




SVSを用いたジェノタイプインピュテーション

フィルジエン株式会社 バイオインフォマティクス部
(biosupport@filgen.jp)

- SNPジェノタイプデータにおいて、未観測SNPを推定し、データの補完を行う手法をジェノタイプインピュテーションとよび、GWAS解析に用いるSNP数を増やすために用いられる
- ジェノタイプインピュテーションを利用することで、少数のSNPをもとに、全ゲノムシーケンスに匹敵する多数のSNPのデータを得ることができ、実験コストの削減や、プラットフォームの異なるSNPマイクロアレイの統合解析などが可能になる
- Golden Helix社SNP & Variation Suite (SVS)ソフトウェアでは、ジェノタイプインピュテーション用のツールが組み込まれており、簡単な操作で実行できる

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

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

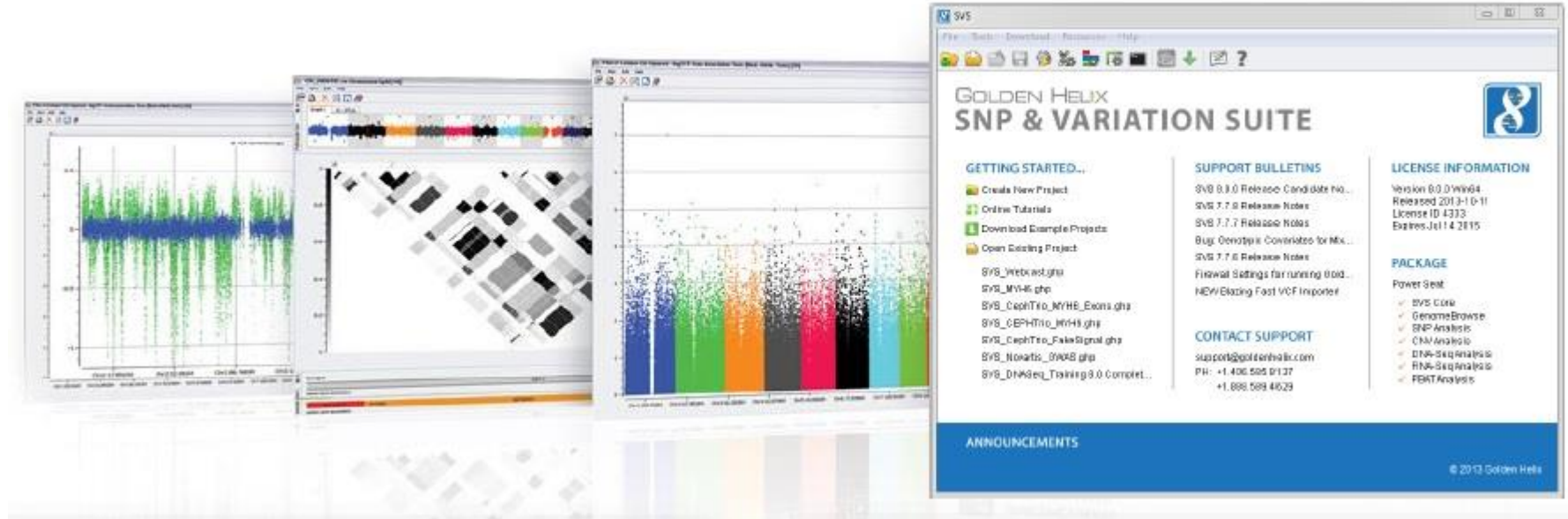
 T --- G --- G --- A --- T
T --- G --- T --- C --- C

