Grenzen

in genealogie en heraldiek

Frontiers in Genealogy and Heraldry

Frontières dans la généalogie et l'héraldique

GENEALOGICA & HERALDICA

XXX

2012

2000 years of genetic variation in Flanders, Brabant and Limburg

(and an early look at European settlements)

Dr. Wim Penninx
Delft University of Technology

Introduction

We have an interesting puzzle. In genealogy we can make our pedigree of the family as far as written history is preserved in the archives. In Western Europe the births, marriages and dead are written during the last four centuries in the municipal and church registers. Before that time the notary and alderman archives, where the testaments and possessions are registered, are sometimes preserved. Before that period it stops for most families. For very few patrician and nobility families the history goes sometimes back until the Middle Ages. Then genealogy stops.

Where the written history ends, the genetic genealogy starts. We use a small change that occasionally occurs in the DNA reproduction process to a person, and after that, the change subsequently is passed to their offspring, and can discriminate between this offspring and the other family members.

The places on the DNA, where we look for the genetic changes are called marker places, shortly markers. On the average a marker place (type STR, see below) shows a mutation in one of the hundred generations. When we test fifty markers we have the statistical probability to find one mutation every two generations.

Because we know the mutation rate of the different markers – ranging from 1 every 30 to 1 every 100,000 generations, we can estimate on the basis of how many and which markers mutated when two families come together.

By using DNA, we can now extend out genealogic data back to a longer timescale. We do not have all data to create a family tree with all details, but we can deduce certain information from our ancestors.

In this article I will show the results of simulations of population growth in Europe (starting 8000 years before present), respectively Flanders, Brabant and Limburg (FBL) and how it affects the variability of Y-DNA. I will show that the general SNP-variability of the European SNP-variability fits the simulations, if we use a higher population growth than was presented in [1]. The measured STR-variability in FBL fits the population growth of the last 2000 years. Figure 1-3 show three "family trees" in different periods in time:

A short summary on DNA data for genealogy purposes

In this paragraph I describe the typical characteristics of changes on the DNA and the research one can do with it; I will not describe exceptional behaviour of DNA. Four types of DNA can be studied for genealogy purposes:

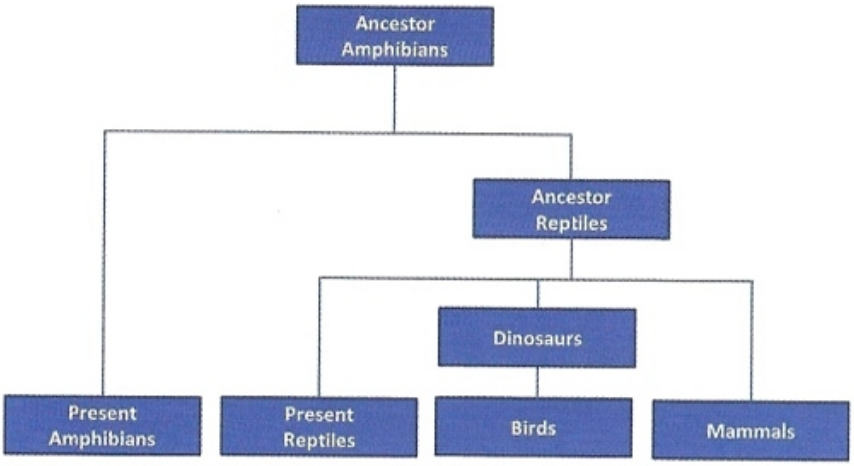


Fig. 1. A family tree from Amphibians to Mammals.

