**Integrating Genomes**

D. R. Zerbino *et al.*

*Science* **336**, 179 (2012);
DOI: 10.1126/science.1216830

complexity but rather moved nimbly within quadrants of the two-dimensional (2D) modeling space charted by two orthogonal axes characterizing model scope and level of realism (Fig. 1B).

A look at this space illustrates that to achieve qualitative insight, simple Boolean (*39*), Bayesian (*40*) (Box 1), ODE (*41*), or stochastic (*13*, *21*) models or physical estimates (*37*) that are focused and conceptual may include few details and make few predictions, but these predictions can be important. A model can be "bigger" and its scope more broad, but the level of realism can stay similar to that of the focused conceptual models. For example, broad and conceptual PDE (*42*), Boolean (*8*), or network (*43*) models can describe mathematically very large interacting systems but use only causal links between genes and/or proteins and so predict just qualitative features of emergent spatial-temporal patterns. However, focused models with relatively few mathematical details (*25*, *36*) can be accurate and mechanistic when precise numbers matter as well as qualitative insight. The broad and mechanistic models (*17*, *19*, *22*, *23*, *44*) are useful when there is a need to mathematically integrate detailed quantitative data and to test precisely formulated hypotheses.

### Outlook

Cell biology is transitioning into a quantitative science characterized by increasing integration of modeling into experiment. In this transition, we have to proceed with numerous, often arbitrary, assumptions about the nature of processes and parameter values governing cell systems. One great future challenge is to improve quantitative experimental methods with an eye toward synchronizing modeling and experiments. Then, frequent back-and-forth between theory and experiment using models of varying scope and level of realism will allow us to overcome the arbitrariness and uncertainty. Another significant challenge is to make switching from one type of model to another a more standard, less ad hoc procedure, to ease modeling use and integration between theory and experiment. Models along this course should be considered impermanent and should be judged by how useful they are and what we can learn from them, not by how close we are to the elusive whole-cell model.

