

# NGSを用いたジェノタイピングを 様々な解析に用いるには？ ～高密度SNPデータ解析の処方箋～

鐘ヶ江 弘美

東京大学大学院 農学生命科学研究科  
生産・環境生物学専攻 生物測定学研究室

# Molecular Markers

- **Simple Sequence Repeats (SSR) markers**

遺伝子のマッピングやQTL解析、マーカー選抜など様々な解析に利用される  
共優性マーカー(codominant marker)、multiallelicマーカー  
再現性が高く、他の種でも利用可能

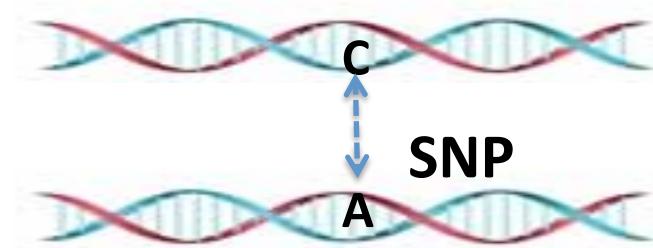
- **Single Nucleotide Polymorphism (SNP) markers**

SSRマーカーと比較するとハイスループット  
連鎖地図やゲノミックセレクション、GWASに利用されている

- **Next generation sequencing (NGS) を使ったジェノタイピング**

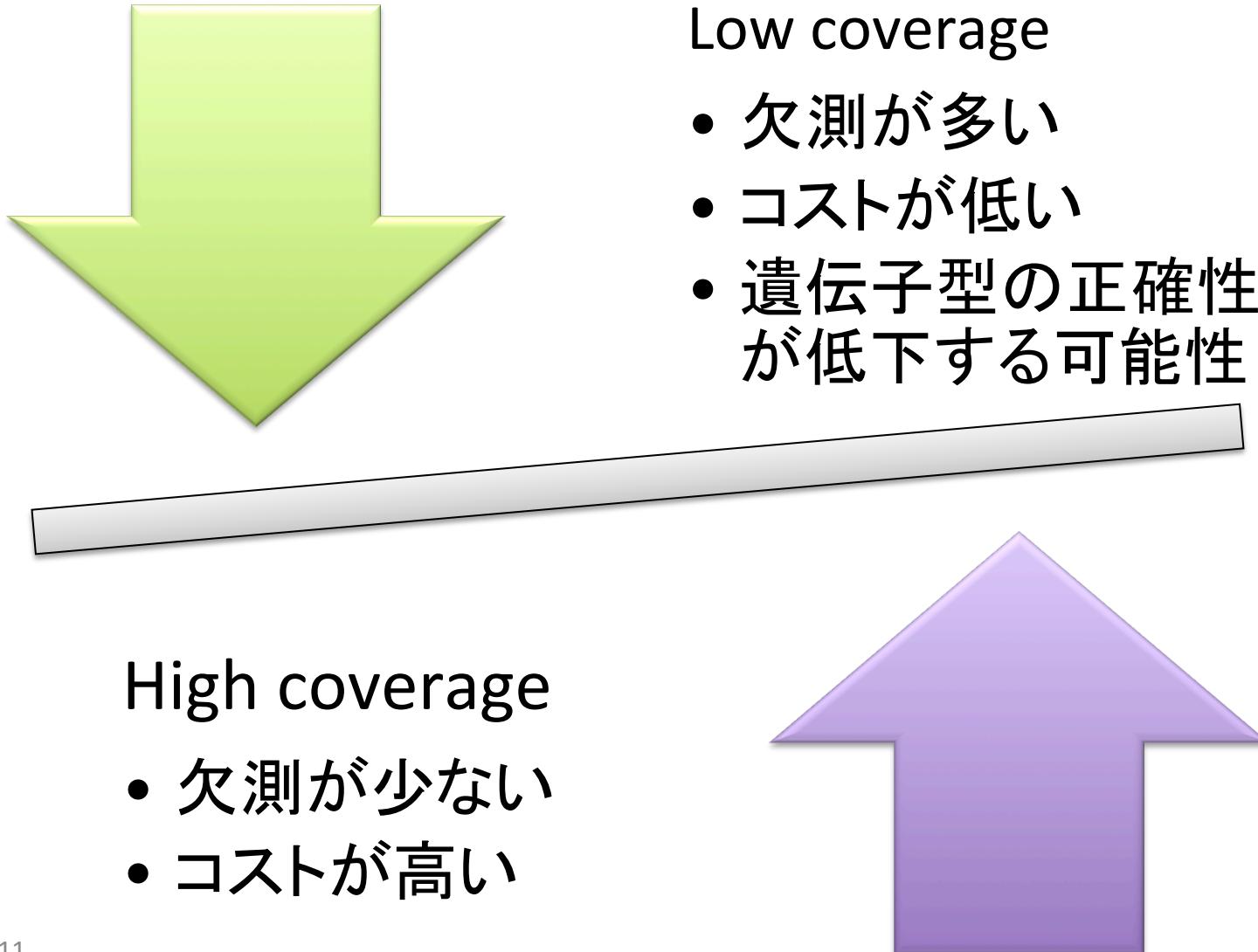
RAD-seq、GBS、low depth WGS

一度に多くのSNPを得ることができる  
マーカーの偏りが少ない  
欠測も多い

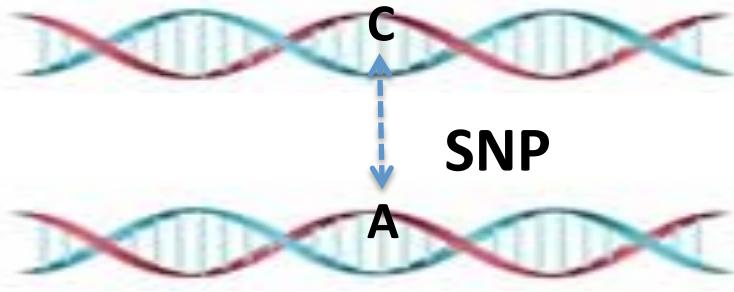




# GWAS やゲノミックセレクションに利用する時の問題点



# ゲノミックセレクション -出穂日の予測



## 予測モデル

$y$ :出穂日

$$y=f(x_1, x_2, x_3, \dots, x_k)$$



## 出穂日予測

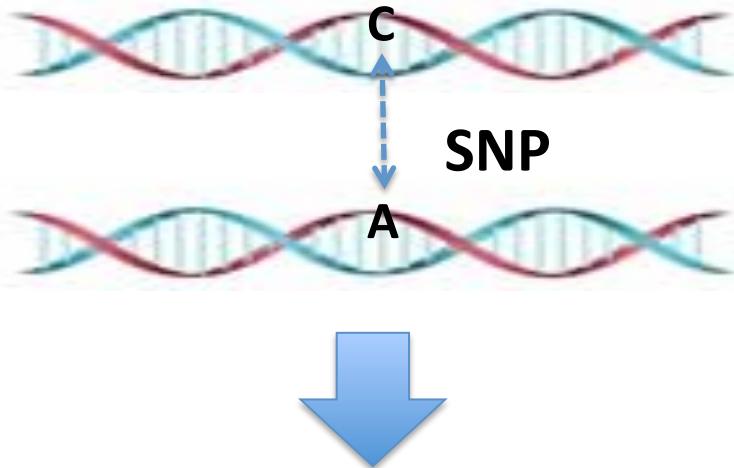
$$y=f(x)$$

実際にはどこに遺伝子があるか分からないので、  
高密度DNA多型を網羅的に予測に利用する

表現型( $y$ )およびDNAマーカー多型を調査し、  
回帰分析を用いて、予測式 $f(x)$ を求めておく



出穂日( $y$ )が未知の個体でも、  
DNAマーカー多型( $x_1, x_2, x_3, \dots, x_k$ )から  
出穂日を予測できる。



**予測モデル**  
 $y$ :出穂日  
 $y = f(x_1, x_2, x_3, \dots, x_k)$



**出穂日予測**  
 $y = f(x)$



欠測が多く、DNA多型が低密度

DNA多型を網羅的に予測に利用できない



検出力の低下

欠測があるマーカーは マーカー遺伝子型から  
行列を計算できない

出穂日      DNA多型

$$\begin{matrix} y_1 \\ y_2 \\ y_3 \\ \dots \end{matrix} = \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \dots \end{matrix}$$

