



Human Genome Center
Institute of Medical Science, University of Tokyo



スパコンが加速するがん研究

宮野 悟

東京大学医科学研究所ヒトゲノム解析センター

miyano@ims.u-tokyo.ac.jp

未来をひらくスーパーコンピュータ
～「京」からその先へ 限りなき挑戦～

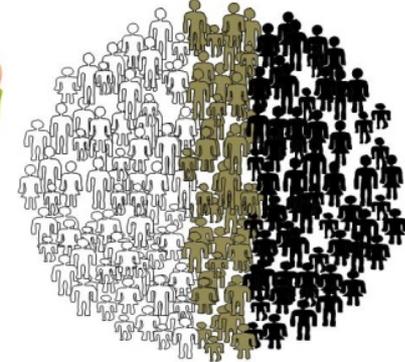
科学技術館

2014年8月23日

「世界に一つだけの花」
私たちは一人一人異なる
ゲノムを持っている

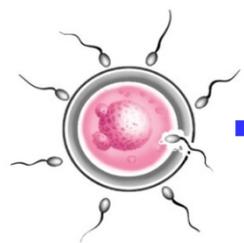


「がんは日本の国民病」日本人
の半分が罹り、3分の1が亡く
なっている



「私のDNA」と「私」

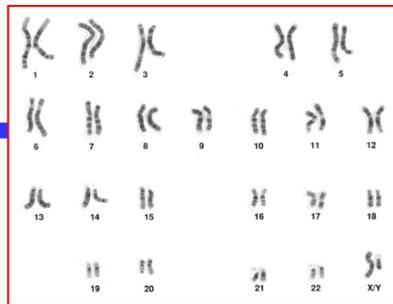
