types present in subjects of different ethnogeographic origins, providing insights into population history and migration patterns. Although such studies have suggested that modern human lineages derive from Africa, many important questions regarding human origins remain unanswered, and more analyses using detailed SNP maps will be needed to settle these controversies. In addition to providing evidence for population expansions, migration, and admixture, SNPs can serve as markers for the extent of evolutionary constraint acting on particular genes. The correlation between patterns of intraspecies and interspecies genetic variation may prove to be especially informative to identify sites of reduced genetic diversity that may mark loci where sequence variations are not tolerated.

The remarkable heterogeneity in SNP density implies that there are a variety of forces acting on polymorphism—sparse regions may have lower SNP density because the mutation rate is lower, because most of those regions have a lower fraction of mutations that are tolerated, or because recent strong selection in favor of a newly arisen allele "swept" the linked variation out of the population (*165*). The effect of random genetic drift also varies widely across the genome. The nonrecombining portion of the Y chromosome faces the strongest pressure from random drift because there are roughly one-quarter as many Y chromosomes in the population as there are autosomal chromosomes, and the level of polymorphism on the Y is correspondingly less. Similarly, the X chromosome has a smaller effective population size than the autosomes, and its nucleotide diversity is also reduced. But even across a single autosome, the effective population size can vary because the density of deleterious mutations may vary. Regions of high density of deleterious mutations will see a greater rate of elimination by selection, and the effective population size will be smaller (*166*). As a result, the density of even completely neutral SNPs will be lower in such regions. There is a large literature on the association between SNP density and local recombination rates in *Drosophila*, and it remains an important task to assess the strength of this association in the human genome, because of its impact on the design of local SNP densities for disease-association studies. It also remains an important task to validate SNPs on a genomic scale in order to assess the degree of heterogeneity among geographic and ethnic populations.

8.4 Genome complexity

We will soon be in a position to move away from the cataloging of individual components of the system, and beyond the simplistic notions of "this binds to that, which then docks on this, and then the complex moves there. . . ." (*167*) to the exciting area of network perturbations, nonlinear responses and thresholds, and their pivotal role in human diseases.

The enumeration of other "parts lists" reveals that in organisms with complex nervous systems, neither gene number, neuron number, nor number of cell types correlates in any meaningful manner with even simplistic measures of structural or behavioral complexity. Nor would they be expected to; this is the realm of nonlinearities and epigenesis (*168*). The 520 million neurons of the common octopus exceed the neuronal number in the brain of a mouse by an order of magnitude. It is apparent from a comparison of genomic data on the mouse and human, and from comparative mammalian neuroanatomy (*169*), that the morphological and behavioral diversity found in mammals is underpinned by a similar gene repertoire and similar neuroanatomies. For example, when one compares a pygmy marmoset (which is only 4 inches tall and weighs about 6 ounces) to a chimpanzee, the brain volume of this minute primate is found to be only about 1.5 cm^3, two orders of magnitude less than that of a chimp and three orders less than that of humans. Yet the neuroanatomies of all three brains are strikingly similar, and the behavioral characteristics of the pygmy marmoset are little different from those of chimpanzees. Between humans and chimpanzees, the gene number, gene structures and functions, chromosomal and genomic organizations, and cell types and neuroanatomies are almost indistinguishable, yet the developmental modifications that predisposed human lineages to cortical expansion and development of the larynx, giving rise to language, culminated in a massive singularity that by even the simplest of criteria made humans more complex in a behavioral sense.

Simple examination of the number of neurons, cell types, or genes or of the genome size does not alone account for the differences in complexity that we observe. Rather, it is the interactions within and among these sets that result in such great variation. In addition, it is possible that there are "special cases" of regulatory gene networks that have a disproportionate effect on the overall system. We have presented several examples of "regulatory genes" that are significantly increased in the human genome compared with the fly and worm. These include extracellular ligands and their cognate receptors (e.g., wnt, frizzled, TGF-β, ephrins, and connexins), as well as nuclear regulators (e.g., the KRAB and homeodomain transcription factor families), where a few proteins control broad developmental processes. The answers to these "complexities" perhaps lie in these expanded gene families and differences in the regulatory control of ancient genes, proteins, pathways, and cells.

8.5 Beyond single components

While few would disagree with the intuitive conclusion that Einstein's brain was more complex than that of *Drosophila*, closer comparisons such as whether the set of predicted human proteins is more complex than the protein set of *Drosophila*, and if so, to what degree, are not straightforward, since protein, protein domain, or protein-protein interaction measures do not capture context-dependent interactions that underpin the dynamics underlying phenotype.

