

FIT2012シンポジウム
e-サイエンス: 超大規模実問題に挑戦するアルゴリズムと計算技術
9月4日(火) 13:00~15:00
法政大学 小金井キャンパス

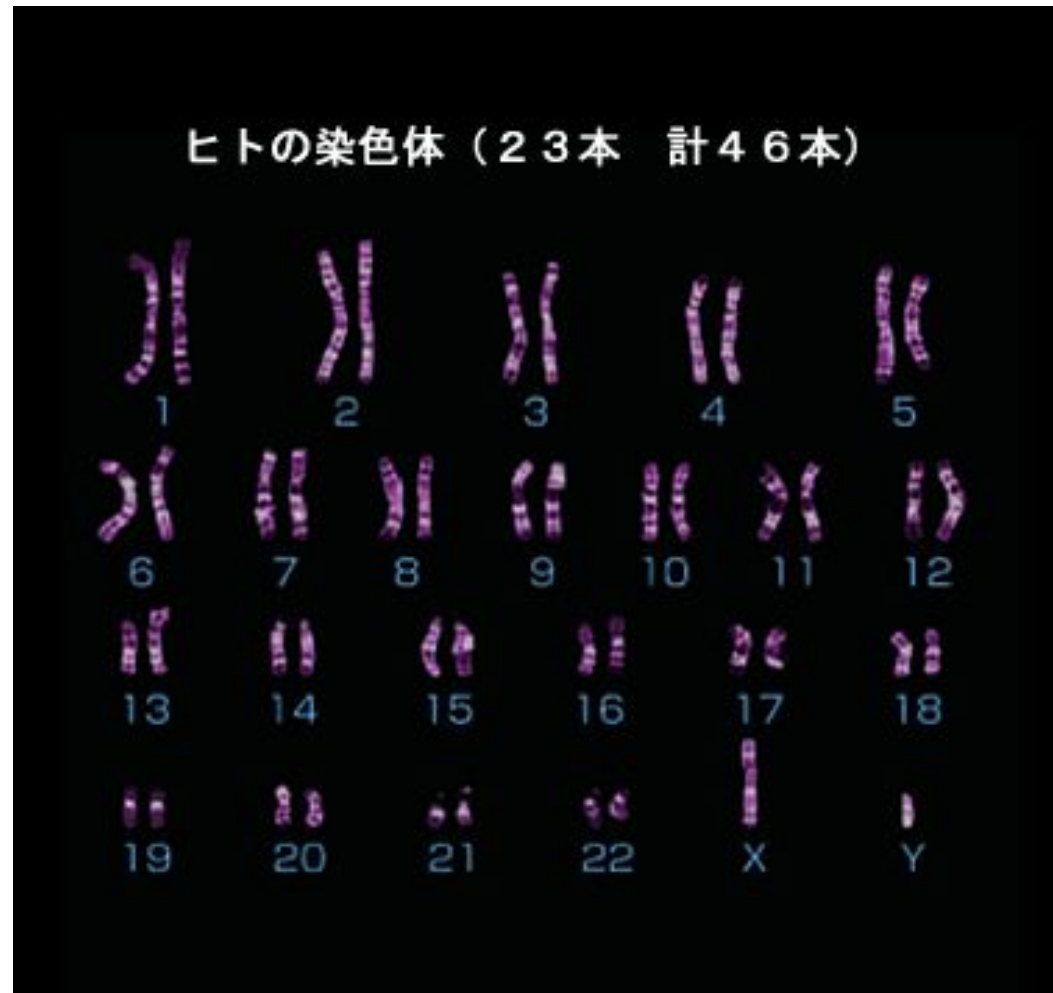
パーソナルゲノム時代に必要とされる 情報科学とスーパーコンピュータ

井元清哉

東京大学医科学研究所
ヒトゲノム解析センター
DNA情報解析分野

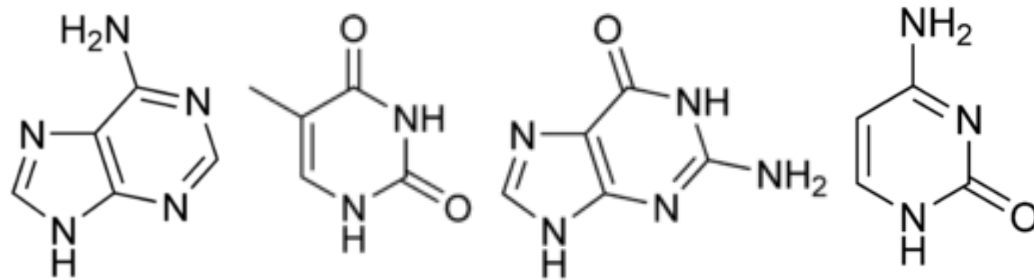
imoto@ims.u-tokyo.ac.jp
<http://bonsai.hgc.jp/~imoto>

ゲノム：生命の設計図

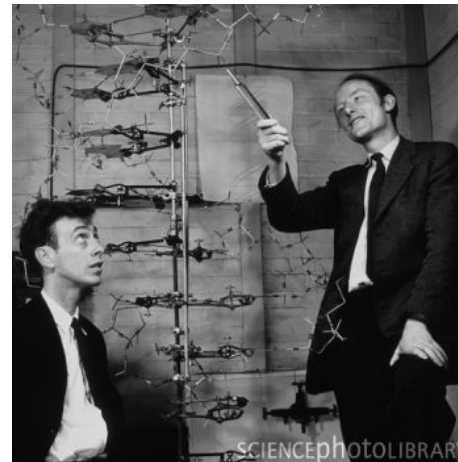
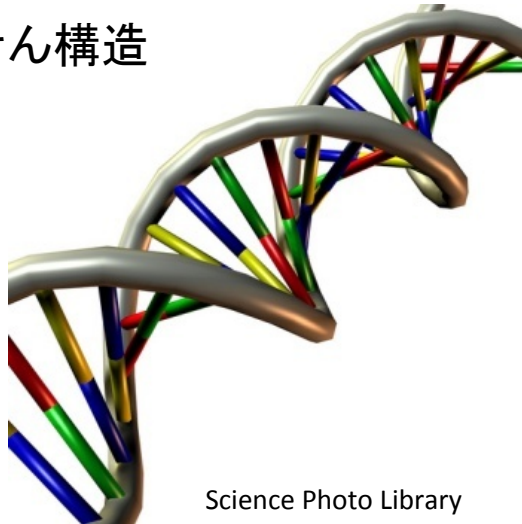


DNA

DNA を構成している塩基は、A (アデニン)、T (チミン)、G (グアニン)、C (シトシン)



二重らせん構造



Watson & Crick

ヒトゲノム
30億塩基対

(約3GB)

例: 1番染色体

塩基数: 約2億7900万
遺伝子数: 約2600個

120,001番目の塩基



CATCTGCAGGTGTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGGGTGCA
AGCTGAGCACTGGAGTGGAGTTTTCTGTGGAGAGGAGCCATGCCTAGAG
TGGGATGGGCCATTGTTCCATCTTCTGGCCCCCTGTTGTCTGCATGTAACCTT
AATACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAG
CATCAACTTCTCTCACAACCTAGGCCAGTAAGTAGTGCTTGTGCTCATCT
CCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTAC
CTGCCGTCTGCTGCCATCGGAGCCCCAAAGCCGGGCTGTGACTGCTCAGAC
CAGCCGGCTGGAGGGAGGGGCTCAGCAGGTCTGGCTTTGGCCCTGGGAGA
GCAGGTGGAAGATCAGGCAGGCCATCGCTGCCACAGAACCCAGTGGATTG
GCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGC

