Chapter 9

Polymorphisms

How do we find a gene that contributes to a disorder or a behavior? The technology for such gene hunting has been revolutionized several times over the past thirty years. In fact, some techniques considered “hot” only a decade are now obsolete. The remarkable progress has been due to one major phenomenon–the ability to detect and cheaply genotype “spelling variations” in the human gnome. To understand this, we must introduce (or refresh) some terminology.

Suppose that we gathered DNA from a large number of individuals and examined the first 1,000 nucleotides on the first chromosome. There would be some sections in which the base pair sequence is the same for all of the DNA strands. These are called *monomorphic* sections. There will be other sections in which the sequence of nucleotides differs. These are called *polymorphic* sections or, simply, *polymorphisms*. Using the analogy of a four-letter DNA alphabet, polymorphisms are really “spelling variations.” The term *allele* refers to a specific spelling variation.

The alleles at a polymorphic section are called either *mutants* or *common polymorphisms* depending on their frequency. A mutant allele has a frequency of less than 1% in the general population. Alleles with a frequency higher than 1% are considered “common.”

Before the 1980s, finding the genotype of an individual usually involved various laboratory assays for the product of a gene—the protein or enzyme–but not the gene itself. The cases of the ABO and Rhesus blood groups are classic examples of how one infers genotypes from the reaction of gene products with certain chemicals. The actual number of known polymorphisms was probably in the low 100s. As a result, for most of the 20th century attempts to find the genes for many Mendelian disorders were unsuccessful.

In the mid 1980s, genetic technology took a great leap forward with the ability to genotype the DNA itself. The geneticist could now examine the DNA directly without going through the laborious process of developing assays to detect individual differences in proteins and enzymes. Direct DNA analysis had the further advantage of being able to identify alleles in sections of DNA that did not code for polypeptide chains. As a result of these new advances, the

Table 9.1: Types of polymorphisms.

I. Protein/enzyme polymorphisms

II. DNA polymorphisms.
- A. Single nucleotide polymorphisms (SNPs)
- B. Tandem repeat polymorphisms
- C. Structural polymorphisms (deletions, inversions, etc.)
- D. Sequence polymorphisms

number of polymorphic regions increased exponentially.

The result was spectacular. The location and nature of the genes for Mendelian disorders like Huntington's disease and cystic fibrosis had remained a mystery for the better part of the 20th century. Within 10 to 15 years, these genes, as well as those for most genetic disorders, had been located and partially characterized.

In this chapter, we outline various types of polymorphisms and the techniques used to detect them.

9.1 Types of polymorphisms

Table 9.1 presents an overview of the major types of polymorphisms. They are divided into two major categories according to how .

9.1.1 Protein/enzyme polymorphisms

In the early days of human genetics, the majority of polymorphisms were those associated with proteins and enzymes. To detect the polymorphism and a person's genotype, one performed assays for the gene product, i.e., the protein or enzyme produced by the genetic blueprint.

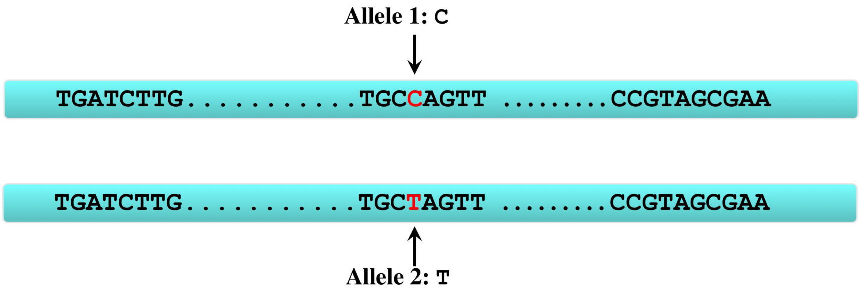
Most of these polymorphisms were detected in blood. When your blood is typed, you are informed that you are blood group O+ or AB- or A+, etc. The letter in this blood group gives your phenotype at the ABO locus, and the plus (+) or minus (-) sign denotes your phenotype at the Rhesus gene.

