Hundreds of Surnames Available for Genetic Matching in a Unique Czech Database

by Marek Blahuš
<marek@blahus.cz>

My current home town of Brno, the second largest city in the Czech Republic and the capital of Moravia, has been called the birthplace of genetics. Here, in the garden of the St. Thomas Abbey on today's *Mendlovo náměstí* (Mendel square), the Augustian friar Gregor Johann Mendel conducted a set of pea plant experiments between 1856 and 1863 that led to the discovery of inheritance principles which later became the core of classical genetics. Mendel presented his results (in German) at two meetings of the local Natural History Society in 1865 and had a paper published in its journal in 1866, but he never received enough attention from the scientific community and only got three citations over the next thirty-five years. Shortly after publishing his discoveries, he was elevated as abbot and had to spend most of his time running the monastery, which he did until his death in 1884. His grave can still be visited at the Brno Central Cemetery and the monastery now hosts a museum that presents his life and heritage (pun intended).

Re-discovery of Mendel's work in 1900 gave birth to the science of genetics, and the advent of molecular biology eventually led in 1953 to the discovery of the molecular structure in which genetic information is stored in our cells – DNA molecules arranged in a double-helix – which earned Crick and Watson, American and British scientists from the University of Cambridge, a Nobel Prize in Medicine in 1962. This new understanding of inheritance has enabled us to do much more than just breed peas more effectively – an individual's genetic information was found to be unique and superseded the fingerprint, with DNA having been used for the first time as evidence in court in 1986. Mendel back in the 1860s had no idea how complex the human genome is, but he has left us the notion that traits can travel through generations as discrete, inheritable phenomena, although he has described only one of the possible patterns that this travel can fall into.

Genealogists are good at tracing family ties, and the essential event that establishes a new individual in the family tree and provides continuity to the family is the birth of an offspring – a new family member, who is not only given a name and a surname, but also a unique genetic makeup. Traditionally in our Central European context, a child's given name has largely been the choice of the parents, who would typically select a name of a saint (at least if the family was Catholic) of their preference, although there were some written and unwritten rules limiting their choice, such as that two siblings could not share a name (this rule is valid till this day), or the habit of naming a child after the saint who was commemorated in the calendar of the Catholic Church on or around the birth date. Surname followed a different inheritance pattern, at least since 1786 when a law declared that everyone must have just one which cannot be changed and should be passed to the next generation; women accepted the surname of their husband at marriage.

St. Thomas Abbey in Brno. Foundations of Mendel's greenhouse in the front. Photo by Misa.Jar & Dominikmatus, CC BY-SA 3.0.

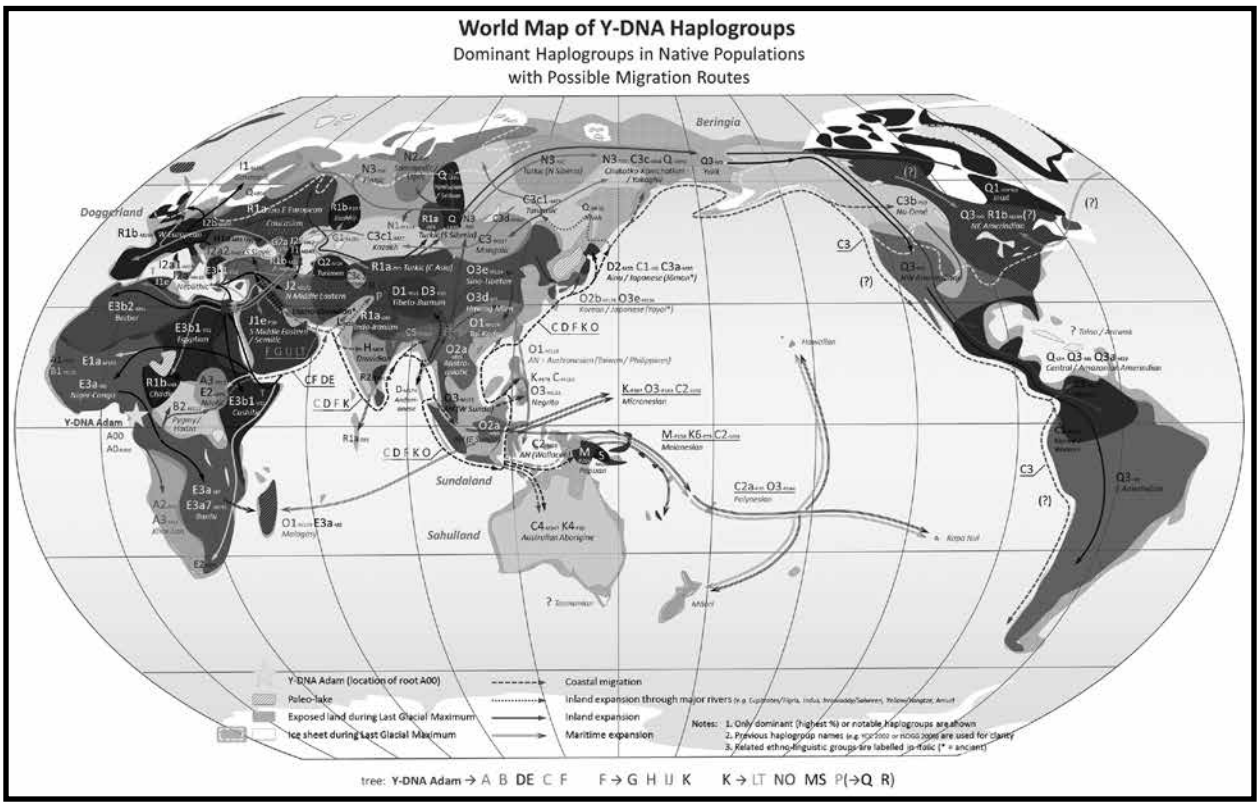
Genetic recombination that is part of human

