Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

5

Finding relationships among biological entities

CHAPTER OUTLINE

It is impossible too often to remind people that, on the one hand, all correct reasoning consists in substituting like things for like things, and inferring that what is true of one will be true of all which are similar to it in the points of resemblance concerned in the matter. On the other hand, all incorrect reasoning consists in putting one thing for another where there is not the requisite likeness. It is the purpose of the rules of deductive and inductive logic to enable us to judge as far as possible when we are thus rightly or wrongly reasoning from some things to others.

William Stanley Jevons, English economist and logician (1835–82)

Section 5.1. Defining relationships and similarities

We are raised to believe that science explains how the universe, and everything in it, works. Engineering and the applied sciences use these scientific explanations to create things, for the betterment of our world. This is a lovely way to think about the roles played by scientists and engineers, but it is not completely accurate. For the most part, we do not understand very much about the universe. Nobody understands the true nature of gravity, or mass, or light, or magnetism, or atoms, or thought. While we don't understand much about anything, we are very good at understanding the relationships among the things we do not understand. We can credibly specify relationships between gravity and mass, mass and energy, energy and light, light and magnetism, atoms and mass, thought and neurons,

Logic and Critical Thinking in the Biomedical Sciences. https://doi.org/10.1016/B978-0-12-821364-3.00005-X

and so on. Karl Pearson, a 19th century statistician and philosopher, wrote that "All science is description and not explanation." Pearson was admitting that we can describe relationships but we cannot explain why those relationships are true.

Before going any further along this line of thought, it is important to draw the distinction between a relationship and a similarity, insofar as one of the most common errors in scientific analysis involves confusing these two concepts. Here is a short story that demonstrates the distinction: You look up at the clouds, and you begin to see the shape of a lion. The cloud has a tail, like a lion's tale, and a fluffy head, like a lion's mane. With a little imagination the mouth of the lion seems to roar down from the sky. You have succeeded in finding similarities between the cloud and a lion. If you look at a cloud and you imagine a tea kettle producing a head of steam and you recognize that the physical forces that create a cloud and the physical forces that produced steam from a heated kettle are the same, then you have found a relationship.

A relationship is a fundamental and unchanging feature that permits us to describe one thing in terms of another thing. A similarity is simply a feature that happens to be present in two compared things; it may or may not help us establish a relationship. In the case of genetic ancestry, we may notice that a child has various biological features that are similar to the biologic features of a parent (e.g., height, weight, and color of eyes). In such cases, we must understand that the similarities are present as a consequence of the genetic relationship between parent and child, not the other way around. Similar traits that one individual may have in common with another individual do not constitute a relationship.

In medicine a good example wherein we benefit from understanding a fundamental relationship, in the absence of superficial similarities, is found in the ciliopathies. The ciliopathies are a clinically diverse set of diseases that are all closely related to one another based on one shared relationship: they all involve an inherited alteration of the primary cilium. The primary cilium, also called the nonmotile cilium, is a single undulipodium (also called flagellum) found in nearly every cell of all vertebrates. Cilipathies affect the kidneys, bones, and eyes and can produce heterotaxies, such as situs inversus, wherein the position of organs is switched from one side of the body to the other.[1,2] There are only a few organs and functional systems of the human body that escape involvement by this strange collection of pathogenetically related but clinically dissimilar diseases. Had we focused on clinical similarities among diseases, we would never have discovered that all of these diseases are related to one another by defects in one cytoarchitectural component. When we study the primary cilium and find new drugs that protect, repair, or otherwise compensate for defects in the primary cilium, we will likely have a treatment that can be useful in all of the ciliopathies. [Glossary Cilia, Primary ciliopathies]

Many of the most popular algorithms available to data analysts are devoted to finding similarities among data objects (e.g., movie preferences, shopping habits, and daily routines). It is important to understand that such algorithms may be very useful for marketeers, but their scientific value is dubious, unless the analyst can successfully discover a set of relationships based upon discovered similarities. In general the similarities we find among data objects are clues at best. Similarities sometimes help us find relationships, but we can never depend on similarities as our basis for drawing scientific inferences.

Section 5.2. Ancestral genes

In the past several decades, we have collected a great deal of information about the gene sequences in many different species of living organisms. The lessons we have learned can be summarized in a few broad statements.

1. **The genome is much more complex than we had imagined, and most of this complexity seems to reside within a variety of genomic control systems that we understand only superficially.**

Some of these control systems include DNA methylations, histone and nonhistone protein complexes with chromatin, noncoding RNAs, microRNAs, RNA splicing factors, nonsequence modifications of RNA that might influence gene expression, and pseudogenes.[3,4] [Glossary CpG island, DNA methylation, Spliceosome, Alternative RNA splicing, Pseudogene]

2. **There is an enormous amount of gene sequence diversity, from species to species and from individual to individual within a species.**

The most studied genetic variations in humans are single nucleotide polymorphisms (SNPs), wherein differences of one nucleotide at various sites in the genome can occur at relatively high frequency within the population of a species. There are millions of SNPs in the human genome, with each SNP indicating a single nucleotide variant that occurs in some individuals and not in others. A SNP is estimated to occur about once in every 300 nucleotides and sometimes even higher.[5,6] It seems that the more individuals you sample for a SNP database, the more variations you can find. Due to the high frequency of genetic variations, we shall never be able to catalog all the SNPs that occur in the full population of humans (Fig. 5.1). [Glossary Molecular targeted drugs, Synonymous SNPs, Silent mutation]

By studying SNPs, we can sometimes establish an association between particular SNP variations, or sets of variations, and an individual's susceptibility to diseases. [Glossary Anonymous variation, Genome wide association study, DNA polymorphism]

SNPs are just one form of genetic polymorphism among individuals. Alterations in chromosome number, deletions of stretches of DNA, insertions of DNA, and a host of subtle complex variations wherein parts of chromosomes are translocated elsewhere within the same chromosomes or to other chromosomes are all sources of genetic variation within the human population.[7] [Glossary Genomic disorder, Microdeletion]

3. **Despite all the diversity among the genomes of individual organisms, we find that nearly all the genes in humans and other species of living organisms can be traced back to genes found in distant ancestors that we all have in common.**

The vast majority of the genes in extant mammalian species are ancient, arising hundreds of millions of years ago, sometimes billions of years ago. Here are a few observations that illustrate the point:

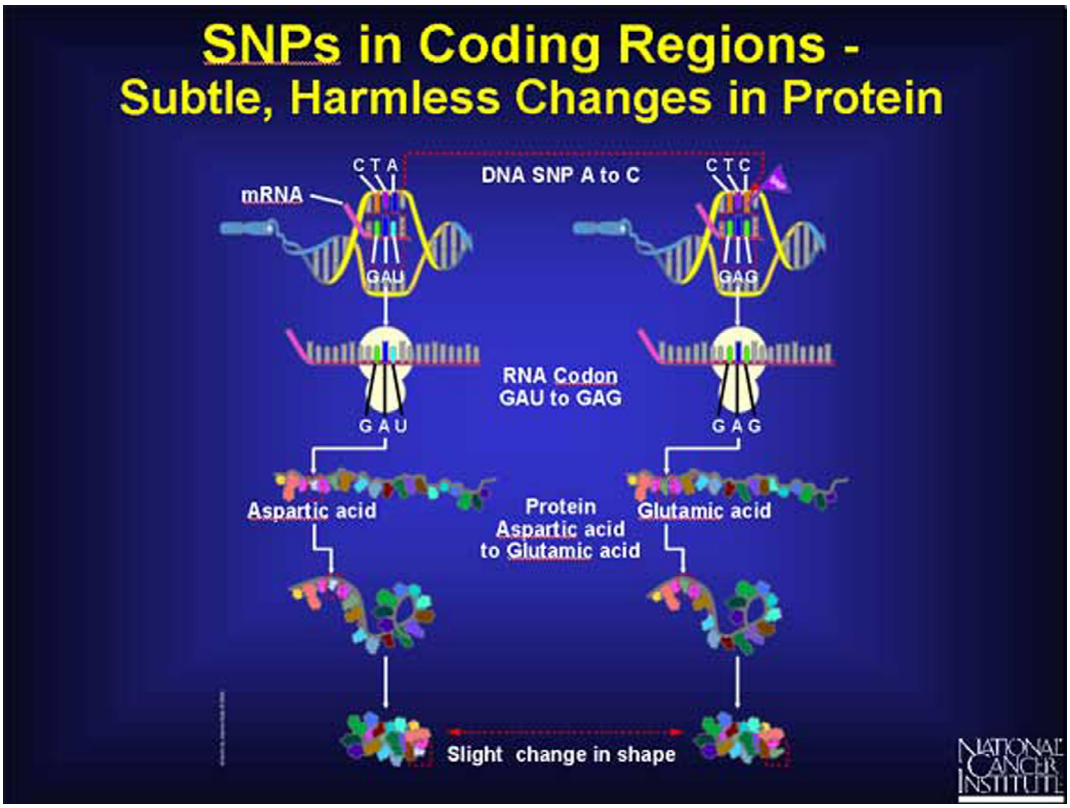

FIG. 5.1 Graphic indicating how a difference of one nucleotide (i.e., single nucleotide polymorphism or SNP) may result in a minor modification of a protein. *Source: U.S. National Cancer Institute, NIH.*

There are over 500 core genes that are found throughout Class Metazoa (i.e., animal species).[8] Not all species have all of these genes, but it would seem that all species have most of these core genes. Even the most basal metazoan, the placozoan *Trichoplax adhaerens*, which lacks neurons and organs, contains 66% of the set of core metazoan genes.[8]

Not only do we find homologs of human genes in single-celled eukaryotic organisms, but we also find homologies that extend all the way up our ancestral lineage and into the bacterial kingdom (as in the case of an actin homolog in the proteobacteria Haliangium ochraceum[9]). [Glossary Homolog]

Approximately 60% of the annotated protein coding genes in the mouse genome originate from prokaryotic and basal eukaryotic ancestors,[10] and there is every reason to believe that the same can be said for all mammals, including humans.

Nearly 75% of human disease-causing genes are believed to have a functional homolog in the fly.[11]

The human genome and the chimpanzee genome are between 97% and 92% alike, indicating that closely related species have nearly the same set of genes.[12]

Pax6 is a regulatory gene that controls eye development. This gene, like so many other basic regulatory genes, has been strictly conserved throughout metazoan evolution. Remarkably, Pax6 in mice is so similar to its homolog in insects that the corresponding genes from either species can be interchanged and function properly in the recipient organism.[13] [Glossary Law of sequence conservation]

Nearly identical gene families (vida infra) are found in many different classes of animals.

The human genome is packed with relic sequences from ancient ancestral organisms, most notably viruses.

You can make a strong argument that so much of the human genome is found in other organisms that we barely have a genome that we can call our own. Grudgingly, we must admit that we are not self-made men (or women). We inherited our genetic fortunes from our ancestors. The news is not all bad. Because our genes are much the same as the genes of other animals, we can infer the purposes of our genes by studying homologous genes in other species. Doing so has helped us to build phylogenetic trees for all living species of organisms. Homologous genes also allow us to study the pathogenesis of human disease based upon observations in other species of organisms (e.g., fish, roundworms, mosquitoes, and yeast). [Glossary Mosquito, Nonphylogenetic signal]

As genes are inherited through the ancestral lineage, gene duplication events permit organisms to repurpose conserved genes.[14] The duplicate gene, no longer serving an essential function (fulfilled by the original gene), can serve some new purpose after it has been modified by mutation and preserved in the species gene pool through natural selection. Because the new gene is a modification of the original gene, you might expect the two genes to have some functional similarity. When this is so, the two genes are said to be members of a gene family. When the members of a gene family are situated close to one another on a chromosome, they are referred to as gene clusters. Further duplications of members of a gene family may produce large gene families. As we might expect, large gene families evolve over long periods of time, and this tells us that large gene families are ancient and may be present (in part or in whole) in all the classes of organisms that descend from the gene family's founder organism. [Glossary Founder effect, Gene pool, Natural selection, Paralog]

An example of an ancient gene family is the globins, whose function in the red blood cells of many animals is the distribution of oxygen from an external source (e.g., air or water) to tissues throughout the body.[15,16] The genes of the globin family are so ancient that they are present in bacteria. It is believed that the original globin molecule evolved in bacteria prior to the evolution of oxygenic photosynthesis (i.e., prior to 2.3 billion years ago) at a time when all organisms were anaerobic. Oxygen is toxic to anaerobic organisms. The purpose of the first globins in bacteria was to scavenge oxygen, hence protecting anaerobic cells from oxygen-induced toxicity. Following the advent of oxygenic photosynthesis, when

the earliest aerobic organisms were evolving, the oxygen-scavenging properties of the early globins were modified to facilitate the extraction of oxygen from the air and the delivery of oxygen to the intracellular enzymes participating in the pathways for oxidative phosphorylation. A variety of hemoglobinopathies affect humans, all involving the globin gene family.[15] [Glossary Oxygen crisis]

Gene acquisitions may occur after viruses successfully integrate into the human human genome (or the genome of our ancestors). About 8% of our human genome is derived from sequences with homology to known infectious retroviruses, and these sequences can usually be recognized by their subsequences that contain viral genes (e.g., gag, pol, and env genes) and by the presence of long terminal repeats. The viral sequences, remnants of ancient retroviral infections, and the occasional nonretroviral infection were branded into our DNA and subsequently amplified.[17–19] [Glossary Retrovirus]

Section 5.3. The significance of gene sequence conservation

The law of sequence conservation is the guiding principle of bioinformatics. Put simply, if a sequence is conserved through evolution (i.e., if you can find a closely similar sequence that is present in a class of organisms and in their line of mutual ancestors), then that sequence must perform a useful function for the organism. Furthermore, if a sequence is highly conserved (i.e., with very little sequence difference among class members and their ancestors), then that sequence is likely to have a vital function.[20] For example, oncogenes are highly conserved, suggesting that oncogenes, despite their carcinogenic potential, serve important purposes in the human genome and in the genomes of our ancestors.

The law of sequence conservation is so useful and so fundamental to the field of genomics and to the design of gene-related computational algorithms that it may as well be known as the first law of bioinformatics. [Glossary Bioinformatics, Sequence similarity]

One of the consequences of the law of sequence conservation is that most monogenic diseases occur in conserved genes. Why is this so? Genes that are not conserved can tolerate mutations without causing disease (that is why they are not conserved). Genes that are highly conserved cannot tolerate mutation without producing disease (that is why they are conserved). This being the case, we can presume that just about every monogenic disease of humans (i.e., nearly all the rare inherited diseases and the rare de novo diseases) is also a monogenic disease of ancestors in the phylogenetic lineage (that is because conserved human genes are genes found in our ancestral lineage). [Glossary Monogenic disease, Single gene disease]

How is gene conservation actually achieved? It's all done by natural selection. A new mutation in a gene is more likely to be deleterious than beneficial. Hence, if a gene is vital to the function of an organism and if a mutation occurs in this gene, then the organism that carries the mutation will be less likely to procreate and pass the mutated gene to the next generation.[20] Conversely, if a gene serves no function, then mutations will take place without reducing the fitness of organisms, and such mutations may greatly expand the size of the species gene pool. Hence, junk DNA gets progressively junkier as time passes.

There are four conceptual points that we need to keep in mind when we invoke the law of sequence conservation:

1. There is no biochemical method by which certain sequences in the genome are marked for conservation while other sequences are marked for mutability.

It's all done through natural selection.

2. Sequences that are not conserved and have no function are the ingredients for new genes (that will be conserved).

Nonconserved sequences, that serve no function essential for survival, may mutate to produce new and useful sequences. In fact, it is only the nonconserved sequences from which new genes may evolve (because the conserved sequences are constrained not to mutate).[20]

3. Regulatory genes are highly conserved. Hence, genomic regulation is important to the species, perhaps as important as are the protein-coding genes. [Glossary Gene regulation]

Many of the most highly conserved sequences occur in noncoding regions (i.e., sequences that are not genes).[21–24] We can infer that conserved noncoding sequences must have an important function. The only known function of noncoding sequences is genomic regulation. Hence, we can infer that conserved noncoding DNA sequences have a regulatory function. Empirically, this seems to be the case.[25]

4. If a mutation in a gene produces a disease in humans, we could reasonably expect that the same mutation might produce some deleterious effect in other species that occupy the human phylogenetic lineage.

Among the conserved genes, we like to presume that a mutation that produces a deleterious effect in one species is likely to produce a similar negative effect in other species of the same ancestral lineage. Of course, there are exceptions. Lesch-Nyhan disease is a rare syndrome caused by a deficiency of hypoxanthine-guanine phosphoribosyltransferase (HGPRT), an enzyme involved in purine metabolism. In humans, HGPRT deficiency results in high levels of uric acid, with resultant renal disease and gout. Various severe neurologic and psychologic symptoms accompany the syndrome in humans, including self-mutilation. Neurologic features tend to increase as the affected child ages. The same HGPRT deficiency of humans can be produced in mice. As far as anyone can tell, mice with HGPRT deficiency are totally normal.[26] Likewise, mice with the most commonly mutated gene known to be the root cause of human nephronophthisis, a clinically serious inherited kidney condition, do not develop the disease.[1]

Here is one additional example. There are rare subtypes of type 2 diabetes that have a monogenic origin. These rare types of diabetes become clinically evident in children and have a Mendelian pattern of inheritance. One such monogenic form of diabetes is MODY-8 (maturity-onset diabetes of the young), caused by a mutation in the carboxyl-ester

lipase gene. This mutation was experimentally inserted into the germ line of a mouse, producing a transgenic strain of mice carrying the MODY-8 mutation. These mice failed to develop any signs of diabetes, pancreatic damage, or any dysfunction caused by the mutated gene.[27] Why not? **In all three cases (murine versions of Lesch-Nyhan disease, nephronophthisis, and monogenic diabetes), we presume that the mouse employs pathways that compensate for the genetic deficiency.**

As it happens, some of the most conserved genes, which have been passed over billions of years from bacteria to primitive eukaryotes and eventually to humans, are a lot more dispensable than most of us had imagined. It would seem that just a few hundred bacterial genes are sufficient to maintain a living organism.[28] In the case of metazoans, 500–600 genes apparently constitute a core collection that are common to most animals.[8] What can we expect to happen when we start knocking out those identified genes from living organisms? Wouldn't any organism die if it were deprived of any of a universally conserved gene? It turns out that such is not the case. In some instances, knocking out universally conserved genes, presumed to encode essential cellular activities, had little effect on cell viability.[28] People with an inflated opinion of their own importance are often reminded that nobody is indispensable. The same advice may apply as a universal truth for genes. Organisms have a way of compensating for missing genes, either by coopting alternate metabolic pathways or by finding alternate enzymes that might replicate the activity of a missing protein.

Section 5.4. Unexpected gene conservation

Despite our sense that anything is possible in the vastness of space, we see an awful lot of sameness throughout the universe. Wherever we aim our telescopes, we see galaxies, most of which are flat and spiral, often having about the same size, and composed of the same objects: stars, planets, gas, dust, black holes. A small set of physical laws impose stability everywhere at once, and the result is the somewhat repetitious world in which we live. Likewise, despite the large number of species living on our planet, they are all variations of a few common themes that can be encapsulated under a simple classification.

About a half billion years ago, the early metazoan classes (i.e., animals) evolved at a rapid rate, producing dozens of body plans that we can examine in ancient shale deposits. This period, which lasted about 40 million years or so, is known as the Cambrian explosion. The same body plans that evolved during the Cambrian explosion account for nearly all of the classes of animals that live today. This is to say that since the Cambrian, no new body plans have gained entry into the metazoan world. A great deal has been written about the Cambrian explosion, much of it focused on why current metazoan body plans are basically limited to the Cambrian models.[29] It seems bizarre that despite the billions of new animal species that have appeared on earth, since the Cambrian, they all seem limited to a small number of ancient body plans. On the plus side the limited repertoire of body plans makes it easy to classify organisms. Hence, by the mid-20th century, the classification of living animals (i.e., members of Class Metazoa) was largely settled.[30,31] There were lots of

debatable class assignments, but taxonomists had every expectation that all quarrels would eventually be resolved. [Glossary Cambrian explosion]

Bacteria, in contrast to animals, can pick and choose their internal structure ad libitum. At the same mid-century mark, bacterial taxonomy was in a sorry state. Infectious organisms were classified by their growth in different culture media, by the staining properties (e.g., gram positive, gram negative, and acid fast), by the speed of their growth, by the temperature at which they grew, and by whether they produced pigment. These diagnostic tests helped us identify pathogenic bacteria, but they did not help the taxonomists understand the biological relationships among different species of bacteria. [Glossary Classification system versus identification system]

When molecular biologists developed methods for sequencing the genomes of various organisms, there was hope that bacterial classification could be based on genomic sequencing. This hope was somewhat deflated when it was determined that horizontal gene transfer (i.e., the exchange of genetic material among different species) was rampant in the bacterial kingdom. At the time, it was feared that bacteria were just a mish-mash of genes popping in and out of organisms, with no stable genetic composition. The thinking went that if the gene pool is constantly being flooded with the genes of other species, then there is no stable gene pool, hence no sensible way of assigning species and no credible classification of bacterial organisms. Bacterial taxonomy seemed hopeless. [Glossary Horizontal gene transfer]

In 1977 the field of bacterial taxonomy changed for the better when Carl Woese and George E. Fox announced that there existed a class of bacteria that contained species that were demonstrably different from all other species of prokaryotes. They named these bacteria Archaebacteria, later known as Archaea (from the Greek meaning original or first in time), the name indicating that the Archaea predated all other classes of prokaryotes (i.e., organisms having no nucleus). When we compare species of Class Eubacteria with species of Class Archaea, we are not likely to notice any big differences. The eubacteria have the same shapes and sizes as the archaea. All species of eubacteria and all species of archaea are single-celled organisms, and they all have a typical prokaryotic structure (i.e. lacking a membrane-bound nucleus to compartmentalize their genetic material). As it happens, Class Eubacteria contains all of the bacterial organisms that are known to be pathogenic in humans. The archaeans, so far as we can currently tell, are nonpathogenic. Many archaeans are extremophiles, capable of living in hostile environments (e.g., hot springs and salt lakes), but some Archaean species occupy less demanding biological niches (e.g., marshland, soil, and human colon). Class Archaea does not hold a monopoly on extremophilic prokaryotes; some eubacterial species live in extreme environments (e.g., the alkaliphile *Bacillus halodurans*).

