HUMAN GENOME INFORMATICS

TRANSLATING GENES INTO HEALTH

EDITED BY

CHRISTOPHE G. LAMBERT

DARROL J. BAKER

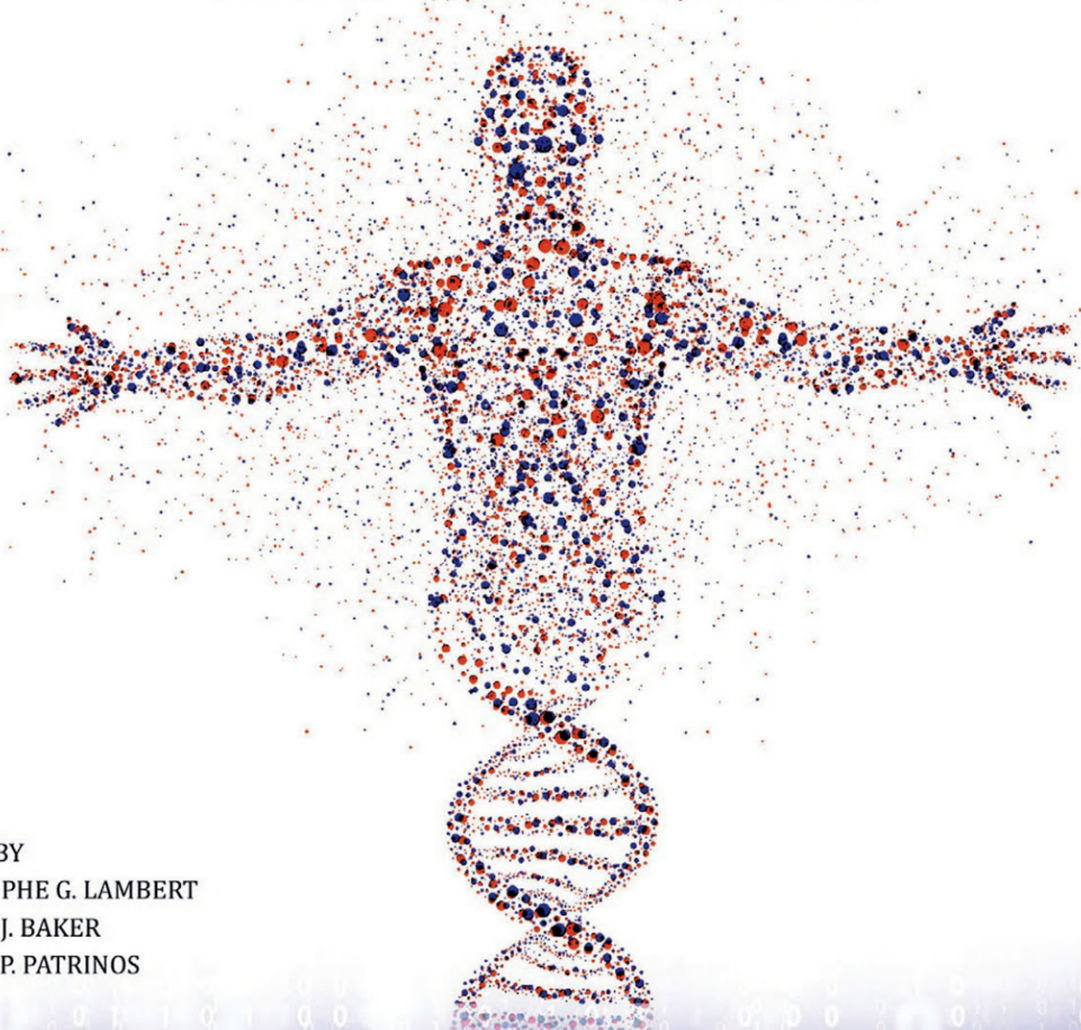GEORGE P. PATRINOS

Human Genome Informatics

Translational and Applied Genomics Series

Human Genome Informatics

Translating Genes into Health

Edited by

Christophe G. Lambert

Center for Global Health, Division of Translational Informatics,
Department of Internal Medicine, University of New Mexico
Health Sciences Center, Albuquerque, NM, United States

Darrol J. Baker

The Golden Helix Foundation, London, United Kingdom

George P. Patrinos

Department of Pharmacy, School of Health Sciences, University of Patras,
Patras, Greece

