Information in this guide is prepared and presented by Bryant McAllister, Associate Professor of Biology at The University of Iowa. This and other resources for understanding the interpretations and uses of personal DNA tests are available at the Genetics for Genealogy website.
An ancient sexual cycle governs the transmission of genetic information from generation to generation in humans and most other sexual organisms. Although there are lots of variations on the common theme, the general principles of inheritance apply to single-celled fungi, plants, and animals including humans. At conception, an individual inherits from the mother a set of chromosomes present in the egg and from the father a matching (or homologous) set of chromosomes in the fertilizing sperm. DNA is the recipe for cellular and organismal function and a unique DNA molecule forms the centerpiece of each chromosome in each maternal and paternal set.

Two basic rules describe the inheritance of chromosomes. These are known as Mendel's Laws or Mendelian Principles named for their discoverer, Gregor Mendel, who performed and studied crosses in the garden pea. Discoveries in the early 20th century revealed the relationship between the rules of inheritance of traits described in 1866 by Mendel and the transmission of whole chromosomes including the exchange between chromosome pairs to produce new combinations between the DNA inherited from each parent. These processes of genetic transmission produce predictable patterns of diminishing DNA identity between increasingly more distant relatives. A property upon which genome-wide SNP identity matching uses to find relatives and predict relationships.
Three primary providers of genome-wide SNP tests assess an individual’s DNA and return to the customer a list of DNA matches with predicted levels of relationship. Comparisons are made among customers within each company database, thus each customer database is private to each company and not shared. Both 23andMe and AncestryDNA require that a new genetic analysis is performed – on DNA extracted from a saliva sample provided by the customer – whereas Family Tree DNA will accept raw genetic data produced by other companies. GEDmatch is an open database that accepts raw data files produced by each of these companies.
Using genetics in genealogical research concerns the intersection between DNA-based inferences of common ancestry and historical inferences of who these ancestors were. Analysis of an individual’s DNA and comparison of those data with others reveals relatives connected through recent common ancestor(s). Comparison of family history with a matching relative may provide evidence on the identity of ancestors potentially responsible for the shared genetic similarity. As the number of genetic relatives increases, each appearing as a DNA match with other descendants of a common identified ancestor(s), the greater the confidence in the identity of the common connection.
DNA is the recipe of the cell. DNA is a molecule consisting of two very long polymers of nucleotides wrapped together in helical form. The informational bits of the polymer are the series of inner pairs of bases Adenine (A) and Thymine (T) and Guanine (G) and Cytosine (C). The goal of The Human Genome Project was to determine a reference sequence of the 3 billion bases present in the DNA molecules passed from a parent to a child. DNA functions primarily to store information in the cell and it is copied with great fidelity each time a cell divides. Products, mostly proteins, are synthesized for cellular activities using information encoded in the sequence of bases in DNA.

Your unique DNA composition was set at your conception by the approximately 3 billion basepairs (Gbp) of DNA in your dad’s sperm combined with a similar set of 3 Gbp in the DNA of your mom’s egg at fertilization. This genome established in the single cell from which you developed has been dutifully copied and is shared by all cells of your body.
DNA molecules are organized into chromosomes - of which humans have 23 pairs - one member of each pair contributed by each parent. One pair of chromosomes, the sex chromosomes, relates to the sex of parent and child. Males have a mismatched pair of sex chromosomes referred to as the X and Y chromosomes. The other 22 pairs are the autosomes. Each autosomal pair contains distinct genetic information encoded in its DNA sequence and often differs in appearance to other pairs. The centromere is a distinct feature located in many cases near the center of the X shape of the chromosome and in other cases near one end.

