The Human Genome

The Sequence of the Human Genome

ヒトゲノムの塩基配列

J. Craig Venter, 1* Mark D. Adams, 1 Eugene W. Myers, 1 Peter W. Li, 1 Richard J. Mural, 1 Granger G. Sutton, 1 Hamilton O. Smith, 1 Mark Yandell, 1 Cheryl A. Evans, 1 Robert A. Holt, 1 Jeannine D. Gocayne, 1 Peter Amanatides, 1 Richard M. Ballew, 1 Daniel H. Huson, 1 Jennifer Russo Wortman, 1 Qing Zhang, 1 Chinnappa D. Kodira, 1 Xiangqun H. Zheng, 1 Lin Chen, 1 Marian Skupski, 1 Gangadharan Subramanian, 1 Paul D. Thomas, 1 Jinghui Zhang, 1 George L. Gabor Miklos, 2 Catherine Nelson, 3 Samuel Broder, 1 Andrew G. Clark, 4 Joe Nadeau, 5 Victor A. McKusick, 6 Norton Zinder, 7 Arnold J. Levine, 7 Richard J. Roberts, 8 Mel Simon, 9 Carolyn Slayman, 10 Michael Hunkapiller, 11 Randall Bolanos, 1 Arthur Delcher, 1 Ian Dew, 1 Daniel Fasulo, 1 Michael Flanigan, 1 Liliana Florea, 1 Aaron Halpern, 1 Sridhar Hannenhalli, 1 Saul Kravitz, 1 Samuel Levy, 1 Clark Mobarry, 1 Knut Reinert, 1 Karin Remington, 1 Jane Abu-Threideh, 1 Ellen Beasley, 1 Kendra Biddick, 1 Vivien Bonazzi, 1 Rhonda Brandon, 1 Michele Cargill, 1 Ishwar Chandramouliswaran, 1 Rosane Charlab, 1 Kabir Chaturvedi, 1 Zuoming Deng, 1 Valentina Di Francesco, 1 Patrick Dunn, 1 Karen Eilbeck, 1 Carlos Evangelista, 1 Andrei E. Gabrielian, 1 Weiniu Gan, 1 Wangmao Ge, 1 Fangcheng Gong, 1 Zhiping Gu, 1 Ping Guan, 1 Thomas J. Heiman, 1 Maureen E. Higgins, 1 Rui-Ru Ji, 1 Zhaoxi Ke, 1 Karen A. Ketchum, 1 Zhongwu Lai, 1 Yiding Lei, 1 Zhenya Li, 1 Jiayin Li, 1 Yong Liang, 1 Xiaoying Lin, 1 Fu Lu, 1 Gennady V. Merkulov, 1 Natalia Milshina, 1 Helen M. Moore, 1 Ashwinikumar K Naik, 1 Vaibhav A. Narayan, 1 Beena Neelam, 1 Deborah Nusskern, 1 Douglas B. Rusch, 1 Steven Salzberg, 12 Wei Shao, 1 Bixiong Shue, 1 Jingtao Sun, 1 Zhen Yuan Wang, 1 Aihui Wang, 1 Xin Wang, 1 Jian Wang, 1 Ming-Hui Wei, 1 Ron Wides, 13 Chunlin Xiao, 1 Chunhua Yan, 1 Alison Yao, 1 Jane Ye, 1 Ming Zhan, 1 Weiqing Zhang, 1 Hongyu Zhang, 1 Qi Zhao, 1 Liansheng Zheng, 1 Fei Zhong, 1 Wenyan Zhong, 1 Shiaoping C. Zhu, 1 Shaying Zhao, 12 Dennis Gilbert, 1 Suzanna Baumhueter, 1 Gene Spier, 1 Christine Carter, 1 Anibal Cravchik, 1 Trevor Woodage, 1 Feroze Ali, 1 Huijin An, 1 Aderonke Awe, 1 Danita Baldwin, 1 Holly Baden, 1 Mary Barnstead, 1 Ian Barrow, 1 Karen Beeson, 1 Dana Busam, 1 Amy Carver, 1 Angela Center, 1 Ming Lai Cheng, 1 Liz Curry, 1 Steve Danaher, 1 Lionel Davenport, 1 Raymond Desilets, 1 Susanne Dietz, 1 Kristina Dodson, 1 Lisa Doup, 1 Steven Ferriera, 1 Neha Garg, 1 Andres Gluecksmann, 1 Brit Hart, 1 Jason Haynes, 1 Charles Haynes, 1 Cheryl Heiner, 1 Suzanne Hladun, 1 Damon Hostin, 1 Jarrett Houck, 1 Timothy Howland, 1 Chinyere Ibegwam, 1 Jeffery Johnson, 1 Francis Kalush, 1 Lesley Kline, 1 Shashi Koduru, 1 Amy Love, 1 Felecia Mann, 1 David May, 1 Steven McCawley, 1 Tina McIntosh, 1 Ivy McMullen, 1 Mee Moy, 1 Linda Moy, 1 Brian Murphy, 1 Keith Nelson, 1 Cynthia Pfannkoch, 1 Eric Pratts, 1 Vinita Puri, 1 Hina Qureshi, 1 Matthew Reardon, 1 Robert Rodriguez, 1 Yu-Hui Rogers, 1 Deanna Romblad, 1 Bob Ruhfel, 1 Richard Scott, 1 Cynthia Sitter, 1 Michelle Smallwood, 1 Erin Stewart, 1 Renee Strong, 1 Ellen Suh, 1 Reginald Thomas, 1 Ni Ni Tint, 1 Sukyee Tse, 1 Claire Vech, 1 Gary Wang, 1 Jeremy Wetter, 1 Sherita Williams, 1 Monica Williams, 1 Sandra Windsor, 1 Emily Winn-Deen, 1 Keriellen Wolfe, 1 Jayshree Zaveri, 1 Karena Zaveri, 1 Josep F. Abril, 14 Roderic Guigo _, 14 Michael J. Campbell, 1 Kimmen V. Sjolander, 1 Brian Karlak, 1 Anish Kejariwal, 1 Huaiyu

Mi, 1 Betty Lazareva, 1 Thomas Hatton, 1 Apurva Narechania, 1 Karen Diemer, 1 Anushya Muruganujan, 1 Nan Guo, 1 Shinji Sato, 1 Vineet Bafna, 1 Sorin Istrail, 1 Ross Lippert, 1 Russell Schwartz, 1 Brian Walenz, 1 Shibu Yooseph, 1 David Allen, 1 Anand Basu, 1 James Baxendale, 1 Louis Blick, 1 Marcelo Caminha, 1 John Carnes-Stine, 1 Parris Caulk, 1 Yen-Hui Chiang, 1 My Coyne, 1 Carl Dahlke, 1 Anne Deslattes Mays, 1 Maria Dombroski, 1 Michael Donnelly, 1 Dale Ely, 1 Shiva Esparham, 1 Carl Fosler, 1 Harold Gire, 1 Stephen Glanowski, 1 Kenneth Glasser, 1 Anna Glodek, 1 Mark Gorokhov, 1 Ken Graham, 1 Barry Gropman, 1 Michael Harris, 1 Jeremy Heil, 1 Scott Henderson, 1 Jeffrey Hoover, 1 Donald Jennings, 1 Catherine Jordan, 1 James Jordan, 1 John Kasha, 1 Leonid Kagan, 1 Cheryl Kraft, 1 Alexander Levitsky, 1 Mark Lewis, 1 Xiangjun Liu, 1 John Lopez, 1 Daniel Ma, 1 William Majoros, 1 Joe McDaniel, 1 Sean Murphy, 1 Matthew Newman, 1 Trung Nguyen, 1 Ngoc Nguyen, 1 Marc Nodell, 1 Sue Pan, 1 Jim Peck, 1 William Rowe, 1 Robert Sanders, 1 John Scott, 1 Michael Simpson, 1 Thomas Smith, 1 Arlan Sprague, 1 Timothy Stockwell, 1 Russell Turner, 1 Eli Venter, 1 Mei Wang, 1 Meiyuan Wen, 1 David Wu, 1 Mitchell Wu, 1 Ashley Xia, 1 Ali Zandieh, 1 Xiaohong Zhu 1

全ゲノム・ショットガン塩基配列決定法を用いて、ヒトゲノムの真正染色質領域の 29.1 億塩基対の配列を決定しコンセンサス配列を作成したので報告する。5 人のヒト被験者 DNA からプラスミドクローン化ライブラリーを作製し、27,271,853 回の高品質な塩基配列解析をクローン両端から行い、全ゲノムの 5.11 倍に及ぶ 148 億塩基対の DNA 塩基配列を 9 ヶ月かけて決定した。アセンブリ戦略には 2 つの方法、すなわち全ゲノムアセンブリ法と染色体部分アセンブリ法を用いた。両法ともに、当セレラ社ならびに国際ゲノムプロジェクトが明らかにした配列データを組み合わせて用いている。既存の公開塩基配列データは、国際ゲノムプロジェクトが使用したクローニング法およびアセンブリ法に内在するバイアスを含めないように、また配列決定されているゲノム領域の 2.9 倍となるように、550 塩基対に断片化した。これにより、アセンブリ可能範囲が全ゲノムの 8 倍までに達し、5.11 倍の場合と比べ、最終アセンブリにおけるギャップ数およびギャップサイズを減少できた。2 つのアセンブリ戦略より得られた結果はよく似ており、独立したマッピングデータともほぼ合致した。従って今回のアセンブリ法は、ヒト染色体の真正染色質領域を有効にカバーするものである。ゲノムの 90%以上が 10 万塩基対以上のサイズの骨格アセンブリ中に見いだされ、25%が 1000 万塩基対以上のサイズの大型骨格アセンブリ中に存在する。ゲノム配列を解析したところ、確実に蛋白質をコードする 26,588 個の翻訳可能領域が存在していた。さらに、

1 Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. 2 GenetixXpress, 78 Paci_c Road, Palm Beach, Sydney 2108, Australia. 3 Berkeley Drosophila Genome Project, University of California, Berkeley, CA 94720, USA. 4 Department of Biology, Penn State Uni-versity, 208 Mueller Lab, University Park, PA 16802, USA. 5 Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. 6 Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287_4922, USA. 7 Rockefeller University, 1230 York Avenue, New York, NY 10021_6399, USA. 8 New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. 9 Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasa- dena, CA 91125, USA. 10 Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520_8000, USA. 11 Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. 12 The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. 13 Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. 14 Grup de Recerca en Informa `tica Me`dica, In-stitut Municipal d'Investigacio _ Me `dica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

連絡先：To whom correspondence should be addressed.
E- mail: humangenome@celera.com

約 12,000 個のコンピュータ推測に由来するマウス相同性遺伝子あるいは不十分であるが存在をうかがわせる遺伝子が見いだされた。遺伝子が密集するクラスターが複数個あることは明確であったが、およそ半数の遺伝子は、明らかに蛋白質をコードしない大きな配列によって分断されている低 GC 塩基配列中に分散していた。エクソンはゲノムの 1.1%に過ぎず、イントロンは 24%、遺伝子と遺伝子の間をつなぐ DNA がゲノムの 75%を占めていた。断片的なブロック配列が重複したものは、染色体全長にまで及ぶものを含め、全ゲノム中に豊富に存在し、複雑な進化の軌跡を残している。比較ゲノム解析から、神経機能、組織特異的発生制御、凝血機構および免疫機構に随伴して、脊椎動物が遺伝子を拡大して来たことが指摘できる。今回のコンセンサス配列と国際ゲノムプロジェクトによるゲノムデータを比較したところ、210 万種もの SNPs（single-nucleotide polymorphisms,単一塩基多型）が存在することが判明した。ヒト半数体ゲノムの任意の対を比較すると、平均して 1250 塩基に 1 塩基の割合で異なるが、全ゲノムを通し、多型性の程度で顕著な不均一性が認められた。全 SNPs 中 1%未満で蛋白質の変異体を生じるが、どの SNPs が機能に影響を与えるかを決定する作業は今後の課題として残っている。

序論

ヒトゲノムを構成している DNA の暗号解読作業は、ヒトの進化、疾患の原因究明、ヒトの存在条件を規定する環境と遺伝形質の相互作用を理解するために貢献できることから、幅広い期待を集めてきた。ヒトゲノムの全塩基配列決定という目標を掲げたプロジェクトが初めて正式に提唱されたのは、1985 年であった[(1)]。その後、このアイデアに対して科学界では賛否両論が展開された[(2)]が、1990 年には、米国で、米国立衛生研究所（NIH）と米エネルギー省の統率下、ヒトゲノム配列決定完了までに 15 年と 30 億ドルをかけるヒトゲノムプロジェクト（HGP）が公式に開始された。1998 年、我々は、独自のゲノム配列解析施設を建設し、3 年かけてヒトゲノム配列を決定する意向を表明した。今回、目標達成への最終段階、すなわちヒトゲノムの真正染色質のほぼ完全な配列を決定したので報告する。配列決定は、全ゲノムを断片化し、断片の塩基配列を決定した後、この断片をアセンブリするランダムショットガン法により行った。

DNA 配列決定の近代史は 1977 年に端を発する。この年、Sanger が、DNA 鎖終結（chain-terminating）ヌクレオチドアナログを用いて、DNA ヌクレオチドの順序を決定する方法を報告した[(3)]。同年、ヒト遺伝子が初めて単離され塩基配列が決定された[(4)]。1986 年、Hood ら[(5)]が Sanger 方式を改良し、ヌクレオチドに蛍光色素を付着させてコンピュータに逐次読み取らせることができる配列解析法を報告した。これを自動化した初の DNA シーケンサーは、カリフォルニア州のアプライドバイオシステム社が 1987 年に開発し、この新技術で 2 つの遺伝子の配列解読に成功したことが示された[(6)]。ヒトゲノムの部分領域の配列解読

[7]が開始された初期から、ゲノム中の遺伝子の存在予測のために、配列注釈をつけ、確認評価するには、cDNA（RNA から逆転写された相補鎖 DNA）塩基配列が必要欠くべからざるものであることが明らかになっていた。これらの研究は、一部は遺伝子同定のための発現配列タグ（Expressed sequence tag: EST）法を開発する基盤となった[8]。EST 法は、任意に断片を選択し、高速塩基配列決定を行うことによって、cDNA ライブラリーの特徴を明らかにする方法である。この EST 法により、ヒト遺伝子の迅速な発見とマッピング（地図作成）が実現した[9]。さらにヒト EST 配列の数が増えるにつれ、大量の配列データを解析する新しいコンピュータルゴリズムの開発も必要になった。そこで 1993 年、ゲノム研究所（The Institute for Genomic Research: TIGR）で、何十万もの EST をアセンブリさせ解析できるアルゴリズムが開発された。これによって、3 万個の EST アセンブリを基に、ヒト遺伝子を特性づけ、配列注釈をつけることが可能となった[10]。1982 年、ショットガン制限酵素消化法を用いて、49 kbp に及ぶバクテリオファージラムダの完全ゲノム配列が決定された[11]。その後 1991 年、痘瘡ウイルスのゲノム配列決定[12]にあたって全ゲノムショットガン法が検討されたが、ゲノムアセンブリに適したソフトウェアがまだなかったため、却下された。しかし 1994 年、微生物のゲノム配列決定プロジェクトが TIGR で企画された時には、TIGR の EST アセンブリアルゴリズムを併用すれば、全ゲノムショットガン法は使用可能とみなされた。そして 1995 年、全ゲノムショットガン法を用いて、1.8 Mbp のインフルエンザ菌（*Haemophilus Influenzae*）のゲノム解読が完了した[13]。その後いくつかのゲノムの配列決定が行われ、本法の汎用性が確立した[14,15]。

このようなメガ塩基対サイズ以上の大きな遺伝子の塩基配列決定法のキーポイントは、paired-end 配列（メイト対とも言う）を用いることである。この相補的塩基対をなす末端配列は、挿入サイズとクローニング特性がそれぞれ異なるサブクローンライブラリーに由来し、適当な長さに処理した二重鎖 DNA クローンの両端から 500～600bp の配列である。バクテリオファージラムダにクローニングした DNA の長い断片（18～20kbp）由来の末端配列を用いて、微生物ゲノムのアセンブリに成功したことから、150 kbp の細菌人工染色体（BAC）[17,18]に由来する末端配列を用いれば、ヒトゲノムをマップすると同時に配列決定できる筋道があることが示唆された[16]。長さが判っている配列の末端配列どうしをつないでいけば、ゲノム全体がつながる長距離連続性を与えてくれる。BAC 末端配列決定（BES）法の改良法により、シロイヌナズナ *Arabidopsis thaliana* 第 2 染色体の解読が成功裏に完了している[19]。

1997 年、Weber と Myers[20]が、ヒトゲノムに対して全ゲノムショットガン塩基配列決定法の適用を提案した。2 人の提案はからなずしも歓迎されたわけではない[21]。しかし 1998 年の初頭までには、5%に足らぬゲノムしか配列決定されていないことから、全世界におけるヒトゲノム塩基配列決定の進捗度は極めて緩慢であり[22]、目標の 2005 年までにゲノム解読を

終了できる見込みは薄かった。

1998 年早々、PE バイオシステムズ (現アプライドバイオシステムズ)社は、自動化高速キャピラリーDNA シーケンサーを開発し、やがて ABI PRISM 3700 DNA アナライザーと名づけた。PE バイオシステムズ社と TIGR の研究陣は検討を重ねた結果、この 3700 アナライザーと TIGR で開発された全ゲノムショットガン法を用いて、ヒトゲノム配列決定を行う計画を立てた[23]。ゲノム配列解析施設における作業原則の多くは、TIGR 側で確立された[24]。しかし、我々Celera 社が夢に描いた解析施設は、TIGR の約 50 倍もの性能を有するものであり、従って試料の調製とトラッキング、および全ゲノムアセンブリ法に新規開発が求められた。*H. influenzae* のゲノムに比べて複雑な反復配列を有するヒトゲノムを解読するには、150 倍のスケールアップが必要であり、それは実現不可能だと論じる者もいた[25]。そこで、大きくて複雑な真核生物のゲノム上で全ゲノムアッセンブリを行えるかどうかの予備試験に、まずキイロショウジョウバエ（*Drosophila melanogaster*）を選んだ。Gerald Rubin とバークレーのショウジョウバエゲノムプロジェクトと協力しながら、ショウジョウバエゲノムの 120-Mbp 真性染色質部の塩基配列を 1 年かけて決定した[26-28]。この結果、重要な知見が 2 点明らかになった。その（i）は、アセンブリアルゴリズムによって染色体アセンブリが極めて正確な順序で行え、かつ実質的に全ゲノムの 10 倍以下の配列があれば、この方法で整列化が可能なこと、その（ii）は、包括的な最終アセンブリの代りに中間アセンブリを何回も行っても意味がない、ということであった。

これらの知見のためばかりでなく、Celera 社のヒトゲノム解読計画に続いて国際ゲノムプロジェクトに大変化が生じたこともあって[29]、我々はヒトゲノムへの全ゲノムショットガン法の適用計画を変更する事にした。すなわち、我々は当初、3 年かけてヒトゲノムの 10 倍の配列をカバーし、中間アセンブリした配列データを 4 分の 1 づつ提供する予定であったが、新しい計画では、ランダムショットガン法で約 5 倍の配列決定を行った後、順序づけも整列化もしていない BAC クローン化ライブラリーの配列断片のデータと国際ゲノムプロジェクト[30]が GenBank に公表したサブアセンブリの結果を用いて、本プロジェクトを加速遂行することにした。さらに、中間アセンブリがないため、4 分の 1 づつ公表することも放棄した。

このやり方は、全ゲノムショットガン法で 8 倍の配列カバー分をアセンブリした場合の結果と一致するような妥当な結果を速やかにもたらしたものの、13 倍の効果的な配列カバーでショウジョウバエゲノムが完結したようには、ヒトゲノム塩基配列決定はいかなかった。しかし、この配列カバー倍数を縮小した戦略でも、1 年未満でセレラ社は正確な順序付け・整列化した骨格配列を作成できることが明らかになった。かくて、ヒトゲノム配列決定作業を 1999 年 9 月 8 日に開始し、2000 年 6 月 17 日に完了した。同年 6 月 25 日に初回のアセ

ンブリが完了し、今回ここに発表するアセンブリは 2000 年 10 月 1 日に完了したものである。ここに我々はヒトゲノムに適用した全ゲノムランダムショットガン法について記述する。我々はホモサピエンスゲノムの 23 対の染色体を構成している約 30 億塩基対の配列をアセンブリするために、2 つの異なる方法を開発した。GenBank 由来のデータはいずれも細かく断片化し、キメラクローンや外来 DNA の夾雑物、アセンブリを間違えた連結（コンティグ）などから生じる配列偏向を、最終配列から除去した。ヒトの遺伝暗号を正確に解析するためには、ゲノム配列が間違うことなく正確に組み立てられ、コンティグが忠実な順序と整列化方向を示すことが必須条件である。そのため、本稿で我々はゲノム再構築の「品質」を立証することに細心の注意を払った。また、コンピュータによる算出法に基づいたヒト遺伝暗号の予備的解析結果についても述べる。図 1（本号についている折込ページ参照）（各染色体のファイルは、Science Online の fig. 1 [www.sciencemag.org/cgi/content/full/291/55071/304/DC2]に掲載）に、ゲノムの概観図とコードされている各遺伝子群の特性をまとめて示した。ゲノムに関する詳細な手引きと解釈は始まったばかりである。（図 1 を本文にリンクさせてください）特定の解析セクションを読者が見つけ出しやすいように、本論文は 7 つのセクションに大別し、各セクションの初めに主な結果の要約を載せた。（各章の目次）

第 1 章 DNA 供給源と塩基配列解析法

要約

ここでは、DNA の抽出とライブラリー構築のための方法と共に、人種差・性差を超えて多様性が確保できるドナー選択に関する合理的・倫理的な規則について論じる。ショットガン法の重要な第一段階は、プラスミドライブラリーの構築である。DNA ライブラリーが一様なサイズでなく、キメラ（起源が異なる）でなく、任意にゲノムを代表していないのであれば、この後いくら段階を重ねても、ゲノム配列を正確に再構築することはできない。我々は、高速大量処理可能な自動化 DNA 塩基配列解析装置と膨大な配列情報量（2730 万個の配列読み取り；149 億個の塩基配列）を効率よく追跡できる大型コンピュータを用いた。コンピュータにゲノムを再構築するには、2、10、50kbp の各ライブラリーから得たプラスミドクローンの両端から、配列を順次解析・追跡していかねばならない。我々の結果から、末端配列の正確な対形成率（pairing rate）は 98%を上回ることが示唆された。

米国ならびに世界医学会のさまざまな政策・方針、中でもヘルシンキ宣言は、ヒトを被験者とする実験を行なうにあたって勧告を出している。そこで我々は、機関内倫理委員会（Institutional Review Board: IRB）[31]を召集した。IRB は、ヒト DNA を採取し使用するためのプロトコールの設定、ならびに今回の DNA 配列解析研究に参加してくれる研究ボランティアからインフォームドコンセントを得る過程を確立することができるよう支援してくれた。被験者（ドナー）のプライバシーと秘密を守るためには、いくつかの措置・手順をとった。例えば、2 段階の同意方式をとったこと、標本・記録用にアルファベットと数字を組合わせた無作為抽出の安全なコードシステムを採用したほか、研究者に対して被験者との接触を制限したこと、ドナーからの連絡は現場以外でも随意に任せたことなどがある。さらに我々は、米厚生省に申請して機密性証明書（Certificate of Confidentiality）の交付を得た。この証明書は、公衆衛生法 42 U.S.C.241(d)のセクション 301(d)に記載されているように、自らの自由意志でボランティアになった個人のプライバシーを守る権限をセレラ社に与えるものである。

セレラ社と IRB は、解読が完了したヒトゲノムの第 1 号は、多様な人種背景を持つ多数のドナー由来のものの混成体であってしかるできであると信ずるものである。期待されるドナーには、自由意志に基づき、自分が属する民族地理的範疇を自ら明らかにしてほしい旨、依頼した（すなわち、アフリカ系米国人、中国人、ラテンアメリカ系米国人、白人、など）。今回は、21 人のドナーが参加した [32]。

各ドナーから得た 3 つの基本情報項目--年齢、性別、自称の民族地理的範疇--を記録し、供与標本との関係は機密コード化した。女性からは、約 130ml のヘパリン添加全血を採取し

た。男性からは、同じく 130ml までのヘパリン添加全血を採取したが、さらに 5 つの精液標本を 6 週にわたり採取した。Epstein-Barr ウイルス不死化法により、リンパ芽球細胞株を樹立した。ゲノムの DNA 配列決定のために、5 人の被験者（図 2）由来の DNA を配列決定のために選択した。内訳は男性 2 人、女性 3 人ーアフリカ系米国人 1 人、アジア系中国人 1 人、ラテンアメリカ系メキシコ人 1 人、白人 2 人ーであった（*Science* Online の fig. 2 [www.sciencemag.org/cgi/content/full/291/5507/1304/DC1]を参照のこと）。誰の DNA を配列解析するかについては、DNA ライブラリーの品質や樹立細胞株の使用可能量などテクニカルな問題はいうまでもなく、多様性を達成するという目標も含め、錯綜する複雑な要因を踏まえて決定した。

表 1. Celera 社が作製したアセンブリ・インプットデータ

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage	A	0	0	0.52	0.52	
(2.9-Gb genome)	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

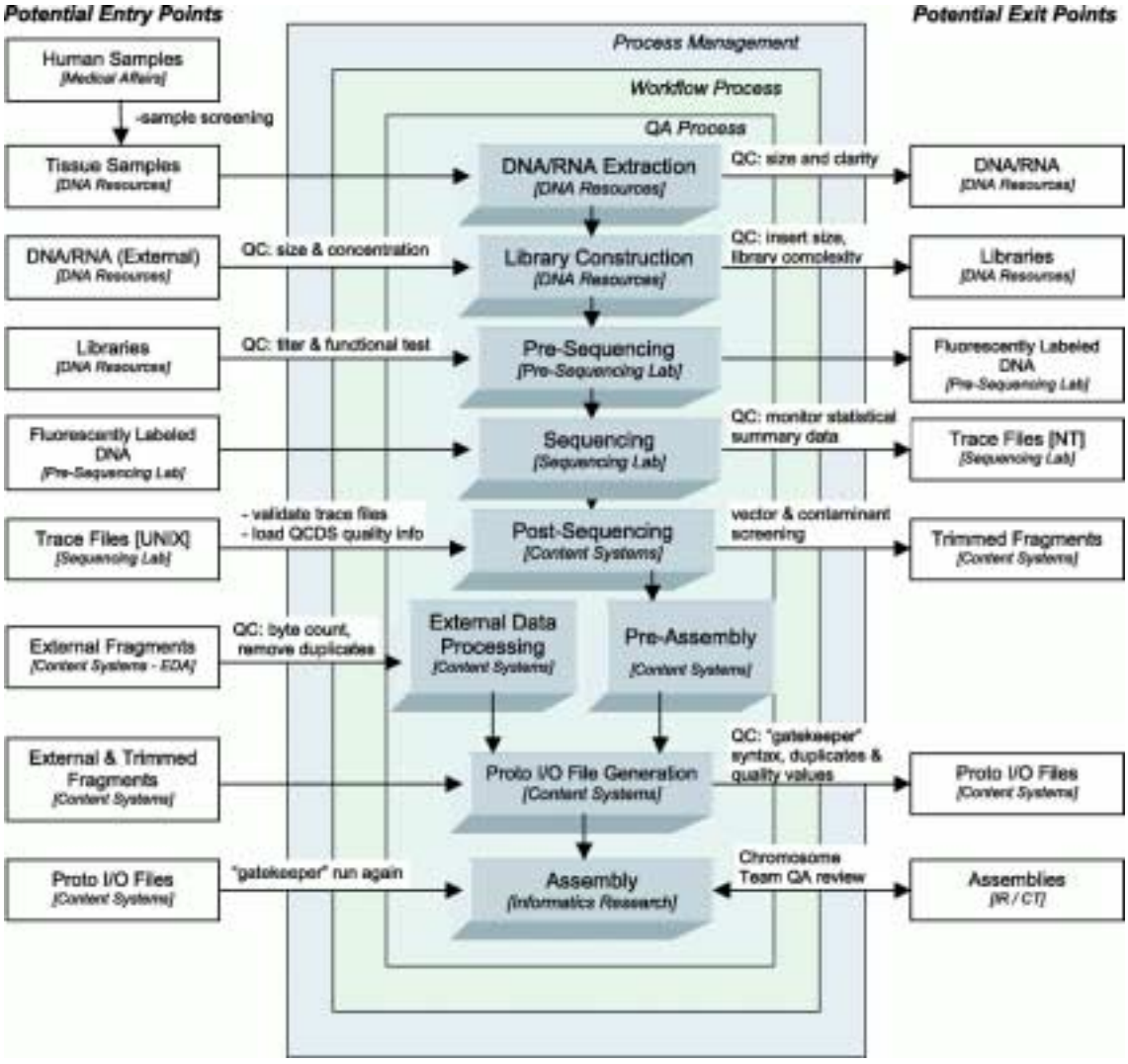
* Insert size and SD are calculated from assembly of mates on contigs.

† % Mates is based on laboratory tracking of sequencing runs.

1.1. ライブラリーの構築と配列解析

全ゲノムショットガン法の要は、変化に富むサイズの挿入片を含む高品質プラスミドライブラリーを作製することにある。そうすると、つき合わせ可能な配列（メイト対）が得られ、各プラスミド挿入片の両端からオーバーラップしたクローンを 1 つづつ読み取れるのである。高品質のライブラリーは、ゲノムのあらゆる領域から出てくる断片を偏りなく含んでおり、挿入片のないクローンが少なく、ミトコンドリアゲノムや大腸菌（*Escherichia coli*）ゲノムの DNA 由来の夾雑物がない。我々は各ドナー由来の DNA から、3 つのサイズ、すなわち 2kbp、10kbp、50kbp のグループのプラスミドライブラリーを構築した(表 1) [(33)]。

図 2. 配列解析パイプラインのフローチャート
SOP（標準操作手順）を遵守し、各部署内および全体を通して品質に焦点をおき、サンプルの受け取り、選別、プロセシングを行なう。各工程で、規定した品質ガイドラインに従って、サンプルとデータを内部のものと外部のもので交換できるようなインプットとアウトプットを規定した。製造パイプライン工程、製品（データ）、品質管理措置、および関係責任者を示し、本文中でさらに言及した。[拡大像（55K GIF ファイル）]



DNA 塩基配列解析工程をデザインする際に、我々は、信頼性と再現性のある方法で実施でき、モニターも効果的にできるシンプルなシステムを開発することに焦点を置いた（図 2）[34]。

現在行なわれている配列解析プロトコールは、ジデオキシ配列解析法 [35] に基づいており、普通は 1 回の反応で 500～750bp の塩基配列しか読めない。読み取り配列長がこのように制限されてしまうことから、真核生物の大型ゲノム解析には、高速処理量を記念碑が立つほど大躍進させることが不可欠であった。これを我々はセレラ社の施設内に広さ約 3 万平方フィートも占拠する研究室を設けて達成した。総読み取り速度は 1 日あたり 175,000 回で、連続的に配列データを吐き出す。この DNA 塩基配列解析施設は高性能コンピュータ施設で支えられている [36]。

DNA 塩基配列解析の過程はモジュール方式とし自動化した。モジュール間にサンプルの未処理分を置くことで、4 つの主要モジュールがそれぞれ独立して操作できた。4 つの主要モジュールは（i）ライブラリー形質転換、プレート作製、コロニー採集、（ii）DNA 鋳型作製、（iii）ジデオキシ法反応設定と精製、（iv）ABI PRISM 3700 DNA アナライザーによる塩基配列決定である。各モジュールのインプット量とアウトプット量を注意して合わせ、未処理分は連続調節したため、配列解析作業は、1999 年 5 月にショウジョウバエのゲノム解読プロジェクトを開始して以来、1 日も中断することなく進行した。ABI 3700 は全自動キャピラリーアレイシーケンサーで、これ自体は最少の手作業時間（現在は一日あたり 15 分程度と推定）で操作することができる。キャピラリーシステムでは、手動によるサンプル添加やスラブゲル（slab gel）の場合にあったレーントラッキングエラーがなくなったため、サンプルと塩基配列解析トレースの正確な連結が可能である。約 65 人の生産スタッフを雇用・訓練し、4 つの生産モジュール間を定期的に輪番制で移動させた。研究室情報中央管理システム（LIMS）により、独自のバーコード識別子によって全てのサンプルプレートを追跡した。この研究施設を支えたのは、原材料および中間過程検査をおこなう品質管理チームと文書管理、バリデーション、施設監査等を担当する品質保証グループである。スケールアップ成功のために緊要であったのは、スケールアップ実施前にソフトウェアと機器のバリデーションを行ない、工程を変更した場合は製造スケールで検査したことである。

1.2. トレースプロセシング

自動化トレースプロセシング・パイプラインは各塩基配列ファイルを処理するために開発したものである [37]。データの品質管理のため、ベクター由来の配列をトリミング（端除去）した。トリミングした後の配列の平均長は 543bp であった。このため配列解析精度は急激に高まって平均 99.5%となり、98%以下の精度となったのは 1000 回の読み取りのうち 1 回

未満である [26]。トリミングした各配列は、ベクターのみの配列、あるいは *E. coli* ゲノムの DNA 配列、ヒトミトコンドリアの DNA 配列など夾雑物と一致しないかどうかスクリーニングした。いずれかの夾雑物とあきらかに一致した配列は、読み取った全てを破棄した。計 713 個の読み取り結果が *E. coli* ゲノムの DNA 配列と一致し、2114 個の読み取り結果がヒトミトコンドリアのゲノムと一致した。

1.3. 品質評価とコントロール

配列データの塩基対レベルでの正確さは、解析対象のゲノムのサイズと反復配列などが増えてくるに従って重要になってくる。各読み取り配列は、ゲノム内で独自の位置を占めねばならない。それほどではないエラー率であっても、アセンブリの有効性を減少させてしまう。さらにつき合わせ可能な配列（メイト対）の情報の正しさを維持することが、下記のアルゴリズムに絶対欠かせない。配列解析反応が各工程を踏んで進んでいくに従って、メイト対配列の正当性を維持するための手順管理法も確立した。これには LIMS のなかに厳重な規則を組み込んだことも含んでいる。セレラ社式工程で作成した配列データの精度は、ショウジョウバエゲノム・プロジェクトの進行過程で評価した [26]。単一の研究施設内で全ヒトゲノムのデータを採取することにより、われわれは均一の品質規格を確保し、自動化・スケール経済性・工程一貫性に伴うコスト効果をあげることができた。

第 2 章 ゲノムアセンブリ戦略と特徴

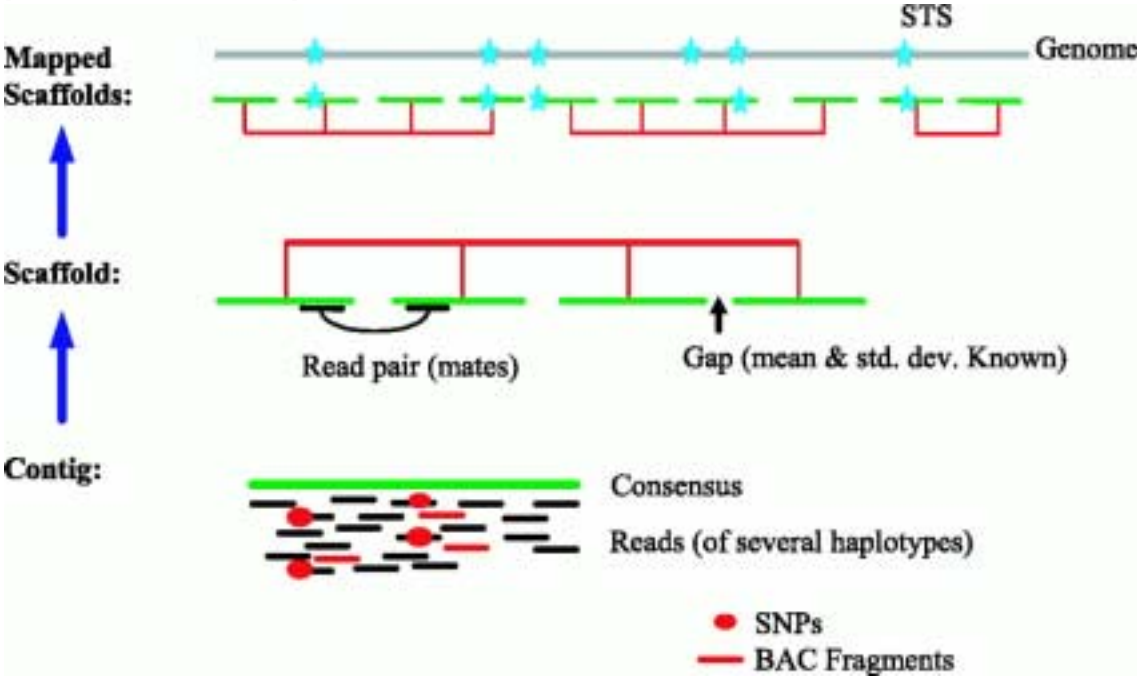
要約

ここでは、ゲノムをアセンブリするために用いた 2 つの手法について述べる。第一の方法では、全配列の読み取り結果と GenBank の断片化されたデータとを、コンピュータにより組み合わせ、一つの独立した偏りのないゲノムの概観を作成する。第二の手法では、マッピング情報に基づいて、全ての断片を、ある領域または染色体にまとめ、次に、まとめたデータを断片化して、コンピュータによるアセンブリを行なう。これらの方法では、正しい順序と正しい方向性でアセンブリされた DNA 塩基配列が、基本的には同じように再構築される。第二の手法では、わずかに大きい配列包括度（sequence coverage）（大きいほどギャップが少なくなる）が得られ、ゲノム解析段階で用いる主要な配列となった。また次に、このアセンブリ過程の完全さと正確さについて論じ、主に独立した BAC ごとのアプローチ（BAC-by-BAC approach）により再構築された国際ゲノムプロジェクトのゲノム配列との比較を行う。我々が行ったアセンブリは、ヒト染色体の真正染色質領域を効率よくカバーしている。ゲノムの 90%以上が、10 万 bp 以上のアセンブリ骨格中にあり、25%が 1000 万 bp 以上のアセンブリ骨格の中にあった。

ショットガン法によるアセンブリは遡行問題（inverse problem）をはらむ典型的な例である。すなわち、解析標的とした配列から無作為にサンプリングした 1 セットの読み取り配列が与えられた時、標的配列の中の順序と位置を再構築せよというわけである。キイロショウジョウバエのために開発したゲノムアセンブリアルゴリズムは、今や約 25 倍も大きいヒトゲノム解析向けに拡張されている。セレラ社のアセンブリは、1 セットのコンティグ（連結断片）からなり、これが配列骨格に順に並べられ方向付けされた後、既知のマーカーによって染色体の位置にマッピングされるのである。コンティグは、読み取り配列断片の集まりであり、配列重複部分があるために、ゲノムの隣接区間のコンセンサス配列を再構築する土台になる。メイト対は、アセンブリ戦略の中心的な構成要素であり、連続したコンティグ間のギャップサイズがかなり精密にわかっているアセンブリ骨格を作成するのに用いる。これは、一方が 1 つのコンティグの中にあり、もう一方が別のコンティグの中にある一対の読み取り配列が、二つのコンティグ間の方向性と距離を暗示していることを観察することにより達成できる（図 3）。最終的には、我々のアセンブリでは、報告した最終セットのアセンブリ骨格に、読み取った配列を全て組み込んだわけではない。組み込まなかった読み取り配列は「チャフ（chaff、おじゃまむし）」と名付けた。チャフの典型例は、高反復領域内から出てきた読み取り配列、多数のゲノムプロジェクトで認められていた他生物由来の配列で様々な経路で持ち込まれたもの、および品質の低い配列やトリミングされていないベクターの配列である。

図 3 全ゲノムアセンブリの解剖図

オーバーラップした断片化 bactig フラグメント（赤線）と、5 人からセレラ社内で得られた読み取り配列（黒線）を組み合わせ、コンティグとコンセンサス配列を作る（緑線）。メイト対の情報を使って、コンティグをつなげてアセンブリ骨格（赤）を作る。アセンブリ骨格は、STS（青星）物理的マップ情報により、ゲノム（灰色線）にマッピングされる。[拡大像（13K GIF ファイル）]



2.1. アセンブリデータセット

アセンブリには、二つの独立したデータセットを用いた。第一のデータセットは、セレラ社で作成した平均長 543bp の、272.7 万個の読み取り配列からなるランダムショットガン・データセットである。これは主に、5 人のドナー由来の DNA サンプルから作成した 16 個のライブラリーのメイト対読み取り配列で構成される。挿入配列サイズが 2、10、50kbp のものを用いた。1 つのライブラリーに由来するメイト対が、ゲノム内の既に配列決定された区域にどのように位置しているかを知ることで、各ライブラリーにおける挿入配列サイズの範囲の特徴がわかり、その平均値と標準偏差を決定することができた。表 1 に、読み取り配列の数、配列カバー倍数、データセットにより達成したクローン配列カバー倍数（clone coverage）の詳細を示した。クローン配列カバー倍数は、両端からの配列を持つ各クローンの挿入片全体を考慮したものだが、クローニングされた DNA 中のゲノムの配列カバー倍数といえる。ひいては、ゲノムの物理的 DNA カバー倍数（DNA coverage）を計る尺度となる。ゲノムサイズを 2.9Gbp と仮定すると、セレラ社のトリミングした配列はゲノムの 5.1 倍であり、クローン配列カバー倍数は 2、10、50kbp ライブラリーのそれぞれ 3.42、16.40、18.84 倍、総計 38.7 倍であった。

第二のデータセットは、公的資金によって展開された国際ヒトゲノムプロジェクト

（PFP）から得たもので、主に BAC クローンを塩基配列解析に用いて決定したものである[30]。アセンブリに入力したこの BAC データは、2000 年 9 月 1 日に GenBank からダウンロードして得たもので（表 2)、全長 4443.3 Mbp の配列であった。各 BAC のデータは、その完成度によって 4 つのフェーズに分類される。フェーズ 0 のデータは、BAC クローンをごく軽いショットガン法にかけて得たものであり、一般的にはアセンブリされていない読み取り配列セットで、全ゲノムの 1 倍未満を特徴としている。フェーズ 1 のデータは、BAC コンティグまたは”bactig”と呼んでいるコンティグが順番に並んでいない（unordered）集合体であり、フェーズ 2 のデータは順番に並んだ（ordered）bactig の集合体である。フェーズ 3 のデータは、完全な BAC 配列である。過去 2 年間、PFP は、各 BAC クローンの 3〜4 倍の配列カバー倍数の軽ショットガン法データから得られるフェーズ 1 データの作成に専念してきた。時間的にはより早くできるが品質と完成度は低いデータの作成になってしまっている。

表 2. アセンブリに入力した GenBank のデータ

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
Washington University, USA	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
Baylor College of Medicine, USA	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867

The Institute of Physical and Chemical Research (RIKEN), Japan	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Sanger Centre, UK	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
Others*	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
All centers combined†	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

* Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare, Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Southwestern Medical Center, University of Washington.