### References and Notes


1. G. T. Reeves, S. E. Fraser, *PLoS Biol.* **7**, e21 (2009).
2. A. D. Lander, *Cell* **144**, 955 (2011).
3. J. M. G. Vilar, C. C. Guet, S. Leibler, *J. Cell Biol.* **161**, 471 (2003).
4. A. M. Turing, *Philos. Trans. R. Soc. Lond. B* **237**, 37 (1952).
5. A. L. Hodgkin, A. F. Huxley, *J. Physiol.* **117**, 500 (1952).
6. S. J. Vayttaden, S. M. Ajay, U. S. Bhalla, *ChemBioChem* **5**, 1365 (2004).
7. R. FitzHugh, *Biophys. J.* **1**, 445 (1961).
8. R. Albert, H. G. Othmer, *J. Theor. Biol.* **223**, 1 (2003).
9. W. J. Nelson, *Nature* **422**, 766 (2003).
10. J. M. Johnson, M. Jin, D. J. Lew, *Curr. Opin. Genet. Dev.* **21**, 740 (2011).
11. M. D. Onsum, C. V. Rao, *Curr. Opin. Cell Biol.* **21**, 74 (2009).
12. A. Jilkine, L. Edelstein-Keshet, *PLOS Comput. Biol.* **7**, e1001121 (2011).
13. R. Wedlich-Soldner, S. Altschuler, L. Wu, R. Li, *Science* **299**, 1231 (2003).
14. E. M. Ozbudak, A. Becskei, A. van Oudenaarden, *Dev. Cell* **9**, 565 (2005).
15. J. E. Irazoqui, A. S. Gladfelter, D. J. Lew, *Nat. Cell Biol.* **5**, 1062 (2003).
16. H. Meinhardt, *J. Cell Sci.* **112**, 2867 (1999).
17. E. Marco, R. Wedlich-Soldner, R. Li, S. J. Altschuler, L. F. Wu, *Cell* **129**, 411 (2007).
18. J. Valdez-Taubas, H. R. B. Pelham, *Curr. Biol.* **13**, 1636 (2003).
19. B. D. Slaughter, A. Das, J. W. Schwartz, B. Rubinstein, R. Li, *Dev. Cell* **17**, 823 (2009).
20. A. T. Layton *et al.*, *Curr. Biol.* **21**, 184 (2011).
21. S. J. Altschuler, S. B. Angenent, Y. Wang, L. F. Wu, *Nature* **454**, 886 (2008).
22. A. B. Goryachev, A. V. Pokhilko, *FEBS Lett.* **582**, 1437 (2008).
23. A. S. Howell *et al.*, *Cell* **139**, 731 (2009).
24. M. Fivaz, S. Bandara, T. Inoue, T. Meyer, *Curr. Biol.* **18**, 44 (2008).
25. D. Seetapun, D. J. Odde, *Curr. Biol.* **20**, 979 (2010).
26. N. Inagaki, M. Toriyama, Y. Sakumura, *Dev. Neurobiol.* **71**, 584 (2011).
27. N. W. Goehring *et al.*, *Science* **334**, 1137 (2011).
28. F. Tostevin, M. Howard, *Biophys. J.* **95**, 4512 (2008).
29. A. T. Dawes, E. M. Munro, *Biophys. J.* **101**, 1412 (2011).
30. M. M. Kozlov, A. Mogilner, *Biophys. J.* **93**, 3811 (2007).
31. F. Ziebert, S. Swaminathan, I. S. Aranson, *J. R. Soc. Interface* (2011).
32. D. Shao, W.-J. Rappel, H. Levine, *Phys. Rev. Lett.* **105**, 108104 (2010).
33. D. Kabaso, R. Shlomovitz, K. Schloen, T. Stradal, N. S. Gov, *PLOS Comput. Biol.* **7**, e1001127 (2011).
34. A. R. Houk *et al.*, *Cell* **148**, 175 (2012).
35. K. Sekimoto, J. Prost, F. Jülicher, H. Boukellal, A. Bernheim-Grosswasser, *Eur. Phys. J. E* **13**, 247 (2004).
36. M. J. Dayel *et al.*, *PLoS Biol.* **7**, e1000201 (2009).
37. J. van der Gucht, C. Sykes, *Cold Spring Harb. Perspect. Biol.* **1**, a001909 (2009).
38. J. S. Bois, F. Jülicher, S. W. Grill, *Phys. Rev. Lett.* **106**, 028103 (2011).
39. Y.-K. Kwon, K.-H. Cho, *Biophys. J.* **92**, 2975 (2007).
40. D. Mortimer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10296 (2009).
41. Z. Zheng, C.-S. Chou, T.-M. Yi, Q. Nie, *Math. Biosci. Eng.* **8**, 1135 (2011).
42. Y. Xiong, C.-H. Huang, P. A. Iglesias, P. N. Devreotes, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17079 (2010).
43. J. T. Gao *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7647 (2011).
44. A. F. M. Marée, A. Jilkine, A. Dawes, V. A. Grieneisen, L. Edelstein-Keshet, *Bull. Math. Biol.* **68**, 1169 (2006).


**Acknowledgments:** We thank R. Li and D. Lew for helpful discussions. This work was supported by NIH grant 2R01GM068952 and NSF grant DMS-1118206 to A.M.

10.1126/science.1216380



REVIEW

# Integrating Genomes

**D. R. Zerbino,[1] B. Paten,[1] D. Haussler[1,2]***

**As genomic sequencing projects attempt ever more ambitious integration of genetic, molecular, and phenotypic information, a specialization of genomics has emerged, embodied in the subdiscipline of computational genomics. Models inherited from population genetics, phylogenetics, and human disease genetics merge with those from graph theory, statistics, signal processing, and computer science to provide a rich quantitative foundation for genomics that can only be realized with the aid of a computer. Unleashed on a rapidly increasing sample of the planet's $10^{30}$ organisms, these analyses will have an impact on diverse fields of science while providing an extraordinary new window into the story of life.**