$b + e$

出穂日の予測精度が低下

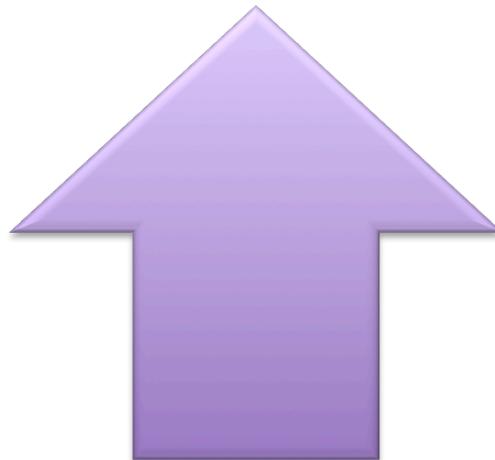
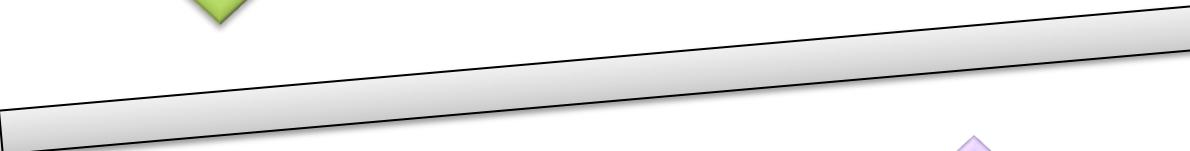
# GWAS やゲノミックセレクションに利用する時の問題点



Low coverage

- 欠測が多い
- コストが低い
- ジェノタイピングの正確性が低下する可能性

欠測した遺伝子型を補完し  
低成本でジェノタイピング

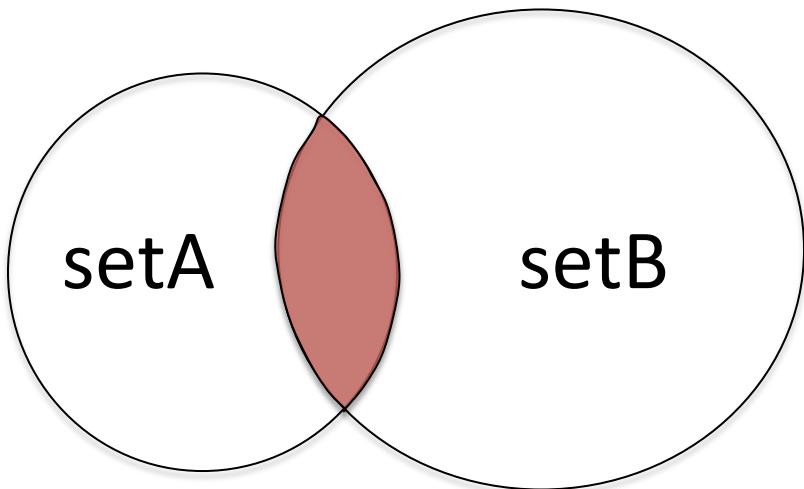


High coverage

- 欠測が少ない
- コストが高い

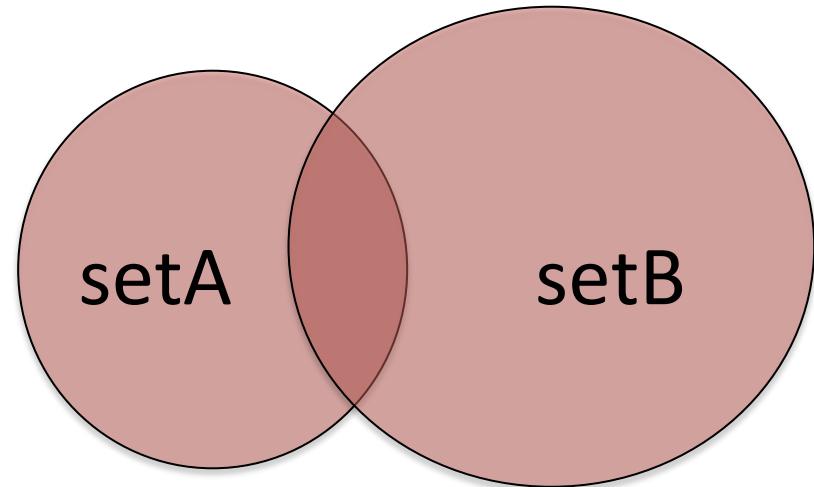
# マーカーセットが異なるデータを用いる場合

Imputationを行わない場合



両方のセットで遺伝子型データのある  
重なったマーカーしか利用できない

Imputationを行う場合



2つのデータセットで共通していない  
マーカーの遺伝子型を補完する  
Imputationを行うことで、  
すべてのマーカーを利用できる



# IMPUTATION ソフトウェア

ソフトウェア名	URL
Beagle	<a href="https://faculty.washington.edu/browning/beagle/beagle.html">https://faculty.washington.edu/browning/beagle/beagle.html</a>
Tassel	<a href="http://www.maizegenetics.net/#!tassel/c17q9">http://www.maizegenetics.net/#!tassel/c17q9</a>
IMPUTE2	<a href="https://mathgen.stats.ox.ac.uk/impute/impute_v2.html">https://mathgen.stats.ox.ac.uk/impute/impute_v2.html</a>
PLINK	<a href="http://pngu.mgh.harvard.edu/~purcell/plink/pimputation.shtml">http://pngu.mgh.harvard.edu/~purcell/plink/pimputation.shtml</a>
minimac2	<a href="http://genome.sph.umich.edu/wiki/Minimac2">http://genome.sph.umich.edu/wiki/Minimac2</a>

# 遺伝子型の予測

## Step 1

→ マーカー

品種A	T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
品種B	C	C	·	A	G	C	T	C	C	G	A	G	C	·	T	C
品種C	T	C	·	A	G	C	G	T	C	G	A	·	G	A	G	C
品種D	C	C	C	A	A	C	G	T	·	G	A	G	C	G	T	C
品種E	A	C	C	A	G	C	T	C	C	·	A	G	G	G	G	T
品種F	C	C	C	A	G	C	T	C	C	G	A	G	G	·	T	T

サンプルで遺伝子型を  
共有している領域を特定

ゲノムに存在する連鎖不平衡と  
ハプロタイプブロック構造を利用

## Step 2

→ マーカー

品種A	T	C	C	A	G	C	G	T	C	G	·	G	G	A	G	C
品種B	C	C	·	A	G	C	T	C	C	G	A	G	C	·	T	C
品種C	T	C	·	A	G	C	G	T	C	G	A	·	G	A	G	C
品種D	C	C	C	A	A	C	G	T	·	G	A	G	C	G	T	C
品種E	A	C	C	A	G	C	T	C	C	G	A	G	G	G	G	T
品種F	C	C	C	A	G	C	T	C	C	G	A	G	G	·	T	T

## Step 3

→ マーカー

品種A	T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
品種B	C	C	C	A	G	C	T	C	C	G	A	G	C	G	T	C
品種C	T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
品種D	C	C	C	A	A	C	G	T	C	G	A	G	C	G	T	C
品種E	A	C	C	A	G	C	T	C	C	G	A	G	G	G	G	T
品種F	C	C	C	A	G	C	T	C	C	G	A	G	G	G	T	T

ハプロタイプの情報から  
欠測している遺伝子型を補完

# リファレンスパネルを用いた遺伝子型の予測

## Step 1

### サンプル

→ マーカー

品種A	T	.	C	A	G	C	.	.	.	G	A	G	G	A	G	.
品種B	C	C	C	A	G	C	.	.	.	G	A	G	C	.	T	C

### リファレンスパネル

品種C	T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
品種D	C	C	C	A	A	C	G	T	C	G	A	G	C	G	T	C
品種E	A	C	C	A	G	C	T	C	C	G	A	G	G	G	G	T
品種F	C	C	C	A	G	C	T	C	C	G	A	G	G	G	T	T

