

# ANCESTRYBYDNA USER MANUAL

## TABLE OF CONTENTS

**[HOW THE TEST WORKS](#)**

**[STATEMENT ON RACE](#)**

**[UNDERSTANDING BIOGEOGRAPHICAL ANCESTRY](#)**

**[WHAT THE RESULTS MEAN](#)**

**[YOUR GENOTYPES AND THE GENOTYPES.DOC FILE](#)**

**[INTERPRETING YOUR RESULTS-INTRODUCTION](#)**

**[YOUR TRIANGLE PLOTS AND THE TRIANGLE.DOC FILE](#)**

**[YOUR BAR GRAPH AND THE BARGRAPH.DOC FILE](#)**

**[SIMPLE VERSUS COMPLEX ADMIXTURE](#)**

**[YOUR DATA TABLE AND THE RESULTS.DOC FILE](#)**

**[YOUR SEQUENCES AND THE SEQUENCES.DOC FILE](#)**

**[THE SCIENTIFIC FOUNDATION AND VALIDATION OF THE TEST](#)**

**[TEST ACCURACY](#)**

**[PERCENTAGES AND PHYSICAL APPEARANCE:](#)**

**[BETTER UNDERSTANDING HUMAN GENETICS](#)**

**[GLOSSARY](#)**

What follows is the User Manual for the genomics test you have purchased. To fully understand your results, we recommend that you read this user manual from the beginning to the end. **To view your data files you will need the Adobe Acrobat Reader 5.0 Program available free of charge at <http://www.adobe.com/products/acrobat/readstep2.html>.**

## **HOW THE TEST WORKS:**

To determine your ancestry, we have extracted DNA from your buccal sample (mouth swab) the *ANCESTRYbyDNA* test to determine the sequence of your DNA at a large number of different positions. The buccal sample you returned to us contained thousands of cells, and each of your cells contains your DNA. Though we are all 99.9% identical at the level of our DNA sequence, there are certain regions of each chromosome that are different from person to person. These regions are called genetic markers or Single Nucleotide Polymorphisms (SNPs), and it so happens that some small fraction of these SNPs are also different amongst the world's continental population groups. These types of markers are best termed Ancestry Informative Markers (AIMs) and they constitute less than 5% of our genetic material, which is related to the very recent common origin of our species. To help you better understand why this is the case, we recommend the book: *The Great Human Diasporas; The History of Diversity and Evolution*, by Cavalli-Sforza. The positions in your DNA we have sequenced are those we have discovered that are different in this way, and they are spread across all of your chromosomes. That is why it is called a "genomics" test – it is a survey of all of your genetic material, which is known as your "genome". In other words, we have sequenced markers from chromosome 1, 2, 3 ... 22,. Until *ANCESTRYbyDNA*, genetics tests for genealogy or personal interest have been restricted to just one chromosome, like Y-chromosome or the mitochondrial DNA. As such, these other tests offer information that is very different from *ANCESTRYbyDNA*. It is not that this information is incomplete or defective, it's simply different information and it is useful for other purposes.

Your DNA was derived from your mother and father, and theirs was derived from their mothers and fathers, and theirs from their parents and so on. You have 23 pairs of chromosomes. Your maternal copy of chromosome 1 could have been passed through your mother from your maternal grandmother OR your maternal grandfather, but which one you received was randomly determined at conception (you could not have received both). Most of the time this chromosome 1 copy that you receive from your mother is actually a chimeric chromosome that includes parts from your grandfather and your grandmother. Recombination is the process by which these chimeric chromosomes are created and occurs at least once on each chromosome every time a new sperm or egg cell is made. As such, although blending does not occur at the level of the gene (the unit of trait expression) our chromosomes are mixed together and so our genomes contain segments of DNA from all of our ancestors. In contrast, the mitochondrial DNA or Y-chromosome test can only provide data on one single lineage of ancestors each generation into the past. For example, 10 generations ago (year 1802 at 20 years per generation), a baby born today has 1024 ancestors. By measuring your ancestral proportions using a genomics method, we are actually measuring the average population affiliations of all of these 1024 ancestors. Since random processes (recombination and independent assortment) at inception determines the mixings and pairings you harbor, two offspring from a set of parents may have different sets of chromosome pairs, and therefore different ancestral proportions even though they were the product of the same male-female union. For example, an employee of our company is a European male who married a Hispanic woman and had three children of mixed descent. Each of the children exhibits their own unique proportionality, which you can see by clicking on <http://www.ancestrybydna.com/casestudy.pdf>. If these two had an infinite number of children, the average would correspond to that proportionality exactly between the mother and father, but each child would deviate from this average by a unique and random amount.

## **STATEMENT ON RACE:**

Race is a defining issue of modern times in the US, Europe, and many other parts of the world. The impact of the European colonial period that started more than 500 years ago has set the tone for the interactions among diverse populations of the world. Colonization, Genocide, Slavery, Legalized Segregation, Apartheid, Jim Crow Laws and Concentration Camps are but a few of the atrocities that are the history of our civilized world and every culture has its own list to be ashamed of. Given the enormity of these events, their long-term consequences will take generations to overcome. Modern conceptions of Race, Racism, and Racialization are some of the fallout of these events.

Part of our mission at DNAPrint is to work towards the abolition of these misconceptions and the social injustice that result from Racism and Racialization. In this light, we are dedicating considerable internal resources to education regarding the different perspectives (Sociocultural, Political, and Biological) on race and the meaning of populations in light of genomic science and biomedical research.

