ヒトゲノム解析—現状と将来

The human genome project—present status and prospect

理研横浜研究所 ゲノム科学総合研究センター

服部 正平 Masahira HATTORI

Key words: human genome; sequencing; chromosome 21

1. ヒトゲノム解読の現状

1980年後半に世界的に開始されたヒトゲノムプロジェクトはゲノム物理地図作製終了後1996年2月に日欧米など16の公的機関よりなるシーケンシングコンソーシアムを設立し本格的なヒトゲノム配列決定を開始した。同コンソー

シアムは当初2005年の完全解読を目指したが、その後米企業セセラ社による2000年解読計画に対抗するかたちで、1998年にその完全解読を2003年へ短縮した概要版（ドラフト）配列を2000年に完成する計画に変更した。そして国際コンソーシアムおよびセセラ社は2000年6月にドラフト完成を世界的にアナウンスし、ついで2001年2月にその解析結果を *Nature*¹⁾ および *Science* 誌²⁾ にそれぞれ発表し

表1 ヒトゲノム染色体別解読完成計画

染色体	サイズ (Mb)	主センター	共同センター	終了年 (予定)
1	263	The Sanger Center	Washington University (50 Mb)	2002
2	255	Washington University	Institute for Systems Biology (small)	2002
3	214	Baylor College (15+130 Mb)	Chinese Academy of Science (40 Mb), Washington University (50 Mb)	2002
4	203	Washington University	The Sanger Center (2 Mb), Myers (11 Mb)	2003
5	194	DOE Joint Genome Institute/Stanford Human Genome Center		2002
6	183	The Sanger Center		2002
7	170	Washington University (140 Mb)	Washington University (30 Mb), Inst. for Systems Biology (small), Jena (small)	2002
8	155	Whitehead Institute	Keio University (10 Mb), Jena (small)	2003
9	145	The Sanger Center	Whitehead Institute (2 Mb), GBF-German Research (3 Mb)	2002
10	144	The Sanger Center	Genome Therapeutics Corporation (35 Mb)	2002
11	144	RIKEN GSC (11q 85 Mb)	Institute for Systems Biology (small), The Sanger Center (11p 6 Mb), Whitehead Institute (11p 60 Mb)	2003
12	143	Baylor College		2002
13	114	The Sanger Center		2002
14	109	Genoscope	Institute for Systems Biology (7 Mb), Washington University (3.7 Mb)	2001
15	106	Whitehead Institute	Institute for Systems Biology (10 Mb)	2002
16	98	DOE Joint Genome Institute/Stanford Human Genome Center	The Sanger Center (1 Mb)	2003
17	92	Whitehead Institute	Max Planck Institute (1.5 Mb)	2002
18	85	Whitehead Institute (18q 60 Mb)	RIKEN GSC (18p 25 Mb)	2002
19	67	DOE Joint Genome Institute/Stanford Human Genome Center		2001
20*	72	The Sanger Center, done		2001
21*	50	RIKEN GSC/Keio/Jena/Max Planck/GBF, done		2000
22*	56	The Sanger Center/Keio/Univ. Oklahoma, done		1999
X	164	The Sanger Center (100 Mb)/Baylor College (35 Mb)	Max Planck Institute (2 Mb), Jena (1 Mb)	2002
Y*	27	Washington University, done		2001

図1 公的16機関によるヒトゲノム解読の現状(2001年9月現在)

青い点は完成データ量を示し、約3000 Mbがヒトゲノムの全長。波線は2001年初頭での予想完成データ生産量、実線は同年9月に修正された予想完成データ生産量。

た。一方、現在までに完全解読が終了し論文として発表されている染色体は公的機関による1999年12月の22番³⁾と2000年5月の21番⁴⁾である。21番染色体解読では筆者らのグループがその中心的役割を果たした^{5), 6)}。2001年に入り、Yおよび20番染色体の完全解読が報告されているが論文としてまだ報告はない。現在国際コンソーシアムは2003年春を目処に残った各染色体の完全解読を目指しドラフトを高精度の完成配列へとそのシーケンシングを進めている。2001年9月における解読状況を図1にまた各染色体別の完成予定を表1に示す。図1から完全解読は早くも2002年中に完了することが期待されまさに今そのカウントダウンが始まっている。