母のDNAと卵細胞



1954年

父のDNA

23組の染色体
30億×2文字の
ATCGのDNA情報



環境因子

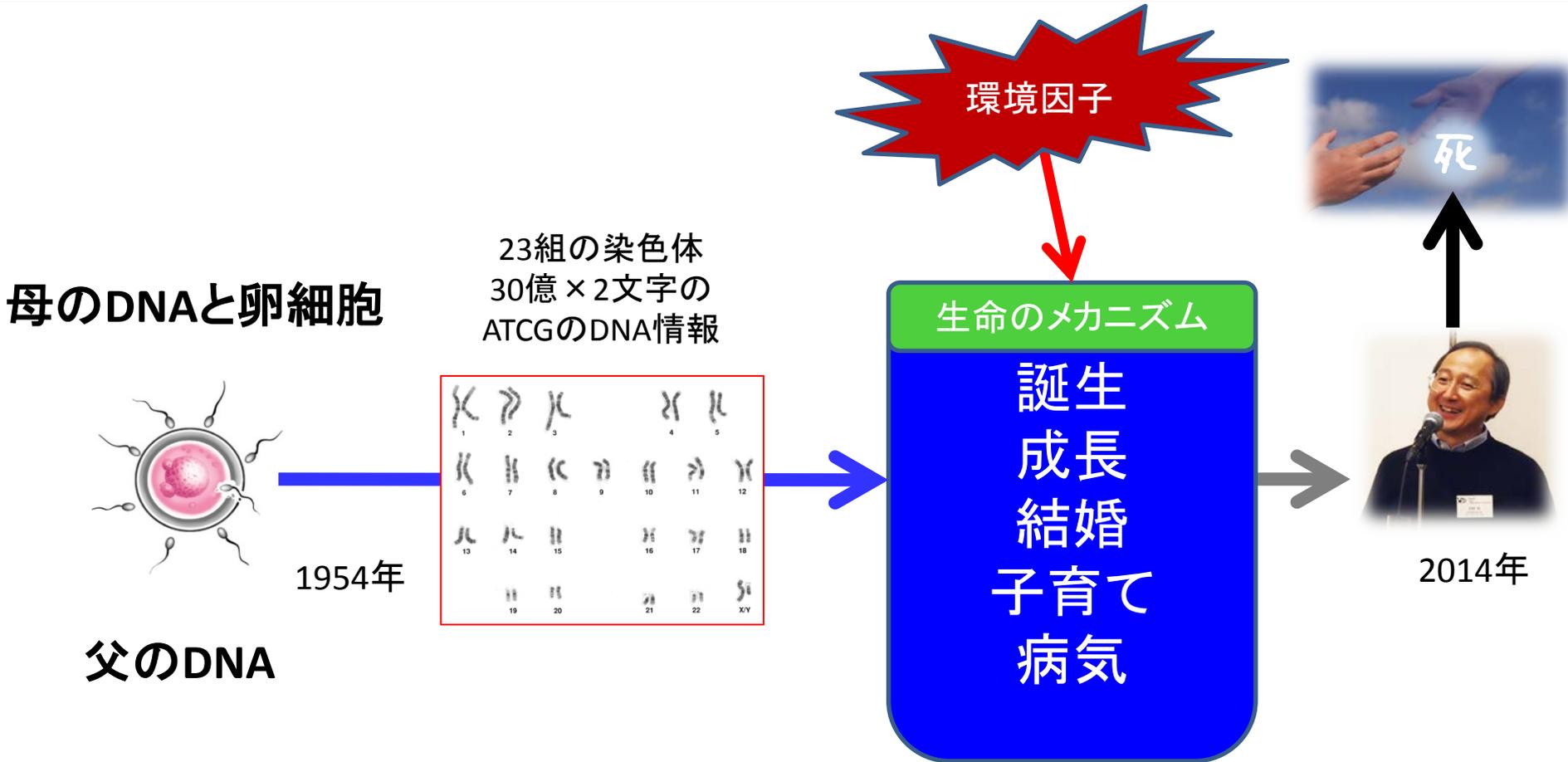
生命のメカニズム

誕生
成長
結婚
子育て
病気

死



2014年

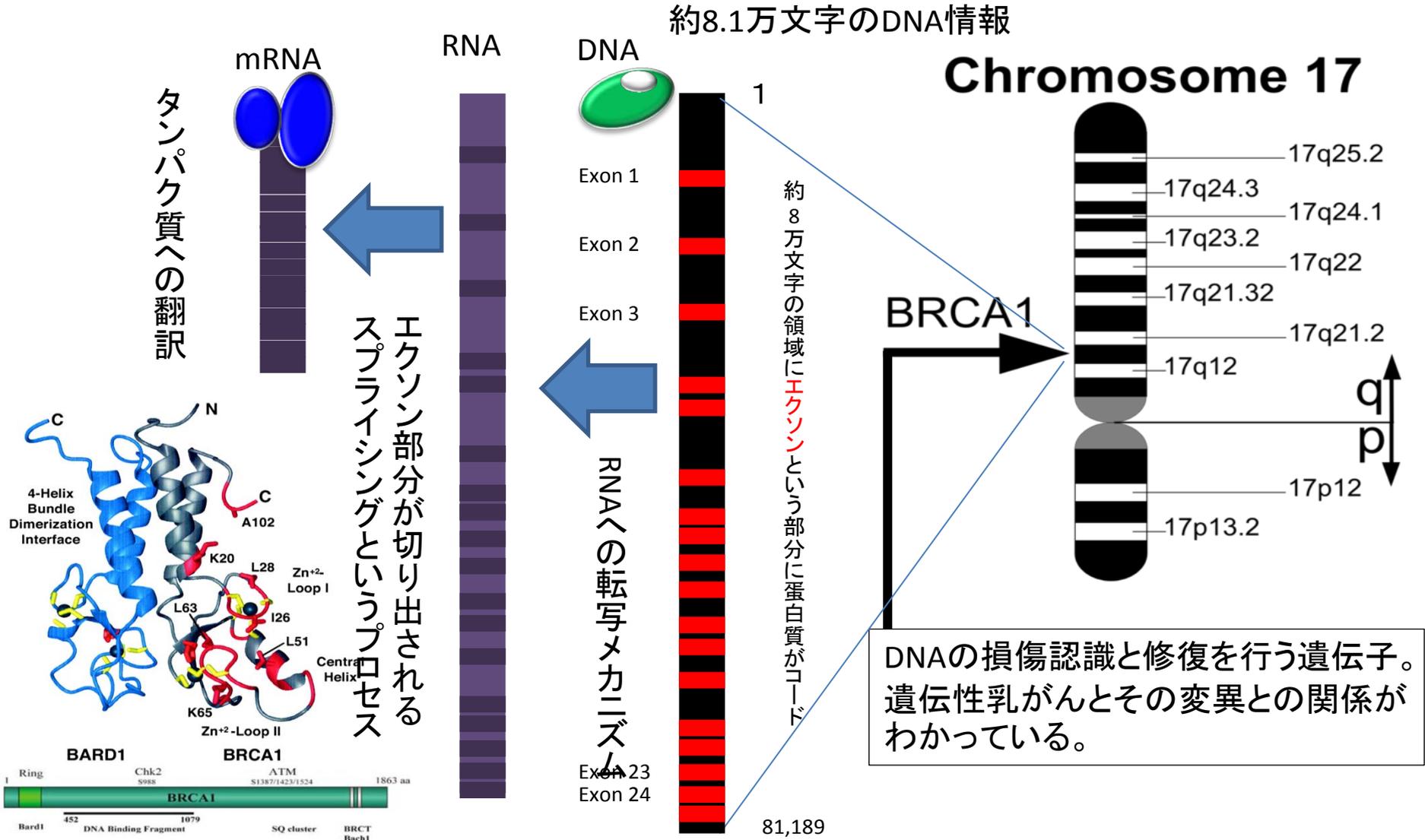


アンジェリーナ・ジョリーさん

2013年5月14日

- ハリウッド女優のアンジェリーナ・ジョリーさんが乳がん予防のために、両胸の乳腺切除及び乳房再建の手術をおこなったと告白した。
- 遺伝子検査の結果、BRCA1と呼ばれる遺伝子に変異があり、乳がん・卵巣がんになる確率が一般の人よりもかなり高いことが分かったため、がんのできる乳腺を切除する手術に踏み切ったと明かした。

