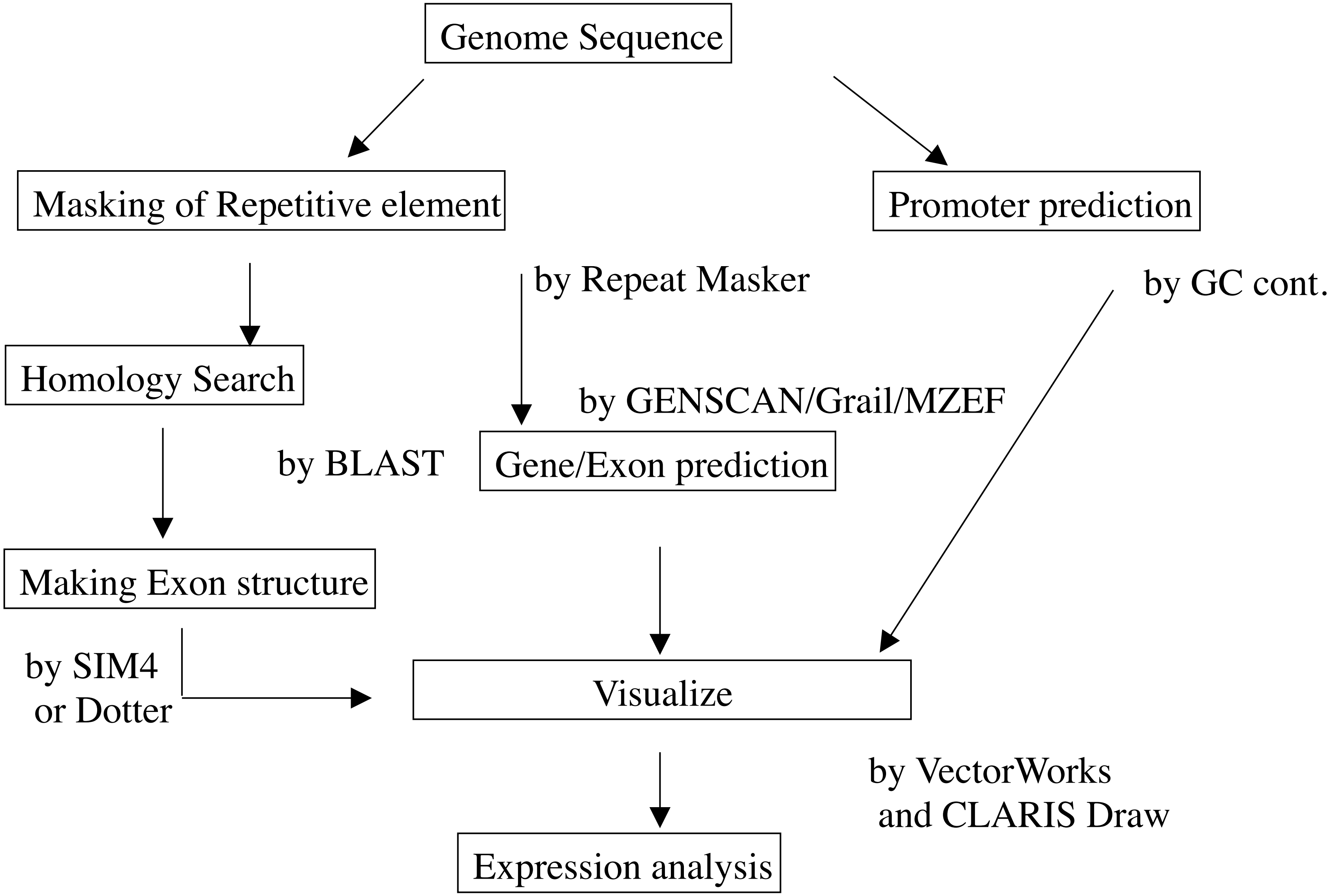
ゲノムアノテーション

ゲノムアノテーションフロー（2002）

Flow chart of gene finding in KEIO



Genome Sequence

Masking of Repetitive element

Promoter prediction

by Repeat Masker

by GC cont.