†The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96× coverage of the genome.

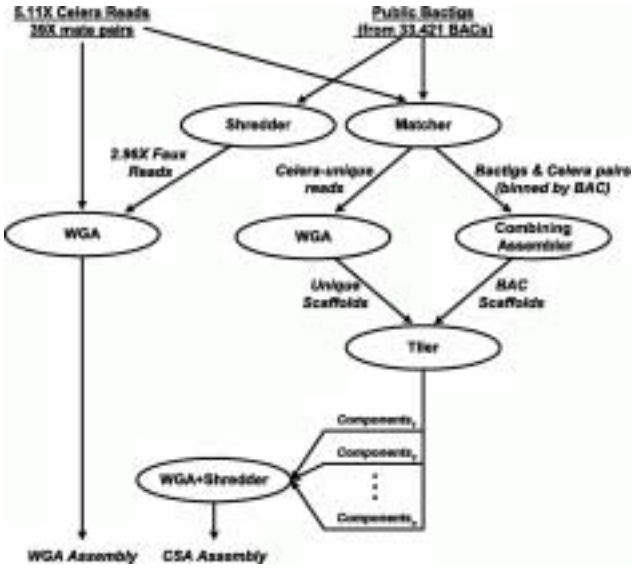
我々は、以下の 3 つのデータセットに対して、BLAST アルゴリズムを用いて、bactig 配列の夾雑物をスクリーニングした。（ｉ）Univec core のベクター配列 (38) について：98%の同一性を持つ配列の配列末端で 25bp が一致するか、配列内部で 30bp が一致するか、（ii）GenBank のハイスループットゲノム配列部門（High Throughput Genomic（HTG）Sequence division）のヒト以外の生物由来の配列 (39) について： 98%同一性を持つ配列で 200 bp が一致するか、（iii）霊長類ウイルスやヒトウイルの侵入遺伝子を含まない GenBank の非反復塩基配列について：98%同一性を持つ配列で 200 bp が一致するか、でふるい分けた。コンティグの末端 50 bp 以内に 25 bp 以上のベクター配列を認め次第、ベクターの部分まで末端を削除した。これらの基準の下で、我々は、夾雑物やベクターと考えられる配列を、フェーズ 3 のデータから 2.6 Mbp 分、フェーズ 1 と 2 のデータから 61.0 Mbp 分、フェーズ 0 のデータから 16.1 Mbp 分削除した（表 2）。これにより、計

4363.7 Mbp の PFP 配列データ（20%が完全状態、75%が粗稿状態（フェーズ 1 と 2）、5%が一回だけの読み取り配列（フェーズ 0））が残った。さらに、104,018 対の BAC 末端配列メイト対をダウンロードし、二つのアセンブリ処理データセットに加えた [18]。

2.2. アセンブリ戦略

2 種類のアセンブリ法を用いた。第一の手法は、セレラ社のデータと PFP データをあわせて合成ショットガンデータとして用いる全ゲノムアセンブリプロセス（whole-genome assembly process）である。第二の手法は、最初にセレラ社と PFP のデータを、大きな染色体分節に局在している組に分けて、各組ごとに最初からショットガンアセンブリを行う区画化アセンブリプロセス（compartmentalized assembly process）である。図 4 にこれらの工程の全体的な流れを示す。

図 4 セレラ社の 2 分岐アセンブリ戦略
長円形はそれぞれのラベルで示される機能を実行するコンピュータプロセス（訳注：本文中では例えばマッチング機などと記載）を示す。長円形間の矢印上のラベルは、プロセスによって作り出される、または使われるオブジェクトの性質を示す。この図は本文中の議論をまとめたものである。使われている用語と成句は、本文中に定義されている。[拡大像（26K GIF ファイル）]



全ゲノムアセンブリでは、PFP データを、最初に、完全に bactig の 2 倍分カバーする 550 bp の読み取り配列の合成ショットガンデータセットに分解または断片化した。これにより、BAC データセットのなかに配列重複があるために、結果的にゲノムを 2.96 倍カバーするのに十分な、1605 万個の「人造（faux)」読み取り配列ができた。しかし PFP アセンブリプロセスに固有の偏りは組み込んでいない。次に、4332 万個の読み取り配列を合わせたデータセット（ゲノムの 8 倍）と全ての関連したメイト対情報に対して、我々の全ゲノムアセンブリアルゴリズムを適用し、ゲノムの再構築を行った。ゲノム内の BAC の位置も、bactig のアセンブリも、このプロセスでは用いなかった。bactig うち 2.13%のアセンブリが間違っているという有力な証拠を見いだしたため [40]、bactig は読み取り配列断片長にまで断片化した。その上、PFP の出した物理地図上に正しく配置されていない BAC がいくつかあったこと、ならびに、おそらくサンプル追跡ミスの結果（後述)、少なくとも 2.2%の BAC が、その BAC のものではない配列データを含んでいるという有力な証拠を見いだしたこと [41] から、BAC 位置情報は無視した。要するに、我々は、外部で作成されたデータからはちょっとばかり良い配列カバー倍数を得ただけであって、メイト対情報やアセンブリされた bactig 情報、ゲノムの位置情報は使わず、本当に最初から全ゲノムアセンブリを行ったのである。

区画化ショットガンアセンブリ（compartmentalized shotgun assembly, CSA）では、セレラ社と PFP のデータを、自信を持って決定できる最も大きい染色体分節と思われるもの、すなわち「コンポーネント（区画)」に分割し、この区画化したサブセットに対してショットガンアセンブリを適用した。この場合、そのコンポーネント個別の、最初からのアセンブリであるということを確実にするため、bactig データを再び人造読み取り配列断片長に断片化した。データをこの方法で分けることによって、全体的なコンピュータ作業が削減され、染色体間の重複の影響が改善された。また、結果としては、全ゲノムアセンブリ法とは相互に独立したゲノムの再構築を行うことになり、一致しているかどうか 2 つのアセンブリを比較することができた。異なるゲノム領域が混らないようにするには、コンポーネント分割の質（quality）が重要である。われわれは、このコンポーネントを、（ｉ）各 BAC から得た最長の配列骨格からと、（ii）セレラ社データセット特有のアセンブリ骨格から作成した。BAC アセンブリは、bactig とこれらの bactig にマッピングした 5 倍数のセレラ社データをインプットデータとして用い、結合アセンブラ（combining assembler）を使って得たものである。この作業は、中間段階として行った。所定の配列に対してアセンブリ骨格が正確で完全であればあるほど、配列重複部位とメイト対情報に基づき、これらの骨格を一層正確な連続したコンポーネントとして並べることができるからという理由による。我々はさらに正確さを増すために、コンポーネント中のアセンブリ骨格の並び方を視覚的に探究し整理した。最終的な CSA アセンブリでは、区画化以外の件は全て無視し、区画化した該当セレラ社データと、区画化

した該当 bactig データを断片化した人造読み取り配列断片に対して、我々の全ゲノムアセンブリアルゴリズムを適用した結果、各コンポーネントで個別に当初段階からの配列再構成を行うことができた。

2.3. 全ゲノムアセンブリ

ヒトゲノムの全ゲノムアセンブリ（WGA）に用いたアルゴリズムは、キイロショウジョウバエゲノムの配列決定［詳細な報告は（28)］に用いた方法を拡大した方法である。

WGA アセンブラは、5 つの主要な段階、すなわち、ふるい機（Screener)、オーバーラップ検出装機（Overlapper)、ユニティグ検出機（Unitigger)、アセンブリ骨格形成機（Scaffolder)、反復配列解析機（Repeat Resolver）で構成されるパイプラインからなる。ふるい機（Screener）は、6-bp 未満のエレメントで構成されるマイクロサテライト反復配列を全て発見してマークし、Alu、Line、リボゾーム DNA など散在する既知の反復配列を全てふるい落とす。残りのマークされた領域はオーバーラップ（塩基配列一致）の有無が検索される。ふるい落とされた領域はオーバーラップ検索をされることはないが、オーバーラップの一部でふるいに引っかからないがつき合わせれば一致するセグメントを含む可能性もある。

オーバーラップ検出機（Overlapper）は、1 つひとつの読み取り配列を他の全読み取り配列と比較し、最短でも 40bp の、端から端まで完全なオーバーラップ（塩基配列一致）がある配列と、つき合わせ配列どうし中の差異が 6%以下しかないものを検出する。全てのデータは、綿密にベクターをトリミングしているので、オーバーラップ検出機は、完全にオーバーラップする配列があるぞあるぞと主張できる。全てのオーバーラップセットを計算するのに、4 ギガバイトの RAM を搭載した 4 プロセッサー Alpha SMP1 台では、約 10,000CPU 時間かかる。このような機械を 40 台並行して作動させたところ、実所要時間は 4〜5 日となった。

上述のように計算処理した 1 つひとつのオーバーラップは、統計学的には 10^{17} 分の 1 の確率の事象であり、従って偶発的なものではない。アセンブリの組み合わせを困難にしているものは、実際にゲノムがオーバーラップしている領域からサンプリングされたオーバーラップ配列は多いが（すなわち、その配列はアセンブリすべきものである)、実際にはもっと多くのオーバーラップが、前述のスクリーニングで除外されなかった低反復配列数エレメントを含む 2 つの異なる配列由来のものである（組合わせると誤りになる）という点である。我々は前者を「真のオーバーラップ」、後者を「反復配列によって引き起こされる間違ったオーバーラップ（repeat-induced overlaps)」と呼ぶ。アセンブラは、特に、工程初期に、この間違ったオーバーラップを選別しないよう注意しなければならない。

我々は、この目的をユニティグ検出機で達成した。先ず、他の読み取り配列アセンブリに比べて、これは一種類しかなく文句なく採択できると思われる読み取り配列アセンブリを見つける。これらのサブアセンブリから構成されるコンティグ群を unitig と呼ぶ（ユニークにアセンブリしたコンティグの意（uniquely assembled contigs））。正式には、これらの unitig は、全てのオーバーラップを示すグラフの中で、より確実に正しいと思われる中間段階のサブグラフである [42]。経験的にはこれらのアセンブリの多くは正しいが（すなわち、真のオーバーラップしか含まない）、残念なことにいくつかは、実際は複数の反復塩基配列由来の配列の寄せ集めで、単一サブアセンブリ（体）になるまで分断しすぎたものとなってしまう。しかし、過剰分断した unitig は、平均カバー倍数の程度（coverage depth）が高くなりすぎて、全体の配列カバー倍数レベルとは一致しないため、容易に見分けることができる。我々は、unitig が、特有の DNA 塩基配列だけか 2 つ以上のコピーを含む反復配列で構成されているかを判定し、オッズ比の対数を算出する単純な統計学的識別プログラムを開発した。この識別プログラムは、閾値を十分に厳しく設定すれば、正しいと確信できる unitig のサブセットを同定できる。また、あまり厳しくない閾値では、正しくアセンブリされた可能性の高い、残りの unitig のサブセットを同定できる。これらのうちで、一貫してアセンブリの足場となり得るものを選別していくと（後述参照）、選別された unitig はほぼ正しいと考えられる。これらの 2 セットをあわせて、U-unitigs と呼ぶ。ヒト第 22 番染色体を 6 倍カバー倍数シュミレーションショットガン法でアセンブリした結果から、経験的に、我々が得た U-unitigs は長さ 2 kbp 以上の特有の DNA 塩基配列の 98%をカバーすることがわかった。我々はさらに、一つの U-unitig の末端で反復配列が始まる境界を同定できた。これを梃子にして、U-unitigs が、バラバラに散在する Alu 配列と他の 100～400 bp の反復配列セグメントの 93%以上を占めるようにした。

ユニティグ検出機（Unitigger）を作動させた結果、推測でヒトゲノムの 73.6%をカバーし、正しくアセンブリされたサブコンティグのセットが得られた。ついでアセンブリ骨格形成機（Scaffolder）が、メイト対情報を用いて、これらをアセンブリ骨格にリンクさせる作業を行う。2 つ以上のメイト対がある場合（所定の U-unitigs のペアがお互いに特定の距離と方向にあることを意味する）、メイト対の間違いが 2%未満であると仮定すると、距離と方向性が誤りとなる確率は、さらに約 10^{10} 分の 1 となる。したがって、少なくとも 2 個の 2 kbp または 10kbp のメイト対で連結できる U-unitigs は、極めて確実に全部リンクさせることができ、中程度の大きさのアセンブリ骨格が作成できる。このアセンブリ骨格を、さらに 50 kpb のメイト対と BAC 末端配列を確認しながら、繰返してお互いに連結させる。このプロセスにより、サイズ的に 100 万 bp の桁のアセンブリ骨格が作られた。これにはコンティグ間のギャップを含んでいるが、それは通常は反復配列に相当する部分で、まれに配列決定できない部分に相当する小さなギャップである。これらのアセンブリ骨格でゲノムの中の特有

の塩基配列の大部分を再構成できる。

キイロショウジョウバエの場合のアセンブリでは、3 段階の反復配列解析戦略（repeat resolution strategy）をとった。各段階は次第に難しくなり、そのためミスを犯しやすくなった。ヒトの場合のアセンブリでも、第 1 段階で「大岩」サブステージを継続使用することにした。この段階では、良好であるが決定的だとはいえない識別プログラムスコアを持つ全ての unitig を、アセンブリ骨格にあるギャップに配置する。これは、2 対以上のメイト対があって､そのメイト対の読み取り配列のうち 1 つは既にアセンブリ骨格上にあるために、unitig を所定のギャップに曖昧さが全くなく配置させ得るという条件で行った。この方法で unitig を正しくないギャップに挿入してしまうのは、確率解析では 10^{-7} 未満であると予測している。

我々は、ヒトアセンブリの次の段階「石」サブステージを改訂し、我々のこれまでの研究から示唆された機構に近づけるようにした [43]。各ギャップについて、アセンブリ骨格のコンティグの中にあり読み取り配列（R）の位置を示すはずのメイト対（M）を使って、ギャップの中に配置され得る全ての読み取り配列(R)を集める。セレラ社のメイト対情報は 99%以上正しいので、こうすればこのセット中の、全てではないがほとんどの読み取り配列は、ギャップに収まるはずである。ある読み取り配列がそのギャップに収まらない時は、セットの中の残りの読み取り配列がそのギャップにうまく収まることはまれである。従って、我々は、単純に、この読み取り配列セットをギャップ内でアセンブルし、このアセンブリのやり方に矛盾するような読み取り配列は排除した。この作業は、キイロショウジョウバエのアセンブリのために行った作業よりも信頼性が高いことが判明した。ヒト第 22 番染色体について、シミュレートしたショットガン法データセットをアセンブリしたところ、全ての「石」が正しく位置された。

ギャップ解消の最終段階は、アセンブリした BAC データ（ギャップをカバーしているはず）でギャップを埋めることである。これを、外部ギャップ「歩行（walking)」と呼ぶ。今回は、キイロショウジョウバエの配列決定で報告した非常に攻撃的な「丸小石」サブステージは適用しなかった。「丸小石」サブステージは、99.62%しか正しくない長い散在配列の反復再構築物(repeat reconstructions)を生み出すミスを犯したからである。ヒトゲノムのためには、99.99%以上の精度にならないことが判っている方法は用いない方が論理的に良い、と我々は決めた。その代価として、サイズが若干大きく、数も若干多いギャップができてしまったが。

アセンブリプロセスの最終段階およびそれに至るいくつかの中間過程では、各コンティグのコンセンサス配列を作製する。我々のアルゴリズムは、各塩基の評価に品質価値加重評

価法（quality-value-weighted measures）を用いつつ、最大節減（maximum parsimony）原則で動いている。この効果は、各過程で、報告されるべき正しい塩基のベイズ推定（Bayesian estimate）に現れる。コンセンサス配列の作製には、セレラ社のデータがあればそれを利用する。セレラ社のデータが所定の領域をカバーしていないときは、BAC データ配列を用いる。

ヒトゲノムの WGA を達成できた鍵となったのは、オーバーラップ検出機と中央コンセンサス配列構築サブルーチン（central consensus sequence-constructing subroutines）を平行して動かすことであった。また、メモリーは重要な問題であった。キイロショウジョウバエのために構築したソフトウエアをそのまま適用すると、600 ギガバイトの RAM を持つコンピュータが必要であった。オーバーラップ検出機にユニティグ検出機を加増することで、最大限の即時データ連絡により、同じ計算を 28 ギガバイトの RAM で済ますことができた。さらに、最初の 3 段階が加増的に処理できるため、データが配達されるたびに、計算のこの部分の状態を連続更新できた。また必要なときはいつでも、アセンブリ骨格構築と反復配列解析を完了するために 7 日間連続稼働させることができた。我々のアセンブリ作業のために使われた計算機インフラストラクチャーは、全体で、クラスターあたりで 4 ギガバイトのメモリーを備えた 10 台の 4 プロセッサー SMP（Compaq 社、ES40、Regatta）と、64 ギガバイトのメモリーを備えた 1 台の 16-プロセッサーNUMA マシン（Compaq 社、GS160、Wildfire）から構成される。アセンブラを 1 回作動させるための総計算時間は、おおよそ 20,000CPU 時間であった。

セレラ社データと、断片化 bactig のデータをアセンブリすることにより、スパンが計 2.848 Gbp で、2.586 Gbp の配列から構成される一組のアセンブリ骨格が作られた。チャフ、すなわちアセンブリに組み込まれなかった読み取り配列のセットは、1127 万個（26%）であった。これはキイロショウジョウバエでの経験と矛盾しない。ゲノムの 84%以上が 100 kbp 長以上のアセンブリ骨格でカバーされ、総計 2.297 Gbp の配列のうち、平均 91%を塩基配列が占め、9%がギャップであった。100 kbp 以上のアセンブリ骨格 1637 個中に、計 93,857 個のギャップがあった。アセンブリ骨格サイズは平均 1.5 Mbp であり、平均コンティグサイズは 24.06 kbp、平均ギャップサイズは 2.43 kbp で、それぞれのサイズ分布は基本的に指数関数的であった。全ギャップの 50%以上が 500 bp よりも短く、全ギャップの 62%以上が 1 kbp よりも短く、100 kbp を超える長さのギャップはなかった。同様に、配列の 65%以上が 30 kbp 以上のコンティグの中にあり、31%以上が 100 kbp 以上のコンティグにあり、最大のコンティグは 1.22 Mbp であった。表 3 に、このアセンブリ構造の統計を詳しくまとめ、区画化ショットガンアセンブリと直接比較した。

表 3 全ゲノムショットガンアセンブリと区画化ショットガンアセンブリのアセンブリ骨格の統計値

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
Compartmentalized shotgun assembly					
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≤1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
Whole-genome assembly					
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≤1 kbp	62,356	60,343	59,156	54,079	49,592
Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

2.4. 区画化ショットガンアセンブリ

WGA アプローチに加えて、部分的にアセンブリするアプローチを行った。これは、ゲノムを区画に分け、それぞれの区画を、個々にショットガン法でアセンブリする方法である。我々は、これが、染色体間でも多数ある重複の解決と、U-unitigs を計算するための統計処理の改良に有用であると考えた。区画化アセンブリプロセスでは、セレラ社の読み取り配列と bactig をゲノムの多数の大メガベース領域にクラスター化する過程と、その後にセレラ社データと bactig データから得られた人造読み取り配列断片に対して WGA アセンブラーを走らせる過程を含んでいる。

CSA ストラテジーの最初の段階は、セレラ社の読み取り配列を、特定の PFP BAC 登録に対して BAC コンティグに一致するものと、公表データと一致しないものに分けることであった。セレラ社の読み取り配列を適切に配置するために、このような相互一致があることは保証されなければならない。そのために、最初に全読み取り配列に対して、一般的反復配列エレメント（既にライブラリーがある）部分を目隠ししておいて、読み取り配列の目隠しされていない部分で 40 bp 以上が一致したときのみ、ヒットありとした。こうしてセレラ社の 2727 万個の読み取り配列中、2076 万個が bactig と一致した。その他の 62 万個の読み取り配列には一致するものがなかったが、にもかかわらずそれらのメイト対が bactig と一致したため、bactig の BAC 自体の領域に所属するものと判断された。残りの読み取り配列のうち、292 万個の読み取り配列は完全にふるい落とされ、一致させることができなかった。しかし、他の 297 万個の読み取り配列は、GenBank データセットには認められない計 1.189 Gbp の目隠しされない配列を持っていた。セレラ社のデータは全体でゲノムの 5.11 倍に及ぶ重複があるため、我々は、240 Mbp のセレラ社特有の配列が、GenBank データセットには含まれていないと推測している。

CSA プロセスの次の段階では、結合アセンブラが、ゲノムに関わる 5 倍分のセレラ社読み取り配列と、BAC に入っている bactig を取り込み、組み合わせデータによりその場所特定のためのアセンブリを作成した。この高品質な再構成配列は一時的なものであり、次の段階でより信頼性の高い情報を提供するため、オーバーラップしたり隣接したりしている骨格配列のセットの中にそれらを埋めていくのに役立てようという単純なものである。概略を述べると、結合アセンブラは、最初に、一致するセレラ社読み取り配列のセットがあるかどうか調べ、ふるい落としで見逃された反復配列であることを示す過剰データ累積部位があるかどうか判断する。ある場合は、反復配列由来の読み取り配列（そのメイト対が本来あるべき位置にまだマッピングされていないもの）を除去する。次に、2 つの bactig の間で相対的にみて同じ位置を一貫して示している全てのメイト対のセットを、1 つのリンクとして束ね、その束の中のメイト対の数に応じて重みを与える。次に「がつがつ（greedy）」戦略で、この重みの順にメイト対の束を選び、bactig に順位をつける。適宜選抜されたメイト対の束は、造形用の 2 つのアセンブリ骨格を一緒に結びつけることができる。その骨格内のコンティグ間で大部分のリンクと一致している場合のみ、単一の骨格に組み上げるのである。一旦アセンブリ骨格が完全になれば、WGA アセンブラのところで前述した「石」戦略によりギャップを埋める。

フェーズ 1 と 2 の BAC に関する GenBank のデータは、BAC1 個あたり平均 19.8 個、平均 8099 bp の bactig から構成されていた。結合アセンブラの適用により、個々のセレラ式 BAC アセンブリで平均 1.83 個の骨格にまとめられた（メディアン値：1 個）。これは、平均サイズが 18,973 bp の、平均 8.57 個のコンティグから構成されている。配列断片の

順序と方向性の決定に加えて、この組み合わせ方式では、ギャップが 57%少ない結果となった。フェーズ 0 のデータに関しては、平均の GenBank 入力データは、平均 784 bp の 91.52 個の読み取り配列から構成されていた。結合アセンブラの適用により、平均サイズが 873 bp の、平均 58.1 個のコンティグから構成される平均 54.8 個のアセンブリ骨格が作られた。基本的には、一部少量のアセンブリが行われたが、典型的なフェーズ 0 の BAC に代表される 0.5〜1 倍のデータセットに合致する十分な数のセレラ社データが正確にアセンブリされたでわけはなかった。結合アセンブラは、フェーズ 3 のデータにも適用され、SNP の同定、アセンブリの確認、セレラ社読み取り配列の位置決定が行われた。フェーズ 0 のデータは、BAC 組み合わせ方式の全ゲノムショットガンデータセットと BAC の 1 倍分の軽ショットガン塩基配列決定からでは、BAC に拾われた領域の良いアセンブリ結果を生み出さないことを示唆している。少なくとも各 BAC について 3 倍の軽ショットガン塩基配列決定が必要である。

GenBank データと適合しなかった 589 万個のセレラ社断片は、われわれの全ゲノムアセンブラでアセンブリした。アセンブリによって、計 442 Mbp のスパンで、326 Mbp の配列から構成される、1 セットのアセンブリ骨格が得られた。この骨格全体の 20%以上が 5 kbp 以上の骨格配列で占められ、計 302 Mbp の配列の中で、平均 63%が配列そのもので 27%がギャップであった。全ての 5 kbp 以上の骨格は、結合アセンブラで作成された全てのアセンブリ骨格とともに、次の、タイル積み段階（tiling phase）に持ち込まれた。

この段階での典型例では、ゲノムに関連した配列の少なくとも 95%を構成する BAC 領域すべてについて、1〜2 個の骨格と、つながり不明のセレラ特有の骨格があった。ゲノム構成を明らかにするための次の段階は、ゲノム全体を通じてこれらの BAC 骨格とセレラ特有骨格の順序づけとオーバーラップしている配列を、タイルを積み上げるようにして並びを決定するタイリングであった。このために、我々はセレラ社の 50 kbp メイト対情報、BAC 末端対情報 [18]、配列標識部位（STS）マーカー [44] を用いて、広範囲にわたるガイド設定と染色体長分の分離を行った。比較的扱いやすい数の骨格であれは、このタイリングを完全に自動化した方法では行わず、最初のタイリングを良好な帰納法で計算し、その後人手（管理責任者）で、矛盾点や未結合点を解決する。このために、我々は、タイリングするオーバーラップのグラフとそれぞれの証拠を表示するグラフィカル・ユーザーインターフェースを開発した。管理責任者は、マッピングされた STS データ、配列オーバーラップのドットプロット、その選択を裏付けるメイト対の証拠のビジュアル表示が暗示するものを探査できた。このプロセスの結果、「コンポーネント」群が得られた。各コンポーネントは、管理責任者が認証した BAC 骨格とセレラ特有骨格が積み上げられ並べられたセットである。このプロセスでは、推定スパンが 2.922 Gbp の、3845 個のコンポーネントが得られた。

最終的な CSA を作成するために、各コンポーネントを、WGA アルゴリズムを用いてアセンブリした。WGA プロセスで行ったのと同様に、アセンブラに、これとは独立にデータをアセンブリする自由度を持たせるために、bactig データを 2 倍の人為的なショットガンデータに断片化した。bactig ではなくこの人造読み取り配列を用いることで、アセンブリアルゴリズムは、もとの bactig のアセンブリの誤りを正し、PFP 登録データの中のキメラ配列を除去できた。キメラ、すなわち、（ゲノムの他の部分から）混入した配列は、もともとそこにはなかったものだということで、コンポーネントの再アセンブリには組み込まなかった。これが効果を現し、CSA プロセスにおけるこれ以前の操作段階は、結局、ゲノムの大きな隣接区画に関係するセレラ社断片と PFP データを一緒にするのに役立つだけとなった。そこでは我々は、WGA に用いたアセンブラを対象領域の当初段階からのアセンブリに用いた。

コンポーネントの WGA アセンブリにより、スパンが計 2.906 Gbp で、2.654 Gbp の配列から構成される一セットの骨格が作られた。チャフ、すなわちアセンブリに組み込まれなかった読み取り配列のセットは、617 万個（22%）であった。ゲノムの 90%以上が 100 kbp 以上の骨格でカバーされ、配列は計 2.492 Gbp で、平均 92.2%が配列そのものであり、7.8%がギャップであった。100 kbp 以上の骨格 1,940 個に属する 107,199 個のコンティグの中に、計 105,264 個のギャップがあった。骨格サイズは平均 1.4 Mbp であり、平均コンティグサイズは 23.24 kbp、平均ギャップサイズは 2.0 kbp で、それぞれのサイズは指数関数的に分布している。またそのため、平均値は、大部分のデータの過小評価値となりやすい。図 5 に、様々なサイズ範囲の骨格中の塩基数ヒストグラムを示す。全ギャップの 49%以上が 500 bp よりも短く、62%以上が 1 kbp よりも短く、全て 100 kbp より短かったことを考慮していただきたい。同様に、配列の 73%以上が 30 kbp 以上のコンティグの中にあり、49%以上が 100 kbp 以上のコンティグにあって、最大のコンティグは 1.99 Mbp であった。表 3 に、WGA アセンブリと直接比較のため、このアセンブリの統計値の要約を示す。

図 5 CSA のアセンブリ骨格サイズの分布

各サイズ範囲に属するアセンブリ骨格が全配列に占める割合（%）を示す。[拡大像（10K GIF ファイル）]

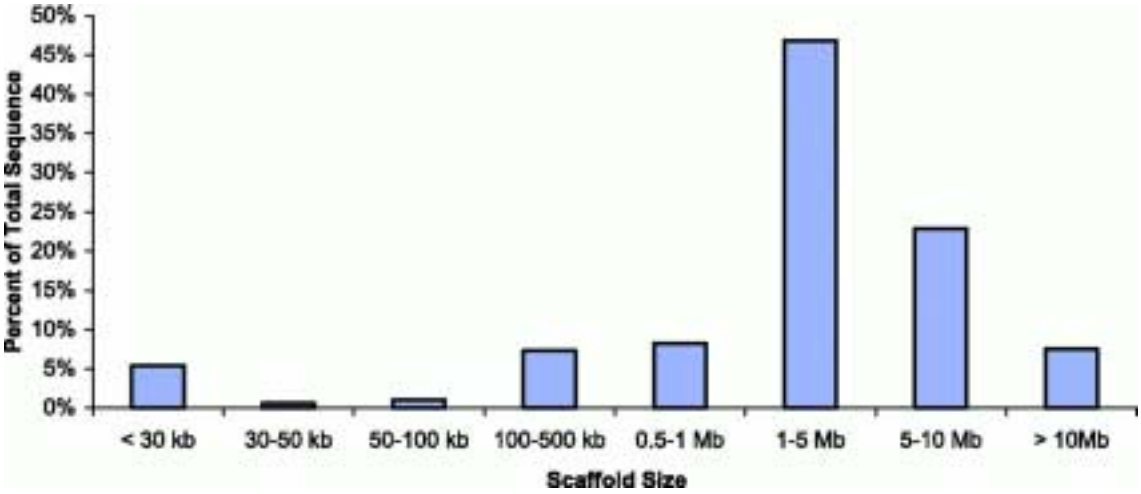


表 3 全ゲノムショットガンアセンブリと区画化ショットガンアセンブリのアセンブリ骨格の統計値

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
Compartmentalized shotgun assembly					
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≦1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
Whole-genome assembly					
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≦1 kbp	62,356	60,343	59,156	54,079	49,592

Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

2.5. WGA と CSA の骨格比較

別々の計算処理プロセス（WGA と CSA）を経て、ヒトゲノムのアセンブリが 2 種類得られたので、これらの完全性、整合性、連結性を調べる別手段として、2 つのアセンブリから得られた骨格を比較した。それぞれのアセンブリから、少なくとも 1000 個の断片を含む基準骨格を 1 セット（セレラ社の読み取り配列または bactig 断片）得た。これらは 2218 個の WGA 骨格と 1717 個の CSA 骨格で、各々計 2.087 Gbp と 2.474 Gbp であった。少なくとも 20 個のフラグメントか小さい方の骨格では少なくとも 20%を占めている一方の基準骨格の配列を、他方のアセンブリから得られた全ての骨格の配列と比較した。各比較に関して、ミスマッチが 2%以下で 200 bp 以上の一致があるものを全て表にした。

この表から、2 つの方法で、各アセンブリの特有の配列量を推定した。第一の方法は、他方のアセンブリの一致部分とは異なる各アセンブリの塩基の数を測定することである。WGA の約 82.5 Mbp（3.95%）が CSA にカバーされておらず、CSA の約 204.5 Mbp（8.26%）が WGA にカバーされていなかった。この推定では、アセンブリの整合性や一致部分のユニークさは必要とされない。したがって、別の解析を行った。その解析では、一対の比較骨格間で 1 kbp 未満の一致配列については、他の一致配列で矛盾のない順序と方向性が確認されない限り、除外した。これにより、一致配列数の測定ができた。より厳密なこの方法によると、WGA の 1.982 Gbp（95.00%）が CSA にカバーされており、CSA の約 2.169 Gbp（87.69%）が WGA にカバーされていた。

WGA を CSA と比較すると、骨格構造上の不一致度合いが評価できる。一方のアセンブリから得られた大型部分骨格で、他方のアセンブリ骨格群にあってはそのうちのたった一つとしか一致しない例で、それでも一致片によって示される配列のオーバーラップする長さが全長にわたらない例を探した。候補となる最初のセットは自動的に同定し、各候補セットを手動で調査した。このプロセスから、そのアセンブリがその場所に限定されるとは限らないと考えられる例を 31 個見いだした。これらの例はさらに調査し、どのアセンブリが間違っていて、なぜなのかを判断した。

さらに、順序と方向性について、局所的な不一致性を評価した。以下に述べる結果は、

一方のアセンブリの 1 つのコンティグが、他方のアセンブリの 2 つ以上のコンティグに一致する場合（後者の順序と方向性が、対応する前者の場所と一致する限り）を除外している。これらの小さな配置ミスは塩基対で数百個程度のものであり、1 kbp 以上のものはめったに関与していない。我々は、CSA アセンブリのうち、計 295 kbp（0.012%）が、局所的に WGA アセンブリと矛盾し、WGA アセンブリのうち、2.108 Mbp（0.11%）が、CSA アセンブリと矛盾することを見いだした。

CSA アセンブリは骨格のカバー倍数の点では数パーセント良好であり、WGA よりもわずかに矛盾が少なかった。というのは、CSA は実際、メガベースサイズの課題のショットガンアセンブリを数千回行っているが、WGA はギガベースサイズの課題のショットガンアセンブリを 1 回行っているためである。課題サイズが 2.5 桁も大きいことを考えれば、この 2 者間の情報ロスは非常に少ない。2 つの結果のうち CSA は論理的に出力しやすくベターで、その後の解析を開始しなければならない時点で入手できたことから、以降の解析は全て、このアセンブリに関して行った。

2.6. アセンブリ骨格のゲノムへのマッピング

ゲノムのアセンブリの最終段階は、染色体上にアセンブリ骨格を並べ、方向付けさせることである。最初に、CSA のコンポーネント中の順序に基づき、アセンブリ骨格をグループ化した。これらのグループ化したアセンブリ骨格を、アセンブリ骨格間に残存しているメイト対データを検討して並べ直した。次に、物理的マッピングデータを用いて、アセンブリ骨格グループを染色体上にマッピングした。この段階は、各アセンブリ骨格が複数のマーカーと重なるような、信頼性の高い高分解能マップ情報を得られるかどうかに左右される。入手可能なゲノム全体にわたるマップ情報は 2 つある。高密度 STS マップと、ワシントン大学で作製された BAC クローンのフィンガープリントマップ（WashU BAC マップ）である [45]。ゲノム全体の STS マップのうち、GeneMap99（GM99）は最多のマーカーを有し、アセンブリ骨格をマッピングするには最も有用であった。2 種類のマッピング法は、お互いに相補的なものである。フィンガープリントマップは、オーバーラップした BAC クローンの比較により作製されたため、部分的な情報が良好である。一方、GM99 は、枠組みとなるマーカーが、よく実証された遺伝子マップに由来しているため、広範囲にわたる情報はより信頼性が高い。両タイプのマップを、座位アセンブリのために入力したコンポーネントの、人手による管理基準として用いたが、これらは、アセンブラにより生み出された配列順を規定するものではない。

フィンガープリントマップと GM99 の、アセンブリ骨格のマッピング効率を調べるため、まず、大きなアセンブリ骨格で比較することでこれらのマップの信頼性を検討した。

GM99 上では、10 個の大きなアセンブリ骨格（9 Mbp 以上）上の STS マーカーのうち 1%のみが異なる染色体にマッピングされた。STS マーカーの 2%が 5 フレームワーク結合分以上位置が違っていた。しかし、フィンガープリントマップでは、2%が異なる染色体上にあり、アセンブリ骨格配列における BAC 位置の平均 23.8%が、5 BAC 以上、フィンガープリントマップの配置と一致しなかった。さらに不一致の根源を検討したところ、不一致のほとんどは 10 個のアセンブリ骨格のうち 4 個に由来することが明らかになった。これは、マップかアセンブリ骨格のいずれかの質にバラツキがあることを示している。4 個のアセンブリ骨格とも、残り 6 個のアセンブリ骨格と同じようにアセンブリしており、クローン配列カバー倍数解析で評価すると、GM99 との不一致率が同じように低いことが示された。このことから、これらの場合ではフィンガープリントマップの全体像は信頼性が低いと結論した。より小さいアセンブリ骨格では、GM99 との不一致率は高かった（STS の 4.21%が 5 フレームワーク結合分以上一致しなかった）が、フィンガープリントマップとの不一致率は低かった（BAC の 11%がフィンガープリントマップで 5 BAC 分以上一致しなかった）。この結果は、セレラ社のアセンブリ骨格構築では、小さいアセンブリ骨格よりも大きいアセンブリ骨格の方が長距離メイト対により、よりうまく合うという、クローン配列カバー倍数解析と一致している[46]。

我々は、これらのマップのマーカー（BAC または STS）に基づき、セレラ社のアセンブリ骨格の並び順を 2 種類作製した。アセンブリ骨格の順序が GM99 と WashU BAC マップで一致する場合は、その順序が正しいことに確信を持ち、これらのアセンブリ骨格には、「アンカーアセンブリ骨格」と名付けた。両方のマップで全般的な不一致性が低く信頼できるアセンブリ骨格のみ、アンカーアセンブリ骨格であると判断した。GM99 側にあるアセンブリ骨格は、大枠順（framework order）を乱さない場合、順序を入れ替えて WashU の順序と一致させることが可能であった。各アセンブリ骨格の方向付けは、マッピングされた複数のマーカー（順序に矛盾がないもの）の存在により決定した。1 つのマーカーしか持たないアセンブリ骨格は、方向を決めるには情報が不十分である。我々は、ゲノムの 70.1%がアンカーアセンブリ骨格中にあることを見いだし、そのアンカーアセンブリ骨格は 99%以上が方向付けされている（表 4）。GM99 は WashU マップよりも解像度が低いため、STS マーカーで一致するもののない多くのアセンブリ骨格を、アンカーアセンブリ骨格に関連づけて並べた。それらが、WashU BAC マップ上の同じか隣接した BAC 由来の配列を持っていたためである。また、WashU BAC マップ上に時折みられるおかしな並べ方のため、WashU マップ上に「マッピングできない」と判断された多くのアセンブリ骨格を、アンカーアセンブリ骨格に関連づけて GM99 で並べることができた。これらのアセンブリ骨格には、「整列化アセンブリ骨格」と名付けた。13.9%のアセンブリ骨格がこれらの補足的方法で並べることができた。このようにして、ゲノムの 84.0%を確実に整然と並べた。

