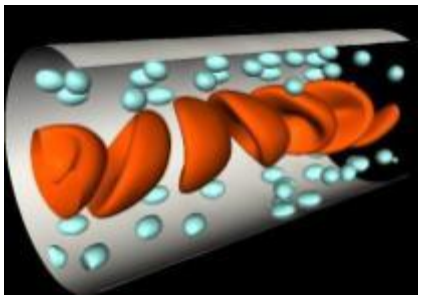
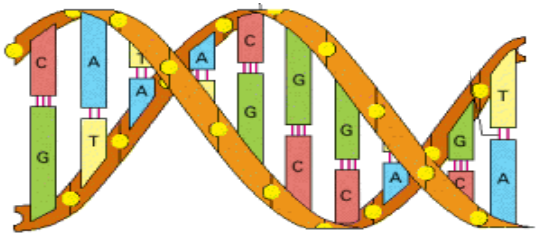
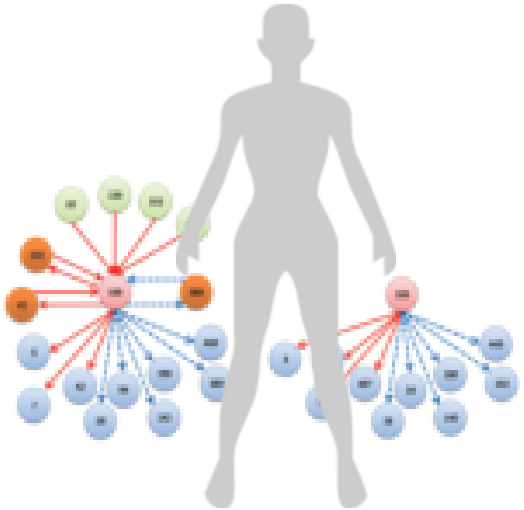
スーパーコンピュータが解き明かす生命の不思議







