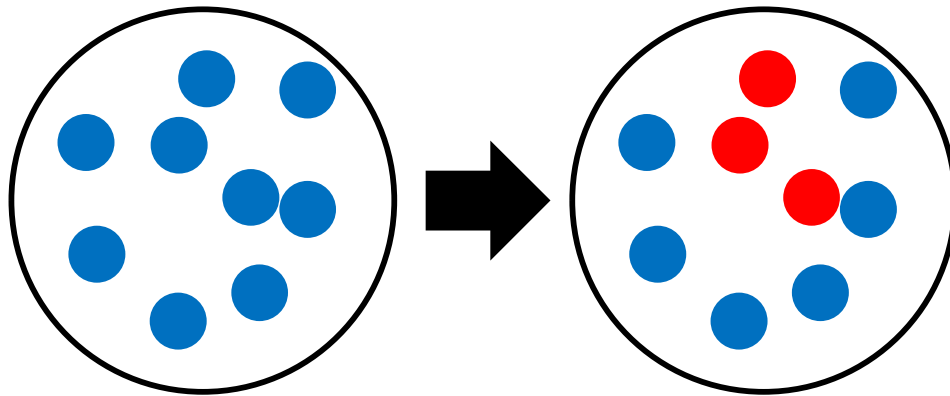
① 遺伝統計学における関連解析



- 疾患発症リスクを定量化するためには、どうすればいいのでしょうか？
- 疾患を発症した人(ケース)と、発症していない人(コントロール)とで、各遺伝子多型のアレル頻度を比較することで達成できます。

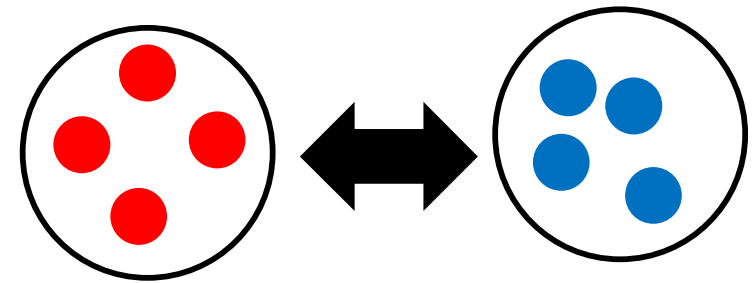
① 遺伝統計学における関連解析

コホート研究



経時的に観測

ケースコントロール研究

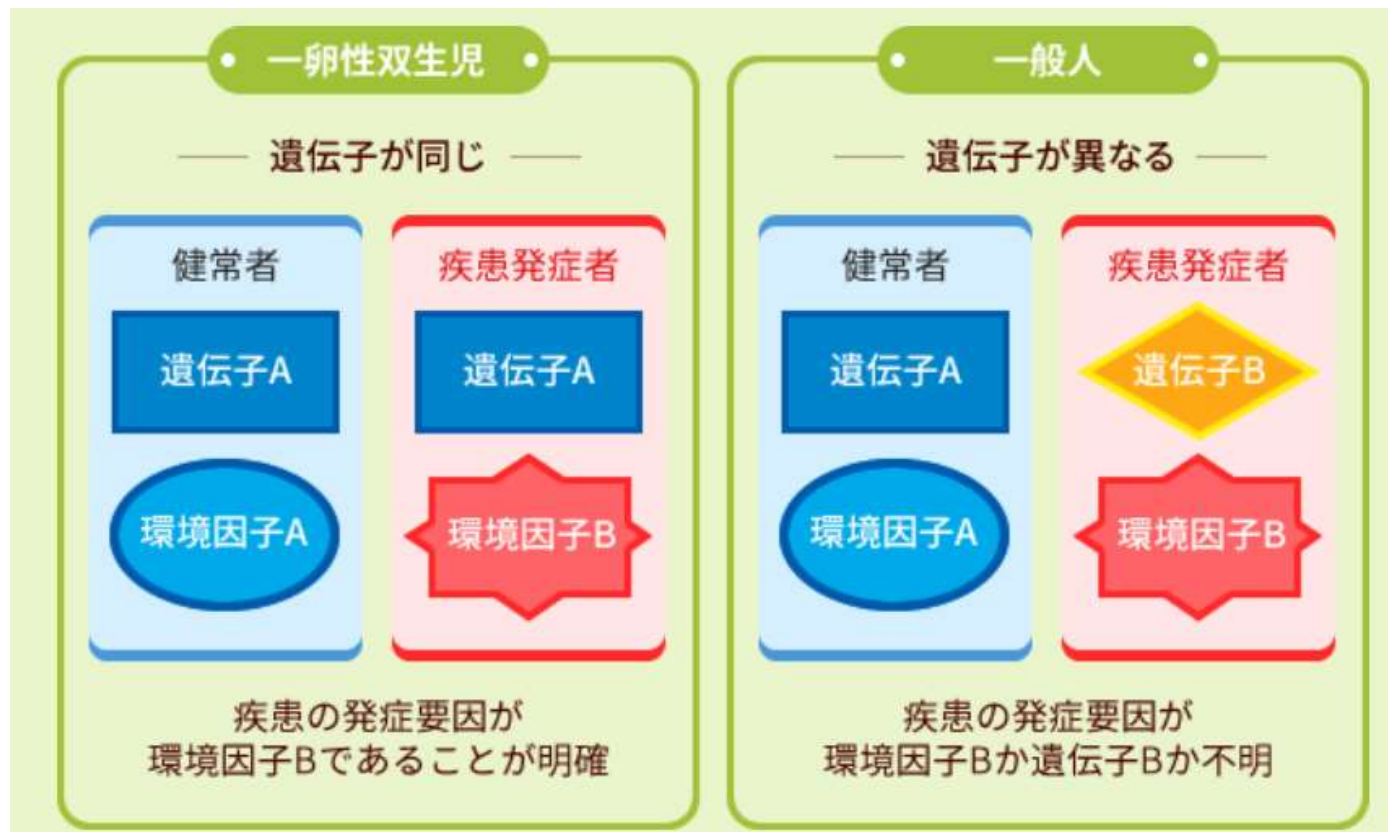


別々に収集し比較

- ・集団を経時的に観測し疾患発症に関わる情報を得るのが**コホート研究**、ケースとコントロールを別々に集めるのが**ケースコントロール研究**です。
- ・遺伝統計解析ではどちらも扱いますが、集団中の発症者の割合が高くない疾患の場合、十分なケース数を確保する目的で、ケースコントロール研究を選択する例が多いです。

① 遺伝統計学における関連解析

双子研究



(大阪大学大学院医学系研究科ツインリサーチセンターHPより)

- 遺伝的背景が疾患発症リスクのどの程度を決定するか、を調べるためには、**双子研究**が有効です。
- 一卵性双生児と二卵性双生児、一般集団を比較することで、**遺伝的背景と環境因子が疾患発症に与える相対的なリスク**を推定できます。⁸

① 遺伝統計学における関連解析

アレル頻度比較 (2×2分割表)

	G	T	Total
ケース	120	80	200
コントロール	100	100	200

カイ二乗検定(df=1)
P = 0.044

ジェノタイプ頻度比較 (2×3分割表)

	GG	GT	TT	Total
ケース	36	48	16	100
コントロール	25	50	25	100

コックラン・アーミテージ検定(df=1)
P = 0.046

※分割表の各群(列)の比率に線型傾向があるかを評価する検定。実質的には線型回帰に等しい。

- 一番シンプルな関連解析は、ケースコントロールの**分割表検定**です。
- アレル頻度(=2×2分割表)やジェノタイプ頻度(=2×3分割表)が、ケース群とコントロール群で異なるかを評価します。
- **カイ二乗検定**や、**コックラン・アーミテージ検定**が使用されます。

① 遺伝統計学における関連解析

odds ratio = ad/bc .

a	b
c	d

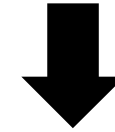
OR = 1.5

	G	T	Total
ケース	60	40	100
コントロール	50	50	100

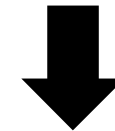
	G	T	Total
ケース	120	80	200
コントロール	100	100	200

	G	T	Total
ケース	240	160	400
コントロール	200	200	400

P = 0.15



P = 0.044



P = 0.0045

- 分割表のオッズ比(odds ratio: OR)を用いて、リスクアレルの保有本数が増える毎に上昇する発症リスクを定量化します。
- 同じオッズ比の場合、サンプル数が多い程、検出力が上昇し、関連解析の結果は有意になり、P値が小さくなります。

① 遺伝統計学における関連解析

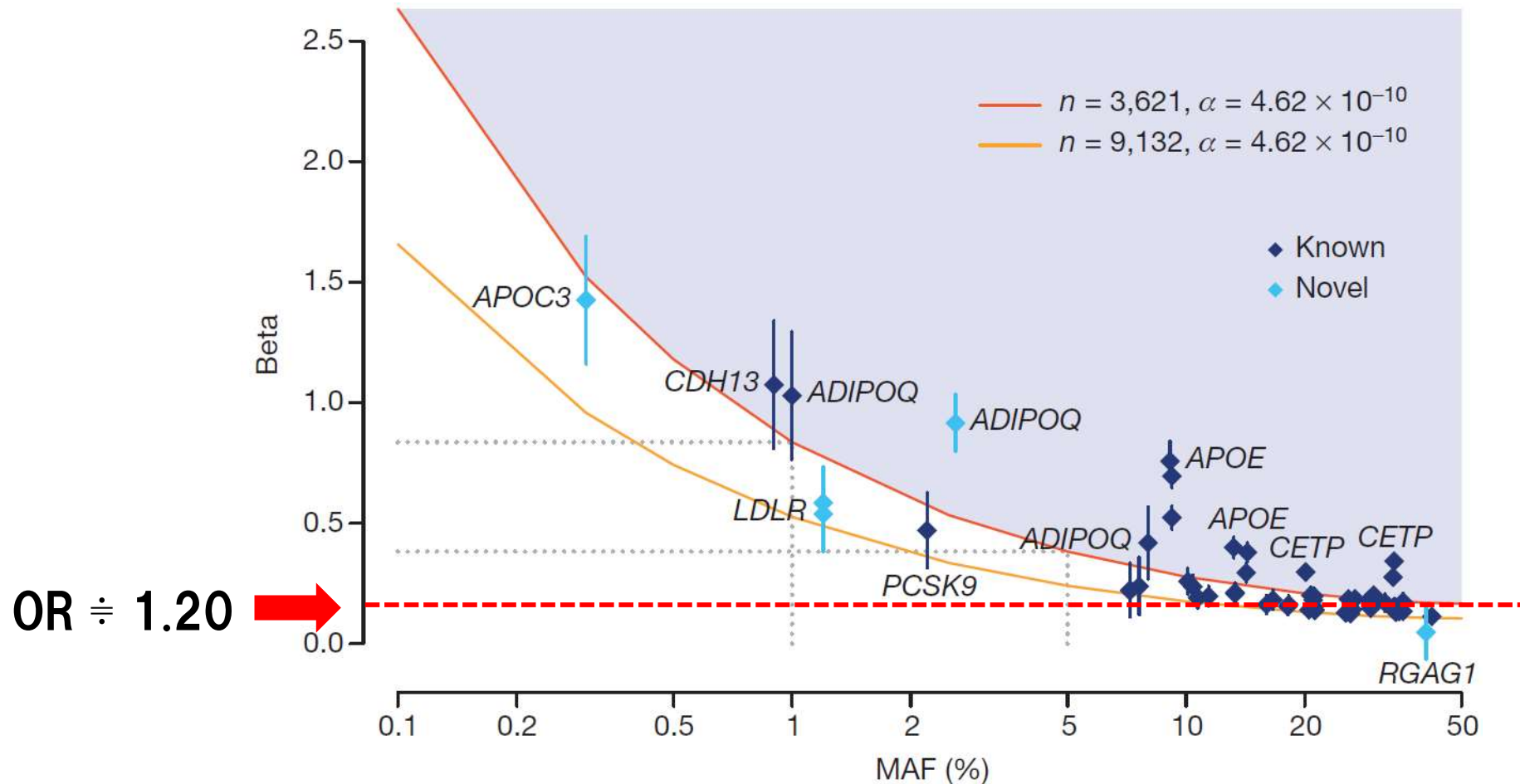


Figure 3 | Summary of association results across the UK10K-cohorts study.

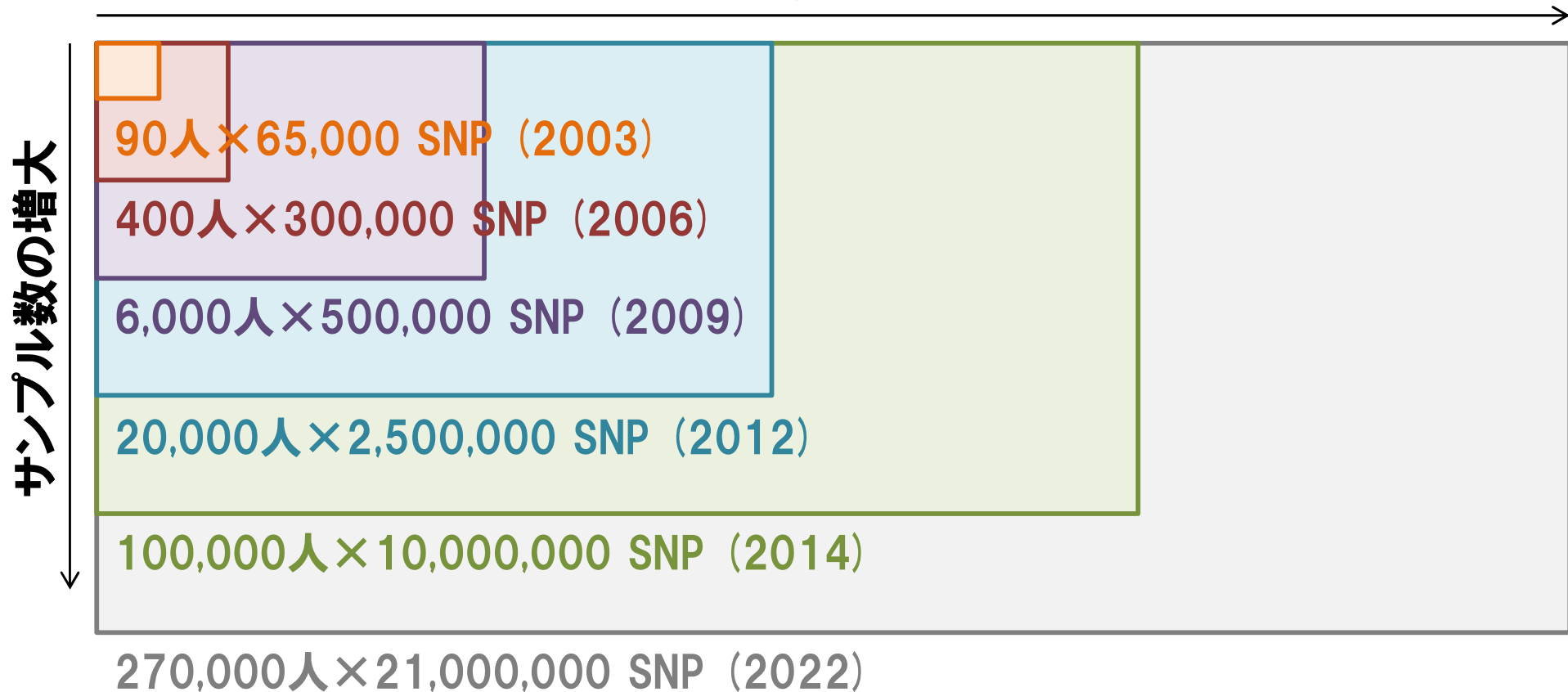
(UK10K. *Nature* 2015)

- 一部の特殊な遺伝子多型(例:HLA遺伝子型)や、遺伝性の希少疾患のリスク遺伝子を除いて、一般的な疾患において、SNPに代表される一般的な遺伝子多型の持つオッズ比は、**大きくても1.20程度**です。

① 遺伝統計学における関連解析

関節リウマチGWASの対象サンプル数・SNP数の推移

SNP数の増大

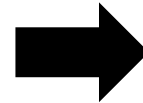


- 疾患発症リスクを有する遺伝子多型をケースコントロール関連解析で同定するには、**大きいサンプルサイズが必要**になります。
- 近年、GWASに代表される疾患ゲノム関連解析に用いられるサンプル数は、増加の一途を辿っています。

① 遺伝統計学における関連解析

ジェノタイプ頻度比較 (2×3分割表)

	GG	GT	TT	Total
ケース	36	48	16	100
コントロール	25	50	25	100



ロジスティック回帰分析

$$\begin{array}{c} \text{疾患} \\ \text{発症} \end{array} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \sim \begin{array}{c} \text{ジェノ} \\ \text{タイプ} \end{array} \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \\ \vdots \\ 2 \\ 0 \\ 1 \\ 2 \end{pmatrix} + \begin{array}{c} \text{年齢} \end{array} \begin{pmatrix} 46 \\ 23 \\ 64 \\ 78 \\ \vdots \\ 72 \\ 39 \\ 24 \\ 19 \end{pmatrix} + \begin{array}{c} \text{性別} \end{array} \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ \vdots \\ 1 \\ 2 \\ 1 \\ 2 \end{pmatrix}$$

- 実際のSNPの関連解析は、**分割表検定よりも回帰分析**で行われます。
- ケースコントロール解析の際は、ロジスティック回帰分析が使われます。
- 年齢や性別等の**共変量**を回帰モデルに組みこむ**重回帰モデル**により、関連解析結果の**バイアス補正が可能になる**点が理由です。

① 遺伝統計学における関連解析

ロジスティック回帰分析

$$\begin{matrix} \text{疾患} \\ \text{発症} \end{matrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \sim \begin{matrix} \text{ジェノ} \\ \text{タイプ} \end{matrix} \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \\ \vdots \\ 2 \\ 0 \\ 1 \\ 2 \end{pmatrix} + \begin{matrix} \text{年齢} \end{matrix} \begin{pmatrix} 46 \\ 23 \\ 64 \\ 78 \\ \vdots \\ 72 \\ 39 \\ 24 \\ 19 \end{pmatrix} + \begin{matrix} \text{性別} \end{matrix} \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ \vdots \\ 1 \\ 2 \\ 1 \\ 2 \end{pmatrix}$$