Homology Search

by GENSCAN/Grail/MZEF

by BLAST

Gene/Exon prediction

Making Exon structure

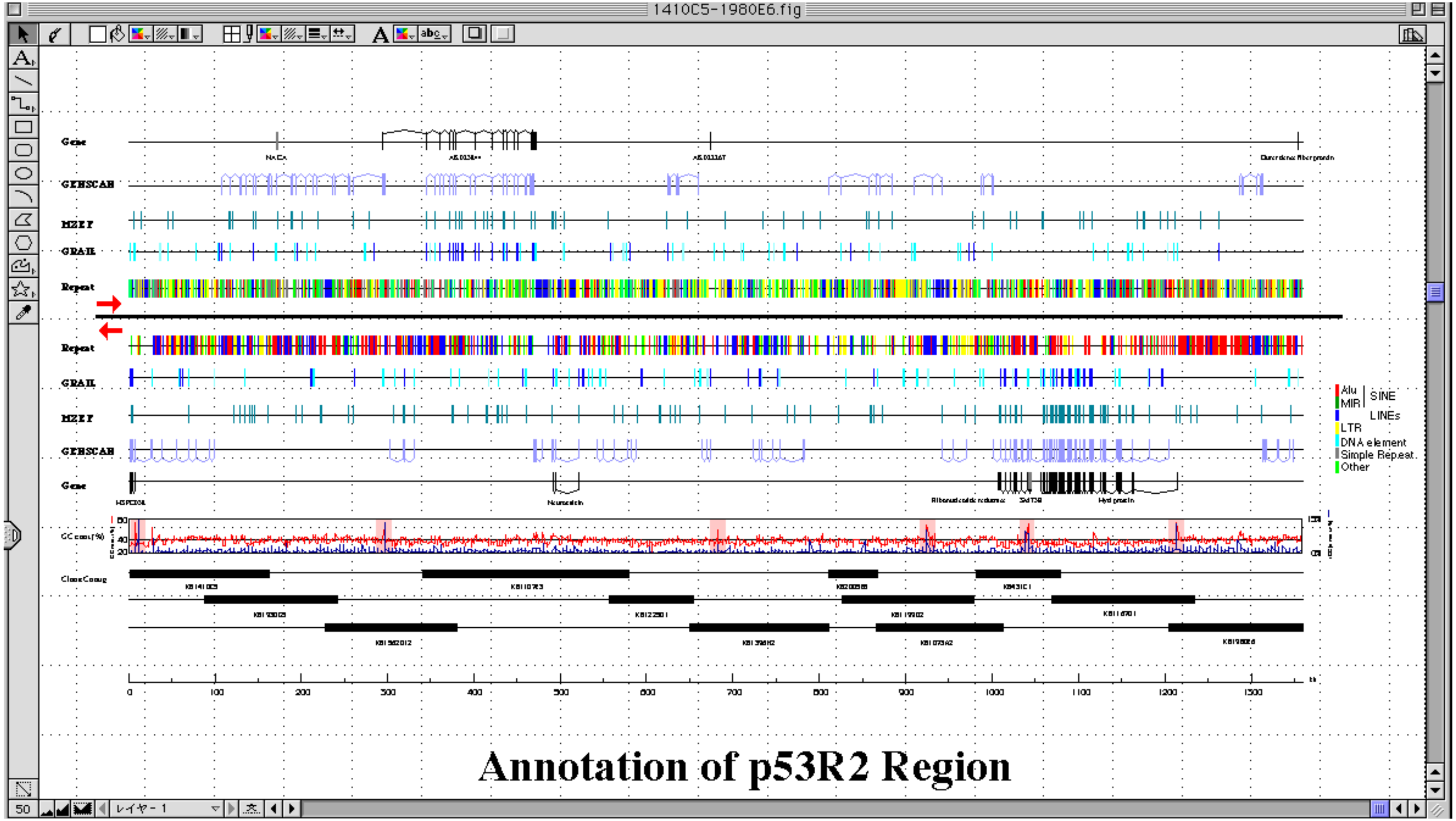
by SIM4
or Dotter

Visualize

by VectorWorks
and CLARIS Draw

Expression analysis

ゲノムアノテーション（2001）

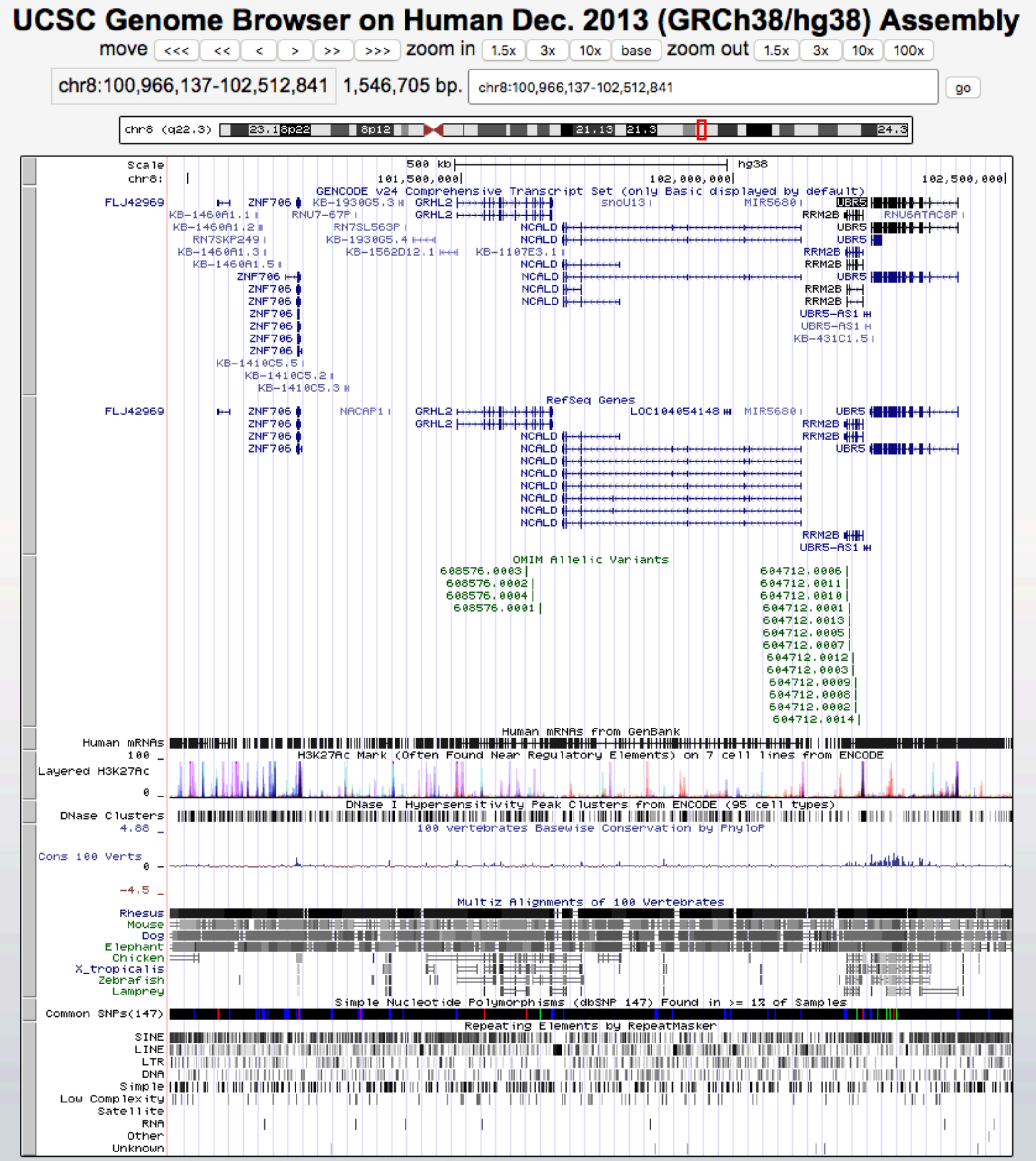


1410C5-1980E6.fig

Gene
GENSCAN
MZEF
GRAIL
Repeat
Repeat
GRAIL
MZEF
GENSCAN
Gene
Alu
MIR
SINE
LINEs
LTR
DNA element
Simple Repeat
Other
GC cont(%)
60
40
20
CloneContig
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
kb
50
レイヤー1

Annotation of p53R2 Region

ゲノムアノテーション（2016）



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

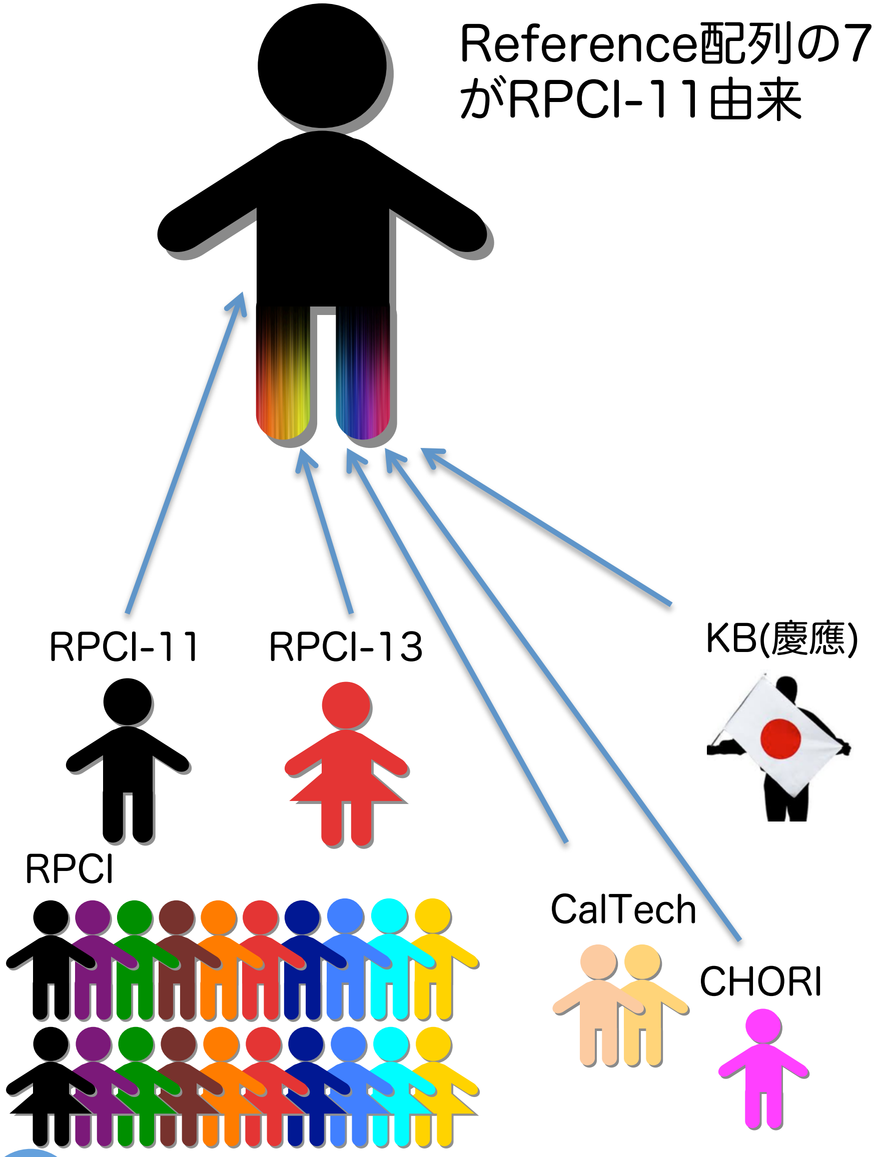
chr8:100,966,137-102,512,841 1,546,705 bp.

move <<< << < > >> >>>
zoom in 1.5x 3x 10x base
zoom out 1.5x 3x 10x 100x
chr8:100,966,137-102,512,841
go
chr8 (q22.3)
Scale
chr8:
500 kb
hg38
101,500,000
102,000,000
102,500,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
FLJ42969
ZNF706
GRHL2
NCALD
UBR5
RRM2B
UBR5-AS1
RefSeq Genes
OMIM Allelic Variants
Human mRNAs from GenBank
Human mRNAs
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
DNase Clusters
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Common SNPs(147)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

https://genome.ucsc.edu/ Oct.5 2016

個人ゲノム研究

ヒトゲノム計画では誰の配列を決定したのか


RPCI-11
RPCI-13
KB(慶應)
RPCI
CalTech
CHORI

Reference配列の72%程度がRPCI-11由来

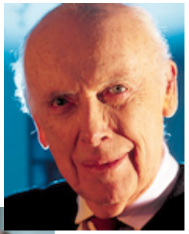
1997年3月23日にNY州バッファロー市の新聞に広告を出し、10名の男性と10名の女性のボランティアを選出した。50mLの血液サンプルを採取し、男性２名、女性２名からDNAライブラリーを作成した。匿名男性１名のライブラリーの質が一番良かったため、RP11がHGPで最も利用されたクローンとなった。

一方、セレラ社のプロジェクトではベンターを含む21名のボランティアから５名の検体を選別して全ゲノムシークエンシングを行った。

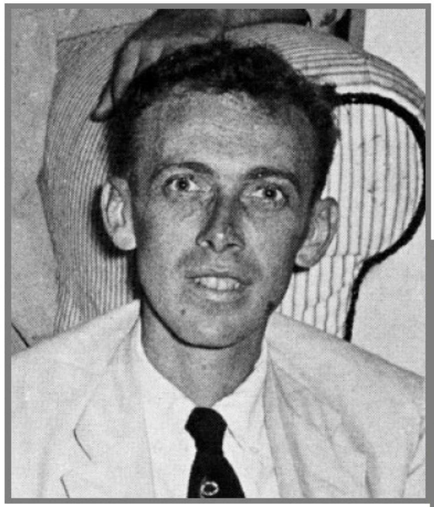
Genome Res. **11**:483-496 (2001)

