報道発表資料

2000 年 6 月 26 日
独立行政法人 理化学研究所
科学技術振興事業団
慶應義塾大学

国際ヒトゲノムシーケンス決定コンソーシアムが ヒトゲノムのドラフトシーケンスを終了

ヒトゲノム完全解読を目指す「国際ヒトゲノムシーケンス決定コンソーシアム」は、ヒトゲノムのドラフトシーケンスを終えました。「国際ヒトゲノムシーケンス決定コンソーシアム」は、日・米・英・仏・独・中国の計 16 センターが参画。日本では、理化学研究所ゲノム科学総合研究センターが 11、18、21 番染色体の一部を担当し、科学技術振興事業団（JST）などの支援を受けた慶應義塾大学医学部のグループが、2、6、8、21、22 番染色体の一部を担当しています。さらに、上記コンソーシアム発足以前より開始されていた JST のプロジェクトのもとに解析を進めていた、東海大学医学部及び、（財）癌研究会による解読結果の一部も、コンソーシアムのデータとして取り込まれており、日本全体では、ヒトゲノム全体の約 7%を担当しています（各センターの分担状況は別添資料を参照）。

決定された配列データは、ヒトゲノム全体の約 90%に相当する 2700Mb。その内訳は、先に詳細に解析を終え論文発表された 21、22 番染色体を含め、2 割が 99.99%以上の高精度で解読を終了しました。さらに、これらよりもやや低い精度（少なくとも 99.9%以上）で読まれた「ドラフトシーケンス」が 7 割です。

本件については、米国、英国において「国際ヒトゲノムシーケンス決定コンソーシアム」の各国代表機関から、それぞれプレス発表されます。

1.背　景

国際ヒトゲノム計画の最大の目標は、ヒトゲノムの全 30 億塩基の配列を決定し、そこに書き込まれた遺伝情報を読み取ることです。全塩基配列決定のため 1996 年より「国際ヒトゲノムシーケンス決定コンソーシアム」が形成され、活動を行ってきました。当初は高精度（99.99%以上）のデータを積み重ねることを戦略としてきましたが、精度は若干落ちても全体像を早く知ることも学問の発展にとって有用と考えられることから、昨年 5 月より約 1 年間の予定で全体像を大ざっぱに掌握するためのドラフトシーケンスプロジェクトを開始しました。

2. 研究成果

今回、ヒトゲノムの約 90%を占める 2700Mb の配列が決定しましたが、このヒトゲノムのドラフトシーケンス決定にあたっては、まず、ヒトゲノムの BAC クローンライブラリーを主に使い、各クローンを染色体上の特定領域に位置付けるマッピング作業が米国ワシントン大学を中心に進められてきました。マッピングされたクローンごとに重複度 4 以上で配列を決定し、そのデータを直ちに公用データベースに登録。一般に公開するとともに、米国の NCBI が世話役となってそれらのデー

タ全体の編集作業を行ってきました。その進捗状況は、NCBI からウィークリーリポートとして各センターに知らされる一方、全チームは、3 ないし 4 カ月ごとに会合を持ち、その進捗状況のチェックや問題点の解決に当たってきました。

日本では、理化学研究所ゲノム科学総合研究センター及び、科学技術振興事業団（JST）などの支援を受けた慶應義塾大学医学部のグループが参画。日本全体では、約 210Mb の配列を決定しました。これは、ヒトゲノム全体の約 7%にあたり、このような人類共有の財産を形成する歴史的プロジェクトにわが国は大いに貢献しました。

3. 今後の展開

今回の成果はあくまでも「ドラフトシーケンス」であり、最終的なデータではありません。今後、国際ヒトゲノム計画ではドラフトシーケンスに十分な意味付け・注釈付け（Annotation）を行った後、9 月にもその全内容を論文として公表します。また、全配列の高精度解読に向けて努力を続け、遅くとも 2003 年春までにはヒトゲノム配列全体の高精度解読を終了する予定です。

ドラフトシーケンスの終了によって（一部不正確を含む）、ヒトゲノム上の大半の遺伝子が同定され、また多数の SNP が発見されるものと思われます。これらの情報は疾患遺伝子の探索を飛躍的にスピードアップし、医学に大きな貢献をもたらす事が期待されています。また、発生・分化や免疫反応、脳・神経系など、ヒトを形成する高次の機能の解明も大きく進展するに違いありません。また、人類の進化についても、これまでとは異なる新しいアプローチが可能となります。

（問い合わせ先）
独立行政法人理化学研究所　ゲノム科学総合研究センター
ゲノム構造情報研究グループ
プロジェクトディレクター　榊　佳之
Tel ： 042-778-9923 / Fax ： 042-778-9924
Tel ： 03-5449-5622 / Fax ： 03-5449-5445
（東大医科研）

