1000 Genomes Project ハプロタイプと高品質古代人ゲノムとの比較

ファイル説明書

2014 年 3 月 26 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. gt_majority.pl

「作業報告書」1.(3)で使用した、unphased の二倍体からハプロタイプを仮定し VCF ファイルを作成するプログラム。

(1) コマンド

```
$ gt_majority.pl <VCF_file>
```

<VCF_file>には入力 VCF ファイル名を指定します。

(2) 機能

VCF ファイルの uphased の二倍体の genotype から、次のような基準で 1 本のハプロタイプを仮定し、その VCF ファイルを標準出力に出力します。出力される形式は、phased の二倍体(homozygous)です。

- 0/0 → 0|0
- 1/1 → 1|1
- 0/1および1/0 → 0|0または1|1（挿入・欠失は0|0。置換はリード数が多いもの（同数であれば1|1）を採用する。）

ただし FILTER フィールドの値が「LowQual」であるサイトは削除します。なおリード数は VCF ファイルの GT フィールドの A・C・G・T のサブフィールドの値を利用します。入力 VCF ファイルが heterozygous であり、かつ A・C・G・T のサブフィールドが存在しないときは、missing value “.”を出力します。

(3) 実行例

```
$ gt_majority.pl Xp11_hX.Nea.recode.vcf
```

この例では、Xp11_hX.Nea.recode.vcf の unphased の二倍体からハプロタイプを仮定し VCF ファイルを標準出力に出力します。

2. make_rdf_from_vcfs.sh

「作業報告書」2.(1)で使用した、VCF ファイルから系統ネットワーク用 RDF ファイルを作成するプログラム。

(1) 必要環境

- VCFtools [1] (vcftools, vcf-merge)
- Tabix [2] (bgzip, tabix)
- gt_majority.pl
- vcf_for_network.pl
- maf2vcf.pl
- vcf2rdf.pl (*)
- uniq_rdf.pl (*)

(*)印…「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」の納品物に含まれます。

(2) コマンド

```
$ make_rdf_from_vcfs.sh <basename_of_VCFs> <Chimpanzee_MAF> ¥
  <chr_name> <hgref_name> <Chimpanzee_name>
```

<basename_of_VCFs>には入力 VCF ファイルの拡張子を除いたファイル名を指定します。

<Chimpanzee_MAF>には入力 MAF ファイル名を指定します。

<chr_name>には染色体名を指定します。

<hgref_name>には<Chimpanzee_MAF>におけるヒトのリファレンス配列の配列名を指定します。

<Chimpanzee_name>には<Chimpanzee_MAF>におけるチンパンジーの配列名を指定します。

(3) 機能

1000 Genomes の再推定後の VCF ファイルとデニソバ人の VCF ファイルと Altai Neandertal の VCF ファイルを結合し、系統ネットワークに使用しないサイトと染色体を除き、RDF ファイルを出力します。

また、MAF ファイルからチンパンジーの情報を取得し、系統ネットワークに使用できるサイトのみをでチンパンジーを含む RDF ファイルを出力します。

① 入力ファイル

- <basename_of_VCFs>.BAM.vcf … 1000 Genomes の再推定後の VCF ファイル
- <basename_of_VCFs>.Deni.recode.vcf … デニソバ人の VCF ファイル
- <basename_of_VCFs>.Nea.recode.vcf … Altai Neandertal の VCF ファイル
- <Chimpanzee_MAF> … チンパンジーを含む MAF ファイル

