遺伝子流動程度の集団遺伝学的統計推定

ファイル説明書

2014 年 9 月 29 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. sstar.pl

「作業報告書」2.(1)で使用した、S*を計算するプログラム。

(1) コマンド

```
$ sstar.pl [--vcf | --pop | --ref | --ref-mac | --size] <args>
```

引数（必須）

--vcf <args>には入力 VCF ファイル名を指定します。

--pop <args>には検査集団に属する個体名一覧のファイル名を指定します。

--ref <args>には参照集団に属するハプロタイプ名一覧のファイル名を指定します。

引数（オプション）

--ref-mac <args>には S*の計算に使用する SNP の、参照集団における minor allele の数の上限を整数で指定します。デフォルト値は 0 です。

--size <args>には 1 個のハプロタイプの S*を計算する際に使用するハプロタイプの数を指定します。デフォルト値は 40 です。

(2) 機能

VCF ファイルに基づいて、指定されたハプロタイプの S*を計算します。

① 入力ファイル

(a) VCF ファイル

S*を計算したい領域の VCF ファイルです。VCF ファイルに含まれるサイト全体を対象に S*を計算します。

(b) 検査集団に属する個体名一覧のファイル

S*を計算したい検査集団に属する個体名を記載したファイルです。個体名は 1 行ずつ記述します。

(c) 参照集団に属するハプロタイプ名一覧のファイル

参照集団（遺伝子移入がないと仮定）に属するハプロタイプ名を記載したファイルです。

ハプロタイプ名は 1 行ずつ記述します。ハプロタイプ名は VCF ファイルの GT フィールドの左側のハプロタイプを指定するときは個体名に“-1”を付加し、右側のハプロタイプを指定するときは個体名に“-2”を付加したものとなります。一倍体の場合は“-1”を付加します。

② 出力内容

検査集団に属するすべての個体のハプロタイプの S*を出力します。ただしヘテロ接合体で相が不明であるサイトが存在する個体については出力しません。

各ハプロタイプの S*の計算において使用するサイトは、以下の条件を満たすものとなります。

- 一塩基置換である。

- Allele の種類が 2 種類である。
- 参照集団における頻度が--ref-mac で指定された値以下である。
- 計算対象のハプロタイプが rare allele である。

検査集団は、--pop <args>で指定された個体のハプロタイプの中から--size <args>で指定された数を無作為に選んだものになります。これは 1 個のハプロタイプの S*を計算するたびに新たに選択されます。なお--pop <args>で指定された個体のハプロタイプが--size <args>で指定された数より少ない場合は、前者のすべてが検査集団となります。

個体名と S*とそのときの SNP の組み合わせをタブ区切りで標準出力に出力します。

S*が同点となるような SNP の組み合わせが複数個存在するときは、それらのすべてを出力しますが、異なる部分のみ SNP 名を出力し、すでに出力した SNP から組み合わせと変更がない部分は"..."で表します。

S*の計算に使用できるサイトが 1 個以下である場合は S*を算出できないため、"///"を出力します。

(3) 実行例

```
$ sstar.pl --vcf 1000g.vcf --pop test_pop.id  --ref ref.hap ¥
  --size 100 --ref-mac 24 > s-star.txt
```

（表示が複数行に渡っているため、改行の直前に¥を記述してあります。改行せずに入力する場合は¥を入力する必要はありません。）

この例では、1000g.vcf に記載された領域において、ref.hap に記載されたハプロタイプを参照集団として、test_pop.id に記載された個体のハプロタイプの S*とそのときの SNP の組み合わせを出力し、s-star.txt に保存します。

参照集団における rare allele の頻度が 24 以下の SNP を使用して、test_pop.id から 100 個のハプロタイプを無作為に選んだものを検査集団として、S*を計算します。

(4) 制限事項

① 入力 VCF ファイル

入力 VCF ファイルは常に領域全体が S*の計算対象となります。VCF ファイルの一部分の領域のみで S*を計算することはできません。

VCF ファイルの仕様上は複数の染色体のサイトを含むことができますが、sstar.pl の入力 VCF ファイルでは、すべてのサイトが同一の染色体上のサイトでなければなりません。2 個以上の染色体のサイトが記載されている場合はエラーとなります。

