2023年8月25-27日
遺伝統計学・夏の学校　講義実習資料

GenomeDataAnalysis6

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

GenomeDataAnalysis6

① GWAS統計量を用いた解析手法とLDSC

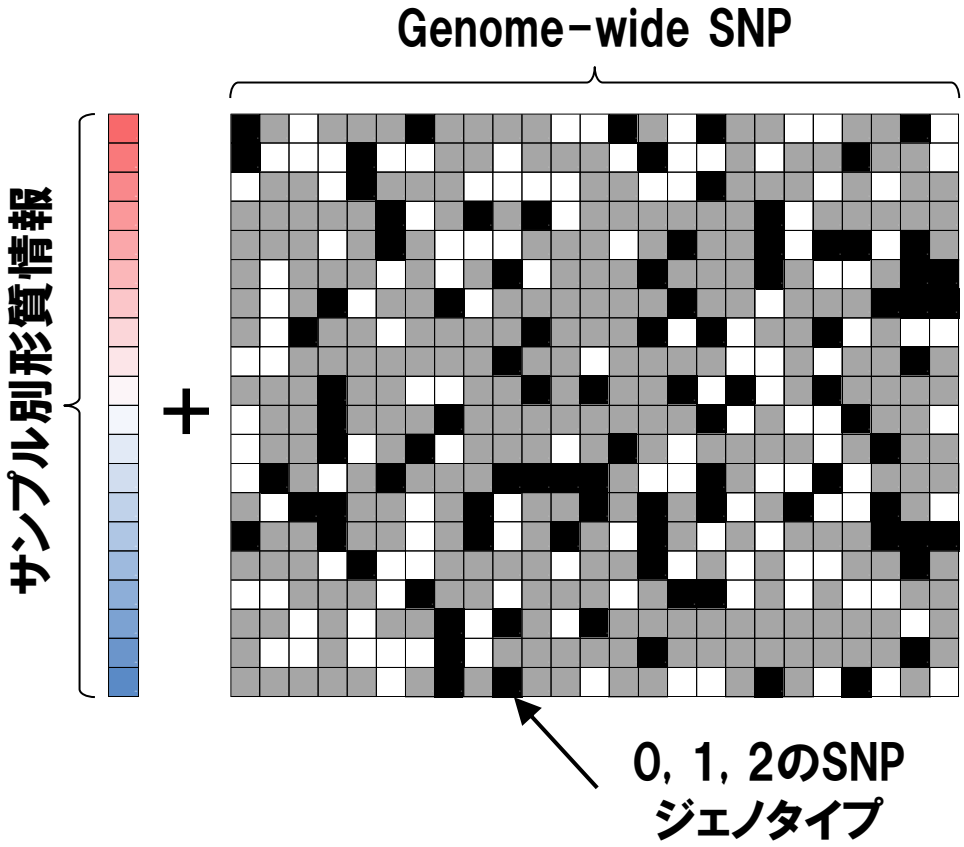
② Anacondaを使ったLDSCのインストール

③ LDSC解析実習

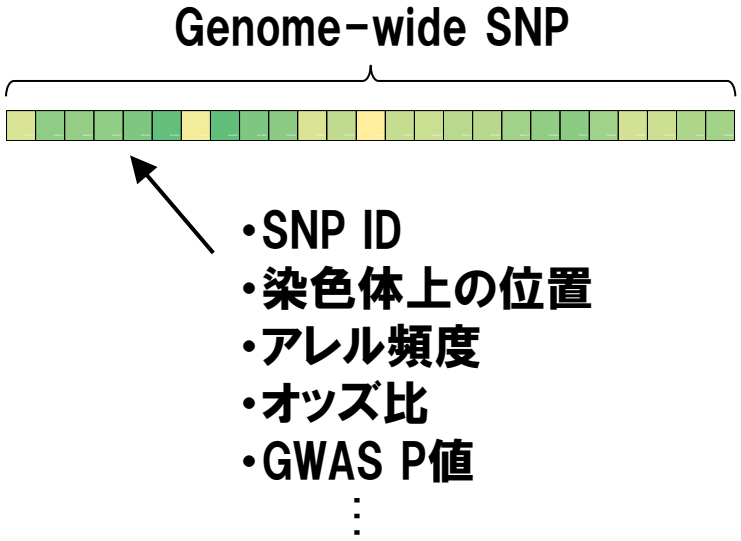
本講義資料は、Windows PC上で
C:¥SummerSchoolにフォルダを配置すること
を想定しています。

① GWAS統計量を用いた解析手法とLDSC

個人別形質・ゲノム情報

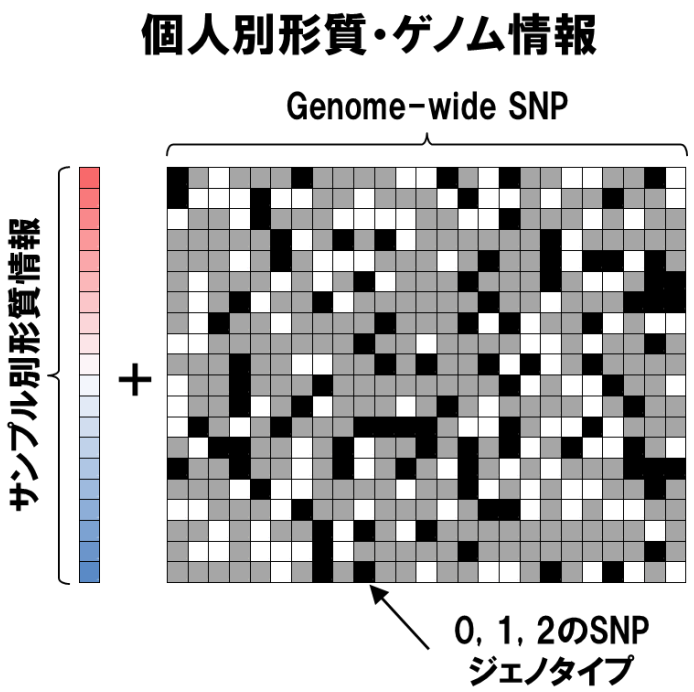


GWAS統計量



- GWASゲノムデータ解析に際して使用する入力データは、大きく分けて2種類存在し、**個人別の形質情報とゲノム情報**(GWAS individual data)と、**GWAS統計量**(GWAS summary statistics)になります。

① GWAS統計量を用いた解析手法とLDSC



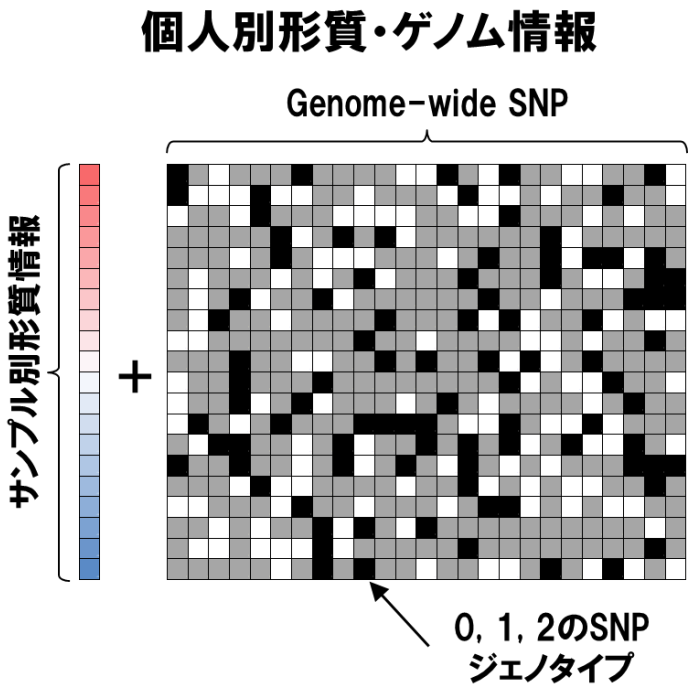
実施可能な解析手法

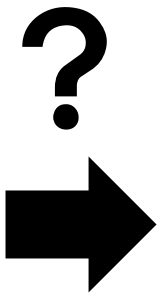
- GWAS
- eQTL解析
- Genotype imputation
- 選択圧解析
- Polygenic risk score推定
- Heritability推定
- Conditional解析
- Haplotype解析

⋮

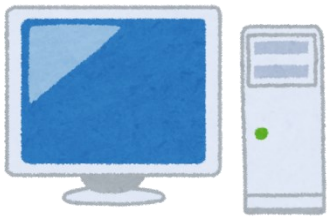
- 個人別の形質・ゲノム情報には、GWASゲノムデータ解析に必要な情報が(ほぼ)全て含まれています。
- GWASそのものも含め、多彩なゲノムデータ解析が実施可能になります。
- GenomeDataAnalysis１～５で実施した演習内容は、いずれも個人別の形質・ゲノム情報を用いた解析手法が対象でした。

① GWAS統計量を用いた解析手法とLDSC





データ共有・公開
の進みにくさ



大規模データ
解析に要する
計算負荷の増大

- 一方、個人別形質・ゲノム情報は個人情報に相当するため、データの共有や公開が進みにくい、というデメリットがあります。
- 特に、複数のGWASを統合するGWASメタアナリシスでは、全てのGWASの個人別情報へのアクセスが、困難になります。
- 数十万人規模の個人別情報解析は、計算機資源の負荷も大きいです。

① GWAS統計量を用いた解析手法とLDSC

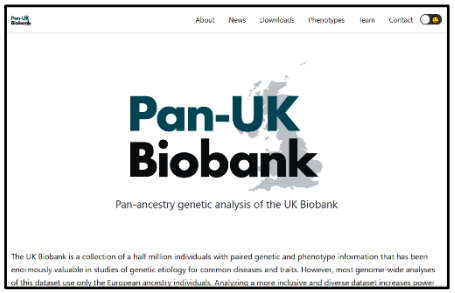
GWAS統計量

Genome-wide SNP

- SNP ID
- 場所
- アレル頻度
- オッズ比
- GWAS P値
- ⋮

データの公開・
入手が容易

公開データベース



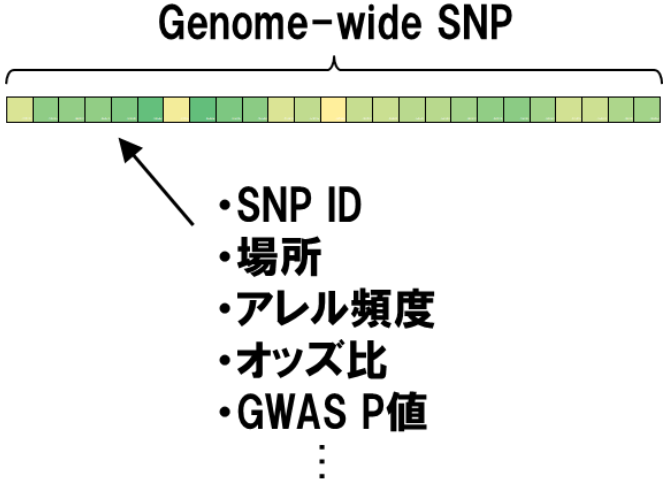
https://pan.ukbb.broadinstitute.org/

https://pheweb.jp/

- 一方、GWAS統計量は個人情報に相当しないため、データの公開や入手が容易という利点があります。
- 現在、GWAS論文の出版時にはGWAS統計量を一般公開(=特定の手続きを行わずに自由に入手可能な状態)することが推奨されています。
- GWAS統計量を一般公開するデータベースの整備も進んでいます。

① GWAS統計量を用いた解析手法とLDSC

GWAS統計量



データの公開・
入手が容易

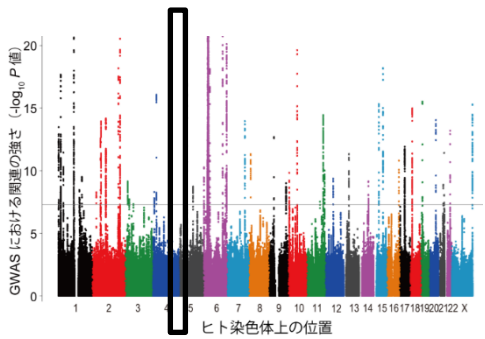
実施可能な解析手法

- GWASメタアナリシス
- Fine-mapping
- Functional annotation
- Gene-based関連解析
- 統計量imputation
- Conditional解析
- TWAS
- PRS予測モデル構築
- Mendelian randomization