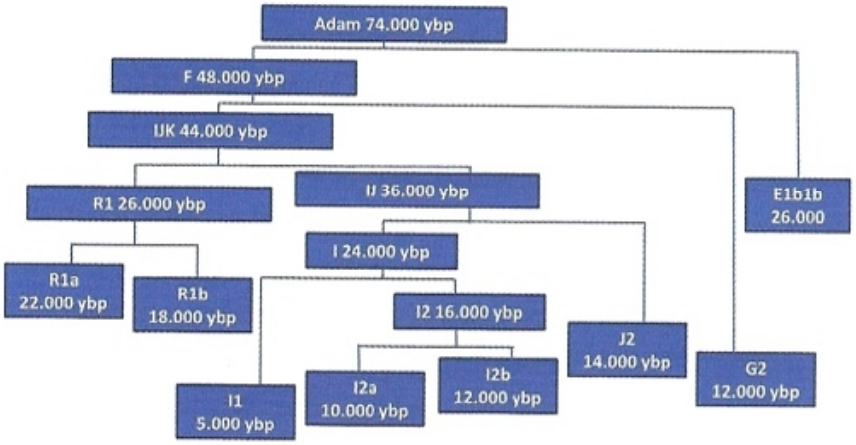


Fig. 2. A family tree from Y-DNA father Adam to haplogroups.

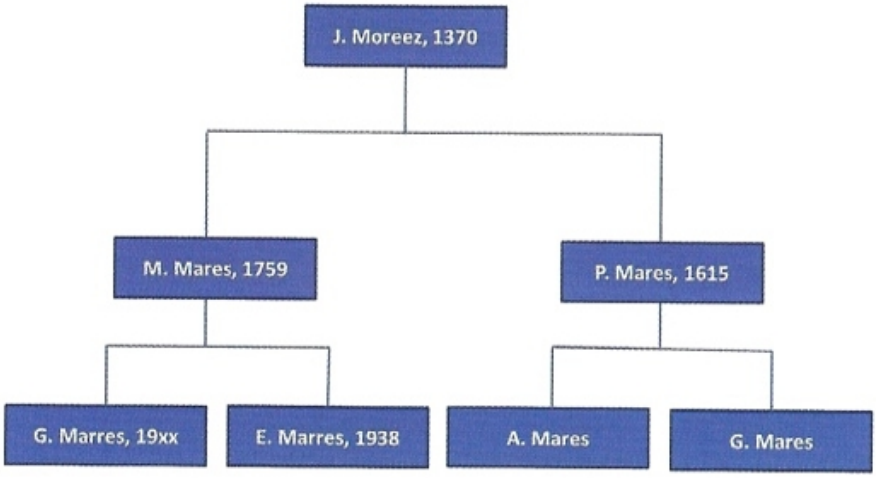


Fig. 3. A family tree from our oldest genealogical ancestor to ourself.

- mtDNA, the mitochondrial DNA, given from mother to her children
- Y-DNA, the Y-chromosome, given from father to son
- X-DNA, a son gets his X-chromosome from his mother, a daughter gets it from both parents.
- auDNA, 22 different autosomal chromosomes, given from parents to children. All chromosomes are different and each have their own length and pattern.

The mtDNA and Y-DNA can be used to track DNA without a mixing from the two parents. Changes in these parts of the DNA can only occur by genetic changes. The Y-DNA is one of the short chromosomes. The mtDNA is very short.

X-DNA and auDNA are given from both parents to a child. A child gets parts (segments) of these chromosomes from both parents. By analysing the DNA of one person, it is not possible to deduce from which grandparents the parts of the chromosomes orginated. If one shares a 3th gr. grandfather (five generations), it is likely one finds shared segments. In case the relatives are closer related, it is very likely one finds shared segments. In case the distance is further, the chances to find a shared segment are small.

DNA has two kinds of genetic changes. SNP's and STR's. I will describe the main characteristics:

- an **SNP** can occur everywhere on a chromosome or mtDNA. SNP's have a very low chance to occur. In most cases an SNP occurred only once in the human history. E.g. the SNP that was called M267, occurred in the Y-chromosome of one man. All sons gave it to their sons etc. Every man that descended in pure male line from him, has this SNP, and nobody else. The people who have this SNP are called haplogroup J1-M267. On average every few generations an SNP occurs on the Y-chromosome. Since the mtDNA is very short, the chance that an SNP mutation occurs on the mtDNA is much smaller.
  Most mutations have no consequences. Some mutations do have consequences. Many of the differences between people in different parts of the world (colour of hair or eyes, tolerance for milk) are the result of SNP's on the auDNA.
- **STR**'s are very short repetitive sequences on a chromosome. One of them on the Y-chromosome is called DYS393. If this repetitive sequence is 13 times present, one says DYS393=13. In case a mutation takes place from father to son, the length of the repetition in generally goes up or down by one. In the case it goes up, we have DYS393=14. The chance that this value changes is low, but very large in comparison to an SNP. In the case of DYS393 this chance is about 0.0008 (and is called mutation rate). So roughly one in 1250 generations (about 45.000-50.000 years) this value of DYS393 changes. If we would have e.g. 62 markers and they would have an average mutation rate like DYS393, we have a total mutation rate of about 0.05. We expect in 20 generations (about 600 years) 1 marker to change. Some of the most often measured markers have higher mutation rates; if we measure 60 STR-markers, we expect a few markers to have changed after 600 years. Since it is statistics, sometimes we see 0 changes, sometimes we see more changes than a few. If we have many measurements, we have better statistics, and the average value can be determined better.

If one finds old human material (graveyards, Ötzi the Iceman, remains in Egyptian Pyramids), one will try to analyse the DNA. If it is well preserved (as Ötzi the Iceman), or the DNA is recent, one can analyse the DNA fairly complete. In case it is old, and not well conserved, one can still sometimes deduce the mtDNA characteristics or a few SNP's. One of the successes of analysing old human material was the analysis of the mummies in the era of Tutanchamon. One was able to deduce family relations

between the mummies that were found. The main conclusions, the parents of Tutanchamon were brother and sister, is accepted by most scientists.

In the rest of this article we use Y-DNA data, so we study male-line family trees. In genetic genealogy (and biology) one uses the word "phylogenetic tree".

Some family trees

In 2012-2013, it is quite expensive to measure a complete Y-chromosome. In 2013 a commercial company offered a complete Y-chromosome analysis for € 1000. Most people, that have their DNA measured for genealogical reasons, have measured a set of SNP and STR markers. For recent history (up to a 1000 years) most research is done using STR-markers. This is where traditional genealogy meets genetic genealogy. Surnames and Y-DNA are used to create family trees.

If we measure different people in a large male-line families, we can trace the DNA changes. The DNA of close relatives are very similar. Suppose a DNA change took place in the year 1700 in one of two brothers, and both brothers have male-line descendants. In that case we can trace which persons are descendants of the brother with the marker change, and who are the descendants of the brother without the marker change. In case we can measure several male-line family members, and we have genealogy information, we can determine in which period the mutation must have taken place. In case we do not have genealogical information, and we have two lines, we can roughly calculate how long ago this specific marker mutation took place. This is done by determining the amount of marker changes after the specific marker change took place. In case few marker changes took place after the specific marker change, the specific marker change took place recently. In case many changes took place after the specific change, the specific marker change took place a long