ヒト17番染色体



私はなぜがんになるの？

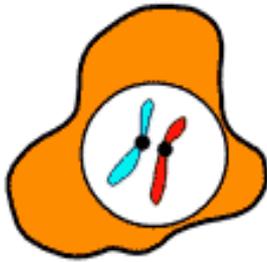


- 個人個人で異なっている遺伝的要因(ゲノムDNA)
- 腫瘍細胞に蓄積した遺伝子変異(がんゲノム)
- 環境要因などによるゲノム修飾の変異(エピゲノム)



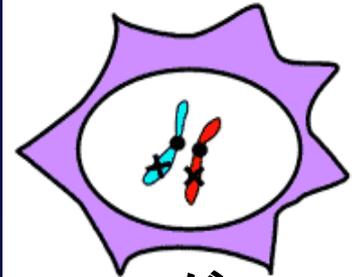
がんの悪性度や治療応答性、副作用の出やすさなどを規定している

がんを捉える第一手は



正常細胞

「私」のゲノム
の情報が
不可欠



がん細胞

My ゲノムを「シーケンス」

生物のゲノム情報を読み取る装置は一般にシーケンサーとよばれ、A, T, C, Gの文字で綴られるゲノム情報(ヒトの場合30億文字の情報)をコンピュータで読めるように取り出すことを「**シーケンス**」とよんでいます。

Dear Dr. Satoru Miyano,

Tue, 08 Jul 2014 09:26:52 -0400

Greetings!

MacroGen, one of the exhibitors at the TCGC in San Francisco, CA, has been the corporate partner of choice on **genomic sequencing** for many academic and commercial organizations.

ヒト全ゲノムシーケンスが1000ドルで!!

Human Whole Genome Sequencing (hWGS) starting at \$1,000!!

MacroGen is one of the few organizations in the US that offers CLIA and/or research hWGS services using the latest technology from Illumina, the **HiSeqX Ten** platform. Our turnaround time is four weeks or less.

We look forward to having the pleasure of personally speaking with you more. Please feel free to contact us if you would like to obtain more technical information or specific pricing for services. Have a great week!

Sincerely,

Susan Chung

Marketing Manager

MacroGen Clinical Laboratory

1330 Piccard Drive Suite 205

Rockville, MD 20850

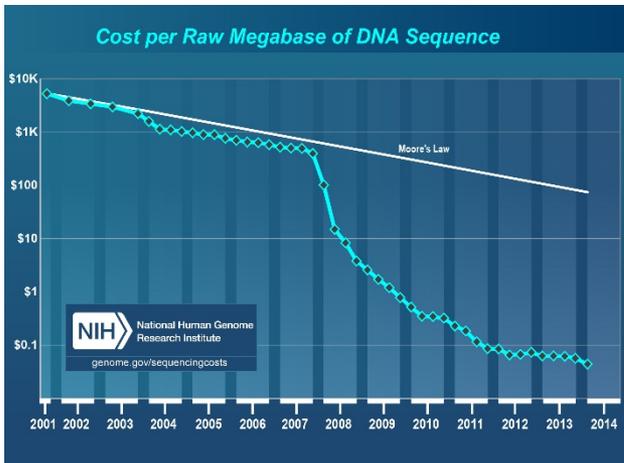
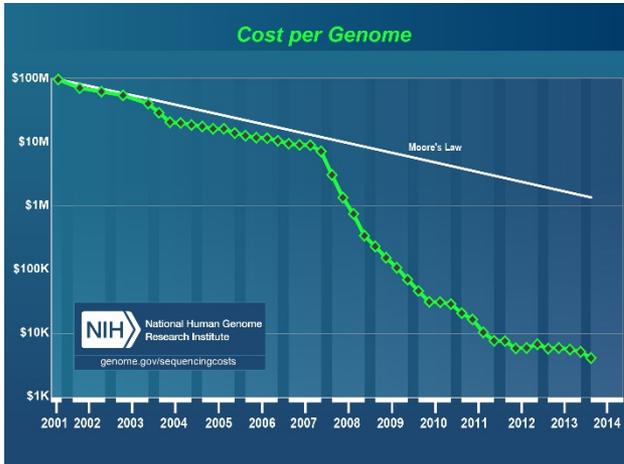
Tel. 301) 637 - 9023