線形回帰分析

$$\begin{matrix} \text{身長} \end{matrix} \begin{pmatrix} 165 \\ 172 \\ 180 \\ 158 \\ \vdots \\ 171 \\ 167 \\ 166 \\ 170 \end{pmatrix} \sim \begin{matrix} \text{ジェノ} \\ \text{タイプ} \end{matrix} \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \\ \vdots \\ 2 \\ 0 \\ 1 \\ 2 \end{pmatrix} + \begin{matrix} \text{年齢} \end{matrix} \begin{pmatrix} 46 \\ 23 \\ 64 \\ 78 \\ \vdots \\ 72 \\ 39 \\ 24 \\ 19 \end{pmatrix} + \begin{matrix} \text{性別} \end{matrix} \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ \vdots \\ 1 \\ 2 \\ 1 \\ 2 \end{pmatrix}$$

- 身長や臨床検査値など、量的な変数を対象としたSNPの関連解析には、**線形回帰分析**が使われます。
- この場合、アレル毎のオッズ比ではなく、アレル毎の**効果量(effect size)**を求めることになります(=アレル本数あたり身長が何cm伸びるか)。

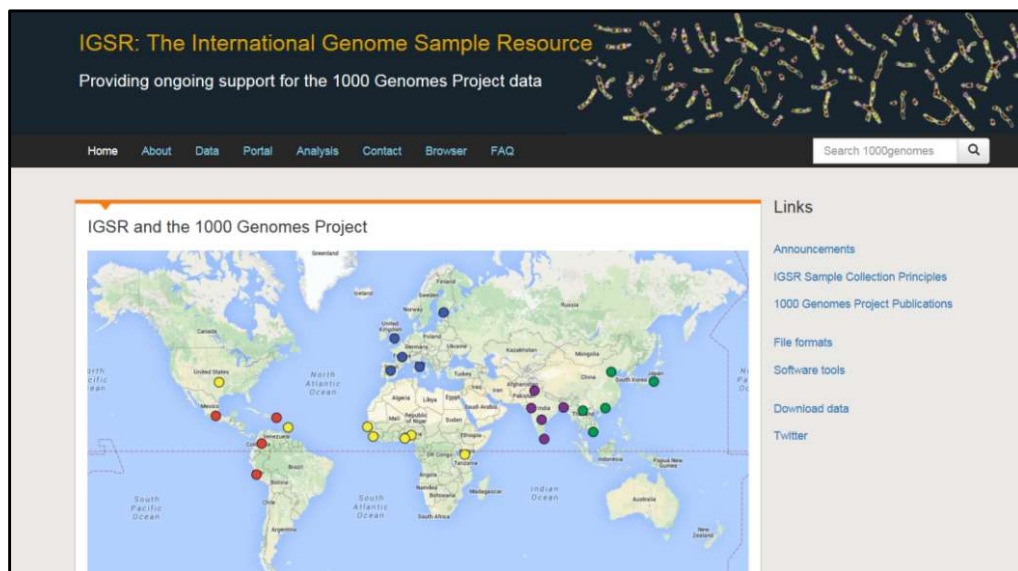
GenomeDataAnalysis2

- ① 遺伝統計学における関連解析
- ② PLINKを使ったゲノムワイド関連解析
- ③ 遺伝子発現量を対象としたeQTL解析

② PLINKを使ったゲノムワイド関連解析

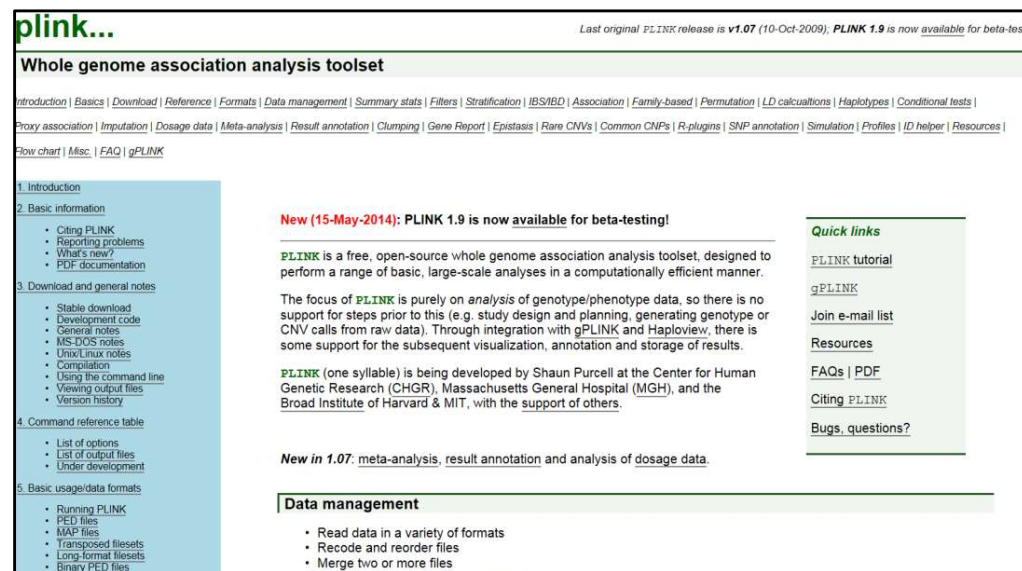
1000 Genomes Project

<http://www.internationalgenome.org/>



PLINK

<http://pngu.mgh.harvard.edu/~purcell/plink/>



- それでは再び、1000 Genomes Projectのジェノタイプデータと、遺伝統計解析ソフトPLINKを用いた、ゲノムデータ実習をやってみましょう。
- 今回はゲノムワイド関連解析を実施してみます。
- Linux上で”./plink”と入力すると、実行されます(復習)。
- PLINKは”./plink --(コマンド)(引数)”で実行します(復習)。

② PLINKを使ったゲノムワイド関連解析

※「GenomeDataAnalysis1」演習内容の復習です。

○:起動

```
./plink
```

○:ファイルの読み込み

```
./plink --bfile 1KG_EUR --out test
```

○:各SNPのアレル頻度の計算

```
./plink --bfile 1KG_EUR --out test1 --freq
```

○:マイナーアレル頻度によるSNPのフィルタリング

```
./plink --bfile 1KG_EUR --out test2 --maf 0.2 --make-bed
```

○:各SNPのHardy-Weinberg平衡の計算

```
./plink --bfile test2 --out test3 --hardy
```

○:サンプル間の遺伝的な近さ(近縁関係)の推定

```
./plink --bfile test2 --out test4 --genome
```

○:サンプルの遺伝的背景の推定

```
./plink --bfile test2 --out test5 --cluster --mds-plot 4
```

• PLINKは”./plink --(コマンド)(引数)”で実行します(復習)。

② PLINKを使ったゲノムワイド関連解析

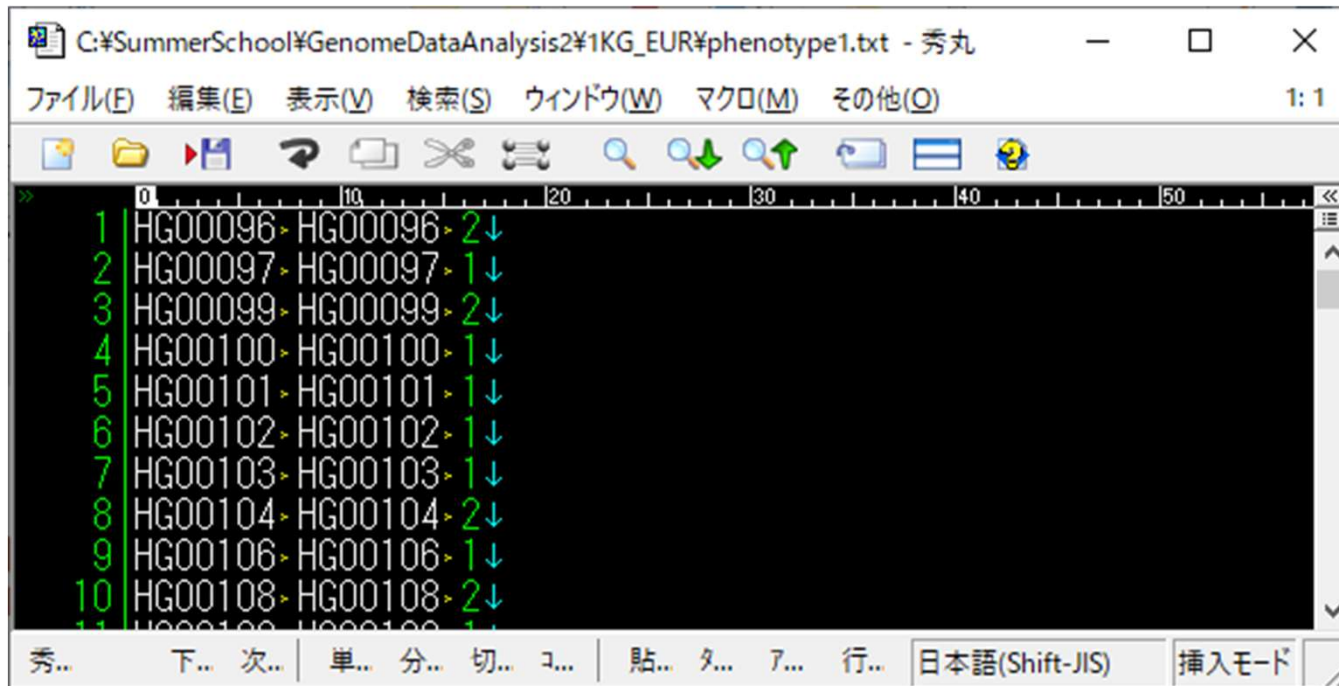
ロジスティック回帰分析

$$\begin{array}{c} \text{疾患} \\ \text{発症} \end{array} \begin{array}{c} \left(\begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right) \sim \begin{array}{c} \text{ジェノ} \\ \text{タイプ} \end{array} \begin{array}{c} \left(\begin{array}{c} 0 \\ 1 \\ 2 \\ 0 \\ \vdots \\ 2 \\ 0 \\ 1 \\ 2 \end{array} \right) + \begin{array}{c} \text{年齢} \\ \left(\begin{array}{c} 46 \\ 23 \\ 64 \\ 78 \\ \vdots \\ 72 \\ 39 \\ 24 \\ 19 \end{array} \right) + \begin{array}{c} \text{性別} \\ \left(\begin{array}{c} 1 \\ 1 \\ 2 \\ 2 \\ \vdots \\ 1 \\ 2 \\ 1 \\ 2 \end{array} \right) \end{array}$$

- ロジスティック回帰分析を用いたケースコントロールGWASを実施します。
- 1000 Genomes Projectのサンプルは一般人集団で構成されており、**疾患罹患情報は付随していない**ため、ランダムにケース群とコントロール群を割り振ることにします。
(もちろん、一般人集団でも何らかの疾患に罹患している可能性は否定できません。)

② PLINKを使ったゲノムワイド関連解析

※ファイル”phenotype1.txt”を開いて、
内容を確認してみてください。



```
C:\SummerSchool\GenomeDataAnalysis2\1KG_EUR\phenotype1.txt - 秀丸
ファイル(F) 編集(E) 表示(V) 検索(S) ウィンドウ(W) マクロ(M) その他(O) 1:1
>>
1 HG00096 HG00096 2↓
2 HG00097 HG00097 1↓
3 HG00099 HG00099 2↓
4 HG00100 HG00100 1↓
5 HG00101 HG00101 1↓
6 HG00102 HG00102 1↓
7 HG00103 HG00103 1↓
8 HG00104 HG00104 2↓
9 HG00106 HG00106 1↓
10 HG00108 HG00108 2↓
11 HG00109 HG00109 1↓
秀.. 下.. 次.. 単.. 分.. 切.. コ.. 貼.. タ.. ア.. 行.. 日本語(Shift-JIS) 挿入モード
```

• PLINKの形質ファイルの構成は、下記の通りです。

1列目: Family ID

2列目: Sample ID

3列目(以降): Phenotype(ケース=2、コントロール=1)

(サンプルの構成や並び順は、元のジェノタイプデータと同一でなくて大丈夫です)

② PLINKを使ったゲノムワイド関連解析

※Linuxコマンド”ls”や”wc”で、ファイルの構成やサンプル数・SNP数を確認して下さい。

- 1KG_EUR.bed
- 1KG_EUR.bim
- 1KG_EUR.fam



- 1KG_EUR_QC.bed
- 1KG_EUR_QC.bim
- 1KG_EUR_QC.fam

381 Sample

8,830,185 SNP

--maf 0.05

--hwe 0.000001

--indep-pairwise 100 5 0.8

381 Sample

1,349,118 SNP

- 演習に伴うデータ計算時間やディスク容量の軽減目的で、前回使ったジェノタイプデータにフィルタリングを実施し、SNP数を減らしています。
(--indep-pairwiseによるLD関係にあるSNPの除去は、一般的には実施しません。)

② PLINKを使ったゲノムワイド関連解析

○:ゲノムワイド関連解析の実施(ロジスティック回帰分析)

```
./plink --bfile 1KG_EUR_QC --out 1KG_EUR_QC_Pheno1 --pheno  
phenotype1.txt --logistic --ci 0.95
```

※Cygwinの場合plinkをplink.exeに変えてください



出力ファイル: 1KG_EUR_QC_Pheno1.assoc.logistic

※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。

※実行ファイルにアクセス権限を与える必要がある場合があります。

- “**--pheno**”で、関連解析に用いる形質ファイルを指定します。
(ped/famファイルに元々書き込まれていた形質情報より優先されます。)
- “**--logistic**”で、各SNPにおけるロジスティック回帰分析を実施します。
- “**--ci**”で、効果サイズの信頼区間を追加で出力します。

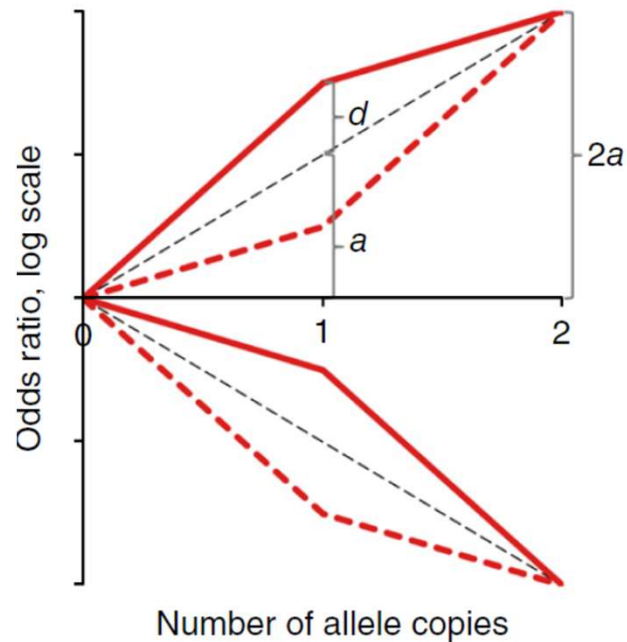
② PLINKを使ったゲノムワイド関連解析

出力ファイル: 1KG_EUR_QC_Pheno1.assoc.logistic

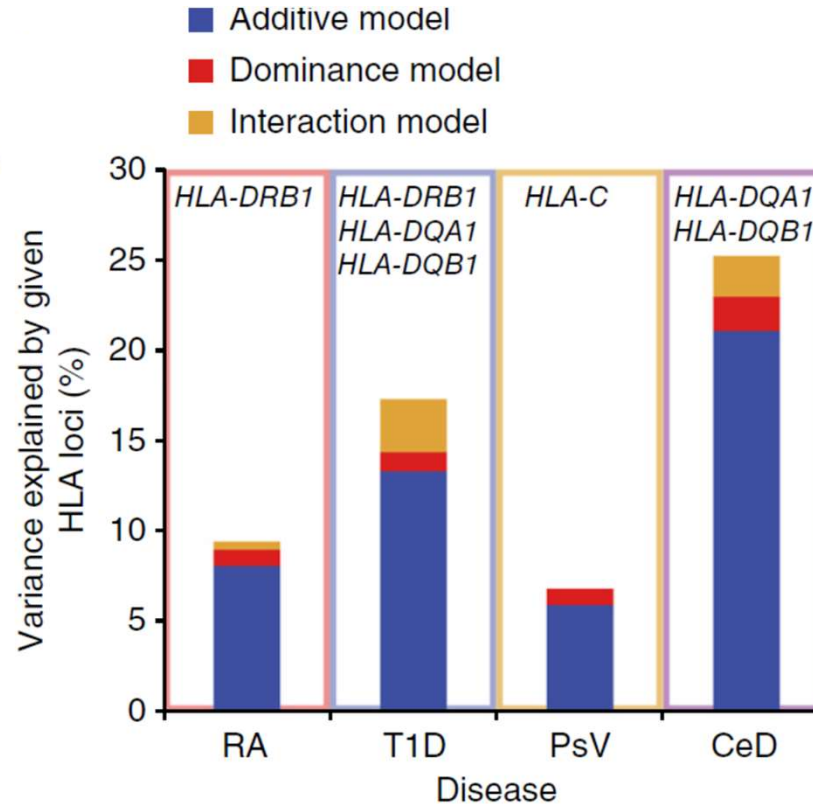
CHR	SNP	BP	A1	TEST	NMISS	OR	SE	L95	U95	STAT	P
1	rs55998931	10492	T	ADD	381	1.304	0.3075	0.7135	2.381	0.862	0.3887
1	chr1:10622	10622	T	ADD	381	1.034	0.1685	0.7429	1.438	0.1968	0.844
1	chr1:10623	10623	T	ADD	381	1.805	0.2519	1.102	2.957	2.344	0.01907
1	chr1:11187	11187	G	ADD	381	0.9023	0.1688	0.6481	1.256	-0.6094	0.5423
1	chr1:11409	11409	G	ADD	381	0.9035	0.1817	0.6328	1.29	-0.5582	0.5767
1	chr1:11457	11457	C	ADD	381	1.131	0.1819	0.7919	1.615	0.6768	0.4985
1	chr1:11508	11508	G	ADD	381	0.9041	0.2154	0.5927	1.379	-0.4678	0.64
1	chr1:11542	11542	A	ADD	381	0.9421	0.2078	0.6269	1.416	-0.2871	0.774
1	chr1:11565	11565	G	ADD	381	1.022	0.1775	0.7217	1.447	0.122	0.9029
1	chr1:11677	11677	G	ADD	381	1.04	0.1566	0.7654	1.414	0.2527	0.8005

- "CHR": 染色体番号
- "SNP": SNP ID
- "BP": 染色体上の位置(base pair)
- "A1": アレル1
- "TEST": ジェノタイプ効果
- "NMISS": サンプル数
- "OR": (アレル1の)オッズ比
- "SE": 効果サイズ(OR対数値)のSE
- "L95": オッズ比95%CI下限
- "U95": オッズ比95%CI上限
- "STAT": 統計量
- "P": P値