- GWAS統計量を対象とした解析手法は、実施に際して要求される**計算機資源が比較的小規模**であり、ノートPCで実施可能な例が多いです。
- 解析手法の発達により、<u>以前は個人別形質・ゲノム情報が必要と考えられていた解析も、GWAS統計量に基づき実施可能</u>になっています。
- その一つに、**LDSCによるheritability推定**があります。

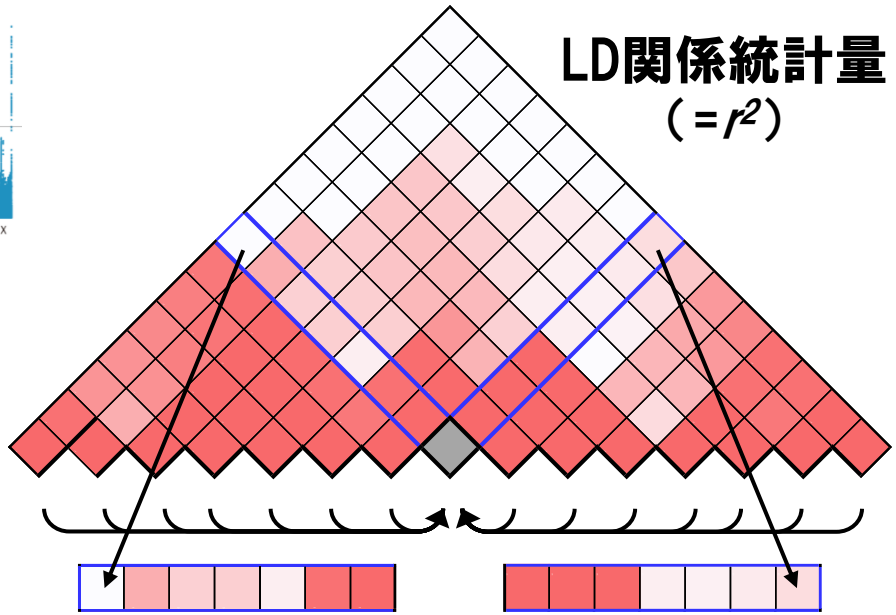
(Pasaniuc B et al. ***Nat Rev Genet*** 2017, Uffelmann E et al. ***Nat Rev Methods Primers*** 2021)

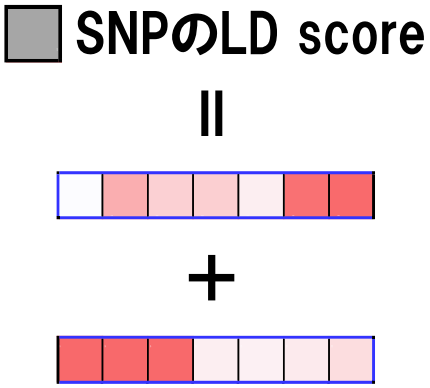
① GWAS統計量を用いた解析手法とLDSC



各領域・SNPで
個別に計算

LD scoreの計算方法





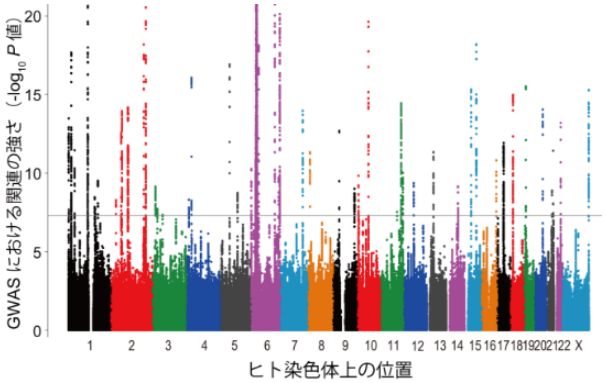
- Linkage Disequilibrium SCore regression（LDSC）は、ゲノム全域に存在するコモンバリアントにおけるLD scoreに基づく解析手法です。
- LD scoreは、予め各SNPにおいて、近傍のLD関係にあるSNPとのLD統計量を足し合わせて計算されています。
- LD scoreが高いSNP程、多くのSNPとLD関係にある、と解釈できます。

（※LD scoreの計算自体には個人別ゲノムデータが必要になります。）

(Bulik-Sullivan BK et al. ***Nat Genet*** 2015)

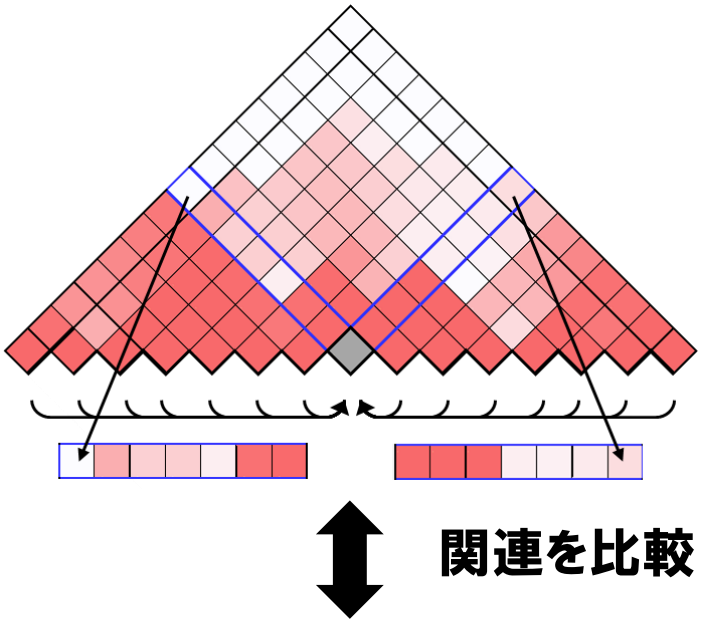
① GWAS統計量を用いた解析手法とLDSC

GWAS統計量

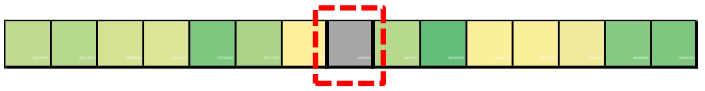


ゲノム全域で評価

LD score



GWAS関連統計量
（= X²値）

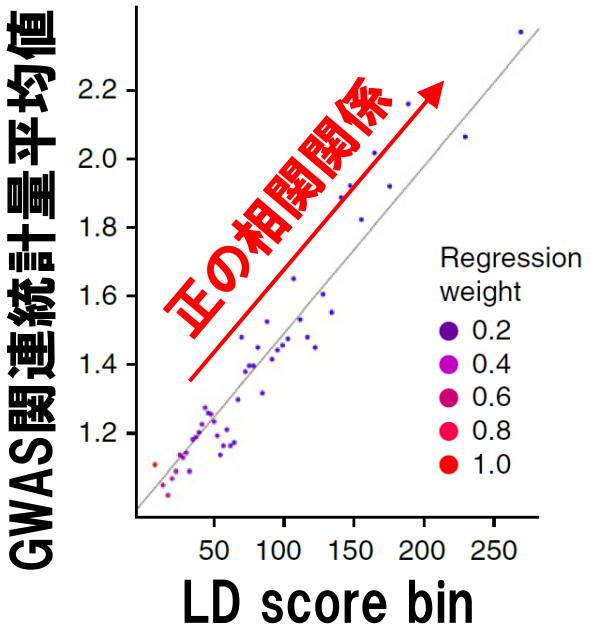


・LDSCでは、各SNPのLD scoreとGWAS関連統計量(=X²値)の関連をゲノム全域に渡って比較(=regression)することで、疾患の遺伝的背景(=heritability)を定量化します。

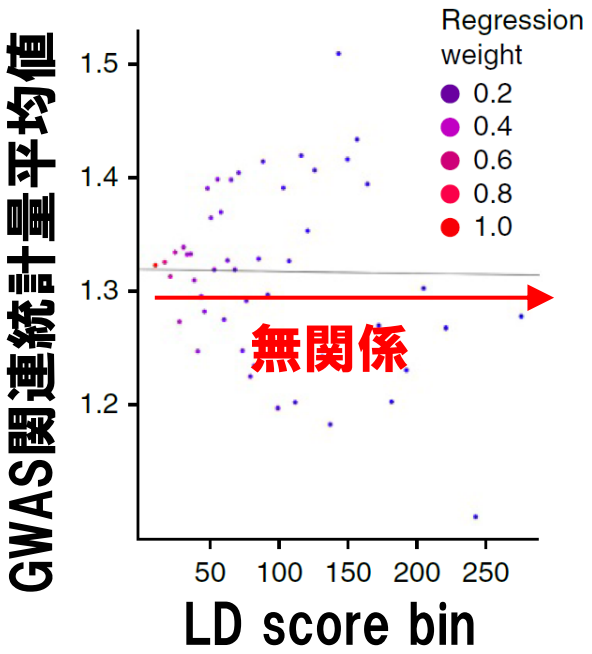
(Bulik-Sullivan BK et al. ***Nat Genet*** 2015)

① GWAS統計量を用いた解析手法とLDSC

Polygenic形質のGWAS



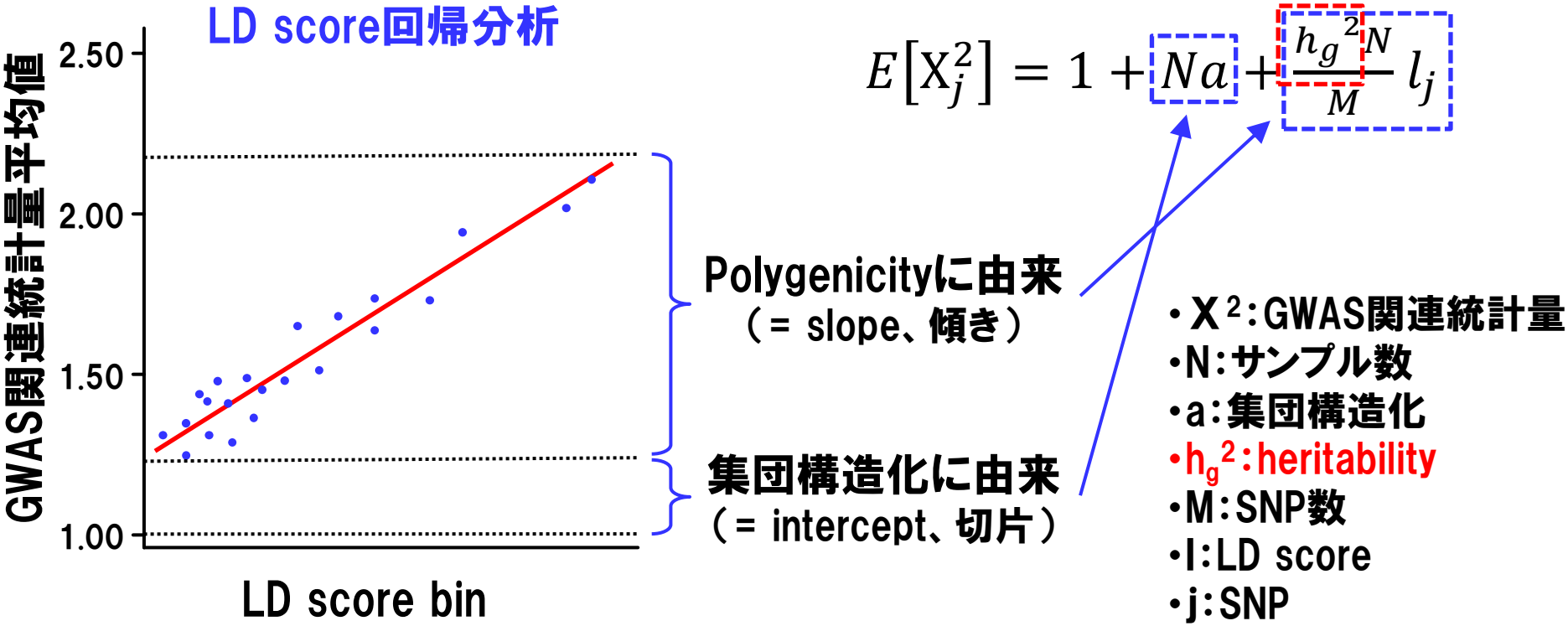
集団構造化のみのGWAS



- Polygenic仮説(=ゲノム全域の無数のSNPが弱い疾患リスクを有する)が成立する形質のGWASでは、LD scoreが高いSNP程、GWAS関連統計量平均値も高くなる傾向(正の相関関係)が認められます。
- GWAS統計量に疾患感受性が反映されず、集団構造化等で見せかけの関連が存在する場合、両者は無関係になります。