Woese and Fox showed that despite extensive horizontal gene transfer in ancient and modern prokaryotes, there are fundamental differences among prokaryotes that establish excellent criteria for assigning biological classes.[32,33] Differences in the sequence and structure of ribosomal RNA (the 16sRNA component in particular) distinguish archaean species from bacteria.[33] Furthermore the differences in 16sRNA among the prokaryotic

classes can be exploited to establish the different subclasses of prokaryotic organisms and the chronology of their evolution. It would seem that prokaryotes never swap their ribosomal RNA and that prokaryotes exist as valid species (i.e., evolving gene pools). Later studies indicate that genes other than ribosomal RNA may serve as keys to the phylogenetic organization of prokaryotes, if we just look for them.[9] [Glossary Molecular clock]

Just as we once believed that it was hopeless to try to create a valid phylogenetic classification of bacteria, many scientists currently believe that it would be impossible to create a phylogenetic classification of viruses. At the current time the classification of viruses is based on phenetics (i.e., based on physical similarities). As discussed previously, good classifications are never based on measuring similarities. In the case of viruses, if we classify based on their genomic molecules (i.e., DNA or RNA, single stranded or double stranded), we find that subclasses with of the same genomic type will have dissimilar structures: envelope, size, shape, proteins, and capsid. When we classify viruses based on method of contagion or by persistence within hosts (i.e., acute, chronic, latent, or persistent), or by toxicity (lytic and immunogenic), or by target cell specificity, no consistent taxonomic order is revealed. [Glossary Capsid, Phenetics]

For the pessimists among us, the barriers to discovering a good phylogenetic classification are overwhelming. Let's take a look at a few of the impediments to establishing a viral phylogeny.

1. Large number of known and unknown viral species

Simply put, the greater the number of species, the more work is required to prepare a taxonomy. Every new species requires a certain irreducible amount of study, and if new species are being discovered at a rate that exceeds our ability to describe and classify known species, then the list of unassigned species will become infinitely long over time.

2. Lack of any accepted concept of a "root" virus

The classification of cellular organisms is built on the premise that each of the major classes (i.e., bacteria, archaeans, and eukaryotes) has a root or founder class, with a hypothesized set of class-defined features, from which all subclasses descended. In the case of viruses, we really have no way of describing the ancestor of all extant viruses.

Furthermore, we define viruses as being obligatory parasites, requiring one of the major classes of cellular organisms as a host. If this were the case, then viruses, as we have come to define them, could not have existed prior to the existence of host organisms to parasitize. Hence, if viruses existed prior to the emergence of cellular life, then the root of the viruses was not a virus, insofar as the root organism could not have parasitized a cellular host. If the root of the viruses was not a virus, then it may have been almost anything, and we could not rule out the existence of multiple root organisms, accounting for the widely varying versions of viral genomes that we observe today.

Toying with phylogenetic logic is harmless fun, but it illustrates how it is impossible to create a top-down classification of viruses, if we know nothing about the biological features that would define the top class. [Glossary Nonphylogenetic property]

3. High rate of mutation in viruses

For the most part, viruses do not repair their genomes. A notable exception is the megavirus *Cafeteria roenbergensis*.[34] Presumably, we will find that other megaviruses have DNA repair pathways, but the small, simple viruses have high rates of mutation in DNA and RNA, with no mechanism to repair the damage. This means that genome-damaged viruses have two choices: to die or to live with and replicate their mutations. Consequently, viral genomes tend to degenerate quickly, producing lots of variants. Species mutability is particularly prevalent among the RNA viruses (e.g., influenza virus, Newcastle disease virus, and foot and mouth disease virus).

Mutational variations of a virus do not produce a new species; variations produce diversity in the viral gene pool of the species. If new mutations do not produce an alteration in the specificity of its host organisms, essentially establishing a separate gene pool for the variant, then the mutational variants will usually preserve their membership in the same viral species.

Still, all those viral genomic variants complicate the job of the viral taxonomist. Basically the high rate of mutation in viruses yields lots of genomic variation among viral populations, making it easy for bioinformaticians to detect new species where none exists. We can easily imagine a situation wherein new species are discovered and old species are declared extinct, because we simply do not have the time and manpower to carefully examine every genomic variant for the structural and physiologic features that determine its correct taxonomic classification. Bioinformaticians off-handedly refer to the variant genomes, resulting from mutations and replication errors, as quasispecies.[35] For the traditional taxonomist, trying to create a simple phylogenetic classification of viruses, the vague concept of "quasispecies" must be particularly exasperating.

4. Multiparental lineage of viruses.

Viral reassortment is a process wherein whole segments (the equivalent of viral chromosomes) are exchanged between two viruses infecting the same cell. Viral reassortment has been observed in four classes of segmental RNA viruses: Bunyaviridae, Orthomyxoviridae, Arenaviridae, and Reoviridae. Following reassortment a new species of virus may appear, and this new species will contain segments of two parental species. This poses a serious problem for traditional taxonomists, who labor under the assumption that each new species has one and only one parental species.[36] It is the uniparental ancestry of biological classifications that accounts for their simplicity and for the concept of lineage, wherein the ancestry of any species can be computed from as an uninterrupted line of classes stretching from the species level to the root level. When a species has more than one parent, then its lineage is replaced by an inverted tree. The tree branches outward with each class reaching to more than one parent class, iteratively producing a highly complex ancestry wherein the individual classes have mixed heritage. [Glossary Reassortment]

In the case of viruses that mutate at a high rate; that exchange large pieces of their DNA; that extract DNA from their host organisms; and that produce an uncountably large number of diverse species, one might think that finding a stable classification would be an

impossible task. Not so. **We are finding that viruses, like all other organisms on earth, contain conserved genes that permit us to trace the ancestral lineage of viral species, hence opening the possibility of creating a true phylogenetic classification of viruses.**[37–40]

Conserved molecular gene motifs help us to classify viruses into biological groups that share phylogenetic origins.[41,42] For example, despite the sequence variations that occur in rapidly mutating viruses, scientists find that the three-dimensional folds of protein molecules are conserved and that viruses can be grouped into so-called fold families, which can in turn be grouped into fold super families that preserve phylogenetic relationships among viral lineages.[41]

Among the retroviruses, it has been shown that viral ancestral lineages can be determined by looking at inherited variations in so-called "global" genomic properties (e.g., translational strategies and motifs in Gag and Pol genes and their associated enzymes).[42]

We shouldn't be surprised that viruses, like every other class of organism, fall into a rather limited set of phylogenetic classes. Because all viruses are parasitic (i.e., depend on host cells for their replication and survival), we can see why all viral species are constrained to evolve in a manner that maintains their host compatibility.[43] A demonstration of host-specific constraints on viruses is found by noting the specificity of viruses for all the so-called kingdoms of organisms. Virus infections are found in Class Archaea, Class Bacteria, and Class Eukaryota, but there is no instance in which any single class of viruses is capable of infecting more than one of these classes of cellular organisms. Furthermore, within a class of cellular organisms, there are only rare instances of classes of virus that can infect distantly related subclasses. For example, there are virtually no viruses that can infect both Class Animalia and Class Plantae (rare exceptions are claimed[44]). Furthermore, as the host evolves, so must the virus. Hence, we might expect to find ancestral lineages of viruses that shadow the lineage of their host organisms.

Highly innovative work in the field of viral phylogeny is proceeding, from a variety of different approaches, including inferring retroviral phylogeny by sequence divergences of nucleic acids and proteins in related viral species,[42] tracing the acquisition of genes in DNA viruses,[45] and dating viruses by the appearance of viral-specific antibodies in ancient host cells.[46] Because viruses evolve very rapidly, it is possible to trace the evolution of some viruses, with precision, over intervals as short as centuries or even decades.[41,47–49] [Glossary Precision]

The tiny world of viruses was recently enlarged by the discovery of nucleocytoplasmic large DNA viruses, NCLDVs popularly known as giant viruses).[50,51] The life of an NCLDV is not much different from that of obligate intracellular bacteria (e.g., rickettsia). The NCLDVs, with their large genomes and complex sets of genes, have provided taxonomists with an opportunity to establish ancestral lineages among some of these viruses.[50,51]

Section 5.5. Relationships between human diseases and orthodiseases

Animal models have been used for centuries to study human diseases and to test the safety and efficacy of drugs. Unfortunately, for ourselves and for the animals sacrificed in such studies, we are learning that these traditional forms of experimentation often lead to disappointment.

In Volume 1, Section 2.5, "All eutherian cell types are equivalent among classes of species," we examined a tragic incident, highlighting the limitations of animal tests for drug safety.[52] In two separate clinical trials, preclinical animal studies failed to predict human toxicity. Why not? Aside from differences in targeted disease pathways, every species has its own methods for dealing with toxic insults. In many cases toxins cause less cellular damage than the "defense" mechanisms created to nullify the damage. In at least one such case, human study participants were ravaged by an exaggerated inflammation pathway reaction known as a cytokine storm.[53–55]

Species react to stimuli in species-typical ways. Humans with unrelated types of injuries will react with a systemic response that varies little from person to person.[56] Whether the injury is trauma, burns, or endotoxemia, the physiologic response in humans is reflected in a shared gene expression profile. The stereotypical human response to injury is not replicated in mice; neither is the accompanying gene expression profile. It is of no surprise, then, that of about 150 candidates, antiinflammatory agents developed from mouse inflammation models, none have passed trials in humans.[56] Perhaps, we should simply stop using animals to predict systemic reactions to toxins, drugs, infections, and injuries.

Although animals respond to changes in their environments in species-specific ways, we know that animals share many of the same genes and that the function of these genes is often very similar in different organisms. If we restrict our attention to genes of known activity in distantly related organisms, we may learn something of value.

As background for this interesting approach to animal-based research, let's review a few definitions. Genes from two different organisms are considered homologous to one another if both originated from a gene in a common ancestral organism. If the mechanism of descent was through speciation (i.e., if the genes in both organisms came through descent from a common ancestor), then homologous genes are also orthologous. If the homologous genes arose after gene duplication within a species, then the homologous genes are paralogous. Orthologous genes tend to have a similar function, in every organism in which they are found. As previously mentioned, Pax6 is a gene that controls eye development and is an ortholog shared by metazoans (i.e., animals). Pax6 in mice is so similar to its homolog in insects in which the corresponding genes from either species can be interchanged and function properly.[13] Likewise, mammalian genes can be switched with their insect homologs and function adequately in the recipient organism.[57–61] [Glossary Ortholog]

Orthodiseases are conditions observed in nonhuman species that result from alterations in genes that are orthologous to the genes known to cause diseases in humans. For example, if a loss-of-function mutation in a particular gene in humans was associated with an inherited blood cell disorder and if a loss-of-function mutation in the homologous gene in a zebrafish resulted in lymphocytosis (proliferation of lymphocytes), then we would consider the condition in zebrafish to be an orthodisease of the human genetic counterpart.

Orthologous genes may serve as important clues to the pathogenesis of diseases in humans, even when the organisms themselves are not closely related. The logical justification for orthodisease models of human pathologic processes is as follows:

1. The mutations that account for genetic diseases in humans occur in conserved genes.

The reason being that conserved genes are essential (otherwise, they would not be conserved). If a malfunction occurs in an essential gene, it is likely to produce some pathological consequences.

2. Conserved genes nearly always have homologs that can be found throughout the eukaryotic lineage.

If a gene is essential for us, it is probably essential for other organisms.

3. The homologous conserved genes of humans that are found in other animals are likely to participate in metabolic pathways that are at least similar to the pathways found in humans.

The reason being that conserved genes tend to have a similar function (i.e., similar substrates and similar products) in every organism. Hence, we might expect that the proteins encoded by conserved genes will participate in conserved metabolic pathways. As it happens, nearly 75% of human disease-causing genes are believed to have a functional homolog in the fly.[11]

4. The metabolic pathways affected by gene mutations are likely to be involved in disease pathogenesis.
5. It is easier and faster to study genetic mutations and their subsequent effects on metabolic pathways and on the development of disease in model organisms (such as worms, flies, fish, and yeast) than in mammalian systems.
6. Drugs that are effective in modifying metabolic pathways in model organisms are likely to have similar biological effects in humans. [Glossary Driver pathway, Druggable driver]

There is abundant observational and experimental evidence that supports all of these logical assertions, and the study of human disease pathogenesis using nonmammalian orthodisease models is currently flourishing.[11,62–66] Where traditional animal models are failing, biologists are finding success with single-cell eukaryotes and insects. Though we can expect disease phenotypes to diverge among species affected by orthologous

genes, we might be able to study specific pathways that have been conserved through most of the history of eukaryotic evolution. For example, the 2013 Nobel Prize in Physiology or Chemistry was awarded for work on vesicular transport disorders. Progress in this area came from studies of human-inherited transport disorders,[67] but the vesicular transport pathway was dissected by studying orthologous genes in yeast.[65]

Though yeast shares many homologous genes with humans, there has been concern that the homologs may not participate in the same pathways in yeast as they do in humans. In a large study of the proteins involved in human spinocerebellar ataxias, it was found that the human proteins participated in pathways that were similar to the yeast pathways followed by their yeast orthologs.[68] Hence, for yeast models of the spinocerebellar ataxias, homologous pathways seem to exist for homologous genes, supporting the relevance of the yeast model for human disease (Fig. 5.2).

The nematode *Caenorhabditis elegans* is a well-studied organism. Despite its distant phylogenetic lineage (i.e., Class Protostomia, not Class Deuterostomia, which contains humans), more than 65% of human disease-causing genes currently identified have a counterpart in the *C. elegans* worm.[69] As a hermaphroditic organism, *C. elegans* has a particularly useful property that enhances its value in disease research. When organisms are exposed to mutagens, the first-generation progeny self-fertilize producing some second-generation worms that are homozygous for the mutation. This has allowed researchers to study how specific mutations disrupt development and cause disease.

When using nematodes (roundworms) to study human disease processes, we must be very careful to remember that biological systems are always complex and that the final phenotype resulting from a gene mutation is an emergent property of the total system that develops over time. For example, the root cause of human retinoblastoma (a cancer of

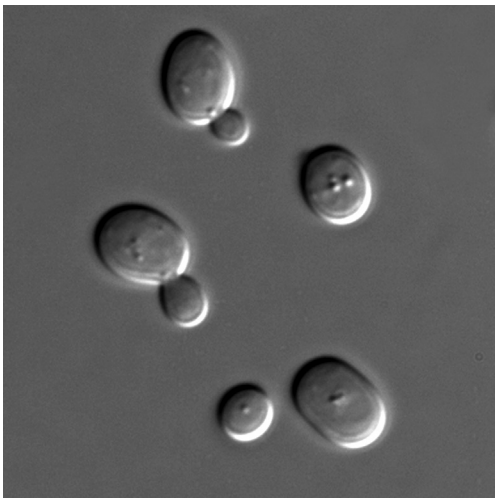

FIG. 5.2 The yeast *Saccharomyces cerevisiae*, a model organism for human disease research. *Source: Wikipedia, entered into the public domain by its copyright holder, Masur.*

retinal stem cells) is a mutation in the RB1 gene. Mutating the homologous gene in the nematode results in ectopic vulvae, a condition that is unmistakenly different from retinoblastoma.[66] Nonetheless, pathways involved in causing retinoblastoma (in humans) and in causing ectopic vulvae (in *C. elegans*) may share many important commonalities, including sensitivity to potentially useful gene-targeted drugs (Fig. 5.3). [Glossary Complex disease, Retinoblastoma, Trilateral retinoblastoma]

The zebrafish (*Danio rerio*) is a species of small freshwater minnow. Their common name comes from the distinctive horizontal stripes on the sides of their bodies, a somewhat inaccurate appellation insofar as most of the stripes on zebras (*Equus quagga*) are nearly vertical. Zebrafish eggs are fertilized outside the mother's body, allowing scientists to inject DNA or RNA into one-cell stage embryos to produce transgenic or knockout strains of zebrafish. Zebrafish is easy to grow, in large or small schools, and their development can be closely monitored, from egg onward (Fig. 5.4).

The zebrafish shares with humans the same evolutionary descent, to the level of Class Euteleostomi. At this point the ancestors of the zebrafish branched to Class Actinopterygii

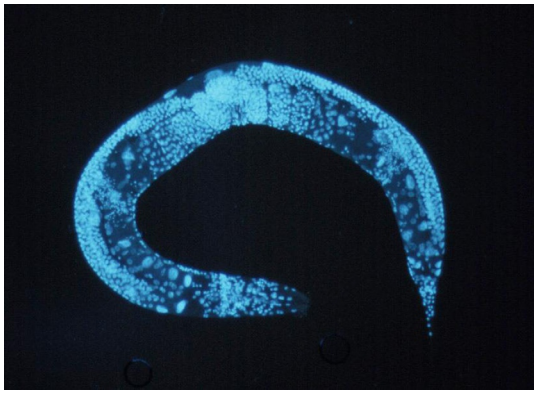

FIG. 5.3 *Caenorhabditis elegans*, a nematode (roundworm) about 1 mm in length. *C. elegans* serves as a model organism for human disease research. *Source: Wikipedia, from a public domain work from the U.S. National Institutes of Health.*

FIG. 5.4 Zebrafish, a model organism for the study of human diseases. *Source: Wikipedia, and entered into the public domain by its author, Azul.*

(i.e., ray-finned fish accounting for at least 30,000 species of extant fish), while the ancestors of humans branched to Class Sarcopterygii (i.e. the lobe-finned fish). En route to Class Euteleostomi, the ancestors of humans and of ray-finned fish descended through Class Metazoa, Class Bilateria, Class Craniata, and Class Vertebrata and Gnathostomata. In doing so, zebrafish acquired nearly all the cell types that are present in humans today. Consequently, zebrafish and humans are not all that distant from one another developmentally, and these two species have similar immunologic systems, central nervous systems, and peripheral nervous systems. This being the case, it should not be surprising to learn that 70% of human genes are found in zebrafish.[70]

In the past few decades, the zebrafish has become a very useful model for studying developmental and disease-related pathways in humans.[71] For example, the zebrafish, as a fellow member of Class Chordata develop chordomas (as do humans). Likewise, as a fellow member of Class Craniata, the zebrafish has served as an appropriate model for human neurocristopathies, including Waardenburg-Shah syndrome and Hirschsprung's disease.[71,72–74] [Glossary Neurocristopathy]

As was the case with *C. elegans*, it is important not to overinterpret experiments using zebrafish. In one large study, a gene in zebrafish was shown to modulate its susceptibility to mycobacterial infection.[75] Naturally, there was hope that the orthologous gene in humans would be associated with human susceptibility to tuberculosis. Despite a large study involving 9115 subjects, no such association was found.[76] Once again we have learned that the genetic root cause of a disease does not account for pathogenesis, which is an emergent property of the system in which the mutation is expressed. [Glossary Multistep process]

Orthodiseases have been found in *Drosophila* (fruit flies). *Drosophila* contains homologs of the genes that cause tuberous sclerosis, a hamartoma cancer syndrome in humans. The brain tubers (hamartomas of the neurectoderm, also called phakomas), for which tuberous sclerosis takes its name, contain large, multinucleate neurons. Loss of function of the same genes in Drosophila produce enlarged cells with many times the normal amount of DNA.[77] The tuberous sclerosis orthodisease in Drosophila is being studied to help us understand cell growth control mechanisms in humans. [Glossary Hamartoma, Orthodisease, Neurectoderm]

Oncogenes and tumor suppressor genes are conserved genes found in many metazoans. We can often find nonvertebrate organisms that recapitulate some of the steps in human carcinogenesis. Loss-of-function mutations in *Drosophila* genes, which are homologous to human tumor suppressor genes, have resulted in neoplasms growing from imaginal disc cells. Hence, we can infer that carcinogenesis involves conserved genes that have been present in very early metazoans (i.e., before the split of Class Deuterostomia, containing humans, and Class Protostomia, containing insects).[78] [Glossary Imaginal disc, Oncogene]

Medical researchers are currently discovering the root mutations responsible for genetic diseases, and they are doing so at a rate that far exceeds our ability to understand how the mutation produces its clinical phenotype. Knowing the operational pathways of

disease development provides us with an opportunity to test new drugs that repair, control, or bypass the metabolic damage caused by a flawed gene. Without this knowledge, we cannot rationally develop new cures for human diseases. **How can we understand the metabolic consequences of thousands of genetic defects known to be associated with the pathogenesis of human diseases, if we continue to rely upon lengthy studies using rodent models that often produce misleading results?** Organisms such as nematodes (e.g., *C. elegans*), the fruit fly (i.e., *Drosophila melanogaster*), the zebrafish (i.e., *D. rerio*), or the yeast (i.e., *S. cerevisiae*) can be propagated, genetically modified, and studied in the laboratory. Experiments on these nonmammalian species are conducted at much lower cost and, in much less time, than comparable experiments on rodents and other mammals.[64] Perhaps more importantly, such studies can be repeated, validated, and extended by other laboratories.