2. ドラフト配列とは

国際コンソーシアムのシーケンシングでは、セントロメアなどヘテロクロマチン領域やサイズの大きな反復領域は解読対象ではなく、真正クロマチン領域のみが主な解読対象となっておりそのサイズは約3000 Mbである。また解読に使われているヒトゲノムの材料は各センターによってリソースが異なり、今回の配列は不特定複数のヒト配列となっている。今回国際コンソーシアムによって発表されたドラフト(2001年2月当時)は約2700 Mbで全体の90%の配列が決定されたことになる。しかしそのうち30%は21番や22番のように完成配列(10,000 bpに1 bp以下の間違いがあり、99.99%以上の精度)であるが、残りの70%は低精度(多く

の未決定領域が存在し、また1 kbに1 bp程度の間違いがあり、99.9%程度の精度)となっている。図1に示したように現時点(2001年9月)での完成配列は約50%となっており、着実に完成へとシーケンシングは進んでいる。低精度のドラフトは約1年でその大部分を決定できたが、それを高精度にするにはさらに2~3年かかり非常に高度なシーケンス技術を必要とする。すなわち決定領域を90から100%にすることと精度を99.9から99.99%にする努力はドラフトを完成する以上に容易でないことが理解できる。

3. ドラフト配列からわかったこと

遺伝子数: ドラフト配列とはいえ、われわれはほぼ全体のヒト遺伝情報を手中にすることができ、配列の解析からわれわれ人類の生命の設計図の中身をおおまかではあるが見ることができる。まず興味があるのはヒトがどれだけの遺伝子数をもっているかである。ヒト以外にこれまで線虫やドロソフィラなどのゲノム配列が決定されており、その遺伝子数からヒトは少なくとも10万程度の遺伝子をもつと推察されていた。しかし配列解析の結果、ヒト遺伝子数は3~4万であることが判明した。この数字はさまざまな解析からも同程度となっている(表2)。筆者らも21番染色体解析から4万弱と見積もっている⁴⁾。ゲノム配列から遺伝子部分を発見する手法には、cDNAデータなどをゲノム配列に張り付けて同定する方法とコンピュータを用いた遺伝子領域の予測による方法がある。ドラフト配列の解析では15,000個の遺伝子が

cDNA などの張り付けから、そして 17,000 個の遺伝子がコンピュータにより発見された (計 32,000 個)。セセラ社の解析でも同じ手法が用いられ、同程度の遺伝子数を報告している。表 3 に各生物の遺伝子数を示す。最近ヒト遺伝子数は 4~5 万であるという別の報告も出てきている^{7),8)}。より確定的な遺伝子数は完全解読終了まで待たねばならない。一方、転写産物 (mRNA) は約 12 万種であることが報告されている⁹⁾。この数字の差は一つの遺伝子が複数種の mRNA を転写していることを示唆している。すなわちヒトにおいてはゲ

ノム上での遺伝子数の増大よりもむしろ転写段階での多様化 (alternative splicing による variant mRNA の生成=1 遺伝子が異なった 2 種類以上のタンパク質を生産する) が新たな機能獲得に重要であることをうかがわせる。このことは今日教科書でいわれている 1 遺伝子=1 タンパク質の概念をもはや否定している。