Department of Pathology, College of medicine and Health Sciences,
United Arab Emirates University, Al-Ain, United Arab Emirates

Department of Pathology—Bioinformatics Unit, Faculty of Medicine and Health
Sciences, Erasmus University Medical Center, Rotterdam, The Netherlands

ELSEVIER

ACADEMIC PRESS
An imprint of Elsevier

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

For information on all Academic Press publications visit our website
at https://www.elsevier.com/books-and-journals

Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Contributors

Darrol J. Baker   The Golden Helix Foundation, London, United Kingdom

Marina Bartsakoulia   Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece

Melek Chaouch   Institute of Pasteur of Tunisia, Tunis, Tunisia

Emilson Chimusa   University of Cape Town, Cape Town, South Africa

Alexandros Kanterakis   Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion; Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece

Theodora Katsila   Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece

Aggeliki Komianou   Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece

J. Kumuthini   Centre for Proteomic and Genomic Research, Cape Town, South Africa

Christophe G. Lambert   Center for Global Health, Division of Translational Informatics, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM, United States

Collen Masimiremwa   African Institute of Biomedical Science and Technology, Harare, Zimbabwe

George P. Patrinos   Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece; Department of Pathology, College of medicine and Health Sciences, United Arab Emirates University, Al-Ain, United Arab Emirates; Department of Pathology—Bioinformatics Unit, Faculty of Medicine and Health Sciences, Erasmus University Medical Center, Rotterdam, The Netherlands

George Potamias   Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion, Greece

Konstantinos Sarris   Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece

Morris A. Swertz   Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Giannis Tzimas   Department of Computer and Informatics Engineering, Technological Educational Institute of Western Greece, Patras, Greece

Emmanouil Viennas   University of Patras, Faculty of Engineering, Department of Computer Engineering and Informatics, Patras, Greece

L. Zass   Centre for Proteomic and Genomic Research, Cape Town, South Africa

Preface

We are delighted to offer this textbook to the scientific community, entitled Human Genome Informatics, which covers a timely topic in the rapidly evolving discipline of bioinformatics.

In the postgenomic era, the development of electronic tools to translate genomic information into a clinically meaningful format is of utmost importance to expedite the transition of genomic medicine into mainstream clinical practice. There are several well-established textbooks that cover the bioinformatics field, and there are also numerous protocols for bioinformatics analysis that one can retrieve from the Internet. However, the field of human genome informatics is a relatively new one that emerged in the postgenomic era, constituting a niche research discipline. As such, there are hardly any books that discuss this important new discipline, despite its broad implications for human health.

We therefore decided to deliver a textbook focused on human genome informatics, in order to first define the field and some of its history, and then provide an overview of the most commonly used electronic tools to analyze and interpret human genomic information into a clinically meaningful format, hence expediting the integration of genomic medicine into the mainstream clinical practice. At the same time, the book will provide an update on related topics, such as genomic data sharing, human genomic databases, and informatics tools in pharmacogenomics. To our knowledge, no other existing book deals exclusively with this topic.

We envision that this textbook will be of particular benefit to graduate and doctoral students, postdoctoral researchers in the field of genome informatics and bioinformatics, and representatives from bioinformatics companies and diagnostic laboratories interested in establishing such tools to translate/interpret the findings from their analyses. Also, this textbook will be ultimately useful as the main course material or supplementary reading in related graduate courses.

Our effort to compile most of the chapters included in this textbook has been assisted by many internationally renowned experts in their field, who have kindly accepted our invitation to share with us and our readers their expertise, experience, and results through contributed chapters. In addition, we made efforts to formulate the book contents using simple language and terminology, along with self-explanatory illustrations, in order that the book be useful not only to experienced professionals and academics, but also to undergraduate medical and life science students.

We are grateful to the publishing editors, Drs. Mariana Kuhl, Rafael Texeira, and Peter Linsley at Elsevier, who helped us in close collaboration to overcome encountered difficulties. We also express our gratitude to all contributors for delivering outstanding compilations that summarize their experience and many years of hard work in their field of research and to those colleagues who provided constructive comments and criticisms on the chapters. We are indebted to the copy editor, Jude Fernando, who has refined the final manuscript prior to letting it into production. Also, we owe special thanks to the academic reviewers for their constructive criticisms of the chapters and their positive evaluation of our proposal for this compilation.

We feel certain that some points in this textbook can be further improved. Therefore, we would welcome comments and criticism from attentive readers, which will contribute to improve the contents of this book even further in its future editions.