Section 5.6. Inferring the relationships between genetic diseases and their phenocopies

The field of molecular genetics has made great advances in the past few decades. We now know the specific genes that cause hundreds of rare diseases. Useful as this knowledge may be, we must keep in mind that knowing the gene that causes a disease is quite different from knowing how the affected gene causes the disease. In many cases, we simply have no understanding of why a particular mutation, in a particular gene, may produce a disease having a specific array of clinical features. Our ignorance is particularly striking in those cases in which a mutation occurs in a gene coding for a transcription factor. These proteins regulate the expression of other genes. Their influence on our bodies may shift radically at various stages of human development, serving one role in the embryo and another role in the adult organism. Furthermore the biological effect of any given transcription factor may be modulated or offset by the concurrent activities of other genetic modulators. The basic problem here is that genes operate in a complex biological system that changes from moment to moment and from cell to cell. As we have seen (vide supra), when we study the role of orthologous genes in nonhuman species, we can begin to understand how particular genes may participate in the pathways that lead to human disease. In many cases, our most inventive molecular biologists remain baffled and are incapable of drawing any logical connection between an altered gene and the disease that it causes.

Let's look at a series of diseases whose clinical phenotype could not be explained by its root genetic cause, until further study yielded new information about the pathways involved in the disease process.

Ligneous conjunctivitis

Consider the example of the rare, inherited disease, ligneous conjunctivitis. The word "ligneous" means hard, like wood, and the term "ligneous conjunctivitis" refers to a thick, hard coating on the conjunctiva of the eye, extending over the sclera and under the eyelids.

In addition to conjunctivitis, affected individuals may also develop middle ear inflammation, inflammation of the trachea and bronchial tree, and blockage of the flow of cerebrospinal fluid.

Ligneous conjunctivitis is associated with a gene mutation causing a deficiency of plasminogen. How does a deficiency of plasminogen lead to conjunctivitis, middle ear inflammation, tracheobronchitis, and blockages in the cerebrospinal canals? Plasmin, the activated form of plasminogen, breaks down fibrin, a protein produced during coagulation and clot formation. In the absence of plasminogen, fibrin accumulates in various sites, and the accumulating fibrin dries and hardens. On the surface of the eyes, dried fibrin elicits inflammation. Accumulating fibrin in the middle ear and the tracheobronchial mucosa leads to plugging and inflammation at these sites. In the brain, an occlusive hydrocephalus may occur, due to fibrin blocking the normal flow of cerebrospinal fluid through the ventricles.

It would be difficult for anyone to predict that a plasminogen deficiency would produce ligneous conjunctivitis. With the benefit of hindsight, the pathogenesis is obvious.

Severe combined immunodeficiency (SCID)

SCID is an aggregate of diseases with varying genetic causes that are all characterized by immunodeficiencies of both arms of the adaptive immune system (i.e., B cells and T cells) in infants. The second most common cause of SCID is associated with defective adenosine deaminase, an enzyme involved in the breakdown of purines. The possible mechanisms whereby a deficiency of adenosine deaminase leads to SCID have been the subject of much discussion in the literature, extending over several decades. Here is one hypothesis, described in a syllogism:

```
Adenosine deaminase is required for the survival of rapidly
dividing cell populations.

Immunocytes (B and T cells) are rapidly dividing cell populations.

Hence,
Deficiencies in adenosine deaminase produce a depletion of
immunocytes.

Immunocytes are necessary for a normal immune response.

Hence,
Deficiencies in adenosine deaminase produce severe combined
immunodeficiency.
```

The argument seems to make sense, but it fails to explain a deficiency of adenosine deaminase that specifically blocks the normal fetal development of T and B cells when we know that every cell population in the developing fetus is rapidly dividing. Why does an inherited deficiency of adenosine deaminase not cause widespread necrosis of all rapidly dividing tissues, including gut epithelium and bone marrow? [Glossary Aggregate disease]

The answer to this question might come from an understanding of the pathogenesis of some of the other forms of SCID. Individuals with deficiencies of proteins participating in nonhomologous end joining (NHEJ) DNA repair account for several of the subtypes of SCID.[79] It happens that the process of lymphocyte differentiation involves enzymatic, breakage followed by enzymatic repair, of V(D)J recombination units that account for the remarkable diversity of recognition sites that characterize the adaptive immune system. If repair does not occur, then T and B cells cannot properly develop.

Knowing the relationship between DNA repair deficiencies and immune deficiencies, we may be able to shed some light on the pathogenesis of SCID in adenosine deaminase deficiency. When resting blood lymphocytes are treated with deoxycoformycin, an adenosine deaminase inhibitor, single strand breaks accumulate.[80] Hence a relationship between deoxyadenosine deaminase deficiency and DNA repair deficiency is established, suggesting that this form of SCID develops much like the other forms of SCID that result from a DNA repair deficiency.

In this case, our understanding of the pathogenesis of SCID, in several subtypes of the disease, has helped us to hypothesize the common pathogenesis of SCID (i.e., the inability to adequately repair V(D)J recombination units) in an enigmatic subtype.

Alpha-1 antitrypsin deficiency with cirrhosis and emphysema

Alpha-1 antitrypsin deficiency is an example of an inherited monogenic disease causing disorders of two different organs (lungs and liver) through two different mechanisms, neither of which is particularly obvious. To understand this unusual disease, we need to take a moment to discuss neutrophils. Neutrophils are naturally invasive cells, penetrating tissues on their route to an inflammatory focus. To invade tissues, neutrophils produce autodigestive proteases. One of the autodigestive proteases produced by neutrophils is elastase, which specifically digests elastin. Elastin provides the lung with a delicate, loose scaffold that allows the lung to expand and contract with each breath. The elastase produced by neutrophils could easily dissolve the elastin fibers of the lung, producing severe emphysema. That being the case, how do we manage to avoid death by autodigestion?

The liver is the only organ that synthesizes and secretes, into the blood, a protein known as alpha-1 antitrypsin, which is a general purpose antiprotease designed to nullify the effects of circulating proteases produced by the pancreas and by white blood cells. When the circulating levels of alpha-1 antitrypsin fall below a certain threshold, lying somewhere between 15% and 40% of normal levels, elastase produced by neutrophils will begin to destroy the delicate elastic tissue that of the lungs. Hence, we can now see why it is that individuals with alpha-1 antitrypsin deficiency may develop emphysema.

Adults who smoke and who have relatively low levels of circulating alpha-1 antitrypsin are particularly prone to develop emphysema, presumably because cigarette smoke deactivates alpha-1 antitrypsin. Furthermore the chronic inflammatory influence of cigarette smoke on the lungs elicits neutrophils to increase their secretion of elastase, hence overwhelming the little alpha-1 antitrypsin activity that remains. For these reasons, smoking may be hazardous to individuals who have alpha-1 antitrypsin deficiencies, even those individuals who are lucky enough to have mild variants of the disorder.

Alpha-1 antitrypsin is encoded by the SERPINA1 gene. Over 75 pathogenic mutations of this gene have been found, but the severe form of the disease, associated with both lung and liver diseases, is the PiZZ variant. In this form of the disease, an abnormal alpha-1 antitrypsin molecule is produced that is harmfully sequestered inside liver cells, leading eventually to liver cell necrosis, cirrhosis, and, sometimes, hepatocellular carcinoma.[81] Hence, we can see how at least one form of alpha-1 antitrypsin deficiency causes liver disease. [Glossary Cirrhosis]

Because the liver is the only organ that produces the altered protein, a liver transplant prior to the development of emphysema would prevent diseases of the liver and of the lung. In its most severe form, the disease may affect infants and may require liver transplantation. In many instances, however, affected individuals will have normal lung and liver function well into adulthood. Such cases, if screened and diagnosed early, provide ample opportunities for interrupting the pathogenic events and steps that precede the development of lung disease, cirrhosis, and hepatocellular carcinoma. For example, an autophagy-enhancing drug has been shown to promote the degradation of the altered alpha-1 antitrypsin that accumulates in liver cells, hence reducing subsequent hepatic fibrosis.[82]

In the case of alpha-1 antitrypsin deficiency disease, knowing the genetic defect underlying the disorder was not very helpful. We needed to study the complex pathogenesis of the disorder before we could be of much assistance to affected individuals. In point of fact, we do not understand the pathogenesis for the great majority of simple, monogenic diseases whose root genetic defect is known. This being the case, how can we ever hope to understand and cure genetic diseases? We must not be discouraged. It happens that circumstances have provided us with a simple way to understand genetic mechanisms of disease that essentially bypasses the field of genetic analysis: the phenocopy.

Phenocopies are acquired diseases that closely mimic inherited genetic diseases. The root cause of a phenocopy diseases is often one particular environmental agent that sets into motion a sequence of events leading to the set of morphologic and clinical features found in its genetic counterpart.

If a phenocopy disease has the same clinical phenotype as its genetic equivalent, then how do we identify a phenocopy when it occurs? Here are a few of the clues that help us distinguish a phenocopy from its genetic counterpart.

1. Absence of disease in family members
2. May develop in a cluster of unrelated individuals or in one particular geographic location
3. Lacks a causal mutation (obviously)
4. Not rare (i.e., more common than its rare genetic equivalent)
5. Often associated with an identified causal agent, such as a drug or environmental toxin
6. Can occur in any age group, with many occurring in middle-aged and elderly individuals
7. May often involve organs known to metabolize or process ingested chemicals (e.g., liver, kidney, and lungs)
8. May have rapid onset of symptoms
9. May have reversible clinical course (i.e., some patients recover normal health)

Most of these listed features of phenocopy diseases are self-explanatory. Point 4 asserts that phenocopy diseases are more common than their genetic equivalents, and the reason that this is true is not obvious. We must remember that genetic diseases typically involve specific mutations in a specific gene, and the likelihood that such a mutation occurs in an individual is small. Hence, most genetic diseases are rare. Phenocopy diseases typically occur when individuals are exposed to a specific agent. The only limit to the number of people affected by a phenocopy diseases is the level of exposure and the size of the at-risk population. Empirically, we can confirm that phenocopies are more prevalent than their genotypic equivalents. This point is discussed at length in Volume II, Section 1.5, "Compositionality: why small outnumbers large."

Point 6 asserts that many phenocopies occur in middle-aged and elderly individuals. At low doses of environmental toxins, it may take many years for the cumulative effects to take their toll on the human body. Put simply, genes account for diseases in the young; environment accounts for their phenocopies in the older population. We can often observe the switchover from genetic diseases to phenocopies, as we age. In the case of all of the inherited cancer syndromes, genes account for the cancers that occur in childhood and early adulthood. As age increases, genes account for fewer and fewer of the cancers. By late adulthood, phenocopies dominate.[83,84]

Point 8 asserts that phenocopy diseases may have rapid onset of symptoms. Many environmental agents have a direct action on cellular constituents. In such cases, we would expect to see rapid onset of symptoms and clusters of reported cases in high-risk locations.

Point 9 asserts that phenocopies are sometimes reversible. Patients may fully recover when we eliminate the causative factor, or we introduce some preventative measure, or we reduce the level of exposure.[85] The genetic diseases are characterized by genes that persist in every cell of the organisms, through the life of the organisms; hence, genetic diseases cannot be easily reversed. There are exceptions of course. Cumulative environmental exposures that produce chronic damage, particularly to nondividing cell populations with limited repair capacity (e.g., neurons) are not generally reversible. Examples might include radiation damage, postradiation ischemia, solar elastosis, heavy metal poisoning to nerves and brain, asbestos-induced mesothelioma, and postinfluenza encephalopathy.

A noteworthy exception to point 2 (i.e., that only phenocopy diseases occur in one particular geographic location) is found in Tangier disease. Tangier disease is endemic to individuals living on Tangier Island, in the Chesapeake Bay. This disease is characterized by hypercholesterolemia, abnormal deposits of cholesterol in various tissues (producing big yellow tonsils and multiple xanthomas of the skin) and a propensity for heart attacks occurring at an early age. The high rate of occurrences of Tangier disease, on one island, would suggest a dietary cause (e.g., sedentary life and shellfish diet). Not so. Tangier disease is a rare genetic disorder, with an autosomal recessive inheritance pattern, characterized by very low levels of high-density lipoproteins.[86] It occurs in high frequency on Tangier Island, because this small island community is highly insular, and has a limited gene pool. A clue to the genetic origin of Tangier disease comes from knowing that cases

of Tangier disease are not restricted to Tangier Island. Like all genetic diseases, isolated cases can occur anywhere on earth, because genetic mutations do not respect geographic boundaries.[87]

Compare the list of features of phenocopy diseases with those of diseases with the same phenotype but whose root cause is a mutation in a particular gene[88]:

1. Familial.
2. Usually progressive and irreversible unless there is medical intervention.
3. Early age of onset.
4. Fatal genetic diseases are seldom common. Common genetic diseases are seldom fatal.
5. Types of organ dysfunctions that are otherwise rare in the general population (tremors, nystagmus, and wasting).
6. Multiple dysmorphisms and multiple organ abnormalities may be encountered.
7. Neurocognitive impairment is sometimes present.

Of the seven listed features of genetic diseases, points 1 to 3 are self-evident. Statement 4, "Fatal genetic diseases are seldom common. Common genetic diseases are seldom fatal," is the result of natural selection. Lethal disease genes tend to eliminate themselves from the gene pool, insofar as affected individuals are unlikely to procreate. The genes causing the most common genetic diseases (e.g., thalassemia, sickle cell disease, and hemochromatosis) often have little deleterious effect in the heterozygous state. In the homozygous state, they may shorten the lifespan of affected individuals, but they are otherwise compatible with life.

Point 7 asserts the empiric observation that genetic diseases are often characterized by cognitive impairment.[88] In addition to many known single gene disorders that are associated with cognitive impairment, virtually every so-called genomic disorder (i.e., disorders of structural rearrangements or deletions or duplications of chromosomes) is associated with cognitive deficits.[89] In addition, conditions that produce deleterious effects on regulators of genomic activity are likely to be manifested, in part, by neurologic deficits.[90–93] Empirically, we find that most of the toxins known to alter the epigenome produce cognitive impairments.[55] At this time, we have no satisfactory explanation for the close association between genomic/epigenomic alterations and cognitive impairments. We can only guess that cognition is complex, requiring the cooperation of multiple pathways performing at high efficiency. We might expect that small perturbations of any of these pathways may reduce cognitive function overall. Alternately, cognitive impairment, resulting from genomic or epigenomic alterations, may be no more prevalent than impairments in organs other than the brain. Perhaps, our expectations of behavioral "normality" are so finely honed that small cognitive impairments draw our attention, while small deficits in organ function go unheeded. In any case the frequent association of cognitive impairments with genomic and epigenomic alterations is a deep mystery.

Many of the most common human diseases are polygenic.[94] This means that multiple genes are involved, with no single gene acting as the root cause of the disease. Many of the polygenic diseases have an acquired component (i.e., some factor or factors from the environment that is involved in pathogenesis) or are triggered by some specific environmental

event (e.g., an allergen, dehydration, extreme cold). The polygenic diseases cannot always be distinguished from phenocopies of rare diseases, but the following features usually apply to polygenic diseases. [Glossary Polygenic disease]

1. Nonfamilial
2. Very common; the most common killers of humans (heart disease and cancer) being polygenic
3. Occurs in an older population than we see for the Mendelian-inherited diseases [Glossary Mendelian inheritance]
4. Typically sporadic [Glossary Sporadic, Sporadic disease versus phenocopy disease]
5. Often triggered by an environmental factor
6. Typically involve mutations in noncoding regions of the genome [Glossary Noncoding mutational diseases]

Point 6 asserts that polygenic diseases often involve mutation in noncoding regions of the genome. This finding, which arose from data collected in many different genome-wide association studies, was a surprise discovery and deserves some explanation. In the monogenic disorders a single altered gene is the root cause of each clinical disease. When there are multiple genes involved in the pathogenesis of disease, we can infer that each of the involved genes contributes to the disease phenotype, but none of the involved genes is sufficient by itself to cause the disease phenotype. This is true because if any of the involved genes were sufficient to cause the disease, then the disease would be monogenic. Presumably, all of the genes in a polygenic disease produce a small effect, and the combination of these small effects is deleterious to the organism. Small effects to the gene often come in the form of modifications to function, not elimination of functions, and it is the noncoding regions of the genome that hold the various modifiers of gene activity. **Hence the common diseases of humans, which tend to be polygenic, are associated with mutations and DNA variations in noncoding regions**.

These listed features of genetic diseases, phenocopies, and polygenic diseases are intended as generalizations that might suggest one category of disease over another and should not be construed as rules.

In the case of the phenocopy, we can expose an organism to the agent and monitor the changes that ensue in cells and tissues. In many cases, we can watch what happens when the agent is removed and the organism recovers. Eventually, we can determine the key pathways involved in the phenocopy disease. From there, we might investigate how an alteration in a specific gene might produce a cascade of events that recapitulates the disease pathways operative in the phenocopy disease. [Glossary Pathway-driven disease]

When we stop and compare the genetic diseases, as a group, with their phenocopies, we learn a great deal about the pathogenesis of both forms of the disease. Here are a few observations that seem to hold for most phenocopies.

1. There is usually one dominant pathway that drives the clinical expression of the phenocopy. As we would expect, the pathway disrupted in the phenocopy disease is almost always the same pathway that is disrupted in the genetic form of the disease.

Hence the phenocopy tells us how the rare disease expresses itself, and this is something that we can seldom infer from our knowledge of the gene mutation associated with the rare disease.

2. Pharmacologic treatments for the phenocopy disease may apply to pathways operative in the genetic form of the disease.
3. When the gene responsible for an inherited disease is unknown, the careful study of its phenocopy will often suggest a set of candidate genes, operative in a pathogenetic pathway, any one of which may serve as the root cause of the inherited disease.
4. Recognizing the cause of a phenocopy disease may curtail potential environmental catastrophes. The phenocopy diseases help us to focus on the cellular pathways leading to disease. If we exclusively study the genetics of disease, we will likely miss out on such opportunities.
5. Phenocopies demonstrate the various ways that heritable and environmental factors can contribute to the converging pathways of related diseases.
6. The pathway involved in a phenocopy disease can contribute to the pathogenesis of any of the common diseases that have phenotypic overlap with the phenocopy disease. This is true because there are a limited number of symptoms and clinical features of observed diseases. Hence, when a particular biological feature is present in a genetic disease, its phenocopy, and in a common disease, we can guess that all three disorders may involve the same metabolic pathway.
7. Genetic diseases and their phenocopies mimic one another. Hence a sudden increase in the occurrences of what is thought to be a genetic disease, within a population of unrelated individuals, should prompt a thorough search for an acquired phenocopy.
8. The phenocopy diseases remind us that we can have a disease without a causal gene, but we cannot have a disease without a causal metabolic pathway. Hence, when a genetic root cause of a disease has been discovered, we must not assume that every instance of the disease has a genetic cause (i.e., some instances may be phenocopies). Hence, phenocopies should be included in the differential diagnosis of every suspected genetic disease. This is particularly the case in those diseases that present themselves in adults, insofar as most Mendelian-inherited genetic diseases are usually manifested in infancy, childhood, or adolescence.