② 出力ファイル

- <basename_of_VCFs>.BAM.vcf.gz … <basename_of_VCFs>.BAM.vcf の bgzip 圧縮
- <basename_of_VCFs>.BAM.vcf.gz.tbi … <basename_of_VCFs>.BAM.vcf.gz のインデックス
- <basename_of_VCFs>.Deni.major.vcf … <basename_of_VCFs>.Deni.recode.vcf から 1 本のハプロタイプを仮定した VCF ファイル
- <basename_of_VCFs>.Deni.nw.log … VCFtools のログ
- <basename_of_VCFs>.Deni.nw.recode.vcf.gz … <basename_of_VCFs>.Deni.major.vcf から挿入・欠失のサイトを除いた VCF ファイルの bgzip 圧縮
- <basename_of_VCFs>.Deni.nw.recode.vcf.gz.tbi … <basename_of_VCFs>.Deni.nw.recode.vcf.gz のインデックス
- <basename_of_VCFs>.Nea.major.vcf … <basename_of_VCFs>.Nea.recode.vcf から 1 本のハプロタイプを仮定した VCF ファイル
- <basename_of_VCFs>.Nea.nw.log … VCFtools のログ
- <basename_of_VCFs>.Nea.nw.recode.vcf.gz … <basename_of_VCFs>.Nea.major.vcf から挿入・欠失のサイトを除いた VCF ファイルの bgzip 圧縮
- <basename_of_VCFs>.Nea.nw.recode.vcf.gz.tbi … <basename_of_VCFs>.Nea.nw.recode.vcf.gz のインデックス
- <basename_of_VCFs>.BAMDeniNea.vcf … <basename_of_VCFs>.BAM.vcf と <basename_of_VCFs>.Deni.nw.recode.vcf と<basename_of_VCFs>.Nea.nw.recode.vcf を連結した VCF ファイル

- <basename_of_VCFs>.BAMDeniNea.nw.vcf
  … <basename_of_VCFs>.BAMDeniNea.vcf で missing or unphased genotype があるハプロタイプをすべて missing value “.”に置き換え、3 種類以上の allele が出現するサイトを削除した VCF ファイル
- <basename_of_VCFs>.BAMDeniNea.nw.rdf
  … <basename_of_VCFs>.BAMDeniNea.nw.vcf を RDF ファイルに変換し、missing value “.”を含むハプロタイプを削除し、デニソバ人と Altai Neandertal の配列を 1 本ずつにした RDF ファイル
- <basename_of_VCFs>.BAMDeniNea.nw.uniq.rdf
  … <basename_of_VCFs>.BAMDeniNea.nw.rdf の同一の配列を 1 つのグループにまとめた RDF ファイル
- <basename_of_VCFs>.BAMDeniNea.nw.group
  … <basename_of_VCFs>.BAMDeniNea.nw.uniq.rdf のグループ名と構成要素の一覧
- <basename_of_VCFs>.BAMDeniNea.nw.Chimp.rdf
  … <basename_of_VCFs>.BAMDeniNea.nw.vcf に<Chimpanzee_MAF>の情報を加え、系統ネットワークに使用できるサイトのみにした VCF ファイルから作成した RDF ファイルから missing value “.”を含むハプロタイプを削除し、デニソバ人と Altai Neandertal の配列を 1 本ずつにした RDF ファイル
- <basename_of_VCFs>.BAMDeniNea.nw.Chimp.uniq.rdf
  … <basename_of_VCFs>.BAMDeniNea.nw.Chimp.rdf の同一の配列を 1 つのグループにまとめた RDF ファイル
- <basename_of_VCFs>BAMDeniNea.nw.Chimp.group
  … <basename_of_VCFs>.BAMDeniNea.nw.Chimp.uniq.rdf のグループ名と構成要素の一覧

(4) 実行例

```
$ make_rdf_from_vcfs.sh Xp11_hX Xp11_hX_multiz46way.maf ¥
  X hg19.chrX panTro2.chrX
```

この例では、Xp11_hX.BAM.vcf と Xp11_hX.Deni.recode.vcf と Xp11_hX.Nea.recode.vcf を結合した VCF ファイルから、系統ネットワークに使用できるサイトのみの RDF ファイルを作成し、同一の配列を 1 つのグループにまとめたものが Xp11_hX.BAM.DeniNea.nw.uniq.rdf に出力されます。

また、Xp11_hX_multiz46way.maf を X 染色体とみなし、hg19.chrX をリファレンス配列として panTro2.chrX の配列を VCF ファイルに加え、系統ネットワークに使用できるサイトのみの RDF ファイルを作成し、同一の配列を 1 つのグループにまとめたものが Xp11_hX.BAMDeniNea.nw.Chimp.uniq.rdf に出力されます。