表 4 アセンブリ骨格マッピングのまとめ
アセンブリ骨格は様々な信頼度でゲノムにマッピングされた（アンカーアセンブリ骨格が最も信頼性が高く、マッピングされないアセンブリ骨格が最も信頼性が低い）。アンカーアセンブリ骨格は、WashU MADC マップと GM99 により、矛盾なく並べられた。整列化アセンブリ骨格は、WashU BAC マップか、GM99 か、コンポーネントタイリング路程のうち 1 つ以上に基づいて並べられた。拘束アセンブリ骨格は、少なくとも 2 つの社外マップの間では順序が矛盾するが、これらの位置は近隣のアンカーアセンブリ骨格または整列化アセンブリ骨格に隣接している。マッピングされないアセンブリ骨格は、所属染色体の記載がなされている程度である。アセンブリ骨格のサブカテゴリーは、それぞれのカテゴリーの下に示してある。

Mapped scaffold category	Number	Length (bp)	% Total length
Anchored	1,526	1,860,676,676	70
Oriented	1,246	1,852,088,645	70
Unoriented	280	8,588,031	0.3
Ordered	2,001	369,235,857	14
Oriented	839	329,633,166	12
Unoriented	1,162	39,602,691	2
Bounded	38,241	368,753,463	14
Oriented	7,453	274,536,424	10
Unoriented	30,788	94,217,039	4
Unmapped	11,823	55,313,737	2
Known chromosome	281	2,505,844	0.1
Unknown chromosome	11,542	52,807,893	2

次に、アンカーアセンブリ骨格の間に配置はできたものの並び順を決めることができなかった全てのアセンブリ骨格を、アンカーアセンブリ骨格間にある間隔体とし、アンカーアセンブリ骨格間に「拘束されている」と考えた。例えば、小さなアセンブリ骨格で、GeneMap の同じ箱（bin）に対して STS 該当部位を持っているか、または同じ BAC に当たる部位があるものは、相対的に並べることができないが、他のアンカーアセンブリ骨格か整列化アセンブリ骨格に関連づけて、境界域に割り当てることができる。残りのアセンブリ骨格は、位置情報や、矛盾情報を持たず、一般的な染色体座位に割り当てることしかできなかった。以上のアプローチを用いて、ゲノムの約 98%が、位置固定されるか、並べられるか、間隔拘束された。

最後に、染色体ごとにアセンブリ骨格を拡散展開することによって、染色体上に配置されたアセンブリ骨格の局在位置を確定した。ゲノムの 2%にあたる、マッピングされなかった残りのアセンブリ骨格は、ゲノム全体を通じて均等に分配されると仮定した。マッピングされなかったアセンブリ骨格長の合計を、マッピングされたアセンブリ骨格数の合計で割ることにより、アセンブリ骨格間のギャップが 1,483 bp であると推定した。このギャップを用いて、各染色体上の全アセンブリ骨格を分別し、染色体内にきちんと

配置できないものをオフセットとして設けた。

アセンブリ骨格をマッピングする作業の間、我々は多くの問題に遭遇し、追加的な品質評価とバリデーション解析を行うはめになった。少なくとも 978 個（33,173 個の 3%）の BAC がゲノム内の 2 ヶ所以上からの配列データを持つことが判明した[47]。これはアセンブリ戦略の項で述べた bactig キメラ解析と一致する。これらの BAC は、CSA アセンブリの中でユニークな位置に割り当てることはできず、そのためアセンブリ骨格を並べるために利用することもできなかった。似たようなもので STS も、ゲノム重複、繰返し配列、偽遺伝子があるために、アセンブリで常にユニークな位置に割り当てることができたわけではない。

完璧なオーバーラップを徹底的に検索するのには時間がかかるため、CSA では、21,607 個のアセンブリ骨格内ギャップを作り出してしまった。メイト対データが当該コンティグはオーバーラップすべきだと示唆しているのに、実際にはオーバーラップが検出されなかったギャップである。これらのギャップは長さ 50 bp の固定長と定義されており、CSA アセンブリの計 116,442 個のギャップの 18.6%を占めている。

アセンブリ骨格を並べる手段として、cDNA や EST データで示されているエキソンの並び順は用いないことにした。このデータを利用しなかったのは、利用した場合、翻訳データにあわせるためにアセンブリ骨格を並べ直すことにより、アセンブリの特定の領域に偏りができ、アセンブリと遺伝子の定義プロセスのバリデーションがより困難になるためである。

2.7. アセンブリとバリデーション解析

我々は、完成度（ゲノムカバー度）と精度（順序と方向の構造的正確さとアセンブリのコンセンサス配列）の観点から、ゲノムのアセンブリを解析した。

完成度

完成度は、アセンブリ中の真正染色質の配列の割合として定義される。これは、真正染色質の配列決定が完了するまでは絶対的な確実性をもって知ることはできない。しかし、次のことに基づいて完成度を推測することができる：（ｉ）アセンブリ骨格内のギャップの推定サイズ、（ii）、既に配列が発表されている第 21、22 番染色体におけるカバー度[48,49]、（iii）アセンブリに含まれる別個のランダム配列のセット（STS マーカー）の割合の解析。全ゲノムライブラリーには、ヘテロクロマチン配列が含まる。それをアセンブリする試みは行われたことがないが、キイロショウジョウバエでみられたように、

ヘテロクロマチンの領域に埋め込まれた特有の配列の例がある可能性がある[50,51]。

ヒト第 21、22 番染色体の配列決定は、高品質で完了し、発表されている[48,49]。この配列もアセンブラに入力したが、配列決定の完了した配列を切断してショットガンデータセットにし、構造的な多型や BAC データのアセンブリミスがあった場合、アセンブラが元の配列とは違うアセンブリをするようにした。特に、アセンブラは、コンポーネント長のスケール（一般的に数メガベースサイズ）で、反復配列を解析できなければならない。しかも、この 2 者比較は、アセンブラが反復配列を解析できる解像度レベルを明らかにする。特定の領域では、アセンブリ構造は、発表されているヒト第 21、22 番染色体のものとは異なる（後述参照）。セレラ社データに基づき、配列決定の“終わったはず”の配列を、この違った形で柔軟にアセンブリした結果、第 21、22 番染色体の配列よりも多い DNA セグメントアセンブリとなった。なぜ第 21、22 番染色体配列よりもセレラ社配列ではギャップが多いのか理由を検討し、それらはゲノムの他領域におけるギャップに特有のものであるかもしれないと期待した。セレラ社アセンブリでは、25 個のアセンブリ骨格があり、それぞれ 10 kbp 以上の配列から構成され、全部集めると第 21 番染色体の 94.3%をカバーしている。また、62 個のアセンブリ骨格が第 22 番染色体の 95.7%をカバーしている。これら 2 つの染色体のセレラ社アセンブリで残っているギャップ長は合計 3.4 Mbp である。これらのギャップ配列は反復配列目隠し機（RepeatMasker）を使い、全ゲノムアセンブリに対して検索を行って解析した[52]。ギャップ配列の約 50%が、反復配列目隠し機で同定される一般的反復配列で構成されていた。残りの半分以上は、コピー数の少ない反復配列であった。

完成度を評価するもっと包括的な方法は、アセンブリに含まれる全く独自の配列データセットの含量を測定することである。Genemap99 由来の 48,938 個の STS マーカー[51]をアセンブリ骨格と比較した。これらのマーカーはアセンブリプロセスでは利用されていなかったため、これらは完成度について真に独自性のある基準となった。ePCR[53]と BLAST[54] を用いて、アセンブリされたゲノムに STS を配置した。我々は、マッピングされたゲノムに、44,524 個（91%）の STS を見いだした。その他に 2648 個のマーカー（5.4%）が、アセンブリされていないデータ、すなわち「チャフ」の検索で見いだされた。我々は、セレラ社配列にも、2000 年 9 月時点の BAC データにも見いだされない 1283 個の STS マーカー（2.6%）を同定した。このことから、これらのマーカーがヒト由来でない可能性が考えられる。もしそうならば、アセンブリされたセレラ社配列はヒトゲノムの 93.4%を表していることになり、アセンブリされていないデータは 5.5%で、カバー度合いは全部で計 98.9%となる。同様に、CSA を 36,678 個の TNG 放射線ハイブリッドマーカー（55a）に対して比較した。32,371 個のマーカー（88%）が、マッピングされた CSA アセンブリ骨格に位置し、2055 個のマーカー（5.6%）が残り

の位置に発見された。この場合は、後者のゲノム全体調査を通じて、ゲノムの 94%のカバー度となる。

精度

精度は、アセンブリの構造的・配列的正確さとして定義される。セレラ社データとGenBank データの配列原材料は異なる人間に由来するため、配列決定の終了した他の配列と比較して、直接、アセンブリのコンセンサス配列のヌクレオチドレベルで正確さを調べることはできない。しかし 6 章に示すように、この作業は多型同定のため行われている。コンセンサス配列の精度は、読みとり配列の品質値から導き出される統計値をもとにすると、99.96%以上である。

アセンブリが構造的に正しいということは、メイト対解析で測定できる。正しいアセンブリでは、読み取り配列の全てのメイト対が、正しい距離と向きをもって、コンセンサス配列に配置されるはずである。ある一組の読みとり配列ペアを見て、それらが正しく方向付けされており、読みとり配列間の距離が、その読みとり配列がサンプリングされたライブラリのインサートサイズ分布の平均値±3 標準偏差値以内であれば、その読みとり配列ペアは「正当（valid）」と呼ばれる。正しく方向付けされていないときは、その読みとり配列ペアは「方向ミス（misoriented）」と呼ばれ、読みとり配列間の距離が正しい範囲にないが正しく方向付けされている場合はその読みとり配列ペアは「分離ミス（misseparated）」と呼ばれる。アセンブラが用いる各ライブラリの平均±標準偏差はこのように決定される。これらを評価するため、第 21 番染色体の決定終了配列 [48] にマッピングされた全ての読みとり配列を検討し、試験室内追跡エラーとキメラ化（ゲノムの 2 つの異なる部位が同一プラスミドにクローニングされている）の結果として不正メイト対がどのぐらいの量あったか、そして、正確な場合はインサートサイズの分布がどの程度きっちりしていたかを決定した（表 5）。セレラ社の全ライブラリーの標準偏差は非常に小さく、少数の 50 kbp ライブラリーを除き、インサート長の 15%以下であった。2 kbp および 10 kbp のライブラリーは不正メイト対を 2%未満しか含まなかったが、50 kbp のライブラリーは幾分多かった（約 10%）。メイト対情報は完璧ではなかったにも関わらず、その精度は、特に数個のメイト対で方向性を確認するか否定するかする場合、対象アセンブリに関して正当・方向ミス・分離ミスの各メイト対量測定は、バリデーションにおいて信頼性の高い手段となると考えられる程であった。

表 5 メイト対のバリデーション
セレラ社のフラグメント配列を、報告されている第 21 番染色体の配列にマッピングした。独自ににマッピングされた各メイト対を、方向付けと位置が正しいかどうか評価した（テストしたメイト対の数）。2 つのメイトの相対方向や位置が間違っていた場合は、不正と判断した（不正メイト対の数）。BES：BAC 末端配列

Library type	Library no.	Chromosome 21						Genome		
		Mean insert size (bp)	SD (bp)	SD/mean (%)	No. of mate pairs tested	No. of invalid mate pairs	% invalid	Mean insert size (bp)	SD (bp)	SD/ mean (%)
2 kbp	1	2,081	106	5.1	3,642	38	1.0	2,082	90	4.3
	2	1,913	152	7.9	28,029	413	1.5	1,923	118	6.1
	3	2,166	175	8.1	4,405	57	1.3	2,162	158	7.3
10 kbp	4	11,385	851	7.5	4,319	80	1.9	11,370	696	6.1
	5	14,523	1,875	12.9	7,355	156	2.1	14,142	1,402	9.9
	6	9,635	1,035	10.7	5,573	109	2.0	9,606	934	9.7
	7	10,223	928	9.1	34,079	399	1.2	10,190	777	7.6
50 kbp	8	64,888	2,747	4.2	16	1	6.3	65,500	5,504	8.4
	9	53,410	5,834	10.9	914	170	18.6	53,311	5,546	10.4
	10	52,034	7,312	14.1	5,871	569	9.7	51,498	6,588	12.8
	11	52,282	7,454	14.3	2,629	213	8.1	52,282	7,454	14.3
	12	46,616	7,378	15.8	2,153	215	10.0	45,418	9,068	20.0
	13	55,788	10,099	18.1	2,244	249	11.1	53,062	10,893	20.5
	14	39,894	5,019	12.6	199	7	3.5	36,838	9,988	27.1
BES	15	48,931	9,813	20.1	144	10	6.9	47,845	4,774	10.0
	16	48,130	4,232	8.8	195	14	7.2	47,924	4,581	9.6
	17	106,027	27,778	26.2	330	16	4.8	152,000	26,600	17.5
	18	160,575	54,973	34.2	155	8	5.2	161,750	27,000	16.7
	19	164,155	19,453	11.9	642	44	6.9	176,500	19,500	11.05
Sum					102,894	2,768	2.7			
						(mean = 2.7)				

ゲノムの配列クローンカバー倍数は 39 倍であった。これは、どの塩基対も平均 39 個のクローンに含まれる、言い換えれば、メイト対でつなぎ合わされた 39 個の読みとり配列のどこかにいることを意味している。配列クローンカバー倍数が低い領域、または不正メイト対の比率が高い領域は、アセンブリに問題がある可能性がある。正当メイト対を用いて、アセンブリの各塩基のカバー倍数を計算した（表 6）。まとめると、30 kbp 以上のアセンブリ骨格では、セレラ社アセンブリのうち配列クローンカバー倍数が 3 倍未満の領域にあるのは 1%未満であった。このように、順序と方向を含めて、この測定をするだけでアセンブリの 99%以上が強い支持を受けるのである。

表 6 区画化ショットガンアセンブリ（CSA）と PFP アセンブリのゲノム全体のメイト対解析[*]

Genome library	CSA			PFP		
	% valid	% mis-oriented	% mis-separated[†]	% valid	% mis-oriented	% mis-separated[†]
2 kbp	98.5	0.6	1.0	95.7	2.0	2.3
10 kbp	96.7	1.0	2.3	81.9	9.6	8.6
50 kbp	93.9	4.5	1.5	64.2	22.3	13.5
BES	94.1	2.1	3.8	62.0	19.3	18.8
Mean	97.4	1.0	1.6	87.3	6.8	5.9

[*] Data for individual chromosomes can be found in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1.

[†]Mates are misseparated if their distance is >3 SD from the mean library size.

全ての方向ミスメイト対・分離ミスメイト対の場所と数を調べた。この解析を CSA アセンブリ（2000 年 10 月 1 日の時点）に関して行ったのに加えて、2000 年 9 月 5 日の時点 [30, 55b] での PFP アセンブリの調査も行った。後者の場合、セレラ社のメイト対を PFP アセンブリにマッピングしなければならなかった。高忠実度反復配列によるマッピングのエラーを避けるために、マッピングするメイト対は、両方の読みとり配列が 6%未満の違いで一箇所だけで一致（match）しているもののみとした。閾値として 5 個以上の不正メイト対セットが存在すれば、そこは潜在的分断点（potential breakpoint）を示すものと設定した。そこでは 2 つのアセンブリの構成が異なっていた。発表されている配列と CSA の第 21 番染色体アセンブリとを図示して比較すれば（図 6A）この方法論のバリデーションの一つとなる。図中の青いチェックマークは分断点を示す。双方の染色体配列に、同様な（少数の）分断点があった。例外はセレラ社のアセンブリの 12 セットのアセンブリ骨格で（212 個の単一コンティグアセンブリ骨格にあって全部で染色体長の 3%）、間違った位置にマッピングされていた。これらは小さすぎて信頼のおけるマッピングができなかったからである。図 6 と 7 と表 6 に、2 つのアセンブリのメイト対と分断点の差を示す。両方のアセンブリとも、大きいインサートのライブラリー（50 kbp と BAC 末端）の方が、小さいインサートのライブラリーより、方向ミス・分離ミスメイト対の割合が高かった。大きいインサートのライブラリーは、ゲノムのより大きな部分に及ぶという単純な理由から、不一致を検出しやすかった。第 8 番染色体の 2 つのアセンブリを図形的に比較（図 6、B と C）したところ、セレラ社アセンブリよりも PFP アセンブリに分断点が多いことが示された。図 7 に、各染色体の両方のアセンブリの分断点地図（青いチェックマーク）を横に並べて示す。セレラ社アセンブリの順序と方向付けは、2 つの配列決定終了染色体を除けば、分断点が実質的に少ない。図 7 には、両方のアセンブリの大きいギャップ（10 kbp 以上）も赤いチェックマークで示してある。

CSA アセンブリでは、全てのギャップのサイズはメイト対データに基づいて推定されている。2 つのアセンブリは、異なるヒトゲノムを用いているため、分断点は、構造多型により生じている可能性がある。これらは、双方のゲノムアセンブリの未完成状態をも反映している。

図 6 CSA アセンブリと PFP アセンブリの比較

（A）第 21 番染色体の全てと（B）第 8 番染色体の全て、（C）セレラ社の一つの骨格である第 8 番染色体の 1-Mb 領域。図を作るために、セレラ社のフラグメント配列を各アセンブリ上にマップした。各パネルの上段に PEP アセンブリ、下段にセレラ社アセンブリを示した。中段には、両アセンブリで順序と方向性が同じであり、かつ一貫して整列された最長の配列であるセレラ社配列を緑色の線で示した。方向性は同一であるが順番が異なる配列ブロックは、黄色の線で示した。方向性が同一でない配列ブロックは、赤線で示した。明確さを期し、黄線と赤線の場合、長さ 50kbp 以上が一致した配列のセグメント間にのみ線を引いた。各パネルの上段と下段には、ライブラリーサイズ別に分類した各アセンブリに対するセレラ社の不正メイト対の程度（赤：方向が不正。黄：メイト間の距離が不正）を示した。（正しい距離にあるメイト対は、平均ライブラリーのインサート片サイズから期待されるように、わかりやすくするため図から削除した）。分断予測点は、同タイプの不正メイト対が山をなしている箇所に相当するが、各アセンブリ軸上に青いチェックマークで示した。1 万個以上の N は、シアンブルーのバーで表した。Science online (www.sciencemag.org/chi/content/full/291/5507/1304/DC1)の図 3 で、24 個の染色体すべてをプロットしたものが見られる。[この図の拡大像（70K GIF ファイル）]

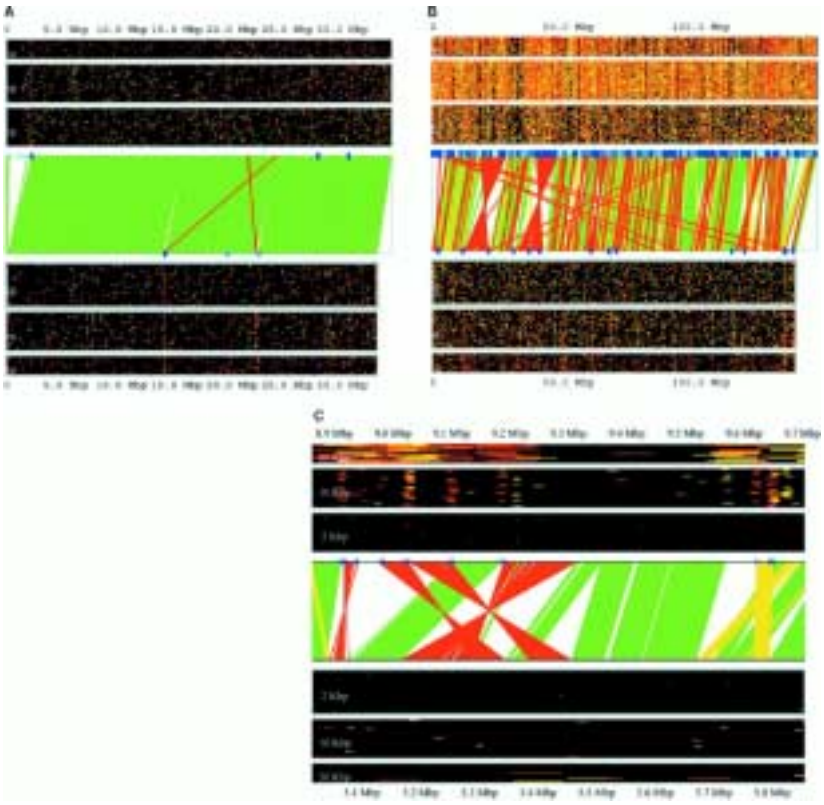
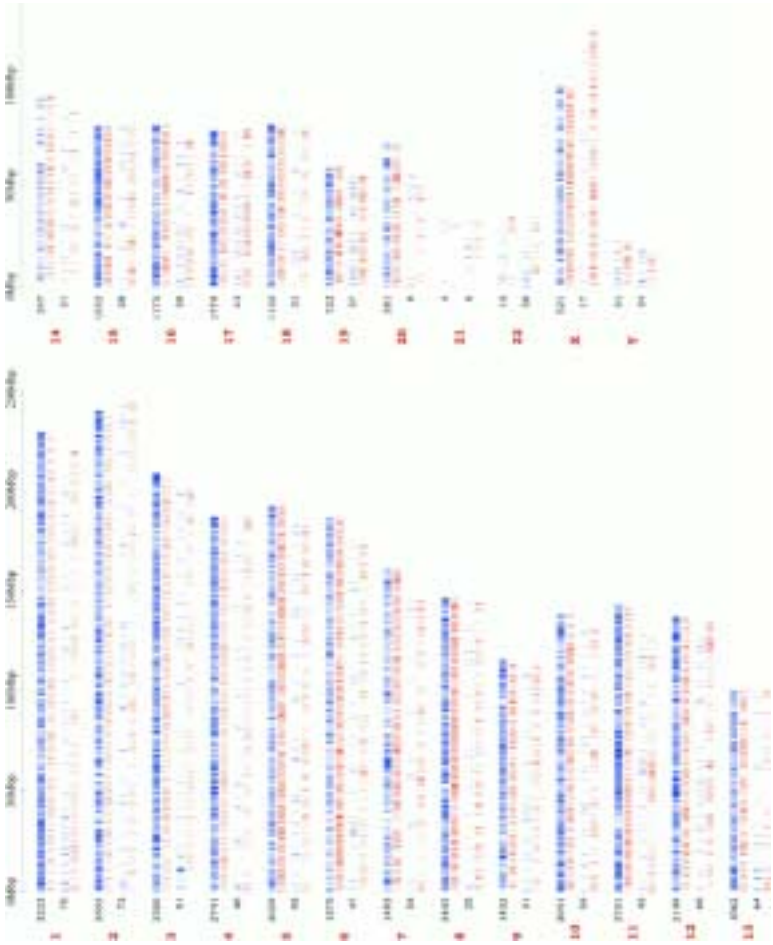


図 7 全染色体の分断点の分布と大ギャップを体系的にみた図
各染色体で上段の対は PFP アセンブリ、下段の対はセレラ社のアセンブリを表している。青いチェックマークは分断点、赤いチェックマークは 1 万 bp 以上のギャップを示す。染色体あたりの分断点の数は黒で、染色体数は赤で示した。[この図の拡大像（55K GIF ファイル）]



第 3 章 遺伝子の予測と注釈

要旨

遺伝子数の計数のため､Otto と名付けた統合的でしかも証拠に基盤をおく方法を開発した。遺伝子同定確率増を図るために証拠として用いたのは、マウスゲノムとヒトゲノムで配列保存されている領域、EST や他の mRNA 由来データとの類似性、他の蛋白への類似性等である。Otto（Otto-RefSeq および Otto homology からなる）を標準的な遺伝子予測アルゴリズムである Genscan と比較した場合、Otto は遺伝子構造決定において高い感度（Otto0.78 対 Genscan 0.50）ならびに高い特異性（Otto 0.93 対 Genscan 0.63）を示した。Otto の予測遺伝子群は、遺伝子発現があるかもしれないという弱いとはいえ依然として有意な証拠を示す 3 つの遺伝子予測プログラムから得られた遺伝子群セットを組み合わせたものである。少なくとも 2 種類の証拠を要するという保守主義的な基準を適用して、後続章の詳細な解析に用いる 26,383 個の遺伝子を、信頼できる遺伝子セットとして確定した。遺伝子構造の詳細な解析法を確立するには、高度な手作業整理によって、この初歩的コンピュータ計算法による結果を改善することが必要となるであろう。

3.1. 自動遺伝子注釈

遺伝子とは、同時に転写されるエクソン群がある部位である。単一遺伝子からでも、異なるスプライシング、異なる転写開始部位・転写終了部位によって多数の転写産物（したがって多数の異なる蛋白と異なる機能）を生じることがある。ヒト細胞は 10 億塩基対もあるゲノム DNA の中から転写開始信号を認識し、数塩基対から数十万塩基対離れたエクソンをスプライシングする信号を識別することができる。ゲノムを解析する第 1 段階は、それぞれの遺伝子単位及び転写単位を決定することである。

哺乳類でタンパクをコードする遺伝子数は当初から議論の的であった。再結合化したデータに基づく当初の予測では、遺伝子数は 30,000 から 40,000 個であった。一方、脳から出された後日の予測では、遺伝子数は 100,000 以上であった [56]。EST、CpG アイランド、転写密度推定に基づく最近のデータでも、企業と公的機関の間の不一致は解決されなかった。Incyte Pharmaceuticals 社が報告した 142,634 個の遺伝子数は近年の最大数で、しかもそれは EST データと CpG アイランド－EST 間相互関係に基づいている [57]。一方、以下に挙げる 3 つの異なった予測値は、対照的に遙かに低い。（ⅰ）約 35,000 個：全ゲノム EST データおよび第 22 番染色体データに連動するサンプリング手段に由来する遺伝子数 [58]、（ⅱ）28,000 から 34,000 個：ヒトおよびフグ（*Tetraodon nigroviridis*）間の保存配列を用いた系統的分類法に由来する遺伝子数 [59]、（ⅲ）35,000 個：第 21 染色体及び第 22 染色体の 67Mbp 中の既知および予測遺伝子 770 個の密度から、単純に約 3Gbp の真正染色質ゲノムに外挿した遺伝

子数。

コンピュータを用いてゲノム DNA 配列における転写単位を同定する際の問題は 2 つに分類できる。第 1 の問題は、個々の遺伝子に相当すると思われる断片に配列を分割する点である。この問題は重大であり、他のほとんどの新規遺伝子発見アルゴリズムが有する弱点である。またこの問題は、ヒト遺伝子目録中の遺伝子数の決定にも重大である。第 2 の挑戦的問題は、ある領域にコード化配置されていて転写産物を作り出すことができると思われる構造を反映した遺伝子モデルを構築することである。この作業は、全長 cDNA の配列決定が完了しているか、高度に相同的なタンパク質配列が知られている場合は、根拠ある精度でできる。新規開発法による遺伝子予測は、これより精度は低くなるものの、相同タンパク質もしくは EST からではわからない遺伝子を発見する唯一の方法である。以下に、タンパク質をコードする遺伝子の予測にあたって、これらの問題を解決するために開発した方法について述べる。

ヒトゲノムにおいて遺伝子を同定し構造決定する目的で、一定のルールに基づく Otto と名付けたシステムを開発した [60]。Otto は、遺伝子を同定し構造を明確にする手作業の過程をソフトウェアの中でシミュレートしようというものである。ゲノムの特定領域に注釈を加える過程で、人間の整理者は、コンピュータパイプライン（後述）が提供する証拠を検証し、いろいろな型の証拠がお互いにどのように関係するのか確かめる。整理者は、いろいろな型の証拠に様々なレベルの信頼度を設け、遺伝子の存在を支持する証拠に特定のパターンがないかを探す。例えば、整理者は多数の EST との相同性を検証し、これらがより長い仮想 mRNA に連結できるか否かを試験するかもしれない。また、スプライシング・ジャンクションにまたがる EST の存在と、推定されるエクソン末端の共通スプライシング部位の存在を問いながら、整理者は相同性の強さ、配列一致の連続性も評価するだろう。この種の手作業による注釈はショウジョウバエゲノムに注釈を加える際に用いられた。

Otto システムは、2 つある方法のいずれかにより観察した証拠から遺伝子存在注釈を行うことができる。第 1 の方法は、既知の遺伝子配列（ここでは、RefSeq データベースの手作業整理が終わったサブセットに示されるヒト遺伝子 [61] に限定）に対して高度の配列一致がみられる場合、Otto はこの事実から遺伝子注釈を行う。第 2 の方法は、Otto は広範囲の証拠を評価し、その証拠が遺伝子注釈を支持するのに適切かを決定する。これらの過程は後述する。

初めに、コンピュータパイプラインが作製した、タンパク配列と EST 配列がオーバーラップする一致セットの検証に基づいて、遺伝子境界を予測する [62]。このパイプラインは、タンパク質、EST、ゲノム配列データベースに対して、骨格配列を検索し、類似配列をもつ領

域を限定し、3 つの新規遺伝子予測プログラムを実行する。

あると思われる遺伝子領域境界を定めるため、BLAST により同定した遺伝子配列一致をもとにして、Otto を用いてゲノム領域を分割した。解析対象の領域内で一致するそれぞれのデータベース配列について、配列型式（タンパク、EST 等）と同様に、配列一致について両者のやり方を取り込むアルゴリズムを用いて比較した。その結果は、遺伝子を確定し遺伝子境界を同定すると思われる関連配列からなる貯蔵箱に分類した。この過程で、同一領域へ多数のヒットがあれば、ある領域のカバー倍数を追跡することで、一貫したデータセットに圧縮整理できる。例えば、もし一連の塩基配列が多数の EST でオーバーラップして示される場合、骨格配列上の EST セットによるその統合領域を、EST 証拠により支持される遺伝子領域としてマークする。これにより、単一遺伝子を含むと信じられる一連の “ 遺伝子貯蔵箱 ” が生みだされた。このアルゴリズムを最初に実施する際の弱点は、重複遺伝子が連続反復する領域における遺伝子境界予測にあった。遺伝子が集まるクラスターでは、度々、互いに結合した相同隣接遺伝子群となるため、これらの遺伝子を人為的にひとつにしてしまう遺伝子注釈が生じる。

次に既知遺伝子（全長 cDNA 配列がゲノムに完全一致するもの）を同定し、cDNA に相当する領域を予測転写領域として注釈をつけた。コンピュータパイプラインで探索したデータセットには、国立生物工学情報センター（NCBI, National Center for Biotechnology Information）の整理済みヒト遺伝子セット RefSeq に由来するサブセットも含めた。もし、ある RefSeq 転写産物が、少なくともその全長の 50%がゲノム・アセンブリに対して 92%以上の配列同一性レベルで一致するならば、解析対象のゲノム領域に対する RefSeq 転写産物の SIM4 [63] アライメントを用いて Otto 遺伝子注釈を進めた。ゲノム配列にはギャップおよびフレームシフト等の配列エラーが存在するため、実験的に決定された cDNA 配列に正確に合致する転写産物を予測することは必ずしも可能ではないのである。このようにして、遺伝子目録として合計 6538 遺伝子を同定、転写産物を予測した。

十分な配列類似性にも関わらず既知の遺伝子に一致しない領域は、転写産物を予測するために配列類似情報を用いる Otto の部分システムを用いて解析した。ここでは、ヒトゲノムにおける潜在遺伝子の予測のため、マウスゲノム DNA とヒトゲノム DNA 間の配列保存状況、ヒトの転写産物（EST 及び cDNA）への類似性、齧歯類の転写産物（EST 及び cDNA）への類似性、ヒトゲノム DNA を翻訳した場合の既知タンパクとの類似性に対応して、コンピュータパイプラインが作製した証拠を Otto が評価する。遺伝子貯蔵箱に入れられたゲノム DNA 領域から、遺伝子配列を抽出し、それに何らかの相同性があるという証拠があるサブ配列（この領域に隣接する 100 塩基分をプラスして）をマークした。

その領域で、何らかの相同性によってカバーされない塩基配列は、N に置換した。この配列断片、すなわち相同ゲノム配列で示された信頼度の高い領域と N で示されたそれ以外の領域を Genscan で評価し、一貫した遺伝子モデルが作製されるか否かを検証した。この手法は、第 1 に遺伝子境界を確立し（ほとんどの遺伝子発見アルゴリズムが得意としない）、支持証拠のない領域を除去することで、遺伝子予測を単純化した。Genscan がもっともらしい遺伝子モデルを作製した場合、Otto 遺伝子注釈に進める前に、遺伝子モデルをさらに評価した。Genscan による最終予測が、同一領域のもとのゲノム配列に対して作製した予測からは非常に異なる場合がしばしば生じた。遺伝子モデルを精密化するために Genscan を利用する場合の弱点は、有効な短いエクソンを最終遺伝子注釈で見失ってしまうことである。

配列類似性に基づく遺伝子構造決定の次のステップは、各エクソンに対する証拠の確かさの評価のために、前段階で用いた相同性に基づいた証拠と、それぞれの予測転写産物を比較することであった。内部エクソンは、末端から±10 塩基以内で相同性証拠によりカバーされる場合には、あるものとした。第 1 エクソンおよび最終エクソンは、10 塩基以内の内側端を必要としたが、外側端は 5’および 3’非翻訳領域（UTRs）を許すため許容範囲を広げた。予測精度維持のため、多数のエクソンからなる遺伝子を予測した際は、上に定義した全“ヒット”数を予測エクソン数で割った値が 0.66 以上であるか、RefSeq 配列に一致するという証拠があることを課した。単一エクソン遺伝子は、少なくとも 3 つの支持ヒット（両端に±10 塩基）によってカバーされ、しかもこれらは、予測されたオープン・リーディング・フレームを完全にカバーしなければならないものとした。さらに単一エクソン遺伝子には、Genscan にて開始コドン及び停止コドンが予測されることも必要とした。これらの基準に合わない遺伝子モデルは無視し、基準に合格したものを Otto 予測に進めた。相同性に基づく Otto 予測は 3’および 5’非翻訳配列を含まない。他の 3 つの新規遺伝子発見プログラム（GRAIL, Genscan, FgenesH [63]）はコンピュータ解析の一部として実行されたが、これらのプログラムの結果は Otto 予測を作製するために直接利用はしなかった。Otto は配列類似性により 11,226 個の遺伝子を追加予測した。

3.2. Otto の実効性評価

Otto の相同性に基づくプロセスと、既知遺伝子の構造決定に Otto が用いる方法の実効性を評価するため、我々は、Otto 予測転写産物と、特異的な SIM4 アライメントがある 4512 個の RefSeq 転写産物セットから取り出された対応する（おそらく正しい）転写産物とを比較した（表 7）。Otto と Genscan の相対的性能を評価するために、3 種の比較を行った。第 1 は、対応する RefSeq 配列を除いた相同性配列データのみを用いて Otto により予測された遺伝子モデルの正確性を確かめた（表 7 の Otto-homology）。感度（正確に予測された塩基数を cDNA 全長で割ったもの）と特異度（正確に予測された塩基数を正確・不正確を含めて予測

された総塩基数で割ったもの）を測定した。第 2 には、RefSeq 配列だけに適用された Otto 予測（Otto-RefSeq)、すなわち既知遺伝子の存在注釈のために Otto が使うプロセスの感度と特異度を試験した。しかも第 3 には、これらの RefSeq 配列に対応する Genscan 予測の正確性を確かめた。期待どおり、アライメントを用いた方法（Otto-RefSeq）が最も正確で、Otto-homology は両基準にて Genscan より有能であった。しかし、真の RefSeq 塩基配列のうちの 2.7%が Otto-refseq 注釈には現れて来ず、Otto-RefSeq 転写産物中の塩基配列のうちの 6.1%がもともとの RefSeq 転写産物には含まれていないものだった。この差異は、セレラ社アセンブリと RefSeq 転写産物との間の論理構造差によるもので、遺伝子多型、セレラ社アセンブリ中の不完全データまたは不正確データ、アライメント過程で Sim4 によって導入されたエラー、または、比較に用いられたデータセット中の異なるスプライシング型の存在によって引き起こされたものであり得る。

表 7 Otto と Genscan の感度と特異度
感度と特異度は、最初に公刊されている RefSeq 転写産物に予測産物をアラインし、特異的にアラインされた RefSEq 塩基数（N）の割合を計算した。感度は公刊されている RefSeq 転写産物長に対する N の比、特異度は予測産物長に対する N の比。全ての差は統計的に有意である（Tukey HSD; p<0.001)。

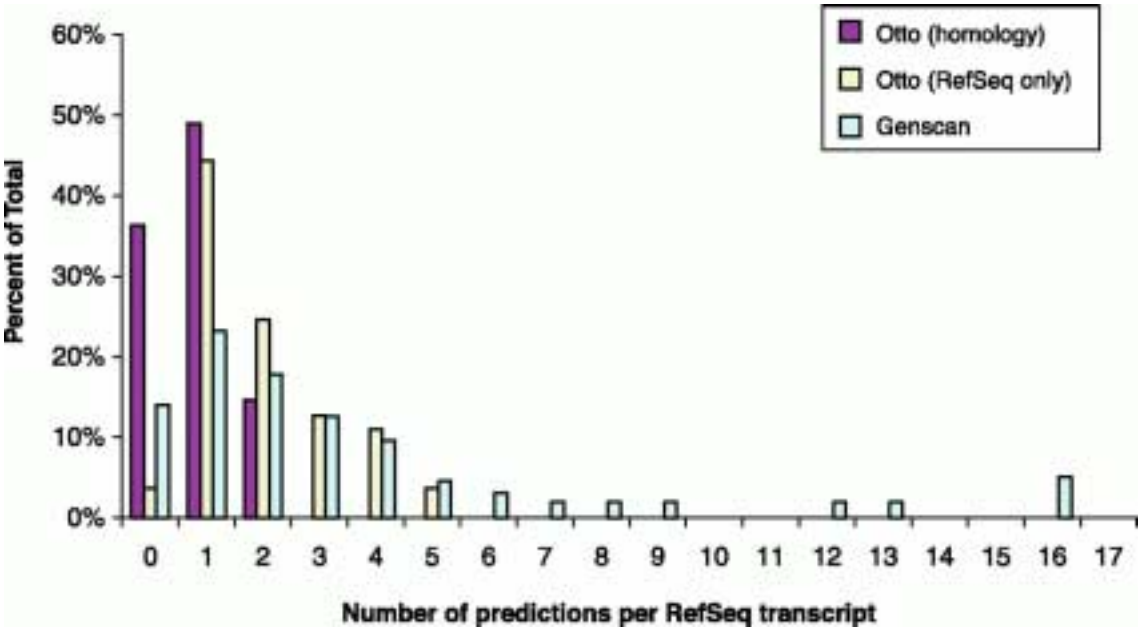
Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)†	0.604	0.884
Genscan	0.501	0.633

* Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction.

†Refers to those annotations produced by supplying all available evidence to Genscan.