GCATAG



G

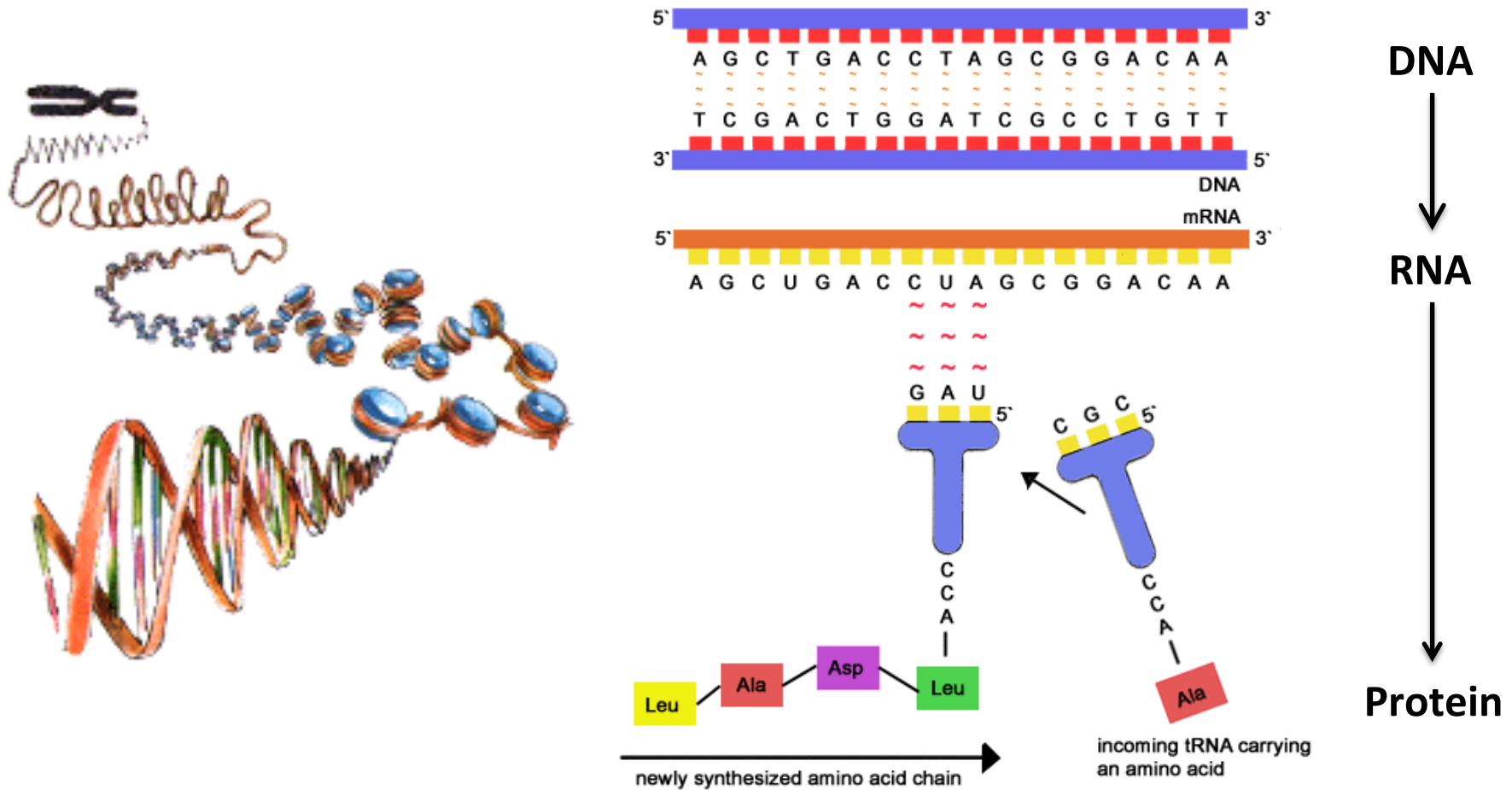
一塩基多型: SNP
集団内で1%以上の頻度
1000塩基に1つ程度

500塩基

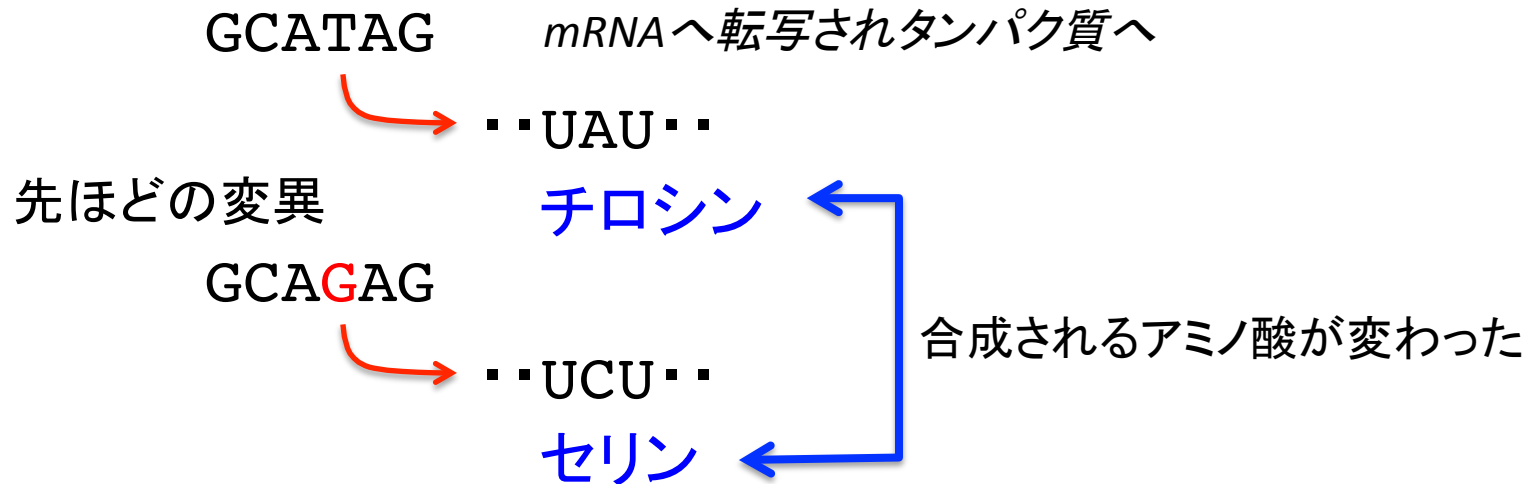


120,500

分子生物学のセントラルドグマ



DNAの変異がタンパク質を変える

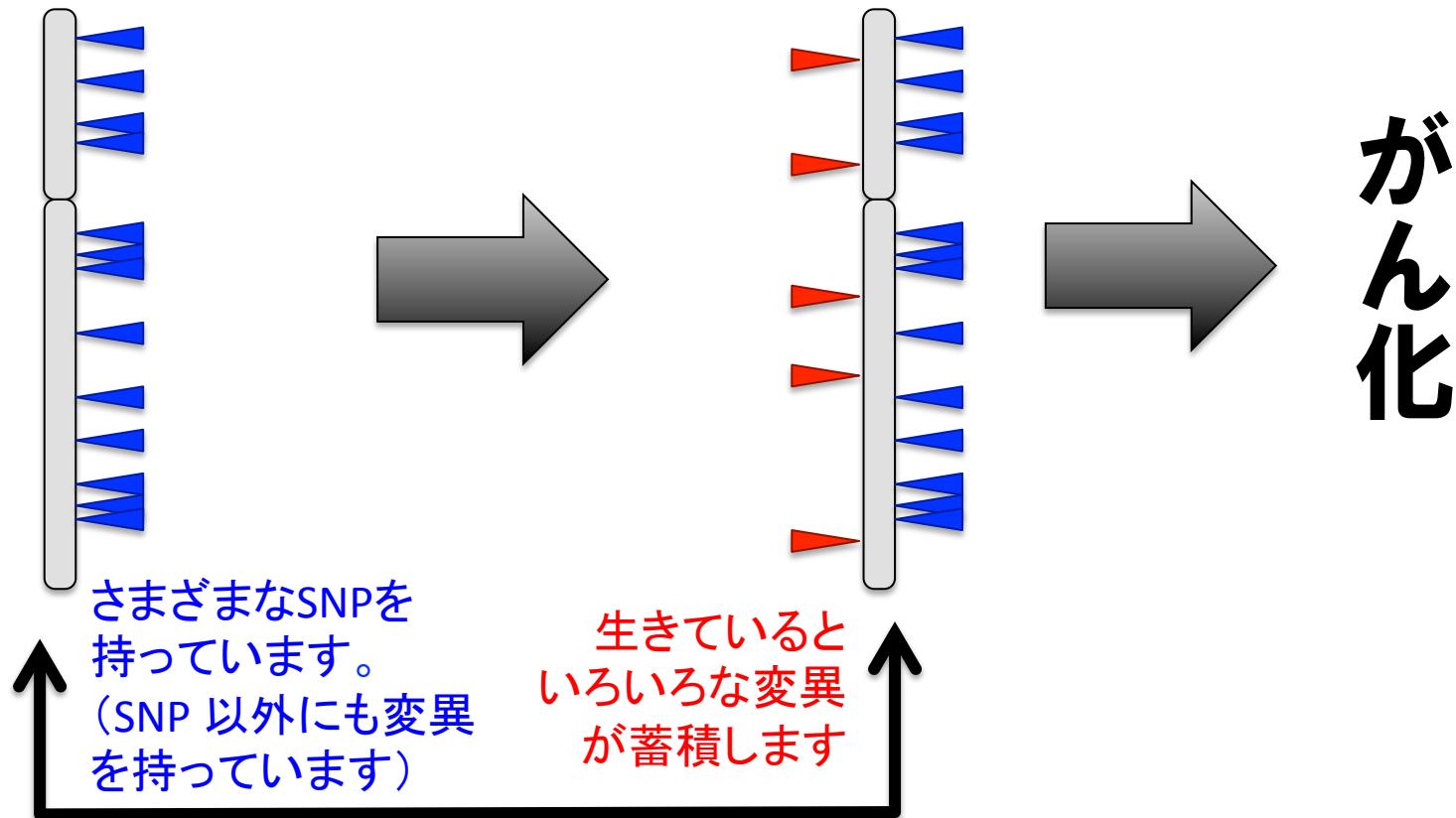


p53 遺伝子: がん抑制遺伝子の一つ。
細胞が「がん化」するとアポトーシスを起こさせる。

がんにおいて最も高頻度に異常(変異)が認められる。

がんゲノム

Aさんのゲノム



比較すればがんの原因となった変異が分かる
個人のゲノムを読むことが必要！！

国際ヒトゲノム計画

- 2003年に30億文字からなるゲノム情報の電子化を完了しました。
- 「ヒトゲノムが解読された」と世界中で報道されましたが、ヒトゲノムという暗号が解読されたわけではありません。