サンプルとリファレンスパネルの中の個体で、遺伝子型を共有している領域を特定

ゲノムに存在する連鎖不平衡とハプロタイプブロック構造を利用

## Step 2

### サンプル

→ マーカー

品種A	T	.	C	A	G	C	.	.	.	G	A	G	G	A	G	.
品種B	C	C	C	A	G	C	.	.	.	G	A	G	C	.	T	C

### リファレンスパネル

品種C	T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
品種D	C	C	C	A	A	C	G	T	C	G	A	G	C	G	T	C
品種E	A	C	C	A	G	C	T	C	C	G	A	G	G	G	G	T
品種F	C	C	C	A	G	C	T	C	C	G	A	G	G	G	T	T

リファレンスパネルの遺伝子型とハプロタイプの情報から欠測している遺伝子型を補完

## Step 3

### サンプル

→ マーカー

品種A	T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
品種B	C	C	C	A	G	C	T	C	C	G	A	G	C	G	T	C

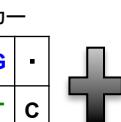
### リファレンスパネル

品種C	T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
品種D	C	C	C	A	A	C	G	T	C	G	A	G	C	G	T	C
品種E	A	C	C	A	G	C	T	C	C	G	A	G	G	G	G	T
品種F	C	C	C	A	G	C	T	C	C	G	A	G	G	G	T	T

# リファレンスパネルを用いた遺伝子型の予測

欠測を含むRAD-seq

	T	C	C	A	G	C	·	C	·	G	A	G	G	A	G	·
品種A	T	·	C	A	G	C	·	C	·	G	A	G	G	A	G	·
品種B	C	C	·	A	G	C	T	C	C	G	A	G	C	·	T	C



リファレンスパネル

T	C	C	A	G	C	G	T	C	G	G	A	G	G	A	G	C
C	C	C	A	A	C	G	T	C	G	A	G	C	G	T	C	
A	C	C	A	G	C	T	C	C	G	A	G	G	G	G	T	C
C	C	C	A	G	C	T	C	C	G	A	G	G	G	T	T	

遺伝子型を補完した  
サンプルデータ

T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
C	C	C	A	G	C	T	C	C	G	A	G	C	G	T	C

リファレンスパネルの遺伝子型に基づいて  
統計学的にサンプルの遺伝子型を予測

ジェノタイピング  
されていないSNP

	T	C	C	A	G	C	·	·	·	G	A	G	G	A	G	C
品種A	T	C	C	A	G	C	·	·	·	G	A	G	G	A	G	C
品種B	C	C	C	A	G	C	·	·	·	G	A	G	C	T	C	

リファレンスパネル

T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
C	C	C	A	A	C	G	T	C	G	A	G	C	G	T	C
A	C	C	A	G	C	T	C	C	G	A	G	G	G	G	T
C	C	C	A	G	C	T	C	C	G	A	G	G	G	T	T

ジェノタイピングされて  
いないSNPの補完

T	C	C	A	G	C	G	T	C	G	A	G	G	A	G	C
C	C	C	A	G	C	T	C	C	G	A	G	C	G	T	C

遺伝子型を補完することにより、  
解析するSNP数を増やすことが可能

# リファレンスパネルの作成

- 自前のデータだけでは、リファレンスパネルとして利用できる系統数が少ない
- 公共データベースで公開されているゲノム配列を解析することで、系統数を増やすことができる
- 公開されたデータを使うことで、シークエンスにコストがかからない
- SNPの遺伝子型情報を利用してリファレンスパネルを作成

## 公開されているゲノム配列を検索

データベース名	URL
DRAsearch	<a href="https://trace.ddbj.nig.ac.jp/DRASearch/">https://trace.ddbj.nig.ac.jp/DRASearch/</a>
DBCLS SRA	<a href="http://sra.dbcls.jp">http://sra.dbcls.jp</a>
NCBI	<a href="http://www.ncbi.nlm.nih.gov/sra/">http://www.ncbi.nlm.nih.gov/sra/</a>
DNAexus	<a href="http://sra.dnanexus.com/">http://sra.dnanexus.com/</a>

# 公開されているゲノム配列を検索

DBCLS SRA

<http://sra.dbcls.jp>

下位を含めて、生物種で検索

- [Oryza sativa](#) (rice) Click on organism name to get more information.
  - [Oryza sativa Indica Group](#) (long-grained rice)
    - [Oryza sativa Aus Group](#)
    - [Oryza sativa Indica Group x Oryza sativa Japonica Group](#)
  - [Oryza sativa Japonica Group](#) (Japanese rice)
    - [Oryza sativa Aromatic Japonica Group](#)
    - [Oryza sativa Temperate Japonica Group](#)
    - [Oryza sativa Tropical Japonica Group](#)
  - [Oryza sativa Japonica Group x Oryza sativa Indica Group](#)

## Project List from taxonomy (β version)

Study Type: Whole Genome Sequencing Platform: Taxon ID: 4530  incl. child taxonomy (ex. strains)  Species name: Oryza sativa

→ back to DBCLS SRA top

Oryza > Oryza sativa

→ TAB-delimited format

Total: 371 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 10 ▾

SRA ID	Study ID	Study Title	Study Type	Taxon ID	Taxon Name	Exps	Runs	Update
DRA000010	DRP00010	Whole genome shotgun sequences of Oryza sativa japonica variety, Koshihikari	Whole Genome Sequencing	39947	Oryza sativa Japonica Group	1	9	2009-08-11
DRA000029	DRP00029	Analysis of somaclonal variation on the genome of regenerated rice	Whole Genome Sequencing	39947	Oryza sativa Japonica Group	1	2	2009-09-15
DRA000307	DRP000308	Whole genome sequencing of Japonica rice cultivar Omachi	Whole Genome Sequencing	39947	Oryza sativa Japonica Group	1	2	2010-10-14
DRA000434	DRP000443	Analysis of somaclonal variation on the genome of regenerated rice	Whole Genome Sequencing	39947	Oryza sativa Japonica Group	1	1	2011-08-18
DRA000470	DRP000489	Oryza sativa cv. Nipponbare whole genome	Whole Genome Sequencing	39947	Oryza sativa Japonica Group	1	1	2011-09-21
DRA000499	DRP000523	Genome sequencing reveals agronomically-important loci in rice using MutMap	Whole Genome Sequencing	39947	Oryza sativa Japonica Group	7	8	2011-09-21
ERA000078	ERP000111	Rice High-throughput genotyping by whole-genome resequencing	Whole Genome Sequencing	4530	Oryza sativa	1	150	
ERA000213	ERP000106	Sequencing 620 rice genomes for genome-wide association studies	Whole Genome Sequencing	4530	Oryza sativa	620	620	
ERA000970	ERP000235	The indica genome sequence by next-generation sequencing	Whole Genome Sequencing	39946	Oryza sativa Indica Group	1	5	
ERA000971	ERP000236	The japonica genome sequence by next-generation sequencing	Whole Genome Sequencing	39947	Oryza sativa Japonica Group	1	8	