Otto は遺伝子再構築に証拠に基づく手法を使うため、介在エクソンに対する実験的証拠がない場合、スプライシングにより転写産物を生じさせることができないはずのエクソンセットを偶然作り出す結果となるのかもしれない。実際に全ての証拠が単一転写産物を指し示している場合にも“遺伝子分割”をしてしまうと思われる。これらの方法が遺伝子予測を誤って分割してしまう傾向についても調べた。その傾向は図 8 に示す。Otto-RefSeq 予測および Otto-homology 予測とも、Genscan のみによる予測に比較し、既知遺伝子を断片化する傾向は低かった。

図 8 異なる遺伝子注釈法による遺伝子分割の分析
ゲノムアセンブリに対する RefSea 転写産物の Sim4 依存アライメント 4517 個が選ばれており（使用基準はテキスト参照）、Genscan 注釈、Sim-4 による精密化 RefSeq アライメントにのみ依存した Otto(RefSeq only)注釈、Otto(homology)注釈（Genscan に対して使える全ての証拠を使用して作り出された注釈）のオーバーラップする数が割り振られている。これらのデータは、単一の RefSeq 転写産物と関係する Gencan 予測、そして／または、Otto 予測の頻度をしめす。ここで示されている Otto-homology 予測が 0 のクラスは、RefSeq 転写産物に対して確認することなしに Otto-homolgy プログラムがコールをかけ、したがって証拠不十分で Otto コールが成立しなかったもの。[拡大像（13K GIF ファイル）]



3.3. 遺伝子数

Otto システムが非常に保守的なものであるので、相同性が弱い領域においては、我々は異なる遺伝子予測戦略を用いてきた。以下では他の新規遺伝子発見プログラムの予測結果を用いている。これらの遺伝子に対しては、解析を進めるにあたって、予測転写産物が以下に挙げる証拠を少なくとも 2 つ有する場合に遺伝子セットに含めることを強調しておきたい。すなわち、タンパク、ヒト EST、齧歯類 EST、またはマウスゲノム断片で一致する断片である。最後のものは、コンピュータパイプラインで使った 3 つの遺伝子発見プログラムによる予測のサブセットである。これらについては、Otto が遺伝子構造があると予測するのに十分な配列類似情報は存在しなかった。3 つの新規遺伝子発見プログラムは約 155,695 個の予測値を出し、そのうちの約 76,410 個の遺伝子が非重複性（お互いにオーバーラップしていない）であった。これら遺伝子のうち、57,935 個は既知遺伝子または Otto による予測遺伝子とオーバーラップを示さなかった。Otto 予測遺伝子とオーバーラップを示さない予測のなかで 21,350 個だけが、少なくとも 1 種類の配列相同性証拠によって部分的に支持され、8,619 個が 2 種類の証拠によって部分的に支持された（表 8）。

表 8 異なる遺伝子注釈法による遺伝子分割の分析

ゲノムアセンブリに対する RefSea 転写産物の Sim4 依存アライメント 4517 個が選ばれており（使用基準はテキスト参照）、Genscan 注釈、Sim4 による精密化 RefSeq アライメントにのみ依存した Otto(RefSeq only)注釈、Otto(homology)注釈（Genscan に対して使える全ての証拠を使用して作り出された注釈）のオーバーラップする数が割り振られている。
これらのデータは、単一の RefSeq 転写産物と関係する Gencan 予測、そして／または、Otto 予測の頻度をしめす。ここで示されている Otto-homology 予測が 0 のクラスは、RefSeq 転写産物に対して確認することなしに Otto-homolgy プログラムがコールをかけ、したがって証拠不十分で Otto コールが成立しなかったもの。

		Total	Types of evidence				No. of lines of evidence*			
			Mouse	Rodent	Protein	Human	≥1	≥2	≥3	≥4
Otto	Number of transcripts	17,969	17,065	14,881	15,477	16,374	**17,968**†	17,501	15,877	12,451
	Number of exons	141,218	111,174	89,569	108,431	118,869	140,710	127,955	99,574	59,804
De novo	Number of transcripts	58,032	14,463	5,094	8,043	9,220	*21,350*	**8,619**	4,947	1,904
	Number of exons	319,935	48,594	19,344	26,264	40,104	79,148	31,130	17,508	6,520
No. of exons per	Otto	7.84	5.77	6.01	6.99	7.24	7.81	7.19	6.00	4.28
transcript	De novo	5.53	3.17	3.80	3.27	4.36	3.7	3.56	3.42	3.16

* Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript.

† This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.

この数値（21,350）と Otto 遺伝子注釈による数値（17,764）の合計、39,114、がヒト遺伝子総数の上限に近いものである。表 8 に示したように、さらなる証拠の要求基準を厳しくすると、この数値は急減し、2 つの証拠を要求すると遺伝子総数は 26,383 個に、3 つの証拠を要求すると約 23,000 個になる。4 つのカテゴリーの証拠をすべて要求するのは厳しすぎる。新しいタンパク質（現在までに記載のないタンパク質ファミリーのメンバー）をコードしている遺伝子群を除外してしまうからである。解析のこの時点では偽遺伝子への修正はなされていない。

我々の自動遺伝子注釈法によっても他の新規遺伝子発見法によっても見つけられなかった遺伝子をさらに同定するために、予測遺伝子外の領域を調べた。EST 配列に類似し、しかもその EST 配列がスプライシングジャンクションを越えてゲノム配列に一致する領域である。予測遺伝子の 3’-UTR にあたる可能性のある領域を修正することで、約 2,500 個の領域が残った。さらに、マウスゲノム配列断片、齧歯類 EST、まはた cDNA との相同性、または既知タンパク質への相同性という証拠のうちの少なくとも一つを満たすという要求基準を追加すると、1,010 個に減少した。前述の数字にこれを加えると、ヒトゲノムでは、要求基準の厳しさに応じて、約 40,000、27,000、24,000 個の推定遺伝子数となった。表 8 は、遺伝子数と支持証拠に基づく信頼度を示す。26,383 個の遺伝子セットにコードされる転写産物を以下の解析のためにアセンブリした。このセットは、既知遺伝子との一致性に基づいて Otto により予測された 6,538 個の遺伝子、相同性に基づいて Otto により予測された 11,226

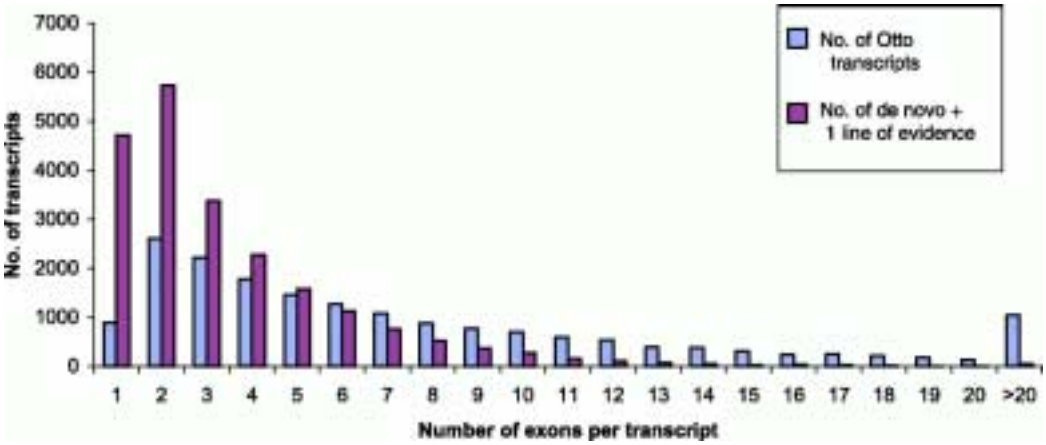
個の転写産物、他の新規遺伝子予測プログラムが 2 種類の支持証拠により予測した転写産物のサブセットに由来する 8,619 個の遺伝子を含んでいる。図 1 に、染色体図に沿って 26,383 個の遺伝子を示す。これらはごく予備的な遺伝子注釈セットであって、自動化プロセスの全ての限界下にある。遺伝子の予測と注釈の全て、そしてここに示す関連証拠は、熟練した整理者によるものではなく、完全にコンピュータプロセスの産物である。支持証拠、すなわち、既知遺伝子、タンパクまたは EST との良好な相同性、適度の相同性によって確認された他の新規遺伝子予測プログラムによる証拠、の量に基づく様々な信頼度レベルにてヒトゲノム中の遺伝子数を出しているのである。

3.4. ヒト遺伝子転写産物の特徴

我々は、ヒト DNA 配列中の“典型的”な遺伝子の平均長は約 27,894 塩基と推定している。この推定は、最も信頼度の高い遺伝子セットを示すために用いた RefSeq 転写産物がカバーする平均長に基づいたものである。

遺伝子注釈に至る転写産物セットは非常に異なる。表 8 および図 9 に示されるように、Otto 予測転写産物はより長く、平均 7.8 個のエクソンがある。しかし他の遺伝子予測プログラム由来の転写産物は平均約 3.7 エクソンである。転写産物中で我々が同定した最大エクソン数は titin mRNA の 234 個である。表 8 では Otto 転写産物と他の予測プログラム転写産物の支持証拠量を比較している。例えば、典型的 Otto 転写産物では、タンパク相同性の証拠によって支持される 7.81 エクソンのうち 6.99 エクソンを有する。予測どおり、Otto 転写産物には、他の新規遺伝子予測プログラムによる転写産物の場合よりも、一般的により多い支持証拠がある。

図 9 17,968 個の Otto 予測転写産物と、他の新規遺伝子予測プログラム由来で少なくとも一つの証拠を持つ Otto 予測とはオーバーラップしない 21,350 個の予測転写産物との間の、転写産物 1 個あたりのエクソン数の比較
両者とも、2 個のエクソン数があるカテゴリーが最大数を示すが、他の新規遺伝子予測プログラム由来の方が小さい転写産物側に寄っている。Otto 予測の場合は 19.7%が 1 個または 2 個のエクソンで、5.7%が 20 個以上ある。他の新規遺伝子予測プログラムでは、49.3%が 1 個または 2 個のエクソンで、0.2%が 20 個以上となっている。[拡大像（13K GIF ファイル）]



第 4 章 ゲノム構造

要旨

このセクションでは、非コーディング領域のアセンブリしたゲノム配列への寄与及び、予測される遺伝子セットとの関係を記載する。これらには、ゲノムの細胞遺伝学的地図における GC 含有率及び遺伝子密度の解析、CpG アイランドの計数解析、また、全ゲノムの反復配列についても簡単に述べる。

4.1. 細胞遺伝学的地図

ゲノム構造要素において最も明白で間違いなく識別可能なものは、おそらくギムザ染色によるバンド・パターンであろう。染色体バンド研究から、ヒト染色体の約 17～20%が C バンド、すなわち構造性異形染色質からなる[64]。この異形染色質のほとんどが、高度な多様性を示し、様々な高次の反復構造を伴った多様なαサテライト DNA ファミリーから構成される。多くの染色体は複雑な染色体内及び染色体間重複を動原体周辺領域に有する [66]。解析配列の約 5%がαサテライト配列であることが同定された為、アセンブリには含めなかった。動原体周辺領域の解析は進行中である。

残り約 80%のゲノム、すなわち真正染質成分は、G バンド、R バンド、T バンドに分けられる [67]。これらの細胞遺伝学的バンドは、塩基配列構成及び遺伝子密度の違いによるものと考えられてきた。しかし、今回の研究で分子レベルにて正確なバンド境界を決定することは不可能であった。T バンドは、最も GC 含有率及び遺伝子密度が高く、G バンドは GC 含有率が低い [68]。Bernardi は真正染色質について分子レベルで叙述しようとし、それは 300 キロ塩基対以上に及ぶ塩基組成の異なる長大な DNA 領域であると記載し、アイソコア（isochore；L、H1、H2、H3 で示される）と名付けた [69]。Bernardi は、低 GC 含有率（43%未満）を L（軽）アイソコア、それ以上を H（重）アイソコアと定義し、それはさらにゲノムの 24%、8%、5%に相当する 3 つの高 GC 含有率クラスに分類できるとした。これまでに、遺伝子密度は L アイソコアにて非常に低く、H2 及び H3 アイソコアにて 20 倍に濃縮されていると主張されてきた [70]。アセンブリ全体を通じて、連続整列化できる 50 キロ塩基対間隔で GC 含有率を調べたところ、GC 含有率 48%以上の領域（H3 アイソコア）は平均長 273.9 キロ塩基対であり、GC 含有率 43%以上 48%未満（H1 及び H2 アイソコア）は平均長 202.8 キロ塩基対、43%未満（L アイソコア）領域は 1078.6 キロ塩基対であることがわかった。アセンブリされた配列で、50 キロ塩基対間隔の GC 含有率と遺伝子密度の相互関係も同様に試験した（表 9 及び図 10、11）。予測どおり、遺伝子密度は低 GC 含有領域よりも高 GC 含有率領域において高かった。しかし、GC 含有率と遺伝子密度間の相互関係には予測されてきたほど偏りがなかった [69]。遺伝子は予測されていたよりも高率に低 GC 含有領域に存在

した。

表 9 isochore の GC 含有率特性

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43-48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

* The predictions were based on Bernardi's definitions (70) of the isochore structure of the human genome.

図 10 GC 含有率と遺伝子密度の相関。
青色のバーは該当する GC 含有率を示すゲノム（50-kbp 枠）の割合を示す。この各 GC 含有率範囲と関連する全遺伝子の割合は黄色のバーで示す。このグラフから、ゲノムの約 5%で GC 含有率は 50～55%であるが、これだけで遺伝子の約 15%を含んでいることが分かる。

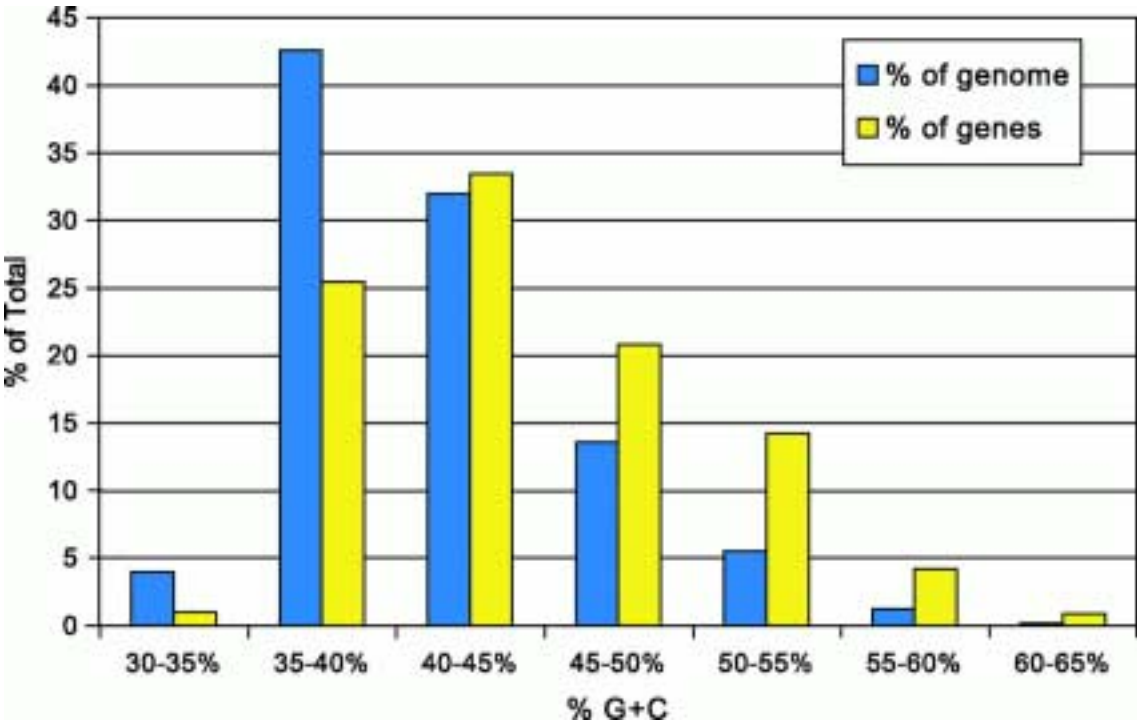
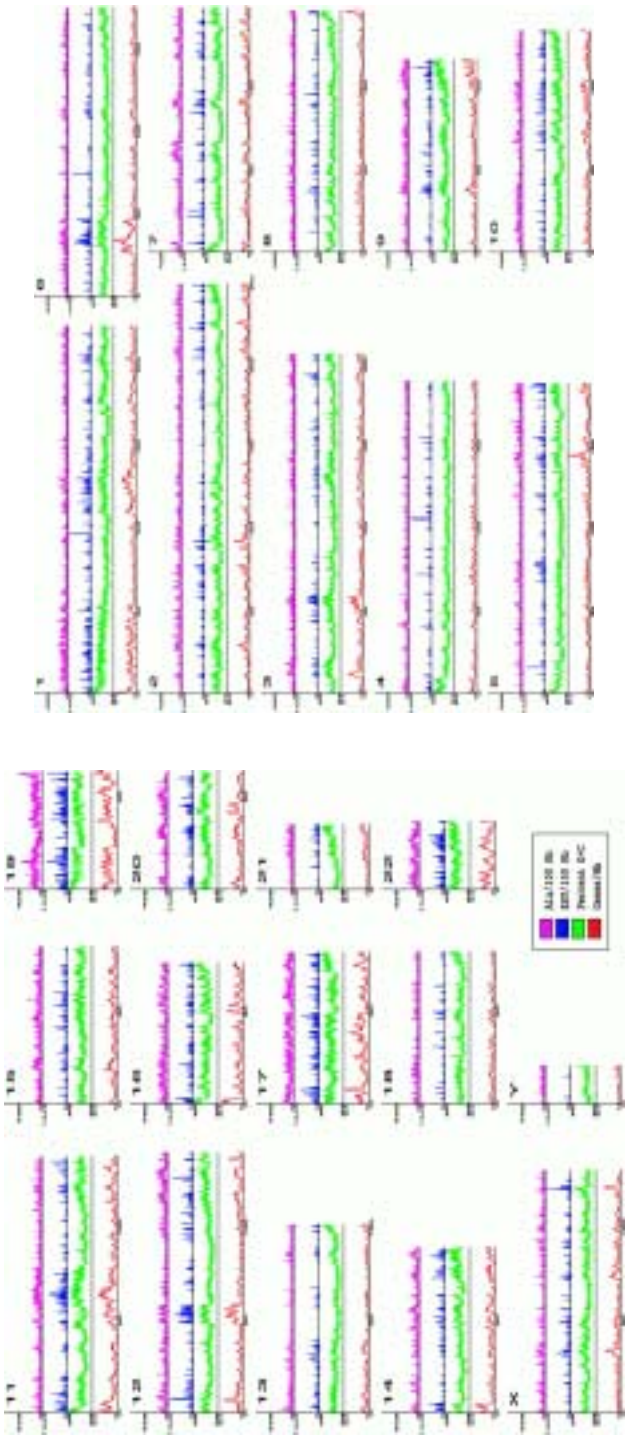


図 11 ゲノムの構造学的特徴

各染色体における遺伝子密度（橙色）、GC 含有率（緑色）、EST 密度（青色）、Alu 密度（桃色）の関係。遺伝子密度は 1-Mbp 枠で算出した。GC ヌクレオチドの割合は 100-kbp 枠で算出した。EST と Alu 配列の数は 100-kbp 枠で算出した。



不均衡な数の H3 含有バンドがある第 17、第 19、第 22 染色体が最も高い遺伝子密度を示した（表 10）。逆に、最も低い遺伝子密度を示すことがわかった X、第 4、第 18、第 13 及び、Y 染色体には H3 バンドはほとんどなかった。第 15 番染色体には H3 バンドがほとんどない

が、我々の解析では顕著な低遺伝子密度を示さなかった。さらに、低遺伝子密度であることがわかった第 8 染色体は H3 バンド分布に異常を示さないようである。

表 10 染色体の特長　De novo/any は、Otto 予測と重複せずその他が明らかに 1 つ以上ある新規予測のすべてをさす。De novo/2x は、Otto 予測と重複せずその他が明らかに 2 つ以上ある新規予測のすべてをさす。Desert（砂漠）は存在注釈遺伝子をもたない配列領域。

Chr.	Sequence coverage (CS assembly)						Base composition			Gene prediction*					Gene density (genes/Mbp)						
	Size (Mbp)	No. of scaf-folds	Largest scaf- fold (Mbp)	No. of scaf- folds >500 kbp	Se- quence covered by scaf-folds >500 kbp	% of total se-quence in scaf-folds >500 kbp	% repeat	% GC	No of CpG islands	Otto	De novo/any	De novo/2×	Total (Otto + de novo/any)	Total (Otto + de novo/2×)	Se- quence in deserts >500/kbp	Se- quence in deserts >1 Mbp	Otto	De novo/ any	De novo/2×	Otto + de novo/ any	Otto + de novo/ 2×
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	2,453	29	6	8	8	3	16	11
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,954	1,816	55	19	5	7	2	12	7
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	1,611	50	12	5	7	3	12	8
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	1,145	55	18	4	6	2	10	6
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	1,366	46	15	5	7	2	11	7
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	1,467	38	9	6	7	3	13	8
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	1,219	26	12	5	7	3	12	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7	2	11	6
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	1,018	22	9	6	7	3	13	8
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7	2	12	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	1,586	27	9	8	8	4	16	12
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	1,342	24	9	7	7	3	14	10
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7	2	10	5
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	873	34	20	7	8	3	14	10
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	8	3	15	10
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	10	9	3	19	12
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	8	4	19	15
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	7	2	10	6
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	20	9	4	29	23
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	8	7	3	16	11
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	9	6	8	3	13	8
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	0	14	9	4	23	17
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	8	5	6	3	11	7
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	2	3	8	2	11	5
U*	75	11,542	1						479	196	278	132	474	328							
Total	2907	53,591		1,059	2,490				28,519	17,764	21,350	8,619	39,114	26,383	606	208					
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	7	3	14	9

* Chromosomal assignment unknown.

哺乳類ゲノムは遺伝子的にみれば何もない砂漠に散らばる遺伝子オアシス群でできているという Ohno の仮説 [71] は妥当なのだろうか。確かにヒトゲノムは遺伝子不在の砂漠、あるいは遺伝子数の貧弱な莫大な領域を含んでいるようである。仮に 500 kbp 以上にわたり遺伝子が存在しない領域を砂漠と定義すると、605 Mbp、すなわちゲノムの約 20%が砂漠である。これら砂漠は、染色体間に一様に分布しているわけではない。遺伝子に富む第 17、第 19、第 22 染色体では、合計 171 Mbp の約 12%のみが砂漠中に存在する。一方、遺伝子があまり存在しない第 4、第 13、第 18、X 染色体は 492Mbp の 27.5%が砂漠中に存在する（表 11）。これらの領域で遺伝子が存在しないと明確に予測されても、生物学的機能が全くないとは必ずしも言えない。

表 11 ゲノムの概観

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8[*]
Percent of base pairs spanned by exons	1.1 to 1.4[*]
Percent of base pairs spanned by introns	24.4 to 36.4[*]
Percent of base pairs in intergenic DNA	74.5 to 63.6[*]
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

[*] In these ranges, the percentages correspond to the annotated gene set (26, 383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively

4.2. 連鎖地図

連鎖地図は遺伝学的解析に基盤を与え、遺伝特性の研究及び遺伝子のポジショナル・クローニングにおいて幅広く用いられている。距離測定基準、すなわちセンチモルガン（cM）は、減数分裂における相同染色体間の組換え率に基づくものである。一般的に、組換え率は雄性より雌性において高率だが、これによる地図書き込みの増大程度は全ゲノムで一様ではない [72]。連鎖地図及び細胞遺伝学的地図は、ゲノム解析及び遺伝学的解析において広く用いられてきた。今回、ほぼ完全なゲノム配列が決定されたことで、究極の物理地図を作製する機会が得られ、これら 2 つの地図との対応を全面的に解析することが可能となった。これより、ゲノムプロジェクトのマッピング相と配列決定相の間の堂々巡りが終焉を迎えることとなった。

今回の研究で、全ゲノムに対し Genethon 連鎖地図の構成マーカーの位置を決定した。表 12 に示すように、3 Mbp ごとに組換え率を計算、1 Mbp あたりの cM で表した。これまで染色体テロメア領域にて組換えが高率に生じることが記録されてきた [73]。今回のマッピング結果より、最高組換え率と最低組換え率には 4.99 cM/Mbp の差が、雄性と雌性間では最大 4.4 cM/Mbp の差（第 16 染色体では 4.99 cM/Mbp 対 0.47 cM/Mbp）が生じた。このことから、ゲノム領域間の組換え率における変化度は、雄性と雌性間の組換え率の差を超えることが示された。ヒトゲノムには組換えを生じやすいホットスポットが存在し、その領域では組換え率が 1 kb 違っただけで 5 倍以上変化する。従って、組換え率の変化度は、どのくらいのサイズの塩基対ごとに検証するかに依存する。不幸にも、CEPH（Centre d'Étude du Polymorphism Humain）の参照家系と他の参照家系では、減数分裂クロスオーバーをほとんど生じないため、約 3 Mbp 以上の解像度を得ることは出来なかった。次の挑戦は染色体レベルにて組換えが起こる配列を決定することであろう。任意のマーカー間で組換え率の変化度を正確に予測することは、ポジショナル・クローニング・プロジェクト等にて連鎖領域を絞りこむマーカー設計上、極めて有用になるものと思われる。

表 12 全ゲノムにおける物理的距離毎の組換え率（cM/Mb）

CSA でマップしたアセンブリ上に Genethon マーカーをおき、相対的物理的距離と組換え率を染色体ごとに 3-Mb 枠で算出した。NA は「該当なし」を示す。

Chrom.	Male			Sex-average			Female		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
1	2.60	1.12	0.23	2.81	1.42	0.52	3.39	1.76	0.68
2	2.23	0.78	0.33	2.65	1.12	0.54	3.17	1.40	0.61
3	2.55	0.86	0.23	2.40	1.07	0.42	2.71	1.30	0.33
4	1.66	0.67	0.15	2.06	1.04	0.60	2.50	1.40	0.77
5	2.00	0.67	0.18	1.87	1.08	0.42	2.26	1.43	0.62
6	1.97	0.71	0.28	2.57	1.12	0.37	3.47	1.67	0.64
7	2.34	1.16	0.48	1.67	1.17	0.47	2.27	1.21	0.34
8	1.83	0.73	0.14	2.40	1.05	0.46	3.44	1.36	0.43
9	2.01	0.99	0.53	1.95	1.32	0.77	2.63	1.66	0.82
10	3.73	1.03	0.22	3.05	1.29	0.66	2.84	1.51	0.76
11	1.43	0.72	0.31	2.13	0.99	0.47	3.10	1.32	0.49
12	4.12	0.76	0.26	3.35	1.16	0.49	2.93	1.55	0.59
13	1.60	0.75	0.01	1.87	0.95	0.17	2.49	1.19	0.32
14	3.15	0.98	0.18	2.65	1.30	0.62	3.14	1.63	0.75
15	2.28	0.94	0.34	2.31	1.22	0.42	2.53	1.56	0.54
16	1.83	1.00	0.47	2.70	1.55	0.63	4.99	2.32	1.12
17	3.87	0.87	0.00	3.54	1.35	0.54	4.19	1.83	0.94
18	3.12	1.37	0.86	3.75	1.66	0.43	4.35	2.24	0.72
19	3.02	0.97	0.10	2.57	1.41	0.49	2.89	1.75	0.87
20	3.64	0.89	0.00	2.79	1.50	0.83	3.31	2.15	1.34
21	3.23	1.26	0.69	2.37	1.62	1.08	2.58	1.90	1.18
22	1.25	1.10	0.84	1.88	1.41	1.08	3.73	2.08	0.93
X	NA	NA	NA	NA	NA	NA	3.12	1.64	0.72
Y	NA	NA	NA	NA	NA	NA	NA	NA	NA
Genome	4.12	0.88	0.00	3.75	1.22	0.17	4.99	1.55	0.32

4.3. CpG アイランドと遺伝子間の相互関係

CpG アイランドとは、ゲノム全体を通じて比較した場合、高頻度に CpG のジヌクレオチド含む、メチル化を受けていない DNA 伸張鎖である [74]。CpG アイランドは遺伝子の転写開始部位に選択的に存在すると信じられ、ほとんどのハウスキーピング遺伝子がトランスクリプト 5’端に CpG アイランドを持っている [75, 76]。更に、実験的証拠から CpG アイランドのメチル化が遺伝子不活化と相関し [77]、遺伝子刷込み [78] および、組織特異的遺伝子発現 [79] に重要であることが示されてきた。

実験的手法を用いることで、ヒト・ゲノムには推定 30,000 個から 45,000 個 [74, 80]、ヒト第 22 染色体には推定 499 個の CpG アイランドがあるとされている [81]。Larsen ら [76]、および Gardiner-Garden と Frommer [75] は、コンピュータを用いた方法で CpG アイランドを同定し、G+C 含有率 50%以上かつ CpG ジヌクレオチドの実測頻度対予測頻度の比が 0.6 以上となる 200 塩基対以上の DNA 領域と定義した。

CpG アイランドの実験的定義と計算的定義との直接比較は困難である。コンピュータを用いた方法はシトシンのメチル化状態を考慮せず、実験的手法では高 GC 含有率を有する領域を直接選別しないからである。しかし、存在注釈を付けられているゲノムの転写産物と全ゲノム配列のセットがあれば、CpG アイランドと遺伝子転写開始部位の相関を決定することができる。我々は、第 22 番染色体の公表されている遺伝子存在注釈について、今回のアセンブリによる全ヒトゲノムと計算機で存在注釈を付けた遺伝子を用いる場合と同様にして解析を行い、CpG アイランド計算の違いを Larsen らと比較した [76]。主要相違点は、我々は 200 塩基対ごとに計算する範囲（ウインドウ）を設けてそれをスライドさせていき、連続するウインドウがオーバーラップした場合のみ、それらを併合した。併合に際しては CpG 値を再計算し、設定閾値以下ではアイランドの可能性があっても却下した。

様々な CpG 統計値を計算するため、CG ジヌクレオチドの見込み率に 2 つの異なる閾値を用いた。元々の閾値 0.6 を用いる場合（方法 1）の他に、CG ジヌクレオチド見込み率の高い閾値 0.8 を用いた（方法 2）結果、第 22 番染色体上の CpG アイランド数が同染色体上で存在注釈が付いている遺伝子数に近づいた。表 13 に主要結果をまとめる。方法 1 で計算した CpG アイランドは CSA 配列のわずか 2.6%を CpG アイランドと予測するが、転写開始部位（開始コドン）の 40%が CpG アイランド内に存在する。これは他グループが報告した比率に匹敵する [82]。表 13 の最後の 2 列は、第 1 エクソンから最近傍 CpG アイランドまでの実測および予測平均距離を示すものである。最近傍 CpG アイランドまでの実測平均距離は対応する予測距離よりも短かく、これは CpG アイランドと第 1 エクソン間に関係があることを確認するものである。

表 13 異なる 2 つの方法を用いた場合に第 22 番染色体（配列長 34-Mbp）および全ゲノム（配列長 2.9-Gbp）において同定された CpG アイランドの特徴

方法 1 では CG 見込み率 0.6 以上、方法 2 では 0.8 以上を用いた。

	Chromosome 22		Whole genome (CS assembly)	
	Method 1	Method 2	Method 1	Method 2
Number of CpG islands detected	5,211	522	195,706	26,876
Average length of island (bp)	390	535	395	497
Percent of sequence predicted as CpG	5.9	0.8	2.6	0.4
Percent of first exons that overlap a CpG island	44	25	42	22
Percent of first exons with first position of exon contained inside a CpG island	37	22	40	21
Average distance between first exon and closest CpG island (bp)	1,013	10,486	2,182	17,021
Expected distance between first exon and closest CpG island (bp)	3,262	32,567	7,164	55,811

同様に、遺伝子間領域、イントロン、エクソン及び第 1 エクソン等の様々な配列クラスにおける CpG アイランド・ヌクレオチド分布を調べた。個々の配列クラスに対し、その配列クラスにおける CpG アイランド・ヌクレオチドの実測分と予測分の比として見込み値を計算した。CSA 配列に方法 1 を用いた結果、遺伝子間領域 0.89、イントロン 1.2、エクソン 5.86、第 1 エクソン 13.2 の値を得た。第 22 番染色体でも、両データセットに対し高閾値を適用しても（方法 2）、同様の傾向が観察された。要するに、全ゲノム解析は初期の解析を拡大したものとなっており、CpG アイランドと第 1 エクソン間には強力な相関があることを示唆するものである。

4.4. 全ゲノム反復配列

様々な種類の反復 DNA がゲノムに占める割合を表 14 に示す。これまでに報告された値 [83] に非常に近い約 35%のゲノムがこれら反復配列クラスに含まれることがわかった。上述したように、セレラ社アセンブリでは反復配列の解像度が不完全な結果、反復配列は少なく示されているものと思われる。アセンブリ骨格長の約 8%分がギャップ中に存在し、これらのほとんどが反復配列と考えられる。第 19 番染色体には最高の遺伝子密度があるが、同様に最高の反復配列密度（57%）がある（表 10）。興味深いことに、反復配列の様々なクラスの中で、Alu 配列と遺伝子密度に明確な相関が見られる。これは LINE 配列では見られない。

表 14 区画化ショットガン・アセンブリ配列における反復 DNA の分布

Repetitive elements	Megabases in assembled sequences	Percent of assembly	Previously predicted (%) (83)
Alu	288	9.9	10.0
Mammalian interspersed repeat (MIR)	66	2.3	1.7
Medium reiteration (MER)	50	1.7	1.6
Long terminal repeat (LTR)	155	5.3	5.6
Long interspersed nucleotide element (LINE)	466	16.1	16.7
Total	1025	35.3	35.6

第 5 章．ゲノムの進化

要約

ゲノムの進化におけるダイナミックな性質はいくつかのレベルで捉えることができる。これらのレベルには、RNA 中間体（レトロトランスポジション）とセグメントゲノム重複（segmental genomic duplication）を介する遺伝子重複などが含まれる。この章では、機能遺伝子（無イントロンパラログ（intronless paralog））や不活性遺伝子（偽遺伝子）を生み出す、ゲノム全体でのレトロトランスポジションの発生について述べる。翻訳プロセスと核の調節に関わる遺伝子は、我々の調査で検出された、全ての無イントロンパラログとプロセシングされた偽遺伝子の約 50%を占める。また、セグメントゲノム重複の範囲を分類整理し、1,077 個の重複したブロック配列が 3,522 種類の遺伝子をカバーしている証拠を見いだした。

5.1. ヒトゲノム中のレトロトランスポジション

スプライシング処理等のプロセシングされた mRNA 転写産物がゲノムに逆に転写され組み込まれる（レトロトランスポジション）と、結果的に、無イントロンパラログと呼ばれる機能遺伝子か、不活性遺伝子（偽遺伝子）が生じる。パラログとは、重複のために、1 つの生物に 2 コピー以上存在する遺伝子のことである。機能的に類似した、または同一なタンパク質をコードする遺伝子が、イントロンを含む形とイントロン無しの両方の形で存在することは、これまでにも報告されている [84,85]。これらの進化的事象をゲノムの全容にわたって分類整理することは、細胞生物学においてこのような遺伝子重複の機能的な結果を理解する上で有用である。マウスやその他の哺乳類ゲノムにおける、配列保存された無イントロンパラログの同定は、これらの転位の進化時代史を理解する基礎となり、哺乳類の進化的放散（mammalian radiation）における遺伝子欠失と蓄積に対する洞察を与える。

Otto 予測による全 901 個の単一エクソン遺伝子（single-exon genes）がコードするタンパク質セットを、その他の複数エクソン遺伝子から予測される転写産物がコードするタンパク質に対して BLAST 解析した。相同性基準を、配列全長の 90%で配列同一性 70%と設定したとき、単一エクソン遺伝子と複数エクソン遺伝子との一致例が 298 個見いだされた。これら 298 個の配列のうち、97 個は、実験的に確認された GenBank の全長遺伝子データセットに、特定の厳密度下でみられ、手作業による調査で確認された。

我々は、これらの 97 個の遺伝子が、既知遺伝子の無イントロンパラログを表しているのではないかと考えている（www.sciencemag.org.cgi/content/full/291/5507/1304/DC2 の Science Online にある Web の表 1 参照）。これらのほとんどには、直接繰り返し配列が接しているが、これらの繰り返し配列の正確な性質はまだ明らかにされていない。我々が自信を持てるケ

ースには全て、レトロトランスポゾンの特徴であるポリアデニン化〔poly(A)〕テールがある。

機能する無イントロンパラログの事象を述べた最近の研究発表では、レトロトランスポジションが X 染色体不活性化を逃れる機構として用いられている可能性を推測している[84,86]。我々は、これらのレトロトランスポジションした遺伝子が X 染色体起源であるという傾向は認めていない。むしろ、これらの結果は、イントロンを含むパラログも、それに対応する無イントロンパラログも、ランダムに染色体に分布していることを示している。我々は、元になる単一の染色体から、複数のターゲット染色体へレトロトランスポジションした例を複数見いだした。興味深い例としては、第 13 染色体上にある 5 つのエクソンを持つリボゾームタンパク質 L21 遺伝子が、別々に、第 1、3、4、7、10、14 番染色体にレトロトランスポジションしていた例があった。レトロトランスポジションの元遺伝子のサイズも多様である。最大の例は、第 11 番染色体上にある、31 個のエクソンを持つジアシルグリセロールキナーゼ・ゼータ遺伝子であり、第 13 番染色体に無イントロンパラログがある。経路に関わりなく、コード領域や非コード領域に二次的な遺伝子変化を伴い、異なる機能発現パターンを導くレトロトランスポジションは、哺乳類の機能的な能力のレパートリーを増やす重要経路の代表である[87]。

我々の、レトロトランスポジションした無イントロンパラログの予備的なセットには、翻訳過程（リボゾームタンパク質 40%、翻訳伸長因子 10%）と、核の調節（HMG 非ヒストンタンパク質 4%)、そして代謝酵素と調節酵素に関わる遺伝子が明らかに多すぎる存在割合を示している。無イントロンパラログのサブセットに特異的に一致する EST 群は、これらの無イントロンパラログが発現していることを示唆している。元遺伝子と、それらの無イントロンパラログ間の上流調節配列の違いが、組織特異的遺伝子発現を説明すると考えられた。もしそうだとしても、これらのプロセシングされた遺伝子が機能するように発現され、翻訳されることを明らかにするためには、さらなる解明と実験による実証が必要である。

5.2. 偽遺伝子

偽遺伝子は機能のないコピーであり、正常遺伝子に非常によく似ているがわずかに異なっていて、発現されない。我々は、ヒトゲノムにあるプロセシングされた偽遺伝子の予備的な解析法を開発した。これは、遺伝子の不活性化の原因である現存する進化圧力を解明するための出発点となる。これらのプロセシングされた偽遺伝子の一般的な構造的特性には、元の機能遺伝子にある介在配列の完全欠損、3’末端の poly（A）系列領域、偽遺伝子配列の前後にある直接反復配列などがある。プロセシングされた偽遺伝子はレトロトランスポジ

ションの結果生じるが、プロセシングされていない偽遺伝子は部分的なゲノム断片の重複により生じる。

無イントロンパラログの検出と同様なホモロジー基準を用いて、BLAST により、17,764 個の Otto 予測転写産物をゲノム配列に対して解析した。そして、転写産物全長の 70%の配列のうち、70%以上の同一性を示す領域を解析した。プロセシングされた偽遺伝子と考えられる例が計 2,909 個検出された。偽遺伝子を検索する特異的な方法はまだ使われたことがないため、この数字は少なく見積もられている可能性が高い。

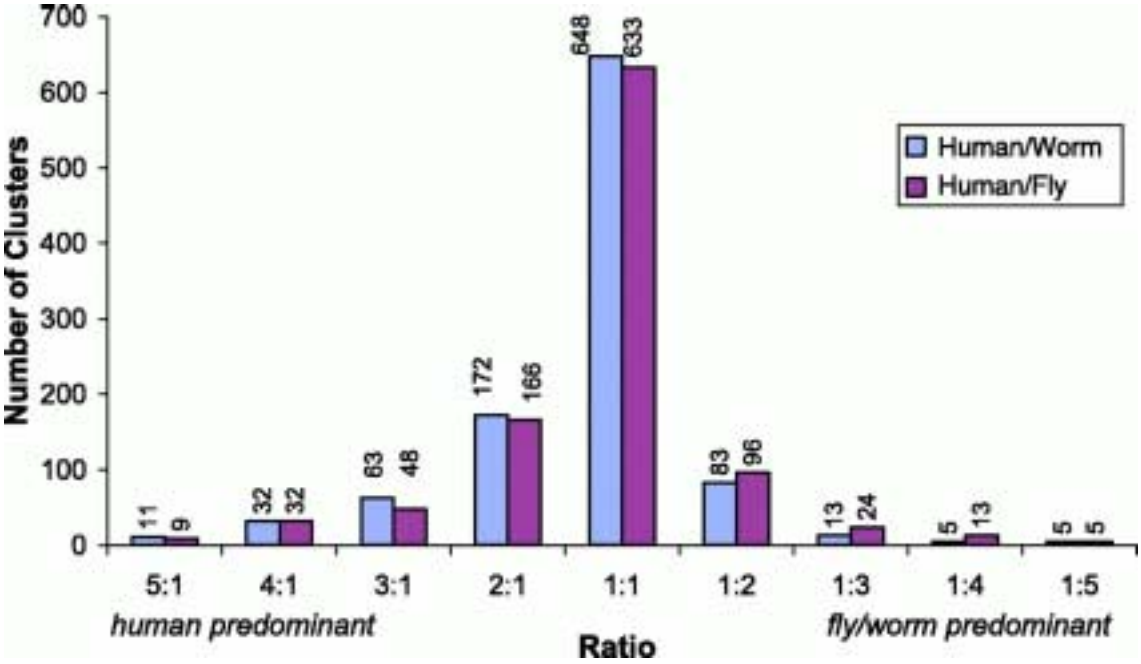
我々は、ヒトゲノムの構造的要素とレトロトランスポジションの傾向との関係を探索した。GC 含量と転写産物の長さを、プロセシングされた偽遺伝子（1177 個の元遺伝子）と、予想される遺伝子セットの残りの遺伝子との間で比較した。プロセシングされた偽遺伝子の存在を予想させる遺伝子の転写産物を、対応する偽遺伝子が検出されなかった遺伝子群のそれと比較したところ、転写産物の平均長が短かった（Otto 予測遺伝子セットでは 1027 bp 対 1594 bp）。最近の報告 [88] とは異なり、全般的な GC 含量には、有意な違いがみられなかった。プロセシングされた偽遺伝子として存在する傾向のある遺伝子ファミリーがあることは明らかである。これらの遺伝子には、リボソームタンパク質（67%）、lamin 受容体（10%）、翻訳伸長因子アルファ（5%）、そして HMG 非ヒストンタンパク質（2%）が含まれる。翻訳と核の調節に関わる遺伝子の間でのレトロトランスポジション発生率の増加（無イントロンパラログとプロセシングされた偽遺伝子を含む）は、これらの遺伝子の転写活性の増加を反映している可能性がある。