A number of other loci such as Kell, Duffy, MN, and Kidd can also be phenotyped from blood. These polymorphisms are still used today to assess suitability of donors and recipients for blood transfusions (ABO locus) and to assess Rhesus incompatibility between a mother and her fetus. However, blood group polymorphisms have given way to other, more sophisticated techniques in modern human genetic research.

9.1.2 DNA polymorphisms

The other large class of polymorphisms are those that detect spelling variations at the level of DNA nucleotides. For our purposes, we can classify them into

Figure 9.1: Example of a single nucleotide polymorphism.


Allele 1: C
TGATCTTG...........TGCCAGTT ....CCGTAGCGAA
TGATCTTG...........TGCTAGTT ....CCGTAGCGAA
Allele 2: T

three types, each of which is discussed below.[1]

9.1.2.1 Single nucleotide polymorphisms

A *single nucleotide polymorphism* or *SNP* is a sequence of DNA on which humans vary by one and only one nucleotide (see Figure 9.1). Because humans differ by one nucleotide per every thousand or so nucleotides, there are millions of SNPs scattered throughout the human genome.

The major advantage of SNPs, however, lies in the fact that they can be detected in a highly automated way using specialized DNA "chips" usually called *DNA arrays*.

9.1.2.2 Tandem repeat polymorphisms

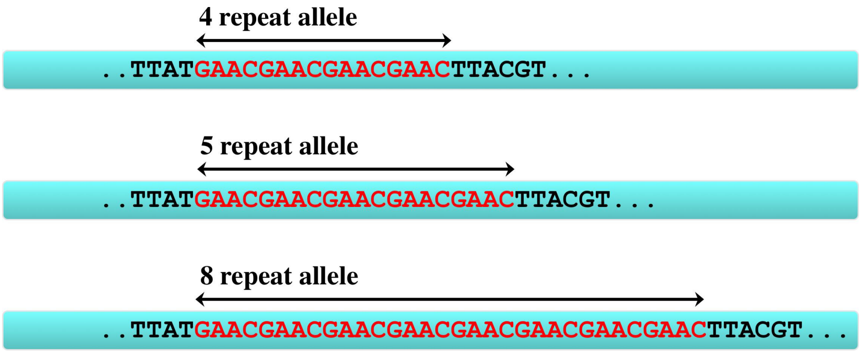
A tandem repeat polymorphism consists of a series of nucleotides that are repeated in tandem (i.e., one time after another). The polymorphism consists of the number of repeats. Figure 9.2 illustrates this type of polymorphism. The repetitive nucleotide sequence is `GAAC` and the figure depicts three alleles–a four-repeat allele, a five-repeat allele, and an eight-repeat allele.

9.1.2.3 Tandem repeat terminology (graduate)

Unfortunately, even though the concept of the tandem repeat is quite simple, the terminology for referring to these polymorphisms can be confusing. When the number of repeats is small, usually five to six or fewer, then the polymorphism may be called a *microsatellite*, *simple sequence repeat* (SSR), or *short tandem repeat* (STR). When the number of repeated nucleotides is larger, then the polymorphism may be called a *minisatellite*, particularly when it is located in a telomere. Finally, the term *variable number of tandem repeats* (VNTR) polymorphism has been used equivocally. Sometimes it is generic and refers to

[1] There are other DNA polymorphisms such as the restriction fragment length polymorphism (RFLP), but these have largely been superseded with new technologies. Hence, they are not discussed.

Figure 9.2: A tandem repeat polymorphism.


4 repeat allele
..TTATGAACGAACGAACGAACTTACGT...
5 repeat allele
..TTATGAACGAACGAACGAACGAACTTACGT...
8 repeat allele
..TTATGAACGAACGAACGAACGAACGAACGAACGAACTTACGT...

any tandem repeat polymorphism. At other times, it refers to repeat with a large involving a large number of nucleotides.