(Bulik-Sullivan BK et al. ***Nat Genet*** 2015)

① GWAS統計量を用いた解析手法とLDSC

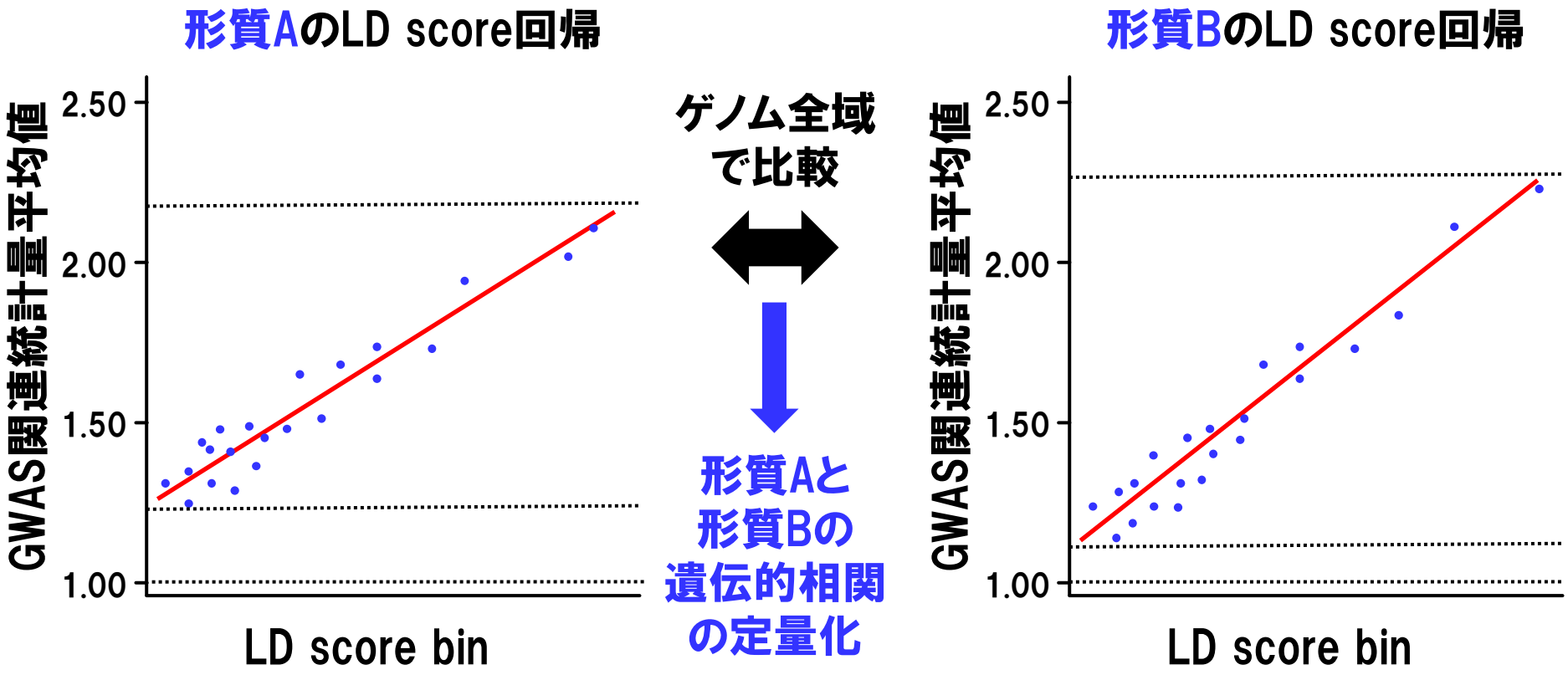


- LDSCにより、GWAS統計量における統計量のinflation（= λ_{GC}の1.00からの乖離）を、集団構造化からの由来（= intercept、切片）と、polygenicityからの由来（= slope、傾き）に分解することができる、とも解釈できます。
- 後者に基づき、heritabilityを推定することができます。

（Heritability=ゲノムワイドな遺伝子多型が説明可能な疾患の遺伝的背景の割合）

(Yang J et al. ***Eur J Hum Genet*** 2011, Bulik-Sullivan BK et al. ***Nat Genet*** 2015)

① GWAS統計量を用いた解析手法とLDSC

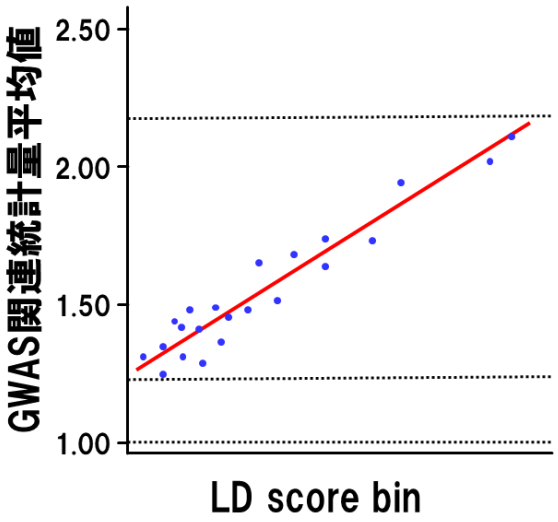


- 二つのpolygenicな形質に対するLD score回帰の結果をゲノム全域で比較することで、遺伝的背景の相関関係(=genetic correlation)を定量化することができます。

(正の相関と負の相関を区別する都合、GWAS関連統計量は正の値のみをとるX^{2}値(=Z値×Z値)でなくZ値を用いて計算されます。) (Bulik-Sullivan BK et al. ***Nat Genet*** 2015)

① GWAS統計量を用いた解析手法とLDSC

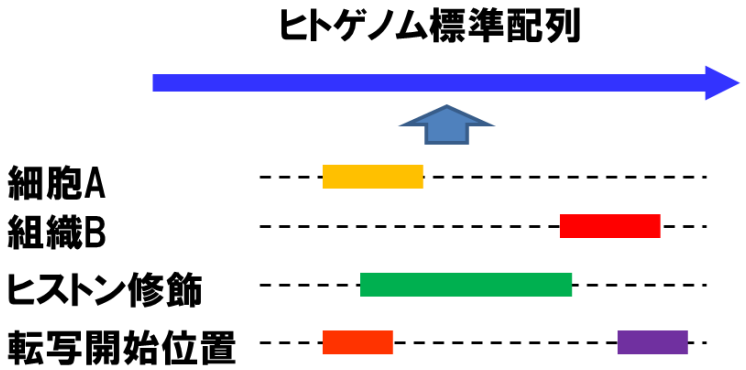
LD score回帰分析



各カテゴリーに属するゲノム領域に層別化した解析

各カテゴリーにおけるheritabilityのenrichmentを定量化

層別化カテゴリー情報



$$E\left[\chi_j^2\right] = 1 + Na + N\sum_{C} \tau_C l(j,C)$$

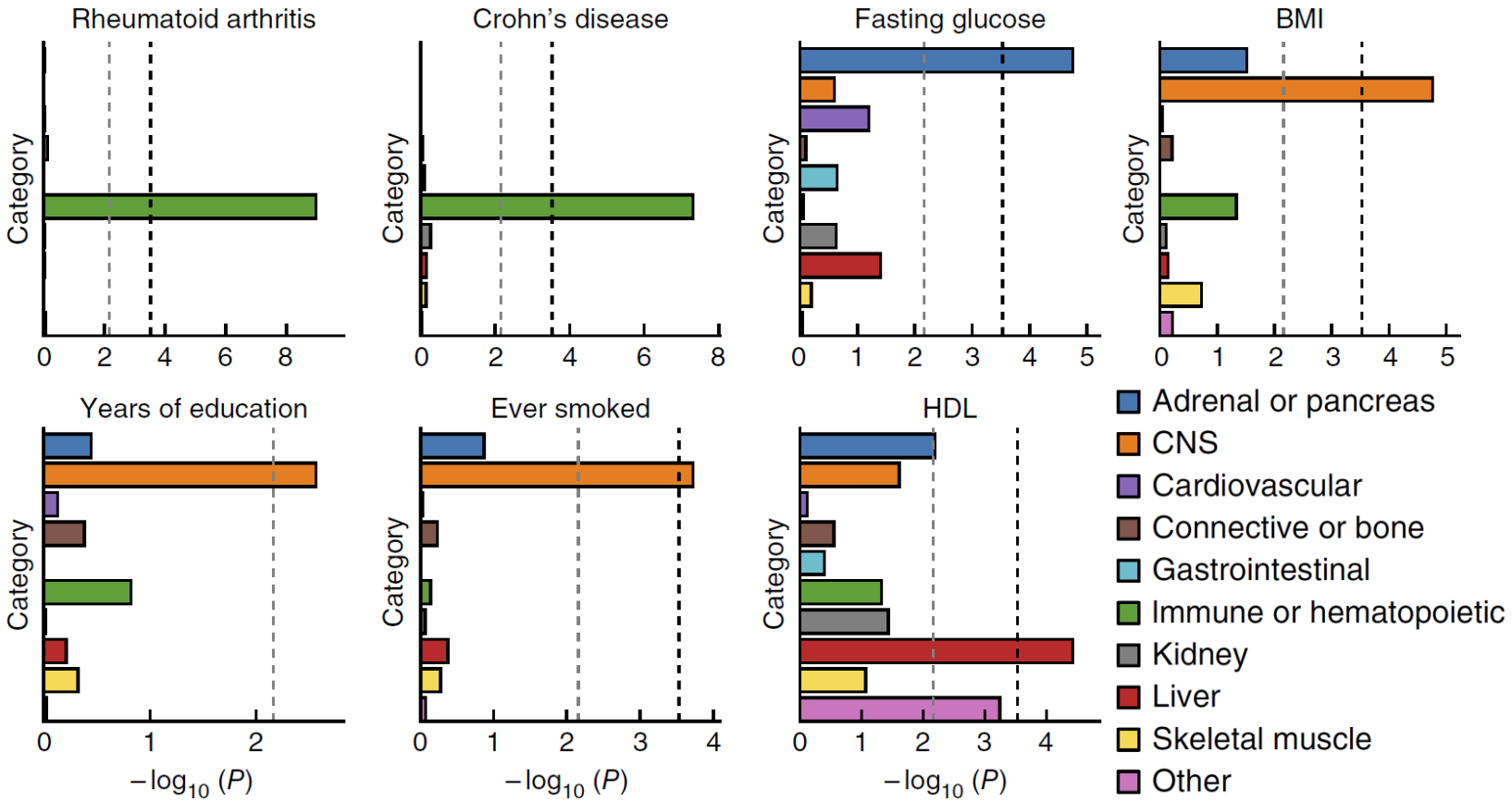
- C：層別化カテゴリー
- τ：SNP毎のheritabilityへの寄与

- ヒトゲノム領域を、細胞組織に特異的なエピゲノム修飾等のカテゴリーに層別化した上でLDSCを実施することで、カテゴリー別のheribtability（=partitioned heritability）のenrichmentを定量化できます（=stratified LDSC）。
- どの細胞組織・エピゲノム修飾が病態に重要か、知ることができます。

(Finucane HK et al. ***Nat Genet*** 2015)

