一細胞ゲノム解析

Single-cell genomics of cancer

加藤　護
Mamoru KATO
国立がん研究センター研究所バイオインフォマティクス部門

◎腫瘍内不均質性は，がん細胞がゲノム・エピゲノム変化を経た進化の結果発生し，その程度は抗がん剤耐性や予後のような臨床像とも関係すると考えられている．これら不均質性やがん細胞進化を，その最小単位である単一細胞レベルで調べるため，近年，一細胞シークエンス技術が開発された．この技術はflow cytometryなどの方法で細胞を分離し，次世代シークエンサーを使って細胞ひとつひとつのゲノムを配列決定するという技術である．一細胞シークエンスは技術的にまだ課題が多いが，組織をそのまま読む通常のシークエンスではわからないがん細胞の分集団を発見し，これまで不明であったがん細胞進化の詳細を明らかにしつつある．この技術の臨床応用として血中循環腫瘍細胞の変異変化をタイピングし，患者内のがんの状態をモニタリングする研究もはじまっている．

Key word：腫瘍内不均質性，がん細胞進化，一細胞シークエンス

がん組織に腫瘍内不均質性(intra-tumor heterogeneity)があることは古くから知られていたが，がんゲノム研究がこの不均質性に本格的に取り組みはじめたのはつい最近のことである．次世代シークエンサーと組み合わされた一細胞シークエンス技術は，この不均質性をもっとも明確に解明できる技術として注目されている．

一細胞シークエンス技術は，多細胞生物において生殖細胞(精子，卵細胞)[1,2]や神経細胞[3]にも適用されているが，本稿ではがん細胞への適用を取り上げる．また，広義にはRNAやエピゲノムのシークエンスも含むが，ここではおもにゲノムDNAシークエンスに着目する．

腫瘍内不均質性

採取されたがん組織はけっして均質ではない．さまざまな種類の間質細胞を含むこともあるし，がん細胞間でも異なる病理像を呈し，異なるマーカーを発現することもある．病変組織をマウスに異種移植しても，増殖を開始できる(能力をもった)がん細胞もあれば，そうでない細胞もある．このような腫瘍内不均質性は診断や治療にも影響を与えると考えられる．たとえば，検査時において悪性度の高い細胞や化学療法に耐性がありそうな細胞の小集団が他細胞の大集団のなかに埋没している場合，不均質性を考慮しない診断では精度が落ちるであろう．ほかにも腫瘍の不均質性(多様性)の度合いが高いと腫瘍が悪性化しやすいという報告もある[4]．このように腫瘍内不均質性は生物学的な追求の対象としてだけではなく，臨床医学的にも重要である(本特集・谷内田「腫瘍内多様性とがんゲノム進化」の稿も参照)．

このようながん細胞の不均質性のうち細胞の有糸分裂による継承がかかわる不均質性に関しては，それが生じる理由として，①細胞間の分化階層構造，②がん細胞ゲノムの進化，が考えられている．①はいわゆるがん幹細胞に関連する概念であり，大ざっぱには(ゲノムが同じであったとしても)エピゲノム的な状態の違い，と言い換えることができる．②はがん細胞のクローン性増殖(clonal expansion)[5]のことであり，腫瘍進展においてさまざまな体細胞変異を得たがん細胞が生き

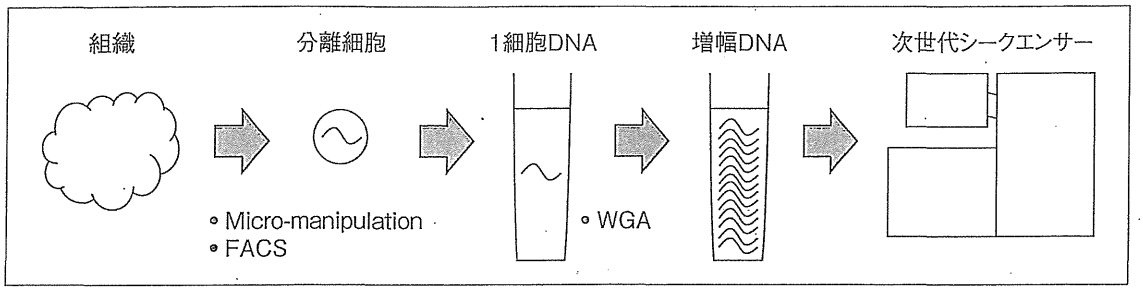

図 1　一細胞シークエンスの原理

残り，多様性(不均質性)が生まれるというダーウィン進化にたとえられる過程である．組織像における複雑な不均質性をこれらの描像によって捨象するとすれば，不均質性は「腫瘍進展に伴うゲノム・エピゲノム状態の時間変化の結果現れる，それら状態の細胞間での違い」ということもできる．

不均質性を解明するシークエンス技術

ゲノム分野で不均質性を研究するおもな手法には2つある．ひとつは組織から細胞塊のままDNAを抽出して次世代シークエンサーでシークエンスし(バルクシークエンス)，リードに含まれる変異文字の割合を数え，それを変異アレル頻度として分析することである．たとえば，組織にがん細胞の分集団Aが60%，がん細胞の分集団Bが30%存在し，Aに特異的な変異アレルaが座位1に，Bに特異的な変異bが座位2に存在するとしよう．その腫瘍組織をシークエンスすると，変異aが座位1に30%(通常ヘテロ接合での変異が期待されるため，60/2)，bが座位2に15%検出されることが期待される．逆にもともとどのような分集団が存在するのかわかっていない場合，これらのデータから分集団が2つ存在し，それぞれ組織中60%，30%を占めていると推測することができる．2つ目の手法は一細胞シークエンスである．ここでは細胞の塊をばらばらにして一細胞にし，一細胞のDNAを増やして(whole genome amplification：WGA)，次世代シークエンサーにかけられるほど十分な量を得，シークエンスして細胞のゲノムをひとつ一つ分析する(図1)．不均質性の最小単位を分析するので，論理は明快である．