ヒトゲノムは今1000ドル、そして100ドル時代へ

◇シーケンス技術の飛躍的進歩◇

◆だれもが自分のゲノム情報を利用できる時代が始まった◆

2013年



2014年、現在

HiSeq X™ Ten
Illumina社



1000ドルゲノムを達成
全ゲノム1万8千人/年
平均6テラベース/日
装置:10億円

Ion Proton™



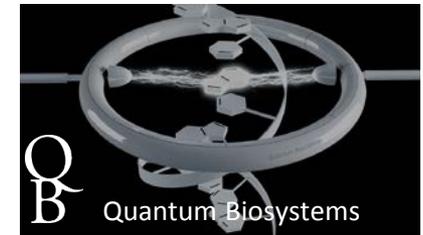
シリコンシーケンサーの普及
シリコンチップは安い・速い

1チップ+試薬=600ドルで15キガベース、6時間
2014年後半には100キガベースのChip IIが登場
装置:1500万円

2015年～

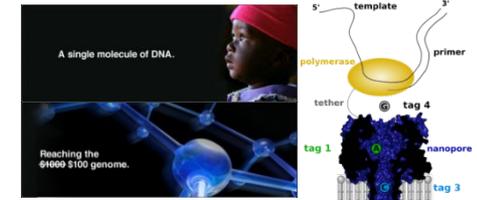
新技術の実用化
ナノポア・シーケンサー
1時間以内、1万円
超安価・高速・高精度

トンネル効果電流を測定することでATCGを識別



我が国独自のシリコンシーケンサー技術
川合FIRSTプロジェクトの成果による
大阪大学発バイオベンチャー企業

DNAをナノポアタンパク質に通してシーケンス



1時間以内、1万円;装置:10万円
100万ナノポア集積
COMSチップ→数百円

システム異常といっても・・・ さらに、がんは複雑

この複雑さをどのように理解すればよいのか？
気持ちには、ムンクの『叫び』

- 原発から浸潤、転移へ、がん細胞は変幻。
- がん幹細胞や、正常細胞である血管内皮細胞、免疫炎症細胞などと「つるみ」ながら、
- 薬剤耐性を獲得し、
- 時空間で進化する不均一な細胞集団
- 複数の原発が進化することもある！

スパコンで加速できた がん研究の事例

スパコンでがんの原因遺伝子と変異を探しだし、なぜがんを引き起こしているメカニズムを探し出す。



新学術領域研究 システムがん

(複合領域: 4201) 2010.6-2015.3

システムの統合理解に基づくがんの先端的診断、治療、予防法の開発

領域代表 宮野 悟 PhD

東大医科学研究所ヒトゲノム解析センター

A01班長 稲澤譲治 MD PhD

東京医科歯科大難治疾患研究所

A02班長 高橋 隆 MD PhD

名古屋大学医学研究科

数学とスパコン

を駆使した
大規模データ解析と
数理モデリング

融合

腫瘍学

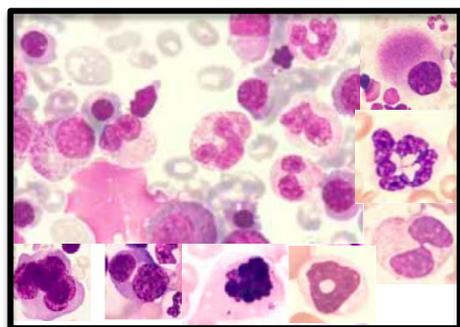
がん病理学
がん生物学

骨髄異形成症候群(MDS)の 原因遺伝子の発見



小川誠司 (おがわ・せいし)

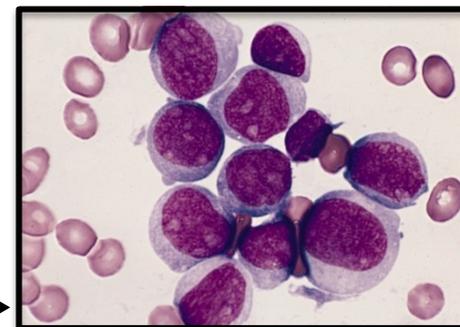
難治性血液がんの克服に向けて



骨髄異形成症候群

A02 小川誠司
(東大医学部附属病院)
(現在、京都大学医学研究科教授)

移行



急性骨髄性白血病

◇いろいろな血液細胞が
作れなくなることで
赤血球減少 → 貧血、動悸、息切れ
白血球減少 → 感染しやすくなる
血小板減少 → 出血しやすい

MDS: myelodysplastic syndromes

【共同研究】

A01 宮野 悟(東大医科学研究所)

歳をとらない人 その後ろで歳をとっていった人

磯野カツオ

(11歳、小学5年生)
テレビ番組「サザエさん」より。
1969年10月放送開始

声の出演

高橋和枝

(2代目:28年間)

1970年1月11日~1998年5月10日

つまり、

『年寄りはいらいよ』

骨髄異形成症候群で、1999年3月23日
東大医学部附属病院で亡くなる(70歳)。

➤ どんな病気？

- 骨髄で正常な血液を作れなくなる、代表的な「血液がん」。
- 急性骨髄性白血病への移行が特徴。
- お年寄りに多い。日本で、推定で毎年新たに5千人が罹患。
- **その原因は不明だった。**

➤ 治療法は？

- 骨髄移植しかない。
- しかし、骨髄移植の適応は一般に60歳まで。

大規模次世代シーケンサー解析、スパコンをフル活用した数理解析チームとの共同により、研究開始から論文発表までわずか1年ほどで(2010年7月～2011年9月11日)、未解決の大問題を解決・発表。