Marek Blahuš speaking live about genetic genealogy on Czech TV in 2017. Source: Czech Television.

reproduction, on the other hand, happens already at the moment of conception and does not depend on human-made decisions at all – other than who the parents are, of course. Half of each parent's genetic information is discarded in the process, so that a new person can be conceived, who is a unique mixture of the father's and mother's genes (even each sibling is different, because other 50% of each parent's information gets passed). It has seemingly worked like that since the first days of humanity, but it took us people a very long time since then till the 20th century to understand how our own biological inheritance works behind the scenes – to be able to read and follow the flow of genetic information in our families. Many genetic genealogists will tell you that you have two trees – a *genealogical* and a *genetic* one – and that there are actually two separate fields of research – *traditional genealogy* and *genetic genealogy*. I elaborate on this by saying that the two work best when combined, and point out that *genes never lie* and as such record a story of our *actual ancestors* from the *ideal world* in which there were no illegitimate children, no misattributed paternities and no hidden adoptions… A world other than the one we know as a society, but which still exists somewhere there, encoded deep in our genes – a parallel universe of the past that makes the biological whole of us, whose parts can be retraced to our actual ancestors. If only those genes had names on them! And if the ancestors did not wash out of them so fast!

Indeed, 98% of human DNA is *autosomal DNA*, which is what people usually think of when hearing about genetic genealogy these days. This is the part of our DNA where the 50% rule applies literally, at least in the transfer of genetic information from parent to child. Approximately 25% of our autosomal DNA is shared with each of our grandparents, but the odds quickly come into play with more generations and there is a 10% chance of us not sharing any detectable DNA with a third cousin – this happens when no piece of DNA has been passed all the way from our closest ancestors (in this case our great-great-grandparents) to *both* us and that cousin. From a different point of view, our DNA is likely to miss any traces of on average 6 out of our 128 g-g-g-g-g-grandparents (people who were born around 1750) and this percentage quickly raises as we proceed further back into history and the number of ancestors doubles with each generation. So, while we can have up to 1024 ancestors in our pedigree born around year 1650 (this is about how far back a Czech genealogist can usually get if church records were preserved), only less than a half of those people have left any genetic trace in us. Moreover, it is a big challenge to decipher who is who in this genetic melting pot – scientists are still struggling to identify just meaningful "ethnicities" in our autosomal DNA, not even distinguish individual ancestors and link them to names known to us from historical records. And as time goes by, even Americans searching for their European homelands who currently get by with an autosomal DNA test (such as the one from AncestryDNA) will eventually need to turn their attention towards other types of testing – due to the limited time autosomal DNA can bring us back into our genetic past.