The 22 autosome pairs are essentially like different books containing unique information encoded in a common language. Each copy of Chromosome 1 is a different edition of a particularly long book with 249,000,000 basepairs (bp) of information in its DNA sequence, whereas chromosome 21 is the shortest book with only 47,000,000 bp. Moreover, the editions of your chromosomes are unique to you, and each of your children has a set of chromosomes unique to them. Although the DNA sequence of the chromosomes is unique for any single person, regions of sequence identity are present in the DNA of close relatives.
The genome-wide SNP analysis examines the bases present at about 0.02% of the sites in the DNA of your genome. These sites are distributed along each of the autosomes, and depending on the company, sites are also distributed along the other chromosomes. The Illumina SNP chip is the common platform for these analyses. Analysis of DNA extracted from cheek cells reveals the identity of the bases at specific sites that are known to vary in the human population. The bases present at these Single Nucleotide Polymorphisms (SNPs) comprise the raw data interpreted from the chip. Given a particular version of the chip used by a company, the same sites referenced by position in the genome are assayed for each customer.
This table represents a small amount of the raw data produced from the SNP chip. Each row contains an identifier (rsid) of the site analyzed, the chromosomal location of the site, and the position along the DNA sequence of that chromosome. The position of each site is addressed to the reference sequence produced by The Human Genome Project. Note that the 16 sites shown here are distributed over a region of Chromosome 1 containing about 15,000 basepairs, so only about 0.1% of the sites in this region are analyzed. A genotype or genetic profile is given for each site. In some cases both copies of Chromosome 1 have the same base (e.g., AA), and in other cases a different base is present on the two chromosomes (e.g., AG). Although the DNA molecule of each chromosome is comprised of basepairs, only a single base (A, C, G, or T) in the pair is identified at each SNP for each chromosome.
Segregation of chromosomes is a fundamental principle of biology, and refers to the transmission of the members of each chromosome pair. Only one member of each chromosome pair is transmitted to a child. Each child inherits a single member from each parent, thus restoring the pairs of chromosomes. In the diagram above, the chromosomes of the parents are colored differently to show the inheritance of one member of each chromosome pair from each parent. Because of segregation, exactly half of the autosomes of each normal child is comprised of genetic information from each parent.
This diagram illustrates the 50% identity between a daughter and her mother by direct comparison of their SNP data distributed across the autosomes and the X chromosome. Just like autosomes, females inherit an X from their mother and their father. All regions for which SNP data are compared exhibit a “half-identical” pattern and are colored light blue in the Genome-Wide Comparison. At each SNP genotype in the daughter, at least one chromosome matches a chromosome of the mother. For example, the daughter could be GG and the mother TG showing that the daughter inherited from her mother the chromosome with the G at that SNP. Comparisons of genotypes at many SNPs distributed along each chromosome provide the information for coloring the chromosomes in this overview of the genome.
Consider that parents and their offspring share 50% genetic identity across the autosomes due to the process of segregation. Thus, the expected identity of a person compared to a grandparent is 25%, and the identity to a great grandparent is 12.5%. Segregation ensures that the identity between parent and child is 50%; exceptions occurring in cases where the parents are close biological relatives. While the genetic contribution is expected to decrease by 50% each subsequent generation, this reduction is not exact. Other genetic processes, independent assortment of different chromosome pairs and exchange within chromosome pairs, produce variability in the actual contributions realized from each grandparent and more distant ancestor.
While each chromosome pair segregates so that only one member of the pair is transmitted to each child, the different chromosome pairs segregate independently. This is illustrated in the transmission of the different colored chromosomes in the intermediate generation. Some children inherit the same colored chromosomes at the two pairs, whereas other children inherit different colored chromosomes. This is due to the independence of the chromosomes in different pairs. Given only two pairs of chromosomes, the 4 illustrated outcomes are all equally likely. The 22 pairs of autosomes in humans produce a greater range of possible outcomes with the overall expectation being that the grandparents contribute equally to the composition of autosomes.
Chromosomes are often not transmitted from one generation to the next as a complete unit. Rather, exchange between the members of each chromosome pair commonly produces a novel DNA sequence combining regions from one parent with regions from the other parent. Through independent assortment of the different chromosome pairs, and the exchange within chromosome pairs, novel combinations of DNA are transmitted to each child. This diagram is targeted on the one intermediate individual inheriting differently colored parental chromosomes and recombining them to produce 6 unique children, but segregation, independent assortment, and exchange occur every generation to produce unique combinations of genetic variation among individuals. These processes also produce distinct patterns of identity within chromosomal segments of close relatives.
This diagram illustrates a comparison of SNP data between a grandmother and her granddaughter. Light blue shading along the chromosomes indicate regions where at each SNP genotype, one chromosome of the granddaughter matches at least one chromosome of the grandmother. The patchwork of light blue segments within chromosomes is the result of recombination. Independent assortment is evident by the distinct patterns of identity with the grandmother among the different chromosome pairs, even including absence of genetic identity between the granddaughter and grandmother at Chromosomes 11, 21, and 22. Due to the absence of genetic identity throughout these three chromosomes, and the fact of segregation that the granddaughter inherited a Chromosome 11, 21, and 22 from her mother, it is expected that the granddaughter inherited versions of these chromosomes with identity to her maternal grandfather.
This comparison using the DNA Relatives DNA tool in 23andMe illustrates how each chromosomal region is identical to one or the other maternal grandparent. Identical matching segments between the grandson and each grandparent are colored differently along each chromosome. Transitions between segments matching one grandparent into segments matching the other grandparent result from exchange within chromosomal pairs in the mother. Experimental measures of the rate of exchange in the human genome reveal that on average 36 exchanges occur between the pairs of autosomes, and in females 1.9 exchanges occur on the X. In this example, 44 exchanges between the autosomes and 2 between the X chromosomes are evident. Comparison of the grandson with a single maternal grandparent and a single paternal grandparent would be sufficient to infer the grandparent of origin for every chromosome region.
Detection of identical segments of matching SNPs occurs using comparisons of genotype data between individuals. Here, the raw data for SNPs on Chromosome 1 are compared between a grandchild and a set of grandparents. Because at each SNP the grandchild inherited only one of the four copies of this chromosome region present in this set of grandparents, large segments where the grandchild always has at least one variant in common with one of the grandparents can be identified. At some of the SNPs in this region, the other grandparent and the grandchild are mismatched with no common variant. Focusing on the extreme SNPs in the large segment of half-identity between the grandchild and grandparent 2, this comparison reveals a 45.8 million basepair (Mbp) segment of identity on Chromosome 1. Between positions 46722389 and 48483197, an exchange between the chromosomes of these grandparents occurred in their child producing the new combination of variants observed in their grandchild.
The Human Genome Project, completed in 2003, produced a reference sequence for each of the 22 autosomes and the X and Y sex chromosomes. The combined sequences of all of these chromosomes encompasses 3 billion pairs of bases (i.e., 3 Gb of DNA sequence). The reference base positions along each chromosome are given for one strand of the DNA molecule, which would be paired with a complementary strand, starting with position 1 at one end of the chromosome and sequentially increasing with each adjacent base. Each SNP has a known position in this reference sequence, and the interval between two SNPs along a chromosome represents the physical length of the DNA molecule measured in basepairs.