② PLINKを使ったゲノムワイド関連解析



(Lentz TL et al. *Nat Genet* 2015)



- ジェノタイプにおいて、**リスクアレル保有数の増加とリスクの増加との関係性が線型**なとき、“**additive model**”と呼ばれます。
- 希少難病の原因遺伝子変異では、線型から外れる“**dominance model**”や“**recessive model**”の存在が知られています。
- 特殊な遺伝子座位(HLA遺伝子)を除き、ほとんどのコモンバリエントは **additive model**に従う例が多いと考えられています。

② PLINKを使ったゲノムワイド関連解析

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis2/1KG_EUR

```
$ awk '{print $2"¥t"($1*100000000000+$3)"¥t"$12}'
```

```
1KG_EUR_QC_Pheno1.assoc.logistic > 1KG_EUR_QC_Pheno1.assoc.logistic.P.txt
```

CHR	SNP	BP	A1	TEST	NMISS	OR	SE	L95	U95	STAT	P
1	rs55998931	10492	T	ADD	381	1.304	0.3075	0.7135	2.381	0.862	0.3887
2	chr1:10622	10622	T	ADD	381	1.034	0.1685	0.7429	1.438	0.1968	0.844
3	chr1:10623	10623	T	ADD	381	1.805	0.2519	1.102	2.957	2.344	0.01907
4	chr1:11187	11187	G	ADD	381	0.9023	0.1688	0.6481	1.256	-0.6094	0.5423
5	chr1:11409	11409	G	ADD	381	0.9035	0.1817	0.6328	1.29	-0.5582	0.5767
6	chr1:11457	11457	C	ADD	381	1.131	0.1819	0.7919	1.615	-0.6768	0.4985
7	chr1:11508	11508	G	ADD	381	0.9041	0.2154	0.5927	1.379	-0.4678	0.64
8	chr1:11542	11542	A	ADD	381	0.9421	0.2078	0.6269	1.416	-0.2871	0.774
9	chr1:11565	11565	G	ADD	381	1.022	0.1775	0.7217	1.447	0.122	0.9029
10	chr1:11677	11677	G	ADD	381	1.04	0.1566	0.7654	1.414	0.2527	0.8005

\$1:染色体番号 \$2:SNP ID \$3:染色体上の位置 \$12:P値

- GWAS結果のP値を取り出してみましよう。
- PLINK結果ファイルは、不定数の半角スペースで区切られていて扱いにくいですが、AWKコマンドを使うと処理が簡単です。

② PLINKを使ったゲノムワイド関連解析

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis2/1KG_EUR

```
$ awk '{print $2"¥t"($1*1000000000000+$3)"¥t"$12}'
```

```
1KG_EUR_QC_Pheno1.assoc.logistic > 1KG_EUR_QC_Pheno1.assoc.logistic.P.txt
```

\$1:染色体番号 **→ 染色体番号×1000億+染色体上の位置**

\$3:染色体上の位置

\$12:P値

1KG_EUR_QC_Pheno1.assoc.logistic.P.txt

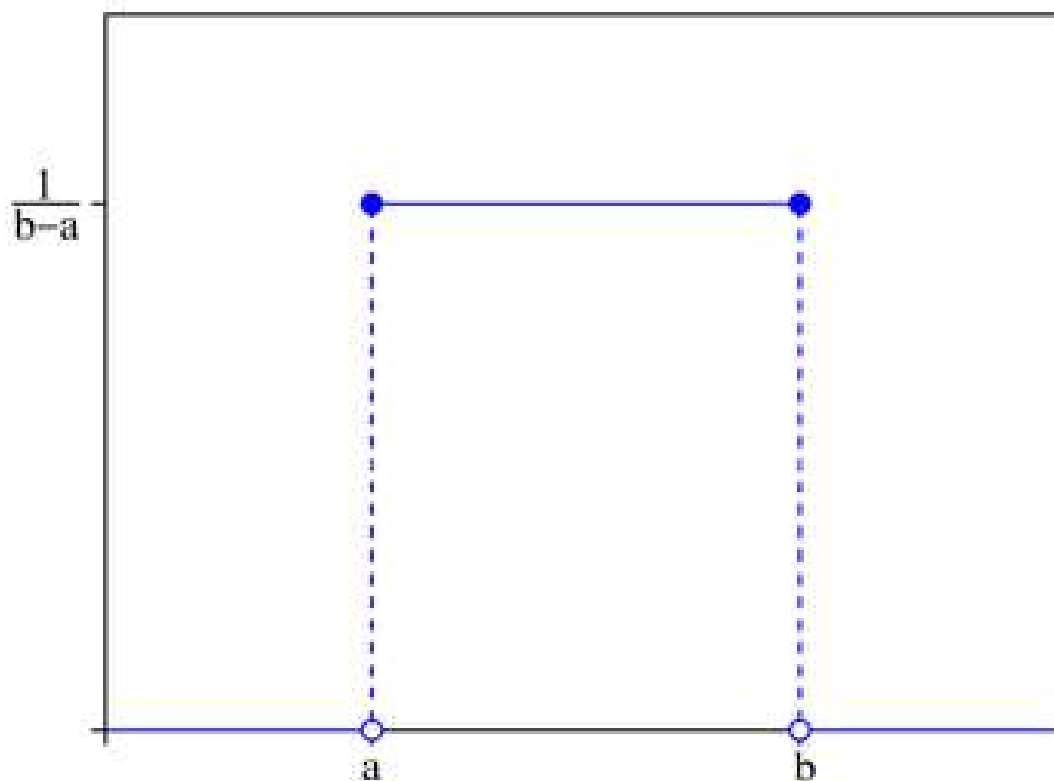
1	SNP	Position	P
2	rs55998931	100000010492	0.3887↓
3	chr1:10622	100000010622	0.844↓
4	chr1:10623	100000010623	0.01907↓
5	chr1:11187	100000011187	0.5423↓
6	chr1:11409	100000011409	0.5767↓
7	chr1:11457	100000011457	0.4985↓

• **SNPの位置情報**を表すのは、**一つの数字で間に合います。**

(×1000億という単位は主催者の恣意的な選択であり、国際標準ではありません。)

② PLINKを使ったゲノムワイド関連解析

連続一様分布

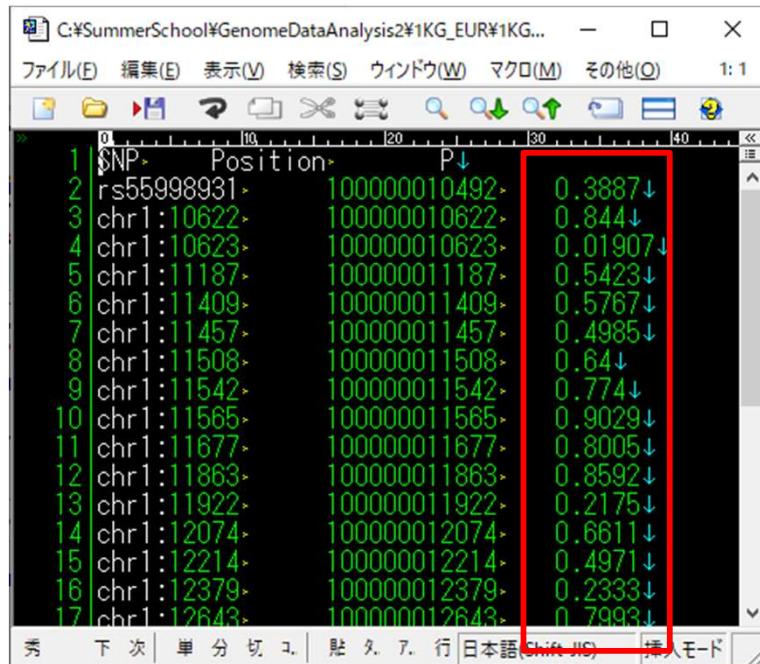


- GWASを実施した結果、**135万個の全SNPに対するP値**が得られました。
- P値の分布をみてみましょう。
- 今回はランダムに形質データを作ったので、帰無仮説下の分布に従う、つまり**P値は0～1の一様分布に従う**、と期待されます。

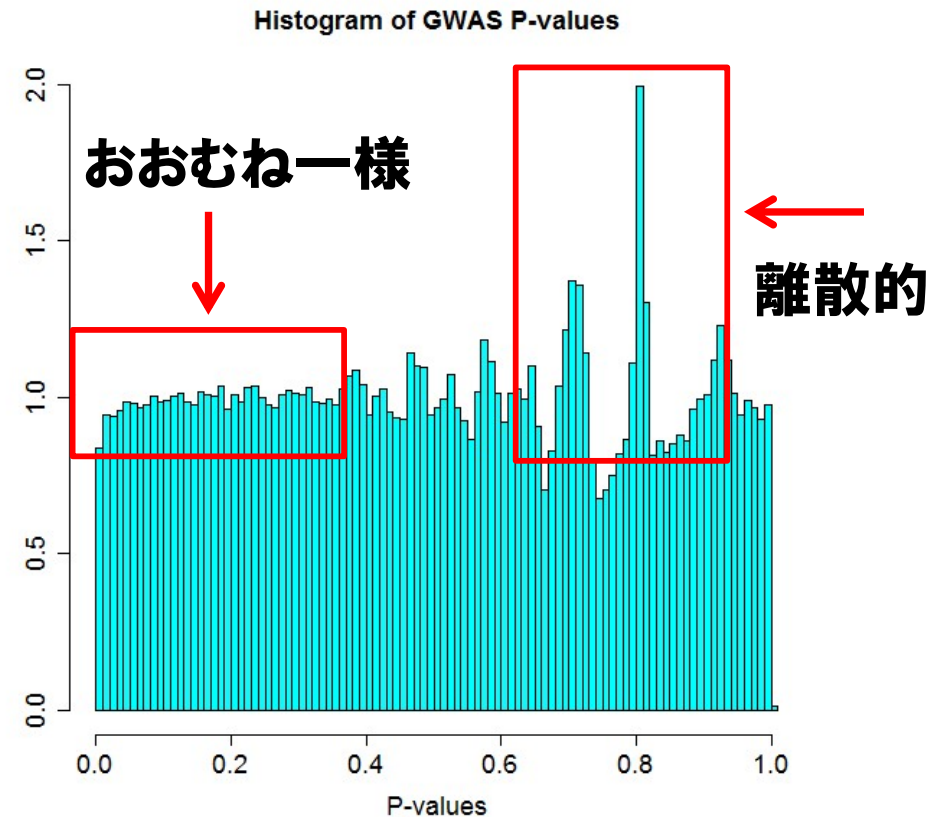
② PLINKを使ったゲノムワイド関連解析

1KG_EUR_QC_Pheno1.assoc.logistic.P.txt

離散一様分布 + 連続一様分布



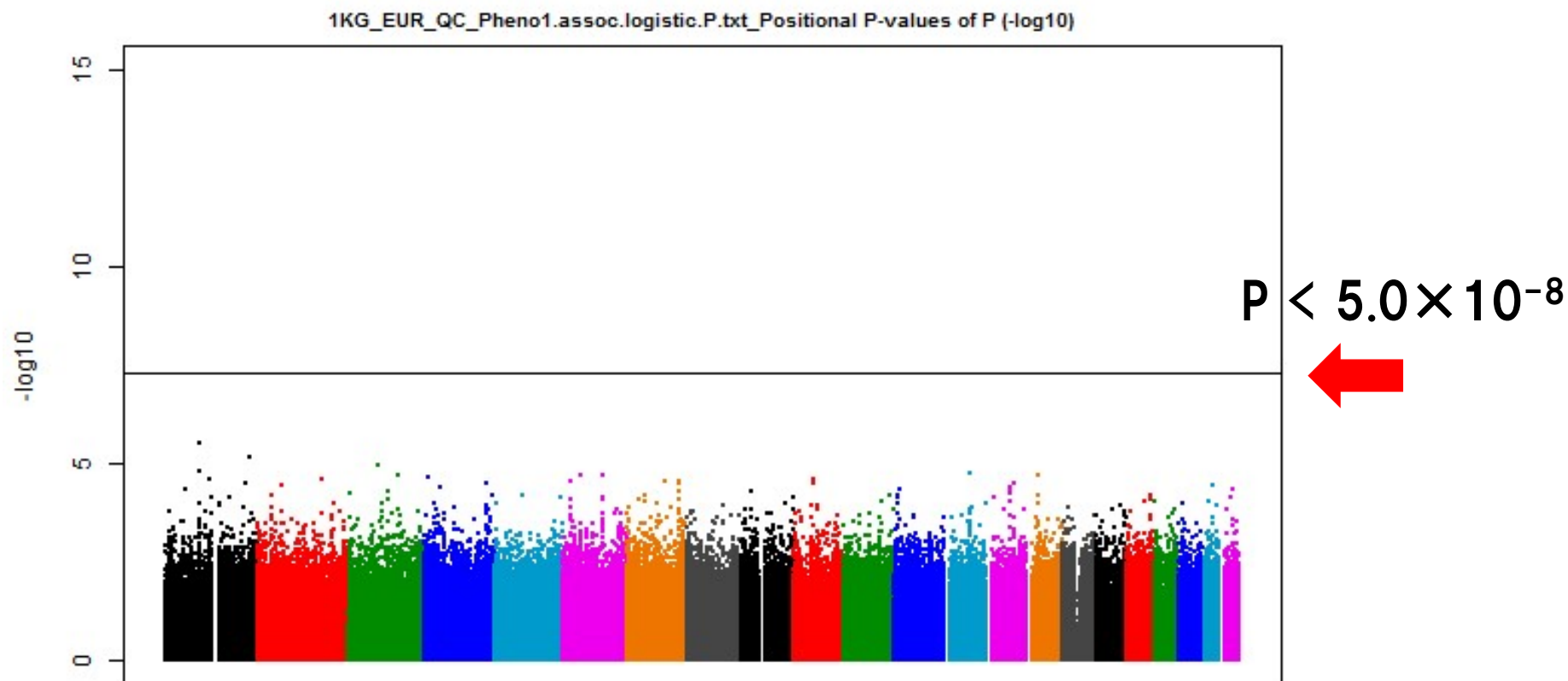
SNP	Position	P
rs55998931	100000010492	0.3887↓
chr1:10622	100000010622	0.844↓
chr1:10623	100000010623	0.01907↓
chr1:11187	100000011187	0.5423↓
chr1:11409	100000011409	0.5767↓
chr1:11457	100000011457	0.4985↓
chr1:11508	100000011508	0.64↓
chr1:11542	100000011542	0.774↓
chr1:11565	100000011565	0.9029↓
chr1:11677	100000011677	0.8005↓
chr1:11863	100000011863	0.8592↓
chr1:11922	100000011922	0.2175↓
chr1:12074	100000012074	0.6611↓
chr1:12214	100000012214	0.4971↓
chr1:12379	100000012379	0.2333↓
chr1:12643	100000012643	0.7993↓



※ファイル”HistogramPlot.R”を開いて、内容をRにコピー&ペーストして下さい。

- 一様分布に近いP値の分布が確認できました。
- 実際には、連続一様分布と離散一様分布を混ぜ合わせた分布になります(381名という少人数で構成される分割表の多様性が離散的なためです)。

② PLINKを使ったゲノムワイド関連解析



※ファイル”ManhattanPlot.R”を開いて、内容をRにコピー&ペーストして下さい。

- マンハッタンプロットを書いてみましょう。
- 今回は、ランダムな形質値を与えているため、**ゲノムワイド水準**($P < 5.0 \times 10^{-8}$)を満たすSNPは認められませんでした。

② PLINKを使ったゲノムワイド関連解析

線形回帰分析

$$\begin{array}{c} \text{Coordinate 1} \\ \left(\begin{array}{c} 0.0068 \\ 0.0031 \\ -0.0017 \\ 0.0080 \\ \vdots \\ 0.0051 \\ -0.0032 \\ 0.0030 \\ -0.0028 \end{array} \right) \sim \left(\begin{array}{c} 0 \\ 1 \\ 2 \\ 0 \\ \vdots \\ 2 \\ 0 \\ 1 \\ 2 \end{array} \right) + \left(\begin{array}{c} 46 \\ 23 \\ 64 \\ 78 \\ \vdots \\ 72 \\ 39 \\ 24 \\ 19 \end{array} \right) + \left(\begin{array}{c} 1 \\ 1 \\ 2 \\ 2 \\ \vdots \\ 1 \\ 2 \\ 1 \\ 2 \end{array} \right)\end{array}$$

- 続いて、線形回帰分析を用いたGWASを実施します。
- 形質情報には、量的変数が必要です。
- 試しに、多次元尺度構成法(MDS)によるサンプルのクラスタリング解析で得た第1座標を形質として使ってみましょう。

② PLINKを使ったゲノムワイド関連解析

○:ゲノムワイド関連解析の実施(線形回帰分析)

```
./plink --bfile 1KG_EUR_QC --out 1KG_EUR_QC_MDS1 --pheno MDS_1.txt --  
linear --ci 0.95
```

※Cygwinの場合plinkをplink.exeに変えてください



1KG_EUR_QC_MDS1.assoc.linear

MDS_1.txt

FID	IID	C1
HG00096	HG00096	0.00686842↓
HG00097	HG00097	0.00310107↓
HG00099	HG00099	-0.00176052↓
HG00100	HG00100	0.00801577↓
HG00101	HG00101	-0.000814574↓
HG00102	HG00102	-0.00375586↓
HG00103	HG00103	0.00245805↓
HG00104	HG00104	-0.00324035↓
HG00106	HG00106	0.00515692↓
HG00108	HG00108	0.000748419↓
HG00109	HG00109	0.00391991↓
HG00110	HG00110	-0.00289634↓
HG00111	HG00111	-0.00125867↓
HG00112	HG00112	-0.000192971↓
HG00113	HG00113	0.00302745↓
HG00114	HG00114	0.00366942↓

• “--linear”で、各SNPにおける線形回帰解析を実施します。

② PLINKを使ったゲノムワイド関連解析

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis2/1KG_EUR

```
$ awk '{print $2"¥t"($1*100000000000+$3)"¥t"$12}' 1KG_EUR_QC_MDS1.assoc.linear  
> 1KG_EUR_QC_MDS1.assoc.linear.P.txt
```