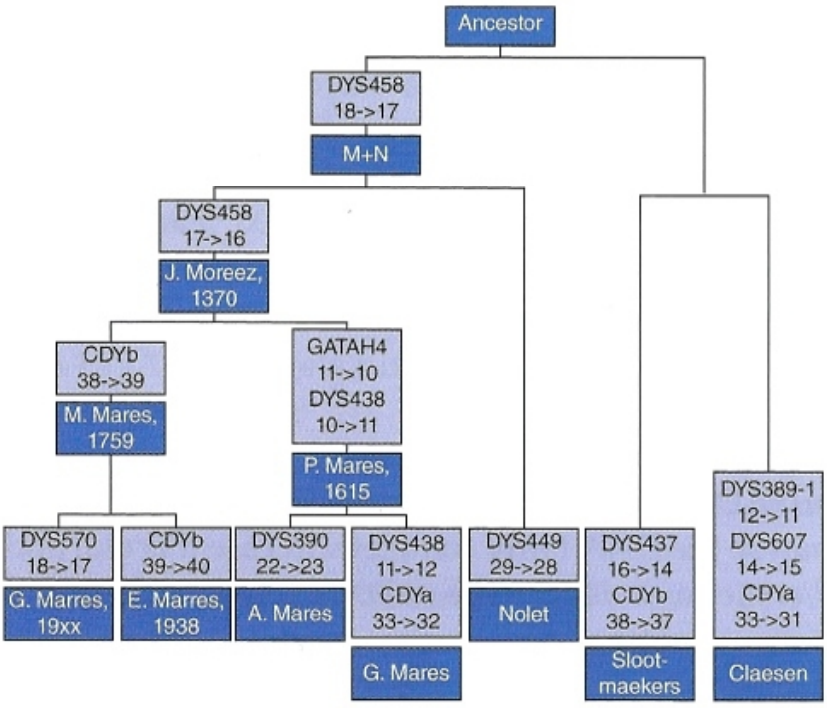


Fig. 4. The positions where the marker changes have taken place is drawn in this male-line family tree. This tree contains both genealogical data, as well as relations purely based on DNA data. The ancestor probably lived between 1000-1370. All persons have their oldest genealogical ancestors in the region of Maastricht-Liege.

time ago. This is determined by an estimate how many changes we expect during a period of e.g. 100 years. An example of a male-line family tree with genetic markers is given in given in figure 4.

Trees, population densities and statistics

The population size has been determined for a long time by the amount of food the population can grow in the area they have available. If an area is very suitable for crops, the population density of people will be large; a mountain area will have a lower population density. By increasing knowledge, we are more efficient to cultivate crops, and the population density increases. In case a small group of a population moves to an area without people, the population density will increase quickly to the density that fits the knowledge and the suitability of the area to grow crops.

Some examples:

- On an isolated island, a group will live during a longer time with a constant population size. After a quick an initial growth the population with stay at a constant population size. The Y-chromosome variability will be small, since the most recent common ancestor of all group members lived recently. This pattern is e.g. found in isolated indigeneous people in southern America and in small islands in the Pacific.
- When a large group settles in a new area, the ratio of different Y-DNA groups will be the same in the new area as in the area they departed. This pattern is e.g. found when the Vikings populated Iceland, where the haplogroups distribution is the same as the founding country Norway.
- If we compare the genetic differences between Europeans and Africans we see a large difference. In Africa the population started to grow a long time ago, which results in large genetic differences between Africans. In Europe the population growth was fairly recent, which results in a much smaller variability today.
  A recent strong increase of population is seen in the Ashkenazi Jews. This population grew strongly (in last 1000 years) which resulted in a small DNA variability in this group.
- If a population is well mixed, the ratio of different genetic characteristics will stay constant; independent of the population growth. If the population is not well mixed, it is likely that certain groups grow differently, faster or slower, than other groups, due to differences in chances to survive. In the case of destructive wars between populations, it is clear that the DNA of some groups will become extinct or will be minimized.
- In case an immigrant settles in a new country, and adjusts to the population that is already living in the country, the population growth of his DNA will follow the population growth of the country.

Some examples of population mixtures

A simulation

In this article I simulate the growth of a population for a period of time. I start with a population of size X as a start of the simulation. I choose a population growth for the total population in one generation, say alpha. I choose an average generation length of 30 years and define it as the average age of a father when a child is born. By assuming that each male has random chance to have male sons, we can create simulated male-line family trees for as many generations as we want. Each man will have an equal chance to have one son. In this model the chance to have one more son will be independent of the already born sons. I will use a Poisson distribution for the generation of male descendants. In this simulation I only look at sons who reach the age to become a father. In case a son will die earlier, or a daughter is born, this data is not used in the calculation.

To determine the ratio of the numbers in our calculation and a real population density, I used historic data, when healthcare was less than today. The oldest data of "Bevolkingsatlas van Nederland" was used (between 1800-1850). In a group of 1000 inhabitants, the birthrate was 11.6 boys per year. On average 61% of the boys reaches the age of 30, and is expected to become a father. From these values one deduces that 22% (=11.6*30*0.61/1000) of the population is expected to be a father in the definition as given above. In case one has 100 potential fathers in the simulation, it represents a population size of 100/0.22=450 people.