9.1.2.4 Structural variants

Here, we use the term structural variants to refer to spelling variations that involve deletions or insertions of a nucleotide sequence, inversions, and translocations. When the structural variant is somewhat large (some geneticists define "large" as 1 kilobase or more, others 10 kb), the polymorphism is called a *copy number variant* or *CNV*.[2]There is considerable research being done on CNVs and medical disorders, including psychopathology (see Section X.X).

Insertion-deletion polymorphisms or *indels*, an example of which is given in Figure 9.3, are intuitive. Whether an allele is called an insertion or deletion, however, depends on the consensus nucleotide sequence of the human genome. If an allele is missing a nucleotide sequence that is present in the consensus sequence, then the allele is a called a *deletion.* If the allele contains a nucleotide sequence that is not in the consensus sequence, then it is an *insertion.*

A particularly important type of insertion occurs when a section of DNA is duplicated and inserted into the same region. Remember pseudegenes from Section X.X? These are sections of DNA with a nucleotide sequence very similar to a known gene but the DNA does not produce a functional polypeptide. Most pseudogenes resulted from duplications.

An *inversion* polymorphism occurs when one allele has a nucleotide sequence that is reversed in another allele. Figure 9.4 presents an example. Assume that the spelling variation in the consensus sequence is the one on the top. The inverted allele has a section that has the same spelling but is "read in reverse" from right to left instead of the ordinary left to right order.

[2] Like many phenomena in molecular biology, the CNV has been defined partly by the laboratory techniques used to detect the polymorphism. A deletion of five nucleotides is difficult to detect with today's technology, but deletions of several thousands of base pairs can be observed with current automated technology. Hence, the size limit for a CNV is a pragmatic issue not a biological one.

Figure 9.3: An insertion-deletion polymorphism.

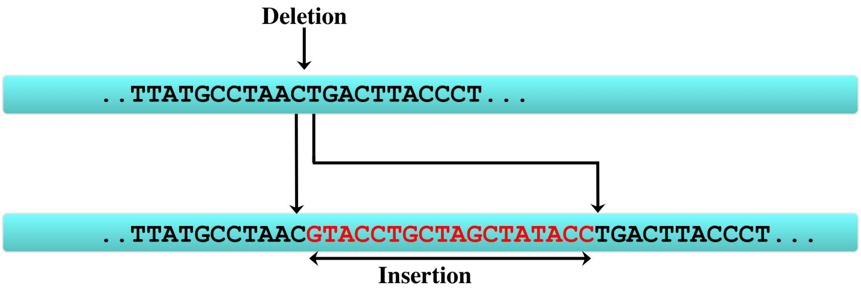
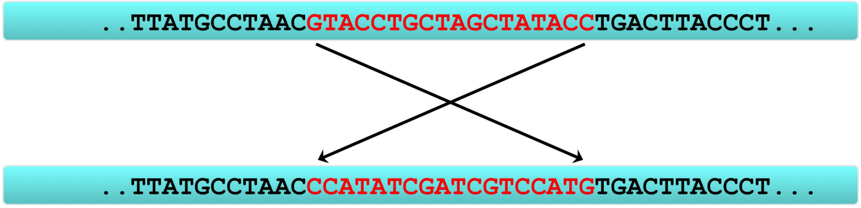

Deletion
..TTATGCCTAACTGACTTACCCT...
..TTATGCCTAACGTACCTGCTAGCTATACCTGACTTACCCT...
Insertion

Figure 9.4: An inversion structural variant.


..TTATGCCTAACGTACCTGCTAGCTATACCTGACTTACCCT...
..TTATGCCTAACCCATATCGATCGTCCATGTGACTTACCCT...

