

Genetics by numbers

ナンバー登録による遺伝学

Nature Vol.451 (516-518) / 31 January 2008

全ゲノム相関解析によって、信頼性の高い疾患マーカーが次々と見つかった。その現況とこの先待ち受けていることについて、Monya Baker が取材報告する。

人の将来の健康を車のナンバープレート・リストと同じように読み取れるなどとは、かつて誰も思わなかっただろう。昨年、一連の研究¹⁻³によって、心疾患を発症する確率の上昇と、rs10757274 や rs1333040 といった一見意味不明の番号との相関性が明らかにされた。これらの番号はそれぞれが、専門用語でいう塩基多型 (SNP; スニップ) にあたるもので、第9染色体上の互いに近い位置にある。これらの SNP が実際にどのような作用を及ぼしているのかはわからない。しかし、このうちのどれかを2コピーもつ人は、心疾患を発症する確率が、タバコを1日10本吸う人と同じくらいになる。その影響は糖尿病やタバコの吸い過ぎが原因のものよりは少ないものの、これ

までの全ゲノム相関解析で見つかった最強のリスクの1つであることは確かである。全ゲノム相関解析 (genomewide association study) とは、あるヒト集団において別のヒト集団よりも高い頻度で存在する遺伝子多型を見つけ出すとする研究である。

rs10757274 や rs1333040 と心疾患の相関性は非常に堅固なものであり、心血管疾患についてより詳細な全ゲノム相関解析を行う際のポジティブコントロールとして既に用いられている。「もしこの SNP と心疾患の相関性が認められない場合、解析過程で何かが間違っているといえます」とオタワ大学心臓研究所 (カナダ) の Ruth McPherson は語る。McPherson は、2007年5月、ゲノム相関解析結果について最初に報告さ

れた2報のうちの1つ¹の論文筆頭著者でもある。もう1本の論文²は、アイスランドのレイキャビクにある deCODE Genetics 社の研究チームによるものだ。これらの「悪い」SNP のどれかを1コピーもっているとリスクは約40%上昇し、2コピーもっているとその人が生涯で心臓発作を起こす確率は倍加する。

しかし、相関性が強くても、その意味を見つけ出すことはむずかしい。全ゲノム相関解析から、集団内のゲノムに高頻度で塩基1個が変化している SNP が見つかったが、それらは近隣 DNA のほかの多型とも相関している可能性がある。SNP は必ずしも疾患の原因となる変異ではない。SNP は、遺伝子のタンパク質コード領域にも発現調節領域にも存在していない可能性があり、そ



ができる。初期の解析では、そうした相関には再現性が認められず偽陽性（誤検出）が多数見つかるとして、芳しくない評判が立った。しかし、統計学的知識が積み重ねられ、より大規模な集団でより多くの SNP が解析されたことで、状況は変わりつつある。2007年には再現研究から、2型糖尿病やクローン病、心血管疾患などの「ありふれた疾患」と相関する遺伝子座が突き止められ⁴（コラム「SNPと疾患の一例」を参照）、この年は全ゲノム相関解析の年とよぶべきものになった。複数の研究チームが、詳細情報がなければ意味のない非常に多数のナンバープレートのような SNP と、遺伝子領域とを相関させた緻密なリストを提示したからである。2008年と2009年はどうなるかと研究者に問えば、さらに膨大なリストについて大いに語ってくれることだろう。より多くの SNP についてより大きな集団を精査すれば、もっと高い再現性で、さらに多くの疾患と関連づけられるはずだと。

だが、この数字と文字から成る長くてまぎらわしい番号が、薬剤の探索や疾患の説明、患者それぞれに合わせた治療法の推奨などの目的に、いったいいつごろ使用できるようになるかを問われれば、まだ数年の研究を待たねばならないというのが妥当な回答だろう。研究者らは、SNP 精査解析では、環境要因や遺伝子の過剰コピーのような疾患の重要な関与因子を見つけることはできないことを、あっさり認めている。しかも、見つかった多型（対立遺伝子）は、全リスクのごくごく一部にしか関与していない。それでも、全ゲノムを精査することで、相関解析は遺伝子と疾患の思いもよらない結びつきを明らかにできる。「疾患の原因の可能性のある SNP を見つけることは、疾患の基盤にある生物学的特性を見つけ出すことと同じではありませんが、疾患説明と治療法開発へ向けた大きな一歩にはなりません」と、ブロード研究所（米国マサチューセッツ州ケンブリッジ）の人類遺伝学者 David Altshuler は語る。