秋山　泰　（Yutaka Akiyama）

東京工業大学　大学院情報理工学研究科

目次

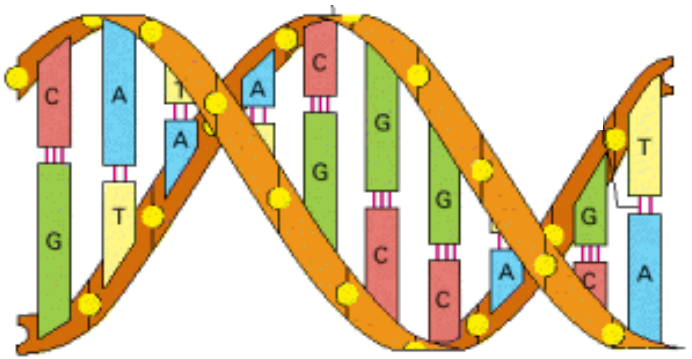
１）生命の設計図（ゲノム）の理解

２）病因の“システム的”な解明

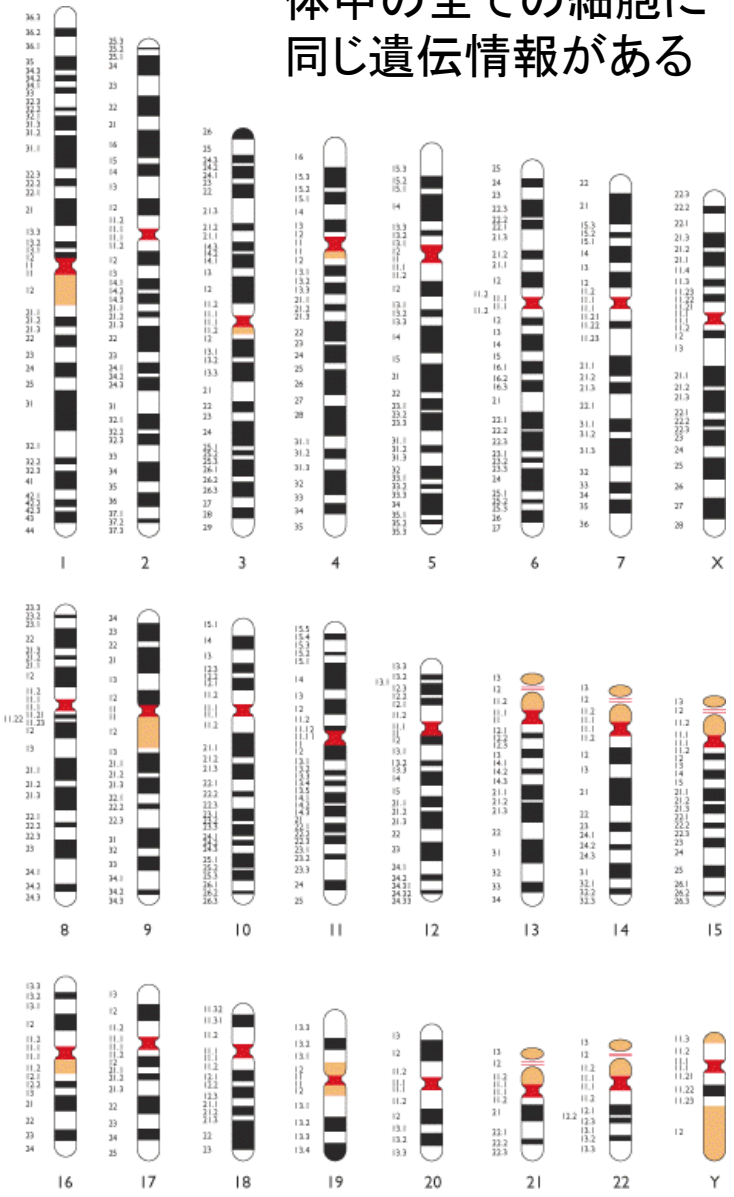
３）創薬・医療の高度化

４）スパコンと私

スーパーコンピュータが切り拓く
生命の設計図（ゲノム）の理解



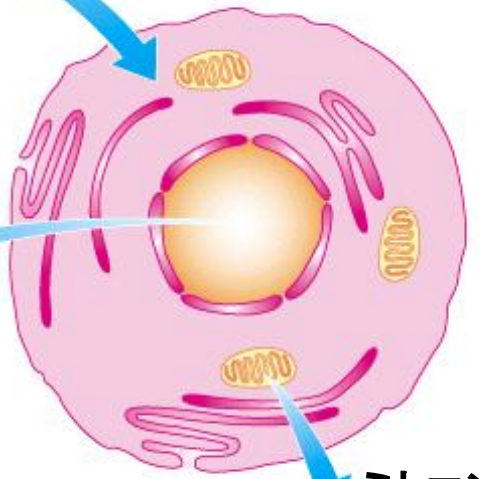
体中の全ての細胞に
同じ遺伝情報がある



ヒトの染色体　1番 ～ 22番染色体
X 染色体、Y 染色体

人間・家族

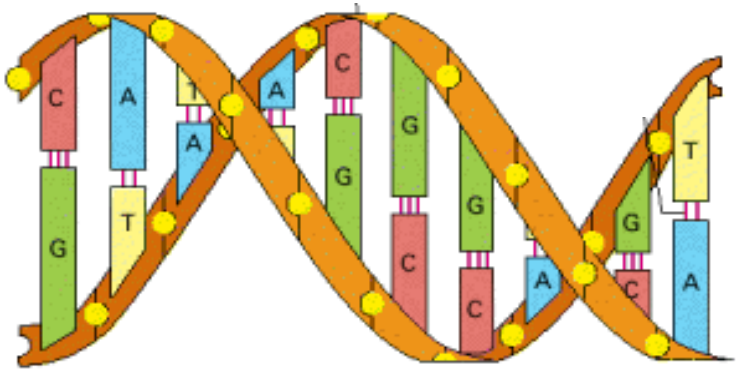
ヒト細胞



核ゲノム

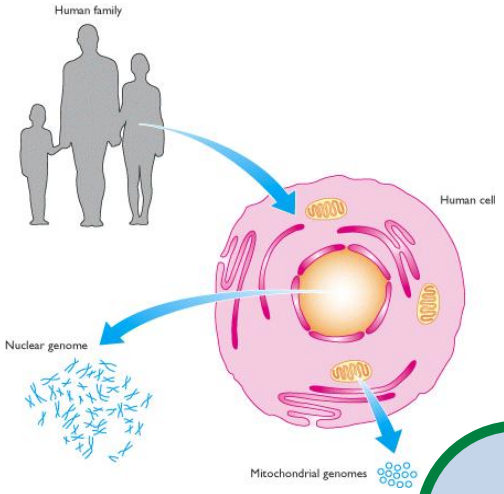
＝

ミトコンドリア
ゲノム



DNA
二重らせん

書籍 “Genomes3”より引用

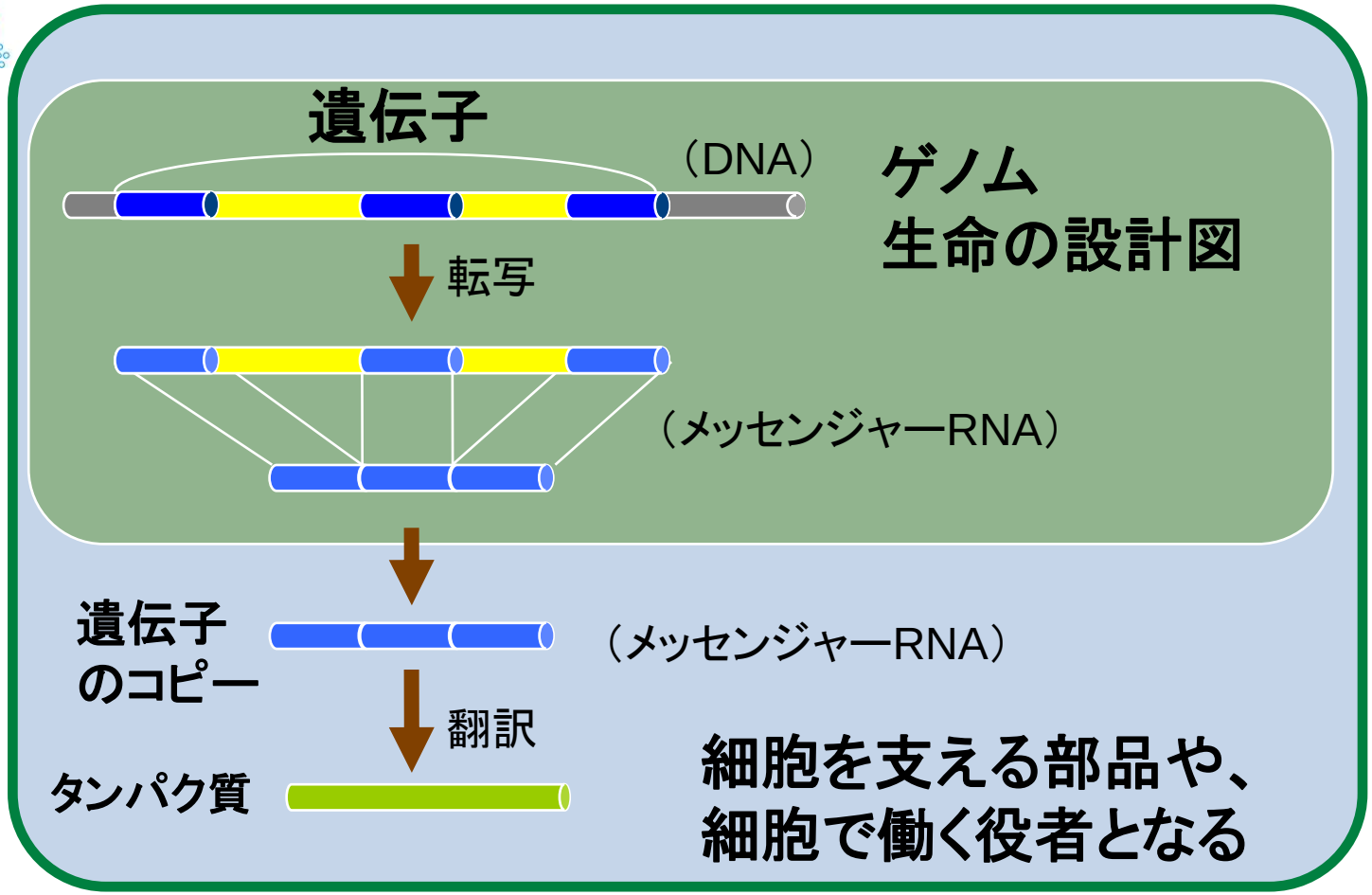


ゲノム：　**遺伝情報の総体**　**細胞内の全DNA**

英語ではGenome（＝Gene 遺伝子 + ome 総体)

遺伝子：　**遺伝情報の要素単位**　英語では Gene

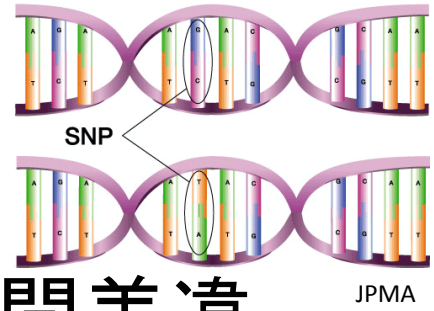
あるタンパク質を合成するなど1つの役割に対応



原図：瀬々潤准教授（東工大）

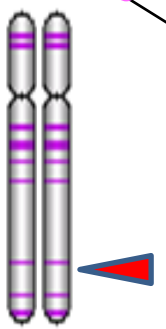
ゲノム配列の個人差

- **SNP**（Single Nucleotide Polymorphism）