Inversions usually occur when a chromosome breaks in two places and DNA repair mechanisms mistakenly splice the middle fragment back but in reverse order. Typically they involve many thousands of base pairs.

A *translocation* occurs when a section of DNA is deleted from one chromosome and then inserted into another chromosome.

9.1.2.5 Sequence polymorphisms

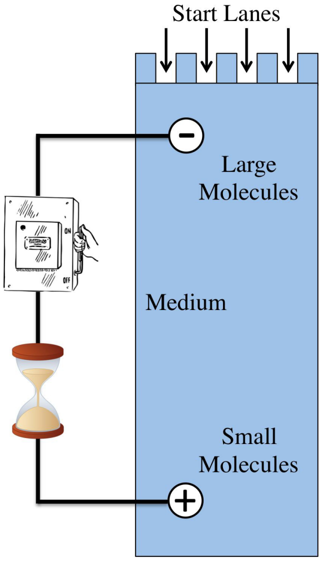
The ultimate polymorphism is to actually have the whole sequence of nucleotides for a region for a large number of DNA strands and then examine all of the differences among the strands. Here, the DNA differences could be a SNP, a tandem repeat or a structural change. There is no accepted term for this phenomenon, so we call them *sequence polymorphisms*. In effect, sequence polymorphisms subsume all known DNA polymorphisms. Section 9.3.4 below provides details on modern sequencing.

9.2 Basic techniques in molecular genetics

This section is short and merely defines several of the major techniques used in molecular genetics. There are many good web resources that cab help you learn more about them.

9.2.1 Electrophoresis

Figure 9.5: Schematic for electrophoresis.



Electrophoresis is a generic chemical technique that separates molecules by their molecular weight and/or electronic charge (see Figure 9.5). One places purified biological material in a starting lane in a viscous liquid medium. An electric current is passed through the medium for a specified time. The molecules with a charge opposite to the electrode at the far end of the medium will migrate to the opposite end of the medium. The viscosity of the liquid, however, will impede the migration of large molecules more than small ones. Hence, at the end of a session, the smaller molecules will have moved further from the start lanes than the larger molecules. Current electrophoretic techniques are so sensitive that they can distinguish two DNA or RNA fragments that differ by only a single nucleotide.

Electrophoresis is used to detect tandem repeat polymorphisms and indel (insertion/deletion) polymorphisms. The logic is straightforward. If one allele in a tandem repeat polymorphism has 11 repeats while another has 16 repeats, then the 11 repeat DNA should move further on the electrophoretic medium than the 16 repeat fragment. The problem is that we require a technology to actually *see* the DNA. This is where our next tool–a DNA probe–comes into play.

9.2.2 Probes

A probe is a manufactured fragment of single-stranded DNA or RNA with a predetermined nucleotide sequence. It is introduced into a medium (such as the electrophoretic medium) so that it may bind to its complementary single-stranded DNA or RNA fragment. Usually the probe is comprised of nucleotides with specially colored fluorescent tags that will glow under appropriate lighting.

To detect desired DNA fragments in electrophoresis, one “baths” the medium in probes, allowing enough time for them to bind to their complements. Remaining single-stranded probes that did not bind are then washed away and the medium is viewed under ultraviolet light. The result are visible bands in the electrophoretic medium. See the section on the U.S. Federal Bureau of Investigation’s CODIS system (Section X.X) for an example of how this technique is used in forensic applications.

9.2.3 Polymerase chain reaction

Imagine that you are a crime scene investigator who finds a tiny droplet of blood at a crime scene. How can you obtain enough DNA from such a small specimen to perform an analysis. The answer is the polymerase chain reaction or PCR. The technique involves a soup comprised of the DNA that you purified from the specimen, a large number of free nucleotides, some of those "replication stuff" enzymes that produce two copies of DNA from a single copy, and a number of *primers* (a DNA fragment with a nucleotide sequence specific to the DNA area you want to copy).