① GWAS統計量を用いた解析手法とLDSC

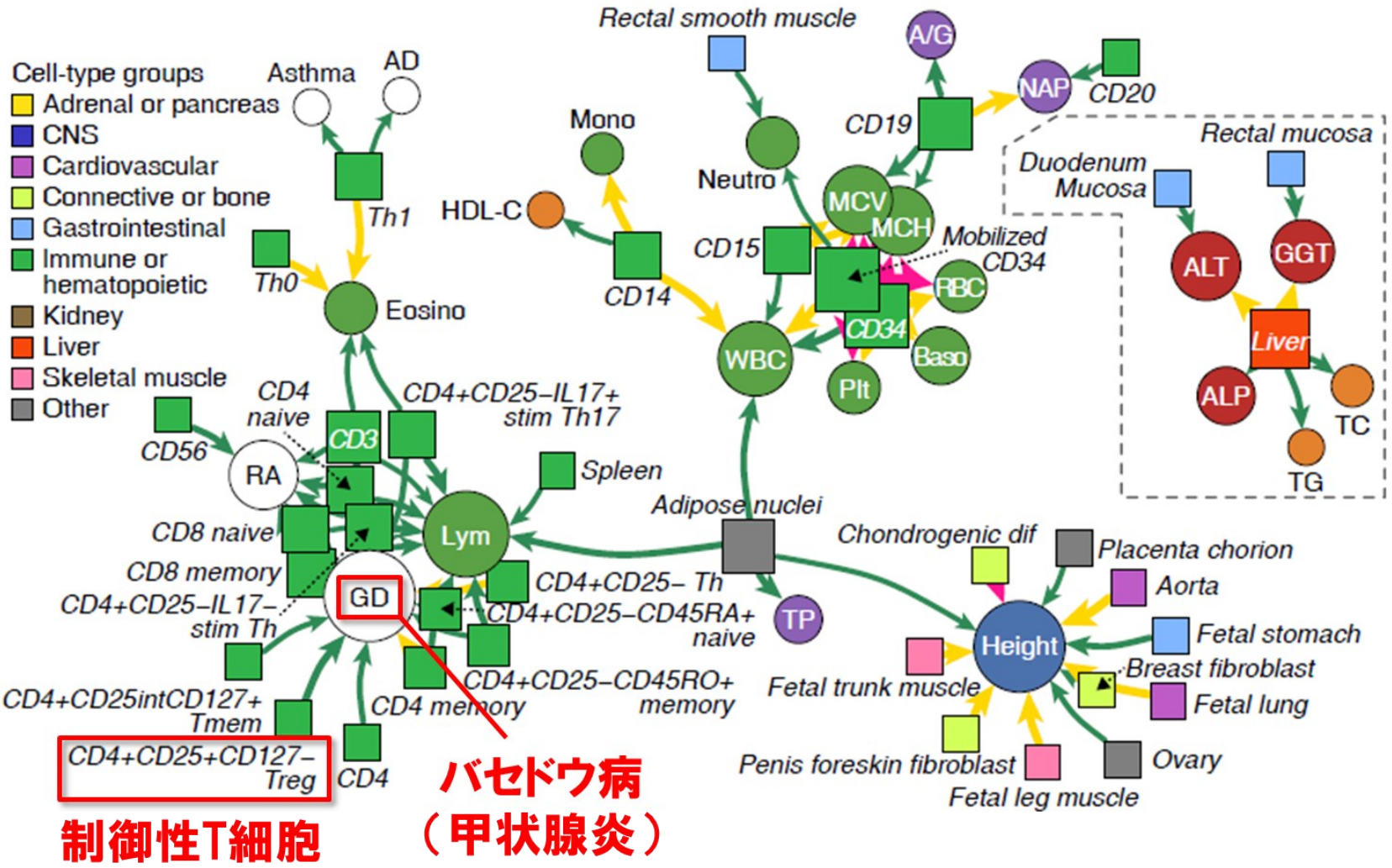
GWAS統計量に対するStratified LDSC解析結果



- Stratified LDSC解析をGWAS統計量に適用した結果、**関節リウマチ・クローン病における免疫細胞、肥満・喫煙歴における中枢神経系**の関与など、疾患の細胞組織特異性が明らかになりました。

(Finucane HK et al. ***Nat Genet*** 2015)

① GWAS統計量を用いた解析手法とLDSC

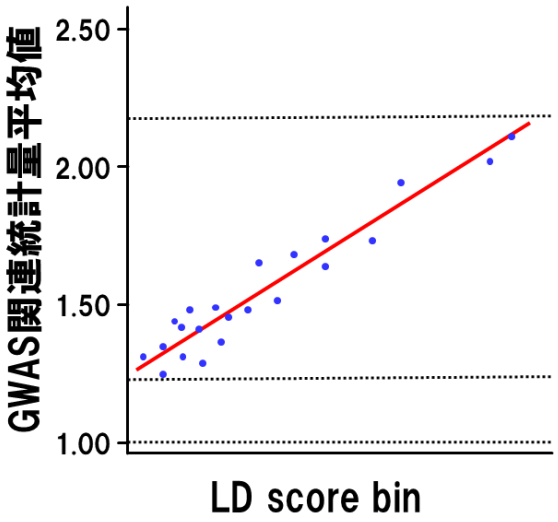


- 日本人集団100形質、200細胞組織を対象にstratified LDSC解析を実施したところ、疾患と細胞組織のネットワークが再構築されました。

(Kanai M et al. ***Nat Genet*** 2018)

① GWAS統計量を用いた解析手法とLDSC

LD score回帰分析



細胞組織別の遺伝子発現を考慮したLDSC解析

各細胞組織における heritabilityのenrichment を細分化して定量化

細胞組織特異的遺伝子発現情報

	細胞組織A	細胞組織B	細胞組織C	細胞組織D	細胞組織E	細胞組織E
遺伝子A						
遺伝子B						
遺伝子C						
遺伝子D						
遺伝子E						
遺伝子F						

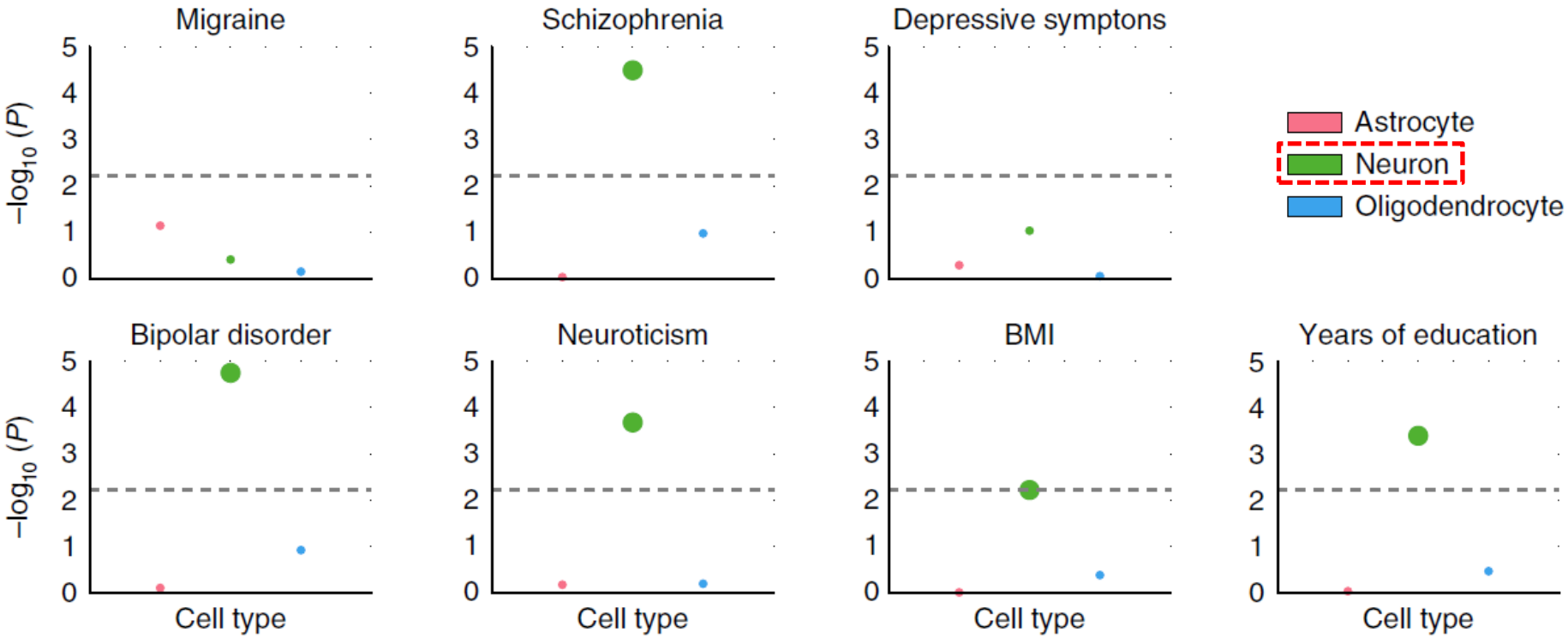
SEG = Specifically Expressed Genes

- 細分化された細胞組織特異性を検討する手法として、細胞組織特異的遺伝子発現情報を考慮したLDSC(=LDSC-SEG)も開発されています。
- どの細胞組織が疾患病態に重要か、より細かい細胞組織分類に基づき検討することができます。

(Finucane HK et al. ***Nat Genet*** 2018)

① GWAS統計量を用いた解析手法とLDSC

GWAS統計量に対するLDSC-SEG解析結果



- 中枢神経系に関わる形質のGWAS統計量にLDSC-SEG解析を適用した結果、Astrocyte、Neuron、Oligodendrocyteの3種類の神経細胞の中で、Neuronの寄与が相対的に大きい事が明らかになりました。

(Cahoy JD et al. ***J Neurosci*** 2008, Finucane HK et al. ***Nat Genet*** 2018)

① GWAS統計量を用いた解析手法とLDSC

- LDSCは、heritabilityの推定や細胞組織特異性の解明など、疾患ゲノムデータ解析の根幹となる情報を、GWAS統計量に基づき小さい計算負荷で推定可能にした点が、画期的な解析手法でした。
- LDSCを実施する際の注意点として、「GWASサンプル数が数千名以上であること」が挙げられます。
- サンプル数が少ないGWASについては、個人別形質・ジェノタイプ情報を用いた手法など、異なるアプローチが必要になります。
- 「GWASのサンプル集団と同一のサンプル集団でLD scoreが計算されていること」、も注意点になります。
- 異なる人種集団間の遺伝的相関の検定にはLDSCは使用できず、異なる解析手法が必要となることに、留意してください。

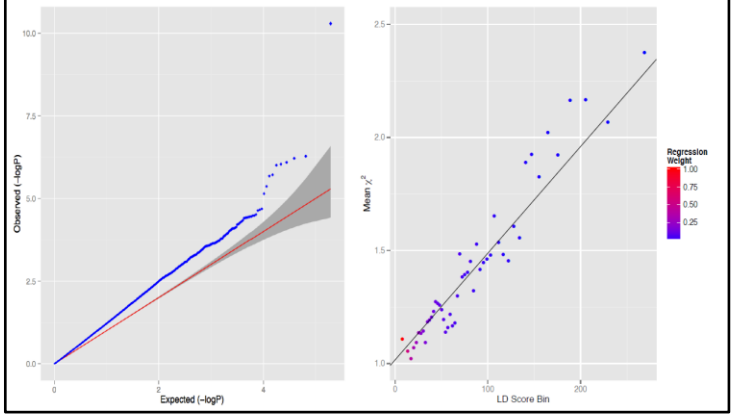
① GWAS統計量を用いた解析手法とLDSC

Use and Interpretation of LD Score Regression

Brendan Bulik-Sullivan
bulik@broadinstitute.org
PGC Stat Analysis Call
BROAD INSTITUTE

Simulated Polygenicity

- λ_{GC} = 1.30; LD Score Regression intercept = 1.02



LD Score Regression

- Regress χ2 statistics against LD Score

$$E[\chi^2 \mid \ell_j] = Nh^2\ell_j / M + Na + 1$$

  - LD Score (L_j) is a property of SNP j, defined as sum r^2, estimated as sum r^2 w/ all other SNPs a 1cM window.
  - N is sample size.
  - M is # SNPs.
  - h^2 is SNP-heritability.
  - a is inflation from pop strat/cryptic relatedness.