CHR	SNP	BP	TEST	NMISS	BETA	SE	L95	U95	STAT	P
1	rs55998931	10492	T	381	0.007537	0.004092	-0.0004838	0.01556	1.842	0.0663
1	chr1:10622	10622	T	381	-0.006048	0.002242	-0.01044	-0.001654	-2.698	0.00729
1	chr1:10623	10623	T	381	-0.01629	0.003212	-0.02258	-0.00999	-5.07	6.235e-07
1	chr1:11187	11187	G	381	-0.005836	0.002244	-0.01023	-0.001439	-2.601	0.009646
1	chr1:11409	11409	G	381	-0.001292	0.002437	-0.006068	0.003484	-0.5303	0.5962
1	chr1:11457	11457	C	381	0.001694	0.002438	-0.003084	0.006473	0.695	0.4875
1	chr1:11508	11508	G	381	-0.007857	0.002862	-0.01347	-0.002246	-2.745	0.006344
1	chr1:11542	11542	A	381	-0.009322	0.002747	-0.01471	-0.003939	-3.394	0.0007609
1	chr1:11565	11565	G	381	0.004328	0.002373	-0.0003223	0.008979	1.824	0.06892

\$1:染色体番号

\$2:SNP ID

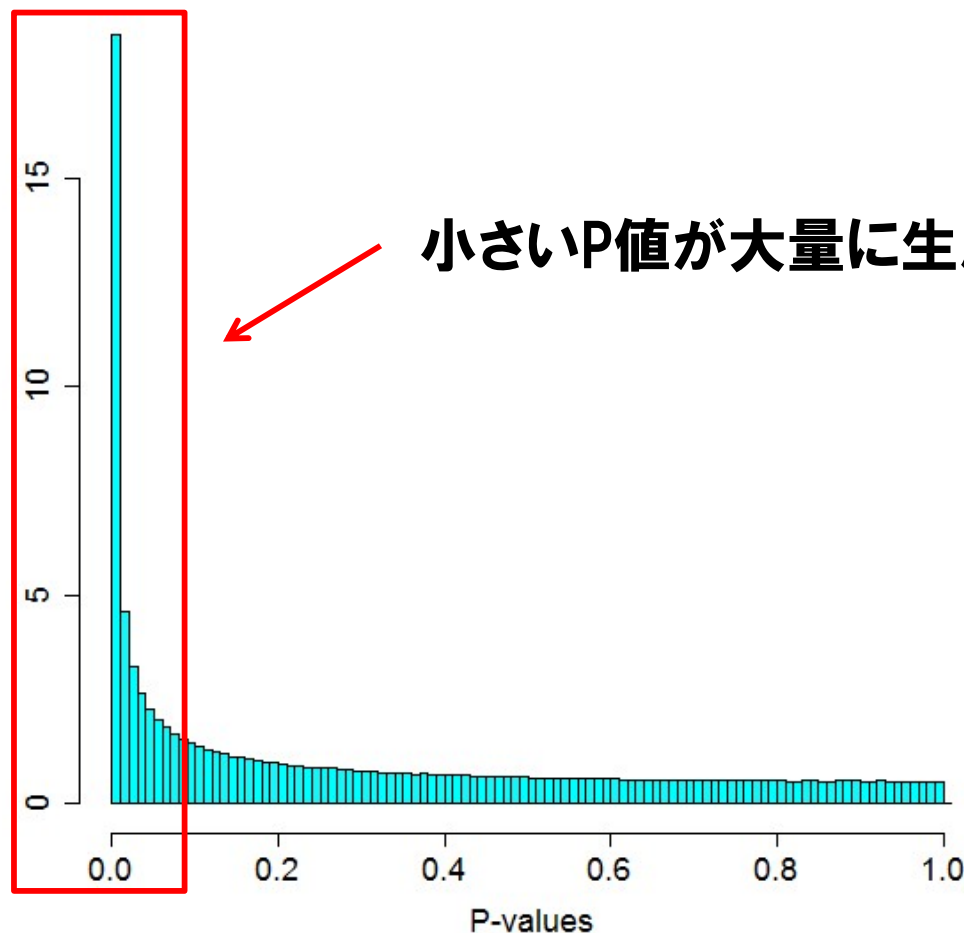
\$3:染色体上の位置

\$12:P値

- 線形回帰分析の結果ファイルは、ロジスティック回帰分析の結果ファイルと類似の構成になっています。
- 同様に、AWKコマンドを使ってP値を取り出してみましょう。

② PLINKを使ったゲノムワイド関連解析

Histogram of GWAS P-values

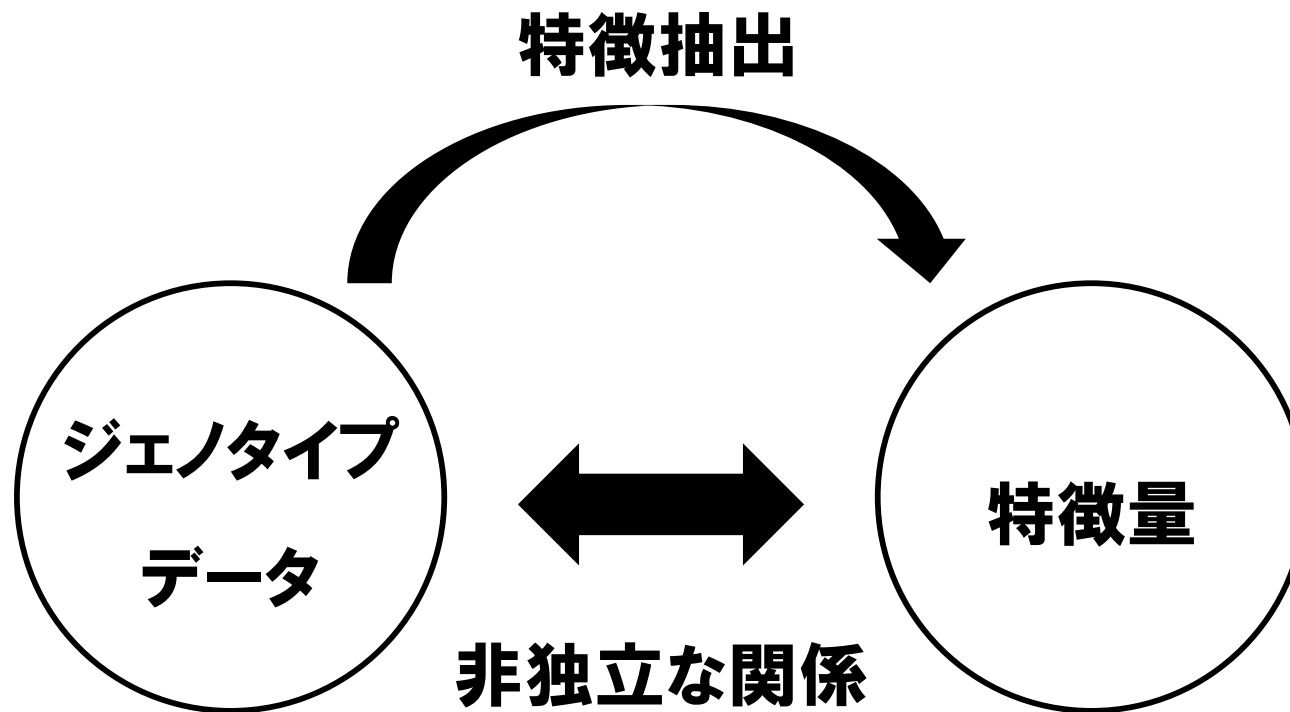


小さいP値が大量に生成されている。

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

- P値のヒストグラムを書いたら、**一様分布に従わない分布**となりました。
- 何故、偏った分布が得られたのでしょうか？

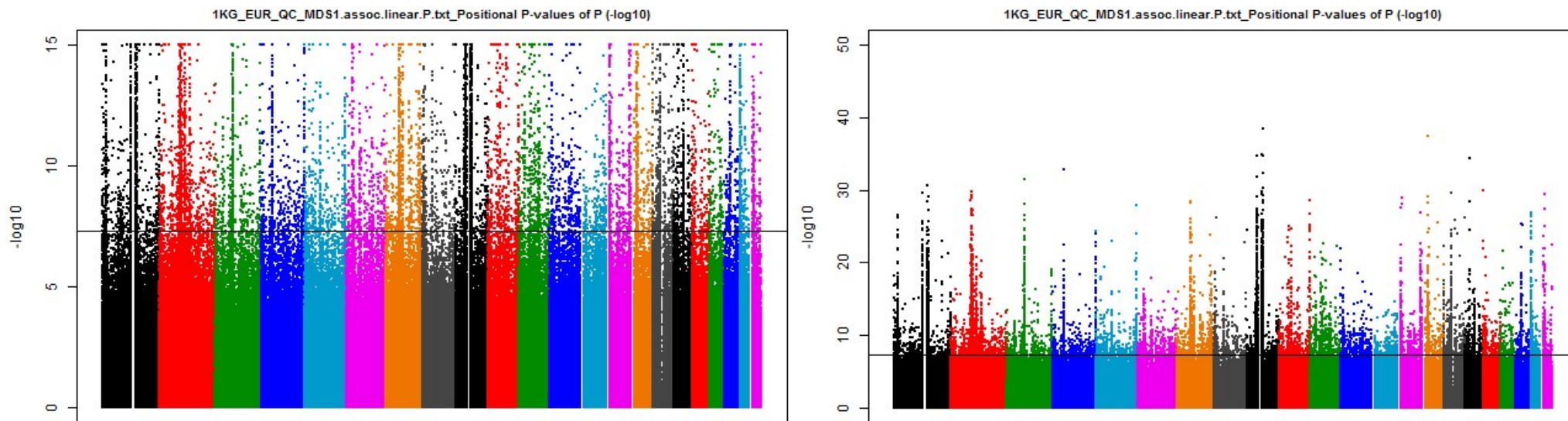
② PLINKを使ったゲノムワイド関連解析



- 形質情報に用いたMDS値は、元々は同じジェノタイプデータから作成されたものでした。
- ジェノタイプデータとその変換値という非独立な変数同士の独立性の検定を行ったため、帰無仮説が成立していなかったことになります。
- そのため、P値の分布が帰無仮説下から大きく外れたわけです。

② PLINKを使ったゲノムワイド関連解析

Y軸スケール変換



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

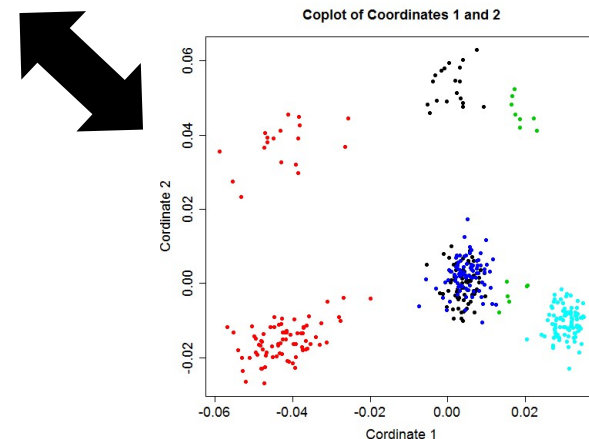
・マンハッタンプロットを書いてみると

①:ゲノム全体における統計量のインフレーション

②:特定の遺伝子領域における関連

の二つが認められることがわかりました。

・後者は、各集団間の遺伝的背景の差を反映する領域と考えられます。

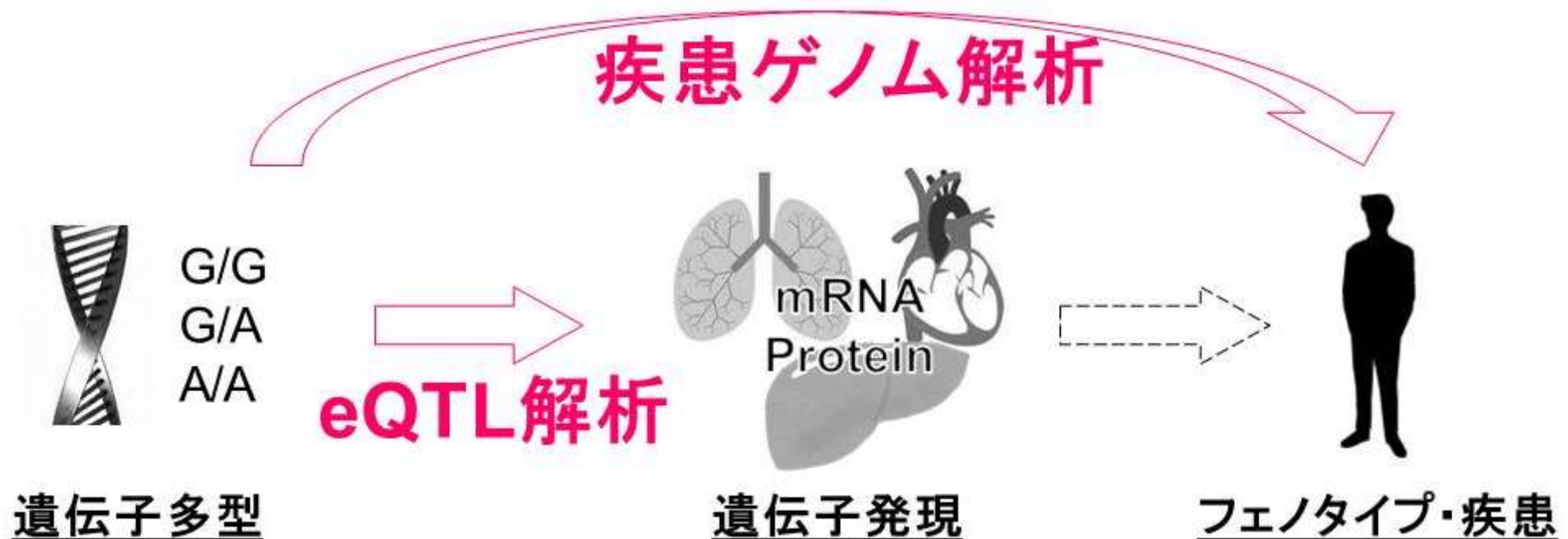


GenomeDataAnalysis2

- ① 遺伝統計学における関連解析
- ② PLINKを使ったゲノムワイド関連解析
- ③ 遺伝子発現量を対象としたeQTL解析

③ 遺伝子発現量を対象としたeQTL解析

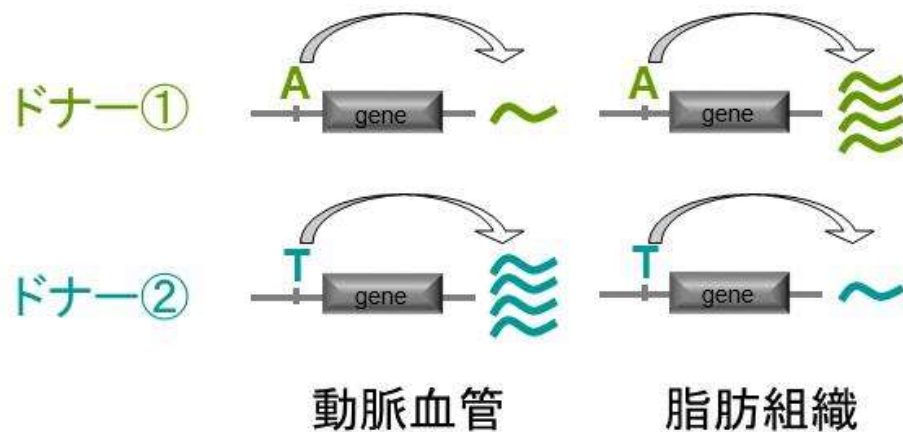
Expression Quantitative Trait Loci (eQTL)



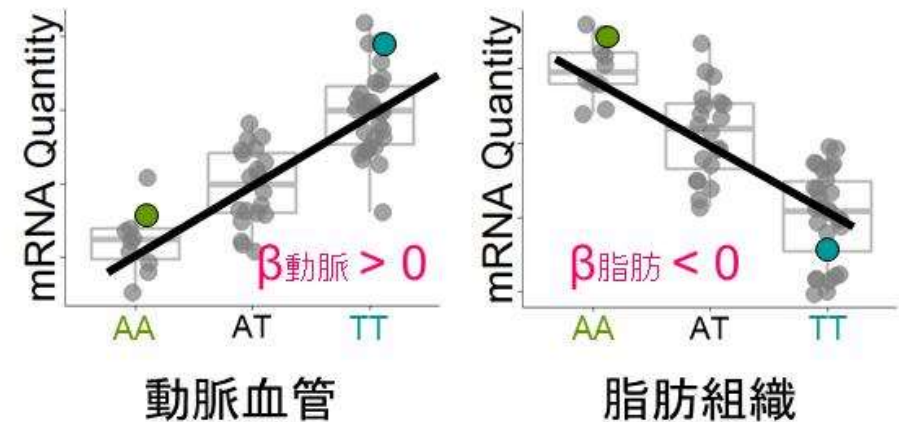
- 遺伝子多型が遺伝子発現に与える量的効果をeQTLといいます。
- 遺伝子発現量の変化は、細胞内・生体内の機能に影響を与える中間形質(endophenotype)であるため、eQTL解析を行うことで、遺伝子多型と疾患発症のつながりを検討できると考えられています。