これら2つの手法を比較するとさまざまな点で一長一短があるが，一細胞シークエンスでの利点はバルクシークエンスでは座位1と座位2での関連がわからないのに対し，一細胞シークエンスではわかるという点である(図2)．たとえば，バルクでは座位1で20%，座位2で20%の頻度があった場合，通常40%の分集団がひとつ存在すると推測してしまうが，実はこれら座位間には関連がなく，2つの分集団がそれぞれ40%存在しているのかもしれない．実験エラーやコピー数変化によって頻度がずれたり，あるいは分集団間の頻度が近接していたりする場合を考えると，実データでこのような例はけっしてtrivialなケースではないことがわかる．また，この違いは生物学的に異なる解釈を導いてしまう．前者のケースでは単一集団がクローナルに存在しているだけであるが，後者のケースでは2集団が協調的に存在している可能性が出てくる．

また，一細胞シークエンスでは低頻度のコピー数変化(copy number alteration：CNA)をより高感度に検出できるという利点もある．CNA検出は次世代シークエンサーを使っても一般にノイズが多く，バルクシークエンスの場合，CNAが細胞集団内で低頻度であるとノイズに埋もれやすい．

一細胞シークエンスにはこのような利点があるが，しかしコストと(後述するように)テクニカルな問題がある．一方，バルクシークエンスはdepthを深く読めばよいだけであり，条件によってはコストを大幅に下げることもできる．分集団の構造を追及せず，単に点変異の有無を調べるには，バルクのほうが適している可能性もある．

一細胞シークエンス技術

一細胞シークエンスのプロセスは，手順①：組

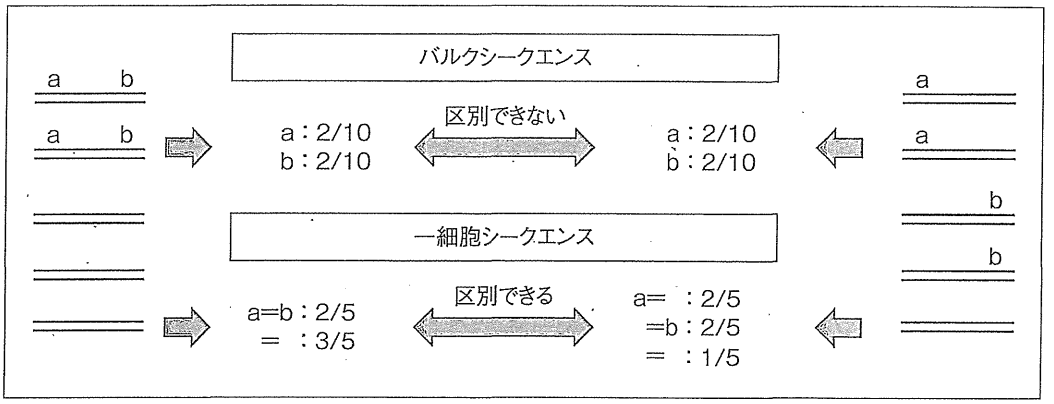

図 2 バルクシークエンスと比較した一細胞シークエンスの利点
バルクシークエンスでの頻度(2/10 など)は DNA コピー数(に比例したリード数)が単位となり，一細胞シークエンスでの頻度は細胞数が単位となる．一細胞シークエンスにおいて“=”は細胞内での連鎖を表す(相，すなわちハプロタイプは決定されない)．

織から細胞を分離し，手順②：一細胞の DNA を増やし，手順③：シークエンスし，手順④：情報処理を行うことである．いずれの工程もバルクほどは確立されておらず，技術的困難が伴う．それは多くの論文[6-8]が，データクオリティの記述に多くのスペースを費やしていることからもうかがえる．

手順①：組織からの細胞分離

手順①では，手動による微小操作(連続希釈とマイクロピペッティング)と fluorescence activated cell sorting(FACS)が使われる[9]．固形腫瘍の場合はその前に酵素によって細胞をばらばらにしなければならない．細胞のサンプリングデザインとしては，ランダムサンプリング[6]とマーカー選別によるターゲットサンプリング(*e. g.*, 腫瘍細胞と正常細胞の分離[10])がある[9]．微小操作は比較的簡単に行えるが，自動化ができず，サンプル汚染にも注意しなければならない[8,9]．FACS はある程度自動化できるが，操作に習熟が必要である．最近では微小流路を利用して細胞分離から核酸増幅まで行う C1(Fluidigm 社)という一括システムが登場している．これまではトランスクリプトームに対応していたが，執筆時点(2014 年 2 月)でゲノムにも対応予定とのことである．自動化されており，閉鎖系のためサンプル汚染の可能性も少ないと考えられる．

手順②：一細胞のDNAを増やす

手順②の WGA には，PCR ベースの WGA[6]，multiple displacement amplification (MDA)[7,10,11]，そして最近開発された MALBAC[8]という手法が使われる．WGA で問題になるのはゲノムが全域にわたって(かつ，均質に)増幅されるか，という問題である．出発の量が一細胞の DNA(倍数体)1 コピーであることから，その難しさが想像できる．CNA を同定する場合は Navin ら[6]が示したように，ゲノムの 6%程度が増幅されていれば解像度 50k bps 程度の CNA が得られるので，(均質に増幅されてさえいれば)増えない領域があっても大きな問題とはならないかもしれない．しかし，重要遺伝子を SNV を通して発見・分析しようとする場合は，増えない領域の SNV を見落とす可能性があるので問題となる．これは allelic dropout(ADO)と呼ばれ，Wang らのグループ[7,10,11]や Zong らはこの率を，(おもに正常細胞の)同じサンプルに対し，バルクでヘテロであるにもかかわらず，シングルでホモとなった座位の数によって近似的に評価している．それによれば，ADO 率は(エキソームシークエンスにおいて)40～43%[7,10]である．Zong ら[8]によれば，MDA は 65%，彼らが開発した MALBAC では 1%とのことである．

手順③：シークエンス

手順③は通常の方法とととくに変わりはない．

手順④：情報処理

手順④では，Narvin ら[6]はマッパーに Bowtie を使い，CNA 検出に独自の Varbin[12]というアルゴリズムを使っている．CNA 検出の原理は depth 分析である(増幅領域には多くのリードがマップ

表 1　一細胞シークエンスの代表的研究

著者	出版年	腫瘍	細胞数*1	細胞の分離	WGA 法	対象変異	対象範囲	カバレジ
Navin ら	2011	乳がん	200	FACS	PCR-based	CNA	WG	6%(<1x)
Hou ら	2012	血液腫瘍	90→58	Micro-manipulation	MDA	SNV	WE	≧5x が≧70%
Hu ら	2012	腎がん	25→22	Hou らに準じる	Hou らに準じる	SNV	WE	≧5x が≧80%
Li ら	2012	膀胱がん	66→55	Hou らに準じる	Hou らに準じる	SNV	WE	≧5x が≧60%
Zong ら	2012	(大腸がん細胞株*2)	13→5	Micro-manipulation	MALBAC	SNV, CNA	WG	Mean 25x

*1：シークエンス成功の歩留まりが報告あるいは推察されるものには矢印を付けた．Zong ら以外では最終的にデータとして使われた細胞数，Zong らの場合，平均 depth 数が約 25x 以上のものを選んだ．
*2：技術開発研究のため，(　)でくくった．
WG：whole genome, WE：whole exome.