いくつかの企業は、疾患と相関する遺伝子多型の検査試薬を企画したり、既に販売したりしている。2007年11月、deCODE社は医師向けに、個人の9p21.3に心疾患発症リスクのある対立遺伝子が何コピーあるかを調べる検査試薬を販売し始めた。この種の検査試薬に懸念を表しているのが、疾病対策センター公衆衛生ゲノミクス局（NOPHG；米国ジョージア州アトランタ）の局長 Muin Khoury である。彼は、この問題に対する警告を含んだコメントをいくつか書いており、「たとえ相関性が多数の研究で再現されたとしても、情報としてはまだ脆弱です」と話す。現状では、家族歴解析のほうが、疾患に対する予測能が高いのである。Khouryによれば、遺伝子検査の結果によって人々がもっと健康なライフスタイルをとるようになるという証拠はなく、逆にそうした検査により、自分には疾患のリスクがないと誤った期待を生み出す可能性さえある。発症リスクのある対立遺伝子がそうした検査で見つからなかった人でも、まだ見つかっていないだけであって、実際にはリスクのある多型をもっている可能性があるからである。これらの検査試薬の有効性を実証するには、まだ時間がかかりそうである。Khouryによれば、「民間企業はこの種の研究をしようとはしていないし、今のところ民間企業がそれをすべきだという意見もない」という。

連携が解決のカギ

こうした現状にもかかわらず、疾患の説明がさらに進むことを期待して、新たな相関を探し出そうと研究が進められている。最終的には、非常に重要な遺伝子すべてで高頻度に見られる多型は、疾患との相関においてあまり影響しないということになるだろう。そして、そうした多型の発見は、データをふるい分けするのに十分な検体数があるかどうかにかかってくる。全ゲノム相関解析には多額の費用がかかる。費用は以前に比べかなり下がったものの、数十万個の

の場合は間接的にでも疾患のリスクに寄与しているとは考えられない。心疾患リスクと相関する SNP からわかるのは、疾患原因の「容疑者」SNP が第9染色体短腕上の9p21.3とよばれる長い領域内に位置していることだけである。この領域内にはタンパク質をコードする遺伝子は見つかっておらず、また近隣の2つの遺伝子（どちらも腫瘍抑制因子）解析からも説明がつかない。

番号の意味を理解する

全ゲノム相関解析に熱心な人々は今も、rs10757274とrs1333040のような謎解きのキーワードをもっと見つけ出そうとしている。1つの疾患もしくは症状に関して何十万個もの SNP を精査することで、数十例の相関を見つげ出すこと

SNP の遺伝子型の解析に、まだ 1 人当たり数百ドル（およそ数万円）もかかり、しかも患者データの収集と管理という費用のかかる込み入った作業は計算に含まれていない。「費用に関しては標本サイズ（解析対象とする集団などの大きさ）が大きな障壁となっており、この問題を解決できる最良の方法はデータを共有することです」と、ワシントン大学（米国シアトル）の統計遺伝学者である Lon Cardon は話す。

標本サイズを拡大するために、ウエルカムトラストは自らケースコントロール・コンソーシアム（WTCCC）を立ち上げ、数種の疾患それぞれについて 2000 例の患者と 3000 例の遺伝的対照群（健康な人）について調査し、合計およそ 1 万 9000 のデータを集めた。この手法は、診断確定済みの集団と診断未確定の集団で SNP を比較することによ

て、遺伝的リスクとの相関がみえてくるはずだという根拠に基づいている。

見落としのおそれ

英国レスター大学で心臓病学研究に携わり、WTCCC と共同で冠状動脈性心疾患を担当する 2 名の研究者の 1 人、Nilesh Samani は、たとえ大きい標本サイズで解析したとしても、疾患に対する影響が小さい多型は見逃されてしまうだろうと述べている。彼によると、仮に 1 つのゲノム内に 1 個につき発症確率を 20% 上げる遺伝子座が 10 個あるとして、2000 件の症例と 2000 件の対照例を解析した場合、統計学的にみて、10 個の遺伝子座のうちおよそ 3 個で見つかるだろうという。同様の標本サイズの独立した別の集団でも、2 個か 3 個の疾患に影響を与える遺伝子座が見つかるかもしれないが、それらは違う遺伝子座である可能性があり、また偽陽性のために結論が出なくなるかもしれない。「これらの解析結果をすべて 1 つにまとめてプールして初めて、原因となる遺伝子座を網羅することが現実的に可能になるでしょう」と Samani はいう。

