FACTS AND OBSERVATIONS IN RELATION TO THE X CHROMOSOME: MISCELLANEOUS MATERIAL RELATING TO ANCESTRY GLEANED FROM THE RESOURCES HERE

http://davidkfaux.org/X_CHROMOSOME_Haploblock_Resources.pdf

by

David K. Faux

Structure of the X

Chromosome X is a sex chromosome (the other is the Y), which, along with mtDNA and the 22 pairs of autosomes make up the genome of an individual. Men have one X (always from the mother; its counterpart being the Y) and women have two (XX), one from each parent.

- While the **male X is haploid (only one)** the **female** is just like the autosomes and so **diploid (she has two)**. Therefore while the nucleotide sequence of the male X comes directly from one of the mother's recombined X chromosomes and will be passed unchanged to his daughters; the female XX will be an arbitrary (alphabetic order) series of A, C, T, G such as AT, CC she will only know which allele came from which parent when a haploblocks sequence of nucleotides is identical such as AA TT CC CC and in this case the haploblock can be inferred directly. For the rest of the series **the data must be "phased"**, parsed into the correct maternal and paternal chromosomes. This can be done by testing relatives, or by using powerful computer programs such as PHASE. Good **programs**, shown to work particularly well in admixed individuals, are SABER, HAPAA, PLINK and ANCESTRYMAP however none are in any way remotely "user friendly" for the "average" genetic genealogist.

- It is about as long as chromosome 7 and so a **relatively large chromosome**.

- The X chromosome is comprised of about **154,000,000 (154 Mb) nucleotide (ACGT) bases**.

- **The X chromosome is composed of** nucleotide base substitutions and other elements which serve as markers for use in ancestry and disease studies. These markers include:

1) **SNPs** (single nucleotide polymorphisms) which are changes / substitutions of 4 nucleotide bases – Adenine, Cytosine, Guanine and Thymine (ACGT). These are changes from the ancestral form (e.g., also found in chips and gorillas) to a derived form such as an original C which becomes a G. Some of these SNPs are AIMs (ancestral informative markers) and occur at frequencies that vary by population.
2) **STRs** (short tandem repeats) where at a particular position on the X there are repeats of for example TAGATAGA for a finite number of times but often varying from person to person (the basis of forensic identification) due to a rate of mutation higher than SNPs. So for example at marker DXS10075 an individual could have allele (variant) repeats of TAGA from 16 to 21 (each X marker having slightly different properties).
3) **LINE** (long interspersed nuclear elements) or SINE (short interspersed nuclear elements). These are further grouped as for example L1 retrotransposons (a LINE variant) which make up about 17% of the human genome. Many of these units tend to be lumped into a category of "junk DNA" which is misleading since they have been implicated in a number of a number of cellular processes. L1 is also composed of **Alu insertions** (a form of "wandering DNA") such as DXS225 which varies by population. The "without Alu" state is the ancestral variety.

- A **haploblock** would contain all of the above elements, which have structural and functional properties and can vary in length from say 3 SNPs to 20 or more Mb (million bytes, bases) of the X.

- A **haplotype** is an individual's inherited version of a haploblock configuration. Some haplotypes will be found in all populations, some unique to certain groups (e.g., southern East Asia), and some perhaps more regional (e.g., northwest France).

- Frequently occurring common **haplotypes can be found in both parents** of a female (so she is homozygous thoughout this block) and is due to shared ancestry. Homozygous haploblocks in females are an excellent way to "find" haploblocks suitable for study. The number of these **shared haplotypes** many be fairly common in a country such as the Netherlands. Typical sizes of these blocks would be in the order of 100 to 200 Kb in size.

- **Haploblocks vary in size**, and to some extent the definition is arbitrary. If a female has matching alleles along a stretch of say 200 Kb of the X, then it is clear that both parents had the same haplotype and that this represents a haploblock with distinct start and stop points.

- **More details** about haploblock characteristics will be included later in this paper.

- There are **regions of high and low SNP density** (or any genetic markers). For example the area around position 135 Mb is high in the number of SNPs but nearby at 145 Mb the number drops precipitously.

- There is a "**p arm**" (for petit or small) and a "**q arm**" on the X separated at a region known as the **centromere** which is a concatenation of proteins and nucleotide bases to which the microtubule spindles are attached during meiosis, and is metacentric (near the middle of the chromosome). The tips of the X are known as **telomeres** and may combine with the pseudo-autosomal part of the Y (if one is present).