- 10年、1000億円かかりました。

今はどうでしょう？ 「私のゲノム」の費用

国際ヒトゲノム計画(～2003年)

10年、1000億円

次世代シーケンサーの登場により

現在 10日、50万円

システム / HiSeqシステム / **HiSeq 2500/1500**Subscribe |   | Follow us: [トップページ](#)[システム](#)[アプリケーション](#)[研究者の声](#)[仕様](#)[ワークフロー](#)[テクノロジー](#)[お問い合わせ](#)**HiSeq 2500****HiSeq 1500**

ランモード	標準ラン	Fastラン*	標準ラン	Fastラン*
データ産出量 (2 ×100bp)	600Gb	120Gb	300Gb	60Gb
ラン時間	約11日	約27時間	約8.5日	約27時間
クラスター形成	cBot	内部搭載	cBot	内部搭載
ペアエンドリード数	60億	12億	30億	6億
シングルリード数	30億	6億	15億	3億
最大リード長**	2 ×100bp	2 ×150bp	2 ×100bp	2 ×150bp
Q30以上の塩基割合***	> 85% (2 ×50bp) > 80% (2 ×100bp)			

データ処理

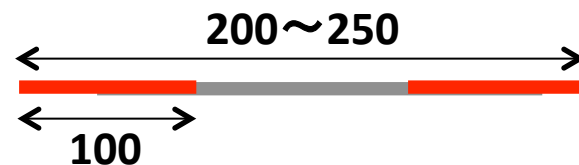
- キーワード

x40

シーケンサーで読んだDNA配列にはエラーが含まれます。
そこで、40コピー分 ($30\text{GB} \times 40 = 120\text{GB}$) 程度読みます。

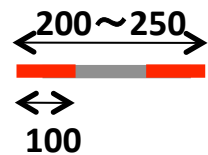
ショートリード (Short Read)

先ほどの HiSeq では、染色体をつくる塩基がずらっと長〜く一度に読めるわけではありません。実際は、200~250bp 程度の長さに裁断された DNA の両端100 b ずつが読めます。