Illumina

スパコンと次世代シーケンサーを使い、がんの新たな原因機構を発見した我が国初の大規模ゲノムシーケンス研究

臨床的・生物学的検証

MDSの原因は不明

2010年

次世代シーケンサーHiSeq2000を用いて、29検体の全エクソン・シーケンスを行う

候補変異を高速リシーケンス技術を用いて、多数検体について効率よくスクリーニングする技術を開発

- 582例の検体を用いた変異解析で再現確認。
- 変異体を導入HeLa細胞でRNAスプライシング異常が起こることを確認。
- 変異したスプライシング因子が、造血機能に及ぼす効果をマウスに骨髄移植して検討。

ヒゲノム解析センタースパコン上に、超高速、効率的に変異遺伝子の候補を検出するデータ解析パイプラインを構築

高精度変異同定アルゴリズムを開発。血液腫瘍試料には正常細胞が混入しているという困難を克服。

268個の体細胞変異を同定
RNAスプライシング経路の4つの遺伝子に異常が頻出
がん研究の歴史で初めての知見

原因を解明

2011年9月
Nature 2011

データ解析パイプラインを公開



当時6000コア(75TFLOPS)
1PB高速ストレージ



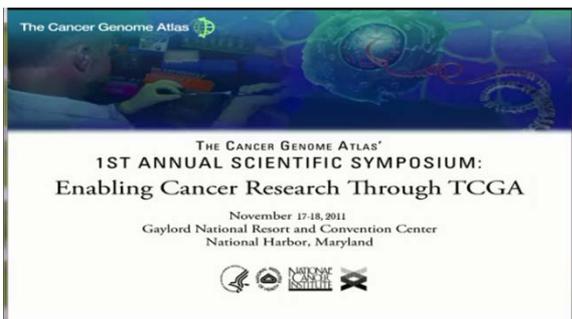
エクソーム解析・変異解析



融合遺伝子解析



がん研究の歴史に刻まれる発見



TCGA: 2005年に開始。これまでに毎年約50M\$の研究費



Keynote Speechで紹介(2011年11月17日)

Eric Lander, Director of Broad Institute of MIT and Harvard

Short Biography
BS: Mathematics
MS: MBA
PhD: Molecular Biology

Insights from Genomic Approaches	
Functional classes	
Protein and lipid kinases	BRAF (melanoma), PIK3CA (breast), PIK3R1 (GBM), EGFR (lung), FGFR2 (endometrial), JAK2 (myeloproliferative)
Lineage survival genes	MITF (melanoma), NKX2-1 (lung), SOX9 (colon), SOX2 (squamous lung and esophageal),
Epigenetic regulators	DNMT3A (AML), EZH2 (DLBCL), UTX (many), MLL2/3
Metabolic enzymes	IDH1/2 (GBM and others)
RNA splicing factors	SRSF2, SF3B1, ZRSR2, U2AF1 (MDS), SF3B1 (sideroblastic anemia, CLL)
Translocations	ERG (prostate), ALK (lung)
Surprises	
Notch	Oncogene (T-ALL) and tumor suppressor (squamous)

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | Full text

Archive > Volume 478 > Issue 7387 > Articles > Article

NATURE | ARTICLE [- previous article](#) [next article >](#)

Frequent pathway mutations of splicing machinery in myelodysplasia

Kenichi Yoshida, Masashi Sanada, Yuichi Shiraiishi, Daniel Nowak, Yasunobu Nagata, Ryo Yamamoto, Yusuke Sato, Aiko Sato-Otsubo, Ayana Kon, Masao Nagasaki, George Chalkidis, Yutaka Suzuki, Masashi Shiosaka, Ryoichiro Kawahata, Tomoyuki Yamaguchi, Makoto Otsu, Naoshi Obara, Mamiko Sakata-Yanagimoto, Ken Ishiyama, Hiraku Mori, Florian Nolte, Wolf-Karsten Hofmann, Shuichi Miyawaki, Sumio Sugano, Claudia Haferlach, H. Phillip Koeffler, Lee-Yung Shih, Torsten Haferlach, Shigeru Chiba, Hiromitsu Nakauchi, Satoru Miyano & Seishi Ogawa
= Show fewer authors

Affiliations | Contributions | Corresponding author

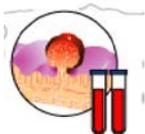
Nature 478, 64–69 (06 October 2011) | doi:10.1038/nature10496
Received 07 June 2011 | Accepted 24 August 2011 | Published online 11 September 2011

この発見は、世界で初めて、骨髄異形成症候群(MDS)の原因遺伝子を発見したという意義だけでなく、「RNAスプライシング」の異常が、がんの発症に関わることを示した世界で初めてのものの