- Both arms are made up of heterochromatin and euchromatin. The former relates to the "**G banding**" seen with certain types of staining, and these bands include areas of tightly packed (condensed) **heterochromatin** (one of which is the centromere, another the telomeres). These areas may be relatively inert for transcription and recombination, but include many microsatellite sequences that could have ancestral significance. These areas are also relatively "gene poor" but include "tissue – specific" genes. The rest is **euchromotin** where most of the activity including gene expression and transcription resides. The G (for Giesma) bands are low in CG repeat content (known as isochores of long GC segments of greater than 300 Kb) and poor in functional genes. The other areas are known as **R bands**, relatively gene – rich especially in "housekeeping genes". The specific relationship between nucleotide base sequences and both of these cytogenic bands is unclear. However in terms of **mapping out the X**, they are used as borders to denote **segments, such as Xq21.1**. A new technique, known as **"in silico" staining**, is now being used to better map these bands. These defined locations, which are in a sense structural, are **part of the architecture of the X**, the scaffolding on which the DNA attaches.

- There is a relationship between areas of **high CG content** (e.g., centromere and telomere regions and Xq28) where **high levels of Alu sequences** occur, and low CG regions where long nearly identical LINE sequences are found. Segments of 1 Mb or more will each have signature of characteristic CG contents, and hence **concentration of genes**.

- The **centromere** in the constricted **mid section** of the X includes large strings of repetitive microsatellite sequences that have no role in coding. It may, hoever, provide the matrix upon which events such as daughter chromatids assembling and splitting occur. The number of generations that the centromere may remain intact is unclear, although as we will see the recombination rate is low here and may represent an "Out of Africa" or earlier sequence.

- The **telomeres** are at the **tips of the arms** of the X, and comprise a region where TTAGGG repeats appear designed to keep the area intact.

Genes on the X

- There are **1098 known protein coding genes** on the X chromosome which is about 4% of the total of all known human genes.

- Over **300 diseases have been mapped on the X**. There are many X – linked disorders where males tend to be more adversely affected because they don't have an extra X chromosome as a "back up". Examples include Duschesne Muscular Dystrophy (lack of ability to produce dystrophin), Fragile X syndrome (mental retardation) and haemophelia (blood coagulation disorder). Less dangerous conditions such as red – green color blindness fall into this category, as well as traits such as male pattern baldness.

- An example of an important marker / gene on the X chromsome is **XIC which functions to "silence" one X of every XX in female somatic cells**. Generally with the expression of the X inactive specific transcript (XIST), one X chromosome (maternal or paternal) folds into a Barr body (seen on staining). This is generally thought to be a random process with a 50:50 chance for each X in each cell to be inactivated.

Mutation Rates

- **On average there is one SNP difference per 2,000 bp on the X**. The X has a **reduced mutation rate relative to the autosomes**.

- **The mutation rate is approximately one bp substitution per 100,000 years in the above sequence**. The average rate per nucleotide base for the chromosomes in general is about 10 to the minus 8 per SNP locus (0.000000001). By one estimate this would result in about 2.567 new mutations each generation over the 154 Mb span of the X. **A much higher rate is going to be observed for most STR markers**, something in the order of 0.039 per transmission from generation to generation. Any sharing in regions such as this is much more likely due to sharing going back many thousands if not millions of years.

- Due to the meiotic processes and transmission factors where an X remains intact in a male, **males have higher rates of mutations** (up to 5 times although other studies report 1.7) on the X than do females. This is often interpreted as being consistent with "male-driven evolution", and the primary explanation for the lower mutation rates of the X in general.

- **Mutation rates are also higher in areas of high recombination** (the process of recombination is probably mutagenic).

Inheritance Pattern

- The X chromosome has **a unique inheritance pattern** different from the autosomes, Y chromosome or mtDNA. Considering males (who are less complicated, what a surprise, than females), no paternal ancestors make any contribution to the X chromosome. Hence even if your father's father was Canadian Cree Native American, unless there is someone with Native American heritage on your mother's side within the circular "**Fibonacci series**", there will be **zero** Native American contribution to the X. Even on your mother's side of the family, any male to male transmission (father's father) will create a break in the line at that place (e.g., your mother's mother's father's father's lineage is not represented on your X).