- 1) Race is not a biological concept. There is not enough genetic differentiation among human populations to consider them zoological races.
- 2) Race is a social construct. This means that these classifications (black, white, Hispanic, Jewish) are defined (and redefined) by the prevailing sociopolitical structure.
  - a. Race is often a great amalgamation of many diverse populations and ethnicities.
  - b. Race is often ascribed only to the minority populations.
  - c. In the US, any minority population ancestry is dominant and the person is completely of the minority group (e.g. “the one-drop rule”).
- 3) Despite the veracity of points one and two above, since there is a correspondence amongst broad racial categories and populations, the conclusion that there are no average biological differences amongst any racially described groups may not be true.
- 4) Racism continues. "In some places, and for some people, overt racism has given way to implicit racialization and “Colorblind Racism” a term coined by Dr. Eduardo Bonilla-Silva (Stanford University)."
- 5) Race should not be used as a surrogate for population. Doing so may lead to over generalization and unfounded stereotypes. A population is the unit of evolution and refers to a group of persons who generally select mates from within the group.
- 6) Being respectful is the first important step in not having a racialized perspective.
  - a. Each person is a human being first and foremost. It is disrespectful, at any level, on the street, in the lab, or in the clinic, to consider his or her population group first.
  - b. Populations should be described (not defined) in precise language that members of the community would use.
- 7) We believe that the physical and cultural diversity of the world’s peoples should be embraced as a valuable and even sacred resource. Indeed, the genomic variation both within and amongst populations is in many ways our Human Biodiversity and will provide important clues as to the origins, our physiological construction, and the possible futures of our fragile species.

## **UNDERSTANDING BIOGEOGRAPHICAL ANCESTRY:**

The test provides a research grade estimation of a person's BioGeographical Ancestry. BioGeographical Ancestry (BGA) is a means of expressing the proportional ancestry of a person that is devoid of the ethnic labels and the dichotomous grouping of persons into racial categories. There are important uses of this in epidemiological and complex diseases mapping research and in forensic science. BGA estimates provide a description of a person in terms of ancestral proportions that are based on the evolutionary and geographical history of our species. Our recommended book, by one of the leaders in the field of Evolutionary Anthropology, Dr. Luca Cavalli-Sforza (Stanford University), details a broadly accepted model of human evolution. It is within this scientific framework of human origins that the BGA estimation can be understood as a description of a person's placement on a *Multi-Dimensional Continuum of Ancestry*<sup>™</sup>.

## **WHAT THE RESULTS MEAN:**

Human beings migrated out of central, sub-Saharan Africa some 200,000 years ago to inhabit various regions of our globe. These migrants established founder groups that gave rise to present-day Europeans, Native Americans, Africans, and East Asians. A map of these human migration patterns can be found in the map.gif file on your CDROM. If your heritage has been derived from more than one of these groups the test results tell you what your mixture ratios are. If you do not have recent admixture, the test identifies which groups you are part of, and confirms that there is no evidence of recent admixture. It so happens that many people from places such as Nigeria, Ireland and Japan are of relatively unmixed heritage (African, Indo-European and East Asian, respectively – see our website at [www.ancestrybydna.com](http://www.ancestrybydna.com)), but many people from places such as the United States, South East Asia or Latin America are admixed. For example, Hispanics from Mexico or elsewhere in Central or South America were derived from the colonial mixture of Europeans, Native Americans, with some proportion of West African. Native Americans inhabited North and South America from Alaska to Patagonia. If your great grandfather was a great Aztec warrior, and of unmixed heritage, you will exhibit at least 12% Native American ancestry on average. If your great-great grandmother was a sanguine Chinese philosopher of unmixed heritage, you would be of at least 6% East Asian heritage on average. Many people from Puerto Rico are heavily admixed – generally showing significant Native American, African and Indo-European mixture. It is important to point out that these results do not give you any information other than your ancestral proportions. You should not use your results to make inferences about your predisposition to respond to a particular drug, or develop a particular disease.

It is notable that some regions of the world have more complicated histories than other regions making the concept of ancestry more complicated and even tedious. For much of our history as a species, we were more mobile than today. The advent of agriculture, in at least four separate global regions about 10,000 years ago changed this for many people, but did not stop the process of migration. Indeed, the largest migrations in human history started only 500 years ago with the European colonial period, the trade in enslaved West Africans, and the colonization of the New World. However, prior to this time and for millennia people have moved about and particular regions of the world show traces of these migrations back and forth into and out of continental and sub-continental regions. Some examples of such regions are East Africa, North Africa, Central Asia, South Asia, and Insular Southeast Asia. Although these populations are distinct groups today with languages, cuisines and cultures that identify them as such, their genetic makeup reflects the long-term history of migrations from more than one region.

**YOUR GENOTYPES AND THE GENOTYPES.DOC FILE:**

Your sequences are provided to you in the genotypes.doc file. Each of the sequences is comprised of two letters. There are two sequences because you received one chromosome from your mother and another from your father. DNA is comprised of nucleotide bases and there are four different types: Guanine (G), Adenine (A), Thymidine (T) and Cytosine (C). Most DNA sites do not differ amongst individuals, but the sites we measure are variable in sequence from person to person, as well as from one population to the next. It is these positions that we measure order to estimate your ancestry. Each site has only two possible letters – they are called bi-allelic sites for this reason. Whereas one person may have two copies of a “G” at one site (represented as “G/G”), another may have a G and a C at this same site (represented as “G/C”). The possible letters for one site may be G or C, but for another they may be C or T, or A or G. Your string of letter pairs across all sites measured are most likely quite unique to you, and in a way they could represent a sort of a genetic bar code for your identification. However, it is the information they give us about your ancestral heritage that is of value here.

It so happens that these letters change slowly over time. When our ancestors migrated to establish the founder groups that gave rise to today’s continental population groups, the nature of their sequences at these sites gradually drifted towards one or the other letter. By measuring the frequency of each letter in each of the descendent ancestral groups of these splinter groups, we can construct a sort of genetic map distinguishing each. We have sequenced your DNA at these sites and made an estimate of the proportional extent to which you share identity with each of four major population groups, sub-Saharan African, Indo-European, East Asian, and Native American.