Total: 371 << first < prev 1 2 3 4 5 6 7 8 9 10 next > last >> 10 ▾

→ back to DBCLS SRA top

検索結果を  
タブ区切りで  
保存可能



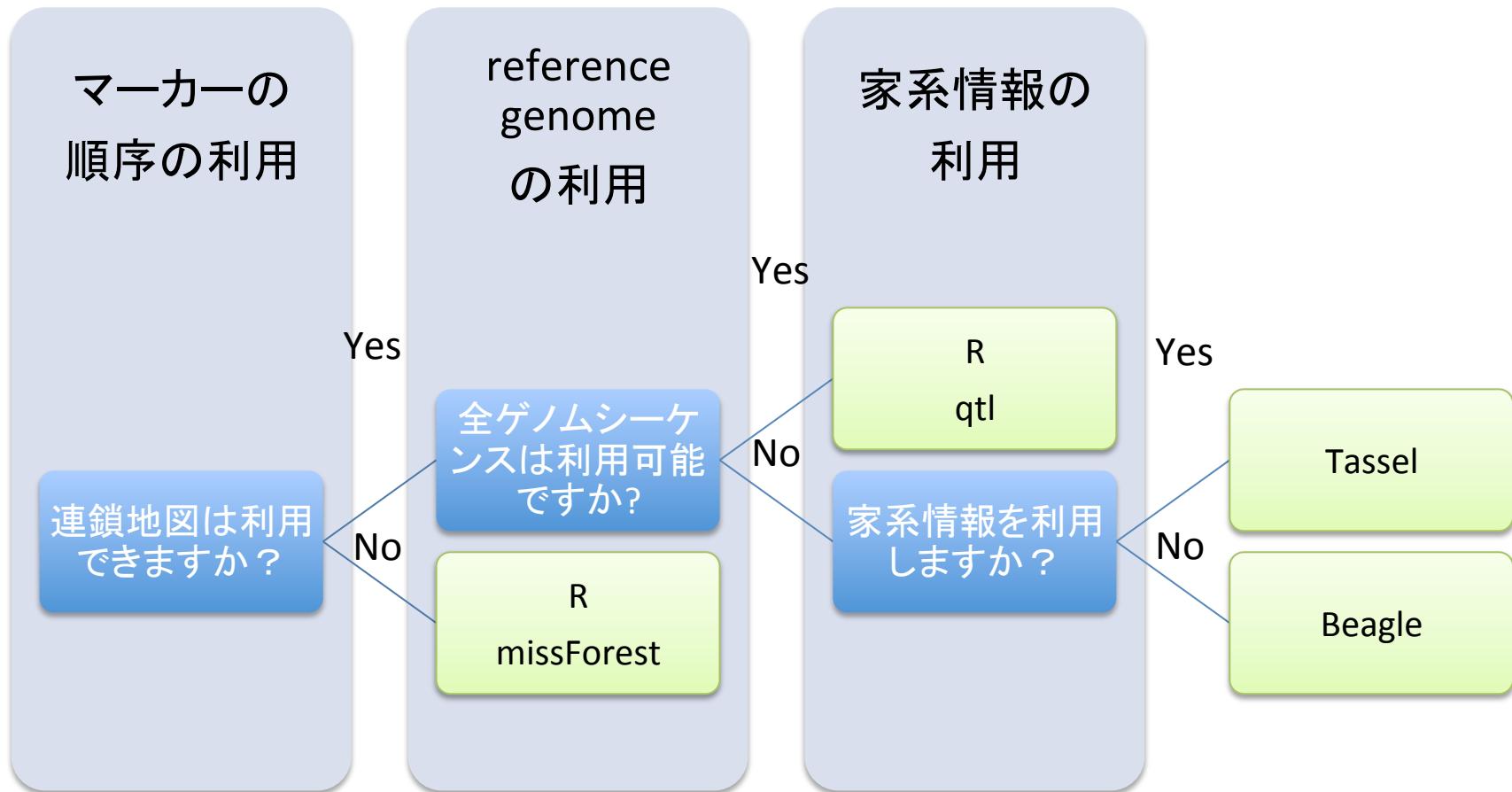
# データのフィルタリング

imputationを行う前に、以下のようなマーカーを除去

- 欠測率の高いマーカー
- 分離比が歪んでいるマーカー
- 実験サンプル間で不一致が多いマーカー
- MAFが非常に低いマーカー

どのような条件でマーカーを除去するか？  
最適な条件はサンプルごとに異なるので、  
それぞれの研究に応じて適切な条件を用いる

# Imputationに用いるソフトの選択の例



どのソフトが適しているのか？はサンプルごとに異なる  
いくつかのソフトで解析し、最適なソフトを選択する

## R missForest

**missForest: Nonparametric Missing Value Imputation using Random Forest**

<https://cran.r-project.org/web/packages/missForest/missForest.pdf>

- ノンパラメトリック
- mixed-type imputation method
- random forestを用いて、実測値から欠測値を予測

## R qtl

<http://www.rqtl.org/manual/qtl-manual.pdf>

地図距離に基づいて、欠測値を予測

- **calc.genoprob** : *Calculate conditional genotype probabilities*

# Beagleとは？

imputation, genotype calling, genotype phasing, IBD segment detectionを行う

- 2015.11.11現在の最新版はBeagle version 4.1
- <http://faculty.washington.edu/browning/beagle/beagle.html>
- Beagle 4.1の場合、Java version 8が必要

## wget を使う方法

```
wget http://faculty.washington.edu/browning/beagle/beagle.21Oct15.abc.jar
```

## HPからのBeagleのダウンロード

<https://faculty.washington.edu/browning/beagle/beagle.html#download>

### Beagle version 4.1

Program: beagle.21Oct15.abc.jar  
Author: Brian Browning  
Email: [browning@uw.edu](mailto:browning@uw.edu)

### Contents

- [Introduction](#)
- [Download Beagle 4.1](#)

# Beagleの入力ファイル

- vcfおよびvcf.gzを利用可能
- Beagleで利用するためには**GT**あるいは**GL**のFORMATが必要

## GT FORMATの例

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	RTx430	Tx642
Chr01	236ss.3	C	T		2256.5	PASS	AC=2;AF=0.043	GT:AD:DP:GQ:PL	0/0 14,0:14:33:0,33,423	0/0 143,0:143:99:0,376,5016
Chr01	284ss.6	T	A		5219.94	PASS	AC=6;AF=0.130	GT:AD:DP:GQ:PL	0/0 14,0:14:36:0,36,491	0/0 135,0:135:99:0,370,4920
Chr01	871ss.10	C	T		24370.1	PASS	AC=32;AF=0.696	GT:AD:DP:GQ:PL	1/1 0,10:10:24:328,24,0	0/0 88,0:88:99:0,244,3212

**GT**: genotype, encoded as allele values separated by either of / or |. The allele values are 0 for the reference allele (what is in the REF field), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on.

**GL** : genotype likelihoods comprised of comma separated floating point log10-scaled likelihoods for all possible genotypes given the set of alleles defined REF and ALT fields.

# Beagleの実行

データサイズが大きい場合は、染色体ごとにvcfファイルを分けて、実行する

## Format GT vcf file

```
java -jar beagle.21Oct15.abc.jar gt="test.sorghum.Nov11.abc.vcf.gz" out="out.gt"
```

test.sorghum.Nov11.abc.vcf.gz

phytozome sorghum v.2.1 SNP数 3,699,951 22系統 のサイズを小さくしたvcf file  
gt=で入力するvcf fileを指定、 out=で出力ファイルを指定