WTCCC が提供するデータは主として、遺伝子型、年齢、性別、疾患の有無に限られている。このデータは一部の研究にとっては非常に貴重であることは確かだが、コレステロール値や血圧といった特定形質に遺伝子多型がどう影響するのかを見つけ出すには、もっと情報量の豊富なデータが必要である。それを満たす、特に情報の豊富な「鉱脈」の 1 つは、フラミンガムの SHARe（SNP Health Association Resources）というデータベースである。これはフラミンガム心臓研究に参加した 9000 人を超える被験者を対象に、55 万個の SNP の遺伝子型を解析したデータである。

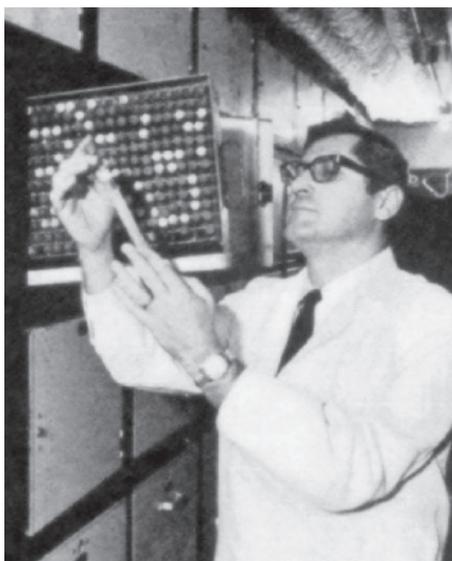
米国立心肺血液研究所（メリーランド州ベセスダ）による資金提供の下、「60 歳フラミンガム研究」とよばれる調査で、マサチューセッツ州の 3 世代の人々が追跡調査された。そこから得られた

データから、喫煙と高コレステロールが心疾患のリスクであることが確実にになった。このデータセットには、すべての被験者についてすべての表現型データが存在しているわけではないが、血液分析から血管造影、ライフスタイル調査までの数千の臨床的変数が含まれており、場合によっては同一個人から数年間にわたって得られたデータもある。このデータベースの構築にあたっては、古い医療記録や刊行物の考古学的な調査をフラミンガム市に要請する必要があったが、完成したデータベースは SHARe という頭字語が意味するように、現在ほかの研究者も利用可能になっている。

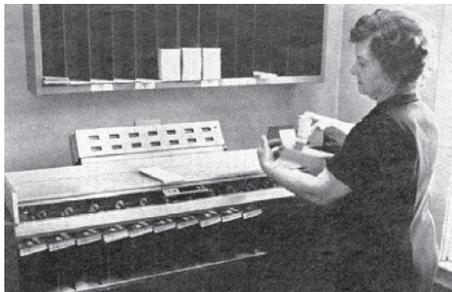
2008 年 1 月 25 日現在、米国立衛生研究所（NIH）を介して資金提供を受けているすべての全ゲノム相関解析は、そのデータを遺伝子型・表現型データベース（dbGaP）におさめて保管するよう通達を受けている。データを収集した研究者は少なくとも 9 か月間、自身の解析結果を独占的に公表する権利を有するが、プライバシー保護の遵守を約束した研究者は、ほかの研究者のデータにいち早くアクセスすることができるようになる。SHARe も同様の猶予期間を設定している。

こうしたやり方はすでにゲノム解析研究では標準になっている。それにもかかわらず、なかには懸念を抱いている人々もいる。新たな興味深い問題を解明するためにはデータが必要だが、もしデータをダウンロードした研究者がそれを解析して結果を発表できるというなら、データ収集の動機づけがほとんどなくなるのではないかと、というのである。また、研究者がダウンロードで得たデータを不適切に扱うことも考えられる。ワシントン大学（米国シアトル）の心臓学者で疫学者の Bruce Psaty は、2 つの大規模解析から入手可能なデータを、研究者たちがどのように利用したかを調査した。すると、一部の研究者はオリジナル研究の承諾の範囲を超えていただけでなく、自身の解析において、オリジナル研究の意図するところから逸脱して

FRAMINGHAM HEART STUDY



FRAMINGHAM HEART STUDY



William Kannel（上）は、手間ひまのかかるフラミンガム心臓研究（下）の第二ディレクターだった。

いるものがあることがわかった。例えば、患者の初回心臓発作の予測因子を探す解析研究なのに、すでに心臓発作を起こしたことがある患者を除外していないものがあつた⁵。大規模な疫学調査の解析が不十分だと、本当の相関が見落とされたり、誤った相関を見つけたりする可能性があるということだ。どちらにしても、情報資源が無駄になってしまう。