In case you are wondering whether phenocopy diseases are common, please be advised that we can usually find them, if we look for them. Here is just a partial listing, wherein the phenocopy is listed on the left, followed by its genotypic twin:

Acquired conduction defect	**Inherited conduction defect**
Acquired porphyria cutanea tarda	Inherited porphyria cutanea tarda
Acquired von Willebrand disease	Inherited von Willebrand disease
Aminoglycoside-induced hearing loss	Inherited mitochondriopathic deafness

Continued

Acquired conduction defect	Inherited conduction defect
Antabuse (disulfiram) treatment	Inherited alcohol intolerance
Drug-induced methemoglobinemia	Inherited methemoglobinemia
Fetal exposure to methotrexate	Miller syndrome[95]
Methylmalonic acidemia caused by severe deficiency of vitamin B12	Inherited methylmalonic acidemia
Osteolathyrism and scurvy	Inherited collagenopathies [Glossary Collagenopathy]
Alcohol-induced sideroblastic anemia	Inherited sideroblastic anemia
B12 deficiency	Inherited pernicious anemia
Cardiomyopathy due to alcohol abuse	Inherited dilated cardiomyopathy[96]
Lead-induced encephalopathy	Inherited tau encephalopathy[97]
Myopathy produced by nucleoside analog reverse transcriptase inhibitors (i.e., HIV drugs)	Inherited mitochondrial myopathy
Pseudo-Pelger-Huet anomaly	Inherited Pelger-Huet anomaly[98,99]
Thalidomide-induced phocomelia	Roberts syndrome and SC pseudothalidomide syndrome[100]
Warfarin embryopathy	Brachytelephalangic chondrodysplasia punctata[101]
Drug-induced cerebellar ataxia[102]	Hereditary spinocerebellar ataxia[103]
Copper poisoning	Wilson disease
Quinacrine-induced ochronosis	Inherited ochronosis (i.e., mutation in the HGD gene for the enzyme homogentisate 1,2-dioxygenase)[104]
Drug-induced Parkinsonism[105,106]	Autosomal-dominant inherited Parkinsonism[107]
Acquired pulmonary hypertension due to hypoxia, thromboembolism, left-sided heart failure, or drugs	Inherited pulmonary hypertension[108]
Amphotericin toxicity	Inherited renal tubular acidosis[109]
Acquired platelet storage pool deficiencies	Hereditary platelet storage pool deficiency[110]
Acquired porphyrias	Inherited porphyrias[111,112]
Acquired iron overload and hemochromatosis	Inherited hemochromatosis
Acquired cirrhosis	Inherited cirrhosis (due to mutation in keratin 18)[113]
Anticoagulant drugs that inhibit thrombus formation	Inherited factor X deficiency (a form of hemophilia)

The phenocopy diseases are seldom taught to medical students or to biomedical scientists, but they should be. In some cases, they provide the means by which we may understand, diagnose, prevent, and cure human diseases.

Section 5.7. The logic of treating disease pathways, not disease genes

Before the 20th century, how did physicians treat disease? Traditional remedies for disease were based on providing a drug that had the opposite effect on the body as the symptoms of the disease. This strategy was intended to provide some relief from the illness, while the

body healed itself. For many diseases the old ways of treatment were reasonably effective. For example, foxglove, the source of digitalis, increases the contractility of the heart. Hence, digitalis provides some benefit to every cardiac disease characterized by reduced cardiac contractility. Likewise, warfarin, a natural anticoagulant originally found in spoiled sweet clover, was useful for thrombotic diseases. Atropine, from belladonna, the deadly nightshade plant, was a naturally occurring drug that depressed the parasympathetic neurons and could be used for any disease associated with an overactivity of the parasympathetic system. Of course aspirin, an antiinflammatory drug brewed from willow bark, was used since antiquity to soothe any and all inflammatory diseases.

In all these cases, one drug was useful for all the diseases that presented with the same set of symptoms. The drawback of this approach was that the drugs, for the most part, treated symptoms without actually curing the disease. Physicians largely depended on time and the natural healing processes of the body to provide the cure. Modern medicine is following a new strategy in its battle against disease. It is basically the following:

1. Determine the gene that causes the disease.
2. Develop a drug that targets, in some fashion, the disease gene.
3. Design a clinical trial to see if the drug is curative.

Simple enough. Some of the earliest and most successful therapies developed in the past two decades have targeted specific mutations occurring in specific subsets of diseases. One such example is ivacaftor, which targets the G551D mutation present in about 4% of individuals with cystic fibrosis.[114] It is seldom wise to argue with success, but it must be mentioned that the cost of developing a new drug is about $5 billion.[115] To provide some perspective, $5 billion dollars exceeds the total gross national product of many countries, including Sierra Leone, Swaziland, Suriname, Guyana, Liberia, and the Central African Republic. Many factors contribute to the development costs, but the most significant is the incredibly high failure rate of candidate drugs. About 95% of the experimental medicines that are studied in humans fail to be both effective and safe. The costs of drug development are reflected in the rising costs of drugs.

When a new drug is marketed to a very small population of affected individuals, the cost of treating an individual may be astronomical. Americans should not pin their hopes on the belief that one day, the FDA or CMS (which administrates Medicare) will step in and put a stop to the price rises. The Food and Drug Administration can approve or reject drugs, but it does not regulate prices. Likewise, Medicare is not permitted to consider cost when it decides whether a treatment can be covered. Knowing this, some notable pharmaceutical companies have raised the prices of medications far beyond their manufacturing costs.[116–118] In effect the cost of curing curable diseases may exceed our ability to pay for those cures.[118]

It is in the interests of society to develop drugs that have the widest possible user market.[119] Developing drugs that target a mutation that is specific for a few individuals with a

rare disease, or a tiny subpopulation of individuals who have a common disease, may not be the most effective strategy.

Disease phenotypes result from the development of altered metabolic pathways active in the cell types that participate in lesions (i.e., the tissues affected by disease). Even in those conditions whose root cause is a specific mutation in a specific gene, the consequent clinical disease can seldom be accounted for solely by any one mutation. We know this because rare diseases that exhibit genetic heterogeneity (i.e. can be caused by a variety of possible mutations) all tend to involve the same metabolic pathways. Likewise, phenocopies of genetic diseases often involve the same pathways that drive their genetic counterparts, without actually involving the protein encoded by the root cause of the genetic form of the disease. We also know that the acquired version of most genetic diseases account for the bulk of disease occurrences. **Hence, if we want to develop treatments that benefit the greatest number of individuals affected by a disease, it would be far more practical to find treatments that target the disease-driving pathways than to design drugs that target a specific gene involved in a small subset of affected patients**.

Here are just a few examples where we have benefited from pathway-targeted drugs.[55]

Losartan as blocker of TGF-beta

Losartan is an angiotensin II receptor antagonist drug used in the treatment of hypertension. More recently, it has been shown that losartan blocks signaling by transforming growth factor-beta (TGF-beta). TGF-beta signaling is involved in fibrogenesis and inflammation and is a convergent pathway for a variety of diseases that were previously considered unrelated. Consequently, losartan is being developed as a potential treatment for Marfan syndrome,[120,121] epilepsy,[122] hypertrophic cardiomyopathy, other myopathic disorders and various inflammatory conditions.[123,124]

Losartan is an FDA-approved drug, currently available as an inexpensive generic. Losartan has been used daily, for years, in millions of patients, for the treatment of hypertension. Healthcare professional have wide experience with losartan's side effects. Hence, clinicians are eager to adopt losartan's new uses, based on updated knowledge of the converged pathways targeted by the drug.

Botox for muscle spasms

The sudden release of excess amounts of acetylcholine at neuromuscular junctions may cause muscle spasms. *Clostridium botulinum* toxin A (botox) blocks the nerve from releasing acetylcholine and relaxes the spasm. Hence, we would expect botox to be effective in a variety of conditions whose pathways converge to produce muscle spasms. This would include cerebral palsy, movement disorders characterized by muscle overactivity and spasticity, spasmodic torticollis, strabismus, local dystonia including laryngeal dystonia, blepharospasm (spasmic eye closure or blinking), and hemifacial spasm.

Angiogenesis inhibitors

Bevacizumab is an angiogenesis inhibitor (i.e., it reduces the formation of new small vessels). All cancers, at some point in their pathogenesis, must vascularize themselves, to deliver oxygen to tumor cells. Hence an angiogenesis inhibitor, such as bevacizumab has the potential of serving as a universal anticancer drug, effective against any kind of cancer, regardless of which specific genetic errors drives its proliferation. Bevacizumab is currently employed in the treatment of common cancers, including cancers of the colon, lung, breast, kidney, ovaries, and brain (i.e., glioblastoma). Bevacizumab produces tumor shrinkage in more than half of vestibular schwannomas occurring in Neurofibromatosis 2.[125] [Glossary Schwannoma]

As you might expect, Bevacizumab has value in treating diseases other than cancer, for which angiogenesis occurs. Two noncancerous diseases of neovascularization (i.e., diseases caused in whole or in part by overgrowth of new vessels) and treated with angiogenesis inhibitors are hereditary hemorrhagic telangiectasia[126] and various forms of ocular neovascularization, including common age-related macular degeneration.[127]

MTOR pathway inhibitors

Tuberous sclerosis is an inherited disease that is characterized by the early development of hamartomatous growths of varying types, in many different organs, including the brain, kidneys, lungs, heart, skin, eyes, and pancreas. Hamartomas are benign, malformative overgrowths of tissues. The tubers of tuberous sclerosis, from which the disorder takes its name, are focally thickened, pale gyri of the brain cortex. Seizures, intellectual delay, and autism are commonly found in individuals with this disorder.

The root genetic mutation associated with the majority of cases of tuberous sclerosis is a mutation of either the TSC1 gene, coding for the protein hamartin, or in the TSC2 gene, which codes for the protein tuberin. Either of these mutations result in the hyperactivation of the mammalian target of rapamycin (mTOR) signaling pathway. Consequently, inhibitors of the convergent mTOR pathway, such as sirolimus, everolimus, and rapamycin, are considered potential drugs for managing the growth of hamartomas and developmental disorders arising from mTOR pathway hyperactivity in this inherited syndrome.[128]

As it happens, the mTOR signaling pathway ties into several other important pathways, and the mTOR inhibitors have been proposed as potential drugs in the treatment of various diseases in which mTOR participates, including Alzheimer's disease, cancers, and aging-related disorders, and as a preventive measure against transplant rejection.[129]

Unidentified pulmonary hypertension pathway

Pulmonary hypertension can occur as an inherited condition or as an acquired condition following hypoxia, thromboembolism, or left-sided heart failure, or may be caused by various drugs. Regardless of the cause, all forms of pulmonary hypertension seem to be associated

with one signaling pathway involving several proteins, including angiopoietin-1, TIE2 (endothelial-specific receptor for angiopoietin-1), bone morphogenetic protein receptor 1A and 2 (BMPR1A and BMPR2).[108] If this is the case, then any of these proteins would serve as candidate targets for new drugs that may prove to be effective against all forms of pulmonary hypertension.

CD1a as a therapeutic target in inflammatory skin diseases

The CD1a molecule is expressed abundantly in Langerhans cells, a reticuloendothelial cell that inhabits skin. Some types of inflammatory responses in skin are triggered by CD1a, and these happen to include the response pathway observed in poison ivy allergy (i.e., reaction to the plant-derived lipid urushiol) and in psoriasis. Treatment with antibodies blocking CD1a reduced the inflammatory responses in both diseases.[130] Hence, drugs targeted to CD1a may be potentially useful for any inflammatory diseases wherein the CD1a-initiated inflammatory pathway operates.

JAK2 inhibitors for myeloproliferative disorders

Janus kinase genes (e.g., AK1, JAK2, JAK3, and TYK2) influence growth and immune response in various types of blood cells, through their effect on cytokines. Mutations of the JAK2 gene are involved in several myeloproliferative conditions, including myelofibrosis and polycythemia vera, and at least one form of hereditary thrombocythemia.[131–133] Inhibitors of JAK genes have been approved for the treatment of a wide range of hematologic disorders characterized by proliferating blood cells, including myeloproliferative disorders and immunologic reactions. For example, Ruxolitinib has been approved in the United States for use in psoriasis, myelofibrosis and rheumatoid arthritis.[134] A host of JAK pathway inhibitors is either approved or under clinical trial for the treatment of allergic diseases, rheumatoid arthritis, psoriasis, myelofibrosis, myeloproliferative disorders, acute myeloid leukemia, and relapsed lymphoma.[135] [Glossary Myelofibrosis, Polycythemia vera]

The specific JAK2 mutation observed in some, but not all, myeloproliferative neoplasms is JAK2V617F.[136,137] Because the JAK2 inhibitors currently in use cannot discriminate between wild-type and mutant JAK2 enzymes, they exert an effect on all proliferating hematopoietic cells, neoplastic or normal, and relieve the debilitating conditions that accompany myeloproliferative disorders, such as splenomegaly and constitutional symptoms that result from inflammation triggered by blood cells (e.g., fever).[138] This serves as an example where drugs targeting a pathway (i.e., the JAK2 pathway) and not the specific mutant protein (i.e., JAK2V617F) provide a wide array of benefits to individuals with a range of JAK2-related diseases. [Glossary Wild-type gene]

Pembrolizumab for any tumor that has microsatellite instability high (MSIH) or is mismatch repair deficient (dMMR) [Glossary Microsatellite, Microsatellite instability]

Regardless of type of tumor, if the tumor demonstrates high microsatellite instability or is mismatch repair deficient, then the tumor may respond to treatment with

pembrolizumab.[139] Microsatellite instabilities and mismatch repair deficiencies are commonly found in colorectal, endometrial, and gastrointestinal cancers.

Cytokine storm inhibitors

Hemophagocytic lymphohistiocytosis is a rare condition characterized by widespread proliferation of lymphocytes and the engorgement of macrophages by red blood cells. This condition is always life threatening and can quickly progress to hyperpyrexia, shock, and multiorgan failure. Hemophagocytic lymphohistiocytosis can occur in genetic form, in infants, or as an acquired condition following infections (e.g., Epstein-Barr virus) or may occur in association with several genetic diseases.[140,141] Every form of hemophagocytic lymphohistiocytosis, regardless of the different pathogenesis, will converge to a pathway that calls into action an inflammatory response characterized by the secretion of large amounts of cytokines, vividly referred to as a cytokine storm. Knowing the converged pathway common to all cases of hemophagocytic lymphohistiocytosis provides us with a therapeutic opportunity. Drug development for the treatment of the convergent cytokine storm pathway is an active area of research that may prove beneficial for the development of various severe inflammatory disorders, including coronavirus infections.[53,54] One such drug is colchicine, which has been successfully used to quell the exaggerated cytokine response in gout and in Familial Mediterranean fever. [Glossary Familial hemophagocytic lymphohistiocytosis]

C-KIT inhibitors

Gastrointestinal stromal tumor (GIST) is a soft tissue tumor that often contains a mutation in the c-KIT gene.[142] As it happens, not all GISTs have the c-KIT mutation.[143,144] An alternate mutation, in the platelet-derived growth factor receptor-alpha gene (PDGFR-alpha), was shown to be the root cause of a minority of GIST cases. Mutations in the gene coding for PDGFR-alpha or c-kit proteins led to identical GIST tumors, causing activation of the same tyrosine kinase pathway. Most importantly, GISTs associated with either gene benefited from imatinib treatments.[145]

Imatinib (trade name Gleevec) inhibits tyrosine kinase, an enzyme involved in a pathway that drives the growth of various rare tumors and proliferative diseases (e.g., chronic myelogenous leukemia, gastrointestinal stromal tumor, hypereosinophilic syndrome).[143,146–149] Pathways with increased tyrosine kinase activity and pathways whose tyrosine kinase activity is particularly sensitive to the inhibiting action of imatinib would make the best drug targets. Because imatinib is targeted to a key protein in a general pathway that contributes to a proliferative phenotype, its use may be of benefit in a variety of different diseases. [Glossary Rare cancer]

In all these examples, major advances in treatment followed an approach that targeted general metabolic pathways involved in a variety of diseases. None of these treatments targeted the specific gene that was involved in one genetic disease or in one subset of one genetic disease. The goal was to find a general disease pathway whose activity could be modified by a drug that was effective against several or many related diseases.

Glossary

Aggregate disease A condition that includes multiple disorders all having the same clinical phenotype. For example, chronic obstructive pulmonary disease (COPD) has been called an aggregate of many "small COPDs" representing individual diseases that happen to be difficult to distinguish from one another.[150] Aggregate diseases would include all of the so-called end-stage conditions (e.g., end-stage kidney and end-stage heart) produced by long-standing or chronic pathologic processes that irreversibly diminish the organ's function, often replacing most of the organ with fibrous tissue. The majority of common diseases (e.g., heart attacks, asthma, common cold, and constipation) are aggregates of conditions that have many different pathogeneses.

Alternative RNA splicing A normal mechanism whereby one gene may code for many different proteins.[151] In humans, about 95% of genes that have multiple exons are alternately spliced. It has been estimated that 15% of disease-causing mutations involve splicing.[152,153] Cancer cells are known to contain numerous splicing variants that are not found in normal cells.[154,155] Normal cells eliminate most abnormal splicing variants through a posttranscriptional editing process. Alternative RNA splicing may result from mutations in splice sites or from spliceosome disorders.

In hereditary thrombocythemia, characterized by in overproduction of platelets, there is a mutation in the gene coding for thrombopoietin protein. This gene mutation leads to mRNAs with shortened untranslated regions that are more efficiently translated than the transcripts that lack the mutation. This, in turn, causes the overproduction of thrombopoietin, which induces an increase in platelet production.[156]

Anonymous variation A genetic variation for which there is no change in gene function. Today the bulk of the 3 billion base-pair sequence comprising the human genome cannot be assigned to any particular function; a randomly occurring mutation is likely to be anonymous; hence, it is assumed that most SNPs are anonymous. Other commonly encountered anonymous markers include the microsatellites, for which there occur variations in the length of repeated sequences within the microsatellites, but these variations cannot be assigned to a gene or to a particular function. Mutations that occur in postmitotic cells (i.e., cells that will never divide) are, for all practical purposes, anonymous and undetectable. A mutation must be passed to a population of progeny cells before it can do much damage and before it can be detected by current molecular biological techniques. Some types of mutations are difficult to find, even when they occur in large numbers of cells. For example, when a mutation consists of a duplicated chromosome, the alteration cannot be detected by methods that find point mutations (because there aren't any).

Bioinformatics The science of the curation and analysis of biological data. The field of bioinformatics had focused on genomic data for several decades. Recently the field has expanded its purview into epigenomics, proteomics, metabolomics, and so on.

The term "bioinformatics" should not be confused with "biomedical informatics," the latter referring to applications of information systems (e.g., computers and hospital databases) to the practice and the science of medicine.

Cambrian explosion Studies of shale strata indicate that something very special happened, in the history of terrestrial animals, in a relatively short period, stretching from about 550–500 million years ago. In this 50 million year span, nearly all of the major phyla of animals that we see today came into existence. We call this era the Cambrian explosion. The word "Cambrian" is a Latinized form of the Welsh language word for Wales, where shale deposits of the Cambrian age were first studied. The word "explosion" tells us that paleontologists have come to think of a span of 50 million years as a blinding flash in earth's history.

Our understanding of the major classes of animals is based almost entirely on fossils found in shale. Animals certainly preceded the Cambrian period, but such animals were soft and uncalcified and would be underrepresented in the fossil record.[157,158] Furthermore, our reliance on body plans, as the only measure of a phyla's emergence, is somewhat presumptuous. It may very well be that the

defining expression of phyla may not have developed until well after the Cambrian explosion.[159–161] In particular the bryozoans (tiny invertebrate aquatics that filter food particles from water) seem to have arisen sometime after the Cambrian.

Capsid The protein shell of a virus that encloses the genetic material of the virus when the virus is outside its host cell. The capsid aids the virus with its attachment to the target host cell and with the penetration of the viral genome into the host cell.

Cilia Many terrestrial organisms contain fine, hair-like cell protrusions that wave back and forth. This wave movement can propel a cell, move food particles toward the cell, or push objects away from the cell. The cilia of respiratory lining cells work in unison, to flush mucus and particulate matter up and out of the tracheobronchial tree.

Cirrhosis A liver condition in which fibrous tissue proliferates in liver acini (the glandular portion of the liver) and around ducts and vessels. This results in the death of liver cells and a reactive overgrowth of hepatocytes with distorted and enlarged liver acini. These changes result in a marked decrease in liver function and a subsequently high mortality rate.

Liver cirrhosis predisposes to liver cancer. In the United States, about 80%–90% of hepatocellular carcinomas cases arise in cirrhotic livers.

