

概要

- AMEDデータ利活用プラットフォームの利用申請でアクセス可能な約2万件について、日本人由来がん細胞株の全ゲノム配列データからGermline由来のSNPs、SVsの除去を目的に本PFのゲノムデータを取得。
- 特にCoding領域について非同義的置換の多くからGermline SNP由来のものを差し引くことができ、またSVsについてもGermline由来のものも多くを除去することができた。

背景

- がん細胞において体細胞突然変異（SNVs）、あるいは構造異常（SVs）の検出を試みる場合、生殖系列多型（Germ line SNPs/SVs）の除去を行う必要があるが、多くのがん細胞株においてはその同一個体のゲノムDNAが入手可能でない。
- 今回、バイオバンクで公開されている日本人ゲノムで登録されているSNPs、SVs情報を取得し、日本人由来がん細胞株の全ゲノム配列データから、それらを除去した。
- さらに我々が収集した、long read解析あるいはimputationがなされているデータを用いて、これらの体細胞変異がどちらの染色体（ハプロタイプ）に生じているのか、いわゆるPhasing解析を行った。
- 同様のアプローチを正常DNA検体および配列情報が入手できないがん検体についても援用した。



AMEDデータ利活用プラットフォーム

3大バイオバンク等のゲノムデータのメタデータの横断検索およびvisiting環境での解析が可能

- サンプル数：約2万件
- サンプルメタデータ：健康者および多様な疾患患者の全ゲノムデータ、臨床基本4情報（性別、年齢階級、出生地/居住地、疾患名・ICD-10コード）

方法

AMEDデータ利活用プラットフォームを利用

PF外での解析



1

PF利用申請

- プラットフォーム内のアレル頻度情報・臨床基本4情報を利用するため、PF利用申請



2

生殖系列多型の除去の実施

- 横断検索によりSNPs、SVs情報を探索、変異Call実施（GATK、NanomonSVを使用）



3

自身で収集したデータを使用したPhasing解析の実施

- Callされた変異の異同検証のため、自身で収集したデータを用いてPhasing解析を実施（WhatsHapを使用）

結果

- 特にCoding領域については非同義的置換の多くからGermline SNPに由来するものを差し引くことができた。
- またSVsについても生殖系列に由来するものも多くを除去することができた。

考察

- 直接のGWAS解析に加えて、基礎研究においてもバイオバンクデータが有効に機能することを示すことができた。