③ 遺伝子発現量を対象としたeQTL解析

1. 遺伝子周辺SNPと遺伝子発現量を把握

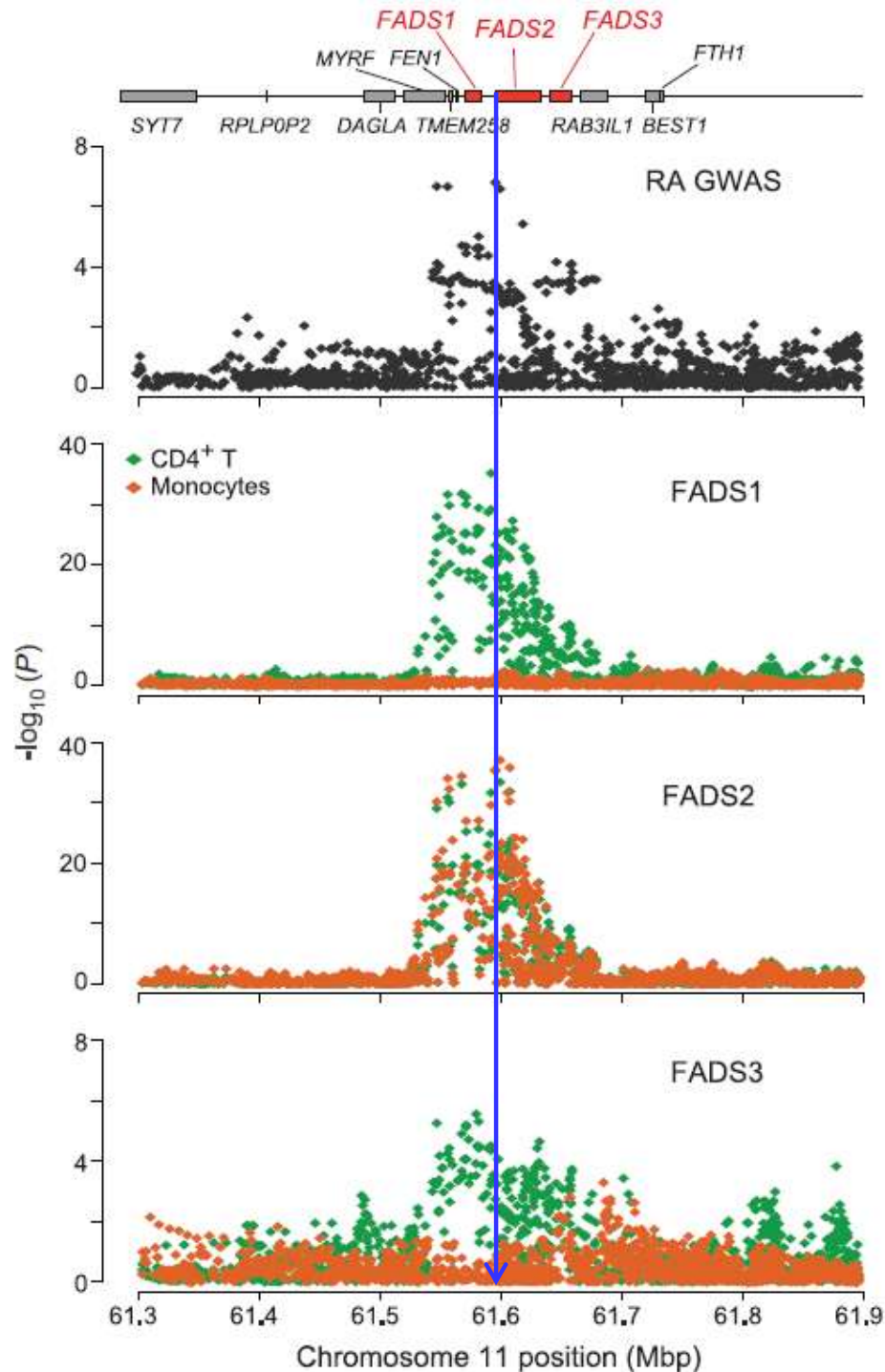


2. アレル別の発現量プロットを臓器間で比較



- 遺伝子発現パターンは組織によって異なるため、SNPのeQTL効果が特定の組織のみで認められたり、組織によって効果の方向性が異なる例が存在します。
- **組織特異的eQTL解析**の実施により、**遺伝子多型と各組織のつながり**が明らかになります。

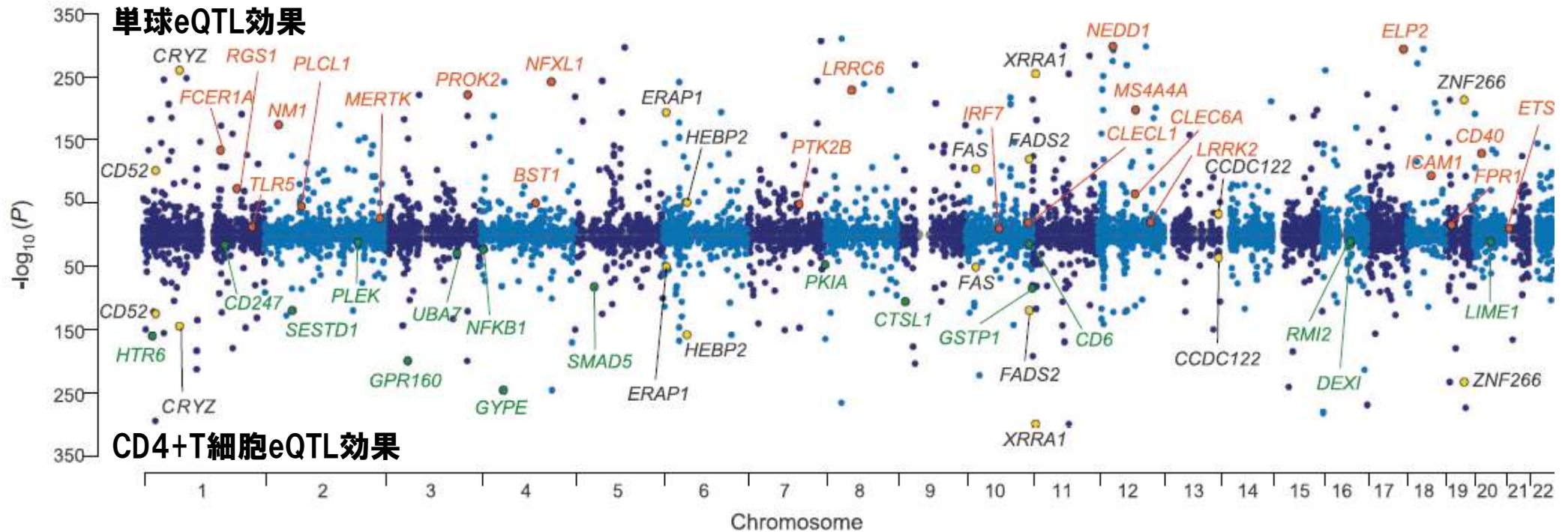
③ 遺伝子発現量を対象としたeQTL解析



- eQTL効果を有するSNPは、GWASで同定され、疾患感受性リスクを有するSNPでもある例が多いです。
- eQTL解析を実施することで、疾患リスクアレルがどのような機序で疾患発症を引き起こすか、の鍵を知ることができます。

③ 遺伝子発現量を対象としたeQTL解析

ゲノムワイドなeQTL解析(単球およびCD4+T細胞)

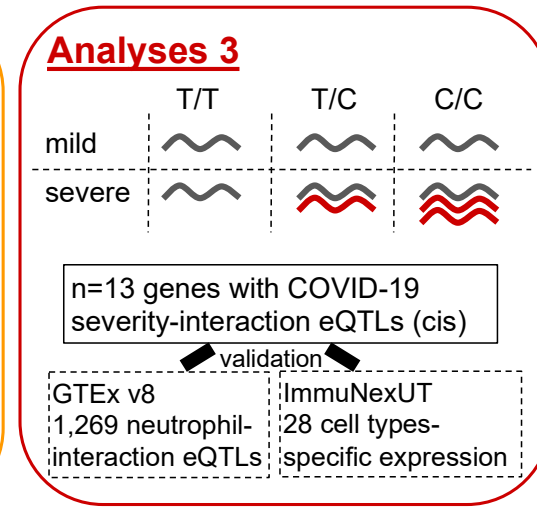
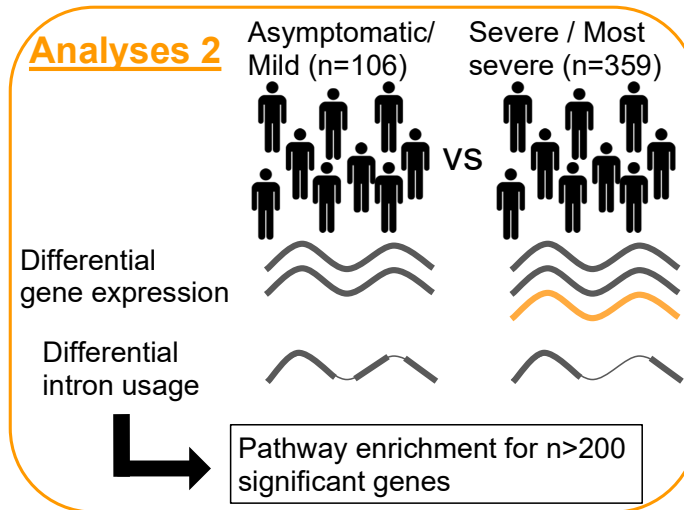
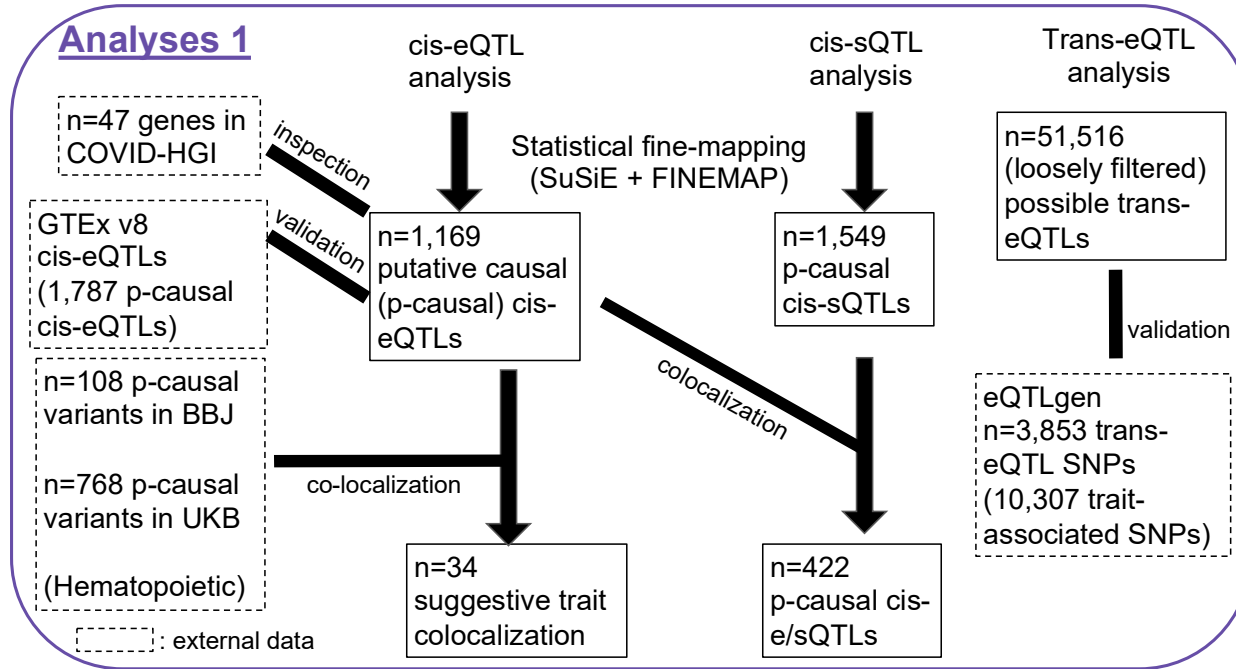


(Raj T et al. *Science* 2014)

- ゲノムワイドの全SNPと全遺伝子の発現量の関連を網羅的に解析するeQTL解析を実施し、疾患GWASの結果と照合することで、遺伝子変異、遺伝子発現量、疾患リスクの関係が明らかになると期待されます。³⁹

③ 遺伝子発現量を対象としたeQTL解析

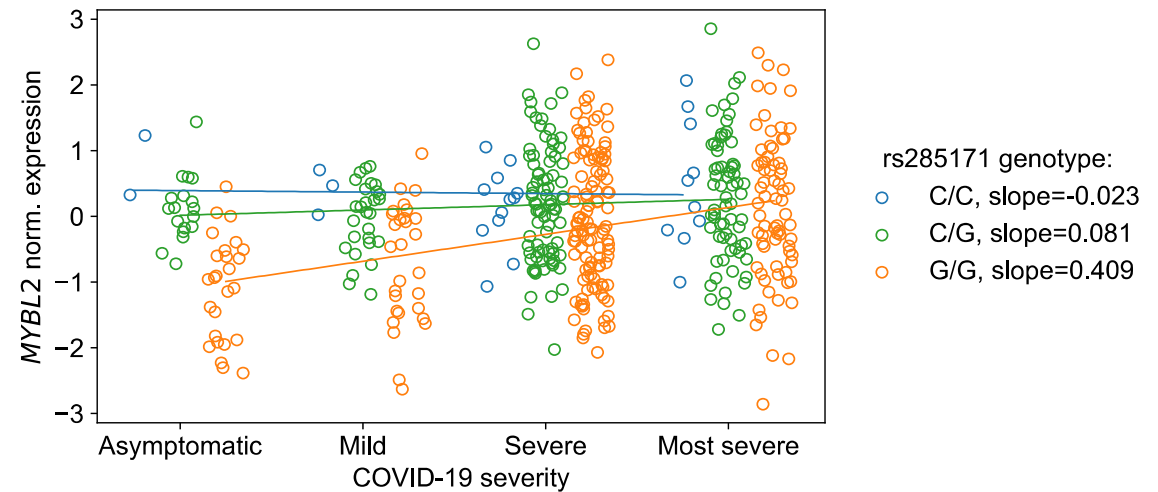
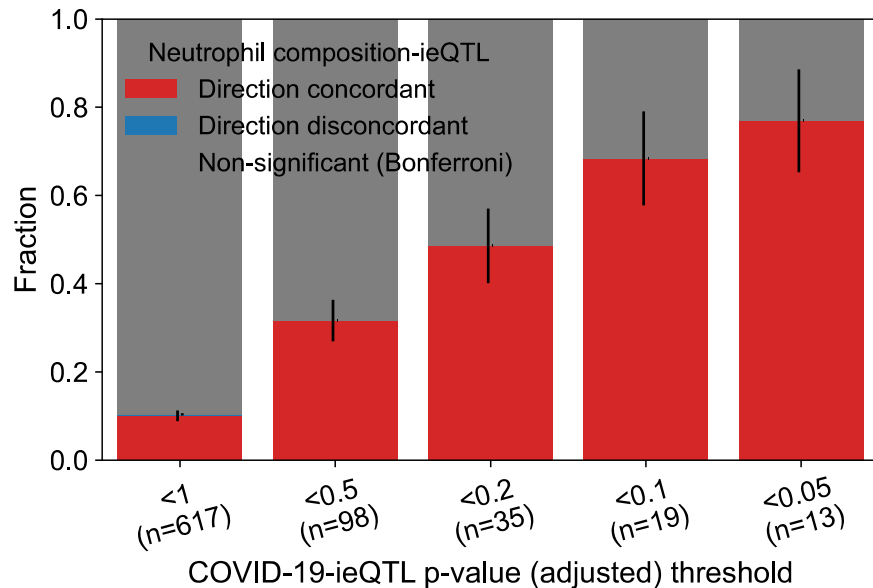
COVID-19:末梢血RNA-seq eQTL解析



• 日本人集団COVID-19感染者PBMCのRNA-seq eQTL解析(n=465)。

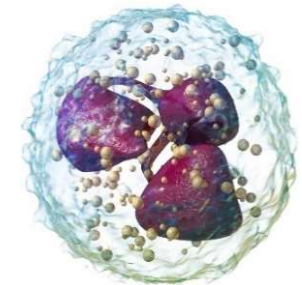
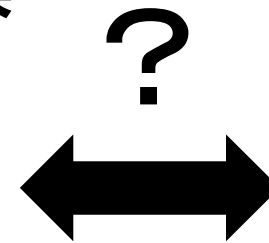
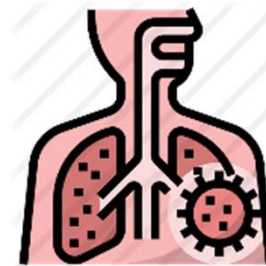
③ 遺伝子発現量を対象としたeQTL解析

COVID-19:末梢血RNA-seq eQTL解析



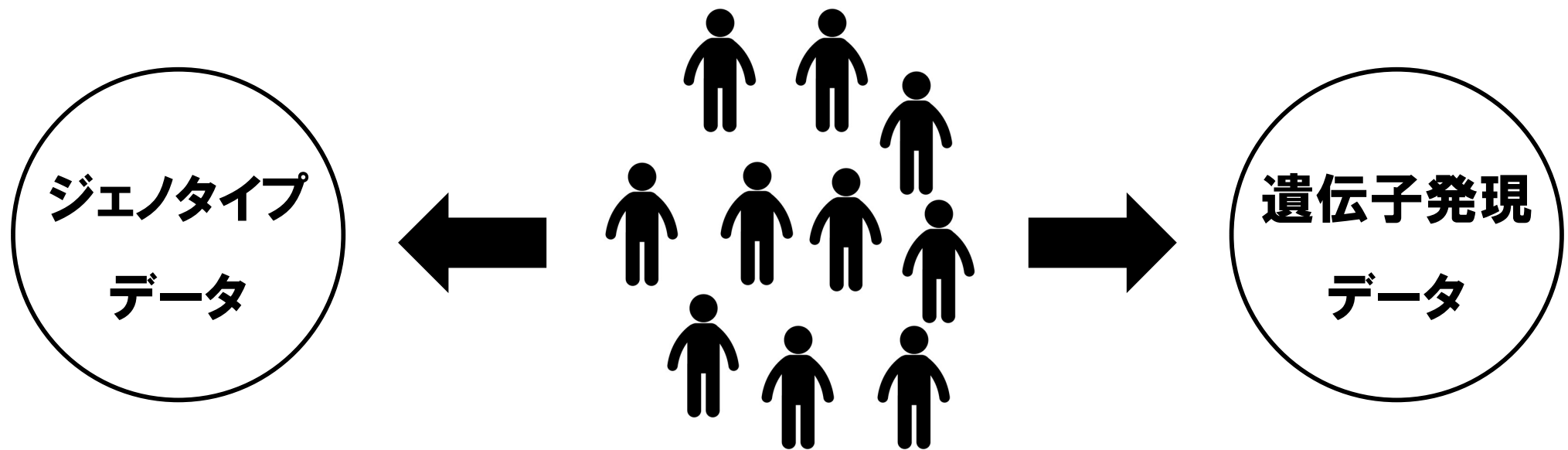
COVID-19重症度
依存的eQTL効果

好中球特異的
eQTL効果



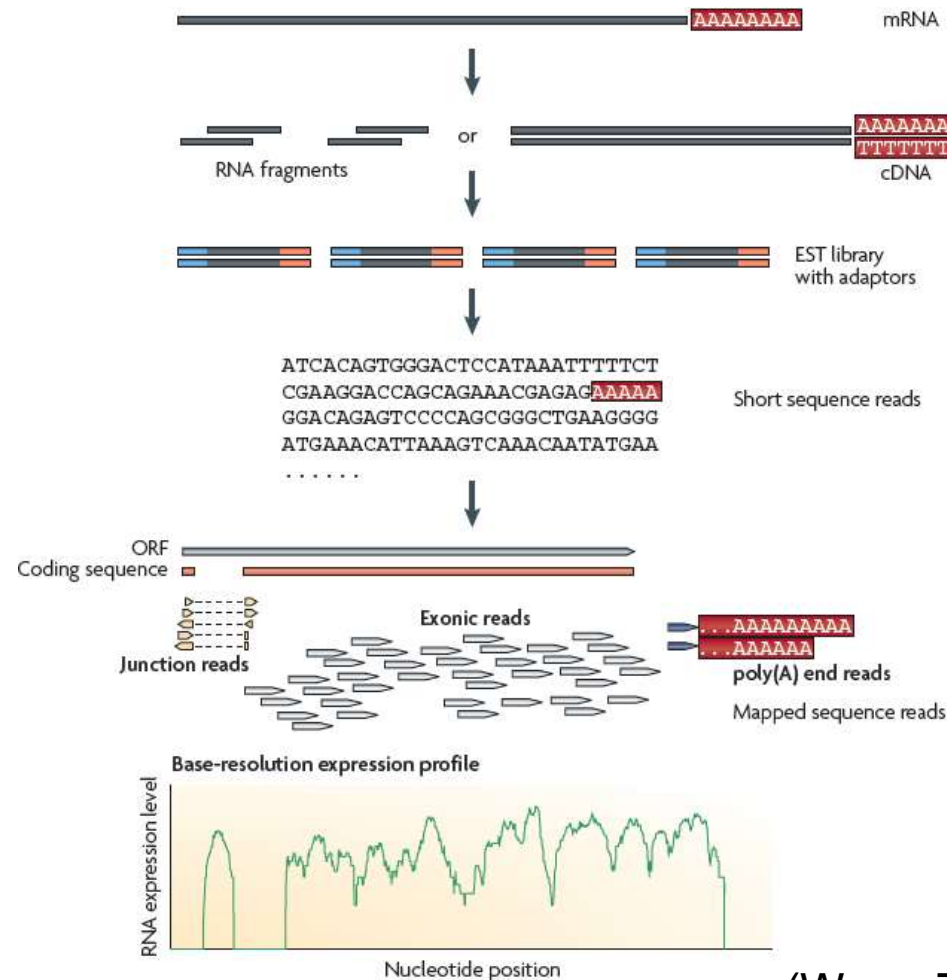
• COVID-19重症度に依存したeQTL効果(interaction eQTL; ieQTL)が存在し、好中球特異的eQTLとの重複が認められ、COVID-19重症化における好中球活性化の関与が示唆されました。