  「単塩基多型」と訳される。スニップ。
  約1000塩基あたり1塩基程度の個人間差違。
  ただし人口の1%以上の頻度で存在するもの。



例：ヒト12番染色体　ALDH2遺伝子（お酒を飲んで赤くなるか？）

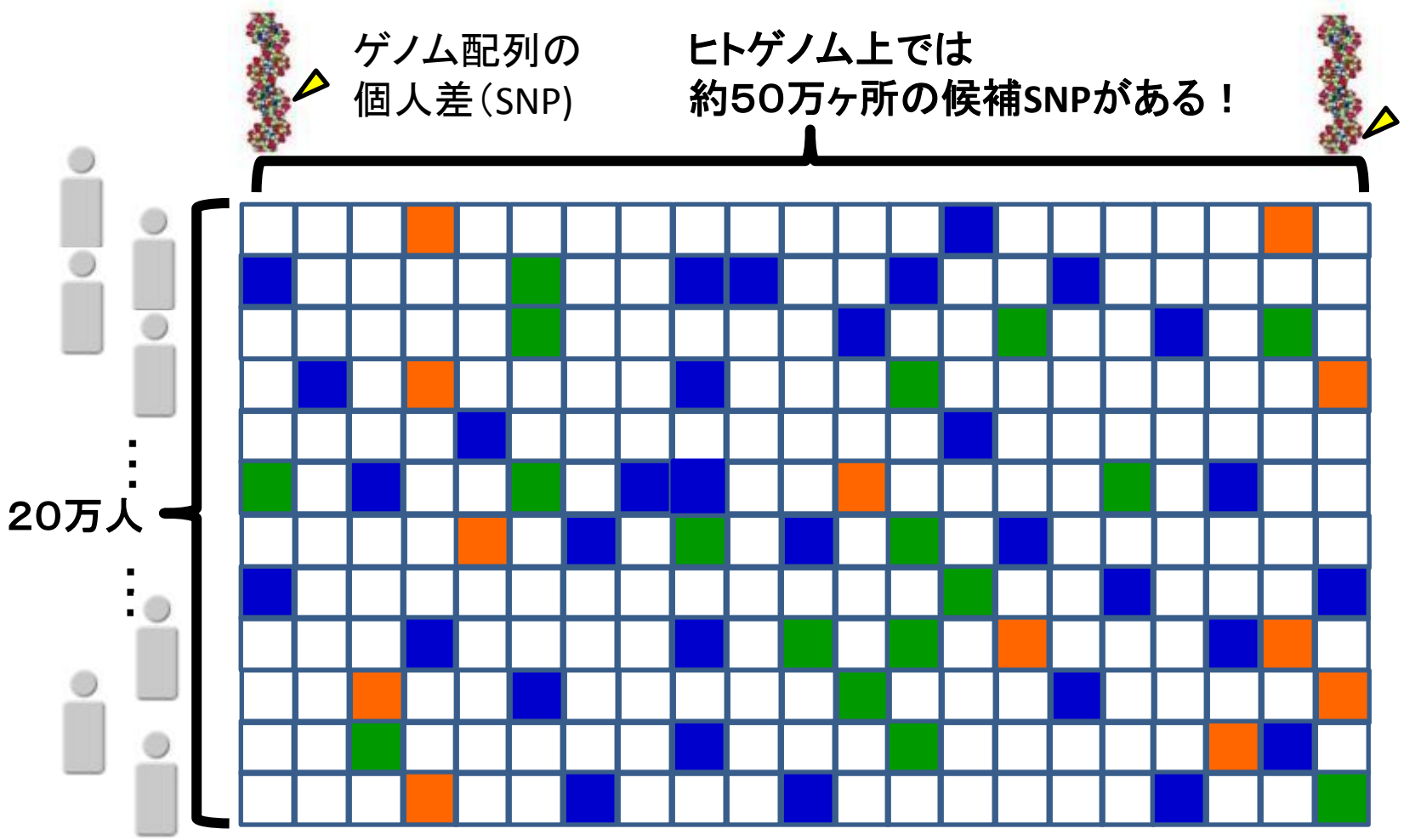
```
SQ   Sequence 135 BP; 31 A; 25 C; 48 G; 31 T; 0 other;
     caaattacag ggtcaactgc tatgatgtgt ttggagccca gtcacccttt ggtggctaca      60
     agatgtcggg gagtggccgg gagttgggcg agtacgggct gcaggcatac actgaagtga     120
     aaactgtgag tgtgg                                                     135
//
```



g (*1型)
a (*2型)

父母から受け継いだ2本の染色体がどちらも*2型だと酒に弱い。
アルコールが分解されてできるアルデヒドの処理に時間がかかる。
日本人では*1/*1型の人が約56%、*1/*2が約38%、*2/*2が約4%。

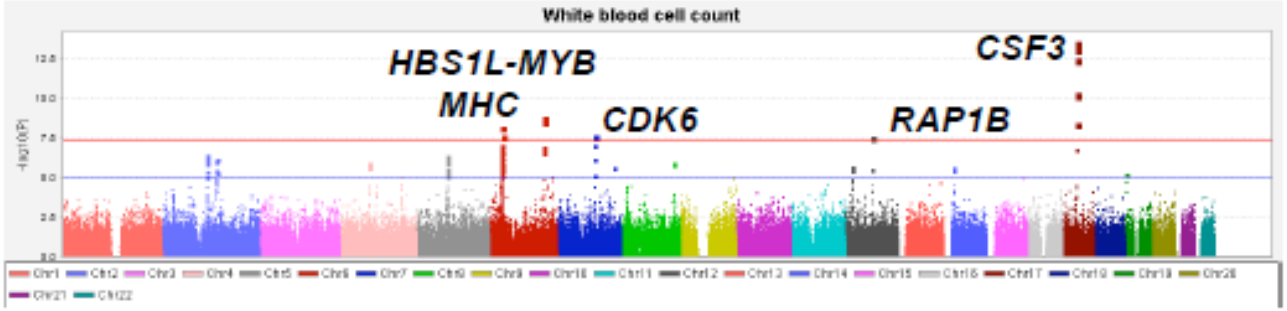
病気と連関するSNPを探したい



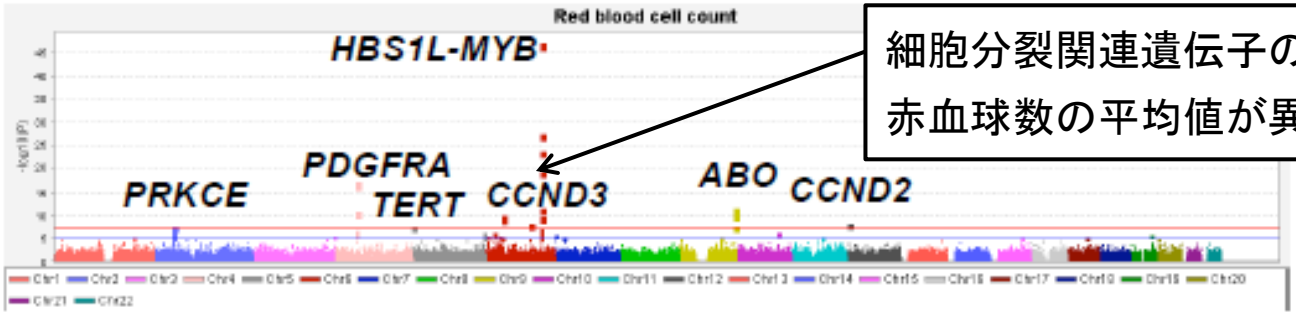
我が国の「バイオバンクジャパン」には約20万人、**47疾患**の貴重なデータがある。
（個人の情報は厳格に切り離されており、どの患者さんの情報かは判らない）