ジェノタイプインピュテーション

 AGCTGTATCGATAAC
AGATCCATCGCTATT

 TGATCCAGCTATAAC
ACCTCTATCGATGAT

 TGATGTAGCTATGAT
TCCTGCATCGCTAAC

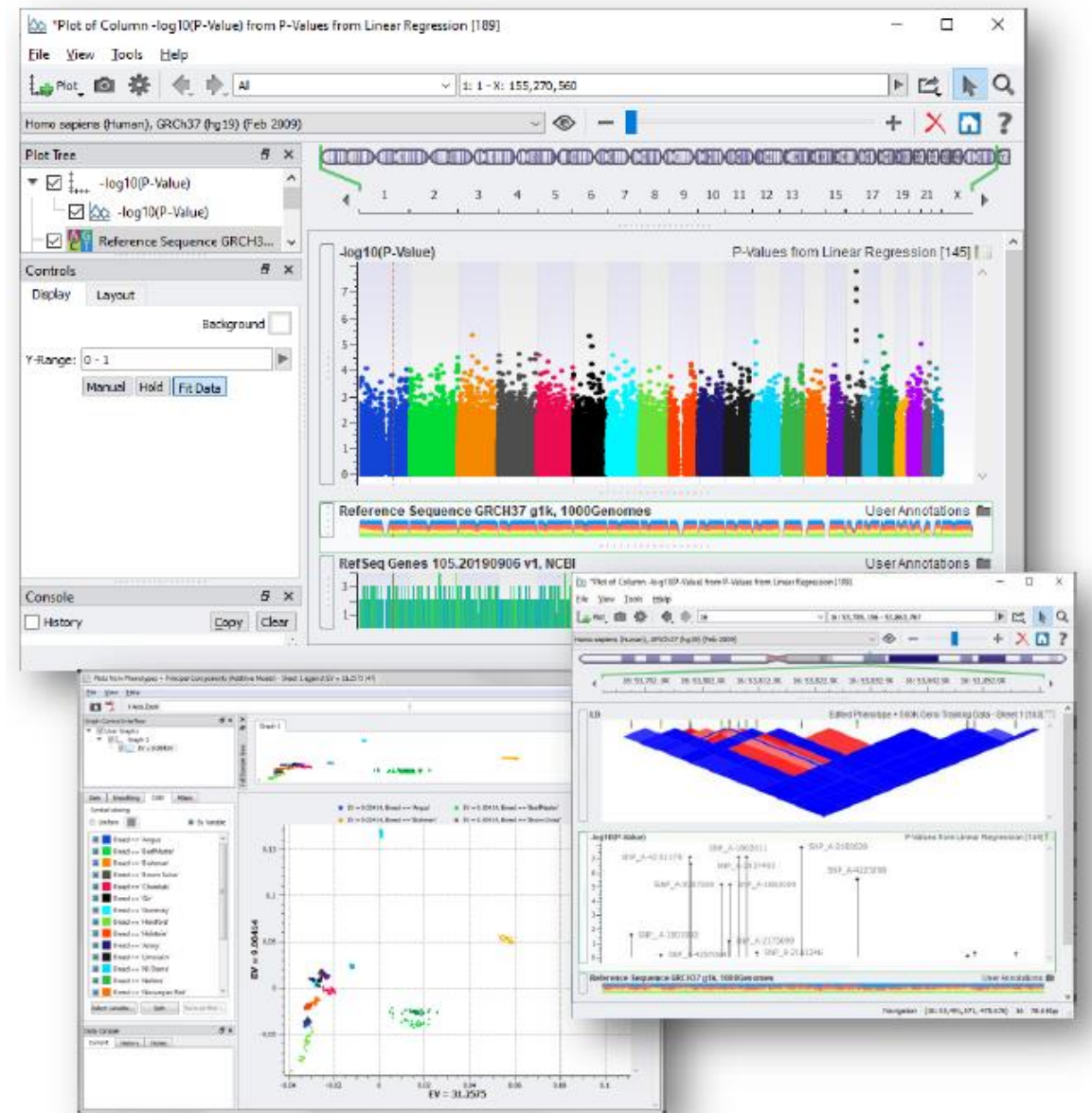
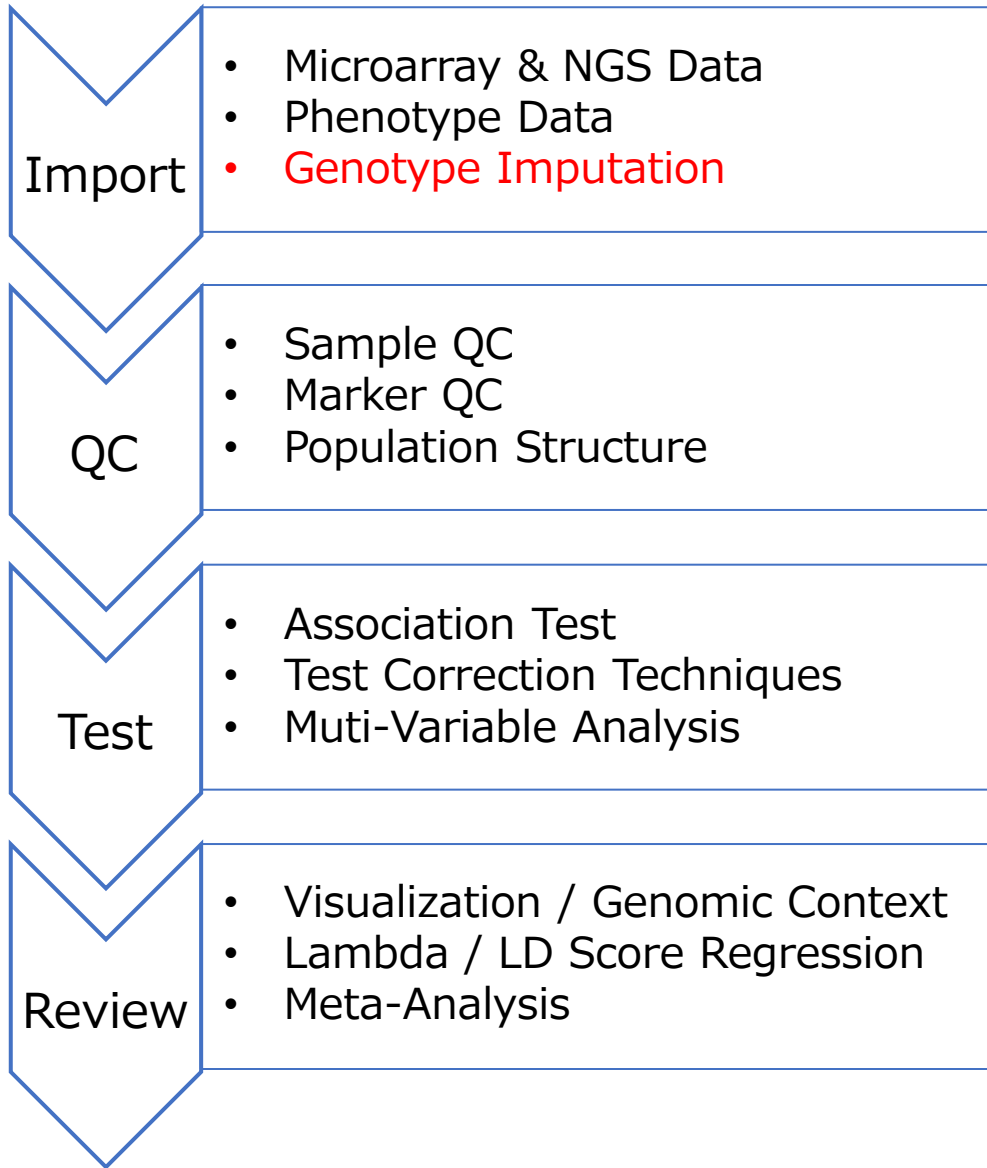