Bulik-Sullivan et al., Nat Genet, 2015

URLs

- ldsc
  - github.com/bulik/ldsc
  - Installation instructions
  - FAQ
- Tutorials / wiki
  - github.com/bulik/ldsc/wiki
- Pre-computed European LD Scores
  - broadinstitute.org/~bulik/eur_ldscores/
- ldsc_users google group:
  - groups.google.com/forum/?hl=en#!forum/ldsc_users

https://www.med.unc.edu/pgc/wp-content/uploads/sites/959/2019/01/pgc_stat_bulik_2015.pdf

・LDSCの作者による説明スライドが公開されており、参考になります。

講義の概要

GenomeDataAnalysis6

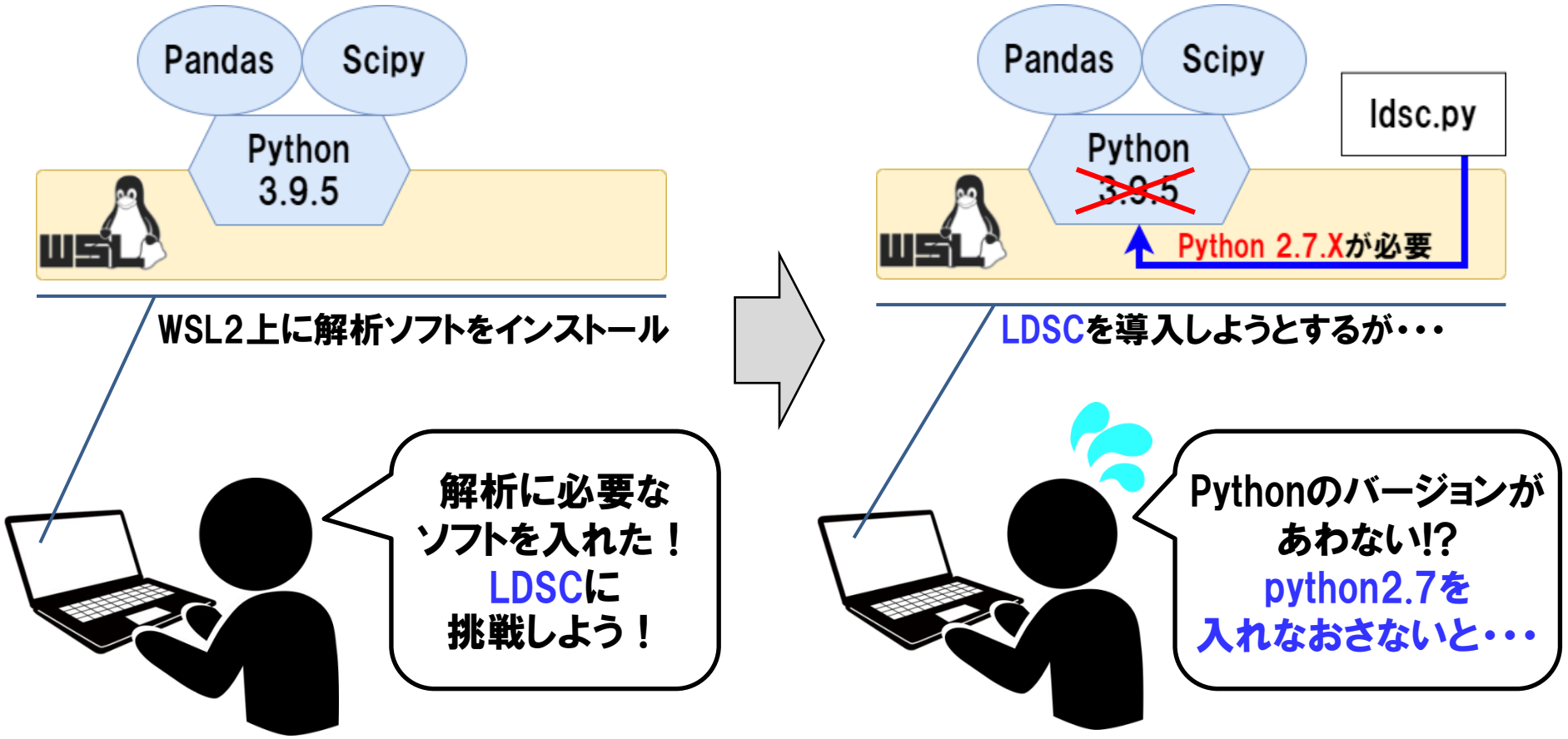
① GWAS統計量を用いた解析手法とLDSC

② Anacondaを使ったLDSCのインストール

③ LDSC解析実習

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。

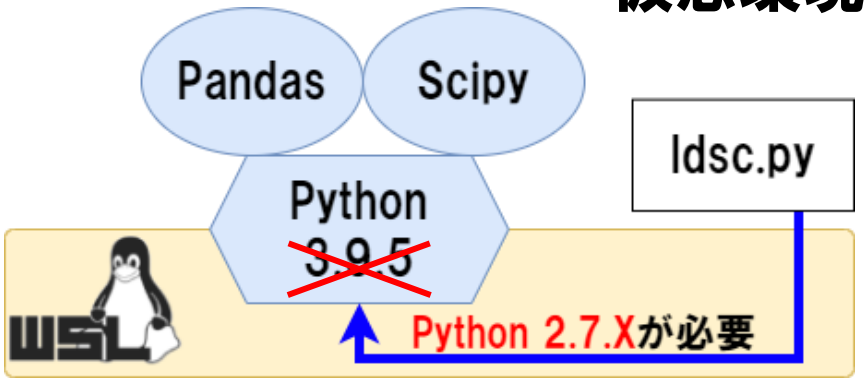
② Anacondaを使ったLDSCのインストール



- ソフトウェアは**他のプログラムやパッケージに依存**している事があります。
- 実際の解析では多くのソフトウェアを使用しますが、**ソフトウェア同士でプログラムのバージョンが合わず**に困ることがよくあります。

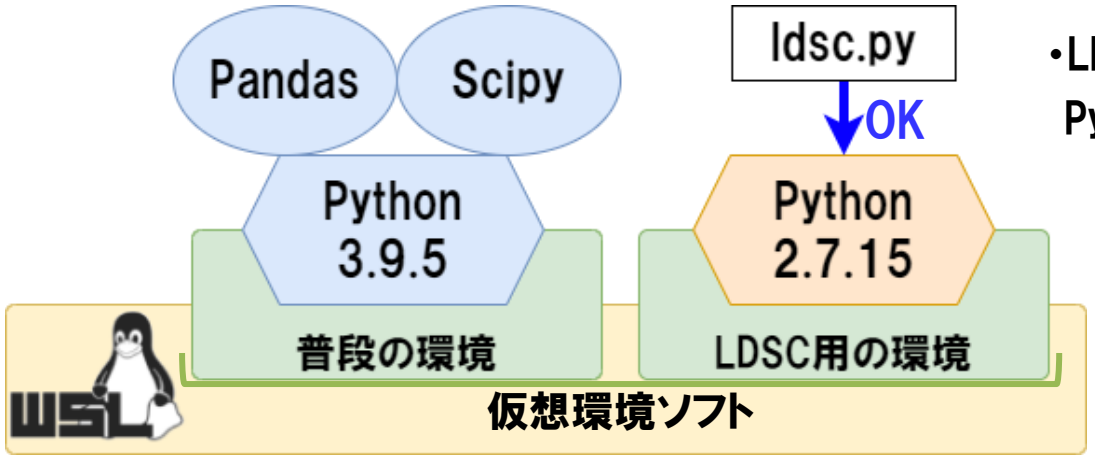
② Anacondaを使ったLDSCのインストール

仮想環境ソフトなし



・Pythonを一度すべて消して入れなおす。
（→今までに構築したモジュールも消える）
・Python2.7を追加で入れる。
（→コマンドが重複するなど、混乱が生じる）

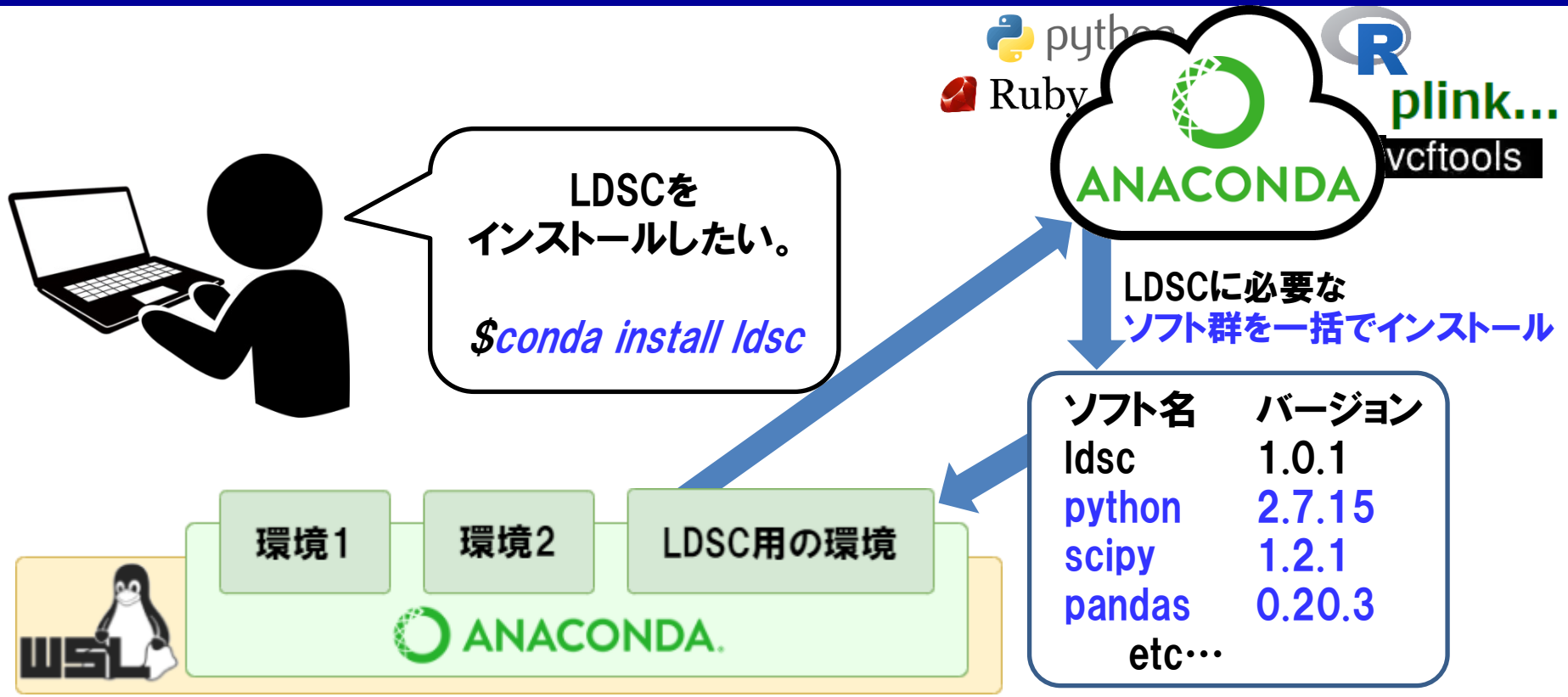
仮想環境ソフトあり



・LDSC用の環境を新規に作成してそこに
Python2.7をインストールする。

・ソフトの依存関係問題を解決する方法の一つに、仮想環境があります。
解析ごとに環境を分けてソフトをインストールできます。
・様々な仮想環境ソフトがありますが、今回はAnacondaを使用します。

② Anacondaを使ったLDSCのインストール



- AnacondaはPythonとそのモジュールを管理するソフトですが、現在はPythonに限らずデータサイエンス用のソフトに広く対応しています。
- 主な機能は、仮想環境を構築すること、Anacondaのリポジトリ(ソフトが揃っているクラウドリソース)からソフトをインストールすることです。
- ソフト名を指定するだけで必要なソフトを一括インストールしてくれます。

② Anacondaを使ったLDSCのインストール

statgen@statgen-PC: ~

$ conda env create -f /mnt/c/SummerSchool/__for_install__/ldsc_setting.yml

ldsc_setting.yml

```
name: ldsc_env
channels:
- bioconda
- conda-forge
dependencies:
- r-base=3.6.1
- r-corrplot=0.84
- r-ggplot2=3.1.1
- python=2.7.18
- bitarray=0.8.3
```

作成する環境名

ソフトを探すAnacondaのリポジトリ名

ダウンロード・インストールするソフト名

- LDSCの実習で必要なAnaconda環境は手順書で説明しています。
- 手順書で実行してもらった上記コマンドは、ymlファイルに記載しているソフトウェアをインストールしつつ、仮想環境を構築するコマンドでした。

