2023年8月25-27日
遺伝統計学・夏の学校　講義実習資料

# GenomeDataAnalysis1

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

## GenomeDataAnalysis1

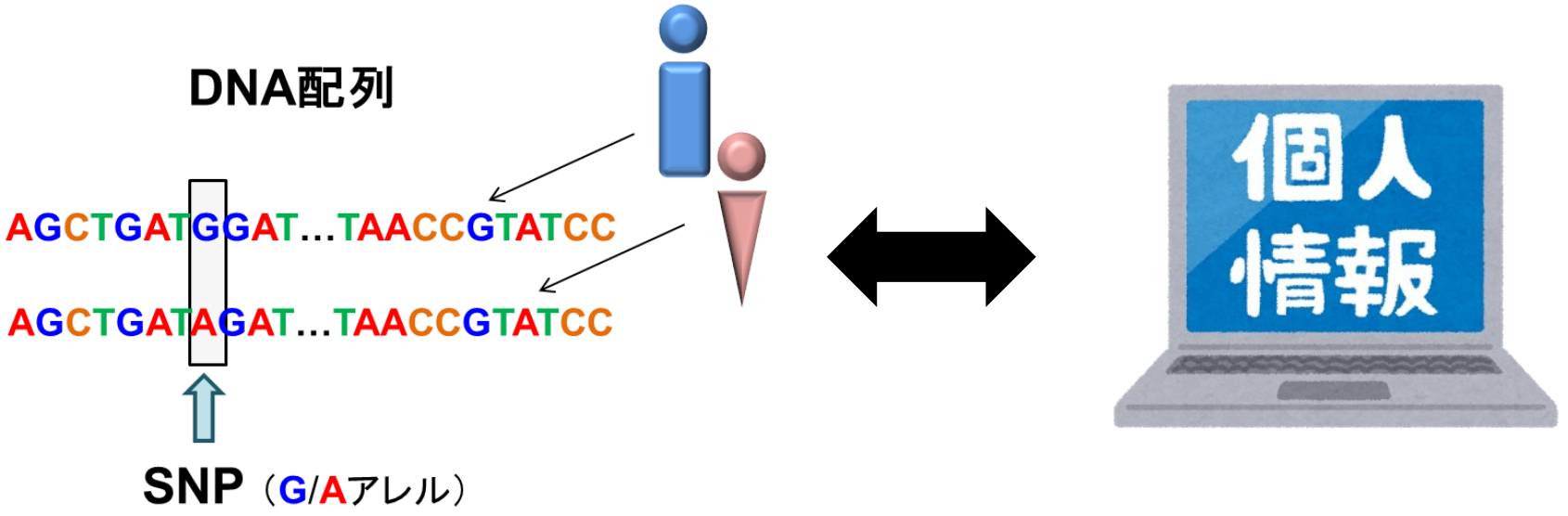
① ヒトゲノムデータの取り扱い

② 1000 Genomes Projectデータ

③ 遺伝統計解析ソフトPLINK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 ヒトゲノムデータの取り扱い




- ヒトゲノムデータ解析の最初のステップは、**ヒトゲノムデータの入手**です。
- 世の中には沢山のヒトゲノムデータがありますが、厳重に管理され、<u>ほとんどは自由にアクセスすることが出来ません</u>。
- なぜなら、ヒトゲノムデータは個人の特定が可能な、**個人情報としての側面**があるからです（平成29年に施行された、改正個人情報保護法におけるゲノムデータ取り扱いについては、本講義では説明しません）。

# ① ヒトゲノムデータの取り扱い

## ヘルシンキ宣言

- 患者・被験者福利の尊重
- 本人の自発的・自由意思による参加
- インフォームド・コンセント取得の必要
- 倫理審査委員会の存在
- 常識的な医学研究であること

## 倫理審査委員会での審議

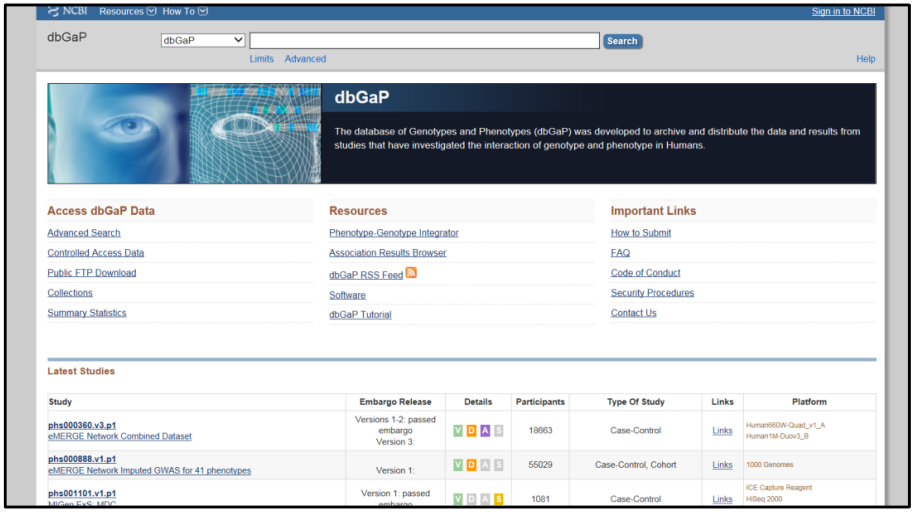
- 何を目的とした研究で
- どの研究施設の
- どの研究者が
- どのような人を対象に
- どのように同意を得て
- どのような研究を行い
- どう結果を管理・公開するか
- 結果は参加者に返却されるのか

- ヒトゲノムを収集・解析するためには、規約を遵守する必要があります。
- ヒトを対象とした科学研究を対象とした**ヘルシンキ宣言**等の指針の遵守や、各研究施設に設置された**倫理審査委員会**での承認が必要です。
- 研究実施時にも、得られたヒトゲノムデータは**厳重な管理**が必要です。

# ① ヒトゲノムデータの取り扱い

## dbGAP

http://www.ncbi.nlm.nih.gov/gap




## NBDCデータベース

http://biosciencedbc.jp/


- 一方で、公的資金を用いて得られたヒトゲノムデータは、**リソースとして公開・共有されるべき**、という意見もあります。
- 公開データの有効な2次利用により、多くの研究成果が生まれています。
- dbGAP/NBDCのような、**公的データベース**にヒトゲノムデータが登録されていて、一定の条件を満たすと、2次利用ができます。

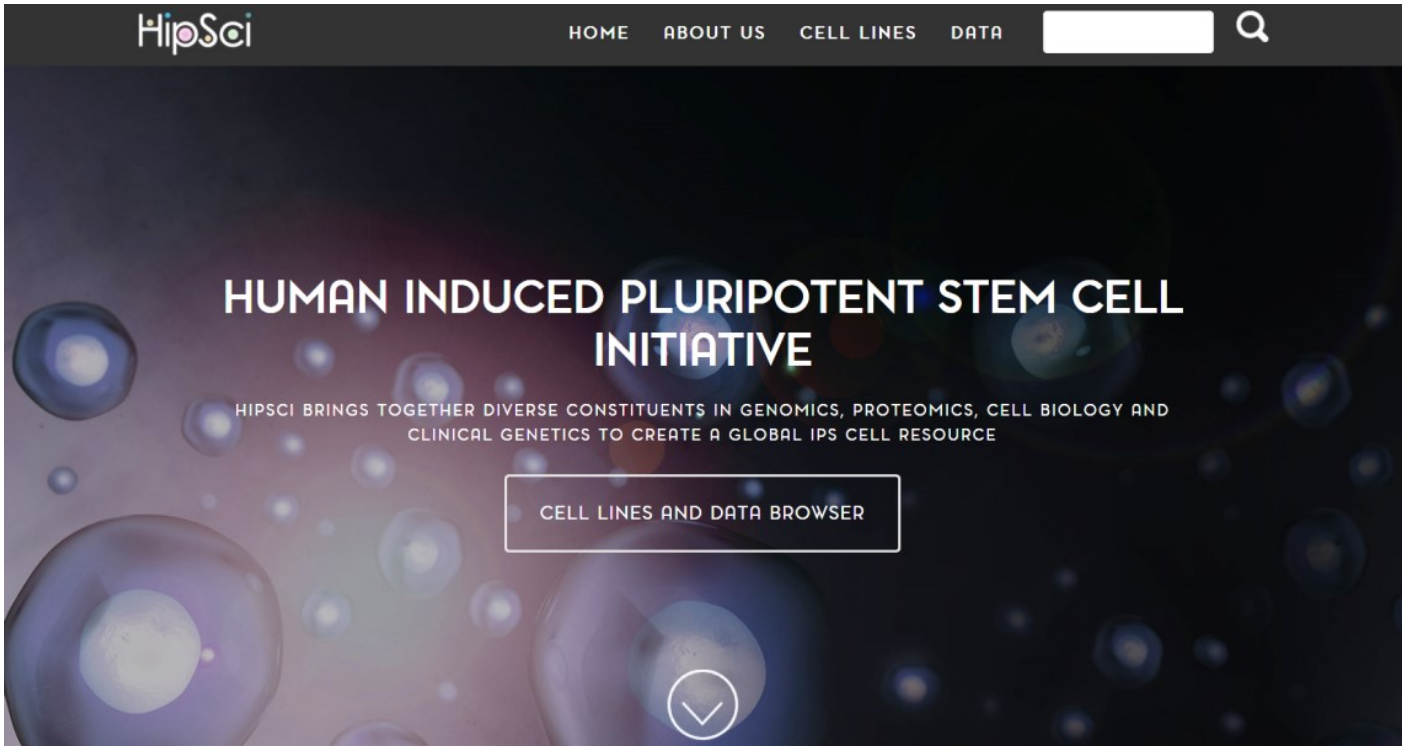
# ① ヒトゲノムデータの取り扱い



- ヒトゲノム研究全般を推進するため、最初からゲノムデータを公開することを目的として実施されたプロジェクトもあります。
- 2000年代初頭に行われた国際HapMap Projectと、2010年代初頭に行われた1000 Genomes Projectが有名です。
- これらのプロジェクトで取得されたヒトゲノムデータは、誰でもアクセスできるように公開されてきました。

（公開データでもヒト由来であることにかわりないので適切に取り扱う必要があります。）

# ① ヒトゲノムデータの取り扱い

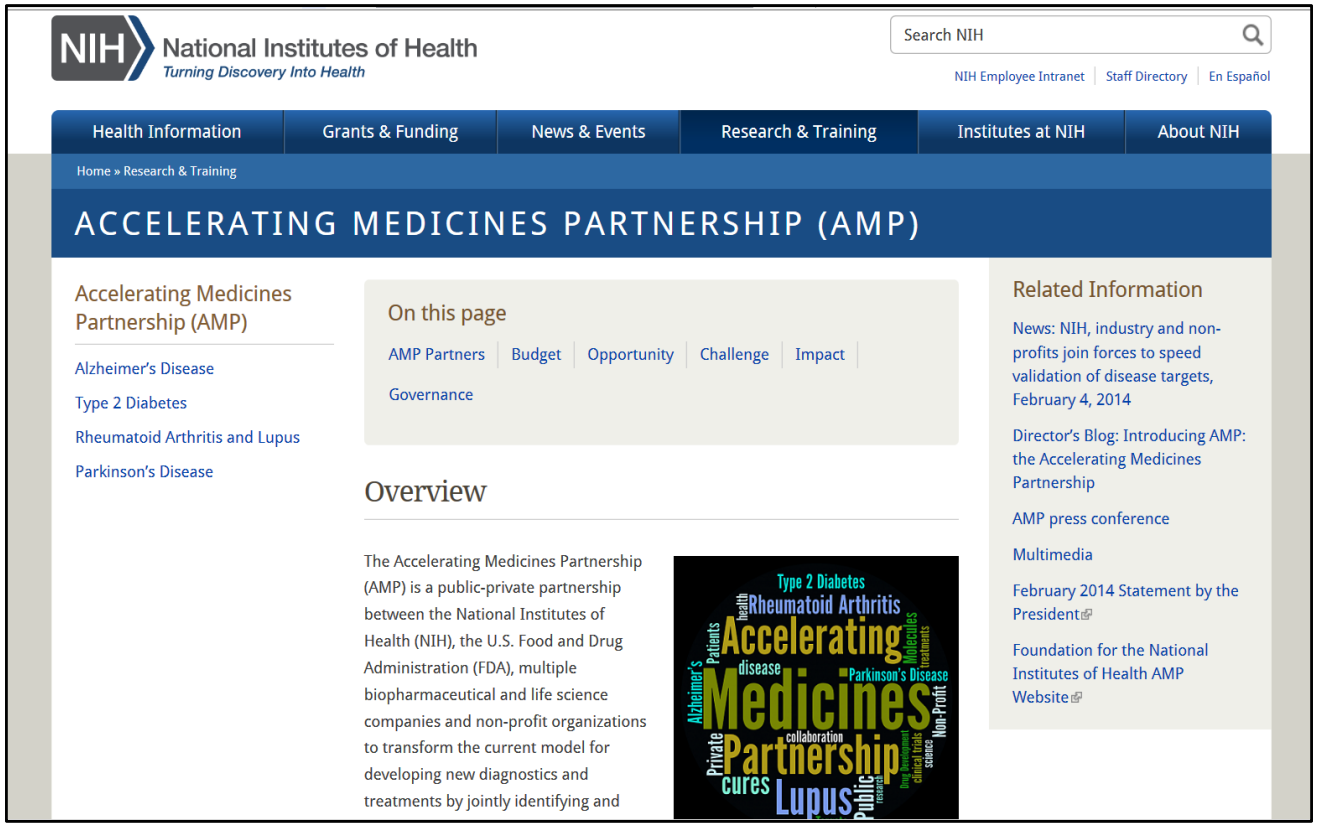



https://www.hipsci.org/

- 欧米人集団数百名からiPS細胞を樹立する、HipSci（Human Induced Pluripotent Stem Cell Initiative）プロジェクトが公開されました。
- iPS細胞由来の全ゲノム・エピゲノム情報は一般公開され、iPS細胞株も分譲手続きを経て入手することができます。

（Kilpinen H et al. ***Nature*** 2017）

# ① ヒトゲノムデータの取り扱い




https://www.nih.gov/research-training/accelerating-medicines-partnership-amp

- アルツハイマー病、糖尿病、自己免疫疾患を対象とした米国の Accelerating Medicines Partnership (AMP) では、エピゲノム・ゲノムデータを一般公開しています。メールアドレス登録で入手可能です。
- シングルセルRNA-seqなど、最新のオミクス情報が公開されています。

