1000 Genomes Project data を用いた

ヒトゲノム遺伝子流動候補領域における、

ハプロタイプ間組み換え推定

ファイル説明書

2013 年 10 月 2 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. vcf2rdf.pl

「作業報告書」2.(1)②および 2.(2)④で使用した、VCF ファイルを RDF ファイルに変換するプログラム。

(1) 機能

1000 Genomes Project の VCF ファイルを Network 4.6.1.1[1]の入力ファイルである RDF ファイルに変換し、標準出力に出力します。

Weight はすべてのサイトが 10（Network 4.6.1.1 のデフォルト値）になります。

Frequency はすべての配列が 1（Network 4.6.1.1 のデフォルト値）になります。

(2) コマンド

```
$ ./vcf2rdf.pl [<input_VCF>]
```

<input_VCF>には入力 VCF ファイル名を指定します。省略した場合は標準入力から読み込みます。

(3) 実行例

```
$ ./vcf2rdf.pl 1000g.vcf
```

この例では、1000g.vcf を RDF ファイルに変換し、標準出力に出力します。

(4) RDF ファイルについての補足

Network 4.6.1.1 の RDF ファイルは 2 種類の異なる書式が存在します。1 つは Example に含まれているファイルの書式であり、もう 1 つは RDF-Editor で保存したときに作成されるファイルの書式です。RDF-Editor はどちらの書式でも開くことができます。詳細は Network 4.6.1.1 の構成ファイルを参照してください。

vcf2rdf.pl では、前者の Example に含まれているファイルの書式のみを扱います。

(5) 制限事項

① 倍数性

vcf2rdf.pl では二倍体を仮定しているため、入力が一倍体であった場合は 2 つ目の配列は“.”となります。

② 文字数

Network 4.6.1.1 では配列名は 6 文字以内、サイト名は 6 文字以内の制限があります（RDF-Editor 内のみでは制限がより緩くなりますが、系統ネットワークを描画する際にはやはり同じ制限がかかります）。したがって vcf2rdf.pl が出力する RDF ファイルも同様となります。

配列名は、VCF ファイルの個体名（7 文字に限る。）の 3 文字目以降に、genotype 欄の左側の配列の

場合は“L”を、genotype欄の右側の配列の場合は“R”を加えたものになります。例えばVCFファイルの個体名が“HG00096”の場合、出力される配列名は“00096L”と“00096R”になります。このため、個体名の先頭の 2 文字(現時点では“HG”と“NA”のみですが他の文字でもかまいません。)を除いた残りの5 文字が重複しないようにする必要があります。

サイト名は、VCF ファイルの座標(999999999 以下に限る。)の下 6 桁が使われます(サイト ID は使われません)。例えば VCF ファイルの座標が“123456789”の場合、出力されるサイト名は“456789”になります。このため、1000000bps を超える範囲を扱う場合はサイト名の重複に注意してください。

2. uniq_rdf.pl

「作業報告書」2.(1)③で使用した、RDF ファイル中で同一のハプロタイプを 1 つのグループに統合するプログラム。

(1) 機能

Network 4.6.1.1[1]の入力ファイルである RDF ファイルで同一のハプロタイプを 1 つのグループに統合し出力します。Frequency はすべての配列が 1 になります。

また、統合したグループ名と統合する前のハプロタイプ名の一覧表を出力します。

複数のハプロタイプが 1 つにまとめられたグループ名は以下のようになります。

- グループ名は 2 つのパートに分かれます: [種類][番号]
- [種類]は含まれる配列の種類により次の文字が入ります。
  - ヒトの reference と“DENISO”のいずれも含まれないグループ … H
  - ヒトの reference が含まれ、“DENISO”が含まれないグループ … HRH
  - Denisova が含まれ、ヒトの“DENISO”が含まれないグループ … DH
  - ヒトの reference と“DENISO”の両方が含まれるグループ … HRDH
- [番号]はグループが見つかった順番に整数が入ります。

(2) コマンド

```
$ ./uniq_rdf.pl <in_RDF_file> [<out_RDF_file> [<out_group_file>]]
```

<in_RDF_file>には入力 RDF ファイル名を指定します。

<out_RDF_file>には出力 RDF ファイル名を指定します。

<out_group_file>には一覧表のファイル名を指定します。

<out_RDF_file>と<out_group_file>の両方を、または<out_group_file>のみを、省略することができます。

<out_RDF_file>が省略されたときのファイル名は、<in_RDF_file>が“.rdf”で終わるときは<in_RDF_file>の“.rdf”を“.uniq.rdf”に変えたものになり、そうでないときは<in_RDF_file>の後に“.uniq.rdf”を加えたものになります。

<out_group_file>が省略されたときのファイル名は、<out_RDF_file> が“.uniq.rdf”で終わるときは<out_RDF_file>の“.uniq.rdf”を“.group”に変えたものになり、そうでなく<out_RDF_file>が“.rdf”で終わるときは<out_RDF_file>の“.rdf”を“.group”に変えたものになり、そうでないときは<out_RDF_file>の後に“.group”を加えたものになります。

(3) 実行例

```
$ ./uniq_rdf.pl 1000g.rdf 1000g.uniq.rdf 1000g.group
```

この例では、1000g.rdf で同一のハプロタイプを 1 つに統合し 1000g.uniq.rdf に出力します。また、統合したハプロタイプ名と統合する前のハプロタイプ名の一覧表を 1000g.group に出力します。