③ 遺伝子発現量を対象としたeQTL解析



- eQTL解析を実施するためには、**同一サンプル由来のヒトゲノムデータと遺伝子発現データ**の、両方が必要になります。
- 両方ともデータ取得にコストがかかるため、通常の疾患ゲノムデータと比較して、eQTL解析データへのアクセスは限定されがちです。

③ 遺伝子発現量を対象としたeQTL解析



(Wang Z et al. *Nat Rev Genet* 2009)

- gEUVADISプロジェクトの遺伝子発現データは、NGSを用いて発現定量を行うRNAシーケンズ(RNA-seq)により取得されています。
- マイクロアレイと比較して、安価・測定レンジが広い・スプライシングパターンを把握可能などの利点があり、近年急速に普及しています。⁴⁴

③ 遺伝子発現量を対象としたeQTL解析

Legend

Data access schema for Geuvadis RNAseq data. The main accession site to the data created and analyzed by the Geuvadis RNA-sequencing project is EBI ArrayExpress, where the data is stored under three accessions: E-GEUV-1 for mRNA post-QC samples used in analyses of this paper, E-GEUV-2 for small RNA post-QC samples, and E-GEUV-3 for all the sequenced data.

1) Raw reads in the form of fastq files are stored in ENA under the accession ERP001942 and ERP001941, but they are accessible also through ArrayExpress (the ENA and FASTQ columns)

[mRNA QC+](#) [miRNA QC+](#) [All QC+/-](#)

2) mRNA mapped reads are stored and accessible from EBI ArrayExpress. Files of mapped small RNA reads are not provided due to the more complex nature of mapping to different references for different analytical purposes and the large number of multimapping reads making file sizes very large.

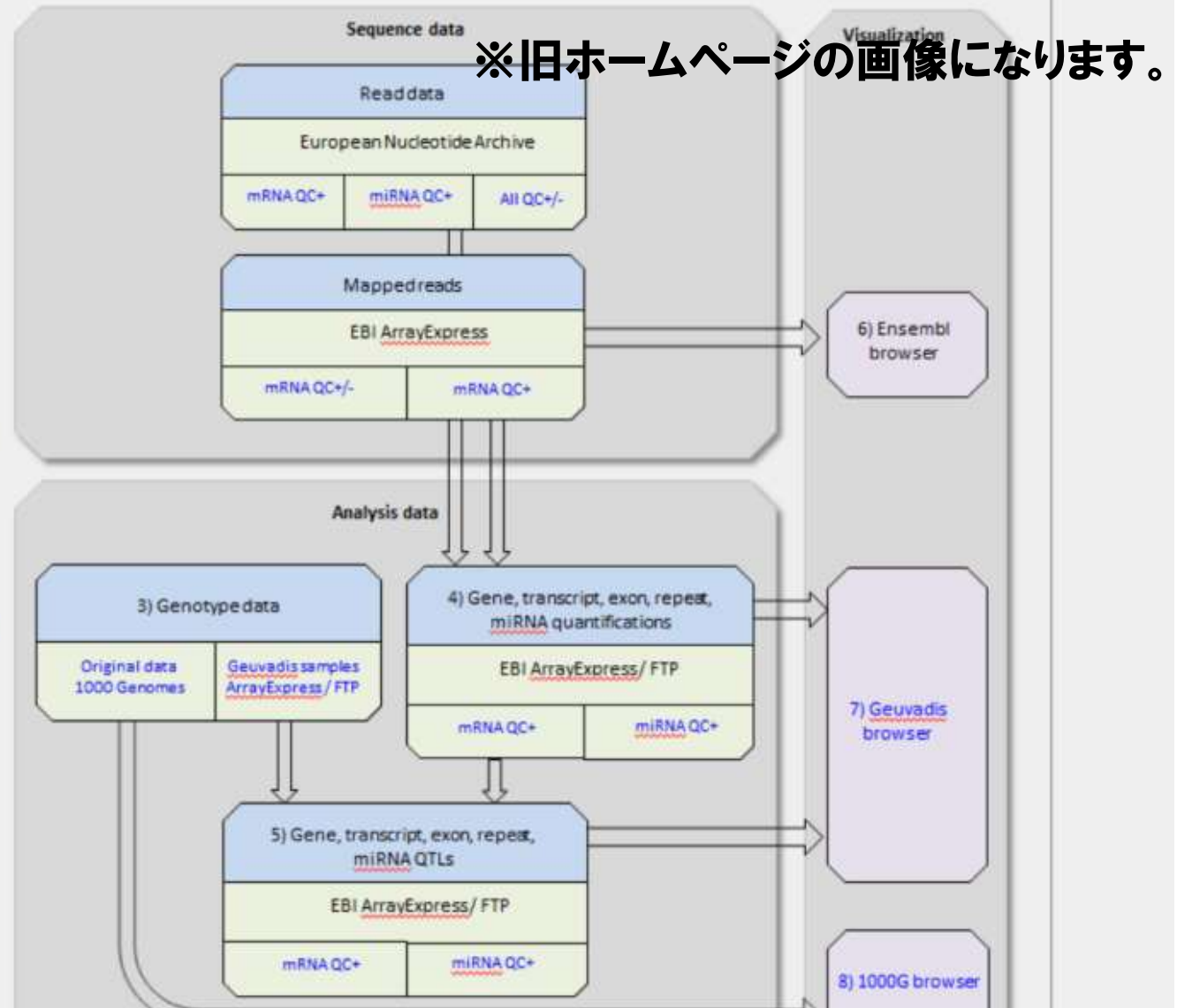
[mRNA QC +/-](#) [mRNA QC+](#)

3) Genotype data that have been used in [Geuvadis data analysis](#) are available from EBI ArrayExpress site under accession E-GEUV-1, and the vcf files include also a functional reannotation of all the variants. The [original data](#) created by 1000 Genomes Project are available in the 1000 Genomes web site.

4 and 5) [Geuvadis analysis results](#) for gene, transcript, exon, and repeat quantifications and QTLs will be available from EBI ArrayExpress site under accession E-GEUV-1, and miRNA quantifications and mirQTLs under accession E-GEUV-2.

6) mRNA mapping results per sample down to the level of individual reads can be visualized using Ensembl Genome Browser using the links from ArrayExpress (the Ensembl icon)

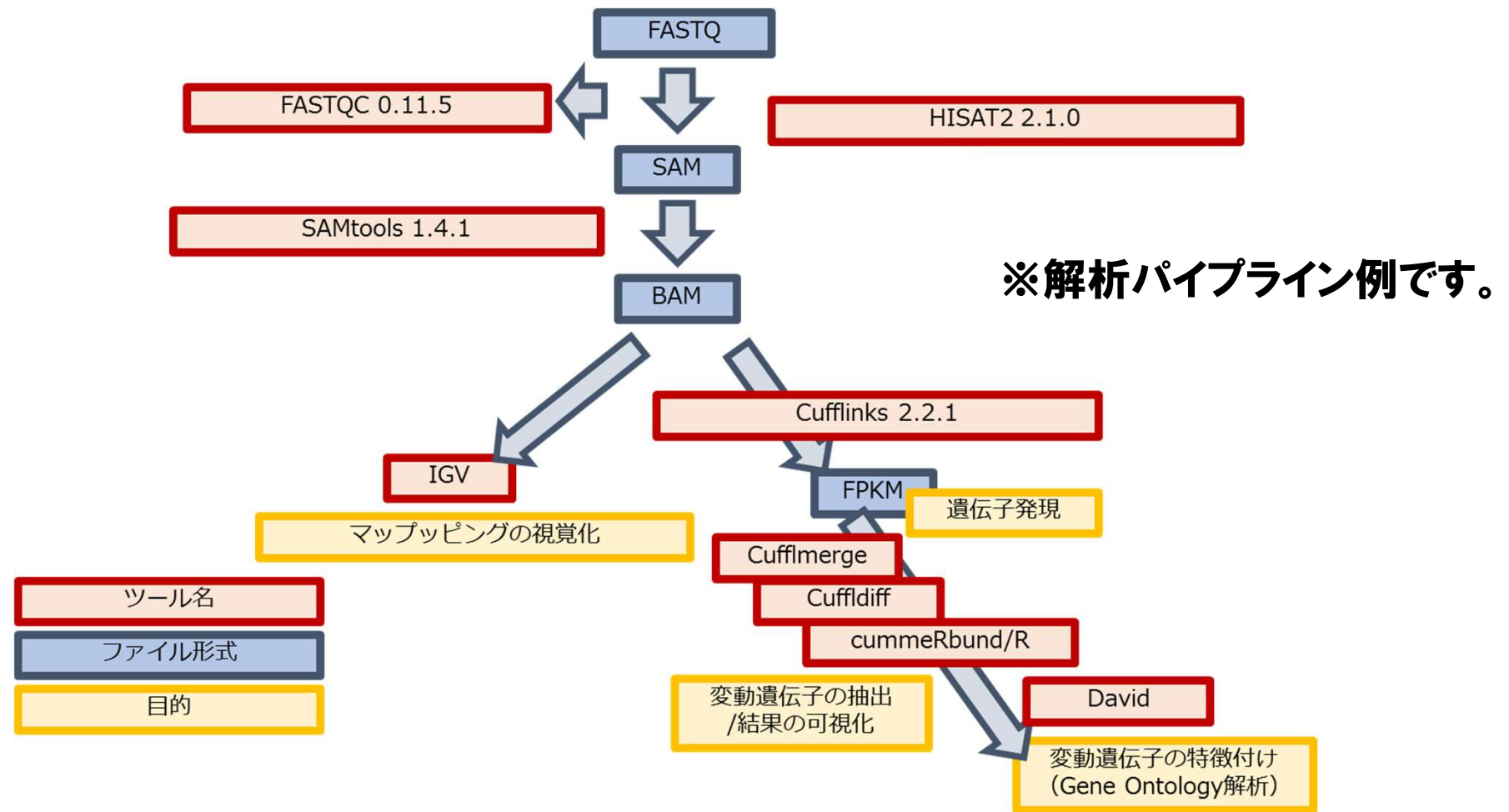
7) [Geuvadis data browser](#) was created specially for the Geuvadis RNA-seq project to visualize quantification and QTL results, and allows searching by variant ID, gene and



<https://www.internationalgenome.org/data-portal/data-collection/geuvadis>

- gEUVADISプロジェクトによる遺伝子発現データは、複数の解析ステップにおけるデータとして、公開されています。

③ 遺伝子発現量を対象としたeQTL解析



- NGSで得られたRNA-seqの結果ファイル(リード情報)から、遺伝子発現量を定量化するには、複数の手順にまたがるデータ解析が必要です。
- 手順の複雑さや計算時間を考慮し、本講義では説明しません。
- 興味を持った方は、各種Webサイトやハウツー本で勉強してください。

③ 遺伝子発現量を対象としたeQTL解析

線形回帰分析

$$\begin{array}{c} \text{遺伝子発現量} \\ \left(\begin{array}{c} 2.55 \\ 3.30 \\ 1.04 \\ 1.90 \\ \vdots \\ 2.70 \\ 4.90 \\ 1.61 \\ 3.21 \end{array} \right) \sim \begin{array}{c} \text{ジェノ} \\ \text{タイプ} \\ \left(\begin{array}{c} 0 \\ 1 \\ 2 \\ 0 \\ \vdots \\ 2 \\ 0 \\ 1 \\ 2 \end{array} \right) + \begin{array}{c} \text{年齢} \\ \left(\begin{array}{c} 46 \\ 23 \\ 64 \\ 78 \\ \vdots \\ 72 \\ 39 \\ 24 \\ 19 \end{array} \right) + \begin{array}{c} \text{性別} \\ \left(\begin{array}{c} 1 \\ 1 \\ 2 \\ 2 \\ \vdots \\ 1 \\ 2 \\ 1 \\ 2 \end{array} \right) \end{array}$$

Exp_BLK.txt

```
C:\SummerSchool\GenomeDataAnalysis...
ファイル(E) 編集(E) 表示(V) 検索(S) ウィンドウ(W) マクロ(M)
その他(O) 1: 1
1 | HG00096-HG00096-1796.033974
2 | HG00097-HG00097-2608.238722
3 | HG00099-HG00099-1533.474469
4 | HG00100-HG00100-1492.286304
5 | HG00101-HG00101-1530.550332
6 | HG00102-HG00102-850.5196705
7 | HG00103-HG00103-1812.941447
8 | HG00104-HG00104-1858.508984
9 | HG00105-HG00105-940.5380347
10 | HG00106-HG00106-1588.041064
11 | HG00108-HG00108-1988.548226
12 | HG00109-HG00109-1093.122425
13 | HG00110-HG00110-609.2191001
```

- eQTL解析の実施には、**線形回帰分析**を使います。
- 定量化された**遺伝子発現量**を**従属変数**としてモデルに組み込みます。
- 今回は、gEUVADISプロジェクトで得られた**BLK遺伝子**の**遺伝子発現量**を対象とします。

③ 遺伝子発現量を対象としたeQTL解析

○:ゲノムワイド関連解析の実施(線型回帰分析)

```
./plink --bfile 1KG_EUR_QC --out 1KG_EUR_QC_Exp_BLK --pheno Exp_BLK.txt  
--linear --ci 0.95
```

※Cygwinの場合plinkをplink.exeに変えてください



1KG_EUR_QC_Exp_BLK.assoc.linear

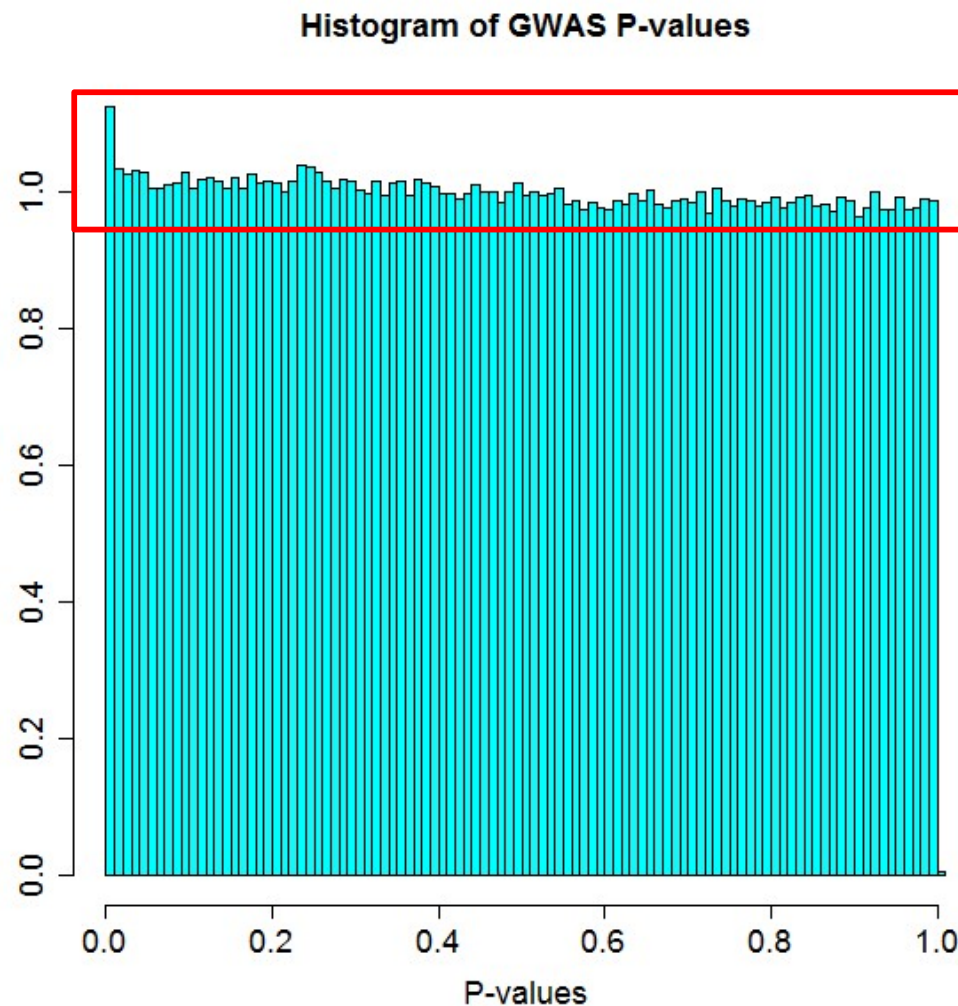
```
statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis2/1KG_EUR
```

```
$ awk '{print $2"¥t"($1*1000000000000+$3)"¥t"$12}'
```

```
1KG_EUR_QC_Exp_BLK.assoc.linear > 1KG_EUR_QC_Exp_BLK.assoc.linear.P.txt
```