Core Features

- Powerful Data Management
- Rich Visualizations (GenomeBrowse)
- Robust Static
- Flexible

Applications

- Genotype Analysis
- Agrigenomics Analysis
- DNA/RNA Sequence Analysis
- CNV Analysis

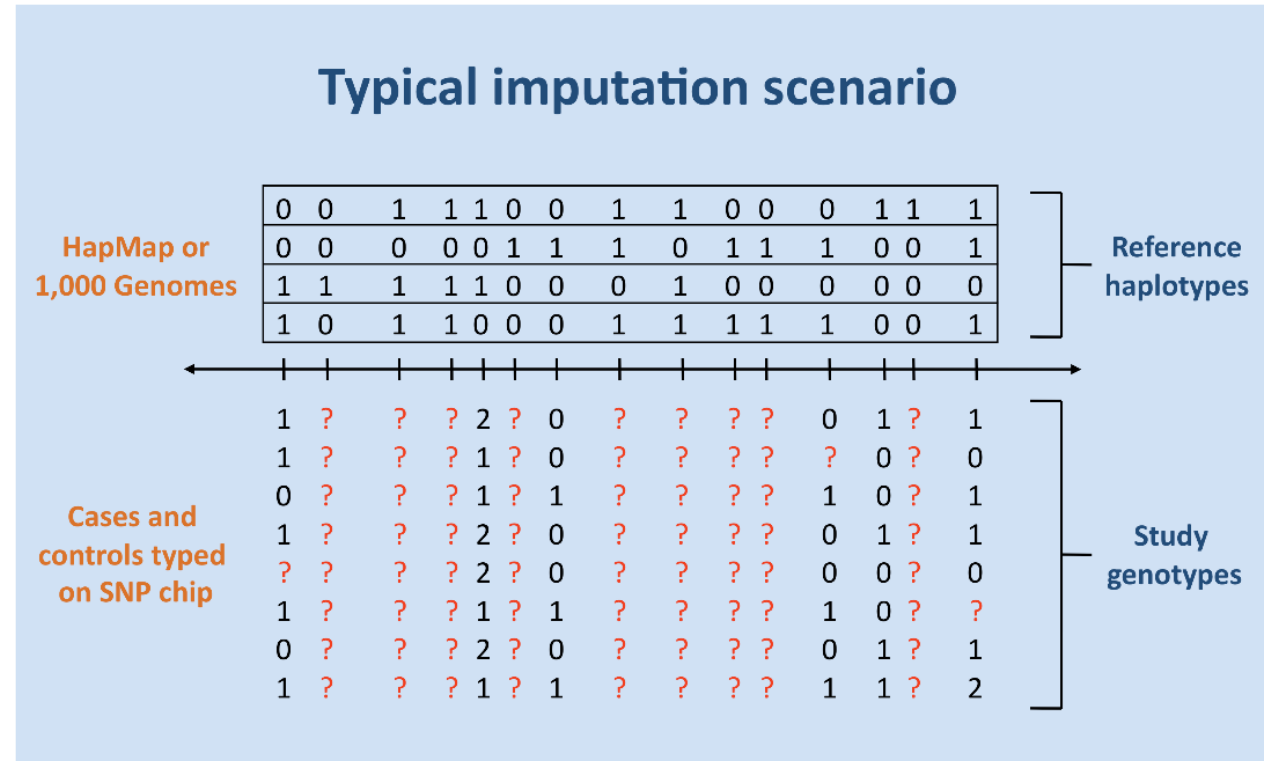


■ インピュテーションを行うことのメリット

- 欠損データの補完
- 異なるプラットフォームのSNPマイクロアレイデータの統合解析
- SNPマーカーの高密度化による、検出力と解像度の向上
- SNPマイクロアレイ搭載プローブ数の節約によるコスト削減

■ SVS上でインピュテーションを行うことのメリット

- グラフィカルインターフェースによる簡単な操作
- Windows, Mac, Linuxコンピュータでの実行が可能
- 多数のSNPデータファイルフォーマットのサポート (PLINK, Illumina, Affymetrix, VCFなど)
- インピュテーション後のGWAS解析へのシームレスな移行





研究用サンプルデータ

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

リファレンスパネル (ハプロタイプデータ)



AGCTGTATCGATAAC
AGATCCATCGCTATT
TGATCCAGCTATAAC
ACCTCTATCGATGAT
TGATGTAGCTATGAT
TCCTGCATCGCTAAC

インピュテーションデータ

AGCTGTATCGATAAC
AGATCCATCGCTATT

- ジェノタイプインピュテーションの実行には、全ゲノムシーケンスなどによって得られた、リファレンスサンプルのSNPハプロタイプデータがリファレンスパネルとして必要
- SNPマイクロアレイなどより得られた研究用サンプルのジェノタイプデータと、リファレンスパネルのハプロタイプデータを比較することで、研究用サンプルで未観測のSNPジェノタイプの推測を行う

- ジェノタイプデータのフェーシングとインピュテーションの実行プログラムとして、BEAGLE 4.1と4.0の2種類のアルゴリズムを搭載
 - BEAGLE 4.1: 独立したサンプルデータ用
 - BEAGLE 4.0: 家系サンプルデータ用
- 100,000人以上の多人数のジェノタイプデータの処理に対応
- オリジナルのプログラムと同様に、計算時の反復回数やサンプルサイズなどの条件を指定可能

Phasing and Imputation Algorithm

Beagle 4.1

4.1 Phasing Iterations:

Max Cluster Size in CM:

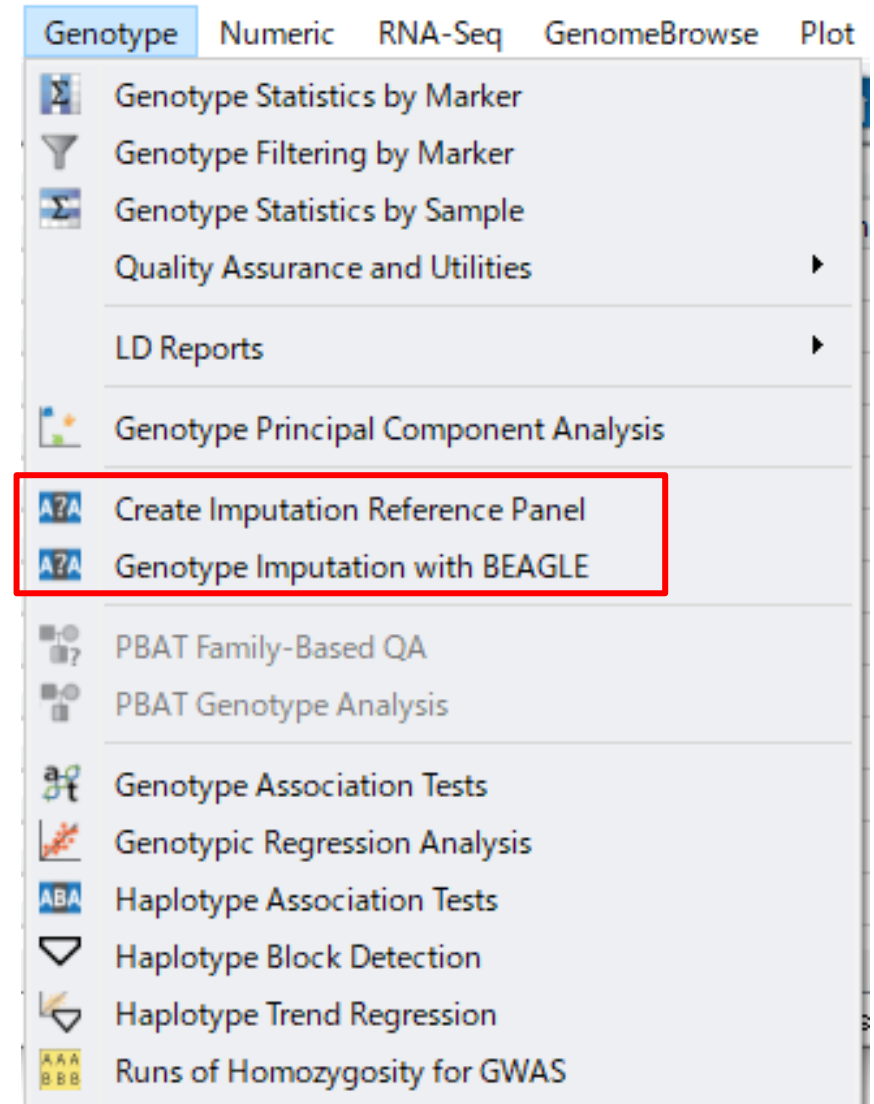
Effective Population Size:

Allele Miscall Rate:

Beagle 4.0

Burn-in Iterations:

Phasing Iterations:



レディメイドのリファレンスパネルを用いる場合

リファレンスパネルのダウンロード

- 1000 Genomesプロジェクトのリファレンスパネルダウンロード

カスタムリファレンスパネルを用いる場合

リファレンスパネルの構築

- リファレンスサンプルのVCFファイルインポート
- ハプロタイプフェージング

リファレンスサンプル
データの処理

データインポート

- SNPマイクロアレイデータのインポート
- マーカーマップデータの付加
- 異なるプラットフォームのデータの統合

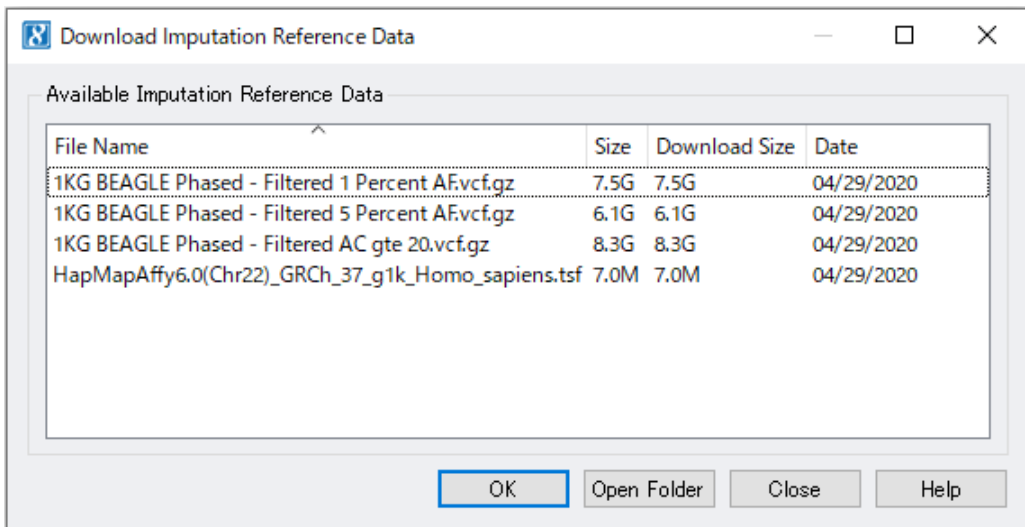
アレルのエンコード

- ABアレルのAGCT変換

研究用サンプルデータの処理

インピュテーションの実行

- BEAGLEプログラムの選択
- 確率データ出力などのオプション設定



- SVSのダウンロードツールより、1000 Genomesプロジェクトのリファレンスパネルをダウンロード可能
- 日本人を含む、様々な人種の2500人分の全ゲノムシーケンスより得られたSNPデータ
- ハプロタイプフェージング実行済みのため、ダウンロード後ただちにインピュテーションに使用可能