(4) RDF ファイルについての補足

Network 4.6.1.1 の RDF ファイルは 2 種類の異なる書式が存在します。1 つは Example に含まれているファイルの書式であり、もう 1 つは RDF-Editor で保存したときに作成されるファイルの書式です。RDF-Editor はどちらの書式でも開くことができます。詳細は Network 4.6.1.1 の構成ファイルを参照してください。

uniq_rdf.pl では、前者の Example に含まれているファイルの書式のみを扱います。

(5) Frequency についての補足

uniq_rdf.pl で出力する配列の Frequency は、入力ファイルに含まれる同一の配列の数ではなく、常に 1 になります。これは Network 4.6.1.1 で Frequency が 1 より大きい配列に対して reduction を行う際の計算方法が異なるため、Frequency を 1 として系統ネットワークの計算を行えるようにするためです。なお、系統ネットワークを作成した後に、fix_fdi.pl を使用して、元の Frequency に応じた大きさのノードで系統ネットワークを描画することができます。

3. fix_fdi.pl

「作業報告書」2.(1)⑦で使用した、Network4.6.1.1[1]形式の系統ネットワークファイル（FDI 形式）を編集しノードの大きさと色をグループの Frequency と地域に合ったものにして出力するプログラム。

(1) 機能

uniq_rdf.pl で作成した、統合したグループ名と統合する前のハプロタイプ名の一覧表をもとに、次のように FDI ファイルを編集して標準出力に出力します。

① 大きさ

一覧表をもとに、FDI ファイルの TAXON_FREQUENCY を編集することで、ノードの大きさを変更します。

② 色

一覧表をもとに、FDI ファイルの TAXON_COLOR_PIE と TAXON_PIE_FREQUENCY を編集することで、ノードの色を変更します。

(2) コマンド

```
$ ./fix_fdi.pl <in_FDI_file> <group_file> [<population_file> <color_file>]
```

<in_FDI_file>には入力 FDI ファイル名を指定します。

<group_file>には統合したグループ名と統合する前のハプロタイプ名の一覧表のファイル名を指定します。

<population_file>には 1000 Genomes Project の個体名と地域名の対応表のファイル名を指定します。

<color_file>には地域名と色の対応表のファイル名を指定します。

<population_file>と<color_file>は両方を同時に省略する場合に限り省略することができます。省略された場合はノードの大きさのみを編集し、色は編集しません。

(3) 入力ファイル形式

① 統合したグループ名と統合する前のハプロタイプ名の一覧表のファイル

統合したグループ名と統合する前のハプロタイプ名の一覧表のファイルは、uniq_rdf.pl で作成したものを使用します。

② 個体名と地域名の対応表のファイル

1000 Genomes Project の個体名と地域名の対応表のファイルは、1 行につき 1 個体の地域を定義し

ます。各行では地域名と個体名をタブ区切りで記述します。

個体名は 1.(5)②と同様の変換をした上で解釈されます。

例は同包の population.txt を参照してください。

③ 地域名と色の対応表のファイル

地域名と色の対応表のファイルは、1 行につき 1 地域の色を定義します。各行では地域名と色をタブ区切りで記述します。

色は 16 進数 6 桁(#000000～#ffffff)、16 進数 3 桁(#000～#fff)、10 進数(0～16777215)のいずれかの形式で指定します。10 進数は FDI ファイルの記述と同一となります。

例は同包の color.txt を参照してください。

(4) 実行例

```
$ ./fix_fdi.pl 1000g.fdi 1000g.group population.txt color.txt
```

この例では、1000g.group に基づいて 1000g.fdi の TAXON_FREQUENCY を編集し、1000g.group と population.txt と color.txt に基づいて TAXON_COLOR_PIE と TAXON_PIE_FREQUENCY を編集し、標準出力に出力します。

4. 1000g.vcf

「作業報告書」2.(1)①で使用した Xp11_hX 領域の 1000 Genomes Project の VCF ファイル。

5. 1000g_noindel.rdf

「作業報告書」2.(1)②で作成した RDF ファイル。

6. 1000g_noindel.uniq.rdf

「作業報告書」2.(1)③で作成した RDF ファイル。

7. 1000g_noindel.group

「作業報告書」2.(1)③で作成した統合したグループ名と統合する前のハプロタイプ名の一覧表のファイル。

8. 1000g_main.rdf

「作業報告書」2.(1)④で作成した RDF ファイル。

9. 1000g_main.fdi

「作業報告書」2.(1)⑥で作成した Network 形式の系統ネットワーク。

10. 1000g_main_fix.fdi

「作業報告書」2.(1)⑦で作成した Network 形式の系統ネットワーク。

11. population.txt

「作業報告書」2.(1)⑦で使用した個体名と地域名の対応表のファイル。

12. color.txt

「作業報告書」2.(1)⑦で使用した地域名と色の対応表のファイル。

13. HumanChimp.rdf

「作業報告書」2.(2)④で作成した RDF ファイル。

14. HumanChimp.fdi

「作業報告書」2.(2)⑤で作成した Network 形式の系統ネットワーク。

15. Human3+root.odg

「作業報告書」2.(3)で作成したルートを書き加えた系統ネットワークの画像。

16. Human9.rdf

「作業報告書」3.(1)の系統ネットワークを作成するために使用した RDF ファイル。

17. Human9.fdi

「作業報告書」3.(1)で作成した Network 形式の系統ネットワーク。

18. Human5.rdf

「作業報告書」3.(2)の系統ネットワークを作成するために使用した RDF ファイル。

19. Human5.fdi

「作業報告書」3.(2)で作成した Network 形式の系統ネットワーク。

[1] Network 4.6.1.1 … URI: http://www.fluxus-engineering.com/sharenet.htm