② Anacondaを使ったLDSCのインストール

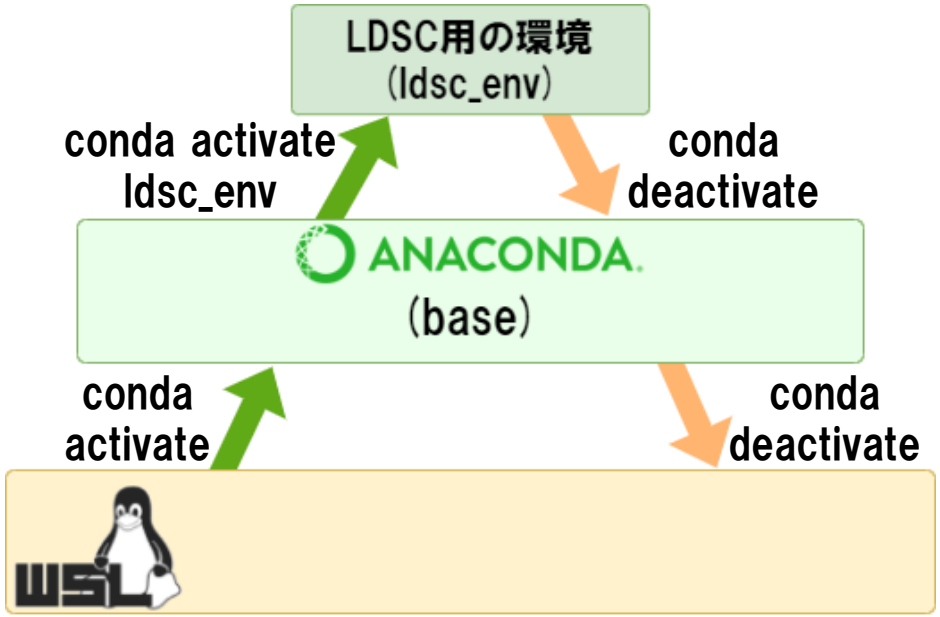
statgen@statgen-PC: ~

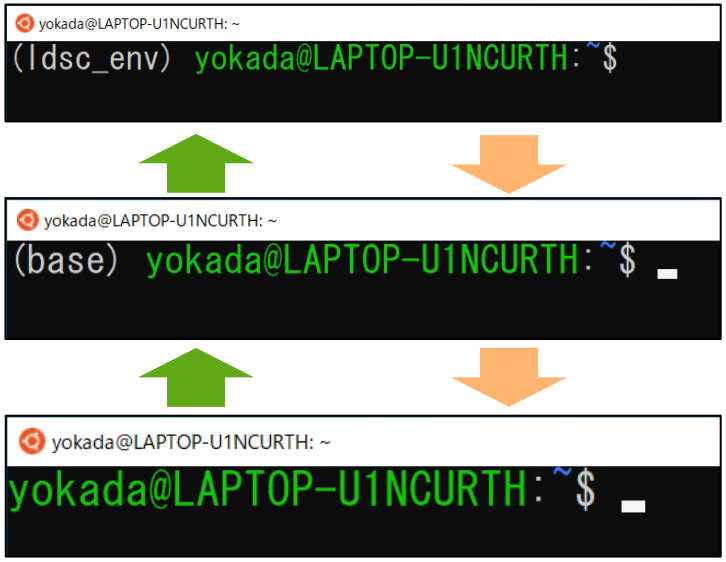
$ conda activate

$ conda activate ldsc_env

$ conda deactivate

$ conda deactivate





- 皆さんのPCには、事前に配布した手順書通りに行っていれば、既に”ldsc_env”という環境ができています。
- コマンドを入力して、環境を切り替えてみましょう。

② Anacondaを使ったLDSCのインストール

M1/M2 MacでAnaconda3がうまく使えない理由？

		x86_x64系		Arm v◯系	
命令セット アーキテクチャ		Intel64	AMD64	Arm v8	
使用ハード		PC・サーバー			スマートフォン
プロセッサ シリーズ	コンシューマPC	IntelCore	AMD Ryzen	Apple M	Apple A Qualcomm Snapdragon
	サーバー用PC	IntelXeon	AMD Ryzen Threadripper AMD EPYC	FUJITSU Processor A64FX	
特徴		高パフォーマンス・設計の高い汎用性		小型・省電力・省スペース・設計の独自性	

殆どのWindows PC
M1以前のMac

M1/M2 Mac

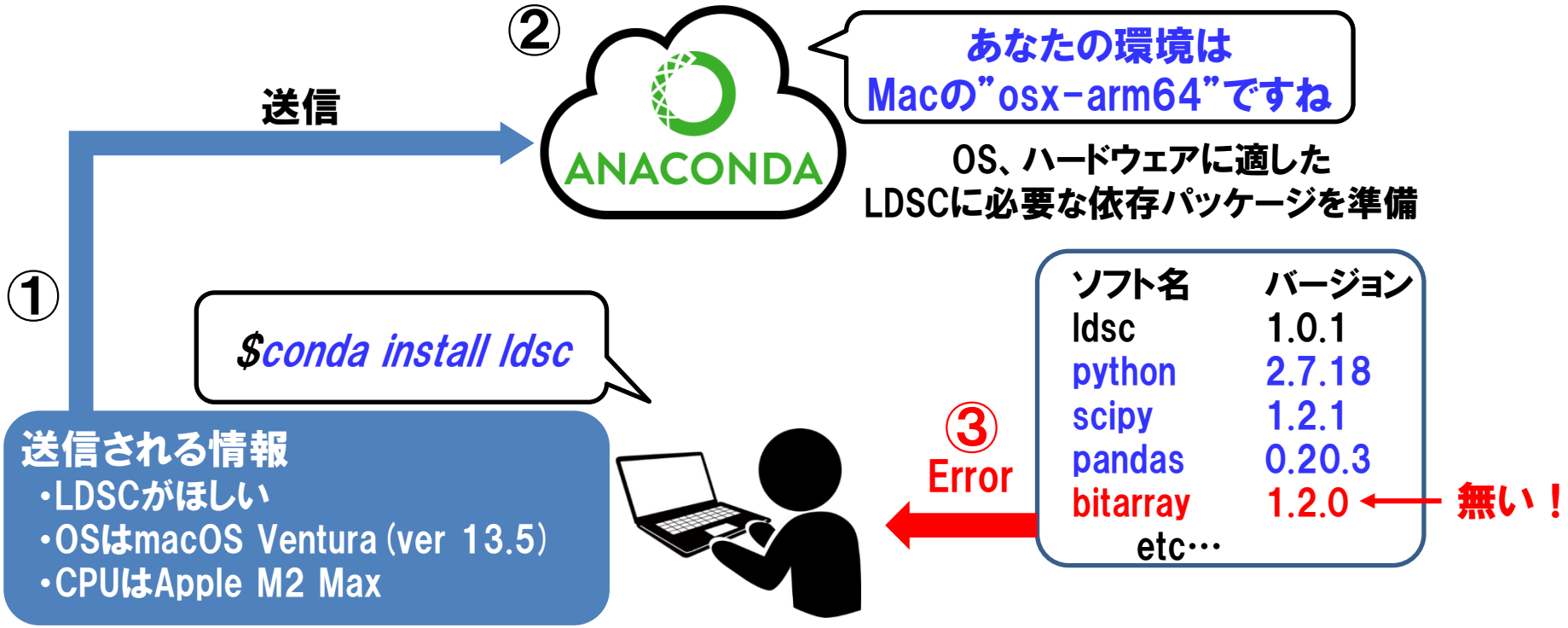
・M1/M2 MacでAnaconda3の実装の不具合が報告されています。

⇒CPU命令セットアーキテクチャ(Instruction Set Architecture：ISA)の違い。

・x84_x64系統とArm系統のCPUにはソフトウェアの互換性はありません。

・Arm版CPUへのソフトウェア実装時には個別CPUにあわせたコンパイルが必要になります。

② Anacondaを使ったLDSCのインストール



- Anacondaは各CPU用のコンパイル済みのパッケージを配布している。
- まだM1/M2 CPUに用意されていないパッケージが相当数ある。
- 根本的な解決策は、必要なパッケージやライブラリを自分で把握してM1/M2 Mac上でソースコードからコンパイルするのが対策方法(エンドユーザーにはまだ敷居が高い？)。
- 今後対応が進んでいく可能性はあります。

講義の概要

GenomeDataAnalysis6

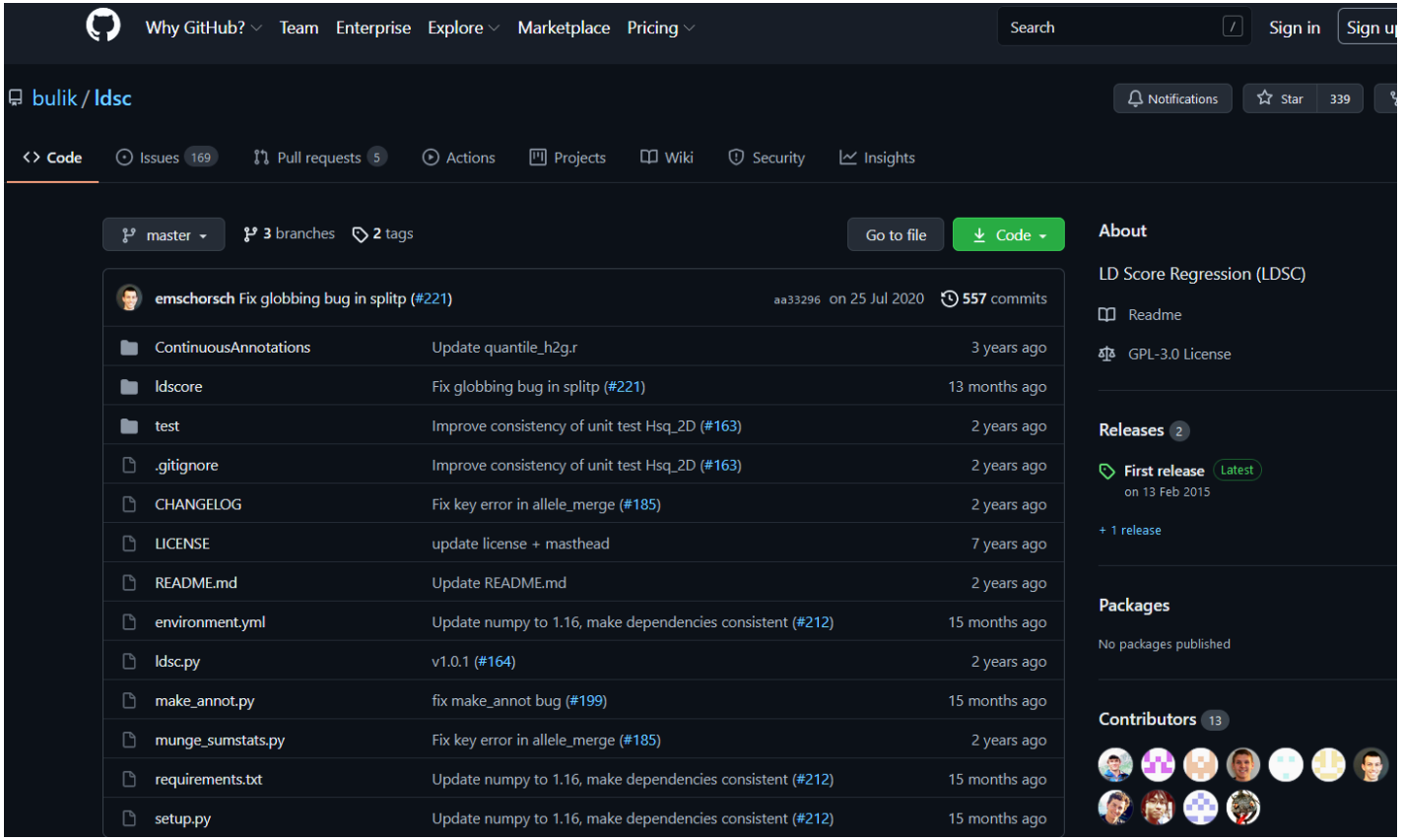
① GWAS統計量を用いた解析手法とLDSC

② Anacondaを使ったLDSCのインストール

③ LDSC解析実習

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 LDSC解析実習



https://github.com/bulik/ldsc

https://www.med.unc.edu/pgc/wp-content/uploads/sites/959/2019/01/pgc_stat_bulik_2015.pdf

- LDSCのソースコードはgithub上で公開されています。
- 実施方法の説明や解析に必要なデータのダウンロードリンク、FAQなど内容が充実しており、手順に沿った一通りの解析も実施可能です。

③ LDSC解析実習

./GenomeDataAnalysis6/Analysis

./Input/

1000G_EUR_Phase3_baseline/
1000G_frq/
1kg_eur/
baseline/
Cahoy_1000Gv3_ldscores/
cell_type_groups/
eas_ldscores/
GWASstats/
weights_hm3_no_hla/

⟺

./Output/

CellTypeSpecificity/
Heritability/
LDscore/

⟺

./ldsc/

.git/
ContinuousAnnotations/
ldscore/
Test/
.gitignore
CHANGELOG
environment.yml
ldsc.log
ldsc.py
ldsc.pyc
LICENSE
make_annot.py
munge_sumstats.py
README.md
requirements.txt
setup.py

解析実習内容

1. LD scoreの計算
2. Heritability推定
3. Genetic correlation解析
4. Stratified LDSC解析
5. LDSC-SEG解析

- LDSCには複数の解析方法があり、また多くのファイル群を扱います。
- 整理の観点から、入力ファイル群(./Input/)、出力ファイル群(./Output/)、実行ファイル群(./ldsc/)、に分類して配置しています。
- 今回の実習では、５種類のLDSC解析を実施します。