Classification system versus identification system It is important to distinguish classification systems from identification systems. An identification system matches an individual object (e.g., organism and disease) with its assigned name. In the case of organisms, identification is based on finding several features that, taken together, can help determine the species name of an organism. For example, if you have a list of characteristic features like large, hairy, strong, African, jungle dwelling, and knuckle-walking, you might correctly identify the organisms as a gorilla. These identifiers are different from the phylogenetic features that were used to classify gorillas within the hierarchy of organisms (Animalia, Chordata; Mammalia, Primates; Hominidae, Homininae; Gorillini, Gorilla). Specifically, you can identify an animal as a gorilla without knowing that a gorilla is a type of mammal. Conversely, you can classify a gorilla as a member of Class Gorillini without knowing that a gorilla happens to be large.

One of the most common mistakes in the biological sciences is to confuse an identification system with a classification system. The former simply provides a handy way to associate an object with a name; the latter is a system of relationships among objects.

Collagenopathy A variety of clinical conditions involving genetic alterations of the various collagen genes or other genes involved in the complex process of collagen synthesis. Examples are Ehlers-Danlos syndrome, osteogenesis imperfecta, familial aneurysmal disorders or aortic dissection disorders, Caffey disease (infantile cortical hyperostosis), and Bruck syndrome. Some of the noncollagen genes involved in collagenopathies include the ACTA2 gene (thoracic aortic aneurysms and aortic dissection), PLOD2 gene (procollagen-lysine dioxygenase 2 involved in Bruck syndrome), and genes associated with familial aneurysms (e.g., smad3, tgfbr1, tgfbr2, and tgfb2).

Complex disease A somewhat vague term generally indicating that the pathogenesis of a disease cannot be understood. The presumption is that our lack of understanding results from the many different factors that contribute to the development of the disease over time. We settle on the idea that it's all just too much to grasp.

When the development of a disease involves numerous environmental factors, some known and others assumed, as well as multiple genetic and epigenetic influences, we have no way of fully understanding how all of these factors interact with one another, and we have no way of fully describing the biological steps that lead to the clinical expression of disease. Likewise, we have no way of predicting how different individuals with a complex disease will respond to treatment.

The term "common disease" is nearly equivalent to the term "complex disease." Nearly all the common diseases of humans (e.g., cancer, heart disease, and asthma) are complex, and nearly all the complex diseases are common.

CpG island DNA methylation is a form of epigenetic modification (i.e., modifications of the genome that do not change the nucleotide sequence of the genome). The most common form of methylation in DNA occurs on cytosine nucleotides, most often at locations wherein cytosine is followed by guanine. These methylations are called CpG sites.

CpG islands are concentrations of CpG sites having a GC content over 50% and ranging from 200 base pairs to several thousand base pairs in length. There are about 29,000–50,000 CpG islands, and most of these are associated with promoters.[162] Various proteins bind specifically to CpG sites. For example, MECP2 is a chromatin-associated protein that modulates transcription. MECP2 binds to CpGs; hence, alterations in CpG methylation patterns can alter the functionality of MECP2. Mutations in MECP2 cause RETT syndrome, a progressive neurologic developmental disorder and a common cause of mental retardation in females. It has been suggested that the MECP2 mutation disables normal protein-epigenome interactions.[163]

DNA methylation DNA methylation is a chemical modification of DNA that does not alter the sequence of nucleotide bases. It is currently believed that DNA methylation plays a major role in cellular differentiation, controlling which genes are turned on and which genes are turned off in a cell, hence determining a cell's "type" (e.g., hepatocyte, thyroid follicular cell, and neuron). Because cells of a particular cell lineage divide to produce more cells of the same lineage, DNA methylation patterns must be preserved with each somatic cell generation. About 1% of DNA is methylated in human somatic DNA, and DNA methylation occurs preferentially on cytosine, most often at sites where cytosine is followed by guanine (designated as "CpG").

DNA polymorphism Differences in the sequence among gene alleles or genomic segments occurring among various members of a population (i.e., variants in the species gene pool). DNA polymorphisms usually consist of small differences in nucleotide sequence or to variable numbers of repeated nucleotide units

Driver pathway A metabolic pathway that develops during the pathogenesis of disease and that persists to play a necessary role in the clinical expression of the disease. A "driver pathway" is distinguished from a passenger pathway, the latter being a pathway that plays a role in the pathogenesis of the disease, but does not serve to maintain the clinical phenotype after the disease has fully developed.

The distinction between driver pathways and passenger pathways may have therapeutic importance. A driver pathway, even if it is present in a small portion of people affected with a disease, is likely to be a valid therapeutic target in the subset of patients who are shown to have the driver pathway. A passenger pathway would be the target for agents that prevent the development of the disease.

Druggable driver A driver pathway that serves as a molecular target for a therapeutic drug. The ideal druggable driver would have the following properties:

1. The pathway target is necessary for the expression for disease, but is not necessary for the survival of normal cells (i.e., you can eliminate the pathway without killing normal cells)
2. There must be a pathway protein that is necessary for the activity of the pathway (i.e. if the protein is removed, the pathway cannot proceed)
3. The protein can be targeted by a drug. Among other properties, this condition informs us that the targeted protein must be chemically stable.
4. The protein target is itself not necessary for the survival of normal cells (i.e., targeting the protein must not kill the patient).

Familial hemophagocytic lymphohistiocytosis A rare disease of early childhood, characterized by lymphocytosis (i.e., increased blood lymphocytes) and rapidly enlarging lymph nodes infiltrated by lymphocytes and histiocytes, many of which are seen engulfing red blood cells (i.e. phagocytosis). The underlying defect in this disease is the absence of functional perforin. Perforin is a cytolytic protein expressed by activated cytotoxic lymphocytes and natural killer cells. Without perforin, lymphocytes

and histiocytes cannot adequately destroy organisms. In such cases, lymphocytes and histiocytes accumulate in an ineffectual response to infection.

Founder effect Occurs when a specific mutation enters the population through the successful procreational activities of a founder and his or her offspring, who carry the founder's mutation. When all of the patients with a specific disease have an identical mutation, the disease may have been propagated through the population by a founder effect. This is particularly true when the disease is confined to a separable subpopulation, as appears to be the case for Navaho neurohepatopathy, in which the studied patients, all members of the Navaho community, have the same missense mutation.

Not all diseases characterized by a single gene mutation arise as the result of a founder effect. In the case of cystic fibrosis, a dominant founder effect can be observed within a genetically heterogeneous disease population. One allele of the cystic fibrosis gene accounts for 67% of cystic fibrosis cases in Europe. Hundreds of other alleles of the same gene account for the remaining 33% of cystic fibrosis cases.[164]

Perhaps the most notable "founder" is the so-called "mitochondrial eve." Mitochondria are inherited whole from the cytoplasm of maternal oocytes. Hence, all of us can, in theory, trace our mitochondria back up to the woman from whom all humans living today have descended.

Gene pool The imagined aggregate collection of genetic material from all of the members of a species.

Gene regulation Gene expression is influenced by many different regulatory systems, including the epigenome (e.g., chromatin packing, histone modification, and base methylation), transcription and posttranscription modifiers (e.g., transcription factors, DNA promoter sites, DNA enhancer sites, cis- and transacting factors, alternative RNA splicing, miRNA and competitive endogenous RNAs, additional forms of RNA silencing, RNA polyadenylation, and mRNA stabilizers), translational modifiers (e.g., translation initiation factors, and ribosomal processing), and posttranslational protein modifications.

Disruptions of any of these regulatory processes produce disease in humans and other metazoans.[165–174] Moreover, anything that modifies any regulatory process (e.g., environmental toxins, substrate availability, and epistatic genes) can influence gene regulation and hence can produce a disease phenotype.

Genome wide association study Abbreviation: GWAS. A method to find single nucleotide polymorphisms (SNPs) that are statistically associated with a polygenic disease. The methodology involves hybridizing DNA from individuals with disease and individuals from a control group, against a DNA array of immobilized fragments of DNA known to contain commonly occurring SNPs (i.e., allele-specific oligonucleotides). The SNPs that hybridize against the DNA extracted from individuals with disease (i.e. the SNPs matching the case samples) are compared with the SNPs that hybridize against the controls. SNPs that show a statistical difference between case samples and control samples are said to be associated with the disease.

Of course, there are many weaknesses to this approach: one being that differences in SNPs do not necessarily imply any functional variance in the gene product.[175] In addition, differences in SNPs may lead to statistically valid results that nonetheless have no relevance to the pathogenesis of disease.[176] Aside from false-positive GWAS associations, the methodology is virtually guaranteed to miss valid SNP associations, simply because SNP arrays are not exhaustive (i.e., do not contain 50+ million SNPs) and are limited to a selected set of commonly occurring polymorphisms. For example, a rare variant of the APOE gene has been shown to be strongly correlated with longevity.[177] This variant, because it is not included among the common APOE variants included in SNP arrays, would have been missed by a GWAS study. True associations are those that can be found repeatedly from laboratory to laboratory and that can be shown to have pathogenetic relevance. To date, very few disease-associated SNPs found in GWAS studies have met these criteria. It has been suggested that the GWAS studies, in toto, have had little scientific merit and have been misleading.[178]

A sympathetic evaluation of GWAS studies is that they help us to see recurrent sets of pathway genes involved in diseases. Knowing that a related set of genes seems to implicate a pathway in the development or expression of a common disease has great value.[179] By focusing attention on a pathway,

scientists can start to dissect the important events in the pathogenesis of a disease.[180] In addition, we should keep in mind that a gene whose variant form plays only a very minor role in the expression of a polygenic disease may actually serve as the target of a new drug that is highly effective for the disease. How so? Small variants in the enzyme (as observed in SNPs) may produce only a small change in the activity of the enzyme, and this may reveal itself as a very small disease association in a GWAS study. Nonetheless, we can imagine situations wherein a new drug may decrease the activity of a minor pathway enzyme by 95%, thus greatly reducing the overall output of the pathway. Such an effect might be crucial in key disease pathways. This seems to be the case observed in statins, a drug that targets one of many enzymes involved in cholesterol synthesis (HMG-CoA reductase) but which produces profound alterations in total cholesterol levels. A small variation in a disease pathway gene (such as the gene coding for HMG-CoA reductase) might not produce a dramatic finding in a GWAS study, but even a small effect might serve as a significant finding, leading to the discovery of a new, major class of drugs.[181]

Genomic disorder Synonymous with genomic disease. Although all genetic diseases are technically genomic diseases, the term "genomic disorder" is usually reserved for diseases arising from the loss or gain of portions of the DNA in chromosomes (i.e., usually involving several megabases and never single nucleotide mutations). Diseases in which there is an increase or decrease of portions of DNA that normally occur as multiple copies (i.e., copy number losses and copy number gains) are included in the genomic disorders. Some genomic disorders arise as sporadic, de novo conditions caused by chromosomal rearrangements that may result in interstitial or terminal deletions or duplications. Meiotic nondisjunction events in phenotypically normal carriers of balanced translocations may be passed on as a disorder of gene dosage in their offspring.[89]

Hamartoma Hamartomas are benign growths that occupy a peculiar zone lying between neoplasia (i.e., a clonal expansion of an abnormal cell) and hyperplasia (i.e., the localized overgrowth of a tissue). Some hamartomas are composed of tissues derived from several embryonic lineages (e.g., ectodermal tissues mixed with mesenchymal tissue). This is almost never the case in cancers, which are clonally derived neoplasms wherein every cell is derived from a single cell type.

Hamartomas occasionally occur in abundance in inherited syndromes, as in tuberous sclerosis. The pathognomonic lesion in tuberous sclerosis is the brain tuber, the hamartoma from which the syndrome takes its name. Tubers of the brain consist of localized but poorly demarcated malformations of neuronal and glial cells. Like other hamartoma syndromes the germ line mutation in tuberous sclerosis produces benign hamartomas and carcinomas, indicating that hamartomas and cancers are biologically related. Hamartomas and cancers associated with tuberous sclerosis include cortical tubers of brain, retinal astrocytoma, cardiac rhabdomyoma, lymphangiomyomatosis (very rarely), facial angiofibroma, white ash leaf-shaped macules, subcutaneous nodules, cafe-au-lait spots, subungual fibromata, myocardial rhabdomyoma, multiple bilateral renal angiomyolipoma, ependymoma, renal carcinoma, and subependymal giant cell astrocytoma.[182]

Another genetic condition associated with hamartomas is Cowden syndrome. Cowden syndrome is associated with a loss-of-function mutation in PTEN, a tumor suppressor gene.[183] Features that may be encountered are macrocephaly, intestinal hamartomatous polyps, benign hamartomatous skin tumors (multiple trichilemmomas, papillomatous papules, and acral keratoses), dysplastic gangliocytoma of the cerebellum, and a predisposition to cancers of the breast, thyroid, and endometrium.[184–186]

Homolog In the field of bioinformatics, "homolog" always refers to homologous genes. Genes from different organisms are considered homologous to one another if both descended from a gene in a common ancestral organism. If the mechanism of descent was speciation, the homologous genes are also orthologous. If the mechanism of descent was due to gene duplication, then the homologous genes are also paralogous. Homologous genes tend to share similar sequences.

Horizontal gene transfer The direct transfer of genetic material between organisms, by mechanisms other than by reproduction (i.e., other than the transfer of DNA from parents to offspring). The very

first eukaryotic ancestors derived their genetic material by horizontal gene transfer from prokaryotes (bacteria and archaean organisms), viruses, and possibly from other now-extinct organisms that might have preceded the eukaryotes. The early eukaryotes almost certainly exchanged DNA between one another, and we see evidence of such exchanges in modern single-celled eukaryotes and fungi.[187,188]

To an unknown extent horizontal gene transfer occurs throughout the animal kingdom. For example, tardigrades, a microscopic animal, has a genome one-sixth of which was derived from bacteria, archaeans, plants, and fungi.[189]

It should be noted that many of the most significant evolutionary advances came from interspecies gene acquisitions. The primordial mitochondrion that helped to create the first eukaryotic cell was an acquisition from bacteria. The very first chloroplast in the most primitive precursor of the plant kingdom was a pilfered cyanobacteria. The big jump in adaptive immunology came with acquisition of the RAG1 gene. This gene enabled the DNA that encodes a segment of the immunoglobulin molecule to rearrange, thus producing a vast array of protein variants.[190] The RAG1 gene, which kicked off adaptive immunity in animals, was derived from a transposon, an ancient DNA element that was acquired through horizontal gene transfer or through infection from another living organism or from a virus.

Imaginal disc Imaginal discs are foci of stem cells that live in insect larvae and that grow to become the various parts of adult insects. Flying insects have an imaginal disc to create a wing and another imaginal disc to create a leg and another for an antenna and so forth. Imaginal discs have been the fundamental tool that developmental biologists have used, for decades, to study the control of organogenesis.[191] In insects, tumors may arise from imaginal disc cells, and these insect neoplasms are transplantable, invasive, metastatic, and hence malignant.[192]

Law of sequence conservation If a sequence is conserved through evolution (i.e., if we can find a closely similar sequence that is present in various animal species having a shared ancestral class), then that sequence must perform a useful function for the organism. Furthermore, sequences that are highly conserved (i.e., with very little difference among class members), are likely to have a very important function. This law is so useful and so fundamental to genomics and to gene-related computational algorithms that it may as well be known as the first law of bioinformatics.

The corollary to the law is that genomic sequences that degenerate over time and for which there are large variations in closely related species must not have a very important function in the organism.

Mendelian inheritance A pattern of inheritance observed for traits inherited from the mother or the father. The modes of Mendelian inheritance are autosomal dominant, autosomal recessive, sex-linked dominant, and sex-linked recessive. The most comprehensive listing and discussion of the Mendelian diseases has been collected, for many decades, in Mendelian Inheritance in Man, currently available online.[182] The number of cataloged Mendelian diseases varies depending on how they are counted (e.g., a smaller number if counted by disease phenotype and a larger number if counted by genotypic subtypes), but it is generally accepted that there are at least 7000 documented Mendelian diseases that occur in humans.

Microdeletion Microdeletions are cytogenetic abnormalities that typically span several megabases of DNA. Microdeletions are too small to be visible with standard cytogenetics, but they can often be detected with fluorescent in situ hybridization (FISH). All of the microdeletion syndromes are rare diseases, and they typically arise as de novo germ line aberrations (i.e., not inherited from mother or father, in most instances). Conditions that occur rarely and sporadically to produce a uniform set of phenotypic features in unrelated subjects may be new cases of microdeletion syndromes.[193] DiGeorge syndrome is a typical microdeletion disease, with a germ line 22q11.2 deletion encompassing about 3 million base pairs on one copy of chromosome 22, containing about 45 genes. Neurofibromatosis I sometimes occurs as a microdeletion syndrome involving a region of chromosome 17q11.2 that includes the NF1 gene.

Microsatellite Also known as simple sequence repeats (SSRs), microsatellites are DNA sequences consisting of repeating units of 1–4 base pairs. Microsatellites are inherited and polymorphic. This means that within a population there will be wide variation in the number of repeats at any microsatellite locus.

Microsatellite instability When there is a deficiency of proper mismatch repair (a type of DNA repair), DNA replication is faulty, and novel microsatellites (repeating short DNA sequences) may appear in chromosomes. This phenomenon is called microsatellite instability, and it is observed to occur at high frequency in cells obtained from various types of cancer. Microsatellite instability is a characteristic feature of colon cancers occurring in hereditary nonpolyposis colorectal cancer syndrome.

Molecular clock The molecular clock is a metaphor describing an analytic method by which the age of phylogenetic divergence of two species can be estimated by comparing the differences in sequence between two homologous genes or proteins. The name "molecular clock" and the basic theory underlying the method were described in the early 1960s, when the amino acid sequence of the hemoglobin molecule was determined for humans and other hominids.[194] It seemed clear enough at the time that if the number of amino acid substitutions in the hemoglobin sequence, compared among two species, was large, then a very great time must have elapsed since the phylogenetic divergence of the two species. The reason being that sequence changes occur randomly over time, and as more time passes, more substitutions will occur. Conversely, if the differences in amino acid sequence between species is very small, then the time elapsed between the species divergence must have been small.

As with all simple and elegant theories in the biological sciences, the devil lies in the details. Today, we know that analyses must take into account the presence or absence of conserved regions (whose sequences will not change very much over time). Indeed analysts must apply a host of adjustments before they can claim to have a fairly calibrated molecular clock.[195,196] At the end of the process, biologists use additional information related to the timing of species divergences, possibly corroborating the prior chronology or tentatively establishing new timelines.

Molecular targeted drugs Chemotherapy that selectively targets the activity of a few molecular species within the cell. For example, Avastin (bevacizumab) binds to and inhibits the biologic activity of human vascular endothelial growth factor (VEGF); Iressa blocks EGFR; Gleevec (imatinib mesylate) inhibits ABL, PDGFR, and KIT kinases; and SU11248 inhibits c-Kit, VEGFR, and PDGFR. Because these agents target a narrow range of cellular pathways that are elevated in cancers, they produce very little systemic toxicity. Most targeted chemotherapies are also called nontoxic chemotherapies.

Monogenic disease Same as single gene disease. The term is used in cases wherein a mutation in a single gene is the root cause of a disease.

Knowing what we now know about the pathogenesis of disease, the term "monogenic disease" is somewhat misleading. All diseases develop in steps and many different genes, pathways, and biological events may contribute to the pathogenetic process. Because we continue to discover "carriers" of monogenic disease genes who never develop the disease, we can infer that such diseases are not truly monogenic. It is easy to imagine that specific polymorphisms of multiple genes, in addition to the "monogenic disease gene" may be necessary to produce the disease phenotype.

It is worth remembering that any disease attributed to the loss of function of a single gene can be mimicked by selective epigenetic silencing. We don't call such conditions "zero-genic" diseases (i.e., diseases with no gene mutation). It is safest to refrain from assigning cardinality to genetic diseases until we fully understand their pathogeneses.

Mosquito Mosquitoes are members of Class Culicidae. Four genera of mosquitoes are vectors for human diseases: Aedes, Anopheles, Armigeres, and Culex. Among these genera, there are hundreds of individual species. Associating specific species of mosquito with specific diseases is a field of medicine unto itself. Mosquitoes are vectors for biologically diverse organisms (animals, protists, and viruses), transmitting viral and eukaryotic pathogens. As yet, mosquitoes are not known to be vectors for bacterial diseases, but this biological oversight may soon be corrected. It has recently been shown that mosquitoes carry pathogenic bacteria, including antibiotic-resistant species.[197]

Multi-step process All of life can be described as a multistep process, wherein each cellular event is directly preceded by some other event. Because every biological event has a preceding event, it can be inferred that every cellular event that occurs in any organism can be iteratively traced backward through history, to the first cellular event that occurred on the planet, some 4 billion years ago. For practical reasons, determining the root cause of a disease requires us to choose an arbitrary cut point where we say that pathogenesis begins, and we call this cut point the root cause.

Myelofibrosis Characterized by a sustained hyperplasia of extramedullary hematopoietic cells (i.e., blood cell lineage maturation occurring outside of the bone marrow). The pathogenesis of myelofibrosis is obscure and somewhat controversial. It is thought that the extramedullary hematopoiesis is caused by a primary fibrosing disease of the bone marrow. A specific JAK2 mutation is found in more than half of patients with myelofibrosis.[132]

Natural selection The tendency for favorable heritable traits to become more common over successive generations. The traits are selected from expressed genetic variations among individuals in the population.