5.3. ヒトゲノムでの遺伝子重複

以前報告した手法 [27] を構築したときに、我々は、予想されるヒトタンパク質セットをグループ化してタンパク質ファミリーにする、Lek と呼ばれるグラフ理論的なアルゴリズムを開発した [89]。Lek クラスター法で作られた完全クラスターは、タンデム重複などの他の方法とは対照的に、タンパク質ファミリーの拡張における、全ゲノムや染色体の重複の役割を比較する基礎となる。各完全クラスターは、相同性からみて閉鎖的な特定の島状集団を表すため、しかも Lek は同時に複数の生物のタンパク質群で補足的なクラスター化ができるため、ある完全クラスター形成に、各動物が貢献できるタンパク質の数は、各動物ゲノムの遺伝子注釈の品質レベルによって決まる信頼度に応じて、予測できる。それから各動物による各クラスターへの貢献度の分散値を計算し、大規模重複と小規模重複との相対的重要性、タンパク質ファミリーの生物特異的な拡張現象と縮小現象（1 つの生物内での個々のタンパク質ファミリーに対する自然選択の結果と考えられる）を評価できるようにする。図 12 に示したように、完全クラスター内で、キイロショウジョウバエおよび線虫と比較し

た時に、ヒトタンパク質の相対数に大きな分散が認められたことは、これら 3 つの動物種の各ゲノムで、遺伝子ファミリーの相対的な拡張が複数回生じたことにより説明できるかもしれない。このような拡張は、観察されるとおり、ヒト対ハエ、ヒト対虫のクラスターの比が 1:1 でピークになるような分布を示し、ヒト優位側もハエ／虫優位側も同じようにカバーするスロープの広がり方を示す（図 12）。さらに、ヒトにはより多いタンパク質があるにもかかわらず、虫とハエのタンパク質が優位を占めるところでは、ほとんど同じだけ多くのクラスターが存在する。額面どおりにとれば、この解析は、個々のタンパク質ファミリーに作用する自然選択が、ヒトのタンパク質セットの少なくともいくつかの構成要素を拡張させるための主な力であったことを示唆する。しかし、我々の解析では、全ゲノムの重複が起こった後、遺伝子の欠失が生じたか、ばらばらに重複が生じたかの差が、容易には区別できない。これらのシナリオを見分けるため、より進んだ解析を行った。

図 12 完全タンパク質クラスターにおける遺伝子重複
ヒト、ムシ、ハエの予測されたタンパク質セットを Lek クラスター解析にかけた [27]。ヒト対ムシ、ヒト対ハエでみたとき、クラスターあたりで異なる比を持つクラスターの数（全体数）をプロットした。



5.4. 大規模重複

2 つの独立した方法を用いて、ヒトゲノムの大規模重複を検索した。第 1 に、高度に保存された重複ブロック配列を同定する、タンパク質ファミリーを基礎とした方法を述べる。第 2 には、全染色体間でブロック配列の重複を同定する包括的な方法を述べる。後者により、全 24 個の染色体をカバーする、多数の重複した染色体分節を同定できた。

第 1 の方法は、ゲノムの 2 箇所以上の場所にある、高度に保存された相同タンパク質のブロック配列を探索するという考え方に基づいている。この比較では、それらのタンパク質が同じファミリー、同じ完全 Lek クラスターに属する場合、2 つの遺伝子は同等であると考えられた（本質的にパラログといってよい遺伝子）[89]。当初、各染色体は、予想される遺伝子のスタートコドンによって、染色体に沿って並べられた遺伝子の紐として表されていた。我々は、2 本の DNA ストランドを単一の紐として考えた。大規模の複製では、部分的逆転が比較的よく起こるためである。各遺伝子に、タンパク質ファミリーと Lek 完全クラスターに従って索引を付けた [89]。索引を付けた遺伝子の紐の全ペアを、Smith-Waterman アルゴリズムを用いて前向き後向き両方の方向に並べた [90]。同じ Lek 完全クラスターの 2 つのタンパク質のマッチ（配列一致）には、スコア 10 を与え、ミスマッチにはスコア–10 を与えた。ギャップ開裂と伸張にはペナルティーとして–4 と–1 を与えた。これらのパラメータを用いたところ、19 個の保存された染色体間重複ブロック配列が認められ、それらは全て、以下に示した包括的な方法によっても検出され展開された。比較的少ない数しかブロック重複の検出ができなかったのは、完全 Lek クラスターの保守的な制約に基づいて、保守的にならざるを得ない方法を用いた結果である。

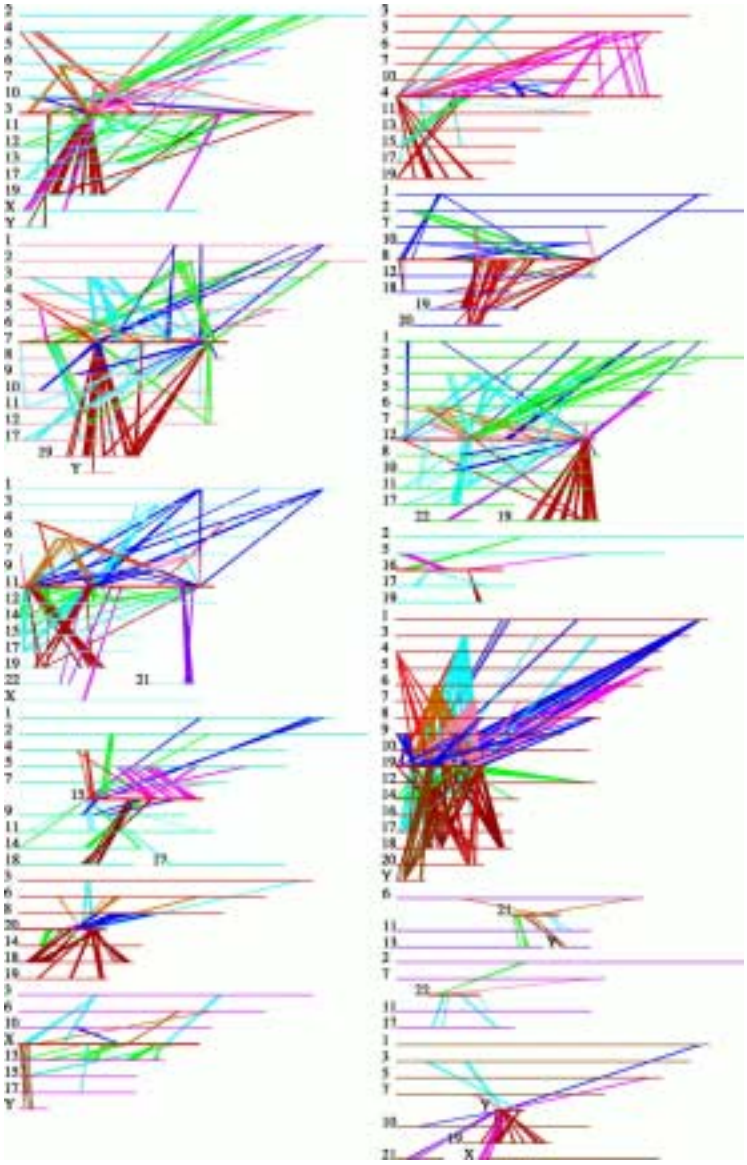
第 2 の、より包括的なアプローチでは、MUMmer システムに基づいたアルゴリズムを用いて、全染色体を直接アライメントした [91]。このアライメント法は、suffix tree データ構造と線形時間アルゴリズムを用いて、長い配列を非常に早くアライメントする。例えば、100 Mbp の染色体 2 本なら、4 ギガバイトのメモリーで、20 分未満でアライメントできる（Compaq Alpha コンピュータを使用）。この手法は最近、シロイヌナズナ *A. thaliana* の 5 つの染色体間の多数の大規模なセグメント重複を同定するのに用いられた [92]。この生物では、この方法により、ゲノムの 60%（66 Mbp）が、24 個の非常に大きい重複セグメントで覆われていることが明らかになった。シロイヌナズナでは、DNA を基礎としたアライメントで、十分、染色体間のセグメント重複を明らかにできたが、ヒトのゲノムでは、全染色体レベルでの DNA アライメントは感度があまり高くない。そこで、以下のような改変法が開発され適用された。第 1 に、26,588 個の全タンパク質（9,675,713,000,000 アミノ酸）を、ストランドの位置に関わりなく、24 個の染色体のそれぞれに沿って生じる通りに、縦に順々につないだ。次に、鎖状につなげられたタンパク質セットを、MUMmer アルゴリズムにより、各染色体に対して並べた。結果として得られた配列マッチをクラスター化し、2 つの異なる染色体上に近接して存在する 3 つ以上のタンパク質の配列マッチを持つ全セットを抽出した [93]。これらはセグメント重複の候補配列である。一連のフィルターが開発され、これらのセットから、偽陽性と考えられるものを除去するために適用された。例えば、多数のタンパク質にわたって広がっている小さなブロック配列は除去された。フィルター方法を改良するため、ごちゃごちゃに並び替えたタンパク質セットを作製した。これは、26,588 個のタンパク質をとり、その順序を無作為化し、真のゲノムと同じ数のタンパク質を含むように、24

個の並び替えた染色体に分けて作製した。このごちゃ混ぜタンパク質セットは、真のゲノムと同じ構成を持つ、すなわち、全てのタンパク質と全てのドメインが同じ回数現れる。次に、完全なアルゴリズムを、真のデータおよびごちゃ混ぜデータに対して適用した。ごちゃ混ぜデータに対する結果は、偽陽性率を推測するために用いた。フィルター処理後のアルゴリズムにより、3,522 種類の識別可能遺伝子を含む 1,077 個の重複ブロック配列中、10,310 個の遺伝子ペアが得られた。ブロック配列の多くにタンデムに重複された拡張領域があることから、遺伝子ペアが識別可能遺伝子よりも過剰なことが説明できる。対照的に、ごちゃ混ぜデータでは、370 個の遺伝ペアのみが見いだされ、偽陽性の推測値は 3.6%であった。1,077 個のブロック配列重複を最も説明できると考えられるのは、過去のセグメント重複である。多くの場合、タンパク質の順序は並び替えられているが、近接性は保存されている。1,077 個のブロック配列のうち、159 個は 3 つの遺伝子しか持たず、137 個は 4 つの遺伝子を持ち、781 個は 5 個以上の遺伝子を持っていた。

検出された重複の程度を説明するために、図 13 に、各染色体に索引付けされた全 1,077 個のブロック配列の重複を、24 個のパネルとして示す。それぞれ、索引を付けた染色体にマッピングされた重複のみを示す。この図により、重複がゲノム中に普遍的にあることが明白になった。この図が示している 1 つの特徴は、多くの比較的小さな染色体の範囲が、1 つから多数へ、という重複関係を持つことであり、図形的に明らかである。解析で得られたそのような例の 1 つは、よく調べられた嗅覚受容体（OR）ファミリーである。これは、ゲノム全体にわたるブロック配列に散在していて、いくつかの進化段階における散開ゲノム再構築（genome-deployment reconstructions）のために解析されている [(94)]。図 13 は、いくつかの染色体、例えば第 2 番染色体が、検出された大規模な重複を他よりも多く含んでいることも示している。実際、最大の重複セグメントの 1 つは、第 2 番染色体上にある 33 個のタンパク質からなる大きなブロック配列であり、2p 上の 8 個のより小さいブロックに広がっており、1 つの転位を含む第 14 番染色体上のパラログ様のセットにアライメントされる（図 13 の第 2、14 番染色体のパネル参照）。タンパク質は隣接していないが、第 2 番染色体の 97 個のタンパク質を含む領域と第 14 番染色体の 332 個のタンパク質を含む領域を占めている。この多数のタンパク質重複が偶然観察される確率は、この長さのスパンを通じても、2.3×10^{-68} である [(93)]。この重複したセットは第 2 番染色体の 20 Mbp、第 14 番染色体の 63 Mbp に占めており、後者の染色体では 70%以上である。第 2 番染色体は、ほぼ同じ大きさのブロック配列重複も持っており、これは、染色体長腕 2q と第 12 番染色体に共有されている。この重複には、4 つの既知の Hox 遺伝子クラスターのうち 2 つが含まれているが、1 対の染色体腕の近位と遠位の間で、重複の程度がかなり拡張している。この重複の幅は、他の 2 つの Hox クラスターがのっている 2 つの染色体上にもみられる。

図 13 ヒトゲノムの染色体間のセグメント重複

24 個のパネルは、合計 10,310 対の遺伝子を含む、1,077 個の重複した遺伝子のブロックを示す。それぞれの線は、1 つのブロックに属する相同遺伝子のペアを示す。全てのブロックは、それらがのっている染色体上に 3 個以上の遺伝子を持っている。各パネルは、一つの染色体と、共有ブロックを持つ他の染色体との重複を全て示す。各パネルの中央にある染色体は、強調するため、太い赤線で示してある。他の染色体は、染色体番号の順に並べて、各パネル内に上から下まで示してある。挿入図（下、中央右）は、第 18 番染色体と 20 番染色体の間の 1 つの重複の拡大図を示す。拡大して示した 64 個の遺伝子対のうち 12 個の遺伝子名を示した。



第 18 番染色体と第 20 番染色体の間のもう一つの大型重複は、他で観察される大型重複に共通する特徴のいくつかを説明するよい例である（図 13、挿入図）。この重複には、検出され並べられた、染色体内の対となる相同遺伝子が 64 個含まれている。大きな挿入配列である可能性が高く、第 20 番染色体との一致がない第 18 番染色体上の 40 Mb の領域（図 13 の第 18 番染色体上にある「Krup rel」および「collagen rel」の遺伝子領域の間）を除外すると、完全に重複したセグメントは第 18 番染色体の 36 Mb をカバーしており、第 20 番染色体の 28 Mb をカバーしている。この測定では、重複セグメントは各染色体の正味の長さのほぼ半分を占めている。最も可能性の高いシナリオは、「この領域の全長が、非常に大きい 1 つのブロック配列として重複され、小規模の組換えによる並べ替えが行われた」というものである。そうならば、重複したセグメント間に認められる相対的な挿入配列と逆位を説明するためには、少なくとも 4 回の後続した組換えが起こることが必要である。このアライメント中の 64 個のタンパク質の対は、第 18 番染色体の 217 個のタンパク質部位の中にあり、そして第 20 番染色体の 322 個のタンパク質部位の中にあって、関与するタンパク質の密度は 20〜30%である。これは、太古に大規模重複が起こり、その後染色体の片方または両方の遺伝子欠失がおこったことと矛盾しない。重複が起きた後、遺伝子対の片方のみが欠失すると、ブロック配列の遺伝子対の発見がうまくいかないであろうし、染色体上の 50%未満の遺伝子が欠失したならば、ここで観察された重複密度となるであろう。検出されたアライメントが意味あるものだといえる別個の確認法としては、この重複におけるアライメントされたタンパク質の対のうちかなりの数が、注釈を付けたものを含めて（図 13）、小さい Lek 完全クラスター（前述参照）に入り込んでいることを見ればよい。これは、それらが非常に小さいパラログファミリーの一員であることを示す。それらがゲノム中で相対的にはわずかしかないということが、そのアライメントが独自性がありしっかりしたものであることを保証する。

大規模重複の多くには、他に 2 つの質的な特徴が観察されている。第一に、疾患に関連したいくつかのタンパク質は、OMIM（Online Mendelian Inheritance in Man）アサイメント（割付確認）で、重複セグメントの一員であることがわかっている（www.sciencemag.org/cgi/content/full/291/5507/1304/DC1 の Science Online にある Web の表 2 参照）。我々は、両方の重複セグメントにあるパラログが、同様な病状に関与している少数の例を見いだした。これらの遺伝子の中で注目に値するのは、出血性疾患に関わるホメオスタシスに関与するタンパク質（凝固因子）、発育障害に関与するホメオボックスタンパク質のような転写調節タンパク質、そして心臓血管の異常に関与するカリウムチャンネルである。これらの疾患遺伝子に関しては、重複セグメント内のパラログ様遺伝子を徹底的に研究することで、疾患の原因に関して新しい洞察が得られる可能性がある。これらの遺伝子が同一のまたは類似した遺伝疾患に関与するかどうかを判断するにはさらなる研究が必要である。第二に、第 18、20 番染色体アライメントには、特異的な大型重複があったと予

想される、保存された多数のタンパク質とコードエクソンがあるが、この特異的領域にある第 18 番染色体のゲノム DNA は、場合によっては対応する第 20 番染色体の 10 倍以上大きい。この、重複染色体領域の一方における、非コード DNA の選択的な蓄積（または逆に、非コード DNA の欠失）は、比較した領域で多数観察されている。どのような機構がこれらのプロセスを促進するのかを説明する仮説を検証しなければならない。

アライメント結果を評価すると、重複の起こった時期に関するいくつかの予見が得られる。前述したように、大規模な古代のセグメント重複が、実際、今回のゲノム全体の解析で検出されたブロック配列のほとんどを最も良く説明できる。前述したように発展した（第 2 番染色体から第 14 番染色体、2 番から 12 番、18 番から 20 番）大規模重複に関与しているヒト染色体の領域は、それぞれが、異なるマウスの染色体領域と相似である。対応するマウスの染色体領域では、そのヒト染色体のシンテニー相手と比べて、配列の保存程度もそして順序までも、ヒトの重複領域がお互い似ているよりも、もっと似ている。さらに、対応するマウス染色体領域はそれぞれ、ヒトの配列重複の割付確認が行われたヒトの遺伝子に対して直系といえる遺伝子を有意な割合で持っている。これらのことに基づいて、対応するマウス染色体領域について、粗略なレベルで解析を行ったところ、ヒトで観察されたのと同じ大規模の重複の生成物であるようである。マウスのより完全なゲノムがアセンブリされた時点で、より詳細な解析を行う必要があるが、これらの基礎となる大規模重複は、二つの種の分岐に先立って起こったと考えられる。これにより、重複が起きた時期は、最も最近でも、霊長類とげっ歯類系列に分かれる前となる。この時期は、ヒト染色体と、ニワトリや日本のフグ（*Fugu rubripes*）、ゼブラフィッシュの染色体とのシンテニーを調べることにより、さらに精細化できる [(95)]。これらの種にマッピングされたヒト重複の両方のペアに対応する唯一の実質的なシンテニー領域は、Hox クラスター領域に限られている。これらの領域（またはその他）のヒト染色体に対するシンテニーが、さらなるマッピングによって拡張されれば、ヒトにみられるほぼ染色体全長にわたる重複の年代が、脊椎動物が多様化していく根幹の時期に定められるだろう。

MUMmer に基づく結果は、大きいブロック重複のサイズが、数遺伝子から染色体のほとんどをカバーするセグメントまでの範囲であることを示している。セグメント重複の程度は、昔起こった全ゲノム重複事件が、無数の重複領域の原因であるかどうか疑問を起こさせるものである [(96)]。重複後、多くの欠損とその後の組換えが起きている。これらの事象により、全ゲノム重複と、多数回の小さな重複との区別が難しくなった。種間のゲノム比較に一部は依存するが、全ブロック重複の推定年代を比較することに特に焦点をおいた今後の解析が、これらの二つの仮説のうちどちらの可能性が高いかの決定に必要であろう。様々な脊椎動物のゲノムの比較、そして属間にわたるゲノム比較まで行われれば、重複の解読が可能になり、やがては、我々のゲノムについて階層区分された歴史が明らかになり、それに

よってヒトを他の生き物と分けている多くの重要な機能の出現の歴史が明らかになるであろう。

第6章A ゲノム全体の配列バリエーションの検討

要約

一塩基変異多型（SNP）を同定するため、コンピュータを用いた手法により、セレラ社で得た配列を他のSNPデータと比較した。その結果、2つの染色体間に見られるSNPの出現率は、およそ1200から1500塩基につき1個の割合であった。SNPはゲノム全体にわたり、ランダムではない分布を示している。コード領域と予測される部分に影響するSNPを機能面から分析すると、全SNPのうち、ごく少数（1%未満）のみがタンパク質機能に影響を与える可能性を有していた。結果として、ヒトタンパク質の構造的な多様性に寄与する遺伝的バリエーションは数百万というより、わずか数千であり得ると推定される。

完全なゲノム配列が得られたことにより、研究者が遺伝子を発見する速度は劇的に加速されると思われるが、ヒトの健康状態に関する個人差の遺伝的基礎は、DNA配列の差異を解析することによってのみ明らかにされ得る。ゲノム全体に対するショットガンシークエンシングは全ゲノムアセンブリとの組み合わせにより、配列バリエーションの検出に特に有効な方法となる。さらに今回は、以下の3つの異なる手法により同定されたSNPの分布と特性の比較を行った：（ｉ）セレラ社コンセンサス配列のPFPアセンブリに対するアライメント、（ii）高品質で読み取られたゲノム配列のオーバーラップ部分（以下、”Kwok”と呼ぶ; 1,120,195個のSNP）[97]、および（iii）リデュースリプレゼンテーションショットガンシークエンシング（以下、”TSC”と呼ぶ; 632,640個のSNP）[98]。これらのデータは一致して、全体的な塩基多様性として約8×10^{-4}の値を示し、また、ゲノム全体でのSNP密度の著しい不均一性および、発現されるタンパク質に変化を生じさせない非コード性のバリエーションが圧倒的に多数を占めることを示していた。

6.1. セレラのコンセンサス配列とPFPアセンブリのアライメントにより見出されたSNP

SNPを見つけるには、すべての部位に対してシークエンシングを繰り返し行って品質を最大限に高め、明瞭なサンプリングモデルを用いて偽陽性判定および偽陰性判定の割合を定量的に処理することが理想的である[99]。しかし、このような詳細が得られないままコンセンサス配列を比較するためには、より特殊なアプローチが必要であった（PFPアセンブリの品質スコアは取得が困難であった）。まず、2つのコンセンサス配列間の差異をすべて同定し、その後、シークエンシングエラーと誤アセンブリングの寄与を減らすために、フィルタリング（濾過）処理を行った。今回、フィルタリング処理の有効性を測る尺度としては、トランジション置換およびトランスバージョン置換の割合を測定した。これは、哺乳類の進化[100]およびヒトのSNPｓ[101, 102]においては、2：1の比が典型的であると文献に記されているためである。実際のフィルタリング処理では、セレラ社コンセンサス配列中の品質ス

コアが30未満の部分および、バリアント密度が400 bp中5個よりも大きい部分のバリアントを除去した。その結果、トランジションとトランスバージョンの比は1.57：1から1.89：1に変化した。フィルタリングを2.3 Gbpのセレラ社とPFPのコンセンサス配列アライメントに適用すると、総計2,778,474個の置換差から、SNPと推定される差異が2,104,820個同定された。これらのSNPと他の手法により得られたSNPとの重複については、以下で述べる。

6.2. 公共SNPデータベースとの比較

PowerBlastプログラム[103]による配列類似性検索（sequence similarity search）を用いて、dbSNP（www.ncbi.nlm.nih.gov/SNP）に含まれる2,536,021個のSNP、およびHGMD（英国ウェールズ大学Human Gene Mutation Database）に含まれる13,150個のSNPをセレラ社コンセンサス配列上にマップした。dbSNPに含まれている最も大きなデータセットはKwokセットとTSCセットであり、それぞれ、dbSNPレコードの47%および25%を占めている。低いカバー倍数のdbSNP配列しかない品質の低いアライメント、およびセレラ社配列とdbSNP接続配列の間の同一性が98%未満のアライメントは除外した。また、セレラ社ゲノム上で複数の位置にマッピングされるdbSNP配列も除外した。総計2,336,935個のdbSNPバリアントが、セレラ社配列上の1,223,038個の特定の位置にマップされた。これは、dbSNPにはかなりの反復性があることを示唆している。なお、TSCセットでは585,811個、Kwokセットでは438,032個のSNPがゲノム上の特定の位置にマップされた。この解析で使用したユニークなSNPの数は、セレラ-PFP、TSCおよびKwokを含め、総計2,737,668個である。表15は、これらの方法の一つの手法により同定されたSNPは、かなりの部分が別の手法によっても見出されることを示している。Kwokとセレラ-PFPのSNP間に見られる非常に高い重複率（36.2%）は、部分的には、Kwokで使用した配列がPFPアセンブリに取り込まれたためであろう。KwokとTSCセット間の重複が著しく低い（16.4%）のは、セットの大きさがともに小さいためである。なお、セレラ-PFPで得られたSNPの24.5%が、セレラのゲノム配列から得られたSNPと重複している[46]。ヒトの集団サンプルによるSNPの検証は費用のかさむ面倒な過程となるため、複数のデータセットに基づく”in silico”（シリコンチップで、すなわちコンピュータ解析による）確認が効率的な当初検証法となるであろう。

表 15. 全ゲノム SNP データベースに含まれる SNP の重複
表中の数値は、各ペアのデータセットで重複している SNP の数である。括弧内の数値は重複の割合であり、これは重複する SNP の数を、比較した 2 つのデータベースのうち小さい方の SNP 総数で割ったものである。各データベースの SNP 総数は以下の通り。セレラ-PFP：2,104,820 個、TSC：585,811 個、Kwok：438,032 個。TSC および Kwok データセットでは、ユニークな SNP のみを含めた。

	TSC	Kwok
Celera-PFP	188,694	158,532
	(0.322)	(0.362)
TSC		72,024
		(0.164)

これら3セットのSNPがヒトのバリエーションに対して同一の実態を表しているのかを評価する方法の一つとしては、各セットにおける6種類のあり得る塩基変異の頻度を調べることが挙げられる（表16）。塩基のバラツキを測定した従来の結果は、ほとんどが特定の遺伝子に対する小スケールの解析に基づくものであるが [101] 、これら3つのデータセットを用いた我々の解析は、従来の結果をゲノム全体的な規模で検証するものといえる。Kwokセット、TSCセットおよび、我々の全ゲノムショットガン [46] から得られたSNPには、この塩基置換パターンに顕著な均一性が見られる。他のデータセットと比較すると、セレラ-PFP組み合わせでは他のSNPセットで観察される2：1のトランジション：トランスバージョン比から、わずかに逸脱している。しかし、コンピュータによるセレラ PFP間の比較で同定されたSNPsの一部は、実際にはシークエンスエラーである可能性があるため、この結果は予想外のものではない。セレラ-PFPセットに見られる配列差の15%がシークエンスエラー（おそらくはランダムに起こった）によるものであると仮定すると、真のSNPに対するトランジション：トランスバージョン比は2：1となる。

表 16. 各 SNP データセットにおける塩基変化の要約

SNP data set	A/G (%)	C/T (%)	A/C (%)	A/T (%)	C/G (%)	T/G (%)	Transition: transversion
Celera-PFP	30.7	30.7	10.3	8.6	9.2	10.3	1.59:1
Kwok*	33.7	33.8	8.5	7.0	8.6	8.4	2.07:1
TSC†	33.3	33.4	8.8	7.3	8.6	8.6	1.99:1

* November 2000 release of the NCBI database dbSNP (www.nci.nlm.nih.gov/SNP/) with the method defined as Overlap SnpDetectionWithPolyBayes. The submitter of the data is Pui-Yan Kwok from Washington University.

†November 2000 release of NCBI dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the methods defined as TSC-Sanger, TSC-WICGR, and TSC-WUGSC. The submitter of the data is Lincoln Stein from Cold Spring Harbor Laboratory.

6.3. 確認されたSNPに基づく塩基多様性の推定

同定されたSNPの数は染色体間で大きく異なっていた。これらの値を染色体の大きさと配列カバー倍数に対して正規化するため、我々は塩基の多様性の標準的な統計値であるπを使用した[104]。塩基多様度とは、特定部位あたりのヘテロ接合性の尺度であり、母集団から任意に選んだ一対の染色体で特定のヌクレオチド部位が異なる確率を定量化したものである。各染色体に対する塩基多様度を計算するには、塩基変異があるかが調べられた部位の塩基数が既知である必要があり、また、リデュースリプレゼンテーションシークエンシングのような手法においては、シークエンスの品質および各部位のカバー倍数も知っている必要がある。これらのデータは容易に入手できるものではないため、TSCからは、塩基多様度を推定することはできなかった。高品質配列がオーバーラップしている部位からの塩基多様性の推定は可能であるはずだが、この場合も全アライメントの詳細についてより多くの情報が必要である。

ショットガンアセンブリからの塩基多様度の推定には、マルチアライメントの各カラムに対して、2つまたはそれ以上の異なるアリルが存在することの確率、および、実際にアリルが異なる配列を有する場合にSNPを検出する確率（すなわち、正しい配列判定の可能性）を計算することが必要である。カバー倍数が大きいほど、また、配列の品質が高いほど、SNPの検出に成功する確率は高くなる[105]。カバー倍数のバラツキを矯正した後においても、塩基多様度は常染色体間で異なっていた。染色体内の変化度合いを推定するため、100-kbpの配列ウインドウに対するπの推定値を使用した分散分析法により、この不均一性の有意性を試験した（セレラ-PFP比較では、F＝29.73、P＜0.0001）。

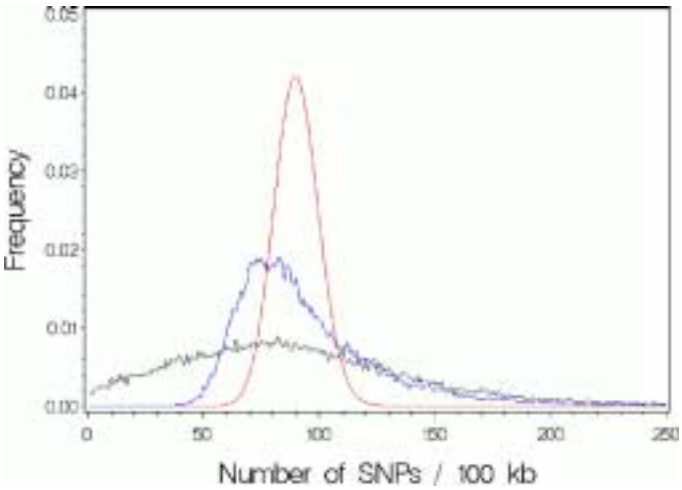
セレラ社配列とPFPの比較から推定される常染色体の平均多様度は8.94×10^{-4}であり、X染色体上の塩基多様度は6.54×10^{-4}であった。X染色体は常染色体に比べ変異性が低いと予想されるが、これは、母集団中の常染色体4コピーごとに、X染色体は3つしかなく、有効母集団サイズが小さいために、ランダムドリフトによって、より速い速度でX染色体から変異が排除されることを意味するためである[106]。

ゲノム全体にわたる塩基の変異を確認した結果からいって、遺伝子サンプリングに基づいて行われたヒトの塩基多様度の従来の推定値は適度に正確であったと思われる[101, 102, 106, 107]。ゲノム全体では、我々が推定したセレラ-PFPアライメントに対する塩基多様度の値は8.98×10^{-4}であったのに対し、10個の高密度に再シークエンシングされたヒト遺伝子を平均して発表された推定値は8.00×10^{-4}であった[108]。

6.4. ヒトゲノム中における塩基多様度の変動

このように、染色体間でのSNP密度の変化度合いが明らかに高いことから、染色体内でのより微細なスケールでも不均一性が存在するか否か、また、この不均一性が偶然から予想されるより大きいか否かという疑問が生じる。もし、SNPが無作為かつ独立な変異により生じるなら、任意の一定の大きさのフラグメントに含まれるSNPの数はポアソン分布に従うはずであると思われる。しかし、100-kbpのフラグメントに含まれるSNPの分布に対して観察される分散は、ポアソン分布から予測されるよりもはるかに大きなものであった（図14）。しかし、この極端に単純化されたモデルでは、ゲノムの各領域における異なる組換え率および集団の履歴は無視されている。集団遺伝学の理論によると、このバリエーションは中立合着（neutral coalescent）と呼ばれる数式によって説明され得る[109]。我々は、組み換えを含む中立合着をシミュレートするために、よく試験されたアルゴリズムを適用して[110]、有効集団サイズとして10,000、1塩基あたりの組換え率は変異率に等しいとし[111]、このモデルによるSNP数の分布も生成した[112]。しかし、観察されたSNPの分布はポアソンモデルまたは合着モデルのいずれよりも、はるかに大きな変化度を有しており、この差異は極めて有意なものであった。これは、ゲノム中のSNP密度には有意な変化があることを示唆しており、この観察には説明が必要とされる。

図 14．100-kbp ごとの SNP 密度をセレラ-PFP の SNP で決めた結果
グラフの色分けは以下の通り。黒：セレラ-PFP における SNP 密度、青：合着モデル、赤：ポアソン分布。この図は、ゲノム上の SNP 分布はランダムではなく、特定領域の履歴の合着モデルにより完全には説明できないことを示している。



DNA配列が持つある種の特性は局所的なSNP密度に影響し得る。これには、DNAポリメラーゼがエラーを生じる頻度やミスマッチ修復の効率などが含まれる。SNP密度に関係する可能性の高い重要な因子の一つとしてはG+C含量がある。この理由の一部は、CpGジヌクレオチド中のメチル化されたシトシンは脱アミノ化を受けてチミンを形成する傾向があるためであり、このためにCpGの変異速度は他のジヌクレオチドに比べ、約10倍ほど高くなってい

る。我々はゲノム全体にわたって100 kbp配列ウィンドウ中のGC含量と塩基多様度を計算し、それらの間の相関性が陽性（r＝0.21）かつ極めて有意（P＜0.0001）であることを見出したが、G+C含量により説明されるバリエーションは全体のごく一部に過ぎなかった。

6.5. ゲノムクラスごとのSNP

機能別クラスごとのSNP密度の均一性を試験するため、我々はNCBI RefSeqデータベースから得た10,239個の既知遺伝子とセレラ社Otto遺伝子注釈から予測されるすべてのヒト遺伝子を対象とし、配列を遺伝子間領域（予測される転写単位から＞5kbと定義）、5’-UTR、エクソン領域（ミスセンスおよびサイレント）、イントロン領域、3’-UTRに分類した。コーディング領域については、SNPをサイレント（アミノ酸配列を変化させないもの）とミスセンス（タンパク質産物を変更するもの）に分類した。セレラ-PFP、TSCおよびKwokセットのコーディング領域におけるミスセンスSNPとサイレントSNPの比（それぞれ1.12、0.91および0.78）は中立な期待値に比べ、ミスセンス変異の頻度が著しく低くなっており、これは、自然淘汰による有害アミノ酸変化部分の排除とよく合っている[112]。これらの比は、Cargil *et al.*[101]およびHalushka *et al.*[102]により報告されている0.88および1.17のミスセンス：サイレント比とほぼ同じといえる。同様な結果は、セレラ社のショットガン配列から得られたSNPにおいても観察された[46]。

タンパク質の機能を障害する恐れのある変更をもたらすSNPの割合がいかに小さいかは、印象的といえる。10,239個のRefSeq遺伝子中のミスセンスSNPはそれぞれ、セレラ-PFP、TSCおよびKwok SNPにおけるSNP総数の約0.12、0.14および0.17%に過ぎなかった。配列非保存的なタンパク質の変化は、ミスセンスSNPのさらに一部に過ぎない（セレラ-PFP、KwokおよびTSCにおいて47、41および40%）。遺伝子間領域は実質的にほとんど研究されていない領域であるが[113]、同定したSNPの75%が遺伝子間領域にあることは注目に値する（表17）。SNP出現率はイントロン内が最も高く、エクソン内が最も低かった。SNP出現率はイントロン内よりも遺伝子間領域の方が低く、これは、これら2つのDNAクラスの速やかな識別材料の一つとなり得る。これらのSNP出現率はセレラ社のSNPにおいても確認され、そこでも、エキソン内の方がイントロン内よりも低く、また、遺伝子外領域の方がイントロン内よりも低かった[46]。これらの遺伝子間領域SNPの多くは、連鎖および相関解析用のマーカーとして貴重な情報を提供すると思われる。また、そのいくつかは制御機能を担っている可能性も高い。

表 17．分類した各ゲノム領域における SNP の分布

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	

第 7 章 ヒトゲノムにおける予測蛋白コード遺伝子の概観

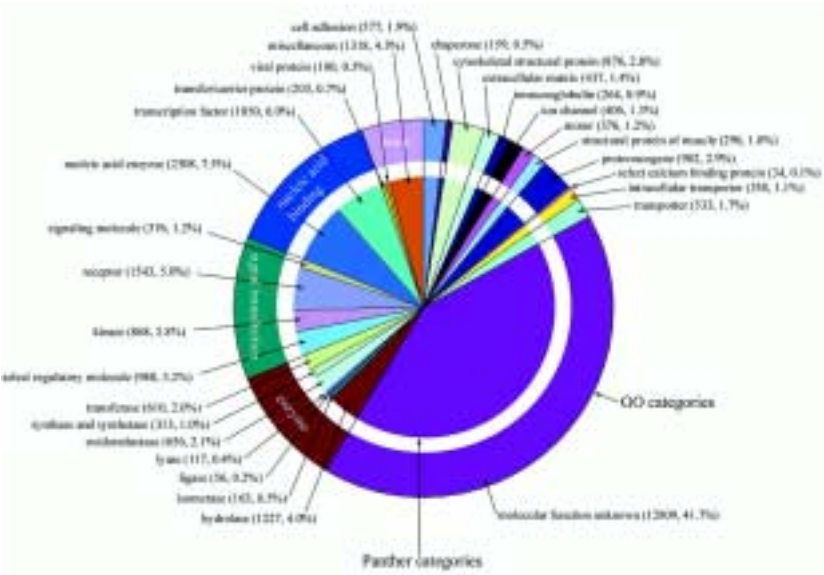
要旨

この章では、完全配列決定されている他の真核生物ゲノムとヒトゲノムを比較した際に、顕著な相違点、類似点を分類整理する目的で、予測蛋白セットの初期コンピュータ解析結果を示す。既知蛋白ファミリーへの分類割当同定法を用いたところ、40%以上のヒト予測蛋白では、分子機能の説明が不可能である。蛋白ドメイン構造に基づく解析から、ハエゲノムおよび線虫ゲノムと比較した際、ヒトゲノムにおける顕著な相違点の詳細カタログができた。特に顕著なものは、神経細胞機能、止血機構、後天的免疫応答、細胞骨格構成等の、発生制御と細胞プロセシングに関与する蛋白でのドメイン領域拡大である。最終的な蛋白ファミリー数および詳細蛋白構造については、今後の実験的研究および包括的な手作業による整理を待たねばならない。

蛋白をコードすると予測されるヒト遺伝子の予備的解析を行った。2 つ手法を用いて、26,588 個の予測蛋白（上述した証拠を最低 2 種類もつ 26,383 個の予測遺伝子に相当）の分子機能を解析、分類した。第一の方法は、国際 Pfam データベース [114,115] およびセレラ社 Panther Classification（CPC）の両方を用いる蛋白ファミリー・レベルでの解析に基づく（図 15）[116]。第二の方法は、Pfam および SMART データベースの両方を用いる蛋白ドメインレベルでの解析に基づく [115, 117]。

