# SAS/Geneticsの紹介

池田匡志　尾崎紀夫

従来の分子遺伝学は，血友病やハンチントン舞踏病のような単一遺伝子疾患の病態生理の解明に大きな功績を残した。最近，多数の遺伝因子と環境因子が発症に関与し，しかも発症頻度の高い複雑疾患（complex disease）である糖尿病，高血圧，うつ病などの病態生理を遺伝子レベルから解明しようとするゲノム研究がなされている。複雑疾患のゲノム研究を支えているのは，簡易に，安価で，大量のSNP genotyping[1]が短時間で可能とする技術革新である。すなわち，複雑疾患のゲノム研究を行う研究者は，以前主流であったPCR-RFLP法（少なくとも筆者のラボでは）では，いかに努力しても１日で1000-1500 genotypingがせいぜいであったが，今ではTaqMan法を使えば，数千genotypingから1-2万のgenotypingが，簡単に，しかもhuman errorが少なく可能となるという技術革新の恩恵を受けている。このような状況下，すなわちwet labo workの比重が軽減され，算出された多数のデータを遺伝統計学的に解析する，dry labo work，bioinformaticsの重要性が高まっている。

Bioinformatics系の雑誌が注目され，impact factorが上昇する中，多くのsoftwareが紹介され，ほとんどがfreeでdownloadできる（とくに我々がしばしば活用しているのは‘Bioinformatics’に載っているsoftwareで，すべてfreeでdownloadできる：http://bioinformatics.oxfordjournals.org/）。しかし，筆者を含め，我々のlabo memberはUnixやLinuxの知識は乏しく，softwareをdownloadしても実際動かせないこともしばしばある（もっとも，最近ではWindowsをサポートしたsoftwareが多くなってきている）。また，たとえ動かせたとしても，data inputのformatが各softwareで異なることが多く，その修正に時間をとられることが多い。

今回紹介するSAS/Geneticsは，とくにgenetic association analysisを行う研究者を対象に作られている。また，user friendly でありcomputerにそれほど詳しくない研究者でも容易に扱うことができ，かつ，遺伝統計の分野で用いられる基本的な解析法はほとんど網羅されている。さらに，excel dataを読み込むだけでdata inputが完了し，簡便に扱うことができる点も優れており，はじめて解析を行うものにとっては魅力的なsoftwareである。

以下に特にお薦めの「特性・利点」を挙げ，紹介する。

## Ⅰ．簡単

SASといえば，macroを使用して難解なイメージがある。しかし，SAS/Geneticsのmanualは丁

1　SNP genotyping：数100塩基に一個程度の割合という高い密度でゲノム上に存在するSingle Nucleotide Polymorphism（一塩基多型：SNP）の有無を確認すること。

寧に，わかりやすく書いてある。さしあたり，exampleのmacroを少し手直しするだけですべての解析が可能となる。一旦macroを作ってしまうと，それを保存しておけば，だいたい同じ解析をすることが多いので，input dataの名前と，SNPsの数を変えるだけで終了してしまう。

また，先にもあげたinput dataのformatがexcelベースで作成，読み込むことが可能である点も，このsoftwareを扱いやすいものとしている。例えばPed fileを作成するにしても，自分で変換プログラムを作成できない筆者は，excelとtextを駆使してPed fileを作成している。この方法も，特別面倒なものではないが，やはりexcelをそのままimportできる点は，なんともありがたい。

### Ⅱ．迅速

Phaseが不明なgenotype dataからhaplotypeを推定する場合，EMアルゴリズムを用いる方法がもっとも一般的である。2個のSNPsからhaplotypeを推定する場合は，だいたいどのsoftwareを用いても一瞬で終了してしまうが，多くのSNPsから推定する場合，SAS/Geneticsではその速さにびっくりしてしまう（もちろん，そんなに多くのSNPsでhaplotypeを推定することは少ないと考えられるし，意義は薄いかもしれないが）。また，haplotype-wiseの解析で，global P値を出すときに用いるpermutation testも，他のsoftwareより数倍早い印象を持っている。これらのことは，SAS/Geneticsが洗練されたprogramで書かれていることが推定される。

### Ⅲ．解析方法の幅広さ

SAS/Geneticsは，有名なB.Weirらのgroupが中心となって開発された解析方法を基に構成されている。彼らはGDA（http://hydrodictyon.eeb.uconn.edu/people/plewis/software.ph）やPOWERMARKER（http://statgen.ncsu.edu/powermarker/）といったfreeのsoftwareを提供している。特にPOWERMARKERはSAS/Geneticsより多い解析方法を展開しているが，基本的にはSAS/Geneticsに構成される解析方法で十分であり，不便さはまったく感じない。また，本SAS/Geneticsのversionでは，以前のversionにはついていなかった機能も多々ある。その中で，いわゆるtagging SNPの選択が注目される。このtagging SNPsは，連鎖不平衡（LD）[2]にある領域（LD blockとかhaplotype blockと呼ばれることもある）のなかで，その領域を代表しうる「tagとなる」SNPsを抽出し，そのSNPsで関連を見ることで，redundantな解析を避け（多重比較によるtype Ⅰ error rateの上昇を防ぎ），randomにpickupしたSNPsを用いた解析よりpowerの高い解析を行うことを目的としたものである。この「tagとなる」SNPsの選出方法は多くのcriteriaが提唱されているが，決定的な方法はいまだ確立されていない。SAS/Geneticsでは，Johonson et al（2001, Nat Genet）の方法を用いている（最近では，LD evaluationにもっとも用いられるsoftware HAPLOVIEW；http://www.broad.mit.edu/mpg/haploview/にのっているtaggerを用いる人が多いかもしれないが，criteriaが厳しすぎる，という意見も聞く）。

---

2　連鎖不平衡（LD）：ある集団において，複数の遺伝子多型が独立して存在せず，何らかの関係を持って存在することである。LDの強さの評価にはLewontin's D'とHill's $r^2$といった指標を使うが，どちらも０－１の範囲をとり，これらの指標が１に近いときLDが強い。

以上，利点を述べたが，SAS/Geneticsにももちろん弱点もある。筆者は主にassociationをhaplotype-wiseに検討する解析を行っているが，例えば，頻度の低いhaplotypesを解析に用いることは，powerが下がることが予想される。そのために，頻度の低いhaplotypesを除外して解析できるsoftwareもあるため（例えばUNPHASED; http://www.hgmp.mrc.ac.uk/~fdudbrid/software/unphased/），そちらを用いることもある。また，最近のHapMap projectの進展に伴い，LDの評価には，HAPLOVIEWが多く用いられている。これはHapMapからdownloadしたdataを直接利用することで，民族のLDを見ることができるものであるが，graphicalにも有利であり，この用途ではSAS/Geneticsは不利な点であろう。

SAS/Geneticsは初めてSNPsのassociation analysis，特にhaplotype-wiseの解析を行う研究者には特にお薦めである。一度使用してみていただきたい。

（いけだ　まさし：名古屋大学大学院医学系研究科精神医学分野）
（おざき　のりお：名古屋大学大学院医学系研究科精神医学分野）