血液検査の項目の平均値がSNP型によって大きく変わる例も発見

白血球数



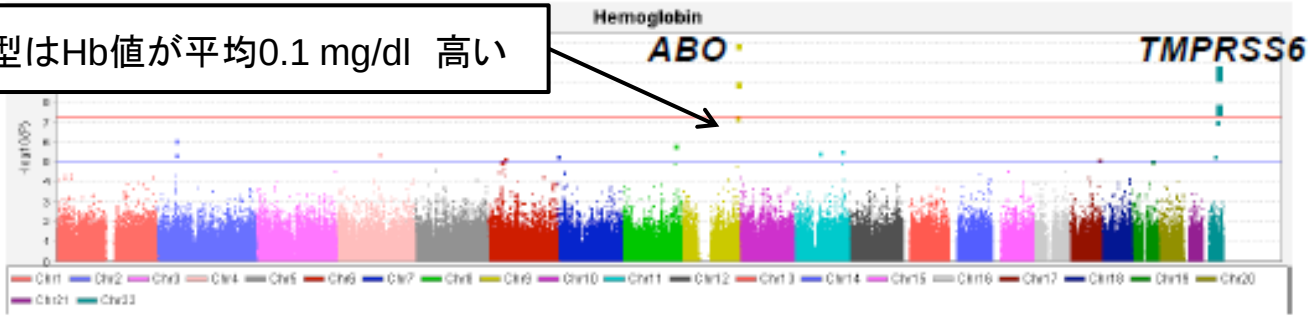
赤血球数



細胞分裂関連遺伝子の個人差で
赤血球数の平均値が異なる

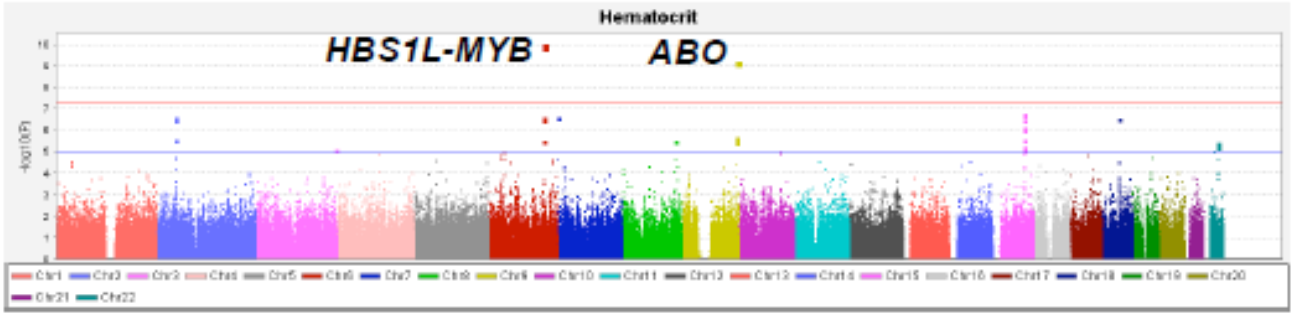
血液型B型はHb値が平均0.1 mg/dl 高い

血色素
ヘモグロビン



血液型 B型は
女性の貧血が
21%少ない

ヘマトクリット

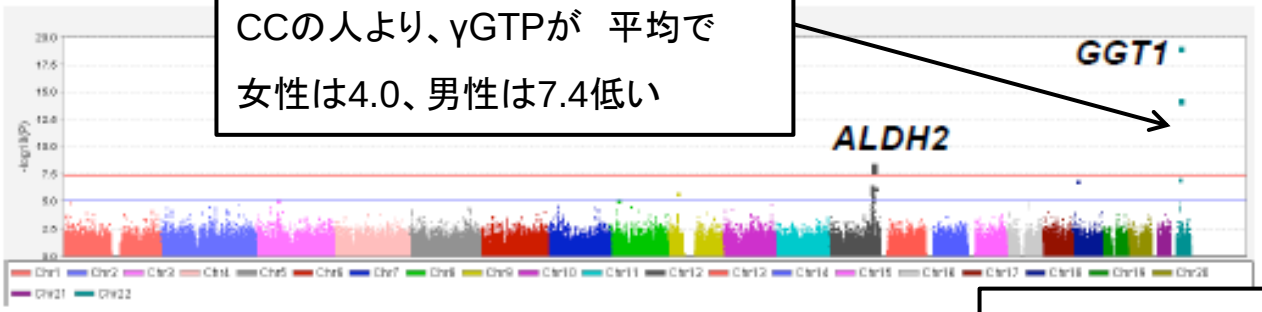


血液細胞数のゲノムワイド関連解析（GWAS），**Kamatani Y. ほか. Nature Genetics, 2010年2月**

GGT1内のあるSNPがTTの人は
CCの人より、γGTPが　平均で
女性は4.0、男性は7.4低い

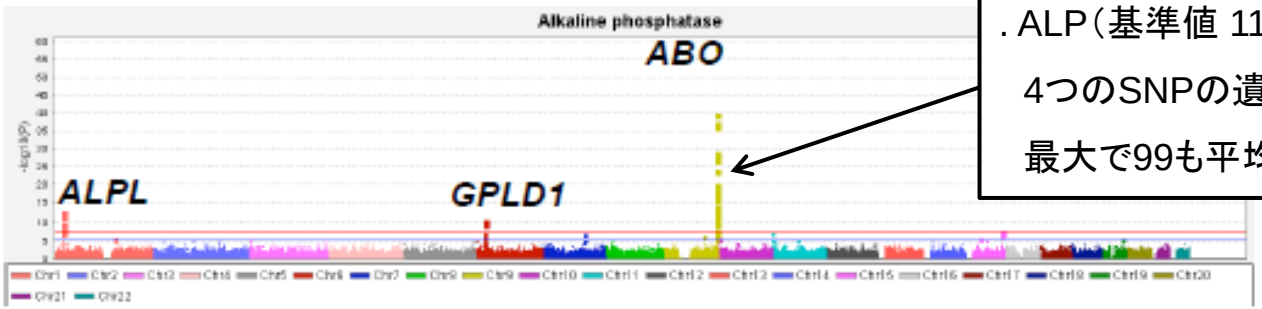
GGT

γ-GPT
肝機能



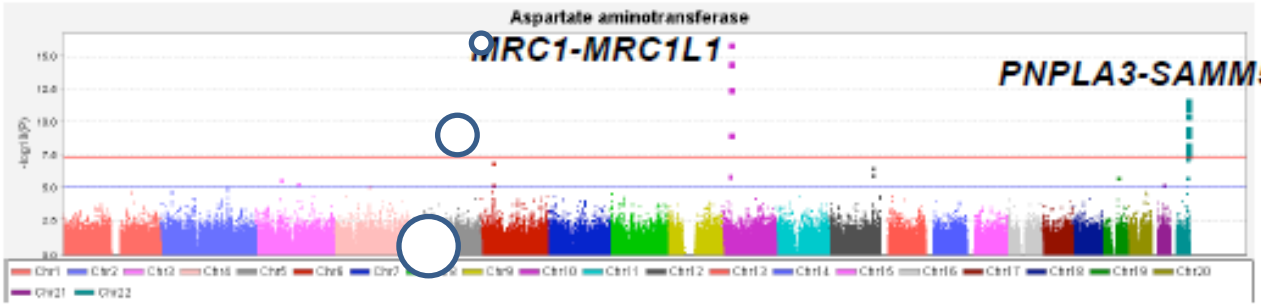
ALP

胆道
肝機能、骨

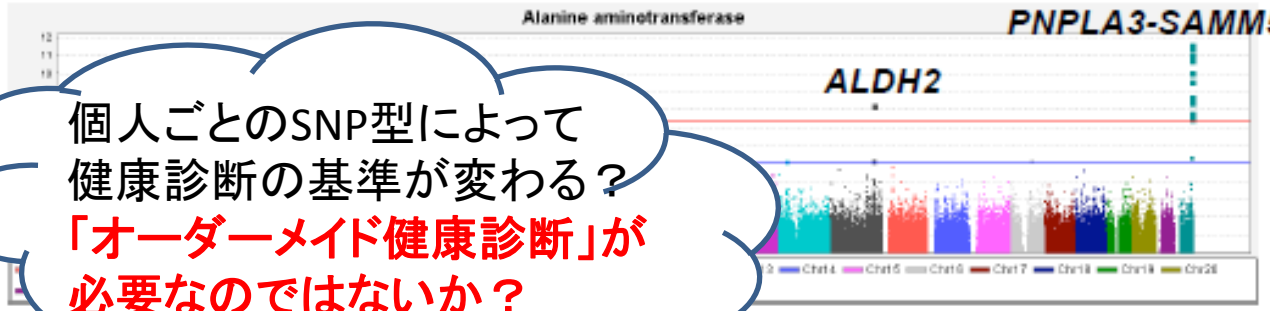


. ALP（基準値 110-354 U/L）
4つのSNPの遺伝型により
最大で99も平均値が違う

AST



ALT



個人ごとのSNP型によって
健康診断の基準が変わる？
「オーダーメイド健康診断」が
必要なのではないか？

提供：鎌谷直之博士（理研／スタージェン）

新型DNAシークエンサの衝撃

新型シークエンサ（文字数／$）
5ヶ月ごとに倍

参考：ハードディスク（MB／$）
14ヶ月ごとに倍

旧型シークエンサ（文字数／$）
19ヶ月ごとに倍

1ドル当たりの
DNA配列読取り文字数
5年ごとに4000倍超

ハードディスク価格（MByte／1ドル）

DNA配列読取りの価格（文字数／1ドル）

西暦

1,000,000
100,000
10,000
1,000
100
10
1
0

100,000,000
10,000,000
1,000,000
100,000
10,000
1000
100
10
1
0.1

1990 1992 1994 1996 1998 2000 2003 2004 2006 2008 2010 2012

Lincoln Stein, Genome Biology, vol. 11(5), 2010 を翻訳

地球上の生命は微生物群と相互作用しながら生存する

外部環境共生系

土壌圏の微生物
水田、畑地、森林等

水圏の微生物
海洋、河川、養殖場等

産業系の微生物
排水処理場、ゴミ産廃場

極限環境の微生物
南極、地下生命圏、熱水

微生物と化学物質が
地球環境の根幹を形成

内部環境共生系

ヒトの微生物
ヒト（健康と病態）の腸内、口腔、皮膚常在菌

植物の微生物
茎、葉、根の常在菌

家畜・魚類・昆虫の微生物
牛や豚、シロアリ等の腸内・ルーメン常在菌

微生物メタゲノム情報

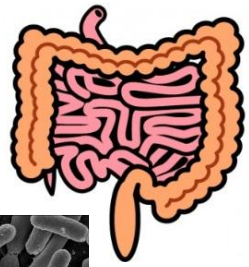
宿主情報

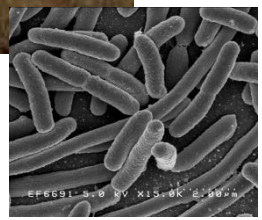
環境情報

大気・河川・森林・農場など　生物と微生物で
埋め尽くされている日本国土の情報化

黒川 顕 教授（東工大）
の原図から一部改変

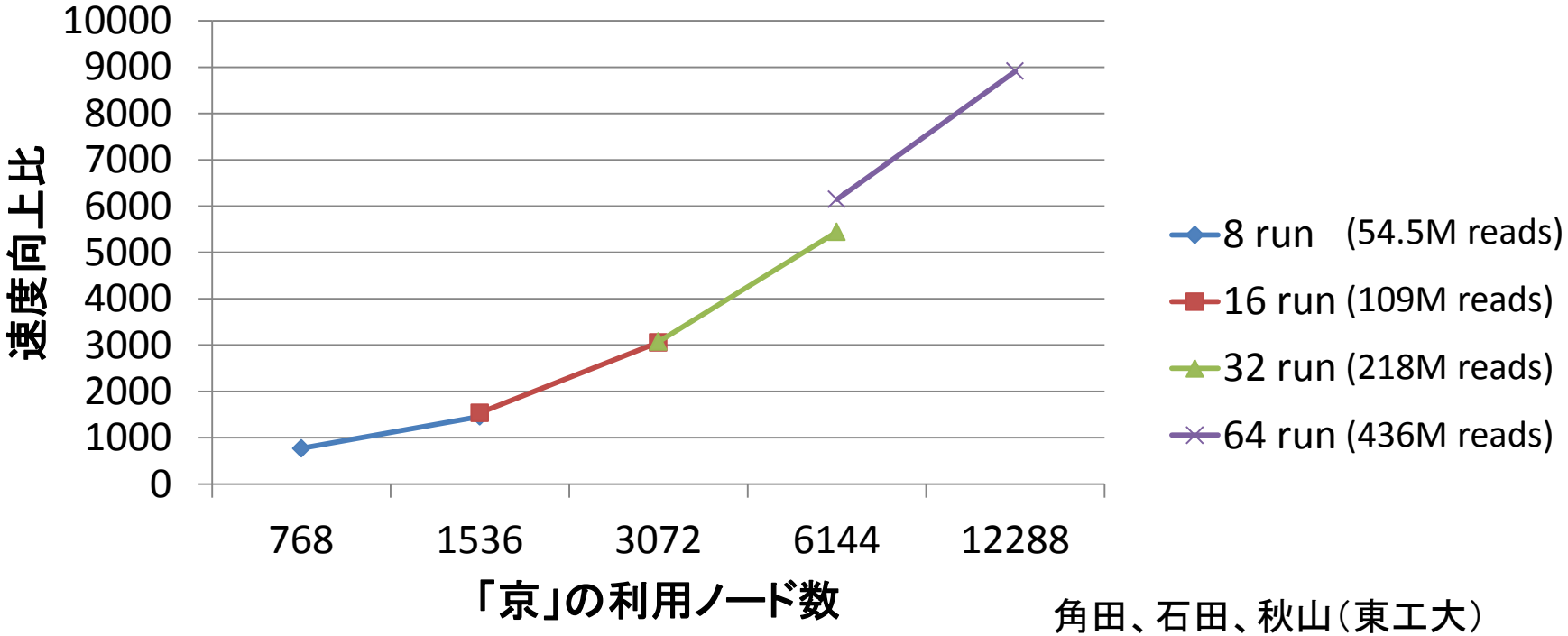
微生物メタゲノム解析の圧倒的な加速が可能に





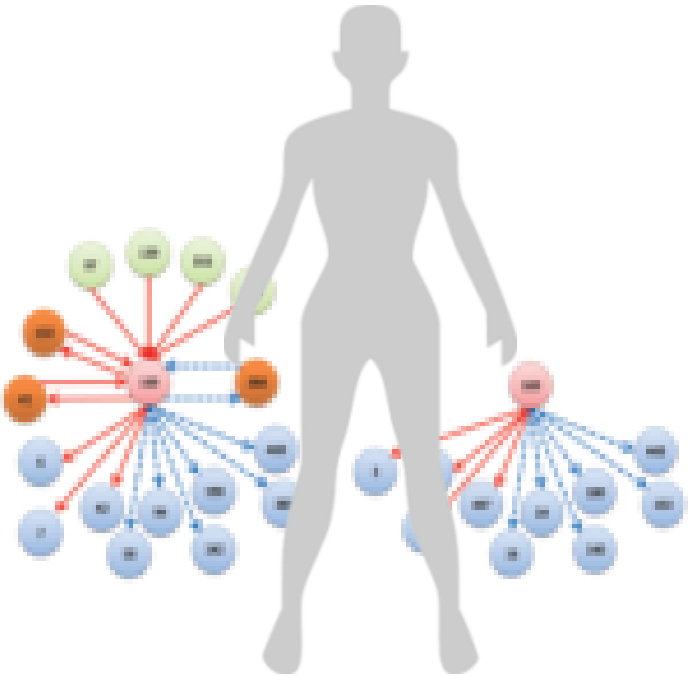
メタゲノム解析とは:
たとえば土壌中や、人の腸内にいる多数の微生物を分離培養せず、一網打尽にそのまま丸ごと解析

従来の並列計算機(144コア): 400時間(17日間)
「京」の1/8 利用: 1時間で解析可能に

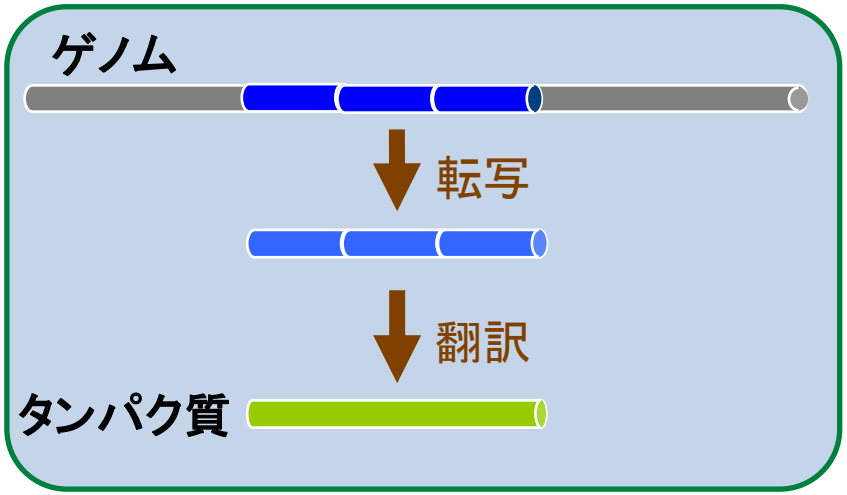


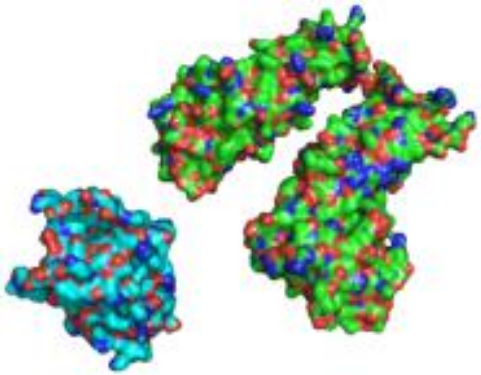
角田、石田、秋山(東工大)