# 講義の概要

## GenomeDataAnalysis1

① ヒトゲノムデータの取り扱い

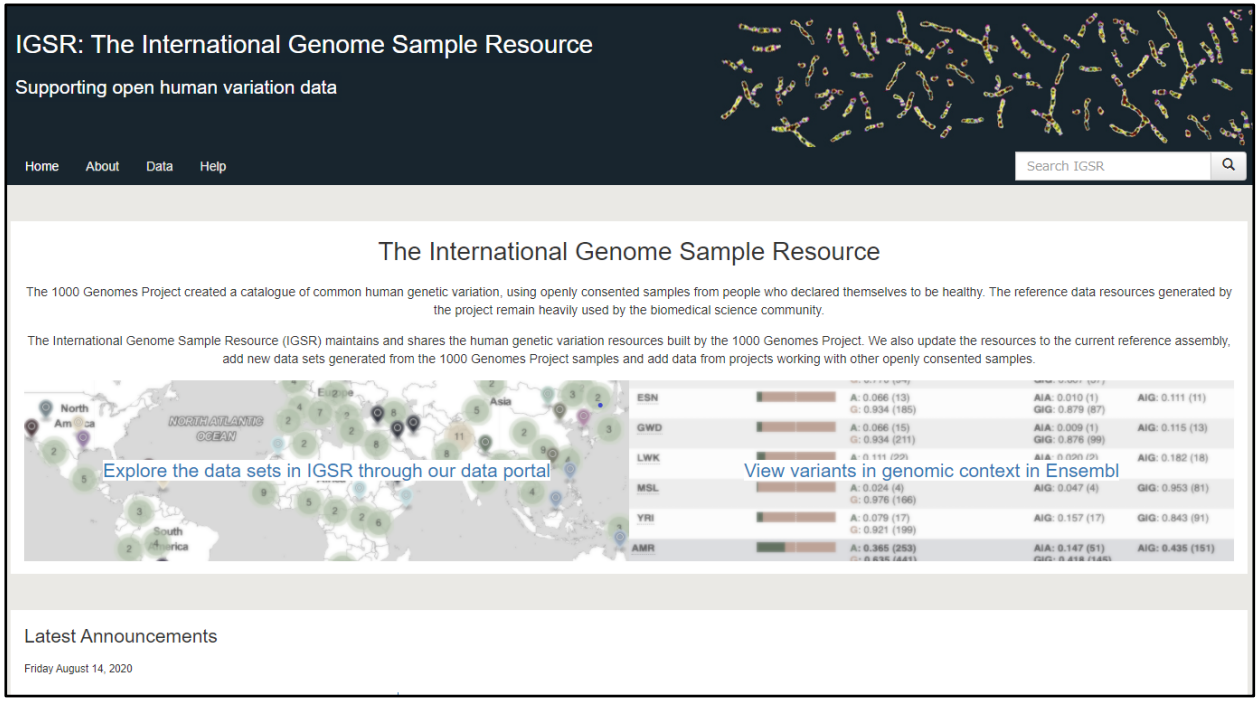
② 1000 Genomes Projectデータ

③ 遺伝統計解析ソフトPLINK実習

# ② 1000 Genomes Projectデータ

## 1000 Genomes Project

http://www.internationalgenome.org/




- 1000 Genomes Projectでは、次世代シークエンサー(NGS)を用いて、多数の人類集団の全ゲノムシークエンス結果を公開しています。
- 2017年に公開されたPhase 3では、2500人、8400万SNPのジェノタイプデータが得られています。

# ② 1000 Genomes Projectデータ

## NGS出力ファイル
## （リード情報、FASTQ形式）

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:N:18:1
```

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ
[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

## ゲノムデータ

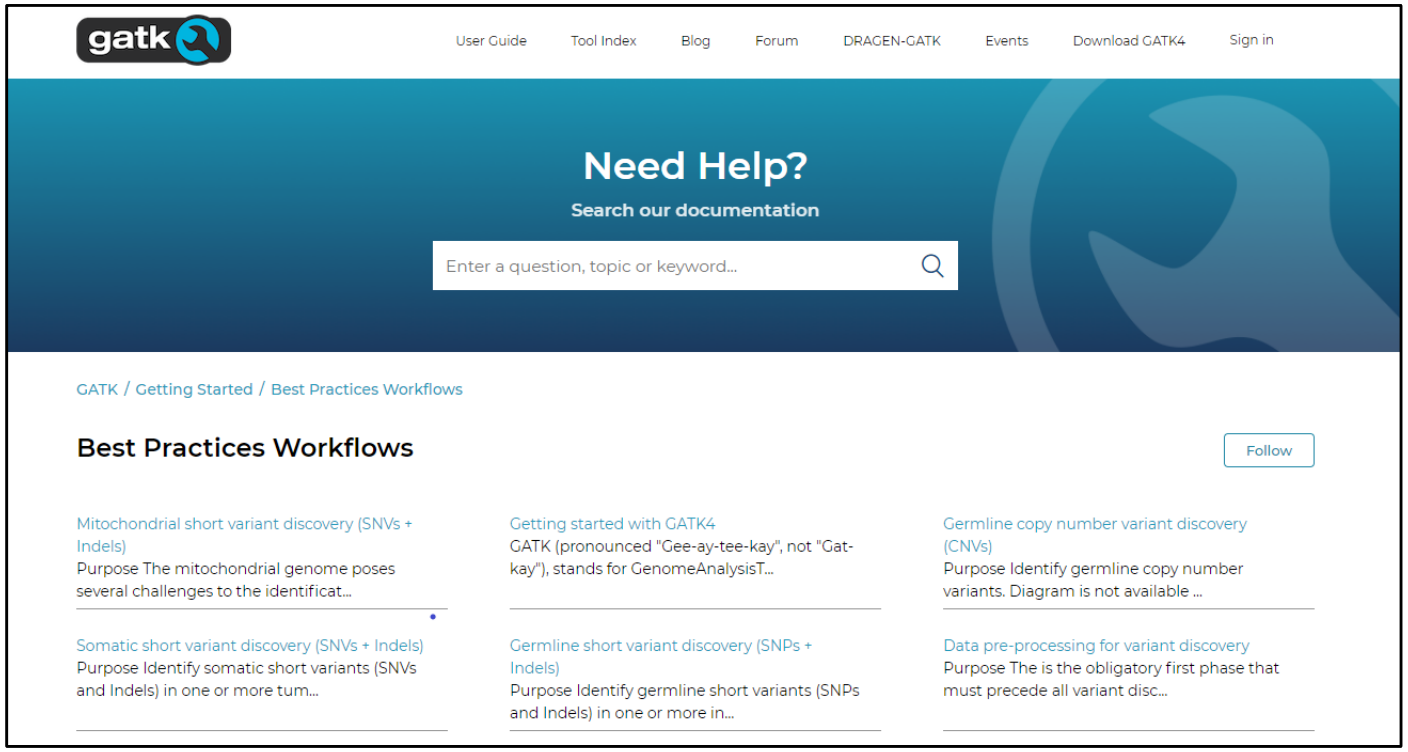
| | SNP1 | SNP2 | SNP3 | SNP4 |
|---|---|---|---|---|
| Sample1 | 1 | 0 | 1 | 1 |
| Sample2 | 1 | 2 | 1 | 2 |
| Sample3 | 0 | 1 | 1 | 1 |
| Sample4 | 0 | 2 | 0 | 2 |
| Sample5 | 1 | 1 | 1 | 0 |
| Sample6 | 1 | 1 | 1 | 1 |

- NGSの出力ファイル(リード情報、FASTQファイル)から、個人のヒトゲノム配列情報を取得するには、**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。

# ② 1000 Genomes Projectデータ

## GATK Best Practices

https://software.broadinstitute.org/gatk/best-practices/




- ゲノム配列を構築するためのNGSデータ解析ソフトとしては、米国Broad研究所が開発したGATKが有名です。
- 推奨パイプラインが、GATK Best Practicesとして公開されています。
- 本講義では、解析後に得られたヒトゲノム配列データを使うことにします。

# ② 1000 Genomes Projectデータ

## 1000 Genomes Project

http://www.internationalgenome.org/

FTP ディレクトリ /vol1/ftp/release/20130502/ / ftp.1000genomes.ebi.ac.uk

エクスプローラーでこの FTP サイトを表示するには、Alt キーを押して、[表示]をクリックし、[エクスプローラーで FTP サイトを開く]をクリックしてください。

1 階層上のディレクトリへ

| | | |
|---|---|---|
| 06/25/2014 12:00午前 | 1,168 | 20140625_related_individuals.txt |
| 05/27/2015 12:00午前 | 1,216,886,729 | ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 224,660 | ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 773,788,987 | ALL.chr10.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 133,086 | ALL.chr10.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 767,084,423 | ALL.chr11.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 133,331 | ALL.chr11.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 740,805,962 | ALL.chr12.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 133,171 | ALL.chr12.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 556,832,848 | ALL.chr13.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 96,652 | ALL.chr13.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 506,573,037 | ALL.chr14.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 88,321 | ALL.chr14.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 457,900,567 | ALL.chr15.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 81,584 | ALL.chr15.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 495,134,005 | ALL.chr16.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 81,026 | ALL.chr16.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 434,623,611 | ALL.chr17.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 79,187 | ALL.chr17.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 436,425,683 | ALL.chr18.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 74,610 | ALL.chr18.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 359,519,603 | ALL.chr19.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 55,477 | ALL.chr19.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 1,312,735,578 | ALL.chr2.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 238,011 | ALL.chr2.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 341,680,844 | ALL.chr20.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 56,516 | ALL.chr20.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 218,612,970 | ALL.chr21.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 35,357 | ALL.chr21.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 214,453,750 | ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 36,078 | ALL.chr22.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 1,105,776,520 | ALL.chr3.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 196,236 | ALL.chr3.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 1,117,054,318 | ALL.chr4.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 188,993 | ALL.chr4.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |
| 05/27/2015 12:00午前 | 989,263,345 | ALL.chr5.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz |
| 05/27/2015 12:00午前 | 176,777 | ALL.chr5.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi |

- "Data"→Download data from the IGSR FTP site項目の"FTP site"をクリックすると、ダウンロード可能なジェノタイプデータが掲載されたFTPサイトに飛ぶことができます。（FTPサイトとは、FTPプロトコルを用いてファイルのアップロード/ダウンロードを行うことができるサーバーのことです。）

# ② 1000 Genomes Projectデータ

```
statgen@statgen-PC: /mnt/c
$ cd /mnt/c/SummerSchool/GenomeDataAnalysis1
```

※Cygwinの場合 /mnt/を/cygdrive/に変えてください
※Macの場合 資料を置いたフォルダを指定してください

```
statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis1
$ wget
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr1.phase3_shapeit2_mvncall_integrated_v5b.20130502.genotypes.vcf.gz
--2016-08-23 21:26:31--
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz
        => `ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz'
ftp.1000genomes.ebi.ac.uk (ftp.1000genomes.ebi.ac.uk) をDNSに問いあわせています...
193.62.192.8
ftp.1000genomes.ebi.ac.uk (ftp.1000genomes.ebi.ac.uk)|193.62.192.8|:21 に接続しています... 接続しました。
anonymous としてログインしています... ログインしました!
==> SYST ... 完了しました。    ==> PWD ... 完了しました。
==> TYPE I ... 完了しました。  ==> CWD (1) /vol1/ftp/release/20130502 ... 完了しました。
==> SIZE ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz ...
```

- **個々のジェノタイプデータは、Linuxに実装されたwgetコマンドでダウンロードできますが、時間がかかるので、本講義では既にダウンロード済のデータを使います。**

# ② 1000 Genomes Projectデータ

## 1 The VCF specification

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

### 1.1 An example

付帯情報の説明

各行が各SNPに対応

各サンプル

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G    A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.
20     17330   .         T    A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T    .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC  G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

- NGSジェノタイプデータは、主に”vcfファイル形式”で保存されます。
- vcf形式は、単にサンプル毎のジェノタイプ情報だけでなく、SNP毎の多彩な付帯情報を含めることができますが、やや複雑です。

# ② 1000 Genomes Projectデータ

example.ped

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

example.map

| | | | |
|---|---|---|---|
| 1 | SNP1 | 0 | 10000 |
| 1 | SNP2 | 0 | 20000 |
| 1 | SNP3 | 0 | 30000 |

第１列：Family ID
第２列：Sample ID
第３列：Paternal ID ・・・ 使用しない(=0)
第４列：Maternal ID ・・・ 使用しない(=0)
第５列：Sex ・・・ 男性=1、女性=2、不明=0
第６列：Phenotype ・・・ ケース=2、コントロール=1、不明=0
第７列以降：各SNPのジェノタイプデータ

第１列：各SNPの染色体番号
第２列：各SNPの名称
第３列：使用しない(=0)
第４列：各SNPの染色体上の位置

- 本講義ではシンプルな”pedファイル形式”に変換したデータを扱います。
- 併せて、SNP情報が記載された”mapファイル”を扱います。
- ”ped”は”pedigree”の略で、家系例を対象とした連鎖解析の際に、家系情報を表現するデータ形式として用いられていました。

# ② 1000 Genomes Projectデータ

バイナリ形式のジェノタイプデータ

SNP情報ファイル

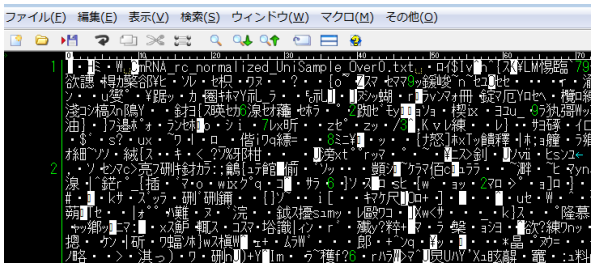
サンプル情報ファイル

- example.ped
- example.map

→

- example.bed
- example.bim
- example.fam

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

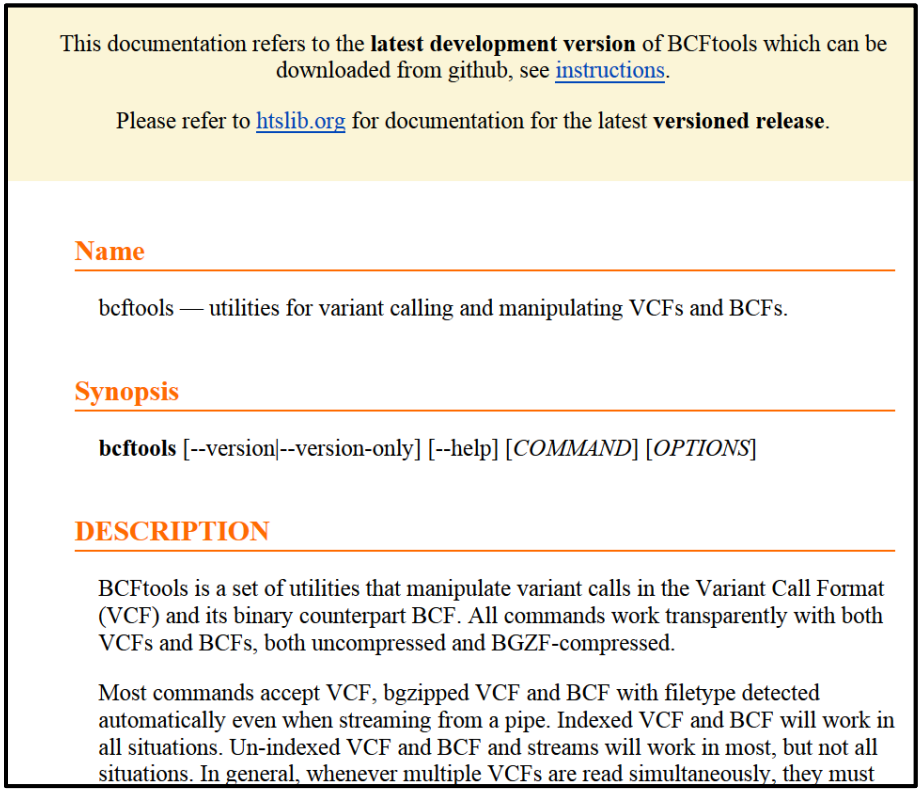



※ファイル"example.ped/map/bed/bim/fam"をエディタで開いてみて、中身を確認してみて下さい。

- ディスク容量節約もかねて、ped/map形式から、ジェノタイプ情報をバイナリ形式で保存した、bed/bim/fam形式に変換することがあります。

（バイナリ形式なのはbedファイルだけで、bim/famファイルはテキスト形式です。）

# ② 1000 Genomes Projectデータ



This documentation refers to the **latest development version** of BCFtools which can be downloaded from github, see instructions.

Please refer to htslib.org for documentation for the latest **versioned release**.

## Name

bcftools — utilities for variant calling and manipulating VCFs and BCFs.

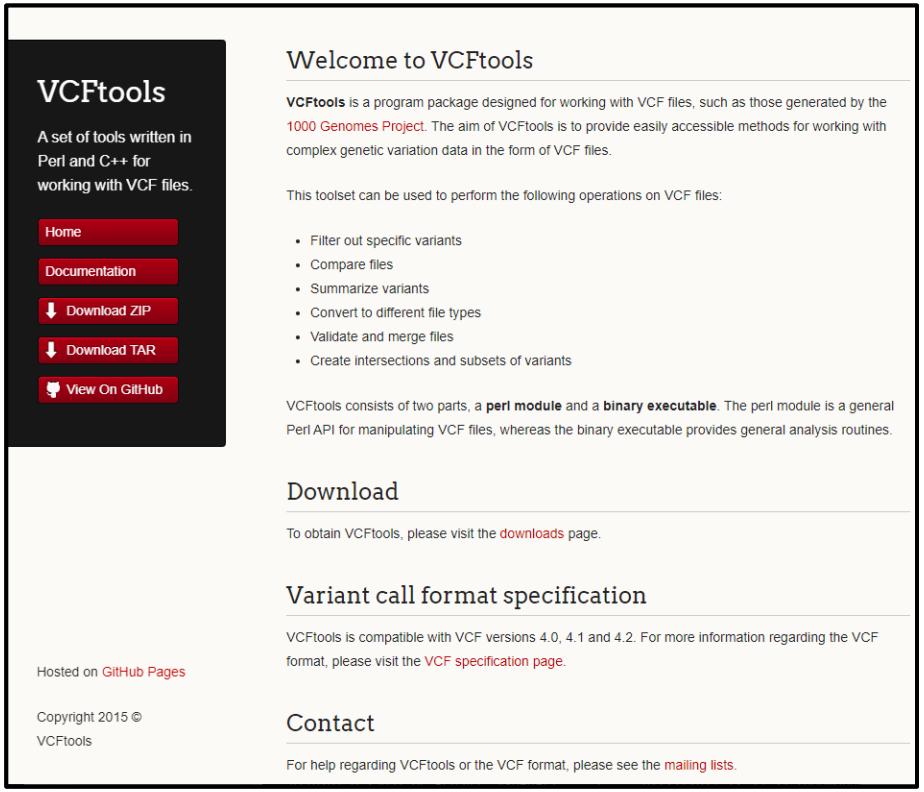
## Synopsis

**bcftools** [--version|--version-only] [--help] [*COMMAND*] [*OPTIONS*]

## DESCRIPTION

BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF. All commands work transparently with both VCFs and BCFs, both uncompressed and BGZF-compressed.

Most commands accept VCF, bgzipped VCF and BCF with filetype detected automatically even when streaming from a pipe. Indexed VCF and BCF will work in all situations. Un-indexed VCF and BCF and streams will work in most, but not all situations. In general, whenever multiple VCFs are read simultaneously, they must



VCFtools

A set of tools written in Perl and C++ for working with VCF files.

Home

Documentation

Download ZIP

Download TAR

View On GitHub

Hosted on GitHub Pages

Copyright 2015 © VCFtools

## Welcome to VCFtools

**VCFtools** is a program package designed for working with VCF files, such as those generated by the 1000 Genomes Project. The aim of VCFtools is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files.

This toolset can be used to perform the following operations on VCF files:

- Filter out specific variants
- Compare files
- Summarize variants
- Convert to different file types
- Validate and merge files
- Create intersections and subsets of variants

VCFtools consists of two parts, a **perl module** and a **binary executable**. The perl module is a general Perl API for manipulating VCF files, whereas the binary executable provides general analysis routines.

## Download

To obtain VCFtools, please visit the downloads page.

## Variant call format specification

VCFtools is compatible with VCF versions 4.0, 4.1 and 4.2. For more information regarding the VCF format, please visit the VCF specification page.

## Contact

For help regarding VCFtools or the VCF format, please see the mailing lists.

https://samtools.github.io/bcftools/bcftools.html

https://vcftools.github.io/index.html

- vcfファイル形式からpedファイル形式への変換については、幾つかのツールやソフトウェア上で実装されています。
- "bcftools"や"vcftools"などの、ソフトウェアが知られています。

# ② 1000 Genomes Projectデータ

