1000 Genomes Project data の

vcf ファイルの修正とハプロタイプ再推定

ファイル説明書

2013 年 11 月 28 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. VCF_BAM.pdf

「作業報告書」3.(1)②で参照した、リードの質の値から genotype を推定する基準の詳細。

2. make_bam_dler.pl

「作業報告書」2.(2)②および 4.(2)②で使用した、指定した個体の指定した領域の BAM ファイルを 1000 Genomes Project のウェブサイトからダウンロードするシェルスクリプトを作成するプログラム。

(1) 必要環境

SAMtools（make_bam_dler.pl を実行する際に SAMtools は不要ですが、make_bam_dler.pl で作成したシェルスクリプトを実行する際に SAMtools が必要となります。）

(2) コマンド

```
$ ./make_bam_dler.pl <alignment_index_file> <id_file> <region>
```

<alignment_index_file>には 1000 Genomes Project のインデックスファイル名を指定します。

<id_file>には BAM ファイルをダウンロードしたい個体名を改行(LF)区切りで記載したファイル名を指定します。

<region>にはダウンロードしたい領域を「Reference sequence name:開始座標-終了座標」の形式で指定します。ここに指定した文字列は、出力される SAMtools のコマンドにそのまま記述されます。詳細は SAMtools のマニュアルを参照してください。

(3) 機能

1000 Genomes Project のインデックスファイルと、個体の一覧と、領域をもとに、それらの BAM ファイルを SAMtools でダウンロードするコマンドを標準出力に出力します。

インデックスファイルに記載されたファイル名が 1000 Genomes Project の命名規則にしたがっていれば、作成したシェルスクリプトを実行した際に保存される BAM ファイル名は

/(.*)¥.(ILLUMINA|SOLID|LS454)¥.(low_coverage|exome)¥.bam$/i

にマッチする名称となります。1 つ目の括弧内は個体名となります。

(4) 実行例

```
$ ./make_bam_dler.pl alignment.index ids.txt X:50521806-50604915 > bam_dl.sh
$ ./bam_dl.sh
```

この例では、1 行目で alignment.index に記載された BAM ファイルのうち ids.txt に記載された個体の BAM ファイルの、X:50521806-50604915 の範囲に全部または一部が含まれるリードの情報をダウンロードするための SAMtools のコマンドを bam_dl.sh に保存します。2 行目でシェルスクリプトを実行すると、カ

レントディレクトリに BAM ファイルが保存されます。

3. quality_read.pl

「作業報告書」3.(2)および 4.(3)で使用した、genotype を再推定するプログラム。

(1) 必要環境

SAMtools

(2) コマンド

```
$ quality_read.pl <BAM_file> [<BAM_file> [...]] <Ref_FASTA_file> <VCF_file>
```

<BAM_file>には 1000 Genomes Project の入力 BAM ファイル名を指定します。複数個の指定が可能です。

<Ref_FASTA_file>にはリファレンス FASTA ファイル名を指定します。

<VCF_file>には入力 VCF ファイル名を指定します。

(3) 機能

指定された VCF ファイルに記載された一塩基置換の変異サイトについて、指定された 1000 Genomes project の BAM ファイルのリードの質にもとづいて genotype を再推定し、VCF 形式で出力します。なお、BAM ファイルのリードのフィルタリングのためにリファレンス配列を使用します。

① 処理する個体

入力 BAM ファイル名が

/([^¥/]+)¥.(ILLUMINA|SOLID|LS454)¥.(low_coverage|exome)¥.bam$/i

にマッチするとき、1 番目の括弧内を個体名とみなします。マッチしない BAM ファイルは無視されます。

ここで決定した個体のうち、入力 VCF ファイルに記載されている個体が処理対象となります。入力 VCF ファイルに記載されていない個体は無視されます。

② 処理する領域

リファレンス FASTA ファイルの配列 ID が

/chr([¥d]{1,2}|X|Y)(:([¥d]+)-([¥d]+))?/

にマッチするとき、1 番目の括弧内を染色体名とみなします。2 番目の括弧内が 1 回存在するとき、その範囲の配列が FASTA ファイルに記載されているとみなします。2 番目の括弧内が存在しないとき、染色体の全体の配列が FASTA ファイルに記載されているとみなします。