スーパーコンピュータが切り拓く病因の“システム的”な解明



遺伝子はネットワークで動く

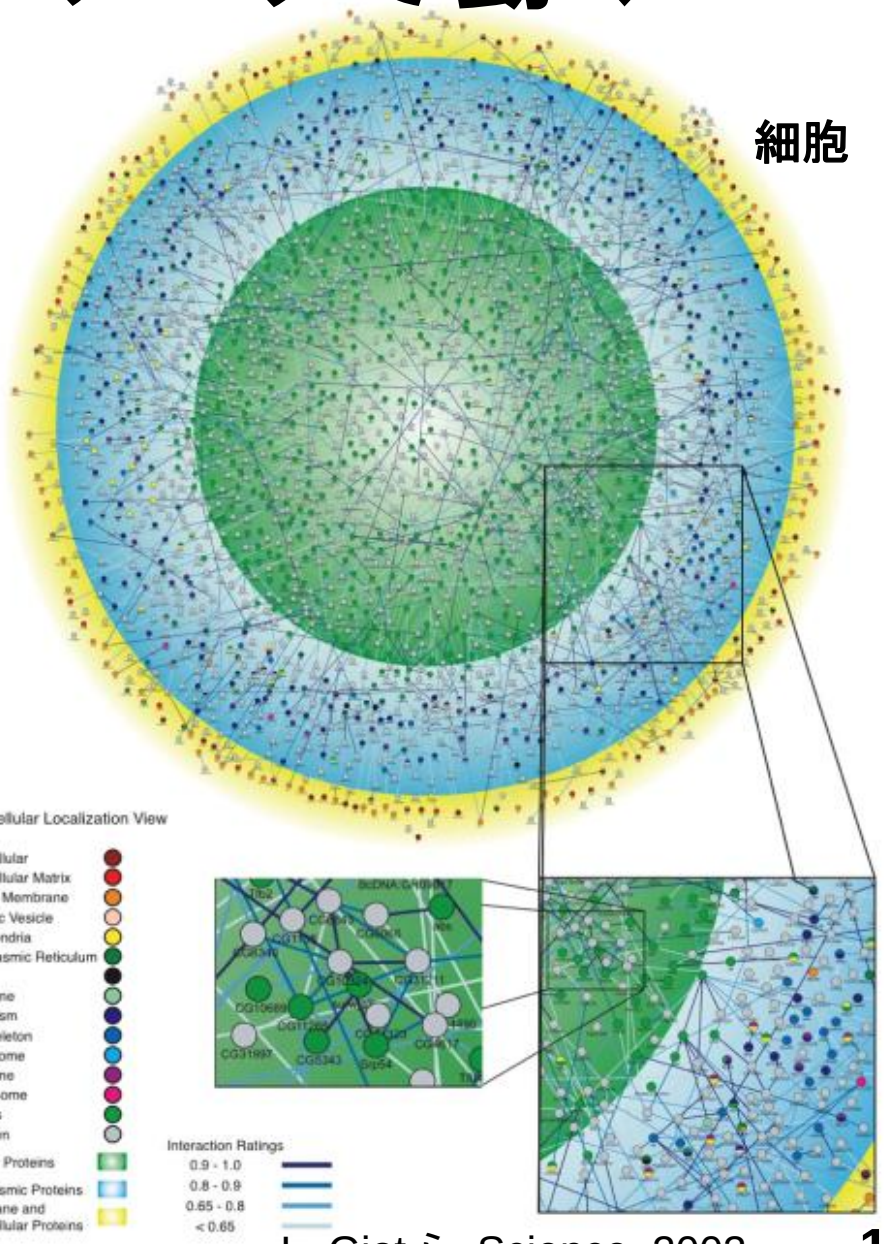




タンパク質は
ドッキングする。

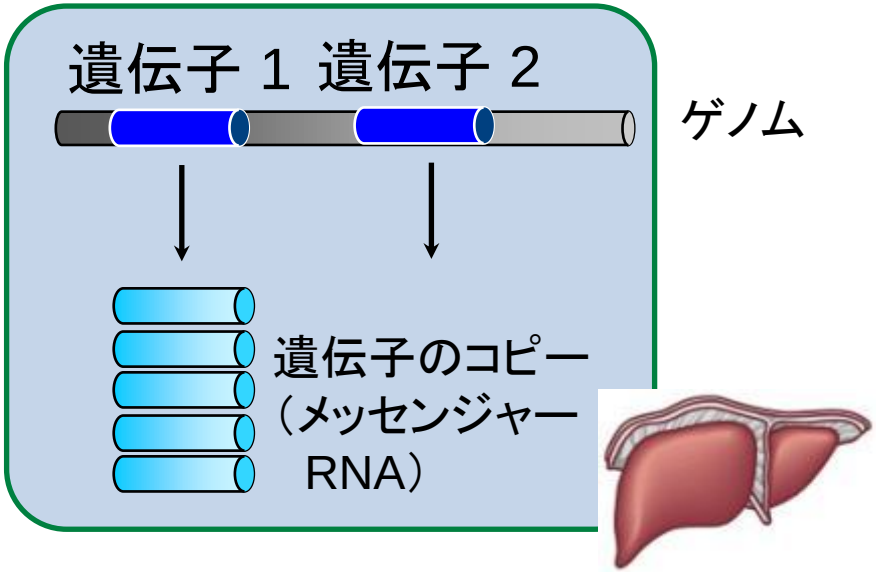
相互作用関係
の複雑な網。

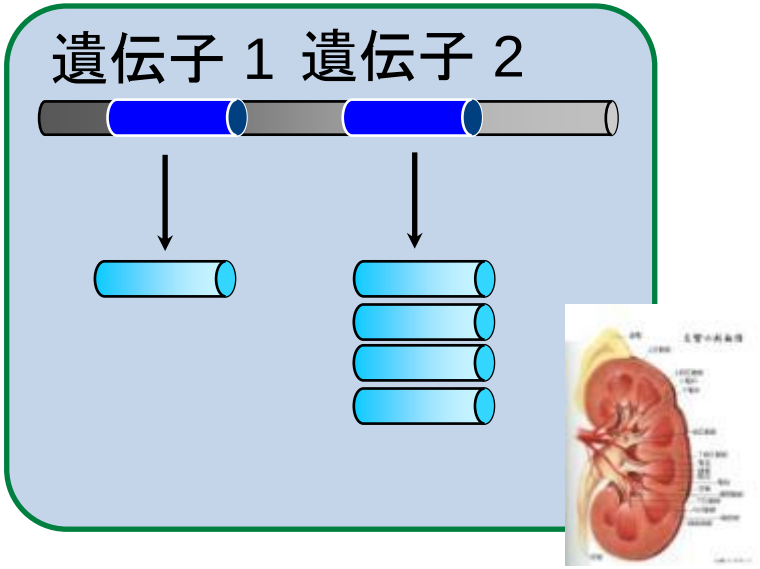
「木を見て森を見ず」では、
病因の解明に到達できない。
抗がん剤に耐性ができる、等。



L. Giot ら Science, 2003.

遺伝子発現の違い





ある個人について、全ての細胞は同じゲノム（生命の設計図）を持っている。肝臓も腎臓も脳も皮膚も、元々は全て同じ設計図。

しかし細胞は役割を認識し、「コピーのしかた」（**遺伝子発現**）が臓器ごとなどに異なっている。がん等の病気では、その種類ごとに特定の遺伝子グループの発現量に変化が見られることがある。