```
statgen@statgen-PC: ~
$ cd /mnt/c/SummerSchool/GenomeDataAnalysis1/1KG_EUR/

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis1/1KG_EUR
$ ls
1KG_EUR.bed  1KG_EUR.bim  1KG_EUR.fam  1KG_EUR_Sample.xlsx

statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis1/1KG_EUR
$ wc 1KG_EUR.*
   1392295    2962444  847697763 1KG_EUR.bed
   8830185   52981110  257807275 1KG_EUR.bim
       381       2286       9144 1KG_EUR.fam
  10222861   55945840 1105514182 合計
```

bimファイルが8,830,185行
→8,830,185SNPのデータ

famファイルが381行
→381サンプルのデータ

※Cygwinの場合 /mnt/を/cygdrive/に変えてください
※Macの場合 資料を置いたフォルダを指定してください

- 今回の講義では、1000 Genomes Projectサイトからダウンロードしたジェノタイプデータを使います(”1KG_EUR.bed/bim/fam” )。
- Phase Ⅰ(α)という少し古いバージョンの、欧米人集団のデータです。
- Linuxコマンド”wc”で、サンプル数やSNP数を確かめてみましょう。

# ② 1000 Genomes Projectデータ

| | |
|---|---|
| • pwd | 自分が作業しているディレクトリを表示します。 |
| • cd /mnt/c/ | 他のディレクトリに移動します。 |
| • ls | ディレクトリの中のファイル一覧を表示します。 |
| • ls -la | ディレクトリの中のファイル一覧を詳細に表示します。 |
| • echo “test” | コンソールに文字を表示します。 |
| • echo “test” > test.txt | コンソールの内容をファイルに書き込みます。 |
| • cp test.txt test_cp.txt | ファイルをコピーします。 |
| • rm test_cp.txt | ファイルを削除します。 |
| • cat test.txt | ファイルの内容を表示します。 |
| • cat test.txt test.txt | ファイルを縦に繋いだ内容を表示します。 |
| • paste test.txt test.txt | ファイルを横に繋いだ内容を表示します。 |

- Linuxコマンドのおさらいです。
- Linux(Shell)で練習してみましょう。

# ② 1000 Genomes Projectデータ

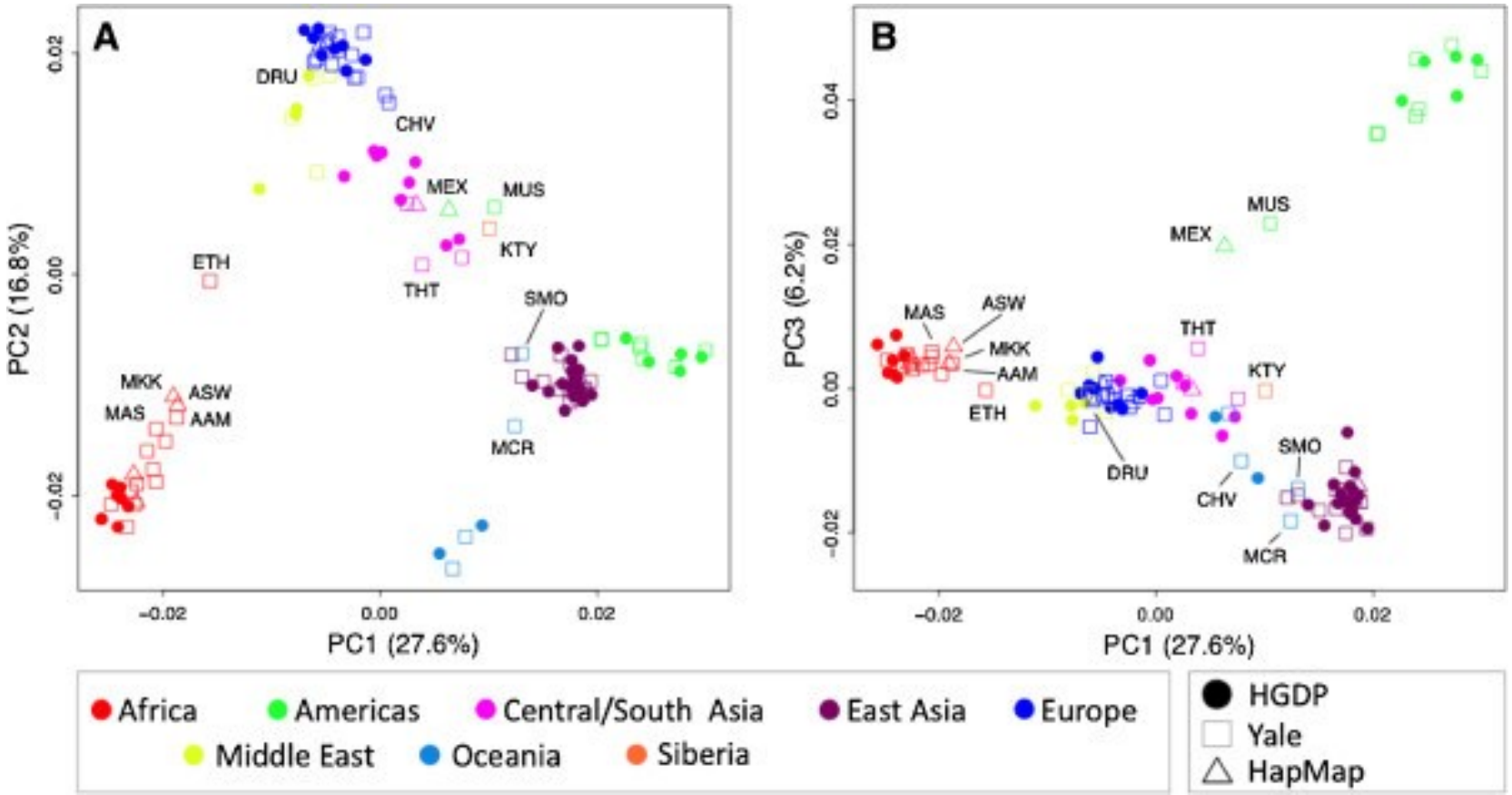
T21

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | |
| 2 | | Family ID | Individual ID | Paternal ID | Maternal ID | Sex | Disease | Population | | Population | No.samples | Description of the population |
| 3 | | HG00096 | HG00096 | 0 | 0 | 1 | 1 | GBR | | GBR | 89 | British individuals from England and Scotland |
| 4 | | HG00097 | HG00097 | 0 | 0 | 2 | 1 | GBR | | FIN | 93 | HapMap Finnish individuals from Finland |
| 5 | | HG00099 | HG00099 | 0 | 0 | 2 | 1 | GBR | | IBS | 14 | Iberian populations in Spain |
| 6 | | HG00100 | HG00100 | 0 | 0 | 2 | 1 | GBR | | CEU | 87 | CEPH individuals |
| 7 | | HG00101 | HG00101 | 0 | 0 | 1 | 1 | GBR | | TSI | 98 | Toscan individuals |
| 8 | | HG00102 | HG00102 | 0 | 0 | 2 | 1 | GBR | | Total | 381 | - |
| 9 | | HG00103 | HG00103 | 0 | 0 | 1 | 1 | GBR | | | | |
| 10 | | HG00104 | HG00104 | 0 | 0 | 2 | 1 | GBR | | | | |
| 11 | | HG00106 | HG00106 | 0 | 0 | 2 | 1 | GBR | | | | |
| 12 | | HG00108 | HG00108 | 0 | 0 | 1 | 1 | GBR | | | | |
| 13 | | HG00109 | HG00109 | 0 | 0 | 1 | 1 | GBR | | | | |
| 14 | | HG00110 | HG00110 | 0 | 0 | 2 | 1 | GBR | | | | |
| 15 | | HG00111 | HG00111 | 0 | 0 | 2 | 1 | GBR | | | | |
| 16 | | HG00112 | HG00112 | 0 | 0 | 1 | 1 | GBR | | | | |
| 17 | | HG00113 | HG00113 | 0 | 0 | 1 | 1 | GBR | | | | |
| 18 | | HG00114 | HG00114 | 0 | 0 | 1 | 1 | GBR | | | | |
| 19 | | HG00116 | HG00116 | 0 | 0 | 1 | 1 | GBR | | | | |
| 20 | | HG00117 | HG00117 | 0 | 0 | 1 | 1 | GBR | | | | |

- GBR：イングランドおよびスコットランド
- FIN：フィンランド
- IBS：スペインのイベリア半島
- CEU：アメリカのユタ州
- TSI：イタリアのトスカーナ地方

- 381サンプルの内訳は、”1KG_EUR_Sample.xlsx”に記載されています。
- ”GBR”、”FIN”、”IBS”、”CEU”、”TSI”という、5つの地域の住民で構成されていることがわかります。

# ② 1000 Genomes Projectデータ




- 異なる地域から集められたのは、「同地域の住民は遺伝的に近い」、「離れた地域や異なる集団は遺伝的に遠い」、という事情からです。
- ヒトゲノム変異のカタログを効率的に収集するためには、多彩な人類集団から幅広くゲノムデータを収集することが重要です。



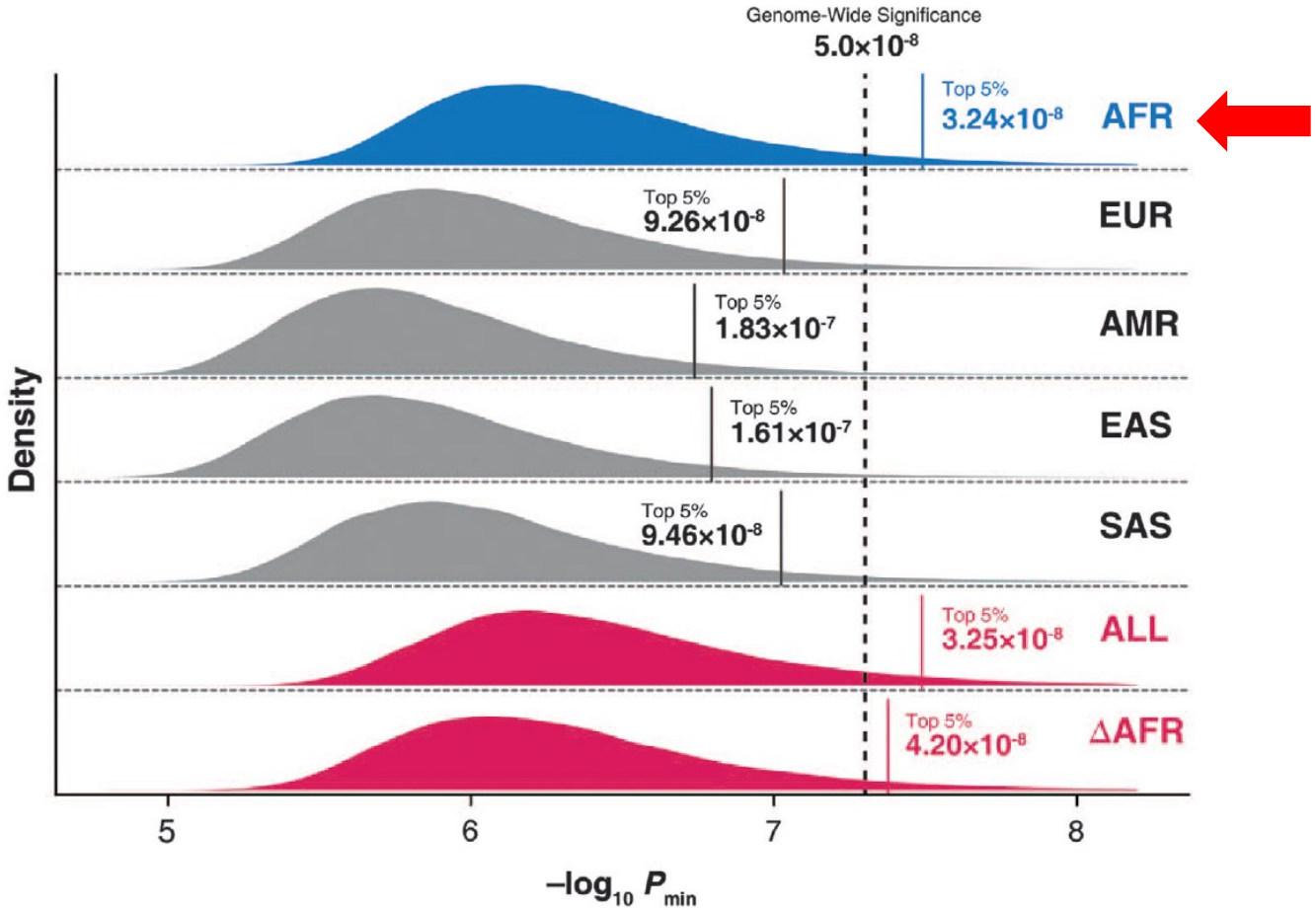
(Nievergelt CM et al.***Investig Genet*** 2013)

# ② 1000 Genomes Projectデータ

| Population | Sub-population | Code | Male | Female | Total |
|---|---|---|---|---|---|
| AFR | African Caribbeans in Barbados | ACB | 47 | 49 | 96 |
| | Americans of African Ancestry in SW USA | ASW | 26 | 35 | 61 |
| | Esan in Nigeria | ESN | 53 | 46 | 99 |
| | Gambian in Western Divisions in the Gambia | GWD | 55 | 58 | 113 |
| | Luhya in Webuye, Kenya | LWK | 44 | 55 | 99 |
| | Mende in Sierra Leone | MSL | 42 | 43 | 85 |
| | Yoruba in Ibadan, Nigeria | YRI | 52 | 56 | 108 |
| | Sub-Total | - | 319 | 342 | 661 |
| EUR | Utah Residents (CEPH) with Northern and Western European Ancestry | CEU | 49 | 50 | 99 |
| | Finnish in Finland | FIN | 38 | 61 | 99 |
| | British in England and Scotland | GBR | 46 | 45 | 91 |
| | Iberian Population in Spain | IBS | 54 | 53 | 107 |
| | Toscani in Italia | TSI | 53 | 54 | 107 |
| | Sub-Total | - | 240 | 263 | 503 |
| AMR | Colombians from Medellin, Colombia | CLM | 43 | 51 | 94 |
| | Mexican Ancestry from Los Angeles USA | MXL | 32 | 32 | 64 |
| | Peruvians from Lima, Peru | PEL | 41 | 44 | 85 |
| | Puerto Ricans from Puerto Rico | PUR | 54 | 50 | 104 |
| | Sub-Total | - | 170 | 177 | 347 |
| EAS | Chinese Dai in Xishuangbanna, China | CDX | 44 | 49 | 93 |
| | Han Chinese in Beijing, China | CHB | 46 | 57 | 103 |
| | Southern Han Chinese | CHS | 52 | 53 | 105 |
| | Japanese in Tokyo, Japan | JPT | 56 | 48 | 104 |
| | Kinh in Ho Chi Minh City, Vietnam | KHV | 46 | 53 | 99 |
| | Sub-Total | - | 244 | 260 | 504 |
| SAS | Bengali from Bangladesh | BEB | 42 | 44 | 86 |
| | Gujarati Indian from Houston, Texas | GIH | 56 | 47 | 103 |
| | Indian Telugu from the UK | ITU | 59 | 43 | 102 |
| | Punjabi from Lahore, Pakistan | PJL | 48 | 48 | 96 |
| | Sri Lankan Tamil from the UK | STU | 55 | 47 | 102 |
| | Sub-Total | - | 260 | 229 | 489 |
| Total | - | - | 1,233 | 1,271 | 2,504 |

- **最新のPhase 3では5集団（26地域）2,504名が対象です（JPT=104）。**