Currently, there are more than 30 different mathematical descriptions of complexity (*170*). However, we have yet to understand the mathematical dependency relating the number of genes with organism complexity. One pragmatic approach to the analysis of biological systems, which are composed of nonidentical elements (proteins, protein complexes, interacting cell types, and interacting neuronal populations), is through graph theory (*171*). The elements of the system can be represented by the vertices of complex topographies, with the edges representing the interactions between them. Examination of large networks reveals that they can self-organize, but more important, they can be particularly robust. This robustness is not due to redundancy, but is a property of inhomogeneously wired networks. The error tolerance of such networks comes with a price; they are vulnerable to the selection or removal of a few nodes that contribute disproportionately to network stability. Gene knockouts provide an illustration. Some knockouts may have minor effects, whereas others have catastrophic effects on the system. In the case of vimentin, a supposedly critical component of the cytoplasmic intermediate filament network of mammals, the knockout of the gene in mice reveals them to be reproductively normal, with no obvious phenotypic effects (*172*), and yet the usually conspicuous vimentin network is completely absent. On the other hand, ~30% of knockouts in *Drosophila* and mice correspond to critical nodes whose reduction in gene product, or total elimination, causes the network to crash most of the time, although even in some of these cases, phenotypic normalcy ensues, given the appropriate genetic background. Thus, there are no "good" genes or "bad" genes, but only networks that exist at various levels and at different connectivities, and at different states of sensitivity to perturbation. Sophisticated mathematical analysis needs to be constantly evaluated against hard biological data sets that specifically address network dynamics. Nowhere is this more critical than in attempts to come to grips with "complexity," particularly because deconvoluting and correcting complex networks that have undergone perturbation, and have resulted in human diseases, is the greatest significant challenge now facing us.

It has been predicted for the last 15 years that complete sequencing of the human genome would open up new strategies for human biological research and would have a major impact on medicine, and through medicine and public health, on society. Effects on biomedical research are already being felt. This assembly of the human genome sequence is but a first, hesitant step on a long and exciting journey toward understanding the role of the genome in human biology. It has been possible only because of innovations in instrumentation and software that have allowed automation of almost every step of the process from DNA preparation to annotation. The next steps are clear: We must define the complexity that ensues when this relatively modest set of about 30,000 genes is expressed. The sequence provides the framework upon which all the genetics, biochemistry, physiology, and ultimately phenotype depend. It provides the boundaries for scientific inquiry. The sequence is only the first level of understanding of the genome. All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain.

Another paramount challenge awaits: public discussion of this information and its potential for improvement of personal health. Many diverse sources of data have shown that any two individuals are more than 99.9% identical in sequence, which means that all the glorious differences among individuals in our species that can be attributed to genes falls in a mere 0.1% of the sequence. There are two fallacies to be avoided: determinism, the idea that all characteristics of the person are "hard-wired" by the genome; and reductionism, the view that with complete knowledge of the human genome sequence, it is only a matter of time before our understanding of gene functions and interactions will provide a complete causal description of human variability. The real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.

References and Notes

1. R. L. Sinsheimer, *Genomics* **5**, 954 (1989); 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).
4. P. H. Seeburg *et al.*, *Trans. Assoc. Am. Physicians* **90**, 109 (1977).
5. E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* **154**, 353 (1986).
6. J. Gocayne *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8296 (1987).
7. A. Martin-Gallardo *et al.*, *DNA Sequence* **3**, 237 (1992); W. R. McCombie *et al.*, *Nature Genet.* **1**, 348 (1992); M. A. Jensen *et al.*, *DNA Sequence* **1**, 233 (1991).
8. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
9. M. D. Adams *et al.*, *Nature* **355**, 632 (1992); M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 256 (1993); M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 373 (1993); M. H. Polymeropoulos *et al.*, *Nature Genet.* **4**, 381 (1993); M. Marra *et al.*, *Nature Genet.* **21**, 191 (1999).
10. M. D. Adams *et al.*, *Nature* **377**, 3 (1995); O. White *et al.*, *Nucleic Acids Res.* **21**, 3829 (1993).
11. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 729 (1982).
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).
14. C. M. Fraser *et al.*, *Science* **270**, 397 (1995).
15. C. J. Bult *et al.*, *Science* **273**, 1058 (1996); J. F. Tomb *et al.*, *Nature* **388**, 539 (1997); H. P. Klenk *et al.*, *Nature* **390**, 364 (1997).
16. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
17. H. Schmitt *et al.*, *Genomics* **33**, 9 (1996).
18. S. Zhao *et al.*, *Genomics* **63**, 321 (2000).
19. X. Lin *et al.*, *Nature* **402**, 761 (1999).
20. J. L. Weber, E. W. Myers, *Genome Res.* **7**, 401 (1997).
21. P. Green, *Genome Res.* **7**, 410 (1997).
22. E. Pennisi, *Science* **280**, 1185 (1998).
23. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
24. M. D. Adams *et al.*, *Nature* **368**, 474 (1994).
25. E. Marshall, E. Pennisi, *Science* **280**, 994 (1998).
26. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
27. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
28. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
29. F. S. Collins *et al.*, *Science* **282**, 682 (1998).
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); J. M. Prober *et al.*, *Science* **238**, 336 (1987).
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).
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).
49. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13239 (2000).
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).
54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).