② 同一の配列の S*

検査集団に属する個体名一覧のファイルに記載された個体のハプロタイプの中にまったく同一の配列のものが存在する場合であっても、S*の計算は別々に行います。そのため、--size で指定された数のハプロタイプを検査集団として選ぶ際の選び方によって S*の値が異なることがあります。

2. sstar2mega_group.pl

「作業報告書」2.(2)で使用した、MEGA[1]用のグループ定義ファイルを出力するプログラム。

(1) コマンド

```
$ sstar2mega_group.pl [-S | -G | -C] <args>
```

引数（必須）

-S <args>には入力 S*ファイル（sstar.pl の出力ファイル）名を指定します。

引数（オプション）

-G <args>にはハプロタイプ名と統合ハプロタイプ名の対応一覧のファイル名を指定します。

-C <args>には S*の階級区分値を整数で指定します。コンマ区切りで複数個の値を指定することができます。

(2) 機能

S*の値を記載した、sstar.pl の出力ファイルに基づき、S*の値を階級に分けたものを MEGA 用のグループ定義ファイルに変換して出力します。

① 入力ファイル

(a) 入力 S*ファイル

sstar.pl の出力を保存したファイルです。

(b) ハプロタイプ名と統合ハプロタイプ名の対応一覧のファイル

系統樹では同じ配列の OTU が 1 つにまとめられていることがあるので、まとめる前の個別のハプロタイプ名とまとめた後の統合ハプロタイプ名との対応を記載します。

各行の1列目がまとめる前の個別のハプロタイプ名、2列目がまとめた後の統合ハプロタイプ名です。列はタブ区切りです。

② 階級区分

S*の値を、コマンドライン引数で指定された階級区分値に従って区分します。なお、"///"(算出不能)、"-inf"(－∞)、"-10000"、"0"は指定の有無にかかわらず区分されます。正の値の S*は階級区分値によって区分され、階級名は当該ハプロタイプの S*の値を超えない最大の階級区分値となります。

S*の値が"-inf"(大文字・小文字は問わない)で始まる場合はその後にどのような文字が続いていてもすべて－∞とみなされます。

S*の値が上記以外の場合は、階級名が"invalid"となります。

③ 出力

OTU 名とその階級名を"="で結んだものを、1 行に 1 個ずつ、標準出力に出力します。

(3) 実行例

```
$ sstar2mega_group.pl -S s-star.txt -G group.txt ¥
  -C 0,5000,20000,50000,100000 > s-star.mega.txt
```

(表示が複数行に渡っているため、改行の直前に¥を記述してあります。改行せずに入力する場合は¥を入力する必要はありません。)

この例では、まず s-star.txt に記載された S*の値に基づいて、各ハプロタイプを次の 8 つの階級に区分します。括弧内は階級名です。

- 算出不能 (NA)
- －∞ (-inf)
- －10000 (-10000)
- 0 (0)
- 5000 以上 20000 未満 (5000)
- 20000 以上 50000 未満 (20000)
- 50000 以上 100000 未満 (50000)
- 100000 以上 (100000)

また、s-star.txt に記載されたハプロタイプ名のうち、group.txt に記載されているものについては、group.txt に基づいて名称を統合ハプロタイプ名に変換します。

その結果を MEGA 用のグループ定義ファイルに変換したものを、s-star.mega.txt に保存します。

(4) 制限事項

S*の値が異なる複数のハプロタイプが、-G で指定されたハプロタイプ名と統合ハプロタイプ名の対応一覧のファイルで同一の統合ハプロタイプに属するとされている場合であっても、sstar2mega_group.pl では特にチェックを行いません。そのため、1 個の統合ハプロタイプ名に複数の異なる階級（グループ名）が付けられて出力される場合があります。

そのような場合、MEGA では先に現れるものが優先されます。

3. S-star/*.txt

各領域での S*の計算結果。

4. S-star/s-star.ppt

各領域の系統樹において、S*≧T1 である OTU にマーカーを表示したもの。

[1] MEGA … URI: http://www.megasoftware.net/