Christophe G. Lambert

Center for Global Health, Division of Translational Informatics, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM, United States

Darrol J. Baker

The Golden Helix Foundation, London, United Kingdom

George P. Patrinos

Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece; Department of Pathology, College of medicine and Health Sciences, United Arab Emirates University, Al-Ain, United Arab Emirates; Department of Pathology—Bioinformatics Unit, Faculty of Medicine and Health Sciences, Erasmus University Medical Center, Rotterdam, The Netherlands

Human Genome Informatics: Coming of Age

Christophe G. Lambert*, **Darrol J. Baker**[†], **George P. Patrinos**[‡,§,¶]

*Center for Global Health, Division of Translational Informatics, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM, United States, †The Golden Helix Foundation, London, United Kingdom, ‡Department of Pharmacy, School of Health Sciences, University of Patras, Patras, Greece, §Department of Pathology, College of medicine and Health Sciences, United Arab Emirates University, Al-Ain, United Arab Emirates, ¶Department of Pathology—Bioinformatics Unit, Faculty of Medicine and Health Sciences, Erasmus University Medical Center, Rotterdam, The Netherlands

1.1 INTRODUCTION

Human genome informatics is the application of information theory, including computer science and statistics, to the field of human genomics. Informatics enlists computation to augment our capacity to form models of reality with diverse sources of information. When forming a model of reality, one engages in a process of abstraction. The word "*abstraction*" comes from the Latin *abstrahere*, which means to "draw away," which is a metaphor, based in human vision, that as we back away from something, the details fall away and we form mental constructs about what we can discern from the more distant vantage point. That more distant vantage point both encompasses a greater portion of reality and yet holds in mind a smaller amount of detail about that larger space.

Given the human mind's limit on the number of variables it can manage, as we form our mental models of reality, we pay attention to certain facets of reality and ignore others, perhaps leaving them to subconscious or unconscious processing mechanisms. When we form models of reality, we have a field of perception that encompasses a subset of reality at a particular scale and a particular time horizon and that includes a subset of the variables at that spatio-temporal scale. Those variables are recursively composed using abstractive processes, for instance, by scale: an atom, a base pair, a gene, a chromosome, a strand of DNA, the nucleus, a cell, a tissue, an organ, an organ system, the human body, a family, a racial group defined by geography and heredity, or all of humanity. Note this abstraction sequence was only spatial and ignored time. Because our perceivable universe is seen through the lens of three spatial and one apparently

nonreversible temporal dimension, the mental models we compose describe the transformations of matter-energy forwards through space-time. Let us relate this to information theory and computer science, then bring it back to genomics.

In the 1930s, Alan Turing introduced an abstract model of computation, called the Turing machine (Turing, 1937). The machine is comprised of an infinite linear blank tape with a tape head that can read/write/erase only the current symbol and can move one space to the left or right or remain stationary. This tape head is controlled by a controller that contains a finite set of states and contains the rules for operating the tape head, based only on the current state and the current symbol on the tape (the algorithm or program). Despite the simplicity of this model, it turns out that it can represent the full power of every algorithm that a computer can perform and is thus a universal model of computation.

Suppose we wanted an algorithm to write down the first billion digits of the irrational number π. We could create a Turing machine that had the billion digits embedded in the finite controller (the program) and we could run that program to write the digits to the tape one at a time. In this case, the length of the program would be proportional to the billion digits of output. This might be coded in a language like C++ as: printf("3.1415926[…]7,504,551"), with "[…]" filled in with the remaining digits. If a billion-digit number was truly random and had no regularity, this would approach being the shortest program that we could write (the information-theoretic definition of randomness). However, π is not a random number, but can be computed to an arbitrary number of digits via a truncated infinite series. An algorithm to perform a series approximation of π could thus be represented as a much shorter set of instructions.

In algorithmic information theory, the *Kolmogorov complexity* or descriptive complexity of a string is the length of the shortest Turing machine instruction set (i.e., shortest computer program) that can produce that string (Kolmogorov, 1963). We can think of the problem of modeling a subset of reality as generating a parsimonious algorithm that prints out a representation of the trajectory of a set of variables representing an abstraction of that subset of reality to some level of approximation. That is, we say, "under such and such conditions, thus and such will happen over a prescribed time period". The idea of Kolmogorov complexity motivates the use of *Occam's razor*, where, given two alternate explanations of reality that explain it comparably well, we will choose the simpler one.