もちよって公平に分ける

もし研究者たちが慎重であって審査する側も厳正である場合、より多くのデータに、より多くの研究者がアクセスするようになれば、今よりも少ない費用でよりすぐれた研究ができるようになるはずだと Altshuler はいう。「これほど大規模なデータセットを、1つの研究グループだけですべて利用し尽くすことはできません。このデータを集めたグループの独創性よりも、世界全体に内在する独創性のほうがはるかに大きいはずですから」。deCODE Genetics 社の最高責任者である Kári Stefánsson によれば、研究者たちは既に共同研究の相手をうまく見つけ出しているというが、NIH 権能の彼いわく、旧ソビエトのような「支配体質」には憤りを感じるという。「私は NIH に命令されたから、皆とデータを共有しようと思っているわけではありません。自らそうすべきだと考えるから、データを共有しようと思うのです」。

米国立ヒトゲノム研究所の集団ゲノミクス局（メリーランド州ベセスダ）で局長を務める Teri Manolio は、そうした懸念があることを認めつつも、科学は適応していくものだと言語。WTCCC も、遺伝子相関情報ネットワーク（WTCCC と同じような目標をもつ米国の官民共同研究組織）も、ヒトゲノム解読計画から生まれたものである。この「ビッグサイエンス」計画が軌道に乗り始めた当時も、自分のデータをウェブ上に載せてしまうと研究者としてのキャリアが終わってしまうのではないかという「大きな懸念の声」が上がったと、Manolio は回想する。だが、それは杞

憂に終わった。「塩基配列データを掲載するだけでは論文発表にならないことや、データの共有が実際に人々に危害を及ぼすようなこともないことがすぐに明白になったからです」と Manolio はいう。そして実際、dbGaP への最初のデータ提供者の中には、データ共有によってさらに共同研究者を得たり認知度が得られたりした研究者もおり、現在もいくつかの学術機関研究グループや企業研究グループが、要請されていないにもかかわらずデータを提供する予定でいる、と彼女は付け加えた。

大規模データの必要性も偽陽性の危険性も現在広く認識されていることを受けて、Cardon は、興味深く説得力のある解析に必要なことをするために、研究者たちが互いに信頼できるようになってほしいと考えている。「dbGaP へデータを上げることは本当に役立つことだと思いますが、だからといってこれがデータを集めた人たちの共同研究の代わりになるわけではありません。データをダウンロードするだけや単独研究よりも、共同研究のほうがずっと威力があるのは確かです」と彼はいう。

数の上での強み

多数のデータが集まった場合、結果は明白なものになる。糖尿病遺伝学イニシアチブのデータを使って調べている研究グループは、血中脂質濃度との相関について 2758 人の数十万個の SNP を探索したのち、196 個の SNP を詳しく解析することを決めた。すると、6058 人で SNP と血中脂質濃度を測定した別の 2 つの研究データを保有する研究グループから問い合わせを受けた。

2 つの研究グループはデータを共有することで、さらに詳細に解析すべき SNP をより多く見つけ出すことができ、どちらの研究グループからも研究報告がなされた^{6,7}。その共同研究で見つかった、第 1 染色体短腕上にある rs599839 という SNP⁷ は、低密度リポタンパク質コレステロールの値が低いことと関連していることがわかり、さらに詳しい解析か

SNP と疾患の例

 RS10811661

糖尿病

rs10811661 と糖尿病の相関は、32500 を超える人から得た試料を使って最初に見つけ出し確認されたもので、2007 年に初めて発表された⁹。rs10811661 は第 9 染色体上のある遺伝子の近くで見つかり、マウスの研究では、この遺伝子が過剰に発現すると膵島細胞の再生を妨げることが明らかになっている¹⁰。その後の研究で、このリスク SNP を保有する人たちは、グルコースを摂取したときのインスリン放出量が正常な人に比べて少ないことがわかった¹¹。

 RS2241880

クローン病

第 2 染色体長腕上にある rs2241880 という SNP は、クローン病に関するいくつかの全ゲノム相関解析¹²⁻¹³で浮かび上がってきたが、潰瘍性大腸炎などほかの炎症性腸疾患とは相関がみられなかった。この SNP は、腸内壁で発現して細胞内細菌を分解処理するのを助ける遺伝子のコード領域内にある。培養細胞では、この遺伝子を発現しない細胞はサルモネラ菌感染を防御する能力が劣っていた。