There is, however, a little part of human DNA which follows a much more timeproof pattern of inheritance: the Y chromosome. It is present only in men, as this is the chromosome that defines masculinity. Apart from 22 pairs of autosomal chromosomes, human nuclear DNA consists of one extra pair of so-called *sex chromosomes*, which are either two X chromosomes (in case of women), or an X chromosome paired with a Y chromosome (in case of men). The Y chromosome is about one third of the size of the X chromosome and it also does not carry many traits – only a minimum of human genes is Y-linked. What makes the Y chromosome special is the way it is passed – from the father



World map of dominant Y-chromosome haplogroups with possible migration routes. Artwork by Chakazul, CC BY-SA 3.0.

to the son. The mother provides an X chromosome, but the father's Y chromosome does not recombine with it, so that it is passed on to the next generation intact. In fact, almost intact – except for random mutations that happen every now and then due to minor inaccuracies in the reproductive processes. Such mutations result in a slightly modified Y chromosome – which is then passed to all future descendants in the *patrilineal* (i.e. *direct paternal* or male-only) line.

Due to the existence of mutations (and their opportune speed), genealogists can use the Y chromosome to trace human patrilineal lines and distinguish individual branches from each other, starting from living individuals and going back through Y-DNA isolated from ancient skeletal remains all the way to the hypothetical "Y-chromosomal Adam," creating a large and deep family tree of mankind (*not* the whole humanity) in the process. Branches in this tree are defined by mutations that have started them, which are called *SNPs* (single-nucleotide polymorphisms), and only people in the last few levels are known to us by names (based on written records). The human Y-chromosome phylogenetic tree, as it is called, spans an estimated 250,000 years and is being constantly improved as more and more men get their Y chromosome tested. Vast majority of these tests are performed by Family Tree DNA in Houston and I can highly recommend that you (or a male relative of yours, such as father or brother, or your husband) take their Big Y-700 test if you are interested in participating in this research and learning something about the deep origin of your patrilineal line. That test is still pricey although its price has been dropping in the last few years (it is now available for $399 in the ongoing Christmas sale), but you can also start with one of the subordinate Y-37, Y-67 or Y-111 tests and get an upgrade later.

It is important to realize, however, that the Y chromosome represents just *one* ancestral line out of many – a particular one that has had the luck of being recorded into our men's DNA in a much more resistant way than all other ones. Another type of resistant DNA information – mitochondrial DNA, which corresponds to the *matrilineal*, i.e. all-women ancestral line – has the unfortunate property of mutating *too slowly*, so that there are *too many* people who share the same mitochondrial DNA and it can only be used to make very

broad estimates about relationships, ranging in the thousands of years. On the other hand, the resolution reachable with modern Y-DNA testing (SNPs) is believed to be as detailed as 80 years (or two or three generations) per mutation.

Research of Y-DNA started earlier than the now ubiquitous autosomal DNA testing. First commercial genealogical tests became available for purchase from Family Tree DNA in 2000 and those were Y-chromosome and mitochondrial tests. AncestryDNA used to test the Y chromosome in the past too, but they discontinued it in 2014 in favor of the more profitable autosomal testing. Some other companies, such as 23andMe and Living DNA perform some Y-DNA testing along with their autosomal test, but the available resolution of the result only goes half the way it could and there is no possibility of an upgrade. A German-based family business called YSEQ DNA Shop offers a plethora of fine-grained Y-DNA tests for a good price, but an important disadvantage for the regular genealogist is that they do not maintain any database of other customers' results to compare against.

Czech genealogists have been always interested in Y-DNA research and its principles are more widely known among them than among their American colleagues. This may be owed in part to the easier availability of the testing equipment (simple Y-DNA tests are today being routinely performed by every other forensic institute using a commercially available technology), but there is definitely also a link to the needs and interests of Czech (and more broadly European) researchers. As alluded above, it is *not that difficult* for a Czech genealogist to research his pedigree all the way back to a thousand ancestors back in the 17th century, provided that records have been preserved. Most of these ancestors were typically Czech and many times the families would live in the same region (or even village!) for centuries. It is therefore of little surprise that autosomal testing results will not be of much value to such a researcher – indeed, the typical Czech gets a "33% Eastern Europe, 33% Western Europe, 33% Southern Europe" ethnicity mix, and would possibly get a "100% Czech" if only the companies and their algorithms supported this (which might not be that easy to achieve because the Czech lands are situated not only at the geographical, but also at the genetic crossroads of Europe). Given this uniform ancestry and lack of interest in locating remote cousins (read Kate Challis' article elsewhere in this issue), it comes natural that Czech genealogists are more interested in researching their deeper past – older than the one that is attested in written sources – with the aim of extending their existing family history research, to the Middle Ages and Ancient History, although at the expense of limiting their line of sight to one particular ancestral line only and the obvious lack of concrete names in those old times.