される．欠失領域には少ないリード)．整数コピー数の同定も行っており，ガウシアン・カーネル密度推定法を使って depth を分析している．SNV に関して Wang のグループ[7,10,11]では，彼らが開発している SOAP を中心にして分析を行っている．Zong らも含め，通常のコールと違う点は複数細胞(Li ら[10]，Zong ら[8]では 3 細胞)で同じ変異が検出されることをコールの条件としていることである．SNV のエラー率(誤検出率)は ADO 率とは逆に，バルクでホモなのにシングルではヘテロ，という原理で近似的に評価している．それによれば，10^{-4}〜10^{-5}程度[7,10,11]，MALBAC では 0(計算母数以下)で[8]，きわめて低い．変異頻度の計算に関し，Hou らと Li らの研究においてはコールされた変異の頻度を細胞数から直接計算するのではなく，細胞ごとのリード数を考慮するベイズ推定によって計算している(手法は文献[13]参照)．これは各細胞において，depth が十分に取れない状況に対処するためである．変異が検出された後の分析においては，標準的な分析法(系統樹分析，主成分分析，変異頻度分析)以外に現在のところとくに深い分析手法はみあたらないが，バルクと違っていくつか注意しなければならない点がある．ひとつは細胞間に系統関係があるため，細胞間で変異が関連していることである．つまり，バルクにおいて 100 症例で頻度 40%であることと，シングルにおいて 100 細胞で頻度 40%であることとでは意味が違う．前者は変異イベントが 40 回起こったことを表すが，後者は共通祖先に変異イベントがたった 1 回起こって子孫細胞に受け継がれたことを表す．このため種内変異を扱う集団遺伝学のアイデアを取り入れた分析手法や解釈を考えに入れなければならない．また，ADO による避けがたい false negative にも注意しなければならない．

一細胞シークエンスの代表的研究

本格的な一細胞シークエンスがはじまったのは，2011 年の Navin らの研究からである[6](表 1)．彼らは FACS で細胞を分離し，PCR ベースの WGA 法で DNA を増幅し，Illumina 社のシークエンサーで，乳がん臨床サンプル 2 症例から 200 細胞をシークエンスした．と同時に，2 つの細胞株から 14 細胞をシークエンスし，バルクシークエンスと比べたデータ品質も調べている．彼らが注目したのは CNA で，このためとくに深い depth は必要がなかった．シークエンスカバレジは 6%程度であるが，全ゲノム領域が対象である．研究デザインでひとつ注目すべきは，細胞がランダムサンプリングされていることである．よって一見腫瘍と無関係かもしれない細胞が，場合によっては 60%もシークエンスされている．しかしこのため，組織学的に推定された腫瘍細胞率との比較が可能になったり，間質細胞に埋もれた pseudo-diploid という新しい細胞分類が発見されたりしている．また彼らは，細胞集団内では均質であるが集団間としては異なるというがんの不均質性を発見した．このことから，がん細胞は漸次的な進化をするのではなく，古生物学でいう断続平衡進化(まったく変化のない期間と突発的に変化する期間を交互に繰り返す進化)をすることを提唱した．

続く 2012 年には Hou らが，1 症例 90 細胞の一

細胞エキソームシークエンシングを行っている[7]．対象は細胞分離が容易な血液腫瘍である．この研究では正常細胞株からの2細胞に対しwhole genomeの一細胞シークエンスも行われ，技術的側面とデータ品質についても多く記述されている．Depthの低さ(～WGA非増幅)とGC-contentとの間には相関があったが，リピートや染色体位置との相関はなかったようである．90細胞をシークエンスしたが，データ品質チェックによって最終的に使われた細胞数は58で，腫瘍50細胞中712 SNVが発見された．変異の集団頻度に関し，細胞数から計算された変異頻度とバルクでリード数から計算された変異頻度はよい相関を示した(R^2=0.75)．がん細胞進化過程での自然選択の痕跡，および分集団はとくに発見されなかった．

Xuら[11]はHouらの方法を踏襲し，腎がん1症例17細胞中260個のSNVを発見したが，分集団は発見されなかった．

Liら[10]は(改良された)Houらの方法を用いて，膀胱がんの1症例66細胞をシークエンスし，さらに99症例をバルクでシークエンスして，それぞれで検出された変異を比較した．データ品質チェックを通った44細胞に3つの分集団と443個の変異を検出し，*NIPBL*や*CFTR*といった新規ドライバー遺伝子がその3集団共通にみられることを発見した．さらに，変異頻度の分析によってnon-synonymous変異がsynonymous変異に比べいくぶん高頻度にずれていることが示された．一細胞分析の場合，進化的に中立であっても遺伝的連鎖(有糸分裂による継承の際の染色体連鎖)があるためnon-synonymousのほうが高い理由はなく，これを説明するためがん細胞には漸次的な自然選択がかかっていることを提唱した(一見するとNavinらの説と相反するが，とくに詳しくは述べられていない)．

おわりに

一細胞シークエンスはバルクシークエンスに比べまだ技術的に確立されておらず，コストもかかる．しかし，腫瘍内不均質性の最小単位を分析するため，不均質性とそれに付随するがん細胞進化にもっとも明確なデータを与える．集団“内”では均質であるが，集団“間”では異質である分集団の発見や，断続平衡進化・漸次進化の提唱にみられるように，腫瘍内不均質性や腫瘍進展の解明を強力に促進する．これまでの研究は一時刻点のスナップショットデータであったが，今後多時刻点でのデータが得られれば，その最小単位の明快なダイナミクスが明らかにされるであろう．また，それに従ってより強力な情報解析手法も開発されねばならない．