 8Q24

がん

複数の研究¹⁴⁻¹⁵から、第 8 染色体長腕上の 8q24 とよばれる、遺伝子のほとんどない領域で SNP 群が見つかった。これらには、乳がん、結腸直腸がん、前立腺がんとの相関がみられた。その 1 つである rs16901979 は、アイスランド集団で最初に見つかり、次に、アフリカ系米国人男性で見つかった。病態の根源にあるメカニズムは何もわかっていないが、この領域はやみくもに相関を探さなければおそらく見つからなかっただろう。

ら、肝臓で3つの遺伝子の発現が増大して、そのうち1つの遺伝子がグルコース取り込みに関与していることが明らかになった。おもしろいことにそのSNPは、わずか数か月前に、9p21.3の心疾患リスクを実証した研究³においても、心疾患と関連づけられていた。

SNPの影響を解明しようとする研究では、もっと情報の豊富な表現型を追跡して、最終的に実験によって機能を確認することになる。最近、deCODE社によって、心疾患と関連づけられた9p21.3領域上にあるSNP群は、脳や腹部の動脈瘤とも関連していることが明らかになった⁸。McPhersonの研究チームは、この遺伝子座が動脈疾患と関連することを見だし、この領域には血管壁の保全に関与する「何か」があることを示唆した。それにもかかわらず、基盤となるメカニズムの解明は、どちらの研究チームにとってもまだ手探り状態にある。「私たちはこの領域の塩基配列を徹底的に解析していますが、何も

新しいことは見つけ出せていません」とStefánssonはいう。McPhersonは現在、遺伝子発現に影響を及ぼす可能性のあるRNA塩基配列を探しているところだが、その領域は長く、原因となる配列を探し出せたとしても、そのメカニズムまでは明らかにならないかもしれないと、考えている。

たとえ見つかった1つの多型の影響が小さいものであっても、新しいメカニズムを見つけ出す重要性はとても大きいだろうし、全ゲノム相関解析によって、誰もそこにあるとは考えないようなゲノム内の場所に目的の配列が見つかる可能性もある。多くの研究者が最も待ち望んでいるのは、「予期せぬ驚き」である。「この20年間というもの、心血管生物学においては、従来からいわれるリスク因子の研究がすべてだった」と、リューベック大学(独)の心臓専門医Heribert Schunkertはいう。彼はSamaniとともに、9p21.3領域が心疾患と関連することを実証した³。現在見つかっている

多くの遺伝子座は、これらのリスク因子と結びつかない。「わくわくしますね。これはつまり、我々が本当の心血管生物学というものをもっとつかめていないということですから」と彼は語った。 ■

Monya Bakerは、Nature Reports Stem Cellsの編集長。サンフランシスコ発のNatureの記事を執筆している。

1. McPherson, R. et al. *Science* **316**, 1488-1491 (2007).
2. Helgadóttir, A. et al. *Science* **316**, 1491-1493 (2007).
3. Samani, N. J. et al. *N. Engl. J. Med.* **357**, 443-453 (2007).
4. The Wellcome Trust Case Control Consortium *Nature* **447**, 661-678 (2007).
5. Psaty, B. M., Arnett, D. & Burke, G. *J. Am. Med. Assoc.* **298**, 2060-2062 (2007).
6. Willer, C. J. et al. *Nature Genet.* doi:10.1038/ng.76 (2008).
7. Kathiresan, S. et al. *Nature Genet.* doi:10.1038/ng.75 (2008).
8. Helgadóttir, A. et al. *Nature Genet.* doi:10.1038/ng.72 (2008).
9. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT et al. *Science* **316**, 1331-1336 (2007).
10. Krishnamurthy, J. et al. *Nature* **443**, 453-457 (2006).
11. Grarup, N. et al. *Diabetes* **56**, 3105-3111 (2007).
12. Hampe, J. et al. *Nature Genet.* **39**, 207-11 (2007).
13. Rioux, J. D. et al. *Nature Genet.* **39**, 596-604 (2007).
14. Amundadóttir, L. T. et al. *Nature Genet.* **38**, 652-658 (2006).
15. Freedman, M. L. et al. *Proc. Natl Acad. Sci. USA* **103**, 14068-14073 (2006).



知ってることは載っていない

nature

www.naturejpn.com

npg nature asia-pacific