# ② 1000 Genomes Projectデータ




- **集団によって、存在するSNPの組成（数、種類、頻度分布）が異なる**ことが知られています。
- 人類の中で長い歴史を持つアフリカ人集団は、一番多くの（独立した）SNPを保有しています。

(Kanai M et al. ***J Hum Genet*** 2016)

# 講義の概要

## GenomeDataAnalysis1

① ヒトゲノムデータの取り扱い

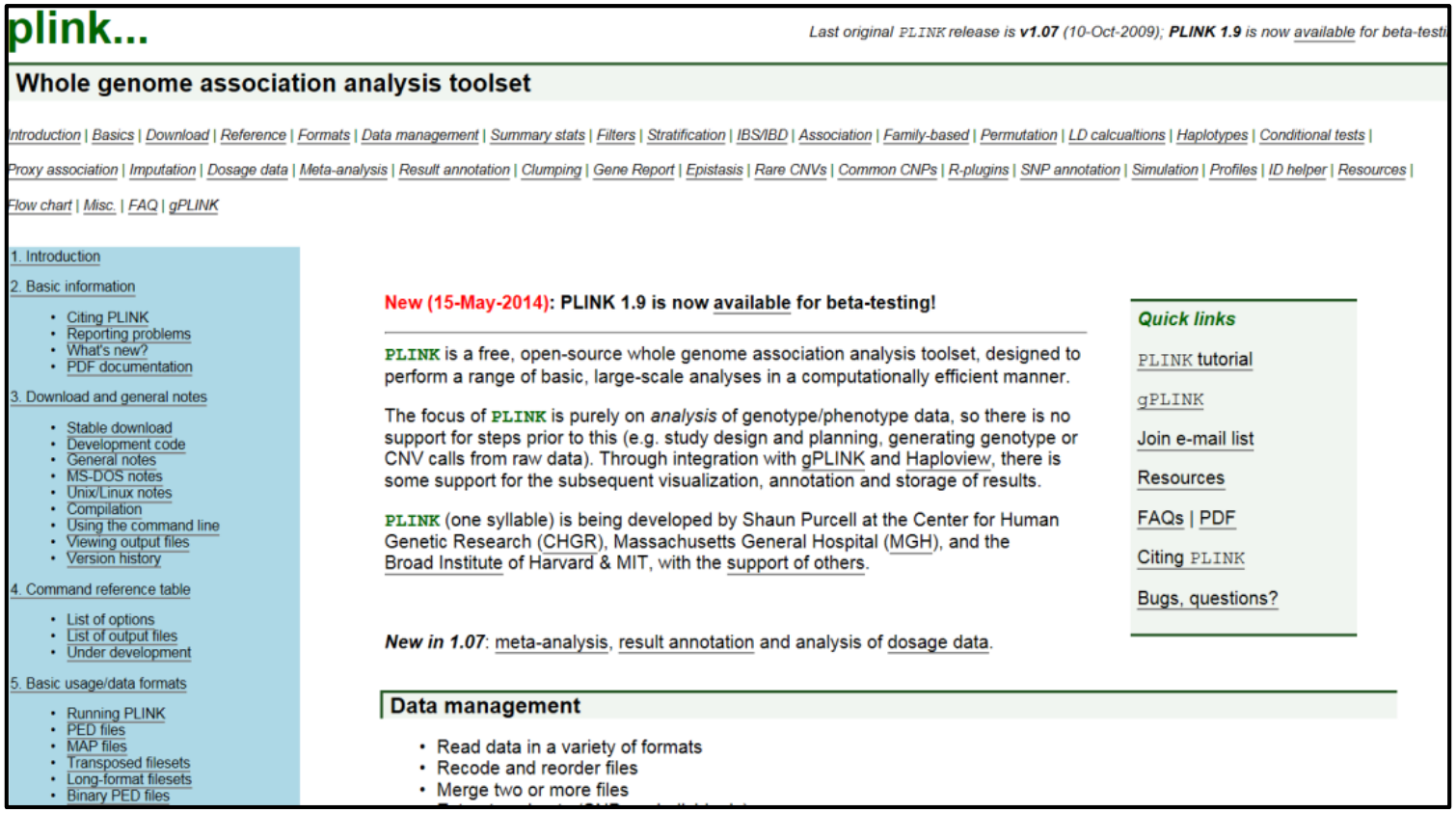
② 1000 Genomes Projectデータ

③ 遺伝統計解析ソフトPLINK実習

# ③ 遺伝統計解析ソフトPLINK実習

## PLINK

http://zzz.bwh.harvard.edu/plink/

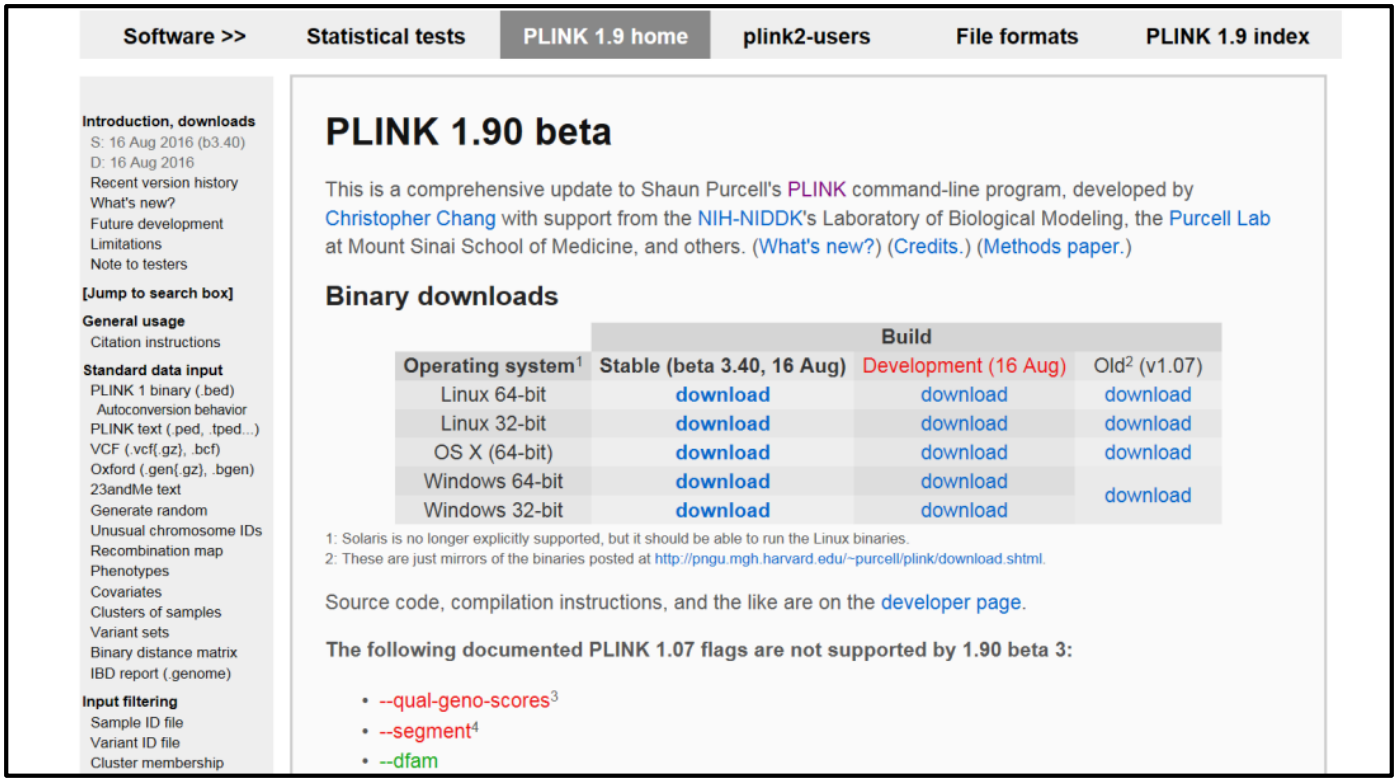



- 1000 Genomes Projectゲノムデータに対して、PLINK(プリンク)という遺伝統計解析ソフトを使って解析を行いましょう。
- PLINKはSNPデータ解析ツールで、米国ハーバード大で開発されました。

# ③ 遺伝統計解析ソフトPLINK実習

## PLINK v1.90

https://www.cog-genomics.org/plink2



Software >> | Statistical tests | PLINK 1.9 home | plink2-users | File formats | PLINK 1.9 index

Introduction, downloads
S: 16 Aug 2016 (b3.40)
D: 16 Aug 2016
Recent version history
What's new?
Future development
Limitations
Note to testers
[Jump to search box]
General usage
Citation instructions
Standard data input
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf{.gz}, .bcf)
Oxford (.gen{.gz}, .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix
IBD report (.genome)
Input filtering
Sample ID file
Variant ID file
Cluster membership

### PLINK 1.90 beta

This is a comprehensive update to Shaun Purcell's PLINK command-line program, developed by Christopher Chang with support from the NIH-NIDDK's Laboratory of Biological Modeling, the Purcell Lab at Mount Sinai School of Medicine, and others. (What's new?) (Credits.) (Methods paper.)

### Binary downloads

| | Build | | |
|---|---|---|---|
| Operating system[1] | Stable (beta 3.40, 16 Aug) | Development (16 Aug) | Old[2] (v1.07) |
| Linux 64-bit | download | download | download |
| Linux 32-bit | download | download | download |
| OS X (64-bit) | download | download | download |
| Windows 64-bit | download | download | download |
| Windows 32-bit | download | download | |

1: Solaris is no longer explicitly supported, but it should be able to run the Linux binaries.
2: These are just mirrors of the binaries posted at http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml.

Source code, compilation instructions, and the like are on the developer page.

The following documented PLINK 1.07 flags are not supported by 1.90 beta 3:

- --qual-geno-scores[3]
- --segment[4]
- --dfam

・バージョン1.90から開発主体が移行し、速度も劇的に速くなりました。

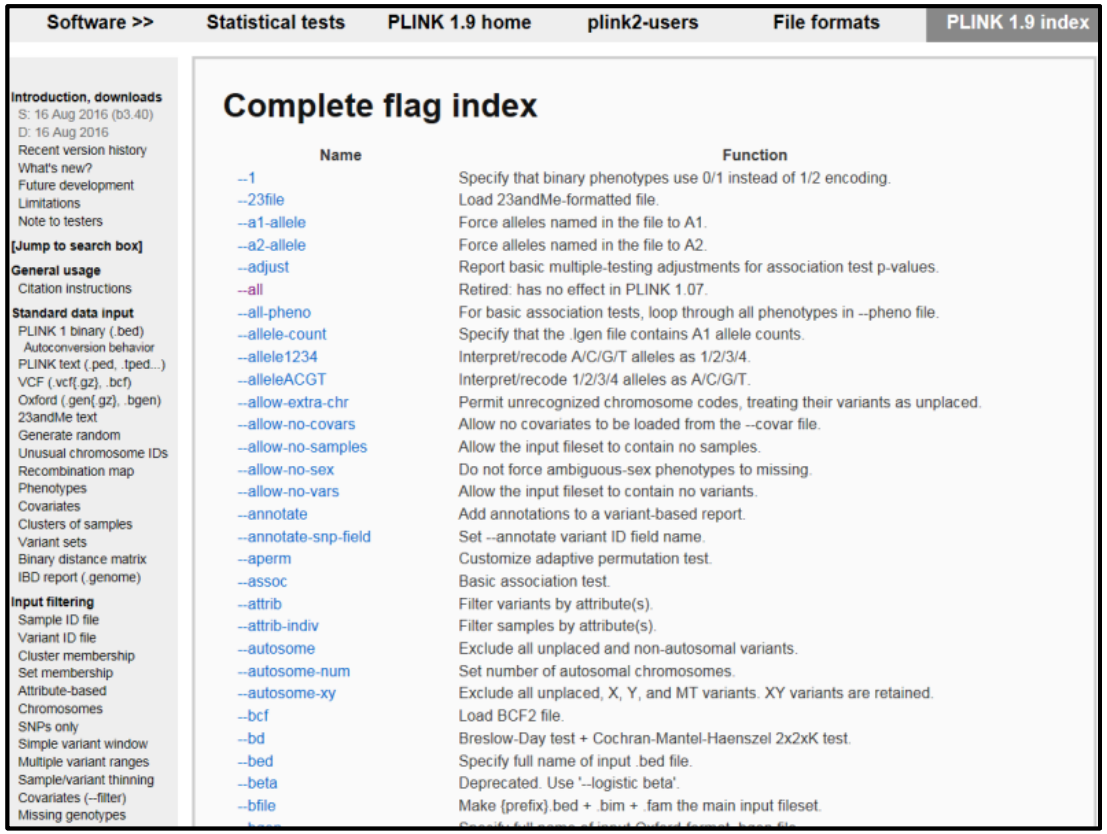
・今回は、このv1.90を使います。

・WSL2用に、”Linux 64-bit, Stable”版(”plink”)がダウンロード済です。
（Cygwinの場合、”Windows 64-bit”版(”plink.exe”)を使用してください。）

# ③ 遺伝統計解析ソフトPLINK実習

## PLINK v1.90 index

https://www.cog-genomics.org/plink2/index



Software >> | Statistical tests | PLINK 1.9 home | plink2-users | File formats | PLINK 1.9 index

**Introduction, downloads**
S: 16 Aug 2016 (b3.40)
D: 16 Aug 2016
Recent version history
What's new?
Future development
Limitations
Note to testers
**[Jump to search box]**
**General usage**
Citation instructions
**Standard data input**
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf{.gz}, .bcf)
Oxford (.gen{.gz}, .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix
IBD report (.genome)
**Input filtering**
Sample ID file
Variant ID file
Cluster membership
Set membership
Attribute-based
Chromosomes
SNPs only
Simple variant window
Multiple variant ranges
Sample/variant thinning
Covariates (--filter)
Missing genotypes

### Complete flag index

| Name | Function |
|---|---|
| --1 | Specify that binary phenotypes use 0/1 instead of 1/2 encoding. |
| --23file | Load 23andMe-formatted file. |
| --a1-allele | Force alleles named in the file to A1. |
| --a2-allele | Force alleles named in the file to A2. |
| --adjust | Report basic multiple-testing adjustments for association test p-values. |
| --all | Retired: has no effect in PLINK 1.07. |
| --all-pheno | For basic association tests, loop through all phenotypes in --pheno file. |
| --allele-count | Specify that the .lgen file contains A1 allele counts. |
| --allele1234 | Interpret/recode A/C/G/T alleles as 1/2/3/4. |
| --alleleACGT | Interpret/recode 1/2/3/4 alleles as A/C/G/T. |
| --allow-extra-chr | Permit unrecognized chromosome codes, treating their variants as unplaced. |
| --allow-no-covars | Allow no covariates to be loaded from the --covar file. |
| --allow-no-samples | Allow the input fileset to contain no samples. |
| --allow-no-sex | Do not force ambiguous-sex phenotypes to missing. |
| --allow-no-vars | Allow the input fileset to contain no variants. |
| --annotate | Add annotations to a variant-based report. |
| --annotate-snp-field | Set --annotate variant ID field name. |
| --aperm | Customize adaptive permutation test. |
| --assoc | Basic association test. |
| --attrib | Filter variants by attribute(s). |
| --attrib-indiv | Filter samples by attribute(s). |
| --autosome | Exclude all unplaced and non-autosomal variants. |
| --autosome-num | Set number of autosomal chromosomes. |
| --autosome-xy | Exclude all unplaced, X, Y, and MT variants. XY variants are retained. |
| --bcf | Load BCF2 file. |
| --bd | Breslow-Day test + Cochran-Mantel-Haenszel 2x2xK test. |
| --bed | Specify full name of input .bed file. |
| --beta | Deprecated. Use '--logistic beta'. |
| --bfile | Make {prefix}.bed + .bim + .fam the main input fileset. |

- PLINKには多くの機能があり、使い方の一覧が説明されています。
- 今回は、代表的な機能について、説明します。

# ③ 遺伝統計解析ソフトPLINK実習