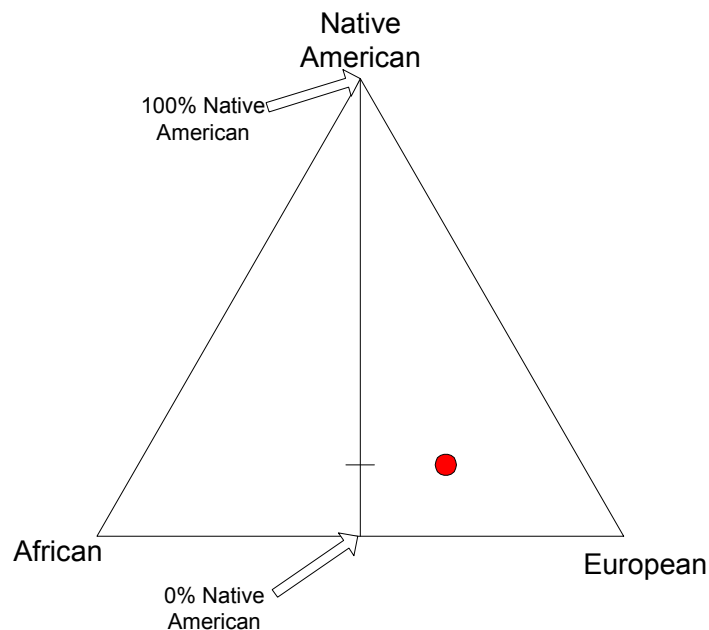
## **INTERPRETING YOUR RESULTS – INTRODUCTION:**

An important concept to understand when interpreting your data is that even though we can always derive a maximum likelihood estimate (MLE) of your ancestry from your DNA, your answer is an estimate of percentages that are similarly likely, not one specific set of percentages. The MLE is the most likely ancestry mix. Other percentages are possible, though less likely, and these are shown in terms of confidence contours on your triangle plot and in your bar graph. This is because with a genetics test, ancestry can only be estimated in a statistical sense, much like the track of a hurricane. Anyone who lives in the South East US knows that it is impossible to project the track of a hurricane exactly, but that the “cones” of most likely migration are extremely accurate. The same is true with *ANCESTRYbyDNA* 2.0 and 2.5.

Your results are expressed in two separate graphical representations. The first we will discuss is the Composite Triangle Plot and the second is the Bar Graph.

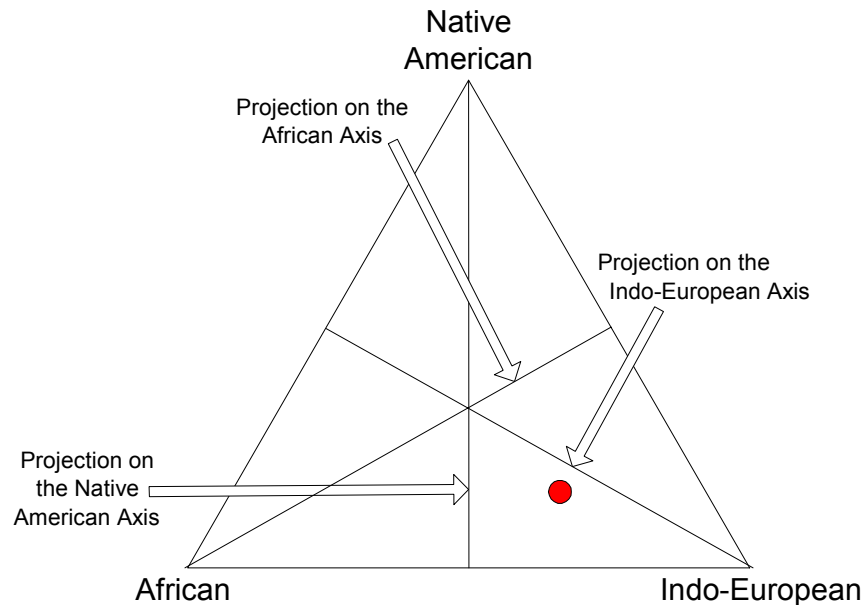
## **YOUR TRIANGLE PLOT AND THE TRIANGLE.DOC FILE:**

We have provided you with a graphical “shapshot” of your results called a composite triangle plot. The red point on this plot is called a Maximum Likelihood Estimate (MLE), and only one is present on the plot. This MLE represents the best estimate of your ancestral proportions. To read a triangle plot, you drop a perpendicular line from the vertex (triangle point) of each triangle to the triangle edge below it.



In the example above, the red dot is the MLE. We have dropped a line from the Native American vertex to the line below, and this particular line serves as a scale for the percentage of Native American ancestry

– from 0% at the base to 100% at the vertex (or tip). Confusing In this example, the individual is about 15% Native American (indicated by the hash mark).



In the example above, we have created the scale line for each of the other two vertices. As with the first figure, for each line the point or vertex represents 100%, and the base near the line represents 0%. The arrows show where the circle projects on to each of these lines. In addition to being able to see that the person is of 15% Native American ancestry in the plot above, we can see that the person is of 60% Indo-European Ancestry and 25% African ancestry as well. You will notice that the three percentages must add up to 100%.

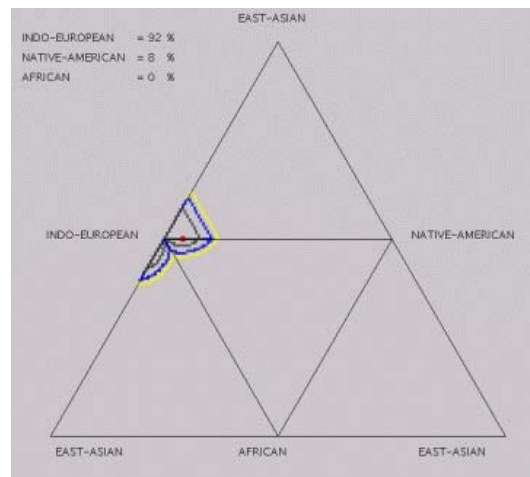
The MLE we have plotted is a statistical estimate, and all statistical estimates have confidence intervals. The MLE tells you what the best estimates of the proportions are, but in reality, there is a small chance that they are something else. The confidence intervals tell you what other values are also likely. In order to know your proportions with 100% confidence, we would have to perform the test for each region of the variable genome, which would make the test very expensive. Since we have not, your results are statistical estimates. We calculate and plot for you all of the estimates that are “2 fold, 5 fold and 10 fold less likely” than the MLE. The way it works is that the MLE is the most likely estimate. Any point within the first contour (inner most ring) is up to “2 fold less likely” than the MLE. the farther from this point, the closer it is to “2 fold less likely”. Any point outside the first and within the second is from “2-5 fold less likely” and any point within the last contour, but outside the second contour is from “5-10 fold less likely”. Any point outside the last contour is AT LEAST “10 fold less likely” and the farther from the MLE, the greater the increase over “10 fold”.