③ LDSC解析実習

./Input/

1000G_EUR_Phase3_baseline/
1000G_frq/
1kg_eur/
baseline/
Cahoy_1000Gv3_ldscores/
cell_type_groups/
eas_ldscores/
GWASstats/
weights_hm3_no_hla/

⟺

./Output/

CellTypeSpecificity/
Heritability/
LDscore/

⟺

./ldsc/

.git/
ContinuousAnnotations/
ldscore/
Test/
.gitignore
CHANGELOG
environment.yml
ldsc.log
ldsc.py
ldsc.pyc
LICENSE
make_annot.py
munge_sumstats.py
README.md
requirements.txt
setup.py

解析実習内容

1. LD scoreの計算
2. Heritability推定
3. Genetic correlation解析
4. Stratified LDSC解析
5. LDSC-SEG解析

- 「LD scoreの計算」を行います。
- 主要な人種集団におけるLD scoreは、既に計算され、githubホームページ等で一般公開されており、必ずしも必要な作業ではありません。
- 独自のLD referenceパネルを持っている場合、LD scoreを改めて計算することで、より高解像度の解析が可能になります。

③ LDSC解析実習

(base) statgen@statgen-PC: ~
$ conda activate ldsc_env

※ファイル"LDSC_Command.txt"を開いて、内容をLinuxコマンドにコピー＆ペーストして下さい。

Anaconda環境の起動

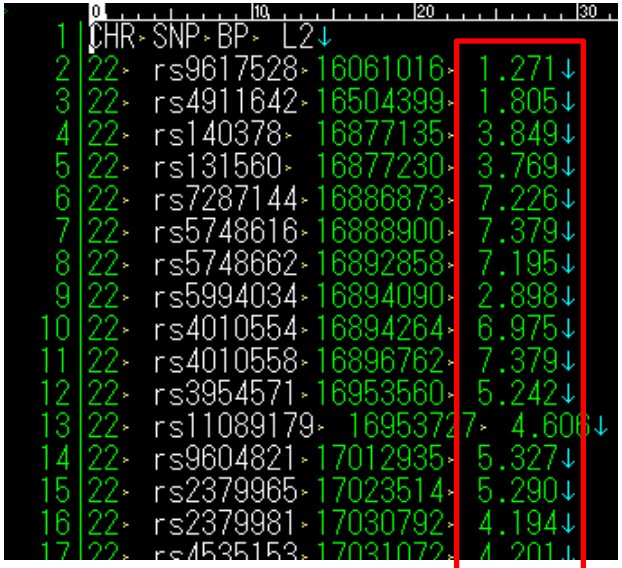
(ldsc_env) statgen@statgen-PC: ~
$ cd /mnt/c/SummerSchool/GenomeDataAnalysis6/Analysis/

(ldsc_env) statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis6/Analysis
$ python ./ldsc/ldsc.py --bfile ./Input/1kg_eur/22 --l2 --ld-wind-cm 1 --out ./Output/LDscore/EUR_22

- AnacondaでLDSCの実行に必要な仮想環境を起動します。
  （LDSC解析演習は、Cygwin環境やM1 M2 macには対応していません。）
- 1000 Genomes Project EURサンプルの22番染色体のSNPジェノタイプファイル(PLINKフォーマット・379名・19,156SNP)を使って、LD scoreを計算します。

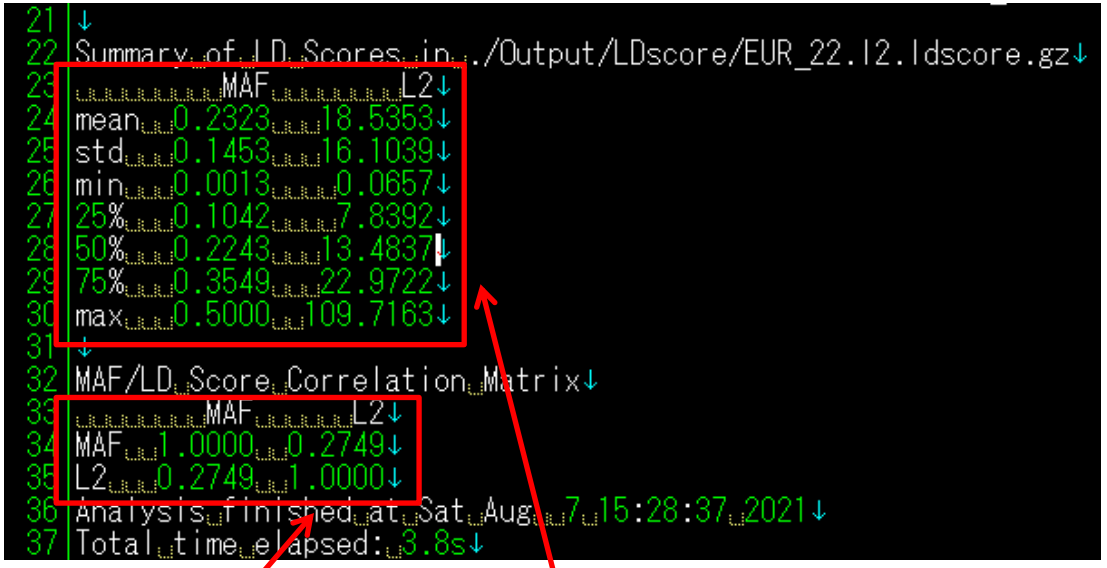
③ LDSC解析実習

"EUR_22.l2.ldscore.gz"
各SNP毎のLD scoreファイル



CHR	SNP	BP	L2
22	rs9617528	16061016	1.271
22	rs4911642	16504399	1.805
22	rs140378	16877135	3.849
22	rs131560	16877230	3.769
22	rs7287144	16886873	7.226
22	rs5748616	16888900	7.379
22	rs5748662	16892858	7.195
22	rs5994034	16894090	2.898
22	rs4010554	16894264	6.975
22	rs4010558	16896762	7.379
22	rs3954571	16953560	5.242
22	rs11089179	16953727	4.606
22	rs9604821	17012935	5.327
22	rs2379965	17023514	5.290
22	rs2379981	17030792	4.194
22	rs4535153	17031072	4.201

"EUR_22.log"
ログファイル



```
Summary of LD Scores in ./Output/LDscore/EUR_22.l2.ldscore.gz
                MAF       L2
mean        0.2323  18.5353
std         0.1453  16.1039
min         0.0013   0.0657
25%         0.1042   7.8392
50%         0.2243  13.4837
75%         0.3549  22.9722
max         0.5000 109.7163

MAF/LD Score Correlation Matrix
            MAF       L2
MAF      1.0000  0.2749
L2       0.2749  1.0000

Analysis finished at Sat Aug  7 15:28:37 2021
Total time elapsed: 3.8s
```

LD score

MAFとLD score
の相関係数

MAFとLD score
の対応関係

- 各SNPに対して、LD scoreが計算されます。
- マイナーアレル頻度(=minor allele frequency; MAF)とLD scoreは**正の相関**があり、MAFが高い程、LD scoreが大きくなる傾向が認められます。

（MAFが大きいSNP間において、LD指標r^2が高い値をとりやすいことに由来します。）

③ LDSC解析実習

./Input/

1000G_EUR_Phase3_baseline/
1000G_frq/
1kg_eur/
baseline/
Cahoy_1000Gv3_ldscores/
cell_type_groups/
eas_ldscores/
GWASstats/
weights_hm3_no_hla/

⇔

./Output/

CellTypeSpecificity/
Heritability/
LDscore/

⇔

./ldsc/

.git/
ContinuousAnnotations/
ldscore/
Test/
.gitignore
CHANGELOG
environment.yml
ldsc.log
ldsc.py
ldsc.pyc
LICENSE
make_annot.py
munge_sumstats.py
README.md
requirements.txt
setup.py

解析実習内容

1. LD scoreの計算
2. Heritability推定
3. Genetic correlation解析
4. Stratified LDSC解析
5. LDSC-SEG解析

- 「Heritability推定」を行います。
- LD scoreには、東アジア人集団を対象に計算済みの値を使用します。
- ./Input/GWASstats/に用意した、日本人集団の身長(Height)・肥満(body mass index; BMI)・2型糖尿病(type 2 diabetes; T2D)のGWAS統計量を使用します。

③ LDSC解析実習

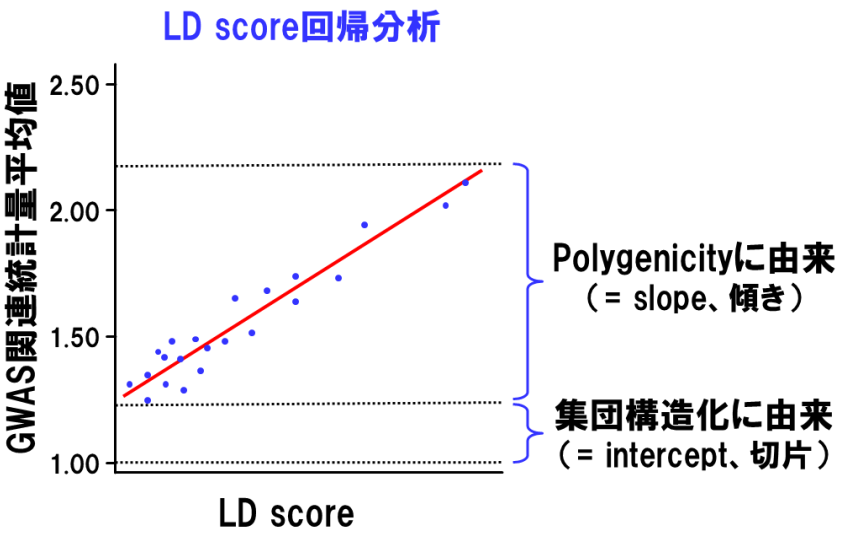
```
(ldsc_env) statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis6/Analysis
$ python ./ldsc/ldsc.py --h2 ./Input/GWASstats/BMI_JPT.sumstats.gz --ref-ld-chr
  ./Input/eas_ldscores/ --w-ld-chr ./Input/eas_ldscores/ --out ./Output/Heritability/Herit_BMI_JPT

$ python ./ldsc/ldsc.py --h2 ./Input/GWASstats/Height_JPT.sumstats.gz --ref-ld-chr
  ./Input/eas_ldscores/ --w-ld-chr ./Input/eas_ldscores/ --out ./Output/Heritability/Herit_Height_JPT

$ python ./ldsc/ldsc.py --h2 ./Input/GWASstats/T2D_JPT.sumstats.gz --ref-ld-chr
  ./Input/eas_ldscores/ --w-ld-chr ./Input/eas_ldscores/ --out ./Output/Heritability/Herit_T2D_JPT
```

※ファイル”LDSC_Command.txt”を開いて、内容をLinuxコマンドにコピー＆ペーストして下さい。

- **GWAS統計量は、データベース（https://pheweb.jp/）からダウンロード後、munge_sumstats.pyを用いてLDSC解析用にフォーマット変更しています。**
- **LDSC解析は、ゲノム全域の（全てではなく）主要なコモンバリアントを対象とし、HapMap3プロジェクトのSNP（約100万箇所）を使用しています。**
  （LDSC解析用フォーマットやSNPの内訳の詳細はgithubを参照してください。）

③ LDSC解析実習

"Herit_XXX_JPT.log"
各形質のLDSC解析結果



GWAS統計量
全体のinflation

集団構造化に
由来するinflation

	$Lambda_{GC}$	Mean X^2	Intercept	h^2
BMI	1.49	1.68	1.09	0.169
Height	1.69	2.42	1.21	0.326
T2D	1.13	1.31	0.96	0.097

Heritability
推定値

- GWAS統計量の帰無仮説からの乖離をinflation（= λ_{GC}の1.00からの乖離）と呼びます。LDSC解析により、集団構造化とpolygenicityに由来するinflationを分離することが可能になりました。
- いずれの形質も、inflationに占める集団構造化の影響は小さそうです。
- GWASデータで説明可能なheritabilityは、身長＞肥満＞2型糖尿病の順に高いようです。

③ LDSC解析実習

./Input/

```
1000G_EUR_Phase3_baseline/
1000G_frq/
1kg_eur/
baseline/
Cahoy_1000Gv3_ldscores/
cell_type_groups/
eas_ldscores/
GWASstats/
weights_hm3_no_hla/
```