General results of the simulations

To simulate the genetic variability in human history, we use a population growth of alpha (growth in a generation). The simulations show that, in case alpha is 1%, only 1% of the males will have a male-line family tree after many generations. A large part of the other 99% will have descendants, but they will not have a male-line family tree. In case the population growth is larger, e.g. 5%. The percentage of the males that will have a pure male lineage will go to 5%. The easiest way to visualize this, is to think of a small island, where the amount of inhabitants is limited. In this article I will call the ratio of Y-DNA ancestors to the amount of potential fathers the "Y-DNA ancestor ratio".

The actual size of the group does not influence the the Y-DNA ancestor ratio. Suppose we have at an initial moment a population size of 1.000.000 people and the growth rate is 1%. After a long period of time after the initial moment we will determine the amount of Y-DNA ancestors from the initial moment, the expected value will be 1.000.000*1% = 10.000. In case the period of time is not long, the Y-DNA ancestor ratio will be between 1% and 100%.

Effects not calculated

In case chances to survive depend on social situation or warfare, the chances to have descendants are not equal. In that case the percentage of the males that will have pure male lineages will go down further.

A comparison between the simulations and measurement.

Two simulations were made for different periods and areas. It is important that in the comparisons with measured values a few criteria are met:

1. The variability is mainly determined by the evolution of the studied group. This requires that the influx of newcomers has a low impact on the statistics, and can be neglected (or calculated). It is easiest if the studied group stayed in the same area. The uncertainties with respect to mixtures are small in that case.
2. The variability can be compared with measured values.

The first study deals with the SNP haplogroups as a comparison for the European population in the early period in Europe (starting 8000 ybp); the second study deals with the STR variability in the area of Flanders, Brabant and Limburg in the last 2000 years.

The prehistoric population densities in Europe between 8000-2000 ybp

I started with the expected population densities of Europe in the period of 8000-2000 ybp (years before present), as determined by [1]. They determined a population growth value of 1.013. Simulations were created with 1.01 and 1.02 respectively, and using a European population of 29 million in the year A.D. 0 [1]. In the rest of this paragraph I discuss the results for a population growth of 0.02 and will argue that this is too low. For Europe this means a growth from 600.000 inhabitants in 8000 ypb (6000 BC) to 1.1 million in 7000 ybp and 29 million in 2000 ybp (the year A.D. 0). The simulation started with 135.000 potential fathers (600.000/0.22) in 8000ypb. I created male-line family trees. The results of the simulations are shown in table 1. In the simulation only a small group (2.565) of the 135.000 potential fathers has male line descendants at present in the simulation; this is about 2% as expected (see the paragraph "General results of the simulations"). This value can also be deduced from the amount of measured European unique Y-DNA SNP's that is older than 8000 ybp. With the present knowledge this value is a lot smaller than the value of 2.565. With the present knowledge about 10-20 are expected in Europe (a few G, one I1, a few I2, a few J2, one R1a, one R1b) [2].

The amount of Y-DNA genetic fathers was in Europe more likely to be on the order of 10-100 than 2.500 in 8000 ybp. Two reasons can explain this difference:

1. A large difference in survival changes for different groups: the likely introduction of lactase-persistency [3] and the introduction of cattle herds in Europe (possibly by R1b) gave this group a much larger population growth than the indigenous people of Europe.
2. The population density was overestimated by [1]. It seems likely that the influence of new knowledge and DNA-changes gave rise to a higher population growth than 0.02, and the initial value of 600.000 people in 8000 ybp is too large. A population growth of 0.03 and a population of 29 million in Europe in 2000 ybp would give a value of 78.000 people in 8000 ybp and an expected value of 500 unique Y-DNA fathers.

Haplogroup R1b is now the largest Y-DNA haplogroup in Western-Europe. The first split of R1b in Europe was probably after 8000ybp, R1b-L11 as parent of both R1b-P312 and R1b-U106, see [4]. I compared the expected population size of R1b from 8000 ybp until 1800 A.D. I first determined the R1b population size in Europe in the year 1800, so before the recent and strong population growth. This was determined by using the present values of R1b percentages as published on Eupedia [5] and using the published populations in the different countries in the year 1800 [6]. It is not expected that survival advantages of Y-DNA haplogroups existed within a country in the last 200 years. This calculation resulted is a population of about 80 million R1b in Europe in the year 1800. If we start with a group with one founding father R1b in Europe in 8000 ybp, this gives us an average population growth of 0.06. This value is not strongly dependent on the value of this value of one founding father R1b; in case we would have three founding fathers, the population growth would still be 0.06.