Neurectoderm Alternate spelling: neuroectoderm. An embryonic derivative of the ectodermal layer that produces the neural tube and neural plate, from which the central nervous system (primarily brain and spinal cord) derive.

Neurocristopathy A disease of cells that derive from the neural crest. Examples include MEN2 (multiple endocrine neoplasm syndrome type 2), a ganglionic diseases of the GI tract and neurofibromatosis.

Non-coding mutational diseases We are all familiar with the concept of a genetic disease, wherein a mutation in a gene serves as the root cause of a disease. We are much less familiar with the idea that a mutation in a noncoding region of the genome (i.e., a nongene) can also produce a disease.

Most single nucleotide mutations in noncoding regions of the genome cannot cause human disease, because most of the DNA in noncoding regions are nonfunctional and neither code for proteins nor serve as regulatory elements. Still, we know that some portion of the noncoding region of DNA is regulatory, and we presume that mutations of noncoding regions that are capable of causing disease must arise in regulatory sequences. Regulatory effects caused by mutations in noncoding regions generally produce mild changes in the level of expression of coding genes (i.e., they do not turn off the expression of genes). Hence a mutation in a single noncoding site is seldom sufficient to produce a serious disorder.

Nonetheless, examples of diseases caused by mutations in noncoding regions can be found, as listed here:

A form of frontotemporal dementia and/or amyotrophic lateral sclerosis (FTDALS1) is caused by a heterozygous hexanucleotide repeat expansion (GGGGCC) in a noncoding region of the C9ORF72 gene.

Lactase persistence is associated with noncoding variation in the MCM6 gene.

Chronic tubulointerstitial nephropathy can be caused by a 5656A-G transition in mitochondrial DNA. This adenine is the single noncoding nucleotide separating the structural genes of 2 tRNAs.

Hyperferritinemia-cataract syndrome is caused by heterozygous mutation in the iron-responsive element in the 5-prime noncoding region of the ferritin light chain gene.

Nonphylogenetic property Properties that do not hold true for a class; hence, nonphylogenetic properties cannot be used by ontologists to create a classification. For example, we do not classify animals by height or weight because animals of greatly different heights and weights may occupy the same biological class. Similarly, animals within a class may have widely ranging geographic habitats; hence, we cannot classify animals by locality. Case in point: penguins can be found virtually anywhere in the

southern hemisphere, including hot and cold climates. Hence, we cannot classify penguins as animals that live in Antarctica or that prefer a cold climate.

Scientists commonly encounter properties, once thought to be class specific, that prove to be uninformative, for classification purposes. For many decades, all bacteria were assumed to be small, much smaller than animal cells. However, the bacterium *Epulopiscium fishelsoni* grows to about 600 μm by 80 μm, much larger than the typical animal epithelial cell (about 35 μm in diameter).[198] *Thiomargarita namibiensis*, an ocean-dwelling bacterium, can reach a size of 0.75 mm, visible to the unaided eye. What do these admittedly obscure facts teach us about the art of classification? Superficial properties, such as size, seldom inform us how to classify objects. The ontologist must think very deeply to find the essential defining features of classes.

Nonphylogenetic signal DNA sequences that cannot yield any useful conclusions related to evolutionary lineage. Because DNA mutations arise stochastically over time (i.e., at random locations in the gene and at random times), two organisms having different ancestors may, by chance alone, achieve the same sequence in a chosen stretch of DNA. When gene sequence data are analyzed, and two organisms share the same sequence in a stretch of DNA, it can be tempting to infer that the two organisms belong to the same class (i.e., that they inherited the identical sequence from a common ancestor). This inference is not necessarily correct. When mathematical phylogeneticists began modeling inferences for gene data sets, they assumed that most of class assignment errors based on DNA sequence similarity would occur where the branches between sister taxa were long (i.e. when a long time elapsed between evolutionary divergences, allowing for many random substitutions in base pairs). They called this phenomenon, wherein nonsister taxa were assigned the same ancient parent class, "long branch attraction." In practice, errors of this type can occur whether the branches are long, short, or in-between. The term "nonphylogenetic signal" refers to just about any pitfall in phylogenetic grouping due to gene similarities acquired through any mechanism other than inheritance from a shared ancestor. This would include random mutational and adaptive convergence.[199–201]

Oncogene Normal genes or parts of genes that when altered to become a more active form, or are overexpressed, contribute toward a neoplastic phenotype in a particular range of cell types. The normal form of the oncogene is called the protooncogene. The altered more active form of the gene is called the activated oncogene. Activation usually involves mutation or amplification (i.e., an increase in gene copy number), translocation, or fusion with an actively transcribed gene or some sequence of events that increases the expression of the gene product. Some retroviruses contain activated oncogenes and can cause tumors by inserting their oncogene into the host genome.

Orthodisease Orthodiseases are conditions observed in nonhuman species that result from alterations in genes that are homologous to the genes known to cause diseases in humans.

Ortholog Refers to a gene found in different organisms that evolved from a common ancestor's gene through speciation. As an empiric observation, orthologs in different species often have the same or similar functionality.[13,57–61] It is assumed that a gene and its encoded protein have greater similarity to their orthologs in another species than to any of the other genes/proteins in its own genome. It is this basic assumption that drives the algorithms designed to determine the evolutionary lineage of orthologs in different species.[202,203] Orthology is a type of homology.

Oxygen crisis Nobody knows with any certainty about the history of terrestrial oxygenation, but a consensus of opinion seems to favor the following scenario:

1. About 3.5 billion years ago, a prokaryote evolved that could produce oxygen from water, through a rather inefficient pathway that did not involve photosynthesis. All life on earth at this time was prokaryotic, and virtually all living organisms were anaerobes for which oxygen was toxic. A relatively small amount of oxygen was produced, all of which was rapidly trapped by substances within the earth, particularly oxidizing minerals such as iron. As an incidental observation, virtually all of the iron found on earth is oxidized. Pure metallic iron is exceedingly rare, and most of the pure elemental iron on this planet arrived recently in meteorites that had traveled billions of miles through airless space.

2. About 2.7 billion years ago, cyanobacteria evolved oxygenic photosynthesis, producing oxygen efficiently using photons. No organism since has independently evolved oxygenic photosynthesis. Early eukaryotes of Class Archaeplastida captured cyanobacteria and adapted them as organelles (chloroplasts) much later, probably no earlier than 2.1 billion years ago. At this point (i.e., about 2.7 billion years ago), a great deal of oxygen was being synthesized, but this oxygen was absorbed by the oceans and the emerging land masses, and very little oxygen made its way into the atmosphere.[204]
3. About 2.5 billion years ago, the earth was saturated with oxygen, and the excess gas bubbled its way into the atmosphere.
4. About 2.3 billion years ago, atmospheric oxygen reached a level that was toxic to most of the existing anaerobic organisms, producing the so-called oxygen crisis or oxygen catastrophe.
5. The really large atmospheric oxygen concentrations, comparable with those we see today, did not come until about 1 billion years ago. By this time, there were many eukaryotes, all containing mitochondria capable of turning oxygen into metabolic energy. The most dramatic examples of organisms exploiting the high-energy metabolism provided by oxygen came about a half billion years later, in the Cambrian explosion.[205]

Paralog A paralogous gene. Refers to genes found in different species of organisms that evolved from a common ancestor's gene through gene duplication. Paralogs permit the organism to get a new functionality from a gene, through natural selection, without losing the functionality of the gene that has been duplicated. A paralog is a type of homolog. All homologs are either orthologs or paralogs.

Pathway-driven disease Refers to disorders whose clinical phenotype is largely the result of a single, identifiable metabolic pathway. Diseases with similar clinical phenotypes can often be grouped together as they share a common, disease-driving pathway. Examples would include the channelopathies (driven by malfunctions of pathways that involving the transport ions through membrane channels), ciliopathies (driven by malfunctions of cilia), and lipid receptor mutations (driven by any of the mutations involving lipid receptors).

At this point, our ability to sensibly assign diseases to pathways is limited because diseases may involve many different pathways, and those pathways may be different for different cell types involved in the disease. For example, it is difficult to speak of a class of diseases all driven by errors in transcription factor pathways. A single transcription factor may regulate pathways in a variety of cell types with differing functions and embryologic origins. Hence the syndromes resulting from a mutation in a transcription factor may involve multiple pathways and multiple tissues and will not have any single, identifiable pathway that drives the clinical phenotype. Still, there is some hope that as more cell-based data become available, modern data analysis techniques will reliably match specific diseases with specific pathways.[206]

Phenetics The classification of organisms by feature similarity, rather than through relationships. Taxonomists are generally opposed to the idea of phenetics-based classifications, preferring to build classifications by finding the relationships that connect one class to another. Taxonomists have long held that a species is a natural unit of biological life and that the nature of a species is revealed through the intellectual process of building a consistent taxonomy, an intellectual process that is not based on phenetics.[207]

Polycythemia vera A sustained increase in circulating red blood cells that is not a physiologic response (e.g., not due to chronic hypoxia). The incidence of polycythemia vera is about 2 persons per 100,000 population. A somatic mutation, JAK2-V617F, is present in most cases of polycythemia vera.[133] Most of the life-threatening consequences of polycythemia vera result from hyperviscosity (i.e., thickening) of the blood, a condition that arises as a consequence of having an overabundance of circulating red blood cells. With aggressive therapy aimed at keeping the number of circulating red cells within physiologic norms, individuals affected by polycythemia vera may live many years with their disease.

Polycythemia vera does not usually follow the course of typical malignant neoplasms (i.e. large masses of invasive tumor metastasizing to many different organs are not typically seen). More often, polycythemia may attain the so-called "spent" phase in which the marrow becomes fibrotic and the patient becomes anemic. In rare cases a patient with polycythemia vera may eventually develop acute myeloid leukemia.

Polygenic disease A disease whose underlying cause involves alterations in multiple genes. In general the development of polygenic diseases is highly dependent upon environmental modifiers that trigger bouts of disease that enhance or reduce susceptibility to disease or that seem to serve as the apparent root cause of the disease.

As an example, consider a patient with no known underlying medical condition who is stung by a bee and immediately succumbs to anaphylactic shock. It is tempting to say that the root cause is the bee sting, but we know that most individuals who are stung by a bee do not develop an anaphylactic response. Clearly, some underlying condition must have predisposed the patient to develop shock. You want to blame the patient's genes, but if there was no parental history of familial history of anaphylaxis, then it would be hard to put the blame on an inherited gene. In such instances, we look toward a polygenic explanation, wherein multiple variants of gene expression together produce a physiological condition that predisposes the individual to anaphylactic shock. Of course, we cannot be certain that we are correct until we identify all of the variant genes and demonstrate the biological mechanism by which they exert their effect. That's a very tall order. In the meantime, we work under the tentative assumption that we are correct. That's science.

Precision Precision is the degree of exactitude of a measurement and is verified by its reproducibility (i.e., whether repeated measurements of the same quantity produce the same result). Accuracy measures how close your data come to being correct. Data can be accurate but imprecise or precise but inaccurate. If you have a 10-pound object and you report its weight as 7.2376 pounds, every time you weigh the object, then your precision is remarkable, but your accuracy is dismal.

What are the practical limits of precision measurements? Let us stretch our imaginations, for a moment, and pretend that we have just found an artifact left by an alien race known throughout the galaxy for its prowess in the science of measurement. As a sort of time capsule for the universe, their top scientists decide to collect the history of their civilization, encoded in binary. Their story looked something like "001011011101000…" extended to about 5 million places. Rather than print the sequence out on a piece of paper or a computer disc, these aliens simply converted the sequence to a decimal length (i.e., .001011011101000…) and marked the length on a bar composed of a substance that would never change its size. To decode the bar and recover the history of the alien race, one would simply need to have a highly precise measuring instrument that would yield the original binary sequence. Computational linguists could translate the sequence to text, and the recorded history of the alien race would be revealed! Of course the whole concept is built on an impossible premise. Nothing can be measured accurately to 5 million places.

We live in a universe with practical limits (e.g., the speed of light, the Heisenberg uncertainty principle, the maximum mass of a star, the second law of thermodynamics, the unpredictability of highly complex systems, and division by zero). There are many things that we simply cannot do no matter how hard we try. The most precise measurement achieved by modern science has been in the realm of atomic clocks, where accuracy of 18 decimal places has been claimed.[208] Nonetheless, many scientific disasters are caused by our ignorance of our own limitations and our persistent gullibility, leading us to believe that precision claimed is precision obtained.

Primary ciliopathies The ciliopathies are a genetically and phenotypically diverse set of diseases that could not have made any biological sense, as a class of related diseases, until quite recently. The key to the ciliopathies lies in the primary cilium, a mostly unary structure that exists in virtually every cell of the body. The primary cilium escaped attention until recently, standing as it does in the shadow of its more populous, more frenetic relative, the motile cilium. Every school child is taught about the structure, cytology, and function of the motile cilia. These organelles protrude from the apical cell wall of epithelial cells lining the respiratory tract, intestines, and ducts throughout the body, pushing

intraluminal materials along their way up a bronchus or down the intestines or through a duct. Motile cilia have a nine-paired microtubule axoneme, with a central pair of microtubules and outer dynein arms. One cell may have many motile cilia. The primary cilium has the appearance of a deformed motile cilia, lacking, as it does, the central pair of microtubules and outer dynein arms that are essential for normal motility. The primary cilium was first observed by microscopists in the 1950s, but because there was only one primary cilium per cell and because it didn't seem to serve any function, it was long dismissed as an evolutionary relic, much like the coccygeal bone or the vermiform appendix.[209] Around the 1990s it was recognized that the primary cilium, which grows from its leading tip, has no synthetic machinery within itself to transport growth substrates up through the cilium to the growth zone. Hence, it was presumed that the primary cilium must rely on some hypothetical transport mechanism to achieve its growth. This hypothetical mechanism was dubbed IFT (the abbreviation for intraflagellar transport). Soon thereafter the genes involved in intraflagellar transport were discovered, and these genes, when knocked out in mice, seemed able to produce heterotaxy (i.e., left-right organ asymmetry), indicating that primary cilia play a regulatory role in embryonic and fetal development.[209] Today the ciliopathies are a well-defined class of phenotypically diverse inherited diseases. Each ciliopathy, regardless of its phenotype, is associated with the proteins whose functions converge upon the primary cilium.[1,210]

All ciliopathies, despite their phenotypic diversities, involve disorders of the primary cilium. For example, primary ciliary dyskinesia features bronchiectasis, sinusitis, otitis media, infertility, and situs defects. Alstrom syndrome features dilated cardiomyopathy, obesity, sensorineural hearing loss, retinitis pigmentosa, endocrine abnormalities, and renal and hepatic disease. It is hard to imagine two diseases less similar to one another than primary ciliary dyskinesia and Alstrom syndrome. Nonetheless, both are caused by aberrations affecting the primary cilia. There is hope that treatments developed for any member of the ciliopathies might be of benefit for every type of ciliopathic disease[1,211] (Fig. 5.5).

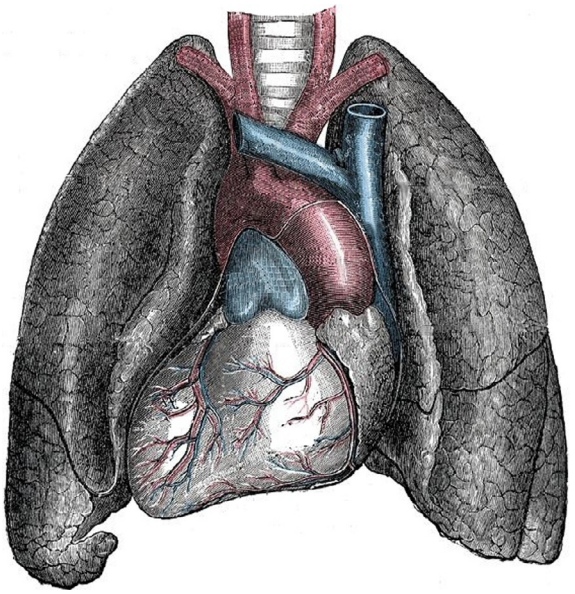

FIG. 5.5 The heart, lungs, and great vessels are switched around to appear the mirror image of normal chest anatomy. The easy giveaway is the normal, posterior location of the trachea, proving that the error is in the anatomy and not the result of a flipped image. It is hard to imagine that this malformation could result from a defect in cilia, but current evidence indicates that situs inversus is a ciliopathy. *Source: 20th U.S. edition of Gray's Anatomy of the Human Body, originally published in 1918.*

Included among the ciliopathies are Joubert syndrome, nephronophthisis, Senior-Loken syndrome, orofaciodigital syndrome, Jeune chondrodysplasia syndrome, autosomal dominant polycystic kidney disease, recessive polycystic kidney disease, Leber congenital amaurosis, Meckel-Gruber syndrome, Bardet-Biedl syndrome, Usher syndrome, Alstrom syndrome, McKusick-Kaufman syndrome, Ellis van Creveld syndrome, cranioectodermal dysplasia (Sensenbrenner syndrome), short rib polydactyly, some forms of retinal dystrophy, and heterotaxy (including visceral situs anomalies, asplenia or polysplenia, congenital heart defects, biliary atresia, and midline defects).[1,2] There are only a few organs and functional systems of the human body that escape involvement by this strange collection of related rare diseases.

Pseudogene A gene (i.e., a coding sequence) that does not actually code for a protein. Theories explaining the origin of pseudogenes are many. Some pseudogenes presumably devolved from genes that acquired mutations that rendered the genes nonfunctional. Other pseudogenes may have been reverse transcribed into DNA via RNA retrotransposons. Pseudogenes are identified from genomic sequence data using computational algorithms that search for stretches of DNA that have some sequence similarities to functional genes, along with sequences that might render the gene nonfunctional (e.g., premature stop codons, frameshift mutations, and a poly-A tail).

It is currently believed that transcribed pseudogene sequences (i.e., pseudogene RNA) is one of a class of competitive endogenous RNA species that compete for microRNA binding sites and consequently diminish the repressive actions of microRNA on target expression. Hence, pseudogenes moderate microRNA activity and provide some level of gene expression enhancement. There are, at a minimum, several thousand pseudogenes in the genome, and some genes, such as actin, may have numerous pseudogenes.[22,212] At present, pseudogenes are thought to play a role in the dysregulation of cancer cells and in cell defects found in neurodegenerative disorders.[90,213]

Rare cancer The definition of rare disease is included in the US Orphan Drug Act of 1983[214] and is included here:

> *For purposes of paragraph (1), the term 'rare disease or condition' means any disease or condition which (A) affects less than 200,000 persons in the United States or (B) affects more than 200,000 in the United States and for which there is no reasonable expectation that the cost of developing and making available in the United States a drug for such disease or condition will be recovered from sales in the United States of such drug.*

With just a few exceptions, all cancers affect fewer than 200,000 people in the United States each year and would be classified as rare diseases. Hence the official US definition of a rare disease is not particularly helpful. The National Cancer Institute has defined a rare cancer as a cancer with under 40,000 new cases in the United States each year. This definition also includes all but a handful of cancer types. In my opinion a more realistic definition for a rare cancer would include any cancer that occurs with an incidence under 1 per 100,000 persons per year (i.e., under 3000 new cases in the United States each year). In practical terms a rare cancer may be encountered once every few years in a busy hospital. For any given rare cancer, the average oncologist might encounter a patient with that tumor once or never within the span of his/her career.

Reassortment Often confused or used interchangeably with "recombination," reassortment is generally reserved for a viral event wherein two similar segmented viruses exchange part of their genomes during the coinfection of a host cell. Reassortment seems to be the major mechanism accounting for new influenza virus strains.

Retinoblastoma Tumor arising from primitive cells that produce the retinal lining epithelium in the eye.

Retrovirus An RNA virus that replicates through a DNA intermediate. The DNA intermediate may become integrated into the host DNA, from which viral RNA is transcribed. When integration of the virus occurs in germ cells, the viral DNA can be inherited by the offspring. Through this mechanism the

human genome carries a legacy of retroviral DNA. Ancient retroviruses account for about 8% of the human genome.[46]

Schwannoma A tumor is composed of neoplastic Schwann cells that are normally found wrapped around the axonal extensions of peripheral nervous system neurons (i.e., of neural crest origin). Schwannomas of the acoustic spinal nerves occur in neurofibromatosis type 2.

Sequence similarity Occurs when two sequences (nucleotides in DNA or RNA or amino acids in proteins) share many of the same components in the same order and roughly the same locations in the sequence. There are many ways to measure sequence similarity. The important thing to remember is that sequence similarity does not always imply that the two sequences share an origin either within the organism (by mutation of one sequence into the other or duplication of a sequence or rearrangement between sequences) or through evolution (derivation from the same sequence in an ancestral organism).