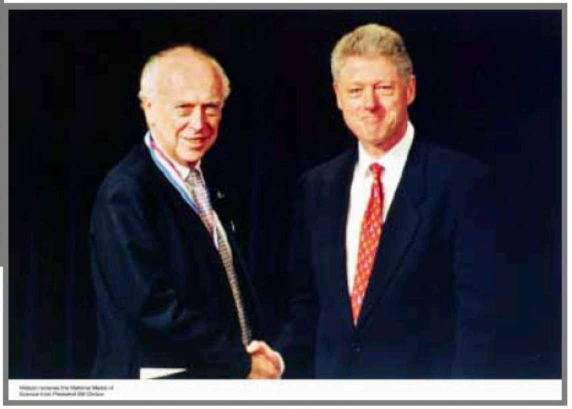
パーソナルゲノムの歴史

2007 May	ワトソン個人ゲノム公開 (by 454) → 論文は2008 Jun
2007 Sep	ベンター個人ゲノム解読論文発表 (by ABI 3730xl)
2008 Jun	1000ゲノム計画スタート
2008 Nov	中国人、アフリカ人個人ゲノム解読 (by GA)
2008 Dec	韓国人個人ゲノム解読 (by GA)
2009 Nov	Complete Genomics 社が$5,000で全ゲノム受託解析を開始
2010 Feb	古代人ゲノム解読 Palaeo-Eskimo (4,000 year-old) (by GAIIx)
2010 May	ネアンデルタール人ゲノム解読 (38,000 year-old) (by 454 with GA)
2010 Jun	1000ゲノム計画パイロット実験データ公開 (by 454 and GA and SOLiD)
2010 Oct	日本人個人ゲノム解読 (by GAIIx)
2011 Nov	23andME 社が $999 で全エキソン受託解析 (by GAIIx)
2012 Feb	古代人ゲノム解読 Tyrolean Iceman (5,000 year-old) (by GAIIx)
2013 Nov	日本人1000人ゲノム解読 (by HiSeq2500)



James D. Watson (1928-)





The 1st personal genome sequencing

Jun. 1, 2007
$1,000,000

nature   Vol 452|17 April 2008|**doi:10.1038/nature06884**

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

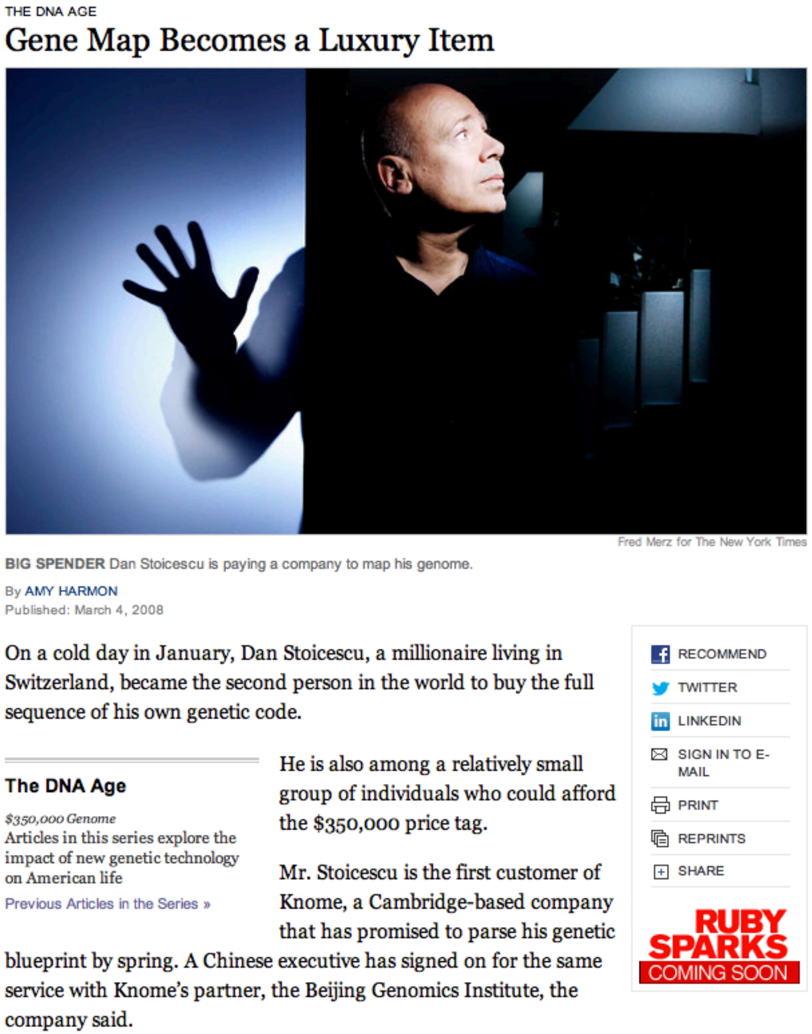
Table 3 | SNPs matching HGMD mutations causing disease or other phenotypes

HGMD accession	Chromosome	Coordinate	HUGO symbol	Gene name	Cytogenetic	Phenotype	Zygosity
CM003589	1	97937679	DPYD	Dihydropyrimidine dehydrogenase	1q22	Dihydropyrimidine dehydrogenase deficiency	Heterozygous
CM950484	1	157441978	FY	Duffy blood-group antigen	1q	Duffy blood group antigen, absence	Homozygous*
CM942034	4	619702	PDE6B	Phosphodiesterase 6B, cGMP-specific, rod, beta	4p16.3	Retinitis pigmentosa 40	Heterozygous
CM021718	9	36208221	GNE	UDP-N-acetylglucosamine 2-epimerase	9p	Myopathy, distal, with rimmed vacuoles	Heterozygous
CM980633	10	50348375	ERCC6	Excision repair cross-complementing rodent repair deficiency, complementation group 6 protein (CSB)	10q	Cockayne syndrome	Homozygous†
CM050716	11	76531431	MYO7A	Myosin VIIA	11q13.5	Usher syndrome 1b	Homozygous†
CM950928	12	46812979	PFKM	Phosphofructokinase, muscle	12q13.3	Glycogen storage disease 7	Homozygous*
CM032029	14	20859880	RPGRIP1	Retinitis pigmentosa GTPase regulator interacting protein 1	14q11	Cone–rod dystrophy	Heterozygous
CM984025	19	18047618	IL12RB1	Interleukin-12 receptor, beta 1	19p13.1	Mycobacterial infection	Heterozygous
CM024138	19	41014441	NPHS1	Nephrosis-1, congenital, Finnish type	19q	Congenital nephrotic syndrome, Finnish type	Heterozygous
CM910052	22	49410905	ARSA	Arylsulphatase A	22q	Metachromatic leukodystrophy	Heterozygous

* Coverage at these SNP positions is less than 5. However, both produce benign phenotypes.
† Coverage at these SNP positions is greater than 5. Both would produce severe phenotypes if they were truly homozygous.

Nature **452**:872-6 Jun (2008)

Dan Stoicescu



THE DNA AGE

Gene Map Becomes a Luxury Item

Fred Merz for The New York Times

BIG SPENDER Dan Stoicescu is paying a company to map his genome.

By AMY HARMON
Published: March 4, 2008

On a cold day in January, Dan Stoicescu, a millionaire living in Switzerland, became the second person in the world to buy the full sequence of his own genetic code.

He is also among a relatively small group of individuals who could afford the $350,000 price tag.

Mr. Stoicescu is the first customer of Knome, a Cambridge-based company that has promised to parse his genetic blueprint by spring. A Chinese executive has signed on for the same service with Knome's partner, the Beijing Genomics Institute, the company said.

The DNA Age
$350,000 Genome
Articles in this series explore the impact of new genetic technology on American life
Previous Articles in the Series »

RUBY SPARKS
COMING SOON

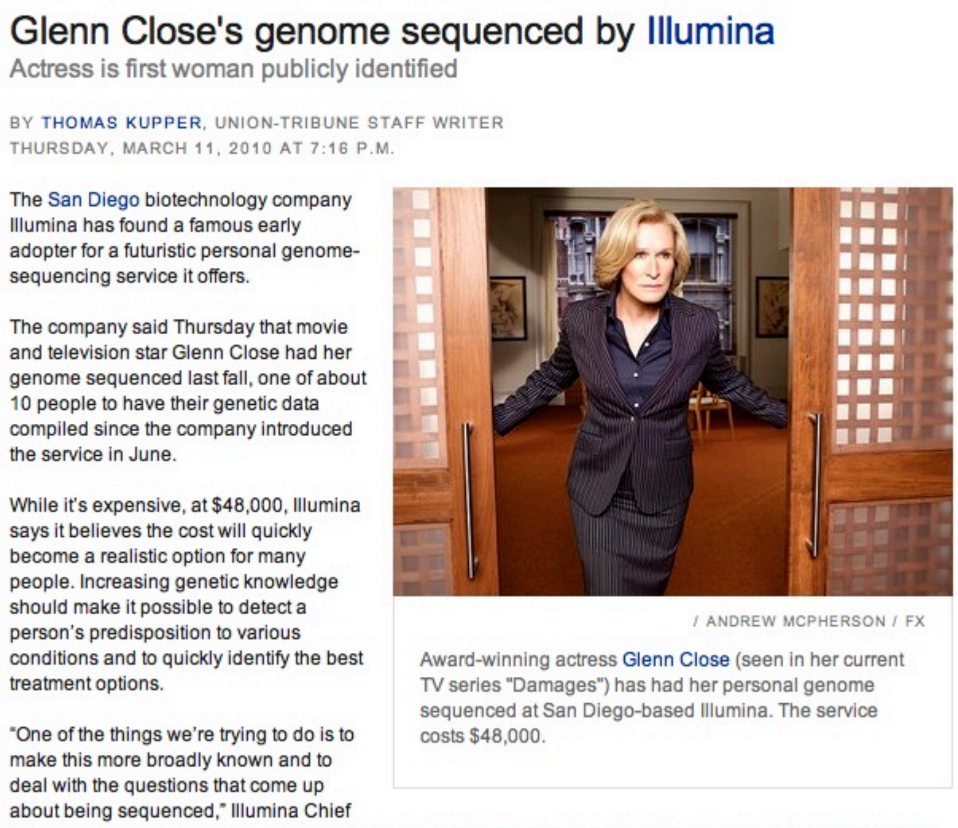
The 1st Customer; Dan Stoicescu

Mar. 4, 2008
$350,000

Bentley

410 uvp (1985)

Glenn Close



Glenn Close's genome sequenced by Illumina

Actress is first woman publicly identified

BY THOMAS KUPPER, UNION-TRIBUNE STAFF WRITER
THURSDAY, MARCH 11, 2010 AT 7:16 P.M.