- BEAGLEのオリジナルウェブサイトからも、同じサンプルセットのデータファイルをダウンロード可能
- 染色体ごとにファイルが分割されており、染色体別にインピュテーションを行う際に利用
- ハプロタイプフェージング実行済みのため、ダウンロードしたファイルを所定のフォルダにコピーするだけで、インピュテーションに使用可能

Name	Date modified	Type	Size
chr1.1kg.phase3.v5a.vcf.gz	1/5/2017 2:03 PM	GZ File	737,074 KB
chr1.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:00 PM	TBI File	205 KB
chr2.1kg.phase3.v5a.vcf.gz	1/5/2017 2:04 PM	GZ File	788,570 KB
chr2.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:00 PM	TBI File	217 KB
chr3.1kg.phase3.v5a.vcf.gz	1/5/2017 2:04 PM	GZ File	671,966 KB
chr3.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:01 PM	TBI File	179 KB
chr4.1kg.phase3.v5a.vcf.gz	1/5/2017 2:04 PM	GZ File	690,066 KB
chr4.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:01 PM	TBI File	173 KB
chr5.1kg.phase3.v5a.vcf.gz	1/5/2017 2:47 PM	GZ File	599,745 KB
chr5.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:44 PM	TBI File	162 KB
chr6.1kg.phase3.v5a.vcf.gz	1/5/2017 2:49 PM	GZ File	624,556 KB
chr6.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:44 PM	TBI File	154 KB
chr7.1kg.phase3.v5a.vcf.gz	1/5/2017 2:49 PM	GZ File	557,489 KB
chr7.1kg.phase3.v5a.vcf.gz.tbi	1/5/2017 2:45 PM	TBI File	143 KB

ハプロタイプフェージング実行前のVCFファイル

Sample01	Sample02	Sample03
0/1	0/1	1/1
0/1	0/1	0/1
1/1	0/1	0/1
1/1	1/1	0/1
0/1	0/1	1/1

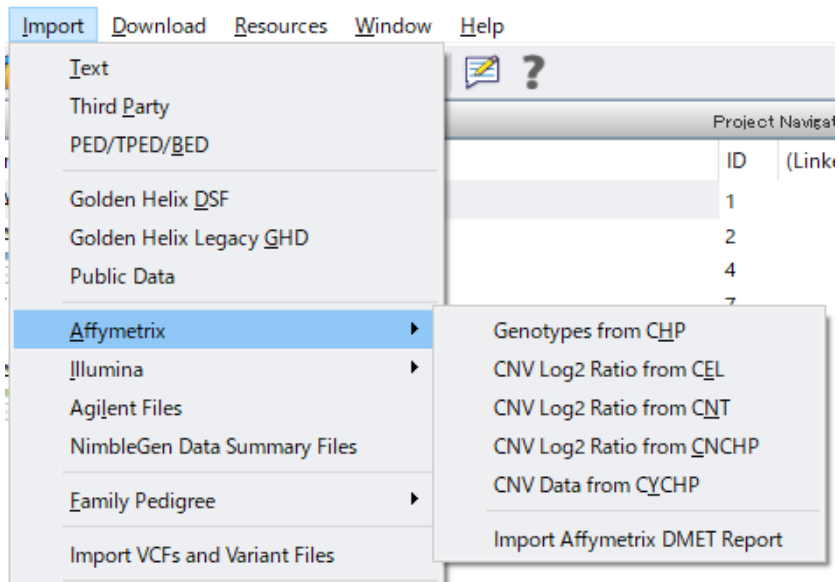
染色体のフェーズが不明

ハプロタイプフェージング実行後のVCFファイル

Sample01	Sample02	Sample03
0 1	0 1	1 1
1 0	1 0	1 0
1 1	0 1	0 1
1 1	1 1	1 0
0 1	0 1	1 1

SNPのフェーズ情報

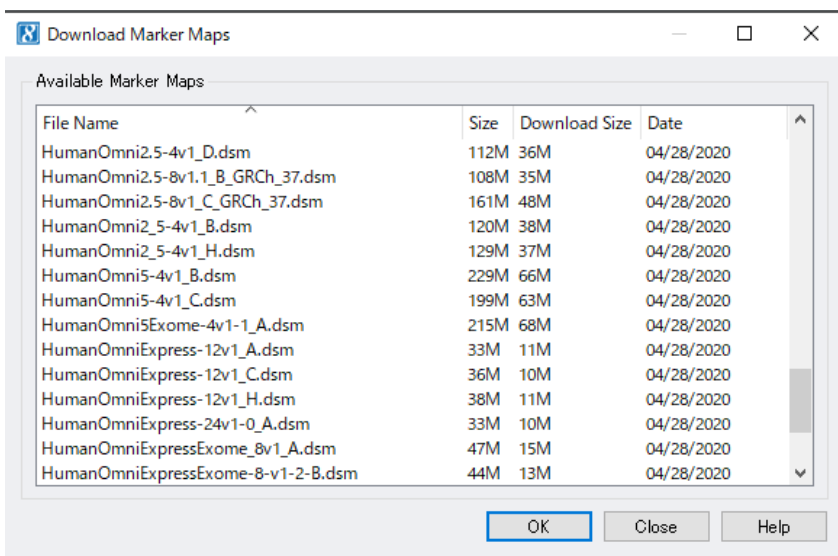
- カスタムリファレンスパネルを作成する場合は、リファレンスサンプルのジェノタイプデータファイル（VCFファイルなど）をインポートした後、ハプロタイプフェージングを実行することが必要
- ハプロタイプフェージングにより、2本の染色体のハプロタイプが決定され、インピュテーションに利用可能となる