The first step in PCR is to heat this soup to just about the boiling point of water. This breaks the bonds for double-stranded DNA, making it single stranded. As the mixture cools, the primers in the soup will join with their complementary single-stranded DNAs from the specimen and the "replication stuff" will attach free nucleotides, making them double stranded.

Hence, if your specimen had, say, 1,000 copies of the person's DNA in the white blood cells, then after one round of PCR, you would have 2,000 copies of the desired region. Need more? Then do a second round bringing your total to 4,000 copies. By 15 rounds, you would have over 30 million copies–plenty for analysis.

9.3 Detecting polymorphisms

Methods used for detecting polymorphisms depend on the type of polymorphism. One technique genotypes SNPs while another detects tandem repeats. A second consideration is the purpose for genotyping. Some research studies require genotyping a million polymorphisms on many thousands of participants. Here, the cost of an individual genotype must be low. In a clinical setting, however, the issue may be to confirm or rule out the diagnosis of a genetic or genetically influenced syndrome. Here, a more expensive–but also more discriminating–techniques may be used.

The following is a highly simplified overview of the major techniques used to detect polymorphisms. The purpose is to present the logic of the techniques. As a result, many important laboratory steps are overlooked and over simplified.

DNA is a very long molecule, so the first step in most procedures is *DNA fragmentation*. This "cuts" the DNA into short fragments that can then be used for the procedures. There are several ways to fragment DNA, and they range from the chemical (slice the DNA using enzymes) to physical (force the DNA through a nebulizer).

9.3.1 Tandem repeat polymorphisms

Traditionally, tandem repeat polymorphisms have been assayed with using electrophoresis and then probes. After fragmentation, the relevant loci are amplified through PCR and the PCR products are then separated by electrophoresis. The medium (actually, something that extracts the DNA fragments from the medium) is bathed in the relevant probes. Single-stranded probes that did not bind to their complementary PCR products are washed away. Electrophoresis will then separate the remaining strands according to size.

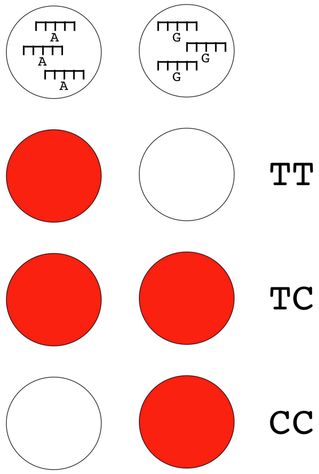
The nucleotides used int he PCR are special–they have fluorescent tags. Hence, the newly synthesizes strands will be visible under the appropriate illumination. A laser sensitive to the fluorescence will scan the output of the electrophoresis and report "hits" to a computer. which records these and saves the data.

See Section X.X for a more detailed explanation of this in the context of forensic science.

9.3.2 SNPs

Today, detection of SNPs is done through large scale *DNA arrays* often termed *microarrays*. An illustration of how they work is presented in Figure 9.6. The SNP of interest has two alleles–`T` and C. The first step is to manufacture a single-stranded DNA section that is both unique to and complementary to the `T` allele. This, of course, will have a adenine (`A`) in the position complementary to the `T`. A second single-stranded DNA probe is manufactures that is unique to and complementary to the `C` allele; it will have a `G`. A technique like PCR is then used to make a very large number of these sections. Then, the `A` strands are glued onto a very tiny area of the array and the `G` strands onto a tiny adjacent area. This gives the top row in Figure 9.6.

Figure 9.6: How DNA arrays detect SNPs.



Next DNA is extracted from a biological specimen taken from the person. The DNA is purified, cut into millions of fragments and then amplified using PCR. But this amplification is done with a twist. Instead of an ordinary batch of free nucleotides, the nucleotides used in the PCR have fluorescent tags that will make them "glow" when viewed under a certain light. The whole DNA array is then bathed in the fluorescently labeled, single-stranded DNA from the PCR.