The San Diego biotechnology company Illumina has found a famous early adopter for a futuristic personal genome-sequencing service it offers.

The company said Thursday that movie and television star Glenn Close had her genome sequenced last fall, one of about 10 people to have their genetic data compiled since the company introduced the service in June.

While it's expensive, at $48,000, Illumina says it believes the cost will quickly become a realistic option for many people. Increasing genetic knowledge should make it possible to detect a person's predisposition to various conditions and to quickly identify the best treatment options.

"One of the things we're trying to do is to make this more broadly known and to deal with the questions that come up about being sequenced," Illumina Chief Executive Jay Flatley said. "The more we can bring prominence and visibility to this, the more people will talk about it."

/ ANDREW MCPHERSON / FX

Award-winning actress Glenn Close (seen in her current TV series "Damages") has had her personal genome sequenced at San Diego-based Illumina. The service costs $48,000.

The 1st Customer service for woman

Mar. 11, 2010
$48,000

Benz C-class

MOONEY M20F (1967)

MICHAEL DOUGLAS
GLENN CLOSE
Fatal Attraction
ANNE ARCHER

Copyright © 1987 by Paramount Pictures.

GLENN CLOSE
Disney's
101 Dalmatians
GLENN CLOSE
Disney's
102 DALMATIANS
Meet Two Unlikely Heroes With A Bone To Pick.

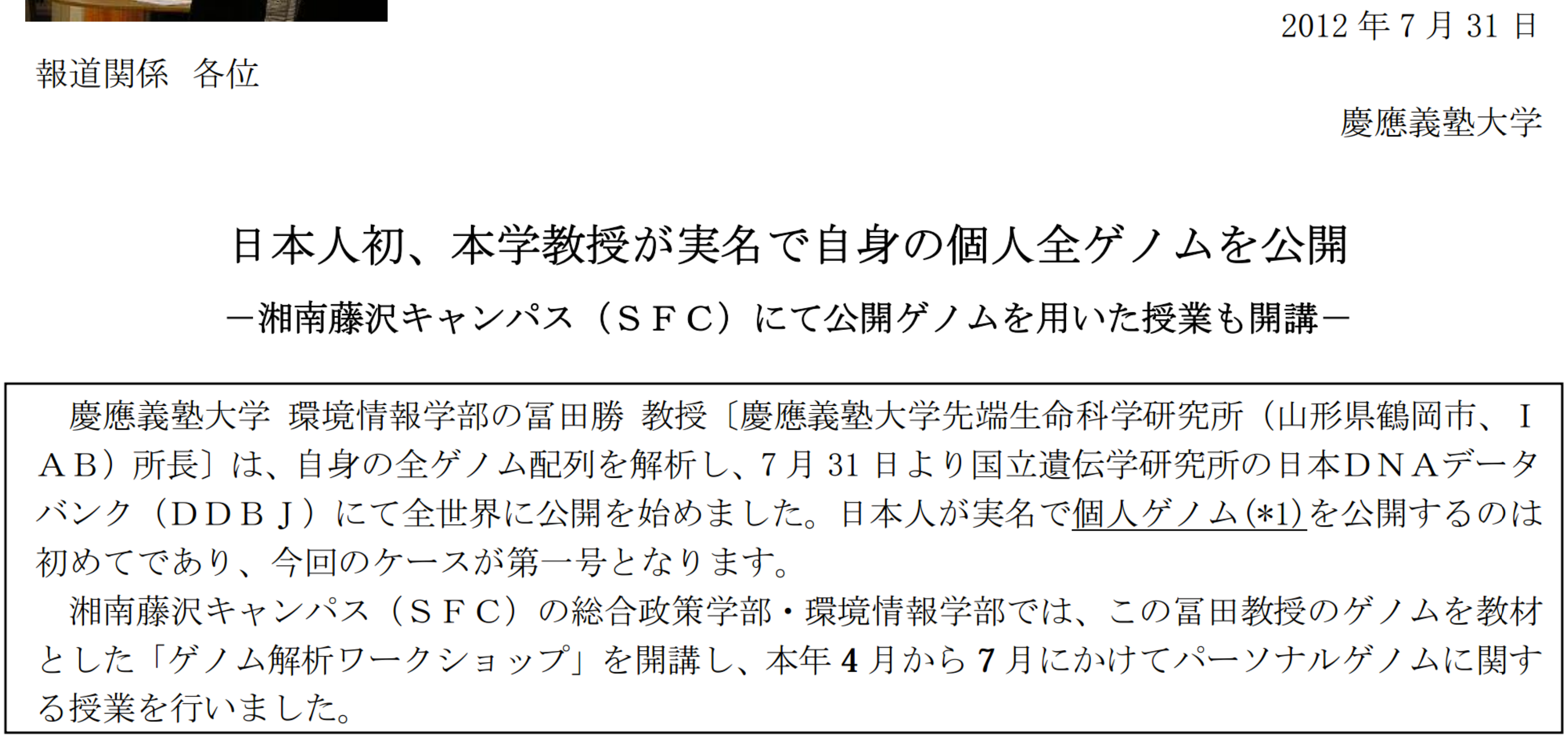
Copyright © Walt Disney Studios Motion Pictures

Masaru Tomita

The 1st Customer service for Japanese

Jul. 31, 2012

$5,000



2012 年 7 月 31 日

報道関係　各位

慶應義塾大学

日本人初、本学教授が実名で自身の個人全ゲノムを公開

－湘南藤沢キャンパス（ＳＦＣ）にて公開ゲノムを用いた授業も開講－

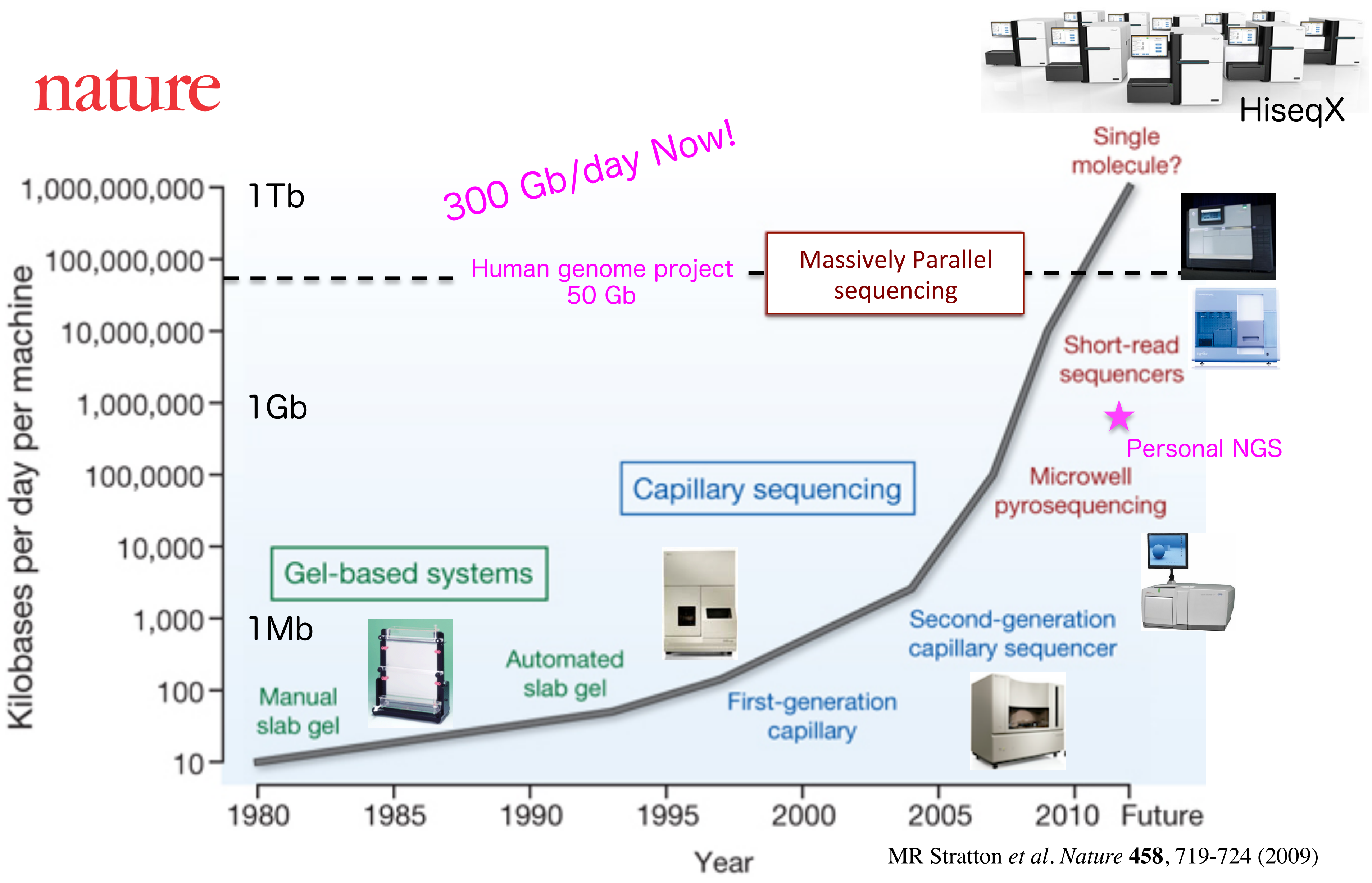
慶應義塾大学 環境情報学部の冨田勝 教授〔慶應義塾大学先端生命科学研究所（山形県鶴岡市、ＩＡＢ）所長〕は、自身の全ゲノム配列を解析し、7 月 31 日より国立遺伝学研究所の日本ＤＮＡデータバンク（ＤＤＢＪ）にて全世界に公開を始めました。日本人が実名で個人ゲノム(*1)を公開するのは初めてであり、今回のケースが第一号となります。