なぜスパコンが必要か？

半端でない大量のゲノム
データの解析

なぜスパコンがゲノムデータの解析に必要なのか その一例



検体

大量の検体

先端技術シーケンサー



11日～
27時間

正常組織 900億文字
がん組織 1200億文字
のデータ

スパコンと医療インフォマティクス
データ解析(スパコンと人材)



1日～

21億ピースのジグソーパズルをとくようなもの

親から受け継いだゲノム

がんを発症する鍵遺伝子の変化

環境因子による変化を受けたゲノム

黒幕X

黒幕Z

細やかなデータ解析が必要

がんの診断・治療のために IBM Watsonの導入・開発

がんでの導入事例

http://www.crainsnewyork.com/article/20140319/HEALTH_CAR E/140319845/ibms-watson-joins-genome-center-to-cure-cancer

New York Genome Center
に脳腫瘍の医療のサポート
に導入されたWatson



- IBM Watson Oncologist
“Watson helping oncologists treat cancer patients”



Memorial Sloan-Kettering
Cancer Center
The Best Cancer Care. Anywhere.

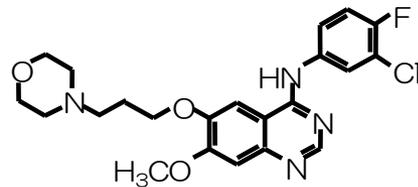
- MD Anderson's Oncology Expert Advisor
“Watson closing the loop of clinical trials,
patient care, and treatment development”



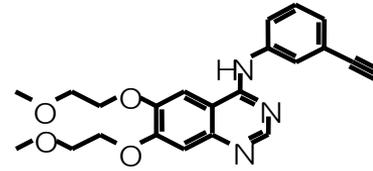
- 「天才的医師」「天才的研究者」「メガファーマ」への依存は大きなリスク
- 「普通の医師」が、バイオメディカルビッグデータとWatsonで「診断サポート」→海外で承認されているが日本では未承認の抗がん剤の特許での適用など
- 「普通の開発者チーム」が低コストで新たな形の薬を開発

スパコンで、抗がん剤が効く人、効かない人の遺伝子制御システムの異常の相違いを大規模データ解析で明らかにする。

肺腺癌の分子標的薬（キナーゼ阻害剤）



ゲフィチニブ
(イレッサ)



エルロチニブ
(タルセバ)

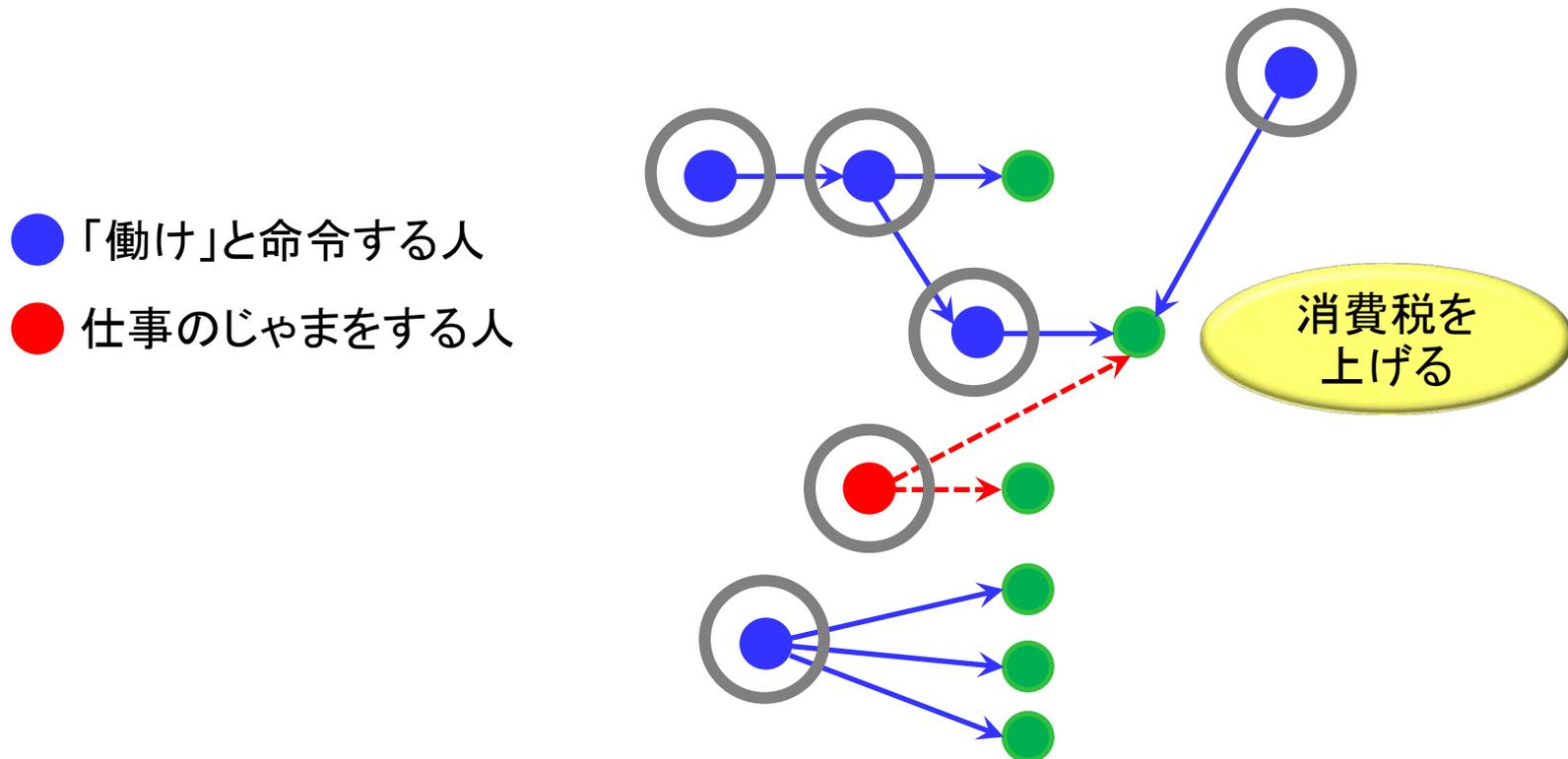
- 上皮成長因子受容体（EGFR チロシンキナーゼ）阻害剤
- EGFR（上皮細胞成長因子受容体）遺伝子に突然変異のある肺腺癌にしばしば著効を示す（奏効率80%）

