デニソバ・ゲノム配列データに関する調査およびパイロット解析

ファイル説明書

2013年5月23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 subst_alt_v3.pl

調査報告書」2(3)③で使用した、VCFファイルに記載された個体の配列をFASTA形式で出力するプログラム。

(1) **機能**

VCFファイルとリファレンスFASTAファイルをもとに、指定された個体の指定されたゲノム領域の配列をFASTAファイルに出力します。プログラムの進行状況は標準出力に出力します。

出力されるFASTAファイルの定義行の名称は、入力VCFファイルに記載された個体名の末尾に -1」・-2」を付けたものになります。一倍体（半数体）では -1」を付けたもののみになります。

(2) **コマンド**

```
$ ./subst_alt_v3.pl <input_fa> <input_vcf> <output_fa> <names> [[start]-[end]]
```

<input_fa>には入力FASTAファイル名を指定します。

<input_vcf>には入力VCFファイル名を指定します。

<output_fa>には出力FASTAファイル名を指定します。

<names>には配列を出力したい個体名を指定します。 ,」で区切って複数指定できます。 --all」を指定すると、VCFファイルに記載されたすべての個体名を対象とします。

[start]には開始座標を整数で指定します。省略された場合は入力FASTAファイルの先頭が開始座標になります。

[end]には終了座標を整数で指定します。省略された場合は入力FASTAファイルの末尾が終了座標になります。

(3) **実行例**

```
$ ./subst_alt_v3.pl hg_ref.fa Deni.vcf Deni.fa --all 50521806-50604915
```

この例では、hg_ref.faをリファレンス配列として、Deni.vcfに記載されたすべての個体について、50521806-50604915の範囲の配列をDeni.faに出力します。

```
$ ./subst_alt_v3.pl hg19_ref.fa Xp11.vcf Xp11.fa HG00096,NA20828,PanTro2
```

この例では、hg19_ref.faをリファレンス配列として、Xp11.vcfに記載された個体のうちHG00096とNA20828とPanTro2について、hg19_ref.faの先頭から末尾までの範囲の配列をXp11.faに出力します。

(4) **制限事項**

① **ゲノム座標**

入力FASTAファイルの定義行には、染色体名と座標がrange=chrN:9999-9999の形式（N: 染色体名、9999: ゲノム座標（桁数は可変））で書かれている必要があります。現在、UCSCテーブルブラウザからダウンロードしたファイルはこの形式になります。

VCFファイルは複数の染色体のレコードを含むことがあります。このプログラムで対象とする領域の染色体名は、入力FASTAファイルの定義行に書かれているものを使用します。

【入力FASTAファイルの定義行の例】

>hg19_ct_Xp11hX_2353_(null) range=chrX:50521806-50604915 5'pad=0 3'pad=0 strand=+

② 個体名

入力VCFファイルに複数の同名の個体が存在する場合は、それらのうちで最後に出現する個体の配列のみが出力されます。これはコマンドライン引数で --all」を指定した場合も同様です。

名称に ,」を含む個体をコマンドライン引数で個別に指定することはできません。コマンドライン引数で --all」を指定することにより、名称に ,」を含む個体の配列も出力できます。

--all」という名称の個体が存在するとき、コマンドライン引数でその1個の個体のみを指定することはできません。 ,」で区切って2個以上の個体を指定する場合は、その中で --all」という名称の個体を指定することもできます。

③ ALTの重複

ある1つの個体について、座標の範囲が重複する複数の変異サイトでgenotypeに1以上の数値[1]が指定されているとき、それらのサイトの中で最初にVCFファイルに出現したもののみが処理されます。

④ <ID>

VCFファイルのREFフィールドまたはALTフィールドに<ID>形式で値が指定されている変異サイトには対応しません。そのサイトは無視されます。

⑤ Phase

指定された領域において、genotypeがunphasedの二倍体であった場合、すなわちVCFファイルのgenotypeの区切り記号[1]が /」の場合、homoのサイトはphasedと同様に処理します。Unphasedでheteroのサイトが存在するときは、配列を決定できないので、その個体については配列を一切出力しません。

⑥ 倍数性

ある1つの個体の染色体が一倍体（半数体）であるか二倍体であるかは、指定された領域において最初にVCFファイルに出現した変異サイトのgenotypeフィールドの区切り記号の有無で判断します。指定された領域内で1つの個体の倍数性が途中で変わった場合は、エラーとして処理を中断します。ただし、unphasedであり配列を出力しない個体については、このエラーは発生せず、次の個体の処理に続きます。

三倍体以上のgenotypeには対応しません。

⑦ リファレンス配列の妥当性

入力FASTAファイルのリファレンス配列とVCFファイルのREFフィールドの値が一致しているかどうかのチェックは行いません。なお、REFからALTに置き換えたときは、そのサイトのリファレンス配列とREFとALTの値を標準出力に出力しますので、これを使用して後で確認することが可能です。

2 Deni_ucsc.vcf

調査報告書」2(3)①でUCSCゲノムブラウザでダウンロードしたデニソバ・ゲノム配列のVCFファイル。

3 Deni.vcf

調査報告書」2(3)②で作成したデニソバ・ゲノム配列のVCFファイル。

4 Deni.fa

調査報告書」2(3)③で作成したデニソバ・ゲノム配列のFASTAファイル。

5 Deni.log

調査報告書」2(3)**③で**Deni.fa**を作成した際の**subst_alt_v3.pl**の標準出力。**

6 all.fa

調査報告書」2(3)**④で作成したデニソバ・ゲノム配列および近縁種の**FASTA**ファイル。**

7 all.aln

調査報告書」2(3)**④で作成した**Clustal**形式のアライメントファイル。**

8 all.meg

調査報告書」2(3)**⑤で作成した**MEGA**形式のアライメントファイル。**

9 all.mts

調査報告書」2(3)**⑤で作成した**MEGA**形式の系統樹ファイル。**

10 all.nwk

調査報告書」2(3)**⑤で作成した**Newick**形式の系統樹ファイル。**

[1] VCF**ファイルの**genotype**フィールドは区切り記号と数値で構成されます。**

区切り記号は「|」がphased**、「/」が**unphased**を表します。**

数値は0**が**REF**の配列、**1**が**ALT**の**1**番目の配列、**2**が**ALT**の**2**番目の配列を表します。**3**以上も同様です。**

(URI: http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41)