湘南藤沢キャンパス（ＳＦＣ）の総合政策学部・環境情報学部では、この冨田教授のゲノムを教材とした「ゲノム解析ワークショップ」を開講し、本年 **4** 月から **7** 月にかけてパーソナルゲノムに関する授業を行いました。

https://www.keio.ac.jp/ja/press_release/2012/kr7a4300000aubbx.html Oct.4 2016

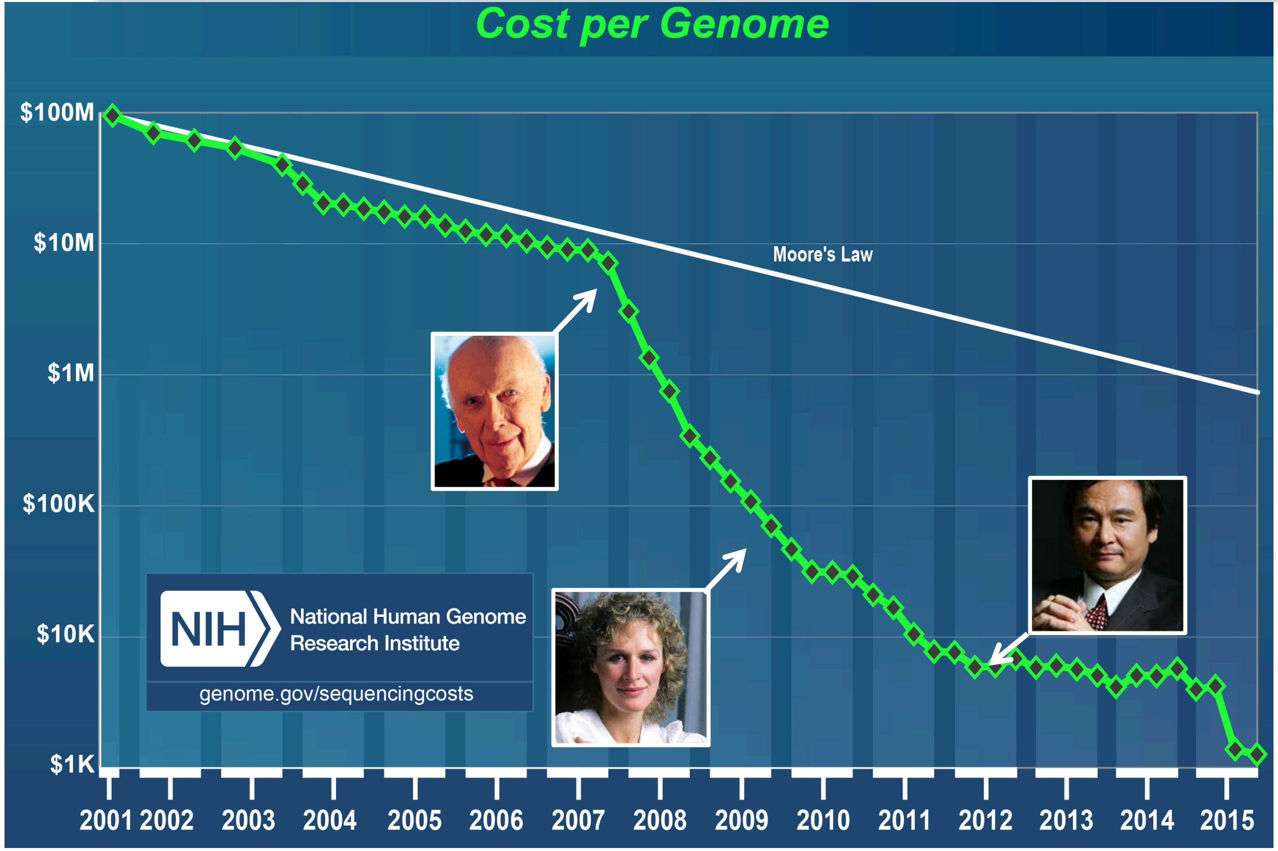
The History of DNA Sequencing Technology

nature



MR Stratton *et al*. *Nature* **458**, 719-724 (2009)

$1000 genome



https://www.genome.gov/sequencingcostsdata/ (Oct. 4, 2016)

試薬	$797
本体	$137
検体処理など	$55-65

⇩

$1000 Genome Era

Nature doi:10.1038/nature.2014.14530

2014年1月にイルミナ社は１つのシステムでDNA抽出から人件費までのコストをすべて含めて1000ドル以下で全ゲノムを 30x カバレッジで解析できるHiSeq Xシステムを発表した。このシステムでは年間1万8千検体を処理することができると公表している。

実際の個人ゲノム解読委託コスト

$2,900 (Nov. 2015)

Understand Your Genome,
provided by Illumina.

http://www.illumina.com/company/events/understand-your-genome.html/

30x WGS
CLIA/CAPラボ
UYG Tokyo では $4,900

$999 (Sep. 2015)

myGenome,
provided by Veritas Genetics.

Veritas Genetics
Login
Products
Services
About
Contact
myGenome
ORDER NOW
LEARN MORE
U.S. price.

We sequence your whole genome to help you improve your health, longevity, and much more. Order now for $999.

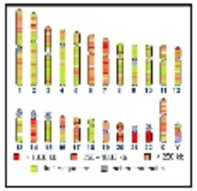
https://www.veritasgenetics.com/

30x WGS
CLIA/CAPラボ
HIPPA法遵守

WGSとiPodの比較

2003 Human Genome Project ending (1990-2003)

2.7 billion dollars



x 0.00000037

1/2,700,000

2014 Illumina launchs HiSeqX Ten

1000 dollars/genome

"Forty such machines would be able to sequence more genomes in one year than had been produced by all other sequencers to date."

シークエンシングのコストは１年あたりおよそ3.84倍安くなっている。

2003	2016
iPod 3rd	iPod touch

Red Hot Chil...
11 von 17
Can't Stop
Red Hot Chili Peppers
By The Way
0:13
-4:16
MENU

Copyright © 2016 Apple Inc.