著者は，国立がんセンターの柴田龍弘博士，新井康仁博士，筆宝義隆博士らの協力を得，多時刻点の一細胞シークエンスや，集団遺伝学の考えを取り入れた新しい分析手法の研究を行っている．

一細胞シークエンス技術の臨床応用としてはcirculated tumor cell(CTC)への応用があり，CTC検出システムで血中から分離された腫瘍細胞をカウントするだけでなく，次世代シークエンサーを用いて細胞の変異検出を行ってモニタリングや治療効果因子の同定に役立てようとする研究が行われている[14,15]．アメリカNIHは一細胞研究の重要性を認識し，一細胞シークエンスを含む単一細胞研究に5年間で約90億円(US$ 90 million)の助成を2012年に公表しており[16]，Single Cell Analysis Programとして助成が開始されている．

文献

1) Hou, Y. et al.：*Cell*, **155**：1492-1506, 2013.
2) Lu, S. et al.：*Science*, **338**：1627-1630, 2012.
3) Evrony, G. D. et al.：*Cell*, **151**：483-496, 2012.
4) Maley, C. C. et al.：*Nat. Genet.*, **38**：468-473, 2006.
5) Nowell, P. C.：*Science*, **194**：23-28, 1976.
6) Navin, N. et al.：*Nature*, **472**：90-94, 2011.
7) Hou, Y. et al.：*Cell*, **148**：873-885, 2012.
8) Zong, C. et al.：*Science*, **338**：1622-1626, 2012.
9) Shapiro, E. et al.：*Nat. Rev. Genet.*, **14**：618-630, 2013.
10) Li, Y. et al.：*Gigascience*, **1**：12, 2012.
11) Xu, X. et al.：*Cell*, **148**：886-895, 2012.
12) Baslan, T. et al.：*Nat. Protocols*, **7**：1024-1041, 2012.
13) Yi, X. et al.：*Science*, **329**：75-78, 2010.
14) Ni, X. et al.：*Proc. Natl. Acad. Sci. USA*, **110**：21083-21088, 2013.
15) Heitzer, E. et al.：*Cancer Res.*, **73**：2965-2975, 2013.
16) Owens, B.：*Nature*, **491**：27-29, 2012.

6. 発がんドライバー変異の同定

David Tamborero, Abel Gonzalez-Perez, Nuria Lopez-Bigas

シークエンシング技術の爆発的進展によって腫瘍ゲノムの特徴がわかり，がんという疾病は臨床でみられるように分子レベルにおいても均質でないことが明らかになってきた．がんの異なるステージで，腫瘍細胞が獲得する表現型を担うさまざまな変異を同定することは，がん化の過程を理解し，新しい治療介入法を開発するために必要なさらなる解析の基礎となる．本稿では，現在の再シークエンシング計画によって生み出される大規模データの分析や，腫瘍形成を促進する変異・遺伝子・パスウェイの同定に役立つ計算機的手法について論ずる．

はじめに

次世代シークエンシング技術によって，腫瘍細胞ゲノムにおける体細胞変化の完全解明への扉が開かれた[1]．がん研究は変容しつつある．たとえ腫瘍の種類が同じでも，がんのサンプルは体細胞異常に関し，きわめて不均質であることもわかってきた[2]．細胞が経験する複製の回数やDNA維持装置の欠陥，環境からの攻撃といった要因が，体細胞変異の変異率とパターンを規定するが[3]，がん細胞に見出されるほんの一握りの体細胞変異だけが腫瘍形成の原因となる．残りのものは，がんが引き起こすゲノム不安定性の二次的偶発事象にすぎない[4]．前者と後者（すなわちドライバー変異とパッセンジャー変異）を区別することは，がん生物学研究および新規治療介入法の開発における最重要課題の1つである．

腫瘍ゲノム再シークエンシングが生み出す大量データの解釈は，近年大きな進歩があったとはいえ，いまだ多くの点で大変な仕事である．本稿では，がんドライバー変異の同定を行う最新の計算機的方法を振り返り，その主要な結果を論ずる．ただし本稿では，体細胞の単一塩基変異および短いフレームシフト変異を対象とする．また，エクソームにおける変異を評価する方法にのみ焦点を当てる．非コード領域（non-coding region）における変異の役割は依然として明らかではないが，近い将来には全ゲノムシークエンスのコストが下がり，その分野で飛躍的進展があろうことは十分に予見できる．

[キーワード&略語]
発がんドライバー，体細胞変異，計算機的方法

ICGC：The International Cancer Genome Consortium（国際がんゲノムコンソーシアム）
TCGA：The Cancer Genome Atlas（がんゲノムアトラス）

1 がんに対する変異の影響を同定する

再シークエンシングによって，腫瘍ゲノムにおける

Identification of oncogenic driver mutations
David Tamborero[1]/Abel Gonzalez-Perez[1]/Nuria Lopez-Bigas[1][2]：Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra[1]/Institució Catalana de Recerca i Estudis Avançats（ICREA）[2]（ポンペウ・ファブラ大学実験健康科学部生物医学情報研究ユニット[1]/カタラーナ高等研究所[2]）

体細胞変異が検出される．その役割を調べる最初のステップは，ゲノム要素という文脈においてそれらを意味づけることである．すなわち，コーディング配列の場合，突然変異をもつタンパク質コード遺伝子を同定し，その影響を調べる．終止変異，フレームシフト変異といったタンパク質産物を切り詰める変異はおそらくタンパク質の不活化を引き起こすだろうし，一方，同義変異はタンパク質機能に対してはるかに穏やかな効果をおよぼす．これら両極端の間にある非同義変異がタンパク質機能に与える影響を調べることが，計算機的手法の主題である（表）[5]．

当初，計算機ツール群は生殖細胞系列変異のために設計されていた．例えば，SIFT[6]，PolyPhen-2[7]，Mutation Assesor[8]は，アミノ酸進化保存の異なる指標を使ってタンパク質傷害の程度を推測した．Condelはこれらのツールを凌駕すべく，それらツールの出力を総合して，独自の標準化スコアの重み付き平均にもとづき，コンセンサス有害スコアを算出した[9]．