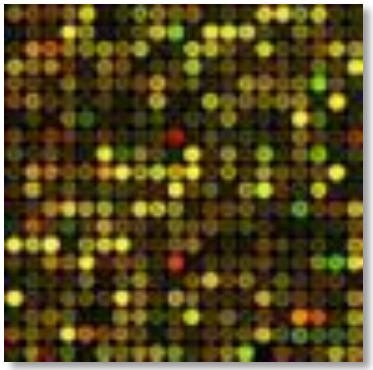
違いを調べる

個人差

組織間差

時間による差

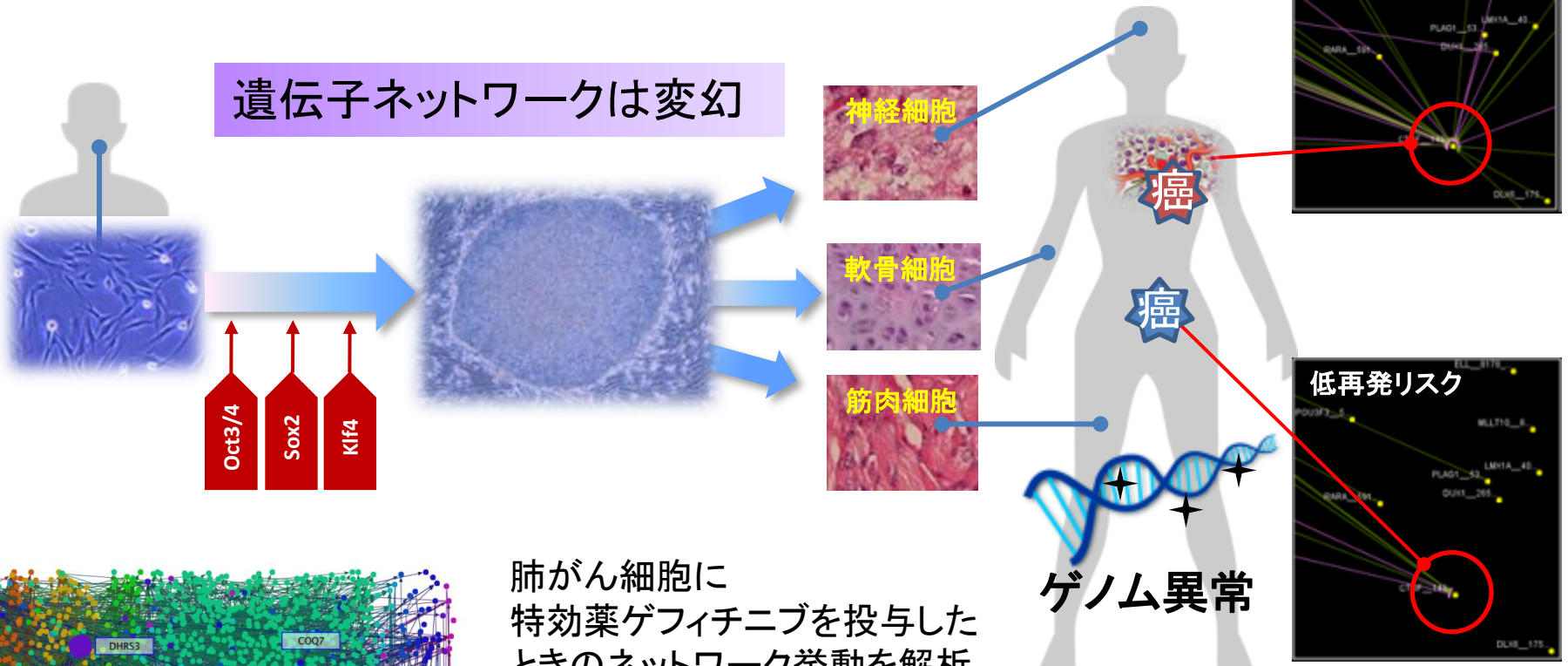
種差

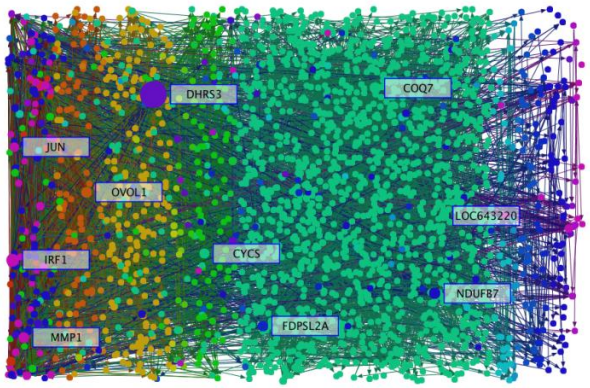


瀬々潤准教授（東工大）の原図を一部改変

ネットワークの違いから診断

がんの「個性」を、遺伝子ネットワーク全体として理解。
がんの再発リスクの高さが評価できるようになってきた。



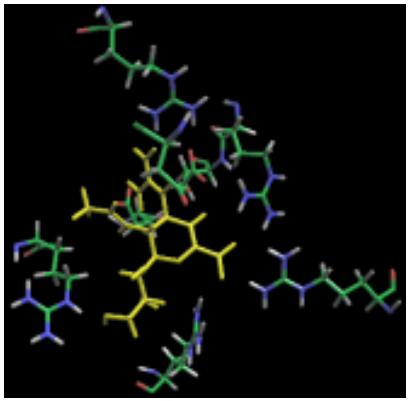


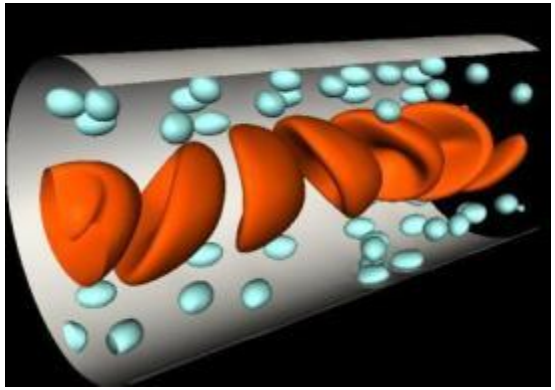
肺がん細胞に
特効薬ゲフィチニブを投与した
ときのネットワーク挙動を解析

数万CPUコアが必要となる
きわめて複雑な確率計算

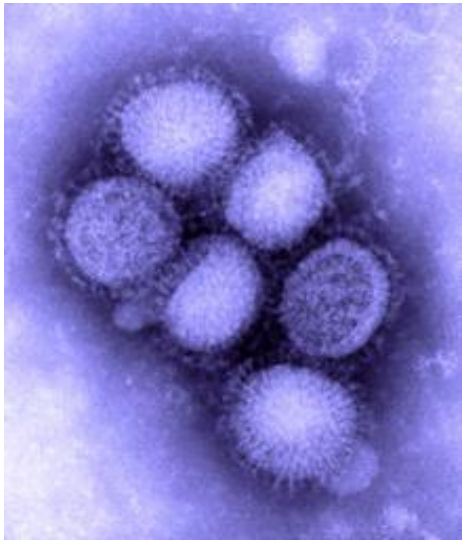
宮野悟教授（東大医科研／理研）の原図を改変

スーパーコンピュータが切り拓く
創薬・医療の高度化

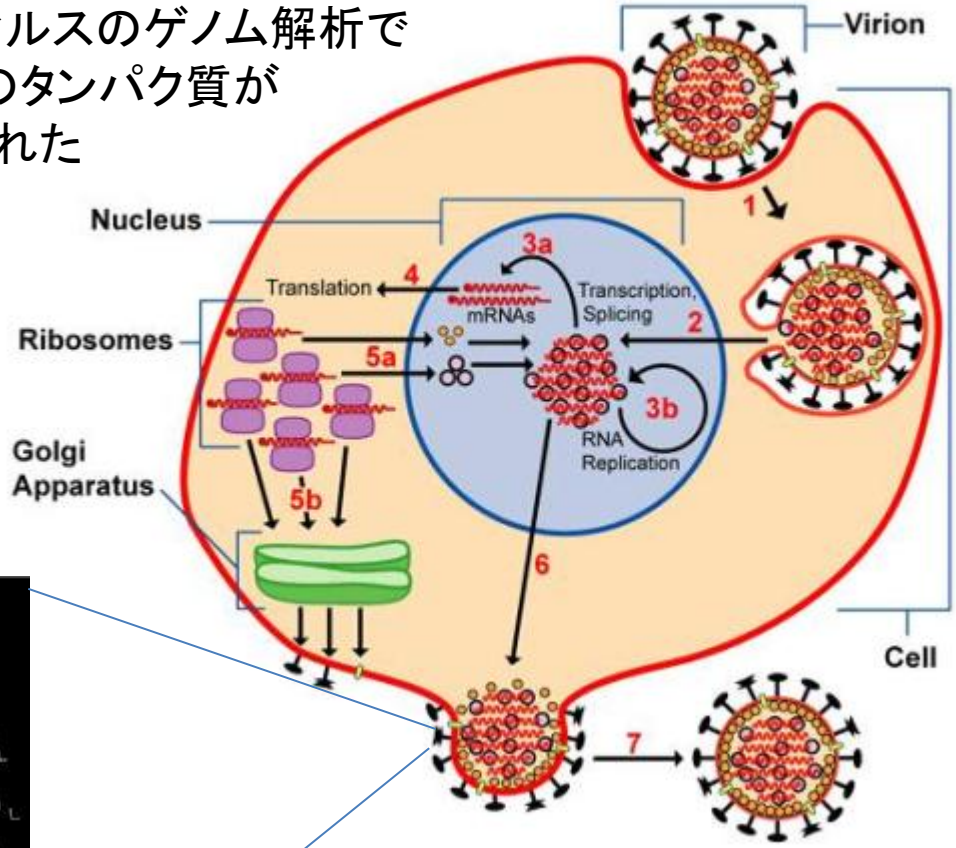




抗インフルエンザウィルス薬の開発

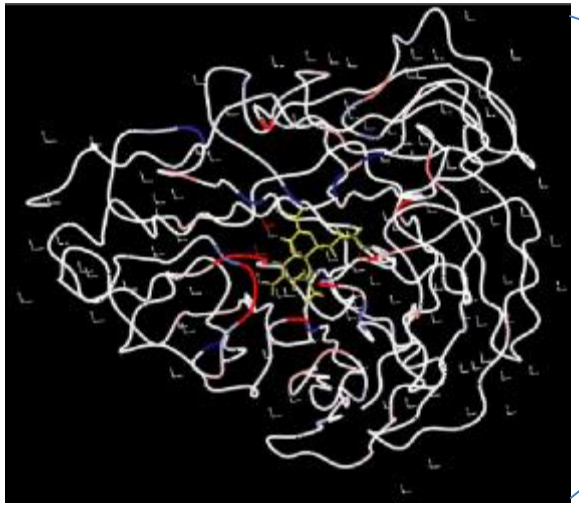


Flu ウィルスのゲノム解析で
10種のタンパク質が
発見された



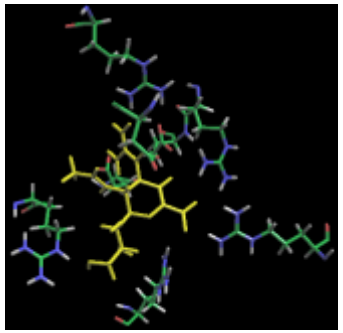
宿主細胞内
でのウィルス
増殖の機構
が解明された

NCBI Web
より引用



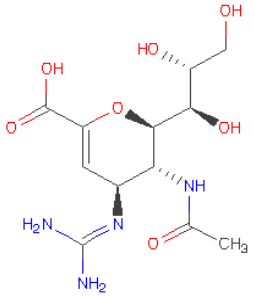
ノイラミニダーゼ酵素　(NA)

インフルエンザウィルスが作り出す
「ノイラミニダーゼ酵素」の機能を邪魔
できれば、たとえ細胞内で増殖しても
宿主の細胞膜から離れて移動できない



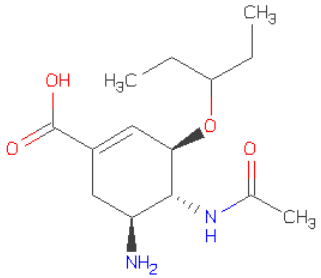
ノイラミニダーゼ酵素(NA) のくぼみ（ポケット）部分にぴったり
はまり込み、NAの役割を邪魔（阻害）するような構造を持った
化合物を、計算機上で、探す競争が世界中で繰り広げられた。

ザナミビル
（商品名　リレンザ）



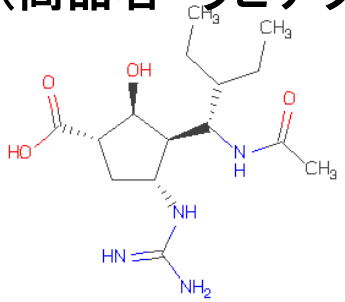
ビオタ社　(1989)

オセルタミビル
（商品名　タミフル）



ギリアド・サイエンス社　(1996)

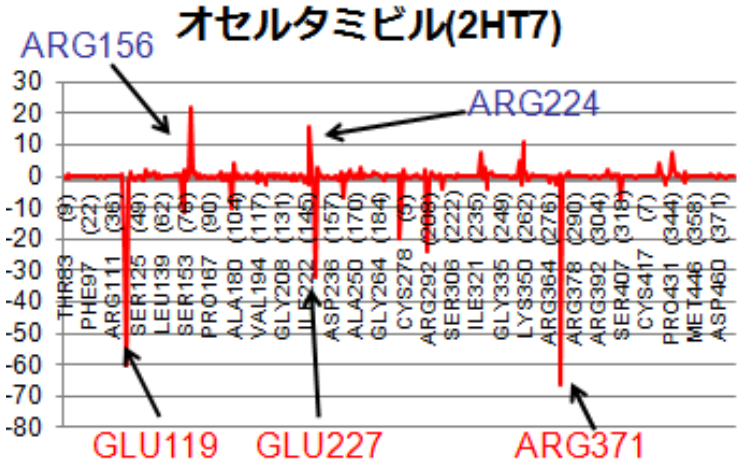
ペラミビル
（商品名　ラピアクタ）



バイオクリスト社

精密な「分子軌道法」計算の理論や、
並列計算のための技法は日本のお家芸！

しかし、あまりにも膨大な計算コストのため
従来までは創薬現場で使いにくかった。

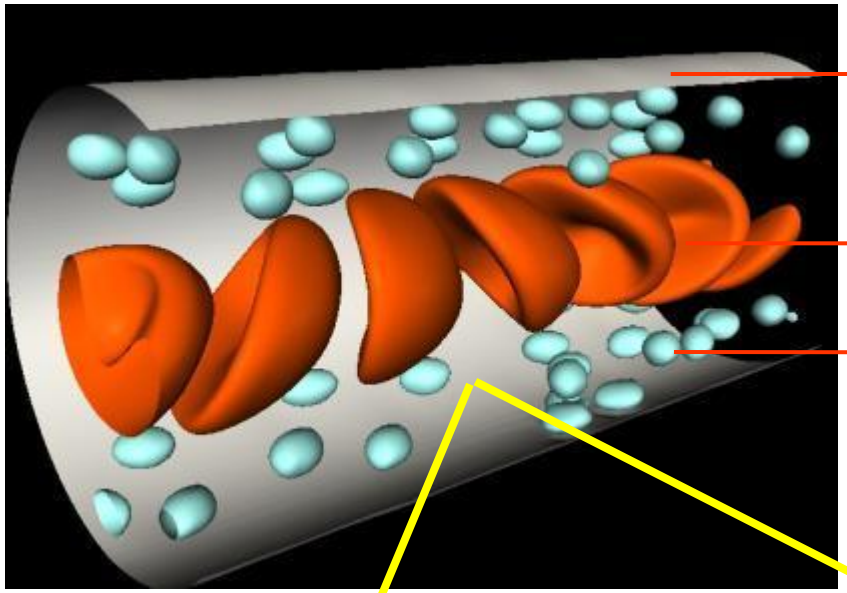


各阻害薬とノイラミニダーゼ酵素の結合予測　分子軌道法シミュレーション　（尾渡・秋山，2009）

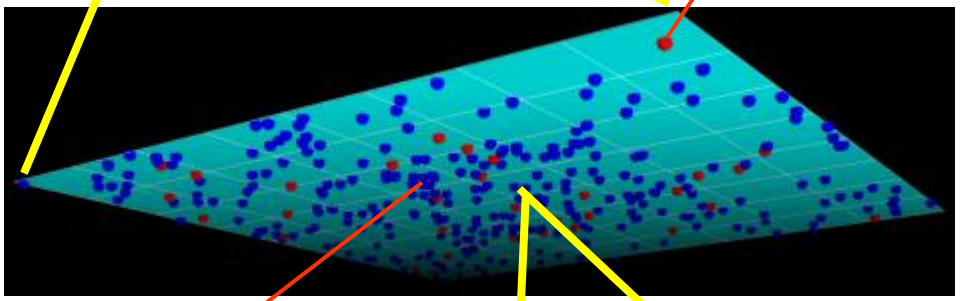
血栓症の形成過程のシミュレーション

提供:
高木周博士(理研)