ジェノタイプデータファイル

Unsort		G_2897	G_2898	G_2899	G_2900	G_2901
Map	Sample	SNP_A-2188530	SNP_A-1842133	SNP_A-1915605	SNP_A-2247158	SNP_A-2044424
Chromosome		1	1	1	1	1
Position		44764547	44767913	44795461	44807164	44820319
dbSNP RS ID		rs325155	rs325150	rs16831920	rs2524353	rs226067
Associated Gene		ERI3	ERI3	ERI3	ERI3	ERI3
Cytoband		p34.1	p34.1	p34.1	p34.1	p34.1
Reference Alleles A/B		[C/T]	[A/G]	[C/G]	[A/G]	[C/T]
Top Alleles		[C/T]	[T/C]	[C/G]	[T/C]	[G/A]
Bottom Alleles		[G/A]	[A/G]	[G/C]	[A/G]	[C/T]
Strand		+	-	+	-	-
Strand Versus dbSNP		reverse	same	same	reverse	reverse
1	GSM233274	A_A	B_B	B_B	A_B	A_A
2	GSM233272	A_A	B_B	B_B	B_B	A_A
3	GSM233270	A_A	B_B	B_B	B_B	A_A
4	GSM233268	A_A	B_B	B_B	A_B	A_A
5	GSM233266	A_A	B_B	B_B	B_B	A_A
6	GSM233264	A_A	B_B	B_B	A_A	A_A
7	GSM233262	A_A	B_B	B_B	A_A	A_A
8	GSM233260	A_A	B_B	B_B	A_A	A_A
9	GSM233258	A_A	B_B	B_B	A_A	A_A
10	GSM233256	A_A	B_B	B_B	A_B	A_A



ジェノタイプデータ + マーカーマップ



マーカーマップファイル

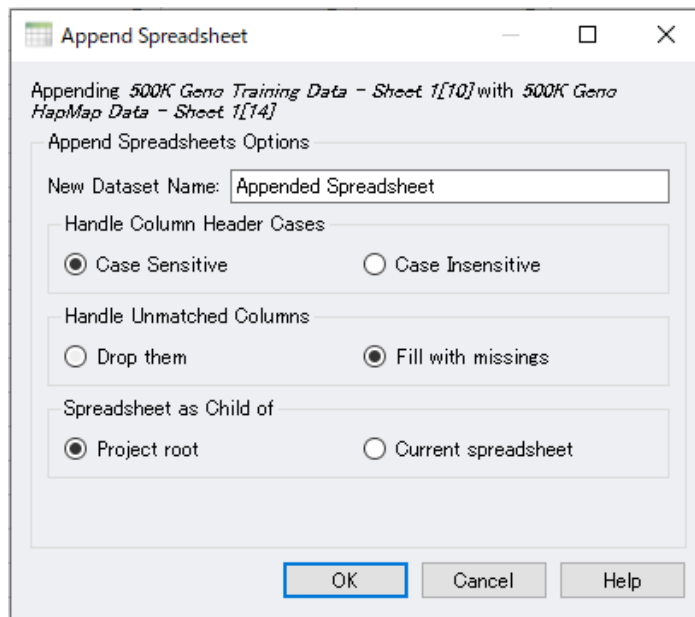
- 研究用サンプルデータとして、SNPマイクロアレイのデータファイルや、PLINK、VCFフォーマットのジェノタイプデータをインポート
- インピュテーションの実行には、各SNPのマーカー情報（染色体上の位置やRS ID）も必要

- マイクロアレイAでジェノタイピングを行ったデータセット


 GA - G - CT -- A
AA - C - GT -- C
 GA - C - GG -- A
AA - C - GT -- A

- マイクロアレイBでジェノタイピングを行ったデータセット

 G - TC - C -- GA
G - TC - G -- AC
 G - TG - C -- GA
A - AC - G -- AC



- マイクロアレイAとBのジェノタイプデータを統合したデータセット

 GA - G - CT -- A
AA - C - GT -- C
 GA - C - GG -- A
AA - C - GT -- A
 G - TC - C -- GA
G - TC - G -- AC
 G - TG - C -- GA
A - AC - G -- AC

- 別の種類のマイクロアレイなど、異なるプラットフォームのジェノタイプデータは、SVS上で統合させることが可能
- 各マイクロアレイで共通して搭載されていないSNPは欠損データとなるため、データの補完にインピュテーションを行う必要がある

Unsort		G 1	G 2	G 3
Map	Sample	SNP A-1886933	SNP A-1902458	SNP A-2131660
	Chromosome	1	1	1
	Position	785989	1130727	1156131
	dbSNP RS ID	rs2980300	rs10907175	rs2887286
	Associated Gene	?	TLL10	SDF4
	Reference Alleles A/B	[A/G]	[A/C]	[C/T]
	Strand	-	+	+
1	GSM233274	B_B	A_B	?_?
2	GSM233272	B_B	A_A	?_?
3	GSM233270	B_B	A_A	A_B
4	GSM233268	A_B	A_B	A_B
5	GSM233266	B_B	A_A	B_B
6	GSM233264	B_B	A_B	A_B
7	GSM233262	A_B	A_A	?_?
8	GSM233260	A_A	A_A	B_B
9	GSM233258	B_B	A_A	B_B
10	GSM233256	B_B	A_A	B_B

rsID Field:

Variant Source:

dbSNPデータベースを利用したアレルのエンコード

ABアレルのAGCT変換

Recode Genotype Column Data by Allele Name

Flip DNA strands for AGCT encoded genotypes

Transcode AB to AGCT encoding using mapping:

Marker map field in format 'A/B':

Transcode using allele mapping:

Marker map field in format 'A:G B:T'

マーカーマップを利用したアレルのエンコード

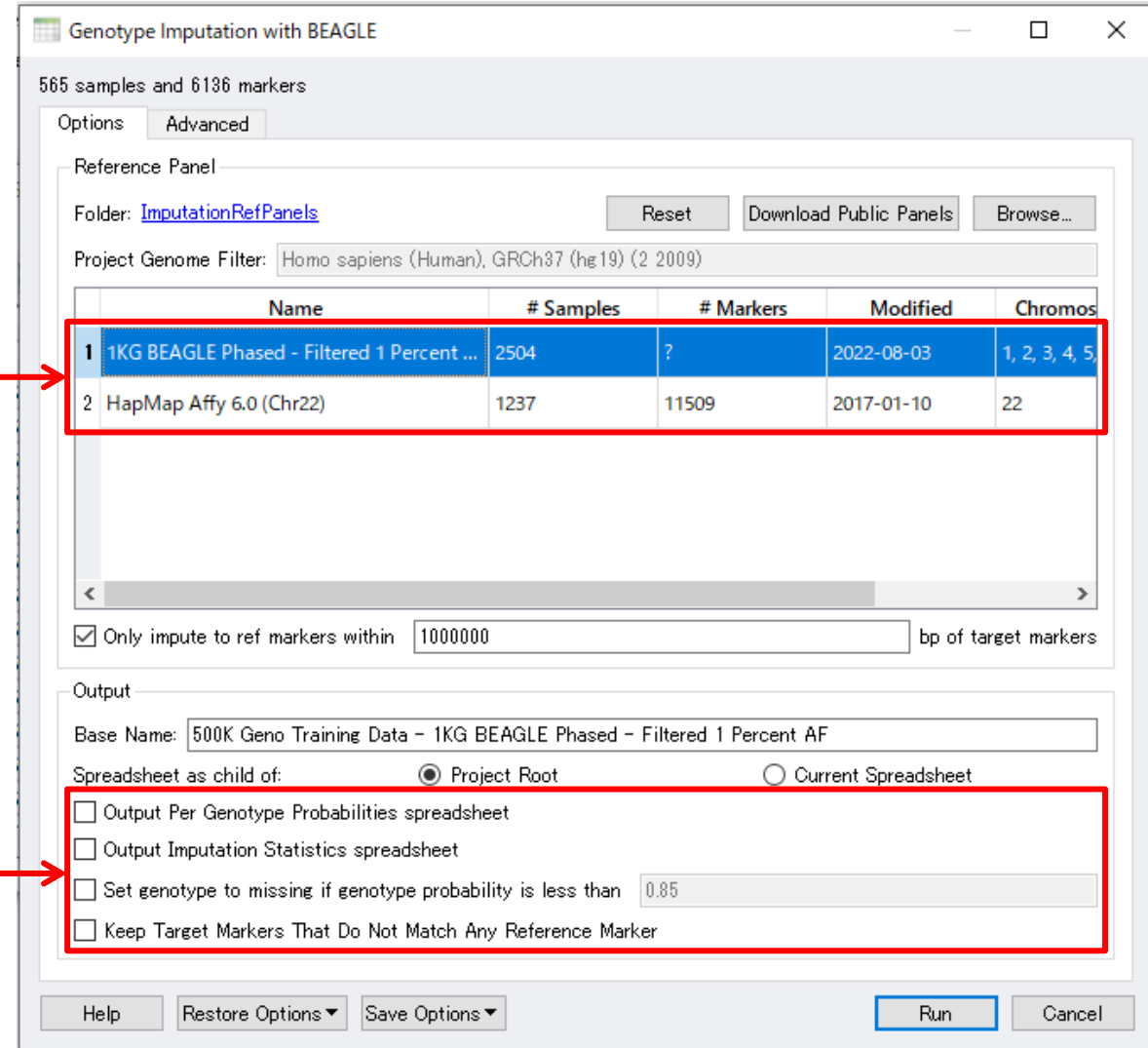
Unsort		G 1	G 2	G 3
Map	Markers	SNP A-1886933	SNP A-1902458	SNP A-2131660
	Chromosome	1	1	1
	Position	785989	1130727	1156131
	Reference	T	A	T
	Alternates	A	C	C
	dbSNP RS ID	rs2980300	rs10907175	rs2887286
	Associated Gene	?	TLL10	SDF4
	Reference Alleles A/B	[A/G]	[A/C]	[C/T]
	Strand	-	+	+
1	GSM233274	T_T	A_C	?_?
2	GSM233272	T_T	A_A	?_?
3	GSM233270	T_T	A_A	C_T
4	GSM233268	A_T	A_C	C_T
5	GSM233266	T_T	A_A	T_T
6	GSM233264	T_T	A_C	C_T
7	GSM233262	A_T	A_A	?_?
8	GSM233260	A_A	A_A	T_T
9	GSM233258	T_T	A_A	T_T
10	GSM233256	T_T	A_A	T_T

- SNPマイクロアレイでは、ジェノタイピング結果がABのアレルデータとして出力されるものがあり、マーカーマップの情報を用いて、AGCTの塩基変換が必要
- マーカーマップにデータがなければ、dbSNPデータベースと直接比較しての変換も可能