So-called *surname projects* have been emerging ever since the commercial availability of Y-DNA testing. They are based on the observation that surnames in many cultures (including Czech society) are passed down from father to son, just like Y chromosomes. Therefore, people who share a surname can use Y-DNA testing to determine whether they share a common ancestor within recent history. Because some surnames obviously emerged independently several times in various times and places – such as *Novák* which was a convenient way to call anyone who was a "newcomer" to the village – the results of a Y-DNA test may not always be the same even if the surnames are.

The occurrence of mismatching Y-DNA results in spite of a genealogically assumed relationship in the direct paternal line is often attributed to a *non-paternity event* that must have taken place somewhere along one of the two compared descendancy lines. A non-paternity event occurs when a child is believed to be fathered by another man than its actual father. In genealogical terms, it may mean anything starting from undocumented surname changes, through illegitimate children, to undisclosed adoptions. Naturally, the discovery of such facts may have social and medical consequences for the family, particularly if pertinent to the deeds of people who are still alive, so certain degree of caution needs to be applied when performing such a research.

On the other hand, many genealogists are stuck at an "unknown father" (in Latin the record would say *pater ignotus*, or the respective field in the church book may simple have been left blank) and cannot continue their research any further because of this missing piece of information. With genetic genealogy, they have at least a chance of once learning more about who the actual father might have been, or at least what his surname could have been, based on comparisons in public databases and provided they find a descendant in the direct paternal line who is ready to take a Y-DNA test. In some lucky cases, even two-century-old family mysteries have already been solved by making use of genetics.

Another goal achievable through Y-DNA testing is verification of already built patrilineal lines, which are otherwise the least certain kind of all ancestral

lines when following the paper trail – already an old Roman-law principle has remarked that, for biological reasons, the mother was always more likely to have been recorded correctly. If you want to verify that all your ancestors in the direct paternal line, up to a certain "family founder," were indeed the sons of their fathers, it is enough to buy two Y-DNA tests – one for yourself and one for a remote cousin (who shares the same surname and therefore is also related to the family founder in a direct paternal line). If the results for the two of you will match (or will be very similar, because small mutations occur naturally with time), you may feel free to assert that your research on your surname matches the biological reality of the past centuries, at least up to the most recent ancestor you and the cousin have in common. I have personally successfully proved the perfect alignment of my patrilineal lines in my genealogical and biological trees by successfully matching against a Czech-American fifth cousin; we are now guaranteed to share our g-g-g-g-grandfather Tomáš Blahuš, born in 1782 in Ostrožská Lhota, Moravia, and to be each descended from a series of five more men of the Blahuš (spelled Blahous in the U.S.) as recorded in written documents.

There are more than 7,000 surname projects run by volunteers within the Family Tree DNA website and you may search for your surname at their homepage www.familytreedna.com (scroll down and look for the "Search Surname" text box).

Even more importantly, if you have Czech ancestors, and in particular if your direct paternal ancestral line is Czech (i.e. if you have a Czech surname), you should pay attention to a unique Czech research project that has been collecting Y-DNA results of specifically Czech ancestral lines already since 2007. The project is called *Genetika a příjmení* ("Genetics and Surnames") and it has been funded by the Prague Institute of Criminology. Researchers there have been interested in learning more about the variety of Czech Y-DNA, in order to answer questions such as which is the most typical Czech Y-DNA *haplogroup* (an ancient subgroup in the phylogenetic tree) and what the probability is of two bearers of the same surname sharing the same Y chromosome. Knowing what "Czech DNA" looks like can have such philanthropic applications as determining the nationality of an unknown victim of a disaster that took place abroad. However, the most enthusiastic audience and the people who most readily volunteered to take part in the project have always been genealogists. They know the origins of their ancestors and they can find their own use of the results in running their all local surname projects, or possibly just small-scale two-person comparisons within the larger research framework.