Since the first genome sequences were obtained in the mid-1970s (*1*, *2*), computers have been necessary for processing (*3*) and archiving (*2*, *4*) sequence data. However, the discipline of computational genomics traces its roots to 1980, when Smith and Waterman developed an algorithm to rapidly find the optimal comparison (alignment) of two sequences of length $n$ among the more than $3^n$ possibilities (*2*, *5*), and Stormo *et al.* built a linear threshold function to search a library of 78,000 nucleotides of *Escherichia coli* messenger RNA sequence for ribosome binding sites (*6*). What seemed large data sets for biology then don't seem so today, as high-throughput, short-read sequencing machines churn out terabytes of data (*2*, *7*). We have seen a 10,000-fold sequencing performance improvement in the past 8 years, far outpacing the estimated 16-fold improvement in computational power under Moore's law (*8*). Using genomics data to model genome evolution, mechanism, and function is now the heart of a lively field.

Every genome is the result of a mostly shared, but partly unique, 3.8-billion-year evolutionary journey from the origin of life. Diversity is created mostly by copy errors during replication. These create single-base changes, which are known as substitutions if spread to the whole population (fixed) or single-nucleotide polymorphisms (SNPs) if not uniformly present in the population (segregating). Replication errors also create insertions and deletions (collectively, indels), as well as tandem duplications where a short sequence is repeated sequentially. Chromosomes often exchange long similar segments through the process of homologous recombination. Specific sequences of DNA, known as transposable elements, have the

[1]Center for Biomolecular Sciences and Engineering, University of California, Santa Cruz, CA 95064, USA. [2]Howard Hughes Medical Institute, University of California, Santa Cruz, CA 95064, USA.

*To whom correspondence should be addressed. E-mail: haussler@soe.ucsc.edu

capacity to replicate themselves within the cell, using machinery analogous to that found in certain viruses, leaving many copies (*9*). Rearrangements lead to patterns such as inversions, segmental deletions and duplications (causing copy number variants), chromosome fusion and fission, and translocations between chromosomes (*10*). At the largest scale, occasionally the whole genome is duplicated, greatly increasing its gene content (*11*). The present diversity of life was created gradually through these edits and is manifest in the germline genotype of each living individual. Starting from the germline genotype, the genomes of the somatic cells continue to experience similar edits during the lifetime of every individual, some undergoing a kind of evolutionary process called somatic selection that plays a role in cancer and immunity (*2*, *12*).

Genomes are the core of the molecular mechanisms of cells, and of the physical properties (or phenotype) of organisms. They contain recipes of the active molecules of the cell, proteins, and their messenger RNAs, as well as other functional RNAs. Sequencing technology is used to determine RNA abundance (*13*), subcellular location (*14*), splicing isoforms (*15*), secondary structure (*16*), and rates of engagement with molecular complexes such as the ribosome (*17*). It is used to assay the epigenetic mechanisms that regulate RNA and protein production and function, including methylation (*2*, *18*), histone modifications (*19*), transcription factor binding (*20*), chromatin accessibility (*21*), and chromatin three-dimensional interactions (*2*, *22*). When applied to these data, computational genomics builds models of epigenetic mechanisms and gene regulatory networks (*2*, *23*), articulating with the broader models of molecular systems biology such as protein signaling cascades, metabolic pathways, and regulatory network motifs (*24*).

Combining evolutionary, mechanistic, and functional models, computational genomics interprets genomic data along three dimensions. A gene is simultaneously a DNA sequence evolving in time (history), a piece of chromatin that interacts with other molecules (mechanism), and, as a gene product, an actor in pathways of activity within the cell that affect the organism (function). Molecular phenotypes from epigenetic state and RNA expression levels are the first stations on the road from genotype to organismal phenotype, where evolutionary selection acts. Beyond the basics of storing, indexing, and searching the world's genomes, the three fundamental, interrelated challenges of computational genomics are to explain genome evolution, model molecular phenotypes as a consequence of genotype, and predict organismal phenotype.