```
statgen@statgen-PC: ~
$ cd /mnt/c/SummerSchool/GenomeDataAnalysis1/1KG_EUR/
```

※Cygwinの場合 /mnt/を/cygdrive/に変えてください。

```
statgen@statgen-PC: /mnt/c/SummerSchool/GenomeDataAnalysis1/1KG_EUR
$ ./plink
```

※Cygwinの場合plinkをplink.exeに変えてください。

```
PLINK v1.90b3.40 64-bit (16 Aug 2016)     https://www.cog-genomics.org/plink2
(C) 2005-2016 Shaun Purcell, Christopher Chang   GNU General Public License v3

  plink [input flag(s)...] {command flag(s)...} {other flag(s)...}
  plink --help {flag name(s)...}

Commands include --make-bed, --recode, --flip-scan, --merge-list,
--write-snplist, --list-duplicate-vars, --freqx, --missing, --test-mishap,
                （中略）
'plink --help | more' describes all functions (warning: long).
```

**※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。**

- **コンソール画面で”./plink”と入力すると、実行されます。**
- **”./ファイル名”は、ファイルをLinux上で直接実行する方法です。**

# ③ 遺伝統計解析ソフトPLINK実習

**※ファイル”PLINK_Command_1.txt”を開いて、内容をShellにコピー＆ペーストして下さい。**

※Cygwinの場合は
”PLINK_Command_1_cygwin.txt”を使用してください

**○: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

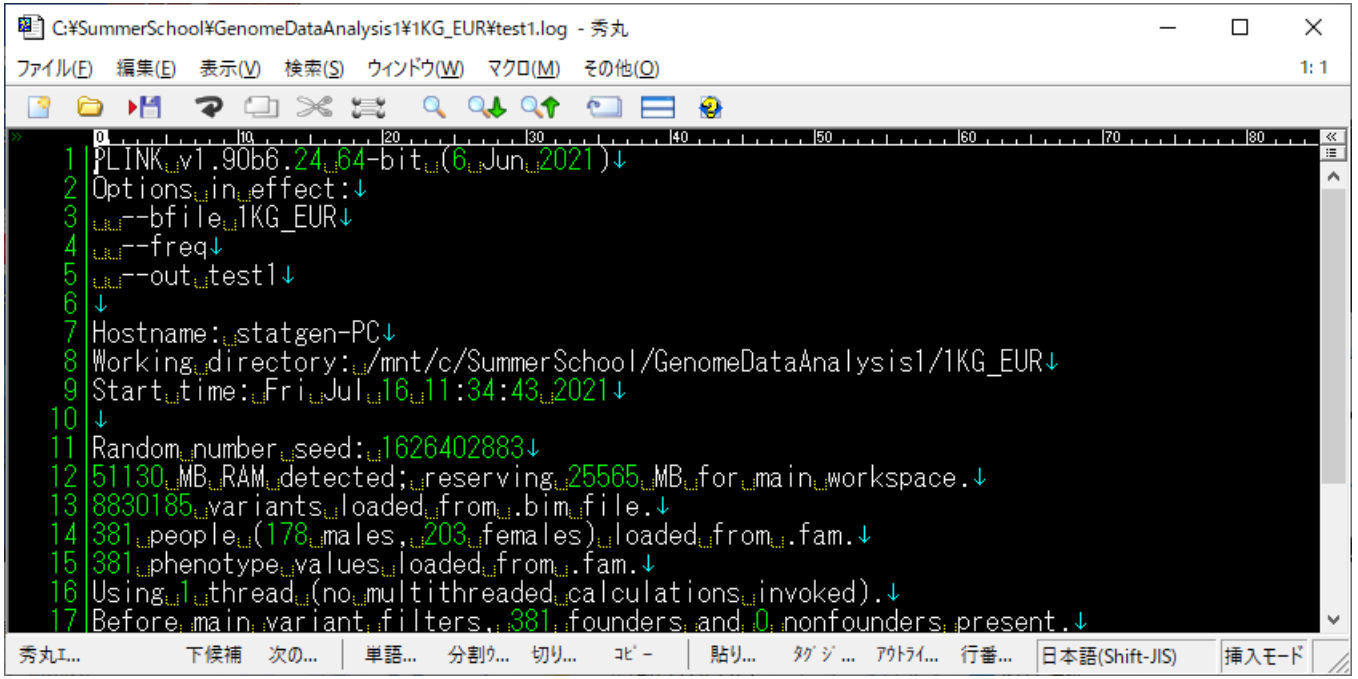
**・PLINKは”./plink --(コマンド)(引数)”という形で実行します。**

# ③ 遺伝統計解析ソフトPLINK実習

## ○:ファイルの読み込み

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

出力ファイル:test.log



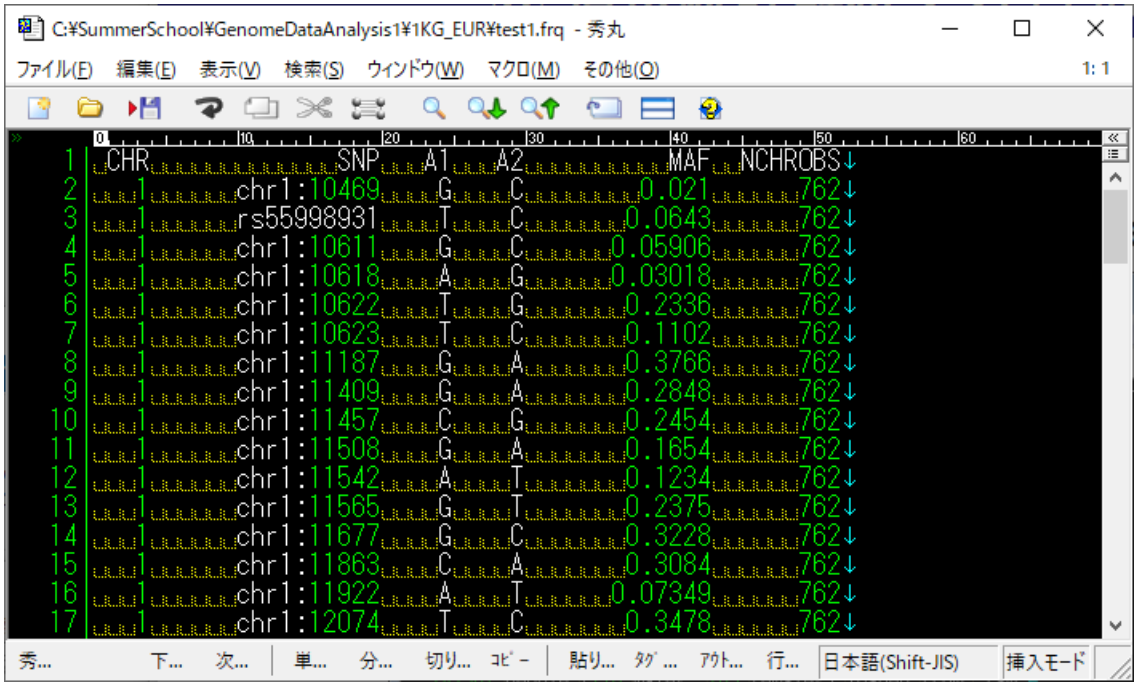

- "--bfile"は、bed/bim/fam形式のジェノタイプデータを読み込みます。
- "--file"だと、ped/map形式のジェノタイプデータを読み込みます。
- "--out"は、出力ファイル群の名前(ヘッダー部分)を指定します。

# ③ 遺伝統計解析ソフトPLINK実習

## ○：各SNPのアレル頻度の計算

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

出力ファイル：test1.frq

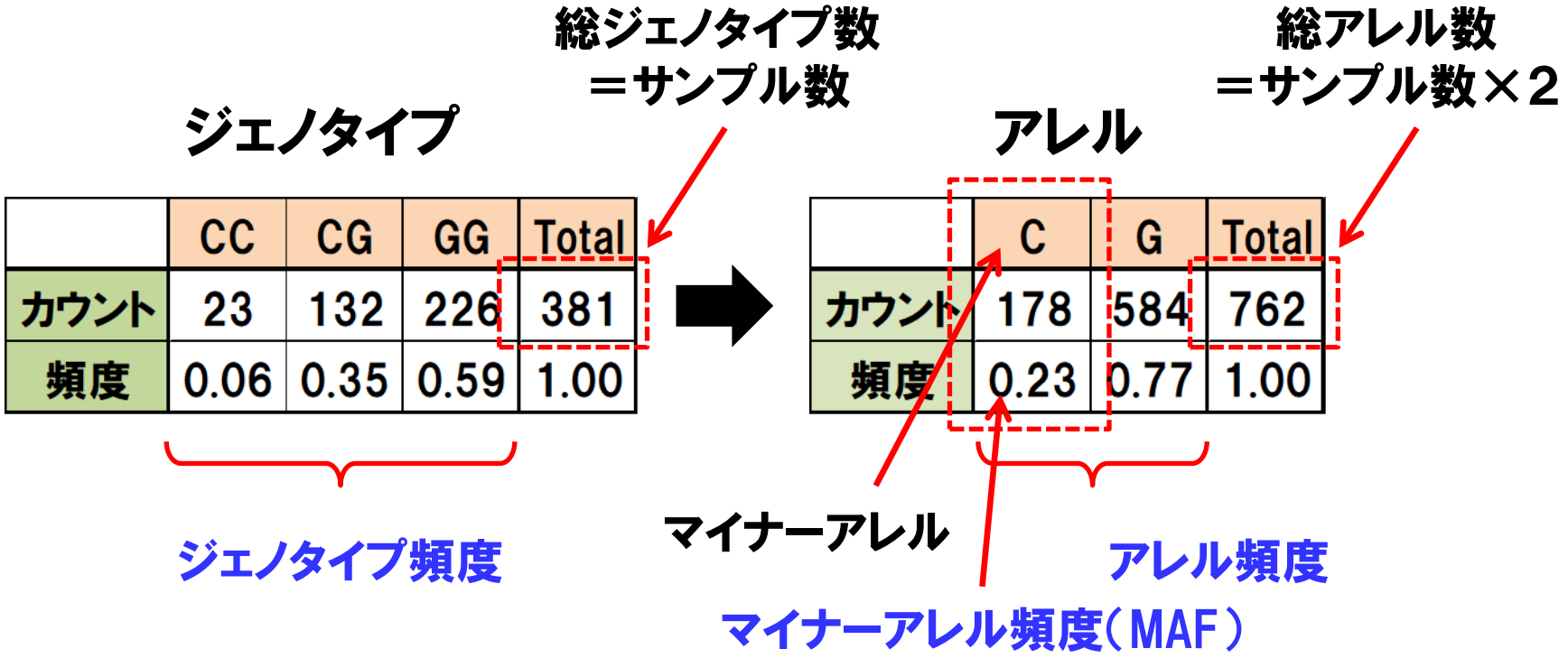



| | CHR | SNP | A1 | A2 | MAF | NCHROBS |
|---|---|---|---|---|---|---|
| 2 | 1 | chr1:10469 | G | C | 0.021 | 762 |
| 3 | 1 | rs55998931 | T | C | 0.0643 | 762 |
| 4 | 1 | chr1:10611 | G | C | 0.05906 | 762 |
| 5 | 1 | chr1:10618 | A | G | 0.03018 | 762 |
| 6 | 1 | chr1:10622 | T | G | 0.2336 | 762 |
| 7 | 1 | chr1:10623 | T | C | 0.1102 | 762 |
| 8 | 1 | chr1:11187 | G | A | 0.3766 | 762 |
| 9 | 1 | chr1:11409 | G | A | 0.2848 | 762 |
| 10 | 1 | chr1:11457 | C | G | 0.2454 | 762 |
| 11 | 1 | chr1:11508 | G | A | 0.1654 | 762 |
| 12 | 1 | chr1:11542 | A | T | 0.1234 | 762 |
| 13 | 1 | chr1:11565 | G | T | 0.2375 | 762 |
| 14 | 1 | chr1:11677 | G | C | 0.3228 | 762 |
| 15 | 1 | chr1:11863 | C | A | 0.3084 | 762 |
| 16 | 1 | chr1:11922 | A | T | 0.07349 | 762 |
| 17 | 1 | chr1:12074 | T | C | 0.3478 | 762 |



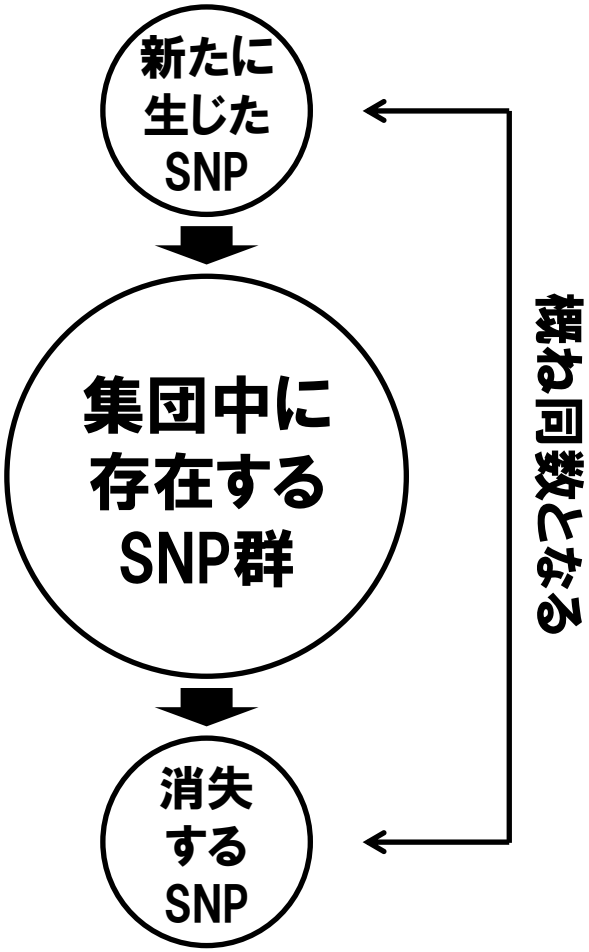
- "--freq"は、各SNPのアレル頻度を計算します。

# ③ 遺伝統計解析ソフトPLINK実習

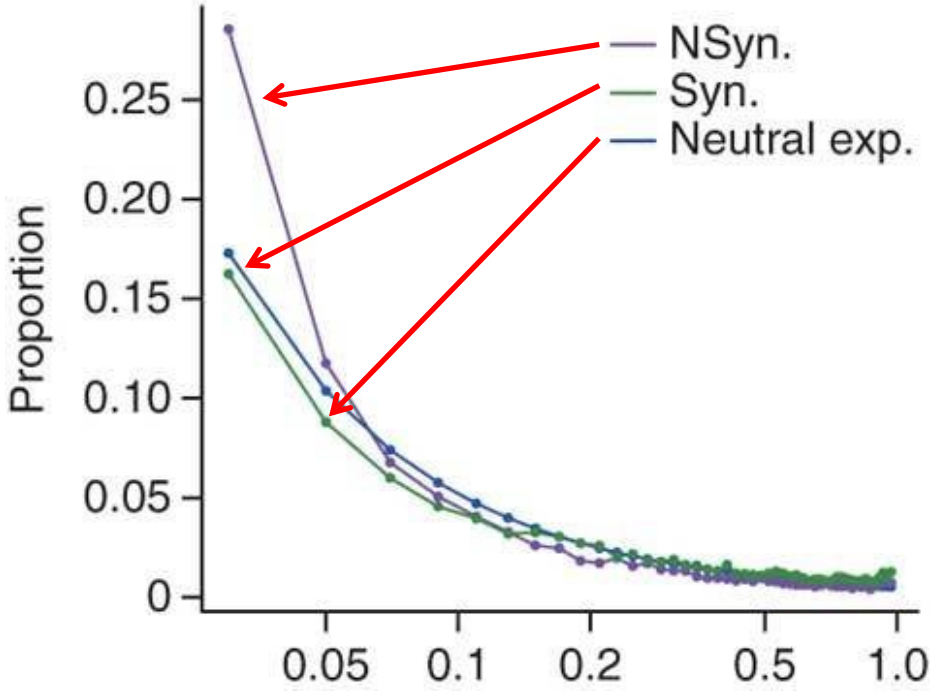



- SNPを構成する塩基変異を”アレル”（例:C/G）といいます。
- ２つのアレルの組み合わせを”ジェノタイプ”（例:CC/CG/GG）といいます。
- 頻度の小さい方のアレルを”マイナーアレル”といいます。
- 集団中のそれぞれの頻度を、”アレル頻度”、”ジェノタイプ頻度”、”マイナーアレル頻度”（minor allele frequency: MAF）といいます。

# ③ 遺伝統計解析ソフトPLINK実習




デンマーク集団における
SNPアレル頻度分布




(Li et al. ***Nat Genet*** 2010)

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

# ③ 遺伝統計解析ソフトPLINK実習

○: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

出力ファイル:test2.bed、test2.bim、test2.fam

サンプル数:381サンプルのまま

SNP数:8,830,185 SNP → MAF>0.2の3,191,128 SNP

- "--maf (数値)"で、MAFが指定した数値以下のSNPを除外できます。
- "--make-bed"で、フィルタリング後のデータを新たなbed/bim/famファイルとして作成します。
- "--recode"だと、新たなped/mapファイルとして作成します。

# ③ 遺伝統計解析ソフトPLINK実習

## ○：各SNPのHardy-Weinberg平衡の計算

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

出力ファイル：test3.hwe