In case we follow the population growth of 0.06 for R1b, we expect an increase from 1 founding father in 8000 ybp to 7 pedigree lines in 7000 ybp and 48 lines in 6000 ybp. The amount of pedigree lines is on the correct order of scale. Since the percentage of time between the appearance of R1b in Europe and the expansion to 50 haplogroup probably took place on a scale of 2000 years, we expect that the STR-variability in the 2000 years is small. This is indeed what is measured: the STR-modal values of the different R1b haplogroups are close together. The variability within each of the R1b subgroups is large, since it spans a time period of about 6000 years, while the difference between the groups span a period of a 1000-2000 years.

Population densities in the region Flanders, Brabant and Limburg

This population was suitable to compare the simulations with recent STR-variability. In this region about 1000 people were measured with 37 STR-markers [7]. The population density is simulated with the same technique and assumptions as for Europe between 8000-2000 ybp. In this case the population growth is set as a function of time. The density is set to 0.02 until 400 A.D. as before and a population density in the FBL-region of 66.000. This values is based on the population growth of [1] in Western Europe, and the population density as estimated for this region between the 10^{th}-14^{th} centuries [8]. The value fits the population size as mentioned in [9] for the parts of Brabant and Limburg in the present Netherlands for the year 400 A.D. The density was probably larger in the most successful years of the Roman Empire [9], and went down to 66.000 when the Romans left the Low Countries. An increase and a later decrease of population has a similar genetic variability as a slow increase, as long as the starting and ending population sizes are the same, and the population growth values are independent of Y-DNA groups. In the case of a decrease of population and later strong increase, the variability in the population decreases. The used values for the population growth are 0.10 (400-1300 A.D.), 0.0 (1300-1500 A.D.), 0.15 (1500-1700 A.D.), 0.03 (1700-1800 A.D.), 0.20 (1800-1900 A.D.) and follow the values in [1], [9] and [10]. These values result in a population size of 5 million in the year 1900 (and 9.4 million in the year 2000 A.D.). The influx of people from foreign areas

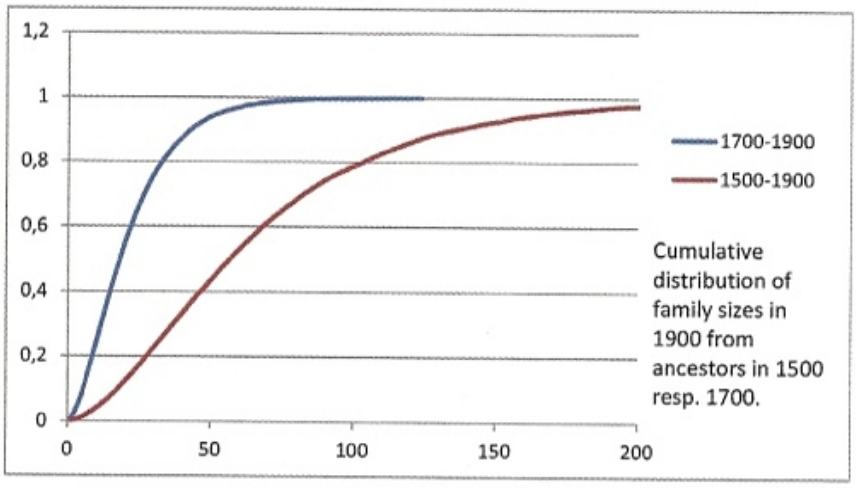
is expected to contribute significantly (up to 40%) in the period 0-800 A.D. (from the Rhine region) and less than 20% in later periods [9].

The simulations produced again a long list of a male line family trees, with the results shown in table 2. We have about 1270 Y-DNA ancestors in 400 A.D. and 16.500 in 1200 A.D; the number of inhabitants are 66.000 and 1 million respectively.