The greater the number of DNA positions we read, the closer these contour lines come to the MLE point. On the triangle plot, the likelihood (probability) that your true value is represented by a different point other than the MLE decreases as you approach the red dot, where the probability is at its maximum (hence, it is called the Maximum Likelihood Estimate or MLE). We could perform the test so that the contour lines are very close to the MLE, however this would require us to sequence a much larger

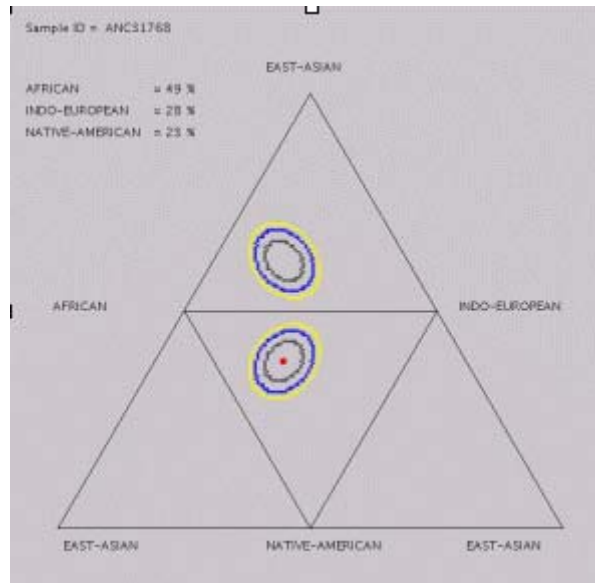
collection of markers. To keep the test affordable, we limit the survey a reasonable number of markers that are sufficient for you to know with good confidence what your proportions are. The yellow circle (10 fold contour) is also referred to as the “one-log interval” and is generally taken as a scientific level of confidence.

The composite triangle plot is read in the same manner as the equilateral triangle previously reported. The composite triangle holds 4 smaller equilateral triangles or sub-triangles, and the percentages for any point in each of these triangles is read as described above. You have all possible 3-population triangles and you can see the extension of the confidence intervals in all possible dimensions. Each of the four equilateral triangles within the composite triangle is called a sub-triangle and the percentage mixture corresponding to any point within the sub-triangle is determined using the same scaling method described in above for a single triangle plot. The most likely result is plotted in the center sub-triangle of the composite triangle. Any point within the composite triangle represents a mixture percentage for three groups at a time, but the reason for expressing the data in this new triangle is that all possible groups of three can be presented on a 2-dimensional sheet of paper. The MLE is found at the center of the 2-fold confidence interval demarked by the black line. Points within each sub-triangle that fall within the black line represent BGA proportions that are from 1.1 to 2 times less likely to be the true value than the MLE (the greater the distance from the MLE the greater the reduced likelihood). Points within each sub-triangle that fall within the black and blue lines represent BGA proportions that are from 2.1 to 5 times less likely to be the true value than the MLE (the greater the distance from the MLE the greater the reduced likelihood). Points within each sub-triangle that fall within the blue and yellow lines represent BGA proportions that are up to 10 times less likely to be the true value than the MLE (the greater the distance from the MLE the greater the reduced likelihood).

In the example plot shown here, the individual is 92% European and 8% Native American. For this individual, there is a noticeable indentation in the confidence contours away from the African axis, showing that there is relatively little chance that the individual has African ancestry. The confidence contours in the Native American direction are broader, and indeed this individual has Native American ancestry. The test is “saying” that there is a chance this person has 16% Native American ancestry, or even 0% Native American ancestry, rather than 8% but that its much less likely (at least 10 times less likely) than the 8% Native American answer. There is also fair spread in the East Asian direction for this individual, indicating that it is possible, though less likely, that this person has some level of East Asian heritage.

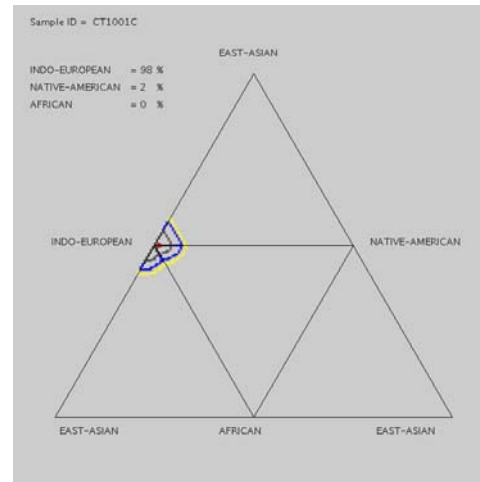
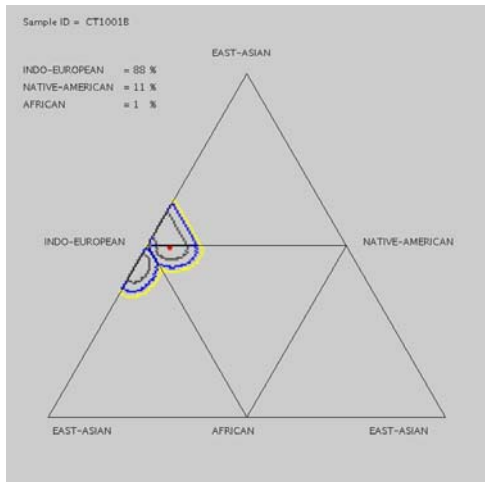


Print this plot out and fold along the lines such that you connect all three vertices of the composite triangle to make a pyramid. You will notice that the confidence contour space is now symmetrical. In fact, this is always the case. The confidence contour lines are on the surface of the pyramid but they connect through the center of the pyramid to form a 3-dimensional shape, which has symmetry, and you can imagine this connection by looking at your pyramid. Points in the interior of the pyramid are points with 4-population admixture. Points on the surface of the pyramid are points with 3-population admixture. How likely it is that a 3-population or a 4-population point represents the true admixture proportions depends on whether the point falls within the yellow confidence contour, blue confidence contour or black confidence contour, and whether the point is on the surface (3-population) or interior of the pyramid (4-population). In this way, you can imagine the likelihood of a variety of 4-population mixture results. In fact, people who are most likely to be 4-population mixture would have discontinuous regions on the 2-dimensional representation of the pyramid. An example of such a person is shown below:



However, you will notice that when the triangles are folded and the intervals on the two sides are connected through the center of the resulting pyramid (in your mind, looking at the pyramid), the three dimensional shape created by the contour lines has a large amount of volume. In fact, for people of 2-population or 3-population admixture, the ratio of the surface area to volume of the 3-dimensional space is relatively high compared to people of 4-population admixture, where there is considerably more volume to the shape.