## Format GL vcf file

```
java -jar beagle.21Oct15.abc.jar gl="test.21Oct15.abc.vcf.gz" out="out.gl"
```

gl=で入力するvcf fileを指定

## リファレンスパネルを利用した欠測の補完

```
java -jar beagle.21Oct15.abc.jar ref=ref.21Oct15.abc.vcf.gz gt=target.21Oct15.abc.vcf.gz  
out=out.ref
```

ref=でリファレンスパネルのvcf fileを指定

# Beagleの実行結果

出力ファイル out.gt.vcf.gz ファイルを解凍

```
gunzip out.gt.vcf.gz
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	RTx430	Tx642	Ajabsido	SC35	SC971	SC265	SC283
Chr01	236ss.3	C	T	.	.	PASS	.	GT:DS	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0
Chr01	284ss.6	T	A	.	.	PASS	.	GT:DS	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0	1 1:2	0 0:0
Chr01	871ss.10	C	T	.	.	PASS	.	GT:DS	1 1:2	0 0:0	1 1:2	0 1:1	1 1:2	1 1:2	0 1:1

imputation前の入力ファイルは 0/0,0/1,1/1

imputation後の出力ファイルは 0|0,0|1,1|1

GT: 0がREF alleleで、1がALT allele

0/0はREFのホモ

0/1はREFとALTのヘテロ

1/1はALTのホモ

| はphasedを示す  
/ はunphasedを示す

# Beagleの実行結果

## imputation後のvcfファイル

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	RTx430	Tx642	Ajabsido	SC35	SC971	SC265	SC283
Chr01	236	ss.3	C	T	.	PASS	.	GT:DS	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0
Chr01	284	ss.6	T	A	.	PASS	.	GT:DS	0 0:0	0 0:0	0 0:0	0 0:0	0 0:0	1 1:2	0 0:0
Chr01	871	ss.10	C	T	.	PASS	.	GT:DS	1 1:2	0 0:0	1 1:2	0 1:1	1 1:2	1 1:2	0 1:1



0/0はREFのホモ -> 0  
0/1はREFとALTのヘテロ -> 1  
1/1はALTのホモ -> 2

GWASやゲノミックセレクションなど、様々な解析に用いるためにスコア化

marker	RTx430	Tx642	Ajabsido	SC35	SC971	SC265	SC283
Chr01:236	0	0	0	0	0	0	0
Chr01:284	0	0	0	0	0	2	0
Chr01:871	2	0	2	1	2	2	1

# Tasselとは？

## *Trait Analysis by aSSociation, Evolution and Linkage*

<http://www.maizegenetics.net/#!tassel/c17q9>

- 作物の解析に最適化
- 欠測の補完だけではなく、様々な機能がある
- コマンドラインからだけではなく、充実したGUIから簡単に解析可能

The screenshot shows the Buckler Lab for Maize Genetics and Diversity website. At the top, there is a banner with a corn cob background and the text "Buckler Lab for Maize Genetics and Diversity". Below the banner is a navigation menu with links to Home, Bioinformatics, Publications, People, Research, and Contact Us. The main content area features a large red stylized letter 'T' followed by the text "TASSEL 5.0". Below this, there is a link to "Tassel Version 5.0 (*Getting Started!*)" and a note "(Build: October 15, 2015 Requires: Java 1.8)". There are also links for "Tassel 5 Mac OS", "Tassel 5 Windows 64 Bit", "Tassel 5 Windows 32 Bit", and "Tassel 5 UNIX". To the right, there is a screenshot of the "Alignment Viewer" software interface, which displays a grid of data. The bottom left corner of the slide contains the text "2015/11" and the bottom right corner contains the number "24".

# Tasselのダウンロード

HPからのダウンロード  
java 1.8

<http://www.maizegenetics.net/#!tassel/c17q9>

Tassel Version 5.0 (*Getting Started!*)  