2003	2016
30 GB	128 GB
59,800円	38,800円

x 0.27

1/3.67

x 0.00000037

1/2,700,000

128 GB

0.14円

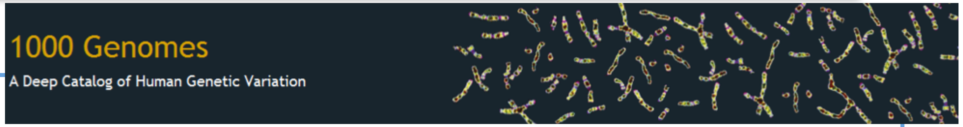
or

52,555 TB

38,800円

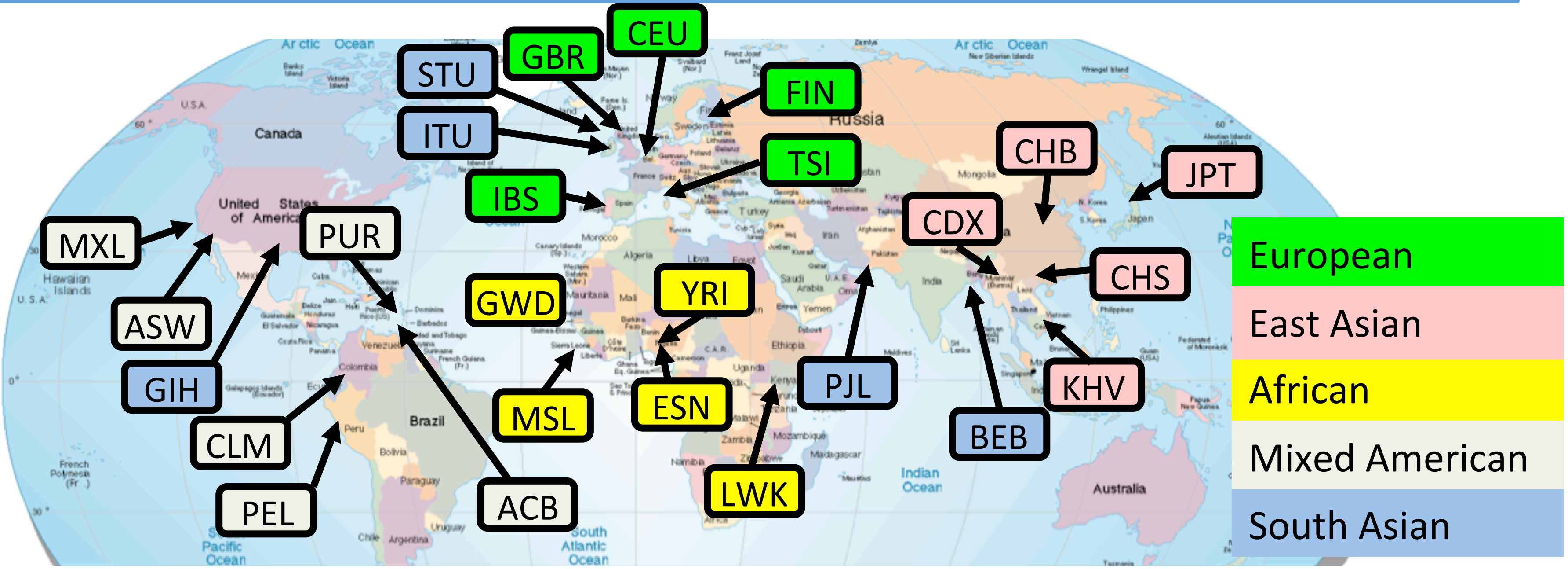
1000 genomes

http://www.internationalgenome.org/


1000 Genomes
A Deep Catalog of Human Genetic Variation

パイロット研究：～2009年
　ヨーロッパ、アフリカの親子（トリオ）全ゲノム解読、179人の全ゲノム解析、697人の全エキソン解析

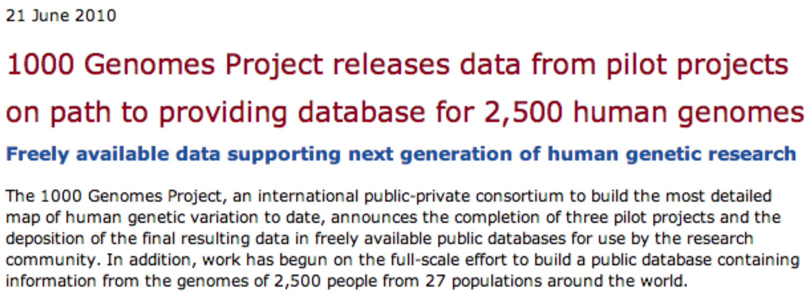
本格解析：2009年～
　27の異なる人種2500人の全ゲノム解析。ヨーロッパ（ユタ州）、イタリア、イギリス、フィンランド、スペイン、日本、ベトナム、中国x3、ナイジェリアx2、ケニヤ、ガンビア、シエラレオネ、アフリカン・アメリカン、バルバドス、メキシコ、プエルトリコ、コロンビア、ペルー、インド（ヒューストン在住）、パキスタン、バングラディッシュ、インド（英国在住）、スリランカ（英国在住）


CEU
GBR
STU
FIN
ITU
TSI
IBS
CHB
JPT
CDX
CHS
MXL
PUR
ASW
GIH
CLM
PEL
ACB
GWD
YRI
MSL
ESN
LWK
PJL
BEB
KHV
European
East Asian
African
Mixed American
South Asian

日本人個人ゲノム

2010年6月22日

日本人の100人ゲノム、Exomeデータの公開



21 June 2010

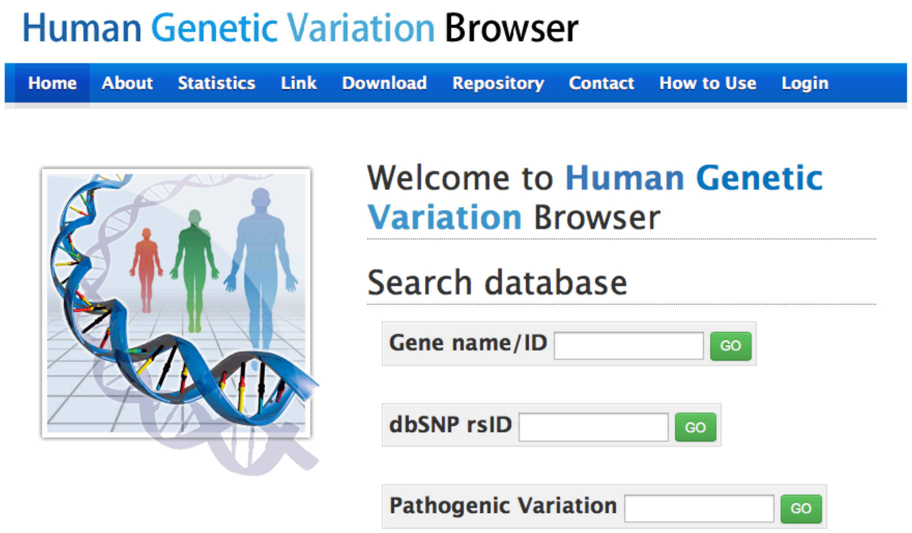
1000 Genomes Project releases data from pilot projects on path to providing database for 2,500 human genomes

Freely available data supporting next generation of human genetic research

The 1000 Genomes Project, an international public-private consortium to build the most detailed map of human genetic variation to date, announces the completion of three pilot projects and the deposition of the final resulting data in freely available public databases for use by the research community. In addition, work has begun on the full-scale effort to build a public database containing information from the genomes of 2,500 people from 27 populations around the world.

2013年11月12日

日本人1208人のExomeデータの公開



Human Genetic Variation Browser

Welcome to **Human Genetic Variation** Browser

Search database

Gene name/ID ☐ GO

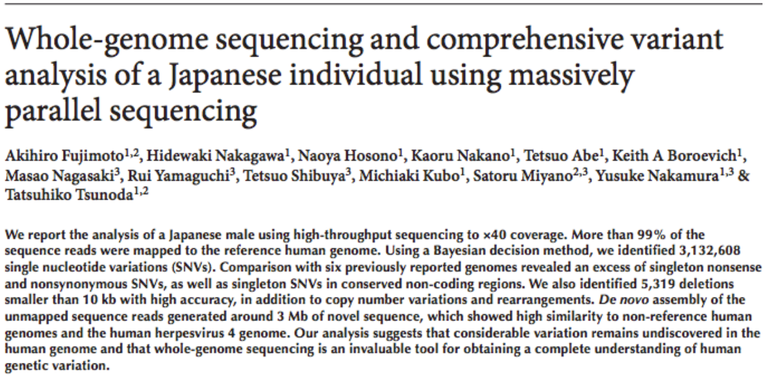
dbSNP rsID ☐ GO

Pathogenic Variation ☐ GO

J Hum Genet, 2016 Feb 25

2010年11月11日

日本人の個人ゲノムの解読



Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing

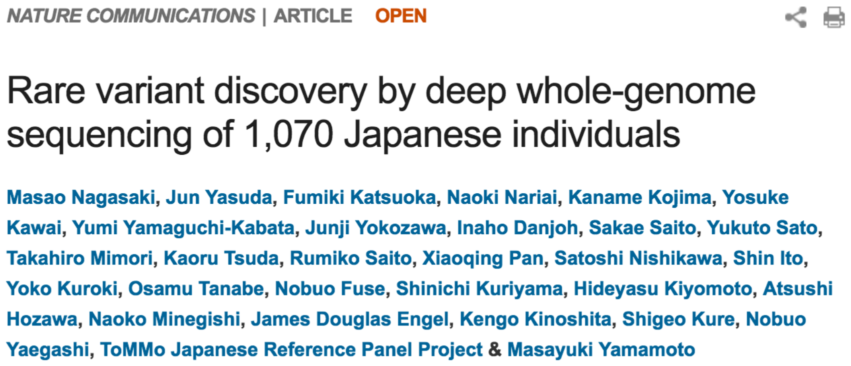
Akihiro Fujimoto[1,2], Hidewaki Nakagawa[1], Naoya Hosono[1], Kaoru Nakano[1], Tetsuo Abe[1], Keith A Boroevich[1], Masao Nagasaki[3], Rui Yamaguchi[3], Tetsuo Shibuya[3], Michiaki Kubo[1], Satoru Miyano[2,3], Yusuke Nakamura[1,3] & Tatsuhiko Tsunoda[1,2]

We report the analysis of a Japanese male using high-throughput sequencing to ×40 coverage. More than 99% of the sequence reads were mapped to the reference human genome. Using a Bayesian decision method, we identified 3,132,608 single nucleotide variations (SNVs). Comparison with six previously reported genomes revealed an excess of singleton nonsense and nonsynonymous SNVs, as well as singleton SNVs in conserved non-coding regions. We also identified 5,319 deletions smaller than 10 kb with high accuracy, in addition to copy number variations and rearrangements. *De novo* assembly of the unmapped sequence reads generated around 3 Mb of novel sequence, which showed high similarity to non-reference human genomes and the human herpesvirus 4 genome. Our analysis suggests that considerable variation remains undiscovered in the human genome and that whole-genome sequencing is an invaluable tool for obtaining a complete understanding of human genetic variation.

Nat Genet. 2010;**42**:931-6

2015年8月21日

日本人1000人分のゲノム配列決定の発表



NATURE COMMUNICATIONS | ARTICLE OPEN

Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals

Masao Nagasaki, Jun Yasuda, Fumiki Katsuoka, Naoki Nariai, Kaname Kojima, Yosuke Kawai, Yumi Yamaguchi-Kabata, Junji Yokozawa, Inaho Danjoh, Sakae Saito, Yukuto Sato, Takahiro Mimori, Kaoru Tsuda, Rumiko Saito, Xiaoqing Pan, Satoshi Nishikawa, Shin Ito, Yoko Kuroki, Osamu Tanabe, Nobuo Fuse, Shinichi Kuriyama, Hideyasu Kiyomoto, Atsushi Hozawa, Naoko Minegishi, James Douglas Engel, Kengo Kinoshita, Shigeo Kure, Nobuo Yaegashi, ToMMo Japanese Reference Panel Project & Masayuki Yamamoto

Nature Commun., **6**:8018 (2015)

100K – 1M genome projects

برنامج الجينوم البشري السعودي
Saudi Human Genome Program

a national program for sequencing the genome ...