Other examples of triangle plots: (each of these individuals is mainly European)



## YOUR BAR GRAPH AND THE BARGRAPH.DOC FILE:

The data is essentially the same as that shown in your triangle plot, it is just shown in a different way. To make the bar graph we plot the MLE values including the values within the 2-fold confidence range, one group at a time.

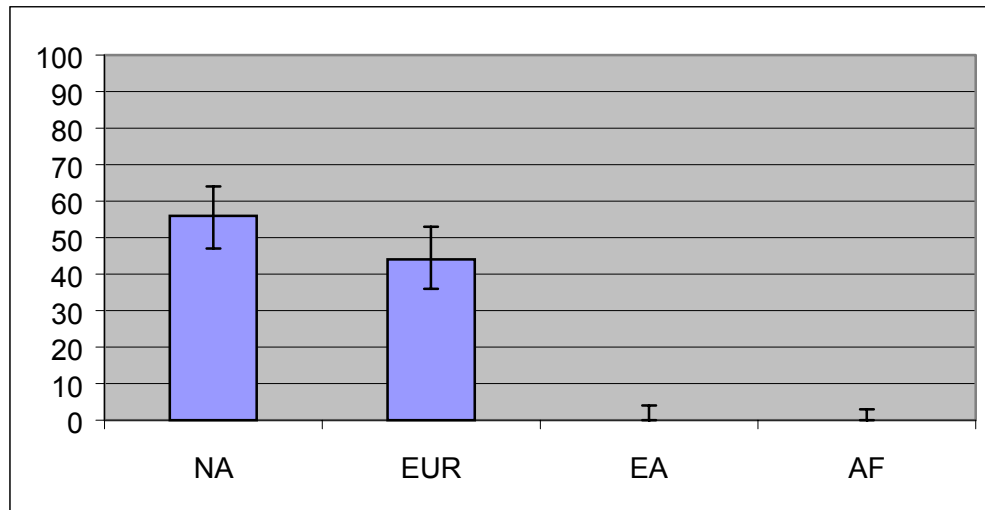
For people of simple or polarized admixture, the bar graph is a useful presentation because it provides a separate, objective view of your possible ancestry percentages for any one particular group – one group at a time. Customers seeking to confirm a great- great-grandparent of Native American ancestry who have obtained an MLE showing 100% European find the bar graph useful in understanding the statistical meaning of their results and how it may still be possible (though less likely) that there is a small amount of Native American ancestry.

For people of more complex admixture, (i.e. 4-population admixture, such as 30% European, 30% African, 20% Native American and 20% East Asian, which might be obtained from a person with a Dominican father and a Philippine mother), the MLE is computed using a more complex 4-D methodology. Because the triangle plot projects the results in terms of most likely 3-population mixture results, and because individuals of 4-way admixture are obviously more complex, on a group-by-group basis, the bar graph is more informative for the latter.

How is the bar graph generated? The bar graph is made by taking the MLE values from the 3-D triangle plot method (for individuals of polarized ancestry) or from the 4-D algorithm (for individuals of more complex, or even admixture). The confidence ranges are established for each ancestry group by searching the likelihood calculation of all 3-population or 4-population admixture combinations and finding the highest and lowest values for each group that fall within the 2-fold likelihood range (0.3 Log base 10) of the MLE.

Example: The bar graph below shows the results for a person of 2-population admixture. This person's most likely ancestry mix (MLE) is 55% Native American, 45% European. There are confidence ranges extending above and below the tip of each bar. These ranges show the other percentages this person could be, though any percentage is up to 2-times less likely than the percentage indicated by the blue bar.

Though this person is most likely to be 55% Native American, 45% European, they could also possibly be 65% Native American, 35% European though it is 2-times less likely than the MLE value of 55% Native American, 45% European. You will notice that there are confidence bars for the East Asian (EA) and African (AF) groups too. Thus, it is possible that this person is 54% Native American, 44% European, 1% East Asian and 1% Western sub-Saharan African, though again, it is significantly less likely than the most likely estimate of 55% Native American, 45% European.



People of 1-population or 2-population admixture would most likely show one or two blue bars, respectively. People with 3-population or 4-population mixture would most likely show 3 or 4 bars, respectively. A person that shows 3 bars may still be a 4-way mix if there is a confidence range for that fourth group, its simply less likely (2 times less likely) that they are a 4-population mix incorporating a value for the fourth group within the range than the indicated 3-population mix.

## **SIMPLE VERSUS COMPLEX ADMIXTURE:**

The composite triangle plot shows your Ancestry mix. To make the triangle plot we assume that your pedigree since the time the major human populations diverged from one another is no more complex than a 3-population mixture of ancestries. The central triangle shows the most likely 3-population mix (each vertex or point of the triangle is a group, and a triangle has 3 such vertices/points). Even if your most likely ancestry mix is a function of only 2 groups (i.e. your MLE falls on an edge of the triangle), for most people, there will be confidence contours extending in the 3<sup>rd</sup> dimension (toward the 3<sup>rd</sup> vertex) in this triangle, and the distance these intervals will extend towards this vertex is greater than the distance they would extend toward the vertex for the 4<sup>th</sup> group (this should be visible in one of the three other triangles around the central triangle). In other words, your DNA sequences optimally fit with one particular 3-population ancestry combination. The particular point/MLE is the most likely outcome given your DNA sequences and is indicated by the red point that is located within the 10-fold, 5-fold and 2-fold confidence contours.