An alternative measure of distance between SNPs along a chromosome is in comparison to the reference linkage map of the human genome produced from measurements of exchange between chromosomes. Experimental measures of exchange between reference points along a chromosome have been used to determine this genetic size of chromosomes measured in centimorgans. Two SNPs located on a chromosome pair so that exactly 1 exchange is expected each generation are separated by 100 cM or 1 Morgan, a measurement named after Thomas Hunt Morgan - the fly geneticist that first described the phenomenon of linkage along chromosomes. Each SNP along a chromosome has a position on the linkage map. Genetic size is particularly important to the use of SNP data to estimate relationships, because the exchange between chromosomes each generation is the mechanism that breaks up segments of identity resulting from common ancestry.
Focusing on a single pair of chromosomes, this diagram illustrates the reduction in the length of a chromosome region inherited from a single ancestor. Exchange between pairs of chromosomes, which can occur each generation, reduces the size of regions inherited from ancestors. The 1st cousins in this diagram each inherited regions of this chromosome from their common grandfather. The brackets encompass an identical segment shared between the cousins in the overlapping region.

Exchange between chromosome pairs gradually reduces the size of regions inherited from ancestors; however, segregation can result in the loss of identity with ancestors. At the 2nd cousin level of this diagram, the grandfather’s chromosome region is transmitted in the lineage on the left, whereas it is lost in the lineage on the right. Segregation produces a 50:50 chance of retention and loss of ancestral chromosome segments each generation.

These processes occur independently across each of the 22 pairs of autosomes. Considering all 22 pairs of autosomes and the amount of exchange within these pairs, full 1st cousins with two common grandparents are expected to share 41 identical segments, on average. The expected number of identical segments shared between full 2nd cousins drops considerably to only 15.
DNA relatives discovered or validated through SNP comparisons are not commonly a biological ancestor, but rather relatives descended from a common ancestor, a pair of ancestors, or several ancestors. The same principle of relatedness decreasing by $\frac{1}{2}$ each generation also applies to these relatives considering every generation in each line back to each common ancestor. This example shows the identical chromosome segments present in two descendants of a single common grandfather. Ten segments of identity spread across ten different chromosomes and totaling 0.33 billion basepairs (Gb) of DNA are identified by light blue regions on the chromosome map. Considering that the principle of segregation produces a 50% relatedness between each parent and child, the relatedness between these half-first cousins is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$. This corresponds to an expected genome-wide identity of 6.25%, which is similar to the 6.04% identity observed.

In cases where more than a single ancestor is in common, identity through each common ancestor is determined and the sum of these individual identities gives the overall expectation. For example, the identity between full first cousins is $\frac{1}{16} + \frac{1}{16} = \frac{1}{8}$, or 12.5% genome-wide identity.
Based on segregation, independent assortment, and exchange between chromosomes each generation, the numbers and lengths of shared identical segments can be predicted for a given relationship. Segment sizes are predicted in the context of genetic size measured in cM. Two individuals with a shared common grandparent are expected to share 20.7 segments with an average size of 21.7 cM for a combined overall size of 448.8 cM. This comparison shows 10 segments with a combined length of 450 cM shared between half 1st cousins. Each of these 10 segments exceeds the threshold size set by 23andMe for identifying an identical segment. Other segments smaller than this threshold may also be present.
Mechanisms of chromosomal inheritance produce decreasing levels of segmental identity between DNA sequences of more distantly related individuals. In relatives that share a pair of common ancestors, overall DNA sequence identity due to common ancestry decreases by ¼ with each generation, considering that full siblings share 50% genome identity. 23andMe reports this percent identity along with estimates of relationship for DNA relatives. Genome-level identity reflects the number and length of matching chromosomal segments. Combined genetic length of identical segments, and the number of identical segments shared, provide data for estimating relatedness and are commonly reported along with a predicted relationship. The number of shared identical segments shows the steepest decline with distance of close relationship, whereas the average length of these shared segments declines more gradually. It also becomes increasingly likely that 4th cousins and beyond have no matching segments of identity inherited from their common ancestors. This limits the effectiveness of genome-wide SNP comparisons to finding relatives connected within about 5 generations (4th cousins and closer).