慶應義塾大学　医学部
分子生物学教室　教授　清水　信義
Tel ： 03-3563-3755 / Fax ： 03-3351-2370

（報道担当）
独立行政法人理化学研究所 広報室　嶋田　庸嗣
Tel ： 048-467-9272 / Fax ： 048-462-4715

<補足説明>

※ Mb

メガベース（100 万塩基）の略

※ BAC ライブラリー

バクテリアの人工染色体を用いてクローニングされた（ヒト）DNA 断片の集団（ライブラリー）

※ NCBI

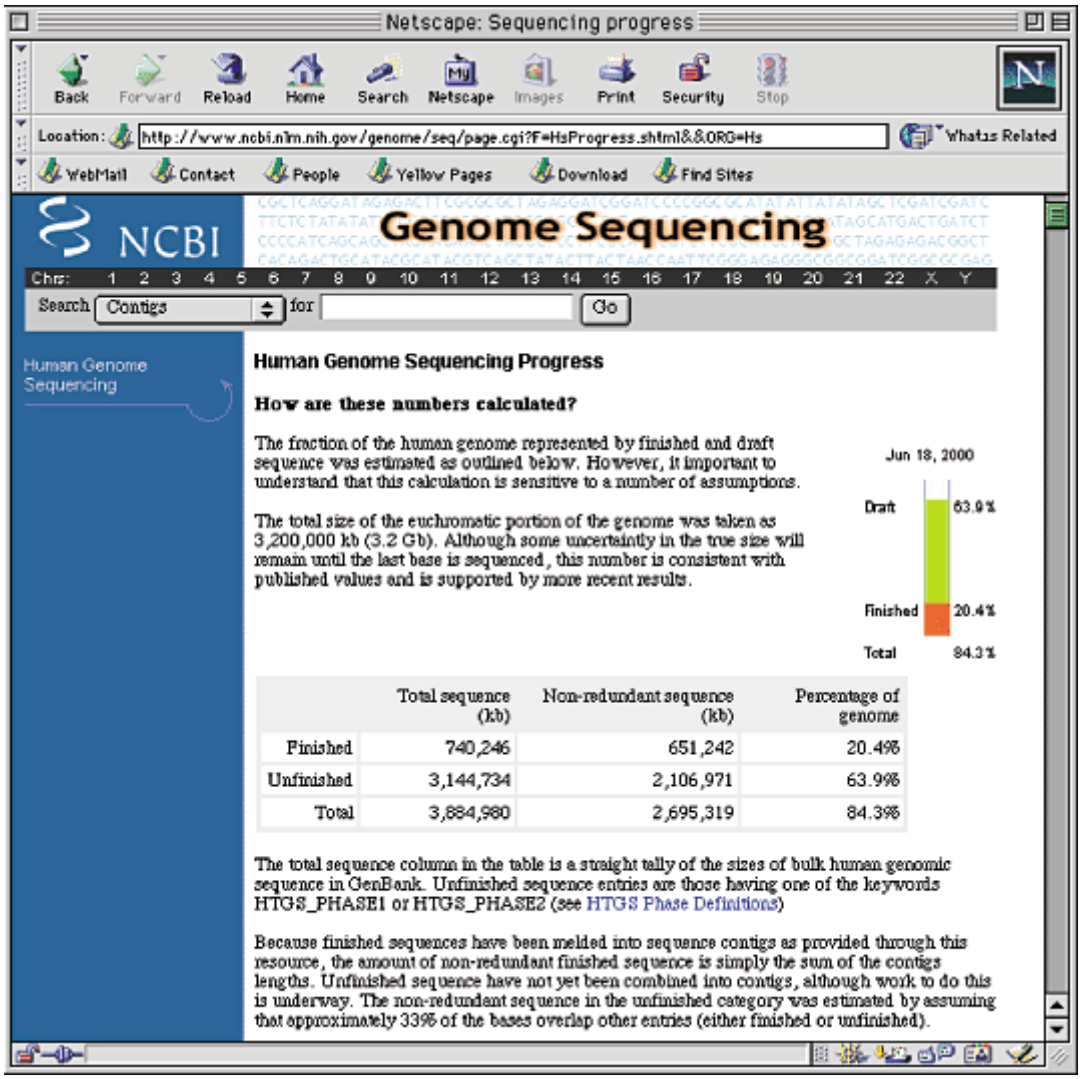
アメリカ国立バイオテクノロジー情報センター（National Center for Biotechnology Information）の略