### Obtaining Genomic Sequences

Current methods in genome analysis start with genome assembly (*2*), the process of reconstructing an entire genome from relatively short random DNA fragments, called reads. Given sufficient read redundancy, or coverage depth, it is possible to detect read overlaps and thereby progressively reconstitute most of the genome sequence (*2*). However, this ideal scenario is complicated by the fact that genomes commonly contain large redundant regions (repeats), or regions where the statistical distribution of bases is significantly biased (low-complexity DNA), leading to coincidental, spurious read overlaps. These create complex networks of read-to-read overlaps that do not all reflect actual overlaps in the genome. The most persistent difficulty of assembly is to determine which overlaps are legitimate and which are spurious. This problem is NP-hard, which means that it is at least as hard as any problem in the class of problems that can be solved in nondeterministic polynomial time (*2*). Therefore, we expect that the only efficient solutions will be heuristic methods that are not guaranteed to find the optimal solution. For this reason, difficult regions of genomes are left as undetermined gaps (*2*), prone to errors (*2*), or costly to finish (*2*). Newer sequencing technologies, producing longer reads (*2*), may alleviate this problem.

After the first complete genome from a species is assembled (the reference genome), new genomes from that species or closely related species are generally not assembled de novo but are assembled using the reference genome as a template, exploiting similarities derived from the common evolutionary ancestor. Reads from the new genome are mapped (aligned) onto the reference genome (*2*), and systematic discordances are detected (*2*). This process may be used simply to enumerate the variants present in the new genome, or to guide the complete assembly of the new genome (called reference-based assembly) (*2*). However, with short reads, mapping algorithms may also be confounded by repeats (*2*), and both mappers and reference-based assemblers tend to bias toward the reference genome, occasionally treating genuine variants as noise (*2*). As technology improves toward longer reads, we expect improvements here too.

### Modeling the Evolution of Genotype

Genomes are compared by alignment: The bases of DNA are partitioned into sets (columns) that are putatively derived from a suitably recent common ancestor. From this, we can analyze what was conserved or changed during the evolution of the genomes from their common ancestor. At a large scale, alignments can indicate changes in segment order and copy number, and at a small scale they can indicate specific base substitutions (see Fig. 1).

As in genome assembly, the primary challenge is to distinguish spurious sequence similarities from those due to common ancestry. Regions of genomes that are subject to purifying selection in which similarity of sequence is conserved, such as orthologous protein-coding regions, can often be reliably aligned across great evolutionary distances, such as between vertebrates and invertebrates. Regions that are neutrally drifting (i.e., not under positive or negative selection) diverge much more quickly, and can be reliably aligned only if they diverged recently (e.g., within the past 100 million years for two vertebrate genomes) (*2*). It is therefore common to distinguish alignments of subregions (local alignments) (*2*) from alignments of complete sequences (global alignments) (*2*) or even complete genomes (genome alignments) (*2*). Local alignments are typically used between conserved functional regions of more


**Fig. 1.** Assembly and alignment. (**A**) Assembly of a number of reads, grouped by pairwise sequence overlap. Because the genomic sequence contains a repeated sequence, the reads coming from the two copies of the repeat overlap and must be separated by the assembly software to produce a linear assembly. (**B**) Alignment of five sequences and an outgroup. Each row is a sequence; each column is a set of bases that descend from a common ancestral base. Six columns are highlighted. Column 4 contains a base that is fixed among the five sequences, whereas the other columns contain segregating SNPs. The trees on the sides represent two alternative ways of representing the phylogeny between the sequences. The left tree is optimal in terms of substitution complexity for columns 1, 2, and 3; the right tree is optimal for columns 4, 5, and 6. Given the difference between the two trees, a recombination event may have occurred between columns 3 and 5.