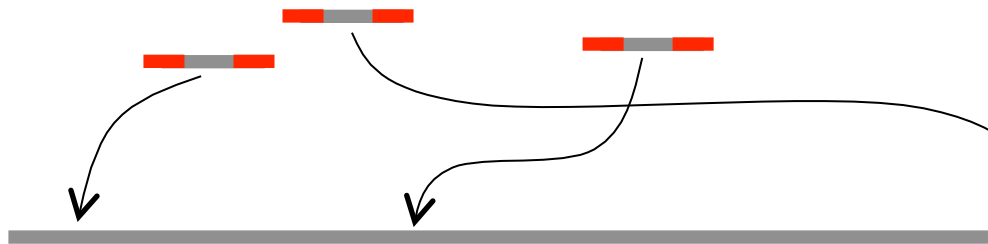
データ処理

1 run 当たり数十億 read (100bp) の断片が出力される



(Pair-End で読んだ場合)

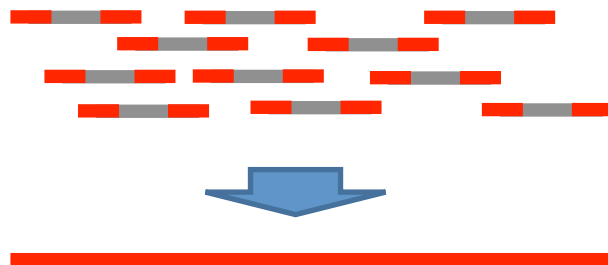
ヒトゲノムリファレンス配列へのマッピング



100文字をエラーを考慮しつつ
30億塩基対のヒトゲノムの
どこにマップできるのかサーチ

数十億ピースのジグソーパズルを
出来上がりの絵の上で作る作業
出来上がりに無いピースもある

マッピングできなかった配列を de novo アセンブリ



正常組織ゲノムにはないがんゲノム特有
の配列を同定

数万ピースのジグソーパズルをどんな
絵が出来上がるのか、それともなにも
出来ないのかも分からないまま作る作業

東大医科研HGCスーパーコンピュータシステム

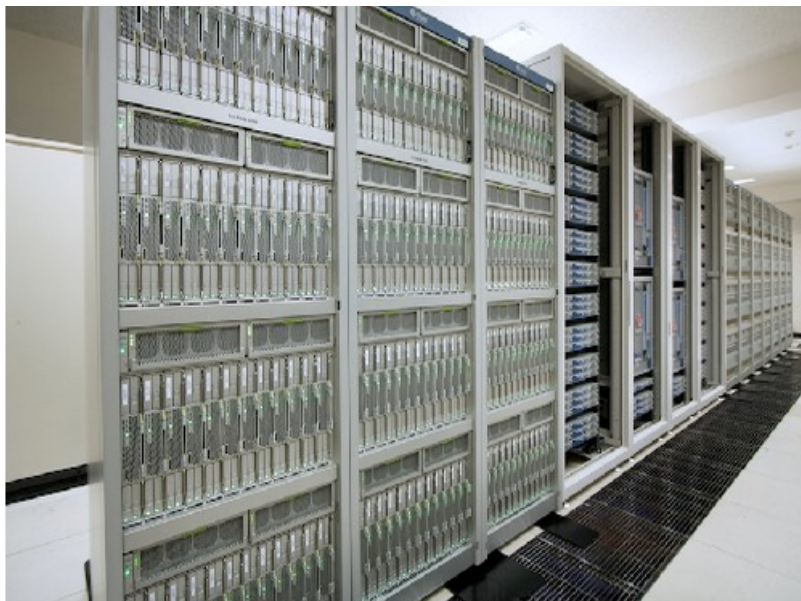
計 155 T Flops (LINPACK)、3 PB (Lustre システム)

● Shirokane1 性能 (2009年1月 運用開始)