- 研究用サンプルデータのアレルエンコードが完了したら、インピュテーション実行用ツールを起動して、インピュテーションを実行する
- インピュテーション実行の際は、リファレンスパネルの選択に加え、データの信頼性の値などの出力オプションの設定も可能
- これとは別に、BEAGLEプログラムのオプション設定も調整可能で、BEAGLE4.1と4.0の選択や、計算の反復回数、サンプルサイズなどの条件を指定可能

リファレンスパネル
の選択

出力オプションの
設定



インプテーション実行結果

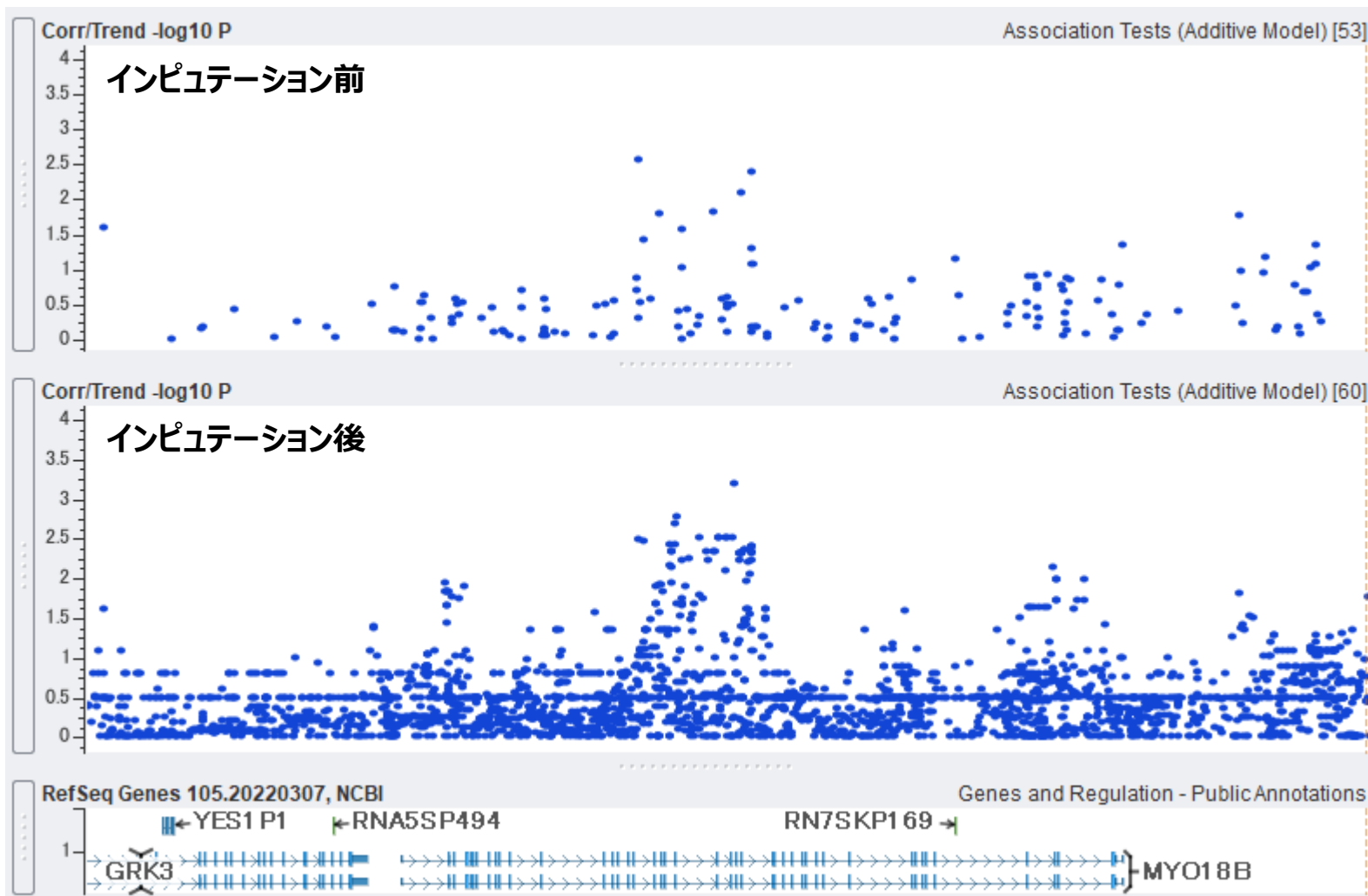
インプテーション前

All: 565 × 6,150			
Active: 565 × 6,150			
G	11	G	12
SNP A-2151351	SNP A-4219868	SNP A-4284341	
22	22	22	
16892143	16894264	16900134	
rs41439550	rs4010554	rs4010550	
?	?	?	
[C/T]	[A/C]	[A/G]	
T_T	C_C	A_A	
C_T	C_C	A_A	
T_T	C_C	A_A	
C_T	C_C	A_A	
T_T	C_C	A_A	
T_T	A_C	A_G	
T_T	?_?	A_G	
T_T	C_C	A_A	

SNP数

インプテーション後

All: 565 × 424,154			
Active: 565 × 424,154			
G	129	G	130
rs9604663	rs587676956	rs143164330	
22	22	22	
16060517	16060544	16060567	
T	T	T	
C	C	G	
C_T	T_T	G_T	
C_T	T_T	G_T	
C_C	T_T	G_G	
T_T	T_T	T_T	
C_T	T_T	G_T	
C_C	T_T	G_G	
C_C	T_T	G_G	
T_T	T_T	T_T	



お問い合わせ先：フィルジエン株式会社

TEL: 052-624-4388 (9:00～18 : 00)

FAX: 052-624-4389

E-mail: biosupport@filgen.jp