Silent mutation A mutation that does not alter phenotype. Silent mutations can occur in noncoding regions or in genes. It has been reported that silent mutations in genes may have subtle effects on the tertiary structure of proteins.[215]

Single gene disease Synonymous with monogenic disease/disorder. A disease for which an aberration of one gene leads to the disease. Such disorders may exhibit Mendelian inheritance patterns (recessive or dominant). In the case of recessive single gene disorders, the two alleles of the single gene are affected. A single gene disorder may have genetic heterogeneity. For example, there may be many different genes that are capable of causing the disorder (e.g., familial forms of dilated cardiomyopathy can result from mutations in any one of the following genes: LDB3 gene, TNNT2 gene, SCN5A gene, TTN gene, DES gene, EYA4 gene, SGCD gene, CSRP3 gene, ABCC9 gene, PLN gene, ACTC gene, MYH7 gene, PSEN1 gene, PSEN2 gene, gene encoding metavinculin, gene encoding fukutin, TPM1 gene, TNNC1 gene, ACTN2 gene, DSG2 gene NEXN gene, MYH6 gene, TNNI3 gene, SDHA gene, BAG3 gene, CRYAB gene, LAMA4 gene, MYPN gene, PRDM16 gene, MYBPC3 gene, TNNI3 gene, and GATAD1 gene. Nonetheless, each case is caused by an aberration of only one of those listed genes; hence, familial dilated cardiomyopathy is considered a single gene disorder. When a single gene causes a disease, regardless of the number of disease-causing variants in the gene, it would be considered a single gene disorder.

Also, in single gene disorders, there may be multiple genes that modify the clinical phenotype. Regardless, if a disease is characterized by a known mutation in a single gene that is sufficient for the expression of the disease, then the disease is considered a single gene disorder (even when other gene variants may contribute to the clinical phenotype).

Spliceosome In eukaryotes, DNA sequences are not transcribed directly into full-length RNA molecules, ready for translation into a final protein. There is a pretranslational process wherein transcribable sections of DNA, called exons, are cut out and spliced together and a single coding sequence is assembled. Alternative splicing is one method whereby more than one protein form can be produced by a single gene.[165] Cellular proteins that coordinate the splicing process are referred to, in aggregate, as the spliceosome. Errors in normal splicing can produce inherited disease, and it estimated that 15% of disease-causing mutations involve splicing.[152,153] Even the unicellular eukaryotes have spliceosomes.[216] Examples of spliceosome diseases are spinal muscular atrophy and some forms of retinitis pigmentosa.[165]

Sporadic Describes a disease or a specific case occurrence of a disease with no known cause and without any discernible pattern of occurrence (e.g., genetic and environmental). Thus diseases that have a familial pattern of inheritance are always considered nonsporadic, even when the root genetic cause is unknown. Likewise, diseases that occur as an epidemic or endemic pattern are always considered nonsporadic, even when the precise environmental cause is unknown. Rare diseases are seldom sporadic, as they typically exhibit some pattern of inheritance.

Common diseases are often sporadic but may contain subsets of disease occurrences that are nonsporadic. An example is schizophrenia. Schizophrenia is a common disease with a prevalence of about 1.1%. This translates to about 51 million individuals worldwide, who suffer from this mental disorder. Many cases of schizophrenia occur in families, and such cases are considered to be inherited and, thus, nonsporadic. Other cases seem to have no familial association and are considered sporadic. Are these sporadic cases caused by environmental factors or are they caused by de novo mutations that arose in the affected individuals? Recent evidence would suggest that many of the so-called sporadic cases arise from new mutations in affected individuals.[217]

When an association is made between a disease and some demographic factor, the distinction between sporadic and nonsporadic may be arbitrary. For example, if a disease occurs predominantly in women, can it be called sporadic? The cause may be completely unknown, but it has a definite pattern.

It should be mentioned that the term "sporadic" is fraught with scientific ambiguity and should probably be abandoned altogether. To label a disease "sporadic" seems to legitimize and perpetuate the dubious notion that diseases can occur without cause. When you read an old textbook of medicine and you see a disease listed as "sporadic," you would likely accept this as a substantiated fact. Of course, this is a terrible way to think about diseases. Many of the diseases that were considered to be sporadic, decades ago, are now known to have specific causes. **Would it not be more accurate to use the phrase "not as yet determined" in place of "sporadic," for occurrences of a disease whose cause is currently unknown?**

Sporadic disease versus phenocopy disease A sporadic disease is a disease with no known cause. A phenocopy disease is a nongenetic disease that mimics a genetic disease. Despite the clear-cut difference in the two definitions, it is often impossible to distinguish sporadic diseases from phenocopy diseases, when the distinction relies upon information that is not available. For example, if a patient presents with the clinical features of a well-described genetic disease but lacks the genetic biomarker for the disease, we might say that the disease is sporadic (i.e., of no known cause). Nonetheless, we know that diseases do not arise spontaneously. Even sporadic diseases have a causal pathogenesis, and if the pathogenesis is not genetic, it must be acquired. Hence, when a genetic cause is ruled out, we are tempted to say that the sporadic disease is a phenocopy (i.e., a clinical mimic without a root genetic cause). But suppose further research shows that the presumptive sporadic disease has a root genetic cause that is different from the genetic mutation that had been previously identified in cases of the inherited disease. In this case the disease is no longer sporadic (i.e., without known cause) and no longer a phenocopy (i.e. without genetic cause). Everything we thought to be true is now false, and it's all because we did not know the full story when we committed errors. These kinds of mistakes arise all the time in modern medicine, but they could all be avoided if we eliminated some of our popular jargon.

Synonymous SNPs Single nucleotide polymorphisms (SNPs) that have different sequences but which produce an equivalent transcriptional result due to triplet redundancy in the genetic code. For example, guu, guc, gua, and gug all code for the amino acid valine and are synonymous with one another

Trilateral retinoblastoma The occurrence of bilateral retinoblastomas is sometimes followed by the occurrence of a pineoblastoma, and this is referred to as trilateral retinoblastoma. The pineal gland has an evolutionary anlage identical to that of the eye. The difference is that neuroectoderm-derived photoreceptors of the pineal gland took up residence deep within a midline recess in the brain, while the photoreceptors of the eyes developed as paired external structures. The evolved pineal gland, like the evolved eyes, reacts to filtered sunlight, though any light reaching the pineal has had a lot more filtering than the light reaching the retinas. In response to light cessation, the pineal gland releases melatonin, a hormone that influences circadian rhythms (e.g., sleep). The same germ line mutation that leads to bilateral retinoblastomas may occasionally cause a pineoblastoma. Pineoblastomas share a common morphology (i.e., histologic appearance) and homology (i.e. development from equivalent embryonic anlagen) with retinoblastomas.[218]

Wild-type gene The functional, nonmutated gene found naturally in a population.

References

[1] Novarino G, Akizu N, Gleeson JG. Modeling human disease in humans: the ciliopathies. Cell 2011;147:70–9.

[2] Ware SM, Aygun MG, Hildebrandt F. Spectrum of clinical diseases caused by disorders of primary cilia. Proc Am Thorac Soc 2011;8:444–850.

[3] Sadikovic B, Al-Romaih K, Squire J, Zielenska M. Cause and consequences of genetic and epigenetic alterations in human cancer. Curr Genomics 2008;9:394–408.

[4] Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, et al. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat Chem Biol 2011;7:885–7.

[5] Genetics Home Reference. National library of medicine, July 1. Available from: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp; 2013. Accessed 6 July 2013.

[6] Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 2012;337:100–4.

[7] Kim HL, Iwase M, Igawa T, Nishioka T, Kaneko S, Katsura Y, et al. Genomic structure and evolution of multigene families: "flowers" on the human genome. Int J Evol Biol 2012;2012:917678.

[8] Frederic MY, Lundin VF, Whiteside MD, Cueva JG, Tu DK, Kang SY, et al. Identification of 526 conserved metazoan genetic innovations exposes a new role for cofactor E-like in neuronal microtubule homeostasis. PLoS Genet 2013;9:e1003804.

[9] Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 2009;462:1056–60.

[10] Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 2013;14:117.

[11] Pandey UB, Nichols CD. Human disease models in *Drosophila melanogaster* and the role of the fly in therapeutic drug discovery. Pharmacol Rev 2011;63:411–36.

[12] Wetterbom A, Sevov M, Cavelier L, Bergstrom TF. Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. J Mol Evol 2006;63:682–90.

[13] Erwin D, Valentine J, Jablonski D. The origin of animal body plans. American Scientist; 1997 March/April.

[14] Britten RJ. Almost all human genes resulted from ancient duplication. PNAS 2006;103:19027–32.

[15] Hardison RC. Evolution of hemoglobin and its genes. Cold Spring Harb Perspect Med 2012;2:a011627.

[16] Storz JF. Gene duplication and evolutionary innovations in hemoglobin-oxygen transport. Physiology (Bethesda) 2016;31:223–32.

[17] Griffiths DJ. Endogenous retroviruses in the human genome sequence. Genome Biol 2001;2:reviews1017.1–reviews1017.5.

[18] Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 2010;463:84–7.

[19] Patel MR, Emerman M, Malik HS. Paleovirology: ghosts and gifts of viruses past. Curr Opin Virol 2011;1(4):304–9.

[20] Alfoldi J, Lindblad-Toh K. Comparative genomics as a tool to understand evolution and disease. Genome Res 2013;23:1063–8.

[21] Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, et al. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 2005;37:766–70.

[22] Ng SY, Gunning P, Eddy R, Ponte P, Leavitt J, Shows T, et al. Evolution of the functional human beta-actin gene and its multi-pseudogene family: conservation of noncoding regions and chromosomal dispersion of pseudogenes. Mol Cell Biol 1985;5:2720–32.

[23] Santangelo AM, de Souza FSJ, Franchini LF, Bumaschny VF, Low MJ, Rubinstein M. Ancient exaptation of a core-sine retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet 2007;3:e166.

[24] Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D, Matthews L, et al. Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. Nat Genet 2008;40:971–6.

[25] Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 2005;438:803–19.

[26] Engle SJ, Womer DE, Davies PM, Boivin G, Sahota A, Simmonds HA, et al. HPRT-APRT-deficient mice are not a model for lesch-nyhan syndrome. Hum Mol Genet 1996;5:1607–10.

[27] Raeder H, Vesterhus M, El Ouaamari A, Paulo JA, McAllister FE, Liew CW, et al. Absence of diabetes and pancreatic exocrine dysfunction in a transgenic model of carboxyl-ester lipase-MODY (maturity-onset diabetes of the young). PLoS One 2013;8:e60229.

[28] Koonin EV. How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet 2000;1:99–116.

[29] Berman JJ. Evolution's clinical guidebook: translating ancient genes into precision medicine. Cambridge, MA: Academic Press; 2019.

[30] Simpson GG. Principles of animal taxonomy. New York: Columbia University Press; 1961.

[31] Simpson GG. The principles of classification and a classification of mammals. Bull Am Mus Nat Hist 1945;85:1–350.

[32] Woese CR. Bacterial evolution. Microbiol Rev 1987;51:221–71.

[33] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. PNAS 1977;74:5088–90.

[34] Fischer MG, Kelly I, Foster LJ, Suttle CA. The virion of cafeteria roenbergensis virus (CroV) contains a complex suite of proteins for transcription and DNA repair. Virology 2014;466:82–94.

[35] Andino R, Domingo E. Viral quasispecies. Virology 2015;479-480:46–51.

[36] Morgan GJ. What is a virus species? Radical pluralism in viral taxonomy? Stud Hist Philos Biol Biomed Sci 2016;59:64–70.

[37] Argos P, Kamer G, Nicklin MJ, Wimmer E. Similarity in gene organization and homology between proteins of animal picornaviruses and a plant comovirus suggest common ancestry of these virus families. Nucleic Acids Res 1984;12:7251–67.

[38] Kamer G, Argos P. Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. Nucleic Acids Res 1984;12:7269–82.

[39] Goldbach R. Genome similarities between plant and animal RNA viruses. Microbiol Sci 1987;4:197–202.

[40] Koonin EV, Dolja VV. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit Rev Biochem Mol Biol 1993;28:375–430.

[41] Nasir A, Caetano-Anolles G. A phylogenomic data-driven exploration of viral origins and evolution. Sci Adv 2015;1:e1500527.

[42] Jern P, Sperber GO, Blomberg J. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology 2005;2:50.

[43] Bandin I, Dopazo CP. Host range, host specificity and hypothesized host shift events among viruses of lower vertebrates. Vet Res 2011;42:67.

[44] Balique F, Lecoq H, Raoult D, Colson P. Can plant viruses cross the kingdom border and be pathogenic to humans? Viruses 2015;7:2074–98.

[45] Hughes AL, Friedman R. Poxvirus genome evolution by gene gain and loss. Mol Phylogenet Evol 2005;35:186–95.

[46] Emerman M, Malik HS. Paleovirology: modern consequences of ancient viruses. PLoS Biol 2010;8: e1000301.

[47] Mohammed MA, Galbraith SE, Radford AD, Dove W, Takasaki T, Kurane I, et al. Molecular phylogenetic and evolutionary analyses of Muar strain of Japanese encephalitis virus reveal it is the missing fifth genotype. Infect Genet Evol 2011;11:855–62.

[48] Bannert N, Kurth R. The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet 2006;7:149–73.

[49] Rasmussen MD, Kellis M. A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol 2011;28:273–90.

[50] Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. PNAS 2011;108:17486–91.

[51] Yoon CK. Reviving the lost art of naming the world. The New York Times 2016; August 2.

[52] Blamont M. French drug trial disaster leaves one brain dead, five injured. Reuters 2016; January 15.

[53] D'Elia RV, Harrison K, Oyston PC, Lukaszewski RA, Clark GC. Targeting the cytokine storm for therapeutic benefit. Clin Vaccine Immunol 2013;20:319–27.

[54] Lee DW, Gardner R, Porter DL, Louis CU, Ahmed N, Jensen M, et al. Current concepts in the diagnosis and management of cytokine release syndrome. Blood 2014;124:188–95.

[55] Berman J. Precision medicine, and the reinvention of human disease. Cambridge, MA: Academic Press; 2018.

[56] Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W, et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc Natl Acad Sci U S A 2013;110:3507–12.

[57] Xu EY, Lee DF, Klebes A, Turek PJ, Kornberg TB, Reijo Pera RA. Human BOULE gene rescues meiotic defects in infertile flies. Hum Mol Genet 2003;12:169–75.

[58] Padgett RW, Wozney JM, Gelbart WM. Human BMP sequences can confer normal dorsal-ventral patterning in the Drosophila embryo. Proc Natl Acad Sci U S A 1993;90:2905–9.

[59] Hamada N, Backesjo CM, Smith CI, Yamamoto D. Functional replacement of Drosophila Btk29A with human Btk in male genital development and survival. FEBS Lett 2005;579:4131–7.

[60] McGinnis N, Kuziora MA, Mc Ginnis W. Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. Cell 1990;63:969–76.

[61] Grifoni D, Garoia F, Schimanski CC, Schmitz G, Laurenti E, Galle PR, et al. The human protein Hugl-1 substitutes for Drosophila lethal giant larvae tumour suppressor function in vivo. Oncogene 2004;23:8688–94.

[62] Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol 2009;7:e1000247.

[63] Chow CY, Reiter LT. Etiology of human genetic disease on the fly. Trends Genet 2017;33:391–8.

[64] Strange K. Drug discovery in fish, flies, and worms. ILAR J 2016;57:133–43.

[65] Novick P, Field C, Schekman R. Identification of 23 complementation groups required for post-translational events in the yeast secretory pathway. Cell 1980;21:205–15.

[66] McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc Natl Acad Sci U S A 2010;107:6544–9.

[67] Gissen P, Maher ER. Cargos and genes: insights into vesicular transport from inherited human disease. J Med Genet 2007;44:545–55.

[68] Rubinsztein DC. Protein-protein interaction networks in the spinocerebellar ataxias. Genome Biol 2006;7:229.

[69] Palikaras K, Tavernarakis N. Caenorhabditis elegans (Nematode). In: Brenner's encyclopedia of genetics. second edition Philadelphia: Elsevier; 2013. p. 404–8.

[70] Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature 2013;496:498–503.

[71] Spitsbergen JM, Kent ML. The state of the art of the zebrafish model for toxicology and toxicologic pathology research – advantages and current limitations. Toxicol Pathol 2003;31 (Suppl):62–87.

[72] Kelsh RN, Eisen JS. The zebrafish colourless gene regulates development of non-ectomesenchymal neural crest derivatives. Development 2000;127:515–25.

[73] Smolowitz R, Hanley J, Richmond H. A three-year retrospective study of abdominal tumors in zebrafish maintained in an aquatic laboratory animal facility. Biol Bull 2002;203:265–6.

[74] Wojciechowska S, van Rooijen E, Ceol C, Patton EE, White RM. Generation and analysis of zebrafish melanoma models. Methods Cell Biol 2016;134:531–49.

[75] Tobin DM, Vary Jr. JC, Ray JP, Walsh GS, Dunstan SJ, Bang ND, et al. The lta4h locus modulates susceptibility to mycobacterial infection in zebrafish and humans. Cell 2010;140:717–30.

[76] Curtis J, Kopanitsa L, Stebbings E, Speirs A, Ignatyeva O, Balabanova Y, et al. Association analysis of the LTA4H gene polymorphisms and pulmonary tuberculosis in 9115 subjects. Tuberculosis (Edinb) 2011;91:22–5.

[77] No attributed author. Tuberous sclerosis complex in flies too? a fly homolog to TSC2, called gigas, plays a role in cell cycle regulation, July 27. Available from: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=coffeebrk.chapter.25; 2000.

[78] Hariharan IK, Bilder D. Regulation of imaginal disc growth by tumor-suppressor genes in drosophila. Ann Rev Genet 2006;40:335–61.

[79] Nagel ZD, Chaim IA, Samson LD. Inter-individual variation in DNA repair capacity: a need for multipathway functional assays to promote translational DNA repair research. DNA Repair (Amst) 2014;19:199–213.

[80] Cohen A, Thompson E. DNA repair in nondividing human lymphocytes: inhibition by deoxyadenosine. Cancer Res 1986;46:1585–8.

[81] Rudnick DA, Perlmutter DH. Alpha-1-antitrypsin deficiency: a new paradigm for hepatocellular carcinoma in genetic liver disease. Hepatology 2005;42:514–21.

[82] Hidvegi T, Ewing M, Hale P, Dippold C, Beckett C, Kemp C, et al. An autophagy-enhancing drug promotes degradation of mutant alpha1-antitrypsin Z and reduces hepatic fibrosis. Science 2010;329:229–32.

[83] Houlston RS, Collins A, Slack J, Morton NE. Dominant genes for colorectal cancer are not rare. Hum Genet 1992;56:99–103.

[84] Whiffin N, Houlston RS. Architecture of inherited susceptibility to colorectal cancer: a voyage of discovery. Genes (Basel) 2014;5:270–84.

[85] Dobzhansky T. Genetics of the evolutionary process. New York: Columbia University Press; 1970.

[86] Hayden MR, Clee SM, Brooks-Wilson A, Genest Jr. J, Attie A, Kastelein JJ. Cholesterol efflux regulatory protein, Tangier disease and familial high-density lipoprotein deficiency. Curr Opin Lipidol 2000;11:117–22.

[87] Huang W, Moriyama K, Koga T, Hua H, Ageta M, Kawabata S, et al. Novel mutations in ABCA1 gene in Japanese patients with Tangier disease and familial high density lipoprotein deficiency with coronary heart disease. Biochim Biophys Acta 2001;1537:71–8.

[88] Solomon BD, Muenke M. When to suspect a genetic syndrome. Am Fam Physician 2012;86:826–33.

[89] McDermid HI, Morrow BE. Genomic disorders on 22q11. Am J Hum Genet 2002;70:1077–88.

[90] Costa V, Esposito R, Aprile M, Ciccodicola A. Non-coding RNA and pseudogenes in neurodegenerative diseases: "The (un)Usual Suspects" Front Genet 2012;3:231.

[91] Kleaveland B, Shi CY, Stefano J, Bartel DP. A network of noncoding regulatory RNAs acts in the mammalian brain. Cell 2018;174:350–62.

[92] Shen E, Shulha H, Weng Z, Akbarian S. Regulation of histone H3K4 methylation in brain development and disease. Philos Trans R Soc Lond B Biol Sci 2014;369:20130514.

[93] Weissman J, Naidu S, Bjornsson HT. Abnormalities of the DNA methylation mark and its machinery: an emerging cause of neurologic dysfunction. Semin Neurol 2014;34:249–57.

[94] Berman JJ. Rare diseases and orphan drugs: keys to understanding and treating common diseases. Cambridge, MD: Academic Press; 2014.

[95] Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010;42:30–5.