• 前回と同じく、“--linear”を使ってGWASを実施してみましよう。

③ 遺伝子発現量を対象としたeQTL解析



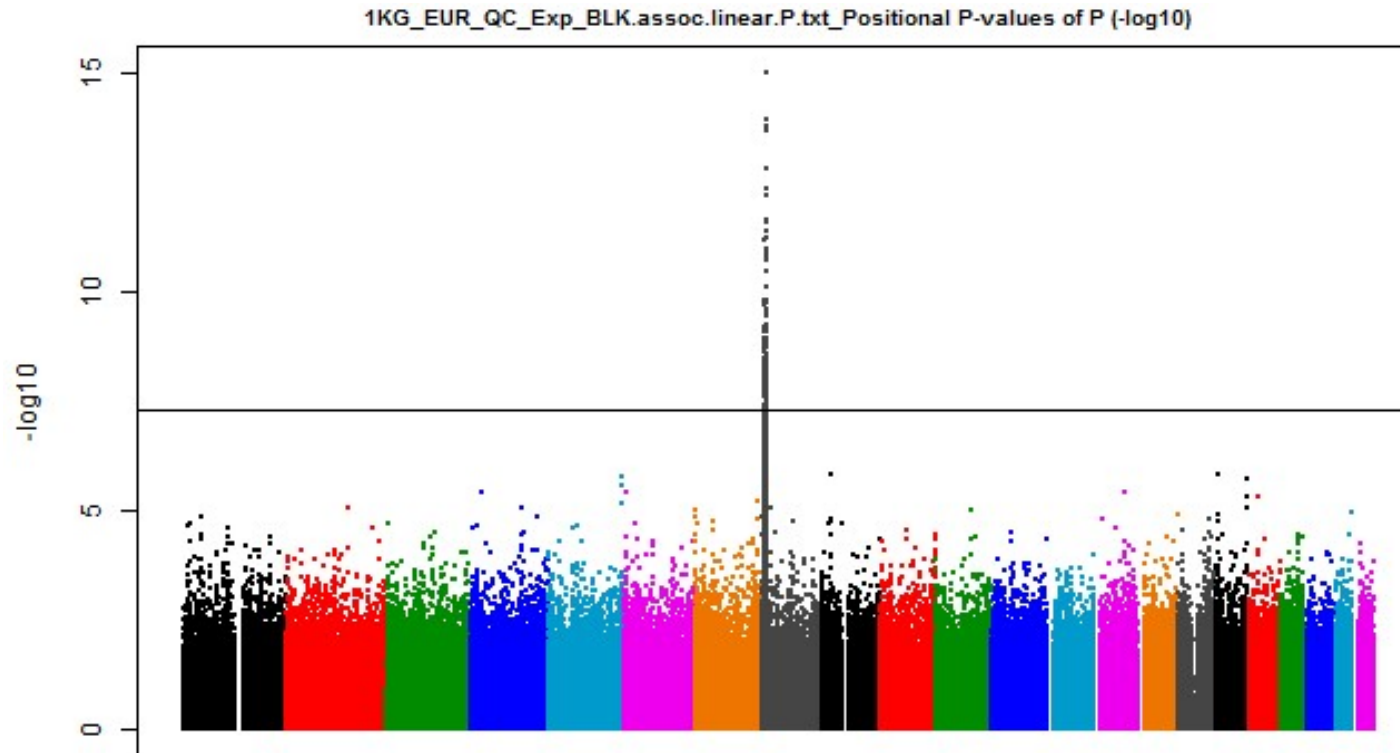
帰無仮説に従い
均一なP値の分布

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

- P値のヒストグラムを書いたところ、**おおむね一様分布に従うことが確認**できました。

③ 遺伝子発現量を対象としたeQTL解析

BLK遺伝子発現量に対するゲノムワイドeQTL解析結果



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

- マンハッタンプロットを描いてみると、一つの遺伝子領域に、ゲノムワイド水準を満たすeQTL効果が認められました。
- どの遺伝子領域のどのSNPか、調べてみましょう。

③ 遺伝子発現量を対象としたeQTL解析

```
statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis2/1KG_EUR
$ awk '$3<=10^-12 {print $0}' 1KG_EUR_QC_Exp_BLK.assoc.linear.P.txt
rs13255193      800011309192    4.539e-13
rs13257831      800011332964    6.545e-13
rs2736345       800011352485    1.707e-14
rs1478898       800011395079    6.882e-16
rs2244894       800011448659    1.497e-13
rs2244648       800011450422    2.068e-14
rs13273172      800011461111    1.188e-14
```

P<10⁻¹²を満たすSNPのみ抽出

最小P値SNP
rs1478898

- AWKコマンドを使って、P値に対する閾値でフィルターをかけることで、上位のeQTL効果を示したSNPを抜き出してみましよう。
- 最小P値を示したSNP: **rs1478898**について、Webツールで調べてみましよう。

③ 遺伝子発現量を対象としたeQTL解析

Current Build 155
Released April 9, 2021

rs1478898

Organism	<i>Homo sapiens</i>	Clinical Significance	Not Reported in ClinVar
Position	chr8:11537570 (GRCh38.p13) ?	Gene : Consequence	BLK : Intron Variant
Alleles	G>A	Publications	0 citations
Variation Type	SNV Single Nucleotide Variation	Genomic View	See rs on genome
Frequency	A=0.392554 (103905/264690, TOPMED) A=0.411244 (57570/139990, GnomAD) A=0.47083 (8894/18890, ALFA) (+ 12 more)		

Variant Details	Genomic Placements ?
Clinical Significance	
Frequency	
HGVS	
Submissions	
History	
Publications	

Sequence name	Change
BLK RefSeqGene	NG_023543.2:g.48559G>A
chr 8 fix patch HG76_PATCH	NW_018654717.1:g.1810526T>C
GRCh37.p13 chr 8	NC_000008.10:g.11395079G>A
GRCh38.p13 chr 8	NC_000008.11:g.11537570G>A

Gene: BLK, BLK proto-oncogene, Src family tyrosine kinase (plus strand)

<http://www.ncbi.nlm.nih.gov/SNP/>

- dbSNPで検索したところ、rs1478898はBLK遺伝子領域に存在するSNPであることがわかりました。

③ 遺伝子発現量を対象としたeQTL解析

Query SNP: **rs1478898** and variants with $r^2 \geq 0.8$

chr	pos (hg38)	LD (r ²)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	NHGRI/EBI GWAS hits	GRASP QTL hits	Selected eQTL hits	GENCODE genes	dbSNP func annot
8	11500647	0.8	0.92	rs55860742	T	C	0.20	0.39	0.00	0.53			BLD, THYM			LUN-1			73 hits	BLK	intronic
8	11534584	0.92	1	rs2248909	A	C	0.39	0.41	0.01	0.53		BLD	BLD, THYM			Egr-1,NF-E2,ZBTB7A			74 hits	BLK	intronic
8	11536255	0.97	1	rs2248699	A	G	0.31	0.40	0.01	0.51			BLD			4 altered motifs			69 hits	BLK	intronic
8	11537570	1	1	rs1478898	G	A	0.30	0.40	0.00	0.51		BLD	6 tissues	BLD, BLD	POL2	6 altered motifs			74 hits	BLK	intronic
8	11539347	0.88	-0.99	rs2409784	A	C	0.40	0.58	0.99	0.46		BLD	BLD, LNG	BLD, BLD	POL2, POL24H8, TBP	NRSF		12 hits	84 hits	BLK	intronic
8	11539365	0.97	0.99	rs2248325	A	G	0.31	0.40	0.00	0.51		BLD	BLD, LNG	BLD, BLD	POL2, TBP, POL24H8	Pax-4			75 hits	BLK	intronic
8	11539564	0.97	0.99	rs2248316	A	C	0.30	0.40	0.00	0.51		BLD	BLD, LNG			4 altered motifs			76 hits	BLK	intronic
8	11539577	0.97	0.99	rs2248315	T	A	0.33	0.40	0.00	0.51		BLD	BLD, LNG			4 altered motifs			76 hits	BLK	intronic
8	11539948	0.94	-0.99	rs2061830	C	G	0.64	0.59	1.00	0.48		BLD	BLD, LNG	LNG		Pou2f2, Pou3f3, Sox			83 hits	BLK	intronic
8	11541356	0.97	0.99	rs2618434	A	G	0.31	0.40	0.00	0.51			BLD	BLD		AIRE, RP58, TAL1			80 hits	BLK	intronic
8	11541444	0.95	0.97	rs2467520	T	C	0.31	0.40	0.00	0.51			BLD			Arid5b, E2A			74 hits	BLK	intronic
8	11541501	0.85	0.96	rs200483144	T	TC	0.40	0.41	0.02	0.53			BLD			4 altered motifs			70 hits	BLK	intronic
		0.83	0.92	rs78481210	T	C	0.39	0.41	0.01	0.50			BLD			HDAC2, Irf			21 hits	BLK	intronic
8	11541975	0.97	0.99	rs2245357	A	T	0.33	0.40	0.00	0.51			BLD			GR			73 hits	BLK	intronic
8	11543119	0.93	-0.99	rs12386974	C	G	0.61	0.59	1.00	0.48			BLD, VAS	KID		7 altered motifs			76 hits	BLK	intronic
8	11543171	0.97	0.99	rs2245250	G	A	0.32	0.40	0.00	0.51			BLD, VAS	IPSC					74 hits	BLK	intronic
8	11543435	0.97	0.99	rs2245232	G	T	0.30	0.41	0.02	0.51			ESDR, BLD, VAS			NERF1a			75 hits	BLK	intronic
8	11543607	0.96	0.99	rs11780851	G	A	0.29	0.40	0.00	0.51			4 tissues	BLD		SIX5, STAT, Znf143			71 hits	BLK	intronic
8	11544838	0.92	-0.99	rs6601599	G	A	0.61	0.59	1.00	0.48			BLD	BLD		6 altered motifs			72 hits	BLK	intronic

<https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php>

- **HaploReg**で検索したところ、rs1478898および連鎖不平衡関係(近傍のSNP間で非独立なアレル分布を有する状態)にあるSNPは、**エンハンサーとして機能するヒストン修飾**位置に存在していて、遺伝子発現制御への関与が示唆されました。

③ 遺伝子発現量を対象としたeQTL解析

```
statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis2/1KG_EUR
$ awk '$2>=800011395079-250000 && $2<=800011395079+250000 {print $0}'
1KG_EUR_QC_Exp_BLK.assoc.linear.P.txt | sort -k 2,2 -n >
1KG_EUR_QC_Exp_BLK.assoc.linear.P.BLK.txt
```

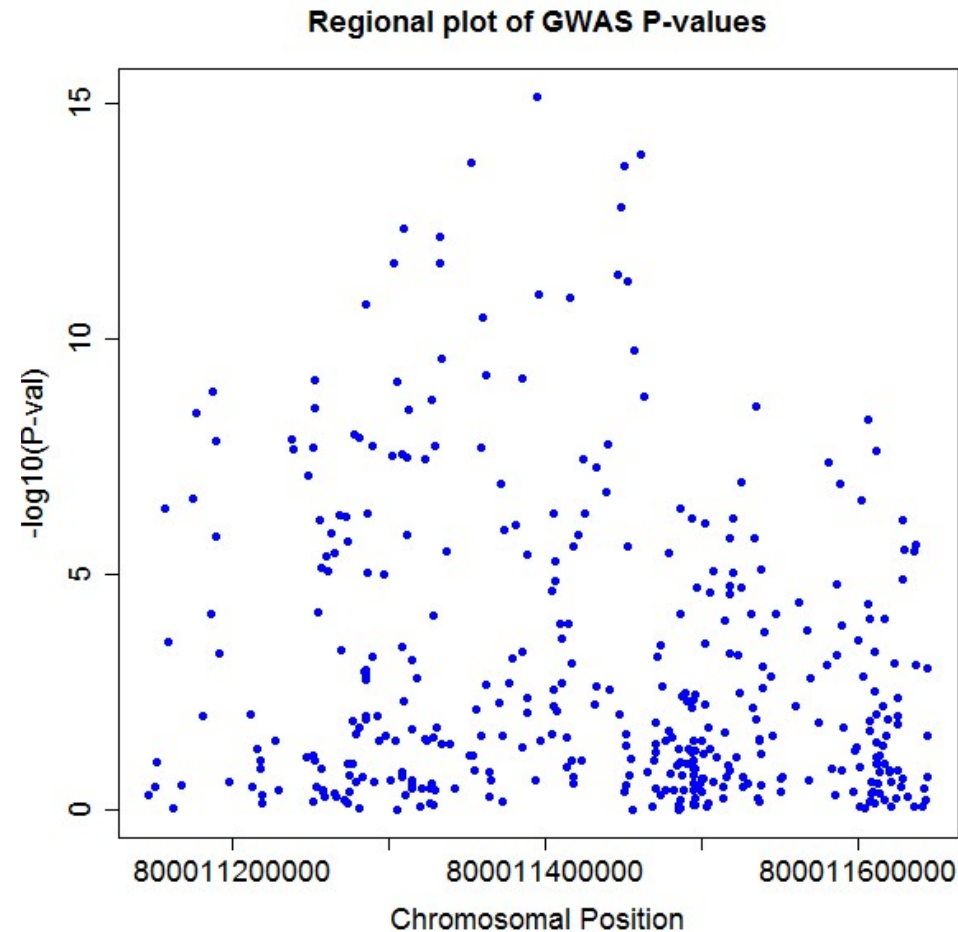
Linuxコマンド: `sort -k 2,2 -n`

数値順で

第2列目を

- rs1478898の周囲±250kbにおける、eQTL解析結果を抜き出してみよう。
- AWKコマンドで抜き出した後、Linuxコマンドsortで第2列を数値順でソートしておきます。

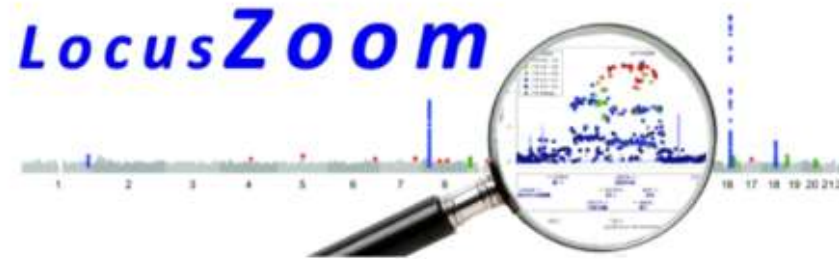
③ 遺伝子発現量を対象としたeQTL解析



※ファイル”RegionalPlot.R”を開いて、Rにコピー&ペーストして下さい。

- rs1478898の周囲±250kbのeQTL解析結果をプロットしてみました。
- 悪くないですが、**遺伝子の位置などの追加情報が欲しい**ところです⁵⁵。

③ 遺伝子発現量を対象としたeQTL解析



LocusZoom is a tool to plot regional association results from genome-wide association scans or candidate gene studies. This is Version 1.1

Report problems to cristen@umich.edu

We are pleased to announce that our paper on *LocusZoom* has been published. [[ABSTRACT](#) | [PDF](#)]

REFERENCE:

Pruim RJ*, Welch RP*, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. (2010) LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* 2010 September 15; 26(18): 2336.2337.

Count of Successful Plots

This week	952
Year 2016	59062
Jan	5936
Feb	8143
Mar	7638
Apr	8015
May	8127
Jun	9472
Jul	7881
Aug	3850
Year 2015	96212
Year 2014	97391
Year 2013	87012
Year 2012	69783
Year 2011	50775
Year 2010	22184

Links

- [Plot Using Your Data](#)
- [Plots Using Your Data and Your Hitspec File](#) Batch mode, results returned via Email