最近開発されたバイオインフォマティクスツールでは，腫瘍の体細胞変異を扱うことができる．fathmmは，がんデータでトレーニングされた隠れマルコフモデルの枠組みで配列保存データを用い，ドライバー変異を識別する[10]．CHASMは配列保存に加え，特定のアミノ酸置換やタンパク質構造変化予測に関連した特徴変数に対し，既知のドライバー変異と人工的につくったパッセンジャー変異でトレーニングされたRandom Forestアルゴリズムを使用して，識別を行う[11]．TransFICはこれらツールのスコアを，遺伝子グループ間での機能変異に対する許容性の違いを考慮に入れて改善し，使用する[12]．なお，スプライス部位変異といった特殊な種類の変異の場合には，他の特別なツールが必要となる[5]．

2 正の選択シグナルを検出する

細胞は増殖するうえでの利点をドライバー変異から受け取り，その結果ドライバー変異は腫瘍のクローン進化過程でより広まっていく．一方，パッセンジャー変異はドライバー変異のそばにいるというだけで増えていく．それゆえ腫瘍形成に関与する遺伝子は，腫瘍サンプルを見渡したとき，正の自然選択と同じシグナルを出しているはずである．正の選択のシグナルをとらえるいくつかの方法が，ドライバー遺伝子候補の同定に利用されている（図1，表）．

1）変異の頻度による方法

最も直感的な方法は，期待されるバックグラウンド変異率よりも遺伝子が頻繁に変異しているかを評価することである（図1A）．このアイデアは，例えばワシントン大学によって開発されたMuSiC[13]や，Broad研究所によって開発されたMutSig[14]といったツールに実装されている．バックグラウンド変異率の推定には，ゲノム上特定の位置に起こる体細胞変異の発生確率に影響を与えるゲノムの特徴が考慮される．例えば，遺伝子の長さ，変異の種類，周りの核酸パターンなどである．MutSig（後にMutSigCVと改名）の最新の実装においては，DNA領域の複製タイミングや遺伝子発現レベルといった因子が，変異率に影響を与える共変量として統計的枠組みに組み込まれ，推定精度を増すために使われている[14]．これらの方法によって確かに，がんで高頻度に変異している遺伝子の検出には成功している．しかし変異が低頻度のドライバーは，ほとんど検出できていない．腫瘍形成の全体像を理解するには，低頻度ドライバーも検出しなければならないが，そのためにはより正確なバックグラウンド変異率の推定と，より大きなサンプルサイズが必要であろう．

2）変異の機能的影響による方法

別の，遺伝子の変異負荷に頼らないアプローチは，サンプル集団において各遺伝子の変異が機能に与える影響を評価する方法である（図1B）．ドライバー変異は，コーディング遺伝子においてタンパク質の機能に影響を与えるはずである．それとは反対に，パッセンジャー変異はランダムに分布するはずである．したがって，機能に大きな影響を与える変異を多くもっていれば，それは正の選択をさし示し，これをもってドライバー遺伝子と判断できる．この方法はバックグラウンド変異率の推定に依存しないため，変異の集団頻度と関係しないドライバー遺伝子の検出に向く．OncodriveFM[15]はこの方法を実装している．すなわち，OncodriveFMは遺伝子ごとに各変異の機能的影響を示すいくつかの指標を用い，ある基準値からのずれを計算する．

3）変異の集中による方法

正の選択シグナルの第三は，タンパク質の一次構造の特定の箇所に変異がまとまっていることである（図1C）．

表　発がんドライバー遺伝子を同定するためのツール一覧

	方法	説明
単一の非同義変異が機能におよぼす影響を推定する	SIFT[6]	ユーザーの定義するデータベースによって類似タンパク質のマルチプル配列アライメントをつくり，アライメントのすべての位置でのすべての可能な置換に関する標準化確率を計算する．この確率にもとづいて，観察された置換が中立か有害かを分類する．
	PolyPhen-2[7]	有害・中立アミノ酸変化からなる2つのデータセットによってトレーニングされたナイーブベイズ分類器．主に野生型と変異型アミノ酸の性質の比較に関連した8個の配列ベース，3つの構造ベースの特徴変数が，分類器の作製に使われる．
	Mutation Assessor[8]	アミノ酸残基の進化的保存を調べることで，非同義SNVの機能的影響を予測する．機能全体の保存性をバックグラウンドにして相同配列のマルチプルアライメントをクラスタリングし，タンパク質サブファミリーを決定して，その進化的保存度を利用する．
	Condel[9]	Condel（コンセンサス有害スコア）は，非同義SNVによる機能的影響スコアを組合わせる方法である．有害および中立非同義SNVのデータセットについて各種ツールが算出するスコアの相補的累積分布から得られた値を，各種ツールの結果を組合わせるための重みとして使う．
	fathmm[10]	がんの“病原性的重み”を変異のモデルにおける許容度で表し，重みをもつ保存タンパク質ドメインと相同配列のアライメントを隠れマルコフモデルで表現する．配列保存性を隠れマルコフモデルと組合わせ，がん体細胞変異が機能におよぼす効果を予測する．
	CHASM[11]	専門家がまとめたCOSMIC由来のドライバー変異と，ランダムシミュレートで得られたパッセンジャー変異にもとづいてトレーニングされたRandom Forest分類器．アミノ酸残基の生理化学的性質や，タンパク質・DNAのマルチプルアライメント由来のスコア，領域ベースのアミノ酸配列組成，UniProtKB特徴表からのタンパク質局所的構造の性質予測および注釈など，86個の多様な特徴変数（SNVBoxデータベースで入手可能）を使う．
	TransFIC[12]	TransFIC（がんの機能的影響変換スコア）は，変異がタンパク質機能に与える影響を評価する他の方法が出す機能的影響スコアを，類似タンパク質における機能的影響変異の許容度を考慮に入れて変換する．変換によって，体細胞変異が細胞に与える影響を補正したと解釈する．
腫瘍集団におけるドライバー遺伝子を同定する	MuSiC[13]	シークエンスカバレッジや遺伝子の長さ，核酸変化の種類といった各観察変異の基準確率を規定する複数の特徴を考慮に入れて，期待されるより多くの変異をもつ遺伝子を同定する．
	MutSigCV[14]	高頻度に変異する遺伝子を同定する．各遺伝子で期待される変異荷重に関連した追加データも取り入れて，バックグラウンドモデルを作製する．その共変量の例としては，がん細胞株で観察されたRNAシークエンスから集めた遺伝子発現データ，HeLa細胞株で測定されたDNA 複製のタイミングがある．
	OncodriveFM[15]	機能的影響をいくつかの指標値の組で計り，大きな影響をもつ変異に富む遺伝子を同定する．
	OncodriveCLUST[16]	サイレント変異を使ってつくられた基準モデルから期待される以上にクラスター化する変異をもつ遺伝子を同定する．
	Active Driver[17]	リン酸化部位に集積する傾向のある変異をもつ遺伝子を同定する．
腫瘍集団におけるドライバー遺伝子のモジュールを同定する	MEMo[18]	与えられたパスウェイがサンプル間で相互排他的な変異を提示するかを調べ，遺伝子を連結してモジュールを見出す．
	HotNet[19]	変異頻度や機能的影響スコアといった指標値が集中する遺伝子モジュールを，与えられたパスウェイ上で連結された遺伝子を通って伝搬する熱拡散モデルによってみつけ出す．