The DNA strands from the PCR will tend to bind (or *hybridize*, as the molecular biologists call it) with their complementary single-stranded DNA fragments on the DNA array. After a suitable time, the remaining single-stranded DNA is washed from the array. The array is exposed to the special light and a laser scans the array. As the array is being scanned, the areas that fluoresce are detected by the laser and recorded in a computer.

Now, return to our polymorphism in Figure 9.6 and ask what the laser would "see" if the person's genotype were `TT`. The single-stranded DNA with the `T` would bind to the complementary A strand on the chip, making the `A` area light up. Since the person lacks the C allele, there would be no strand complementary to the `G` section of the array. Hence, this area would not light up. This pattern is seen in the second row in Figure 9.6.

If the person were homozygous for the `C` allele, then we would observed the opposite pattern, i.e., the one on the last row of the figure. The area with the `G` DNA strand would fluoresce while that with the `A` strand would remain unlighted.

Finally, the DNA from a `TC` heterozygote would bind with both the `A` and the `G` sections, giving the pattern in the third row of the figure.

9.3.3 CNVs

There are many different ways to detect copy number variants (CNVs). Here, the purpose is paramount. Consider testing for a microdeletion in clinical cytogenetics. A *microdeletion* is a deletion involving many kb but is too small to detect using traditional karyotypes. Usually, the medical doctor suspects that an infant or young child may have a specific syndrome due to a microdeletion and requests tests to confirm or rule out that syndrome. Hence, the test is for one CNV and there is no need to use a method for cataloging all of the thousands of known CNVs.

There are many techniques used to detect CNVs in research designs intended to see which CNVs may be associated with a disorder or trait. One strategy is digital or virtual karyotyping (Wang et al., 2002). Here, one uses existing strategies for detecting polymorphisms and then applies software to look for CNVs. For example, a CNV deletion could be detected in a DNA array that assays SNPs by looking for a region where there is no heterozygosity.

9.3.4 Next generation sequencing

The Holy Grail for genotyping an individual is to obtain the complete nucleotide sequence of the person's genome. The Human Genome Project sequenced one human genome. It took about 10 years and cost three billion dollars. Today a variety of new technologies are emerging to sequence an individual's genome (Koboldt et al., 2010; Mardis, 2013). Collectively, they are called *next generation sequencing* (*NGS*) technologies or *massive parallel sequencing.* It is too early to predict which ones will prevail, but early results on the potential of NGS are striking. The current goal is the $1K genome, i.e., a procedure to obtain an individual's genome for $1,000 US.

Despite using very different laboratory methods, the logic of most NGS strategies is the same. The DNA is fragmented and then amplified. The PCR products are then sequenced in parallel. That is, millions to billions of the fragments are sequenced at the same time and the results stored into a computer.

Finally, computer algorithms are used to “align” the short segments into a long sequence.

There are several varieties of NGS. Although we have spoke of sequencing a whole genome, *targeted sequencing* selects specific regions of the genome for sequencing (Koboldt et al., 2010). A particular type of targeted sequencing selects the exons and nearby regions, providing a sequence of what is called the *exome*. Another NGS strategy is to focus on the various types of RNA, particularly mRNA in the study of gene expression in animal brains (Hitzemann et al., 2013).

Because so much of the methodology for NGA is still in the development stage, there are some rough spots with the technology. Standard and protocols for NGS in research have been proposed (Goldstein et al., 2013), but it will take several years of data collection to come up with accepted standards.

Eventually, NGS will supersede all of the methods mentioned above for detecting polymorphisms. The major reason is that with a genome at hand, all one needs is software to search it, compare it to a genomic library of variants, and spit out out the tandem repeats, SNPs, and structural variants. That said, such a technology is not readily available today. Currently, supercomputing resources are required to store the data, align the fragments, and arrive at an unambiguous sequence. Still, advances in technology–on both the chemistry and the data side–will make is possible for individuals to obtain their own genomic sequence in the future.