Due to its early conception and the in-house character of the testing, Genetics and Surnames have been relying on STR markers only, which is an older testing technique, inferior to the modern SNP testing mentioned above. *STR* stands for "short tandem repeats" and it is basically concerned with counting the number of repetitions of short fixed sequences (such as "G-A-T-A") in certain places of the Y chromosome's genetic code. The number of repetitions in each place is called a *marker* and a set of standardized markers (called a *haplotype*) measured in one person's Y-DNA can be compared to the markers of another person. The closer two people are to each other in their patrilineal lines, the more similar their haplotypes should be (the values of markers could be identical or some of them could differ by one or two). As haplotypes are just relatively short series of numbers, STR results are easy to handle, although comparing them results only in probabilities, such as "if you share 25 out of 25 markers, you are 85% likely to share a relative in the direct paternal line at most 8 generations ago."

The Genetics and Surnames project uses the PowerPlex Y23 System set of 23 STR markers, which means that people who have tested their Y-DNA through this project have received marker values for a particular set of 23 locations (it used to be only 17 in an early phase of the project). Fortunately, this set of markers is mostly compatible with what Family Tree DNA is testing for (all 23 are covered in their Y-111 product, but even the cheaper Y-37 and Y-67 allow a basic comparison). Also, importantly, the state-of-the-art Big Y-700 test does not only cover all of SNP testing (the modern technique), but it comes combined with a thorough test of 700 STR markers (as the number in the name suggests), which is more than enough to make the test results compatible with those of any other inferior tests including the one performed by Genetics and Surnames.

Since the Genetic and Surnames project's inception, a Czech genealogical website called *Genebáze* (www.genebaze.cz) has been hosting its results within its set of Czech (and Slovak) DNA (Y chromosome and mitochondrial) databases. The website has been partially translated into English. Direct link to the Czech Y-DNA database in English is: https://www.genebaze.cz/cgi-bin/cyd.cgi?lang=us Direct link to the Slovak Y-DNA database in English is: https://www.genebaze.cz/cgi-bin/ydsk.cgi?lang=us

At the time of writing, the Czech Y-DNA database contained almost 3000 Y-STR results linked to Czech patrilineal lines/surnames, out of which most but not all come from the Genetics and Surnames project. The Slovak Y-DNA database is much smaller – merely 300 records – because the research is concentrated on Czech samples. Nobody – other than the employees performing the tests in the lab in Prague aside of their regular work – is paid for the collection and processing of samples from donors, and donors need to document their direct paternal line at least back into the 19th century to qualify for participation. Earlier criterion also requested donors to have a less frequent (100 to 200 occurrences) surname of Czech-language origin, but this requirement has since been discontinued. What is very important in Czech context is that the testing is free for the donor *and* they are going to get their results when it has completed (everyone gets their haplotype and their approximate position on the phylogenetic tree – a haplogroup – which can give them a basic idea of their direct paternal line's ancient history).

For these reasons, many Czechs' first experience with genetic genealogy has been through this project, and many people in the Czech Republic and Slovakia have first heard about genetic genealogy through reports about this project in the media (TV, radio, newspapers, magazines, news websites) that were the result of the project's active promotion since 2015 and which have helped to grow the database size rapidly, although the testing process has always been rather slow as it includes many logistically demanding steps and lot of communication, samples handling and data processing. Apart from the project leader Vlastimil Stenzel and the long-time administrator of Genebáze, Ludvík Urban, much of this success is owed to the work of Martin Kotačka – a young genealogist, current director of the Brno University of Technology Archives, and a good friend of mine. I have been happy to provide him some help in the process, but the biggest part of the often-tedious work spent on the project administration has been carried out by him.

The Czech and Slovak Y-DNA databases at Genebáze are open to external submissions too, so if your direct paternal line is of Czech or Slovak origin and you have done a Y-DNA test, please consider submitting your STR markers into the database. If you found the process too difficult (it is not always straightforward to locate the correct markers in the form), you may send me your results by email and request for assistance. The database welcomes any relevant submissions.

Once your profile has been entered in the database, you will be able to search for matches among all the other samples and if you find a match that might be relevant, you can consider contacting the person and getting a more advanced test done for both of you to explore the possible relationship more closely.