がんは薬剤耐性を獲得する

残念ながら、しばらくするとEGFR遺伝子に、さらにもう1つ別の耐性変異が生じたりして効かなくなる。

遺伝子ネットワーク

～「職場・社会における人間関係」のようなもの～

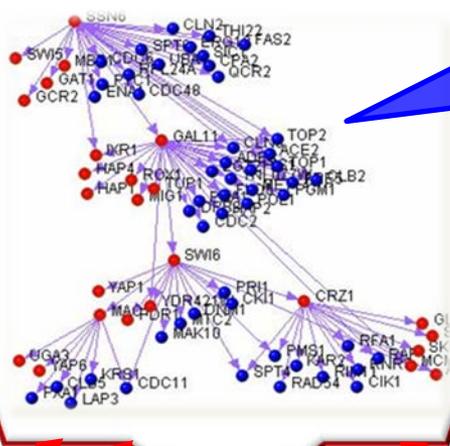


「遺伝子ネットワーク地図」 を描くためのデータとアルゴリズム

- 全遺伝子の発現状態をmRNAなどの転写産物量としてDNAチップなどで計測したものです。マイクロRNAも含まれています。先ほどのシーケンサーによっても定量計測できるようになりました。
- 先ほどのたとえで言いますと、それぞれの人の働きぶりを数値化したものです。
- 遺伝子ネットワークを推定するアルゴリズムは、職場・社会における人の働きぶりデータから、人間の制御構造とその影響力を推定します。

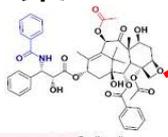
薬剤感受性・耐性

遺伝子の転写を
コントロールするのが
遺伝子ネットワーク



薬の効き方でこのネットワークがどのように異なっているかを解明したい

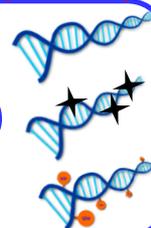
薬



Paclitaxel



- 個人個人で異なっている遺伝的要因 (DNA)
- 腫瘍細胞に蓄積した遺伝子変異 (がんゲノム)
- 環境要因によるDNAの修飾 (エピゲノム)



50%阻害濃度 (IC50)

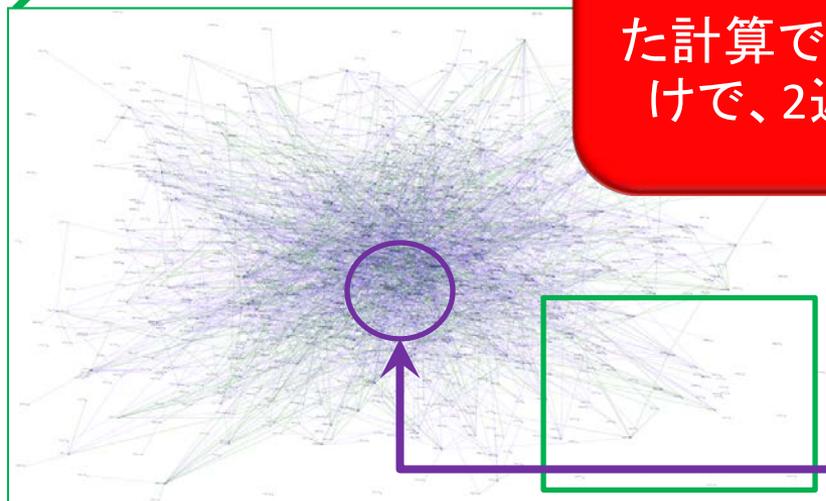
- 特定の薬物もしくは阻害剤などが注目する生物学的プロセス(がん細胞が死滅するなど)の半数を阻害するにはどれだけの薬剤濃度が必要かを示した値。
- IC50の値が小さいがん細胞に対しては、抗がん剤はよく効くが、高い値だと効きにくい(耐性をもっている)と考えられる。

700種のがん細胞株に対する抗がん剤AのIC50スコアと遺伝子発現データから個々の遺伝子ネットワークを推定できるが...