<http://locuszoom.sph.umich.edu/genform.php?type=yourdata>

- “LocusZoom”を使って、eQTL解析結果の、BLK遺伝子領域内SNP P値の図を描いてみましょう。

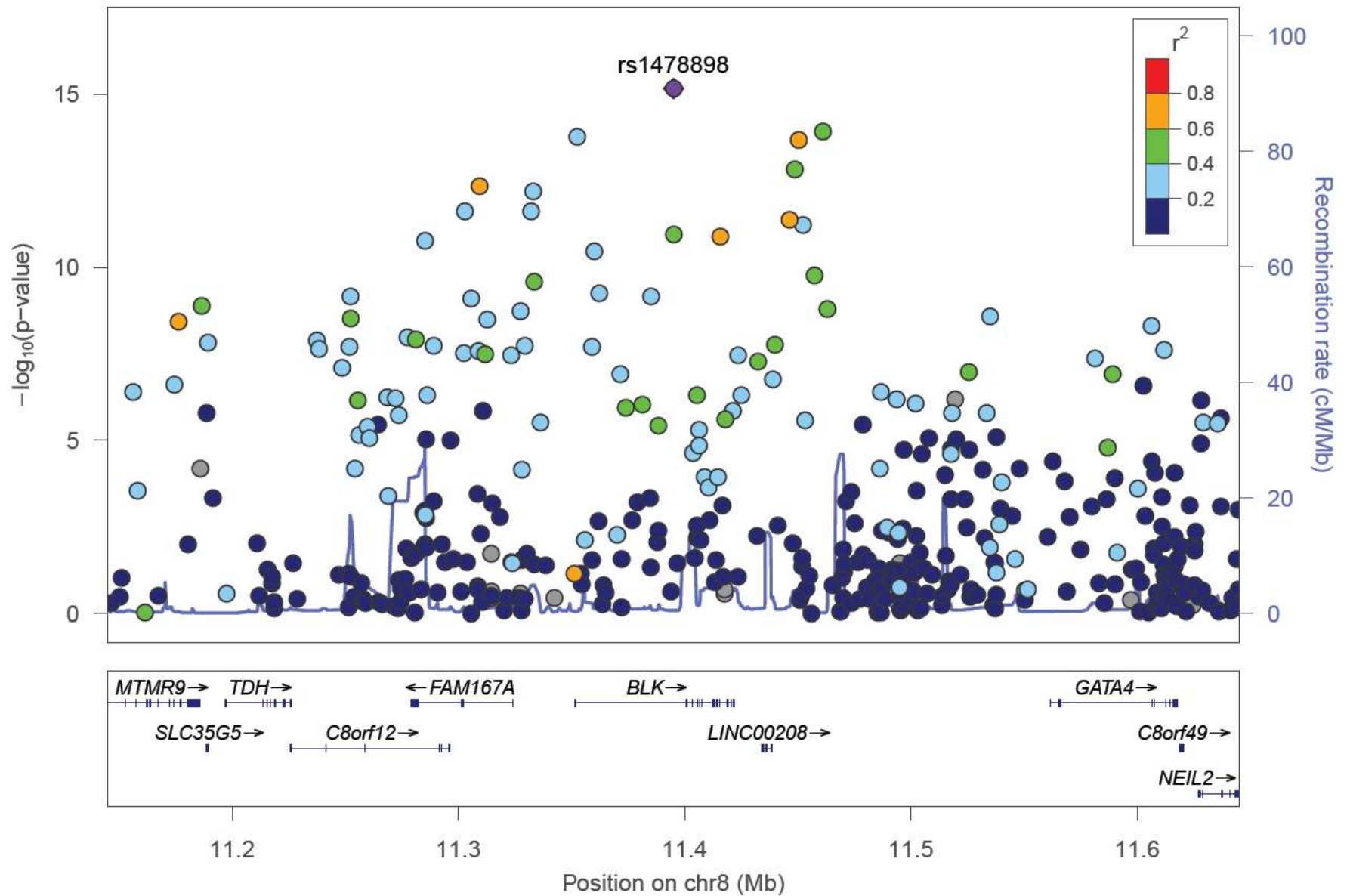
③ 遺伝子発現量を対象としたeQTL解析

Provide Details for Your Data	Path to Your File	<input type="text" value="ファイルを選択 1KG_EUR_Q...c.linear"/>	File will be sent to server and used for plotting (Maximum 2GB)			
	P-Value Column Name	<input type="text" value="P"/>	Set for <input type="text" value="PLINK data"/> or WikiGWA data			
	Marker Column Name	<input type="text" value="SNP"/>	Default is MarkerName			
	Column Delimiter	<input type="text" value="WhiteSpace"/>	Default is tab			
Specify Region to Display	SNP	<input type="text"/>	+/-	<input type="text" value="400"/> Kb		
	Gene	<input type="text" value="BLK"/>	+/-	<input type="text" value="200"/> Kb	<input type="text"/>	
	Region	Chr: <input type="text" value="None"/>	<input type="text"/>	Mb through	<input type="text"/>	Mb

Required: Fill in Only ONE of These Three

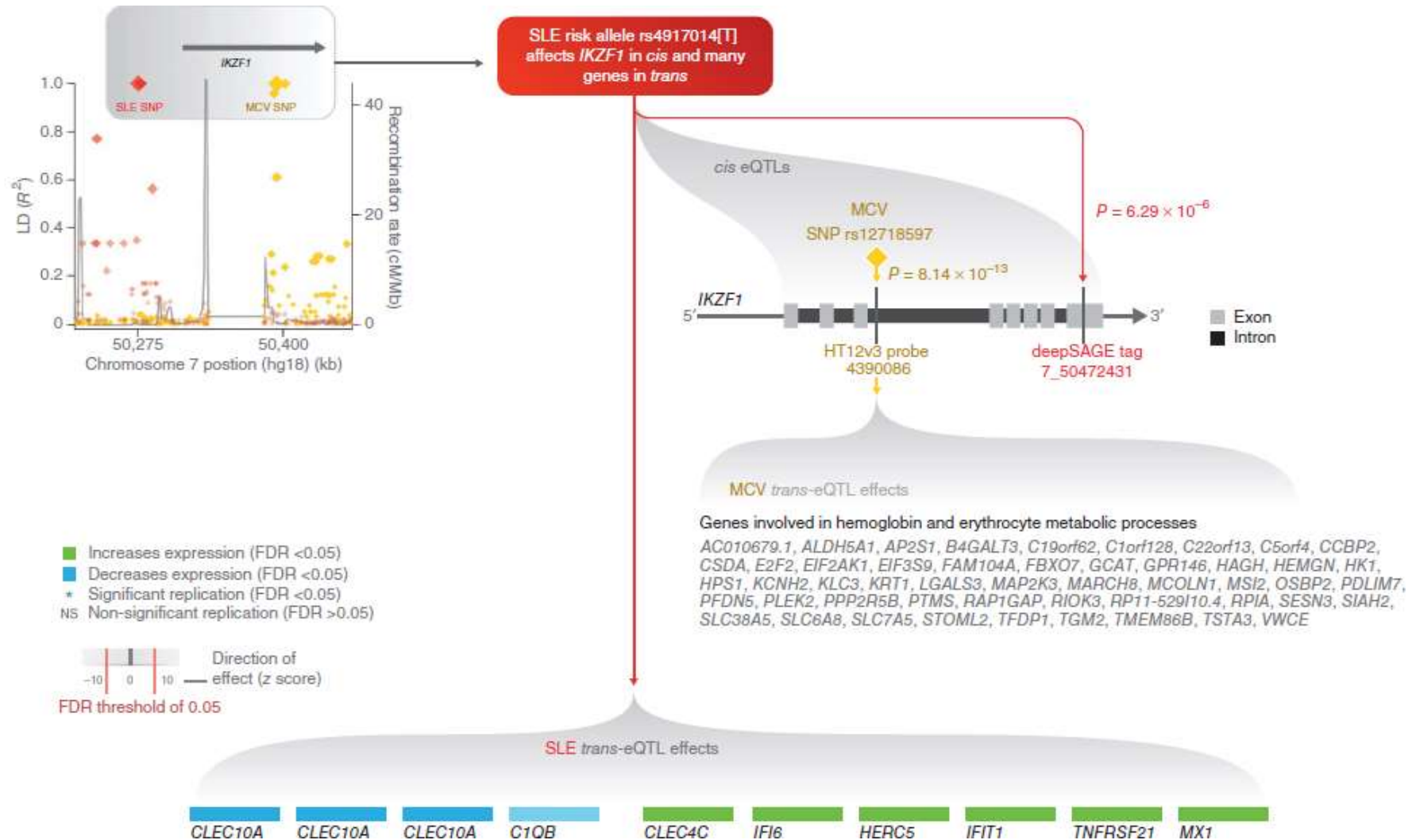
1KG_EUR_QC_Exp_BLK.assoc.linear

③ 遺伝子発現量を対象としたeQTL解析



- eQTL効果を示したSNPと周囲のSNPの連鎖不平衡関係や各遺伝子との位置関係が記載された図が描けました。

③ 遺伝子発現量を対象としたeQTL解析

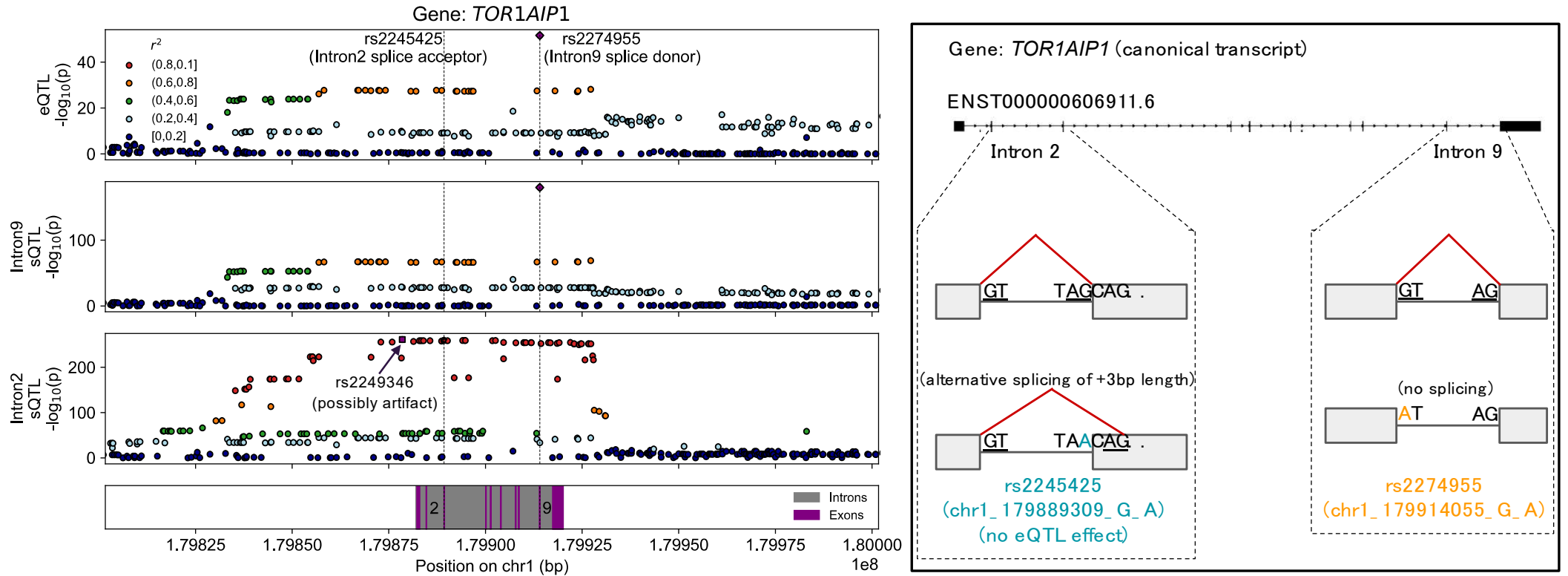


- **cis-eQTL**: 遺伝子変異が近傍の遺伝子の発現量に影響を与える現象。
- **trans-eQTL**: 異なる染色体上の遺伝子の発現量に影響を与える現象。
- 少数ですが、これまでの研究でtrans-eQTLの存在も報告されています。

(Westra HJ et al. *Nat Genet* 2014)

③ 遺伝子発現量を対象としたeQTL解析

TOR1AIP1遺伝子スプライス部位の変異によるsQTL効果



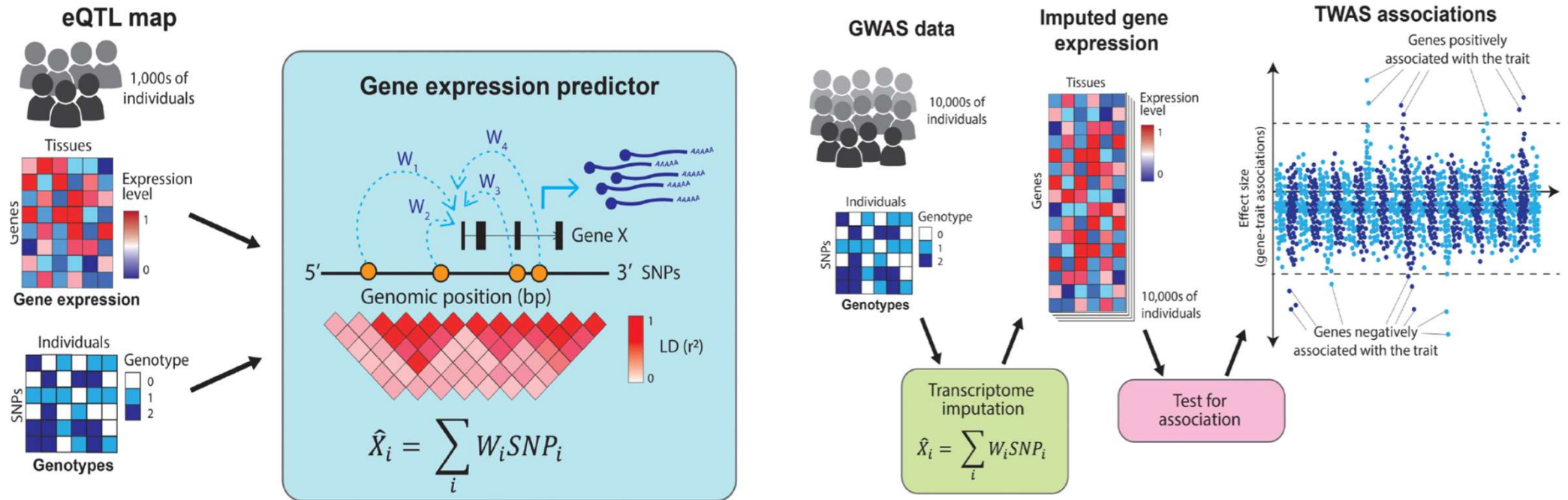
- **splicing QTL (sQTL)**: 遺伝子変異が**選択的スプライシング**に影響を与える現象。遺伝子配列のスプライス部位の変異に基づく例が多いです。
- **eQTL結果とsQTL結果を統合**することで、**機能性遺伝子変異の詳細な絞り込みに貢献**すると考えられています。

③ 遺伝子発現量を対象としたeQTL解析

TWAS: eQTL効果に基づく遺伝子発現予測モデル

eQTLに基づく遺伝子発現予測モデル

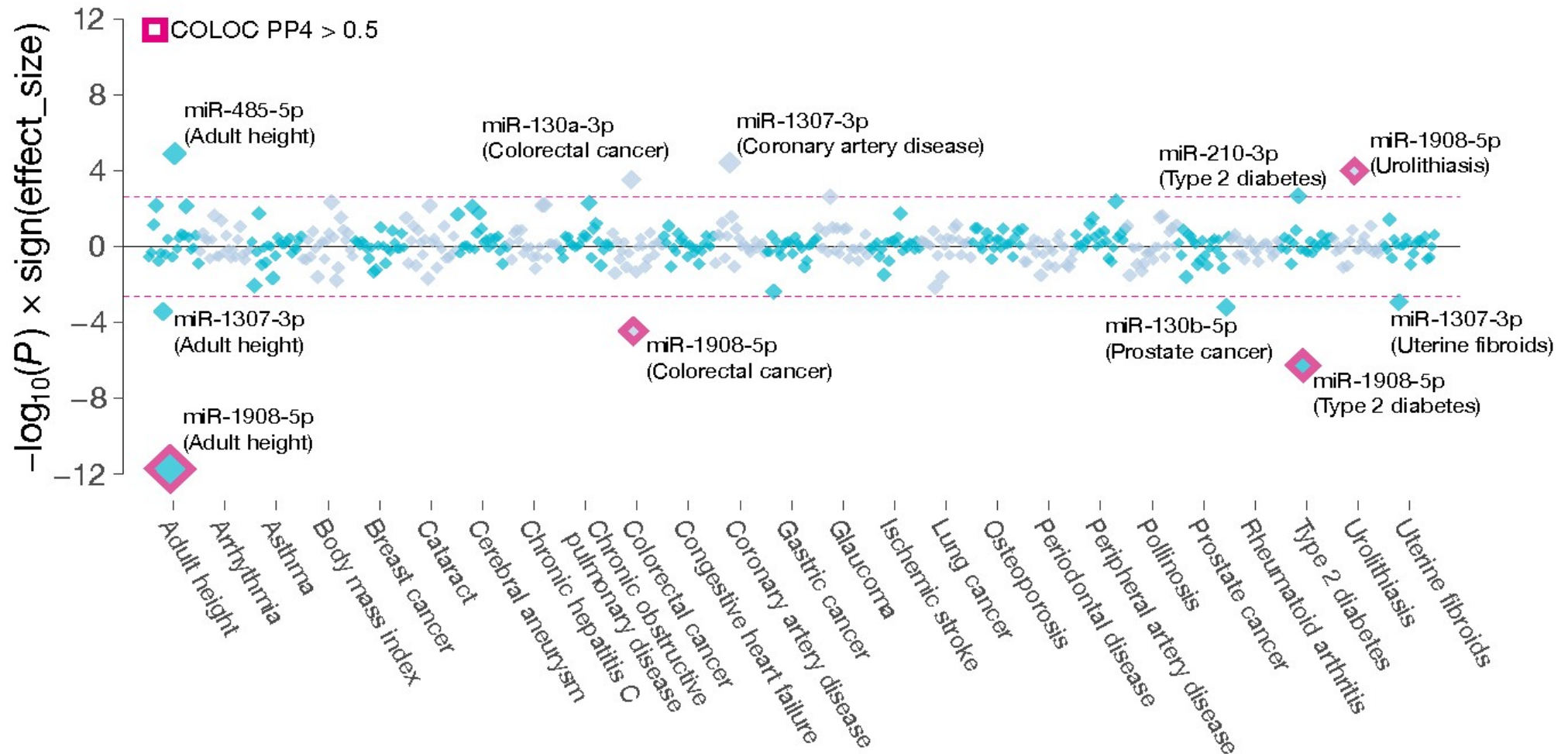
TWAS = GWAS + eQTL



- eQTL効果に基づき、個人の遺伝子変異情報から**特定の遺伝子の発現量を予測モデル**を構築することが可能。
- **TWAS (Transcriptome-Wide Association Study)**: 疾患GWAS情報とeQTL予測モデルを統合し、**ケースコントロール間の遺伝子発現量変化を予測**。

③ 遺伝子発現量を対象としたeQTL解析

日本人集団におけるマイクロRNA TWAS解析



- 日本人集団マイクロRNA eQTLデータベースと多彩な疾患のGWAS情報を統合する**マイクロRNA TWAS解析**を実施。
- 疾患発症予測バイオマーカーマイクロRNAを複数同定(例:miR-1908-5p:糖尿病・大腸癌・尿路結石・身長)。(Sonehara K et al. *Hum Mol Genet* 2022)

終わりに

- 1000 Genomes Projectのデータを使って、GWASを何パターンか実施してみました。
- PLINKに実装された機能を使うと、GWASの実施自体は簡単です。
- 一方で、統計量の全体的な分布に偏りがいないか、などGWASの結果を検証して、適切に解釈することが重要です。
- eQTL解析は、Webツールでの結果の公開が充実しています。
- 興味をもったSNPがあったら、機能的な意義について色々と調べてみて下さい。

- MDS結果で得られた第2～第4座標を形質としてGWASを実施し、結果の検討や解釈を行って下さい。

./1KG_EUR/test5.mds

- 公開された疾患GWASの結果をダウンロードして、マンハッタンプロットや各領域の作図、リスクSNPの機能的意義を検討して下さい。

./1KG_EUR/GWASpval_GIANT/GIANT_HEIGHT_2014_B37.txt.gz

./1KG_EUR/GWASpval_GIANT/GIANT_BMIA_2015_B37.txt.gz

./1KG_EUR/GWASpval_GIANT/GIANT_BMIE_2015_B37.txt.gz