The bar graph will communicate approximately the same ratios as presented in your composite triangle plot. However, the bar graph provides you a completely different view of your results. For the bar graph, we calculate the likelihood of ancestry affiliation using the triangle plot algorithm and plot the values from your most likely triangle. If the value for another triangle is within a log base 10 (within 10-fold) of the MLE from the best triangle, the level of the 4<sup>th</sup> group is also plotted in the bar graph. The data is basically the same as that presented in the triangle plot. Your most likely estimate, or “answer”, is still the set of percentages provided in the results table, and these percentages should be about the same as those provided in the bar graph, but the view provided by the bar graph is different. The reason we provide you with the bar graph is that it may be useful to some customers because:

- 1) Looking at ancestry percentages on a bar graph is easier to understand for some customers and,
- 2) The representation of ancestry levels within one log of likelihood in terms of bars illustrates the equivalency of readings that are similar in likelihood and that the answer is really not only one simple set of percentages but a range of likely percentages with one that happens to be the most likely. This is a concept many of our customers have had trouble with in the past.
- 3) Showing the fourth group if its likelihood score is within 10-fold of the third allows for greater sensitivity in detecting affiliation for some groups in some customers.

For example, a person who believes their great-great-grandfather was an American Indian but who obtained an MLE result on their triangle plot of 100% European with confidence contours that extend towards the Native American (and/or East Asian) axes only narrowly may find that in the bar graph the 2-fold confidence range is more substantial. This may lend credence to their hypothesis.

Each point in the triangle plot is a specific and unique 3-population ancestry mix.

What if your ancestry is more complex – that is, individuals from all four ancestry groups have contributed to your pedigree since the time the major human populations diverged from one another? In this case, the MLE from the triangle plot will be an oversimplification of your ancestry. Instead we calculate your MLE using a more complex 4-way algorithm, but we still provide you the triangle plot that carries information that is diagnostic of 4-population admixture. For the individuals of 4-population admixture the following apply:

- 1) If your heritage is best explained as a function of 4 groups you will most likely see discontinuous (separated) confidence contour regions on your composite triangle plot. In other words, you will see contours in more than one of the 4 sub-triangles. After you print out your plot and fold it up into a 3-D pyramid, you will see that the confidence regions for the different triangles are now shapes on faces of a pyramid that can be united within the pyramid to form a 3-dimensional shape or space. This space would be rather large compared to that for individuals of homogeneous ancestry (their spaces are normally confined to the tips of the pyramid, and the surface area to volume ratio is higher for their spaces than yours). Points within the 3-D space (within the pyramid) are likelihood estimates for 4-dimensional admixture. Your MLE is determined using the 4-dimensional algorithm, and the percentage values should roughly correspond to the center of this shape. Whether any other point within this space is 10-fold, 5-fold or 2-fold less likely than the MLE at the center is determined by which of the 3 confidence contours they fall within (the black contour which is the 2-fold likelihood space, the blue shape which is the 5-fold likelihood space or the yellow shape which is the 10-fold likelihood space).
- 2) If your heritage is best explained as a function of 4 groups you would see four blue bars on your bar graph. Individuals of homogeneous ancestry, or 2 or 3-population admixture will always have confidence ranges (the lines above and below the tip of the blue bar) for each of the 4 groups, though they may be very small, but they never show 4 blue bars.
- 3) If your heritage is best explained as a function of 4 groups, it will have been determined using the 4-D algorithm and the bar graph will exhibit 4 blue bars with corresponding confidence ranges. This is your most likely estimate (MLE) of ancestry admixture.

If a 4-D algorithm works best for people of complex admixture then why don't we simply calculate the 4-population admixture percentages for each customer whether they are of homogeneous, 2-population, 3-population or 4-population admixture? Mainly because the 4-D algorithm is not as robust for individuals with 3-population or simpler admixture; for these individuals the simultaneous consideration of 4 groups produces too much statistical noise that would impact the quality of the results by a couple of percent. We have concluded from internal research and development that when the number of classification groups increases from 3 to 4, a larger number of genetic "markers" is needed to provide results of comparable confidence as evidenced by the comparison of *ANCESTRYbyDNA 2.5* (a 175 marker test) from the previous version of the test, *ANCESTRYbyDNA 2.0* (a 71 marker test).

**YOUR DATA TABLE AND THE RESULTS.DOC FILE:**

The composite triangle and bar graph are tools for you to visualize your data. The results of the MLE are reported in the data table as in this example:

**SAMPLE-ID = XXXXXXXX**

ESTIMATE	ANCESTRY
90%	Indo-European
10%	Native American
0%	Sub-Saharan African
0%	East Asian

The MLE percentages of Indo-European , Sub-Saharan African, East Asian, and Native American are shown. In the example above, the person was determined to be 90% Indo-European, 10% Native American, and the MLE is plotted at the position of the triangle plot indicated by these proportions. The bar graph would show solid blue bars at 90% for Indo-European and 10% Native American. Unlike the table however, the triangle plot and bar graph show the confidence regions around the MLE, which is, as we have mentioned, very important for properly interpreting your results.

The ancestral proportions we report to you are a function of probability, and the MLE is the best estimate. Going back to the triangle, the further from the point estimate on the triangle, the less likely that point represents your true values. The farther from the red dot (the MLE) in any direction, the lower the probability. Therefore, your plot is an estimate rather than a precise fixed number, but it is the most likely estimate based on the data produced from many regions of your genome.

Why is it that it is not possible to determine what your proportions are *exactly*? When drawing conclusions from DNA one must use statistics. To do this without using statistics, we would have to go back in time 200,000 years and keep track of every one of your ancestors since the origin of our species in Africa. Clearly, this is impossible. Since you inherited your DNA from your ancestors, your ancestral proportions are written in your DNA, but this information must be statistically inferred from the DNA sequence. If we measured 1,000 or 10,000 genetic sites in your DNA as opposed to the 200 or so we measure today, your confidence intervals would be smaller but the MLE would probably be very similar (in the triangle plot, the confidence ring would shrink around the MLE point). The costs of measuring more genetic sites can quickly become out of hand making the price to the consumer unrealistic. Thus, in conclusion, the result of our test is your MLE. Though the MLE is a statistical estimate, and there is a small chance your true proportions are slightly different from that of the MLE, the MLE is the best estimate. If you want to keep it simple, simply use this MLE when describing your genetic heritage. An alternative to understanding more in terms of your families heritage is to have more persons in your family (parents, grandparents, siblings) take the standard **ANCESTRYbyDNA** test.

## **YOUR SEQUENCES AND THE SEQUENCES.DOC FILE:**

In the sequences.doc file, we have provided you some markers so that you can cut and paste them into various types of gene search engines. By doing this, you can learn about:

- a) The human genome project
- b) How our human DNA sequences are similar to, and differ from those of other organisms.
- c) How cancer research is conducted through genetic research
- d) And other things...