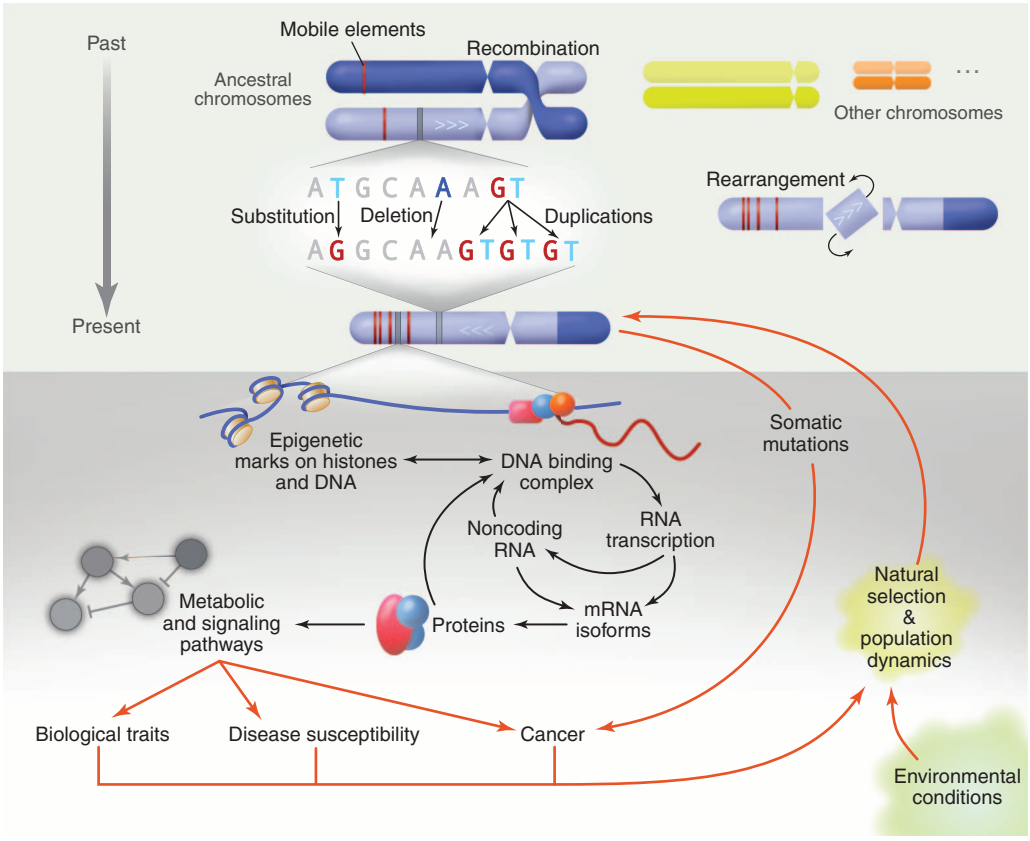


**Fig. 2.** The dynamic processes that affect and are affected by the genome. **Top**: The genome changes as it is modified by random mutations. At the larger scale, homologous recombination events swap equivalent pieces of DNA, rearrangements reconnect different regions of DNA, and transposable elements can self-reproduce. At the finer scale, small modifications such as substitutions and insertion/deletion events occur. **Bottom**: The genome affects the molecular processes in the cell, namely the transcription of genes and functional RNA, which through pathways affect the phenotype of the organism by causing phenotypes such as disease and other specific traits. Through natural selection, the phenotypes condition the selective pressure on the genome favoring or disfavoring specific mutations.

distantly related genomes (*2*). Conversely, full genome alignments become practical when comparing genomes from closely related species.

When applied to more than two species or to multiple gene copies within a species, phylogenetic methods provide an explicit order of gene descent through shared ancestry. When the model of evolution is restricted to consider only indels and substitutions (the most common events), the phylogeny is represented by a single tree in which the terminal (leaf) nodes represent the observed (present-day) sequences, the branches represent direct lines of descent, and the internal nodes represent the putative ancestral sequences (*2*). Finding the optimal phylogeny under probabilistic or parsimony models of substitutions (and also of indels) is NP-hard (*2*), and considerable effort has been devoted to obtaining efficient and accurate heuristic solutions.