C:¥SummerSchool¥GenomeDataAnalysis1¥1KG_EUR¥test3.hwe - 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)

| | CHR | SNP | TEST | A1 | A2 | GENO | O(HET) | E(HET) | P |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | chr1:10622 | ALL | T | G | 23/132/226 | 0.3465 | 0.3581 | 0.5665 |
| 3 | 1 | chr1:10622 | AFF | T | G | 0/0/0 | nan | nan | 1 |
| 4 | 1 | chr1:10622 | UNAFF | T | G | 23/132/226 | 0.3465 | 0.3581 | 0.5665 |
| 5 | 1 | chr1:11187 | ALL | G | A | 35/217/129 | 0.5696 | 0.4696 | 3.251e-05 |
| 6 | 1 | chr1:11187 | AFF | G | A | 0/0/0 | nan | nan | 1 |
| 7 | 1 | chr1:11187 | UNAFF | G | A | 35/217/129 | 0.5696 | 0.4696 | 3.251e-05 |
| 8 | 1 | chr1:11409 | ALL | G | A | 14/189/178 | 0.4961 | 0.4074 | 1.499e-05 |
| 9 | 1 | chr1:11409 | AFF | G | A | 0/0/0 | nan | nan | 1 |
| 10 | 1 | chr1:11409 | UNAFF | G | A | 14/189/178 | 0.4961 | 0.4074 | 1.499e-05 |
| 11 | 1 | chr1:11457 | ALL | C | G | 13/161/207 | 0.4226 | 0.3704 | 0.005626 |
| 12 | 1 | chr1:11457 | AFF | C | G | 0/0/0 | nan | nan | 1 |
| 13 | 1 | chr1:11457 | UNAFF | C | G | 13/161/207 | 0.4226 | 0.3704 | 0.005626 |
| 14 | 1 | chr1:11565 | ALL | G | T | 16/149/216 | 0.3911 | 0.3622 | 0.1562 |
| 15 | 1 | chr1:11565 | AFF | G | T | 0/0/0 | nan | nan | 1 |
| 16 | 1 | chr1:11565 | UNAFF | G | T | 16/149/216 | 0.3911 | 0.3622 | 0.1562 |
| 17 | 1 | chr1:11677 | ALL | G | C | 38/170/173 | 0.4462 | 0.4372 | 0.7265 |

秀丸ｴﾃﾞ... 下候補 次の結果 単語をｺ... 分割ｳｨﾝ... 切り抜き ｺﾋﾟｰ 貼り付け ﾀｸﾞｼﾞｬﾝ... ｱｳﾄﾗｲﾝ... 行番号... 日本語(Shift-JIS) 挿入モード

- "--hardy"は、各SNPのHardy Weinberg平衡の統計量(P値)を計算します。

# ③ 遺伝統計解析ソフトPLINK実習

## アレル

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

×

○

## ジェノタイプ

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

## HWEが成立する条件

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

## HWEの法則

- アレル1頻度 = p
- アレル2頻度 = 1-p

↓

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

- Hardy Weinberg平衡(HWE)とは、一定の条件下で、**アレル頻度からジェノタイプ頻度を推定できる**ことを指します。

（ジェノタイプ頻度からアレル頻度を計算することは、仮定なしでいつでもできます。）

# ③ 遺伝統計解析ソフトPLINK実習

**アレル**

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

一意に
計算可能

**ジェノタイプ（実測）**

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

HWE成立
を仮定

HWE検定で
乖離を評価
（P = 0.57）

・アレル1頻度 = p
・アレル2頻度 = 1−p
↓
・ジェノタイプ11頻度 = p*p
・ジェノタイプ12頻度 = 2*p*(1−p)
・ジェノタイプ22頻度 = (1−p)*(1−p)

**ジェノタイプ（推定）**

| | CC | CG | GG | Total |
|---|---|---|---|---|
| 頻度 | 0.05 | 0.36 | 0.59 | 1.00 |

- ジェノタイプ推定値と実測値の乖離を調べるのが、HWE検定です。
- 実験により実測されたジェノタイプ結果が不正確な時、HWE検定で実測値と推定値に乖離が生じやすいため、SNPデータのクオリティ・コントロール（Quality Control: QC）の一環として実施されます。

# ③ 遺伝統計解析ソフトPLINK実習

## ジェノタイプ

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

ホモ接合型　　ヘテロ接合型　　ホモ接合型

## HWEからの乖離
（ヘテロ接合型が減少）


## HWEからの乖離
（ヘテロ接合型が増加）


- ジェノタイプ実測値とHWE成立時のジェノタイプ推定値の間の乖離は、ヘテロ接合型ジェノタイプが増える場合と減る場合の2通りがあります。
- 有意なHWE検定P値を示すSNPジェノタイプが同定された場合、ヘテロ接合型の増減を確認する必要があります。

# ③ 遺伝統計解析ソフトPLINK実習

## ADH1B/ALDH2変異による死亡率への影響

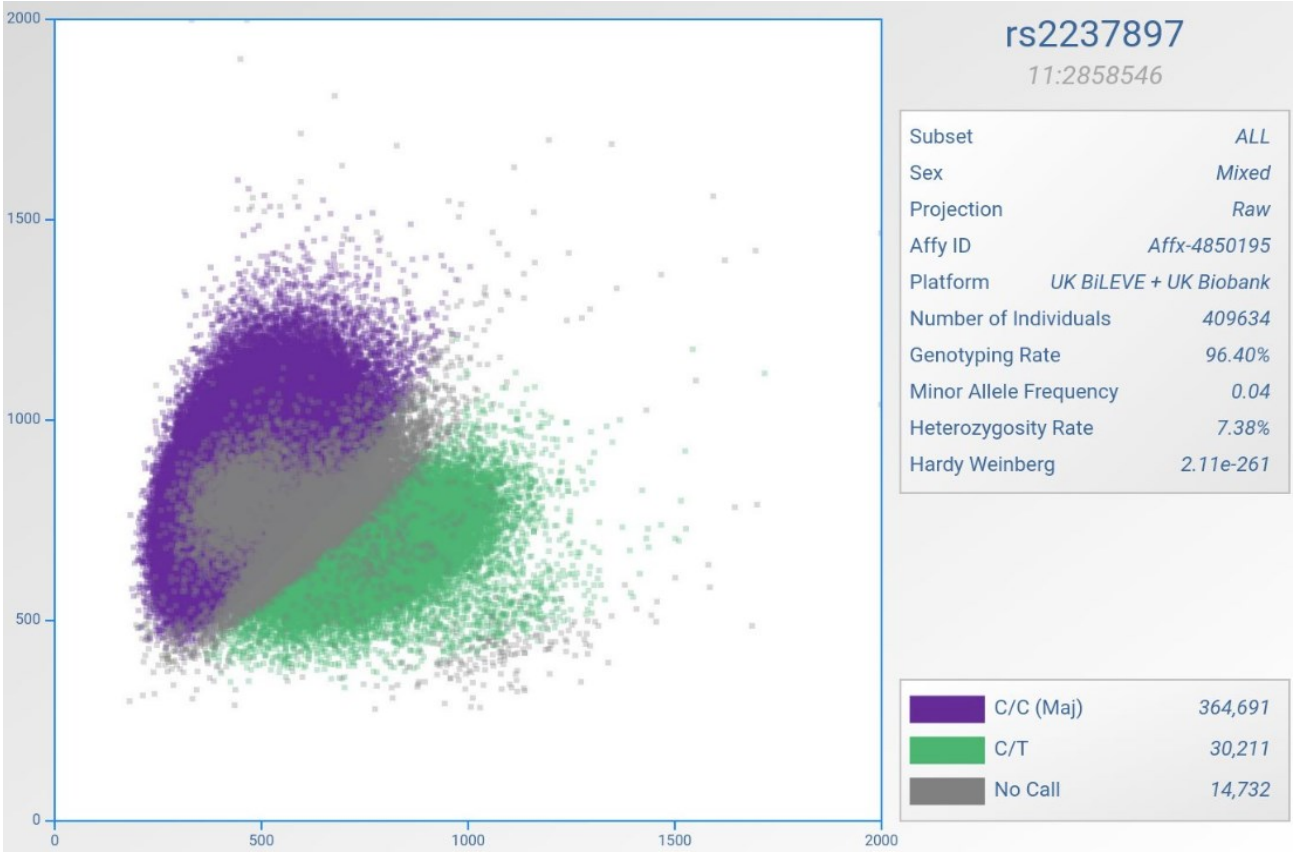
As both rs1229984 and rs671 were significantly deviated from the QC threshold of Hardy-Weinberg equilibrium ($P_{HWE} < 1.0\times10^{-6}$), …, indicating that the observed deviation from HWE was not caused by genotyping error but by heterogeneity in allele frequency spectra among the regions of Japan.

- 均一な集団内において、原則として、HWE平衡はゲノムワイドの全SNPで成立していると仮定されていますが、一部SNPで例外もあります。