The sequences are just a few of those that we measure in order to determine your ancestral proportions. For any genetic position we measure, the variant sequence is flanked by invariant sequence (remember, 99.9% of our DNA is identical from person to person!). For example, if we measured whether you have a C or a T at position 25 in the following sequence:

**GACCTACCATGATAAATTCCTAGG[C/T]GAGGAAGCTACTACACTGAGTTTAT**

The site we actually sequence is indicated by the [C/T], and all of the other letters represent DNA nucleotides that are invariant. That is to say, they are the same from person to person. Since these invariant nucleotides have no impact on the estimate of your ancestral proportions, there is no reason to measure them. We provide you with this view of your sequences because you can use them to have fun and learn more about genomics at the same time – here is how.

The human genome project has spawned the development of myriad DNA databases, and you use the sequences in a search through these databases. We have all paid for the construction of these databases, and they are yours to use freely. Their main function is as a tool for scientists to better understand why it is that certain people are afflicted with diseases or respond to drugs differently.

Many of the databases can be reached through the website of our partner, Geospiza Corp. at <http://www.geospiza.com/outreach/organelles/index.html>. Geospiza operates this site as a community service, and its main function is to help educate the layperson. In addition to searching with your sequences, you can click on links that will help you understand how the databases were constructed, and what information each offers.

For a basic slide show introduction to genomics and DNA databases please visit:  
<http://www.geospiza.com/outreach/organelles/mutation/slide1.html>

To screen your sequences against the human genome, all you need to do is copy and paste your sequences into the database query forms provided through the <http://www.ncbi.nlm.nih.gov/> site (which can be accessed through the <http://www.geospiza.com/outreach/organelles/index.html> site). Click on the “BLAST” selection on the top menu and you have entered the human genome database search engine. Select “Standard nucleotide-nucleotide BLAST”, and then simply paste your sequence (one sequence per query) into the Search box and click BLAST! A progress page will appear, and you will need to click the “Format!” Button to see your results. Learn more about the region of the chromosome corresponding to this sequence by clicking on the blue links.

## **THE SCIENTIFIC FOUNDATION AND VALIDATION OF THE TEST:**

The science behind the test has been published in the scientific literature (for references, please see our web site at [www.ancestrybydna.com](http://www.ancestrybydna.com)). We have determined the frequency of DNA sequence variants in the various human populations (some of this data can be seen on the website), and by determining your sequence for each; we can determine the probability that you identify with each group. The test has been evaluated using a large number of people from a wide range of ancestral groups, and the estimates correspond well to what is known from anthropological and historical data. For example, Hispanics are known to have arisen as an ethnic group from the blending of colonial Europeans with Native Americans, and the hundreds of Hispanics we have tested align with these two groups almost exclusively, as expected. As another example, though most Nigerians plot as of unmixed African BioGeographical Ancestry (BGA), African Americans plot more as a mixture between this group and Europeans, which is also what would be expected from what we know about the admixture between Africans and Europeans in the US. The method has also been validated through pedigree challenge; when the BGA is determined from a mother and father, that of their children should plot somewhere between the two. To date, we have tested numerous family pedigrees, and the ancestral proportions of offspring always plot somewhere amongst those of their parents. When outside agencies blindly test the MLE estimates, they prove to be excellent estimates of ancestral proportions.

When tested against known pedigrees, the *ANCESTRYbyDNA 2.5* test performs quite well. The data for individual A is presented below. His wife, individual B, is Hispanic and she was determined to be of mostly Native American ancestry but with some Indo-European and African heritage. This was also expected based on what we know from anthropological origin of the Hispanics (which were derived from the union of Spanish explorers, Native Americans, and West Africans in Colonial Caribbean and Latin America). Each of the 3 children are plotted roughly half way amongst both parents, as expected. None of the children exhibit East Asian ancestry. The results of the children were consistent with those of the parents, and the MLE's are accurate estimates when tested against what is known from biographical data.

Experiment: Ancestrybydna 2.5 Blind trials on samples from families.

Purpose: To determine how well the test results agree with expectations formed from appreciation of a family pedigree.

Summary of Results:

Family 1

Individual A	INDO-EUROPEAN	93	EAST-ASIAN	0	NATIVE-AMERICAN	7
Individual B	INDO-EUROPEAN	7	AFRICAN	22	NATIVE-AMERICAN	71
C1	INDO-EUROPEAN	47	AFRICAN	15	NATIVE-AMERICAN	38
C2	INDO-EUROPEAN	60	AFRICAN	2	NATIVE-AMERICAN	38
C3	INDO-EUROPEAN	57	AFRICAN	6	NATIVE-AMERICAN	37
S	INDO-EUROPEAN	86	EAST-ASIAN	0	NATIVE-AMERICAN	14
M	INDO-EUROPEAN	81	EAST-ASIAN	7	NATIVE-AMERICAN	12
F	INDO-EUROPEAN	92	EAST-ASIAN	0	NATIVE-AMERICAN	8

Summary of Results:

Family 1: The father is F and mother is M. Both exhibit some Native American admixture (8% and 12%, respectively). Their children, Individual A and S also both exhibit some Native American admixture (7% and 14%, respectively). Due to the law of independent assortment (the test markers span 22 chromosomes), these results are reasonable. For example, if one of the children measured with 75% Native American, or 20% East Asian, these results would be unreasonable. Individual A married his wife, who is Mexican and they had three children, C1, C2, C3 Each of these three children have 38%, 38% and 37% Native American ancestry, respectively, with the balance Indo-European and African. Again, these results are quite reasonable given the fact that the father was mostly Indo-European with slight Native American admixture and the mother was Hispanic, and mostly Native American with Indo-European and African admixture.

For more information on how MLE's read for individuals of various mixed populations, please see <http://www.ancestrybydna.com/> (product info section). You will find that the results are in good agreement with what is known from the anthropological history of each population and that all of the evidence we have to date suggests that *ANCESTRYbyDNA* test results are highly accurate.

