

3-3-2-4 集団間の民族学的遺伝的差(階層化を含む、その問題点と存在の有無の検定)

3-3-2-4-1 集団間の民族学的遺伝的差について

初めに、“集団間の民族学的遺伝的差”について説明する。ある集団とある集団が民族学的遺伝的に等しいという場合、その両集団は十分な期間、互いに交流して無作為な交配が行われ続けてきたことが必要である。このことを裏返せば、どんなに長い期間、距離的に近い関係にあったとしても無作為交配が行われてこなかった集団の間には何らかの民族学的遺伝的差が残ったり、蓄積したりする。このような集団間の民族学的遺伝的差は、アレル頻度・genotype 頻度の異なる多型がゲノムワイドに多数存在することで確認される。もし、比較する集団間の民族学的遺伝的差が小さい場合には、アレル頻度・genotype 頻度が異なる多型の数が少なく、また、アレル頻度・genotype 頻度の差が小さくなる。

ここで、話をケース・コントロール関連解析に戻す。統計的有意差を持つ疾患関連マーカーが検出されるということは、ケース群とコントロール群のマーカーのアレル頻度・genotype 頻度が異なり、その差が検出されるということである。しかしながら、このようなマーカーの genotype 頻度・アレル頻度の差は必ずしも疾患と関連しているとは限らない。このように疾患関連マーカーを検出しようとして、疾患と関連のないマーカーが誤って疾患と関連があるように検出された場合、その検出結果を偽陽性と呼ぶ。

偽陽性の原因は簡単に大きく分けて 2 つある。

偽陽性 A

ケース検体とコントロール検体をそれぞれの母集団から抽出したときに、母集団を代表する検体が選ばれなかったための偽陽性である。これはサンプリングバイアスによる偽陽性と呼ばれ、偶然に支配されているため、回避できない。

偽陽性 B

ケース母集団とコントロール母集団との間に民族学的遺伝的差が存在したために検出されてしまったという場合である。こちらはケース母集団とコントロール母集団の選び方を理想化すれば、回避可能であるし、本来回避すべき状況である。

しかしながら、厳密にケース母集団とコントロール母集団を同一の一般母集団の中に設定し、偽陽性 B を完全に回避することは難しい。その理由も大きく分けて 2 つある。1 つは、対象疾患に罹りやすい集団と罹りにくい集団が混在している集団において、ケース・コントロール検体抽出を行う場合である。この場合はサンプル抽出の結果として得られるケース検体に罹患率の高い集団由来の検体が多くなり、逆にコントロール検体に罹患率の低い集団由来の検体が多くなる。このようにある疾患の罹患率について民族学的にムラがあるとき、その疾患についてこの集団は階層化しているという。従って、罹患率の異なる 2(複数)集団が存在していることが知られており、検体抽出時にその 2(複数)集団を区別することができれば、その 2(複数)集団を区別して検体抽出を行うことが必須である。しかしながら 2(複数)集団が区別できない場合には、この混合母集団を用いる限り、階層化の影響を排除することは不可能であり、その結果として生じる偽陽性の検出も避け得ない。

もう 1 つは、検体抽出の方法が不完全であるために、ケース母集団とコントロール母集団とを厳密に共通な一般母集団内に設定できない場合である。これは、検体抽出の現実的方法に由来する問題である。

ケース検体は通常、ある医療機関において治療を受ける罹患者集団から選ばれる。一方、コントロール検体はそれとは異なる理由で選ばれる(研究のために募られたボランティア、健康診断などのために集まった集団からの抽出、ケース抽出医療機関における他疾患罹患者集団からの抽出など、その方法は様々である)ため、ケース母集団と同一の母集団からのコントロール検体の抽出というのは難しい。RIKEN SRC では、大都会を中心とした複数地域から、罹患者を複数の疾患について個別に収集し、その混合を日本人の一般集団とみなしてコントロールとしている。この設定の背景には日本人は均一であるとの仮定、もしくは都会は多くの地域から流入した人々の混合であるので日本国内の民族学的遺伝的差が消失しているという仮定がある。

このようにケース・コントロール関連解析には常に両群の民族学的遺伝的差が存在するための偽陽性 B の可能性を疑う必要がある。ゲノムワイド関連解析の場合には、回避不可能なサンプリングバイアスによる偽陽性 A も多いことから、陽性結果の解釈は難しい。従ってケース・コントロール集団間の民族学的遺伝的差による偽陽性 B はできるだけ避けたいものである。

実は、複数の集団の間に民族学的遺伝的差があるか否かは多数の遺伝的マーカーについてのタイピング結果を統合して検討することにより、偽陽性 B が混入しているか否かを検定できる。これは、次のような論理に基づく。以下は、単純化のために 2 群間に民族学的遺伝的差があるか否かの解析について述べてある。

ゲノムワイドに多数のマーカーについて集団間有意差検定を行った場合、2 群間に民族学的遺伝的差が無いならば、有意差を認めるマーカーは疾患と関連するマーカーであるか、サンプリングバイアスに由来する偽陽性 A であるはずである。疾患と真に関連する遺伝子の数は限られており、真に関連するマーカーの数は無視できるほど少ない。従って、有意差を認めるマーカーはサンプリングバイアスに由来するもので全てが占められると考えればよい。このような場合、確かに相当数の“偽陽性”が検出されるが、その検出頻度は確率論的に予想された値に近くなるはずである。このことを利用して、多数のマーカーについての群間有意差検定値を集計してその集計値がサンプリングバイアスで説明されるべきか、それ以外の原因(両群間の民族学的遺伝的差)に由来するものと判断されるべきかの検定が可能である。簡単に言うと、有意差のあるマーカーの数が、偶然生じる個数程度であれば民族学的遺伝的差は無いが、有意差のあるマーカーの個数が偶然では説明できないほど多ければ民族学的遺伝的差があると判断される。

具体的には、全てのマーカーに関して得られる²検定値の総和を、そのマーカーの数の自由度の²分布に照らして両群間に民族学的遺伝的差が無いという帰無仮説が真である確率を求めることで検定する。RIKEN SRC のデータは定期的にこの検定を行い、民族学的遺伝的差が示唆される検体群を検出する予定である。