In our modeling of reality, we are not generally trying to express the state space transitions of the universe down to the level of every individual atom or quark in time intervals measured by Planck time units, but rather at some level of

abstraction that is useful with respect to the outcomes we value in a particular context. Also, because reality has constraints (i.e., laws), and thus regularity, we can observe a small spatial-temporal subset of reality from models that not only describe that observed behavior, but also that generalize to predict the behavior of a broader subset of reality. That is, we don't just model specific concrete observables in the here and now, but we model abstract notions of observables that can be applied beyond the here and now.

The most powerful models are the most universal, such as laws of physics, which are hypothesized to hold over all of reality and can thus be falsified if any part of reality fails to behave according to those laws, and yet, cannot be proven because all reality would have to be observed over all time. This then forms the basis of the scientific method where we form and falsify hypotheses but can never prove them. Unlike with hydrogen atoms or billiard balls where the units of observation may be considered in most contexts as near-identical, when we operate on abstractions such as cells, or people, we create units of observation that may have enormous differences.

1.2 FROM INFORMATICS TO BIOINFORMATICS AND GENOME INFORMATICS

In biology, we often blithely assume that the notion of *ceteris paribus* (all things being equal) holds, but it can lead us astray (Lambert and Black, 2012; Meehl, 1990). For instance, while genetics exists at a scale where *ceteris paribus* generally holds, we are nevertheless trying to draw relations with genetic variations at the molecular scale, with fuzzy phenotypes at the level of populations of nonidentical people.

So unlike our previous example of writing a program to generate the first billion digits of π, which has a very precise answer, our use of abstraction to model biology involves leaving out variables of small effect, which nevertheless, when left unaccounted for, may result in error when we extrapolate our projections of the future with abstract models. We would do well to mind the words of George Box, "all models are wrong, but some are useful":

> Since all models are wrong, the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary, following William of Occam, he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist, overelaboration and overparameterization is often the mark of mediocrity (Box, 1976).

How then do we choose what variables to study at what level of abstraction over what time scale? To begin to answer this question, it is useful to talk about

control in the context of goal-directedness and to turn to a field that preceded and contributed to the development of computer science, namely *Cybernetics*. In 1958, Ross Ashby introduced the *Law of Requisite Variety* (Ashby, 1958). *Variety* is measured as the logarithm of the number of states available to a system. *Control*, when stripped of its negative connotations of coercion, can be defined as restricting the variety of a system to a subset of states that are valued and preventing the other states from being visited. For instance, an organism will seek to restrict its state space to healthy and alive ones. For every disturbance that can move a system from its current state to an undesirable one, the system must have a means of acting upon or regulating that disturbance. Ashby's example of a fencer staving off attack is helpful:

> Again, if a fencer faces an opponent who has various modes of attack available, the fencer must be provided with at least an equal number of modes of defense if the outcome is to have the single value: attack parried. (Ashby, 1958)

The law of requisite variety says that "variety absorbs variety," and thus that the number of states of the regulator or control mechanism whose job is to keep a system in desirable states (i.e., absorb or reduce the variety of outcomes) must be at least as large as the number of disturbances that could put the system in an undesirable state. All organisms engage in goal-directed activity, the primary one being sustaining existence or survival. The fact that humanity has dominated as a species reflects our capacity to control our environment—to both absorb and enlist the variety of our environment in the service of sustaining health and life.

In computing, a *universal Turing machine* is a Turing machine that can simulate any Turing machine on arbitrary input. If DNA is the computer program for the "Turing machine of life," the field of human genome informatics is metaphorically moving towards the goal of a universal Turing machine that can answer "what-if" questions about modifying the governing variables of life. Note, the computer science concept of self-modifying code also enriches this metaphor. In particular, cancer genomics addresses the situation where the DNA program goes haywire, creating cancer cells with distorted copies where portions of the genome are deleted, copied extra times, and/or rearranged. Self-modifying code in computer science is enormously difficult to debug and is usually discouraged. Similarly, in cancer, we acknowledge that it is too difficult to repair rapidly replicating agents of chaos, and thus, most treatments involve killing or removing the offending cancer cells. Also, with the advent of emerging technologies such as CRISPR genome editing, humanity is now poised on the threshold of directly modifying our genome (Cong et al., 2013). Such technologies, guided by understanding of the genome, have the potential to recode portions of the program of life in order to cure genetic diseases.