⇔

./Output/

```
CellTypeSpecificity/
Heritability/
LDscore/
```

⇔

./ldsc/

```
.git/
ContinuousAnnotations/
ldscore/
Test/
.gitignore
CHANGELOG
environment.yml
ldsc.log
ldsc.py
ldsc.pyc
LICENSE
make_annot.py
munge_sumstats.py
README.md
requirements.txt
setup.py
```

解析実習内容

1. LD scoreの計算
2. Heritability推定
3. Genetic correlation解析
4. Stratified LDSC解析
5. LDSC-SEG解析

- 「Genetic correlation解析」を行います。
- 引き続き、東アジア集団のLD scoreを使用します。
- 3形質の日本人集団のGWAS統計量(身長・肥満・2型糖尿病)の遺伝的相関をペアワイズで推定します。

③ LDSC解析実習

```
(ldsc_env) statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis6/Analysis
$ python ./ldsc/ldsc.py --rg
  ./Input/GWASstats/BMI_JPT.sumstats.gz,./Input/GWASstats/Height_JPT.sumstats.gz --ref-ld-chr
  ./Input/eas_ldscores/ --w-ld-chr ./Input/eas_ldscores/ --out ./Output/Heritability/Cor_BMI_Height_JPT

$ python ./ldsc/ldsc.py --rg
  ./Input/GWASstats/BMI_JPT.sumstats.gz,./Input/GWASstats/T2D_JPT.sumstats.gz --ref-ld-chr
  ./Input/eas_ldscores/ --w-ld-chr ./Input/eas_ldscores/ --out ./Output/Heritability/Cor_BMI_T2D_JPT

$ python ./ldsc/ldsc.py --rg
  ./Input/GWASstats/Height_JPT.sumstats.gz,./Input/GWASstats/T2D_JPT.sumstats.gz --ref-ld-chr
  ./Input/eas_ldscores/ --w-ld-chr ./Input/eas_ldscores/ --out ./Output/Heritability/Cor_Height_T2D_JPT
```

※ファイル”LDSC_Command.txt”を開いて、内容をLinuxコマンドにコピー＆ペーストして下さい。

- **Genetic correlation解析を実施する場合、2形質のGWAS統計量を一つのコマンドライン上で指定します。**

③ LDSC解析実習

”Cor_XXX_XXX_JPT.log”
2形質間のLDSC解析結果

”Cor_BMI_T2D_JPT.log”

```
47 Genetic Covariance
48 ----------------
49 Total Observed scale gencov: 0.0339 (0.0063)
50 Mean z1*z2: 0.144
51 Intercept: 0.0215 (0.0122)
52
53 Genetic Correlation
54 ------------------
55 Genetic Correlation: 0.2524 (0.0538)
56 Z-score: 4.6937
57 P: 2.6834e-06
58
```

P値　　遺伝的相関

3形質間の
遺伝的相関関係

	BMI	Height	T2D
BMI	-	-0.070	0.252
Height	-0.070	-	-0.008
T2D	0.252	-0.008	-

- 各形質ペアごとに、遺伝的相関(genetic correlation)が計算されます。
- 正の相関か、負の相関か、相関は有意かどうか、を確認します。
- 肥満と2型糖尿病の場合、正の遺伝的相関(r_g=0.252)が、有意に確認されました(P = 2.6×10^{-6})。
- 一方、肥満と身長、2型糖尿病と身長の遺伝的相関は小さいようです。

③ LDSC解析実習

./Input/

```
1000G_EUR_Phase3_baseline/
1000G_frq/
1kg_eur/
baseline/
Cahoy_1000Gv3_ldscores/
cell_type_groups/
eas_ldscores/
GWASstats/
weights_hm3_no_hla/
```

⇔

./Output/

```
CellTypeSpecificity/
Heritability/
LDscore/
```

⇔

./ldsc/

```
.git/
ContinuousAnnotations/
ldscore/
Test/
.gitignore
CHANGELOG
environment.yml
ldsc.log
ldsc.py
ldsc.pyc
LICENSE
make_annot.py
munge_sumstats.py
README.md
requirements.txt
setup.py
```

解析実習内容

1. LD scoreの計算
2. Heritability推定
3. Genetic correlation解析
4. Stratified LDSC解析
5. LDSC-SEG解析

- 「Stratified LDSC解析」を行います。
- UKバイオバンクの肥満GWAS統計量(UKBB BMI)と、欧米人集団を対象に計算されたLD score、細胞組織別のエピゲノム修飾状況ファイルを使用します。

③ LDSC解析実習

```
(ldsc_env) statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis6/Analysis
$ python ./ldsc/ldsc.py --h2 ./Input/GWASstats/BMI_UKBB.sumstats.gz --w-ld-chr
  ./Input/weights_hm3_no_hla/weights. --ref-ld-chr ./Input/cell_type_groups/CNS.,./Input/baseline/baseline.
  --overlap-annot --frqfile-chr ./Input/1000G_frq/1000G.mac5eur.
  --out ./Output/CellTypeSpecificity/BMI_UKBB_CNS --print-coefficients

$ python ./ldsc/ldsc.py --h2 ./Input/GWASstats/BMI_UKBB.sumstats.gz --w-ld-chr
  ./Input/weights_hm3_no_hla/weights. --ref-ld-chr ./Input/cell_type_groups/Other.,./Input/baseline/baseline.
  --overlap-annot --frqfile-chr ./Input/1000G_frq/1000G.mac5eur.
  --out./Output/CellTypeSpecificity/BMI_UKBB_Other --print-coefficients
```

※ファイル”LDSC_Command.txt”を開いて、内容をLinuxコマンドにコピー＆ペーストして下さい。

- **中枢神経系(”central nerve system; CNS”)とその他の細胞組織(”Other”)の二つの細胞組織を対象に、stratified LDSC解析を実施します。**
- **実際には、より多くの細胞組織に対して並行してstratified LDSC解析を実施します。**

③ LDSC解析実習

"BMI_UKIBB_XXX.results"
Stratified LDSC解析結果

Category	Proportion of SNPs	Proportion of h^2 (SE)	Enrichment (SE)	P値
CNS	0.149	0.399 (0.021)	2.68 (0.14)	6.2E-28
Other	0.203	0.323 (0.027)	1.59 (0.13)	7.5E-06

- CNS特異的なエピゲノム修飾がゲノム全体のSNPの14.9%を占める一方、それらのSNPはheritabilityの39.9%を説明し、約2.68倍（=0.399/0.149）、heritabilityがenrichしていると考えられました。
- 一方、Other特異的なエピゲノム修飾は、そこまで強くenrichしていないようです。
- 相対的に、CNSの方が疾患病態への関与が強いと考えらえます。

③ LDSC解析実習

./Input/

1000G_EUR_Phase3_baseline/
1000G_frq/
1kg_eur/
baseline/
Cahoy_1000Gv3_ldscores/
cell_type_groups/
eas_ldscores/
GWASstats/
weights_hm3_no_hla/

⇔

./Output/

CellTypeSpecificity/
Heritability/
LDscore/

⇔

./ldsc/

.git/
ContinuousAnnotations/
ldscore/
Test/
.gitignore
CHANGELOG
environment.yml
ldsc.log
ldsc.py
ldsc.pyc
LICENSE
make_annot.py
munge_sumstats.py
README.md
requirements.txt
setup.py

解析実習内容

1. LD scoreの計算
2. Heritability推定
3. Genetic correlation解析
4. Stratified LDSC解析
5. LDSC-SEG解析

- 「LDSC-SEG解析」を行います。
- UKバイオバンクの肥満GWAS統計量(UKBB BMI)と、欧米人集団を対象に計算されたLD score、細胞組織特異的遺伝子発現情報を使用します。

③ LDSC解析実習

```
(ldsc_env) statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis6/Analysis
$ python ./ldsc/ldsc.py --h2-cts ./Input/GWASstats/BMI_UKBB.sumstats.gz --ref-ld-chr
  ./Input/1000G_EUR_Phase3_baseline/baseline. –out
  ./Output/CellTypeSpecificity/BMI_UKBB_CellTypeSpecificity --ref-ld-chr-cts
  ./Input/Cahoy_1000Gv3_ldscores/Cahoy.v2.ldcts  --w-ld-chr ./Input/weights_hm3_no_hla/weights.
```

※ファイル”LDSC_Command.txt”を開いて、内容をLinux コマンドにコピー＆ペーストして下さい。

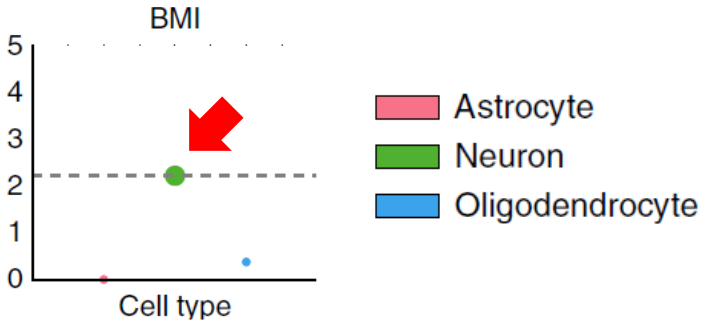
- **今回は、マウスの3種類の中枢神経細胞（Astrocyte、Neuron、Oligodendrocyte）における遺伝子発現情報を使用します。**

(Cahoy JD et al. ***J Neurosci*** 2008)

③ LDSC解析実習

”BMI_UKBB_CellTypeSpecificity.cell_type_results.txt”
LDSC-SEG解析結果

Cell type	Coefficient (SE)	P値
Neuron	7.9E-09 (3.0E-09)	0.0044
Oligodendrocyte	7.3E-10 (3.5E-09)	0.42
Astrocyte	5.8E-09 (2.6E-09)	0.99

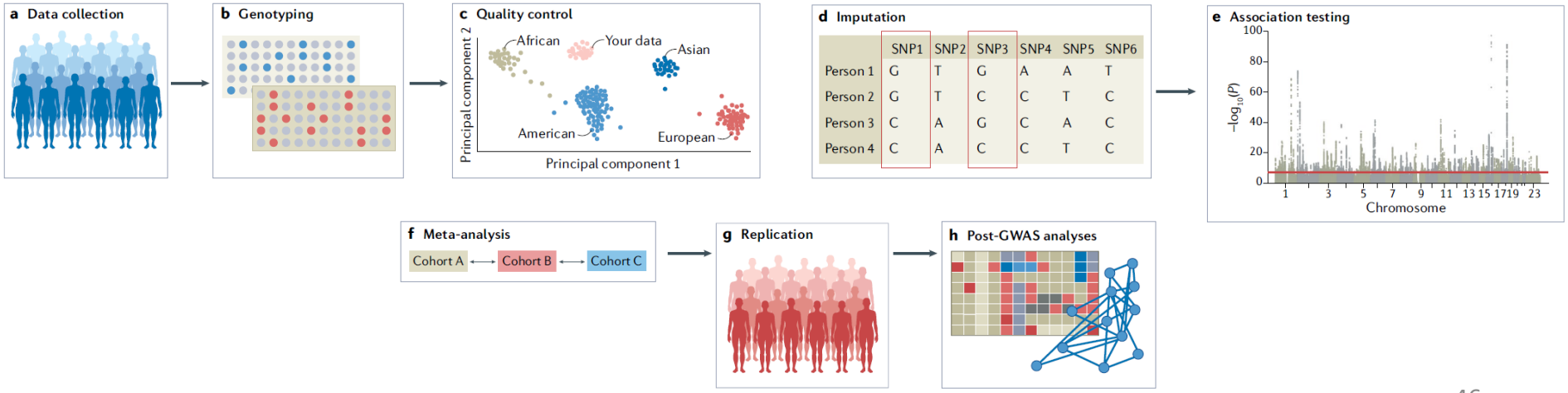


・肥満のGWAS統計量にLDSC-SEG解析を適用した結果、Astrocyte、Neuron、Oligodendrocyteの3種類の神経細胞の中で、Neuronの寄与が相対的に大きい事が確認できました。

(Cahoy JD et al. ***J Neurosci*** 2008, Finucane HK et al. ***Nat Genet*** 2018)

終わりに

- GWASゲノムデータ解析に際して使用する入力データのうち、GWAS統計量を用いた解析手法として、LDSCを取り上げてみました。
- LDSCは、heritabilityの推定や細胞組織特異性の解明など、GWAS統計量を用いたデータ解析の応用可能性を広げた点が画期的でした。
- 現在、GWAS統計量を用いた多くのデータ解析手法が開発されています。
- どのような手法で、どのような解析が可能か、総説等で最新の状況をアップデートしてみると、いいかもしれません。



(Uffelmann E et al. ***Nat Rev Methods Primers*** 2021)