مدينة الملك عبد العزيز للعلوم والتقنية KACST

http://shgp.kacst.edu.sa/site/

サウジアラビアでは糖尿病、難聴、心疾患、神経疾患、遺伝病などの原因遺伝子解明のために健常人を含む10万人規模の全ゲノム解析を行う。

イギリスでは2017年までにがん疾患、希少疾患および感染症患者10万人の全ゲノム配列を解読し、医療に結びつけることを目指している。

Genomics england

https://www.genomicsengland.co.uk/

Veterans Health Administration
Research & Development
Improving Veterans' Lives www.research.va.gov

Million Veteran Program:
A Partnership with Veterans
MVP

http://www.research.va.gov/mvp/

米国退役軍人省では退役軍人のため、医療情報や健康管理情報と照合できる100万人のゲノム解析を行う。

http://www.jcvi.org/cms/home/

J. Craig Venterは長寿と老化防止を目指す疾病診断と幹細胞治療のベンチャー企業を設立し、10万人規模のゲノム解析を行う。

J. Craig Venter™
INSTITUTE

東北メディカル・メガバンク事業

岩手と宮城で15万人のバイオバンクを構築

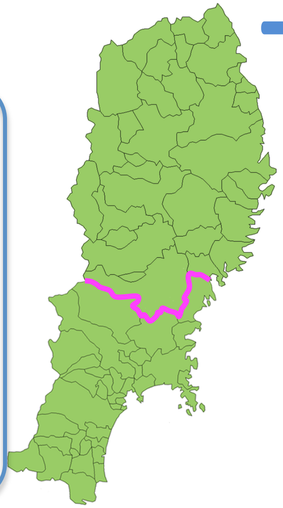
150,000 participants

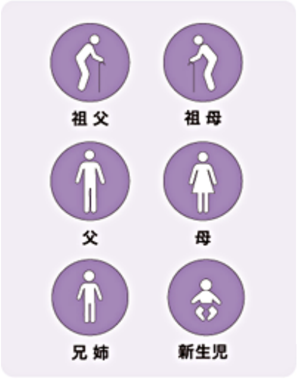
Three generations

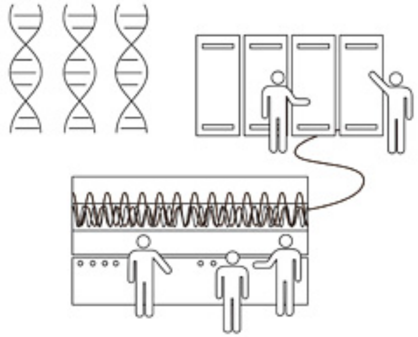
(children, parents and grandparents)

70,000 participants

(10,000 families)



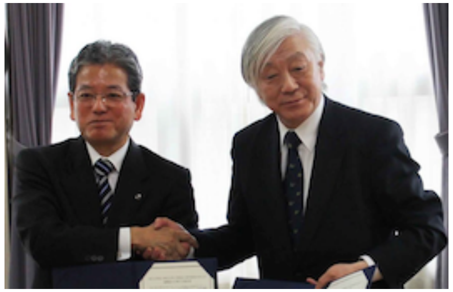




Residents

50,000 participants

Tohoku



30,000 participants

Iwate

ヒトゲノムの多様性

ヒト遺伝子の多様性

http://science.nationalgeographic.com/science/enlarge/skin-colors.html
Photograph by Sarah Leen Oct. 4 2016

99.92%

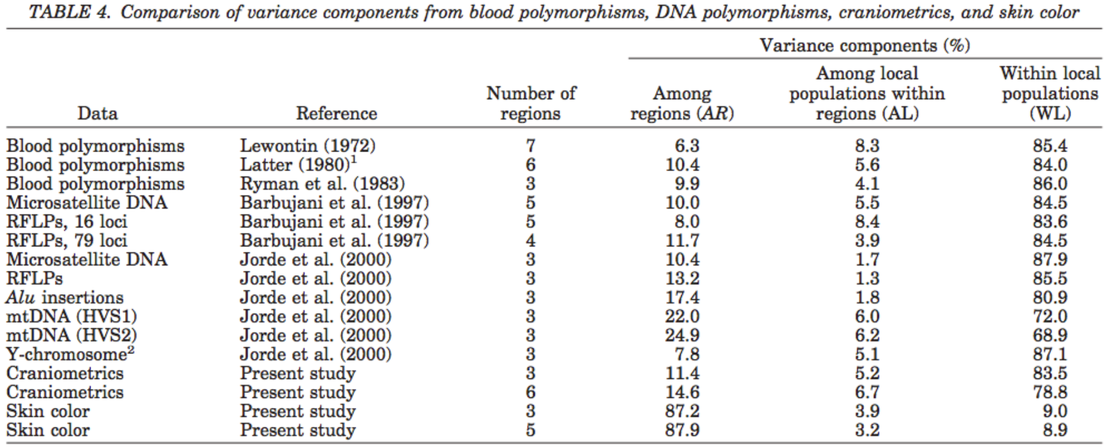


TABLE 4. Comparison of variance components from blood polymorphisms, DNA polymorphisms, craniometrics, and skin color

			Variance components (%)		
Data	Reference	Number of regions	Among regions (*AR*)	Among local populations within regions (AL)	Within local populations (WL)
Blood polymorphisms	Lewontin (1972)	7	6.3	8.3	85.4
Blood polymorphisms	Latter (1980)[1]	6	10.4	5.6	84.0
Blood polymorphisms	Ryman et al. (1983)	3	9.9	4.1	86.0
Microsatellite DNA	Barbujani et al. (1997)	5	10.0	5.5	84.5
RFLPs, 16 loci	Barbujani et al. (1997)	5	8.0	8.4	83.6
RFLPs, 79 loci	Barbujani et al. (1997)	4	11.7	3.9	84.5
Microsatellite DNA	Jorde et al. (2000)	3	10.4	1.7	87.9
RFLPs	Jorde et al. (2000)	3	13.2	1.3	85.5
Alu insertions	Jorde et al. (2000)	3	17.4	1.8	80.9
mtDNA (HVS1)	Jorde et al. (2000)	3	22.0	6.0	72.0
mtDNA (HVS2)	Jorde et al. (2000)	3	24.9	6.2	68.9
Y-chromosome[2]	Jorde et al. (2000)	3	7.8	5.1	87.1
Craniometrics	Present study	3	11.4	5.2	83.5
Craniometrics	Present study	6	14.6	6.7	78.8
Skin color	Present study	3	87.2	3.9	9.0
Skin color	Present study	5	87.9	3.2	8.9

Am J Phys Anthropol. 2002;118:393-8.

　ある２人の個人のゲノムの違いはおおよそ1,300文字につき１文字（0.1%）であり、そのうちの94%は人種を超えて見つかる多型である。すなわち、人種特異的に見つかる多型はわずか0.006%である。