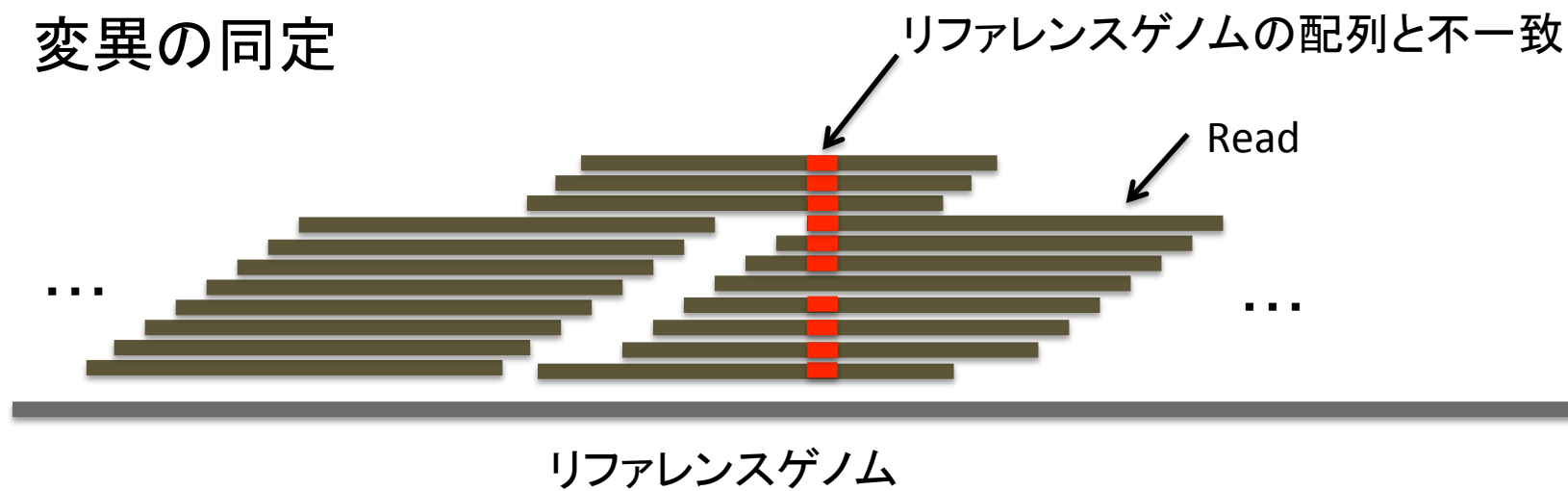
構成名	機種名	Rpeak (Gflops)	CPU	OS	CPU コア数	メモリ /node	ディス ク容 量
分散メモ リ型 Type-A	Sun Fire X4440	75,494	AMD Quad Core Opteron 8356 (2.3GHz)	Cent OS	192	128GB	-
分散メモ リ型 Type-B	Sun Blade X6250		Intel Quad Core Xeon E5450 (3.0GHz)		6,144	32GB	-
共有メモ リ型 (2011年 12月 まで稼働)	SGI Altix 4700	849	Intel Itanium (1.66GHz)	SUSE Linux	128	2048GB	58TB
ディス クア レイ	-	-	-	-	-	-	1PB

● Shirokane2 性能 (2012年1月 運用開始)

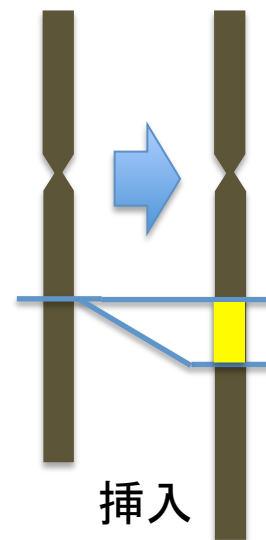
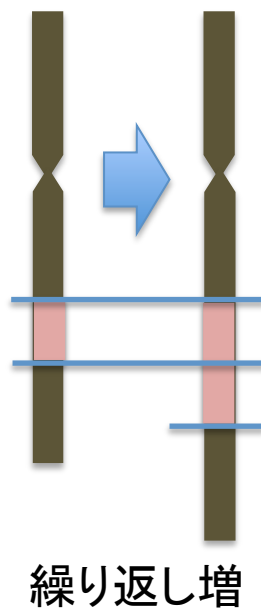
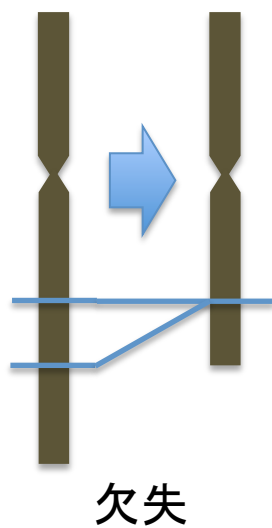
構成名	機種名	Rpeak (Gflops)	CPU	OS	CPU コア数	メモリ /node	ディス ク容 量
分散メモ リ型 Type-A	Hitachi HA8000 /RS210	150,141	Intel Xeon X5675 (3.06GHz)	Red Hat Enterprise Linux	144	144GB	-
分散メモ リ型 Type-B	Hitachi HA8000- tc /HT225		AMD Opteron 6276 (2.3GHz)		16,128	32GB	-
共有メモ リ型	SGI Altix UV 100	1,361	Intel Xeon E7 8837 (2.66GHz)		128	2TB	58TB
ディス クア レイ	-	-	-	-	-	-	2PB