With the human genome having a state space of three billion base pairs times two sets of chromosomes, compounded by epigenetic modifiers that can vary by tissue, compounded by replication errors, compounded by a microbiome living in synergy with its host, compounded by effects of the external environment, the complete modeling of the time evolution of the state space of a human organism at a molecular level appears intractable. Suppose we wanted to perform molecular dynamics simulations of the human body at a femtosecond time scale and do so for an hour. A human body contains approximately 7×10^{27} atoms that we would want to simulate over 3.6×10^{18} timesteps. Such a simulation might require 10,000 floating point operations per atom per timestep to account for various molecular forces, taking on the order of 10^{50} floating point operations. The world's fastest supercomputer at the time of this writing approaches 100 petaflops, or 10^{17} floating point operations per second (Fu et al., 2016). Such a simulation would take 10^{33} s or 3.17×10^{25} years on such a computer. However, if Moore's Law (Moore, 2006) holds and computation capacity is able to double indefinitely every 2 years, in 220 years we could perform such a simulation in 1 s.

By the law of requisite variety, it would seem that our goal of controlling the complete "computer program" of human life is doomed until long after the hypothesized *technological singularity* when human computation capacity is exceeded by computers (Vinge, 1993). We don't appear to have the capacity to model all possible deviations from health at a molecular level and form appropriate responses. However, the successes in modern treatment of numerous diseases suggest that there is enough constraint and regularity in how the building blocks of life assemble to form recognizable and treatable categories of processes and outcomes that we may hope that the set of variables we need to control may not be ultimately intractable.

The evolution of a field of knowledge towards becoming a science begins with classification (e.g., taxonomies), followed by searching for correlations (e.g., genome-wide association studies), followed by forming cause and effect models (e.g., well-characterized molecular mechanisms), and theories (e.g., Darwinian evolution). Other than the successes of understanding monogenic disease processes, much of the past 15–20 years of molecular genetic research has been in the classification and correlation stage. We are still figuring out the relevant variables in the field of genomic health, and only baby steps have been taken to form dynamic causal models of complex systems.

To better understand causation, we need to measure and model the time evolution of systems. This means that, in addition to understanding the germline DNA, we need to understand the time evolution of epigenetic modifications, gene expression, and the microbiome and how these all function together at many orders of magnitude of temporal-spatial scale, including the variation

we observe in human populations. Unfortunately, the field has barely broached studies of this kind and will thus not be covered in this book.

1.3 INFORMATICS IN GENOMICS RESEARCH AND CLINICAL APPLICATIONS

As indicated above, informatics plays a vital role not only in genomics research by interpreting high-throughput genotyping and gene expression and deep DNA sequencing analysis data, but also in clinical applications of these new technologies. This not only include informatics tools for genomics, and reciprocally proteomics and metabolomics, analysis, but also tools for the proper annotation, including but not limited to variant nomenclature. It is also of equal importance to establish incentives for openly sharing genomics research results with the scientific community.

1.3.1 Genome Informatics Analysis

Creating transparent and reproducible pipelines is essential for developing best practices for tools, data, and workflow management systems. In Chapter 2, we present a deeper dive into the software tools for managing genomic analysis pipelines, including coverage of such systems as Galaxy and TAVERNA, emphasizing the importance of creating a reliable reproducible workflow upon which others can build and contribute. Recommendations are made on coding standards, code testing and quality control, project organization, documentation, data repositories, data ontologies, virtualization, data visualization, crowdsourcing, and the support of metastudies. Discussion is made of the tradeoffs between the modularity and maintainability of command line tools versus the usability of graphical user interfaces and how modern workflow management systems combine the two.

Similarly, in this book, we present examples on how cytogenetics paradigms shape decision making in translational genomics (Chapter 3). In particular, we describe how early genomic technologies, such as low-resolution cytogenetic testing, shaped worldviews which continue to influence our mental models of genomic medicine. In particular, as our capacity to sequence the human genome has grown exponentially, our capacity to turn this data into understanding has not kept pace. As a result, in medical decision making, the two poles of the central conflict may be verbalized as: "give me only the information I am sure about so that I don't make errors of omission" versus "give me as much information as possible so that I don't make errors of omission." The advent of high-resolution genome-wide microarrays, followed by whole-exome and whole-genome sequencing, has only exacerbated this conflict, as cytogenetics gives way to cytogenomics and as physicians attempt to perform decision making under ever more uncertainty.