(Build: October 15, 2015 Requires: Java 1.8)

[Tassel 5 Mac OS](#)  
[Tassel 5 Windows 64 Bit](#)  
[Tassel 5 Windows 32 Bit](#)  
[Tassel 5 UNIX](#)

Tassel Version 5.0 Standalone  
(GBS Pipeline V2 - Preferred Version)

[Using Git - Recommended!](#)  
[Download \(Click on "Tags"\)](#)



ここからOSに合わせてダウンロード

YouTube JP

1. TASSEL installation and increasing heap size

Panzea

チャンネル登録 55 視聴回数 615 回

The screenshot shows a YouTube video player with the following details:

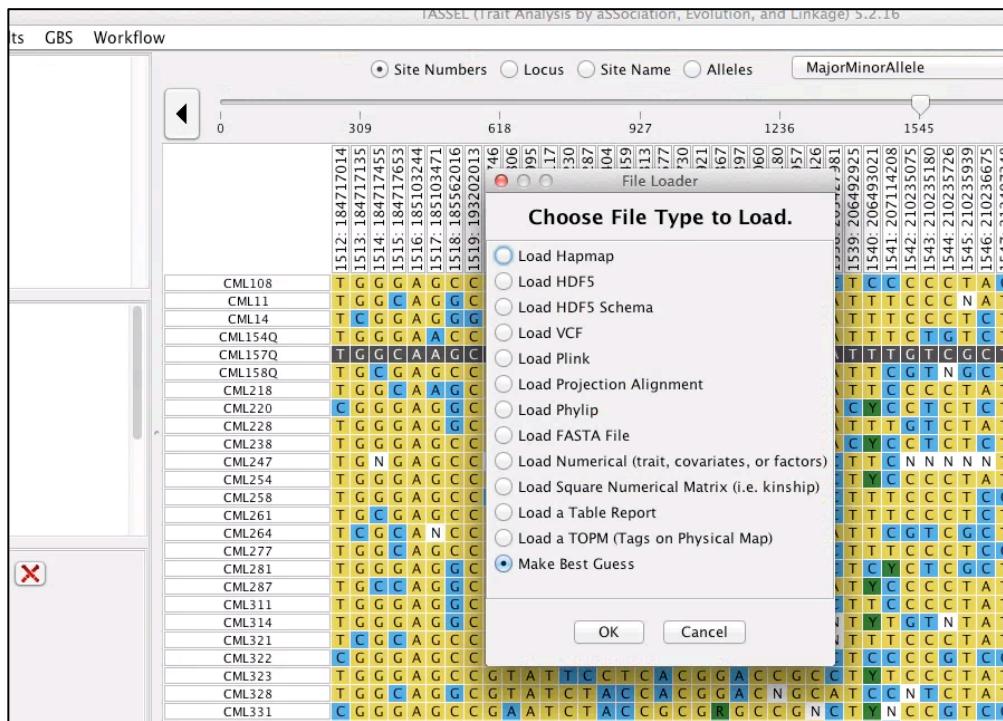
- Video title: 1. TASSEL installation and increasing heap size
- Uploader: Panzea
- Views: 615
- Comments: 55

The video content itself shows a close-up view of a maize field with young plants growing in rows.

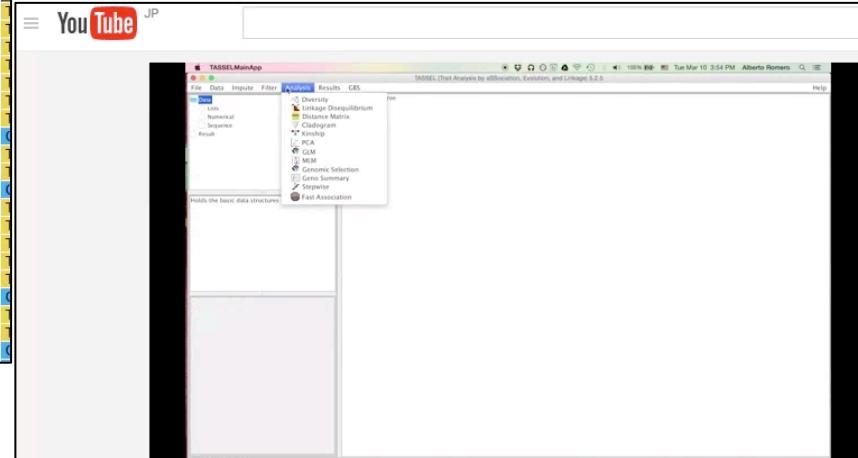
YouTubeでの解説

# Tasselの入力ファイル

メニューのDataからLoadを選択 → Make Best Guess  
vcf fileを入力ファイルとして利用



vcf fileだけでなく、Hapmapや  
Plinkなどのファイルを利用可能



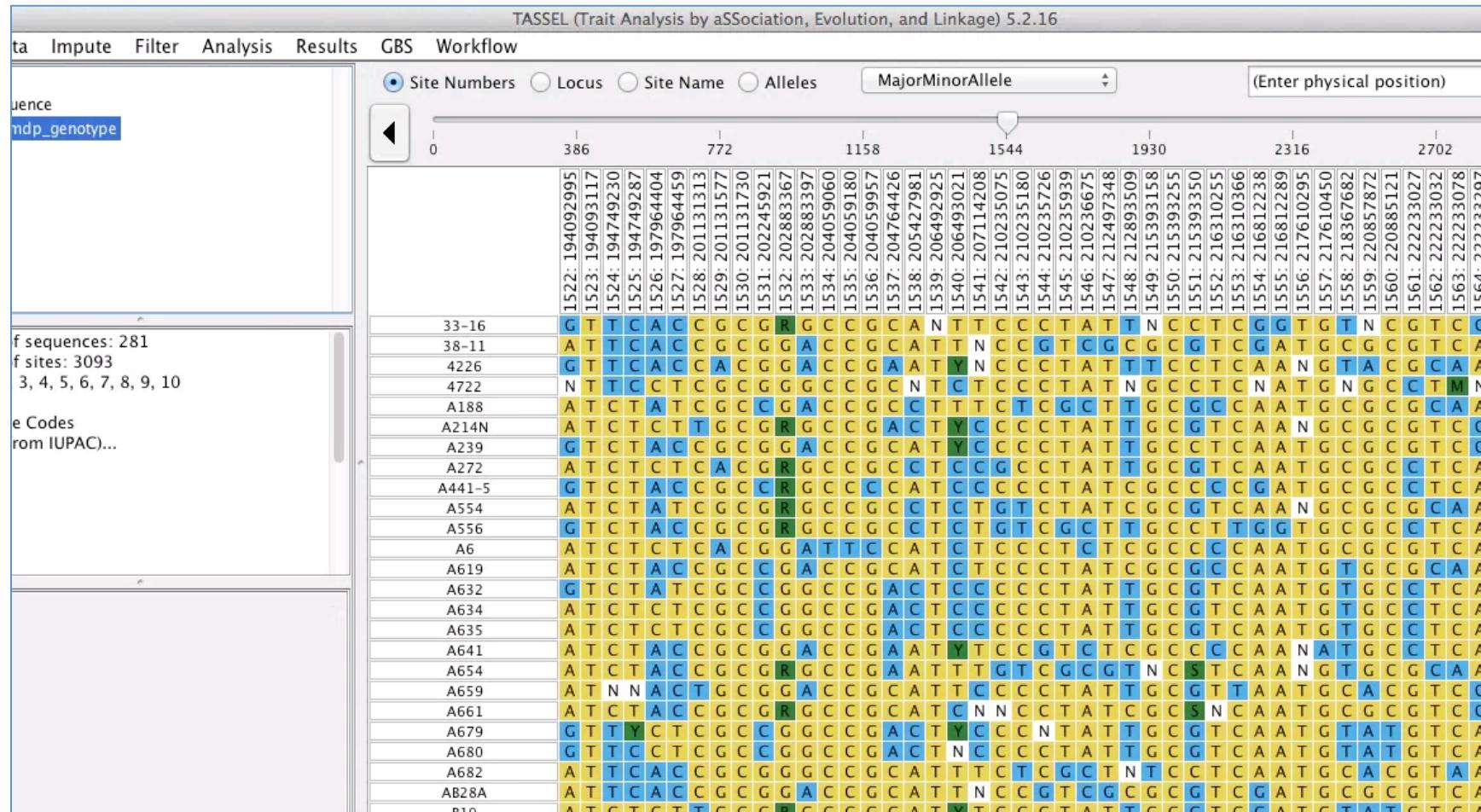
## 2. TASSEL menus



視聴回数 279 回

YouTubeでの解説

# Tassel 遺伝子型データの表示



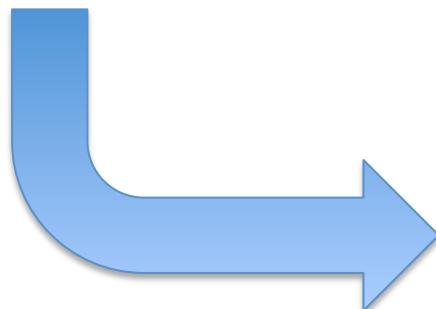
Major alleleとMinor alleleを色分けして表示させることで、データの確認が可能

# Tassel 染色体ごとに分けたファイルを作成

The screenshot shows the TASSEL software interface. The menu bar includes File, Data, Impute, Filter, Analysis, Results, GBS, and Workflow. The 'Data' menu is open, displaying various options like Load, Export, Get Taxa List, etc., with 'Separate' highlighted and circled in blue. Below the menu is a table of genotype data for several individuals (RTx430, Tx642, Ajabsido, SC35, SC971, SC265, SC283, Segaolane, Macia, SC1345) across different genomic positions (Site Numbers). The table uses color coding for alleles (e.g., A, T, C, G).

Dataから  
Separateを選択

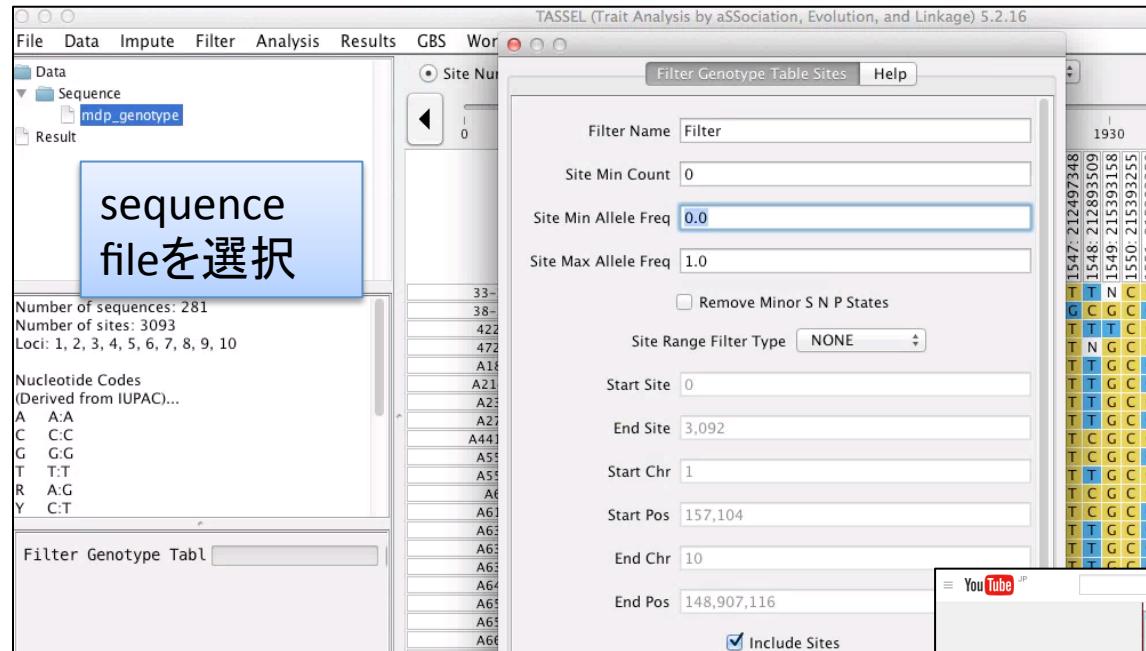
元のファイルの下に  
染色体ごとに分けた  
ファイルが作成される



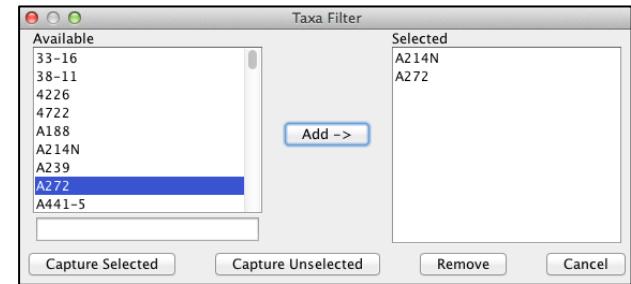
The screenshot shows a file browser window with a 'Sequence' folder expanded. Inside the 'Sequence' folder are files named mdp\_genotype, Sbicolor\_255, Sbicolor\_255\_chrom01 (which is highlighted in blue), Sbicolor\_255\_chrom02, and Sbicolor\_255\_chrom03. There is also a 'Result' folder.

# Tassel データのフィルタリング

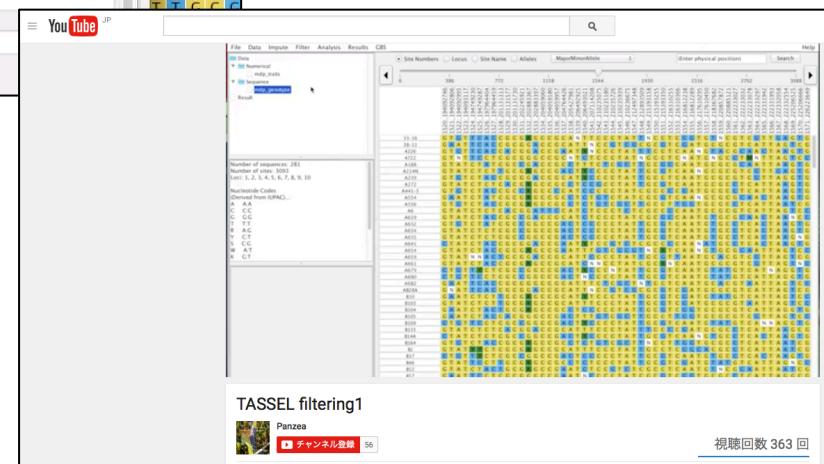
Filter → Filter Genotype Table Sitesを選択



Filter → Taxa Names  
系統名を検索しながら  
フィルタリングが可能



MAFでフィルタリング



YouTubeでの解説

# Tassel LD plotの作成

Step 1

The screenshot shows the Tassel software interface. The menu bar includes Data, Impute, Filter, Analysis, Results, GBS, and Workflow. The Analysis menu is highlighted. A blue oval surrounds the "Linkage Disequilibrium" option. On the left, there's a tree view of project files under "Data" and "Sequence". The main panel displays sequence data with columns for Locus, Position, Site, Number of sites, States, and Frequency. Below the table, a message box shows: "Number of sequences: 281", "Number of sites: 3093", "Loci: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10", "Nucleotide Codes (Derived from IUPAC)", and "A C T G". A blue box highlights the text "Analysis → Linkage Disequilibriumを選択".