- After 7 generations, of the 128 ancestors, only 21 contribute to the genomic structure of the X. The female at the top of the Fibonacci based fan chart contributes **1/8**, and the mtDNA line ancestor shown at the bottom contributes **1/64**, and those scattered in between contribute between 1/16 and 1/32. Reference to the **fan chart** created by Blaine Bettinger will show how this works for females (basically the principle is the same but there are a greater number of ancestors represented on the two X chromosomes of women).

- The **probable contribution from an ancestor 7 generations back** who is in the 1/16 category would be 6% and if it came through a male in the previous generation it would still be 6% (the male transmitting his X unchanged except for a possible mutation to all his daughters). This is probability only, so there is a chance that more than this amount (e.g., 12%) will be directly attributable to that ancestor on paper. It is also possible that despite what is an "air tight" paper genealogy, the ancestor may contribute zero to the X. The various possibilities can be seen by examining the recombination results of full siblings and other close family members. The present author is attempting to obtain simulation data from 23andme showing what percentage (including confidence intervals) of for example Native American ancestry 5 generations ago is likely to be present with each subsequent generation toward the present.

X Chromosome Functional Aspects and Features Related to Ancestry

- During each meiosis, the **probable cross over (recombination) rate** is about one event on the short arm and one to two events on the long arm of the X (or sometimes one event on each arm irrespective of size). Going back 7 generations there could be anything from 3 recombination events back to the woman at the top of the fan; and 6 such events to the woman at the bottom of the fan chart.

- **Returning to haploblocks and SNPs, some haploblocks are "deserts"** with no SNPs and often equivalent to the chimp and gorilla. These "SNP free" regions are more common in Europeans. These can be 1 Mb in length with zero or only a couple of SNPs. Calculations show that 40% of all polymorphisms (e.g., SNPs) showed no variation in a sample of East Asian X chromosomes; whereas only 10% in a West African sample showed no variation. Other areas have been termed “jungles” since they are sites where there is frequent activity. The difference in recombination rates between these two groups of sites may be in the order of recombination being 1000 times more likely in certain hotspots.

- **Other haploblocks can be 750,000 years old** (the “age of the X” according to some estimates), or **120,000 years ago when modern mankind emerged**, or relate to the **migration from Africa circa 70,000 years ago**, or of **more recent ancestral significance** (generally haploblocks of longer lengths) a few hundred or more recent relationships. Bottlenecks can well make for strange combinations but matching an African at 24 SNPs over 100 Kb does not mean that there is necessarily any recent connection. The likelihood of a true (genealogical) connection grows as a function of block size. If the match is 10 Mb then the link is likely recent. Hence the X includes **multiple “shapshots” of the population history of humankind**.

- There appear to be many haplotypes on the X which serve to form haploblocks and the latter may be for example **1 Kb to 10 Mb with little to no recombination for many generations**.

- There is more among – region, and less within – population, molecular variance on the **X** relative to the autosomes and thus the former **has greater potential to elaborate on the between population structure**.

- **In population genetics** this means that in theory there will be three X chromosomes for every four autosomes (assuming that there are equal numbers of males and females). Hence between any two unrelated individuals the **time to most recent common ancestor (TMRCA) should be ¾ of what is found on the autosomes**; and the frequency of change involving the autosomes should happen at ¾ the rate of the X chromosome.

- **Diversity on the X is lower** than on the autosomes, and so populations should differ more on their X chromosomes than their autosomes. One study found three times as many ancestral informative SNPs on the X than the autosomes.

- **Genetic distances between human populations are larger on the X than the autosomes**.

- It is therefore possible to be 100% European, but match only a few other Europeans, no Asians, but a majority of Africans on a particular haploblock, however the ancestral connection could be immediately subsequent to the “Out of Africa” event or even before.

- **The X chromosome has a recombination rate that is in the order of 1 centriMorgan = 1 Mb**. However, while this is correct for the autosomes, a somewhat more accurate estimate for the X is .08 Mb. The X chromosome has a lower recombination rate relative to the autosomes. A centriMorgan (CM) is the 1% probability that two loci on the same chromosome will recombine at a given meiosis. This **varies substantially by region**. For example at Xq13.3 the rate of recombination is 0.16 cM/Mb in contrast to the average of 1.3 cM/Mb on the chromosome as a whole. One study contrasted Xq13 (0.166 cM/Mb) to Xp22 (1.3 cM/Mb) in 15 Eurasian populations. Hence due to the low rate of reshuffling of blocks, it is expected that matches here are very ancient. Many studies consider a region of the X which is > 0.5 cM/Mb to be low LD with >0.0 cM/Mb medium to low combination and hence high recombination rates. Thus it is likely that any demographic even will be better preserved or conserved in the haplotypes of haploblocks in zones with low cross over rate, which can be selected to show snapshots of different points in the history of a population.