全体像: ヒト (一般には乳動物) ゲノムの特徴の一つはゲノムのかかなりの領域が反復配列と呼ばれる繰り返し配列で占められていることである。ヒトゲノムの場合は約 50% が反復配列であった。これら反復配列の 90% 程度は分散型反復配列である。SINE (150 万コピー) や LINE (85 万コピー) などが代表的であり、RNA を介した逆転写により増幅についてゲノムへ挿入する機構が明らかとなっている。この反復配列はヒトの機能に直接かかわるタンパク質をコードしておらずその存在する意味は不明である。しかしその増幅は現在においても起こっており、遺伝子部分への *de novo* 挿入が病気を起こした例がいくつか報告されている。これら反復配列の分散や増幅時期の詳細な解析はヒトゲノムの進化や染色体構築の機構を推察するうえで多くの知見を提供する。ヒトゲノムの 20 kb ほどの GC 含量は 30~70% の範囲 (平均は 41%) でゲノム中をモザイク状に変動している。GC 含量と上記の SINE と LINE の分布には相関がある。GC 含量が高い領域には SINE が多く、低い領域には LINE が相対的に多い。反復配列を除いた残り 50% はいわゆるユニーク配列であり遺伝子のほとんどがこの領域に存在する。遺伝子の密度も染色体によって高低がある。一般に GC 含量の高い染色体は遺伝子密度が高く、低い染色体は密度が低い傾向にある。

表 2 ゲノム配列からヒト遺伝子数の推定

	遺伝子数	推定の根拠
国際 Consortium	32,000	Blast+コンピュータ予測
HGREP	34,000	Blast+コンピュータ予測
Ensemble	39,000	Blast+コンピュータ予測
#21 Consortium	<40,000	#21+#22 完成データ
	36,000	マウスゲノムとの比較
France	31,000	フグゲノムとの比較
	35,000	EST のクラスタリング
平均	30,000~40,000	

表 3 ゲノム配列から推定された遺伝子数の種間比較

ヒト	32,000
シロイヌナズナ	25,706
ショウジョウバエ	13,338
線虫	18,256
酵母	6,144
大腸菌/枯草菌	4,000

図 2 ヒト 21 番染色体 (34 Mb) の構造解析

Clone はシーケンスに用いた BAC, PAC, コスミドなど。CpG islands は 250 bp 以上、50% 以上の GC 含量、CpG の実際の出現頻度/期待出現頻度 > 0.6 で一般に遺伝子と位置的に相関がある。NotI は制限酵素サイトで CpG サイトをもつ 8 塩基カッター。Alu および LINE1 はそれぞれ代表的な反復配列。

表 4 ヒト概要ゲノム配列からのヒトゲノム全体像

既知遺伝子と一致: 14,882 (17,764)	tRNA: 497
コンピュータ予測: 16,896 (~21,350)	SnRNA: 78
計 Protein-coding genes: 31,778	
平均エクソンサイズ: 145 bp	エクソン: 1.1%
平均エクソン数/遺伝子: 8.8	イントロン: 24%
平均イントロンサイズ: 3,365 bp	
平均 Protein coding サイズ: 1340 bp (447 aa)	
平均遺伝子サイズ: 27 kb	
Gene density highest: Chr.19 (23 genes/Mb)	
lowest: Chr.13 (5 genes/Mb)	

国際コンソーシアムからのデータ。上段かこ内はセラ社の数字。

同様な傾向は同一染色体内においても観察され、例えば図2に示した21番染色体の全体像の解析結果からもGC含量が高い領域に多くの遺伝子が偏って存在していることがわかる。この他染色体内や染色体間にユニーク領域が重複した相同領域がかなり存在することが明らかとなっている²⁾。これら相同領域の存在は遺伝子重複による新たな遺伝子(機能)の獲得に関する機構の解明に重要である。最近ヒトを含めた霊長類ゲノムの解析からこれら重複領域に存在する遺伝子についての興味ある結果が報告されている¹⁰⁾。今回発見された遺伝子のうち約60%は機能未知であった。またヒト(せきつい動物)に特異的な遺伝子も見つかっており、免疫系遺伝子や神経細胞で働く遺伝子群などがその大部分を占める¹⁾。表4に遺伝子の平均サイズなどヒトゲノムの各種統計値を挙げておく。

4. 解読終了後のゲノム研究

ヒトゲノムはその完全解読に向かってシーケンスが進められているが、すでにその後に続く研究が開始されている。これらを以下に列挙しておく。

1. 全遺伝子の発見(遺伝子 annotation)—比較ゲノム解析(表5)
2. 各遺伝子の機能解析(タンパク質 motif)—Proteome
3. 遺伝子間相互作用(Transcriptome)—マイクロアレイ技術
4. 各遺伝子産物(タンパク質)の相互作用研究(Interactome)—3次元立体構造
5. 医学・医療への応用研究—SNP, 病気遺伝子の同定, 遺伝子治療, 創薬など