SNV：一塩基変異

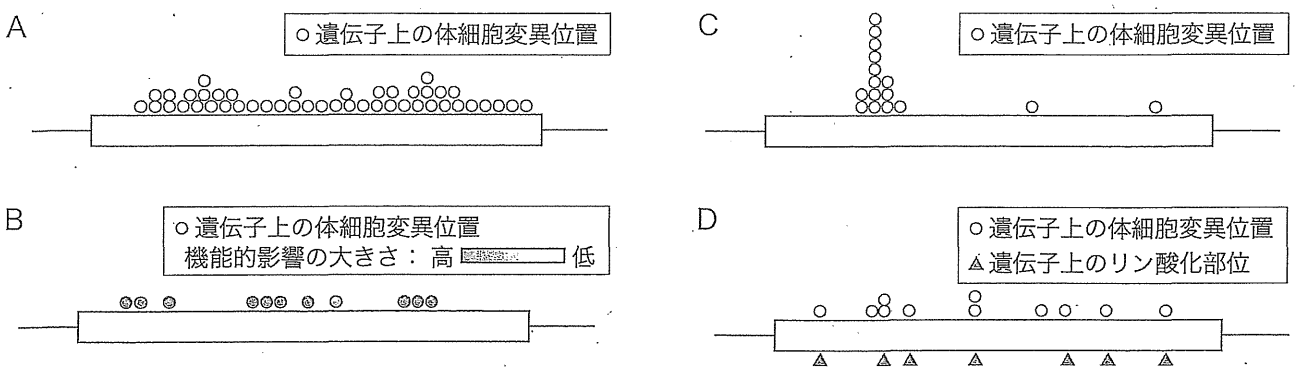

図1　がんサンプル集団における，異なる正の選択シグナルを示す遺伝子の例
各点（○）は，集団中異なるサンプルに対する体細胞変異の遺伝子上の位置を示す．A）偶然から期待されるより多くの変異が蓄積する．そのような遺伝子はその腫瘍集団に対し高頻度で変異する遺伝子として同定される．この方法においては，バックグラウンドの変異確率に影響を与えることがわかっている複数の因子を考慮に入れる必要がある．B）この遺伝子に観察される変異は強い機能的影響を引き起こすものに偏っており，このタイプの変異が選択されたこと，したがってドライバーであることを示唆している．機能的影響をスコア化するため使われる指標自体の能力が，このアプローチの性能を規定する．例えば，PIK3CA残基1,047を変える変異はがん原的であることが知られているが，保存にもとづく機能的影響の指標では過小評価になる[16]．なぜならこの残基は種にわたってかなり変化に富むからである．C）本文で述べているように，この遺伝子で観察された変異は特定の領域にきわめて集中している．BRAF 残基600 のがん原変異のような機能獲得型変異は，この方法で高感度に検出される[16]．一方，機能欠失を起こす切り詰め型変異は，遺伝子上で散らばってしまう．D）本文で述べているように，このケースではタンパク質のリン酸化部位に変異が偏って発生する

がんに狙われた変異は，タンパク質のある箇所に集積する．OncodriveCLUST[16] はこの考えを取り入れ，変異の確率がタンパク質配列にわたって均質かを調べる．この方法では，サイレント変異を腫瘍変異の凝集度の基準線としている．この基準線を越えて非同義変異を集積している遺伝子が検出される．

4）その他の方法

他にも正の選択シグナルを検出する方法が開発されている．例えばActiveDriver は，リン酸化部位に偏って起こる体細胞変異をもつ遺伝子を検出し，それによってリン酸化ネットワークを乱すドライバーイベントを浮き彫りにする（**図1D**）[17]．他に，遺伝子モジュールにおける変異を調べる方法もある．MEMo はパスウェイデータを使って，サンプル間で相互排他的なパターンにしたがう変異をもつ遺伝子の，連結クリークを検出する．この背後にある考え方は，すでに変異したパスウェイ上の別の遺伝子に変異が起きると，腫瘍細胞にさらなる選択上の利益を与えることなく，むしろ合成致死を引き起こしてしまうことである[18]．他に，HotNet アルゴリズムは遺伝子相互作用マップ上で熱拡散モデルを使って，変異頻度のような指標値に富む遺伝子を連結しながら，かき集めてくる[19]．

5）ベストな方法

いうまでもないことだが，手法の性能を規定するのは，つくり上げた統計的枠組みのなかで潜在的因子の複雑さがどれだけ考慮できているか，さらにはそこでの基準自体がどこまで適用できるか，である．例えば変異頻度にもとづく方法は，低頻度なドライバー変異を見過ごす傾向がある．機能への影響を調べる方法は，機能喪失イベントにはより明確な結果を出すだろう．凝集している変異を同定する方法は，がん遺伝子を同定するにはよい．ベストな考えは，いくつかの方法を組合わせることだろう．それによって，各方法の利点と欠点のバランスをとり，完全かつ確実なドライバー遺伝子のリストを得ることができる．われわれは最近このアイデアにもとづいて12の異なるがん腫の3,205サンプルを解析し，その結果291個の変異ドライバー候補を得た[20]．つまり，正の選択シグナルを相互補完的に調べる方法を組合わせることで，真のがん遺伝子の抽出を改善できることを実証した．

最後になるが，がんドライバーとして同定された遺伝子に起こる変異のすべてが，腫瘍形成に関与しているわけではないことを強調しておく．確かにドライバー遺伝子は腫瘍表現型につながる潜在能力をもっており，ド