③ 処理する変異サイト

入力 VCF ファイルに記載された変異サイトのうち、リファレンス FASTA 配列の範囲内にあり、REF や ALT に<ID>が使われておらず、一塩基置換である変異サイトが処理対象となります。そうでない変異サイトは無視されます。

④ リード情報の読み込み

入力 BAM ファイルに対して SAMtools を実行してリード情報を取得します。なお PCR or optical duplicate の flag がオンであるリードは使用しません。

このリード配列とリファレンス FASTA 配列を比較し、連続する 10 塩基(挿入・削除は含めない。)内にミスマッチが 3 塩基以上存在するとき、その 10 塩基すべての allele を N に、質を 0 に書き換えます。

⑤ Genotype の再推定

対象とする変異サイトについて、Brad Chapman の方法[1]を用いて質の値を計算し、genotype を再推定します。

⑥ ハプロタイプの再推定

⑤の結果と入力 VCF ファイルとを比較し、ハプロタイプを再推定して出力します。出力ファイル名は、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".new.vcf"に変えたものになり、そうでないときは入力 VCF ファイル名の後に".new.vcf"を加えたものになります。

(4) 実行例

```
$ ./quality_read.pl HG0009[67].*.*.bam chrX.fa 1000g.vcf
```

この例では、1000g.vcf に記載された一塩基置換の変異サイトについて、HG0009[67].*.*.bam にマッチするファイル、例えば HG00096.ILLUMINA.exome.bam と HG00097.SOLID.exome.bam と HG00096.ILLUMINA.low_coverage.bam と HG00097.SOLID.low_coverage.bam などのファイルに記載されたリードの質に基づいて genotype を再推定し、VCF 形式で 1000g.new.vcf に出力します。なお、BAM ファイルのリードはあらかじめ chrX.fa を用いてフィルタリングされます。

(5) 制限事項

quality_read.pl では 1000 Genomes Project の 1000 人を超えるサンプルの個体名をコマンドラインや別途入力ファイルで指定することなく処理できるようにするため、BAM ファイルのファイル名から自動で個体名を判定することとし、BAM ファイルは命名規則に従ったもののみを扱います。BAM ファイルの具体的なファイル名は

/¥/?([^¥/]+)¥.(ILLUMINA|SOLID|LS454)¥.(low_coverage|exome)¥.bam$/i

にマッチする必要があり、この 1 番目の括弧内を個体名とみなします。これにマッチしない名前の BAM ファイルは無視されます。

4. phase_read.pl

BAM ファイルと VCF ファイルに基づいて、BAM ファイルのそれぞれのリードがどちらの phase のものであるかを判定するプログラム。

(1) 必要環境

SAMtools

(2) コマンド

```
$ phase_read.pl <BAM_file> [<BAM_file> [...]] <Ref_FASTA_file> <VCF_file>
```

<BAM_file>には 1000 Genomes Project の入力 BAM ファイル名を指定します。複数個の指定が可能です。

<Ref_FASTA_file>にはリファレンス FASTA ファイル名を指定します。

<VCF_file>には入力 VCF ファイル名を指定します。

(3) 機能

指定された BAM ファイルに記載されたそれぞれのリードの配列と、指定された VCF ファイルに記載された変異サイトの phased genotype とを比較し、リードがどちらの phase のものであるかを判定します。