図 15 26,383 個のヒト遺伝子の分子機能分布

26,383 個のヒト遺伝子の分子機能分布。各スライスには、特定の分子機能カテゴリに割当て同定したヒト遺伝子の機能名を示し、その数と割合は括弧内に示す。外円は Gene Ontology(GO)[179]の分子機能カテゴリの割当同定分、内円はセレラ社の Panther 分子機能カテゴリ [116] 割当同定分である。



今回の結果は予備的なもので、いくつかの制限がある。熟練した生物学者が Panther、Pfam、SMART における統計モデルを構築、注釈、評価を行ってはいるが、遺伝子存在予測および機能割当同定はコンピュータツールを用いて作製されたものである。コンピュータ予測遺伝子セットでは、偽陽性予測（一部は非活性偽遺伝子）および偽陰性予測（一部はコンピュータ予測不能ヒト遺伝子）があることを予期した。エクソンと遺伝子の境界を明確にする際の誤差もあることを予期した。同様に、自動的な方法による蛋白機能割当同定においても偽陽性予測、偽陰性予測の両方を予期した。機能割当同定プロトコールは、いくつかの生物種にわたって見られる蛋白ファミリーおよび既知ヒト遺伝子ファミリーに焦点を置いている。従って、機能が知られていても、大きな蛋白ファミリーに属さない多数の遺伝子に機能割当同定を行わない場合がある。明記のない限り、Panther、Pfam、SMART におけるモデルのために定義された統計的カットオフ・スコアを用いて、機能を割当同定した 26,588 個の予測蛋白セットから、任意ファミリーもしくは任意機能カテゴリーにある遺伝子の全数計測を行った。

今回のヒト予測蛋白セットの初期検査では、概略的な 3 項目を問題にした。すなわち、(i)予測遺伝子産物のあり得る分子機能は何か。そして、現行の分類方法を用いた際にこれら蛋白はどのように分類されるか。(ⅱ)動物種間で共通すると思われる中核機能は何か。(ⅲ)配列が決まっている他の真核生物の全蛋白とヒト全蛋白はいかに異なるか。

7.1. ヒト予測蛋白の分子機能

図 15 に最低 2 個以上の存在を示す証拠をもつ 26,588 個のヒト予測蛋白について、推定分子機能の概観を示す。遺伝子産物の約 41%（12,809 個）が初期解析で分類不能で、機能不明蛋白と名付けた。我々の自動分類法は比較的大きな蛋白ファミリーのみ処理するため、実際には既知機能もしくは予測機能をもつ“非分類”配列が多数存在する。自動機能予測された蛋白の 60%について、概略的クラスに特異蛋白機能を納めた。ここでは、できるだけ多数の蛋白を分類するため、高次の細胞プロセシングよりも、分子機能に焦点をおいた。これらの機能予測は、既知の機能配列に対する類似性に基づくものである。

さらに 12,731 個の（存在を示す証拠が一つしかない）低信頼度予測遺伝子を解析したところ、これらの付加的推定遺伝子の 636 個（5%）のみが自動法を用いた場合に遺伝子機能があると割当同定された。636 個の予測遺伝子の 3 分の 1 が内在性レトロウイルス蛋白であった。このことは、これらの未知機能遺伝子の大多数が実遺伝子ではないことを示唆している。これらの付加的遺伝子 12,095 個ののほとんどが今日までに配列決定されたゲノムにおいてユニークな配列であることから、単純に大部分が偽陽性遺伝子という予測結果なのであろう。

最も一般的な分子機能は転写因子および核酸代謝に関与する蛋白（核酸関連酵素）である。ヒトゲノムにおいて高い割合で存在する他の分子機能は、受容体、燐酸化酵素、加水分解酵素である。驚くにはあたらないが、加水分解酵素のほとんどが蛋白分解酵素である。前癌遺伝子ファミリーのメンバーであるのみならず、“ 選択的調節分子 ” ファミリーのメンバーである蛋白も多数存在した。すなわち、(i)ヘテロ三量体 GTP 結合蛋白（G 蛋白）、細胞周期調節因子等のシグナルトランスダクションの特定ステップに関与する蛋白、(ⅱ)燐酸化酵素、G 蛋白、脱燐酸化酵素の活性を調節する蛋白である。

7.2. 進化的に保存されたコア・プロセス

様々な “ モデル生物 ” ゲノム配列プロジェクトがすでに完了しているので、ヒトゲノム進化解析を開始するために妥当な比較情報が入手可能である。S.cerevisiae (“ パン酵母 ”)[118] および、二種類の異なる無脊椎動物 C.elegans（線虫）[119]、D.melanogaster（ハエ）[26]のみならず、最近完了した初の植物ゲノム A.thaliana[92]は、ゲノム間比較に多様な基盤を提供するものである。

動物種を越えて一般則だと思われるコア機能は何か、という問題を解明するため、ヒト・ハエ間およびヒト・線虫間で保存されている “ 厳格オルソログ ” を列挙してみた（図 16）。2 つの遺伝子がオルソログ（“ 進化上保存されている蛋白セット ”）である場合、遺伝的にたどることにより 2 つの生物の共通祖先に遡ることが可能であり、従って異なる生物においても同様の保存された機能を担うものと思われるため、オルソロジーの概念は重要である。この解析において、オルソログ（共通祖先からの遺伝によって 2 つの生物に存在する遺伝子）をパラログ（重複という事件により、ある生物に 2 つ以上コピーが存在する遺伝子）から分離することは重要である。なぜなら、パラログはやがて機能的に分岐していくと思われるからである。文献[120]における酵母－線虫オルソログ比較に従い、それぞれの比較ペア（ヒト－ハエおよびヒト－線虫）において 2 つの異なる事例を同定した。第一の事例は、各生物とも 1 つの遺伝子をもっていて、どちらの生物においても他の近縁ホモログが存在しない遺伝子のペアとなるものであった。これら遺伝子は、オルソログとパラログの区別を複雑にする追加的メンバーが他にいないため、簡単にオルソログと同定された。第二の事例は、比較した生物の片方、もしくは両方に 2 つ以上のファミリーメンバーをもつ遺伝子ファミリーである。Chervitz ら[120]は、2 つの生物における全配列間の関係を示す系統樹を解析することでこの事例を処理し、系統樹中の最近傍にある遺伝子ペアを探した。もし最近傍遺伝子ペアが異なる生物種由来ならば、それらはオルソログと推定された。我々はこれらの最近傍遺伝子ペアが、系統樹を検索しなくても、ペア間の配列比較によって、自信をもって同定できることを指摘したい（図 16 の脚注参照）。もし最近傍遺伝子ペアが異な

る生物由来でないとしたら、種としての進化（および／または一方の生物による遺伝子欠失）後に一方もしくは両方の生物でパラログ的な増大があったはずである。この一対一対応関係がないならば、オルソログの定義は不明瞭になってしまう。ヒト予測蛋白セットの初期コンピュータ概観では、各予測蛋白に対してこの問題に回答することはできなかった。そこで、“ 厳格オルソログ ”、すなわち、不明瞭さがない一対一対応関係があるもののみを考察する（図 16）と、この基準に従えば、ヒト ハエ間では 2,758 個、ヒト 線虫間では 2,031 個の厳格オルソログがある（このうち 1,523 個は共通している）。我々は、*D. melanogaster* と *C.elegans* でも厳格オロソログがあるこれら 1,523 個のヒト蛋白を進化的に保存されたセットとして定義することにする。

図 16 脊椎動物および無脊椎動物にわたる推定オルソログ機能

脊椎動物および無脊椎動物にわたる推定オルソログ機能。各スライスには、特定の分子機能カテゴリに関与しているヒト、ハエ、線虫のゲノムの”厳格オルソログ”機能名を示し、その数と割合を括弧内に示す。”厳格オルソログ”は、ここでは、各オルソログ対が(i) 10^{-10} 以下の BLAST P 値を有している、(ii) どの生体でもどのようなパラログより有意な BLAST P 値を有している、など双方向性 BLAST にベストヒット [180] した場合、すなわち、オルソロジーをあいまいにしかねない種分化につづく重複が見られないような場合に限定した。この尺度は全く厳格であり、そのためオルソログ数は少なくなっている。この基準により、ヒトとハエの間のオルソログ数は 2758 個、ヒトと線虫のオルソログ数は 2031 個ある（ヒトとハエ、ヒトと線虫で共通しているのは 1523 個）。

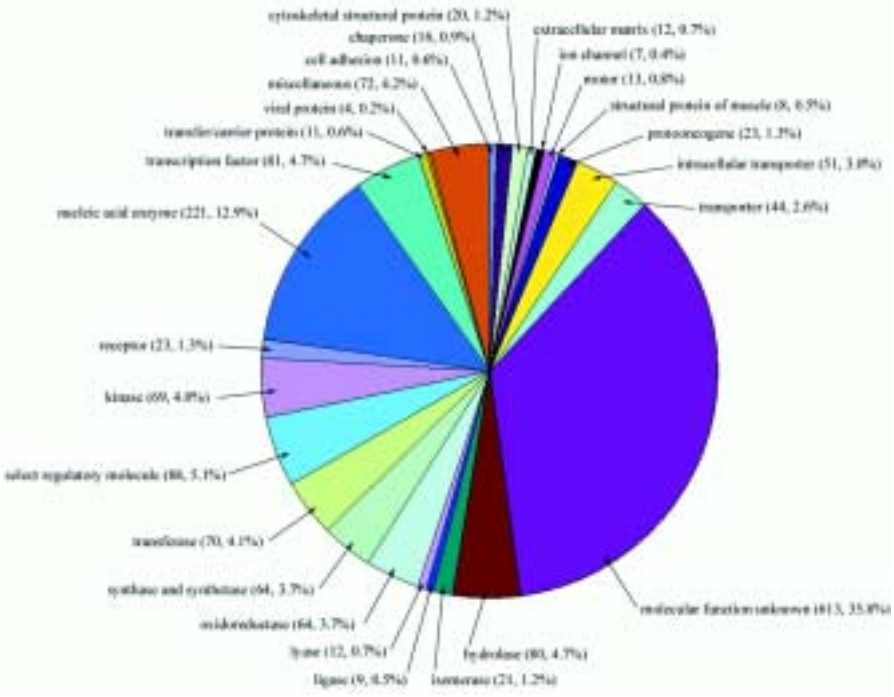


表 18 ヒト *H. sapiens*（H）、ショウジョウバエ *D.melanogaster*（D）、線虫 *C.elegans*（C）、酵母 *S.cervisiae*（Y）、アラビドプシス *A.thaliana*（A）蛋白におけるドメインに基づく比較解析　ヒト *H. sapiens*（H）、ショウジョウバエ *D.melanogaster*（D）、線虫 *C.elegans*（C）、酵母 *S.cervisiae*（Y）、アラビドプシス *A.thaliana*（A）蛋白におけるドメインに基づく比較解析。これら真核生物のそれぞれの予測蛋白セットは、E 値 0.001 をカットオフ値とする Pfam バージョン 5.5 で解析した。特定した Pfam ドメインを含む蛋白数を、ドメイン総数（括弧内）と共に各欄に示す。発表のために、細胞過程別にドメインを分類した。いくつかのドメイン（SH2）は、ひとつ以上の細胞過程に含まれている。大規模な自動的分類にはさまざまな制約があるため、Pfam 解析の結果は、蛋白ファミリーのヒト curation に基づいて得た結果とは異なっている場合がある。今回の解析に用いた E 値の厳しいカットオフ値のために数が少なくなったドメインの代表例には、2 つの星印（＊＊）をつけている。代表例は、短い分岐ドメイン、αヘリックスが優勢なドメイン、システインに富むジンクフィンガー蛋白の数種などである。

Accession number	Domain name	Domain description	H	F	W	Y	A
		Developmental and homeostatic regulators					
PF02039	Adrenomedullin	Adrenomedullin	1	0	0	0	0
PF00212	ANP	Atrial natriuretic peptide	2	0	0	0	0
PF00028	Cadherin	Cadherin domain	100 (550)	14 (157)	16 (66)	0	0
PF00214	Calc_CGRP_IAPP	Calcitonin/CGRP/IAPP family	3	0	0	0	0
PF01110	CNTF	Ciliary neurotrophic factor	1	0	0	0	0
PF01093	Clusterin	Clusterin	3	0	0	0	0
PF00029	Connexin	Connexin	14 (16)	0	0	0	0
PF00976	ACTH_domain	Corticotropin ACTH domain	1	0	0	0	0
PF00473	CRF	Corticotropin-releasing factor family	2	1	0	0	0
PF00007	Cys_knot	Cystine-knot domain	10 (11)	2	0	0	0
PF00778	DIX	Dix domain	5	2	4	0	0
PF00322	Endothelin	Endothelin family	3	0	0	0	0
PF00812	Ephrin	Ephrin	7 (8)	2	4	0	0
PF01404	EPh_Ibd	Ephrin receptor ligand binding domain	12	2	1	0	0
PF00167	FGF	Fibroblast growth factor	23	1	1	0	0
PF01534	Frizzled	Frizzled/Smoothened family membrane region	9	7	3	0	0
PF00236	Hormone6	Glycoprotein hormones	1	0	0	0	0
PF01153	Glypican	Glypican	14	2	1	0	0
PF01271	Granin	Grainin (chromogranin or secretogranin)	3	0	0	0	0
PF02058	Guanylin	Guanylin precursor	1	0	0	0	0
PF00049	Insulin	Insulin/IGF/Relaxin family	7	4	0	0	0
PF00219	IGFBP	Insulin-like growth factor binding proteins	10	0	0	0	0
PF02024	Leptin	Leptin	1	0	0	0	0
PF00193	Xlink	LINK (hyaluron binding)	13 (23)	0	1	0	0
PF00243	NGF	Nerve growth factor family	3	0	0	0	0
PF02158	Neuregulin	Neuregulin family	4	0	0	0	0
PF00184	Hormone5	Neurohypophysial hormones	1	0	0	0	0
PF02070	NMU	Neuromedin U	1	0	0	0	0
PF00066	Notch	Notch (DSL) domain	3 (5)	2 (4)	2 (6)	0	0

PF00865	Osteopontin	Osteopontin	1	0	0	0	0
PF00159	Hormone3	Pancreatic hormone peptides	3	0	0	0	0
PF01279	Parathyroid	Parathyroid hormone family	2	0	0	0	0
PF00123	Hormone2	Peptide hormone	5 (9)	0	0	0	0
PF00341	PDGF	Platelet-derived growth factor (PDGF)	5	1	0	0	0
PF01403	Sema	Sema domain	27 (29)	8 (10)	3 (4)	0	0
PF01033	Somatomedin_B	Somatomedin B domain	5 (8)	3	0	0	0
PF00103	Hormone	Somatotropin	1	0	0	0	0
PF02208	Sorb	Sorbin homologous domain	2	0	0	0	0
PF02404	SCF	Stem cell factor	2	0	0	0	0
PF01034	Syndecan	Syndecan domain	3	1	1	0	0
PF00020	TNFR_c6	TNFR/NGFR cysteine-rich region	17 (31)	1	0	0	0
PF00019	TGF-β	Transforming growth factor β-like domain	27 (28)	6	4	0	0
PF01099	Uteroglobin	Uteroglobin family	3	0	0	0	0
PF01160	Opiods_neuropep	Vertebrate endogenous opioids neuropeptide	3	0	0	0	0
PF00110	Wnt	Wnt family of developmental signaling proteins	18	7 (10)	5	0	0
		Hemostasis					
PF01821	ANATO	Anaphylotoxin-like domain	6 (14)	0	0	0	0
PF00386	C1q	C1q domain	24	0	0	0	0
PF00200	Disintegrin	Disintegrin	18	2	3	0	0
PF00754	F5_F8_type_C	F5/8 type C domain	15 (20)	5 (6)	2	0	0
PF01410	COLFI	Fibrillar collagen C-terminal domain	10	0	0	0	0
PF00039	Fn1	Fibronectin type I domain	5 (18)	0	0	0	0
PF00040	Fn2	Fibronectin type II domain	11 (16)	0	0	0	0
PF00051	Kringle	Kringle domain	15 (24)	2	2	0	0
PF01823	MACPF	MAC/Perforin domain	6	0	0	0	0
PF00354	Pentaxin	Pentaxin family	9	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF02210	TSPN	Thrombospondin N-terminal-like domains	14	1	0	0	0
PF01108	Tissue_fac	Tissue factor	1	0	0	0	0
PF00868	Transglutamin_N	Transglutaminase family	6	1	0	0	0
PF00927	Transglutamin_C	Transglutaminase family	8	1	0	0	0
PF00594	Gla	Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain	11	0	0	0	0
		Immune response					
PF00711	Defensin_beta	Beta defensin	1	0	0	0	0
PF00748	Calpain_inhib	Calpain inhibitor repeat	3 (9)	0	0	0	0
PF00666	Cathelicidins	Cathelicidins	2	0	0	0	0

PF00129	MHC_I	Class I histocompatibility antigen, domains alpha 1 and 2	18 (20)	0	0	0	0
PF00993	MHC_II_alpha**	Class II histocompatibility antigen, alpha domain	5 (6)	0	0	0	0
PF00969	MHC_II_beta**	Class II histocompatibility antigen, beta domain	7	0	0	0	0
PF00879	Defensin_propep	Defensin propeptide	3	0	0	0	0
PF01109	GM_CSF	Granulocyte-macrophage colony-stimulating factor	1	0	0	0	0
PF00047	Ig	Immunoglobulin domain	381 (930)	125 (291)	67 (323)	0	0
PF00143	Interferon	Interferon alpha/beta domain	7 (9)	0	0	0	0
PF00714	IFN-gamma	Interferon gamma	1	0	0	0	0
PF00726	IL10	Interleukin-10	1	0	0	0	0
PF02372	IL15	Interleukin-15	1	0	0	0	0
PF00715	IL2	Interleukin-2	1	0	0	0	0
PF00727	IL4	Interleukin-4	1	0	0	0	0
PF02025	IL5	Interleukin-5	1	0	0	0	0
PF01415	IL7	Interleukin-7/9 family	1	0	0	0	0
PF00340	IL1	Interleukin-1	7	0	0	0	0
PF02394	IL1_propep	Interleukin-1 propeptide	1	0	0	0	0
PF02059	IL3	Interleukin-3	1	0	0	0	0
PF00489	IL6	Interleukin-6/G-CSF/MGF family	2	0	0	0	0
PF01291	LIF_OSM	Leukemia inhibitory factor (LIF)/oncostatin (OSM) family	2	0	0	0	0
PF00323	Defensins	Mammalian defensin	2	0	0	0	0
PF01091	PTN_MK	PTN/MK heparin-binding protein	2	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00048	IL8	Small cytokines (intecrine/chemokine), interleukin-8 like	32	0	0	0	0
PF01582	TIR	TIR domain	18	8	2	0	131 (143)
PF00229	TNF	TNF (tumor necrosis factor) family	12	0	0	0	0
PF00088	Trefoil	Trefoil (P-type) domain	5 (6)	0	2	0	0
		PI-PY-rho GTPase signaling					
PF00779	BTK	BTK motif	5	1	0	0	0
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00609	DAGKa	Diacylglycerol kinase accessory domain (presumed)	9	4	7	0	6
PF00781	DAGKc	Diacylglycerol kinase catalytic domain (presumed)	10	8	8	2	11 (12)
PF00610	DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin (DEP)	12 (13)	4	10	5	2
PF01363	FYVE	FYVE zinc finger	28 (30)	14	15	5	15
PF00996	GDI	GDP dissociation inhibitor	6	2	1	1	3
PF00503	G-alpha	G-protein alpha subunit	27 (30)	10	20 (23)	2	5
PF00631	G-gamma	G-protein gamma like domains	16	5	5	1	0
PF00616	RasGAP	GTPase-activator protein for	11	5	8	3	0

		Ras-like GTPase					
PF00618	RasGEFN	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	9	2	3	5	0
PF00625	Guanylate_kin	Guanylate kinase	12	8	7	1	4
PF02189	ITAM	Immunoreceptor tyrosine-based activation motif	3	0	0	0	0
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF00130	DAG_PE-bind	Phorbol esters/diacylglycerol binding domain (C1 domain)	45 (56)	25 (31)	26 (40)	1 (2)	4
PF00388	PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	12	3	7	1	8
PF00387	PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	11	2	7	1	8
PF00640	PID	Phosphotyrosine interaction domain (PTB/PID)	24 (27)	13	11 (12)	0	0
PF02192	PI3K_p85B	PI3-kinase family, p85-binding domain	2	1	1	0	0
PF00794	PI3K_rbd	PI3-kinase family, ras-binding domain	6	3	1	0	0
PF01412	ArfGAP	Putative GTP-ase activating protein for Arf	16	9	8	6	15
PF02196	RBD	Raf-like Ras-binding domain	6 (7)	4	1	0	0
PF02145	Rap_GAP	Rap/ran-GAP	5	4	2	0	0
PF00788	RA	Ras association (RalGDS/AF-6) domain	18 (19)	7 (9)	6	1	0
PF00071	Ras	Ras family	126	56 (57)	51	23	78
PF00617	RasGEF	RasGEF domain	21	8	7	5	0
PF00615	RGS	Regulator of G protein signaling domain	27	6 (7)	12 (13)	1	0
PF02197	RIIa	Regulatory subunit of type II PKA R-subunit	4	1	2	1	0
PF00620	RhoGAP	RhoGAP domain	59	19	20	9	8
PF00621	RhoGEF	RhoGEF domain	46	23 (24)	18 (19)	3	0
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01017	STAT	STAT protein	7	1	1 (2)	0	0
PF00790	VHS	VHS domain	4	2	4	4	8
PF00568	WH1	WH1 domain	7	2	2 (3)	1	0
		Domains involved in apoptosis					
PF00452	Bcl-2	Bcl-2	9	2	1	0	0
PF02180	BH4	Bcl-2 homology region 4	3	0	1	0	0
PF00619	CARD	Caspase recruitment domain	16	0	2	0	0
PF00531	Death	Death domain	16	5	7	0	0
PF01335	DED	Death effector domain	4 (5)	0	0	0	0

PF02179	BAG	Domain present in Hsp70 regulators	5 (8)	3	2	1	5
PF00656	ICE_p20	ICE-like protease (caspase) p20 domain	11	7	3	0	0
PF00653	BIR	Inhibitor of Apoptosis domain	8 (14)	5 (9)	2 (3)	1 (2)	0
		Cytoskeletal					
PF00022	Actin	Actin	61 (64)	15 (16)	12	9 (11)	24
PF00191	Annexin	Annexin	16 (55)	4 (16)	4 (11)	0	6 (16)
PF00402	Calponin	Calponin family	13 (22)	3	7 (19)	0	0
PF00373	Band_41	FERM domain (Band 4.1 family)	29 (30)	17 (19)	11 (14)	0	0
PF00880	Nebulin_repeat	Nebulin repeat	4 (148)	1 (2)	1	0	0
PF00681	Plectin_repeat	Plectin repeat	2 (11)	0	0	0	0
PF00435	Spectrin	Spectrin repeat	31 (195)	13 (171)	10 (93)	0	0
PF00418	Tubulin-binding	Tau and MAP proteins, tubulin-binding	4 (12)	1 (4)	2 (8)	0	0
PF00992	Troponin	Troponin	4	6	8	0	0
PF02209	VHP	Villin headpiece domain	5	2	2	0	5
PF01044	Vinculin	Vinculin family	4	2	1	0	0
		ECM adhesion					
PF01391	Collagen	Collagen triple helix repeat (20 copies)	65 (279)	10 (46)	174 (384)	0	0
PF01413	C4	C-terminal tandem repeated domain in type 4 procollagen	6 (11)	2 (4)	3 (6)	0	0
PF00431	CUB	CUB domain	47 (69)	9 (47)	43 (67)	0	0
PF00008	EGF	EGF-like domain	108 (420)	45 (186)	54 (157)	0	1
PF00147	Fibrinogen_C	Fibrinogen beta and gamma chains, C-terminal globular domain	26	10 (11)	6	0	0
PF00041	Fn3	Fibronectin type III domain	106 (545)	42 (168)	34 (156)	0	1
PF00757	Furin-like	Furin-like cysteine rich region	5	2	1	0	0
PF00357	Integrin_A	Integrin alpha cytoplasmic region	3	1	2	0	0
PF00362	Integrin_B	Integrins, beta chain	8	2	2	0	0
PF00052	Laminin_B	Laminin B (Domain IV)	8 (12)	4 (7)	6 (10)	0	0
PF00053	Laminin_EGF	Laminin EGF-like (Domains III and V)	24 (126)	9 (62)	11 (65)	0	0
PF00054	Laminin_G	Laminin G domain	30 (57)	18 (42)	14 (26)	0	0
PF00055	Laminin_Nterm	Laminin N-terminal (Domain VI)	10	6	4	0	0
PF00059	Lectin_c	Lectin C-type domain	47 (76)	23 (24)	91 (132)	0	0
PF01463	LRRCT	Leucine rich repeat C-terminal domain	69 (81)	23 (30)	7 (9)	0	0
PF01462	LRRNT	Leucine rich repeat N-terminal domain	40 (44)	7 (13)	3 (6)	0	0
PF00057	Ldl_recept_a	Low-density lipoprotein receptor domain class A	35 (127)	33 (152)	27 (113)	0	0
PF00058	Ldl_recept_b	Low-density lipoprotein receptor repeat class B	15 (96)	9 (56)	7 (22)	0	0
PF00530	SRCR	Scavenger receptor cysteine-rich domain	11 (46)	4 (8)	1 (2)	0	0

PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF00090	Tsp_1	Thrombospondin type 1 domain	41 (66)	11 (23)	18 (47)	0	0
PF00092	Vwa	von Willebrand factor type A domain	34 (58)	0	17 (19)	0	1
PF00093	Vwc	von Willebrand factor type C domain	19 (28)	6 (11)	2 (5)	0	0
PF00094	Vwd	von Willebrand factor type D domain	15 (35)	3 (7)	9	0	0
		Protein interaction domains					
PF00244	14-3-3	14-3-3 proteins	20	3	3	2	15
PF00023	Ank	Ank repeat	145 (404)	72 (269)	75 (223)	12 (20)	66 (111)
PF00514	Armadillo_seg	Armadillo/beta-catenin-like repeats	22 (56)	11 (38)	3 (11)	2 (10)	25 (67)
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00027	cNMP_binding	Cyclic nucleotide-binding domain	26 (31)	21 (33)	15 (20)	2 (3)	22
PF01556	DnaJ_C	DnaJ C terminal region	12	9	5	3	19
PF00226	DnaJ	DnaJ domain	44	34	33	20	93
PF00036	Efhand**	EF hand	83 (151)	64 (117)	41 (86)	4 (11)	120 (328)
PF00611	FCH	Fes/CIP4 homology domain	9	3	2	4	0
PF01846	FF	FF domain	4 (11)	4 (10)	3 (16)	2 (5)	4 (8)
PF00498	FHA	FHA domain	13	15	7	13 (14)	17
PF00254	FKBP	FKBP-type peptidyl-prolyl cis-trans isomerases	15 (20)	7 (8)	7 (13)	4	24 (29)
PF01590	GAF	GAF domain	7 (8)	2 (4)	1	0	10
PF01344	Kelch	Kelch motif	54 (157)	12 (48)	13 (41)	3	102 (178)
PF00560	LRR**	Leucine Rich Repeat	25 (30)	24 (30)	7 (11)	1	15 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00989	PAS	PAS domain	18 (19)	9 (10)	6	1	13 (18)
PF00595	PDZ	PDZ domain (Also known as DHR or GLGF)	96 (154)	60 (87)	46 (66)	2	5
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF01535	PPR**	PPR repeat	5	3 (4)	0	1	474 (2485)
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01740	STAS	STAS domain	5	1	6	2	13
PF00515	TPR**	TPR domain	72 (131)	39 (101)	28 (54)	16 (31)	65 (124)
PF00400	WD40**	WD40 domain	136 (305)	98 (226)	72 (153)	56 (121)	167 (344)
PF00397	WW	WW domain	32 (53)	24 (39)	16 (24)	5 (8)	11 (15)
PF00569	ZZ	ZZ-Zinc finger present in dystrophin, CBP/p300	10 (11)	13	10	2	10
		Nuclear interaction domains					
PF01754	Zf-A20	A20-like zinc finger	2 (8)	2	2	0	8
PF01388	ARID	ARID DNA binding domain	11	6	4	2	7
PF01426	BAH	BAH domain	8 (10)	7 (8)	4 (5)	5	21 (25)

PF00643	Zf-B_box**	B-box zinc finger	32 (35)	1	2	0	0
PF00533	BRCT	BRCA1 C Terminus (BRCT) domain	17 (28)	10 (18)	23 (35)	10 (16)	12 (16)
PF00439	Bromodomain	Bromodomain	37 (48)	16 (22)	18 (26)	10 (15)	28
PF00651	BTB	BTB/POZ domain	97 (98)	62 (64)	86 (91)	1 (2)	30 (31)
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	3 (4)	1	0	0	13 (15)
PF00385	Chromo	chromo' (CHRomatin Organization MOdifier) domain	24 (27)	14 (15)	17 (18)	1 (2)	12
PF00125	Histone	Core histone H2A/H2B/H3/H4	75 (81)	5	71 (73)	8	48
PF00134	Cyclin	Cyclin	19	10	10	11	35
PF00270	DEAD	DEAD/DEAH box helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
PF01529	Zf-DHHC	DHHC zinc finger domain	15	20	16	7	22
PF00646	F-box**	F-box domain	16	15	309 (324)	9	165 (167)
PF00250	Fork_head	Fork head domain	35 (36)	20 (21)	15	4	0
PF00320	GATA	GATA zinc finger	11 (17)	5(6)	8 (10)	9	26
PF01585	G-patch	G-patch domain	18	16	13	4	14 (15)
PF00010	HLH**	Helix-loop-helix DNA-binding domain	60 (61)	44	24	4	39
PF00850	Hist_deacetyl	Histone deacetylase family	12	5 (6)	8 (10)	5	10
PF00046	Homeobox	Homeobox domain	160 (178)	100 (103)	82 (84)	6	66
PF01833	TIG	IPT/TIG domain	29 (53)	11 (13)	5 (7)	2	1
PF02373	JmjC	JmjC domain	10	4	6	4	7
PF02375	JmjN	JmjN domain	7	4	2	3	7
PF00013	KH-domain	KH domain	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
PF01352	KRAB	KRAB box	204 (243)	0	0	0	0
PF00104	Hormone_rec	Ligand-binding domain of nuclear hormone receptor	47	17	142 (147)	0	0
PF00412	LIM	LIM domain containing proteins	62 (129)	33 (83)	33 (79)	4 (7)	10 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
PF02344	Myc-LZ	Myc leucine zipper domain	1	0	0	0	0
PF01753	Zf-MYND	MYND finger	14	14	9	1	7
PF00628	PHD	PHD-finger	68 (86)	40 (53)	32 (44)	14 (15)	96 (105)
PF00157	Pou	Pou domain--N-terminal to homeobox domain	15	5	4	0	0
PF02257	RFX_DNA_binding	RFX DNA-binding domain	7	2	1	1	0
PF00076	Rrm	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)
PF02037	SAP	SAP domain	15	8	5	5	6 (7)
PF00622	SPRY	SPRY domain	44 (51)	10 (12)	5 (7)	3	6
PF01852	START	START domain	10	2	6	0	23
PF00907	T-box	T-box	17 (19)	8	22	0	0
PF02135	Zf-TAZ	TAZ finger	2 (3)	1 (2)	6 (7)	0	10 (15)
PF01285	TEA	TEA domain	4	1	1	1	0
PF02176	Zf-TRAF	TRAF-type zinc finger	6 (9)	1 (3)	1	0	2

PF00352	TBP	Transcription factor TFIID (or TATA-binding protein, TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
PF00567	TUDOR	TUDOR domain	9 (24)	9 (19)	4 (5)	0	2
PF00642	Zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	17 (22)	6 (8)	22 (42)	3 (5)	31 (46)
PF00096	Zf-C2H2**	ZInc finger, C2H2 type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
PF00097	Zf-C3HC4	Zinc finger, C3HC4 type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
PF00098	Zf-CCHC	Zinc knuckle	9 (17)	6 (10)	17 (33)	7 (13)	68 (91)

表 19 ヒト *H. sapiens* (H) および ショウジョウバエ *D melanogaster* (D)、線虫 *C elegans* (C)

Panther family/subfamily*	H	F	W	Y	A
Neural structure, function, development					
Ependymin	1	0	0	0	0
Ion channels					
Acetylcholine receptor	17	12	56	0	0
Amiloride-sensitive/degenerin	11	24	27	0	0
CNG/EAG	22	9	9	0	30
IRK	16	3	3	0	0
ITP/ryanodine	10	2	4	0	0
Neurotransmitter-gated	61	51	59	0	19
P2X purinoceptor	10	0	0	0	0
TASK	12	12	48	1	5
Transient receptor	15	3	3	1	0
Voltage-gated Ca^{2+} alpha	22	4	8	2	2
Voltage-gated Ca^{2+} alpha-2	10	3	2	0	0
Voltage-gated Ca^{2+} beta	5	2	2	0	0
Voltage-gated Ca^{2+} gamma	1	0	0	0	0
Voltage-gated K^{+} alpha	33	5	11	0	0
Voltage-gated KQT	6	2	3	0	0
Voltage-gated Na^{+}	11	4	4	9	1
Myelin basic protein	1	0	0	0	0
Myelin PO	5	0	0	0	0
Myelin proteolipid	3	1	0	0	0
Myelin-oligodendrocyte glycoprotein	1	0	0	0	0
Neuropilin	2	0	0	0	0
Plexin	9	2	0	0	0
Semaphorin	22	6	2	0	0
Synaptotagmin	10	3	3	0	0
Immune response					
Defensin	3	0	0	0	0
Cytokine†	86	14	1	0	0
GCSF	1	0	0	0	0

GCSF	1	0	0	0	0
GMCSF	1	0	0	0	0
Intercrine alpha	15	0	0	0	0
Intercrine beta	5	0	0	0	0
Inteferon	8	0	0	0	0
Interleukin	26	1	1	0	0
Leukemia inhibitory factor	1	0	0	0	0
MCSF	1	0	0	0	0
Peptidoglycan recognition protein	2	13	0	0	0
Pre-B cell enhancing factor	1	0	0	0	0
Small inducible cytokine A	14	0	0	0	0
Sl cytokine	2	0	0	0	0
TNF	9	0	0	0	0
Cytokine receptor†	62	1	0	0	0
Bradykinin/C-C chemokine receptor	7	0	0	0	0
Fl cytokine receptor	2	0	0	0	0
Interferon receptor	3	0	0	0	0
Interleukin receptor	32	0	0	0	0
Leukocyte tyrosine kinase receptor	3	0	0	0	0
MCSF receptor	1	0	0	0	0
TNF receptor	3	0	0	0	0
Immunoglobulin receptor†	59	0	0	0	0
T-cell receptor alpha chain	16	0	0	0	0
T-cell receptor beta chain	15	0	0	0	0
T-cell receptor gamma chain	1	0	0	0	0
T-cell receptor delta chain	1	0	0	0	0
Immunoglobulin FC receptor	8	0	0	0	0
Killer cell receptor	16	0	0	0	0
Polymeric-immunoglobulin receptor	4	0	0	0	0
MHC class I	22	0	0	0	0
MHC class II	20	0	0	0	0
Other immunoglobulin†	114	0	0	0	0
Toll receptor-related	10	6	0	0	0
Developmental and homeostatic regulators					
Signaling molecules†					
Calcitonin	3	0	0	0	0
Ephrin	8	2	4	0	0
FGF	24	1	1	0	0
Glucagon	4	0	0	0	0
Glycoprotein hormone beta chain	2	0	0	0	0
Insulin	1	0	0	0	0
Insulin-like hormone	3	0	0	0	0
Nerve growth factor	3	0	0	0	0

Neuregulin/heregulin	6	0	0	0	0
neuropeptide Y	4	0	0	0	0
PDGF	1	1	0	0	0
Relaxin	3	0	0	0	0
Stannocalcin	2	0	0	0	0
Thymopoeitin	2	0	1	0	0
Thyomosin beta	4	2	0	0	0
TGF-β	29	6	4	0	0
VEGF	4	0	0	0	0
Wnt	18	6	5	0	0
Receptors†					
Ephrin receptor	12	2	1	0	0
FGF receptor	4	4	0	0	0
Frizzled receptor	12	6	5	0	0
Parathyroid hormone receptor	2	0	0	0	0
VEGF receptor	5	0	0	0	0
BDNF/NT-3 nerve growth factor receptor	4	0	0	0	0
Kinases and phosphatases					
Dual-specificity protein phosphatase	29	8	10	4	11
S/T and dual-specificity protein kinase†	395	198	315	114	1102
S/T protein phosphatase	15	19	51	13	29
Y protein kinase†	106	47	100	5	16
Y protein phosphatase	56	22	95	5	6
Signal transduction					
ARF family	55	29	27	12	45
Cyclic nucleotide phosphodiesterase	25	8	6	1	0
G protein-coupled receptors‡	616	146	284	0	1
G-protein alpha	27	10	22	2	5
G-protein beta	5	3	2	1	1
G-protein gamma	13	2	2	0	0
Ras superfamily	141	64	62	26	86
G-protein modulators†					
ARF GTPase-activating	20	8	9	5	15
Neurofibromin	7	2	0	2	0
Ras GTPase-activating	9	3	8	1	0
Tuberin	7	3	2	0	0
Vav proto-oncogene family	35	15	13	3	0
Transcription factors/chromatin organization					
C2H2 zinc finger-containing†	607	232	79	28	8
COE	7	1	1	0	0
CREB	7	1	2	0	0
ETS-related	25	8	10	0	0

Forkhead-related		34	19	15	4	0
FOS		8	2	1	0	0
Groucho		13	2	1	0	0
Histone H1		5	0	1	0	0
Histone H2A		24	1	17	3	13
Histone H2B		21	1	17	2	12
Histone H3		28	2	24	2	16
Histone H4		9	1	16	1	8
Homeotic†		168	104	74	4	78
ABD-B		5	0	0	0	0
Bithoraxoid		1	8	1	0	0
Iroquois class		7	3	1	0	0
Distal-less		5	2	1	0	0
Engrailed		2	2	1	0	0
LIM-containing		17	8	3	0	0
MEIS/KNOX class		9	4	4	2	26
NK-3/NK-2 class		9	4	5	0	0
Paired box		38	28	23	0	2
Six		5	3	4	0	0
Leucine zipper		6	0	0	0	0
Nuclear hormone receptor†		59	25	183	1	4
Pou-related		15	5	4	1	0
Runt-related		3	4	2	0	0
ECM adhesion						
Cadherin		113	17	16	0	0
Claudin		20	0	0	0	0
Complement receptor-related		22	8	6	0	0
Connexin		14	0	0	0	0
Galectin		12	5	22	0	0
Glypican		13	2	1	0	0
ICAM		6	0	0	0	0
Integrin alpha		24	7	4	0	1
Integrin beta		9	2	2	0	0
LDL receptor family		26	19	20	0	2
Proteoglycans		22	9	7	0	5
Apoptosis						
Bcl-2		12	1	0	0	0
Calpain		22	4	11	1	3
Calpain inhibitor		4	0	0	0	1
Caspase		13	7	3	0	0
Hemostasis						
ADAM/ADAMTS		51	9	12	0	0
Fibronectin		3	0	0	0	0
Globin		10	2	3	0	3

Matrix metalloprotease	19	2	7	0	3
Serum amyloid A	4	0	0	0	0
Serum amyloid P (subfamily of Pentaxin)	2	0	0	0	0
Serum paraoxonase/arylesterase	4	0	3	0	0
Serum albumin	4	0	0	0	0
Transglutaminase	10	1	0	0	0
Other enzymes					
Cytochrome p450	60	89	83	3	256
GAPDH	46	3	4	3	8
Heparan sulfotransferase	11	4	2	0	0
Splicing and translation					
EF-1alpha	56	13	10	6	13
Ribonucleoproteins†	269	135	104	60	265
Ribosomal proteins†	812	111	80	117	256

* The table lists Panther families or subfamilies relevant to the text that either (i) are not specifically represented by Pfam (Table 18) or (ii) differ in counts from the corresponding Pfam models.

†This class represents a number of different families in the same Panther molecular function subcategory.

‡This count includes only rhodopsin-class, secretin-class, and metabotropic glutamate-class GPCRs.