※ SNP

1 塩基多型（Single Nucleotide Polymorphism）の略

※ JST の取り組み

科学技術振興事業団による JST シークエンシングプロジェクトは、1995 年より開始され、北里大学、慶應義塾大学、東海大学、癌研究会が参加し、詳細塩基配列を行ってきた。



「ヒトゲノム計画」の解析状況

（URL：http://www.ncbi.nlm.nih.gov/genome/seq）

各センターの解析状況

（2000年6月25日現在）

機関名	ドラフトシーケンス(kb)	完成データ(kb)	合計(kb)
ベーラー医科大（米）	212,125	46,633	258,758
中国科学院 ヒトゲノムセンター（中国）	41,762	782	42,544
ゲノムテクノロジーセンター（独）	2,050	2,531	4,581
ジェノスコープ（仏）	33,868	47,837	81,705
ゲノム治療コーポレーション（米）	62,445	5,530	67,975
分子ゲノムテクノロジー 研究所（独）	24,088	22,721	46,809
エネルギー省合同ゲノム 研究所（米）	305,456	55,896	361,352
慶応大（日）	5,867	14,238	20,105
マックスプランク研究所（独）	4,711	4,020	8,731
理研GSC（日）	169,194	18,862	188,056
サンガーセンター（英）	634,611	208,739	843,350
スタンフォード大学 DNA　シーケンスセンター（米）	29,436	867	30,303
ワシントン州立大 ゲノムセンター（米）	15,388	11,275	26,663
ワシントン州立大 シーケンスセンター（米）	18,550	10,112	28,662
MIT ホワイトヘッド 医科学研究所（米）	1,042,089	46,585	1,088,674
ワシントン大学 ゲノムシーケンスセンター（米）	508,795	145,148	653,943
その他※	72,414	110,259	182,673
合計	3,182,849	752,035	3,934,884

※JSTの支援による東海大 癌研究会の11783kbを含む

International Sequencing Consortium Announces Working Draft of the Human Genome

The Human Genome Project public consortium today announced that it hasassembled a working draft of the sequence of the human genome-the geneticblueprint for a human being.

This major milestone involved two tasks： placing large fragments of DNA in the proper order to cover all of the human chromosomes, and determining the DNA sequence of these fragments.

The assembly reported today consists of overlapping fragments covering 97 percent of the human genome, of which sequence has already been assembled for 85 percent of the genome. The sequence has been threaded together into a string of A,T,C, and Gs arrayed along the length of the human chromosomes.

Production of genome sequence has skyrocketed over the past year, with more than 60 percent of the sequence having been produced in the past six months alone. During this time, the consortium has been producing 1000 bases a second of raw sequence - 7 days a week, 24 hours a day.

The average quality of the working draft sequence far exceeds the consortium's original expectations for this intermediate product. Consortium centers have produced far more sequence data than expected (over eighteen billion bases of raw sequence data, comprising overlapping fragments totaling 3.9 billion bases). As a result, the working draft is substantially closer to the ultimate finished form than the consortium expected at this stage. Specifically, more than 50 percent of the genome sequence is in near-finished form or better, and 20 percent of it is in completely finished form. Across the genome, the average DNA segment resides in a continuous gapless sequence "contig" of 200,678 bases.

The sequence information from the public project has been continuously, immediately and freely released to the world, with no restrictions on its use or redistribution. The information is scanned daily by scientists in academia and industry, as well as by commercial database companies providing information services to biotechnologists. Already, many tens of thousands of genes have been identified from the genome sequence. More than a dozen disease genes have been pinpointed by access to the working draft.

Consortium goals. The consortium's goal for the Spring of 2000 was to produce a working draft version of the human sequence, an assembly containing overlapping fragments that cover 90 percent of the genome and that are

sequenced in 'working draft' form, i.e- with some gaps and ambiguities. The consortium's ultimate goal is to produce a completely "finished" sequence, i.e. one with no gaps and 99.99 percent accuracy. The target date for this ultimate goal had been 2003, but today's results mean that the final, stand-the-test-of-time sequence will likely be produced considerably ahead of that schedule.

Complementary approaches. In a related announcement, Celera Genomics announced today that it has completed it's own first assembly of the human genome DNA sequence.

The public and private projects use similar automation and sequencing technology, but different approaches to sequencing the human genome. The public project uses a 'hierarchical shotgun' approach in which individual large DNA fragments of known position are subjected to shotgun sequencing (i.e, shredded into small fragments that are sequenced, and then reassembled on the basis of sequence overlaps).

The Celera project uses a "whole genome shotgun" approach, in which the entire genome is shredded into small fragments that are sequenced and put back together on the basis of sequence overlaps.

