人類集団のゲノム・リシークエンスデータにおける imputation

ファイル説明書

2014 年 1 月 28 日

「pack」フォルダに格納したファイルは以下のとおりです。

1. vcf2beagle.pl

「作業報告書」2.(1)で使用した、未知の情報を含む VCF ファイルを Beagle 入力用ファイルに変換するプログラム。

(1) コマンド

```
$ ./vcf2beagle.pl <VCF_file> [<Beagle_file_prefex>]
```

<VCF_file>には入力 VCF ファイル名を指定します。

<Beagle_file_prefex>には出力する Beagle 入力用ファイルの接頭辞を指定します。

(2) 機能

VCF ファイルを Beagle 入力用のファイルに変換して Unphased genotypes file と Phased genotypes file と Markers file を出力します。

① 出力ファイル名

出力ファイル名は、Unphased genotypes file は“.unphased”を、Phased genotypes file は“.phased”を、Markers file には“marker”を、それぞれ接頭辞の後に付けた名前になります。

接頭辞は<Beagle_file_prefex>で指定されたものになります。<Beagle_file_prefex>が省略された場合は、<VCF_file>が“.vcf”で終わるときは<VCF_file>から“.vcf”を除いたものが接頭辞になり、そうでないときは<VCF_file>が接頭辞になります。

② 出力データ

すべての変異サイトにおいて相が推定済みである個体は Phased genotypes file に出力され、1 個以上の変異サイトにおいて相が未推定である個体は Unphased genotypes file に出力されます。

入力 VCF ファイルの GT フィールドに 0 以上の整数が指定されているときは、該当する allele を出力します。入力 VCF ファイルの GT フィールドが missing value (.)であるときは、“N”を出力します。したがって出力されたファイルを Beagle で使用するときは `missing=N` を指定してください。

(3) 実行例

```
$ ./vcf2beagle.pl Xp11_hX.vcf
```

この例では、Xp11_hX.vcf を Beagle 入力用のファイルに変換して、Unphased genotypes file が Xp11_hX.unphased に保存され、Phased genotypes file が Xp11_hX.phased に保存され、Markers file が Xp11_hX.marker に保存されます。

2. beagle2vcf.pl

「作業報告書」2.(3)で使用した、Beagle の出力に既知の相推定結果を反映させて VCF ファイルを作成プログラム。

(1) コマンド

```
$ ./beagle2vcf.pl <phased_VCF_file> <missing_VCF_file> <Beagle_output_file> ¥
  <output_VCF_file>
```

<phased_VCF_file>には既知の相の情報を記載した入力 VCF ファイル名を指定します。

<missing_VCF_file>には相や allele が未知である変異サイトを含む入力 VCF ファイル名を指定します。

<Beagle_output_file>には Beagle の出力結果である phased file のファイル名を指定します。ここで指定する phased file の先頭行は“I”で始まるタイトル行でなければなりません。

<output_VCF_file>には出力する VCF ファイル名を指定します。

(2) 機能

<missing_VCF_file>において allele や相が未知である情報を、<Beagle_output_file>に従って補完し、<output_VCF_file>に出力します。その際、<phased_VCF_file>の相の情報との整合性を確認し、矛盾があるときは<Beagle_output_file>の内容を採用せず未知のままとします。

<missing_VCF_file>において既知である情報は変更せずにそのまま出力します。

① 処理対象の個体

<missing_VCF_file>に記載された個体を以下のように出力します。

<phased_VCF_file>と<Beagle_output_file>の両方にも記載された個体は、情報を補完して出力します。

<phased_VCF_file>と<Beagle_output_file>のどちらかまたは両方に記載されていない個体は、<missing_VCF_file>の内容をそのまま出力します。

② 処理対象の変異サイト

<missing_VCF_file>と<phased_VCF_file>と<Beagle_output_file>のすべてに記載された変異サイトを出力します。

<missing_VCF_file>と<Beagle_output_file>のどちらか一方のみに記載された変異サイトが存在するとエラーになります。

<missing_VCF_file>と<Beagle_output_file>の両方に記載されており<phased_VCF_file>に記載されていない変異サイトが存在するとエラーになります。

③ 相の判定

<Beagle_output_file>の homozygous である変異サイトは<Beagle_output_file>の出力を採用します。

<Beagle_output_file>が heterozygous である変異サイトは、直前の heterozygous の変異サイトと当該の変異サイトと直後の heterozygous の変異サイトの 3 つについて、<phased_VCF_file>と<Beagle_output_file>とで相を比較します。これらが矛盾しないときは<Beagle_output_file>を採用します。これらが矛盾するときは<Beagle_output_file>を採用せず、未知のままとします。

(3) 実行例

```
$ ./beagle2vcf.pl 1000g_Xp11_hX.vcf modified_Xp11_hX.vcf ¥
  o1.Xp11_hX.phased o1.Xp11_hX.vcf
```

この例では、modified_Xp11_hX.vcf において allele や相が未知である情報を、o1.Xp11_hX.phased に従って補完し、o1.Xp11_hX.vcf に出力します。その際、1000g_Xp11_hX.vcf の相の情報との整合性を確認し、矛盾があるときは o1.Xp11_hX.phased の内容を採用せず未知のままとします。

modified_Xp11_hX.vcf において既知である情報は変更せずにそのまま出力します。

3. Xp11_hX/

(1) Xp11_hX.vcf

作業報告書 2.(1)で使用した、個体によって相の推定済みと未推定が混合している場合や片方の allele のみが既知である SNP サイトを含む VCF ファイル。

(2) Xp11_hX.unphased, Xp11_hX.phased, Xp11_hX.marker

作業報告書 2.(1)で作成した Beagle の入力ファイル。

(3) o1.Xp11_hX.unphased.phased, o1.Xp11_hX.phased.phased

作業報告書 2.(2)で実行した Beagle の出力ファイル。

(4) o1.log

作業報告書 2.(2)で実行した Beagle のログファイル。

(5) o1.Xp11_hX.vcf

作業報告書 2.(3)で作成した、o1.Xp11_hX.unphased.phased と o1.Xp11_hX.phased.phased の情報を用いて Xp11_hX.vcf の未知の情報を補完した VCF ファイル

(6) o1_phased.Xp11_hX.vcf

作業報告書 2.(3)で作成した、Xp11_hX.vcf を作成する際にまず o1.Xp11_hX.phased.phased の情報のみを反映させた VCF ファイル。

(7) o1.Xp11_hX.unphased, o1.Xp11_hX.phased, o1.Xp11_hX.marker

作業報告書 2.(4)で作成した Beagle（2 回目）の入力ファイル。

(8) o2*

作業報告書 2.(4)で作成した Beagle（2 回目）の出力ファイルおよび Beagle（3 回目）の入力ファイル。

(9) o3*

作業報告書 2.(4)で作成した Beagle（3 回目）の出力ファイル。