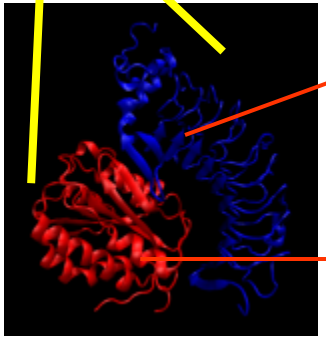
多数の赤血球や血小板を含んだ血流計算



血小板，血管壁間の分子間結合に関する動的モンテカルロ計算



リガンド-受容体分子の相互作用に関する分子動力学計算

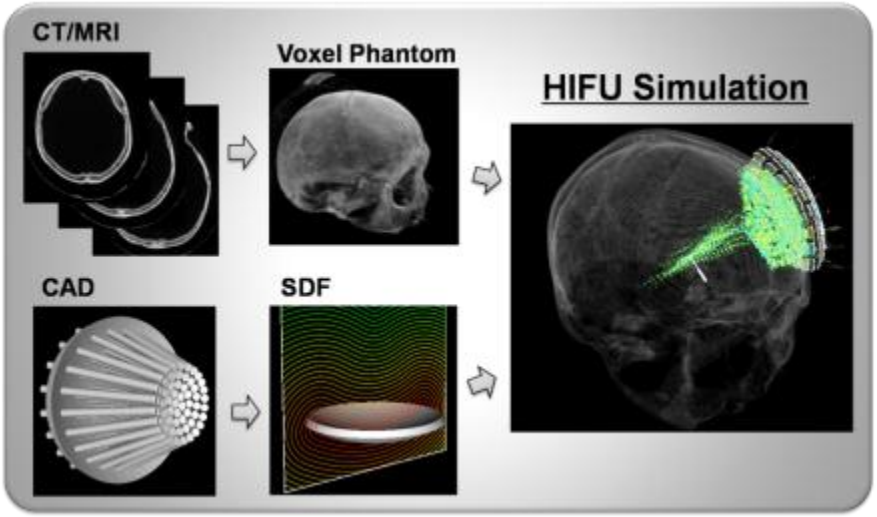


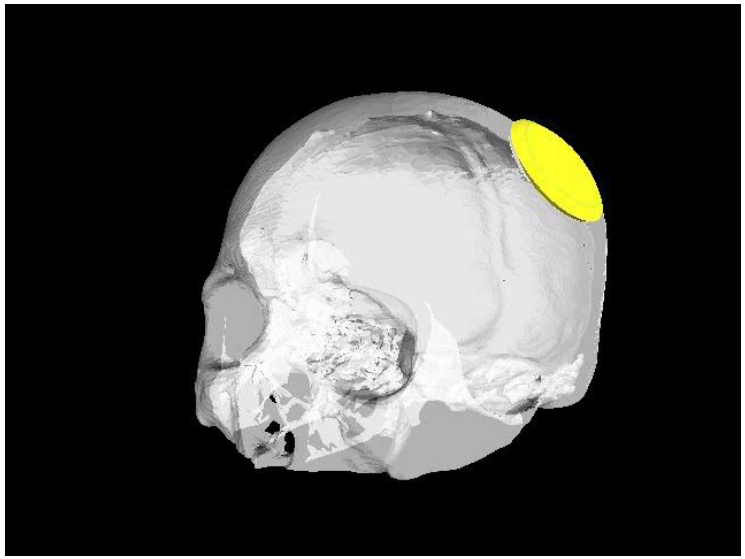
現状:**毛細血管**が対象

「**京**」では:**心筋梗塞**の対象となる直径数mm程度の**冠動脈**の計算が可能

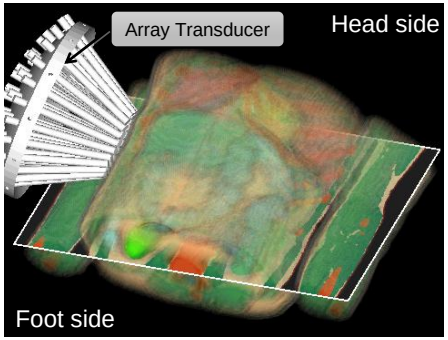
超音波治療器開発用シミュレーション

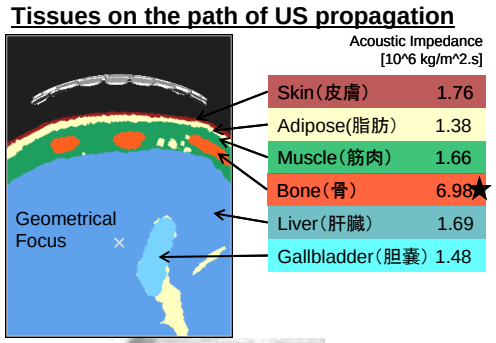
提供:
高木周博士(理研)

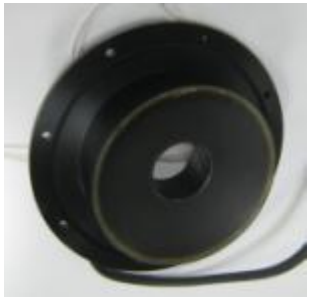




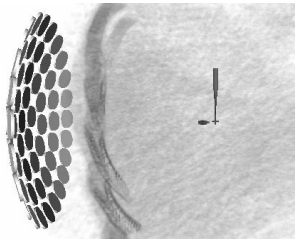
医用画像データとCADデータを直接利用するHIFUシミュレータ



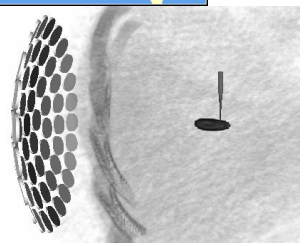




国産初
実機設計のための
詳細シミュレーション



焦点制御なし



焦点制御あり

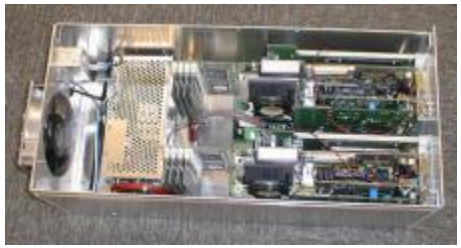
時間反転法による肝腫瘍焼灼シミュレーション

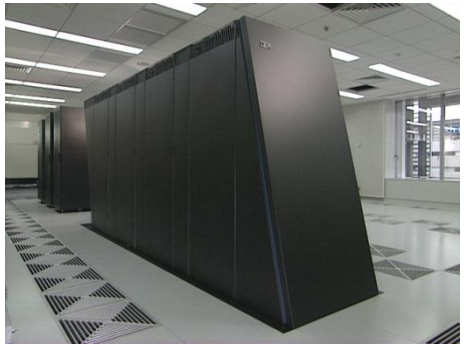
現状:低解像度・ミリ秒スケール超音波伝播シミュレーション

「京」では:高解像度・実機設計用詳細計算&秒スケール腫瘍焼灼シミュレーション

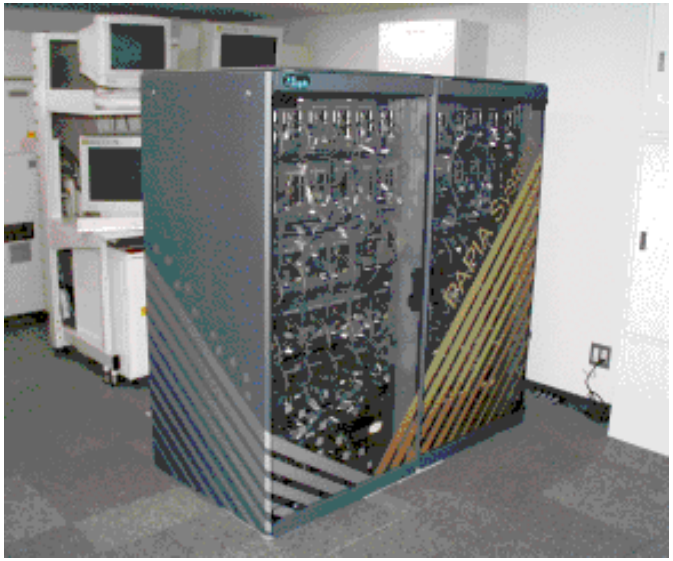
スパコンと私：

大規模バイオ計算への２０年





日本　RWCプロジェクト（1997）



愛称　PAPIAクラスタ　(Akiyama, GIW'98)

