組み換え率および連鎖不平衡(LD)ブロック保存確率の算出

ファイル説明書

2014 年 7 月 31 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. vcf2rehh.pl

「作業報告書」2.(1)①で使用した、VCF ファイルから rehh 用の入力ファイルを作成するプログラム。

(1) コマンド

```
$ vcf2rehh.pl <file>
```

引数（必須）

<file>には VCF ファイル名を指定します。

(2) 機能

VCF ファイルから、rehh 用の入力ファイル（ハプロタイプファイル・SNP ファイル）を作成します。

① 出力ファイル名

(a) ハプロタイプファイル

入力 VCF ファイル名が“.vcf”で終わるときは、入力 VCF ファイル名の“.vcf”を“.hap”に変えたものになり、そうでないときは入力 VCF ファイル名の末尾に“.hap”を加えたものになります。

(b) SNP ファイル

入力 VCF ファイル名が“.vcf”で終わるときは、入力 VCF ファイル名の“.vcf”を“.inp”に変えたものになり、そうでないときは入力 VCF ファイル名の末尾に“.inp”を加えたものになります。

② 出力対象

(a) サイト

VCF ファイルに記載されたサイトのうち、ALT allele が 1 種類であり、REF allele と ALT allele のいずれも 1 塩基であるサイトを出力します。

(b) サンプル

VCF ファイルに記載されたサンプルのうち、一倍体と二倍体が混在していないサンプルを出力します。

③ 出力内容

(a) ハプロタイプファイル

ハプロタイプ名は VCF ファイルの見出し行の値に“-1”を付加したものおよび“-2”を付加したものになります。ただし一倍体の場合は前者のみとなります。

ハプロタイプは、VCF ファイルに記載された genotype に応じて、ancestral allele は“1”、derived allele は“2”、missing value または unphased genotype の場合は“0”となります。Ancestral allele は INFO フィールドの AA サブフィールドの値から判別します。AA サブフィールドの値が REF allele または ALT allele と一致するときは、それを ancestral allele とします。一致しないか、または AA サブフィールドが存在しないときは、REF allele を ancestral allele であるとみなします。Ancestral allele ではない他方の allele を derived allele とします。

(b) SNP ファイル

SNP 名は VCF ファイルの ID フィールドの値を使用します。ただし ID フィールドが missing "." であるときは、染色体名とコロン":"と座標をつなげた文字列とします。

染色体名は VCF ファイルの CHROM フィールドの値を使用します。

座標は VCF ファイルの POS フィールドの値を使用します。

Ancestral allele は常に"1"となります。

Derived allele は常に"2"となります。

(3) 実行例

```
$ vcf2rehh.pl input.vcf
```

この例では、input.vcf から rehh 用の入力ファイルを作成し、ハプロタイプファイルを input.hap に、SNP ファイルを input.inp に出力します。

(4) 制限事項

VCF ファイルでは CHROM フィールドの値としてコロン":"とスペースを除く任意の文字列を使用することができます。一方、rehh の入力ファイルの染色体名は数字でなければエラーとなります。

vcf2rehh.pl では VCF ファイルの CHROM フィールドに記載された染色体名が rehh に適合するかどうかのチェックは行わず、そのまま SNP ファイルに出力します。そのため、この値が rehh に適合しない場合は、rehh に入力する前に染色体名を修正する必要があります。

2. rehh/*

各領域で EHH の計算と bifurcation graph の作成に使用したデータと結果、および領域間を比較したまとめ。