表5に示した他の生物種のゲノム解析では、これらゲノム配列を相互に比較することによりゲノム中の機能領域を同定しようというものである。機能領域(配列)の多くは進化の過程で一般によく保存されるので、進化上離れた生物種間(例えばヒト・マウス)での比較ゲノム解析は未知遺伝子、発現制御領域、複製などの機能領域を、より詳細にゲノム配列上にアノテートすることができる。

表 5 現在ゲノムシーケンス進行中の生物種

Mouse	Washington Univ., MIT (Celera)
Rat	Baylor, Celera
Zebrafish	Sanger
Pig	China-Beijin
Chimpanzee	Japan
Rice	Japan, USA
Fugu	JGI
Tetraodon	Genoscope
<i>Ciona intestinalis</i>	JGI
<i>Ciona savignyi</i>	MIT
各種微生物	Gold database: http://ergo.integratedgenomics.com/GOLD/

5. データベースと情報 URL

上記のポストシーケンス研究の推進にはさまざまなゲノムデータを整理整頓したデータベースの構築および情報生物学の手法の進歩が必須となる。たとえばヒトゲノム研究では、ヒトゲノムの配列、遺伝子、SNPなどの情報がデータベース化されておりこれらデータは各種URLからダウンロードできる。代表的なURLを下記に挙げておく。詳細は文献11を参照されたい。

・遺伝子や文献情報:

LocusLink: <http://www.ncbi.nlm.nih.gov/LocusLink/>
PubMed: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>

・ヒトゲノム情報:

HGREP: <http://hgrep.ims.u-tokyo.ac.jp/>
Ensembl: <http://www.ensembl.org/>
Mapview: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch

・病気等の情報:

OMIM: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

・SNP 情報

dbSNP: <http://www.ncbi.nlm.nih.gov/SNP/index.html>
JSNP Database: <http://snp.ims.u-tokyo.ac.jp>

6. おわりに

あらゆる生物種のゲノム情報は21世紀の生命科学の新たな展開において中心的・基盤的な役割を果たすことは間違いなく、その根幹にあるものがゲノム配列情報である。上述したDNAマイクロアレイによる遺伝子発現のグローバルな解析やタンパク質レベルでのプロテオーム解析などのゲノム機能解析も、ゲノム配列情報が存在してこそ強力な研究手法になりうるのである。全ゲノム配列の決定のもたらす第一の恩恵は、各生物がどのような遺伝子をもっているか、またもっていないかを一挙に明らかにできるところにある。この「全

ゲノムを知る' ことが出発点となりあらゆる方面での研究が大きく展開される。最後に、微生物からヒトに至るゲノム配列決定に関する技術についての成書¹²⁾を紹介しておく。

- 1) International Human Genome Sequencing Consortium: *Nature*, **409**, 860-921 (2001).
- 2) J. C. Venter *et al.*: *Science*, **291**, 1304-1351 (2001).
- 3) I. Dunham *et al.*: *Nature*, **402**, 489-495 (1999).
- 4) M. Hattori *et al.*: *Nature*, **405**, 311-319 (2000).
- 5) 服部正平: 蛋白質・核酸・酵素, **45**, 1978-1985 (2000).
- 6) 榊 佳之: 科学, **70**, 630-635 (2000).
- 7) J. B. Hogenesch *et al.*: *Cell*, **106**, 413-415 (2001).
- 8) A. A. Camargo *et al.*: *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 12103-12108 (2001).
- 9) F. Liang *et al.*: *Nature Genet.*, **25**, 239-240 (2000).
- 10) M. E. Johnson *et al.*: *Nature*, **413**, 514-518 (2001).
- 11) 高木利久編: 「ゲノム医科学と基礎からのバイオインフォマティクス」, 実験医学増刊, 羊土社, 2001.
- 12) 小笠原直毅ほか: 「ホールゲノムショットガン法によるゲノム解析とアノテーション」, 学会出版センター, 東京, 2001.