変異の同定



さまざまな変異



実例

がん患者 A さん

がん組織

正常組織(末梢血)

シーケンサーから
出てくるファイルサイズ

約600GB

約450GB

リファレンスゲノムへ
マッピング、BAM作成

両方で33時間

シングルコアだと
600時間

使用したコア数

最大1016コア

BAMファイル

160GB

110GB

x40

x30

一人分
1.3TB



千人分
1.3PB

変異数

全ゲノム

タンパクコード領域

がん組織 vs 正常組織

8万7千

1400

正常組織 vs Reference Genome

166万

2万4千

統計
解析へ

2013年、「私」のDNAシーケンスを 10万円で1時間以内で得られる時代が始まる

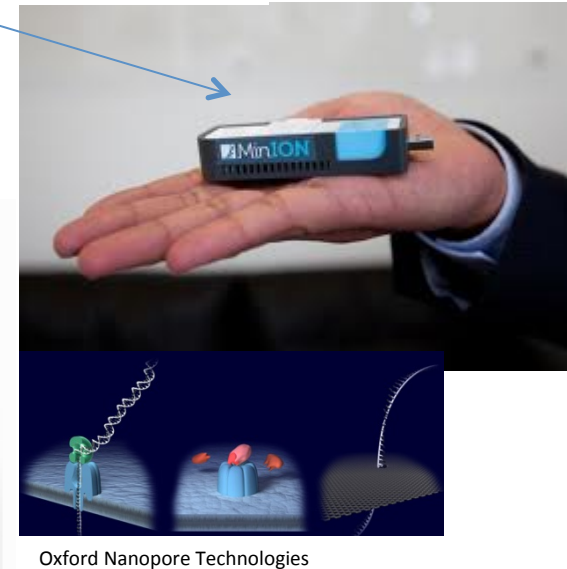
参考: ヒトゲノム計画は1000億円、13年を費やした

だれもが自分のゲノム情報
を利用できる時代が到来

シリコンシーケンサーの登場

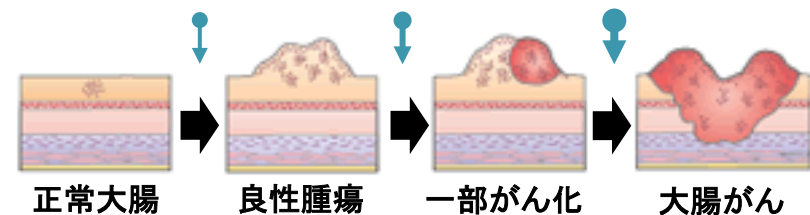


State-of-the-art electronics
Integrated reagent delivery



さらに、エピゲノム、トランスクリプトームもこの装置で解析

個人個人のがんのシステムの
統合理解が大きく進む



まとめ

- 次世代シーケンサーの発展に伴い、パーソナルゲノム時代の幕が開けた
- ただし、データ量は膨大であり、その解析には、スーパーコンピュータの利用が不可欠である
- また、生物学者や治療にあたる医学者との綿密な連携が必要
- ゲノム医科学への貢献がなければ「遊びの研究」と自らを律してやっています