I picked randomly 1000 persons out of the simulated database. We determined the STR-distances between the 1000 picked persons, as was done in the measurements. This was done by a simulation of the individual STR mutations of each of the 37 markers. In report [7] only close matches of up to 6 STR-marker differences are reported, so we concentrated on close matches. This means automatically that the simulation only gives information on the short time scale (up to 2000 years). This means that a possible significant influx of Germanic people between 0-800 A.D. has limited influence on the statistics. Another systematic effect might influence the measured statistics. Family members might stimulate close relatives to join in the measurements. A set of equal surname members is present in the dataset. This set has some impact on the set of 0-1 STR-marker differences, and low impact on the 2-6 STR-marker differences. Similar Y-DNA with the same surnames is also expected, if we have random contributions to the dataset. The surnames in the FBL-region have their origin in 1200-1600, see [11], and represent a shared ancestor in the given timeframe. The measured and simulated STR-marker differences are given in table 3. The model fits the genetic variability that is measured, see table 3.

Small and large families that share surnames.

Another result from the simulation is the distribution of descendants from ancestors at the time surnames became common. Ann Marynissen (Universität von Köln, [11]) showed that the first surnames were given from fathers to children starting in 13th century in the FBL-region. In the cities of Brugge and Gent most citizens have surnames in the 14th and 15th century. In Brabant and Limburg the surnames started later than in Flanders. In Holland the surnames started in the 16th and 17th century and in the Dutch provinces of Groningen, Friesland and Drenthe the higher percentages of surnames are reached in the 18th and 19th century. If surnames were uniquely given at a moment in time, our simulations would give the distribution of the family sizes of the surnames. In figure 5, we show two cumulative distributions. The "male line family size" is defined as the number of males reaching 30 in a period of thirty years starting the year 1900, and who descended from the same male ancestor, who lived in the year 1500 respectively 1700. The blue line gives the cumulative distribution starting an ancestor in the year 1700. The expected value of the male line family size is 18. The 0.05-0.95 percentages are found at 3-53 men, which shows that both small size and large size families are expected. The red lines give the cumulative distribution for a shared ancestor in the year 1500. The expected family size is 57, and the 0.05-0.95 percentages are at 12-166 persons. People who have done genealogical surname research in the FBL-region will not be surprised by these numbers.



(Figure 5. Cumulative distribution of family sizes (males reaching 30 in a period of thirty years) in 1900 from ancestors in 1500 respectively 1700 for the region of Flanders, Brabant and Limburg.)

General conclusions

In this article I made simulations to show the effect of population growth to the genetic variability in the early years of European settlements and of the region of Flanders, Brabant and Limburg. From this we drew the following conclusions:

- The model is able to describe the genetic SNP-variability for the Europe (8000-2000 ybp) and STR-variability in Flanders, Brabant and Limburg (0 A.D.-present).
- The population growth is larger in the period 8.000-2.000 ybp than reported by [1]
- The present genetic STR-variation is consistent with a model that starts with a population size of 66.000 in Flanders, Brabant and Limburg in the year 400 and a population growth that is independent of the haplogroup.
- We can estimate the amount of Y-DNA ancestors in Flanders, Brabant and Limburg.
- The STR-variability between the R1b subgroups is as expected from the simple model.

Acknowledgements: I thank E.C.W.L. Marres MSc. and E. Maryniak MSc. for reviewing an earlier version of this article.

Tables

In the columns are given the following values: moment in time (1), the period that indicates the type of information we have (2), the growth rate as taken from the literature (3), the number Y-DNA ancestors (deduced in our simulation) (4), the number of potential fathers (as deduced from population size) (5) and the population size (as deduced from the literature) (6).

Table 1. The results of the simulation of male line family trees for Europe.

Period until	*Period*	*Growth rate*	*Y-DNA Ancestor*	*Potential Fathers*	*Population*
8000ybp	prehistory	0.02	2.565	135.000	600.000
7000ybp	prehistory	0.02	4.819	250.000	1.100.000
6000ybp	prehistory	0.02	9.382	470.000	2.100.000
5000ybp	prehistory	0.02	19.534	870.000	4.000.000
4000ybp	prehistory	0.02	45.003	1.670.000	7.600.000
3000ybp	prehistory	0.02	133.126	3.330.000	15.000.000
2000ybp	prehistory	0.02	6.450.000	6.450.000	29.000.000

Table 2. The results of the simulation of male line family trees for Flanders, Brabant and Limburg.