その後の解析で使用するため、リードの情報を個体ごとに次の 5 種類の SAM ファイルに分けて出力します。

① 1 番目の phase のみにマッチするリード

変異サイトを含んでおり、1 番目の phase のみにマッチするリードを出力します。ファイル名は、元の BAM ファイルの個体名の部分の末尾に"-1"を加え、拡張子を"sam"にしたものになります。

② 2 番目の phase のみにマッチするリード

変異サイトを含んでおり、2 番目の phase のみにマッチするリードを出力します。ファイル名は、元の

BAM ファイルの個体名の部分の末尾に“-2”を加え、拡張子を“sam”にしたものになります。

③ 変異サイトを含んでおり、どちらの phase にもマッチするリード

変異サイトを含んでおり、両方の phase にマッチするリードを出力します。ファイル名は、元の BAM ファイルの個体名の部分の末尾に“-H”を加え、拡張子を“sam”にしたものになります。質の値は元の BAM ファイルの質の値の半分にしたものを出力します。

④ 変異サイトを含んでいないリード

変異サイトを含んでいないので phase を判定できないリードを出力します。ファイル名は、元の BAM ファイルの個体名の部分の末尾に“-N”を加え、拡張子を“sam”にしたものになります。今後の解析で使用しないので、質の値は③のような半分にする処理はしていません。

⑤ どちらの phase にもマッチしないリード

変異サイトを含んでおり、どちらの phase にもマッチしないリードを出力します。ファイル名は、元の BAM ファイルの個体名の部分の末尾に“-U”を加え、拡張子を“sam”にしたものになります。

(4) 実行例

```
$ ./phase_read.pl HG0009[67].*.*.bam chrX.fa 1000g.vcf
```

この例では、HG0009[67].*.*.bam にマッチするファイル、例えば HG00096.ILLUMINA.exome.bam と HG00097.SOLID.exome.bam と HG00096.ILLUMINA.low_coverage.bam と HG00097.SOLID.low_coverage.bam などに記載されたそれぞれのリードの配列と、1000g.vcf に記載された変異サイトの phased genotype とを比較し、リードがどちらの phase のものであるかを判定します。

出力ファイル名は、“HG00096-1.ILLUMINA.exome.bam”、“HG00097-N.SOLID.exome.bam”などのようになります。

(5) 制限事項

phase_read.pl では 1000 Genomes Project の 1000 人を超えるサンプルの個体名をコマンドラインや別途入力ファイルで指定することなく処理できるようにするため、BAM ファイルのファイル名から自動で個体名を判定することとし、BAM ファイルは命名規則に従ったもののみを扱います。BAM ファイルの具体的なファイル名は

/([^¥/]+)¥.(ILLUMINA|SOLID|LS454)¥.(low_coverage|exome)¥.bam$/i

にマッチする必要があり、この 1 番目の括弧内を個体名とみなします。これにマッチしない名前の BAM ファイルは無視されます。

5. Xp11_hX.vcf

「作業報告書」4.で使用した Xp11_hX 領域(X:50521806-50604915)の VCF ファイル。

6. Xp11_hX_exome.sh

「作業報告書」4.(2)②で作成した、Exome の BAM ファイルをダウンロードするためのシェルスクリプト。

7. Xp11_hX_lowcoverage.dl.sh

「作業報告書」4.(2)②で作成した、Low Coverage の BAM ファイルをダウンロードするためのシェルスクリプト。

8. qual/*.qual

「作業報告書」4.(3)で作成した、変異サイトごとの allele ごとの質の値および再推定したハプロタイプを個体ごとに保存したファイル。

9. Xp11_hX.regen.vcf

「作業報告書」4.(3)で作成した VCF ファイル。

[1] Brad Chapman の方法 … URI:
https://bitbucket.org/chapmanb/synbio/src/7b1b3a972b7e/SynBio/Sequencing/ConstructAnalysis.py
同一プラットホームの同一 strand のリードは、質の値が大きい順に並べて、n 番目のものに 1/n を掛けたものを合計する。それ以外のリードは質の値を合算する。