### **TEST ACCURACY:**

The genotypes (nucleotide letters) we have determined for you are quite accurate. Because we use the latest genetic reading equipment available, we routinely achieve a greater than 99% accuracy for each site. If an accurate value was not obtained for you at a particular site, you will see an "FL" instead of your letters for that site. Having a few of these generally does not prevent us from making a good ancestry estimate, but of course having too many would. Some reasons you may have an "FL" for a site include

- a) A small region of your chromosome around this site is missing or is of different sequence character than for most. This result is not uncommon given the highly variable nature of the chromosomal positions we measure, and it certainly does not imply you have any sort of defect in any way whatsoever (in fact, it may be an indication of your uniqueness).

- b) We did not get enough DNA from your swab. Some markers are more sensitive to this than others. If there are too many “FLs” for your read-out, we will not be able to determine your ancestry proportions to a degree of accuracy that we would like, and in this case we will have to asked you to submit another sample for a second try.

Experiment: Repeated estimation from the same samples with *ANCESTRYbyDNA 2.0*(BD).

Purpose: To determine how reproducible the results are by measuring the proportions in the same individuals on different occasions.

Results:

plate3-BD101-Data.INP	INDO-EUROPEAN	100	EAST-ASIAN	0	NATIVE-AMERICAN	0
plate5-BD101-Data.INP	INDO-EUROPEAN	100	EAST-ASIAN	0	NATIVE-AMERICAN	0
plate3-BD304-Data.INP	INDO-EUROPEAN	85	AFRICAN	15	NATIVE-AMERICAN	0
plate5-BD304-Data.INP	INDO-EUROPEAN	86	AFRICAN	14	NATIVE-AMERICAN	0
plate3-BD316-Data.INP	INDO-EUROPEAN	72	AFRICAN	27	EAST-ASIAN	1
plate5-BD316-Data.INP	INDO-EUROPEAN	79	AFRICAN	20	EAST-ASIAN	1
plate3-BD3162-Data.INP	INDO-EUROPEAN	100	EAST-ASIAN	0	NATIVE-AMERICAN	0
plate5-BD3162-Data.INP	INDO-EUROPEAN	89	EAST-ASIAN	5	NATIVE-AMERICAN	6
plate3-BD317-Data.INP	INDO-EUROPEAN	79	EAST-ASIAN	21	NATIVE-AMERICAN	0
plate5-BD317-Data.INP	INDO-EUROPEAN	84	EAST-ASIAN	16	NATIVE-AMERICAN	0

Summary of Results: Variation in percentages is a result of failed markers. This test shows a 5-6% variation for the absolute percentage in any one group. Since this experiment we have been using 5 samples on each run as internal controls. The average variation is 2-3% for these controls. Another group of 11 have also been tested repeatedly, and these show an average 2-3% variation for 10 samples, and an average 5% variation for the other sample. Best estimate from all of the data on repeated measurements is on order of 3-4% variations for most determinations if the individual tested has failed markers. Thus, if your profile came back as 96% Indo-European and 4% East Asian, it is debatable whether the 4% East Asian is significant and would also be addressed by the confidence contours.

**PERCENTAGES AND PHYSICAL APPEARANCE:**

We have noted that individuals exhibiting physical characteristics of a population group generally have at least 30-35% identity with that group. For example, persons with an 85% Indo-European and 15% African generally exhibit few if any physical features characteristic of the African group, such as darker skin. Why is this? The genes that determine physical appearance are but a very small percentage of the total number of genes in the genome. Thus, for all of these genes to have sequences characteristic of one group, the person would need to be of relatively high proportions for that group. The higher the



percentage of African a person is, the more likely the areas of the genome that determines physical appearance will be of African origin.

### **BETTER UNDERSTANDING HUMAN GENETICS:**

First of all, let us admit that genetics is a complex subject. You do not have to be a scientist to understand your results, and if you do not understand your results from thoroughly reading all of the materials supplied to you, you may want to do some basic reading on the internet.

The U.S. Department of Energy Human Genome Program is an excellent place to begin the process of understanding Human genetics. <http://www.ornl.gov/hgmis/publicat/primer2001/index.html>

Other informative sites: <http://www.genome.gov/>  
[http://www.sc.doe.gov/ober/hug\\_top.html](http://www.sc.doe.gov/ober/hug_top.html)  
<http://www.gene.ucl.ac.uk/hugo>  
<http://www.wellcome.ac.uk>  
<http://www.nhgri.nih.gov>  
<http://www.doegenomestolife.org/>  
<http://www.ncbi.nlm.nih.gov/ncicgap>  
<http://www.jgi.doe.gov/>  
<http://www.tigr.org/>

For further reading, we recommend: The Great Human Diasporas; The History of Diversity and Evolution, by Cavalli-Sforza. It is an informative tool in understanding migration patterns and ancestral evolution.

People interested in discussing genealogy from DNA should visit this interactive forum:  
<http://archiver.rootsweb.com/th/index/GENEALOGY-DNA>

## **GLOSSARY:**

**Ancestry Informative Marker (AIM):** AIMS are the subset of genetic markers that are different in allele frequencies across the populations of the world. Most polymorphism is shared among all populations and for most loci the most common allele is the same in each population.

**Allele:** Alternate sequences for a particular position in the genome. For example, a common variation in the genome is for some forms of the sequence to have Cytosine (C) while other forms have Thymidine (T). Thus, since we have two copies of each chromosome, there are three genotypes at this position CC, CT, and TT.

**Chromosome:** The physical units of heredity: long linear strands of DNA. Humans have 22 autosomal chromosome pairs, plus two sex chromosomes, X and Y. Men have two copies of each autosome, 1, 2, ..., 22, X, Y. Women have two copies of each chromosome 1, 2, 3, ..., 22, X, X. Each person thus has a total of 46 chromosomes.

**Genomics:** The study of the complete complement of genetic material in a species.

**Genome:** All of the genetic material in a species. The human genome is approximately 3,300,000,000 base pairs in length.

**Locus (pl. loci):** The name for a physical position on the genome. Can either refer to a large region such as a complete gene or a very specific region, like a particular base pair position.

**Polymorphism:** The property of having more than one state or alternate sequence at a particular position. The alternate states are called alleles.

**Single Nucleotide Polymorphism (SNP; pronounced snip):** A precise base pair position where different people are found to vary in sequence. Generally two alternate alleles are found at a particular SNP. At least 2,000,000 SNPs are now known and there may be over 30,000,000 in the human genome.