多様性

SNP; Single Nucleotide Polymorphism
SNV; Single Nucleotide Variation

-CGCGAGTGCTATC-

-CGCGAGAGCTATC-

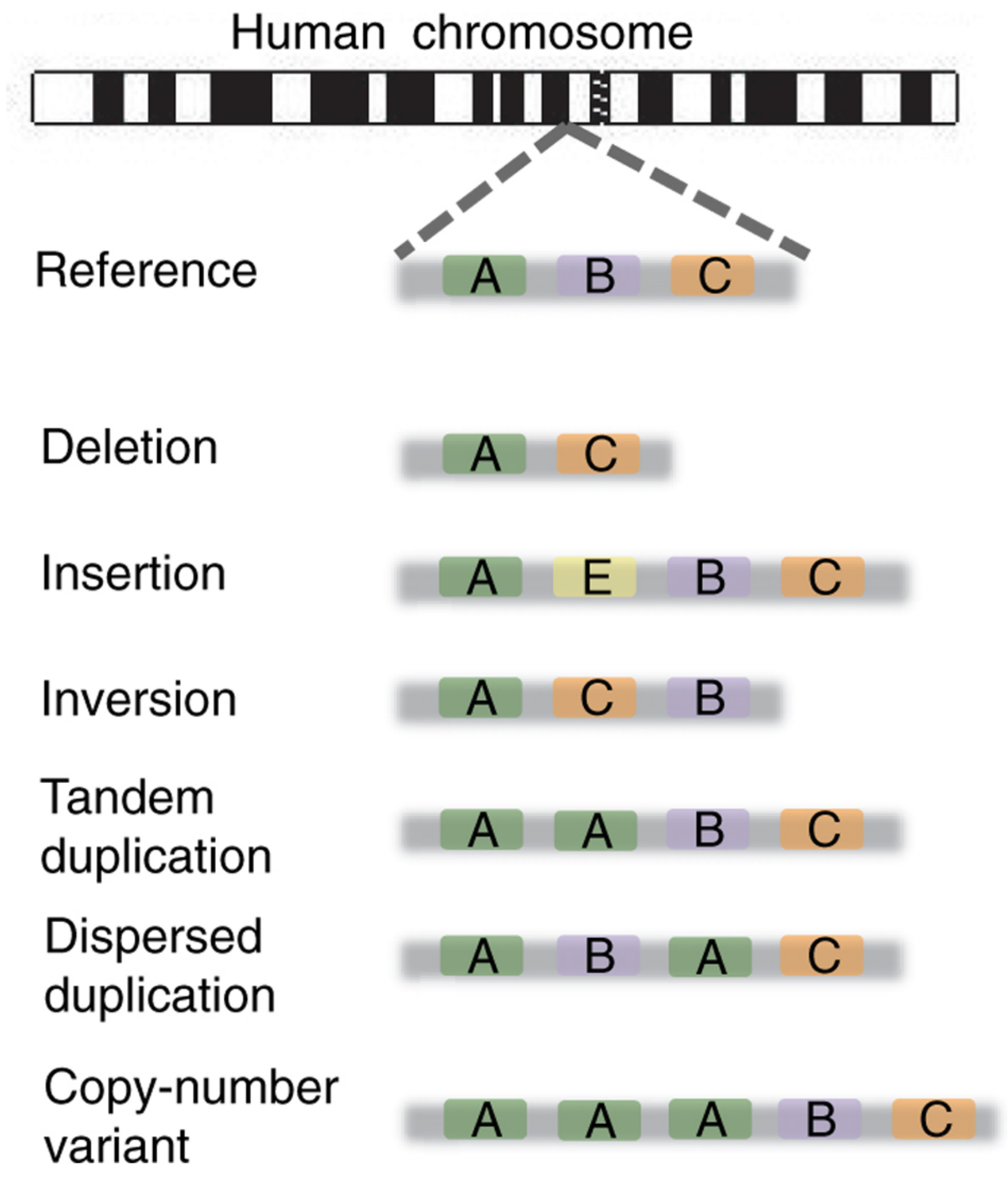
INDEL; Insertion Deletion

-CGCGAGTGCTATC-

-CGCGAG GCTATC-
-CGCGAGTGCTATC-
T

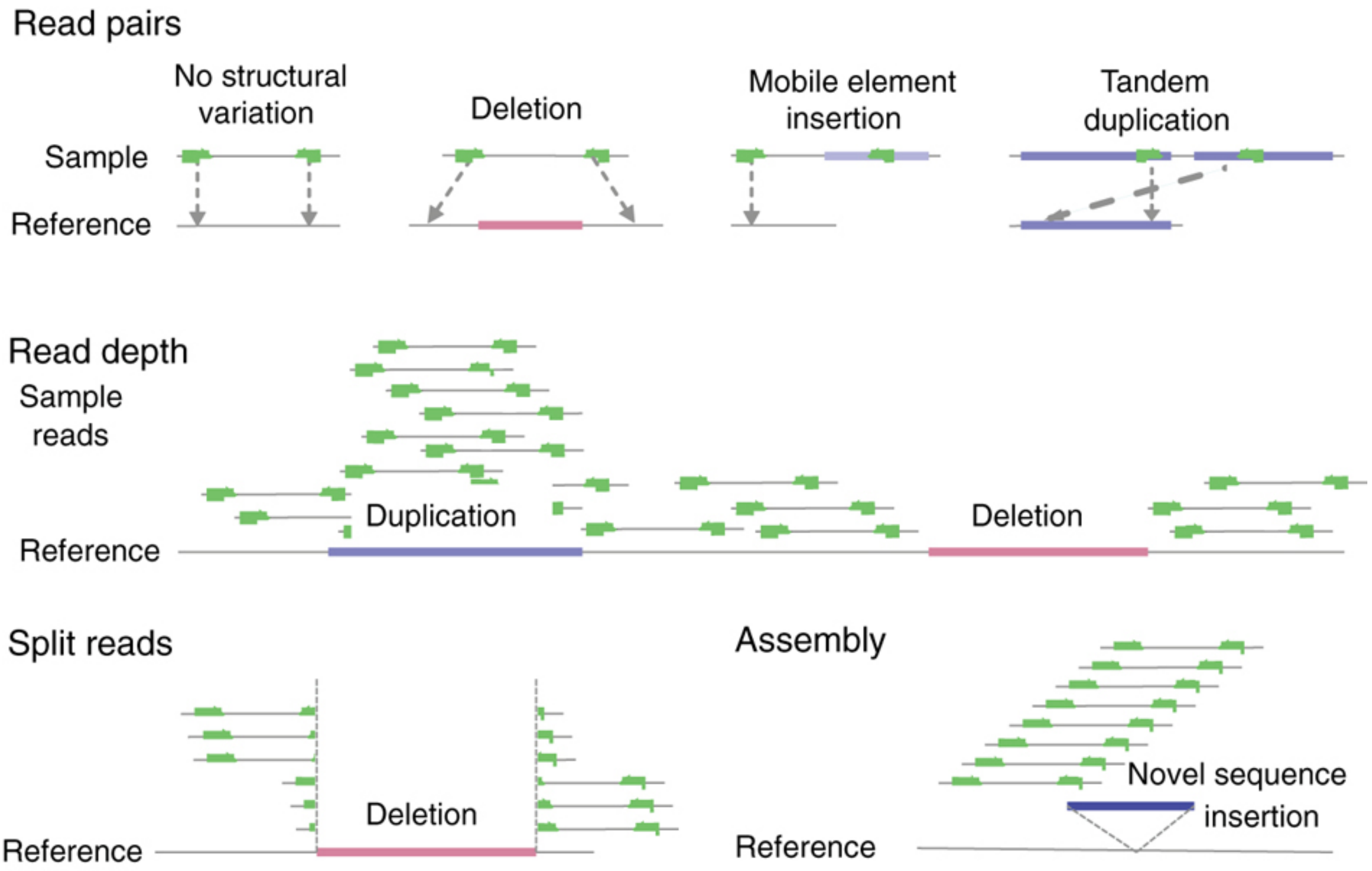
SV; Structural Variation
CNV; Copy Number Variation

Structural Variation


Human chromosome
Reference
A B C
Deletion
A C
Insertion
A E B C
Inversion
A C B
Tandem duplication
A A B C
Dispersed duplication
A B A C
Copy-number variant
A A A B C

Nature Methods 9, 133–137 (2012)

Structural Variation


Read pairs
No structural variation
Deletion
Mobile element insertion
Tandem duplication
Sample
Reference
Read depth
Sample reads
Duplication
Deletion
Reference
Split reads
Deletion
Reference
Assembly
Novel sequence insertion
Reference

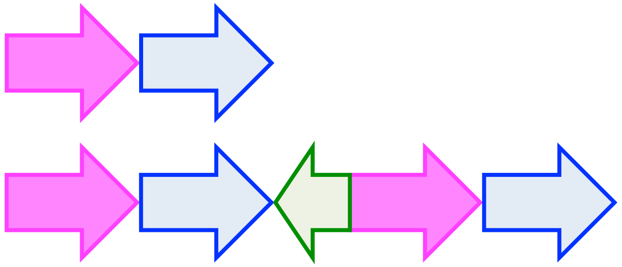
Nature Methods 9, 133–137 (2012)

ゲノムアノテーション

-CGCGAGTGCTATC-

-CGCGAGAGCTATC-

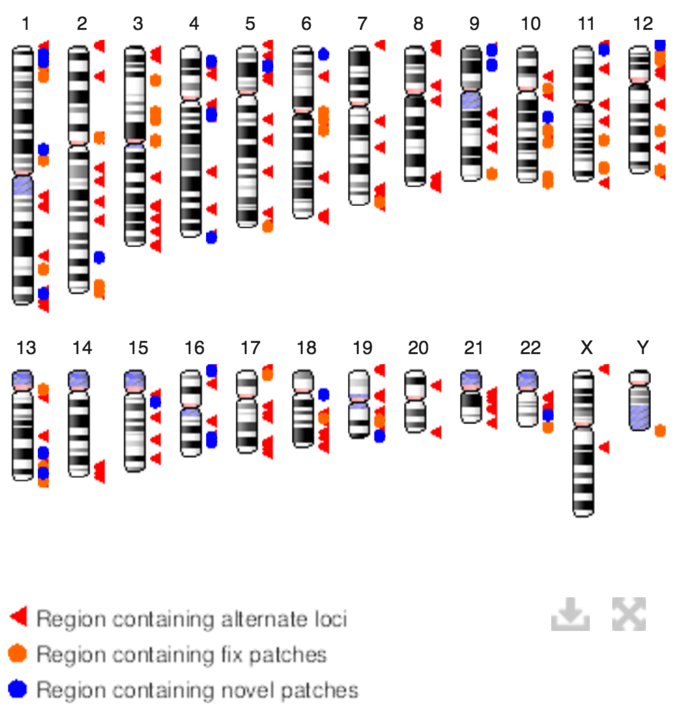
遺伝子機能
遺伝子制御



遺伝子機能
遺伝子制御

長鎖DNAシークエンサーによるゲノム解析

ヒトゲノム参照配列（GRCh38）


1 2 3 4 5 6 7 8 9 10 11 12
13 14 15 16 17 18 19 20 21 22 X Y
◀ Region containing alternate loci
● Region containing fix patches
● Region containing novel patches

An ideogram representation of the latest human assembly, GRCh38.p8 (not showing unplaced or unlocalized sequences).

https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Oct. 10 2016

783か所のGapが残っている。

PacBio RS II

101x genome data

ユークロマチンに位置する190か所のGapの
105か所を埋めることができた。

Seo J et al. *Nature* (2016) PMID 27706134

最新機種では
50x で530万円ほど

Sequel

ゲノム解読の課題

- 技術、コスト
- データ解釈