- **Linkage disequilibrium (LD) is the tendency of SNPs or markers proximal to each other to "ride together" avoiding recombination**. If separated by only a few thousand bp it is highly likely that two markers will stay together during recombination when large blocks of each arm of the X exchange genetic sequences of bp. Hence large haploblocks are formed each time a cross over occurs and sometimes former haploblocks will be separated. LD is greater on the X than the autosomes (recombination is 2/3 of what is found on the autosomes); and the size of the regions with a single genetic history will be larger. As expected, in the above noted study of 15 Eurasian populations, there was much stronger LD in all populations between markers on Xq13. Here the rate of recombination is 0.16 cM/Mb hence little "reshuffling" likely to occur here. Sequence variation suggests and African origin, with 9 vranches as off-shoots of the original ancestor (but with Africans found having each variant). Despite the relatively small size (162 Kb) there were haplotypes characteristic of Europeans and others of Asians, but it was typically that some of each were found in both populations. However haplotypes do show a cline suggesting admixture in zones of population contact and extending beyond in somewhat unpredictable ways. Hence to see the effects of an ancestor population by carefully chosing the region based on recombination rate one could "look for" ancestral relationships in theory that are "Out of Africa", Paleolithic, Neolithic, and more recent. This would have important implications in the search for the "best match" Asian populations to the Native Americans whose shared ancestry would extend back 10 to 20 KYA.

- For an individual, the series of recombination events over the past few hundred years are likely most important in the LD in haploblocks; but further back in time it is population demographic history, often mediated by **genetic drift**, that contributes most to LA patters. Hence back in time an uncommon crossover event could have occurred which was amplified dramatically by the chance factors of genetic drift.

- An unexpected increase in the LD rate on the X could be in indication of a **founder effect** in a population.

- It must be noted that there is what is termed an **ascertainment bias** in the SNPs used for analysis. Many were obtained from studies involving Europeans and so it should be no surprise that people of European descent typically have the highest levels of heterozygosity (diversity). Also the samples of Native Americans available for research (such as in the HGDP-CEPH panel) lack heterozygosity due to demographic factors. For example the Karitiana of South America are essentially an extended family of about 150 persons – all related in overlapping ways.

- It is likely that the X follows most of the same "rules" as the autosomes and if so then the magnitude of **LD can vary significantly between populations**. In general it seems that Africans have the least LD, while East Asians and Native Americans have the greatest LD – likely due to founder effects via "Out of Africa". Although differences do occur within regions most of the vaiance is within geographic populations is relatively similar. In general Africans in particular and Europeans would show less LD (and often shorter haploblocks) than East Asians, Pacific Islanders or Native Americans. Africans seldom show any significant LD for haplotypes greater than 20 Kb; relative to Europeans at 35 Kb and Native Americans up to 70 Kb. The latter three cluster together and seldom would for example European haploblocks of any great length be found in any of these three populations; but for example it is more likely that there would be similarity between East Asians and **Native Americans** with the latter **being a subset, but somewhat unique due to bottleneck, drift and founder effects**.

- It is possible that **the X chromosome may better help to identify the nearest Central and East Asian relatives of Native Americans** (relative to the Y or mtDNA).

- Even when mapping 60 Kb haplotypes on the autosomes, it is frequently observed that the most **common** African **haplotype** is "uncommon" elsewhere, and also that the most common European haplotype is possibly rare or absent in most other populations.

- **The X and the autosomes have different population histories**, with for example a higher percentage of SNPs of notably high derived allele frequency (but few rare derived alleles in comparison with the autosomes), especially for the East Asian and European populations. Some interpretations include greater drift in these two populations relative to the African group. This may be consistent with a more **"dramatic" bottleneck at the time of the "Out of Africa" event**. There is evidence of greater bottleneck events also.