[96] Piano MR. Alcoholic cardiomyopathy: incidence, clinical characteristics, and pathophysiology. Chest 2002;121:1638–50.

[97] Zhu H-L, Meng S-R, Fan J-B, Chen J, Liang Y. Fibrillization of human tau is accelerated by exposure to Lead via interaction with His-330 and His-362. PLoS ONE 2011;6:e25020.

[98] Wang E, Boswell E, Siddiqi I, Lu CM, Sebastian S, Rehder C, et al. Pseudo-Pelger-Huet anomaly induced by medications: a clinicopathologic study in comparison with myelodysplastic syndrome-related pseudo-Pelger-Hu t anomaly. Am J Clin Pathol 2011;135:291–303.

[99] Juneja SK, Matthews JP, Luzinat R, Fan Y, Michael M, Rischin D, et al. Association of acquired Pelger-Huet anomaly with taxoid therapy. Brit J Haemat 1996;93:139–41.

[100] Schule B, Oviedo A, Johnston K, Pai S, Francke U. Inactivating mutations in ESCO2 cause SC phocomelia and Roberts syndrome: no phenotype-genotype correlation. Am J Hum Genet 2005;77:1117–28.

[101] Franco B, Meroni G, Parenti G, Levilliers J, Bernard L, Gebbia M, et al. A cluster of sulfatase genes on Xp22.3: mutations in chondrodysplasia punctata (CDPX) and implications for warfarin embryopathy. Cell 1995;81:1–20.

[102] Van Gaalen J, Kerstens FG, Maas RP, Harmark L, van de Warrenburg BP. Drug-induced cerebellar ataxia: a systematic review. CNS Drugs 2014;28:1139–53.

[103] Rossi M, Perez-Lloret S, Doldan L, Cerquetti D, Balej J, Millar Vernetti P, et al. Autosomal dominant cerebellar ataxias: a systematic review of clinical features. Eur J Neurol 2014;21:607–15.

[104] Penneys NS. Ochronosislike pigmentation from hydroquinone bleaching creams. Arch Dermatol 1985;121:1239–40.

[105] Langston JW, Ballard P, Tetrud JW, Irwin I. Chronic parkinsonism in humans due to a product of meperidine-analog synthesis. Science 1983;219:979–80.

[106] Priyadarshi A, Khuder SA, Schaub EA, Shrivastava S. A meta-analysis of Parkinson's disease and exposure to pesticides. Neurotoxicology 2000;21:435–40.

[107] Zimprich A, Biskup S, Leitner P, Lichtner P, Farrer M, Lincoln S, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron 2004;44:601–7.

[108] Du L, Sullivan CC, Chu D, Cho AJ, Kido M, Wolf PL, et al. Signaling molecules in nonfamilial pulmonary hypertension. N Engl J Med 2003;348:500–9.

[109] DuBose Jr. TD. Experimental models of distal renal tubular acidosis. Semin Nephrol 1990;10:174–80.

[110] Weiss HJ, Rosove MH, Lages BA, Kaplan KL. Acquired storage pool deficiency with increased platelet-associated IgG: report of five cases. Am J Med 1980;69:711–7.

[111] Bleiberg J, Wallen M, Brodkin R, Applebaum I. Industrially acquired porphyria. Arch Derm 1964;89:793–7.

[112] Cam C, Nigogosyan G. Acquired toxic porphyria cutanea tarda due to hexachlorobenzene. JAMA 1963;183:88–91.

[113] Ku NO, Wright TL, Terrault NA, Gish R, Omary MB. Mutation of human keratin 18 in association with cryptogenic cirrhosis. J Clin Invest 1997;99:19–23.

[114] Kotha K, Clancy JP. Ivacaftor treatment of cystic fibrosis patients with the G551D mutation: a review of the evidence. Ther Adv Respir Dis 2013;7:288–96.

[115] Herper M. The cost of creating a new drug now $5 billion, pushing big pharma to change. Forbes Magazine 2013; August 11.

[116] Goldberg P. An old drug's 21st century makeover begins with 84-fold price increase. Cancer Lett 2005; May 13.

[117] Berenson AA. Cancer drug's big price rise is cause for concern. New York Times 2006; March 12.

[118] Vanchieri C. When will the U.S. flinch at cancer drug prices? J Natl Cancer Inst 2005;97:624–6.

[119] Hurley D. Why are so few blockbuster drugs invented today? The New York Times 2014; November 13.

[120] Gelb BD. Marfan's syndrome and related disorders – more tightly connected than we thought. N Engl J Med 2006;355:841–4.

[121] Singh MN, Lacro RV. Recent clinical drug trials evidence in Marfan syndrome and clinical implications. Can J Cardiol 2016;32:66–77.

[122] Bar-Klein G, Cacheaux LP, Kamintsky L, Prager O, Weissberg I, Schoknecht K, et al. Losartan prevents acquired epilepsy via TGF-beta signaling suppression. Ann Neurol 2014;75:864–75.

[123] Lim DS, Lutucuta S, Bachireddy P, Youker K, Evans A, Entman M, et al. Angiotensin II blockade reverses myocardial fibrosis in a transgenic mouse model of human hypertrophic cardiomyopathy. Circulation 2001;103:789–91.

[124] Cohn RD, van Erp C, Habashi JP, Soleimani AA, Klein EC, Lisi MT, et al. Angiotensin II type 1 receptor blockade attenuates TGF-beta-induced failure of muscle regeneration in multiple myopathic states. Nat Med 2007;13:204–10.

[125] Plotkin SR, Merker VL, Halpin C, Jennings D, McKenna MJ, Harris GJ, et al. Bevacizumab for progressive vestibular schwannoma in neurofibromatosis type 2: a retrospective review of 31 patients. Otol Neurotol 2012;33:1046–52.

[126] Bose P, Holter JL, Selby GB. Bevacizumab in hereditary hemorrhagic telangiectasia. N Engl J Med 2009;360:2143–4.

[127] Eyetech Study Group. Anti-vascular endothelial growth factor therapy for subfoveal choroidal neovascularization secondary to age-related macular degeneration: phase II study results. Ophthalmology 2003;110:979–86.

[128] Curatolo P, Moavero R. mTOR inhibitors in tuberous sclerosis complex. Curr Neuropharmacol 2012;10:404–15.

[129] Tsang CK, Qi H, Liu LF, Zheng XF. Targeting mammalian target of rapamycin (mTOR) for health and diseases. Drug Discov Today 2007;12:112–24.

[130] Kim JH, Hu Y, Yongqing T, Kim J, Hughes VA, Le Nours J, et al. CD1a on Langerhans cells controls inflammatory skin disease. Nat Immunol 2016;17:1159–66.

[131] Mead AJ, Rugless MJ, Jacobsen SEW, Schuh A. Germline JAK2 mutation in a family with hereditary thrombocytosis. N Engl J Med 2012;366:967–9.

[132] Barosi G, Bergamaschi G, Marchetti M, Vannucchi AM, Guglielmelli P, Antonioli E, et al. JAK2 V617F mutational status predicts progression to large splenomegaly and leukemic transformation in primary myelofibrosis. Blood 2007;110:4030–6.

[133] Zhang L, Lin X. Some considerations of classification for high dimension low-sample size data, Stat Methods Med Res 2011; November 23. Available from: http://smmsagepubcom/content/early/2011/11/22/0962280211428387long. Accessed 26 January 2013.

[134] Mesa RA, Yasothan U, Kirkpatrick P. Ruxolitinib. Nat Rev Drug Discov 2012;11:103–4.

[135] Pesu M, Laurence A, Kishore N, Zwillich SH, Chan G, O'Shea JJ. Therapeutic targeting of Janus kinases. Immunol Rev 2008;223:132–42.

[136] McLornan D, Percy M, McMullin MF. JAK2 V617F: a single mutation in the myeloproliferative group of disorders. Ulster Med J 2006;75:112–9.

[137] Steensma DP, Dewald GW, Lasho TL, Powell HL, McClure RF, Levine RL, et al. The JAK2 V617F activating tyrosine kinase mutation is an infrequent event in both "atypical" myeloproliferative disorders and myelodysplastic syndromes. Blood 2005;106:1207–9.

[138] Verstovsek S. Therapeutic potential of JAK2 inhibitors. Hematol Am Soc Hematol 2009;2009:636–42.

[139] US FDA. FDA grants accelerated approval to pembrolizumab for first tissue/site agnostic indication. U.S. Food and Drug Administration; 2017 May 23.

[140] Dufourcq-Lagelouse R, Pastural E, Barrat FJ, Feldmann J, Le Deist F, Fischer A, et al. Genetic basis of hemophagocytic lymphohistiocytosis syndrome (review). Int J Mol Med 1999;4:127–33.

[141] Janka G, Zur Stadt U. Familial and acquired hemophagocytic lymphohistiocytosis. Hematol Am Soc Hematol Educ Program 2005;2005:82–8.

[142] Fletcher CD, Berman JJ, Corless C, Gorstein F, Lasota J, Longley BJ, et al. Diagnosis of gastrointestinal stromal tumors: a consensus approach. Int J Surg Pathol 2002;10:81–9.

[143] Berman J, O'Leary TJ. Gastrointestinal stromal tumor workshop. Hum Pathol 2001 Jun;32(6):578–82.

[144] O'leary T, Berman JJ. Gastrointestinal stromal tumors: answers and questions. Hum Pathol 2002;33:456–8.

[145] Burger H, den Bakker MA, Kros JM, van Tol H, de Bruin AM, Oosterhuis W, et al. Activating mutations in c-KIT and PDGFRalpha are exclusively found in gastrointestinal stromal tumors and not in other tumors overexpressing these imatinib mesylate target genes. Cancer Biol Ther 2005;4:1270–4.

[146] Heinrich MC, Joensuu H, Demetri GD, Corless CL, Apperley J, Fletcher JA, et al. Phase II, open-label study evaluating the activity of imatinib in treating life-threatening malignancies known to be associated with Imatinib-sensitive tyrosine kinases. Clin Cancer Res 2008;14:2717–25.

[147] Heinrich MC, Corless CL, Demetri GD, Blanke CD, von Mehren M, Joensuu H, et al. Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. J Clin Oncol 2003;21:4342–9.

[148] Selvi N, Kaymaz BT, Sahin HH, Pehlivan M, Aktan C, Dalmizrak A, et al. Two cases with hypereosinophilic syndrome shown with real-time PCR and responding well to imatinib treatment. Mol Biol Rep 2013;40:1591–7.

[149] Cools J, DeAngelo DJ, Gotlib J, Stover EH, Legare RD, Cortes J, et al. A tyrosine kinase created by fusion of the PDGFRA and FIP1L1 genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. N Engl J Med 2003;348:1201–14.

[150] Rennard SI, Vestbo J. The many "small COPDs", COPD should be an orphan disease. Chest 2008;134:623–7.

[151] Sorek R, Dror G, Shamir R. Assessing the number of ancestral alternatively spliced exons in the human genome. BMC Genomics 2006;7:273.

[152] Pagani F, Baralle FE. Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 2004;5:389–96.

[153] Fraser HB, Xie X. Common polymorphic transcript variation in human disease. Genome Res 2009;19:567–75.

[154] Venables JP. Aberrant and alternative splicing in cancer. Cancer Res 2004;64:7647–54.

[155] Srebrow A, Kornblihtt AR. The connection between splicing and cancer. J Cell Sci 2006;119:2635–41.

[156] Wiestner A, Schlemper RJ, van der Maas AP, Skoda RC. An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. Nat Genet 1998;18:49–52.

[157] Erwin DH. The origin of bodyplans. Amer Zool 1999;39:617–29.

[158] Valentine JW, Jablonski D, Erwin DH. Fossils, molecules and embryos: new perspectives on the Cambrian explosion. Development 1999;126:851–9.

[159] Bromham L. What can DNA tell us about the Cambrian explosion? Integr Comb Biol 2003;43:148–56.

[160] Budd GE, Jensen S. A critical reappraisal of the fossil record of the bilaterian phyla. Biol Rev Camb Philos Soc 2000;75:253–95.

[161] Love GD, Grosjean E, Stalvies C, Fike DA, Grotzinger JP, Bradley AS, et al. Fossil steroids record the appearance of Demospongiae during the Cryogenian period. Nature 2009;457:718–21.

[162] Bogler O, Cavenee WK. Methylation and genomic damage in gliomas. In: Zhang W, Fuller GN, editors. Genomic and molecular neuro-oncology. Sudbury, MA: Jones and Bartlett; 2004. p. 3–16.

[163] Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet 1999;23:185–8.

[164] Estivill X, Bancells C, Ramos C. Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. Hum Mutat 1997;10:135–54.

[165] Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev 2003;17:419–37.

[166] Tanackovic G, Ransijn A, Thibault P, Abou Elela S, Klinck R, Berson EL, et al. PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. Hum Mol Genet 2011;20:2116–30.

[167] Horike S, Cai S, Miyano M, Chen J, Kohwi-Shigematsu T. Loss of silent chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet 2005;32:31–40.

[168] Preuss P. Solving the mechanism of Rett syndrome: how the first identified epigenetic disease turns on the genes that produce its symptoms. Research News Berkeley Lab; 2004 December 20.

[169] Soejima H, Higashimoto K. Epigenetic and genetic alterations of the imprinting disorder Beckwith-Wiedemann syndrome and related disorders. J Hum Genet 2013;58:402–9.

[170] Agrelo R, Setien F, Espada J, Artiga MJ, Rodriguez M, P rez-Rosado A, et al.. Inactivation of the lamin A/C gene by CpG island promoter hypermethylation in hematologic malignancies, and its association with poor survival in nodal diffuse large B-cell lymphoma. J Clin Oncol 2005;23:3940–7.

[171] Bartholdi D, Krajewska-Walasek M, Ounap K, Gaspar H, Chrzanowska KH, Ilyana H, et al. Epigenetic mutations of the imprinted IGF2-H19 domain in Silver-Russell syndrome (SRS): results from a large cohort of patients with SRS and SRS-like phenotypes. J Med Genet 2009;46:192–7.

[172] Chen J, Odenike O, Rowley JD. Leukemogenesis: more than mutant genes. Nat Rev Cancer 2010;10:23–36.

[173] Martin DIK, Cropley JE, Suter CM. Epigenetics in disease: leader or follower? Epigenetics 2011;6:843–8.

[174] McKenna ES, Sansam CG, Cho YJ, Greulich H, Evans JA, Thom CS, et al. Loss of the epigenetic tumor suppressor SNF5 leads to cancer without genomic instability. Mol Cell Biol 2008;28:6223–33.

[175] Ikegawa S. A short history of the genome-wide association study: where we were and where we are going. Genomics Inform 2012;10:220–5.

[176] Platt A, Vilhjalmsson BJ, Nordborg M. Conditions under which genome-wide association studies will be positively misleading. Genetics 2010;186:1045–52.

[177] Beekman M, Blanch H, Perola M, Hervonen A, Bezrukov V, Sikora E, et al. Genome-wide linkage analysis for human longevity: genetics of healthy aging study. Aging Cell 2013;12:184–93.

[178] Couzin-Frankel J. Major heart disease genes prove elusive. Science 2010;328:1220–1.

[179] Field MJ, Boat T. Rare diseases and orphan products: accelerating research and development. Institute of Medicine (US) Committee on Accelerating Rare Diseases Research and Orphan Product Development, Washington, DC: The National Academics Press; 2010. Available from: http://www.ncbi.nlm.nih.gov/books/NBK56189/.

[180] Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 2010;86(1):6–22.

[181] Panagiotou OA, Evangelou E, Ioannidis JP. Genome-wide significant associations for variants with minor allele frequency of 5% or less – an overview: a HuGE review. Am J Epidemiol 2010;172:869–89.

[182] Omim. Online Mendelian inheritance in man, Available from: http://omim.org/downloads. Accessed 20 June 2013.

[183] Salmena L, Carracedo A, Pandolfi PP. Tenets of PTEN tumor suppression. Cell 2008;133:403–14.

[184] Brownstein MH, Mehregan AH, Bikowski JBB, Lupulescu A, Patterson JC. The dermatopathology of Cowden's syndrome. Brit J Derm 1979;100:667–73.

[185] Haibach H, Burns TW, Carlson HE, Burman KD, Deftos LJ. Multiple hamartoma syndrome (Cowden's disease) associated with renal cell carcinoma and primary neuroendocrine carcinoma of the skin (Merkel cell carcinoma). Am J Clin Pathol 1992;97:705–12.

[186] Schrager CA, Schneider D, Gruener AC, Tsou HC, Peacocke M. Clinical and pathological features of breast disease in Cowden's syndrome: an underrecognized syndrome with an increased risk of breast cancer. Hum Pathol 1998;29:47–53.

[187] Fitzpatrick DA. Horizontal gene transfer in fungi. FEMS Microbiol Lett 2012;329:1–8.

[188] Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 2008;9:605–18.

[189] Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Nishimura EO, et al. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci U S A 2015;112:15976–81.

[190] Kapitonov VV, Jurka J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 2005;3:e181.

[191] Baker NE. Patterning signals and proliferation in Drosophila imaginal discs. Curr Opin Genet Dev 2007;17:287–93.

[192] Woodhouse E, Hersperger E, Shearn A. Growth, metastasis, and invasiveness of Drosophila tumors caused by mutations in specific tumor suppressor genes. Dev Genes Evol 1998;207:542–50.

[193] Harmon A. The DNA age: searching for similar diagnosis through DNA. The New York Times 2007; December 28.

[194] Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 2012;109:1193–8.

[195] Schwartz JH, Maresca B. Do molecular clocks run at all? A critique of molecular systematics. Biol Theory 2006;1:357–71.

[196] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol 2006;4:e88.

[197] Eby J. The prevalence of antibiotic-resistant bacteria on mosquitoes collected from a recreational park, In: Proceedings of the National Conference on Undergraduate Research, April 15–17; 2010.

[198] Angert ER, Clements KD, Pace NR. The largest bacterium. Nature 1993;362:239–41.

[199] Bergsten J. A review of long-branch attraction. Cladistics 2005;21:163–93.

[200] Berman JJ. Taxonomic guide to infectious diseases: understanding the biologic classes of pathogenic organisms. First edition Cambridge, MA: Academic Press; 2012.

[201] Berman JJ. Data simplification: taming information with open source tools. Waltham, MA: Morgan Kaufmann; 2016.

[202] Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 2004;5:R7.

[203] Koonin EV, Galperin MY. Sequence, evolution, function: computational approaches in comparative genomics. Boston: Kluwer Academic; 2003.

[204] Holland HD. The oxygenation of the atmosphere and oceans. Philos Trans R Soc Biol Sci 2006;361:903–15.

[205] Sperling EA, Frieder CA, Raman AV, Girguis PR, Levin LA, Knoll AH. Oxygen, ecology, and the Cambrian radiation of animals. Proc Natl Acad Sci U S A 2013;110:13446–51.

[206] Greene CS, Troyanskaya OG. Chapter 2: Data-driven view of disease biology. PLoS Comput Biol 2012;8:e1002816.

[207] DeQueiroz K. Ernst Mayr and the modern concept of species. PNAS 2005;102(suppl 1):6600–7.

[208] Bloom BJ, Nicholson TL, Williams JR, Campbell SL, Bishof M, Zhang X, et al. An optical lattice clock with accuracy and stability at the 10-18 level. Nature 2014;506:71–5.

[209] Satir P. CILIA: before and after. Cilia 2017;6:1.

[210] Jakobsen L, Vanselow K, Skogs M, Toyoda Y, Lundberg E, Poser I, et al. Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. EMBO J 2011;30:1520–35.

[211] Tang Z, Zhu M, Zhong Q. Self-eating to remove cilia roadblock. Autophagy 2014;10:379–81.

[212] Zhang Z, Harrison P, Gerstein M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 2002;12:1466–82.

[213] Poliseno L. Pseudogenes: newly discovered players in human cancer. Sci Signal 2012;5:5.

[214] U.S. Orphan Drug Act. Comment. Prior to the passage of this act, there was virtually no development of new drugs for the treatment of rare diseases. This ignored, rare diseases were effectually orphaned by the drug research community. The Orphan Drug Act of 1983 encouraged the development of new drugs for rare diseases by offering tax advantages and an extended period of exclusive marketing right for participating companies. The Orphand Drug Act is considered a legislative triumph, as the development of new drugs for rare diseases increased after passage of the bill, Available from: http://www.fda.gov/orphan/oda.htm; 1983.

[215] Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, et al. A “silent” polymorphism in the MDR1 gene changes substrate specificity. Science 2007;315(5811):525–8.

[216] Baldauf SL. An overview of the phylogeny and diversity of eukaryotes. J Syst Evol 2008;46:263–73.

[217] Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, et al. Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat Genet 2011;43:864–8.

[218] Kivela T. Trilateral retinoblastoma: a meta-analysis of hereditary retinoblastoma associated with primary ectopic intracranial retinoblastoma. J Clin Oncol 1999;17:1829–37.