Phylogenetic analysis is complicated by homologous recombination, which creates DNA molecules whose parts have different evolutionary histories (Fig. 2). The coalescent model with recombination (*2*) models the evolutionary history of a gene with both substitutions and homologous recombination. Individual histories of parts are represented in an alignment with a separate phylogenetic (coalescent) tree for each base (*2*).

Evolutionary relationships between DNA sequences may also include balanced structural rearrangements that change the order and the orientation of the bases in the genome, as well as segmental duplications, gains, and losses that alter the number of copies of homologous bases (*2*). Unfortunately, these processes are usually modeled and treated separately from one another, and separately from substitutions and short indels. The construction of a mathematically and algorithmically tractable unified theory of genome evolution, in which stochastic processes jointly describe base substitution, recombination, rearrangement, and the various forms of duplication, gain, and loss, remains a major challenge for the field (*2*). With incomplete knowledge of the mathematical difficulties inherent in such a model, it is hard to predict when, if ever, such a model will be forthcoming. The only thing we are assured of is that projects such as the 1000 Genomes Project (*2*) will be producing massive amounts of data from which to build and test (via likelihood methods) a variety of approximate models.

### From Genotype to Phenotype

Geneticists have correlated genomic mutations to phenotypic differences for many years, but today they do so at an unprecedented scale. Sequencing surveys across vertebrates [Genome 10K (*2*)], insects [i5K (*2*)], plants (*2*), microorganisms (*2*), cell lineages (*2*), and "metagenomes" (obtained by sequencing DNA from environmental samples containing an unknown collection of organisms) (*2*) present us with tens of thousands of genomes and challenge us to rework and deepen our methods. To date, such studies have given us concrete examples of the unfolding history and diversity of life, explored the ties between the body's microbial populations and our health, and investigated the response of species to current environmental changes such as climate shift, disease, and competitors (*2*). Future studies could be coupled with experimental data derived from an expansion of cell culture resources for diverse species and tissues (*2*) and newer single-cell assay methodologies (*2*), allowing deeper comparisons.

When studying the population genetics of a single species, the recombination rate determines how likely it is that proximal sequence variants share the same coalescent tree (*2*). Lack of recombination leads to linkage disequilibrium, in which nearby segregating variants are correlated. This phenomenon is exploited in correlating specific segregating variants with phenotypic traits or diseases—for example, in genome-wide association studies conducted with microarrays or incomplete sequencing data (*2*). However, this same phenomenon limits the resolution of these approaches in finding the actual causal variant. Genome-wide association studies are also blind to the patterns of allele segregation in close relatives. Future genotype-phenotype studies using complete genomes will increasingly use genotypic context in related as well as unrelated cases and controls, combined with better prediction of the possible effects of genome variants, to identify causal variants (*2*).

Large projects such as ENCODE (*2*), modENCODE (*2*), and the Epigenomics Roadmap (*2*) are providing data on the epigenome and the transcriptional machinery needed to construct models of molecular phenotypes involving epigenetic state, RNA expression, and (inferred) protein levels, requiring specialized analysis tools. Genome browsers such as Ensembl (*2*) and the UCSC Genome Browser (*2*) provide an integrated view

of these data, along with background knowledge and various modeling results. Because many key elements of epigenetics, RNA expression, and protein production cannot be directly measured and therefore must be inferred, the mathematical models of these processes contain numerous latent (hidden) variables, often one for every site in the genome. Approaches include hidden Markov models (*2*), factor graphs (*2*), Bayesian networks (*2*), and Markov random fields (*2*). Model inference (parameter estimation) and model application (computation of conditional and marginal probabilities) are large-scale computational tasks.