7.3. ヒトゲノムと配列決定済みの他の真核生物ゲノムとの間の相違

脊椎動物分類上の分子基盤を探求するため、ヒトゲノムを他の配列決定済みの真核生物ゲノムと 3 つのレベルにて比較した。すなわち、分子機能、蛋白ファミリー、蛋白ドメインである。

脊椎動物に特徴的な発生学的プロセス、細胞学的プロセスを明らかにするため、分子の相違を表現型相違に関係づけることが可能である。表 18 および表 19 は、抜粋した蛋白ファミリー/ドメインファミリー（配列類似性により定義。例、セリン スレオニン蛋白燐酸化酵素）および、スーパーファミリー（配列関連ファミリーをいくつか含むと思われる共通分子機能により定義。例、サイトカイン）に関して、配列決定済みの全真核生物ゲノム間の比較をしめす。これらの表において、非常に大きい（スーパー）ファミリー、もしくは、他の配列決定済みの真核生物ゲノムと比較しヒトでは有意に異なる（スーパー）ファミリーに焦点を置いた。最も顕著なヒトゲノムでの増幅は、以下に関与する蛋白群にて生じることを発見した。すなわち、(i)後天性免疫、(ⅱ)神経発生、神経構造、神経機能、(iii)発生および恒常性維持における細胞間および細胞内シグナル経路、(iv)止血、(ⅴ)アポトーシスである。

後天性免疫

ヒトゲノムとショウジョウバエ *Drosophila* ゲノムもしくは線虫 *C.elegans* ゲノム間における最も顕著な相違点は、後天性免疫に関与する遺伝子の出現である（表 18 および表 19）。後天性免疫応答は脊椎動物にてのみ生じる防御系であるため、これは予測された。ヒトゲノムにおいて、22 個のクラス I 主要組織適合性複合体（MHC, major histocompatibility complex）抗原遺伝子、22 個のクラス II MHC 抗原遺伝子の他に 114 個の免疫グロブリン遺伝子を発見した。さらに、同一起源免疫グロブリン受容体ファミリーにおいて 59 個もの遺伝子が存在する。ドメインレベルでは、MHC 等の分子構成のために古代免疫グロブリン類が、そして免疫エフェクター細胞と細胞外マトリックス間相互作用に介在するいくつかの細胞接着分子を構成するためにインテグリン類が、拡張・補充されていることをみればこの点は例証されている。脊椎動物特異的な蛋白には、分泌型 4-α螺旋束蛋白群、すなわちサイトカインおよびケモカインからなるパラクライン型免疫調節因子ファミリーが含まれる。サイトカイン受容体シグナル伝達に関連する細胞質シグナル伝達コンポーネントも、同様にハエおよび線虫にはわずかしか存在しない。これらの蛋白ドメインには、転写時のシグナルトランスデューサーとアクチベーター（STAT, signal transducer and activator of transcription）、サイトカインシグナルのサプレッサー（SOCS, suppressors of cytokine signaling）、活性化 STAT の蛋白インヒビター（PIAS, protein inhibitors of activated STATs）が含まれる。対照的に、Toll 受容体のような先天性免疫応答に役割を果たす動物特異的蛋白ドメインは、ヒトゲノムにおいて有意に増幅して いるとは思えない。

神経発生、神経構造、神経機能

ヒトゲノムでは、線虫ゲノムおよびハエゲノムと比較して、神経発生に関与する蛋白ファミリーメンバー数に顕著な増加がみられる。これらの例には、エペンデミン、神経増殖因子（NGF）等の神経栄養因子、セマフォリン等のシグナル分子のみならず、ミエリン蛋白、電位依存型イオンチャネル、シナプトタグミン等のシナプス蛋白といった神経構造および機能に直接関与している多数の蛋白が含まれる。これらの結果は、これらの動物分類の神経系間において知られている表現型の既知相違点と高い相関を示す。顕著なものでは、(i) ニューロン数とコネクション数の増加、(ⅱ)神経細胞タイプ数の増加（ハエ、線虫では数百タイプであるのに比較して、ヒトでは千以上のタイプが存在する）[121]、(ⅲ)個々の神経軸索長の増加、(iv)グリア細胞の有意な増加、特に、ニューロンと同じ幹細胞から分化するが電気的には不活性な支持細胞となる、髄鞘を形成するグリア細胞の出現である。多くの顕著な蛋白増幅が神経発生には関与している。細胞接着を仲介する細胞外ドメインにおいて、コネキシンドメインを有する蛋白 [122] はヒトにのみ存在する。ハエゲノムもしくは線虫ゲノムに存在しないこれらの蛋白は、細胞間チャネルの構成的サブユニットおよび、電気的カップリングの構造基盤となっている。軸索誘導および神経細胞ネットワーク形成は、エフリンのサブセット、および、それらと同一起源の、位相投射を樹立するための位置標識として働く受容体型チロシン燐酸化酵素を介している [123]。セマフォリン（ハエでは 6 個、

線虫では 2 個のメンバーが存在するのと比較して、ヒトでは 22 個の遺伝子が存在する）およびその受容体（ニューロピリンおよびプレキシン）の生物学的役割は、軸索誘導分子であると思われる[124]。神経栄養因子や一部のサイトカインのようなシグナル分子は、神経細胞の生存、増殖、軸索誘導を調節することが示されてきた[125]。Notch 受容体およびリガンドはグリア細胞運命決定およびグリア新生に重要な役割を担っている[126]。

ヒトにて増幅された他の遺伝子ファミリーは、神経構造および機能において鍵となる役割を直接的に担う。シナプス小胞の膜融合と放出に関与するカルシウムセンサー（もしくは受容体）として機能するシナプス伝達調節蛋白として当初は発見されたシナプトタグミン（無脊椎動物に比較してヒトでは 2 倍以上に増幅した遺伝子ファミリー）は、この様な例のひとつである[127]。神経細胞特異的アダプター分子中の PDZ ドメインおよび SH3 ドメインがヒトにて同時増大したことは興味深い。例としては、シナプス間隙でのチャネル機能を調節すると思われる蛋白が含まれる[128]。同様に、EAG サブファミリー（サイクリックヌクレオチド依存性チャネルに関連)、電位依存性カルシウム／ナトリウムチャネル・ファミリー、内向き整流カリウムチャネル・ファミリー、電位依存性カリウムチャネルのαサブユニット・ファミリーを含む、いくつかのイオンチャネル・ファミリーにおける増幅（表 19）も指摘される。電位依存性ナトリウムチャネルおよび電位依存性カリウムチャネルは神経細胞において活動電位を生み出すことに関与している。電位依存性カルシウムチャネルと合わせ、これらは神経伝達物質の放出、神経突起の成長、短期記憶に活動電位をカップリングすることで重要な役割を担っている。最近の知見によると、カルシウムが調節するナトリウムチャネルとシナプトタグミンの結合により、神経細胞興奮性が樹立、調節されると思われる[129]。

ミエリン塩基性蛋白およびミエリン結合糖蛋白は、脊椎動物の中枢神経系および末梢神経系における主要構成蛋白である。ミエリン P0 は末梢神経髄鞘の主要構成蛋白であり、ミエリン蛋白脂質およびミエリン乏突起神経膠細胞（oligodendrocyte）糖蛋白は中枢神経系に見られる。これらのミエリン蛋白のいずれかに突然変異が生じると、髄鞘の喪失および神経線維結合の重度障害という病理的に重篤な脱髄鞘が起こる[130]。ヒトでは、ミエリン形成に関与する異なる 4 つのファミリーに属する遺伝子が少なくとも 10 個（ミエリン P0 5 個、ミエリン蛋白脂質 3 個、ミエリン塩基性蛋白、ミエリン乏突起神経膠細胞糖タンパク（MOG, myelin-oligodendrocyte glycoprotein)、そして遠縁にある MOG ファミリー関連メンバーが恐らく存在すると思われる。ハエはミエリン蛋白脂質をただ一つ、線虫は全く持たない。

発生と恒常性維持機能における細胞間および細胞内シグナル伝達経路

ヒトにおいて、無脊椎動物に比較して増幅した多くの蛋白ファミリーはシグナル伝達過程に関与する。特に、発生および分化に応答したシグナル過程に関与する蛋白ファミリーが

挙げられる（表 18 および表 19）。これらには、分泌ホルモンや細胞成長因子、受容体、細胞内シグナル分子、転写因子が含まれる。

ヒトゲノムで強化されている発生関与シグナル分子には、wnt、トランスフォーミング成長因子（TGF, transforming growth factor）- β、線維芽細胞成長因子（FGF, fibroblast growth factor）、神経成長因子（NGF, nerve growth factor）、血小板由来細胞成長因子（PDGF）およびエフリン等の細胞成長因子が含まれる。これらの成長因子は組織分化およびアクチン 細胞骨格制御および核機能制御に関与する広範な細胞学的プロセスに影響を及ぼす。ヒトにおいて、これら発生関与リガンドに対応する受容体も同様に増幅している。例えば、今回の解析から少なくとも 8 個のヒト・エフリン遺伝子（ハエ 2 個、線虫 4 個）および 12 個のエフリン受容体（ハエ 2 個、線虫 1 個）が存在すると示唆された。Wnt シグナル経路においては、18 個の wnt ファミリー遺伝子（ハエ 6 個、線虫 5 個）および 12 個の frizzled 受容体（ハエ 6 個、線虫 5 個）を発見した。Wnt 経路の下流にある転写コリプレッサーの Groucho ファミリーは、ヒトでは 13 個と予測され（ハエ 2 個、線虫 1 個）、さらに顕著に増幅している。

シグナル伝達に関与する細胞外接着分子はヒトゲノムにて増幅している（表 18 および表 19）。これら接着分子ドメインのいくつかが細胞外マトリクス・プロテオグリカンと結合することは、宿主防御、形態形成、組織修復に重大な役割を演じる [131]。これらの結合調節というヘパラン硫酸プロテオグリカンの明確な役割 [132] と一致して、ヒトゲノムでは線虫およびハエに比べてヘパラン硫酸の硫酸転移酵素の増幅があることを発見した。ヘパラン硫酸転移酵素は組織分化を調節する [133]。同様にヒトでは、アクチン 細胞骨格構造蛋白にも増幅が見られる。ハエおよび線虫と比較して、ヒトで爆発的に増幅しているのは、ネブリン反復配列（蛋白あたり平均 35 ドメイン）、アグレカン反復配列（蛋白あたり平均 12 ドメイン）、プレクチン反復配列（蛋白あたり平均 5 ドメイン）である。これら反復配列は、アクチン細胞骨格調節に関与し、神経、筋肉、脈管組織に著明な発現が認められる蛋白に含まれている。

配列決定済みの 5 つの真核生物間の比較により、細胞質シグナル伝達に関与したいくつかの蛋白ファミリーおよびドメインが増幅していることが明らかになった（表 18）。特に、発生制御および後天性免疫において役割を担うシグナル伝達経路が実質的に強化されている。Ras スーパーファミリーGTPase および、これらに付随する GTPase 活性因子（GAP）、GTP 交換因子（GEF）は、ヒトにおいて 2 倍もしくはそれ以上の増幅をしめす。ヒトゲノムおよび線虫 *C.elegans* ゲノムには、ほぼ同数のチロシン燐酸化酵素が存在するが、ヒトでは燐酸化チロシン・シグナル伝達に関与する SH2 ドメイン、PTB ドメイン、ITAM ドメインの増加が見られる。さらに、ハエもしくは線虫ゲノムと比較した際、ヒトゲノムでは燐酸 2 エステラーゼに 2 倍以上の増幅が見られる。

細胞内シグナル分子の下流エフェクターには、発生上の運命を伝達する転写因子が含まれる。ハエ・ゲノムと比較した際、転写因子であるリガンド結合型核内ホルモン受容体群は顕著に増幅していることが示されている。ただし、線虫と比べれば増幅程度は顕著ではない（表 18 および表 19）。ヒトにて最も強烈な増幅があるのは、恐らく C2H2 ジンクフィンガー転写因子であろう。Pfam プログラムにて、564 個のヒト蛋白において合計 4500 個の C2H2 ジンクフィンガー・ドメインが検出された。これに対し、234 個のハエ蛋白では 771 個である。これは、C2H2 転写因子数のみならず転写因子あたりの DNA 結合モチーフ数（ヒト平均 8 個、ハエ平均 3.3 個、線虫平均 2.3 個）が劇的に拡大してきたことを意味する。さらに、これら転写因子の多くが、ハエや線虫では見られない KRAB もしくは SCAN ドメインのどちらかをもっている。これらドメインは、転写因子の重合体形成に関与し、転写因子の結合組み合わせを増大するものである。一般的に、転写因子ドメインのほとんどは 3 種類の動物で共通だが、これらドメインを再分類してみると生物種特異的転写因子ファミリーがあるという結果となった。ヒト、ハエ、線虫にみられるドメインの組み合わせは、ハエおよびヒトにおける BTB ドメインと C2H2 ドメインの組み合わせ、3 つの動物ゲノムにおけるホメオドメイン単独もしくは Pou ドメインおよび LIM ドメインとの組み合わせである。しかし植物では、異なるセットの転写因子が増幅している。すなわち、myb ファミリーおよび、VP1 ドメインと AP2 ドメインを含むユニークな蛋白セットである [(134)]。酵母ゲノムは多細胞真核生物と比較した際、転写因子を少数しか持たず、そのレパートリーは代謝制御に関与する酵母特異的 C6 転写因子ファミリーの増幅に限られる。

ここまで他の真核生物ゲノムと比較した際にヒトゲノムにて拡張を示すシグナル伝達分子のサブセットについて説明してきたが、ほとんどの蛋白ドメインが非常によく遺伝子的に保存されていることを述べておくべきであろう。興味深いことに、線虫とヒトはおおよそ同数のチロシン燐酸化酵素およびセリン・チロシン燐酸化酵素をもっている（表 19）。しかし、これらは単に触媒ドメインを数えあげているだけであるという点は重要である。というのも、これらドメインを有する蛋白はまた、意味のある組合せに多様性がある蛋白の結合ドメインにも、広範なレパートリーがあることを示すからである。

止血

止血は、凝固経路の血漿蛋白分解酵素および、血管内皮と血小板の相互作用により、主に調節される。脊椎動物と無脊椎動物間に知られる解剖学的および生理学的差異に一致して、止血に重要な蛋白を構成する細胞外接着ドメインは、ヒトではハエおよび線虫に比べ増幅している（表 18 および表 19）。血球系細胞と血管マトリクス間の表面相互作用に関与する FIMAC、FN1、FN2、C1q 等のドメインの進化を記しておきたい。さらに、VWA、VWC、VWD、クリングル、FN3 等のいっそう古くから存在する動物特異的ドメインが、止血調節

に関与する多ドメイン蛋白に活発に取り込まれてきた。セリン蛋白分解酵素の総数に大きな増幅は認めないが、この酵素ドメインは血管系構成全体で蛋白分解調節を担ういくつかの多ドメイン蛋白へ特異的に取り込まれてきた。これらは、キニンおよび補体経路に属する血漿蛋白において示される。ADAM（a disintegrin and metalloprotease）および MMPs（matrix metalloproteases）の 2 つのマトリックス・メタロプロテアーゼファミリーにおいて有意な増幅が見られる（表 19）。細胞外マトリックス（ECM）蛋白の分解は、癌、関節炎、アルツハイマー氏病、種々の炎症状態等の疾患において、組織発生および組織分解に重要である。[135、136]。ADAM は、フィブリノーゲン分解および血球系コンポーネントと血管マトリックス・コンポーネント間の相互作用に重要な役割をもつ膜貫通型蛋白ファミリーである。これらの蛋白は、マトリックス蛋白、さらにはシグナル分子を切断することが示されてきた。ADAM-17 は TNF（tumor necrosis factor、腫瘍壊死因子）－α を転換し、ADAM-10 は Notch シグナル経路に関与すると考えられてきた [135]。今回、マトリックスメタロプロテアーゼ・ファミリーメンバーを 19 個、ADAM および ADAM-TS ファミリーメンバーを合計 51 個を同定した。

アポトーシス

真核生物間でアポトーシス経路構成因子の一部が進化過程で保存されることは、発生制御において、そして病原体やストレス・シグナルへの応答において、アポトーシスが中枢的役割を担うことと矛盾しない。プログラム細胞死、すなわちアポトーシスに関与するシグナル伝達経路には、細胞外ドメイン、アダプター（蛋白―蛋白相互作用の）ドメイン、エフェクター酵素や調節酵素にみられるドメインを含んだ、よく解析されたドメイン間の相互作用により仲介される [137]。真核生物間の多様性と、ハエと線虫に対して比較した際、ヒトにおける相対的な増幅程度の推定値を出すため、アポトーシス経路にのみ存在する中枢的なアダプターとエフェクター酵素ドメインの蛋白数を列挙してみた（表 18）。アポトーシス調節に限定される蛋白における DED ドメイン等のアダプタードメインは、脊椎動物特異的であったが、BIR、CARD、Bcl2 等はハエや線虫にも存在した（ただし、ヒトでの Bcl2 ファミリー・メンバー数は有意に増幅している）。植物および酵母にはカスペースが存在しないが、カスペース様分子、すなわちパラ・カスペースおよびメタ・カスペースの存在が報告されている [138]。他の動物ゲノムと比較して、ヒトゲノムでは、カスペースおよびカルパイン・ファミリー等のアポトーシス・カスケードに関与する蛋白のみならず、アポトーシスに関与するアダプターおよびエフェクター・ドメインを含む蛋白の増幅がみられる。

他の蛋白ファミリーの増幅

代謝酵素： ハエもしくは線虫と比較してヒトにはチトクロム P450 遺伝子が少ない。一方、リポオキシゲナーゼ（ヒト 6 個）は脊椎動物および植物特異的であるが、リポオキシゲナーゼ活性化蛋白（ヒト 4 個）は脊椎動物特異的と思われる。リポオキシゲナーゼはアラキ

ドン酸代謝に関与し、その活性化蛋白はアレルギー応答から癌にいたる様々なヒト病理に関与すると考えられてきた。最も驚くべきヒトにおける遺伝子増幅の一つは、グリセロアルデヒド-3-燐酸脱水素酵素（GAPDH, glyceraldehyde-3-phosphate dehydrogenase）遺伝子数（ヒト 46 個、ハエ 3 個、線虫 4 個）である。しかし、多くの逆転写された GAPDH 偽遺伝子が存在すること [139] が、この見かけ上の増幅を説明する証拠もある。しかし、細菌からヒトにいたる全ての生物種に見られ、基礎代謝に関与する進化的に保存された酵素として長らく知られてきた GAPDH に、他の機能があることが近年示されてきたことは大変興味深い。GAPDH は第二活性 [140] を示し、ウラシル DNA グライコシラーゼとして作用する。これは細胞周期調節因子として機能し [141]、アポトーシスに関与すると考えられてきた [142]。

翻訳：ヒトで著明に増幅しているもう一つのセットは、翻訳機構に関与するファミリー群にある。今回、ゲノム中にそれぞれのサブユニットが少なくとも 10 個のコピーをもつ 28 個の異なるリボゾーム・サブユニットを同定した。全てのリボゾーム蛋白遺伝子は、線虫もしくはハエに比較して平均約 8〜10 倍増幅している。逆転写された偽遺伝子がこれらの増幅の多くを占めると思われる（上述の考察と（143）を参照）。近年の知見では、リボソーム蛋白の多くが蛋白合成とは別個の 2 次的機能をもつと示唆されている。例えば、L14a および関連 L7 サブユニット（ヒト 36 個）はアポトーシスを誘導することが示されている [144]。同様に延長因子 1α ファミリー（eEF1A、 ヒトで 56 遺伝子）では 4〜5 倍の増幅が存在する。この増幅の多くは、逆転写に由来すると思われるイントロンを持たないパラログであるようだ。さらに、これらの多くが偽遺伝子であると思われる証拠がある [145]。しかし、この延長因子の 2 つ目の型である eEF1A2 は筋肉で組織特異的に発現され、偏在的に発現する eEF1A と相補的発現様式を示す [146]。

リボヌクレオ蛋白：オルタナティブスプライシングにより、単一遺伝子から多数の転写産物が生じる。従って、生物の全蛋白に付加的多様性をもたらすことができる。今回、269 個の遺伝子がリボヌクレオ蛋白であることを同定した。この数は線虫リボヌクレオ蛋白遺伝子数の 2.5 倍以上、ハエの 2 倍、アラビドプシス・ゲノムで同定された 265 個とおよそ同じである。ヒトでのリボヌクレオ蛋白遺伝子の多様性が、スプライシング・レベルもしくは翻訳レベルで遺伝子調節に寄与しているかは不明である。

翻訳後修飾：このプロセスに関わるセットにて最も顕著な増幅がみられるのは、止血やアポトーシス等の細胞学的プロセスにおいて蛋白架橋触媒を行うカルシウム依存性酵素であるトランスグルタミナーゼである [147]。ビタミン K 依存的γカルボキシラーゼ遺伝子産物は、凝固因子、オステオカルシン、マトリクス GLA 蛋白に見られる GLA ドメイン(ハエ、線虫には存在しない)に作用する [148]。チロシン化蛋白硫酸転移酵素は、凝集因子およびケモカイン受容体を含む炎症および止血に関与した蛋白の翻訳後修飾に関与する [149]。核蛋白

の修飾に関与するドメイン数には有意な増加はないが、現時点で配列がわかっている他のゲノムには存在しないヒト予測蛋白において、ドメイン・アレンジメントが多数存在する。これらには、ユビキチン・フィンガードメインをもつ HD6 にて、ヒストン脱アセチル化酵素ドメインが直列的に存在することが含まれる。これはハエゲノムには存在しない特徴である。さらに重要な核調節酵素 PARP（poly-ADP ribosyl transferase、ポリ ADP リボシル転移酵素）ドメインが、ヒトでは蛋白結合ドメイン BRCT および VWA にそれぞれ融合するという例も挙げられる。

まとめ

ハエおよび線虫と比較した際、ヒトで見られる表現型の複雑さの差については、いくつかの解釈が可能である。これらの一部は、免疫系、止血、神経、脈管、細胞骨格の複雑性における顕著な差異に相関する。ヒトゲノムがこれまでの予測に比べ少数の遺伝子しかもたないという点は、蛋白構造や転写・翻訳調節、蛋白翻訳後修飾、翻訳後調節レベルで、組み合わせの多様さにより補われると思われる。組み合わせの多様さを増大もしくは変化させるためにドメインを混ぜ合わせることは、蛋白数の絶対数を劇的に増やすことなく、蛋白－蛋白相互作用を介在する能力を指数関数的に増大することができる[150]。明らかな新規性がある（配列解析の展望からみて）蛋白ドメインの進化、および、量的かつ質的なドメイン融合（既存ドメインへの新規ドメインの補充）によって増大する制御上の複雑性の 2 つが、今回、ヒトにて観察された特徴である。おそらくこの傾向を示す最良の例示となるのは、C2H2 ジンク・フィンガーをもった転写因子群であろう。そこでは、KRAB や SCAN 等の脊椎動物に特異的なドメインと共に、蛋白あたりのドメイン数拡張が見られる。

特異的蛋白クラスの翻訳調節のため、ヒトゲノムにおいては内在性リボゾーム・エントリー部位が顕著に使用されているという近年の報告から、このプロセスが用いられる程度を完全に同定すべく、この分野でのさらなる研究がヒトゲノムにおいて要求されると思われる[151]。これら修飾に関与する蛋白ファミリーの一部に増幅例が存在することを示したが、翻訳後レベルで、蛋白プロセッシングにおける複雑性の増大と相関しているかどうかの評価には、さらなる実験的証拠が要求される。ヒトにおける転写後プロセッシングおよびイソフォーム発生の程度については、全面的な目録化作業が残っている。スプライソゾーム機構の保守的性質から、このレベルにおける調節機構を解剖するためには更なる解析が必要であろう。

第 8 章 結論

8.1. 全ゲノム配列解析法対 BAC-by-BAC 法

さまざまなゲノムサイズと反復配列を有する多様な生物に全ゲノムショットガン法を適用した経験から、我々は本法の利点と弱点を評価することができる。多数の微生物ゲノム、ショウジョウバエ、そして今回ヒトに対して用いて成功したことで、本法の有用性に疑いの余地はない。本法によって配列決定した多数の微生物ゲノム [15, 80, 152] から、メガ塩基対サイズのゲノムでは、de novo のメイト対配列以外をインプットしなくても効率よく配列決定できることが実証された。ショウジョウバエやヒトのようにもっと複雑なゲノムでは、順序よく整列化されたマーカーの物理地図情報が、配列骨格の広域にわたる整列に重要であった。配列骨格を染色体へ導入するには、マーカーの数自体よりもマップの品質（マーカーの整列化）の方がもっと重要である。このマッピングは配列解析と同時に行なうこともできたが、マッピングデータが予め存在したことが役に立った。アラビドプシス（*A. thaliana*）ゲノムの配列解析においては、個々の BAC クローンを配列解析していくことで、動原体領域に配列がうまく延びていき、複雑な反復領域の高解像度解析がなされた。同じように、ショウジョウバエでは、反復性の高い動原体やテロメアの近傍領域で BAC 物理地図が非常に役に立った。WGA 法は、ゲノムのユニークな領域で高品質の再構築を行なうために有用であることが判った。ゲノムサイズと、もっと重要なのは反復配列であるが、この二つが大きく増えてくるにつれ、WGA 法では反復配列の再構築が難しくなってくる。

個々のクローン配列決定法では、コストと全体の効率を考慮すると、今後の大規模なゲノム配列決定プロジェクトに対し無比の戦略として正当化することは難しい。しかし、配列アセンブリの曖昧さを解消するために、BAC や他のクローンに基づくマッピングと配列解析戦略を選択して適用する方式は、コンピュータによる計算法のみで解消するわけではないが、明らかに探求する価値はある。全ゲノムショットガン段階と BAC クローン配列解析段階の両段階で充分量の配列カバー倍数がある場合に限り、全ゲノム配列解析への混成アプローチはうまく機能するであろう。ヒトゲノムアセンブリにおける我々の経験では、全ゲノム配列データと BAC ショットガン配列データの両方で少なくとも 3 倍のカバー倍数が必要であることが示唆されている。

8.2. ヒトの遺伝子数は少ない

我々は、ホモサピエンスの真正染色質の約 95%まで配列決定し、アセンブリを行った。さらに、新しい自動化遺伝子予測法を用いて、ヒト遺伝子の予備カタログを作成した。この結果、一つの大きな驚くべき事実が判明した。すなわち、遺伝子数はこれまでの分子予測

値（5 万～14 万個）よりも遙かに少ない（26,000～38,000 個）ことが分かったのである。この格差がいかなる理由であれ、詳しい注釈をつけ、比較ゲノム学（特に *Mus musculus* ゲノムを用いて）を駆使し、複雑な表現型を注意深く分子レベルで分析することによってのみ、我々のゲノムの基本的な「パーツリスト」の重要問題は解明されるであろう。確かに、こうした分析はまだ完全なものではなく、各転写単位のより精密な構造が決定されてくれば、ここ数年以内にかなりの改善が得られると思われる。踏み出すべき第一歩は、何故 EST データ由来の推定遺伝子数が我々の推定値とこれほどまでに一致しないのかを明らかにすることである。EST 由来の遺伝子数が大きくなってしまったのは、次に述べる理由によるのではなかろうか。（ⅰ）翻訳されない 3'と 5'のリーダー配列とトレーラー配列に様々な長さのものが存在している、（ⅱ）RNA プロセシングではしばしばイントロン領域がスプライスされずに残ってしまうような場合が生じるが、このような予測できない変動についてはほとんど分かっていない、（ⅲ）ヒト遺伝子の約 40%が別途にスプライスされているという知見 [153] がある、（ⅳ）最後になったが、異種由来の核 RNA やゲノム DNA からの夾雑物が珍しくない EST ライブラリーの構築において、まだ解決されていない問題があるためである。もちろん、以上のことを裏づける EST データや蛋白質データがないため、予測されずにいる遺伝子が存在している可能性もある。ただし、この数は遺伝子予測にマウスゲノムのデータを利用すれば、制限されるはずである。ゲノム配列決定が始まったばかりであることは真実であるが、究極的には、ある遺伝子の存在を示すために特異的な細胞種の中の mRNA を測定することが必要になるであろう。

J.B.S. Haldane は、1937 年に、生物の集団はそれが持ち得る遺伝子数に対して代価を払わねばならないであろうと推測した。彼は、遺伝子数があまりに多くなると、各接合子は大変多くの有害な新規突然変異を受け、集団自体が単純に自らを維持できなくなるとの説を立てた。この前提条件を踏まえ、さらに特定の遺伝子座における判っている突然変異の発生率と X 線誘発性突然変異の発生数などを考慮した Muller [154] は、1967 年に哺乳類のゲノムは最大でも 3 万を超えないであろうと計算した [155]。

ヒトの遺伝子座が 30,000 という推定値も Crow と Kimura [156] によって提唱された。*D. Melanogaster* に対する Muller の推定値は､遺伝子注釈付けから出されたハエゲノムの 13,000 に対して、10,000 である [26,27]。遺伝子数の最大理論値に対するこのような議論は、遺伝子負荷という単純化されたアイデアに基づいている。すなわち、全ての遺伝子には、有害な状態へ突然変異する率が、低いけれども一定レベルはある、という考え方である。とはいえ、多くのマウス・ハエ・虫・酵母のノックアウト突然変異モデルで、判別できるほど表現型が変ることは殆んどないことは、はっきりしている。

ヒト遺伝子がこの程度であったということは、ヒトの発達に固有の複雑性を生むメカニズ

ムや恒常性を維持する精巧なシグナル伝達系のメカニズムを我々自身が別に探さねばならないことを意味している。一つひとつの遺伝子・遺伝子産物の機能を調節している機構の数は多い。例えば、クロマチン構造の「開放」の程度とそれに伴う転写活性は、ヒストンと DNA の酵素的な修飾に関与する蛋白質複合体によって制御されている。表 19 に、核における制御に関与すると思われる多くの蛋白質を列挙した。

転写の位置・時期・品質は、核のシグナル伝達事象と密接につながっているが、こうした多くの蛋白質の組織特異的な発現とも関係しているのである。同じく重要なのは、インシュレータ、繰り返し配列、内因性ウイルスなどを含む調節 DNA エレメント [157]、刷り込み現象に置ける CpG アイランドのメチル化 [158]、転写活性を変化させるプロモーター・エンハンサーとイントロン領域である。またスプライセオソーム機構は、マルチサブユニットの蛋白質（表 19）に加えて、構造的・触媒的 RNA エレメント [159] から成り立っており、後者は開始・終結に関するいろいろな部位とスプライシングによって転写構造を制御している。従って、さまざまなクラスの RNA 分子を研究する必要がある [160]。すなわち、小さな核小体 RNA、アンチセンスリボレギュレーターRNA、X-遺伝子量代償に関する RNA、その他遺伝子発現制御で明確な役割を正当に評価できる構造的 RNA などである。RNA の編纂は、コード変化がmRNA レベルで直接生じている現象であるが、臨床的・生物学的に関連がある [161]。最後になったが、翻訳制御の例として、インターナルリボゾーマルエントリーサイトなどが挙げられる。これは、細胞周期の調節とアポトーシスに関与している蛋白質で見つかっている [162]。蛋白質レベルでは、蛋白 蛋白間相互作用の性質、蛋白修飾、局在化などにおける微小な変化が、細胞の生理的特性に劇的な影響を及ぼし得る [163]。それ故、このダイナミックな系は、活動を調整する多くの方法を有しており、そのことから考えれば、単独の遺伝子ごとに解析することによって複雑系を明らかにしていくのは、全く成功するとは思われない。

遺伝子の in situ 研究から、ヒトゲノムは、＜G+C＞含量、CpG アイランド、および遺伝子数において非対称的に構成されていることが明らかになった [68]。しかし遺伝子は、これまで予想されていたほど不均等に分布しているわけではない（表 9）[69]。ヒトゲノムの中の G+C が最も多い分画である H3 アイソコアは、従来考えられていたより多くの割合を占めており（約 9%）、最も遺伝子密度の高い分画であるが、予想の 40%弱ほどはなく、たかだか遺伝子の 25%を含んでいるに過ぎない。G+C の少ない L アイソコアは、ゲノムの 65%を占め、遺伝子は 48%である。この不均一性は、哺乳類における遺伝子複製の数百万年にわたる総括的結果であるが、脊椎動物ゲノムの「砂漠化」として述べられている [71]。何故、遺伝子密度の異なるクラスター領域があちこち存在しているのであろうか。これらは、歴史上の不慮の出来事だったのか、それとも淘汰と進化によってもたらされたのであろうか。このような不毛領域が必要でないのなら、ヒトゲノムよりサイズがはるかに小さい哺乳類

のゲノムを見つけ出すことができるはずである。事実、多くの種類のコウモリは、ヒトより格段に小さいサイズのゲノムを持っている。例えば、*Miniopterus* は、イタリアコウモリの一種であるが、ゲノムサイズはヒトゲノムのわずか 50%である [164]。同じように、アジアのホエジカの一種 *Muntiacus* は、ゲノムサイズがヒトゲノムの約 70%である。

8.3. ヒト DNA 配列のばらつきとゲノム全域に渡るその分布

今回のヒトゲノムは、多型についてほぼ一様な確認が完了した初の真核生物ゲノムである。我々は、300 万を上回る SNP を同定しマッピングしたものの、これは、SNP を見つけ一覧表を作成する仕事が申し分なく完全であることを示唆するものでは決してない。ただ、これらの SNP は、全体としてのヒト集団中に存在している SNP の 1 群を表しているに過ぎない。それにもかかわらず、ゲノム全域にわたるばらつきを初めて一瞥すれば、ゲノム全体に散らばった SNP の強い不均一性が目に付く。DNA の多型性は、突然変異・遺伝子移動・淘汰・遺伝的ドリフトなど、これまでに見られた集団の遺伝力を示すスナップショットを携えているのである。SNP の高密度アレイを用いることができれば、このような因子のそれぞれに関わる疑問について、ゲノム全体ベースで取り組むことができるであろう。SNP 研究により、民族地理学的に異なる起源をもつ被験者に存在していたハプロタイプの範囲を確立することができ、ひいては民族の歴史と移住パターンに洞察を加えることもできる。こうした研究から、近代人の系統がアフリカに端を発していることが示唆されているが、ヒトの起源に関する多くの重大な疑問にはまだ答えが出ていない。さらに、このような論争に決着をつけるには、詳細な SNP マップを用いたもっと多くの解析が必要とされるであろう。民族の増大・移住・混合をうかがわせる証拠を提供してくれることに加え、SNP は、特定の遺伝子に対して働く進化抑制の程度を量るマーカーとなりうるのである。配列の多様性が損なわれた部位を同定するのに、種内と種間における遺伝子ばらつきパターンの相関関係が特に有益であると判明するかもしれない。

SNP 密度の目立った不均一性が物語っているのは、多型に作用するさまざまな力が存在していることである。すなわち、SNP 密度の低い領域がちらほら見られるのは、突然変異率が低いため、あるいはごく一部の受容できる変異だけ受け入れているため、新たに生じた対立遺伝子に有利なように強力な淘汰が近年行なわれた結果、それに関連する変異が集団から「一掃」されたため、と考えられる [165]。遺伝子がランダムに漂流した結果現れる影響も、ゲノム全体にわたってさまざまである。Y 染色体の非組換え部は、ランダムな遺伝子漂流から最も強力な圧力をうけているが、これは常染色体と同じように約 4 分の 1 程度、集団内に Y 染色体が存在しており、それに応じて染色体 Y 上の多型レベルが低いためである。同様に、X 染色体は常染色体に比べ有効な集団サイズが小さく、塩基多様性も低い。しかし、1 つの常染色体だけでも、有害な突然変異の密度にばらつきがあるため、有効な集団サイズ

にもばらつきがある。有害な突然変異の密度が高い領域は、淘汰による排除率が大きくなり、有効な集団サイズがさらに小さくなるであろう[166]。その結果、そのような領域では、完璧な中立の SNP でさえも密度が下がると考えられる。*Drosophila* の SNP 密度と局所組換え率との関連については文献が豊富にあるが、同じような関連がヒトゲノムにおいてどれほど強力であるかを判定するのは、今後の重要な課題である。何故なら、疾病と関連する研究では、局所の SNP 密度を設計する上でこの関連が大きな影響を及ぼすからである。地理的・民族的集団内に不均一性がどの程度あるかを判定するために、ゲノムスケールで SNP を確認することもまた、今後なすべき重要な課題である。

8.4. ゲノムの複雑性

我々は程なく、このゲノム体系の個々の成分をカタログ化する場から離れ、「これはあれと結合する、だからこれとドッキングさせ、そうすれば複合体はそちらに動く」[167] という単純な考えを超え、ネットワークの揺れという刺激的な場へ、非線形の反応や閾値へ、そしてヒトの疾患で果たしている中心的な役割へと進んでいくことになろう。

その他の「パーツリスト」を列挙していけば、複雑な神経系を有する生体では、遺伝子数、ニューロン数、細胞の種類数は、構造や行動の複雑性をはかる簡便な物差しとは（どんなに簡便なものであっても）、相関しないことが明らかになる。相関すると期待されてもいない。これは、非線形と後成の領分なのである[168]。5 億 2000 万という普通のタコのニューロン数は、マウスの脳内のニューロン数を一桁超えている。マウスとヒトをゲノムのデータで比較し、哺乳類の比較神経解剖学を見てみると[169]、哺乳類で認められる形態学的・行動学的多様性が、同じような遺伝子のレパートリーや同じような神経解剖学的特性で支えられているのは、明らかである。例えば、ピグミーマーモセット（キヌザル、身長わずか 10cm、体重約 170 g）をチンパンジーと比較してみると、キヌザルの脳の容積はおよそ 1.5 cm^3 に過ぎず、チンパンジーの大きさから 2 桁少ない数値であり、ヒトより 3 桁少ないことがわかる。しかし、この 3 者の脳の神経解剖学的特徴は驚くほど同じで、小さなキヌザルの行動特性は、チンパンジーの行動特性と殆んど異ならない。ヒトとチンパンジーとでは、遺伝子の数、遺伝子の構造と機能、染色体とゲノムの組織、細胞の種類、神経解剖学的特徴は殆んど識別できないが、ヒトという系統を大脳皮質拡大と喉頭の発生へと促した発達上の変化が言語をもたらし、結果的には極めて独特のものにしてしまった、すなわち、基準の最も単純なもので比べても、行動という面ではヒトをより複雑にさせてしまったのである。

ニューロンの数、細胞の種類の数、あるいは遺伝子ないしゲノムサイズの数を単純に調べるだけでは、我々が認めている複雑さの違いを説明できない。それどころか、このような

大きな差をもたらしたのは、これらのセット内・セット間（すなわちニューロン同士、細胞同士、遺伝子（ないしゲノムサイズ）同士、あるいはニューロンと細胞、ニューロンと遺伝子(ゲノムサイズ)、細胞と遺伝子等）の相互作用である。さらに、全体のシステムに不均衡な影響を与える制御遺伝子ネットワークの「特殊例」が存在している可能性もある。我々は、ハエや線虫に比べて、ヒトゲノムではっきり増加している「調節遺伝子」の例をいくつか提示した。例えば、細胞外リガンドやそれらと同起源の受容体（wnt、frizzled、TGF‐ 、エフリン、コネキシンなど)、ならびに核調節因子（KRAB ファミリー、ホメオドメイン転写因子ファミリーなど）が含まれるが、そこでは数種の蛋白質が幅広い発生過程を制御している。こうした「複雑性」が何故生じたのかに対する回答は、おそらくこうした拡大遺伝子ファミリーの中に、ひいては古代の遺伝子や蛋白質、反応経路、細胞の調節制御における差の中にあるのであろう。

8.5. 単一の成分を超えて

アインシュタインの脳が *Drosophila* の脳より複雑である、と直感的に断定しても異を唱える者は殆んどいないであろうが、予測されたヒト蛋白質の組み合わせが *Drosophila* のそれより複雑であるかどうか、複雑であるのならどの程度、といったもっと厳密な比較をおこなうのは簡単ではない。蛋白や蛋白のドメイン、蛋白と蛋白の相互作用などを量る物差しは、表現型の根底にある動的機能を支えている「状況に応じた」相互作用の実態を把握してはいないからである。