ライバー変異を蓄積しうる．しかしそれらはまた，パッセンジャー変異も蓄積しうる．個々の腫瘍の変異を評価するとき，このことは心に留めておくべきである．

③ 発がんドライバーの完全なカタログ化に向けて

1）ドライバー変異のないサンプル

もちろんすべての腫瘍形成がドライバー遺伝子の変異で説明できるわけではないが，常識的な予想では，がんはその進行過程で獲得される変異によって引き起こされるはずである．しかし変異の少ないサンプルのなかには，ドライバー遺伝子にほとんど，あるいは，全く変異がないサンプルもある．この理由としては以下があげられる．まずは，変異性ドライバー遺伝子を，すべては検出できていないことである．次は，発がんイベントが非コード領域に起きていることである．3番目は，腫瘍形成が変異とは別のメカニズム（例えば，転座，コピー数変化，高メチル化）によって引き起こされているということである．とはいえ，他のメカニズムで腫瘍形成を促す遺伝子はたいてい点変異上のドライバーでもあろうから，多くの腫瘍サンプルの変異を分析すれば，がんドライバーの包括的なカタログは作製できるはずである．例えば，エピジェネティクスによるサイレンシングや遺伝子欠失の標的となるドライバーはまた，遺伝子切り詰め型の変異でもおそらく標的のはずである．コピー数増幅や遺伝子融合によるドライバーはまた，活性化変異を介したドライバーとして機能しているはずであろう[21]．

2）低頻度のドライバー

最近の研究において，がんゲノムアトラス（The Cancer Genome Atlas：TCGA）や国際がんゲノムコンソーシアム（The International Cancer Genome Consortium：ICGC）などが提供する大規模データセットから得られたドライバー遺伝子のカタログが報告されている[2) 20) 22) 23)]．それらが示しているのは，高頻度で変異する遺伝子というものがほとんど存在しないということである[2)]．むしろ，頻度の分布を描いたとき低頻度側に長い尾を引く（したがって検出の信頼性もまた低い），低リカレントなドライバーによって，がんの全体像は支配されているということである[2)]．いまだ体系的に研究されたことのないまれながんに関するものを除いて，比較的高頻度で変異する遺伝子は発見し尽くされたといえる[23)]．対照的に，低リカレントなドライバーのカタログは，調べる症例の数が増えるほど，長くなっている．一方で，依然としてなぜ，ある種の遺伝子は他よりはるか頻繁にがんの標的となるのか，その理由はわかっていない．しかしこの疑問は，腫瘍の形成過程や腫瘍細胞が依存するメカニズムを深く理解するうえできわめて重要である．

④ IntOGen-Mutations

がんゲノム研究における大きな障壁の1つは，がん再シークエンシング計画によって得られる，長大な変異のカタログを容易に分析できるバイオインフォマティクスパイプラインが不足していることである．この課題に答えるために，われわれはIntOGen-Mutations（http://www.intogen.org/web/mutations/v04/search?1）をつくった（**図2**）．これはウェブベースのツールであり，腫瘍サンプル集団のデータから発がんドライバーを同定することを目的としている．さらに，大規模な国際コンソーシアムから個々の研究室まで，それらが提供する可能な限りのがん研究データが体系的に分析されており，その結果をブラウズすることもできる．

IntOGen-Mutationsの最初のリリースでは，31の異なる大規模がんプロジェクトから，4,600以上の腫瘍サンプルのデータを分析した[24)]．例えば各腫瘍集団の正の選択シグナルを調べ，遺伝子・パスウェイ・組織レベルで結果を得た．加えて，興味を引く外部データベースへのリンクも提供した．

このツールは，新しいがんゲノム再シークエンシングデータを使って，定期的にアップデートされている．次回のリリースでは，7,000サンプル以上の分析を網羅する予定である．すでに分析されたデータセットの結果をブラウズすることに加え，ユーザーはこのツールで自身がもつ腫瘍サンプル集団や単一個体の変異を分析することもできる．分析パイプラインはわれわれのサーバー上オンラインで動かすこともできるし，ユーザーのコンピュータ上でローカルに動かすこともできる．将来のバージョンでは，治療方針決定に役立つ分子標的薬の情報も含める予定である．