- 日本人集団の飲酒量や死亡率に関連するADH1B変異(rs1229984)とALDH2変異(rs671)では、HWEからの乖離が認められますが、SNPジェノタイピングエラーに起因する乖離ではないことが知られています。

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

# ③ 遺伝統計解析ソフトPLINK実習

## 大規模バイオバンクSNPにおけるHWE乖離

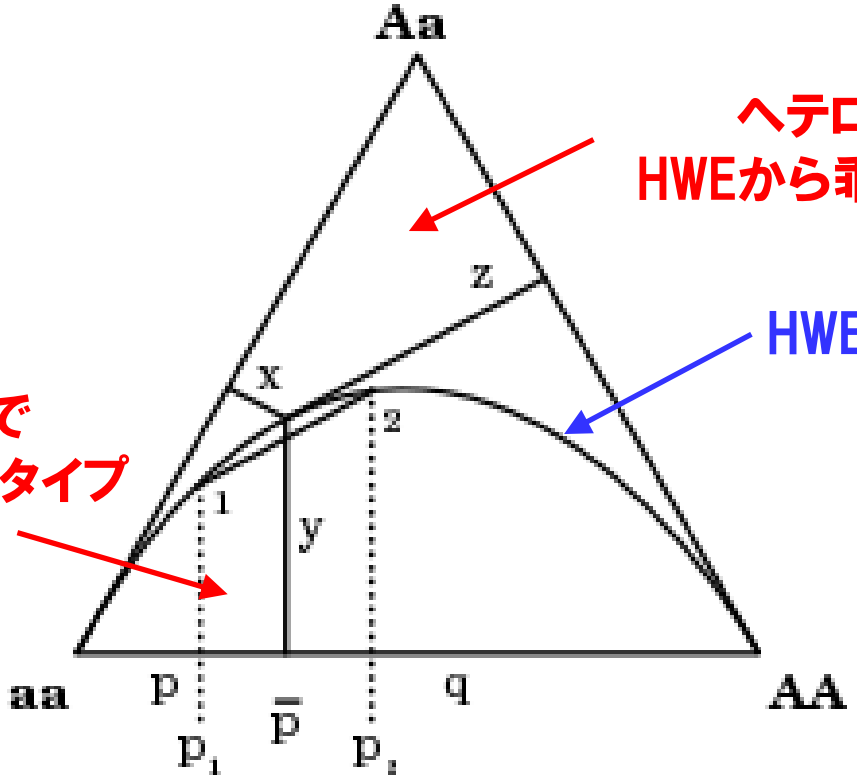



・複数の実験単位でタイピングされたSNPジェノタイプを、大規模バイオバンク全体で統合すると、新たに実験エラーでHWEからの有意な乖離を示すSNPの存在も明らかになってきています。

(http://www.nealelab.is/blog/2019/9/17/genotyped-snps-in-uk-biobank-failing-hardy-weinberg-equilibrium-test)

# ③ 遺伝統計解析ソフトPLINK実習

## de Finitte diagram




ヘテロ接合型減少で
HWEから乖離するジェノタイプ

HWEに従うジェノタイプ

ヘテロ接合型増加で
HWEから乖離するジェノタイプ

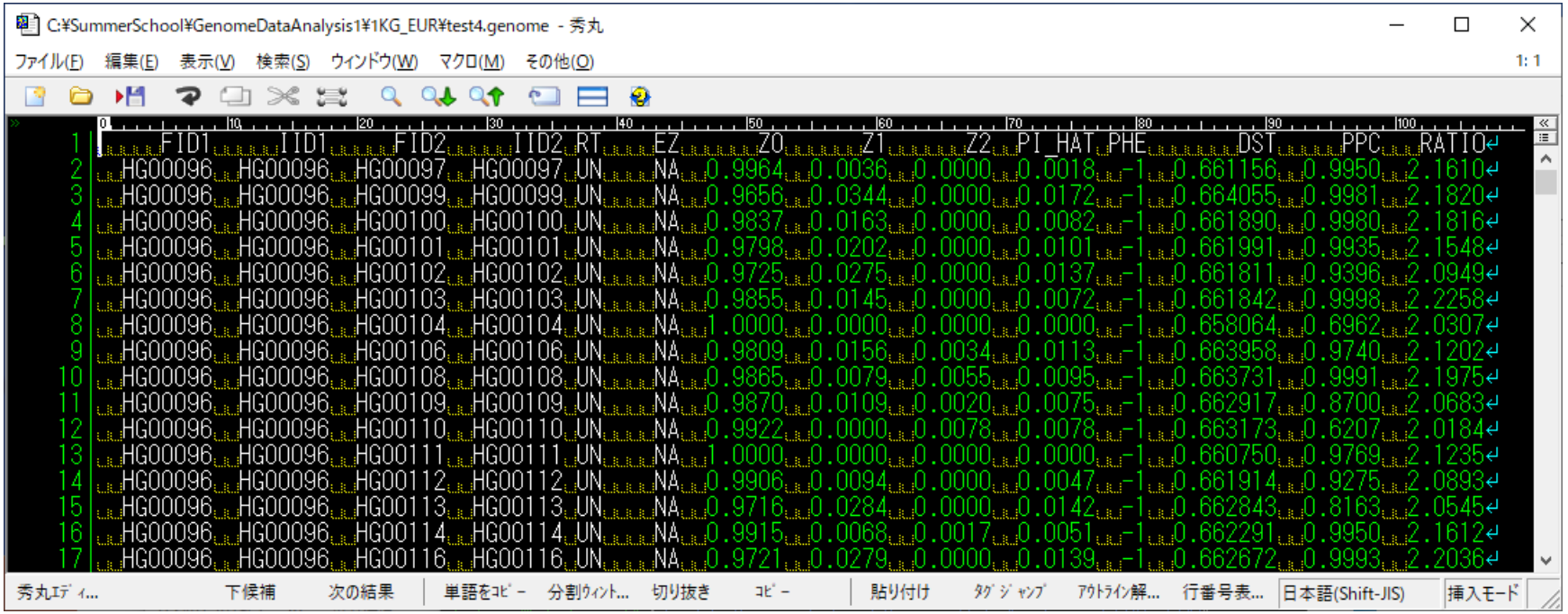
- ジェノタイプは自由度2の分割表であり、即ち任意のジェノタイプは2次元空間上の特定の座標で表すことができます。
- 3角形のプロット内の座標としてジェノタイプを表す図を、de Finitte diagramといい、HWEからの乖離を目視で確認するのに便利です。

# ③ 遺伝統計解析ソフトPLINK実習

## ○：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

出力ファイル：test4.genome


C:¥SummerSchool¥GenomeDataAnalysis1¥1KG_EUR¥test4.genome - 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)

| | FID1 | IID1 | FID2 | IID2 | RT | EZ | Z0 | Z1 | Z2 | PI_HAT | PHE | DST | PPC | RATIO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | HG00096 | HG00096 | HG00097 | HG00097 | UN | NA | 0.9964 | 0.0036 | 0.0000 | 0.0018 | -1 | 0.661156 | 0.9950 | 2.1610 |
| 3 | HG00096 | HG00096 | HG00099 | HG00099 | UN | NA | 0.9656 | 0.0344 | 0.0000 | 0.0172 | -1 | 0.664055 | 0.9981 | 2.1820 |
| 4 | HG00096 | HG00096 | HG00100 | HG00100 | UN | NA | 0.9837 | 0.0163 | 0.0000 | 0.0082 | -1 | 0.661890 | 0.9980 | 2.1816 |
| 5 | HG00096 | HG00096 | HG00101 | HG00101 | UN | NA | 0.9798 | 0.0202 | 0.0000 | 0.0101 | -1 | 0.661991 | 0.9935 | 2.1548 |
| 6 | HG00096 | HG00096 | HG00102 | HG00102 | UN | NA | 0.9725 | 0.0275 | 0.0000 | 0.0137 | -1 | 0.661811 | 0.9396 | 2.0949 |
| 7 | HG00096 | HG00096 | HG00103 | HG00103 | UN | NA | 0.9855 | 0.0145 | 0.0000 | 0.0072 | -1 | 0.661842 | 0.9998 | 2.2258 |
| 8 | HG00096 | HG00096 | HG00104 | HG00104 | UN | NA | 1.0000 | 0.0000 | 0.0000 | 0.0000 | -1 | 0.658064 | 0.6962 | 2.0307 |
| 9 | HG00096 | HG00096 | HG00106 | HG00106 | UN | NA | 0.9809 | 0.0156 | 0.0034 | 0.0113 | -1 | 0.663958 | 0.9740 | 2.1202 |
| 10 | HG00096 | HG00096 | HG00108 | HG00108 | UN | NA | 0.9865 | 0.0079 | 0.0055 | 0.0095 | -1 | 0.663731 | 0.9991 | 2.1975 |
| 11 | HG00096 | HG00096 | HG00109 | HG00109 | UN | NA | 0.9870 | 0.0109 | 0.0020 | 0.0075 | -1 | 0.662917 | 0.8700 | 2.0683 |
| 12 | HG00096 | HG00096 | HG00110 | HG00110 | UN | NA | 0.9922 | 0.0000 | 0.0078 | 0.0078 | -1 | 0.663173 | 0.6207 | 2.0184 |
| 13 | HG00096 | HG00096 | HG00111 | HG00111 | UN | NA | 1.0000 | 0.0000 | 0.0000 | 0.0000 | -1 | 0.660750 | 0.9769 | 2.1235 |
| 14 | HG00096 | HG00096 | HG00112 | HG00112 | UN | NA | 0.9906 | 0.0094 | 0.0000 | 0.0047 | -1 | 0.661914 | 0.9275 | 2.0893 |
| 15 | HG00096 | HG00096 | HG00113 | HG00113 | UN | NA | 0.9716 | 0.0284 | 0.0000 | 0.0142 | -1 | 0.662843 | 0.8163 | 2.0545 |
| 16 | HG00096 | HG00096 | HG00114 | HG00114 | UN | NA | 0.9915 | 0.0068 | 0.0017 | 0.0051 | -1 | 0.662291 | 0.9950 | 2.1612 |
| 17 | HG00096 | HG00096 | HG00116 | HG00116 | UN | NA | 0.9721 | 0.0279 | 0.0000 | 0.0139 | -1 | 0.662672 | 0.9993 | 2.2036 |

秀丸エディ... 下候補 次の結果 単語をコピー 分割ウィント... 切り抜き コピー 貼り付け タグジャンプ アウトライン解... 行番号表... 日本語(Shift-JIS) 挿入モード

- "--genome"は、全サンプルペアの組み合わせについて、遺伝的な近さ（近縁関係）を推定します。

# ③ 遺伝統計解析ソフトPLINK実習

## ○：IBS（identity-by-state）

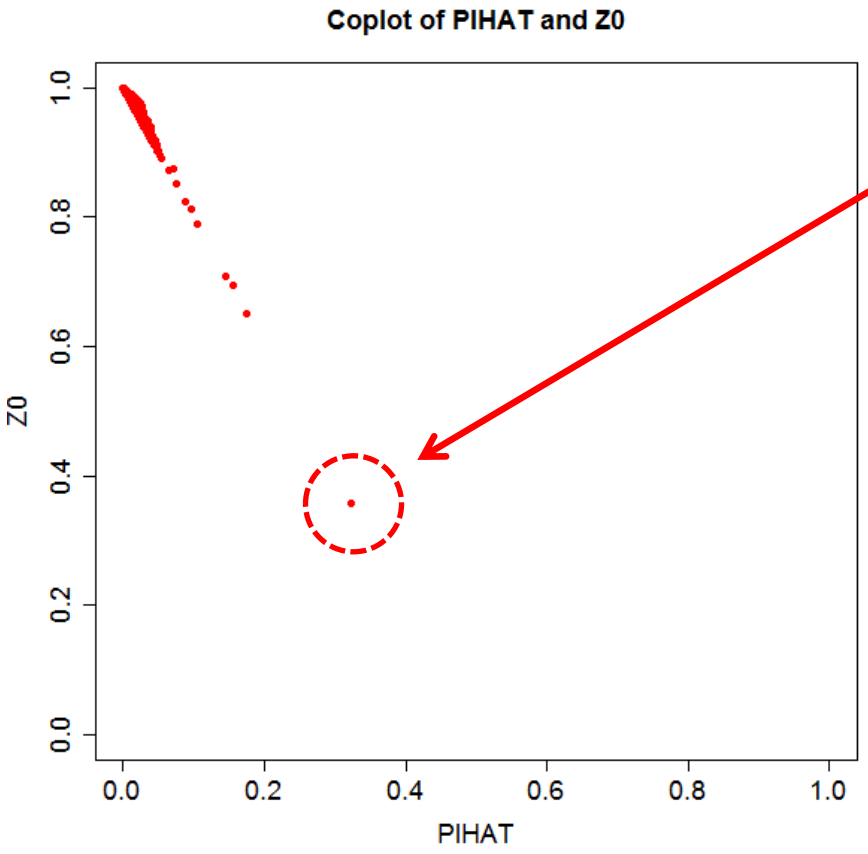