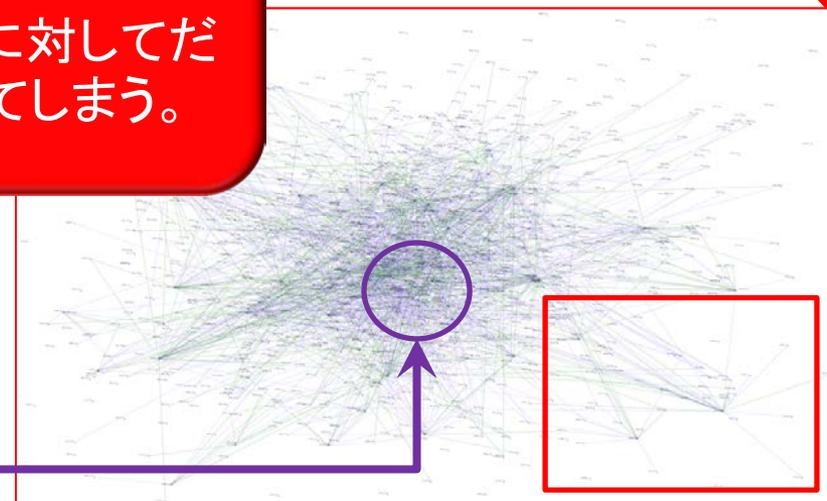
効く ← 抗がん剤Aの効き具合(IC50スコア) → 効かない



SiGN-L1というソフトウェアで
1000CPU(12.5TFLOPS)を使った
計算では、1薬剤に対してだけ
で、2週間かかってしまう。



抗がん剤Aの効く患者さん



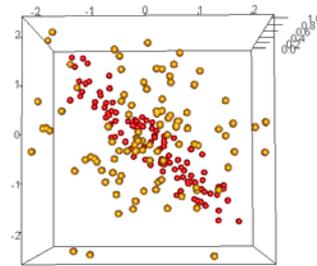
抗がん剤Aの効かない患者さん

抗がん剤Aの効く・効かない度合いで変化する遺伝子因果の関係を
ネットワークとして抽出

「京」を使うことではじめて実現

728のがん細胞株と101の薬剤に対して、薬剤感受性・耐性を遺伝子ネットワークの違いとして捉えることに成功

- がんは抗がん剤に対してやがて耐性を獲得する。「京」によるかつてない規模のデータ解析により、抗がん剤に対する薬剤耐性とがんのシステム異常を約402億個の遺伝子因果関係として抽出し、薬剤感受性・耐性の制御の姿を浮かび上げること成功。



「京」以外では発想できない解析

試験的解析

未実行の本格解析(Rを利用したRecursive Elastic Net)

1000x142x1h= 142,000ノード時間

10000x101x2h=2,020,00ノード時間

巨大化するバイオメディカル データ

米・英・カナダ

100万人のゲノムと臨床情報のデータシェアリングが
はじまった。

米国Broad Institute of MIT and Harvardでは、2014年のシーケン
スデータ量が300ペタバイトに達する

参考:クラウドGoogleは700ペタバイト、マイクロソフトは110ペタバイト

もっとメディカルな研究に利用しやすいスパコンは

- ゲノム、RNAシーケンス、エピゲノムなどのオミクスデータ解析には、メモリを多く積み、高速ディスクアレイからなるストレージが不可欠。ファイルの生成・アクセスが多く(1ジョブが5万ファイルを生成/アクセス)、小さな大量のジョブ処理に対応することが鍵。
- バランスのとれたCPU、メモリ、ストレージ、アーカイブのトータルシステムへの期待
- 演算性能偏重(TOP500)の潮流から、サイエンス重視へシフト
- 特殊なプログラミングや言語を使わずとも、それなりの性能が出るマシンを待望

ヒトゲノム解析センター (225 TFLOPS)	58,582,518ジョブ (2013年総数)
東大情報基盤センター (>1 PFLOPS)	685,879ジョブ (2013年総数)
Texas Advanced Computing Center, University of Texas (500 TFLOPS)	550,000ジョブ (2012年総数)



2桁の違い

「京」は1回メモリにアクセスする間、20演算分は休みになっている。また、ストレージも全体で30ペタバイトしかなく、既に限界になっている。

クラウド計算が癌の治療に

次の大きなブレークスルーは
試験管の中ではなく
ビッグデータの中に
見つかるかもしれない

<http://www.ink-live.com/emagazines/hemispheres-united-airlines/1699/august-2014/>

ユナイテッド航空機内誌
“Hemisphere”
2014年8月号、最終ページより