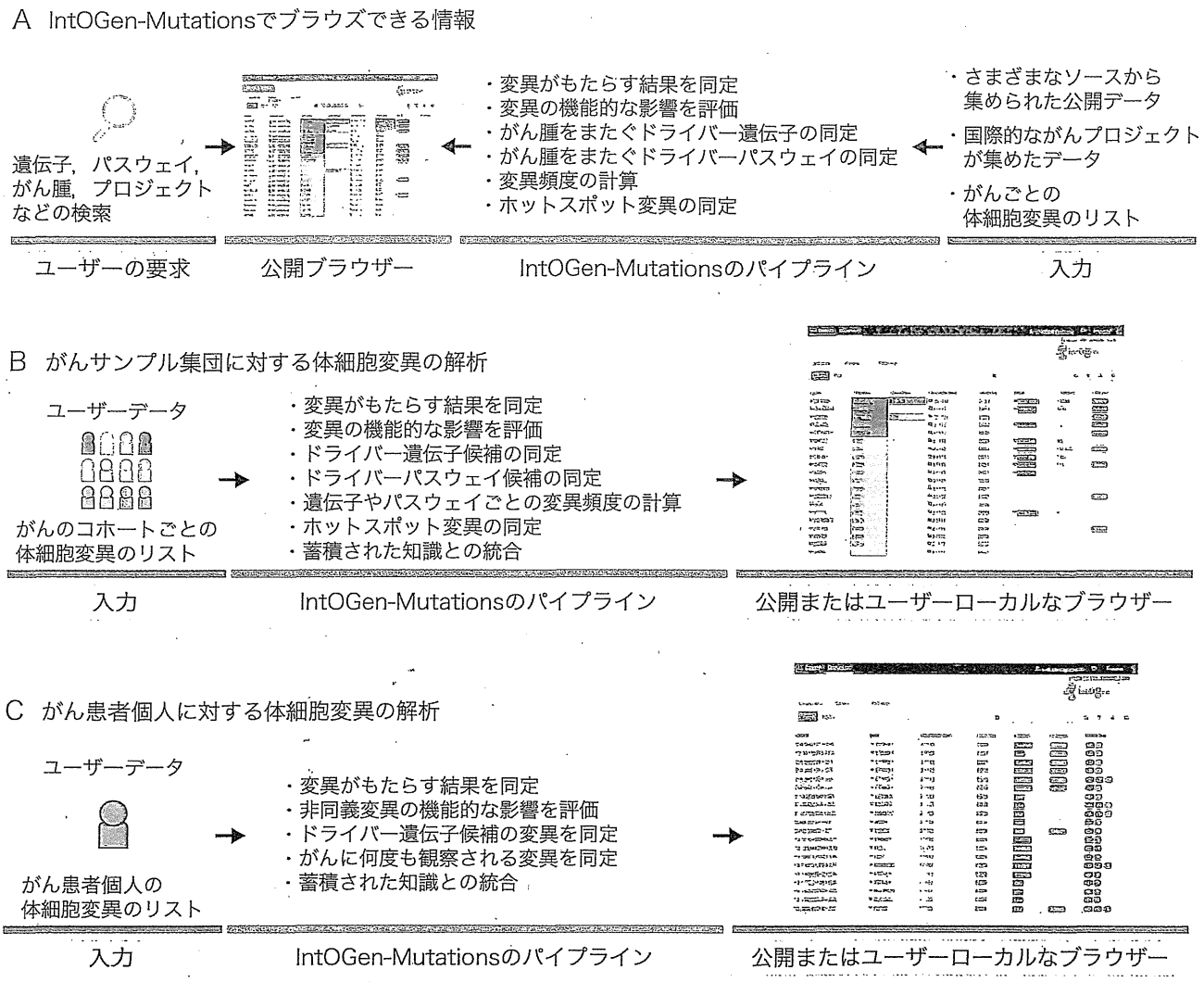

図2　IntOGen-Mutationsの使用例

A）再シークエンシング計画によって提供される，可能な限りのデータが体系的に分析されており，その結果をブラウズすることができる．B）腫瘍集団における新しい体細胞変異の分析．C）腫瘍一症例に対する体細胞変異の分析（文献25より転載）

おわりに

がんは無数の体細胞変異がその性質を決定する不均質な病気であり，がんを理解するためにはドライバー変異をパッセンジャー変異から分離しなければならない．次世代シークエンシング技術によって，大規模な腫瘍集団のシークエンス分析が可能となり，それによって正の選択シグナルを介してドライバー遺伝子が検出されてきた．各々の方法には，結果を解釈する際注意すべき前提がある．ベストなアプローチは，相互補完的な方法の結果を組合わせて，各方法の利点と欠点のバランスをとることであろう[20]．この戦略を使って最近，既知がん遺伝子の役割が確かめられ，その知見を他の腫瘍へ拡張したり，腫瘍進化に関連する新しい遺伝子候補や生物学的過程を発見したりすることが可能となった．この流れのなかで，主要ながんの標的となるありふれたドライバー遺伝子の探索はほぼ終了したといえる．一方，頻度分布上，長い尾を引く低頻度ドライバーは，より大規模なデータセットを分析することでさらに発見されていくであろう．

ドライバー遺伝子の包括的なカタログを作製することは，腫瘍形成を深く理解し，患者ごとに最適な新しい治療戦略を開発するためのさらなる解析への最初のステップとなる．ドライバーイベントが，時間と空間

でどう作動するかを理解する必要もあるだろう．その理解のうえに，腫瘍細胞が特異的にもつ弱点を同定できれば，腫瘍細胞のクローナルな性質と，それが正常細胞と行う相互作用をも考慮した選択的治療法が設計できるようになる．究極的には実験による検証が必要であろうが，これらの結果は，分子標的薬による個別化医療の実施，免疫療法の使用，よりよい早期発見法の開発といった，より合理的で効果的ながんの病勢管理を可能とする次世代治療戦略を生み出す重要なステップとなるだろう．

（翻訳：加藤　護）

文献

1) Stratton MR：Science, 331：1553–1558, 2011
2) Vogelstein B, et al：Science, 339：1546–1558, 2013
3) Alexandrov LB, et al：Nature, 500：415–421, 2013
4) Stratton MR, et al：Nature, 458：719–724, 2009
5) Gonzalez-Perez A, et al：Nat Methods, 10：723–729, 2013
6) Ng PC & Henikoff S：Nucleic Acids Res, 31：3812–3814, 2003
7) Adzhubei IA, et al：Nat Methods, 7：248–249, 2010
8) Reva B, et al：Nucleic Acids Res, 39：e118, 2011
9) González-Pérez A & López-Bigas N：Am J Hum Genet, 88：440–449, 2011
10) Shihab HA, et al：Bioinformatics, 29：1504–1510, 2013
11) Carter H, et al：Cancer Res, 69：6660–6667, 2009
12) Gonzalez-Perez A, et al：Genome Med, 4：89, 2012
13) Dees ND, et al：Genome Res, 22：1589–1598, 2012
14) Lawrence MS, et al：Nature, 499：214–218, 2013
15) Gonzalez-Perez A & Lopez-Bigas N：Nucleic Acids Res, 40：e169, 2012
16) Tamborero D, et al：Bioinformatics, 29：2238–2244, 2013
17) Reimand J & Bader GD：Mol Syst Biol, 9：637, 2013
18) Ciriello G, et al：Genome Res, 22：398–406, 2012
19) Vandin F, et al：Genome Res, 22：375–385, 2012
20) Tamborero D, et al：Sci Rep, 3：2650, 2013
21) Tamborero D, et al：PLoS One, 8：e55489, 2013
22) Kandoth C, et al：Nature, 502：333–339, 2013
23) Lawrence MS, et al：Nature, 505：495–501, 2014
24) Gonzalez-Perez A, et al：Nat Methods, 10：1081–1082, 2013

＜著者プロフィール＞

Nuria Lopez-Bigas：スペイン・バルセロナのポンペウ・ファブラ大学（Universitat Pompeu Fabra）・生物医学ゲノムグループ（Biomedical Genomics group）リーダー．2002年，バルセロナがん研究所（the Oncologic Research Institute）にて聴力障害の分子的基礎の研究で博士号を取得．イギリスのケンブリッジのEuropean Bioinformatics Institute に移り，コンピュータによる疾病およびがん遺伝子の研究プロジェクトに参加．'06年より，ポンペウ・ファブラ大学でグループリーダーとなり，がんのゲノミクスとバイオインフォマティクスを研究．'11年より，ICREA（Institució Catalana de Recerca i Estudis Avançats）Research Professor.