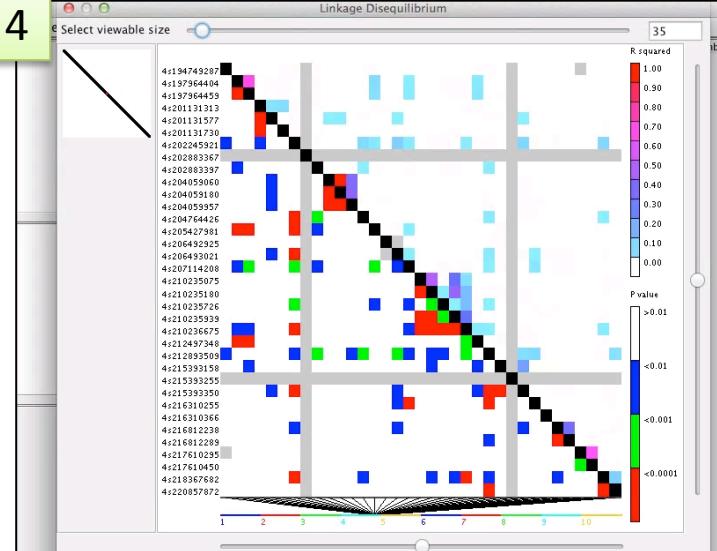
Step 2

The screenshot shows the Tassel software interface with the Results menu highlighted. A blue oval surrounds the "LD" option under the "Result" folder. The main panel displays a table titled "Table Title: Linkage Disequilibrium" with columns for Locus, Position, Site, Number of sites, States, and Frequency. A blue box highlights the text "Result フォルダにLDの結果が表示される".

Step 3

The screenshot shows the Tassel software interface with the Results menu highlighted. A blue oval surrounds the "LD Plot" option under the "Result" folder. The main panel displays a table titled "Table Title: Linkage Disequilibrium" with columns for Locus, Position, Site, Number of sites, States, and Frequency. A blue box highlights the text "Result → LD plot".

Step 4



LD plot が表示される

# Tassel Phenotype の欠測を補完

**TASSEL (Trait Analysis by aSSociation)**

File Data Impute Filter Analysis Results GBS Workflow

Data Numerical mdp\_phenotype Result

Table Title: mdp\_phenotype  
Number of columns: 8  
Number of rows: 563  
Number of elements: 4504

Taxa	location	EarHT	dpoll	EarDia	EarWT	dpoll
33-16	A	64.75	64.5			64.75
38-11	A	92.25	68.5			92.25
4226	A	65.5	59.5			59.5
4722	A	81.13	71.5			71.5
A188	A	27.5	62			62
A214N	A	65	69			69
A239	A	47.88	61			61
A272	A	35.63	70			70
A441-5	A	53.5	67.5			67.5
A554	A	38.5	66			66
A556	A	28	65			65
A6	A	109.5	80.5			80.5
A619	A	36	61			61
A632	A	60	61			61
A634	A	54	59			59
A635	A	37	64			64
A654	A	39	64			64
A659	A	46.5	58.5			58.5
A661	A	51.5	59			59
A679	A	65	66			66
A680	A	68	65.5			65.5
A682	A	47	57.5			57.5

**Extract Inbred Haplotypes by FILLIN**  
**Impute By FILLIN**  
**Impute By FSFHap**  
**Numerical Impute**  
**Remove indels for input to Beagle v.4**  
**LD KNNi Imputation**  
**Evaluate Imputation Accuracy**

Impute → Numerical Imputeを選択

A441-5

欠測を補完する方法を選択

Imputation by mean  
Number of nearest neighbors to be evaluated: 5  
Choose Distance type: Euclidean

Ok Cancel Defaults

元のファイルの下に  
Imputed ファイルが作成される

TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.2.16

File Data Impute Filter Analysis Results GBS Workflow

Data Numerical mdp\_phenotype imputed\_mdp\_phenotype Result

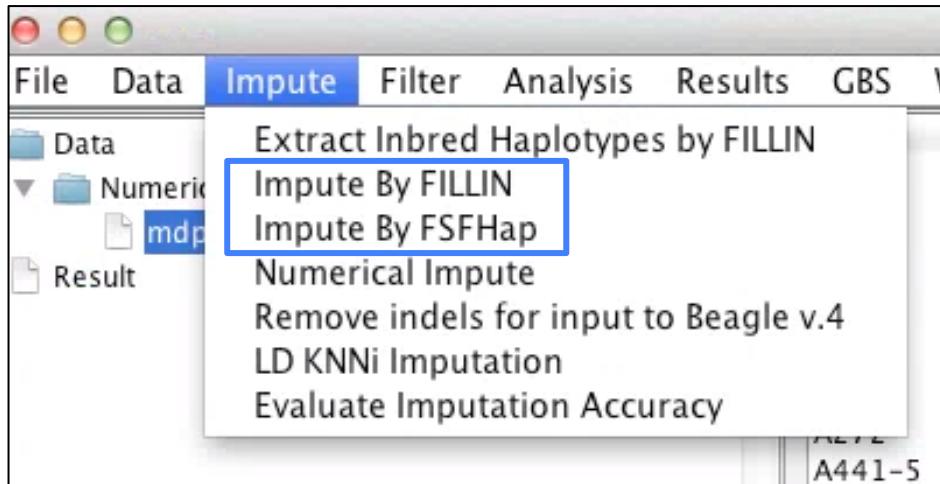
Table Title: Phenotype  
Number of columns: 8  
Number of rows: 563  
Number of elements: 4504

Taxa	location	EarHT	dpoll	EarDia	EarWT	dpoll	Q1	Q2	Q3