- There appears to be a second occasion when **severe drift and bottleneck** occurred in the non-African, East Asian and European populations. This was perhaps during the **Last Glacial Maximum**.

- In a stable population (which has reached **"Hardy – Weinburg Equilibrium"** - HWE), with a 50/50 ratio of men (each with one X chromosome) to women (each with two X chromosomes), there will be a set of 3 X chromosomes to every 4 of the 22 paired autosomes, and hence there should be .75 diversity on the X relative to the autosomes.

- **Estimates of effective population sizes of the X chromosome** appear to be higher than "expected" in relation to the autosomes (3/4 the diversity), at least in recent (2008) studies. Specifically, rates of .85 were found in the San (African) and 1.08 for the Basque (European). However in another study the estimates of the effective population sizes for the X can be half that of an autosome (in this case chromsome 10). The latter rates occur when the authors take into consideration the fact that in a population in HWE will have 3 X chromosomes for every 4 autosomes; of the 3 X only two are in females and recombining and thus at any given time only half of the X chromosomes will be undergoing recombination (which is precisely what is shown in some studies).

- Estimates of the **ratio of female to male effective population sizes** range from 1.8 in the San to 14.0 in the Basque.

- **Africans have the greatest rate of variability (diversity)** among all Continental groups. As a rule, East Asians resemble each other more than they resemble Africans. This likely is related to the greater diversity in the latter and events immediately subsequent to the "Out of Africa" event(s) – such as bottlenecks.

- It appears that a **"burst" of X chromosome drift** occurred before the East Asian and European populations split. African chromosome X's are more differentiated (show a higher degree of variation) than autosomes relative to all non African populations. This could reflect a more severe male bottleneck(s) or some type of selective pressure around the time of the "Out of Africa" migration.

- Looking at data examining **STR (short tandem repeat) markers** on region Xq13 (an area of very low crossing over so high LD), the closest pair or markers at 162 Kb apart (DXS1225-DXS8082) were in very high LD irrespective of population structure or historical ties. Hence recombination events between these markers were extremely rare, but there were many haplotypes (e.g., 27 in a population and each population often had dramatically different diversity levels). **Haplotypes can be very common in a group, rare in others** with for example some East Asian types being "almost absent" in Europe but do increase in certain populations such as tribes in the Volga-Ural region suggesting a gradation or cline tapering off to the west.

- There are **“ethnically variant” SNPs** on all chromosomes. Using a pool of over 3 Mb SNPs, one group identified over 100 Kb of SNPs which appear uniquely in a particularly group (using HapMap data). Broken down by Continental group, 74% were linked to the African group, 15% to the European cohort, and only 7% related specifically to the East Asians and evenly distributed across all chromosomes. However, looking at the top 1% of these, there were three times more SNPs (152) on the X than on any of the others. Hence the degree of population differentiation is greater on the X than any of the 22 other chromosomes.

- There are **recombination “hotspots” and “deadspots” on the X**. In other words some regions are less likely than others to experience cross-over or recombination and hence will exhibit more conservation of haploblocks. Recombination on the X is particularly likely to occur at these “hotspots” with little activity elsewhere. A study of 11 mother – son meioses (one mother) showed that in this family on the q arm two regions of high recombination occur, Xq21.3 (15 Mb) and Xq27 (8 Mb). No recombination hotspots were seen on the p arm, orcentromeric chiasm.

- The above analysis gives **an average of 2 recombination events per meiosis** (1.0 p arm; 1.27 q arm), translating to **blocks 19.3 Mb in length after 4 meioses**. Similar, more general, observations can be made by referring to the demonstration at decodeme and the “Mendel Family”, the 23andme demo family.

- When comparing oneself to samples such as decodeme, for example with their “Compare Me” feature, it is imperative to determine whether the reference sample is male or female. Generally if there are a lot of green bands (medium sharing) this is a female when looking at the X chromosome.

- There appear to be **“stitching zones” or “anchors” at a single SNP where a block begins, and these regions can vary significantly by population**. In an examination of the author’s output of an analysis from the program PLINK using the HGDP-CEPH data, the 6.5 Mb match with a Xibo from northeastern China began at 80,509,534. The only individuals of 1045 woldwide who had the same start position were two Han – Xibo, 15 matches between groups of Native Americans, and one match between two San.