Period until	*Period*	*Growth rate*	*Y-DNA Ancestor*	*Potential Fathers*	*Population*
-7600ybp	-	-	5	400	4.800?
-5600ybp	prehistory	0.02	24	1.259	10.000?
-3600ybp	prehistory	0.02	116	3.932	23.000?
-1600ybp	prehistory	0.02	1270	14.500	66.000
-700ybp	estimated	0.10	16.500	255.996	1.000.000
-500ybp	estimated	0.00	31.024	255.967	970.000
-300ybp	estimated	0.15	91.994	388.233	2.520.000
-200ybp	Data	0.03	209.191	741.684	2.800.000
-100ybp	Data	0.20	1.072.291	1.072.291	5.040.000
-present			-	-	9.380.000

Table 3. The number of relationships in the dataset as a function of the number of STR differences. The tables gives the measured values of [7] and the simulated values with two values for the added mutation rates as mentioned in [7]. In the model the

individual mutations are simulated according to the model of [12], whereby for each individual marker the individual mutation rates are used. It includes chances to mutate back to the original STR-value. Two values for the added mutation rate are used in the simulations. This represents the range of these mutation rates as available in the literature, see e.g. [13].

STR differences	*# measured in [7]*	*# in simulation with mutation rate 0.0055*	*# in simulation with mutation rate 0.0045*
0	9	8	11
1	20	13	22
2	31	22	40
3	51	35	66
4	51	52	104
5	81	76	157
6	168	109	222

Notes and references

1 Klein Goldewijk, K. , A. Beusen, M. de Vos and G. van Drecht, 'The HYDE 3.1 spatially explicit database of human induced land use change over the past 12,000 years', *Global Ecology and Biogeography* 20(1) (2011): 73-86.DOI: 10.1111/j.1466-8238.2010.00587.x., see also http://themasites.pbl.nl/en/themasites/hyde/landusedata/landcover/index.html.

2 A collection of literature, 'Origins, spread and ethnic association of European haplogroups and subclades' (2013), collected at http://www.eupedia.com/europe/origins_haplogroups_europe.shtml.

3 Andrew Curry, Archaeology: 'The milk revolution', *Nature* 500, 20–22 (2013) doi: 10.1038/500020a, (http://www.nature.com/news/archaeology-the-milk-revolution-1.13471).

4 A collection of literature, 'Haplogroup R1b (Y-DNA)' (2013), collected at http://www.eupedia.com/europe/Haplogroup_R1b_Y-DNA.shtml.

5 A collection of literature, 'European haplogroups' (2013), collected at http://www.eupedia.com/europe/european_y-dna_haplogroups.shtml.

6 A collection of literature, 'Population values in the countries in Europe in the year 1800' (2013), collected at http://www.populstat.info/.

7 Marc van den Cloot, Hir. M.B.A., *DNA-Project 2010 België, exclusief oud hertogdom Brabant* (2010) (ISBN 9789080430525).

8 Paolo Malanima, 'Energy and population in Europe, The Medieval Growth (10th-14th Centuries)', *L'Europa come problema*, (2009).

9 J.C. Schalekamp, *Bataven en Buitenlanders*, (2009) (ISBN 9789073299412).

10 Leo Lucassen en Jan Lucassen, *Winnaars en Verliezers*, (2011) (ISBN 9789035136434).

11 Ann Marynissen, 'Waar, waar en wanneer zijn onze familienamen ontstaan' (2013), collected at http://users.telenet.be/ann.marynissen/waaromontstaan.html,

12 Sandy Paterson, 'TMRCA, an attempt at realistic pairwise TMRCA estimation' (2013), collected at http://www.tmrca.com.

13 Kaye N. Ballantyne et al., 'Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications', *The American Journal of Human Genetics*, (2010), doi:10.1016/j.ajhg.2010.08.006.