The hierarchical shotgun method has the advantage that the global location of each individual sequence is known with certainty, but it requires constructing a map of large fragments covering the genome. The whole shotgun method does not require this step, but presents other challenges in the assembly phase.

"The two approaches are quite complementary. The public project and Celera plan to discuss the relative scientific merits of the methods employed by the two projects. In the end, the best approach may well be to use a combination of the methods for sequencing future genomes," said Francis Collins, Director of the National Human Genome Research Institute. In fact, current plans by the public project to sequence the genome of the laboratory mouse involve this hybrid strategy.

Next phase. The Human Genome Project will now focus on converting the draft and near-finished sequences to a finished form. This will be done by filling the gaps in the working draft sequence and by increasing the overall sequence accuracy to 99.99 percent. Although the working draft version is useful for the most biomedical research, a highly accurate sequence that is as close to perfect as possible is critical for obtaining all the information there is to get from human sequence data. This has already been achieved for chromosomes 21 and 22.

Human DNA variation. The greater-than-expected sequence production has also yielded a bumper crop of human genetic variations - called single nucleotide polymorphisms or SNPs. The Human Genome Project had set a goal of discovering 100,000 SNPs by 2003. Already, with today's assembled sequences and other data accumulated by The SNPs Consortium, scientists have now found more than 300,000 SNPs and will likely have 1 million SNPs by year end. These SNPs provide a powerful tool for studies of human disease and human history.

<Background>

Sequencing, which is determining the exact order of DNA's four chemical bases, commonly abbreviated A, T, C and G, has been expedited in the Human Genome Project by technological advances in deciphering DNA and the collaborative nature of the effort, which includes about 1,000 scientists worldwide working together effectively.

The Human Genome Sequencing Project aims to determine the sequence of the euchromatic portion of the human genome. The euchromatic portion excludes certain regions consisting of long stretches of highly repetitive DNA that encode little genetic information, and that are not recovered in thevector systems used by the genome project. Such regions account for about 10% of the genome, and are said to be heterochromatic. (For example, the center of chromosomes, called centromeres, consists of heterochromatic DNA.)

The international Human Genome Sequencing consortium includes scientists at 16 institutions in France, Germany, Japan, China, Great Britain and the United States. The five largest centers are located at： Baylor College of Medicine, Houston, Texas; Joint Genome Institute in Walnut Creek, CA; Sanger Centre near Cambridge, England; Washington University School of Medicine, St. Louis; and Whitehead Institute, Cambridge, Massachusetts. Together, these five centers have generated about 82% of the sequence. The attached table provides more detail about the 16 centers and their individual contributions to the Human Genome Project.

The project has been tightly coordinated so that no region of the genome is left unattended to, and duplication is minimized. Participants in the international consortium have all adhered to the project's quality standards, and to the daily data release policy. The project is funded by grants from government agencies and public charities in the various countries. These include the National Human Genome Research Institute at the US National Institutes of Health, the Wellcome Trust in England, and the US Department of Energy.

The total cost for the working draft is approximately $300 million worldwide, with roughly half ($150 million) being funded by the US National Institutes of Health. The cost of sequencing the human genome is sometimes reported as $3 billion. However, this figure refers to the original estimate of total funding for the Human Genome Project over a 15-year period (1990-2005) for a wide range of scientific activities related to genomics. These include studies of human diseases, experimental organisms (such as bacteria, yeast, worms, flies and mice), development of new technologies for biological and medical research, computational methods to analyze genomes, and ethical, legal and social issues related to genetics.

The sixteen institutions that form the Human Genome Sequencing Consortium include：

1.Baylor College of Medicine, Houston, Texas,
2.Beijing Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing, China
3.Gesellschaft fur Biotechnologische Forschung mbH, Braunschweig, Germany
4.Genoscope, Evry, France
5.Genome Therapeutics Corporation, Waltham, MA, USA
6.Institute for Molecular Biotechnology, Jena, Germany
7.Joint Genome Institute, U.S. Department of Energy, Walnut Creek, CA, USA
8.Keio University, Tokyo, Japan
9.Max Planck Institute for Molecular Genetics, Berlin, Germany
10.RIKEN Genomic Sciences Center, Saitama, Japan
11.The Sanger Centre, Hinxton, U.K.
12.Stanford DNA Sequencing and Technology Development Center, Palo Alto, CA, USA
13.University of Washington Genome Center, Seattle, WA, USA
14.University of Washington Multimegabase Sequencing Center, Seattle, WA, USA
15.Whitehead Institute for Biomedical Research, MIT, Cambridge, MA, USA
16.Washington University Genome Sequencing Center, St. Louis, MO, USA