- **Two particularly well conserved areas are on the area on both sides of the centromere, where the p and q arms of the X join at about 60 Mb from the top of the p segment. A second area is the region on both sides of 80 Mb from the top of the p, which is about half way down the chromosome**. This area of about 16.8 Mb includes a 2.7 cM/Mb over an enormous distance. Hence it is the region least likely to recombine over for example 4 meioses. Concerning the centromere it is not unusual to find haploblocks of over 100 SNPs which in turn span 2 or even more Mb.

- **After 500 or even 200 years it is unclear how many haploblocks will be retained from any particular ancestor**. However the medical literature using pedigree analysis shows that for certain types of colon cancer, a haploblock of for example 8 Mb can be conserved for 300 or more years.

- It is of particular interest to determine **how large a haploblock might potentially be after say 500 years**. In other words could a relatively large block from a single ancestor be conserved over a period of half a millennium. Even at 300 years, although some ancestors may have "dropped off" the X, the contribution of one or more from 9 or more generations may still be substantial. This question has relevancy when trying to ascertain the specific contribution of any one ancestor (e.g., a Native American gggggg grandparent). Good evidence comes from disease studies, specifically of those with a genetically determined form of colon cancer. In families with a founder mutation of MSH2, haploblocks of 0.6 to 2.3 Mb have been observed in relation to the gene and flanking regions on chromosome 2. The "core haplotype" here was 0.59 Mb. As yet this mutation has not been observed in Europe and has been attributed to a couple who came to the USA about 1700. However estimates of the age of the mutation range around 500 years, and it is suspected that the source is Scotland. These data are rather similar to that relating to another form of colon cancer involving the APC gene. In this instance the mutation has been traced to a couple who arrived in the USA about 1630, and that some descendants have a haploblock of 7.17 Mb surrounding this gene. These situations likely give an upper range back 400 years. Much more common would likely be smaller haploblocks – but this would depend on the specific region of the chromosome since some areas are well conserved.

- Some **haploblocks** with specific haplotypes including SNPs and STRs **have been converved for many thousands of years** and may or may not have any obvious connection to higher survivability or selective advantage. In the case of the lactase persistence allele that has reached virtual fixation in some regions of Europe, the C-13910>T allele is embedded in an extended haplotype of 500 Kb or longer. Here the most common haplotype is "A" which extends well beyond the LCT gene to include long flanking regions resulting in a large percentage of Europeans having a 1 Mb haplotype in this region. It is likely that something similar would occur on the X. It is generally assumed that a block of this magnitude conserved in this way means a "recent origin", but "recent" here is likely to be 7000 YBP when agriculture was introduced to northwestern Europe (and older if one was to explore the origins thousand of years before this). In another study authors sequenced a 10 Kb noncoding region, with "little evidence of recombination", of region Xp11.22 (bases 50,583,087-50,593,170). In testing areas within this zone in populations around the world 67 haplotypes emerged, with 7 which occurred in more than one geographic area (3 haplotypes made up the majority of the world's diversity), and the estimated date of these haplotypes was 1.44 MYX (some common within and without Africa, others rare in one place or the other). In a wider sampling they found that even the one haplotype not seen in and African sample (e.g., from the Human Diversity Panel) was located there – suggesting founder effects. Had the authors not "gone searching" for the one haplotype not found in Africa, false conclusions could have been reached and this alerts researchers as to the limitations of the HGDP-CEPH samples. It is also noteworthy that the authors could not rule out a contribution of an Archaic human population (e.g., Homo Erectus) to this region.

- It should be noted that while the centromere exhibits, in general, a tendency towards extended homozygosity this does not occur in every individual.

- In these **regions of "extended homozygosity"** there may be few haplotypes. Those over 1 Mb are common, and may average about 34 per individual, and average 73 homozygous markers. The non-random distribution of these units are thought to reflect regions of the chromosome where recombination is rare. Here there are frequently two to four unique haplotypes.

- One study defined **extended homozygosity** as a minimum of 10 consecutive homogeneous markers (e.g., SNPs) extending over 1 MB (although the average number was 73 markers). They were looking at diploid chromosomes so used female subjects. **The X appears to be the most "dense" chromosome in terms of these zones with an average spacing of only 35 Mb between them**. Hence the 4.3 regions of the X contrasted sharply to 0.2 on chromosome 22 (which has a distance of 146.7 Mb between regions of extended homozygosity). At some locations up to 30% of their subjects were homozygous here (suggesting very common haplotypes in that population).