Genotype determines phenotype via epigenetic, transcriptional, and proteomic state. Classification and regression methods that are used to predict phenotype from genotype can take advantage of estimates of these intermediate states as additional or alternate inputs (*2*). These methods include general linear models, neural networks (*2*), and support vector machines (*2*), preferred in part because of their ability to cope with very high-dimensional input feature spaces (i.e., with many measured variables). There is currently more to be gained in predicting phenotype by incorporating biological knowledge to improve the input feature space—for example, by substituting inferred transcript levels or inferred protein activity levels for raw gene expression measurements (*2*)—than by using yet more sophisticated techniques of classification and regression.

## Looking Ahead to Applications

Understanding the shared evolutionary history of life starts by storing, indexing, and comparing genomes. It requires tools to rapidly produce evolutionarily related segments of DNA according to a model of genome evolution when prompted with a query segment. How will this be accomplished as we collectively grow from petabytes ($10^{15}$ bytes) of genome data today to exabytes ($10^{18}$ bytes) tomorrow? One possibility may be to use differential compression based on the inferred evolutionary trajectory of genomes, where each sequence is represented as a set of differences from its inferred parent (*2*). This may allow us to create a new web of genetic information that is compact, rapidly searchable, and directly reflects the natural origin of genomic relatedness.

Genomics has had a profound effect on medicine and will continue to do so. Cancer therapeutics are expected to advance as a result, because genomic modifications are the source of nearly all cancers (*2*). Within the body's somatic cells, genomic changes occur at random, from environmental impacts, or as a result of treatment; subpopulations of genetically distinct cancer cells expand and compete (*2*). Sequencing a sample of a cancer patient's noncancerous tissue reveals the patient's genome at birth (i.e., germline genome). Comparing this to the genome obtained from a tumor biopsy then reveals the mutations that have occurred subsequently in the patient's cancer cells. Tracking tumor genomes in this manner from early disease through each stage of treatment will become the norm and will inform therapeutic decisions (*2*). Changes that are readily detectable only through computational methods in genotype, epigenetic state, gene expression pattern, and activated pathway structure will provide crucial information on the state of the tumor during initial tumor growth and during the emergence of resistance to therapy (*2*). Recurring tumor-specific genomic variants and intermediate molecular phenotypes that drive cancer and determine patient response to therapy will come more clearly to light (*2*) and will be translated into better-targeted cancer diagnosis and treatment (*2*).

Other fields of medicine will also benefit from computational methods and findings. For example, immune cells undergo specific mutations through rounds of somatic selection (*2*), accompanied by changes in epigenetic state, gene expression pattern, and activated pathway structure. Deep sequencing of T cell receptors and B cell antibodies (*2*), coupled with genome-wide measurements of genetic variation, epigenetic state, and gene expression pattern in immune cells, will be used to model immune cell function and correlate immune response with antigen. High-throughput genomics data will be used in vaccine design (including cancer immunotherapy) and the treatment of infectious diseases (*2*), autoimmune diseases, and compromised immune systems resulting from chemotherapy, transfusions, transplants, and stem cell therapies (*2*).

Genomic variants, epigenetic state, and expression pattern play key roles in stem cell therapies and basic science applications of stem cells that can only be discerned through the use of computational tools (*2*). Induced pluripotent stem (iPS) cells and lineage-specific directly reprogrammed cells are made from somatic cells (*2*) that have already incurred somatic mutations and are cultured in conditions that may select for further mutations (*2*). These mutations will soon be assessed with whole-genome analysis. Measurements of epigenetic modification and gene expression will confirm the pluripotent or lineage-specific status of the reprogrammed cells and verify that the epigenetic memory of the tissue from which they were derived is erased. Because every batch of reprogrammed cells will show some unexpected genetic mutations, epigenetic changes, and expression differences on a genome-wide level, some with consequences, the interpretation of these data will be of critical importance. In summary, the future of research into cancer, immunology, and stem cells involves all three key challenges of computational genomics: explaining (somatic) evolution, modeling molecular phenotype, and predicting organismal phenotype.