55a. M. Olivier *et al.*, *Science* **291**, 1298 (2001).

55b. See http://genome.ucsc.edu/.

56. N. Chaudhari, W. E. Hahn, *Science* **220**, 924 (1983); R. J. Milner, J. G. Sutcliffe, *Nucleic Acids Res.* **11**, 5497 (1983).
57. D. Dickson, *Nature* **401**, 311 (1999).
58. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).
59. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
60. M. Yandell, in preparation.
61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* **16**, 44 (2000).
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 $\sim$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); C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997); R. J. Mural, *Methods Enzymol.* **303**, 77 (1999); A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000); Floreal *et al.*, *Genome Res.* **8**, 967 (1998).
64. G. L. Miklos, B. John, *Am. J. Hum. Genet.* **31**, 264 (1979); U. Francke, *Cytogenet. Cell Genet.* **65**, 206 (1994).
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).
67. W. A. Bickmore, A. T. Sumner, *Trends Genet.* **5**, 144 (1989).
68. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992).
69. G. Bernardi, *Gene* **241**, 3 (2000).
70. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996).
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).
73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* **34**, 331 (2000).
74. A. Bird, *Trends Genet.* **3**, 342 (1987).
75. M. Gardiner-Garden, M. Frommer, *J. Mol. Biol.* **196**, 261 (1987).
76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992).
77. S. H. Cross, A. Bird, *Curr. Opin. Genet. Dev.* **5**, 309 (1995).
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).
80. F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995 (1993).
81. S. H. Cross *et al.*, *Mamm. Genome* **11**, 373 (2000).
82. D. Slavov *et al.*, *Gene* **247**, 215 (2000).
83. A. F. Smit, A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995).
84. D. J. Elliott *et al.*, *Hum. Mol. Genet.* **9**, 2117 (2000).
85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* **274**, 24849 (1999).
86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* **59**, 282 (1999).
87. P. Nouvel, *Genetica* **93**, 191 (1994).
88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* **10**, 672 (2000).
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, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
91. A. L. Delcher *et al.*, *Nucleic Acids Res.* **27**, 2369 (1999).
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); D. Sharon *et al.*, *Genomics* **61**, 24 (1999).
95. W. B. Barbazuk *et al.*, *Genome Res.* **10**, 1351 (2000); A. McLysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* **17**, 22 (2000); D. W. Burt *et al.*, *Nature* **402**, 411 (1999).
96. Reviewed in L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998).
97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* **8**, 748 (1998); P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, *Genome Res.* **9**, 499 (1999).
98. D. Altshuler *et al.*, *Nature* **407**, 513 (2000).
99. G. T. Marth *et al.*, *Nature Genet.* **23**, 452 (1999).
100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
101. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
102. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
103. J. Zhang, T. L. Madden, *Genome Res.* **7**, 649 (1997).
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).
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); D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000).
109. S. Tavare, *Theor. Popul. Biol.* **26**, 119 (1984).
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).
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).
114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).
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).
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).
123. D. G. Wilkinson, *Int. Rev. Cytol.* **196**, 177 (2000).
124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* **44**, 219 (2000).
125. P. J. Horner, F. H. Gage, *Nature* **407**, 963 (2000); P. Casaccia-Bonnefil, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* **468**, 275 (1999).
126. S. Wang, B. A. Barres, *Neuron* **27**, 197 (2000).
127. M. Geppert, T. C. Sudhof, *Annu. Rev. Neurosci.* **21**, 75 (1998); J. T. Littleton, H. J. Bellen, *Trends Neurosci.* **18**, 177 (1995).
128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* **274**, 24453 (1999).
129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3666 (2000).
130. G. Lemke, *Glia* **7**, 263 (1993).
131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* **68**, 729 (1999).
132. N. Perrimon, M. Bernfield, *Nature* **404**, 725 (2000).
133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* **273**, 24979 (1998).
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).

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

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

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).

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).

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

143. N. Kenmochi *et al.*, *Genome Res.* **8**, 509 (1998).

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

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

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

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

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

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

150. T. Pawson, P. Nash, *Genes Dev.* **14**, 1027 (2000).

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

152. C. M. Fraser *et al.*, *Science* **281**, 375 (1998); H. Tettelin *et al.*, *Science* **287**, 1809 (2000).

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

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

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).

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

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

161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* **290**, 1765 (2000).

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

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

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

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

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

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).

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

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

171. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).

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

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

174. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).

175. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).

176. A. Krogh, K. Sjölander, *J. Mol. Biol.* **235**, 1501 (1994).

177. K. Sjölander, *Proc. Int. Soc. Mol. Biol.* **6**, 165 (1998).

178. A. Bairoch, R. Apweiler, *Nucleic Acids Res.* **28**, 45 (2000).

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).

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, 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