3. vcf_for_network.pl

2.make_rdf_from_vcfs.sh で使用した、系統ネットワークに使用しないサイトと染色体を VCF ファイルから除くプログラム。

(1) コマンド

```
$ vcf_for_network.pl <site|hap> [<VCF_file>]
```

<site|hap>には 3 種類以上の allele が現れるサイトの処理方法を指定します。

<VCF_file>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が使用されます。

(2) 機能

VCF ファイルから、系統ネットワークに使用できないサイトや染色体の情報を除去します。

① Missing value

missing value “.”がある染色体の全サイトを missing value “.”とします。

② 3 種類以上の allele が現れるサイト

<site|hap>で site を指定すると 3 種類以上の allele が現れるサイトを削除します。<site|hap>で hap を指定すると 3 種類以上の allele が現れるサイトでは allele count が多い 2 種類を REF と ALT とし、それ以外の allele を持つ染色体の全サイトを missing value “.”とします。

(3) 実行例

```
$ vcf_for_network.pl site Xp11_hX.BAMDeniNea.vcf
```

この例では、Xp11_hX.BAMDeniNea.vcf で missing value “.”がある染色体の全サイトを missing value “.”とし、3 種類以上の allele が現れるサイトを削除した VCF ファイルを標準出力に出力します。

4. maf2vcf.pl

2.make_rdf_from_vcfs.sh で使用した、VCF ファイルに MAF ファイルの情報を加えるプログラム。

(1) コマンド

```
$ maf2vcf.pl [<VCF_file>] <MAF_file> <chr_name> <reference_OTU_name> ¥
  <output_OTU_names>
```

<VCF_file>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が使用されます。

<MAF_file>には入力 MAF ファイル名を指定します。

<chr_name>には染色体名を指定します。

<reference_OTU_name>には<Chimpanzee_MAF>におけるリファレンス配列の配列名を指定します。

<output_OTU_names>には<Chimpanzee_MAF>における出力したい種の配列名をコンマ区切りで指定します。

(2) 機能

VCF ファイルに記載されたサイトについて、MAF ファイルの指定された種の genotype を加えて VCF ファイルに出力します。MAF ファイルの配列が VCF ファイルの REF とも ALT とも異なるときは、そのサイトは出力しません。

(3) 実行例

```
$ maf2vcf.pl Xp11_hX.BAMDeniNea.nw.vcf Xp11_hX_multiz46way.maf ¥
  X hg19.chrX panTro2.chrX,ponAbe3.chrX
```

この例では、Xp11_hX.BAMDeniNea.nw.vcf に記載されたサイトについて、Xp11_hX_multiz46way.maf の hg19.chrX をリファレンス配列として panTro2.chrX と ponAbe3.chrX の genotype を加えて VCF ファイルを標準出力に出力します。

5. grep_rdf

「作業報告書」2.(1)で使用した、RDF ファイルから特定の配列を抜き出すプログラム。

(1) コマンド

```
$ grep_rdf <in|ex> [--id <ids> | --idfile <file>] [<RDF_file>]
```

<in|ex>には配列名の指定方法を指定します。

--id <ids>には配列名をコンマ区切りで指定します。

--idfile <file>には配列名を改行区切りで記載したファイルのファイル名を指定します。

<RDF_file>には入力 RDF ファイル名を指定します。省略した場合は標準入力が使用されます。

(2) 機能

RDF ファイルから特定の配列を抜き出します。

<in|ex>で in を指定すると、--id または--idfile で指定された配列のみが出力されます。<in|ex>で ex を指定すると、--id または--idfile で指定されなかった配列のみが出力されます。

--id と--idfile はどちらか一方のみ指定できます。

--id または--idfile で指定された配列が<RDF_file>に存在しない場合はその配列名は無視されます。

(3) 実行例

```
$ grep_rdf.pl in --id H1,H6,H10,DENISO Xp11_hX.BAMDeniNea.nw.rdf
```

この例では、Xp11_hX.BAMDeniNea.nw.rdf に記載された配列のうち（存在するならば）H1 と H6 と H10 と DENISO のみの RDF ファイルが標準出力に出力されます。

6. Xp11_hX/*

Xp11_hX 領域の VCF ファイル・MAF ファイルと、それらから構築した系統ネットワークおよび中間ファイル。

7. dys44/*

dys44 領域の VCF ファイル・MAF ファイルと、それらから構築した系統ネットワークおよび中間ファイル。

8. rrm2p4/*

RRM2P4 領域の VCF ファイル・MAF ファイルと、それらから構築した系統ネットワークおよび中間ファイル。

[1] VCFtools … URI: http://vcftools.sourceforge.net/

[2] Tabix … URI: http://samtools.sourceforge.net/