Ultimately, next-generation sequencing is gaining momentum in all aspects of genomics research as well as clinical applications with several different platforms existing today, making data resulting from deep DNA sequencing impossible to interpret without dedicated tools and databases. As such, we opted to present a range of available tools and databases accompanied by practical guidelines for next-generation sequencing analysis (Chapter 4). In particular, we present the modular steps involved in the processing and secondary analysis of next-generation sequencing pipelines, including both DNA- and RNA-seq, while touching upon the follow-on tertiary analysis that may be applied once genomic variants have been identified. We describe common formats for storing raw sequencing data, such as FASTQ and SAM/BAM, as well as some popular online data repositories for sequencing data. We describe the basic building blocks of next-generation sequencing pipelines, including sequence alignment as well as approaches to annotation. We close reviewing some of the limitless applications of next-generation sequencing as a modular technology.

Apart from genomics application, proteomics and metabolomics are also coming of age in the postgenomic era, and as such, proteomics and metabolomics data analysis is also key for translational medicine. In Chapter 5, we underscore the importance and pitfalls of large-scale proteomics and metabolomics measures in the clinic for characterizing biological processes and objectively characterizing phenotypes in translational medicine. Discussion is made of ways and means for managing the complexity of these datasets as genome/proteome/metabolome interactions are considered and what challenges remain for broader adoption of these technologies in clinical practice.

1.3.2   Genomics Data Sharing

The continued deposition of genomic data in the public domain is essential to maximize both its scientific and clinical utility. However, rewards for data sharing are currently very few, representing a serious practical impediment to data submission. Moreover, a law of diminishing returns currently operates both in terms of genomic data publication and submission since manuscripts describing a single or few genomic variants cannot be published alone. To date, two main strategies have been adopted as a means to encourage the submission of human genomic variant data: (a) database journal linkups involving the affiliation of a scientific journal with a publicly available database and (b) microattribution, involving the unambiguous linkage of data to their contributors via a unique identifier. The latter could, in principle, lead to the establishment of a microcitation-tracking system that acknowledges individual endeavor and achievement (Giardine et al., 2011; Patrinos et al., 2012).

In Chapter 6, we discuss an important trend in science that started early in the field of genomics, as an outgrowth of public funding. That is, data generated for

one research purpose by one organization can be shared with the entire field to augment our collective capacity to model the complexity of the human genome and genomic processes. We also discuss both social and technical challenges of realizing the full potential of collaborative science, with an emphasis on reward systems that enable credit and attribution to be made to genomic data contributors, which could eventually become more widely adopted as novel scientific publication modalities.

1.3.3 Genomic Variant Reporting and Annotation Tools

Advances in bioinformatics required to annotate human genomic variants and to place them in public data repositories uniformly have not kept pace with their discovery. At present, there are a handful of tools that are used to annotate genomic variants so that they are reported with a constant nomenclature. In Chapter 7, we discuss the Human Genome Variation Society (HGVS) nomenclature system for variant reporting and its challenges with ambiguous reporting of variants. We then describe a new tool, MutationInfo, to automatically infer chromosomal positions from dbSNP and HGVS genetic variants. This tool combines existing tools with a BLAST-like alignment tool (BLAT) search in order to successfully locate a much larger fraction of genomic positions for HGVS variants. Finally, we compare the available tools for checking the quality of variants documented in HGVS resources and dbSNP and we highlight the challenge of consistently representing genomic mutations across databases due to multiple versions of different coordinate systems in use.

1.4 PHARMACOGENOMICS AND GENOME INFORMATICS

Pharmacogenomics aim to rationalize drug use by delineating adverse reactions and lack of drug response with the underlying genetic profile of an individual. Since there is a documented lack of (pharmaco)genomics knowledge from clinicians (Mai et al., 2014), there is an urgent need to develop informatics solutions and tools to translate genomic information into a clinically meaningful format, especially in the case of individualization of drug treatment modalities. In other words, a tool that would be able to translate genotyping information from a few or more pharmacogenomic biomarkers into recommendations for drug use.

Chapter 8 touches upon the challenges in translating pharmacogenomics knowledge—the relationship between genetic factors and drug safety and efficacy—to decision making in a clinical setting. We present pharmacogenomics information system approaches to managing both the complexity and uncertainty of pharmacogenomic decision making towards achieving better health outcomes through personalized drug treatment, in the context of a