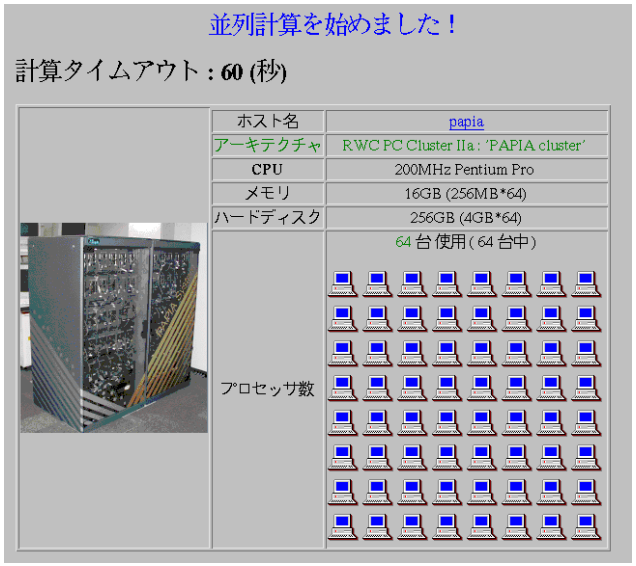
Pentium Pro 200MHz　×　64ノード

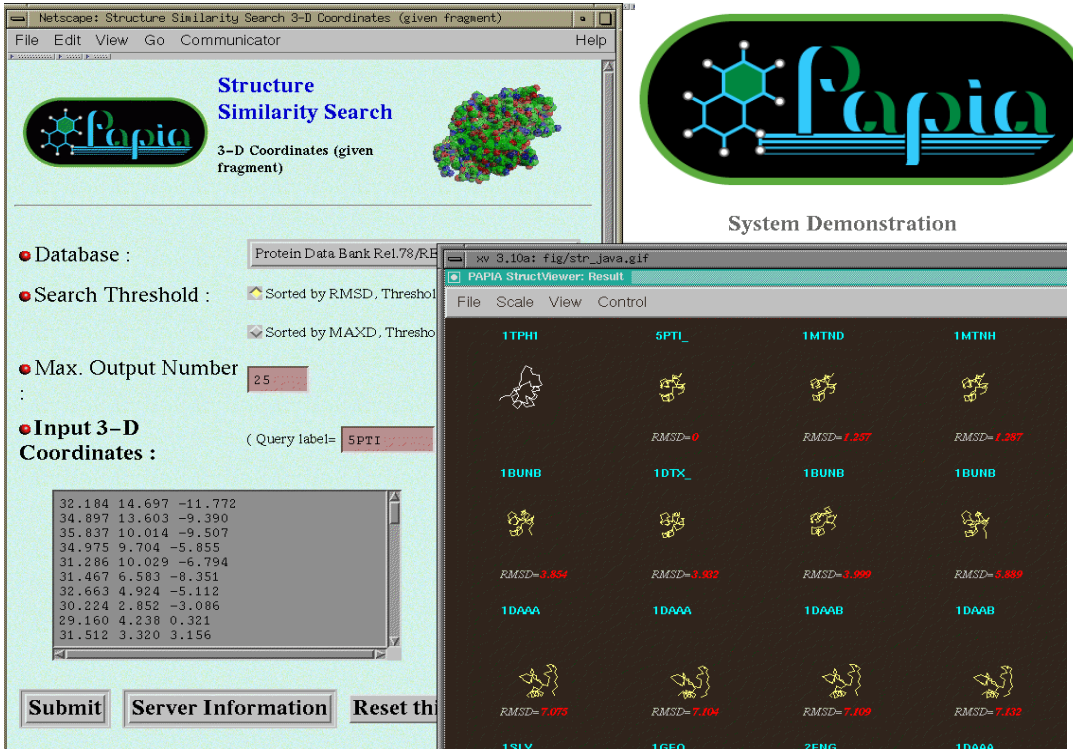
理論最大性能　12.8　GFLOPS

ローカルディスク：　4GB／ノード　　OS: NetBSD + **SCore**

世界初

インターネット上で無料の並列計算　　70カ国超から利用





System Demonstration

米国 セレラ・ジェノミクス社（1999）

私企業でありながら
国際共同研究プロジェクト
を出し抜いてヒトゲノム解読
を行ったベンチャー企業
（クレイグ・ベンター社長）

当時としては破格の計算機投資

DEC Alpha 1200CPU

理論最大性能
1.3 TFLOPS
20 TBディスク

管制センターのような司令室

フロアを埋め尽くす計算機群

日本　産総研CBRC　（2001）

愛称 **MAGIクラスタ（NEC製）**

CPU コア数:　1040

ノード数:　520

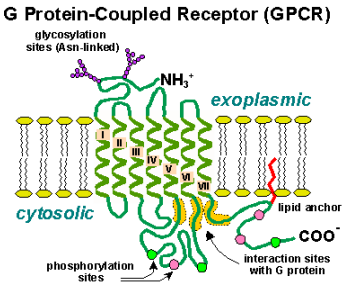
CPU:　Pentium III 933MHz，Dual

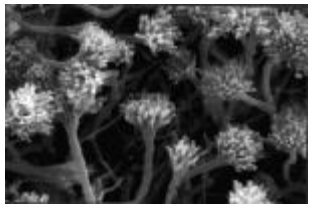
Memory:　520GB

Disk:　19,152GB

OS:　RedHat Linux 7.1

並列OS:　SCore 4.1





2001年　秋

32bit 級PCクラスタ限定では

米国NCSAを抜きLinpack世界一

653.8 GFLOPS

（ピーク性能の67%）

成果例:

単粒子画像解析、GPCR遺伝子予測、麹菌ゲノム解析

BESPA　　**SEVENS**　　**GeneDecoder**

日本　産総研CBRC　（2005）

愛称　**Blue Protein（IBM製）**

Core数:　8192

Node数:　4098

CPU:　P440改

Memory:　2TB

Disk:　20TB

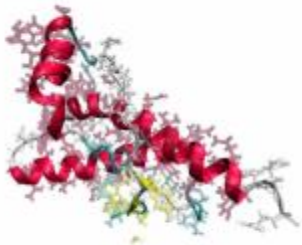
バイオ向けはこれではダメ！と交渉
米国側設計よりもI/Oノードを倍増

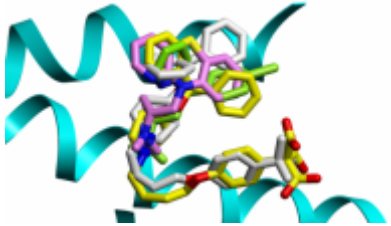
2005年　春

18.2 TFLOPS

（ピーク性能の80%）

世界総合第8位

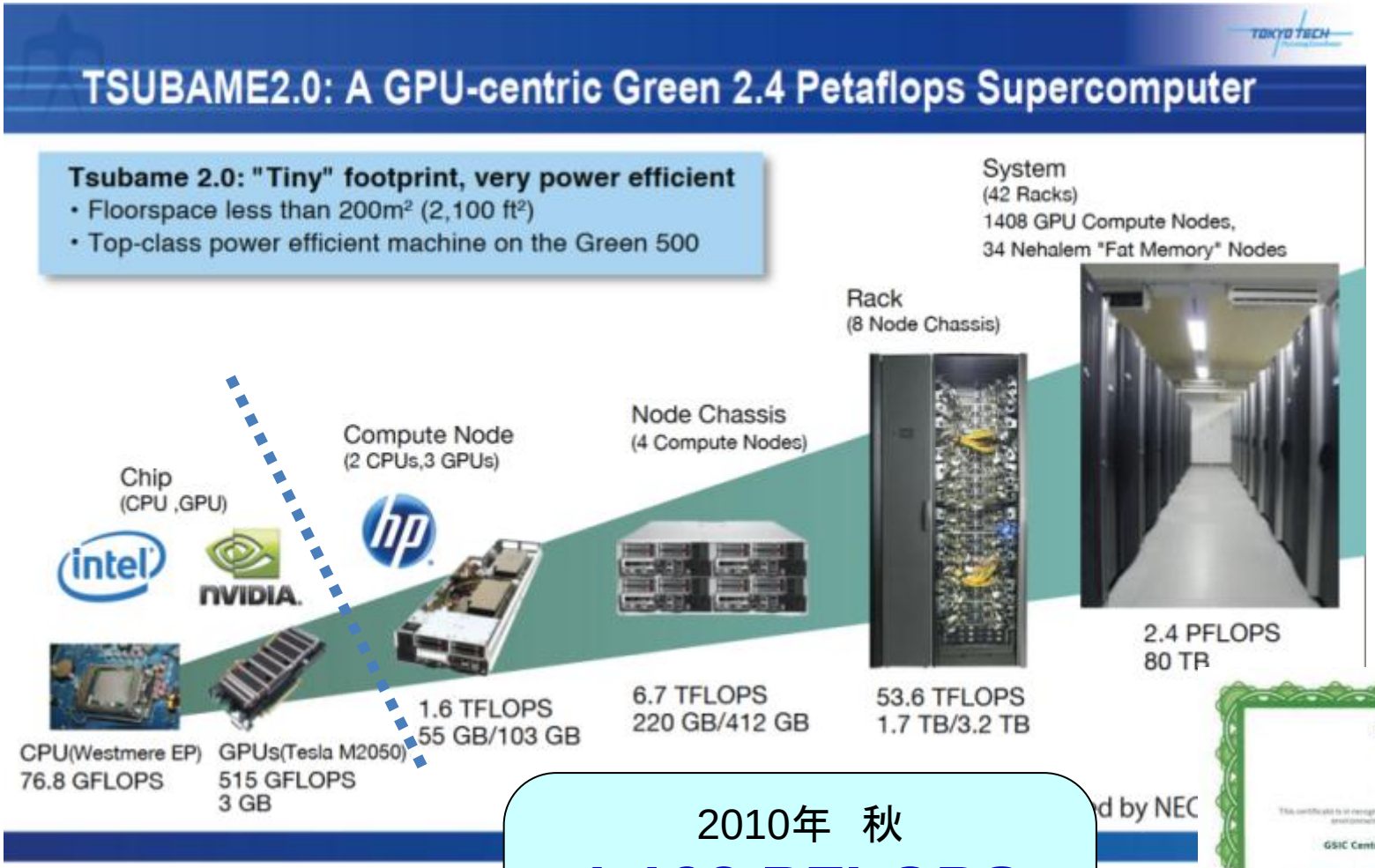




成果例:
プリオン
解析

創薬候補化合物スクリーニング　CoLBA

日本　東京工業大学　(2010)



計算ノードより上位の層は東工大が設計そのものに関与した。（単に購入したものではない）

愛称　TSUBAME2.0

（NEC／HP製）

我が国初の“ペタコン”

2010年　秋

1.192 PFLOPS

（ピーク性能の52%）

世界第4位（現在5位）

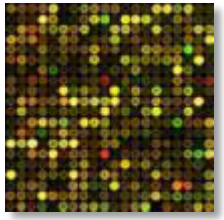
省エネ　世界第2位

生命情報解析へのハードル

- 高い計算能力
  - 膨大なデータの組み合わせ、複雑な確率計算
- メモリやストレージ（ハードディスク）も大量に必要
  - １つのジョブがTB(テラバイト）級のデータを利用する
- 高機能言語やライブラリが必要
  - 数個～数十個のアプリを複合して問題を解くことも

バイオ系研究
問題の構造が**複雑**
パソコンのクラスタから出発

高機能言語や
ライブラリを多用する



物理系研究
問題の構造は**シンプル**
手作り並列計算機から出発

アセンブリ言語による
究極の高速化

まとめ

「京」の世界一高速なハードウェアと、神戸（計算科学研究機構など）に集まる人材

写真提供：（独）理化学研究所

ゲノムの理解
病因の明解
創薬・医療