現時点では、複雑性について述べられた数学的理論は 30 編を越える [(170)]。しかし、遺伝子の数と生体の複雑性を関連させて数学的理論で説明していくことはまだこれからである。さまざまな異なる成分（蛋白質、蛋白質複合体、相互作用する細胞系、相互作用するニューロン群）で構成される生体システムを解析するための実用的なアプローチの 1 つは、グラフ理論 [(171)] を用いることであろう。このシステムの各成分は、複雑なトポグラフィの交点で表すことができ、それらの相互作用はエッジ(辺)で表せられる。大きなネットワークを調べてみると、各ネットワークが自律的に組織化できることがわかるものの、それより重要なことは、各ネットワークがとりわけ強固になりうることである。この強固さは、成分の余剰に起因するのではなく、不均一に張り巡らされたネットワークが有する 1 つの性質と言えよう。こうしたネットワークのエラー寛容性には、代償を払わねばならない。各ネットワークは、ネットワークの安定性に不釣合いなほど寄与しているいくつかの結び目（交点）を取捨選択されることに、弱いのである。1 例として、遺伝子ノックアウトが挙げられる。僅少な影響しか及ぼさないノックアウトがある一方、劇的変化を組織体に及ぼすノックアウトもある。哺乳類の細胞質における中間フィラメントネットワークでおそらく欠かせない一員と思われるビメンチン（vimentin）を取り上げてみよう。マウスでこの遺伝子を

ノックアウトさせると、繁殖面では正常であり、表現型として現れる特徴にも影響はないが[172]、正常マウスで目立つビメンチンネットワークは完全に欠落している。一方、*Drosophila* とマウスでは、ノックアウトの約 30%が決定的な結び目に相当しており、遺伝子産物での減少、あるいは全滴によってネットワークそのものが時間の大半をつぶしてしまう。ただし、このような場合でも、適度の遺伝的背景があれば、表現型の正常性は保たれることがある。従って、「良い遺伝子」、「悪い遺伝子」が存在しているのではなく、さまざまなレベルで、さまざまな連携を持ち、混乱に対する感度がさまざまであるようなネットワークが存在しているだけなのである。精巧な数学的解析は、特にネットワークの動的機能に焦点をあてた確固たる生物学的データセットに対して、絶えず評価されなくてはいけない。“ 複雑性 ”を把握するための試みの中で、これ以上重大な箇所はない。というのも、とりわけ、混乱を受けてヒトに疾患を起こしてしまった複雑なネットワークを解きほぐし修正することこそ、今我々が直面している最大の有意義な挑戦的課題なのであるから。

ヒトゲノム全解析によって、ヒトの生物学的研究に対する新しい戦略が切り開かれるであろう、医学に対して、ひいては医療・公衆衛生を通じて、社会に対しても大きな影響がおよぶであろうと、この 15 年来予測されてきた。生物医学研究への影響は既に感じられている。ヒト生物学におけるゲノムの役割を理解すべく出発した長い刺激的な旅にあって、このようにヒトゲノム配列を組み立てることは、初めてのこととはいえ、踏み出しにくい一歩であった。これが実現したのは、ほかでもない、機器とソフトウェアに革新的なものが現れ、その結果、DNA 調製から注釈付けまでの過程のほぼ全ての段階で自動化が可能となったためである。次にとるべき行動は、あきらかである。すなわち、比較的中庸の数である約 3 万の遺伝子が発現される時、必ず生じる複雑性とは何かを明確に定義することである。今回提示した配列は、遺伝学、生化学、生理学、究極的には表現型に依存するもの全てを囲む枠組みとなる。科学的な疑問に答える最前線となるものである。ゲノムを理解するにあたっては、初期段階にすぎない。あらゆる遺伝子とそれらを制御するあらゆる因子を同定しなくてはいけない。これらの機能も、単独でも協調状態でも、確認されなくてはならない。世界中のさまざまな人種間の配列変異を記述し、ゲノムの変異と特定の表現型との繋がりも確定しなくてはいけない。今や我々は、何を説明しなくてはならないかがわかったのである。

もう 1 つ最重要の挑戦的課題が待機している。すなわち、今回のゲノム情報についてだけでなく、個人の健康を向上させるためにどのような可能性がゲノムにあるのか、を一般市民が議論することである。多種多様なデータ供給源から、どんな 2 人でも、99.9%以上同じ塩基配列を持っていることが判明した。このことは、我々ヒトという種では、個人間の遺伝子に起因しうる差は、どれほど栄光ある差であれ全て、解読された配列の 0.1%に過ぎないことを意味している。ここで、避けねばならない誤った考え方は 2 つある。決定主義と

還元主義である。前者は、個人の特性が全てゲノムによって“がっちりと繋がれている”とする考え方であり、後者は、ヒトゲノム配列に関する完璧な知識を持った今、遺伝子の機能と相互作用を我々が理解することによって、ヒトの多様性について完璧な因果関係を記載できるようになるのは時間の問題である、とする考え方である。ヒト生物学への真の挑戦は、遺伝子がどのような編成を組んで身体の驚くべき機構を構築し維持しているのかを見つけ出す仕事を超え、我々自身の存在を探究するために我々の精神がどれほど見事に考え方を組織化するようになったか、の説明を探し求めている我々の前に今や立ちはだかっている。

参考文献

1. R. L. Sinsheimer, *Genomics* **5**, 954 (1989) [Medline]; U.S. Department of Energy, Office of Health and Environmental Research, *Sequencing the Human Genome: Summary Report of the Santa Fe Workshop*, Santa Fe, NM, 3 to 4 March 1986 (Los Alamos National Laboratory, Los Alamos, NM, 1986).
2. R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1996).
3. F. Sanger, *et al*., *Nature* **265**, 687 (1977) [Medline].
4. P. H. Seeburg, *et al*., *Trans. Assoc. Am. Physicians* **90**, 109 (1977) [Medline].
5. E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* **154**, 353 (1986) [Medline].
6. J. Gocayne, *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8296 (1987) [Medline].
7. A. Martin-Gallardo, *et al*., *DNA Sequence* **3**, 237 (1992) ; W. R. McCombie, *et al*., *Nature Genet.* **1**, 348 (1992) [Medline]; M. A. Jensen , *et al*., *DNA Sequence* **1**, 233 (1991) .
8. M. D. Adams, *et al*., *Science* **252**, 1651 (1991) [Medline].
9. M. D. Adams, *et al*., *Nature* **355**, 632 (1992) [Medline]; M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 256 (1993) [Medline]; M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 373 (1993) [Medline]; M. H. Polymeropoulos, *et al*., *Nature Genet.* **4**, 381 (1993) [Medline]; M. Marra, *et al*., *Nature Genet.* **21**, 191 (1999) [Medline].
10. M. D. Adams, *et al*., *Nature* **377**, 3 (1995) [Medline]; O. White, *et al*., *Nucleic Acids Res.* **21**, 3829 (1993) [Abstract].
11. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 729 (1982) [Medline].
12. B. W. J. Mahy, J. J. Esposito, J. C. Venter, *Am. Soc. Microbiol. News* **57**, 577 (1991) .
13. R. D. Fleischmann, *et al*., *Science* **269**, 496 (1995) [Medline].
14. C. M. Fraser, *et al*., *Science* **270**, 397 (1995) [Abstract].
15. C. J. Bult, *et al*., *Science* **273**, 1058 (1996) [Abstract]; J. F. Tomb, *et al*., *Nature* **388**, 539 (1997) [Medline]; H. P. Klenk, *et al*., *Nature* **390**, 364 (1997) [Medline].
16. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996) [Medline].
17. H. Schmitt, *et al*., *Genomics* **33**, 9 (1996) [Medline].
18. S. Zhao, *et al*., *Genomics* **63**, 321 (2000) [Medline].
19. X. Lin, *et al*., *Nature* **402**, 761 (1999) [Medline].
20. J. L. Weber and E. W. Myers, *Genome Res.* **7**, 401 (1997) [Full Text].
21. P. Green, *Genome Res.* **7**, 410 (1997) [Full Text].
22. E. Pennisi, *Science* **280**, 1185 (1998) [Full Text].
23. J. C. Venter, *et al*., *Science* **280**, 1540 (1998) [Full Text].
24. M. D. Adams, *et al*., *Nature* **368**, 474 (1994) [Medline].
25. E. Marshall and E. Pennisi, *Science* **280**, 994 (1998) [Full Text].
26. M. D. Adams, *et al*., *Science* **287**, 2185 (2000) [Abstract/Full Text].
27. G. M. Rubin, *et al*., *Science* **287**, 2204 (2000) [Abstract/Full Text].
28. E. W. Myers, *et al*., *Science* **287**, 2196 (2000) [Abstract/Full Text].
29. F. S. Collins, *et al*., *Science* **282**, 682 (1998) [Abstract/Full Text].
30. International Human Genome Sequencing Consortium (2001), *Nature* **409**, 860 (2001).
31. Institutional review board: P. Calabresi (chairman), H. P. Freeman, C. McCarthy, A. L. Caplan, G. D. Rogell, J. Karp, M. K. Evans, B. Margus, C. L. Carter, R. A. Millman, S. Broder.
32. Eligibility criteria for participation in the study were as follows: prospective donors had to be 21 years of age or older, not pregnant, and capable of giving an informed consent. Donors were asked to self-define their ethnic backgrounds. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed on all samples at the clinical laboratory prior to DNA extraction in the Celera laboratory. All samples that tested positive for transmissible viruses were ineligible and were discarded. Karyotype analysis was performed on peripheral blood lymphocytes from all samples selected for sequencing; all were normal. A two-staged consent process for prospective donors was employed. The first stage of the

consent process provided information about the genome project, procedures, and risks and benefits of participating. The second stage of the consent process involved answering follow-up questions and signing consent forms, and was conducted about 48 hours after the first.

33. DNA was isolated from blood (173) or sperm. For sperm, a washed pellet (100 µl) was lysed in a suspension (1 ml) containing 0.1 M NaCl, 10 mM tris-Cl-20 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was extracted with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to Bst XI adapters (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3'-CACA overhangs, were inserted into Bst XI-linearized plasmid vector with 3'-TGTG overhangs. Libraries with three different average sizes of inserts were constructed: 2, 10, and 50 kbp. The 2-kbp fragments were cloned in a high-copy pUC18 derivative. The 10- and 50-kbp fragments were cloned in a medium-copy pBR322 derivative. The 2- and 10-kbp libraries yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and inserts were unstable. To remedy this, the 50-kbp libraries were digested with Bgl II, which does not cleave the vector, but generally cleaved several times within the 50-kbp insert. A 1264-bp Bam HI kanamycin resistance cassette (purified from pUCK4; Amersham Pharmacia, catalog no. 27-4958-01) was added and ligation was carried out at 37°C in the continual presence of Bgl II. As Bgl II-Bgl II ligations occurred, they were continually cleaved, whereas Bam HI-Bgl II ligations were not cleaved. A high yield of internally deleted circular library molecules was obtained in which the residual insert ends were separated by the kanamycin cassette DNA. The internally deleted libraries, when plated on agar containing ampicillin (50 µg/ml), carbenicillin (50 µg/ml), and kanamycin (15 µg/ml), produced relatively uniform large colonies. The resulting clones could be prepared for sequencing using the same procedures as clones from the 10-kbp libraries.

34. Transformed cells were plated on agar diffusion plates prepared with a fresh top layer containing no antibiotic poured on top of a previously set bottom layer containing excess antibiotic, to achieve the correct final concentration. This method of plating permitted the cells to develop antibiotic resistance before being exposed to antibiotic without the potential clone bias that can be introduced through liquid outgrowth protocols. After colonies had grown, QBot (Genetix, UK) automated colony-picking robots were used to pick colonies meeting stringent size and shape criteria and to inoculate 384-well microtiter plates containing liquid growth medium. Liquid cultures were incubated overnight, with shaking, and were scored for growth before passing to template preparation. Template DNA was extracted from liquid bacterial culture using a procedure based upon the alkaline lysis miniprep method (173) adapted for high throughput processing in 384-well microtiter plates. Bacterial cells were lysed; cell debris was removed by centrifugation; and plasmid DNA was recovered by isopropanol precipitation and resuspended in 10 mM tris-HCl buffer. Reagent dispensing operations were accomplished using Titertek MAP 8 liquid dispensing systems. Plate-to-plate liquid transfers were performed using Tomtec Quadra 384 Model 320 pipetting robots. All plates were tracked throughout processing by unique plate barcodes. Mated sequencing reads from opposite ends of each clone insert were obtained by preparing two 384-well cycle sequencing reaction plates from each plate of plasmid template DNA using ABI-PRISM BigDye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Sequencing reactions were prepared using the Tomtec Quadra 384-320 pipetting robot. Parent-child plate relationships and, by extension, forward-reverse sequence mate pairs were established by automated plate barcode reading by the onboard barcode reader and were recorded by direct LIMS communication. Sequencing reaction products were purified by alcohol precipitation and were dried, sealed, and stored at 4°C in the dark until needed for sequencing, at which time the reaction products were resuspended in deionized formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sample sheets were created at load time using a Java-based application that facilitates barcode scanning of the sequencing plate barcode, retrieves sample information from the central LIMS, and reserves unique trace identifiers. The application permitted a single sample sheet file in the linking directory and deleted previously created sample sheet files immediately upon scanning of a sample plate barcode, thus enhancing sample sheet-to-plate associations.

35. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977) [Medline]; J. M. Prober , *et al*., *Science* **238**, 336 (1987) [Medline].

36. Celera's computing environment is based on Compaq Computer Corporation's Alpha system technology running the Tru64 Unix operating system. Celera uses these Alphas as Data Servers and as nodes in a Virtual Compute Farm, all of which are connected to a fully switched network operating at Fast Ethernet speed (for the VCF) and gigabit Ethernet speed (for data servers). Load balancing and scheduling software manages the submission and execution of jobs, based on central processing unit (CPU) speed, memory requirements, and priority. The Virtual Compute Farm is composed of 440 Alpha CPUs, which includes model EV6 running at a clock speed of 400 MHz and EV67 running at 667 MHz. Available memory on these systems ranges from 2 GB to 8 GB. The VCF is used to manage trace file processing, and annotation. Genome assembly was performed on a GS 160 running 16 EV67s (667 MHz) and 64 GB of memory, and 10 ES40s running 4 EV6s (500 MHz) and 32 GB of memory. A total of 100 terabytes of physical disk storage was included in a Storage Area Network that was available to systems across the environment. To ensure high availability, file and database servers were configured as 4-node Alpha TruClusters, so that services would fail over in the event of hardware or software failure. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping with parity (RAID-5).

37. Trace processing generates quality values for base calls by means of Paracel's TraceTuner, trims sequence reads according to quality values, trims vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in design and algorithm to the phred program (174), TraceTuner reports quality values that reflect the log-odds score of each base being correct. Read quality was evaluated in 50-bp windows, each read being trimmed to include only those consecutive 50-bp segments with a minimum mean accuracy of 97%. End windows (both ends of the trace) of 1, 5, 10, 25, and 50 bases were trimmed to a minimum mean accuracy of 98%. Every read was further checked for vector and contaminant matches of 50 bp or more, and if found, the read was removed from consideration. Finally, any match to the 5' vector splice junction in the initial part of a read was removed.

38. National Center for Biotechnology Information (NCBI); available at www.ncbi.nlm.nih.gov/.

39. NCBI; available at www.ncbi.nlm.nih.gov/HTGS/.

40. All bactigs over 3 kbp were examined for coverage by Celera mate pairs. An interval of a bactig was deemed an assembly error where there were no mate pairs spanning the interval and at least two reads that should have their mate on the other side of the interval but did not. In other words, there was no mate pair evidence supporting a join in the breakpoint interval and at least two mate pairs contradicting the join. By this criterion, we detected and broke apart bactigs at 13,037 locations, or equivalently, we found 2.13% of the bactigs to be misassembled.

41. We considered a BAC entry to be chimeric if, by the Lander-Waterman statistic (175), the odds were 0.99 or more that the assembly we produced was inconsistent with the sequence coming from a single source. By this criterion, 714 or 2.2% of BAC entries were deemed chimeric.

42. G. Myers, S. Selznick, Z. Zhang, W. Miller, *J. Comput. Biol.* **3**, 563 (1996) [Medline].

43. E. W. Myers, J. L. Weber, in *Computational Methods in Genome Research*, S. Suhai, Ed. (Plenum, New York, 1996), pp. 73-89.

44. P. Deloukas *et al.*, *Science* **282**, 744 (1998).

45. M. A. Marra *et al.*, *Genome Res.* **7**, 1072 (1997).

46. J. Zhang *et al.*, data not shown.

47. Shredded bactigs were located on long CSA scaffolds (>500 kbp) and the distribution of these fragments on the scaffolds was analyzed. If the spread of these fragments was greater than four times the reported BAC length, the BAC was considered to be chimeric. In addition, if >20% of bactigs of a given BAC were found on a different scaffolds that were not adjacent in map position, then the BAC was also considered as chimeric. The total chimeric BACs divided by the number of BACs used for CSA gave the minimal estimate of chimerism rate.

48. M. Hattori, *et al.*, *Nature* **405**, 311 (2000) [Medline].

49. I. Dunham, *et al.*, *Nature* **402**, 489 (1999) [Medline].

50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13239 (2000) [Abstract/Full Text].

51. The International RH Mapping Consortium, available at www.ncbi.nlm.nih.gov/genemap99/.

52. See http://ftp.genome.washington.edu/RM/RepeatMasker.html.

53. G. D. Schuler, *Trends Biotechnol.* **16**, 456 (1998) [Medline].

54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990) [Medline].

55a. M. Olivier, *et al.*, *Science* **291**, 1298 (2001) . 55b. See http://genome.ucsc.edu/.

56. N. Chaudhari and W. E. Hahn, *Science* **220**, 924 (1983) [Medline]; R. J. Milner and J. G. Sutcliffe, *Nucleic Acids Res.* **11**, 5497 (1983) [Abstract].

57. D. Dickson, *Nature* **401**, 311 (1999) [Medline].

58. B. Ewing and P. Green, *Nature Genet.* **25**, 232 (2000) [Medline].

59. H. Roest Crollius, *et al.*, *Nature Genet.* **25**, 235 (2000) [Medline].

60. M. Yandell, in preparation.

61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* **16**, 44 (2000) [Medline].

62. Scaffolds containing greater than 10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds greater than one megabase, the sequence was cut into single megabase pieces before computational analysis. All sequence was masked for complex repeats using Repeatmasker (52) before gene finding or homology-based analysis. The computational pipeline required ~7 hours of CPU time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the NCBI. Nucleotide searches were performed against human, mouse, and rat Celera Gene Indices (assemblies of cDNA and EST sequences), mouse genomic DNA reads generated at Celera (3×), the Ensembl gene database available at the European Bioinformatics Institute (EBI), human and

rodent (mouse and rat) EST data sets parsed from the dbEST database (NCBI), and a curated subset of the RefSeq experimental mRNA database (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (54) optimized for the Compaq Alpha compute-server and an effective database size of 3×10^9 for BLASTN searches and 1×10^9 for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of $<1 \times 10^{-4}$, human nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >94% identity, and rodent nucleotide BLAST results having an expectation score of $<1 \times 10^{8}$ with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *C. elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs, and with Lap for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (63).

63. E. C. Uberbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* **266**, 259 (1996) [Medline]; C. Burge and S. Karlin, *J. Mol. Biol.* **268**, 78 (1997) [Medline]; R. J. Mural, *Methods Enzymol.* **303**, 77 (1999) [Medline]; A. A. Salamov and V. V. Solovyev, *Genome Res.* **10**, 516 (2000) [Abstract/Full Text]; Floreal *et al.*, *Genome Res.* **8**, 967 (1998).
64. G. L. Miklos and B. John, *Am. J. Hum. Genet.* **31**, 264 (1979) [Medline]; U. Francke, *Cytogenet. Cell Genet.* **65**, 206 (1994) [Medline].
65. P. E. Warburton, H. F. Willard, in *Human Genome Evolution*, M. S. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121-145.
66. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* **10**, 839 (2000) [Abstract/Full Text].
67. W. A. Bickmore and A. T. Sumner, *Trends Genet.* **5**, 144 (1989) [Medline].
68. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992) [Medline].
69. G. Bernardi, *Gene* **241**, 3 (2000) [Medline].
70. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996) [Medline].
71. S. Ohno, *Trends Genet.* **1**, 160 (1985) .
72. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am. J. Hum. Genet.* **63**, 861 (1998) [Medline].
73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* **34**, 331 (2000) [Abstract/Full Text].
74. A. Bird, *Trends Genet.* **3**, 342 (1987) .
75. M. Gardiner-Garden and M. Frommer, *J. Mol. Biol.* **196**, 261 (1987) [Medline].
76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992) [Medline].
77. S. H. Cross and A. Bird, *Curr. Opin. Genet. Dev.* **5**, 309 (1995) [Medline].
78. J. Peters, *Genome Biol.* **1**, reviews1028.1 (2000) (http://genomebiology.com/2000/1/5/reviews/1028).
79. C. Grunau, W. Hindermann, A. Rosenthal, *Hum. Mol. Genet.* **9**, 2651 (2000) [Abstract/Full Text].
80. F. Antequera and A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995 (1993) [Abstract].
81. S. H. Cross, *et al.*, *Mamm. Genome* **11**, 373 (2000) [Medline].
82. D. Slavov, *et al.*, *Gene* **247**, 215 (2000) [Medline].
83. A. F. Smit and A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995) [Abstract].
84. D. J. Elliott, *et al.*, *Hum. Mol. Genet.* **9**, 2117 (2000) [Abstract/Full Text].
85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* **274**, 24849 (1999) [Abstract/Full Text].
86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* **59**, 282 (1999) [Medline].
87. P. Nouvel, *Genetica* **93**, 191 (1994) [Medline].
88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* **10**, 672 (2000) [Abstract/Full Text].
89. Lek first compares all proteins in the proteome to one another. Next, the resulting BLAST reports are parsed, and a graph is created wherein each protein constitutes a node; any hit between two proteins with an expectation beneath a user-specified threshold constitutes an edge. Lek then uses this graph to compute a similarity between each protein pair *ij* in the context of the graph as a whole by simply dividing the number of BLAST hits shared in common between the two proteins by the total number of proteins hit by *i* and *j*. This simple metric has several interesting properties. First, because the similarity metric takes into account both the similarity and the differences between the two sequences at the level of BLAST hits, the metric

respects the multidomain nature of protein space. Two multidomain proteins, for instance, each containing domains A and B, will have a greater pairwise similarity to each other than either one will have to a protein containing only A or B domains, so long as A-B-containing multidomain proteins are less frequent in the proteome than are single-domain proteins containing A or B domains. A second interesting property of this similarity metric is that it can be used to produce a similarity matrix for the proteome as a whole without having to first produce a multiple alignment for each protein family, an error-prone and very time-consuming process. Finally, the metric does not require that either sequence have significant homology to the other in order to have a defined similarity to each other, only that they share at least one significant BLAST hit in common. This is an especially interesting property of the metric, because it allows the rapid recovery of protein families from the proteome for which no multiple alignment is possible, thus providing a computational basis for the extension of protein homology searches beyond those of current HMM- and profile-based search methods. Once the whole-proteome similarity matrix has been calculated, Lek first partitions the proteome into single-linkage clusters (27) on the basis of one or more shared BLAST hits between two sequences. Next, these single-linkage clusters are further partitioned into subclusters, each member of which shares a user-specified pairwise similarity with the other members of the cluster, as described above. For the purposes of this publication, we have focused on the analysis of single-linkage clusters and what we have termed "complete clusters," e.g., those subclusters for which every member has a similarity metric of 1 to every other member of the subcluster. We believe that the single-linkage and complete clusters are of special interest, in part, because they allow us to estimate and to compare sizes of core protein sets in a rigorous manner. The rationale for this is as follows: if one imagines for a moment a perfect clustering algorithm capable of perfectly partitioning one or more perfectly annotated protein sets into protein families, it is reasonable to assume that the number of clusters will always be greater than, or equal to, the number of single-linkage clusters, because single-linkage clustering is a maximally agglomerative clustering method. Thus, if there exists a single protein in the predicted protein set containing domains A and B, then it will be clustered by single linkage together with all single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the number of real clusters must always be less than or equal to the number of complete clusters, because it is impossible to place a unique multidomain protein into a complete cluster. Thus, the single-linkage and complete clusters plus singletons should comprise a lower and upper bound of sizes of core protein sets, respectively, allowing us to compare the relative size and complexity of different organisms' predicted protein set.

90. T. F. Smith and M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981) [Medline].

91. A. L. Delcher, *et al*., *Nucleic Acids Res.* **27**, 2369 (1999) [Medline].

92. *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).

93. The probability that a contiguous set of proteins is the result of a segmental duplication can be estimated approximately as follows. Given that protein A and B occur on one chromosome, and that A' and B' (paralogs of A and B) also exist in the genome, the probability that B' occurs immediately after A' is $1/N$, where N is the number of proteins in the set (for this analysis, N = 26,588). Allowing for B' to occur as any of the next J-1 proteins [leaving a gap between A' and B' increases the probability to $(J - 1)/N$; allowing B'A' or A'B' gives a probability of $2(J - 1)/N$]. Considering three genes ABC, the probability of observing A'B'C' elsewhere in the genome, given that the paralogs exist, is $1/N^2$. Three proteins can occur across a spread of five positions in six ways; more generally, we compute the number of ways that K proteins can be spread across J positions by counting all possible arrangements of $K - 2$ proteins in the $J - 2$ positions between the first and last protein. Allowing for a spread to vary from K positions (no gaps) to J gives

$$L = \sum_{X=K-2}^{J-2} \binom{X}{K-2}$$

arrangements. Thus, the probability of chance occurrence is L/N^{K-1}. Allowing for both sets of genes (e.g., ABC and A'B'C') to be spread across J positions increases this to L^2/N^{K-1}. The duplicated segment might be rearranged by the operations of reversal or translocation; allowing for M such rearrangements gives us a probability $P = L^2M/N^{K-1}$. For example, the probability of observing a duplicated set of three genes in two different locations, where the three genes occur across a spread of five positions in both locations, is $36/N^2$; the expected number of such matched sets in the predicted protein set is approximately $(N)36/N^2 = 36/N$, a value $\ll 1$. Therefore, any such duplications of three genes are unlikely to result from random rearrangements of the genome. If any of the genes occur in more than two copies, the probability that the apparent duplication has occurred by chance increases. The algorithm for selecting candidate duplications only generates matched protein sets with $P \ll 1$.

94. B. J. Trask, *et al*., *Hum. Mol. Genet.* **7**, 13 (1998) [Abstract/Full Text]; D. Sharon, *et al*., *Genomics* **61**, 24 (1999) [Medline].

95. W. B. Barbazuk, *et al*., *Genome Res.* **10**, 1351 (2000) [Abstract/Full Text]; A. McLysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* **17**, 22 (2000) [Medline]; D. W. Burt, *et al*., *Nature* **402**, 411 (1999) [Medline].

96. Reviewed in L. Skrabanek and K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998) [Medline].

97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* **8**, 748 (1998) [Abstract/Full Text]; P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, *Genome Res.* **9**, 499 (1999) [Abstract/Full Text].

98. D. Altshuler, *et al*., *Nature* **407**, 513 (2000) [Medline].

99. G. T. Marth, *et al*., *Nature Genet.* **23**, 452 (1999) [Medline].

100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).

101. M. Cargill, *et al*., *Nature Genet.* **22**, 231 (1999) [Medline].

102. M. K. Halushka, *et al*., *Nature Genet.* **22**, 239 (1999) [Medline].

103. J. Zhang and T. L. Madden, *Genome Res.* **7**, 649 (1997) [Abstract/Full Text].

104. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).

105. From the observed coverage of the sequences at each site for each individual, we calculated the probability that a SNP would be detected at the site if it were present. For each level of coverage, there is a binomial sampling of the two homologs for each individual, and a heterozygous site could only be ascertained if both homologs are present, or if two alleles from different individuals are present. With coverage x from a given individual, both homologs are present in the assembly with probability $1 - (1/2)x^{-1}$. Even if both homologs are present, the probability that a SNP is detected is <1 because a fraction of sites failed the quality criteria. Integrating over coverage levels, the binomial sampling, and the quality distribution, we derived an expected number of sites in the genome that were ascertained for polymorphism for each individual. The nucleotide diversity was then the observed number of variable sites divided by the expected number of sites ascertained.

106. M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, *Genetics* **150**, 1133 (1998) [Abstract/Full Text].

107. D. A. Nickerson *et al.*, *Nature Genet.* **19**, 233 (1998); D. A. Nickerson, *et al*., *Genomic Res.* **10**, 1532 (2000) ; L. Jorde, *et al*., *Am. J. Hum. Genet.* **66**, 979 (2000) [Medline]; D. G. Wang, *et al*., *Science* **280**, 1077 (1998) [Abstract/Full Text].

108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000) [Medline].

109. S. Tavare, *Theor. Popul. Biol.* **26**, 119 (1984) [Medline].

110. R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, D. J. Futuyma, J. D. Antonovics, Eds. (Oxford Univ. Press, Oxford, 1990), vol. 7, pp. 1-44.

111. A. G. Clark, *et al*., *Am. J. Hum. Genet.* **63**, 595 (1998) [Medline].

112. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).

113. H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, *Nature Genet.* **22**, 78 (1999) [Medline].

114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997) [Medline].

115. A. Bateman, *et al*., *Nucleic Acids Res.* **28**, 263 (2000) [Abstract/Full Text].

116. Brief description of the methods used to build the Panther classification. First, the June 2000 release of the GenBank NR protein database (excluding sequences annotated as fragments or mutants) was partitioned into clusters using BLASTP. For the clustering, a seed sequence was randomly chosen, and the cluster was defined as all sequences matching the seed to statistical significance (E-value < 10^{-5}) and "globally" alignable (the length of the match region must be >70% and <130% of the length of the seed). If the cluster had more than five members, and at least one from a multicellular eukaryote, the cluster was extended. For the extension step, a hidden Markov Model (HMM) was trained for the cluster, using the SAM software package, version 2. The HMM was then scored against GenBank NR (excluding mutants but including fragments for this step), and all sequences scoring better than a specific (NLL-NULL) score were added to the cluster. The HMM was then retrained (with fixed model length) and all sequences in the cluster were aligned to the HMM to produce a multiple sequence alignment. This alignment was assessed by a number of quality measures. If the alignment failed the quality check, the initial cluster was rebuilt around the seed using a more restrictive E-value, followed by extension, alignment, and reassessment. This process was repeated until the alignment quality was good. The multiple alignment and "general" (i.e., describing the entire cluster, or "family") HMM (176) were then used as input into the BETE program (177). BETE calculates a phylogenetic tree for the sequences in the alignment. Functional information about the sequences in each cluster were parsed from SwissProt (178) and GenBank records. "Tree-attribute viewer" software was used by biologist curators to correlate the phylogenetic tree with protein function. Subfamilies were manually defined on the basis of shared function across subtrees, and were named accordingly. HMMs were then built for each subfamily, using information from both the subfamily and family (K. Sjölander, in preparation). Families were also manually named according to the functions contained within them. Finally, all of the families and subfamilies were classified into categories and subcategories based on their molecular functions. The categorization was done by manual review of the family and subfamily names, by examining SwissProt and GenBank records, and by review of the literature as well as resources on the World Wide Web. The current version (2.0) of the Panther molecular function schema has four levels: category, subcategory, family, and subfamily. Protein sequences for whole eukaryotic genomes (for the predicted human proteins and annotated proteins for fly, worm, yeast, and *Arabidopsis*) were scored against the Panther library of family and subfamily HMMs. If the score was significant (the NLL-NULL score cutoff depends on the protein family), the protein was assigned to the family or subfamily function with the most significant score.

117. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, *Nucleic Acids Res.* **27**, 229 (1999) [Abstract/Full Text].

118. A. Goffeau *et al.*, *Science* **274**, 546, 563 (1996).

119. *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).

120. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).

121. E. R. Kandel, J. H. Schwartz, T. Jessell, *Principles of Neural Science* (McGraw-Hill, New York, ed. 4, 2000).

122. D. A. Goodenough, J. A. Goliger, D. L. Paul, *Annu. Rev. Biochem.* **65**, 475 (1996) [Abstract].

123. D. G. Wilkinson, *Int. Rev. Cytol.* **196**, 177 (2000) [Medline].

124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* **44**, 219 (2000) [Medline].

125. P. J. Horner and F. H. Gage, *Nature* **407**, 963 (2000) [Medline]; P. Casaccia-Bonnefil, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* **468**, 275 (1999) [Medline].

126. S. Wang and B. A. Barres, *Neuron* **27**, 197 (2000) [Medline].

127. M. Geppert and T. C. Sudhof, *Annu. Rev. Neurosci.* **21**, 75 (1998) [Abstract/Full Text]; J. T. Littleton and H. J. Bellen, *Trends Neurosci.* **18**, 177 (1995) [Medline].

128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* **274**, 24453 (1999) [Abstract/Full Text].

129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3666 (2000).

130. G. Lemke, *Glia* **7**, 263 (1993) [Medline].

131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* **68**, 729 (1999).

132. N. Perrimon and M. Bernfield, *Nature* **404**, 725 (2000) [Medline].

133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* **273**, 24979 (1998) [Full Text].

134. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).

135. T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, *J. Biol. Chem.* **274**, 25555 (1999) [Abstract/Full Text].

136. R. A. Black and J. M. White, *Curr. Opin. Cell Biol.* **10**, 654 (1998) [Medline].

137. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* **24**, 47 (1999) [Medline].

138. A. G. Uren *et al.*, *Mol. Cell* **6**, 961 (2000).

139. P. Garcia-Meunier, M. Etienne-Julan, P. Fort, M. Piechaczyk, F. Bonhomme, *Mamm. Genome* **4**, 695 (1993) [Medline].

140. K. Meyer-Siegler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8460 (1991).

141. N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, *Nucleic Acids Res.* **21**, 993 (1993) [Abstract].

142. N. A. Tatton, *Exp. Neurol.* **166**, 29 (2000) [Medline].

143. N. Kenmochi, *et al.*, *Genome Res.* **8**, 509 (1998) [Abstract/Full Text].

144. F. W. Chen and Y. A. Ioannou, *Int. Rev. Immunol.* **18**, 429 (1999) [Medline].

145. H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, *Nucleic Acids Res.* **18**, 1513 (1990) [Abstract].

146. D. M. Chambers, J. Peters, C. M. Abbott, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4463 (1998) [Abstract/Full Text]; A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, *Dev. Dyn.* **216**, 267 (1999) [Medline].

147. D. Aeschlimann and V. Thomazy, *Connect. Tissue Res.* **41**, 1 (2000) [Medline].

148. P. Munroe, *et al.*, *Nature Genet.* **21**, 142 (1999) [Medline]; S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, *Science* **254**, 1634 (1991) [Medline]; B. Furie, *et al.*, *Blood* **93**, 1798 (1999) [Full Text].

149. J. W. Kehoe and C. R. Bertozzi, *Chem. Biol.* **7**, R57 (2000) .

150. T. Pawson and P. Nash, *Genes Dev.* **14**, 1027 (2000) [Full Text].

151. A. W. van der Velden and A. A. Thomas, *Int. J. Biochem. Cell Biol.* **31**, 87 (1999) [Medline].

152. C. M. Fraser, *et al.*, *Science* **281**, 375 (1998) [Abstract/Full Text]; H. Tettelin, *et al.*, *Science* **287**, 1809 (2000) [Abstract/Full Text].

153. D. Brett, *et al.*, *FEBS Lett.* **474**, 83 (2000) [Medline].

154. H. J. Muller and H. Kern, *Z. Naturforsch. B* **22**, 1330 (1967) [Medline].

155. H. J. Muller, in *Heritage from Mendel*, R. A. Brink, Ed. (Univ. of Wisconsin Press, Madison, WI, 1967), p. 419.

156. J. F. Crow, M. Kimura, *Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).

157. K. Kobayashi *et al.*, *Nature* **394**, 388 (1998).

158. A. P. Feinberg, *Curr. Top. Microbiol. Immunol.* **249**, 87 (2000) [Medline].

159. C. A. Collins and C. Guthrie, *Nature Struct. Biol.* **7**, 850 (2000) [Medline].

160. S. R. Eddy, *Curr. Opin. Genet. Dev.* **9**, 695 (1999) [Medline].

161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* **290**, 1765 (2000) [Abstract/Full Text].

162. M. Holcik, N. Sonenberg, R. G. Korneluk, *Trends Genet.* **16**, 469 (2000) [Medline].

163. T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, *Nature* **408**, 106 (2000) [Medline].

164. E. Capanna and M. G. M. Romanini, *Caryologia* **24**, 471 (1971) .

165. J. Maynard Smith, *J. Theor. Biol.* **128**, 247 (1987) [Medline].

166. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* **141**, 1619 (1995) [Abstract].

167. J. E. Bailey, *Nature Biotechnol.* **17**, 616 (1999) .

168. R. Maleszka, H. G. de Couet, G. L. Miklos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3731 (1998) [Abstract/Full Text].

169. G. L. Miklos, *J. Neurobiol.* **24**, 842 (1993) [Medline].

170. J. P. Crutchfield and K. Young, *Phys. Rev. Lett.* **63**, 105 (1989) ; M. Gell-Mann and S. Lloyd, *Complexity* **2**, 44 (1996) .

171. A. L. Barabasi and R. Albert, *Science* **286**, 509 (1999) [Abstract/Full Text].

172. E. Colucci-Guyon, *et al*., *Cell* **79**, 679 (1994) [Medline].

173. J. Sambrook, E. F. Fritch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).

174. B. Ewing and P. Green, *Genome Res.* **8**, 186 (1998) [Abstract/Full Text]; B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998) [Abstract/Full Text].

175. E. S. Lander and M. S. Waterman, *Genomics* **2**, 231 (1988) [Medline].

176. A. Krogh *et al*., *J. Mol. Biol.* **235**, 1501 (1994) [Medline].

177. K. Sjölander, *Proc. Intell. Syst. Mol. Biol.* **6**, 165 (1998) .

178. A. Bairoch and R. Apweiler, *Nucleic Acids Res.* **28**, 45 (2000) [Abstract/Full Text].

179. GO, available at www.geneontology.org/.

180. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33 (2000) [Abstract/Full Text].

181. We thank E. Eichler and J. L. Goldstein for many helpful discussions and critical reading of the manuscript, and A. Caplan for advice and encouragement. We also thank T. Hein, D. Lucas, G. Edwards, L. Foster, D. Bhandari, P. Davies, T. Safford, J. Schira, and the Celera IT staff for outstanding computational support. The cost of this project was underwritten by the Celera Genomics Group of the Applera Corporation. We thank the Board of Directors of Applera Corporation: J. F. Abely Jr. (retired), R. H. Ayers, J.-L. Bélingard, R. H. Hayes, A. J. Levine, T. E. Martin, C. W. Slayman, O. R. Smith, G. C. St. Laurent Jr., and J. R. Tobin for their vision, enthusiasm, and unwavering support and T. L. White for leadership and advice. Data availability: The genome sequence and additional supporting information are available to academic scientists at the Web site (www.celera.com). Instructions for obtaining a DVD of the genome sequence can be obtained through the Web site. For commercial scientists wishing to verify the results presented here, the genome data are available upon signing a Material Transfer Agreement, which can also be found on the Web site.

5 December 2000; accepted 19 January 2001
10.1126/science.1058040
Include this information when citing this paper.

This article has been cited by other articles:

- Hazbun, T. R., Fields, S. (2001). Networking proteins in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 98: 4277-4278 [Full Text]
- Olivier, M., Aggarwal, A., Allen, J., Almendras, A. A., Bajorek, E. S., Beasley, E. M., Brady, S. D., Bushard, J. M., Bustos, V. I., Chu, A., Chung, T. R., Witte, A. D., Denys, M. E., Dominguez, R., Fang, N. Y., Foster, B. D., Freudenberg, R. W., Hadley, D., Hamilton, L. R., Jeffrey, T. J., Kelly, L., Lazzeroni, L., Levy, M. R., Lewis, S. C., Liu, X., Lopez, F. J., Louie, B., Marquis, J. P., Martinez, R. A., Matsuura, M. K., Misherghi, N. S., Norton, J. A., Olshen, A., Perkins, S. M., Perou, A. J., Piercy, C., Piercy, M., Qin, F., Reif, T., Sheppard, K., Shokoohi, V., Smick, G. A., Sun, W.-L., Stewart, E. A., Fernando, J., Tejeda, , Tran, N. M., Trejo, T., Vo, N. T., Yan, S. C. M., Zierten, D. L., Zhao, S., Sachidanandam, R., Trask, B. J., Myers, R. M., Cox, D. R. (2001). A High-Resolution Radiation Hybrid Map of the Human Genome Draft Sequence. *Science* 291: 1298-1302 [Abstract] [Full Text]

Volume 291, Number 5507, Issue of 16 Feb 2001, pp. 1304-1351.