33-16	A	64.75	64.5			64.5	0.014	0.972	0.014
38-11	A	92.25	68.5	37.897	0.003	68.5	0.003	0.993	0.004
4226	A	65.5	59.5	32.219	0.071	59.5	0.071	0.917	0.012
4722	A	81.13	71.5	32.421	0.035	71.5	0.035	0.854	0.111
A188	A	27.5	62	31.419	0.013	62	0.013	0.982	0.005
A214N	A	65	69	32.006	0.21	69	0.21	0.997	0.002
A239	A	47.88	61	36.064	0.002	61	0.002	0.963	0.003
A272	A	35.63	70	37.062	0.059	70	0.059	0.959	0.019

Missing values imputed as mean of trait.  
from mdp\_phenotype

Taxa	location	EarHT	dpoll	EarDia	EarWT	dpoll	Q1	Q2	Q3
33-16	A	64.75	64.5			64.5	0.014	0.972	0.014
38-11	A	92.25	68.5	37.897	0.003	68.5	0.003	0.993	0.004
4226	A	65.5	59.5	32.219	0.071	59.5	0.071	0.917	0.012
4722	A	81.13	71.5	32.421	0.035	71.5	0.035	0.854	0.111
A188	A	27.5	62	31.419	0.013	62	0.013	0.982	0.005
A654	A	38.5	66	37.062	0.059	66	0.059	0.959	0.019
A659	A	28	65	38.846	0.003	65	0.003	0.993	0.004
A661	A	109.5	80.5	39.323	0.037	80.5	0.037	0.997	0.002
A679	A	54	59	42.471	0.11	59	0.11	0.997	0.003
A680	A	37	64	41.152	0.004	64	0.004	0.997	0.003
A682	A	51.5	59	35.928	0.001	59	0.001	0.997	0.002
A828A	A	47	57.5	32.504	0.222	57.5	0.222	0.997	0.003
B10	A	65.5	66	36.561	0.001	66	0.001	0.997	0.002
B103	A	68	65.5	37.062	0.008	65.5	0.008	0.997	0.003
B104	A	65	64	44.773	0.001	64	0.001	0.997	0.002
B105	A	66	65.5	56.25	0.037	65.5	0.037	0.997	0.003
B109	A	67	66	64.5	0.002	66	0.002	0.997	0.003
B115	A	68	65.5	64.5	0.092	65.5	0.092	0.997	0.003

# Tassel Genotype の欠測を補完



Tasselには2種類のImpute方法

集団に合わせて使い分ける

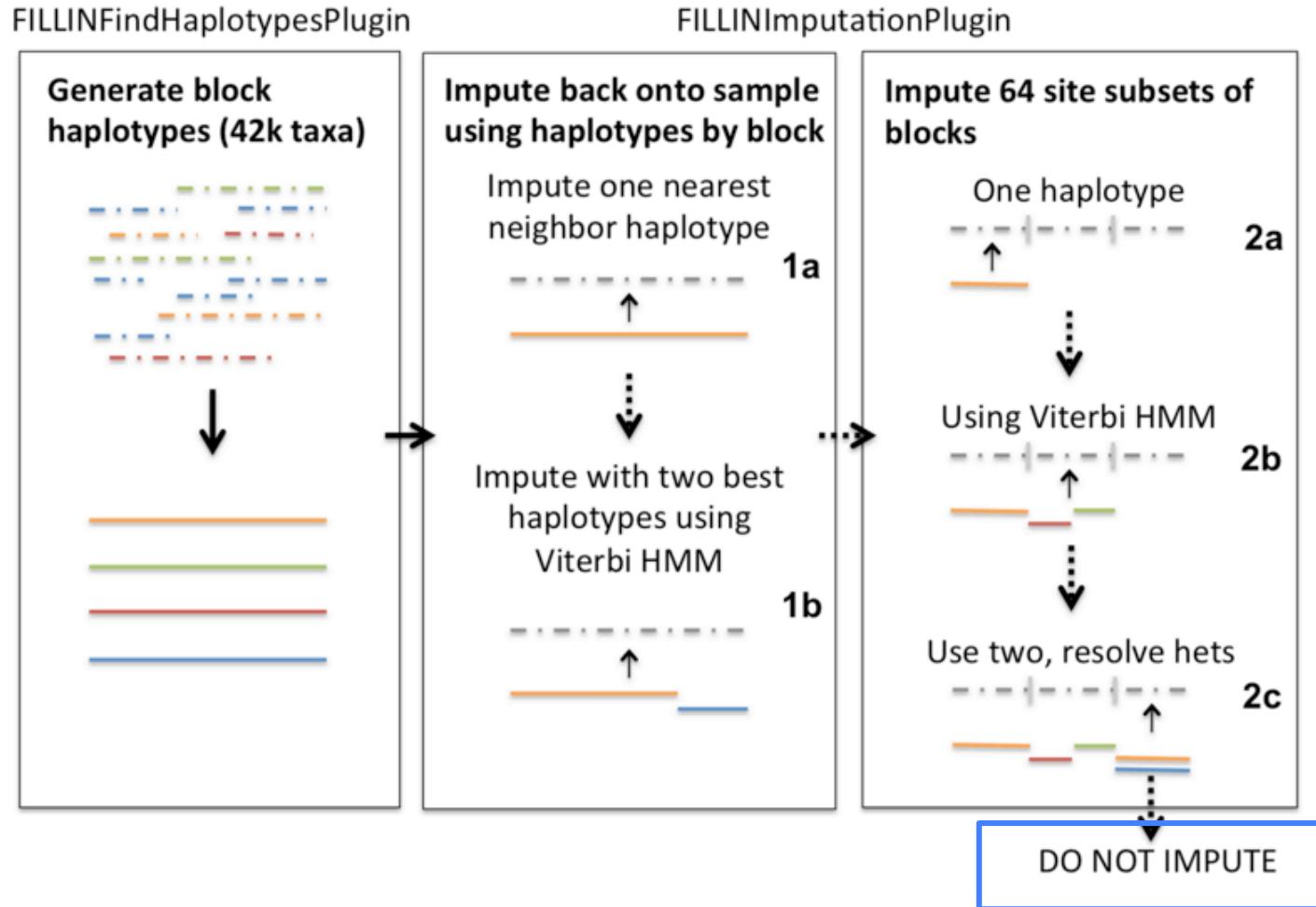
## Impute By FILLIN

Fast, Inbred Line Library ImputatioN  
generalized approach

## Impute BY FSFHap

impute missing data in full sib families (bi-parental families)

# Tassel - Impute By FILLIN



<https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual/FILLIN/FILLIN> より引用

Beagleとは異なり、欠測を埋められないマーカーがある

# Tassel - Impute BY FSFHap

- Inbred の両親と後代のImputation
- 欠測率が高く、ヘテロ率の高いGBSデータ用に開発
- 親の遺伝子型と後代の遺伝子型データが必要
- ヘテロの両親の $F_1$ には利用できない
- 両親の遺伝子型が正確である場合にはこれを利用した方が良い

<https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual/FSFHaplImputation/FSFHaplImputation>

# Genotype-Imputation Accuracy

Consensus imputation improves accuracy

	R <sup>2</sup>		
	Diverse Landraces	Diverse Inbreds	Temperate Inbreds
FILL IN	0.488	0.988	0.981
Beagle	0.662	0.883	0.942
Both agree	0.713	0.994	0.994
Beagle, when FILL IN does not impute	0.626	0.767	0.833

# Genotype-Imputation Accuracy

- 正確に遺伝子型を補完することができなければ、その後の解析に影響する
- 遺伝子型に矛盾がないか？を確認することでimputationの正確性を調べることが可能
- 既知の遺伝子型をマスクすることで、遺伝子型の正確性を解析  
既知の遺伝子型と補完された遺伝子型を比較
- Tasselの場合はすべての遺伝子型を補完することができない  
このため、Tasselで遺伝子型を予測したのち、補完できなかった遺伝子型をBeagleを用いて補完する方法もある
- BeagleとTasselを両方用いて、一致する遺伝子型だけを利用することで正確性が高くなる
- Inbredに関してはTasselの方が優れているが、ヘテロな集団の場合はBeagleの方が適している