In addition, two institutions played a key role in providing computational support and analysis for the Human Genome Project over the course of the past eighteen months. These include：

The National Center for Biotechnology Information at NIH
The European Bioinformatics Institute in Cambridge, UK

The assembly of the genome sequence across chromosomes was also assisted by scientists at the University of California, Santa Cruz, and Neomorphic,Inc

国際ヒトゲノムコンソーシアムがドラフトシーケンスの終了を発表（要約）

国際ヒトゲノムコンソーシアムは人間の遺伝設計図であるヒトゲノムのドラフト配列の決定と編集を終了したことをここに発表する。この大きな成果は　1）比較的大きく断片化したヒト DNA をゲノム全体にわたって整列化する、2）その断片の配列を決定する、の 2 つのステップによって達成された。

国際ヒトゲノムコンソーシアムではヒトゲノムの 97%を断片化したヒト DNA でカバーし、それをもとにヒトゲノム全体の 85%に相当する配列を決定した。

国際ヒトゲノムコンソーシアムのシーケンス決定能力はここ 1 年で急上昇し、データの 60%はここ半年に生産されたものである。我々はこの半年間、毎秒 1000 塩基のペースで、毎週 7 日間、1 日 24 時間、データを生産したことになる。

データの精度は当初の予想より高く、20%は完成データ、50%もほぼ完タである。連続した配列の平均のながさは 200、678 塩基である。国際ヒトゲノムコンソーシアムは世界の科学者、企業などにむけてデータを即時に公開し、その使用について制限を加えていない。その結果、これまでに数万の遺伝子が見い出され、そこから少なくとも 12 の疾患遺伝子が特定された。

国際ヒトゲノムコンソーシアムの目標 : 我々は 2000 年春までにヒトゲノムの 90%をカバーするドラフトシーケンスを、2003 年春までに全体を 99。99%以上の精度でカバーする完成シーケンスを発表することを目標とした。しかし現在のペースを考えると、それよりかなり早く完了すると予想される。

相補的な 2 つのアプローチ : 本日、セレーラ社よりヒトゲノム全体の第一次編集を終了したとの我々と関連する発表がおこなわれる。両者は類似のオートメイションと DNA シーケンステクノロジーを用いているが、アプローチは異なる。 国際コンソーシアムは DNA 断片をゲノム上に位置ずけ、それをショットガン法でシーケンス解析する”階層的ショットガン法”を、セレーラ社は全ゲノムを小断片化して、それをランダムにシーケンス決定し、最後に全体を編集する”全ゲノムショットガン法”を採用している。”階層的ショットガン法”はゲノム断片の位置を決める手間がかかるが、シーケンスの位置情報がわかる長所がある。一方、”全ゲノムショットガン法”では、大量のデータ生産が可能であるが編集過程が大きな挑戦である。ｔ2 つのアプローチは相補的で、両者は互いのメリットについて検討する計画である。両者を融合するのがベストである。”と国際コンソーシアムの代表の F. Collins は述べている。事実、公的機関は次のマウスゲノムの解析に 2 つの方法の融合を考えている。

次のフェーズ : 国際ヒトゲノムコンソーシアムでは今後、ドラフトシーケンスで残されたギャップをうめ、さらに精度をあげて 99.99%以上の精度で全配列を決定する。21、22 染色体では既にこれを終了している。

ヒトゲノムの多様性 : 予想以上のシーケンスデータが生産され、その結果きわめて大量の SNP（1 塩基多型）が生産された。ゲノム計画では 2003 年までに 10 万の SNP をみいだす予定であったが、既に 30 万を見い出し、今年の終わりまでに 100 万 SNP

を見い出すと予想される。これはヒトの病気や進化の研究の強力なツールとなる。

<背景>

国際ヒトゲノムコンソーシアムでは、ほぼ全ての遺伝子を含み、ヒトゲノムの 90%を占める真正クロマチンと呼ばれる領域を対象としている。残りの 10%はヘテロクロマチンと呼ばれ、くり返し配列が大半を占め、遺伝子がほとんど存在しない。セントロメアー（動原体）はその例である。

国際ヒトゲノムコンソーシアムは米、英、日、仏、独、中国の 16 のセンター、1000 人を超える科学者が参加している。米、英の 5 つのセンターが全体の 82%のデータを生産している。（日本は全体の 7%を生産し米、英に次いでいる。）

ドラフトシーケンスの経費は世界全体で 300 百万 USD と見積もられる。NIH がその半分をしめている。ヒトゲノム計画全体は 3000 百万 USD と見積もられる。

（訳：榊　佳之）