One current parameter underlying current sequencing is the *number of reads* denoted as *X*. Having fractionated and amplified the DNA, the number of reads is, roughly speaking, the number of times that this biological material is put through the sequencing step that “reads” the DNA sequence. A 1X sequence is the cheapest and does it once. A 25X sequence performs it 25 times. There is no gold standards for X. The number of reads all depends on the purpose at hand.

The 1,000 Genomes Project is the latest, large-scale, international attempt to catalog human genetic variation (The 1,000 Genomes Project Consortium, 2012). Using several NGS technologies, it has reported the nucleotide sequence of 1,092 people of different ethnic groups throughout the world. It estimates that the human genome contains 38 million single nucleotide polymorphisms, 1.4 million short indels (insertion/deletion polymorphisms), and 14,000 copy number variants.

9.3.4.1 NGS and personalized medicine

There is considerable speculation about the implications of the $1K genome for personalized medicine. Personalized medicine involves customization of medical procedures and therapeutics so that they apply to the individual, not to the collection of individuals with a certain disorder. We have all experienced it to a certain degree. For example, hay fever (allergic rhinitis) sufferers respond differently to the antihistamines used to mange the problem. The typical course of treatment is to try this drug and then that one until, by chance, the patient arrives at one that controls the symptoms with a minimum of annoying side effects. The goal of personalized medicine is to develop tests that predict how a patient will respond to each drug and then start with the one likely to be the most efficacious.

Given the role that genetics play in the etiology of disorders and in drug responses, it is natural that a person's genotypes on many loci will be relevant information for personalized medicine. There are already plans to develop protocols and educational programs for both patient and health care providers (see Biesecker et al., 2012; Katsanis and Katsanis, 2013). It is premature to speculate on how genomics and personalized medicine will evolve, but there are some scenarios that pose ethical, moral, and economic questions. Think about the following case.

Consider the economics of genotyping on an an hoc basis for a number of different medical conditions over the lifespan of a person. A neonatal screen for common Mendelian disorders may cost \$180; a diabetes screen, \$125; a profile for cholesterol transport and potential build up, \$200; and so on. How much would it cost over a person's lifetime to perform these specific tests?

If a person's whole genome could be sequenced for \$1, 000, would it be better to just do that at birth and record the information in the newborn's medical record? Computer algorithms can search the sequence as needs arise during the person's life.

Such a scenario raises issues about confidentiality, consent and authorization, and perhaps even intellectual property (who "owns" the neonate's genome?) Will rising health care costs force us into this strategy as a cost-cutting measure? We must begin thinking about such issues.

9.4 References

Biesecker, L. G., Burke, W., Kohane, I., Plon, S. E., and Zimmern, R. (2012). Next-generation sequencing in the clinic: are we ready? *Nature reviews.Genetics*, 13(11):818–824.

Goldstein, D. B., Allen, A., Keebler, J., Margulies, E. H., Petrou, S., Petrovski, S., and Sunyaev, S. (2013). Sequencing studies in human genetics: design and interpretation. *Nature reviews.Genetics*, 14(7):460–470.

Hitzemann, R., Bottomly, D., Darakjian, P., Walter, N., Iancu, O., Searles, R., Wilmot, B., and McWeeney, S. (2013). Genes, behavior and next-generation rna sequencing. *Genes, brain, and behavior*, 12(1):1–12.

Katsanis, S. H. and Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature reviews.Genetics*, 14(6):415–426.

Koboldt, D. C., Ding, L., Mardis, E. R., and Wilson, R. K. (2010). Challenges of sequencing human genomes. *Briefings in bioinformatics*, 11(5):484–498.

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 6:287–303.

Wang, T. L., Maierhofer, C., Speicher, M. R., Lengauer, C., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002). Digital karyotyping. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):16156–16161.