types present in subjects of different ethnographic 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...” 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 cm³, 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 of the genome 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 homogeneously 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 reproducibly 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 ge-
The Human Genome

18. Z. Zhao et al., Genomics 63, 321 (2000).
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 giving an informed consent. Donors were asked to self-define their ethnic background. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed. Liquid transfers were performed 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. Kapferer’s polymerase chain reaction for perypheral 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 concerned 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, 0.2 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was precipitated with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, fragments were selected, end- polished with consequent BAL31 nuclease and T4 DNA polymerase treatment, and size-selected by electrophoresis on 1% low-melt-point agarose. After ligation of adapters, XI adapters (Invitrogen, catalog no. N8095-1), DNA polymerase treatments, and size-selected by electrophoresis to remove excess adapters, and the fragments, now with 3–C2A 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 yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and insertions 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 pUCV; American Pharm.; 5′-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 sequenced using the same procedures as from the 10-kbp libraries.
34. Transformed cells were plated on agar diffusion plates prepared with a agar plate 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 denaturation 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 a clone 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 reconstituted in denatured formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sequence chromatograms were treated 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 presents a detailed view of the linking directory and deleted previously created sample sheet files immediately upon scanning of a
The Human Genome

37. Trace processing generates quality values for base calls by means of Paracel’s TraceTrimmer, trim sequence reads according to quality values, trim vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in tor and adapter sequence from high-quality reads, and screens sequences for contaminants. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping (RAID-5).


40. All bactigs over 3 kb were examined for coverage by a set of non-overlapping 50-bp segments. Each segment 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. The Virtual Compute Farm is composed of 440 nodes in a Virtual Compute Farm, all of which are used for CSA gave the minimal estimate of chimerism.


64. P. Novuel, Genetica 93, 191 (1994).


68. The similarity 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
The Human Genome

share at least one significant BLAST hit in common. This is an especially interesting property of the method 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 (a) based 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 one or more 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 classification, i.e., one with perfectly overlapping paralogous proteins that are perfectly partitioned into 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 other single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the members 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 number of 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.

93. The probability that a contiguous set of proteins is the result of a single duplication can be approximated 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 A and B occur 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 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^3. 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 spread across J positions with possible arrangements K proteins in the J – 2 positions between the first and last protein. Allowing for a spread to vary from 0 positions (no gaps) to J gives:

L = \sum_{k=0}^{J} \binom{J}{k} \frac{K!}{(K-k)!}

arrangements. Thus, the probability of chance occurrence is LN^-1. Allowing for both sets of genes (e.g., ABC and A-B-C) to be spread across J positions increases this to LN^-1. The duplicated segment might be rearranged by the operations of reversion or translocation; allowing for M such rearrangements gives us a probability P = LNM^-1. For example, the probability of observing a duplicated set of three genes in different two locations, where the three genes occur with 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 (a) based 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 one or more 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 classification, i.e., one with perfectly overlapping paralogous proteins that are perfectly partitioned into 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 other 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 number of 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.

93. The probability that a contiguous set of proteins is the result of a single duplication can be approximated 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 A and B occur 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 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^3. 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 spread across J positions with possible arrangements K proteins in the J – 2 positions between the first and last protein. Allowing for a spread to vary from 0 positions (no gaps) to J gives:

L = \sum_{k=0}^{J} \binom{J}{k} \frac{K!}{(K-k)!}

arrangements. Thus, the probability of chance occurrence is LN^-1. Allowing for both sets of genes (e.g., ABC and A-B-C) to be spread across J positions increases this to LN^-1. The duplicated segment might be rearranged by the operations of reversion or translocation; allowing for M such rearrangements gives us a probability P = LNM^-1. For example, the probability of observing a duplicated set of three genes in different two locations, where the three genes occur with
The Human Genome

27. B. Ewing, P. Green, Genome Res. 8, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, Genome Res. 8, 175 (1998).
35. A. 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. Abey 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.
36. 5 December 2000; accepted 19 January 2001