In addition to other medical applications, similar scenarios are playing out in applications of genomics in a wide range of fields, such as agriculture (*2*) and the study of human prehistory (*2*). The increasing availability of data is leading to the development of elaborate multidimensional analysis tools incorporating DNA sequences, alignments, phylogenetic trees, lists of variants, epigenomic and functional assays, phenotypic changes, etc. To face the challenges of obtaining the maximum information from every sequencing experiment, we must borrow advances from a spectrum of different research fields and tie them together into foundational mathematical models implemented with numerical methods. There is a tension between the comprehensiveness of models and their computational efficiency. As this plays out, a comprehensive but computable model of genome evolution and its functional repercussions on organisms is taking shape, embodied in computational genomics. Yet we still await a formulation that is both simple and expressive enough to compare models, store information, and communicate results in an exabyte age. As a common language develops, shaped by our increasing knowledge of biology, we anticipate that computational genomics will provide enhanced ability to explore and exploit the genome structures and processes that lie at the heart of life.

### References and Notes


1. W. Fiers *et al.*, *Nature* **260**, 500 (1976).
2. For a full list of references by subject, see table S1 in the supplementary materials.
3. T. R. Gingeras, J. P. Milazzo, D. Sciaky, R. J. Roberts, *Nucleic Acids Res.* **7**, 529 (1979).
4. G. H. Hamm, G. N. Cameron, *Nucleic Acids Res.* **14**, 5 (1986).
5. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
6. G. D. Stormo, T. D. Schneider, L. Gold, A. Ehrenfeucht, *Nucleic Acids Res.* **10**, 2997 (1982).
7. E. R. Mardis, *Trends Genet.* **24**, 133 (2008).
8. L. D. Stein, *Genome Biol.* **11**, 207 (2010).
9. C. Feschotte, *Nat. Rev. Genet.* **9**, 397 (2008).
10. L. Feuk, A. R. Carson, S. W. Scherer, *Nat. Rev. Genet.* **7**, 85 (2006).
11. P. Dehal, J. L. Boore, *PLoS Biol.* **3**, e314 (2005).
12. L. M. F. Merlo, J. W. Pepper, B. J. Reid, C. C. Maley, *Nat. Rev. Cancer* **6**, 924 (2006).
13. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, *Nat. Methods* **5**, 621 (2008).
14. J. C. Simpson, R. Wellenreuther, A. Poustka, R. Pepperkok, S. Wiemann, *EMBO Rep.* **1**, 287 (2000).
15. C. Trapnell *et al.*, *Nature* **28**, 511 (2010).
16. J. G. Underwood *et al.*, *Nat. Methods* **7**, 995 (2010).
17. N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, *Science* **324**, 218 (2009).
18. A. L. Brunner *et al.*, *Genome Res.* **19**, 1044 (2009).
19. T. S. Mikkelsen *et al.*, *Nature* **448**, 553 (2007).
20. G. Robertson *et al.*, *Nat. Methods* **4**, 651 (2007).
21. A. P. Boyle *et al.*, *Cell* **132**, 311 (2008).
22. M. J. Fullwood *et al.*, *Nature* **462**, 58 (2009).
23. P. J. Mitchell, R. Tjian, *Science* **245**, 371 (1989).
24. U. Alon, *Nat. Rev. Genet.* **8**, 450 (2007).



**Acknowledgments:** We thank D. A. Earl for designing the figures, and E. Green, M. Häussler, J. Ma, D. Earl, H. Zerbino, R. Kuhn, G. Hickey, T. Pringle, K. Pollard, A. Krogh, R. Shamir, M. Waterman, and R. Durbin for their corrections and comments. Supported by the Howard Hughes Medical Institute (D.H.), National Human Genome Research Institute Data Analysis Center for the Encyclopedia of DNA Elements grant U01 (B.P.), and the American Association for Cancer Research (Stand Up To Cancer/An Integrated Approach to Targeting Breast Cancer Molecular Subtypes and Their Resistance Phenotypes) (D.R.Z.).


### Supplementary Materials

www.sciencemag.org/cgi/content/full/336/6078/179/DC1
Table S1

10.1126/science.1216830