- 2サンプルにおいて、あるアレルが、同じであること。
- アレルの由来は問わない。
- IBSは2サンプルにおける各SNPに対して実測可能。
- IBS=0のSNPが少なく、IBS=2のSNPの個数が多いほど、近縁関係にある。

## ○：IBD（identity-by-descent）

- 2サンプルにおいて、あるアレルを、同じ祖先から受け継いでいること（同祖由来）。
- IBDは直接観測不可能のため、IBSやアレル頻度分布から推定する。
- PLINKでは、PI_HATという値で、各サンプルペア間のIBDの値を推定できます。
- PI_HAT＝0（近縁関係なし）、PI_HAT＝0.25（おじ・おば）、PI_HAT＝0.5（親子/兄弟）、PI_HAT＝1（本人/一卵性双生児）と、IBD推定値に基づきサンプルペア間の近縁関係を知ることができます。

- 遺伝情報に基づくサンプルペア間の近縁関係を表す指標として、”IBS”と”IBD”があります。

# ③ 遺伝統計解析ソフトPLINK実習




サンプルペア

- "HG00119"
- "HG00124"

PI_HAT = 0.32

Z0 = 0.36

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

- "test4.genome"ファイル中の、"PI_HAT"と"Z0"（IBS=0のSNPの割合）をプロットすると、近縁関係がわかります。
- サンプルペア"HG00119"と"HG00124"は、おじ・おば程度の近縁関係にあるか、ゲノムが混入している可能性が浮上しました。

# ③ 遺伝統計解析ソフトPLINK実習

## ○:サンプルの遺伝的背景の推定

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

出力ファイル:test5.mds

C:¥SummerSchool¥GenomeDataAnalysis1¥1KG_EUR¥test5.mds - 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

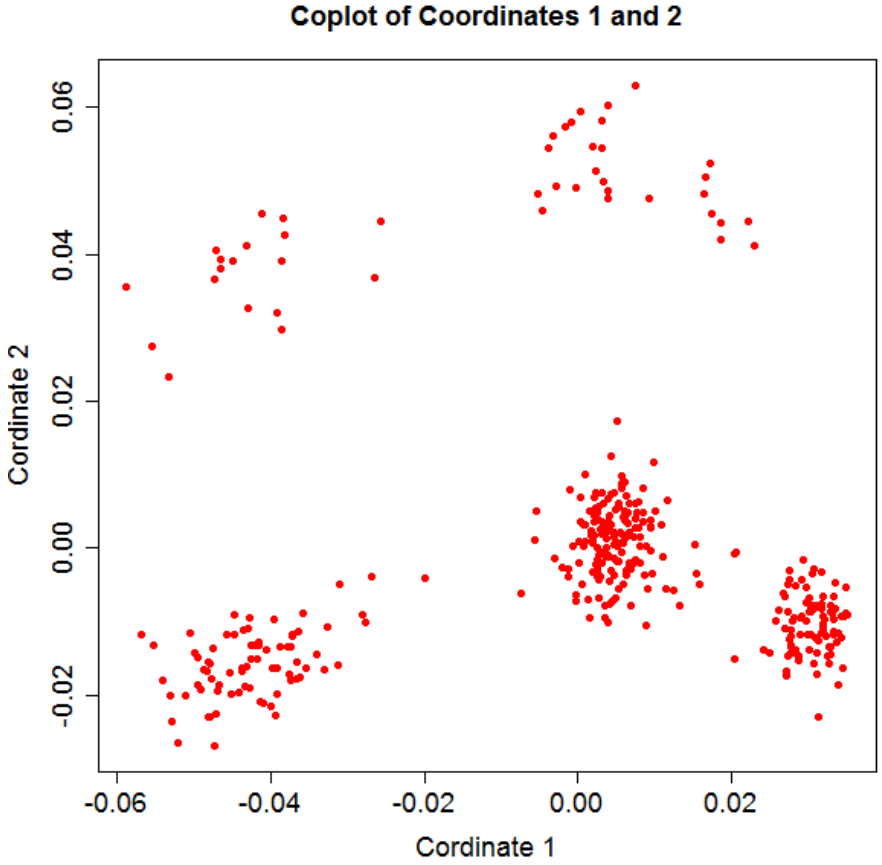
| | FID | IID | SOL | C1 | C2 | C3 | C4 |
|---|---|---|---|---|---|---|---|
| 2 | HG00096 | HG00096 | 0 | 0.00686842 | -0.00244403 | -0.0185996 | 0.0097876 |
| 3 | HG00097 | HG00097 | 0 | 0.00310107 | 0.0545643 | 0.000953304 | 0.0109884 |
| 4 | HG00099 | HG00099 | 0 | -0.00176052 | 0.0574844 | -0.00426765 | -0.00516236 |
| 5 | HG00100 | HG00100 | 0 | 0.00801577 | 0.0015642 | -0.0305084 | 0.00915341 |
| 6 | HG00101 | HG00101 | 0 | -0.000814574 | 0.0579398 | -0.00185522 | 0.00736288 |
| 7 | HG00102 | HG00102 | 0 | -0.00375586 | 0.0545004 | 0.00115038 | -0.012981 |
| 8 | HG00103 | HG00103 | 0 | 0.00245805 | -0.00256974 | -0.0189964 | -0.0174293 |
| 9 | HG00104 | HG00104 | 0 | -0.00324035 | 0.0562325 | 0.00837995 | 0.00107564 |
| 10 | HG00106 | HG00106 | 0 | 0.00515692 | -0.00178675 | -0.0304097 | -0.0148813 |
| 11 | HG00108 | HG00108 | 0 | 0.000748419 | 0.00326841 | -0.0260576 | 0.0317262 |
| 12 | HG00109 | HG00109 | 0 | 0.00391991 | 0.0487702 | 0.00120484 | -0.00146337 |
| 13 | HG00110 | HG00110 | 0 | -0.00289634 | 0.0493008 | -0.00388013 | -0.0131919 |
| 14 | HG00111 | HG00111 | 0 | -0.00125867 | -0.00278667 | -0.0276712 | 0.00546834 |
| 15 | HG00112 | HG00112 | 0 | -0.000192971 | -0.00632142 | -0.0222364 | 0.0171959 |
| 16 | HG00113 | HG00113 | 0 | 0.00302745 | 0.0581966 | 0.000741724 | 0.017252 |
| 17 | HG00114 | HG00114 | 0 | 0.00366942 | 0.000120276 | -0.0263168 | -0.0305194 |

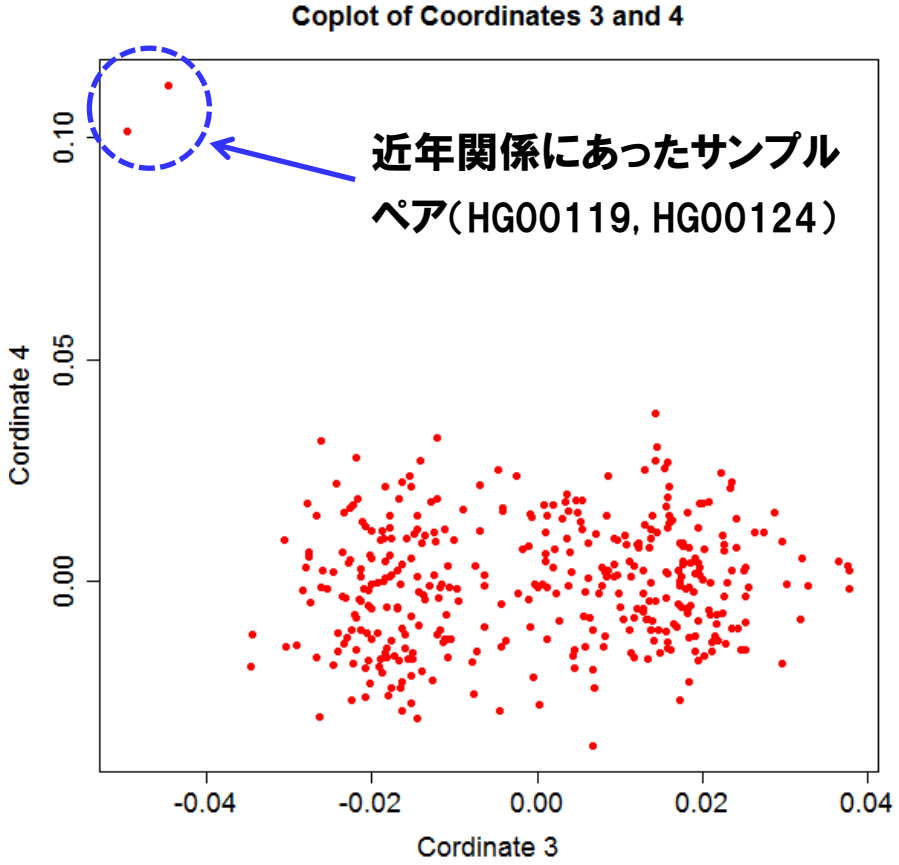
秀丸エ... 下候補 次の... 単語... 分割ウ... 切り... コピー 貼り... タグジ... アウトライ... 行番... 日本語(Shift-JIS) 挿入モード

- "--cluster"と"--mds-plot" で、多次元尺度構成法(MDS: multi-dimensional scaling)による、サンプルのクラスタリングを実施できます。
- 一般には主成分分析(PCA: principal component analysis)が広く使われます。

# ③ 遺伝統計解析ソフトPLINK実習

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

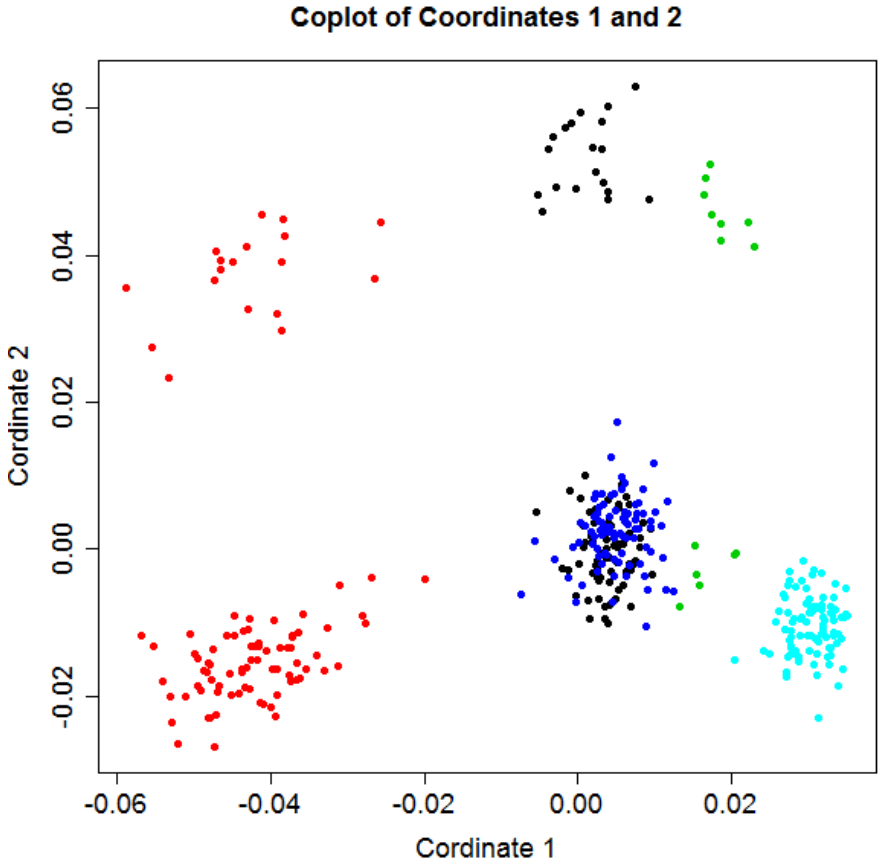


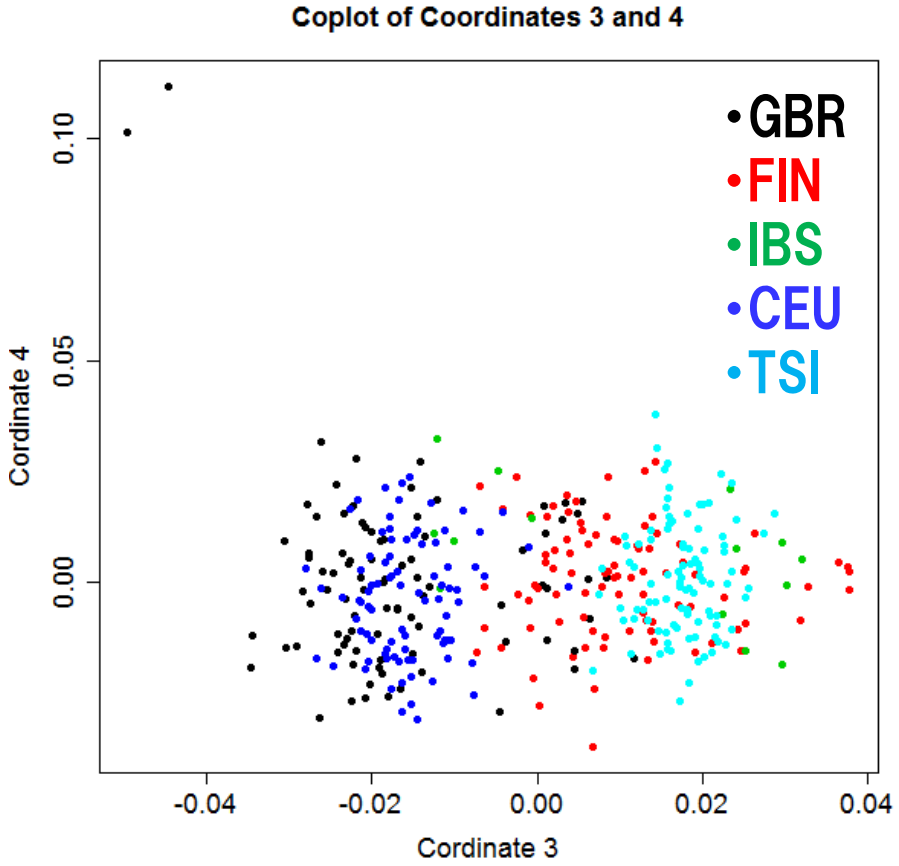




・遺伝的背景に基づき、サンプルをクラスタリングすることができました。

# ③ 遺伝統計解析ソフトPLINK実習

※ファイル"PlotMDS2.R"を開いて、内容をRにコピー＆ペーストして下さい。




- 地域集団で色をつけてみると、地域集団ごとや、近縁関係に基づきクラスタリングされていることがわかります。
- つまり、遺伝情報から出身地域を推定することが可能です。

# 終わりに

- ヒトゲノムデータの入手方法から解析まで、駆け足でなぞってみました。
- 数千人、数千万SNPもの大規模なゲノムデータが一般公開され、数多くの研究のリソースとして活用されています。
- 特別なプロジェクトを担当しなくても、公開ゲノムデータを使うだけで、誰でもゲノムデータ解析や研究ができることを感じて頂ければと思います。
- PLINKをはじめ、ゲノムデータ解析のソフトウェアは充実していて、Linux環境を確保すれば誰でも簡単に実施できる状況にあります。
- 興味をもったソフトウェアやテーマについて、色々と試してみて下さい。