- Zones of extended homozygosity appear in specific chromosomal locations, regions where there appears to be a **suppression of recombination** (the X having more of these than any other chromosome).

- In admixture studies, **changes in ancestry only occur at "stitch points"**, and the size of "ancestry blocks" will reflect when the indigenous populations came into contact so more recent gene flow will on average result in longer ancestry blocks.

- **It is unclear how small an ancestry block could be and still reflect ancestral significance (true block from an ancestral group)**. This means generally using empirically based definitions as to where to set the bar to both avoid false positives and minimize false negatives.

- At the moment there are a number of **different algorithms and programs that will calculate percentage of majority and minority ancestry**. **Decodeme** (circa 1 Mb SNPs) and 23andme (circa 500 Kb SNPs) use different approaches and provide different output. The former has a browser and allows the customer to adjust the amount of sharing along any chromosome from 10 Mb to 1 Mb to .1 Mb. They may be more "generous" in giving percentages of minority ancestry (including too much noise) and so too many false positives. On the other hand **23andme** at present only provides a visual output and no browser, and no painting of the X. For the autosomes they seem to be "cautious" but it is unclear how many bp constitute a block worthy of "painting". It seems that the reference samples show a lot more small (1 Mb) haploblocks than show in the customer's output. Hence there may be more false negatives here. It is unclear how they are going to approach painting the X. With the "Family Inheritance" feature they use a .01 Gb criteria of sharing for determination of a relationship within a genealogical timeframe.

- **The author's experience**: My goal is to be able to identify a specific ancestor's signature on the X chromosome. This might be a challenge with a majority of ancestors who are European, and many from the same regions (e.g., East Anglia, England). The one exception is a well documented Six Nations Native American ancestor who was born in 1747 at Canajoharie Upper Mohawk Village, and a member of the Astawenserontha Bear Clan of the Lower Mohawks (Iroquoian speaking people). This lineage has been traced on both sides to the mid 1600s. What is known, however, is that "true Mohawks" were rather rare even by the mid 1700s due to the adoption of many war captives (mostly other Iroquoian peoples such as the Hurons), and even the Siouan – speaking Catawba (Flathead) of North Carolina would be well – represented in the Mohawk genome of the time. Since this ancestor made only a 1/256 or .004 to the autosomal part of the author's genome one might not expect much more in relation to the X. However, chance factors have resulted in the ancestor's son making a 1/16 or a relatively large 6% (statistically, but could be more or less) contribution to the X. Hence the assumption is that there will be Mohawk sequences to the tune of about 5 Mb someone on the chromosome (either in one block, or scattered in smaller units or both). As it turns out, based on the decodeme data which has twice as many SNP markers tested as it competitor, the author's largest haploblock is 364 SNP markers and about 5 Mb of nucleotide bases in the region between position 80,920,459 and 85,610,696 and matches exactly one of the 672 males in the HGDP-CEPH panel of 53 population groups worldwide. This individual is a Xibo from Northeastern China (Manchuria), with the next largest match here being 307 markers with a Yakut (Lake Baikal, Siberia). These two areas are those proposed as the homeland of the ancestors of the Native Americans. Since there are no Native American sequences from the Northeast or areas north of Mexico it is not necessarily likely that there will be exact matches to these groups, but rather to the "root stock" – either way the separation time would likely be over 10,000 years.

- There is another area from about 86 Mb to 102 Mb where even with dropping the bar to the most liberal (but questionable) level of "match", there is nothing, not even fragmentary, in the HDGP-CEPH panel. A reasonable hypothesis considering the flanking block below 86 Mb ("Xibo Block") is that this part is a Native American block, found at low frequency in East Asia (if at all), but via a founder effect would show if there were reference samples from the Great Lakes region (where the author's ancestors originate). There is an interesting parallel with mtDNA, where haplogroup X2e is one of 5 Native American founding lineages. It is only found in the Great Lakes region, and not in Asia.

- See http://davidkfaux.org/Autosomal_NA_Analysis.html for more specifics. The data on the X analysis is continued at the end of this page.

Please note that the above "fact sheet" is a work in progress and is in its preliminary stage. Not only does more information need to be incorporated, also what is already noted needs to edited and organized in a more meaningful way.

David K. Faux, Ph.D.

Changes and additions please contact the author at fauxdk [at] gmail.com
Version: 8 February 2009